I. Conceptual/Theoretical Material

Controlled, randomized, experiments (Chapter 1)
Vocabulary: e.g. randomization, placebo, bias
Observational Studies (Chapter 2)

Ideas: There are issues of objectivity & bias. Cost (time and money). Ethics. Advantages/Disadvantages of each kind of study. Review pages 10-11, 27-28

Sample Surveys (Chapter 19)
Vocabulary: bias, simple random sample. A type of observational study.

II. Graphical Summary (Chapter 3.1-3.3)

Histogram
- Know how to read one
- Know how to recognize a good one vs. a bad one. (e.g. total area should equal 100% or 1.0)

III. Descriptive Statistics (Chapter 4)
- Average and Median -- calculating these statistics from a list, but also trying to get you to see averages and medians in terms of areas (to prepare you for Chapter 5)
- Histogram and the Average and the Median (pp 63-65)
- Standard deviation (4.5) -- calculating these statistics from a list
- Chapter review (pp. 76-77)

IV. The Normal (Chapter 5)

- Z score -- know the formula
- Percentiles -- what does it mean to be at the 90^{th} percentile? The 75^{th} percentile?
- Chapter 5.5 is the most important -- reading areas
- ignore 5.6 on transformations
- review on page 96

V. Probability (Chapter 13 and 14)

Laws:
- 0% <= probability <= 100%
- probability of A = 100% - probability of NOT A
- definition of chance (p. 222)
- sampling w/ and w/o replacement (p. 225)
- no conditional probability on final
- multiplication rule -- p. 229, the chance that two things WILL BOTH happen equals the chance of the first multiplied by the chance the second one will happen given that the first has already happened. Also,

if events are independent, multiply the unconditional probabilities. (p. 232 top). Look for the word AND.
- addition rule (ask yourself if the events are mutually exclusive) p. 241. Look for the word OR.
- summary: know when to add or multiply (pages 243-245)

VI. The Expected Value and Standard Error (Chapter 17) for a SUM/NUMBER/COUNT

a. Concept of the Expected Value -- also known as the mean of a random variable or I like to think of it as the "most likely outcome" with the understanding that in the "short run" or when you are sampling, almost any outcome is possible.

FORMULA : (number of draws) * (average of the box) (page 289)

b. Concept of the Standard Error. Any outcome is a function of the expected value + some chance error. The standard error gives you an estimate of how big the chance error is likely to be Chance error and standard error should be similar in size.

FORMULA : (square root of number of draws) * (SD of the box) (page 291)

Important idea: WHEN YOU HAVE A SAMPLE, YOU ARE IN A SITUATION WHERE YOU WILL NEED TO FIND AN S.E. On page 291, Freedman tells you that the S.E. is for a chance process. On page 292 he goes on to say that your observed values are rarely larger than 2 or 3 S.E. away from the expected value.

c. The E.V., the S.E. and the Normal (17.3)

This is the most important section of Chapter 17. The idea behind statistics is that if we were to sample and re-sample and then plot our results, we would see a normal looking curve (top of page 295). This happens because the "truth" or the true parameter value is at the center (the expected value) and we're going to get estimates all around it, but centered on this value.

The remainder of Chapter 17 just has different methods for different situations (that is, different kinds of boxes).

BE SURE THAT YOU KNOW HOW TO CALCULATE SIMPLE AVERAGES AND BOX AVERAGES FOR THE FINAL.

VII. The Expected Value for a Percentage (Chapter 20)

This is a special case of Chapter 17. Suppose you have a box with only 1's and 0's in it. The proportion or percentage of 1's is the population percentage and it is also the expected value for the sample percentage (that is, the outcome of our samples).

The standard error of a percentage will be:

$$\frac{\sqrt{n} * \sqrt{\text{proportion of 1's} * \text{proportion of 0's}}}{n} * 100$$

where n= sample size

And again, in Chapter 20.3, the percentages will distributed normally, with the center at the expected value and the spread measured by the S.E. of the percentage.

You are expected to know how to calculate the expected value (just know the percentage of 1's), the standard error and work with the normal curve and an observed outcome, specifically:

$$\frac{sample\% - \exp ected\%}{S.E.percentage} = Z$$

VIII. The Confidence Interval (Chapter 21)

The confidence interval is the plus-or-minus figure usually reported in newspaper or television opinion poll results. For example, 47% percent of your sample picks an answer with range of +4 or -4 around it, you can be "confident" that if you had asked the question of the entire relevant population the interval 43% (47-4) to 51% (47+4) would contain or cover the true percentage who would have picked the answer if you had asked them

.

