Chapter 10: Describing Data. For histograms we begin with a frequency table which has the following columns:

| Class Interval | Width \((w)\) | Freq. \((f)\) | Rel. Freq. \((rf = f/n)\) | Density \((rf/w)\) |

The class intervals must encompass all data; the endpoints of consecutive intervals touch—i.e. they do not overlap and there are no gaps. The width of a class interval is the distance between its endpoints. The frequency of a class interval is the number of observations in it; remember the endpoint convention when counting—each interval includes its left endpoint but not its right endpoint. The relative frequency of a class interval is its frequency divided by the sample size, \(n\). The density of a class interval is its relative frequency divided by its width.

A histogram is a collection of rectangles. Each class interval has a rectangle and the height of the rectangle is the frequency (for a frequency histogram), relative frequency (for a relative frequency histogram) or density (for a density scale histogram) of its class interval.

When all class intervals have the same value of \(w\), then the three histograms have identical shapes. When the class intervals have variable \(w\)’s, you must use the density scale histogram b/c the other two are misleading.

Measures of center. We denote a data set by \(x_1, x_2, x_3, \ldots, x_n\). After sorting the data from smallest to largest we denote the data by \(x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}\). The mean is denoted by \(\bar{x}\) and is calculated by \(\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}\). The median is denoted by \(\tilde{x}\); its computation is a special case of the rule given below for percentiles.

Measures of position. The median is a measure of position. The median can be generalized to percentiles, of which there are 99. For \(j = 1, 2, \ldots, 99\) we calculate the \(j\)th percentile as follows.

- Compute \(k = jn/100\).
- If \(k\) is an integer, then the \(j\)th percentile is \((x_{(k)} + x_{(k+1)})/2\).
- If \(k\) is not an integer, then round it up to the next integer, \(k'\). The \(j\)th percentile is \(x_{(k')}\).

The median is the 50th percentile. Two other percentiles have names. The 25th (75th) percentile is called the first (third) quartile and is denoted by \(Q_1 (Q_3)\).

Measures of spread. The interquartile range, denoted IQR, equals \(Q_3 - Q_1\). Very roughly speaking, it is the range of the center half of the data. A much more useful measure of spread is given by either of the closely related variance, denoted \(s^2\), or standard deviation, denoted \(s\).

We begin by associating with each observation, \(x_i\) its deviation, \(x_i - \bar{x}\). There are \(n\) deviations but b/c they sum to 0, we say that they have \((n-1)\) degrees of freedom, written \(df = n - 1\). The variance is \(s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n - 1)\), and the standard deviation is \(s = \sqrt{s^2}\). The empirical rule gives us an interpretation of \(s\): Approximately 68% of the data are in the interval \(\bar{x} \pm s\). The quality of the empirical rule approximation is related to the shape of the data set. If the data set is approximately bell-shaped then the empirical rule approximation is good. If the data are strongly skewed (to the left or right) then the interval \(\bar{x} \pm s\) usually contains much more than 68% of the data. Finally, if the data have two prominent peaks with a valley between them, then the interval \(\bar{x} \pm s\) usually contains much less than 68% of the data.
Chapter 11: Inference for One Population. We assume that our data, \( x_1, x_2, x_3, \ldots, x_n \), are the observed values of i.i.d. trials or are a random sample from a finite population. We summarize the data with \( \bar{x} \) and \( s \).

Gosset’s CI for the population mean, \( \mu \), is given by \( \bar{x} \pm t \left( \frac{s}{\sqrt{n}} \right) \). You need to know how to find the value of \( t \) from the online calculator: remember \( df = n - 1 \).

We can test the null hypothesis that \( \mu = \mu_0 \), where \( \mu_0 \) is a known number specified by the researcher. The observed value of the test statistic is

\[
t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.
\]

You need to know how to use the online calculator to obtain the P-value for any of the three possible alternatives.

