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. Write your work in the provided space for each problem (including any R work if appropriate). You may not use personal computers, only use the classroom computer at your desk. 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 10 questions for a total of 46 points on 10 pages.

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

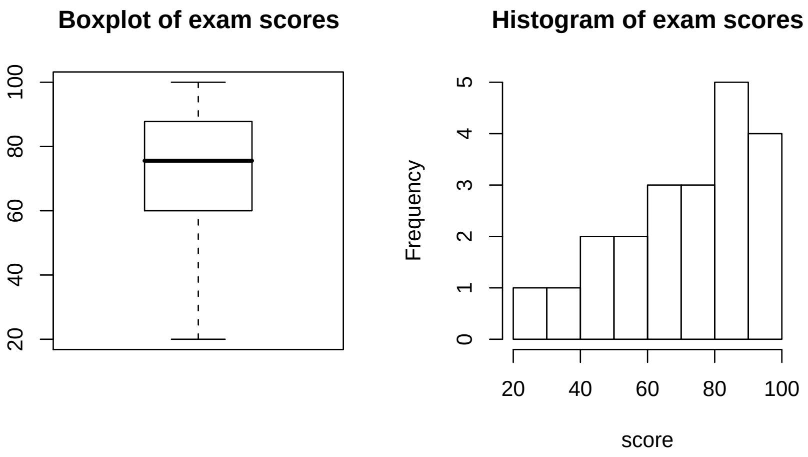
Points Earned: ____________ **out of 46 total points**

Exam Score: ____________

Solution:

Exam Results. Below are the summary statistics.

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   60.00   75.56   71.53   87.78  100.00
> par(mfrow = c(1, 2))
> boxplot(score, main = "Boxplot of exam scores")
> hist(score, main = "Histogram of exam scores")
```



1. Provide **short succinct** written answers to the following conceptual questions.

   (a) (1 point) Would mass in grams be classified as a nominal, ordinal, interval, or ratio level of measurement?

   Solution: Mass is a ratio level of measurement since 0 mass means there is no mass at all.

   (b) (1 point) Which of the following measures of variation is most effected by outliers:

   IQR, standard deviation, range

   Solution: The range is most effected by outliers

   (c) (1 point) What percent of data is greater than Q_3?

   Solution: 25%

   (d) (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?

Solution: If all three measures of center are the same, the distribution is symmetrical. (Not necessarily a normal distribution, all we know is that it is symmetrical.)

(e) (1 point) If the median is greater than the mode for a data set, what can you conclude about the data's distribution?

Solution: The data set is positively skewed.

(f) (1 point) What does the standard deviation represent conceptually **in words**? (Be concise but don't simply state the equation in words verbatim.)

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

(g) (1 point) Give an example of sampling error?

Solution: Use a simple random sample of 5 students from a class to estimate the mean weight of students in the class. Each time you repeat the process, you don't necessarily get the same sample mean because you don't always have the same people in the sample.

(h) (2 points) A box plot is a useful tool that can quickly communicate many traits about a set of data. List 4 useful pieces of information that an observer can easily assess using a box plot.

Solution: A box plot can be used to get an approximation of:

1. central tendency
2. variation in the data
3. IQR, Q_1, Q_2, median
4. shape of the data
5. assess if outliers exist
6. min
7. max

(i) (1 point) You scored in the 78th percentile on the GRE. If 8,000 people took the GRE, how many people had a score at least as high as your score?

Solution: If you were in the 78th, then 78% of the people scored lower than you and 22% did at least as well as you. Therefore, $8000 \cdot 0.22 = 1760$ people who had a score at least as high as yours.

2. A survey conducted in our class asked 18 students how many hours they work per week. Use the R output below to answer the following questions.

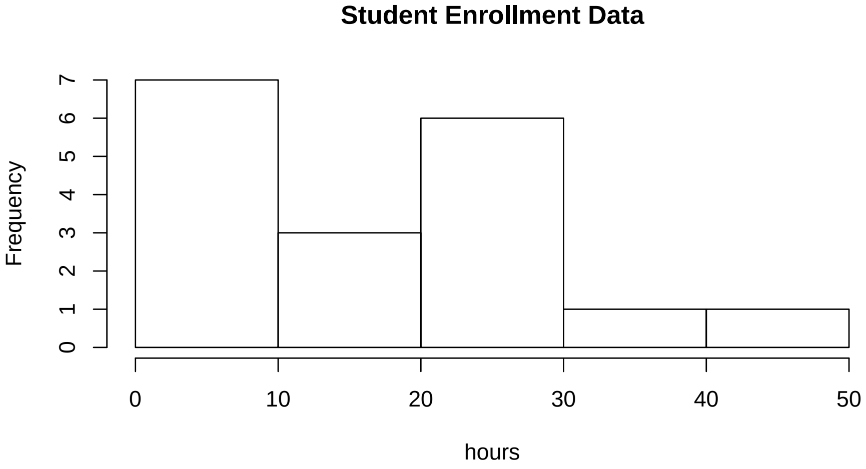
   There are 18 data points stored in the variable x, below is the sorted data:

```
> sort(x)
 [1]  0  0  0  0  8  8 10 12 13 20 21 30 30 30 30 30 35 50
```

The basic descriptive statistical analysis is as follows:

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   16.50   18.17   30.00   50.00
> var(x)
[1] 215.6765
> sd(x)
[1] 14.68593

> hist(x, xlab = "hours", main = "Student Enrollment 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] 12.5
```

The range rule is a rough estimate of σ, it is relatively close to the actual standard deviation shown in the output.

