The University of Texas at Austin
McCombs School of Business

Final Exam Review

What's included in the Final?

The final will be on Wednesday, May 13, 7 pm

2 hour exam (probably 25 questions)

- Material since Test 2 - about 50%
  - Chapter 12 Testing Hypotheses
  - Chapter 13 More about Tests and Intervals
  - Chapter 14 Comparing Two Means
  - Chapter 16 Inference for Regression
  - Chapter 18 Multiple Regression
- Test 1 and Test 2 material - about 50%

Ground Rules:

- Test will be on Zoom
- Open book/open notes
- You must stay on video and have Zoom controller or Chat window open so that host may contact you if there is a problem with your video
- You must show your work in Rscript for any problem requiring calculations

Final covers chapters 1 - 8, 10 - 14, 16, and 18

Skip the following:

- Transforming Skewed Data (Section 3.11)
- Regression to the mean and non-linear relationships (in Chapter 4)
- Probability trees and Bayes' rule (in Chapter 5)
- The Binomial formula; Uniform, Geometric, and Poisson distributions (in Chapter 6)
- Continuity correction and 7.6 (in Chapter 7)
- Stratified, Cluster, Multistage, and Systematic Samples (in Chapter 8)
- Bootstrapping (in Chapter 11)
- Bootstrap Hypothesis Tests and Intervals (in Chapter 12)
- Critical Values (Section 13.3)
- The Pooled t-Test (Section 14.5)
- Sections 16.3 and 16.4
- Adjusted R^2 and 18.6 (in Chapter 18)

Suggestions

- Work problems (**especially from textbook**)
- Review the "What have I learned?" sections at the end of each chapter
- Review problems on Test 1 and Test 2
- Make notes (even though test is open book)
- Work the Sample Final Exam (using your notes)
- Don’t stay up all night!!

Data

- Variables and Cases
- Variables can be categorical or quantitative
  Quantitative data can be nominal or ordinal
  Data can be cross sectional or time series
- Distributions describe the values of the variable and how often they occur

Categorical variables

- Graph with bar charts and pie charts
  `plot` or `pie`

- Contingency tables
  `table`

- Joint, marginal, and conditional distributions
  `prop.table` and `margin.table`
  `barplot` and `mosaicplot`

Quantitative Data

- Histograms, stemplots, and time series plots
  `hist`, `stem`, and `plot( data$x, type = "l")`
- Look for overall pattern and deviations from that pattern
- **Describe:** center, spread, and shape
  - Symmetric or skewed
  - Outliers

Numerical Descriptions

- **Center:** mean and median
- **Spread or variability:**
  - Range
  - Quartiles and IQR
  - Variance
  - Standard deviation

`mean`, `median`, `sd`

Numerical Descriptions

- Five number summary
  `fivenum`
- Boxplots
  `boxplot`

Examining Relationships

- Explanatory and response variables

  Independent and dependent variables

- Scatterplots `plot(x,y)`
  - Positive or negative association
  - Outliers
  - Linear patterns

- Correlation `cor`
  - Measures strength and direction of a linear relationship
  - $-1 \leq r \leq +1$
  - $r = \pm 1$ only for perfect linear relationships
  - Correlation does not imply a cause and effect relationship

Regression

- Least squares regression `lm(y ~ x)`
- Regression coefficients and their interpretation
- Standard error of the estimate
- Percentage of variation explained: R^2

Randomness and Probability

- **Random -** individual outcomes are uncertain but there is a regular distribution of outcomes in the long term.
- Probability of a random phenomenon
- Empirical (relative frequency) probabilities
  Personal or subjective probabilities

Probability Models

- Sample Spaces
- Probability Rules
  - $0 \leq P(A) \leq 1$ for any event A
  - $P(S) = 1$
  - $P(A^C) = 1 - P(A)$
  - Addition rule for disjoint events
    General addition rule
  - Multiplication rule for independent events
    General multiplication rule
- Discrete vs Continuous models

General Probability Rules

- Additional rule for disjoint events
  P(at least one of events A, B, C,... occurs)
  = P(A) + P(B) + P(C) + ...

