Sampling and Hypothesis Testing

Lesson 4

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Topics

- **Review Probability**
- **Sampling Methods**
- **Hypothesis Testing**
- **Detecting outliers**

Probability
Round 2

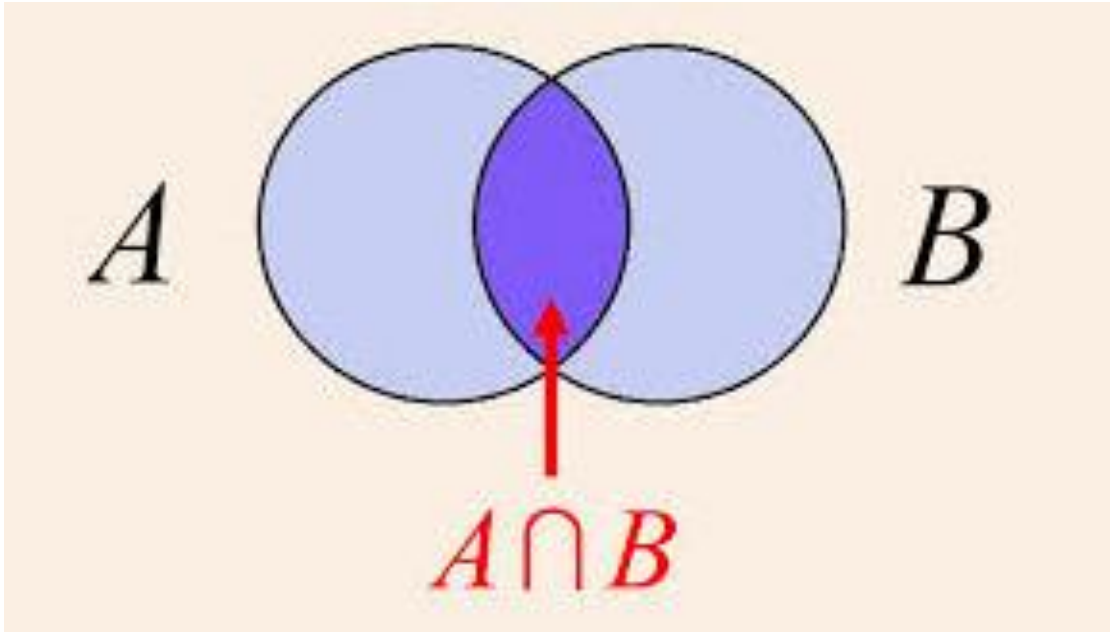
Rule	Notation	Description
General Probability	P(A) = (# outcomes A) / (# possible Outcomes)	Probability of event when outcomes are equally likely
Probability Assignment Rule	P(S) = 1	The set of all possible outcomes must have P = 1
Completement Rule	P(A) = 1 – P(Ac)	P of event occurring is 1 minus P of it not occurring

Independent vs Disjoint

- **Disjoint =**
  - **mutually exclusive events**
  - Cannot occur at the same time
  - Cannot be independent
  - Events have nothing in common, so if one occurs, the other one does not
- **Independent =**
  - the occurrence of one does not change the probability of the other occurring

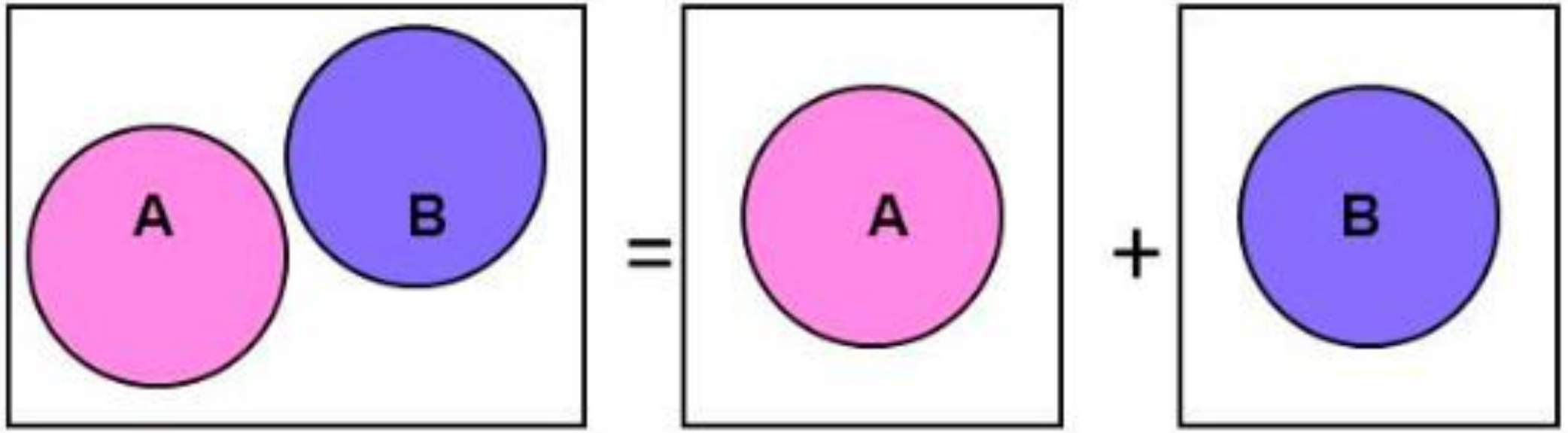
Intersection $P(A \cap B)$

- **P(A and B) = P(A) * P(B)**
  - Independent Events-
    - occurrence of one does not change the probability of the other occurring
  - P(All points in both A and B)
  - Joint probability



