MAT 167: Statistics

Test I: Chapters 1-4

Instructor: Anthony Tanbakuchi

Spring 2008

Name: ____________________

Computer / Seat Number: ____________

No books, notes, or friends. **Show your work.** You may use the attached equation sheet, R, and a calculator. No other materials. Using any other program or having any other documents open on the computer will constitute cheating.

You have until the end of class to finish the exam, manage your time wisely.
If something is unclear quietly come up and ask me.
If the question is legitimate I will inform the whole class.

Express all final answers to 3 significant digits. Probabilities should be given as a decimal number unless a percent is requested. Circle final answers, ambiguous or multiple answers will not be accepted. Show steps where appropriate.

The exam consists of 8 questions for a total of 35 points on 7 pages.

This Exam is being given under the guidelines of our institution's
Code of Academic Ethics. You are expected to respect those guidelines.

Points Earned: ____________ **out of 35 total points**

Exam Score: ____________

1. Provide short written answers to the following conceptual questions.
   (a) (1 point) Is the range rule very susceptible to outliers?

   (b) (1 point) What percent of data lies within the IQR?

   (c) (1 point) What does the z-score represent in words?

   (d) (1 point) What does the standard deviation represent **in words**?

   (e) (1 point) A student needs to quantitatively describe the variation of the heights of students in a class. In comparison to variance, what important characteristic of standard deviation makes it more useful for communicating the amount of variation in the heights?

   (f) (2 points) Give an example of sampling error.

   (g) (1 point) If the mean, median, and mode for a data set are different, what can you conclude about the data's distribution?

2. A survey conducted in our class asked 27 students how far they travelled to school (in miles). Use the R output below to answer the following questions.

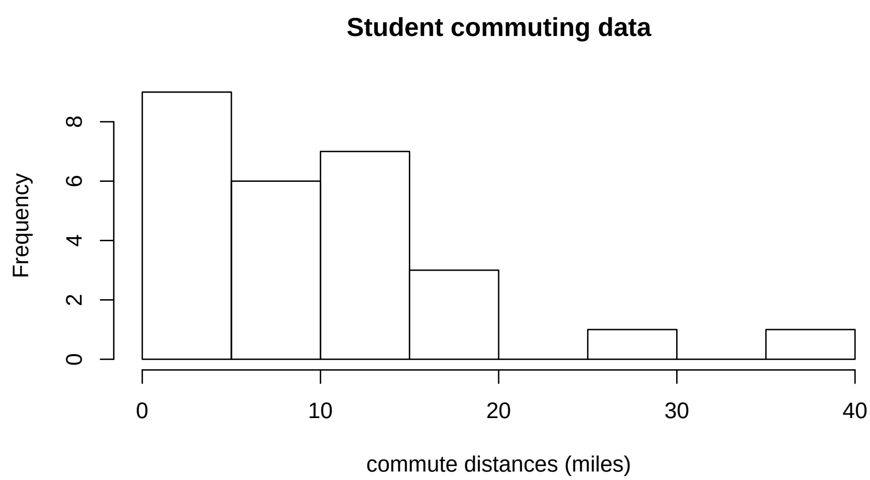
   There are 27 data points stored in the variable x, below is the sorted data:

```
> sort(x)
 [1]  0.1  0.1  3.5  4.0  4.5  5.0  5.0  5.0  5.0  6.0  6.0  7.3  8.0
9.0 10.0
[16] 11.8 12.0 12.5 13.0 13.0 13.0 15.0 16.0 20.0 20.0 27.0 40.0
```

   The basic descriptive statistical analysis is as follows:

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.10    5.00    9.00   10.81   13.00   40.00
> var(x)
[1] 73.91148
> sd(x)
[1] 8.597179

