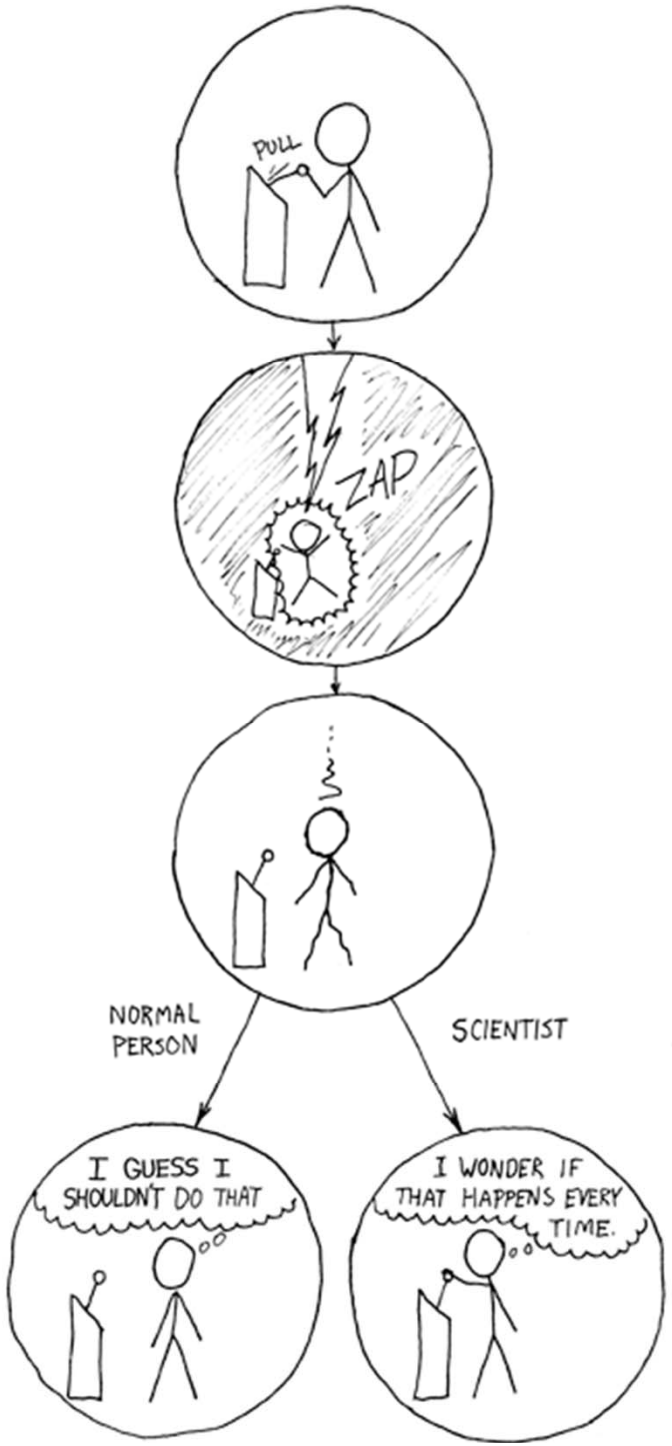
UNIVERSITY of WASHINGTON

Data Science UW
Methods for Data Analysis

Hypothesis Testing and Outliers
Lecture 3
Steve Elston


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

Topics

- Review
- Conditional Probability Trees
- Sampling Methods
- Hypothesis Testing
- Dealing with outliers

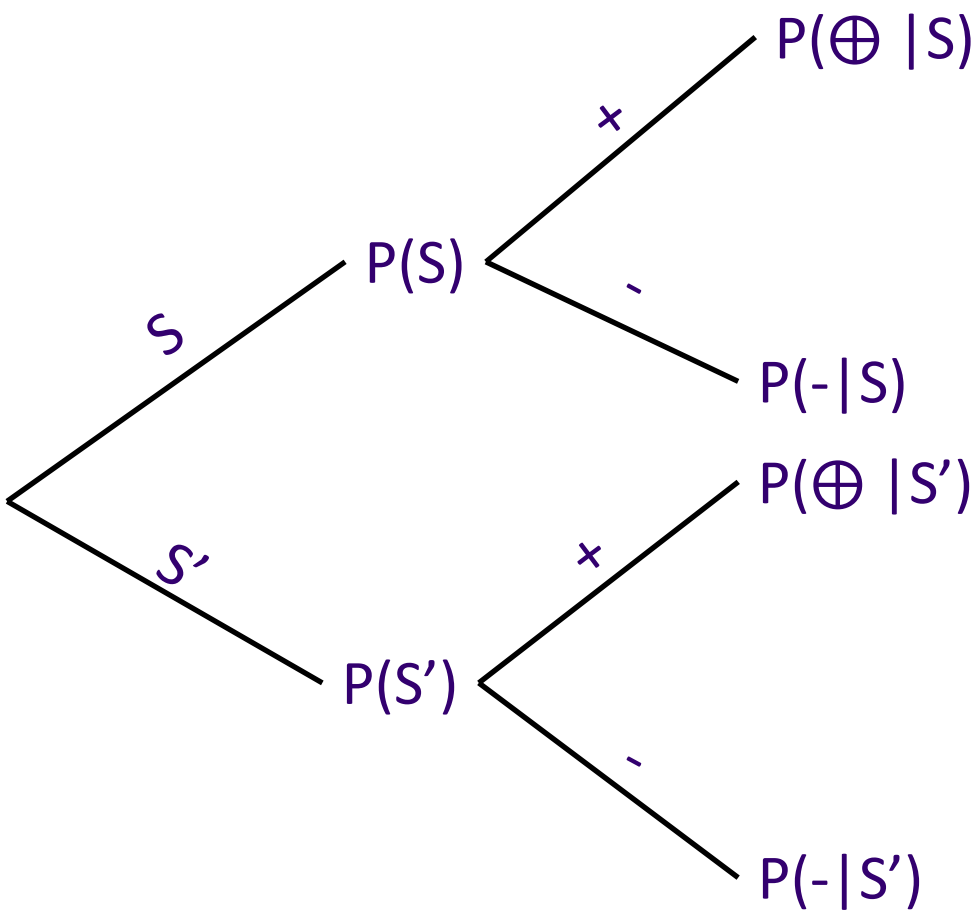
Review

- Counting
  - Factorials
  - Permutations
  - Combinations
- Probability
  - The 3 axioms
    - Probability is bounded between 0 and 1.
    - Probability of the Sample Space = 1.
    - The probability of finite *mutually exclusive* unions is the sum of their probabilities.
  - Conditional probability
  - Independent events
- Missing Data

