Stats 104-E Spring 2015

Study Notes

David Wihl

April 27, 2015

Contents

1 The Nature of Statistics **4**
1.1 Definitions . . . 5
1.2 Experiments vs. Observations. Causality vs. Association . . . 6
1.3 Bias in Sampling . . . 6

2 Descriptive Statistics **7**
2.1 Average or Typical Values . . . 7
2.1.1 Mean . . . 7
2.1.2 Median . . . 8
2.1.3 Quartiles . . . 8
2.2 Measures of Dispersion . . . 9
2.2.1 Range . . . 9
2.2.2 Interquartile Range . . . 9
2.2.3 Spread, MAD . . . 9
2.2.4 Variance . . . 10
2.2.5 Standard Deviation . . . 10

3 Missing Lectures **11**

4 Decision Analysis **11**
4.1 Maximax and Maximin . . . 11
4.2 Decision Tree Analysis . . . 11

5 Sampling and Sample Distributions **12**
5.1 Population to Sample to Population . . . 12
5.2 Law of Large Numbers . . . 13

6 Central Limit Theorm **14**
6.1 For Continuous Data . . . 14
6.2 For a Proportion . . . 14
6.3 Bias of an Estimator . . . 15

7 Confidence Intervals **16**
7.1 Intro . . . 16
7.2 For a Proportion . . . 17
7.2.1 Interpretation of Confidence Intervals . . . 18
7.2.2 Determine Sample Size . . . 18
7.2.3 Small Samples - Agresti Confidence Intevals . . . 19
7.2.4 One-Sided Confidence Intervals . . . 19
7.3 For Continuous Data . . . 20
7.3.1 In Stata . . . 21
7.3.2 Margin of Error for Continuous Data . . . 21
7.3.3 Small Sample Sizes - Student t . . . 21
7.3.4 Continuous One-Sided Confidence Intervals . . . 22
7.3.5 Summary . . . 23

8 Hypothesis Testing **23**
8.1 Intro to Hypothesis Testing . . . 23
8.2 The Null Hypothesis . . . 24
8.3 Tail Tests . . . 24
8.3.1 Types of Errors . . . 24
8.3.2 Level of Significance . . . 25
8.3.3 Two Tailed Test Using a Confidence Interval . . . 25
8.3.4 One Sided Tests . . . 26
8.3.5 The Test Statistic Approach . . . 26
8.3.6 Small Sample Sizes ($n < 30$) . . . 27
8.4 Testing a Proportion . . . 28
8.5 P-values . . . 28
8.5.1 Intro to P-values . . . 28
8.5.2 How P-values are Calculated . . . 29
8.5.3 P-values in Stata . . . 30

9 Two Sample Testing **30**
9.1 Intro . . . 31
9.2 Hypothesis Testing Two Proportions . . . 32
9.3 Comparing Two Means (large sample) . . . 32
9.4 Matched Pairs . . . 33

10 Chi-Square Test **34**
10.1 Goodness of Fit . . . 34
10.2 Chi-Squared Test of Independence . . . 35

11 Regression Analysis **35**
11.1 Least Squares Method . . . 36
11.2 R^2 Errors . . . 37
11.3 Commands in Stata for Regression . . . 38
11.4 Understanding Regression Output . . . 38
11.4.1 Expressing Uncertainty . . . 38
11.4.2 Estimates of Population Parameters . . . 39
11.4.3 Estimating Error Variance s_e . . . 39
11.4.4 Estimating a Confidence Interval using s_e . . . 40
11.4.5 Understanding ϵ . . . 40
11.4.6 s_y vs s_e . . . 41
11.4.7 Recap: Three Step Plan . . . 41
11.5 Regression Hypothesis Testing . . . 42
11.6 Diagnostics . . . 43
11.7 Reporting Results . . . 43
11.8 Diagnostics Revisited . . . 44
11.9 Predicting Residuals in Stata . . . 44
11.10 Fixing Problems with the Model . . . 44
11.11 Recap . . . 45

12 Multiple Regression **45**
12.1 Comparison with Simple Linear Regression . . . 45
12.2 Interpretation of Multiple Regression . . . 46
12.3 Adjusted R^2 . . . 46
12.4 The Overall F Test . . . 47
12.5 Advanced F Test . . . 48

13 Dummy Variables **48**
13.1 Variable Selection . . . 48
13.2 Dummy Variables . . . 48
13.3 Interaction Term Model . . . 49

14 Regression Diagnostics **50**
14.1 Residuals vs. Fitted . . . 50
14.2 Normality, Heteroskedasticity and Multicollinearity . . . 51
14.2.1 Normality . . . 52
14.2.2 Heteroskedasticity . . . 53

14.2.3 Multicollinearity . . . 54
14.3 Nonlinearities . . . 55
14.4 Finding Outliers . . . 56
14.5 Summary . . . 56

15 Non-parametric Tests **57**
15.1 Signtest . . . 57
15.2 Runs Test for Randomness . . . 58

16 Econometrics (Optional) **58**

17 Final Project **59**

18 Taking the Exams **59**
18.1 How to Study . . . 59
18.2 Exam Shortcuts . . . 59
18.3 Materials to bring . . . 60
18.4 Exam Curves . . . 60

19 Summary of Commands for Tools **62**
19.1 Stata . . . 62
19.2 TI *n*spire Calculator . . . 64

1 The Nature of Statistics

Class 1, Lecture 1

Most people are not very good at analyzing probabilities or understanding data. We have inherent biases when assigning probabilities to events and like to use heuristics (short cuts that lead to decisions that are not necessarily correct).

Statistical Data Discovery in General:

1. Start with a Question / Hypothesis
2. Design an Experiment
3. Collect Data
4. Analyze
5. Does it make sense? Reasonable?
6. Repeat? Publish? Scrap? Redesign?

Why compare averages? It leads to a loss of meaning due other more important factors like variance.

Framing a question is very important.

Most people are risk seeking for losses, risk avoidance for gains, e.g. Sure $50 or coin flip for $100 and most people pick sure $50. $50 sure loss, or coin flip for $100 loss, and most people pick coin flip. This is even though all the probabilities are the same. Probabilistic laws and statistical tools will help understand what's going on in order to lead to better decisions.

1.1 Definitions

Class 2, Lecture 2

Statistics is a collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty.

One goal of statistics is to describe and understand sources of variability. If data didn't vary, it wouldn't be interesting.

A **population** (big N) is the entire collection of objects or individuals about which information is desired. A **sample** (little n) is a subset that is being studied. We assume in this class that the sample is $< 10\%$ of population, and that the population is large.

Descriptive statistics consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs. A **statistic** is a numerical summary based on a sample.

Inferential statistics uses methods that take results from a sample, extends them to the population, and measures the reliability of the result. This is how predictions are made. (This is where all the money is.) Inferential Statistics uses **Estimation** and **Hypothesis Testing**.

A **parameter** is a numerical summary of a population. A **statistic** is a numerical summary of a sample. Usually, we use statistics to make guesses about parameters.

Confounding in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.

A **lurking variable** is an explanatory variable that was not considered in a study, but that affect the value of the response variable in the study. In addition, lurking variables are typically related to any explanatory variables considered in the study.

1.2 Experiments vs. Observations. Causality vs. Association

Two ways of collecting data:

1. Experiments: try to figure out what people do.
2. Direct observation: seeing what people really do. Three types: Observational - time stopped, Retrospective - going back in time, Prospective - going forward or longitudinal

Key elements of experiments:

1. Control effects of variables through double blind, blocks, completely randomized experimental design, rigorously controlled experimental design
2. Replication
3. Randomization

You can only determine causality with an experiment. Observational studies do not allow a researcher to claim causation, only association. Observations do not explain why.

Basic Randomized Controlled Experiment: Individuals are randomly assigned to groups, then the groups are randomly assigned to treatments. If the assignment was truly random, they you can claim causality.

Disadvantages of Experiments:

- Expensive. Might take years or decades to complete.
- Might be unethical
- Subjects could be difficult to monitor
- Animal results may not generalize to humans

Experiments are unnecessary if you don't care about causality.

Most of this class is Observational data. How data is collected is at least as important as how it is analyzed. In particular, a sample should be representative of the population, and random sampling (everyone in the population is equally likely to be selected), is often the best way to achieve this.

1.3 Bias in Sampling

A sampling method is **biased** if the sample favors some parts of the population over others.

Examples of bad sampling:

- Voluntary response - allowing individuals to choose to be in the sample
- Convenience - choosing individuals that are easiest to reach
- Selection: some groups are over- or under-represented
- Nonresponse - some cannot be reached, refuse to participate, or fail to answer some questions
- Response - subject gives an incorrect response when question is confusing or misleading
- Wording-Deliberate - taint response based on question
- Wording-unintentional - question is vague or ill-defined

Data can be time dependent, e.g. a poll on the presidential election that changes rapidly and significantly as the election looms.

Remember: "Garbage in = Garbage out"

2 Descriptive Statistics

We need methods to analyze data both graphically and numerically from a random sample from some population.

Univariate data: one variable per case. *Multi-variate* data: more than one variable per case.

Data may be *discrete* or *continuous*.

Dotplots and histograms are quick visual ways to examine the data, and potentially find outliers.

Histograms are problematic: the bin width may hide valuable information. The *skewness* follows the tail (e.g. right or left).

2.1 Average or Typical Values

2.1.1 Mean

Assuming there is a series of n univariate data points $x_1, x_2, x_3, ..., x_n$.

The Mean or Arithmetic Average is:

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

Mean is affected by extreme values (ie. outliers). It is *not* robust (ie. sensitive to extreme values).

It cannot be less than the minimum value or greater than the maximum value. The mean does not need to be one of the values in the set.

2.1.2 Median

In an ordered list, the *median* is the value right in the middle of the list. If the list has an even number of members, it is the average of the two middle numbers. It is more robust than mean (less sensitive to outliers).

Mean can be calculated in *O(n)* time. Since the median requires sorting, it takes *O(n log(n))* time to calculate.

There is no accepted notation for the median. Also known as 50% point.

If the mean and median are close, the data is symmetric. If there is a significant difference, then there are likely outliers.

Example:

{1,2,3,4,5} mean = 3, median = 3

{1,2,3,4,100} mean = 22, median = 3

Mean is used more often, but the median is often more useful. Mean by itself is not that useful.

