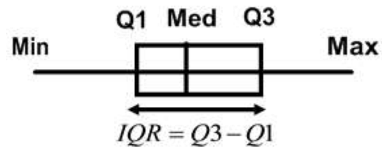
Unit 1: One-variable data **AP Statistics Course Review Study Guide**

***Statistics Facts*: Descriptive Statistics**

Describing distributions: **SOCS = S**hape, **O**utliers/unusual, **C**enter, **S**pread (use comparison language if comparing)

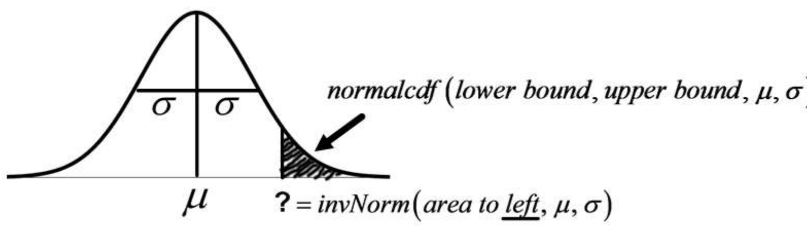
Outliers: Data point is an outlier if it is $< Q1 - 1.5IQR$ or $> Q3 + 1.5IQR$
For data that follows $N(\mu, \sigma)$ an outlier is a point more (or less) than $\mu \pm 2\sigma$



Standardizing Normal data: If $N(\mu, \sigma)$ we can standardize to $N(0,1)$

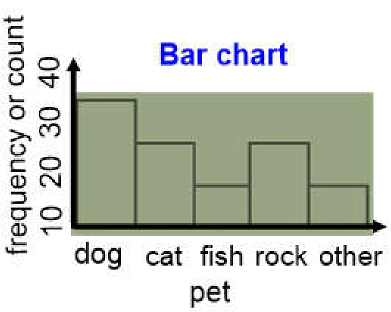
by finding z-score:

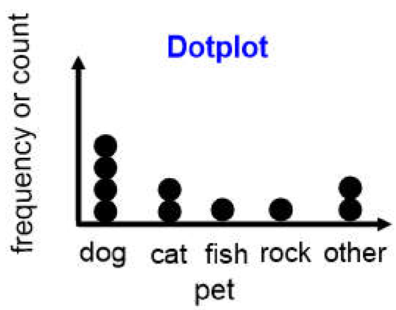
$$z = \frac{x - \mu}{\sigma}$$

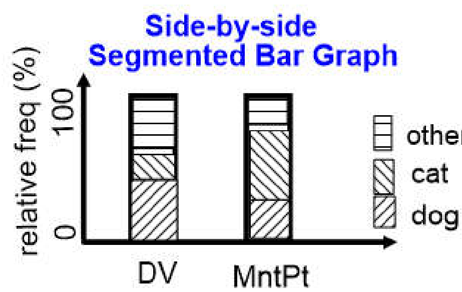


Standard deviation: Measures the average distance between individual data values and their mean.

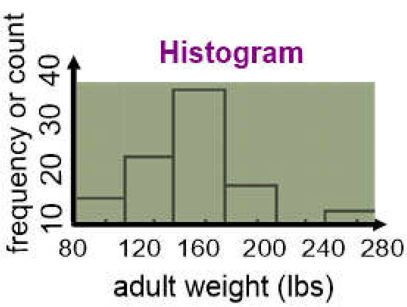
Categorical Data

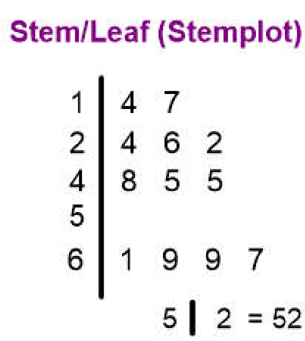


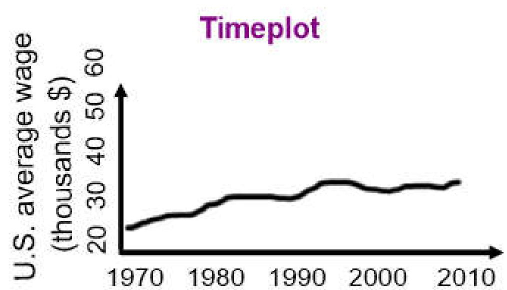




Numerical Data







Unit 2: Two-variable (x-y) data

Regression:

Least-Squares Regression Line (LSRL): $\boxed{\hat{y} = a + bx}$ or $\boxed{response\ variable = a + b(explanatory\ variable)}$

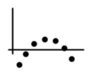
r: correlation coefficient (no units) $-1 \le r \le 1$

$b = r\frac{s_y}{s_x}$ *(given on AP formula sheet)*

r^2: coefficient of determination (fraction or percent of variation in y that is explained by the LSRL)

If a line is a good fit to data, then residuals are in a random pattern (no pattern).

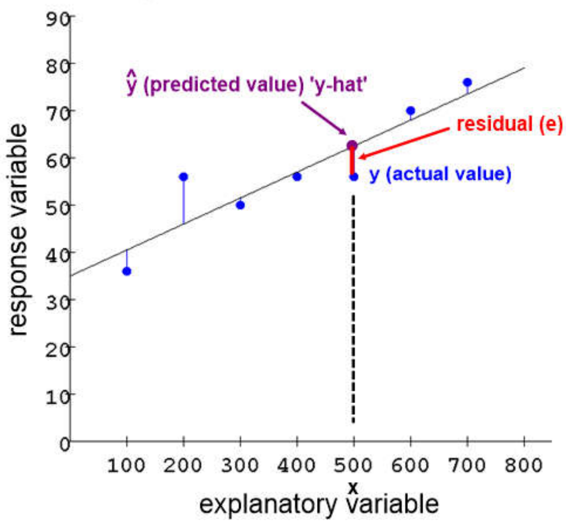
$\boxed{residual = y_{observed} - y_{predicted}}$

If residuals display a pattern  then data is not linear and follows:

$\boxed{Note: correlation \nRightarrow causation}$

Linear Regression

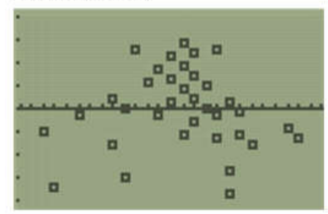
Associations which are approximately linear on a scatterplot can be modelled with a line called the **Least Squares Regression Line (LSRL).** (Also known less accurately as 'linear model', 'line of best fit', or 'trend line').