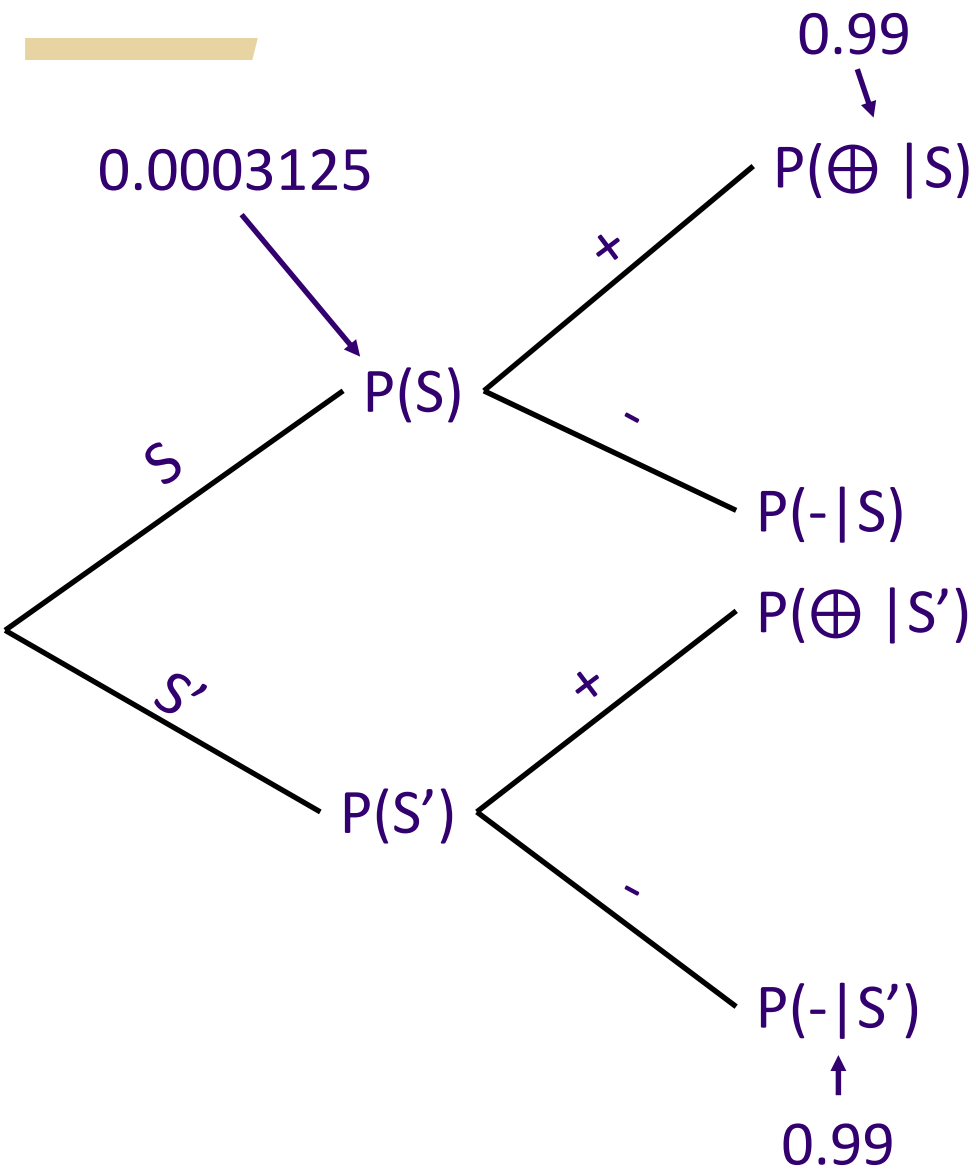
Conditional Probability Trees

- Let's consider a test for a Sickle Cell Anemia.
- Events:
  - S = patient has Sickle Cell Anemia
  - S' = patient does not have Sickle Cell Anemia
  - $\oplus$ = patient tests positive
  - - = patient tests negative
- Rate in US = 1/3200. P(S) = 1/3200 = 0.0003125.
- Medical company tells us that a test is 99% accurate.
  - $P(\oplus \mid S) = 0.99$
  - $P(- \mid S') = 0.99$

