INST314 Class Notes

250902 Day 1

"Anomaly Detection Across Multi Scale Temporal Data Streams for Human Behavior Modeling" - Dr. Faisal Quader

Statistics

The study if the collection, analysis, interpretation, organization, and presentation of data

Types of Statistics

Descriptive Statistics

- Measures of central tendency, variance, range
- Scatter plots, histograms, bar charts

Inferential Statistics

- Interpret the meaning of descriptive statistics
- Extrapolate sample to populations
- Hypothesis testing, correlational analysis, prediction
- Rely on probability distributions

Important Terms

- Population
- Sample
- Parameter
- Variable
- Constant
- Variable types
  - Numeric
    - Continuous (interval, float, numeric)
    - Discrete (integer, count)
  - Categorical (enums, enumerated, factors nominal)
    - Binary (dichotomous, logical, indicator, Boolean)
    - Ordinal (ordered factor)
    - Nominal (character)

Population vs. Sample

- These are measures: numbers or categorical labels
- You cannot have a normal distribution of chickadee beaks, not are we piling human beings up in a bell curve-shaped heap!

Descriptive Statistics

Measures of central tendency/estimates of location

- Mean (average)
- Median (50th percentile)
- Mode
- Range
- Weighted mean
- Trimmed mean (truncated)

Description terms

- Robust/resistant
- Outlier/extreme value
- Anomaly detection

Mean

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

$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$

Modal Class

$M_O = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$

Where
l = lower limit of the modal class
h = size of the class interval (assuming all class sizes to be equal)
f_1 = frequency of the modal class
f_0 = frequency of the class receding the modal class
f_2 = frequency of the class succeeding the modal class

Distributions

- Normal distribution - parametric
- Skew - kurtosis
- Normal is not so normal
- Bimodal
- Binomial
- Poisson
- Normal is not so normal, but sometimes we can make it normal
- Central Limit Theorem

- n>30, preferably much greater
- n<30, sample mean follows population mean distribution
- https://vimeo.com/75089338

Why the emphasis on normality?

- The math is easy
- Most things in nature can be made to follow it via the central limit theorem
- But is it always the best choice?

Statistics in Information Science

Statistics in Science

- Question phrased as a *testable* hypothesis
  - Null hypothesis and research hypothesis
    - Null: nothing is happening here, no effect/difference
      - H_0
    - Alternative or Research: Something is happening here. There is an effect/these samples are different
      - H_a or H_R
- Collect data (sampling method)
- Analyze data
  - Descriptive statistics: mean, median, mode, range, variance, standard deviation
  - Inferential statistics: confidence intervals, regression, t-test, prediction
- Conclusion: what is the probability of a result this extreme based on pure chance? Significance/p-value
- Presentation, refinement, next experiment

Experiments vs. Observational Studies

- Controls
- A true controlled experiment is the only way to prove causation
  - Ethics
  - Practicality
- Correlation vs. Causation

Statistics in Information Science

- Sample size
- Experimental design?
- Sometimes what we are interested in is the anomalies!
- Statistical significance and real-world difference are not always the same thing
- Does the distribution matter?
  - What part of the curve are you interested in?

Statistics in Science vs. Information Science

Science	Information Science
How strong is the evidence?	How reliable is this model/conclusion?
Bias toward the null	Bias may be towards the research hypothesis
Control over sampling, study design	Often have little/no control over sampling
Frequently using a small sample to infer information about a large population	Residuals or subtle trends may be the most interesting
- Type 1 error: mistakenly reject the null - Type 2 error: mistakenly accept the null	P-value: in an interesting model result easily within the bounds of normal variation? How strongly predictive is it?

250909 Day 2

Statistics is the study of the collection, analysis, interpretation, organization, and presentation of data
We need to be able to visualize our data in order to communicate it (for example, to businesspeople)

Qualitative vs. Quantitative - describe the *quality* ("I did pretty well on the test") as a *quantity* (85%-95%)

Variable Types

Why do we care?

- Knowing the data type tells the software how to behave (i.e. just because locations have been entered as zip codes does not mean that it makes sense to do math on them)
- Storage and indexing can be optimized (relational databases)
- Possible values/operations enforced in the software

When is it a problem?

