MAT 167: Statistics

Test II: Chapters 4-7

Instructor: Anthony Tanbakuchi

Fall 2007

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, 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 4 questions for a total of 25 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 25 total points**

Exam Score: __________

1. Assume that men's waists are normally distributed with $\mu = 35$ in and $\sigma = 2.3$ in.
   (a) (1 point) If 1 man is randomly selected, find the probability that his waist size is greater than 34 in.

   (b) (1 point) If 20 men are randomly selected, find the probability that their mean waist size is less than 34 in.

   (c) (1 point) You are designing sweat pants that are "one size fits all". In reality, the pants only stretch out to fit the bottom 90% of the male waist sizes, what is the maximum waist size that the pants will stretch to?

2. The clothing manufacturer's association (CMA) publishes data that manufacture's use to determine what sizes of clothing they should make. As mentioned before, the CMA states that men's waists are normally distributed with $\mu = 35$ in and $\sigma = 2.3$ in. Lately, you are getting a lot of returns on your one size fit's all sweat pants (that you designed in a previous question) because they are too small.

   (a) (1 point) You would like to conduct a study to estimate (at the 95% confidence level) the mean waist size of men with a margin of error of 1 in. Assuming that the standard deviation of waist sizes is $\sigma = 2.3$ in, what sample size should you use for this study?

   (b) (1 point) A study was conducted (and they ignored your recommendation of sample size!) of 5 randomly selected men and the following waist sizes were measured:

   36.8 38.3 37.8 40 38.3

   Construct a 95% confidence interval for the true population mean waist size using the above data. (**Assume σ is unknown.**)

   (c) (1 point) Why is the margin of error in your calculated confidence interval greater than our original desired margin of error of 1 in?

3. As mentioned before, the CMA states that men's waists are normally distributed with $\mu = 35$ in. Lately, you are getting allot of returns on your one size fit's all sweat pants (that you designed in a previous question) because they are too small.

   You believe that the mean waist size of men is actually greater than 35 in. Using the same study data from the previous question of 5 randomly sampled men (shown below again), conduct a hypothesis test to test your claim. Use a significance level of 0.01 and **assume σ is unknown**.

   36.8 38.3 37.8 40 38.3

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

   (b) (1 point) What are the test's requirements?

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

   (d) (1 point) What are the hypothesis?

   (e) (1 point) What α will you use?

   (f) (1 point) Conduct the hypothesis test. What is the p-value?

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

(h) (1 point) State your final conclusion in words.

(i) (1 point) What is the *actual* probability of a Type I error for this study data?

(j) (1 point) Assume a Type I error has occurred, state what the formal decision was and what the error is.

(k) (1 point) If the researcher had an α of 0.005 and failed reject H_0, what have they proven?

(l) (1 point) Assume a Type II error has occurred, state what the formal decision was and what the error is.

(m) (1 point) In general, 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?

4. A newly married couple does not want to get pregnant during their first year of marriage and decides to only use condoms as a contraceptive device. Studies[1] have shown that the probability of pregnancy per use of a condom is 0.15% (individual probability of pregnancy). During the first year the couple use a condom every time for a total of 162 times.

   (a) (1 point) Find the probability of 0 pregnancies over 1 year.

   (b) (1 point) Find the probability of 1 or more pregnancies over 1 year.

   (c) (1 point) Would it be unusual to become pregnant during 1 year of use when only using this form of contraceptive?

   (d) (1 point) What would be the average number of pregnancies that the couple should expect to have given the above information?

   (e) (1 point) Using the equation for the mean of the binomial distribution, solve for the number of uses n required to have an average (mean) of 1 pregnancy.

   (f) (1 point) Based on your above results, if you were the primary care physician for this couple, what would you tell them about their plan to prevent pregnancy? Would it be effective?

[1] The data presented in this problem are approximate values derived from actual published data based on *typical* use effectiveness.

Basic Statistics: Quick Reference & R Commands

by Anthony Tanbakuchi. Version 1.4
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())**

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

2 Descriptive Statistics

2.1 Numerical

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

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

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 **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!)

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

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

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

4.2 Poisson Distribution

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

5 Continuous random variables

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

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

5.1 Uniform Distribution

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

5.2 Normal Distribution

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

5.3 t-Distribution

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

5.4 χ^2-Distribution

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

5.5 F-Distribution

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

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

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

7.2 CI Critical Values (two sided)

$$z_{\alpha/2} = P(z > \alpha) = \texttt{qnorm(1-alpha/2)} \tag{66}$$
$$t_{\alpha/2} = P(t > \alpha) = \texttt{qt(1-alpha/2, df)} \tag{67}$$
$$\chi_L^2 = P(\chi^2 < \alpha) = \texttt{qchisq(alpha/2, df)} \tag{68}$$
$$\chi_R^2 = P(\chi^2 > \alpha) = \texttt{qchisq(1-alpha/2, df)} \tag{69}$$

7.3 Required Sample Size

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

8 Hypothesis Tests

Test statistic and R function (when available) are listed for each.
Optional arguments:
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"`.

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.

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

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**)**

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

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}, \quad \bar{q} = 1 - \bar{p} \tag{76}$$

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.

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

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.

$$t = \frac{\bar{d} - \mu_{d0}}{s_d / \sqrt{n}}, \quad d_i = x_i - y_i, \quad df = n - 1 \tag{78}$$

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.

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

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

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.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \quad t = \frac{r - \rho_0}{\sqrt{\frac{1-r^2}{n-2}}}, \quad df = n - 2 \tag{81}$$

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`

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

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())