AP Stats Notes

Last updated 06/09/2021. Calculator commands are exclusively for TI-84 Plus CE. General advice before we begin:

1. Always include context in your answers!

1 One-Variable Data

1.1 Definitions

Categorial Variable. Takes on values of category names or group labels. Described with counts (frequencies) or proportions (relative frequencies)

Can be represented with:

1. bar graph
2. dot plot
3. pie chart

Quantitative Variable. Takes on numerical values for measured/counted quantity.

Can be represented with:

1. dot plot
2. histogram
3. stem plot/stem and leaf display
4. cumulative frequency plot (*ogive*)
5. boxplots

Calculator Command 1: Histogram

1. STAT → EDIT: enter in the list of quantitative data
2. 2nd → STATPLOT: choose histogram (third from left), turn plot on
3. ZOOM → ZoomStat: adjust the window
4. GRAPH

Relative Frequency. Frequency divided by total number in population.

1.2 Describing Quantitative Variable Distributions

CUSS:

1. center (mean/median)
2. unusual features (e.g. clusters, gaps, outliers)
3. shape (unimodal, bimodal, uniform, and symmetric, skewed right/left)

4. spread (e.g. range, standard deviation)

When looking at cumulative frequency plots:

1. a distribution skewed left rises slowly first, then steeply later
2. a distribution skewed right rises steeply first, then slowly later

Centers:

Median. Middle value (for odd number), or average of two middle values (for even number of data points). Not affected much by outliers

Mean. $\sum x_i/N$. Not good for skewed distributions, is pulled toward direction of skew.

Spreads:

Range. $\max - \min$

Interquartile Range (IQR). $Q_3 - Q_1$, resistant to extreme values. Numerical rule for outliers: $x_i > Q_3 + 1.5 \cdot \text{IQR}$, $x_i < Q_1 - 1.5 \cdot \text{IQR}$

Variance. $\sum(x_i - \bar{x})^2/N$. For a sample, use $N - 1$.

Standard Deviation. $\sigma = \sqrt{\sum(x_i - \bar{x})^2/N}$. For a sample, use $N - 1$

Residual. $x - \bar{x}$

Designating Position:

Simple Ranking. which number data point in order

Percentile Ranking. what percentage of values fall at or below

z-score. $(x - \mu)/\sigma$

Other:

Boxplot/Box and Whisker display. Shows five number summary (min, Q_1, median, Q_3, max).

Normal Distribution. A bell shaped curve, with mean μ and standard deviation σ. Its point of inflection lie at $\mu \pm \sigma$. Around 68-95-99.7% of values lie within 1σ, 2σ, and 3σ respectively.

2 Two-Variable Data

2.1 Two Categorical Variables

Contingency table. Also called a two-way table, shows two categorical variables. Can be displayed in **segmented bar chart** or **mosaic plots**.

Marginal Frequencies. Found on the margins of contingency tables, they are the frequencies of variable 2, summed over all instances of variable 1.

Conditional Frequencies. Distribution of variable two, within a single category of variable one.

2.2 Two Quantitative Variables

To describe a scatterplot, say

1. form (linear or non-linear)
2. direction (positive or negative)
3. strength (weak, moderate, strong)
4. unusual features (outliers, clusters)

Correlation (coefficient). strength of linear relationship. y is response variable, x is explanatory variable. Correlation does not imply causation!!

$$\begin{aligned} r &= \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\ &= \sum\frac{z_x z_y}{n-1} \\ -1 &\le r \le 1 \end{aligned}$$

Coefficient of Determination. Written as R^2 and having a value of r^2. R^2 is the percentage of variation in y explained by the variation in x.

(Least Squares) Regression Line. Best fit line obtained by minimizing $\sum(y_i-\hat{y}_i)^2$ (hat indicates predicted value), passes through $(\bar{x},\bar{y})$ and has slope:

$$b = r\frac{s_y}{s_x}$$

Residuals (again). $\hat{e}_i = y_i - \hat{y}_i$. Plotting residuals with x can show nonlinear relationships or that a linear model is appropriate. Some relations:

$$\sum \hat{e}_i = 0$$

$$s_e = \sqrt{\frac{\sum \hat{e}_i^2}{n-2}}$$

Simpson's Paradox. Subgroups may show one trend, but combining them may show a different trend.