Union $P(A \cup B)$

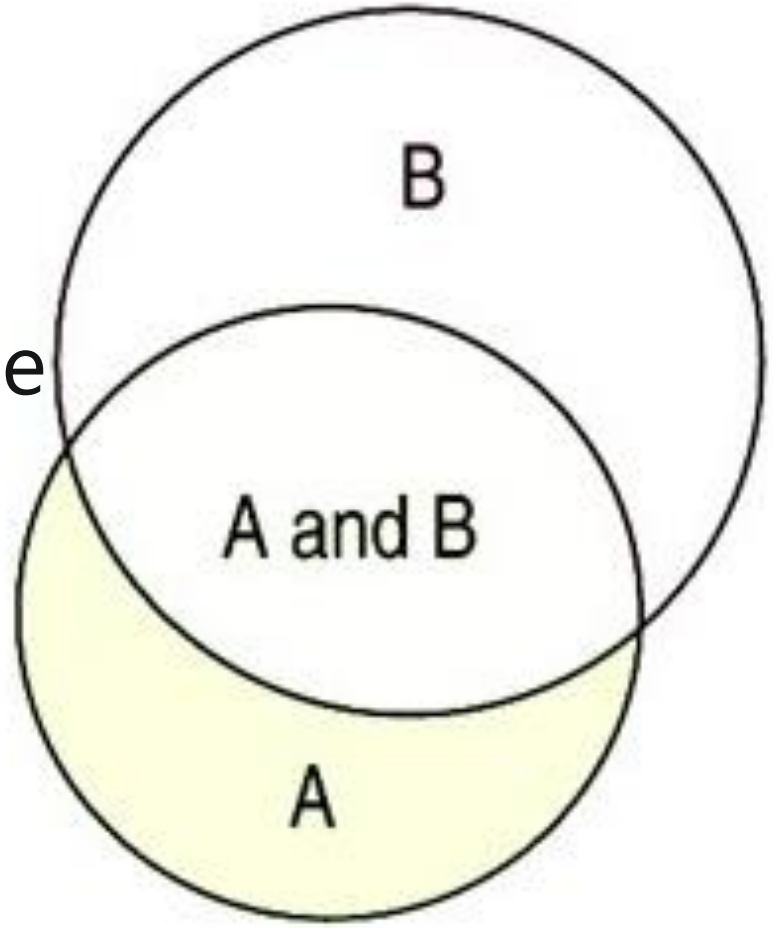
- **P(A or B) = P(A) + P(B)**
  - Disjoint (Mutually Exclusive)



Union $P(A \cup B)$

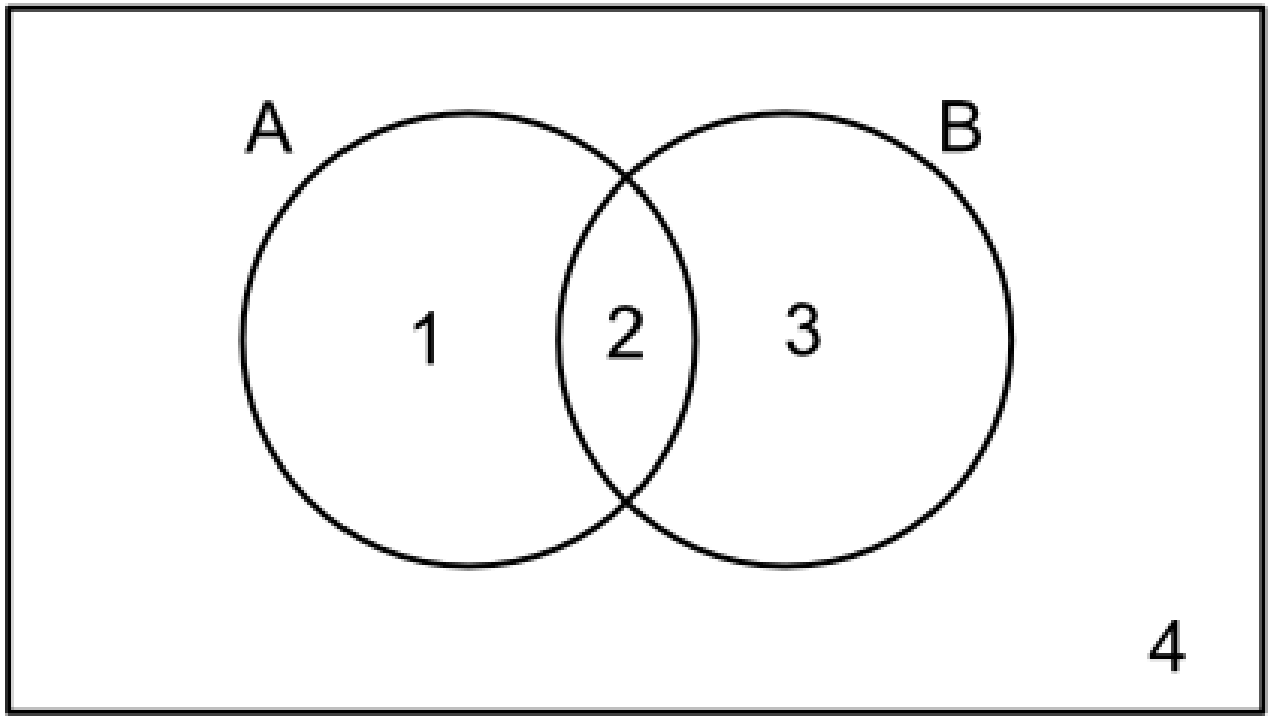
- **P(A or B) = P(A) + P(B) - P(**$A \cap B$ **)**
  - Non-disjoint events will double count the intersection, so we must remove the overlap