It is also possible to estimate the median of a pdf, denoted by \( \nu \). There is an exact method for \( n \leq 20 \), but it won’t be on the exam. The approximate method uses \( z \) which you recall is given in the following table.

| \( z \) | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |
| Confidence Level | 80% | 90% | 95% | 98% | 99% |

First, you sort the data as discussed earlier. Then you select a desired confidence level from the five above, giving you your \( z \). Next, calculate

\[
k' = \frac{n + 1}{2} - \frac{z\sqrt{n}}{2}.
\]

If \( k' \) is an integer (never happens) set \( k = k' \). Otherwise, round \( k' \) down to the nearest integer \( k \). The CI is \([x_{(k)}, x_{(n+1-k)}]\).

The logic of CI’s. For concreteness, assume that we are estimating the mean of a population, \( \mu \). A particular CI, denoted by \([L, U]\), is either: correct, too small or too large. Correct means, of course, that the parameter value is within the CI; i.e. \( L \leq \mu \leq U \). Too small means that \( U < \mu \). Too large means that \( L > \mu \).

Chapter 12: Inference for Two Populations. We want to compare two populations by comparing their means. The first (second) population has mean \( \mu_1 \) (\( \mu_2 \)), standard deviation \( \sigma_1 \) (\( \sigma_2 \)) and variance \( \sigma_1^2 \) (\( \sigma_2^2 \)). (I created a lot of confusion in the Course Notes etc., by writing \( \mu_X \) (\( \mu_Y \)) when I meant \( \mu_1 \) (\( \mu_2 \)). Sorry about that.)

Data. We assume that we have independent random samples from the two populations. We denote the data from the first population by: \( x_1, x_2, \ldots, x_{n_1} \), and denote the data from the second population by: \( y_1, y_2, \ldots, y_{n_2} \). We summarize our two sets of data by computing their means and standard deviations, which are denoted by:

\[
\bar{X}, S_1, \bar{Y} \text{ and } S_2.
\]

Our point estimate of \( \mu_1 - \mu_2 \) is \( \bar{X} - \bar{Y} \). The standardized version of this estimator is

\[
W = \frac{(X - Y) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}.
\]
For the exam, you need to know only Case 1, in which we must assume that \( \sigma_1^2 = \sigma_2^2 \). We call this common value \( \sigma^2 \). The best estimate of \( \sigma^2 \) is

\[
s^2_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.
\]

With this estimate, \( W \) is replaced by \( W_1 \):

\[
W_1 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}}.
\]

Case 1 tells us to use the \( t \) curve with \( df = n_1 + n_2 - 2 \) as our reference. This leads to Gosset’s CI for \( \mu_1 - \mu_2 \):

\[
(\bar{x} - \bar{y}) \pm ts_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.
\]

Remember: You need to know how to use the online calculator to find the value of \( t \).

We can test \( H_0 : \mu_1 = \mu_2 \). The observed value of the test statistic is

\[
t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{1/n_1 + 1/n_2}}.
\]

You need to know how to use the online calculator to obtain the P-value for any of the three possible alternatives.

**Paired data.** Pair \( i \) gives us \( x_i \) and \( y_i \). We calculate the difference, \( d_i = x_i - y_i \) and analyze the \( d \)'s using the methods of Chapter 11.

**Chapter 13: Two Dichotomies.** For a finite population there is a table of population counts and population proportions/probabilities:

<table>
<thead>
<tr>
<th></th>
<th>( B )</th>
<th>( B^c )</th>
<th>Total</th>
<th>( B )</th>
<th>( B^c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>( N_{AB} )</td>
<td>( N_{AB}^c )</td>
<td>( N_A )</td>
<td>( A )</td>
<td>( p_{AB} )</td>
<td>( p_{AB}^c )</td>
</tr>
<tr>
<td>( A^c )</td>
<td>( N_{A'}B )</td>
<td>( N_{A'B}^c )</td>
<td>( N_{A'} )</td>
<td>( A^c )</td>
<td>( p_{A'B} )</td>
<td>( p_{A'B}^c )</td>
</tr>
<tr>
<td>Total</td>
<td>( N_B )</td>
<td>( N_{B'} )</td>
<td>( N )</td>
<td>Total</td>
<td>( p_B )</td>
<td>( p_{B'} )</td>
</tr>
</tbody>
</table>

For trials, we have the table of probabilities. We define conditional probability as follows: \( P(A|B) = P(AB)/P(B) \). For a finite population, if you prefer you may use the formula \( P(A|B) = N_{AB}/N_B \). This definition gives us the multiplication rule for conditional probabilities: \( P(AB) = P(A)P(B|A) \).

