PS699: Statistical Methods II

Quick Review of Fundamental Probability & Distributions, and Statistics, Estimation, & Inference

I. Basic Probability Review:

A. *Preamble*:

1. We can view the outcome of every aspect of the physical & social world as the product of some "random" "experiment".

   a) "Random" here means that prior to the outcome, some degree (from zero to infinite) uncertainty exists about what the outcome will be

      (1) Now, zero & infinite uncertainty both, in some sense, strain the meaning of "uncertainty". We are usually uninterested in either case:

         (a) Zero, because there is nothing to explain; it just is. Infinite uncertainty is even little hard to grasp & probably not empirically relevant in vast majority of cases. We usually have some idea of range of possible outcomes if nothing else.

         (b) Because of this, most texts exclude 0 and ∞ uncertainty from definition of "random".

      (2) Do remember, though, that something happens with probability 0 or with probability 1, either of which correspond to 0 uncertainty, is actually possible.

2. Another key term: ***Data-Generating Function (DGF)*** – the mechanism that produces these random outcomes.

   a) Usually, in social science, DGF is outside analysts' control.

   b) One view of the object of our empirical-estimation exercises: to estimate the parameters of the (ideally, substantively & theoretically determined) DGF.

B. *Probability Function*:

So, let A represent the outcome of some random experiment, as in:
$A \equiv$ some "event", a possible outcome of the experiment
then let $S \equiv$ set of all possible outcomes, called the Sample Space
and let $Pr(A) \equiv$ some function of A defined over the Sample Space

⇒ Axiomatic Definition of a Probability Function (P.F.):

(from Set Theory:)

① $\emptyset \leq Pr(A) \leq 1 \quad \forall A \in S$
- any event among permiss. has prob. b/w 0 & 1

② $Pr(S) = 1$
- some event among permiss. must happen

③ $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$ $\forall (A, B) \in S$
- this defines how we manipulate probabilities mathematically

Notes:

on ① → Notice that "nothing happens" is always an element of S b/c $\emptyset$ (null set) always a subset of every other set, including any sample space.

on ③ → if A & B disjoint, that is A & B have "nothing in common" or "no overlap" then

$$Pr(A \cup B) = Pr(A) + Pr(B)$$

C. *Random Variables*: assigning numeric values to outcomes probability functions gives definitions discrete and continuous probability functions

Now suppose for every possible outcome (ie, $\forall A \in S$) we assign a numeric value, e.g. ① on a {roll of a die} there are {six possible faces that might come up}, we could assign to each {up face} the {numeric value on that face}

(random experiment; sample space; event or outcome; random variable)

② on a {coin toss} there are two possibilities, {heads or tails}, for each face that may come up, {H} or {T}, we assign a {value H=1, T=∅}

(random experiment; sample space; event; random variable)

Random Variables: are mathematically defined thusly $X \equiv \begin{cases} 1 & \text{if coin heads up} \\ \emptyset & \text{if coin tails up} \end{cases}$

1. If outcomes *finite* or *countably infinite*, then R.V. is ***discrete***.

define probability functions over (discrete) random variables axiomatically:

$$f(x) \equiv Pr(X = x)$$

probability function (pf); random variable; some placeholder for the specific values that X may take

Axioms:

① $\emptyset \leq Pr(X = x) \leq 1$ or $\emptyset \leq f(x) \leq 1$

② $\sum_{x_i \in S} Pr(X = x_i) = 1$ or $\sum_{x_i \in S} f(x_i) = 1$

notice by these axioms that random variables are taken to be defined over disjoint events. That is, for any outcome only one value is assigned to X & for any value assigned to X only one outcome corresponds. (There is a "one-to-one" correspondence, or, identically, a function.)

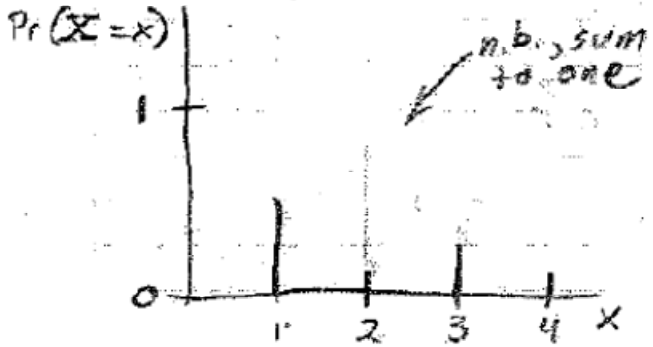
2. If outcomes neither *finite* nor *countably infinite*, then they are *infinitely divisible*, & the RV is ***continuous***. ⇒ no way to sum probabilities of values ⇒ integrate, and the axiomatic definition of a ***probability density function*** (*pdf*) becomes:

① $\Pr(a \le X \le b) = \int_a^b f(x)\,dx \ge \emptyset$

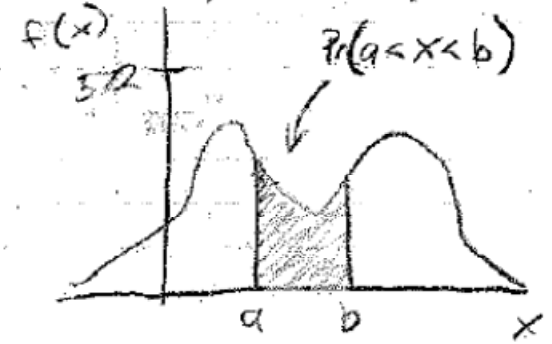
①a $\Pr(X = a) = \int_a^a f(x)\,dx = \emptyset$ (by definition of integrals)

② $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Probability Function



Probability Density Function



- n.b. f(x) not Pr(X=x)
- n.b. not nec. 0-1 on y axis
- n.b. can be any shape at all
- n.b. unbroken line (usu., not nec.)

Notes: a) perfectly analogous to discrete case, except that probability at any specific point in the sample space is zero.

b) Thus $f(x)$ is not a probability function (pf) but i.e. a probability density function (pdf) — something else, we call it...

c) Since $Pr(X=a) = \emptyset$, $Pr(a \le X \le b) = Pr(a < X \le b)$
$= Pr(a \le X < b) = Pr(a < X < b)$

d) Recall that integration is "scooping" up the area under a curve. So, to find the probability of some range of possible outcomes, say from a to b, we "scoop" the area under the pdf, $f(x)$, from a to b, i.e. we integrate from a to b.

D. *Examples*:

1. Discrete:

① Bernoulli (Binary Experiment) e.g. coin toss; unemployed/not; war/not war; individ. vote/not; gov't run deficit/not; child is girl/boy } often, but not always, an $X \sim$ not X dichotomy

- distributions have pf's or pdf's these have argument(s) they're called parameter(s) of the distribution. ← what you need to know about the spec. example

Bernoulli: $f(x) = Pr(X = x)$ $\quad x \in \{0, 1\}$

Bernoulli's parameter $\equiv \pi$

$$f(x) \equiv Pr(X = x) = \begin{cases} \pi^x (1-\pi)^{1-x} & \text{for } x = 0 \text{ or } x = 1 \\ \emptyset & \text{for all else} \end{cases}$$

First Principles of Bernoulli (i.e., use it when what you're interested in follows or most closely follows these)

ⓐ only two possible outcomes
ⓑ outcomes mutually exclusive (disjoint)
<—> but you can always redefine outcomes so that this is true

② Discrete Uniform (equiprobable) Distribution

$$f(x) \equiv \Pr(X = x) = \begin{cases} \frac{1}{N} & \text{for } X = 1, 2, \ldots, N \\ \emptyset & \text{for all else} \end{cases}$$

parameter: N

eg. → coin toss
→ die roll
→ names drawn from hat (which may approx. some selection process)

First Principles ① N disjoint events
② all equally likely

(Not used in our branches of stats much; used in formal theory a lot.)

③ Binomial Distribution:

$$f(x) \equiv \Pr(X = x) = \begin{cases} \frac{N!}{x!(N-x)!} \pi^x (1-\pi)^{N-x} & \text{for } X = 0, 1, 2, \ldots, N \\ \emptyset & \text{for all else} \end{cases}$$

parameters: N, π

First Principles ① N Bernoulli trials
② Each trial has constant, same probability, π

(note: $N!/[x!(N\text{-}x)]!$ is "N choose x", written $\binom{N}{x}$)

(4) **Poisson Distribution**: Limit as the number of trials $\to\infty$ at same time $\pi\to 0$ (e.g., because time-slice $\to 0$) such that $n\pi$, which is the mean of a binomial, is stable, we get continuous-time version of Binomial, the limiting distribution of the Binomial, known as the Poisson:

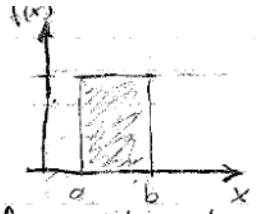
$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

First-Principles: It's a count distribution. Each event is independent. Probability of event within observation period is λ.

2. Continuous:

① Continuous Uniform Distribution [a, b]

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for all } a<x<b \\ \phi & \text{else} \end{cases}$$



parameters: a, b -- the range of possible values

First Principles: ① fixed range of possible values (bounded)

② nothing "more likely" than anything else in this range

(even less/more commonly used in stats/theory than discrete)

② Normal (Gaussian) Function:

$$f(x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

parameters: σ^2 (variance), μ (mean)

"the bell-shape"

- symmetric
- "thin-tailed"
- no intuitive 1st principles (near) regularity

for now, just empirical

a) You'll see this *pdf* written other ways as well; for example:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2}\right]} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{(x-\mu)}{\sigma}\right]^2} = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{\frac{-\left(x-\mu\right)^2}{2\sigma^2}\right\}$$

b) And, for reference, the ***standard-normal*** is the normal with mean 0 and s.d. 1, so:

$$x \sim N(0,1) \Leftrightarrow f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} = (2\pi)^{-.5} e^{-\frac{x^2}{2}}$$

3. A squared *standard-normal* is *Chi-Squared*, χ_1^2; sum of *n* <u>*independent*</u> std-norms squared $\sim \chi_n^2$, sum of *n* <u>*independent*</u> $\chi_{n_i}^2 \sim \chi_{\Sigma_i n_i}^2$:

1. Chi-Squared Distribution
   - if $Z \sim N(0,1)$, then, letting $X \equiv Z^2$, $X \sim \chi_1^2$
     ("X is distributed Chi-Squared with 1 degree of freedom")
   - $E(X)$ for any Chi-Squared r.v. X is it's degrees of freedom

Aside: for any $X \sim N(\mu, \sigma^2)$, $z = (x-\mu)/\sigma \sim N(0,1)$

- if $z_1, z_2, \ldots, z_n$ each dist. "independently" $N(0,1)$, then $\left(\sum_{i=1}^{n} z_i^2\right) \sim \chi_n^2$
- if $X_1, X_2, \ldots, X_N$ each dist "independently" $\chi_{n_i}^2$, then $\left(\sum_{i=1}^{N} X_i\right) \sim \chi_{\Sigma n_i}^2$

Example: if the e_i From $y_i = X_i b + e_i$ distributed "independently" normal with variance σ_e^2, then

$$\left(\sum_{i=1}^{n} \left(\frac{e_i}{\sigma_e}\right)^2\right) \sim \chi_n^2$$

4. Ratio of two χ^2, each divided by its *degrees of freedom* is *F* with those º free:

F Distribution: if X_1 & X_2 are "independent" chi-squared r.v.'s with n_1 & n_2 degrees of freedom respectively, then

$$F = \left(\frac{X_1/n_1}{X_2/n_2}\right) \sim F_{n_1,n_2}$$ → "F is distributed F with n_1 & n_2 degrees of freedom"

Example: if e_i from $y_i = X_i b + e_i$ "independent" Normal$(0, \sigma_e^2)$ } "independent" & "independent" from each other
and e_i^* from $y_i^* = X_i^* b + e_i^*$ "independent" Normal$(0, \sigma_{e^*}^2)$ }

Then $$\left\{\left[\sum_{i=1}^{n_1}(e_i/\sigma_e)^2\right]/n_1\right\} \Big/ \left\{\left[\sum_{i=1}^{n_2}(e_i^*/\sigma_{e^*})^2\right]/n_2\right\} \sim F_{n_1,n_2}$$

5. Standard normal, divided by square root of a χ^2 divided by its degrees of freedom is *t* with those deg's free:

t-Distribution (also called "student" t):

• if $Z \sim N(0,1)$ & $X \sim \chi^2_n$ is "independent" of Z, then

$$\left(t \equiv Z/\sqrt{X/n}\right) \sim t_n \qquad (\text{"t is distributed t with n deg. free."})$$

Example: if $\hat{b}$ from $y_i = x_i\hat{b} + e_i$ is $N(0,1)$ and e_i is "independent" of $\hat{b}$, & also Normal$(0,1)$

then $$\hat{b}/\sqrt{\Sigma e_i^2/n} \sim t_n$$

• it can be shown that $(t_n)^2 \sim F_{1,n}$ for any r.v. distributed t_n

6. Some useful relations & limiting distributions:

a) Square of a t_{n-k} is $F_{1,n-k}$: that is, $t^2_{n-k} \sim F_{1,n-k}$

b) Two other useful limits: $\lim_{n\to\infty} t_{n-k}$ is N(0,1); and $\lim_{n2\to\infty} (n_1 \times F_{n1,n2}) = \text{Chi-squared}[n_1]$

E. *Cumulative Distribution Functions*:

Cumulative Distribution Functions: (cdf)

$$F(x) \equiv \sum_{X \le x} f(x) \equiv Pr(X \le x) \quad \text{(discrete)}$$

$$\Rightarrow f(X_i) = F(X_i) - F(X_{i-1})$$

continuous:

$$F(x) \equiv Pr(X \le x) \equiv \int_{-\infty}^{x} f(t)\,dt$$

math note: since we want to integrate to X, we don't want to use x as the index (dx) by which we integrate, so we just use a different letter, say t

$$\Rightarrow \frac{\partial F(x)}{\partial x} = f(x)$$

(a good test: you understand integration & differentiation well enough for our purposes if you can see why this is so)

Further Implications:

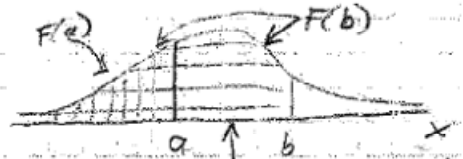
① $0 \le F(x) \le 1$ (by def. of prob. (dens.) function)

② if $x > y$, $F(x) \ge F(y)$ (b/c same def: x beyond y so "includes" it, $Pr(X) \ge \emptyset$)

