Cheatography

Basic statistics with R Cheat Sheet
by xeonkai via cheatography.com/31513/cs/9886/

Descriptive statistics

Base installation

`summary(), mean(), sd(), var(), min(), max(), median(), length(), range(), quantile(), fivenum()`

`Hmisc` package

`describe()`

`pastecs` package

`stat.desc(x, basic=TRUE, desc=TRUE, norm=FALSE, p=0.75)`

`basic=TRUE` - no. of values, null values, missing values, min, max, range, sum

`desc=TRUE` - median, mean, std error of mean, 95% CI for mean, variance, std dev, coefficient of variation

`norm=TRUE` - skewness, kurtosis, Shapiro-Wilk test of normality

`psych` package

`describe()`

To call function that has been masked, use`Hmisc::describe(x)`

Descriptive statistics by group

`aggregate()`

Single value function - `aggregate(mtcars[vars], by=list(am=mtcars$am), mean)`

Several functions - `by(data, INDICES, FUN)`

```
dstats <- function(x)(c(mean=mean(x), sd=sd(x)))
```

```
by(mtcars[vars], mtcars$am, dstats)
```

`doBy` package

`summaryBy(formula, data=dataframe, FUN=function)`

Formula - `var1 + var2 ... ~ groupvar1 + groupvar2 + ...`

```
summaryBy(mpg+hp+wt~am, data=mtcars, FUN=mystats)
```

Descriptive statistics by group (cont)

`psych` package

`describe.by(mtcars[vars], mtcars$am)`

Frequencies and contingency tables

`table(var1,var2,-...,varN)`	Creates an N-way contingency table from N categorical variables (factors). Ignores missing values (NAs) by default. `useNA="ifany"` to include NA as a valid category.
`xtabs(formula, data)`	Creates an N-way contingency table based on a formula and a matrix or data frame
`prop.table(table-, margins)`	Expresses table entries as fractions of the marginal table defined by the `margins`
`margin.table(table, margins)`	Computes the sum of table entries for a marginal table defined by the `margins`
`addmargins(table-, margins)`	Puts summary `margins` (sums by default) on a table
`ftable(table)`	Creates a compact "flat" contingency table

Example code

One way table

`mytable <- with(Arthritis, table(Improved))`

`prop.table(mytable) # turn frequencies into proportions`

`prop.table(mytable)*100 # turn frequencies into percentages`

Two way table

`mytable <- table(Treatment, Improved)`

`mytable <- xtabs(~ Treatment + Improved, data = Arthritis)`

`margin.table(mytable, 1) # generate marginal frequencies, 2 generates column sums`

`prop.table(mytable, 1) # generate marginal proportions, 2 generates column proportions`

`prop.table(mytable) # cell proportions`

`addmargins(mytable) # adds a sum row and column`

`addmargins(prop.table(mytable))`

`addmargins(prop.table(mytable, 1), 2) # adds a sum column`

`addmargins(prop.table(mytable, 2), 1) # adds a sum row`

Two way tables can be created using `Crosstable()` function in `gmodels` package

Three way table

`ftable()` function can print multidimensional tables

Chi-square test of independence (Two-way table)

```
> library(vcd)
> mytable <- xtabs(~Treatment+Improved, data=Arthritis)
> chisq.test(mytable)

        Pearson's Chi-squared test

data:  mytable
X-squared = 13.1, df = 2, p-value = 0.001463

> mytable <- xtabs(~Improved+Sex, data=Arthritis)
> chisq.test(mytable)

        Pearson's Chi-squared test

data:  mytable
X-squared = 4.84, df = 2, p-value = 0.0889

Warning message:
In chisq.test(mytable) : Chi-squared approximation may be incorrect
```

1 Treatment and Improved not independent

2 Gender and Improved independent

Measures of association (Two-way table)

```
> library(vcd)
> mytable <- xtabs(~Treatment+Improved, data=Arthritis)
> assocstats(mytable)

                    X^2 df  P(> X^2)
Likelihood Ratio 13.530  2 0.0011536
Pearson          13.055  2 0.0014626

Phi-Coefficient   : 0.394
Contingency Coeff.: 0.367
Cramer's V        : 0.394
```

Covariances / correlations

`x`	Matrix or data frame
`use`	Specifies the handling of missing data. Options are `all.obs` (assumes no missing data - missing data will produce an error), `everything` (any correlation involving a case with missing values will be set to `missing`), `complete.obs` (listwise deletion), and `pairwise.complete.obs`(pairwise deletion).
`method`	Specifies the type of correlation. The options are `pearson`, `spearman`, or `kendall`.

