Chapter 2: Review of Statistical Inference

Statistics: The science of decision making in the presence of uncertainty.

In this chapter we are going to talk about tools for making statistical inference, that is drawing inference about a population based upon a sample from that population. The characteristics of the population associated with the random variable of interest are called parameters. These procedures will be of two primary types. 1) Estimation of a parameter and 2) testing a hypothesis about a parameter.

Hopefully some of these concepts are familiar to you. These concepts are covered in courses which are prerequisites for this class. Please see me if the majority of this material does not appear familiar to you.

We will begin by talking about estimation of a parameter. Two common parameters that we will wish to estimate are the mean (center) of the data and the variance (spread) of the data.
The Mean and Variance of a Random Variable

The **mean** or the **expected value** of a random variable \( Y \) is denoted \( \mathbb{E}(Y) = \mu_Y \). For a discrete random variable,

\[
\mathbb{E}(Y) = \sum_i y_i \Pr(Y = y_i).
\]

If we have a continuous random variable, then the mean has the form

\[
\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f(y) \, dy,
\]

where \( f(y) \) is the probability density function of \( Y \).

The **variance** of a random variable is the long-term average of squared deviations from the mean. We will use the notation \( \text{Var}(Y) = \mathbb{E}\{(Y - \mu_Y)^2\} = \sigma_Y^2 \). For discrete and continuous random variables the forms are

\[
\text{Var}(Y) = \sum_i (y_i - \mu_Y)^2 \Pr(Y = y_i)
\]

and

\[
\text{Var}(Y) = \int_{-\infty}^{\infty} (y - \mu_y)^2 f(y) \, dy,
\]

respectively.

Some important properties of the expectation are summarized below. Suppose that \( k \) is a constant, \( Y, Y_1, Y_2 \) are all random variables. Then,

\[
\begin{align*}
\mathbb{E}(k) &= k \\
\mathbb{E}(kY) &= k \mathbb{E}(Y) \\
\mathbb{E}(Y_1 + Y_2) &= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) \\
\text{Var}(Y) &= \mathbb{E}\{(Y - \mu_Y)^2\} = \mathbb{E}(Y^2) - \mu_Y^2.
\end{align*}
\]
Even if we don’t know about the population, we can calculate quantities called the sample mean and the sample variance. We will define the sample mean as

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n},$$

and the sample standard deviation as

$$S^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1}.$$  

Notice that in both of these formulas we have used capital letters. Throughout this class, we will use capital letters to denote random variables and lower case letters to indicate particular values. Thus, if we were to take a sample that contained $n$ values, then the observed sample mean and observed sample variance would be

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}.$$  

Also, the positive square root of the (sample) variance is called the (sample) standard deviation and can be denoted by $S$, while is observed value is denoted by $s$.

$\bar{Y}$ and $S^2$ are statistics that we will use as point estimates of the population parameters $\mu_Y$ and $\sigma^2_Y$. Next, we will discuss some of the properties of these point estimates and talk about what they tell us of the underlying population of interest.
To refresh your mind about how to calculate these, suppose that $Y$ is discrete with $Y = 0, 1$. Suppose that $\Pr(Y = 1) = .25$. Calculate the mean and variance of $Y$.

Now, suppose that we gather the following 5 data points: 9, 2, 4, 3, 7. Find the sample mean and variance for this data.
Estimation

We will define an **unbiased statistic** as one which has the following property: if \( W \) is a point estimate of the population parameter \( \theta \), then \( W \) is an unbiased statistic if \( E(W) = \theta \). That is, in a long-term average, the value of the statistic is equal to what we are trying to estimate.

The quantities \( \bar{Y} \) and \( S^2 \) are unbiased estimators of the population quantities \( \mu_Y \) and \( \sigma^2_Y \). \( \bar{y} \) and \( s^2 \) are the corresponding unbiased estimates. I will show you the derivation for the unbiasedness of \( \bar{Y} \); the derivation for \( S^2 \) is more difficult.