③ $F(+\infty) = 1$ (should be obvious) new stuff

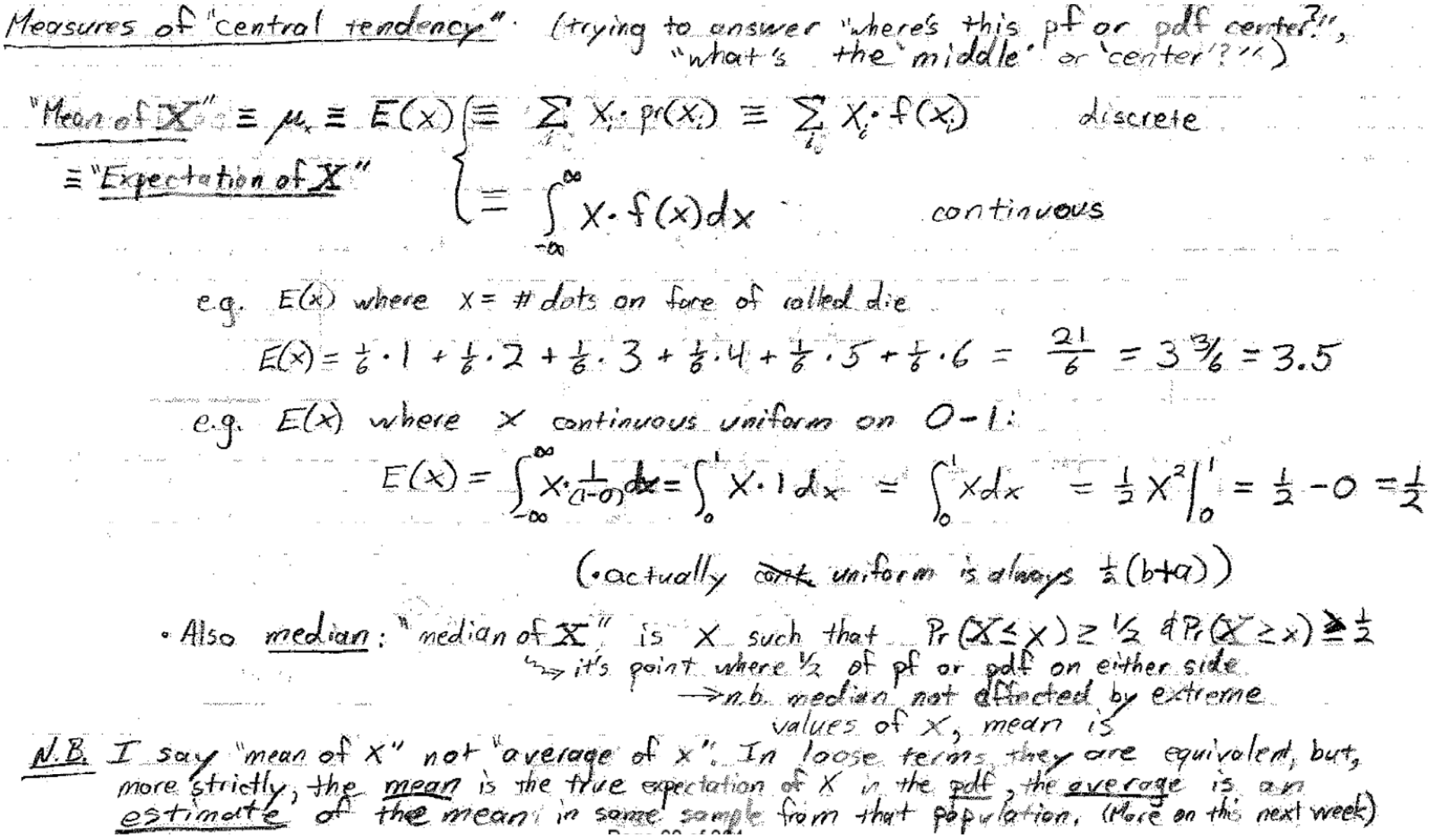
④ $F(-\infty) = \emptyset$ (this too)

d ⑤ $Pr(a \le x \le b) = F(b) - F(a)$

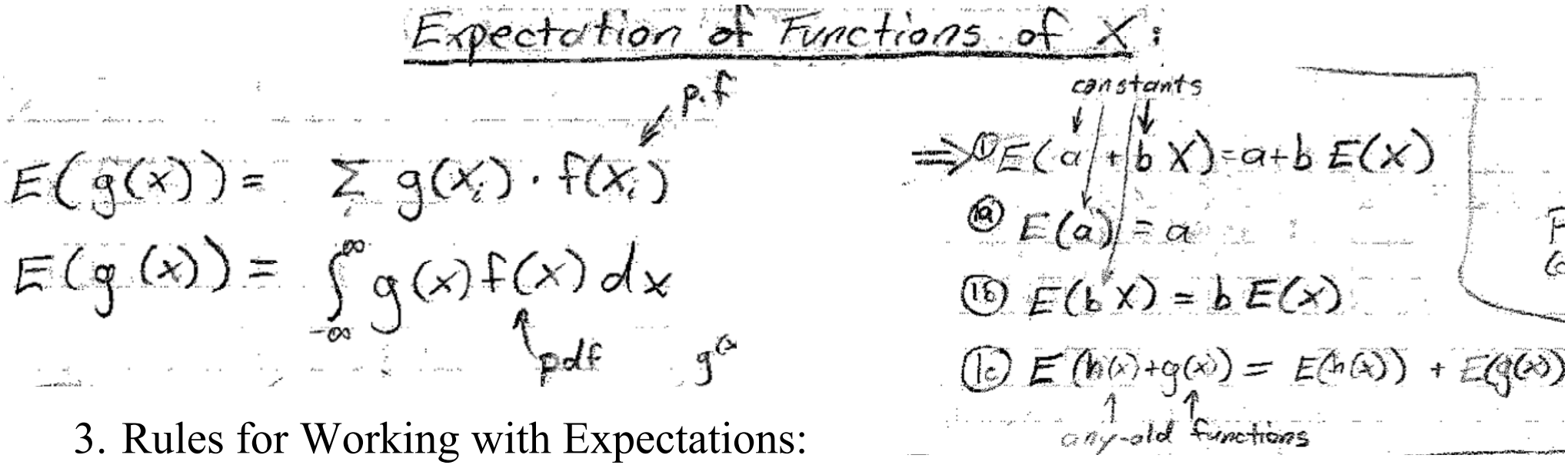


$F(b) - F(a) = \int_a^b f(x)\,dx$

F. **Expectations (Means)**:



1. *Median*: not so commonly used in our branches of stats; again, more in theory
2. *Mode*: "most common value". That x at which $f(x)$ maximized, i.e., Argmax $f(x)$.



3. Rules for Working with Expectations:

a) *Expectation of a sum is sum of the expectations*. (So, e.g., can move E in-&-out of ∑.)

b) Expectation operator *slides through constants, snagging on random variables*:

$$E(a + bX + cZ + dXZ) = a + bE(X) + cE(Z) + dE(XZ)$$

c) Note: *slides through constant coefficients, but cannot slide through product of 2 RV's*

4. How about a proof of this crucial property: $E(a+bX)=a+bE(X)$

Proof: (continuous)

$$E(a+bX) = \int_{-\infty}^{\infty} (a+bX) f(x)\,dx$$

($f(x)$: pdf; $g(x)$: pdf)

$$= \int_{-\infty}^{\infty} \left(a f(x) + bx f(x)\right) dx$$

can do this by integral rules

$$= \int_{-\infty}^{\infty} a f(x)\,dx + \int_{-\infty}^{\infty} bx f(x)\,dx$$

can pull constants out by integral rules

$$= a\int_{-\infty}^{\infty} f(x)\,dx + b\int_{-\infty}^{\infty} x f(x)\,dx$$

$\int_{-\infty}^{\infty} f(x)\,dx = 1$ by def of pdf

$\int_{-\infty}^{\infty} x f(x)\,dx = E(x)$ by def of $E(x)$

$$= a \cdot 1 + bE(x)$$

$$= a + bE(x) \quad \square$$

+ $E(g(x))$

5. **Variance**, crucial special case of $E(f(X))$:

VARIANCE: $V(x) \equiv \mathrm{Var}(x) \equiv \sigma_x^2 \equiv E[(x-\mu)^2] = \begin{cases} \sum (x-\mu)^2 \Pr(x) & \text{discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx & \text{continuous} \end{cases}$

a) Some preliminary properties:

$V(x) \equiv E[(x-\mu_x)^2] \equiv E(x^2 - 2x\mu_x + \mu_x^2) = E(x^2) - 2\mu_x E(x) + \mu_x^2$
$= E(x^2) - \mu_x^2$

① $V(x) = E(x^2) - \mu_x^2$
$= E(x^2) - [E(x)]^2$

constants

→ useful: makes calc variances

② $\mathrm{Var}(a + bX) = b^2 \mathrm{Var}(x)$

② $\mathrm{Var}(\text{any constant}) = 0$

③ $\mathrm{Var}(g(x))$
$= \int_{-\infty}^{\infty} (g(x) - E(g(x)))^2 f(x)\,dx$
$\approx [g'(\mu)]^2 \mathrm{Var}(x)$

b) $E(x^2) = \mu^2 + \sigma^2$

b) Note from (2): "constants neither vary nor *covary* [to be defined shortly…]"

c) Note from (3): use Taylor-series linear-approximation for V(nonlinear f(RV))…

6. **Mean Squared-Error**, another crucial special case of $E(f(X))$:

$$MSE_{(x \text{ about } c)} \equiv E[(x-c)^2] = \underbrace{\sigma_x^2}_{\text{variance of } X} + \underbrace{(c-\mu)^2}_{\left(\text{"bias" of } c \text{ relative to } \mu\right)^2}$$

7. While at it, 3rd & 4th ***moments***, **Skew** & **Kurtosis** also special cases $E(f(X))$:

Just some FYI definitions: skew: measure the symmetry of a distribution

$f_1(x)$ is skewed left (or negative)
$f_2(x)$ is skewed right (or positive)

→ as Greene says, skew is direction of longer tail

$$\text{skew} \equiv E[(X-\mu)^3]$$

kurtosis: measures "how fat the tails" are (n.b. this not necessarily a simple relationship to variance)

e.g. t is more ikurtotic than std normal.

$$\text{kurtosis} = E[(X-\mu)^4]$$

G. *Joint (Multivariate) Distributions*:

A. Suppose you have two R.V.'s, X_1 & X_2 (e.g. $X_1 = \begin{pmatrix} \text{female} \\ \text{or} \\ \text{male} \end{pmatrix}$ & $X_2 = \begin{pmatrix} \text{taller 6'} \\ \text{or} \\ \text{shorter 6'} \end{pmatrix}$ or $X_1 = \begin{pmatrix} \text{GDP} \\ \text{/capita} \end{pmatrix}$ & $X_2 = \begin{pmatrix} \text{Degree} \\ \text{of Democ.} \end{pmatrix}$

Define: Joint (Bivariate) Distribution of X_1 & X_2

$$\text{Prob}(a \le X_1 \le b, c \le X_2 \le d) = \begin{cases} \sum_{a \le X_1 \le b} \sum_{c \le X_2 \le d} f(X_1, X_2) & \text{discrete} \\ \int_a^b \int_c^d f(X_1, X_2)\, dX_2\, dX_1 & \text{continuous} \end{cases}$$

Aside on doing "double integrals" & "double sums": just do them one at a time. First, sum or integrate over one variable, treating the other as a constant. Then, sum or integrate the other. (This generalizes to "multiple" sums & integrals") Changing the order of integration or summation makes no difference.

Axiomatic Definition:

① $f(X_1, X_2) \ge 0$

② $\begin{cases} \sum_{X_1} \sum_{X_2} f(X_1, X_2) = 1 & \text{discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(X_1, X_2) = 1 & \text{continuous} \end{cases}$

Cumulative Distribution Function (joint c.d.f.)

$$F(X_1, X_2) \equiv \text{Prob}(X_1 \le a, X_2 \le b) \quad (\text{"," means "and"})$$

$$= \begin{cases} \sum_{X_1 \le a} \sum_{X_2 \le b} f(X_1, X_2) \\ \int_{-\infty}^{a} \int_{-\infty}^{b} f(X_1, X_2)\, dX_2\, dX_1 \end{cases}$$

1. *Marginal Distributions*:

Marginal Distributions: (i.e., getting the univariate distributions back out of the bivariate distribution)

$$f_{X_1}(x_1) = \begin{cases} \sum_{X_2} f(X_1, X_2) \\ \int_{-\infty}^{\infty} f(X_1, X_2)\, dX_2 \end{cases} \qquad \text{d.v.v.} \quad f(X_2) = \begin{cases} \sum_{X_1} f(X_1, X_2) \\ \int_{-\infty}^{\infty} f(X_1, X_2)\, dX_1 \end{cases}$$

Intuition: A Snow-Covered Driveway

Consider a driveway covered with snow:
→ think of the length of the driveway as the X_1 dimension (length dimension, x-axis) & the width of the driveway as the X_2 dimension (width, y-axis)

Then the height of the snow at any point on the driveway of as reflecting the "probability (density)" that any randomly given snowflake will have fallen at that (X_1, X_2) coordinate. In general, the height could vary from spot to spot of course.

First of all, we're defining snowflakes that fall on the driveway as the only relevant ones, so, by definition, the "probability density" at any point on the driveway is zero (no snow) or positive (some snow)

$$\Rightarrow f(X_1, X_2) \geq 0$$

Since we've defined the driveway as the whole universe of relevant snow (seems that way shoveling it), any relevant snowflake must have fallen on the driveway somewhere

$$\Rightarrow \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(X_1, X_2)\, dX_2\, dX_1 = 1$$

↳ or just integrate (shovel, snow-plow) over length & width of driveway

So, if we took a shovel or a plow & plowed all the snow down to a line at the end or side of the driveway (that's the first integral), then turned the plow & plowed all the snow on that line into a point in the corner (that's the second integral after having done the first), we would have all the snow on that point. In probability, we call all : one.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)\,dx_2\,dx_1 = 1$$

snow on driveway

plow all to one side

plow that side to one point

$\Rightarrow$ all the probability, or 1.

A discrete example:

s \ h	6' and over	Under 6'	sum of (height) columns
Male	18%	30%	48%
Female	10%	42%	52%
sum of (sex) rows	28%	72%	

in % of population = probability randomly chosen person will be in that category

$$\sum_{sex}\sum_{height} f(s,h) = \sum_{sex}(\text{sum of height})$$

$$= 48\% + 52\% = 100\%$$

$$= 1$$

n.b. order of summation irrelevant

Marginal Distributions: (i.e., getting the univariate distributions back out of the bivariate distribution)

$$f_{X_1}(x_1) = \begin{cases} \sum_{x_2} f(x_1, x_2) \\ \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \end{cases} \quad \text{& v.v.} \quad f(x_2) = \begin{cases} \sum_{x_1} f(x_1, x_2) \\ \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 \end{cases}$$

- back to the driveway analogy for intuition:
  in general, snow falls all over the driveway, (X_1, X_2) coordinates all over. Suppose you wanted to know the probability (or proportionate amount) of snow falling at X_2, not caring where it fell on relative to the X_1 axis! Say, proportionate amount in line with some tree next to the driveway. Well, get out your snow plow and plow the whole driveway (parallel to the street) so that it heaps up in a line along the side of the driveway (the line is perpendicular to the line from the tree of interest across the driveway). Wherever the snow was generally higher across the driveway will now be the highest point along the line you just plowed, won't it? (rhetorical) That height of that line of snow now represents the marginal (i.e. univariate) distribution of X_1 as it gives the "probability density" that X_1 coordinate was that point what ever X_2 was. (i.e., the "accumulated" hight of snow along the line at X_1)

Example: in the sex & height table above, height = $\begin{cases} \text{over six foot w/ } p = 28\% \\ \text{under six foot w/ } p = 72\% \end{cases}$ is the marginal distribution of height. It's obtained by summing rows, i.e. summing over sexes. ← marginal

2. *Expectations of RV's & Functions of RV's in Multivariate Distributions*:

Expectations From Joint p.d.f.'s & p.f.'s: from the definition just given, we can show a whole bunch of other useful things.

