Bioinformatics and Biostatistics BB2440: Biostatistics Lecture 5: Statistical Tests Timo Koski

TK

12.09.2013

Outline of Lecture 5.

- Statistical Test: What is it?

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
    - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
    - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
    - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means
    - Two groups, t-test for difference of means

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means
  - Two groups, t-test for difference of means
  - Matched pairs or the paired t-test

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means
  - Two groups, t-test for difference of means
  - Matched pairs or the paired t-test
- Structure and Logic of a Statistical Test (2)

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means
  - Two groups, t-test for difference of means
  - Matched pairs or the paired t-test
- Structure and Logic of a Statistical Test (2)
  - Type I Error, Type II Error, Power of a Test

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means
  - Two groups, t-test for difference of means
  - Matched pairs or the paired t-test
- Structure and Logic of a Statistical Test (2)
  - Type I Error, Type II Error, Power of a Test
- Two Additional Tests

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means
  - Two groups, t-test for difference of means
  - Matched pairs or the paired t-test
- Structure and Logic of a Statistical Test (2)
  - Type I Error, Type II Error, Power of a Test
- Two Additional Tests
  - **"goodness of fit"**: χ^2-test

KTH Matematik

Outline of Lecture 5.

- Statistical Test: What is it?
- Structure and Logic of a Statistical Test (1)
  - hypothesis, null hypothesis, alternative hypothesis, test statistic, significance level, critical value, critical region, rare event rule
  - Decision: Critical Region, P-value, Confidence Interval, Alternative
- Test of Differences of Means
  - Two groups, t-test for difference of means
  - Matched pairs or the paired t-test
- Structure and Logic of a Statistical Test (2)
  - Type I Error, Type II Error, Power of a Test
- Two Additional Tests
  - **"goodness of fit"**: χ^2-test
  - Nonparametric: Wilcoxon rank-sum test

KTH Matematik

Statistical Test: What is it ? Recall Mendel's Peas

Gregor Mendel obtained by crossing peas with green pods and peas with yellow pods the offspring of 580 peas. Among them 428 had green pods, and 152 had yellow pods.
We wanted to infer, e.g., the **proportion of yellow pods that would be obtained in similar experiments**. We know this proportion as the population parameter denoted p. The best estimate of p based on this data is $\widehat{p} = \frac{152}{580} = 26.2\%$ The theory or **hypothesis** of Mendel was that 25 % of the peas would have yellow pods.

What is it ? Recall the Statement

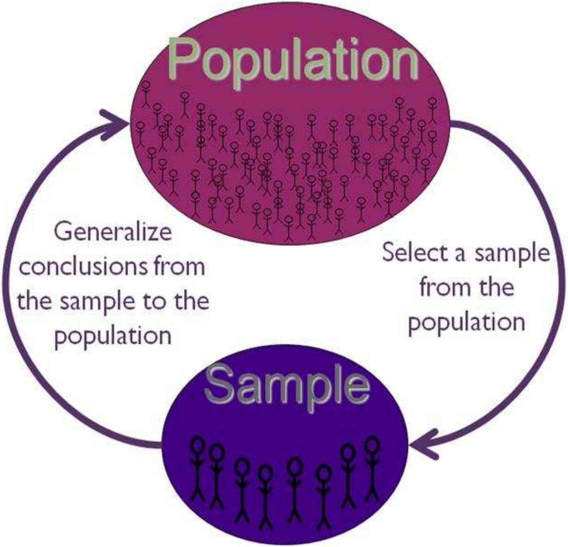
The CI was found as

$$0.226 < p < 0.298$$

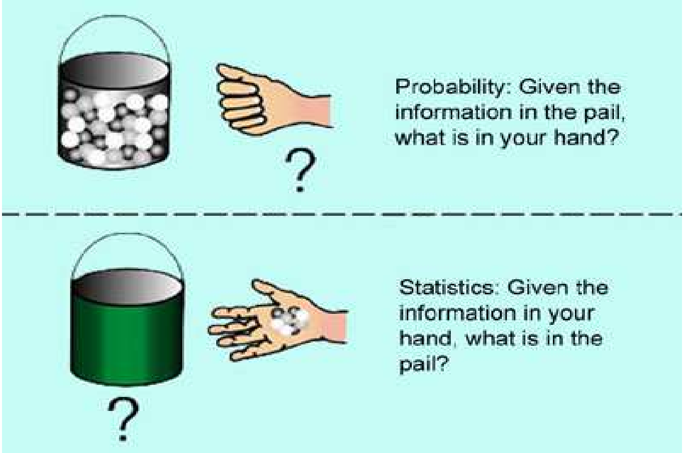
This is 95 % confidence that the limits 22.6 % and 29.8 % contain the true percentage of offspring peas with yellow pods. That interval includes 25%, so Mendel's expected value of 25% cannot be described as wrong. The results do not apper to provide significant evidence against the 25% rate in Mendel's hypothesis.

This is an example of what is called a **statistical test**. We proceed by finding out in which sense the conclusion above contains the structure and logic of a statistical test.

