Dataanalyse - Sampling & estimation - Kursusgang 2

Ege Rubak - rubak@math.aau.dk
http://www.math.aau.dk/∼rubak/teaching/2010/nano4

12. februar 2010

Terminology

- Population: All the individuals we are interested in.
  - E.g.: All companys in Denmark
- Sample: A subset of the population.
  - E.g.: 50 randomly chosen companys.
- Parameter: A descriptive measure of the population.
  - E.g.: Mean or variance.
  - E.g.: The average number of employees in Danish companies.
- Sample statistic: A descriptive measure of the sample.
  - E.g.: The average number of employees in the sample.
- Goal: Make conclusion about population by using sample.
  - Method: Make conclusion about parameter from sample statistic.

Terminology

- Population: All the individuals we are interested in.
  - E.g.: All companys in Denmark
- Sample: A subset of the population.
  - E.g.: 50 randomly chosen companys.
- Parameter: A descriptive measure of the population.
  - E.g.: Mean or variance.
  - E.g.: The average number of employees in Danish companies.
- Sample statistic: A descriptive measure of the sample.
  - E.g.: The average number of employees in the sample.
- Goal: Make conclusion about population by using sample.
  - Method: Make conclusion about parameter from sample statistic.

Terminology

- Population: All the individuals we are interested in.
  - E.g.: All companys in Denmark
- Sample: A subset of the population.
  - E.g.: 50 randomly chosen companys.
- Parameter: A descriptive measure of the population.
  - E.g.: Mean or variance.
  - E.g.: The average number of employees in Danish companies.
- Sample statistic: A descriptive measure of the sample.
  - E.g.: The average number of employees in the sample.
- Goal: Make conclusion about population by using sample.
  - Method: Make conclusion about parameter from sample statistic.

Terminology

- Population: All the individuals we are interested in.
  - E.g.: All companys in Denmark
- Sample: A subset of the population.
  - E.g.: 50 randomly chosen companys.
- Parameter: A descriptive measure of the population.
  - E.g.: Mean or variance.
  - E.g.: The average number of employees in Danish companies.
- Sample statistic: A descriptive measure of the sample.
  - E.g.: The average number of employees in the sample.
- Goal: Make conclusion about population by using sample.
  - Method: Make conclusion about parameter from sample statistic.

Terminology

- Population: All the individuals we are interested in.
  - E.g.: All companys in Denmark
- Sample: A subset of the population.
  - E.g.: 50 randomly chosen companys.
- Parameter: A descriptive measure of the population.
  - E.g.: Mean or variance.
  - E.g.: The average number of employees in Danish companies.
- Sample statistic: A descriptive measure of the sample.
  - E.g.: The average number of employees in the sample.
- Goal: Make conclusion about population by using sample.
  - Method: Make conclusion about parameter from sample statistic.

Sampling

- We want to do calculations with data.
- Observations are realizations of stocastics variables.
- We need to know the distribution of data.

$$
\begin{array}{cccc}
X_1 & X_2 & \cdots & X_n \\
\downarrow & \downarrow & & \downarrow \\
x_1 & x_2 & \cdots & x_n
\end{array}
$$

Terminology:

- $X_1, \ldots, X_n$ is a sample.
- $x_1, \ldots, x_n$ is an observed sample.
  We also call this observations.

Sampling

- We want to do calculations with data.
- Observations are realizations of stocastics variables.
- We need to know the distribution of data.

$$\begin{array}{cccc} X_1 & X_2 & \cdots & X_n \\ \downarrow & \downarrow & & \downarrow \\ x_1 & x_2 & \cdots & x_n \end{array}$$

Terminology:

- $X_1, \ldots, X_n$ is a sample.
- $x_1, \ldots, x_n$ is an observed sample.
  We also call this observations.

Sampling

- We want to do calculations with data.
- Observations are realizations of stocastics variables.
- We need to know the distribution of data.

$$\begin{array}{cccc} X_1 & X_2 & \cdots & X_n \\ \downarrow & \downarrow & & \downarrow \\ x_1 & x_2 & \cdots & x_n \end{array}$$

Terminology:

- $X_1, \ldots, X_n$ is a sample.
- $x_1, \ldots, x_n$ is an observed sample.
  We also call this observations.

Sampling

- We want to do calculations with data.
- Observations are realizations of stocastics variables.
- We need to know the distribution of data.

$$
\begin{array}{cccc}
X_1 & X_2 & \cdots & X_n \\
\downarrow & \downarrow & & \downarrow \\
x_1 & x_2 & \cdots & x_n
\end{array}
$$

Terminology when $X_i \sim N(\mu, \sigma)$:

- $X_1, \ldots, X_n$ is a sample from a normal distribution $N(\mu, \sigma)$.
- $x_1, \ldots, x_n$ is an observed sample from a normal distribution $N(\mu, \sigma)$.
  We also call this observations.

