Statistics 111 - Lecture 27
Final review

Administrative Notes

- Final Exam is **Tuesday, May 10th (3-5pm)**
  - Covers Chapters 1-8 and 10 in textbook
  - Bring ID cards to final!
  - Allowed: Calculators, **double-sided 8.5 x 11 cheat sheet**
- Exam Rooms:
<table>
<thead>
<tr>
<th>Stat 111 Lecture</th>
<th>Last Name</th>
<th>Final Exam Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>11am – 12pm</td>
<td>Everyone</td>
<td>MEYERSON HALL B1</td>
</tr>
<tr>
<td>2 – 3pm</td>
<td>Everyone</td>
<td>COHEN HALL G17</td>
</tr>
</tbody>
</table>

- Office hours will be held throughout the exam period up until the final exam on May 10th
- List of additional textbook study problems from second half of the course will be also be posted on the course website

Outline

- Collecting Data (Chapter 3)
- Exploring Data - One variable (Chapter 1)
- Exploring Data - Two variables (Chapter 2)
- Probability (Chapter 4)
- Sampling Distributions (Chapter 5)
- Introduction to Inference (Chapter 6)
- Inference for Population Means (Chapter 7)
- Inference for Population Proportions (Chapter 8)
- Inference for Regression (Chapter 10)
- Urban Analytics Case Study

Experiments

- Try to establish the **causal effect** of a treatment
- Key is reducing presence of **confounding** variables
- **Matching**: ensure treatment/control groups are very similar on observed variables eg. race, gender, age
- **Randomization**: randomly dividing into treatment or control leads to groups that are similar on observed and unobserved confounding variables
- **Double-Blinding**: both subjects and evaluators don’t know who is in treatment group vs. control group

Sampling and Surveys

- Just like in experiments, we must be cautious of potential sources of bias in our sampling results
- Voluntary response samples, undercoverage, non-response, untrue-response, wording of questions
- **Simple Random Sampling**: less biased since each individual in the population has an equal chance of being included in the sample
Different Types of Graphs
• A **distribution** describes what values a variable takes and how frequently these values occur
• Boxplots are good for **center**, **spread**, and **outliers** but don’t indicate shape of a distribution
• Histograms much more effective at displaying the **shape** of a distribution

Measures of Center and Spread
• **Center:** Mean
  \[ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \]
  
  • **Spread:** Standard Deviation
    \[ s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \]
  
    • For outliers or asymmetry, median/IQR are better
      - **Center:** Median - “middle number in distribution”
      - **Spread:** Inter-Quartile Range \( IQR = Q3 - Q1 \)
    
    • We use mean and SD more since most distributions are symmetric with no outliers (eg. Normal)

Relationships between continuous var.
• Scatterplot examines relationship between **response** variable \( Y \) and a **explanatory** variable \( X \):
  
  - Positive vs. negative associations
  - Correlation is a measure of the strength of linear relationship between variables \( X \) and \( Y \)
    - \( r \) near 1 or -1 means strong linear relationship
    - \( r \) near 0 means weak linear relationship
  - Linear Regression: come back to later…

Probability
• **Random** process: outcome not known exactly, but have **probability distribution** of possible outcomes
  
  • **Event:** outcome of random process with prob. \( P(A) \)
  
  • **Probability calculations:** combinations of rules
    - Equally likely outcomes rule
    - Complement rule
    - Additive rule for disjoint events
    - Multiplication rule for independent events
  
  • **Random variable:** a numerical outcome or summary of a random process
    - Discrete r.v. has a finite number of distinct values
    - Continuous r.v. has a non-countable number of values
    - Linear transformations of variables

The Normal Distribution
• The Normal distribution has center \( \mu \) and spread \( \sigma^2 \)
  
  • Have tables for any probability from the standard normal distribution \( \mu = 0 \) and \( \sigma^2 = 1 \)
  
  • Standardization: converting \( X \) which has a \( N(\mu, \sigma^2) \) distribution to \( Z \) which has a \( N(0,1) \) distribution:
  \[ Z = \frac{X - \mu}{\sigma} \]

  • **Reverse standardization:** converting a standard normal \( Z \) into a non-standard normal \( X \)
  \[ X = \sigma \cdot Z + \mu \]

Inference using Samples
• Continuous: pop. mean estimated by sample mean
  - **Sampling Distributions:** Distribution of values taken by statistic in all possible samples from the same population
  
  • Discrete: pop. proportion estimated by sample proportion
  
  • Key for inference: **Sampling Distributions**
  
  • Continuous: pop. mean estimated by sample mean
    - **Sampling Distributions:** Distribution of values taken by statistic in all possible samples from the same population
  
  • Discrete: pop. proportion estimated by sample proportion
    - **Sampling Distributions:** Distribution of values taken by statistic in all possible samples from the same population
Sampling Distribution of Sample Mean

- The center of the sampling distribution of the sample mean is the population mean: $	ext{E}(ar{X}) = \mu$
- Over all samples, the sample mean will, on average, be equal to the population mean (no guarantees for 1 sample)
- The standard deviation of the sampling distribution of the sample mean is
  \[
  \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}
  \]
- As sample size increases, standard deviation of the sample mean decreases!
- Central Limit Theorem: if the sample size is large enough, then the sample mean $\bar{X}$ has an approximately Normal distribution

Binomial/Normal Dist. For Proportions

- Sample count $Y$ follows Binomial distribution which we can calculate from Binomial tables in small samples
- If the sample size is large ($np$ and $n(1-p)$ ≥ 10), sample count $Y$ follows a Normal distribution:
  \[
  \text{mean}(Y) = np
  \]
  \[
  \text{SD}(Y) = \sqrt{np(1-p)}
  \]
- If the sample size is large, the sample proportion also approximately follows a Normal distribution:
  \[
  \text{mean}(\hat{p}) = p
  \]
  \[
  \text{SD}(\hat{p}) = \frac{p(1-p)}{n}
  \]

