Political Science 599 – Fall 2005
Robert J. Franzese, Jr. (Rob)
(following Jake Bowers, following Nancy Burns, following John Jackson...)

Administrative Preliminaries

• **Instructions:** While we wait for everyone to arrive and settle into seats, please read the syllabus to see if you have questions, and then read this introductory note. Also, please add your contact and background information to the sheet circulating as it comes to you.

• **Class Meetings:** This class meets Tuesdays and Thursdays, 10:00-12:00, in 3451 Mason Hall (as you already know, apparently).

• **Office Hours:** My office hours are Tuesdays, 12:15-1:45 (in a Haven office TBA), and by appointment (in 4256 I.S.R.). The best way to contact me, and I encourage you to do so, is by email: franzese@umich.edu.

• **Your G.S.I.:** Your GSI is Bryce Corrigan (becorrig@umich.edu). In addition to guiding you adeptly and intrepidly through all the confusion I generate, he will lead section, which meets in 7603 Haven Hall, Thursdays, 1:00-2:00, and hold some office hours in **TBA, TBA**.

Graduate School, Political Science, & This Course

• Learning takes multiple exposures & practice.
  - No kicking yourself for not getting it on first read or first exposure.
  - No *math gene*; it’s just a language, one designed for expressing logic; it’s cumulative & you may be foreign or you may be rusty.
  - Don’t wait to seek help or ask questions.
  - My experience as an illustrative example...

• Be here (in Pol Sci Ph.d. Program) only if want to be. If so, then probably want to be in this class also.
– *Political Science* = *scientific* study of politics, meaning it endeavors to discover, expound, explain, and explore *systematic relationships* between aspects of the (socio-economic and) political world.

– I.e., *Political Science* develops theoretical propositions, primarily positive (i.e., not normative) ones about systematic relationships that should obtain b/w features of (socio-economic & political) world and, of course, evaluates those propositions empirically.

– This course begins (for most of you) your formal training in the methodology of such empirical evaluation.

• Meaning & nature of *smart*. Interested & interesting as important. Cultivate curiosity.

**Review Syllabus & Preview Course**

• **Textbooks**
  – No single book does all or works for all. Wonnacott and Wonnacutt our core text, but explore.
  – Most req’d readings in books available at Shaman Drum; those & rest on electronic or physical reserve.

• **Assignments & Documents**: Bring to class printed or send in Adobe PDF or equivalent.

• **Groups**
  – Work in groups, but, ultimately, submit your own work.
  – Grades increasingly irrelevant; only learning matters; need to know *you* can do it on own.
  – If copying b/c time or difficulty, *come see us for help* (see above).

• **Computing**
  – *Stata* = workhorse; some prog-ing perhaps, in *R*, to see more *1st hand*.
  – Be flexible:
∗ Want something u like, does what u want, & enough others use.
∗ I use E-Views, Stata as needed, Gauss for sims or prog-ing.
− Profit from department’s computing class. Exposure to key software and hardware. I don’t work with Stata too often, won’t be tremendously useful on Stata questions. Sorry.
− Word Processing:
  ∗ For some (not all) assignments, required that documents have professional-looking...
    ...equations, with greek letters & mathematical symbols;
    ...graphs, charts, tables w/ informative titles & headings, embedded where appropriate;
    ...bibliographies and appropriate citation;
    ...footnotes, etc., all with appropriate cross-referencing.
  ∗ Can do in Word, Wordperfect, LaTeX, or whatever.
  ∗ Pay close attention to how the articles and other materials we read are written, including how graphics and tables and equations are formatted, presented, and discussed.

• Data
  − Begin working with data starting next week.
  − Data almost never exist in exactly the right format and arrangement to begin analyzing immediately...
  − ...so not provided polished Stata file ready to use (=Lesson One).

Positive Political Science and the Practice of Data Analysis

• Two Kinds of Positive Questions
  − Factual: Descriptive, Historical, Photographic: What Happened? What is or was the truth? (History)
    ∗ Who voted for Bush, Gore, Nader, Buchanan? How many voted for...? Who actually won that election? For whom did Palm-Beach-county voters intend to vote?
* What share of some physical population (e.g., current U.S., people living in democracies since World War II, Chinese citizens 60 years ago) believes or did something?
* What was the relationship between money growth-rates and real interest-rates under the Bretton Woods system?
* What effect did the Ghent system have on early unionization? ...did telephone and door-to-door get-out-the-vote drives have on U.S. voters over the past 30 years?


- **The Differences Lie in the Goals, To What We Aim to Infer.**

• Some (hopefully) clarifying discussion:

Consider smoking & cancer: Determining whether smoking & cancer are related, & whether, in particular, smoking causes cancer, is a theoretical question. Discovering whether smoking & cancer are/were related in some particular group that’s observed (the folks in this or that study) or observable (say, physical population of the town, country, or globe actually living at some time) is a factual question. The scientific/theoretical question is whether smoking causes cancer, full stop (it does). The concepts of hyperpopulation & TRUTH are key here. Whether smoking & cancer happen to have been related in a specific observed or observable sample/population is a factual question (true history, if you will). So, historical science stops with one inferential step, to what is/was true (e.g., arson investigators, forensic pathologists: what caused this or that fire or death). Theoretical science adds an additional inferential step, from that historical truth to universal TRUTH (e.g., actuary, medical researcher: what causes fires or deaths). Of course, both are science, & it’s rather impossible, to do either sort without the other... Can’t de-
termine what caused this fire without some good theory about what causes fires, & can’t gain much empirical confidence in what causes fires without knowing what caused (a bunch of) specific fires. So it’s not a ‘this is better or higher than that’ kind of distinction, but a distinction in goals.

• **Empirical Leverage on Answers**: The practice of using empirical info to help answer important substantive Q’s of either sort involves:

  1. substantive & theoretical knowledge: nec. to ask meaningful Q’s of the evidence, to measure & compile evidence effectively;
  2. understanding principles stats & research design: nec. to know how infer powerfully & honestly from evidence the answers it suggests & in form comparable across aspects of Q’s, studies, & scholars;
  3. skill writing persuasive arguments (always);
  4. some adeptness with mathematics (usually) & skill programming computers (occasionally);
  5. ability to construct pro-looking doc’s & -sounding presents (always).

**Introducing Ourselves & Interests**

**Empirical Research Designs**

• **Types of Research Designs**
  
  – Experimental Study (Experiments):

    * What & How
      
      · *experiment* ≡ *exogenous control of treatment* [Why important?]
      
      · what is *treatment effect*? [E(τ_i|T=1)-E(τ_i|T=0), but...]
      
      · usually randomize treatment & control group membership [Why?]
      
      · Complications: what are *confounds*? [Other complications exist...]

    * Examples
      
      · Hutchings & Valentino
      
      · Morton & Lohmann
* Laboratory
  · **Strength**: Internal Validity; treat known-exog ⇒ known-causal
  · **Weaknesses**: multifarious, incl. external val. usu. (& ethics)

– Observational Study (Observations):
  * What & How [(comparative) history: for most part, all we have...]
  * Examples
    · Election outcomes, wars, governments, policies, etc.
    · Surveys (just another way of recording history)
  * Strength: External Validity; Weaknesses: multifarious, including internal validity usually.

– Field, ‘Natural’, or ‘Quasi-’ Experiments (mixture):
  * What & How
    · Field Experiments:
    · Natural Experiments:
    · ‘Quasi’ Experiments:
  * Examples
    · Field Exp: Gerber&Green; Wantchekon; *survey exp.* (TESS)
    · Natural Exp: Achen&Bartels’ sharks; ‘quake, T’Wave, Meteor
    · ‘Quasi’ Exp: i.e., ‘critical cases’ (≈ obs. study)
  * Strengths: mixture; Weaknesses: multifarious, including combination of above & some of own (e.g., feasibility and ethics).

• **How to Collect Data**
  – **Population**: Determine & Characterize
  – **Sample**: from Pop, from Existent/Feasible sample from Pop, or from Exper’ly-Gen’d sample from Pop
  – **Sample, How?**: Simple Random Sample (S.R.S.)
    * S.R.S. ⇒ Representativeness & Comparability ⇒ ability Infer to Pop from Samp relatively directly.
Perfect S.R.S. rarely obtainable & might be highly inefficient even if so ⇒ select sample strategically [how?] &/or feasibly ⇒ need more statistical theory & work to infer from Samp to Pop.

**Considering Some Preliminary Examples**

- **Do phone calls or door knocking 'get out the vote’ better?**
  - Hypothetical Field Experiment:
    * I phone everyone in my neighborhood and my colleague knocks on every door in his neighborhood?
    * Problems?
  - Gerber & Green’s Field Experiments:
    * Randomized sample 30,000 registered voters in New Haven.
    * Randomized application of canvassing, phone, or mail.
    * Large effect of canvassing, small for mail, none for phone.
    * **Concl:** Turnout ↓ partly due ↓ face-to-face pol mobilization.
    * Problems?
  - Observational Study:
    * How might you construct obs study to explore these Q’s? Probs?
    * What about a laboratory experiment? Problems?

- **(How Much) Do SAT Coaching Classes Help?**
  - Obs Study: Performance differences b/w class takers & non-takers
    * But class-takers differ in lots of ways from non-takers...how?...some relevant to SAT performance?
    * So what might you try to do about that?
  - How might you design a field experiment? Could you conduct one? What about a lab experiment?

- **How did they discover that cigarette smoking caused cancer?**
  - Simple obs comparison cancer rates among smokers & non-smokers.
– But smokes & non-smokes diff many ways, some cancer-relevant ⇒
– Controls... then theory & lab exp’s & more & better obs studies.

• **Methods We’ll Be Covering Used in Both Kinds of Studies**
  – Multivariate methods at end of term designed specifically to address confounding in observational studies.
  – Descript & hypoth-test meth’s we work through *en route* in both.

### Sampling

• **Sampling** b/c we do not usually observe all relevant observations.

• **Population** is term for “all the relevant observations.”

• **NOTE:** For theoretical Q’s (of *Truth*), population not obtainable even in principle. Concept/Notion of *Superpopulation* or *Hyperopulation*. For factual questions (of what is/was *true*), we could, in principle, observe the entire population to which we’re trying to infer.

• **So How Sample?**
    – If we’re observing comparative history (e.g., election outcomes across polities &/or over time), do we have random selection? [No.]
    – If we’re surveying people, can we have random selection? [One sense, yes; another, no. Depends on pop to which intend to infer.]
    – Even if could, random sampling might be devastatingly inefficient.
  – Alternatives?
    * Sampling Strategically: what might you want in a sample?
    * Stats Adjusts for certain patterns non-randomness. [See ps699]

• Example: See http://www.pages.drexel.edu/ pa34/ptn1.htm
  – Alf Landon (R) v. FDR (D)
  – *Gallup (N=5K)* v. *Literary Digest (N=10M w/ ≈2.5M response)*
Sample Size

- OK, random=good, but how big is big enough?

...The new poll found that a slight majority of registered voters – 53 percent – say Bush is more qualified than Kerry to be commander in chief, while 43 percent say they prefer the Democratic nominee. At the end of the Democratic convention, Kerry enjoyed an eight-point advantage over Bush on that question. Taken together, the results of the poll suggest that Bush’s recent gains have come from eroding perceptions of Kerry and not as a consequence of improved views of Bush’s performance as president.

“I like the way [Bush] has handled [Iraq] – he just did what he had to do, didn’t pussyfoot around,” said Joy L. Crockett, 52, a manicurist in Hammond, La. She said she worries that Iraq and the war on terrorism make it a bad time to change presidents and believes that Bush offers the best hope to “get the country back to better than it was.”

But others worried that Bush’s go-it-alone leadership style has isolated the United States from the rest of the world at precisely the time that the country needs help from its allies to stabilize Iraq and fight the international war on terrorism.

“He’s alienated the U.N. – if anyone in government thinks they’re going to get any countries like France to pay for some of this, I want some of what they’re smoking. They’ll be laughing up their sleeves,” said William Thomas, 66, a retired electrician who lives in a suburb of Cleveland.
A total of 1,207 randomly selected adults was interviewed Aug. 26-29, including 945 registered voters and 775 likely voters. The margin of sampling error for the subsample of likely voters is plus or minus three percentage points; it is slightly smaller for all voters.

- So, what’s this article saying about distribution of Kerry & Bush supporters?
  - 1 to 1, 50%-50%, Joy C. to William T.? \((\text{Even-handedness} \text{ biases impressions toward 50-50, huh?})\)
  - No, obv’ly: says 53%-43%. Crockett&Thomas just e.g.’s, not just small sample but \textit{unrepresentative} one. [Selected on dep var.]
  - How certain is this result (sample) as estimate of what’s true about US voting public on that day (population): a factual Q.
  - Somehow sense that, say, surveying 10 people, even randomly, & finding, say, 60-40 not give comfortably reliable pic US voting pub,
  - but 1,207 enough? How certain pic that offer (assume random)?

- Simplify e.g.: Say 55-45 (so total 100%). Is 45%-55% a reliable difference in a sample this size (1207)? Clarify Q: If took many more samples of size 1207 from this pop, how frequently results near 45-55?

- **Using Computer Simulation to Answer Such Questions**
  - Computer Code (in R) for the simulation:
    ```r
    > n1207 <- c(rep(0, 0.45 * 1207), rep(1, 0.55 * 1207))
    > n1207results <- vector(length = 1000)
    > for (i in 1:1000) { n1207results[i] <- mean(sample(n1207, size = 1207, replace = TRUE)) }
    ```
  - This code is ‘resampling’, with replacement, from the sample 1000 times, thereby treating the original sample as the population from which we sample, which is safe, i.e. allows inference to the population if the original sample is \textit{representative}. What’s this look like?
    ```r
    > classnames <- read.table("classdata.csv", sep = ",", header = TRUE)
    > firstandmiddle <- sapply(strsplit(as.character(classnames$Name), ","), function(x) x[2])
    > firstnames <- sapply(strsplit(firstandmiddle, " "), function(x) x[1])
    > namesamps <- data.frame(origname = sort(firstnames))
    ```
> for (i in 2:6) { namesamps[[i]] <- sort(sample(namesamps$origname, replace = TRUE)) }
> print(namesamps)

<table>
<thead>
<tr>
<th>origname</th>
<th>Samp1</th>
<th>Samp2</th>
<th>Samp3</th>
<th>Samp4</th>
<th>Samp5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aleksandra</td>
<td>Aleksandra</td>
<td>Aleksandra</td>
<td>Andrea</td>
<td>Aleksandra</td>
</tr>
<tr>
<td>2</td>
<td>Andrea</td>
<td>Daniel</td>
<td>Aleksandra</td>
<td>Andrea</td>
<td>David</td>
</tr>
<tr>
<td>3</td>
<td>Daniel</td>
<td>David</td>
<td>Andrea</td>
<td>Hemanth</td>
<td>David</td>
</tr>
<tr>
<td>4</td>
<td>David</td>
<td>David</td>
<td>Daniel</td>
<td>Andrea</td>
<td>Jane</td>
</tr>
<tr>
<td>5</td>
<td>Hemanth</td>
<td>David</td>
<td>Hemanth</td>
<td>Daniel</td>
<td>Jennifer</td>
</tr>
<tr>
<td>6</td>
<td>Jane</td>
<td>Hemanth</td>
<td>Hemanth</td>
<td>Hemanth</td>
<td>Katherine</td>
</tr>
<tr>
<td>7</td>
<td>Jennifer</td>
<td>Hemanth</td>
<td>Jane</td>
<td>Maria</td>
<td>Jennifer</td>
</tr>
<tr>
<td>8</td>
<td>Jiaan</td>
<td>Jennifer</td>
<td>Jiaan</td>
<td>Katherine</td>
<td>Maria</td>
</tr>
<tr>
<td>9</td>
<td>Katherine</td>
<td>Jennifer</td>
<td>Krysha</td>
<td>Maria</td>
<td>Jiaan</td>
</tr>
<tr>
<td>10</td>
<td>Krysha</td>
<td>Katherine</td>
<td>Maria</td>
<td>Maria</td>
<td>Katherine</td>
</tr>
<tr>
<td>11</td>
<td>Maria</td>
<td>Krysha</td>
<td>Matias</td>
<td>Michael</td>
<td>Krysha</td>
</tr>
<tr>
<td>12</td>
<td>Matias</td>
<td>Maria</td>
<td>Michael</td>
<td>Michael</td>
<td>Krysha</td>
</tr>
<tr>
<td>13</td>
<td>Michael</td>
<td>Maria</td>
<td>Michael</td>
<td>Michael</td>
<td>Krysha</td>
</tr>
<tr>
<td>14</td>
<td>Michelle</td>
<td>Michelle</td>
<td>Michael</td>
<td>Michio</td>
<td>Matias</td>
</tr>
<tr>
<td>15</td>
<td>Michio</td>
<td>Michelle</td>
<td>Papia</td>
<td>Michio</td>
<td>Michio</td>
</tr>
<tr>
<td>16</td>
<td>Papia</td>
<td>Michelle</td>
<td>Papia</td>
<td>Papia</td>
<td>Michio</td>
</tr>
<tr>
<td>17</td>
<td>Patrick</td>
<td>Michelle</td>
<td>Papia</td>
<td>Rachel</td>
<td>Patrick</td>
</tr>
<tr>
<td>18</td>
<td>Rachel</td>
<td>Michio</td>
<td>Patrick</td>
<td>Rachel</td>
<td>Shabana</td>
</tr>
<tr>
<td>19</td>
<td>Shabana</td>
<td>Patrick</td>
<td>Patrick</td>
<td>Shabana</td>
<td>Shabana</td>
</tr>
<tr>
<td>20</td>
<td>Tarah</td>
<td>Rachel</td>
<td>Shabana</td>
<td>Tarah</td>
<td>Rachel</td>
</tr>
<tr>
<td>21</td>
<td>Thomas</td>
<td>Tarah</td>
<td>Thomas</td>
<td>Uzma</td>
<td>Shabana</td>
</tr>
<tr>
<td>22</td>
<td>Uzma</td>
<td>Valenta</td>
<td>Thomas</td>
<td>Valenta</td>
<td>Valenta</td>
</tr>
<tr>
<td>23</td>
<td>Valenta</td>
<td>Valenta</td>
<td>Valenta</td>
<td>Valenta</td>
<td>Tarah</td>
</tr>
</tbody>
</table>

• Back to Bush-Kerry poll-e.g. *Quantiles* of resulting distribution:

```r
> quantile(n1207results, prob = c(0, 0.05, 0.5, 0.95, 1))

       0%      5%     50%     95%    100%
0.5045568 0.5252278 0.5492958 0.5724938 0.5932063
```

• Let’s draw those 1000 samples of size 1207 again (simulation results change, but only slightly):

```r
> for (i in 1:1000) { n1207results[i] <- mean(sample(n1207, size = 1207, replace = TRUE)) }
> quant1207 <- quantile(n1207results, prob = c(0, 0.05, 0.5, 0.95, 1))
> print(quant1207)

       0%      5%     50%     95%    100%
0.5053853 0.5269263 0.5492958 0.5733223 0.5915493
```

• OK, taking 1K rndm samps size 1207, about 90% of samples have b/w approx 53% & 57% pro-Bush (& 47% and 43% pro-Kerry).

  – How does this help answer Q about comparison b/w 45% & 55%?
  – If 1207 sample truly rep, then can sample from *it* as if were pop.
So, although it remains possible that only 50% would support Bush in any one particular sample of size 1207, that is extremely unlikely.

None (or <0.5% of first 1K rndm samps we drew, had such low %)]

• Not quite our question. Actual question: whether 55%-45% in our one survey is credible evidence, & how credible, of a true gap. So redo for difference between the two numbers.

• What if sample size were cut to 100 from 1207? More simulations!

  Code:
  ```r
  > n100 <- c(rep(0, 0.45 * 100), rep(1, 0.55 * 100))
  > n100results <- vector(length = 1000)
  > for (i in 1:1000) { n100results[i] <- mean(sample(n100, size = 100, replace = TRUE)) }
  > quant100 <- quantile(n100results, prob = c(0, 0.05, 0.5, 0.95, 1))
  > print(quant100)
  
  Quantiles:
  0% 5% 50% 95% 100%
  0.40 0.46 0.55 0.63 0.69
  
  Let’s look at the entire distribution of results (across the 1000 hypothetical samples of size 100)
  ```

  ```r
  > par(mfrow = c(1, 2), pty = "s", mar = c(2, 2, 1, 0), oma = c(0, 0, 0, 0), mgp = c(1.5, 0.5, 0))
  > hist(n100results, breaks = 50)
  > abline(v = quant100[c(2, 3, 4)], lwd = 2)
  > text(quant100[c(2, 3, 4)], -1, round(quant100[c(2, 3, 4)], 2), font = 2)
  > hist(n1207results, breaks = 50)
  > abline(v = quant1207[c(2, 3, 4)], lwd = 2)
  > text(quant1207[c(2, 3, 4)], -1, round(quant1207[c(2, 3, 4)], 2), font = 2)
  ```
• **Answering Same Q Using Analytic Formula**

• much quicker and easier (and more precise if our assumptions (what assumptions?) are correct)!

  - 95% **confidence interval** of a proportion is:
    \[
    \pi = P \pm 1.96 \sqrt{\frac{P(1 - P)}{n}} \quad (1)
    \]

  - Mechanically (for now; theory later): \( P \equiv \) sample proportion, \( \pi \equiv \) population proportion (that we don’t observe), \( n \equiv \) sample size, & the 1.96 has something to do with 95% (we’ll get to this later...).

  - So, we’ve got:
    \[
    \pi = .55 \pm 1.96 \sqrt{\frac{.55(1 - .55)}{1207}} \quad (2)
    \]
    \[
    .55 \pm 1.96 \sqrt{\frac{.55(1 - .55)}{1207}} \quad (3)
    \]
    \[
    .55 \pm 1.96 \sqrt{\frac{.2475}{1207}} \quad (4)
    \]
    \[
    .55 \pm 1.96 \sqrt{.00021} \quad (5)
    \]
    \[
    .55 \pm 1.96 \cdot .0145 \quad (6)
    \]
    \[
    .55 \pm .028 \quad (7)
    \]
Or, for the sample-size n=100:

\[
\pi = .55 \pm 1.96 \sqrt{\frac{.55(1-.55)}{100}}
\]  

(8)

\[
.55 \pm 1.96 \sqrt{\frac{.55(1-.55)}{100}}
\]  

(9)

\[
.55 \pm 1.96 \sqrt{\frac{.2475}{100}}
\]  

(10)

\[
.55 \pm 1.96 \sqrt{.002475}
\]  

(11)

\[
.55 \pm 1.96 \cdot .04975
\]  

(12)

\[
.55 \pm .0975
\]  

(13)

So, how do we interpret this \textit{confidence interval} thing?

Precisely=Tricky: Depends on whether you’re a classical or Bayesian.

− Would like to say: 95% confident that true % supporting Bush is between 52% and 58%. Bayesian basically can, but have to give some other info first.

− Classically, can only say “actual % supporting Bush would be w/in 3± percentage pts of 55% in 95 of 100 samples of size 1207, & within about 10 pts of 55% in 95% of samples of size 100.”

So, go back to \textit{Wash Post} art & reread that bit about “sampling error”

\textbf{Note: Choosing a sample size involves solving (1) for n given a substantively meaningful choice of P. For P=.5, the 95\% c.i. size}

\[
= \pm 1.96 \sqrt{\frac{.5(1-.5)}{n}} = \pm 1.96 \sqrt{\frac{.5\cdot 5}{n}} = \pm 1.96\cdot \frac{5}{\sqrt{n}} = \pm \frac{98}{\sqrt{n}}.
\]

Let’s plot that.

> par(pty = "s", mar = c(3, 2, 1, 1), oma = c(0, 0, 0, 0), mgp = c(1.5, 0.5, 0))
> n <- seq(10, 10000, 10)
> m <- 0.98/sqrt(n)
> plot(n, m, type = "l", ylim = c(0, 0.1))
> abline(v = c(100, 200, 400, 800, 1600, 3200), lwd = 0.5, col = "gray")
• We’ll return to consider this more later.

**Other Examples: Observational Studies, Theoretical Questions, Regression Analyses**

• Democracy $\leftrightarrow$ Development

---

- Hypotheses?
- Problems? [Direction Causality, Selection, Confounds, Outliers, ...]
- What about those outliers? More Hypotheses?
- **Note:** Theories (theoretical hypotheses) don’t have proper nouns or proper adjectives.
- This was generated in SPSS, by the way.

- **Government Partisanship ⇒ (⇐?) Size of Government**

```
CTRYS GovPart Spend
US  6.16   22.83
JA  8.86   16.39
GE  5.71   30.44
FR  5.82   42.95
IT  5.09   38.70
UK  6.06   34.56
CA  5.09   21.99
AU  4.75   40.16
BE  5.88   50.29
DE  5.21   39.49
FI  5.13   30.40
GR  7.52   40.11
IR  6.00   38.74
NE  5.68   54.86
NO  3.99   39.42
PO  8.45   41.53
SP  8.38   23.70
SW  3.64   40.02
SZ  5.78   8.98
AL  6.38   25.79
NZ  5.64   42.38
```

\[ \text{GovPart} = 1950-87 \text{ average of an index of left-right (0-10) partisanship of government} \]
\[ \text{Spend} = 1988 \text{ total government expenditure as a percent of GDP} \]
These tables & figures were generated in Lotus 1-2-3.

**Interpreting Regression in 5 Min** (so can read results encountered in substantive courses starting now):

- Output gives *eqtn* or *model* where LH$S$, “Spend”=*dep.var.* (effect) & RH$S$=*indep.vars.* (causes), “GPart”, “NPart”. (We hope/claim cause & effect...)

- Estimated *coefficients* on those indep.vars. tell how much dep.var. tends to move for 1-unit increase in independent variable. Sign tells direction of relationship & size tells magnitude (given substantive scales independent & dependent variables). [So, the above estimated equation is...] (Coefficients ⇔ Effects/Relationships so directly only for linear-additive models: $y=mx+b$.)

- Next row contains estimated *standard errors* of coefficients. They gauge the *(im)precision* or *(un)certainty* of estimated coefficient/relationship. Can read these loosely like ± number given in polls. E.g., “when average govt partisanship increases (right) by 1, govt spending tends to decrease by about 2 % GDP [coefficient
on GPart] ± (give or take) 1.9 [standard error on that coefficient].”

- Next row gives **t-stats**. Std Errs at very least < coefficient for any credence coeff not =0. Prefer s.e. not > fract1/2 as large ⇒ |t| ≥ 2. Most common t-stat is coefficient divided by std err.

- Next row is probability, under certain assumptions, of having estimated a coefficient so far or farther from zero if *true relationship* were zero. These **test** whether relationship (positive or negative) exists with acceptable certainty. We like these **p-levels** or **α-levels** to be near or < 0.10.

- This is **Multiple Regression** (which we begin to cover at end of term). Like scatter-plots for 2 variables, but gives relationship between set of possible independent variables & dependent variable. Thus, each coefficient is “effect of X on Y, holding all else constant” or “effect of X on Y after having ‘netted out’ all relationships between other x’s & Y.”

- The $R^2$ term is just square of correlation between predictions based on estimated equation and actual outcomes. Thus, it summarizes how well set of independent variables predict dependent variable. What share of total variation in dep var explicable by the model, i.e., linear-additively by the indep vars.

**Choosing a Sample Size**

- Example research problem: Prof’s Cara Wong & Nick Valentino want to know how much the newspapers are talking about genetics, and if they’re doing so in the context of race.

  - What is their population?
  - Should they take every single newspaper published over that period and read the whole thing? Or what?
  - Oh! They should sample? [Cool!] How should they select their sample? At random? [Cool!]
– How big of a sample should they draw? Depends on how much information they think they will need to say something meaningful.

– Let’s talk in proportionate terms: do they want to determine with, say, 95% confidence, a ±1% pt share of articles mentioning genetics? And a, say, ±1% pt of those mentioning race in that context?

• We learned how to answer this sort of question already. Cool!

– Option 1: Simulate and Resample:
  * Choose a 1% difference in a plausible range, like 4.5% and 5.5% of articles mentioning genetics;
  * Generate a fake dataset with this characteristic (P=.05) and sample size, say, 100;
  * Resample 1K times & see where 95% of results fall. Do they overlap 4.5% and 5.5%? If so, then try sample size 200, & so...

– Option 2: Solve Analytically:
  * We know the 95% c.i. will be \( \pi = P \pm 1.96 \sqrt{\frac{P(1-P)}{n}} \)
  * So, the relevant part here is \( \pm 1.96 \sqrt{\frac{P(1-P)}{n}} \)
  * Which, for P=.05= \( \pm 1.96 \sqrt{\frac{.05.95}{n}} = \pm 1.96 \sqrt{\frac{.0475}{n}} = \pm \frac{1.96 \cdot 2179}{\sqrt{n}} = \pm \frac{4272}{\sqrt{n}} \Rightarrow \)
  * We want \(.005 = \frac{4272}{\sqrt{n}} \Rightarrow n = \left(\frac{4272}{.005}\right)^2 \approx 8,938 \) obs (newspapers)!

– So, Cara & Nick have got some serious reading to do! And that was just for gauging with 95% confidence within 1% pt the percentage of newspaper articles that mention genetics, assuming that percentage about 5%. To ascertain with analogous 95% confidence within 1% pt the percentage of those genetics-mentioning articles that also mention race in that context, assuming that to be, say, 20% of genetics articles, they’d need \( n = \left(\frac{1.96 \sqrt{.2 \cdot .8}}{.05}\right)^2 \approx 24,586 \) genetics articles (!!), which, if they’re about 5% of total (1 per 20) as assumed, means 20(24,586)=491,725 newspapers! Maybe they
should hire some grad students, &/or use some automated hardware & software for textual analysis, &/or lower their sights for statistical certainty. How might they lower their sights?

- Some general lessons:
  - How does required \( n \) relate to \( P \)?
    * The formula (function) \( n = \left( \frac{1.96 \sqrt{P(1-P)}}{.005} \right)^2 \) increases \( P(1-P) \).
    * So, when is \( P(1-P) \) max’d? For \( P = .5 \). [How did I get this?]
    * So, need more observations for a given degree of certainty (95%) for a given size of confidence interval (±1% pt) as \( P \to .5 \).
  - How does required \( n \) relate to the desired certainty (confidence)?
    * As we demand greater certainty (say 99% confidence), that 1.96 grows (to 2.576), so \( n \) must be higher.
    * If we will accept lower certainty (say 90%), that 1.96 shrinks (to 1.645), so sufficient \( n \) is lower.
  - How does required \( n \) relate to acceptable width of that conf. int.?
    * The denominator in above formula is .5(width) c.i., so, as it shrinks (smaller range desired), \( n \) grows & v.v..
  - How does variability of our estimates of \( P \) respond to sample size?
    * Does it shrink linear-proportionately? I.e., e.g., if we double \( n \), do we halve the ±?
    * No. Notice it’s not \( n \), but \( \sqrt{n} \) in denominator of c.i. formula.
    * So, variability of \( P \) estimates ↓ in \( \sqrt{n} \), i.e., e.g., variability halved when double \( \sqrt{n} \) (4=twice2, but \( \frac{1}{2} \sqrt{4} < \sqrt{2}; \frac{1}{2} \sqrt{8} = \sqrt{2} \)).
    * I.e., we must *quadruple* the sample size to *halve* the variability. More generally, we must multiply the sample sample-size \( n \) times to reduce variability by a (multiplicative) factor of \( \frac{1}{\sqrt{n}} \).
    * Recall our graph of this relationship from before:
Work & Think Through Exercises on Samples & Causality

- Samples: WW #1-3, p. 10. In each case, ask...
  - What is the population to which the researcher would like to infer?
  - Is this a factual question or a theoretical one?
  - Is the sampling strategy likely to yield a representative sample of that population or hyperpopulation?

- Causality: WW #1-9, pp. 17-18. For each of these (rhetorical) cases...
  - ...the alternative causal directions have opposite sign effect. Need not be so. What would happen if similar $Y \Leftrightarrow X$ relation and both negative or positive?
  - ...what other kinds of info would you need to help sort this out?
Descriptive Statistics

- What’s a **Variable**?
  - We’ll be talking vars a lot. (Used word 12 times already, not counting abbreviations.) Maybe we should define them clearly?
  - One good def *Political Science*: *study of systematic relationships b/w (socio-econo-)political variables*, so better get this down!

  - Definition: For present purposes (can be more math’ly precise later),
    - A **variable** is column of data in a dataset, perhaps labeled $X$ or $Y$ or something else clever, that has more than one value, or
    - An algebraic ‘placeholder’, perhaps labeled $X$ or $Y$ or something else clever, that could have more than one value.

- Example: A variable in a dataset capturing a respondents’ Bush support (like in our previous polling example) would look like a long column of 1s and 0s (each row being a respondent).

- A **constant**, by the way, is the clever name for a variable that doesn’t vary (e.g., only 1s or only 0s).

- **Summarizing Variables**
  - Motivation:
    * We’ve seen already that we’re prob’ly going to need lots of obs on vars to obtain acceptable certainty in our empirical evaluations of these theories about relations b/w variables,
    * So, we’re going to have (at least we hope so!) big, long columns of 0s and 1s (or other values) like this: $01010010101010010101001010000011...$
    * So, we’re going to want to be able to summarize all this info (respondent 1 dislikes Bush, 2 likes, 3 dislikes, etc. etc.) somehow.
  - Proportions:
    * We’ve summarized already. I said 55% support Bush & you knew I meant “proportion of 1s is 55%”.

Fall 2005 – Rob Franzese
* What else did we give in summary without even realizing it? n
What was that?
* For simple, **binary** variables (i.e., \{0,1\}), this summary gives you almost all the info:
  * If told 75% of a sample of 100 people supported Bush, that’s 75 ones and 25 zeros.
  * Only missing info on ordering of 1s & 0s (which unimportant if random, anonymous sample).
– What if a variable has...[wait for it]...more than two values?! **Discrete** or, even...dare I say it?...**Continuous**?!
– Yeah, we’re going to want some other ways of summarizing too.
  Here’re two other useful concepts:

- The **Distribution** of a variable refers to its (range of) possible values and the probability or likelihood or **Density** of those values in that distribution.
  – Say I have a variable with 7 values (Likert scale, e.g.) for 1000 respondents.
  – Could graph it, but this just one dimension: 1000 dots on 1 line, at 1,2,3,4,5,6,7, but repeated values just overprinting repeatedly. Not very helpful.
  – How about a **Frequency Table**?

<table>
<thead>
<tr>
<th>Value</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
</tr>
</tbody>
</table>
Such a table gives the *density* of the variable, so a very nice start.

A bar chart of same info even more informative in some ways; called a *density plot*:

- Height (or length) of bars gives frequency or density of that x-axis (or y-axis) value.
- So, here we have 1 bar of height (frequency, density) 500 over the value 3, 4 of height 100 over values 1,2,4, & 5, and 2 50-high over 6 & 7, like so (2 seconds in Lotus):

![Bar chart example](image)

What if 100 or 1000 values, or 1,000,000? ⇒ *histogram*

- Like a bar chart, but grps values into ranges called *bins*.
- For histograms in our sample-size examples, I spec’d 51 bins. More bins gives finer-grained, smoother, more continuous look to distribution’s density plot, & fewer gives lumpier one.
- What if just 1 bin? What if $n$ bins? How would these look?
- Can do histograms/density plots with bar-heights of # or % (frequency, proportion) of observations w/ that (those) values.

Let’s consider real-world example, verbal IQ scores of sample of 8th-graders in Netherlands, plotted at several granularities bin-sizes).

### A Unimodal, Symmetric Distribution

```r
summary(nlschools$IQ)
Min. 1st Qu. Median Mean 3rd Qu. Max.
4.00 10.50 12.00 11.83 13.00 18.00
```

```r
table(nlschools$IQ)
4 4.5 5 5.5 6 6.5 7 7.5 8 8.5 9 9.5 10 10.5 11 11.5
```
max(table(nlschools$IQ)) # Mode is 12, i.e. most frequent value
280

table(nlschools$IQ)[table(nlschools$IQ)==max(table(nlschools$IQ))]
12
280

par(mfrow=c(1,3), pty="s")
hist(nlschools$IQ, main="", sub="8th Graders in the Netherlands", xlab="Verbal IQ", breaks=2)
hist(nlschools$IQ, main="", sub="8th Graders in the Netherlands", xlab="Verbal IQ")
hist(nlschools$IQ, main="", sub="8th Graders in the Netherlands", xlab="Verbal IQ", breaks=30)
mtext(outer=TRUE, "Verbal IQs", side=1, line=-1)

• Discuss the tradeoffs involved in choice of bin-size.
• Let’s consider 2 other vars & their empirical density functions.
  – A density (or distribution) function is a mathematical formula that tells you how many observations are over any single value.
  – The familiar, bell-shaped normal curve is a probability density function (not an empirical one), so it’s heights are probabilities (strictly: probability densities) not observations.
The IQ’s above look roughly normal, judging by their empirical density functions, but our histograms are of empirical observations. This is (some of) what we use them for. To see how well some empirical observations approximate a theoretical truth or proposition.

**A Bimodal Distribution**

data(geyser) names(geyser)
> names(geyser)
[1] "waiting"  "duration"

summary(geyser$duration)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  0.8333 2.0000 4.0000 3.4610 4.3830 5.4500

table(geyser$duration) length(unique(geyser$duration)) [1] 118

par(mfrow=c(1,3),pty="s")
hist(geyser$duration,main="",xlab="Duration (Minutes)",breaks=2)
hist(geyser$duration,main="",xlab="Duration (Minutes)"")
hist(geyser$duration,main="",xlab="Duration (Minutes)",breaks=200)
mtext(outer=TRUE,side=1,"Eruption Times of Old Faithful",line=-1)

Eruption Times of Old Faithful
A Skewed Distribution

data(UN) names(UN) [1] "infant.mortality" "gdp"

summary(UN$gdp)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
  36   442  1779  6262  7272 42420   10

length(unique(UN$gdp)) [1] 189
table(UN$gdp)[table(UN$gdp)==max(table(UN$gdp))] 321

hist(UN$gdp,main="GDP of Nations in 1998",xlab="GDP (1998 US Dollars)",breaks=2)


GDP of Nations in 1998

• Often we want to summarize with 1 #, like, say, just give ‘middle’ of distrib. Often good & useful & parsimonious, but sometimes insufficient or misleading (bimodal & skewed distributions above?).

• Several useful such summary or descriptive statistics exist:

• Summaries (statistics) of ‘middle’ location or central tendency: mean, median, mode.
  – Formula for (sample) mean:

\[
\bar{x} = \frac{1}{n} \sum x
\]
• Summaries spread, variability: IQR, variance, std dev.

Measures of ‘Location’ or ‘Central Tendency’

• Review of ‘Stuff’ Mentioned So Far:
  – ‘Philosophical’ Underpinnings of Political Science
    * Positive Political Science
    * Empirical Methods & Logic of Empirical Evaluation
  – Sampling, random select, rndm assign (in Exps.)
  – Inference, w/ rndm assign, w/o rndm assign,
  – Confidence Intervals and Uncertainty
    * Mostly presaged b/c not even talked about probability yet!
  – Resampling & bootstrapping (also advanced material, but not conceptually more difficult than rest).
  – Density of a data distribution
    * Likely, common, frequent var vals, where dist piles up (or not).
    * Key features of such densities:
      · Number of peaks or modes
      · Skewed or symmetric Distributions
        · Display emp dists: histograms; bin-width choice & impl’s
          – Obv’ly, we’ll come back to these in more rigorous way soon enough.

• Statistics that characterize such distributions of data:
  – Measures of location, specifically of central tendency.
  – Measures of spread, e.g. variance.

• Let $x \equiv$ a variable; then,
  
  **Mean:** $\bar{x} \equiv \frac{1}{n} \sum x$
  
  **Median:** value of $x$ such that 50% of realizations of $x \geq$ & 50% of $\leq$ than this value.
Given order statistics where \( y_1 = \min x, \ldots, y_n = \max x \), median = \( Y_{(N+1)/2} \) if length of \( x \) = odd, \& \( \frac{1}{2}(Y_{N/2} + Y_{1+N/2}) \) if length of \( x \) even.

**Mode:** The most frequently occurring value of \( x \).

- **EXAMPLE & PRACTICE:**
  - WW number 2-1, page 30
  - What operation do you need to do to get a median?
  - What about mean?
  - How do mean & median relate to each other?
  * See previous distributions (unimodal & bimodal; skewed & symmetric; etc.)... Also,

\[
\begin{align*}
&> \text{skew1<-c}(1,95,96,97,98,99,100) \\
&> \text{skew2<-c}(1,2,3,4,5,6,100) \\
&> \text{summary(skew1)} \\
&\quad \text{Min. 1st Qu. Median Mean 3rd Qu. Max.} \\
&\quad 1.00 \quad 95.50 \quad 97.00 \quad 83.71 \quad 98.50 \quad 100.00 \\
&> \text{summary(skew2)} \\
&\quad \text{Min. 1st Qu. Median Mean 3rd Qu. Max.} \\
&\quad 1.00 \quad 2.50 \quad 4.00 \quad 17.29 \quad 5.50 \quad 100.00
\end{align*}
\]

**Spread**

- **Spread:**
  - What is it? [Dispersion]
  - Draw examples of data distributions low & high dispersion.
  - How about some substantive questions regarding means? Regarding variances? See, both can indicate interesting stuff.
  - What might be some good measures of spread?

- **Inter-quartile Range (IQR):** \( x_{75^{\text{th}} \text{percentile}} - x_{25^{\text{th}} \text{percentile}} \)
  - What are percentiles?
So what does IQR mean?
Notice that IQR & Median are related. Both defined by **quantiles** of data dists. [Which quantile is the median?]
Quantiles another way to talk about order stats, but rather than “smallest, next smallest, etc...,” talk about number obs below or above or b/w values. I.e., another way to talk about histograms.

• How about some summary typical or avg distance each obs from mean?
  E.g., subtract each obs from mean, & sum those distances; like total distance b/w obs & center. Call this statistic **Francine**.

\[
\text{Spread Measure Francine} \equiv (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_1 + x_2 + \ldots) - (\bar{x} + \bar{x} + \ldots)
\]
\[
\sum x - n\bar{x} \quad (16)
\]
but \( \bar{x} \equiv \frac{1}{n} \sum x \) so
\[
\sum x - n\left(\frac{1}{n} \sum x\right) \quad (18)
\]
\[
0 \quad (19)
\]

• Do problem 2-15 (p 39):
  a. Calculate the mean of 3,7,8,12,15: \((3+7+8+12+15)/5=9\).
  b. Calc deviations from mean: -6 -2 -1 3 6, Avg Deviation: 0
  c. (skip)
  d. already did it.

• So, **Francine** is always 0. Drat. Try absolute values? Squares?

  - **Mean Absolute Deviation**: \( \frac{1}{n} \sum |x - \bar{x}| \)
  - **Mean Squared Deviation**: \( \frac{1}{n} \sum (x - \bar{x})^2 \)

  * Similar to the **Variance** of a sample: \( \frac{1}{(n-1)} \sum (x - \bar{x})^2 \)
* Standard deviation ≡ \sqrt{\text{variance}} of a sample.

* So, what was diff? *Sample v. Pop.* Divide by \( n - 1 \) or \( n \)?
  
  · Where does this \( (n - 1) \) come from? **Degrees of Freedom**, in this case, of the variance. Represents amount of info available — \( n \) observations, less info used.
  
  · If had just 1 obs, what spread mean? Nada. Need at least 2.
  
  · So must first calc (sample) mean, then calc dev from it.
  
  · That one mean calculation “consumes” one piece of information, b/c, once have mean & n-1 obs, that last obs could only have one possible value for that mean to be accurate.
  
  · So, when calc mean, have \( n \) free pieces of info to start, when calc variance etc., already used 1, so \( (n - 1) \) remain.

• Some More Practice & Example w/ Std. Dev. Measure of Spread.

  – National Health & Nutrition Examination Survey 1976-80
    
    * 20,322 Americans aged 1 to 74. We’ll talk about height.
    
    * What question should you be wondering when I describe sample to you? [representative?]
  
  – 6,588 women. Avg height (i.e. (sample) mean) 63.5” & s.d. 2.5”.
    
    * Tells us most women around 63.5” tall, but some deviations of course. [what are you assuming about shape of histogram?]
    
    * “The SD says how far away numbers on a list are from their average. Most entries on the list will be around one SD away from the average. Very few will be more than two or three SDs.”
    
    * So what does an SD of 2.5 inches say?

  – Why std dev instead of variance or IQR? Mostly not matter, but
    
    * If normal dist, then ±s.d. = bounds w/in which about 68% of cases fall, & ±2s.d. = bounds w/in which about 95% Ah Hah! So that 1.96 number we saw before, kinda close to 2, & Rob mumbled something about 95% when we talked about that formula. Hmmmm. The plot thickens...
ASIDE: In world of normal distributions, mean & variance are all you need to know. Called *moments* of the dist. I’m not sure why, but mean & variance are first two *moments*. Third’s skew. Then kurtosis, which is fourth, and has something to do with one’s breath, I assume, and so on, I assume. There’s something called a moment-generating function for a dist. I don’t know anything about them either, other than that they exist. So lots of terms: *distribution* of data, *normal distribution*, which is not like a *boring* or *plain-old* or *ordinary* dist in contrast to an *abnormal* one. The normal distribution is a conceptual tool. It is theoretical and was discovered and some good reasons to suppose some things normally distributed (something called *Central Limit Theorem*), but often just a useful abstraction. One cool thing about it (among many many cool things) is that it is completely characterized by these two parameters, a mean and a variance.

ASIDE 2: What do we know about probability? Just numbers. Always between 0 & 1, inclusive. And the total probability of all possible values must sum to 1. Probability distributions are formulas that generate numbers that have this characteristic. Any formula that has these properties can be called a probability distribution. The area under a probability density function [which shows the “density of probability” at each value of an imaginary and theoretical variable] sums to one.

**Measurement**

- **Measurement & Conceptualization**: among most important parts of research design.
  - **Concept**: a word or sentence that helps us make sense of/place into pleasing order, other (sets of) words or sentences
  - **Measurement**: assigning numbers; act of recognition of manifestations of concepts in the world.

