STATISTICS II with MATHEMATICA [Lecture Notes,Topics and Lab Assignments]

Research Proposal · March 2017

DOI: 10.13140/RG.2.2.31857.28002

CITATIONS
0

READS
20,047

1 author:

Dimitris Apostolos Sardelis

53 PUBLICATIONS **122** CITATIONS

SEE PROFILE

THE AMERICAN COLLEGE OF GREECE

STATISTICS II
with MATHEMATICA

Lecture Notes
Topics
Lab Assignments

by Dr. Dimitris Sardelis

Athens 2006

THE HYPOTHESIS TESTING PROCEDURE

A. What do I test?

According to the problem context, formulate two appropriate (competing) hypotheses one of which can be true and the other false:

Null Hypothesis (H_0) : A general proposition of no difference or equality (translated to relations $=$, $\leq$, $\geq$) that is assumed to be true throughout the statistical analysis.

Alternative Hypothesis (H_1) : A general proposition of difference or of inequality (translated to relations $\neq$, $<$, $>$).

B. Which (testing) tool do I use?

Under the assumption that H_0 is true, the variable of interest is called **test statistic**. All test statistic formulas mingle population parameters with sample characteristics and their possible values form **sampling probability distributions**.

C. How do I reach a decision?

The **level of significance α** ultimately defines which extreme test statistic values are to be considered "very unlikely", and the **degree of confidence $1-\alpha$** which test statistic values are to be considered likely. To specify the demarkation line(s) between probable and improbable test statistic values□henceforth called as **critical test statistic values**□α must be set as the area of the region at one or both extremes/tails of the sampling distribution graph, according to the relation stated in H_1:

- ★ If the relation in H_1 is $<$, α is set at the extreme left region of the sampling distribution, the test is called **left tailed**, and for any test statistic value **less** than the (left) critical test statistic value, H_0 must be **rejected.**
- ★ If the relation in H_1 is $\neq$, α is set at the extreme left and right regions of the sampling distribution, the test is called **two tailed**, and for any test statistic value either **less** than the (left) critical test statistic value or **more** than the (right) critical test statistic value, H_0 must be **rejected.**
- ★ If the relation in H_1 is $>$, α is set at the extreme right region of the sampling distribution, the test is called **right tailed**, and for any test statistic value **greater** than the (right) critical test statistic value, H_0 must be **rejected.**
- ◆ **P- value** (p) is the probabilility of obtaining a value of the test statistic at least as extreme as that actually obtained given that H_0 is true. For one sided/tailed tests, H_0 is to be rejected if $p < \alpha$. For two tailed tests and symmetrical sampling distributions, H_0 is to be rejected if $2\,p < \alpha$.

TESTS FOR ONE MEAN

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: NORMAL

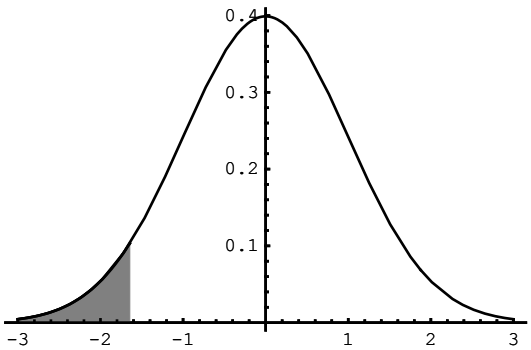
Test Statistic

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}}$$

Probability Distribution Function Form

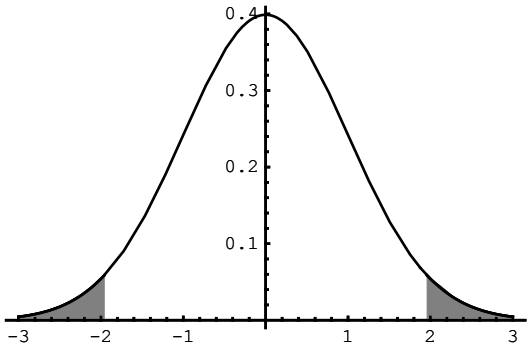
$$\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\,\pi}}$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : \mu < \mu_0$)



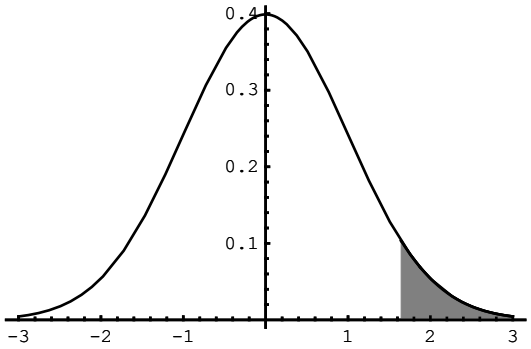
The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : \mu \neq \mu_0$)



Both shaded regions have an area α and define the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : \mu > \mu_0$)



The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

SAMPLING DISTRIBUTION: STUDENT T-DISTRIBUTION

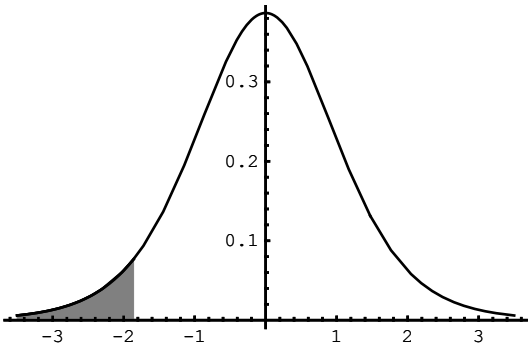
Test Statistic

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}; \; d.f = n - 1$$

Probability Distribution Function Form

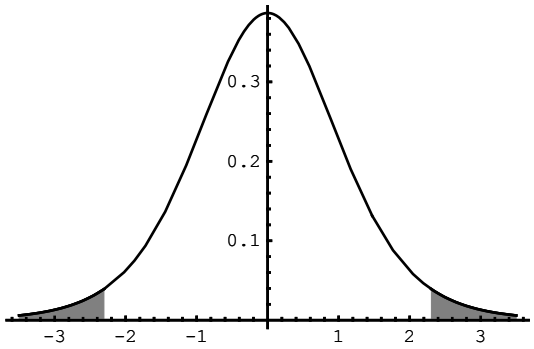
$$\frac{\left(\frac{n-1}{t^2+n-1}\right)^{n/2}}{\sqrt{n-1}\;\text{Beta}\left[\frac{n-1}{2}, \frac{1}{2}\right]}$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : \mu < \mu_0$)



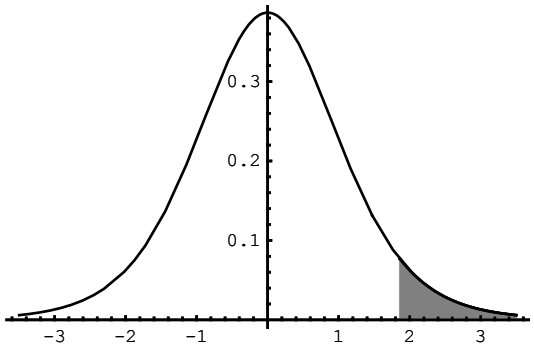
The shaded region has an area α and defines the so called **rejection region,** meaning that if t falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : \mu \neq \mu_0$)



Both shaded regions have an area α and define the so called **rejection region,** meaning that if t falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : \mu > \mu_0$)



The shaded region has an area α and defines the so called **rejection region,** meaning that if t falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if (Test Stat : z)	Reject H_0 if (Test Stat : t, $d.f = n-1$)
left – tailed	$\mu \geq \mu_0$	$\mu < \mu_0$	$z < z_\alpha = -z_{1-\alpha}$	$t < t_\alpha = -t_{1-\alpha}$
two – tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	either $z < z_{\alpha/2} = -z_{1-\alpha/2}$ or $z > z_{1-\alpha/2}$	either $t < t_{\alpha/2} = -t_{1-\alpha/2}$ or $t > t_{1-\alpha/2}$
right – tailed	$\mu \leq \mu_0$	$\mu > \mu_0$	$z > z_{1-\alpha}$	$t > t_{1-\alpha}$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if ($variable : Z$, Test Stat : z)	Reject H_0 if ($variable : T$, Test Stat : t, $d.f = n-1$)
left – tailed	$\mu \geq \mu_0$	$\mu < \mu_0$	$p = prob[Z < z] < \alpha$	$p = prob[T < t] < \alpha$
two – tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	$p = 2\ prob[Z > \lvert z \rvert] < \alpha$	$p = 2\ prob[T > \lvert t \rvert] < \alpha$
right – tailed	$\mu \leq \mu_0$	$\mu > \mu_0$	$p = prob[Z > z] < \alpha$	$p = prob[T > t] < \alpha$

ILLUSTRATIONS

Exercise	H_0	H_1	Facts	Actual Decision	Conclusion for H_0
7.3 .1	$\mu = 100$	$\mu \neq 100$	$\alpha = 0.01$; pop.normal$(\sigma = 50)$ $n = 25$, $\bar{x} = 70$	$z = -3 < z_{0.005} = -2.58$ True $p = 2\,(0.00135) = 0.0027 < 0.01$ True	reject

7.3 .2	$\mu \le 4.5$	$\mu > 4.5$	$\alpha = 0.01$; pop.normal $(\sigma = 0.02)$ $n = 16$, $\bar{x} = 4.512$	$z = 2.4 > z_{0.99} = 2.33$ True $p = 0.0082 < 0.01$ True	reject
7.3 .3	$\mu = 6$	$\mu \neq 6$	$\alpha = 0.05$; $\sigma = 1$ $n = 9$, $\bar{x} = 7$	$z = 3 > z_{0.975} = 1.96$ True $p = 2\,(0.00135) = 0.0027 < 0.05$ True	reject
7.3 .4	$\mu \le 80$	$\mu > 80$	$\alpha = 0.05$; pop.normal $(\sigma = 10)$ data $(n = 20, \bar{x} = 81.65)$	$z = 0.74 > z_{0.95} = 1.645$ False $p = 0.2296 < 0.05$ False	do not reject
7.4 .1	$\mu \le 190$	$\mu > 190$	$\alpha = 0.05$; pop.normal $n = 16$, $\bar{x} = 195$, $s = 8$	$t = 2.5 > t_{0.95}[\text{d.f} = 15] = 1.753$ True $p = 0.0123 < 0.05$ True	reject
7.4 .2	$\mu_d \le 0$	$\mu_d > 0$	$\alpha = 0.05$; $d \equiv x\,(\text{with}) - x\,(\text{without})$ data $(n = 16, \bar{d} = 0.893750, s_d = 0.304344)$	$t = 11.747 > t_{0.95}[\text{d.f} = 15] = 1.753$ True $p = 2.892 * 10^{-9} < 0.05$ True	reject
7.4 .3	$\mu_d \ge 0$	$\mu_d < 0$	$\alpha = 0.05$; $d \equiv x\,(\text{after}) - x\,(\text{before})$ data $(n = 12, \bar{d} = -2.5, s_d = 2.7798)$	$t = -3.115 < t_{0.95}[\text{d.f} = 11] = 1.7959$ True $p = 0.0049 < 0.05$ True	reject
7.4 .4	$\mu \ge 4$	$\mu < 4$	$\alpha = 0.05$; pop.normal $n = 25$, $\bar{x} = 3.8$, $s = 0.5$	$t = -2 < t_{0.05}[\text{d.f} = 24] = -1.7109$ True $p = 0.02847 < 0.05$ True	reject
7.4 .5	$\mu \le 160$	$\mu > 160$	$\alpha = 0.05$; data $(n = 10, \bar{x} = 178.9, s = 26.3837)$	$t = 2.2653 > t_{0.05}[\text{d.f} = 9] = 1.8331$ True $p = 0.02487 < 0.05$ True	reject
7.4 .6	$\mu \le 25$	$\mu > 25$	$\alpha = 0.01$; data $(n = 15, \bar{x} = 33.8, s = 12.5653)$	$t = 2.71242 > t_{0.99}[\text{d.f} = 14] = 2.624$ True $p = 0.0084 < 0.01$ True	reject
7.5 .1	$\mu = 175$	$\mu \neq 175$	$\alpha = 0.05$; $n = 100$, $\bar{x} = 170$, $s = 25$	$t = -2 < t_{0.0250}[\text{d.f} = 99] = -1.984$ True $p = 2\,(0.02412) = 0.04824 < 0.05$ True	reject
7.5 .2	$\mu \le 45$	$\mu > 45$	$\alpha = 0.05$; $n = 36$, $\bar{x} = 48$, $s = 12$	$t = 1.5 > t_{0.95}[\text{d.f} = 35] = 1.6896$ False $p = 2\,(0.0713) = 0.1426 < 0.05$ False	do not reject
7.5 .3	$\mu \le 90$	$\mu > 90$	$\alpha = 0.05$; $n = 100$, $\bar{x} = 96$, $s = 30$	$t = 2 > t_{0.95}[\text{d.f} = 99] = 1.6602$ True $p = 0.0241 < 0.05$ True	reject
7.5 .4	$\mu \le 60$	$\mu > 60$	$\alpha = 0.05$; pop.not normal $(\sigma^2 = 280)$ data $(n = 40, \bar{x} = 71.1, s = 16.7332)$	$z = 4.195 > z_{0.95} = 1.645$ True $p = 0.0000136 < 0.05$ True	reject
7.5 .5	$\mu \ge 3$	$\mu < 3$	$\alpha = 0.01$; data $(n = 30, \bar{x} = 2.391676, s = 1.41982)$	$t = -2.347 < t_{0.01}[\text{d.f} = 29] = -2.462$ False $p = 0.012989 < 0.01$ False	do not reject

TESTS FOR TWO MEANS

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: NORMAL

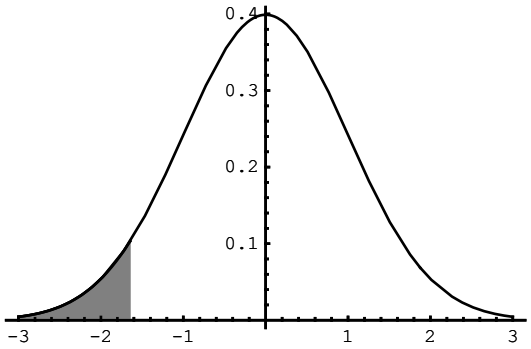
Test Statistic

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Probability Distribution Function Form

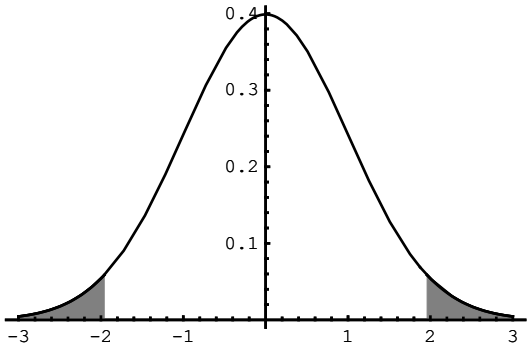
$$\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\,\pi}}$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : \mu_1 - \mu_2 < 0$)



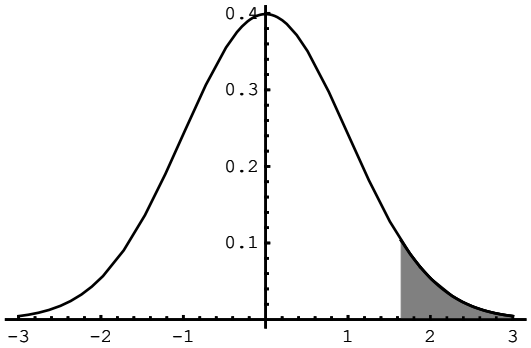
The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : \mu_1 - \mu_2 \neq 0$)



Both shaded regions have an area α and define the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : \mu_1 - \mu_2 > 0$)



The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

SAMPLING DISTRIBUTION: STUDENT T-DISTRIBUTION

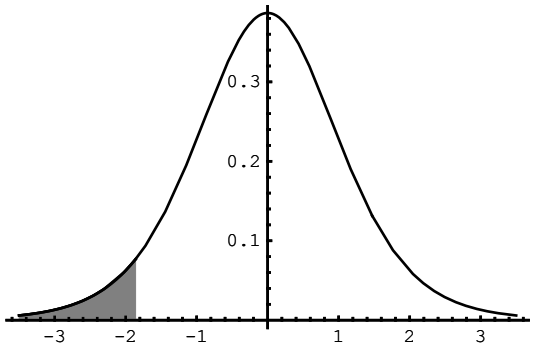
Test Statistic

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}; \; d.f = n_1 + n_2 - 2; \; s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Probability Distribution Function Form

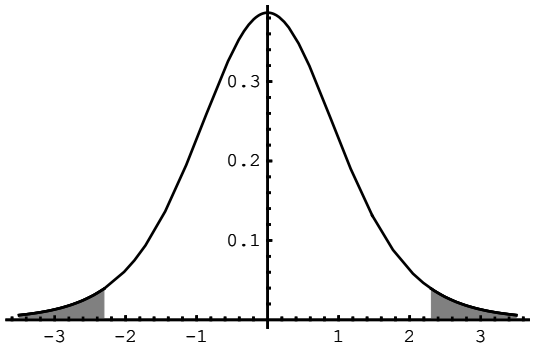
$$\frac{\left(\frac{n1+n2-2}{t^2+n1+n2-2}\right)^{\frac{1}{2}(n1+n2-1)}}{\sqrt{n1 + n2 - 2}\ \text{Beta}\left[\frac{1}{2}(n1 + n2 - 2),\ \frac{1}{2}\right]}$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : \mu_1 - \mu_2 < 0$)



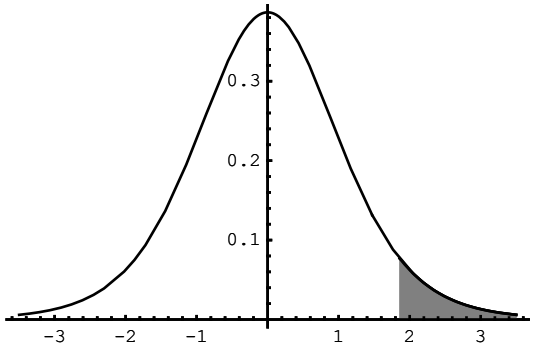
The shaded region has an area α and defines the so called **rejection region,** meaning that if t falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : \mu_1 - \mu_2 \neq 0$)



Both shaded regions have an area α and define the so called **rejection region,** meaning that if t falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : \mu_1 - \mu_2 > 0$)



The shaded region has an area α and defines the so called **rejection region,** meaning that if t falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if (Test Stat : z)	Reject H_0 if (Test Stat : t, $d.f = n1 - n2 - 2$)
left – tailed	$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 < 0$	$z < z_\alpha = -z_{1-\alpha}$	$t < t_\alpha = -t_{1-\alpha}$
two – tailed	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	either $z < z_{\alpha/2} = -z_{1-\alpha/2}$ or $z > z_{1-\alpha/2}$	either $t < t_{\alpha/2} = -t_{1-\alpha/2}$ or $t > t_{1-\alpha/2}$
right – tailed	$\mu_1 - \mu_2 \leq 0$	$\mu_1 - \mu_2 > 0$	$z > z_{1-\alpha}$	$t > t_{1-\alpha}$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if $(\text{variable}: Z, \text{Test Stat}: z)$	Reject H_0 if $(\text{variable}: T, \text{Test Stat}: t, d.f = n1 - n2 - 2)$
left – tailed	$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 < 0$	$p = prob[Z < z] < \alpha$	$p = prob[T < t] < \alpha$
two – tailed	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$p = 2\, prob[Z > \lvert z \rvert] < \alpha$	$p = 2\, prob[T > \lvert t \rvert] < \alpha$
right – tailed	$\mu_1 - \mu_2 \leq 0$	$\mu_1 - \mu_2 > 0$	$p = prob[Z > z] < \alpha$	$p = prob[T > t] < \alpha$