Mode is the value that occurs most frequently. It is not covered in this class, but is often the most intuitive value for the typical data value in a given set.

2.1.3 Quartiles

Split the data into 4 equal groups by number of values, or 25% percentiles.

```
25% | 25% | 25% | 25%
    Q1    Q2    Q3
```

Q2 is the median. A box and whisker plot displays this nicely.

2.2 Measures of Dispersion

"The mean and median give us information about the The mean and median give us information about the central tendency of a set of observations, but these numbers shed no light on the dispersion, or spread of the data."

Example:

{5,5,5,5,5} mean = 5

{1,3,5,8,8} mean = 5

Measures of variation or dispersion give information on the spread or variability of the data.

2.2.1 Range

This is simplest measure of dispersion:

$$Range = x_{max} - x_{min}$$

This ignores how the data is distributed. Not robust - highly dependent on outliers.

2.2.2 Interquartile Range

This eliminates the extreme values, making it somewhat more robust. Based on quartiles, how spread out is the middle 50%?

$$IQR = Q3 - Q1$$

2.2.3 Spread, MAD

Measuring average distance to the mean is a horrible measure of dispersion because it is always = 0!

$$spread = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

The distance is supposed to always be positive but spread does not do that. So an improvement is Mean Absolute Deviation (MAD) by using Absolute Value.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

This is very useful but underutilized in practice.

If you can calculate several different metrics (mean, median, MAD) and they all somewhat agree, then it is reasonable to assume that the actual average value is approximately there and outliers are limited.

2.2.4 Variance

Variance is defined as:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The difference is squared to make sure the difference is positive. Highly susceptible to outliers. The units of variance are $units^2$ which makes it hard to interpret.

2.2.5 Standard Deviation

Most common used measure of variance. Shows variation around the mean. Has the same units as the original data.

$$\begin{aligned} s_x =& \sqrt{s_x^2} \\ =& \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

Standard Deviation is not robust - it is very susceptible to outliers (due to the square of the terms). Use MAD when possible as it is easier to explain.

It is often used as a measure of risk. Measure return $\bar{x}$ (dependent) vs risk s (independent).

Standard Deviation measures dispersion equally in both directions. For financial considerations, it is more important to consider downside risk.

Example: Trampoline Math

{5,-5,5,-5,5, -5} $s = 5.48$

{9, -1, 9, -1, 9, -1} $s = 5.48$

but for an investment +9 / -1 is preferable as it will yield a higher rate of return.

3 Missing Lectures

NEED TO COMPLETE LECTURES 4-15

4 Decision Analysis

4.1 Maximax and Maximin

Lecture 16 / Class 17

Many large decisions in corporate environments are made, scarily, without any use of probability. Maximax and Maximin is an example of decision making without the necessity of probability. *Maximax* (optimistic) and *Maximin* (pessimistic) use non probabilistic decision criteria.

A *payoff table* shows alternatives, states of nature and payoffs. See Packet 1, Slides 481 and 490.

Maximax takes the maximum of each row, and then the maximum of the resulting set. "What is the best that can happen?" Most aggressive.

Maximin takes the minimum of each row, and then the maximum of the resulting set. "What is the worst that can happen?" Most conservative.

Adding probabilities to the columns of the payoffs is much more useful. Try to maximize *Expected Monetary Value* (EMV). Example on Packet 1, slide 496. This prediction is not the actual expected amount, but over a large set of samples, this is the average expected amount.

4.2 Decision Tree Analysis

Decision Trees show the problem with all possible outcomes and payoffs. Squares are decision nodes. Circles are uncertain external events (probabilistic node, like a coin). Walk the tree to find which gives the best expected value. "Fold back the tree" walking from right to left. (In CompSci, this is called a Depth First Search). Multiply the end

states times the probabilities and then aggregate to one level up in the tree. At any given branch, the best path can be determined by the aggregated value of the path.

Expected Monetary Value does not include utility and risk. Since it involves a predicted average over repetition, it may not be appropriate for one-off decisions. It also factors in only Monetary value so it does not take into account other objectives (e.g. environment, aesthetic, social)

See Packet 1, slide 518 for an example of a multi-step decision tree.

Sensitivity Analysis: How sensitive is the decision to the assumptions? What new probability would change the decision?

5 Sampling and Sample Distributions

Lecture 17 / Class 18

5.1 Population to Sample to Population

Inference: how to go from a sample to the population.

The *Central Limit Theorm* says that $\bar{x}$ (the guessed average) follows a bell shaped curve. As the sample size increases, the distribution becomes more normal and narrower and the sample mean becomes closer to the population mean. This applies to bimodal and other variables that do not follow a normal distribution (e.g. dragon wings).

p is the proportion of a population with a certain characteristic. $\hat{p}$ is the sample statistic used to estimate p.

μ is mean value of a population variable. $\bar{x}$ is the sample statistic used to estimate μ.

It is assumed that both p and μ are fixed but not known.

Every time you collect a sample of data, you get different $\hat{p}$ and $\bar{x}$. This is called *sample variation*. For this reason, $\bar{x}$ is a *random variable* so it has a probability distribution.

Once we know the sampling distribution of $\bar{x}$ (shape, center, spread of the different samples), we can determine how far $\bar{x}$ is from μ without even knowing μ.

$$
\begin{aligned}
E(\overline{X}) =& E\left(\frac{1}{n}\sum X_i\right) \\
=& \frac{1}{n}E\left(\sum X_i\right) \\
=& \frac{1}{n}\sum E(X_i) \\
=& \frac{1}{n}\sum \mu \\
=& \mu \\
mean(\overline{X}) =& \mu \\
Variance(\overline{X}) =& \sigma^2/n
\end{aligned}
$$

Here's how:

$$
\begin{aligned}
\overline{X} =& \frac{1}{n}\sum_{i=1}^{n} X_i \\
Var(\overline{X}) =& \frac{1}{n^2}\sum Var(X_i) \\
=& \frac{n\sigma^2}{n^2} \\
=& \frac{\sigma^2}{n}
\end{aligned}
$$

So μ is removed from the equation. The standard deviation tells us how far $\overline{X}$ is from the truth. It is determined by the variance of the sample and the number of samples, not by what the actual μ value is.

$$StdDev(\overline{X}) = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt[+]{n}}$$

5.2 Law of Large Numbers

As the sample size increases, the variance of the sample mean decreases (again, the distribution becomes more normal and narrower). As $n \to \infty$, $Var(\overline{X}) \to 0$ and $Mean(\overline{X}) \to \mu$.

6 Central Limit Theorm

6.1 For Continuous Data

1. If samples of size $n \geq 30$ are drawn from any population with mean $= \mu$ and standard deviation $= \sigma$, then the sampling distribution of the sample means approximates a normal distribution. The greater the sample size, the better the approximation.

2. If the population is normally distributed, then the sampling distribution of sample means is normally distributed for any sample size n (not just ≥ 30).

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

Standard error of the mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Unless the distribution is explicitly told, it is not possible to evaluate $P(a < X < b)$.

However, with n sufficiently large, the CLT allows one to evaluate $P(a < \overline{X} < b)$ irrespective of the underlying population.

Lecture 18

This class focuses on estimating two things about a population: the mean and the proportion. Ref: the German Tank problem to estimate the size of the population.

6.2 For a Proportion

For discrete data, we want to know the proportion of a set of values. Assume p is the population proportion of a given characteristic. We now want to estimate p using a sample of the population. This estimate is called $\hat{p}$.

$$\hat{p} = \frac{x}{n} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$

Since $\hat{p}$ is a random variable, the sample has mean and variance

$$\begin{aligned}
\hat{P} =& \frac{1}{n}\sum_{i=1}^{n} X_i \\
E(\hat{p}) =& E\left(\frac{1}{n}\sum X_i\right) \\
=& \frac{1}{n}\sum E(X_i) \\
=& \frac{1}{n}\sum p \\
=& p \\
Var(\hat{p}) =& Var\left(\frac{1}{n}\sum X_i\right) \\
=& \frac{1}{n^2} Var\left(\sum X_i\right) \\
=& \frac{1}{n^2}\sum Var(X_i) \\
=& \frac{1}{n^2}(npq) = \frac{pq}{n} \\
StdDev(\hat{p}) =& \sqrt{\frac{pq}{n}} \\
=& \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
\hat{p} \sim& \mathcal{N}\left(n, \frac{p(1-p)}{n}\right)
\end{aligned}$$

6.3 Bias of an Estimator

For discrete data $n \cdot p \geq 5$ and $n(1-p) \geq 5$ for the CLT to "kick in." In practice n has to relatively much larger like > 100.

		Sample statistic	Population Parameter
Recap:	Mean	$\bar{x}$	μ
	Variance	s^2	σ^2
	Correlation	r	ρ
		Guess	True, but unknown

Guesses should be *unbiased* and have *minimum variance*. MVUE (Minimum Variance, Unbiased Estimates).

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Unbiased if $bias = 0$ (expected value equals true, not a particular value of $\bar{x}$).

For samples, we divide by $n-1$ instead of n to make an unbiased estimator. The guess would otherwise be too low.

$$E(\bar{x}) = \sum_{i=1}^{\infty} x_i p_i = \mu$$

$$E(s^2) = \sigma^2$$

$$\begin{aligned} E[X+c] =& E[X] + c \\ E[X+Y] =& E[X] + E[Y] \\ E[aX] =& aE[X] \\ E[aX+bY+c] =& aE[X] + bE[Y] + c \end{aligned}$$

Example: Roulette has a \$1 bet with a \$35 payoff for $\frac{1}{38}$ odds.

$$E[\text{gain from a \$1 bet}] = -\$1 \cdot \frac{37}{38} + \$35 \cdot \frac{1}{38} = -\$0.0526$$

7 Confidence Intervals

7.1 Intro

Section 5

95% confidence interval means that 95% of samples of this size will produce confidence intervals that capture the true proportion, *or* we are 95% confident that the true proportion lies in our interval.

Formula for Confidence Interval of a mean:

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

The whole $\pm$ term is called the *margin of error* or e. The square root term is called the *standard error*.

7.2 For a Proportion

Lecture 19 - Class 20

In this class, we stop at population μ mean and p proportion. Median, variance, and other inferential statistics are not covered.

Instead of the prior estimated point values, we will now estimate a range of values (aka *Confidence Interval*), which is preferable.

Narrower intervals are better as they are likely more accurate. This works for both continuous and discrete data.

