Probability Theory and Statistics

Lecture 6

November 5, 2013

Robert Dahl Jacobsen
robert@math.aau.dk

Department of Mathematical Sciences
Aalborg University

AALBORG UNIVERSITY
DENMARK

Agenda

Estimation

Two means

Likelihoods

Matlab

Statistics in a nutshell

- Model:

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

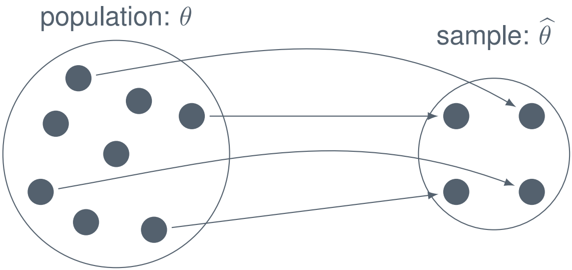
- Estimation:

$$\widehat{\mu} = \overline{x}, \quad \widehat{\sigma}^2 = s^2$$

- Hypothesis test:

$$\mu = \mu_0, \quad \sigma^2 = \sigma_0^2$$

Estimation



- Point estimate: Estimate of population parameter (θ) from sample ($\widehat{\theta}$).
- Estimator: Corresponding random variable ($\widehat{\Theta}$).

parameter	estimate	estimator
μ	$\overline{x}$	$\overline{X}$
σ^2	s^2	S^2

Unbiased estimate

- Unbiased estimator:

$$\mathrm{E}(\widehat{\Theta}) = \theta$$

- Example:

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

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\Big(\mu, \frac{\sigma^2}{n}\Big)$$

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

Then:

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

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

$$\mathrm{E}(S^2) = \sigma^2$$

Confidence interval for mean

Known variance

- Sample:

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, n$$

- Notation:

$$z_\alpha = \alpha \text{ fractile of } N(0,1)$$

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

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

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

- Shorthand:

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Confidence interval: Interpretation

- We are $(1-\alpha)100\%$ confident that μ is in the CI.
- 20 samples with 100 observations from $N(0,2)$:

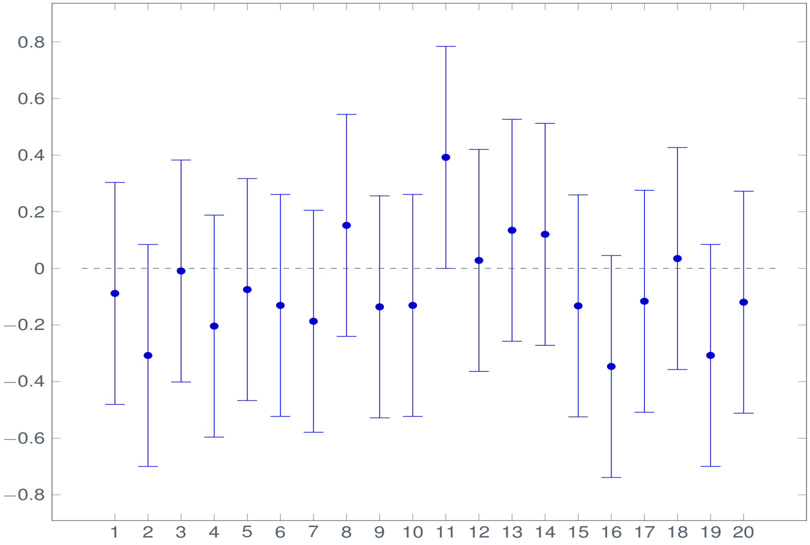
$$
\begin{array}{lllll}
1 & : x_{1,1}, x_{1,2}, \ldots, x_{1,100} & \rightarrow & \overline{x}_1 \\
2 & : x_{2,1}, x_{2,2}, \ldots, x_{2,100} & \rightarrow & \overline{x}_2 \\
 & \vdots & & \\
20 & : x_{20,1}, x_{20,2}, \ldots, x_{20,100} & \rightarrow & \overline{x}_{20}
\end{array}
$$

- 95% confidence interval:

$$\overline{x} \mp 1.96 \cdot \frac{2}{10}$$

- Expect one $\overline{x}_k$ outside confidence interval:
`http://xkcd.com/882`

Confidence intervals



Chocolate bars

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

- the corresponding random variable is approx. normally distributed.
- the population standard deviation is 10 calories.
- the sample mean is 224 calories.

Calculate 90% and 95% confidence intervals for the mean. Which one is larger?