ILLUSTRATIONS

Exercise	H_0	H_1	Facts	Actual Decision	Conclusion for H_0
7.6 .1	$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 < 0$	$\alpha = 0.05$; pops. normal $(\sigma_1^2 = \sigma_2^2)$ $n_1 = 10$, $\bar{x}_1 = 94$, $s_1^2 = 14$ $n_2 = 12$, $\bar{x}_2 = 98$, $s_2^2 = 9$	$t = -2.785 < t_{0.05}[d.f = 20] = -1.7247$ $p = 0.0057 < 0.05$ both True	reject
7.6 .2	$\mu_1 - \mu_2 \leq 0$	$\mu_1 - \mu_2 > 0$	$\alpha = 0.05$; pops. normal $(\sigma_1^2 = \sigma_2^2)$ $n_1 = 10$, $\bar{x}_1 = 26.1$, $s_1^2 = 144$ $n_2 = 8$, $\bar{x}_2 = 17.6$, $s_2^2 = 110$	$t = 1.577 > t_{0.95}[d.f = 16] = 1.7459$ $p = 0.0672 < 0.05$ both False	do not reject
7.7 .1	$\mu_A - \mu_B \leq 0$	$\mu_A - \mu_B > 0$	$\alpha = 0.05$; $n_A = 50$, $\bar{x}_A = 28.25$, $s_A^2 = 25$ $n_B = 50$, $\bar{x}_B = 22.50$, $s_B^2 = 16$	t (or z) $= 6.35 > 1.6602$ (or 1.645) $p = 3.38 \times 10^{-9} < 0.05$ both True	reject
7.7 .3	$\mu_A - \mu_B \geq 0$	$\mu_A - \mu_B < 0$	$\alpha = 0.05$; $n_A = 100$, $\bar{x}_A = 33$, $s_A^2 = 900$ $n_B = 150$, $\bar{x}_B = 49$, $s_B^2 = 1050$	$t = -3.939 < t_{0.05}[d.f = 148] = -1.656$ $z = -4 < -1.645$ $p = 0.000053 < 0.05$ all True	reject
7.7 .5	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$\alpha = 0.05$; data$\left[n_1 = 30, \bar{x}_1 = 54.2767, s_1 = 3.1297, n_2 = 35, \bar{x}_2 = 49.0457, s_2 = 1.2472\right]$	$t = 9.091 > t_{0.975}[d.f = 63] = 2$ $z = 8.589 > 1.96$ $p = 4.51 \times 10^{-13} < 0.05$ all True	reject

TESTS FOR ONE PROPORTION

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: NORMAL

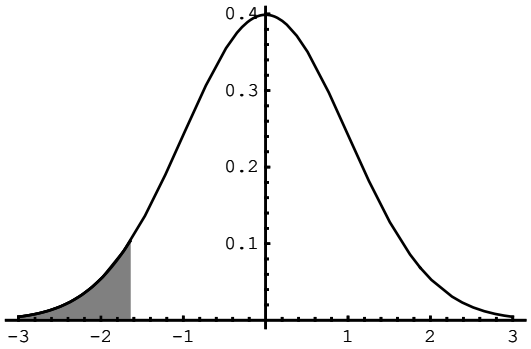
Test Statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}$$

Probability Distribution Function Form

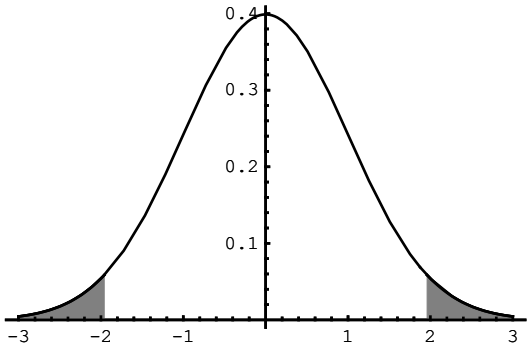
$$\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\,\pi}}$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : p < p_0$)



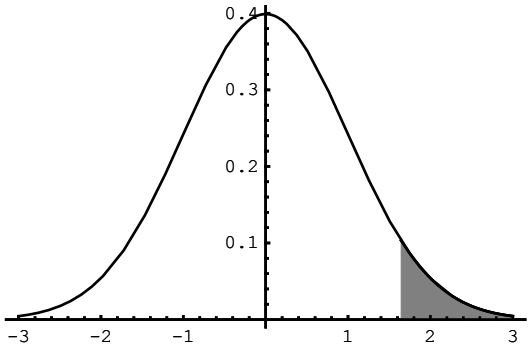
The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : p \neq p_0$)



Both shaded regions have an area α and define the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : p > p_0$)



The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if (Test Stat : z)
left – tailed	$p \geq p_0$	$p < p_0$	$z < z_\alpha = -z_{1-\alpha}$
two – tailed	$p = p_0$	$p \neq p_0$	either $z < z_{\alpha/2} = -z_{1-\alpha/2}$ or $z > z_{1-\alpha/2}$
right – tailed	$p \leq p_0$	$p > p_0$	$z > z_{1-\alpha}$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if ($\text{variable}: Z$, $\text{Test Stat}: z$)
left – tailed	$p \geq p_0$	$p < p_0$	$p = prob[Z < z] < \alpha$
two – tailed	$p = p_0$	$p \neq p_0$	$p = 2\, prob[Z > \lvert z \rvert] < \alpha$
right – tailed	$p \leq p_0$	$p > p_0$	$p = prob[Z > z] < \alpha$

ILLUSTRATIONS

Exercise	H_0	H_1	Facts	Actual Decision	Conclusion for H_0
7.8 .1	$p \geq 0.25$	$p < 0.25$	$\alpha = 0.05$; $p_0 = 0.25$, $p = 42/200$;	$z = -1.306 < -1.645$ False $p = 0.0957 < 0.05$ False	do not reject
7.8 .2	$p \leq 0.6$	$p > 0.6$	$\alpha = 0.05$; $N = 10000$; $p_0 = 0.6$; $p = 165/250$	$z = 1.93649 > 1.645$ True $p = 0.0264 < 0.05$ True	reject

TESTS FOR TWO PROPORTIONS

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: NORMAL

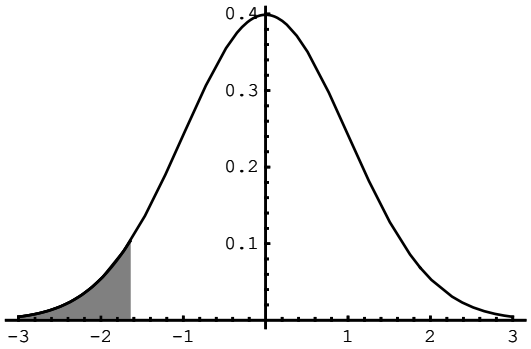
Test Statistic

$$z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\,(1-\bar{p})}{n_1} + \frac{\bar{p}\,(1-\bar{p})}{n_2}}}\,;\ \bar{p} = \frac{n_1\,\hat{p}_1 + n_2\,\hat{p}_2}{n_1 + n_2}$$

Probability Distribution Function Form

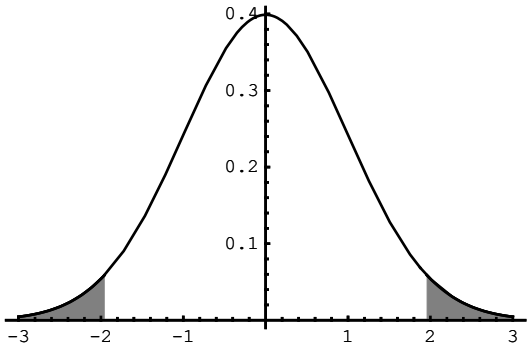
$$\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\,\pi}}$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : p_1 - p_2 < 0$)



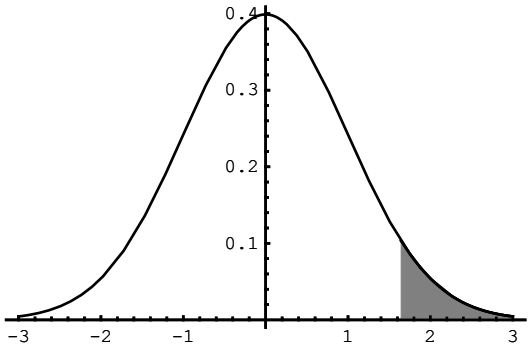
The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : p_1 - p_2 \neq 0$)



Both shaded regions have an area α and define the so called **rejection region,** meaning that if z falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : p_1 - p_2 > 0$)



The shaded region has an area α and defines the so called **rejection region,** meaning that if z falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if (Test Stat : z)
left – tailed	$p_1 - p_2 \geq 0$	$p_1 - p_2 < 0$	$z < z_\alpha = -z_{1-\alpha}$
two – tailed	$p_1 - p_2 = 0$	$p_1 - p_2 \neq 0$	either $z < z_{\alpha/2} = -z_{1-\alpha/2}$ or $z > z_{1-\alpha/2}$
right – tailed	$p_1 - p_2 \leq 0$	$p_1 - p_2 > 0$	$z > z_{1-\alpha}$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if (variable : Z, Test Stat : z)
left – tailed	$p_1 - p_2 \geq 0$	$p_1 - p_2 < 0$	$p = prob[Z < z] < \alpha$
two – tailed	$p_1 - p_2 = 0$	$p_1 - p_2 \neq 0$	$p = 2\, prob[Z > \lvert z \rvert] < \alpha$
right – tailed	$p_1 - p_2 \leq 0$	$p_1 - p_2 > 0$	$p = prob[Z > z] < \alpha$

ILLUSTRATIONS

Exercise	H_0	H_1	Facts	Actual Decision	Conclusion for H_0
7.9 .1	$p_1 - p_2 = 0$	$p_1 - p_2 \neq 0$	$\alpha = 0.05$; $n_1 = 225$; $x_1 = 36$; $n_2 = 225$; $x_2 = 45$;	$z = -1.104 < -1.96$ $p = 0.26946 < 0.05$ both False	do not reject

7.9 .2	$p_1 - p_2 \geq 0$	$p_1 - p_2 < 0$	$\alpha = 0.05$; $n_1 = 160$; $p_1 = 0.58$; $n_2 = 150$; $p_2 = 0.63$;	$z = -0.8996 > -1.645$ $p = 0.18416 < 0.05$ both False	do not reject
7.9 .3	$p_B - p_A \leq 0$	$p_B - p_A > 0$	$\alpha = 0.05$; $n_B = 200$; $p_B = 0.15$; $n_A = 200$; $p_A = 0.12$;	$z = 0.8779 > 1.645$ $p = 0.19 < 0.05$ both False	do not reject
7.9 .4	$p_1 - p_2 \leq 0$	$p_1 - p_2 > 0$	$\alpha = 0.05$; n1 = 400; x1 = 288; n2 = 400; x2 = 260;	$z = 2.13 > 1.645$ $p = 0.0331 < 0.05$ both True	reject

TESTS OF INDEPENDENCE AND HOMOGENEITY

Hypotheses

H_0: *The R-row and C-column factors are mutually independent / Populations are homogeneous with respect to a specific category.*

H_1: *The R-row and C-column factors are mutually dependent / Populations are not homogeneous with respect to a specific category.*

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: CHI SQUARE

Test Statistic

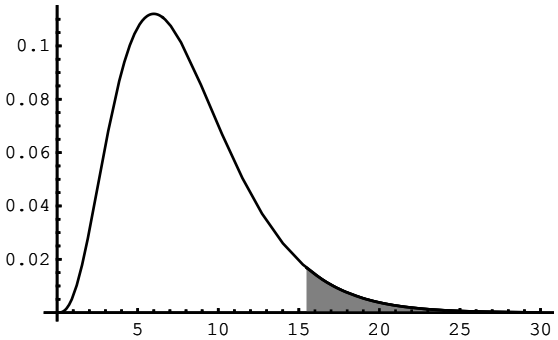
$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \; ; \; E_{ij} = \left(\sum_{j=1}^{c} O_{ij}\right)\left(\sum_{i=1}^{r} O_{ij}\right) \Big/ \left(\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\right)$$

r: number of rows, c: number of columns, O_{ij}: observed frequencies, E_{ij}: expected frequencies

Probability Distribution Function Form

$$\frac{2^{\frac{1-(r-1)\,(c-1)}{2}}\, e^{-\frac{x}{2}}\, x^{\frac{(r-1)\,(c-1)-1}{2}-1}}{\text{Gamma}\left[\frac{(r-1)\,(c-1)-1}{2}\right]}$$

Acceptance and Rejection Regions:right tailed tests only



The shaded region has an area α and defines the so called **rejection region,** meaning that if χ^2 falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if $\left(\text{variable : } X^2, \text{ Test Stat : } \chi^2\right)$
right – tailed	independence homogeneity	dependence nonhomogeneity	$\chi^2 > \chi^2_{1-\alpha}$ $\text{d.f} = (r-1)\,(c-1)$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if $\left(\text{variable : } X^2, \text{ Test Stat : } \chi^2\right)$
right – tailed	independence homogeneity	dependence nonhomogeneity	$p = prob\left[X^2 > \chi^2\right] < \alpha$ $\text{d.f} = (\text{r} - 1)\,(\text{c} - 1)$

ILLUSTRATIONS

Exercise	H_0	H_1	Facts	Actual Decision	Conclusion for H_0
12.4 .2	Skill level and gender are independent	Skill level and gender are dependent	$O_{11} = 106$, $O_{12} = 6$, $O_{21} = 93$, $O_{22} = 39$, $O_{31} = 215$, $O_{32} = 73$	$\chi^2 = 24.186 > 9.210$ $p = 5.598 \times 10^{-6}$	reject
12.5 .4	The two populations do not differ with respect to the proportion of innovative executives	The two populations differ with respect to the proportion of innovative executives	$O_{11} = 120$, $O_{12} = 90$, $O_{21} = 50$, $O_{22} = 120$	$\chi^2 = 29.223 > 3.841$ $p = 6.45 \times 10^{-8}$	reject

GOODNESS OF FIT TESTS

Hypotheses

H_0: ***The sample data comes from a population that follows a given probability distribution.***

H_1: ***The sample data does not come from a population that follows a given probability distribution.***

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: CHI SQUARE

Test Statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_k - E_k)^2}{E_k}$$

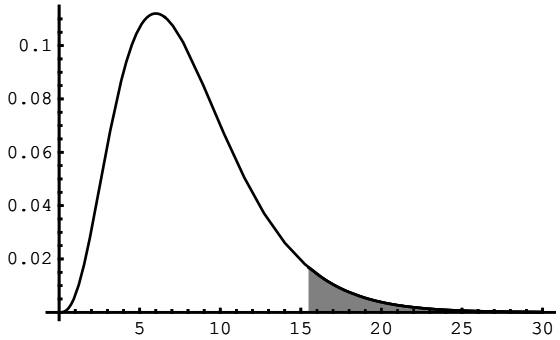
k: number of mutually exclusive and exhaustive categories, O_k: observed frequency of occurrence of sample values in category k, E_k: expected frequency in category k according to the hypothesized probability distribution model.

Probability Distribution Function Form

$$\frac{2^{\frac{1-(k-r)}{2}}\, e^{-\frac{x}{2}}\, x^{\frac{k-r-1}{2}-1}}{\text{Gamma}\left[\frac{k-r-1}{2}\right]}$$

r: number of constraints/restrictions imposed on the data.

Acceptance and Rejection Regions:right tailed tests only



The shaded region has an area α and defines the so called **rejection region,** meaning that if χ^2 falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if $\left(\text{variable}: X^2, \text{Test Stat}: \chi^2\right)$
right – tailed	The sample data comes from a population that follows a given probability distribution.	The sample data does not come from a population that follows a given probability distribution.	$\chi^2 > \chi^2_{1-\alpha}$ $\text{d.f} = \text{k} - \text{r}$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if $\left(\text{variable}: X^2, \text{Test Stat}: \chi^2\right)$
right – tailed	The sample data comes from a population that follows a given probability distribution.	The sample data does not come from a population that follows a given probability distribution.	$p = prob\left[X^2 > \chi^2\right] < \alpha$ $\text{d.f} = \text{k} - \text{r}$

ILLUSTRATIONS

THE NORMAL DISTRIBUTION MODEL (12.3.1)

Null Hypothesis (H_0): *The sample data comes from a population that is normally distributed.*

Alternative Hypothesis (H_1): *The sample data does not come from a population that is normally distributed.*

Finding the Expected Frequencies and the Test Stat Value

Classes (k)	O_k	z_+	$Prob[z \le z_+]$	$Prob[z_- \le z \le z_+]$	$E_k \ge 1$	$\frac{(O_k - E_k)^2}{E_k}$
< 40.0	10	−1.75	0.0401	0.0401	8.94	0.1257
40.0 – 49.9	12	−1.24	0.1075	0.0674	15.03	0.6108
50.0 – 59.9	17	−0.74	0.2296	0.1221	27.23	3.8433
60.0 – 69.9	37	−0.23	0.4090	0.1794	40.01	0.2264
70.0 – 79.9	55	0.28	0.6103	0.2013	44.89	2.2769
80.0 – 89.9	51	0.79	0.7852	0.1749	39.00	3.6923
90.0 – 99.9	34	1.30	0.9032	0.1180	26.31	2.2477
100.0 – 109.9	5	1.81	0.9649	0.0617	13.76	5.5769
≥ 110.0	2	∞	1.0000	0.0351	7.83	4.3409
Totals	***223***			***1.0000***	***223.00***	$\chi^2 = 22.941$

Restrictions/Constraints: Three[sample size: 223, sample mean: $\bar{x} = 74.47$, sample standard deviation: s=19.66].

Decision/Conclusion:

$\chi^2 = 22.941 > \chi^2_{0.95}[\text{d.f} = 9 - 3 = 6] = 12.592, \text{p} = 0.000816482 \ll 0.05$

Therefore, H_0 must be rejected, i.e., the sample data does not come from a population that is normally distributed.

THE POISSON DISTRIBUTION MODEL (12.3.2)

Null Hypothesis (H_0): *The sample data comes from a population that is Poisson distributed.*

Alternative Hypothesis (H_1): *The sample data does not come from a population that is Poisson distributed.*

Finding the Expected Frequencies and the Test Stat Value

k	O_k	$k\,O_k$	$prob[k]$	$E_k \ge 1$	$\frac{(O_k - E_k)^2}{E_k}$
0	3	0	0.006	1.8	0.800
1	9	9	0.028	8.4	0.043

2	21	42	0.075	22.5	0.100
3	38	114	0.129	38.7	0.013
4	46	184	0.168	50.4	0.384
5	54	270	0.175	52.5	0.043
6	49	294	0.151	45.3	0.302
7	34	238	0.113	33.9	0.000
8	20	160	0.073	21.9	0.165
9	16	144	0.042	12.6	0.197
10	6	60	0.022	6.6	0.055
11	3	33	0.011	3.3	0.027
≥ 12	1	12	0.007	2.1	0.576
Totals	***300***	***1560***	***1.0000***	***300***	$\chi^2 = 3.425$

Restrictions/Constraints: Two[sample size: 300, sample mean: $\lambda = 5.2$].

Decision/Conclusion:

$$\chi^2 = 3.425 < \chi^2_{0.95}[\text{d.f} = 13 - 2 = 11] = 19.675,\ p = 0.983835 > 0.05$$

Therefore, H_0 must not be rejected, i.e., the sample data comes from a population that is Poisson distributed.

THE BINOMIAL DISTRIBUTION MODEL (12.3.3)

Null Hypothesis (H_0): *The sample data comes from a population that is binomially distributed with $k = 0, 1, \square, 20$ and $p = 0.2$.*

Alternative Hypothesis (H_1): *The sample data does not come from a population that is binomially distributed with $k = 0, 1, \square, 20$ and $p = 0.2$.*

Finding the Expected Frequencies and the Test Stat Value

k	O_k	$k\,O_k$	$prob[k]$	$E_k \geq 1$	$\frac{(O_k - E_k)^2}{E_k}$
0	1	0	0.0115	1.153	0.0203
1	5	5	0.0576	5.765	0.1014
2	12	24	0.1369	13.691	0.2088
3	15	45	0.2054	20.536	1.4926
4	30	120	0.2182	21.820	3.0666
5	21	105	0.1746	17.456	0.7195
6	8	48	0.1091	10.910	0.7762
7	4	28	0.0545	5.455	0.3881
8	2	16	0.0222	2.216	0.0211
≥ 9	2	19	0.0100	1.000	1.0055
Totals	***100***				$\chi^2 = 7.8001$

Restrictions/Constraints: One[sample size:100].

