Estimation and Hypothesis Testing
Estimation vs Testing

Often the questions of interest are ones of estimation

What is the mean fuel-economy of this engine?
By how much does this new drug delay relapse?
How old is the species *Homo sapiens*?

Hypothesis testing is concerned with questions such as

Does this sample come from a population with mean $\mu_0$?
Are these two samples from the same population?
Does the new drug delay relapse?

There is a null hypothesis $H_0$ and usually some specification of the alternative hypothesis $H_1$. 
Parameters and Statistics

Both estimation and testing are concerned with a parameter $\theta$, which should (if possible) be a meaningful quantity.

We assume we have some sample $x = (x_1, x_2, \ldots, x_n)$ from a random variable $X$ whose distribution depends on $\theta \in \Theta$.

A statistic $t = t(x)$ is any number calculated from the sample. Since the sample is a random observation of $X_1, X_2, \ldots, X_n$, we can regard $t$ as a sample of the random variable $T = t(X)$. The distribution of $T$ is called the sampling distribution.
Estimation

A statistic $T$ is an estimator of $\theta$ if its intention is to be close to the (unknown) value of $\theta$. To perform statistical inference for an estimator $T$ of $\theta$ we will often need to derive its distribution.

Suppose the population has mean $\mu$ and variance $\sigma^2$. Then we can often use

$$\bar{X}_n \sim N(\mu, \sigma^2/n).$$

This holds exactly if the population is normal, and approximately otherwise (by the Central Limit Theorem).
Normal Theory

If $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

are independent random variables. In addition,

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Hence, in the normal case, we know the sampling distributions of $\bar{X}_n$ and $S^2$ if $\sigma^2$ is known.
Student’s t distribution

From the first result,

\[
\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).
\]

However, in general \( \sigma^2 \) will not be known. We can not simply replace \( \sigma \) by its estimate \( s \), as this will increase the variability of the left-hand-side.

It can be shown that

\[
\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}.
\]

The density function of a \( t \)-distribution is similar to a normal distribution but with fatter tails.
Confidence Intervals

Suppose we know the distribution of $t(X) - \theta$ for some estimator $T$. Then given an estimate $t(x)$ we can ask how near it is likely to be to the ‘true’ value, $\theta$.

Denote by $q_\alpha$ the $\alpha$th quantile of the distribution of $t(X) - \theta$. Then we have

$$P \left( q_{\alpha/2} < t(X) - \theta < q_{1-\alpha/2} \right) = 1 - \alpha,$$

so the probability

$$P \left( t(X) - q_{1-\alpha/2} < \theta < t(X) - q_{\alpha/2} \right) = 1 - \alpha.$$

The interval

$$(t(x) - q_{1-\alpha/2}, t(x) - q_{\alpha/2})$$

is called a $100(1 - \alpha)\%$ confidence interval for $\theta$. 

Example

Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$,
\[ \mu = 0.7, \sigma = 1, n = 10 \]
\[
\mu = 0.7, \sigma = 1, n = 30
\]
Robust Estimation

Generally we assume that the distribution $f$ generating the data has a particular form. If this assumption is not correct, then the conclusions we draw from inference may not be accurate.

In robust estimation we seek an estimator which is almost as efficient as the best estimator if the assumptions made about $f$ are true, but does as well as possible in the presence of departures from the assumed models, in particular long-tailed distributions which give rise to outliers.

Outliers are sample values that cause surprise in relation to the majority of the sample. Outlying observations may be correct, but they should always be checked for transcription errors.

It can be difficult or impossible to spot outliers, and if we remove outliers for no reason then we will underestimate variances.
Breakdown Points

Estimates such as the sample mean and variance are highly sensitive to outliers, and can be made arbitrarily large by changing just one sample value.

The sensitivity of an estimator can be measured by the breakdown point. Informally, the *breakdown point* $\epsilon^*$ is the largest fraction of the data that can be moved arbitrarily without perturbing the estimator to the boundary of the parameter space. The higher the breakdown point, the more robust the estimator is against outliers.

A common robust estimator of location is the trimmed mean. The *trimmed mean* is the mean of the remaining observations when the largest and smallest $\alpha n$ observations are removed. The breakdown point of the $\alpha$-trimmed mean is $\alpha$. Note that the median is a special-case of a trimmed mean.

When assessing a robust estimator it is important to consider properties other than breakdown point as well.
M-estimators

Maximum likelihood estimators (MLEs) are not always very robust, and this inspired the more general *M-estimators*.

