Statistics Quick Reference Card & R Commands

by Anthony Tanbakuchi. Version 1.8.2
http://www.tanbakuchi.com
ANTHONY@TANBAKUCHI·COM

Get R at: http://www.r-project.org

R commands: **`bold typewriter text`**

1 Misc R

To make a vector / store data: **`x=c(x1, x2, ...)`**
Help: general **`RSiteSearch("Search Phrase")`**
Help: function **`?functionName`**
Get column of data from table:
`tableName$columnName`
List all variables: **`ls()`**
Delete all variables: **`rm(list=ls())`**

$$\sqrt{x} = \texttt{sqrt(x)} \tag{1}$$

$$x^n = \texttt{x^n} \tag{2}$$

$$n = \texttt{length(x)} \tag{3}$$

$$T = \texttt{table(x)} \tag{4}$$

2 Descriptive Statistics

2.1 Numerical

Let **`x=c(x1, x2, x3, ...)`**

$$\text{total} = \sum_{i=1}^{n} x_i = \texttt{sum(x)} \tag{5}$$

$$\text{min} = \texttt{min(x)} \tag{6}$$

$$\text{max} = \texttt{max(x)} \tag{7}$$

$$\text{six number summary}: \texttt{summary(x)} \tag{8}$$

$$\mu = \frac{\sum x_i}{N} = \texttt{mean(x)} \tag{9}$$

$$\bar{x} = \frac{\sum x_i}{n} = \texttt{mean(x)} \tag{10}$$

$$\tilde{x} = P_{50} = \texttt{median(x)} \tag{11}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \tag{12}$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \texttt{sd(x)} \tag{13}$$

$$CV = \frac{\sigma}{\mu} = \frac{s}{\bar{x}} \tag{14}$$

2.2 Relative standing

$$z = \frac{x-\mu}{\sigma} = \frac{x-\bar{x}}{s} \tag{15}$$

Percentiles:

$$P_k = x_i, \quad (\text{sorted } x)$$

$$k = \frac{i-0.5}{n} \cdot 100\% \tag{16}$$

To find x_i given P_k, i is:

1. $L = (k/100\%)n$
2. if L is an integer: $i = L + 0.5$; otherwise i=L and round up.

2.3 Visual

All plots have optional arguments:

- **`main=""`** sets title
- **`xlab="", ylab=""`** sets x/y-axis label
- **`type="p"`** for **p**oint plot
- **`type="l"`** for **l**ine plot
- **`type="b"`** for **b**oth points and lines

Ex: `plot(x, y, type="b", main="My Plot")`

Plot Types:

`hist(x)` histogram
`stem(x)` stem & leaf
`boxplot(x)` box plot
`plot(T)` bar plot, `T=table(x)`
`plot(x,y)` scatter plot, `x`, `y` are ordered vectors
`plot(t,y)` time series plot, `t`, `y` are ordered vectors
`curve(expr, xmin,xmax)` plot `expr` involving x

2.4 Assessing Normality

Q-Q plot: **`qqnorm(x); qqline(x)`**

3 Probability

Number of successes x with n possible outcomes. (Don't double count!)

$$P(A) = \frac{x_A}{n} \tag{17}$$

$$P(\bar{A}) = 1 - P(A) \tag{18}$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{19}$$

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{if } A, B \text{ mut. excl.} \tag{20}$$

$$P(A \text{ and } B) = P(A) \cdot P(B|A) \tag{21}$$

$$P(A \text{ and } B) = P(A) \cdot P(B) \quad \text{if } A, B \text{ independent} \tag{22}$$

$$n! = n(n-1)\cdots 1 = \texttt{factorial(n)} \tag{23}$$

$${}_nP_k = \frac{n!}{(n-k)!} \quad \text{Perm. no elem. alike} \tag{24}$$

$$= \frac{n!}{n_1! n_2! \cdots n_k!} \quad \text{Perm. } n_1 \text{ alike}, \ldots \tag{25}$$

