Chapter 5

Statistical models, estimation, and confidence intervals
A statistical model describes the outcome of a variable, the response variable, in terms of explanatory variables and random variation. For linear models the response is a quantitative variable. The explanatory variables describe the expected values of the response and are also called covariates or predictors. An explanatory variable is quantitative with values having an explicit numerical interpretation as in linear regression model. On the other hand, there are no explanatory variables for single variable data. The response is the measurement of variable, but there is no additional information about the measurement as in one sample model.
The response $y$ is a function $\mu$ of explanatory variables plus a remainder term $e$ describing the random variation. In general, let

$$y_i = \mu_i + e_i, \quad i = 1, \ldots, n.$$  

Here, $\mu_i$ includes the information from the explanatory variables and is a function of parameters $\theta_1, \ldots, \theta_p$ and explanatory variable $x$

$$\mu_i = f(x_i; \theta_1, \ldots, \theta_p). \quad (5.1)$$

The remainder terms $e_1, \ldots, e_n$ are assumed to be independent and $N(0, \sigma^2)$ distributed, all with mean zero and standard deviation $\sigma$. 
Linear regression model

The linear regression has the size $p = 2$ of parameters, and $(\theta_1, \theta_2)$ correspond to $(\alpha, \beta)$. The function $f(x_i; \alpha, \beta) = \alpha + \beta \cdot x_i$ determines the model

$$y_i = \alpha + \beta \cdot x_i + e_i \quad i = 1,\ldots,n,$$

(5.3)

where $e_1, \ldots, e_n$ are independent and $N(0,\sigma^2)$ distributed. Or, equivalently, the model assumes that $y_1, \ldots, y_n$ are independent. The parameters of the model are $\alpha$, $\beta$, and $\sigma$. The slope parameter $\beta$ is the expected increment in $y$ as $x$ increases by one unit, whereas $\alpha$ is the expected value of $y$ when $x = 0$. This interpretation of $\alpha$ does not always make biological sense as the value zero of $x$ may be nonsense. The remainder terms $e_1, \ldots, e_n$ represent the vertical deviations from the straight line. The assumption of variance homogeneity means that the typical size of these deviations is the same across all values of $x$. 
One sample model

In *one sample* case, there is no explanatory variable (although there could have been one such as sex or age). We assume that $y_1, \ldots, y_n$ are independent with

$$y_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n. \quad (5.5)$$

All observations are assumed to be sampled from the same distribution. Equivalently, we may write

$$y_i = \mu + e_i, \quad i = 1, \ldots, n. \quad (5.5^*)$$

where $e_1, \ldots, e_n$ are independent and $N(0, \sigma^2)$ distributed. The parameters are $\mu$ and $\sigma^2$, where $\mu$ is the expected value (or population mean) and $\sigma$ is the average deviation from this value (or the population standard deviation).
Estimation (technical)

We are interested in the values of the parameters, and we will use the data to compute their estimates. Estimation is the process of finding the values of the parameters that make the model fit the data the best. We already carried out least squares estimation for linear regression. Here we introduce a general idea by means of squared residuals. Given values of the parameters $\theta_1, \ldots, \theta_p$, the expected value of $y_i$ becomes

$$f(x_i; \theta_1 \ldots, \theta_p).$$

If the model fits well to the data, then the observed values $y_i$ and the expected values should all be “close” to each other. Hence, the sum of squared deviations is a measure of the model fit:

$$Q(\theta_1, \ldots, \theta_p) = \sum_{i=1}^{n} (y_i - f(x_i; \theta_1, \ldots, \theta_p))^2$$
Estimation (technical)

The least squares estimates

\[ \hat{\theta}_1, \ldots, \hat{\theta}_p \]

are the values for which

\[ Q(\theta_1, \ldots, \theta_p) = \sum_{i=1}^{n} (y_i - f(x_i; \theta_1, \ldots, \theta_p))^2 \]

is the smallest possible. Expressions for the estimators are obtained by deriving partial derivatives and solving the equations. It turns out that the solution actually corresponds to a minimum point. Note that we actually use the data to compute their estimates by the least squares method. Different samples lead to different estimates, and so we need to be concerned with the variability and precision of the estimates.
Residual variance (technical)

