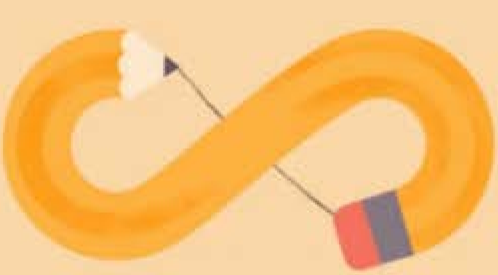
STUDYRESOURCES DISCORD

AP EXAM 2020

Notes

https://t.me/studyresources3



StudyResources AP Statistics Review Sheet. https://t.me/apresources

BEFORE WE START:

Calculator Tips

- Press “MODE” and confirm that “Stat Diagnostics” are turned on.

- Make sure to update your calculator to the latest operating system. Non-color TI-84s should be at least version 2.55 MP, and the TI-84 Plus CE should be at least 5.3.0. You can check your calculator version with [2nd] [+] [Enter]

1. Exploring One-Variable Data

1.1 analyzing categorical data

- Places an individual into of several categories
- Bar graphs, pie charts and venn-diagrams can be used to present this data

1.2 analyzing quantitative data

- Takes numerical values for which arithmetic operations make sense (ex: # of siblings)
- Dot plots, box plots, stem plots, scatter plots and histograms
- A stem plot gives a quick picture of the shape of a distribution while including numerical values
    - Separate each observation into a stem and a leaf
        - Write leaf to right of stem. Stems in vertical column and draw a vertical line to right of column
    - Histograms break the range of data values into classes and displays the count/% of observations that fall into that class
        - To make a histogram:
            - Divide range of data into equal-width classes
            - Count observations in each class→ “frequency”
            - Draw bars to represent classes→ height=frequency
            - Bars should touch unlike bar graphs
            - Should have between 4-12 bins
    - Boxplot: Displays 5 number summary
        - Min, Q1, median, Q3, max
        - A value is an outlier if it is less than Q1-1.5(IQR) or greater than Q3+1.5(IQR)
            - Interquartile range (IQR) → Q3-Q1
        - Modified Boxplot displays outliers as separate dots.

1.3 Describing Distributions SOCS

- Look for overall pattern and for striking deviation from that pattern
- There are 3 important things to note when looking at distributions

  - Shape: Symmetric or skewed (left or right)
    - Symmetric → mean=median
    - Skewed Right → mean>median
    - Skewed Left → mean<median
  - Center: What number typifies the data (mean vs. median)
  - Spread: How variable are the data values (Range)
    - Highest-Lowest
  - Outliers

Lower Outlier = Q1 – (1.5 x IQR)

Higher Outlier = Q3 + (1.5 x IQR)

- Median is resistant to extreme values because it is not easily affected by outliers
- Mean is not resistant because it is easily affected by outliers
- Percentile: Another measure of relative standing is a percentile rank
  - 'P'th percentile → value w/ p% observations below it
  - Median is the 50th percentile
- Quartiles: Q1 + Q3 represent the 25th and 75th percentiles
  - Find them by ordering data min to max and determine the median. Q1 is middle of bottom half and Q3 is middle of top half
- Variance and Standard Deviation
  - Standard Deviation: typically how far away a value is from the mean of a data set (average distance from mean)
  - $variance\ = \sigma^2$
  - $\sqrt{}$var. = st.dev.
  - $\sigma = population\ standard\ deviation$
  - $\mu = population\ mean$
  - $S =\ sample\ standard\ deviation$
  - $\overline{x} =\ sample\ mean$
  - $\sigma^2 =\ population\ variance$
  - $S^2 =\ sample\ variance$

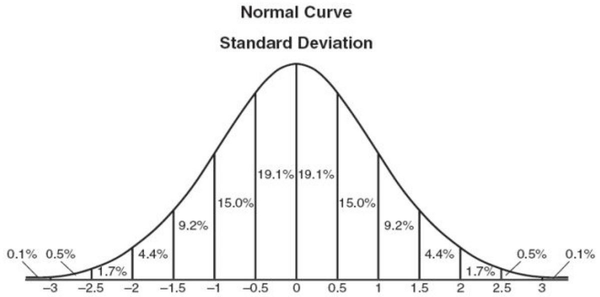
→ CALCULATOR: find by placing in lists, run 1-Var Stats STAT → CALC → 1:1-Var Stats

- Standardized Value
  - One way to describe relative position in a data set is to tell how many st. devs. above or below the mean the observation is.
  - Standardized Value: Z-score
  - $z = \frac{x-\mu}{\sigma}$

1.4 density curves and normal distribution

- Density Curve:
  - A density curve is an idealized description of the overall pattern of a distribution

  - Area underneath density curve = 1 and it represents 100% of observations
  - Density curve is valid only if it is on or above the horizontal axis and its area is 1
  - The mean of the density curve is at its balance point
  - The area of a density curve represents the % of observations that fall inside it
  - Median of density curve cuts area in half
- Normal Distribution
  - Empirical Rule 68-95-99.7
    - 68% of observations lie within 1 st. dev. of mean
    - 95% lie within 2 st. dev.
    - 99.7% lie within 3 st. dev.



CALCULATOR:

FIND Z-SCORE: 2nd → VARS → 3:InvNorm(area:, μ : , σ :)

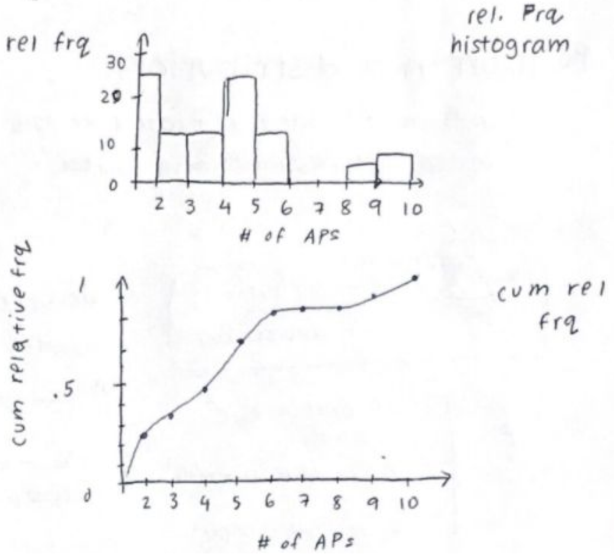
FIND AREA: 2nd → VARS → 2:normalcdf(L:, U:, μ : , σ :)

- Normal Probability Plot (NPP)
  - A NPP plots values of a data set vs. its standardized values (z-scores)
  - If the NPP is relatively linear we can say the distribution is relatively normal
  - A non-linear pattern indicates the data is not normal and outliers will be far from the overall pattern…. Not associated with correlation