1) $$E[X_1] = \sum_{X_1} X_1 \left(f_{X_1}(X_1)\right)$$

$$= \sum_{X_1} X_1 \left(\sum_{X_2} f(X_1, X_2)\right)$$ • from the definition of the marginal dist., f_{X_1}

$$\boxed{= \sum_{X_1}\sum_{X_2} X_1 f(X_1, X_2)}$$ •• X_1 is not changed as we sum over X_2, i.e., relative to summing over X_2 it is a constant, thus we can move it in or out of summing over X_2 at will.

• The Continuous Analogue is

$$\boxed{E(X_1) = \int_{-\infty}^{\infty}\int X_1 f(X_1, X_2)\, dX_2\, dX_1}$$

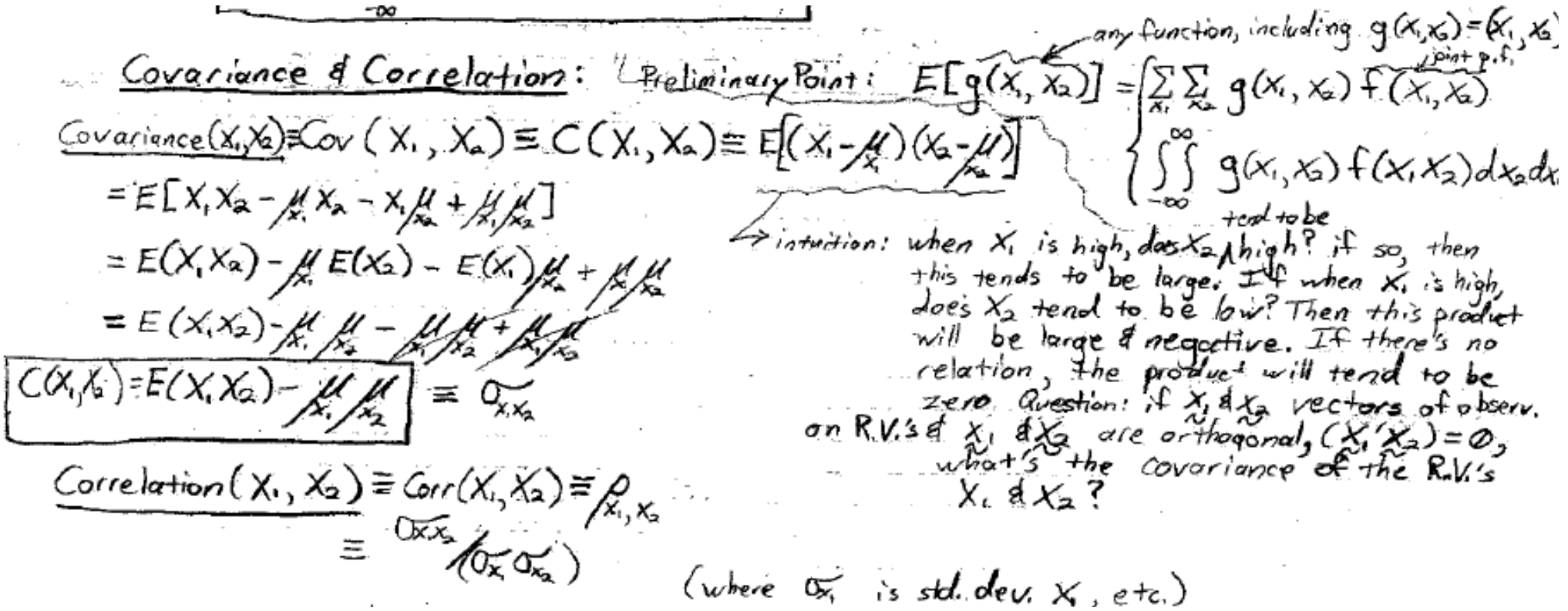
(notational note: I may from time to time write it this way if both (or all) integrals over same range. If I ever leave it off, or write $\int_{X_2} I$ mean $-\infty$ to ∞ (not indefinite integral) or over the range of X_2.)

2) $$\boxed{V(X_1) \equiv E\left[(X_1 - \mu_{X_1})^2\right]}$$

$$= \sum_{X_1} (X_1 - \mu_{X_1})^2 f_{X_1}(X_1)$$

$$\boxed{\begin{cases} = \sum_{X_1}\sum_{X_2} (X_1 - \mu_{X_1})^2 f(X_1, X_2) & \text{• logic for this exactly as above} \\ = \int_{-\infty}^{\infty}\int (X_1 - \mu_{X_1})^2 f(X_1, X_2)\, dX_2\, dX_1 & \text{• by analogy} \end{cases}}$$

3. *Covariance & Correlation* (special cases of $E(f(X))$ in multivariate dists):



a) Both covariance and correlation are measures of simple (i.e., linear) association, with larger positive (negative) values indicating greater positive (negative) association, smaller indicating smaller, and 0 indicating no (simple, linear) association.

b) Covariance, however, is sensitive to the scale of the variables; correlation nets those scales to produce a single measure $-1 \leq \rho \leq 1$, with -1 (+1) indicating perfect negative (positive) association and 0 indicating no association.

4. *Conditional Distributions & Expectations* (very special cases of $E(f(X))$…):

Conditional Distributions

$$f(X_2 \mid X_1) \equiv f(X_1, X_2) / f_{X_1}(X_1)$$

writing this $f(X_1, X_2) = f(X_2 \mid X_1) f_{X_1}(X_1)$ also has important uses.

(This also goes by "Bayes' Law")

(definition works for discrete or continuous; in fact also for set theory analog:

$P(A \mid B) = P(A \cap B) / P(B)$

also written $P(A \mid B) = P(AB) / P(B)$)

conditional distribution: "f of X_2, conditional on X_1"

"f of X_2 given X_1"

$\rightarrow \forall f_{X_1}(X_1) \neq 0$

- obvious since you can't be "given" impossible information (given X_1 when $p(X_1) = \emptyset$)

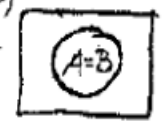
- So, within any joint distribution, $f(X_1, X_2)$ there are 2 marginal distributions, $f_{X_1}(X_1)$ & $f_{X_2}(X_2)$, (or n of them in multivariate), & two conditional distributions, $f(X_1 \mid X_2)$ & $f(X_2 \mid X_1)$, (or a whole boatload (I think $N(N!)$) in the multivariate case).

analogous

- $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\emptyset}{P(B)} = \emptyset$

"p of A given B" is obviously zero, so check!

Think about a) 

& b) 

in b) $P(A \text{ given } B) = 1$

check $= P(A \cap B)/P(B) = \frac{P(B)}{P(B)}$

in a) $\frac{P(A \cap B)}{P(B)}$ is the shaded as a fraction of B which reasons out well enough: once we're given that we're in B, we can think of B as the whole sample space →

Conditional Expectations

- We may very well be interested in the expectation of X_2 given X_1, or renaming the variables for obvious reasons, y given x.

$$E(y \mid x) = \begin{cases} \sum_{y} y \cdot f(y \mid x) \\ \int_{-\infty}^{\infty} y \cdot f(y \mid x)\, dy \end{cases}$$

($E(y|x)$ sometimes written $\mu_{y|x}$)

Greene also calls this "regression of y on x"

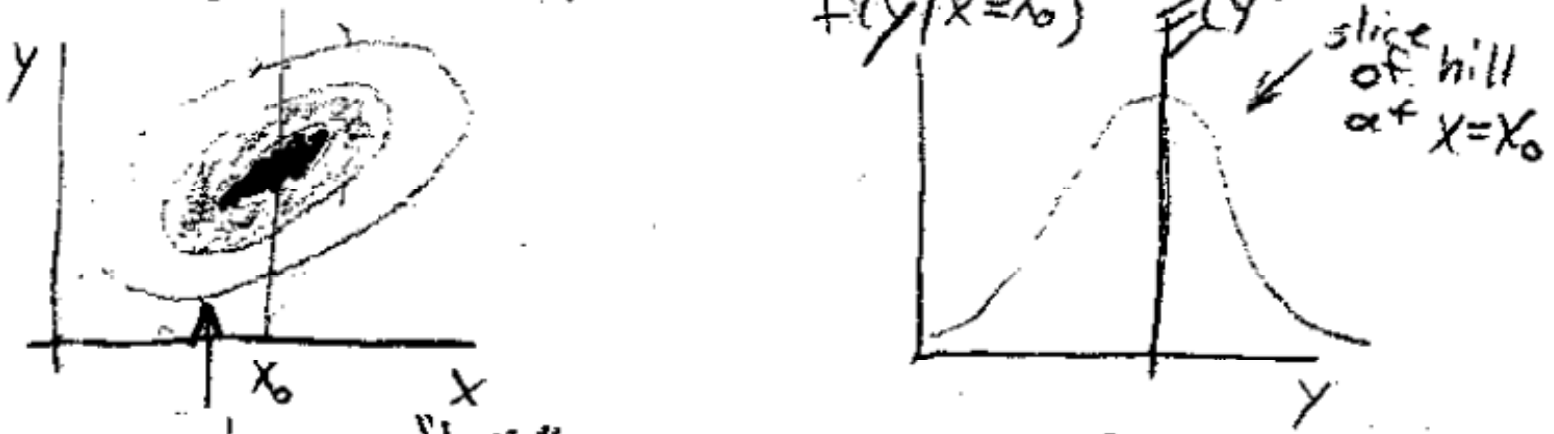
Example: Go back to our sex & height table.

that's right, but note not necessarily "linear regression of $y = nx$"

- The condition distribution of (sex | height $\geq 6'$) is $f(y|x) = \begin{cases} \Pr(\text{male} \mid 6'+) / \Pr(6'+) = .18/.28 = 9/14 \\ \Pr(\text{female} \mid 6'+) / \Pr(6'+) = .10/.28 = 5/14 \end{cases}$
- The conditional expectation, calling female = 1 & male = 0, is

$$E(y|x) = \sum_y y \cdot f(y|x) = 0 \cdot 9/14 + 1 \cdot 5/14 = 5/14$$

Harder to draw for continuous, but suppose…



These are "hills" are coming out of the page, darker is higher, higher represents "more probable" i.e. larger probability density a slice of the hill at x_0, say, would be 2 dimensional & could be drawn as in the second chart. That's $f(y|x)$ $E(y|x_0)$ is just the mean of that slice.

Perhaps the most important example is the *expected squared, conditional deviation* function or the *conditional-variance* function, i.e., **the *skedastic* function**:

$$V(y|x) \equiv E\left[(y - E(y|x))^2 \,|\, x\right] = \begin{cases} \sum_y [y - E(y|x)]^2 f(y|x) \\ \int_{-\infty}^{\infty} (y - E(y|x))^2 f(y|x)\, dy \end{cases}$$

writing it out, we can simplify:

$$V(y|x) \equiv E[(y - E(y|x))^2 \,|\, x] = E\left[y^2 - 2y\,E(y|x) + \{E(y|x)\}^2 \,|\, x\right]$$

(Expect. of sum is sum of expects.) & n.b., "givens" get distributed

$$= E[y^2|x] - 2E[y\,E(y|x)|x] + E[\{E(y|x)\}^2|x]$$

law of iterated expectations & "giving X" twice is redundant

$$= E[y^2|x] - 2E[y\,\mu_{y|x}|x] + [E(y|x)]^2$$

$$= E[y^2|x] - 2\mu_{y|x}E(y|x) + \mu_{y|x}^2$$

$$\Rightarrow V(y|x) = E(y^2|x) - [E(y|x)]^2$$

Conditional Variance:

$$V(y|x) \equiv E\left[\{(y - E(y|x)\}^2 \,|\, x\right]$$

$$\equiv \sum_y (y - \mu_{y|x})^2 \cdot f(y|x) \quad \text{or} \quad \int_{-\infty}^{\infty} (y - \mu_{y|x})^2 f(y|x)\, dy$$

$$\equiv E(y^2|x) - \mu_{y|x}^2$$

- Returning to the (beautiful) contour map of the "hills" & the slice from that map: the conditional variance of y given $X = x_0$ is the variance of $f(y|x)$ which is the variance of that slice of the "hill" i.e., a measure of how "spread" the distribution is, how "broad" or "fat" the "hill" is at that point. In general, the cond. var. depends on which x you give bc cross-sectional (conditional) shape of the hill is going to differ by where you slice it, unless the hill is very regularly shaped. No "true" hill would, for example, have identically shaped cross-sections.

FYI: The conditional-variance function is called the skedastic function. A (joint) distribution with identical $V(y|x)\ \forall\ x$, i.e. the same skedastic function for any X is called homoskedastic. If $V(y|x)$ varies by X we give, then the (joint) distribution is heteroskedastic.

- N.b.: Expectations of $y|x$: $E(y|x)$, $V(y|x)$ are generally going to be functions of x or simply constants. Thus, they "use" x to "describe" (give central tendency & spread of) y
- FYI: As thinking about the hill-crossection analogy may make clear, there are conditional skews, conditional kurtoses, etc. (i.e., conditional moments)

- Alternative Analogy for Conditional Distributions: Think of a sheet hovering in a room over air jets blowing at various strengths.
  - The strength with which a particular part of the sheet is being "blown" is actually the force of frequency with which it is hit by particles in the air stream. Thus the height of the sheet is actually a pretty good analogy for a joint p.d.f. The coordinates, x & y, give the layout of the room, & the height of the sheet at (x,y) is $f(x,y)$, the p.d.f. (& strength with which it's being "blown").
    - The conditional distribution $f(y|x=x_0)$, say, would be by the thread running through the sheet at x_0.
    - $f(x|y=y_0)$ is analogously, the thread running the other way at y_0.

5. *Summary*: Key Properties of Expectations, (Co)Variances, & Conditioning

(A) Expectations

1) Laws of Iterated Expectations

X, Y, Z = R.V.'s

a) $E(E(X)) = E(X)$

b) $E(E(Y|X)) = E(Y)$

a, b, c = constants

2) Expectation of a sum is sum of the expectations

$$E\left(\sum_{i=1}^{n} g(x)\right) = \sum_{i=1}^{n} E(g(x))$$

3) Expectation of a constant is the constant

a) $E(a) = a$ b) $E(X|X) = X$

↖ to "give" something renders it constant (if "conditions on that info")

4) Rob's Rule of Expectations:

"E slides on through constants, snagging around random variables"

a) $E(a+bX+cY) = a+bE(x)+cE(y)$

b) $E\{(X'X)^{-1}X'[ee']X(X'X)^{-1}\} = (X'X)^{-1}X'E(ee')X(X'X)^{-1}$

(constant matrix; random vector)

(B) Conditioning:

1) Conditioning is distributive:

$$E(a+bX+cY|X) = E(a|X)+bE(X|X)+cE(Y|X)$$
$$= a+bX+cE(Y|X)$$

2) Conditioning More than Once is Redundant:

$E[E(Y|X)|X] = E(Y|X)$ (on same info!)