Decision/Conclusion:

$$\chi^2 = 7.8001 < \chi^2_{0.95}[\text{d.f} = 10 - 1 = 9] = 16.919,\ p = 0.55441 > 0.05$$

Therefore, H_0 must not be rejected, i.e., the sample data comes from a binomially distributed population with $k = 0, 1, 2, \square, 20$ and $p = 0.2$.

THE UNIFORM DISTRIBUTION MODEL (12.3.5)

Null Hypothesis (H_0): *All days are equally preferred by the sample of 300 grocery shoppers.*

Alternative Hypothesis (H_1): *All days are not equally preferred by the sample of 300 grocery shoppers.*

Finding the Expected Frequencies and the Test Stat Value

k	O_k	E_k	$\frac{(O_k - E_k)^2}{E_k}$
M	10	300 / 7	25.1905
Tu	20	300 / 7	12.905
W	40	300 / 7	0.1905
Th	40	300 / 7	0.1905
F	80	300 / 7	32.1905
Sa	60	300 / 7	6.8571

Su	50	300 / 7	1.1905
Totals	*300*		$\chi^2 = 78.0000$

Restrictions/Constraints: One[sample size:300].

Decision/Conclusion:

$\chi^2 = 78.00 > \chi^2_{0.95}[\text{d.f} = 7 - 1 = 6] = 12.592,\ p = 9.21485 \times 10^{-15} \ll 0.05.$

Therefore, H_0 must be rejected, i.e., all days are not equally preferred by the sample of 300 grocery shoppers.

TESTS FOR ONE VARIANCE

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: CHI SQUARE

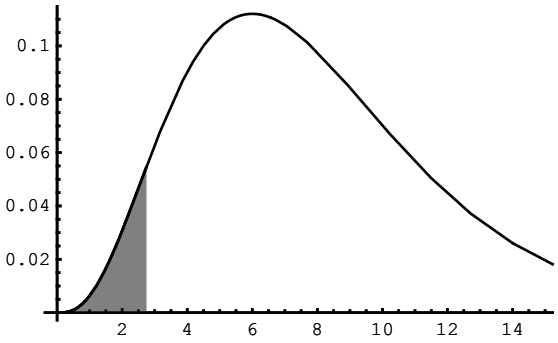
Test Statistic

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

Probability Distribution Function Form

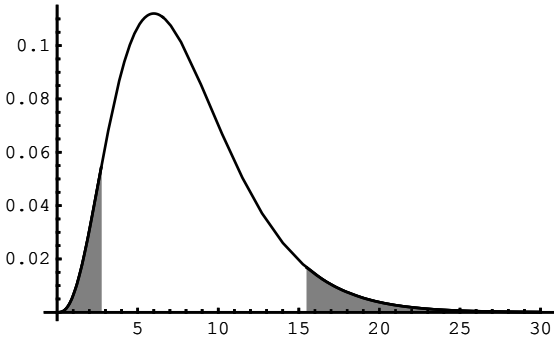
$$\frac{2^{\frac{1-n}{2}}\, e^{-\frac{x}{2}}\, x^{\frac{n-1}{2}-1}}{\text{Gamma}\left[\frac{n-1}{2}\right]}$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : \sigma^2 < \sigma_0^2$)



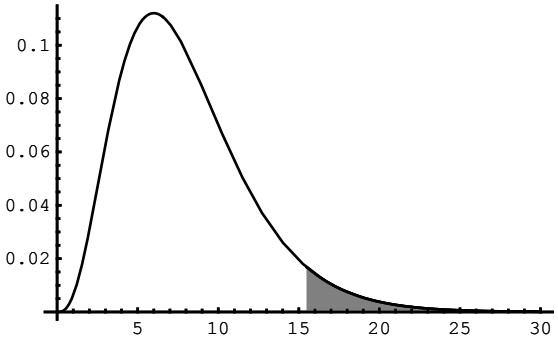
The shaded region has an area α and defines the so called **rejection region,** meaning that if χ^2 falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : \sigma^2 \neq \sigma_0^2$)



Both shaded regions have an area α and define the so called **rejection region,** meaning that if χ^2falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : \sigma^2 > \sigma_0^2$)



The shaded region has an area α and defines the so called **rejection region,** meaning that if χ^2 falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if $\left(\text{Test Stat} : \chi^2\right)$
left – tailed	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{\alpha}$
two – tailed	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	either $\chi^2 < \chi^2_{\alpha/2}$ or $\chi^2 > \chi^2_{1-\alpha/2}$
right – tailed	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{1-\alpha}$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if $\left(\text{variable} : X^2, \text{Test Stat} : \chi^2\right)$
left – tailed	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$p = prob\left[\boldsymbol{x}^2 < \chi^2\right] < \alpha$
two – tailed	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$p = prob\left[\boldsymbol{x}^2 < \chi^2\right] < \alpha/2$, if $\chi^2 \leq n-1$ $p = prob\left[\boldsymbol{x}^2 > \chi^2\right] < \alpha/2$, if $\chi^2 > n-1$
right – tailed	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$p = prob\left[\boldsymbol{x}^2 > \chi^2\right] < \alpha$

ILLUSTRATIONS

Exercise	H_0	H_1	Facts	Actual Decision	Conclusion for H_0
7.10 .1	$\sigma^2 = 15$	$\sigma^2 \neq 15$	$\alpha = 0.05$; $n = 21$; $s^2 = 10$;	$\chi^2 = 13.33 < 10.851$ $p = 0.1374 < 0.0250$ both False	do not reject
7.10 .2	$\sigma^2 \leq 0.00005$	$\sigma^2 > 0.00005$	$\alpha = 0.05$; $n = 31$; $s^2 = 0.000061$;	$\chi^2 = 36.6 > 43.773$ $p = 0.1891 < 0.05$ both False	do not reject

TESTS FOR TWO VARIANCES

DECISION MAKING:VISUAL APPRECIATION

SAMPLING DISTRIBUTION: F DISTRIBUTION

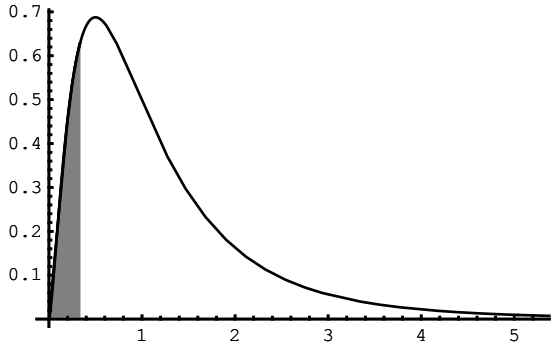
Test Statistic

$$F = \frac{s_1^2}{s_2^2}$$

Probability Distribution Function Form

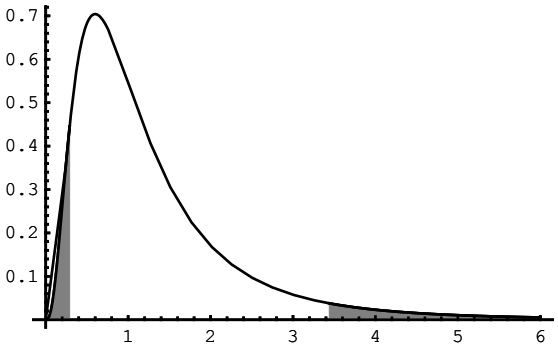
$$\left(f^{\frac{n1-1}{2}-1}\,(n1-1)^{\frac{n1-1}{2}}\,(n2-1)^{\frac{n2-1}{2}}\,(f\,(n1-1)+n2-1)^{\frac{1}{2}(-n1-n2+2)}\right)\Big/\,\text{Beta}\left[\frac{n1-1}{2},\,\frac{n2-1}{2}\right]$$

Acceptance and Rejection Regions:left tailed tests ($H_1 : \sigma_1^2 - \sigma_2^2 < 0$)



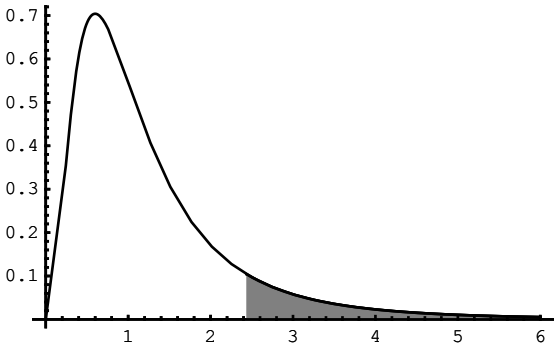
The shaded region has an area α and defines the **rejection region,** meaning that if F falls there, one must reject H_0.

Acceptance and Rejection Regions:two tailed tests ($H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$)



The shaded region has an area α and defines the **rejection region,** meaning that if F falls there, one must reject H_0.

Acceptance and Rejection Regions:right tailed tests ($H_1 : \sigma_1^2 - \sigma_2^2 > 0$)



The right shaded region has an area α and defines the **rejection region,** meaning that if F falls there, one must reject H_0.

TEST STAT DECISION RULES

Test Type	H_0	H_1	Reject H_0 if (Test Stat : F, d.f num : $n_1 - 1$, d.f den : $n_2 - 1$)
left – tailed	$\sigma_1^2 - \sigma_2^2 \geq 0$	$\sigma_1^2 - \sigma_2^2 < 0$	$F < F_{1-\alpha}[n_1 - 1, n_2 - 1] = F_{\alpha}^{-1}[n_2 - 1, n_1 - 1]$
two – tailed	$\sigma_1^2 - \sigma_2^2 = 0$	$\sigma_1^2 - \sigma_2^2 \neq 0$	either $F < F_{\alpha/2}[n_1 - 1, n_2 - 1] = F_{1-\alpha/2}^{-1}[n_2 - 1, n_1 - 1]$ or $F > F_{1-\alpha/2}[n_1 - 1, n_2 - 1]$
right – tailed	$\sigma_1^2 - \sigma_2^2 \leq 0$	$\sigma_1^2 - \sigma_2^2 > 0$	$F > F_{1-\alpha}[n_1 - 1, n_2 - 1]$

P-VALUE DECISION RULES

Test Type	H_0	H_1	Reject H_0 if (variable : Ö, Test Stat : F, d.f num : $n_1 - 1$, d.f den : $n_2 - 1$)
left – tailed	$\sigma_1^2 - \sigma_2^2 \geq 0$	$\sigma_1^2 - \sigma_2^2 < 0$	$p = prob[ö < F] < \alpha$
two – tailed	$\sigma_1^2 - \sigma_2^2 = 0$	$\sigma_1^2 - \sigma_2^2 \neq 0$	$p = prob[ö > F] < \alpha / 2$
right – tailed	$\sigma_1^2 - \sigma_2^2 \leq 0$	$\sigma_1^2 - \sigma_2^2 > 0$	$p = prob[ö > F] < \alpha$

ILLUSTRATIONS

Exercise	H_0	H_1	Facts	Actual Decision	Conclusion for H_0
7.11 .2	$\sigma_1^2 - \sigma_2^2 \leq 0$	$\sigma_1^2 - \sigma_2^2 > 0$	$\alpha = 0.05$; n1 = 25; $s_1^2 = 96$ n2 = 121;; $s_2^2 = 144$;	$F = \frac{144}{96} = 1.5 < 1.79$ $p = 0.1253 < 0.05$ both False	do not reject
7.11 .3	$\sigma_1^2 - \sigma_2^2 = 0$	$\sigma_1^2 - \sigma_2^2 \neq 0$	$\alpha = 0.05$; n1 = 8; $s_1^2 = 2916$ n2 = 8;; $s_2^2 = 4624$;	$F = \frac{4624}{2916} = 1.586 > 4.99$ $p = 0.278904 < 0.0250$ both False	do not reject

ONE-WAY ANALYSIS OF VARIANCE

Hypotheses

H_0: *The samples selected come from populations with equal means.*

H_1: *The samples selected come from populations with means not all equal.*

Summary Data

k : number of samples / treatments; j = 1, 2, 3, □ , k;

n_j : size of sample (j); i = 1, 2, 3, □ , n_j; $N = \sum_{j=1}^{k} n_j$

sample totals : $T_j = \sum_{i=1}^{n_j} x_{ij}$; grand total : $T = \sum_{j=1}^{k} T_j$;

sums of squares : $\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2$;

Sums of Squared Differences

$$SSTr = \sum_{j=1}^{k} \left(\frac{T_j^2}{n_j} \right) - \frac{T^2}{N};$$

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{k} \left(\frac{T_j^2}{n_j} \right);$$

$$SST = SSTr + SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2 - \frac{T^2}{N}$$

Anova Table

Source of Variation	Sums	d.f	Mean Squares	F - Ratio
Between Treatments	SSTr	$k - 1$	$MSTr = \frac{SSTr}{k-1}$	$F = \frac{MSTr}{MSE}$
Within Treatments	SSE	$N - k$	$MSE = \frac{SSE}{N-k}$	
Total	SST	$N - 1$		

Decision

Reject H_0 if

$$F > F_{1-\alpha}[d.f\ num: k-1,\ d.f\ den: N-k]$$

or if

$$prob[Ö > F] < \alpha$$

Illustration: 8.2.4

Hypotheses

H_0: The three gas brands yield the same mileage.
H_1: At least one gas brand yields a different mileage.

Summary Data

$k = 3$; $T_A = 120$; $T_B = 135$; $T_C = 164$; $T = 419$; $Ó x^2 = 8563$;

Anova Table/Decision

Source of Variation	Sums	d.f	Mean Squares	F - Ratio
Between Treatments	$SSTr = 8503 - \frac{419^2}{21} = 142.9524$	$3 - 1 = 2$	$MSTr = 71.4762$	$F = 21.4429 > F_{0.99}[2, 18] = 6.01$ $p = 0.0000173 < 0.05$
Within Treatments	$SSE = 8563 - 8503 = 60$	$21 - 3 = 18$	$MSE = 3.3333$	$\therefore$ reject H_0
Total	$SST = 8563 - \frac{419^2}{21} = 202.9524$	$21 - 1 = 20$		

TWO-WAY ANALYSIS OF VARIANCE

Hypotheses

H_0: *The samples selected (columns-treatments, rows-blocks) come from populations with equal means / There is no column-trearments effect.There is no row-blocks effect.*

H_1: *The samples selected (columns-treatments, rows-blocks) come from populations with means not all equal./ There is column-trearments effect.There is row-blocks effect.*

Summary Data

k : number of treatments (columns); j = 1, 2, 3, □ , k;

n : number of blocks (rows); i = 1, 2, 3, □ , n

$$\text{treatment totals : } T_j = \sum_{i=1}^{n} x_{ij}; \quad \text{block totals : } T_i = \sum_{j=1}^{k} x_{ij};$$

$$\text{Grand Total : } T = \sum_{j=1}^{k} T_j = \sum_{i=1}^{n} T_i$$

$$\text{Sums of Squares : } \sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}^2;$$

Sums of Squared Differences

$$\mathbf{SSTr} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{j=1}^{k} \mathbf{T_j^2} - \frac{\mathbf{T^2}}{\mathbf{n\,k}}; \quad \mathbf{SSB} = \frac{\mathbf{1}}{\mathbf{k}} \sum_{j=1}^{n} \mathbf{T_i^2} - \frac{\mathbf{T^2}}{\mathbf{n\,k}};$$

$$\mathbf{SSE = SST - SSTr - SSB;}$$

$$\mathbf{SST} = \sum_{j=1}^{k}\sum_{i=1}^{n} \mathbf{x_{ij}^2} - \frac{\mathbf{T^2}}{\mathbf{n\,k}};$$

Two-Way Anova Table

Source of Variation	Sums	d.f	Mean Squares	F - Ratio
Between Treatments (columns)	SSTr	k - 1	$MSTr = \frac{SSTr}{k-1}$	$F_c = \frac{MSTr}{MSE}$
Between Blocks (rows)	SSB	n - 1	$MSB = \frac{SSB}{n-1}$	$F_r = \frac{MSB}{MSE}$
Error	SSE	(k - 1) (n - 1)	$MSE = \frac{SSE}{(k-1)\,(n-1)}$	
Total	SST	k n - 1		

Decisions

Reject H_0 for **treatments (columns)** if

$$\mathbf{F_c > F_{1-\alpha}[d.f\ num: k-1,\ d.f\ den: (k-1)\,(n-1)]}$$

or if

$$\mathbf{prob[Ö > F_c] < \alpha}$$

Reject H_0 for **blocks (rows)** if

$$\mathbf{F_r > F_{1-\alpha}[d.f\ num: n-1,\ d.f\ den: (k-1)\,(n-1)]}$$

or if

$$\mathbf{prob[Ö > F_r] < \alpha}$$

Illustration: 8.4.2

Hypotheses

H_0 : There is no difference in detection ability among the three methods (treatments/columns).

H_1 : There is a difference in detection ability among the three methods (treatments/columns).

Summary Data

$k = 3\ ;\ n = 4;\ T_A = 16;\ T_B = 20;\ T_C = 33;\ T_I = 17;\ T_{II} = 17;\ T_{III} = 18;\ T_{IV} = 17;$
$\sum T_i^2 = 1191;\ \sum T_j^2 = 1745;\ T = 69;\ \acute{O}x^2 = 443;$

Two-Way Anova Table/Decision

Source of Variation	Sums	d.f	Mean Squares	F - Ratio
Between Treatments (columns)	$SSTr = \frac{1}{4}\,1745 - \frac{69^2}{3 \cdot 4} = 39.5$	$3 - 1 = 2$	$MSTr = 19.75$	$F_c = 18.23 > 5.14$ $p = 0.0028 < 0.05$ $\therefore$ reject H_0
Between Blocks (rows)	$SSB = \frac{1}{3}\,1191 - \frac{69^2}{3 \cdot 4} = 0.25$	$4 - 1 = 3$	$MSB = 0.0833$	$F_r = 0.0769$
Error	$SSE = 46.25 - 39.5 - 0.25 = 6.5$	$2 \cdot 3 = 6$	$MSE = 1.0833$	
Total	$SST = 443 - \frac{69^2}{3 \cdot 4} = 46.25$	$3 \cdot 4 - 1 = 11$		

THE KRUSKAL WALLIS H-TEST:

The Non-Parametric Alternative to ONE-WAY ANOVA

Ground of Applications To test whether or not several population means are equal when the sampled populations are neither normally distributed nor their variances are equal.

Basic Idea Rank all observations in the combined set of data from low to high. For ties, associate to each tied observation the corresponding arithmetic mean rank corresponding to the tied observations.

Notation k: the number of independent samples; n_j:the number of observations in the jth sample; $n = \sum n_j$: the total number of observations in all samples; R_j: the sum of the ranks in the jth sample.

Test Statistic

$$H = \frac{12}{n\,(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3\,(n+1)$$

Decision

Reject H_0 if

$$H > \chi^2_{1-\alpha}[d.f : k-1]$$

or if

$$prob\left[X^2 > H\right] < \alpha$$

Illustrations

Exercise13.9.1 (ranked data)

(A) Hypotheses

H_0 : The three coffee brands are equally preferred.

H_1 : The three coffee brands are not equally preferred.

(B) Rank Sums/Test Statistic

$n_A = n_B = n_C = 5;\ n = 15$

$R_A = 9 + 10 + 11 + 12 + 13 = 55$

$R_B = 14 + 1 + 5 + 7 + 8 = 35$

$R_C = 2 + 3 + 4 + 15 + 6 = 30$

Note that $R_A + R_B + R_C = 120[1 + 2 + \cdots + 15 = 15 \cdot 16/2]$.

$$H = \frac{12}{15 \cdot (15+1)}\left[\frac{55^2}{5} + \frac{35^2}{5} + \frac{30^2}{5}\right] - 3 \cdot (15+1) = 3.5$$

(C/D) Decision/Conclusion

Since $H = 3.5 < \chi^2_{0.95}[3-1=2] = 5.991$, one can not reject H_0 (p-value decision: since $p = 0.173774 > 0.05$, one can not reject H_0). Therefore, we conclude with 95 % confidence that the three coffee brands are equally preferred.