- Default behavior of important functions is to automatically convert a text column into a factor with only existing values as valid

Descriptive Statistics

Measures of central tendency/estimates of location:

- Mean (average)
- Median (50th percentile)

- Mode
- Range
- Weighted mean
- Trimmed mean (truncated) - remove highest/lowest value to reduce skewing

Related terms:

- Robust/resistant
- Outlier/extreme value
- Anomaly detection

Estimates of variability:

- Deviations (errors, residuals): difference between the observed values and the *estimate of location*
- Estimate of location - whatever measure of central tendency is being used

> **Standard Deviation Formulas**
>
> **For a Population**
>
> $$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$
>
> **For a Sample**
>
> $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

- Variance (mean squared error/difference): squared deviations from the mean

> **Variance Formulas**
>
> **For a Population**
>
> $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$
>
> **For a Sample**
>
> $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Mean absolute deviation(LI-norm, Manhattan norm): absolute value of deviations from the mean
- Median absolute deviation - robust to outliers
- Range
- Order statistics: rank
- Percentile of quantile

- Interquartile range (IQR)

Assumptions

- If there is a calculation, there are assumptions
- Is this a variable you can do math on?
- You need to know the distribution of the data to calculate variance, standard deviation, etc.
- Just because you can tell the computer to do the math does not mean that the result is meaningful
- There are ways to calculate all the descriptive statistics for different sampling methods and different distributions
- Unless otherwise noted, the formula shown is for a simple random sample of a normally distributed variable
- Every statistical model is a probability model built on assumptions about the distribution of the data and the sampling method

What happens when we violate assumptions?

- Assuming a normal distribution can lead to underestimation of extreme events (black swans)
- Predictive models may perform very poorly
- Multivariate models will give inaccurate results
- Some analyses are sensitive to very minor violations of their assumptions
  Why run an analysis that is guaranteed to give you a wrong answer?

How do we avoid violating assumptions?

- Choose the correct model/test for the data type
- Choose models with few assumptions
- Large sample size - not a cure-all but helps
- Visualize the data!

Distributions

Measure the spread of the data
When can you substitute a normal distribution and why would you?
Normal is not so normal
If you can describe it as a function you can derive the statistics for it
Calculus:
Area under the curve (integral)
Specific points of interest? (derivatives!): central tendency

Normal

Binomial

The binomial distribution with n trials and success probability p has:

- Mean = np
- Variance $\sigma^2 = np(1-p)$
- True/False, +/-, Heads/Tails
- Large sample size effect
- P near 0.5 - approximates normal

Poisson

Rate data
Becomes more normal as the rate increases

Bimodal

Combined populations or a third factor that is influencing the variable measured

- If you can describe it as a function you can derive the statistics for it
  - Calculus:
    - Area under the curve (integrals!): standard deviations
    - Specific points of interest? (derivatives!): central tendency
- Options if your variable of interest is not normally distributed
  - Central limit theorem
  - Is it normal where it matters?
  - Can I make it normal (enough)?
  - Use the right distribution

Sampling Methods

Probability Sampling

Simple random

Every member of a population have an equal chance of being selected
Usually representative

Stratified random

Split population into groups, randomly select from each group
Ensures every group is represented even if it is a small group

Cluster random

Split population into clusters, randomly select clusters and include every member of the selected clusters
Useful when clusters are reflective of population as a whole

Systemic random

Put every member of the population in some order, choose a random starting point and select every nth member
Usually representative

Non-Probability Sampling

Convenience

Choose members that are readily available
Under coverage

Voluntary response

Researcher puts out a request for volunteers
Non-response bias

Snowball sample

Researchers recruit initial subjects and then ask them to recruit additional subjects
Sampling bias thus findings cannot be extrapolated to larger population

Purposive sample

Researchers recruit individuals based on who they think will be most useful based on the purpose of the study
Non representative sample

Sampling methods and assumptions

- Every probability sampling method has associated methods to calculate valid statistics
- If your sample is not a probability sample, realize that your conclusions may only apply to your specific sample
- Do not confidently state probabilities for a non-probability based sample!
  - "*If* this sample represents the population we would expect..."

Visualizing the Data

Graphing the Distribution

Qualitative

- Categorical variables
- Bar graph
- Additive bar graph
- Pie chart
- When there are many categories, use a horizontal bar chart
- Use frequencies instead of percentages when you have relatively few observations as percentages can make a small difference look large
- DON'T use a continuous line to represent non-continuous data!