3) Conditioning on X renders X constant (see Ab)

4) Conditioning & Expectation sort of undo each other (see Ab, Ab)

Rob's Rules for Variances & Covariances

a) Constants neither vary nor covary.

$$V(a) = 0 \qquad C(a, X) = 0$$

b) "Variance slides through sums, eliminating constants, snagging on random variables and squaring their coefficients, while spitting out 2 times all the covariances times their associated coefficients."

$$\rightarrow V(a + bX + cY) = b^2 V(X) + c^2 V(Y) + 2bc \times Cov(X, Y)$$

→ Notice that in Matrix notation, with X a constant vector and $\underset{\sim}{b}$ a random vector, this becomes just:

$$V(\underset{\sim}{X}' \underset{\sim}{b}) = \underset{\sim}{X}' V(\underset{\sim}{b}) \underset{\sim}{X}$$

V-cov mat

"X squared" in matrix-land

→ Special Cases of the Rule:

- $V(X + Y) = V(X) + V(Y) + 2 \times Cov(X, Y)$
- $V(X - Y) = V(X) + V(Y) - 2 \times Cov(X, Y)$

$(-1)^2$; $2 \times (-1)$

c) The Analogous for Covariances (becomes a kind of FOIL):

$$Cov(a + bX + cY, d + eW + fV) =$$

$$be\,Cov(X, W) + bf\,Cov(X, V) + ce\,Cov(Y, W) + cf\,Cov(V, Y)$$

(D) Properties of Note Combining the Above:

1) Any R.V. y can be written: $y = E(y|x) + (y - E(y|x))$

$\equiv E(y|x) + \varepsilon$

$\equiv \hat{y} + \varepsilon$ (n.b., not necessarily linear $\hat{y}$, & true $E(y|x)$ not literally $\hat{y}$)

2) Properties that follow:

a) $Cov(X, Y) = Cov(X, E(Y|X))$

b) $V(Y) = E(V(Y|X)) + V(E(Y|X))$

$= \hat{\sigma}_\varepsilon^2 + V(\hat{y})$ (error, regression) "variance decomposition"

$\Rightarrow TSS = RSS + ESS = ESS + RSS$ (residual, explained)

$= SSR + SSE = SSE + SSR$

Note: $\Rightarrow V(\hat{y}) \leq V(y)$ (and also that $V(\hat{\varepsilon}) \leq V(y)$ though that less notable)

I.D.2. (cont.) Properties that follow from $E(y) = E(y|x) + (y - E(y|x))$

$= \hat{y} + \varepsilon$

c) $E(XY) = E(X \cdot E(Y|X))$

d) $E(\varepsilon) = E(\varepsilon|X) = E(\varepsilon|Y) = 0$

i) $E(\varepsilon) = E(Y - E(Y|X)) = E(Y) - E(E(Y|X)) = E(Y) - E(Y) = 0$ □

ii) $E(\varepsilon|X) = E(Y - E(Y|X)|X) = E(Y|X) - E((Y|X)|X) = E(Y|X) - E(Y|X) = 0$ □

iii) $E(\varepsilon|Y) = E(Y - E(Y|X)|Y) = E(Y|Y) - E(E(Y|X)|Y) = Y - Y = 0$ □

e) $Cov(X, \varepsilon) = 0$; in fact, $Cov(h(X), \varepsilon) = 0 \ \forall \ h(\cdot)$

↳ $E(X\varepsilon) - E(X)E(\varepsilon) = E(XY - XE(Y|X)) - E(X) \cdot 0$

$= E(XY) - E(XE(Y|X))$

$= E(XY) - E(XY) = 0$ □

f) $V(\varepsilon) = E(V(Y|X))$

↳ $V(Y - E(Y|X)) = V(Y) + V(E(Y|X)) - 2Cov(Y, E(Y|X))$

$= V(Y) + V(E(Y|X)) - 2V(E(Y|X))$

$= V(Y) - V(E(Y|X))$ □

using variance-decomposition result.

I. D.3. For Special Case where $E(Y|X)$ is linear, i.e., $E(Y|X) = a + bX$

a) $a = E(Y) - bE(X)$

b) $b = Cov(X,Y)/V(X)$ proof: $Cov(X,Y) = Cov(X, E(Y|X))$

$= Cov(X, a + bX)$

$= bV(X)$

$\Rightarrow b = Cov(X,Y)/V(X)$ □

c) $V(y) = E(V(Y|X)) + V(E(Y|X))$

$= E(V(Y - a - bX)) + V(a + bX)$

$= E(V(\varepsilon)) + b^2 V(X)$

↑ recall that "giving" X_i fixes it: if constants don't vary or covary

d) $R_{X,Y} = \sqrt{b_{xy} b_{yx}} = \sqrt{\frac{Cov(X,Y)}{V(y)} \times \frac{Cov(Y,X)}{V(x)}} = \sqrt{\frac{[Cov(X,Y)]^2}{V(Y)V(X)}}$

↑ x reg y ↑ y reg x

$= \frac{Cov(X,Y)}{sd(X)\,sd(Y)}$ □

e) $E(y) = a + bE(x)$; i.e., $\bar{y} = a + b\bar{x}$

f) "Coefficient of Determination," R^2; $= \{\rho(y, \hat{y})\}^2$

H. *Independence*:

① Stochastic (Statistical) Independence

$f(y|x) = f(y)$ ⟺ $f(x,y) = f(x)f(y)$ ⟺ $f(x|y) = f(x)$

proofs:

$f(y|x) \equiv \frac{f(x,y)}{f(x)}$ so, if $f(y|x) = f(y)$, then $f(x,y) = f(x)f(y)$

$f(x|y) \equiv \frac{f(x,y)}{f(y)}$ so, if $f(x|y) = f(x)$, then $f(x,y) = f(x)f(y)$

uni-directional

if $f(x,y) = f(x)f(y)$, then

$f(x) = \frac{f(x,y)}{f(y)} \equiv f(x|y)$

& $f(y) = \frac{f(x,y)}{f(x)} \equiv f(y|x)$

knowing X doesn't tell you one darn thing about the distribution of y & v.v. Not mean, not variance, nothing

② Mean Independence: $E(Y|X) = E(Y)$

$E(Y|X) = E(Y) \not\Rightarrow E(X|Y) = E(X)$

corollary: if $E(Y|X) = E(Y)$ then $E(Y|h(X)) = E(Y)$ $\forall\, h(\cdot)$

knowing X tells you $E(Y)$, the mean of X; may not tell you $V(Y)$ or other stuff about the distribution of y

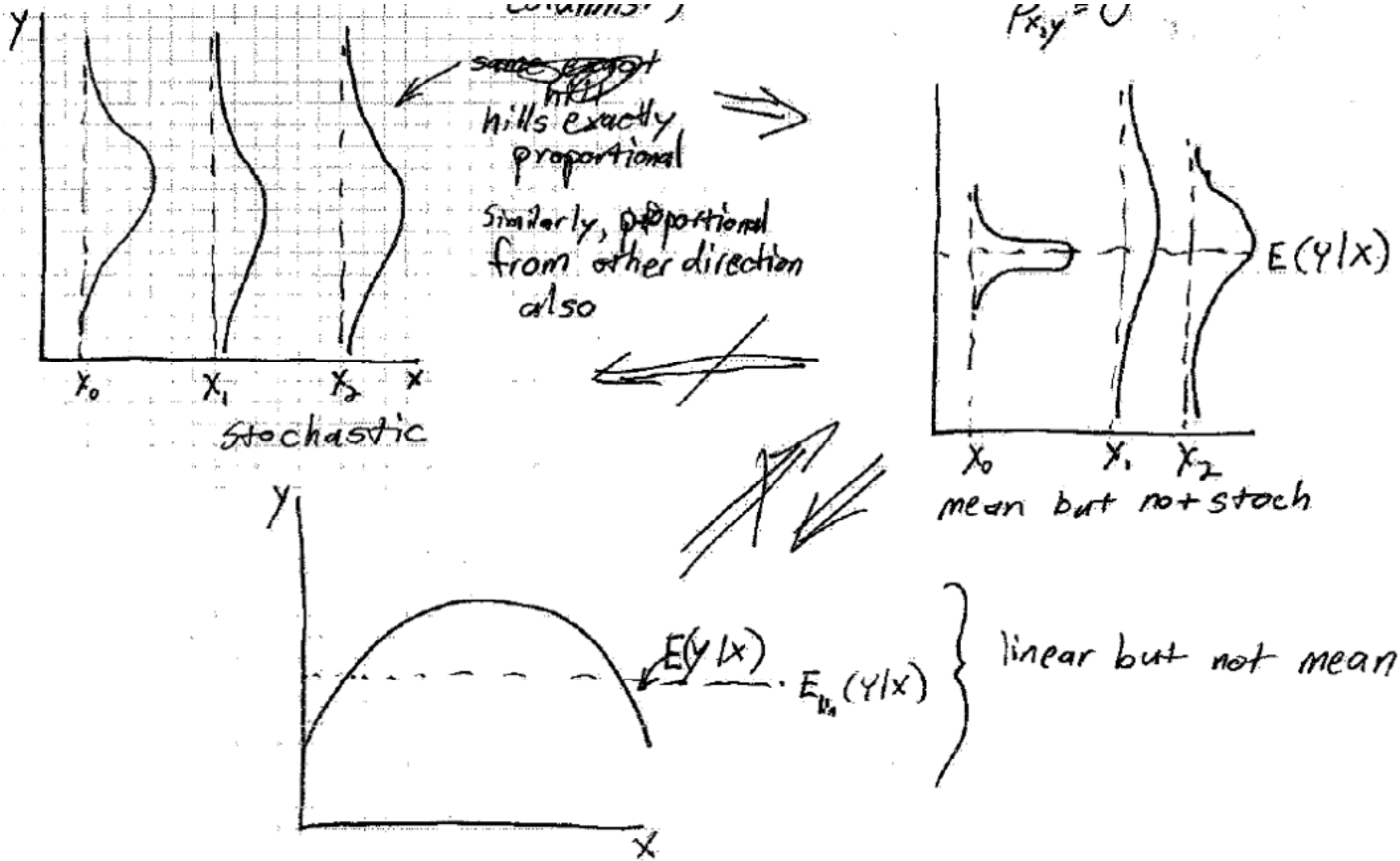
③ "Linear Independence" or Uncorrelatedness

~~(not quite same thing~~

(clarify relation to 'linear independent columns')

$Cov(X,Y) = 0$ ⟺ linear independence

$\rho_{X,Y} = 0$


Y
X
X_0
X_1
X_2
stochastic
hills exactly proportional
Similarly, proportional from other direction also
$\rho_{X,Y} = 0$
E(Y|X)
mean but not stoch
Y
X
E(Y|X)
$E_{U_i}(Y|X)$
linear but not mean

$$E(X) = \begin{bmatrix} E(X_{11}) & \cdots & E(X_{1n}) \\ \vdots & & \vdots \\ E(X_{n1}) & \cdots & E(X_{nn}) \end{bmatrix} \qquad E(\underset{\sim}{\varepsilon}) = \begin{bmatrix} E(\varepsilon_1) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix}$$

$$V(\underset{\sim}{\varepsilon}) = E\left[(\underset{\sim}{\varepsilon} - \underset{\sim}{\mu}_{\varepsilon})(\underset{\sim}{\varepsilon} - \underset{\sim}{\mu}_{\varepsilon})'\right] = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{bmatrix}$$

$n \times 1 \quad 1 \times n$

symmetric

$$= \sum_{\varepsilon} \quad \text{or} \quad \Omega$$

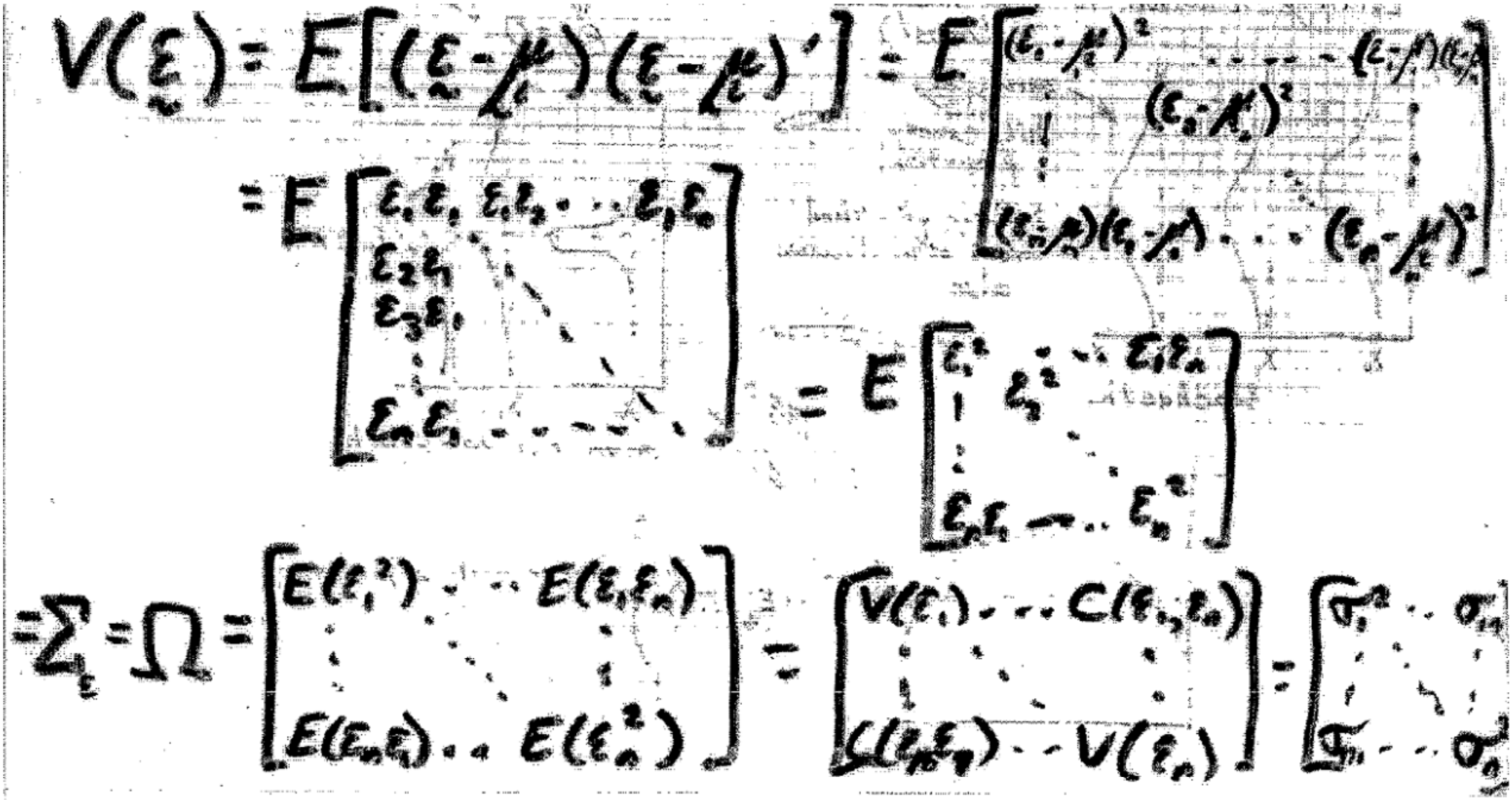
$$V(\underset{\sim}{b}) = \begin{bmatrix} V(b_1) & \cdots & C(b_1, b_k) \\ \vdots & \ddots & \vdots \\ C(b_k, b_1) & \cdots & V(b_k) \end{bmatrix}$$

confusion in notes regarding $V(X)$

Special Case: if $X_1, X_2, \ldots, X_n$ multivariate normal,

then linear $\Rightarrow$ mean $\Rightarrow$ stochastic

Longer-hand demonstration of V(vector)=matrix:



CRUCIAL IMPLICATION of INDEPENDENCE:

Given independence, joint distribution is just product of the marginal distributions:

Since in SRS, each $X_i \sim$ independently $f(x_i, \underset{\sim}{\theta})$, the joint distribution of the X_i's is

$$f(\underset{\sim}{X}, \underset{\sim}{\theta}) = f(x_1, \underset{\sim}{\theta}) f(x_2, \underset{\sim}{\theta}) f(x_3, \underset{\sim}{\theta}) \cdots$$

$\underset{\sim}{X}$: vector of the X_i's or matrix of the vectors $\underset{\sim}{X}_i$

$$= \prod_{i=1}^{n} f(x_i, \underset{\sim}{\theta})$$

II. ***Estimation & Statistical Inference***

A. *Preamble:*

We will think of outcomes or events in the world as having been generated by some underlying (and usually unobservable) probabilistic process.