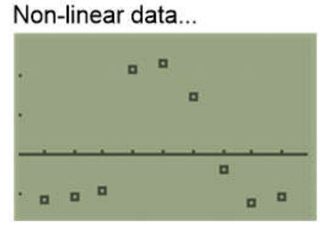
A **residual** is the error between what the LSRL line predicts the y value will be (for a given x) and the actual y value.

$$e = y - \hat{y}$$

Linear data...



Non-linear data...



Residual Plots

-

Straightening non-linear data:

Exponential model $(y = a^x)$

$y = x^a$

$\boxed{\log y} = a\boxed{\log x}$ ← We straighten data by taking logs

$\log\ of\ \ x\ and\ \ y$

-or-

$y = a^x$

Power model $(y = x^a)$ $\boxed{\log y} = \boxed{x}\log a$

$\log\ of\ \ y\ only$

The 3 ways to find the equation of an LSRL...

1) <u>If you have the full data set</u>:
Use calculator. Enter data in L1, L2 and run LinReg.

2) <u>If you have the output of a software analysis</u>:
Use the values in the table:

- Look for the word 'coefficient' or 'estimate':
- One row is for the y-intercept, labelled 'constant' or 'intercept'. (The coefficient of this row is the y-intercept, 'a')
- One row is for the x term, labelled the name of the x-variable. (The coefficient of this row is the slope, 'b')

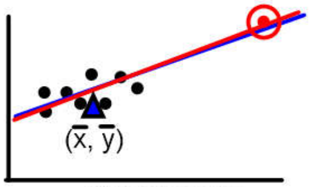
3) <u>If you have no data or software output, but summary data on x and y</u>:

- Use the formula $b = r\frac{s_y}{s_x}$ to find the slope, b.
- Solve for y-intercept, a, by plugging in the centroid $(\bar{x}, \bar{y})$ (the only point that is always on the LSRL) and solving for a.

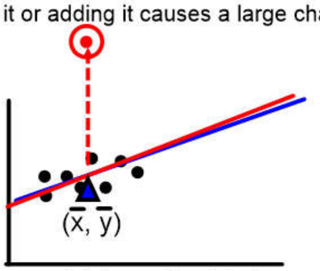
<u>Outlier effect on slope - the concept of 'leverage'</u>

The point $(\bar{x}, \bar{y})$ is always on the LSRL and you can think of it like a 'fulcrum' of a lever.
A data point whose x value is the same as the fulcrum will have no effect on the LSRL, but a line whose x value is far away from the fulcrum has a large effect and we say this is a **high leverage point**.

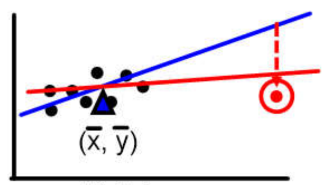
But for the point to cause a change in slope of the LSRL, it needs to have a high residual. A point already near the LSRL won't 'push' the LSRL and cause much change. If a point has high leverage and high residual (so that removing it or adding it causes a large change in the LSRL slope) we say that point is **influential.**



high leverage
small residual, not influential
(on slope)



high residual but
low leverage, not influential
(on slope)

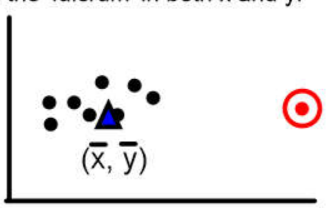


high leverage
large residual, influential
(on slope)

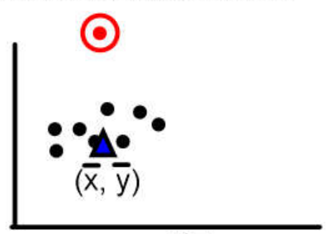
<u>Outlier effect on correlation</u>

- Correlation is the measure of the strength of the association and is high (close to + 1 or -1) if the points are grouped tightly around the LSRL.
- Correlation is calculated as sum of products of standardized distances x and y from the mean, so a point has a large effect on correlation if it is far from the 'fulcrum' in both x and y.

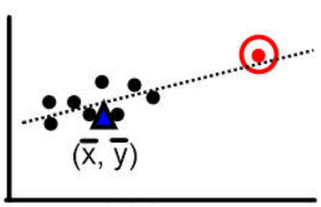
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}$$



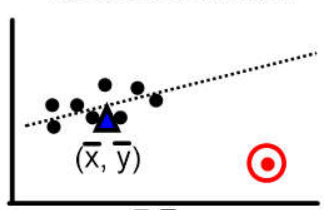
close to $(\bar{x}, \bar{y})$ in y
little effect on correlation



close to $(\bar{x}, \bar{y})$ in x
little effect on correlation



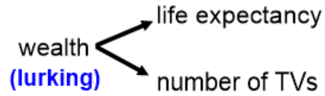
far from $(\bar{x}, \bar{y})$ in both x and y
strongly affects correlation
including this point <u>strengthens</u> correlation