Estimator

- We have a identically distributed sample $X_1, \dots, X_n$.
- An estimator of a population parameter is a sample statistic used to estimate the parameter.
- Estimators for mean and variance is $\overline{X}$ and S^2, respectively:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\Big)$$

- $\overline{X}$ and S^2 are also stocastic variables.
- If $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$:

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{og} \quad E(S^2) = \sigma^2.$$

Estimator

- We have a identically distributed sample $X_1, \ldots, X_n$.
- An estimator of a population parameter is a sample statistic used to estimate the parameter.
- Estimators for mean and variance is $\overline{X}$ and S^2, respectively:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

- $\overline{X}$ and S^2 are also stocastic variables.
- If $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$:

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{og} \quad E(S^2) = \sigma^2.$$

Estimator

- We have a identically distributed sample $X_1, \dots, X_n$.
- An estimator of a population parameter is a sample statistic used to estimate the parameter.
- Estimators for mean and variance is $\overline{X}$ and S^2, respectively:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

- $\overline{X}$ and S^2 are also stocastic variables.
- If $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$:

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{og} \quad E(S^2) = \sigma^2.$$

Estimator

- We have a identically distributed sample $X_1, \dots, X_n$.
- An estimator of a population parameter is a sample statistic used to estimate the parameter.
- Estimators for mean and variance is $\overline{X}$ and S^2, respectively:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\Big)$$

- $\overline{X}$ and S^2 are also stocastic variables.
- If $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$:

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{og} \quad E(S^2) = \sigma^2.$$

Estimator

- We have a identically distributed sample $X_1, \dots, X_n$.
- An estimator of a population parameter is a sample statistic used to estimate the parameter.
- Estimators for mean and variance is $\overline{X}$ and S^2, respectively:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\Big(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\Big)$$

- $\overline{X}$ and S^2 are also stocastic variables.
- If $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$:

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{og} \quad E(S^2) = \frac{n-1}{n}\sigma^2.$$

Estimates

- We have observed a sample $x_1, \ldots, x_n$:

$$X_i \sim N(\mu, \sigma^2)$$

- An estimate of a parameter is a certain value of a sample statistic.
- Estimator $\rightarrow$ estimate by $X_i \rightarrow x_i$:

$$\begin{array}{ccccccc} X_1 & X_2 & \cdots & X_n & & \overline{X} & S^2 \\ \downarrow & \downarrow & & \downarrow & & \downarrow & \downarrow \\ x_1 & x_2 & \cdots & x_n & & \overline{x} & s^2 \end{array}$$

- We estimate μ and σ^2 with $\overline{x}$ and s^2, respectively:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big)$$

Estimates

- We have observed a sample $x_1, \ldots, x_n$:

$$X_i \sim N(\mu, \sigma^2)$$

- An estimate of a parameter is a certain value of a sample statistic.
- Estimator $\rightarrow$ estimate by $X_i \rightarrow x_i$:

$$\begin{array}{cccccc} X_1 & X_2 & \cdots & X_n & \overline{X} & S^2 \\ \downarrow & \downarrow & & \downarrow & \downarrow & \downarrow \\ x_1 & x_2 & \cdots & x_n & \overline{x} & s^2 \end{array}$$

- We estimate μ and σ^2 with $\overline{x}$ and s^2, respectively:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

Estimates

- We have observed a sample $x_1, \dots, x_n$:

$$X_i \sim N(\mu, \sigma^2)$$

- An estimate of a parameter is a certain value of a sample statistic.
- Estimator $\rightarrow$ estimate by $X_i \rightarrow x_i$:

$$\begin{array}{ccccccc} X_1 & X_2 & \cdots & X_n & & \overline{X} & S^2 \\ \downarrow & \downarrow & & \downarrow & & \downarrow & \downarrow \\ x_1 & x_2 & \cdots & x_n & & \overline{x} & s^2 \end{array}$$

- We estimate μ and σ^2 with $\overline{x}$ and s^2, respectively:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

Normally distributed data

... is our favorite situation!

- Easy calculations.
- Beatiful theory :-)
- So we also use it if data is approximately normal distributed.

Remember from last time:

- Mean and variance characterises the normal distribution.
- If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent:

$$aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

Normally distributed data

... is our favorite situation!

- Easy calculations.
- Beatiful theory :-)
- So we also use it if data is approximately normal distributed.

Remember from last time:

- Mean and variance characterises the normal distribution.
- If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent:

$$aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

Normally distributed data

... is our favorite situation!

- Easy calculations.
- Beatiful theory :-)
- So we also use it if data is approximately normal distributed.

Remember from last time:

- Mean and variance characterises the normal distribution.
- If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent:

$$aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

Estimators for normal data

- We have a normal distributed sample with independent observations:

$$X_i \sim N(\mu, \sigma^2)$$

- Estimators for μ and σ^2:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim N(\mu, \frac{\sigma^2}{n})$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$$

- We have:

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{og} \quad E(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

$\overline{X}$ and S^2 are independent.