A. Thus, we will think of outcomes as having been generated by some probability function or probability density function (in 2-way competitions)
   1. E.g., we might think of election outcomes as having been produced by some Bernoulli probability function.
   2. E.g., we might similarly think of how much the two candidates spend on the campaign as coming from p.d.f's
      a) The sum of their spending might be considered Normal (μ, σ^2)
      b) Each candidate's spending might be considered to have been drawn from a bivariate normal $N(\underset{\sim}{\mu}, \Sigma)$

B. This does not mean that we think of the world as random in the sense of "completely unpredictable" or "having nothing systematic about it"
   1. For one thing, the statement that the outcome in question comes from some particular type of p.f. or p.d.f. is itself a statement of "probabilistic regularity"
   2. For another thing, we usually go on to say that we know or think we know something about what determines the parameters of the distribution
      a) We may think, for example, that Π_i, probability the incumbent wins is some function of the amount of pork s/he brought to her/his district
      b) Total spending we might think to be a function of how close to 50-50 the partisanship of voters in the district, for example

C. We might, then, wish to use the outcomes that actually occured in some real-world sample of like events to infer something about the (parameters of the) underlying distribution.

D. This is the whole point: we are interested in inferring from what we observe how the world actually works; the underlying systematic elements of the political $, more broadly, the socio-politico-economic world.
   1. This course is about how best to use the information available from what has happened.
   2. "Best" here refers to the "science" of accurate inference; it is decidedly an applied science as we are interested in it

III. Samples, Sampling Theory, & Sample Statistics

A. (Simple) Random Sample (≡RS): n observations, $X_1, X_2, X_3, \ldots, X_n$ drawn independently from the same population (i.e. from the universe of possibilities as described by a p.f. or p.d.f.) is a S.R.S. from that p.f. or p.d.f.:

$\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ a S.R.S. from $f(\underset{\sim}{X}_i, \underset{\sim}{\theta})$

- $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$: may be an observation on a single R.V. or on a random vector or even a random matrix
- f: the p.f. or p.d.f.
- $\underset{\sim}{\theta}$: the parameter or parameters of that p.d.f.

e.g. $X_1, \ldots, X_n$ a S.R.S. from $N(\underbrace{\mu, \sigma^2}_{\underset{\sim}{\theta}})$

≡ each X_i a draw from an indepent $N(\mu, \sigma^2)$

1) In practice, our samples will be in the form of a

a) cross-section (CS): a set of observations on various units (individuals, states, countries, governments, etc.) at some time, t.

b) time-series (TS): a set of observations on a single unit across some number of (usually evenly spaced) time periods, $t = 1, 2, 3, \ldots, T$

c) time-series-cross-section (TSCS): the combination of (a) & (b) -- a set of cross-sectional observations across some number of time periods, or equivalently a set of time-serial observations across some cross-section of units

FYI:
- i) a TSCS of data with (many) more CS than TS is usually called panel data
- ii) a TSCS of data with (many) more TS than CS is usually called pooled data

→ this nomenclature is not all that terribly adhered to though

2) Since in SRS, each $X_i \sim$ independently $f(X_i, \underset{\sim}{\theta})$, the joint distribution of the X_i's is

$$f(\underset{\sim}{X}, \underset{\sim}{\theta}) = f(X_1, \underset{\sim}{\theta}) f(X_2, \underset{\sim}{\theta}) f(X_3, \underset{\sim}{\theta}) \cdots$$

($\underset{\sim}{X}$: vector of the X_i's or matrix of the vectors $\underset{\sim}{X}_i$)

$$= \prod_{i=1}^{n} f(X_i, \underset{\sim}{\theta})$$

B) (Sample) Descriptive Statistics

1) A <u>statistic</u> is any function computed from the data.
- of course we usually restrict our attention to functions of the data that are informative about the distribution of the data (& hopefully ∴, by some inference process, informative about the underlying population).

2) Some common <u>"descriptive" statistics</u> (i.e. they purport to help describe or summarize the entire sample distribution & again by inference, the population distribution

III.B.2. Descriptive Statistics:

a) First, generically, a statistic, γ is some function of the data

$$\gamma = g(x_1, x_2, \ldots, x_n)$$

b) Examples:

i) The Sample Mean or Average

(a) The mean, i.e. μ or $E(X)$, is a population feature, the sample mean or average is an estimate of it.

(b) Def. $\bar{X} \equiv g(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$

(c) Other measures of the sample's central tendency, the sample median, mode, midrange, etc. can be defined analogously to their population cousins as well.

ii) The Sample Standard Deviation:

(a) usually, defined $s_x \equiv \left[\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\right]^{1/2}$

occasionally, we use (for reasons to be explained)

$$\hat{\sigma}_x = \left[\frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\right]^{1/2}$$ instead

or divide by some other quantity (related to n) for other reasons

(b) other measures of the sample spread, sample mean absolute deviation or sample range etc can again be defined similarly to their population analogs. Sample variance, e.g., is square of sample std. dev.

iii) Sample Relationships

(a) sample covariance:

$$s_{xy} \equiv \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

(b) sample correlation:

$$r_{xy} \equiv \frac{s_{xy}}{s_x s_y}$$

(c) it is often useful to organize & arrange sample variances & covariances &/or correlations in matrices:

(1) Sample Covariance Matrix

$$\underset{\sim}{S_x} = \begin{bmatrix} s_{x_1x_1} & \cdots & s_{x_1x_n} \\ \vdots & \ddots & \vdots \\ s_{x_nx_1} & \cdots & s_{x_nx_n} \end{bmatrix}$$

(2) Sample Correlation Matrix

$$\underset{\sim}{R_x} = \begin{bmatrix} r_{x_1x_1} & \cdots & r_{x_1x_n} \\ \vdots & \ddots & \vdots \\ r_{x_nx_1} & \cdots & r_{x_nx_n} \end{bmatrix}$$

n.b.: • Diagonal elements of $\underset{\sim}{S_x}$ are variances of x_i, right?

• Diagonal elements of $\underset{\sim}{R_x}$ are all one (1), right?

III.B.2.c. Some useful short-cuts for calculating sample statistics.

i) $S_x^2 \equiv$ sample variance $= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2\right] = \sum \frac{1}{n-1} X_i^2 - \frac{n}{n-1}\bar{X}^2$

($\& \; S_x = \sqrt{S_x^2}$)

ii) $S_{xy} \equiv$ sample cov. of $x \& y = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} X_i Y_i\right) - n\bar{X}\bar{Y}\right] = \sum \frac{1}{n-1} X_i Y_i - \frac{n}{n-1}\bar{X}\bar{Y}$

iii) $r(a \cdot X, b \cdot Y) = \frac{a \cdot b}{|a \cdot b|} r_{X,Y}$

(constants; absolute value)

[e.g. $r_{X,Y} = .5$, $\Rightarrow r(2 \cdot X, -1 \cdot Y) = -.5$]

iv) $S(a + b \cdot X, c + d \cdot Y) = b \cdot d \cdot S_{XY}$

(constants)

[e.g. $S(2 + 3X, 4 + 2Y) = 12 \cdot S_{X,Y}$

• implies, for example, $S^2(a + bX) = b^2 \cdot S_X^2 \Rightarrow S(a + bX) = |b| S_X$ (abs. val.)

• in short, all the useful properties of expectations, variances, & covariances we used for population R.V.'s last week work exactly analogously for their sample statistic analogs.

III. C. Sampling Distributions (of Sample Statistics)

1) The random sample itself can be seen as presenting a "sample distribution". This is <u>not</u> to be confused with the <u>sampling distribution</u> of a statistic.

2) Since each observation is a draw from a probability distribution, if we drew another sample the observations would be different. Thus, <u>across repeated samples</u>, sample statistics calculated from those samples will vary in some way. The <u>sampling distribution</u> of a statistic is the p.f. or p.d.f. which describes how the statistic varies across these (usually hypothetical) samples.

3) Thus, we can speak, for example, of $\bar{X} \sim f(\bar{x}, \underset{\sim}{\theta})$; i.e. of the sample average being distributed across repeated samples, according to $f(\bar{x}, \underset{\sim}{\theta})$. We might, furthermore, be interested in the expected value or variance, etc., of $\bar{X}$ or whatever other statistic across these repeated samples.

3) Thus, we can speak, for example, of $\bar{X} \sim f(\bar{x}, \underset{\sim}{\theta})$; i.e. of the sample average being distributed, across repeated samples, according to $f(\bar{x}, \underset{\sim}{\theta})$. We might, furthermore, be interested in the expected value or variance, etc, of $\bar{X}$ or whatever other statistic across these repeated samples.

Eg. $E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$ by definition of $\bar{X}$

$= \frac{1}{n} E\left(\sum_i X_i\right)$ b/c can pull constants out of $E(\cdot)$

$= \frac{1}{n}\left(E(X_1) + E(X_2) + E(X_3) + \ldots\right)$ writing out the sum

$= \frac{1}{n}\left(\mu_x + \mu_x + \mu_x + \ldots + \mu_x\right)$ each X_i is a S.R.S. from population with some mean μ_x)

$= \frac{1}{n}(n \cdot \mu_x)$

$= \mu_x$

$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 V\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(X_i)$

n.B. in this step, since X_i independent, $Cov(X_i, X_j) = 0$

$= \frac{1}{n^2}(n \cdot \sigma_x^2)$

$= \sigma_{\bar{x}}^2 / n$

III.C.3. Mean & Variance of the Sampling Distribution of $\bar{X}$

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

$$V(\bar{X}) = \frac{1}{n} \cdot \sigma_x^2$$

Notes: a) This is true for X's drawn (independently) from any p.f. or p.d.f. with defined means & variances.

b) Variance of averages across repeated samples is $\frac{1}{n}$ • (the variance of the underlying random variable)

$\Rightarrow$ "averaging reduces variance" in this sense & by a factor of $\frac{1}{n}$ where n is the sample size being averaged.

An example: if $X_i \sim N(\mu_x, \sigma_x^2)$ $\forall i$ ($\Rightarrow$ the X_i are indep.)

$$\bar{X} \sim N(E(\bar{X}), V(\bar{X}))$$ b/c sum of normals is normal

$$\bar{X} \sim N(\mu_x, \sigma_x^2/n)$$

4. We will often be interested in how a statistic behaves across repeated samples as those (hypothetically repeated) samples get larger & larger:

c.g.a) $$\lim_{n\to\infty} V(\bar{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

b) $$\lim_{n\to\infty} E(\bar{X}) = \lim_{n\to\infty} \mu_x = \mu_x$$ (μ_x constant, doesn't vary with n)

c) so, as $n \to \infty$, the distribution of $\bar{X}$ goes to "a "spike" over μ_x
• more on this in a bit.

IV. (Point) Estimation of Parameters

A. First some Definitions

1. <u>point estimate</u> a statistic that gives a single-value estimate or "guess" for a parameter, θ, or for each of a set of parameters, $\underset{\sim}{\theta}$
2. <u>standard error</u> the standard deviation of the sampling distribution of some point estimate
   - the sampling standard deviation is the standard error; the square of that is the sampling variance
3. <u>interval estimate</u> a statistic that gives some range of values as a guess of θ or $\underset{\sim}{\theta}$. Usually, the interval is constructed in an appropriate (i.e. useful) manner from a point estimate and its standard error.

   e.g. "<u>margin of error</u>"; 40% favor Clinton +/- 2: $\theta \equiv$ true % favoring Clinton, 40% is the point estimate, $\hat{\sigma}$ is the standard error, "margin of error" $\equiv 2 \cdot \hat{\sigma}$ (so here, $\hat{\sigma}$ must be 1) & the interval estimate is 40% +/- 2 or $\{38\%, 42\%\}$
4. <u>Estimator</u>: an estimator is a rule or a strategy for guessing at a parameter; i.e. it is the function that produces an estimate:

   e.g. $g(\underset{\sim}{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ is an estimator for μ_x (that we often call $\bar{x}$)
5. <u>Finite</u> (or "small") <u>Sample</u>: any fixed sample size less than ∞

   <u>Asymptotic</u> (or "large" or "infinite" <u>sample</u>: a sample of infinite or "going toward infinite" size

IV.B. Criteria for Judging/Comparing Estimators

- An estimator for (a) parameter(s), θ, can be any old function of the data, but surely some are "better" in some sense than others as "guesses" of θ. The "better" we are usually concerned with relates to how they perform across repeated samples; i.e., we are compare them in terms of their sampling distributions.