Summary of Sampling Distribution

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Unknown Parameter</th>
<th>Statistic</th>
<th>Variability of Statistic</th>
<th>Distribution of Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>$\mu$</td>
<td>$\bar{X}$</td>
<td>$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$</td>
<td>Normal (if $n$ large)</td>
</tr>
<tr>
<td>Count $X_i = 0$ or $1$</td>
<td>$p$</td>
<td>$\hat{p}$</td>
<td>$\text{SD}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$</td>
<td>Binomial (if $n$ small)</td>
</tr>
</tbody>
</table>

Introduction to Inference

- Use sample estimate as center of a confidence interval of likely values for population parameter
- All confidence intervals have the same form:
  \[
  \text{Estimate} \pm \text{Margin of Error}
  \]
- The margin of error is always some multiple of the standard deviation (or standard error) of statistic
- Hypothesis test: data supports specific hypothesis?
  1. Formulate your Null and Alternative Hypotheses
  2. Calculate the test statistic: difference between data and your null hypothesis
  3. Find the p-value for the test statistic: how probable is your data if the null hypothesis is true?

Inference: Single Population Mean $\mu$

- Known SD $\sigma$: confidence intervals and test statistics involve standard deviation and normal critical values
  \[
  \left(\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)
  \]
  \[
  Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}
  \]
- Unknown SD $\sigma$: confidence intervals and test statistics involve standard error and critical values from a $t$ distribution with $n-1$ degrees of freedom
  \[
  
  \left(\bar{X} - t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}\right)
  \]
  \[
  T = \frac{\bar{X} - \mu}{s / \sqrt{n}}
  \]
- $t$ distribution has wider tails (more conservative)

Inference: Comparing Means $\mu_1$ and $\mu_2$

- Known $\sigma_1$ and $\sigma_2$: two-sample $Z$ statistic uses normal distribution
  \[
  Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}
  \]
- Unknown $\sigma_1$ and $\sigma_2$: two-sample $T$ statistic uses $t$ distribution with $\min(n_1-1, n_2-1)$ degrees of freedom
  \[
  T = \frac{\bar{X}_1 - \bar{X}_2}{s_p \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
  \]
  \[
  s_p^2 = \frac{\left( n_1 - 1 \right) s_1^2 + \left( n_2 - 1 \right) s_2^2}{n_1 + n_2 - 2}
  \]
- Matched pairs: instead of difference of two samples $X_1$ and $X_2$, do a one-sample test on the difference $d$
  \[
  T = \frac{\bar{X} - 0}{s' / \sqrt{n'}}
  \]
  \[
  s'^2 = \frac{\left( n_1 - 1 \right) s_1^2}{n_1}
  \]
Inference: Population Proportion $p$

- **Confidence interval** for $p$ uses the Normal distribution and the *sample proportion*:
  \[
  \hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
  \]

- **Hypothesis test** for $p = p_0$ also uses the Normal distribution and the *sample proportion*:
  \[
  Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}
  \]

Inference: Comparing Proportions $p_1$ and $p_2$

- **Hypothesis test** for $p_1 - p_2 = 0$ uses Normal distribution and complicated test statistic
  \[
  Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}
  \]
  \[
  Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}
  \]

  *Pooled standard error*:
  \[
  SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}
  \]

- **Confidence interval** for $p_1 = p_2$ also uses Normal distribution and sample proportions
  \[
  \left[\hat{p}_1 - Z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + Z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right]
  \]

Linear Regression

- Use *best fit line* to summarize linear relationship between two continuous variables $X$ and $Y$:
  \[
  Y_i = \alpha + \beta \cdot X_i
  \]
  \[
  Y_i = \alpha + \beta \cdot X_i
  \]
- The slope ($\beta = \frac{\sum X - \bar{X} \cdot \sum Y - \bar{Y}}{\sum X^2 - n \bar{X}^2}$): average change you get in the $Y$ variable if you increased the $X$ variable by one
- The intercept ($\alpha = \bar{Y} - \bar{X} \cdot \beta$): average value of the $Y$ variable when the $X$ variable is equal to zero
- Linear equation can be used to predict response variable $Y$ for a value of our explanatory variable $X$

Significance in Linear Regression

- Does the regression line show a significant linear relationship between the two variables? $H_0: \beta = 0$ versus $H_1: \beta \neq 0$
- Uses the *t distribution* with $n-2$ degrees of freedom and a test statistic calculated from JMP output
  \[
  T = \frac{b}{SE(b)}
  \]
- Can also calculate *confidence intervals* using JMP output and $t$ distribution with $n-2$ degrees of freedom
  \[
  \left(b \pm t \cdot SE(b)\right) \quad \left(a \pm t \cdot SE(a)\right)
  \]

Urban Analytics in Philadelphia

- Quantitative analysis of the economic and social functioning of local areas within large cities
- Philadelphia is an interesting case study for contemporary issues in urban revival and gentrification
- Creating empirical measures for concepts like *urban vibrancy* that have been difficult to quantify
- Examined associations between crime, poverty, demographics and land use
- It is important to do quantitative analysis of large cities carefully and at the correct level of resolution
  \[
  \text{What we see when we look at the city in the aggregate can be quite different than specific neighborhoods}
  \]
- Both sides of the classic Jane Jacobs vs. Urban renewal fight were based on empirical arguments
- Jacobs’ key innovation was basing her observations at a high resolution: individual streets and blocks rather than aggregating over entire cities
Last Class!

- Thanks everyone for a great semester!
- See you on May 10th for the final exam!