Estimators for normal data

- We have a normal distributed sample with independent observations:

$$X_i \sim N(\mu, \sigma^2)$$

- Estimators for μ and σ^2:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim N(\mu, \frac{\sigma^2}{n})$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$$

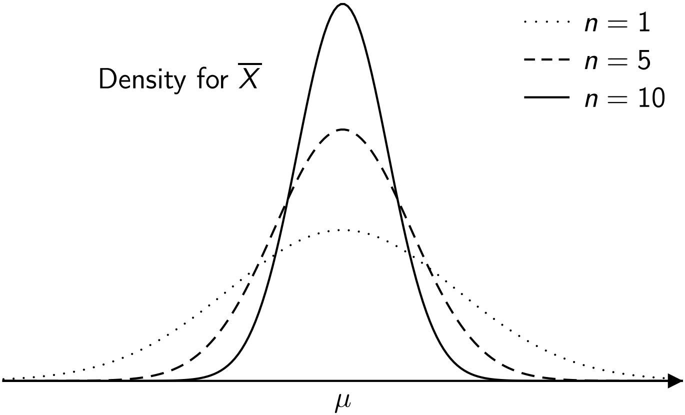
- We have:

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{og} \quad E(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

$\overline{X}$ and S^2 are independent.

Effect of more observations

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$



Estimates

- We have a normal distributed sample $x_1, \ldots, x_n$:

$$X_i \sim N(\mu, \sigma^2)$$

- We estimate μ and σ^2 with $\overline{x}$ and s^2:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

- We replace X_i with the observations x_i.

$$\begin{array}{cccccc} X_1 & X_2 & \cdots & X_n & \overline{X} & S^2 \\ \downarrow & \downarrow & & \downarrow & \downarrow & \downarrow \\ x_1 & x_2 & \cdots & x_n & \overline{x} & s^2 \end{array}$$

Confidence interval

- A point estimate is not interesting alone.
- We want to say something about the uncertainty of the estimate.
- We need the distribution of the estimate.
- We are going to look at 2 confidence intervals:
    1. μ in normal dsitribution with known σ.
    2. μ in normal distribution with unknown σ.

Confidence interval

- A point estimate is not interesting alone.
- We want to say something about the uncertainty of the estimate.
- We need the distribution of the estimate.
- We are going to look at 2 confidence intervals:
    1. μ in normal dsitribution with known σ.
    2. μ in normal distribution with unknown σ.

Confidence interval

- A point estimate is not interesting alone.
- We want to say something about the uncertainty of the estimate.
- We need the distribution of the estimate.
- We are going to look at 2 confidence intervals:
    1. μ in normal dsitribution with known σ.
    2. μ in normal distribution with unknown σ.

Confidence interval for μ with known σ

- Sample $(X_1, \ldots, X_n)$, $X_i \sim N(\mu, \sigma^2)$.
- Remember from last time: If $Y \sim N(\mu, \sigma^2)$ then $\frac{Y-\mu}{\sigma} \sim N(0,1)$:

$$P(-1.96 \leq \frac{Y-\mu}{\sigma} \leq 1.96) = 0.95$$

- Remember: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$:

$$P\big(-1.96 \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\big) = 0.95$$

$$\Updownarrow$$

$$P\big(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\big) = 0.95$$

- What can we learn from

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95?$$

- The probability that $\overline{X}$ takes a value $\overline{x}$, such that the interval $[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ contains μ, is 0.95.
  - This interval is called a 95% confidence interval for μ.
- The interval is stocastic.
- Generally: $100(1-\alpha)\%$ confidence interval for μ:

$$[\overline{x} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}; \overline{x} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}].$$

  $z_{\alpha/2}$ is the $\alpha/2$ fractile for standard normal distribution.
- Note the notation in the book is with $\eta = (1-\alpha)$.

- What can we learn from

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95?$$

- The probability that $\overline{X}$ takes a value $\overline{x}$, such that the interval $[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ contains μ, is 0.95.
  - This interval is called a 95% confidence interval for μ.
- The interval is stocastic.
- Generally: $100(1-\alpha)\%$ confidence interval for μ:

$$[\overline{x} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}; \overline{x} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}].$$

  $z_{\alpha/2}$ is the $\alpha/2$ fractile for standard normal distribution.
- Note the notation in the book is with $\eta = (1-\alpha)$.

- What can we learn from

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95?$$

- The probability that $\overline{X}$ takes a value $\overline{x}$, such that the interval $[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ contains μ, is 0.95.
  - This interval is called a 95% confidence interval for μ.
- The interval is stocastic.
- Generally: $100(1-\alpha)\%$ confidence interval for μ:

$$[\overline{x} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}; \overline{x} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}].$$

  $z_{\alpha/2}$ is the $\alpha/2$ fractile for standard normal distribution.
- Note the notation in the book is with $\eta = (1-\alpha)$.