Statistical Inference


Population
Select a sample from the population
Sample
Generalize conclusions from the sample to the population

Statistical Inference: Random sample: no rigging, no selection bias



Probability: Given the information in the pail, what is in your hand?

Statistics: Given the information in your hand, what is in the pail?

The Basics of a Statistical Test

Definition

In statistics, a **hypothesis** *is a claim or statement about a property of a population. A* **hypothesis test** *(or a significance test) is a standard procedure for checking a claim about a property of a population.*

Example of a hypothesis:

- Mendel claims that under certain circumstances, the percentage of the offspring peas with yellow pods is 25 %.

Standard procedure contains: null hypothesis, alternative hypothesis, test statistic, significance level, critical value, decision rule, proceed assuming null hypothesis is true, with the logic of the rule of rare event.

The Individual Components of a Statistical Test

- Given a claim identify a statistical model for your population and the **null hypothesis** and the **alternative hypothesis**.
- Given a claim and sample data, compute the value of the **test statistic**
- Given **significance level**, identify the **critical values**.
- Given a value of a test statistic, identify significance level, identify the P-**value**.

We shall now explain what this is.

The Structure and Logic a Statistical Test: Null and Alternative Hypothesis

- The **null hypothesis** (denoted by H_o) is a statement that the value of a population parameter (such as proportion, mean, standard deviation) is equal to some claimed value. Examples:

  $$H_0 : p = 0.5 \quad H_0 : \mu = 0.86$$

  We **assume that** H_0 **is true and reach a conclusion either to reject** H_0 **or fail to reject** H_0.

- The **alternative hypothesis** (denoted by H_1) is a statement that the value of a population parameter is somehow different from the null hypothesis.

  $$H_1 : \mu > 0.5 \quad H_1 : \mu < 0.5 \quad H_1 : \mu \neq 0.5$$

KTH Matematik

The Structure and Logic a Statistical Test: Test Statistic

The **test statistic** is a value computed from the sample data and it is used to make the decision about the rejection of the null hypothesis. The test statistic is used to for determining whether there is significant evidence against the null hypothesis. Examples:

- Test statistic for proportion: $z = \frac{\widehat{p}-p}{\sqrt{\frac{pq}{n}}}$.
- Test statistic for mean: $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ or $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$.

The Structure and Logic of a Statistical Test: a Statistical Test: Critical Region, Significance Level

- The **critical region** (or **rejection region**) is the set of all values of the test statistic that cause us to reject the null hypothesis.
- The **significance level** denoted by α is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true. Common choices of α are 0.05, 0.01 and 0.10.
- A **critical value** is any value that separates the critical region, where we reject the null hypothesis, from the values of the test statistic that do not lead to rejection of the null hypothesis.

Two-tailed, Left-tailed, Right-tailed

The *tails* in a density curve are the regions bounded by critical values. Some hypothesis tests are two-tailed, some are right-tailed and some are left-tailed.

- **Two-tailed test:** the critical region in two parts, significance level is a sum of two areas under the density curve. The significance level is divied equally between the two tails. (Most of the tests in this lecture are of this kind).
- **Left-tailed test:** the critical region in the left tail of the density curve.
- **Right-tailed test:** the critical region in the right tail of the density curve.

KTH Matematik

Example: Mendel's Peas & Hypothesis Again

$$H_0 : p = 0.25 \quad H_1 : p \neq 0.25$$

We have $\widehat{p} = \frac{152}{580} = 26.2\%$. We evaluate the test statistic using the null hypothesis, i.e., assuming that the null hypothesis is true, i.e., $p = 0.25$

$$z = \frac{\widehat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.262 - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{580}}} = 0.67$$

Let us take $\alpha = 0.05$. The critical value is $z_{0.025} = 1.96$ (a two tailed critical region).

Step One: Critical values

Critical values

$P(Z > z_\alpha) = \alpha$ där $Z \in N(0,1)$

α	z_α	α	z_α
0.10	1.2816	0.001	3.0902
0.05	1.6449	0.0005	3.2905
0.025	1.9600	0.0001	3.7190
0.010	2.3263	0.00005	3.8906
0.005	2.5758	0.00001	4.2649

Example: Mendel's Peas & Hypothesis Again

$$H_0 : p = 0.25 \quad H_1 : p \neq 0.25$$

We have, if H_o is true (c.f. Lecture 3.)

$$Z = \frac{\widehat{p} - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{n}}} \quad \text{approximatively} \sim N(0,1)$$

Let us take $\alpha = 0.05$. The critical value is $z_{0.025} = 1.96$ (a two tailed critical region). Then

$$Pr\left(-1.96 \leq Z \leq 1.96\right) = 1 - 0.05 = 0.95.$$

Example: Mendel's Peas & Hypothesis Again

The critical region has two parts (tails)

$$z < 1.96 \quad \text{or } z > 1.96$$

Since the computed value of test statistic is $z = 0.67$, it does not lie in the critical region, and we fail to reject H_o, which was Mendel's claim made hypothesis. □

Recall: **Rare Event Rule**

If, under a given assumption, the probability of an observed event is very small, we conclude that the assumption is likely not correct.