Exercise13.9.4 (ranked data)

(A) Hypotheses

H_0 : The three teaching methods are equally effective

H_1 : The three teaching methods are not equally effective

(B) Rank Sums/Test Statistic

$n_A = n_B = n_C = 5;\ n = 15$

$R_A = 5 + 7 + 9 + +\ \ 3 + 6 = 30$

$R_B = 4 + 1 + 8 + 2 + 10 = 25$

$R_C = 15 + 13 + 11 + 12 + 14 = 65$

Note that $R_A + R_B + R_C = 120[1 + 2 + \cdots + 15 = 15 \cdot 16/2]$.

$$H = \frac{12}{15 \cdot (15+1)}\left[\frac{30^2}{5} + \frac{25^2}{5} + \frac{65^2}{5}\right] - 3 \cdot (15+1) = 9.5$$

(C/D) Decision/Conclusion

Since $H = 9.5 > \chi^2_{0.95}[3-1=2] = 5.991$, reject H_0 (p-value decision: since $p = 0.0233314 < 0.05$, reject H_0). Therefore, we conclude with 95 % confidence that the three teaching methods are not equally effective.

Exercise 13.9.2 (data to be ranked)

(A) Hypotheses

H_0 : The four new formulas for cake mix are not different.

H_1 : The four new formulas for cake mix are different.

(B) Ranking of Data, Rank Sums/Test Statistic

Formula A	72 (3)	88 (10)	70 (1)	87 (7.5)	71 (2)
Formula B	85 (5)	89 (12.5)	86 (6)	82 (4)	88 (10)
Formula C	94 (18)	94 (18)	88 (10)	87 (7.5)	89 (12.5)
Formula D	91 (14)	93 (16)	92 (15)	95 (20)	94 (18)

```
Sort[{72, 88, 70, 87, 71, 85, 89, 86, 82, 88, 94, 94, 88, 87, 89, 91, 93, 92, 95, 94}]
```

```
{70, 71, 72, 82, 85, 86, 87, 87, 88, 88, 88, 89, 89, 91, 92, 93, 94, 94, 94, 95}
```

$n_A = n_B = n_C = n_D = 5;\ n = 20$

$R_A = 3 + 10 + 1 + 7.5 + 2 = 23.5$

$R_B = 5 + 12.5 + 6 + 4 + 10 = 37.5$

$R_C = 18 + 18 + 10 + 7.5 + 12.5 = 66$

$R_D = 14 + 16 + 15 + 20 + 18 = 83$

Note that $R_A + R_B + R_C + R_D = 210[1 + 2 + \cdots + 20 = 20 \cdot 21/2]$.

$$H = \frac{12}{20 \cdot (20+1)}\left[\frac{23.5^2}{5} + \frac{37.5^2}{5} + \frac{66^2}{5} + \frac{83^2}{5}\right] - 3 \cdot (20+1) = 12.4486$$

(C/D) Decision/Conclusion

Since $H = 12.4486 > \chi^2_{0.95}[4-1=3] = 7.815$, reject H_0 (p-value decision: since $p = 0.00599428 < 0.05$, reject H_0). Therefore, we conclude with 95 % confidence that the four formulas are different.

SIMPLE LINEAR REGRESSION, DETERMINATION, CORRELATION AND PREDICTION

Linear Regression: To determine that linear function between any two variables y (dependent) and x(independent) that best fits a given bivariate data for these variables.
Correlation/Determination: To measure the strength of the linear relationship between the variables x and y and reach a decision as to whether this linear relationship has a significant explanatory power, i.e., whether it accounts for a substantial part of the variation in the given data.

Variables and Bivariate Data

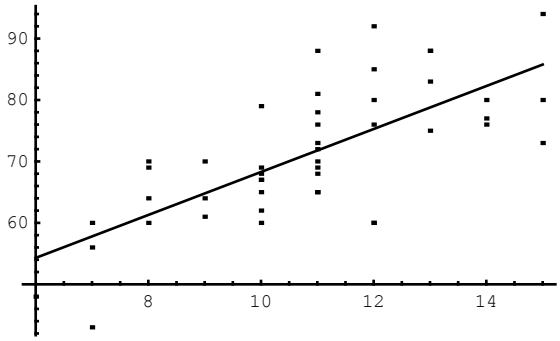
Let x and y be the independent and dependent variables under investigation, respectively. The bivariate sample data consists of measurements of both x and y which come in pairs. Let $(x_1, y_1), (x_2, y_2), (x_3, y_3), \square, (x_i, y_i), \square, (x_n, y_n)$ be n such measurements.

Scatter Diagram

A first approach in the study of the relationship between x and y, is to to visualize the data by plotting the observed pairs of observations as points on the x – y plane.

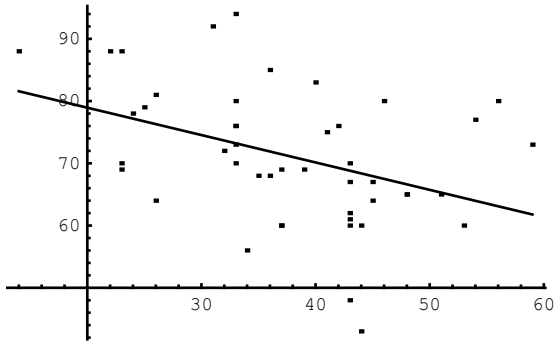
x,y can be positively correlated

The figure below shows data with a positive-upward trend between the variables:



x,y can be negatively correlated

The figure below shows data with a negative-downward trend between the variables :



Summary Data

sample size : n

simple sums : $\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i$

sums of squares : $\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} x_i y_i$

Variation Measures

variation in x : $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\,\overline{x}^2$

variation in y : $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\,\overline{y}^2$

co – variation in x and y : $S_{xy} = \sum_{i=1}^{n} x_i\,y_i - n\,\overline{x}\,\overline{y}$

S_{xx} and S_{yy} are positive definite quantities. The co-variation S_{xy} can be negative, zero or positive.

The Method of Least Squares and the Regression Line y on x

The problem: Among all possible linear functions $\hat{y} = a + bx$, find the one which minimizes the sum of squares

$$\sum_{i=1}^{n} (a + bx_i - y_i)^2$$

The solving procedure of this optimization problem involves the use of differential calculus.

The solution: The equation of the line of best fit/regression line is $\hat{y} = a + bx$ with

$$b = \frac{S_{xy}}{S_{xx}},\ a = \overline{y} - b\,\overline{x}$$

and the least sum of squares, also called error sum of squares or sum of squares for error , is

$$SSE = S_{yy}\left(1 - \frac{S_{xy}^2}{S_{xx}\,S_{yy}}\right).$$

In other words, the expected values $\hat{y}_i$ of the dependent variable (determined from the regression line) and the observed values y_i for all values x_i, differ by an amount which is the least possible.

The Coefficient of Determination

$$r^2 \equiv \frac{S_{xy}^2}{S_{xx}\,S_{yy}}$$

Since $S_{yy} > 0$, and the least sum of squares can not be negative, we must have that $r^2 \le 1$. For the particular case where $r^2 = 1$, the least sum of squares becomes zero , i.e., all sample pair observations fall right on the regression line. Therefore, the coefficient of determination is a percentage grade (ranging from 0 to 100 %) and expresses how close data points are to the regression line, i.e., it is a measure of the linearity of the data points, of how perfect the fit of the regression line is to the sample data.

The least sum of squares can be written as $SSE = S_{yy}\left(1 - r^2\right)$. Solving for r^2 we have

$$r^2 = \frac{S_{yy} - SSE}{SSE}$$

The quantity SSE may be interpreted as the variation in the dependent variable "left unexplained" by the regression function. Then the above formula implies that r^2 may also be interpreted as the % of the total variation in y that is explained by the regression of y on x.

Pearson's Correlation Coefficient

$$r \equiv \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}$$

Since $0 \le r^2 \le 1$, it obviously follows that $-1 \le r \le 1$. If $S_{xy} < 0$, $r < 0$ and the variables are said to be negatively correlated. If $S_{xy} > 0$, $r > 0$ and the variables are said to be positively correlated. Finally, If $S_{xy} = 0$, $r = 0$ and the variables are said to be non correlated. Therefore, Pearson's Correlation Coefficient measures the strength of the relationship between x and y without discriminating which of the two variables is the dependent and which the independent.

Anova for Regression

To reach a decision on whether two variables x(independent) and y(dependent) are linearly related , it is necessary that the number of paired observations must also be taken into account. The modern decision making on this issue is effected by applying the analysis of variance to regression, henceforth called ANOVA for regression and is outlined below.

Hypotheses

H_0 : The variables x and y are not linearly related

H_1 : The variables x and y are linearly related

Source of Variation	Sums	d.f	Mean Squares	F - Ratio
Regression	$SSR = r^2 S_{yy}$	1	$MSR = r^2 S_{yy}$	$F = \frac{n-2}{1-r^2} r^2$
Error	$SSE = (1 - r^2) S_{yy}$	n - 2	$MSE = \frac{1-r^2}{n-2} S_{yy}$	
Total	$SST = S_{yy}$	n - 1		

Decision

Reject H_0 if

$$F = \frac{n-2}{1-r^2} r^2 > F_{1-\alpha}[\text{d.f num} = 1, \text{d.f den} = n-2]$$

Prediction of y for a given x

One may get a trivial point-prediction of the value that y assumes on average for a given x by means of the regression equation. But what is more important and extremely more interesting is to make non-trivial interval predictions taking into account that all possible y-values for any given x, say x_k, follow a t-distribution with the specifications:

degrees of freedom : $n - 2$

mean : $\hat{y}_k = a + b x_k$

standard deviation : $s_{\hat{y}_k} = \sqrt{\frac{1-r^2}{n-2}\left(1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{S_{xx}}\right) S_{yy}}$

In particular, confidence intervals for the most probable y-values can be constructed : We are $100\ (1-\alpha)\ \%$ confident that for a given value of $x = x_k$, the feasible y –values are contained in the interval $\hat{y}_k \pm t_{1-\alpha/2}\, s_{\hat{y}_k}$.

An Illustration: Exercise 9.5.11

Variables: midterm grade(x), final grade(y)

Summary Data

$$n = 21,\ \sum_{i=1}^{n} x_i = 1749,\ \sum_{i=1}^{n} y_i = 1706,\ \sum_{i=1}^{n} x_i^2 = 147\,051,\ \sum_{i=1}^{n} y_i^2 = 141\,308,\ \sum_{i=1}^{n} x_i\, y_i = 143\,312$$

Variation Measures

$$S_{xx} = 147\,051 - 21\left(\frac{1749}{21}\right)^2 = \frac{9690}{7},\ S_{yy} = 141\,308 - 21\left(\frac{1706}{21}\right)^2 = \frac{57\,032}{21},$$

$$S_{xy} = 143\,312 - 21\left(\frac{1749}{21}\right)\left(\frac{1706}{21}\right) = \frac{8586}{7}$$

The Coefficient of Determination

$$r^2 = \frac{18\,429\,849}{46\,053\,340} = 0.400185$$

Anova for Regression

H_0 : The variables x and y are not linearly related

H_1 : The variables x and y are linearly related

Source of Variation	Sums	d.f	Mean Squares	F - Ratio
Regression	SSR = 1086.83	1	MSR = 1086.83	$F = 12.6764$
Error	SSE = 1628.98	19	MSE = 85.736	
Total	SST = 2715.81	20		

Since $F = 12.6764 > F_{0.95}[1, 19] = 4.38$, reject H_0. Hence,we are 95% confident that the variables x and y are linearly related.

The Regression Line y on x

$$\hat{y} = 7.44128 + 0.886068\,x$$

Prediction of y for a given x: If a student's midterm grade is 65, (1) predict his final exam score with 95% confidence, and (2) find the probability that his final exam score will be at least 90.

(1) **expected mean :** $y\,(x = 65) = 65.037$; **prediction standard error :** $s = 10.5132$;
critical t – value : $t_{0.9750}[19] = 2.0930$; **C.I**$_{95\%}$ **:** $65.037 \pm (2.0930)\,(10.5132)$
min predicted final score : 43.03, **max predicted final score :** 87.04
(2) **Using *Mathematica* :** $\text{prob}[y \geq 89.5 \mid x = 65] = \text{StudentTPValue}[10.66, 19] = 9.26397 \times 10^{-10}$, **i.e., one in a billion!**

MULTIPLE LINEAR REGRESSION, DETERMINATION, CORRELATION AND PREDICTION

Linear Regression: To determine that linear function between any number of variables $y \equiv x_0$ (dependent) and $x_1, x_2, \square, x_k$(independent) that best fits a given multivariate data for these variables.
Correlation/Determination: To measure the strength of the linear relationship between the variables $x_0, x_1, x_2, \square, x_k$ and reach a decision as to whether this linear relationhip has a significant explanatory power, i.e., whether it accounts for a substantial part of the variation in the given multivariate data.

Variables and Multivariate Data

Let $x_1, x_2, \square, x_k$ and $y \equiv x_0$ be the independent and dependent variables under investigation, respectively. The multivariate sample data consists of simultateous measurements of all $k + 1$ variables. Let

$$(x_{11}, x_{21}, x_{31,\square}, x_{01}),\ (x_{12}, x_{22}, x_{32,\square}, x_{02}),\ (x_{13}, x_{23}, x_{33,\square}, x_{03}),$$
$$\square,\ (x_{1\,i}, x_{2\,i}, x_{3\,i,\square}, x_{0\,i}),\ \square,\ (x_{1\,n}, x_{2\,n}, x_{3\,n,\square}, x_{0\,n}),$$

be n such measurements. In most of what follows, we shall restrict to the case of two independent variables ($k = 2$).

Summary Data

For the case of two independent variables, the summary data consists of 3 simple sums and of 6 sums of squares:

sample size : $\mathbf{n}$

simple sums : $\sum_{i=1}^{n} \mathbf{x}_{0\,i},\ \sum_{i=1}^{n} \mathbf{x}_{1\,i},\ \sum_{i=1}^{n} \mathbf{x}_{2\,i},$

sums of squares : $\sum_{i=1}^{n} \mathbf{x}_{0\,i}^2,\ \sum_{i=1}^{n} \mathbf{x}_{1\,i}^2,\ \sum_{i=1}^{n} \mathbf{x}_{2\,i}^2,\ \sum_{i=1}^{n} \mathbf{x}_{0\,i}\,\mathbf{x}_{1\,i},\ \sum_{i=1}^{n} \mathbf{x}_{0\,i}\,\mathbf{x}_{2\,i},\ \sum_{i=1}^{n} \mathbf{x}_{1\,i}\,\mathbf{x}_{2\,i}$

In the general case of k independent variables, the summary data will consist of $k + 1$ simple sums and of $(k + 1) \cdot (k + 2) / 2$ sums of squares.

Variation Measures

variation in x_1 : $\mathbf{S_{11}} = \sum_{i=1}^{n} \mathbf{x}_{1\,i}^2 - \mathbf{n}\,\overline{\mathbf{x}_1}^2$

variation in x_2 : $\mathbf{S_{22}} = \sum_{i=1}^{n} \mathbf{x}_{2\,i}^2 - \mathbf{n}\,\overline{\mathbf{x}_2}^2$

variation in y : $\mathbf{S_{00}} = \sum_{i=1}^{n} \mathbf{x}_{0\,i}^2 - \mathbf{n}\,\overline{\mathbf{x}_0}^2$

co – variation of x_1 and y : $\mathbf{S_{01}} = \sum_{i=1}^{n} \mathbf{x}_{1\,i}\,\mathbf{y}_i - \mathbf{n}\,\overline{\mathbf{x}_{1\,i}}\,\overline{\mathbf{y}}$

co – variation of x_2 and y : $\mathbf{S_{02}} = \sum_{i=1}^{n} \mathbf{x}_{2\,i}\,\mathbf{y}_i - \mathbf{n}\,\overline{\mathbf{x}_{2\,i}}\,\overline{\mathbf{y}}$

co – variation of x_1 and x_2 : $\mathbf{S_{12}} = \sum_{i=1}^{n} \mathbf{x}_{1\,i}\,\mathbf{x}_{2\,i} - \mathbf{n}\,\overline{\mathbf{x}_{1\,i}}\,\mathbf{x}_{2\,i}$

S_{11}, S_{22} and S_{00} are positive definite quantities. The co-variations S_{01}, S_{02},S_{12} can be negative, zero or positive.

The Method of Least Squares and the Regression Plane y on x_1, x_2

The problem: Among all possible linear functions $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$, find the one which minimizes the sum of squares

$$\sum_{i=1}^{n} (b_0 + b_1 x_{1\,i} + b_2 x_{2\,i} - y_i)^2$$

The solving procedure of this optimization problem also involves the use of differential calculus and is quite lengthy.

The solution:

The equation of the plane of best fit/regression plane is $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ with

$$b_1 = \frac{S_{01}\, S_{22} - S_{02}\, S_{12}}{S_{11}\, S_{22} - S_{12}^2},\ b_2 = \frac{S_{02}\, S_{11} - S_{01}\, S_{12}}{S_{11}\, S_{22} - S_{12}^2},\ b_0 = \overline{y} - b_1\, \overline{x_1} - b_2\, \overline{x_2}.$$

The least sum of squares, also called error sum of squares or sum of squares for error , is

$$SSE = S_{00} \left(1 - \frac{r_{01}^2 + r_{02}^2 - 2\, r_{01}\, r_{02}\, r_{12}}{1 - r_{12}^2}\right),$$

where r_{01}, r_{02}, r_{12} are the Pearson's correlation coefficients

$$r_{01} = \frac{S_{01}}{\sqrt{S_{00}\, S_{11}}},\ r_{02} = \frac{S_{02}}{\sqrt{S_{00}\, S_{22}}},\ r_{12} = \frac{S_{12}}{\sqrt{S_{11}\, S_{22}}}.$$

In other words, the expected values $\hat{y}_i$ of the dependent variable (determined from the regression plane) and the observed values y_i for all values $x_{1\,i}$, $x_{2\,i}$, differ by an amount which is the least possible.

The Coefficient of Multiple Determination

$$R^2 \equiv \frac{r_{01}^2 + r_{02}^2 - 2\, r_{01}\, r_{02}\, r_{12}}{1 - r_{12}^2}.$$

Since $S_{00} > 0$, and the least sum of squares can not be negative, we must have that $R^2 \leq 1$. For the particular case where $R^2 = 1$, the least sum of squares becomes zero , i.e., all sample observations fall right on the regression plane. Therefore, the coefficient of multiple determination too is a percentage grade (ranging from 0 to 100 %) and expresses how close data points are to the regression plane, i.e., it is a measure of the linearity of the data points, of how perfect the fit of the regression plane is to the sample data.

The least sum of squares can be written as $SSE = S_{00}\left(1 - R^2\right)$. Solving for R^2 we have

$$R^2 = \frac{S_{00} - SSE}{SSE}.$$

The quantity SSE may be interpreted as the variation in the dependent variable "left unexplained" by the regression function. Then the above formula implies that R^2 may also be interpreted as the % of the total variation in y that is explained by the regression of y on x_1, x_2.

If $r_{12} = 0$, i.e., the independent variables x_1, x_2 are independent, the R^2 form siplifies to

$$R^2 = r_{01}^2 + r_{02}^2$$

Therefore, when using truly independent variables, the % of the total variation in y that is explained by the regression of y on x_1 and the % of the total variation in y that is explained by the regression of y on x_2 add up to the % of the total variation in y that is explained by the regression of y on both x_1 and x_2.

Anova for Multiple Regression

To reach a decision on whether any number of variables are linearly related , it is necessary that the number of observations must also be taken into account. The modern decision making on this issue is effected by extending in a natural way the analysis of variance of simple regression to the multiple regression case.

Hypotheses

H_0 : The variables $x_0, x_1, x_2, \square, x_k$ are not linearly related.

H_1 : The variables $x_0, x_1, x_2, \square, x_k$ are linearly related.