- What can we learn from

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95?$$

- The probability that $\overline{X}$ takes a value $\overline{x}$, such that the interval $[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ contains μ, is 0.95.
  - This interval is called a 95% confidence interval for μ.
- The interval is stocastic.
- Generally: $100(1-\alpha)\%$ confidence interval for μ:

$$[\overline{x} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}; \overline{x} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}].$$

  $z_{\alpha/2}$ is the $\alpha/2$ fractile for standard normal distribution.
- Note the notation in the book is with $\eta = (1-\alpha)$.

- What can we learn from

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95?$$

- The probability that $\overline{X}$ takes a value $\overline{x}$, such that the interval $[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}; \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ contains μ, is 0.95.
  - This interval is called a 95% confidence interval for μ.
- The interval is stocastic.
- Generally: $100(1-\alpha)\%$ confidence interval for μ:

$$[\overline{x} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}; \overline{x} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}].$$

  $z_{\alpha/2}$ is the $\alpha/2$ fractile for standard normal distribution.
- Note the notation in the book is with $\eta = (1-\alpha)$.

Interpretation

- An experiment with sample size n is repeated k times:

$$\begin{array}{rcl} 1\colon x_{1,1}, x_{1,2}, \ldots, x_{1,n} & \longrightarrow & \overline{x}_1 \\ 2\colon x_{2,1}, x_{2,2}, \ldots, x_{2,n} & \longrightarrow & \overline{x}_2 \\ & \vdots & \\ k\colon x_{k,1}, x_{k,2}, \ldots, x_{k,n} & \longrightarrow & \overline{x}_k \end{array}$$

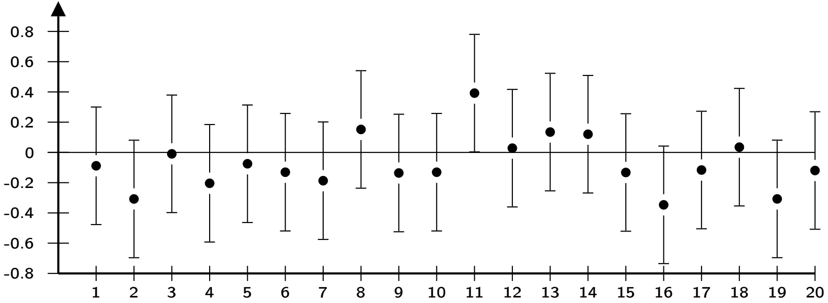
- Evaluate 95% confidence interval for each of $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$.
- We expect that 95% of confidence intervals contains μ.

Illustration of confidence intervals

20 samples with 100 observations:

$$(x_{1,1},\dots,x_{1,100}),\dots,(x_{20,1},\dots,x_{20,100}),\quad X_{i,j}\sim N(0,2)$$

$$\overline{x}_{i,\cdot}=\frac{1}{100}\sum_{j=1}^{100}X_{i,j}\sim N(0,\frac{2}{10})$$



Facts about confidence intervals

- The smaller the better.
- More observations give smaller confidence intervals.
- Larger % gives larger confidence interval (95% CI is contained in 99% CI).

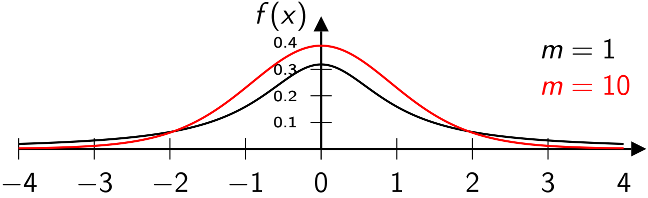
t distribution

- $U \sim N(0,1)$
- $W \sim \chi^2(k)$
- U and W are independent.
- Then

$$T = \frac{U}{\sqrt{W/k}}$$

is t distributed with k degrees of freedom (notation: $T \sim t(k)$).

Density for $t(k)$:

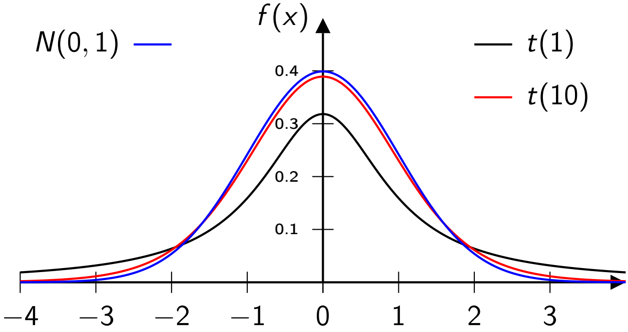


t distribution

- For $T \sim t(k)$ we have

$$E(T) = 0 \qquad \text{og} \qquad \text{Var}(T) = \frac{k}{k-2}, \quad \text{for } k > 2.$$

- The larger k, the more $t(k)$ looks like $N(0,1)$.