Rare Event Rule, Hypothesis Testing

If, under $H_0 : p = 0.25$, the probability of the test statistic falling in the critical region is very small, we conclude that $H_0 : p = 0.25$ is likely not correct.

We have a under $H_0 : p = 0.25$ that

$$Z = \frac{\widehat{p} - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{n}}} \quad \text{approximatively} \sim N(0,1)$$

and therefore

$$Pr\left(-1.96 \leq Z \leq 1.96\right) = 1 - 0.05 = 0.95 = 1 - \alpha.$$

or the probability of finding Z in the critical region is

$$Pr\left(Z > -1.96\right) + Pr\left(Z \leq 1.96\right) = 0.025 + 0.025 = 0.05 (= \alpha).$$

Decision Criteria

- **Basic**: *Reject H_o if the test statistic falls within the critical region. Fail to reject H_o if the test statistic does not fall within the critical region.*
- **P-value method**: *Reject H_o if P-value $\leq \alpha$, where α is the significance level, e.g., 0.05. Fail to reject H_o if P-value $> \alpha$*
- **Another option**: *Instead of using a significance level like $\alpha = 0.05$ identify the P-value and leave the conclusion to the reader or your boss.*
- **Confidence Intervals**: *Reject a claim that the population parameter has a value that is not included in the interval.*

P-value

The **P-value** *(or* **p-value** *or* **probability value***) is the probability of getting a value of the test statistic which is* at least as extreme *as the one representing the sample data. The null hypothesis is rejected if the P-value is very small, such as 0.05 or less.*

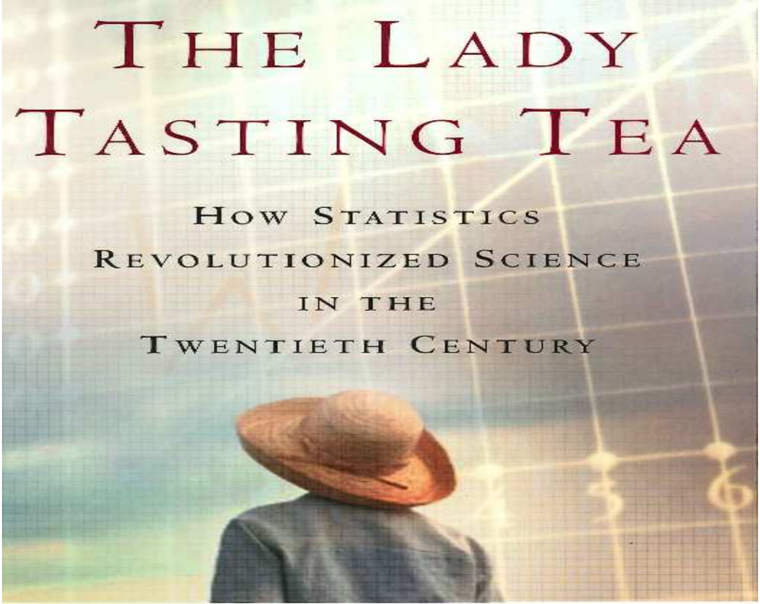
Example

With Mendel's peas,
$Pr(Z > 0.67) = 1 - Pr(Z \leq 0.67) = 1 - 0.749 = 0.251$, *and* **P-value** $= 2 \cdot 0.251 = 0.502$. *We multiply by two because we have a two-tailed test. Hence we fail to reject Mendel's H_o by P-value.*

Decision Criteria

When we used the 95 % confidence interval with the limits 22.6 % and 29.8 % and noted that this interval includes 25%, we failed to reject Mendel's expected value of 25% using the criterion of confidence intervals.

Lady tasting tea



Lady tasting tea

David Salsburg: *Lady Tasting Tea - How Statistics Revolutionized Science in the Twentieth Century* Holt McDougal, 2002-05-01.

Brief commercial:

(The book is) saluting the spirit of those who dared to look at the world in a new way, this insightful, revealing history explores the magical mathematics that transformed the world.

Test of Differences of Means

- Two groups, t-test for difference of means
- Matched pairs or the paired t-test

Two samples & The t-Test

We shall introduce methods for using sample data from two independent samples to test the hypotheses about population means via CI estimates of the difference between two population means.

Definition

Two samples are **independent** *if the sample values selected from one population are not related to or somehow paired or matched with the sample values selected from the other population.*

KTH Matematik

Two samples

Statistical model:

$$\begin{array}{lcll} X_1, X_2, \ldots, X_{n_1} & \sim & N(\mu_1, \sigma_1) & \text{(sample 1)} \\ Y_1, Y_2, \ldots, Y_{n_2} & \sim & N(\mu_2, \sigma_2) & \text{(sample 2)} \end{array}$$

where all Xs and Ys are independent.

Two samples, test statistic for the difference between means: Case (a)

(a) σ_1 and σ_2 **known**

We plan to get a CI for $\mu_1 - \mu_2$, denoted by $I_{\mu_1 - \mu_2}$. A natural estimate of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$. We get the test statistic (Z score)

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

KTH Matematik

Two samples, significance level, critical region

(a) σ_1 and σ_2 **known**

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

implies that the two-tailed critical region with confidence level $1 - \alpha$ is

$$Z < -z_{\alpha/2} \quad \text{or} \quad Z > z_{\alpha/2}$$

Here $z_{\alpha/2}$ is a critical value obtained by the table above.