> hist(x, xlab = "commute distances (miles)", main = "Student commuting data")
```



   (a) (1 point) Use the range rule of thumb to estimate the standard deviation. Is it close to the actual standard deviation?

   (b) (1 point) What is P_{25} equal to?

   (c) (1 point) What is the IQR (inter quartile range) equal to?

(d) (1 point) For the student who commutes 4.5 miles to school, what is their approximate percentile?

(e) (1 point) What is the z-score for the student who commutes 40 miles to school?

(f) (1 point) Is 40 miles an unusual (outlier) distance based on it's z score?

(g) (1 point) Which measure of center would you use to describe this data? **Why**?

(h) (1 point) Is the data positively skewed, negatively skewed, or symmetrical?

(i) (1 point) Construct an interval using the Empirical Rule which you would expect 68% of the data to fall within.

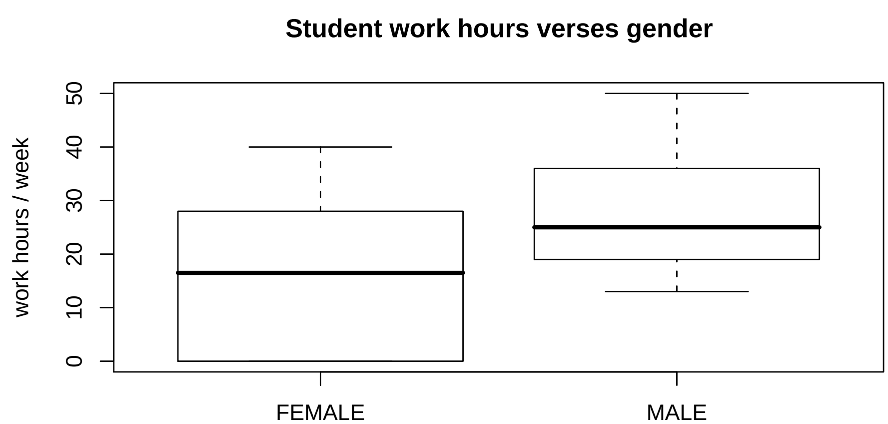
(j) (1 point) Would the Empirical Rule be appropriate to use for this data set? **Why?**

3. "The average commute distance of US community college students is 10.8 miles." This conclusion was reached by a student who had surveyed his statistics class.

   (a) (1 point) What type of sampling did the study use?

   (b) (1 point) Briefly state what is wrong with the student's conclusions.

4. Use the below box plot to answer the following questions.



   (a) (1 point) Which gender has a higher median number of work hours?

   (b) (1 point) What is the approximate median work hours / week for the females?

   (c) (1 point) Which gender has a larger variation in work hours for the middle 50% of individuals?

   (d) (1 point) What is the maximum hours per week observed for the male data?

5. Using the below table for our class to answer the following questions.

	BLACK	BLOND	BROWN	RED
FEMALE	1	5	12	2
MALE	1	0	5	1

(a) (1 point) Find the probability of selecting a person with red hair.

(b) (1 point) Would it be unusual to randomly select a person with red hair?

(c) (1 point) Find the probability of randomly selecting three males without replacement.

(d) (1 point) If you randomly select 5 people with replacement, what is the probability that at least one has red hair?

(e) (1 point) Find the probability of selecting a male student or a student with red hair.

(f) (1 point) Find the probability of selecting a person with red hair given that they are male.

6. (1 point) With one method of a procedure called acceptance sampling, a sample of items is randomly selected without replacement and the entire batch is accepted if every item in the sample is okay. The Niko Electronics Company has just manufactured 10,000 CDs, and 500 are defective. If 5 of the CDs are randomly selected for testing without replacement, what is the probability that the entire batch will be accepted?

7. Given the following frequency table summarizing data from a study:

age.years	frequency
0-9	5.00
10-19	8.00
20-29	12.00
30-39	2.00

   (a) (1 point) Construct a cumulative frequency table.

   (b) (1 point) What is the probability of randomly selecting someone from the study who 19 years or younger?