The general estimation process: there is a real but unknown p population proportion. We take a random sample to generate $\hat{p}$. Then using a little formula, we determine our confidence of $\hat{p}$'s accuracy.

Recap 1:
If $X \sim \mathcal{N}(Mean, Variance)$ then $Z = \frac{X-mean}{\sqrt{variance}} \sim \mathcal{N}(0,1)$, resulting in $P(-1.96 \leq Z \leq 1.96) = 0.95$. In other words, 95% of the area is ± 1.96 of Z. 95% is 1.96 standard deviations.

Recap 2 (CLT):
If n is large, then $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ where $n\hat{p} \geq 5, n\hat{q} \geq 5$. By the standardization rule (anything that follows a normal curve can be standardized):

$$\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \sim \mathcal{N}(0,1)$$

$$P\left(-1.96 < \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} < 1.96\right) = 95\%$$

$$P\left(\hat{p} - 1.96\sqrt{\frac{pq}{n}} < p < \hat{p} + 1.96\sqrt{\frac{pq}{n}}\right) = 95\%$$

where p is the true but still unknown value. Via magic, we define the confidence interval as:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

to the find true, but unknown p value. This assumes that n is large, $n\hat{p} > 5$, $n\hat{q} > 5$ and our sample size is less than 10% of the population. This presumes that the population is large (like the number of people in the US).

The real formula includes a finite population correction:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{1-\frac{n}{N}}$$

we can ignore this additional term if $n \ll N$, like less than 10% of the population.

So now know the likely interval. The true p's location in the interval remains unknown. The easiest way to reduce the variance is to get more data.

In Stata, use the "`cii` *number of trials, number of successes*, wald" command. This called the Wald Interval. There are five different interval calculators in Stata. This interval calculator breaks down for large or small p (because the square root term tends to 0).

7.2.1 Interpretation of Confidence Intervals

The interval isn't a probability, but is a way of building confidence in the process. If you created many intervals, only a few (e.g. less than 5%) would not contain the true value.

$z_{\alpha/2}$ is the area under the probability distribution such that the area to the right of $z_{\alpha/2} = \frac{\alpha}{2}$. If $a = 0.05$, this is a 95% confidence interval. If $a = 0.10$, this is a 90% confidence interval. If $a = 0.01$, this is a 99% confidence interval. α is given, typically 0.95.

There is a trade-off between confidence and precision. The higher the confidence, the less precise the interval. Generally, 95% confidence is considered sufficient.

7.2.2 Determine Sample Size

Margin of error: the $\pm$ term.

Now, in reverse, you can start with the margin of error and then solve for n.

$$n = (1.96)^2\hat{p}(1-\hat{p})/(0.03)^2$$

assuming a 3% margin of error is desired.

But this requires $\hat{p}$. How to determine $\hat{p}$?

Option 1: Do a *pilot poll* to determine an appropriate $\hat{p}$, and then determine n.

Option 2: Find the worst case scenario. $0 < \hat{p} < 1.0$ because it is a proportion. Graph $\hat{p}$ vs. $\hat{p}(1-\hat{p})$ to see a local maxima of $\hat{p} = 0.5$. Then plug this to the above formula (with

the supplied margin of error) to determine the sample size.

$$\begin{aligned} n =& (1.96)^2 \hat{p}(1-\hat{p})/(0.03)^2 \\ =& (3.84)(0.5)(0.5)/(0.03)^2 \\ =& 1066.66 \end{aligned}$$

so at worst case, 1067 people are required to get a 3% margin of error. This can be approximated as $\frac{1}{e^2}$ ($1.96 \sim 2$. $n = 2^2 \cdot (1/2)^2/e^2 = 1/e^2$).

7.2.3 Small Samples - Agresti Confidence Intevals

What if there is a small sample with no successes? As a heuristic, use an interval of $\frac{3}{n}$.

The Agresti confidence interval should be used instead of the more popular Wald. Wald doesn't work for small intervals or p is close to 0 or 1.

Recap of Wald:

$$\hat{p} = \frac{x}{n} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$

In Agresti, use:

$$\hat{p} = \frac{x+2}{n+4}$$

to keep p away from edges case of $p = 0, 1$ or when n is small. Then use the same formula:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

7.2.4 One-Sided Confidence Intervals

Puts the 5% uncertainty on one side only.

95% Upper one-side CI (confident of that value or lower):

$$(-\infty, \hat{p} + 1.64\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

95% Lower one-side CI (confident of that value or higher):

$$(\hat{p} - 1.64\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \infty)$$

7.3 For Continuous Data

Lecture 20 / Class 21

The Confidence Interval is constructed for continuous data as with proportional data.

Recall: the CLT for large n:

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

then by the Standardization Rule (Z-score):

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

95% of the data will be within ± 1.96 of the Z score.

$$P\bigg(-1.96 < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96 \bigg) = 95\%$$

By doing some algebra, we get μ within a range:

$$P\bigg(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \bigg) = 95\%$$

so this defines the upper and lower band. We are 95% confident that the true mean μ is within that interval. Out of 100 samples, 95 will contain the true mean.

This problem assumes that σ, the population variance, is known. In practice, σ is never known. If $n \geq 30$, we can substitute s (the sample standard deviation) for σ.

$$\overline{X} \pm 1.96\frac{s}{\sqrt{n}}$$

The n is small ($< 10\%$ of the population) and the data is normally distributed and we know s, then the Confidence Interval is:

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

The 1.96 is for 95%. This coefficient (*margin of error*) changes depending on the desired confidence level. As you get more confident, the interval gets larger. Of course, a confidence of 100% could be achieved with a confidence interval of $(-\infty, \infty)$.

7.3.1 In Stata

When using Stata's `summarize` command, it provides the standard deviation $\overline{X} \pm 1.96s$. This shows how variable the data is (Empirical Rule).

When using the `ci` command, Stata provides the Standard Error, which is $\overline{X} \pm 1.96\frac{s}{\sqrt{n}}$. This is how variable the sample $\overline{X}$ is. This tells you where μ is.

The Confidence Level (α) can be specified in Stata "`ci` *variable*`, level(99.9)`" for a 99.9% confidence interval. The leads to a high degree of confidence, but a very wide range of possible values.

7.3.2 Margin of Error for Continuous Data

Recall for Discrete data, we used:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

which is maximized when $p = 0.50$. However with Continuous Data, the variation is determined by σ and is not fixed at 0.50.

For continuous data, the margin of error (e):

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Factors affecting the margin of error:

- data variation σ. Direct relation
- sample size n. Inverse relation
- level of confidence, $1 - \alpha$. Direct relation

7.3.3 Small Sample Sizes - Student t

This applies when 1) σ is not known, 2) $n < 30$, 3) the data is approximately normally distributed.

If n is small, replacing σ with s results in **more uncertainty**. So the CLT does not hold. Instead, we have a *student's t distribution*:

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

t is a wider / fatter distribution than a normal Z distribution to account for this additional uncertainty. It is centered at 0, but it's degrees of freedom $= n - 1$.

As the sample size or degrees of freedom increases, the t distribution looks like N(0,1) distribution.

$$t_{n-1} \quad \xrightarrow{n \to \infty} \quad \mathcal{N}(0,1)$$

So determine the confidence interval by:

$$\overline{X} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

Determine t by lookup in a table. In the t-distribution table, rows are degrees of freedom which is $n - 1$ and columns represent percentage of confidence interval (α).

Stata always uses t-distribution for confidence intervals irrespective of n because it is more accurate. If $n > 30$ it doesn't matter much. At $n > 1000$, the t value and the coefficient (e.g. 1.96) are the same.

In summary, here are the overall assumptions:

- data is independent and sample is random
- sample size is $< 10\%$ of the population
- if the sample size is small, the data needs to be approximately normally distributed
- for large datasets, skewness doesn't matter due to CLT

See flow chart, packet 2, slide 261.

7.3.4 Continuous One-Sided Confidence Intervals

95% upper bound one-side CI (this value or lower):

$$(-\infty, \overline{X} + 1.64\frac{s}{\sqrt{n}})$$

95% lower bound one-side CI (this value or higher):

$$(\overline{X} - 1.64\frac{s}{\sqrt{n}}, \infty)$$

For 95% one-sided intervals, use $z_\alpha = 1.64$.

7.3.5 Summary

Always correct:

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Approximation for large n:

$$\overline{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

For homework and exams, always assume a fixed approximated 95% CI:

$$\overline{X} \pm 1.96\frac{s}{\sqrt{n}}$$

8 Hypothesis Testing

Lecture 21 / Class 25

8.1 Intro to Hypothesis Testing

Point estimates give us a single "guess."

Confidence Intervals give us a region.

Alternatively, the info in a sample can to used to perform a *test hypothesis* about the parameters in a population.

A *hypothesis* is a statement about a characteristic of one or more population values. It is a question. Means, populations, correlations, covariances can all be types of hypotheses. E.g. "Here is the old summary. Has it changed?"

One-sided means "has it gone up (or down)?" Two-sided means "has it changed (in either direction)?"

We are always using sample data, so there is always a chance of a mistake. Confidence Intervals give a range. A hypothesis test can only give a true/false result of a specific question. If $\mu \neq 20$, we still don't know the real value of μ.

In this class, we use only mean μ and proportion p and no other summary stats.

8.2 The Null Hypothesis

H_0 is a statement to be tested. The *null hypothesis* is a statement of no change, no effect or no difference (status quo). It is **assumed true** until evidence indicates otherwise. By default, we assume that nothing has changed, and then try to disprove it.

H_a is the *alternative hypothesis*, which we are trying to find evidence to support. Since we only have sample data, we can really only disprove a theory, not prove it, since we haven't seen all the data. (Basically we are trying to find the exception to the H_0 null hypothesis). It is always easier to disprove than to prove something.

E.g. H_0: All cats have four legs. To prove it, you need to find all cats in the world. To disprove, you need only find one cat that does not have four legs.

8.3 Tail Tests

Setting up H_0 and H_a (ho and ha):

- Two tail test: $H_0 =$ some value; $H_a \neq$ some value
- Left tail test: $H_0 =$ some value; $H_a <$ some value
- Right tail test: $H_0 =$ some value; $H_a >$ some value

Most statisticians do not like two-sided because it applies to only a single value.