Source of Variation	Sums	d.f	Mean Squares	F - Ratio
Regression	$SSR = R^2\, S_{00}$	k	$MSR = R^2\, S_{00}$	$F = \frac{n-k-1}{1-R^2}\, R^2$
Error	$SSE = \left(1 - R^2\right) S_{00}$	n - k - 1	$MSE = \frac{1-R^2}{n-k-1}\, S_{00}$	
Total	$SST = S_{00}$	n - 1		

Decision

Reject H_0 if

$$F = \frac{n-k-1}{1-R^2} R^2 > F_{1-\alpha}[d.f\,num = k,\ d.f\,den = n-k-1]$$

Prediction of y for given $x_1, x_2, \ldots x_k$

One may get a trivial point-prediction of the value that y assumes on average for given values of the independent variables by means of the regression equation. But what is more important and extremely more interesting is to make non-trivial interval predictions taking into account that all possible y-values for any given $x_1, x_2, \ldots$ values, say $\{x_{1\,p}, x_{2\,p}, \ldots, x_{kp}\}$, follow a t-distribution with the specifications:

degrees of freedom : $n-k-1$

mean : $\hat{y}_p = b_0 + b_1 x_{1\,p} + b_2 x_{2\,p} + \cdots + b_k x_{k\,p}$

standard deviation : $s_{\hat{y}_p} = \sqrt{\left(\frac{1-R^2}{n-k-1}\left(1+\frac{1}{n}+\sum_{i,j=1}^{k}\frac{\left(x_{i\,p}-\overline{x_i}\right)\left(x_{j\,p}-\overline{x_j}\right)}{S_{ij}}\right)S_{00}\right)}$

In particular, confidence intervals can be constructed for the most probable y-values : We are $100\,(1-\alpha)\,\%$ confident that for the given values $\{x_{1\,p}, x_{2\,p}, \ldots, x_{k\,p}\}$ of the independent (predictor) variables, the feasible y –values are contained in the interval $\hat{y}_p \pm t_{1-\alpha/2}\, s_{\hat{y}_p}$.

Multiple Regression with *Mathematica*

To deal nowadays manually -with pencil,paper, and a calculator- beyond simple linear regression, is a tremendous task of almost no educational value. We present instead the print-outs of *Mathematica* for two regression exercises: one with two independent variables, and the other with five independent variables.Both print-outs may be appreciated by the general treatment given above.

Exercise 10.3.3

Variables/Data:

Sales in $1000 (y) on advertising expenditures in$1000 (x1) and on population density in people per square mile (x2).The data consists of 10 simultaneous observations of y, x1, x2.

MAthematica print-out

```
data1033={{0.2,50,20},{0.2,50,25},{0.2,50,24},{0.3,60,30},{0.3,60,32},{0.4,70,40},{0.3,50,28},{0.5,75,50},{0.4,70,40},{0.5,74,50}};

function1033=Fit[data1033,{1,x1,x2},{x1,x2}]

-3.81678+68.9469 x1+0.245719 x2

regress1033=Regress[data1033,{1,x1,x2},{x1,x2}]
```

	""	"Estimate"	"SE"	"TStat"	"PValue"
ParameterTable →	1	−3.81678	6.30948	−0.604928	0.564314
	x1	68.9469	17.1563	4.01876	0.00506759
	x2	0.245719	0.187513	1.31041	0.231414

RSquared → 0.976636, AdjustedRSquared → 0.969961,

EstimatedVariance → 3.39408, ANOVATable →

	"DF"	"SumOfSq"	"MeanSq"	"FRatio"	"PValue"
"Model"	2	993.141	496.571	146.305	1.94938×10^{-6}
"Error"	7	23.7586	3.39408	""	""
"Total"	9	1016.9	""	""	""

```
Regress[data1033,{1,x1,x2},{x1,x2},RegressionReport→{SinglePredictionCITable}]
```

SinglePredictionCITable →

"Observed"	"Predicted"	"SE"	"CI"
20.	22.2586	2.05269	{17.4047, 27.1124}
25.	22.2586	2.05269	{17.4047, 27.1124}
24.	22.2586	2.05269	{17.4047, 27.1124}
30.	31.6104	1.96511	{26.9637, 36.2572}
32.	31.6104	1.96511	{26.9637, 36.2572}
40.	40.9623	2.04674	{36.1226, 45.8021}
28.	29.1533	2.4849	{23.2774, 35.0291}
50.	49.0856	2.13567	{44.0356, 54.1357}
40.	40.9623	2.04674	{36.1226, 45.8021}
50.	48.8399	2.15524	{43.7436, 53.9362}

Exercise 10Review (25,26,27,28)

Variables/Data:

Tromp Electronics is interested in knowing what variables are associated with consumer's knowledge of a new personal computer that the company recently placed on the market. In a survey of potential purchasers of the computer, information was collected on the following variables:

0. Knowledge of the computer (y)
1. Education (x1)
2. Age (x2)
3. Knowledge of current events (x3)
4. Distance of residence from major retailing center (x4)
5. Income of household (x5).

The data consists of 46 simultaneous observations of `y, x1, x2, x3, x4, x5.`

MAthematica print-out

```
data10025={{12,33,65,11,19,76},{10,51,74,6,21,65},{15,59,86,15,40,73},{11,33,67,15,21,76},{10,35,65,19,28,68},{8,23,55,16,12,69},{7,34,59,12,33,56},{11,43,73,11,27,70},{12,43,50,17,33,60},{11,33,76,16,40,73},{10,53,68,15,24,60},{8,26,56,12,30,64},{14,56,91,4,31,80},{13,22,69,6,40,88},{9,43,68,9,30,61},{12,33,73,12,28,80},{11,39,72,13,32,69},{13,41,68,11,33,75},{6,43,55,16,24,48},{10,25,80,13,44,79},{10,43,53,5,21,62},{15,46,82,21,31,80},{10,37,66,8,26,69},{10,43,68,1,35,67},{9,23,53,4,36,70},{11,26,74,9,40,81},{7,44,39,8,23,43},{11,14,64,1,36,88},{7,37,64,15,17,60},{11,32,64,14,36,72},{9,45,72,10,22,64},{12,31,97,3,34,92},{12,36,94,6,32,85},{10,45,74,9,23,67},{11,48,73,10,42,65},{15,33,81,2,38,94},{14,54,83,9,27,77},{13,40,82,13,31,83},{8,33,68,5,19,70},{11,24,64,5,42,78},{11,36,65,19,28,68},{12,44,50,17,33,60},{13,23,69,7,41,88},{14,42,68,11,33,76},{8,37,64,15,17,60},{11,48,74,10,42,65}};
```

```
function10025=Fit[data10025,{1,x1,x2,x3,x4,x5},{x1,x2,x3,x4,x5}]
```

```
43.4326+3.04658 x1-0.678529 x2+0.421256 x3-0.299127 x4-0.184945 x5
```

```
regress10025=Regress[data10025,{1,x1,x2,x3,x4,x5},{x1,x2,x3,x4,x5}]
```

ParameterTable →	""	"Estimate"	"SE"	"TStat"	"PValue"
	1	43.4326	1.34373	32.3224	0.
	x1	3.04658	0.106423	28.627	0.
	x2	−0.678529	0.0192925	−35.1705	0.
	x3	0.421256	0.0189092	22.2779	0.
	x4	−0.299127	0.0363732	−8.22381	3.94104×10^{-10}
	x5	−0.184945	0.0263418	−7.02097	1.7536×10^{-8}

, RSquared → 0.989949, AdjustedRSquared → 0.988692, EstimatedVariance → 1.34753, ANOVATable →

""	"DF"	"SumOfSq"	"MeanSq"	"FRatio"	"PValue"
"Model"	5	5308.71	1061.74	787.917	0.
"Error"	40	53.9012	1.34753	""	""
"Total"	45	5362.61	""	""	""

```
Regress[data10025,{1,x1,x2,x3,x4,x5},{x1,x2,x3,x4,x5},RegressionReport→{SinglePredictionCITable}]
```

	"Observed"	"Predicted"	"SE"	"CI"
	76.	78.1774	1.24518	{75.6608, 80.694}
	65.	64.7878	1.233	{62.2958, 67.2798}
	73.	73.4415	1.27756	{70.8594, 76.0235}
	76.	74.407	1.21719	{71.9469, 76.867}
	68.	66.6697	1.21401	{64.2161, 69.1233}
	69.	68.3628	1.29638	{65.7427, 70.9829}
	56.	56.8501	1.23478	{54.3545, 59.3457}
	70.	70.2361	1.18012	{67.8509, 72.6212}
	60.	60.6893	1.26381	{58.1351, 63.2436}
	73.	74.3852	1.23252	{71.8942, 76.8762}
	60.	57.6562	1.21261	{55.2055, 60.107}
	64.	64.616	1.20972	{62.171, 67.0609}
	80.	79.4917	1.26919	{76.9265, 82.0568}
	88.	87.9847	1.23334	{85.492, 90.4773}
	61.	62.08	1.19696	{59.6609, 64.4992}
	80.	79.5839	1.19249	{77.1737, 81.994}
	69.	71.0059	1.17931	{68.6225, 73.3894}
	75.	74.4703	1.19467	{72.0558, 76.8848}
	48.	46.4797	1.25537	{43.9425, 49.0169}
	79.	78.6095	1.28367	{76.0151, 81.2039}
	62.	61.6688	1.25752	{59.1272, 64.2103}
	80.	80.4471	1.28109	{77.8579, 83.0362}
{SinglePredictionCITable →	69.	69.3942	1.18275	{67.0037, 71.7846}}
	67.	66.5949	1.24597	{64.0767, 69.1131}
	70.	69.7177	1.24435	{67.2028, 72.2327}
	81.	80.3863	1.20908	{77.9426, 82.8299}
	43.	44.6856	1.29578	{42.0668, 47.3045}
	88.	87.4488	1.27494	{84.8721, 90.0256}
	60.	58.9825	1.23643	{56.4836, 61.4815}
	72.	71.3467	1.19531	{68.9309, 73.7625}
	64.	63.5884	1.20509	{61.1528, 66.024}
	92.	92.6335	1.28065	{90.0452, 95.2218}
	85.	87.4496	1.24645	{84.9305, 89.9688}
	67.	67.5917	1.19831	{65.1698, 70.0136}
	65.	64.3684	1.23998	{61.8623, 66.8745}
	94.	93.2355	1.2528	{90.7035, 95.7675}
	77.	76.7228	1.23679	{74.2232, 79.2225}
	83.	80.8181	1.19946	{78.3939, 83.2423}
	70.	69.0496	1.23575	{66.5521, 71.5472}
	78.	78.3574	1.22813	{75.8752, 80.8395}
	68.	69.0377	1.2134	{66.5854, 71.4901}
	60.	60.0108	1.26578	{57.4525, 62.569}
	88.	86.8221	1.22892	{84.3383, 89.3058}
	76.	76.8384	1.21726	{74.3782, 79.2985}
	60.	62.0291	1.22186	{59.5596, 64.4986}
	65.	64.7896	1.24029	{62.2829, 67.2963}

STATISTICS II TOPICS

by Dr. Dimitris Sardelis

TOPIC 1: Test for one mean / normally distributed population with known variance

A psychologist is conducting a project in which the subjects are employees with a certain type of physical handicap. On the basis of past experience, the psychologist believes that the mean sociability score of the population of employees with this handicap is greater than 80. The population of scores is known to be approximately normally distributed, with a standard deviation of 10 . A random sample of 20 employees selected from the population yields the following results: 99, 69, 91, 97, 70, 99, 72, 74, 74, 76, 96, 97, 68, 71, 99, 78, 76, 78, 83, 66. The psychologist wants to know whether this sample result provides sufficient evidence to indicate that this belief about the population mean sociability score is correct. Let $\alpha = 0.05$ **[Daniel&Terrell, Exercise 7.3.4 , p.341].**

TOPIC 2: Test for one mean / normally distributed population with unknown variance

According to the label attached on a particular item sold at supermarkets, its net weight is 200 gr. Over a weekly period, a customer buys 9 such items, (s)he weights them and finds that the average net weight is 195 gr with a standard deviation of 9 grams. Is the customer justified to complain that these items are systematically underweighted? Assume that the weights of all such items are normally distributed and make your decision at a 0.05 level of significance**.**

TOPIC 3: Test for one mean / large sample

The consumer's average expenditures in a supermarket has been estimated to be \$280 per person with a standard deviation of \$120. Aiming to increase spending, the sales manager decides to introduce background music. On the first day, 200 customers spend \$300 on average. Is the direction justified to conclude from this that background music increases sales? Let $\alpha = 0.05$.

TOPIC 4: Test for one proportion

A college with an enrollment of approximately 10,000 students wants to build a new student parking garage.The administration believes that more than 60% of the students drive cars to school. If, in a random sample of 250 students, 165 indicate that they drive a car to school, is the administration's position supported? Let $\alpha = 0.05$ **[Daniel&Terrell, Exercise 7.8.2, p.362].**

LAB ASSIGNMENT I

(A) An athlete finds that his times for running the 100m race are normally distributed with a mean of 10.8 seconds. He trains intensively for a week and then runs 100m on each of five consecutive days. His times were: 10.70, 10.65, 10.75, 10.80, 10.60 seconds. Is the athlete justified to conclude that the training has improved his times? Let $\alpha = 0.05$. (B) 200 women are each given a sample of butter and a sample of margarine and asked to identify the butter; 120 of them do so correctly. Can women tell butter from margarine ?Let $\alpha = 0.01$.

TOPIC 5: Test for two means/ paired samples

A study is conducted to investigate how effective street lighting placed at various locations is in reducing automobile accidents in a certain town. The following table shows the median number of nighttime accidents per week at 12 locations one year before and one year after the installation of lighting.

Location	A	B	C	D	E	F	G	H	I	J	K	L
# before	8	12	5	4	6	3	4	3	2	6	6	9
# after	5	3	2	1	4	2	2	4	3	5	4	3

Do these data provide sufficient evidence to indicate that lighting does reduce nighttime automobile accidents? Let $\alpha = 0.05$ **[Daniel&Terrell, Exercise 7.4.3, p.349].**

TOPIC 6 : Test for two means of independent samples /normally distributed populations with unknown variances

McKinley Drugs is comparing the effectiveness of a new sleeping pill formula, B, with formula A, which is now on the market. For three nights, 25 subjects try Formula B, and 25 subjects in an independent sample try Formula A. The variable of interest is the average number of additional hours of sleep(compared with the nights when no drug is taken)the subjects get for the three nights. The results are as follows:

Medicine	**A**	**B**
mean ($\bar{x}$)	1.4	1.9
variance (s^2)	0.09	0.16

Do these data provide sufficient evidence to indicate that Formula B is better than Formula A? Let $\alpha = 0.01$ **[Daniel&Terrell, Exercise 7.6.3, p.357].**

TOPIC 7: Test for two means of independent large samples

The gasoline mileage for 50 type A cars is 17 miles/gallon with a standard deviation of 2.5 miles/gallon.The gasoline mileage for 70 type B cars is 18.6 miles/gallon with a standard deviation of 3.0 miles/gallon. It appears then that type B cars consume less gasoline than type A cars and should therefore be preferred. Is such a conclusion justified at the 5% level of significance?

TOPIC 8: Test for two proportions

Of 600 people who had been inoculated against common cold, 49 caught severe colds within the next three months; of 400 people who had not been inoculated, 50 caught severe colds in the same period. Do these results provide sufficient evidence for the effectiveness of inoculation against common cold? Let $\alpha = 0.001$

TOPIC 9 : Test for Independence

A government agency surveys unemployed persons who are seeking work. It prepares the following tabulation, by sex and skill level, of 532 interviews.

Skill Level	Male	Female	**Totals**
Skilled	**106** (87.16)	**6** (24.84)	**112**
Semiskilled	**93** (102.72)	**39** (29.28)	**132**

Unskilled	**215** (224.12)	**73** (63.88)	**288**
Totals	**414**	**118**	**532**

Do these data provide sufficient evidence to indicate that skill status is related to sex? Let $\alpha = 0.01$ **[Daniel&Terrell, Exercise 12.4.2, p.684].**

TOPIC 10 : Test for Homogeneity

A consultant with Nash Management Concepts wishes to know whether he can conclude that executives with a graduate degree in business administration are more innovative in the performance of their duties than executives who do not hold such a degree. A sample is selected from each of the two populations, and the performance of each executive in the two samples is analyzed by a management consulting firm. On the basis of these analyses, each executive is then categorized as either innovative or not innovative. The results are as follows.

Innovative ?	with Degree	without Degree	**Totals**
Yes	**120** (93.95)	**90** (116.05)	**210**
No	**50** (76.05)	**120** (93.95)	**170**
Totals	**170**	**210**	**380**

Can we conclude from these data that the two populations differ with respect to the proportion of innovative executives? Let $\alpha = 0.05$ **[Daniel&Terrell, Exercise 12.5.4, p.692].**

LAB ASSIGNMENT 2

(A) A market- research survey conducted by a multinational company indicates that in a randomly chosen sample of 1000 of its customers , 517 of them favor product X . A second market- research survey independently interviews a random sample of 2000 customers, of whom 983 favor product X. Decide whether the results of the two surveys differ significantly at the 5 % level of significance.

(B) Over a long period of time, a research team monitored the number of car accidents which occurred in a country. Each accident was classified as being trivial (minor damage and no personal injuries), serious (damage to vehicles and passengers, but no deaths) or fatal (damage to vehicles and loss of life).The colour of the car which, in the opinion of the research team, caused the accident was also recorded and the following data was collected:

	Trivial	**Serious**	**Fatal**
White	50	25	16
Black	35	39	18
Green	28	23	13
Red	25	17	11
Yellow	17	20	16
Blue	24	33	10

Decide at the 5 % level whether this data provides evidence of association between the colour of the car and the type of accident.

TOPIC11 : Goodness of Fit (Uniform Distribution Model)

The Table below shows the number of employees absent for a single day during a particular period of time. Test, at the 5% level, the hypothesis that the number of absentees does not depend on the day of the week.

Day	**M**	**T**	**W**	**Th**	**F**
# of absentees	121	87	87	91	114

TOPIC 12 : Goodness of Fit (Normal Distribution Model)

The following table shows the distribution of a sample of 223 employees by score on an aptitude test.The mean and variance computed from the sample data are 74.47 and 386.4252, respectively.

Score	**# of Employees**
< 40.0	10
40.0 – 49.9	12
50.0 – 59.9	17
60.0 – 69.9	37
70.0 – 79.9	55
80.0 – 89.9	51
90.0 – 99.9	34
100.0 – 109.9	5
≥ 110.0	2
Total	**223**

Test the goodness of these data to a normal distribution. Let $\alpha = 0.05$ **[Daniel&Terrell, Exercise 12.3.1, p.676].**

TOPIC 13 : Goodness of Fit (Poisson Distribution Model)

The manager of Hillmart Variety Stores, during a 300-day period, finds that a particular item is called for as follows.

# of times called for	0	1	2	3	4	5	6	7	8	9	10	11	12	13
# of days called for	3	9	21	38	46	54	49	34	20	16	6	3	1	0

Test the goodness of fit of these data to a Poisson distribution. Let $\alpha = 0.05$ **[Daniel&Terrell, Exercise 12.3.2, p.676].**

TOPIC 14: Goodness of Fit (Binomial Distribution Model)

The Table below presents some well-known data on the number of boys and girls in families of eight children obtained in Saxony in the period 1876-1885.

# of boys	# of families
0	215
1	1485
2	5331
3	10 649
4	14 959
5	11 929
6	6678
7	2092
8	342
Total	**53 680**

If the probability of a male birth is constant and is the same in all families, then the number of boys in a family of fixed size should follow the binomial distribution. If, however, the probability of a male birth varies for any reason, either from one family to another or within the same family due to such factors as parental age, then there should be more families in the extreme classes than the binomial formula predicts.Test whether or not these data follow a binomial distribution by using the value 0.5147 for the actual proportion of boys. Let $\alpha = 0.001$.