→ CALCULATOR: to draw a NPP: 2nd → y= → type: (last one)

Cumulutive Relative Frequency

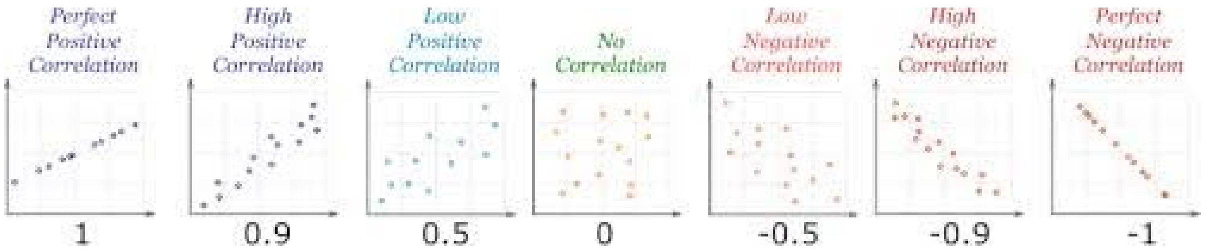
	APs	rel. Frq	percentile
2	6	0,25	0.25
3	3	0.125	0.375
4	3	0.125	0.5
5	6	0.25	0.75
6	3	0.125	0.875
7	0	0	0.875
8	0	0	0.875
9	1	0.04	0.92
10	2	0.083	1.0



" all models are wrong but some are useful "

2. Exploring Two-Variable Data

2.1 scatterplots and correlation



Scatterplots (quantitative data)

- Explanatory Variable; Horizontal Axis “x-axis” - independent variable
- Response Variable; Vertical Axis “y-axis” - dependent variable

Correlation

- r = correlation coefficient
  - r is positive there is a positive correlation
  - r is negative there is a negative correlation
  - The closer it is to 1 or -1 the stronger the correlation.
- r^2 = coefficient of determination (the variation in y explained x)
  - Closer to 1 = strong correlation
  - Interpreting r^2: “ _____ % of the variation in [response variable name] can be accounted for by the linear relationship with [explanatory variable name].”

HOW TO DESCRIBE SCATTER PLOTS:

- Strength (strong, moderately strong, moderately weak, weak)
- Direction (positive, negative)
- Shape (linear, nonlinear [curved, cluster, etc])
- Outliers (any points not conforming to the overall pattern)

2.2 least squares regression

A Line of best fit can be found for Linear Models using least-squares regression. This line minimizes the sum of squared errors. On a Ti-84 the line of best fit can be found by using CALCULATOR: find line of best fit: STAT → CALC → 4:LinReg(ax+b)

$\bar{y} = a + b\bar{x}$ → passes through $(\bar{x}, \bar{y})$

- Extrapolation: using predictions outside of data range (avoid this, as you don't know what happens to the trend afterwards)
- Residual = $y - \hat{y}$

HOW TO SOLVE FRQ:

- If it asks for the model:
    - $\hat{y} = b + ax$
    - Define $\hat{y}$ and x
- Interpreting slope:
    - For every 1 [unit] increase in [x in context], there is a [a] increase in expected value of [y in context].
- Is the linear model appropriate:
    - See if there is a pattern in residual plot.
        - If yes, not appropriate
        - If no, appropriate

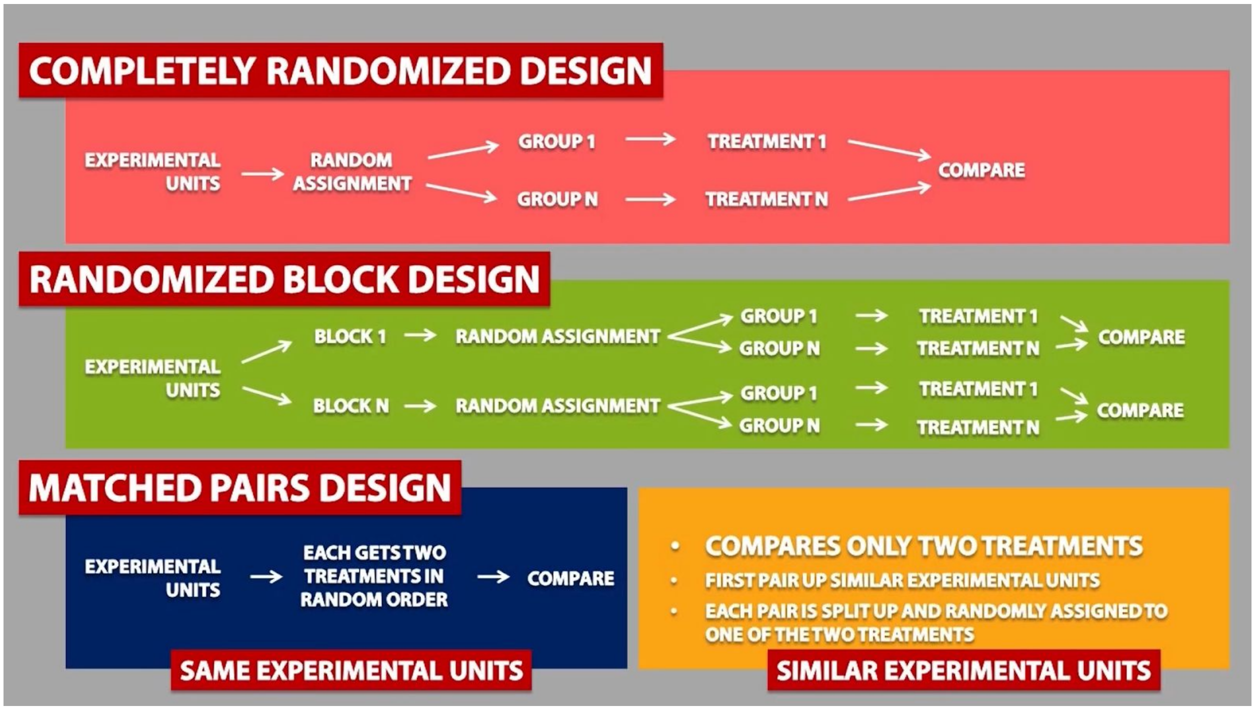
3. Collecting Data

3.1 sampling and surveys

- Sample:
    - A smaller group of individuals selected from a population
    - To *represent* the entire population, individuals are selected *at random*
        - The best defense against bias
    - Sample surveys
        - Surveys designed to ask questions of a small group of people in the hope of learning something about the entire population
    - Sample size
        - You need a large enough sample to be representative of the population
- Census
    - The entire population is the "sample"
    - Difficult to complete and impractical; some people might be difficult to get a hold of and populations rarely stand still
- Biases
    - Sampling methods that tend to over or underemphasize some characteristics of a population
    - Voluntary response bias
        - When individuals can choose to respond or not
        - ALWAYS invalid and can't be recovered
        - Form of nonresponse bias
    - Convenience sample