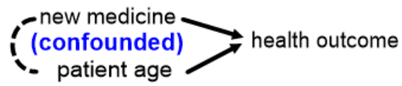
far from $(\bar{x}, \bar{y})$ in both x and y
strongly affects correlation
including this point <u>weakens</u> correlation

Terminology

Lurking variable: When one variable causes two other variables to change together, making them appear associated. (Used by our textbook, not an official term)



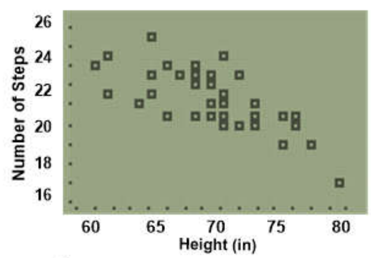
Confounded variables: When the effect of multiple explanatory variables on a response variable can't be separated. (Official term, used on AP Statistics Exam)



Association: General term meaning there appears to be some relationship between variables.
(Official term, used on AP Statistics Exam)

Correlation: Precise term describing the strength and direction of a linear relationship (usually taken to mean the correlation coefficient, r).
(Official term, used on AP Statistics Exam)

Standard explanation wordings:



$$\hat{y} = (53.8471) - (0.5728)x$$

$$r = -0.873 \qquad r^2 = 0.763$$

$$s = 1.58$$

slope, b:

"For every 1 added inch in height, the number of steps decreases by 0.5728 steps, on average."

intercept, a:

"A person who is zero inches tall is predicted to take 53.8471 steps, on average."

association / explaining r (correlation coefficient):

"There is a linear, negative, fairly strong association between number of steps and height".

r^2 (coefficient of determination):

"About 76% of the variation in number of steps is explained by the LSRL which relates number of steps to height."

s (standard deviation of the residuals):

"The actual number of steps (for a given height) are 1.58 steps away from the predicted number of steps, on average." or "The average error between actual and predicted number of steps for a given height is 1.58 steps."

Unit 3: Experiments, Studies, Biases

Observational Study: No treatment is imposed.
Experimental Study: A treatment is imposed.
No cause/effect relationship can be concluded from an observational study. Why not? Correlation does not imply association due to possible lurking variables.

Sampling: Selecting a portion of a population for analysis.
Simple Random Sample (SRS): Each *set* of n individuals has an equal chance of being selected.
Best way to obtain: Draw names from a hat.
Random Sample: Each *individual* has an equal chance of being selected.
Stratified Random Sample: Population is divided into groups or strata, take SRSs from each stratum (Stratified is used when you expect some difference between strata and want to include some from each).
Cluster Sample: Population is divided into non-homogeneous groups called clusters. SRS take from some of the clusters (usually clusters separated by geographic location). (Cluster is used when you don't expect difference between clusters, but want to subdivide for convenience.)
Systematic Sample: Employing an algorithm for selecting e.g. choose every 5^{th} individual from a list.
Convenience Sample: A potentially biased sample which was taken in some way 'convenient' to the researchers (e.g. everyone who leave a particular building, researchers ask everyone in their neighborhood).

Bias: Anything which causes a sample to be not representative of the population from which it is sampled.
Voluntary response: Asking for volunteers instead of selecting the participants.
Non-response: Researchers choose the participants, but they may choose not to participate.
Response bias: Anything in the survey design or procedures which might induce a particular response (attractive interviewer, boss will find out answers, wording of survey questions)
Undercoverage bias: Anything which results in some portion of the population not being included in the right proportion (landline phones for survey, surveying only North part of city)

Some experiment terms...

Subject, participant, experimental unit: One individual object or person to which treatment is applied and response data is measured.

Group: A collection of experimental units.

Factor: An explanatory variable whose levels are controlled.

Treatment: Applying different levels of the factor to a group.

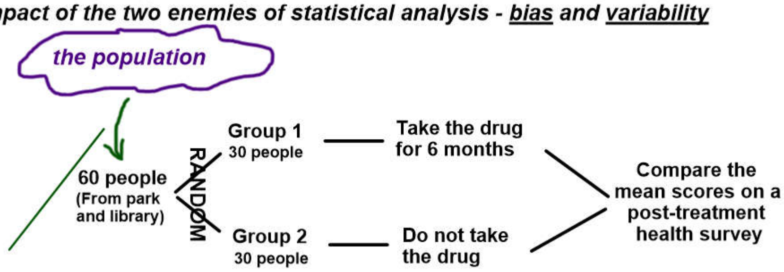
Response variable: The variable which is measured to determine the effect of the treatment.

Only a well-designed experiment can conclude cause-and-effect

For a study to be called an experiment, technically, only one thing is required:

- **Researchers apply a treatment to multiple groups.**

But a "well-designed" experiment takes further measures to reduce the impact of the two enemies of statistical analysis - bias and variability



Controlling bias: Use an appropriate sampling technique to select the subjects from the population under study. This allows the conclusion to be applied as broadly as possible.

To reduce variability...

1) [Random assignment of subjects to more than one group].

Random assignment to groups controls for (removes the variability of) differences between the subjects (known and unknown).

Note: If one group receives no treatment sometimes this is called a 'control group' but this is *not* required, you just must have any two or more groups.

2) Control of the factors.

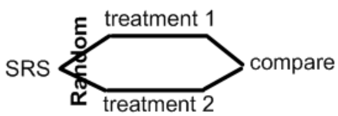
At least one factor must be under control and *imposed* as a treatment by the experimenters on the subjects.

3) Replication. Two different meanings of replication, both important:

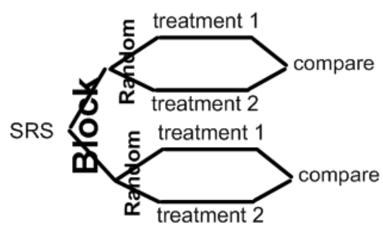
- ***Replication of treatment*:** Randomization is how we control for differences in the subjects we don't know about. **There must be *enough subjects in each group*** for the 'averaging out' to work.
- ***Replication of experiment*:** Because experiments can sometimes randomly have unusual results, **the *entire experiment should be replicated*,** preferably by different researchers.

