

Appendix A

Collection of formulas and R commands

Contents

- A Collection of formulas and R commands**
 - A.1 Introduction, descriptive statistics, R and data visualization 1
 - A.2 Probability and Simulation 3
 - A.2.1 Distributions 5
 - A.3 Statistics for one and two samples 9
 - A.4 Simulation based statistics 11
 - A.5 Simple linear regression 12
 - A.6 Multiple linear regression 14
 - A.7 Inference for proportions 15
 - A.8 Comparing means of multiple groups - ANOVA 16
- Glossaries** 18
- Acronyms** 19

This appendix chapter holds a collection of formulas. All the relevant equations from definitions, methods and theorems are included – along with associated R functions. All are included in the same order as in the book, except for the distributions which are listed together.

A.1 Introduction, descriptive statistics, R and data visualization

	Description	Formula	R command
1.4	Sample mean The mean of a sample.	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	mean(x)
1.5	Sample median The value that divides a sample in two halves with equal number of observations in each.	$Q_2 = \begin{cases} x_{(\frac{n+1}{2})} & \text{for odd } n \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}}{2} & \text{for even } n \end{cases}$	median(x)
1.7	Sample quantile The value that divide a sample such that p of the observations are less than the value. The 0.5 quantile is the Median.	$q_p = \begin{cases} \frac{x_{(np)} + x_{(np+1)}}{2} & \text{for } pn \text{ integer} \\ x_{(\lceil np \rceil)} & \text{for } pn \text{ non-integer} \end{cases}$	quantile(x,p,type=2),
1.8	Sample quartiles The quartiles are the five quantiles dividing the sample in four parts, such that each part holds an equal number of observations	$Q_0 = q_0 = \text{"minimum"}$ $Q_1 = q_{0.25} = \text{"lower quartile"}$ $Q_2 = q_{0.5} = \text{"median"}$ $Q_3 = q_{0.75} = \text{"upper quartile"}$ $Q_4 = q_1 = \text{"maximum"}$	quantile(x, probs,type=2) where probs=p
1.10	Sample variance The sum of squared differences from the mean divided by $n - 1$.	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	var(x)
1.11	Sample standard deviation The square root of the sample variance.	$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$	sd(x)
1.12	Sample coefficient of variance The sample standard deviation seen relative to the sample mean.	$V = \frac{s}{\bar{x}}$	sd(x)/mean(x)
1.15	Sample Inter Quartile Range IQR: The middle 50% range of data	$IQR = Q_3 - Q_1$	IQR(x)

	Description	Formula	R command
1.18	Sample covariance Measure of linear strength of relation between two samples	$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$	<code>cov(x,y)</code>
1.19	Sample correlation Measure of the linear strength of relation between two samples between -1 and 1.	$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}$	<code>cor(x,y)</code>

A.2 Probability and Simulation

	Description	Formula	R command
2.6	Probability density function (pdf) for a discrete variable fulfills two conditions: $f(x) \geq 0$ and $\sum_{\text{all } x} f(x) = 1$ and finds the probability for one x value.	$f(x) = P(X = x)$	dnorm, dbinom, dhyper, dpois
2.9	Cumulated distribution function (cdf) gives the probability in a range of x values where $P(a < X \leq b) = F(b) - F(a)$.	$F(x) = P(X \leq x)$	pnorm, pbinom, phyper, ppois
2.13	Mean of a discrete random variable	$\mu = E(X) = \sum_{i=1}^{\infty} x_i f(x_i)$	
2.16	Variance of a discrete random variable X	$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2]$	
2.32	Pdf of a continuous random variable is a non-negative function for all possible outcomes and has an area below the function of one	$P(a < X \leq b) = \int_a^b f(x) dx$	
2.33	Cdf of a continuous random variable is non-decreasing and $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$	$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$	
2.34	Mean and variance for a continuous random variable X	$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$ $\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$	
2.54	Mean and variance of a linear function The mean and variance of a linear function of a random variable X .	$E(aX + b) = a E(X) + b$ $V(aX + b) = a^2 V(X)$	
2.56	Mean and variance of a linear combination The mean and variance of a linear combination of random variables.	$E(a_1 X_1 + a_2 X_2 + \dots + a_n X_n) =$ $a_1 E(X_1) + a_2 E(X_2) + \dots + a_n E(X_n)$ $V(a_1 X_1 + a_2 X_2 + \dots + a_n X_n) =$ $a_1^2 V(X_1) + a_2^2 V(X_2) + \dots + a_n^2 V(X_n)$	

	Description	Formula	R command
2.58	Covariance The covariance between two random variables X and Y .	$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$	

A.2.1 Distributions

Here all the included distributions are listed including some important theorems and definitions related specifically with a distribution.