  –General Addition Rule



Conditional Probability

- **P(B|A) = P(A and B) / P(A)**
- **P(A) cannot = 0**



P (A | B) is A *given* B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{2+3} = \frac{2}{5}$$

Intersection $P(A \cap B)$

- **P(A and B) = P(A) * P(B | A)**
- **P(A and B) = P(B) * P(A | B)**
  - Dependent Events
  - General Multiplication Rule

Sampling

Sample vs. Population

- **Almost impossible to examine the entire population**
- **Need a representative sample**
  - **Proportionate to population**
- **Bias- samples that over or under emphasize characteristics**
  - **Height in Netherlands**

	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.

Sampling Methods

- **Random Sampling-**
  - **Protects us from influences of all features of our data**
  - **Make sure sample is representative**
  - **Sampling Variability- each random sample drawn is different**
- **Stratified Sampling-**
  - **Slice the data into homogenous groups called strata**
    - **Example male vs female, or age groups**
  - **Sample randomly from each strata**
  - **Reduces sampling variability**
- **Cluster**
  - **Splitting data into representative clusters and then randomly sample a few clusters**
  - **Clusters are not homogenous like strata, but representative of population**
- **Multistage = using many methods**
- **Systematic-**
  - Randomly order your sample, then select every nth person

Large Samples and Law of Large Numbers

- When you run many trials, the theoretical probability will get closer and closer to the actual probability

- **Relative frequencies only even out in the long run.. AKA infinity**
- **Sequences of random events don't need to balance out**

– **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.**

•There is no such thing as the Law of Averages

- **"It's bound to happen again"**
- **Nope, it's not**

Standard Deviation vs. Standard Error

Standard Deviation: Measures dispersion from the mean (aka variability of the data)

- **Measures spread and deviation of the data**
- **How different are data points from each other?**

Standard Error of mean: Measures how precise our estimate of the mean (or other measurement) is

- **Variability in estimator from sample to sample**
- **Used for confidence intervals**
- **Testing null hypothesis of differences between means**
- **SEM is always smaller than SD**

Sampling Lab

Hypothesis Testing

Steps in Hypothesis Testing

Identify a hypothesis that can be tested.

- "Increasing the size of our webstie logo will drive more customers to our sales section."

Select a criteria to evaluate the hypothesis.

- If our sample has a chance **>= 90%** that we'll see an increase in at least 10,000 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 and the resulting probability.

Hypothesis Testing

1. State problem as null hypothesis:

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

2. Decide on a significance level (probability cutoff) that would allow you to reject the null hypothesis and accept the alternative hyptothesis.

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

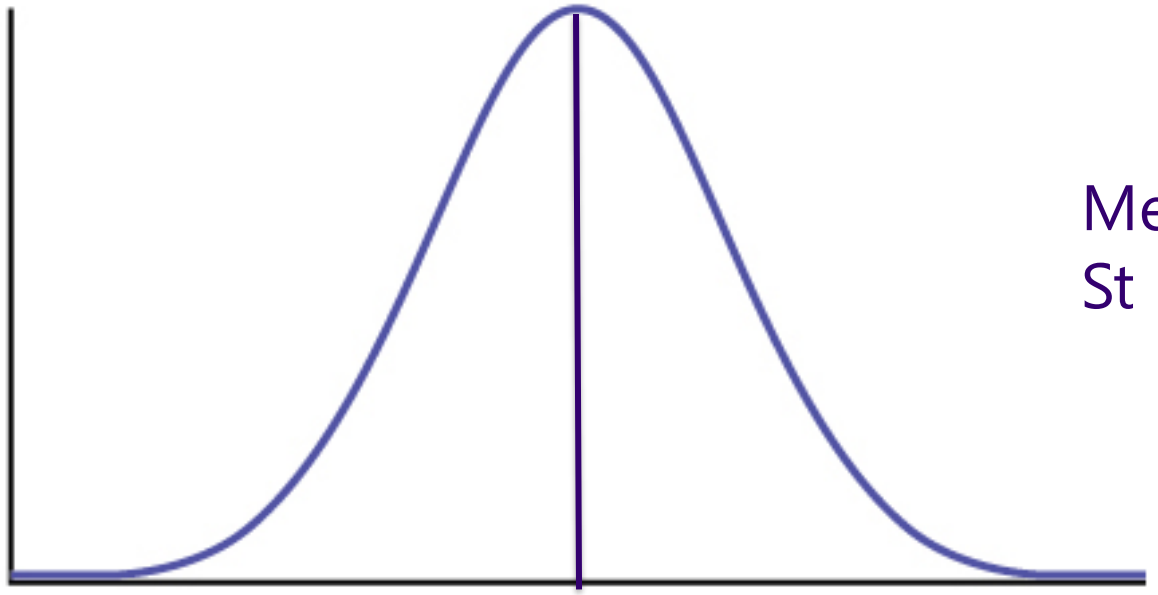
–0.9, 0.95, and 0.99 are common (problem specific)