8.3.1 Types of Errors

Four possible outcomes of hypothesis testing:

- Reject null hypothesis when alternative is correct. Correct decision
- Do not reject null hypothesis when alternative is correct. Also a correct decision
- **Type I error**: Reject null when null is correct
- **Type II error**: Do not reject null when alternative is correct

Usually, you can minimize either Type I or Type II errors, but not both due to an inverse relationship. It is worse to make Type I errors, so that is usually minimized. Do not want to move away from Null hypothesis unless there is strong evidence against it.

$$\begin{aligned}\alpha =& P(\text{Type I error})\\ =& P(\text{rejecting } H_0 \text{ when it is true})\\ =& P(\text{reject } H_0 | H_0 \text{ is true})\\ \beta =& P(\text{Type II error })\\ =& P(\text{not rejecting } H_0 \text{ when } H_a \text{ is true})\end{aligned}$$

8.3.2 Level of Significance

The **level of significance** is α, the probability of making a Type I error.

Normally use $\alpha = 0.05$. The greater the cost of a Type I error, the smaller α should be.

We never "accept" the null hypothesis because we don't have access to the entire population. Instead we don't reject the null hypothesis. A legal analogy: we don't say the defendant is "innocent", we say "not guilty."

"There is sufficient sample evidence that x is not true."

"There is insufficient evidence to conclude that $x > y$."

Three ways to do hypothesis testing:

- Confidence Interval Method (classic)
- Test statistic method (classic)
- P-values (computer-based)

8.3.3 Two Tailed Test Using a Confidence Interval

The *Level of Confidence* is directly related to Level of Significance: $(1 - \alpha)x100\%$. It represents the percentage of intervals that will contain the unknown parameter (if repeated samples are taken).

1. Define Hypothesis

$$\begin{aligned}H_0 :& \theta = \theta_0\\ H_a :& \theta \neq \theta_0\end{aligned}$$

2. Construct Confidence Interval

3. Accept or Reject
   If θ_0 falls within this interval, we fail to reject the null ("not guilty!"). If θ_0 is outside the interval, we reject the null ("guilty!").

θ_0 is the hypothesized value, which could be a mean, median, etc.

Find the 95% confidence interval and determine if μ_0 (the true median value) is within this interval.

Since this is a 95% confidence interval, there is a 5% chance of error.

In Stata, use `cii` to create the Confidence Interval.

See example on packet 2, slide 308.

8.3.4 One Sided Tests

Lecture 22 / Class 26

Five steps of Hypothesis Testing:

1. State the null and alternative hypotheses
2. Choose a significance level α (usually 5%)
3. Choose a test statistic and use the significance level to establish a decision rule
4. Compute the value of the test statistic
5. Apply the decision rule and make the decision

8.3.5 The Test Statistic Approach

"Unwrapping the confidence interval." Not in the interval is the same as saying that μ_0 is either less than or greater than the interval, which is the same as:

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -1.96$$

or

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > 1.96$$

(see packet 2, slide 317 for details)

The test statistic is then very simply

$$t_{stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

which is the same as:

$$\mu_0 \text{ not in } \bar{x} \pm 1.96\frac{s}{\sqrt{n}}$$

Test Statistic	Decision Rule
$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$	If $\lvert t_{stat} \rvert > 1.96$, reject H_0
$H_0 : \mu \geq \mu_0$ $H_a : \mu < \mu_0$	If $t_{stat} < -1.64$, reject H_0
$H_0 : \mu \leq \mu_0$ $H_a : \mu > \mu_0$	If $t_{stat} > 1.64$, reject H_0

We use 1.96 because it is 2.5% on either side. We use 1.64 because it is 5% on a single side.

We don't know μ, we have to use the proxy of $\bar{x}$.

If $\bar{x} - \mu_0$ is small, then H_0 is true. If $\bar{x} - \mu_0$ is large, then H_a is true.

If $t_{stat} \ll 0$, then we have case 2 and H_a is true. If $t_{stat} \gg 0$, then we have case 3 and H_a is true.

But this depends on the units, which is removed when dividing by $(\sigma/\sqrt{n})$.

All this assumes that $n > 30$; must use t distribution if n is small.

$\bar{x}$ samples still follow a standard normal distribution (via the Central Limit Theorem). If the alternative is two sided, a very unlikely value is on either of the sample distribution. A one-sided alternative means we care about an unlikely value only on a far single side of the distribution.

8.3.6 Small Sample Sizes ($n < 30$)

The cutoff values of 1.96 and 1.64 have to be adjusted using the t distribution.

8.4 Testing a Proportion

The test statistic is simply

$$t_{stat} = \frac{(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)/n}}$$

Test Statistic	Decision Rule
$H_0 : p = p_0$ $H_a : p \neq p_0$	If $\lvert t_{stat} \rvert > 1.96$, reject H_0
$H_0 : p = p_0$ $H_a : p < p_0$	If $t_{stat} < -1.64$, reject H_0
$H_0 : p = p_0$ $H_a : p > p_0$	If $t_{stat} > 1.64$, reject H_0

For proportions we always assume a lot of data $n \gg 30$.

Variances work differently. Recall:

$$Var(\bar{x}) = \frac{s^2}{n}$$

but for proportions

$$Var(\hat{p}) = \frac{p(1-p)}{n}$$

There are two ways to write the test statistic (which are equivalent for large samples):

$$t_{stat} = \frac{(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)/n}}$$

Most people use the above, but the lower one is a technically correct alternative.

$$t_{stat} = \frac{(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

8.5 P-values

8.5.1 Intro to P-values

Probability Values (P-values) in range [0,1] that provide strength of evidence. *If P is low, H_0 must go.* It is a measure of how much statistical evidence exists.

If P is high, there is evidence for H_0.

If P is low, there is evidence for H_a.

It is a measure of how consistent the data is with the null hypothesis.

p-value < 0.01 corresponds to "highly statistically significant, very strong evidence" to reject H_0.

$0.01 <$ p-value < 0.05 corresponds to "statistically significant and adequate evidence" to reject H_0.

p-value > 0.05 corresponds to "insufficient evidence" against H_0.

Good things about p-values:

- automatically adjusts for large and small datasets
- don't have to worry about one-sided or two-sided
- just simply read the value and plug into the appropriate range

Current fashion is to publish and provide the p-value, rather than the interpretation.

In Stata, use:

```
ttest distance=230
one-sample t test
```

and look at `Pr(T > t)` value. E.g. 0.07 means there is insufficient evidence to reject (but it is close).

8.5.2 How P-values are Calculated

μ_0 is where we want to go. $\bar{x}$ is where we ended up. How far are we? We divide $\bar{x} - \mu_0$ by $s/\sqrt{n}$ only to get rid of the units. If we are close, the null is probably true. If we are far, the null is probably not true.

The P-value is the Cumulative Distribution Function (the shaded area under the Gaussian distribution curve). The smaller it is, the further $\bar{x}$ is from μ_0. If $\bar{x} = \mu_0$ then the P-value $= 0.5$ (half of area under normal distribution would be filled in).

[By avoiding Calculus, this class is talking about bunnies and parachutes instead of precise terminology.]

Need to calculate:

$$P(\bar{x} < value)$$

$$where \;\; \bar{x} \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$$

This is more work by hand, but much easier with a computer.

8.5.3 P-values in Stata

In Stata, if you have only the summary statistics, not the raw data, use

`ttesti` $n\ \bar{x}\ s\ \mu_0$

(Stata always uses the t distribution regardless of sample size).

For the $>$ one-sided hypothesis, the P-value represents the area on the right side of the Gaussian distribution.

When $H_a : \mu \neq \mu_0$, the P-value is defined as *twice* the area to right of observed |sample mean|.

See decision tree on packet 2, slide 391.

In Stata for proportions, use:

`prtesti` $n\ n_1\ p_0$, *count*

where

n is the sample size

n_1 is the sample proportion (or count)

p is the proportion to test

count n_1 is a count not a proportion

9 Two Sample Testing

Lecture 24 / Class 28

9.1 Intro

Testing two means and two proportions, instead of one.

What about more than two? Use Chi-squared tests. [Chi-squared is on homework, but not the exam]

ANOVA (Analysis of Variants)

Notation for Two Proportions: for population $i = \{1, 2\}$

p_i population proportion

n_i size of the sample

x_i number of successes in the sample

$\hat{p}_i = \frac{x_i}{n_i}$ (the sample proportion)

Requirements:

1. We have proportions from two independent simple random samples
2. For each of the two samples, the number of successes and failures is at least 5 each

Create a two sample Confidence Interval. The 95% confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

In Stata, use `prtesti` n_1 x_1 n_2 x_2`,` *count*

If the interval is all positive then $\hat{p}_1 > \hat{p}_2$.

If the interval is all negative then $\hat{p}_1 < \hat{p}_2$.

If the interval spans 0, then one is not significantly bigger than the other (or cannot be determined).

As long as $n > 30$, it doesn't matter if the sample size is different between the random variables.

Use the Stata `by` command to group by value (aka "Stacked data"). Creates two columns for each different value of a random variable. e.g. `prtest snap, by(gender)`

9.2 Hypothesis Testing Two Proportions

Decision Rules for Testing Two Proportions:

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \text{ where } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$\hat{p}$ is called the *pooled proportion*.

The null hypothesis is always $H_0 : p_1 = p_2$.

Test Statistic	Decision Rule	Stata Diff $(p_1 - p_2)$
$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$	If $\|T\| > 1.96$, reject H_0	$H_a :$ diff $\neq 0$
$H_0 : p_1 = p_2$ $H_a : p_1 < p_2$	If $T < -1.64$, reject H_0	$H_a :$ diff < 0
$H_0 : p_1 = p_2$ $H_a : p_1 > p_2$	If $T > 1.64$, reject H_0	$H_a :$ diff > 0

The formula for the test statistic uses a pooled variance estimate because under the null hypothesis we assume that the two proportions are equal.

Recall:

Confidence Interval is *guess* $\hat{p} \pm 1.96\sqrt{\text{Var(guess)}} = \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Test Statistic is $t = \frac{\text{guess}-\text{hypvalue}}{\sqrt{\text{Var(guess)}}} = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ could also be written as $\frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$.

The two Var(guess) do not have to be the same.

