YMS.4e CONTENT REVIEW
(Chapters 1-12)

I. Exploring Data

A. Analyzing Categorical Data

1. Categorical Data – nominal scale, names (e.g. male/female or eye color or breeds of dogs)

2. Displaying distributions
Bar graphs (bars do not touch)
Pie charts (percentages must sum to 100%)

3. From a two-way table of counts, find marginal and categorical distributions (in percents), describe relationship between two categorical variables by comparing percents and be able to recognize/explain Simpson's paradox

B. Displaying Quantitative (one variable univariate) Data with Graphs

1. Quantitative Data – rational scale, numbers where an average can be calculated (e.g. weights of hamsters or amounts of chemicals in beverages)

2. Displaying distributions
Dot plots – can resemble probability curves
Stem (& leaf) plots – remember to put in the key (e.g. 8|2 means 82 mg. of salt)
Split stems if too many data points
Back–to–back for comparison of two samples
Histogram – put // for breaks in axis, use no fewer than 5 classes (bars)

3. Describe distributions using SOCS (Shape, Outliers, Center and Spread)

C. Describing Quantitative Data with Numbers (**1-variable stats**)

1. Measures of central tendency (center)
mean ($\bar{x}, \mu$)
median (middle)
mode (most)

2. Measures of dispersion (spread)
range (max – min)
quartile (25% = Q_1, 75% = Q_3)
interquartile range ($Q_3 - Q_1$)
variance $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} or \sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$
standard deviation (s, σ) = square root of variance

3. Mean, range, variance, and standard deviation are non-resistant measures (strongly influenced by outliers). Use mean and standard deviation with approximately normal distributions; use median and IQR for skewed distributions (where the mean chases the tail).

4. 5-number summary (min, Q_1, M, Q_3, max) shown using *boxplots* (modified shows outliers)

II. Modeling Distributions of Data

A. Describing Location in a Distribution

1. *Ogive* – cumulative relative frequency plot

2. Adding a constant *a* to a data set increases mean by *a* but has no effect on standard deviation; multiplying a constant *b* to a data set multiplies the mean and standard deviation by *b*

3. Density curves are models which smooths out irregularities of actual data (use μ, σ) where area under curve always equals 1

B. Normal Distributions

1. If the mean = 0 and the standard deviation = 1, this is a standard, normal curve

2. Use with z scores (standard scores), z = $\frac{x-\mu}{\sigma}$, where +z are scores are above the mean and –z are scores below the mean.

3. To compare two observations from different circumstances, find the z score of each, then compare

4. Use z scores to find the *p* value, the probability (or proportion or percent) of the data that lies under a portion of the bell curve, *p* values represent area under the curve. Use **normalcdf** (to find the *p* value), or **invnorm** (to find z score) on the calculator

5. ALWAYS DRAW THE CURVE and shade to show your area.

6. 68% – 95% – 99.7% rule for area under the curve

7. To assess Normality for a given data set, graph the data, apply the 68-95-99.7 rule, compare the mean/median and construct a Normal Probability Plot

III. Describing Relationships

A. Scatterplots and Correlation

1. To graph two variable (bivariate) data – DATA MUST BE QUANTITATIVE. Graph the explanatory variable (independent) on the *x* axis, the response variable (dependent) on the *y* axis

2. Scatterplots look for relationships between the variables (linear, exponential or power)

3. Look for clusters of points and gaps. Two clusters indicate that the data should be analyzed to find reasons for the clusters.

4. If the points are scattered, draw an ellipse around the plot. The more elongated, the stronger the linear relationship. Sketch the major axis of the ellipse. This is a good model of the linear regression line.

5. Linear correlation coefficient (r) measures the strength of the linear relationship $-1 \le r \le 1$

   $r = 0$ indicates no relationship (the ellipse is a perfect circle)
   $-r$ indicates an inverse relationship
   r is a non–resistant measure (outliers strongly affect r)

   $$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$ (**2–variable stats**)

6. *Correlation does not imply causation.* Only a well-designed, controlled experiment may establish causation

B. Least-Squares Regression

1. Least squares regression line (LSRL) – used for prediction; minimizes the vertical distances from each data point to the line drawn. (**Linreg a+bx**)

   y varies with respect to *x*, so choose the explanatory and response axes carefully (*y* is dependent on *x*)

   $\hat{y}$ = predicted y value

   $\hat{y} = a + bx$ is the equation of the LSRL where $b = r\left(\frac{s_y}{s_x}\right)$, $a = \bar{y} - b\bar{x}$ and the point $(\bar{x}, \bar{y})$ is always on the line.

