Inference!
CI Formula
One-sample t-interval for μ

Inference!
Test Statistic Formula
One-sample t-test for μ

Inference!
CI Formula
One-sample z-interval for p

Inference!
Test Statistic Formula
One-sample z-test for p

Inference!
CI Formula
Two-sample t-interval for
$\mu_1 - \mu_2$

Inference!
Test Statistic Formula
Two-sample t-test for
$\mu_1 - \mu_2$

Inference!
CI Formula
Two-sample z-interval for
$p_1 - p_2$

Inference!
Test Statistic Formula
Two-sample z-test for
$p_1 - p_2$

Inference!
Test Statistic Formula
Chi-Square Test for
Homogeneity/Independence

Inference!
Test Statistic Formula
Chi-Square Test for
Goodness of Fit

Remember!
$df = n - 1$

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Remember!
$df = n - 1$

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Remember to use p_0 when checking the Large Counts condition!

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Remember to use $\hat{p}$ when checking the Large Counts condition!

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Pro Tip – You can also use 2-SampTTest on your calculator!

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Pro Tip – You can also use 2-SampTInt on your calculator!

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}}$$

Reminder:

$$\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2}$$

Time saver! Use 2-PropZTest on your calculator!

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Time saver! Use 2-PropZInt on your calculator!

Remember!
df = number of groups – 1

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Sum the components over all categories!

Remember!
df = (rows – 1)(columns – 1)

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Sum the components over all cells in the two-way table

Inference!
CI Formula
t-interval for slope

Inference!
Test Statistic Formula
t-test for slope

Inference!
Point Estimate and Margin of Error

Inference!
How to find the point estimate and margin of error given a confidence interval

Inference!
Type I and Type II error

Inference!
Power

Inference!
Interpret the confidence interval

Inference!
Interpret the confidence level

Inference!
Interpret the *P*-value

Inference!
What is an unbiased estimator?

Remember!
$df = n - 2$

$$t = \frac{b - \beta_0}{SE_b}$$

Remember!
$df = n - 2$

$$b \pm t^* SE_b$$

Suppose a confidence interval is (A, B).

The point estimate is the average of A and B. The point estimate is the exact center of the confidence interval!

To find the margin of error, subtract the point estimate from the upper bound of the confidence interval!
margin of error = B – (the point estimate)

Confidence intervals are of the form:
point estimate ± margin of error
statistic ± (critical value)(standard error of statistic)

Ex: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$

$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

The **power** of a test is the probability a test will correctly reject the null hypothesis, given the alternative hypothesis is true.

Remember!
P(Type II error) = 1 – *P*(power)

A **Type I error** occurs when the null hypothesis is true and is rejected (false positive).

A **Type II error** occurs when the alternative hypothesis is true and the null hypothesis is not rejected (false negative).

In repeated random sampling with the same sample size, approximately C% of confidence intervals created will capture the population parameter.

The population parameter could be:
- population proportion
- difference in population proportions
- population mean
- difference in population means
- population mean difference

We are C% confident that the confidence interval from ___ to ___ captures the population parameter (in context).

The population parameter could be:
- population proportion
- difference in population proportions
- population mean
- difference in population means
- population mean difference

When estimating a population parameter, a statistic is **unbiased** if, the center of the sampling distribution for the statistic is *equal to the population parameter*.

A *P*-value is the probability of obtaining a test statistic as extreme or more extreme than the observed test statistic when the null hypothesis is assumed to be true.

Inference!
Conditions for a one-sample t-test and t-interval for μ

Inference!
Conditions for a one-sample z-test and z-interval for p

Inference!
Conditions for a two-sample t-test and t-interval for $\mu_1 - \mu_2$

Inference!
Conditions for a two-sample z-test and z-interval for $p_1 - p_2$

Inference!
Conditions for a chi-square test

Inference!
Conditions for a t-test or t-interval for slope

Inference!
Why do we check the Random, 10%, and Normal/Large Sample conditions?

Inference!
What is the difference between a parameter and a statistic?

Inference!
What is the difference between the population distribution, the sample distribution, and the sampling distribution?

Inference!
What is the Central Limit Theorem?

