Testing Hypothesis

Much of statistics is concerned with testing hypotheses against data using several standard techniques. At the core of these tests lies the concept of the "null hypothesis". A null hypothesis is set up and we use our tests to see if we can reject the null hypothesis, $H_0$. In other words, if we want to test whether two rock samples have different densities, we form the null hypothesis that they have equal densities and test if we can reject $H_0$. We will illustrate all this with an example:

It is claimed that the density of a particular sandstone is $2.35 \text{ gcm}^{-3}$. We are handed a sample of 50 specimens from an outcrop in the same area and decide to set the criteria that the samples are from another lithological unit if the sample mean is less than 2.25 or larger than 2.45. This is a clear-cut criterion for accepting or rejecting the claim that the samples are from the same units, but it is not infallible. Since our decision will be based on a sample, there is the possibility that the sample mean may be $< 2.25$ or $> 2.45$ even though the population mean $\mu$ is 2.35. We will therefore want to know what the chances are that we make a wrong decision.

We will investigate what the probability is that $x$ will be $< 2.25$ or $x > 2.45$ even if $\mu = 2.35$. Here, $s (= \sigma) = 0.42$. This probability is given by the area under the tails in Fig.2-1.

![Fig. 2-1. We reject the null hypothesis when the computed statistic falls in the tail area.](image)

Since $n = 50 \gg 30$ we will treat our sample as of infinite size. Then we have

$$s_x = \frac{s}{\sqrt{n}} = \frac{0.42}{\sqrt{50}} = 0.06$$

We can now evaluate the normal scores
We find the area under each tail to be \( 0.5 \left[ 1 + \text{erf} \left( \frac{-1.67}{\sqrt{\sigma_0^2}} \right) \right] \). Thus, the probability of getting a sample mean that falls in the tail area of the distribution is

\[
p = 2 \cdot 0.0475 = 0.095 \text{ or } 9.5%.
\]

This result means there is a 9.5% chance we will erroneously reject the hypothesis that \( \mu = 2.35 \) when it is in fact true. We call this committing a type I error.

Let us look at another possibility, where our test will fail to detect that \( \mu \) is not equal to 2.35. Suppose for the sake of argument that the true mean is 2.53. Then, the probability of getting a sample mean in the range 2.25 - 2.45 and hence erroneously accept the claim that \( \mu = 2.35 \) is given by the tail area in Fig. 2-2.

As before, \( s_\overline{x} = 0.06 \) so the normal scores become

\[
z_0 = \frac{2.25 - 2.53}{0.06} = -4.67 \quad z_1 = \frac{2.45 - 2.53}{0.06} = -1.333
\]

It follows that the area \( A = 0.5 \left[ 1 + \text{erf} \left( \frac{-1.333}{\sqrt{\sigma_0^2}} \right) - \text{erf} \left( \frac{-4.67}{\sqrt{\sigma_0^2}} \right) \right] = 0.092 \) or 9.2%. This is the risk we run of accepting the incorrect hypothesis \( \mu = 2.35 \). We call this committing a type II error.

We recognize that there are several possibilities when testing the null hypothesis. The table below summarizes the variations:

<table>
<thead>
<tr>
<th></th>
<th>Accept ( H_0 )</th>
<th>Reject ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 ) is TRUE</td>
<td>Correct Decision</td>
<td>Type I Error</td>
</tr>
<tr>
<td>( H_0 ) is FALSE</td>
<td>Type II Error</td>
<td>Correct Decision</td>
</tr>
</tbody>
</table>

If the hypothesis is true, but is rejected, we have committed a Type I error, and the probability of doing so is designated \( \alpha \). In our example, \( \alpha \) was 0.095. If our hypothesis is incorrect, but we still accept it, then we have committed a Type II error, and the probability of doing so is designated \( \beta \). In our case, with \( \mu = 2.53 \), \( \beta \) was 0.092.
Significance test

We saw in our example that the type II error probability depended on the value of $\mu$. Since $\mu$ is often not known, it is common to simply either reject $H_0$ or reserve judgment (i.e., never accept $H_0$). This way we avoid committing a type II error altogether, at the expense of never accepting $H_0$. We call this a significance test and say that the results are statistically significant if we can reject $H_0$. If not, the results are not statistically significant, and we attempt no further decisions. Hence, in statistics we can only disprove hypotheses, but never prove them...