- Adcock and Collier Figure 1; think of examples.
Random variables and Probability Distributions

- **Random Variables:**
  - people’s height, level of political interest, whether coin is heads, # deaths from horse kicks during Franco-Prussian War, # radioactive particles emitted by piece of uranium per minute.
  - These are all theoretical constructs. They refer to stuff in the real world. They can be represented by mathematical formulas.
  - You give me some other examples...
• “A random variable is a function that assigns a number to each outcome of the sample space of an experiment.”
  – Experiment: repeatable (in thry) method generating outcomes
  – Outcome: a possible observation from the experiment. You DO experiments to OBSERVE outcomes.
  – Sample space: the possibilities for values of this variable.
    * Sometimes continuous (i.e. infinite values — examples?).
    * Sometimes discrete (i.e. countable values — examples?)
  – Function: some math that changes some inputs to outputs.

• Some Examples & Practice
  – An obs (or measure) of person’s ht as “realization of an RV”
  – Coin flip. War, not War. Employed, not. Alive, Dead. # Parties.
  – Can (& usu. will) know some things re: RV before obs:
    * Coin Flips? (only 2 values; heads, tails)
    * Height? (Non-neg.; most val’s possible; overwhelmingly b/w 36” and 84” for adults.)
    * For your examples?
  – Height:
    * I’m 68”. I sample one of you... Another... Say 5 of you. Let’s graph dist of this RV in this sample.
    * Can calculate (sample) mean & std. dev...
    * But only 6 pieces info not a lot about overall dist of height.
    * Imagine slowly filling histogram until begins to look continuous...
    * What will it look like? [Central Tendency?] [Symmetric or Skewed?] [Unimodal or Bimodal?] [Shape?]
    * You work for airline; how answer: “What is the probability that a man is over 7 feet tall?”
    * Say got whole pop globe & 1% > 84”. Want to qualify this statement somehow for your purposes?
– If RV followed or exactly followed a known dist, e.g., normal, then wouldn’t need whole pop to make probability statements. Could use properties of that known dist.

– *Normal* needs only mean & s.d. *Bernoulli* (coin toss) needs even less, just p=prob(1), i.e., mean. Number heads in N tosses is a *Binomial*: just p & N. Counts of events occurring at constant probability rate (e.g., horse-kicks to heads in Prussian War) are *Poissons*; need rate (mean again) and time. Etc. Etc.

• More on probability & prob distributions after this summary.

**Summarizing Measurement**

• Distinction between *Concept* and *Measure*

• *Construct Validity* — plausible mapping b/w concept & measure.

• Multiple Measures & Meas. Thry (“Sampling the Universe of Content”)

• *Empirical Validity Tests* (Convergent & Discriminant Validity)

• *Levels of Measurement*: nominal, categorical, ordinal, interval, ratio. Defined by kinds of algebraic operations each allows.

  – *Nominal* allowable operations: =, ≠
  – *Categorical* allowable operations: =, ≠
  – *Ordinal* allowable operations: =, ≠, <, >
  – *Interval* allowable operations: =, ≠, <, >, +, –
  – *Ratio* allowable operations: =, ≠, <, >, +, –, ×, ÷

• Two hierarchies:

  – 1\(^{st}\) Hierarchy: rom *concept* to *operationalization* to *measurement*: (“heat” to “molecular motion” to “mercury rising in a tube”).

    * Concepts are ideas embedded in webs of other ideas.

    * Operationalization translates these abstract ideas into observable variables; how we scientists see the imprints of concepts.
Can’t “see” molecular motion, so we use “instruments” to gauge its imprint (imprint of its operationalization)

Want to measure “heat”, but, in fact, can only gauge effect of molecular motion (& imprecisely at that),

Even debate about how to do that (What is motion? What is a molecule?) [In PS, key concern of folks like David Collier.]

* Measures gauge how op’d concepts imprint upon world.

* Basically, A hierarchy of questions here:
  
  (1) What is the concept? How useful/used? How relate to other concepts? (Rob’s Rule: A concept that takes more words to define & explain than saves longrun is not useful.)
  
  (2) How would we recognize it when it happens? How does concept manifest in world?
  
  (3) What instruments can we use to recognize/gauge? How can we see manifestations of concept?

2nd Hierarchy: type of information your instruments give & what can do w/ this info: what kind of basic statements are meaningful.

* Always, “lowest level” to “highest level”, have “equals” statements.

* Lowest Level? Nominal: can only name values & so can only say equal/not, same/not. Examples?

* Next Level? Ordinal: can order, but distances b/w not meaningful. Examples?

* Next Level? Interval: distance b/w values meaningful, but not nec. multiplicative/division. Examples?

* Highest Level? Ratio: distance & products & ratios & all mathematical relations have full meaning.

* Most folks will just collapse highest two & for most purposes OK to do so. Strictly speaking, ratio must be invariant to additive/multiplicative or other rescaling, so, e.g., Calendar dates & Fahrenheit/Celsius not strictly ratio (e.g., create new scale by
subtracting 32 & ratios in that new will not be equal those in
original), BUT w/in same scale they are (64 really is twice 32 in
that scale), so must folks satisfied w/ that. Need to remember
‘twice’ here only meaningful w/in the same scale, though. By
strict def, most PS scales not ratio, by looser one many are.
– Seems much ado about near nada, philosophical taxonomies of mea-
surement dancing about on pinheads.... But important b/c:
  * Don’t confuse operationalization w/ concepts — (operational-
   ism: “intelligence”≡IQ)
  * Note: not all vectors of numbers can bear all mathematical op-
erations & so convey as much meaning. So, beware! & be aware!

Programming with Data

• Commands to language interpreter in plain text file. In ideal world...
  – Each task has own file — but what’s 1 task? Good Q; tradeoffs.
  – Each command either perfectly intelligible in plain language or
    commented by programmer to become so.

• Command file sent to interpreter, which processes into output. Usually,
  commands not compiled but interpreted.

• What does a GUI have to do with this process?

• Example: File called “pctdems4800.r”. At shell prompt, could
  – R CMD BATCH pctdems4800.r pctdems4800.rout to run file.
  – Could use file to collect commands fed R one at time.
  – Same procedures available for Stata.

##This file reads data from the 1948-2000 National Election Studies
##Cumulative file. It also tells me what percent voted democrat in
##each of the elections between 1948 and 2000.

##First, read the data file: Since the datafile is in SPSS portable
##format, this requires using the "foreign" library.
library(foreign)
nes4800<-read.spss("nes48_00.por",use.value.labels=FALSE,to.data.frame=TRUE)

##The codebook tells me that variable VCF0704 is party of vote and
##VCF0004 is year.

##Here are the relevant pieces from the codebook, just to help me
##make sense of things:

##VAR CF0704   PARTY OF PRES VOTE- ALL MAJOR CANDIDATES ##
##1. Democrat ##2. Republican ##3. Major third party candidate
##(Wallace 1968/Anderson 1980/Perot 1992,1996) ##0. Did not vote....

##VAR CF0004 YEAR OF STUDY ##1948-2000 coded.

attach(nes4800)

##Check to see if the variables make sense

table(VCF0004) table(VCF0704)

##I'm including some output here to remind me of what is going on.
##VCF0704

## 0 1 2 3
##28456 7598 8081 580
##
##I want percentages so what do I do?

totalsbyyear<-colSums(votebyyear)

##VCF0004 ##VCF0704

0  272  664 1139  496 1450  283 1297  460 1291  530 1507 1118  1575  926
1  212  518  511  0 446  0 750  0 421  0 566  0 670  
2  178  717  0 755  0 452  0 361  0 490  0 1021  0 652  
3  0 0 0 0 0 0 0 0 0 0 116 0 0 0 

## VCF0004 ##VCF0704

0 2304  656 1418  881 2176  845 1980  827 1795  598 1281  687 
1  0 383  0 575  0 563  0 793  0 600  0 590  
2  0 494  0 801  0 632  0 564  0 434  0 530  
3  0 81  0 0 0 0 0 0 301  0 82 0 0 

##I want percentages so what do I do?
totalsbyyear<-colSums(votebyyear)
1976 1978 1980
2248 2304 1614

1418 2257 2176 2040 1980 2485 1795 1714 1281 1807

##Divide second row of table called votebyyear by column totals

```r
pctdemvotebyyear<-votebyyear[2,]/totalsbyyear > pctdemvotebyyear
```

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3202417</td>
<td>0.2727751</td>
<td>0.0000000</td>
<td>0.2900114</td>
<td>0.0000000</td>
<td>0.3776461</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000000</td>
<td>0.4774029</td>
<td>0.0000000</td>
<td>0.2703918</td>
<td>0.0000000</td>
<td>0.2092421</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000000</td>
<td>0.2980427</td>
<td>0.0000000</td>
<td>0.2372986</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000000</td>
<td>0.2547630</td>
<td>0.0000000</td>
<td>0.2759804</td>
<td>0.0000000</td>
<td>0.3191147</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000000</td>
<td>0.3500583</td>
<td>0.0000000</td>
<td>0.3265080</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

##Is this right? Check for 1948 using numbers in votebyyear.

\[
\frac{212}{272+212+178} = 0.3202417
\]

###Probability

- **Overview**
  - Example: designing seats for airline. Need to know something re: shapes of people.
  - On what attributes should we focus?
  - How gather info on that? What do we want to know?
  - What is the population? Not as simple as seems.
    * Likely airplane riders?
    * What of laws that can’t turn people away? so perhaps whole pop? Of which country? All?
  - Say have sample of values on variable of interest. What might density (histogram) look like? What info do you want from histogram?
– Say you’re president of airline, and not a corp, so you’re personally liable if sued. Now what info do you want?

* Difficulty here is you want to know future. Don’t care about this histogram per se, but what it suggests re: chances be sued.
* If we got absolutely everybody on globe, & asked, “what is probability man>7 feet tall”.
* Could say, “only 2% of male pop globe > 7’, so, guess safe to say probability that any given male passenger > 7’ is 2%”
* Do you want to qualify this statement somehow?
  · If everyone has equal chance of flying and
  · If my current obs have captured TRUE pop w/distrib on this variable never changing.
* I.e., we care not about height as empirical fact but height as Random Variable. I.e, not as variable w/ certain values in dataset, but as theoretical entity (placeholder) assigning values to outcomes. R.V. height has many possible values; density of those possible values is our concern.
* A particular obs or measurement of person’s height is “realization of a random variable”. A particular flip of a coin, etc.
* So, as may have noticed, can know some things re: R.V. before gather any obs. Example: about coin flips? About height?
  · Esp. nice if R.V. approx’d by well-known probability distribution, e.g. normal dist, b/c then don’t need whole pop to make probability statements.
  · More about how easy this is later, but, if, say, height is an R.V. drawn from a normal probability distribution, then we know tons about height.
  · Much easier to know about a mathematical formula than about real world. [Maybe really tall people hide from interviewers? Or are mistaken for trees or something?] But formula tells you prob seeing someone > 7’ just by plugging in some numbers.
· E.g., coin flips are known to be realizations of a Bernoulli dist. And number of heads in, say, 5 flips is a realization from a Binomial distribution. The horsekicks in Franco-Prussian war (another quotidian example), are known realizations from Poisson distribution. And many more.
· Lots of things (not all), e.g. height, normally distributed.

• **Frequentist Definition of Probability**
  
  – *Probability*: Limiting relative frequency: \( \text{Pr} \equiv \lim_{n} \frac{f}{n} \)
  
  *Relative Frequency*: \# times certain value of R.V. realizes.
  *Limiting*: if you looked after infinite realizations.
  *Example*: Rolled 6-sided die 10 times. Get some 1s, 2s, etc... Unequal relative frequencies at first. But, if fair die, & rolled it over & over forever, each value equally likely, so, over time, *limiting relative frequency* settles to \( \frac{1}{6} \) each. So, “probability of rolling 1 = \( \frac{1}{6} \) = probability of rolling 2,3,4,5, or 6”
  *Example*: Coin-flips. Proportion of tosses of fair coin yielding heads changes w/ more tosses. Eventually, proportion approaches .5. Plot shows two trials of 5000 tosses each.

* FYI, people have actually physically done this:
  · Count Buffon (1707-1788) 4040 times, finding 50.69% heads;
· Karl Pearson (??) 24,000 times, finding .5005;
· As German prisoner WWII, John Kerrich (S.African statistician) 10K times: .5067.

− So how & why does # of obs we matter? [more obs, more certainty, usually; settles toward true or limiting probability].
− The point(s)?
  * (1) To infer from data observed to unobserved or even unobservable: TRUTH or future or democracy in China — to try to understand how patterns we observe produced so can say what might happen in new circumstances.
  * (2) To talk as coherently as possible, as scientists, about our uncertainty re: results & theories in widely communicable & commonly understood ways. A shared language of uncertainty.
  * Sometimes some conflicts b/w these two goals.

− So Probability any outcome of random phenomenon=proportion times outcome would occur in very long series repetitions phenomenon: long-term relative frequency.
  * By this def, probability inherent in real-world phenomenon.
  * Confident p(heads)=.5 b/c, in reality, true rel.freq.=.5.

− Talking about probabilities more rigorously requires math. Until 1930s, not very good set math tools to talk probability in general way, but then Kolmogorov realized set theory allows talk about probability in logically consistent way.

− So, some big list terms & ideas to follow:

  • Describe coin tossing math’ly (probability model coin tossing) req’s:
    − (1) List of possible outcomes of coin toss
    − (2) Probability (relative frequency) each outcome.
    − Want describe math’ly b/c coin-toss-like things (yes/no survey Q, democracy/non) of interest can’t do physical act to gen the random phenom enough for rel. freq. reveal self).
Plus, even if could (e.g., guys above tossing coins), not exactly give true probability (an ideal coin). Just estimate in sample that size.

- **Outcome Set** or **Sample Space** ($S$): exhaustive catalog of all possible outcomes of experiment.
  - Sample Space of single coin toss? $S = \{H, T\}$
  - Sample Space of three coin tosses?

  $$S = \{HHH, HHT, HTT, HTH, THH, THT, TTH, TTT\}$$

  > expand.grid(flip1=c(0,1),flip2=c(0,1),flip3=c(0,1),flip4=c(0,1))

<table>
<thead>
<tr>
<th>flip1</th>
<th>flip2</th>
<th>flip3</th>
<th>flip4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

  - Each toss $H(=1)$ or $T(=0) \Rightarrow 2^{16} = 65,536$ outcomes.
  - If outcome of interest “# heads” $\Rightarrow 17$ outcomes.
  - If interest # of heads in 4 tosses: $S = \{0, 1, 2, 3, 4\}$

- **Event** $\equiv$ outcome or set outcomes of random phenom.
  - **Simple Events**: e.g., any single, specific outcome; e.g., 4 coin tosses, $\{0000\}$ or any other 1.
  - **Compound Events**: any combo outcomes, e.g., exactly 2 heads...?
• Characteristics of Probabilities:
  – Probabilities lie b/w 0 & 1 (Chances b/w 0% & 100%).
  – Probabilities of any outcome set must sum to 1.
    * **Probability Function**: Any function satisfying $0 \leq P(A_i) \leq 1$ & $\sum_i P(A_i) = 1 \forall$ events $A_i$.
  – Corollary: $P(G) = 1 - P(\text{not } G)$ (“chance of something = 100% - chance of opposite thing,” FPP p. 209)

• Some Probability Thoughts & Examples
  – Toss fair coin 10,50,100 times & record how often heads. What is best guess of probability of heads after each time?
  – 2 boxes, get $1 if draw red marble & 0 if draw blue. $A$ contains 3 red & 2 blue; $B$ 30 red & 20 blue. Which box?
  – What matters is ratio (relative frequency). If drew marbles again & again with replacement: $\frac{3}{2+3} = \frac{30}{30+20} = .6$. What about w/o replacement? (difficult problem; discuss)
  – Roll 2 fair dice:
    * Sample Space? $S = \{11, 12, \ldots, 56, 66\} \Rightarrow \text{36 outcomes}$. Histogram appearance?
    * Sum: Min & Max? Sample Space? $S = \{2, \ldots, 12\}$. Mean? Histogram Appearance?
    * Probability (rel. freq.) 2 as sum?
      > expand.grid(roll1=c(1:6),roll2=c(1:6))
      \[
      \begin{array}{ccc}
      \hline
      \text{roll1} & \text{roll2} \\
      \hline
      1 & 1 & 1 \\
      2 & 2 & 1 \\
      3 & 3 & 1 \\
      4 & 4 & 1 \\
      5 & 5 & 1 \\
      6 & 6 & 1 \\
      7 & 1 & 2 \\
      8 & 2 & 2 \\
      9 & 3 & 2 \\
      10 & 4 & 2 \\
      11 & 5 & 2 \\
      \hline
      \end{array}
      \]
> rowSums(expand.grid(roll1=c(1:6), roll2=c(1:6)))

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

> table(rowSums(expand.grid(roll1=c(1:6), roll2=c(1:6))))

2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 5 4 3 2 1

* Just 1 way to get sum=2; just 1 sum=12. But 6 ways sum=7.

− 3-child family:

\[ S = \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\} \]

Sum probs=1. Not nec. =, but \( \approx \) here. (WW Fig.3-4, p.75)
* Can calculate probability of any collection of outcomes.
* An event, e.g. \(BBB\), labeled \(e_1\).

* What is probability at least one boy?
  
  - At least one B in 7 of 8, so \(\frac{7}{8}\) or \(.875\) or 87.5%.
  
  - Or, what set of outcomes comprise this event? \((e_1 \ldots e_7)\). So, \(E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\} = 7/8\)

  - Probability of event \(E\) is sum of probabilities of outcomes that comprise it. \(Pr(E) = \sum Pr(e)\) (for disjoint events)

> expand.grid(k1=c("B","G"),k2=c("B","G"),k3=c("B","G"))

<table>
<thead>
<tr>
<th>event</th>
<th>k1</th>
<th>k2</th>
<th>k3</th>
<th>Prob(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>1/8</td>
</tr>
<tr>
<td>2</td>
<td>G</td>
<td>B</td>
<td>B</td>
<td>1/8</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>G</td>
<td>B</td>
<td>1/8</td>
</tr>
<tr>
<td>4</td>
<td>G</td>
<td>G</td>
<td>B</td>
<td>1/8</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>B</td>
<td>G</td>
<td>1/8</td>
</tr>
<tr>
<td>6</td>
<td>G</td>
<td>B</td>
<td>G</td>
<td>1/8</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>G</td>
<td>G</td>
<td>1/8</td>
</tr>
<tr>
<td>8</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>1/8</td>
</tr>
</tbody>
</table>

- WW Problem 3-5.

  * a. In learning exp., subject attempts some task twice. Each time, chance failure=.40. Calc. chance 1 failure using probability tree?
Outcome  Probs (sum=1)
SS  \( .6 \times .6 = .36 \)
SF  \( .6 \times .4 = .24 \)
FS  \( .4 \times .6 = .24 \)
FF  \( .4 \times .4 = .16 \)

\* \( \Pr(E) = \Pr(SF) + \Pr(FS) = .24 + .24 = .48 \)

\* What if subject has to do this 3 times? [Add branch to tree]
Outcome  Probs (sum=1)
e1  SSS  \( .6 \times .6 \times .6 = .216 \)
e2  SSF  \( .6 \times .6 \times .4 = .144 \)
e3  SFS  \( .6 \times .4 \times .6 = .144 \)
e4  SFF  \( .6 \times .4 \times .4 = .096 \)
e5  FSS  \( .4 \times .6 \times .6 = .144 \)
e6  FSF  \( .4 \times .6 \times .4 = .096 \)
e7  FFS  \( .4 \times .4 \times .6 = .096 \)
e8  FFF  \( .4 \times .4 \times .4 = .064 \)

\* \( \Pr(E) = \Pr(e2) + \Pr(e3) + \Pr(e5) = .144 + .144 + .144 = .432 \)

• Compound Events
  – Set Theory: Say 2 sets, G & H w/ some elements/outcomes.
    * Union: What is “union of G & H”?
      · Elements in G, H, or both; i.e., “G OR H”. Written: \( G \cup H \).
    * Intersection: What is “intersection of G & H”?
      · Elements in both G & H; i.e., “G AND H”] Written: \( G \cap H \).
    * So, \( \cup = OR \) and \( \cap = AND \)
  – Examples:
    * G is 1 failure, & H is 2 failures. What are events in each set?
      * \( G = \{ e2, e3, e5 \} \) and \( H = \{ e4, e6, e7 \} \)
      * What is \( G \cup H \)? ... \( G \cup H = \{ e2, e3, e4, e5, e6, e7 \} \)
      * What is \( G \cap H \) ... \( G \cap H = \{ \} \)
New set, \( J \), at least 1 fail: \( J = \{e_2, e_3, e_4, e_5, e_6, e_7, e_8\} \)

- What is \( G \cup J \)? ... \( \{e_2 \ldots e_8\} \)
- \( G \cap J \)? ... \( \{e_2, e_3, e_5\} \)

So what is \( \Pr(G \cup H) \) i.e. \( \Pr(G \text{ OR } H) \)?

**Addition Rule:**
\[
\Pr(G \text{ OR } H) = \Pr(G) + \Pr(H) - \Pr(G \text{ AND } H)
\]

Subtract \( \Pr(G \text{ AND } H) \) so not double count. If \( G \) & \( H \) mut. excl. — i.e., no shared events (disjoint) — then not need subtract.

**Conditional Probability**

- Same idea prob \( \equiv \) limiting rel. freq. but set relevant outs reduced, redefining universe poss outs. (Draw Venn Diagram.)

\[
\Pr(H \mid G) = \frac{\Pr(H \text{ and } G)}{\Pr(G)} = \frac{\Pr(H \cap G)}{\Pr(G)}
\]

- Roll 2 dice seq’ly. \( \Pr(\text{sum}=3)=\frac{2}{36} \) If 1\(^{st}\) roll 1, then \( \Pr(\text{sum}=3)=? \) \( S \) had 36 outs b4 1\(^{st}\) roll; now just 6, & only 1 \( \Rightarrow \text{sum}=3: \) sample space reduced. \( \Pr(\text{sum}=3 \mid 1^{st} \text{ roll}=1')=\frac{1}{6}. \)

- Prob 3-15 WW, p. 88 (Stat Abs US ‘87)

<table>
<thead>
<tr>
<th></th>
<th>M(ale)</th>
<th>F(emale)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E(\text{mployed}) )</td>
<td>51.9%</td>
<td>40.9%</td>
<td>92.8%</td>
</tr>
<tr>
<td>( U(\text{nemployed}) )</td>
<td>3.9%</td>
<td>3.3%</td>
<td>7.2%</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>55.8%</strong></td>
<td><strong>44.2%</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>

- a. Unemp. rate? \( \Pr(U) \) p(wrkr drawn at rndm unemp’d? (.072)
- b. \( \Pr(U \mid M) \)? (How to read this?) (Male U-rate) \( .039/.558=.06989 \)
- c. \( \Pr(U \mid F) \)? (Female U-rate) \( .033/.442=.07467 \)

- Prob 3-16 WW, p.88. ‘74 Gallup poll: abortion attitudes. Q text: “US Sup Ct has ruled that a woman may go to a doctor to end pregnancy at any time during the first 3 months of pregnancy. Do you favor or oppose this ruling?”

<table>
<thead>
<tr>
<th></th>
<th>Favor</th>
<th>Oppose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>e1=.27</td>
<td>e2=.21</td>
</tr>
<tr>
<td>Female</td>
<td>e3=.24</td>
<td>e4=.28</td>
</tr>
</tbody>
</table>
- What’s probability that individual drawn at random would be . . .
  * a. in favor of abortion? (label e’s here). So this is asking about
    \[ \Pr(e_1, e_3) = .27 + .24 = .51 \]
  * b. in favor of abortion if male? So this is asking about
    \[ \Pr(A | M) = \frac{\Pr(A \cup M)}{\Pr(M)} = \frac{.27}{.27 + .21} = \frac{.27}{.48} = .5625 \]
  * b. in favor of abortion if female? So this is asking about
    \[ \Pr(A | F) = \frac{\Pr(A \cup F)}{\Pr(F)} = \frac{.24}{.24 + .28} = \frac{.24}{.52} \approx .462 \]

- Statistical Independence

  - No rule says when restrict analysis to subgrps, diff answers for each. If same answer for each subgrp, then *statistical independence*.
  - Formally if F & E statistically independent, \( \Pr(F|E) = \Pr(F) \) or \( \Pr(F) \perp \Pr(E) \). Knowing something re: E tells nothing re: F.
  - **If F & E independent events, then:** \( \Pr(E \cap F) = \Pr(E) \Pr(F) \)
  - Famous Independence Example: betting on coin tosses, 1$ if heads.
    * If first toss heads, what chance win $ on 2\textsuperscript{nd}? If first toss tails, what chance win $ on 2\textsuperscript{nd}? \Rightarrow Tosses independent.
  - Prob 3-23, WW p.92: ’85 Labor Force Participation (from \textit{CPS}).
    \[
    \begin{array}{ccc}
    & <25 & \geq 25 & \text{Total} \\
    \text{E(mployed)} & 20.4 & 86.8 & 107.2 \\
    \text{U(nemployed)} & 3.2 & 5.1 & 8.3 \\
    \text{Totals (M)} & 23.2 & 91.9 & 115.5 \\
    \end{array}
    \]
    - a. What is \( \Pr(U) \)? \( 8.3/115.5 \approx .072 \)
    - b. What is \( \Pr(U | Y) \)? \( \Pr(U \text{ AND } Y) / \Pr(Y) = 3.2/23.6 \approx .136 \)
    - b. What is \( \Pr(U | O) \)? \( \Pr(U \text{ AND } O) / \Pr(O) = 5.1/91.9 \approx .055 \)
    - c. Is unemployment independent of age?
    - What if only observed 115 people instead of 115M? Concerns about uncertainty of diff? — i.e. is \(.136-.055=0.081\) big or not?]
3. THE MULTIPLICATION RULE

This section will show how to figure the chance that two events happen, by multiplying probabilities.

Example 3. A box has three tickets, colored red, white and blue.

\[ \frac{R}{W} | \frac{B}{H} \]

Two tickets will be drawn at random without replacement. What is the chance of drawing the red ticket and then the white?

Solution. Imagine a large group of people. Each of these people holds a box \( \frac{R}{W} | \frac{B}{H} \) and draws two tickets at random without replacement. About one third of the people get \( \frac{R}{W} \) on the first draw, and are left with

\[ \frac{W}{R} | \frac{H}{B} \]

On the second draw, about half of these people will get \( \frac{W}{R} \). The fraction who draw \( \frac{R}{W} \) is therefore

\[
\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}
\]

The chance is 1 in 6, or \( \frac{1}{6} \).

For instance, suppose you start with 600 people. About 200 of them will get \( \frac{R}{W} \) on the first draw. Of these 200 people, about 100 will get \( \frac{W}{R} \) on the second draw. So \( \frac{100}{600} = \frac{1}{6} \) of the people draw the red ticket first and then the white one. (In figure 4, the people who draw \( \frac{R}{W} | \frac{B}{H} \) are at the top left.)

Statisticians usually multiply the chances in reverse order:

\[
\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}
\]

The reason: \( \frac{1}{3} \) refers to the first draw, and \( \frac{1}{2} \) to the second.

Figure 4. The multiplication rule. (Each stick figure corresponds to 100 people.)

---

The method in example 3 is called the multiplication rule.

**Multiplication Rule.** The chance that two things will both happen equals the chance that the first will happen, multiplied by the chance that the second will happen given that the first has happened.

Example 4. Two cards will be dealt off the top of a well-shuffled deck. What is the chance that the first card will be the seven of clubs and the second card will be the queen of hearts?

Solution. This is like example 3, with a much bigger box. The chance that the first card will be the seven of clubs is \( \frac{1}{52} \). Given that the first card was the seven of clubs, the chance that the second card will be the queen of hearts is \( \frac{1}{51} \). The chance of getting both cards is

\[
\frac{1}{52} \times \frac{1}{51} = \frac{1}{2,652}
\]

This is a small chance: about 4 in 10,000, or 0.04 of \( \frac{1}{100} \).

Example 5. A deck of cards is shuffled, and two cards are dealt. What is the chance that both are aces?

Solution. The chance that the first card is an ace equals \( \frac{4}{52} \). Given that the first card is an ace, there are 3 aces among the 51 remaining cards, so the chance of a second ace equals \( \frac{3}{51} \). The chance that both cards are aces equals

\[
\frac{4}{52} \times \frac{3}{51} = \frac{12}{2,652} = \frac{1}{221}
\]

This is about 1 in 200, or \( \frac{1}{2} \) of \( \frac{1}{100} \).

Example 6. A coin is tossed twice. What is the chance of a head followed by a tail?

Solution. The chance of a head on the first toss equals \( \frac{1}{2} \). No matter how the first toss turns out, the chance of tails on the second toss equals \( \frac{1}{2} \). So the chance of heads followed by tails equals

\[
\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}
\]

Exercise Set C

1. A deck is shuffled and two cards are dealt. (a) Find the chance that the second card is a heart given that the first card is a...
Example (FPP, 234): In *Bomber*, Len Deighton argues WWII pilots 2% chance being shot down each mission. So, in 50 missions “mathematically certain” shot down: $50 \times 2\% = 100\%$. Correct?

How set up?

* 2% chance shot down. Like drawing from box 100 cards, 2 of which “shot down” & 98 say “survive”.
* Survive 2 missions? (if indep!) $98/100 \times 98/100 = .98^2 = .96$
* 50 missions? (if indep!) $0.98^{50} = .36$.

**Bayes Theorem**

- Probability as lim.rel.freq. mostly regards going from data to pop’s — under some assumtpts re: repeatability exper’s (e.g, dice rolls)
- But other situations where want to think about probability — what is probability *this* car is lemon? Or *this* defendant is telling truth? Or *this* particular lab outcome is right?
- In each case, too little data just to count relative freq’s. (In some cases, too much prior knowledge, which is oddly similar, actually.)
- *Bayesian & Classical (Frequentist) Statistics:*
  * From frequentist view, $p$ heads for some coin is given, constant, & exists in TRUTH & will be found if just flip it enough.
  * From Bayesian view, $p$ a random variable itself (!), about which have prior info that helps refine estimates of it (new info causes us to adjust those priors, but no reference to any TRUE $p$).
  * Assessing probability Bayesian-style involves specifying this prior knowledge as a probability, then reassessing this prior given what observed in world, to get posterior probability of events.
  * If prior and posterior are different, then the new information changed beliefs about probability. We *updated.*
Bayes’ Theorem & Bayes Rule, though, equally central to both Frequentist & Bayesian Statistics. Just diff deep philosophical interp.

Example from WW about the Q-car (page 93).

* Know in advance (prior) that 30% of cars at Honest Ed’s faulty.
* To refine assess this car, so gamble less, call mechanic (evidence).
* Know that, when this mechanic drives faulty car, says “faulty” 90% of time & 10% says “ok” mistakenly.
* Also know that, when this mechanic drives good car, says “ok” 80% of time, & 20% says “faulty” mistakenly.
* What’s chance this car is faulty?
  · a. Before he hires the mechanic?
  · b. After the mechanic pronounces it “faulty”?

\[
\begin{align*}
\Pr(F|\text{"f"}) &= \frac{\Pr(F \cap \text{"f"})}{\Pr(\text{"f"})} = \frac{.3 \times .9}{.3 \times .9 + .7 \times .2} = \frac{.27}{.27 + .14} = .27/.41 = .66 \\
\Pr(F|\text{"OK"}) &= \frac{\Pr(F \cap \text{"OK"})}{\Pr(\text{"OK"})} = \frac{.3 \times .2}{.3 \times .1 + .7 \times .8} = \frac{.03}{.03 + .56} = .03/.59 = .05
\end{align*}
\]

Fall 2005 – Rob Franzese
Prob 3-26, WW p. 97: A barometer manufacturer tests one & finds it erroneously predicts no rain 10% of time & erroneously predicts rain 30% of time. It actually rains 40% of time in some town in Sept. Barom says rain on Lab Day, what probability actually rain?

\[ \Pr(R|\text{“rain”}) = \frac{\Pr(R \text{ AND } \text{“rain”})}{\Pr(\text{“rain”})} = \frac{.4 \times .9}{.4 \times .9 + .6 \times .3} = \frac{.36}{.36 + .18} = .36/.54 = .67 \]

• Why Probability?

– Isaac, p. 169: “Statistics can be roughly characterized as the mathematical theory of making decisions in the face of uncertainty.”
– Need way to talk about what likely and what not.
– Probability doesn’t say much about particular coin toss(es), but about behavior of coin(s). Percentage of time expect heads if flipped coin(s) repeatedly in same circumstances.
• Probability Distributions

  - P(certain values) in R.V. governed by prob dist that R.V. P-Dist of R.V. X lists poss values X & p each:

<table>
<thead>
<tr>
<th>Value of X</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>...</td>
<td>$p_k$</td>
</tr>
</tbody>
</table>

  - Folks talk of p-dists as “generating” obs, meaning “that set values looks awfully like set that such-&-such p-dist would gen.” Or “that RV follows 1st principles of such-&-such p-dist”.

    * What’s probability again? Limiting relative frequency.
    * What’s R.V.? Device maps $S$ onto $\mathbb{R}$ (assigns #’s to exp.outs.)
    * What’s a discrete R.V.? Continuous R.V.?

  - E.g.: Fair die roll. Prob obs @ value=1/6. Can write:
    
    $x$ | $p(x)$ | $x$ | $p(x)$ | $x$ | $p(x)$ | $x$ | $p(x)$ | $x$ | $p(x)$ | $x$ | $p(x)$ |
    1   | 1/6   | 2   | 1/6   | 3   | 1/6   | 4   | 1/6   | 5   | 1/6   | 6   | 1/6   |

    * So, can write: $\Pr(X < 2) = p(1) = 1/6$ or $\Pr(1 < X < 4) = p(2) + p(3) = 2/6 = 1/3$, &

    * w/ this notation, can rewrite (pop) mean & variance formulas:

    Mean: $\mu = \sum xp(x)$

    Variance: $\sigma^2 = \sum (x - \mu)^2 p(x)$

• Discrete Probability Distributions:

  - Bernoulli Prob. Dist. Examples Bernoulli R.V.?

    $p(x) = \pi^x (1 - \pi)^{1-x}$

  - Binomial Prob. Dist. Examples Binomial R.V.?

    * # successes, $s$, in # trials, $n$ (finite).
    * Each trial independent.

    $p(s) = \binom{n}{s} \pi^s (1 - \pi)^{n-s}$
\(* \binom{n}{s} \), read “n choose s”, called “binomial coefficient” \( \Rightarrow \# \) of possible combinations of \( s \) things chosen from \( n \).

\[
\binom{n}{s} = \frac{n!}{s!(n-s)!}, \text{ where } n! = n(n-1)(n-2) \ldots 1
\]

\(* \) Binomial pf’s w/ \( \pi = \frac{1}{4} \) & \( n \in 5, 10, 20 \)

\[\begin{array}{c|cccc}
\hline
x & 0 & 5 & 10 & 15 & 20 \\
\hline
f & 0.05 & 0.25 & 0.35 & 0.1 & 0.1
\end{array}\]

\[\begin{array}{c|cccc}
\hline
x & 0 & 5 & 10 & 15 & 20 \\
\hline
f & 0.05 & 0.25 & 0.35 & 0.1 & 0.1
\end{array}\]

- Examples: \# heads in \( n \) flips. \# days/wk read paper. Others?
- So, e.g., probability 1 girl out of 3 children (from before):

\[
p(s = 1|n = 3) = \binom{3}{1}(.48)^1(.52)^2 = \frac{3!}{1!2!}(.13)
\]

\[
= \frac{3 \cdot 2 \cdot 1}{1 \cdot 2 \cdot 1}(.13) = 3(.13) = .39
\]

- Binomial Mean & Variance:

\[
\mu = \pi_1 + \pi_2 + \ldots + \pi_n = n\pi
\]

\[
\sigma^2 = n\pi(1 - \pi)
\]

- Cool! No huge list & counting! Just formula.
- Example: \( P(400H) \) in 1K flips?

\[
p(400) = \binom{1000}{400}(.5)^{400}(.5)^{1000-400} = \frac{1000!}{400!600!}(.5)^{400}.5^{600} = 4.63 \times 10^{-11}
\]

- I.e., extremely rare [decimal, 11 zeros, & 463]. So, formulas really cool & useful!
Note: Bernoulli=Binomial w/ \( n = 1, p(1) = \pi, p(0) = 1 - \pi \):

\[
p(1) = \binom{1}{1}(.48)^1(1 - .48)^{(1-1)}
\]

\[
= \frac{1!}{1!(1 - 1)!}(.48)^1(.52)^0
\]

\[
= 1 \times .48 \times 1 = .48
\]

Let’s calculate mean & s.d. of Bernoulli using formulas:

- **Mean** \( \mu = \sum xp(x) = 1 \times \pi + 0 \times (1 - \pi) = \pi \)

- **Variance** \( \sigma^2 = \sum (x - \mu)^2 p(x) \)

\[
= (0 - \mu)^2 \times (1 - p) + (1 - \mu)^2 \times (p) = (0 - p)^2 \times (1 - p) + (1 - \mu)^2 \times p
\]

\[
= p^2(1 - p) + (1 - 2p + p^2)p = p^2 - p^3 + p - 2p^2 + p^3
\]

\[
= p - p^2 = p(1 - p)
\]

So Bernoulli var, \( \sigma^2 = \pi(1 - \pi) \), has NO info beyond mean.

- Also notice: if \( p=1 \) or \( 0 \), variance is 0. Makes sense, right?

---

**Poisson Distribution**: # (indep) events in fixed period, at fixed rate (\( \lambda \)). (Binomial → Poisson as \( n \rightarrow \infty \) & \( \pi \rightarrow 0 \).)

\[
p(x) = \frac{e^{-\lambda}\lambda^x}{x!}
\]

\[
\mu = \lambda
\]

\[
\sigma^2 = \lambda
\]

* Poisson pf’s w \( \lambda \in 2, 5, 10 \).
\[
\mu = \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} = 0 \cdot \frac{e^{-\lambda} \lambda^0}{0!} + \sum_{x=1}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!}
\]

\[
= 0 + \sum_{x=1}^{\infty} x e^{-\lambda} \frac{\lambda^{x-1}}{(x-1)!} \cdot \frac{\lambda}{x} = \sum_{x=1}^{\infty} \frac{x \lambda e^{-\lambda} \lambda^{x-1}}{x (x-1)!}
\]

\[
= \lambda \cdot \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^{x-1}}{(x-1)!} = \lambda \cdot 1 = \lambda
\]

· Last line b/c probs must sum 1 & \(\sum\) here=Poisson p.f. again.
· (Derivation of variance (omitted) is similar.)

* \(\pi\) in Bernoulli & Binomial was p(success) in 1 trial; \(\lambda\) similarly =success rate (per period).

* Example (DeGroot, p.257) Store owner believes mean # customers arriving per hr is 4.5. What \(Pr(\geq 12\text{\ customers2hrs})\)?
· In 2 hrs, expect \(2 \cdot \lambda = 2 \cdot 4.5 = 9\) customers.

\[
\begin{array}{c|c|c|c|c|c|c|c|c}
\hline
x & 0 & 5 & 10 & 15 & 20 & 5 & 10 & 15 \\
\hline
f_p & 0.02 & 0.04 & 0.06 & 0.08 & 0.1 & 0.12 & 0.1 & 0.06 \\
\hline
\end{array}
\]

· For \(x \in 0, \cdots, 11\), \(p(x) = \{0.00012, 0.00111, 0.00500, 0.01499, 0.03374, 0.06073, 0.09110, 0.11712, 0.13176, 0.13176, 0.11858, 0.09702\}, \& \ 1 - \sum x p(x) = 1 - .80301 = .19699.
· Could also add height spikes @ 12,13,..., but how far? \([p(100)=(3.5 \times 10^{-67})\), so maybe ignore past some pt, but what is domain of poisson? \(0 \rightarrow \infty\), so much easier just to add \(p(0..11)\)]
• **Continuous PDFs**

  – How conceive probability of outcomes for continuous RV’s?
    * For any specific $x$, $p(X = x) = 0$ (!)
    * but, somehow, $p(x_0 < X < x_1) \geq 0$ for $x_1 > x_0$.
    * **Relative Frequency (Probability) Density**:

      $$\text{RelativeFrequencyDensity} = \frac{\text{RelativeFrequency}}{\text{CellWidth}}$$

  ![Relative Frequency Histogram](image)

  * Bar-height in 4-3a is proportion cases in that bin.
  * Fig 4-3b rescales same info so total area sums to 1.
  * $\Rightarrow$ chunky histogram, so add obs doing same.
FIGURE 4-4
How relative frequency density may be approximated by a probability density as sample size increases, and cell size decreases. (a) Small n, as in Figure 4-3b. (b) Large enough n to stabilize relative frequencies. (c) Even larger n, to permit finer cells while keeping relative frequencies stable. (d) For very large n, this becomes (approximately) a smooth probability density curve.
Prob 4-17, WW p. 126)

4-17 The total time T that I wait for buses, on a long trip that includes transfer, has the following probability distribution: Note that the area of a triangle = base \times height/2, so that the total area or probability is $20 \times .10/2 = 1.00$.

![Probability Distribution Graph]

a. If I wait more than 15 minutes, I will be late for my appointment. What is the chance of this?
b. What is the mean waiting time?

* a. Wait > 15min ⇒ late; prob = what share area mins > 15
* a. Area triangle: base=5, ht=.05 ⇒ 5(.05)/2 = 5(.025) = .125
* b. Mean waiting time ⇒ mass balance-point of dist ⇒ 10.

• (Standard) Normal Distribution

- Synonyms:
  * Z dist—& z values we’ll discover & explore later.
  * Gaussian dist: Karl Gauss described dist errors astronomical obs (prob’ly known to LaPlace, De Moivre, & Jacob Bernoulli)
    - Why might Gauss have used it for astron-obs errs? What story might suggest such bell w/ 0 mean?
    - Prob’ly something like “most errors small, but some possibility big ones, increasingly unlikely as err size grows but never 100% impossible, equal chance +/- error, error sizes continuous.
  * Standard Normal≡Normal Dist w/ mean=0 & var=1.
- Famously Bell-Shaped. Symmetric (around 0). Unimodal (at 0). Mean=Median=Mode=0. Var=1. Thin tails.
- CONTINUOUS⇒ Probability Density Function
Rather ugly pdf, so use tables (or, nowadays, computers).

E.g., suppose R.V. Z generated by std norm dist. What prob observe value of $Z > 1.6$. Use table:

$$\Pr(Z > 1.60) = .055$$
– Some more examples:
\[
\Pr(1.60 < Z < 2.3) = \Pr(Z > 1.60) - \Pr(Z > 2.3) = .055 - .011 = .044
\]
\[
\Pr(Z < -2.50) = \Pr(Z > 2.50) = .006
\]
\[
\Pr(-1.96 < Z < 1.96) = 1 - 2(.025) = 1 - .05 = .95
\]
– But how often mean & sd gonna be 0 & 1?! Luckily, can rescale any norm dist [any mean, any variance] so becomes std norm.

– **Standardizing, Normalizing, Standard-Normalizing:**
  * Suppose R.V. X has mean $\mu$ & sd $\sigma$.
  * Make new R.V. $Z \equiv (X - \mu)/\sigma$
  * Note: Numerator centers to 0 mean, denominator squishes in or stretches out to var=sd=1, keeping area under sum to 1.

\[
Z = \frac{X - \mu}{\sigma}
\]

**FIGURE 4-7**

Standardization: A general normal variable (men’s heights) rescaled to a standard normal.

**FIGURE 4-8**

General normal rescaled to a standard normal.
* So, $Z$ measures # s.d.'s $X$ is from its mean.
* E.g., $X$ normal w/ $\mu = 16$ & $\sigma = 5$, what is $p(X > 20)$?

$$\Pr(X > 20) = \Pr(Z > \frac{20 - 16}{5}) = \Pr(Z > .8) = .212$$

- **Standard Normal PDF**: formula giving height (prob density) of std norm bell curve @ value $x$:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}, \text{ for } -\infty < z < \infty$$

* Other ways to write it: $p(z) = \phi(x) = f(x|0,1)$
* How might Gauss have guessed this?!
(General) Normal PDF: formula giving height (prob density) of norm bell curve @ value x

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}, \text{ for } -\infty < x < \infty \]

* Other ways to write it: \( p(x) = f(x|\mu, \sigma^2) \)

Other Continuous Distributions & PDFs

* PDF ≡ fnctn giving probability density at any outcome value.

* Continuous Uniform Distribution:

\[
 f(x) = \begin{cases} 
 \frac{1}{b-a}, & \text{for } a \leq x \leq b \\
 0, & \text{otherwise} 
\end{cases}
\]

* Example: \( a = 0 \) & \( b = 2 \), then pdf:

\[
 f(x) = \begin{cases} 
 \frac{1}{2-0} = \frac{1}{2}, & \text{for } 0 \leq x \leq 2 \\
 0, & \text{otherwise} 
\end{cases}
\]

* Calc some probs for U(0..2) distributed R.V.:

\[ \Pr(X > 1) = ? = .5 \]
\[ \Pr(X > 2) = ? = 0 \]

* Another Example:

\[
 f(x) = \begin{cases} 
 \frac{1}{5}, & \text{for } -1 \leq x \leq 4 \\
 0, & \text{otherwise} 
\end{cases}
\]
What’s \( \Pr(0 \leq X \leq 2) \)? \( 2 \times \frac{1}{5} = \frac{2}{5} \)

With curvier (tech term) dists, like normal, we’d use calculus (integrals) rather than algebra & geometry (sums, graphs) to find areas under curves, but same thing:

\[
\Pr(0 \leq X \leq 2) = \int_{0}^{2} f(x) \, dx = \int_{0}^{2} \frac{1}{5} \, dx = \frac{1}{5} x |_{0}^{2} = \frac{2}{5} - (0)(5) = \frac{2}{5}
\]

**Cumulative Distribution Functions (CDFs)**

- When sum (integrate) prob (density) fnctns, i.e., calc areas under curves up to some pt, \( x \), \( \Rightarrow \) **Cumulative Distribution Functions**.
- \( \text{CDF} \equiv \text{area under } f(x) \text{ up to } x; \ F(x) \equiv \text{CDF of RV } X \text{ eval’d @ } x. \)

\[
F(x) = \Pr(X \leq x), \quad \text{for } -\infty < x < \infty
\]

\[
F(a) = \Pr(X \leq a) = \sum_{\text{all } x_i \leq a} p(x_i)
\]

- Using these CDFs already, just not explicitly. Some Useful Rules:

\[
\Pr(X > x) = 1 - F(x) = 1 - \Pr(X \leq x)
\]

- For any given values \( x_1 \) and \( x_2 \) such that \( x_1 < x_2 \):

\[
\Pr(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)
\]
– CDF for std norm, using that $\phi$ notation:

* $\phi(x)$ draws curve. It gives @ value $z$ a curve height. $f(x|\mu, \sigma^2)$ does same for general normal dist.