(b) (1 point) What is P_{50} equal to?

Solution: You can find the value in the summary output for the 1st quarter, it is equal to the median 16.5.

(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 22.

(d) (1 point) What percent of the data falls within the IQR?

Solution: The IQR contains the data between Q_1 and Q_3 or equivalently P_{25} and P_{75}, thus 50% of the data falls inside the IQR.

(e) (1 point) What is the mode for the data?

Solution: The mode is 30 hours per week (it occurs 5 times).

(f) (1 point) What is the percentile for a student who works 8 hours per week?

Solution: The percentile is the percent of data points less than the given data point. A student working 8 hours per week could be position 5 or 6, use $i = 5.5$, thus:

$$\begin{aligned} P_k &= x_i, \quad \text{(sorted } x\text{)} \\ k &= \frac{i - 0.5}{n} \cdot 100\% \\ &= \frac{5.5 - 0.5}{18} \cdot 100\% \\ &= 27.8\% \end{aligned}$$

(g) (1 point) What is the z-score for the student who is working 50 hours per week?

Solution:

$$z = \frac{x - \bar{x}}{s}$$

```
> x.bar
[1] 18.16667
> s
[1] 14.68593
> z = (50 - x.bar)/s
> signif(z, 3)
[1] 2.17
```

(h) (1 point) Is 35 hours an unusual (outlier) value based on its z score? (Why)

Solution:

```
> x.bar
[1] 18.16667
> s
```

```
[1] 14.68593
> z = (35 - x.bar)/s
> signif(z, 3)
[1] 1.15
```

No, since $|z| \not< 2$.

(i) (2 points) If someone asked what the typical number of hours worked per week for this class was, why would just reporting the mean be misleading? **What would you report in addition to the mean?**

Solution: Since the mean can be biased by outliers, reporting the mean might mislead someone to think that most students work quite a few hours per week. You also need to give people an idea how much the number of hours per week varies. You should also report the standard deviation or you could use the Empirical Rule to report an interval that contains approximately 95% of the class work hours.

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

Solution: Since the histogram has a longer right tail, it is positively skewed.

(k) (1 point) Construct an interval using the Empirical Rule which you would expect 95% of the data to fall within.

Solution: 95% of the data falls within $\mu \pm 2\sigma$ which for this data set is: $18.2 \pm 29.4 = (-11.2, 47.5)$.

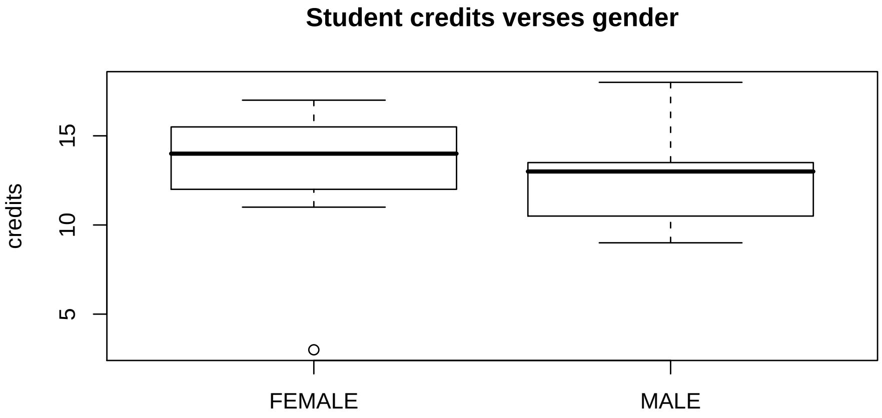
(l) (1 point) Would the Empirical Rule be accurate when used for this data set? **Why?**

Solution: No. The Empirical Rule describes bell shaped symmetrical data (normally distributed). Since this data is quite skewed, the Empirical Rule would not provide an accurate estimate of an interval that contains 95% of the data.

3. (1 point) When plotting frequencies tabulated from categorical data, why would a bar graph be preferable to a pie chart?

Solution: Bar graphs present the frequencies in terms of bar heights, pie charts present the frequencies in wedges of area. Humans can easily compare lengths (the bar height), but they cannot compare areas as easily. A bar graph makes interpreting the data straight forward.

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



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

Solution: Females

(b) (1 point) What is the approximate median number of credits for the females?

Solution: 14

(c) (1 point) What is the approximate IQR of credits for the males?

Solution: $Q_1 \approx 10$, $Q_3 \approx 10$, $\therefore IQR = 3$

(d) (1 point) Which gender has a larger range of credits?

Solution: Females (min$\approx$ 4, max$\approx$ 16, $\therefore$ range$\approx$ 16). Many students forgot to take into account the outlier and claimed that the males had a larger range. Don't ignore the outliers, they are part of the range.