Differences between means

We will often want to know if an observed difference in sample means can be attributed to chance. We will again use Student's $t$-test. It is assumed that the two distributions have the same variance but possibly different means.

We are interested in the distribution of $\bar{x}_1 - \bar{x}_2$, the difference in sample means. If the samples are independent and random, the difference distribution will be approximately normal with mean $\mu_1 - \mu_2$ and standard deviation

$$\sigma_\varepsilon = \sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(2.1)

where $\sigma_p^2$ is called the pooled variance:

$$\sigma_p^2 = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$$

(2.2)

We find the $t$-statistic by evaluating

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(2.3)

and test the hypothesis $H_0$: $\mu_1 = \mu_2$ based on the $t$-distribution for $\nu = n_1 + n_2 - 2$ degrees of freedom. For large $n_1$, $n_2$, the $t$-distribution becomes very close to a normal distribution and we may instead use $z$-statistics based on

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(2.4)

We will illustrate the two-sample $t$-test with an example: We have obtained random samples of magnetites from two separate outcrops. The measured magnetizations in Am²kg⁻¹ are

Outcrop 1: \{87.4, 93.4, 96.8, 86.1, 96.4\} \hspace{1cm} n_1 = 5
Outcrop 2: \{106.2, 102.2, 105.7, 93.4, 95.0, 97.0\} \hspace{1cm} n_2 = 6

We state our null hypothesis $H_0$: $\mu_1 = \mu_2$; the alternative hypothesis is of course $H_1$: $\mu_1 \neq \mu_2$. We decide to use 95% significance level, so $\alpha = 0.05$. In this case, $\nu = 5 + 6 - 2 = 9$, and a $t$-
statistics table (Appendix 2.4) shows that the critical $t$ value is 2.262, and we will reject $H_0$ is our $t$ exceeds this critical value. From the data we find

\[ \bar{x}_1 = 92.0 \text{ with } s_1 = 5.0 \]
\[ \bar{x}_2 = 99.9 \text{ with } s_2 = 5.5 \]

Using Eq. (2.3) we obtain

\[ t = \frac{99.9 - 92.0}{\sqrt{\frac{5 \cdot 5^2 + 4 \cdot 5.5^2}{9} \left( \frac{1}{5} + \frac{1}{6} \right)}} = 2.5 \]

Since $t > 2.262$ we must reject $H_0$. We conclude that the magnetizations at the two outcrops are not the same. We have now put confidence limits on sample means and compared sample means to investigate whether two populations have different means. We will turn our attention to inferences about the standard deviation.

Inferences about the standard deviation

The most popular way of estimating $\sigma$ is to compute the sample standard deviation. When investigating properties of $s$ and $\sigma$ we will be using the "chi-square" statistic

\[ \chi^2 = \frac{(n-1)s^2}{\sigma^2} \] (2.5)

The $\chi^2$ distribution depends on the degrees of freedom $\nu = n - 1$ and is restricted to positive values because of the power of 2. It portrays how the sample standard deviation would be distributed if we selected random samples of $n$ items. Fig. 2-3 shows a typical $\chi^2$ curve

In the same way we used $z_\alpha$ and $t_\alpha$, we now use $\chi^2_\alpha$ as the value for which the area to the right of $\chi^2_\alpha$ equals $\alpha$. Because the distribution is not symmetrical, we must evaluate the $\alpha/2$ and $1 - \alpha/2$ critical values separately. And in the same way we put confidence intervals on $\mu$, we now use (2.5) to find

\[ \chi^2_{n-1,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2} \]

or

\[ \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \] (2.6)
which gives the \( \alpha \) confidence interval on the variance. For large samples \((n > 30)\) this can be simplified to

\[
\frac{s}{1 + \frac{z_{\alpha/2}}{\sqrt{2n}}} < \sigma < \frac{s}{1 - \frac{z_{\alpha/2}}{\sqrt{2n}}}
\]  

(2.7)

Note that the confidence interval is not symmetrical about the sample standard deviation.