        - Consists of individuals who are conveniently available
        - Every individual in the population is not equally convenient to sample
    - Undercoverage
        - Part of the population is less represented than it is in the population
    - Nonresponse bias
        - When a large fraction of those sampled fails to respond
        - May occur for a variety of reasons
            - Ex. survey calls home phone during the day, but the individual is not home during the day
    - Response bias
        - Anything in a survey design that influences responses
        - Ex. wording of questions to favor one response
- Population and Parameters
    - Population parameter
        - A numerically valued attribute of a model for a population
        - We hope to estimate it from sampled data
        - Ex. the mean income of all employed people in the country
    - Sample statistic
        - Values calculated for sampled data
        - Ex. the calculated mean income of all employed people in a representative sample
    - A representative sample reflects the corresponding parameters accurately
- Simple Random Samples
    - Simple Random Sample (SRS)
        - A sample in which each set of *n* elements in the population has an equal chance of selection
    - Sampling frame
        - A list of individuals from whom the sample is drawn
    - Sampling variability
        - The natural tendency of randomly drawn samples to differ
        - Not an error
- Stratified Sampling
    - Strata
        - Homogenous groups
    - Stratified random sampling
        - When the population is sliced into strata before the sample is selected
        - Then simple random sampling is used within each stratum
    - Benefit(s)
        - Reduced sampling variability
- Cluster and Multistage Sampling
    - Cluster sampling

        - Works by splitting the population into representative clusters and then selecting one or a few clusters at random and performing a census within each of them
        - Clusters should be representative of the population
    - Multistage sampling
        - Sampling schemes that combine several methods
- Systematic sampling
    - A sample drawn by selecting individuals systematically from a sampling frame
    - Can be representative when there is no relationship between the order of the sampling frame and the variables of interest
- CLUSTER vs. STRATIFIED SAMPLING
    - Strata are internally homogeneous but differ from one another. Clusters can be heterogeneous; we want them to represent the population



COMPLETELY RANDOMIZED DESIGN

EXPERIMENTAL UNITS → RANDOM ASSIGNMENT → GROUP 1 → TREATMENT 1 → COMPARE
RANDOM ASSIGNMENT → GROUP N → TREATMENT N → COMPARE

RANDOMIZED BLOCK DESIGN

EXPERIMENTAL UNITS → BLOCK 1 → RANDOM ASSIGNMENT → GROUP 1 → TREATMENT 1 → COMPARE
RANDOM ASSIGNMENT → GROUP N → TREATMENT N → COMPARE
EXPERIMENTAL UNITS → BLOCK N → RANDOM ASSIGNMENT → GROUP 1 → TREATMENT 1 → COMPARE
RANDOM ASSIGNMENT → GROUP N → TREATMENT N → COMPARE

MATCHED PAIRS DESIGN

EXPERIMENTAL UNITS → EACH GETS TWO TREATMENTS IN RANDOM ORDER → COMPARE

SAME EXPERIMENTAL UNITS

- COMPARES ONLY TWO TREATMENTS
- FIRST PAIR UP SIMILAR EXPERIMENTAL UNITS
- EACH PAIR IS SPLIT UP AND RANDOMLY ASSIGNED TO ONE OF THE TWO TREATMENTS

SIMILAR EXPERIMENTAL UNITS

3.2 experiments

- Experiments are when an actual treatment is DONE to a group of units. (Like samples but active).
- Experimental Units - individuals on which the experiment is done
- Subjects - if experimental units are humans
- Treatment - a specific experimental condition applied to the units
- Factors - Explanatory variables
- Level - a specific value or amount of factor

- Placebo - dummy bill that does nothing (usually used as control, preserves blindness)
- Experiments give good evidence for causation.
- Comparative experiments
  - Treatment -> Observe response
  - Gives more control, so can reduce lurking variables
- Bias can exist if an outcome in the experiment is more favored
  - Such as a pharmacy pushing the good effects of a drug
- Randomization
  - (see previous section for designs)
  - Methods
    - Table of random digits
      - assign each unit a number and go through each digit
    - Papers in hat
    - Ti84: MATH → PRB → 5:randInt(L: , U:)
- Principles of Experimental design
  - Control
    - Is there a control? Are lurking variables prevented?
  - Randomization
    - Are the units and treatments assigned randomly?
  - Replication
    - Can this be done multiple times to reduce chance variation?
- Statistically significant - An effect that couldn't have been chance.
- Double blind - neither subjects or doctor knows which treatment is given to which subject.
  - But a 3rd party needs to know.
- Matched Pairs
  - Pair up similar subjects and give each a different treatment. Observe any differences.
- Confounding Variable - adds effect to response and cannot be distinguished from the independent variable

3.3 responsible stats (causation, scope, ethics)

- Random assignment → ensures two groups are as similar as possible before treatment is imposed
  - Allows us to make a causal inference
- Challenges of establishing causation:
  - Lack of realism → if experiment results are due to specific environment and specific individuals
  - Consider having randomized comparative experiment
- How to establish causation without experiment
  - Association is strong and consistent
  - Larger values of experimental variable associated with stronger response

  - Alleged cause precedes the effect of time
  - Cause is plausible
- Data Ethics:
  - Do no harm
  - No discrimination
  - Planned studies must be reviewed by institutional review board
  - Informed consent from all individuals
  - All individual data is confidential

4. Probability, Random Variables, and Probability Distributions

4.1 randomness, probability, and simulation

- Chance behavior is unpredictable in the short run, but has a regular and predictable pattern in the long run
- Law of large numbers: guarantees that as more and more repetitions are made, the proportion of times that a specific outcome occurs approaches a single value aka the true probability.
- The probability of any outcome of a chance process is a number between 0 and 1.
- random =/= "haphazard" in chance, but rather "by chance."
- Runs that seem "not random" to our intuition can be quite common
- FALSE: Law of averages: must lose to win so that the "wins" and "losses" balance out
- Simulation: imitation of chance behaviour, based on a model that accurately reflects the situation
- FRQ Performing a simulation
  - STATE: ask a question of interest about a chance process
  - PLAN: describe how to use a chance device to imitate one repetition of the process. Tell us what you will record at the end of each repetition.
  - DO: perform many repetitions of the simulation
  - CONCLUDE: use the results of your simulation to answer the question from the STATE step.

4.2 probability rules

- Two events are mutually exclusive or disjoint if they cannot occur at the same time.
- The probability that Event A occurs, given that Event B has occurred, is called a conditional probability. The conditional probability of Event A, given Event B, is denoted by the symbol P(A|B).
- The complement of an event is the event not occurring. The probability that Event A will not occur is denoted by $P(A^C)$.