Now, we will define the **sum of squares** of a sample \( Y_1, Y_2, ..., Y_n \) to be the statistic

\[
SS_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.
\]

We will use the notation \( SS_Y \) to mean either the random variable (above) or the estimate found by plugging in the observed values of the \( y_i \)'s.
Notice that $SS_Y = (n - 1)S_Y^2$ when $Y_1, Y_2, ..., Y_n$ is a random sample from an unknown distribution. This means that $E(SS_Y) = (n - 1)\sigma_Y^2$, where we will call the $(n - 1)$ the **degrees of freedom (df)** associated with $SS_Y$. If we divide the sum of squares by its degrees of freedom, we get a quantity called **mean square**, denoted $MS_Y$. Note: the degrees of freedom are not always $n - 1$, especially in more complicated experimental designs.

**Interval Estimators and Interval Estimates**

Define a $100(1 - \alpha)\%$ **Confidence Interval Estimator** of a population parameter as a random interval that should contain the parameter of interest with probability $1 - \alpha$ for some $\alpha$ such that $0 < \alpha < 1$. When we use a sample to obtain the endpoints of such an interval, we will call the result a $100(1 - \alpha)\%$ **Confidence Interval (Estimate)**. For example, you may have constructed a 95% confidence interval for the mean of a normal distribution based upon a sample of size $n$ and a known standard deviation $\sigma$, according to the formula

$$
\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}.
$$

This interval may be interpreted in the following way: if we construct many interval according to this procedure for many different samples of size $n$, about 95% of them should contain the true value of the parameter $\mu$. However, once we collect a sample and construct an interval, that interval either does or does not contain $\mu$. 
You may be familiar with the following two interval estimators for the mean. Suppose that $Y$ is normally distributed with mean $\mu$ and standard deviation $\sigma$. If we observe a random sample $y_1, y_2, ..., y_n$ from $Y$ and let $\Pr(Z \leq Z_{1-\alpha/2}) = 1 - \alpha/2$, where we can get $Z_{1-\alpha/2}$ from Table A, we can construct the confidence interval for $\mu$ as

$$\bar{y} \pm Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

If we do not know that value of $\sigma$, we have to estimate it using $s$. In this case, we would look up $t_{1-\alpha/2,n-1}$ in Table B, and construct the interval as

$$\bar{y} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}}.$$

To help visualize some of these ideas, I used a computer to simulate 50 samples of size $n = 9$ from a normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 1$.

I have made graphs which depict the means of the 50 samples, the estimated variance of the samples, the estimated standard deviations of the samples, and the 90% confidence intervals assuming first that $\sigma$ is known and second that it is unknown.

What do you notice from looking at the pictures?
Plot of the Mean of the 50 Sims

Plot of the Variance of the 50 Sims

Plot of the St. Dev. of the 50 Sims

90% CIs with Sigma Known

90% CIs with Sigma Unknown
Tests of Hypotheses  Suppose that you are interested in answering a particular question about a population. For example, suppose you are interested in determining if your new SUV does, in fact, get on average the 23 miles per gallon on the highway that the manufacturer claims. In statistical terms, you wish to answer the questions if the mean mpg, \( \mu \geq 23 \), or if the true average mpg is less than the manufacturer claims, \( \mu < 23 \).

Define a **Statistical Hypothesis** as an assertion about a population that we wish to test using our sample. Define a **Test of a Hypothesis** as a rule by which we decide whether or not to reject a hypothesis. For instance, we could decide that if you get 20 mpg or less when driving on the highway, we would reject the hypothesis that the mean highway mpg is 23.

A **Test Statistic** is just a statistic calculated from the sample that we collect to assess the validity of our hypothesis. In many cases, we might use the sample mean as a test statistic. In our example, suppose that we take ten highway drives and determine the average mpg that we see on these drives.

The **Critical Region** for a test statistic is the values of the test statistic for which we are going to reject the null hypothesis. In this case, we would say that all sample means, \( \bar{y} \mid \bar{y} \leq 20 \) define the critical region for our example.
We only get to see a sample from the population, not the entire population, so there is always the chance that we will make an incorrect decision about our hypothesis.

First, suppose that our SUV does, in fact, get an average of $\mu = 23$ mpg. However, when we do a sample, we see an average of $\bar{y} = 20$. Thus, we would decide to reject the hypothesis, even though the hypothesis is true. This is called a **Type I error**. We will denote the probability of such an error by $\alpha$. Note that this is the same $\alpha$ as in confidence interval construction.

On the other hand, suppose that our SUV only gets an average of $\mu = 22$ mpg. When we do a sample, suppose that we observe an average of $\bar{y} = 23$. Thus, we would decide not to reject the null hypothesis, even though it is false. We will call this a **Type II error**. The probability of committing a type II error depends upon the true value of the parameter of interest. We will denote this probability by $\beta(\mu)$.