$${}_nC_k = \frac{n!}{(n-k)!k!} = \texttt{choose(n,k)} \tag{26}$$

4 Discrete Random Variables

$$P(x_i): \text{ probability distribution} \tag{27}$$

$$E = \mu = \sum x_i \cdot P(x_i) \tag{28}$$

$$\sigma = \sqrt{\sum (x_i - \mu)^2 \cdot P(x_i)} \tag{29}$$

4.1 Binomial distribution

$$\mu = n \cdot p \tag{30}$$

$$\sigma = \sqrt{n \cdot p \cdot q} \tag{31}$$

$$P(x) = {}_nC_x p^x q^{(n-x)} = \texttt{dbinom(x, n, p)} \tag{32}$$

4.2 Poisson distribution

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!} = \texttt{dpois(x, }\mu\texttt{)} \tag{33}$$

5 Continuous random variables

CDF $F(x)$ gives area to the left of x, $F^{-1}(p)$ expects p is area to the left.

$$f(x): \text{ probability density} \tag{34}$$

$$E = \mu = \int_{-\infty}^{\infty} x \cdot f(x)\,dx \tag{35}$$

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx} \tag{36}$$

$$F(x): \text{ cumulative prob. density (CDF)} \tag{37}$$

$$F^{-1}(x): \text{ inv. cumulative prob. density} \tag{38}$$

$$F(x) = \int_{-\infty}^{x} f(x')\,dx' \tag{39}$$

$$p = P(x < x') = F(x') \tag{40}$$

$$x' = F^{-1}(p) \tag{41}$$

$$p = P(x > a) = 1 - F(a) \tag{42}$$

$$p = P(a < x < b) = F(b) - F(a) \tag{43}$$

5.1 Uniform distribution

$$p = P(u < u') = F(u')$$

$$= \texttt{punif(u', min=0, max=1)} \tag{44}$$

$$u' = F^{-1}(p) = \texttt{qunif(p, min=0, max=1)} \tag{45}$$

5.2 Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{46}$$

$$p = P(z < z') = F(z') = \texttt{pnorm(z')} \tag{47}$$

$$z' = F^{-1}(p) = \texttt{qnorm(p)} \tag{48}$$

$$p = P(x < x') = F(x')$$

$$= \texttt{pnorm(x', mean=}\mu\texttt{, sd=}\sigma\texttt{)} \tag{49}$$

$$x' = F^{-1}(p)$$

$$= \texttt{qnorm(p, mean=}\mu\texttt{, sd=}\sigma\texttt{)} \tag{50}$$

5.3 t-distribution

$$p = P(t < t') = F(t') = \texttt{pt(t', df)} \tag{51}$$

$$t' = F^{-1}(p) = \texttt{qt(p, df)} \tag{52}$$

5.4 χ^2-distribution

$$p = P(\chi^2 < \chi^{2\prime}) = F(\chi^{2\prime})$$

$$= \texttt{pchisq(}X^{2\prime}\texttt{, df)} \tag{53}$$

$$\chi^{2\prime} = F^{-1}(p) = \texttt{qchisq(p, df)} \tag{54}$$

5.5 F-distribution

$$p = P(F < F') = F(F')$$

$$= \texttt{pf(F', df1, df2)} \tag{55}$$

$$F' = F^{-1}(p) = \texttt{qf(p, df1, df2)} \tag{56}$$

6 Sampling distributions

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{57}$$

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \tag{58}$$

7 Estimation

7.1 Confidence Intervals

$$\text{proportion: } \hat{p} \pm E, \quad E = z_{\alpha/2} \cdot \sigma_{\hat{p}} \tag{59}$$

$$\text{mean } (\sigma \text{ known}): \bar{x} \pm E, \quad E = z_{\alpha/2} \cdot \sigma_{\bar{x}} \tag{60}$$

$$\text{mean } (\sigma \text{ unknown, use } s): \bar{x} \pm E, \quad E = t_{\alpha/2} \cdot \sigma_{\bar{x}}, \tag{61}$$

$$df = n - 1$$

$$\text{variance: } \frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}, \tag{62}$$

$$df = n - 1$$

$$\text{2 proportions: } \Delta\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \tag{63}$$

$$\text{2 means (indep): } \Delta\bar{x} \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{64}$$