For the Confidence Interval, the proportions are not assumed to be equal, so the Variance is estimated by:

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$$

9.3 Comparing Two Means (large sample)

Notation for population $i = \{1, 2\}$:

μ_i population mean

σ_i population standard deviation

n_i size of the first sample

$\bar{x}_i$ sample mean

s_i sample standard deviation

Requirements:

1. σ_1 and σ_2 are unknown. No assumption made about their equality.
2. The two samples are independent.
3. Both samples are simple random samples.
4. The two samples size are both large (ie. > 30) or both populations have normal distributions.

A confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

[Easy to compute manually - so it is probably on the test]

In Stata: `ttesti n1 m1 s1 n2 m2 s2, unequal` . Defaults to t distribution.

The possible hypotheses are the same:

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

Test Statistic	Decision Rule	Stata Diff $(\mu_1 - \mu_2)$
$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$	If $\lvert T\rvert > 1.96$, reject H_0	$H_a :$ diff $\neq 0$
$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 < \mu_2$	If $T < -1.64$, reject H_0	$H_a :$ diff < 0
$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 > \mu_2$	If $T > 1.64$, reject H_0	$H_a :$ diff > 0

9.4 Matched Pairs

This when there are two samples that are **not** independent, e.g. Weight Watchers, Before / After or matched, shared characteristics.

Is the data matched or independent?

If we don't take into account the match, the results are wrong.

To account for this, take the difference between $\overline{X}_1 - \overline{X}_2$ and then do a hypothesis test on the *difference*.

$$H_0 : \mu_D = 0$$
$$H_a : \mu_D > 0$$

10 Chi-Square Test

A class of two tests: *goodness of fit* and *statistical independence*.

10.1 Goodness of Fit

Tests several proportions at the same time, aka the multinomial setting.

k categories of interest with $p_1, p_2, ..., p_k$ probabilities that a value is in a particular cell. All p's add up to 1, as usual.

$$H_0 : p_1 = a_1, p_2 = a_2, ..., p_k = a_k$$

where $a_1, a_2, ..., a_k$ are the values to be tested.

H_a : at least one p_i is not equal to the specified value.

O observed frequency of an outcome, given

E expected frequency of an outcome, calculated

k number of different categories

n number of trials

s_i sample standard deviation

Calculate Observed and Expected to see if they are consistent. Known as Chi-Squared Goodness of Fit (GOF) Test.

$$e_i = n \cdot p_i$$

$$\chi^2 = \sum_{\text{all } i} \frac{(o_i - e_i)^2}{e_i}$$

Smallest possible value is zero. Smaller χ^2 means H_0 is plausible. Larger χ^2 means reject the null.

Use table to determine cut-off values (determined by degrees of freedom $k-1$). As before, we typically use $\alpha = 5\%$ level of significance.

If $\chi^2 > \chi^2_{\alpha,k-1}$, then reject the null in favor of H_a. Something has changed (but we don't know what or which direction).

Online calculator at vassar

Requirements:

1. Data is random
2. Data has frequency counts per category
3. $e_i \geq 5, o_i$ can be anything. Might need to group smaller categories.

10.2 Chi-Squared Test of Independence

aka Two-way Chi-Squared Test.

Tests if r rows and c columns are independent or not. H_0 is independent, H_a is dependent.

In Stata, use `tabi r1c1 r1c2 r1c3 \r2c1 r2c2 r2c3, chi2`. Look for P value. Again, if P is low, H_0 must go.

Need to figure out the probabilities in order to determine e_i. Recall, for independent variables:

$$P(A \text{ and } B) = P(B)P(A)$$

If $e_{ij} = P(r_i)P(c_j)$ then independent

In Stata `tabulate row col, chi2`, which generates a P-value (like the others).

11 Regression Analysis

Class 32 / Lecture 28

Predict the dependent variable Y on the basis of the independent variables $x_1, x_2, ...$

Simple Linear Regression: One X is used to explain Y. Multiple Regression: more than one X is used to explain Y.

k is the dimensionality (number of x's).

n is the number of rows.

In simple Linear Regression, the predicted value is:

$$\hat{y} = b_0 + b_1 x$$

The observed value is (plain) y.

In this case for simplicity, the error is defined Mean Absolute Error:

$$e = \frac{1}{n} \sum_{all\ i} |y_i - \hat{y}_i|$$

11.1 Least Squares Method

This is a simple fitting method. Find b_0 and b_1 which will minimize

$$\sum_{i=1}^{n} (y_i - b_0 - b_i x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

By squaring the errors, we "penalize" large residuals.

Using calculus, we can find that

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{r \cdot s_y}{s_x}$$

So the slope of b_1 has the same sign as the correlation r. The intercept is:

$$b_0 = \bar{y} - b_1 \bar{x}$$

We usually don't care about the intercept, as all it provides is the value of y when $x = 0$.

Least Squares has two interesting properties:

1) The mean of the residuals $\frac{1}{n} \sum e_i = 0$. This implies that $\sum e_i = 0$.

2) The mean of the fitted values equals the mean of the original values ie. $\bar{y} = \bar{\hat{y}}$.

Since $\hat{y}$ is defined as a linear relationship with x, there is a perfect correlation, ie. $corr(\hat{y}, x) = 1$.

$corr(E, X) = 0$ as there is no linear relationship between the errors and x values.

Some basic tautologies:

$$\begin{aligned} Y &= \hat{Y} + (Y - \hat{Y}) \\ Y &= \hat{Y} + e \\ Var(Y) &= Var(\hat{Y} + e) \\ &= Var(\hat{Y}) + Var(e) + 2Cov(\hat{Y}, e) \end{aligned}$$

We know that $corr(\hat{Y}, e) = 0$. There is no correlation between the predicted values and the errors. If the correlation = 0, then the covariance = 0.

$$Var(Y) = Var(\hat{Y} + e) = Var(\hat{Y}) + Var(e)$$

So,

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n-1}\sum_{i=1}^{n} e_i^2$$

or

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

The sum of squares total (SST) = regression (SSR) + error (SSE)

$$Var(Y) = Var(\hat{Y}) + Var(e)$$

SST is the total information in y. SSR is total information explained by x. SSE is the information in y not explained by x. We want to maximize SSR and minimize SST. If SST = SSR, then SSE = 0 and we have a perfect fit.

11.2 R^2 Errors

R^2: what percentage of the variation of the predicted y is explained by the variation in x. The rest is unexplained by the model.

This is the *coefficient of determination*:

$$R^2 = \frac{SSR}{SST}$$

R^2 is in the range $[0, 1]$. The closer R^2 is to 1, the better the fit. If it is 0, the model explains none of the variability of the response data around its mean.

Since we have determined the model over a small range of x's, do not extrapolate beyond the range of x's we've already seen. Do not try to predict y outside of this, such as $x = 0$, as this will lead to inaccurate results.

The most accurate guess is around $\bar{x}$. Further away from this point, the errors become quadratically wrong.

R^2 does not indicate whether a regression model is adequate. You can have a low R^2 for a good model, or a high R^2 value for a model that does not fit the data!

Adding "junk" x variables will increase R^2.

The Root MSE is more important than R^2 (more on this later).

11.3 Commands in Stata for Regression

Command	Purpose
`regress y x`	Perform a regression
`predict yhat, xb`	predict values based on the model
`predict resid, r`	predict residuals

11.4 Understanding Regression Output

Class 33 / Lecture 29

11.4.1 Expressing Uncertainty

How we express uncertainty in exact vs. inexact relationship? Many things in Statistics are inexact relationships. How do we know that there is an inexact relationship? 1) There are errors when predicting y from X, 2) There may be two y values for a given x. When doing regression, we are modeling an average of y rather than the real y, or

$$E(y|x) = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is the error, or noise. This is a *data generating process.*

The noise determines the quality of relationship between x and y. We assume (in this class), the noise mean = 0, is normally distributed:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

The distribution of ϵ does not depend on X (in fact it can't by definition!). In general:

$$y = f(x) + \epsilon$$

Where $f(x)$ is the part of y determined by x and ϵ is the part of y not determined by x. In a simple linear regression $f(x) = \beta_0 + \beta_1 x$.

11.4.2 Estimates of Population Parameters

	Estimate	Population Parameter
Intercept	b_0	β_0
Slope	b_1	β_1
Noise	e_i	ϵ_i
Variance	s_e^2	σ^2
	Guess	True, but unknown

There are $n+2$ unknowns (all ϵ_i plus slope and intercept) for n rows of data. Recall:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

11.4.3 Estimating Error Variance s_e

To estimate variance and standard deviation of the error:

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

s_e^2 is our estimate of σ^2. $s_e = \sqrt{s_e^2}$ is our estimate of σ. The standard deviation of the noise, aka s_e or **Root MSE**, is important as it is good estimate and unbiased. It is more important than R^2 as a measure of quality of regression.

Use s_e to form bands around the regression line:

Percent of y Values	Band
68%	$b_0 + b_1X \pm 1s_e$
95%	$b_0 + b_1X \pm 1.96s_e$
99%	$b_0 + b_1X \pm 3s_e$

11.4.4 Estimating a Confidence Interval using s_e

Assuming a 95% confidence interval, use:

$$\hat{y} = b_0 + b_1 \times x \pm 1.96s_e$$

For 68%, use 1 instead of 1.96. For 99% use 3 instead of 1.96.

Example: a Honda accord with 50,000 miles. What is the 95% confidence interval of the true price?

$$\begin{aligned} x_0 &= 50000 \\ b_0 &= 17066 \\ b_1 &= -.06 \\ s_e &= 303.14 \\ 17066 - 0.06(50000) &\pm 1.96(303.14) = (13472, 14660) \end{aligned}$$

An interval is a lot more useful in practice than R^2. As rule of thumb, use $\pm 2 \times s_e$ to get a quick idea of the model.

11.4.5 Understanding ϵ

The variance of noise ϵ is key. As more noise is added, the prediction will be increasingly incorrect.

The *standard errors* say how good the guesses are, in other words the amount of uncertainty in our estimates of β_i. $\text{Var}(b_i) = s_{b_i}^2$. The smaller the variance, the better the guess.

A confidence interval for β_1 is:

$$b_1 \pm 1.96(s_{b_1})$$

where:

$$Var(b_1) = s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2}$$

Badness occurs with small n, big s_e, small s_x^2. We actually want more variance in the x's so that less of the overall variance is due to noise.

A confidence interval for β_0 is:

$$b_0 \pm 1.96(s_{b_0})$$

where:

$$Var(b_0) = s_{b_0}^2 = s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)$$

11.4.6 s_y vs s_e

After performing a regression, check s_y (via Stata `summary`).

