MAT 167: Statistics

Final Exam

Instructor: Anthony Tanbakuchi

Summer 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. Show your work in the space provided. 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 and turn it in with the test. 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 16 questions for a total of 80 points on 9 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 80 total points**

Exam Score: _____

1. The following is a partial list of statistical methods that we have discussed:

   1. mean
   2. median
   3. mode
   4. standard deviation
   5. z-score
   6. percentile
   7. coefficient of variation
   8. scatter plot
   9. histogram
   10. pareto chart
   11. box plot
   12. normal-quantile plot
   13. confidence interval for a mean

   14. confidence interval for difference in means
   15. confidence interval for a proportion
   16. confidence interval for difference in proportions
   17. one sample mean test
   18. two independent sample mean test
   19. one sample proportion test
   20. two sample proportion test
   21. test of homogeneity
   22. test of independence
   23. linear correlation coefficient & test
   24. regression
   25. 1-way ANOVA

   For each situation below, which method is most applicable?
   - If it's a hypothesis test, **also state what the null and alternative hypothesis are**.
   - If it's a graphical method, **also describe what you would be looking for**.
   - If it's a statistic, how susceptible to outliers is it?

   (a) (2 points) A researcher at the department of labor wants to determine if the proportion of men who work in sociology, psychology, and anthropology is the same from recent study data on the subject.

   (b) (2 points) A math department committee wants to award $50 to the student who received the best score on their calculus final exam this past semester. However, the three faculty who taught calculus last semester gave different final exams. What method could help identify the top student amongst the three different exams?

   (c) (2 points) A drug research would like to test the claim that the mean absorption of 1 gram of vitamin E is the same for four methods of delivery: topical, intravenous, oral, and nasal

Instructor: Anthony Tanbakuchi   Points earned: _____ / 6 points

spray.

(d) (2 points) A researcher is using a statistical hypothesis test that requires the population which the sample was drawn from to have a normal distribution. How could the researcher check this assumption?

(e) (2 points) A drug researcher wants to determine if a new growth hormone drug can increase the mean weight of mice as compared to a control group of mice.

2. (1 point) The test of homogeneity can be thought of as a generalization of what two sample test?

3. (1 point) If the mean, median, and mode for a data set are all the same, what can you conclude about the data's distribution?

4. (2 points) Under what conditions can we approximate a binomial distribution as a normal distribution?

5. (2 points) Give **a clear specific example** of when you would use a population distribution.

6. (2 points) Give **a clear specific example** example of when you would use a sampling distribution.

7. (1 point) What percent of data lies within one standard deviation as stated by the Empirical Rule?

8. 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) Do we use the population distribution or the sampling distribution when calculating the p-value?

    (c) (1 point) If you fail to reject H_0 but H_0 is false. 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) What variable is used to represent a Type II error?

    (f) (1 point) What does the power of a hypothesis test represent?

    (g) (1 point) In the one sample mean test with σ unknown, what is the distribution of the test statistic?

    (h) (2 points) Why is it important to use random sampling?

9. A consumer advocate group believes that Crystal Springs Sparkling Mineral Water contains more than the advertised 35 mg of sodium per serving. They randomly sample 40 servings

Instructor: Anthony Tanbakuchi   Points earned: _____ / 12 points

and measure the amount of sodium contained in each sample. The collected data has a sample mean of 37.5 mg, and a sample standard deviation of 8.2 mg.

(a) (2 points) What is the null and alternative hypothesis?

(b) (1 point) What hypothesis test should be used?

(c) (2 points) What are the requirements for the hypothesis test? Are they met?

(d) (1 point) What significance level will you use?

(e) (2 points) Manually compute the p-value.

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

(g) (2 points) What is the final conclusion?

10. A craps table at a local casino has been losing more money than normal. It seems that bets involving a one on the face of the dice (such as "snake eyes") are appearing more than usual. The casino manager thinks that the dice have been weighted to cause the side with one to have a higher probability of occurring than a fair dice.

(a) (2 points) The casino manager takes one of the dice from the table and flips it 100 times, the side with a value of one appears 26 times. Construct a 95% confidence interval for the

true probability of getting a one with this die.

(b) (1 point) If the die is fair, what should the true probability be for getting a one?

(c) (1 point) Why can't the casino manager conclude that the die is unfair simply based on the fact that he observed one appear more than it theoretically should have appeared? What other explanation could account for this?

(d) (1 point) Now the casino manager would like to conduct a hypothesis test to determine if the die is unfair. What type of hypothesis test should he use?

(e) (2 points) What are the manager's hypothesis for the test?

(f) (2 points) What are the test's requirements? Are they satisfied?

(g) (2 points) Conduct the hypothesis test, what is the p-value?

(h) (1 point) What is the formal decision?

(i) (1 point) What is the conclusion?

(j) (1 point) What is the probability of a Type I error for this study?

11. Eighteen students were randomly selected to take the SAT after having either no breakfast or a complete breakfast A researcher would like to test the claim that students who eat breakfast score higher than students who do not.

