UPPSALA
UNIVERSITET

Uppsala Child & Baby Lab

Statistics for Psychology 1

R for Psychology Research

Marcus Lindskog, Docent
2019-10-14

Overview

1. Descriptive statistics
2. Testing one mean and one proportion.
3. Testing two means.
4. Test of Independence
5. Correlations

Descriptive statistics

Descriptive statistics in R.

- There are many ways to do descriptive statistics in R.
- The most simple is to use functions from Base R to calculate `mean()`, `sd()`.
- Another is to use the `summary()`-function in Base R to get basic descriptive statistics on all variables in a data frame.
- One way that gives you a comprehensive overview is `describe()` from the `psych` package
- If you want to customize, then go with `summarise()` in `dplyr`

```r
# code chunk here
library(psych)
test_data <- bfi
describe(test_data)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
A1	1	2784	2.41	1.41	2	2.23	1.48	1	6	5	0.83	-0.31	0.03
A2	2	2773	4.80	1.17	5	4.98	1.48	1	6	5	-1.12	1.05	0.02
A3	3	2774	4.60	1.30	5	4.79	1.48	1	6	5	-1.00	0.44	0.02
A4	4	2781	4.70	1.48	5	4.93	1.48	1	6	5	-1.03	0.04	0.03
A5	5	2784	4.56	1.26	5	4.71	1.48	1	6	5	-0.85	0.16	0.02
C1	6	2779	4.50	1.24	5	4.64	1.48	1	6	5	-0.85	0.30	0.02
C2	7	2776	4.37	1.32	5	4.50	1.48	1	6	5	-0.74	-0.14	0.03
C3	8	2780	4.30	1.29	5	4.42	1.48	1	6	5	-0.69	-0.13	0.02
C4	9	2774	2.55	1.38	2	2.41	1.48	1	6	5	0.60	-0.62	0.03
C5	10	2784	3.30	1.63	3	3.25	1.48	1	6	5	0.07	-1.22	0.03
E1	11	2777	2.97	1.63	3	2.86	1.48	1	6	5	0.37	-1.09	0.03
E2	12	2784	3.14	1.61	3	3.06	1.48	1	6	5	0.22	-1.15	0.03
E3	13	2775	4.00	1.35	4	4.07	1.48	1	6	5	-0.47	-0.47	0.03
E4	14	2791	4.42	1.46	5	4.59	1.48	1	6	5	-0.82	-0.30	0.03
E5	15	2779	4.42	1.33	5	4.56	1.48	1	6	5	-0.78	-0.09	0.03
N1	16	2778	2.93	1.57	3	2.82	1.48	1	6	5	0.37	-1.01	0.03
N2	17	2779	3.51	1.53	4	3.51	1.48	1	6	5	-0.08	-1.05	0.03
N3	18	2789	3.22	1.60	3	3.16	1.48	1	6	5	0.15	-1.18	0.03
N4	19	2764	3.19	1.57	3	3.12	1.48	1	6	5	0.20	-1.09	0.03
N5	20	2771	2.97	1.62	3	2.85	1.48	1	6	5	0.37	-1.06	0.03
O1	21	2778	4.82	1.13	5	4.96	1.48	1	6	5	-0.90	0.43	0.02
O2	22	2800	2.71	1.57	2	2.56	1.48	1	6	5	0.59	-0.81	0.03
O3	23	2772	4.44	1.22	5	4.56	1.48	1	6	5	-0.77	0.30	0.02
O4	24	2786	4.89	1.22	5	5.10	1.48	1	6	5	-1.22	1.08	0.02
O5	25	2780	2.49	1.33	2	2.34	1.48	1	6	5	0.74	-0.24	0.03
gender	26	2800	1.67	0.47	2	1.71	0.00	1	2	1	-0.73	-1.47	0.01
education	27	2577	3.19	1.11	3	3.22	1.48	1	5	4	-0.05	-0.32	0.02
age	28	2800	28.78	11.13	26	27.43	10.38	3	86	83	1.02	0.56	0.21

Grouping is also possible