Confidence interval for μ with unknown σ

Sample with $(X_1, \ldots, X_n)$ independent, $X_i \sim N(\mu, \sigma^2)$.

- We have: $\overline{X} \sim N(\mu, \sigma^2/n)$ and $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$.
- Remember from before:

$$\frac{\overline{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1)$$

- Like when σ is known:

$$P\left(-|t_{\alpha/2}| \leq \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \leq |t_{\alpha/2}|\right) = 1 - \alpha$$

$$\Updownarrow$$

$$P\left(\overline{X} - t_{|\alpha/2|}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + |t_{\alpha/2}|\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$t_{\alpha/2}$ is $\alpha/2$-fractile in $t(n-1)$ distribution.

Confidence interval for μ with unknown σ

Sample with $(X_1, \ldots, X_n)$ independent, $X_i \sim N(\mu, \sigma^2)$.

- We have: $\overline{X} \sim N(\mu, \sigma^2/n)$ and $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$.
- Remember from before:

$$\frac{\overline{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1)$$

- Like when σ is known:

$$P\left(-|t_{\alpha/2}| \leq \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \leq |t_{\alpha/2}|\right) = 1 - \alpha$$

$$\Updownarrow$$

$$P\left(\overline{X} - t_{|\alpha/2|}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + |t_{\alpha/2}|\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$t_{\alpha/2}$ is $\alpha/2$-fractile in $t(n-1)$ distribution.

- $100(1-\alpha)\%$-confidence interval for μ when σ is unknown:

$$[\overline{x} - |t_{\alpha/2}|\frac{s}{\sqrt{n}}; \overline{x} + |t_{\alpha/2}|\frac{s}{\sqrt{n}}]$$

- Note:
  - $|t_\alpha| > |z_\alpha|$ regardless of the degree of freedom.
    - The confidence interval is greater than when we know σ
    - Natural to introduce more uncertainty when two parameters are unknown.
  - When the number of degrees of freedom grows, $t_\alpha \rightarrow z_\alpha$.
    - With many observations, it doesn't matter if σ is known or unknown.

Hypothesis testing

- A hypothesis is a statement that is either true or false
  - The average income in Aalborg is at least 100.000 kr.
  - The average height of males is the same in Sweden and Denmark.
  - The proportion of female students is the same on computer science and sociologi.

Formulation of hypothesis

- We start with quantitative hypothesis:
  - We are interested in a parameter θ.
  - θ_0 is a number.
- 3 kinds of hypothesis:

$$\begin{array}{lll} H_0 : \theta = \theta_0 & H_0 : \theta \geq \theta_0 & H_0 : \theta \leq \theta_0 \\ H_1 : \theta \neq \theta_0 & H_1 : \theta < \theta_0 & H_1 : \theta > \theta_0 \end{array}$$

- H_0 is called the null hypothesis.
  H_1 (sometimes noted H_A) is called the alternative hypothesis.
- The sign by H_1 determines if the test is 1- og 2-sided:
  - "$\neq$": 2-sided test – we have 2 directions if H_0 is rejected.
  - "$\geq$", "$\leq$": 1-sidet test – we have 1 direction if H_0 is rejected.

Formulation of hypothesis

- We start with quantitative hypothesis:
  - We are interested in a parameter θ.
  - θ_0 is a number.
- 3 kinds of hypothesis:

$$\begin{array}{lll} H_0 : \theta = \theta_0 & H_0 : \theta \geq \theta_0 & H_0 : \theta \leq \theta_0 \\ H_1 : \theta \neq \theta_0 & H_1 : \theta < \theta_0 & H_1 : \theta > \theta_0 \end{array}$$

- H_0 is called the null hypothesis.
  H_1 (sometimes noted H_A) is called the alternative hypothesis.
- The sign by H_1 determines if the test is 1- og 2-sided:
  - "$\neq$": 2-sided test – we have 2 directions if H_0 is rejected.
  - "$\geq$", "$\leq$": 1-sidet test – we have 1 direction if H_0 is rejected.

Formulation of hypothesis

- We start with quantitative hypothesis:
  - We are interested in a parameter θ.
  - θ_0 is a number.
- 3 kinds of hypothesis:

$$
\begin{array}{lll}
H_0 : \theta = \theta_0 & H_0 : \theta \geq \theta_0 & H_0 : \theta \leq \theta_0 \\
H_1 : \theta \neq \theta_0 & H_1 : \theta < \theta_0 & H_1 : \theta > \theta_0
\end{array}
$$

- H_0 is called the null hypothesis.
  H_1 (sometimes noted H_A) is called the alternative hypothesis.
- The sign by H_1 determines if the test is 1- og 2-sided:
  - "$\neq$": 2-sided test – we have 2 directions if H_0 is rejected.
  - "$\geq$", "$\leq$": 1-sidet test – we have 1 direction if H_0 is rejected.