Random: Data come from a random sample **10%:** When sampling without replacement, n < 10% of the population size **Large Counts:** • **Test:** $np_0 \geq 10$ and $n(1-p_0) \geq 10$ • **Interval:** $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$	**Random:** Data come from a random sample **10%:** When sampling without replacement, n < 10% of the population size **Normal:** Population distribution is normal, large sample ($n \geq 30$), or a dotplot of the sample data shows no strong skewness or outliers.
Random: Data come from independent random samples or 2 groups in a randomized experiment. **10%:** when sampling without replacement: n < 10% of the population size for both samples. **Large Counts:** • Test: $n_1\hat{p}_C \geq 10$, $n_1(1-\hat{p}_C) \geq 10$, $n_2\hat{p}_C \geq 10$, $n_2(1-\hat{p}_C) \geq 10$ • CI: $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, $n_2(1-\hat{p}_2) \geq 10$	**Random:** Data come from independent random samples or 2 groups in a randomized experiment. **10%:** when sampling without replacement: n < 10% of the population size for both samples. **Normal:** For both populations, either the population distribution is normal, large sample ($n \geq 30$), or a dotplot of the sample data shows no strong skewness or outliers.
Linear: True relationship between the variables is linear. **Independent** observations, 10% condition when sampling without replacement **Normal:** Responses vary normally around the regression line for all x-values **Equal Variance** around the regression line for all x-values **Random:** Data from a random sample or randomized experiment	**Random:** Data from a random sample, separate random samples, or groups in a randomized experiment. **10%:** when sampling without replacement: n < 10% of the population size for all samples. **Large Counts**: All expected counts must be at least 5.
A **parameter** is a number that describes the population. Ex: μ, p, σ A **statistic** is a number that describes the sample. Ex: $\bar{x}$, $\hat{p}$, s	**Random…**so we can generalize to the population from which the sample was selected. **10% condition…**so sampling without replacement is OK and we can use the stated formula for standard deviation. **Normal/Large Sample…**so the sampling distribution is approximately Normal.
The **central limit theorem** (CLT) states that when the sample size is sufficiently large, a sampling distribution of the mean of a random variable will be approximately normally distributed.	The **population distribution** is the distribution of responses for *every individual of the population*. The **sample distribution** is the distribution of responses for a *single sample*. The **sampling distribution** is the distribution of values for the statistic for *all possible samples* of a given size from a given population.

Inference!
What calculator function is used for a one-sample t-interval for μ?

Inference!
What calculator function is used for a one-sample t-test for μ?

Inference!
What calculator function is used for a one-sample z-interval for p?

Inference!
What calculator function is used for a one-sample z-test for p?

Inference!
What factors affect the width of a confidence interval?

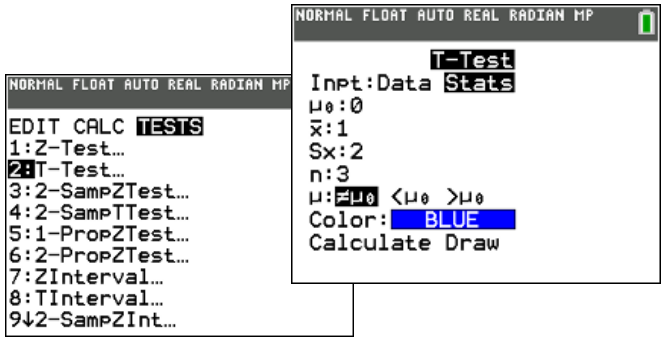
Inference!
How do I make a decision based on a *P*-value?

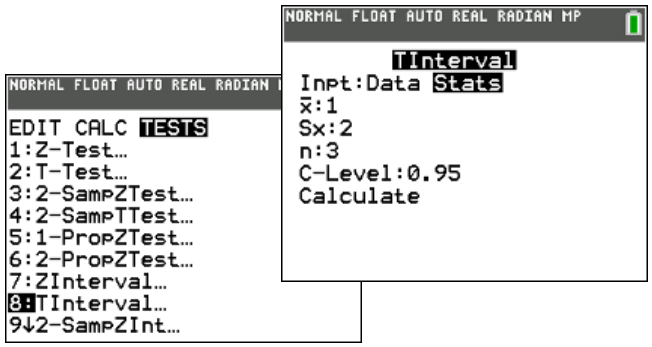
Inference!
What does it mean if we reject (or fail to reject) the null hypothesis?

