Statistics Final Exam Review Notes

How to Study for the Final?

- Study Exams
- Study Notes
- Key Terms used in Stats
- Final Exam Review Sheet

Topic 1: Collecting Data And Bias

Population: The collection of all people or objects you want to study.

Census: Collecting data from every person or object in the population

Sample: Collecting data from a subgroup of the population

Random: Everyone in the population has an equal chance to be in the data

Various ways of collecting data:

- Convenience (asking friends and family)
- Voluntary Response (putting a survey out into the world and allowing anyone to fill it out.)
- Simple Random Sample (picking individual people or objects randomly usually with a random number generator)
- Cluster (collecting data from groups of people in a population instead of one at a time, selecting classes and getting data from everyone in those classes)
- Stratified (comparing two or more populations, so collecting sample data from each population, comparing a sample of women to a sample of men)
- Systematic (getting data from every 20^{th} person on a list or every 5^{th} person that comes in a store)

Bias: When a data set does not represent the population.

Various kinds of bias

- Sampling Bias (Not using randomization when you collect a sample, Voluntary Response, Convenience)
- Response Bias (Controversial topics, people will not answer truthfully)
- Non-Response Bias (people refuse to answer or take part in the data collecting)
- Deliberate Bias (Deliberate lies about data, deliberately leave out certain groups of the population)
- Question Bias (phrasing a question in a specific way in order to make people answer the way you want)

The goal of data collecting is to get unbiased data that represents the population!!

Topic 2: Experimental Design

Related, Associated, Correlation ≠ Cause and Effect

Why? Confounding Variables

Confounding Variables: Variables that might influence the response variable (Y) other than the explanatory variable (X).

Experiment: Scientific process for controlling confounding variables in order to prove cause and effect.

How to control confounding variables? Make similar groups. (Either the same group of people measured twice or similar groups with random assignment)

Random Assignment: Randomly putting people into two or more groups in order to make the groups alike.

How does experiment prove cause and effect?

- Let's suppose we need to prove that taking medicine (X) causes a person's blood pressure to go down (effect).
- Showing that taking medicine is related to blood pressure does not prove cause and effect, because lots of things can influence blood pressure other than medicine.
- Confounding Variables? Genetics, Age, Stress, Diet, Exercise, Placebo Effect…
- If we create similar groups with random assignment. One group gets the medicine (treatment group) and one group gets a placebo (control group). The groups have similar ages, race, ethnicity, diets, exercise, etc. Since the only difference between the groups is medicine or not, if the medicine group shows significantly lower blood pressure, we have proved it is only the medicine that could have caused it.

Placebo: Fake medicine or fake treatment

Placebo Effect: The capacity of the human brain to regulate physical responses based on the person believing something is true.

Topic 3: Data Analysis

Categorical Data Analysis

Sample Proportion ($\hat{p}$) = $\frac{Amount}{Total} = \frac{x}{n}$

Sample Percentage = $\frac{Amount}{Total}$ x 100%

Example: A data set of 127 people found that 14 were left handed.

Sample Percentage of left handed ($\hat{p}$) = $\frac{Amount}{Total}$ x 100% = $\frac{14}{127}$ x 100% ≈ 0.1102 x 100% ≈ 11.0%

Amount = sample proportion x Total = $\hat{p}$ x n

We collected data from n = 262 people and found that the sample proportion of people in the data that drink beer was $\hat{p} \approx 0.63$. How many people in the data set drink beer?

Amount = sample proportion x Total = 262 x 0.63 = 165.06 ≈ 165 people in the data set drink beer. (Round amount of people to ones place.)