We get the fitted values, the residuals, and the residual sum of squares (denoted SS).

\[ \hat{y}_i = \hat{\mu}_i = f(x_i; \hat{\theta}_1, \ldots, \hat{\theta}_p), \quad r_i = y_i - \hat{y}_i, \quad SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \] (5.6)

The fitted value is the value we would expect for the \( i \)th observation if we repeated the experiment, and we can think of the residual \( r_i \) as a “guess” for \( e_i \) since \( e_i = y_i - \mu_i \). Therefore, it seems reasonable to use the residuals to estimate the variance \( \sigma^2 \).

\[ s^2 = \hat{\sigma}^2 = \frac{SS_e}{n - p} = \frac{1}{n - p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \] (5.7)

The number \( n - p \) in the denominator is called the residual degrees of freedom, or residual df, and \( s^2 \) is often referred to as the residual mean squares, the residual variance, or the mean squared error (denoted MS).
The estimate of standard deviation $\sigma$ is given by the residual standard deviation $s$; i.e., the square root of $s^2$.

$$\hat{\sigma} = s = \sqrt{s^2} = \sqrt{\frac{SS_e}{df_e}} \quad (5.8)$$

The statistical model results in the estimate of parameters $\theta_1, \ldots, \theta_p$. The standard deviation of the estimate is called the standard error, and given by

$$SE(\hat{\theta}_j) = s \sqrt{k_j} \quad (5.12)$$

The constant $k_j$ depends on the model and the data structure—but not on the observed data. The value of $k_j$ could be computed even before the experiment was carried out. In particular, it decreases when the sample size $n$ increases.
Student t distribution

If we replace $\sigma$ with its estimate $s$ and consider instead

$$T = \frac{\sqrt{n}(\bar{y} - \mu)}{s},$$

- **Symmetry.** The distribution of $T$ is symmetric around zero, so positive and negative values are equally likely.

- **Dispersion.** Values far from zero are more likely for $T$ than for $Z$ due to the extra uncertainty. This implies that the interval should be wider for the probability to be retained at 0.95.

- **Large samples.** When the sample size increases then $s$ is a more precise estimate of $\sigma$, and the distribution of $T$ more closely resembles the standard normal distribution. In particular, the distribution of $T$ should approach $N(0,1)$ as $n$ approaches infinity.
Density functions are compared for the $t$ distribution with $r = 1$ degree of freedom (solid) and $r = 4$ degrees of freedom (dashed) as well as for $N(0,1)$ (dotted). The probability of an interval is the area under the density curve, illustrated in the figure below.
Student t distribution

It can be proven that these properties are indeed true. The distribution of $T$ is called the *t distribution with n−1 degrees of freedom* and is denoted $t(n−1)$ or $t_{n−1}$, so we write

$$T = \frac{\sqrt{n}(\bar{y} - \mu)}{s} \sim t_{n−1}.$$  

The *t* distribution is often called *Student’s t distribution* because the distribution result was first published in 1908 under the pseudonym “Student”. The author was a chemist, William S. Gosset, employed at the Guinness brewery in Dublin. Gosset worked with what we would today call quality control of the brewing process. Due to time constraints, small samples of 4 or 6 were used, and Gosset realized that the normal distribution was not the proper one to use. The Guinness brewery only let Gosset publish his results under pseudonym.
Density for the $t_{10}$ distribution. The 95% quantile is 1.812, as illustrated by the gray region which has area 0.95. The 97.5% quantile is 2.228, illustrated by the dashed region with area 0.975.
The following table shows the 95% and 97.5% quantiles for the \( t \) distribution for a few selected degrees of freedom for illustration. The quantiles are denoted \( t_{0.95,r} \) and \( t_{0.975,r} \) as illustrated for \( r = 10 \). For data analyses where other degrees of freedom are in order, you should look up the relevant quantiles in a statistical table, called the \( t \) distribution table.

