THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Analysis of Engineering and Scientific Data

Semester 1 – 2019

Sabrina Streipert s.streipert@uq.edu.au

Chapter 9:
Hypothesis testing

Question

Is the observed *data* is due to **chance**, or due to **effect**?

Formulate two(mutually exclusive) hypothesis:

Research Hypotheses

1. The *null* hypothesis H_0, which stands for our initial assumption about the data.
2. The *alternative hypothesis* H_1, (sometimes called H_A).

Hypothesis testing

Procedure:

1. Collect the data.
2. Formulate H_0 and H_1.
3. Based on the data, decide whether to reject or not reject H_0.

Open Question: How to decide whether to reject H_0?

	True state	
Decision	H_0 true	H_1 true
Accept H_0	OK	Type II Error (false negative)
Reject H_0	Type I Error (false positive)	OK

Statistical Test Errors

- The probability of a Type I Error is called the **significance level of the test**, denoted by α.

$$\alpha = P(\text{Type I Error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

(Common choice: $\alpha = 0.05$, i.e., accepting to have a 5% probability of incorrectly rejecting the null hypothesis.)

- The probability of a Type II Error is called the **power of the test**, denoted by β.

$$\beta = P(\text{type II error}) = P(\text{accept } H_0 \mid H_1 \text{ is true})$$

Power of the test = $1 - \beta$

$\longrightarrow$ **AIM:** α is low and power $(1 - \beta)$ is large.

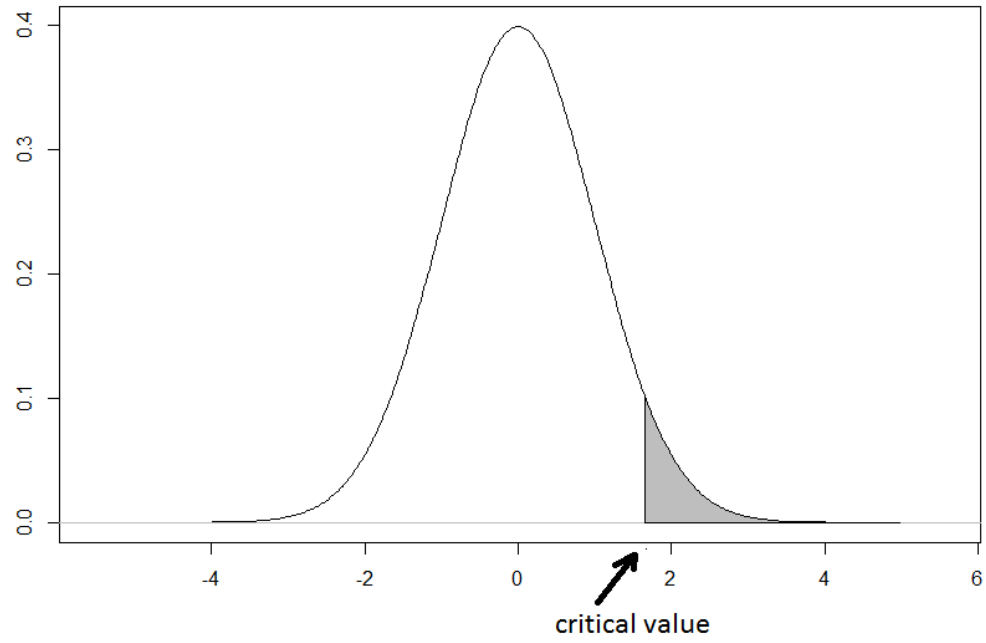
Open Question

How to decide whether to reject H_0?

- Let X be a random variable with range $\mathcal{X}$.
- Find an "appropriate" subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**.
  Often
  $$R = \{x : T(x) > c\},$$
  where T is some **test statistic** and c is called a **critical value**.
- Then decide via the rule:
  $$\begin{cases} X \in R \Rightarrow & \text{reject the null hypothesis } H_0 \\ X \notin R \Rightarrow & \text{do not reject the null hypothesis.} \end{cases}$$

$H_0: \mu = \mu_0$

$H_1: \mu > \mu_0$



- The area of the shaded area is α!
- If the observed test statistics falls into the shaded area, we **reject the null hypothesis**.

Example

An alien empire is considering taking over planet Earth, but they will only do so if the portion of rebellious humans is less than 10%. They abducted a random sample of 400 humans, performed special psychological tests, and found that 14% of the sample are rebellious. The true standard deviation is 0.25.

