| Numerical (quantitative) | | Categorical (qualitative) | |
|---|---|---|---|
| Continuous | Discrete | Regular Categorical | Ordinal |
| Take on any of an infinite number of values within a given range | Take on one of a specific set of numeric values | Categories that could be represented with numbers but difficult to do arithmetic operations on | Levels have inherent ordering |
| Height, Distance (Often rounded to look like a whole #) | Floors of a building, number of cars (can't have half a car) | Gender (0 for women, 1 for men) | Least Likely, Likely, Most Likely |

| Observational Study | Experimental Study |
|---|---|
| Collect data in a way that does not directly interfere with how the data arises. Only establish an association<br><br>Two types:<br>1. Retrospective: uses past data<br>2. Prospective: data collected throughout the study | Randomly assign subjects to treatments<br><br>Establish causal connections |

| Sampling and Bias | |
|---|---|
| General Terms | ● Census: entire population<br>● Exploratory analysis<br>● Representative sample<br>● Inference |
| Sources of Bias | 1. **Convenience sample -** individuals who are easily accessible are more likely to be included in the sample<br>2. **Non-response** - only a non-random fraction of the **random** sampled people respond to the survey (initially a random sample)<br>3. **Voluntary response** - sample consist of people who respond because of strong opinion (no initial random sample) |
| Sampling Method | **1. Simple Random Sampling (SRS) -** each case equally likely to succeed<br>**2. Stratified Sampling** - divided population into homogeneous strata, then randomly sample from each stratum<br>3. **Cluster Sample** - divided population into cluster, randomly |

| | |
|---|---|
| | sample a few clusters, then sample all observations within these few clusters<br>4. **Multistage Sampling -** divided population into cluster, randomly sample a few clusters, then random sample the observations within these few clusters |

| Experimental Design | |
|---|---|
| 4 Properties | 1. Control - control group<br>2. Random - equal probability<br>3. Replicate (collect a large enough sample or replicate the entire study)<br>4. Block - block variables known or suspected to affect the outcome |
| Explanatory vs Blocking Variables | Explanatory - conditions we can impose on the experimental units<br>Blocking Variables - characteristics the samples come with that we would like to control for (gender) |
| More Terms | ● Placebo/Placebo effect<br>● Blinding (sample don't know which group they're in)<br>● Double Blind (both sample and researcher don't know group assignment) |

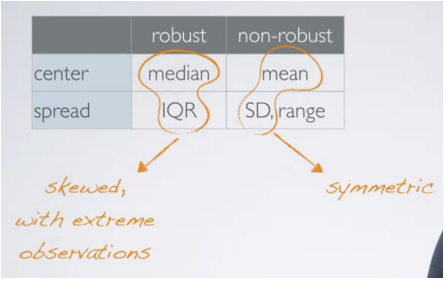
Random Sample Assignment

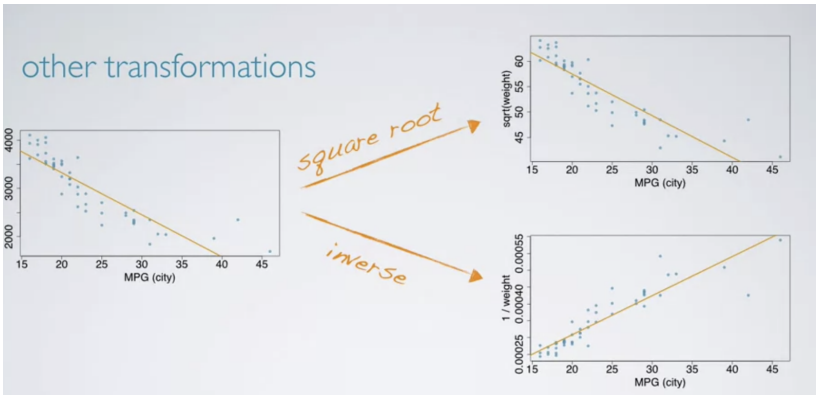
Causal - Effect on the experimental group is due to the treatment being applied
Generalizable - what's happening to the sample can be generalized for the entire population

| | Random Assignment | No Random Assignment |
|---|---|---|
| Random Sampling | Causal/Generalizable<br>(ideal experiment) | Not Causal/Generalizable<br>(most observational studies) |
| No Random Sampling | Causal/Not Generalizable<br>(most experiments) | Not Causal/Not Generalizable<br>(bad experiment) |

| Frequently used Visualizations | |
|---|---|
| Scatter Plots | Trends<br>Identify outliers to observe |
| Histograms | Data density<br>Shape of distribution:<br>- Skewness: Left skewed, right skewed, symmetric<br>- Modality: unimodal, bimodal, uniform, multimodal<br>-  |
| Dot Plot | When individual values are of interest, may become busy |
| Box Plot | Useful for highlighting outliers, media, IQR |
| Intensity Map | For highlighting spatial distribution (heat map) |