* Neither tells area since just height (1D), not area (2D).

* So, how would CDF of std norm look? (often written $\Phi(x)$).

* Hey, that table stuff before was using $\Phi(x)$! Height of $F(x)$ (cdf) gives area under curve up to that value, $x$, of R.V.

– Example:

* X Normal w/ $\mu = 5$ & $\sigma^2 = 2$ (often written: $X \sim N(5, 2)$).

* What is $\Pr(1 < X < 8)$?

  · Draw $f(x)$, the areas, and our goal.
· But, first, rescale so can use std norm
\[
\Pr(1 < X < 8) = \Pr\left(\frac{1 - 5}{2} < \frac{x - 5}{2} < \frac{8 - 5}{2}\right) = \Pr(-2 < Z < 1.5)
\]
· Now, convert to \(\Phi\)'s conformable to Table:
\[
\Pr(Z < 1.5) - \Pr(Z < -2) = \Phi(1.5) - \Phi(-2) = \Phi(1.5) - [1 - \Phi(2)]
\]
\[
= (1 - .067) - [1 - (1 - .023)] = .933 - [1 - (.977)] = .933 - .023 = .91
\]
* So, been doing this; just diff notation & more explicit math now.
* How on computer? (in general)
\[
\text{pnorm}(8, \text{mean}=5, \text{sd}=2) - \text{pnorm}(1, \text{mean}=5, \text{sd}=2)
\]
* For a standard normal random variable:
\[
\Pr(-1 < Z < 1) = 1 - 2(0.16) = 1 - .32 = .68
\]
\[
\Pr(-2 < Z < 2) = 1 - 2(0.023) = 1 - .046 = .95
\]
\[
\Pr(-3 < Z < 3) = 1 - 2(0.00135) = 1 - .002701 = .9973
\]
* How interpret these numbers?
· 1\textsuperscript{st} line: 68% of area under std norm curve w/in 1 sd of mean,
· 2\textsuperscript{nd} line: 95% w/in 2 s.d. of mean,
· 3\textsuperscript{rd} line: 99.7% w/in 3 sds.
* I sense that plot thickening again...

• Two Quick Reviews

- The exponential, or natural, or \(e\)
  * \(p \equiv \text{principal}; r \equiv \text{annual rate}; n \equiv \# \text{ times/yr compounded}
  * At end 1\textsuperscript{st} interest period, have principal+interest: \(p + i = p + p \frac{r}{n} = p(1 + \frac{r}{n})\).
  * At end 1 yr, have \(A = p \left(1 + \frac{r}{n}\right)^n\).
  * E.g., \(p = $1 & r = 100\% = 1\), so \(A = 1(1 + \frac{1}{n})^n\) & interest compounded:
    · Bi-(semi-)annually (NOT biennially) \(\Rightarrow A = (1 + \frac{1}{2})^2 = 2.25\)
    · Quarterly \(\Rightarrow A = (1 + \frac{1}{4})^4 = 2.44141\)
    · Monthly \(\Rightarrow A = (1 + \frac{1}{12})^{12} = 2.61303\)
Daily ⇒ \( A = (1 + \frac{1}{365})^{365} = 2.71457 \)
Hourly ⇒ \( A = (1 + \frac{1}{8760})^{8760} = 2.71813 \)
Every second ⇒ \( A = (1 + \frac{1}{31,536,000})^{31,536,000} = 2.71828 \)
Instantly? Could do that. Leonhard Euler (/oiler/) did:
\[
\Rightarrow e \equiv \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = 2.718281828459045\ldots
\]
* In other words, \( e \) is a useful constant related to compounding.
* Logarithm w/ \( e \) as base (Log base-e; written \( \ln \) or often just \( \log \) in many fields, including stats, b/c nearly only log-base used).
* \( \pi \) is the ratio circle circumference to diameter: \( \pi \equiv 3.1415926\ldots \)
* \( e \) & \( \pi \) appear in many p-dists. For example: (std.) norm. p.d.f.:
\[
f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \text{ for } -\infty < z < \infty
\]
\[
f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \text{ for } -\infty < x < \infty
\]

− Distribution Functions
* Notation: \( \Pr(X = x) \) “probability R.V. \( X = \text{value } x \)”. Also written \( P(X = x) \) or \( p(x) \) or \( f(x) \).
* \( p(x) \) is probability function (pf) for discrete or probability density (or mass) function (pdf or pmf) of \( X \) eval’d @ \( x \).
* So \( p(1) \) means function \( p \) eval’d @ \( x = 1 \).
* \( F(x) \) is cumulative dist fnctn (cdf) of R.V. \( X \) eval’d @ \( x \).
\[
F(a) = \Pr(X \leq a) = \sum_{\text{all } x_i \leq a} p(x_i)
\]

Expectations & Linear Fnctns & Fnctns of Rndm Vars
* Preliminary Discussion:
− Say person’s ht related to vitamin A s/he ingests. Can write:
  − \( H = g(A) \) where \( H \equiv \text{ht}, A \equiv \text{vit A ingested}, \) & \( g() \) some function relating the 2 vars.
  − Maybe v.A & ht linearly related, so \( g(A) \) specific form (draw it):
    \[
    H = a + bA
    \]
– What does \( a \) do/mean here? And \( b \)?
– Hmmm, getting closer... Imagine: \( \text{Policy} = a + b(\text{MedVoterPref}) \)

**Expectations**

– So, more notation & definitions: \( \mu \equiv \text{expected value} \) of var; i.e., our best guess of value before realized, in min squared mistake sense.
– Written: \( \mathbb{E}(R) \); said “the expected value of \( R \).” So:

\[
\mathbb{E}(X) = \sum_{x:p(x)>0} xp(x) \quad \text{for Discrete RVs}
\]
\[
\int_{-\infty}^{\infty} xf(x)dx \quad \text{for Continuous RVs}
\]

– Hmmm, we’re getting closer... Imagine: \( \mathbb{E}(Y) = a + bX \) ...

**Some Properties**

– For \( a,b \equiv \text{constants}, X \equiv \text{RV} \)

\[
\mathbb{E}(b) = b \quad \mathbb{E}(a + bX) = a + b\mathbb{E}(X)
\]

– Expectation of Sum = Sum of Expectations: \( \mathbb{E}(\sum X) = \sum \mathbb{E}(X) \)

\[
\mathbb{E}(a + b_1X_1 + \ldots + b_nX_n) = a + b_1\mathbb{E}(X_1) + \ldots + b_n\mathbb{E}(X_n)
\]

– **If** 2 rvs are independent, then:

\[
\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \quad \text{and} \quad \mathbb{E}\left( \prod_{i=1}^{n} X_i \right) = \prod_{i=1}^{n} \mathbb{E}(X_i)
\]

– If \( X \) has pdf \( f(x) \) and \( g(X) \) is any function of \( X \) then:

\[
\mathbb{E}(g(X)) = \sum_{x} g(X)f(x), \text{if } X \text{ Discrete}
\]
\[
\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(X)f(x)dx, \text{if } X \text{ Continuous}
\]

**Elaborations, Illustrations, & Examples**
- \( H = a + bX \) (Ht = a + bVA). Then \( E(H) = E(a + bX) = E(a) + E(bX) = E(a) + bE(X) \) or \( \mu_H = a + b\mu_X \)

- If \( g(X) = X^2 \), then \( E(g(X)) = E(X^2) \) = \( \begin{cases} \sum_x x^2 f(x), & \text{discrete} \\ \int_{-\infty}^{\infty} x^2 f(x) dx, & \text{continuous} \end{cases} \)

- **Expectation of Functions of Random Variables**

- **WW Prob 4-29.** \( X \) discrete RV w/:

<table>
<thead>
<tr>
<th>x</th>
<th>p(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.5</td>
</tr>
<tr>
<td>4</td>
<td>.5</td>
</tr>
</tbody>
</table>

- So following true or false?
  - a. \( E(X + 10) = E(X) + 10 \)
  - b. \( E(X/10) = E(X)/10 \)
  - c. \( E(10/X) = 10/E(X) \)
  - d. \( E(X^2) = [E(X)]^2 \)
  - e. \( E(5X + 2)/10 = (5E(X) + 2)/10 \)

- So, if \( E(X) = 5 \), what is \( E(-3X + 15) \)?

- **Sums of RVs**

  - Sum of expectations of \( n \) RVs is expectation of their sums:
    \[
    E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n)
    \]

  - True regardless of whether RVs independent.
  
  - Imagine red & blue in box /w proportion \( p \), \( 0 \leq p \leq 1 \), red.
  
  - Draw w/ replace \( n \) times. Explain how figure expected # red.
    - What kind p-dist gens this RV? Relevant info for that RV?
      - Binomial. \( E(x) = np \). That was easy, eh?

- **Products of RVs**

  - What true re: sums not nec. true re: products. **Iff** indep:
    \[
    E\left( \prod_{i=1}^{n} X_i \right) = \prod_{i=1}^{n} E(X_i)
    \]
* Suppose 3 RVs, each from dist w/ mean 5.
  \[ E(2X_1 - 3X_2 + X_3 - 4) = ? \ldots \]
  \[ \ldots 2E(X_1) - 3E(X_2) + E(X_3) - 4 = 2 \cdot 5 - 3 \cdot 5 + 5 - 4 = -4 \ldots \]
  \[ E(2X_1 X_2 X_3) = ? \ldots \] dunno. Depends. Indep? Intuition?
* What if \( X_1, X_2, X_3 \) ind, \& spec’ly, \( E(X_i) = 0, E(X_i^2) = 1 \)
  \[ E \left[ X_1^2(X_2 - 4X_3)^2 \right] = ? \]
  \[ \text{Well, since indep, can pull them apart:} \]
  \[
  E \left[ X_1^2(X_2 - 4X_3)^2 \right] = E(X_1^2)E(X_2 - 4X_3)^2
  \]
  since \( E(X_i^2) = 1, = E(X_2 - 4X_3)^2 \)
  \[ = E(X_2^2 - 8X_2X_3 + 16X_3^2) \]
  \[ = E(X_2^2) - E(8X_2X_3) + E(16X_3^2) \]
  \[ = 1 - 8E(X_2)E(X_3) + 16 \cdot 1 \]
  \[ = 1 - 0 + 16 = 17 \]

**The Binomial & Normal & Latter Approx to Former**
* Remember Gauss wanted describe rel. freq. errs in astro. obs?
* He wanted a p-dist w/ some reasonable characteristics:
  · Sym @ 0, contin, unimod, most mass near 0, long-thin tails.
* ...but why/how on earth:
  \[
  f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}
  \]
  \[
  f(x) = \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \text{ for } -\infty < x < \infty
  \]
* Before Gauss got his name stuck to this dist by using it in least squares regression, some bored French aristocrats gambled a lot.
* Some games involved bets on complicated dice rolls or coin flips; e.g., “bet you I can get less than 10 of 40 heads.”
* OK, so, fair coin, \& independent flips; what is, say, \( P(\text{between 495 and 510 heads}) \) in 1000?
· Sum: $S_{1000} = H_1 + H_2 + \ldots + H_{1000}$, where $H_i$ indicates head given value of 1 & tail given 0.
· So, exper=coin flip. Event heads=H=1. Bernoulli w/ p=.5.
· Sum $n$ indep Bernoulli = Binomial:

$$p(s) = \binom{n}{s} \pi^s (1 - \pi)^{n-s}$$

$$\binom{n}{s} = \frac{n!}{s!(n-s)!}, \text{ where } n! = n(n-1)(n-2)\ldots 1$$

$$p(495 \leq S \leq 510) = p(495) + p(496) + p(497) + \ldots + p(510)$$

$$= \frac{1000!}{495!(1000 - 495)!} \cdot 0.5^{495} \cdot 0.5^{1000-495} + \ldots$$

· But no calculator or computer for these guys!!
· Hmmm, an approximation of some kind? Some formula to plug in 495 and 510 and 1000 and .5 get something close?
* Abraham De Moivre (ca. 1730s) mathematician of French extraction, but protestant so fled to England.
· De Moivre playing around w/ Binomial. What look like when $n = 3, 5, 10$?
· What about $n = 50, 100$?

Hmmm...Looking familiar? Instead of summing big factorials, could we integrate (area under)...something? But what? A parabola? Sorta, some quadratic function, but needs inflection points which parabola lacks, needs some sort of “swooping” tail-thinning/hill-piling.

· Or $n = 1000$?
Yup. Definitely getting normal-looking dist, no? So, say normal dist, & want prob values b/w 495 & 510. What do? [Tables, Computer] What need to know? expected value (aka weighted avg, mean) and std dev or var.

OK, what expected value of this binomial: sum of 1000 flips?

Well, we know \( E(S) = E(H_1 + H_2 + \ldots + H_{1000}) \)

And we also know expectation of sum equals sum of expectations. (Nice not to have to use summation or integrals here!).

So, \( E(S) = E(H_1) + \ldots + E(H_{1000}) \). So just add all the expects.

Let’s see, that’ll be \( E(H_1) = .5 \) plus \( E(H_2) = .5 \) plus... (Oh! \( p(\text{heads}) \) constant @ .5 \( \forall \) flips!).

So, \( E(S) = .5 \times 1000 = 500 \). (I.e., generally: \( E(S) = np \).)

And now need var. What’s var? [\( \approx \) avg dispersion or spread obs from mean] So it a kind of expected value too! Defined: VARIANCE: \( \text{Var}(X) = \sigma^2(X) = E(X - E(X))^2 \)

What does this mean? Say it in words...[expected value of squared differences b/w each X value & expected value of X].
· So, want \( \text{Var}(S) = \text{Var}(H_1 + H_2 + \ldots + H_{1000}) \). Well, know expectation of discrete RV is: \( \text{E}(X) = \sum_{x:p(x)>0} xp(x) \).

· So, for any discrete RV:

\[
\text{Var}(X) = \sum_{i} (x_i - \mu)^2 p(x_i)
\]

· And now some useful rules for algebra of variances:

\textbf{If } Y = a + bX, \textbf{ then } \text{Var}(Y) = b^2 \text{Var}(X)

\textbf{Iff 2 RVs indep., then: } \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)

· Let’s see how former derives from our expectations rules:

\[
\begin{align*}
\text{E}(Y - \text{E}(y))^2 & = \text{E}(a + bX - \text{E}(a + bX))^2 \\
& = \text{E}(a + bX - a - b\text{E}(X))^2 \\
& = \text{E}(bX - b\text{E}(X))^2 \\
& = \text{E}(b^2 X^2 - 2b^2XE(X) + b^2(X^2)) \\
& = \text{E}(b^2(X - \text{E}(X))^2) \\
& = b^2\text{E}(X - \text{E}(X))^2 \\
& = b^2\text{Var}(X)
\end{align*}
\]

· Back to the problem. So, because flips indep, can write:

\[
\text{Var}(S) = \text{Var}(H_1+H_2+\ldots+H_{1000}) = \text{Var}(H_1) + \ldots + \text{Var}(H_{1000})
\]

· And each of the \( H \) are Bernoulli RVs. So we know:

\[
\text{Var}(H) = \text{E}(H - \text{E}(H))^2 = \sum(H - \mu)^2 p(H)
\]

\[
= (1 - p)^2 \times p + (0 - p)^2 \times (1 - p)
\]

\[
= p^2 - p^3 + p - 2p^2 + p^3 = p(1 - p)
\]

· And so \( \text{Var}(S) = \sum \text{Var}(H) = np(1 - p) \).

* Whew. So, finally, expected value of this dist is \( \text{E}(S) = np = 1000 \times .5 = 500 \) & variance is \( np(1-p) = 250 \).
* So, we can find area under normal curve w/ mean 500 & var 250 b/w 495 & 510:
  
  \[
  \frac{510 - 500}{\sqrt{250}} \leq Z \leq \frac{510 - 500}{\sqrt{250}} \approx Pr(-.32 \leq Z \leq .63)
  \]

* Look it up in WW’s table at back: (1-(.374+.264))= .36.

* OK. So now know where normal curve comes from, and how used. Plus, mean & variance binomial RV.

### Joint distributions

- Sometimes interested how 2 (or more!) vars distributed together.
  - E.g., political interest & political participation. [more examples?]
  - WW’s example less interesting: X=# girls in family 3. Y=# of runs where run is an unbroken string of children’s gender. So, BBB=1 run (Y=1); BBG has Y=2. (Fig. 5-1, p. 155)

<table>
<thead>
<tr>
<th>(e)vent</th>
<th>p(e)</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBB</td>
<td>.14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BBG</td>
<td>.13</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>BGB</td>
<td>.13</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>BGG</td>
<td>.12</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GBB</td>
<td>.13</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GBG</td>
<td>.12</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>GGB</td>
<td>.12</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GGG</td>
<td>.11</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

- So, \( Pr(X = 1, Y = 2) = p(1, 2) = .13 + .13 = .26 \) More notation:

- **Joint distribution**, \( p(x, y) \equiv Pr(X = x \text{ and } Y = y) \)

<table>
<thead>
<tr>
<th>y</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>.14</td>
</tr>
<tr>
<td>1</td>
<td>.26</td>
</tr>
<tr>
<td>2</td>
<td>.24</td>
</tr>
<tr>
<td>3</td>
<td>.11</td>
</tr>
</tbody>
</table>
• **RULE:** \( p(x, y) = p(x)p(y) \iff x \& y \text{ independent.} \)
  
  – Now, \( p_x(2) \)? 0 + .24 + .12 = .36 (Write in table *margins...*)
  
  – **In general:** \( p(x) = \sum_y p(x, y) \)
    
    * Meaning? Summing over \( y \). Summing over values of \( y \) at particular value \( x \). Called: **marginal distribution** of \( x \).
    
    * So, can use joint distribution to restate some ideas from before.
      
      E.g., independence: \( X \& Y \text{ indep} \iff p(x, y) = p(x)p(y) \forall x, y \).
    
    * WW 5-4 a. Are \( X \& Y \text{ indep?} \) [fill in marginal probabilities]

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.1</td>
<td>.2</td>
</tr>
<tr>
<td>2</td>
<td>.3</td>
<td>.4</td>
</tr>
<tr>
<td>.4</td>
<td>.6</td>
<td></td>
</tr>
</tbody>
</table>

* If they were indep, joint probs would be proportional to marginal
  
  \[
  \begin{array}{ccc}
  1 & 2 \\
  .4(.3) = .12 & .6(.3) = .18 & \neq .2 & .3 \\
  \end{array}
  \]

  * So, in this table, \( p(x, y) \neq p(x)p(y) \iff X \& Y \text{ not indep.} \)
  
  – How depict joint distributions? Tables or graphs.

  * 2 discrete RVs: what look like? (Seen some; more later)
    
    * 2 continuous RVs: what look like? (Seen some; more later)

• **Functions of More than One RV: Discrete**

  \( r \equiv g(x, y) \Rightarrow E(r) = \sum rp(x, y) ; E(g(x, y)) = \sum_x \sum_y g(x, y)p(x, y) \)

  – Familiar? If just \( g(x) \& p(x) \), then just \( E(g(x)) = \sum_x g(x)p(x) \).

  \[
  0 \quad 2 \quad 4 \\
  \begin{array}{ccc}
  0 & .1 & .1 \\
  2 & .1 & .4 \\
  4 & 0 & .1 \\
  \end{array}
  \]

  – WW Problem 5-8: \( r = x^2 + y^2 \). And \( p(x, y) \) is:
* What is $E(r)$? first: $r$ has following values:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>20</td>
<td>32</td>
</tr>
</tbody>
</table>

* And:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>16</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.2</td>
<td>20</td>
<td>.2</td>
</tr>
<tr>
<td>8</td>
<td>.4</td>
<td>32</td>
<td>.1</td>
</tr>
</tbody>
</table>

* So: $E(r) = 0(.1) + 4(.1) + 16(0) + 4(.1) + 8(.4) + 20(.1) + 16(0) + 20(.1) + 32(.1) = 11.2$

* Reverse order ⇒ same:

$$
\sum \sum g(x, y)p(x, y)
$$

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(0+0).1</td>
<td>(4).1</td>
<td>(16)0</td>
</tr>
<tr>
<td>2</td>
<td>(4+0).1</td>
<td>(4+4).4</td>
<td>(4+16).1</td>
</tr>
<tr>
<td>4</td>
<td>(16+0)0</td>
<td>(16+4).1</td>
<td>(16+16).1</td>
</tr>
</tbody>
</table>

* Down columns: $(0*.1) + (4*.1) + 0 + (4*.1) + (8*.4) + (20*.1) + (20*.1) + (32*.1) = 11.2$

- **Functions of More than One RV: Continuous**

  * Example: $X$ & $Y$ uniform over $(0,1)$, $(0,1)$:
    * Uniform over square $S$; area under $S = 1$ ⇒ joint pdf $X$ & $Y$:
      \[
      f(x, y) = \begin{cases} 
      1, & \text{for } (x, y) \in S \\
      0, & \text{otherwise}
      \end{cases}
      \]

  * Now, we now that:

  \[
  E(X^2 + Y^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x^2 + y^2)f(x, y)dxdy \\
  = \int_0^1 \int_0^1 (x^2 + y^2) \cdot 1dxdy
  \]
\[ = \int_0^1 \left[ \int_0^1 (x^2 + y^2) \, dx \right] \, dy \]

\[ = \int_0^1 \left[ \int_0^1 x^2 \, dx + \int_0^1 y^2 \, dx \right] \, dy \]

\[ = \int_0^1 \left( \left[ x^3 / 3 \right]_0^1 + \left[ xy^2 \right]_0^1 \right) \, dy \]

\[ = \int_0^1 \left( (1/3) + y^2 - 0 \right) \, dy \]

\[ = \int_0^1 (1/3) \, dy + \int_0^1 y^2 \, dy \]

\[ = (1/3)1 + (1/3)1^3 - 0 = (1/3) + (1/3) = 2/3 \]

* So, this pt is what? Balancing pt., mean, \( \mu(x^2 + y^2) \)...again.

**COVARIANCE**

- We’ve used variance for single var; not big step to ask how 2 vars might **covary**: vary together.
- \( \text{Var}(X) \equiv \sigma_x^2 \equiv E(X - E(X))^2 = E(X - \mu)^2 \) Covar similar.
- **COVARIANCE**: \( \sigma_{x,y} = \text{Cov}(X, Y) = E(X - \mu_x)E(Y - \mu_y) \)

\[
\begin{array}{c|ccc}
 & 0 & 1 \\
\hline
0 & .2 & 0 & .2 \\
1 & .4 & .2 & .6 \\
2 & 0 & .2 & .2 \\
\hline
 & .6 & .4
\end{array}
\]

* \( \text{Cov}(x, y) \) aka \( \sigma_{x,y} \)? First need \( E(y) = \mu_y \) & \( E(x) = \mu_x \).
* So \( E(X) = \sum xp(x) = 0 \times .2 + .6 \times 1 + .2 \times 2 = 1 \), and \( E(Y) = .6 \cdot 0 + .4 \cdot 1 = .4 \)
* \( \sum (X - E(X))(Y - E(Y))p(x, y) \)

\[
\begin{align*}
&= (0-1)(0-.4).2 + (0-1)(1-.4)0 \\
&+ (1-1)(0-.4).4 + (1-1)(1-.4).2 = .08 + .12 = .2 \\
&+ (2-1)(0-4)0 + (2-1)(1-.4).2
\end{align*}
\]
Another example; different X & Y values:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>.2</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>.6</th>
<th>.4</th>
</tr>
</thead>
</table>

* New vars, $X^*$ & $Y^*$, same joint dist as X & Y. Just rescaled.

\[
E(X^*) = \sum xp(x) = 0 \times .2 + .6 \times 10 + .2 \times 20 = 10
\]

\[
E(Y^*) = .6 \times 0 + .4 \times 10 = 4
\]

* And so covariance...same marginal dist, but values change:

\[
= (0-10)(0-4) \cdot 2 + (0-10)(10-4) \cdot 0
+ (10-10)(0-4) \cdot 4 + (10-10)(10-4) \cdot 2 = -10(-4) \cdot .2 + (20-10)(10-4) \cdot 2 = 20
\]

* Are $X^*$ & $Y^*$ more strongly related? No, but Cov is higher.

* So, Cov \(\Rightarrow\) neg. or pos. relation, but magnitude=scale-sensitive.

* A standardizing rescaling would be helpful. How stdize univar. dists? Similar is helpful for multivar. & Cov. See below.

\[
\begin{array}{c|ccc}
X & 1 & 2 & 3 \\
\hline
0 & 1/8 & 0 & 0 \\
1 & 1/8 & 3/8 & 1/8 \\
2 & 1/8 & 3/8 & 1/8 \\
3 & 1/8 & 0 & 0 \\
\end{array}
\]

- One more to illustrate one more (5-12d, WW p. 168):

\[
\begin{array}{c|ccc}
Y & 1 & 2 & 3 \\
\hline
0 & 1/8 & 0 & 0 \\
1 & 1/8 & 3/8 & 1/8 \\
2 & 1/8 & 3/8 & 1/8 \\
3 & 1/8 & 0 & 0 \\
\end{array}
\]

* Cov? 1st find $E(X)$ & $E(Y)$ using marginal dists:

\[
E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2}
\]

\[
E(Y) = 1 \cdot \frac{2}{8} + 2 \cdot \frac{4}{8} + 3 \cdot \frac{2}{8} = 2
\]

* Can use $\sum (X - E(X))(Y - E(Y))p(x, y)$ to calc for discrete


\[
\begin{array}{c|ccc|c}
\text{x} & 1 & 2 & 3 & \\
0 & \left(0-\frac{3}{2}\right)(1-2)\left(\frac{1}{8}\right) & 0 & 0 & \left(\frac{1}{8}\right) \\
1 & 0 & \left(1-\frac{3}{2}\right)(2-2)\left(\frac{2}{8}\right) & \left(1-\frac{3}{2}\right)(3-2)\left(\frac{1}{8}\right) & \left(\frac{3}{8}\right) \\
2 & 0 & \left(2-\frac{3}{2}\right)(2-2)\left(\frac{2}{8}\right) & \left(2-\frac{3}{2}\right)(3-2)\left(\frac{1}{8}\right) & \left(\frac{3}{8}\right) \\
3 & \left(3-\frac{3}{2}\right)(1-2)\left(\frac{1}{8}\right) & 0 & 0 & \left(\frac{1}{8}\right) \\
\hline
\end{array}
\]

* What do we have here? \( (\frac{3}{16}) + (\frac{-1}{16}) + (\frac{1}{16}) + (\frac{-3}{16}) = 0 \)
  – If X & Y indep., then \( \text{Cov} = \text{Corr} = \sigma_{xy} = \rho_{xy} = 0 \). NOT vice versa.

**CORRELATION: Rendering Covariance Scale-Invariant**

\[
\rho \equiv \text{Cor}(X, Y) \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y} \equiv \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}
\]

– Convenient Fact/Property: \(-1 \leq \rho \leq 1\)
  * As correlation coefficient, *rho*, \( \rho \), goes from -1 to 1, say strongly negatively related or weakly positive, or whatever.
    · See WW for more on how this std-ization works.
    · Measure of strength of *linear* relationship b/w 2 vars. Extremely handy. Used every day in research, past, pres, future.
  – Clarify/Recall from before: \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \) *iff* X & Y indep. How/why:

  * Note: \( \text{Var}(X) = \text{Cov}(X, X) \)
  * So, \( \text{Var}(X + Y) = \text{Cov}(X + Y, X + Y) \) With me? OK, then:

\[
\begin{align*}
\text{Var}(X + Y) &= \text{Cov}(X + Y, X + Y) = E[(X + Y) - E(X + Y)](X + Y) - E(X + Y)] \\
&= E[(X + Y)^2 - E((X + Y)^2)] \\
&= E[X^2 + 2XY + Y^2] - 2XE(X) - 2XE(Y) - 2YE(X) - 2YE(Y) + \\
&\quad E(X)^2 + 2E(X)E(Y) + E(Y)^2] \\
&= E[(X^2 - 2XE(X) + E(X)^2) + (Y^2 - 2YE(Y) + E(Y)^2)] + \\
&\quad (2XY - 2XE(Y) - 2YE(X) + 2E(X)E(Y))] \\
&= E[(X - E(X))^2 + (Y - E(Y))^2 + 2(X - E(X))(Y - E(Y))] \\
&= E[(X - E(X))^2 + E[(Y - E(Y))^2] + E[2(X - E(X))(Y - E(Y))] \\
&= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)
\end{align*}
\]
* Cov\((X, Y)\) = 0 iff \(X\&Y\) indep, so formula reduces only then.

\[
\begin{array}{c|ccc}
\text{X} & 0 & 1 & .2 \\
Y & .2 & 0 & .6 \\
\text{Cov}(X,Y) & .6 & .4 & .2
\end{array}
\]

So, back to example. What’s the correlation?

\[
\text{Cov}(X,Y) = .2; \text{E}(X) = 1; \text{E}(Y) = .4. \text{ Var}(X) \text{ and Var}(Y)?
\]

\[
\text{Var}(X) = \sum(x - \mu)^2 p(x)
\]

\[
\text{So, for } X: (0 - 1)^2 .2 + (1 - 1)^2 .6 + (2 - 1)^2 .2 = .2 + 0 + .2 = .4.
\]

\[
\text{For } Y: (0 - .4)^2 (.6) + (1 - .4)^2 (.4) = .16(.6) + .36(.4) = .24
\]

\[
\text{So } \rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = .2/\sqrt{(.4)(.24)} = .65
\]

What about for the rescaled version?

\[
\text{Cov}(X,Y) = 20; \text{E}(X) = 10; \text{E}(Y) = 4. \text{ So, what’s Var}(X) \text{ & Var}(Y)?
\]

\[
\text{So } \rho = \frac{20}{\sqrt{(40)(24)}} = .65
\]

So Corr. Coeff. (\(\rho\)) scale-indep. or dimensionless.

**Visualizing Discrete & Continuous Joint PDFs**

Discrete:

* Let RVs \(X\) & \(Y\) have joint pdf

\[
h(x,y) = \frac{x + 1 - y}{54}
\]

* with domain of support

\[
\Lambda = \{(x,y) : x \in \{3, 5, 7\}, y \in \{0, 1, 2, 3\}\}
\]

\[
\begin{array}{c|cccc}
\text{Y} & 0 & 1 & 2 & 3 \\
\text{X} & 3 & 5 & 7 \\
\text{P} & \frac{4}{54} & \frac{3}{54} & \frac{2}{54} & \frac{1}{54} \\
\text{X} & 5 & 7 \\
\text{P} & \frac{6}{54} & \frac{5}{54} & \frac{4}{54} & \frac{3}{54} \\
\text{X} & 7 \\
\text{P} & \frac{8}{54} & \frac{7}{54} & \frac{6}{54} & \frac{5}{54}
\end{array}
\]
* Continuous (Specifically: Bivariate Normal)

- $x_1$ & $x_2$ jointly distributed bivariate standard normal (i.e., both std norm, w/ $\mu = 0$, $\sigma^2 = 1$, & $\sigma_{x_1, x_2} = \rho$

$$\mathbf{X} \sim N(\mathbf{0}, \Sigma) \text{ with } \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- This gives the following pdf:

$$f(x_1, x_2) = \frac{e^{-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2 + 2\rho^2}}}{2\pi \sqrt{1 - \rho^2}} \quad \left\{ (x_1 : -\infty \ldots \infty), (x_2 : -\infty \ldots \infty), (-1 < \rho < 1) \right\}$$
· What does it look like?
• What are the marginal densities (individ dists of \( x_1 \) & \( x_2 \))?

\[
    f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) \, dx_2
\]

\[
    f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) \, dx_1
\]

• And covariance:

\[
    \text{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) \, dx \, dy
\]

• Remember, integration & summation basically same, just 1 for discrete & 1 for continuous.

• **Correlation: Wrap-Up**

  – X & Y indep. \( \Rightarrow \) cov=corr=0. BUT Cov(X,Y)=0 \( \neq > \) X & Y indep. Not nec’ly. Only *linearly* indep. Could be non-linearly related.

  – Full Chain:

    * X & Y (Stochastically) Independent \( \Rightarrow \) X & Y Mean-Independent

      \( \Rightarrow \) X & Y Linearly-Independent (Cov=Corr=0).

      - \( f(X|Y)=f(X) \) & \( f(Y|X)=f(Y) \) \( \Rightarrow \) E(X|Y)=E(X) & E(Y|X)=E(Y) \( \Rightarrow \) \( \sigma_{xy} = \rho_{xy} = 0 \).

    * X & Y Linearly-Independent (Cov=Corr=0) \( \neq > \) X & Y Mean-Independent \( \neq > \) X & Y (Stochastically) Independent.

      - \( \sigma_{xy} = \rho_{xy} = 0 \) \( \neq > \) E(X|Y)=E(X) & E(Y|X)=E(Y) \( \neq > \) f(X|Y)=f(X) & f(Y|X)=f(Y).

  – Association only; not causation. “Correlation \( \neq > \) causation”

    * But association still very important & informative.

    * Who/why invented? Many people; many reasons.

      - Closely associated (corr’d: ha ha!) w/ Francis Galton, who obsessed w/ heredity studies, anthropology & like. Do tall parents produce tall kids. If so, why ethnicities long since divided into short & tall?

      - Bunch of English statisticians in mid-1800s. Some stats packages call \( \rho \) “Pearson’s correlation coefficient” after Karl Pearson (Galton’s protege) who derived some statistical tests using it.
Populations, Samples, Inferences: Back to Beginning

- Simple Random Sample (SRS): \( n \) obs, \( X_1, X_2, \ldots, X_n \) w/ \( f_1(\cdot) = f_2(\cdot) = \ldots = f_n(\cdot) = f(\cdot) \)
  - What does this mean? Key is pdf same for all obs. Meaning?
  - We use samples to draw inferences about populations.
    * “You don’t have to eat whole ox to know that it is tough.”
      - What is Samuel Johnson’s population here?
      - What is he trying to do by tasting a bit of ox?
      - How should he sample the ox?
  - So, what want next:
    * Want to be more precise now.
    * Want way to evaluate the sample estimate we get.
    * Will be handy for this if estimates of sample means, variances, covariances, correlations (\& coefficients we’ll get later) actually have distributions of their own.

Sampling Distribution

- Set-up
  - Say we had a normally distributed RV [chalkboard], w/ mean \( \mu \)
  - Say we draw sample of \( n = 5 \) & calculated sample mean \( \bar{x} \).
  - What is \( \bar{x} \)? It’s an estimate of \( \mu \).
  - What is \( \mu \)? The population-distribution mean. \( \text{TRUTH} \).
  - So: sample statistic as estimate of population parameter.
  - Now, get new stat-cum-est if draw another sample from same pop.
  - Imagine gathering many such est’s from \( n = 5 \) & plot histogram.
  - Now, do it all again, repeated sampling, but larger \( S, n = 10 \).
  - All again, lrgr sample, repeat many times \& new hist @ time.
  - Another, larger. Another, larger. What happens to hists as \( n \uparrow \)?
Hmm, if hists began to look like recognized probability (density) distributions, then we’d say the “sample statistic had a limiting **sampling distribution**” of that pdf it resembled.

Would all of these stats/ests, $\bar{x}_n$, equal $\mu$? [Rhetorical: *no*]

But, what’s mean of @ hist? Especially accurately so as $n \uparrow$, right?

**Law of Large Numbers**

First, sample mean is unbiased estimate of population mean:

$$E(\bar{x}) = \mu_x$$

$$E(\bar{x}) = E\left(\frac{1}{n} \sum x_i\right) = \frac{1}{n} \sum E(x_i) = \frac{1}{n} \sum \mu_x = \frac{1}{n}(n\mu_x) = \mu_x$$

I.e., Expected value (mean) of sample mean = population mean.

I.e., “on average” samp mean as *estimator* gets pop mean right.

* Still clueless whether any partic est any good. Just: if we could draw bazillion SRS, we’d be right on average across those.

* Anything else might like to know re: *sampling distribution* of this estimate?

**Weak Law of Large Numbers (LLN)**

Many theorems re: how sample stats behave as sample sizes grow.

* Most important involves sample mean as $n \to \infty$.

* Implication of LLN: more sample info really does tell more re: pop. (Crazy world if not)

**Weak Law of Large Numbers (LLN) Theorem:**

* Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed RVs w/ finite means, $E(X_i) = \mu, -\infty < \mu < \infty$, & finite variances, $Var(X_i) = \sigma^2, 0 < \sigma^2 < \infty$, then $\bar{X} \xrightarrow{p} \mu$.

  * I.e., sample mean *converges in probability* to pop mean.

  * I.e., as $n \uparrow$, $\bar{X} \xrightarrow{p} \mu_\cdot$, w/o error (i.e., exactly, w/ 0 var).
Preliminaries: Before proving theorem, need some prior results:

- First, notice that \( E[(\bar{X} - \mu)^2] \) is \( \text{Var}(\bar{X}) \)
- and \( \text{Var}X_i \) can be written as \( \sigma^2 \).

\[
\text{Var}(\bar{X}) = \text{Var}\left(\frac{\sum_i X_i}{n}\right) \\
= \text{Var}\left(\frac{X_1}{n}\right) + \text{Var}\left(\frac{X_2}{n}\right) + \ldots \\
= \frac{1}{n^2}\text{Var}(X_1) + \frac{1}{n^2}\text{Var}(X_2) + \ldots \\
= \frac{1}{n^2}(\sum_i \text{Var}(X_i)) \\
= \frac{1}{n^2}(n \cdot \sigma^2) = \frac{\sigma^2}{n}, \tag{20}
\]

- and \( \sqrt{\text{Var}(\bar{X})} \) is called **standard error** of \( \bar{X} \).
- and, finally, recall that:

\[
E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} \sum E(X_i) = \frac{1}{n}(n \cdot \mu) = \mu
\]

- Second, introduce **Chebyshev’s Inequality** which states

\[
\text{Pr}[(|X - \mu|) > t] \leq \frac{\sigma^2}{t^2} = \frac{\text{Var}(X)}{t^2} \tag{21}
\]

- Explanation: “if \( \sigma^2 \) is very small, there is a high probability that \( X \) will not deviate much from \( \mu \).” (Rice ‘95, p. 125)

- Proof of Weak LLN:

  * First, notice \( \bar{X} \xrightarrow{p} \mu \) is equivalent to \( \text{Pr}(|\bar{X} - \mu| > \varepsilon) \to 0 \) as \( n \to \infty \).
  * By Chebyshev’s Inequality, (21), we can write:

\[
\text{Pr}(|\bar{X} - \mu| > \varepsilon) \leq \frac{\text{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n} \leq \frac{\sigma^2}{n\varepsilon^2} \to 0, \text{ as } n \to \infty
\]

  * So, since \( \text{Pr}(|\bar{X} - \mu| > \varepsilon) \leq \) something that \( \to 0 \) as \( n \to \infty \),
then it too must \( \to 0 \) as \( n \to \infty \). QED.
• **Standard Error**
  
  – Consider 2 cases:
    * What if we took samples of size \( n = 1 \)?
    * What if we took samples of size \( n = 1,000,000 \)?
    * How would the 2 sampling distributions of sample mean differ?
  
  – General Rule: typical deviation a lot smaller when huge random samples than when tiny ones.
    * If random sample of 4 people, could be representative or *weird*.
    * If rndm samp 1M, much more likely to be rep & much harder to be weird b/c more certain to pull info from all over pop.
  
  – The typical deviation of an estimate is *standard error*:

\[
\text{Standard Error } (\bar{X}_n) = \frac{\sigma}{\sqrt{n}}
\]

  * See derivation above (i.e., \( \sqrt{20} \))
  
  – Example: WW 6-6 and 6-7 (page 198)
    * Pop of US men 1975 income avg \( \mu = 10,000 \) w/ \( \sigma = 8,000 \).
      * If rndm samp \( n = 100 \) drawn to est \( \mu \), what is std.err.\( (\bar{x}) \)?

\[
\sigma/\sqrt{n} = 8000/\sqrt{100} = 8000/10 = $800
\]

      * Male pop in ’75 was 78M. Say 1% sample taken (\( n = 1\% \cdot 78M = 780,000 \)). What is s.e.\( (\bar{x}) \)?

\[
\frac{8000}{\sqrt{780000}} \approx \frac{8000}{883} \approx $9
\]

      * Say male pop CA was 1/10 of US male pop. Say 1% sample CA men drawn? What is s.e.\( (\bar{x}) \)?

\[
n = \frac{780000}{10} = 78000 \Rightarrow 8000/\sqrt{78000} \approx 8000/279 \approx $29
\]
• **Central Limit Theorem**

  – Set-up Discussion

  * So, as sample size ↑, sampling distribution gets *skinnier*.
  * More amazingly, sampling dist of sample means gets skinnier in very partic way: it increasingly well approximates normal dist.
  * Very useful b/c can characteristics of Normal dists to get much understanding from s.e.(sample means), for example.
    · E.g., how much area in sampling distribution of sample mean lies w/in 1 se of mean?
    · If sample size huge, then overwhelming share probability in sampling distribution is where?
    · So, with huge sample, likelihood that our 1 sample mean (*est*) close to (unknown, *true*) pop mean ≫ than w/ tiny sample.
  * Phenomenon has name: **Central Limit Theorem (CLT)**. (Saw it in action w/ coin flips etc. above.) It’s Meaning:
    · “Whenever a random sample of size *n* is taken from *any* distribution with mean *µ* and variance *σ*², the sample mean *x̄* with have a distribution that is *approximately* normal with mean *µ* and variance *σ*²/*n*.” (Degroot, 282)
  * So what?
    · If sampling dist of est matches well-known prob dist, esp. norm known so well, can make prob statements about ests.
    · (Proving sampling dists for estimators: what statisticians do.)
    · w/o computers, how calc these probs? w/ them how calc?
  * So why do we use it?
    · The CLT (or the *Generalized CLT*) gives leverage to evaluate pretty much any # we est — to evaluate most formulas we use to estimate things, i.e. to evaluate *estimators*.
    · So ability to draw inferences — *any inferences* — from results will depend crucially on fact that this CLT phenom occurs.
· Gives ability to talk about how sure we are of #s we estimate.
  * So how do we use it? What do we need?
    · Random Sample.
    · Large Sample.
    · o/w, very difficult (not imposs.) get from sample to pop.
    · (Uh-Oh. Discuss.) How large is large enough? (Discuss)
  * OK. Random (or “fixed” thereto) & Large (enough). Now What?
    · Use z-tables to compare sample to pop: 
      \[ z = \frac{\bar{x} - \mu}{se} \]
  * Comparing LLN & CLT (first go-through):
    · LLN: samp means ever better approx of pop means as \( n \) bigger.
    · Nothing re: samp dist samp means, so can’t make p-states.
  * CLT: history
    · Abraham de Moivre found RV “sum of heads” (Binomial) indistinguishable from normal dist when # flips large.
    · 150yrs(!) before Lyapanov able prove more gen thrm: samp dist of samp means any RV converges in dist to Normal.\(^1\)
    · So long b/c proof requires show everything re: p-dist of sample means becomes identical to everything re Normal p-dist, but sampling dist of sample means is unknown!
    · Need more general ways characterizing shapes of p-dists.
    · Mathematicians have several tools that allow close approximation shape unknown functions, e.g. Fourier Transform.
    · Statisticians adapted \( FT \), & similar tools, for stat purposes.
    · Easiest proof of CLT uses moment generating functions or generalization, characteristic functions, both much like \( FT \).
    · Not going to cover these, so won’t prove any CLT, rather try get thrm intuition, & show example comparing CLT & LLN.

\(^1\)An RV converges in distribution to some specified dist means that p-dist of RV ever more like specified. I.e., not just RV converges in probability (i.e. becomes ever closer to some specified single value), but that all characteristics of RV’s p-dist (mean, variance, & all other moments) ever closer to those of specified dist.
– **Central Limit Theorem (CLT):** \(X_1, X_2, \ldots = \text{sequence iid RVs w/ finite } E(X_i) = \mu \text{ & finite } Var(X_i) = \sigma^2 > 0.\) Then:
\[
\sqrt{n}(\bar{X} - \mu) \overset{d}{\to} N(0, \sigma^2).
\]

* This version CLT regards centered \(\bar{X}\) (i.e. mean of resulting RV=0).
* Can restate after stdzng deviations from \(\mu\) familiar way:
\[
\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \overset{d}{\to} N(0, 1).
\]

– Note: Some disjuncture b/w way folks talk CLT & how thrm works.
  * “Sampling distribution of sample mean is Normal,” but...
  * LLN \(\Rightarrow\) sampling dist \(\bar{X}\) *degenerate*, i.e. collapses to 1 pt, \(\mu\);
  * plus, thrm stated in diff b/w smpl & pop mean, scaled by \(\sqrt{n}\).

– Discussion: Consider using just mean if goal is make p-states:
  * E.g., think population is US pop today. As \(n \to\) US pop (full census), \(\bar{x} \to\) true US pop.
  * E.g., think coin flips. As \(n \to \infty, \bar{x} \to \infty\) (i.e., \(\to \mu = np\))
  * So, as per LLN, as \(n \to \infty, \bar{x} \to 1\ #, not a dist \(\Rightarrow\) no p-states.
  * So, two problems to redress (to pursue our goal):
    · Prevent \(\bar{x}\) from drifting to \(\infty\) if wants, & from
    · collapsing to point, which it will want (LLN).
  * Solutions reasonably intuitive:
    · Focus on distance b/w samp & pop means, \(\bar{X} - \mu\), thus fixing & centering dist at mean zero., so it cannot go to infinity.
    · But, \(\bar{X} \overset{p}{\to} \mu\) (LLN), so this now \(\bar{X} - \mu \overset{p}{\to} 0\).
    · To keep \(\bar{X} - \mu\) from 0 as \(n \to \infty\) suggests multiply by \(n\) somehow, so as LLN shrinks \(\bar{X} - \mu\), the \(\times f(n)\) reinflates it.
    · Want inflate by \(n \atop \text{same rate that } n \text{ shrinks } \bar{X} - \mu\).
    · Oh! We know that. Standard errors shrink in \(n\) at rate \(\sqrt{n}\!\!\).!
    · \(\Rightarrow\) particular function of sample mean that not collapse as \(n \to \infty\), so non-degenerate sampling dist (i.e. a p-distribution w/o all its density at single pt).
CLT formula looks lot like numerator we used to stdz an RV: \( \frac{x_i - \bar{x}}{sd_x} \).