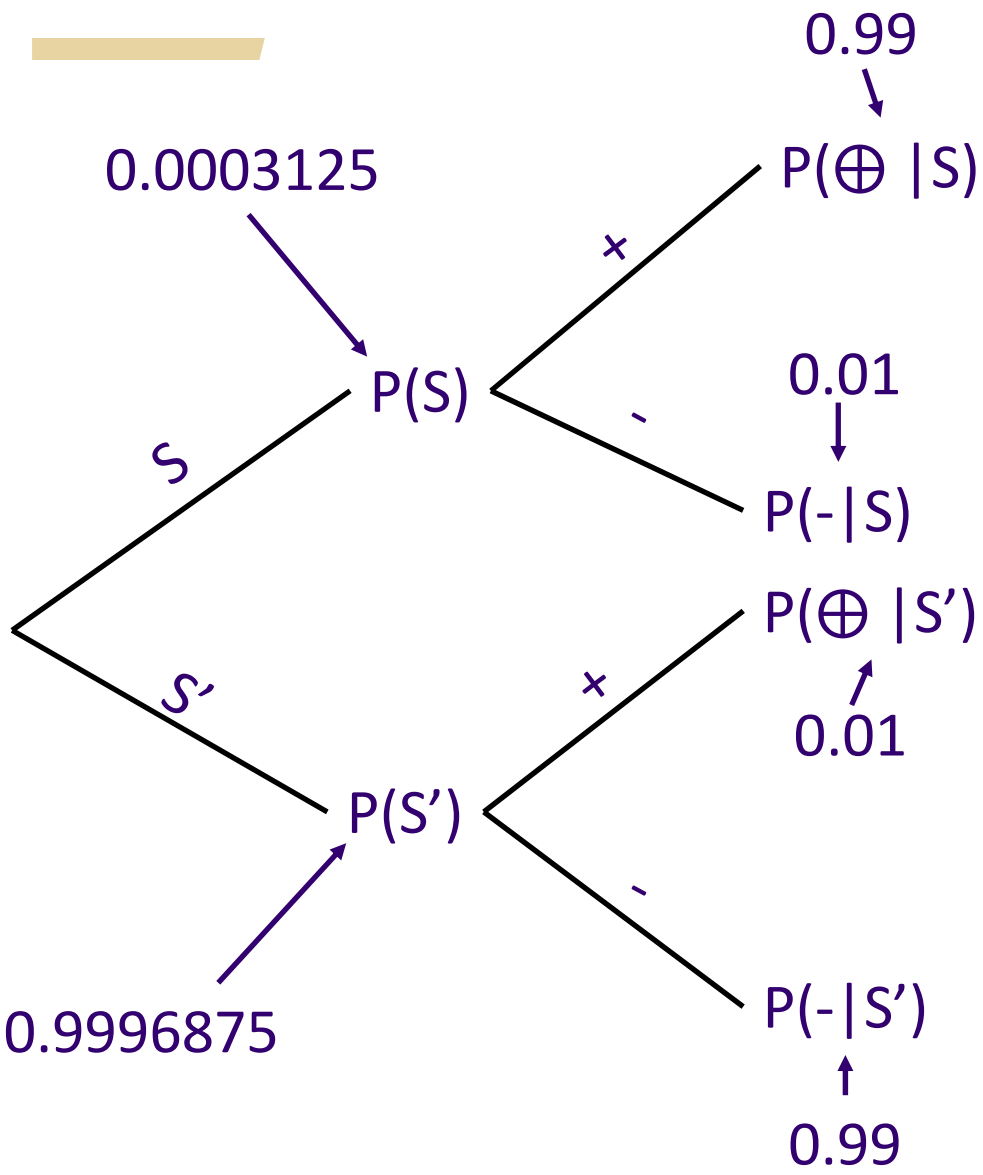
Conditional Probability Trees



Conditional Probability Trees



Conditional Probability Trees



Conditional Probability Trees

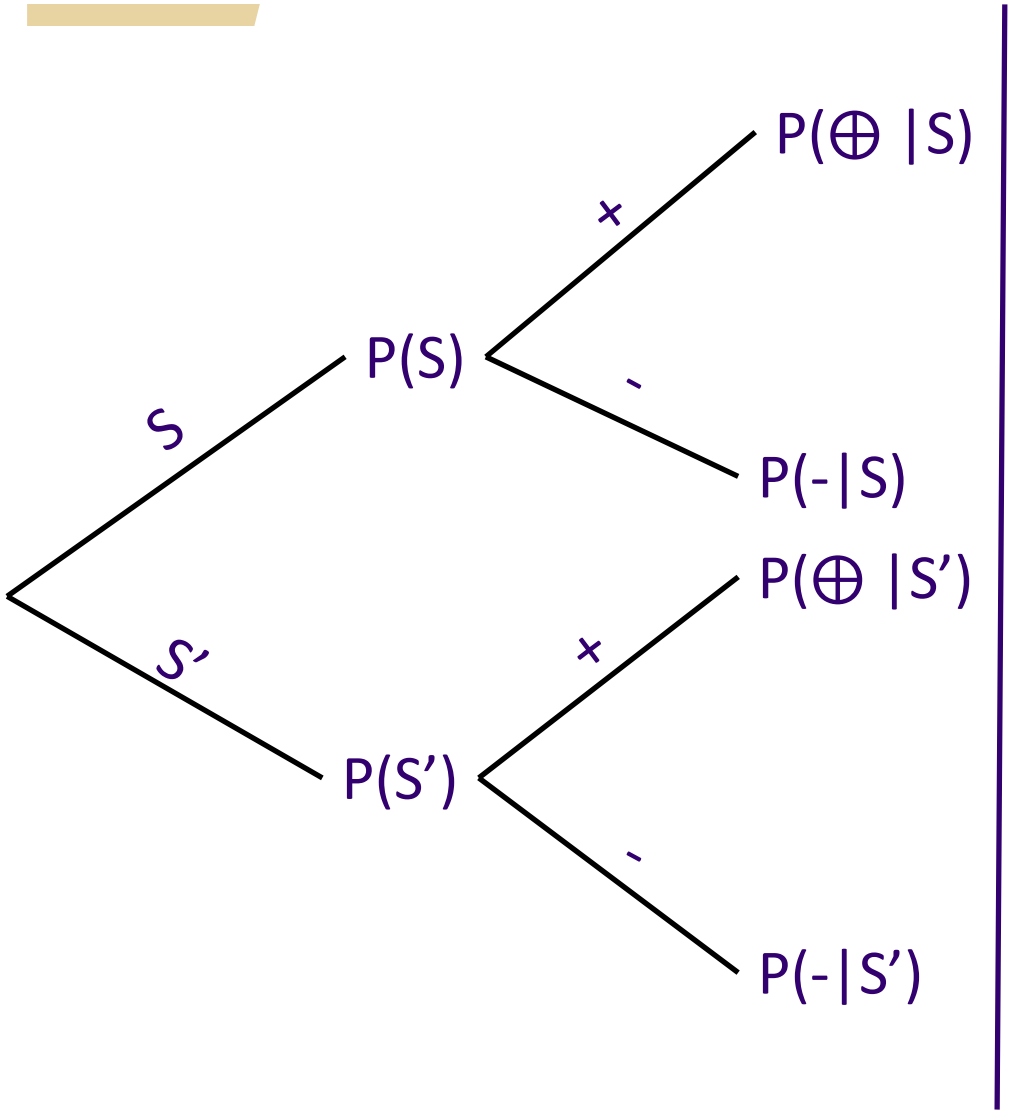
- What we really want to know is:
  - What is the P(S| $\oplus$)?
  - Also important to know: P(S|-)?
- From conditional probability definition:

$$P(S|\oplus) = \frac{P(S \cap \oplus)}{P(\oplus)}$$

- We also know that

$$P(\oplus) = P(\oplus \cap S) + P(\oplus \cap S')$$

Conditional Probability Trees



$P(\oplus \cap S) = P(S)P(\oplus|S)$

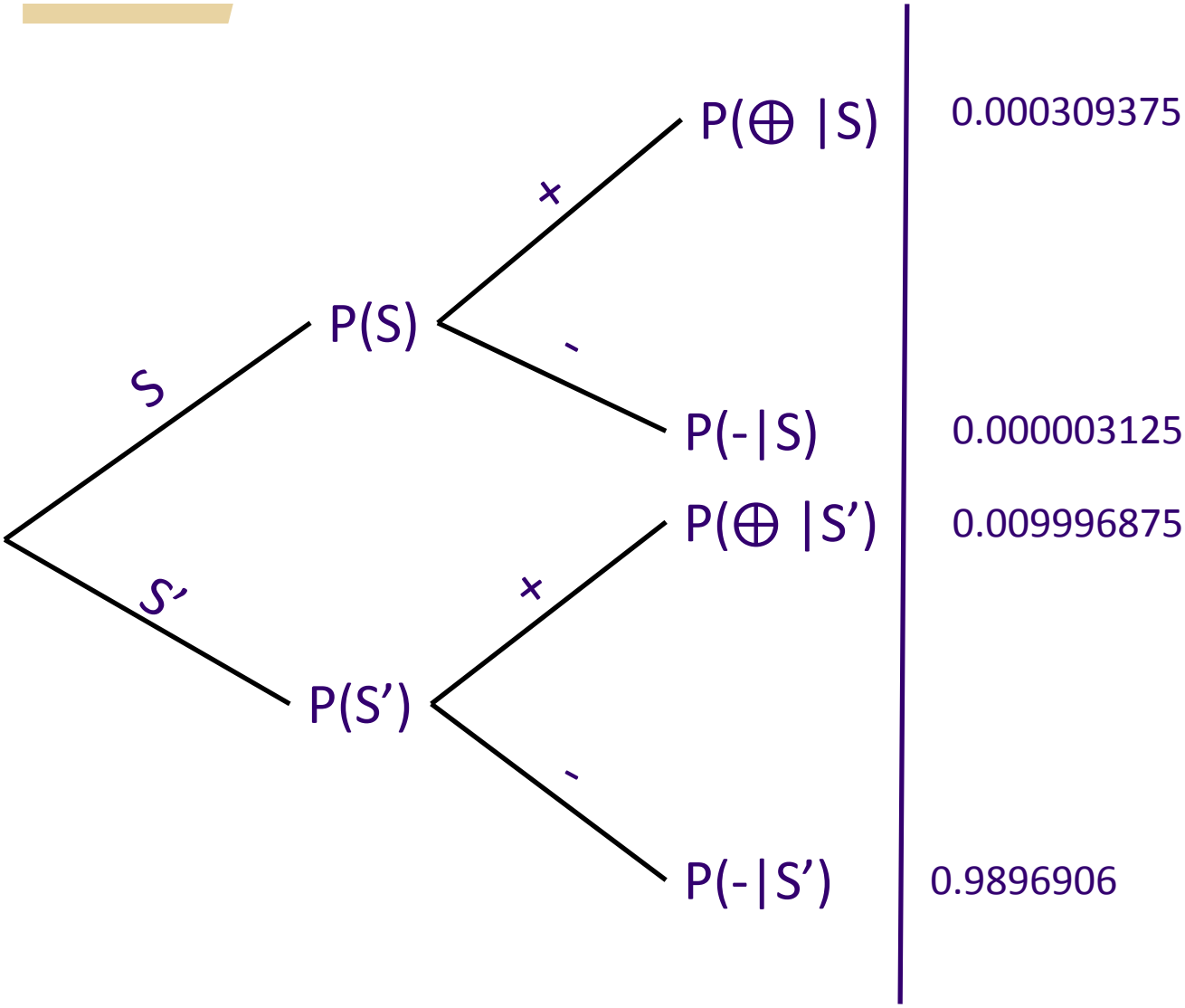
0.0003125*0.99=0.000309375

$P(-\cap S) = P(S)P(-|S)$

$P(\oplus \cap S') = P(S')P(\oplus|S')$

$P(-\cap S') = P(S)P(-|S')$

Conditional Probability Trees



Conditional Probability Trees

$$P(\oplus) = P(\oplus \cap S) + P(\oplus \cap S')$$

$$P(\oplus) = 0.01030625$$

0.000309375 = $P(\oplus \cap S)$

$$P(S|\oplus) = \frac{P(S \cap \oplus)}{P(\oplus)}$$

0.000003125 = $P(-\cap S)$

0.009996875 = $P(\oplus \cap S')$

$$P(S|\oplus) = \frac{0.000309375}{0.01030625}$$

$$\boldsymbol{P(S|\oplus) = 0.03001819}$$

Similarly,

$$P(S|-) = 0.000003157543$$

0.9896906 = $P(-\cap S')$

*Now see the probability interview question

Sample vs. Population

- Sampling is important because we can almost never look at the whole population.
- We use inferences on the sample to say something about the population.
- We need estimates of variances on the sample calculations to say something about the population.

	Sample	Population
AB Testing	The users we show A and B versions of the website.	All users that visit our site. (Past, present and future)
World Cup Soccer	Only 32 teams post qualification in one season.	All national teams in the world for four years.
Average height of Data Science Students	UW Methods for DS Class	All DS students.

