You will not need to access any websites on the exam; indeed, you are not allowed to access any websites during the exam.

You will not be asked any questions about output from vassarstats. You need to know how to use the Normal Curve Area Calculator website. This website is used often in Statistics: To obtain an approximate P-value for the Sum of Ranks test and the Runs test. You need to know what to enter in the Mean and SD boxes and what to enter in either the Above or Below box. Remember to always use the continuity correction.

Chapters 1 and 2: The CRD with a Numerical Response. Suppose we have a set of $m$ observations, denoted by $w_1, w_2, \ldots, w_m$. The mean of these $m$ numbers is equal to their sum divided by $m$; it is denoted by $\bar{w}$. Observation $w_i$ has deviation $w_i - \bar{w}$. The variance, $s^2$, and standard deviation, $s$, are measures of spread that are functions of the deviations. (You will not be asked to compute $s^2$ or $s$ for a set of data.)

The Empirical Rule (ER) provides an interpretation of the numerical value of the standard deviation, $s$; in particular, it states that:

1. Approximately 68% of the observations lie in the closed interval $[\bar{w} - s, \bar{w} + s]$,
2. Approximately 95% of the observations lie in the closed interval $[\bar{w} - 2s, \bar{w} + 2s]$, and
3. Approximately 99.7% of the observations lie in the closed interval $[\bar{w} - 3s, \bar{w} + 3s]$.

The accuracy of this approximation is related to the shape of the distribution of the data: it is accurate for bell-shaped distributions and might not be accurate for other distribution shapes. Also, you need to know how to evaluate the ER for a particular set of data; i.e., by counting the number of observations in the interval of interest.

The sorted data (smallest to largest) are denoted by: $w_(1), w_(2), \ldots, w_(m)$. (Note the parentheses in the subscripts.) The median is the number in the center position and is denoted by $\tilde{w}$. Indeed:

- If the sample size $m$ is an odd integer, define $k = (m + 1)/2$, which will be an integer. Then $\tilde{w} = w_(k)$.
- If the sample size $m$ is an even integer, define $k = m/2$, which will be an integer. Then $\tilde{w} = (w_(k) + w_(k+1))/2$.
Regarding histograms of data, you need to know:

- For a frequency histogram, the height of a rectangle equals the frequency of its class interval;
- For a relative frequency histogram, the height of a rectangle equals the relative frequency of its class interval; and
- For a density histogram, the height of a rectangle equals the relative frequency of its class interval divided by the width of its class interval. Equivalently, the area of a rectangle equals the relative frequency of its class interval.

Remember the endpoint convention: Each class interval includes its left endpoint, but not its right endpoint. Exception: The last class interval—i.e., the class interval with the largest lower [upper] bound—includes both of its endpoints.

**Chapter 3: Randomization, Probability and Sampling Distributions.** Suppose that we plan to perform a completely randomized design (CRD) with with \( n_1 \) \( n_2 \) units assigned to treatment 1 [2] and a total of \( n = n_1 + n_2 \) units being studied. Number the units 1, 2, \ldots, \( n \). An assignment is any collection of \( n_1 \) numbers from the collection 1, 2, \ldots, \( n \). For example, if \( n_1 = n_2 = 3 \), then 1,4,6 is an assignment; namely, the assignment that places units 1, 4 and 6 on treatment 1 and the remaining units—2, 3 and 5—on treatment 2.

A test statistic is a rule that takes as input the data generated by a CRD and yields as output a number that summarizes the data in a meaningful way. The number we obtain from a particular set of data is called the observed value of the test statistic. The test statistic introduced in Chapter 3 has observed value \( u = \bar{x} - \bar{y} \) and is denoted by \( U \).

The sampling distribution of the test statistic \( U \) is a compilation of the values of \( U \), taken over all possible assignments. The sampling distribution cannot be obtained without introducing an assumption. The assumption we need is that the Skeptic’s Argument is correct. (The Skeptic’s Argument is stated below under Chapter 5.)

**Chapter 4: Approximating a Sampling Distribution.** A (computer) simulation experiment looks at some assignments instead of all assignments. Each examined assignment is selected at random from the collection of all possible assignments. The relative frequency of occurrence of the event \( B \) in a simulation experiment is an approximation of the probability of the event \( B \). The precision of this approximation can be investigated by computing the nearly certain interval. In particular, let \( \hat{r} \) denote the relative frequency of occurrence of the event \( B \) in a simulation experiment with \( m \) runs. Let \( r \) denote the probability of the event \( B \). The nearly certain interval for \( r \) is

\[
\hat{r} \pm 3 \sqrt{\frac{\hat{r}(1 - \hat{r})}{m}}.
\]

**Chapter 5: A Statistical Test of Hypotheses.** Given the exact (or an approximate) sampling distribution for \( U \), you need to know how to apply the formulas given below for finding a P-value. In particular, if \( u \) is the actual observed value of the test statistic \( U \), then the rules are:
• For the alternative $>$, the P-value is equal to $P(U \geq u)$.

• For the alternative $<$, the P-value is equal to $P(U \leq u)$.