```r
# code chunk here
library(psych)
test_data <- bfi
by_group_summary <- describeBy(test_data,test_data$gender)
by_group_summary[[1]]
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
A1	1	918	2.73	1.43	2	2.62	1.48	1	6	5	0.56	-0.69	0.05
A2	2	908	4.50	1.26	5	4.65	1.48	1	6	5	-0.89	0.34	0.04
A3	3	912	4.34	1.33	5	4.47	1.48	1	6	5	-0.75	-0.11	0.04
A4	4	916	4.43	1.48	5	4.60	1.48	1	6	5	-0.76	-0.39	0.05
A5	5	915	4.38	1.32	5	4.51	1.48	1	6	5	-0.72	-0.20	0.04
C1	6	913	4.48	1.24	5	4.61	1.48	1	6	5	-0.80	0.17	0.04
C2	7	913	4.24	1.34	5	4.34	1.48	1	6	5	-0.61	-0.42	0.04
C3	8	910	4.20	1.32	4	4.30	1.48	1	6	5	-0.63	-0.35	0.04
C4	9	911	2.72	1.41	2	2.61	1.48	1	6	5	0.41	-0.86	0.05
C5	10	914	3.51	1.65	4	3.52	1.48	1	6	5	-0.13	-1.23	0.05
E1	11	913	3.27	1.66	3	3.21	1.48	1	6	5	0.15	-1.24	0.06
E2	12	913	3.27	1.61	3	3.21	1.48	1	6	5	0.12	-1.18	0.05
E3	13	908	3.91	1.40	4	3.97	1.48	1	6	5	-0.41	-0.63	0.05
E4	14	915	4.27	1.50	5	4.41	1.48	1	6	5	-0.70	-0.58	0.05
E5	15	915	4.29	1.37	5	4.40	1.48	1	6	5	-0.64	-0.36	0.05
N1	16	910	2.83	1.56	3	2.73	1.48	1	6	5	0.40	-1.04	0.05
N2	17	915	3.30	1.52	3	3.27	1.48	1	6	5	0.04	-1.07	0.05
N3	18	917	2.94	1.54	3	2.86	1.48	1	6	5	0.31	-1.08	0.05
N4	19	910	3.19	1.59	3	3.13	1.48	1	6	5	0.15	-1.14	0.05
N5	20	909	2.48	1.49	2	2.30	1.48	1	6	5	0.76	-0.54	0.05
O1	21	913	4.98	1.08	5	5.15	1.48	1	6	5	-1.15	1.22	0.04
O2	22	919	2.65	1.56	2	2.48	1.48	1	6	5	0.66	-0.70	0.05
O3	23	915	4.50	1.22	5	4.63	1.48	1	6	5	-0.87	0.56	0.04
O4	24	913	4.90	1.24	5	5.11	1.48	1	6	5	-1.21	0.97	0.04
O5	25	915	2.45	1.37	2	2.27	1.48	1	6	5	0.83	-0.14	0.05
gender	26	919	1.00	0.00	1	1.00	0.00	1	1	0	NaN	NaN	0.00
education	27	838	3.18	1.19	3	3.22	1.48	1	5	4	-0.10	-0.62	0.04
age	28	919	28.02	11.03	25	26.49	8.90	3	74	71	1.23	1.22	0.36

Testing one mean and one proportion

Single sample t-test

- Single sample t-test is available in Base R.

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

```
t_test_data <- rnorm(100, mean = .6, sd =.2)

critical_mu <- .5

test_statistics <- t.test(t_test_data, mu = .5,
                          alternative = "greater")
```

T-test output

```
print(test_statistics)
```

```
##
##      One Sample t-test
##
## data:  t_test_data
## t = 5.44, df = 99, p-value = 1.929e-07
## alternative hypothesis: true mean is greater than 0.5
## 95 percent confidence interval:
##  0.5759811       Inf
## sample estimates:
## mean of x
## 0.6093601
```

Access parameters

```
#t-value
test_statistics$statistic
```

```
##        t
## 5.439952
```

```
#p-value
test_statistics$p.value
```

```
## [1] 1.928896e-07
```

```
#p-value
test_statistics$conf.int
```

```
## [1] 0.5759811       Inf
## attr(,"conf.level")
## [1] 0.95
```