Sample vs. Population

- If we sampled 4 beers and the ABV was [4%,5%,5%,6%], then the sample mean would be 5%.
- While there is variance in the sample, there is NO variance in the mean of that sample!!!!!!
- But if we want to say something about the population, we provide the mean with a variance statistic.
  - This allows us to say something to the effect of ‘There is a 90% chance that the mean of all beers lies between 4.5% and 5.5%’.
  - In order to say something about the population we have to know how the sample was generated.

Sampling Strategies

- Convenience or Accidental Sampling (This is bad).
  - Grabbing whatever is easier.
- Bernoulli Sampling
  - Every point subjected to a probability of being selected. Simple Random Sample (most common)
  - Fixed size Bernoulli sampling.
  - Example randomly sample weight of product to ensure quality
- Stratified Sampling
  - Sampling subpopulations in a representative fashion.
  - This is sometimes important for class imbalance problems.
  - Example sample equal numbers of men and women
  - Example sample equal numbers of people in different income categories

Sampling

- Cluster Sampling
  - Divide data into clusters.
  - Select some clusters
  - Sample from selected clusters
  - Can stratify sample from clusters
  - Example, select a few selected store locations and sample customers at these locations
- Systematic Sampling*
  - Sampling every k-th element of a population.
  - Can lead to bias from organization of data

* Can be bad.

Sampling

- Sampling plan
  - Know number of clusters, strata, samples in advance
  - Don’t stop sampling when desired result is achieved: e.g. error measure!
- Note that random sampling, if done properly, controls for database effects, like indexing.
- R demo

Large Samples and Law of Large Numbers

- If we roll a die 60 times and 600 times, which of the dice will more likely have exactly $1/6^{th}$ of the rolls equal to 6 appearing?
  - P(x=10|60trials)=?
  - P(x=100|600trails)=?
  - (Check in R)
- Which die will be more likely to be within 5%?
  - P((1/6-1/20) < x < (1/6+1/20))=P(7/60 < x < 13/60)?
    - P(7 < x < 13 | 60 trails)=?
    - P(70 < x < 130 | 600 trials)=?
  - (Check in R)

Law of Large Numbers!!!

- Sample statistics converge to the population statistics as more unbiased experiments are performed.
  - E.g. The mean of 50 coin flips (0,1)=(T,H) is usually farther away from the true mean of 0.5 than 5,000 coin flips.
  - First proof by Jacob Bernoulli in 1713
- R demo of coin flips

Standard Deviation vs. Standard Error

- Standard Deviation: Measure of variability in a sample or population.
- Standard Error: Measure of variability in the statistics of the sample.
- For example:
  - Standard deviation of a sample.
  - Standard deviation of a set of means calculated from multiple samples.
    - You can imagine that the larger my sample, the more confident we can be about the mean.
  - Standard error of a statistics decreases by a rate of 1/sqrt(n) , where n is your sample size.
- Proof by R demo

Hypothesis Testing

- Identify a hypothesis that can be tested.
  - "Changing our web-site logo to be bigger on the front page will drive more than 100,000 customers to our site per day."
- Select a criteria to evaluate the hypothesis.
  - If our sample has a probability of **>= 90%** chance that there are more than 100,000 customers per day, we accept the hypothesis.
- Select a random sample from the population.
  - Randomly assign a cookie to new site users that tells the server to show A or B website.
- Compare observations to what we expect to observe and calculate statistic.

Hypothesis Testing

- We first state our population assumptions in the *null hypothesis.* (H_0)
- We state our new *alternative hypothesis* as an alternative to the null. (H_a)
- The null + alternative should make up all possible outcomes and be mutually exclusive

H_0: The old website drives equal amount of traffic or more.

H_a: The old website drives less traffic than the new one.

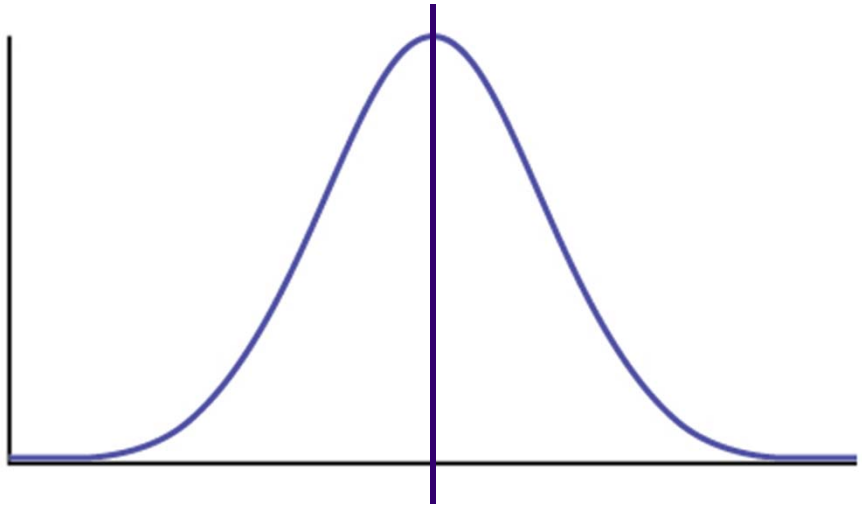
- Decide on a significance level (probability cutoff)
  - 0.9, 0.95, and 0.99 are common (problem specific)

Hypothesis Testing

- Based on our findings we can only do two things
  - We reject the null-hypothesis.
    - Since the alternative covers all other possibilities, we can say we accept the alternative hypothesis.
  - We *fail* to reject the null hypothesis.
    - We do not accept the null hypothesis because we have already believed our null hypothesis from the start.
    - We could have failed for two reasons:
      - The alternative hypothesis was false to begin with.
      - We did not collect enough evidence for the alternative hypothesis.

Hypothesis Testing

- We know that the average time a user spends on a page has a mean of 15 seconds and a s.d. of 4 seconds.
- If we assume normality, how do we test if a change to the page has a higher view time?



Mean = 15 s, s.d. = 4 s

Hypothesis Testing

- We know that the average time a user spends on a page has a mean of 15 seconds and a s.d. of 4 seconds.
- If we assume normality, how do we test if a change to the page has a higher view time?

H_0: The old website has the same or more viewership than the new website.

H_a: The old website has less viewership than the original.