Your book discusses the classical approach to hypothesis testing. This concept is something you may have seen in other courses, where you determine the critical region first, and then determine whether or not the test statistic falls into the rejection region. Please read the section in the book about the classical approach, as we will not be using it much in this course.
For the classical approach, we can test a hypothesis by forming the appropriate $Z$ or $t$ statistic, and then comparing to the critical values from Tables A & B. Use the forms

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

We will be using the **p-value** method for determining whether or not to reject a null hypothesis. The p-value of a test is the probability of seeing a value of the test statistic at least as extreme as we observed when the null hypothesis is true. We would then reject the hypothesis in the event that the p-value of our test is smaller than the $\alpha$ level agreed upon. Usually, scientists choose either the $\alpha = .01, .05$.

Suppose that we have the sample $5, 7, 7, 8, 9, 10$, which we believe comes from a normal distribution with standard deviation $\sigma = 2$. First, find the mean of the sample and form a 95% confidence interval.
Now, form the confidence interval assuming that the value of $\sigma$ is unknown.

Now, perform a test of the hypothesis that the true mean, $\mu = 8$, at $\alpha = .05$.

Finally, perform a test of the hypothesis assuming that the value of $\sigma$ is unknown. Still use $\alpha = .05$. 
The Operating Characteristic Curve

Recall that we said that the probability of making a Type II error depends upon the true value of the parameter of interest. As it is usually of interest to us to reject hypotheses, we would like to make this error as small as possible.

Define the Operating Characteristic Curve (or OC curve) for a particular hypothesis test as the plot of $\beta(\mu)$ for a variety of values of $\mu$, for a fixed sample size and hypothesis.

For example, suppose that we are interesting in determining if a new type of laptop battery has longer life than the one currently in use. We will say that the current battery can run a particular laptop for an average of 4.7 hours, with a standard deviation of 20 minutes. Let us define our null hypothesis as $\mu \leq 4.7$, as we are only interested in knowing if the new battery is better. Suppose that we will use a significance level $\alpha = .05$ and a sample size of $n = 10$.

First, find the value which would be the largest observed average for which we would not reject the hypothesis.
Now, find the probability that we observe an average lifetime of less than 4.87 given that the true average is \( \mu = 4.8, 4.9, 5.0, 5.1 \).

Finally, plot the results on a set of axes.
A plot related to the OC curve is called the **Power Curve** because the quantity $1 - \beta(\mu)$ is called the **Power**. In general, we are interested in having large power. This is because power gives us the chance of correctly rejecting the hypothesis when it is false. This is usually the goal of our experiment, so we want this probability to be as large as possible.

Notice that the power curve will just look like the OC curve "upside-down". Draw it below for the example:

In both of these cases, we have considered both $\sigma$ and $n$ to be fixed. Decreasing $\sigma$ and increasing $n$ would both serve to increase our power. Unfortunately, $\sigma$ is a characteristic of the process that we are looking at and cannot be changed. We can increase $n$, but this will sometimes be very expensive. In the next section we will figure out how to determine how big $n$ needs to be to achieve certain goals.
Sample Size Calculations

Suppose that we want to determine how many samples we are going to need before we collect data to assess a hypothesis. If we can answer three questions, we can get an idea how many samples we will need.

- How large of a difference from the hypothesized value do you wish to detect? [For our example, perhaps we are interested in a difference of $\Delta = 15$ minutes.]

- How much variability is found in the population? [Usually based upon historical data, in our example I gave you $\sigma = 20$ minutes.]

- What size risks are acceptable to you? [For example, we want a hypothesis test of level $\alpha = .05$. Perhaps we also want to detect the difference with probability $.99$, or $\beta = .01$.]

There are two formulas which now contain $n$; the one relating to $\alpha$ and the one relating to $\beta$. We know that our cutoff value must satisfy

$$\frac{\bar{Y}_c - \mu_0}{\sigma / \sqrt{n}} = \pm Z_{1-\alpha} \text{ or } \pm Z_{1-\alpha/2}$$

and

$$\frac{\bar{Y}_c - \mu_1}{\sigma / \sqrt{n}} = \pm Z_{1-\beta}$$
Let’s do this for the example that I presented. Recall that we are interested in detecting an increase in battery life of at least 15 minutes, when the hypothesized value is 4.7 hours with a standard deviation of 20 minutes. Recall that we have $\alpha = .05$ and $\beta = .01$.