(e) (1 point) How many outliers are there in this data set as indicated by the box plots?

Solution: The circles indicate outliers, there is 1.

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

	BLACK	BLOND	BROWN
CAR	0	2	11
PUBLIC	2	0	0
TRUCK	1	0	2

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

Solution:
$P(\text{brown}) = \frac{13}{18} = 0.722$

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

Solution: No since $P \nleq 0.05$.

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

Solution:
$P(\text{car \& car \& car}) = \frac{13}{18} \cdot \frac{12}{17} \cdot \frac{11}{16} = 0.35$

(d) (1 point) If you randomly select 4 people with replacement, what is the probability that at least one uses public transportation?

Solution:
$P(\text{at least one public}) = 1 - P(\text{none public}) = 1 - P(\text{not public})^4 = 1 - \left(\frac{16}{18}\right)^4 = 0.376$

(e) (1 point) Find the probability of selecting a student who drives a truck or a student with brown hair.

Solution:
$P(\text{brown or truck}) = \frac{14}{18} = 0.778$

(f) (1 point) Find the probability of selecting a person with blond hair given that they are a truck driver.

Solution:
$P(\text{blond}|\text{truck}) = \frac{0}{3} = 0$

6. A researcher needs to estimate the mean height of corn in a field containing 10,000 corn plants. The corn plants are planted 100 to a row, and there are 100 rows. The researcher wants to take a sample of 100 corn plant heights.

   (a) (1 point) Describe a method for sampling this population that would be classified as a convenience sample.

   Solution: The researcher could sample 100 plants in the first row.

   (b) (1 point) Describe a method for sampling this population that would be classified as a simple random sample.

   Solution: The researcher could assign each corn plant a different number and write each number on a card. Then 100 cards could be picked after shuffling the cards sufficiently to ensure they are in a random order. Preferably, a computer could be used to select 100 of the numbers randomly using a sufficiently randomized number generator.

7. The weather report for this work week (Monday through Friday) states that the probability of rain is 5% for each day.

   (a) (1 point) What is the probability that it will rain at least once this week?

   Solution: The events are independent.
   $P(\text{at least a day of rain}) = 1 - P(\text{none}) = 1 - P(\text{no rain on a day})^5 = 1 - 0.95^5 = 0.226$

   (b) (1 point) What is the probability that it won't rain this week?

   Solution: The events are independent.
   $P(\text{no rain}) = P(\text{no rain on 1 day})^5 = 0.95^5 = 0.774$
   This is just the compliment of the previous question.

   (c) (1 point) If you decided not to use an umbrella this week, would it be unusual for you to get wet?

   Solution: No, since $P(\text{at least 1 day of rain}) \nleq 0.05$

8. (2 points) Car tires must not deform or explode when inflated up to their maximum pressure rating. Before distributing the tires, they must be tested. To test the safety of tires, an inspector randomly samples 50 tires (without replacement) from a batch of 5,000 that have been manufactured. The inspector inflates each of the fifty tires until they explode or deform to make sure they meet the minimum safety requirements. If none of the sampled tires fails the test, the tires will be distributed to dealers. If the batch contains 15 defective tires that will explode if selected, what is the probability that the batch will be rejected?

Solution: We are randomly selecting $n = 50$ tires from $n = 5000$. Since we are sampling without replacement these are dependent trials, but $n/N \leq 0.05$ so we can simplify the problem by approximating it as independent.

$$\begin{aligned} P(\text{batch rejected}) &= P(\text{at least one tire defective}) \\ &= 1 - P(\text{None defective out of 50}) \\ &= 1 - P(\text{not defective})^{50} && \text{approx. as indep} \\ &= 1 - (1 - 15/5000)^{50} \\ &= 0.139 \end{aligned}$$

9. (2 points) The quadratic mean is defined as

$$\text{quadratic mean} = \sqrt{\frac{\sum x_i^2}{n}}$$

find the quadratic mean of x. If $x = \{2, 8, 7, 4, 9\}$

Solution:

```
> x = c(2, 8, 7, 4, 9)
> n = length(x)
> n
 [1] 5
> quad.mean = sqrt(sum(x^2)/n)
> signif(quad.mean, 3)
 [1] 6.54
```

10. (2 points) Given $y = \{a, -2a, 4a\}$, completely simplify the following expression. Assume a is an unknown constant.

$$\frac{\left(\sum(y_i - 2a)\right)^2}{9a}$$

Solution:

$$\begin{aligned}\frac{\left(\sum(y_i - 2a)\right)^2}{9a} &= \frac{(a - 2a + -2a - 2a + 4a - 2a)^2}{9a} \\ &= \frac{(-3a)^2}{9a} \\ &= a\end{aligned}$$

Statistics Quick Reference Card & R Commands

by Anthony Tanbakuchi. Version 1.8.2
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, ...)`**
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())`**

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

2 Descriptive Statistics

2.1 Numerical

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

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

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

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

7.2 CI Critical Values (two sided)

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

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

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

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 seperated 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)** to see the column names 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
```