- General addition rule
  P(A or B) = P(A) + P(B) - P(A and B)

- Multiplication rule for independent events
  P(A and B) = P(A)P(B)

- General multiplication rule
  P(A and B) = P(A)P(B|A)
  Conditional probability: $P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$

Random Variables

- Mean

$$\mu_X = x_1 P_1 + x_2 P_2 + \ldots + x_k P_k$$

- Variance

$$\sigma_X{}^2 = (x_1 - \mu_X)^2 P_1 + (x_2 - \mu_X)^2 P_2 + \ldots + (x_k - \mu_X)^2 P_k$$

- Standard Deviation

$$\sigma_X = \sqrt{\sigma_X{}^2}$$

Adding and subtracting random variables:

$E(X \pm Y) = E(X) \pm E(Y)$

$Var(X \pm Y) = V(X) + Var(Y)$ (if X and Y are **independent**)

Bernoulli

B is the number of successes from one trial
where p is the probability of success
$E[B] = p$ and $Var(B) = p(1 - p)$

Binomial

X is the number of successes from n independent trials
where p is the probability of success
$E[X] = np$ and $Var(X) = np(1 - p)$

Sample Proportion

$\hat{p} = \frac{X}{n}$ is the proportion of successes from n independent trials
where p is the probability of success
$\mathrm{E}[\hat{p}] = \frac{E[X]}{n} = p$ and $Var(\hat{p}) = \frac{Var(X)}{n^2} = \frac{p(1-p)}{n}$

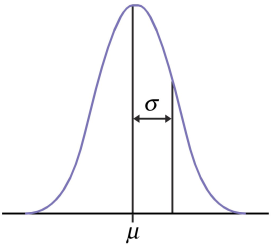
We did not cover Uniform, Geometric, or Poisson dist

Binomial Distributions

- The total number of observations *n* **is fixed** in advance.
- The outcomes of all *n* observations are statistically **independent.**
- Each observation falls into just one of 2 categories: **success** and **failure.**
- Same **probability of success** for each trial

We did not cover calculating Binomial probabilities

Normal Distribution

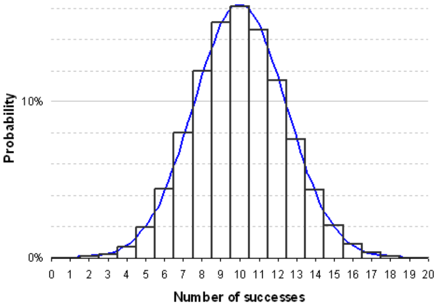


- 68 - 95 - 99.7 rule
- Standardized observations
- Normal distribution calculations:
  - Find areas - Forward Calculations
  - Find a value when given an area - Backward Calculations

Calculate probabilities for the Normal distribution:

- `pnorm(x, mean, standard_dev)` finds areas to the left of x
- `qnorm(probability, mean, standard_dev)` finds the value with the specified probability to the left
- Use `hist` and `qqnorm` to check if data is normal
- For calculations about the sample mean, $\bar{x}$, use $\frac{s}{\sqrt{n}}$ for standard_dev

Normal approximation for the Binomial



If $np \geq 10$ and $n(1-p) \geq 10$ the Binomial distribution is approximately Normal with

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

We did not cover Uniform or Exponential distributions

Sampling

- Populations
  Parameters describe populations

- Samples
  Statistics describe samples

- We hope that sample data is representative of the population

- Sampling variability - sample to sample differences (also called sampling error)

- Non-sampling errors - due to voluntary response, non-response, poorly worded questions, etc.

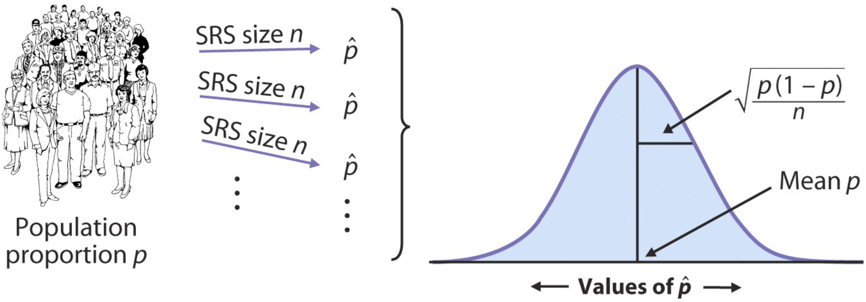
Big Ideas

- Sample - Examine a Part of the whole
- Randomize - to avoid bias
- The Sample Size is what matters

Possible causes of Bias

- Voluntary response samples
- Under coverage
- Non-response
- Behavior or appearance of interviewer
- Poorly worded questions
- Interviewer fabrications

Sampling Distribution of $\hat{p}$



The mean of the sampling distribution is p

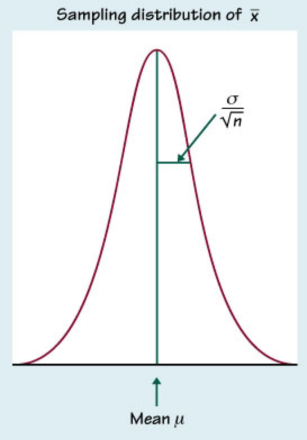
The standard deviation is $\sqrt{\frac{p(1-p)}{n}}$