We will not do too many sample size calculations of this exact form, but they can be an important part of experimental design. Sometimes experiments fail not because the hypotheses are incorrect, but rather because the sample sizes are too small to detect reasonable differences.
Summary of Hypothesis Tests - Means

Tests on a Single Mean

- Testing $H_0: \mu = \mu_0$ vs i. $H_1: \mu \neq \mu_0$, ii. $H_1: \mu > \mu_0$, or iii. $H_1: \mu < \mu_0$.

- Testing when $\sigma^2$ is known.
  - Test Statistic: $Z = (\bar{Y} - \mu_0)/(\sigma/\sqrt{n})$. This follows a standard normal distribution when the sampled population is normally distributed. Otherwise don’t use unless $n$ is large (say $n \geq 25$).
  - Decision Rule: Reject $H_0$ if i. $z \geq Z_{1-\alpha/2}$ or $z \leq Z_{\alpha/2}$, ii. $z \geq Z_{1-\alpha}$, or iii. $z \leq Z_{\alpha}$.
  - $p$-value: i. $2pr(Z \geq |z|)$, ii. $pr(Z \geq z)$, or iii. $pr(Z \leq z)$.
  - Confidence Interval for $\mu$: $\bar{y} \pm Z_{1-\alpha/2}\sigma/\sqrt{n}$. If $Y$ is not normal, use only for large sample size.

- Testing when $\sigma^2$ is unknown.
  - Test Statistic: $t = (\bar{Y} - \mu_0)/(S/\sqrt{n})$. This follows a $t$ distribution with $n - 1$ degrees of freedom when the sampled population is normally distributed. Otherwise don’t use unless $n$ is large (say $n \geq 30$).
  - Decision Rule: Reject $H_0$ if i. $|t| \geq t_{1-\alpha/2}$, ii. $t \geq t_{1-\alpha}$, or iii. $t \leq -t_{1-\alpha}$. 

18
- p-value: i. $2 \Pr(t \geq |t_{\text{obs}}|)$, ii. $\Pr(t \geq t_{\text{obs}})$, or iii. $\Pr(t \leq t_{\text{obs}})$.

- Confidence Interval for $\mu$: $\bar{y} \pm t_{1-\alpha/2}s/\sqrt{n}$, using the $t$ distribution with $n-1$ degrees of freedom. If $Y$ is not normal, use only for large sample size.

**Tests on Two Means**

- Testing $H_0 : \mu_1 = \mu_2$ vs. i. $H_1 : \mu_1 \neq \mu_2$, or ii. $H_1 : \mu_1 > \mu_2$.

- Testing when samples are independent and $\sigma_1^2$ and $\sigma_2^2$ are known.

  - Test Statistic: $Z = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$. This follows a standard normal distribution when the sampled populations are normally distributed and $H_0$ is true. Otherwise don’t use unless $n_1 \geq 25$ and $n_2 \geq 25$.

  - Decision Rule: Reject $H_0$ if i. $z \geq Z_{1-\alpha/2}$ or $z \leq Z_{\alpha/2}$, or ii. $z \geq Z_{1-\alpha}$.

  - p-value: i. $2 \Pr(Z \geq |z|)$, or ii. $\Pr(Z \geq z)$.

  - Confidence Interval for $\mu_1 - \mu_2$: $(\bar{y}_1 - \bar{y}_2) \pm Z_{1-\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. If either distribution is not normal, use only for large sample size.
• Testing when samples are independent and $\sigma_1^2$ and $\sigma_2^2$ are unknown but equal.

  – Test Statistic:
    \[
    t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.
    \]
    This follows a $t$ distribution with $n_1 + n_2 - 2$ degrees of freedom when the sampled populations are normally distributed and $H_0$ is true. Otherwise don’t use unless $n_1 \geq 30$ and $n_2 \geq 30$.

  – Decision Rule: Reject $H_0$ if i. $|t| \geq t_{1-\alpha/2}$, or ii. $t \geq t_{1-\alpha}$.

  – $p$-value: i. $2pr(t \geq |t_{obs}|)$, or ii. $pr(t \geq t_{obs})$.