Quantitative Data Analysis (Normal)

Shape = Normal (Bell Shaped)

Average (Center) = Mean ($\bar{x}$)

Spread = Standard Deviation (s)

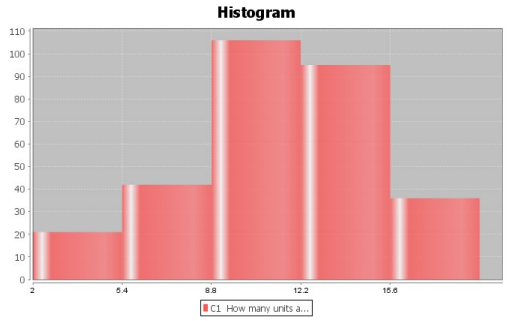
Typical Values = Between $\bar{x} - s$ (mean – standard deviation) and
$\bar{x} + s$ (mean + standard deviation) *(About the middle 68%)*

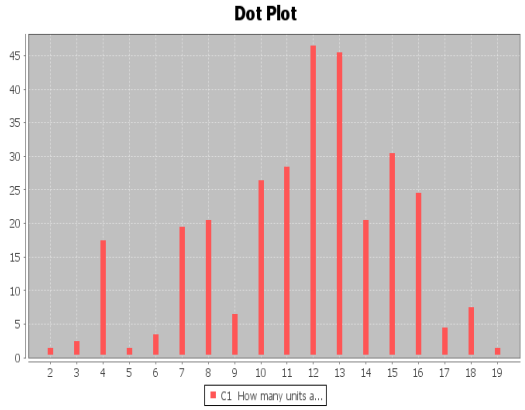
Unusually High Values (High Outliers) = Values in the data set greater than or equal to $\bar{x} + 2s$ (mean + 2 x standard deviation) *(Use dot plot to identify) (Top 2.5%)*

Unusually Low Values (Low Outliers) = Values in the data set less than or equal to $\bar{x} - 2s$ (mean – 2 x standard deviation) *(Use dot plot to identify) (Bottom 2.5%)*

Example:

Number of Units enrolled this semester (Math 140 Survey Data)





Descriptive Statistics

Variable	Mean	Standard Deviation
C1 How many units are you planning to enroll in this semester?	11.593	3.479

Variable	Min	Max
C1 How many units are you planning to enroll in this semester?	2.0	19.0

Variable	N total
C1 How many units are you planning to enroll in this semester?	300

Shape: Nearly Normal (not perfect, but close to bell shaped)

Average (Center) = 11.593 units (mean)

Spread = 3.479 units (standard deviation)

$11.593 - 3.479 \leq$ Typical Number of Units $\leq 11.593 + 3.479$

8.114 units $\leq$ Typical Number of Units $\leq$ 15.072 units

Unusual High Cutoff = 11.593 + (2x3.479) = 18.551 (So any person taking more than 18.551 units is considered unusually high)

Unusual High Values in Dot-Plot: 19 units

Unusual Low Cutoff = 11.593 − (2x3.479) = 4.635 (So any person taking less than 4.635 units is considered unusually low)

Unusual Low Values in Dot-Plot: 4 units, 3 units, 2 units

Quantitative Data Analysis (Skewed or Not Normal)

Shape = Skewed Right, Skewed Left, or any non-bell shaped graph

Average (Center) = Median

Spread = Interquartile Range (IQR)

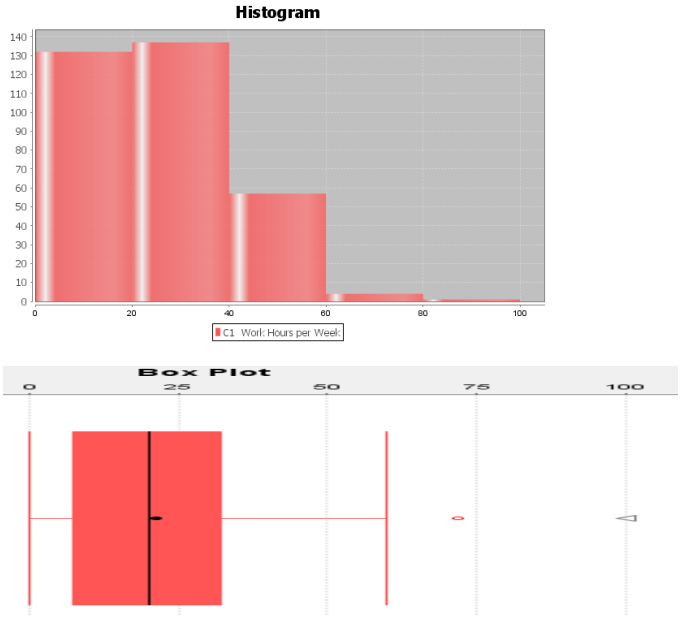
Typical Values = Between Quartile 1 (Q1) and Quartile 3 (Q3) *(About the middle 50%)*

