A Very Incomplete Survey of Basic Statistical Tests in R

Packages needed and a Note about Icons

Please load up the following packages (or install and then load them as needed)

```r
library(tidyverse)
library(car)
library(foreign)
library(lme4)
library(MASS)
library(CCA)
library(psych)
```

You may come across the following icons. The table below lists what each means.

Icon	Description
	Indicates that an example continues on the following slide.
	Indicates that a section using common syntax has ended.
	Indicates that there is an active hyperlink on the slide.
	Indicates that a section covering a concept has ended.

A Side Note About R

Its a big deal that you have come this far with R especially since it was rough at times. It may not be apparent, but developing coding skills like the ones in this course have benefits, not least of all in simply understanding the structure of a given data set. There are too many examples of students and even professionals who run an analysis on data without considering the data itself. R and other syntax-based software packages like it to their credit make you explore your data whether it be through checks or frustration.

While proprietary softwares such as SPSS, SAS, Minitab, etc. may be easier to learn, R and others like Python are free, open, and widley used. Picking on SPSS, as of this writing users pay for different tiers depending on needs (such as machine learning, methods to deal with missing data, etc.) they want to use. s

With that said, learning R is a lifelong process and assisting student learning and growth should never be confined to a single course so please FEEL FREE to contact me if you have questions regarding R (or Python if you go there) at any time. Again, I will always make time for students.

Purpose

This walk-through will provide you with information on how to perform a number of statistical tests using R. Some of these will look familiar while others you will be exposed to in future statistics courses if that is your path. In either case, hopefully these will be helpful if for no other reason than to provide a check or confirmation of results.

Decisions Decisions Decisions

When deciding which test is appropriate to use, it is important to consider the type of variables that you have. Please load in the following data sets (and look at them by using `View()` or `head()`

```r
some_ed_data <-
  read_csv("some_ed_data.csv")
```

```r
some_exercise_data <-
  read_csv("some_exercise_data.csv")
```

```r
some_survey_data <-
  read_csv("some_survey_data.csv")
```

Source

A majority of the information included in this survey of approached was scraped from the web using `R` via the UCLA Institute for Digital Research & Education site using the `xml2` package. They also fully support SAS, SPSS (for those of you moving on to EDP 614), Stata, and Mplus.

An Incomplete Table of Approaches (1/4)

Number of Dependent Variables	Number and Type of Independent Variables	Type of Dependent Variables	Test(s)
1	0 IVs (1 population)	interval & normal	one-sample t-test
1	0 IVs (1 population)	ordinal or interval	one-sample median
1	0 IVs (1 population)	categorical (2 categories)	binomial test
1	0 IVs (1 population)	categorical	Chi-square goodness-of-fit
1	1 IV with 2 levels (independent groups)	interval & normal	2 independent sample t-test
1	1 IV with 2 levels (independent groups)	ordinal or interval	Wilcoxon-Mann Whitney test
1	1 IV with 2 levels (independent groups)	categorical	Chi-square test
1	1 IV with 2 levels (independent groups)	categorical	Fisher’s exact test

An Incomplete Table of Approaches (2/4)

Number of Dependent Variables	Number and Type of Independent Variables	Type of Dependent Variables	Test(s)
1	1 IV with 2 or more levels (independent groups)	interval & normal	one-way ANOVA
1	1 IV with 2 or more levels (independent groups)	ordinal or interval	Kruskal Wallis
1	1 IV with 2 or more levels (independent groups)	categorical	Chi-square test
1	1 IV with 2 levels (dependent/matched groups)	interval & normal	paired t-test
1	1 IV with 2 levels (dependent/matched groups)	ordinal or interval	Wilcoxon signed ranks test
1	1 IV with 2 levels (dependent/matched groups)	categorical	McNemar
1	1 IV with 2 or more levels (dependent/matched groups)	interval & normal	one-way repeated measures ANOVA
1	1 IV with 2 or more levels (dependent/matched groups)	ordinal or interval	Friedman test

An Incomplete Table of Approaches (3/4)

Number of Dependent Variables	Number and Type of Independent Variables	Type of Dependent Variables	Test(s)
1	2 or more IVs (independent groups)	interval & normal	factorial ANOVA
1	2 or more IVs (independent groups)	ordinal or interval	ordered logistic regression
1	2 or more IVs (independent groups)	categorical (2 categories)	factorial logistic regression
1	1 interval IV	interval & normal	correlation
1	1 interval IV	interval & normal	simple linear regression
1	1 interval IV	ordinal or interval	non-parametric correlation
1	1 interval IV	categorical	simple logistic regression

An Incomplete Table of Approaches (4/4)