Experiment Designs: (Single blind, double blind – placebos aide in blinding)

Completely randomized



Blocked randomized



Matched pair

Compare differences before and after treatment or pre vs. post test

Note: We block on differences we know about, and we randomize to take care of differences we don't know about.

Blinding and Placebos

To eliminate issues of subject and/or researchers biases affecting results, humans can be prevented from knowing which experimental units are assigned to which groups. This is called **blinding.**

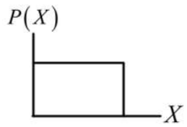
Single-blind: When one class (subject or researcher) is blinded.
Example: Researcher knows which cola is which, but brand is hidden from subject.

Double-blind: When everyone in both classes (subject and researcher) are blinded.
Example: A 3rd party prepares the cola samples so both the researcher and subject do not know the brands. Codes are used and only revealed after the results of the experiment are final.

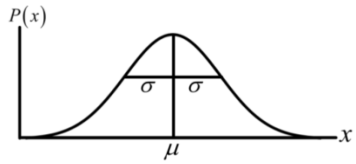
Placebo: Sometimes, subjects are subconciously expecting a particular result and if they can detect that they are in the control group, that can bias the results. So the subject can be given a 'fake treatment' which replicates the experience of receiving the treatment without actually doing anything. This fake treatment is called a **placebo**.

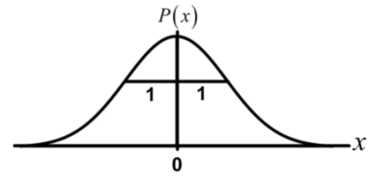
Unit 4: Probability and Data Analysis

Distributions (Continuous variables)

1) <u>Uniform</u>: 

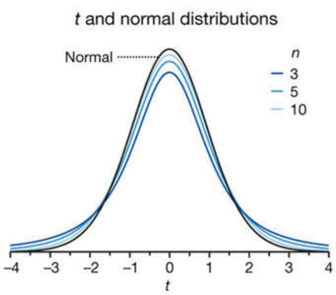
All events equally likely

2) <u>Normal</u>: 

standardized: 

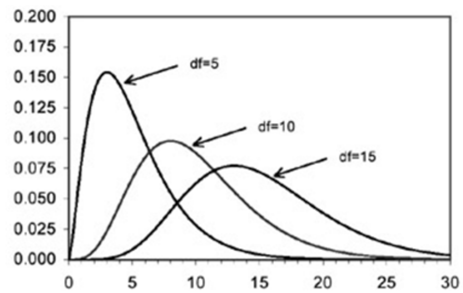
$$z = \frac{x - \mu}{\sigma}$$

3) <u>Student t-distribution</u>:



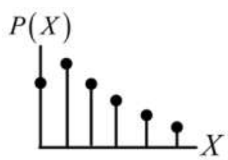
- More area in tails than Normal.
- Depends upon degrees of freedom
- As n (df) increases, approaches a Normal distribution.

4) <u>Chi-squared</u> χ^2:



- Depends upon df.
- Skewed right.
- Mode is at df-2.
- Median is at df.
- Lower df = higher peak.

Distributions (Discrete variables)

5) <u>Binomial</u>: 

X	0	1	2	3	4	5
$P(X)$	.3	.4	.2	.05	.01	.001

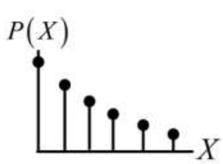
Binomial Setting:
1) 2 outcomes
2) Probability of success does not change
3) Trials are independent
] Bernoulli Trials
4) Fixed number of trials

(skew depends upon p, if p=0.5 symmetric)

$B(\mu,\sigma)$
where $\mu = np$
$\sigma = \sqrt{npq} \quad (q = 1 - p)$

$$P(x) = {}_nC_k\,(p)^k\,(q)^{n-k} = binompdf(n, p, k) \quad \text{'exactly' } k \text{ successes}$$
$$P(x) \quad (cumulative) \quad = binomcdf(n, p, k) \quad \text{'at most' } k \text{ successes}$$

If $np \geq 10$ *Binomial distribution can be approximately by Normal distribution* $N(\mu,\sigma)$ *where* $\mu = np, \sigma = \sqrt{npq}$

6) <u>Geometric</u>: 

X	1	2	3	4	5 ...
$P(X)$	.2	.16	.13	.10	.08

Geometric Setting:
1) 2 outcomes
2) Probability of success does not change
3) Trials are independent
] Bernoulli Trials
4) Non-Fixed number of trials

$G(\mu,\sigma)$
where $\mu = \frac{1}{p}$

$$P(x) = (q)^{k-1}(p) = geometpdf(p, k) \quad \text{success 'on exactly' } k\text{th trial}$$
$$P(x)\,(cumulative) = geometcdf(p, k) \quad \text{success 'on or before' } k\text{th trial}$$

1: Definitions

Law of Large Numbers

For a small number of trials, *anything* can happen.
As number of trials increases, the ***experimental probability*** approaches the ***theoretical probability.***

Definitions

- **Trial**: One complete 'occurrence' of a situation .
- **Outcome**: One possible result that can occur when a trial is conducted.
- **Sample space**: Set of all possible outcomes that can occur in a trial.
- **Event**: Any subset of the sample space (the "desired" outcomes).
- **Union**: $A \cup B$ (A ***OR*** B)
- **Intersection**: $A \cap B$ (A ***AND*** B)
- **Parameter**: Number describing a *population (or model)*, e.g. μ, σ
- **Statistic**: Number describing a *sample*, e.g. $\bar{x}, s$