Testing standard deviations

We might want to test whether our sample standard deviation \( s \) is equal to or different from a given population \( \sigma \). In such a case the null hypothesis becomes \( H_0: s = \sigma \) with the alternative hypothesis \( H_1: s \neq \sigma \). As usual, we select our level of significance to be \( \alpha = 0.05 \).

Assume we have 15 estimates of temperatures with \( s = 1.3 \, ^\circ C \) and we want to know if \( s \) is any different from \( \sigma = 1.5 \, ^\circ C \) based on past experience. From \( \alpha = 0.05 \) and \( v = 14 \) we find the critical \( \chi^2 \) values from a table to be \( \chi^2_{0.025} = 26.119 \) and \( \chi^2_{0.975} = 5.63 \). Based on our sample statistic we compute

\[
\chi^2 = \frac{14 \cdot 1.3^2}{1.5^2} = 10.5
\]

We see that we cannot reject \( H_0 \) at the 95% significance level. Instead we may accept \( H_0 \) or reserve the judgment. This was a two-sided test since we must check that \( \chi^2 \) did not land in either of the two tails.

For large samples \( n \geq 30 \), the \( \chi^2 \) does not vary much with \( \nu \) and we may use the simpler statistic

\[
z = \frac{s - \sigma}{\sigma \sqrt{\frac{2}{n}}}
\]

and use the standard \( z \)-statistics table.

Testing two standard deviations

In the \( t \)-test for differences between means we assumed that the standard deviation of the two samples were the same. Often this is not the case and one should first test whether this assumption is valid. We want to know whether the two variances are different or not. The statistic that is most appropriate for such tests is called the \( F \)-statistic, defined as

\[
F = \begin{cases} 
\frac{s_1}{s_2}, & s_1 > s_2 \\
\frac{s_2}{s_1}, & s_2 > s_1 
\end{cases}
\]

(2.8)

For normal distributions this variance ratio is a continuous distribution called the \( F \) distribution. It depends on the two degrees of freedom \( v_1 = n_1 - 1 \) and \( v_2 = n_2 - 1 \). As before, we will reject the null hypothesis \( H_0: \sigma_1 = \sigma_2 \) at the \( \alpha \) level of significance and [possibly] accept the alternative \( \sigma_1 \neq \sigma_2 \) when our observed \( F \) statistic exceeds the critical value \( F_{\alpha/2} \).

Example: In our case of magnetic magnetizations we assumed that the \( \sigma \)'s were approximately the same. Let us now show that this is actually justified. We find
\[ F = \frac{5.5^2}{5.0^2} = 1.21 \]

From the table we find \( F_{0.025}(v_1 = 5, v_2 = 4) = 9.36 \). Hence we cannot reject \( H_0 \) and conclude that the difference in sample standard deviations is not statistically significant at the 95% level.

**The \( \chi^2 \) Test**

The last parametric test we shall be concerned with is the chi-squared test. It is a sample-based statistic using normal scores that is squared and summed up:

\[
\chi^2 = \sum_{i=1}^{n} \left( \frac{x - \mu}{\sigma} \right)^2
\]  
(2.9)

If we draw all possible samples of size \( n \) from a normal population and plotted \( \Sigma z^2 \), they would form the \( \chi^2 \) distribution mentioned earlier. The \( \chi^2 \) test is used to compare the shape of our data distribution to a distribution of known shape (usually a normal distribution).

The \( \chi^2 \) test is most often used on data that have been categorized or binned. Assuming that our observations have been binned into \( k \) bins, the test statistics is found as

\[
\chi^2 = \sum_{i=1}^{n} \left( \frac{O_j - E_j}{E_j} \right)^2
\]  
(2.10)

where \( O_j \) and \( E_j \) is the number of observed and expected values in the \( j \)'th bin. Note that this \( \chi^2 \) still is non-dimensional since we are using counts, even if the denominator is not squared. With counts, the probability that \( m \) out of \( n \) counts will fall in a given bin \( j \) is determined by the binomial distribution, with

\[
\mu = E_j = np_j
\]

and

\[
\sigma = \sqrt{np_j(1-p_j)} = \sqrt{np_j} = \sqrt{E_j}
\]

