Dataanalyse - Repetition - Kursusgang 5

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

5. marts 2010

Stocastic variables

- X is discrete if it only takes countably many values.
- X is continuous if it takes uncountably many values.
- Examples:

Experiment	Stocastic variable	Type
Throw a die	# eyes	discrete
Throw two dice	$\sum$ eyes	discrete
Weigh a person	Weight	continuous
Measure men in DK	height	continuous

Discrete stochastic variables

- Probability function, $f(x)$:

$$f(x) = P(X = x)$$

- Mean value/**E**xpected value:

$$E(X) = \sum_{\text{outcome}} xf(x)$$

- Variance – the expected deviation from the mean value:

$$\begin{aligned}
\text{Var}(X) &= E\big((X - E(X))^2\big) = E(X^2) - E(X)^2 \\
&= \sum_{\text{outcome}} (x - E(X))^2 f(x) \\
&= \sum_{\text{outcome}} x^2 f(x) - E(X)^2
\end{aligned}$$

Continous stochastic variables

- Density function, $f(x)$:

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

- Mean value/**E**xpected value:

$$E(X) = \int_{\text{outcome}} xf(x)\,\mathrm{d}x$$

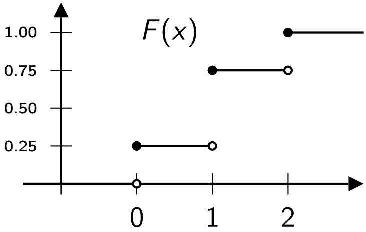
- Variance – the expected deviation from the mean value:

$$\begin{aligned}\mathrm{Var}(X) &= E\big((X - E(X))^2\big) = E(X^2) - E(X)^2 \\ &= \int_{\text{outcome}} (x - E(X))^2 f(x)\,\mathrm{d}x \\ &= \int_{\text{outcome}} x^2 f(x)\,\mathrm{d}x - E(X)^2\end{aligned}$$

Distribution function

- X is a discrete stocastic variable with probability function $f(x)$
- The cumulative distribution function for X:

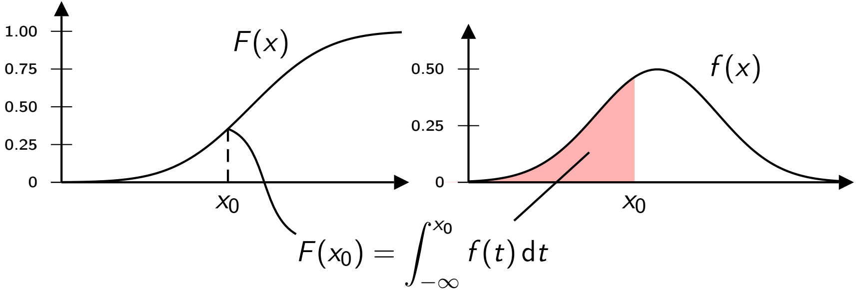
$$F_X(x) = P(X \leq x) = \sum_{y \leq x} f(y)$$



Distribution function

- X is a continuous stocastic variable with density function $f(x)$
- The cumulative distribution function for X:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$



Quantiles

- For $0 < p < 1$ the p quantile is the number(s) x_p where

$$F(x_p) = P(X \leq x_p) = p.$$

- The quantiles can be found in tables. E.g. for the standard normal distribution $N(0,1)$:

p	0.005	0.01	0.025	0.05	0.10	0.25	0.50
x_p	-2.58	-2.33	-1.96	-1.64	-1.28	-0.67	0.00
p	0.75	0.90	0.95	0.975	0.99	0.995	
x_p	0.67	1.28	1.64	1.96	2.33	2.58	

Distributions

- Uniform-distribution
- Binomial-distribution
- Normal-distribution
- χ^2-distribution
- t-distribution

Binomial distribution

- $X \sim B(n, p)$ if it is a sum of n independent "success/failure" experiments with succes probablility $0 \leq p \leq 1$.
- Probability function:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

- Mean and variance:

$$E(X) = n \cdot p \quad \text{and} \quad \mathrm{Var}(X) = n \cdot p \cdot (1 - p)$$