2: Equally-likely outcomes

If all outcomes are equally likely:

$$P(E) = \frac{\text{number of outcomes in event } E}{\text{number of outcomes in the sample space } S}$$

$$= \frac{\text{number of 'desired' outcomes}}{\text{total number of outcomes}}$$

2: Counting Strategies

Strategy	When to Use
1. List out all the cases and just count them up. (can also use tree diagrams, grids for 2 dice, methodical listing system to help)	1. Best strategy, but only good for small numbers.
2. Multiplication Principle (one box per choice, fill in with number of ways to make that choice, multiply)	2. Multiple choice to make, every possible choice in each box can be paired with every other choice.
3. Permutations (use calculator) (special case: 'choose all' = n! ways)	3. A set of distinct objects (no repeats), choosing some or all, and objects are 'used up' as you choose them, and order matters.
4. Combinations (use calculator)	4. A set of distinct objects (no repeats), choosing some or all, and objects are 'used up' as you choose them, and order does not matter.
5. Multiplication Principle w/Combinations	5. Multiple choices to make, but each is a choice of a number of items out of a set.
6. Distinguishable permutations $\#\text{ distinguishable permutations} = \frac{n!}{n_1!n_2!n_3!\ldots}$	6. Number of ways to arrange all items in a set if there are repeats.

3: Conditional Probability

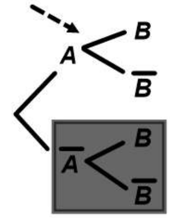
$$P(\text{video games} \mid \text{girl}) = \frac{12}{60} = .20$$

event: The event is always contained within the conditional sample space.

condition: The condition is always just a portion of the sample space (the *conditional sample space*).

The **conditional sample space** is a portion of the **sample space**.
The **event** is a portion of the **conditional sample space**.

	girls	boys	
read a book	18	4	22
video games	12	20	32
watch Netflix	30	16	46
	60	40	100



3: Conditional Probability

The event goes in the numerator of the fraction.

$$P(\text{video games} \mid \text{girl}) = \frac{12}{60} = .20$$

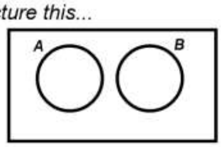
The condition goes in the denominator of the fraction.

4: Compound Events (OR)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Special Case: If two events have no overlap, they are called 'mutually exclusive' or 'disjoint' events.

Picture this...

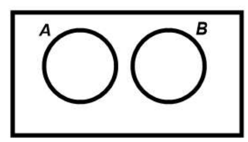


So the OR formula is simplified...

$$P(A \cup B) = P(A) + P(B) - \cancel{P(A \cap B)}$$

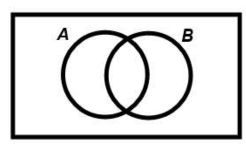
$$\left(P(A \cap B) = 0\right)$$

4: Disjoint Events



A and B are mutually-exclusive
A and B are disjoint events

$$P(A \cap B) = 0$$



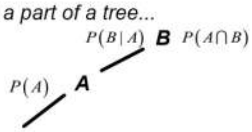
A and B are non mutually-exclusive
A and B are not disjoint events
A and B are joint events

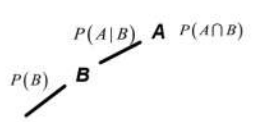
$$P(A \cap B) \neq 0$$

5: Compound Events (AND)

$$P(A \cap B) = P(A) \bullet P(B \mid A) \qquad P(A \cap B) = P(B) \bullet P(A \mid B)$$

Picture a part of a tree...





Special Case: If the probability of an event does not change regardless of whether or not another event happens, then the events are independent events.

For independent events: $P(B) = P(B \mid A)$

So... $P(A \cap B) = P(A) \bullet P(B \mid A)$ ...simplifies to... $P(A \cap B) = P(A) \bullet P(B)$

5: Independent Events

Test for independent events:

Two events are independent if:

$$P(B) = P(B \mid A) = P(B \mid \overline{A})$$

(check any two)

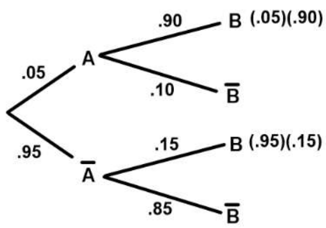
Note: Some books also use the simplified version of the AND formula as a 'test for independence'...

If $P(A \cap B) = P(A) \bullet P(B)$
then A and B are independent

...but this is more a consequence of independence, not the reason.

6: AND/OR together

We often need to use the AND and OR rules together:

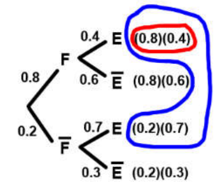


$$B = (A \text{ and } B) \text{ or } (\overline{A} \text{ and } B)$$
$$P(B) = P(A \text{ and } B) \text{ or } P(\overline{A} \text{ and } B)$$
$$P(B) = P(A \text{ and } B) + P(\overline{A} \text{ and } B)$$
$$P(B) = (P(A) \cdot P(B|A)) + (P(\overline{A}) \cdot P(B|\overline{A}))$$
$$P(B) = (.05)(.90) + (.95)(.15)$$
$$P(B) = .1875$$

6: Bayes' Formula

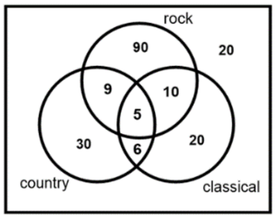
$$P(A|E) = \frac{P(A) \cdot P(E|A)}{P(E)}$$

But use probability of paths on a tree diagram:



$$P(F|E) = \frac{(0.8)(0.4)}{(0.8)(0.4) + (0.2)(0.7)}$$

7: Venn Diagrams



Venn diagrams are great for word problems with lots of information.