Example 1: Simpson's Paradox

Suppose Randy Johnson strikes out 25/50 lefties and 9/10 of righties, and you strike out 2/5 lefties and 8/10 righties. Randy Johnson seems to be better than you in both categories, but if you take the total, the Big Unit strikes out $34/60 \approx 57\%$ while you strike out $10/15 \approx 66\%$, making you think you're better

.

2.3 Outliers

(Regression) Outlier. A point that falls far away from regression line (large residual).

Influential Point. A point that changes the regression line a lot if removed. Most often has extreme values of x. May not have a large residual.

High leverage point. A point that has an extreme value of x. If it happens to lie close to the existing regression line, it could simply strengthen correlation and R^2. If it does not lie close to the existing regression line, it can also be an influential point.

Important note: if we plot x vs. y instead of y vs. x, we *do not* get the same line. rs_x/s_y is not the reciprocal of rs_y/s_x.

3 Collecting Data

To get conclusions about larger population *parameters* (μ, σ), must take representative samples and calculate *sample statistics* ($\bar{x}, s$). Must be randomized, must have appropriate sample size/treatment group size.

3.1 Types of Observational Studies

Observational Study. Gather information without disturbing population.

Retrospective studies. Examine existing data, look backward.

Prospective studies. Watch for outcomes, track individuals into the future.

Types of Bias (samples consistently do not represent population):

1. *Voluntary response surveys*: too much emphasis on strong opinions, undersample those who don't care as much.
2. *Convenience surveys*: choose individuals that are easy to reach, but typically misses certain groups.
3. *Undercoverage bias*: some groups in population left out
4. *Response bias* question itself leads to misleading results
5. *Nonresponse bias*: individuals chosen for sample aren't reached, hard to know which part of population is responding

Sampling Methods

1. *census*: collecting data from every individual in the population
2. *simple random sample (SRS)*: every possible sample has equal chance of being selected. Most often, this means without replacement
3. *stratified sampling*: divide into homogeneous groups called *strata*, pick random samples in each strata, combine. Have reduced sampling variability.
4. *Cluster sampling*: divide population into heterogeneous groups called *clusters*, pick one or more clusters. Typically more convenient.
5. *Systematic sampling*: pick a random point in some order (e.g. alphabetical), then sample every kth individual.

Sampling variability. variation in sample statistics across different samples of the same population. While bias is to be avoided, sampling variability is unavoidable and is to be quantified.

3.2 Experiments

Can show casual relationships, unlike observational studies. An experiment is performed on *experimental units*, called *subjects* if they are people. Experiments involve *experimental variables/factors* that have an effect on *response variables*. A group is treated with some *level* of the explanatory variable.

Example 2: Experiment

Lab rats are given either 0, 1, or 2 cheese cubes and go through either 0 or 1 hour of exercise per day. Their weight gain is measured after 7 days. In the jargon of experiments, this corresponds to:

1. units: lab rats
2. 2 factors: number of cheese cubes (3 levels), and amount of exercise (2 levels)
3. total of 6 treatments
4. response variable: weight gain

Control group. Collection of experimental units, either not given any treatment, given a current treatment, or given a treatment with inactive substance (placebo).

Blinding. Subjects don't know which treatment they receive. *Double blinding* is when neither subjects nor response evaluators know who is receiving which treatment.

Matched Pairs Design. Two treatments are compared based on paired subjects, one who receives one treatment and other receives other treatment. Often, really just subjects who are given both treatments, one at a time in a random order.

Blocking. Like stratification in observational studies, blocking divides population into representative *blocks* before randomly assigning to treatments.

An experiment should be *replicated* on a sufficient number of subjects, so that real response differences are apparent. A random sample must be taken in order to *generalize* to whole population.

4 Probability, Random Variables

4.1 Probability

Law of Large Numbers. When an experiment is performed a large number of times, the relative frequency of an event becomes closer to the true probability of the event.

Conditions:

1. The event should not change from trial to trial
2. Conclusion should be based on large number of observations

Probability. Likelihood that a particular event occurs. For equally likely events:

$$P(A) = \frac{\text{number of events in A}}{\text{total number of events}}$$

Some probability rules:

1. Complement:

$$P(A^C) = P(A') = 1 - P(A)$$

2. A or B ("general addition rule"):

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

$P(A \cap B) = 0$ if A and B are mutually exclusive.