The confidence level tells you how sure you can be. It is expressed as a percentage and represents what percentage of all possible intervals calculated by this method would cover or contain the "truth". The 95% confidence level means you can be 95% certain; the 99% confidence level means you can be 99% certain. Your hope is that the one interval you have is one of the "good" ones and the level of confidence just tells the reader how certain you are.

To calculate the confidence interval for a percentage:

$$\text{sample } \% \pm Z * (\text{Standard Error of the Percentage})$$

The Z will depend on your level of confidence. To be 95% confident, use a Z=2, to be 99% confident use a Z=3 (2.58 to be exact, but on this test, I will accept 3, it's close enough). The formula for the standard error of the percentage is given in the previous section.

Example: A telephone company serves 600,000 households. From these they take a random sample of 1000 households and get the following data. The percentage of households in the sample with caller ID is 53.2% and the percentage with speed dial is 26.7%.

Calculate a 99% confidence interval for the percentage of households with caller ID in the population.

$$53.2\% \pm 3 * (\frac{\sqrt{1000} * \sqrt{.532 * .468}}{1000} * 100) = 53.2\% \pm 3 * (1.577) = 53.2\% \pm 4.73$$

IX. The Expected Value and Confidence Interval for Averages (Chapter 23)

All we want you to know in Chapter 23 is that the sample average can vary from sample to sample. This chapter is just an extension of Chapter 17 and Chapter 20, but for averages.

The box model really has no tickets, instead, just understand that the population average (if known) is the same as the box average. If the population average is unknown, this means the box average should be substituted with the sample average.

The expected value for the average = box average = population average.

The standard error for the average is:

$$\frac{\sqrt{n} * \text{Standard Deviation of the Box}}{n}$$

where n = sample size. The box standard deviation is always given.

So a question might be like question #6 in the "problem solving" part of your practice final.

And just like Chapter 21, you can calculate the confidence interval for the population average if all you have is a sample average. Example:

A telephone company serves 600,000 households. From these they take a random sample of 1000 households and get the following data. In the sample the average number of telephones is 3.10 per household with an SD of 1.10.

Calculate a 95% confidence interval for the <u>average</u> number of telephones per household in the population.

$$3.1 \pm 2 * (\frac{\sqrt{1000} * 1.10}{1000}) = 3.1 \pm 2 * (.0348) = 3.1 \pm .0695$$

X. Hypothesis Testing (Chapter 26.1-26.4)

Statistical inference is that branch of statistics in which one typically makes a statement about a population based upon the results of a sample. The researcher conducts a test of hypotheses. This test uses the sample data to decide between two competing hypotheses or claims about a population.

(a) The Hypotheses

A statistical hypothesis is a statement about one or more parameters. For example, the average amount of time people need to sleep is 7.1 hours per day.

There are two kinds of statistical hypotheses: Null and Alternative. The symbol for the null hypothesis is H_0 and for the alternative hypothesis it is H_a. The alternative hypothesis is a counter to the null hypothesis or is the competing claim. In the sleep example, researchers might believe people are sleeping less than ever now. The null and alternative hypotheses may be phrased:

H_0: mean = 7.1
H_a: mean < 7.1

Notice the > symbol in the alternative hypothesis. The researchers believe that people are getting less sleep and hence the mean time should be less than 7.1.

Setup the null and alternative hypotheses for the following:

1. The U.S. Department of Health has set an average bacteria count of 70 bacteria per cu. cm. of water as its minimum acceptable level for clam digging waters. An average value larger than 70 is felt to be dangerous, for eating clams taken from such waters may cause hepatitis. The Department the bacteria count is at the dangerous level because of a recent environmental accident.

$H_{0:}$ mean = 70
H_a: mean > 70

2. A recent study of the ecosystem in a deciduous forest indicated that in the natural forest, the average net change in nitrate nitrogen is an increase of 2 kilograms per hectare per year. Foresters believe that a defoliation of forest undergrowth would lead to a decrease in this value.