Solution:

$T = \bar{X}$, $H_0 : \mu = 0.1$, $\quad H_1 : \mu > 0.1$, $\quad \alpha = 0.05$

Step 1: Find the critical value:

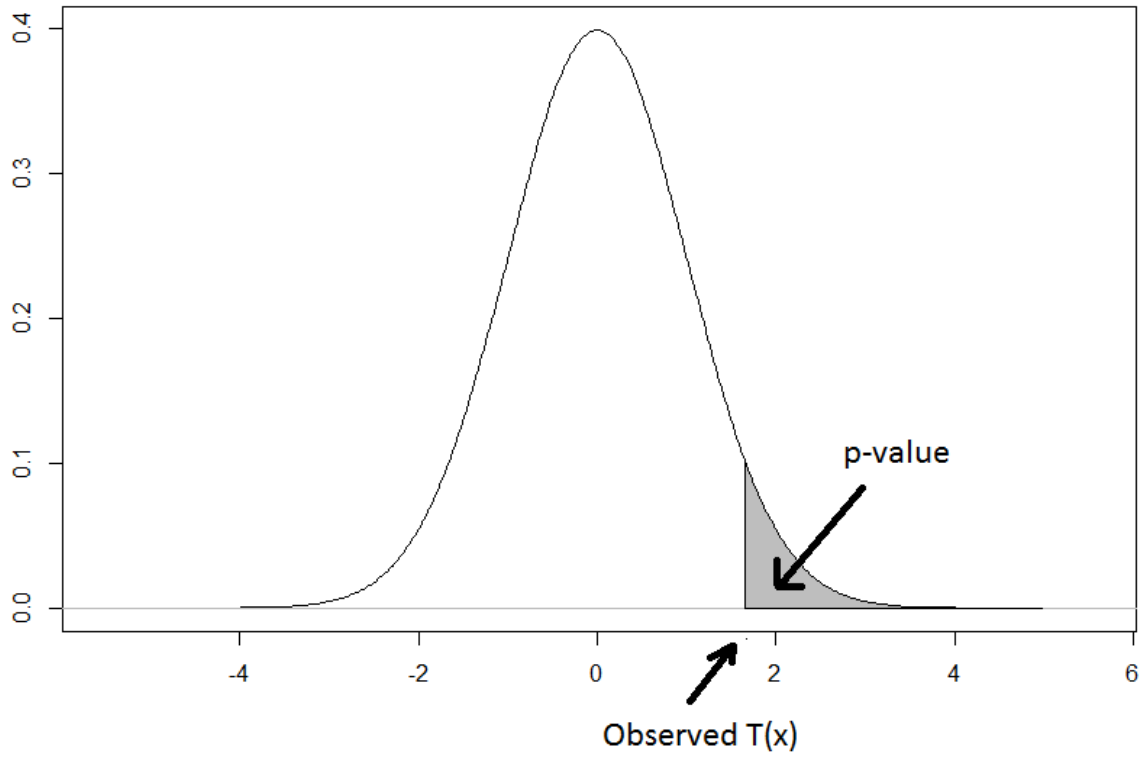
$$0.05 = 1 - \Phi\left(\frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \quad \Leftrightarrow \quad \Phi\left(\frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = 0.95$$

$$\Leftrightarrow \quad 1.65 = \frac{c - 0.1}{\frac{0.25}{\sqrt{400}}} \quad \Leftrightarrow \quad c = 0.1206$$

$\bar{x} = 0.14 > 0.1206 \Longrightarrow$ REJECT H_0

p-value Approach

The *p*-**value** is smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.



p-value Method

1. Collect data.
2. Give the null and alternative hypotheses (H_0 and H_1).
3. Choose an appropriate test statistic.
4. Determine the distribution of the test statistic under H_0.
5. Evaluate the outcome of the test statistic.
6. Calculate the p-value.
7. Accept or reject H_0 based on the p-value.

p-value low $\Rightarrow$ reject H_0

p-value Rule

P-value $\approx$ the probability of seeing data like the ones used for the test statistic given that H_0 is true.

p-value low $\Rightarrow$ reject H_0

p-value	evidence
< 0.01	very strong evidence against H_0
$0.01 - 0.05$	moderate evidence against H_0
$0.05 - 0.10$	suggestive evidence against H_0
> 0.1	little or no evidence against H_0

Example revisited

- Let $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, ($\sigma$ is known).
- We would like to test $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$.

$$p = P(\text{reject } H_0 \mid H_0 \text{ true}) = P_{H_0}(T > \bar{x}) = P\left(\frac{T-\mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$

$Z \sim N(0,1)$

$$= 1 - P\left(Z \le \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$

Alien example - revisited

Recall $\sigma = 0.25$

$n = 400$

Recall $\overline{X} = 0.14$. Define

$$T = \bar{X}, \quad H_0 : \mu = 0.1, \quad H_1 : \mu > 0.1$$

$$p = P(T > \bar{x}) = P\left(\frac{T - \mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{0.14 - 0.1}{0.25/\sqrt{400}}\right)$$

$$= 1 - \Phi\left(\frac{0.04}{0.25/20}\right) = 1 - \Phi(3.2) = 0.00069$$

$p < 0.01 \Rightarrow$ reject H_0

Would the alien's opinion change if they would have picked a different sample?

if $n=40 \Rightarrow$ p-value $= 1-\Phi\left(\frac{0.14-0.1}{0.25/\sqrt{40}}\right) =$

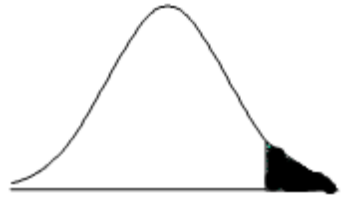
$= 1-\Phi(1.0119) = 0.1563$

p-value $> 0.1 \Rightarrow$ do not reject H_0

Types of tests

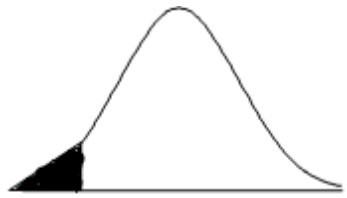
- **Right one-sided test:** where H_0 is rejected for the p-value defined by $P_{H_0}(T \geq t)$.