3. A and B:

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

where $P(A|B)$ means the probability of A given B. For independent events, $P(A \cap B) = P(A)P(B)$. Note that this "general multiplication rule" can also be formulated as a definition for conditional probability:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

4.2 Random Variables

Expected Value (mean). For a random variable X,

$$\mu_X = \mathbb{E}(X) = \sum xp(x)$$

Variance:

$$\begin{aligned} \text{var}(X) = \sigma^2 &= \sum (x_i - \mu_x)^2 p_i \\ &= \langle x^2 \rangle - \langle x \rangle^2 \end{aligned}$$

Suppose you made a new random variable $W = X \pm Y$. Then:

$$\mu_W = \mu_X \pm \mu_Y \text{ (always, by linearity of expectation)}$$
$$\sigma_W^2 = \sigma_X^2 + \sigma_Y^2 \text{ (if X and Y are independent)}$$

Transforming random variables:

1. adding constant ($X \to X + c$) only changes mean
2. multiplying by constant ($X \to cX$) changes mean and standard deviation

4.3 Binomial Distribution

Two possible outcomes with constant probabilities, repeated multiple times. Success probability p, and failure probability $q = 1 - p$.

Probability of exactly k success in n trials:

$$P(k, n) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{(k!)(n-k)!} p^k q^{n-k}$$

Calculator Command 2: Binomial

For $P(X = k, n)$, 2nd → DISTR → binompdf(n, p, k), where X is the number of successes.

For $P(X \leq k, n)$: 2nd → DISTR → binomcdf(n, p, k)

For all probability calculations, must show:

1. distribution name (e.g. binomial)
2. parameters (e.g. $n = 5$, $p = 0.325$)
3. boundary and direction (e.g. $X \geq 1$)

For a binomial random variable X that is the number of successes in n trials of success probability p:

$$\mu_X = np$$
$$\sigma_X = \sqrt{npq}$$

4.4 Geometric Distribution

Probability of success p and probability of failure $q = 1 - p$, the probability of first success on trial $X = k$ is:

$$q^{k-1}p$$

Mean and variance for random variable X:

$$\mu_X = \frac{1}{p}$$
$$\sigma_X = \sqrt{\frac{q}{p^2}}$$

Calculator Command 3: Geometric

To find $P(X = k)$, go to 2nd → DISTR → geometpdf(p,k).

To find $P(X \leq k)$, go to 2nd → DISTR → geometcdf(p,k).

5 Sampling Distributions

5.1 Normal Distribution

Calculator Command 4: Normal Distribution

To find $P(\text{lower} < X < \text{upper})$, go to 2nd → DISTR → normalcdf(lower, upper, μ, σ).

If using z-scores to find probability, simply set $\mu = 0$ and $\sigma = 1$.

To find the limiting value (z-score) for a given probability, go to 2nd → DISTR → invNorm(area, μ, σ, tail). Note that the tail for older models is LEFT by default.

Approximating Binomial as normal:

1. If mean is far from edges ($\mu - 3\sigma \geq 0$ and $\mu + 3\sigma \leq n$), good for normal approximation.
2. Rule of thumb: $np, nq \geq 10$

In general, when checking if a distribution is close to normal, look to see if unimodal and roughly symmetric. An advanced plot: normal probability plot, which is the last type of graph in STATPLOT. Those closer to linear lines are more normal.

Central Limit Theorem: For a population with mean μ, standard deviation σ, and any distribution, take all samples of size n (the sampling distribution). If $n \geq 30$:

1. the distribution of sample means is approximately normal
2. mean of sample means is μ
3. standard deviation of sample means is $\sigma/\sqrt{n}$

In sampling distributions, *bias* means the sampling distribution is not centered on population parameter.

5.2 Sampling Distribution for Sample Proportions

Sampling distribution of sample proportion: starting with a population of given proportion p, and taking all samples of size n.

1. The set of all $\hat{p}$ is approximately normal if $np, nq \geq 10$ and $n < 10\%N$, where N is the size of the population.
2. $\mu_{\hat{p}} = p$
3. $\sigma_{\hat{p}} = \sqrt{pq/n}$

Sampling distribution of difference of sample proportions: starting with two populations of proportions p_1 and p_2. Take all samples of sizes n_1 and n_2 respectively.