8. (2 points) Given $x = \{4c, 2c, -2c\}$, where c is a constant, completely simplify the following expression:

$$\sqrt{\frac{\sum(x_i - 2c)^2}{5}}$$

Basic Statistics: Quick Reference & R Commands

by Anthony Tanbakuchi. Version 1.7
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, ...)**
Get help on function: **?functionName**
Get column of data from table: **tableName$columnName**
List all variables: **ls()**
Delete all variables: **rm(list=ls())**

$$\sqrt{x} = \texttt{sqrt(x)} \quad (1)$$
$$x^n = \texttt{x\^{}n} \quad (2)$$
$$n = \texttt{length(x)} \quad (3)$$
$$T = \texttt{table(x)} \quad (4)$$

2 Descriptive Statistics

2.1 Numerical

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

$$\text{total} = \sum_{i=1}^{n} x_i = \texttt{sum(x)} \quad (5)$$
$$\text{min} = \texttt{min(x)} \quad (6)$$
$$\text{max} = \texttt{max(x)} \quad (7)$$
$$\text{six number summary}: \texttt{summary(x)} \quad (8)$$
$$\mu = \frac{\sum x_i}{N} = \texttt{mean(x)} \quad (9)$$
$$\bar{x} = \frac{\sum x_i}{n} = \texttt{mean(x)} \quad (10)$$
$$\tilde{x} = P_{50} = \texttt{median(x)} \quad (11)$$
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \quad (12)$$
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \texttt{sd(x)} \quad (13)$$
$$CV = \frac{\sigma}{\mu} = \frac{s}{\bar{x}} \quad (14)$$

2.2 Relative Standing

$$z = \frac{x-\mu}{\sigma} = \frac{x-\bar{x}}{s} \quad (15)$$

Percentiles:

$$P_k = x_i, \quad (\text{sorted } x)$$
$$k = \frac{i-0.5}{n} \cdot 100\% \quad (16)$$

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!)

$$P(A) = \frac{x_A}{n} \quad (17)$$
$$P(\bar{A}) = 1 - P(A) \quad (18)$$
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (19)$$
$$P(A \text{ or } B) = P(A) + P(B) \quad \text{if } A, B \text{ mut. excl.} \quad (20)$$
$$P(A \text{ and } B) = P(A) \cdot P(B|A) \quad (21)$$
$$P(A \text{ and } B) = P(A) \cdot P(B) \quad \text{if } A, B \text{ independent} \quad (22)$$
$$n! = n(n-1)\cdots 1 = \texttt{factorial(n)} \quad (23)$$
$${}_nP_k = \frac{n!}{(n-k)!} \quad \text{Perm. no elem. alike} \quad (24)$$
$$= \frac{n!}{n_1! n_2! \cdots n_k!} \quad \text{Perm. } n_1 \text{ alike}, \ldots \quad (25)$$
$${}_nC_k = \frac{n!}{(n-k)!k!} = \texttt{choose(n,k)} \quad (26)$$

4 Discrete Random Variables

$$P(x_i): \text{ probability distribution} \quad (27)$$
$$E = \mu = \sum x_i \cdot P(x_i) \quad (28)$$
$$\sigma = \sqrt{\sum (x_i - \mu)^2 \cdot P(x_i)} \quad (29)$$

4.1 Binomial Distribution

$$\mu = n \cdot p \quad (30)$$
$$\sigma = \sqrt{n \cdot p \cdot q} \quad (31)$$
$$P(x) = {}_nC_x p^x q^{(n-x)} = \texttt{dbinom(x, n, p)} \quad (32)$$

4.2 Poisson Distribution

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

5 Continuous random variables

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

$$f(x): \text{ probability density} \quad (34)$$
$$E = \mu = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \quad (35)$$
$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx} \quad (36)$$
$$F(x): \text{ cumulative prob. density (CDF)} \quad (37)$$