Examples of hypothesis

Examples from before:

- The average income in Aalborg is at least 100.000 kr.

$$H_0 : \mu_{\text{income}} \geq 100.000$$
$$H_1 : \mu_{\text{income}} < 100.000$$

- The average height of males is the same in Sweden and Denmark.

$$H_0 : \mu_S = \mu_D$$
$$H_1 : \mu_S \neq \mu_D$$

- The proportion of female students is the same on computer science and sociologi.

$$H_0 : p_{cs} = p_s$$
$$H_1 : p_{cs} \neq p_s$$

Types of errors

- We can make two types of errors:
  - Type I error: Reject a true hypothesis.
  - Type II error: Accept a false hypothesis.

Choice	H_0 is true	H_0 is false
Reject H_0	Type I error	No error
Accept H_0	No error	Type II error

- Type I is the worst error: "We would rather let a criminal go free than put an innocent in prison".
- Ideally we want a test where it is difficult to make errors.

Types of errors

- We can make two types of errors:
  - Type I error: Reject a true hypothesis.
  - Type II error: Accept a false hypothesis.

Choice	H_0 is true	H_0 is false
Reject H_0	Type I error	No error
Accept H_0	No error	Type II error

- Type I is the worst error: “We would rather let a criminal go free than put an innocent in prison”.
- Ideally we want a test where it is difficult to make errors.

Types of errors

- We can make two types of errors:
  - Type I error: Reject a true hypothesis.
  - Type II error: Accept a false hypothesis.

Choice	H_0 is true	H_0 is false
Reject H_0	Type I error	No error
Accept H_0	No error	Type II error

- Type I is the worst error: "We would rather let a criminal go free than put an innocent in prison".
- Ideally we want a test where it is difficult to make errors.

Errors

- Tests without errors do not exist!
- Furthermore:
  - If a test rarely makes Type I errors, it (more) often makes Type II errors.
  - If a test rarely makes Type II errors, it (more) often makes Type I errors.
- The chance of making errors decrease when the sample size increase.

Errors

- Tests without errors do not exist!
- Furthermore:
  - If a test rarely makes Type I errors, it (more) often makes Type II errors.
  - If a test rarely makes Type II errors, it (more) often makes Type I errors.
- The chance of making errors decrease when the sample size increase.

Level of significance

- Level of significance:

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$
$$= P(\text{reject } H_0 \text{ when } H_0 \text{ is true}).$$

- α is chosen before we test.
  - Commonly: $\alpha = 5\%$.
  - Generally: Adapt α to the situation.
- We don't control

$$\beta = P(\text{Type II error}) = P(\text{accept } H_0 \text{ when } H_0 \text{ is false}).$$

Level of significance

- Level of significance:

$$\begin{aligned}\alpha &= P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true}) \\ &= P(\text{reject } H_0 \text{ when } H_0 \text{ is true}).\end{aligned}$$

- α is chosen before we test.
  - Commonly: $\alpha = 5\%$.
  - Generally: Adapt α to the situation.
- We don't control

$$\beta = P(\text{Type II error}) = P(\text{accept } H_0 \text{ when } H_0 \text{ is false}).$$

Level of significance

- Level of significance:

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$
$$= P(\text{reject } H_0 \text{ when } H_0 \text{ is true}).$$

- α is chosen before we test.
  - Commonly: $\alpha = 5\%$.
  - Generally: Adapt α to the situation.
- We don't control

$$\beta = P(\text{Type II error}) = P(\text{accept } H_0 \text{ when } H_0 \text{ is false}).$$

Consequences of controlling

$$\alpha = P(\text{Type I error})$$

and not

$$\beta = P(\text{Type II error})$$

- We have faith in our decision if we reject H_0
- If H_0 is <u>not</u> rejected, we cannot conclude that H_0 is true.
- Terminology if H_0 <u>can't</u> be rejected:

  Data does not allow us to reject the hypothesis H_0.

  We <u>don't</u> say:

  Data confirms the hypothesis H_0.

Consequences of controlling

$$\alpha = P(\text{Type I error})$$

and not

$$\beta = P(\text{Type II error})$$

- We have faith in our decision if we reject H_0
- If H_0 is <u>not</u> rejected, we cannot conclude that H_0 is true.
- Terminology if H_0 <u>can't</u> be rejected:

  Data does not allow us to reject the hypothesis H_0.

  We <u>don't</u> say:

  Data confirms the hypothesis H_0.

Decision rule

A decision rule is a rule, that tell us when to reject H_0.

- Test statistic: Function that tells us if data supports H_0.
- Critical values: Where the test statistic rejects H_0.

Hypothesis: "Is the average height (μ_h) in Denmark 180 cm?"

$$H_0 : \mu_h = 180$$
$$H_1 : \mu_h \neq 180$$

Procedure:

- We have observations from 100 people, $(x_1, \ldots, x_{100})$:

x_1	x_2	$\cdots$	x_{100}
178 cm	183 cm	$\cdots$	175 cm

- Idea: See if the average $\overline{x}$ is close to 180.
  - But what is "close"?