Plugging in for \( \chi^2 \) we find

\[
\chi^2 = \sum_{i=1}^{n} \left( \frac{x - \mu}{\sigma} \right)^2 = \sum_{i=1}^{n} \left( \frac{O_j - E_j}{\sqrt{E_j}} \right)^2 = \sum_{i=1}^{n} \left( \frac{O_j - E_j}{E_j} \right)^2
\]

As an example, consider the 48 measurements of salinity from Whitewater Bay in Florida (Table 2-1). We would like to know if these observations come from a normal distribution or not. The answer might have implications for models of mixing salt and freshwater.

The first step is to normalize the data into normal scores. We find \( x = 49.59 \) and \( s = 9.27 \) thus transfer all values to

\[
z_i = \frac{x_i - 49.54}{9.27}
\]

We choose to bin the data into 5 bins chosen such that the area under the curve for each bin is the same, i.e., 0.2. Using tables for the normal distribution, we find that the corresponding \( z \)-values for the intervals are \((-\infty, -0.84), (-0.84, -0.26), (-0.26, +0.26), (0.26, 0.84), (0.84, \infty)\). Counting
the values in Table 2-1 we find the observed number of samples for each of the 5 bins are 10, 11, 10, 5, and 12. These are \(O\)'s. The expected values \(E\) are all

\[ E_j = \frac{n}{k} = \frac{48}{5} = 9.6 \]

Using (2.10) we find the observed value \(\chi^2 = 3.04\)

The \(\chi^2\) - distribution depends on \(\nu\), the degrees of freedom, which normally is \(\nu = n - 1 = 4\) in our case. However, we used our observations to compute \(\bar{x}\) then \(s\). This reduces \(\nu\) by 2, leaving 2 degrees of freedom. From Appendix 2.6 we find the critical \(\chi^2\) for \(\nu = 2\) and \(\alpha = 0.05\) to be 5.99. Since this is much larger than our computed value we conclude that we cannot reject the null hypothesis that the salinities were drawn from a normal distribution at the 95% significance level.

We repeat that while we used a normal distribution in this example, the \(E_j\) could have represented any other distribution.