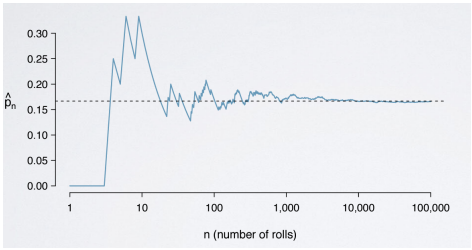
| Measures of Data | |
|---|---|
| Measures of Center | - Mean, median, mode are measures of distribution. Each have different symbols for measures taken for a sample(x) vs taken for a population (u)<br>- In terms of skewness, remember that the mean chases the tail. |
| Measures of Spread (variability) | - Range<br>- Variance (the average squared deviation from the mean). We square the variance to get rid of negatives so pos/neg doesn't cancel each other. Also so larger deviations are weighed more heavily than smaller numbers<br>- Standard Deviation: (square root of variance)<br>- Interquartile Range - 75th percentile - 25th percentile |

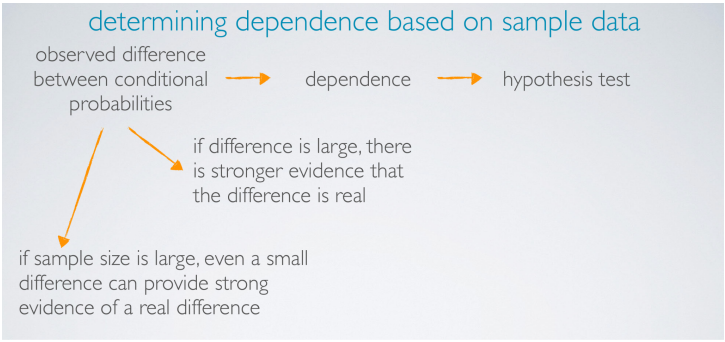
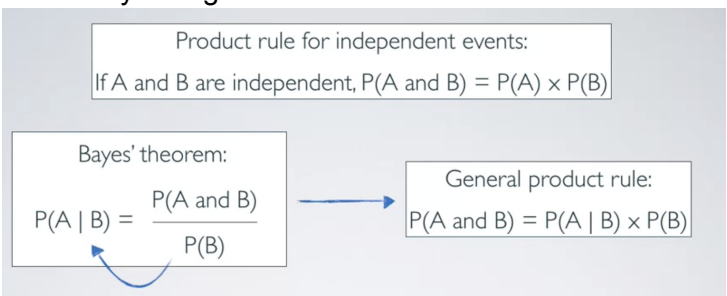
| | |
|---|---|
| Robust Statistics | - Measures on which extreme observations have little effect<br>  <br>- |

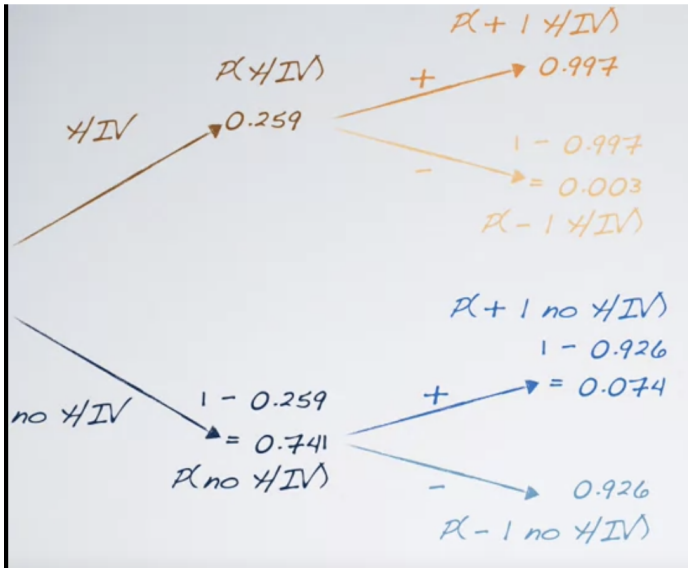
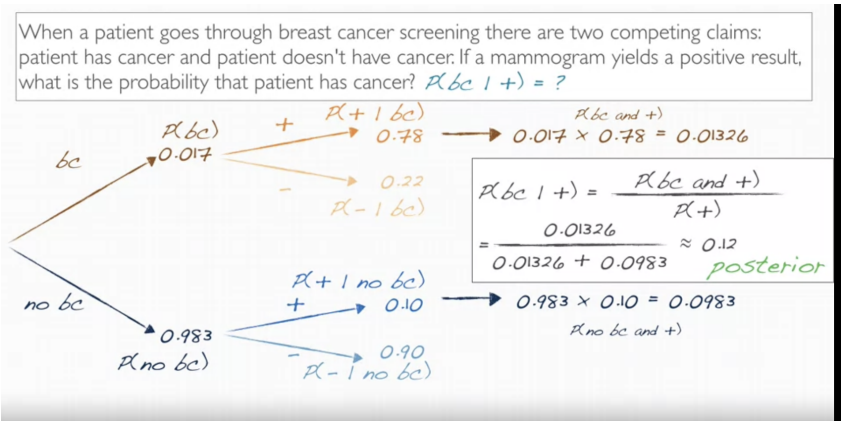
| | |
|---|---|
| Data Transformation (rescale the data, especially when it is strong skewed)<br>- See the data structure differently<br>- Reduce skew in modeling<br>- Straighten a non-linear relationship | |
| Natural Log Transformation | - Applied when much of the data cluster near zero relative to the larger values in the data set and all observations are positive<br>- To make relationship between variables more linear |
| Square root transformation<br>Inverse transformation |  |