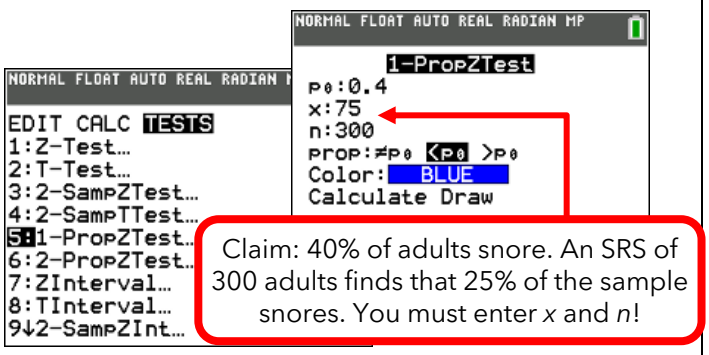
Inference!
What is the probability that a specific confidence interval captures the population parameter?

Inference!
How to calculate expected counts in a chi-square test for homogeneity/independence

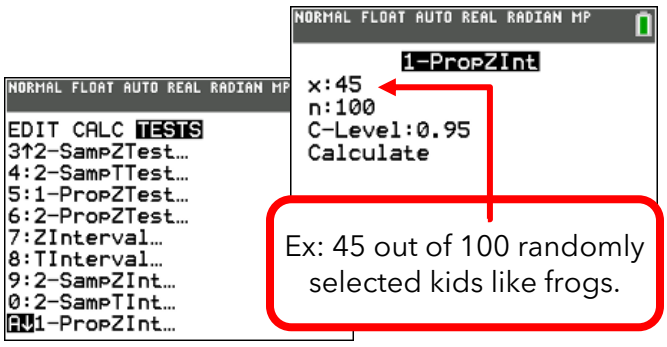
Inference!
How to choose the right inference procedure?







Claim: 40% of adults snore. An SRS of 300 adults finds that 25% of the sample snores. You must enter *x* and *n*!



Ex: 45 out of 100 randomly selected kids like frogs.

If the *P*-value ≤ α, reject the null hypothesis.
If the *P*-value > α, fail to reject the null hypothesis.

The width of a confidence interval:

- Decreases as *n* increases
- Increases as the confidence level increases

0 or 1

A confidence interval, calculated from sample data either does or does not capture the population parameter!
Don't say there is a 95% chance/probability of capturing the population parameter!!!
(say 95% confident)

Rejecting the null hypothesis means there is convincing statistical evidence to support the alternative hypothesis.

Failing to reject the null means there is not convincing statistical evidence to support the alternative hypothesis.

Ask yourself:

- Does the scenario describe mean(s), proportion(s), counts, or slope?
- Does the scenario describe one sample, two samples, or paired data?
- Does the scenario describe a test or a confidence interval?

$$\text{expected count} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

Exploring Data!
Describe a Distribution

Exploring Data!
Outlier Rule

Exploring Data!
How can we use a graph to compare the mean and the median?

Exploring Data!
Interpret the standard deviation

Exploring Data!
How do we describe the relationship between two variables (like in a scatterplot)?

Exploring Data!
Compare two distributions

Exploring Data!
How to find the mean, SD, and 5-number summary using a graphing calculator

Exploring Data!
How to calculate a LSRL using a graphing calculator.

Exploring Data!
What is the IQR?

Exploring Data!
How do I calculate the percentile of a particular value in a data set?

An outlier is any value that falls more than 1.5IQR above Q_3 or below Q_1.

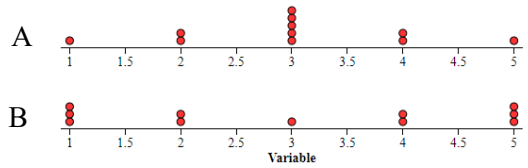
$$\text{Lower outliers} < Q_1 - 1.5(\text{IQR})$$
$$\text{Upper outliers} > Q_3 + 1.5(\text{IQR})$$

Note: Some consider any value that is more than 2SD away from the mean to be an outlier.

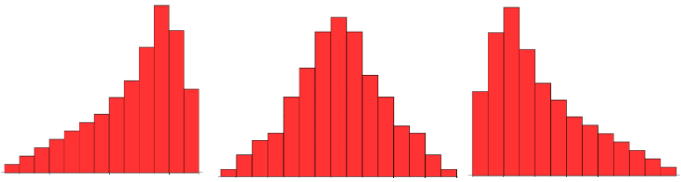
SOCV!