$R = \{T > c\}$



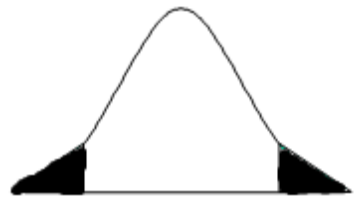
- **Left one-sided test:** where H_0 is rejected for the p-value defined by $P_{H_0}(T \leq t)$.

$R = \{T < c\}$



- **Two-sided test:** where H_0 is rejected for the p-value defined by $P_{H_o}(T \geq t) + P_{H_o}(T \leq -t) = 2P_{H_o}(T \geq t)$.

$R = \{T < -c\} \cup \{T > c\}$



Example:

A manufacturer produces crankshafts for an automobile engine. The wear of the crankshaft after 100.000 miles in 0.0001 inches is of interest due to warranty claims. A random sample of 15 shafts is tested with a sample mean of 2.78 10^{-4} inches and known standard deviation of 0.9. Test the claim that the expected wear is not equal to 0.0003 inches.

(2-sided test)

$H_0: \mu = 3$ $H_1: \mu \neq 3$

$\Rightarrow$ in units of 10^{-4} inches

$n=15$, $\sigma=0.9$ Pick $\alpha=0.05$

$R = \{T < \mu_0 - c\} \cup \{T > \mu_0 + c\}$

$$0.05 = P(\text{reject } H_0 \mid H_0 \text{ true}) = P_{\mu_0}(T < \mu_0 - c) + P_{\mu_0}(T > \mu_0 + c)$$

$$= P\left(\frac{T-\mu_0}{\sigma/\sqrt{n}} < \frac{-c}{\sigma/\sqrt{n}}\right) + P\left(\frac{T-\mu_0}{\sigma/\sqrt{n}} > \frac{c}{\sigma/\sqrt{n}}\right) =$$

$$= \underbrace{\Phi\left(\frac{-c}{0.9/\sqrt{15}}\right)}_{1-\Phi\left(\frac{c}{0.9/\sqrt{15}}\right)} + 1 - \Phi\left(\frac{c}{0.9/\sqrt{15}}\right) = 2\left(1-\Phi\left(\frac{c}{0.9/\sqrt{15}}\right)\right)$$

$$\Rightarrow \Phi\left(\frac{c}{0.9/\sqrt{15}}\right) = 1 - \frac{0.05}{2} = 0.975 \Rightarrow \frac{c}{0.9/\sqrt{15}} = 1.96$$

$$\Rightarrow c = 0.4555 \Rightarrow R = \{T < 2.5445\} \cup \{T > 3.4555\}$$

Since $\bar{x} = 2.78 \notin R \Rightarrow$ don't reject H_0

➢ **Testing Hypotheses on the Mean, Variance Known (Z-Tests)**

Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with μ unknown but σ^2 known.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}, \qquad Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\big[1 - \Phi\big(\lvert z\rvert\big)\big]$	$z > z_{1-\alpha/2}$ or $z < z_{\alpha/2}$
$H_1 : \mu > \mu_0$	$P = 1 - \Phi\big(z\big)$	$z > z_{1-\alpha}$
$H_1 : \mu < \mu_0$	$P = \Phi\big(z\big)$	$z < z_{\alpha}$

Example

Given $H_0 : \mu = 5$, $H_1 : \mu < 5$, σ known. Calcualte the p-value for the test statistic $z_0 = 0.4$.

$$p\text{-value} = P(T < \bar{x} \mid H_0 \text{ true}) = P\left(\frac{T-5}{\sigma/\sqrt{n}} < \frac{\bar{x}-5}{\sigma/\sqrt{n}}\right)$$

$$= \Phi\left(\frac{\bar{x}-5}{\sigma/\sqrt{n}}\right) = \Phi(0.4) = 0.6554$$

$\frac{\bar{x}-5}{\sigma/\sqrt{n}} = z_0$

$\Rightarrow p > 0.1$

$\Rightarrow$ do not reject H_0

We distinguish between the two cases:

- Unrealistic (but simpler): The population variance, σ^2, is known.
- More realistic: The variance is not known and estimated by the sample variance, s^2.

Private school - revisited

Private school claims that its students have a higher IQ. Recall $IQ_{all} \sim N(100, 16)$ and $IQ_A \sim N(105, 25)$.

- Should we try to place our child in this school?
- Is the observed result *significant* **(can be trusted?)**, or due to a *chance*?