Binomial distribution

- $X \sim B(n, p)$ if it is a sum of n independent "success/failure" experiments with succes probablility $0 \le p \le 1$.
- Probability function:

number of ways to choose k of n

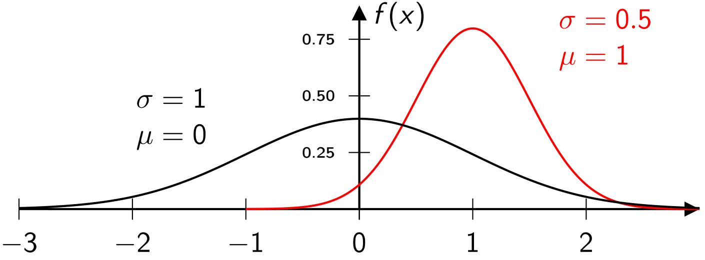
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

- Mean and variance:

$$E(X) = n \cdot p \quad \text{and} \quad \mathrm{Var}(X) = n \cdot p \cdot (1-p)$$

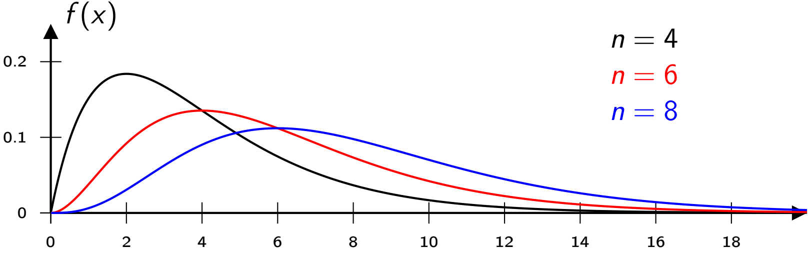
Normal distribution

- $X \sim N(\mu, \sigma^2)$, μ is the mean and σ^2 is the variance.
- $N(0,1)$ is the standard normal distribution.
- If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$.



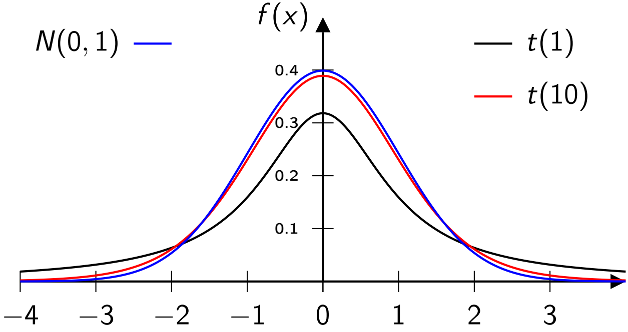
χ^2 distribution

- $X_1, X_2, \ldots, X_n$ are independent, standard normal distributed.
- The sum
$$X = \sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$
is χ^2 distributed with n degrees of freedom (notation: $X \sim \chi^2(n)$).
- Remember X is positive and $E(X) = n$.



t distribution

- For $X \sim t(n)$ we have
$$E(X) = 0 \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{n}{n-2}, \quad \text{for } n > 2.$$
- The larger n, the more $t(n)$ looks like $N(0,1)$.



χ^2 test for goodness-of-fit

- Data:

					Total
Class	1	2	...	k	
Observation	o_1	o_2	...	o_k	n
Expected observation	e_1	e_2	...	e_k	n

- Hypothesis:

$$H_0 : \text{Data follows certain distribution}$$
$$H_1 : \text{Data doesn't follow this distribution}$$

- Under H_0: o's $\approx$ e's.
- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(k - m - 1)$$

- Large values of χ^2 are critical for H_0.

χ^2 test for goodness-of-fit

- Data:

					Total
Class	1	2	...	k	
Observation	o_1	o_2	...	o_k	n
Expected observation	e_1	e_2	...	e_k	n

- Hypothesis:

$$H_0 : \text{Data follows certain distribution}$$
$$H_1 : \text{Data doesn't follow this distribution}$$

- Under H_0: o's $\approx$ e's.
- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(k-m-1)$$

- Large values of χ^2 are critical for H_0.

χ^2 test for goodness-of-fit

- Data:

					Total
Class	1	2	...	k	
Observation	o_1	o_2	...	o_k	n
Expected observation	e_1	e_2	...	e_k	n

- Hypothesis:

H_0 : Data follows certain distribution

H_1 : Data doesn't follow this distribution

- Under H_0: o's $\approx$ e's.
- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(k - m - 1)$$

- Large values of χ^2 are critical for H_0.