- Keep it simple. Overly complicated graphs are hard to read

Quantitative

- Stem and leaf: small datasets
- Histogram: large datasets
- Frequency polygon: compare distributions
- Cumulative frequency polygon
- Box plot: identifying outliers and comparing distributions
- Box and whiskers

Name	Formula
Upper Hinge	75th Percentile
Lower Hinge	25th Percentile
H-Spread	Upper Hinge - Lower Hinge
Step	1.5(H-Spread)
Upper Inner Fence	Upper Hinge + 1 Step
Lower Inner Fence	Lower Hinge - 1 Step
Upper Outer Fence	Upper Hinge + 2 Steps
Lower Outer Fence	Lower Hinge - 2 Steps
Upper Adjacent	Largest value below Upper Inner Fence
Lower Adjacent	Largest value above Lower Inner Fence
Outside Value	A value beyond an Inner Fence but not beyond and Outer Fence

- Bar chart
- Line graph: both x and y axes display ordered variables
- Dot plots

Misuse of Statistics

- Unfair or impractical criteria of comparison
  - Apples vs. oranges, false analogy, logical fallacy
- Cognitive biases
- Biased sample
  - leading question
  - Selecting the sample or portion of data to tell the story you want to tell
- Correlation vs. Causation
- Data dredging (data snooping, vast search effect)
- Overfitting

- Biased labeling
- Estimation error
- Low quality data
- Omitting a controversy
- Out of context data
- Overcomplexity
- Overfitting
- Prosecutor's fallacy - invalid interpretation of a valid statistic
- Regression toward the mean
  - Your first result does not necessarily predict your second
- Significance: not all p-values are created equal
- Subject matter interpretations: Do the results make sense?
- Tyranny of averages: impact outliers

250916 Day 3

Hypotheses and Hypothesis Testing

Hypothesis

- Could these observations have occurred by chance? (assumes a random sample!!!)
- Null hypothesis: an apparent effect is purely due to chance
- Research hypothesis/Alternative hypothesis is the opposite of the null: there is a real difference and the observations are the result of the real effect plus chance variation
- Rejecting the null

Null Hypotheses vs. Alternative Hypotheses

- Company ABC manufactures calculators with an average mass of 450g. An engineer believes that the average weight to be different and decides to calculate the average mass of 50 calculators.
  - Null: $H_0 : \mu = 450g$
  - Alternative: $H_a : \mu \neq 450g$
- The teachers in a school believe that at least 80% of the students will complete high schools. A student disagrees with this values and decides to conduct a test.
  - Null: $H_0 : p \geq 0.8$
  - Alternative: $H_a : p < 0.8$
- The percentage of residents who own a vehicle in town XYZ is no more than 75%. A researcher disagrees with the value and decides to survey 100 residents asking them if they own a vehicle.
  - Null: $H_0 : p \leq 0.75$
  - Alternative: $H_a : p > 0.75$

Probabilities

- Your calculated probability is the likelihood of getting a value this extreme by chance alone
- An example: Panels of juror in the south in the 1950s-1960s
  - 50% of eligible citizens were black
  - On an 80 person panel of potential jurors supposedly pulled at random from the voter rolls, only 4 were black
  - Binomial variable with n=80 trials and p=0.5: Pr(x</=4) = 0.00000000000000000014
- α level
  - α level - probability of making the wrong decision when the null hypothesis is true. Also known as the significance level. Type I error.
  - if the p-value is < α level, then we reject the null

What does that p-value mean???

- is NOT the likelihood that your alternative hypothesis is correct!
- IS highly dependent on sample size
- CANNOT compare p-values for samples of different sizes
- Does not "prove" anything
- The alpha is the threshold. Very small values are not stronger proof. Accept or reject, not strongly reject.

Decision Theory: Errors

- Type I error: Reject the null (conclude something is going on) when it really isn't
- Type II error: Accept the null (conclude nothing is going on) when it really is
- Significance testing emphasizes the probability of a type I error: all the alarm bells are genuine
- But what if the opposite is more important?