### Table 2-1 Standardized scores of salinity measurements from Whitewater Bay

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>Original Sample</th>
<th>Standardized Sample</th>
<th>Sample Number</th>
<th>Original Sample</th>
<th>Standardized Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46.00</td>
<td>-0.38</td>
<td>25</td>
<td>35.00</td>
<td>-1.57</td>
</tr>
<tr>
<td>2</td>
<td>37.00</td>
<td>-1.35</td>
<td>26</td>
<td>49.00</td>
<td>-0.06</td>
</tr>
<tr>
<td>3</td>
<td>62.00</td>
<td>1.34</td>
<td>27</td>
<td>48.00</td>
<td>-0.17</td>
</tr>
<tr>
<td>4</td>
<td>59.00</td>
<td>1.02</td>
<td>28</td>
<td>39.00</td>
<td>-1.14</td>
</tr>
<tr>
<td>5</td>
<td>40.00</td>
<td>-1.03</td>
<td>29</td>
<td>36.00</td>
<td>-1.46</td>
</tr>
<tr>
<td>6</td>
<td>53.00</td>
<td>0.37</td>
<td>30</td>
<td>47.00</td>
<td>-0.27</td>
</tr>
<tr>
<td>7</td>
<td>58.00</td>
<td>0.91</td>
<td>31</td>
<td>59.00</td>
<td>1.02</td>
</tr>
<tr>
<td>8</td>
<td>49.00</td>
<td>-0.06</td>
<td>32</td>
<td>42.00</td>
<td>-0.81</td>
</tr>
<tr>
<td>9</td>
<td>60.00</td>
<td>1.13</td>
<td>33</td>
<td>61.00</td>
<td>1.24</td>
</tr>
<tr>
<td>10</td>
<td>56.00</td>
<td>0.70</td>
<td>34</td>
<td>67.00</td>
<td>1.88</td>
</tr>
<tr>
<td>11</td>
<td>58.00</td>
<td>0.91</td>
<td>35</td>
<td>53.00</td>
<td>0.37</td>
</tr>
<tr>
<td>12</td>
<td>46.00</td>
<td>-0.38</td>
<td>36</td>
<td>48.00</td>
<td>-0.17</td>
</tr>
<tr>
<td>13</td>
<td>47.00</td>
<td>-0.27</td>
<td>37</td>
<td>50.00</td>
<td>0.05</td>
</tr>
<tr>
<td>14</td>
<td>52.00</td>
<td>0.27</td>
<td>38</td>
<td>43.00</td>
<td>-0.71</td>
</tr>
<tr>
<td>15</td>
<td>51.00</td>
<td>0.16</td>
<td>39</td>
<td>44.00</td>
<td>-0.60</td>
</tr>
<tr>
<td>16</td>
<td>60.00</td>
<td>1.13</td>
<td>40</td>
<td>49.00</td>
<td>-0.06</td>
</tr>
<tr>
<td>17</td>
<td>46.00</td>
<td>-0.38</td>
<td>41</td>
<td>46.00</td>
<td>-0.38</td>
</tr>
<tr>
<td>18</td>
<td>36.00</td>
<td>-1.46</td>
<td>42</td>
<td>63.00</td>
<td>1.45</td>
</tr>
<tr>
<td>19</td>
<td>34.00</td>
<td>-1.68</td>
<td>43</td>
<td>53.00</td>
<td>0.37</td>
</tr>
<tr>
<td>20</td>
<td>51.00</td>
<td>0.16</td>
<td>44</td>
<td>40.00</td>
<td>-1.03</td>
</tr>
<tr>
<td>21</td>
<td>60.00</td>
<td>1.13</td>
<td>45</td>
<td>50.00</td>
<td>0.05</td>
</tr>
<tr>
<td>22</td>
<td>47.00</td>
<td>-0.27</td>
<td>46</td>
<td>78.00</td>
<td>3.07</td>
</tr>
<tr>
<td>23</td>
<td>40.00</td>
<td>-1.03</td>
<td>47</td>
<td>48.00</td>
<td>-0.17</td>
</tr>
<tr>
<td>24</td>
<td>40.00</td>
<td>-1.03</td>
<td>48</td>
<td>42.00</td>
<td>-0.81</td>
</tr>
</tbody>
</table>
Non-parametric tests

Last time we finished up looking at the standard parametric tests, i.e., the $t$, $F$, and $\chi^2$ tests. We justified using these tests by either having large samples and invoke the central limits theorem, or simply assuming that the distribution we have sampled is approximately normal. Sometimes, however, none of these conditions are met. The two cases are:

- Small samples ($n < 30$) and you cannot assume population is normal
- Any size sample of ordinal data (which can only be ranked, not operated on numerically)

In those cases we apply non-parametric methods which make no assumptions about the form of the data distribution.

Mann - Whitney test

This test is a non-parametric alternative to the two-sample Student $t$-test. It also goes by the names Wilcoxon test and the $U$-test. The Mann-Whitney test is performed by combining the two data sets we want to compare, sort them into ascending order, and assign each point a rank: Smallest value is given rank = 1; the largest observation is ranked $n_1 + n_2$. Should some of the observations be identical, one assigns the average rank to all these values. E.g. if the 7th and 8th sorted values are identical, we assign to each the rank 7.5. The idea here is that if the samples consist of random drawings from the same population one would expect the ranks for both samples to be scattered more-or-less uniformly through the sequence.

After arranging the data, we add up the ranks for each data set into rank sums which we denote $W_1$ and $W_2$. The sum of $W_1 + W_2$ must obviously equal the sum of the first $(n_1 + n_2)$ integers which is

$$\frac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1).$$

Many early rank sum tests were based on $W_1$ or $W_2$ but now it is customary to use the statistic $U$ defined as

$$U_1 = n_1n_2 + \frac{1}{2}n_1(n_1 + 1) - W_1$$

or

$$U_2 = n_1n_2 + \frac{1}{2}n_2(n_2 + 1) - W_2$$

or simply $U$, the smallest of $U_1$ and $U_2$. This statistic takes on values from 0 to $n_1n_2$ and its sampling distribution is symmetrical about $n_1n_2/2$. The test then consists of comparing the calculated $U$ statistic to a critical $U$ value given the sample sizes and desired level of significance $\alpha$.