Since p is unknown,

we use the standard error instead which is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Sampling Distribution of $\bar{x}$

- If x_i has mean μ and standard deviation σ, $\bar{x} = \sum \frac{x_1+\cdots+x_n}{n}$ has mean μ
- If the x_is are independent, the standard deviation $= \frac{\sigma}{\sqrt{n}}$
- Since σ is typically unknown, it will be estimated with the sample standard deviation, s
- The **standard error is** $\frac{s}{\sqrt{n}}$



Statistical Inference

- Methods for drawing conclusions about a population from sample data are called statistical inference

- Methods
  1. **Confidence Intervals** – estimating a value of a population parameter
  2. **Tests of significance** – assess evidence for a claim about a population

Checklist for Inference

- **Independence:**

  The sampled values must be independent of each other.
- **Randomization:**

  The sample is a simple random sample from the relevant population.
- **10% condition:**

  Sample size is less than 10% of the population size.
- **Sample size condition:**

  Success\failure condition for proportions

  Nearly normal condition for means

Specific sample size conditions:

- **Confidence Interval for proportion:**
  Both $n\hat{p}$ and $n(1-\hat{p})$ are at least 10.

- **Hypothesis test for proportion:**
  Both np_0 and $n(1-p_0)$ are at least 10.

- **CI and Tests for means:**
  n is greater than both 10(skewness)^2 and 10|kurtosis-3|

`library(moments)`

`skewness`, `kurtosis`

Confidence Intervals

A confidence interval has the form:

$$\text{Estimate} \;\pm\; \text{Margin of Error}$$

$$\text{Estimate} \;\pm\; (z^* \; or \; t^*) \times \text{Standard Error (SE)}$$

Confidence Intervals

- The confidence level *C*, shows how confident we are that the procedure will catch the true population parameter.

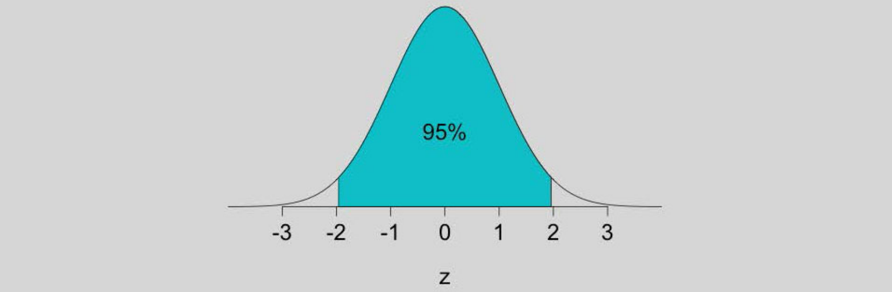
- The procedures give confidence intervals that C% of the time will include the true population parameter

- Type of Problems
  - Proportions
  - Means
  - Matched pairs – (Same as means if you calculate the difference for each pair)
  - Two Independent Samples

Confidence Intervals for: Population Proportion, p

- Estimate: $\hat{p} = \frac{X}{n} = \frac{\text{Number of successes}}{\text{number of trials}}$
- Standard Error: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Critical Value: z^*
- Margin of Error: $z^*\,SE = z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Finding $\pm z^*$ for 95% confidence

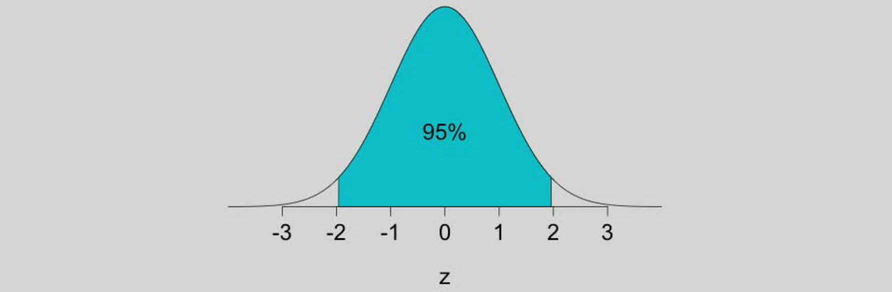


`qnorm`(0.025) and `qnorm`(0.975)

Confidence Intervals for: Population Mean

- Estimate: $\bar{x}$ (`mean`)
- Standard Error: $SE = \frac{s}{\sqrt{n}}$
  where $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ (`sd`)
- Degrees of freedom: $k = n - 1$
- Critical Value: t^* for a distribution $t(k)$
- Margin of Error: $t^* \, SE = t^* \frac{s}{\sqrt{n}}$

Finding $\pm t^*$ for 95% confidence and df = n - 1



`qt`(0.025, df) and `qt`(0.975, df)

Hypothesis Tests

Hypotheses State claims, H_0 and H_a, about a population in terms of the population parameter

Model Are the Independence, Randomness, 10% Condition,
and Sample Size Conditions satisfied?