A good model will have $s_e \ll s_y$, because $\bar{y} \pm 2s_y$ is very basic. A good regression model will do significantly better as $\hat{y} \pm 2s_e$ should produce a much narrower interval.

11.4.7 Recap: Three Step Plan

1. Assume there is a model:
$$y = \beta_0 + \beta_1 x + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$
2. Collect data: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
3. Estimate the truth $(\beta_0, \beta_1, \sigma)$ by guesses (b_0, b_1, s_e)

Example:

$$\begin{aligned} b_0 &= 404 \\ b_1 &= 0.214 \\ s_e &= 30.39 \\ s_x &= 0.3921 \end{aligned}$$

CI for slope:

$$0.214 \pm 1.96(0.03921) = (0.122, 0.278)$$

In this case, the resulting interval is all positive (or all negative), so we know there is a relationship between x and y. As before, if the interval spans 0, we are unsure of the relationship.

11.5 Regression Hypothesis Testing

Recap: Assumptions: linear model, noise is modeled by a standard distribution.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

These tests work identically for β_1 and β_0.

We want to test whether β_1 equals a proposed value. We always use a two-sided test.

$$H_0 : \beta_1 = \beta_1^*$$
$$H_a : \beta_1 \neq \beta_1^*$$

To know if x affects y, test if $\beta_1 = 0$. Use the following test statistic:

$$T = \frac{b_1 - \beta_1^*}{s_{b_1}}$$

Test Statistic	Decision Rule
$H_0 : \beta_1 = \beta_1^*$ $H_a : \beta_1 \neq \beta_1^*$	If $\lvert T \rvert > 1.96$, reject H_0
$H_0 : \beta_1 \geq \beta_1^*$ $H_a : \beta_1 < \beta_1$	If $T < -1.64$, reject H_0
$H_0 : \beta_1 \leq \beta_1^*$ $H_a : \beta_1 > \beta_1^*$	If $T > 1.64$, reject H_0

As before if $n < 30$, use t distribution.

Example:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \\ b_1 &= 1.611 \text{ (from Stata Coef.)} \\ s_{b_1} &= .297 \text{ (from Stata Std. Err.)} \\ t &= \frac{b_1 - 0}{s_{b_1}} = \frac{1.61}{.297} = 5.41 \end{aligned}$$

Since $5.41 > 1.96$, we reject the null. x is related to y.

This can be double-checked because the Confidence Interval is all positive (all negative would work too) — zero is not in the interval. Stata also automatically generates this test statistic and the p-values.

β is used in financial planning to see how a security compares to the market, where $\beta = 1$ means the same risk as the market. There are Exchange Traded Funds (ETF) which have a set beta of $+3$ or -3 the market, e.g. FAZ (+3), FAS (-3), SDS (-2), SSO (+2), etc. These are actively managed.

Correlation is a measure of comparison. β is a measure of risk but a poor one due to typically large Confidence Interval.

The search for α, meaning a potential return even if the markets are returning zero. Most funds have $\alpha < 0$ because "they are charging you and they're idiots!"

11.6 Diagnostics

To determine the quality of the model, check the following:

1. What is b_n - the slope? It should be non-zero.
2. What is R^2 (or "dumb and dumber," the adjusted R^2)?
3. What is the Root MSE, $\pm 2s_e$?
4. Check the t-test, p-value (< 0.05) or Confidence Interval (should not include 0)

11.7 Reporting Results

Boiler plate for reporting results:

- This effect is substantial and statistically significant ("You need x").
- A regression on the basis of a random sample of *widgets* indicates that an additional 1 *x variable* increases / decreases *y variable* by $\hat{y}$.
- The regression line explains t% of the variation in *y variable* (R^2).
- A *widget* with *x value* is predicted to have *y value*, with a 95% confidence of $\pm 2s_e$.

11.8 Diagnostics Revisited

Lecture 31

So far, we assume the following models hold (we could be wrong): linear relationship, single x, normally distributed error. All of the assumptions of the model are really statements about the regression error terms ϵ. Since we cannot observe the errors directly, we rely on diagnostics such as the basic least squares residual:

$$e_i = Y_i - \hat{Y}_i$$

We "pretend" that the least squares residuals are the same as true regression errors, *with some limitations*. Occasionally, we use **Standardized Residuals** for convenience:

$$r_i = \frac{e_i}{s_e} \approx \frac{\epsilon_i}{\sigma} \sim \mathcal{N}(0, 1)$$

r_i is standardized: has the correct units (same as y) and 95% are in the range [-2,+2].

Stata has a quick graphical way to determine if these residuals are normally distributed. `kdensity` *var*, *normal* smoothes the histogram and compares it against a normal distribution (based on the sample data's mean and variance $\mathcal{N}(\bar{x}, s^2)$).

Stata has numerical normally distributed tests (see appendix). Typically use `sktest`. H_0 is that the data is normal. H_a is that the data is not normal. The Stata commands generates a p-value to test the hypothesis.

If the residuals do not follow a normal distribution, it might mean several things, one of which is that the relationship between and x and y is not linear. Other options: unusual noise in the system, transform the dependent variable. These are beyond the scope of this class.

11.9 Predicting Residuals in Stata

1. Run regression
2. Calculate residuals `predict res, r`. (May need to `drop res` between runs of different diagnostics.)
3. Do normalcy test on residuals, graphically and numerically.

11.10 Fixing Problems with the Model

"log() is the duct tape of statisticians."

Things that we can do in this class (alone or in combination):

1. Log transform the independent variable
2. Drop outliers

If the x is transformed or values are dropped, the R^2 value and normality tests must be re-calculated as they no longer has the same semantic meaning.

11.11 Recap

Things to know:

- Examine regression output to determine if there is a relationship between X and y.
- Determine Confidence Intervals for regression parameters
- Perform hypothesis tests for regression parameters
- Examine diagnostics

12 Multiple Regression

- Allows several variables at once to explain the variation in a continuous dependent variable.
- Isolate the unique effect of one variable on the continuous dependent variable while taking into consideration that other variables are affecting it too.
- Write a mathematical equation that tells us the overall effects of several variables together and the unique effects of each on a continuous dependent variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

where k is the number of x's (some texts, like Hastie, use p instead of k).

12.1 Comparison with Simple Linear Regression

- intercept is the same
- Slope b_i is the change in y given a unit change in x_i, while holding all other variables constant
- SST, SSE, SSR and R^2 are the same. Instead of variance of y explain by a single x, it is explained by a set of x's

- s_e has a new formula: $s_e = \sqrt{SSE/(n-k-1)}$
- Slope coefficient confidence intervals are the same
- p-values (one for each x_i) are the same
- *Interpretation* is different due to multiple x_i's

Confidence Intervals are the same:

$$b_j \pm 1.96 s_{b_j}$$

The hypothesis test:

$$H_0 : \beta_j = \beta_j^*$$

when

$$t = \left| \frac{b_j - \beta_j^*}{s_{b_j}} \right| \geq 1.96$$

or p-value < 0.05.

12.2 Interpretation of Multiple Regression

When interpreting the sign of the coefficient of a particular variable, you must also assume that the other variables stay fixed. (The "held fix" concept).

Simpson's Paradox is the reversal of signs of directional associations that sometimes occurs when data is aggregated. (See SAT examples in Packet 3, slides 168-171). When other variables are held fixed, there may be a negative association even though there is a positive association in the aggregate.

12.3 Adjusted R^2

By adding dimensions (x's), the error sum of squares (SSE) will decrease so R^2 will always increase. R^2 becomes even less useful in multiple regression. To counteract, the *adjusted R-squared* is available:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Adjusted R-squared imposes an "artificial" penalty for adding dimensions:

$$R^2 = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n-k-1}SSE}{\frac{1}{n-1}SST}$$

Adjusted R^2 can be negative and doesn't have an interpretation (as it has no real but unknown population equivalent). When $k = 1$, the adjusted value does not give the same value as R^2.

When there is a large gap between R^2 and Adjusted R^2 (such as 10%), it usually means there are extra unnecessary variables in the model.

12.4 The Overall F Test

This tests the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

ie. you need nothing. The alternative hypothesis is that at least one x_n is required.

If there are no x's in the model, then SSR=0, and SST=SSR+SSE = SSE.

However, it at least one x is useful, then SSR $\neq 0$, and ideally if some x's are useful, then SSR>SSE. So we compare SSR to SSE in some fashion.

The test statistic is:

$$f = \frac{(SSR)/k}{SSE/(n-k-1)}$$

and reject for large values of f (the x's explain a significant portion of the model).

The f distribution is a series of tables (like χ^2). It tells us when to reject the null by

$$f \sim F_{k,n-k-1}$$

The decision rule is to reject H_0 if $f \geq f_{k,n-k-1,\alpha}$.

(In this class, we simply read the resulting p-value from the regression output and don't use the tables. See Stata's $Prob > F$).

Mathematically, $f = t^2$ in the one variable case.

12.5 Advanced F Test

The *general linear hypothesis* allows testing of any combination of β_i's, e.g. $H_0 : \beta_1 + \beta_2 + 3\beta_3 = 0$ or $H_0 : \beta_5 = \beta_{12}$.

Stata can test any linear hypothesis using regression coefficients using the `test` command, e.g.

```
test size = 0
test size = 5
test size = age
test size - 2 age - lotsize = 0
test size + age + lotsize = 1234
```

Stata will return the p-value of the test.

13 Dummy Variables

13.1 Variable Selection

Forward Stepwise Regression starts with no variables and then adds one at a time.

Backward Stepwise Regression starts with all variables and then delete the least important one based on largest p-value > 0.05. Refit and repeat. Stop when all variables are significant.

Use the p-value or the Confidence Interval rather than the t stat for small ($n < 30$) datasets. The p-value is automatically adjusted by Stata, whereas the correct t stat must be looked up in a table.

The smallest p-value determines the most important variable (do not use the coefficients!).

Stata can perform stepwise backward regression automatically

```
sw regress x1 x2 x3, pr(.05) begin with full model
```

Predictions on the full model might still be better than the simplified model. Backward regression may not result in the same results as forward regression.

13.2 Dummy Variables

Binary - takes on a value of 0 or 1.

Convert x variables with discrete values into a set of binary / dummy variables.

This effectively allows two regressions at once.