- The probability that Events A and B *both* occur is the probability of the intersection of A and B. The probability of the intersection of Events A and B is denoted by P(A ∩ B). If Events A and B are mutually exclusive, P(A ∩ B) = 0.
- The probability that Events A or B occur is the probability of the union of A and B. The probability of the union of Events A and B is denoted by P(A ∪ B) .
- If the occurrence of Event A changes the probability of Event B, then Events A and B are dependent. On the other hand, if the occurrence of Event A does not change the probability of Event B, then Events A and B are independent.

Formulas

- Test for Independence
  - P(A) = P(A I B)
- Addition Rule
  - P(A or B) = P(A) + P(B) - P(A and B)
- Multiplication Rule for All Events:
  - P(A and B) = P(A) * P(B I A)
- Multiplication Rule for Independent Events:
  - P(A and B) = P(A) * P(B)
- Conditional Probability
  - P(A I B) = P(A and B) / P(B)
- At least once
  - P(A happens at least once) = 1 - P(A does not happen)

4.3 Conditional Probability and Independence

- Conditional probability: P(A|B) This is read, find probability of A, given B.
- Independence: Suppose P(A)>0 and P(B)>0. Then events A and B are independent events if and only if, P(A|B)=P(A) or, equivalently P(B|A)= P(B).
- If two events are independent we can use: P(A and B)= P(A) · P(B)
- To prove independence for a joint probability distribution P(A|B)=P(A) must be proven for every event probability. If A and B are not independent then they are dependent.
- Only if A and B are independent:
  - $SD(a+b) = SD(a-b) = \sqrt{(SD(a))^2 + (SD(b))^2}$
  - $var(a+b) = var(a-b) = var(a) + var(b)$

4.4 discrete vs continuous random variables

- Random variable: variable that has a single numerical value for each trial/ experiment. There are two main types, these are Discrete Random Variable and the Continuous Random Variable.

- Discrete Random Variable - Something that has a countable number of outcomes (number of touchdowns scored in a football game)
- Continuous Random Variable - Measurable outcomes; infinitely many (The time it takes a player to complete the video game Far Cry 4)

4.4.1 Discrete Random Variables

To show probability, we can use a *probability histogram* to represent the probability model.

- This is because each outcome has a specific probability to it.
  - Number of heads flipped, baby genders, etc.

4.4.2 Continuous Random Variables

To show probability, we can use a *density curve* to represent the probability model.

4.5 Transforming and Combining Random Variables

4.5.1 Rules for Means

- $E(x \pm C) = E(x) \pm C$
- $E(ax) = a \times E(x)$
- $E(x \pm y) = E(x) \pm E(y)$

4.5.2 Rules for Variances

- $var(x \pm C) = var(x)$
- $var(ax) = a^2 \times var(x)$
- $var(x + y) = var(x - y) = var(x) + var(y)$

4.5.3 Rules for SD

- $SD(x \pm C) = SD(x)$
- $SD(ax) = a \times SD(x)$
- $SD(x + y) = SD(x - y) = \sqrt{(SD(x))^2 + (SD(y))^2}$

4.6 binomial and geometric variables

4.6.1 Binomial Variables

For a statistical experiment to be a binomial experiment, it must meet these four conditions (BINS):

- Binary - Each outcome is either a success (P) or a failure (Q).
- Independence - All trials are independent of each other
- Number - There are a fixed number, n, of trials.

- Success - The probability of success, p, is the same for each trial.

Definitions:

- When describing a binomial distribution, use the notation B(n, p).
- Binomial random variable: The number of successes, x, in n repeated trials of a binomial experiment.
- Binomial distribution: The probability distribution of a binomial random variable.

Examples of binomial distributions:

- The number of male or female births in the next 20 births at a local hospital
- The amount of successful free throws in the next 30 free throws
- The total correct answers on a 10-question multiple-choice exam when guessing at the answers
- The number of people who show up for a flight when 100 tickets are sold

Test checklist:

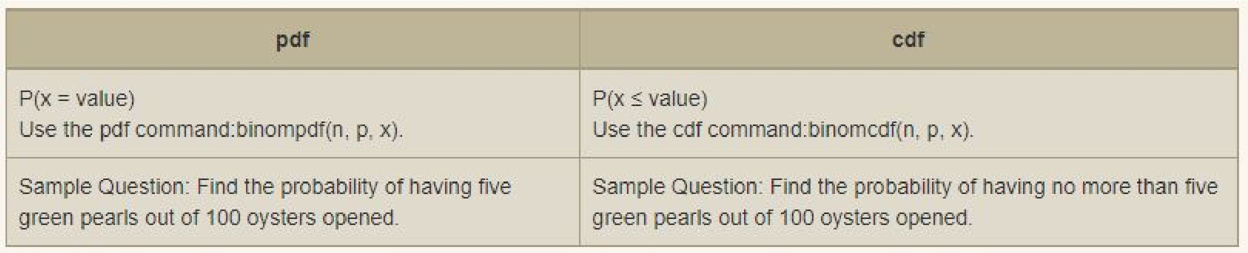
1. Check the 4 conditions
2. Verify the conditions in the text with context, eg: “Yes. There is success or failure (win a prize or don't win a prize), there is a fixed number of trials (n = 52), each lottery ticket's outcome is independent of one another, and there is a constant probability (p = 0.1).”
3. Describe your data using the notation B(n, p) -- where n is number and p is probability. Eg: B(52, 0.1)

Crucial formulas -- Binomial Distribution: Where $\binom{n}{x}$ is equal to the factorial $\frac{n!}{x!(n-x)!}$

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

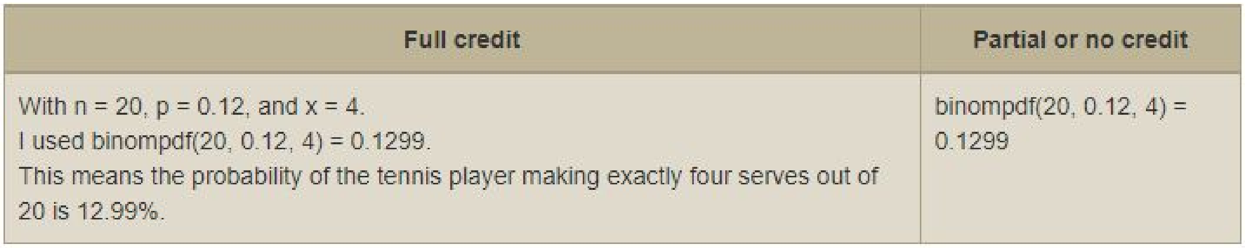
Cumulative Binomial Distribution: Add the probabilities. For instance, P(X > 3) with n = 5 is P(X = 3) + P(X = 4) + P(X = 5)