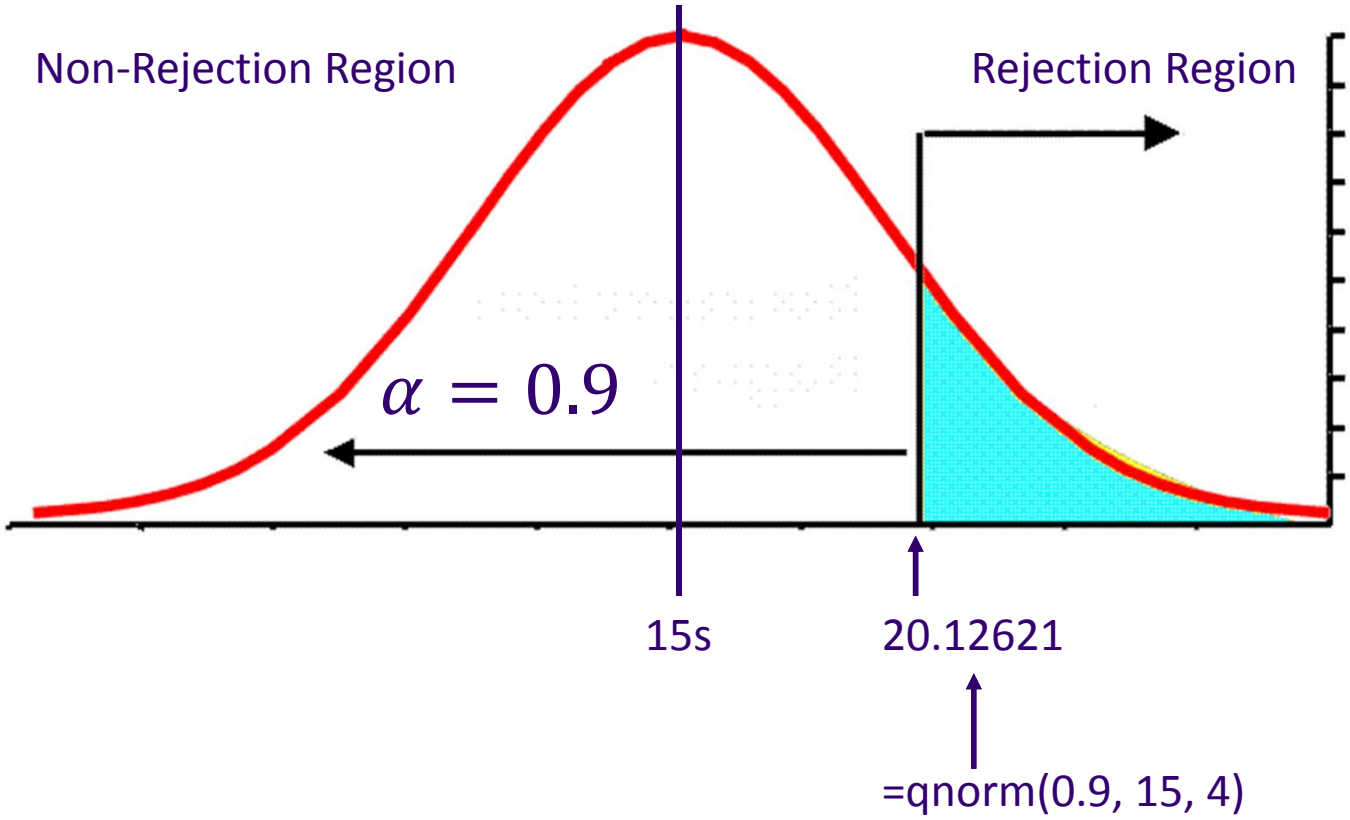
Or…

H_0: The new website has the same or less viewership than the original.

H_a: The new website has more than the original website.

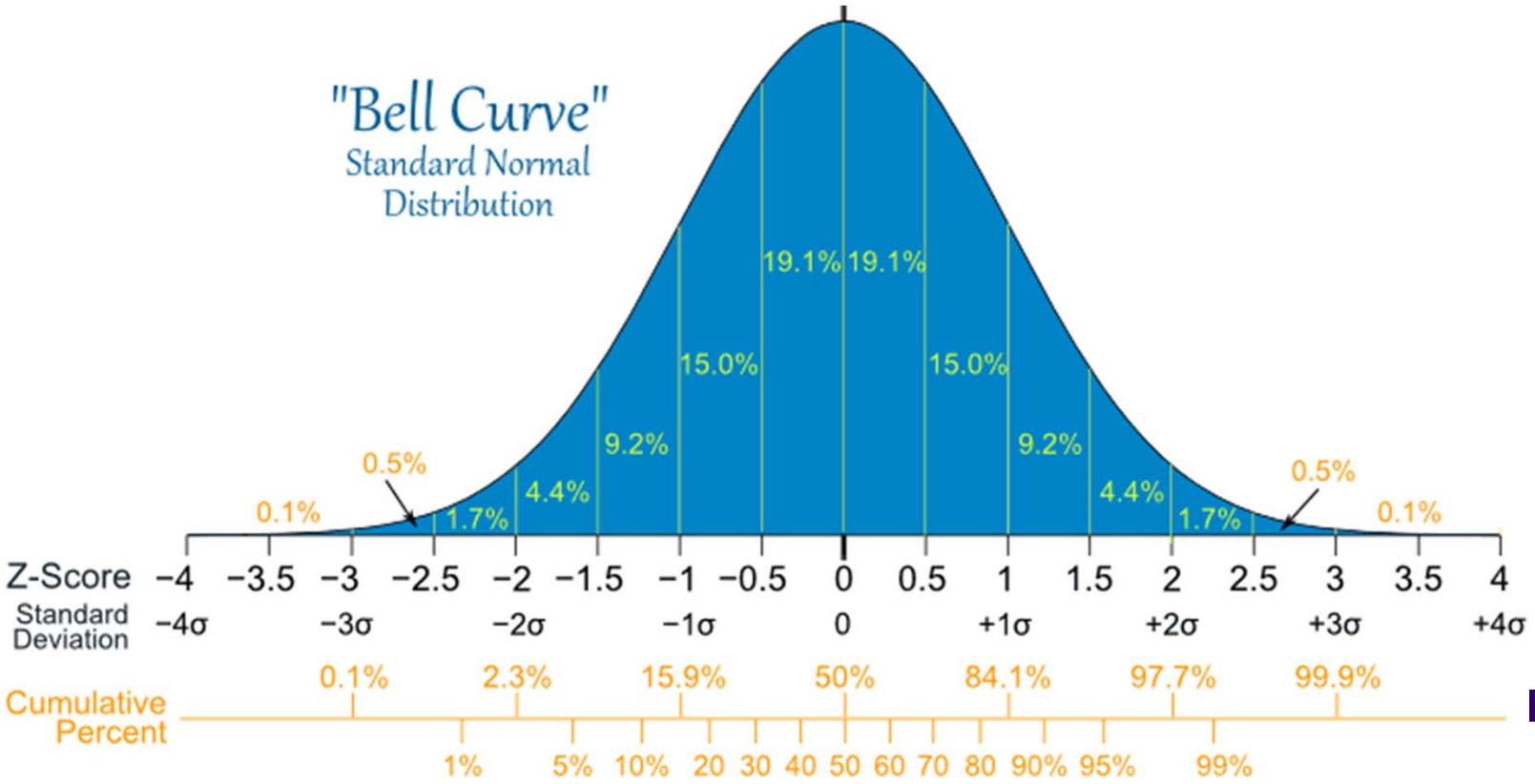
Hypothesis Testing

- We now select a confidence value.
- An event in the **blue** region will have a 10% chance or less of occurring.



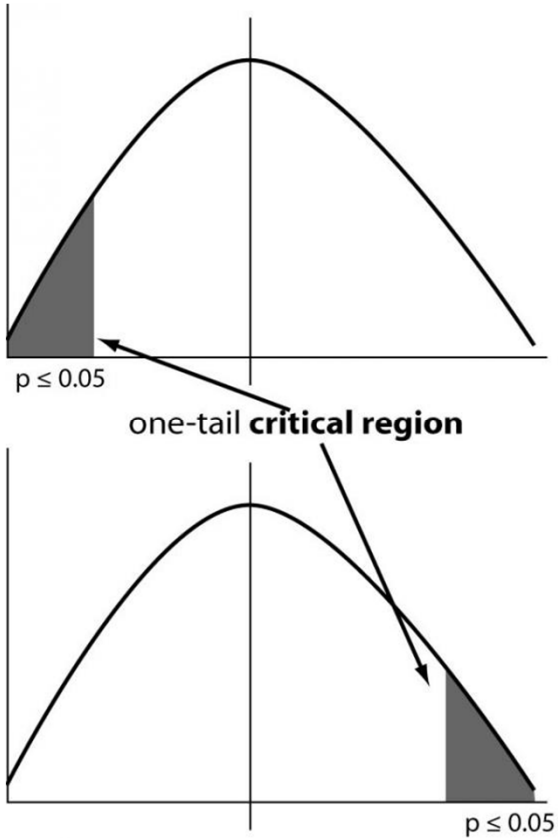
Hypothesis Testing

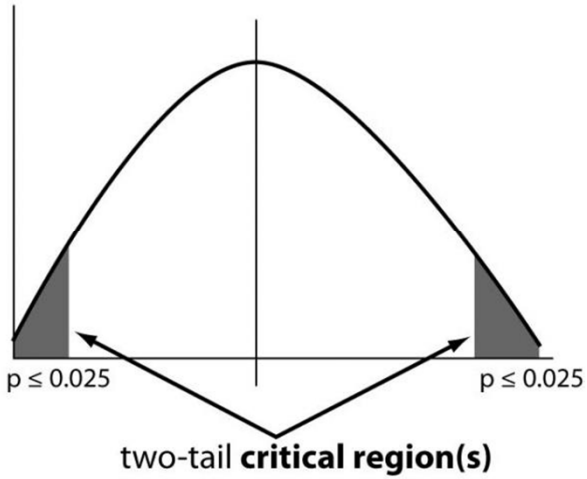
> Probability areas on the normal curve are directly related to the distance to the mean.