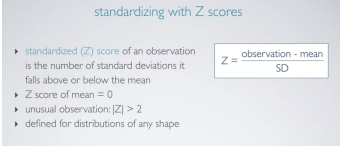
| Intro to Inference | |
|---|---|
| Null Hypothesis (H0) | There is nothing going on, no relation |
| Alternative Hypothesis (HA) | There is something going on, a relationship. Has the burden of proof to “fail to reject H)) |
| Process | - Set a null and alternative hypothesis.<br>- Simulate the experiment assuming that the null hypothesis is true<br>- Evaluate the probability of observing an outcome at |

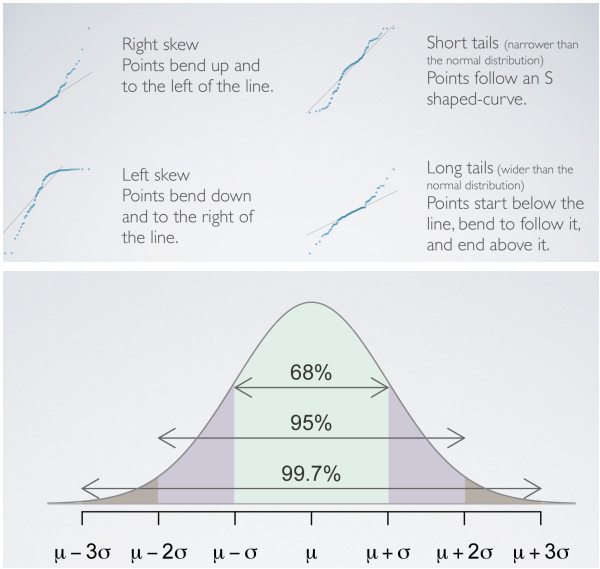
| | |
|---|---|
| | least as extreme as the one observed in the original data (p-value)<br>- If probability is low, reject the null hypothesis |

| Probability | |
|---|---|
| Frequentist Interpretation | - Probability of an outcome is the proportion of times the outcome would occur if we observe the random process an infinite number of times |
| Bayesian Interpretation | - Interprets probability as a subjective degree of belief<br>- When presented with new facts, the probability changes (posterior probability) |
| Law of Large Numbers | - As more observations are collected, the proportion of occurrences with a particular outcome converses to the probability of that outcome<br>- ex) if you toss a coin 1000 times, expect heads 500 times than only 5 times<br>-  |
| Gambler's Fallacy | - Common misunderstanding of law of large numbers<br>- When an individual erroneously believes that a certain random event is less/more likely to happen based on outcome of previous events<br>- ex) roll heads 3 times, higher chance of head a 4th time. |
| Disjoint (Mutually Exclusive) | - Cannot happen at the same time<br>- ex) cannot have both a head and a tail<br>- For disjoint events A and B, P(A or B) = P(A) + P(B) |
| Non-disjoint | - Events can happen at the same time<br>- For non-disjoint events A and B, P(A or B) = P(A) + P(B) - P(A and B) |
| Sample Space | - Collection of all possible outcomes of a trial |

