Design and analysis of experiments
Lecture 2

Jakob G. Rasmussen
Department of Mathematics
Aalborg University
Denmark

Populations and samples

- Population: the set of all individuals of interest
- Sample: a subset of the population - our observed data

- Parameter: quantity describing the population - e.g. mean or variance
- Estimator: quantity estimating a parameter using a sample (this is a random variable)
- Estimate: particular value of an estimator obtained from a sample (this is a realisation of a random variable)

Parameters and estimators - examples

- Sample: $y_1, \ldots, y_n$ - independent and identically distributed

	Parameter	Estimator
Mean	$\mu = \int_{-\infty}^{\infty} y f(y) dy$	$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
Variance	$\sigma^2 = \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy$	$\bar{y} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$
Std. dev.	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Sample mean

- The mean of $\bar{y}$:

$$\mathbb{E}[\bar{y}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[y_i] = \frac{1}{n}n\mu = \mu$$

- Unbiased estimator: the mean of the estimator (sample mean) equals the parameter (population mean)
- The variance of $\bar{y}$:

$$\mathrm{Var}(\bar{y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[y_i] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

- Consistent estimator: the variance of the estimator goes to zero when n goes to infinity
- By the central limit theorem we get for large n

$$\bar{y} = N\left(\mu, \frac{\sigma^2}{n}\right)$$

One series of observations

- The simplest experiment is one series of observations.
- Observations/sample: $y_1, \ldots, y_n$
- Assumptions:
  - Independence
  - Normally distributed, $N(\mu, \sigma^2)$
- Typical questions:
  - Can we estimate μ and σ^2 from the data?
  - How precise are these estimates?
  - Can we answer hypotheses using the estimates?

Estimation of variance

- Standardizing: $z_i = \frac{y_i - \mu}{\sigma} \sim N(0,1)$
- $$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \sim \chi_n^2$$
- We do not know μ, so instead we use $\bar{y}$:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{SS}{\sigma^2} \sim \chi_{n-1}^2$$

- Remark: When we exchange a parameter with estimate we lose one degree of freedom.

Estimation of variance

- Estimate of variance:

$$s^2 = \frac{SS}{n-1} = \frac{\sigma^2}{n-1}\frac{SS}{\sigma^2} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

- s^2 is an unbiased estimate of the variance σ^2:

$$\mathbb{E}[s^2] = \frac{\sigma^2}{n-1}\mathbb{E}[\chi^2] = \sigma^2$$

- If we had used $1/n$ instead of $1/(n-1)$ in the definition of s^2, then the estimator would be biased:

$$\tilde{s}^2 = \frac{SS}{n} \qquad \mathbb{E}[\tilde{s}^2] = \frac{n-1}{n}\sigma^2$$

Estimation of mean

- Ideally we would use:

$$z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- But we do not know σ, so we exchange it by s:

$$\begin{aligned} t &= \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{s^2/\sigma^2}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(SS/\sigma^2)/(n-1)}} \\ &\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = t_{n-1} \end{aligned}$$

- Since the t-distribution has mean 0, $\bar{y}$ is an unbiased estimator for μ

Confidence intervals

- Now we have point estimates $\bar{y}$ and s^2 of μ and σ^2, but we do not know how precise these estimates are.
- For this we can use confidence intervals.
- A confidence interval is an interval such that for a small value α (typically 5%), there is only a probability of $\alpha/2$ that the actual parameter lies outside either end of the interval.
- In other words, we are $(1-\alpha)$ (typically 95%) confident that the true value of the parameter lies inside this interval.

Confidence intervals for mean and variance

- Calculations for mean:

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}} \Rightarrow \mu = \bar{y} - t\frac{s}{\sqrt{n}}$$

- Confidence interval for μ:

$$\bar{y} - t_{n-1;\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}$$

- Calculations for variance:

$$s^2 = \frac{\sigma^2}{n-1}\chi^2 \Rightarrow \sigma^2 = \frac{(n-1)s^2}{\chi^2} = \frac{SS}{\chi^2}$$

- Confidence interval for σ^2

$$\frac{SS}{\chi^2_{n-1;\alpha/2}} < \sigma^2 < \frac{SS}{\chi^2_{n-1;1-\alpha/2}}$$

Hypothesis testing

- Often we want to test whether a hypothesis is true
- We formulate this as a null hypothesis (H_0) vs. an alternative hypothesis (H_1)
- For example, a hypothesis could be:

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

- We calculate a test statistic from the sample and compare with its theoretical distribution to check whether to accept the hypothesis or not

Hypothesis testing - mean

- Hypothesis:

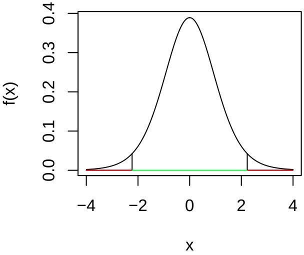
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- Test statistic:

$$t_0 = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

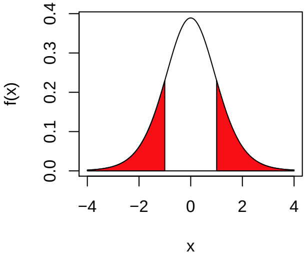
- We accept H_0 if the test statistics is within acceptance region:

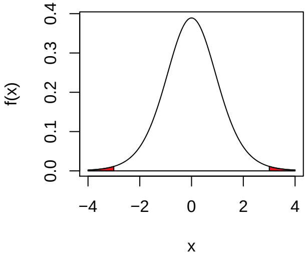
$$-t_{n-1;\alpha/2} < t_0 < t_{n-1;\alpha/2}$$



Hypothesis testing - the p-value

- An alternative to comparing the test statistic with the acceptance region: Calculate the p-value
- Definition: the p-value is the probability of a more extreme result than the one observed, assuming H_0 true
- A small p-value means the test statistic is rather far out, i.e. we reject H_0
- More precisely: reject H_0 if $p < \alpha$, otherwise accept H_0





Hypothesis testing - the p-value

- Different p-values tell different stories - how significant is a rejection of H_0?
- This is usually the standard:

$p \geq 0.10$	Not significant
$0.05 \leq p < 0.10$	Almost significant
$0.01 \leq p < 0.05$	Significant
$0.001 \leq p < 0.01$	Highly significant
$p < 0.001$	Very highly significant

- Note: choosing $\alpha = 0.05$ is the same as rejecting H_0 when it is "significantly" wrong - choosing $\alpha = 0.01$ means we want to be much more sure before rejecting a hypothesis.

Hypothesis testing - variance

- Hypothesis:

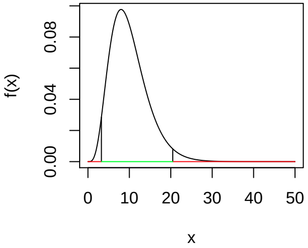
$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

- Test statistic:

$$\chi_0^2 = \frac{SS}{\sigma_0^2} \sim \chi_{n-1}^2$$

- Acceptance region:

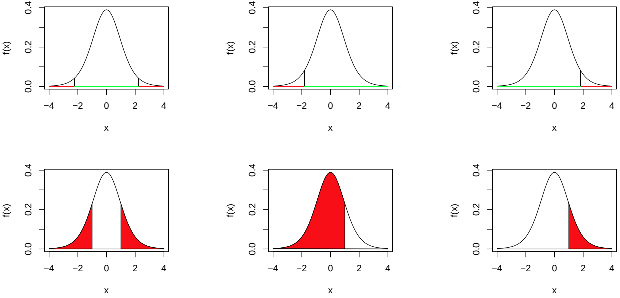
$$\chi^2_{n-1;\alpha/2} < \chi_0^2 < \chi^2_{n-1;1-\alpha/2}$$



One and two-sided tests

- Most tests come in both one- and two-sided versions.
- There are few differences between these, fx. the test statistics are the same.
- Critical values and p-values are found in different ways:

Two-sided:	One-sided 1:	One-sided 2:
$H_0 : \mu = \mu_0$	$H_0 : \mu \geq \mu_0$	$H_0 : \mu \leq \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$



R

- R-demo, part 1
- Exercise

Two series of observations

- Observations: $y_{11}, \ldots, y_{1n_1}$ and $y_{21}, \ldots, y_{2n_2}$
- Assumptions:
  - Independence within samples
  - Independence between samples
  - Normally distributed data:

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

- Typical questions:
  - Do the two samples have different means?
  - Do the two samples have different variances?

Test for difference of means

- For now, we assume variance homogeneity, i.e. $\sigma^2 = \sigma_1^2 = \sigma_2^2$
- Hypothesis of equal means:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

- Consider $\bar{y}_1 - \bar{y}_2$
- Mean: $\mathbb{E}[\bar{y}_1 - \bar{y}_2] = \mu_1 - \mu_2$
- Variance: $\mathrm{Var}(\bar{y}_1 - \bar{y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$
- Thus

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

- But again we do not know σ - we need to estimate this.

Test for difference of means

- We estimate σ from s_1 and s_2.
- Pooled variance estimate:

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

- Thus

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Test and confidence interval

- Test: accept H_0 (equality of means), if

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in [-t_{n_1+n_2-2;\alpha/2}, t_{n_1+n_2-2;\alpha/2}]$$

- Confidence interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2;\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Test for equal variances

- We assumed that the variances were equal - this should be tested!
- Hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- We use the following:

$$F = \frac{s_1^2}{s_2^2} \sim \frac{\sigma_1^2/(n_1-1)\chi^2_{n_1-1}}{\sigma_2^2/(n_2-1)\chi^2_{n_2-1}} = \frac{\sigma_1^2}{\sigma_2^2} F_{n_1-1,n_2-1}$$

- Acceptance region:

$$F_0 = \frac{s_1^2}{s_2^2} \in [F_{n_1-1,n_2-1;\alpha/2}, F_{n_1-1,n_2-1;1-\alpha/2}]$$

Difference of means, unequal variances

- If the variances are not equal, we need to change the t-test slightly:

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

- Degrees of freedom:

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

- This is often called Welch's test, and the t-distribution used is only approximate.

R

- R-demo, part 2
- Exercise