1. **Shape** – Skewed right? Skewed left? Fairly symmetric? Two distinct peaks?
2. **Outliers** – If you are estimating, call them "potential outliers."
3. **Center** – What is the mean? If the distribution is skewed, identify the median.
4. **Variability** – Remember, SD goes with the mean and IQR goes with the median.

The standard deviation gives that typical distance that the values are away from the mean.



The SD of Distribution A < The SD of Distribution B!



Skewed left
mean < median

Roughly symmetric
mean ≈ median

Skewed right
mean > median

SOCV

(shape, outliers, center, variability)

Remember to use COMPARISON words when describing the center and variability.

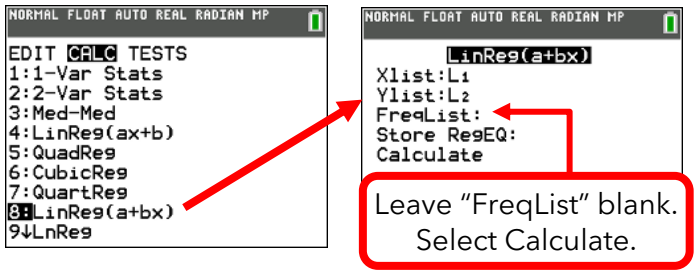
"is similar to"

"is less than" *"is greater than"*

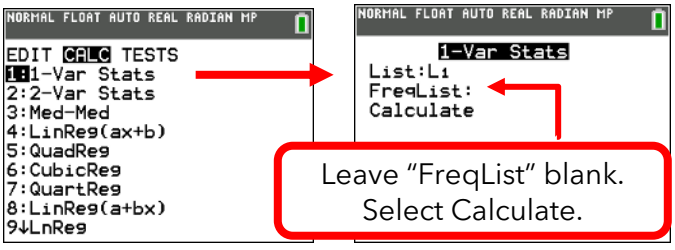
When describing the relationship between 2 quantitative variables, be sure to address:

1. **Direction** – positive or negative
2. **Unusual** values – outliers, influential observations
3. **Form** – Linear or curved
4. **Strength** – Weak → Strong

Enter the x-values in L1 and the y-values in L2.



Enter the data in List 1



- Order the data
- Count the number of values that are <u>less than or equal to</u> the value of interest.
- Count the number of values in the data set.

$$\text{Percentile} = \frac{\text{Number of values less than or equal to the value of interest}}{\text{Number of values in the data set}}$$

Express the decimal as a percentile.
Ex: 0.58 is the 58th percentile.

The **interquartile range** (IQR) is defined as the difference between the third and first quartiles: $Q_3 - Q_1$.

Q_1 and Q_3 form the boundaries for the middle 50% of values in an ordered data set.

Exploring Data!
Interpret the y-intercept of the Least Squares Regression Line

Exploring Data!
Interpret the slope of the Least Squares Regression Line

Exploring Data!
Interpret the coefficient of determination
r^2

Exploring Data!
Properties of correlation r

Exploring Data!
Regression Outlier

Exploring Data!
Correlation
r

Exploring Data!
High-Leverage Point

Exploring Data!
Influential Point

Exploring Data!
What is the difference between categorical and quantitative variables?

Exploring Data!
What is the difference between discrete and continuous variables?

slope

$$\hat{y} = a + bx$$

Interpretation:
For every increase of 1 unit in x context, the predicted y context increases/decreases by slope.

y-intercept

$$\hat{y} = a + bx$$

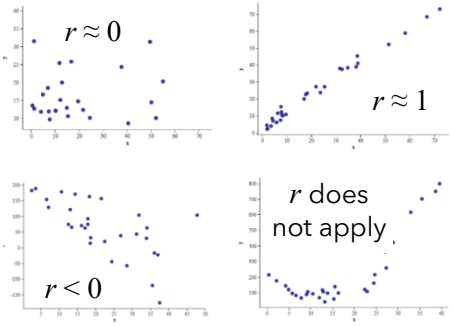
Interpretation:
The predicted value of y context when x context is 0 is y-intercept value.

Correlation, r

- r is unitless.
- r is always between –1 and 1.
- r is greatly affected by regression outliers.
- If the direction is negative, then $r < 0$.
- If the direction is positive, then $r > 0$.
- The closer r is to –1 or 1, the stronger the relationship.
- The closer that r is to 0, the weaker the relationship.