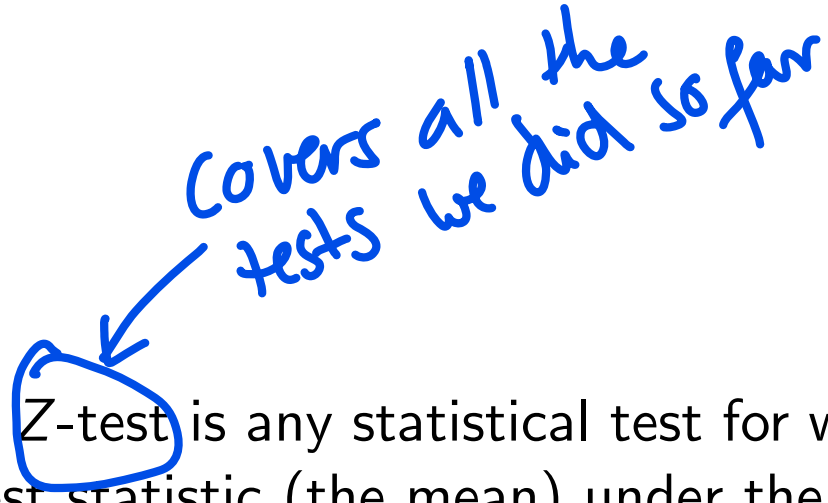
Variance is known to be 16.

14 subjects were randomly assigned to control or treatment group of the experimental medical treatment. The survival times (in days) were

	Data	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12, 18	43.285

The variance is not known and estimated by the sample variance, s^2.

Known variance — the Z-test



- A Z-test is any statistical test for which the distribution of the test statistic (the mean) under the null hypothesis can be approximated by a normal distribution (with known variance).
- Thanks to the central limit theorem, many test statistics are approximately **normally distributed** for large enough samples.

Problem and Solution

Z **test uses CLT** (required: σ is known and n is "large")

- If σ^2 is unknown $\rightarrow$ estimate using sample variance s^2
  Recall: Sample variance $= s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  $\longrightarrow$ the test-statistic reads as

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}},$$

- If n small or σ unknown $\rightarrow$ T no longer follows a Normal distribution!

- **Good news:** T follows The Student-t Distribution.

The t-distribution

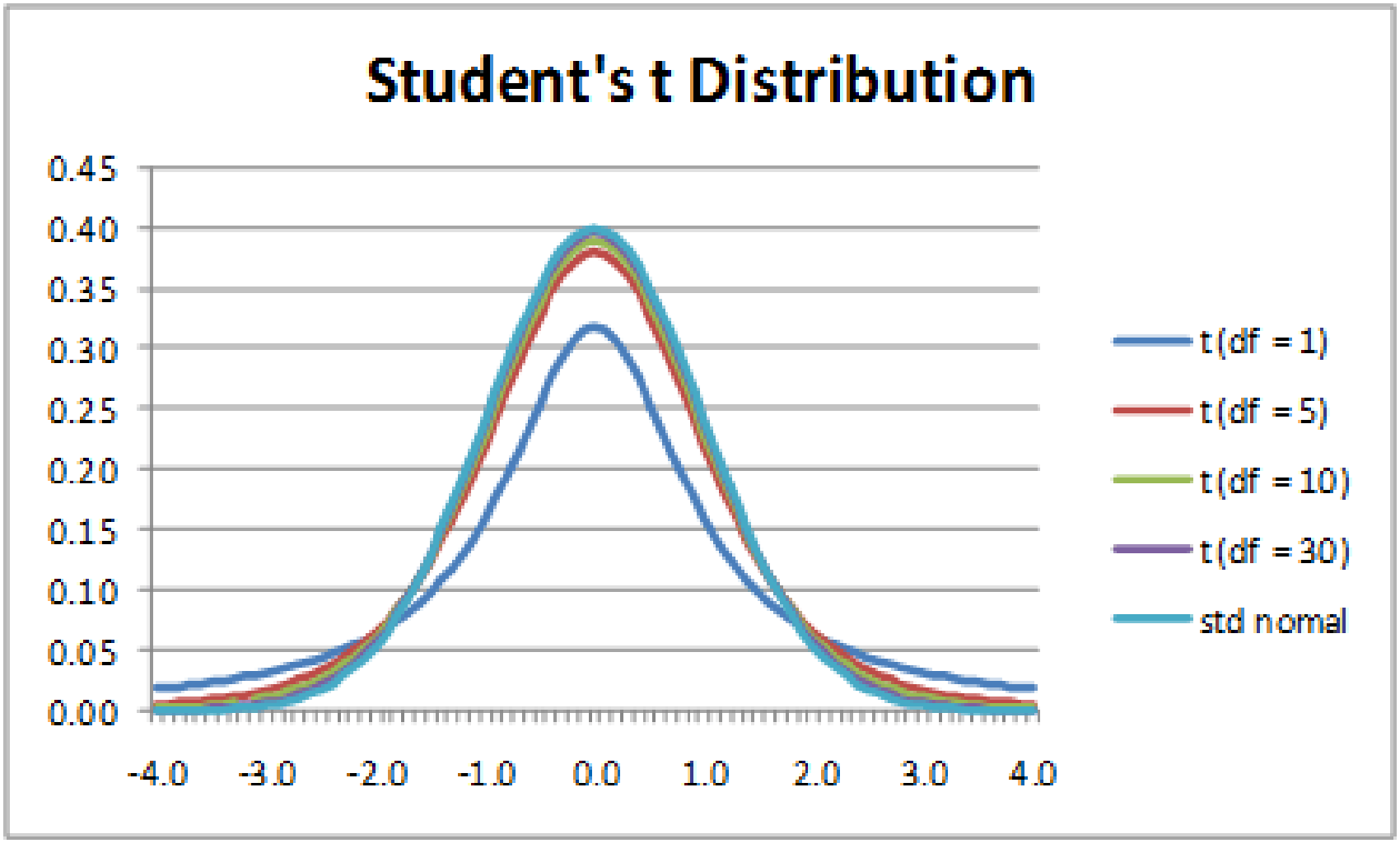
- The pdf of Student-t Distribution with parameter k (degrees of freedom) is:

$$f(x) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\Gamma(k/2)} \frac{1}{[(x^2/k)+1]^{(k+1)/2}}, \quad -\infty < x < \infty,$$

  where $\Gamma(\cdot)$ is the Gamma-function:

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} \mathrm{d}x.$$

- It is a symmetric distribution about 0 and as $k \to \infty$, it approaches a standard Normal distribution.


Student's t Distribution
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
t(df = 1)
t(df = 5)
t(df = 10)
t(df = 30)
std nomal

t-test

The t statistic (introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Ireland) with $n-1$ degrees of freedom is

$$T_{n-1} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}},$$

where s is the estimated standard deviation:

Sample variance $= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$

- For large n, t-test is indistinguishable from the z-test.

Confidence and prediction intervals

recall $\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

- If $\overline{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a $100(1-\alpha)$ confidence interval on μ is given by:

$$\overline{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}},$$

where $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ quantile of the t distribution with n - 1 degrees of freedom.

where $t_{1-\frac{\alpha}{2},n-1}$ is the value such that

$\hat{\Phi}$ = cdf of t-distribution $\quad \hat{\Phi}(t_{1-\frac{\alpha}{2},n-1}) = 1 - \frac{\alpha}{2}$

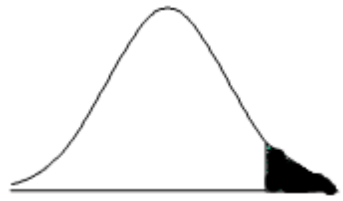
- A related concept is a $100(1-\alpha)$ prediction interval (PI) on a single future observation from a normal distribution is given by

$$\overline{x} - t_{1-\alpha/2,n-1} s\sqrt{1+\frac{1}{n}} \leq X_{n+1} \leq \overline{x} + t_{1-\alpha/2,n-1} s\sqrt{1+\frac{1}{n}}$$

  This is the range where we expect the $n+1$ observation to be, after observing n observations and computing $\overline{x}$ and s.

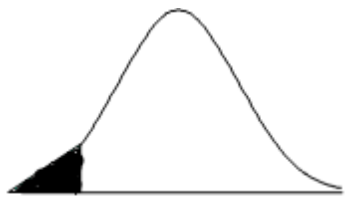
Types of tests

Right one-sided test: where H_0 is rejected for the p-value defined by $P_{H_0}(T \geq t)$.



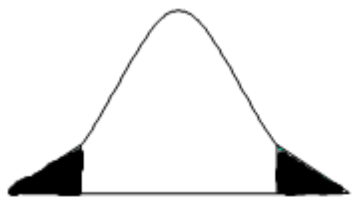
Types of tests

Left one-sided test: where H_0 is rejected for the p-value defined by $P_{H_0}(T \leq t)$.



Types of tests

Two-sided test: where H_0 is rejected for the p-value defined by $P_{H_o}(T \geq t) + P_{H_o}(T \leq -t) = 2P_{H_o}(T \geq t)$.



Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with both μ and σ^2 unknown.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}.$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\left[1 - F_{n-1}(\lvert t \rvert)\right]$	$t > t_{1-\alpha/2,n-1}$ or $t < t_{\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	$P = 1 - F_{n-1}(t)$	$t > t_{1-\alpha,n-1}$
$H_1 : \mu < \mu_0$	$P = F_{n-1}(t)$	$t < t_{\alpha,n-1}$

Note: Since pdf of t-distr. is symmetric,

$t_{\frac{\alpha}{2},n-1} = -t_{1-\frac{\alpha}{2},n-1}$

- In the p-value calculation, $F_{n-1}(\cdot)$ denotes the CDF of the t-distribution with $n-1$ degrees of freedom.
- As opposed to $\Phi(\cdot)$, the CDF of t is not tabulated in standard tables. So to calculate p-values, we use software.

Example

Five samples of a material were tested in a structure, and the average interior temperature in Celcius reported:

23.01 22.22 22.04 22.62 22.59

$\bar{x} = 22.496$

$s \approx 0.3783$

Test the claim that the interior temperature is 22.5.