Consider estimators of a location parameter $\mu$. The MLE would solve

$$
\min_{\mu} \sum_i \log f(x_i - \mu).
$$

However, we could solve this equation for more general functions $\rho$ which do not correspond to density functions. An *M-estimator* solves

$$
\min_{\mu} \sum_i \rho(x_i - \mu),
$$

and their properties will depend on the choice of function $\rho$. 
Let $\psi = \rho'$ if this exists. Then the M-estimator $\hat{\mu}$ satisfies

$$\sum_{i} \psi(x_i - \hat{\mu}) = 0.$$

In general, the function $\rho$ is chosen to be a compromise between the squared difference and the absolute difference.
Common choices of $\rho$ result in the following functions $\psi$: 

- **Trimmed mean**
- **Huber**
- **Tukey bisquare**
- **Hampel**
Hypothesis Testing

Suppose we are interested in the value of some parameter $\theta$, and we have a hypothesis that $\theta \in \Omega_0$. This will often be the hypothesis of “no change” or “no effect”, in which case it is called the null hypothesis, denoted $H_0$.

For example, suppose we have measured the IQ of a number of university students. Our null hypothesis might be that the mean IQ of university students is equal to the population average of 100.

The alternative hypothesis is that $\theta \in \Omega \setminus \Omega_0$. For example, it might be simple, one-sided or two-sided.

A test of a statistical hypothesis is a rule for rejecting $H_0$, based on the observations $x_1, x_2, \ldots, x_n$.

We only consider tests of simple hypotheses, $H_0 : \theta = \theta_0$, say.
For example, suppose the IQ scores of university students are assumed to be normally distributed. Then, if the null hypothesis is true, we know that

\[ T = \frac{\bar{X} - 100}{S/\sqrt{n}} \]

has a \( t \)-distribution with \( n - 1 \) degrees of freedom. Hence if the null hypothesis is true, the observed value of \(|T|\) will be greater than the quantile \( t_{n-1,1-\alpha/2} \) with probability \( \alpha \).

\( T \) is an example of a test statistic - a statistic whose distribution under the null-hypothesis is known.

Given a test statistic \( T \) and an observed value \( t \), we can then proceed in either of two equivalent ways:
1. Calculate the *p*-value

\[ p = P(|T| > t|\theta = \theta_0), \]

the probability under the null-hypothesis of observing a value of the test statistic at least as extreme as \( t \). A small value of \( p \) is evidence against the null-hypothesis.

2. Set a *significance level* \( \alpha \) for the test and determine the *critical value* \( c \) such that

\[ P(|T| > c|\theta = \theta_0) \leq \alpha. \]

Reject \( H_0 \) if the observed value of \(|T|\) is greater than \( c \).

We reject the null hypothesis at level \( \alpha \) if and only if the p-value is less than or equal to \( \alpha \). A \( 100(1 - \alpha)\% \) confidence interval for \( \theta \) will not contain \( \theta_0 \) if and only if we reject the null-hypothesis at level \( \alpha \).
Note that there is no “magic” significance level at which null hypotheses should be automatically rejected. Often people will use a significance level of $\alpha = 0.05$ at which they reject the null hypothesis, however this is completely arbitrary. The significance level used should always take into account the consequences of rejecting the null-hypothesis.

At the least you should qualify any statements, for example

0.05 < $p$ ≤ 0.06  “weak evidence for rejection”
0.03 < $p$ ≤ 0.05  “reasonable evidence for rejection”
0.01 < $p$ ≤ 0.03  “good evidence for rejection”
0.001 < $p$ ≤ 0.01  “strong evidence for rejection”
p ≤ 0.001  “very strong evidence for rejection”

Beware of confusing statistical and practical significance.
One-sample Test of Location

Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, where both $\mu$ and $\sigma^2$ are unknown. To test $H_0 : \mu = \mu_0$ against the alternative, we use the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Under $H_0$, $T \sim t_{n-1}$.

For example, the IQ measurements of 8 university students are

$$\text{IQ} : 118, 121, 96, 102, 93, 110, 117, 131$$

We wish to test $H_0 : \mu = 100$ against $H_1 : \mu > 100$. The value of the test statistic is 2.36 and the p-value is

$$P(T > 2.36 | \mu = 100) = 0.025.$$  

Hence, there is good evidence for rejection of the null hypothesis.
Two-sample tests

The paired t-test

Suppose we have paired data \((X_i, Y_i), i = 1, 2, \ldots, n\) and that their differences \(D_i = X_i - Y_i\) are iid \(N(\mu, \sigma^2)\), with \(\mu\) and \(\sigma^2\) unknown. To test the null hypothesis \(H_0: \mu = 0\) against the alternative we can use the test statistic

\[
T = \frac{\bar{D}}{S_D/\sqrt{n}},
\]

which under \(H_0\) has a \(t_{n-1}\) distribution.