Number of Dependent Variables	Number and Type of Independent Variables	Type of Dependent Variables	Test(s)
1	1 or more interval IVs and/or 1 or more categorical IVs	interval & normal	multiple regression
1	1 or more interval IVs and/or 1 or more categorical IVs	interval & normal	analysis of covariance
1	1 or more interval IVs and/or 1 or more categorical IVs	categorical	multiple logistic regression
1	1 or more interval IVs and/or 1 or more categorical IVs	categorical	discriminant analysis
2+	1 IV with 2 or more levels (independent groups)	interval & normal	one-way MANOVA
2+	2+	interval & normal	multivariate multiple linear regression
2+	0	interval & normal	factor analysis
2 sets of 2+	0	interval & normal	canonical correlation

Tests

ANCOVA (Analysis of Covariance)

```
summary(aov(some_ed_data$write ~
              some_ed_data$prog +
              some_ed_data$read))
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)    
## some_ed_data$prog   1    586     586    10.2 0.00164 ** 
## some_ed_data$read   1   5965    5965   103.7 < 2e-16 ***
## Residuals         197  11327      57                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Binomial Test

```r
prop.test(sum(some_ed_data$female),
          length(some_ed_data$female),
          p = 0.5)
```

```
## 
## 	1-sample proportions test with continuity correction
## 
## data:  sum(some_ed_data$female) out of length(some_ed_data$female), null probability 0.5
## X-squared = 1.445, df = 1, p-value = 0.2293
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4733037 0.6149394
## sample estimates:
##     p 
## 0.545
```

Canonical Correlation

```
cc(cbind(some_ed_data$read,
         some_ed_data$write),
   cbind(some_ed_data$math,
         some_ed_data$science))
```

```
## $cor
## [1] 0.7728409 0.0234784
##
## $names
## $names$Xnames
## NULL
##
## $names$Ynames
## NULL
##
## $names$ind.names
## NULL
##
##
## $xcoef
##             [,1]       [,2]
## [1,] -0.06326131 -0.1037908
## [2,] -0.04924918  0.1219084
##
## $ycoef
```

Chi-square Test

```
chisq.test(table(some_ed_data$female,
                 some_ed_data$schtyp))
```

```
##
##     Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(some_ed_data$female, some_ed_data$schtyp)
## X-squared = 0.00054009, df = 1, p-value = 0.9815
```

Chi-square Goodness of Fit

```r
chisq.test(table(some_ed_data$race),
           p = c(10, 10, 10, 70)/100)
```

```
##
##      Chi-squared test for given probabilities
##
## data:  table(some_ed_data$race)
## X-squared = 5.0286, df = 3, p-value = 0.1697
```

Correlation

```r
cor(some_ed_data$read,
    some_ed_data$write)
```

```
## [1] 0.5967765
```

```r
cor.test(some_ed_data$read,
         some_ed_data$write)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  some_ed_data$read and some_ed_data$write
## t = 10.465, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4993831 0.6792753
## sample estimates:
##       cor 
## 0.5967765
```

Discriminant Analysis

```r
lda(factor(some_ed_data$prog) ~
      some_ed_data$read +
      some_ed_data$write +
      some_ed_data$math,
    data = some_ed_data)
```

```
## Call:
## lda(factor(some_ed_data$prog) ~ some_ed_data$read + some_ed_data$write +
##     some_ed_data$math, data = some_ed_data)
##
## Prior probabilities of groups:
##     1     2     3
## 0.225 0.525 0.250
##
## Group means:
##   some_ed_data$read some_ed_data$write some_ed_data$math
## 1          49.75556           51.33333          50.02222
## 2          56.16190           56.25714          56.73333
## 3          46.20000           46.76000          46.42000
##
## Coefficients of linear discriminants:
##                           LD1         LD2
## some_ed_data$read  0.02919876  0.04385321
## some_ed_data$write 0.03832289 -0.13698224
## some_ed_data$math  0.07034625  0.07931008
```

Factor Analysis

```r
fa(r = cor(model.matrix(~read + write + math + science + socst - 1,
                        data = some_ed_data)),
   rotate = "none",
   fm = "pa", 2)
```

```
## maximum iteration exceeded
```

```
## Factor Analysis using method =  pa
## Call: fa(r = cor(model.matrix(~read + write + math + science + socst - 
##     1, data = some_ed_data)), nfactors = 2, rotate = "none", 
##     fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PA1   PA2   h2   u2 com
## read    0.81  0.06 0.66 0.34 1.0
## write   0.76  0.00 0.58 0.42 1.0
## math    0.80  0.17 0.67 0.33 1.1
## science 0.75  0.26 0.62 0.38 1.2
## socst   0.79 -0.48 0.85 0.15 1.6
## 
##                        PA1  PA2
## SS loadings           3.06 0.33
## Proportion Var        0.61 0.07
## Cumulative Var        0.61 0.68
## Proportion Explained  0.90 0.10
## Cumulative Proportion 0.90 1.00
```

Factorial ANOVA (Analysis of Variance)

