SOLUTIONS
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 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.

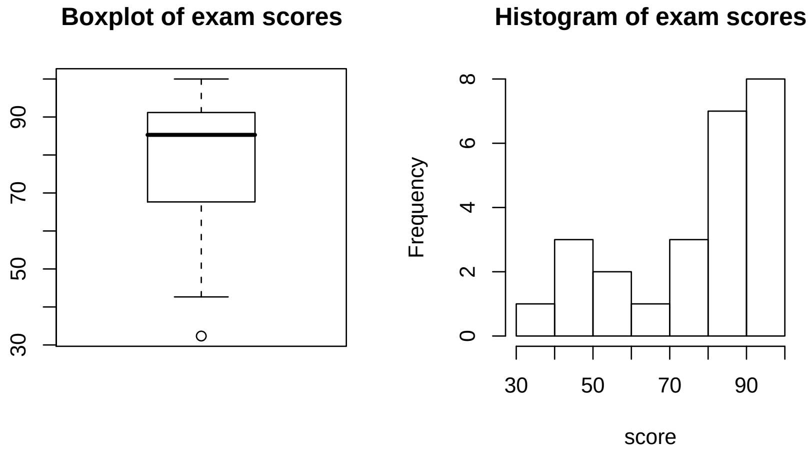
Points Earned: __________ **out of 35 total points**

Exam Score: __________

Solution:

Spring 2008 results. Since question 1(e) may have been ambiguous and under emphasized in class I dropped it. However, make sure you know the answer in the future. The exam score was out of 34 points. Below are the summary statistics.

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.35   67.65   85.29   77.29   91.18  100.00
> par(mfrow = c(1, 2))
> boxplot(score, main = ”Boxplot of exam scores”)
> hist(score, main = ”Histogram of exam scores”)
```



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

   Solution: Yes, it is highly susceptible because it only uses the min and max value. If an outlier exists, it will be the min or max value causing a large change in the range rule estimate.

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

   Solution: 50%

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

   Solution: The z-score represents the number of standard deviations the data point lies from the mean.

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

Solution: The standard deviation represents the average variation of the data from the mean.

(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?

Solution: Standard deviation has the same units as the data whereas variance has squared units. Therefore, the standard deviation would measure the variation of the heights in **inches** whereas the variance would measure the variation in **inches**^{2}.

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

Solution: Sampling errors are errors caused by chance fluctuations. An example would be randomly sampling 10 people and computing their mean height. If you performed this random sample multiple times you would find that the mean height would vary somewhat from sample to sample. Sampling error is the natural fluctuation due to sampling, it is not a human error or misuse of statistics as a non-sapling error would be.

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

Solution: If all three measures of center different, the data is skewed.

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.

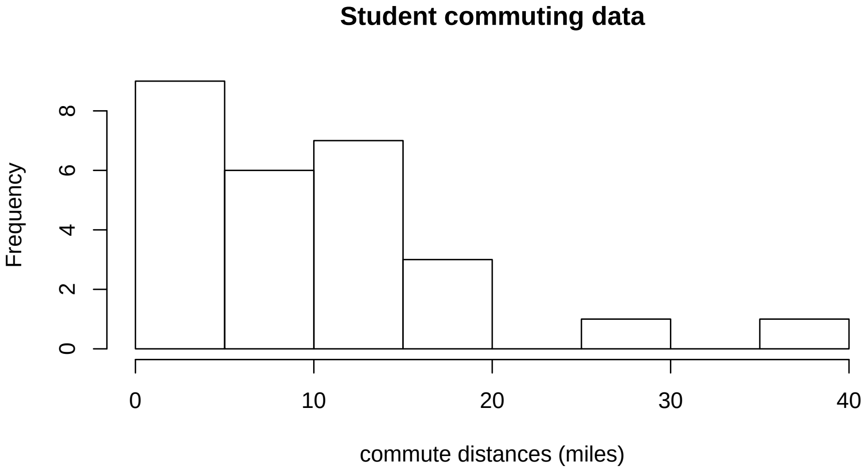
   There are 27 data points stored in the variable x, below is the sorted data:

```
> 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?

   Solution:

```
> s.est = (max(x) - min(x))/4
> signif(s.est, 3)
[1] 9.98
```

   The range rule estimate is close to the actual standard deviation shown in the output.

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

Solution: You can find the value in the summary output for the 1st quarter, it is equal to 5.

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

Solution: $IQR = Q_3 - Q_1$, using the summary output, the IQR is 8.

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

Solution: The percentile is the percent of data points less than the given data point, in this case there are 4 values less than the data point, so the percentile is $4/27 = 15\%$ or P_{15}.

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

Solution:

```
> x.bar
[1] 10.80741
> s
[1] 8.597179
> z = (40 - x.bar)/s
> signif(z, 3)
[1] 3.4
```

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

Solution: Yes, since $|z| > 2$ it is unusual.

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

Solution: Use a measure of center that is resistant to outliers. Since the the data is continuous, the mode would not work, so the next best choice would be the **median**.

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

Solution: Positively skewed.

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

Solution: 68% of the data falls within $\mu \pm \sigma$ which for this data set is: $10.8 \pm 8.6 = (2.21, 19.4)$.

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

Solution: No. The Empirical Rule describes bell shaped symmetrical data (normally distributed). Since this data is quite skewed, the Empirical Rule would not provide accurate estimates.

Points earned: ______ / 1 points

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.