	Description	Formula	R command
2.20	Binominal distribution n is the number of independent draws and p is the probability of a success in each draw. The Binominal pdf describes the probability of x successes.	$f(x; n, p) = P(X = x)$ $= \binom{n}{x} p^x (1 - p)^{n-x}$ <p>where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$</p>	<code>dbinom(x,size,prob)</code> <code>pbinom(q,size,prob)</code> <code>qbinom(p,size,prob)</code> <code>rbinom(n,size,prob)</code> <i>where</i> <code>size=n, prob=p</code>
2.21	Mean and variance of a binomial distributed random variable.	$\mu = np$ $\sigma^2 = np(1 - p)$	
2.24	Hypergeometric distribution n is the number of draws without replacement, a is number of successes and N is the population size.	$f(x; n, a, N) = P(X = x)$ $= \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}}$ <p>where $\binom{a}{b} = \frac{a!}{b!(a-b)!}$</p>	<code>dhyper(x,m,n,k)</code> <code>phyper(q,m,n,k)</code> <code>qhyper(p,m,n,k)</code> <code>rhyper(nn,m,n,k)</code> <i>where</i> <code>m=a, n=N - a, k=n</code>
2.25	Mean and variance of a hypergeometric distributed random variable.	$\mu = n \frac{a}{N}$ $\sigma^2 = n \frac{a(N-a)}{N} \frac{N-n}{N-1}$	
2.27	Poisson distribution λ is the rate (or intensity) i.e. the average number of events per interval. The Poisson pdf describes the probability of x events in an interval.	$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$	<code>dpois(x,lambda)</code> <code>ppois(q,lambda)</code> <code>qpois(p,lambda)</code> <code>rpois(n,lambda)</code> <i>where</i> <code>lambda=λ</code>
2.28	Mean and variance of a Poisson distributed random variable.	$\mu = \lambda$ $\sigma^2 = \lambda$	
2.35	Uniform distribution α and β defines the range of possible outcomes. random variable following the uniform distribution has equal density at any value within a defined range.	$f(x; \alpha, \beta) = \begin{cases} 0 & \text{for } x < \alpha \\ \frac{1}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{for } x > \beta \end{cases}$ $F(x; \alpha, \beta) = \begin{cases} 0 & \text{for } x < \alpha \\ \frac{x - \alpha}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{for } x > \beta \end{cases}$	<code>dunif(x,min,max)</code> <code>punif(q,min,max)</code> <code>qunif(p,min,max)</code> <code>runif(n,min,max)</code> <i>where</i> <code>min=α, max=β</code>

	Description	Formula	R command
2.36	Mean and variance of a uniform distributed random variable X .	$\mu = \frac{1}{2}(\alpha + \beta)$ $\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$	
2.37	Normal distribution Often also called the Gaussian distribution.	$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	dnorm(x,mean,sd) pnorm(q,mean,sd) qnorm(p,mean,sd) rnorm(n,mean,sd) <i>where</i> mean= μ , sd= σ .
2.38	Mean and variance of a normal distributed random variable.	μ σ^2	
2.43	Transformation of a normal distributed random variable X into a standardized normal random variable.	$Z = \frac{X - \mu}{\sigma}$	
2.46	Log-normal distribution α is the mean and β^2 is the variance of the normal distribution obtained when taking the natural logarithm to X .	$f(x) = \frac{1}{x\sqrt{2\pi}\beta} e^{-\frac{(\ln x - \alpha)^2}{2\beta^2}}$	dlnorm(x,meanlog,sdlog) plnorm(q,meanlog,sdlog) qlnorm(p,meanlog,sdlog) rlnorm(n,meanlog,sdlog) <i>where</i> meanlog= α , sdlog= β .
2.47	Mean and variance of a log-normal distributed random variable.	$\mu = e^{\alpha + \beta^2/2}$ $\sigma^2 = e^{2\alpha + \beta^2}(e^{\beta^2} - 1)$	
2.48	Exponential distribution λ is the mean rate of events.	$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$	dexp(x,rate) pexp(q,rate) qexp(p,rate) rexp(n,rate) <i>where</i> rate= λ .
2.49	Mean and variance of an exponential distributed random variable.	$\mu = \frac{1}{\lambda}$ $\sigma^2 = \frac{1}{\lambda^2}$	
2.78	χ^2-distribution $\Gamma(\frac{\nu}{2})$ is the Γ -function and ν is the degrees of freedom.	$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}; \quad x \geq 0$	dchisq(x,df) pchisq(q,df) qchisq(p,df) rchisq(n,df) <i>where</i> df= ν .