• For the alternative $\neq$ the P-value is equal to: 1 if $u = 0$; otherwise:

\[ P(U \geq |u|) + P(U \leq -|u|). \]

If the sampling distribution is symmetric around zero, then these two terms are the same number; if not, then they might be different numbers.

When exact probabilities are not available, they are replaced in the formulas above by the relative frequencies obtained from a simulation. For the alternative $\neq$ note that even though the probabilities might be equal—because of symmetry—the relative frequencies usually won’t be.

If the CRD is balanced—i.e., if $n_1 = n_2$—then the sampling distribution of $U$ is symmetric around zero. If the CRD is not balanced, then the sampling distribution of $U$ might or might not be symmetric. In other words, balance is sufficient for symmetry, but not necessary.

It is instructive and useful to imagine studies that are not performed. In particular, the All-treatment 1 [All-treatment 2] study assigns all units to treatment 1 [2]. If we could perform both of these All-treatment studies, we would be performing the clone-enhanced study. Without additional assumptions, we do not know what a clone-enhanced study would yield.

The constant treatment effect assumption is a popular possibility for an alternative. It states that if the clone-enhanced study could be performed (i.e., each unit gives a response to both treatments), then for every unit the response on treatment 1 minus the response on treatment 2 would be a constant. If, for example, we denote this constant by the number $c$, then we say that we have a constant treatment effect of (size) $c$. Note that $c$ can be a positive number or a negative number, but it cannot be zero; if it were zero then the Skeptic’s Argument—and, hence, the null hypothesis—would be true.

Chapter 6: The Sum of Ranks Test. Given data from a CRD, you need to know how to assign ranks: combine the data into one data set; sort the data; assign positions; and assign ranks (all tied observations receive the same rank, with the rank equal to the mean of the positions). After assigning ranks, you need to know how to calculate $r_1$ and $r_2$ and the proper way to compare them descriptively, namely with $\bar{r}_1$ and $\bar{r}_2$.

For the sum of ranks test, the test statistic is $R_1$ with observed value $r_1$. As with the test statistic $U$, balance in the CRD is sufficient for the sampling distribution of $R_1$ to be symmetric, but not around 0. Also, for an unbalanced CRD, if there are no ties in the combined data set, then the sampling distribution of $R_1$ is symmetric, but again not around 0.

For $r_1$ the observed value of the test statistic $R_1$, then:

• For the alternative $>$, the exact P-value is equal to $P(R_1 \geq r_1)$.

• For the alternative $<$, the exact P-value is equal to $P(R_1 \leq r_1)$.

• You will not be asked for the exact P-value for the alternative $\neq$ and the sum of ranks test.
Chapter 7: Visualizing a Sampling Distribution. Given the sampling distribution of a random variable $X$, calculate the distances between successive sorted possible values. For example, if the sorted possible values of $X$ are:

$0, 2, 4, 7, 9$ and $10$

then the distances are $2, 2, 3, 2$ and $1$. The minimum of these distances is called $\delta$. For my example, $\delta = 1$.

A probability histogram is a collection of rectangles on the number line. There is a rectangle centered at every possible value of $X$. The rectangle centered at the number $x$ extends from $x - \delta/2$ to $x + \delta/2$ and its height equals $P(X = x)/\delta$.

The mean of the sampling distribution for $R_1$ is

$$\mu = n_1(n + 1)/2.$$  

The variance of the sampling distribution for $R_1$ is

$$\sigma^2 = \frac{n_1 n_2 (n + 1)}{12} - \frac{n_1 n_2 \sum(t_i^3 - t_i)}{12 n (n - 1)}.$$  

The standard deviation of the sampling distribution for $R_1$ is the square root of the variance. You need to be able to evaluate these expressions for the mean and variance. In particular, given data from a CRD, you need to know how to calculate the values of the $t_i$'s. Remember all $t_i$'s that equal one can be ignored because they do not affect the variance. The mean and variance of $R_1$ are needed to obtain the normal curve approximation to the P-value.

Chapters 8 and 9: Dichotomous Responses; Critical Regions and Statistical Power. In Chapter 8 we consider the situation in which the response is a dichotomy. In this situation, we present our data in a table of counts:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Total</th>
<th>Row Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S$</td>
<td>$F$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>1</td>
<td>$a$</td>
<td>$b$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>2</td>
<td>$c$</td>
<td>$d$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Total</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

The observed value of the test statistic is $x = \hat{p}_1 - \hat{p}_2$. As shown in the Course Notes, this test statistic, $X$, is equivalent to our earlier test statistic $U$. Thus, if $n_1 = n_2$, then the sampling distribution of $X$ is symmetric around 0.

Exact P-values for $X$ can be obtained from the internet. You need to remember the following properties, stated below for $X$, but also true for $U$:

- For any number $x$, $P(X \geq x) + P(X \leq x) = 1 + P(X = x)$.
If the sampling distribution is symmetric around 0, then for any number \( x \),

\[ P(X \geq x) = P(X \leq -x). \]

You will not be asked to find a critical region on the midterm. (Before you celebrate too much, this topic likely will be on the final exam.)