  – Confidence Interval for $\mu_1 - \mu_2$:
    \[
    (\bar{y}_1 - \bar{y}_2) \pm t_{1-\alpha/2}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.\]
    If either distribution is not normal, use only for large sample size.

• Testing when samples are independent and $\sigma_1^2$ and $\sigma_2^2$ are unknown and unequal.

  – Often called the Behrens-Fisher problem. We will usually say that if two variances are different, that the two populations must also be different.

  – See the text for a way to test if the means are different in this situation.
• Testing when samples are dependent or correlated. Can be used if the elements of the two samples are paired in some way (ex. before and after some treatment). Define \( D_i = Y_{1i} - Y_{2i} \), where \( Y_{1i} \) and \( Y_{2i} \) are correlated. Then we will test \( H_0 : \mu_D = 0 \) vs. i. \( H_1 : \mu_D \neq 0 \), or ii. \( H_1 : \mu_D > 0 \).

  – Test Statistic: \( t = \frac{\bar{D}}{S_D / \sqrt{n}} \). This follows a \( t \) distribution with \( n-1 \) degrees of freedom when the population of differences is normally distributed and \( H_0 \) is true. Otherwise don’t use unless \( n \geq 30 \).

  – Decision Rule: Reject \( H_0 \) if i. \( |t| \geq t_{1-\alpha/2} \), or ii. \( t \geq t_{1-\alpha} \).

  – p-value: i. \( 2 \text{pr}(t \geq |t_{\text{obs}}|) \), or ii. \( \text{pr}(t \geq t_{\text{obs}}) \).

  – Confidence Interval for \( \mu_D \): \( \bar{d} \pm t_{1-\alpha/2} s_D / \sqrt{n} \). If \( D \) is not normal, use only for large sample size.

Your text contains examples of using some of these procedures. Please look over them, and come see me if they give you any difficulty. Notice that the software used in the examples is Excel, not SAS.
Summary of Hypothesis Tests - Variances

Tests on a Single Variance

- Testing $H_0: \sigma^2 = \sigma_0^2$ vs i. $H_1: \sigma^2 \neq \sigma_0^2$, ii. $H_1: \sigma^2 > \sigma_0^2$, or iii. $H_1: \sigma^2 < \sigma_0^2$.
  
  -- Test Statistic: $W = \frac{(n-1)S^2}{\sigma_0^2}$. This follows a chi-square distribution with $\nu = n - 1$ degrees of freedom when the sampled population is normally distributed, $\mu$ is unknown, and $H_0$ is true.

  -- Decision Rule: Reject $H_0$ if i. $w \geq W_{1-\alpha/2}$ or $w \leq W_{\alpha/2}$, ii. $w \geq W_{1-\alpha}$, or iii. $w \leq W_{\alpha}$.

  -- p-value: i. $\min\{2\Pr(W \geq w), 2\Pr(W \leq w)\}$, ii. $\Pr(W \geq w)$, or iii. $\Pr(W \leq w)$.

  -- Confidence Interval for $\sigma^2$: $\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]$, from Table C with $\nu = n - 1$ degrees of freedom, when $Y$ is normal.

Tests on Two Variances: Independent Samples

- Testing $H_0: \sigma_1^2 = \sigma_2^2$ vs i. $H_1: \sigma_1^2 \neq \sigma_2^2$, or ii. $H_1: \sigma_1^2 > \sigma_2^2$. 
Test Statistic: \( F = \frac{S_1^2}{S_2^2} \). This follows an \( F \) distribution with \( \nu_1 = n_1 - 1 \) numerator degrees of freedom and \( \nu_2 = n_2 - 1 \) denominator degrees of freedom when the sampled populations are normally distributed, the samples are independent, and \( H_0 \) is true.

Decision Rule: Reject \( H_0 \) if i. \( f \geq F_{1-\alpha/2} \) or \( f \leq F_{\alpha/2} \), of ii. \( f \geq F_{1-\alpha} \).

p-value: i. \( \min\{2\Pr(F \geq f), 2\Pr(F \leq f)\} \), or ii. \( \Pr(F \geq f) \).

Confidence Interval for \( \sigma_1^2/\sigma_2^2 \): 
\[ [(s_1^2/s_2^2)F_{\alpha/2}, ((s_1^2/s_2^2)F_{1-\alpha/2})], \]
from Table D with \( \nu_1 = n_2 - 1 \) degrees of freedom, \( \nu_2 = n_1 - 1 \) degrees of freedom, and assuming independent samples from normally distributed populations.