* CLT based on that formula, where RV for CLT is \( \bar{X} \) & use std err (i.e. \( \sqrt{\text{Var} \bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \)), rather than std. dev (i.e., \( \sigma \)).

* So formula for stdzd version of \( \bar{X} \), called \( Z \), is

\[
Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}.
\]

* So CLT says that \( Z \xrightarrow{d} N(0, 1) \), i.e. as \( n \to \infty \), \( Z \) converges in distribution to the standard normal.

Figure below contrasts action of LLN & CLT using series of histos.

* Histo at top, shows population dist (here \( X \) is \( \chi^2(2) \) dist’d RV).

* LLN:

  - Left column shows how sampling dist of \( \bar{X} \) changes as \( n \) increases 10 → 10K (each histo from 1000 samples size \( n \)).
  - Clear that sampling dist \( \bar{X} \) by itself eventually collapses to pop mean of 2 as \( n \to \infty \) as per LLN.

* CLT:

  - Right column shows sampling dist of \( \sqrt{n}(\bar{X} - \mu) \).
  - Does not collapse as \( n \to \infty \).
  - Each panel shows RV histo in blocks, smooth estimate density function as line, & Normal pdf as other, smoother line.
  - As \( n \) increases, ragged curve approximates smoother one ever more closely as per CLT.

**Other Statements of the CLT**

* Our way above follows Goldberger (1991): \( \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \xrightarrow{d} N(0, 1) \)

* DeGroot (1986, p. 275): “If a large random sample is taken from any distribution with mean \( \mu \) and finite variance \( \sigma^2 \), regardless of whether this distribution is discrete or continuous, then the distribution of the random variable \( n^{-1/2}(\bar{X}_n - \mu)/\sigma \) will be approximately normally dist’d with mean \( \mu \) and variance \( \sigma^2/n \).”
Figure 1: Sampling Distributions of $\bar{x}$ and $\sqrt{n}(\bar{x} - \mu)$
∗ So, DeGroot basically replaces “asymptotically standard-normal” /w “approximately normal”

∗ Ross (2000, p. 77): “Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables each with mean $\mu$ and variance $\sigma^2$. Then the distribution of the random variable

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma \sqrt{n}}$$

Tends to the standard normal as $n \to \infty$. That is,

$$\Pr \left\{ \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma \sqrt{n}} \leq a \right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$
as as $n \to \infty$.”

∗ NOTE: $\sum X_i = n\bar{x}$, so $\frac{\sum x - n\mu}{\sigma \sqrt{n}} = \frac{n(\bar{x} - \mu)}{\sigma \sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$.

∗ Grimmett (1992, p. 175) “Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite means $\mu$ and finite non-zero variances $\sigma^2$, and let $S_n = X_1 + X_2 + \ldots + X_n$. Then (equiv. by same steps as above):

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} N(0, 1) \text{ as } n \to \infty.$$”

• Standardized Variables

$$z = \frac{\bar{x} - \mu}{\text{SE}}$$

− Question: WW Prob 6-11 (p. 206):

∗ “1-lb” packages filled by well-worn machine have weights that vary Normally around mean 16.2 oz, w/ sd of .12 oz.

∗ Inspector randomly selects few packages from production line to see whether avg wt $\geq 16$ oz. If not, firm faces $500 fine.

∗ What is chance of fine if $n \in (1, 4, 16)$

− Answer:

∗ Start: What want to know? That sample mean (inspector’s test) no more than .2 oz below mean. I.e., want $\Pr(\bar{x} < 16)$. 

Fall 2005 – Rob Franzese
* To use tables, convert what want know to z-scores \( \Pr(z < \frac{16-16.2}{se}) \).
* Note: Stdzng thus not just crazy back-of-bookishness. Folks use critical values of std norm dist so often come to carry probability info by selves: “\( z > 1.96 \)” means “probability < 95%”. So, kind of universal scientific language re: uncertainty. Often-used p-dists become sort of universal, common language. Helpful.
* So, to stdz, need mean & std err (spread of samp dist of est). Given mean & std dev (spread of RV itself). Can go from there:

\[
\begin{align*}
\text{n=1} & \quad \text{se=?} \quad (\frac{.12}{\sqrt{1}} = .12) \\
\text{n=4} & \quad \text{se=?} \quad (\frac{.12}{\sqrt{4}} = .06) \\
\text{n=16} & \quad \text{se=?} \quad (\frac{.12}{\sqrt{16}} = .03)
\end{align*}
\]

\[
\begin{align*}
\text{n=1} & \quad \Pr(z < 2/12) = \Pr(z < -1.67) = \Pr(z > 1.67) = .047 \\
\text{n=4} & \quad \Pr(z < 2/06) = \Pr(z < -3.33) = \Pr(z > 3.33) = .000483 \\
\text{n=16} & \quad \Pr(z < 2/03) = \Pr(z < -6.66) = \Pr(z > 6.66) \approx 0
\end{align*}
\]

– Conclusions?
* Inspector should draw large enough sample, so odds of fining when shouldn’t would be really small.
* Producer should set mean for machine high enough given sd & inspector smpl-sizes. (etc.)

**Proportions**

– What is relationship b/w proportions & means?
– Proportions are means of Bernoulli RVs
– So prop is a mean. So no shock to learn CLT works for props too.
– WW many pp trying persuade you what already know (u so smart!).
– That’s why things worked as did for De Moivre, btw.

**Monte Carlo Simulations**
– Sampling dists so wonderful that people devised ways to view them even when math too complicated.

– *Monte Carlo Simulation* does by generating many RV obs & many random samples. Quite common now b/c computers⇒quick&easy.
  * Say you know something about pop; a pdf would be very useful.
  * Have comp gen rndm samp of whatever size from that pdf.
  * Calculate $\bar{x}$ (or whatever stats/ests of interest for each sample.
  * Have computer draw another sample & calc $\bar{x}$ again.
  * Keep doing it until can draw sufficiently detailed histo for needs.
  * Look at empirical dist $\bar{x}$ & “Hey, looks normal”, “biased”, or...
  * If sampled enough, could make p-states right there just by counting numbers of samples where whatever interests you happens.

• **Review and Cleaning Up:**

– Sampling dists used to give info re: gen properties of ests.
  * W/o them, can’t say much about how to evaluate ests.

– **Estimator**: a function (formula) that produces from data estimate(s) for unknown parameter(s) of dist that produced the data.
  * So formulas for sample mean (avg) or sample standard error or a coefficient in a regression equation are all estimators.

– **Estimate**: one realization of estimator applied to data ($\Rightarrow$ a #).

– So, how use this stuff in practice?
  * When we get our fabulous rndm samp & use its data to calc samp mean (or coeffs, or...), probably only get 1 samp. Not 1 bazillion samples & 1B ests. 1 sample & 1 (set of) est(s).
  * So, when we calc that est, we seek ways to ask whether it’s good & convey meaningful answers to that Q to other scientists.
  * Knowing sampling distributions (or estimating them too empirically or by simulation) is indispensable in that.

– **Asymtopia**: No one lives there, so what good is it?
Sampling Distributions & Properties of Estimators:

• Background & Preliminaries
  – On being an academic: mostly learning on own. E.g., PS 599 & 699, p-sets cover topics not directly or fully covered in lects or sects.
    * ⇒ frustration, but pt to learn self-sufficiently & gain self-sufficiency.
    * In practice, will face your own stat & research-design probs...
    * Can ask help, & friendly neighborhood methodologist: “Never seen exactly this, but seems like that, so look into selection...”
    * Off you go to read & learn about ..., make own judge, figure out.
    * After 599 & 699, expect able pick up anything (w/ work & help).
  – Reminder Questions:
    * What is LLN; what does it say?
      · As sample size $n \to \infty$, $\bar{x}$, i.e., avg/smpl-mean, $\to$ **TRUTH**, exactly, w/o error, no variance
    * What is CLT; what does it say?
      · As $n \to \infty$, some partic fnctn of $\bar{x} \to$ a normally dist’d RV.
    * Why do we care?
      · ability talk about $\mu$ & $\sigma^2$ (across rptd samples) of ests/stats
    * What is sampling distribution?
      · distribution across repeated samples of those ests/stats
    * What is an estimator?
      · a function, a formula estimating some parameter of a distribution when applied to data drawn from that dist.
    * What is a parameter?
      · “a characteristic or combo char’s that determine dist generating data called **parameter** of dist” (DeGroot&Shervish). E.g.?
    * Why care re: sampling dists of estimators for parameters?
      · This is what empirical research does: estimate parameters of **truth** (true dists) using data.
To say ests dist’d such way allows to say how strong evidence, & beyond “is not”/“is so”, he said/she said, w/ might⇒right.

– So: properties of (i.e., evaluating candidate) estimators...

• Bias

– **Parameters**, θ, ˜θ, or Θ, rep some property(s) of p-dist, e.g. mean.
– An estimator (formula), label it U, is an **unbiased** estimator, i.e. yields unbiased estimates, of θ iff E(U) = θ.

\[ U \text{ unbiased estimator/estimate of } \theta \iff E(U) = \theta. \]

– What does this mean?
  * Estimator U = *true* θ ON AVERAGE (across what?).
  * So, unbiasedness not re: any 1 given est U. Re: sampling dist U; Sampling Dist centered on true θ.
– So: **bias(U) = E(U) − θ**.
  * Bias can arise from bad formula (estimator) [e.g.?], or
  * poor match of est assumpts to data-gen conditions [e.g.?].
– Another common way write **point estimate** parameter: ˆθ.
– Example biased estimator: ˆθ = 42 ... biased whenever θ ≠ 42.
– Examples:
  * Estimate # voters pro-“Arnie for Pres in ‘12” in sample (n) data.
    · What is outcome RV’s distribution?
    · What is/are the relevant parameters of that distribution?
    · Which do we wish to estimate?
    · Can you suggest an estimator? How about:

\[ \hat{\pi} \equiv p = \text{number of Arnie voters/total number of voters} = x/n \]

· Assuming x, number Arnie-backers, from some Binomial p-dist, \( B(n, \pi) \), i.e. \( X \sim B(n, \pi) \), what’s true mean?
So, proposed estimator \( p \) really yield \( \pi \) on avg (i.e., in expect)?

\[
E(\hat{p}) = E(X/n) = \frac{1}{n}E(X) = \frac{1}{n}n\pi = \pi
\]

So, sample prop = unbiased est pop probability.

* What about sample mean \( \bar{X} \equiv \frac{\sum x_i}{n} \) in gen, for \( \mu \)?
  * For any RV: \( E(X) \equiv \mu \equiv \sum x_ip(x_i) \). So \( E(\bar{X}) = E(X) \)?

\[
E(\bar{x}) = E(\frac{\sum x_i}{n}) = \frac{1}{n}E(\sum x_i) = \frac{1}{n}(E(x_1) + \ldots + E(x_n)) = \frac{1}{n}n\mu = \mu
\]

* What if new estimator for \( \mu \): \( \bar{x}^* \equiv \frac{\sum x_i + 3}{n} \). Unbiased?

\[
E(\bar{X}^*) = E(\frac{\sum x_i + 3}{n}) = \frac{1}{n}E(\sum x_i + 3)
\]

\[
= \frac{1}{n}(E(x_1 + 3) + \ldots + E(x_n + 3))
\]

\[
= \frac{1}{n}(E(x_1) + 3 + \ldots + E(x_n) + 3)
\]

\[
= \frac{1}{n}n(\mu + 3) = \mu + 3
\]

So bias? \( E(\hat{\theta} - \theta) = E(\bar{x}^*) - \mu = \mu + 3 - \mu = 3 \).

– Sum: Bias=“right on avg.” Nice prop for est, but not only 1...

– **Famous Unbiasedness Joke:** 3 econometr-/statist-icians hunting, 1\(^{st}\) shoots high/left/right/ahead of prey by \( x \), 2\(^{nd}\) opp dir by \( x \), 3\(^{rd}\) yells, “Got ‘im”!

  * Point(s) of joke:
    * Unbiasedness ain’t everything.
    * **Unbiased** notion depends on multiple-samples notion. In only 1, can be cold comfort to know “rt on avg” across mult-samples.
    * Key other feature you’d like for warmer comfort?

**Efficiency**

– Another desirable characteristic estimators: skinny sampling dist. I.e., small variance across repeated samples. Called **efficiency**.
* What’s so good about that?
* What can you do to increase efficiency?
* So, is \( U = 3 \) efficient?
* So efficiency ain’t everything either. Unbiased & efficient: still not quite everything, but really nice!

- Example (WW 7-3): Economist gathers random sample 500 obs & loses records of last 180, leaving 320 obs from which calc sample mean. What’s efficiency relative to whole sample?
  * Need an assumpt even to start: *Missing At Random* — if not, lost representativeness, so biased.
  * Efficiency ≡ variance (across repeated samples), so:
  * Efficiency of sample mean w/ 320 obs = \( \text{Var}(\bar{X}_{320}) \)
  * Efficiency of full-sample sample-mean: \( \text{Var}(\bar{X}_{500}) \)
  * Relative Efficiency (n.b., efficiency (almost) always relative):
    - Same \( \sigma^2 \) since from same population (Missing at Random):
      \[
      \frac{\text{Var}(\bar{X}_{500})}{\text{Var}(\bar{X}_{320})} = \frac{\sigma^2/500}{\sigma^2/320} = \frac{\sigma^2}{500} \cdot \frac{320}{\sigma^2} = .64 \text{ or } 64% 
      \]
    - So smaller-sample mean 64% as efficient as full-sample mean; i.e., larger-sample variance 64% of smaller-sample variance.
    - NEVER lose/waste information!

- Example (WW 7-7): For long-term plan, auto exec commissioned 2 ind. SRS to est prop, \( \pi \), owners intending buy smaller next car.
  * 1st survey prop: \( p_1 = 60/200 = 30% \).
  * 2nd survey prop: \( p_2 = 240/1000 = 24% \).
  * Proposed overall estimate=avg: \( (p_1 + p_1)/2 = p^* = .27 \).
  * What are variances (efficiency) of these estimates?
    - Assume SRS, so \( \text{Var}(p) = p(1 - p)/n \)
    - Prop’d Est: \( U = \frac{1}{2}(p_1) + \frac{1}{2}(p_2) \), so \( V(U) = V[\frac{1}{2}(p_1) + \frac{1}{2}(p_2)] = ? \)
      Remember: \( V(aX + bY) = a^2V(X) + b^2V(Y) + 2abC(X, Y) \)
\[ V\left[ \frac{1}{2}(p_1) + \frac{1}{2}(p_2) \right] = \frac{1}{4} \left( \frac{.30(.70)}{200} \right) + \frac{1}{4} \left( \frac{.24(.76)}{1000} \right) + 0 \]

\[ = \frac{1}{4}(.00105 + .0001824) = .0003081 \]

\[ \cdot V(1^{st}) = .30(.70)/200 = .00105 \]
\[ \cdot V(2^{nd}) = .24(.76)/1000 = .00018 \]
\[ \cdot V(.5(1^{st}) + .5(2^{nd}) = .0003081 \]

* Hmmm, \( \frac{\text{Var}(p^*)}{\text{Var}(p_2)} = \frac{.0003081}{.0001824} \approx 1.69 \), so 2\textsuperscript{nd}-sample only 69\% more efficient avgd! Avg of ests inefficient rel. to 2\textsuperscript{nd}, lrgr sample, so discard 1\textsuperscript{st}, smaller smpl? What of “Waste no info?”

* Best estimate uses all info, counting each \textit{obs} equally, not each \textit{sample} equally! Pooling samples:

\[ p^{**} = \frac{60 + 240}{200 + 1000} = .25\% \]

\[ \text{Var}(p^{**}) = \frac{.25(.75)}{1200} = .00015625 \]

Relative efficiency \( p^{**}/p^* : \frac{\text{Var}(p^*)}{\text{Var}(p^{**})} = \frac{.0003081}{.00015625} \approx 1.97 \Rightarrow 97\% \).

* True or false: Important to know reliability of your sources. As seen here, if unreliable source not discounted appropriately, can be worse than simply discarding. So true.

– Summary:

* Efficiency≡(relative) variance of est across repeated samples.
* So, efficiency helps choose among unbiased estimators.
* WW compare smpl med&avg here: both unbiased (as ests \( \mu \)) if symmetric dists (if asymm?), mean more efficient some circumstances (normal dists) median more in others (fat-tailed dists).

**MSE**

– Combining Unbiasedness & Efficiency:
For estimator $U$ of population parameter $\theta$:
- $U$ is an **unbiased** estimator of $\theta$ $\iff$ $E(U) = \theta$
- Bias $\equiv E(U) - \theta$
- $U$ is a relatively **efficient** estimator compared to $V$ (another estimator) $\iff$ $\text{Var}(U) < \text{Var}(V)$

Comparing biased & unbiased ests:
- Silly always prefer least-variance: Thin sampling dist around wrong value not comforting.
- Almost as silly always pref unbiased: Right on avg but really flat sampling dist $\Rightarrow$ no idea where any given est might be.
- So, criterion that combines bias & efficiency somehow?
- Mean Squared Error (MSE) $\equiv E(\hat{\theta} - \theta)^2$

* MSE of est about some val, usu. about truth, $\theta$, & just say MSE.

* Some math $\Rightarrow$ MSE=$(\text{Est’s Variance})+(\text{Est’s Bias})^2$

  - First, another way to express $\text{Var}(X)$:
    $$\text{Var}(X) = E(X - E(X))^2 = E(X^2 - 2E(X)X + (E(X))^2)$$
    $$= E(X^2) - 2E(X)E(X) + (E(X))^2$$
    $$= E(X^2) - 2(E(X))^2 + (E(X))^2$$
    $$= E(X^2) - (E(X))^2$$

  - So $\text{Var}(X) + (E(X))^2 = E(X^2)$.

Now, can show MSE as def’d above actually a combo of var & bias (as claimed).

$$\text{MSE} \equiv E[(\hat{\theta} - \theta)^2]$$

Given the above, this $E(X^2)$ can be written $V(X) + [\text{bias}(X)]^2$:

$$= \text{Var}(\hat{\theta} - \theta) + (E(\hat{\theta} - \theta))^2$$
$$= \text{Var}(\hat{\theta}) - \text{Var}(\theta) + \text{bias}^2$$, but

$\theta$ = truth, a constant, so $\text{Var}(\theta) = 0$ and so

$$= \text{Var}(\hat{\theta}) + \text{bias}^2$$

- Example (WW 7-9): Large chain tune-up shops must choose 1 of 4 gauges measure gap in spark plug.

<table>
<thead>
<tr>
<th>Gauge</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>bias</td>
<td>0</td>
<td>-10</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>SE</td>
<td>10</td>
<td>0</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

* Tested, @ gauge showed errors (in .01mm):  $\begin{align*}
MSE_A &= 10^2 + (0)^2 = 100 \\
MSE_B &= (0)^2 + (-10)^2 = 100 \\
MSE_C &= 5^2 + 5^2 = 25 + 25 = 50 \\
MSE_D &= 8^2 + 2^2 = 64 + 4 = 68
\end{align*}$

* Which gauge smallest MSE (see why shop want MSE)?
So, C. A little biased, little more spread, but better combo than others (if cost fnct weighs mistakes quadratically).

- Other desirable features sampling dist of est? ∃ at least 1 more:

**Consistency**

- Unbiased: right on avg. across repeated samples. Efficient: small var. across rep. samps. **Consistent**: as sample-size goes to infinity, bias & variance go to zero.
  
  * So consistency manages say something re: 1 sample, albeit ∞, even though 2 across-repeated-samples props in (1 of its) def.

- Consistent estimator concentrates (across repeated samples) in narrower & narrower band around target as sample size → ∞.

  * “as n goes to infinity” = “in limit” = “asymptotically” ≈ “approximately”.
  
  * So, consistent ⇒ **asymptotically unbiased** (not quite v.v.).

  * **consistency** ≡ asymptotically unbiased + variance → 0.

- Yet another definition: \( \lim_{n \to \infty} \Pr(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 \)

  * \( \hat{\theta}_n \equiv \) estimate of \( \theta \) in sample size \( n \);

  * \( \varepsilon \) some arbitrary (arbitrarily small) positive real #.

  * “\( \hat{\theta} \) converges in probability to \( \theta \)” , i.e.:

    \[ \hat{\theta} \to^p \theta \iff \hat{\theta} \text{ consistent.} \]

    \[ \hat{\theta}_n \text{ consistent est } \theta \text{ if } \lim_{n \to \infty} \text{Var}(\hat{\theta}_n) = 0 \]

- So, consistency = **asymptotic** prop of an est; bias, efficiency, MSE small-sample (& asymptotic) props. Meaning?

- How draw consistency?

**Estimation: Introducing Maximum-Likelihood (ML) Estimates/ors/ion (MLE)**

- Preliminary & Background Discussion

  - So far talked about:
* Probability distributions and sample statistics of RVs,
  · including joint distributions multiple RVs.
* Sampling distributions of sample statistics/estimates of RVs,
  · including estimators that produce those estimates.
* Properties of (sampling distributions) of estimators:
  · Bias, Efficiency, MSE, Consistency.
* Maybe time start constructing estimators & obtaining estimates?
  · One Common & Powerful Method: Maximum Likelihood.

• Verbal Description ML:
  – Since not just describing (what happened), but want know how data
    observed was generated (how world works), so want something like
    probability of observing what we’ve observed (data/empirics) given
    some hypothesized data-generation process (DGP) (theory).
  – So we’ll be maximizing \( \Pr(\text{Data}|\text{HypothesizedDGP}) \), then likely
    have specific substantive hypoths to explore that will amount to / may be restated as
    statements about parameters from a probability distribution (the DGP).
  – General Q ML poses: “given my choice of a particular prob-dist as DGP, what specific values of
    its parameters make what actually observed in real world (my data) most likely to have been
    generated”?

• An MLE example:
  – Assume Binomial DGP for occurrence boys & girls in 2-child family;
  – 2 data sets of 1 family each (big samples, huh?):
    * \( R_1 = \) family of 1 boy and 1 girl,
    * \( R_2 = \) family of 2 boys.
  – Let \( p = \) probability(male birth).
  – Consider 2 hypotheses: \( H_1 \), that \( p = 1/4 \), and \( H_2 \), that \( p = 1/2 \).
  – Maximum Likelihood (of having observed sample 1, \( R_1 = (1, 1) \) or
    sample 2, \( R_2 = (2, 0) \)): 
4 key probabilities: \( \Pr(R_1|H_1), \Pr(R_2, H_1), \Pr(R_1|H_2), \Pr(R_2|H_2) \). 
* Let’s organize them into table & calculate the entries:

\[
p(n, s) = \binom{n}{s} \pi^s (1 - \pi)^{n-s} = \frac{n!}{s!(n-s)!} \pi^s (1 - \pi)^{n-s}
\]

### Data

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>( R_1 ): 1boy,1girl</th>
<th>( R_2 ): 2boys</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_1 ): ( p = \frac{1}{4} )</td>
<td>( \frac{2!}{1!(1!)} \left( \frac{1}{4} \right)^1 \left( \frac{3}{4} \right)^1 = \frac{3}{8} )</td>
<td>( \frac{2!}{2!0!} \left( \frac{1}{4} \right)^2 \left( \frac{3}{4} \right)^0 = \frac{1}{16} )</td>
</tr>
<tr>
<td>( H_2 ): ( p = \frac{1}{2} )</td>
<td>( \frac{2!}{1!(1!)} \left( \frac{1}{2} \right)^1 \left( \frac{1}{2} \right)^1 = \frac{1}{2} )</td>
<td>( \frac{2!}{2!0!} = \left( \frac{1}{2} \right)^2 \left( \frac{1}{2} \right)^0 = \frac{1}{4} )</td>
</tr>
</tbody>
</table>

* Can now say things like:
  * Given \( R_1 \), the likelihood of hypothesis \( H_1 \) is \( \frac{3}{4} \) that of \( H_2 \)
  * Given \( R_2 \), the likelihood of \( H_1 \) is \( \frac{1}{4} \) that of \( H_2 \).

* Need more data to answer this question well.

* Another: Shipment radios sampled for quality. 3 of 5 defective. What should we estimate as prop \( \pi \) of defectives in whole shipment?
  
  – ML: Consider whole range of possible \( \pi \) we might choose, and then pick one that best explains sample.
  
  – So, is \( \pi = .1 \) a sensible guess? How to know?
    * If \( \pi = .1 \), \( p(s=3) \) defectives in sample \( n=5 \) is: \( \binom{5}{3}.1^3.9^2 \approx .008. \)
    * If \( \pi = .1 \), about 8 chances in 1000 of sample we observed.]
  
  – Hey! 3 of 5 in our data, so I bet most-supported value of \( \pi \) is...

![_likelihood.png](attachment:likelihood.png)

  
  – Try all other possible \( \pi \):
  
  – Yup! Our sample has maximum probability (\( \approx 35\% \)) when \( \pi = .6. \)
• ML Formally:
  
  – ML formally, in general:

  
  \[
  \text{Likelihood of } \theta = L(\theta) = p(X_1, X_2, \ldots, X_n | \theta)
  \]
  
  and if iid = \(p(X_1|\theta)p(X_2|\theta) \ldots p(X_n|\theta)\)

  – ML Formally, Binomial Defective-Radios Example:

  \[
  Max_p L(p) = \binom{5}{3} p^3 (1-p)^{(5-3)}
  \]

  \[
  \Rightarrow \frac{\partial L(p)}{\partial p} = 3 \binom{5}{3} p^2 (1-p)^2 - 2 \binom{5}{3} p^3 (1-p) = 0
  \]

  \[
  \Rightarrow \frac{\partial L(p)}{\partial p} = 3p^2 (1-p)^2 - 2p^3 (1-p) = 0
  \]

  \[
  \Rightarrow \frac{\partial L(p)}{\partial p} = 3p^2 (1-p) - 2p^3 = 0
  \]

  \[
  \Rightarrow \frac{\partial L(p)}{\partial p} = 3p^2 - 3p^3 - 2p^3 = 0
  \]

  \[
  \Rightarrow \frac{\partial L(p)}{\partial p} = 3p^2 - 5p^3 = 0
  \]

  \[
  \Rightarrow \frac{\partial L(p)}{\partial p} = 3 - 5p = 0
  \]

  \[
  \Rightarrow \hat{p}_{ml} = \frac{3}{5}
  \]

  – Refer back to previous figure:

  * Likelihood=f(p). Max any function has slope (derivative)=0. Slope=0 @ p=\(\frac{3}{5}\) So that’s “Maximum Likelihood” estimate \(p\).

  * Variance(estimate) across repeated samples relates to how flat or peaked likelihood @ that pt, i.e. \(2^{nd}\) derivative (change in slope).
• Estimation. Beginning again, systematically:

- ML: a way to create consistent estimators. [consistency?]

  * Life’d be pretty difficult if re-prove CLTs & LLNs & o/w derive sampling dists of each estimator needed/proposed.

  * Plus, debates about what can hope infer from summaries of sample data (even well-collected) & how, but, usually:

  “the context of inference can be summarized by thinking about $y_{obs}$ as a realization of $Y \sim p^0(y), y \in \mathcal{Y}$., where $p^0(y)$ represents the unknown probability density function with respect to a suitable measure, and where $\mathcal{Y}$ is the sample space. The aim of statistical analysis is to reconstruct $p^0(y)$ on the basis of both data and suitable assumptions and, possibly, on the grounds of previous information, in order to obtain a concise description of the phenomenon being studied which will permit both interpretation and prediction.” (Pace&Salvan: 2)

- Key: what do w/ pdf when get it & how get it. 4 philosophies:

  * **Personalistic Bayesian Paradigm:**
    - pdf describes analyst’s state knowledge re: $Y$.
    - Must describe state knowledge b4 data analysis: *prior* pdf.
    - “Inference is formalization of how initial (prior) pdf changes w/ empirical evidence...according to only updating scheme that maintains internal consistency (Bayes Law)” (4).
    - So, depends crucially on analyst introspection, @ least @ start; some dislike that aspect strongly.

  * **Non-Personalistic Bayesian Paradigm:** (from Laplace)
    - Use prior dists that represent ignorance (*flat/diffuse priors*)
    - Esp. if can all agree can represent ignorance & how do so.
    - Helps reconcile b/c can all start same place en route find $p^0(y)$.

  * **Fisherian/Frequentist/Classical Paradigm:** (R.A. Fisher, 1920s)
    - Frequentist/Experimentalist inference totally based on probs emerge from experimental manipulation.
    - p-dist dice rolls don’t depend on prior beliefs; if assign plots of land some treatment, by dice, then state mind irrelevant.
· Only whether, in repeated rolls (experiments), see similar differences in crop yields according to treatment assigned.
· ⇒ **Likelihood** Notion ≈ “prob that, w/in hypothetical re-runs experiment, various competing stochastic mechanisms assign to re-observation of data produced in actual experiment” (5)
· Pr(event) must be conditional on everything known; i.e., prob as used must match as directly poss what act’ly obs’d.

* **Frequency-Decision:** (Neyman, Pearson, Wald, Lehmann 30s/40s)
  · Inference moved from Fisher’s summaries of experiment-outcomes to series procedures that would produce decisions...
  · ...represented as constrained optimization problems: e.g., Min sum (errs)$^2$ s.t. prediction linear fnct data.
  · Must identify cost fnctn (i.e. stat test yielding most relevant info for your decision—e.g. squared errs) in advance.

* This Class & PS:
  · Mostly follows mixture of last 2.
  · However, all 4 paradigms (& diff typology poss), lead to:
    · Inference requires represent “DGP” as some p-dist.
    · P-Dists: models used to guide interpretation of results (i.e. what does this “mean”) & possible prediction (i.e. what would I expect to see in different place, sample, time, etc.)
  
    – **Constructing (Classical) Estimators:** (no Bayes (explicitly) 599/699)

- **Method of Moments — Estimators by Analogy**
  
  – The $k$th moment of a p-dist is defined: $\mu_k = E(X^k)$
  – And the analogous sample moment is defined: $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$
  – Could derive ests based on such analogy. (but 599/699 gen’ly not)

- **Method of Maximum Likelihood — Estimators by Assumption**
  
  – Starting point for ML:
* Likelihood of $\theta = L(\theta) = p(X_1, X_2, \ldots, X_n|\theta)$

and if iid =

$$p(X_1|\theta)p(X_2|\theta)\ldots p(X_n|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$$

– Then, verbally:

* “Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes.” (from mathworld)

* “The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.” (Fisher 1922 [in Pace and Salvan, 14])

* The **maximum likelihood estimate** provides the value of $\theta$ for which the data provide the most support — that is makes the observed data “most probable” or “most likely”.

– To obtain ML estimate, usually easier work w/ **log-likelihood**:

  * $l(\theta) \equiv \ln(L(\theta)) = \ln(\prod_{i=1}^{n} p(X_i|\theta)) = \sum_{i=1}^{n} \ln[p(X_i|\theta)]$

  * Same b/c max f(X)=max g(f(X)) for g() monoton’ly ↑, e.g. ln()

  * Easier: derivative sum=sum derivatives easier than products.

– Maximum Likelihood Estimate (MLE): param value(s) most likely to have generated sample act’ly obs’d (given some pdf as DGP).

* WW put it this way:
  
  · Consider data $(x_1 \ldots x_n)$ rndm samp from pop w/ $p(\text{d})f p(X|\theta)$
  
  · Realizations $x_i=$draws from pdf, so $p(\text{obs any given } x): p(x_i|\theta)$.

  · iid, so joint $p(\text{d})f$ whole sample=product marginal dist @ obs:

  $$p(X_1, X_2, \ldots, X_n|\theta) = p(X_1|\theta)p(X_2|\theta)\ldots p(X_n|\theta)$$
· N.b., joint p(d)f data (=product individ obs p(d)f’s) = f(θ).
· So sample act’ly obs’d characterized by joint p(d)f.
· Indep key here to allow express that joint p(d)f as simple product of marginal p(d)f’s.

$$p(X_1, X_2, \ldots, X_n|\theta) = p(X_1|\theta)p(X_2|\theta) \ldots p(X_n|\theta)$$

· Again: MLE=value of θ most supported by data. I.e., value that maximizes joint p(d)f conditional on θ.
· Call this joint p(d)f the **likelihood function** & maximize it (or, usu., it’s ln() easier) w.r.t. θ ⇒ \(\hat{\theta}_{mle}\).

∗ MLE properties in large samples (i.e., asymptotically): **BANC**
· Asymptotically efficient (“**Best**”).
· Sampling distribution asymptotically Normal. (**AN**)
· Asymptotically unbiased (in fact, **Consistent**).

− MLE: Steps
  ∗ (1) Choose p(d)f as good candidate DGP for data,
  ∗ (2) Generate joint p(d)f (i.e., likelihood) for its parameter(s),
  ∗ (3) Max (log-)like w.r.t. θ (i.e., 1\(^{st}\)∂ = 0, check 2\(^{nd}\)∂ < 0.

− MLE: Examples
  ∗ **Binomial Example** above. Obvious: \(\hat{\pi}_{mle} = .6\) maxes L(θ).
  ∗ **WW Example 18-2, p. 577: Poisson Distribution**

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$
Parameter(s) \( \theta \)? Here: \( \lambda \), instantaneous rate even occurrence.

Poisson mostly used describe RVs capturing discrete, rarish events in large pops. [Examples?] Looks lot like Binomial, in fact: \( B(n=\text{large}, \ p=\text{small}) \approx \text{Poisson} \).

Here, tiny sample: \( X = (15, 8, 13) \). So \( n = 3 \).

Calc how log-like varies for \( \theta \) goes 5 to 25 by 5. What’s MLE?

\[
L(x|\lambda) = \prod_{x \in \{15, 8, 13\}} \left( \frac{e^{-\lambda} \lambda^x}{x!} \right)
\]

Want log of L. How write?

General form:

\[
l(\theta) = \sum_{i=1}^{n} \log[p(X_i|\theta)]
\]

Begin by taking log of Poisson (part inside summation):

\[
\ln(p(X|\theta)) = \ln \left( \frac{e^{-\theta} \theta^x}{x!} \right)
= \ln(e^{-\theta}) + \ln(\theta^x) - \ln(x!)
= -\theta + x \ln \theta - \ln x!
\]

So log-likelihood function:

\[
\log L(\theta) = \mathcal{L} = \sum \ln(p(X|\theta)) = \sum (-\theta + x \ln \theta - \ln x!)
= -n\theta + \sum x_i \ln \theta - \sum \ln(x!)
\]

With \( n = 3 \& x = (15, 8, 13) \): \( \sum x_i = 15 + 8 + 13 = 36 \).

Only care about stuff that varies w/ \( \theta \) so can dump \( \sum \ln(x!) \):

\[
\log L(\theta) = -3\theta + 36 \ln \theta
\]

\( \log L(\theta) = 42.94, 52.89, 52.49, 47.85, 40.88 \) as \( \theta \) goes 5 \( \rightarrow \) 25 by 5. Plot:
So, MLE seems somewhere around 12.

**WW 18-9: Exponential**

- Problem: Waiting time \( X \) until next telephone call at switchboard has exponential distribution:

\[
p(X) = \theta e^{-\theta X}, \quad 0 \leq X
\]

- To estimate parameter \( \theta \), sample of 5 waiting times (in min.) obs'd: \( X \in \{1.2, 7.5, 1.8, 3.7, .8\} \).
- a. How does loglik vary as \( \theta \) varies from .1, .2, \ldots, .5?
- Well, likelihood function:

\[
L(\theta) = \prod \theta e^{-\theta X}
\]

- Log-likelihood function:

\[
\text{logL}(\theta) = \ln \left( \prod \theta e^{-\theta X} \right) = \sum \ln (\theta e^{-\theta X_i})
\]

\[
= \ln(\theta e^{-\theta X_1}) + \ldots + \ln(\theta e^{-\theta X_n})
\]

\[
= \ln(\theta) + \ln(e^{-\theta X_1}) + \ldots + \ln(\theta) + \ln(e^{-\theta X_n})
\]

\[
= n \cdot \ln(\theta) - \theta \sum X_i
\]
· Plug in what know \((n = 5; \sum X_i = 15)\) ⇒ 
\[
\log L(\theta) = 5 \ln(\theta) - \theta \cdot 15
\]
· How does this vary as \(\theta\) goes from .1 to .5?
\[
\begin{align*}
\log L(.1) &= 5 \ln(.1) - 15(.1) = 5(-2.30) - 1.5 = -13.00 \\
\log L(.2) &= 5 \ln(.2) - 15(.2) = 5(-1.61) - 3 = -11.05 \\
\log L(.3) &= 5 \ln(.3) - 15(.3) = 5(-1.20) - 4.5 = -10.52 \\
\log L(.4) &= 5 \ln(.4) - 15(.4) = 5(-0.92) - 6 = -10.58 \\
\log L(.5) &= 5 \ln(.5) - 15(.5) = 5(-0.69) - 7.5 = -10.97
\end{align*}
\]
· b. So MLE of \(\theta\) \(\approx\)? (n.b., log-likes negative)
· How find max of curve exactly? Slope=0. How find that?
\[
\frac{\partial \log L}{\partial \theta} = \frac{5}{\theta} - 15 = 0 \Rightarrow \frac{5}{\theta} = 15 \Rightarrow \hat{\theta}_{mle} = 1/3
\]
· Is it a max or a min? (I.e., 2\textsuperscript{nd} deriv pos or neg?)
\[
\frac{\partial^2 \log L}{\partial \theta^2} = -5/\theta^2, \hat{\theta}_{mle} = 1/3 \Rightarrow -45
\]
- But is MLE a good estimator?
* Asymptotically unbiased, indeed: consistent, asymptotically efficient, converges in dist to normal, & std errs of MLE’s easy to find (direct from \(2^{nd}\) deriv., i.e. Hessian, of) like fnctn).

* Why std errs related to \(2^{nd}\) deriv. L? (draw wide & narrow L). So, info in given MLE rests on tightness curve around (est’d) max.

* Drawbacks:
  · Asymptotic properties only.
  · Need p(d)f up-front, & can’t be uncertain about it as in bayesian.
  · Also: can be hard find max in multivar dists, not always pretty, smooth surface: local max &/or mistake bumps for max.

– One more, **An Example Application**:

* Deaths/day from *The Times* of London, 1910-1912.

  ```r
  deathcount<-c(162,267,271,185,111,61,27,8,3,1)
  deathsperday<-rep(0:9,deathcount)
  table(deathsperday)
  deathsperday
  0 1 2 3 4 5 6 7 8 9
  162 267 271 185 111 61 27 8 3 1
  * Seems Poisson, but maybe a mixture of 2 Poissons:

  \[
  \log L = \ln \left( \prod_{i=1}^{n} \left[ p e^{-\theta_1} \theta_1^{X_i} \left( 1 - p \right) e^{-\theta_2} \theta_2^{X_i} \right] \right)
  \]

  * So, try estimate \(p, \theta_1, \theta_2\).

  * First, write log-likelihood function:

    ```r
    pmix<-function(p,x){
    e<-p[1]*dpois(x,p[2])+(1-p[1])*dpois(x,p[3])
    if(any(e<=0)) Inf else -sum(log(e))
    }
    * Then find its maximum:

    ```r
    themle<-optim(c(p=.5,t1=1,t2=1),pmix,x=deathsperday,hessian=TRUE)
    round(themle$par,4)
    p \quad t1 \quad t2
    0.3598 1.2565 2.6626
    > sqrt(diag(solve(themle$hessian)))
    ```
$\text{hessian}$

\[
\begin{array}{ccc}
p & t1 & t2 \\
p & 906.0709 & -270.0137 & -341.14697 \\
t1 & -270.0137 & 113.33850 & 61.77306 \\
t2 & -341.1470 & 61.77306 & 192.90984
\end{array}
\]

\[
> \text{sqrt(diag(solve(themle$hessian$)))}
\]

\[
\begin{array}{ccc}
p & t1 & t2 \\
0.1942547 & 0.3494568 & 0.2495489
\end{array}
\]

**• Other Methods to Construct Estimates**

- **Method of Moments (Using the Analogy Principle)**
  
  * Been using (simplest cases of) MoM from start. (Well, “Trust me/book” method: I/book say/s “sample mean is reasonable estimator of the population mean,” & you nod or scribble. “Reasonable” b/c look, defined, calc’d, etc. same (parallel) in sample rel to pop ≡ Analogy Principle).
  
  * Been asking “what p-dist generated these data?”, which ≈
  
  * “what are the characteristics of some p-dist that provides best model for these data?”
  
  * A moment of a dist ≡ a characteristic or property of the dist.
  
  * If we knew all moments of a p-dist, would know everything.

  * So, MoM:
    
    * $k^{th}$ moment of a population p-dist ≡ $\mu_k = E(X^k)$
    
    * And analogous sample moment ≡ $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$

  * Example:
    
    * $1^{st}$ & $2^{nd}$ moments of Normal dist are? ...
    
    * How defined/calc’d? ...
    
    * So, $1^{st}$ & $2^{nd}$ sample moment? ...

  * Not used much (directly) in PS; some in econ, finance esp.
**Example:** SRS of $Y_1, Y_2, \ldots, Y_n$ selected from pop w/ $Y_i \sim U(0, \theta)$, w/ $\theta$ unknown. Use MoM to estimate $\theta$:

- What does $U(0, \theta)$ look like? (DRAW)
- Suppose $p(Y) = 1$ over $(0,1)$. (DRAW). What’s $\mu$? $E(\bar{Y})$?
- Suppose $p(Y) = 1/2$ over $(0,2)$. (DRAW). $\mu = ?$ $E(\bar{Y}) = ?$
- Suppose $p(Y) = 1/100$ over $(0,100)$. (DRAW). $\mu = ?$ $E(\bar{Y}) = ?$
- So, formula for mean ($\equiv E(Y) \equiv \mu$) for $U(0,\theta)$:

$$\mu_1 = \mu = \frac{\theta}{2}$$

- So, first sample moment (analogous):

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}$$

- So, use sample mean to est relevant parameter in pop:
- First, equate population & sample moments:

$$\mu_1 = \frac{\theta}{2} = \bar{Y}$$

- Then, solve for $\theta$ ($\hat{\theta}$) as f(sample moment):

$$\hat{\theta} = 2\bar{Y}$$

* Idea behind MoM: use possible correspondence b/w sample est & pop values to squeeze info from data.

* MoM: robust consistency, good @ multi-eq. models, but effic. probs, esp. rel to ML’s usu. high effic. if $n$ big & p-dist right.

**Least-Squares/Regression/Projection:**

* Alt. strategy starts w/ thinking prediction or regression/projection.
- Obs as draws from some dist. Minimize mistakes using some kind of known info (ind vars) to predict.

$$\min_{\hat{\mu}} \sum_{i=1}^{n} (Y_i - \hat{\mu})^2$$
\[ \Rightarrow 2 \sum_{i=1}^{n} (Y_i - \hat{\mu}) = 0 \]
\[ \Rightarrow \sum_{i=1}^{n} Y_i = n\hat{\mu} \Rightarrow \hat{\mu} = \bar{Y} \]

• (Try LS Reg w/ thry/model for E(Y) = \hat{\mu}, say E(Y)=f(X)=bX?)
• Or: Project info from some info (dep vars) onto some other info (indep vars), how scale X info to get close possible to Y? which turn out to be same thing & same as above.
• (DRAW: vectors X & Y in 2x2, using X vector as x-axis, Y diagonal. How scale X w/ coeff. b to get projection from Y?)

– Notice: MoM, ML, LS all gave same estimator for mean. Not always so, but \( \bar{X} \) for \( \mu_X \) = pretty basic, so odd if not.

**Confidence Intervals**

• Transition from point estimation to interval estimation.
  – Just intro’d 3 gen methods estimate some point. Other ways, but MLE, MME, LS among most powerful & common.
  – Just prior, discussed properties/criteria. List/define: ...
  – So, knowing point est, which=mean of sampling dist of estimator, is not only thing need/want to know to talk probability.
  – Need/want sense also variability of our point ests b/c want say whether data support that est strongly or weakly.
  – One way to talk about this variability is to consider an interval around the point est. And then how likely interval contains the pop value (Bayesian), or something like that (classical).

• Example
  – Let’s begin w/ simple proportions again: Say, want know what prop students in AA live at home. How learn this?
* Sample 400 students. Find 317/400, 79% of sample, live @ home.

* Q: how far could this sample number be from true pop number?
  · Many possible errs sample design, but, say managed perfectly rep sample: is # from 1 samp likely = pop #?
  · No. Some chance deviation. What, then, is avg deviation from true pop number might expect? I.e., need std err of estimate.

* Know SE of sample mean/prop is $\frac{\sigma}{\sqrt{n}}$. Know $\sigma^2 = \pi(1 - \pi)$.

* But not observe $\pi$; just have $p$.
  · In stats, don’t know $\Rightarrow$ replace w/ est., $\pi$ w/ $p$ to est $\sigma$, so:
    · $\hat{\sigma}^2 = .79(1 - .79) = .79(.21) = .1659$. And $\hat{\sigma} = \sqrt{.1659} = .407$. So, $\text{SE} = \frac{.407}{\sqrt{400}} = \frac{.407}{20} = .02$.
    · So, expect chance errors of size 2% on avg. Likely size chance error $\approx$ likely deviation from true pop prop=2%.

* But then: what prob. off by 1SE? How likely, say, pop prop 83%/81% or 75%/77% (i.e. 2 or 1 SE away from sample prop)?
  · Estimate of mean: $\bar{X}$. Sampling distribution?
  · Cntrd @ $\mu$ (unbiased), & variance of this normal dist is SE($\bar{X}$).
  · How know smpl mean/prop normally distrib’d?

* Now what can say about relation b/w $\bar{X}$ & $\mu$, in prob. terms?
  · $\Pr(\bar{X} > \mu + SE \text{ or } \bar{X} < \mu - SE) =$? Well, how much area w/in $\pm 1$ SD of mean of normal dist?
  · $\Pr(\bar{X} > \mu + 1.96SE \text{ or } \bar{X} > \mu - 1.96SE) =$? Well, how much w/in $\pm 1.96$ SD?
    · (Remember: SEs are estimated SDs for estimators/estimates.)
  
* So, prob. $\bar{X}$ more than 1.96SE from $\mu \approx 5\%$ (n.b., Bayesian). Formally: $\Pr(\mu - 1.96SE < \bar{X} < \mu + 1.96SE) = 95\%$

* Can also rewrite to center interval on $\bar{X}$, which we know, rather than $\mu$, which we don’t:
  \[
  \Pr(\bar{X} - 1.96SE < \mu < \bar{X} + 1.96SE) = 95\%
  \]
* From this, can write gen formula 95% confidence interval (c.i.):

\[ \mu = \bar{X} \pm z_{0.025} \text{SE} = \bar{X} \pm z_{0.025} \frac{\hat{\sigma}}{\sqrt{n}} \]

- What parts c.i. vary across repeated samps? Which constant?
- What happens to c.i. as n, SE, or conf level ↑?

* So, from example survey students, can say:
  - \( \pi = .79 \pm z_{(1-\alpha)/2} \times .02 \) (\( \alpha \) =confidence level)
  - So, for \( \alpha = .95 \), \( z_{0.025} \approx 1.96 \), so 95% c.i. = .79 ± .04

* No rule that must pick 95% confidence:
  - Want 99%? Then \( z_{0.005} = 2.58 \) & c.i. wider by \( \approx 2.6/2 = 30\% \).
  - Satisfied w/ 50% conf? \( z_{0.25} = .675 \) & c.i. 33.375% narrower.

* Tend to say c.i. from “about a” to “about b” or ±“about c” b/c:
  - Est’d mean & std errs from data rather than knowing \( \mu \) & \( \sigma \).
  - Used normal approx., but usu. only know norm asym’ly.
− Confidence-Interval Interpretation:
  * Classical:
    · Probability≡relative frequency across repeated randomness producing operations, e.g. sampling
    · @ sample of size 400⇒diff # for prop living @ home & SE.
    · If sampling rightly, then std.err. of est. tells likely amount off.
    · So, c.i. center & bounds vary, not \( \mu \) & \( \sigma \). So interp:
      · “For about 95% of all samples, the interval \( p \pm 1.96 \cdot SE \) covers the pop mean, & for 5% of samples, will not.”
    · Kinda backward from what would like to say. E.g., 25 samples from same population produced these 95% c.i.’s:
* Bayesian:
  - Probability\[posterior,\] post-Bayes updating by evidence, state of knowledge/degree certainty re: parameters.
  - So, can say what want directly: 95% c.i., \( \hat{\theta} \pm 1.96SE = .75....83 \) gives range that contains \( \mu \) w/ 95% probability.

- **Example** (WW Problem 8-3): Rndm srvy 225 of 2700 instits higher learning in US conducted ‘76 by Carnegie Commission. Sample mean enrollment 3700 studs/instit, sample std.dev. 6000 studs. Construct 95% c.i. for total studs in all 2700 instits.
  * So, have (sample) mean=3700, & sd=6000, & know that:

\[
\mu = \bar{X} \pm z_{0.05} \frac{\sigma}{\sqrt{n}}
\]

\[
2700\mu = 2700(3700 \pm 1.96 \frac{6000}{\sqrt{225}})
\]

\[
= 2700(3700 \pm 1.96 \frac{6000}{15})
\]

\[
= 2700(3700 \pm 784)
\]

\[
= 9,990,000 \pm 2,116,800
\]

• **Confidence Intervals for MLE**
  - Back to our Poisson & Death Notices example. If have an MLE est, how get std err (& so get c.i., etc.)?

\[
I(\theta) \equiv E \left[ \frac{\partial \log f(X|\theta)}{\partial \theta} \right]^2 = -E \left[ \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \right]
\]

* (=Huh?) Known as *information matrix*. (Oh! Well that explains... nothing. Huh?)

* Can be shown that:

\[
\text{Var}(\hat{\theta}_{mle}) = \frac{1}{nI(\hat{\theta}_{mle})} = -\frac{1}{E(l''(\hat{\theta}_{mle}))}, \text{ as } n \to \infty
\]
So, if any idea what meant by above greek or information matrix, it’d help w/ variances & SE’s & such.

How about this from Jake: “asymptotic variance of the estimator is the inverse of the information matrix, or the inverse of the matrix of second derivatives of the likelihood function (known as the Hessian). Rice says, ‘the asymptotic variance is [the formula above] which may be interpreted as the average radius of curvature of \( l(\theta) \) at \( \hat{\theta}_{mle} \). When this radius of curvature is small, the estimate is relatively well resolved and the asymptotic variance is small.’ (265)

My version: 2nd derivatives indicate curvature, remember?

[DRAW a likelihood] Notice how greater curvature indicates tighter range around MLE w/ high prob-density? So, greater curvature, larger (more negative) 2nd derivative likelihood (at max), means more certainty of that partic. est’d value, less variance around it in repeated samples. So formula must have variance inversely proportional to 2nd derivative magnitude, but 2nd derivative negative, so negative that, so -(1/I(MLE)).

Matrix of 2nd derivatives of log-likelihood \( \equiv \text{Hessian (Matrix)}: \)

\[
I(\theta) \equiv \mathbb{E} \left[ \left( \frac{\partial \log f(X|\theta)}{\partial \theta} \right)^2 \right] = -\mathbb{E} \left[ \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \right]
\]

So, we want s.e.’s of our estimates, which were these:

```r
#Find min of negative log-likelihood for mixture dist:
themixmle<-optim(c(p=.5,t1=1,t2=1),pmixll,x=deathsperday,hessian=TRUE)

#Find minimum of negative log-likelihood for single parameter dist:
thepoisnle<-optim(c(t1=1),justpoisll,x=deathsperday,hessian=TRUE,method="BFGS")

#What are the estimates? themixmle$par
p  t1  t2
0.3598354 1.2564646 2.6626142

thepoisnle$par
t1
2.156931
```
* We could derive s.e.’s analytically & find c.i. using CLT:

\[ L(x) = \prod \left( \frac{e^{-\theta} \theta^x}{x!} \right) \]

\[ LL(x) = \sum \ln \left( \frac{e^{-\theta} \theta^x}{x!} \right) \]

\[ LL(x) = \sum \left[ -\theta + x \ln \theta - \ln(x!) \right] \]

\[ LL(x) = -n\theta + \ln \theta \sum x - n \ln(x!) \]

\[ \frac{\partial LL}{\partial \theta} = -n + \sum \frac{x}{\theta} \Rightarrow \hat{\theta}_{mle} = \bar{x} \]

\[ \frac{\partial^2 LL}{\partial \theta^2} = -\sum \frac{x}{\theta^2} \]

\[ \Rightarrow I(\theta) = \mathbb{E} \left( -\sum \frac{x}{\theta^2} \right) = -\frac{n\theta}{\theta^2} = -\frac{n}{\theta} \]

* So, est var here is \(-\left( -\frac{n}{\theta} \right)^{-1} = \frac{\hat{\theta}}{n} = \frac{\bar{X}}{n} \), giving c.i.: \( \bar{X} \pm z_{\alpha/2} \sqrt{\frac{\bar{X}}{n}} \).

* Or, just ask comp prog for hessian & sqrt diagonal of inverse:

```plaintext
## What are the standard errors of the estimates?
sqrt(diag(solve(themixmle$hessian)))
p t1 t2
0.1942547 0.3494568 0.2495489
```

```plaintext
sqrt(diag(solve(thepoismle$hessian$))
t1
0.04436214
```

* So: \( 2.16 \pm 1.96 \cdot .044 = 2.16 \pm .086 \) for single-param model.

– So, how interpret this interval?

* If kept reading, & London Times Death Notices random samples, 95% of time our interval would contain pop value.

* Or, from Rice (202), “a 95% confidence interval for \( \mu \) is a random interval that contains \( \mu \) with probability .95; if we were to take
many random samples and form a confidence interval from each one, about 95% of these intervals would contain $\mu$.”

* I.e., c.i.’s vary, not $\mu$; remember this graphic:

- **Ex.: NYTimes Headline:** “Zagats List...Small Brooklyn Spot”
  - Zagat’s 7th best restaurant in NYC = The Grocery (Brooklyn).
  - Headline b/c no dinky mom & pop in Brooklyn top 7 in NYC!
  - Grocery got 28/30 for food $\Rightarrow$ top 7. How?
    - Rating system: 0 to 3 pts times 10 $\Rightarrow$ 0 to 30 scale
    - Sample = folks who rate rest’s online; min. 100 raters:
      - 100-650 rated Grocery, 2600± Daniel (in ’02). Total: 29K±.
  - So, Zagats’ might want report some est of certainty of ratings...
    - Might also want rndm, rep sample (how? rep of whom?)
    - What c.i? 95% is fine, but maybe 75% OK here, not so big deal if est some restaurant rating off.
  - OK, say, 100 rated Grocery on avg 28; so c.i.:
    $$\mu = 28 \pm z_{.125} \frac{\sigma}{\sqrt{n}}$$
    - Where $\sigma$, is s.d. of pop of Grocery eaters’ ratings. Don’t know it, can’t even est it w/o data; let’s pretend sd=10.
    - Critical value of std norm for .125 area to rt=1.15. So, c.i.:
      $$\mu = 28 \pm 1.15 \frac{10}{\sqrt{100}} = 28 \pm 1.15$$
    - I.e., if took 100 samples, 75 of 100 would contain true pop value w/in interval like 26.85 to 29.15.
Daniel:

- Say Daniel rated 29 by sample 2500 & its pop sd is also 10 (pretended heterogeneity of NYC fancy-dinner eaters). C.I.?

\[
\mu = 29 \pm 1.15 \sqrt{\frac{10}{2500}} = 29 \pm 1.15(1/5) = 29 \pm .23
\]

- So, if 100 samples, 75 of 100 get true pop value b/w 28.77 & 29.23. Act’ly, can’t quite say it that way (classically), right?

Example: One more!

- SRS, 37% of people ≤ 30 support Bush. What need to calc c.i.?
  - n=931. So, what do now? \( \mu = \bar{X} \pm z_{.025} \frac{\sigma}{\sqrt{n}} \)
  - What’s \( \sigma \)? Dunno⇒Estimate!
    - \( \hat{\sigma}^2 = \hat{p}(1 - \hat{p}) = .37(.63) = .23 \Rightarrow \sqrt{\frac{.23}{931}} = .015 \Rightarrow SE = .015 \)
  - What’s a std err of an estimate again? OK. What now? Oh, right, how big a c.i. you want? I.e., \( 1 - \alpha \). Say 75% (so \( \alpha = .25 \)).
  - So, want 75% of area around mean, i.e. want 37.5 on @ side 50, so 50-37.5=12.5 in each tail.
  - So want crit. val. \( x \) on normal CDF s.t. \( \Phi(x) = 1 - .125 = .875. \)
  - R gives this with command: \texttt{qnorm(.872)}, which is about 1.15.
  - For 95% dist’d symmetrically about mean: \texttt{qnorm(.975)}=1.96.
  - So, 37\% ± (1.15)(.015) ≈ 37\% ± .017; i.e., 75% c.i.=(.353,.387).

- How interpret?
  - 95\% prob. interval would contain true value if repeated sample & procedure. Not (classically) correct to say “95\% probability true value in c.i.”; either is or isn’t. “In long run” or “on avg” we’d expect such intervals to contain true value.
  - People sometimes say true value lies in interval w/ 95\% confidence, but this just being sneaky.

- Bayesian could say “I’m 95\% confident true value lies b/w \( \textbf{a} \) & \( \textbf{b} \),” but calc c.i. bit differently (using priors).

- For frequentist/classical, (bounds of) intervals vary, not true pop value. Say it backward: “95\% of intervals (constructed thusly from repeated samples) will contain true parameter.”
Hypothesis Testing & Student’s t

• Overview:
  – May have occurred to you: don’t usu. know true pop value $\sigma$. We est w/ sample value, $s$. That adds uncertainty, which widens intervals.
  – History: Lord Guiness (the younger) hires William Gosset to help scientifically revolutionize beer-making in 1908.
    * Gosset, a chem & math student, aka Student, spent many after-hrs taking mean & s.d. (& pints?) of small (but repeated!) samples & plotting (drinking?), trying to discern a sampling dist.
    * Gosset notices that, with s.d.’s estimated, empirical dists much fatter tailed than normals, esp. in smaller samples.
    * $\Rightarrow$ Student’s $t$ distribution: fatter tails than normal, and so you have to go farther out in tails to get 95% of area under curve.
  * Fatness of tails depends on sample size. Big samples, tails quite thin, eventually as thin as Normal’s. Small samples, tails fat rel. normal b/c less info, more uncertainty.

  $$\mu = \bar{X} \pm t_{0.025} \frac{s}{\sqrt{n}}$$

  $$s^2 \equiv \frac{1}{n-1} \sum (X - \bar{X})^2$$

  – $t$-dist changes w/ its degrees of freedom ($n - 1$ here), fatter or skinnier tails as $n \uparrow \downarrow$.
  – In practice, $t$ w/ $n < 100$ perhaps discernable from normal, but $n \geq 100$ impossible to tell at many decimals.
  – Gosset derived $t$-dist assuming a Normal pop, but, in practice, $t$-dist works well for any mound-shaped dist.

• Formal Definition $t$-Distribution
  – “If $Z \sim N(0,1)$ & $U \sim \chi^2_n$, and $Z$ & $U$ independent, then $\frac{Z}{\sqrt{U/n}}$ is $t$-distributed with $n$ degrees of freedom.” (Rice, p. 178).
– The pdf of a $t$-distribution is:

$$f(x) = \left(1 + \frac{x^2}{n}\right)^{-\frac{1-n}{2}} \frac{\Gamma\left(\frac{1+n}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}$$

– The pdf of a $\chi^2$ distribution is:

$$f(x) = \frac{x^{-\frac{1+n}{2}}}{2^{\frac{n}{2}}e^{\frac{x}{2}}\Gamma\left(\frac{n}{2}\right)}$$

– The $\chi^2$ distribution arises from squaring Std.Norm. RVs.

– The $\Gamma()$ function is:

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du, \ x > 0$$

– For positive integers, $n$, $\Gamma(n) = (n - 1)!$. So, the $\Gamma()$ function generalizes a factorial to work for non-integer & complex numbers.

– E.g., Bin. coeff. ${n \choose m} \equiv \frac{n!}{m!(n-m)!} = \frac{\Gamma(n+1)}{\Gamma(m+1)\Gamma(n-m+1)}$, for integer n,m.

– By $n = 30$ the $t$ looks a lot like the normal. The long dashed lines is for $n=5$, the dots for $n=30$, the solid line is for a standard normal.
• **Hypothesis Testing: Differences of Means**
  
  – Say want to know whether 2 means differ. E.g., mean political participation different for women & men?
  
  – The formula (already know it) is (for indep. samples (meaning?)):

  \[
  (\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm z_{(\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
  \]

  – And, if the 2 samples have same variance, then formula becomes:

  \[
  (\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm z_{(\alpha/2)} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
  \]

  – Now, usu. don’t know \(\sigma^2\), so use \(s\) to denote est. \(\sigma\) & it becomes:

  \[
  (\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{(\alpha/2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
  \]

  – Where \(p\) in \(s_p\) is “pooled” since pooling pops (w/ same \(\sigma^2\)). E.g.?

  \[
  s_p^2 = \frac{\sum (X_1 - \bar{X})^2 + \sum (X_2 - \bar{X})^2}{(n_1 - 1) + (n_2 - 1)}
  \]

  – The degrees of freedom here are \((n_1 - 1) + (n_2 - 1)\).

• **Difference of Means: Examples**

  – Kerry-Bush Gender Gap

---

### The New York Times/CBS NEWS Poll

**A Kerry Rebound**

As campaign 2004 heads toward its finale, Senator John Kerry has recovered the strength he had before the conventions with many groups of voters. A nationwide CBS News poll of 909 registered voters conducted Sept. 6-8, after the conclusion of the Republican convention, showed erosion in Mr. Kerry’s support among the groups shown below. In the aftermath of the debates, Mr. Kerry has improved his standing with those voters, according to the latest nationwide New York Times/CBS News Poll, conducted Oct. 14-17 with 901 registered voters.

<table>
<thead>
<tr>
<th>Kerry</th>
<th>Bush</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>47%</td>
</tr>
<tr>
<td>Oct</td>
<td>50</td>
</tr>
<tr>
<td>Independents</td>
<td>50</td>
</tr>
<tr>
<td>Moderate</td>
<td>47</td>
</tr>
<tr>
<td>Under 30</td>
<td>54</td>
</tr>
<tr>
<td>College Graduates</td>
<td>41</td>
</tr>
<tr>
<td>YEAR</td>
<td>47</td>
</tr>
</tbody>
</table>
Say the 2 samples indep. Claim women going Kerry. Assess:

\[(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}\]

* Our case 1 is Sept and 2 is Oct. We know \(\bar{X}_1 = .43\) and \(\bar{X}_2 = .50\).
* What is \(\hat{\sigma}_1^2\)? \((p(1-p)=.43(1-.43)=.2451.\) For other sample is \(.50(1-.50)=.25.\)
* The n’s are 909 and 931.

\[(\mu_1 - \mu_2) = (.43 - .50) \pm t_{\alpha/2} \sqrt{\frac{.2451}{909} + \frac{.25}{931}}\]

* Say, want 90% c.i., so \(\alpha = .10\). \(df = (n_1 - 1) + (n_2 - 1) = 908 + 930 = 1838.\)
* So, critical value of \(t\) w/ 1838 df, at \(\alpha/2 = .10/2 = .05\) & \(t_{.03, df=1838} = 1.65.\) So:

\[(\mu_1 - \mu_2) = (.43 - .50) \pm 1.65 \sqrt{\frac{.2451}{909} + \frac{.25}{931}}\]

\[= -.07 \pm 1.65(.0231984) = -.07 \pm .038\]

* So 90% c.i. is -.108 to -.032. (How interpret?)

– **WW 8-11, p. 268** Large US university in ‘69, male & female profs’ (indep’ly sampled) salaries (in $1K current):

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

* a. Calc 95% c.i. for mean salary difference by gender. 1\(^{st}\) formula:

\[(\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t_{(\alpha)/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\]
\[ \bar{X}_w = (9 + 12 + 8 + 10 + 16)/5 = 11 \]
\[ \bar{X}_m = (16 + 19 + 12 + 11 + 22)/5 = 16 \]

\[ s_p^2 = \frac{\sum(X_w - \bar{X})^2 + \sum(X_m - \bar{X})^2}{(n_w - 1) + (n_m - 1)} \]
\[ = \frac{[(9 - 11)^2 + (12 - 11)^2 + (8 - 11)^2 + (10 - 11)^2 + (16 - 11)^2]}{4 + 4} \]
\[ + \frac{[(16 - 16)^2 + (19 - 16)^2 + (12 - 16)^2 + (11 - 16)^2 + (22 - 16)^2]}{4 + 4} \]
\[ = \frac{[4 + 1 + 9 + 1 + 25] + [0 + 9 + 16 + 25 + 36]}{8} \]
\[ = 126/8 = 15.75 \]

\[ \sqrt{s_p^2} = \sqrt{15.75} = 3.97 \]

\[ \sqrt{\frac{1}{n_w} + \frac{1}{n_m}} = \sqrt{\frac{1}{5} + \frac{1}{5}} = \sqrt{\frac{2}{5}} \approx .63 \]

\[ df = (n_w - 1) + (n_m - 1) = 4 + 4 = 8 \]

Critical value \( t \) for \( df=8 \) & prob in one tail=.025 (i.e. \( t_{.025,8} \))=2.31.

So, putting it all together we have:

\[ (\mu_w - \mu_m) = (11 - 16) \pm 2.31 \cdot 3.97 \cdot .63 = -5 \pm 5.78 \]

\[ \Rightarrow 95\% c.i. = (-10.78, .78). \]

* b. How well does this show U’s discrimination against women?
* How could you enhance the power of this empirical test?
* General problems w/ this data/study/research-design?

**Paired Samples**

- Indep. samples not always poss/avail. E.g., same people/states/ctrys @ diff times⇒samples *dep*, so indep-assuming formula = lying to computer (Horrible thing to do to poor computer/reader!) Each person/state/ctry/etc (i) has difference score, \( D_i = X_{i1} - X_{i2} \)
Conf. int. for avg diff w/ such matched/paired samples is:

\[ \Delta = \bar{D} \pm t(\alpha)/2 \frac{s_D}{\sqrt{n}} \text{ with } df = (n - 1) \]

**WW 8-15:** How much does interesting environ affect brain dev’p?

* Rosenweig et al.: 10 litters of purebred rats. Select (randomly) 1 rat from @ litter for treatment and 1 for control group. Same except treated rats live in cage w/ interesting playthings, & control rats lived in bare isolation. After month, cortexes of rats brains weighed. Results for the 10 pairs of litter mates in centigrams:

<table>
<thead>
<tr>
<th>Litter</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>68</td>
<td>65</td>
<td>66</td>
<td>66</td>
<td>67</td>
<td>66</td>
<td>66</td>
<td>64</td>
<td>69</td>
<td>63</td>
</tr>
<tr>
<td>Control</td>
<td>65</td>
<td>62</td>
<td>64</td>
<td>65</td>
<td>65</td>
<td>64</td>
<td>59</td>
<td>63</td>
<td>65</td>
<td>58</td>
</tr>
<tr>
<td>Difference</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

* Construct an appropriate 95% confidence interval.

\[ \Delta = \bar{D} \pm t(\alpha)/2 \frac{s_D}{\sqrt{n}} \]

\[ \bar{D} = \frac{3 + 3 + 2 + 1 + 2 + 2 + 7 + 1 + 4 + 5}{10} = \frac{30}{10} = 3 \]

\[ s_D^2 = \frac{(3 - 3)^2 + (3 - 3)^2 + (2 - 3)^2 + (1 - 3)^2 + (2 - 3)^2}{10 - 1} \]

\[ + \frac{(2 - 3)^2 + (7 - 3)^2 + (1 - 3)^2 + (4 - 3)^2 + (5 - 3)^2}{10 - 1} \]

\[ = \frac{0 + 0 + 1 + 4 + 1 + 1 + 16 + 4 + 1 + 4}{9} = \frac{32}{9} = 3.55 \]

\[ s_D = \sqrt{3.55} \approx 1.88 \]

\[ t_{.025,df} = t_{.025,9} = 2.26 \]

\[ \Delta = 3 \pm 2.26 \frac{1.88}{\sqrt{10}} = 3 \pm 2.26 \frac{1.88}{3.16} = 3 \pm 2.26 \cdot .59 = 3 \pm 1.33 \]

* So, 95% c.i. of difference=(1.67,4.33). Conclusion?
• Differences of Proportions

– Recall: For large n, 95% c.i. around proportion is:

\[ \pi = p \pm 1.96 \sqrt{\frac{p(1-p)}{n}} \]

* What’s \( p(1-p) \) doing in numerator here?
* Why do we need n large here?

– So, might want compare 2 proportions (in large, indep samples):

\[ (\pi_1 - \pi_2) = (p_1 - p_2) \pm 1.96 \sqrt{\frac{p_1(1-p_1)}{n_2} + \frac{p_2(1-p_2)}{n_2}} \]

– WW 8-21, p. 276 In ‘54, large exper. to test efficacy new polio vaccine. Of 740K US 2\textsuperscript{nd}-graders selected, 400K volunteer. 1/2

<table>
<thead>
<tr>
<th>Group</th>
<th># Kids</th>
<th># Polio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaccine</td>
<td>200,000</td>
<td>57</td>
</tr>
<tr>
<td>Placebo</td>
<td>200,000</td>
<td>142</td>
</tr>
<tr>
<td>Refused</td>
<td>340,000</td>
<td>157</td>
</tr>
</tbody>
</table>

(random) vaccine. Rest: placebo. Results:

* a. For each grp, calc polio rate (in cases/100K)

Vaccine \( 57/2=28.5 \) 28.5/10E5=.000285=57/200K
Placebo \( 142/2=71.0 \) 71/10E5=.00071=71/200K
Refused \( 157/3.4=46.18 \) 46/10E5=.00046=157/340K

* b. Est reduction in polio rate that vaccine produced, w/ 95% c.i.

\[ (\pi_1 - \pi_2) = (p_1 - p_2) \pm 1.96 \sqrt{\frac{p_1(1-p_1)}{n_2} + \frac{p_2(1-p_2)}{n_2}} \]

\[ (\pi_1 - \pi_2) = (.00071 - .000285) \pm 1.96 \sqrt{\frac{.0007(.9993)}{200,000} + \frac{.000285(.9997)}{200,000}} \]

\[ = .000415 \pm 1.96 \sqrt{3.5474795 \times 10^{-9} + 1.42459 \times 10^{-9}} \]

\[ = .000425 \pm 1.96 \times .0000705129 = .000425 \pm .000138205 \]
* In cases per 100K: 42.5 ± 13.8, i.e. 95% c.i. of (28.7, 56.3).

**Bootstrapping (Simulating)**

- Sometimes, SE of est unknown or difficult compute analytically...
  - b/c CLT or asym thry not reassuring (sample too small, e.g.)
  - b/c substantively meaningful est not a mean or sum iid vars, but else, so sampling dist hard to derive even asym’ly.
- Another option is to simulate many random draws from hypothetical universes, using your 1 & only 1 sample repeatedly (w/ replace)
  - Efron’s idea: assess variability of \( \hat{\theta} \) around \( \theta \) using variability of \( \hat{\theta}^* \), set of re-estimates of \( \theta \), around \( \hat{\theta} \).
  - Example: ratio of means (analytic sol’n for SE difficult/unavail).
    - Pop 10 US cities in 1920 (u) & 1930 (x). Use ratio of means to est total pop US in 1930 from 1920. Then, SE/c.i.? (Other ways to forecast, some easier SE calc, but this for sake e.g.)

```r
> t(city)
   1  2  3  4  5  6  7  8  9 10
   u 138 93 61 179 48 37 29 23 30 2
   x 143 104 69 260 75 63 50 48 111 50
```

- Estimate \( E(\theta) = E(X)/E(U) \). No obvious pdf for \( \theta \), so use \( \bar{X}/\bar{U} = T \) (MME) to est \( \Rightarrow 1.52 \).
  - C.I.? Well, CLT only says num&denom @ asym’ly norm, but ratio of means: not certain what sampling dist, could try derive, but easier to bootstrap/simulate &, besides, sample size worrisome for asym thry:

```r
bootrat<-vector(length=1000) for(i in 1:1000){
    thesamp<-sample(1:10,replace=TRUE)
    bootrat[i]<-mean(city[thesamp,"x"])/mean(city[thesamp,"u"]))
}quantile(bootrat,p=c(.025,.975)) ##Basic bootstrap CI
2.5%    97.5%
1.249582 2.032397
mean(bootrat) ##Note: some bias v. actual sample mean
1.539735
```
Just a check, does sampling dist look normal?

What is this qqplot thingy?

Example (has analytic, but smpl small & skewed): Failure times b/w AC repairs on Boeing 720. Want mean or 1/mean=fail rate.

```r
> t(aircondit)
     1  2  3  4  5  6  7  8  9 10 11 12
   hours  3  5  7  18  43  85  91  98 100 130 230 487
mean(aircondit$hours) [1] 108.0833

# Large sample theory CI
airci<-c(108.0833-qt(.975,df=11)*sqrt(var(aircondit$hours))/sqrt(12),
        108.0833+qt(.975,df=11)*sqrt(var(aircondit$hours))/sqrt(12))

Can calc asym thry CI, but reason to distrust in small skewed sample. So, bootstrap:

```r
bootair<-vector(length=1000) for(i in 1:1000){
    thesamp<-sample(1:12,replace=TRUE)
    bootair[i]<-mean(aircondit$hours[thesamp])}

quantile(bootair,p=c(.025,.975))
45.24375 191.33750
```
This time, bootstrap c.i. tighter than asym-thry ci; either poss.

Lot more to learn re: bootstrap/sim; really helpful, but some probs:

- Hard to bootstrap time-series (or similar settings/models)
- Debate about best way to create sim’d c.i.’s. These’re Efron’s original proposal, called “basic bootstrap c.i.”, but
  - some bias of bootstrapped parameter ests rel to smpl ests.
  - simulation itself adds some random noise; other c.i. bootstrap procedures try to account such.

Another use simulation relies on analytic results for SE’s.

- Mentioned b4 that coefficients=effects only in lin-add.
- Sometimes effects complicated functions of coefficients: \( f(\mathbf{b}) \).
- Two options to calc s.e. \( f(\mathbf{b}) \)
  - Delta Method: linearize (1\(^{st}\)-Order Taylor-Series Approx) @ \( \mathbf{b}_{opt} \):
    \[
    [\nabla_{\mathbf{b}} f(\mathbf{b})]' V(\mathbf{b}) [\nabla_{\mathbf{b}} f(\mathbf{b})]
    \]
  - Simulation Method: Draw randomly from \( \mathbf{b} \sim (\mathbf{b}, \hat{\Sigma}_b) \). Calc \( f(\mathbf{b}) \) each draw. Then calc \( V[f(\mathbf{b})] \) across draws.

Sum: very powerful, flexible, & practical approach.

- Always want report certainty of ests.
- May be hard calc s.e. analytically, or may have reasons doubt analytic or asymptotic conditions apply well.
- Good to know can pretty much always simulate such things.
Return to Visualize/Understand Confidence Intervals

- Imagine standing behind target, seeing arrow-heads coming through, but don’t know when person hit bullseye.
- Know archer unbiased, know variance of shots, & by what dist.
- Then, gray dots below rep places arrows came through (estimates).
- With your knowledge of archer’s skill (unbiased, s.d., & distrib), you know that 95% of time a circle of radius 1.96(s.d.) contains bullseye.
- So, if you drew circles of that radius around each arrow-head, 95% of time your circle should contain bullseye.

```r
> x <- rnorm(10000, mean = 0.5, sd = 1)
> y <- rnorm(10000, mean = 0.5, sd = 1)
> plot(c(0, 1), c(0, 1), type = "n", xlab = "X", ylab = "Y")
> points(0.5, 0.5, pch = 19)
> for (i in 1:20) {
+   xsamp <- sample(x, size = 100)
+   ysamp <- sample(y, size = 100)
+   xbar <- mean(xsamp)
+   ybar <- mean(ysamp)
+   cix <- 2 * sqrt(sd(xsamp)/100)
+   ciy <- 2 * sqrt(sd(ysamp)/100)
+   points(xbar, ybar, pch = 19, col = gray(0.6))
+   symbols(xbar, ybar, circles = c(max(cix, ciy)), add = TRUE) + }
```
Hypothesis Testing (cont.)

• So Far:
  – How estimate parameters of dist’s & how describe sample data.
  – Talked about uncertainty of such ests/descripts/summaries.
  – Such ests & uncertainties enable inference from sample ests to pop.
  – Conf int’s & hypoth test culmination of those efforts.
    * Hypoth Test = related Q to Conf Int’s:
      · Not \( p(a < \mu < b) \)... (c.i., Bayesian)
      · But now: “Is \( \mu \geq 0 \)” or “Is \( \mu \geq a \).
      · Care about 0 or \( a \) b/c some hunch (thry?) how exper or obs
        study unfold: call it a hypothesis (i.e., theories⇒hypotheses).
      · Yet another way to think of c.i.’s is set of acceptable hypothe-
        ses @ that c.i.% level, i.e. set of hypotheses supported by data
        for particular parameter.
  – Say interest in diff b/w 2 params (means, rates, coeff’s...)
    * Have hypothesis about whether diff observe (say, from substract-
      ing ests) should be zero, or positive, or whatever.
    * Want evaluate that hypoth with data (evidence). Key factor, as
      would guess, is how much info data offer for each est.
      · If lot, then small diffs might be detectable @ high conf levels.
      · If little, then even large diffs might not be reliably detectable.
    * Say hyp no diff, but find (est) lrg smpl-diff, & tons info, then:
      · “With a 5% chance of error, we can reject hypoth of no diff.”
      · diff is “statistically significant” @ 5% (signif., alpha, p-) level.
      · Hypoth test statements. Equivalent c.i. statement?
  – SUM:
    * We use c.i.’s to talk about how close sample ests to pop param,
    * Hypoth tests to ask if data provide strong evidence to support
      claims re: pop params.
Null Hypothesis

- Some good ways to conceive Null Hypotheses:
  * “null hyp is ritualized exercise of devil’s advocacy” (Abelson, 9)
  * “null hyp is interesting when discrediting it implies...” what you wanted/argued/theorized. (A, 10)

- **Substantive versus Statistical Significance**
  * Statistical significance regards distinguishing b/w population claims (based on sample data) using some prespecified comfort level of error. I.e., can you detect a difference, how reliably?
  * As/more important: **substantive significance**. Even if can detect diff, does diff matter **substantively**? Big **substantively**?
  * Example of stat signif but not subst signif:
    - In 100K-person CPS, 50% of pop say 10, 51% say 10.05 on 20-pt scale=highly stat’ly sig (95% c.i.’s on order of $1/\sqrt{100,000}$).
    - Really, who cares if such a tiny difference exists?
  * Lesson: You estimated parameters for substantive & theoretical reasons. Explore those in interp results!

Example; WW Prob9-2: Real-estate broker employs 2 appraisers to est value of houses for sale, & wonders whether less-experienced (A) as good as (B). As test, each appraises same 5 houses.

<table>
<thead>
<tr>
<th>House</th>
<th>Value$_a$</th>
<th>Value$_b$</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94</td>
<td>81</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>55</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>39</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>116</td>
<td>106</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>136</td>
<td>121</td>
<td>15</td>
</tr>
</tbody>
</table>

- a. Calc 95% c.i. to est diff avg appraisals A, B. Differ at $\alpha = 5\%$?

\[
\bar{X}_a = \frac{94 + 60 + \ldots}{5} = 89, \quad \bar{X}_b = \frac{81 + 55 + \ldots}{5} = 79
\]
\[ \Delta = (\bar{X}_a - \bar{X}_b) \pm t_{.025,n-1} \cdot \left( s_p \sqrt{\frac{1}{n_a} + \frac{1}{n_b}} \right) \]

\[ = 10 \pm t_{.025,4} \cdot \left( s_p \sqrt{\frac{2}{5}} \right) \]

\[ s_p^2 = \frac{\sum (X_a - \bar{X}_a)^2 + \sum (X_b - \bar{X}_b)^2}{(n_a - 1) + (n_b - 1)} \]
\[ = \frac{(94 - 89)^2 + (60 - 89)^2 + (39 - 89)^2 + (116 - 89)^2}{8} = \frac{11,586}{8} = 1448.25 \]
\[ s_p = \sqrt{s_p^2} = \sqrt{1448.25} \approx 38 \]

\[ \Rightarrow \Delta = 10 \pm 2.31(38)\sqrt{\frac{2}{5}} = 10 \pm 2.31 \cdot 38 \cdot .63 = 10 \pm 55.30 \]

\[ \Rightarrow 95\% c.i. = (-45.3, 65.3). \]

- What about avg diff appraisals as paired c.i.? (Houses same, so \( D = X_a - X_b \) more approp).

\[ \Delta = \bar{D} \pm t_{.025} \frac{s_d}{\sqrt{n}}; \text{df} = n - 1 \Rightarrow \bar{D} = 50/5 = 10 \]
\[ s_D^2 = \frac{(13 - 10)^2 + (5 - 10)^2 + (7 - 10)^2 + (10 - 10)^2 + (15 - 10)^2}{4} \]
\[ = \frac{3^2 + 5^2 + 3^2 + 0^2 + 5^2}{4} = \frac{9 + 25 + 9 + 0 + 25}{4} = \frac{68}{4} = 17 \]
\[ \Delta = 10 \pm t_{.025,4} \frac{\sqrt{17}}{\sqrt{5}} = 10 \pm 2.78 \frac{4.12}{2.24} = 10 \pm 5.12 \]
\[ \Rightarrow 95\% c.i. = (4.88, 15.12) \]
- b., c., d. Is diff avg or avg (paired) diff stat’ly distinct from 0?
- e. What was Q motivating this study? Null hypoths? One version of latter rejected, but relation of either to Q?

• Hypothesis Testing: Piece by Piece

- p-values (significance- or alpha-levels)
  * p-value: Jargon. Refers to probability. DEF: \( p \) observe sample value so large as or larger than observed if \( H_0 \) true.
    - 1-sided \( H_0 \): only true-pop vals > (or only <) smpl val weighed.
    - 2-sided \( H_0 \): true-pop val > or < smpl vals weighed.
  * KNOW \( \theta \geq 0 \); test \( H_0 : \theta = 0 \Rightarrow \) only weigh \( p(\hat{\theta} > \hat{\theta}_0 | \theta = 0) \).
  * How generate these probabilities? Very sim to c.i.’s:
    - Info needed same: estimator, estimate, & s.d. & sampling dist
    - Same sampling dists (z & t) under same conditions [i.e. (?)].

- t-value (more generally, test statistic)
  * For H-tests of form \( H_0 : a \geq b \), e.g. \( \mu \geq \mu_0 \):
    
    \[
    Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu_0}{\text{exact SD}}
    \]
  * So result for t-statistic (or t-value) is:
    
    \[
    t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{\bar{X} - \mu_0}{\text{estimated SE}}
    \]
  * where \( \mu_0 \) is null-hypothesized \( \mu \), and \( \bar{X} \) the est’d \( \mu \).
  * So, test is gen’ly \( t = \frac{\text{est-hyp}}{s.e.} \); if \( H_0 : \mu = 0 \), then just \( t = \frac{\text{est}}{SE} \).

- Verbally: if have est (e.g., a mean), & est. s.d. of that est (i.e., s.e.), then calc how many s.e.’s is est from hypoth \( \Rightarrow \) t-value, from which can calc p-value giving probability of observing an est that high or higher if null hyp about population were true.
  * If p=50% (when would this happen?), null quite likely sensible
  * If p=.01%, then very improbable null correct, etc.
Classical H-Test: Reject/Fail-to-Reject (*accept*)

* Classically:
  - Set (α-)level of test, i.e. prob falsely reject null, *a priori*.
  - If est > implied crit val $t$ or $z$, then reject @ that signif lvl.
  - If est $\leq$ crit val, "fail to reject", *NOT accept*.
  - Possible verdicts: guilty/not-guilty, not guilty/innocent.

* H-test considers $p(\text{est}|H_0)$ not $p(H_0|\text{est})$. If $p(\text{data}|H_0)$ small, then, if $H_0$ were true, data unlikely⇒More backward speak: "reject null," but, info regards likelihood obs data/est.

**An Example:** Racial bias in jury selection, 1960-80 in US South.

* Evidence in 1 case: about 50% eligible cits AfrAm, but in 80-person jury panel (potential jurors), only 4 AfrAm.
* If selection truly random (i.e., equal prob), what prob(4 in 80)?
  - What probability dist good model for this?
  - So, given this model fair selection, what prob(4 or fewer AfrAm)?

\[
\Pr(X \leq 4) = \sum_{x=1}^{4} \binom{80}{x} .5^x (1-.5)^{80-x} = 1.3788935 \times 10^{-18}
\]

One way think of c.i. v. H-test: (illustrates that logical inverses)

* c.i.: given 1 sample, what rndm sys gen’d it’s stats;
* H-test: posit rndm sys, & ask whether data likely gen’d by it.

**Note:** C.I.⇒ set acceptable Null Hyps @ p-/α-lvl.

Type-I and Type-II (& Type-III) Errors

* 2 Kinds of Errors in H-Test: (e.g., smoke detector; $H_0$=No fire)
  - Type 1: Reject $H_0$ when true (alarm w/o fire)
  - Type 2: Fail reject $H_0$ when false (fire w/o alarm)

* b/c est’ing, not know truth, can’t know if/which err commit⇒P(errs)?
  - Already know $p(\text{reject } H_0|\text{true})$. What was it?
  - $P(\text{fail reject|false}) \equiv \beta$. [$\text{power} = 1 - \beta = p(\text{reject|false})$].
* How to think of these: The Trial Analogy
  · $H_0$: defendant innocent; $H_A$ defendant guilty.
  · Set $\alpha = .05$ says comfortable rejecting $H_0$ (i.e. convicting) when $H_0$ act’ly true (innocent) about 5% of time.
  · Avoiding Type I Errors: In US sys, much more concern innocents not convicted than guilty let free (“innocent until proven guilty beyond reasonable doubt” instruction) ⇒ sets $\alpha$ very small. How set $\alpha = 0$ in fire alarm example?
  · Type II Errors: What happens if minimize $p$(Type I) in Fire-Alarm by removing batteries?
    · In general, tradeoff b/w $p$(Type I) & $p$(Type II)—no way lower one w/o raise other if using all data efficiently & accurately.
* Type III Error most common: Not know/forget diff Type I & II.

- Formal Steps H-Test (Jerzy Neyman & Egon Pearson ca. 1930)
  1. State Null & Alternative Hypotheses ($H_0$ & ($H_A$)
  2. Choose Significance Level, $\alpha$, probability at which willing accept false reject true null (Type I).
  3. Given fixed $\alpha$, design test & test stat that max $1 - \beta$, prob reject false $H_0$, i.e. power.
  4. Calculate p-value, probability that, if null were true, est a test stat so extreme or more.

• Hypothesis Testing: EXAMPLES
  - Poll says Bush had 52% likely voters, want know if can rest easy.
    * Good null hypothesis? (I.e., thing wants reject.) Alternative?
    \[
    t = \frac{.52 - .5}{\sqrt{.5(1-.5)/699}} = \frac{.02}{.019} = 1.0583
    \]
    * How t-stat converts p-level depends on $H_0$ & $H_A$.
      · If 1-sided ($H_0 : p \leq .5$, $H_A : p > .5$), so: $p(t > 1.0583) = p_1$
If 2-sided ($H_0: p = .5, H_A: p \neq .5$): $p(|t| > 1.0583) = 2 \times p_1$

* In R: `pt(1.0583, df=699, lower.tail=FALSE)=0.145`

* For 2-sided, `pt(1.0583, df=699, lower.tail=TRUE)=0.29`.

– **WW 9-23** In Gallup poll of 1.5K in '75, 45% ‘yes’ to “any area around here (w/in 1 mi) where afraid to walk alone at night?” In poll of 1.5K in '72, only 42% said ‘yes’. Different?

* a. Construct 2-sided 95% c.i. for change in prop afraid:

$$(\pi_1 - \pi_2) = (p_1 - p_2) \pm 1.96 \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$(\pi_1 - \pi_2) = (.45 - .42) \pm 1.96 \sqrt{\frac{.45(.55)}{1500} + \frac{.42(.58)}{1500}} = .03 \pm 1.96\sqrt{.000165 + .0001624} = .03 \pm .035$$

⇒ 95% c.i. = (−.005, .065)

* b. Calc 2-sided p-value for null hypoth of no change:

$H_0: \pi_1 - \pi_2 = 0 \Rightarrow t = \frac{\text{difference}}{\text{estimated SE}}$

\[ s = \sqrt{.000165 + .0001624} = .018 \Rightarrow t = .03/.018 = 1.67 \]

• (DRAW t.) 1-sided p-value for 1.67 (w/ df=3000-2)=.0475. That 1 side, other side=.0475 too, so 2-sided p=2*.0475=.095.

* Is $H_0$ in CI? (yes). Does p-value for $H_0$ fall below 5%? (yes 1-sided; no 2-sided). Note: ConfidenceLevel=1-TestLevel.

– **WW 9-9** Purchaser of several new ships waterproof gloves hopes as good as old ships’ 10% defective rate. Takes rndm sample 100 pairs & counts prop $p$ defective. Level of test determined by relevant factors (costs bad shipments, alternative suppliers, etc.) as $\alpha = .09$.

* a. State null & alternative hypotheses:

$$H_0: p_n \geq p_o, H_A: p_n < 10\%$$
* b. What critical value? I.e., how large must \( p \) be to reject null hypoth (& shipment)?

\[
t = \frac{\text{estimate} - \text{null}}{SE}
\]

\[
t = \frac{p_n - p_o}{\sqrt{\frac{p(1-p)}{N}}} = \frac{p_n - .10}{\sqrt{\frac{.1(.9)}{100}}} = p_n - .10
\]

· Now, what is \( t \)? We want \( t_{.09,99} = 1.34 \), so:

\[
1.34 = \frac{p_n - .10}{.03} \Rightarrow p_n = .14
\]

* c. If \( p \) next 6 ships=\((12,25,8,16,24,21)\%\). Which rejected?

- **WW 9-11** In manufacturing machine bolts, quality control engineer finds sample \( n=100 \) nec. to detect change of .5mm in mean length bolts. Wants precision to detect .1mm, w/ same \( \alpha \) & \( \beta \). How much larger must sample be? (Hint: rephrase problem in c.i. terms. What smpl-size increase nec. to make c.i. 5 times smaller?)

\[
\mu = \bar{X} \pm z_{.025} \frac{\sigma}{\sqrt{n}}
\]

* So, SE relates to precision. Want \( 5\times \) precision of current SE, which is \( \frac{\sigma}{\sqrt{100}} \). Set up to solve for \( n \):

\[
\frac{1}{5} \left( \frac{\sigma}{\sqrt{100}} \right) = \frac{\sigma}{\sqrt{n}} \Rightarrow \frac{\sigma}{\sqrt{100}} = \frac{5\sigma}{\sqrt{n}}
\]

\[
\sqrt{n} = 5\cdot 10 = 50 \Rightarrow n = 50^2 = 2500
\]

- **WW 9-21** Electronics firm buys printed circuits in lots of 50K & in past has found proportion substandard=\( p = 2\% \). To guard against worsening quality, decide to sample each lot: if more than 25 in 1000 defective, they will reject & return lot.

* a. Calc \( \alpha \)-level of this test.

· \( \alpha \equiv \text{chance reject null hypothesis when act'ly true.} \)
Null here that $\pi$ remains 2%. If null true, expect how many defective in 1K circuits? [Binomial: $E(y) = np = 1000(.02) = 20$].

Use normal approx (b/c smpl lrg: CLT), so like norm w/ mean .02 & reject if .025 or above. Still need know variance.

Luckily, know that formula too: $SE = \sqrt{\frac{p(1-p)}{n}}$. Population $\sigma = \sqrt{.02(.98)/1000} = .0044$.