TOPIC15: Test for one variance / normally distributed population

The tensile strength of a synthetic fiber must have a variance of 5 or less before it is acceptable to a certain manufacturer. A random sample of 25 specimens taken from a new shipment gives a variance of 7. Does this provide sufficient grounds for the manufacturer to refuse shipment? Let $\alpha = 0.05$ and assume that tensile strength of the fiber is approximately normally distributed[**Daniel&Terrell, Exercise 7.10.3, p.368].**

TOPIC16: Test for two variances / normally distributed populations

A study is designed to compare two drugs for relieving tension among employees in stressful jobs. A medical team collects data on levels of tension of the subjects in two treatment groups at the end of the first two months of treatment. The variances computed from the sample data are $s_1^2 = 2916$ and $s_2^2 = 4624$. There are 8 subjects in each group. At the 0.05 level of significance, do these data provide sufficient evidence to suggest that the variability in tension levels is different in the two populations represented by the samples? State all necessary assumptions[**Daniel&Terrell, Exercise 7.11.3, p.371].**

TOPIC 17 : One- Way A N O V A

The following table shows the results, in miles per gallon, of an experiment conducted to compare three brands of gasoline. Each brand was used with seven different cars of the same weight and engine size, driven under similar conditions.

Brand A	14	19	19	16	15	17	20
Brand B	20	21	18	20	19	19	18
Brand C	20	26	23	24	23	25	23

Do these data provide sufficient evidence at the 0.01 level of significance to indicate a difference between brands of gasoline?[**Daniel&Terrell, Exercise 8.2.4, p.415**]

TOPIC 18: Two- Way A N O V A

Batches of homogenous raw material are analyzed in four laboratories for the presence of lead.Three methods are used.The following table shows the reported amounts of lead per unit volume of raw material.

Method (columns) Lab (rows)	A	B	C
I	4	5	8
II	5	5	7
III	3	6	9
IV	4	4	9

After eliminating laboratory effects,do these data suggest a difference in detection ability among the three methods?Letα=0.05[**Daniel&Terrell, Exercise 8.4.2, p.428**].

TOPIC 19: The Sign Test

Castlereagh International wants to study the effect of piped-in music on the productivity of employees.One department of a certain factory is selected at random to receive piped-in music for 30 working days.There are 10 employees in the department.The following table shows the average daily output for 30 days before the introduction of music and the average daily output for the 30 days during which music is piped into the department.

Employee	A	B	C	D	E	F	G	H	I	J
Before	90	80	92	85	81	85	72	85	70	88
During	99	85	98	83	88	99	80	91	80	94
Sign	+	+	+	-	+	+	+	+	+	+

Can we conclude from these data that music increases productivity?Letα=0.05.Determine the p-value[**Daniel&Terrell, Exercise 13.8.1, p.737**].

LAB ASSIGNMENT 3

(A) A random table of 250 digits shows the following distribution of the digits 0, 1, 2,□,9.

Digit	0	1	2	3	4	5	6	7	8	9
O_i	17	31	29	18	14	20	35	30	20	36
E_i	25	25	25	25	25	25	25	25	25	25

Using a significance level of `0.01` *, test the hypothesis that the observed distribution of digits does not differ significantly from a uniform distribution.*

(B)Researchers conducted an experiment designed to evaluate the effectiveness of four different methods-A, B, C, and D- of teaching problem solving. The following table shows by teaching method, the scores

```
A 48 38 20 16 95
B 91 37 53 91 80 38
C 67 61 33 85 99 95 81
D 57 62 50 43 59 60 70
```

made by the participating subjects (who were randomly assigned to one of the treatments) when they were forced to solve problems following the training. Do these data provide sufficient evidence to indicate that the four teaching methods differ in effectiveness? Let α = `0.05`.

TOPIC 20 : The Mann-Whitney Test

A team of industrial psychologists draws a sample of the records of those applicants for a certain job who have completed high school(G-Graduates). They select an independent random sample of the records of applicants for the same job who were high school dropouts(D-Dropouts). The following table shows the emotional maturity test scores of the applicants in the two groups.

G	89	97	69	71	67	64	86	79	56	82	72	79	78	83	62	63	96	94	62	69	57	85	67	56	77
D	85	59	85	58	66	65	64	72	78	47	49	54	74	49	65	51	57	63	41	58	67				

Do these data provide sufficient evidence to indicate that the two sampled populations have different medians? Let α = `0.05`. Determine the p-value[**Daniel&Terrell, Exercise 13.7.4, p.732].**

TOPIC 21 : The Kruskal-Wallis Test

Clarendon Mills wants to compare four new formulas for cake mix. Five cakes are baked from each of the four formulas. A panel of judges, unaware of the differences in formulas, gives each cake a score, as shown in the following table.

Formula A	72 (3)	88 (10)	70 (1)	87 (7.5)	71 (2)
Formula B	85 (5)	89 (12.5)	86 (6)	82 (4)	88 (10)
Formula C	94 (18)	94 (18)	88 (10)	87 (7.5)	89 (12.5)
Formula D	91 (14)	93 (16)	92 (15)	95 (20)	94 (18)

If the scores of the judges are not normally distributed, test the null hypothesis of no difference among cake mixes. Let α = `0.05`[**Daniel&Terrell, Exercise 13.9.2, p.742].**

TOPIC 22 : Simple Linear Regression, Correlation, Determination and Prediction

The following are the midterm and final examination scores for 21 students enrolled in an elementary statistics course.

Midterm (x)	Final (y)	x^2	y^2	x · y
98	91	9604	8281	8918
80	91	6400	8281	7280
85	91	7225	8281	7735
65	58	4225	3364	3770
85	73	7225	5329	6205
88	97	7744	9409	8536
78	85	6084	7225	6630
82	85	6724	7225	6970
95	88	9025	7744	8360
100	100	10000	10000	10000
70	79	4900	6241	5530
88	73	7744	5329	6424
82	81	6724	6561	6642
82	67	6724	4489	5494
85	88	7225	7744	7480
80	76	6400	5776	6080
78	55	6084	3025	4290
80	79	6400	6241	6320
88	91	7744	8281	8008
85	79	7225	6241	6715
75	79	5625	6241	5925
1749	**1706**	**147051**	**141308**	**143312**

(1) Define variables. **(2)** Calculate summary data. **(3)** Find variation measures. **(4)** Obtain the sample regression equation/line of best fit. **(5)** Compute and interpret Pearson's correlation coefficient and the determination coefficient. **(6)** Perform a hypothesis test to decide whether or not data provides sufficient evidence to indicate that the linear relationship between the variables involved is significant at the $\alpha = 0.05$ level. **(7)** Predict the expected value of the dependent variable for a given value of the independent variable[**Daniel&Terrell, Data from Exercise 9.5.11, p.516].**

TOPIC 23: Rank Correlation

A panel of five men and another panel of five women are asked to rank ten ideas for a new television program on the basis of their relative appeal to a general audience. The results appear in the following table.

Program idea	1	2	3	4	5	6	7	8	9	10
Men	6	4	8	7	2	1	3	5	9	10
Women	6	10	8	2	7	1	5	9	4	3
d_i	0	−6	0	5	−5	0	−2	−4	5	7
d_i^2	0	36	0	25	25	0	4	16	25	49

Test the null hypothesis that the rankings (of the ideas made by men&women) are mutually independent against the alternative that they are inversely correlated. Let $\alpha = 0.05$**[Daniel&Terrell, Exercise 13.11.2, p.751].**

LAB ASSIGNMENT 4

The following data consist of the scores which ten students obtained in an examination, their IQ's, and the numbers of hours they spent studying for the examination:

IQ	Hours studied	Exam Score
112	5	79
126	13	97
100	3	51
114	7	65
112	11	82
121	9	93
114	8	81
103	4	38
111	6	60
124	2	86

(A) Based on these data test the null hypothesis that the exam scores and the IQ's of the students taking the exam are not linearly correlated and use the regression equation of the exam scores on the IQ's to predict the exam score of a student with an IQ of 106. (B) Test the null hypothesis that the exam scores and the hours that students spend studying for the exam are not linearly correlated and use the regression equation of the exam scores on the hours studied to predict the exam score of a student who studies 6 hours. (C) Test the null hypothesis that the exam scores, the IQ's of the students taking the exam and the hours they spend studying for the exam are not linearly correlated and use the regression equation of the exam scores on the IQ's and the hours studied to predict the exam score of a student with an IQ of 106 who studies 6 hours. Let $\alpha = 0.05$ for (A), (B) and (C).

TOPIC 24: Multiple Linear Regression, Correlation, Determination and Prediction

The following table shows, for a particular week, the sales (y) of a certain product, advertising expenditures (x_1), and population density (x_2) for ten market areas. Sales and advertising expenditures are in tens of thousands of dollars,and population density is in people per square mile.

y	x_1	x_2	y^2	x_1^2	x_2^2	$y \cdot x_1$	$y \cdot x_2$	$x_1 \cdot x_2$
20	0.2	50	400	0.04	2500	4	1000	10
25	0.2	50	625	0.04	2500	5	1250	10
24	0.2	50	576	0.04	2500	4.8	1200	10
30	0.3	60	900	0.09	3600	9	1800	18
32	0.3	60	1024	0.09	3600	9.6	1920	18
40	0.4	70	1600	0.16	4900	16	2800	28
28	0.3	50	784	0.09	2500	8.4	1400	15
50	0.5	75	2500	0.25	5625	25	3750	37.5
40	0.4	70	1600	0.16	4900	16	2800	28
50	0.5	74	2500	0.25	5476	25	3700	37
339	**3.3**	**609**	**12 509**	**1.21**	**38 101**	**122.8**	**21 620**	**211.5**

(1) Define variables. **(2)** Calculate summary data. **(3)** Find variation measures. **(4)** Obtain the multiple sample regression equation/plane of best fit. **(5)** Compute and interpret the simplePearson's correlation coefficients between any two variables and the coefficient of multiple determination. **(6)** Perform a hypothesis test to decide whether or not data provides sufficient evidence to indicate that the linear relationship between the dependent variable and the independent variables involved, is significant at the $\alpha = 0.05$ level. **(7)** Predict the expected value of the dependent variable for given values of the independent variables[**Daniel&Terrell, Data from Exercise 10.3.3, p.560]**.

TOPIC 25: Forecasting

The manager of a sporting goods in northern Vermont finds that the equation which describes the trend in sales is $\hat{y} = 230 + 6.9\,x$ (origin, 1988; x units, 1 year; y, total annual sales in thousands of dollars). The **seasonal index** of sales is

Jan.	Febr.	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
171	128	83	80	78	69	64	65	67	74	104	217

Draw up a monthly sales forecast for the store for 1993**.**

The Store's Monthly Sales Forecast for the year 1993:

Month	x	$\hat{y}$ (x)[in$]	S.I (%)	(S.I) $\cdot \hat{y}$ ($)
Jan.	0	21 778.13	1.71	37 240.59
Febr.	1	21 826.04	1.28	27 937.33
March	2	21 873.96	0.83	18 155.39
April	3	21 921.88	0.80	17 537.50
May	4	21 969.79	0.78	17 136.44
June	5	22 017.71	0.69	15 192.22
July	6	22 065.63	0.64	14 122.00
August	7	22 113.54	0.65	14 373.80
Sept.	8	22 161.46	0.67	14 848.18
Oct.	9	22 209.38	0.74	16 434.94
Nov.	10	22 257.29	1.04	23 147.58
Dec.	11	22 305.21	2.10	48 402.30

STATISTICS II COMPUTER LAB ASSIGNMENT 1

INSTRUCTOR: Dr.D.A SARDELIS

LAB ASSIGNMENT IA

(A) An athlete finds that his times for running the 100m race are normally distributed with a mean of 10.8 seconds. He trains intensively for a week and then runs 100m on each of five consecutive days. His times were: 10.70, 10.65, 10.75, 10.80, 10.60 seconds. Is he justified to conclude that the training has improved his times? Let $\alpha = 0.05$.

Hypotheses

H_0: Training has not improved the athlete's times: $\mu \geq 10.80$ **seconds**

H_1: Training has improved the athlete's times: $\mu < 10.80$ **seconds**

Test Statistic

The population variance being unknown, one is constrained to use the t-distribution with d.f = 5 − 1 = 4 under the assumption that population running times are normally distributed. The test statistic is

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where $\overline{x}$ and s are the sample mean and standard deviation respectively, n is the sample size and $\mu_0 = 10.80$ seconds.

Test Statistic Value(Calculations)

First we must load the Descriptive Statistics package by writing

```
<< Statistics`DescriptiveStatistics`
```

and press SHIFT ENTER.

Then we must write the sample data as a list

```
datatimes = {10.7, 10.65, 10.75, 10.8, 10.6}
```

```
{10.7, 10.65, 10.75, 10.8, 10.6}
```

and then press SHIFT ENTER.

To find the sample size, mean and standard deviation we write (and each time press SHIFT ENTER) the operations

```
n = Length[datatimes]
```

```
5
```

```
m = Mean[datatimes]
```

```
10.7
```

```
s = StandardDeviation[datatimes]
```

```
0.0790569
```

Finally, we find t by defining the population parameter

```
μ0 = 10.8
```

```
10.8
```

if H_0 is true, write also in an input form the t-formula setting m in the place of $\overline{x}$ (remember: to activate every instruction we always press SHIFT ENTER after writing)

$$t = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}}$$

```
-2.82843
```

Critical Test Statistic Value

Given that the level of significance is $\alpha = 0.05$ and that the alternative hypothesis indicates a left-tailed test, one must find the t-critical value: $t_{0.05}^{[d.f=4]} = -t_{0.95}^{[d.f=4]}$. The traditional way is to find this from Tables. The *Mathematica* "way" is much more general and accurate to *any* approximation desired.

First load the package

```
<< Statistics`NormalDistribution`
```

Then activate the family of t-distributions by writing

```
studentt = StudentTDistribution[n - 1]
```

```
StudentTDistribution[4]
```

```
cdfstudentt = CDF[studentt, t]
```

```
0.0237103
```

which is the area under the t-distribution for all t-values up to the one found ($T < t = -2.82843$). To find the critical value $t_{0.05}^{[d.f=4]}$, one must try values greater than -2.82843 so that the corresponding cumulative probability comes as close as possible to 5%. By trial and error we find

```
CDF[studentt, -2.131846]
```

```
0.05
```

Formal Decision

The appropriate general decision rule is to reject H_0 if

$$t < t_{\alpha}^{[d.f]} = -t_{1-\alpha}^{[d.f]}$$

Since

$$-2.82843 < -2.131846$$

is true, one must indeed reject H_0.

Graphical Display of Decision

Mathematica helps us visualize the decision rule. One must first load the package

```
<< Graphics`FilledPlot`
```

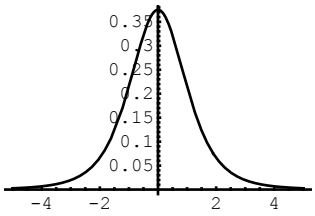
,clear the t variable from its numerical value found above

```
Clear[t]
```

,graph the t-probability distribution function over some reasonable range

```
pdfstudentt = PDF[studentt, t];
```

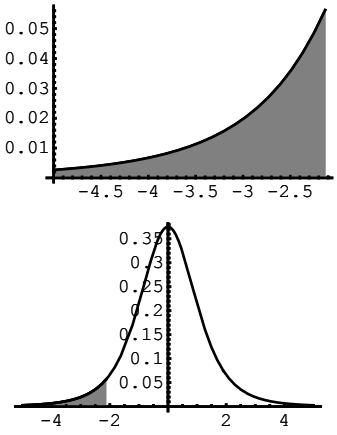
```
graphpdfstudentt = Plot[pdfstudentt, {t, -5.00, 5.00}]
```



- Graphics -

and shade the left tail up to the critical t-value found above

```
Show[FilledPlot[pdfstudentt, {t, -5.00, -2.131846}], graphpdfstudentt]
```



- Graphics -

For any t-value that happens to fall in the shaded region, one must reject the null hypothesis.

P-Value Decision

P-value is defined as the probability of the sample estimate being as extreme as it is given that the hypothesized population parameter is true. In our case, the p-value has already been found to be `0.0237103`, which as we have seen, corresponds to a t-value that falls in the shaded region. *Mathematica* provides the package

```
<< Statistics`HypothesisTests`
```

to derive the p-value directly:

```
MeanTest[datatimes, 10.80]
```

```
OneSidedPValue → 0.0237103
```

By comparing the p-value with the level of significance `0.05`, the same decision as before is reached in almost no time:

The p-value decision rule for one tailed tests is to reject H_0 if

$$\mathbf{p-value < \alpha}$$

Since

$$0.0237103 < 0.05$$

is true, one must reject H_0.

Probabilistic Conclusion

Therefore, the athlete may conclude with a ***confidence*** of at least 95% that the training has indeed improved his times. The *probability*/***risk*** that his running times have nothing to do with his training but are purely coincidental, is less than 5% . In fact this risk is identical with the p-value found, i.e., it is 2.37%.

LAB ASSIGNMENT IB

(B) 200 women are each given a sample of butter and a sample of margarine and asked to identify the butter; 120 of them do so correctly. Can women tell butter from margarine ?Let $\alpha = 0.01$.

Hypotheses

H_0: Women cannot tell butter from margarine: $\mathbf{p \le 0.5}$

H_1: Women can tell butter from margarine: $\mathbf{p > 0.5}$

Test Statistic

This is a binomial case with `n = 200` and `p = 0.5`. Since

$$n\,p_0 = n\,(1-p_0) = 200\,(0.5) = 100 \gg 5,$$

one can use the normal probability distribution with the test statistic

$$\mathbf{z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}}$$

where $\hat{p}$ `( = 120 / 200)` and p_0 `( = 0.5)` are the sample and population proportions respectively and `n` is the sample size .

Test Statistic Value(Calculations)

In order to find `z` we must set in an input form what is given setting `p` in place of $\hat{p}$

```
n = 200; p = 120 / 200; p0 = 0.5;
```

and activate the z-formula also written in an input form

$$z = \frac{p - p_0}{\sqrt{\frac{p_0 (1-p_0)}{n}}}$$

```
2.82843
```

Critical Test Statistic Value

Given that the level of significance is $\alpha = 0.01$ and that the alternative hypothesis indicates a right-tailed test, one must find the z-critical value: $z_{0.99}$. The traditional way is to find this from Tables. The *Mathematica* "way" is much more general and accurate to *any* approximation desired.

The package <<Statistics‘NormalDistribution‘ been already loaded,one must activate the standard normal distribution by writing

```
standard = NormalDistribution[0, 1]
```

```
NormalDistribution[0, 1]
```

```
cdfstandard = CDF[standard, z]
```

```
0.997661
```

which is the area under the normal probability distribution for all z-values up to the one found ($Z < z = 2.82843$). To find the critical value $z_{0.99}$, one must try values less than 2.82843 so that the corresponding cumulative probability comes as close as possible to 99%. By trial and error we find

```
CDF[standard, 2.32636]
```

```
0.99
```

Formal Decision

The appropriate general decision rule is to reject H_0 if

$$z > z_{1-\alpha}$$

Since

$$2.82843 > 2.32636$$

is true, one must indeed reject H_0.

Graphical Display of Decision

Mathematica helps us visualize the decision rule. The package <<Graphics‘FilledPlot‘ been already loaded

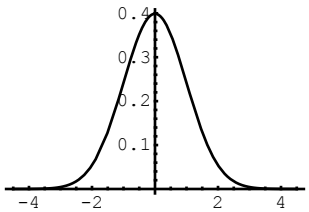
,one must clear the z variable from its numerical value found above

```
Clear[z]
```

,graph the standard normal probability distribution function over some reasonable range

```
pdfstandard = PDF[standard, z];
```

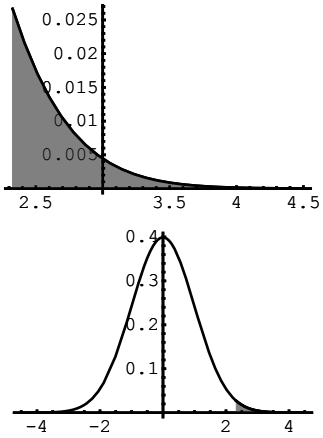
```
graphpdfstandard = Plot[pdfstandard, {z, -4.5, 4.5}]
```



- Graphics -

and shade the right tail above the critical z-value found above

```
Show[FilledPlot[pdfstandard, {z, 2.32636, 4.50}], graphpdfstandard]
```



- Graphics -

For any z-value that happens to fall in the shaded region, one must reject the null hypothesis.