Group without breakfast: SAT Score	480	510	530	540	550	560	600	620	660
Group with breakfast: SAT Score	460	500	530	520	580	580	560	640	690

(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 H_0 and H_a?

(d) (1 point) What α will you use?

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

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

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

12. The following table lists the the fuel consumption (in miles/gallon) and weight (in lbs) of a
vehicle.

Weight	3180	3450	3225	3985	2440	2500	2290
MPG	27	29	27	24	37	34	37

(a) (2 points) Upon looking at the scatter plot of the data, the relationship of fuel consumption
and milage looks linear. Is the linear relationship statistically significant? (**Justify your
answer with an analysis.**)

(b) (1 point) What percent of a vehicle's fuel consumption can be explained by its weight?

(c) (2 points) You are designing a new vehicle and would like to be able to predict its fuel
consumption. Write the equation for fitted model (with the actual values of the coeffi-
cients).

(d) (1 point) What range of vehicle weights is the model valid for making predictions of fuel
efficiency?

(e) (1 point) What is the best predicted fuel consumption for a new vehicle that weights 3200
lbs?

(f) (1 point) If the liner relationship had not been statistically significant, what is the best
predicted fuel consumption for a new vehicle that weights 3200 lbs?

Instructor: Anthony Tanbakuchi   Points earned: _____ / 10 points

13. (2 points) A random sample of 5 chihuahuas was conducted to determine the mean tong length of the breed. Below is the study data in inches.

$$1.6, \ 2.1, \ 1.5, \ 1.9, \ 2$$

Construct a 90% confidence interval for the true population mean using the above data. (**Assume σ is unknown.**)

14. (2 points) A ski resort is designing a new super tram to carry 40 people. If the mean weight of humans is approximately 165 lbs with a standard deviation of 25 lbs, what should the tram's maximum weight limit be so that it can carry the desired capacity 95% of the time?

15. (2 points) You would like to conduct a study to estimate (at the 90% confidence level) the mean weight of brown bears with a margin of error of 5 lbs. A preliminary study indicates that bear weights are normally distributed with a standard deviation of 22 lbs, what sample size should you use for this study?

16. (2 points) Given $y = \{a, -2a, 4a\}$, where a is a constant, completely simplify the following expression:

$$\left(\sum y_i\right)^2 - 2$$

****
End of exam. Reference sheets follow.

Instructor: Anthony Tanbakuchi   Points earned: _____ / 4 points

Statistics Quick Reference Card & R Commands

by Anthony Tanbakuchi. Version 1.8.2

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())**

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

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)

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

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.

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 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!)

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

4 Discrete Random Variables

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

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

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

5 Continuous random variables

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

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

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 χ^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 (σ known): $\bar{x} \pm E$, $E = z_{\alpha/2} \cdot \sigma_{\bar{x}}$ (60)

mean (σ unknown, use s): $\bar{x} \pm E$, $E = t_{\alpha/2} \cdot \sigma_{\bar{x}}$, (61)
$$df = n - 1$$

variance: $\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$, (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: $\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$, $d_i = x_i - y_i$, (65)
$$df = n - 1$$

7.2 CI CRITICAL VALUES (TWO SIDED)

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

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")`

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \quad (72)$$

8.2 1-SAMPLE MEAN (σ KNOWN)
$H_0 : \mu = \mu_0$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad (73)$$

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.

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}, \quad df = n - 1 \quad (74)$$

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 `n=c(n_1, n_2)`
where: `x=c(x_1, x_2)` and `n=c(n_1, n_2)`

$$z = \frac{\Delta \hat{p} - \Delta p_0}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}} \quad \Delta \hat{p} = \hat{p}_1 - \hat{p}_2 \quad (75)$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}, \quad \hat{q} = 1 - \hat{p} \quad (76)$$

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.

$$t = \frac{\Delta \bar{x} - \Delta \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df \approx \min(n_1 - 1, n_2 - 1), \quad \Delta \bar{x} = \bar{x}_1 - \bar{x}_2 \quad (77)$$

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 vectors of sample 1 and sample 2 data.

$$t = \frac{\bar{d} - \mu_{d_0}}{s_d / \sqrt{n}}, \quad d_i = x_i - y_i, \quad df = n - 1 \quad (78)$$

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()`
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.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad df = (\text{num rows - 1})(\text{num cols - 1}) \quad (79)$$

$$E_i = \frac{(\text{row total})(\text{column total})}{(\text{grand total})} = np_i \quad (80)$$

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.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \quad t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}}, \quad df = n - 2 \quad (81)$$

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	$y = b_0 + b_1 x_1 + b_2 x_1^2$	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

$$y = b_0 + b_1 x_1 \quad (82)$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (83)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (84)$$

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

$$F = \frac{MS(\text{treatment})}{MS(\text{error})}, \quad df_1 = k - 1, df_2 = N - k \quad (85)$$

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 file) 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)` or type `head(tableName)` to see the first few rows of data.
4. To access a column of data type `tableName$columnName`

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
```