Decision rule

A decision rule is a rule, that tell us when to reject H_0.

- Test statistic: Function that tells us if data supports H_0.
- Critical values: Where the test statistic rejects H_0.

Hypothesis: "Is the average height (μ_h) in Denmark 180 cm?"

$$H_0 : \mu_h = 180$$
$$H_1 : \mu_h \neq 180$$

Procedure:

- We have observations from 100 people, $(x_1, \ldots, x_{100})$:

x_1	x_2	$\cdots$	x_{100}
178 cm	183 cm	$\cdots$	175 cm

- Idea: See if the average $\overline{x}$ is close to 180.
  - But what is "close"?

Decision rule

A decision rule is a rule, that tell us when to reject H_0.

- Test statistic: Function that tells us if data supports H_0.
- Critical values: Where the test statistic rejects H_0.

Hypothesis: "Is the average height (μ_h) in Denmark 180 cm?"

$$H_0 : \mu_h = 180$$
$$H_1 : \mu_h \neq 180$$

Procedure:

- We have observations from 100 people, $(x_1, \ldots, x_{100})$:

x_1	x_2	$\cdots$	x_{100}
178 cm	183 cm	$\cdots$	175 cm

- Idea: See if the average $\overline{x}$ is close to 180.
  - But what is "close"?

- We have observations $(x_1, x_2, \ldots, x_{100})$, $X_i \sim N(\mu_h, \sigma^2)$. Assume that we know $\sigma^2 = 25$.
- Estimate:

$$\overline{x} = \frac{1}{100}(178 + 183 + \cdots + 175) = 178.$$

- Remember:

$$Z = \frac{\overline{X} - \mu_h}{\sigma/\sqrt{n}} = \frac{\overline{X} - 180}{1/2} \sim N(0,1)$$

  We assume H_0 is true.
- Therefore:

$$P(180 - |z_{0.025}|\frac{1}{2} \leq \overline{X} \leq 180 + |z_{0.025}|\frac{1}{2}) \approx P(179 \leq \overline{X} \leq 181) = 0.95.$$

- We have observations $(x_1, x_2, \dots, x_{100})$, $X_i \sim N(\mu_h, \sigma^2)$. Assume that we know $\sigma^2 = 25$.
- Estimate:

$$\overline{x} = \frac{1}{100}(178 + 183 + \cdots + 175) = 178.$$

- Remember:

$$Z = \frac{\overline{X} - \mu_h}{\sigma/\sqrt{n}} = \frac{\overline{X} - 180}{1/2} \sim N(0,1)$$

We assume H_0 is true.

- Therefore:

$$P(180 - |z_{0.025}|\frac{1}{2} \leq \overline{X} \leq 180 + |z_{0.025}|\frac{1}{2}) \approx P(179 \leq \overline{X} \leq 181) = 0.95.$$

- We have observations $(x_1, x_2, \ldots, x_{100})$, $X_i \sim N(\mu_h, \sigma^2)$. Assume that we know $\sigma^2 = 25$.
- Estimate:

$$\overline{x} = \frac{1}{100}(178 + 183 + \cdots + 175) = 178.$$

- Remember:

$$Z = \frac{\overline{X} - \mu_h}{\sigma/\sqrt{n}} = \frac{\overline{X} - 180}{1/2} \sim N(0, 1)$$

We assume H_0 is true.

- Therefore:

$$P(180 - |z_{0.025}|\frac{1}{2} \leq \overline{X} \leq 180 + |z_{0.025}|\frac{1}{2}) \approx P(179 \leq \overline{X} \leq 181) = 0.95.$$

2.5% quantile
in $N(0, 1)$

Conclusion

Putting the pieces together:

- Our hypothesis:

$$H_0 : \mu_h = 180$$
$$H_1 : \mu_h \neq 180$$

- If H_0 is true, 95% of all samples with 100 persons has an average between 179 and 181 cm.
- In our experiment:
  1. The average is 178 cm.
  2. This is an event that occurs in at most 5% of the samples
  3. Conclusion: Our observation is very unlikely!
     We reject H_0.
- The level of significance is

Conclusion

Putting the pieces together:

- Our hypothesis:

$$H_0 : \mu_h = 180$$
$$H_1 : \mu_h \neq 180$$

- If H_0 is true, 95% of all samples with 100 persons has an average between 179 and 181 cm.
- In our experiment:
  1. The average is 178 cm.
  2. This is an event that occurs in at most 5% of the samples
  3. Conclusion: Our observation is very unlikely!
     We reject H_0.
- The level of significance is

Conclusion

Putting the pieces together:

- Our hypothesis:

$$H_0 : \mu_h = 180$$
$$H_1 : \mu_h \neq 180$$

- If H_0 is true, 95% of all samples with 100 persons has an average between 179 and 181 cm.
- In our experiment:
  1. The average is 178 cm.
  2. This is an event that occurs in at most $\boxed{\text{5\% of the samples}}$
  3. Conclusion: Our observation is very unlikely! We reject H_0.
- The level of significance is 5%.