P-Value Decision

P-value is defined as the probability of the sample estimate being as extreme as it is given that the hypothesized population parameter is true. In our case, the complement of the p-value has been found to be 0.997661, so the p-value is 0.00233887 and corresponds to a z-value that falls in the shaded region. *Mathematica* provides the package <<Statistics‘HypothesisTests‘

to derive the p-value directly: COPY PASTE the test stat value found above into the command

```
NormalPValue[2.8284271247461894`]
```

```
OneSidedPValue → 0.00233887
```

By comparing the p-value with the level of significance 0.01, the same decision as before is again reached in almost no time.

The p-value decision rule for one tailed tests is to reject H_0 if

$$\mathbf{p-value < \alpha}$$

Since

$$0.00233887 < 0.01$$

is true, one must reject H_0.

Probabilistic Conclusion

Therefore, we may conclude with a ***confidence*** of at least 99% that women can indeed tell the difference between butter and margarine. The *probability*/***risk*** that sample evidence is purely circumstantial/ coincidental, is less than 1% . In fact this risk is identical with the p-value found, i.e., it is 0.23%.

STATISTICS II COMPUTER LAB ASSIGNMENT 2

INSTRUCTOR: Dr.D.A SARDELIS

LAB ASSIGNMENT 2A

(A) A market- research survey conducted by a multinational company indicates that in a randomly chosen sample of 1000 of its customers , 517 of them favor product X . A second market- research survey independently interviews a random sample of 2000 customers, of whom 983 favor product X. Decide whether the results of the two surveys differ significantly at the 5 % level of significance.

Hypotheses

H_0: The two surveys do not differ significantly in their results: $\mathbf{p_1 - p_2 = 0}$

H_1: The two surveys differ significantly in their results: $\mathbf{p_1 - p_2 \neq 0}$

Test Statistic

The surveys refer to two large independent samples. Consequently, the sampling distribution may be considered to be normally distributed with the test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\bar{p}\,(1-\bar{p})}{n_1} + \frac{\bar{p}\,(1-\bar{p})}{n_2}}}$$

where $\bar{p}$ is the average sample proportion

$$\bar{p} = \frac{n_1\,\hat{p}_1 + n_2\,\hat{p}_2}{n_1 + n_2}$$

with n_1, n_2 the sample sizes of the two surveys.

Test Statistic Value(Calculations)

To calculate z we write the data and the z-formula in an input form and activate the corresponding cells by pressing SHIFT ENTER right after:

```
n1 = 1000; n2 = 2000; p̂1 = 517/1000; p̂2 = 983/2000; p̄ = (n1 p̂1 + n2 p̂2)/(n1 + n2);
z = (p̂1 - p̂2 - 0)/Sqrt(p̄ (1-p̄)/n1 + p̄ (1-p̄)/n2)
```

$$\frac{17\sqrt{\frac{3}{5}}}{10}$$

The decimal expression for z is found by the command

```
N[z]
```

```
1.31681
```

Critical Test Statistic Value

Given that the level of significance is $\alpha = 0.05$ and that the alternative hypothesis indicates a two-tailed test, one must find the z-critical values: $z_{0.0250} = -z_{0.9750}$. The traditional way is to find this from Tables. The *Mathematica* "way" is much more general and accurate to *any* approximation desired.

First load the package

```
<< Statistics`NormalDistribution`
```

Then one must activate the standard normal distribution by writing

```
standard = NormalDistribution[0, 1]
```

```
NormalDistribution[0, 1]
```

```
cdfstandard = CDF[standard, N[z]]
```

```
0.90605
```

which is the area under the normal probability distribution for all z-values up to the one found ($Z < z = 1.31681$). To find the critical value $z_{0.975}$, one must try values larger than 1.31681 so that the corresponding cumulative probability comes as close as possible to 97.5%. By trial and error we find

```
CDF[standard, 1.95997]
```

```
0.975
```

Formal Decision

The appropriate general decision rule is reject H_0 if

$$\text{either } z < z_{\alpha/2} \text{ or } z > z_{1-\alpha/2}$$

and since

$$1.31681 > 1.95997$$

is false, H_0 cannot be rejected .

Graphical Display of Decision

Mathematica helps us visualize the decision rule. First load the package <<Graphics'FilledPlot'

```
<< Graphics`FilledPlot`
```

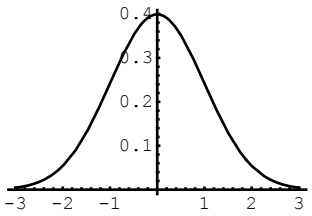
, clear the z variable from its numerical value found above

```
Clear[z]
```

then graph the standard normal probability distribution function over some reasonable range:

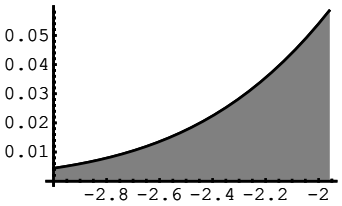
```
pdfstandard = PDF[standard, z];
```

```
An = Plot[pdfstandard, {z, -3.00, 3.00}]
```



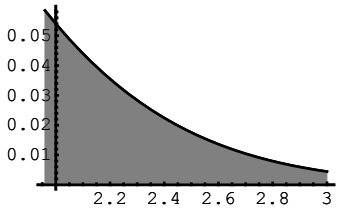
- Graphics -

```
Ln = FilledPlot[pdfstandard, {z, -3.00, -1.95997}]
```



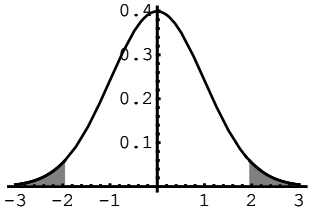
- Graphics -

```
Rn = FilledPlot[pdfstandard, {z, 1.95997, 3.00}]
```



- Graphics -

```
Show[Ln, An, Rn]
```



- Graphics -

For any z-value that happens to fall in the shaded region, one must reject the null hypothesis.

P-Value Decision

For two-tailed tests the P-value is defined as twice the probability of the sample estimate being as extreme as it is given that the hypothesized population parameter is true. In our case, the complement of half the p-value has been found to be 0.90605, so the p-value is 2(1-0.90605)= 0.1879 and corresponds to a z-value that falls in the unshaded region. *Mathematica* provides the package

```
<< Statistics`HypothesisTests`
```

to derive the p-value directly: COPY PASTE the test stat value found above into the command

```
NormalPValue[1.3168143377105217`]
```

```
OneSidedPValue → 0.0939504
```

and double the result

```
2 (0.0939504328384918`)
```

```
0.187901
```

The p-value is greater than the level of significance 0.05 and the same decision as before is again reached in almost no time.

Probabilistic Conclusion

Therefore, we may conclude with a ***confidence*** of 95% that the two surveys do not differ significantly in their results. The *probability*/***risk*** that sample evidence is purely circumstantial/ coincidental, is more than 5% . In fact this risk is identical with the p-value found, i.e., it is 18.79%.

LAB ASSIGNMENT 2B

(B) Over a long period of time, a research team monitored the number of car accidents which occurred in a country. Each accident was classified as being trivial (minor damage and no personal injuries), serious (damage to vehicles and passengers, but no deaths) or fatal (damage to vehicles and loss of life).The color of the car which, in the opinion of the research team, caused the accident was also recorded and the following data was collected:

	Trivial	Serious	Fatal
White	50	25	16
Black	35	39	18
Green	28	23	13
Red	25	17	11
Yellow	17	20	16
Blue	24	33	10

Decide at the 5 % level whether this data provides evidence of association between the color of the car and the type of accident.

Hypotheses

H_0: There is no association between the color of the car and the type of accident.

H_1: There is association between the color of the car and the type of accident.

Test Statistic

The results of the survey are organized in a 6×3 contingency table with the row-factor indicating 6 car colors and the column factor indicating 3 accident degradations ($r = 6$, $c = 3$). The appropriate test stat is chi square with $(6 - 1) \times (3 - 1) = 10$ degrees of freedom

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where O_{ij} and E_{ij} denote the observed and expected frequencies in entry (i, j) respectively. The expected frequencies E_{ij} are derived from the null hypothesis via the rule

$$E_{ij} = \frac{\left(\sum_{j=1}^{c} O_{ij}\right)\left(\sum_{i=1}^{r} O_{ij}\right)}{\left(\sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij}\right)}$$

Test Statistic Value(Calculations)

To calculate χ^2 one must first write the observed frequencies in an input form and activate the cell by pressing SHIFT ENTER right after:

```
O11 = 50; O12 = 25; O13 = 16; O21 = 35; O22 = 39; O23 = 18; O31 = 28; O32 = 23; O33 = 13;
O41 = 25; O42 = 17; O43 = 11; O51 = 17; O52 = 20; O53 = 16; O61 = 24; O62 = 33; O63 = 10;
```

Next let us define the six row totals:

```
TR1 = O11 + O12 + O13; TR2 = O21 + O22 + O23; TR3 = O31 + O32 + O33;
TR4 = O41 + O42 + O43; TR5 = O51 + O52 + O53; TR6 = O61 + O62 + O63;
```

and the three column totals

```
TC1 = O11 + O21 + O31 + O41 + O51 + O61; TC2 = O12 + O22 + O32 + O42 + O52 + O62; TC3 = O13 + O23 + O33 + O43 + O53 + O63;
```

Both the sum of the row totals and the sum of the column totals must give the total number of the elements in the survey, henceforth denoted as "Tot":

```
TR1 + TR2 + TR3 +
 TR4 + TR5 + TR6
```

```
420
```

```
TC1 + TC2 + TC3
```

```
420
```

```
Tot = 420
```

```
420
```

We may now find the expected frequencies

```
E11 = TR1 * TC1 / Tot; E12 = TR1 * TC2 / Tot; E13 = TR1 * TC3 / Tot; E21 = TR2 * TC1 / Tot;
E22 = TR2 * TC2 / Tot; E23 = TR2 * TC3 / Tot; E31 = TR3 * TC1 / Tot; E32 = TR3 * TC2 / Tot;
E33 = TR3 * TC3 / Tot; E41 = TR4 * TC1 / Tot; E42 = TR4 * TC2 / Tot; E43 = TR4 * TC3 / Tot; E51 = TR5 * TC1 / Tot;
E52 = TR5 * TC2 / Tot; E53 = TR5 * TC3 / Tot; E61 = TR6 * TC1 / Tot; E62 = TR6 * TC2 / Tot; E63 = TR6 * TC3 / Tot;
```

Finally, we write in an input form the appropriate chi square test stat sum

$$\text{chi} = \frac{(O11-E11)^2}{E11} + \frac{(O12-E12)^2}{E12} + \frac{(O13-E13)^2}{E13} + \frac{(O21-E21)^2}{E21} + \frac{(O22-E22)^2}{E22} + \frac{(O23-E23)^2}{E23} + \frac{(O31-E31)^2}{E31} + \frac{(O32-E32)^2}{E32} + \frac{(O33-E33)^2}{E33} + \frac{(O41-E41)^2}{E41} + \frac{(O42-E42)^2}{E42} + \frac{(O43-E43)^2}{E43} + \frac{(O51-E51)^2}{E51} + \frac{(O52-E52)^2}{E52} + \frac{(O53-E53)^2}{E53} + \frac{(O61-E61)^2}{E61} + \frac{(O62-E62)^2}{E62} + \frac{(O63-E63)^2}{E63}$$

$$\frac{213\,334\,479\,081\,445}{13\,367\,572\,801\,856}$$

The decimal expression for χ^2 is found by the command

```
N[chi]
```

```
15.9591
```

Critical Test Statistic Value

Given that the level of significance is $\alpha = 0.05$ and that test is right tailed, one must find the χ^2-critical value: $\chi^2_{0.95\ (d.f\ =\ 10)}$. The traditional way is to find this from Tables. The *Mathematica* "way" is much more general and accurate to *any* approximation desired.

Having already loaded the package <<Statistics‘NormalDistribution‘, one must activate the family of the χ^2 distributions

```
chisq = ChiSquareDistribution[n - 1]
```

```
ChiSquareDistribution[-1 + n]
```

and,in particular , the one with 10 degrees of freedom

```
chisq10 = ChiSquareDistribution[11 - 1]
```

```
ChiSquareDistribution[10]
```

and evaluate the area under the curve up to the test stat value found by the command

```
CDF[chisq10, N[chi]]
```

```
0.899191
```

which is the area under the normal probability distribution for all χ^2-values up to the one found ($X^2 < \chi^2 = 15.9591$). To find the critical value $\chi^2_{0.95}$, one must try values larger than 15.9591 so that the corresponding cumulative probability comes as close as possible to 95%. By trial and error we find

```
CDF[chisq10, 18.30705]
```

```
0.95
```

Formal Decision

The appropriate general decision rule is reject H_0 if

$$\chi^2 > \chi^2_{1-\alpha}$$

and since

$$15.9591 > 18.30705$$

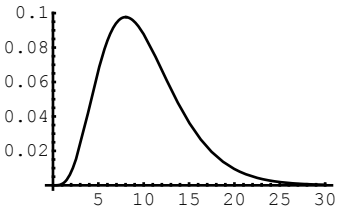
is false, H_0 cannot be rejected .

Graphical Display of Decision

Mathematica helps us visualize the decision rule. The package <<Graphics'FilledPlot' being already loaded, graph the chi square probability distribution function over some reasonable range by the following commands:

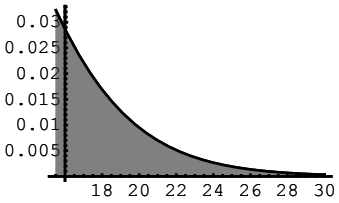
```
pdfchisq = PDF[chisq10, x];
```

```
Ax = Plot[pdfchisq, {x, 0.00, 30.00}]
```



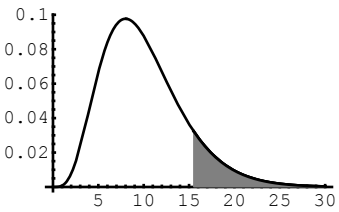

```
- Graphics -
```

```
Rx = FilledPlot[pdfchisq, {x, 15.507, 30.00}]
```




```
- Graphics -
```

```
Show[Ax, Rx]
```




```
- Graphics -
```

For any χ^2-value that happens to fall in the shaded region, one must reject the null hypothesis.

P-Value Decision

The P-value is defined as the probability of the sample estimate being as extreme as it is given that the hypothesized population parameter is true. In our case, the complement of the p-value has been found to be 0.899191, so the p-value is `0.100809` and corresponds to a χ^2-value that falls in the unshaded region. *Mathematica* provides the package **<< Statistics`HypothesisTests`** to derive the p-value directly by introducing the test stat value found above into the command

```
ChiSquarePValue[15.959103589233859`, 10]
```

```
OneSidedPValue → 0.100809
```

The p-value is greater than the level of significance `0.05` and the same decision as before is again reached in almost no time.

Probabilistic Conclusion

Therefore, we may conclude with a ***confidence*** of 95% that there is no association between the color of the car and the type of accident. The *probability*/***risk*** that sample evidence is purely circumstantial/ coincidental, is more than 5% . In fact this risk is identical with the p-value found, i.e., it is 10.08%.

STATISTICS II COMPUTER LAB ASSIGNMENT 3

INSTRUCTOR: Dr.D.A SARDELIS

LAB ASSIGNMENT 3A

A random table of 250 digits shows the following distribution of the digits 0, 1, 2,□,9.

Digit	0	1	2	3	4	5	6	7	8	9
O_i	17	31	29	18	14	20	35	30	20	36
E_i	25	25	25	25	25	25	25	25	25	25

Using a significance level of `0.01` , test the hypothesis that the observed distribution of digits does not differ significantly from a uniform distribution.

Hypotheses

H_0:The observed distribution of digits does not differ significantly from a uniform distribution.

H_1: The observed distribution of digits differs significantly from a uniform distribution.

Test Statistic

The discrepancy between the observed and expected frequencies is measured by the chi square defined as

$$\chi^2 = \sum_{i=0}^{9} \frac{(O_i - E_i)^2}{E_i}$$

Test Statistic Value (Calculations)

To calculate χ^2 one must write the observed and expected frequencies in an input form and activate the cell by pressing SHIFT ENTER right after:

```
O0 = 17; O1 = 31; O2 = 29; O3 = 18; O4 = 14; O5 = 20; O6 = 35; O7 = 30; O8 = 20; O9 = 36;
E0 = 25; E1 = 25; E2 = 25; E3 = 25; E4 = 25; E5 = 25; E6 = 25; E7 = 25; E8 = 25; E9 = 25;
```

Then we write also in an input form the chi square test stat sum and press SHIFT ENTER

$$\text{chi} = \frac{(O0 - E0)^2}{E0} + \frac{(O1 - E1)^2}{E1} + \frac{(O2 - E2)^2}{E2} + \frac{(O3 - E3)^2}{E3} + \frac{(O4 - E4)^2}{E4} + \frac{(O5 - E5)^2}{E5} + \frac{(O6 - E6)^2}{E6} + \frac{(O7 - E7)^2}{E7} + \frac{(O8 - E8)^2}{E8} + \frac{(O9 - E9)^2}{E9}$$

$$\frac{582}{25}$$

The decimal expression for χ^2 is found by the command

```
N[chi]
```

```
23.28
```

Critical Test Statistic Value

Given that the level of significance is $\alpha = 0.01$ and that test is right tailed, one must find the χ^2-critical value: $\chi^2_{0.99}\ (d.f = 9)$. The traditional way is to find this from Tables. The *Mathematica* "way" is much more general and accurate to *any* approximation desired.

One must first load the package

```
<< Statistics`NormalDistribution`
```

, activate the family of the χ^2 distributions

```
chisq = ChiSquareDistribution[n - 1]
```

```
ChiSquareDistribution[-1 + n]
```

and,in particular , the one with 9 degrees of freedom

```
chisq09 = ChiSquareDistribution[10 - 1]
```

```
ChiSquareDistribution[9]
```

and evaluate the area under the curve up to the test stat value found by the command

```
CDF[chisq09, N[chi]]
```

```
0.994403
```

which is the area under the normal probability distribution for all χ^2-values up to the one found ($X^2 < \chi^2 = 23.28$). To find the critical value $\chi^2_{0.99}$, one must try values smaller than 23.28 so that the corresponding cumulative probability comes as close as possible to 99%. By trial and error we find

```
CDF[chisq09, 21.666]
```

```
0.99
```

Formal Decision

The appropriate general decision rule is reject H_0 if

$$\chi^2 > \chi^2_{1-\alpha}$$

and since

$$23.28 > 21.666$$

is true, H_0must be rejected .

Graphical Display of Decision

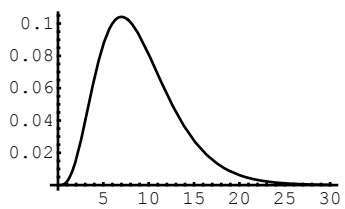
Mathematica helps us visualize the decision rule. First we load the package

```
<< Graphics`FilledPlot`
```

and graph the chi square probability distribution function over some reasonable range by the following commands:

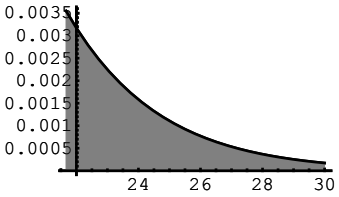
```
pdfchisq = PDF[chisq09, x];
```

```
Ax = Plot[pdfchisq, {x, 0.00, 30.00}]
```



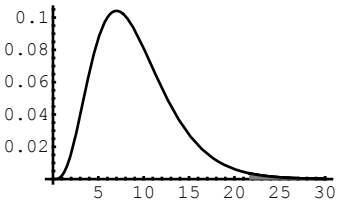

```
- Graphics -
```

```
Rx = FilledPlot[pdfchisq, {x, 21.666, 30.00}]
```




```
- Graphics -
```

```
Show[Ax, Rx]
```




```
- Graphics -
```

For any χ^2-value that happens to fall in the shaded region, one must reject the null hypothesis.