1. **Unbiasedness:** An estimator of θ, call it $\hat{\theta}$, is unbiased iff $E(\hat{\theta}) = \theta$

   a. Examples: i. $\bar{X}$ is an unbiased estimator of μ_x

   $$E(\bar{X}) = E\left(\sum_{i=1}^{n} X_i/n\right) = \frac{1}{n} E\left(\sum X_i\right) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum \mu_x$$
   $$= \frac{1}{n} \cdot n\mu_x = \mu_x \quad \square$$

   ii. $\hat{\theta}$ = 1st observation of X, X_1, is an unbiased estimator of μ_x

   $$E(\hat{\theta}) = E(X_1) = \mu_x \quad \square$$

   b. $\text{Bias}(\hat{\theta}) \equiv E(\hat{\theta}) - \theta$ (= 0 in above two examples)

   $\hat{\theta} = \frac{1}{3}X_1$ is a biased estimate of μ_x, its bias is

   $$E\left(\tfrac{1}{3}X_1\right) - \mu_x = \tfrac{1}{3}E(X_1) - \mu_x = \tfrac{1}{3}\mu_x - \mu_x = -\tfrac{2}{3}\mu_x$$

2. **Efficiency:** An unbiased estimator, $\hat{\theta}_1$, is more efficient than another unbiased estimator, $\hat{\theta}_2$, iff $V(\hat{\theta}_1) < V(\hat{\theta}_2)$

   e.g. $\hat{\theta}_1 \equiv \bar{X}$ is more efficient than $\hat{\theta}_2 = X_1$: $V(\bar{X}) = \frac{\sigma_x^2}{n} < V(X_1) = \sigma_x^2$

Efficiency in the multivariate case:

$\hat{\boldsymbol{\theta}}_1$ is more efficient than $\hat{\boldsymbol{\theta}}_2 \Leftrightarrow V(\hat{\boldsymbol{\theta}}_2) - V(\hat{\boldsymbol{\theta}}_1)$ is positive definite;

$$i.e.,\ \mathbf{x}'\left[V(\hat{\boldsymbol{\theta}}_2) - V(\hat{\boldsymbol{\theta}}_1)\right]\mathbf{x} > 0\ ,\ \forall \mathbf{x}.$$

Note: efficiency is unambiguously desirable only when comparing two estimators with equal bias or consistency properties/performance. More generally, a tradeoff arises between bias/consistency and efficiency. A measure that combines these desiderata is:

3. Mean-Squared Error (MSE):

$$MSE(\hat{\theta}) \overset{①}{\equiv} E\left[(\hat{\theta}-\theta)^2\right] \overset{⑥}{=} Var(\hat{\theta}) + (bias(\hat{\theta}))^2$$

$$\overset{②}{=} E\left[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2\right]$$

$$\overset{③}{=} E(\hat{\theta}^2) - 2E(\hat{\theta}\cdot\theta) + \theta^2$$

$$\overset{④}{=} \left[E(\hat{\theta})\right]^2 + V(\hat{\theta}) - 2E(\hat{\theta})\cdot\theta + \theta^2$$

for any R.V., $E(X^2) = \mu_X^2 + \sigma_X^2$

for any 2 R.V.'s, X & Y, $E(XY) = Cov(X,Y) + E(X)E(Y)$

$Cov(\hat{\theta},\theta) = 0$ b/c θ is a constant

$$⑤\ Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$$\Rightarrow [bias(\hat{\theta})]^2 = [E(\hat{\theta})]^2 - 2E(\hat{\theta})\theta + \theta^2 \quad ;\ \text{so } ④+⑤$$

Two Ways to Show $E[(\hat{\theta}-\theta)^2] = V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$

① $E[(\hat{\theta}-\theta)^2] = V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$

$E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) = E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2$

$E(\hat{\theta}^2) - 2\theta^2 + \theta^2 = E(\hat{\theta}^2 - 2E(\hat{\theta})\hat{\theta} + [E(\hat{\theta})]^2) + E(\hat{\theta})^2 - 2E(\hat{\theta})\theta + \theta^2$

b/c for any R.V. $\hat{\theta} = \theta + (\hat{\theta} - \theta)$

$\theta \cdot [\theta + (\hat{\theta} - \theta)]$

$\theta \cdot \theta + 0$

$= E(\hat{\theta}^2) - 2(E(\hat{\theta}))^2 + [E(\hat{\theta})]^2 + [E(\hat{\theta})]^2 - 2\theta^2 + \theta^2$

(taking $E(\cdot)$ through)

b/c for any R.V. $E(\hat{\theta}) = \theta + (E(\hat{\theta}) - \theta)$

$\theta \cdot [\theta + (E(\hat{\theta}) - \theta)]$

$\theta^2 + 0$

$E(\hat{\theta}^2) - \theta^2 = E(\hat{\theta}^2) - \theta^2$

just add & subtract $E(\hat{\theta})$

① $E\{(\hat{\theta}-\theta)^2\} = E\{[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2\}$

$= E\{(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)\}$

$= V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2 + 2 \cdot E(\hat{\theta} \cdot E(\hat{\theta}) - \hat{\theta} \cdot \theta - E(\hat{\theta})^2 + E(\hat{\theta})\theta)$

$+ 2 \cdot [[E(\hat{\theta})]^2 - E(\hat{\theta})\theta - [E(\hat{\theta})]^2 + E(\hat{\theta})\theta]$

$+ 2 \times 0$

$= V(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$

(Minimum) Mean Squared-Error ("(best) combined min-bias & max-effic."):

$$E\left[\left(\hat{\theta}-\theta\right)^2\right] = V\left(\hat{\theta}\right) + \left[bias\left(\hat{\theta}\right)\right]^2$$

4. More generally, we can assume / or it is appropriate to apply some other function reflecting the costs of "mistakes".

   E.g. Min MSE assumes/applies costs of mistakes according to the function $(\hat{\theta}-\theta)^2$. We could imagine scenarios where costs of mistakes are zero for small mistakes & constant for any large error e.g.

   $$L(\hat{\theta}) = \begin{cases} 0 & \forall\ \hat{\theta}-\theta < c \\ 4 & \forall\ \hat{\theta}-\theta \geq c \end{cases}$$

   or any-old loss function we like

5. Minimum Variance Unbiased Estimator (MVUE): one which is unbiased & has the lowest variance of any estimator that is likewise unbiased.

   Minimum Variance Linear Unbiased Estimator (MVLUE) or "Best" Linear Unbiased Estimator (BLUE): one which is unbiased & linear & has the lowest variance of any estimator that is likewise linear & unbiased.

- in general, we cannot know if some estimator is MVUE or MVLUE/BLUE. (Similarly, we do not generally know if some estimator is Minimum MSE.) However, Cramér & Rao proved an interesting result:

a) Proving ***minimum-variance*** (*most-efficient*) among some class of estimators (among unbiased estimators, or among linear & unbiased estimators…) is hard. One way can know you have min-var (or ***best***) among unbiased is via ***Cramér-Rao Lower-Bound***

Cramér-Rao Lower Bound → for any "well-behaved" p.d.f., any unbiased estimator of parameter θ, call it $\hat{\theta}$, will have variance at least as large as:

$$\left[I(\theta)\right]^{-1} \equiv \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]\right)^{-1}$$

- $L(\theta)$ here is the "likelihood function" of the data, i.e., loosely, it is

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta) \quad ; \text{ e.g. suppose } X_i \text{ drawn from exponential distribution,}$$

θ = hazard rate

$$X_i \sim f(X_i, \theta) = \theta e^{-\theta X_i}$$

$\Rightarrow$ life-expectancy $= \frac{1}{\theta}$ = mean

$\frac{1}{\theta^2}$ = var

$$\text{then, } L(\theta) = \prod_{i=1}^{n} f(X_i, \theta) = \prod_{i=1}^{n} \theta e^{-\theta X_i} = \theta^n e^{-\theta \Sigma X_i}$$

$$\text{thus, } \ln L(\theta) = n \ln \theta - \theta \sum_{i=1}^{n} X_i \rightarrow \hat{\theta}_{ML} = \left(\frac{\Sigma X_i}{n}\right)^{-1}$$

$$\text{i.e., } \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = \frac{\partial\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)}{\partial \theta} = \frac{\partial\left(\frac{n}{\theta} - \Sigma X_i\right)}{\partial \theta} = -\frac{n}{\theta^2}$$

$$\text{So, the C-R lower bound} = \left(-E\left(-\frac{n}{\theta^2}\right)\right)^{-1} = \left[E\left(\frac{n}{\theta^2}\right)\right]^{-1} = \left[\frac{n}{\theta^2}\right]^{-1}$$

$$= \frac{\theta^2}{n}$$

Thus, C-R tells us that any unbiased estimator of the parameter θ in an exponential distribution must have a variance (across repeated samples) of at least $\left[\frac{\theta^2}{n}\right]$. That implies, for example, that if we can find an unbiased estimator with that variance, we can stop looking: no other unbiased estimator will do any better.

6. Large-Sample Criteria for Estimators:

a. Convergence in Probability

i. **Def** X_n converges in probability to a constant, c, written $X_n \xrightarrow{p} c$,
$\iff \lim_{n\to\infty} \text{Prob}(|X_n - c| > \varepsilon) = 0$ for any positive ε.
(abs. value)

- we also write this $\operatorname{plim} X_n = c$ ("plim" ↔ "probability limit")

ii. **Def** X_n converges in [quadratic mean] (something [mean square]) (assuming X_n has mean μ_n & var σ_n^2)
$\iff \lim_{n\to\infty} \mu_n = c$ & $\lim_{n\to\infty} \sigma_n^2 = 0$ & so $\operatorname{plim} X_n = c$

iii. 1 & 2 are not quite the same (2 is stricter), but for our purposes don't sweat the difference

IV.B.6.b. **Consistency**: an estimator, $\hat{\theta}$, of a parameter, θ, is consistent
$\iff \operatorname{plim} \hat{\theta} = \theta$

- The point: a consistent estimator goes to exactly the right answer (i.e. exactly the parameter you are estimating) as the sample size goes to infinity

- Example: $\bar{X}$ is a consistent estimator of μ

① $E(\bar{X}) = \mu$
② $V(\bar{X}) = \sigma^2/n$
$\Rightarrow$ ① $\lim_{n\to\infty} E(\bar{X}) = \mu$
② $\lim_{n\to\infty} V(\bar{X}) = 0$

$\therefore \bar{X}$ converges in quadratic mean to μ. (which is stricter than and therefore) $\Rightarrow \operatorname{plim} \bar{X} = \mu$

- Extensions: ① $\text{plim } \frac{1}{n}\sum_i g(x_i) = E(g(x))$

i.e. the average of any function evaluated at each observation is a consistent estimator of the mean of that function

e.g. $E(X^2) = \Theta$

$\hat{\Theta} = \frac{1}{n}\sum_i x_i^2$

$\Rightarrow \hat{\Theta}$ is consistent

② (Slutsky Theorem) for any continuous $g(X_n)$, $\text{plim } g(X_n) = g(\text{plim } X_n)$

examples: $\text{plim } 3x^2 = 3(\text{plim } X)^2$

$\text{plim } \frac{\bar{X}^2}{S^2} = \frac{(\text{plim } \bar{X})^2}{(\text{plim } S)^2} = \mu^2/\sigma^2$

C. Convergence in Distribution & Limiting Distribution

i. Def X_n converges in distribution to a (different) random variable X with cumulative probability function, $F(X)$ $\Leftrightarrow$

$\lim_{n\to\infty} |F_n(X) - F(X)| = 0$ (written $X_n \xrightarrow{d} f(X)$ or $X_n \xrightarrow{d} X$)

↑ abs. value.

i.e. as $n \to \infty$, the distribution of X_n becomes ever more like, & in the limit exactly equal to, some (other) distribution

- The distribution being "approached" in this way is called the limiting or asymptotic distribution of X_n

ii. The most common & most important example: $X_n \sim t_n \Rightarrow X_n \xrightarrow{d} N(0,1)$

iii. Useful Combinations of the Two Convergences:

① $X_n \xrightarrow{d} X$ & $\operatorname{plim} y_n = c$

② $X_n \xrightarrow{d} X$ & $g(X_n)$ continuous

$\Rightarrow g(X_n) \xrightarrow{d} g(X)$

$\Rightarrow$
$X_n y_n \xrightarrow{d} cX$
$X_n + y_n \xrightarrow{d} X + c$
$X_n / y_n \xrightarrow{d} X/c$ $(c \overset{\text{if}}{\neq} 0)$

iv. Another Useful Example: $F_{1,n} \xrightarrow{d} \chi^2_1$

v. Convergence in Distribution is also sometimes written

$X \overset{A}{\sim} f(X)$ "X is assymptotically distributed $f(X)$"

IV. B.6.d. Best (ie. Minimum Variance) Asymptotic Normal (BAN) & Consistent (BANC) Estimator

$\equiv \begin{cases} \text{①} \ \hat{\theta} \xrightarrow{d} N(\theta, \phi^2/n) \\ \text{②} \ \phi^2 \leq \phi^{*2} \end{cases}$ for any alternative estimator which converges in distribution to $N(\theta, \phi^{*2}/n)$

- The Reason we care about the CLT & limiting distributions is that we (more or less, speaking a bit loosely) "use" them to talk about the asymptotic distributions of statistics we care about.

e.g., as already noted, $\bar{X}_n \overset{A}{\sim} N(\mu_x, \sigma^2/N)$

& $Z = \left((\bar{X}_n - \mu)/\sigma/\sqrt{N}\right) \overset{A}{\sim} N(0,1)$

- Z is not exactly distributed $N(0,1)$, but

1) As n (sample size gets larger) it is more & more like it

2) The closer X (i.e. the actual, exact distribution of X) is to normal to begin with, the sooner & quicker the distribution of Z goes to $N(0,1)$

Consider, forexample, the statistic: $(\hat{\beta} - \beta_0)/s.e.(\hat{\beta}) = t$

i.e., the stat we use to test the hypothesis that $\hat{\beta} = \beta_0$. The CLT & Asymptotic Distribution theory tell us that if

a) $\hat{\beta}$ is an unbiased estimator of β & if

b) s.e. $(\hat{\beta})$ is a consistent estimator of $\sigma_{\hat{\beta}}$, then

$t \overset{A}{\sim} N(0,1)$

- Now, if, as in fact we often will, we know a bit more about $\hat{\beta}$, & s.e.$(\hat{\beta})$ we might know their actual distribution (e.g., under OLS assumptions, $t \sim t_n$), but we don't actually need anything more than "good" estimates of $\hat{\beta}$ & s.e.$(\hat{\beta})$ + the C.L.T. to get "asymptotic normality" for the "t" statistic.

B. *The Law of Large Numbers (LLN) & the Central Limit Theorem (CLT)* (the two most-important theorems in probability & statistics):

The Law of Large Numbers: If simple random sampling from any population with $E(X) = \mu$ & $V(X) = \sigma^2$, then the average $\boxed{\bar{X} \xrightarrow{p} \mu}$

example: X = R.V. equal to a die roll's upward face;

$\bar{X}_n \xrightarrow{p} 3.5$ (i.e. $\lim_{n \to \infty} \frac{1}{n} \sum (\text{die rolls}) = E(\text{die roll})$)

(n = number of die rolls)

Central Limit Theorem: in S.R.S. from any population w/mean μ & variance σ^2,

$$\left\{ Z = \frac{\bar{X}_n - \mu}{(\sigma/\sqrt{n})} \right\} \xrightarrow{d} N(0,1)$$

1. This is the most useful form of the CLT, notice also, though, that the left-hand side can be rewritten:

$$\sqrt{n} \cdot \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \xrightarrow{d} N(0,1)$$

2. Given our rule that $V(bX) = b^2 V(X)$, we can also rewrite it:

$$\sqrt{n} \cdot (\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \quad \text{(as Greene Does)}$$

3. Given our rule that $E(b+X) = b + E(X)$, we can also rewrite it

$$\bar{X}_n \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

if X is drawn from population w/ mean μ, but variance which is different for every observation, σ_{X_i}, then

$$\sqrt{n} \cdot (\bar{X}_n - \mu) \xrightarrow{d} N(0, \bar{\sigma}^2)$$

↑ $\bar{\sigma}^2$ is the "average variance," $\frac{1}{n}\sum \sigma_i^2$

Multivariate Extensions of the CLT:

1. if $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ are a S.R.S. from a joint distribution with mean $\underset{\sim}{\mu}$ and variance Q, then $\sqrt{n}\,(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) \xrightarrow{d} N(0, Q)$
   ↑ matrix (positive definite)
   & all the permutations analogous to the above.
2. if $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ S.R.S. with mean $\underset{\sim}{\mu}$, & variance$(\underset{\sim}{X}_i) = Q_i$ (differs for each i)

$$\sqrt{n}\,(\bar{\underset{\sim}{X}}_n - \underset{\sim}{\mu}) \xrightarrow{d} N(0, Q)$$

where $\bar{Q}_n = \frac{1}{n}\sum_i Q_i$

& $Q = \lim_{n\to\infty} \bar{Q}_n$

3. That is, the CLT generalizes in a more or less transparent way to the multivariate case

III. Basic Statistics Review, Summary:

A. A ***statistic***, or an ***estimate***, is any quantity calculated from the data.

B. An ***estimator*** is formula or strategy applied to calculate an *estimate*.

C. Properties of estimators:

1. ***Unbiasedness*** ("right on average, across samples"): $E(\hat{\theta}) = \theta$.