χ^2 test for goodness-of-fit

- Data:

					Total
Class	1	2	...	k	
Observation	o_1	o_2	...	o_k	n
Expected observation	e_1	e_2	...	e_k	n

- Hypothesis:

$$H_0 : \text{Data follows certain distribution}$$

$$H_1 : \text{Data doesn't follow this distribution}$$

- Under H_0: o's $\approx$ e's.
- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(k - m - 1)$$

- Large values of χ^2 are critical for H_0.

χ^2 test for goodness-of-fit

- Data:

Class	1	2	...	k	Total
Observation	o_1	o_2	...	o_k	n
Expected observation	e_1	e_2	...	e_k	n

- Hypothesis:

H_0 : Data follows certain distribution

H_1 : Data doesn't follow this distribution

- Under H_0: o's $\approx$ e's.
- Test statistic:

num. of parameters, estimated

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(k - m - 1)$$

- Large values of χ^2 are critical for H_0.

χ^2 test for goodness-of-fit

- Data:

					Total
Class	1	2	...	k	
Observation	o_1	o_2	...	o_k	n
Expected observation	e_1	e_2	...	e_k	n

- Hypothesis:

H_0 : Data follows certain distribution

H_1 : Data doesn't follow this distribution

- Under H_0: o's $\approx$ e's.
- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(k - \underbrace{m}_{\text{num. of parameters, estimated}} - 1)$$

- Large values of χ^2 are critical for H_0.

χ^2 test example

- A hospital performs a certain surgery on 5 new patients every day. The table below summarises for one year (365 days) the number of patients that survived the surgery each day:

0	1	2	3	4	5
1	10	56	117	131	50

- Assignment: Test at a 5% significance level the hypothesis that data come from a binomial distribution with $n = 5$ and $p = 0.72$.
- Under H_0 the expected table is calculated as $e_i = 365 \cdot p(i)$, where $p(i) = \binom{n}{i} p^i (1-p)^{n-i}$:

0	1	2	3	4	5
0.6	8.1	41.5	106.8	137.3	70.6

- Since the expected value is less than 5 in the first group, we collapse the two first groups and get:

≤ 1	2	3	4	5
8.7	41.5	106.8	137.3	70.6

χ^2 test example

- A hospital performs a certain surgery on 5 new patients every day. The table below summarises for one year (365 days) the number of patients that survived the surgery each day:

0	1	2	3	4	5
1	10	56	117	131	50

- Assignment: Test at a 5% significance level the hypothesis that data come from a binomial distribution with $n = 5$ and $p = 0.72$.
- Under H_0 the expected table is calculated as $e_i = 365 \cdot p(i)$, where $p(i) = \binom{n}{i} p^i (1-p)^{n-i}$:

0	1	2	3	4	5
0.6	8.1	41.5	106.8	137.3	70.6

- Since the expected value is less than 5 in the first group, we collapse the two first groups and get:

≤ 1	2	3	4	5
8.7	41.5	106.8	137.3	70.6

χ^2 test example

- A hospital performs a certain surgery on 5 new patients every day. The table below summarises for one year (365 days) the number of patients that survived the surgery each day:

0	1	2	3	4	5
1	10	56	117	131	50

- Assignment: Test at a 5% significance level the hypothesis that data come from a binomial distribution with $n = 5$ and $p = 0.72$.
- Under H_0 the expected table is calculated as $e_i = 365 \cdot p(i)$, where $p(i) = \binom{n}{i} p^i (1-p)^{n-i}$:

0	1	2	3	4	5
0.6	8.1	41.5	106.8	137.3	70.6

- Since the expected value is less than 5 in the first group, we collapse the two first groups and get:

≤ 1	2	3	4	5
8.7	41.5	106.8	137.3	70.6

χ^2 test example (cont'd)

- Calculate test statistic (Remeber to add the two first observations):

$$\chi^2 = \frac{(11-8.7)^2}{8.7} + \frac{(56-41.5)^2}{41.5} + \cdots + \frac{(50-70.6)^2}{70.6} \approx 12.9$$