<table>
<thead>
<tr>
<th>Quantile</th>
<th>( t_{1} )</th>
<th>( t_{2} )</th>
<th>( t_{5} )</th>
<th>( t_{10} )</th>
<th>( t_{20} )</th>
<th>( t_{50} )</th>
<th>( t_{100} )</th>
<th>( N(0,1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td>6.314</td>
<td>2.920</td>
<td>2.015</td>
<td>1.812</td>
<td>1.725</td>
<td>1.676</td>
<td>1.660</td>
<td>1.645</td>
</tr>
<tr>
<td>97.5%</td>
<td>12.706</td>
<td>4.303</td>
<td>2.571</td>
<td>2.228</td>
<td>2.086</td>
<td>2.009</td>
<td>1.984</td>
<td>1.960</td>
</tr>
</tbody>
</table>
In one sample case let $y_1, \ldots, y_n$ be independent and $N(\mu, \sigma^2)$ distributed. The estimate of parameters $\mu$ and $\sigma^2$ are given by

$$\hat{\mu} = \bar{y} \sim N(\mu, \sigma^2 / n), \quad \hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

and hence the standard error of the parameter $\mu$ is given by

$$\text{SE}(\hat{\mu}) = s \sqrt{\frac{1}{n}} = \frac{s}{\sqrt{n}}.$$

By standardization we obtain

$$T = \frac{\sqrt{n}(\bar{y} - \mu)}{s} \sim t_{n-1}.$$
Confidence interval: one sample model

If we denote the 97.5% quantile in the $t$ distribution by $t_{0.975,n-1}$, then

$$P\left(-t_{0.975,n-1} < \frac{\sqrt{n}(\bar{y} - \mu)}{s} < t_{0.975,n-1}\right) = 0.95. \quad (5.18)$$

If we move around terms in order to isolate $\mu$, we get

$$0.95 = P\left(-t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}} < \bar{y} - \mu < t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

$$= P\left(\bar{y} - t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}}\right).$$

Therefore, the interval

$$\left(\bar{y} - t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{y} + t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

includes the true parameter value $\mu$ with a probability of 95%. The interval is called a 95% confidence interval for $\mu$. 
Confidence interval in general

Using the estimate of $\mu$ and its corresponding standard error, we can write the 95% confidence interval for $\mu$ by

$$\hat{\mu} \pm t_{0.975,n-1} \cdot SE(\hat{\mu}).$$

That is, the confidence interval has the form

$$\hat{\theta}_j \pm t_{0.975,n-p} \cdot SE(\hat{\theta}_j).$$

(5.22)

In general, confidence intervals for any parameter in a statistical model can be constructed in this way.
Confidence interval in general

If we repeated the experiment or data collection procedure many times and computed the interval

\[ \bar{y} \pm t_{0.975, n-1} \cdot \frac{s}{\sqrt{n}} \]

then 95% of those intervals would include the true value of \( \mu \). We have drawn 50 samples of size 10 with \( \mu = 0 \), and for each of these 50 samples we have computed and plotted the confidence interval. The true value \( \mu = 0 \) is included in the confidence interval 95% of the time. The 75% confidence interval for \( \mu \) is given by

\[ \hat{\mu} \pm t_{0.875, n-1} \cdot \text{SE}(\hat{\mu}). \]

The 75% confidence intervals are more narrow such that the true value is excluded more often, with a probability of 25% rather than 5%. This reflects that our confidence in the 75% confidence interval is smaller compared to the 95% confidence interval.
Confidence interval in general

Confidence intervals for 50 simulated data generated for the true value $\mu = 0$. Each shows 95% confidence intervals of size $n = 10$, 75% confidence intervals of size $n = 10$, and 95% confidence intervals of size $n = 40$, respectively.
Confidence interval in general

The true value $\mu_0$ is either inside the interval or it is not, but we will never know. We can, however, interpret the values in the confidence interval as the values for which it is reasonable to believe that they could have generated the data. If we use 95% confidence intervals,

- the probability of observing data for which the corresponding confidence interval includes $\mu_0$ is 95%
- the probability of observing data for which the corresponding confidence interval does not include $\mu_0$ is 5%

As a standard phrase we may say that the 95% confidence interval includes those values that are in agreement with the data on the 95% confidence level.
Summary: Confidence interval

**Construction.** In general we denote the confidence level $1-\alpha$, such that 95% and 90% confidence intervals corresponds to $\alpha = 0.05$ and $\alpha = 0.10$, respectively. The relevant $t$ quantile is $1-\alpha/2$, assigning probability $\alpha/2$ to the right. Then $(1-\alpha)$-confidence interval for a parameter $\theta$ is of the form

$$\hat{\theta}_j \pm t_{r,1-\alpha/2} \cdot \text{SE}(\hat{\theta}_j),$$

where $r$ is the degrees of freedom.