2. ***Consistency*** ("exactly right, w/o variance, as sample-size$\to\infty$"): $\lim_{n\to\infty} \hat{\theta} = \theta$.

3. ***Efficiency*** ("all useful information used; no wasted info"):

$\hat{\theta}_1$ is more efficient than $\hat{\theta}_2 \Leftrightarrow V(\hat{\theta}_2) - V(\hat{\theta}_1)$ is positive definite;

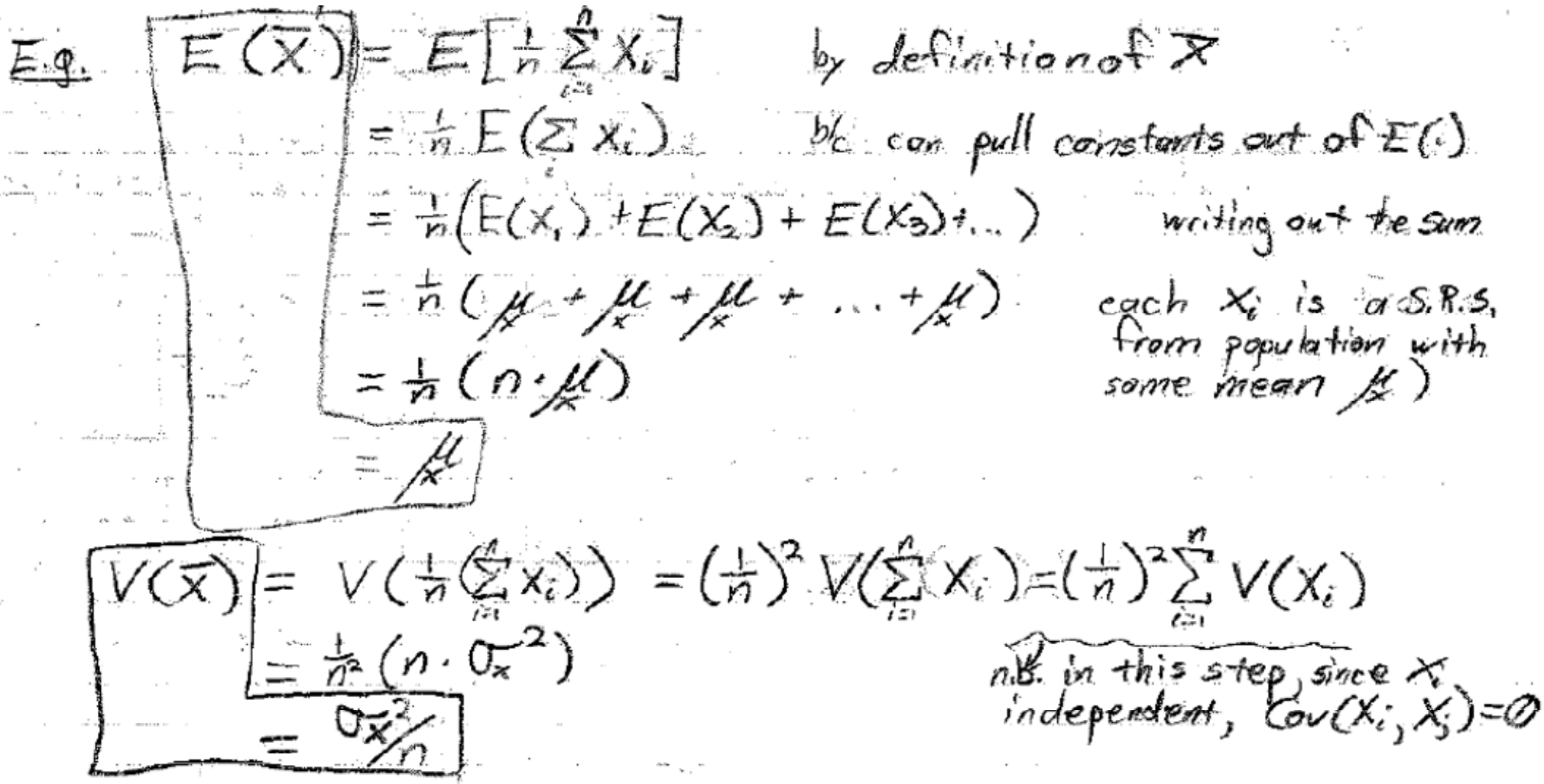
i.e., $\mathbf{x}'\left[V(\hat{\theta}_2) - V(\hat{\theta}_1)\right]\mathbf{x} > 0$, $\forall \mathbf{x}$.

4. ***(Minimum) Mean Squared-Error*** ("(best) combined min-bias & max-effic."):

$$E\left[(\hat{\theta} - \theta)^2\right] = V(\hat{\theta}) + \left[bias(\hat{\theta})\right]^2$$

5. ***Minimum-Variance (Best) Unbiased Estimator*** (BUE) & ***Minimum-Variance (Best) Linear Unbiased Estimator*** (BLUE), ***Minimum-Variance (Best) Asymptotically Normal & Consistent Estimator*** (BANC): Self-explanatory…

6. Following can exemplify ***estimator, estimate, unbiased, consistent, efficient***:



IV. Maximum-Likelihood Estimation:

A. We now actually have more or less all the tools we need for Maximum Likelihood Estimation (MLE)

1. We view outcomes as being produced by some p.f. or p.d.f., $f(X_i, \underset{\sim}{\theta})$
2. If each outcome is i.i.d. (i.e., is from a S.R.S.), then their joint distribution is the product of their marginal distributions:

$$f(X_1, X_2, X_3, \ldots, X_n, \underset{\sim}{\theta}) = f(X_1, \underset{\sim}{\theta}) f(X_2, \underset{\sim}{\theta}) f(X_3, \underset{\sim}{\theta}) \cdots f(X_n, \underset{\sim}{\theta})$$

3. What we do now is choose $\underset{\sim}{\theta}$ (i.e. estimate it) so that it maximizes the "probability" or likelihood of our having observed the data we actually have observed.

B. That is, given the data, $\underset{\sim}{X}$, & the model $f(\cdot)$, define the (joint) conditional distribution of $\underset{\sim}{\theta}$ given $\underset{\sim}{X}$. Maximize this with respect to $\underset{\sim}{\theta}$.

C. Examples:

1. Each $X_i \sim \text{Poisson}(\theta)$: $f(X_i, \theta) = \dfrac{e^{-\theta}\theta^{X_i}}{X_i!}$

(A Poisson Distribution describes a count (R.V.) of events happening at average rate θ per fixed period)

a. $f(X_1, \ldots, X_n, \theta) = \prod_{i=1}^{n} \dfrac{e^{-\theta}\theta^{X_i}}{X_i!}$

b. $\max_{\theta} f(\underset{\sim}{X}, \theta) = \max_{\theta} \ln f(\underset{\sim}{X}, \theta) = \max_{\theta} \sum_{i=1}^{n} [\ln(e^{-\theta}\theta^{X_i}) - \ln(X_i!)]$

$= \max_{\theta} \underbrace{\sum_{i=1}^{n} (-\theta + X_i \ln\theta - \ln(X_i!))}_{\text{call this } L(\theta)}$

C. First-order Condition $\dfrac{\partial L(\theta)}{\partial \theta} = 0$

Aside: $\max f(x) = \max g(f(x))$ for any $g(\cdot)$ which is a strictly monotonic increasing function such as $g(\cdot) = \ln(\cdot)$

Aside #2: "strictly monotonic increasing" means that, in $g(x)$, if x goes up so does $g(x)$

VI. C. 5. d. Maximizing Log-Likelihood of Poisson Model

$$L(\theta) = \sum_{i=1}^{n} (-\theta + x_i \ln\theta - \ln x_i) = -n\theta + \left(\sum_{i=1}^{n} x_i\right)\ln\theta - \sum_{i=1}^{n} \ln(x_i!)$$

F.O.C. $\frac{\partial L(\theta)}{\partial \theta} = 0 \Rightarrow -n + \sum_{i=1}^{n} x_i / \theta^* = 0$

$$\Rightarrow \theta^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

e. Second-Order Condition:

$$\theta > \frac{\partial\left(\frac{\partial L(\theta)}{\partial \theta}\right)}{\partial \theta} \qquad \frac{\partial(-n + \theta^{-1}\cdot\Sigma x_i)}{\partial\theta} = -\theta^{-2}\left(\sum x_i\right)$$

↳ greater than zero

$$\Rightarrow \frac{\partial^2 L(\theta)}{\partial \theta^2} < 0 \ \checkmark$$

f. Thus, MLE for Poisson parameter, θ, is $\hat{\theta} = \bar{x}$.

VI. C. 2. MLE Example 2: MLE for Normal

a. $f_n(x, \underset{\sim}{\theta}) = (2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/2\sigma^2}$

b. $L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/(2\sigma^2)}$

$$\ln L(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \ln\left[(2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/(2\sigma^2)}\right]$$

$$= \sum_{i=1}^{n} \left[-\tfrac{1}{2}\ln(2\pi\sigma^2) - (x_i-\mu)^2/(2\sigma^2)\right]$$

$$= \sum_{i=1}^{n} \left(-\tfrac{1}{2}\ln 2 - \tfrac{1}{2}\ln\pi - \tfrac{1}{2}\ln\sigma^2 - \tfrac{1}{2}(x_i-\mu)^2\sigma^{-2}\right)$$

$$= -\tfrac{n}{2}\ln 2 - \tfrac{n}{2}\ln\pi - \tfrac{n}{2}\ln\sigma^2 - \tfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(\sigma^2)^{-1}$$

c. F.O.C. $\frac{\partial \ln L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = 0$:

i) $\frac{\partial \ln L(\underset{\sim}{\theta})}{\partial \mu} = (\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu^*) = 0$

ii) $\frac{\partial \ln L(\underset{\sim}{\theta})}{\partial \sigma^2} = \frac{n}{2\sigma_*^2} + \frac{1}{2}\cdot(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$

- Multiply both sides of (i) by $(\sigma^2)^{-1}$ $\Rightarrow \sum_{i=1}^{n}(x_i-\mu) = 0$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\mu^* = 0$$

$$\Rightarrow \mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

- Substitute that into (ii) & you'll find $\sigma_*^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

VI. C. 3. We note (in passing for now) that Maximum Likelihood Estimation is especially useful when we have some substantive model for the parameter(s), such as $\mu = a + bX$ or something. It's especially especially useful when the probability function underlying the outcomes we want to model is non-linear.

VI. D. Properties of Max Like Estimators: 1) Consistent ($\operatorname{plim} \hat{\theta}_{ML} = \theta$)
2) Asymptotically Normal
$[\hat{\theta}_{ML} \overset{A}{\sim} N(\theta, \{I(\theta)\}^{-1})]$

4) Invariant: if $\hat{\theta}_{ML}$ is MLE for θ, then $g(\hat{\theta}_{ML})$ is MLE for $g(\theta)$
↳ eg. in normal example above, MLE for σ is square root of MLE for σ^2

3) Asymptotically Achieves Cramér-Rao Lower Bound for Consistent Estimators (i.e. is "best")

VII. Interval Estimation & Hypothesis Testing

A. Some Preliminary Facts We Use a lot in Interval Est. & Hypoth. Testing:

- Note: the process for hypoth. testing is to determine the distribution of a test statistic under the null hypothesis & see how "odd" the actual stat. is in the case at hand. For intervals, it's to see how wide an interval must be to be "reasonably sure to contain the true parameter" - more on this in a bit

1. As we know, in sampling X from a normal, with known mean μ & variance σ^2,

$$\frac{X-\mu}{\sigma} \sim N(0,1)$$

- this is exact, not an approximate distribution

2. When taking a sample average from a known $N(\mu, \sigma^2)$:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- also exact

3. When sampling X from an unknown distribution with known variance,

$$\frac{X-\mu}{\sigma} \not\to N(0,1)$$

- not exact, but converges to Normal

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

4. When sampling from a normal distribution with unknown variance:

$$\frac{X-\mu}{S} \sim t_{n-1} \xrightarrow{d} N(0,1)$$

$$(\bar{X}-\mu)/(S/\sqrt{n}) \sim t_{n-1} \xrightarrow{d} N(0,1)$$

5. When sampling from an unknown distribution with unknown parameters

$$\frac{X-\mu}{S} \xrightarrow{d} N(0,1)$$

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1)$$

upshot: Only $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ always $N(0,1)$ in the limit (by CLT) a question of how large sample must be for $N(0,1)$ to be reasonable approx. Under some circumstances we know the $\xrightarrow{d} N(0,1)$ happens along $\sqrt{n}$

6. $V = \sum_{i=1}^{n} X_i^2$ when each $X_i \sim N(0,1) \Rightarrow V \sim \chi^2_n$

- if (each $X_i \overset{A}{\sim} N(0,1)$) then $V \overset{A}{\sim} \chi^2_n$

7. $W = \left\{ (V_1/n_1) / (V_2/n_2) \right\}$ where $V_1 \sim \chi^2_{n_1}$ & $V_2 \sim \chi^2_{n_2} \Rightarrow W \sim F_{n_1, n_2}$

- if each $V \overset{A}{\sim} \chi^2_n$, then $W \overset{A}{\sim} F_{n_1, n_2}$

e.g.: $R^2 = 1 - (\Sigma e_i^2)/(\Sigma y_i^2)$

→ under certain conditions, $R^2 \sim F$

B. Constructing Confidence Intervals:

1. We start with some estimate (statistic) which we "standardize" or "normalize" (i.e. subtract mean & divide by standard deviation):

estimate → $\bar{x}$ $\quad \left\{ (\bar{x} - \mu) / (s/\sqrt{n}) \right\}$ ← standardization

$s/\sqrt{n}$ ← std dev. of $\bar{x}$ (estimated)

2. Then we ask what "bounds" would contain this with some probability, given its distribution (here t_{n-1} by (4) above); i.e., we look for z's that satisfy $\text{Prob}\left(-z < \frac{\bar{x}-\mu}{s/\sqrt{n}} < z\right) = (1-\alpha)$

2. $\text{Prob}\left(-z < \left\{\frac{\bar{x}-\mu}{s/\sqrt{n}}\right\} < z\right) = (1-\alpha)$

$(1-\alpha)$ → confidence level, e.g. $\alpha = .10$ ⇒ 90% confidence interval for $\{\cdot\}$ is $(-z, z)$

3. More usually, we are interested in an interval anchored on our estimate, $\bar{x}$ here, & going +/− on either side of that so that the interval might encompass the parameter we're trying to estimate.