Mechanics Calculate the test statistic and *P*-value

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \qquad t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

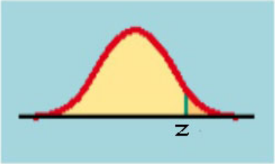
Conclusion Compare the *P*-value and the significance level, α
If the *P*-value $\leq \alpha$, reject H_0. Say the results are statistically significant.

P-value for Proportions

One-tailed tests

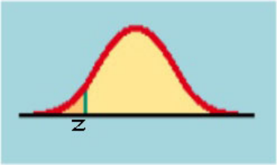
- H_A: $p > p_0$

  `1-pnorm(` $\hat{p}, p_0, \sqrt{\frac{p_0(1-p_0)}{n}}$)



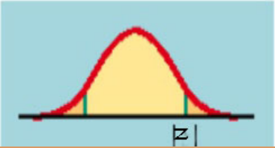
- H_A: $p < p_0$

  `pnorm(`$\hat{p}, p_0, \sqrt{\frac{p_0(1-p_0)}{n}}$)



Two-tailed test

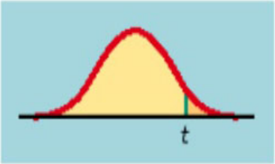
- H_A: $p \neq p_0$
  Double one of the above



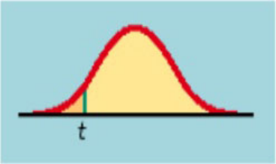
P-value for Means

One-tailed tests

- H_A: $\mu > \mu_0$
  Use `1-pt`(t,deg_freedom)

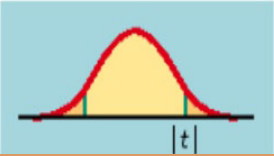


- H_A: $\mu < \mu_0$
  Use `pt`(t,deg_freedom,1)



Two-tailed test

- H_A: $\mu \neq \mu_0$
  Use `2*pt(-abs`(t),deg_freedom)



Errors in Hypothesis Tests

Type I - we reject H_0 (accept H_a) when in fact H_0 is true

Type II - we do not reject H_0 (accept H_a) when in fact H_0 is not true

α = P(Type I error)
= P(reject H_0 when H_0 is true)
= significance level

β = P(Type II error)
= P(do not reject H_0 when H_a is true)
depends on the true value of p

Power = $1 - \beta$

Comparing Two Means

- Matched Pairs

  Find the differences and treat as One-sample

- Two Independent Samples

  Groups must be independent

  $$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

R functions

`ci.prop` for confidence interval for poroportion

`ztest.p` for tests for proportions

`t.test` for inference for means

(ci.prop, ztest.p, and datasets will be given in an .RData file)

Examining Relationships

- Explanatory and response variables

  Independent and dependent variables

- Scatterplots – `plot`
  - Positive or negative association
  - Outliers
  - Linear patterns

- Correlation – `cor`
  - Measures strength and direction of a linear relationship
  - $-1 \leq r \leq +1$
  - $r = \pm 1$ only for perfect linear relationships
  - Correlation does not imply a cause and effect relationship

Regression

- Simple and multiple regression
- Least squares regression
  `lm` and `summary`
- Regression coefficients and their interpretation
  (In multiple regression we interpret the coefficient for one independent variable with the others held constant)
- Standard error of the estimate
- Percentage of variation explained: R^2

Inference for Regression

- Confidence intervals for regression coefficients
  `confint(model, level=0.95)`
- Hypothesis Tests for Regression Coefficients
- F Test for the overall Fit

Assumptions for Inference

- Linear relationship
  Residual plot should not have a curve
  Tilt in the residual plot indicates influential observations
  `plot(x, residuals(model))` and
  `plot(predict(model), residuals(model))`

- Normal errors
  Use a histogram or Normal quantile plot of the residuals
  Use `skewness` and `kurtosis` for the residuals
  `qqnorm`

- Constant standard deviation
  Residual plot should not show increasing scatter

- Independence – Satisfied for randomly selected cross-sectional data