Hypothesis Testing

- Asking if a new value is “greater than” or “less than” the null creates a **one-tailed hypothesis test**.
- Asking if a new value is “not equal to” the null creates a **two tailed hypothesis test.**





Definition of P Value

In technical terms, a P value is the probability of obtaining an effect **at least as extreme** as the one in your sample data, assuming the null hypothesis is true.

For example, for a vaccine study with a P value of 0.04, you'd obtain the observed difference or more in 4% of studies due to random sampling error.

P values address **only one question: how likely are your data, assuming a true null hypothesis?**

P value does not measure support for the alternative hypothesis!

Misuse of P Value

The most common mistake: interpreting a P value as the probability of mistakenly rejecting a true null hypothesis (a Type 1 error).

P values calculations assume the null hypothesis is true for the population and the difference in the sample is entierly from random chance. **P values can't tell you the probability that the null is true or false!**

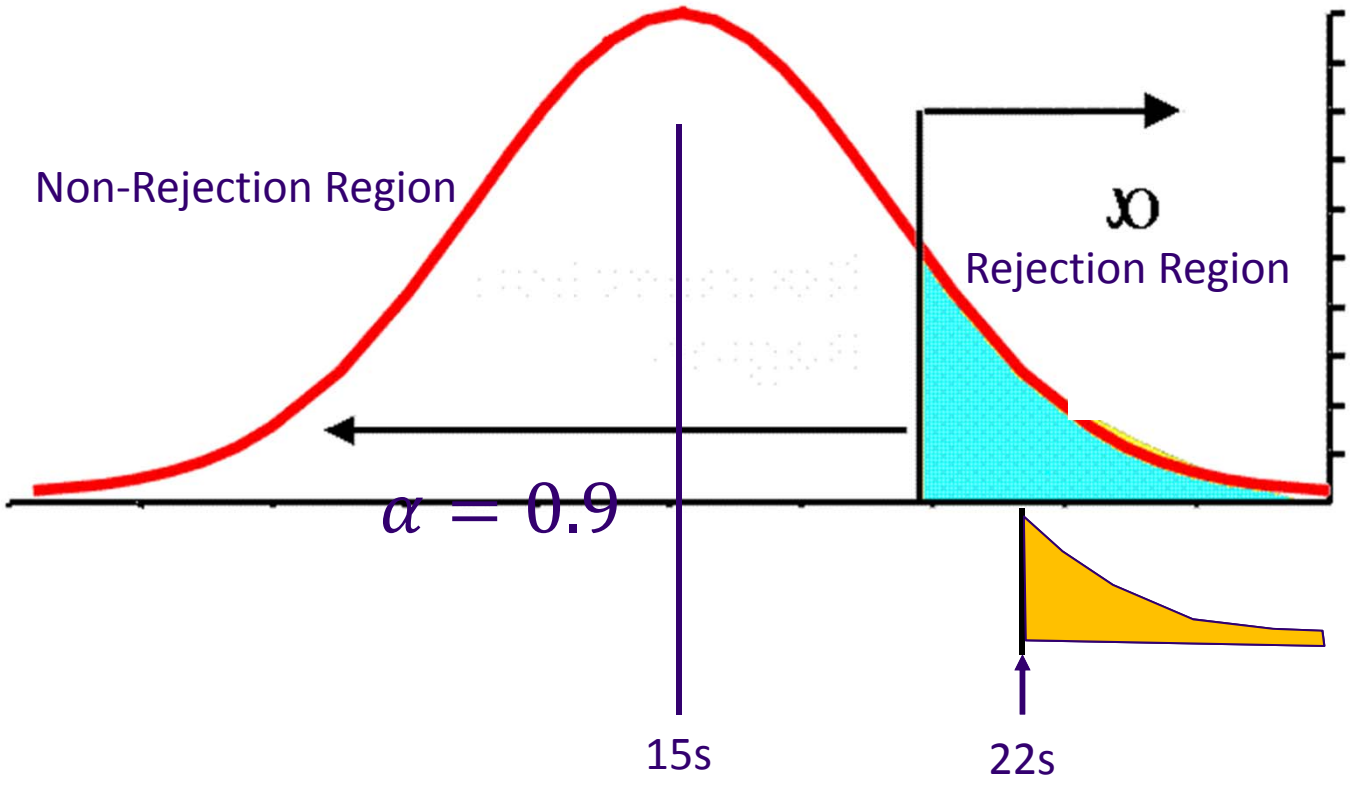
For the vaccine study, correct and incorrect way to interpret a P value of 0.04:

Correct: If vaccine has no effect, **the observed difference or more** arises solely from random sampling error in 4% of studies.

Incorrect: By rejecting the null hypothesis, there's a 4% chance of Type 1 error.

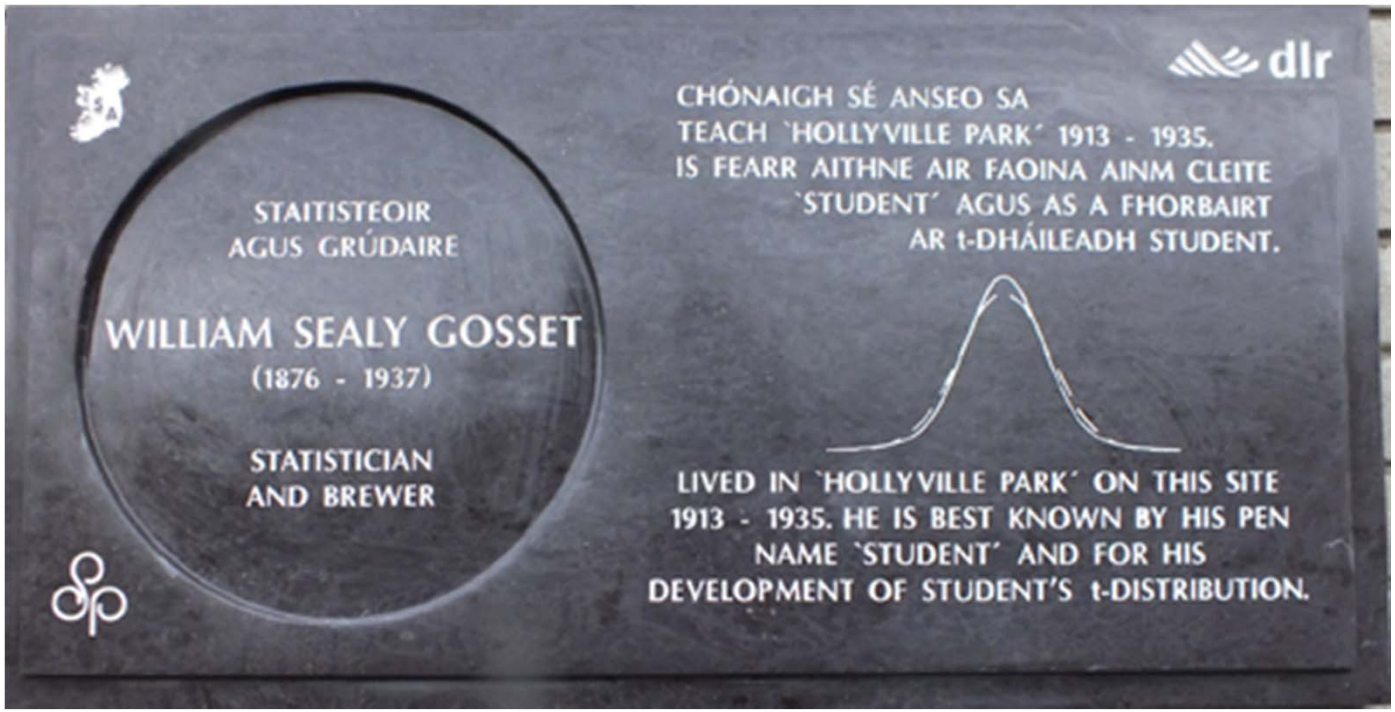
Hypothesis Testing

> Example, what is the p-value of a sample mean of 22 seconds?