$H_0: \mu = 22.5$ $H_1: \mu \neq 22.5$ Pick $\alpha = 0.05$

Step1: Get $t^* = t_{1-\alpha/2, n-1} = \Phi(t^*) = 1 - \frac{\alpha}{2} = 0.975$

$\Rightarrow t^* = 2.776 \Rightarrow$ reject if $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > 2.776$

or $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -2.776$

Step2: $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = -0.0236 \Rightarrow$ do not reject.

t-Distribution Quantiles

ν	$Q(.9)$	$Q(.95)$	$Q(.975)$	$Q(.99)$	$Q(.995)$	$Q(.999)$	$Q(.9995)$
1	3.078	6.314	12.706	31.821	63.657	318.317	636.607
2	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.849
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674

Example

A golf club producer claims that the mean coefficient of restitution exceeds 0.82. Suppose you take a sample with the following restitution coefficient measurements:

0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

$n = 15$
$\bar{x} = 0.83725$
$s = 0.02456$

Pick $\alpha = 0.05$ $\quad H_0: \mu = 0.82$ $\quad H_1: \mu > 0.82$

Step 1: Get $t_{1-\alpha, n-1} = t^*$

$\hat{\Phi}(t^*) = 1-\alpha = 0.95 \Rightarrow$ $t^* = 1.761$

$\Rightarrow$ reject H_0 if $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} > t^* = 1.761$

Step2: $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{0.83725-0.82}{0.02456/\sqrt{15}} = 2.72$

Step3: Since $2.72 > t^{\gamma}$ $\Rightarrow$ Reject H_0

Chapter 11

Simple Linear Regression

Simple Linear Regression

- Aim: Study or analysis of the relationship between two or more variables
  e.g. Pressure of a gas in a container versus its temperature

  > We examine a dependent variable and one or more independent variables (= predictors).
  > $\Longrightarrow$ Regression Analysis

- Key importance (conditional expectation):

$$\mathbb{E}[Y \mid x] = \mu_{Y|x}$$

Suppose for now, the variable Y depends linearly on only one predictor, i.e.:

$$\mathbb{E}[Y \mid x] = \mu_{Y|x} = \beta_0 + \beta_1 x$$

$\Longrightarrow$

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where:

- x is a (non-random) predictor, and
- ϵ is a R.V.(=noise) with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$.

Assumptions:

- Normality of residuals,
- Constant variance, and,
- Independence of observations

Method:

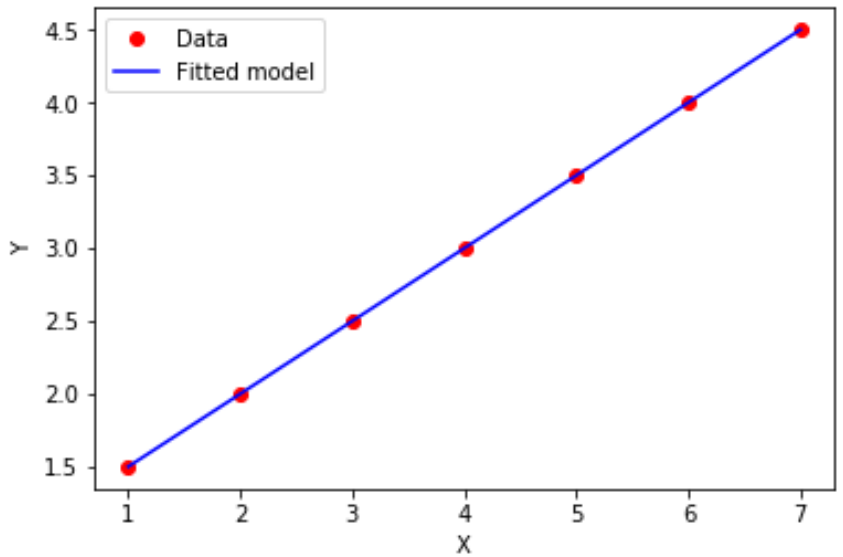
- Collect data:
$$(x_1, y_1), \ldots, (x_n, y_n).$$
- Assume linear relation:
$$y \approx \beta_0 + \beta_1 x \qquad \leftrightarrow \qquad y = \beta_0 + \beta_1 x + \epsilon$$
- Since we do not have all possible tuples, we can only estimate β_0 and β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, i.e.,
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, \ldots, n.$$
- $e_i =$ **residual**.

Use $\hat{\beta}_0, \hat{\beta}_1$ for predictions.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Note that we can also compute predicted observations for our data $(x_i, y_i)_{\{1 \leq i \leq n\}}$.

Ideally, we would like to find $\hat{\beta}_0$ and $\hat{\beta}_1$, such that $y_i = \hat{y}_i$, that is, $e_i = 0$.