	Description	Formula	R command
2.81	Given a sample of size n from the normal distributed random variables X_i with variance σ^2 , then the sample variance S^2 (viewed as random variable) can be transformed to follow the χ^2 distribution with the degrees of freedom $\nu = n - 1$.	$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$	
2.83	Mean and variance of a χ^2 distributed random variable.	$E(X) = \nu$ $V(X) = 2\nu$	
2.86	t-distribution ν is the degrees of freedom and $\Gamma()$ is the Gamma function.	$f_T(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	
2.87	Relation between normal random variables and χ^2 -distributed random variables. $Z \sim N(0,1)$ and $Y \sim \chi^2(\nu)$.	$X = \frac{Z}{\sqrt{Y/\nu}} \sim t(\nu)$	<code>dt(x,df)</code> <code>pt(q,df)</code> <code>qt(p,df)</code> <code>rt(n,df)</code> where <code>df=ν</code> .
2.89	For normal distributed random variables X_1, \dots, X_n , the random variable follows the t -distribution, where \bar{X} is the sample mean, μ is the mean of X , n is the sample size and S is the sample standard deviation.	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$	
2.93	Mean and variance of a t -distributed variable X .	$\mu = 0; \quad \nu > 1$ $\sigma^2 = \frac{\nu}{\nu-2}; \quad \nu > 2$	
2.95	F-distribution ν_1 and ν_2 are the degrees of freedom and $B(\cdot, \cdot)$ is the Beta function.	$f_F(x) = \frac{1}{B(\frac{\nu_1}{2}, \frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot x^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}$	<code>df(x,df1,df2)</code> <code>pf(q,df1,df2)</code> <code>qf(p,df1,df2)</code> <code>rf(n,df1,df2)</code> where <code>df1=ν₁, df2=ν₂</code> .
2.96	The F -distribution appears as the ratio between two independent χ^2 -distributed random variables with $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$.	$\frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2)$	

	Description	Formula	R command
2.98	X_1, \dots, X_{n_1} and Y_1, \dots, Y_{n_2} with the mean μ_1 and μ_2 and the variance σ_1^2 and σ_2^2 is independent and sampled from a normal distribution.	$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$	
2.101	Mean and variance of a F -distributed variable X .	$\mu = \frac{\nu_2}{\nu_2 - 2}; \quad \nu_2 > 2$ $\sigma = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}; \quad \nu_2 > 4$	

A.3 Statistics for one and two samples

	Description	Formula	R command
3.2	The distribution of the mean of normal random variables.	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$	
3.4	The distribution of the σ -standardized mean of normal random variables	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$	
3.4	The distribution of the S -standardized mean of normal random variables	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$	
3.6	Standard Error of the mean	$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$	
3.8	The one sample confidence interval for μ	$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$	
3.13	Central Limit Theorem (CLT)	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	
3.18	Confidence interval for the variance and standard deviation	$\sigma^2 : \left[\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{\alpha/2}^2} \right]$ $\sigma : \left[\sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}}, \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}} \right]$	
3.21	The p -value	The p-value is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.	$P(T > x) = 2(1 - pt(x, n-1))$
3.22	The one-sample t -test statistic and p -value	$p\text{-value} = 2 \cdot P(T > t_{\text{obs}})$ $t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ $H_0 : \mu = \mu_0$	
3.23	The hypothesis test	Rejected: $p\text{-value} < \alpha$ Accepted: <i>otherwise</i>	
3.28	Significant effect	An effect is significant if the $p\text{-value} < \alpha$	
3.30	The critical values: $\alpha/2$ - and $1 - \alpha/2$ -quantiles of the t -distribution with $n - 1$ degrees of freedom	$t_{\alpha/2}$ and $t_{1-\alpha/2}$	
3.31	The one-sample hypothesis test by the critical value	Reject: $ t_{\text{obs}} > t_{1-\alpha/2}$ accept: <i>otherwise</i>	