H_0: mean = 2
H_a: mean < 2

3. The average total blood protein in a healthy adult is 7.25 grams. (advanced topic, not for Stat 10, but some of you should know how to do this)

H_0: mean = 7.25
H_a: mean ≠ 7.25

These examples lead to three general forms for the null and alternative hypotheses about a Population mean (average):

H_0: mean = some value	H_0: mean = some value	H_0: mean = some value
H_a: mean > same value	H_a: mean ≠ same value	H_a: mean < same value

The middle one is an advanced topic (not for Stat 10) and is called a "two-sided hypothesis test". The other two are for Stat 10 and are called "one-sided tests"

(b) The TEST

In this class, there was only one test used, a Z-test or you could say that the test statistic is Z. (For advanced classes, look at Chapter 26.5, there is something called a t-test)

It is the same Z that you worked with in Chapter 23. The result of the test will determine whether you reject or not reject the null hypothesis.

(c) The p-value or probability value

This is the resulting "tail probability" (that is, the area as extreme or more extreme than the Z you get from part (b) above). So suppose you were testing:

H_0: mean = 70
H_a: mean > 70

Let's say you got a sample outcome of a mean = 77 and a standard error of the average was equal to 7. The test result would be $\frac{77-70}{7}$ = Z = +1

The area in the "tail" that is more extreme than your Z score is 16%. This is your p-value or probability value.

(d) The interpretation of a p-value

The p-value tells us the probability that we would observe the data we have if the null were true. So in our example, if the null is really 70, we have a 16% chance of getting a sample outcome as larger or larger than the one we got.

If the p-value is less than or equal to some pre-specified level of significance, we reject the null and favor the alternative saying that the difference between what we got (the sample outcome) and what we were expecting (the value of the null hypothesis) is larger than what could reasonably happen by chance. This is evidence against the null hypothesis. Typical pre-specified levels of significance are 5% and 1%.

If the p value is greater than the pre-specified significance level, as in our example, we fail to reject the null. In this class, we estimate p values from table A. In the real world, researchers use statistical software which calculates the exact p value.

(e) Some observations about hypothesis testing:

1. The = is always in the null hypothesis H_0.
2. If the sample evidence strongly suggests that H_0 is false, then H_0 will be rejected in favor of Ha.
3. Often statistical hypotheses are set up in the hopes of being able to reject H_0 and thereby having evidence which suggests Ha is more correct .
4. We can never prove H_0 to be 100% true so we cannot "accept" the null hypothesis.
5. If H_0 is not rejected, we are not supporting H_0 or accepting it; but rather simply saying that there is insufficient evidence to reject it.
6. "Statistically significant" simply means that the Z test resulted in a probability lower than some pre-specified level. All it means is that the chance you could get an outcome this extreme was low. This tells you that the evidence suggests the alternative is more accurate than the null.

XI. Bivariate Relationships (Chapter 8, 9, 10.1-10.3, 12.1)

Graphical -- Scatter Plot or Scatter Diagram
Know its Direction (positive or negative)
And shape – circular, egg, football, cigar, pencil, line)

Numerical -- correlation coefficient, r
-1 <= r <= +1
supposed to be used on LINEAR relationships only

Properties of r (Chapter 8)
Invariant under addition or multiplication
Can be strongly affected by a single point

Problems with r (Chapter 9)
Weirdly sensitive (see Chapter 8 note above)
Beware non-linear scatterplots
Beware of correlations calculated using points that are averages themselves.

Regression Method for Individuals (10.3)

Pick a person at random, have information on X that you would like to take into account when predicting Y for the person. (1) find out how many SDs the person is above (or below) average for X (2) multiply this value by the correlation to convert it to how many SDs the person is away from the average for Y (3) convert the SDs to a value and add it to the average for Y to get the prediction. (p. 165)

Conversely, if you were told that a person was at the 80% percentile on the X variable, you could convert this into a corresponding Y value if you know the correlation and the average and SD of Y (p. 166)

Regression Line (12.1)

You could examine the relationship between X and Y more generally than in Chapter 10.3 by constructing a line based on information about X, Y and their correlation. The regression line has the equation:

Y = (slope*x) + intercept (page 205)

Where slope = $r * \frac{\text{Standard Deviation of Y}}{\text{Standard Deviation of X}}$ and intercept = (slope*average of X) – average of Y

The slope is interpreted as the rate of change in Y given a one unit change in X.

The intercept is the value of Y when X=0

This line will pass through the mean of X and Y so your equation Y = (slope*x) + intercept looks like this as a check on your work. In other words (mean of Y) = (slope * mean of X) + intercept