Lecture 35, recap of Dummy Variables

β_1 describes the difference the two discrete values, then use hypothesis testing of β_1. Could be used for discrimination testing. The intercept or baseline is when $\beta_1 = 0$. When $\beta_1 \neq 0$, it explains the difference in value of the discrete variable.

β_0, the intercept, represents when all dummy variables have a value of zero.

Remember to use **"while holding everything else the same."**

Example Recoding problem: what if the baseline group was 'other' instead of 'white'?

Original regression:

$$\hat{y} = 30 - 4f + 5b - 2o + 0.3e$$

Recoding the baseline w for o:

$$\hat{y} = 28 - 4f + 7b + 2w + 0.3e$$

Just to be clear, each of the non-baseline values already have the baseline included, i.e. the coefficient of x_1 is $\beta_0 + \beta_1$.

We could use a t-test to compare just one variable:

```
test salary, by (males) unequal
```

The regression allows more variables to be tested than the t-test.

13.3 Interaction Term Model

Models two variables that interact, meaning that the coefficients can combine. Multiply a dummy variable by another variable in the model to create a new variable called the *interaction variable*.

This allows the effect of a continuous variable to differ depending on the value of discrete variable.

It is possible to create an interaction variable with two continuous variables, but then partial derivatives are required for interpretation.

When a discrete variable has multiple values, leave one out to create a baseline and then have dummy variables for each additional category. The baseline represents when all other categories are zeroed out and equals β_0.

14 Regression Diagnostics

Lecture 33

14.1 Residuals vs. Fitted

Use Stata's *residual vs. fitted values plot* `rvfplot` to graph $\hat{y}$ (independent) vs e (dependent). It should look like a random blob.

Even though the p-values, R^2 and Root MSE all look fine, there may still be problems. Generate the residuals vs. fitted values plot. (like Homer Simpson plot).

Need to check our assumptions using residuals from the regression. Ref the famous Anscomb data set, with the same parameters and output even though the datasets are completely different.

Least squares residuals $e_i = y_i - \hat{y}_i$ are good. Standardized residuals are better (see Section 10.8).

Steps:

1. Run the regression `regress y x`
2. Calculate the predicted values `predict yhat`
3. Calculate the residuals e_i `predict res, r`
4. Calculate the standardized residuals r_i `generate sres = res /` s_e

The resulting plot should have no obvious pattern, with half above and below 0. Plot e_i vs $\hat{y}$ or r_i vs $\hat{y}$. The x scale is the same, the y scale is different between these two plots.

The standardized residuals should be in the range $[-2, +2]$ 95% of the time.

Unit recap: The following are all in units of y: y_i, e_i, s_e. r_i is unitless since division by the standard deviation removes the units. If $y = b_0 + b_1 x$, then b_0 must be in units of y and b_1 must be in units of y/x. (This may appear on an exam question as a hypothetical "if the units of x changed, how would b_1 or b_0 change).

How to detection violations of the regression assumption?

Take out outliers because the regression lines wants to go there. It can dramatically change the line. Not all residuals are bad. Look at the residuals. (Extreme points in the y space). Leverage points are extreme in the x space.

Influential observations (extreme in x and y space) are the worst because they exert great influence on the regression line. Use Cook's distance to find them. Applies to project. Should not have to remove more than 10-15% of the data.

Plot the residuals $e = y - \hat{y}$ and look for residuals there.

Non-linearities. Plot x vs y. Look for curves instead of lines. Look for residuals that go -, +, - wrt to the regression line.

Omnibus plot: yhat vs e. Could also plot x vs e. Visually it should look the same if there is only one x dimension. There should be a plot for each dimension, each x. Fix each as appropriate.

e.g. fitting 1/x vs y to make it more linear. Use the lower Root MSE (or R^2) to find which transform works better.

could also log(y), but it makes comparisons between models (R^2, s_e) very difficult.

could of course take the log of an x variable. (other options are splines, fourier transforms, neural nets...)

rule of thumb for transforming x: log(x), sqrt(x), 1/x, x^2. when transforming log(y) make sure it is interpretable.

if you put in x^2, you also need x too. (something I forgot when doing 1/x)

make sure s_e goes down. Could outfit by adding too many variables.

find transformations by plotting y vs each x. make each relationship linear.

1) look for transformations 2) look for outliers 3) check regressions repeat

See flow chart slide 334

14.2 Normality, Heteroskedasticity and Multicollinearity

Lecture 34

Omnibus plot: If all assumptions are met, (meaning $y = \hat{y} + e$, all the information about the x's is stored in $\hat{y}$ and everything else is in e), then the correlation$(\hat{y}, e)$ should be $= 0$. Plotting e vs. $\hat{y}$ should be a random blob.

For multiple variables, plots should be done for each e vs x_i. (This should be done for the project - and include the graphs).

In Stata:

```
regress y x1 x2 x3 x4
rvfplot
predict res,r
generate sres = res / root-mse
scatter res x1
scatter res x2
scatter res x3
...
rvfplot (a final time)
```

After the model is reasonably complete, a final `rvfplot` is a good sanity check.

14.2.1 Normality

Normality: a huge assumption that errors in our regression model are normally distributed. This enables constructing confidence intervals and doing hypothesis tests. This assumptions *must* be validated.

Three different normality tests in Stata. H_0 : errors are normally distributed. H_a : errors are not normally distributed.

```
predict res, r
swilk res
sfrancia res
sktest res
```

We want to fail to reject the null hypothesis, looking for high p-values. If a transformation is done, this cycle must be restarted.

Look at standardized residuals, e.g.:

```
sysuse auto
regress mpg, weight
predict res, r
sktest res
scatter mpg weight
/* looks like an inverse reverse relationship. check it */
generate invweight = 1/weight
drop res /* Stata-ism, need to explicitly drop res */
regress mpg invweight
plot mpg invweight
hist res
/* now standardize */
```

```
generate sres = res / rootmse
hist sres
drop if sres > 2
sktest sres
/* so drop the outliers in the original dataset */
```

(Stata does not allow undo, so make copies of the dataset or reload)

If normality is not satisfied:

- try transforming the dependent variable
- log transform either an x_i or less preferably, the y
- remove outliers

If none of these work, then there are other more advanced techniques but beyond the scope of this class.

14.2.2 Heteroskedasticity

Heteroskedasticity means non-constant variance. We assume that the variance is consistent across all values of all x's. The may not be the case.

Homoskedastic noise:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

A plot would have a straight band around the data, ie. $\text{Var}(\epsilon_i) = \sigma^2$. Irrespective of x_i, the variance is the same.

If we assume homoskedasticity, we have verify our assumption as always.

Heteroskedastic noise:

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

Note the subscript $\sigma_\mathbf{i}^2$ which means the variance changes for a given x_i. This implies that there are a lot of different variances to estimate, complicating the model. A plot of either y vs x or residuals vs. fitted values would have a spread of data points in a fan or tunnel shape. A frequently occurring situation is that $\text{Var}(\epsilon_i) = x^2\sigma^2$ meaning that the variance increases with larger values of x.

With heteroskedasaticity, the estimates are ok, but the standard errors are incorrect, which is a problem in the model. The coefficients are useful, but the p-values are wrong.

An easy way (at this level) to resolve this is by taking a log(y). This makes comparison of models more difficult.

In Stata:

```
generate lsuper = log(super)
scatter lsuper worker
regress lsuper worker
```

but now we can't compare s_e or R^2 to the previous model. The units of y have changed.

Might also be necessary to generate log(x) as well.

```
generate lworker = log(worker)
scatter lsuper lworker
```

x transforms can be compared against prior models because the units of y have not changed.

Test for homoskedasticity in Stata: `hettest`. H_0 is constant variance, and H_a is heteroskedasaticity. Once again, we want a large p-value, if possible.

Fix #1: take log(y), and refit and retest. Then restart model building.

Fix #2: Robust Standard Error, which is beyond scope of class. It assumes that *all* variables have heteroskedasaticity. All modern economic regressions make this assumption.

```
regress y x1 x2 x3, robust
```

14.2.3 Multicollinearity

Multicollinearity: some or all x's are related to each other, when they should be related to y. Two highly related x's confuse the regression algorithm. The standard deviation of the regression coefficients (s_{b_i}) will be disproportionately large, resulting in t ratios that are too small because, recall:

$$t_{stat} = \frac{b_i}{s_{b_i}}$$

When the t stat is too small, it will seem that some variables are not needed when in fact they *are* needed.

"The regression coefficient estimates will be unstable. Because of the high standard errors, reliable estimates are hard to obtain. Signs of the coefficients may be opposite of what is intuitive or reasonable. Dropping one variable from the regression will cause large changes in the estimates of the other variables."

The overall F test is another way of quickly finding that at least one variable is necessary. If the individual p-values say all the variables aren't necessary but the F test disagrees, then there is multicollinearity.

Early on, do a correlation of all x's. If two or more x's are highly related (corr > 0.8), then some should be thrown out. (Also part of the project). Keep the variables that have a better correlation with y. Use Stata's `corr` command.

Dealing with multicollinearity:

- Throw out redundant explanatory variables
- Get more data
- Redefine variables, such as creating an index, e.g. $\frac{x_1+x_2}{2}$
- Step-wise regression

Variance Inflation Factors (VIF) is a good automated way of discovering this.

Take regression of all permutations of x, ie. Take regress of x_1 on x_2, x_3. Take regress of x_2 on x_1, x_3. Take regress of x_3 on x_1, x_2.

Note R^2 on each regression, and then calculate:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Interpretation: If there is no relationship, $R_j^2 = 0$, so $VIF_j = 1$. As R_2 increases (due to a better fit on the regression), VIF_j also increases. If $R_j^2 = 0.90$, then $VIF_j = 10$.

As a rule of thumb, if $VIF_j > 10$, then multicollinearity of x_j may be a problem.

Use Stata's `vif` command to calculate. Drop the x_j with the highest VIF value, refit and re-test. VIF may be better than correlation. (Mention in project that test for multicollinearity was done and move on).

14.3 Nonlinearities

By default, regresses y on $\hat{y}^2, \hat{y}^3, \hat{y}^4$. A significant value means that polynomial terms should be added. H_0: no transformations of x needed, H_a: polynomial or other transformations are needed. A small p-value means one of the x's (but not which) needs to be transformed.

In Stata, use `ovtest`. Re-run again. Might need a x^2 and x^3 and maybe other terms. (Adding a cubic term might need to drop the linear term). As usual, be wary of overfitting especially with polynomials.