Hypthoesis Lab: Up to your turn 1

Hypothesis Testing

We know that the average time a user spends on a page has a mean of 15 seconds and a std dev. of 4 seconds.

If we assume normality, how do we test if a change to the page has a higher view time?



Mean = 15 seconds
St Dev = 4 s

Hypothesis Testing- Define Ho and Ha

We know that the average time a user spends on a page has a mean of 15 seconds and a std dev. 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 new.

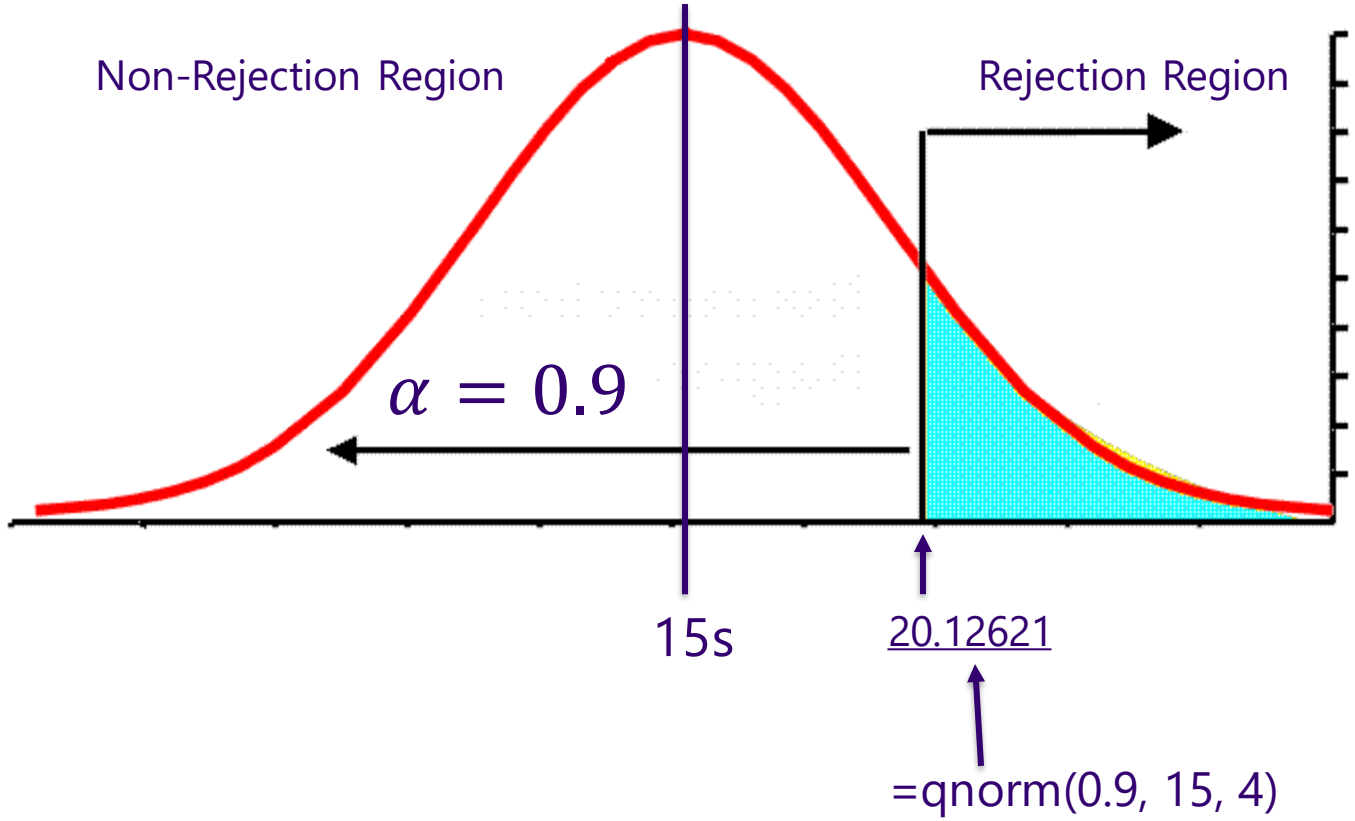
H_0: The new website has the same or less viewership than the original.