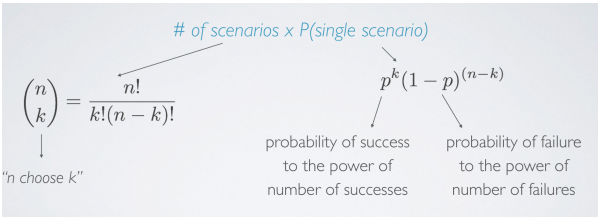
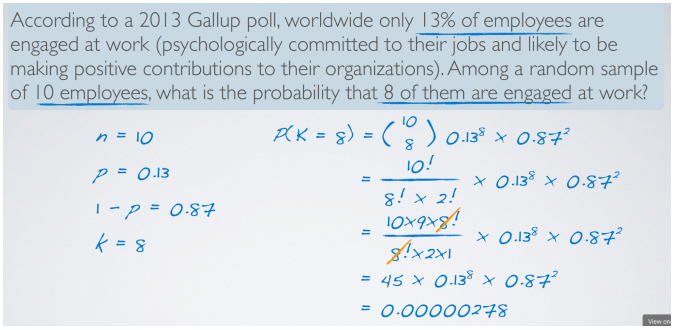
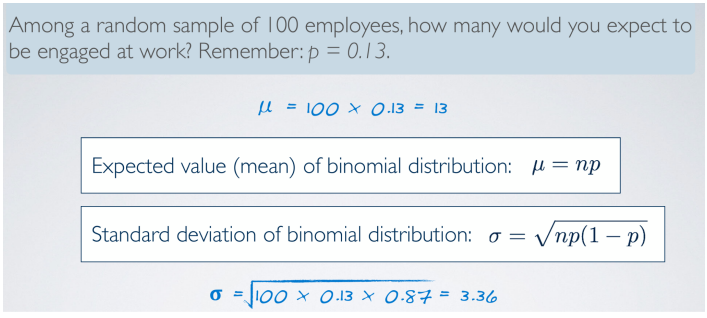
| | |
|---|---|
| Probability Distribution | - Lists all possible outcomes in the sample space and the probabilities with which they occur<br>- Rules:<br>1. Events listed must be disjoint<br>2. Each probability must be between 0 and 1<br>3. The probabilities must total to 1 |
| Complementary Events | - Two mutually exclusive events whose probabilities add up to 1<br>  <br>-<br>- Disjoint vs Complementary<br>- Disjoint are not necessarily complementary but complementary are always disjoint |
| Independence |  <br>  |
| Baye's Theorem (Conditional Probability) | Probability of A given B<br>  |

| | |
|---|---|
| Probability Trees |  |
| Bayesian Inference | Steps:<br>1. Setting a prior<br>2. Collecting Data<br>3. Obtaining a Posterior<br>4. Updating the prior with the previous posterior<br>  |

| Normal Distribution | |
|---|---|
| z -score (only under normal curves) | - How many standard deviations above/below the mean<br>  <br>- |

| | |
|---|---|
| percentiles | - Area below the probability distribution curve<br>Three ways to find it.<br>1. Can be used with z-score if normally distributed. Find the z-score and look up on a table. otherwise need calculus<br>2. Can be calculated with R<br>R<br>`> pnorm(-1, mean = 0, sd = 1)`<br>`[1] 0.1586553`<br>3. Using a web applet |
| Graphs | Percentiles are symmetric and unimodal<br>Scatter plots will be a straight line<br>  <br> |

| Binomial Distribution | |
|---|---|
| Bernoulli Random Variable | When an individual trial has only two possible outcomes.<br><br>Ex) Milgam's experiment where success is if a person doesn't administer a shock and failure if she does administer a shock |
| Binomial Distribution | The probability of having exactly k successes in n independent Bernouilli trials with probability of success p.<br><br>Number of combinations * probability of success and failure |

| | |
|---|---|
| |  <br>Note that the combination formual is the "choose(,)" function in r<br><br>Note that binomial calculation is "dbinom(k,size=sample size, p=%success)" |
| Binomial Conditions | 1. Trials must be independent<br>2. Number of trials (n) must be fixed<br>3. Each trial outcome must be classified as a success or failure<br>4. The probability of success, p, must be the same for each trial |
| Example Problem |  |
| Expected Value and Standard Deviation of BinomD |  |
| Binomial vs Normal Distribution | If a sample size is large enough, then the binomial distribution will closely resemble a normal distribution |

| | |
|---|---|
| | Success-failure rule: A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution.<br>$np \geq 10$<br>$n(1-p) \geq 10$<br><br>Normal approximation to the binomial: If the success-failure condition holds,<br>Binomial(n,p) ~ Normal($\mu$,$\sigma$)<br>where $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ |