Statistics with R

Hypothesis Testing

Zhuanghua Shi (Strongway)

28 May 2018

Hypothesis testing

Today we mainly cover

1. t-distribution and t-tests
   - Simple t-test
   - Paired t-test
2. Analysis of Variance (ANOVA)
   - one-way ANOVA
   - two-way ANOVA
   - repeated-measures ANOVA

t-distribution

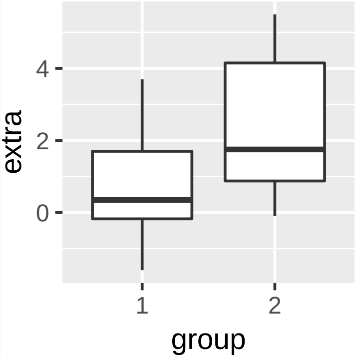
Suppose we got n samples (X_i) from a unknown population $N(\mu, \sigma^2)$, and we are interested in comparison the mean $\bar{X}$ to μ. Often we replace σ with the estimated standard deviation S, then $(\bar{X} - \mu)/(S/\sqrt{n})$ is a t-distribution with $n - 1$ degree of freedom.

- t-distribution is bell shaped with thicker tails than the normal distribution
- `t.test()` is the command

Example: a simple t-test

- `sleep` data from R datasets:
  - the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients

```
ggplot(sleep, aes(x=group,
    y = extra)) +
  geom_boxplot()
```



Example: a simple t-test

- A simple non-paired test using formula

```
stat_t = t.test(extra ~ group, data = sleep)
tidy(stat_t)
```

```
##   estimate estimate1 estimate2 statistic    p.value para
## 1    -1.58      0.75      2.33 -1.860813 0.07939414  17.
##   conf.high                  method alternative
## 1 0.2054832 Welch Two Sample t-test   two.sided
```

Example: a paired t-test

- For paired t-test you need to submit both x and y
  1. reshape the table using `tidyverse`
  2. using paired t-test(. . . , paired = T)

```
sleep %>% spread(group, extra, sep = '_') %>%
  t.test(.$group_1, .$group_2, paired = T, data = .) %>%
  tidy()
```

```
##   estimate statistic    p.value parameter  conf.low  con
## 1    -1.58 -4.062128 0.00283289         9 -2.459886 -0.7
##          method alternative
## 1 Paired t-test   two.sided
```

ANOVA Test

You need ANOVA when

- If you have number of groups > 2 (categorical variable)
- A single *continuous* dependent variable
- Separate, independent group of subjects
- more than 2 measures from same subjects (repeated measures ANOVA)

Null Hypothesis:

- All groups are equal

Alternative Hypothesis:

- At least one group has significant difference from the other

Variance partitioning in ANOVA

- Total variance can be divided into
  - Variability that can be attributed to differences between groups
  - Variability attributed to all other factors - within group variability

Source	SS	df	MS	F
A	$n\,\Sigma(Y_j - Y_T)^2$	a - 1	SS_A/df_A	$MS_A\,/MS_{S/A}$
S/A	$\Sigma(Y_{ij} - Y_j)^2$	a(n -1)	$SS_{S/A}/df_{S/A}$	-
Total	$\Sigma(Y_{ij} - Y_T)^2$	N - 1	-	-

Example: A simple ANOVA

- command aov()

```
sleep %>% aov(extra ~ group, data = .) %>%
  tidy()
```

```
##        term df  sumsq    meansq statistic    p.value
## 1     group  1 12.482 12.482000  3.462627 0.07918671
## 2 Residuals 18 64.886  3.604778        NA         NA
```

Example: A simple ANOVA

- Now we consider subjects ID as a random effects
    - using Error(ID) marked ID as random factor

```
sleep %>% aov(extra ~ group + Error(ID), data = .) ->aov1
summary(aov1)
```

```
##
## Error: ID
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals  9  58.08   6.453
##
## Error: Within
##           Df Sum Sq Mean Sq F value  Pr(>F)
## group      1 12.482  12.482    16.5 0.00283 **
## Residuals  9  6.808   0.756
```

R formula

R formulas are used in various modeling and statistics packages.

- A typical formula: y is a function of x, a, and b

  $y \sim x + a + b$

- The sepal width is a function of petal width, conditioned on species

  $Sepal.Width \sim Petal.Width \mid Species$

R forumula

Symbols used in formula

- + for adding independent variables
- − for removing terms
- `:` for interaction
- ∗ for crossing
- `%in%` for nesting

```
y ~ x1 - x2 # ignor x2

y ~ x1*x2 # same as y ~ x1 + x2 + x1:x2
```

Conditions for ANOVA

- Independence
- Approximate normality: distribution of the response variable shoudl be nearly normal within each group
- Equal variance: groups should have roughly equal variability

ezANOVA

- ez package by Michael Lawrence facilitates easy analysis of factorial experiments.
- It also contains a simulated data from Attention Network Test (ANT)

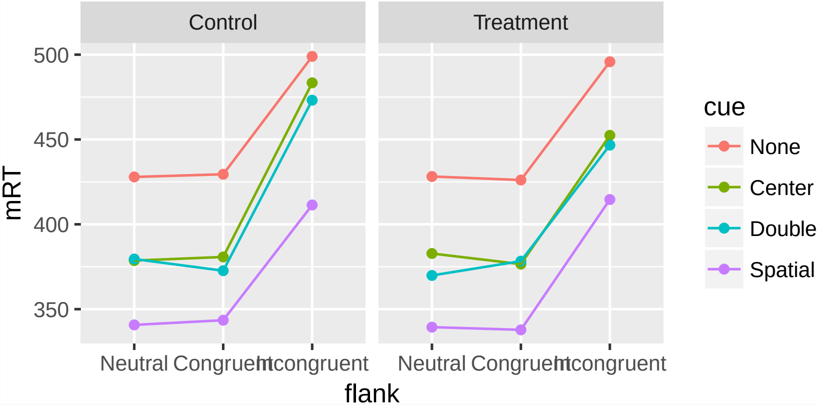
```
library(ez)
data(ANT)
ANT %>% dplyr::filter(error == 0) %>%
  group_by(group, cue, flank) %>%
  summarise(mRT = mean(rt)) -> mRTs
head(mRTs,3)
```

```
## # A tibble: 3 x 4
## # Groups:   group, cue [1]
##   group   cue   flank         mRT
```

ezANOVA and ANT

- Visualize the data

```
mRTs %>% ggplot(aes(flank, mRT, color = cue, group = cue)) +
  geom_point() + geom_line() + facet_wrap(~group)
```



ezANOVA and parameters

- ezANOVA parameters
  - data - data.frame table
  - dv - dependent variable
  - wid - subject id
  - within - within factors, multiple using .() list
  - between - between factors
  - between_covariates - covariates
- Return
  - Mauchly's test for specifity
  - Sphericity corrections
  - Levene's test for Homogeneity
  - AOV

Test on ANT data

```r
results <- ezANOVA(filter(ANT, error == 0),
                   rt, subnum, within=.(cue,flank),
                   between = group)
knitr::kable(results$ANOVA)
```

	Effect	DFn	DFd	F	p	p<.05	
2	group	1	18	18.430592	0.0004378	*	0
3	cue	3	54	516.605213	0.0000000	*	0
5	flank	2	36	1350.598810	0.0000000	*	0
4	group:cue	3	54	2.553236	0.0649749		0
6	group:flank	2	36	8.768499	0.0007901	*	0
7	cue:flank	6	108	5.193357	0.0000994	*	0
8	group:cue:flank	6	108	6.377225	0.0000090	*	0

Practice session

Now we apply those tests for the *search data*.