Power Analysis

- Power
  - How much overlap between two distributions
  - Sample size
- How certain am I to correctly reject the null: how big does my sample need to be to have the power to not accidentally reject my research hypothesis
- Take home: p-values are not magical. If your sample size is too small you will make mistakes no matter how stringent your p

Binomial Distribution

- Boolean, +/-, yes/no, 0/1
- Bernoulli Trial

- Binomial Distribution/Bernouli Distribution: number of successes (n) in x trials
- As the number of trials becomes large the PMF (Probability Mass Function) approximates a normal distribution
- The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes: success or failure.
- A random variable X follows a binomial distribution if:
  - There are a fixed number of trials (n)
  - Each trial has only two outcomes (success/failure)
  - The probability of success is constant (p)
  - The trials are independent of each other

Binomial Probability Formula

$P(x) = \frac{n!}{(n-x)!x!}p^x q^{n-x}$

Mean

$\mu = np$

Standard Deviation

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

Example

- Suppose you flip a fair coin 5 times ($n = 5, p = 0.5$). Let $X =$ number of heads.
  - $P(X = 3) = (\frac{5}{3})(0.5)^3(0.5)^2 = 10 \cdot 0.125 \cdot 0.25 = 0.3125$
  - So the probability of getting exactly 3 heads in 5 flips is 31.25

Shape of the Distribution

- Symmetric if $p = 0.5$
- Skewed right if $p < 0.5$
- Skewed left if $p > 0.5$
- As n increases, it starts to look like a normal distribution (by the Central Limit Theorem)

250923 Day 4

Binomials

Coding Resources

- Binomial test: https://www.statology.org/binomial-test-r/

- Binomial Confidence Interval: https://www.statology.org/binomial-confidence-interval-r/

Calculate 95% confidence interval for a sample of 100.

```
prop.test(x=56, n=100, conf.level=0.95, correct=False)
```

Binomial Distribution

Binomial Experiment Assumptions

- X ~ Binom(n,p): 3 & 4 frequently violated but have minimal impact with large sample size and p ~ 0.5

1. The experiment consists of **n** identical trials and the **n** is always **fixed**
2. Each trial results in one of the two outcomes, called success and failure
3. The probability of success, denoted **p**, remains the same from trial to trial
4. The **n** trials are **independent**. That is, the outcome of any trial does not affect the outcome of the others. This is when there is sample with replacement.

R and Binomial Fun

Probability of getting 25 or more heads

```
mean(rbinom(100000, 50, 0.5) >= 25)
```

Probability of getting 35 or more heads

```
mean(rbinom(10000, 50, 0.5) >= 35)
```

Probability of getting 49 or more heads

```
mean(rbinom(10000, 50, 0.5) >= 49)
```

Confidence Interval

- A range of values that is likely to contain the parameter of interest within a certain level of confidence
- Use a confidence interval when you want to estimate the value of a population parameter
- If I construct a 95% confidence interval for a random sample of data, I am confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval if I redo the test again.
- This does *not tell* you how likely the true value is to lie within the interval
- Related to a hypothesis test, but for a hypothesis test we are comparing two samples and asking if some parameter is different
  - Use a hypothesis test when you want to determine if some hypothesis about a population parameter is likely true or not
  - Hypothesis tests account for the variability of each sample and the relationship between samples. That is *really important*!

- Comparing confidence intervals is a sloppy way to do a hypothesis test that does not account for sample variance
- Sensitive to variance
- Confidence interval *is not* the same as confidence level

How to Calculate Confidence Intervals

- $\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$
- $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ (if population standard deviation σ is known)
  - a is the area under the normal curve outside the confidence interval area
  - $\frac{a}{2}$ is the area under one tail of the distribution

Correlation vs Causation

The errors

- Post hoc fallacy: regression towards the mean
- Hawthorne effect
- Gambler's fallacy/ Monte Carlo fallacy
- Hidden cause
- Big data and fishing for correlations
- Humans favor anecdotes, see patterns where there are none, and are susceptible to bias

---

250930 Day 5

Z-Score

- Standard normal
- Transformation: $Z = \frac{(X-\mu)}{\sigma}$
- The z-score is the standard deviation
  - Z-score of 2 is the value at a standard deviation of 2
- ANY normal distribution can be converted into a z-distribution
- Area under the curve

Confidence Interval

- A range of values that is likely to contain the mean within a certain level of confidence
- Use a confidence interval when you want to estimate the value of a population parameter
- Use a hypothesis test when you want to determine if some hypothesis about a population parameter is likely true or not
- Comparing confidence intervals is a sloppy way to do a hypothesis test that does not account for sample variance

Binomial data confidence interval

$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Normal data confidence interval

$\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$

> ⚠ **Confidence intervals are not the same as a t-test!**
>
> T-tests account for the variability of each sample and the relationship between samples.

