Test 2 will be given in class on Thursday April 5.

Material Covered

The test is cumulative, but will emphasize the recent material (Chapters 6–8, 10–11, and Sections 12.1–3). Note that we omitted a couple sections (8.6 and 11.5) and that there was some additional information about computing means and variances for linear transformations and combinations of random variables.

You are responsible for material covered in the text, in the problem sets, and in class.

Format

Test questions will be designed to try to see how well you understand the material, not how well you can perform various procedures mindlessly. A variety of question formats may be used. Some items may be be tested using "short answers" (a couple sentences to a paragraph), multiple choice, or true/false.

Instructions

Read through these prior to coming to the test and follow them when you take your test.

1. Always show your work and explain your reasoning. Answers without work or reasoning will not receive full credit.
   - Use mathematical notation (especially the equals sign) correctly.
   - Don’t be afraid to use words in your explanations.
   - If you get an unreasonable answer, be sure to say so. Give a brief explanation about how you know your answer is wrong (for example, the mean I calculated is less than 10, but I can see from the data that it should be at least 20). Then go on to other problems and come back and try to fix the error if you have time at the end of the test period.
   - Even if you cannot do a problem completely, show me what you do know.

2. Test restrictions.
   - The test is closed book. No notes are allowed.
   - You may use RStudio (bring your own laptop) or your calculator. There may be portions of the test where you are not allowed to use technology.
   - Do not write in purple on the exam. (The exam will be graded in purple.)
Content

Here is a list of things you should be sure you know how to do. It is not intended to be an exhaustive list, but it is an important list.

You should be able to:

- Understand, use and explain the statistical vocabulary/terminology.
  Here are some examples: population, sample, parameter, statistic, sampling distribution, standard error, p-value, confidence interval, confidence level, critical value, significance level, z-score, observational study, experiment, paired design (paired t, for example), independent samples design (2-sample t, for example), resistant, robust, . . .

- Work with random variables. This includes:
  - Using the basic rules of probability to determine probabilities of events.
  - Computing the mean and standard deviation of a random variable and knowing what they tell you.
  - Interpreting area as probability in a graph of a distribution.
  - Recognizing situations that are described by binomial, normal, and t distributions.
  - Being able to use rules for means and variances to determine the mean and variance of a more complicated random variable from means and variances of simpler random variables.

- Understand the issues involved in collecting good data and the design of studies, including
  - the distinctions between observational studies and randomized experiments.
  - matching study designs with appropriate analysis methods.

- Work with normal, t, and binomial distributions. This includes being able to use the 68-95-99.7 Rule and/or technology to find percentages, z-scores, critical values, etc.

- Understand the basic framework for hypothesis testing and how to interpret p-values.

- Understand the basic framework for confidence intervals and how to interpret the confidence level.

- Perform and interpret all of the confidence intervals and hypothesis tests covered so far. (You should be able to do these using RStudio and “by hand”.)

- Be aware of the assumptions that must be true to make use of various statistical procedures and the degree to which the procedures are robust.

- Understand how to make and interpret graphical representations of data (histogram, density plot, boxplot, bar graph, scatter plot, normal-quantile plot) and when each might be appropriate or inappropriate to use.

Note that the test will be a sample from the possible topics, it will not be exhaustive.
**Example Problems**

A number of extra problems have been assigned with each problem set.

The following problem is one that I have used very frequently on tests.

1. **What do I do?** In each of the following situations, pretend you want to know some information and you are designing a statistical study to find out about it. Give the following THREE pieces of information for each: (i) what variables you would need to have in your data set (ii) whether they are categorical or quantitative, and (iii) what statistical procedure you would use to analyze the results.

Select your procedures from the following list: 1-proportion (a.k.a. binomial test), Chi-squared goodness of fit, 1-sample \( t \), Paired \( t \), 2-sample \( t \), none of these.

Record your answers in the table. Part a) has been done as an example.

(a) You want to know if boys or girls score better on reading tests in Kent County grade schools.

**Answer Table:**

<table>
<thead>
<tr>
<th>Variable(s)</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading score on standardized test</td>
<td>2-sample ( t )</td>
</tr>
<tr>
<td>gender (male or female)</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**

- Often more than one design is possible, so there may be multiple correct answers. But you should not choose a design that is clearly inferior to another design we have already studied.

- You should not choose “none of these” if there is a reasonable design that can be analyzed by a method we already know about. “None of these” should mean that none of the listed procedures will suffice.
Inferential Statistics Summary Sheet (z and t)

Usually we will use the following notation (subscripts indicate multiple populations/samples):

- **Parameters (population)**
  - $p$, proportion (of a categorical variable)
  - $\mu$, mean (of quantitative variable)
  - $\sigma$, standard deviation

- **Statistics (sample)**
  - $n$, sample size
  - $x$, count (of a categorical variable)
  - $\hat{p} = \frac{x}{n}$, proportion (of a categorical variable)
  - $p' = \frac{\hat{p} + 2}{n+4}$, Plus-four proportion (of a categorical variable)
  - $\bar{x}$, mean (of quantitative variable)
  - $s$, standard deviation

- **Sampling distribution**
  - $SE$, standard error (standard deviation of the sampling distribution)
  - $\mu_{\hat{p}}$, $\mu_{\bar{x}}$, mean of sampling distribution (for $\hat{p}$ and $\bar{x}$, respectively)

The procedures involving the $z$ (normal) and $t$ distributions are all very similar.