So, have $\pi \sim N(.02, .0044^2)$. Could ask computer for $p$ to right of .025 on this dist ($Pr(\bar{X} > .025) = .13$) ($\Rightarrow \alpha = .13$).

Or, could stdz it & find correspond area on std norm dist: $(.025-.020)/.0044=1.13636 \Rightarrow Pr(z > 1.13636) = .13$.

*b. Calculate $\beta$ for several alternate hypotheses.

$\beta \equiv$ prob fail reject $H_0$ when $H_A$ true

So, pretend each alternate hypothesis true, i.e. treat as null.

i. $\pi = 2.5%$

$$SE = \sqrt{.025(1-.025)/1000} = .0049$$

$$z = \frac{.025 - .025}{.0049} = 0, Pr(\bar{X} < 25) = Pr(z < 0) = .5$$

ii. $\pi = 3%$

$$SE = \sqrt{.03(1-.03)/1000} = .0054$$

$$z = \frac{.025 - .03}{.0054} = -.93, Pr(\bar{X} < 25) = Pr(z < -.93) = .177$$

iii. $\pi = 3.5%$

$$SE = \sqrt{.035(1-.035)/1000} = .0058$$

$$z = \frac{.025 - .035}{.0058} = -1.72414, Pr(\bar{X} < 25) = Pr(z < -1.72) = .043$$

iv. $\pi = 4%$

$$SE = \sqrt{.04(1-.04)/1000} = .0062$$

$$z = \frac{.025 - .04}{.0062} = -2.42, Pr(\bar{X} < 25) = Pr(z < -2.42) = .008$$
c. Sketch the graph of $\beta$ (I added .02):

\[
\begin{array}{c|c|c|c|c}
\text{beta} & 0.2 & 0.4 & 0.6 & 0.8 \\
\hline
\text{mua} & 0 & 0.02 & 0.03 & 0.04 & 0.05
\end{array}
\]

- So, as proportion defective ↑, prob mistakenly failing reject that only .025 defective ↓.

d. Instead of graphing $\beta$ (probability mistakenly accepting lot), could graph $1 - \beta$ (probability correctly rejecting lot). Called the power curve.
· As proportion defective ↑, probability correctly rejecting null when false ↑. I.e., for given sample-size, tests like this get more powerful the larger the effect they’re trying to detect.

* e. Suppose 50 shoddy lots shipped, each having 3% defective. About how many detected & rejected?
· Know p(mistakenly failing reject shoddy lot|\(\pi = .03\))=.177.
· So probability detecting shoddy lot is 1-.177=.823.
· Assuming p(detect shoddy) in 1 lot indep of others, can mult to get # bad lots caught of 50: .823*50=41.15.

**Elaboration: Relation b/w C.I.’s & H-Tests**

– (1 − \(\alpha\))% C.I.’s give set of hypothesized values that fail reject at \(\alpha\) level (i.e., set of hypotheses supported (not rejected) by data).
– Say have \(X_1, \ldots, X_n\) SRS from \(N(\mu, \sigma^2)\). Let’s test:

\[
H_0 : \mu = \mu_0; H_A : \mu \neq \mu_0
\]

– Let test have \(\alpha\) that rejects for:

\[
|\bar{X} - \mu_0| > x_0, \ldots \text{where...Pr}(|\bar{X} - \mu_0| > x_0) = \alpha
\]

\[
x_0 = \text{SE}_{\bar{X}} \cdot z_{\alpha/2}
\]

– The test thus accepts the null when:

\[
|\bar{X} - \mu_0| < x_0, \ldots \text{or...}|\bar{X} - \mu_0| < \text{SE}_{\bar{X}} \cdot z_{\alpha/2}
\]

– Or, removing the absolute value sign:

\[
-\text{SE}_{\bar{X}} \cdot z_{\alpha/2} < \bar{X} - \mu_0 < \text{SE}_{\bar{X}} \cdot z_{\alpha/2}
\]

– Or, subtracting from both sides, flipping < to < and flipping back:

\[
\bar{X} - \text{SE}_{\bar{X}} \cdot z_{\alpha/2} < \mu_0 < \bar{X} + \text{SE}_{\bar{X}} \cdot z_{\alpha/2}
\]

– And so, a 100(1 − \(\alpha\))% confidence interval for \(\mu_0\) is:

\[
[\bar{X} - \text{SE}_{\bar{X}} \cdot z_{\alpha/2}, \bar{X} + \text{SE}_{\bar{X}} \cdot z_{\alpha/2}]
\]
Thus, \( \mu_0 \) lies outside (inside) \( 1 - \alpha \% \) c.i. \( \iff \) \( \alpha \)-level hypothesis test rejects (fails reject). “In other words, the confidence interval consists of precisely all of those values of \( \mu_0 \) for which the null hypothesis \( H_0 : \mu = \mu_0 \) is accepted” (Rice, 307).

**Some (Classical) Perspectives on Hypothesis Testing**

- WW present H-testing as kind of hybrid b/w two approaches:
  
  * (1) Fisher: null hypoth; reject null if “significance” very small.
    - “Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis” (Fisher (1935: 16) in The Empire of Chance (Gigerenzer et al: 96).
    - Fisher’s *Design of Experiments* same yr as Popper’s *Logic of Scientific Discovery*: both essentially falsificationism.
    - Lots of published small p’s to discredit previously-thought-sound null.
    - A little fuzzy on how choose text-stat & sig level & what’s “lot” & “sound” & what do then...
  
  * (2) Neyman & Pearson: offer 2 H’s \( (H_0 & H_A) \) w/ 1 must-true
    - Then can specify \( R \), “rejection region”, obs in which \( \Rightarrow \) reject \( H_1 \) in favor of \( H_2 \) outside which reject \( H_2 \) in favor of \( H_1 \).
    - Here, “The purpose of observing is to distribute, on basis of obs, praise & blame over these 2 hyp’s, viz. to reject 1 & accept other” (Gigerenzer et al., p. 99)
    - Given this, N&P showed could choose test that both minimizes \( \alpha \) & maximizes \( 1 - \beta \) for a pair of simple hypotheses.
    - Problem: “If obs discordant w/ both \( H_1 & H_2 \), N-P will accept hyp w/ which data least discordant” (Gigerenzer, 104).
    - I.e., Type IV error: “giving right answer to wrong question, asked by wrong model” (Gigerenzer 104, from Kimball 1957).
    - And no way to assign probability to this kind of error and so to make the kinds of trade offs between type I & II errs.
* And this just among frequentists; Bayesians have own debate.
* Point: stats texts often present h-testing as if like CLT, a clear, proven, undeniable mathematical result, but some disagreement still on philosophical foundations and practical details.
* In PS/SocSci, people use smattering of both (plus some Bayesian), so need know nulls & significance as well as alternates & power.
* Upshot: avoid being mechanical w/ p-values, t-values, etc. Think of them as summary stats for posterior distributions of estimates.

**Another Example**

- VP-sales for lrg corp claims sales people avg no more than 15 contacts/wk. Wants ↑.
- As check on his claim, n=36 rndm salespersons’ contacts recorded for 1 rndmly-selected wk.
- Sample mean: 17 contacts; Sample varaince: 9.
- Does evidence contradict VP’s claim? Use $\alpha = .05$.
  * So $H_0 : \mu = 15$ and $H_A : \mu > 15$. We know that:
    $$z = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$
  * At $\alpha = .05$ we also know that $z = 1.645$.
  * So, we’d reject null if $z > 1.645$.
    $$z = \frac{17 - 15}{3/\sqrt{36}} = 2/.5 = 4$$
  * So, we’d reject null @ p-value: 0.0000316712.
- Same VP wants to be able to detect diff of 1 call in mean # calls/wk. I.e., wants test $H_0 : \mu = 15$ against $H_A : \mu = 16$. Find $\beta$.
  * First, find crit val $\bar{Y}$ boundary b/w reject & not reject. Know
    $$z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} > 1.645$$
\[ \bar{Y} - \mu_0 > 1.645 \frac{\sigma}{\sqrt{n}} \]
\[ \bar{Y} > \mu_0 + 1.645 \frac{\sigma}{\sqrt{n}} \]

* So substitute \( \mu_0 = 15 \) and \( n = 36 \):
\[ \bar{Y} > 15 + 1.645 \frac{3}{\sqrt{36}} \]

\[ \beta = \Pr \left( \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{15.8225 - 16}{3/\sqrt{36}} \right) = \Pr(z \leq -0.355) = 0.3613 \]

* Meaning? Probability fairly large. Sample size \( n=36 \) frequently fail detect 1-unit diff from hyp’d mean. Just calc’d that 36% chance fail reject null 15 when real mean 16.

* Note: To design such test, must choose sample size.

- In fact, simple formula for choosing sample size:

  * Say want test \( H_0 : \mu = \mu_0 \) against \( H_A : \mu = \mu_A \).
* For specific $\alpha$ & $\beta$, then just 2 unknowns: $k=$ crit val & $n$:

$$\alpha = \Pr(\bar{Y} > k \text{ when } \mu = \mu_0)$$

$$= \Pr\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} > \frac{k - \mu_0}{\sigma/\sqrt{n}} \text{ when } \mu = \mu_0\right)$$

$$= \Pr(z > z_\alpha)$$

$$\beta = \Pr(\bar{Y} \leq k \text{ when } \mu = \mu_A)$$

$$= \Pr\left(\frac{\bar{Y} - \mu_A}{\sigma/\sqrt{n}} \leq \frac{k - \mu_A}{\sigma/\sqrt{n}} \text{ when } \mu = \mu_A\right)$$

$$= \Pr(z \leq -z_\beta)$$

* So from eq. for $\alpha$ know that

$$\frac{k - \mu_0}{\sigma/\sqrt{n}} = z_\alpha$$

* and from eq. for $\beta$ know that

$$\frac{k - \mu_A}{\sigma/\sqrt{n}} = -z_\beta$$

* Solve both for $k$, then set equal:

$$k = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

$$k = \mu_A - z_\beta \frac{\sigma}{\sqrt{n}}$$

$$\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_A - z_\beta \frac{\sigma}{\sqrt{n}}$$

$$z_\alpha \frac{\sigma}{\sqrt{n}} + z_\beta \frac{\sigma}{\sqrt{n}} = \mu_A - \mu_0$$

$$(z_\alpha + z_\beta) \frac{\sigma}{\sqrt{n}} = \mu_A - \mu_0$$

$$(z_\alpha + z_\beta)\sigma = \sqrt{n}(\mu_A - \mu_0)$$

$$\frac{(z_\alpha + z_\beta)\sigma}{(\mu_A - \mu_0)} = \sqrt{n}$$
\[
\left( \frac{(z_\alpha + z_\beta)\sigma}{(\mu_A - \mu_0)} \right)^2 = n
\]

* So, for hypoth like this: \( H_0 : \mu = \mu_0 \) against \( H_A : \mu > \mu_0 \), sample size as just calc’d.

– Say same VP wants test: \( H_0 : \mu = 15 \) against \( H_A : \mu = 16 \), w/ \( \alpha = \beta = .05 \). Find \( n \) that ensures this level accuracy:

* If \( \alpha = \beta = .05 \), then \( z_\alpha = z_\beta = z_{.05} = 1.645 \), so

\[
 n = \frac{(1.645 + 1.645)^2(3)^2}{(16 - 15)^2} = 97.4
\]

* So, he should have 98 observations.

- Tests for Contingency Tables \( \chi^2 \)

  – A Contingency Table:

* Survey to eval effect new flu vaccine administered in small town.
* Vaccine free in 2-shot sequence over 2wks to those wishing it.
* Some people received both shots, some only 1st, others neither.
* Survey next spring 1K local inhabitants ⇒ following table.
<table>
<thead>
<tr>
<th>No Vaccine</th>
<th>1 shot</th>
<th>2 shots</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>No Flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>313</td>
<td>109</td>
<td>578</td>
<td>1000</td>
</tr>
</tbody>
</table>

- Do data present sufficient evidence to indicate dependence b/w vaccine & flu? i.e., is this diff than expect w/o flu shots? (Would expect/hope flu shots & flu outcome not independent.)

  * Might start by checking if prop sick varies w/ treatment (i.e., if \( \Pr(X \mid Y) = \Pr(X) \)): \( \frac{24}{313} = 7.66\% \), \( \frac{9}{109} = 8.26\% \), \( \frac{13}{565} = 2.3\% \). So, seems some relationship.

  * But are these #'s sig’ly diff? Might next ask how many obs expect in each cell given indep. How?

    - Well, marginal percentages: \( \frac{46}{1000} = .046 \), and \( .954 \). So, if overall flu rate same w/in each treatment, expect \( 313 \cdot .046 = 14.398 \), etc. Another way: multiply row & column marginal frequencies and divide by total # obs: \( \frac{\text{row total} \times \text{col total}}{\text{total # obs}} \)

    * So, if no relation/indep: \( \Pr(X, Y) = \Pr(X)\Pr(Y) \). Check:

      | No Vaccine | 1 shot | 2 shots |
      |------------|--------|---------|
      | Flu        | 46(313)/1000=14.398 | 5.014  | 26.588 |
      | No Flu     | 298.602 | 103.986 | 551.412 |

    * Clearly our Q is “how diff are these 2 tables?”

- Test for “how diff are 2 tables?” called \( \chi^2 \) test:

  \[
  \chi^2 = \sum_{\text{rows}} \sum_{\text{cols}} \frac{(O - E)^2}{E}
  \]

  * where \( O \) stands for “observed” and \( E \) for “expected”.

  * Divide by \( E \) to stdz so lrgst contribs not always from lrgst cells.

  * Obv’ly, lrgr diffs b/w obs’d & expect freqs \( \Rightarrow \) lrgr \( \chi^2 \).

- Here: \( \chi^2 = \frac{(24-14.4)^2}{14.4} + \frac{(289-298.6)^2}{298.6} + \cdots = 17.35. \)

  * So, is this \( \chi^2 \) stat big or little? Need know sampling dist.
* Well, called *chi-squared* b/c samp dist=χ². Now can talk prob!

− χ² distribution:
  * Can have a location parameter like normal’s mean, but usu. just “degrees of freedom” like Student’s t.
  * Degrees of freedom related to obs & # things est’d or tied down, so here dep on # cols & rows.
  * If know all but one of entries in a col or row, know everything (right?). So, df=(c-1)(r-1)=(3-1)(2-1)=2.

− So, crit val of χ² for df=2 & say, α = .05?
  * Tables in back of the book, or ask a computer.
  * For α = .05, reject \( H_0 \) of independence when \( \chi^2_2 > 5.991 \)
    
    \[
    qchisq(.05, 2, \text{lower.tail}=\text{FALSE})=5.991
    \]
    
    \[
    pchisq(17.35, \text{df}=2, \text{lower.tail}=\text{FALSE})= 0.0001708030
    \]
  * Here: reject null of indep @ p<.0002.

− Note: In this case, \( H_0 : \pi_{ij} = \pi_i \pi_j \).

− Used lot in SocSci. In fact, “analysis of contingency tables” an entire branch of sociological methodology (an expert in UM Soc dept: Yu Xie). Methodology also often called “log-linear models” (for reasons that may become clear toward end this or next semester).

---

**ANOVA and F-test**

− Now can compare counts in table cells & compare means, but what if want compare >2 means, e.g., \( H_0 : \mu_1 = \mu_2 = \mu_3 \)?

− If can rep diffs b/w these means somehow, which can, can do this:
  * To start, need find mean of these means (the *Grand Mean*).
  * Then calc variation of these little means around grand mean

\[
\text{GrandMean} \equiv \bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3}
\]

\[
\text{Variance(GrandMean)} = s^2_{\bar{X}} = \frac{1}{a-1} \sum (\bar{X}_i - \bar{X})^2
\]
* a here is number of means, and so a-1=df.

– So that’s one aspect of data’s variance: variance across the 3 means, i.e. the variance *between* the 3 means.

– Another type of variance in the data too: variance *within* the 3 samples themselves.

– Can develop pooled estimate of total variability within the 3 samples by adding deviations from the 3 sample means and dividing by each sample’s degrees of freedom:

\[
sp^2 = \frac{\sum (X_{1i} - \bar{X}_1)^2 + \sum (X_{2i} - \bar{X}_2)^2 + \ldots}{a(n - 1)}
\]

– So, now Q: “Is \(s^2_{\bar{X}}\) large relative to \(sp^2\)?” Suggests considering whether ratio, \(s^2_{\bar{X}}/sp^2\), is large. Gonna need a sampling distribution...

– Sampling dist of this stat is F-dist, so to compare multiple means (or other ests) often can calc an F-stat & do an F-test:

\[
F = \frac{ns^2_{\bar{X}}}{sp^2}
\]

* n in numerator so it will equal denom. on avg, so \(F \geq 1\) key pivot.

– Discussion:

* So F-test used test whether \(>2\) ests (e.g., means) differ.
  · Compares how ests vary cross (sub)samples to how vary w/in.
  · If lots variability w/in smpls, ability to disting diffs cross ↓.
  · (Consider figure; then imagine varying variance of dists.)
  · Lots w/in-smpl var \(\Rightarrow\) @ est (mean) crummy summary of @ individ smpl, so tough tell if ests (means) differ.
  · That’s how hypoth re: means became test stat of ratio vars.
Standard **Analysis of Variance (ANOVA)** Table:

* ANOVA≈a version of regression (or v.v. if psychologist). ANOVA (& rels MANOVA, ANCOVA, ...) used lot in experimental work.

The **ANOVA Table** for a One-Way Comparison

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares, SS</th>
<th>df</th>
<th>Mean Square, MS</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor A</td>
<td>$SS_A = n \sum_{i=1}^{a} (\bar{X}_i - \bar{X})^2$</td>
<td>$(a - 1)$</td>
<td>$MS_A = SS_A/(a - 1) = ns_\bar{X}^2$</td>
<td>$F = \frac{MS_A}{MS_E} = \frac{n s_\bar{X}^2}{s_p^2}$</td>
</tr>
<tr>
<td>Error or Residual</td>
<td>$SS_E = \sum_{i=1}^{a} \sum_{t=1}^{n} (X_{it} - \bar{X}_i)^2$</td>
<td>$a(n - 1)$</td>
<td>$MS_E = SS_E/a(n - 1) = s_p^2$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$SS = \sum_{i=1}^{a} \sum_{t=1}^{n} (X_{it} - \bar{X})^2$</td>
<td>$na - 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• – *$SS_A$*: **Sum of Squares** for factor A.
  • Researcher thinks factor A helps explain diff b/w means.
  • $SS_A \equiv$ variance of $\bar{X}_i$ around Grand Mean $\bar{X}$.
  • What’s $a$? [# values of explanatory variable A, # categories]
  • $n =$? (# of observations).

* $SS_E \equiv$ **Sum of Squared Errors**.
  • Notice from formula: $SS_E$ =w/in-smpl variance.
Here, w/in-smpl var, being var around w/in-smpl means, reveals extent means by that categorization poor descript data.

N.b., “residual” = “error”: “Residual” refers to variation remaining, not explained (& not sought to be). Don’t want (well, sometimes can’t, sometimes don’t want) to explain error, it’s stochastic (random) part of world (or its unexplained/inexplicable part, which treat as stochastic).

* Note col for degrees of freedom:
  - $(a - 1)$ for # levels of factor or values of explanatory variable, or categories minus one for $SS_A$
  - $a(n - 1)$ for n sample size w/in @ of a levels (see how first calc var w/in level then sum over a?).

* Then Mean Squares for factor & residual: $SS/DegFree$,

* Which just as need for F-stat/test (F for Sir Ronald Fisher):
  \[
  F = \frac{\text{explained variance}}{\text{unexplained variance}}
  \]

  - And F fluctuates around 1. Why? (WW p. 328, fn2)

  - Null: @ subsmpl drawn from same pop. If so, 2 ways calc var:
    - (1) pooled var (denominator) & (2) estimate pooled var from var across subsmpl means [since $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$, so $\sigma^2 = \sigma^2_{\bar{X}} \cdot n$.
    - Under null, either method gives (same) right answer, so $F=1$.
    - If null false, in partic if subsmpl means differ, then numerator est of variance will ↑↑ & so reject.

* $SS \equiv Total \ Sum \ of \ Squares$ = total variability in data
  - So $TSS$ = sum of cross-sample variability ($Explained \ Sum \ of \ Squares$) & w/in-smpl variability ($Residual \ Sum \ of \ Squares$).
  - $TSS$ = all there to explain; $ESS$ = explained; $RSS$ = unexplained.

  - Notice: “explain” here is “account for using the different means”

* Can also sum degrees of freedom by parts: df assoc w/ explained var, w/ residual (unexplained, error) var, & total.
* Notice: we’re talking *explanations* all of a sudden.

---

**EXAMPLE WW 10-3:** Suppose rndm smpl of 5 single-family home sales in several cities in ‘85 ⇒ (in 1K US$):

<table>
<thead>
<tr>
<th>City</th>
<th>Population (millions)</th>
<th>Sample of Home Prices</th>
<th>( \bar{X} )</th>
<th>( \sum (X - \bar{X})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>2.8</td>
<td>(110,160,93,206,171)</td>
<td>148</td>
<td>8,510</td>
</tr>
<tr>
<td>Indianapolis</td>
<td>1.2</td>
<td>(72,38,45,108,42)</td>
<td>61</td>
<td>3,480</td>
</tr>
<tr>
<td>Rochester</td>
<td>1.0</td>
<td>(88,66,112,47,52)</td>
<td>73</td>
<td>2,910</td>
</tr>
<tr>
<td>San Diego</td>
<td>2.1</td>
<td>(57,81,181,165,106)</td>
<td>118</td>
<td>11,410</td>
</tr>
</tbody>
</table>

---

So what is substantive Q?

So what is the null hypothesis?

Start with a plot? A box plot? A density plot?

---

- Look different. Let’s ANOVA & F-test it to check:

\[
\bar{X} = \frac{148 + 61 + 73 + 118}{4} = \frac{400}{4} = 100
\]

\[
SS_A = \sum_{i=1}^{4} 5(\bar{X}_i - \bar{X})^2 = 5(148 - 100)^2 + 5(61 - 100)^2 + 5(73 - 100)^2 + 5(118 - 100)^2
\]

\[
= 5(48)^2 + 5(-39)^2 + 5(-27)^2 + 5(18)^2 = 5(2304) + 5(1521) + 5(729) + 5(324) = 11520 + 7605 + 3645 + 1620 = 24390
\]

\[
SS_E = \sum_{i=1}^{4} \sum_{t=1}^{n} (X_{it} - \bar{X}_i)^2
\]
but they’ve already given us the inside part of this so

\[ = 8510 + 3480 + 2910 + 11410 = 26310 \]

* So, \( MS_A = \frac{24390}{3} = 8130 \), & \( MS_E = \frac{26310}{44} = 1644.375 \).

* So \( F = \frac{8130}{1644} = 4.95 \), which way over 1; meaning? If 1 or close?

* How about a p-value? How far is “way over”? F-dist w/ numer & denom df (ddf=16, ndf=3).

\[
\begin{align*}
F_{.25} &= 1.51 \\
F_{.10} &= 2.46 \\
F_{.05} &= 3.24 \\
F_{.01} &= 5.29 \\
F_{.001} &= 9.00
\end{align*}
\]

* Here’s rel F-table; what p-level? meaning?

* The ANOVA Table for this; explain in ordinary language?

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cities</td>
<td>24,390</td>
<td>3</td>
<td>8,130</td>
<td>4.94</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>Residual</td>
<td>26,310</td>
<td>16</td>
<td>1,644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50,700</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Two-Way ANOVA**

* Could even look for >1 reason for diffs means. E.g., in ex., not just city but...? Usu good idea to explore >1 reason (explanation) b/c could be mistaken why some relation exists in data.

* Example: shoe size & spelling ability. Smpls 5 people, @ all ages (0-90). Want understand/explain spelling ability.


· Let’s try age too. Works! AND, suddenly shoe size doesn’t seem to matter... Hmm

· Often need >1 explan var to get right answer, or at least identify & disentangle wrong ones.

* How do it: “Predict/Explain” shoe size by age & vv:

\[
\hat{X}_{ij} = \bar{X} + (\bar{X}_i - \bar{X}) + (\bar{X}_j - \bar{X})
\]
$$RSS : X_{ij} - \hat{X}_{ij}$$

$$SS = SS_A + SS_B + SS_E$$

F for @ factor is then the same as before.

*EXAMPLE: WW 10-15 - Experiment to compare 3 varieties potatoes, assigning @ variety rndmly to 3 equal-sized plots w/ diff soil types. Yields, in bushels/plot:

<table>
<thead>
<tr>
<th>Variety of Potato</th>
<th>Soil</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>21</td>
<td>20</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Clay</td>
<td>16</td>
<td>18</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Loam</td>
<td>23</td>
<td>31</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

· Substantive question: What enhances potato yield: kind of potato or kind of land?
· Null: nothing matters; neither appreciably helps.
· What would boxplots showing Null=true/false look like?
· ANOVA:

$$\bar{X} = \frac{21 + 16 + 23 + \ldots}{9} = 20$$

Sand : $$\frac{21 + 20 + 16}{3} = 19$$
Clay : $$\frac{16 + 18 + 11}{3} = 15$$
Loam : $$\frac{23 + 31 + 24}{3} = 26$$

$$\bar{X}_{i,soil} = (19 + 15 + 26)/3 = 20$$

TypeAPot : $$\frac{21 + 16 + 23}{3} = 20$$
TypeBPot : $$\frac{20 + 18 + 31}{3} = 23$$
\[ \text{TypeCPot} : \frac{16 + 11 + 24}{3} = 17 \]

\[ \bar{X}_{\text{type},j} = \frac{(20 + 23 + 17)}{3} = 20. \]

\( SS_E = \text{residual}. \) Let’s calc using fitted/predicted vals subtracted from obs’d (notice: obs’d-expect’d idea here). Fitted:

<table>
<thead>
<tr>
<th>Soil</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>20+(19-20)+(20-20)=19</td>
<td>20+(19-20)+(23-20)=22</td>
<td>20+(19-20)+(17-20)=16</td>
</tr>
<tr>
<td>Clay</td>
<td>20+(15-20)+(20-20)=15</td>
<td>20+(15-20)+(23-20)=18</td>
<td>20+(15-20)+(17-20)=12</td>
</tr>
<tr>
<td>Loam</td>
<td>20+(26-20)+(20-20)=26</td>
<td>20+(26-20)+(23-20)=29</td>
<td>20+(26-20)+(17-20)=23</td>
</tr>
</tbody>
</table>

Residuals:

<table>
<thead>
<tr>
<th>Soil</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>21-19=2</td>
<td>20-22=-2</td>
<td>16-16=0</td>
</tr>
<tr>
<td>Clay</td>
<td>16-15=1</td>
<td>18-18=0</td>
<td>11-12=-1</td>
</tr>
<tr>
<td>Loam</td>
<td>23-26=-3</td>
<td>31-29=2</td>
<td>24-23=1</td>
</tr>
</tbody>
</table>

\[ SS_{\text{variety}} = 3[(20-20)^2+(23-20)^2+(17-20)^2] = 3[0+9+9] = 54 \]
\[ SS_{\text{soil}} = 3[(19-20)^2+(15-20)^2+(26-20)^2] = 3(1+25+36) = 186 \]
\[ SS_E = 2^2 + (-2)^2 + \cdots = (4 + 4 + 0 + 1 + 1 + 9 + 4 + 1) = 24 \]

So: Two-Way ANOVA Table:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variety</td>
<td>54</td>
<td>2</td>
<td>54/2=27</td>
<td>27/6=4.5</td>
<td>p&lt;.10</td>
</tr>
<tr>
<td>Soil</td>
<td>186</td>
<td>2</td>
<td>186/2=93</td>
<td>93/6=15.5</td>
<td>p&lt;.05</td>
</tr>
<tr>
<td>Residual</td>
<td>24</td>
<td>4</td>
<td>24/4=6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>264</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What does this mean?

Just did **Variance Decomposition**: Part due variety, part soil, & rest residual.

Some R code to implement this:

```r
###TWO WAY ANOVA###
yield<-c(21,20,16,16,18,11,23,31,24)
```
```r
potatAnova <- aov(yield ~ varieties + soils)
summary(potatAnova)
```

```
Call:
potatAnova <- aov(yield ~ varieties + soils)

Residuals: Df Sum Sq Mean Sq F value Pr(>F)
varieties 2 54 27 4.5 0.09467 .
soils 2 186 93 15.5 0.01306 *
Residuals 4 24 6 ---- -----

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1
```

• **Transition to Regression**

- Another technique to sort role variables play in gen 1-another: **regression**, or (subset) **ordinary least squares**.

- In **OLS** regression, variable to explain or understand, called “response” or **dependent variable**, usually labeled **Y**.

- Then have stuff we think explains dep var, called “explanatory” vars, “covariates”, or **independent variables**, usu. lab’d **X**.

- And, usu., want know not just whether X & Y covary, whether X variation X can account Y variation, which ANOVA tells, but also:

  * **THE EFFECT OF X ON Y**: i.e., “When X ↑ certain amount, on avg, what happens to Y?” I.e., \( \frac{\partial Y}{\partial X} \), or \( \frac{\Delta Y}{\Delta X} \), or \( E(Y | X = X_1) - E(Y | X = X_0) \).

  * Find this by something not much different from ANOVA:
    - Offer some \( E(Y) = f(X) \).
    - If linear, i.e., OLS, then that’s \( E(Y) = a + b X \), or \( Y = a + b X + e \).
    - Then find a & b such that min sum-squared-e (“least squares”).
    - So, \( \min_{a,b} [Y - (a + bX)]^2 \), which ~like ANOVA’s \( (X - \bar{X})^2 \) stuff

• **Comparing ANOVA & Regression**

- ANOVA & OLS can do same things w/ categorical vars; just diff norms re: what aspects results important to display.
- Cities & house prices example from above:

* ANOVA:

```r

> cities <- factor(rep(c("Boston", "Indianapolis", "Rochester", "San Diego"), rep(5, 4)))

> theanova <- aov(prices ~ cities)

> summary(theanova)

```

```
Df    Sum Sq Mean Sq F value Pr(>F)
Cities    3  24390.0   8130.0   4.9449  0.01287 *
Residuals 16  26306.0   1644.1 

Signif. codes:  `***' 0.001 `**' 0.01 `*' 0.05 `. ' 0.1
```

```r
> model.tables(theanova, type = "means")
```

Tables of means
Grand mean 100
Cities' means:
Boston Indianapolis Rochester San Diego
148 61 73 118

* OLS:

```r
> theols <- lm(prices ~ cities)

> summary(theols)

```

```
Call: lm(formula = prices ~ cities)

Residuals:
```
Min 1Q Median 3Q Max
-61.00 -23.75 -9.50 27.00 63.00

Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 148.00   | 18.13      | 8.162   | 4.27e-07 |
| citiesIndianapolis  | -87.00   | 25.64      | -3.393  | 0.00372  |
| citiesRochester     | -75.00   | 25.64      | -2.925  | 0.00992  |
| citiesSanDiego      | -30.00   | 25.64      | -1.170  | 0.25920  |

Residual standard error: 40.55 on 16 degrees of freedom
Multiple R-Squared: 0.4811, Adjusted R-sq: 0.3838
F-statistic: 4.945 on 3 and 16 DF, p-value: 0.01287

- Two-way ANOVA & Multiple Regression
  * Two-Way ANOVA shows signif tests for factors “soils” “varieties”
  * OLS shows coeffs assoc’d w/ @ factor, & F-stat for “overall fit”.
  * However, can also ask for anova table for “linear model” object.

  * **Two-Way ANOVA:**
    ```
    > yield <- c(21, 20, 16, 16, 18, 11, 23, 31, 24)
    ```
    ```
    ```
    ```
    ```
    ```
    > potatAnova <- aov(yield ~ varieties + soils)
    ```
    ```
    > summary(potatAnova)
    ```

    Df  Sum Sq Mean Sq  F value    Pr(>F)
    ---
    Fall 2005 – Rob Franzese
varieties  2   54   27   4.5  0.09467 .
soils   2   186   93  15.5  0.01306 *
Residuals  4   24   6   ----
Signif. codes:  0.001 `**' 0.01 `*' 0.05 `. ' 0.1 `.' 1

* Multiple (OLS) Regression

> potatOLS <- lm(yield ~ varieties + soils)
> summary(potatOLS)

Call: lm(formula = yield ~ varieties + soils)

Residuals:
    1     2     3     4
2.000e+00 -2.000e+00 2.313e-16 1.000e+00
    5     6     7     8
-1.767e-15 -1.000e+00 -3.000e+00 2.000e+00
    9
1.000e+00

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        15.0000    1.8264   8.2168  0.00120 **
varietiesB          3.0000    2.0000   1.5000  0.20800
varietiesC         -3.0000    2.0000  -1.5000  0.20800
soilsLoam        11.0000    2.0000   5.5000  0.00533 **
soilsSand          4.0000    2.0000   2.0000  0.11612

---
Signif. codes:  0.001 `**' 0.01 `*' 0.05 `. ' 0.1 `.' 1

Residual standard error: 2.449 on 4 degrees of freedom
Multiple R-Squared: 0.9091,  Adjusted R-sq: 0.8182
F-statistic: 10 on 4 and 4 DF,  p-value: 0.02329
* ANOVA Output from OLS Regression:

\[ > \text{anova(potatOLS)} \]

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>varieties</td>
<td>2</td>
<td>54</td>
<td>27</td>
<td>4.5</td>
<td>0.09467 .</td>
</tr>
<tr>
<td>soils</td>
<td>2</td>
<td>186</td>
<td>93</td>
<td>15.5</td>
<td>0.01306 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>4</td>
<td>24</td>
<td>6</td>
<td>---</td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes:  0.001 `**'  0.01 `*'  0.05 `.'  0.1 ` '  1

• Linear Regression: Overview/Description

− Regression used “[to understand] as far as possible with the available data how the conditional distribution of the response \( y \) varies across subpopulations determined by the possible values of the predictor variables.” (Cook & Weisberg 1999, p. 27 in Berk 2004, p. 10)

− Usu. “conditional distribution of the response” refers to just 1 feature of this dist: the mean. So, write:

\[
y|x = \bar{y}|x + (y|x - \bar{y}|x) = \bar{y}|x - e|x
\]

− where \( e|x \) is “error” but not nec. in causal sense, it just the part of \( y|x \) that not the mean.

− In linear regression, let mean of \( y \) given \( x \) be straight line:

\[
\bar{y}|x = \beta_0 + \beta_1 x
\]

− So can write conditional mean of \( y \) as linear function of \( x \):

\[
y|x = \beta_0 + \beta_1 x + e|x
\]

− If true that mean \( y \) given \( x \) = straight line, then \( e|x = 0 \) (on avg, \( y > \bar{y} \) & \( y < \bar{y} \) parts cancel. If not true, then linear inapprop.
Ordinary Least-Squares (OLS) Regression

- Want to find $a$ & $b$ in $\hat{Y} = a + bX$.
  - I.e., a line (straight), with intercept $a$ and slope $b$.
  - Say data on $X$ & $Y$. Think related. Scatterplot $Y$ vs. $X$. [DRAW]
  - How best summarize? Could do means $Y$ by categories $X$, but say $X$ too many values b/c interval or ratio.
  - Let’s choose straight line that best fits data in some way. How?
  - Legendre & Gauss: find $a$ & $b$ minimize sum of squared residuals:
    \[
    \min_{a,b} \sum (Y_i - \hat{Y}_i)^2
    \]
    \[
    \hat{Y}_i = a + bX_i
    \]
  - So, how? Since said minimize, should think derivative, set equal 0, check 2\(^{nd}\) derivative $> 0$.
  - Since two parameters, $a$ & $b$, want partial derivatives w.r.t @ set equal 0 & both 2\(^{nd}\) derivatives $> 0$.
  - Some think easier if first “mean deviate” $X$, i.e. subtract $\bar{X}$:
    \[
    \hat{Y} = a + bX
    \]
    now, add same amount, $b\bar{X}$, to both sides, rearrange, group, define:
    \[
    \hat{Y} + b\bar{X} = a + bX + b\bar{X}
    \]
    \[
    \hat{Y} = a + bX + b\bar{X} - b\bar{X}
    \]
    \[
    \hat{Y} = (a + b\bar{X}) + b(X - \bar{X})
    \]
    \[
    \hat{Y} = a^* + bx
    \]
  - with $a^* \equiv a + b\bar{X}$ and $x \equiv X - \bar{X}$.
  - Now minimize:
    \[
    \frac{\partial \sum (Y - \hat{Y})^2}{\partial a^*} = \frac{\partial \sum (Y - a^* - bx)^2}{\partial a^*}
    \]
    \[
    \sum 2(Y - a^* - bx)^1(-1) \text{ by Power & Chain Rule}
    \]
– Now, solve for $a^*$ that makes this derivative 0.

\[ \sum (Y - a^* - bx) = 0, \] by dividing both sides by -2 to remove it

\[ \sum Y - na^* - b \sum x = 0 \] now, remember that \( \sum x = 0 \)

\[ \sum Y = na^* \]

\[ \frac{\sum Y}{n} = \bar{Y} = a^* \]

– Check 2\textsuperscript{nd} derivative: \( \frac{\partial^2}{(\partial a^*)^2} = 2n \). Positive; so minimum.

– Now, solve for $b$:

\[ \frac{\partial \sum (Y - \hat{Y})^2}{\partial b} = \frac{\partial \sum (Y - a^* - bx)^2}{\partial b} \]

\[ \sum 2(Y - a^* - bx)^1(-x) \] by Power & Chain Rule again

\[ \sum x(Y - a^* - bx) = 0, \] dividing both sides by -2 again

\[ \sum xY - a^* \sum x - b \sum x^2 = 0 \] recall again \( \sum x = 0 \)

\[ \sum xY - b \sum x^2 = 0 \]

\[ \sum xY = b \sum x^2 \]

\[ \frac{\sum xY}{\sum x^2} = b \]

– Check 2\textsuperscript{nd} derivative: \( \frac{\partial^2}{\partial b^2} = 2 \sum x^2 \): Positive, so minimum.

– Still, want $a$ not $a^*$, so: $a^* = a + b\bar{X}$, so $a = a^* - b\bar{X} = \bar{Y} - b\bar{X}$.

– With just $X$ mean deviated, get $b = \frac{\sum xY}{\sum x^2}$, sub: $Y = (y + \bar{Y})$: 
\begin{align*}
b &= \frac{\sum xy}{\sum x^2} \\
    &= \frac{\sum x(y + \bar{Y})}{\sum x^2} \\
    &= \frac{\sum xy + \bar{Y} \sum x}{\sum x^2}, \text{ but } \sum x = 0 \text{ so} \\
    &= \frac{\sum xy}{\sum x^2} \\
\end{align*}

– So, w/ BOTH \( X \) & \( Y \) mean-deviated: \( b = \frac{\sum xy}{\sum x^2} \).

– When neither are mean-deviated, we get:

\[
\min_{a,b} \sum (Y_i - a + bX_i)^2
\]

\[
\frac{\partial}{\partial a} \sum (Y - a - bx)^2 = \sum 2(Y_i - a - bX_i)(-1) = 0
\]

\[
\Rightarrow \sum (Y_i - a - bX_i) = 0 \Rightarrow na = \sum Y_i - b \sum X_i
\]

\[
\Rightarrow a = \frac{\sum Y_i - b \sum X_i}{n} = \bar{Y} - b\bar{X}
\]

\[
\frac{\partial}{\partial b} \sum (Y - a - bx)^2 = \sum 2(Y_i - a - bX_i)(-X_i) = 0
\]

\[
\Rightarrow \sum [(Y_i - a - bX_i)X_i] = 0 \Rightarrow b \sum X_i^2 = \sum Y_iX_i - a \sum X_i
\]

\[
\Rightarrow b = \frac{\sum Y_iX_i - a \sum X_i}{\sum X_i^2}
\]

– which can be solved by substitution for:

\[
a = \frac{(\sum X_i^2)(\sum Y_i) - (\sum X_i)(\sum X_iY_i)}{n \sum X_i^2 - (\sum X_i)^2}
\]

\[
b = \frac{n \sum X_iY - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2}
\]
• Inference:
  – So, can fit line to scatterplot of y vs. x, by min sum sq resids; great, but also want draw inferences & assess certainty re: inferences.
  – To infer, must have assumpts (args, thry) re: dists. 699 will relax some partic assumpts nec for OLS, but for now understand them.

• Assumptions:
  * 1st: all obs must from dists w/ same var. E.g., 3 obs, then $y_1 \ldots y_3$ 3 RVs w/ same var $p(Y_1|X_1), p(Y_2|X_2), \ldots$: same var.
  * 2nd: mean of $Y_i$, i.e. $E(Y_i)$, lies on population regression line: $E(Y_i) = \mu_i = \alpha + \beta X_i$.
  * 3rd: RVs $Y_1, Y_2, \ldots$ stat’ly indep. Know $Y_i \Rightarrow$ nothing re: $Y_j$.
  * These assumptions give mean each obs as $E(Y_i) = \mu_i = \alpha + \beta X_i$ & variance residual, $V(Y_i - \mu_i) \equiv V(\varepsilon_i) \equiv V(Y_i|X) \equiv \sigma^2_{\varepsilon}$.

• Terminology:
  · $E(Y_i) \equiv$ expected, predicted, estimated or fitted value
  · $Y_i = y_i$ realized, observed value
  · $\varepsilon \equiv$ (true) residual, (just like in ANOVA), error, or stochastic term/component or shock or disturbance.
  · We use $\varepsilon$ for true errors & e or sometimes $\hat{\varepsilon}$ for estimated error.

• Correct forms (notation) of regression equation:
  · True (Population) Equation/Model: $Y_i = \alpha + \beta X_i + \varepsilon_i$
  · Estimated Equation/Model: $Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\varepsilon}_i = a + bX_i + e$
  · Expected Equation/Model, Systematic Component: $E(Y_i) = \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = a + bX_i$

  – WW re: $\varepsilon_i$: “the purely random part [i.e., stochastic component] of $Y$;” i.e., contrasted to systematic component. Idea here: a model; models don’t explain everything, so got some stuff that can’t explain, e.g. measurement error, inherent variability, etc...
  – Now, if this variability dist’d normally, then $Y_i$ dist’d normally (b/c
rest systematic, i.e., non-stochastic), then we’ll know how a & b dist’d [sampling dist] (normally) & away we go!

– Study these figs (WW, 12-1,2). Assumpts 1&2, & Pop v. Est Reg:

FIGURE 12-1
(a) General populations of Y, given X. (b) The special form of the populations of Y assumed in simple linear regression.
- Sampling Distributions of a and b

  - Happily, w/ lrg samples, sampling dists a & b normal. Why? Look at formulae again: lots sums ind RVs ($\epsilon$), so:
    * If $\epsilon \sim N(0, \sigma^2)$, then a & b = sum bunch normals, so normal.
    * If $\epsilon$ dist elsewise, then a & b = sum bunch ind RVs, so CLT, so asym normal.
  - If b unbiased est $\beta$, then $E(b) = \beta$, & know **std dev of b** too:
    \[
    SD(b) = \frac{\sigma}{\sqrt{\sum x^2}}
    \]
  - So, how make SE smaller? [DISCUSS]
    * (1) reducing variability of Y (i.e., $\sigma$),
    * (2) increase n,
    * (3) increase variability of X.
  - Alas, don’t know $\sigma^2$. So what do? (Rhetorical: Estimate it!)
    \[
    s^2 \equiv \frac{1}{n-2} \sum (Y - \hat{Y})^2
    \]
- Why \( n - 2 \)? Degrees Freedom. How many things est’d? 2, a & b. So std err of b:

\[
SE(b) = \frac{s}{\sqrt{\sum x^2}}
\]

- w/ unbiased b, & SE, what can we do?

* CI: \( b \pm t_{0.025}SE \)

* H-test: p-values from \( t = \frac{b - \beta_{null}}{SE} \), for null \( \beta = 0 \), \( t = \frac{b}{SE} \)

• Taking Stock: Tons of cool info now:
  - Can fit line to scatter relating \( Y \) to \( X \), w/ ests of slopes & intercepts.
  - In linear model, slope tells: when \( x \uparrow \), how much \( E(y) \uparrow \downarrow \), i.e., \( \frac{\partial y}{\partial x} \), i.e., EFFECT of \( X \) on \( Y \)!!
  - Then can draw inferences about TRUE \( \beta \) from est’d b:
    * b is correct on avg (unbiased).
    * And know what SD(b) is f() of, & can est it w/ SE(b).
    * And know how b dist across repeated samples, so...
    * Can CI & test & gen’ly say how strong evidence is that \( \beta \) near b or far from zero, etc.
  - Later, we’ll extend to >1 \( X \) & story even more exciting!
  - Now, key Q remaining: how do & use all this stuff?

• EXAMPLE:
  - Data: \[
    \begin{array}{c}
    Y & 0 & 0 & 1 & 1 & 3 \\
    X & -2 & -1 & 0 & 1 & 2 \\
    \end{array}
  \]
  - How calc OLS coeffs? Let’s use non-mean-deviated forms:

\[
a = \bar{Y} - b \bar{X}
\]

\[
b = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2}
\]

* So, need \( X_i Y_i, X_i^2, \sum X_i, \sum Y_i \) (last 2 for \( \bar{Y} & \bar{X} \)).
\[ Y_i \quad X_i \quad X_i \cdot Y_i \quad X_i^2 \]

\[
\begin{array}{cccc}
0 & -2 & 0 & 4 \\
0 & -1 & 0 & 1 \\
1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 \\
3 & 2 & 6 & 4 \\
\end{array}
\]

\[
\sum Y_i = 5 \quad \sum X_i = 0 \quad \sum X_i Y_i = 7 \quad \sum X_i^2 = 10
\]

* And \( \bar{Y} = 1 \) and \( \bar{X} = 0 \). So:

\[
b = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2} = \frac{5(7) - 0 \cdot 5}{5(10) - 0^2} = \frac{35}{50} = \frac{7}{10} = .7
\]

\[
a = \bar{Y} - b\bar{X} = 1 - .7(0) = 1
\]

\[
\Rightarrow \hat{Y}_i = 1 + .7X_i
\]

* Let’s work with (interpret) this a bit:
  · When \( X = 0 \), \( E(Y) = ? \)
  · When \( X = 1 \), \( E(Y) = ? \)
  · If \( X \uparrow 1 \), \( E(Y) \uparrow ? \)

* Here’s scatterplot w/ fitted line. What are vertical lines?