1. The set of all $\hat{p_1} - \hat{p_2}$ is approximately normally distributed, if
   (a) $n_1p_1, n_1q_1 \geq 10$
   (b) $n_2p_2, n_2q_2 \geq 10$
   (c) $n_1 < 10\%N_1$, $n_2 < 10\%N_2$
   (d) each sample is independent
2. $\mu_{\hat{p_1}-\hat{p_2}} = p_1 - p_2$
3. $\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{p_1q_1/n_1 + p_2q_2/n_2}$

5.3 Sampling Distribution for Sample Means

Sampling distribution for sample means: starting with a population of mean μ and standard deviation σ, and taking all samples of size n:

1. the mean of sample means $\mu_{\bar{x}} = \mu$. Does not require $n \geq 30$.
2. the standard deviation of sample means $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Does not require $n \geq 30$.
3. If population is normal, or if $n \geq 30$, the sampling distribution for sample means is normal.

Sampling distribution for differences in means: starting with two populations of means μ_1 and μ_2, and standard deviations σ_1 and σ_2. Take all samples of sizes n_1 and n_2, respectively.

1. Mean of difference of sample means $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$
2. Standard deviation of differences of sample means is $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
3. If populations are normal, or if $n_1, n_2 \geq 30$, sample distribution for differences in means is normal.

6 Inference for Categorical Data

Confidence Interval. Estimate ± margin of error
Confidence Level. Proportion of times repeated applications of a method would capture true population parameter.

6.1 Conditions for Inference, One-Proportion

1. Independence Assumption: individuals in a sample/experiment are independent.
   (a) sample should be simple random sample or representative sample
   (b) if without replacement, $n < 0.1N$
2. Normality assumption: for proportions, binomial approximated as normal requires $np, nq \geq 10$. For means, called the *Normal/Large Sample* condition, requires $n \geq 30$ or the population distribution to be normal.

6.2 One-Proportion z-Interval

Estimate standard deviation with *standard error*:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence interval:

$$p = \hat{p} \pm z^* \times SE(\hat{p})$$

where $z^* \times SE(\hat{p})$ is the margin of error, and the critical value z^* is:

$$z^* = \left|\text{invNorm}\left(\frac{1 - \text{confidence}}{2}, \mu = 0, \sigma = 1, \text{LEFT}\right)\right|$$

Interpretation (in context): I am 95% confident that the proportion of all auto accidents that involve teenage drivers is between 12.7% and 18.6%.

Note: when asked minimum value of n necessary to be within a margin of error, use $\hat{p} = 0.5$, which gives the worst possible scenario for $SE(\hat{p})$.

6.3 One-Proportion z-Test

Null hypothesis H_0, Alternative Hypothesis H_A.

P-value. The probability of getting a sample statistic as extreme or more extreme, given the null hypothesis is true.

Type I error. Mistakenly rejecting a true null hypothesis (false positive). Probability given by α, significance level.

Type II error. Mistakenly failing to reject a false null hypothesis (false negative). Probability given by β, different value of β for each population parameter. The *power* of a test is $1 - \beta$.

For a null hypothesis $H_0 : p = p_0$, and an alternative hypothesis $H_A : p \neq p_0$ (two-sided) or $H_A : p > p_0 / H_A : p < p_0$ (one-sided):

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0 q_0}{n}}$$

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}}$$

$$\text{pval} = (2\times)\text{normalcdf}(z, \infty, \mu = 0, \sigma = 1)$$

$\times 2$ for p-value if 2-tailed.

Interpretation (in context): The high p-value, $0.0816 > 0.05$, indicates that these results could be reasonably explained by sampling error, so I fail to reject the null hypothesis. We do not have evidence that the true percentage of union members who support a strike is less than 75%.

6.4 Important Note

Due to time constraints, I will only be typing notes for things that are *not already on the AP Stats Inference Guides* from Mr. Iams.

6.5 Two-Proportion z-Test

When checking assumptions and conditions, use $n_1 p_{\text{pooled}} \geq 10$, etc., i.e. the expected counts, to check.

7 Other Stuff

7.1 Not In Inference Guides

When venturing into new territory, often simulations are provided in the problem. To estimate p-values, look at the proportion of simulated data points that fall within a certain range.

Points on *power*:

1. reducing α reduces power and increases β
2. power is a function of the correct population parameter value
3. $p - p_0$ is called the *effect*; a greater effect results in a greater power.
4. a larger sample size results in smaller α and β and larger power.

Confidence Interval or Hypothesis Test? If asking whether or not there is evidence $\rightarrow$ Hypothesis Test. If asking how much or how effective $\rightarrow$ Confidence Interval.