**AP Statistics Notes: CH P – Ch 9; Know how to use your calculator**! Name ________________________

**One Variable Data**

1) Read Chapter and Section Summaries – pages 28 – 29, 67, 99 – 100, 103, 130, 157, 161

2) Graphs of quantitative data (one variable) – dot plot, stem plot, histogram, boxplot; Interpret graphs using center, spread, shape, gaps, outliers. An outlier is any number $> Q_3 + 1.5(IQR)$ or $< Q_1 - 1.5(IQR)$

Graphs for categorical data (one variable) include pie charts, bar graphs.

3) Summarizing distributions:

- **Measuring center**: median or mean – Use 1-var stat L1 <u>or</u> 1-var stat L1,L2. Recall: $\bar{x} = \dfrac{\sum x_i}{n}$
- **Measuring spread**: range = maximum – minimum, inter-quartile range = $Q_3 - Q_1$, standard deviation of a sample is s, $s = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}$, variance is $s^2$.
- **Measuring position:** Quartiles (25% of data is $\leq Q_1$, etc.), $P_{90}$ represents the 90th percentile which has 90% of the data $\leq$ to it, standardized scores (z-scores)

$$z = \frac{x-\mu}{\sigma},\ z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}},\ z = \frac{\hat{p}-\mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

4) Effects of changing units on summary measures – Linear transformations of the form **Y= *a* + *b*X**

- $\mu_Y = a + b\mu_x$ and $\sigma_y^2 = b^2\sigma_x^2$. Note that adding a constant "*a* " has <u>no</u> effect on the variance.
- Shape of distribution is not changed when a linear transformation is performed.

5) Shape: Skewed left, skewed right, symmetric (special kind of symmetric is uniform). Mean is pulled towards the tail. Skewed right indicates that most of the data has smaller values but a few pieces of data are large. (Mean $>$ median).

6) Use the mean and standard deviation from the mean to describe sets of data that are roughly symmetric <u>and</u> have no outliers. Otherwise, use the 5 # summary which consists of the: min, $Q_1$, median, $Q_3$, max The boxplot!

7) Properties of the standard deviation, *s*:

- Degrees of freedom are n – 1.
- Measures the average spread about the mean.
- If *s* = 0 then there is no spread in the sample. Otherwise, $s > 0$.
- The more spread out the data is from the mean, the greater the value of *s*.

8) Strongly influenced by extreme observations: $\bar{x}, s, s^2, range, r, r^2\,\hat{y},$

9) **Density Curve** is always on or above the horizontal axis and the total area = 1. Area under the curve represents the proportions of observations. Used for continuous random variables (measurements) and also under certain conditions used to <u>approximate</u> the distribution of a discrete random variable. Density Curves: normal curve and uniform curve.

10) The **standard** normal has $\mu = 0,\ \sigma = 1,$ with axis labelled z. **All** normal curves follow the 68-95-99.7 Rule.

11) To get a probability (area) when we have a normal distribution use NormalCdf(LB,UB, mean, standard deviation). To get the value of z or x when you know it is a normal distribution use InvNormal(left area, $\mu, \sigma$)

**Two-variable Data – Quantitative**

1) Read chapter and section summaries – pages 192 – 193, 195 – 196, 229, 233 – 234, 243, 302 (omit Simpson's Paradox), 311, 314 – 315.

2) Scatterplot, explanatory variable (independent variable), response variable (dependent variable)

3) Association can be positive, negative, or no association.

4) **Correlation** is **r**: measures the strength <u>and</u> direction of a **linear** relationship.

- $r = \frac{1}{n-1}\sum\left[\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)\right]$ Correlation standardized x and y; z and r have **no** unit of measure.
- $-1 \le r \le 1$, and an **r** near zero suggests that there is no linear relationship. (There might be a curved rel.)
- **r** does not change when we use a linear transformation on x or y, interchanging x and y does not affect **r.**

5) Least squares regression line minimizes the sum of the squares of the vertical deviations from the data points to the line. $\hat{y} = b_0 + b_1 x;\ b_1 = r\frac{s_y}{s_x};\ b_0 = \overline{y} - b_1\overline{x}$. The point $(\overline{x}, \overline{y})$ is always on the LSR line. **Read computer printout**!

6) The error between the predicted y and the observed y components is the residual = observed y – predicted y. $(y - \hat{y})$

The mean of the residuals is 0, if we use the LSR line. Look at the pictures of the residuals on pages 216 – 218.

7) The standard deviation of the residuals, the standard error of the line, s, represents the average error we can expect when we use the regression line to predict y for a given x.

8) The coefficient of determination, $\mathbf{r}^2$, represents the <u>percent</u> of variation in the response variable that is explained by the linear relationship with the explanatory variable.