Testing proportions - Binomial test

```r
binomial_data <- sample(c(0,1), 100, replace = TRUE)
mean(binomial_data)
```

```
## [1] 0.48
```

```r
binom.test(x,n,p=0.5,alternative=c("two.sided","less","greater"),
    conf.level=0.95)
```

Binomial test

```
binom_test <- binom.test(sum(binomial_data),
                         length(binomial_data),p=0.5)
print(binom_test)
```

```
##
##      Exact binomial test
##
## data:  sum(binomial_data) and length(binomial_data)
## number of successes = 48, number of trials = 100, p-value = 0.7644
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3790055 0.5822102
## sample estimates:
## probability of success
##                   0.48
```

```
binom_test$estimate
```

```
## probability of success
##                   0.48
```

Testing two means

Independent t-test

```
two_mean_data_g1 <- rnorm(100, 100, 10)
two_mean_data_g2 <- rnorm(100, 110, 10)
two_mean_data_groups <- rep(c(1,2), each = 100)
two_mean_data <- tibble(Group = two_mean_data_groups,
                        IQ = c(two_mean_data_g1,
                               two_mean_data_g2))
```

Group	IQ
1	105.09253
1	118.24950
1	85.98776
1	95.51200
1	91.91916
1	108.65098
1	112.88644
1	99.09051
1	85.06170
1	84.10629

Independent t-test

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0,
       paired = FALSE,
       var.equal = FALSE,
       conf.level = 0.95, ...)
```

Unequal variance

```
t.test(two_mean_data$IQ~two_mean_data$Group)
```

```
##
##      Welch Two Sample t-test
##
## data:  two_mean_data$IQ by two_mean_data$Group
## t = -8.8878, df = 197.61, p-value = 3.89e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.426396  -9.823851
## sample estimates:
## mean in group 1 mean in group 2
##         97.8682        110.4933
```

Equal variance

```
t.test(two_mean_data$IQ~two_mean_data$Group, var.equal = TRUE)
```

```
##
##      Two Sample t-test
##
## data:  two_mean_data$IQ by two_mean_data$Group
## t = -8.8878, df = 198, p-value = 3.851e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.426363  -9.823885
## sample estimates:
## mean in group 1 mean in group 2
##         97.8682        110.4933
```

Paired t-test

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0,
       paired = FALSE,
       var.equal = FALSE,
       conf.level = 0.95, ...)
```

Paired t-test

```
paired_t_test_statistics <- t.test(two_mean_data$IQ~
                                    two_mean_data$Group,
                                  paired = TRUE)

paired_t_test_statistics
```

```
##
##  Paired t-test
##
## data:  two_mean_data$IQ by two_mean_data$Group
## t = -9.5156, df = 99, p-value = 1.247e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.257739  -9.992509
## sample estimates:
## mean of the differences
##               -12.62512
```

```
paired_t_test_statistics$stderr
```

```
## NULL
```

Non-parametric

Mann-Whitney U-test

- Aka. Wilcoxon rank-sum test,

```
wilcox.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95, ...)
```

```
wilcox.test(two_mean_data$IQ~two_mean_data$Group)
```

```
##
##      Wilcoxon rank sum test with continuity correction
##
## data:  two_mean_data$IQ by two_mean_data$Group
## W = 1870, p-value = 2.064e-14
## alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon signed-rank test

```
wilcox.test(two_mean_data$IQ~two_mean_data$Group,
            paired = TRUE)
```

```
##
##      Wilcoxon signed rank test with continuity correction
##
## data:  two_mean_data$IQ by two_mean_data$Group
## V = 408, p-value = 3.408e-13
## alternative hypothesis: true location shift is not equal to 0
```

Test of Independence

χ^2 - test of Independence

```
var_1 <- sample(c(1:4), 200, replace = TRUE)
var_2 <- sample(c(1:4), 200, replace = TRUE)
var_1 <- factor(var_1, levels = c(1,2,3,4),
                labels = c("very small", "small",
                           "medium", "large" ))
var_2 <- factor(var_2, levels = c(1,2,3,4),
                labels = c("very low", "low",
                           "high", "very high"))
table_of_data <- table(var_1, var_2)
table_of_data
```