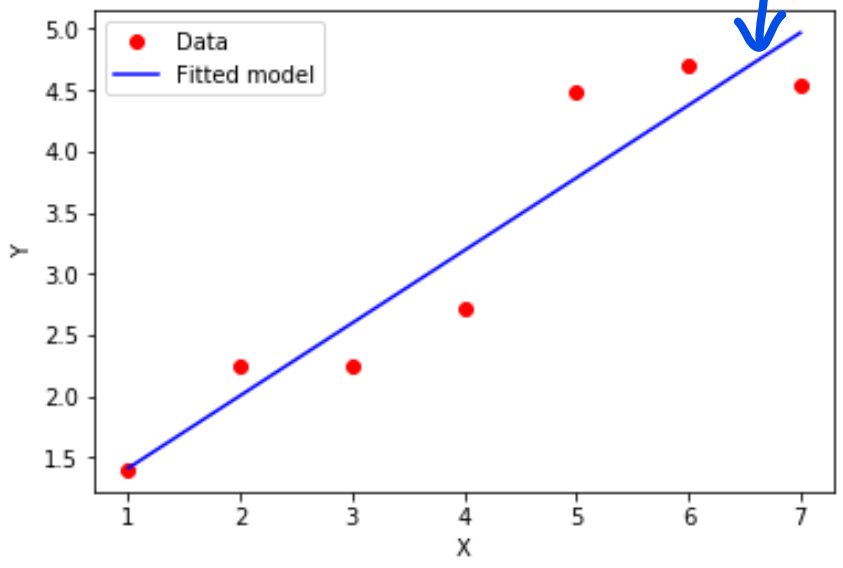


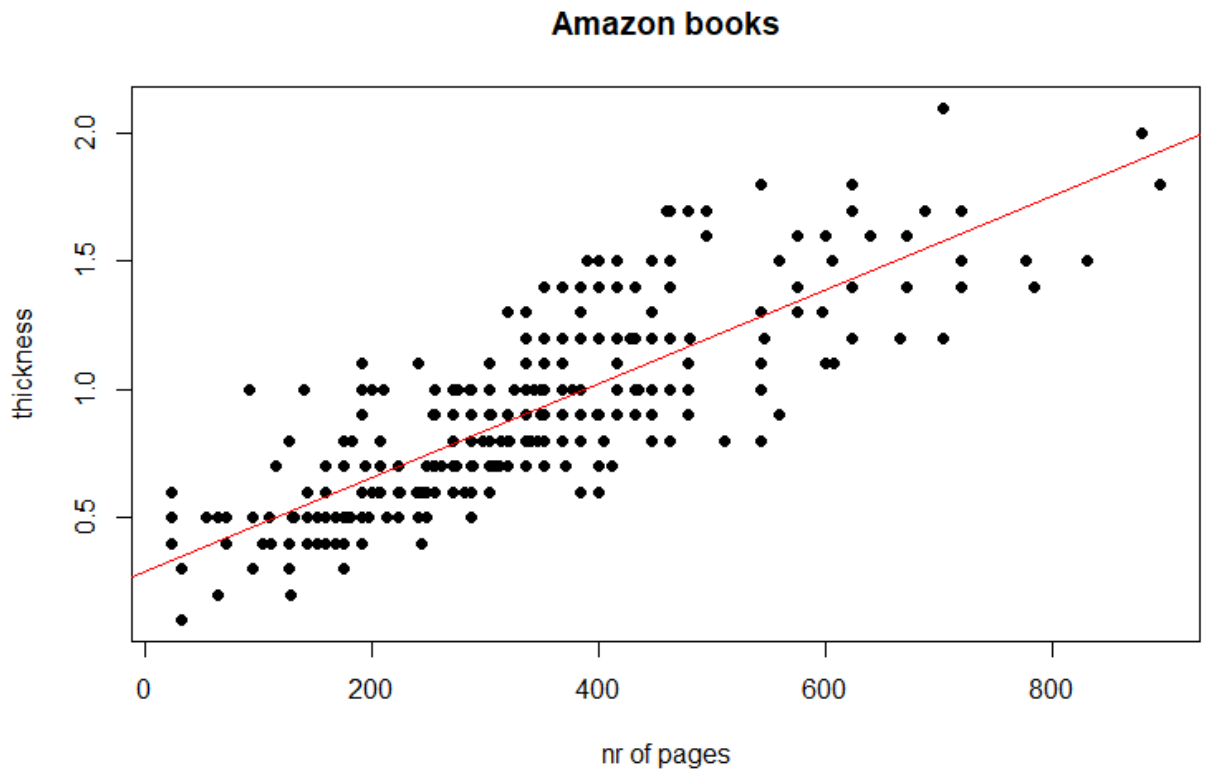
$\hat{\beta}_0 = \beta_0$

$\hat{\beta}_1 = \beta_1$

Much more likely:

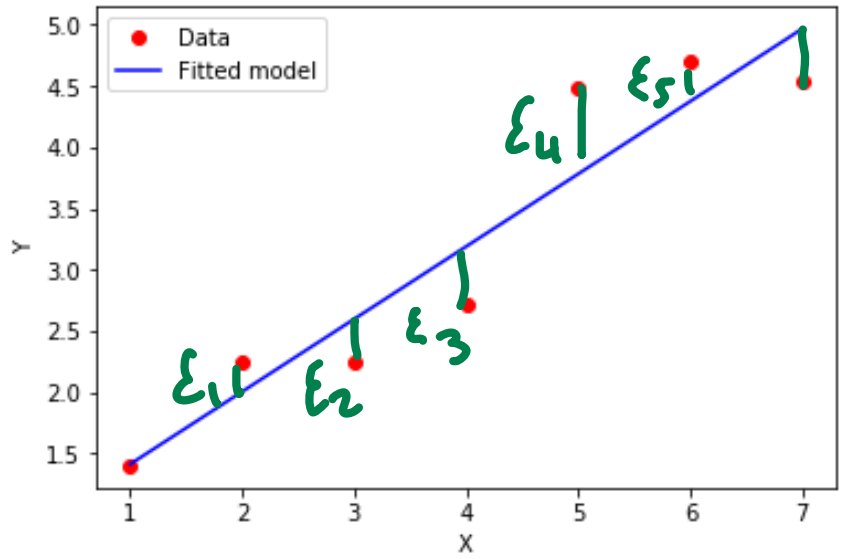
line: $\hat{\beta}_0 + \hat{\beta}_1 x$






```
D <- read.delim("amazon-books.txt")
plot(D$NumPages, D$Thick)
abline(lm(D$Thick~D$NumPages), col='red')
```