You will be given the results from a computer simulation experiment. For example, you might be given the following table obtained under the assumption that the Skeptic is correct.

<table>
<thead>
<tr>
<th>( U )</th>
<th>( R_1 )</th>
<th>( \text{Fail to Reject } H_0 )</th>
<th>( \text{Reject } H_0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to reject ( H_0 )</td>
<td>( a = 9,374 )</td>
<td>( b = 126 )</td>
<td>( a + b = 9,500 )</td>
<td></td>
</tr>
<tr>
<td>Reject ( H_0 )</td>
<td>( c = 128 )</td>
<td>( d = 372 )</td>
<td>( c + d = 500 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( a + c = 9,502 )</td>
<td>( b + d = 498 )</td>
<td>( m = 10,000 )</td>
<td></td>
</tr>
</tbody>
</table>

Make sure you know what all of these numbers/cells represent. For example, for 9,374 assignments both tests correctly fail to reject; for 498 assignments test \( R_1 \) incorrectly rejects; and for 128 assignments \( R_1 \) correctly fails to reject while \( U \) incorrectly rejects.

If instead of the Skeptic being correct, a particular version of the alternative is correct, we get a table with the same labels, but now the topic is power.

<table>
<thead>
<tr>
<th>( U )</th>
<th>( R_1 )</th>
<th>( \text{Fail to Reject } H_0 )</th>
<th>( \text{Reject } H_0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to reject ( H_0 )</td>
<td>( a = 7,500 )</td>
<td>( b = 100 )</td>
<td>( a + b = 7,600 )</td>
<td></td>
</tr>
<tr>
<td>Reject ( H_0 )</td>
<td>( c = 300 )</td>
<td>( d = 2,100 )</td>
<td>( c + d = 2,400 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( a + c = 7,800 )</td>
<td>( b + d = 2,200 )</td>
<td>( m = 10,000 )</td>
<td></td>
</tr>
</tbody>
</table>

As with the earlier table, you need to know what each number/cell represents. In this table, the approximate powers are 0.2400 for \( U \) and 0.2200 for \( R_1 \). If we denote the exact powers by \( r_U \) and \( r_{R_1} \), respectively, then the nearly certain interval for \( r_U - r_{R_1} \) is:

\[
\frac{(c - b)}{m} \pm \frac{(3/m)}{\sqrt{b + c}}.
\]

The nearly certain interval for \( r_{R_1} - r_U \) is

\[
\frac{(b - c)}{m} \pm \frac{(3/m)}{\sqrt{b + c}}.
\]

Note that both of these intervals are approximations to and simplifications of the intervals given in the Course Notes. In my opinion the approximation is good provided \( m \geq 1,000 \).

Chapter 10: Populations: Getting Started. There are two types of problems in Chapter 10.

First, I will describe a Chance Mechanism and you determine probabilities using some combination of multiplying and adding. Below are two examples.
**Chance Mechanism:** I have a box with four cards, numbered 1, 2, 3 and 4. I plan to select two cards at random with replacement. Let $T$ be the total of the two cards and define $X_1$ and $X_2$ as in the *Course Notes*. I want to determine $P(T = 7)$. This is equal to

$$P(X_1 = 3 \text{ and } X_2 = 4) + P(X_1 = 4 \text{ and } X_2 = 3) = (1/4)(1/4) + (1/4)(1/4) = 2/16 = 1/8.$$ 

**Chance Mechanism:** Same as above, but I will select two cards at random without replacement. Again, I want $P(T = 7)$ which can be written the same as above. When we calculate, however, we now get:

$$(1/4)(1/3) + (1/4)(1/3) = 2/12 = 1/6.$$ 

Second, you are given a table of joint probabilities for two random variables $X_1$ and $X_2$, you need to be able to determine the probability of a variety of events. Identify the cells in the table that correspond to the event of interest and sum their probabilities; using the multiplication rule typically does not work.

**Chapter 11: Bernoulli Trials (BT).** If we have BT, then, in addition to the multiplication rule being true, we know that if $X$ is the total number of successes in $n$ BT, then the sampling (or probability) distribution of $X$ is binomial, written $\text{Bin}(n, p)$, where $p$ is the probability of success on a trial. For ($n \leq 6$) make sure you can evaluate the following by hand.

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

for $x = 0, 1, 2, \ldots, n$.

If $X$ has a binomial distribution, its mean is $np$ and its variance is $npq$.

You need to know about the random variables $R$ (the number of runs), $V$ (the length of the longest run of successes) and $W$ (the length of the longest run of failures). For example, suppose that 15 dichotomous trials yield:

11010 11100 00101.

Then the observed values of these random variables are $r = 9$, $v = 3$ and $w = 4$. Also, the observed number of successes is $x = 8$.

Inference for $R$ is performed conditional on the value of $x$. In particular, define $c = 2x(n-x)$. With this definition, the mean and standard deviation of the sampling distribution of $R$ is:

$$\mu = 1 + (c/n) \text{ and } \sigma = \sqrt{\frac{c(n-c)}{n^2(n-1)}}.$$