Big Ideas in Stat 350: Sahar Zangeneh

Big Ideas – Numerical Summaries
- Different types of data lead to different types of summaries/plots
  - For quantitative (numerical) data, consider using a histogram, boxplot, or 5-number summary.
  - For qualitative (descriptive) data consider using a frequency table or bar chart.

Big Ideas - Probability
- Be careful to identify the events you are considering.
  - When using abbreviated notation, CLEARLY denote what A, B, etc. mean.
- Simply giving the definitions is not enough!
  - Use proper, well defined notation and show some (numerical) work.

Big Ideas – Random Variables
- Show all work (completely!)
- Remember to include the units of measurement, when appropriate.
- When working with a continuous probability distribution, always draw the picture!

Big Ideas – Standard Deviation Interpretation
- The empirical rule for bell-shaped data is NOT (!!!!) the definition of SD
  - This is a property of bell-shaped distributions only
  - You can be asked to interpret SD even if the distribution is not bell-shaped

Key Elements to SD Interpretation
- "average distance"
  - Need at least 2 quantities for there to be a distance between them
    - The individual observations
    - The mean
  - "roughly" or "approximately" or "about"

Good Examples – SD Interpretation
- "Roughly, the number of hours a student slept last night was an average distance of 1.94658 hours from the mean."
- "On average, the hours of sleep per night for college students is roughly 1.9 hours away from the mean time of 7.1 hours."
Big Ideas – Time Series Plots

- Look at when data is collected over time
- Overall stability = stable mean + stable variance
- Be careful with your language:
  - There is evidence of an increasing mean
  - NOT: The mean is increasing
- If there appears to be a pattern over time, it may not be appropriate to treat the data as a random sample

Big Ideas – QQ-Plots

- Departure from straight line = non-normal data
  - Just need to see a departure, don’t need to say what type
- Make both a histogram and a QQ-plot:
  - Histogram gives general shape of distribution (bell-shaped; skewed; etc.)
  - QQ-plot gives normality of distribution

Big Ideas – Sampling Distribution for the Proportion (See page 74 of notes for details)

- Theory relies on idea of repeated sampling
  - Take all possible random samples of size n and compute the sample proportion for each sample
  - For a LARGE sample size n, the distribution of the sample proportion will be APPROXIMATELY a NORMAL distribution with a mean of p and a standard deviation of \( \sqrt{\frac{p(1-p)}{n}} \)
- Notation:
  \[ \hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \]

Big Ideas – Sampling Distribution for the Mean

- Theory relies on idea of repeated sampling
  - Take all possible random samples of size n and compute the sample mean for each sample
  - If the parent population is a normal distribution with mean \( \mu \) and standard deviation \( \sigma \), then for any sample size (small or large), the sample mean will have a NORMAL distribution with a mean of \( \mu \) and a standard deviation of \( \frac{\sigma}{\sqrt{n}} \)
- Notation:
  \[ X \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \]

Big Ideas – CLT

- Theory relies on idea of repeated sampling
  - Take all possible random samples of size n and compute the sample mean for each sample
  - If the parent population is NOT a normal distribution but with mean \( \mu \) and standard deviation \( \sigma \), then for a large sample size, the sample mean will have approximately a NORMAL distribution with a mean of \( \mu \) and a standard deviation of \( \frac{\sigma}{\sqrt{n}} \)
- Notation:
  \[ X \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \]

Big Ideas – Confidence Intervals

- Provide a method of stating both how close the value of a statistic is likely to be to the value of a parameter and the accuracy of it being that close
- Basic Structure:
  \[ \text{estimate} \pm \text{error} \]
- You choose the multiplier to get the desired level
  - For (1-\( \alpha \))% CI for population proportion choose \( Z_{\alpha/2} \) multiplier (find it in the last row of table A.2 where df=\( \infty \))
**Big Ideas – Confidence Intervals**

- **Principles for using Confidence Intervals to Guide Decision Making**:
  - **Principle 1**: A value not in a CI can be rejected as possible value of the population parameter.
  - **Principle 2**: When the CIs for parameters for two different populations do not overlap, it is reasonable to conclude that the parameters for the two populations are different.