As with tests for means, there are a couple of examples in the text about using these procedures. Please look them over, and ask me if you have any difficulty in following them. I will expect that you can perform one of these tests if the situation demands it.
Assessing Normality

As you may have noticed, all of the procedures that we have talked about for testing hypotheses and forming confidence intervals rely upon the fact that the underlying population is normal. However, in practice, there are many populations of interest to us which are not normal. Some of the procedures that we have talked about are Robust, meaning that departures from normality must be severe before the procedures are significantly effected. However, others, such as tests of variance, are much more sensitive. Thus, we must consider techniques for assessing if a particular population is normal.

Sample Quantiles

Consider a $q$ which is a number between 0 and 1. The $100q$th sample quantile, $y_q$, is an estimate of the $100q$th quantile of $Y$, denoted by $Y_q$, which is defined to be a value such that $\text{pr}(Y \leq Y_q) = q$. To find $y_q$:

- Order the $y_1, y_2, ..., y_n$ from smallest to largest. We will use the notation $y_{(1)}, y_{(2)}, ..., y_{(n)}$ to indicate the values in the ordered list. Thus, $y_{(1)}$ is the smallest value, $y_{(2)}$ is the second smallest, and so on.

- Write $(n + 1)q$ as a decimal. NOTE THE ERROR IN THE TEXT! IT SAYS $nq$! Let $w$ be the integer part, and $d$ be the decimal part.

- Then, the $100q$th sample quantile is $y_q = y_{(w)} + d(y_{(w+1)} - y_{(w)})$. 

24
• Notice that if \( w \) is 0 or \( n \), this quantity does not exist.

• If \( 100q \) is an integer, then \( y_q \) is called the \( 100q \)th percentile.

• Three of the quartiles have special names: the lower quartile \( (q_1 = y_{25}) \), the median \( (q_2 = y_{50}) \), and the upper quartile \( (q_3 = y_{75}) \).

• The interquartile range (IQR) is defined as \( q_3 - q_1 \).

To ensure that you remember how to do this, let's find \( q_1, q_2, q_3 \) for the sample 2, 6, 3, 5, 5, 4, 7, 3, 12.
Box Plots

One plot based upon the sample quantiles is called the box plot. This plot provides a simple, graphical way to look for outliers in the data. To construct a boxplot, we need to find the three quartiles, the interquartile range, and four fences. The four fences are the upper/lower inner/outer fences. They are defined as LOF = \( q_1 - 3\text{IQR} \), LIF = \( q_1 - 1.5\text{IQR} \), UIF = \( q_3 + 1.5\text{IQR} \), and UOF = \( q_3 + 3\text{IQR} \). Once these numbers are calculated, form a box plot by

- Make vertical lines at \( q_1, q_2, q_3 \). Connect the upper and lower endpoints with horizontal lines.

- Find the smallest data point larger than LIF and the largest data point smaller than UIF. Draw vertical lines at each. Connect these lines to the box with a single horizontal line.

- Mark each data point between LIF and LOF and those between UIF and UOF with a "*". Mark points outside of LOF and UOF with a "o".

- Points marked "*" are called mild outliers, while those marked "o" are called severe outliers.

- Outliers should be checked by the experimenter to determine if they represent real data or if they might have been measured or recorded incorrectly.
Let's construct a box plot for the data on slide 25.

**Normal Quantile Plots**

We can also compare the sample quantiles from data to those for a normal distribution. This can be done by plotting the sample quantiles from the data against the same quantiles from a standard normal distribution. If the points lie nearly on a line, the data appear normal; if the sample quantiles do not lie nearly on a line, this suggests that the data are not normal.

The approximate slope and intercept from a normal quantile plot have an interpretation, as well. We would expect the slope to be approximately . Similarly, we would expect the intercept to be approximately .

27
Look at the normal quantile plots below. Which samples appear to be from normal distributions?

There also exists a statistical test to determine if a sample is consistent with normality. This test is called the **Shapiro-Wilk test**. We will not talk about how it is derived, but we will look at the p-values for this test from software. For the four plots about these values are .9123, .1948, 1.58x10⁻⁷, and .7113, across the rows. For the record, only the first plot is actually data from a normal distribution.