Example: We want to compare the grain size of sand obtained from two different locations on the moon on the basis of measurements of grain diameters in mm as follows.

Location 1: 0.37, 0.70, 0.75, 0.30, 0.45, 0.16, 0.62, 0.73, 0.33 \hspace{1cm} n_1 = 9
Location 2: 0.86, 0.55, 0.80, 0.42, 0.97, 0.84, 0.24, 0.51, 0.92, 0.69 \hspace{1cm} n_2 = 10
We do not know what distribution the grain sizes of sand on the moon follow so we choose the U-test to see if the mean grain size differs in the two samples.

Computing the means gives 0.49 and 0.68. If we wanted to use the t-test we would have to assume that the underlying distributions are normal. The U-test requires no such assumptions. We start by arranging the data jointly in ascending order and keep track of which sample each point originated from:

<table>
<thead>
<tr>
<th>Data</th>
<th>0.16</th>
<th>0.24</th>
<th>0.30</th>
<th>0.33</th>
<th>0.37</th>
<th>0.42</th>
<th>0.45</th>
<th>0.51</th>
<th>0.55</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.62</th>
<th>0.69</th>
<th>0.70</th>
<th>0.73</th>
<th>0.75</th>
<th>0.80</th>
<th>0.84</th>
<th>0.86</th>
<th>0.92</th>
<th>0.97</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
</tbody>
</table>

We first evaluate the rank sum for sample 1, giving $W_1 = 69$, from which it follows that

$$W_2 = \frac{19 \cdot 20}{2} - W_1 = 190 - 69 = 121$$

We now form the null hypothesis $H_0: \mu_1 = \mu_2$, with $H_1: \mu_1 \neq \mu_2$, and state the level of significance $\alpha = 0.05$. From a table with critical values for $U$ we find $U_{\alpha}(9, 10) = 20$. We will reject the null hypothesis if $U$ is $\leq 20$. From $W_1$ and $W_2$ we find

$$U_1 = 9 \cdot 10 + \frac{9 \cdot 10}{2} - 69 = 66$$

$$U_2 = 9 \cdot 10 + \frac{10 \cdot 11}{2} - 121 = 24$$

and hence $U = \min(66, 24) = 24$. This is larger than the critical value of 20, suggesting we cannot reject the null hypothesis. In other words, the observed difference in grain size means is not statistically significant at the 95% significance level.

For large samples ($n_1, n_2 > 30$) things again simplify and it can be shown that the mean and standard deviation of the $U_1$ sampling distribution are

$$\mu_U = \frac{n_1 n_2}{2} \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2)}{12}} \quad (2.13)$$

We could then form the z-score as $z = \frac{U - \mu_U}{\sigma_U}$ and use the familiar critical values $\pm z_{\alpha/2}$, or simply use the standard t-test since we have a large sample.

**Kolmogorov - Smirnov**

Another very useful non-parametric method is the Kolmogorov - Smirnov (K-S) test. It is a test for goodness of fit or shape, and is often used instead of the $\chi^2$-test. A big advantage of the K-S test over the $\chi^2$ is that one does not have to bin the data, which is an arbitrary procedure anyway (how do you select bin size and why?). In the K-S test we convert the data distribution to a cumulative distribution $S(x)$. $S(x)$ then gives the fraction of data points to the "left" of $x$. 
While different data sets will generally have different distributions, all cumulative distributions agree at the smallest \( x \) \((S(x) = 0)\) and the largest \( x \) \((S(x) = 1)\). Thus, it is the behavior between these points that sets distributions apart. There is of course an infinite number of ways to measure the overall difference between two cumulative distributions: We could look at absolute value of the area between the curves, the mean square difference, etc. The K-S statistic is very simple: It consists of the maximum value of the absolute difference between the two cumulative curves. Thus, comparing two cumulative distributions \( S_1(x) \) and \( S_2(x) \) one K-S statistic becomes

\[
D = \max_{-\infty < x < \infty} |S_1(x) - S_2(x)|
\]

(2.14)

Note that \( S_2 \) may be another or a given cumulative probability function like the normal distribution. The distribution of the K-S statistic itself can be calculated under the assumption that \( S_1 \) and \( S_2 \) are drawn from the same distribution, thus providing critical values for \( D \). We will use the K-S test on the salinity measurements we looked at previously.