- Degrees of freedom: $k-1=5-1=4$. Below is a $\chi^2(4)$ table:

p	0.10	0.25	0.5	0.75	0.90	0.95	0.975	0.99	0.995
x_p	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86

- Find critical value c such that $P(\chi^2 \geq c) = 0.05$. This is equivalent to $P(\chi^2 \leq c) = 0.95$. From the table we see that we reject H_0 if χ^2 is bigger than $c = 9.49$. (I.e. in this case we reject.)
- To approximate the p-value we use the table the other way. Since 12.9 is between 11.14 and 13.28 the p-value must be between $1-0.975=2.5\%$ and $1-0.99=1\%$ (the exact p-value from Matlab is 1.2%).

χ^2 test example (cont'd)

- Calculate test statistic (Remeber to add the two first observations):

$$\chi^2 = \frac{(11-8.7)^2}{8.7} + \frac{(56-41.5)^2}{41.5} + \cdots + \frac{(50-70.6)^2}{70.6} \approx 12.9$$

- Degrees of freedom: $k-1=5-1=4$. Below is a $\chi^2(4)$ table:

p	0.10	0.25	0.5	0.75	0.90	0.95	0.975	0.99	0.995
x_p	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86

- Find critical value c such that $P(\chi^2 \geq c) = 0.05$. This is equivalent to $P(\chi^2 \leq c) = 0.95$. From the table we see that we reject H_0 if χ^2 is bigger than $c = 9.49$. (I.e. in this case we reject.)
- To approximate the p-value we use the table the other way. Since 12.9 is between 11.14 and 13.28 the p-value must be between $1-0.975 = 2.5\%$ and $1-0.99 = 1\%$ (the exact p-value from Matlab is 1.2%).

χ^2 test example (cont'd)

- Calculate test statistic (Remeber to add the two first observations):

$$\chi^2 = \frac{(11-8.7)^2}{8.7} + \frac{(56-41.5)^2}{41.5} + \cdots + \frac{(50-70.6)^2}{70.6} \approx 12.9$$

- Degrees of freedom: $k-1 = 5-1 = 4$. Below is a $\chi^2(4)$ table:

p	0.10	0.25	0.5	0.75	0.90	0.95	0.975	0.99	0.995
x_p	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86

- Find critical value c such that $P(\chi^2 \geq c) = 0.05$. This is equivalent to $P(\chi^2 \leq c) = 0.95$. From the table we see that we reject H_0 if χ^2 is bigger than $c = 9.49$. (I.e. in this case we reject.)
- To approximate the p-value we use the table the other way. Since 12.9 is between 11.14 and 13.28 the p-value must be between $1-0.975 = 2.5\%$ and $1-0.99 = 1\%$ (the exact p-value from Matlab is 1.2%).

χ^2 test example (cont'd)

- Calculate test statistic (Remeber to add the two first observations):

$$\chi^2 = \frac{(11-8.7)^2}{8.7} + \frac{(56-41.5)^2}{41.5} + \cdots + \frac{(50-70.6)^2}{70.6} \approx 12.9$$

- Degrees of freedom: $k-1 = 5-1 = 4$. Below is a $\chi^2(4)$ table:

p	0.10	0.25	0.5	0.75	0.90	0.95	0.975	0.99	0.995
x_p	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86

- Find critical value c such that $P(\chi^2 \geq c) = 0.05$. This is equivalent to $P(\chi^2 \leq c) = 0.95$. From the table we see that we reject H_0 if χ^2 is bigger than $c = 9.49$. (I.e. in this case we reject.)
- To approximate the p-value we use the table the other way. Since 12.9 is between 11.14 and 13.28 the p-value must be between $1 - 0.975 = 2.5\%$ and $1 - 0.99 = 1\%$ (the exact p-value from Matlab is 1.2%).

χ^2 test example (cont'd)

- What if the assignment is: Test at a 5% significance level the hypothesis that data come from a binomial distribution with $n = 5$?
- Now the probability of success has to be estimated:

$$\hat{p} = \frac{\text{succeses}}{\text{trials}} = \frac{10 \cdot 1 + 56 \cdot 2 + 117 \cdot 3 + 131 \cdot 4 + 50 \cdot 5}{365 \cdot 5} = 0.68$$

- Then we recalculate expected counts with $p = 0.68$. The new test statistic is $\chi^2 = 1.2$. Now we have to compare with a $\chi^2(3)$ distribution since we have estimated a parameter. In this distribution the critical value is $c = 7.81$, and we therefore cannot reject H_0.