Options for `cov/cor=(x, use=, method= )`

Partial correlations

```
> library(ggm)
> # partial correlation of population and murder rate, controlling
> # for income, illiteracy rate, and HS graduation rate
> pcor(c(1,5,2,3,6), cov(states))
[1] 0.346
```

In this case, 0.346 is the correlation between population and murder rate, controlling for the influence of income, illiteracy rate, and HS graduation rate. The use of partial correlations is common in the social sciences.

Testing correlations for significance

```
cor.test(x, y, alternative = , method = )
```

where *x* and *y* are the variables to be correlated, `alternative` specifies a two-tailed or one-tailed test (`"two.side"`, `"less"`, or `"greater"`) and `method` specifies the type of correlation (`"pearson"`, `"kendall"`, or `"spearman"`) to compute. Use `alternative="less"` when the research hypothesis is that the population correlation is less than 0. Use `alternative="greater"` when the research hypothesis is that the population correlation is greater than 0. By default, `alternative="two.side"` (population correlation isn't equal to 0) is assumed. See the following listing for an example.

Listing 7.18 Testing a correlation coefficient for significance

```
> cor.test(states[,3], states[,5])

        Pearson's product-moment correlation

data:  states[, 3] and states[, 5]
t = 6.85, df = 48, p-value = 1.258e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.528 0.821
sample estimates:
  cor
0.703
```

Independent t-test

```
t.test(y ~ x, data)
```

where y is numeric and x is a dichotomous variable, or

```
t.test(y1, y2)
```

where *y1* and *y2* are numeric vectors (the outcome variable for each group). The optional `data` argument refers to a matrix or data frame containing the variables. In contrast to most statistical packages, the default test assumes unequal variance and applies the Welsh degrees of freedom modification. You can add a `var.equal=TRUE` option to specify equal variances and a pooled variance estimate. By default, a two-tailed alternative is assumed (that is, the means differ but the direction isn't specified). You can add the option `alternative="less"` or `alternative="greater"` to specify a directional test.

In the following code, you compare Southern (group 1) and non-Southern (group 0) states on the probability of imprisonment using a two-tailed test without the assumption of equal variances:

```
> library(MASS)
> t.test(Prob ~ So, data=UScrime)

        Welch Two Sample t-test

data:  Prob by So
t = -3.8954, df = 24.925, p-value = 0.0006506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03852569 -0.01187439
sample estimates:
mean in group 0 mean in group 1
     0.03851265      0.06371269
```

You can reject the hypothesis that Southern states and non-Southern states have equal probabilities of imprisonment ($p < .001$).

NOTE Because the outcome variable is a proportion, you might try to transform it to normality before carrying out the t-test. In the current case, all reasonable transformations of the outcome variable (`Y/1-Y`, `log(Y/1-Y)`, `arcsin(Y)`, `arcsin(sqrt(Y))`) would've led to the same conclusions. Transformations are covered in detail in chapter 8.

Dependent t-test

```
t.test(y1, y2, paired=TRUE)
```

where *y1* and *y2* are the numeric vectors for the two dependent groups. The results are as follows:

```
> library(MASS)
> sapply(UScrime[c("U1","U2")], function(x)(c(mean=mean(x),sd=sd(x))))
       U1    U2
mean 95.5 33.98
sd   18.0  8.45

> with(UScrime, t.test(U1, U2, paired=TRUE))

        Paired t-test

data:  U1 and U2
t = 32.4066, df = 46, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 57.67003 65.30870
sample estimates:
mean of the differences
               61.48936
```

The mean difference (61.5) is large enough to warrant rejection of the hypothesis that the mean unemployment rate for older and younger males is the same. Younger males have a higher rate. In fact, the probability of obtaining a sample difference this large if the population means are equal is less than 0.0000000000000022 (that is, 2.2e–16).