KTH Matematik

Two samples, confidence interval for the difference between means: Case (a)

Then, we are in the same situation as in Lect. 4, we are led to the CI for $\mu_1 - \mu_2$

$$\underline{I_{\mu_1-\mu_2} = \overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Here $z_{\alpha/2}$ is a critical value obtained by the table above.

KTH Matematik

Two samples, confidence interval for the difference between means: Case (a)

If $\sigma_1 = \sigma_2 = \sigma$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

and

$$I_{\mu_1 - \mu_2} = \overline{x} - \overline{y} \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

$z_{\alpha/2}$ is a critical value obtained by the table above.

Two samples, confidence interval for the difference between means: Case (a)

The Confidence Interval Estimate *of* $\mu_1 - \mu_2$ *with* $\sigma = \sigma_1 = \sigma_2$ *known is*

$$\overline{x} - \overline{y} - E < \mu_1 - \mu_2 < \overline{x} - \overline{y} + E$$

where

$$E = z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

where $z_{\alpha/2}$ *is a critical value obtained by the table above. With these we can use the confidence interval criterion of testing a hypothesis on* $\mu_1 - \mu_2$ *with confidence degree* $1 - \alpha$.

KTH Matematik

Two samples, confidence interval for the difference between means: Confidence interval criterion

$$H_o : \mu_1 - \mu_2 = 0 \quad H_o : \mu_1 - \mu_2 \neq 0$$

Find

$$(\overline{x} - \overline{y} - E, \overline{x} - \overline{y} + E)$$

If this interval includes zero, we fail to reject H_o.

KTH Matematik

Two samples, confidence interval for the difference between means: Case (b)

(b) $\sigma_1 = \sigma_2 = \sigma$ **unknown**

The estimate s of σ is, e.g., the square root of the *pooled* sample variance (s_1^2 and s_2^2 are the sample variances of sample one and two, respectively):

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

and the test statistic is

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has **t-distribution** with $n_1 + n_2 - 2$ **degrees of freedom**.

Two samples, confidence interval for the difference between means: Case (b)

The Confidence Intervall Estimate *of* $\mu_1 - \mu_2$ *is*

$$\overline{x} - \overline{y} - E < \mu_1 - \mu_2 < \overline{x} - \overline{y} + E$$

where

$$E = t_{\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

where $t_{\alpha/2}$ *is a critical value of* t *-distribution with* $n_1 + n_2 - 2$ *degrees of freedom and*

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

KTH Matematik

Two samples, confidence interval for the difference between means: Case (b)

The Simplified Confidence Intervall Estimate *of* $\mu_1 - \mu_2$ *is*

$$\overline{x} - \overline{y} - E < \mu_1 - \mu_2 < \overline{x} - \overline{y} + E$$

where

$$E = t_{\alpha/2} s$$

where $t_{\alpha/2}$ *is a critical value of t -distribution with* $n_1 + n_2 - 2$ *degrees of freedom and*

$$s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Simplified → *no pooling in* s.

Two samples, confidence interval for the difference between means: Case (b), Example

In an experiment designed to test the effectiviness of paroxetine for treating bipolar depression, subjects were measured using the Hamilton Depression scale with results as follows:

Placebo group	$n_1 = 43$	$\overline{x} = 21.57, s_1 = 3.87$
Paroxenite treatment group	$n_2 = 33$	$\overline{y} = 20.38, s_2 = 3.91$

Two samples, confidence interval for the difference between means: Case (b), Example of t-Test

In an experiment designed to test the effectiviness of paroxetine for treating bipolar depression.
Null hypothesis: there is no difference between treatment and placebo

$$H_o : \mu_1 = \mu_2$$

Alternative hypothesis: there is a difference between treatment and placebo

$$H_1 : \mu_1 \neq \mu_2$$

We now proceed under H_o, or $\mu_1 - \mu_2 = 0$. Significance level is $\alpha = 0.05$.

Two samples, confidence interval for the difference between means: Case (b), Example

We now proceed under H_o, or $\mu_1 - \mu_2 = 0$. We assume that we have independent samples and normal distributions. We must assume this, we have just summaries of data, no chance of looking at histograms or boxplots or normplots. We use the test statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two samples, confidence interval for the difference between means: Case (b), Example

Insert data and use $H_o : \mu_1 - \mu_2 = 0$ and the simplifed computation (no pooling)

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(21.57 - 20.38) - 0}{\sqrt{\frac{3.87^2}{43} + \frac{3.91^2}{33}}} = 1.321$$

Two samples, confidence interval for the difference between means: Case (b), Example

$$t = 1.321$$

Since we are using t-distribution, the critical numbers are found from the t-distribution with $43 + 33 - 2 = 74$ degrees of freedom. This gives us $t_{0.025} = 1.993$ (by $>>$tinv(0.025, 74) in Matlab). Hence the critical region is

$$t < -1.993 \quad \text{or} \quad t > 1.993$$

Two samples, confidence interval for the difference between means: the Statistical Statement

The critical region is

$$t < -1.993 \quad \text{or} \quad t > 1.993$$

Since the test statistic $= 1.321$ does not fall within the critical region, we fail to reject the null hypothesis $H_o : \mu_1 - \mu_2 = 0$.
Based on these results paroxenite treatment does not have a significant effect as a treatment for bipolar depression.