After computing the normal scores, we plot the cumulative function on the same graph as that of a normal cumulative distribution. Inspecting the graph we find the maximum absolute difference at \( z = 0.37 \), which corresponds to the 53 ppt sample. The \( D \) estimate is 10.70 - 0.641 = 0.06. Based on a significance level of \( \alpha = 0.10 \) and \( n = 48 \), the critical K-S value is 0.17, much larger than observed. Hence we cannot reject the null hypothesis that the samples were collected from a normally distributed population.

Tests of Correlation Coefficients

There are both parametric and non-parametric tests for the linear correlation coefficient \( r \). We will look at both kinds.

Traditional (Least-squares) Correlation

We recall that the conventional correlation coefficient was defined by

\[
r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

(2.15)

Often, we need to test if \( r \) is significant. In such tests, \( r \) is our sample-derived estimate of \( \rho \), the actual correlation of the population. The most useful null hypothesis is \( H_0: \rho = 0 \). It can be shown that the sampling distribution of \( r \) for a population that has zero correlation \( (\rho = 0) \) has mean \( \mu = 0 \) and \( \sigma = \sqrt{1 - r^2 / n - 2} \). Hence, a \( t \)-statistic can be calculated as

\[
t = \frac{r - \mu}{\sigma} = \frac{r}{\sqrt{(1 - r^2) / (n - 2)}} = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}
\]

(2.16)

The degrees of freedom, \( \nu \), is \( n - 2 \). Suppose we rolled a pair of dice, one red and one green (Table 2-2). Using (2.15) we obtain \( r = 0.66 \) which seems quite high, especially since there is no reason to believe a correlation should exist at all. Let us test to see if the correlation is significant. Choosing \( \alpha = 0.05 \) we find critical \( t_{\alpha/2} = 3.182 \). Applying (2.16) gives the observed
$t = 1.52$, hence the correlation of 0.66 is most likely caused by random fluctuations of small samples and we cannot reject $H_0$.

<table>
<thead>
<tr>
<th>Red (x)</th>
<th>Green (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2-2. Examples of rolling a pair of dice.

How high would $r$ have to be for us to find it significant and commit a type I error by rejecting the (true) null hypothesis? We must solve for $r$ in

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \Rightarrow 3.182 = \frac{3r}{1-r^2} \Rightarrow r = \pm 0.88$$

So, if $r$ equals or exceeds $\pm 0.88$ we would find ourselves concluding that red and green dice give correlated pairs of values...

Non-parametric Correlation

Finally, we will look at non-parametric correlation called rank correlation or Spearman's rank correlation, denoted by $r_s$. The rank correlation is carried out by ranking the $x_i$'s and $y_i$'s separately, then finding the difference in rank $d_i$ between $x_i$ and $y_i$ pairs, and evaluate $r_s$ as

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2.17)$$

In the case where the null Hypothesis $H_0$: no correlation is true, the sampling distribution of $r_s$ has mean $r_s - 0$ and standard deviation $\sqrt{n-1}$. We can therefore base our statistics on

$$z = \frac{r_s - 0}{\sqrt{n-1}} = r_s \sqrt{n-1} \quad (2.18)$$

and compare this $z$-value to critical $z_{\alpha/2}$ values. Ranking the dice data gives

<table>
<thead>
<tr>
<th>Red (x)</th>
<th>Rank x</th>
<th>Green (y)</th>
<th>Rank y</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3.5</td>
<td>5</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>3.5</td>
<td>6</td>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Using (2.17) we find $r_s = 0.65$ (surprisingly similar to what we found using (2.15)). The $z$-statistic from (2.18) becomes $z = 1.3$, which is way inside the 95% confidence limits for a normal distribution ($\pm 2$). Hence, we again arrive at the same conclusion that we cannot reject $H_0$. 