	Description	Formula	R command
3.32	Confidence interval for μ	$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$ acceptance region/CI: $H_0 : \mu = \mu_0$	
3.35	The level α one-sample t -test	Test: $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ by $p\text{-value} = 2 \cdot P(T > t_{\text{obs}})$ Reject: $p\text{-value} < \alpha$ or $ t_{\text{obs}} > t_{1-\alpha/2}$ Accept: <i>Otherwise</i>	
3.62	The one-sample confidence interval (CI) sample size formula	$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME} \right)^2$	
3.64	The one-sample sample size formula	$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)} \right)^2$	
3.41	The Normal q-q plot with $n > 10$	naive approach: $p_i = \frac{i}{n}$, $i = 1, \dots, n$ commonly approach: $p_i = \frac{i-0.5}{n+1}$, $i = 1, \dots, n$	
3.48	The (Welch) two-sample t -test statistic	$\delta = \mu_2 - \mu_1$ $H_0 : \delta = \delta_0$ $t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	
3.49	The distribution of the (Welch) two-sample statistic	$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$ $\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$	
3.50	The level α two-sample t -test	Test: $H_0 : \mu_1 - \mu_2 = \delta_0$ and $H_1 : \mu_1 - \mu_2 \neq \delta_0$ by $p\text{-value} = 2 \cdot P(T > t_{\text{obs}})$ Reject: $p\text{-value} < \alpha$ or $ t_{\text{obs}} > t_{1-\alpha/2}$ Accept: <i>Otherwise</i>	
3.51	The pooled two-sample estimate of variance	$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	
3.52	The pooled two-sample t -test statistic	$\delta = \mu_1 - \mu_2$ $H_0 : \delta = \delta_0$ $t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$	
3.53	The distribution of the pooled two-sample t -test statistic	$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$	
3.46	The two-sample confidence interval for $\mu_1 - \mu_2$	$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$	

A.4 Simulation based statistics

	Description	Formula	R command
4.3	The non-linear approximative error propagation rule	$\sigma_{f(X_1, \dots, X_n)}^2 = \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2$	
4.4	Non-linear error propagation by simulation	1. Simulate k outcomes 2. Calculate the standard deviation by $s_{f(X_1, \dots, X_n)}^{\text{sim}} = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (f_j - \bar{f})^2}$	
4.7	Confidence interval for any feature θ by parametric bootstrap	1. Simulate k samples 2. Calculate the statistic $\hat{\theta}$ 3. Calculate CI: $[q_{100(\alpha/2)\%}^*, q_{100(1-\alpha/2)\%}^*]$	
4.10	Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ by parametric bootstrap	1. Simulate k sets of 2 samples 2. Calculate the statistic $\hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$ 3. Calculate CI: $[q_{100(\alpha/2)\%}^*, q_{100(1-\alpha/2)\%}^*]$	

A.5 Simple linear regression

	Description	Formula	R command
5.4	Least square estimators	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}$ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ <p>where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$</p>	
5.7	Variance of estimators	$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}$ $V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$ $\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x} \sigma^2}{S_{xx}}$	
5.11	Tests statistics for $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$	$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}$ $T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}$	
5.13	Level α t -tests for parameter	<p>Test $H_{0,i} : \beta_i = \beta_{0,i}$ vs. $H_{1,i} : \beta_i \neq \beta_{0,i}$ with $p\text{-value} = 2 \cdot P(T > t_{\text{obs},\beta_i})$ where $t_{\text{obs},\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$. If $p\text{-value} < \alpha$ then <i>reject</i> H_0, otherwise <i>accept</i> H_0</p>	<pre>D <- data.frame(x=c(), y=c()) fit <- lm(y~x, data=D) summary(fit)</pre>
5.14	Parameter confidence intervals	$\hat{\beta}_0 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_0}$ $\hat{\beta}_1 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_1}$	<code>confint(fit, level=0.95)</code>
5.17	Confident and prediction interval	<p>Confidence interval for the line:</p> $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$ <p>Interval for a new point prediction:</p> $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$	<pre>predict(fit, newdata=data.frame(), interval="confidence", level=0.95) predict(fit, newdata=data.frame(), interval="prediction", level=0.95)</pre>
5.22	The matrix formulation of the parameter estimators in the simple linear regression model	$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ $V[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ $\hat{\sigma}^2 = \frac{\text{RSS}}{n-2}$	
5.24	Coefficient of determination R^2	$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$	

	Description	Formula	R command
5.7	Model validation of assumptions	<p>> Check the normality assumption with a q-q plot of the residuals.</p> <p>> Check the systematic behavior by plotting the residuals e_i as a function of fitted values \hat{y}_i</p>	<pre>qqnorm(fit\$residuals) qqline(fit\$residuals) plot(fit\$fitted.values, fit\$residuals)</pre>