Total mean squared error



Total Mean Squared Error:

$$L = SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

The least squares estimators

Find the "best possible" $\hat{\beta}_0, \hat{\beta}_1$

- To find the best estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, we would like to minimize

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- Specifically, solve

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- The solution, called the *least squares estimators* must satisfy:

$$\frac{\partial}{\partial \beta_0} L = 0 = \frac{\partial}{\partial \beta_1} L$$

$$0 \overset{(1)}{=} \frac{\partial}{\partial \beta_0} L = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i)$$

$$0 \overset{(2)}{=} \frac{\partial}{\partial \beta_1} L = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i)$$

to be continued next lecture

The least squares solution

Using the sample means, x and y

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

the estimators are:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Additional quantities of interest

$$S_{XX} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

$$S_{XY} = \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

That is,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{XY}}{S_{XX}}.$$

Additional quantities of interest

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \_\_\_\_\_$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = \_\_\_\_\_$$

$$SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \_\_\_\_\_$$

It holds that

$$SS_T = SS_R + SS_E,$$

The Analysis of Variance

- We did not consider the final unknown parameter in our regression model:

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

namely, the $\mathrm{Var}(\epsilon) = \sigma^2$.

- We use the residuals $e_i = \hat{y}_i - y_i$, to obtain an estimate of σ^2.
- Specifically,

$$SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2,$$

and it can be shown that

$$\mathbb{E}[SS_E] = (n-2)\sigma^2,$$

so:

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}.$$

How good is my regression model?

A widely used measure for a regression model is the following ratio of sum of squares, which is often used to judge the adequacy of a regression model:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T},$$

Properties of least square estimator

$$\mathbb{E}[\hat{\beta}_0] = \beta_0, \quad \mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right],$$

$$\mathbb{E}[\hat{\beta}_1] = \beta_1, \quad \mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{XX}},$$

Therefore, the estimated standard error of the slope and the estimated standard error of the intercept are:

$$se\left(\hat{\beta}_0\right) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]},$$

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\sigma^2}{S_{XX}}}.$$

Example

A study considers the microstructure of the ultrafine powder of partially stabilized zirconia as a function of temperature. The data is as follows:

x (Temperature)	1100	1200	1300	1100	1500	1200	1300
y (Porosity)	30.8	19.2	6.0	13.5	11.4	7.7	3.6

a) Find the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

b) Estimate the porosity for a temperature of 1400 degrees Celcius.

c) Find SS_E (= error sum of squares).

d) Find the least square estimates for y with respect to the predictor $x_i^* = x_i + \bar{x}$.

Hypothesis tests in linear regression

- Suppose we would like to test:

$$H_0 : \beta_1 = \beta_{1,0}, \quad H_1 : \beta_1 \neq \beta_{1,0}.$$

- The Test Statistic for the Slope is

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\sigma^2}{S_{xx}}}}.$$

- Under H_0, the test statistic T follows a t - distribution with $n - 2$ degree of freedom.

- Suppose we would like to test:

$$H_0 : \beta_0 = \beta_{0,0}, \quad H_1 : \beta_0 \neq \beta_{1,0}.$$

- The Test Statistic for the intercept is

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}.$$

- Under H_0, the test statistic T follows a t - distribution with $n - 2$ degree of freedom.

An important special case of the hypotheses is:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0.$$

If we fail to reject $H_0 : \beta_1 = 0$, this indicates that there is no linear relationship between x and y.

Example

Suppose we have 20 samples regarding oxygen purity (y) with respect to hydrocarbon levels (x) such that

$$\sum_{i=1}^{20} x_i = 23.92, \qquad \sum_{i=1}^{20} y_i = 1,843.21, \qquad \bar{x} = 1.1960, \qquad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170,044.5321, \qquad \sum_{i=1}^{20} x_i^2 = 29.2892, \qquad \sum_{i=1}^{20} x_i y_i = 2,214.6566$$

Test: $H_0 : \quad \beta_1 = 0 \qquad H_1 : \quad \beta_1 \neq 0$
for $\alpha = 0.01$.

The F distribution

- An alternative is to use the F statistic as is common in ANOVA (Analysis of Variance) (not covered fully in the course).
- Under H_0, the test statistic

$$F = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E},$$

  follows an F - distribution with 1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator.
- Here,

$$MS_R = SS_R/1, \quad MS_E = SS_E/(n-2).$$

Analysis of Variance Table for Testing Significance of Regression

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_E	
Total	SS_T	$n-1$		