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

Hypothesis Testing- Select Confidence Value

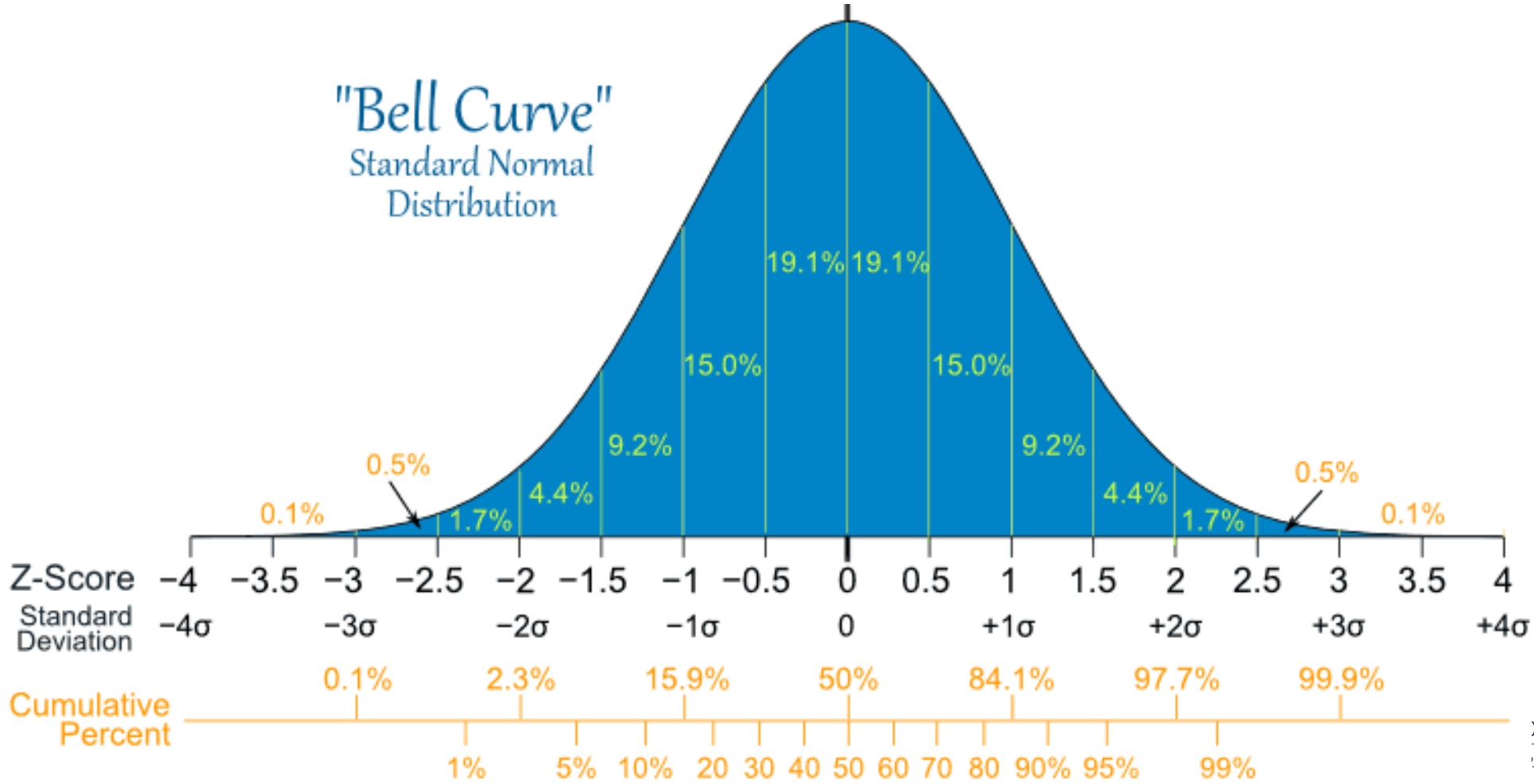
We now select a confidence value, alpha = significance level

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.