The **coefficient of determination** gives the percent of the variation of y-context that is explained by the least-squares regression line using x = x-context.

Correlation (r)
Gives the **strength** and **direction** of the linear relationship between 2 quantitative variables.



An **outlier** in regression is a point that *does not follow the general trend* shown in the rest of the data and *has a large residual*.

An **influential point** in regression is any point that, if removed, changes the relationship substantially (creates big changes to slope and/or y-intercept)

Outliers and high-leverage points are often influential.

A **high-leverage** point in regression has a *substantially* larger or smaller x-value than the other observations have.

A discrete variable can take on a countable number of values. The number of values may be finite or infinite.
THINK: Discrete = Countable
Ex. Number of students

A continuous variable can take on infinitely many values, but those values cannot be counted.
THINK: Continuous = Must be measured
Ex. Height

A **categorical** variable takes on values that are category names or group labels.

A **quantitative** variable is one that takes on numerical values for a measured or counted quantity.

Sampling/Experiments!

What is a control group?
What is the purpose of a control group?

Sampling/Experiments!

What is single blind and double blind?

Sampling/Experiments!

What are two poor sampling methods?

Sampling/Experiments!

Explanatory Variable
Response Variable

Sampling/Experiments!

Experimental Units
Subjects

Sampling/Experiments!

Types of Bias

Sampling/Experiments!

What is bias?

Sampling/Experiments!

Can increasing the sample size correct issues that arose from a biased sampling method?

Sampling/Experiments!

Can the results be generalized to a larger population?

Sampling/Experiments!

What is the difference between an observational study and an experiment?

In a **single-blind** experiment, subjects do not know which treatment they are receiving, but members of the research team do, or vice versa.

In a **double-blind** experiment neither the subjects nor the members of the research team who interact with them know which treatment a subject is receiving.

A **control group** is a collection of experimental units that are either not given a treatment of interest or given a treatment with an inactive substance (placebo).

The **purpose** of a control group is to provide a baseline to which the treatment groups can be compared, so it can be determined if the treatments have an effect.

An **explanatory variable** (or factor) in an experiment is a variable whose levels are manipulated intentionally.

A **response variable** in an experiment is an outcome from the experimental units that is measured after the treatments have been administered.

(1) convenience sampling and
(2) voluntary response sampling

These non-random sampling methods introduce potential for bias because they do not use chance to select the individuals.

Types of Bias

Nonresponse – selected people do not respond
Undercoverage – systematically excluding people from being able to be selected
Response bias – providing inaccurate responses (on purpose or by accident)
Wording Issues – confusing wording or question is slanted towards a particular response

When an experiment is performed on animals or objects, we call those animals or objects **experimental units**.

When an experiment is performed on humans, we call them **subjects**.

NO! NO! NO!

A flawed sampling design can never be improved by increasing the sample size…you'll just get a BIGGER flawed sample.

Note: A larger sample size does reduce variability when the sampling design is NOT flawed. ☺

Bias is the systematic tendency to overestimate or underestimate the true population parameter.

Did the researchers impose a treatment?

If NO → observational study
If YES → experiment

The results of a survey or experiment can only be generalized **to the population from which the sample/subjects were randomly selected**.

If the sample/subjects were not randomly selected then the results can only be generalized to "people like the ones in the study."

Sampling/Experiments!
When can we make conclusions about cause and effect?

Sampling/Experiments!
How to carry out a random assignment by selecting from a hat

Sampling/Experiments!
Simple Random Sample (SRS)

Sampling/Experiments!
Stratified random sample

Sampling/Experiments!
Confounding Variable

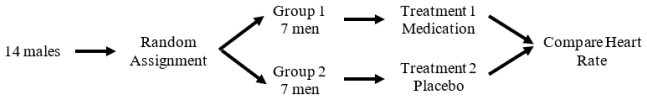
Sampling/Experiments!
Systematic random sample

Sampling/Experiments!
What should a well-designed experiment include?

Sampling/Experiments!
Completely Randomized Design

Sampling/Experiments!
What is a randomized block design and what is the purpose?