```
##              var_2
## var_1         very low low high very high
##   very small        13   8   11        12
##   small             16  10   17        16
##   medium             9   9   18        18
##   large              6  13   15         9
```

χ^2 - test of Independence

```
chi_sq_res <- chisq.test(table_of_data)
chi_sq_res
```

```
##
##      Pearson's Chi-squared test
##
## data:  table_of_data
## X-squared = 8.955, df = 9, p-value = 0.4414
```

Access the test statistics

```
chi_sq_res$observed
```

```
##               var_2
## var_1         very low low high very high
##   very small        13   8   11        12
##   small             16  10   17        16
##   medium             9   9   18        18
##   large              6  13   15         9
```

```
chi_sq_res$expected
```

```
##               var_2
## var_1         very low  low   high very high
##   very small      9.68  8.8 13.420    12.100
##   small          12.98 11.8 17.995    16.225
##   medium         11.88 10.8 16.470    14.850
##   large           9.46  8.6 13.115    11.825
```

Access the test statistics

```
chi_sq_res$statistic
```

```
## X-squared
##  8.954965
```

```
chi_sq_res$p.value
```

```
## [1] 0.4414417
```

Correlations

Pearson correlation

```
cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

```
x <- rnorm(100)
y <- rnorm(100)
cor(x, y, method = "pearson")
```

```
## [1] -0.1001111
```

```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95, continuity = FALSE, ...)
```

Pearson correlation

```
pearson_results <- cor.test(x, y, method = "pearson")
pearson_results
```

```
##
##      Pearson's product-moment correlation
##
## data:  x and y
## t = -0.99605, df = 98, p-value = 0.3217
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.29081084  0.09823872
## sample estimates:
##        cor
## -0.1001111
```

```
pearson_results$p.value
```

```
## [1] 0.3216772
```

Kendaull's tau

```
cor.test(x, y, method = "kendall")
```

```
##
## 	Kendall's rank correlation tau
##
## data:  x and y
## z = -1.0781, p-value = 0.281
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.07313131
```

Spearman's rho

```
cor.test(x, y, method = "spearman")
```

```
##
##      Spearman's rank correlation rho
##
## data:  x and y
## S = 184320, p-value = 0.2932
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1060306
```

Adjusted correlations

```
library(psych)

psych_corr_test <- corr.test(bfi[1:5])
```

```
psych_corr_test$r
```

```
##            A1         A2         A3         A4         A5
## A1  1.0000000 -0.3401932 -0.2652471 -0.1464245 -0.1814383
## A2 -0.3401932  1.0000000  0.4850980  0.3350872  0.3900836
## A3 -0.2652471  0.4850980  1.0000000  0.3604283  0.5041411
## A4 -0.1464245  0.3350872  0.3604283  1.0000000  0.3075373
## A5 -0.1814383  0.3900836  0.5041411  0.3075373  1.0000000
```

Adjusted correlations

```
psych_corr_test$n
```

```
##       A1   A2   A3   A4   A5
## A1 2784 2757 2759 2767 2769
## A2 2757 2773 2751 2758 2757
## A3 2759 2751 2774 2759 2758
## A4 2767 2758 2759 2781 2765
## A5 2769 2757 2758 2765 2784
```

```
psych_corr_test$p
```

```
##                A1            A2            A3           A4            A5
## A1 0.000000e+00  7.027155e-75  3.611010e-45 9.899828e-15  1.276542e-21
## A2 1.171192e-75  0.000000e+00 1.963460e-161 1.192214e-72 5.516048e-100
## A3 1.203670e-45 2.181622e-162  0.000000e+00 1.442527e-84 9.679490e-177
## A4 9.899828e-15  2.384428e-73  2.060752e-85 0.000000e+00  4.817510e-61
## A5 6.382709e-22 6.895059e-101 9.679490e-178 1.204378e-61  0.000000e+00
```

That's all folks!