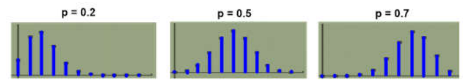
Always start with most overlapped region, and don't forget to subtract what has already been accounted for.

You can fill with either counts or probabilities (but be consistent).

8: Discrete Probability Models

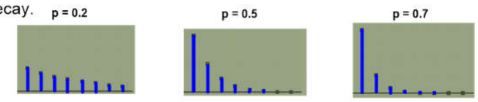
Binomial Shape depends upon p.

$\mu = np \qquad \sigma = \sqrt{npq}$



Geometric Shape is always exponential decay.

$\mu = \frac{1}{p} \qquad \sigma = \frac{\sqrt{1-p}}{p}$



General Discrete Models

$\mu = \text{'expected value'} = \sum X \cdot P(X)$

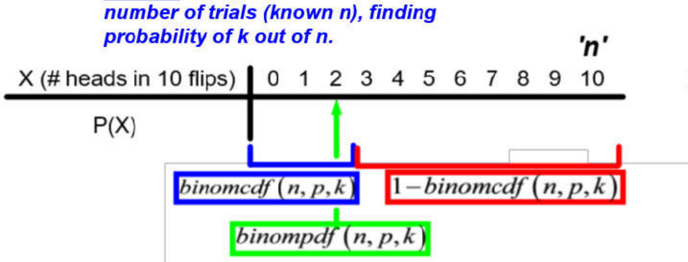
σ *(and* μ*) found using L1(data), L2(freqList), 1−Var Stats*

8: Discrete Probability Models

Binomial

- Only 2 outcomes
- Probabilities must be the same each trial.
- Probabilities of trials must be independent.
- Must have fixed number of trials, n

***Best for**: independent trials, fixed number of trials (known n), finding probability of k out of n.*

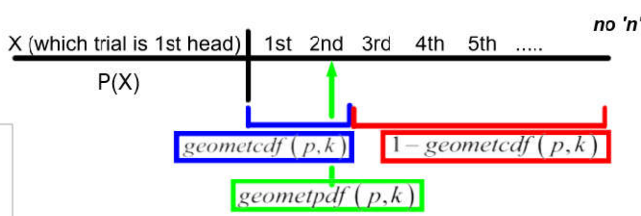


$$P(\text{exactly } k \text{ successes out of } n \text{ trials}) = {}_nC_k (p)^k (q)^{n-k}$$

Geometric

- Only 2 outcomes
- Probabilities must be the same each trial.
- Probabilities of trials must be independent.
- May or may not have fixed number of trials, n

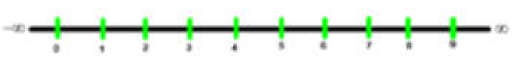
***Best for**: independent trials, non-fixed number of trials (unknown n), finding probability of 'when' the 1st success occurs.*



$$P(\text{success on the } k^{th} \text{ trial}) = (q)^{k-1}(p)$$

9: Discrete vs. Continuous

Discrete



Discrete variables take on only specific values

Discrete situations often involve 'counting' the number of items in specific categories so discrete variables are sometimes referred to as 'categorical' or 'qualitative'

We can use Binomial or Geometric models to analyze probability, and the discrete expected value formula to find the mean of a discrete distribution

Continuous

Continuous variables can take on any value (sometimes within specified limits)

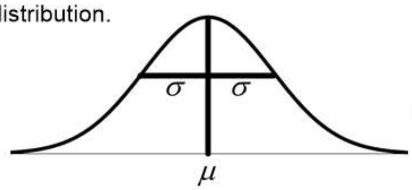
Continuous situations always involve a variable that is numerical (and usually includes units). Continuous variables are sometimes referred to as 'numerical' or 'quantitative'

We use an integral to find the area under the curve between boundaries to find probability.

The normalcdf function finds the area for a normal model.

10: The Normal Model

Many values which have a continuous, infinite number of possible outcomes, especially quantities found in natural systems, can be modeled with a Normal distribution.

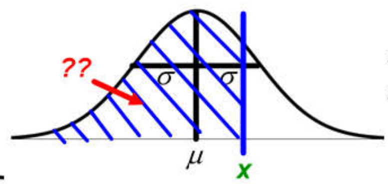


Normal distributions are symmetrical, centered at a mean μ

The average distance data is from this mean (on both sides) is called the standard deviation σ

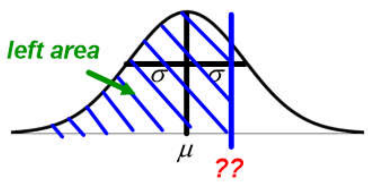
2 calculator functions for use with a Normal distribution:

<u>Have boundaries ⟶ Need area</u>



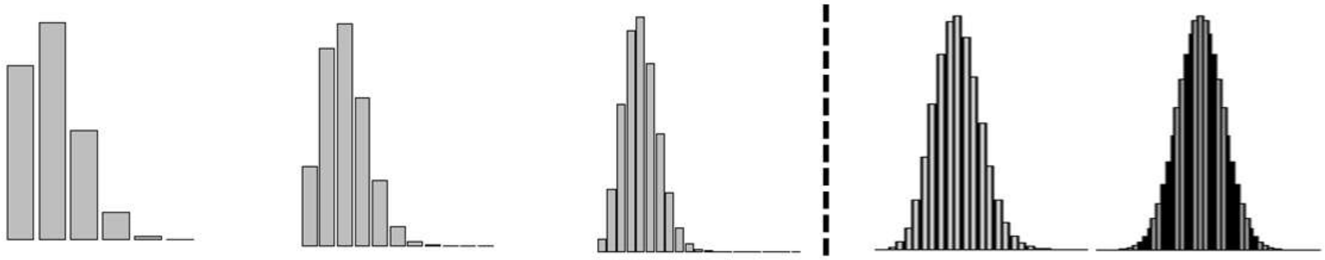
$$area = normalcdf\left(left\ boundary, right\ boundary, \mu, \sigma\right)$$
$$area = normalcdf\left(-999, x, \mu, \sigma\right)$$