Matched Pairs

We shall deal with a testing situation " matched pairs " that is treacherously close to the comparison of two means above.
We present this by an example. Suppose that we want to test the effectiveness of a low-fat diet. The weight of n subjects is measured before and after the diet. The results are $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, respectively. Obviously x_i and y_i would be dependent, but the samples corresponding to different subjects are independent.

Matched Pairs

We have

Diet	Subject			
	1	2	…	n
weight before diet	x_1	x_2	…	x_n
weight after diet	y_1	y_2	…	y_n

Matched Pairs

Let us assume x_j for the jth subject is a sample from $N(\mu_j, \sigma_1)$ and y_j a sample from $N(\mu_j + \Delta, \sigma_2)$. Δ is the population mean difference for *all* matched pairs. Δ is the population parameter for the effectiveness of the low-fat diet.

Matched Pairs

- There is, as in case of two means, two series of observations. But the model for two means is inapplicable, the pairs x_j, y_j are now matched to each other, (two measurements of the weight of one and the same person). The data consists of n matched pairs.
- The unknown parameters are $\mu_1, \ldots, \mu_n$, σ_1, σ_2 och Δ.
- $\mu_1, \ldots, \mu_n$ reflect differences between subjects, whereas Δ reflects the systematic difference between the weights before and after the low fat diet. If $\Delta < 0$ then the weight after diet is in average lower than before the diet.
- Note that σ_1 and σ_2 can be different.

Matched Pairs

We are primarily interested in Δ. To do the analysis we need a trick, which is best illustrated by another example.

Matched Pairs: Example

A laboratory in a brewery takes daily samples of beer to analyse. Two chemists A and B analyse the alcoholic percentage in the samples. One asks if there was a systematic difference between A's and B's results. Daily, for *n* days, we let A and B, independently of each other, to analyse the same sample of beer, new sample per day.

Matched Pairs

Chemist	Beer sample			
	1	2	...	n
A	x_1	x_2	...	x_n
B	y_1	y_2	...	y_n

The Statistical Model:

$$\begin{array}{rcll} X_1, X_2, \ldots, X_n & \sim & N(\mu_i, \sigma_A) & \text{(A's results)} \\ Y_1, Y_2, \ldots, Y_n & \sim & N(\mu_i + \Delta, \sigma_B) & \text{(B's results)} \end{array}$$

$$\Delta = \text{ a systematic difference}$$

The trick

The trick is to form the **differences**

$$Z_i = Y_i - X_i$$

since then $Z_i \sim N(\Delta, \sigma)$ with $\sigma \left(= \sqrt{\sigma_A^2 + \sigma_B^2} \right)$. But now we have reduced the problem to a case with one sample and we can form confidence interval for Δ as we did for μ in Lect. 4.

Hypothesis testing for matched pairs

$\overline{z}$ is the mean value (=arithmetic mean of the samples z_i) of the differences $z_i = y_i - x_i$ of the matched pair data.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2}.$$

is the standard deviation for differences z_i of the matched data.
The test statistic is

$$t = \frac{\overline{z} - \Delta}{\frac{s}{\sqrt{n}}}$$

Hypothesis testing for matched pairs: critcal region

The test statistic is

$$t = \frac{\overline{z} - \Delta}{\frac{s}{\sqrt{n}}}$$

If

$$H_o : \Delta = 0, \quad H_1 : \Delta \neq 0$$

then

$$t = \frac{\overline{z}}{\frac{s}{\sqrt{n}}}$$

The critical values $t_{\alpha/2}$ are obtained from a t-distribution with $n-1$ degrees of freedom (se Lect. 4) The critical region is

$$t < t_{\alpha/2} \quad \text{or} \quad t > t_{\alpha/2}$$

KTH Matematik

Confidence Interval for Matched Pairs

$$\overline{z} - E < \Delta < \overline{z} + E$$

where

$$E = t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

If

$$H_o : \Delta = 0, \quad H_1 : \Delta \neq 0$$

Reject H_o if 0 is not in the interval.

Structure and Logic of Statistical Tests (General Discussion)

We take another look at the general principles.