Test with unknown variance

- We have observations $(x_1, x_2, \dots, x_n)$, $X_i \sim N(\mu, \sigma^2)$. σ^2 is unknown
- Estimates:

$$\overline{x} = \frac{1}{100}(178 + 183 + \cdots + 175) = 178$$
$$s^2 = \frac{1}{99}\left((178 - 178)^2 + \cdots + (175 - 178)^2\right) = 25.$$

- Assume H_0 is true. Remember:

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{178 - \mu}{1/2} \sim t(99).$$

- Hence:

$$P(180 - |t_{0.025}|\frac{1}{2} \leq \overline{X} \leq 180 + |t_{0.025}|\frac{1}{2}) \approx P(179 \leq \overline{X} \leq 181) = 0.95.$$

Test with unknown variance

- We have observations $(x_1, x_2, \ldots, x_n)$, $X_i \sim N(\mu, \sigma^2)$. σ^2 is unknown
- Estimates:

$$\overline{x} = \frac{1}{100}(178 + 183 + \cdots + 175) = 178$$
$$s^2 = \frac{1}{99}\left((178 - 178)^2 + \cdots + (175 - 178)^2\right) = 25.$$

- Assume H_0 is true. Remember:

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{178 - \mu}{1/2} \sim t(99).$$

- Hence:

$$P(180 - |t_{0.025}|\frac{1}{2} \le \overline{X} \le 180 + |t_{0.025}|\frac{1}{2}) \approx P(179 \le \overline{X} \le 181) = 0.95.$$

Test with unknown variance

- We have observations $(x_1, x_2, \ldots, x_n)$, $X_i \sim N(\mu, \sigma^2)$. σ^2 is unknown
- Estimates:

$$\overline{x} = \frac{1}{100}(178 + 183 + \cdots + 175) = 178$$

$$s^2 = \frac{1}{99}\left((178 - 178)^2 + \cdots + (175 - 178)^2\right) = 25.$$

- Assume H_0 is true. Remember:

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{178 - \mu}{1/2} \sim t(99).$$

- Hence:

$$P(180 - |t_{0.025}|\frac{1}{2} \leq \overline{X} \leq 180 + |t_{0.025}|\frac{1}{2}) \approx P(179 \leq \overline{X} \leq 181) = 0.95.$$

2.5% fractile in t(99)

General decision rule for normal distributed data

Hypothesis:

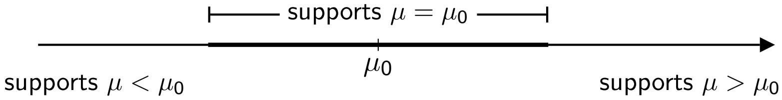
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Procedure for sample $(x_1, \ldots, x_n)$ with known variance σ^2.

- Choose level of significance α.
- Calculate sample mean $\overline{x}$.
- Check if

$$\mu_0 - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}} \leq \overline{x} \leq \mu_0 + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}$$

If yes: We cannot reject H_0.
If no: Reject H_0.



General decision rule for normal distributed data

Hypothesis:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

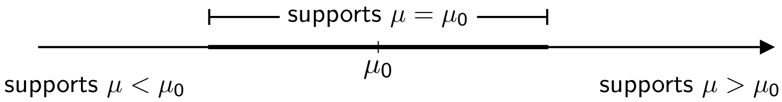
Procedure for sample $(x_1, \ldots, x_n)$ with unknown variance.

- Choose level of significance α.
- Calculate sample mean $\overline{x}$ and standard deviation s.
- Check if

$$\mu_0 - |t_{\alpha/2}|\frac{s}{\sqrt{n}} \leq \overline{x} \leq \mu_0 + |t_{\alpha/2}|\frac{s}{\sqrt{n}}$$

If yes: We cannot reject H_0.

If no: Reject H_0.



Decision rule with confidence interval

Hypothesis:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Procedure for sample $(x_1, \ldots, x_n)$ with known variance σ^2 and level of significance α:

- Calculate sample mean $\overline{x}$.
- Calculate confidence interval for μ:

$$[\overline{x} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}; \overline{x} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}]$$

- Is μ_0 in the confidence interval?

  If yes: We cannot reject H_0.

  If no: Reject H_0.

Decision rule with confidence interval

Hypothesis:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Procedure for sample $(x_1, \ldots, x_n)$ with unknown and level of significance α:

- Calculate sample mean $\overline{x}$ and standard deviation s.
- Calculate confidence interval for μ:

$$[\overline{x} - |t_{\alpha/2}|\frac{s}{\sqrt{n}}; \overline{x} + |t_{\alpha/2}|\frac{s}{\sqrt{n}}]$$

- Is μ_0 in the confidence interval?

  If yes: We cannot reject H_0.

  If no: Reject H_0.