14.4 Finding Outliers

Recall: influential observations are the worst (extreme in x and y).

Cook's distance is (approx) $|e_i| * |x_i - \bar{x}|$ (high residual and far away from mean(x)). Looking for extreme Cook values to know which to drop.

In Stata,

```
predict D, cooksd
hist D /* optional */
summ D /* optional */
graph twoway spike D obs
list make if D> .1 /* 0.1 chosen from graph */
```

And as usual, refit and re-check model.

14.5 Summary

Tests to perform when model building:

- Understand the data. Optional: write down some assumptions on anticipated effects of the column semantics
- Multicollinearity check of columns (`corr`)
- Naïve regression as baseline. Check significant columns, R^2, $\pm 2\times$Root-MSE, Overall F-test, $s_e \ll s_y$...
- Optional: check assumptions against significant columns
- Fitted line plot
- Generate residuals (`predict res,r` and `scatter res, yhat`)
- Generate standardized residuals (`generate sres = res / root-mse` and plot it)
- Look for outliers in residuals. sres should be [-2,+2]. Remove outliers.
- Multicollinearity after regression (`vif`)
- Residual diagnostic plots (`rvfplot`)
- Normality of error terms (`sktest` or `swilk res`)
- Heteroskedasticity (`hettest`)
- Nonlinearities (`ovtest`)

- Create interaction variables as needed
- Outliers (histogram, table, dotplot)
- Lather, rinse, repeat...

15 Non-parametric Tests

(not on exam)

Hypothesis Testing has two paths: parametric and non-parametric. So far, we've covered only parametric, ie t test.

Non-parametric tests include the signtest, the runtest, etc. Non-parametric tests are useful when assumptions and requirements (normally distributed, > 30) break down like small data sets, skewed / asymmetric tests. These tests are typically less powerful than parametric tests. They have very few assumptions such as continuous values.

As an alternative to non-parametric tests, transformations (e.g. log, polynomial) are possible.

15.1 Signtest

`signtest` checks a hypothesis by a value around the median.

$$H_0 : \text{population median } \theta = 0$$
$$H_a : \text{population median } \theta \neq 0$$

Assumes data values are independent and continuous. No assumption about distribution.

1. Sort the data
2. Find the median
3. If the value is $<$ median, assign a negative sign
4. If the value is $>$ median, assign a positive sign
   (If the value equals the median, it is ignored and the sample size is reduced)
5. Check that the same number of values are above and below the median
6. Under H_0, number of positive signs follows a binomial distribution with n and p=0.5

In Stata, use `signtest` *var* = *value* and examine the resulting p-value to test the hypothesis. This could be used one-sided or two-sided.

Advantages:

- Simple and logical
- Widely applicable with few assumptions
- Robust to outliers, as only signs are used not values

Disadvantages:

- Severe loss of information
- No confidence interval available

15.2 Runs Test for Randomness

Determine if a set of data is random or not. (See Richard Thaler books, esp. the Winner's Curse)

A *run* is a sequence of similar events, e.g. number of consecutive coin flips coming up heads, number of female patients in a row. It is unlikely that the runs are too large or too small. When the data is quantitative, the median can be used to determine the number of runs above or below the median.

H_0: the data is in a random sequence.

H_a: the data is not in a random sequence.

The math is a little gnarly.

In Stata, `runtest` *var*`,thresh(0)`. Stata automatically splits continuous data at the median or optionally at the mean.

16 Econometrics (Optional)

Ordinary Least Squares estimates are "BLUE" - Best Linear Unbiased Estimates because according to the Gauss - Markov Theorm:

1. **Unbiased** $E(\hat{\beta}) = \beta$
2. **Minimum Variance** the sampling distribution is as small as possible

3. **Consistent** As $n \to \infty$, the estimators converge to true parameters because the variance gets smaller
4. **Normally distributed** statistical tests can be applied

17 Final Project

Make sure to test the model assumptions, including normal distribution. Should include e vs $\hat{y}$ or x_n. Should evaluate e vs each x variable to find the problem.

Check that 95% of residuals are within $[-2, +2]$.

Don't want month_sold in the project as is. Either eliminate or turn into dummy variables (automatic in Stata) or make it seasonable or perhaps summer / not summer.

Don't expect more than 15% of projects to have interaction variables. They are hard to find.

18 Taking the Exams

18.1 How to Study

The best way to study is to do all the old exams. There is also a Study Guide containing many previous exam questions.

The exam is always the same format: a series of multiple choice questions (no partial credit) and a few short answer questions (with partial credit).

Don't hesitate to ask clarification questions of wording during the exam.

Definitely attend the Exam Review prior to the Exam. Lots of useful information.

18.2 Exam Shortcuts

Any time you hear coin tosses, this involves a binomial variation.

There are quartile questions. Know the difference mean and median for each quartile.

Midterm 1: Test for independence is always the last question on the last page. The answer is almost always NOT independent. Hardest of the three exams.

Second half of course: lots of regression.

Midterm 2: there is always a question about unbiased estimators. It is more straightforward and mechanical than the first exam. Topics include: hypothesis test for mean and proportion, Confidence Interval for a mean, Sample size calculation. The swan problem from the homework is always there. Easiest of the three exams.

Final Exam: always a question on dummy variables, such as "here are the variables, which one would you remove first?" or "how would you interpret the coefficient of this dummy variable?" or "what if you recoded the baseline?" Difficulty is between Exam 1 and Exam 2.

18.3 Materials to bring

- *calculator* – Basic functions. No phones / tablets allowed.
- *cheat sheets* – Two sheets, double sided. No limit on page size - poster size ok
- *scrap paper.* – No limit

18.4 Exam Curves

In the Spring 2015 Exam 1, the A- cutoff was 77% for the College and 54% for the Extension School. (On top of that, the Extension school had one less short answer question and 7 more minutes).

In Exam 2, the Extension school A- cutoff was 87%, and A cutoff was 92% (And the Extension school had two less long answer questions and 7 more minutes). The College cutoffs were not mentioned.

From a Facebook post from Prof. Parzen for the Extension School:

> I never explicitly know the curve ahead of time but generally tell everyone not to worry. There is not a fixed curve for each semester and there is no administrative pressure to give a certain number of As, Bs and Cs. We rank everyone in the class [extension by themselves] at the end of the semester and than start drawing cut offs.
>
> The following is stolen from E-1920 and similar to my thinking in giving out grades-just to give you an indication that this isn't a completely painful exercise-those who see this through to the end will be rewarded for their hard work. In general I don't like giving out C's so will usually go lower for the B's than below. The more I get a sense you tried the more the benefit of the

doubt you get, so ask questions, post here and be an active learner as much as possible.

[from E-1920]To determine your semester grade, we will employ the (very generous) curve given below:
Letter Grade - Percentile Rank - Range of Letter Grade
A 60th Percentile - 100th Percentile
A- 45th Percentile - 60th Percentile
B+ 40th Percentile - 45th Percentile
B 35th Percentile - 40th Percentile
B- 28th Percentile - 35th Percentile
C+ 25th Percentile - 28th Percentile
C 22nd Percentile - 25th Percentile
C- 20th Percentile - 22nd Percentile
D 5th Percentile - 20th Percentile
E/F 0th Percentile - 5th Percentile

"In order to give a D, a professor needs to fill in additional paperwork."

19 Summary of Commands for Tools

19.1 Stata

Viewing and Summarizing Data

Command	Purpose
`describe` *var*	quick description of variables
`list` *var* `[if` *cond*`]`	List a variable with optional conditions
`clear [all]`	delete all data and variables
`drop` *var*	remove one column from a dataset
`drop if` *var* > 2	remove one column from a dataset based on a conditional
`tabstat`	

Packages

Command	Purpose
`ADO`	
`findit` *package*	Searches for a package to ease installation
`ssc install weathr`	
`weathr 02459`	
`stockquote` *symbol*	Download stock quotes. May need to munge dates with `gen date2=(date,"YMD")` `format date2 %td`

Basic Statistics

Command	Purpose
`summarize` $x1$ $x2$`,detail`	Show mean, σ, min, max, with optionally more detail
`table` $x1$ $x2$	generate a frequency table
`binomial` (n, x, p)	Generate a binomial
`generate lmpg=log(mpg)`	generate a new column with the log values of a previous column
`corr`	create correlation matrix of all variables

Regression

Command	Purpose
`regress` *dep indep*	Linear regression
`predict` *yhat*, *xb*	Make predictions based on a prior model
`predict` *res*, *r*	predict residuals
`qreg`	
`swilk score`	Shapiro-Wilk normalcy test
`sfrancia score`	Shapiro-Francia normalcy test
`sktest score`	Skewness/Kurtosis normalcy test, e.g. `sktest res`
`test` *var* = *n*	After having run a regression, find the probability of β_1
`test` *expression*	Using the F statistic, test any combination of coefficients
`sw regress` *var*	stepwise regression
`vif`	Calculate Variance Inflation Factors (VIF)

Plotting

Command	Purpose
`dotplot` *var*	quick univariate display that would show outliers
`beamplot` *var* `, xtitle("`*Title*`")`	quick horizontal univariate display that would show outliers
`twoway scatter` *dep indep* `|| lfit` *dep indep*	show a scatter and associated regression
`twoway scatter` *dep indep* `|| lfitci` *dep indep*	show a scatter and associated regression with 95% confidence interval
`histogram` *yhat*	Histogram, but be careful of bin width
`graph box` *var*, `marker(1, mlabel(`*var*2`))`	box and whisker plot
`scatter` *dep indep*	scatter plot
`scatter` *dep indep* `|| line`	scatter plot with line of best fit
`kdensity` *var*, *normal*	plots smoothened histogram of the data against a normal distribution for a quick visual comparison

19.2 TI *n*spire Calculator

Command	Purpose
`invt (1-`α, df`)`	For an area $[0, 1]$ and degrees of freedom, find the Z value. Typically df is large, like 1000
`invNorm(1-`α`)`	For an area $[0, 1]$, find the Z value
`normCDF(`*`Lower, Upper,`* μ, σ`)`	Find the area of an interval
`tTest` $\mu_0, \sigma, \bar{x}, n, \{+1, 0, -1\}$	Perform a Mean t-test. Use $+1$ for $>$, 0 for $=$, -1 for $<$. Find results in `stat.results`