**Interpretation.** The $1-\alpha$ confidence interval includes the values of $\theta$ for which it is reasonable, at confidence degree $1-\alpha$, to believe that they could have generated the data. If we repeated the experiment many times then a fraction $1-\alpha$ of the corresponding confidence intervals would include the true value $\theta$. 
In the linear regression model, the estimate of parameters are given by

\[ \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}. \]

In order to simplify formulas, define \( SS_x \) as the denominator in the definition.

\[ SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2. \]

The standard errors are given by

\[ SE(\hat{\beta}) = \frac{s}{\sqrt{SS_x}}, \quad SE(\hat{\alpha}) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}, \]

where

\[ s = \sqrt{\frac{SS_e}{df_e}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \bar{\alpha} - \hat{\beta} \cdot x_i)^2} \]
In regression analysis we often seek to estimate the prediction for a particular value of $x$. Let $x_0$ be such an $x$-value of interest. The predicted value of the response is denoted $\mu_0$; that is, $\mu_0 = \alpha + \beta \cdot x_0$. It is estimated by

$$\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} \cdot x_0.$$ 

The predicted value has the standard error

$$SE(\hat{\mu}_0) = s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}. \quad (5.14)$$
Researchers examined the digestibility of fat with different levels of stearic acid. The average digestibility percent was measured for nine different levels of stearic acid proportion. Data are shown in the table below, where $x$ represents stearic acid and $y$ is digestibility measured in percent.

<table>
<thead>
<tr>
<th>$x$</th>
<th>29.8</th>
<th>30.3</th>
<th>22.6</th>
<th>18.7</th>
<th>14.8</th>
<th>4.1</th>
<th>4.4</th>
<th>2.8</th>
<th>3.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>67.5</td>
<td>70.6</td>
<td>72.0</td>
<td>78.2</td>
<td>87.0</td>
<td>89.9</td>
<td>91.2</td>
<td>93.1</td>
<td>96.7</td>
</tr>
</tbody>
</table>
Example: Stearic acid and digestibility

Consider the linear regression model which describes the association between the level of stearic acid and digestibility.

\[
\hat{\alpha} = 96.5334, \quad \hat{\beta} = -0.9337
\]

The residual sum of square \( SS_e = 61.7645 \) is used to calculate

\[
s^2 = \frac{61.7645}{9 - 2} = 8.8234, \quad s = \sqrt{8.8234} = 2.970,
\]

\[
SE(\hat{\beta}) = \frac{2.970}{\sqrt{1028.549}} = 0.0926, \quad SE(\hat{\alpha}) = 2.970 \cdot \sqrt{\frac{1}{9} + \frac{14.5889^2}{1028.549}} = 1.6752.
\]
Example: Stearic acid and digestibility

There are $n = 9$ observations and $p = 2$ parameters. Thus, we need quantiles from the $t$ distribution with 7 degrees of freedom. Since $t_{0.95,7} = 1.895$ and $t_{0.975,7} = 2.365$, we compute the 90% and the 95% confidence interval

$90\% \text{ CI: } -0.9337 \pm 1.895 \cdot 0.0926 = -0.9337 \pm 0.1754 = (-1.11, -0.76)$

$95\% \text{ CI: } -0.9337 \pm 2.365 \cdot 0.0926 = -0.9337 \pm 0.2190 = (-1.15, -0.71)$.

for the slope parameter $\beta$. Hence, decrements between 0.76 and 1.11 percentage points of the digestibility per unit increment of stearic acid level are in agreement with the data on the 90% confidence level.
Example: Stearic acid and digestibility

If we consider a stearic acid level of $x_0 = 20\%$, then we will expect a digestibility percentage of

$$\hat{\mu}_0 = 96.5334 - 0.9337 \cdot 20 = 77.859,$$

which has standard error

$$SE(\hat{\mu}_0) = 2.970 \cdot \sqrt{\frac{1}{9} + \frac{(20 - 14.5889)^2}{1028.549}} = 1.1096.$$ 

The 95% confidence interval for $\mu_0$ is given by

$$77.859 \pm 2.36 \cdot 1.1096 = 77.859 \pm 2.624 = (75.235, 80.483),$$

In conclusion, the predicted values of digestibility percentage corresponding to a stearic acid level of 20% between 75.2 and 80.5 are in accordance with the data on the 95% confidence level.
Example: Stearic acid and digestibility

We can calculate the 95% confidence interval for the expected digestibility percentage for other values of the stearic acid level.
An experiment with two different salmon stocks, from River Conon in Scotland and from River Ätran in Sweden, was carried out as follows.