For example (Box, Hunter & Hunter) two types of rubber were randomly assigned to the left and right shoes of 10 boys and the relative wear on each measured. We wish to test the hypothesis of no difference in wear, \(H_0: \mu = 0\), against \(H_1: \mu \neq 0\).
<table>
<thead>
<tr>
<th>Boy</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>13.2</td>
<td>8.2</td>
<td>10.9</td>
<td>14.3</td>
<td>10.7</td>
<td>6.6</td>
<td>9.5</td>
<td>10.8</td>
<td>8.8</td>
<td>13.3</td>
</tr>
<tr>
<td>B</td>
<td>14.0</td>
<td>8.8</td>
<td>11.2</td>
<td>14.2</td>
<td>11.8</td>
<td>6.4</td>
<td>9.8</td>
<td>11.3</td>
<td>9.3</td>
<td>13.6</td>
</tr>
<tr>
<td>Diff</td>
<td>-0.8</td>
<td>-0.6</td>
<td>-0.3</td>
<td>0.1</td>
<td>-1.1</td>
<td>0.2</td>
<td>-0.3</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.3</td>
</tr>
</tbody>
</table>

From the data we have $\bar{d} = -0.41$ and $s_D^2 = 0.15$, so $T = -3.34$. The p-value

$$P(|T| > 3.34|\mu = 0) = 0.0085,$$

so there is strong evidence to reject the hypothesis that wear is the same for both types of rubber.
The two-sample t-test

Suppose we want to compare the locations of two samples $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$, and assume that $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_1 \sim N(\mu_2, \sigma_2^2)$.

Firstly, we will consider the case that we know (or have tested) that the variances are the same, so $\sigma_1^2 = \sigma_2^2 = \sigma$, say. We are interested in testing the hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$.

In this case

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

so the test statistic

$$T = \frac{\bar{x} - \bar{y} - \delta_0}{s \sqrt{1/n_1 + 1/n_2}},$$

where $s^2$ is the pooled variance $[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$, has a $t_{n_1+n_2-2}$ distribution under $H_0$. 
If the variances are not the same, then we can use the test statistic

\[
\frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},
\]

whose distribution under \( H_0 \) can be approximated by a \( t_\nu \) distribution, where \( \nu \) is given by an expression involving the \( n_i \) and \( s_i^2 \) (this is the default in R).

For example, consider the results of two experiments (which we know should have the same variance) to measure the concentration of a chemical:

<table>
<thead>
<tr>
<th>Exp. 1</th>
<th>22 19 35 11 21 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 2</td>
<td>33 11 20 38</td>
</tr>
</tbody>
</table>

We wish to test \( H_0 : \mu_1 = \mu_2 \) against \( H_1 : \mu_1 \neq \mu_2 \). The test statistic \( T = -0.87 \) which on comparing with a \( t_8 \) distribution gives a p-value of 0.41. Hence there is not enough evidence to reject \( H_0 \).
Test of equal variances

Suppose that we have two samples as before, where $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_1 \sim N(\mu_2, \sigma_2^2)$. We wish to test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$. In this case we can use the test statistic

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

which has an $F_{n_1-1, n_2-1}$ distribution under the null hypothesis (“Snedecor’s F”). The further the value of the test statistic is from 1, the stronger the evidence for unequal variances.

Returning to the chemical example, we might wish to test $H_0 : \sigma_1 = \sigma_2$ against $H_1 : \sigma_1 \neq \sigma_2$. In this case the test statistic $s_1^2 / s_2^2 = 0.545$, and comparing this with a $F_{5,3}$ distribution gives a p-value of 0.52.
Robustness of the t-tests

We have demonstrated that the sample mean can be quite close to normality even if the original population is not too close to normal. In general, the “t-tools” work well unless outliers upset the calculation of $s$.

Nevertheless, alternatives have been developed for continuous distributions. These are generally referred to as *non-parametric tests* or *distribution free tests*. 
Tests of Location Zero

Consider again pairs of data \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\). Assume that the differences \(X_i - Y_i\) are iid and come from a continuous distribution which is symmetric about \(\theta\). The null hypothesis is that \(\theta = 0\).

**sign test**

Under the null hypothesis, positive and negative differences are equally likely, so the number of positive values follows a binomial distribution with parameters \(n\) and 0.5. Hence the p-value can be easily calculated.
Wilcoxon signed-rank test

In this test we first calculate the absolute value of the differences and rank them. The test statistic is the sum of the ranks for the differences with positive sign. Extreme values of this statistic (large or small) indicate departure from the null hypothesis.

The p-value can be calculated exactly for small samples using the permutation distribution, whilst for large samples a normal approximation to the sampling distribution is used.
For example, To test the hypothesis that IQ is not intrinsic, a group of students were instructed to take an IQ test before and after completing an “IQ test training course”. We wish to test the null hypothesis that the differences come from a distribution which is symmetric about $\theta = 0$ against the alternative that $\theta > 0$. The data are presented in the table below.