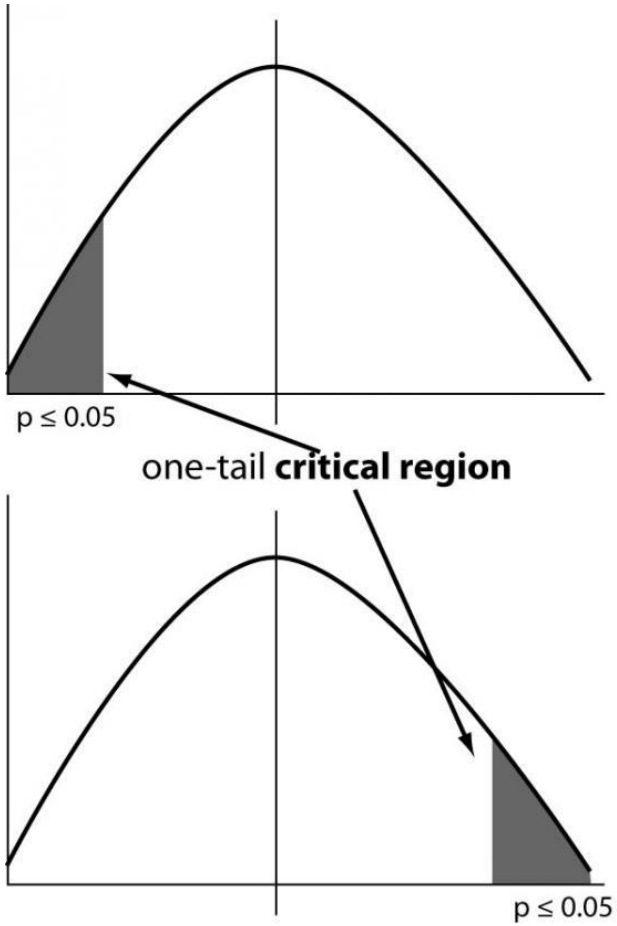


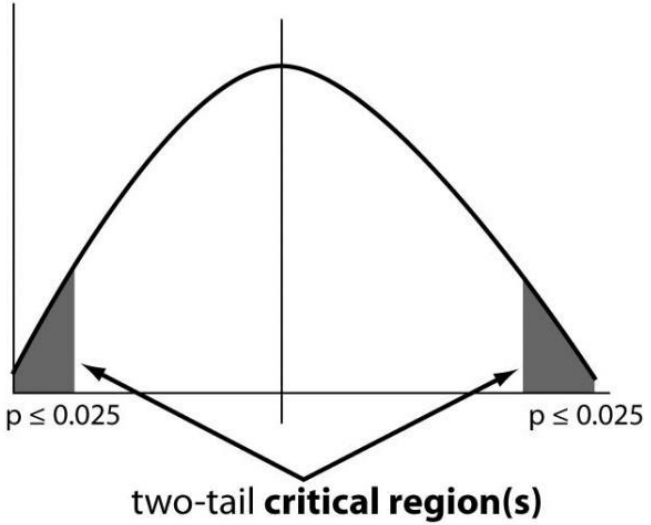
Hypothesis Testing

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.

Use two-sided unless you have a specific reason not to.





The accept-rejection criteria for the null hypothesis is different in each case.

- One-tail test with value >> the cutoff
- One-tail test with value << the cutoff
- Two-tail test with value < −cutoff/2 or > cutoff/2

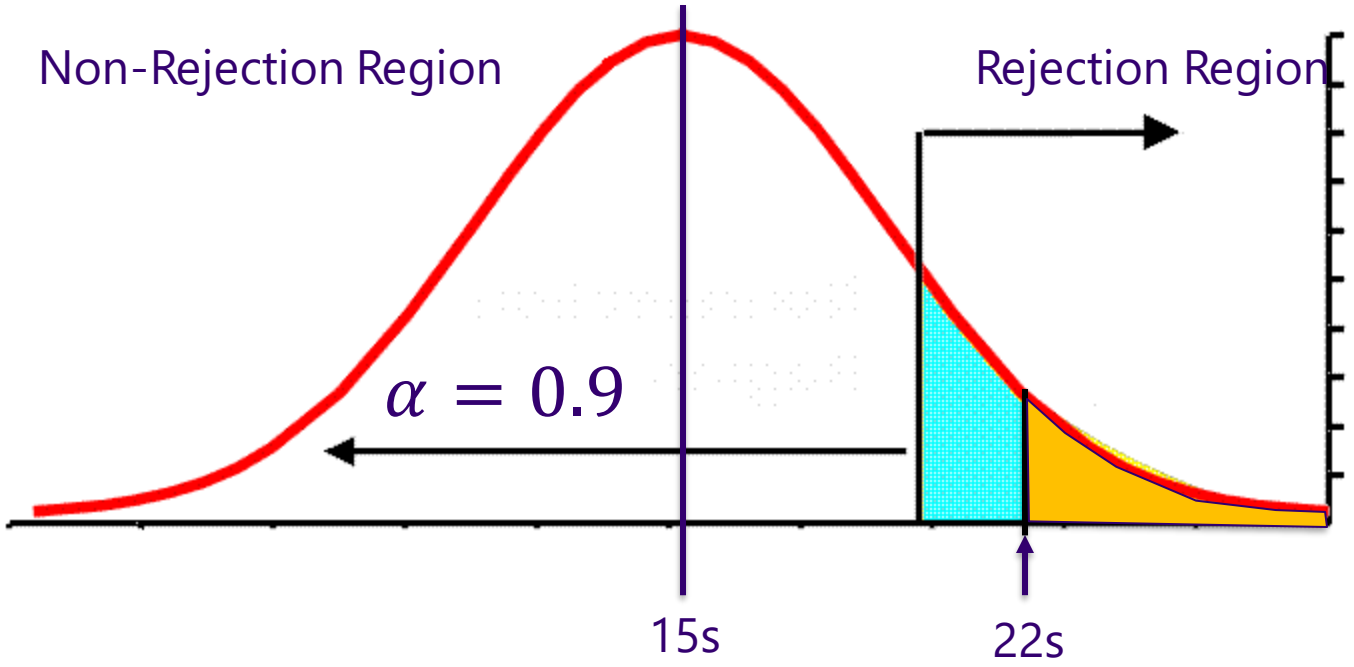
Hypothesis Testing- P values

The p-value is the probability of obtaining the sample results or worse, given the null hypothesis is true.

If the p-value is less than your chosen significance level, alpha, you reject the null.

Ex. P < 0.001 (less than one in a thousand chance of being wrong)

What is the p-value of a sample mean of 22 seconds?