   Do not *extrapolate* (predict a *y* value when the *x* value is far from the other *x* values).

2. Coefficient of Determination (r^2) gives the proportion (%) of variation in the values of y that can be explained by the regression line. The better the line fits, the higher the value of r^2.

   To judge "fit of the line" look at *r* and r^2. If $r = 0.7$, then $r^2 = .49$, so about half the variation in y is accounted for by the least squares regression line.

3. Residual = observed *y* value – predicted *y* value ($y - \hat{y}$); residuals sum to 0

4. Residual plot – scatterplot of (*x*, residuals)
   no pattern $\rightarrow$ good linear relationship,
   curved pattern $\rightarrow$ no linear relationship,

5. Outliers are *y* values far from the regression line (have large residuals)

6. Influential points are *x* values far from the regression line (may have small residuals) which significantly change the LSRL slope

IV. Designing Studies

A. Sampling and Surveys

1. A census contacts every individual in the population to obtain data; a sample survey collects data from part of a population in order to learn something about the entire population.

2. Bad sampling designs result in bias in different forms

   voluntary response sample – participants choose themselves
   convenience sample – investigators choose to sample those people who are easy to reach

3. Good sampling designs

   simple random sample – a group of *n* individuals chosen from a population in such a way that every set of *n* individuals has an equal chance of being the sample actually chosen; use a random number table or **randint** on the calculator
   stratified random sample – divide the population into groups (strata) of similar individuals (by some chosen category) then choose a simple random sample from each of the groups
   cluster sampling- divide the population into groups (clusters); randomly select some of these clusters; all individuals in chosen clusters are included in the sample.

4. Sampling errors

   Biased sampling design- the design systematically favors certain outcomes or responses
   Undercoverage- when some groups of the population are left out, often because a complete list of the population from which the sample was chosen (sampling frame) was not accurate or available.

5. Nonsampling errors

nonresponse – when an individual appropriately chosen for the sample cannot or does not respond

response bias – when an individual does not answer a question truthfully, e.g. a question about previous drug use may not be answered accurately

wording of questions – questions are worded to elicit a particular response, e.g. One of the Ten Commandments states, "Thou shalt not kill." Do you favor the death penalty?

B. Experiments

1. An observational study observes individuals in a population or sample, measures variables of interest, but does not in any way assign treatments or influence responses

2. An experiment deliberately imposes some treatment on individuals (experimental units or subjects) in order to observe response. *Can* give evidence for causation *if* well designed with a control group. 3 necessities:

Control – for lurking variables by assigning units to groups that do not get the treatment

Lurking variables (variables not identified or considered) may explain a relationship between the explanatory and response variables by either confounding (a third variable affects the response variable only) or by common response (a third variable affects both the explanatory and response variables

Randomize – use simple random sampling to assign units to treatments/control groups

Replicate – use the same treatment on many units to reduce the variation due to chance

3. The "best" experiments are double blind – neither the investigators nor the subjects know which treatments are being used on which subjects. Placebos are often used.

4. Designs

Between groups (independent samples)- sometimes uses blocking where subjects are grouped before the experiment based on a particular characteristic or set of characteristics, then simple random samples are taken within each block.

Within groups (repeated measures)

Matched pairs

C. Using Studies Wisely

1. Inference about the population requires that the individuals in a study be randomly selected

2. Correlational studies *can* provide evidence of causation but it’s tricky

3. Do not automatically accept a study is true without analysis

V. Probability: What Are the Chances?

A. Randomness, Probability and Simulation

1. Probability only refers to "the long run" (law of large numbers) never short run
2. A probability is a number between 0 and 1
3. Simulations can be used to determine probabilities