* What’s SE of \( b \)?

\[
\sqrt{\frac{1}{n-2}(Y-\hat{Y})^2} \]

\[
\sum X^2
\]
\[
\begin{array}{cccccc}
Y & \hat{Y} & Y - \hat{Y} & (Y - \hat{Y})^2 \\
0 & 1+.7(-2) & = -0.4 & 0.4 & 0.16 \\
0 & 1+.7(-1) & = 0.3 & -0.3 & 0.09 \\
1 & 1+.7(0) & = 1.0 & 0.0 & 0.00 \\
1 & 1+.7(1) & = 1.7 & -0.7 & 0.49 \\
3 & 1+.7(2) & = 2.4 & 0.6 & 0.36 \\
\end{array}
\]

\[
\sum (Y - \hat{Y})^2 = 1.1 \Rightarrow s^2 = \frac{1.1}{3} \approx .37 \Rightarrow SE = \sqrt{\frac{.37}{10}} = \frac{.61}{3.16} = .19
\]

* So, \( b = .7, \) \( SE = .19, \) & \( t_{df=(n-2)=3} = b/SE = .7/.19 = 3.68. \)

* So, for 1-tailed, \( p = .017 \) (prob est this \( b \) or greater if null hypoth of \( \beta = 0 \) true); for 2-tailed \( p = 2(.017) = .034 \) (prob est \( b \) this far or farther from 0 if null that \( \beta = 0 \) true). Conclusion?

- **ANOVA Table for Example Regression**
  - Can show that Total Deviations=Explained + Unexplained:
    \[
    (Y - \bar{Y}) = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})
    \]
  - Also true that Total Sum Squares=Explained SS+Unexplained SS:
    \[
    \sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2
    \]
  - This is so b/c cross-product in quadratic =0; can you see why?
  - With these, can make ANOVA table from OLS, & corresp. \( F \)-stat=
    \[
    F = \frac{\text{variance explained by the regression}}{\text{variance not explained}}
    \]
  - Our e.g.: \( \sum (\hat{Y} - \bar{Y})^2 = (-.4 - 1)^2 + (.3 - 1)^2 + \ldots = 4.9, \) and \( \sum (Y - \hat{Y})^2 = (0 - (-.4))^2 + (0 - .3)^2 + \ldots = 1.1 \)

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>4.9</td>
<td>1</td>
<td>4.9/1=4.9</td>
<td>F=4.9/.37=13.24</td>
</tr>
<tr>
<td>Residual</td>
<td>1.1</td>
<td>(5-1-1)=3</td>
<td>1.1/3=.37</td>
<td></td>
</tr>
</tbody>
</table>

- p-value for this \( F \)-stat is .036. What hypoth does it test?
• Regression via MLE

– MLE ⇒ another route to OLS ests & sampling dists thereof.
  * (Linear) model \( \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \) ⇒ LS estimators:
    \[
    \hat{\beta}_{0,LS} = \bar{Y} - \hat{\beta}_{1,LS} \bar{X}
    \]
    \[
    \hat{\beta}_{1,LS} = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}
    \]

* Sampling dist LS estimator?
  · If \( \varepsilon \sim N(0, \sigma^2) \), then \( \hat{\beta}_{LS} = \sum \text{(bunch of normals)} \sim N \).
  · If \( \varepsilon \sim \text{else} \), then \( \hat{\beta}_{LS} = \sum \text{(bunch of normals)} \sim a \ N \) by CLT.

* What properties MLE? [BA\text{NC}]

– So, first step of ML? Need a dist from which obs drawn. I know: how about Normal?

\[
Y_i \sim N(\mu_i, \sigma^2) \equiv f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \mu_i)^2}{2\sigma^2} \right)
\]

– So, want straight line through means of \( y \) given \( x \), right? So, \( \mu_i = \) straight-line function of \( X_i \). So, parameterize \( \mu_i \):

\[
\mu_i = \beta_0 + \beta_1 X_i
\]

– So, substitute this into previous equation:

\[
f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right)
\]

– Or we can equivalently write:

\[
y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)
\]

– So, that’s probability of one obs, want likelihood our set of obs. How? If indep, joint prob/like all \( y_i = \) product individ \( p(\cdot) \):

\[
f(y_1, \ldots, y_n|\beta_0, \beta_1, \sigma^2) = f(y_1|\mathbf{B}, \sigma^2) f(y_2|\mathbf{B}, \sigma^2) \ldots f(y_n|\Theta) = \prod_{i=1}^n f(y_i|\Theta) \Rightarrow L(\Theta|y_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right)
\]

\[
\ln L(\beta_0, \beta_1, \sigma^2 | y_i) = \sum_{i=1}^{n} \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) \right)
\]

Logs: \( \ln(ab) = \ln a + \ln b; \ln \left( \frac{a}{b} \right) = \ln a - \ln b; \ln a^b = b \ln a; \ln e^x = x. \)

\[
\ln L(\Theta | y_i) = \sum_{i=1}^{n} \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) =
- \frac{1}{2} \sum_{i=1}^{n} \left( \ln(2\pi) + \ln(\sigma^2) + \frac{1}{\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2 \right) =
- \frac{1}{2} \sum_{i=1}^{n} \left( \ln 2 + \ln \pi + \ln \sigma^2 + \frac{1}{\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2 \right) =
- \frac{n(\ln 2 + \ln \pi + \ln \sigma^2)}{2} - \frac{1}{2\sigma^2} \sum (\beta_0^2 + 2\beta_0\beta_1 x_i +
\beta_1^2 x_i^2 - 2\beta_0 y_i - 2\beta_1 x_i y_i + y_i^2) =
- \frac{n(\ln 2)}{2} - \frac{n\ln \pi}{2} - \frac{n\ln \sigma^2}{2} - \frac{n\beta_0^2}{2\sigma^2} - \frac{2\beta_0\beta_1}{2\sigma^2} \sum x_i - \frac{\beta_1^2}{2\sigma^2} \sum x_i^2
+ \frac{2\beta_0}{2\sigma^2} \sum y_i + \frac{2\beta_1}{2\sigma^2} \sum x_i y_i - \frac{1}{2\sigma^2} \sum y_i^2
\]

Set partial derivatives w.r.t. 3 params = 0, i.e. max:

\[
\frac{\partial \ln L(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = -\beta_0 n - \beta_1 \sum x_i + \sum y_i = 0
\]

\[
\Rightarrow -\beta_0 n - \beta_1 \sum x_i + \sum y_i = 0 \Rightarrow \hat{\beta}_{0ML} = \bar{y} - \hat{\beta}_{1ML} \bar{x}
\]

\[
\frac{\partial \ln L(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = -\beta_0 \sum x_i - \beta_1 \sum x_i^2 + \sum x_i y_i = 0
\]

\[
\Rightarrow -\hat{\beta}_{0ML} \sum x_i - \hat{\beta}_{1ML} \sum x_i^2 + \sum x_i y_i = 0
\]

\[
\frac{\partial \ln L(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = \ldots
\]
- Gives (after some more algebra) MLE (look familiar?):

\[
\hat{\beta}_{0,ML} = \frac{\sum x_i \sum y_i + \sum x_i \sum x_i y_i}{(\sum x_i)^2 - n \sum x_i^2} = \bar{y} - \hat{\beta}_{1,ML} \bar{x} = \hat{\beta}_{0,LS}
\]

\[
\hat{\beta}_{1,ML} = \frac{\sum x_i \sum y_i - n \sum x_i y_i}{(\sum x_i)^2 - n \sum x_i^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = \hat{\beta}_{1,LS}
\]

\[
\hat{\sigma}^2_{ML} = \frac{1}{n} \sum (y_i - \hat{\beta}_{0,ML} - \hat{\beta}_{1,ML} x_i)^2 \rightarrow \hat{\sigma}^2_{LS}
\]

- And that Hessian thingee, 

\[-H^{-1} = \frac{1}{nI(\theta)} = -[L'']^{-1},\]

gives the standard errors...

• **Inference: Properties & Sampling Dists** \( \hat{\beta}_{ML} \) and \( \hat{\beta}_{LS} \)

- MLE:

  * Assumptions:
    - \( Y \sim \text{Normal} \)
    - \( E(Y|X) = \beta_0 + \beta_1 X \)
    - \( V(Y|X) = \sigma^2 \)
    - \( Y|X \) (stochastically) independent.
    - (Act’ly, if normal, constant var, & indep, then \( E(Y|X) \) linear.)

  * \( \Rightarrow \) BANC: asym’ly consistent, normal, efficient [& invariant].

  * \( \Rightarrow \) \( \hat{\beta}_{ML} = (\bar{y} - \hat{\beta}_{1,LS} \bar{x}, \frac{\text{Cov}(x,y)}{\text{Var}(x)}) \sim a \ N(\beta, -H^{-1}) \)

- LSE:

  * Assumptions:
    - \( Y = \beta_0 + \beta_1 X + \varepsilon \) (Linear additive mean/systematic & additively separable stochastic component.
    - \( E(\varepsilon) = 0; V(\varepsilon) = \sigma^2 \)
    - \( \varepsilon \) (weakly) independent.

  * \( \Rightarrow \) \( \hat{\beta}_{LS} = (\bar{y} - \hat{\beta}_{1,LS} \bar{x}, \frac{\text{Cov}(x,y)}{\text{Var}(x)}) \)

  - Dist’d norm’ly if \( \varepsilon \sim N \); else, dist’d asym. norm’ly by CLT.
  - Mins sum sq’d errs, but not shown any other props yet.
Inference:

* Population model: \( Y_i = \alpha + \beta X_i + \varepsilon_i \)

* Will show soon: \( E(\hat{\mathbf{B}}) = \mathbf{B} \) & \( SD_{\hat{b}_1} = \sigma \sqrt{\sum X_i^2} \) (asym’mly for ML).

* Also know that: \( \hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum (Y_i - \hat{Y}_i)^2 \) & \( SE_{\hat{b}_1} = \frac{s}{\sqrt{\sum X_i^2}} \)

* SO, we now have:

  \[
  \cdot (1 - \alpha)\% \ CI: \beta = b_1 \pm T_{\alpha/2,n-2} SE_{\hat{b}_1} \\
  \cdot T-stat for H_0 : \beta = 0 is t_{n-1} = \frac{b_1}{SE}.
  \]

- Returning to our very-small-dataset example:

  - So the \( SE(\hat{b}_1) = \sqrt{\frac{1}{n-2}(Y_i - \hat{Y}_i)^2}{\sum X_i^2} \)

<table>
<thead>
<tr>
<th>y</th>
<th>( \hat{y} )</th>
<th>( y - \hat{y} )</th>
<th>( (y - \hat{y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1+.7(-2)</td>
<td>-0.4</td>
<td>0.16</td>
</tr>
<tr>
<td>0</td>
<td>1+.7(-1)</td>
<td>0.3</td>
<td>0.09</td>
</tr>
<tr>
<td>1</td>
<td>1+.7(0)</td>
<td>1.0</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>1+.7(1)</td>
<td>1.7</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>1+.7(2)</td>
<td>2.4</td>
<td>0.36</td>
</tr>
</tbody>
</table>

\[
\sum (Y - \hat{Y})^2 = 1.1; \ldots s^2 = \frac{1.1}{3} \approx .37; \ldots SE_{\hat{b}_1} = \sqrt{\frac{.37}{10}} = \frac{.61}{3.16} = .19
\]

\[
\Rightarrow \hat{b}_1 = .7; \ldots SE_{\hat{b}_1} = .19 \ldots and \ldots t_{df=(n-2)=3} = \frac{\hat{b}}{SE} = \frac{.7}{.19} = 3.68
\]

- So, 1-tailed p=.017 (prob \( \hat{b} \geq \) to this est if null hyp \( \beta = 0 \) true); 2-tailed p=2(.017)=.034 (prob \( \hat{b} \) so far or farther from 0 if null true).

- Substantive interpretation of estimate & certainty?

- Properties of OLS Estimator:

  - Determine properties by stating what assumed true, then prove how estimator behaves across repeated samples under those conditions.

  * Diff assumpts OLS pertain to diff parts of estimation exercise:
· Weakest assumpts req’d for unbiased coeff ests: \( E(\hat{\Theta}) = \Theta \).
· Stronger assumpts for SD & good SE props: \( V(\hat{\Theta}) \) & \( V(\Theta) \).
· Yet more assumpts for hypothesis tests: \( pdf(\hat{\Theta}) \).
· At strongest, OLS = MLE.

**Gauss-Markov Theorem**

– Statements of Theorem:

* Rob/Jake/Nancy/John: Under certain assumpts, \( \hat{B}_{OLS} \) is **minimum variance** (Best) **Linear Unbiased Estimator** of B. I.e., **OLS is B.L.U.E.**

\[
E(\hat{\beta}_{1,OLS}) = \beta_1 \ldots \text{and} \ldots E(\hat{\beta}_{0,OLS}) = \beta_0 \\
V(\hat{\beta}_{1,OLS}, \hat{\beta}_{0,OLS}) < V(\hat{\beta}^*_1, \hat{\beta}^*_0)
\]

* where, \( \hat{\beta}_{1,OLS} \) is OLS est & \( \hat{\beta}^*_1 \) some other lin-unbiased est.

* Gujarati, 73: “Given the assumptions of the classical linear regression model [CLRM], the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.”

* I.e., among all poss linear (straight line through data) & unbiased (right on avg) ests, OLS has minimum variance.

– Proof steps: assumpts; show LSE linear; find \( E(LSE) \); show unbiased; find \( V(LSE) \); show \( V(LSE) \leq V(\text{lin & un-b est}) \).

* **Assumptions** (\( C(N)LRM \)): (simple set: some redundancy, overlap; not all nec. for all props; etc.):

  · \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \)
  · \( E(\epsilon_i) = 0 \quad \forall i \)
  · \( V(\epsilon_i) = \sigma^2 \quad \forall i \)
  · \( E(\epsilon_i | \epsilon_j) = 0 \quad \forall i \neq j \)
  · \( X \) fixed in repeated samples (non-stochastic).
  · \( (\epsilon_i \sim N(0, \sigma^2) \quad \forall i) \)
**Linear:** tech’ly, means est is linear funct of the data (Y). For our purposes, enough to note that $Y = \hat{b}_0 + \hat{b}_1 X + e$ is line.

- If like more, recall that $\varepsilon$ & so y varies across repeated samples, but x does not. So, linear funct data means coefficient non-varying across samples times thing that does.

$$\hat{\Theta}_{OLS} = \left( \bar{y} - \hat{\beta}_1 \bar{x}, \frac{\sum x_i \sum y_i - n \sum x_i y_i}{(\sum x_i)^2} \right)$$

- Both parameters are thus linear functions of data (some x-term coefficients times y-term data, separated by add/subtract).

**Unbiased:** $E(\hat{\Theta}_{OLS}) = \Theta$

- Simplify: $E(X = \bar{X} = 0 \& E(Y) = \bar{Y} = 0$. Then:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\hat{\beta}_{1,OLS} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum XY}{\sum X^2}$$

- Now, to find $E(\hat{\beta}_{1,OLS})$, sub true/assumed formula for Y:

$$E(\hat{\beta}_{1,OLS}) = E \left( \frac{\sum X_i Y_i}{\sum X_i^2} \right)$$

$$= E \left( \frac{\sum X_i (\beta_0 + \beta_1 X_i + \varepsilon_i)}{\sum X_i^2} \right)$$

$$= E \left( \frac{\sum (X_i \beta_0 + \beta_1 X_i^2 + X_i \varepsilon_i)}{\sum X_i^2} \right)$$

$$= E \left( \frac{\beta_0 \sum X_i}{\sum X_i^2} \right) + E \left( \frac{\sum (\beta_1 X_i^2)}{\sum X_i^2} \right) + E \left( \frac{\sum X_i \varepsilon_i}{\sum X_i^2} \right)$$

Since avg X=0, first term=0 — i.e. no intercept

$$= E \left( \frac{\beta_1 \sum X_i^2}{\sum X_i^2} \right) + E \left( \frac{\sum X_i \varepsilon_i}{\sum X_i^2} \right)$$
\[ \sum X_i^2 \text{ terms cancel, } \& \beta_1 \text{ constant, so} \]
\[ = \beta_1 + E \left( \frac{1}{\sum X_i^2} \sum (X_i \epsilon_i) \right) \]

Now, recall interested in \( Y | X, \& X \text{ fixed, so} \)
\[ = \beta_1 + \frac{1}{\sum X_i^2} E \left( \sum X_i \epsilon_i \right) = \beta_1 + \frac{1}{\sum X_i^2} \sum X_i E(\epsilon_i) \]

\[ \cdot \text{ So, OLS unbiased by assumpt } E(\epsilon_i) = 0. \text{ (Note: w/o fixed-X assumpt, need } E(\epsilon | X) = 0.) \]

* Variance:
\[ V(\hat{B}_{OLS}) = V \left( \sum \frac{X_i \beta_0}{\sum X_i^2} + \sum \frac{\beta_1 X_i^2}{\sum X_i^2} + \sum \frac{X_i \epsilon_i}{\sum X_i^2} \right) \]
\[ \cdot X \text{ fixed (non-stochastic, i.e. constant) in repeated samples, so how } X \text{ or } X^2 \text{ vary or covary } w/ \text{ only stochastic part, } \epsilon? \]
\[ V(\hat{B}_{OLS}) = 0 + 0 + 2(0) + 2(0) + 2(0) + V \left( \sum \frac{X_i \epsilon_i}{\sum X_i^2} \right) \]
\[ V \left( \sum \left[ \frac{X_i}{\sum X_i^2} \epsilon_i \right] \right) \]
\[ \cdot \text{ Now, remember } V() \text{ sliding through sums, squaring constant-coefficients (like } X), \text{ and } 2 \text{ times all covariances (which zero)}: \]
\[ V(\hat{B}_{OLS}) = \sum \left[ \left( \frac{X_i}{\sum X_i^2} \right)^2 V(\epsilon_i) \right] = \sum \left[ \left( \frac{X_i}{\sum X_i^2} \right)^2 \sigma^2 \right] \]
\[ V(\hat{B}_{OLS}) = \left[ \frac{\sum X_i^2}{(\sum X_i^2)^2} \right] \sigma^2 = \frac{\sigma^2}{\sum X_i^2} \]

* Minimum Variance (among linear-unbiased):
\[ \cdot \text{ Proof tough } \& \text{ not very revealing, so: Trust us.} \]
Nature of one proof strategy is to show any linear est, incl. OLS, is some \( \sum c_i y \); then, define alternative as \( c_i + \nu_i \). Must have certain prop’s if also unbiased. Then, \( V(b_{OLS}) = (\sum c_i)^2 V(y) \) and \( V(b_{alt}) = (\sum c_i + \nu_i)^2 V(y) \), which \( \geq V(b_{OLS}) \).

Another proof shows OLS, under CNLRM, hits Cramer-Rao lower bound, which = lowest poss var for ests of certain prop’s.

- **Multiple Regression**
Finding Coefficients in Multiple Regression

\[ S(b_0, b_1, b_2) = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2})^2 \]

* “Min” ⇒ set 1\(^{st}\) deriv \(S\) w.r.t. each param to 0

\[ \frac{\partial S}{\partial b_0} = \sum (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2}) = 0 \]

\[ \frac{\partial S}{\partial b_1} = \sum X_{i1} (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2}) = 0 \]

\[ \frac{\partial S}{\partial b_2} = \sum X_{i2} (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2}) = 0 \]

* Above called “normal” equations (confusing name). Solve:

\[ \sum Y_i = nb_0 + b_1 \sum X_{i1} + b_2 \sum X_{i2} \]

\[ \sum Y_i X_{i1} = b_0 \sum X_{i1} + b_1 \sum X_{i1}^2 + b_2 \sum X_{i2} X_{i1} \]

\[ \sum Y_i X_{i2} = b_0 \sum X_{i2} + b_1 \sum X_{i1} X_{i2} + b_2 \sum X_{i2}^2 \]

* Which, with some more algebra gives:

\[ b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \] (dividing by \(n\))

* “Reg line (plane) runs through mean (centroid) of data.”

\[ b_1 = \frac{(\sum Y_i X_{i1})(\sum X_{i1}^2) - (\sum Y_i X_{i2})(\sum X_{i1} X_{i2})}{(\sum X_{i1}^2)(\sum X_{i2}^2) - (\sum X_{i1} X_{i2})^2} \]

* Notice: First part of num & denom basically the bivariate formula. Second part adjusts that by essentially covariance \(X_1\) & \(X_2\) and covariance of \(X_2\) w/ \(Y\). Sensible, right?

\[ b_2 = \frac{(\sum Y_i X_{i2})(\sum X_{i2}^2) - (\sum Y_i X_{i1})(\sum X_{i1} X_{i2})}{(\sum X_{i1}^2)(\sum X_{i2}^2) - (\sum X_{i1} X_{i2})^2} \]

* Notice: The analogous to the above, in reverse.
• **OLS in Matrix Format:**

\[ Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i, \, i = 1, \ldots, n \]

- I.e., possibly *many* explanators (“multicausality”). Still assume:
  * \( e_i \) rndm draws from some p-dist, w/ \( \text{E}(e_i) = 0, \text{Var}(e_i) = \sigma^2 \), \( \text{Cov}(e_i, e_j) = 0 \) \( \forall i \neq j \)
  * \( X \)'s fixed, w/ no 1 or linear combo of some “perfectly colinear” (perf’ly corr’d) w/ any other(s).

- Now define these vectors & matrices:

\[
\begin{align*}
\mathbf{y} &= \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \\
\mathbf{X} &= \begin{bmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1k} \\ x_{20} & x_{21} & \cdots & & \vdots \\ \vdots & & & & \vdots \\ x_{n0} & x_{n1} & \cdots & & x_{nk} \end{bmatrix} \\
\mathbf{\beta} &= \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \\
\mathbf{e} &= \begin{bmatrix} e_0 \\ \vdots \\ e_n \end{bmatrix}
\end{align*}
\]

- Allows write eq very compactly as:

\[
\mathbf{y} \in \mathbb{R}^n \times 1 = \mathbf{X} \in \mathbb{R}^{n \times (k+1)(k+1)} \times 1 \mathbf{\beta} \in \mathbb{R}^{(k+1) \times 1} + \mathbf{e} \in \mathbb{R}^{n \times 1}
\]

- And can show that \( \hat{\mathbf{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \) mins sum sq’d errs.

\[
\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ 1 & \cdots & 1 \\ \vdots & & \vdots \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}
\]

\[
\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}
\]

- Now restate assumptions in Matrix form:
\( E(e) = 0, \)
\( E(ee') = \sigma^2 I \)
\( (\text{All of}) \mathbf{X} \) is fixed (nonstochastic)
\( \mathbf{X}_{n \times k} \) is of rank \( k \) (no perfect colinearity), where \( k = \# \text{ ind vars} \)
\& \( k < n. \)
\( \Rightarrow \) Rice (537): "In words, the \( y \) are equal to the true values of the function plus random, uncorrelated errors with constant variance. Note that in this model the \( X \)'s are fixed, not random."

**Mean of \( \hat{\beta}_{\text{OLS}} \): Unbiasedness**, i.e., \( E(\hat{\beta}_{\text{OLS}}) = \beta \)

- Remember: \( \hat{\beta} = (X'X)^{-1}X'y \)

\[
E(\hat{\beta}) = E[(X'X)^{-1}X'y] \\
= E[(X'X)^{-1}X'(X\beta + \varepsilon)] \\
= E[(X'X)^{-1}X'X\beta] + E[(X'X)^{-1}X'\varepsilon] \\
= \beta + E[(X'X)^{-1}X'\varepsilon]
\]

- But \( X \) fixed, constant, unchanged in repeated smpls, so:

\[
= \beta + (X'X)^{-1}X'E[\varepsilon] \text{ , but } E(\varepsilon) = 0
\]

- So, \( \beta + (X'X)^{-1}X'0 = \beta + 0 = \beta: \text{ Unbiased} \)

- What assumed for this?

- Values \( X \) not change across samples. Rndm part not \( X \), but \( \varepsilon \)

- On avg (mean) across smpls, dev \( y \) vals (@ @ \( X \)) from reg line=0.

- Some Meanings:
  - Factors not in model not systematically matter for dep var in way corr’d w/ \( X \)
  - Stuff in \( \varepsilon \) = only noise, nothing systematic (that related \( X \)).
  - \( X \) not correlate w/ \( \varepsilon \)
  - Have **correct specification** of model (gee, that’s all?)
* BUT, notice, no assumpts re: constant variance, normality, covar.
  errs. Just $E(\varepsilon) = 0$ & $X$ fixed or $\text{Cov}(X, \varepsilon) = 0$.

- What if $X$ not fixed?
  * Change assumpt $E(\varepsilon) = 0$ to $E(\varepsilon|X) = 0$.
  * Add that $y$ does not cause $X$ (not deal w/ this in 599).
  * Then, line in proof $E((X'X)^{-1}X'\varepsilon)$ becomes $E((X'X)^{-1}X'\varepsilon|X)$,
    but that still 0, so $E(\hat{\beta}|X) = \beta$: unbiased.

- **Variance of OLS Estimator**
  - Answer first, then prove, then examine it:
    $$\text{Var}(\hat{\beta}_{\text{OLS}}) = \sigma^2(X'X)^{-1}$$
  - Proof:
    $$\text{Var}(\hat{\beta}) = E\left([\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\right)$$
    $$= E\left([\hat{\beta} - \beta][\hat{\beta} - \beta]'\right)$$
  * Notice that $\hat{\beta} - \beta = (X'X)^{-1}X'\varepsilon$
    $$= E\left([(X'X)^{-1}X'\varepsilon][(X'X)^{-1}X'\varepsilon]'ight)$$
  * Now remember that $(AB)' = B'A'$ and $((X'X)^{-1})' = (X'X)^{-1}$,
    so:
    $$= E\left((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right)$$
  * Since $X$ fixed across repeated samples (nonstochastic):
    $$= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1}$$
  * Now, what about $E(\varepsilon\varepsilon')?...=V(\varepsilon)$:
    $$= (X'X)^{-1}X'\sigma^2I(X'X)^{-1}$$
    $$= \sigma^2I(X'X)^{-1}X'X(X'X)^{-1}$$
    $$= \sigma^2I(X'X)^{-1}$$
  * Examine it:
First, $E(ee')$ looks like?

$$E(ee') = E \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \begin{bmatrix} e_1 & \ldots & e_n \end{bmatrix}$$

$$= E \begin{bmatrix} e_1^2 & e_1e_2 & e_1e_3 & \ldots & e_1e_n \\ e_2e_1 & e_2^2 & e_2e_3 & \ldots & e_2e_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_ne_1 & e_ne_2 & e_ne_3 & \ldots & e_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(e_1^2) & E(e_1e_2) & E(e_1e_3) & \ldots & E(e_1e_n) \\ E(e_2e_1) & E(e_2^2) & E(e_2e_3) & \ldots & E(e_2e_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E(e_ne_1) & E(e_ne_2) & E(e_ne_3) & \ldots & E(e_n^2) \end{bmatrix}$$

Now what do our assumptions do here?

$$\begin{bmatrix} \sigma^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma^2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \sigma^2 \mathbf{I}$$

Now, what needed assume to get this result?

- Constant variance, independent observations.
- Note, still needn’t say “normal”.
- What mean in real world? What kinds of samples might not have these props? Think examples where $V(\varepsilon) \neq \sigma^2 \mathbf{I}$.

E.g., srvy people clustered in towns (say, 100 towns, 100 peo-
ple/town) asked about town politics, what $E(\varepsilon \varepsilon')$ look like?

* If $X$ not fixed, but varies, across repeated samples, then:

$$E(\varepsilon \varepsilon' | X) = \sigma^2 E(X'X)^{-1}$$

* **Estimating $\sigma^2$**
  
  · Don’t know $V(\text{pop dist})$ from which $\varepsilon_i$ drawn, $\sigma^2$, so how est?
  
  · **Unbiased**: $\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-k-1} = \frac{\hat{e}'\hat{e}}{n-k-1}$
  
  · **BANC**: $\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n} = \hat{e}'\hat{e}$

  - For h-tests & CI’s, need sampling dist.

* If $y \sim N(X\beta, \sigma^2)$, or, equivalently, $\varepsilon \sim N(0, \sigma^2)$, then $\hat{\beta}$=linear-additive fnctn $y$, so $\beta|X \sim N_{k+1}[\beta, \sigma^2(X'X)^{-1}]$

* Is $y \sim (X\beta, \sigma^2)$, then $\hat{\beta}|X \sim^a N_{k+1}[\beta, \sigma^2(X'X)^{-1}]$ by CLT.

**Asymptotic Properties: Consistency**

- $\text{plim}(\hat{\beta}_{OLS}) = \beta$
- Already shown $\hat{\beta}$ unbiased=right on average.
- Already shown $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- So, as sample size grows, what happens to this sampling dist?

**A Regression Example:** ‘Explaining’ Prestige of Occupation

- Data:
  
  * **education**: Average education of occupational incumbents, years, in 1971.
  
  * **income**: Average income of incumbents, dollars, in 1971.
  
  * **women**: Percentage of incumbents who are women.
  
  * **prestige**: Pineo-Porter prestige score for occupation, from social survey conducted in mid-1960s.
  
  * **census**: Canadian Census occupational code.
  
  * **type**: Type of occupation. (bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar).
Models (n.b., “Stepwise Regression,” might be called “Step-Un-Wise”, but pedagogical purpose here)

```r
> plm1 <- lm(prestige ~ income, data = Prestige)
> summary(plm1)
```

```
Call: lm(formula = prestige ~ income, data = Prestige)
Residuals:
   Min     1Q   Median     3Q    Max
-33.007 -8.378  -2.378   8.432  32.084
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.714e+01  2.268e+00  11.97   <2e-16 ***
income    2.897e-03  2.833e-04 10.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 12.09 on 100 degrees of freedom
Multiple R-Squared: 0.5111,    Adjusted R-squared: 0.5062
F-statistic: 104.5 on 1 and 100 DF,  p-value: < 2.2e-16
```

```r
> plm2 <- lm(prestige ~ income + education, data = Prestige)
> summary(plm2)
```

```
Call: lm(formula = prestige ~ income + education, data = Prestige)
Residuals:
   Min     1Q   Median     3Q    Max
-19.404  -5.330   0.015   4.980  17.689
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.848e+00  3.219e+00  -2.13   0.036 *
income    0.0013612  0.0002242   6.07  2.36e-08 ***
education  4.138e+00  0.349e+00  11.85  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7.81 on 99 degrees of freedom
Multiple R-Squared: 0.798,    Adjusted R-squared: 0.7939
F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

```r
> summary(prestigelm)
```

```
Call: lm(formula = prestige ~ income + education + women, data = Prestige)
Residuals:
   Min     1Q   Median     3Q    Max
-19.825  -5.333  -0.136   5.159  17.504
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.794e+00  3.239e+00  -2.09   0.038 *
income    0.0013136  0.0002778   4.72  7.58e-06 ***
education  4.187e+00  0.389e+00  10.77  < 2e-16 ***
women    -8.905e-03  3.041e-02  -0.29   0.770
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 17.51 on 98 degrees of freedom
Multiple R-Squared: 0.9, Adjusted R-squared: 0.899
F-statistic: 230.6 on 3 and 98 DF,  p-value: < 2.2e-16
```
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `. ' 0.1 ` ' 1
Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-Squared: 0.7982,   Adjusted R-squared: 0.792
F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16

library(lattice)
xyplot(prestige~income|equal.count(education),data=Prestige,panel=function(x,y,...)
{panel.xyplot(x,y,...)panel.lmline(x,y,...)})
• Another Regression Example: TouchScreen Voting & Bush Vote in Florida, 2004: Did hackers try to rig ‘04 elects?
  
  – **Variables:**

  **b04pc** “% Bush Vote in the County in 2004”
  **etouch** “Electronic touch screen voting used”
  **income** “median income of the county”

  ```r
  > election <- read.table("election.csv", sep = ",", header = TRUE)
  > electionfl <- election[election$fl == 1, ]
  > summary(electionfl[, c("b04pc", "etouch", "income")])
  ```

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>b04pc</td>
<td>0.2991</td>
<td>0.5390</td>
<td>0.5913</td>
<td>0.5999</td>
<td>0.6838</td>
<td>0.7824</td>
</tr>
<tr>
<td>etouch</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.2239</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>income</td>
<td>26032</td>
<td>30029</td>
<td>33779</td>
<td>35385</td>
<td>40249</td>
<td>52244</td>
</tr>
</tbody>
</table>

```r
> etouch.lm <- lm(b04pc ~ etouch, data = electionfl)
> medinc.lm <- lm(b04pc ~ income, data = electionfl)
> summary(etouch.lm)
```  
**Call:** lm(formula = b04pc ~ etouch, data = electionfl)

**Residuals:**

<table>
<thead>
<tr>
<th>Residuals:</th>
<th>Min.</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>-0.3152</td>
<td>-0.05572395</td>
<td>-0.00077549</td>
<td>0.08004172</td>
<td>0.1853497</td>
</tr>
</tbody>
</table>

**Coefficients:**

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|---------|
| (Intercept) | 0.61438 | 0.01459 | 42.123 | <2e-16 *** |
| etouch     | -0.06447 | 0.03083 | -2.091 | 0.0404 * |

---

**Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `. 0.1 ` ' 1**

**Residual standard error:** 0.1052 on 65 degrees of freedom
**Multiple R-Squared:** 0.06305,
**Adjusted R-squared:** 0.04864
**F-statistic:** 4.374 on 1 and 65 DF, p-value: 0.0404
> summary(medinc.lm)
Call: lm(formula = b04pc ~ income, data = electionfl)

Residuals:
       Min        1Q   Median        3Q       Max
-0.30549 -0.05858 -0.01026  0.07833  0.18875

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.350e-01  7.565e-02   8.394  5.82e-12 ***
income     -9.917e-07  2.105e-06  -0.471    0.639
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `. ' 0.1`` 1

Residual standard error: 0.1085 on 65 degrees of freedom
Multiple R-Squared: 0.003404,   Adjusted R-squared: -0.01193
F-statistic: 0.222 on 1 and 65 DF,  p-value: 0.6391

> par(mfrow = c(1, 2), pty = "s")
> plot(electionfl$etouch, electionfl$b04pc)
> abline(etouch.lm)
> plot(electionfl$income, electionfl$b04pc)
> abline(medinc.lm)

> par(mfrow = c(2, 2), pty = "s")
> plot(etouch.lm)
> par(mfrow = c(2, 2), pty = "s")
> plot(medinc.lm)
– *Interpretation?*

* Something rotten in state of FL...again. Touchscreens messing up...

* Our “findings”:
  · Bush significantly poorly in TouchScreen counties,
  · Even controlling for income.
  · Problem(s) with our study?

* Hout’s actual findings:
  · (See Olbermann @ http://www.msnbc.msn.com/id/6368819/)
  · In highly pro-Kerry counties, Bush overvote highly signif.
  · Still, not by enough to have reversed Florida this time.

* Example: Fertilizer & Rain

  – Fertilizer applied 100 to 1st plot, 200 next, etc. as go down hill.
  – Rainfall runs downhill, so yield may relate to rain, not fert.
  – Fertilizer reg line: \( a=36.42857143, b=0.05892857 \). Interp?
  – But fert also pos corr’d rain; fert not cause rain, rain not cause fert, but 3rd, unobserved factor here [altitude]: *spurious relationship.*

![Graphs of Fertilizer vs Yield and Rain vs Fertilizer](image-url)

  – So, check relationship b/w fert & yield due to rainfall (kind of proxy for alt). Well, b’s diff by rain (.05,.04,.03).
So, want rel b/w fert & yield “holding constant” rainfall (altitude).
Could avg these 3 lines:

Could redress issue by ANOVA (2-way). That would tell var in yield due to rain v. fert, & test null that yield indep. of each.
– Multiple regression another way; gives more info: also effect of each variable holding others constant. Write it thus:

\[
\begin{align*}
E(Y_i) &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \\
y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i
\end{align*}
\]

– This fits a plane to points, rather than line:

\[
\begin{align*}
S(b_0, b_1, b_2) &= \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2
\end{align*}
\]

– So, slope relating fert to yield “controlling for rain”, & v.v. from “shadows” of plane on fert & rain axes.

– So among assumpts re: \( \varepsilon \) (iid), need that \( \varepsilon \) unrelated to any \( X \).
  * Fixed \( X \) assures that regarding included \( X \), but...
  * if rndm \( X \) or if something omitted changes way \( X \) relate to \( Y \), as in this problem,
    * then condition violated & OLS biased.

– So, w/ > 1\( X \), minimize SSE:
− Some calculus then gave 3 eq w/ 3 unknowns:

\[ \sum Y_i = nb_0 + b_1 \sum X_{1i} + b_2 \sum X_{2i} \]
\[ \sum Y_iX_{1i} = b_0 \sum X_{1i} + b_1 \sum X_{1i}^2 + b_2 \sum X_{2i}X_{1i} \]
\[ \sum Y_iX_{2i} = b_0 \sum X_{2i} + b_1 \sum X_{1i}X_{2i} + b_2 \sum X_{2i}^2 \]

− Which we solved for:

\[ b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \]
\[ b_1 = \frac{(\sum Y_iX_{1i})(\sum X_{2i}^2) - (\sum Y_iX_{2i})(\sum X_{1i}X_{2i})}{(\sum X_{1i}^2)(\sum X_{2i}^2) - (\sum X_{1i}X_{2i})^2} \]
\[ b_2 = \frac{(\sum Y_iX_{2i})(\sum X_{1i}^2) - (\sum Y_iX_{1i})(\sum X_{1i}X_{2i})}{(\sum X_{1i}^2)(\sum X_{2i}^2) - (\sum X_{1i}X_{2i})^2} \]

− Same denominator. Symmetric. Adjusted from bivariate essentially by corr of this X w/ others times of others w/ Y.

− For inference:

\[ \beta = b \pm t_{df,0.025} SE, \text{ where } df = n-k-1 \text{ & } k=\# \text{ params (excl. intercept)} \]

**Problems 13-5, 13-6, 13-7, p. 405**

− Midterm cycle: presidents’ party usu. loses cong seats in off-yr.

− Use avg cong vote past 8 elects as base ⇒ stdzd vote loss, \( Y \).

− \( Y \) depends on 2 big things: \( X_1 \), % approving of pres & \( X_2 \), changes over previous yr in real disp inc per cap.

<table>
<thead>
<tr>
<th>year</th>
<th>( y ) approval</th>
<th>income change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1946</td>
<td>0.073</td>
<td>0.32</td>
</tr>
<tr>
<td>1950</td>
<td>0.020</td>
<td>0.43</td>
</tr>
<tr>
<td>1954</td>
<td>0.023</td>
<td>0.65</td>
</tr>
<tr>
<td>1958</td>
<td>0.059</td>
<td>0.56</td>
</tr>
<tr>
<td>1962</td>
<td>-0.008</td>
<td>0.67</td>
</tr>
<tr>
<td>1966</td>
<td>0.017</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Fall 2005 – Rob Franzese
- Tuft estimated this (y as %, i.e. $Y \times 100$):

$$\hat{Y} = 10.9 - .13X_1 - .034X_2$$

* a. On $(X_1, Y)$ plane, graph 6 pts & tag @ w/ its $X_2$ val. Then plot predicted lines for $X_2 = (100, 50, 0, -50)$.

![Graph showing 6 points and predicted lines]

* b. If $X_2$ kept constant, est change in $Y$ for unit change in $X_1$.
   
   \[ \hat{Y} = \text{coef for } X_1 = - .13 \text{ % pts} \]

* c. If $X_1$ kept constant, est change in $Y$ for unit change in $X_2$
   
   \[ \hat{Y} = b_2 = - .034 \text{ % pts} \]

* d. Est vote loss $Y$ for midterm elect when $X_1 = 60\%$ & $X_2 = +$50.

$$\hat{Y} = 10.9 - .13(60) - .034(50) = 10.9 - 7.8 - 1.7 = 1.4 \text{ % pts}$$

- 13-6

* a. From graph, find expected 1946 vote loss ($Y$) given that $X_1 = 32\%$ & $X_2 = -$40.

$$\hat{Y} = 10.9 - .13(32) - .034(-40) = 12.22$$

* What’s error, residual @ this pt?

$$Y - \hat{Y} = 7.3 - 12.22 = -4.92$$
Just before 1970 midterm elect, Nixon’s rating was $X_1 = 56\%$ & change real inc was $X_2 = $70. Predicted vote loss?

\[ \hat{Y} = 10.9 - .13(56) - .034(70) = 1.24 \]


13-7 In est eq 13-5, suppose know a priori that $Y$ no relation to $X_2$. So $\beta_2 = 0$. What’s $b_1$?

**Once More in Matrix Notation:**

- The Notation Defined:

\[
\begin{bmatrix}
    y \\
    \vdots \\
    y_n
\end{bmatrix}
= 
\begin{bmatrix}
    x_0 & x_{11} & x_{12} & \cdots & x_{1k} \\
    x_0 & x_{21} & \cdots & \vdots  \\
    \vdots & \vdots & \ddots & \vdots \\
    x_0 & x_{n1} & \cdots & x_{nk}
\end{bmatrix}
\begin{bmatrix}
    \beta_0 \\
    \vdots \\
    \beta_k
\end{bmatrix}
+ 
\begin{bmatrix}
    e_0 \\
    \vdots \\
    e_n
\end{bmatrix}
\]

- Least-Squares: Minimize Sum of Squared Errors

\[
\min_b SSE \equiv (Y - Xb)'(Y - Xb)
\]

\[
\min_b SSE = (Y'Y - 2b'X'Y - b'X'Xb)
\]

\[
\Rightarrow \frac{\partial SSE}{\partial b} = -2X'Y + 2X'Xb = 0
\]

\[
\Rightarrow X'Y = X'Xb \Rightarrow 
\]

\[
\hat{\beta}_{OLS} = (X'X)^{-1}X'Y
\]

- Implementing it w/ a smpl (small) dataset:

\[
\begin{bmatrix}
    0 \\
    0 \\
    1 \\
    1 \\
    3
\end{bmatrix}
= 
\begin{bmatrix}
    1 & -2 \\
    1 & -1 \\
    1 & 0 \\
    1 & 1 \\
    1 & 2
\end{bmatrix}
\]
− The $X'X$ part (V-Cov of $X$):

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} (1+1+1+1+1) & (-2 - 1 + 9 + 1 + 2) \\ (-2 - 1 + 0 + 1 + 2) & (4 + 1 + 1 + 4) \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}$$

− The $X'y$ part (Cov of $X$ & $y$):

$$X'y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} (0 + 0 + 1 + 1 + 3) \\ (0 + 0 + 0 + 1 + 6) \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

− The $(X'X)^{-1}$ Part:

$$(X'X)^{-1} = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{bmatrix}$$

− Put It All Together To Get the $\hat{\beta}$ Part:

$$\hat{\beta} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{bmatrix} \begin{bmatrix} 5 \\ 7 \end{bmatrix} = \begin{bmatrix} \frac{1}{5} \cdot 5 + 0 \\ 0 + \frac{1}{10} \cdot 7 \end{bmatrix} = \begin{bmatrix} 1 \\ .7 \end{bmatrix}$$

− Thus: $\hat{y} = 1 + .7x$.
− More Generally:

$$X'X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \\ \vdots & \vdots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$
\[
X'y = \left[ \frac{\sum y_i}{\sum x_i y_i} \right]
\]

\[
(X'X)^{-1} = \begin{bmatrix}
\frac{\sum x_i^2}{n} & -\frac{\sum x_i}{n}^2 + n\left(\frac{\sum x_i^2}{n}\right) \\
-\frac{\sum x_i}{n}^2 + n\left(\frac{\sum x_i^2}{n}\right) & \frac{n}{n}^2 + n\left(\frac{\sum x_i^2}{n}\right)
\end{bmatrix}
\]

\[
(X'X)^{-1}X'y = \begin{bmatrix}
-\left(\frac{\sum x_i^2}{n}\right)\left(\frac{\sum y_i}{n}\right) + \frac{\sum x_i}{n}\left(\frac{\sum x_i y_i}{n}\right) \\
\frac{n}{n}^2 - n\left(\frac{\sum x_i^2}{n}\right) \end{bmatrix}
\]

\[
= \hat{\beta}
\]

- **Accounting for Variance: “Goodness of Fit”**
  
  - Goodness of Fit:
    
    * Degree scatterplot pts clustered tightly around reg line (plane/hyperpl.).
    * Share var dep-var (lin-add’ly) accounted by \( X \) (i.e., ANOVA).
    * Like SE/t-stat, but relates to tightness/explan power of whole set \( X \), not just 1 \( x \) ind’ly.
  
  - ANOVA (Variance Accounting):
    
    \[
    (Y - \bar{Y}) = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})
    \]
    
    * Deviation from Mean=Predict Deviation from Mean+Deviation from Predict
    
    \[
    \sum_{i=1}^{n} (y_i - \bar{y}) = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (y_i - \hat{y}_i)
    \]
Total Sum Deviations from Mean = Sum Devs Predicts from Mean + Sum Devs Totals from Predicts

\[ \sum_{i=1}^{n} (Y - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2 + \sum_{i=1}^{n} (Y - \hat{Y})^2 \]

Total Sum Squares (Total Variation) = Sum Squared Predict Devs Mean (Explained Variation) + Sum Squared Devs Predicts (Residual/Unexplained Variation).

I.e., TSS = ESS + RSS (or TSS = RSS + ESS or SST = SSE + SSR or SST = SSR + SSE)

N.b., “Explained” means due to/lin’ly accounted by explan vars.

So, can make ANOVA table & its F-stats from this:

\[ F = \frac{\text{variance explained by the regression}}{\text{variance not explained}} \]

So, to test Y & X related, many roughly or perf’ly equiv ways: ANOVA F, t-test \( \hat{\beta}_x \), conting table; F test nice b/c works for X.

Linear-Regression Measures of “Goodness of Fit”

\( R^2 \): share var(Y) explained (lin-add’ly acc’d) by X

\[ R^2 = \frac{\text{SS Explained by all regressors}}{\text{total SS}} = 1 - \frac{\text{SS Unexplained}}{\text{total SS}} \]

Note: weakly increasing in \( k \), # regressors. [Why?] Adjusted \( R^2 \) or \( \bar{R}^2 \) adjusts for this, penalizing for degrees of freedom: \( 1 - \frac{\text{Unexpl SS}/(n-k-1)}{\text{Total SS}/(n-1)} \)

Other ways to write it:

\[ R^2 = \frac{(\hat{y} - \bar{y})'(\hat{y} - \bar{y})}{(y - \bar{y})'(y - \bar{y})} \]

\[ R^2 = 1 - \frac{e'e}{(y - \bar{y})'(y - \bar{y})} \]

\[ \bar{R}^2 = 1 - \frac{e'e}{n - k - 1 \left[ \frac{(y - \bar{y})'(y - \bar{y})}{n - 1} \right]^{-1}} \]
\[ R^2 = \frac{\text{Var} (\hat{y}_i)}{\text{Var} (y_i)} = \frac{\hat{\beta}^2 \text{Var} (x)}{\hat{\beta}^2 \text{Var} (x) + \sigma^2} \]

\[ = \frac{\text{(causal strength)}^2 \times \text{Var} (x)}{\text{(causal strength)}^2 \text{Var} (x) + \text{(lack of fit)}} \]

- Not try choose reg model that maxes \( R^2 \), in part b/c can always ↑ by add more \( X \), or just put LHS on RHS, but pt was (usu.) to est \( \partial y / \partial x \), not predict \( y \), &, even so, to predict \( y \) want fit only systematic part of \( y \) in your sample, not random (stoch., i.e. \( \varepsilon \)) part too! So, \( R^2 \) is est’d not max’d; a true \( R^2 \) (share of dep var lin-add’ly syst in \( X \)) exists. Still, can compare \( R^2 \) in same smpl of diff models to see if can explain sig’ly more

* **Standard Error of the Regression (S.E.R.)** or **Standard Error of the Estimate (S.E.E.)**; i.e., \( \hat{\sigma}^2 \)
- So, seen it. Roughly: how far avg obs from reg line.
- BLUE (or BANC) estimate of true std dev of residuals, & so of dist from which obs drawn.

\[ \hat{\sigma}^2_{OLS} \equiv \frac{\mathbf{e}'\mathbf{e}}{n - k - 1} \]

\[ \hat{\sigma}^2_{MLE} \equiv \frac{\mathbf{e}'\mathbf{e}}{n} \]

- In units of dep var. So can easily talk about & summarize errors substantively meaningfully.

- **F-Test of Joint Hypotheses**
  - Suppose wanted to test whether 2+ var both/all had coeff=0, or, equivalently, whether 2+ var’s added explan power to model?

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \]

* How diff from t-test on each var individually:
  - Set of t-tests: \( H_0 : \beta_1 = 0 \), and, then, \( H_0 : \beta_2 = 0 \)
· Has alternatives: \( H_A : Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \), and \( H_A : Y = \beta_0 + \beta_1 X_2 + \beta_3 X_3 + \varepsilon \), respectively.

· Whereas joint null is: \( H_0 : \beta_1 = 0 \) and \( \beta_2 = 0 \) at same time; i.e., \( H_0 : \beta_1 = \beta_2 = 0 \). So alternative is: \( Y = \beta_0 + \beta_3 X_3 + \varepsilon \)

* OK, so different; why might want joint hypoth? Suppose vars grp by thry, \( X_{soc} \), \( X_{pol} \), \( X_{eco} \), & want argue all matter (or not):

\[
Y = \beta_0 + X_{soc}\beta_{soc} + X_{pol}\beta_{pol} + X_{eco}\beta_{eco} + \varepsilon
\]

* OK, so how test? Claim \( H_0 : \beta_{soc} = 0 \) equiv “\( H_0 : Y = \beta_0 + X_{pol}\beta_{pol} + X_{eco}\beta_{eco} + \varepsilon \)” sacrifices no explanatory power relative to \( H_A : Y = \beta_0 + X_{pol}\beta_{pol} + X_{eco}\beta_{eco} + \varepsilon \), for example.

* \( R^2 \equiv ? \) Right; so, if null true, then \( R^2 \) not ↑ significantly from \( H_0 \) to \( H_A \) model. Hmm, so if only knew how \( R^2 \) dist’d across rep’d samps under null, could base test on \( \Delta R^2 \)…somehow.

\[- R^2 \equiv 1 - \frac{e'e}{(y-y')(y-y')}
\]

* So that’s like a sum of squared normals or asym. normals, which \( \chi^2 \) divided by something.

* So, \( R^2 \) from \( H_0 \) model divided by \( R^2 \) from \( H_A \) model will have denom’s cancel, and leave one \( \chi^2 \) divided by another.

* So happens that \( (\chi^2_n \text{ divided by } n)/(\chi^2_m \text{ divided } m) \) dist’d F, w/ numerator \( n \) & denominator \( m \) degrees free, so...

\[
\frac{\Delta R^2}{\Delta(df)} \sim F_{\Delta k,n-k_{big}}
\]

\[
\frac{1-R^2_{big}}{n-k_{big}} \sim F_{\Delta k,n-k_{big}}
\]

– So, can test joint hypotheses by comparing \( R^2 \) from bigger model to \( R^2 \) from smaller and asking if explain significantly more. Significantly will be big F in this test.

– Example:

```r
> plm1<-lm(prestige~income,data=Prestige)
> summary(plm1)
```
Call: `lm(formula = prestige ~ income, data = Prestige)`:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 2.714e+01 | 2.268e+00 | 11.97 | <2e-16 *** |
| income | 2.897e-03 | 2.833e-04 | 10.22 | <2e-16 *** |
---
100 degrees of freedom; Multiple R-Squared: 0.5111
F-statistic: 104.5 on 1 and 100 DF, p-value: < 2.2e-16

Call: `lm(formula = prestige ~ income + education, data = Prestige)`:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -6.8477787 | 3.2189771 | -2.127 | 0.0359 * |
| income | 0.0013612 | 0.0002242 | 6.071 | 2.36e-08 *** |
| education | 4.1374444 | 0.3489120 | 11.858 | < 2e-16 *** |
---
99 degrees of freedom; Multiple R-Squared: 0.7980
F-statistic: 195.6 on 2 and 99 DF, p-value: < 2.2e-16

Call: `lm(formula = prestige ~ income + education + women, data=Prestige)`:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -6.7943342 | 3.2390886 | -2.098 | 0.0385 * |
| income | 0.0013136 | 0.0002778 | 4.729 | 7.58e-06 *** |
| education | 4.1866373 | 0.3887013 | 10.771 | < 2e-16 *** |
| women | -0.0089052 | 0.0304071 | -0.293 | 0.7702 |
---
98 degrees of freedom; Multiple R-Squared: 0.7982
F-statistic: 129.2 on 3 and 98 DF, p-value: < 2.2e-16

\[
H_0 : \beta_{edu} = \beta_{women} = 0
\]

\[
\Rightarrow F_{2,98} = \frac{0.7982 - 0.5111}{2} \times \frac{1 - 0.7982}{98} = \frac{0.14355}{0.002059} = 69.71 \Rightarrow p < 5 \times 10^{-13}
\]

(Multi)Collinearity & Micronumerosity

- If explan vars lin’ly related (correlated): “multicollinearity”
  - If perf corr, can’t get coeff ests b/c they’re undefined.
  - If corr’d but not perf, SEs “blow up” & “strange things happen”. W&W: “the estimating procedure is very unstable; it becomes very sensitive to random errors.” However, no violation G-M, so BLUE. Std Errs rightly big. Sensitivity sensible. Practical prob, but not much can do.
• Ways see prob:
  – We’ll need \((\mathbf{X}'\mathbf{X})^{-1}\). If any cols \(\mathbf{X}\) linear functions any other(s), can’t invert.
  – Also, can write \(V(\hat{\beta}_j)\):
    \[
    \text{Var}(\beta_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)\sum x_j^2}
    \]
    – where \(R_j^2 = R^2\) from reg \(x_j\) on all other \(X’\)s in model. So, if can perf’ly pred \(x_j\) using \(x_{\sim j}\), \(R_j^2 = 1\), so s.e. \(\hat{\beta}_j = \infty\). As \(R_j^2 \uparrow\), s.e. explodes.
  – See how works for 2 explan vars:
    \[
    b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2
    \]
    \[
    b_1 = \frac{(\sum Y_i X_{1i})(\sum X_{2i}^2) - (\sum Y_i X_{2i})(\sum X_{1i} X_{2i})}{(\sum X_{1i}^2)(\sum X_{2i}^2) - (\sum X_{1i} X_{2i})^2}
    \]
    \[
    b_2 = \frac{(\sum Y_i X_{2i}) (\sum X_{1i}^2) - (\sum Y_i X_{1i})(\sum X_{1i} X_{2i})}{(\sum X_{1i}^2)(\sum X_{2i}^2) - (\sum X_{1i} X_{2i})^2}
    \]
    \[
    \text{Var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)\sum x_j^2}
    \]
    * Now, remember corr 2 vars (mean=0):
      \[
      \text{Cor}(X_1, X_2) = \frac{\sum X_{1i}X_{2i}}{\sqrt{(\sum X_{1i}^2)(\sum X_{2i}^2)}} = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}
      \]
    * So, can rewrite formula for reg coeff:
      \[
      b_2 = \frac{\text{Cor}(Y_i, X_{2i}) - \text{Cor}(X_{1i}, X_{2i})\text{Cor}(Y_i, X_{1i})}{(1 - \text{Cor}(X_{1i}, X_{2i})^2)} \left(\frac{\sqrt{\text{Var}(Y_i)}}{\sqrt{\text{Var}(X_2)}}\right)
      \]
    * So, if corr \(X_1 \& X_2 = 1\), then coeff undefined.
    * Say in words & see why: “Effect of x1 holding x2 constant...,” but you have no info on that if corr perf’ly.
– So: Perf multi-C means can’t est coeffs at all:

```r
> summary(lm(y ~ x))
Call: lm(formula = y ~ x)

Residuals:
  1       2       3       4
4.920e-16 -5.000e-01  5.000e-01  1.238e-16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.50000   0.35937  12.522  0.00632 **
x           0.25000   0.03442   7.263  0.01843 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5 on 2 degrees of freedom
Multiple R-Squared: 0.9635, Adjusted R-squared: 0.9452
F-statistic: 52.75 on 1 and 2 DF,  p-value: 0.01843
```

```r
> newx <- cbind(1, x, x + 1)
> newx
   x
[1,] 1 2 3
[2,] 1 4 5
[3,] 1 4 5
[4,] 1 20 21

> newxTx <- t(newx) %*% newx
> try(solve(newxTx))
Error in solve.default(newxTx) : system is computationally singular:
  reciprocal condition number = 1.04056e-18
```

– If not perf multi-C, but nearly perf:

```r
> newx <- cbind(1, x, x + runif(x, 0, 0.5))
> newx
   x
[1,] 1 2 2.385615
[2,] 1 4 4.495137
[3,] 1 4 4.241912
[4,] 1 20 20.277707

> cor(newx[, 2:3])
   x
x 1.000000 0.999923
  0.999923 1.000000

> newxTx <- t(newx) %*% newx
> solve(newxTx)
```
> summary(lm(y ~ newx - 1))
Call: lm(formula = y ~ newx - 1)

Residuals:
1     2     3     4
0.013738 -0.008039 -0.007416 0.001717

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
newx  6.05411   0.04081 148.34 0.00429 **
newxx 4.17360   0.09794  42.61 0.01494 *
newx  -3.94660   0.09851 -40.06 0.01589 *

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01764 on 1 degrees of freedom
Multiple R-Squared:  1,  Adjusted R-squared:  1
F-statistic: 1.887e+05 on 3 and 1 DF,  p-value: 0.001692

– Now, small change in data & slopes change dramatically:

> newx[1, 3] <- newx[1, 3] + 0.1
> summary(lm(y ~ newx - 1))
Call: lm(formula = y ~ newx - 1)

Residuals:
1     2     3     4
0.20908 -0.20491 -0.03030 0.02613

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
newx  5.9288   0.6906   8.585 0.0738 .
newxx 3.4820  1.4867   2.342 0.2569
newx  -3.2595  1.4992  -2.174 0.2744

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2955 on 1 degrees of freedom
Multiple R-Squared: 0.9995,  Adjusted R-squared: 0.998
F-statistic: 672.6 on 3 and 1 DF,  p-value: 0.02834

– In gen., multi-C becomes big prob once, say, >95% of var in 1 explan var accounted by lin combo.
– Magnitude of prob, therefore, fnctn of $R^2_j$, spec’ly var explodes w/
\[\frac{1}{1-R_j^2}\] (aka “Variance Inflation Factor” or “VIF”):

- (see Fox and Gujarati for nice disc multi-C / micro-numerosity)

• How occur?Usu. just not realize put mult versions same var in models, as we’ll see in dummy-variables & interaction-terms discussion coming.]

• How detect? Easy:
  - If reg F-test rejects, but t-tests all low, may be a problem; but no need test; just measure:
  - Check corr matrix of regressors, but can mask combo relations.
  - Just reg each on others to get \( R_j^2 \) or VIF. Some software (e.g, Stata) automates.

• What to do about (near) multi-C? Not Much:
  - (1) Add assumptions/theory
  - (2) Combine vars that essentially same, e.g. \( y = \beta_0 + \beta_1(X_1 + X_2) \)
  - (3) Get more data: prob is diff disting effects one var & another, so need more info.
  - Note: more theory/assumpts or more data pretty much universal reaction to big s.e./lack info prob.
Nonlinear & “Extra-Linear” Specifications

• Indicator Variables (Dummies)
  – Preliminaries
    * DEF: Binary (0,1) variable.
    * What assume re: $X$ in C(N)LRM?
      · Fixed (not nec.) & full column rank.
      · Nothing re: what type var X’s—contin., discrete, how dist’d...
      · So, statistical issues w/ using dummies RHS? None. Interpretation & practical issues only.
    * What assume re: $y$?
      · Independent, constant variance, (aym) normal.
      · If dummy dep var, latter 2 somewhat problematic, but...
      · ...not need norm or homosked (or indep) for unbiased & asym normal OK, eventually, but...
      · ...$E(y) = X\beta$, i.e., linearity & practical & interpretation issues really important & these stat ones somewhat too, so...
      · ...see section at end for intro binary dep var.
  – Suppose want know whether age related to pol partic, but age only coded “young”, “middle aged”, and “old”.
    * What need assume just to est $y=a+bAGE+e$ where Age=$(0,1,2)$?
    * What if not want make those assumptions?
      · Suppose create new vars: YNG=$(1$ if $Z=0,0$ else), MID=$(1$ if $Z=1,0$ else), OLD=$(1$ if $Z=2,0$ else)
      · How use? What assumpts, what sorts age-part rel’s poss now?
      · So one use dums to allow arbitrary, “step-fnctn” rel’s $X$ to $y$
    * Now say just “young” & “old”, & let $X \equiv$education level.
      \[ Y_i = \beta_0 + \beta_1 EDU_i + \beta_2 D_i + e_i \]
      · where D is “young” or “old”. [Why or?] [Matter which?]
· What effect EDU? Age? How plot \( \hat{Y} \) v. EDU look?
· So another use/interp=intercept shift.

* Now let \( Y \equiv \text{Income}, \; X \equiv \text{EDU}, \; D \equiv \text{Gender} \; (0: \text{M}, \; 1: \text{F}) \)
  · If Reg \( Y \) on Constant only, \( Y_i = \beta_0 + \varepsilon \), what \( \hat{\beta}_0 \)?
  · So, if Reg \( Y \) on \( D \) only in female smpl only, what est coeff?
  · So, if Reg \( Y \) on \( D \) & \((1 - D)\) only, what est coeffs, \( \hat{\beta}_f \) & \( \hat{\beta}_m \)?
  · Hmm, so H-test \( \hat{\beta}_f = \hat{\beta}_m \) is test what?
  · How do that test? \( \frac{(\hat{\beta}_f - \hat{\beta}_m) - 0}{SE(\hat{\beta}_f - \hat{\beta}_m)} \sim t_{n-k} \).
  · \( SE(\hat{\beta}_f - \hat{\beta}_m) = [V(\hat{\beta}_f - \hat{\beta}_m)]^{0.5} = [V(\hat{\beta}_f) + V(\hat{\beta}_m) - 2C(\hat{\beta}_f, \hat{\beta}_m)]^{0.5} \).
  · Or, easier, Reg \( Y \) on constant \& \((D \text{ or } (1 - D))\). Then \( E(Y|D = 0) = \beta_0, \; E(Y|D = 1) = \beta_0 + \beta_1 \), so \( \hat{\beta}_1 \) is difference of means, so t-test \( H_0 : \hat{\beta}_1 = 0 \) is diff means test!
  · So, another use/interp=conditional means.

- Example: Education, Gender, Income. Hypothetical Data:
  * Say expect more edu \( \Rightarrow \) more income, so Reg INC on EDU:

\[
IN_{C_i} = \beta_0 + \beta_1 EDU_i + \varepsilon_i
\]

· Consider data in left plot. Overall reg line (dashed) positive. i’s gender indicated by \( \circ \) & box.
· Here, slopes same for men & women, \& EDU not corr’d w/ gender (how tell?), but intercepts diff.
· As can see, such case not bias reg slope to ignore diff’s, but, by not accounting diffs intercept, leave it in (est’d) error, so SE’s bigger than need be (inefficiency).

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 5.500000 | 0.418933   | 13.128  | 1.171871e-10 |
| education2          | -0.090991| 0.126313   | -0.7197| 4.809436e-01 |

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 1.5      | 0.414578   | 3.618136| 6.800947e-03 |
* If EDU & gender corr’d, e.g., if women higher EDU on avg than men as in right, but incomes diff’d by fix amount, e.g., men’s higher as shown, * then overall reg line (dashed) shows slight neg rel INC & EDU, even if slopes still same for 2 grps: * Omitted-Variable (gender) b/c rel b/w what omitted (gender) & in-
cluded (edu), so biased coeff est.

* How fix?
  · Separate-sample reg? Probs: (1) lower df so wider SE (ineff); (2) how H-test diffs b/w regs?; (3) slopes same: ineff to est 2 params (sep slopes) when 1 would do. (1 & 3 same, actually)
  · Dummy intercept-shift? More efficient & can test easily.

* In this case, estimate:

```
  Estimate Std. Error t value Pr(>|t|)
  (Intercept) 1.5  0.3091735 4.851645  1.496327e-04
  education   1.0  0.0857493 11.661904  1.556076e-09
  gender      2.0  0.2425356  8.246211  2.409830e-07
```

- Interpretation:
  * When GEN=0 [i.e. female], equation is:
    \[ I\hat{N}C_i = (1.5 + 0) + 1 \times EDU_i \]
  * When GEN=1 [i.e. male], equation is:
    \[ I\hat{N}C_i = (1.5 + 8) + 1 \times EDU_i = 9.5 + 1 \times EDU_i \]
  * So coeff on dummy for gender gives the constant vertical diff b/w the 2 parallel lines; i.e., conditional mean; i.e., intercept shift.

- So, in general, for 1 dummy var:
  \[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + e_i \]
  \[ Y_i = \begin{cases} 
    \beta_0 + \beta_1 X_i, & \text{if } D_i = 0 \\
    (\beta_0 + \beta_2) + \beta_1 X_i, & \text{if } D_i = 1 
  \end{cases} \]

- Notes:
  * Not matter whether women=0 or men=0 for GEN. Coefs change (\( \beta_0 \) cond mean men or women, \( \beta_0 + \beta_1 \) for other), but effects (either way \( \beta_1 = \text{diff}; \hat{\beta}_0 \) in one model (say, cond mean women) will exactly equal \( \hat{\beta}_0 + \hat{\beta}_1 \) in other (cond mean women).
  * Choice D & (1-D) versus constant & (D or (1-D)) also makes no substantive diff in effects, math’ly identical; only changes interp; so use most convenient &/or presentationally effective.
More than 2 categories?

* Same idea, but use multiple dummies.
* Age example from above: concerned that some qualitative diffs b/w young, mid-aged, old.
* If these qual diffs not modify effects of X, but just shift \( E(Y) \uparrow \downarrow \), can do w/ 2 dummies ⇒
  
  \[
  Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \varepsilon_i
  \]
  
  where \( D_{1i} = 1 \) if young, & 0 o/w,
  
  and \( D_{2i} = 1 \) if old & 0 o/w.

\[
Y_i = \begin{cases} 
\beta_0 + \beta_1 X_i, & \text{if } D_{1i} = 0 \text{ and } D_{2i} = 0 \\
(\beta_0 + \beta_2) + \beta_1 X_i, & \text{if } D_{1i} = 1 \text{ and } D_{2i} = 0 \\
(\beta_0 + \beta_3) + \beta_1 X_i, & \text{if } D_{1i} = 0 \text{ and } D_{2i} = 1 
\end{cases}
\]

Can’t have \( D_{1i} = D_{2i} = 1 \), right?

So, \( \beta_0 \) = intercept for MID, \( \beta_2 \) is fixed diff b/w lines (intercept-diffs) b/w MID & YNG lines, & \( \beta_3 \) sym’ly diff MID & OLD.

**Interaction Terms**

- What if diff slopes? I.e., effects edu on inc diff by age or gen
  
  * (1) Could est separately but... (same probs as before, plus all slopes (coeffs) diff by age or by edu, & maybe not want that.
  
  * (2) Use age-eduction or gender-education interaction terms.

- (Multiplicative) Interaction Terms (mult vars together) allow specify how effects one var depend on levels other & v.v.

- E.g., if think effect \( X \) (say, EDU) on \( Y \) (INC) vary by, depend upon, are moderated (not mediated) by \( D \), gender, then could model:

  \[
  Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i D_i) + e_i
  \]

  \[
  Y_i = \begin{cases} 
\beta_0 + \beta_1 X_i, & \text{if } D_{1i} = 0 \\
(\beta_0 + \beta_2) + \beta_1 X_i + \beta_3 X_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i, & \text{if } D_{1i} = 1
\end{cases}
\]
- Easiest & surest way to interp, as always, $\partial y / \partial x$ (for any type var):
  \[
  \begin{align*}
  \frac{\partial Y}{\partial X} &= \beta_1 + \beta_3 D \\
  \frac{\partial Y}{\partial D} &= \beta_2 + \beta_3 X
  \end{align*}
  \]
- Note: works same regardless what sort variable X & or D.


* Variables:
  - **sex**: 'F', female; 'M', male.
  - **weight**: measured weight in kg.
  - **height**: measured height in cm.
  - **repwt**: reported weight in kg.
  - **repht**: reported height in cm.

```
<table>
<thead>
<tr>
<th>sex</th>
<th>weight</th>
<th>height</th>
<th>repwt</th>
<th>repht</th>
</tr>
</thead>
<tbody>
<tr>
<td>F:112</td>
<td>Min. : 39.0</td>
<td>Min. : 57.0</td>
<td>Min. : 41.00</td>
<td>Min. :148.0</td>
</tr>
<tr>
<td>M: 88</td>
<td>1st Qu.: 55.0</td>
<td>1st Qu.:164.0</td>
<td>1st Qu.: 55.00</td>
<td>1st Qu.:160.5</td>
</tr>
<tr>
<td></td>
<td>Median : 63.0</td>
<td>Median :169.5</td>
<td>Median : 63.00</td>
<td>Median :168.0</td>
</tr>
<tr>
<td></td>
<td>Mean : 65.8</td>
<td>Mean :170.0</td>
<td>Mean : 65.62</td>
<td>Mean :168.5</td>
</tr>
<tr>
<td></td>
<td>3rd Qu.: 74.0</td>
<td>3rd Qu.:177.2</td>
<td>3rd Qu.: 73.50</td>
<td>3rd Qu.:175.0</td>
</tr>
<tr>
<td></td>
<td>Max. :166.0</td>
<td>Max. :197.0</td>
<td>Max. :124.00</td>
<td>Max. :200.0</td>
</tr>
<tr>
<td></td>
<td>NA's :17.00</td>
<td>NA's :17.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

****with M=1 dummy********repwt=depvar****

```
Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 41.3227620 2.16785082 19.061626 5.933960e-45
weight 0.2644593 0.03631928 7.281515 1.007941e-11
sexM -39.9641221 3.92932151 -10.170744 1.823304e-19
weight*sexM 0.7253627 0.05598043 12.957435 1.631447e-27
```

****with F=1 dummy****

```
Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 1.3586399 3.27719248 0.4145743 6.789499e-01
weight 0.9898221 0.04259951 23.2355263 6.242642e-56
sexF 39.9641221 3.92932151 10.1707437 1.823304e-19
weight*sexF -0.7253627 0.05598043 -12.9574351 1.631447e-27
```

* What does reg suggest about rel wt & repwt in men v. women?
  - Intercept is repwt among women @ 0kg.
  - So, @ smpl min wt 39kg, women rep 41+.26*39=51kgs;
  - For @ kg wt among women, rep .26, about 1/4 of actual wt;
So, @ max (166kg), predicted repwt=85kg.
Far from 1-to-1, & over-report (?) @ low, underreport @ high.
Men: @ min wt (39kg), 41+.26*39-39+.072*39=40. For @ +1kg wt, rep 0.264459+0.725362=.989.
- Here’s plot of this interactive regression...

- Woops, wait a sec, what’s up w/ that obs (#12)?
  * Ht & wt for that woman switched in data entry!!
  * So, try again w/o obs 12:

| Estimate  | Std. Error  | t value | Pr(>|t|) |
|-----------|-------------|---------|----------|
| (Intercept) | 1.35863993  | 1.58142305 | 0.8591249 | 3.914271e-01 |
| weight    | 0.98982207  | 0.02055657 | 48.1511201 | 5.204720e-104 |
| sexF      | 1.98928622  | 2.45693465 | 0.8096618  | 4.192156e-01   |
| weight*sexF | -0.05670878 | 0.03855672 | -1.4707885 | 1.431139e-01   |

- Oh, never mind... Whew; good thing not publish that...

• Standard Errors on Linear Combinations of Terms
  - Already know how! Say, interaction model like:

$$ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon_i $$
– Often most interested in effect $X_1$ on $Y$, as fnctn of $X_2$ & v.v.:
  
  $\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2$

  $\frac{\partial Y}{\partial X_2} = \beta_2 + \beta_3 X_1$

– Very likely might want to plot these effect lines, or calc & report these conditional effects @ several values vars.

– But then, immediately, you want to see some std errs, conf intervals, h-tests & p-levels, not just point ests, so:

$$SE\left( \frac{\partial Y}{\partial X_2} \right) = \sqrt{V(\hat{\beta}_2 + \hat{\beta}_3 X_1)} = \sqrt{V(\hat{\beta}_2) + X_1^2 V(\hat{\beta}_3) + 2X_1C(\hat{\beta}_2, \hat{\beta}_3)}$$

– So, just need get est’d variances coeffs (square of s.e.’s, i.e. diagonal element of “vce”) and their est’d covar’s (off-diags “vce”).

– Note: Effect differs at val’s other var; so do s.e.’s & so c.i.’s & t-stats & p-lvls

• **Dependent- &/or Independent-Variable Transformations**

  – **Rescaling: Linear Rescaling**

    * Sometimes convenient rescale (some/all) $X$ to comm; e.g., 0-1.
    * Renders coefficients directly comparable in a sense. [Elab.]
    * How: $\frac{X - \min X}{\max X - \min X}$
    * Notes:

      - Changes meaning $\beta$ (effect unit-change rescaled $X$=effect range-sized change $X$), & so s.e. changes too.
      - Each $X$ contains exactly same info as before, so not change t- or $R^2$ (unless rescaled Y too) or S.E.R. (then it rescaled equiv’ly).

  – **Rescaling: Standardized Regression**

    * Sometimes convenient rescale $X$ relative to own sample SD.
    * Renders coefficients directly comparable in a sense. [Elab.]
    * Notes:
Changes meaning $\beta$ (effect unit-change rescaled $X$=effect S.D.-sized change $X$), & so s.e. changes too.

Here, usu. would rescale $Y$ also, to get $\beta$=effect SD change $X$ in terms of SD of $Y$.

Sometimes, therefore, relative size stdzd $\beta$ considered relative “importance” of effects; but problematic as such in that conflates info on slope & on $X$-variance in sample.

Each $X$ contains exactly same info as before, so not change t-; num & denom $R^2$ changed same way, so no change either; but S.E.R. definitely refer to something new, so diff.

---

**Linear & Stdzd Rescaling Examples:**

**Raw Data:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2.387312</td>
<td>12.48777</td>
<td>0.191172</td>
<td>0.8503</td>
</tr>
<tr>
<td>PSUPGPW</td>
<td>0.495912</td>
<td>0.235925</td>
<td>2.101990</td>
<td>0.0484</td>
</tr>
<tr>
<td>NPGOVVPW</td>
<td>-2.847637</td>
<td>2.237494</td>
<td>-1.272690</td>
<td>0.2177</td>
</tr>
</tbody>
</table>

R-squared 0.186944 Mean dependent var 25.24783
Adjusted R-squared 0.105638 S.D. dependent var 9.655468
S.E. of regression 9.131244 Akaike info criterion 7.382389
Sum squared resid 1667.592 Schwarz criterion 7.530497
Log likelihood -81.89747 F-statistic 2.299272
Durbin-Watson stat 1.887032 Prob(F-statistic) 0.126243

Rescaling vars in example to 0=min, 1=max:

genr dgov01=(dgovpw-@min(dgovpw))/(@max(dgovpw)-@min(dgovpw))
genr psupg01=(psupgpw-@min(psupgpw))/(@max(psupgpw)-@min(psupgpw))
genr npgov01=(npgovpw-@min(npgovpw))/(@max(npgovpw)-@min(npgovpw))

show dgovpw dgov01 psupgpw psupg01 npgovpw npgov01

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Maximum</th>
<th>Minimum</th>
<th>S.D.</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGOVPW</td>
<td>25.24783</td>
<td>26.60000</td>
<td>45.10000</td>
<td>11.00000</td>
<td>9.655468</td>
<td>0.158789</td>
</tr>
<tr>
<td>DGOV01</td>
<td>0.417825</td>
<td>0.457478</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.283152</td>
<td>0.158789</td>
</tr>
<tr>
<td>PSUPGPW</td>
<td>57.38261</td>
<td>57.20000</td>
<td>80.40000</td>
<td>41.10000</td>
<td>9.103878</td>
<td>0.821165</td>
</tr>
<tr>
<td>PSUPG01</td>
<td>0.414316</td>
<td>0.409669</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.231651</td>
<td>0.821165</td>
</tr>
<tr>
<td>NPGOVVPW</td>
<td>1.965217</td>
<td>1.800000</td>
<td>4.300000</td>
<td>1.000000</td>
<td>0.959928</td>
<td>0.907747</td>
</tr>
<tr>
<td>NPGOV01</td>
<td>0.292490</td>
<td>0.242424</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.290887</td>
<td>0.907747</td>
</tr>
</tbody>
</table>
Kurtosis 2.421182 2.421182 3.698725 3.698725 2.929475 2.929475
Jarque-Bera 0.417725 0.417725 3.052737 3.052737 3.163451 3.163451
Probability 0.811507 0.811507 0.217323 0.217323 0.205620 0.205620
Sum 580.7000 9.609971 1319.800 9.529262 45.20000 6.727273
Sum Sq. Dev. 2051.017 1.763846 1823.373 1.180566 20.27217 1.861540

ls dgovpw c psupg01 npgov01

Dependent Variable: DGOVPW Method: Least Squares
Date: 12/06/05 Time: 00:49
Sample: 1 23 Included observations: 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>19.92167</td>
<td>4.004635</td>
<td>4.974653</td>
<td>0.0001</td>
</tr>
<tr>
<td>PSUPG01</td>
<td>19.48935</td>
<td>9.271860</td>
<td>2.101990</td>
<td>0.0484</td>
</tr>
<tr>
<td>NPGOV01</td>
<td>-9.397202</td>
<td>7.383731</td>
<td>-1.272690</td>
<td>0.2177</td>
</tr>
</tbody>
</table>

R-squared 0.186944 Mean dependent var 25.24783
Adjusted R-squared 0.105638 S.D. dependent var 9.655468
S.E. of regression 9.131244 Akaike info criterion 7.382389
Sum squared resid 1667.592 Schwarz criterion 7.530497
Log likelihood -81.89747 F-statistic 2.299272
Durbin-Watson stat 1.887032 Prob(F-statistic) 0.126243

Stdzd Rescaling of Vars:

gern dgovsd=(dgovpw-@mean(dgovpw))/(@stdev(dgovpw))
gern psupgsd=(psupgpw-@mean(psupgpw))/(@stdev(psupgpw))
gern npgovsd=(npgovpw-@mean(npgovpw))/@stdev(npgovpw)

show dgovsd psupgsd npgovsd

<table>
<thead>
<tr>
<th></th>
<th>DGOVSD</th>
<th>PSUPGSD</th>
<th>NPGOVSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.000000</td>
<td>1.20E-16</td>
<td>1.93E-17</td>
</tr>
<tr>
<td>Median</td>
<td>0.140042</td>
<td>-0.020058</td>
<td>-0.172114</td>
</tr>
<tr>
<td>Maximum</td>
<td>2.056055</td>
<td>2.528306</td>
<td>2.432247</td>
</tr>
<tr>
<td>Minimum</td>
<td>-1.475623</td>
<td>-1.788535</td>
<td>-1.005510</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.158789</td>
<td>0.821165</td>
<td>0.907747</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.421182</td>
<td>3.698725</td>
<td>2.929475</td>
</tr>
</tbody>
</table>

Jarque-Bera 0.417725 3.052737 3.163451
Probability 0.811507 0.217323 0.205620
Sum 0.000000 2.38E-15 4.44E-16
Sum Sq. Dev. 22.00000 22.00000 22.00000
Dependent Variable: DGOVSD  
Method: Least Squares  
Date: 12/06/05  
Time: 00:55  
Sample: 1 23  
Included observations: 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-5.53E-17</td>
<td>0.197194</td>
<td>-2.80E-16</td>
<td>1.0000</td>
</tr>
<tr>
<td>PSUPGSDF</td>
<td>0.467582</td>
<td>0.222447</td>
<td>2.101990</td>
<td>0.0484</td>
</tr>
<tr>
<td>NPGOVSD</td>
<td>-0.283107</td>
<td>0.222447</td>
<td>-1.272690</td>
<td>0.2177</td>
</tr>
</tbody>
</table>

R-squared 0.186944  
Mean dependent var 0.000000  
Adjusted R-squared 0.105638  
S.D. dependent var 1.000000  
S.E. of regression 0.945707  
Akaike info criterion 2.847340  
Schwarz criterion 2.995448  
F-statistic 2.299272  
Prob(F-statistic) 0.126243

Example how stdzd reg coeffs can mislead:

```r
> x <- c(2, 4, 4)
> y <- c(5, 5, 6)
> reg1 <- lm(y ~ x)
> coef(reg1)

(Intercept)     x
4.5000  0.2500

> coef(reg1)[2] * sd(x)/sd(y)

x
0.5
```

Now, new obs exactly same line but X far from current range (x=20), so 4.50+.25*(20)=9.5. Unstdzd coeff not change (i.e., slope same), but stdzd co-efficient does — even though new pt not change slope!

```r
> x[4] <- 20
> reg2 <- lm(y ~ x)
> coef(reg2)

(Intercept)     x
4.5000  0.2500

> coef(reg2)[2] * sd(x)/sd(y)

x
0.9815652
```

- **Transforming Indep &/or Dep Vars for Nonlinearity:**

  * Lin Reg more flex than sounds: lin in params, not nec’ly in vars.
    - \( \hat{Y} = \beta_0 + \beta_1 X^2 \) or \( \ln(\hat{Y}) = \beta_0 + \beta_1 X \) are OK;
    - \( \hat{Y} = \beta_0 + \beta_1 X_1 + \ln(\beta_1 + \beta_2 X_2) \) NOT (at least not by OLS)
<table>
<thead>
<tr>
<th>Country</th>
<th>PR</th>
<th>DM</th>
<th>STS</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>94</td>
<td>1.0</td>
<td>435</td>
</tr>
<tr>
<td>JA</td>
<td>91</td>
<td>3.9</td>
<td>512</td>
</tr>
<tr>
<td>GE</td>
<td>98</td>
<td>260</td>
<td>520</td>
</tr>
<tr>
<td>FR</td>
<td>79</td>
<td>1.0</td>
<td>577</td>
</tr>
<tr>
<td>IT</td>
<td>95</td>
<td>20.3</td>
<td>630</td>
</tr>
<tr>
<td>UK</td>
<td>85</td>
<td>1.0</td>
<td>650</td>
</tr>
<tr>
<td>CA</td>
<td>88</td>
<td>1.0</td>
<td>195</td>
</tr>
<tr>
<td>AU</td>
<td>98</td>
<td>20.3</td>
<td>183</td>
</tr>
<tr>
<td>BE</td>
<td>91</td>
<td>7.1</td>
<td>212</td>
</tr>
<tr>
<td>DE</td>
<td>99</td>
<td>9.7</td>
<td>179</td>
</tr>
<tr>
<td>FI</td>
<td>98</td>
<td>13.3</td>
<td>200</td>
</tr>
<tr>
<td>GR</td>
<td>88</td>
<td>5.4</td>
<td>300</td>
</tr>
<tr>
<td>IC</td>
<td>97</td>
<td>6.7</td>
<td>60</td>
</tr>
<tr>
<td>IR</td>
<td>96</td>
<td>4.0</td>
<td>166</td>
</tr>
<tr>
<td>LU</td>
<td>90</td>
<td>14.8</td>
<td>60</td>
</tr>
<tr>
<td>NE</td>
<td>96</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>NO</td>
<td>91</td>
<td>8.2</td>
<td>165</td>
</tr>
<tr>
<td>PO</td>
<td>94</td>
<td>12.5</td>
<td>250</td>
</tr>
<tr>
<td>SP</td>
<td>83</td>
<td>6.7</td>
<td>350</td>
</tr>
<tr>
<td>SW</td>
<td>98</td>
<td>12.0</td>
<td>349</td>
</tr>
<tr>
<td>SZ</td>
<td>96</td>
<td>7.7</td>
<td>200</td>
</tr>
<tr>
<td>AL</td>
<td>87</td>
<td>1.0</td>
<td>148</td>
</tr>
<tr>
<td>NZ</td>
<td>80</td>
<td>1.0</td>
<td>97</td>
</tr>
<tr>
<td>Variable</td>
<td>Coefficient</td>
<td>Std. Error</td>
<td>t-Statistic</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>C</td>
<td>91.00292</td>
<td>1.310502</td>
<td>69.44128</td>
</tr>
<tr>
<td>L MAG</td>
<td>0.033297</td>
<td>0.020716</td>
<td>1.607365</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.109552</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean dep. var</td>
<td>91.82609</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adj. R-sq</td>
<td>0.067149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.D. dep. var</td>
<td>5.989780</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.E. reg.</td>
<td>5.785182</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Akaike</td>
<td>6.431418</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum sq. resid</td>
<td>702.8348</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schwarz</td>
<td>6.530156</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log like.</td>
<td>-71.96131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-stat.</td>
<td>2.583622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Durbin-Watson stat</td>
<td>1.905516</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prob(F-stat.)</td>
<td>0.122907</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent Variable: LPROP Method: Least Squares Date: 12/06/05
Time: 04:53 Sample: 1 23 Included observations: 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>86.96900</td>
<td>1.589869</td>
<td>54.70199</td>
<td>0.0000</td>
</tr>
<tr>
<td>@LOG(LMAG)</td>
<td>2.568056</td>
<td>0.663349</td>
<td>3.871352</td>
<td>0.0009</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.416462</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean dep. var</td>
<td>91.82609</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adj. R-sq</td>
<td>0.388674</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.D. dep. var</td>
<td>5.989780</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.E. reg.</td>
<td>4.683249</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Akaike</td>
<td>6.008802</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum sq. resid</td>
<td>460.5892</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schwarz</td>
<td>6.107541</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log like.</td>
<td>-67.10123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-stat.</td>
<td>14.98736</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Durbin-Watson stat</td>
<td>1.742240</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prob(F-stat.)</td>
<td>0.000883</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent Variable: @LOG(LPROP) Method: Least Squares Date: 12/06/05
Time: 04:54 Sample: 1 23 Included observations: 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4.463754</td>
<td>0.017841</td>
<td>250.1987</td>
<td>0.0000</td>
</tr>
<tr>
<td>@LOG(LMAG)</td>
<td>0.028570</td>
<td>0.007444</td>
<td>3.838059</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Fall 2005 – Rob Franzese
Diagnostics & Sensitivity: Leverage & Influence

• Preliminaries
  – Since LS finds vector coeffs to min sum \textit{squared} resids, pts far from line can have \textit{↑↑} influence (prop to square of distance from reg line).
  – Influence on coeff ests=leverage \((x_i)\) distance from centroid \(\times\) discrepancy \((y_i)\) distance from reg line, excl. self).

• Hat Values — Measures of Leverage

\[ \hat{y} = X(X'X)^{-1}X'y = Hy \]

– where \(H \equiv "\text{Hat matrix}" = "\text{fitted-value maker}" \) since \(\hat{y} = X\hat{\beta} \) & \(\hat{\beta} = (X'X)^{-1}X'y\), so \(Hy = X(X'X)^{-1}X'y = X\hat{\beta} = \hat{y}\)

– Call diagonals of Hat matrix \(h_i\). They summarize leverage (potential inflience) that given obs may have on fitted value.

– In simple bivariate regression:

\[ h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \]

– So, basically \(h_i = \text{distance obs from }\bar{x}\), though tends ↓ as \(n \uparrow\).

– In multivar, \(h_i = \text{distance from centroid (pt @ vector means)} X\).

– So, high leverage pts=pts far from center data, \(\Rightarrow\text{↑↑potential influ.}\)

• DFBETA and DFBETAS — Measures of Influence

– Can gauge actual influence pt directly by est w/ v. w/o that pt:

\[ D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \text{ for } i = 1, \ldots, n \text{ and } j = 0, 1, \ldots, k. \]
Useful to scale $D_{ij}$ by std errs (so scales of coefs net out):

$$D^*_{ij} = \frac{D_{ij}}{\hat{SE}_{(-i)}(\hat{\beta}_j)}$$

$D_{ij}$ often called DFBETA$_{ij}$ & $D^*_{ij}$ DFBETAS$_{ij}$.

FYI: some matrix alg exists do this in 1 est w/o actually est n+1 times.

Measures sensitivity each $\beta$ to each obs.

**Cook’s Distance — Summary Measure of Influence**

$$D_i = \frac{(e_i^*)^2}{k + 1} \times \frac{h_i}{1 - h_i}$$

where $e_i^* \equiv \frac{e_i}{s\sqrt{1-h_i}} \equiv$ *(internally) standardized or studentized residuals*

* Externally studentized (uses $e_i$ excluding obs in Q) exists also.

* Useful to plot to check C(N)LRM assumpts, outliers.

Cooks D(istance) also expressible as, basically, sum squared $D^*_{ij}$, so summary in that sense.

• Example of why sensitivity & related explorations good idea.
In regression plotted, $h_{12} = .714$, but avg $\bar{h} = (3 + 1)/183 = .0219$. Avg $h$ in general is $\bar{h} = (k + 1)/n$. So clearly tons leverage.
- Now try reg omitting obs 12 (data-entry error).

| Estimate     | Std. Error | t value | Pr(>|t|) |
|--------------|------------|---------|----------|
| (Intercept)  | 1.35863993 | 1.58142305 | 0.8591249 | 3.914271e-01 |
| weight       | 0.98982207 | 0.02055657 | 48.1511201 | 5.204720e-104 |
| sexF         | 1.98928622 | 2.45693465 | 0.8096618 | 4.192156e-01 |
| weight:sexF  | -0.05670878 | 0.03855672 | -1.4707885 | 1.431139e-01 |

- Wow. Giant DFBETA(S) and Cook’s d.
- So, did removing obs 12 seem to fix matters?
– Much better.

**Binary Dependent-Variables: Logit & Probit**

• Suppose $Y \in 0, 1$. [Examples?] [Draw example plot]
  – What if remain w/ lin-reg?
    * Nonsense predictions: all but $(0,1)$, esp. $\leq 0$ & $\geq 1$
      • So, interpret as model underlying propensity/prob $Y = 1$,
      • and $\leq 0$ & $\geq 1$ just really, **really** (un)likely.
* Heteroskedasticity: \( Y \sim Bernoulli \Rightarrow E(p) = f(X\beta), \& V(Y) = [f(X\beta)][1 - f(X\beta)] \)
  
  · Nonconstant and, worse (699), het related to \( f(X) \).
  
  · But, not need homosked for unbiased, \& easy fixes for efficiency \& s.e. est exist.

* Non-normal \( \varepsilon \): Duh, Bernoulli; But only C(N)LRM \& CLT.

* Linearity: \( p \) linearly related to \( X \); just not plausible substantively \& easy, better exists.

  – So what do about it?

* Model \( E(p) = E(Prob Y=1) \) as some S-shaped (sigmoid) \( f(XB) \).
* Then likelihood:

\[
L = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} = \prod_{i=1}^{n} f(x_i\beta)^{y_i} [1 - f(x_i\beta)]^{1-y_i}
\]

\[
LL = \sum_{i=1}^{n} (y_i) \ln f(x_i\beta) + (1 - y_i) \ln [1 - f(x_i\beta)]
\]

* Two main workhorse sigmoid (link) functions \( x\beta \) to \( y \):
  
  · Logit: \( p = \frac{\exp(X\beta)}{1+\exp(X\beta)} \)
  
  · Probit: \( p = \Phi(X\beta) \) (i.e., cumulative std norm)

  · Work through values @ \( X\beta = 0, \infty, -\infty, \& \) shape.

* Effects:
  
  · NOT lin-add, so NOT equal coeff’s, although same sign, so t-tests of coeff=0 is also of effect=0.
  
  · Effects found by \( \frac{\partial p}{\partial x} \) or \( \frac{\Delta p}{\Delta x} \), as usual.
  
  · Logit: \( \frac{\partial p}{\partial x_j} = \hat{\beta}(1 - \hat{p}) \hat{\beta}_j = \frac{\exp(X\hat{\beta})}{1+\exp(X\hat{\beta})} (1 - \frac{\exp(X\hat{\beta})}{1+\exp(X\hat{\beta})}) \hat{\beta}_j \)

  · Probit: \( \frac{\partial p}{\partial x_j} = \phi(X\hat{\beta}) \beta_j \)

* Discuss