P-Value Decision

The P-value is defined as the probability of the sample estimate being as extreme as it is given that the hypothesized population parameter is true. In our case, the complement of the p-value has been found to be 0.994403, so the p-value is 0.005597 and corresponds to a χ^2-value that falls in the shaded region. *Mathematica* provides the package

```
<< Statistics`HypothesisTests`
```

to derive the p-value directly by introducing the test stat value found above into the command

```
ChiSquarePValue[23.28`, 9]
```

```
OneSidedPValue → 0.00559718
```

Thus the p-value is less than the level of significance `0.01` and the same decision as before is again reached in almost no time.

Probabilistic Conclusion

Therefore, we may conclude with a ***confidence*** of 99% that the observed distribution of digits differs significantly from a uniform distribution. The *probability*/***risk*** that sample evidence is purely circumstantial/ coincidental, is less than 1% . In fact this risk is identical with the p-value found, i.e., it is 0.56%.

LAB ASSIGNMENT 3B

Researchers conducted an experiment designed to evaluate the effectiveness of four different methods-A, B, C, and D- of teaching problem solving. The following table shows by teaching method, the scores

A	48	38	20	16	95		
B	91	37	53	91	80	38	
C	67	61	33	85	99	95	81
D	57	62	50	43	59	60	70

made by the participating subjects (who were randomly assigned to one of the treatments) when they were forced to solve problems following the training. Do these data provide sufficient evidence to indicate that the four teaching methods differ in effectiveness? Let $\alpha = 0.05$.

Hypotheses

H_0: The four teaching methods are equally effective, i.e the respective mean population scores are all equal: $\mu_A = \mu_B = \mu_C = \mu_D$

H_1: The four teaching methods are not equally effective, i.e not all mean population scores are equal

Test Statistic

Summary Data

The four sets of observed data constitute four independent random samples from the specified populations of scores. To get all summary data essentials we must load the package

```
<<Statistics`DataManipulation`
```

and introduce the four samples in an input form

```
sample1 = {48, 38, 20, 16, 95}; sample2 = {91, 37, 53, 91, 80, 38};
sample3 = {67, 61, 33, 85, 99, 95, 81}; sample4 = {57, 62, 50, 43, 59, 60, 70};
```

The sample size and the sum/total of sample elements are then found for each sample by the commands

```
n1 = Length[sample1]; T1 = Apply[Plus, sample1]; n2 = Length[sample2]; T2 = Apply[Plus, sample2];
n3 = Length[sample3]; T3 = Apply[Plus, sample3]; n4 = Length[sample4]; T4 = Apply[Plus, sample4];
```

To find the sums of sample elements squared, one must first define the function

```
f[x_] := x^2
```

and apply for each sample the commands

```
S1 = Apply[Plus, Map[f, sample1]]; S2 = Apply[Plus, Map[f, sample2]];
S3 = Apply[Plus, Map[f, sample3]]; S4 = Apply[Plus, Map[f, sample4]];
```

We may display our findings in a table form as follows:

```
TableForm[{{n1, T1, S1}, {n2, T2, S2}, {n3, T3, S3}, {n4, T4, S4}}]
```

```
5    217    13 429
6    390    28 584
7    521    41 911
7    401    23 423
```

All the elements of these data are

```
n = n1 + n2 + n3 + n4
```

```
25
```

,their sum/grand total is

```
T = T1 + T2 + T3 + T4
```

```
1529
```

and the sum of all the elements squared is

```
S = S1 + S2 + S3 + S4
```

```
107 347
```

Sums of Squared Differences

To reach a decision for the case of many sample means, the following sums have to be found

$$\mathbf{SSTr} = \sum_{j=1}^{k} \left(\frac{T_j^2}{n_j} \right) - \frac{T^2}{N}$$

$$\mathbf{SSE} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{k} \left(\frac{T_j^2}{n_j} \right);$$

$$\mathbf{SST} = \mathbf{SSTr} + \mathbf{SSE} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2 - \frac{T^2}{N}$$

where

k : number of samples / treatments; j = 1, 2, 3, □ , k;

n_j : size of sample (j); i = 1, 2, 3, □ , n_j; $N = \sum_{j=1}^{k} n_j$

sample totals : $T_j = \sum_{i=1}^{n_j} x_{ij}$; grand total : $T = \sum_{j=1}^{k} T_j$;

sums of squares : $\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2$;

SSTr expressed in terms of the summary data and written in an input form, is

```
SSTr = N[(T1^2/n1) + (T2^2/n2) + (T3^2/n3) + (T4^2/n4) - (T^2/n)]
```

```
3003.02
```

Similarly, SSE expressed in terms of the summary data and written in an input form, is

```
SSE = N[S - (T1^2/n1) - (T2^2/n2) - (T3^2/n3) - (T4^2/n4)]
```

```
10 830.3
```

Finally, SST expressed in terms of the summary data and written in an input form, is

```
SST = N[S - (T^2/n)]
```

```
13 833.4
```

and the above sums are found to satisfy the sum rule: **SST = SSTr + SSE**.

```
SSTr + SSE
```

```
13 833.4
```

Test Statistic Value

The appropriate test stat is the F distribution with numerator degrees of freedom: k - 1 = 4 - 1 = 3 and denominator degrees of freedom: n - k = 25 - 4 = 21. The F-formula is

```
k = 4; n = 25; F = (SSTr/(k - 1))/(n - k)/SSE
```

```
1.94095
```

Critical Test Statistic Value

Given that the level of significance is α = 0.05 and that test is right tailed, one must find the F-critical value: $F_{0.95}$ (d.f num = 3, d.f den = 21) . The traditional way is to find this from Tables. The *Mathematica* "way" is much more general and accurate to *any* approximation desired.

The package <<Statistics‘NormalDistribution‘ being already loaded, the family of the F distributions must be activated

```
fratio = FRatioDistribution[3, 21]
```

```
FRatioDistribution[3, 21]
```

and evaluate the area under the curve up to the test stat value found by the command

```
CDF[fratio, F]
```

```
0.846074
```

which is the area under the normal probability distribution for all F-values up to the one found . To find the critical value $F_{0.95}$, one must try values larger than 1.94095 so that the corresponding cumulative probability comes as close as possible to 95%. By trial and error we find

```
CDF[fratio, 3.07246]
```

```
0.95
```

Formal Decision

The appropriate general decision rule is reject H_0 if

$$F > F_{1-\alpha}$$

and since

$$1.94095 > 3.07246$$

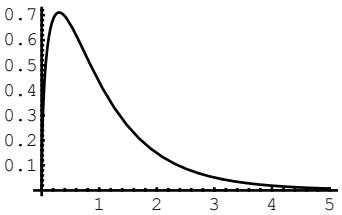
is false, H_0must not be rejected .

Graphical Display of Decision

Mathematica helps us visualize the decision rule. Having the package <<Graphics‘FilledPlot‘ already loaded, one may graph the F-probability distribution function over some reasonable range by the following commands:

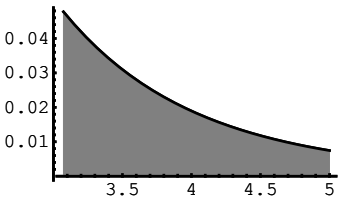
```
pdfratio = PDF[fratio, x];
```

```
AAx = Plot[pdfratio, {x, 0.00, 5.00}]
```



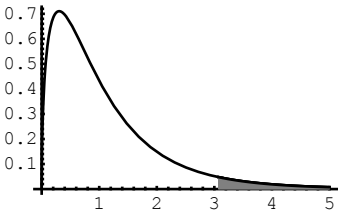
- Graphics -

```
RRx = FilledPlot[pdfratio, {x, 3.07246, 5.00}]
```



- Graphics -

```
Show[AAx, RRx]
```



- Graphics -

For any F-value that happens to fall in the shaded region, one must reject the null hypothesis.

```
1 - 0.8460738846820769`
```

```
0.153926
```

P-Value Decision

The P-value is defined as the probability of the sample estimate being as extreme as it is given that the hypothesized population parameter is true. In our case, the complement of the p-value has been found to be 0.846074 , so the p-value is 0.153926 and corresponds to a F-value that falls in the unshaded region. *Mathematica* provides the package <<Statistics‘HypothesisTests‘

to derive the p-value directly by introducing the test stat value found above into the command

```
FRatioPValue[1.94095, 3, 21]
```

```
OneSidedPValue → 0.153926
```

Thus the p-value exceeds the level of significance 0.05 and the same decision as before is again reached in almost no time.

Probabilistic Conclusion

Therefore, we may conclude with a ***confidence*** of 95% that the four teaching methods are equally effective. The *probability*/***risk*** that sample evidence is purely circumstantial/ coincidental, exceeds 5% . In fact this risk is identical with the p-value found, i.e., it is 15.39%.

STATISTICS II COMPUTER LAB ASSIGNMENT 4

INSTRUCTOR: Dr.D.A SARDELIS

The following data consist of the scores which ten students obtained in an examination, their IQ's, and the numbers of hours they spent studying for the examination:

IQ	Hours studied	Exam Score
112	5	79
126	13	97
100	3	51
114	7	65
112	11	82
121	9	93
114	8	81
103	4	38
111	6	60
124	2	86

(A) Based on these data test the null hypothesis that the exam scores and the IQ's of the students taking the exam are not linearly correlated and use the regression equation of the exam scores on the IQ's to predict the exam score of a student with an IQ of 106. (B) Test the null hypothesis that the exam scores and the hours that students spend studying for the exam are not linearly correlated and use the regression equation of the exam scores on the hours studied to predict the exam score of a student who studies 6 hours. (C) Test the null hypothesis that the exam scores, the IQ's of the students taking the exam and the hours they spend studying for the exam are not linearly correlated and use the regression equation of the exam scores on the IQ's and the hours studied to predict the exam score of a student with an IQ of 106 who studies 6 hours. Let $\alpha = 0.05$ *for (A), (B) and (C).*

Variables

Dependent Variable: exam scores (**y**)
Independent Variables: IQ (**x1**), the number of hours of study for the exam (**x2**)

Data

The data consists of 10 simultaneous observations of **x1, x2, y**. Let us open an input cell and introduce data as a list of three dimensional points in the order : x1, x2, y.

```
data = {{112, 5, 79}, {126, 13, 97}, {100, 3, 51}, {114, 7, 65},
    {112, 11, 82}, {121, 9, 93}, {110, 8, 81}, {103, 4, 38}, {111, 6, 60}, {124, 2, 86}};
```

Check that all ten data points have been introduced by the activation of the command

```
Length[data]
```

```
10
```

These multiple data x1 – x2 – y can be decomposed into two simple data components: (A) one component displaying the x1 – y observations only and (B) a component displaying the x2 – y observations only.

Data A can be extracted by writing and activating the command:

```
dataA01 = data[[All, {1, 3}]]
```

```
{{112, 79}, {126, 97}, {100, 51}, {114, 65}, {112, 82}, {121, 93}, {110, 81}, {103, 38}, {111, 60}, {124, 86}}
```

Likewise, data B can be extracted by writing and activating the command:

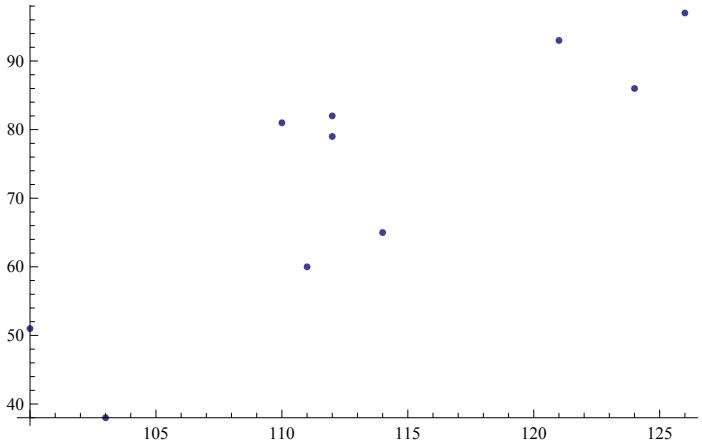
```
dataB02 = data[[All, {2, 3}]]
```

```
{{5, 79}, {13, 97}, {3, 51}, {7, 65}, {11, 82}, {9, 93}, {8, 81}, {4, 38}, {6, 60}, {2, 86}}
```

Scatter Diagrams

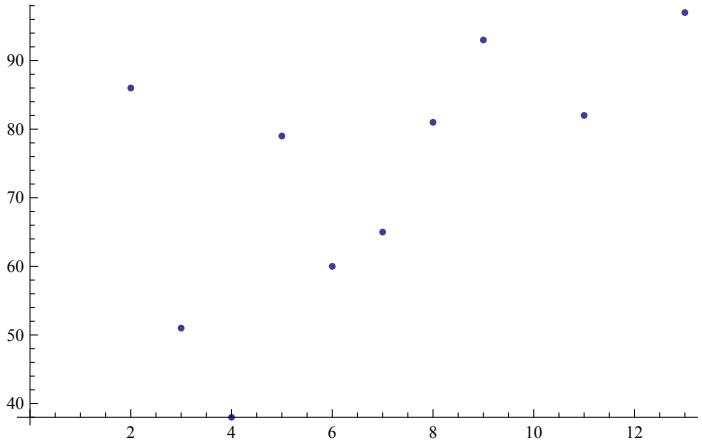
To visually display data (A), we write

```
plotdataA01 = ListPlot[dataA01, PlotStyle → PointSize[0.01]]
```



Similarly, to visually display data (B), we write

```
plotdataB02 = ListPlot[dataB02, PlotStyle → PointSize[0.01]]
```



Both scatter diagrams indicate a positive/upward trend for the `x1 – y`, `x2 – y` relations.

(A) Exam Scores versus IQ

We first load the linear regression package of Mathematica

```
<< Statistics`LinearRegression`
```

ANOVA for Simple Linear Regression

We would like to test the *null hypothesis* that exam scores and the IQ's of students taking the exam are not linearly correlated against the *alternative hypothesis* that exam scores and the IQ's of students taking the exam are linearly correlated. Also, we would like to find the linear regression equation that best fits data(A) and use it for prediction.

A full report of the regression characteristics of the `x1 – y` relation, can be given with the activation of a single command

```
regrAO1 = Regress[dataA01, {1, x1}, {x1}]
```

```
                        Estimate    SE          TStat       PValue
{ParameterTable → 1     -145.09     47.6639     -3.04403    0.0159652 ,
                  x1    1.92666     0.419646    4.59115     0.00177588

 RSquared → 0.724884, AdjustedRSquared → 0.690495, EstimatedVariance → 112.371,

                        DF    SumOfSq    MeanSq     FRatio     PValue
 ANOVATable → Model     1     2368.63    2368.63    21.0787    0.00177588}
              Error     8     898.968    112.371
              Total     9     3267.6
```

The F-ratio in the ANOVA Table above corresponds to a p-value of `0.00177588` which is much less than $\alpha = 0.05$. Therefore, one must reject the null hypothesis, i.e., conclude that the exam scores and the IQ's of students are significantly linearly correlated.

Simple Linear Regression Equation

The first column entries in the parameter table of the full regression report are the constants of the linear regression function. This function can be formed by using COPY PASTE for each of these entries and writing

```
functionAO1 = (-145.0902679830736`) + (1.9266572637517403`) x1
```

```
-145.09 + 1.92666 x1
```

Prediction

For a student with an IQ of 106, one expects his/her exam performance to be(COPY PASTE the function above and set `x1 = 106`)

```
(-145.0902679830736`) + (1.9266572637517403`) * 106
```

```
59.1354
```

on the average.

Model Evaluation

The goodness of the model is expressed by the coefficient of determination (denoted in the output as RSquared) which as we see is `0.724884` or `72.49 %`.

(B) Exam Scores versus Hours of Study

Having loaded the StatisticsLinearRegression package of *Mathematica* we proceed as before:

ANOVA for Simple Linear Regression

We would like to test the null hypothesis that exam scores and the hours of study students spend preparing the exam are not linearly correlated against the alternative hypothesis that exam scores and the hours of study students spend preparing the exam are not linearly correlated. Also, we would like to find the linear regression equation that best fits data(B) and use it for prediction.

A full report of the regression characteristics of the `x2 - y` relation, can be given with the activation of a single command

```
regrBO2 = Regress[dataB02, {1, x2}, {x2}]
```

```
                        Estimate    SE          TStat       PValue
{ParameterTable → 1     51.4229     11.6831     4.4015      0.00228218,
                  x2    3.20251     1.54206     2.07678     0.0714722

 RSquared → 0.350281, AdjustedRSquared → 0.269066, EstimatedVariance → 265.378,

                        DF    SumOfSq    MeanSq     FRatio     PValue
 ANOVATable → Model     1     1144.58    1144.58    4.31301    0.0714722}
              Error     8     2123.02    265.378
              Total     9     3267.6
```

The F-ratio in the ANOVA Table above corresponds to a p-value of `0.0714722` which is larger than $\alpha = 0.05$. Therefore, one must not reject the null hypothesis, i.e., there is no evidence to conclude that the exam scores and the hours of study students spend preparing the exam are linearly correlated.

Simple Linear Regression Equation

The first column entries in the parameter table of the full regression report are the constants of the linear regression function. This function can be formed by using COPY PASTE for each of these entries and writing

```
functionBO2 = (51.42293906810043`) + (3.2025089605734727`) x2
```

```
51.4229+ 3.20251 x2
```

Prediction

For a student with 6 hours of study, one expects his/her exam performance to be(COPY PASTE the function above and set x2 = 6)

```
(51.42293906810043`) + (3.2025089605734727`) 6
```

```
70.638
```

on the average.

Model Evaluation

The goodness of the model is expressed by the coefficient of determination (denoted in the output as RSquared) which as we see is 0.350281 or 35.03 % .

(C) Exam Scores versus IQ and Hours of Study

Having loaded the StatisticsLinearRegression package of *Mathematica* we proceed as before:

ANOVA for Multiple Linear Regression

We would like to test the null hypothesis that the exam scores, the IQ's of students and the hours of study they spend preparing the exam are not linearly correlated against the alternative hypothesis that the exam scores,the IQ's of students and the hours of study they spend preparing the exam are linearly correlated . Also, we would like to find the multiple linear regression equation that best fits data and use it for prediction.

A full report of the regression characteristics of the x1 - x2 - y relation, can be given again with the activation of a single command

```
regr = Regress[data, {1, x1, x2}, {x1, x2}]
```

```
{ParameterTable → |      | Estimate   | SE       | TStat    | PValue    |
                  | 1    | -124.568   | 47.8742  | -2.60199 | 0.0353227 |
                  | x1   | 1.65914    | 0.446139 | 3.71889  | 0.00746694 |
                  | x2   | 1.4393     | 0.6668   | 1.34917  | 0.21929   |

 RSquared → 0.781661, AdjustedRSquared → 0.719279, EstimatedVariance → 101.921,

 ANOVATable → |       | DF | SumOfSq | MeanSq  | FRatio  | PValue     |
              | Model | 2  | 2554.16 | 1277.08 | 12.5301 | 0.00486362 |
              | Error | 7  | 713.444 | 101.921 |         |            |
              | Total | 9  | 3267.6  |         |         |            }
```

The F-ratio in the ANOVA Table above corresponds to a p-value of 0.00486362 which is smaller than $\alpha = 0.05$. Therefore, one must reject the null hypothesis, i.e., it must be concluded that the exam scores,the IQ's of students and the hours of study they spend preparing the exam are linearly correlated.

Multiple Linear Regression Equation

The first column entries in the parameter table of the full regression report are the constants of the linear regression function. This function can be formed by using COPY PASTE for each of these entries and writing

```
funct = (-124.56815350155742`) + (1.6591432471214065`) x1 + (1.4392975886326242`) x2
```

```
-124.568 + 1.65914 x1 + 1.4393 x2
```

Prediction

For a student with an IQ of 106 and 6 hours of study , one expects his/her exam performance to be(COPY PASTE the function above and set x1 = 106, x2 = 6)

```
(-124.56815350155742`) + (1.6591432471214065`) 106 + (1.4392975886326242`) 6
```

```
59.9368
```

on the average.

Model Evaluation

The goodness of the model is expressed by the multiple coefficient of determination (also denoted in the output as RSquared) which as we see is 0.781661 or 78.17 % .