fiveable

📊 AP Statistics – 2023 Cheat Sheet | 👉🏽 See all AP Stats study guides | @thinkfiveable

👆Unit 1
Exploring One-Variable Data

- **Categorical data** (not numerical) is shown in two-way tables & bar graphs, analyzing proportions
- **Quantitative data** is displayed in histograms, dotplots, box plots, stem and leaf plots, and scatterplots.
- Mean: non-resistant (affected by outliers)
- Median: resistant (affected by outliers)
- Unimodal = one clear peak, Bimodal = two clear peaks, Uniform = no clear peaks, flat
- Use **comparison** words when comparing distributions
- **For a histogram ->** make sure you approximate the mean (500-750 units) and use words like "no more" / "approximately" when describing range
- When analyzing distributions, always **CUSS** & **BS** in context - Center, Unusual features, Shape, Spread (remember skew pulls mean), BE SPECIFIC and always contextualize your answer.
- **Normal distribution:** mound-shaped and symmetric. Its parameters are mu (μ) for mean and sigma σ for standard deviation.
- Calculate z-score (value-mean / SD), measuring how many SD a value is from a mean
- The Standard Normal Distribution has a Mean of 0 and a SD of 1
- Empirical Rule: 68% of observations within 1 S

✌️Unit 2
Exploring Two-Variable Data

- For categorical data to be independent, conditional
- frequency = marginal frequency
- For quantitative data, always describe associations with
- direction, strength, form.
- DIrection - positive / negative (slope)
- Form - linear / non-linear
- r (correlation coefficient) measures strength & direction, NOT FORM
- Least Squares regression line (LSRL) predicts values of response variable (y) given explanatory variable (x)
- LSRL written as ŷ = a + bx
- ŷ = predicted value of rsp variable
- a = y-int, b = slope
- Residual = predicted - actual
- Look for random scatter on residual plot!
- Using LSRL to make predictions outside the interval of values of x used to make the equation of the line=extrapolation
- S & R-sq influenced by outliers (s ↑, r-sq ↓)

<u>**Key Interpretations: (For FRQ Writing)**</u>
- **Slope/b**: As the [exp var.] increases by 1 [unit], the [rsp var.] is predicted to increase by b [units].
- **Y-intercept**: When there are zero [exp var], the predicted [rsp var.] is y-int.
- **s**: When using LSRL to predict [rsp var] from [exp var] we are typically off by [value of s].
- **r^2 (in %)**: About [r-sq]% of variation in [rsp var] is explained by the LSRL using [exp var].
- **a residual**: The actual (rsp var) is about [residual] more/less than the predicted (rsp var).

🔎 Unit 3
Collecting Data

- **Simple Random Sample (SRS)**=every group of a certain size has an equal chance of being selected
- **Cluster Sample**=Divide pop. into heterogeneous groups [all from some]
- **Stratified Random Sample**=Divide pop. into strata of homogeneous groups [some from all]
- Why stratify by X, explain why indivs in those strata would have different rsps as opposed to some other variable
- Stratifying: ↓ variability, ↑ precision
- Bias types=**undercoverage, nonresponse, response bias** (inaccurate)-Always say if it leads to over/underestimate of a rsp
- EXPERIMENTS ASSIGN TREATMENTS
- Confounding-When a variable and the exp. variables are associated in a way that their effects on a rsp. Variable can't be distinguished from one another
- Experiments have comparison, random assignment (creates roughly equiv. groups of exp. units by balancing the effects of other variables among treatment groups), control (helps avoid confounding & ↓ variability in rsp var.), & replication (any diffs in effects of treatments can be distinguished from chance differences b/w groups)
- Randomized block design: random assignment of treatments is carried out separately in each block
- Blocks share a var that may impact rsp ↓ variability in rsp var, allows for easier comparison of treatments
- Matched pairs = compare 2 treatments in block size 2

🎲Unit 4
Probability, Random Variables, & Probability Distributions

- **Probability =** the chance of an event occurring, expressed in a decimal (0-1)
- P(event) = successful outcomes / total outcomes
- Complement of an event P(not event) is equal to 1 - P(event)
    - Common complements = at least, at most, greater/less than
- P(A and B = P(A∩B) = probability that BOTH events A and B occur
    - In a 2-way table = intersection of the two events divided by full total of the two events
    - Using condt'l probability = P(A) * P(B|A)
- P(A or B) = P(A U B) = probability that either events A or B occur
    - P(A) + P(B) - P(A and B)
- Conditional Probability = P(A|B) = P(A given B) = probability that event A occurs given B already happened
    - P(A|B) = P(A and B) / P(B)
- Events are mutually exclusive if P(A or B) = P(A) + P(B)
- **Events are indep if P(A|B) = P(A) OR if P(A and B) = P(A) * P(B)**
- Random variables are quantitative and take numerical values determined by the outcome of a chance event.
- Discrete = has a set number of values like number of coins that land heads or composite ACT score
    - P(X = event) must add up to 1
- **Expected Value (OR MEAN) of discrete random variable** is calculated using (x_1* p_1) + (x_2*p_2)..+(x_i*p_i)
- Continuous random var takes on all values in an interval of numbers, represented by normal distribution
- Binomial Random Variables: multiple trials of the same event
    - Conditions: **B**inary, **I**ndependent (10%), **N**umber of fixed trials, **S**ame probability of success p
    - Parameters: n (number of trials) & p (prob of success on any trial), x = successes
- Geometric Random Variables are STILL independent with fixed probability of success without trials set previously
    - Check Binary, Indep, same prob of success p

📊Unit 5
Sampling Distributions

Sample **s**tatistics will help estimate population parameters.
- **X-bar estimates mu, p-hat estimates p** & When describing parameters, include words like TRUE!
- **A sampling distribution is a distribution of values taken by a statistic in all possible samples of the same size from the same pop.** It shows how a <u>statistic</u> varies in many samples in a pop.
- Larger random samples produce estimates closer to the true pop. value - increasing the sample size will make an estimate more **precise** b/c it ↓ variab

<u>When describing a sampling distrib. of p-hat</u>, **it's approx normal if np >= 10 AND n(1-p) >= 10 (Large Counts Condition) → Large Counts ensures that sampling distribution is approx. normal** & helps us find z stat for other purposes (like finding a p-value). <u>Center & spread are found on the formula chart!</u> Make sure to check that the sampling/assignment was random & the 10% condition for indep when it isn't an exp.