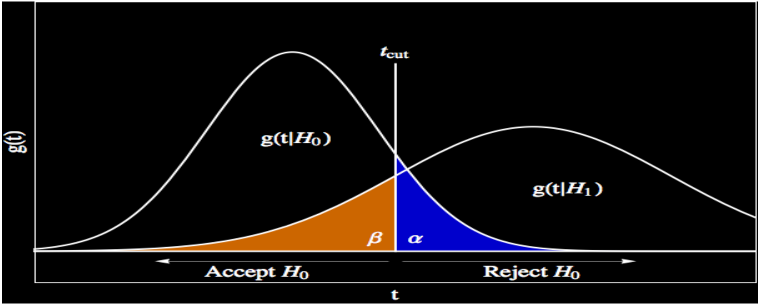
Structure and Logic of Statistical Tests (General Discussion)

When testing a null hypothesis, the statement is to either reject or fail to reject that null hypothesis. Such conclusions are sometimes correct and sometimes wrong, even if we do everything correctly in terms of computing margins of errors e.t.c., and have a good/perfect model of the population. *Thre are two types of errors:*

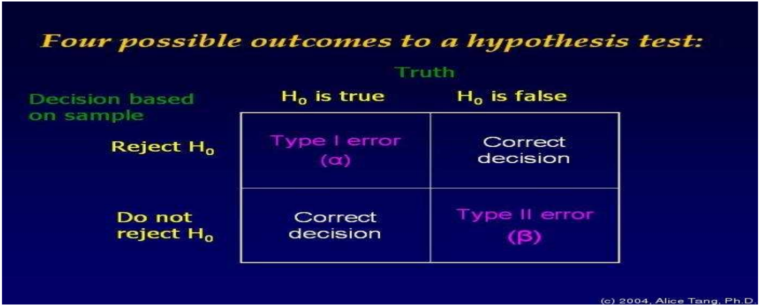
- **Type I error** *The mistake of rejecting the null hypothesis when it actually is true. The symbol α is used to represent the probability of type I error.*
- **Type II error** *The mistake of failing to reject the null hypothesis when it actually is false. The symbol β is used to represent the probability of type II error.*

α and β

Here we read $g(t \mid H_o)$ as the density curve in t, when H_0 is true, and $g(t \mid H_1)$ has a analogous meaning. Here, for e.g. $H_o : X \sim N(\mu, \sigma)$, $H_1 : X \sim N(\mu_1, \sigma_1)$, i.e., the alternative hypothesis is not a composite one. t_{cut} is the critical value.



Four possible outcomes of a test

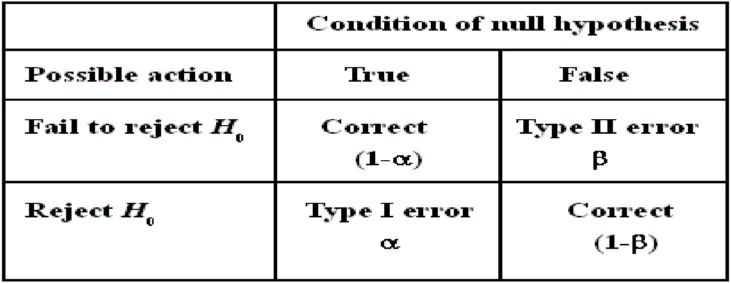


Four possible outcomes to a hypothesis test:

	Truth	
Decision based on sample	H_0 is true	H_0 is false
Reject H_0	Type I error (α)	Correct decision
Do not reject H_0	Correct decision	Type II error (β)

(c) 2004, Alice Tang, Ph.D.

Four possible outcomes of a test; c.f. rule of court



	Condition of null hypothesis	
Possible action	True	False
Fail to reject H_0	Correct $(1-\alpha)$	Type II error β
Reject H_0	Type I error α	Correct $(1-\beta)$

Controlling Type I error and Type II error (General Discussion)

- *For fixed α increase n, then β will decrease.*
- *For fixed n, decrease α, then β will increase.*
- *Increase n, decreases both α and β.*

Power of a Test

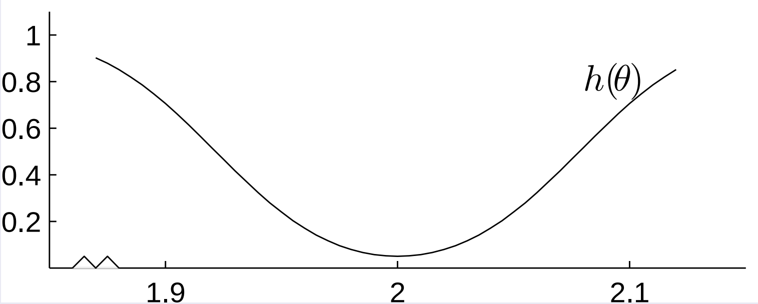
Definition

The **power** *of a hypothesis test is the probability* $1 - \beta$ *of rejecting a false null hypothesis. Power is computed by using a particular significance level* α*, a particular sample size* n*, the value of the population parameter used in the null hypothesis, and a particular value of the population parameter that is an alternative to the value assumed in the null hypothesis.*

Power of a Test: An Example

Power $h(\theta)$ of the test for $N(\theta, 0.04)$ when the level of significance is $\alpha = 0.05$

$$H_o : \theta = 2.0 \quad H_1 : \theta \leq 1.9 \quad \text{or} \quad \theta \geq 2.1$$



Two Additional tests

Two Additional Tests

- **"goodness of fit"**: χ^2*-test*
- *Nonparametric: Wilcoxon rank-sum test*

These are not tests of difference between means, but follow the logic and structure of a hypothesis test.