<table>
<thead>
<tr>
<th>Stock</th>
<th>No. of parasites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ätran</td>
<td>31  31  32  22  41  31  29  40  41  39  36  17  29</td>
</tr>
<tr>
<td>Conon</td>
<td>18  26  16  20  14  28  18  27  17  32  19  17  28</td>
</tr>
</tbody>
</table>

The statistical model for the salmon data is given by

\[ y_i = \alpha_{g(i)} + e_i, \quad i = 1, \ldots, 26, \]

where \( g(i) \) is either “Ätran” or “Conon” and \( e_1, \ldots, e_{26} \) are from \( N(0, \sigma^2) \). In other words, Ätran observations are \( N(\alpha_{\text{Ätran}}, \sigma^2) \) distributed, and Conon observations are \( N(\alpha_{\text{Conon}}, \sigma^2) \) distributed.
Example: Parasite counts for salmons

We can compute the group means and the residual standard deviation.

\[
\hat{\mu}_{\text{Atran}} = \bar{y}_1 = 32.23, \quad \hat{\mu}_{\text{Conon}} = \bar{y}_2 = 21.54
\]

\[
\hat{\sigma}^2 = s^2 = \frac{1}{24} (12 \cdot s_1^2 + 12 \cdot s_2^2) = \frac{1}{2} (7.28^2 + 5.81^2) = 43.40, \quad s = 6.59.
\]

The difference in parasite counts is estimated by

\[
\hat{\mu}_{\text{Atran}} - \hat{\mu}_{\text{Conon}} = 32.23 - 21.54 = 10.69
\]

with a standard error of

\[
\text{SE}(\hat{\mu}_{\text{Atran}} - \hat{\mu}_{\text{Conon}}) = s \sqrt{\frac{2}{13}} = 2.58.
\]

The 95% confidence interval for the difference is given by

\[
10.69 \pm 2.064 \cdot 2.58 = (5.36, 16.02)
\]

We see that the data is not in accordance with a difference of zero between the stock means. Thus, the data suggests that Atran salmons are more susceptible than Conon salmons to parasites during an infection.
Unpaired samples with different SD’s

Assume that the observations \( y_1, \ldots, y_n \) are independent and come from two different groups, group 1 and group 2. Observations from group 1 are assumed to be \( N(\mu_1, \sigma^2_1) \) and observations from group 2 are assumed to be \( N(\mu_2, \sigma^2_2) \). The standard error is estimated by

\[
SE(\hat{\mu}_2 - \hat{\mu}_1) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.
\]

The 95% confidence interval for the difference is given by

\[
\hat{\mu}_2 - \hat{\mu}_1 \pm t_{1-\alpha/2, r} \cdot SE(\hat{\mu}_2 - \hat{\mu}_1)
\]

with degrees of freedom

\[
r = \frac{(SE_1^2 + SE_2^2)^2}{\frac{SE_1^4}{n_1-1} + \frac{SE_2^4}{n_2-1}}
\]
Unpaired samples with different SD’s

Under the assumption of different SD’s we get the standard error

$$SE(\hat{\mu}_2 - \hat{\mu}_1) = \sqrt{\frac{7.28^2}{13} + \frac{5.81^2}{13}} = 2.58.$$  

For the degrees of freedom we get $r = 22.9$. So the 95% confidence interval for the difference becomes

$$10.69 \pm 2.069 \cdot 2.58 = (5.35, 16.04),$$

almost the same as before, where the SD’s are assumed to be the same for the two stocks. In general the results based on the assumption of equal SD’s are quite robust as long as the samples are roughly of the same size and not too small.