<u>Have area ⟶ Need boundary</u>



$$upper\ boundary = invNorm\left(left\ area, \mu, \sigma\right)$$
$$x = invNorm\left(left\ area, \mu, \sigma\right)$$

10: Normal Approximation of Binomial Model



$p = 0.2, n = 5$ $(np = 1)$ | $p = 0.2, n = 10$ $(np = 2)$ | $p = 0.2, n = 20$ $(np = 4)$ | $p = 0.2, n = 50$ $(np = 10)$ | $p = 0.2, n = 100$ $(np = 20)$

Can use Normal approximation for the Binomial distribution.

If $np \ge 10$ *and* $nq \ge 10$

a Binomial distribution can be approximated with a Normal distribution with:

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

11: Combining Multiple Distributions

Define an algebraic expression for how the source distributions are used to build the new distribution:

$$E = A + B - C - D$$

The means are always determined by <u>the defining algebraic expression</u>:

$$\mu_E = \mu_A + \mu_B - \mu_C - \mu_D$$

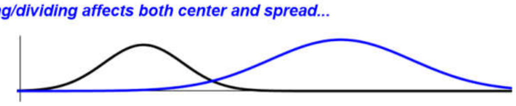
But because each source of variability increases <u>overall variation, the variances always add</u>:

$$\sigma_E^{\ 2} = \sigma_A^{\ 2} + \sigma_B^{\ 2} + \sigma_C^{\ 2} + \sigma_D^{\ 2}$$

However, we must know for certain that the variables are all varying independently of one another. (If not independent, we can find mean but not standard deviation).

11: Transforming a Single Distribution

Multiplying/dividing affects both center and spread...



Adding/Subtracting affects only center...

If $Y = aX \pm b$ $\quad \mu_Y = a\mu_X \pm b$ $\quad \sigma_Y = a\sigma_X$

Inference

'Canned' Interpretations:

<u>Slope of a regression line</u>: For each increase of 1 unit of the explanatory variable, there is an increase(decrease) of b units of the response variable (where b is the slope).

<u>Correlation Coefficient (r)</u>: (ex: if r=.758) There is a moderately strong positive association between the ______(explanatory variable) and the _______ (response variable).

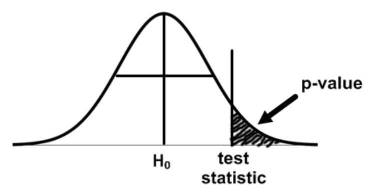
<u>Coefficient of determination (r^2)</u>: Percentage or fraction of variation in y that is explained by the LSRL which relates the explanatory variable to the response variable (note: $1-r^2$ = % of variability in y that is left in the residuals).

<u>Interpretation of a Confidence Interval</u>: We are __% confident that the true population _____ (mean, proportion, difference of means, etc.) lies within the interval (,).

<u>Interpretation of a Confidence Level</u>: If we were to repeat this study many times on many samples of size n, and constructed confidence intervals for each, __% of the confidence intervals would contain the true population ___.

<u>Conclusion of an Inference Test</u>: If H_0 is rejected (low p): We have significant statistical evidence to conclude (H_A). If H_0 is not rejected (high p): We do not have significant statistical evidence to conclude (H_A).

<u>p-value</u>: The probability that if H_0 was true, we would observe a test statistic as far or further from H_0. (or: The probability that the observed statistic value (or an even more extreme value) could occur if H_0 was correct.



<u>Common z* values</u>: 90%: z*=1.64, 95%: z*=1.96, 99%: z*=2.576

<u>How to conduct inference</u>:

<u>Confidence intervals</u>: 1) Check assumptions (conditions).
2) Construct Confidence Interval.
3) Interpret Confidence Interval in context of the problem.

<u>Hypothesis Test</u>: 1) State H_0, H_A, and <u>Type of Test</u>.
2) Check assumptions (conditions).
3) Conduct test, report all necessary results including significance level $\alpha \left(usually\ \alpha = .05 \right)$.
4) Report decision (p-value< α , reject H_0) or (p-value> α , fail to reject H_0).
5) State conclusion in context of the problem.

<u>Errors</u>:

		Null Hypothesis is: True	Null Hypothesis is: False
Decision:	Reject	Type I error $P(I)=\alpha$	$Power = 1-\beta$
	Not Reject		Type II error $P(II)=\beta$

Power of test is the probability that the test correctly rejects a false null hypothesis (the probability that the test detects the observed difference if that difference is statistically significant).

Increase power of a test by:

<u>Increasing n</u>: $n \nearrow,\ \sigma \searrow,\ both\ \alpha, \beta \searrow,\ power = 1-\beta \nearrow$
(but may increase cost, put more people at testing risk)

<u>Increasing α</u>: $\alpha \nearrow,\ \beta \searrow,\ power = 1-\beta \nearrow$
(but increases chance of a Type I error)