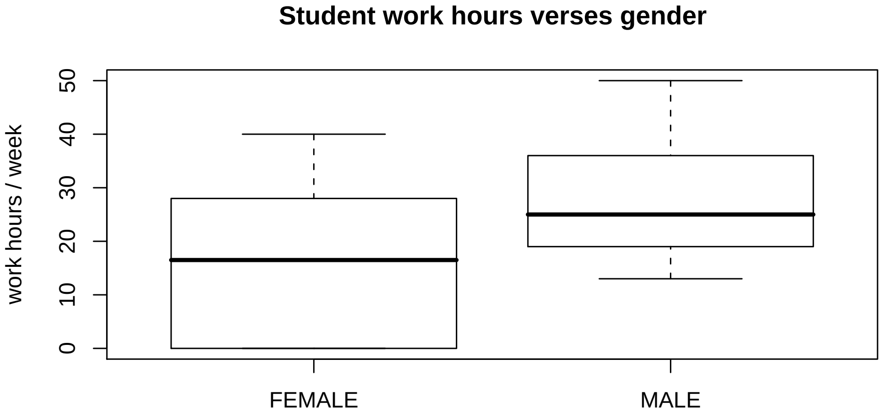
   (a) (1 point) What type of sampling did the study use?

   Solution: Convenience sampling.

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

   Solution: The study is not representative of all US community college students. It only represents the students in one statistics class.

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



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

   Solution: Males

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

   Solution: About 17 hours / week

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

   Solution: Females

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

   Solution: 50 hours/week

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.

Solution:
$P(red) = \frac{3}{27} = 0.111$

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

Solution: No since $P \nleq 0.05$.

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

Solution:
$P(\text{male \& male \& male}) = \frac{7}{27} \cdot \frac{6}{26} \cdot \frac{5}{25} = 0.012$

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

Solution:
$P(\text{at least one red}) = 1 - P(\text{none red}) = 1 - P(\text{not red})^5 = 1 - \left(\frac{24}{27}\right)^5 = 0.445$

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

Solution:
$P(\text{red or male}) = \frac{9}{27} = 0.333$

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

Solution:
$P(\text{red}|\text{male}) = \frac{1}{7} = 0.143$

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?

Solution: Since $n/N \leq 0.05$ we can simplify this problem by treating it as independent even thought it is sampling without replacement. Find probability that all 5 CDs are good.

```
> n = 5
> N = 10000
> p.good = (10000 - 500)/10000
> p = p.good^5
> signif(p, 3)
[1] 0.774
```

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.

Solution:

cumulative.age	cumulative.frequency
0-9	5.00
0-19	13.00
0-29	25.00
0-39	27.00

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

Solution:

```
> p = (13/27)
> signif(p, 3)
[1] 0.481
```

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

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

Solution:

$$\begin{aligned}
\sqrt{\frac{\sum(x_i - 2c)^2}{5}} &= \sqrt{\frac{(4c-2c)^2 + (2c-2c)^2 + (-2c-2c)^2}{5}} \\
&= \sqrt{\frac{4c^2 + 0 + 16c^2}{5}} \\
&= \sqrt{\frac{20c^2}{5}} \\
&= \sqrt{4c^2} \\
&= 2c
\end{aligned}$$

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())`**

$$\sqrt{x} = \texttt{sqrt(x)} \tag{1}$$

$$x^n = \texttt{x^n} \tag{2}$$

$$n = \texttt{length(x)} \tag{3}$$

$$T = \texttt{table(x)} \tag{4}$$

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

$$\begin{aligned} p &= P(u < u') = F(u') \\ &= \texttt{punif(u', min=0, max=1)} \end{aligned} \tag{44}$$

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

5.2 Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{46}$$

$$p = P(z < z') = F(z') = \texttt{pnorm(z')} \tag{47}$$

$$z' = F^{-1}(p) = \texttt{qnorm(p)} \tag{48}$$

$$\begin{aligned} p &= P(x < x') = F(x') \\ &= \texttt{pnorm(x', mean=}\mu\texttt{, sd=}\sigma\texttt{)} \end{aligned} \tag{49}$$

$$\begin{aligned} x' &= F^{-1}(p) \\ &= \texttt{qnorm(p, mean=}\mu\texttt{, sd=}\sigma\texttt{)} \end{aligned} \tag{50}$$

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

$$\begin{aligned} p &= P(\chi^2 < \chi^{2'}) = F(\chi^{2'}) \\ &= \texttt{pchisq(}X^{2}\texttt{', df)} \end{aligned} \tag{53}$$

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

5.5 F-distribution

$$\begin{aligned} p &= P(F < F') = F(F') \\ &= \texttt{pf(F', df1, df2)} \end{aligned} \tag{55}$$

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

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)

$$z_{\alpha/2} = F_z^{-1}(1-\alpha/2) = \texttt{qnorm(1-alpha/2)} \tag{66}$$

$$t_{\alpha/2} = F_t^{-1}(1-\alpha/2) = \texttt{qt(1-alpha/2, df)} \tag{67}$$

$$\chi_L^2 = F_{\chi^2}^{-1}(\alpha/2) = \texttt{qchisq(alpha/2, df)} \tag{68}$$

$$\chi_R^2 = F_{\chi^2}^{-1}(1-\alpha/2) = \texttt{qchisq(1-alpha/2, df)} \tag{69}$$

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 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.

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

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

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.

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

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())**