<table>
<thead>
<tr>
<th>IQ before</th>
<th>118</th>
<th>121</th>
<th>96</th>
<th>102</th>
<th>93</th>
<th>110</th>
<th>117</th>
<th>131</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ after</td>
<td>110</td>
<td>122</td>
<td>110</td>
<td>104</td>
<td>102</td>
<td>105</td>
<td>115</td>
<td>132</td>
</tr>
<tr>
<td>abs(diff)</td>
<td>8</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>sign(diff)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>rank</td>
<td>6</td>
<td>1.5</td>
<td>8</td>
<td>3.5</td>
<td>7</td>
<td>5</td>
<td>3.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

The test statistic is 21.5. Using tables of critical values we find that the p-value $> 0.05$. Using R, the p-value is given as 0.34.
Two-sample non-parametric Tests

Wilcoxon rank-sum test

This test (also known as the Mann-Whitney test) is used to test for location shifts between two independent samples.

Firstly we combine the observations from the two samples and rank them. If the samples are of size $n_1$ and $n_2$ respectively, then the test statistic is the sum of the ranks of the observations from the first sample, minus $n_1(n_1 + 1)/2$.

For the chemical data, we have

<table>
<thead>
<tr>
<th>ordered results</th>
<th>10</th>
<th>11</th>
<th>11</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>33</th>
<th>35</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>experiment</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>rank</td>
<td>1</td>
<td>2.5</td>
<td>2.5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Hence the value of the test statistic is $29.5 - 21 = 8.5$. 
For small samples without ties we can find the permutation distribution of
the test statistic exactly, otherwise a large-sample normal approximation
is used. In this example the p-value is 0.52, so we do not reject the null
hypothesis.

Both the Wilcoxon signed-rank and rank-sum tests can be implemented in
R using the function `wilcox.test`.
Goodness-of-fit tests

These are used to test if the data came from some hypothesised distribution. For continuous data one can use the Kolmogorov-Smirnov test (ks.test), the Anderson-Darling test or the Shapiro-Wilk test (shapiro.test) for normality. However, visual inspection is generally adequate.

Chi-square test

The chi-square test is used to test the hypothesis that the data came from some specific distribution. It requires binned data and so can be applied to discrete distributions. For the chi-square approximation to be valid there should be no less than about 5 observations in each bin.

The test statistic is

\[ X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \]

where \( O_i \) denotes the observed frequency in bin \( i \) and \( E_i \) denotes the frequency expected in bin \( i \) under \( H_0 \).
Under $H_0$, $X^2$ has approximately a $\chi^2$ distribution with $k - p$ degrees of freedom, where $p$ is the number of parameters of the distribution.

For example, suppose we wish to test if the number of defective items coming off a manufacturing plant per day can be assumed to have a Poisson distribution. The data collected over the previous two months can be summarised as

<table>
<thead>
<tr>
<th>count</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>freq</td>
<td>0</td>
<td>7</td>
<td>15</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

and the MLE of the parameter $\mu$ is 3.8. Hence we have

<table>
<thead>
<tr>
<th>count</th>
<th>0–1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>$\geq 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_i$</td>
<td>7</td>
<td>15</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>$E_i$</td>
<td>6.6</td>
<td>9.9</td>
<td>12.5</td>
<td>11.9</td>
<td>9.0</td>
<td>5.7</td>
<td>5.4</td>
</tr>
</tbody>
</table>

and the test statistic $X^2 = 5.73$. This gives a p-value of 0.45.
Contingency Tables

Contingency tables are counts of cross-tabulations on two variables. For example, the caffeine data

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1-150</th>
<th>151-300</th>
<th>&gt; 300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Married</td>
<td>652</td>
<td>1537</td>
<td>598</td>
<td>242</td>
</tr>
<tr>
<td>Prev. Married</td>
<td>36</td>
<td>46</td>
<td>38</td>
<td>21</td>
</tr>
<tr>
<td>Single</td>
<td>218</td>
<td>327</td>
<td>106</td>
<td>67</td>
</tr>
</tbody>
</table>

We may want to test the null-hypothesis that the row and column categories are independent.

Under the null hypothesis, the expected count in cell $i, j$ is $E_{ij} = n p_i.p_j$, where $p_i.$ is the proportion in row $i$ and $p.j.$ is the proportion in column $j$.

We can then apply a goodness-of-fit test with test statistic for an $r \times c$ table

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(r-1)(c-1).$$

For the caffeine data $X^2 \approx 51.65$ and the p-value is essentially zero.