Z-Statisitic	T-Statistic
$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$	$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$
uses population standard deviation	uses sample standard deviation

Hypotheses and Hypothesis Testing

Decision Theory: Errors

- Type I error: Reject the null (conclude something is going on) when it really isn't
- Type II error: Accept the null (conclude nothing is going on) when it really is
- Significance testing emphasizes the probability of a type I error: all the alarm bells are genuine

What does that p-value mean?

- Is NOT the likelihood that your alternative hypothesis is correct
- Is highly dependent on sample size
- CANNOT compare p-values for samples of different sizes

Jury Pool Example

- What kind of tests are we running here?
  - Another way of wording this is that the research hypothesis is that there are fewer black jurors than we would expect based on chance alone. That would imply that jurors that looked like the defendants were excluded from the jury pool deliberately (that is not something the statistics could tell you)
  - So we are doing a left-handed test. We suspect that the number of jurors was fewer than would be expected
- P-value and a level (not p from binomial!) if p <= a then we REJECT the null
  - Null hypothesis - H_0: The jury was randomly chosen from the population with $p = 0.50$, $p_0 = p$

- Alternative/Research hypothesis - H_R: The jury was not chosen randomly, $p < 0.5$ or (0.1, or 0.05 etc.) $p < P_0$
- Pr (x </=4| p = 0.5 and n = 80) = 1.4 x 10^-18
- a = 3.6 x 10^-18
- So, reject the H_0 if p </= a, 1.4 x 10^{-18} < 3.6 x 10^{-18} therefore we reject the null

One Proportion Tests

Test for	H_0	Test Statistic	Use when
Pop. mean μ	$\mu = \mu_0$	$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$	Normal dist. or n > 30, σ known
Pop. mean μ	$\mu = \mu_0$	$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$	n <30 and/or σ unknown
Pop. prop. p	$p = p_0$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$n\hat{p} \geq 10,\ n(1-\hat{p}) \geq 10)$

One-Proportion Z-test

- Sample needs to be large enough for the normal approximation to work. If it isn't, there are other methods (i.e. binomial exact test). See the Practical Statistics for Data Scientists
- Treating the proportion like a mean
- Example:
  - Current political views vs. historic
  - "Success"

Two Proportion Z-test

- Samples are independent
- Large sample size so that the distribution approximates normal
- Again, treating these as if they are a test of means
- Examples:
  - Rates of cancer/dementia type in different populations
  - Lyme disease in different states
  - Church attendance for Democrats vs. Republicans
  - Aspirin and survival rates for those with CVD and without CVD
    - Success = death (yes, I know)

One Sample Tests of Means

One Sample Comparison of Means

- General formula: (sample statistic - population parameter) / (standard deviation of statistic)
- Comparing an observed (sample) mean to a population mean: is this group different?

Test Statistic when σ Known	Test Statistic when σ Unknown
$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$	$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$, $df = n-1$

$\bar{x}$: sample mean
μ : population mean
σ : population standard deviation
s : sample standard deviation
n : sample size
df : degrees of freedom

Z-Test: Large Sample & Population Standard Deviation Known

- When might this be?
- Assumptions:
  - Data are continuous
  - Simple random sample
  - Normally distributed
  - Population standard deviation is known
- Example:
  - Dietary supplement: label states 3mg of melatonin per gummy bear $\mu = 3$
  - The supplement was validated by an outside group as consistently having an average of 3mg of melatonin per gummy near in a sample of 100,000 gummies from randomly selected batches. They observed standard deviation (σ) of 0.5
  - Production was moved to a new facility, and consumers suspect that the gummy bear no longer contains the correct amount of active ingredient
    - This could be a one or toe tailed test (do they suspect it is low, high, or just different?)
    - A random sample of 100 gummy bears was taken from the most recent batch of supplements. The sample mean was 2.8
    - The denominator is known as the Standard Error of the Mean (SE) or standard sampling error
    - H_0 : the amount of melatonin in the gummies has not changed
      - $H_0 : \mu = x_{bar}$ if $p \geq 0.01$
    - H_A : the amount of melatonin in the gummies has declined
      - $H_A : \mu = x_{bar}$ if $p < 0.01$
    - $a = 0.01$
    - z = (2.8 - 3)/(0.5/sqrt100) = -4
      - Pr(z < -0.37) | H0 = 0.000032
      - Reject the null. These gummies do not have the same concentration of melatonin

