# MAT 167: Statistics

## Test II

### Instructor: Anthony Tanbakuchi

### Fall 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. If you choose to use R, write what you typed on the test or copy and paste your work into a word document labeling the question number it corresponds to. When you are done with the test print out the document. Be sure to save often on a memory stick just in case. 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 18 questions for a total of 70 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 70 total points**

**Exam Score:** _____

1. (1 point) Why is it important to use random sampling?

2. For the following statements, determine if the calculation requires the use of a **population distribution** or a **sampling distribution**.

   (a) (1 point) Computing a confidence interval for a mean.

   (b) (1 point) Determining a p-value for a one sample proportion hypothesis test.

   (c) (1 point) Calculating the probability than an individual weights more than 100 lbs.

   (d) (1 point) Computing an interval that contain 95% of individual's weights.

3. (1 point) What type of error does a sampling distribution characterize?

4. (1 point) Under what conditions can we approximate a binomial distribution as a normal distribution?

5. (1 point) If the normal approximation to the binomial is valid, write what the following binomial probability statement is approximately equal to in terms of the normal distribution.

$$P_{\text{binom}}(10 < x \le 15) \approx$$

6. In regards to $\bar{x}$ and the Central Limit Theorem:

   (a) (2 points) What are the two conditions under which the CLT applies?

   (b) (2 points) If the conditions are met, what type of distribution will $\bar{x}$ have?

Instructor: Anthony Tanbakuchi                    Points earned: _____ / 12 points

7. (1 point) Which distribution (normal, binomial, both, or neither) would be appropriate for describing:

   The distribution for the sample mean incomes when taking a random sample of 15 individuals and the distribution of incomes has a strong positive skew.

8. (1 point) For the one sample proportion hypothesis test, what is the distribution of the test statistic? (Give the specific name.)

9. Let $x$ be a random variable with a normal distribution where $\mu = 20$ and $\sigma = 2$.

   (a) (2 points) Make a meaningful sketch that represents $P(x > 23)$.

   (b) (2 points) Find $P(x > 23)$.

   (c) (1 point) Would it be unusual to observe $x > 23$?

10. The following questions regard hypothesis testing in general.

    (a) (1 point) When we conduct a hypothesis test, we assume something is true and calculate the probability of observing the sample data under this assumption. What do we assume is true?

    (b) (1 point) If you are using a hypothesis test to make a decision where the effect of a Type I error may negatively effect human lives, should you increase or decrease $\alpha$?

    (c) (1 point) You reject $H_0$ but $H_0$ is true. What type of error has occurred? (Type I or Type II)

(d) (1 point) What variable represents the actual Type I error.

(e) (1 point) Two studies were conducted, study A had a power of 0.9 and study B had a power of 0.10. Which study would be more likely to support a true alternative hypothesis?

(f) (1 point) A researcher takes a sample, conducts a hypothesis test, and rejects the null hypothesis since the p-value was sufficiently small. The researcher concludes that "the sample data proves that the mean height of men is greater than 5.5 feet." What is wrong with this conclusion?

11. (2 points) What is a normal Q-Q plot used for?

12. (2 points) Eight different second-year medical students at Bellevue Hospital measured the blood pressure of the same person and the results are shown below.

$$138, 130, 135, 140, 120, 125, 120, 130$$

Construct a 95% confidence interval estimate for the mean blood pressure assuming the data has a normal distribution.

13. (2 points) A hypothesis test was conducted for $H_0 : p = 0.5$ and $H_a : p \neq 0.5$. The test statistic is $z = 1.8$. Find the p-value.

14. For women aged 18-24, systolic blood pressure (in mm Hg) are normally distributed with a mean of 114.8 and a standard deviation of 13.1. (Based on data from a National Health Survey). Hypertension is commonly defined as a systolic blood pressure above 140.

   (a) (2 points) If a woman between the ages of 18 and 24 is randomly selected, find the probability that her systolic blood pressure is greater than 140.

   (b) (2 points) A doctor tells a female patent who is in the age range of 18 to 24 that her systolic blood pressure is in the $25^{\text{th}}$ percentile. What is her blood pressure?

   (c) (2 points) If 4 women are randomly selected and their mean blood pressure is computed, what type of distribution would the sample means have and **why**?

   (d) (2 points) If 4 women in that age bracket are randomly selected, find the probability that their mean systolic blood pressure is greater than 140.

15. (2 points) The music industry must adjust to the growing practice of consumers downloading songs instead of buying CDs. It therefore becomes important to estimate the proportion of songs that are currently downloaded. How many randomly selected song purchases must be surveyed to determine the percentage that were obtained by downloading? Assume that we want to be 95% confident that the sample percentage is within one percentage point of the true population percentage of songs that are downloaded.

16. You believe that the true mean head breadth for men is greater than 6.0 in. A study of 8 randomly selected men (shown below) was conducted to test this claim. Use a significance level of 0.025 and assume that men's head breadths are normally distributed.

$$6.2, \ 6.7, \ 5, \ 7.6, \ 7, \ 7, \ 6.8, \ 6.3$$

   (a) (1 point) What type of hypothesis test will you use?

   (b) (2 points) What are the test's requirements?

   (c) (1 point) Are the requirements satisfied? **State how they are satisfied.**

   (d) (2 points) What are the hypothesis?

   (e) (1 point) What $\alpha$ will you use?

   (f) (2 points) Conduct the hypothesis test. What is the $p$-value?

   (g) (1 point) What is your formal decision?

   (h) (2 points) State your final conclusion in words.

   (i) (1 point) If we reject $H_0$, what is the *actual* probability of a Type I error for this study data?

17. A pharmaceutical company has developed a new drug that they believe can be used to increase the probability a student will pass a statistics exam. To test the drug's effectiveness, they randomly select 50 students, give them the drug, and have them take the AP statistics exam of which 30 students pass. For those students who did not get the drug, only 58% passed. The drug company hopes to support the claim that the new drug increases the standard passing rate of 58%.

   (a) (1 point) What type of hypothesis test will you use?

   (b) (2 points) What are the test's requirements?

   (c) (2 points) What are the hypothesis?

   (d) (1 point) What $\alpha$ will you use?

(e) (2 points) What is the $p$-value.

(f) (1 point) What is your formal decision?

(g) (2 points) State your final conclusion in words.

18. Over the past 55 years, data from the National Oceanic and Atmospheric Administration (NOAA) indicates the the probability that the minimum daily temperature is at most 32 degrees fahrenheit[1] on a given day in Tucson is 0.048. (Use 365 days in a year.)

   (a) (1 point) Find the mean number of days per year with a minimum daily temperature of at most 32 degrees fahrenheit.

   (b) (1 point) Find the standard deviation for the number of days per year with a minimum daily temperature of at most 32 degrees fahrenheit.

   (c) (2 points) What is the probability of 20 or more days with a minimum daily temperature of at most 32 degrees fahrenheit.

   (d) (1 point) Would it be unusual to have 20 or more days with a minimum daily temperature of at most 32 degrees fahrenheit? (Why)

<div align="center">

************

End of exam. Reference sheets follow.

************

</div>

---

[1]Data from `http://www.wrcc.dri.edu/cgi-bin/clilcd.pl?az23160`.

# Statistics Quick Reference Card & R Commands

by Anthony Tanbakuchi. Version 1.8.1
http://www.tanbakuchi.com
ANTHONY@TANBAKUCHI.COM
Get R at: http://r-project.org
R commands: **bold typewriter text**

## 1 Misc R

To make a vector / store data: **x=c(x1, x2, ...)**
Help: general **RSiteSearch("Search Phrase")**
Help: function **?functionName**
Get column of data from table:
**tableName$columnName**
List all variables: **ls()**
Delete all variables: **rm(list=ls())**

$$\sqrt{x} = \textbf{sqrt(x)} \tag{1}$$
$$x^n = \textbf{x\^n} \tag{2}$$
$$n = \text{length(x)} \tag{3}$$
$$T = \textbf{table(x)} \tag{4}$$

## 2 Descriptive Statistics

### 2.1 NUMERICAL

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

$$\text{total} = \sum_{i=1}^{n} x_i = \textbf{sum(x)} \tag{5}$$
$$\min = \textbf{min(x)} \tag{6}$$
$$\max = \textbf{max(x)} \tag{7}$$

six number summary : **summary(x)** (8)

$$\mu = \frac{\sum x_i}{N} = \textbf{mean(x)} \tag{9}$$
$$\bar{x} = \frac{\sum x_i}{n} = \textbf{mean(x)} \tag{10}$$
$$\tilde{x} = P_{50} = \textbf{median(x)} \tag{11}$$
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \tag{12}$$
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \textbf{sd(x)} \tag{13}$$
$$CV = \frac{\sigma}{\mu} = \frac{s}{\bar{x}} \tag{14}$$

### 2.2 RELATIVE STANDING

$$z = \frac{x - \mu}{\sigma} = \frac{x - \bar{x}}{s} \tag{15}$$

Percentiles:

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

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.