Unusually High Values (High Outliers) = Look for stars, circles or triangles on the right side of a horizontal box-plot (on top if vertical box-plot). *(Note: May need to zoom out the box-plot)*

Unusually Low Values (Low Outliers) = Look for stars, circles or triangles on the left side of a horizontal box-plot (on bottom if vertical box-plot). *(Note: May need to zoom out the box-plot)*

Example:

Work Hours per Week (Math 140 Survey Data)



Descriptive Statistics

Variable	Q1	Median	Q3	IQR
C1 Work Hours per Week	7.0	20.0	32.0	25.0

Variable	Min	Max
C1 Work Hours per Week	0	100.0

Variable	N total
C1 Work Hours per Week	331

Shape = Skewed Right
Average (Center) = 20 hours per week (median)
Spread = 25 hours per week (IQR)
7 hours (Q1) ≤ Typical ≤ 32 hours (Q3)
Unusual Low Values = None (no circles or triangles on left side of box-plot
Unusual High Values = 72 hours and 100 hours (on right side of box-plot)

Topic 4: Sampling Variability and Confidence Intervals

Statistic: A number calculated from sample data.

Parameter: A number describing a population value. Often a guess, but can be a calculation from a census.

Sampling Distribution: Take lots of random samples from a population, calculate a sample statistic (mean or percent) and graph all of the sample statistics on the same graph.

What do we learn from sampling distributions?

- Sampling Variability: Random samples are always different and always different than the population value.
- The center of the sampling distribution is usually a very good estimate of the population value.
- We can calculate the standard error from the sampling distribution. Very useful in confidence intervals and test statistics.

- Standard Error: The standard deviation of a sampling distribution.

Point Estimate: When someone takes a sample value (statistic) and tells the world that it is the population value (parameter).

- Why is point estimating bad? Because of sampling variability, we know the sample value can be dramatically off from the population value (margin of error).
- Margin of Error: How far off the sample value (statistic) is from the population value).

What are the best ways to estimate a population value?

- Take an unbiased census and calculate population value directly.
- Take lots of random samples, create a sampling distribution, and look at the center of the distribution.

In the real world we often cannot take a census and may only be able to take 1 random sample. If all we have is one random sample, how can we estimate the population value?

- Create a confidence interval.
- Confidence Interval: Two numbers that we think the population value is in between.

How is a confidence interval created?

- We can use a formula approach
  Sample Value ± margin of error = Sample Value ± (critical value x standard error)
- We can create a bootstrap distribution and find the middle 90%, 95% or 99% directly without a formula.
- Bootstrapping: Take lots of random samples from one original sample with replacement.
- Bootstrap Distribution: Take lots of random samples from one original sample with replacement, calculate a sample statistic (mean or percent) and graph all of the sample statistics on the same graph.

What are the confidence levels commonly used to create a confidence interval?

- 95% Confident: 95% of confidence intervals contain the population value and 5% of confidence intervals do not contain the population value.
- 90% Confident: 90% of confidence intervals contain the population value and 10% of confidence intervals do not contain the population value.
- 99% Confident: 99% of confidence intervals contain the population value and 1% of confidence intervals do not contain the population value.
- 95% is most common.
- The higher the confidence level the larger the margin of error and the wider the confidence interval.

- The lower the confidence level, the smaller the margin of error and the narrower the confidence interval.

Does sample size effect confidence intervals? Definitely

- More Random Data = Less Error (better accuracy)
- Less Random Data = More Error (worse accuracy)
- A larger random data set will have a smaller standard error, smaller margin of error and a narrower the confidence interval.
- A smaller random data set will have a larger standard error, a larger margin of error and a wider the confidence interval.