Success/Fail? Percentages? **Inference for Proportions** 1 Proportion — 2 Proportions Z-statistics Normal distributions (no df)	Means of numbers? **Inference for Means** **1 Mean** df = n - 1 **2 Means** **2 Sample** — **Diff. of means** df = TI calc **Matched Pair** — **Mean of diffs.** df = n - 1 t-statistics, t distributions or if n>25: Z-statistics, Normal distributions	Bivariate (y vs. x) data? **Inference for Regression LSRL Slope** t-distributions df = n - 2 t-statistic: $t = \frac{b - \beta}{s_b}$ (statistic: b) (parameter: β) S_b = standard error of slope S = standard error of residuals $usually\ \beta_0 = 0, so\ t = \frac{b}{s_b}$	Counts? **Inference for Counts** χ^2 - statistics χ^2 distributions **1 col (or row)** (compared to expected %) **Goodness of Fit** df = #categories - 1 $\chi^2 = \sum \frac{(obs - exp)^2}{exp}$ **>1 col (or row)** **1 population** — **Independence** **>1 population** — **Homogeneity** df = (#rows - 1)(#cols - 1) expected cell count $= \frac{(row\ total)(col\ total)}{grand\ total}$
Hypotheses: <u>1 proportion</u>: 1PropZTest/Int $H_0 : p = p_0$ $H_A : p > p_0 (or <, \neq)$ <u>2 proportions</u>: 2PropZTest/Int $H_0 : p_1 = p_2\ (p_1 - p_2 = 0)$ $H_A : p_1 > p_2\ (p_1 - p_2 > 0)(or <, \neq)$	<u>1 mean</u>: T-Test/T-Interval $H_0 : \mu = \mu_0$ $H_A : \mu > \mu_0 (or <, \neq)$ <u>2 mean (independent)</u>: 2SampTTest/Int $H_0 : \mu_1 = \mu_2\ (\mu_1 - \mu_2 = 0)$ $H_A : \mu_1 > \mu_2\ (\mu_1 - \mu_2 > 0)(or <, \neq)$ <u>2 mean (matched pairs)</u>: TTest/Int on diffs $H_0 : \mu_D = 0$ $H_A : \mu_D > 0 (or <, \neq)$ $\mu_D = mean\ of\ diffs$	<u>slope</u>: LinRegTTest/Int $H_0 : \beta = 0 (no\ association)$ $H_A : \beta \neq 0 (or <, >)(association)$ $CI : b \pm (t^*)(s_b)$	<u>GOF</u>: X^2GOF-test (obs in L1, exp in L2) H_0: Observed distribution of counts same as expected. H_A: Observed distribution of counts <u>not</u> same as expected. <u>Independence</u>: X^2-Test (2D data in matrix A) H_0: Row and column variables are independent. H_A: Row and column variables are <u>not</u> independent. <u>Homogeneity</u>: X^2-Test (2D data in matrix A) H_0: The distribution of ___ is the same among all populations. H_A: The distribution of ___ is <u>not</u> the same among all populations
Conditions: <u>1 proportion</u>: SRS, n<10%pop, success/fail >10 <u>2 proportions</u>: For each group... SRS, n<10%pop, success/fail >10 Groups independent of each other	<u>1 mean</u>: SRS, n<10%pop, Nearly Normal <u>2 means (indep)</u>: Groups independent For each group... SRS, n<10%pop, Nearly Normal <u>2 means (matched)</u>: How matched? SRS, n<10%pop, diffs are Nearly Normal	Straight enough Residuals show no pattern or fanning Residuals are Nearly Normal	All cell expected counts are > 5 - or - 80% of cells' expected counts are > 5 and none of the expected counts are 0

Sampling distributions for proportions:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population:			
$\hat{p}$	$\mu_{\hat{p}} = p$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$	$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
For two populations:			
$\hat{p}_1 - \hat{p}_2$	$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$	$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$s_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ *When* $p_1 = p_2$ *is assumed*: $s_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}_C(1-\hat{p}_C)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ *where* $\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2}$

Sampling distributions for means:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For one population:			
$\overline{X}$	$\mu_{\overline{X}} = \mu$	$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$	$s_{\overline{X}} = \frac{s}{\sqrt{n}}$
For two populations:			
$\overline{X}_1 - \overline{X}_2$	$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2$	$\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$s_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

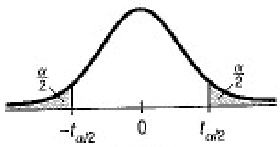
Sampling distributions for regression:

Random Variable	Parameters of Sampling Distribution		Standard Error* of Sample Statistic
For slope:			
b	$\mu_b = \beta$	$\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$ where $\sigma_x = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$	$s_b = \frac{s}{s_x \sqrt{n-1}}$ where $s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$ and $s_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$

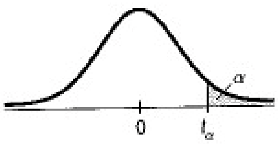
* Standard deviation is a measure of variability from the theoretical population. Standard error is the estimate of the standard deviation. If the standard deviation of the statistic is assumed to be known, then the standard deviation should be used instead of the standard error.

Table T

Values of t_α



Two tails



One tail

Two tail probability	0.20	0.10	0.05	0.02	0.01	
One tail probability	0.10	0.05	0.025	0.01	0.005	
df						df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
30	1.310	1.697	2.042	2.457	2.750	30
32	1.309	1.694	2.037	2.449	2.738	32
35	1.306	1.690	2.030	2.438	2.725	35
40	1.303	1.684	2.021	2.423	2.704	40
45	1.301	1.679	2.014	2.412	2.690	45
50	1.299	1.676	2.009	2.403	2.678	50
60	1.296	1.671	2.000	2.390	2.660	60
75	1.293	1.665	1.992	2.377	2.643	75
100	1.290	1.660	1.984	2.364	2.626	100
120	1.289	1.658	1.980	2.358	2.617	120
140	1.288	1.656	1.977	2.353	2.611	140
180	1.286	1.653	1.973	2.347	2.603	180
250	1.285	1.651	1.969	2.341	2.596	250
400	1.284	1.649	1.966	2.336	2.588	400
1000	1.282	1.646	1.962	2.330	2.581	1000
∞	1.282	1.645	1.960	2.326	2.576	∞
Confidence levels	80%	90%	95%	98%	99%	