Confidence interval for mean

Unknown variance

- Sample:

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

- Notation:

$$t_\alpha = \alpha \text{ fractile of } t(n-1)$$

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

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

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

or

$$\overline{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

- Note: $t_\alpha < z_\alpha$

Normal or t distribution?

General form of confidence interval for mean:

$$\bar{x} \pm \text{fractile} \, \frac{\text{std}}{\sqrt{n}}$$

Situation 1

- Observations from $N(\cdot, \cdot)$ (unknown mean and variance)
- Estimate:

$$\text{mean} = \bar{x}, \qquad \text{variance} = s^2$$

- Use
  - fractile from t distribution
  - std $= s^2$

Situation 2

- Observations from $N(\cdot, \sigma^2)$ (unknown mean)
- Estimate:

$$\text{mean} = \bar{x}$$

- Use
  - fractile from normal distribution
  - std $= \sigma^2$

More chocolate bars

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

- the corresponding random variable is approx. normally distributed.
- the sample standard deviation is 10 calories.
- the sample mean is 224 calories.

Calculate 90% and 95% confidence intervals for the mean.
How are the confidence intervals compared to the ones with known variance?

Confidence interval for variance

- Sample:

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

- Notation:

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

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\chi^2_{\alpha,n-1} = \alpha \text{ fractile of } \chi^2(n-1)$$

- $(1-\alpha)100\%$ confidence interval for s^2:

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

Variating chocolate bars

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

- the sample standard deviation is 10 calories.

Calculate 90% and 95% confidence intervals for the variance.

Difference in means
Known variances

- Two populations:

$$X_{1,i} \sim N(\mu_1, \sigma_1^2)$$
$$X_{2,i} \sim N(\mu_2, \sigma_2^2)$$

- Two samples:

$$x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}$$
$$x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}$$

- Estimate of $\mu_1 - \mu_2$:

$$\overline{x}_1 - \overline{x}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1,i} - \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2,i}$$

- Confidence interval:

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

Test of two means

Unknown & equal variances

- Degrees of freedom: $\nu = n_1 + n_2 - 2$
- Pooled variance estimate:

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

- Confidence interval:

$$(\overline{x}_1 - \overline{x}_2) + t_{\alpha/2,\nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{x}_1 - \overline{x}_2) - t_{\alpha/2,\nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Test of two means

Unknown & unequal variances

- Degrees of freedom:

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

- Confidence interval:

$$(\overline{x}_1 - \overline{x}_2) + t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{x}_1 - \overline{x}_2) - t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Likelihood function
The general approach

- Joint density function of $X_1, X_2, \ldots, X_n$:

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

- θ is the parameter (vector) of f = parameter of interest.
- The likelihood function:

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

- The log-likelihood function:

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

- Notice:

$$\text{Density: } (x_1, x_2, \ldots, x_n) \mapsto f(x_1, x_2, \ldots, x_n; \theta) \quad (\theta \text{ fixed})$$
$$\text{Likelihood: } \theta \mapsto f(x_1, x_2, \ldots, x_n; \theta) \quad (\text{data fixed})$$

Likelihood function

- Maximum likelihood estimate (MLE):

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(x_1, x_2, \ldots, x_n; \theta)$$

- MLE is not necessarily unique
- Exact optimization can be difficult
- Numerical optimization can be
  - time consuming to run
  - time consuming to program
- Easier with independent observations:

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Likelihood function: Example

- Independent observations: $x_1, x_2, \ldots, x_n$, $X_i \sim N(\mu, \sigma^2)$
- Parameter vector: $\theta = (\mu, \sigma^2)$.
- Likelihood function:

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2}\Big)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\Big)$$

- Log-likelihood function:

$$l(\theta; x_1, x_2, \ldots, x_n) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

- Maximum likelihood estimate:

$$\mu = \overline{x}, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \neq s^2$$

Matlab

- $(1-\alpha)100\%$ Confidence interval for mean with known variance:
  ```
  mean(x) + [-1 1] * norminv(1-alpha/2) * std(x) / sqrt(n)
  ```
- $(1-\alpha)100\%$ Confidence interval for mean with unknown variance:
  ```
  mean(x) + [-1 1] * tinv(1-alpha/2, n-1) * std(x) / sqrt(n)
  ```
- $(1-\alpha)100\%$ Confidence interval for variance:
  ```
  (n-1)*std(x)^2 ./ chi2inv( [alpha/2 1-alpha/2], n-1 )
  ```