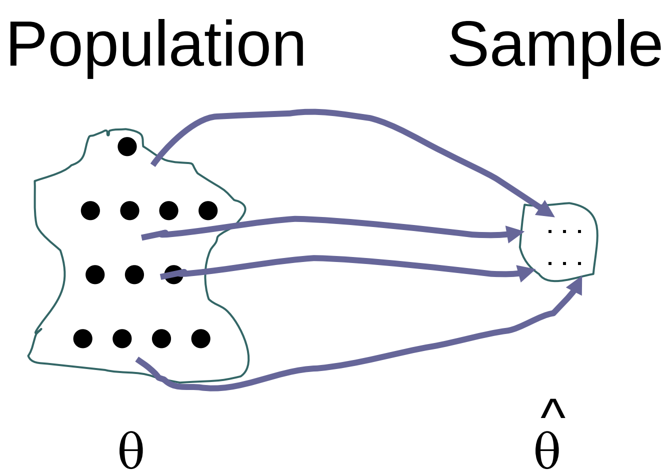
Statistics

Statistics:

1. Model $X_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, n$ iid.
2. Estimation $\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = s^2$
3. Hypothesis test $\mu = \mu_0, \quad \sigma^2 = \sigma_0^2$

Estimation
Estimate

Definition:
A (point) estimate $\hat{\theta}$ of a parameter, θ, in the model is a "guess" at what θ can be (based on the sample). The corresponding random variable $\hat{\Theta}$ is called an estimator.



parameter	estimate	estimator
μ	$\hat{\mu} = \bar{x}$	$\bar{X}$
σ	$\hat{\sigma}^2 = s^2$	S^2

Estimation
Unbiased estimate

Definition:
An estimator $\hat{\Theta}$ is said to be unbiased if

$$E(\hat{\Theta}) = \theta$$

We have

$E(\overline{X}) = \mu$ unbiased

$E(S^2) = \sigma^2$ unbiased

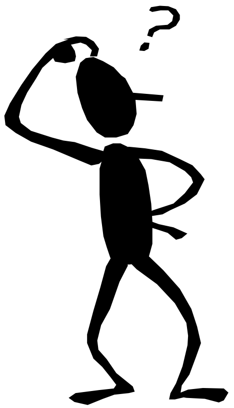
Confidence interval for the mean

Example:

In a sample of 20 chocolate bars the amount of calories has been measured:

- the sample mean is 224 calories

How certain are we that the population mean μ is close to 224?



The confidence interval (CI) for μ helps us here!

Confidence interval for μ
Known variance

Let $\overline{x}$ be the average of a sample consisting of n observations from a population with mean μ and variance σ^2 (known).

A (1-α)100% confidence interval for μ is given by

$$\overline{x} - z_{\alpha/2} \sigma/\sqrt{n} < \mu < \overline{x} + z_{\alpha/2} \sigma/\sqrt{n}$$

From the standard normal distribution N(0,1)

- We are (1-α)100% confident that the unknown parameter μ lies in the CI.
- We are (1-α)100% confident that the error we make by using $\overline{x}$ as an estimate of μ does not exceed $z_{\alpha/2}\sigma/\sqrt{n}$ (from which we can find n for a given error tolerance).

Confidence interval for μ
Known variance

Confidence interval for known variance is a results of the Central Limits Theorem. The underlying assumptions:

- sample size $n > 30$, or
- the corresponding random var. is (approx.) normally distributed

$(1 - \alpha)100\%$ confidence interval :

- typical values of α: $\alpha = 10\%$, $\alpha = 5\%$, $\alpha = 1\%$
- what happens to the length of the CI when n increases?

Confidence interval for μ
Known variance

Problem:

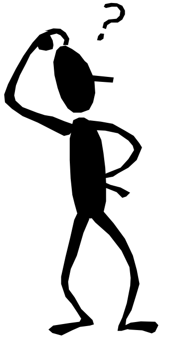
In a sample of 20 chocolate bars the amount of calories has been measured:

- the sample mean is 224 calories

In addition we assume:

- the corresponding random variable is approx. normally distributed
- the population standard deviation $\sigma = 10$

Calculate 90% and 95% confidence interval for μ



Which confidence interval is the longest?

Confidence interval for μ
Unknown variance

Let $\bar{x}$ be the mean and s the sample standard deviation of a sample of n observations from a normal distributed population with mean μ and unknown variance.

A (1-α)100 % confidence interval for μ is given by

$$\bar{x} - t_{\alpha/2, n-1} \, s/\sqrt{n} < \mu < \bar{x} + t_{\alpha/2, n-1} \, s/\sqrt{n}$$

From the t distribution with n-1 degrees of freedom t(n -1)

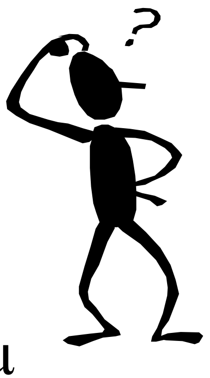
- Not necessarily normally distributed, just approx. normal distributed.
- For n > 30 the standard normal distribution can be used instead of the t distribution.
- We are (1-α)100% confident that the unknown μ lies in the CI.

Confidence interval for μ
Unknown variance

Problem:

In a sample of 20 chocolate bars the amount of calories has been measured:

- the sample mean is 224 calories
- the sample standard deviation is 10



Calculate 90% and 95% confidence intervals for μ

How does the lengths of these confidence intervals compare to those we obtained when the variance was known?