Hypothesis Testing

- What if we don't know our population's standard deviation?
- There are involved probability rules that relax the normality assumption in favor of "heavier tails".
- This distribution is known as the Student's T-distribution.



t-test

- Student's T-test: tests a hypothesis about the difference of two sets of data:
  - **Test whether a population mean has a specified value.**
  - **Test the difference between two means (equal, unknown variances).**
  - **Test a paired-response difference from zero.**
    - **E.g. a before/after drug treatment on patients.**
  - **Test whether the slope of a line is not zero.**
    - **Important for testing the importance of variables (later in class).**
- Use 'Welch's T-test' for testing the difference between two means (unknown variances, potentially different).
- Picking the different tests changes test's results.
- The more assumptions we make, the easier it is to tell the difference between populations.

t-test in R

> t.test() demo in R

Summary

- The normal test and t-test are used for testing values from a continuous distribution (or approximately so).

- What if we wanted to test occurrences or count data?

Chi-squared Test (Pearson's)

- Unpaired test for counts in different categories.
- These categories must be mutually exclusive.
  - Does the patient have cancer? (yes/no)
    - Test if the two categories differ in WBC count.
  - Rolling a die. (1,2,3,4,5,6)
    - Test if the six categories occur the same (fair die).
  - Does a tweet contain a specific word? (yes/no)
    - Test if the two categories differ in tweet length or word count.
- This tests whether the different categories differ in some specific value.
- In order to do this test, we have to specify the 'degrees of freedom' in the Chi-squared test.
  - This is equal to n-1. Where n equals the number of different categories.
- The test looks at the sum of the outcome differences from expectations.

Chi-squared Test (Pearson's)

> Example: A-B test with three different outcomes.

	Occurrence	Expectation %	Expectation Counts	Difference	Squared Difference	(Squared Difference)/Expected
Leave Page	55	0.6	=0.6*120=72	=55-72=-17	289	=289/72=4.014
Continue Purchase	43	0.3	=0.3*120=36	=43-36=7	49	=49/36=1.361
Add More to Purchase	22	0.1	=0.1*120=12	=22-12=10	100	=100/12=8.333
Totals	120					13.708

> Test statistic is 13.708 on a chi-squared distribution with (3-1)=2 degrees of freedom.

> Degree of freedom is (# of options minus 1).

> R Demo

Chi-squared Test (Pearson's)

- Chi-squared is also used for a 'goodness of fit' test.
- Test if sample is representative of population.
  - Test if your sample has expected make up of categories.
  - E.g. If our population is 50-50 men-women, then we test if our sample is different from those expected probabilities.
- If our total sample size is small, we see a breakdown of the Chi-squared test. (a subgroup size ~ < 10)
  - We switch to a Fisher's Exact test in these cases.

Fisher's Exact Test

Lady sipping tea test - 1911

Tea-Tasting Distribution

Success count	Permutations of selection	Number of permutations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
Total		70

$$\frac{8!}{4!(8-4)!} = 70$$

Fisher's Exact Test

- Tests for difference between two groups based on ratios.
- Exact test, because it calculates the probability of observing the sample under the null or worse in all possible cases.
  - Not as much statistical 'power' as Chi-Squared.
  - If you have larger sample sizes, and the two categories are sufficiently different, both tests should give similar p-values.
- Probability of observing a specific outcome:

	Cat. 1	Cat. 2
Successes	A	B
Failures	C	D

$$prob = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}}$$

Fisher's Exact Test

- For example, consider:

2	3
3	4

- We sum up the probability of that outcome occurring or worse:

0	5	1	4	2	3	3	2	4	1	5	0
5	2	4	3	3	4	2	5	1	6	0	7

Same outcome or worse, with the same marginals (row sums).

- R-demo

Fisher’s Exact Test

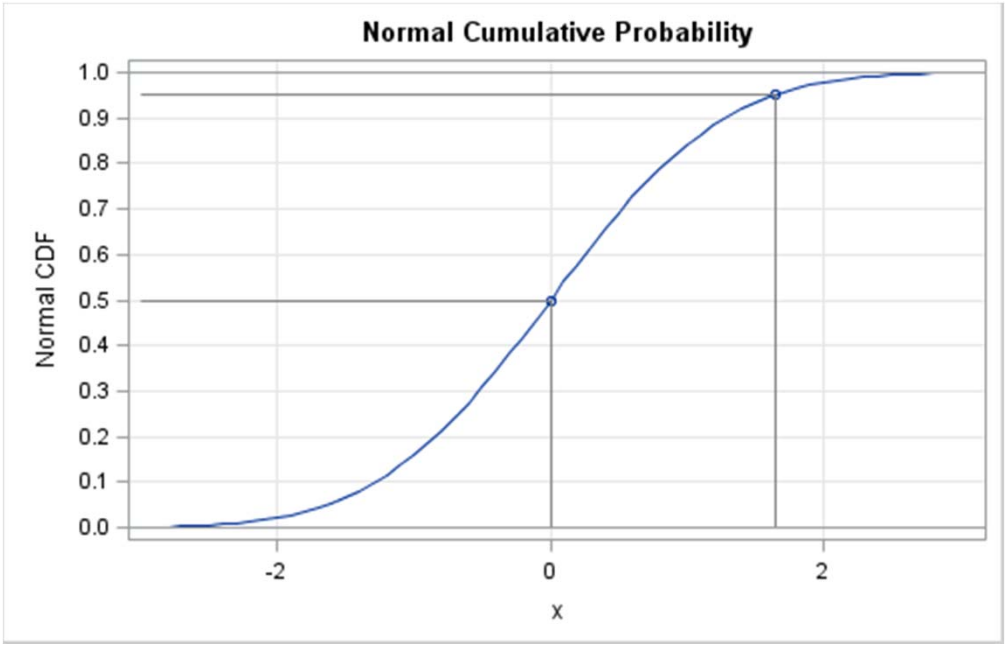
- For paired data, look into McNemar’s test (based on the binomial distribution).

Hypothesis Testing Summary (so far)

- If data is normal,
  - If you know population mean and variance,
    - Use standard normal ‘z-test’.
  - If you just know population mean,
    - Use t-test (unpaired data).
    - Use Welch’s t-test (paired data).
- For categorical comparison tests,
  - If the sample/subgroup size is large enough,
    - Use Chi-squared test
  - If the sample/subgroup size is small,
    - Use Fisher’s Exact test.
- How do we know the data is normal?

Testing for Normality

- Kolmogorov-Smirnov test (K-S test).
  - Tests if two distributions are similar.
- Consider the Normal Cumulative Distribution Function (CDF).



- Any similar distribution should have a similar CDF.

Testing for Normality

- The K-S statistic is just the maximum vertical distance between two CDFs.
- Note: the K-S test can test departure from any hypothetical distribution, not just normal.
- R-demo

Testing for Normality

- Also, the Shapiro-Wilk test can tell us a test statistic for normality.
  - Tests the difference in expected and sample ‘moments’.
  - Moments:
    - 1st moment = mean
    - 2nd moment = variance
    - 3rd moment = skewness
    - 4th moment = kurtosis
    - …
  - Slightly more powerful than the K-S test.

Outliers

- Outlier causes:
  - Bad data
    - Sensor misread, human error, software error
  - Non-representative data
    - Real data that can be argued to be out of our interest. E.g. a sample of annual salaries that includes Warren Buffet.
  - Must provide a legitimate argument to consider as outlier.
  - Or, an **interesting aspect of the dataset** previously overlooked?