B. Probability Rules

1. All probabilities for one event must sum to 1
2. $P(A^C) = 1 - P(A)$ where A^C is the complement of A
3. *Mutually Exclusive* (*disjoint*) *Events*– events which cannot occur at the same time; mutually exclusive events ALWAYS have an effect on each other so they can never be independent.
4. If $P(A + B) = 0$ then A and B are mutually exclusive and:

$$P(A \text{ or } B) = P(A) + P(B) \longrightarrow P(A \cup B) = P(A) + P(B)$$

5. For events that are *not* mutually exclusive:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \longrightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

6. Venn diagrams can be used to find probabilities

C. Conditional Probability and Independence

1. *Independent Events* – the probability of one event does not change (have an effect on) the probability of another event
2. If *A* and *B* are independent then $P(A \text{ and } B) = P(A) \cdot P(B) \longrightarrow P(A \cap B) = P(A) \cdot P(B)$
3. To prove that 2 events *A* and *B* are independent, show $P(A \text{ and } B) = P(A) \cdot P(B)$ *or* $P(B|A) = P(B)$
4. For events that are *not* independent:

$$P(A \text{ and } B) = P(A) \cdot P(B|A) \longrightarrow P(A \cup B) = P(A) \cdot P(B|A)$$

5. Conditional probability formula (use when working with probabilities):

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

VI. Random Variables

A. Discrete and Continuous Random Variables

1. X = variable whose value is a probability (discrete or continuous)

2. To get the *expected value* or *mean* of a discrete random variable, multiply the number of items by the probability assigned to each item (usually given in a probability distribution table), then sum those products, $\mu = \sum x_i p_i$

3. To get the variance of a discrete random variable, use $\sigma^2 = \sum (x_i - \mu)^2 p_i$ where p is the probability assigned to each item, x.

B. Transforming and Combining Random Variables

1. If $Y = a + bX$ then $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b| \sigma_X$

2. To find the sum or difference ($\pm$) using two random variables, add or subtract the means to get the mean of the sum or difference of the variables, $\mu_{x \pm y} = \mu_x \pm \mu_y$

3. To get the standard deviation ($\pm$) using two random variables, **always add** the *variances* then take the square root of the sum, $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$

C. Binomial and Geometric Variables

1. Conditions for binomial distribution:
   Bi- 2 outcomes (success or failure)
   Nom- Number of observations fixed
   I- Observations independent
   Al- Probability of success is always the same

2. Binomial probability of observing k success in n trials (**binompdf** or **binomcdf**):

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

3. The mean of a binomial distribution is $\mu = np$ where p is the probability and n is the number of observations in the sample.

4. The standard deviation of the binomial distribution is $\sigma = \sqrt{np(1-p)}$

5. The graph of a binomial distribution is strongly right skewed (has a long right tail) unless $n(p) \geq 10$ *and* $n(1-p) \geq 10$ then the distribution becomes approximately normal.

6. Conditions for geometric distribution are the same as for the binomial except there is not a fixed number of observations because the task is to find out how many times it takes before a success occurs. This is sometimes called a waiting time distribution.

7. The mean of the geometric distribution is $\mu = \frac{1}{p}$

8. The standard deviation of the geometric distribution is $\sigma = \sqrt{\frac{1-p}{p^2}}$

9. The graph of the geometric distribution is strongly right skewed always

10. Geometric probability = $P(Y = k) = (1-p)^{k-1} p$ (**geometpdf** or **geometcdf**)

VII. Sampling Distributions

A. What is a Sampling Distribution

1. Sampling distributions are probability distributions which are made from the *statistic* of several different samples; larger sample sizes yield less variability (spread)

2. Look for bias and/or variability. Bias is when the statistic is not close to the population parameter; variability is when the sample statistics are widely scattered. Remember the targets on page 426.

B. Sample Proportions

1. Used when working with a percentage value or proportion (e.g. 38% of the population thinks frogs are icky. What is the probability that, in a sample of 120 people, only 29% thinks the same?)