Essential Calculator Usage: BinomCDF vs. BinomPDF (Ti84)



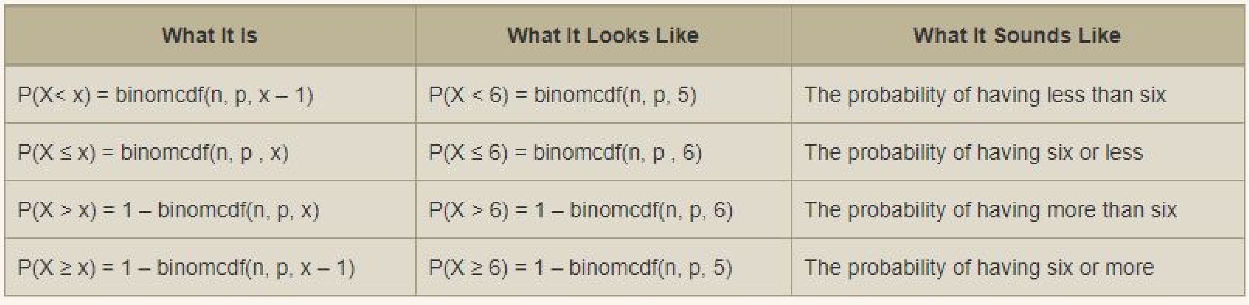
pdf	cdf
P(x = value) Use the pdf command:binompdf(n, p, x).	P(x ≤ value) Use the cdf command:binomcdf(n, p, x).
Sample Question: Find the probability of having five green pearls out of 100 oysters opened.	Sample Question: Find the probability of having no more than five green pearls out of 100 oysters opened.

Test tip: Explain all inputs and outputs in words and proper symbolization. Include the context of the question



Full credit	Partial or no credit
With n = 20, p = 0.12, and x = 4. I used binompdf(20, 0.12, 4) = 0.1299. This means the probability of the tennis player making exactly four serves out of 20 is 12.99%.	binompdf(20, 0.12, 4) = 0.1299

Translator:



What It Is	What It Looks Like	What It Sounds Like
P(X< x) = binomcdf(n, p, x – 1)	P(X < 6) = binomcdf(n, p, 5)	The probability of having less than six
P(X ≤ x) = binomcdf(n, p , x)	P(X ≤ 6) = binomcdf(n, p , 6)	The probability of having six or less
P(X > x) = 1 – binomcdf(n, p, x)	P(X > 6) = 1 – binomcdf(n, p, 6)	The probability of having more than six
P(X ≥ x) = 1 – binomcdf(n, p, x – 1)	P(X ≥ 6) = 1 – binomcdf(n, p, 5)	The probability of having six or more

HOW TO ANSWER FRQ:

- STATE: define variables
- PLAN:
  - State model binom(n,p)
  - Check conditions (BINS)
- DO: find probability using calculator functions above

4.6.2 Geometric Distribution

- Continue until success occurs
- Check BIFS
  - Binary
  - Independent
  - First success
  - Same probability of success
- Shape
  - Always right skewed
- Center
  - 1/probability of success

- Basically on average how many tries it will take
- Variability
    - Standard deviation is square root of 1-probability of success over probability of success
- To calculate probability that the number of tries equals x, you multiply the probability that he does not succeed the number of times that the try failed, and then you multiply it again by the probability that it succeeded
    - If the probability of success is .4, and you want to calculate the probability that it takes 4 tries, multiply .6 three times and then multiply it by .4
    - geometpdf(prob of success, number of tries)
- To calculate probability that the number of tries is less than or equal to x, you can add up all the individual probabilities
    - geometcdf(prob of success, number of tries that can be equal or less than)
- No combinations
- Conditions:
    - Binary - fail or success
    - Independence - 10% condition
    - Success? - prob is always the same
- Formulas
    - Mean: $\frac{1}{p}$
    - Standard deviation: $\frac{\sqrt{(1-p)}}{p}$

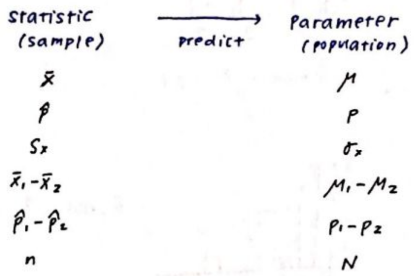
HOW TO ANSWER FRQ

- STATE: define variables, define what you want to find (e.g. P(x≤3))
- PLAN: conditions + state model geom(p,x)
- DO: using technology geometcdf(p,x) = __
- CONCLUDE: The probability that ___ is found within _____ is ______.

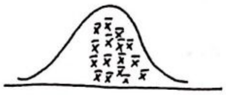
5. Sampling Distributions

5.1 What is a Sampling Distribution

- Sampling distribution of a statistic is the distribution of all the possible samples of a given size from a given population



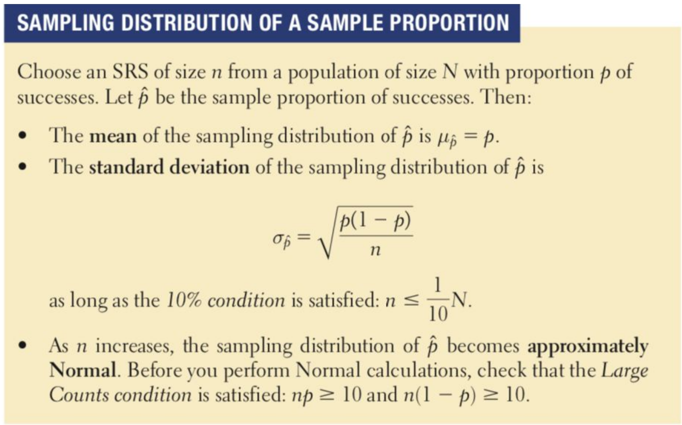
Statistic (sample)	→ predict	Parameter (population)
$\bar{x}$		μ
$\hat{p}$		p
S_x		σ_x
$\bar{x}_1-\bar{x}_2$		$\mu_1-\mu_2$
$\hat{p}_1-\hat{p}_2$		p_1-p_2
n		N



- Unbiased estimator: if on average the value of the estimator is equivalent to the given population parameter
- Biased estimator: systematically underestimating or overestimating the parameter.

5.2 Sample Proportions

- *Shape:* In some cases, the sampling distribution of can be approximated by a Normal curve. This seems to depend on both the sample size *n* and the popula- tion proportion *p*.
- *Center:* The mean of the distribution is $m_{p^\wedge} = p$. This makes sense because the sample proportion *p^* is an *unbiased estimator* of *p*.
- *Spread:* For a specific value of *p,* the standard deviation $s_{p^\wedge}$ gets smaller as *n* gets larger. The value of $s_{p^\wedge}$ depends on both *n* and *p*.
- To sort out the details of shape and spread, we need to make an important connection between the sample proportion $\widehat{p}$ and the number of "successes" *X* in the sample.



SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

Choose an SRS of size n from a population of size N with proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. Then:

- The **mean** of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$.
- The **standard deviation** of the sampling distribution of $\hat{p}$ is

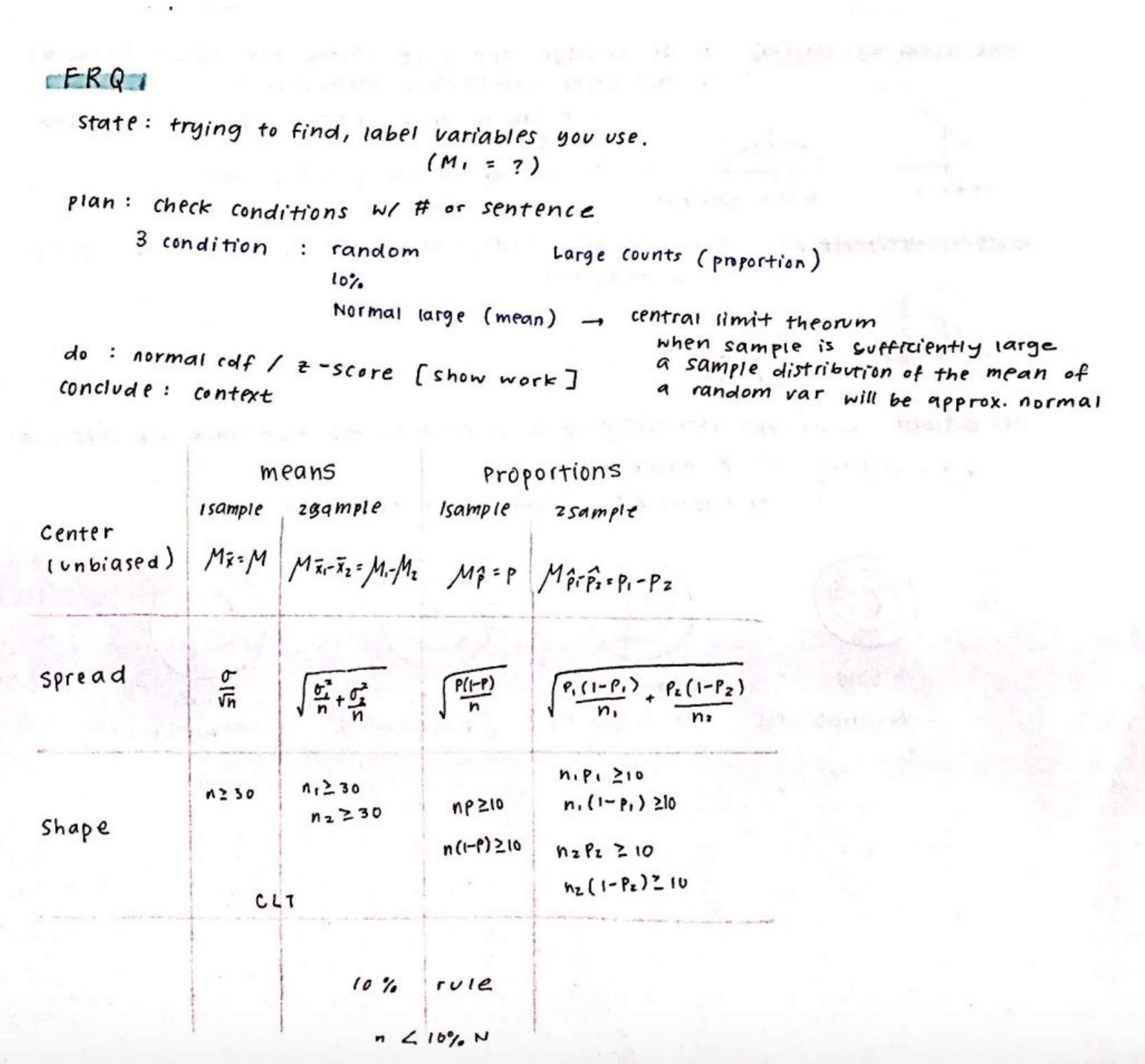
  $$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

  as long as the *10% condition* is satisfied: $n \leq \frac{1}{10}N$.
- As n increases, the sampling distribution of $\hat{p}$ becomes **approximately Normal**. Before you perform Normal calculations, check that the *Large Counts condition* is satisfied: $np \geq 10$ and $n(1-p) \geq 10$.

HOW TO SOLVE FRQ:

- STATE:
  - Point estimator
  - Point estimate
- PLAN:
  - If needed: check 10% condition to find standard deviation
- DO: whatever the problem asks for
- CONCLUDE: write results in sentence form

5.3 Sample Means



FRQ

state : trying to find, label variables you use.
($\mu_1 = ?$)

plan : check conditions w/ # or sentence

3 condition : random
10%
Normal large (mean) → central limit theorum

Large counts (proportion)

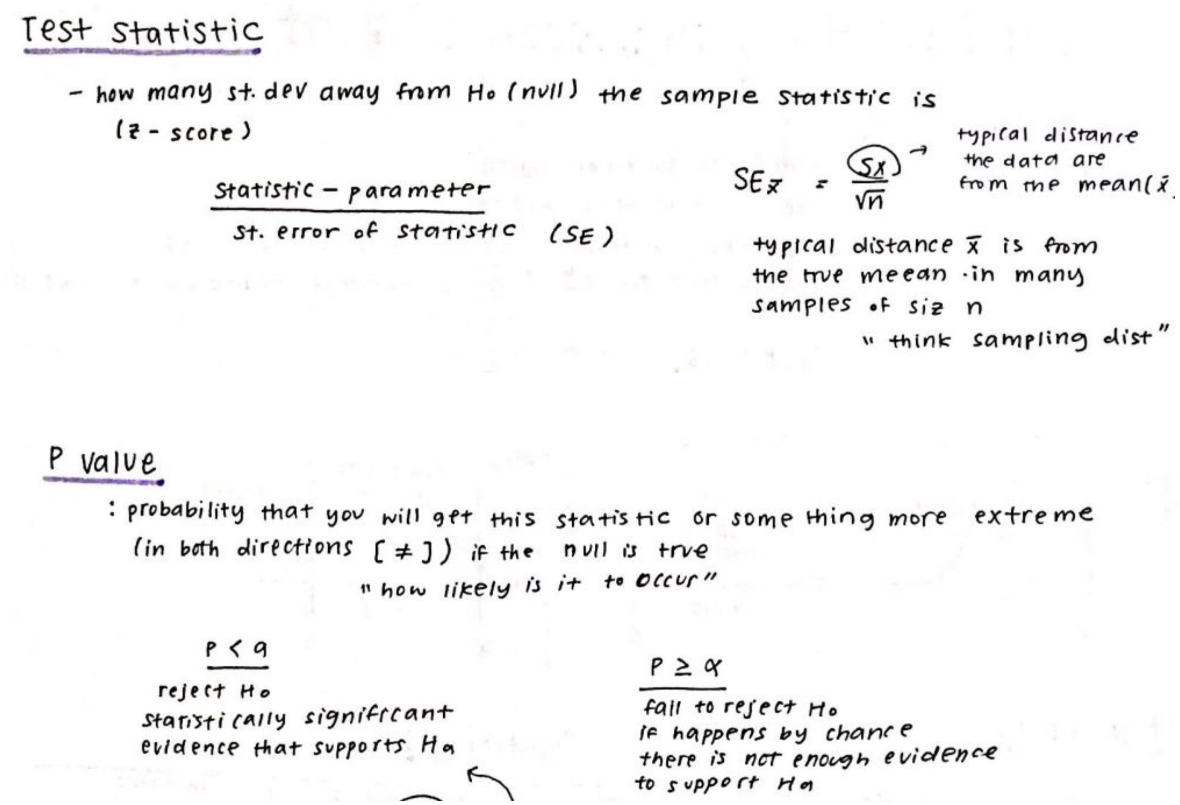
when sample is sufficiently large a sample distribution of the mean of a random var will be approx. normal