So we can rearrange (2.):

$$\text{Prob}\left(\bar{x} - \frac{z \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{z \cdot s}{\sqrt{n}}\right) = (1-\alpha)$$

lower bound · parameter · upper bound · confidence level

Steps of that rearrangement:

$$\Pr\left(-z < \frac{\bar{x}-\mu}{s/\sqrt{n}} < z\right) = \Pr\left(-\frac{z \cdot s}{\sqrt{n}} < \bar{x} - \mu < \frac{z \cdot s}{\sqrt{n}}\right) = \Pr\left(-\frac{z \cdot s}{\sqrt{n}} - \bar{x} < -\mu < \frac{z \cdot s}{\sqrt{n}} - \bar{x}\right) = \Pr\left(\bar{x} - \frac{z \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{z \cdot s}{\sqrt{n}}\right)$$

4. Notes:

a. We're implicity assuming that we won't a symmetric confidence ~~level~~ interval
   i.) if distribution of the statistic is symmetric, then the symmetric C.i. is also the smallest c.i. from lower to upper bound
   ii.) if stat. is not symmetricly distributed, may be reasonable to get the smallest bound-range instead: this is complicated (& rarely done because of that, but we may try it later)

b. Notice that it is the bounds that vary across repeated samples in 3, not the parameter. (The "truth" doesn't vary.) Thus the confidence interval tells us that in x% of repeated samples, a C.i. so constructed would contain the true parameter. (Tortuously close to what we'd like to say, but not, huh? In Bayesian analysis, FYI, you can say exactly what you want to say, but that's beyond us right now.)

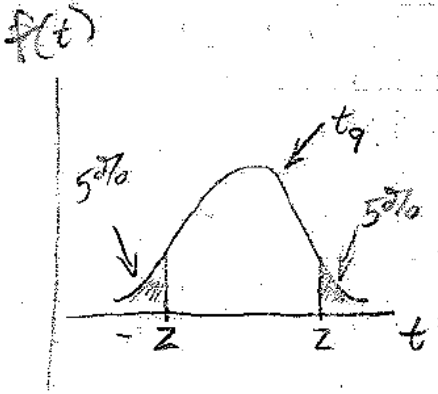
5. Mechanics (Example)

$\bar{X}$ taken from $x_i \sim$ i.i.d. $N(\mu, \sigma^2)$ $i = \{1, \ldots, 10\}$ (unknown)

$\bar{X} = .02$ $\quad s = .1$ from our sample estimates

$$Pr\left(\bar{x} - \frac{z \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{z \cdot s}{\sqrt{n}}\right) = (1 - \alpha)$$

$$Pr\left(.02 - z\frac{.1}{\sqrt{10}} < \mu < .02 + z\frac{.1}{\sqrt{10}}\right) = (1 - \alpha)$$



Suppose we want 90% c.i. (i.e., $\alpha = .1$), where does z come from? What is distrib. of $(\bar{x} - \mu)/(s/\sqrt{n})$? t_{n-1} or t_q here. t is symmetric, so we want 5% on each side $\Rightarrow$ z value of 1.833 from a t-table (or computer program)

$$\Rightarrow \left[.02 - 1.833 \cdot \frac{.1}{\sqrt{10}}, .02 + 1.833\frac{.1}{\sqrt{10}}\right]$$

$[-.038, .078]$ is the 90% c.i.

So, 90% of intervals likewise constructed across repeated samples would contain the true mean.

practical rule of thumb: when in doubt, use the t

6. **Final Note:** Often when the distribution of the sample statistic is asymptotically normal, we use the t distribution rather than jump right to using the normal as an approximation. The t is clearly more conservative than $N(0, 1)$, but we can't actually say this is warranted nonetheless, except in the limit (asymptotically).

VII. C. Hypothesis Testing:

(First we will stress the formal, rigid process of hypothesis testing & its interpretation; Then, & throughout the course, I will press a more intuitive approach & interpretation of hypothesis test results.)

1. The broadly described steps:

a. Formally the testing procedure begins by specifying a null hypothesis, H_0, (i.e. the default hypothesis & usually the thing we would like to "disprove"), and an alternative hypothesis, H_A (the thing we would like to "prove") (or H_1).

- by far the most common hypotheses are $H_0: \beta = c$ (c some constant, usually 0; β: some parameter of interest)

b. Then a "size" or "α-level" or "p-level" for the test is pre-determined.

$H_1: \beta \neq c$

c. Then a test statistic is computed. The test-stat. is some transformation of the estimate of the parameter the distribution of which statistic we know or can approximate under the condition that the null is true. The stat is then checked on this distribution against the pre-determined α.

2. Type I & Type II Errors, Size, Power, Bias & Consistency of Tests

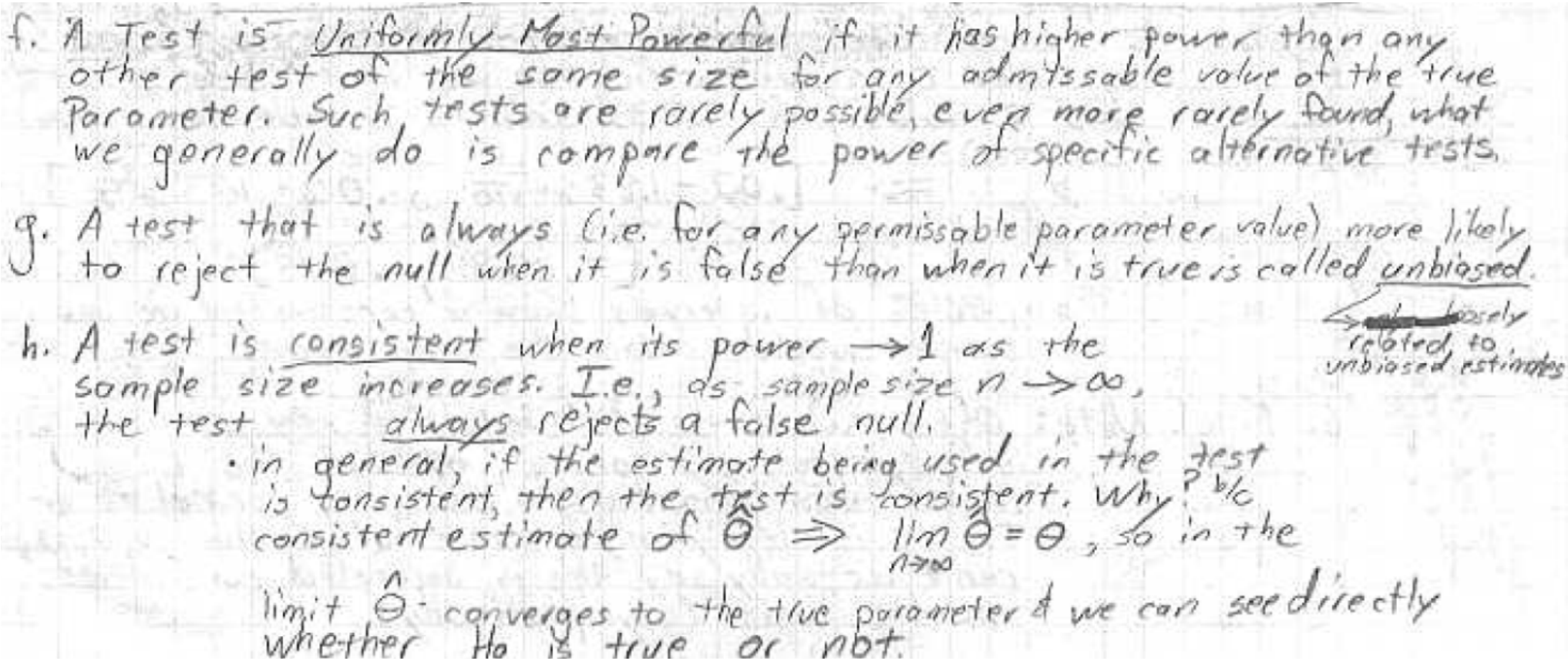
a. Type I Error: Rejection of Null Hypothesis when it is in fact true.
b. Type II Error: Failure to reject the Null Hypothesis when it is in fact false.
(Type III Error: confusing a Type I & Type II error -- (a geek joke))

c. The probability of a Type I error is the size (α-level, p-level, significance level) of a test.

d. The probability that a test will correctly reject the null, i.e. 1−Prob(Type II Error), (call this β) is the power of a test.

e. In general, if we are using all available info. correctly in estimating, & we have correctly constructed the test to make best use of this information & estimate, then there is no way to make Type I errors less likely without making Type II errors more likely, and vice versa.

Example: if the criminal justice system is efficiently using available evidence, there is no way to make conviction of criminals more likely without also making conviction of the innocent more likely & v.v., there is no way to make conviction of innocent less likely without also making conviction of guilty less likely.
Similarly, for fixed efficiency ⇒ barring changes in efficiency, Miranda Rights, for example, leads to more innocent going free and more guilty going free.



Additional notes:

f. "…for specific (ranges of) parameter values."

g. As with h. regarding consistency: essentially, if estimate unbiased, and verdict directly related to estimate, then test unbiased.

3 Mechanics of Conducting Hypothesis Tests:

a. $H_0: \theta = \theta_0$ Example: $H_0: \beta = 0$
$H_1: \theta \neq \theta_0$ $H_1: \beta \neq 0$

b. Convert Estimate, $\hat{\theta}$, to standardized form, as in conf. intervals. → under null being true

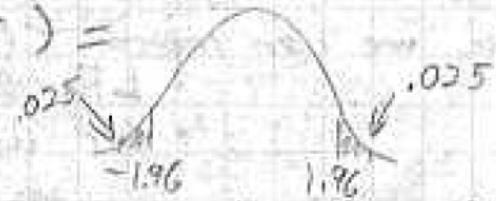
$$T = \left\{ \frac{\hat{\theta} - \theta_0}{\text{std.err.}(\hat{\theta})} \right\} \overset{A}{\sim} N(0,1)$$

Example: $T = \frac{\hat{\beta} - 0}{se(\hat{\beta})} \overset{A}{\sim} N(0,1)$ → as before, we usually use t_{n-1} to approximate rather than the asymptotic distribution $N(0,1)$

or $T = \hat{\beta} / se(\hat{\beta})$

c. Check that test statistic, t, against its distribution or approximate distribution, here t_{n-1}, record the probability of T's more extreme (greater in absolute value) than your actual T

Example $T = 1.96$, distribution t_{1000},

Prob.$(|T| \geq 1.96) =$  $= .05$

d. This is your p-level. It is the probability of having observed a test statistic, here T, this far or farther from the null, here 0, if the null were in fact true.

- If the p-level is $\leq$ the predetermined size or α-level of the test, we reject the null (in favor of the alternative).

- note the 1-for-1 correspondence of hypoth. testing & confidence intervals: if reject, then c.i. did not contain null and vice versa.

e. Interpretational Note:
- The p-level does not tell us the probability the null is true (in non-Bayesian analysis). The null is either true (with probability 1) or not true (with probability 1). The p-level tells us the probability we would have observed a test stat. this far or farther from the null if the null were true. From this we infer that the null was likely untrue or not; the facts, however, are that the null either is or is not true (with p=1)

↑ n.b. we use likely not probably

f. Rob's Rules: the standard procedure is to set the size of the test, α-level, to .10 or .05 or .01 & then star anything that makes this or lower p-level. That's fine, but with modern tech., it's simple (the computer'll give it) to know the "exact" p-level. This is more info., we have the info, therefore give it.

↳ under distributional assumptions/approximations

g. The Broad (& Better by Rob's Rules) Interpretation:
- The question is, "Is the estimated quantity, e.g. the relationship between X & y, 'far' from the null, e.g. 0?"
- "Far" is defined in standard deviation units & how "significant" any increment of "far"ness is, i.e. how tortuous the straining of credulity given the evidence, i.e. how "long" a standard deviation is, is measured in increments of likelihood,
- Thus, p-levels tell us just how far we're straining credulity given the evidence of what has happened if we cling to the null. The lower the p-level, the harder is the stretch of imagination required to continue believing the null to swallow.

VII. D. Three Common Types of Tests in Regression Analysis

1. Think of hypotheses regarding (a) parameter(s), θ, such as

$H_0: c(\theta) = 0$

↳ some function of theta

$H_1: c(\theta) \neq 0$

2. Likelihood-Ratio (LR) Tests:

"Degradation of Fit" tests: LR, ΔR^2, ΔSSE

a) if $c(\underset{\sim}{\theta}) = \underset{\sim}{q}$ is true (constants), then imposing it to begin with (prior to estimation) should make little difference to the likelihood of the data (or fit of the regression)

b) ∴, we ought to be able to base a test on $L(\underset{\sim}{\theta})$ ← $c(\underset{\sim}{\theta}) = \underset{\sim}{q}$ not imposed / $L_R(\underset{\sim}{\theta})$ ← imposed

or, equivalently, $\ln L(\underset{\sim}{\theta}) - \ln L_R(\underset{\sim}{\theta})$ (just applying ln rules)

c) in fact, under certain conditions we won't get into now

$$LR = 2 \cdot [\ln L(\underset{\sim}{\theta}) - \ln L_R(\underset{\sim}{\theta})] \overset{A}{\sim} \chi^2_n$$

(n = number of conditions given by $c(\underset{\sim}{\theta}) = \underset{\sim}{q}$)

d) So, we can reject $c(\underset{\sim}{\theta}) = \underset{\sim}{q}$ if
- LR larger than χ^2_n for some predetermined α
- or we can report the p-level of the test as before

3. Wald Tests

a) if $c(\underset{\sim}{\theta}) = \underset{\sim}{q}$ (constant(s)) true, then $c(\hat{\underset{\sim}{\theta}})$ should not be too "far" from $\underset{\sim}{q}$ a "good" estimate, such as the MLE

b) ∴ we ought to be able to find a test based on $c(\hat{\underset{\sim}{\theta}}) - \underset{\sim}{q}$ not being too far from zero

c) in fact, under certain conditions we won't go into now:

$$W = \{(c(\hat{\underset{\sim}{\theta}}) - \underset{\sim}{q})' (\mathrm{Var}[c(\hat{\underset{\sim}{\theta}}) - \underset{\sim}{q}])^{-1} (c(\hat{\underset{\sim}{\theta}}) - \underset{\sim}{q})\}$$

$$W \sim \chi^2_n$$

← again, number of conditions given by $c(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{q}$

d) reject or not, or report p-level from χ^2_n exactly as before

4. Lagrange Multiplier Tests:

a) if $c(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{q}$ true, then imposing restraint should not have much effect in our optimizing the likelihood. In particular, the first-order condition that $\frac{\partial L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = 0$ should not be "far" from holding.

b) ∴ should be able to base a test on $\left\{\left[\partial \ln L_R(\underset{\sim}{\theta})\right] / \partial \underset{\sim}{\theta}\right\}$ with $c(\underset{\sim}{\theta}) = \underset{\sim}{q}$ imposed being "near" zero.

c) in fact, under conditions we're not going into, etc. etc.

$$LM = \left[\frac{\partial \ln L_R(\hat{\underset{\sim}{\theta}}_R)}{\partial \underset{\sim}{\theta}_R}\right]' \left[I(\hat{\theta}_R)\right]^{-1} \left[\frac{\partial \ln L_R(\hat{\underset{\sim}{\theta}}_R)}{\partial \underset{\sim}{\theta}_R}\right]$$

↑ info matrix from Cramér Rao

$$LM \sim \chi^2_n$$ ← number of conditions again.

And, as mentioned previously, can sometimes also conduct LM tests by clever auxiliary regression in which coefficients are the Lagrangian multipliers for the constraints.