2. The mean of the sampling distribution equals the mean of the population, $\mu_{\hat{p}} = p$

3. The standard deviation of the sampling distribution $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ if $N > 10n$

4. If $np > 10$ and $n(1-p) > 10$ the sampling distribution is approximately normal

C. Sample Means

1. Used when working with the $\bar{x}$'s of samples.

2. The mean of the sampling distribution of the means equals the mean of the population

3. The standard deviation/standard error of the sampling distribution equals $\frac{\sigma}{\sqrt{n}}$ if $N > 10n$

4. Central limit theorem – when the sample size is large ($n \geq 30$) the graph of the sampling distribution of the sample means is approximately normal

VIII. Estimating with Confidence

A. Confidence Intervals: The Basics

1. Confidence intervals = statistic ± (critical value) · (margin of error)

2. Confidence intervals – "I am ___% confident that the population mean lies between the upper and lower levels of the confidence interval" Remember it either does, or does not! Review the picture on page 474 – if many samples are taken, the means of ___% of them should lie within the confidence interval. Confidence level means ___%, confidence interval gives the lower and upper upper bounds.

3. Always verify that the data came from a **random** sample from the population of interest, that the sampling distribution is approximately **normal** and the observations are **independent**.

4. Confidence intervals are *always* two–tailed

B. Estimating a Population Proportion

1. Remember sampling distribution of $\hat{p}$ is approximately normal when $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

2. Confidence interval for a population proportion $= \hat{p} \pm z^* \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$

3. The sample size needed to obtain a confidence level with approximate margin of error (MOE) for a population proportion involves solving: $z^* \left(\frac{\hat{p}(1-\hat{p})}{n} \right) \leq$ MOE where $\hat{p}$ =.50

C. Estimating a Population Mean (Using *t* distribution)

1. If the sample size is very small (<15) use this test only when the sample is normal (be sure to show normality). If the sample size is between 15 and 30, omit outliers, show normality, and use this test. If the sample size is >30, simply state that the test can be used because of a large sample size (Central Limit Theorem).

   Note: The *t* procedures are relatively robust when the population is non-Normal but not robust against outliers.

2. Confidence interval for a population mean = $\bar{x} \pm t^* \left(\frac{s}{\sqrt{n}} \right)$ with $n - 1$ degrees of freedom

3. To decrease the margin of error, increase the sample size or decrease the confidence interval.

IX. Testing a Claim

A. Significance Test: The Basics

1. Require a null hypothesis containing μ or p and = and an alternative hypothesis containing μ or p and either $\neq, <$, or $>$; e.g. H_o: $\mu = 0$, H_a: $\mu < 0$. Some hypotheses are words only.

2. Determine whether the test is two tailed (H_a: $\mu \neq 0$) or one tailed (H_a: $\mu < 0$ or H_a: $\mu > 0$). Choose a rejection α level, e.g. $\alpha = .05$. Reject the null hypothesis if the test statistic falls in the tail or if the *P*-value is smaller than a chosen α level. Fail to reject H_o if the test statistic is not in the tail or if the *P*-value is larger than a chosen α level.

3. Choose the test carefully and always state the conditions of the test before beginning to test. Memorize conditions for each test! Always refer to the problem in context when addressing conditions, such as random sample. Write the conclusion based of your test (reject or fail to reject the null hypothesis) in context of the problem situation.

4. Type I error = α level, the probability the null hypothesis will be rejected when it is true

5. Type II error = β level, the probability the null hypothesis will be accepted when it is false.

6. The power of any test is $1 - \beta$, the probability of rejecting the null hypothesis when it is false (making the correct decision to reject). The power increases when the α level is larger, therefore a test with $\alpha = .05$ is more powerful than a test with $\alpha = .01$ because an investigator is more likely to reject the null hypothesis. That means that the power increases when the confidence interval gets smaller, since a larger α level means a smaller confidence interval. The power also increases when the sample size is increased because the sample more accurately reflects the population.

B. Tests about a Population Proportion (One –Sample *z* Test)

1. Conditions

   a. Random sample/assignment
   b. N > 10*n*
   c. Normal sampling distribution: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$

2. If the population proportion is given, use $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$ }Use **1-PropZTest** on calculator

3. Confidence intervals provide additional information that significance tests do not (a range of plausible values for the true population proportion *p*)

   a. A two-sided test at α = .05 gives roughly the same conclusion as a 95% confidence interval
   b. A one-sided test at α = .05 gives roughly the same conclusion as a 90% confidence interval