do : normal cdf / z-score [show work]

conclude : context

	means		Proportions	
	1sample	2sample	1sample	2sample
Center (unbiased)	$\mu_{\bar{x}}=\mu$	$\mu_{\bar{x}_1-\bar{x}_2}=\mu_1-\mu_2$	$\mu_{\hat{p}}=p$	$\mu_{\hat{p}_1-\hat{p}_2}=p_1-p_2$
Spread	$\frac{\sigma}{\sqrt{n}}$	$\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{n}}$	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$
Shape	$n\geq30$	$n_1\geq30$ $n_2\geq30$	$np\geq10$ $n(1-p)\geq10$	$n_1p_1\geq10$ $n_1(1-p_1)\geq10$ $n_2p_2\geq10$ $n_2(1-p_2)\geq10$
	CLT			
	10% rule $n < 10\% N$			

6. Inference for Categorical Data: Proportions



Test statistic

- how many st. dev away from Ho (null) the sample statistic is (z - score)

$$\frac{\text{statistic} - \text{parameter}}{\text{st. error of statistic}} \quad (SE)$$

$$SE_{\bar{x}} = \frac{\boxed{S_x}}{\sqrt{n}} \rightarrow$$

typical distance the data are from the mean($\bar{x}$.

typical distance $\bar{x}$ is from the true meean ·in many samples ·of siz n

" think sampling dist"

P value

: probability that you will get this statistic or some thing more extreme (in both directions [≠]) if the null is true

" how likely is it to occur"

$p < a$

reject Ho
statistically signifccant
evidence that supports Ha

$P \geq \alpha$

fail to reject Ho
if happens by chance
there is not enough evidence
to support Ha

6.1 Confidence Intervals for Proportions

An interval of values should be used to estimate parameters, in order to account for uncertainty.

Overview: construct a confidence interval for a proportion, p, as the statistic, p^, plus and minus a margin of error.

CONDITIONS TO BE MET FOR THESE PROCEDURES:

- Random? - random sample/random assignment
- Independent? - 10% rule
- Normal? - Large counts (np and nq both greater than or equal to 10) (for two sample remember to check both!)

Confidence INTERVAL	Confidence LEVEL "C"
A range of results from a poll, experiment, or survey that would be expected to contain the population parameter of interest.	The probability that if a poll/test/survey were repeated over and over again, the results obtained would be the same.

Key Words: Range, an interval, from this to that.	Key words: “Capture” the true…

6.1.1 1 Proportion

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

(1-p hat) is equivalent to q hat! This is the goated equation. Remember that with proportions we only use z intervals!

Margin of error:

$$ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$$

Table - Z-Scores for Commonly Used Confidence Intervals

Desired Confidence Interval	**Z Score**
90%	1.645
95%	1.96
99%	2.576

HOW TO ANSWER FRQ:

- STATE:
  - Define parameters and C-level
- PLAN:
  - Check conditions
  - State procedure that will be used
- DO:
  - Find z*: This can be done through your invNorm function on a TI-84! 2ND-VARS-3. Set the appropriate boundary and make sure the mean is zero and the SD is one!
  - You can also use 1-PropZInt on a TI-84! X is the number of successes rounded to a whole number, n is sample size, and c-level is confidence level.
  - TI-84: STAT → TEST → A: 1-PropZInt
  - Finding sample size for margin of error? USE 0.5 FOR P AND Q!!!
  - For a population proportion, the width of the interval is proportional to 1/sqrt(n)

- CONCLUDE:
    - “We are (c-level)% confident that the true proportion of (context) is between (lower bound, upper bound). ”

6.1.2 2 Proportion

$$(\hat{p}_1 - \hat{p}_2) \pm \text{margin of error}$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The procedure is practically the same as the one sample!

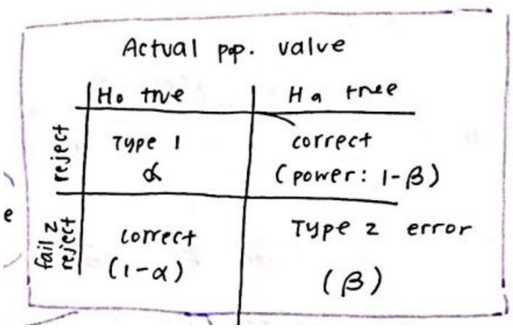
CONDITIONS TO BE MET FOR THESE PROCEDURES:

- Random? - random sample/random assignment
- Independent? - 10% rule
- Normal? - Large counts (n_1p,n_1q, n_2p, and n_2q are greater than or equal to 10)

HOW TO SOLVE FRQ

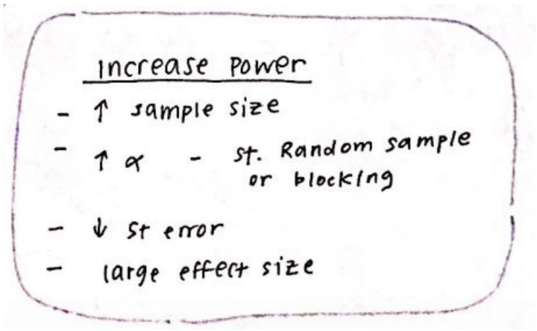
- STATE:
    - Define parameters and C-level
- PLAN:
    - Check conditions
    - State procedure that will be used
- DO:
    - 2-PropZInt on TI-84 can also solve this! STAT-TEST-B
    - Write out formula, but do calculations in calculator
    - TI-84: STAT → TEST → B: 2-PropZInt
    - x=number of successes, n=sample size, and c-level = confidence level.
- CONCLUDE: “We are (c-level)% confident that the true difference in proportion of (context) is between (lower bound, upper bound).”