## 3 Probability

Number of successes $x$ with $n$ possible outcomes.
(Don't double count!)

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

## 4 Discrete Random Variables

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

### 4.1 BINOMIAL DISTRIBUTION

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

### 4.2 POISSON DISTRIBUTION

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

## 2.3 VISUAL

All plots have optional arguments:

- **main=""** sets title
- **xlab="", ylab=""** sets x/y-axis label
- **type="p"** for point plot
- **type="l"** for line plot
- **type="b"** for both 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)**

## 5 Continuous random variables

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

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

### 5.1 UNIFORM DISTRIBUTION

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

### 5.2 NORMAL DISTRIBUTION

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

### 5.3 $t$-DISTRIBUTION

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

### 5.4 $\chi^2$-DISTRIBUTION

$$p = P(\chi^2 < \chi^{2'}) = F(\chi^{2'})$$
$$= \textbf{pchisq(}X^{2'}\textbf{, df)} \tag{53}$$
$$\chi^{2'} = F^{-1}(p) = \textbf{qchisq(p, df)} \tag{54}$$

### 5.5 $F$-DISTRIBUTION

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

## 6 Sampling distributions

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

## 7 Estimation

### 7.1 CONFIDENCE INTERVALS

proportion: $\hat{p} \pm E$, $E = z_{\alpha/2} \cdot \sigma_{\hat{p}}$ (59)
mean ($\sigma$ known): $\bar{x} \pm E$, $E = z_{\alpha/2} \cdot \sigma_{\bar{x}}$ (60)
mean ($\sigma$ unknown, use s): $\bar{x} \pm E$, $E = t_{\alpha/2} \cdot \sigma_{\bar{x}}$, (61)
$$df = n - 1$$
variance: $\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$, (62)
$$df = n - 1$$
2 proportions: $\Delta \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$ (63)
2 means (indep): $\Delta \bar{x} \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, (64)
$$df \approx \min(n_1 - 1, n_2 - 1)$$
matched pairs: $\bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$, $d_i = x_i - y_i$, (65)
$$df = n - 1$$

### 7.2 CI CRITICAL VALUES (TWO SIDED)

$$z_{\alpha/2} = F^{-1}(1 - \alpha/2) = \textbf{qnorm(1-alpha/2)} \tag{66}$$
$$t_{\alpha/2} = F^{-1}(1 - \alpha/2) = \textbf{qt(1-alpha/2, df)} \tag{67}$$
$$\chi_L^2 = F^{-1}(\alpha/2) = \textbf{qchisq(alpha/2, df)} \tag{68}$$
$$\chi_R^2 = F^{-1}(1 - \alpha/2) = \textbf{qchisq(1-alpha/2, df)} \tag{69}$$

### 7.3 REQUIRED SAMPLE SIZE

proportion: $n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$, (70)
$(\hat{p} = \hat{q} = 0.5$ if unknown)
mean: $n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$ (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"** or **"one.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=$\mu_0$, 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 − β, 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) and nc=c($n_1$, $n_2$)**
where: **x=c($x_1$, $x_2$) and nc=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 − β, 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 − β, 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.

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}}, \quad d_i = x_i - y_i, \quad df = n - 1 \tag{78}$$

Required Sample size:
**power.t.test(delta=$h$, sd =σ, sig.level=α, power=1 − β, 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 |
| ...interaction | $y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$ | 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 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 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 and using external data and tables

## 11.1 LOADING EXCEL DATA
1. Export your table as a CSV file (comma separated value) from Excel.
2. Import your table into MyTable in R using:
   **MyTable=read.csv(file.choose())**

## 11.2 LOADING AN .RDATA FILE
You can either double click on the .RData file or use the menu:
  • Windows: *File→Load Workspace...*
  • Mac: *Workspace→Load Workspace File...*

## 11.3 USING TABLES OF DATA
1. To see all the available variables type: **ls()**
2. To see what's inside a variable, type its name.
3. If the variable tableName is a table, you can also type **names(tableName)** to see the column names or **head(tableName)** to see the first few rows of data.
4. To access a column of data type **tableName$colName**

An example demonstrating how to get the women's height data and find the mean:

```
> ls() # See what variables are defined
[1] "women" "x"
> head(women) #Look at the first few entries
  height weight
1     58    115
2     59    117
3     60    120
> names(women) # Just get the column names
[1] "height" "weight"
> women$height # Display the height data
 [1] 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
> mean(women$height) # Find the mean of the heights
[1] 65
```