**Confidence Interval for a Population Proportion Summary**

- **General** Confidence Interval for the Population Proportion $p$:
  \[
  \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
  \]

- **Approximate 95%** Confidence Interval for the Population Proportion $p$:
  \[
  \hat{p} \pm 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
  \]

  (Standard error for $\hat{p}$ is largest when $\hat{p}=0.5$)

- **Conservative** Confidence Interval for the Population Proportion $p$:
  \[
  \hat{p} \pm \frac{z^*}{\sqrt{2n}}
  \]

**Big Ideas – Hypothesis Testing**

- **Basic 5 steps to hypothesis testing are the same no matter what the scenario**
  - Parameters, assumptions, and test statistic change depending on the scenario
  - Result of differences in structure of data

**5 Inference Scenarios**

1. 1-sample inference for pop. proportion $p$
2. 1-sample inference for pop. mean $\mu$
3. Paired samples inference for a population mean difference $\mu_0$
4. 2 indep. samples inference for the difference between 2 pop. means $\mu_1 - \mu_2$
5. 2 indep. samples inference for the difference between 2 pop. prop. $p_1 - p_2$

**5 Steps in Hypothesis Testing**

1. State $H_0$ and $H_a$
2. Check assumptions and calculate test statistic
3. Find p-value
4. Make decision
5. State conclusion
Step 1 – Basic Truths

- Use the same population parameter in both $H_0$ and $H_a$.
- Use the same null value in both $H_0$ and $H_a$.
  - The number the researchers want to investigate, or 0 if the parameter involves a difference.
- ALWAYS have equality in $H_0$ and NEVER have equality in $H_a$.
  - Sign used in $H_a$ depends on what the researchers want to investigate.

Step 1 – Determining the Scenario

- Good questions to ask yourself:
  - Mean (quantitative data) or proportion (categorical data)?
  - One group or two?
  - If means, two groups: paired or independent?

Step 2 - Assumptions

1. $p$: The data are a random sample and the sample size is large ($np \geq 10$, $n(1-p) \geq 10$; for large sample $z$-test)
2. $\mu$: The data are a random sample from a normal population
3. $\mu_2$: The differences are a random sample from a normal population of differences
4. $\mu_1 - \mu_2$: Each sample is a random sample from a normal population, the 2 samples are independent, and the pop. standard deviations are equal (for pooled version)
5. $p_1 - p_2$: Each sample is a random sample and each sample size is large

Example: $z$-test for $p$

1. $H_0: p = p_0$ vs. $H_a: p \neq p_0$
   - or $H_a: p > p_0$
   - or $H_a: p < p_0$
2. Verify assumptions
   - Random sample
   - Large sample: $np_0 \geq 10$ and $n(1-p_0) \geq 10$
   - and calculate $z$-statistic

Example: $z$-test for $p$

3. Find $p$-value
   - Assuming $H_o$ is true, the $p$-value is the probability of observing a $z$-statistic as extreme or more extreme than what we actually saw.
   - $H_a: p \neq p_0 \Rightarrow 2P(Z \leq -z)$
   - $H_a: p > p_0 \Rightarrow P(Z \geq z)$
   - $H_a: p < p_0 \Rightarrow P(Z \leq z)$

4. Decision
   - $p$-value $\leq \alpha \Rightarrow$ Reject $H_o$
   - $p$-value $> \alpha \Rightarrow$ Fail to Reject $H_o$

5. Conclusion
   - Reject $H_o \Rightarrow "There is sufficient evidence that $H_a$ is true."
   - Fail to Reject $H_o \Rightarrow "There is insufficient evidence that $H_a$ is true."
Big Ideas - ANOVA

- Extension of the 2-indep. Samples t-test for 3 or more groups
  - Now the assumption of equal population variances is required
  - IF there is evidence of a difference in group means, Tukey’s MC can help determine which means are different

Big Ideas - Regression

- Looks for linear relationship between 2 quantitative variables
  - Scatterplot of Y vs. X shows
    - Form of relationship (linear or not)
    - Direction (positive or negative)
    - Strength (strong or weak)
    - Outliers (apparent or not)

Big Ideas - Regression

- Scatterplot of residuals vs. X checks assumption of constant standard deviation over the range of X
  - Look for random scatter of points around zero

Big Ideas - Chi-square tests

- For categorical variables only
  - Three tests
    - Same assumptions / test statistic
    - Different hypotheses
    - Different definitions for the expected cell count (E)

Big Ideas - Chi-square tests

1. Goodness of fit: tests population proportions for more than 2 groups
2. Homogeneity: 1 response, 2 populations
   - H0: response has same distribution in each pop.
   - Ha: response does not have same distribution
3. Independence: 1 population, 2 responses
   - H0: 2 variables are not related for the pop.
   - Ha: 2 variables are related for the pop.

Different Type of Errors

<table>
<thead>
<tr>
<th>Truth</th>
<th>Reject $H_0$</th>
<th>Fail to reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>Type I error</td>
<td>😞</td>
</tr>
<tr>
<td>$H_a$ true</td>
<td>😊</td>
<td>Type II error</td>
</tr>
</tbody>
</table>