9) An outlier lies outside the overall pattern. An influential point is an outlier that, if removed, markedly changes the LSR line. An outlier in the x-direction often has a small residual because it has pulled the LSR line towards it.

10) Beware of extrapolation – using the LSR line for prediction outside the values of x of the data points.

11) Lurking variables – a type of confounding variable that has an important effect on the relationship among the variables studied but is not included among the variables studied.

12) Association $\neq$ Causation. Only a well designed experiment in which x is changed and confounding variables are kept under control can show a cause-effect relationship!

13) Beware: Correlations based on averages are usually too high when applied to individuals.

**Two-variable data: Categorical – Two way tables (rows X columns)**

14) Margins are the row totals and column totals

15) To describe relationships among categorical variables use **percents,** since the sample sizes are not usually equal!

16) Rules: A and B are **independent** if and only if $P(A) = P(A|B)$. $P(A|B) = \frac{P(A \cap B)}{P(B)};\ P(notA) = 1 - P(A)$

Another way to check for independence is to use $P(A \cap B) = P(A)P(B)$

17) Transforming to achieve linearity.

- If (x, y) transformed to (x, log y) achieves linearity then an exponential function is an appropriate model. $y = ab^x$
- If (x, y) transformed to (log x, log y) achieves linearity then a power function is an appropriate model. $y = ax^b$

## Chapter 5 Producing Data – Random Sampling Design and Experimental Design

1) Read Chapter and Section Summaries – pages 348 – 349, 373 – 374, 378 – 379

2) In an observational study we do not attempt to influence the responses, we observe individuals and measure a variable of interest. But, in an experiment we deliberately impose some treatment on the individuals in order to observe their responses.

3) Population vs. Sample; Parameter vs. Statistic $(\mu, \bar{x}, \sigma, s, \sigma^2, s^2, p, \hat{p})$

4) The design of a study is biased if it favors some part of the population over another or if it systematically favors certain outcomes. Ex: geographic bias, economic bias.

5) Random sampling designs should give us samples that are representative of the population:

- Simple Random Sample of size n consists of n individuals from the population chosen so that **every set** of n individuals has an equal chance to be the one selected. Use table of random digits to do the selection. Choosing an SRS: label, table, stopping rule, identify the sample.
- Stratified Random Sample – divide into groups of similar individuals and then choose an SRS from each group (called strata) to form the full sample. Reduces variation. Blocking in an experiment is like stratification.
- Systematic Random Sample – the first individual is randomly selected and then every nth one after that.

6) Sources of bias: interviewer, non-response, under-coverage, response, measurement, wording of a question.

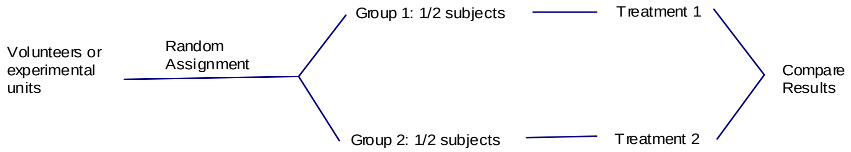
7) Convenience sampling and voluntary response sampling (respond to a general appeal) are poor sampling designs.

8) Because we use **chance** to create the sample, the results obey the laws of probability!

9) Characteristics of a well designed experiment: randomization of subjects to treatments, replication (repeat treatments on **many** units), comparison of 2 or more treatment, control.

10) Terms: factors, levels, treatments, placebo, placebo effect, control group, experimental units (subjects), blinding

11) Outline of a completely randomized experiment:



12) A second form of control is to restrict randomization by forming **blocks** of experimental units that are similar in some way that is important to the response (homogeneous blocks). Randomization of the treatments is carried out within each block. Matched pairs is a common form of blocking (right and left hand, right and left foot, taste-tests)

13) Blocking **reduces** variability and enables us to control the variables we can see.

14) Randomization allows us to equalize the effects of unknown or uncontrollable sources of variation. It does not eliminate the effects of these sources, but it spreads them out across treatment groups.

## Chapters 6 – 8: Probability and Random Variables

1) Read section and chapter summaries – Pages 403, 431, 452 – 453, 476 – 477, 500 – 501, 535 – 536, 551 (up to the mean only).

2) Law of Large Numbers: If we record the value of x each time we repeat a random phenomenon, and average these observed values, this average will get closer and closer to $\mu$ as we make more and more repetitions.