$$df \approx \min(n_1 - 1, n_2 - 1)$$

$$\text{matched pairs: } \bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}, \quad d_i = x_i - y_i, \tag{65}$$

$$df = n - 1$$

7.2 CI Critical Values (two sided)

$$z_{\alpha/2} = F_z^{-1}(1 - \alpha/2) = \texttt{qnorm(1-alpha/2)} \tag{66}$$

$$t_{\alpha/2} = F_t^{-1}(1 - \alpha/2) = \texttt{qt(1-alpha/2, df)} \tag{67}$$

$$\chi^2_L = F_{\chi^2}^{-1}(\alpha/2) = \texttt{qchisq(alpha/2, df)} \tag{68}$$

$$\chi^2_R = F_{\chi^2}^{-1}(1 - \alpha/2) = \texttt{qchisq(1-alpha/2, df)} \tag{69}$$

7.3 Required sample size

$$\text{proportion: } n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2, \tag{70}$$

$$(\hat{p} = \hat{q} = 0.5 \text{ if unknown})$$

$$\text{mean: } n = \left(\frac{z_{\alpha/2} \cdot \hat{\sigma}}{E}\right)^2 \tag{71}$$

8 Hypothesis Tests

Test statistic and R function (when available) are listed for each.

Optional arguments for **hypothesis tests**:

`alternative="two.sided"` can be:
`"two.sided"`, `"less"`, `"greater"`

`conf.level=0.95` constructs a 95% confidence interval. Standard CI only when `alternative="two.sided"`.

Optional arguments for **power calculations & Type II error**:

`alternative="two.sided"` can be:
`"two.sided"` or `"one.sided"`

`sig.level=0.05` sets the significance level α.

8.1 1-Sample Proportion

$H_0 : p = p_0$

`prop.test(x, n, p=`p_0`, alternative="two.sided")`

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \tag{72}$$

8.2 1-Sample Mean (σ known)

$H_0 : \mu = \mu_0$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \tag{73}$$

8.3 1-Sample Mean (σ unknown)

$H_0 : \mu = \mu_0$

`t.test(x, mu=`μ_0`, alternative="two.sided")`

Where `x` is a vector of sample data.

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}, \quad df = n - 1 \tag{74}$$

Required Sample size:

`power.t.test(delta=`h`, sd =`σ`, sig.level=`α`, power=`$1-\beta$`, type ="one.sample", alternative="two.sided")`

8.4 2-Sample Proportion Test

$H_0 : p_1 = p_2$ or equivalently $H_0 : \Delta p = 0$

`prop.test(x, n, alternative="two.sided")`

where: `x=c(`x_1`, `x_2`)` and `n=c(`n_1`, `n_2`)`

$$z = \frac{\Delta\hat{p} - \Delta p_0}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}, \quad \Delta\hat{p} = \hat{p}_1 - \hat{p}_2 \tag{75}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}, \quad \bar{q} = 1 - \bar{p} \tag{76}$$

Required Sample size:

`power.prop.test(p1=`p_1`, p2=`p_2`, power=`$1-\beta$`, sig.level=`α`, alternative="two.sided")`

8.5 2-Sample Mean Test

$H_0 : \mu_1 = \mu_2$ or equivalently $H_0 : \Delta\mu = 0$

`t.test(x1, x2, alternative="two.sided")`

where: `x1` and `x2` are vectors of sample 1 and sample 2 data.

$$t = \frac{\Delta\bar{x} - \Delta\mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df \approx \min(n_1 - 1, n_2 - 1), \quad \Delta\bar{x} = \bar{x}_1 - \bar{x}_2 \tag{77}$$

Required Sample size:

`power.t.test(delta=`h`, sd =`σ`, sig.level=`α`, power=`$1-\beta$`, type ="two.sample", alternative="two.sided")`

8.6 2-Sample Matched Pairs Test

$H_0 : \mu_d = 0$

`t.test(x, y, paired=TRUE, alternative="two.sided")`

where: `x` and `y` are ordered vectors of sample 1 and sample 2 data.