Example:

Suppose the sample mean of a data set describing medicine amounts is 74 mg and the margin of error is 13 mg. What would the confidence interval be?
(Assume we used a 99% confidence interval)

Sample Value ± margin of error = 74 ± 13

$74 - 13 < \text{Population Mean } (\mu) < 74 + 13$

61 mg < Population Mean (μ) < 87 mg (inequality notation)

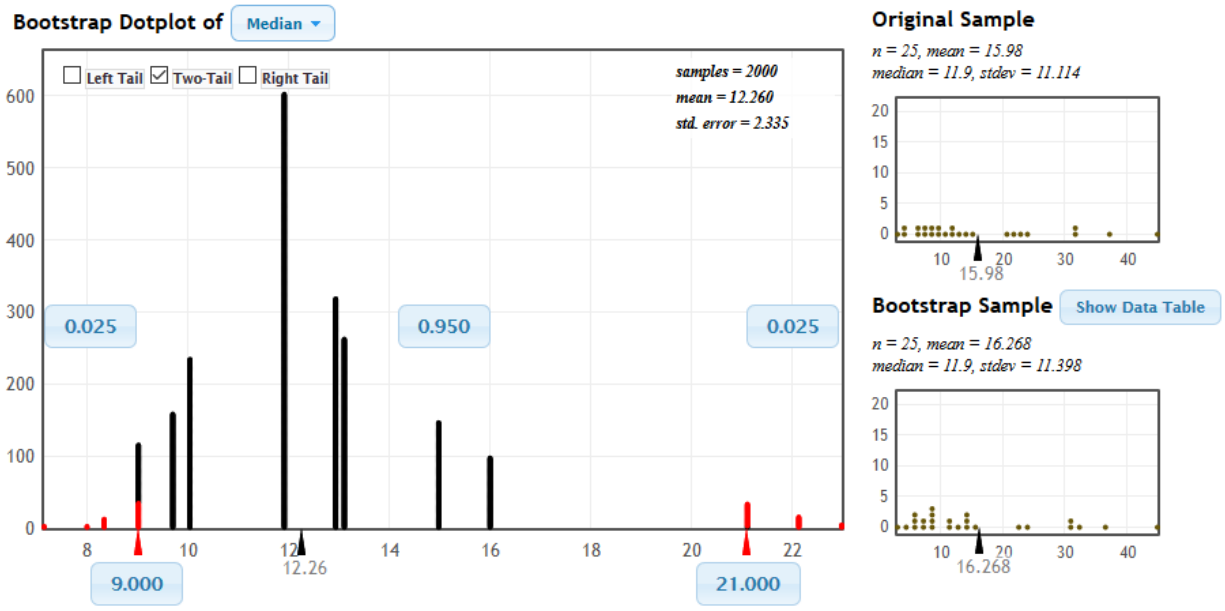
Or

(61 mg , 87 mg) (interval notation)

Sentence to explain: We are 99% confident that the population mean average amount of medicine is in between 61 mg and 87 mg.

Example:

b) Create the bootstrap sampling distribution for the median average price of a used mustang car. Use the bootstrap distribution to find a 95% confidence interval for the population median average price of used Mustang cars.



Conf Int: (9 thous $, 21 thous $)

Sentence: We are 95% confident that the population median average price of a used mustang car is in between 9 thousand dollars and 21 thousand dollars.

What is the advantage of bootstrapping?

- Bootstrapping does not need as many assumptions as formula methods.
- Bootstrapping can be used to find confidence intervals for other statistics besides mean or proportion (median, standard deviation, range, variance, etc.)
- Traditional formula approaches are tied to bell shaped sampling distributions, bootstraps can be used even if the sampling distribution is not bell shaped.

What are the assumptions for creating confidence intervals with formulas?