Finding Outliers – Statistical methods

- Alpha trimmed mean – aka truncated mean
  - Trim percentage (alpha) of outliers
  - Upper, lower or balanced trimming
  - Iterative method
  - Biased estimator
  - Windsor mean – replace outlier values with trim point values

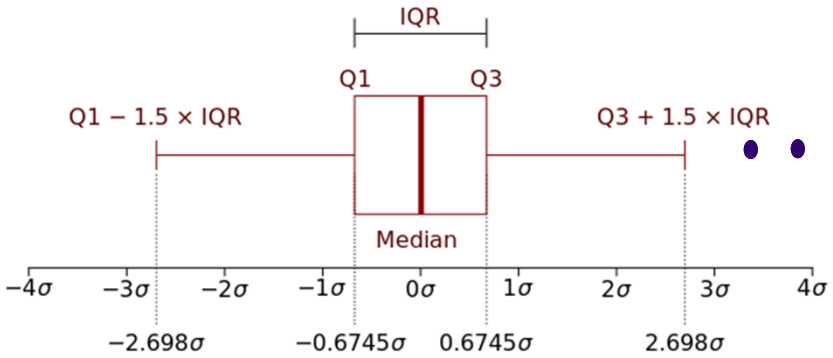
    Tukey et. al. 1947

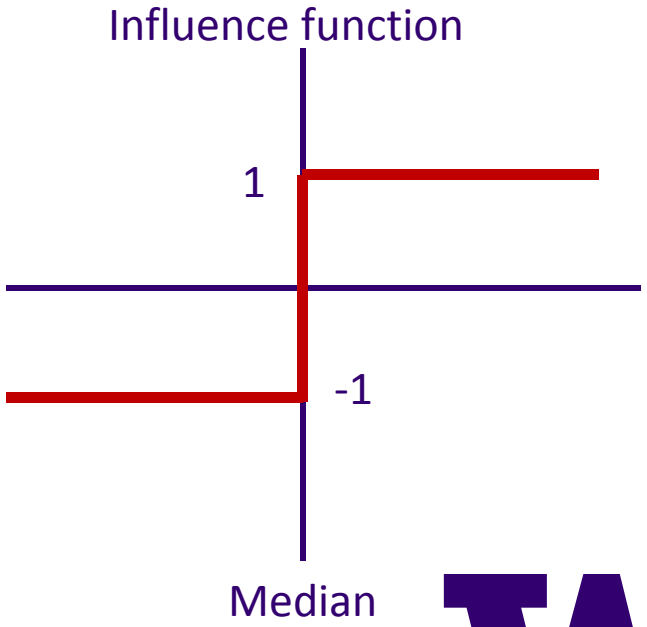
Example with two-sided alpha = 1/3

$$\frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10} + X_{11} + X_{12}}{\text{---}}$$

Finding Outliers – Statistical methods

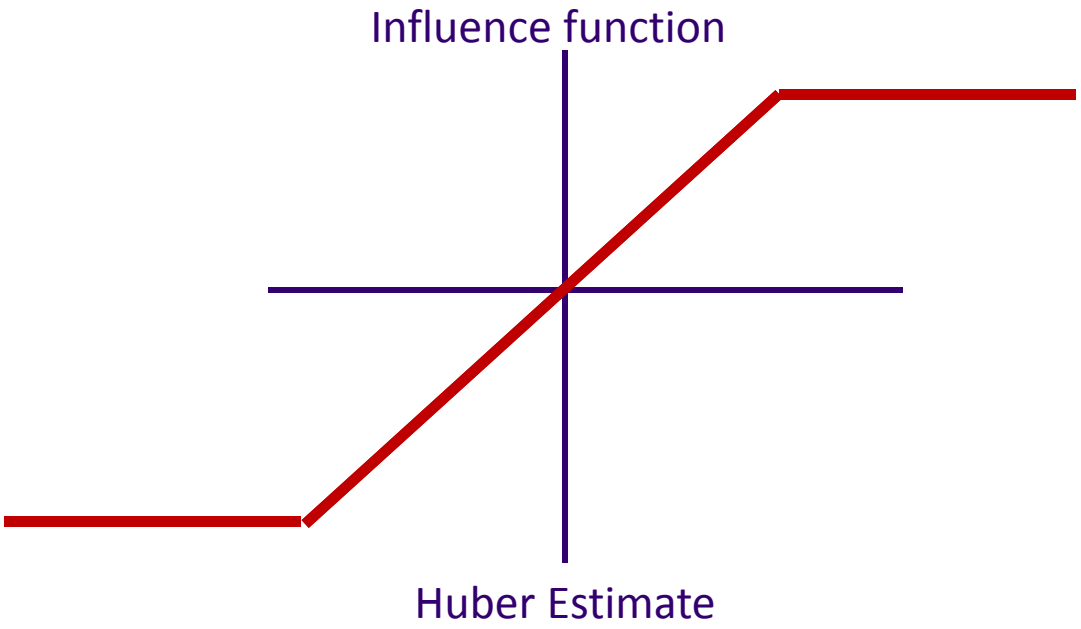
- Median
  - Median is a robust estimator
  - Use interquartile range to detect outliers
  - Biased estimator



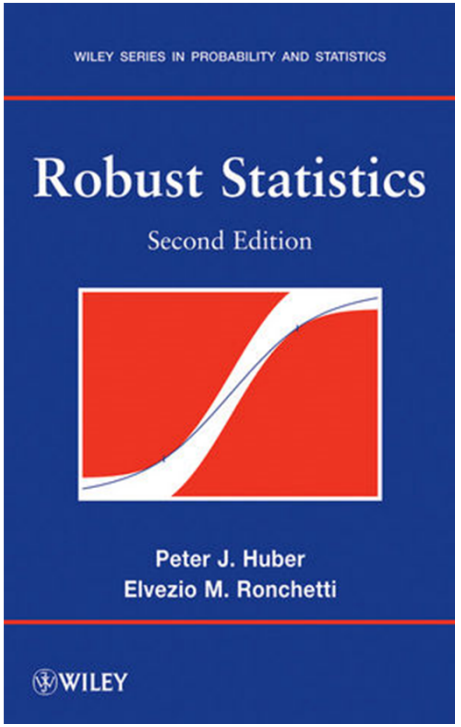


Finding Outliers – Statistical methods

- Huber estimator
  - Attempt to reduce bias
  - Limit influence of outliers
  - Use piecewise influence function



First Edition 1981



Finding Outliers – Statistical methods

- Resampling
  - Find points with exceptional ‘influence’ on the estimate
  - Special case of Jackknife method
  - Computationally intensive

Validating Outliers

Is an outlier an error or a valuable case?

> Investigate multiple relationships in dataset to validate outlier

> Think what interesting or important relationship the ‘outlier’ might represent.

Treating Outliers

- Censor
- Interpolate new value
- Use substitute values

Demo of testing statistical function

Forgot this last time

Final Project

Time to start thinking about your project

> Identify and interesting problem and dataset

https://archive.ics.uci.edu/ml/datasets.html

https://cran.r-project.org/web/packages/nycflights13/index.html

Lots of others!

> Good projects have non-obvious solutions. The real-world is open-ended!

> Let me know what dataset you plan to use and what problem you are investigating

Final Project

- You must present your findings in a professional style!
  - Clear discussion of results, supported by tables and charts
- At the minimum:
  - Explore the dataset with charts and summary statistics
  - Examine several aspects of the dataset; what is important and why?
  - Apply some models; chose for models discussed in course - hypothesis tests, linear models, time series models, etc.
  - Explain choice of models and results
- Professional coding standards and techniques
  - Documented code
  - Maintainable code: use functions, minimize redundant code
  - Show test/validation cases

Assignment

- Complete Homework 3:
  - Test hypotheses for the price of automobiles:
    - Compare and test normality the distributions of price and log price – two tests
    - Test significance of price (log price) stratified by a) fuel type, b) aspiration, and c) rear vs. front wheel drive.
  - You should submit:
    - **R-script** using clear professional style.
    - A text document summarizing and supporting your findings – tables and charts as required
  - Read Intro to Data Science Chapter 6.
  - EMAIL ME! A summary of ideas and/or data sources for your final project in next two weeks.