A.6 Multiple linear regression

	Description	Formula	R command
6.2	Level α t -tests for parameter	Test $H_{0,i} : \beta_i = \beta_{0,i}$ vs. $H_{1,i} : \beta_i \neq \beta_{0,i}$ with p -value $= 2 \cdot P(T > t_{\text{obs},\beta_i})$ where $t_{\text{obs},\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\hat{\beta}_i}}$. If p -value $< \alpha$ the <i>reject</i> H_0 , otherwise <i>accept</i> H_0	<pre>D<-data.frame(x1=c(), x2=c(),y=c()) fit <- lm(y~x1+x2, data=D) summary(fit)</pre>
6.5	Parameter confidence intervals	$\hat{\beta}_i \pm t_{1-\alpha/2} \hat{\sigma}_{\hat{\beta}_i}$	<code>confint(fit,level=0.95)</code>
6.9	Confident and prediction interval (in R)	Confident interval for the line $\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_p x_{p,\text{new}}$ Interval for a new point prediction $\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_p x_{p,\text{new}} + \varepsilon_{\text{new}}$	<pre>predict(fit, newdata=data.frame(), interval="confidence", level=0.95) predict(fit, newdata=data.frame(), interval="prediction", level=0.95)</pre>
6.17	The matrix formulation of the parameter estimators in the multiple linear regression model	$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ $V[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ $\hat{\sigma}^2 = \frac{RSS}{n - (p + 1)}$	
6.16	Model selection procedure	Backward selection: start with full model and stepwise remove insignificant terms	

A.7 Inference for proportions

	Description	Formula	R command
7.3	Proportion estimate and confidence interval	$\hat{p} = \frac{x}{n}$ $\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	<code>prop.test(x=, n=, correct=FALSE)</code>
7.10	Approximate proportion with Z	$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0, 1)$	
7.11	The level α one-sample proportion hypothesis test	Test: $H_0 : p = p_0$, vs. $H_1 : p \neq p_0$ by p -value $= 2 \cdot P(Z > z_{\text{obs}})$ where $Z \sim N(0, 1^2)$ If p -value $< \alpha$ the <i>reject</i> H_0 , otherwise <i>accept</i> H_0	<code>prop.test(x=, n=, correct=FALSE)</code>
7.13	Sample size formula for the CI of a proportion	Guessed p (with prior knowledge): $n = p(1-p) \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$ Unknown p : $n = \frac{1}{4} \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$	
7.15	Difference of two proportions estimator $\hat{p}_1 - \hat{p}_2$ and confidence interval for the difference	$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ $(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$	
7.18	The level α one-sample t -test	Test: $H_0 : p_1 = p_2$, vs. $H_1 : p_1 \neq p_2$ by p -value $= 2 \cdot P(Z > z_{\text{obs}})$ where $Z \sim N(0, 1^2)$ If p -value $< \alpha$ the <i>reject</i> H_0 , otherwise <i>accept</i> H_0	<code>prop.test(x=, n=, correct=FALSE)</code>
7.20	The multi-sample proportions χ^2 -test	Test: $H_0 : p_{11} = p_{12} = \dots = p_{1c} = p$ by $\chi^2_{\text{obs}} = \sum_{i=1}^2 \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$	<code>chisq.test(X, correct = FALSE)</code>
7.22	The $r \times c$ frequency table χ^2 -test	Test: $H_0 : p_{i1} = p_{i2} = \dots = p_{ic} = p_i$ for all rows $i = 1, 2, \dots, r$ by $\chi^2_{\text{obs}} = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ Reject if $\chi^2_{\text{obs}} > \chi^2_{1-\alpha}((r-1)(c-1))$ Otherwise accept	<code>chisq.test(X, correct = FALSE)</code>