"goodness of fit" & χ^2-test

Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. For example, if you expected 10 of 20 offspring from a cross to be male and the actual observed number was 8 males, then you might want to know about the " goodness to fit" between the observed and expected. Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors. The chi-square test is always testing the **null hypothesis, which states that there is no significant difference between the expected and observed result** .

χ^2-test

A New Topic: χ^2-test is a **"goodness of fit"** -test.
χ^2 is pronounced like " chi square ".

χ^2-test

The simplest sitation:

A trial can have r different outcomes: $A_1, A_2, \ldots, A_r$. Let $x_1, x_2, \ldots, x_r$ be the frequencies by which the alternatives $A_1, A_2, \ldots, A_r$ occur in n trials.

χ^2-test

Let $p_1, p_2, \ldots, p_r$ be given probabilities, i.e. $p_i \geq 0$ and $\sum_{i=1}^{r} p_i = 1$. We want to test the hypotheses

$$H_o: \; P(A_i) = p_i \text{ för } i = 1, \ldots, r$$

against

$$H_1: \; \text{not all } P(A_i) = p_i.$$

χ^2-test

The test statistic is

$$Q = \sum_{i=1}^{r} \frac{(x_i - np_i)^2}{np_i} .$$

Note: $x_i =$ observed frequency, $np_i =$ expected frequency. Then $(x_i - np_i)^2$ is a 'fit'. It is not straightforward to give an intuitive justification of the np_is in the denominator.

KTH Matematik

χ^2-test

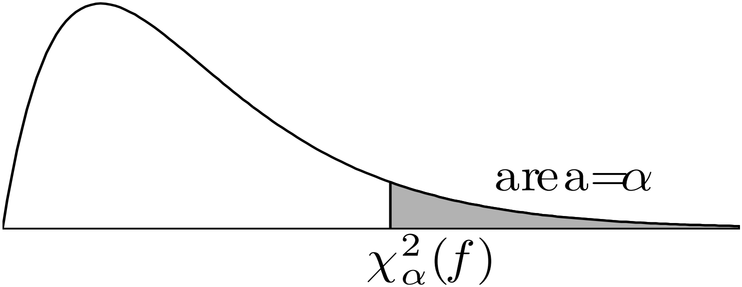
The test statistic is

$$Q = \sum_{i=1}^{r} \frac{(x_i - np_i)^2}{np_i}.$$

Q approximately $\sim \chi^2(r-1)$ if H_0 is true. $r-1$ is the number of degrees of freedom.

χ^2-distribution

$P(X > \chi^2_\alpha(f)) = \alpha$, where $X \in \chi^2(f)$.



χ^2-test

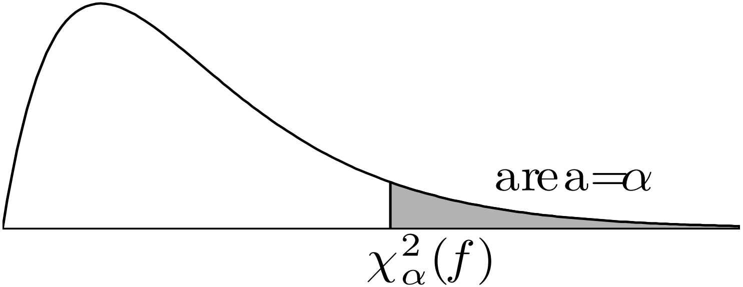
The critical region and decision are one-tailed:

- Reject H_o if
$$Q > \chi^2_\alpha(r-1).$$
- Otherwise you fail to reject H_o

If n is large, the approximative level of significance is α. Practical requirement is that all $np_j \geq 5$.

χ^2- distribution

$P(X > \chi^2_\alpha(f)) = \alpha$, where $X \in \chi^2(f)$.



χ^2-test: example

We have crossed two types of peas with round yellow seeds and wrinkled green seeds respectively. According to Mendelian genetics one should get four types of seeds with the following probabilities:

$$P(\text{round yellow}) = \frac{9}{16}, \qquad P(\text{wrinkled yellow}) = \frac{3}{16},$$
$$P(\text{round green}) = \frac{3}{16}, \qquad P(\text{wrinkled green}) = \frac{1}{16}.$$

χ^2-test: example

When an experiment was performed with 560 crossings of the round yellow and wrinkled green type the following data were obtained

330 round yellow,	100 wrinkled yellow,
112 round green,	18 wrinkled green.

Test, using the χ^2-method on the significance level 5 %, whether Mendelian genetics seem to be consistent with the experimental results.