Sampling/Experiments!
Matched Pairs Design

• Write each subject's name on **equal sized** slips of paper. • Put all the slips of paper in a hat. **Mix well**. • Select as many names as needed for each treatment group, **without replacement**.	**Did the researchers randomly assign the subjects to treatment groups?** If YES → YAY! You can make conclusions about cause and effect. If NO → You CANNOT say that the explanatory variable CAUSED the change in the response variable.
A **stratified random sample** involves the division of a population into separate groups, called strata, based on shared attributes or characteristics (homogeneous grouping). Within each stratum a simple random sample is selected, and the selected units are combined to form the sample. Ex: Randomly select 25 seniors and 25 juniors.	A **simple random sample** (SRS) is a sample in which <u>every group of a given size</u> has *an equal chance* of being chosen. To obtain an SRSs (1) number individuals and use a random number generator to select which ones to include in the sample, ignoring repeats, (2) use a table of random numbers, ignoring repeats, or (3) draw names from a hat, without replacement.
A **systematic random sample** is a method in which sample members from a population are selected according to a random starting point and a fixed, periodic interval. Ex: Select a random person between 1 and 5 to start the process. Survey every 5th person to enter a building thereafter.	A **confounding variable** is a variable that is related to the explanatory variable and influences the response variable and makes it challenging to determine cause and effect.
In a **completely randomized design**, treatments are assigned to experimental units completely at random. Random assignment tends to create roughly equivalent groups, so that differences in responses can be attributed to the treatments.  	a. **Comparisons** of at least two treatment groups, one of which could be a control group. b. **Random assignment** of treatments to experimental units. c. **Replication** enough experimental units in each treatment group to be able to detect a difference. d. **Control** of potential confounding variables, where appropriate.
A **matched pairs design** is a special case of a randomized block design. Using a blocking variable, subjects are arranged in pairs matched on one or more relevant factors. Every pair receives both treatments by randomly assigning one treatment to one member of the pair and subsequently assigning the remaining treatment to the second member of the pair. Alternately, each subject may get both treatments.	For a **randomized block design**, treatments are assigned completely at random within each block. For each block, individuals are similar to each other with respect to at least one blocking variable. The purpose of blocking is: • to reduce the variability of results within each treatment group • to eliminate the possibility of the blocking variable as a confounding variable.

Probability!
Mean and Standard Deviation of a Binomial Distribution

Probability!
Mean and Standard Deviation of a Geometric Distribution

Probability!
Formula for the Binomial probability $P(X = x)$

Probability!
Formula for the Geometric probability $P(X = x)$

Probability!
Conditions for a Binomial Random Variable

Probability!
Conditions for a Geometric Random Variable

Probability!
What is the Law of Large Numbers?

Probability!
How do I calculate the probability of "at least 1"?

Probability!
How do I calculate a conditional probability?

Probability!
How can I tell if two events are independent?

$$\mu_X = \frac{1}{p}$$

Remember that p is the probability of success!

$$\sigma_X = \frac{\sqrt{1-p}}{p}$$

$$\mu_X = np$$

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

Remember that n is the number of trials and p is the probability of success!

$$P(X = x) = (1-p)^{x-1} p$$

p is the probability of success
x is the number of successes

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This is ${}_nC_x$ on the calculator!

n is the number of trials
p is the probability of success
x is the number of successes

GEOMETRIC RANDOM VARIABLE

1. **Binary:** two outcomes for each trial (success or failure)
2. **Independent:** Each trial is independent of the next
3. **Trials** UNTIL a success (not fixed number)
4. **Same probability** of success for each trial (p)

Remember: Keep going until a success.

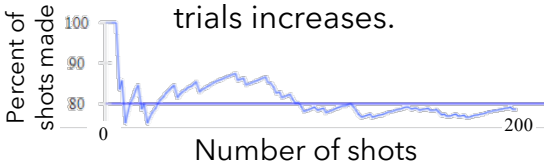
BINOMIAL RANDOM VARIABLE

1. **Binary:** two outcomes for each trial (success or failure)
2. **Independent:** Each trial is independent of the next
3. **Number of trials** is a fixed number (n)
4. **Same probability** of success for each trial (p)

Remember: Fixed number of trials.

P(At least 1) = 1 – P(none)

The **law of large numbers** states that simulated (empirical) probabilities tend to get closer to the true probability as the number of trials increases.



Two events are **independent** if any of the following are true. You only need to check one of these!