```
anova(lm(write ~ female * ses,
         data = some_ed_data))
```

```
## Analysis of Variance Table
##
## Response: write
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## female       1  1176.2 1176.21 14.7212 0.0001680 ***
## ses          1  1042.3 1042.32 13.0454 0.0003862 ***
## female:ses   1     0.0    0.04  0.0005 0.9827570
## Residuals  196 15660.3   79.90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Factorial Logistic Regression

```
summary(glm(female ~ prog * schtyp,
            data = some_ed_data,
            family = binomial))
```

```
##
## Call:
## glm(formula = female ~ prog * schtyp, family = binomial, data = some_ed_data)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -1.698  -1.247   1.069   1.109   1.572
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2765     1.8857  -1.207    0.227
## prog          1.2303     0.9398   1.309    0.191
## schtyp        2.2405     1.7017   1.317    0.188
## prog:schtyp  -1.1313     0.8622  -1.312    0.189
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 275.64  on 199  degrees of freedom
## Residual deviance: 273.65  on 196  degrees of freedom
## AIC: 281.65
##
```

Friedman Test

```r
friedman.test(cbind(some_ed_data$read,
                    some_ed_data$write,
                    some_ed_data$math))
```

```
##
##  Friedman rank sum test
##
## data:  cbind(some_ed_data$read, some_ed_data$write, some_ed_data$math)
## Friedman chi-squared = 0.64491, df = 2, p-value = 0.7244
```

Kruskal Wallis Test

```r
kruskal.test(some_ed_data$write,
             some_ed_data$prog)
```

```
##
##      Kruskal-Wallis rank sum test
##
## data:  some_ed_data$write and some_ed_data$prog
## Kruskal-Wallis chi-squared = 34.045, df = 2, p-value = 4.047e-08
```

McNemar Test

```r
# Some made up data in matrix form
made_up_matrixdata <-
  matrix(c(150, 22, 21, 12), 2, 2)

mcnemar.test(made_up_matrixdata)
```

```
##
##     McNemar's Chi-squared test with continuity correction
##
## data:  made_up_matrixdata
## McNemar's chi-squared = 0, df = 1, p-value = 1
```

Multiple Regression

```r
lm(some_ed_data$write ~
     some_ed_data$female +
     some_ed_data$read +
     some_ed_data$math +
     some_ed_data$science +
     some_ed_data$socst)
```

```
##
## Call:
## lm(formula = some_ed_data$write ~ some_ed_data$female + some_ed_data$read +
##     some_ed_data$math + some_ed_data$science + some_ed_data$socst)
##
## Coefficients:
##          (Intercept)   some_ed_data$female     some_ed_data$read
##               6.1388                5.4925                0.1254
##    some_ed_data$math  some_ed_data$science    some_ed_data$socst
##               0.2381                0.2419                0.2293
```

Multivariate Multiple Regression

```r
mmrlm <-
  lm(cbind(write, read) ~
       female + math + science + socst,
     data = some_ed_data)

summary(Anova(mmrlm))
```

```
##
## Type II MANOVA Tests:
##
## Sum of squares and products for error:
##          write     read
## write 7258.783 1091.057
## read  1091.057 8699.762
##
## ------------------------------------------
##
## Term: female
##
## Sum of squares and products for the hypothesis:
##           write       read
## write 1413.5284 -133.48461
## read  -133.4846   12.60544
##
## Multivariate Tests: female
```

Non-parametric Correlation

```r
cor.test(some_ed_data$read,
         some_ed_data$write,
         method = "spearman")
```

```
## Warning in cor.test.default(some_ed_data$read, some_ed_data$write, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  some_ed_data$read and some_ed_data$write
## S = 510993, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.6167455
```

One Sample *t*-test

```
t.test(some_ed_data$read,
       mu = 50)
```

```
##
##     One Sample t-test
##
## data:  some_ed_data$read
## t = 3.0759, df = 199, p-value = 0.002394
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  50.80035 53.65965
## sample estimates:
## mean of x
##     52.23
```

One-way Analysis of Variance (ANOVA)

```
summary(aov(some_ed_data$read ~
              some_ed_data$prog))
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## some_ed_data$prog   1    381   381.1   3.674 0.0567 .
## Residuals         198  20538   103.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-way Multivariate Analysis of Variance (MANOVA)

```
summary(manova(cbind(some_ed_data$read,
                     some_ed_data$write,
                     some_ed_data$math) ~
                 some_ed_data$prog))
```