Confidence interval for μ
Using the computer

MATLAB: If `x` contains the data we can obtain a `(1-alpha)100%` confidence interval as follow:

```
mean(x) + [-1 1] * tinv(1-alpha/2,size(x,1)-1) *
                              std(x)/sqrt(size(x,1))
```

where

- `size(x,1)` is the size n of the sample
- `tinv(1-alpha/2,size(x,1)-1)` = $t_{\alpha/2}(n-1)$
- `std(x)` = s (sample standard deviation)

R:

```
mean(x) + c(-1,1) * qt(1-alpha/2,length(x)-1) *
                              sd(x)/sqrt(length(x))
```

Confidence interval for σ^2

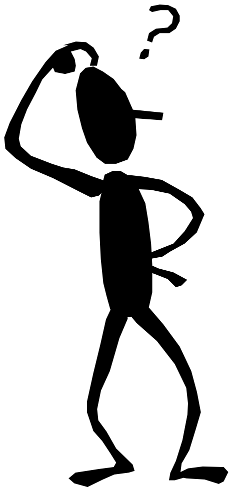
Example:

In a sample of 20 chocolate bars the amount of calories has been measured:

- sample standard deviation is 10

How certain are we that the population variance σ^2 is close to 10^2?

The confidence interval for σ^2 helps us answer this!



Confidence interval for σ^2

Let s be the standard deviation of a sample consisting of n observations from a normal distributed population with variance σ^2.

A (1-α) 100% confidence interval for σ^2 is given by

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

From χ^2 distribution with n-1 degrees of freedom

We are (1-α)100% confident that the unknown parameter σ^2 lies in the CI.

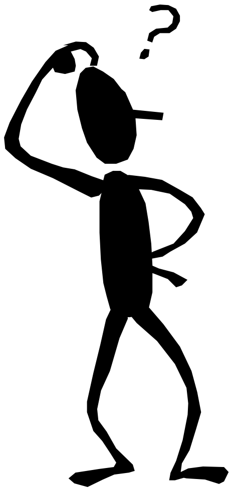
Confidence interval for σ^2

Problem:

In a sample of 20 chocolate bars the amount of calories has been measured:

- sample standard deviation is 10

Find the 90% and 95% confidence intervals for σ^2



Confidence interval for σ^2
Using the computer

MATLAB: If `x` contains the data we can obtain a `(1-alpha)100%` confidence interval for σ^2 as follow:

```
(size(x,1)-1)*std(x)^2./
     chi2inv([1-alpha/2 alpha/2],size(x,1)-1)
```

R:

```
(length(x)-1)*sd(x)^2)/
     qchisq(c(1-alpha/2,alpha/2), length(x)-1)
```

Difference in means
Estimation (known variances)

Consider two populations with means μ_1 and μ_2 and known variances $\sigma^2{}_1$ and $\sigma^2{}_2$, and two samples of sizes n_1 and n_2.

Estimate of $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2$$

Confidence interval:

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Test of two means
Known variances (two-sided)

Hypotheses:

$$H_0 : \mu_1 - \mu_2 = d_0$$
$$H_1 : \mu_1 - \mu_2 = d_0$$

Significance level:

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1 - \alpha$$

Test statistic:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

Critical values:

$$-z_{\alpha/2}, z_{\alpha/2}$$

Decision: Reject H_0 if z does not lie between the critical values

Test of two means

Unknown & equal variances (two-sided)

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{1/n_1 + 1/n_2}}$$

Pooled variance estimate:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Critical values:

$$-t_{\alpha/2,v}, t_{\alpha/2,v}$$

Degrees of freedom

$$v = n_1 + n_2 - 2$$

Test of two means

Unknown & unequal variances (two-sided)

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

Critical values:

$$-t_{\alpha/2,v}, t_{\alpha/2,v}$$

Degrees of freedom:

$$v = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}$$

Maximum Likelihood Estimation
The likelihood function

Assume that $X_1,\ldots,X_n$ are random variables with joint density/probability function

$$f(x_1, x_2, \ldots, x_n; \theta)$$

where θ is the parameter (vector) of the distribution.

Considering the above function as a function of θ given the data $x_1,\ldots,x_n$ we obtain the likelihood function

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n; \theta)$$

Maximum Likelihood Estimation
The likelihood function

Reminder: If $X_1,\ldots,X_n$ are independent random variables with identical marginal probability/ density function $f(x;\theta)$, then the joint probability / density function is

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta)$$

Definition: Given independent observations $x_1,\ldots,x_n$ from the probability / density function $f(x;\theta)$ the **maximum likelihood estimate** (MLE) θ is the value of θ which maximizes the likelihood function

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta)$$

Maximum Likelihood Estimation Example

Assume that $X_1,\ldots,X_n$ are a sample from a normal population with mean μ and variance σ^2. Then the marginal density for each random variable is

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2 \right]$$

Accordingly the joint density is

$$f(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \right]$$

The logarithm of the likelihood function is

$$\ln L(\mu,\sigma^2; x_1, x_2, \ldots, x_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2$$

Maximum Likelihood Estimation Example

We find the maximum likelihood estimates by maximizing the log-likelihood:

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2; \mathbf{x}) = \frac{1}{\sigma^2} \sum_i (x_i - \mu) = 0$$

which implies $\hat{\mu} = \frac{1}{n} \sum_i x_i = \bar{x}$. For σ^2 we have

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_i (x_i - \mu)^2 = 0$$

which implies $\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$

Notice $E[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2$, i.e. the MLE is biased!