χ^2 test example (cont'd)

- What if the assignment is: Test at a 5% significance level the hypothesis that data come from a binomial distribution with $n = 5$?
- Now the probability of success has to be estimated:

$$\hat{p} = \frac{\text{succeses}}{\text{trials}} = \frac{10 \cdot 1 + 56 \cdot 2 + 117 \cdot 3 + 131 \cdot 4 + 50 \cdot 5}{365 \cdot 5} = 0.68$$

- Then we recalculate expected counts with $p = 0.68$. The new test statistic is $\chi^2 = 1.2$. Now we have to compare with a $\chi^2(3)$ distribution since we have estimated a parameter. In this distribution the critical value is $c = 7.81$, and we therefore cannot reject H_0.

χ^2 test example (cont'd)

- What if the assignment is: Test at a 5% significance level the hypothesis that data come from a binomial distribution with $n = 5$?
- Now the probability of success has to be estimated:

$$\hat{p} = \frac{\text{succeses}}{\text{trials}} = \frac{10 \cdot 1 + 56 \cdot 2 + 117 \cdot 3 + 131 \cdot 4 + 50 \cdot 5}{365 \cdot 5} = 0.68$$

- Then we recalculate expected counts with $p = 0.68$. The new test statistic is $\chi^2 = 1.2$. Now we have to compare with a $\chi^2(3)$ distribution since we have estimated a parameter. In this distribution the critical value is $c = 7.81$, and we therefore cannot reject H_0.

Estimators in $N(\mu, \sigma^2)$

- Estimator 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 stocastic variables with $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$.
- If the variance σ^2 is known:

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

- If the variance σ^2 is unknown:

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

Estimators in $N(\mu, \sigma^2)$

- Estimator 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 stocastic variables with $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$.
- If the variance σ^2 is known:

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

- If the variance σ^2 is unknown:

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

Estimators in $N(\mu, \sigma^2)$

- Estimator 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 stocastic variables with $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$.
- If the variance σ^2 is known:

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

- If the variance σ^2 is unknown:

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

Estimators in $N(\mu, \sigma^2)$

- Estimator 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 stocastic variables with $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$.
- If the variance σ^2 is known:

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

- If the variance σ^2 is unknown:

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

Example z-test

- H_0: Mean height is 180.
- We have observations from $n = 100$ people, $(x_1, \ldots, x_{100})$, and calculate $\overline{x} = 178$. Assume we know $\sigma^2 = 25$.
- Critical value (at 5% sig. level): 1.96 (found by looking up the 97.5% quantile in $N(0,1)$ table).
- Test statistic:

$$z = \frac{178 - 180}{\sqrt{25/100}} = -4$$

- Since $|z| = 4 > 1.96$ we reject H_0.
- Approximation of p-value: From the table we only know $P(Z \leq -2.58) = 0.005$ and $P(Z \leq 2.58) = 0.995$. Therefore $P(|Z| \geq 2.58) = 0.01$. I.e. we can only say the p-value is less than 1%.
- 95% confidence interval: $\overline{x} \pm z_{0.975} \cdot \sqrt{\sigma^2/n} = 178 \pm 0.98$.

Example t-test

- Assume σ^2 is unknown and $s^2 = 64$.
- Critical value (at 5% sig. level): 1.98 (found by looking up the 97.5% quantile in $t(99)$ table).
- Test statistic:

$$t = \frac{178 - 180}{\sqrt{64/100}} = -2.5$$

- Since $|z| = 2.5 > 1.98$ we reject H_0.
- Approximation of p-value: From a table we know $P(T \leq -2.63) = 0.005$ and $P(T \leq 2.63) = 0.995$. Therefore $P(|T| \geq 2.63) = 0.01$. I.e. the p-value is between 1% and 5%.
- 95% confidence interval: $\overline{x} \pm t_{0.975} \cdot \sqrt{s^2/n} = 178 \pm 1.59$.