One Sample t-test

- Small sample size
- Population variance unknown
- As with the z-test, we will be comparing the result of our test statistic to a pre-determined alpha
- One tailed, two tailed
- Example:
  - Company wants to claim their batteries last more than 40 hours
  - Simple random sample of 15 batteries
  - Sample mean = 44.9, SD = 8.9
  - Choose a of 0.05
  - $H_0 : \mu = 40$
  - $H_a : \mu > 40$
  - $\hat{x} = 44.9,\ \mu = 40,\ s = 8.9,\ n = 15,\ df = 15 - 1 = 14$
  - test statistic: $t = \frac{44.9 - 40}{\frac{8.9}{\sqrt{15}}} = 2.13$
  - p-value: $P(t_{df=14} > 2.13) = 0.026$
  - Because $p = 0.026 < a = 0.05$ we reject H_0

Two Sample Tests of Means

- General formula: (sample statistic - population parameter) / (standard deviation of statistic)
- Comparing the means of two populations? Are these two population means equal?

Two Sample t-test (Student's t-test)

- Samples are independent (observations in one sample have no connection to the observations of the other) otherwise perform matched pairs t-test
- Data should be approximately normally distributed
- Samples should have approximately the same variance (if not, Welch's t-test)
- Pooled variance is rare in real life, but the calculation is simpler and has tighter tolerances
- Degrees of freedom (df)

> **Student's t-test**
>
> $t = \frac{\mu_1 - \mu_2}{s_p^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

F-Statistic

> **F-Statistic**
>
> $F = \frac{s_1^2}{s_2^2} = \frac{largervariance}{smallervariance}$
> If $F \leq t$, then pool

> If $F > t_{(a,df)}$, then don't pool

- This is a hypothesis test of it's own (and therefore costs us a degree of freedom)!
- We can use a F-test to determine if:
  - Two samples come from populations with equal variances?
  - Does a new condition (treatment/process etc.) reduce or increase the variability of some existing condition.
- We will use this test again in the ANOVA and for regression analysis
- Example:
  - Study habits of students in electrical engineering and physics. Each claims their students study more.
  - $H_0 : \mu_1 = \mu_2$
  - $H_a : \mu_1 > \mu_2$
  - School 1: $\bar{x}_1 = 16.85,\ s_1 = 4.31,\ n_1 = 65$
  - School 2: $\bar{x}_2 = 15.79,\ s_2 = 4.97,\ n_2 = 75$
  - $F = \frac{s_1^2}{s_2^2} = \frac{largervariance}{smallervariance} \rightarrow F = \frac{4.97^2}{4.31^2} = 1.33$
  - Since $F = 1.33 < t = 1.65$, pool variances
  - test statistic: $t = \frac{(16.85-15.79)-(0)}{(\sqrt{\frac{(65-1)(4.31)^2+(75-1)(4.97)^2}{65+75+-1}})(\sqrt{\frac{1}{65}+\frac{1}{75}})} = 1.34$
  - p-value = $P(t > 1.34) = 0.09$
  - Because $p = 0.09 > a = 0.05$ we fail to reject H_0

Welch's t-test (non-pooled variances)

- Samples are independent
- Data should be approximately normally distributed
- Samples DO NOT (or at least are not assumed to) have the same variances
- Use as default: results will be equal if the variances are the same and is more accurate if they are not
- Fewer degrees of freedom (degrees of freedom: $n_1 + n_2 - 2$)

> **Welch's t**
>
> $t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Mann-Whitney U Test (Wilcoxon rank sum)

- Samples are independent
- Data ARE NOT normally distributed
- Sample sizes are small
- Null: the two populations are equal