P(A | B) = P(A)

P(B | A) = P(B)

P(A and B) = P(A) · P(B)

$$P(\text{A} \mid \text{B}) = \frac{P(\text{A and B})}{P(\text{B})}$$

Probability! Mean and Standard Deviation of the Sum of 2 Independent Random Variables	**Probability!** Mean and Standard Deviation of a Discrete Random Variable
Probability! Mean and Standard Deviation of the Difference of 2 Independent Random Variables	**Probability!** Formula and interpretation of a z-score
Probability! Calculator function to find the area under a Normal distribution	**Probability!** Calculator function to find a value from a given area under a Normal distribution
Probability! What is the empirical rule?	**Probability!** What is the difference between binompdf and binomcdf?
Probability! Mutually Exclusive Events	**Probability!** How do I calculate P(A and B)?

$$\mu_X = E(X) = \sum x_i \cdot P(x_i)$$

$$\sigma_X = \sqrt{\sum (x_i - \mu_x)^2 \cdot P(x_i)}$$

$$\mu_T = \mu_X + \mu_Y$$

$$\sigma_T = \sqrt{{\sigma_X}^2 + {\sigma_Y}^2}$$

These formulas are NOT on the AP Exam formula sheet.

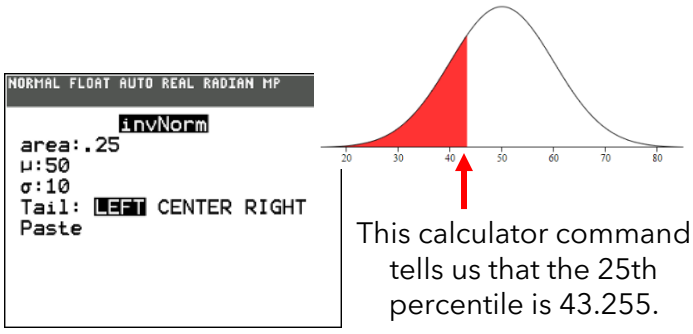
$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

A ***z*-score** tells us how many standard deviations a value falls away from the mean and in what direction.

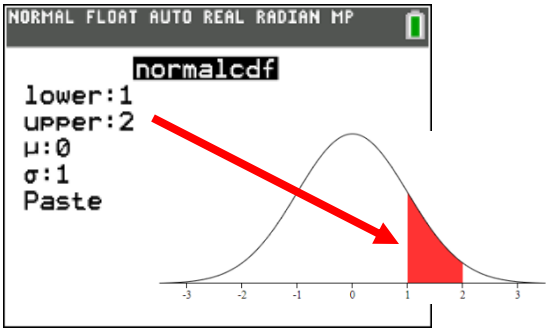
$$\mu_D = \mu_X - \mu_Y$$

$$\sigma_D = \sqrt{{\sigma_X}^2 + {\sigma_Y}^2}$$

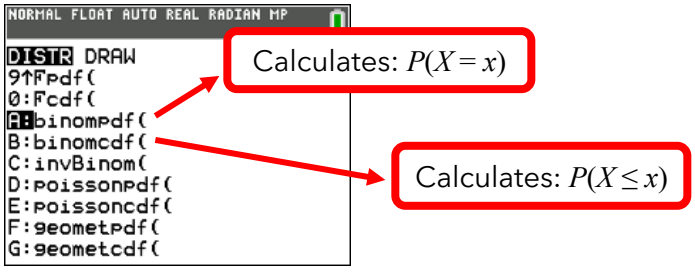
These formulas are NOT on the AP Exam formula sheet.

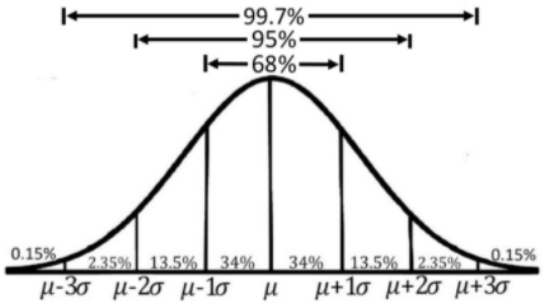


This calculator command tells us that the 25th percentile is 43.255.



binompdf(n, p, x) binomcdf(n, p, x)





$P(\text{A and B}) = P(\text{A}) \cdot P(\text{B} \mid \text{A})$

Two events are **mutually exclusive** (disjoint) if they cannot occur at the same time.

$P(\text{A and B}) = 0$