Paired t-test

- Data:

$$\begin{array}{lllll} \text{Sample 1:} & x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ \text{Sample 2:} & x_{2,1} & x_{2,2} & \dots & x_{2,n} \end{array}$$

- Assumptions:
  - Observations occur in pairs, $(x_{1,i}, x_{2,i})$.
  - Each sample consists of independent, normally distributed observations, $X_{i,j} \sim N(\mu_i, \sigma_i^2)$.
- Note:
  - The two samples do not need to be independent.
  - Is often used in before-after experiments.
- Hypothesis:

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

Paired t-test

- Data:

$$\begin{array}{lcccc} \text{Sample 1}: & x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ \text{Sample 2}: & x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \hline \text{Difference:} & d_1 & d_2 & \dots & d_n \end{array}$$

$d_i = x_{1,i} - x_{2,i}$

- We have a new data set of differences $d_1, \dots, d_n$ which are normally distributed with unknown mean δ and unknow variance σ^2.
- Hypothesis:

$$H_0 \quad : \delta = 0$$
$$H_1 \quad : \delta \neq 0$$

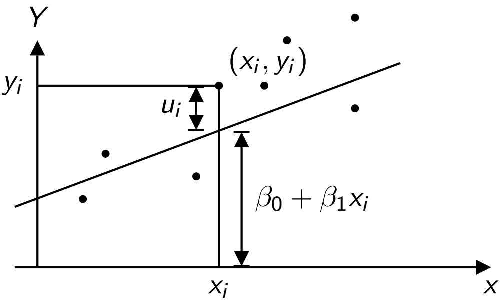
- Use usual t-test to test if $\delta = 0$.

Simple linear regression

- We assume a model where the stochastic variable Y depends linearly on the ordinary variable x:

$$Y = \beta_0 + \beta_1 x + U.$$

U is an error term with $U \sim N(0, \sigma^2)$.

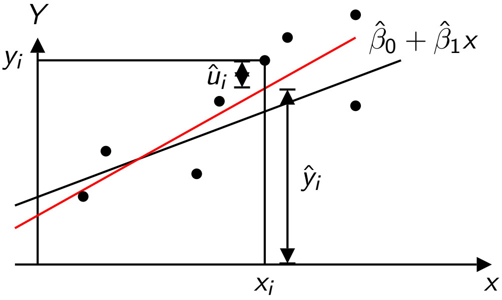


Estimates and estimated regression line

- The "least squares"-estimates are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}.$$

- The regression line is estimated by $\hat{y} = \hat{\beta}_0 + \beta_1 x$.
- **Predicted value:** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value for y_i.
- **Residual:** $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.



Coefficient of determination

- The proportion of the total variation that is explained is called *the coefficient of determination*. The easiest formula to calculate it is

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

Remember:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

- Interpretation: If for example $R^2 = 0.7$ we have that the model explains 70% of the variation of the y_i's. The remaining 30% correspond to random unexplained variation.

Test statistics

- The test statistics are:

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_0} \sim t(n-2),$$

where

$$\hat{\sigma}_0^2 = \frac{((n-1)s_x^2 + n\overline{x}^2)(s_y^2 - \hat{\beta}_1^2 s_x^2)}{n(n-2)s_x^2}.$$

and

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_1} \sim t(n-2),$$

where

$$\hat{\sigma}_1^2 = \frac{s_y^2 - \hat{\beta}_1^2 s_x^2}{(n-2)s_x^2}.$$

Hypothesis test and confidence interval

- We want to test the hypothesis
  - $H_0 : \beta_1 = K$
  - $H_1 : \beta_1 \neq K$

  Often we test with $K = 0$. This corresponds to x not having any influence on Y.
- Under H_0:
$$T_1 = \frac{\hat{\beta}_1 - K}{\hat{\sigma}_1} \sim t(n-2).$$
- Now we do exactly as before: Find a critical value c from the $t(n-2)$ table, and reject H_0 if $|T_1| > c$. The p-value is approximated as before by using the table in the reverse direction.
- A 95% confidence interval for β_1 is given by
$$\hat{\beta}_1 \pm t_{0.975}\hat{\sigma}_1,$$
  where $t_{0.975}$ is the 97.5% quantile found using the $t(n-2)$ table.