- One Population Mean Average
  1. Random
  2. Sample size at least 30 or nearly normal
  3. Population size at least 10 times larger than the sample size

- Two Population Mean Average
  1. Random
  2. Both Sample sizes at least 30 or nearly normal
  3. Population sizes at least 10 times larger than the sample size
  4. Groups can be matched pair or Not matched pair

- One Population Proportion
  1. Random
  2. Sample data has at least 10 success and at least 10 failures
  3. Population size at least 10 times larger than the sample size

- Two Population Proportion
  1. Random
  2. Sample data has at least 10 success and at least 10 failures
  3. Population size at least 10 times larger than the sample size
  4. Groups are Not Matched Pair

How do you interpret two population confidence intervals? (Assume 95% confidence level)

- Two population confidence intervals do not measure each population value separately. Instead, they estimate the <u>difference</u> between the population values ($\mu 1 - \mu 2$ or $p1 - p2$)
- (negative , negative)
  Sentence: We are 95% confident that Population Value 1 is between # and # lower than Population Value 2.
  (Significant Difference between population values!!)
- (positive , positive)
  Sentence: We are 95% confident that Population Value 1 is between # and # higher than Population Value 2.
  (Significant Difference between population values!!)
- (negative , positive)
  Sentence: We are 95% confident that there is no significant difference between population value 1 and population value 2.
  (NO significant difference between population values!!)

Example:

Population 1: Mean average salary for women at the company in dollars per hour

Population 2: Mean average salary for men at the company in dollars per hour

90% Confidence Interval from computer: $(-3.58, -2.24)$

Significant Difference? YES!! (negative,negative) Population mean for women is significantly lower than population mean for men.

Sentence: We are 90% confident that the population mean average salary for women is between $2.24 and $3.58 lower than the mean average salary for men at the company.

Example:

Population 1: proportion of people from United Kingdom with high cholesterol

Population 2: proportion of people from Israel with high cholesterol

99% confidence interval from computer: $(-0.027, +0.016)$

Significant difference? NO!! (negative, positive)

Sentence: We are 99% confident that there is no significant difference between the population proportion (%) of people with high cholesterol in the UK and Israel.

Topic 5: Hypothesis Tests

How do we find the null hypothesis and the alternative hypothesis?

- Write down the claim in symbolic language
- Write down the opposite of the claim
- The statement with $=$, $\leq$ or $\geq$ is the null hypothesis (Ho)
- The statement with $\neq$, $>$ or $<$ is the alternative hypothesis (Ha)
- Many people write Ho with $\leq$ or $\geq$ as just $=$
- The statement with an Ha of $>$ is a "right tailed test"
- The statement with an Ha of $<$ is a "left tailed test"
- The statement with an Ha of $\neq$ is a "two tailed test"

Example:

Write the null and alternative hypothesis. We claim that the percentage of seniors with this disease is higher than the percentage of children with this disease.

Claim: $p1 > p2$

Opposite of claim: $p1 \leq p2$

So what is Ho? $p1 \leq p2$

What is Ha? $p1 > p2$ (Since Ha is greater than, it is a right tailed test)

Ho: $p1 \leq p2$ or $p1 = p2$
Ha: $p1 > p2$ (CLAIM)

Test Statistic: A number calculated to determine if the sample data significantly disagrees with the null hypothesis. There are many different test statistics since data takes on my forms.

Read your Test Statistic

If |Test Statistic| ≥ |Critical Value| , the sample data significantly disagrees with Ho.
If |Test Statistic| < |Critical Value| , the sample data does NOT significantly disagree with Ho.

P-value: If Ho is true, it is the probability of getting the sample data or more extreme by random chance.

Read Your P-value

Low P-value (less than or equal to the sig level) : Reject Ho, Is Significant, Unlikely to be random chance

High P-value (more than sig level) : Fail to reject Ho, Is NOT significant, Could be random chance

Writing your Conclusion

Conclusion Sentence: There (is or is not) significant evidence to (reject or support) the claim.