A.8 Comparing means of multiple groups - ANOVA

	Description	Formula	R command
8.2	One-way ANOVA variation decomposition	$\underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}_{\text{SS(Tr)}}$	
8.4	One-way within group variability	$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + \dots + (n_k-1)s_k^2}{n-k}$ $s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	
8.6	One-way test for difference in mean for k groups	$H_0: \alpha_i = 0; \quad i = 1, 2, \dots, k,$ $F = \frac{SS(\text{Tr})/(k-1)}{SSE/(n-k)}$ <p>F-distribution with $k-1$ and $n-k$ degrees of freedom</p>	<code>anova(lm(y~treatm))</code>
8.9	Post hoc pairwise confidence intervals	$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{n-k} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$ <p>If all $M = k(k-1)/2$ combinations, then use $\alpha_{\text{Bonferroni}} = \alpha/M$</p>	
8.10	Post hoc pairwise hypothesis tests	<p>Test: $H_0: \mu_i = \mu_j$ vs. $H_1: \mu_i \neq \mu_j$ by $p\text{-value} = 2 \cdot P(T > t_{\text{obs}})$ where $t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$</p> <p>Test $M = k(k-1)/2$ times, but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$</p>	
8.13	Least Significant Difference (LSD) values	$LSD_{\alpha} = t_{1-\alpha/2} \sqrt{2 \cdot MSE / m}$	
8.20	Two-way ANOVA variation decomposition	$\underbrace{\sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\mu})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu})^2}_{\text{SSE}} + \underbrace{l \cdot \sum_{i=1}^k \hat{\alpha}_i^2}_{\text{SS(Tr)}} + \underbrace{k \cdot \sum_{j=1}^l \hat{\beta}_j^2}_{\text{SS(BI)}}$	

	Description	Formula	R command
8.22	Test for difference in means in two-way ANOVA grouped in treatments and in blocks	$H_{0,Tr} : \alpha_i = 0, \quad i = 1, 2, \dots, k$ $F_{Tr} = \frac{SS(Tr)/(k-1)}{SSE/((k-1)(l-1))}$ $H_{0,Bl} : \beta_j = 0, \quad j = 1, 2, \dots, l$ $F_{Bl} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}$	<pre>fit<-lm(y~treatm+block) anova(fit)</pre>

One-way ANOVA

Source of variation	Degrees of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p -value
Treatment	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{obs} = \frac{MS(Tr)}{MSE}$	$P(F > F_{obs})$
Residual	$n - k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n - 1$	SST			

Two-way ANOVA

Source of variation	Degrees of freedom	Sums of squares	Mean sums of squares	Test statistic F	p -value
Treatment	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{Tr} = \frac{MS(Tr)}{MSE}$	$P(F > F_{Tr})$
Block	$l - 1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{Bl} = \frac{MS(Bl)}{MSE}$	$P(F > F_{Bl})$
Residual	$(l - 1)(k - 1)$	SSE	$MSE = \frac{SSE}{(k-1)(l-1)}$		
Total	$n - 1$	SST			

Glossaries

cumulated distribution function [Fordelingsfunksjon] The cdf is the function which determines the probability of observing an outcome of a random variable below a given value [3](#),

confidence interval [Konfidensinterval] The confidence interval is a way to handle the uncertainty by the use of probability theory. The confidence interval represents those values of the unknown population mean μ that we believe is based on the data. Thus we believe the true mean in the statistics class is in this interval

Continuous random variable [Kontinuert stokastisk variabel] If an outcome of an experiment takes a continuous value, for example: a distance, a temperature, a weight, etc., then it is represented by a continuous random variable [3](#)

Correlation [Korrelasjon] The sample correlation coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Covariance [2](#)

Covariance [Kovarians] The sample covariance coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Correlation [2](#), [4](#)

Inter Quartile Range [Interkvartil bredde] The Inter Quartile Range (IQR) is the middle 50% range of data [1](#)

Median [Median, stikprøvemedian] The median of population or sample (note, in text no distinction between *population median* and *sample median*) [1](#)

probability density function The pdf is the function which determines the probability of every possible outcome of a random variable [3](#),

Quantile [Fraktil, stikprøvefraktil] The quantiles of population or sample (note, in text no distinction between *population quantile* and *sample quantile*) [1](#)

Quartile [Fraktil, stikprøvefraktil] The quartiles of population or sample (note, in text no distinction between *population quartile* and *sample quartile*) [1](#)

Sample variance [Empirisk varians, stikprøvevarians] [1](#)

Sample mean [Stikprøvegennemsnitt] The average of a sample [1](#)

Standard deviation [Standard avvigelse] [1](#)

Acronyms

ANOVA Analysis of Variance *Glossary:* [Analysis of Variance](#)

cdf cumulated distribution function *Glossary:* [cumulated distribution function](#)

CI confidence interval *Glossary:* [confidence interval](#)

CLT Central Limit Theorem *Glossary:* [Central Limit Theorem](#)

IQR Inter Quartile Range *Glossary:* [Inter Quartile Range](#)

LSD Least Significant Difference *Glossary:* [Least Significant Difference](#)

pdf probability density function *Glossary:* [probability density function](#)