- **To do a hypothesis test**, compute
  
  $$ t \text{ or } z = \frac{\text{data value} - \text{hypothesis value}}{SE} $$

  and compare with the appropriate distribution (using tables or a computer).

- **To compute a confidence interval**, determine the critical value for the desired level of confidence ($z^*$ or $t^*$), then the confidence interval is
  
  $$ \text{data value} \pm (\text{critical value})(SE) $$

Note: Each of these procedures is only valid when certain assumptions are met. In particular, remember

- The sample must be an simple random sample (or something very close to it).
- The sample sizes must be “large enough.”
- Procedures involving means are generally sensitive to outliers, because outliers can have a large effect on the mean (and standard deviation).
- Procedures involving the $t$ statistic generally assume a population that is normal or nearly normal. These procedures are sensitive to to skewness (for small sample sizes) and to outliers (for any size).
- You can’t do good statistics from bad data. The margin of error in a confidence interval, for example, only accounts for random sampling variability, not for errors in experimental design.
<table>
<thead>
<tr>
<th>Situation</th>
<th>parameters</th>
<th>statistics (computed from data)</th>
<th>distributions (usually approximate)</th>
<th>$SE$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Proportion</td>
<td>$p$</td>
<td>$x = \text{count}$</td>
<td>Binom($n, p$) $\approx$ Norm($np, \sigma\sqrt{n}$)</td>
<td>$SE = \sqrt{\frac{p(1-p)}{n}}$ [for hyp. test]</td>
</tr>
<tr>
<td>1 Categorical Variable (2 levels) proportion $p$</td>
<td>$\sigma = \sqrt{p(1-p)}$</td>
<td>$\hat{p} = \frac{x}{n}$</td>
<td>Norm($\hat{p}, SE$)</td>
<td>$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ [Wald CI]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p' = \frac{x+2}{n+4}$</td>
<td>Norm(0,1)</td>
<td>$SE = \sqrt{\frac{p'(1-p')}{n+4}}$ [Plus-4 CI]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n' = n + 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Proportion</td>
<td>$p_1, p_2$ ($\sigma_1, \sigma_2$)</td>
<td>$x_1, x_2$ (success counts)</td>
<td>Norm($p_1 - p_2, SE$)</td>
<td>$SE = \sqrt{\frac{x_1^2}{n_1} + \frac{x_2^2}{n_2}}$</td>
</tr>
<tr>
<td>2 Categorical Variables diff. of two proportions $p_1 - p_2$</td>
<td></td>
<td>$\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$,</td>
<td></td>
<td>$= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$s_1 = \sqrt{\hat{p}_1(1-\hat{p}_1)}$, $s_2 = \sqrt{\hat{p}_2(1-\hat{p}_2)}$</td>
<td></td>
<td>$[SE = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}]$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \text{pooled perc.}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Sample $t$</td>
<td>$\sigma, \mu$</td>
<td>$\bar{x}, s, df = n - 1$</td>
<td>$T(df)$</td>
<td>$SE = \frac{s}{\sqrt{n}}$</td>
</tr>
<tr>
<td>1 Quantitative Variable mean value: $\mu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paired $t$</td>
<td>$\sigma_{\text{diff}}, \mu_{\text{diff}}$</td>
<td>$\bar{x}<em>{\text{diff}}, s</em>{\text{diff}}, df = n - 1$</td>
<td>$T(df)$</td>
<td>$SE = \frac{s_{\text{diff}}}{\sqrt{n}}$</td>
</tr>
<tr>
<td>2 Quantitative Variables mean value of differences: $\mu_{\text{diff}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Welch 2-Sample $t$</td>
<td>$\sigma_1, \sigma_2$</td>
<td>$\bar{x}_1, s_1, \bar{x}_2, s_2,$</td>
<td>$T(df)$</td>
<td>$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$</td>
</tr>
<tr>
<td>1 Quantitative Variable difference of two means: $\mu_1 - \mu_2$</td>
<td></td>
<td>$df \geq \min(n_1 - 1, n_2 - 1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$df \leq n_1 + n_2 - 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Sample $t$</td>
<td>$\sigma = \sigma_1 = \sigma_2$</td>
<td>$\bar{x}_1, s_1, \bar{x}_2, s_2$</td>
<td>$T(df)$</td>
<td>$SE = sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$</td>
</tr>
<tr>
<td>1 Quantitative Variable difference of two means: $\mu_1 - \mu_2$</td>
<td></td>
<td>$sp = \sqrt{\frac{(df_1)s_1^2 + (df_2)s_2^2}{df_1 + df_2}}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pooled est. of st dev</td>
<td>$df_1 = n_1 - 1$ $df_2 = n_2 - 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$df = df_1 + df_2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>