C. Tests about a Population Mean (One-Sample *t* Test)

1. Conditions

   a. Random sample/assignment
   b. N > 10*n*
   c. Normal sampling distribution

2. Simple *t* Test

   Use $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ with $n - 1$ degrees of freedom } Use **T-Test** on calculator

3. Dependent or matched pairs *t* test

   a. Conditions – same as *t* above but the same subjects in the sample receive two different treatments.

   b. First find the differences (subtract) between the two treatments for each subject, then find the mean and standard deviation of the differences. The null hypothesis is always that the differences equal 0.

   c. Use $t = \dfrac{\bar{x} - 0}{\dfrac{s}{\sqrt{n}}}$ with $n - 1$ degrees of freedom } Use **T-Test** on calculator

4. Use significance tests wisely

   a. Statistical significance is not the same as practical significance
   b. Reinforce results with appropriate confidence intervals
   c. Lack of statistical significance does not mean that H_0 is true
   d. Tests are only valid if conditions are met

X. Comparing Two Populations or Groups

A. Comparing Two Proportions

1. Conditions (for both samples)

   a. Random samples/assignment
   b. Normal: $n_1\hat{p}_1 > 10,\ n_1(1-\hat{p}_1) > 10,\ n_2\hat{p}_2 > 10,\ n_2(1-\hat{p}_2) > 10$
   c. Independent- the two populations are at least 10 times as large as their corresponding samples

2. Confidence Intervals

$$CI = (\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$ } Use **2-PropZInt** on calculator

3. Significance Tests

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$ } Use **2-PropZTest** on calculator

$$\hat{p} = \frac{\textit{total number of successes in both samples}}{\textit{total number of observations in both samples}}$$

B. Comparing Two Means (from 2 *independent* samples)

1. Conditions (for both samples)

   a. Random samples/assignment
   b. Normal
   c. Independent

2. Confidence Intervals

$$CI = (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$ } Use **2-SampTInt** on calculator

3. Significance Tests

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

4. Use smaller of $n_1 - 1$ or $n_2 - 1$ for degrees of freedom

XI. Inference for Distributions of Categorical Data

A. Chi-Square Goodness of Fit Tests

1. Tests the null hypothesis that a categorical variable has a specified distribution; hypotheses written in words

2. Conditions

   a. Random
   b. All expected counts ≥ 5
   c. Independent

3. Chi-square statistic

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$ with degrees of freedom = number of categories – 1

B. Inference for Relationships

1. Chi-square tests for homogeneity (different samples) or association/independence (same sample) analyzes two variable categorical data with data in a matrix. Elements in each cell are frequencies – the number of individual (objects) with those particular characteristics

2. Conditions same as above

3. Expected count in any cell = $\frac{(row\ total)(column\ total)}{table\ total}$

4. Chi-square statistic

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$ with *df* = (number of rows – 1)(number of columns – 1)

With a matrix of observed values, put matrix [A] in your calculator. Use χ^2-Test on calculator. The expected counts will appear in matrix [B]

XII. More about Regression

A. Inference for Linear Regression

1. Conditions

   a. Random sample/assignment
   b. Normal- response variable *y* varies according to a normal distribution
   c. Independent observations
   b. No pattern in residual plot

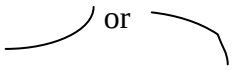
2. Significance Tests

$t = \frac{b - \beta_0}{SE_b}$ with $n - 2$ degrees of freedom } Use **LinRegTTest** on calculator

3. Confidence Intervals (for true population slope)

$CI = b \pm t^* SE_b$ } Use **LinRegTInt** on calculator

B. Transforming to Achieve Linearity

1. If there is a pattern in the residual plot when analyzing a linear relationship between *x* and *y* or if the data appears curved in the shape of an exponential function or a power function,
 or or  then test for these relationships

2. To test for exponential regression, analyze *x* versus log *y* and transform equation if there is no pattern in the residual plot:

   If $\log \hat{y} = 4.6 + 2.3x$ then $\hat{y} = 10^{4.6} \cdot (10^{2.3})^{x}$

3. To test for power regression, analyze log *x* versus log *y* and transform equation if there is no pattern in the residual plot:

   If $\log \hat{y} = 4.6 + 2.3(\log x)$ then $\hat{y} = 10^{4.6} \cdot x^{2.3}$