6.2 Z-Tests for Proportions



Actual pp. value

	Ho true	Ha true
reject	Type 1 α	correct (power: 1-β)
fail to reject	correct (1-α)	Type 2 error (β)



Increase power

- ↑ sample size
- ↑ α - st. Random sample or blocking
- ↓ st error
- large effect size

Errors and its significance when dealing with Significance Tests.

- Type 1 Error: Reject H_0 when H_0 is true

Memory Cue: Hoes are bad girls. You reject a girl, thinking you have rejected a hoe. But actually, they are not hoes. They are your soulmate. You committed the primary and important error! It is a type ONE error! NOOO!

- Type 2 Error: Fail to reject H_0 when H_0 is not true

Memory Cue: You FAIL TO (2) like TWO reject the Ho, when it is false! Oh no!

CONDITIONS TO BE MET FOR THESE PROCEDURES:

- Random? - random sample/random assignment
- Independent? - 10% rule
- Normal? - Large counts (np and nq both greater than or equal to 10)

6.2.1 1 Proportion

HOW TO RESPOND ONE PROPORTION TEST FRQ

- STATE:
  - Population → optional
  - Parameter of interest (we want to find...) → optional
  - Ho
  - Ha + Context and defining variables
  - State α *level*
- PLAN:
  - Prove conditions
  - State procedure that will be used
- DO:

    - State z score
    - $$z = \frac{(\hat{p} - p_0)}{SD(\hat{p})}$$
    - $$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$$
    - State p-value
    - Calculate using normalcdf
- CONCLUDE:
- We fail to reject/reject H_0 as there is (not) convincing evidence that H_0 is false/true. + context

6.2.2 2 Proportion

HOW TO RESPOND TWO PROPORTION TEST FRQ

- STATE:
    - Population → optional
    - Parameter of interest (we want to find...) → optional
    - Ho
    - Ha + Context and defining variables
    - State α *level*
- PLAN:
    - Prove conditions
    - State procedure that will be used
- DO:
    - State z score
    - $$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
    - State p-value
    - Calculate using normalcdf
- CONCLUDE:
- We fail to reject/reject H_0 as there is (not) convincing evidence that H_0 is false/true. + context

7. Inference for Quantitative Data: Means

Conditions

- Randomization Condition: The data arise from a random sample or suitably randomized experiment. Randomly sampled data (particularly from an SRS) are ideal.

- 10% Condition: When a sample is drawn without replacement, the sample should be no more than 10% of the population.
- Nearly Normal Condition: The data come from a distribution that is unimodal and symmetric. Check this condition by making a histogram or Normal probability plot.

7.1 Confidence Intervals for Means

Now that we know how to create confidence intervals and test hypotheses about proportions, it'd be nice to be able to do the same for *means*. Just as we did before, we will base both our confidence interval and our hypothesis test on the sampling distribution model.

Confidence INTERVAL	Confidence LEVEL "C"
A range of results from a poll, experiment, or survey that would be expected to contain the population parameter of interest. Key Words: Range, an interval, from this to that.	The probability that if a poll/test/survey were repeated over and over again, the results obtained would be the same. Key words: "Capture" the true…

7.1.1 1 Sample

One sample t-interval for the mean.

FIRST, make sure the conditions are met. After they are, then we are ready to find the confidence interval for the population mean, μ.

The confidence interval is $\bar{y} \pm t^*_{n-1} \times SE(\bar{y})$

When the standard error of the mean is

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

The critical value depends on the particular confidence level, C, that you specify and on the number of degrees of freedom, n – 1, which we get from the sample size.

7.1.2 2 Sample

Two-Sample t-interval

When the conditions are met, we are ready to find the confidence interval for the difference between means of two independent groups, 1 – 2.

The confidence interval is

$$\left(\bar{y}_1 - \bar{y}_2\right) \pm t_{df}^{*} \times SE\left(\bar{y}_1 - \bar{y}_2\right)$$

where the standard error of the difference of the means is

$$SE\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The critical value t*df depends on the particular confidence level, C, that you specify and on the number of degrees of freedom, which we get from the sample sizes and a special formula.

7.2 T-Tests for Sample

Errors and its significance when dealing with Significance Tests.

- Type 1 Error: Reject H_0 when H_0 is true

Memory Cue: Hoes are bad girls. You reject a girl, thinking you have rejected a hoe. But actually, they are not hoes. They are your soulmate. You committed the primary and important error! It is a type ONE error! NOOO!

- Type 2 Error: Fail to reject H_0 when H_0 is not true

Memory Cue: You FAIL TO (2) like TWO reject the Ho, when it is false! Oh no!

Steps

1) STATE Parameter Of Interest (optional)
2) Define parameters (NOT OPTIONAL)
3) Write the appropriate hypotheses
4) Check conditions/assumptions
5) State procedure to be used (one/two sample t test)
6) Find mean and standard deviation
7) Find the p-value
8) Explain the p-value in context
9) State your conclusion

7.2.1 1 Sample

One sample t-test for the mean

STATE:

- H_0 (always =)
- H_A + context (≠, <, >)
- Define parameters
- State α *level*

PLAN:

- Check conditions
- State procedure to be used

DO:

- We test the hypothesis H_0: = 0 using the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})}$$

The standard error of the sample mean is $SE(\bar{y}) = \frac{s}{\sqrt{n}}$

When the conditions are met and the null hypothesis is true, this statistic follows a Student's t model with n – 1 df. We use that model to obtain a P-value.

- Ti84: STAT → TEST → 2: T-Test

CONCLUDE:

- "Since p value = __ [logic symbol] α =___, we fail to reject/reject H_0. There is not enough/enough statistically significant evidence that ________."

7.2.2 2 Sample

Two-sample T-test for Means

STATE:

- H_0 (always =)
- H_A + context
- Define parameters
- State α *level*

PLAN:

- Check conditions
- State procedure to be used

We test the hypothesis H0:1 – 2 = 0, where the hypothesized difference, 0, is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$$

The standard error is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

When the conditions are met and the null hypothesis is true, this statistic can be closely modeled by a Student’s t-model with a number of degrees of freedom given by a special formula. We use that model to obtain a P-value.

- Ti84: STAT → TEST → 4: 2-SampleTTest

CONCLUDE:

- “Since p value = __ [logic symbol] α =___, we fail to reject/reject H_0. There is not enough/enough statistically significant evidence that ________.”