Claim is Ho (Reject or not reject)
Claim is Ha (Support or not support)
Low P-value is evidence (Yes)
High P-value is not evidence (No)

Type 1 and Type 2 Errors

Type 1 Error (False Positive)

- Rejecting Ho and supporting Ha by mistake
- Low P-value from biased data
- To stop Type 1 error, lower the significance level
- Probability of Type 1 error? Significance Level (alpha level)

Type 2 Error (False Negative)

- Failing to Rejecting Ho by mistake
- High P-value from biased data
- To stop Type 2 error, raise the sample size
- Probability of Type 2 error? Beta level

5% significance level: both type 1 and type 2 errors have relatively low probability of occurring (alpha and beta levels are pretty low)

1% significance level: Probability of type 1 error (alpha level) is very low, but the probability of type 2 error (beta level) is higher.

10% significance level: Probability of type 1 error (alpha level) is higher, but the probability of type 2 error (beta level) is lower.

Various Hypothesis Tests

One Population Mean Average

Null and Alternative Hypothesis? (Either can be claim)

Ho: $\mu \geq 36$ thousand dollars
Ha: $\mu < 36$ thousand dollars

Assumptions?

Random
Sample size at least 30 or nearly normal
Population size at least 10 times larger than the sample size

Test Statistic: T – test statistic
Test Statistic Sentence: The number of standard errors that the sample mean is above or below the population mean in Ho.

Two Population Mean Average

Null and Alternative Hypothesis? (Either can be claim)

Ho: $\mu 1 = \mu 2$ (Categorical variable is not related to quantitative variable)
Ha: $\mu 1 \neq \mu 2$ (Categorical variable is related to quantitative variable)

Assumptions?

Random
Both Sample sizes at least 30 or nearly normal
Population sizes at least 10 times larger than the sample size
Groups can be matched pair or Not matched pair

Test Statistic: T – test statistic
Test Statistic Sentence: The number of standard errors that the sample mean from group 1 is above or below the sample mean from group 2.

One Population Proportion

Null and Alternative Hypothesis? (Either can be claim)

Ho: $p = 0.7$
Ha: $p \neq 0.7$

Assumptions?

Random
Sample data has at least 10 success and at least 10 failures
Population size at least 10 times larger than the sample size

Test Statistic: Z – test statistic
Test Statistic Sentence: The number of standard errors that the sample proportion is above or below the population proportion in Ho.

Two Population Proportion

Null and Alternative Hypothesis? (Either can be claim)

Ho: $p1 \leq p2$ (% not related to groups)
Ha: $p1 > p2$ (% is related to groups)

Assumptions?

Random
Sample data has at least 10 success and at least 10 failures
Population size at least 10 times larger than the sample size
Groups are Not Matched Pair

Test Statistic: Z – test statistic
Test Statistic Sentence: The number of standard errors that the sample proportion from group 1 is above or below the sample proportion in group 2.

Goodness of Fit Test
(multiple P test - categorical variable percentage is same in lots of groups)

Null and Alt hypothesis? (Either can be claim)

Ho: p1 = p2 = p3 = p4 = p5 (% is not related to groups)
Ha: at least one is not equal (% is related to groups)

Assumptions?
Random
Population sizes at least 10 times larger than sample sizes
Expected values are at least 5
Individuals Independent of each other.

Test Statistic?
Chi-Squared Test Statistic $\left(\chi^2\right)$: The sum of the averages of the squares of the differences between the observed sample data and the expected values from the null hypothesis.

Independence / Homogeneity Test Categorical / Categorical relationship test

Data: Quantitative Variables (more than two responses) gives a contingency table (two-way table)

Null and Alternative hypothesis? *(Same for both Independence or Homogeneity)*
(Either Ho or Ha can be claim)

Ho: The categorical variables are not related (independent, not associated)
Ha: The categorical variables are not related (independent, not associated)

Assumptions for Independence test?
Random
Population sizes at least 10 times larger than sample sizes
Expected values are at least 5
Individuals are independent of each other

Assumptions for Homogeneity test?
Random
Population sizes at least 10 times larger than sample sizes
Expected values are at least 5
Individuals and samples are independent of each other

Test Statistic?
Chi-Squared Test Statistic $\left(\chi^2\right)$: The sum of the averages of the squares of the differences between the observed sample data and the expected values from the null hypothesis.