Calculating PValues

Normal Distribution = One Proportion Ztest

- **In a normal distribution, we know the mean and standard deviation**
- **From the SD, we can find the Zscore**
- **From the Zscore, we can find the P value**

- **But what if we don't know the SD?**

T Distribution

Discovered by Guiness QA Engineer, William Sealey Gosset

- **Central Limit Theorem**
  - All we need to model sampling dist of $\hat{y}$ (sample mean) is a random sample and the true population std dev..
  - Avoid this by using Standard Error, an estimate of the pop std dev
  - Gosset was sampling beers and realized that Standard Error only works well on large samples, due to CLT
  - Invented the Student's T

Student's T-test

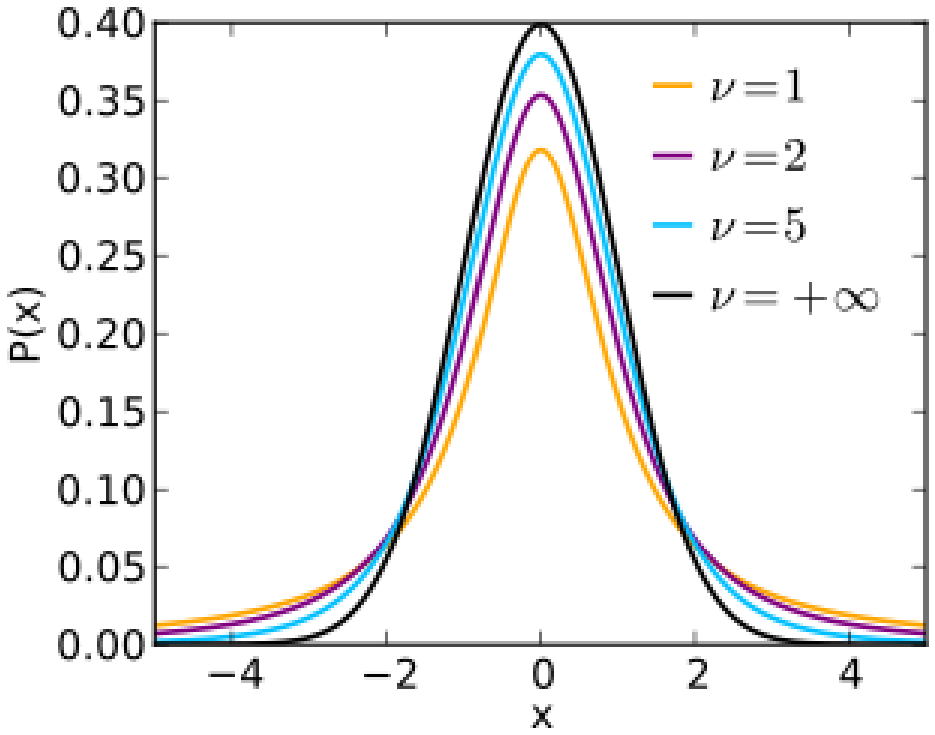
Are farmed salmon contaminated beyond the level permitted by the EPA?

- **Tests a hypothesis about the difference between two datasets**
  - **Test whether a population mean has a specified value.**
  - **Test the difference between two means with equal, unknown variances**
  - **Test a paired-response difference from zero.**
    - **E.g. a before/after drug treatment on patients.**
- **Use 'Welch's T-test' for testing the difference between two means (unknown variances, potentially different).**

Student's T

Degrees of Freedom

- Bell Shaped
- With n-1 degrees of freedom (df)
- Compensate for extra variance with
  - Wider confidence values
  - Larger P values



One Sample T-Test for the Mean

Compares difference between an observed statistic and a hypothesized value, to the standard error of the observed value.

- **Must have mu**
- **Ex. Are farmed salmon contaminated beyond the level permitted by the EPA?**
  - EPA screening value is 08ppm
  - H0: mu = 0.08
  - Ha: mu > 0.08

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

Where

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

μ = Proposed constant for the population mean

$\bar{x}$ = Sample mean

n = Sample size (i.e., number of observations)

s = Sample standard deviation

$s_{\bar{x}}$ = Estimated standard error of the mean (s/sqrt(n))

Hypthoesis Lab: Up to your turn 6

Chi-squared Test (Pearson's Correlation)

Is the distribution of birth months the same for major league baseball players as it is for the rest of the population?

Assumptions:

- **Data must be counts of categorical variables**
- **Counts should be independent of each other**
- **Sample should be random**
- **Should have at least 5 individuals in each counted category**

Chi-squared Test (Pearson's Correlation)

- **Unpaired test for counts in different, mutually exclusive, categories.**
- **Tests whether the different categories differ in some specific value**
  - **Are the differences between observed and expected counts significant?**
- **In order to do this test, we specify the 'degrees of freedom' in the Chi-squared test.**
  - This is equal to n-1. Where n equals the number of different categories, not sample size
- **The test looks at the sum of the difference between observed data and the counts given by the null model**
  - **Squared residuals, hence chi-squared**
  - **Gives us positive values**

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Squared Table

Table 5.3. Chi-square value

Degrees of Freedom	Probability								Non-significant	Signi-ficant	Highly significant
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	

Chi-squared Test (Pearson's)

RECAP: We're looking at data on the birth months of major league baseball players. We've checked the assumptions and conditions for performing a χ^2 test.