```
##                    Df   Pillai approx F num Df den Df  Pr(>F)
## some_ed_data$prog   1 0.035319    2.392      3    196 0.06984 .
## Residuals         198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-way Repeated Measures Analysis of Variance (ANOVA)

```r
model <-
  lm(gender ~ item_1 + item_2,
     data = some_survey_data)

analysis <-
  Anova(model,
        idata = factor_surveydata,
        idesign = ~s)

print(analysis)
```

```
## Anova Table (Type II tests)
##
## Response: gender
##           Sum Sq Df F value Pr(>F)
## item_1    0.0601  1  0.2396 0.6307
## item_2    0.7268  1  2.8974 0.1069
## Residuals 4.2642 17
```

Ordered Logistic Regression

```r
# Create ordered variable write3 as
# a factor with levels 1, 2, and 3
some_ed_data$write3 <-
  cut(some_ed_data$write,
      c(0, 48, 57, 70),
      right = TRUE,
      labels = c(1,2,3))

table(some_ed_data$write3)
```

```
##
##  1  2  3
## 61 61 78
```

```r
# fit ordered logit model and store results 'some_write_data'
some_write_data <-
  polr(write3 ~
         female + read + socst, data = some_ed_data,
       Hess=TRUE)

summary(some_write_data)
```

```
## Call:
## polr(formula = write3 ~ female + read + socst, data = some_ed_data,
```

Principal Components Analysis (PCA)

```
princomp(formula = ~read + write + math + science + socst,
         data = some_ed_data)
```

```
## Call:
## princomp(formula = ~read + write + math + science + socst, data = some_ed_data)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 18.252929  7.677044  6.213371  5.774331  5.429881
##
##  5  variables and  200 observations.
```

Repeated Measures Logistic Regression

```r
glmer(highpulse ~ diet + (1 | id),
      data = some_exercise_data,
      family = binomial)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace  Approximation)
##  [glmerMod]
##  Family: binomial  ( logit )
## Formula: highpulse ~ diet + (1 | id)
##    Data: some_exercise_data
##      AIC      BIC   logLik deviance df.resid
## 105.4679 112.9674 -49.7340  99.4679       87
## Random effects:
##  Groups Name        Std.Dev.
##  id     (Intercept) 1.821
## Number of obs: 90, groups:  id, 30
## Fixed Effects:
## (Intercept)         diet
##      -3.148        1.145
```

Simple Linear Regression

```r
lm(some_ed_data$write ~
     some_ed_data$read)
```

```
##
## Call:
## lm(formula = some_ed_data$write ~ some_ed_data$read)
##
## Coefficients:
##       (Intercept)  some_ed_data$read
##           23.9594             0.5517
```

Simple Logistic Regression

```r
glm(some_ed_data$female ~
      some_ed_data$read,
    family = binomial)
```

```
##
## Call:  glm(formula = some_ed_data$female ~ some_ed_data$read, family = binomial)
##
## Coefficients:
##       (Intercept)  some_ed_data$read
##           0.72609           -0.01044
##
## Degrees of Freedom: 199 Total (i.e. Null);  198 Residual
## Null Deviance:       275.6
## Residual Deviance: 275.1     AIC: 279.1
```

Two Independent Samples *t*-test

```
t.test(some_ed_data$read ~
         some_ed_data$female)
```

```
##
##      Welch Two Sample t-test
##
## data:  some_ed_data$read by some_ed_data$female
## t = 0.74506, df = 188.46, p-value = 0.4572
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.796263  3.976725
## sample estimates:
## mean in group 0 mean in group 1
##        52.82418        51.73394
```

Wilcoxon-Mann-Whitney Test

```
wilcox.test(some_ed_data$read ~
              some_ed_data$female)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  some_ed_data$read by some_ed_data$female
## W = 5300, p-value = 0.4029
## alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon Signed Rank Sum Test

```r
wilcox.test(some_ed_data$write,
            some_ed_data$read,
            paired = TRUE)
```

```
## 
## 	Wilcoxon signed rank test with continuity correction
## 
## data:  some_ed_data$write and some_ed_data$read
## V = 9261, p-value = 0.3666
## alternative hypothesis: true location shift is not equal to 0
```

Other Approaches

There are so many other approaches that are for specific cases or use statistical approaches, but aren't themselves statistics. With that said, the ones given in this overview are overkill for most of you and should cover any statistics you come across.

Additional Things

Reporting

After running a statistical test successfully, it can be difficult to know how to report the results. The `report` package automatically produces reports of models and dataframes according to best practices guidelines. Click here for more information.

Visualizations

Interested in making incredible visuals? Check out #tidytuesday on Twitter. You do not need an account for access.

Something useless

If you are a fan of the show Rick & Morty, consider downloading the most pointless package `mortyr` to do pointless statistics on pointless data. More about the package here.

Thats it!