> **Mann-Whitney U Test (Wilcoxon Rank Sum)**

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Paired Samples/Matched Pairs Test

- Samples are not independent: before and after, etc.
- Data should be approximately normally distributed
- Becomes a one-sample t-test of the differences

> **Matched Pairs Test**
>
> $t = \frac{\bar{x}_d - \mu_d}{(\frac{s_d}{\sqrt{n}})}$, $df = n - 1$
>
> $\bar{x}_d$: sample mean difference
> μ : population mean difference
> s : sample difference standard deviation
> n sample size

Wilcoxon Signed Rank Test

- Samples are not independent
- Data ARE NOT normally distributed (paired t is fairly robust to departures from normality)
- Calculated similarly to Mann-Whitney U

251007 Day 6

Storytelling and the Art of the Presentation

Why is storytelling important?

- Why has the data been collected?
- What can it tell us?
- Why do we care?

A good story has

- A beginning
- A middle
- An end
- So does good code! And a good research paper and...

The story of the data is the CONTEXT

- This is why we put a hypothesis in words, not just in mathematical notation
  - Does the question we are asking make sense?
  - Is it important?
- Random analysis without contest is meaningless

Step one: A good story has a hook

- A clear, concise explanation
- A picture
- A graphic
- Why should I care?

The importance of good visuals

- Simple analysis presented clearly > complicated/difficult presented poorly
- What did you do?
- Why did you do it?
- What does it show?
- What is the explanation?
- What is the next step?

Keep it simple!

- Your graph skeleton shouldn't be spooky

251021 Day 7

Assessing Bias

- https://mediabiasfactcheck.com/
- Who is the audience?
- What is the purpose?
- Who are the authors?
- Is the source reputable?
- How long ago was it published?
- Supporting documentation?

Bad Practices

- Misleading graphs
- Wrong test
- Start with the conclusion and work to the answer
- P-hacking

- Data dredging/mining

> "If you torture the data long enough, it will confess."

Harking

- Hypothesizing after the results are known

P-Hacking

Fishing, Data Dredging

- Exploratory Data Analysis
- Can provide

Cherry Picking

- All of your data must be reported - if there were errors, explain them, but do not fail to report
- All analyses conducted must be reported including the statistically insignificant ones whether or not they support your narrative

Chi-Square

Pearson's Chi-Square Tests

- Non-parametric
- Test a hypothesis about the distribution of a categorical variable

Chi-Squared Assumptions

Other Chi-Squared Tests

- Test of homogeneity is a special case of the test of independence

McNemar's Test

- Uses the chi-square test statistic but is not a Pearson's Chi Square
- Used when you have a related pair of categorical variable with two groups each and you want to determine if the proportion of the variables is equal

---

251104

Comparing Multiple Means

Multiple Means

- T-tests: allow us to compare 2 means to one another
- What if you want to compare multiple means?
  - Average carbon dioxide emissions for all 50 states or for different power generation

Why not just do multiple t-tests?

Probability of making a Type-I error increases with each test

Solution 1: Control for the error rate

- ANOVA: Analysis of variance
- This is effectively fitting a regression model to the data and seeing if the independent variables differ (generalized linear models/GLMs)
- The good - simple to compute, easy to interpret
- The bad - tells us if at least one group is different, but does not tell us which group is different ot in what way

---

251111

Reporting Results of Transformed Data

> ⚠ **Danger Zone!**
>
> DO NOT use a transformation to make categorical data fit a normal curve. There are appropriate

---

251125

Simple Linear Regression

- Regression: Best fit line, least squares regression line
- Standardized variables

Assumptions

- Linearity: there is a linear relationship between the data
- Independence: no correlation between consecutive residuals in the time series
- Homoscedasticity: residuals have constant variance at every level of x

Running a Simple Linear Regression in R

```r
#fit simple linear regression model
model <- lm(score~hours)

#view model summary
summary(model)
```

Interpreting the Results:

- Residual standard error:
- Multiple R-squared:
- p-value associated with model coefficients:

Assessing the Fit of the Regression Line

- Coefficient of determination: R^2 : Proportion of the variance in the response variable that can be explained by the predictor variable
  - Ranges from 0 -1
  - 0: response variable cannot be explained by the predictor variable at all
  - 1:

Multivariate Linear Regression