3) Rules of probability:

- Addition Rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. If A and B are mutually exclusive then $P(A \cup B) = P(A) + P(B)$.
- Multiplication Rule: $P(A \text{ and } B) = P(A)P(B|A)$. If A and B are independent then $P(A \cap B) = P(A)P(B)$.
- If A and B are independent then knowing that B has occurred does not change the probability of A.
- If A and B are mutually exclusive, then they are **not** independent, provided $P(A) \neq 0, P(B) \neq 0$
- $P(X \geq 1) = 1 - P(X = O)$ and $P(X \geq k) = 1 - P(X < k)$

4) Know how to draw probability tree diagrams, Venn diagrams, and two-way tables to help you answer a problem.

5) Discrete Random Variables – a count. Example: # of siblings

- Expected value of X = E(X) = $\mu = \sum \left( x_i p_i \right)$; Variance of X = $\sigma_x^2 = \sum \left[ \left( x_i - \mu \right)^2 p_i \right]$ or use 1-varstat L1,L2
- Probability distributions list all possible values of X with their corresponding probabilities. Sum of prob. = 1

6) Combining Random Variables X and Y.

- The following is always true: $\mu_{x+y} = \mu_x + \mu_y$ and $\mu_{x-y} = \mu_x - \mu_y$.
- If X and Y are **independent** then $Var(X \pm Y) = Var(X) + Var(Y)$.
- Combining our rules for RV and linear transformations: $Var(aX \pm bY + c) = a^2 Var(X) + b^2 Var(Y)$

7) Binomial Distribution: parameters are *n* and *p*.

- Binomial Setting: 1) fixed # of observations, *n* 2) observations are independent 3) two outcomes, *S* and *F* and 4) the probability of a success, *p*, is the same for all observations.
- X = the count of successes. X = 0,1,2,3…*k*,…*n*; Shape is symmetric ( $np \geq 10; n(1-p) \geq 10$ ), skewed right, or skewed left.
- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = binomialpdf(n, p, k)$
- $P(X \leq k) = binomialcdf(n, p, k)$.
- $\mu_x = np, \ \sigma_x = \sqrt{np(1-p)}$. If $np \geq 10$ *and* $n(1-p) \geq 10$, then we can use the normal approximation.

8) Geometric Distribution – Waiting Time: parameter is *p*.

- X = position of the first success. X = 1, 2, 3, 4…*k*…
- $P(X = k) = p(1-p)^{n-k}$; Always skewed right; $\mu = \dfrac{1}{p}$

**Chapter 9 – Sampling Distributions**

1) Read the summaries found on pages 578 – 579, 589, 602 – 603, 605 – 606

2) The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population. Look at its center, spread, shape, and outliers.

3) If the mean of the sampling distribution is **equal** to the true value of the parameter being estimated then the statistic is unbiased. $\mu_{\bar{x}} = \mu$, $\mu_{\hat{p}} = p$. Therefore, $\bar{x}$ and $\hat{p}$ are unbiased estimates for $\mu$ and p, respectively.

4) The spread of the sampling distribution is determined by the sampling design and the sample size. Larger samples give smaller spread. As long as the population is much larger than the sample size (at least 10 times larger), then the spread of the sampling distribution does not depend on the population size.

5) Choose an SRS of size n from a large population:

| **Sampling Distribution of sample proportions, $\hat{p}$** | **Sampling Distribution of sample means, $\bar{x}$** |
|---|---|
| $\hat{p} = \frac{x}{n} = \frac{\text{count in sample with characteristic}}{\text{size of the sample}}$ | $\bar{x}$, the sample mean, is the statistic |
| Center: $\mu_{\hat{p}} = p$ Always!! | Center: $\mu_{\bar{x}} = \mu$ Always!! |
| Spread: If pop size $\geq$ 10n, then $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ | Spread: If pop size $\geq$ 10n, then $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$,<br>Therefore, $\sigma_{\bar{x}} < \sigma$ |
| Shape: If $np \geq 10$ *and* $n(1-p) \geq 10$, then the sampling distribution is approximately normal. | Shape:<br>• If the population is normal then the sampling distribution is also normal.<br>• If the population is not normal:<br>Small sample sizes have a shape similar to the pop, but more “mound” shaped.<br>If sample size is large, $n \geq 30$, then the sampling distribution is approximately normal. |
| After you have verified the conditions:<br>$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ Probability statement: P( $\hat{p}$ ....) = P(z.…) | After you have verified the conditions:<br>$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ Probability statement: P( $\bar{x}$ ……) = P(z……) |
| Use the table or normalcdf to compute probabilities. | Use the table or normalcdf to compute probabilities. |

6) An observed effect so large that it would rarely occur by chance is called *statistically significant*. Flipping a coin 100 times an observing 80 heads is statistically significant.

7) To divide the standard deviation of the sampling distribution by *k*, we need to multiply the sample size by $k^2$.