$$F^{-1}(x): \text{ inv. cumulative prob. density} \quad (38)$$
$$F(x) = \int_{-\infty}^{x} f(x')\, dx' \quad (39)$$
$$p = P(x < x') = F(x') \quad (40)$$
$$x' = F^{-1}(p) \quad (41)$$
$$p = P(x > a) = 1 - F(a) \quad (42)$$
$$p = P(a < x < b) = F(b) - F(a) \quad (43)$$

5.1 Uniform distribution

$$p = P(u < u') = F(u') = \texttt{punif(u', min=0, max=1)} \quad (44)$$
$$u' = F^{-1}(p) = \texttt{qunif(p, min=0, max=1)} \quad (45)$$

5.2 Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \quad (46)$$
$$p = P(z < z') = F(z') = \texttt{pnorm(z')} \quad (47)$$
$$z' = F^{-1}(p) = \texttt{qnorm(p)} \quad (48)$$
$$p = P(x < x') = F(x') = \texttt{pnorm(x', mean=}\mu\texttt{, sd=}\sigma\texttt{)} \quad (49)$$
$$x' = F^{-1}(p) = \texttt{qnorm(p, mean=}\mu\texttt{, sd=}\sigma\texttt{)} \quad (50)$$

5.3 t-distribution

$$p = P(t < t') = F(t') = \texttt{pt(t', df)} \quad (51)$$
$$t' = F^{-1}(p) = \texttt{qt(p, df)} \quad (52)$$

5.4 χ^2-distribution

$$p = P(\chi^2 < \chi^{2\prime}) = F(\chi^{2\prime}) = \texttt{pchisq(}X^{2\prime}\texttt{, df)} \quad (53)$$
$$\chi^{2\prime} = F^{-1}(p) = \texttt{qchisq(p, df)} \quad (54)$$

5.5 F-distribution

$$p = P(F < F') = F(F') = \texttt{pf(F', df1, df2)} \quad (55)$$
$$F' = F^{-1}(p) = \texttt{qf(p, df1, df2)} \quad (56)$$

6 Sampling distributions

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (57)$$
$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \quad (58)$$

7 Estimation

7.1 Confidence Intervals

$$\text{proportion: } \hat{p} \pm E, \quad E = z_{\alpha/2} \cdot \sigma_{\hat{p}} \quad (59)$$
$$\text{mean } (\sigma \text{ known}): \bar{x} \pm E, \quad E = z_{\alpha/2} \cdot \sigma_{\bar{x}} \quad (60)$$
$$\text{mean } (\sigma \text{ unknown, use } s): \bar{x} \pm E, \quad E = t_{\alpha/2} \cdot \sigma_{\bar{x}}, \quad (61)$$
$$df = n-1$$
$$\text{variance: } \frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}, \quad (62)$$
$$df = n-1$$
$$\text{2 proportions: } \Delta\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} \quad (63)$$
$$\text{2 means (indep): } \Delta\bar{x} \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad (64)$$
$$df \approx \min(n_1 - 1, n_2 - 1)$$
$$\text{matched pairs: } \bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}, \quad d_i = x_i - y_i, \quad (65)$$
$$df = n-1$$

7.2 CI Critical Values (two sided)

$$z_{\alpha/2} = F_z^{-1}(1-\alpha/2) = \texttt{qnorm(1-alpha/2)} \quad (66)$$
$$t_{\alpha/2} = F_t^{-1}(1-\alpha/2) = \texttt{qt(1-alpha/2, df)} \quad (67)$$
$$\chi_L^2 = F_{\chi^2}^{-1}(\alpha/2) = \texttt{qchisq(alpha/2, df)} \quad (68)$$
$$\chi_R^2 = F_{\chi^2}^{-1}(1-\alpha/2) = \texttt{qchisq(1-alpha/2, df)} \quad (69)$$

7.3 Required sample size

$$\text{proportion: } n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2, \quad (70)$$
$$(\hat{p} = \hat{q} = 0.5 \text{ if unknown})$$
$$\text{mean: } n = \left(\frac{z_{\alpha/2} \cdot \hat{\sigma}}{E}\right)^2 \quad (71)$$

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

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

11 Loading external data

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