If you are given (usually) three probabilities/conditional probabilities you should be able to build a table of probabilities. Once you have the table, you can calculate any conditional probability that you want.

**Testing and estimation for comparing** \( p_A \) **and** \( p_B \). We present our data as:

<table>
<thead>
<tr>
<th></th>
<th>( B )</th>
<th>( B^c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>( a )</td>
<td>( b )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>( A^c )</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>
We can test \( H_0 : p_A = p_B \). Define \( m = b + c \). We obtain the exact P-value by using the online calculator: enter 0.5 in the first box; enter \( m \) in the second box; enter \( x = b \) in the third box. In the output, \( P(X \geq x) \) is the P-value for \( >; \) \( P(X \leq x) \) is the P-value for \( <; \) and for \( \neq \), the P-value is the smallest of the following three numbers: 1, \( 2P(X \geq x) \); and \( 2P(X \leq x) \).

For estimation, the CI for \( p_A - p_B \) is

\[
\frac{(b - c)}{n} \pm z \sqrt{\frac{n(b + c) - (b - c)^2}{(n - 1)n^2}}.
\]

Chapter 14: Regression. We observe/determine two numbers per case (subject, trial), denoted by \( (x_i, y_i) \) for case \( i \). We view the \( x_i \)'s as fixed numbers and the \( y_i \)'s as observed values of random variables. (There are issues, but we will ignore them b/c we did not have time to cover them.) A new idea is that we impose a model on the structure of our data/populations. The model is

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, 2, 3, \ldots n. \]

The \( \epsilon_i \)'s are i.i.d. random variables with mean 0 and variance \( \sigma^2 \). The numbers \( \beta_0 \), \( \beta_1 \) and \( \sigma \) are parameters, whose values are known to Nature and unknown to the researcher.

We use the Principle of Least Squares to obtain point estimates of \( \beta_0 \) (denoted by \( b_0 \)) and \( \beta_1 \) (\( b_1 \)). This gives us the regression line \( \hat{y} = b_0 + b_1 x \). Thus, for any case in the data set, we can use its \( x \) to calculate a predicted value for \( y \). To measure the quality of these predictions, we define the residual \( e_i = y_i - \hat{y}_i \). The residuals are sample versions of the \( \epsilon_i \)'s and we use them to estimate \( \sigma \).

The residuals have two linear restrictions: \( \sum e_i = 0 \) and \( \sum (x_i e_i) = 0 \), so we say that they have \( df = n - 2 \). We define SSE to be \( \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \). Our estimate of \( \sigma^2 \) is denoted by \( s^2 \) and is equal to SSE divided by its df, which is also called MSE. Our estimate of \( \sigma \) is \( s = \sqrt{s^2} \).

You need to know how to use computer output to handle three inference problems: CI for \( \beta_1 \); CI for the mean of \( Y \) for a given value of \( X \); PI of value of \( Y \) for a given value of \( X \). The formulas are, respectively,

\[
\text{Coef} \pm t(\text{SE Coef}); \text{Fit} \pm t(\text{SE Fit}); \text{ and Fit} \pm t(\text{SE Pred}),
\]

where \( \text{SE Pred} \) equals the square root of \( \text{VAR Pred} \), which is given by

\[ \text{VAR Pred} = s^2 + (SE \text{ Fit})^2. \]

In all cases, \( t \) is obtained from the online calculator with \( df = n - 2. \)

Finally, there is the ANOVA table. The headings are

\[ \text{Source} \quad DF \quad SS \quad MSE \]

The sources are: Regression, Residual Error, Total; the \( DF \) are \( 1 \), \( (n - 2) \) and \( (n - 1) \). SSE is above; \( \text{SST} = \sum (y - \bar{y})^2 \); the SS’s add. MS is SS divided by \( DF \), but we don’t calculate it for total.

The Coefficient of Determination is denoted by \( R^2 \) and equals \( (\text{SST} - \text{SSE}) \) divided by \( \text{SST} \).