$$t = \frac{\bar{d} - \mu_{d0}}{s_d / \sqrt{n}}, \quad d_i = x_i - y_i, \quad df = n - 1 \tag{78}$$

Required Sample size:

`power.t.test(delta=`h`, sd =`σ`, sig.level=`α`, power=`$1-\beta$`, type ="paired", alternative="two.sided")`

8.7 Test of Homogeneity, Test of Independence

$H_0 : p_1 = p_2 = \cdots = p_n$ (homogeneity)

$H_0 : X$ and Y are independent (independence)

`chisq.test(D)`

Enter table: `D=data.frame(c1, c2, ...)`, where `c1, c2, ...` are column data vectors.

Or generate table: `D=table(x1, x2)`, where `x1, x2` are ordered vectors of raw categorical data.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad df = (\text{num rows - 1})(\text{num cols - 1}) \tag{79}$$

$$E_i = \frac{(\text{row total})(\text{column total})}{(\text{grand total})} = np_i \tag{80}$$

For 2×2 contingency tables, you can use the Fisher Exact Test:

`fisher.test(D, alternative="greater")`

(must specify alternative as greater)

9 Linear Regression

9.1 Linear Correlation

$H_0 : \rho = 0$

`cor.test(x, y)`

where: `x` and `y` are ordered vectors.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}, \quad t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}}, \quad df = n - 2 \tag{81}$$

9.2 Models in R

Model type	Equation	R model
linear 1 indep var	$y = b_0 + b_1 x_1$	`y~x1`
...0 intercept	$y = 0 + b_1 x_1$	`y~0+x1`
linear 2 indep vars	$y = b_0 + b_1 x_1 + b_2 x_2$	`y~x1+x2`
...inteaction	$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$	`y~x1+x2+x1*x2`
polynomial	$y = b_0 + b_1 x_1 + b_2 x_2^2$	`y~x1+I(x2^2)`

9.3 Regression

Simple linear regression steps:

1. Make sure there is a significant linear correlation.
2. `results=lm(y~x)` Linear regression of `y` on x vectors
3. `results` View the results
4. `plot(x, y); abline(results)` Plot regression line on data
5. `plot(x, results$residuals)` Plot residuals

$$y = b_0 + b_1 x_1 \tag{82}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{83}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{84}$$

9.4 Prediction Intervals

To predict y when $x = 5$ and show the 95% prediction interval with regression model in `results`:

`predict(results, newdata=data.frame(x=5), int="pred")`

10 ANOVA

10.1 One way ANOVA

1. `results=aov(depVarColName~indepVarColName, data=tableName)` Run ANOVA with data in `TableName`, factor data in `indepVarColName` column, and response data in `depVarColName` column.
2. `summary(results)` Summarize results
3. `boxplot(depVarColName~indepVarColName, data=tableName)` Boxplot of levels for factor

$$F = \frac{MS(\text{treatment})}{MS(\text{error})}, \quad df_1 = k - 1, df_2 = N - k \tag{85}$$

To find required sample size and power see `power.anova.test(...)`

11 Loading and using external data and tables

11.1 Loading Excel Data

1. Export your table as a CSV file (comma seperated file) from Excel.
2. Import your table into `MyTable` in R using:
   `MyTable=read.csv(file.choose())`

11.2 Loading an .RData file

You can either double click on the `.RData` file or use the menu:

- Windows: *File→Load Workspace...*
- Mac: *Workspace→Load Workspace File...*

11.3 Using tables of data

1. To see all the available variables type: `ls()`
2. To see what's inside a variable, type its name.
3. If the variable `tableName` is a table, you can also type `names(tableName)` to see the column names or type `head(tableName)` to see the first few rows of data.
4. To access a column of data type `tableName$columnName`

An example demonstrating how to get the women's height data and find the mean:

```
> ls() # See what variables are defined
[1] "women" "x"
> head(women) #Look at the first few entries
  height weight
1     58    115
2     59    117
3     60    120
> names(women) # Just get the column names
[1] "height" "weight"
> women$height # Display the height data
 [1] 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
> mean(women$height)  # Find the mean of the heights
[1] 65
```