QUESTIONS: What are the hypotheses, and what does the test show?

H_0: The distribution of birth months for major league ballplayers is the same as that for the general population.

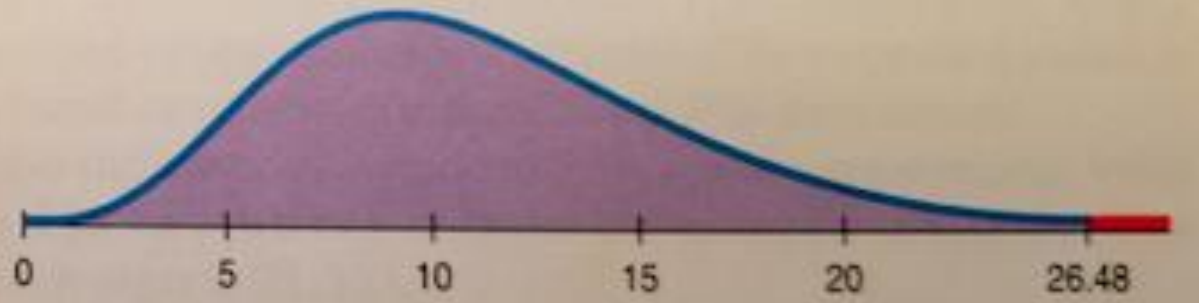
H_A: The distribution of birth months for major league ballplayers differs from that of the rest of the population.

$$df = 12 - 1 = 11$$

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

$$= \frac{(137 - 118.24)^2}{118.24} + \frac{(121 - 103.46)^2}{103.46} + \cdots$$

$$= 26.48 \text{ (by technology)}$$



$$\text{P-value} = P(\chi^2_{11} \geq 26.48) = 0.0055 \text{ (by technology)}$$

Because of the small P-value, I reject H_0: there's evidence that birth months of major league ballplayers have a different distribution from the rest of us.

Hypthoesis Lab: Up to your turn 7

Fisher's Exact Test

- **Similar to Chi-Squared, but for small sample sizes**
- **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:

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

Fisher's Exact Test

Example: Nood et al. (2013) studied patients with *Clostridium difficile* infections, which cause persistent diarrhea.

One nominal variable was the treatment: some patients were given the antibiotic vancomycin, and some patients were given a fecal transplant. The oth nominal variable was outcome: each patient was either cured or not cured.

The percentage of people who received one fecal transplant and were cured (13 out of 16, or 81%) is higher than the percentage of people who received vancomycin and were cured (4 out of 13, or 31%), which seems promising, but the sample sizes seem kind of small. Fisher's exact test will tell you whether this difference between 81 and 31% is statistically

	fecal	vancomycin
sick	3	9
cured	13	4

P of these exact numbers: 0.00772

RC Table:
R = rows
C = columns

Fisher's Exact

Next you calculate the probability of more extreme ways of distributing the 12 sick people:

To calculate the probability of 3, 2, 1, or 0 sick people in the fecal-transplant group, you add the four probabilities together to get *P*=0.00840

	fecal	vancomycin
sick	2	10
cured	14	3

P of these exact numbers: 0.000661

	fecal	vancomycin
sick	1	11
cured	15	2

P of these exact numbers: 0.0000240

	fecal	vancomycin
sick	0	12
cured	16	1

P of these exact numbers: 0.000000251

Fisher's Exact

Observed data

	1	2	⋯	k	
1	n_{11}	n_{12}	⋯	n_{1k}	n_{1+}
2	n_{21}	n_{22}	⋯	n_{2k}	n_{2+}
⋮	⋮	⋮	⋱	⋮	⋮
r	n_{r1}	n_{r2}	⋯	n_{rk}	n_{r+}
	n_{+1}	n_{+2}	⋯	n_{+k}	n

- Assume H_0 is true.
- Condition on the marginal counts
- Then Pr(table) $\propto 1 / \prod_{ij} n_{ij}!$

- Consider all possible tables with the observed marginal counts
- Calculate Pr(table) for each possible table.
- P-value = the sum of the probabilities for all tables having a probability equal to or smaller than that observed.

Fisher's Exact

	Men	Women	*Row Total*
Studying	***a***	***b***	*a + b*
Non-studying	***c***	***d***	*c + d*
Column Total	*a + c*	*b + d*	*a + b + c + d (=n)*

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;b!\;c!\;d!\;n!}$$

	Men	Women	*Row Total*
Studying	**0**	**10**	*10*
Non-studying	**12**	**2**	*14*
Column Total	*12*	*12*	*24*

For this table (with extremely unequal studying proportions) the probability is $p = \binom{10}{0}\binom{14}{12}/\binom{24}{12} \approx 0.000033652$.

Hypthoesis Lab: Finish Lab

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.

Outliers

Outlier Issues

- Identification
  - > Test whether or not an observation(s) is an outlier
- Accomodation
  - > Using robust statistical techniques that can deal with outliers. E.g. Using the median instead of the mean.
- Dealing/fixing
  - > Correcting data to not have outliers influence statistical conclusions.

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? TBD next class