<u>When describing a sampling distrib of x-bar</u>:
It's approx. normal if the population is normal or if there are at least 30 values in each sample (Central Limit Theorem).

This ensures that the sampling distrib. Is approx normal! Center & spread are found on the formula chart. Make sure to check that sampling/assignment was random & the 10% condition for indep when it isn't an experim.

⚖️Unit 6
Proportions

3 Major Conditions
- Normal/Large Counts Condition: np>10, n(1-p)>10
    - In a CI, replace p with p-hat
    - Unless it states Normal
- Independence = n(10%)<N
    - NOT FOR EXPERIMENT
- Random = Should be stated in the problem or STATE assumed **[Or Indep. Random Samples for 2-sample CI/test]**

A Conf. Interval = Point Estimate +/- Margin of Error
**Larger sample size decreases margin of error!
For a Confidence Interval of one sample - state as a **one-sample C% Z-Interval for p**, hypothesis test = **one sample z-test for p**
For a confidence interval of two samples - state as a **two-sample C% Z-Interval for p_1 - p_2**, hypothesis test = **two-sample z-test for p_1-p_2**
Note: there is NOT **paired data for proportions. For tests, state parameter(s) & hypotheses!**
For a <u>2-sided</u> alternative hypothesis (means or proportions), we can use a confidence interval to make a decision about Ho. If Ho isn't in the interval, reject Ho. If Ho is in the interval, fail to reject Ho.
[Conf level = opp of signif level] (95% CI = 0.05 sig)

<u>Key Interpretations:</u>
- **P-value:** Assuming the **[Ho in context]** is true, the probability that the [observed statistic - x-bar or p-hat] will take a value as or more extreme than it does is **[p-value].** {MAKE SURE TO USE CONTEXT}
- **Confidence level**: If we were to select many random samples of the same size # **[from the problem]** from the same population of **[problem]** and construct a C% confidence interval using each sample, about C% of the intervals would capture the **[parameter in context]**.
- **Confidence intervals:** We are C% confident that the interval from __ to ___ captures the [true parameter in context]

🐱 Unit 7
Means

3 Major Conditions
- Normal = CLT: n>30 OR SKETCH SAMPLING DISTRIBUTION
    - Unless it states the population is Normal
- Independence = n(1/10)<N
    - NOT FOR EXPERIMENT
- Random = Should be stated in the problem or STATE assumed **[Or Indep. Random Samples for 2-sample CI/test]**

For a confidence interval for one sample - state as a **one-sample C% t-interval for mu**, hypothesis test = **one-sample t-test for mu**

For a confidence interval of two samples - state as a **2-sample C% t-interval for mu_1 - mu_2**, hyp. test = **two-sample t-test for mu_1-mu_2**

Note: there **IS** paired data for mean an (experimental unit received 2 treatments). For paired data, label the parameter mu_diff as the **true mean difference of [context - context]**
CI = One-sample t C% CI for mu_diff
Hyp test = One-sample t test for mu_diff

<u>Type I Error</u> - Rejecting Null hypothesis (Ho) & finding convincing evidence for the alt. hyp. (Ha) **when we should've failed to reject Ho and not found convincing evidence for Ha**

<u>Type II error</u> - Failing to reject the Null hypothesis (Ho) & not finding convincing evidence for the alt. hyp. (Ha) when **we should've rejected Ho and found convincing evidence for Ha**

When interpreting Type I/II - replace Ho/Ha with the actual null/alternative hypothesis!

Power = 1 - P(Type II error) - probability we correctly reject Ho when the reality is that Ho is false <u>[correctly detecting a false Ho]</u>

To increase power, you can increase sample size [decreases standard error & dec. p-values], increase significance level [more likely to reject Ho], and increase the diff b/w Ho and true Ha.

✳️Unit 8 Chi-Squares / 🦢Unit 9
Slopes

<u>Unit 8 - Chi-Square Inference:</u>
1 Major Formula:

$$\chi 2 = \Sigma((observed - expected)^2/expected)$$

3 Tests: GoF, Independence, and Homogeneity

1 Type of Statistical Inference: Hypothesis Test
- Chi-Squared is a non-parametric test meaning we do not make assumptions
- Remember to calculate your df (n-1)

3 Major Conditions: random, indep., at least 5 success/fail

Remember to name the correct type of test

<u>Unit 9 - Inference for Slopes:</u>

LSRL Equation for Inference for Slopes: $\mu = a + \beta x$

5 Major Conditions: Linear, Indep, Normal, equal SD, Random

For a confidence interval - use this equation: b±t*(SEb)
For unit 9, df=n-2
For a hypothesis test - use this equation: (B-βo)/SEb

FRQ Tips: **Allocate your time wisely for answering open-ended questions / Keep in mind that your answers will be evaluated as a whole / Familiarize yourself with statistical terminology and use it accurately / Highlight the keywords in the questions by underlining or circling them / Identify and verify all underlying assumptions / Be proficient in creating graphs manually and analyzing data presented in different forms.**