ANOVA Hypothesis Test (showing mean average is equal in lots of different groups) – Categorical / Quantitative Relationship study.

Type of Data (Quantitative / Categorical)

Null and Alternative Hypothesis? (Either can be claim)

Ho: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5$ (Categorical variable is not related to quantitative variable)
Ha: at least one $\neq$ (Categorical variable is related to quantitative variable)

Assumptions?

Random
Population sizes at least 10 times larger than sample sizes
At least 30 or bell shaped
Groups should be independent of each other
Similar Variances (no standard deviation is more than twice as large as any other)

Test Statistic for ANOVA?

F test statistic: The ratio of the variance between the groups to the variance within the groups.

Three Relationship / Association Hypothesis Tests

Independence test (Categorical / Categorical)

ANOVA test (Categorical / Quantitative)

Correlation test (Quantitative/ Quantitative)

Example:

What test would we use if we wanted to see if a person's political viewpoint is related to the person's blood pressure in mm of Hg?

ANOVA

Example:

What test would we use if wanted to show that believing in UFO's is related to where you live?

Independence / Homogeneity Test

Example:

What test would we use if we wanted to show that the number of miles on the car you drive is related to the amount of money in thousands of dollars you have spent on repairs to that car?

Correlation Test

Correlation Hypothesis Test (is there a linear relationship between two different quantitative variables)

Type of Data? (Quantitative / Quantitave)

Null and Alternative Hypothesis? (Either can be claim)

Ho: Slope = 0 (no correlation)
Ha: Slope ≠ 0 (is correlation)

Test Stastic for Correlation? T – test statistic

T-test statistic (for correlation): The number of standard errors that the slope of the regression line is above or below zero.

Assumptions for Correlation Hyp Test?

Random quantitative ordered pair data
Population sizes at least 10 times larger than sample sizes
Scatterplot and correlation coefficient (r) show some linear trend
No influential outliers
Sample size at least 30
Histogram of residuals nearly normal
Histogram of residuals centered close to zero
Residual plot should be evenly spread out (NO fan or V shape).

What are some of the sample statistics we look at in Correlation and Regression?

Correlation: A linear relationship between two different quantitative varaibles.

***Remember Correlation is not causation

Correlation Coefficient (r): gives the strength and direction of the linear correlation.

If r close to +1 : Strong Positive Correlation (points in scatterplot close to line going up from left to right)
If r close to −1 : Strong Negative Correlation (points in scatterplot close to line going down from left to right)

If r close to 0 : No Correlation (points in scatterplot do not follow any linear pattern)

r-squared: percentage of variability in y that can be explained by the linear relationship with x.

x –variable (explanatory variable)

y-variable (response variable)

Note: y should respond to x. Also y should be the variable you are most interested in, or want to make predictions about.

Ex) temperature vs cases of pneumonia
temperature does not respond to cases of pneumonica (x explanatory)
cases of pneumonia may respond to temperature. We are also more interested in predicting how many cases of pneumonia we are likely to see.

Residual: The vertical distance that each point in the scatterplot is above or below the regression line

Standard deviation of the residual errors s_e : (Two meanings?)

1) The average vertical distance points are from the regression line.
2) The average prediction error.

Making Predictions

Ex) #cases of pneumonia (thousands) = 47.1 + -0.135 temperature (farenheit)

Predict the number of cases of pneumonia if the temperature is 35 degrees F?

#cases of pneumonia (thousands) = 47.1 + -0.135 (35)

= 47.1 + -4.725 = 42.375 (thousands of cases of pneumonia)

Slope: The amount of increase or decrease in the Y variable per 1 unit of X variable.

What does the slope of -0.135 mean? (rate of change)

The number of cases of pneumonia is decreasing 0.135 thousand (135 cases) for every degree the temperature increases.

Y-intercept: The predicted Y value when X is zero.

What does the y-intercept represent in this problem? When x is zero (initial value)

(0 , 47.1)

If the temperature is 0 degrees F, we predict 47.1 thousand cases of pneumonia.