χ^2-test

$$H_0 : P(\text{round yellow}) = \frac{9}{16}, \ldots$$

Note that $np_j \geq 5$ The test statistic is

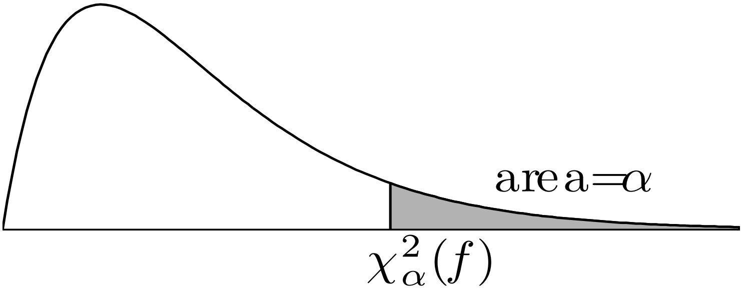
$$Q = \frac{(330 - 560\frac{9}{16})^2}{560\frac{9}{16}} + \frac{(100 - 560\frac{3}{16})^2}{560\frac{3}{16}} + \frac{(112 - 560\frac{3}{16})^2}{560\frac{3}{16}} + \frac{(18 - 560\frac{1}{16})^2}{560\frac{1}{16}}$$

$$= 9.68 > \chi^2_{0.05}(4-1) = \chi^2_{0.05}(3) = 7.81.$$

Table of $\chi^2_{0.05}(f)$ not shown. $\underline{H_0 \text{ is rejected.}}$

Recall: **Rare Event Rule**

If, under a given assumption, the probability of an observed event is very small, we conclude that the assumption is likely not correct.



Wilcoxon rank-sum test

The Wilcoxon rank-sum test is a non-parametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other.

Wilcoxon rank-sum test

$x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ are independent observations of the (continuous) random variables X and Y. We consider

H_0 : X och Y have the same density curve

H_1 : The density curve of of X:s is shifted w.r.t. the density curve of Y.

Note that this is **non-parametric**, since no parameters are involved.

Wilcoxon rank-sum test: the cookbook recipe

- First, arrange all the observations into a single ranked series. That is, rank all the observations without regard to which sample they are in. Replace the smallest with 1. The second smallest is replaced with 2, e.t.c, the largest is replaced with $n_1 + n_2$
- Now the x-observations are replaced by their *ranks* $r_1, \ldots, r_{n_1}$.
- The test statistic is the rank sum of the x-values

$$U \stackrel{\text{def}}{=} r_1 + \cdots + r_{n_1}$$

Wilcoxon rank sum test: the cookbook recipe

For large samples, U is approximately normally distributed, if H_o is true. The Z score is then

$$Z = \frac{U - \mu_U}{\sigma_U}, \text{ approximatively } \sim N(0,1)$$

where μ_U and σ_U are the mean and standard deviation of U.

$$\mu_U = \frac{n_1(n_1 + n_2 + 1)}{2}.$$

$$\sigma_U = \sqrt{\frac{n_1(n_1 + n_2 + 1)}{12}}.$$

Wilcoxon rank-sum test: Example

A dentist tried two anaesthetics *A* and *B* by injecting them in given doses in two different groups with 22 and 30 patients, respectively. Thereafter the time of anaesthesia in the surrounding soft body parts was registered for each patient. Results (unit: min):

A	195	240	154	95	65	82	132	155	125	119	155	345
	145	200	130	223	145	207	183	190	137	210		
	88	73	165	188	145	158	195	165	140	145	203	196
B	230	225	128	190	170	158	72	135	105	155	165	120
	138	125	188	145	208	75						

Wilcoxon rank-sum test: Example

Here the same data have been ranked:

A	65	82	95	119	125	130	132	137	145	145	154	155
	155	183	190	195	200	207	210	223	240	345		
B	72	73	75	88	105	120	125	128	135	138	140	145
	145	145	155	158	158	165	165	165	170	188	188	190
	195	196	203	208	225	230						

Wilcoxon rank-sum test: Example

The sum of the ranks of A-samples becomes

$$U = \sum_{i=1}^{22} r_i = 1 + 5 + 7 + 9 + 11.5 + 14 + \cdots + 52 = 602.5.$$

(Ties replaced mean rank.) If there is no difference in the times of anaesthesia $U \sim N(22 \cdot 53/2, \sqrt{22 \cdot 30 \cdot 53/12}) = N(583.0, 54.0)$.

KTH Matematik

Wilcoxon rank-sum test: Example

If you do not know in which direction the possible difference between A and B lies, we should make the test two-sided. With P-**value** we get, if H_o is true,

$$P(U > 602.5) \approx 1 - \Phi\left(\frac{602.5 - 583.0}{54.0}\right) = 0.36,$$

and $P = 2 \cdot 0.36 = 0.72 > 0.05$ follows. Hence we fail to reject the hypothesis that the times of anaesthesia are equal.

Wilcoxon rank-sum test: Example

In Norman and Streiner *Biostatistics, Bare Essentials* p. 261, $n_1 \leftrightarrow n$, $n_2 \leftrightarrow m$, $N = n + m = n_1 + n_2$. When they take the z-score, the use a **continuity correction** of size 1/2, which means that they evaluate $\Phi\left(\frac{602.5-583.0}{54.0}\right)$ as $\Phi\left(\frac{602.5-583.0-\frac{1}{2}}{54.0}\right)$

SUMMARY

We have seen four different tests of claims made in various situations: t-test for difference of two means, t-test for matched pairs, goodness-of-fit test (χ^2-test), Wilcoxon rank-sum test.
These tests are different but follow the standard procedure: null hypothesis, alternative hypothesis, test statistic, significance level, critical value, decision rule, proceed assuming null hypothesis is true, with the logic of the rule of rare event.