SOLUTIONS
MAT 167: Statistics

Final Exam

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. 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 12 questions for a total of 70 points on 11 pages.

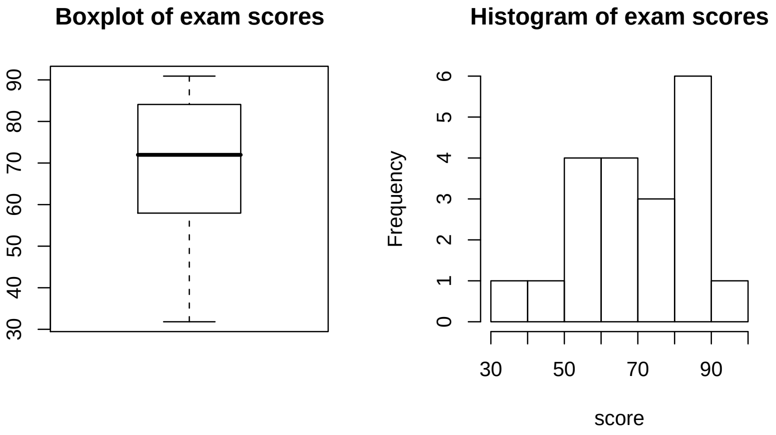
This Exam is being given under the guidelines of our institution's **Code of Academic Ethics**. You are expected to respect those guidelines.

Points Earned: ____________ **out of 70 total points**

Exam Score: ____________

Solution: Spring 2008 results.

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  31.82   58.14   71.97   69.32   83.71   90.91
> par(mfrow = c(1, 2))
> boxplot(score, main = "Boxplot of exam scores")
> hist(score, main = "Histogram of exam scores")
```



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. match pair test
    20. one sample proportion test
    21. two sample proportion test
    22. test of homogeneity
    23. test of independence
    24. linear correlation coefficient & test
    25. regression
    26. 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) Ten pairs of chicks were selected to test the effect of a vitamin supplement on early growth. The chicks in each pair were siblings of high birth weight. One chick in each pair was given the supplement and the other was not. After two weeks, the weight of each chick was recorded. The researcher would like to test the research hypothesis that the supplement increases the growth rate of chicks in the first weeks after hatching against the null hypothesis that it has no effect.

> **Solution:** Matched pair hypothesis test. $H_0 : \mu_d = 0$ supplement does not change growth rate, $H_a : \mu_d > 0$ (assuming the groups that gets the supplements is group 1, or x) supplement increases growth rate. (This is a matched pair test because *pairs* of chicks who were siblings. Since siblings are not independent, they would not be considered independent.)

(b) (2 points) A researcher would like to compare the distribution of incomes of individuals in the four major regions of the US: the west, the south, the midwest, the east.

> **Solution:** A box plot. A box plot makes it easy to compare distributions for multiple categories.
>
> Histograms are not easy to compare since you can't easily put many of them side by side. Just looking at the standard deviation wouldn't give you a full picture of each category's distribution.

(c) (2 points) A researcher wants determine if the mean income in the four major regions of the US — the west, the south, the midwest, the east — are not all the same.

> **Solution:** 1-way ANOVA. $H_0 : \mu_{\text{west}} = \mu_{\text{south}} = \mu_{\text{midwest}} = \mu_{\text{east}}$ income is the same in all 4 regions, H_a : income is different in at least one region.

(d) (2 points) An investigator is interested in the success of a job training program for current welfare recipients. If fewer than 30% of participants in the program are able to find work within three months, the program will be discontinued.

> **Solution:** 1-sample proportion test. $H_0 : p = 0.3$, $H_a : p < 0.3$
>
> A confidence interval would not be appropriate in this case because we need to test if there are fewer than 30% rather than simply estimate the success rate.

(e) (2 points) A high school principal is interested in how well she can predict the number of days that her students miss school as a function of their GPA.

> **Solution:** Linear correlation coefficient. Specifically, look at r^2 if the correlation is statistically significant since we want to determine *how well* we can predict.
>
> If the principal had wanted to make a prediction, then regression would be used.

(f) (2 points) A manufacturer needs to measure how consistently a new machine can cut nails to the desired length.

Solution: Standard deviation. Take a sample of nails and determine the sample standard deviation. A small standard deviation would indicate the machine is very consistent.

2. (1 point) 1-Way ANOVA can be thought of as a generalization of what two sample test?

Solution: The two sample t 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?

Solution: If all three measures of center are the same, the distribution is symmetrical.

4. The following questions regard hypothesis testing in general.

(a) (2 points) 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?

Solution: We assume the null hypothesis H_0 is true.

(b) (2 points) Do we use the population distribution or the sampling distribution when calculating the p-value?

Solution: Sampling distribution.

(c) (2 points) If you reject H_0 but H_a is false. What type of error has occurred? (Type I or Type II)

Solution: Type I

(d) (2 points) In the one sample proportion test, what is the distribution of the test statistic?

Solution: The standard normal distribution (z distribution)

(e) (2 points) In the one sample proportion test, what requirements must be met so that the test statistic's distribution is valid?

Solution: The data must have a binomial distribution and the normal approximation to the binomial (np and $nq \geq 5$) must be satisfied.

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

Solution: To prevent bias in the results.

(g) (2 points) A two sample mean hypothesis test was conducted with $H_0 : \Delta\mu = 0$ and $H_a : \Delta\mu > 0$. The first and second samples had respective sample sizes of 18 and 20. The test statistic calculated from the sample data is $t = 1.89$. Find the p-value.

Solution: Find the area to the right of the test statistic since H_a uses $>$ using the t distribution:

```
> df = min(18 - 1, 20 - 1)
> df
[1] 17
> p = 1 - pt(1.89, df)
> signif(p, 3)
[1] 0.038
```

5. Nine students were randomly selected who had taken the SAT twice. A researcher would like to test the claim that students who take the SAT test a second time score higher than their first test.

Student	A	B	C	D	E	F	G	H	I
First SAT Score	480	510	530	540	550	560	600	620	660
Second SAT Score	460	500	530	520	580	580	560	640	690

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

Solution: Use a matched pair hypothesis test.

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

Solution: (1) Simple random samples, (2) the sampling distribution of $\bar{d}$ is normally distributed (CLT must apply to d_i)

(c) (2 points) What are the hypothesis H_0 and H_a?

Solution: $H_0 : \mu_d = 0$, $H_0 : \mu_d < 0$, where $d_i = (\text{first SAT score})_i - (\text{second SAT score})_i$. If we expect the second set to be better then the scores should be higher on average so the difference would be negative (Thus $H_0 : \mu_d < 0$).

(d) (1 point) What α will you use?

Solution: $\alpha = 0.05$

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

Solution:

```
> first = c(480, 510, 530, 540, 550, 560, 600, 620, 660)
> second = c(460, 500, 530, 520, 580, 580, 560, 640, 690)
> res = t.test(first, second, paired = TRUE, alternative = ”less”)
> res
        Paired t−test

data:  first and second
t = −0.1322, df = 8, p−value = 0.4491
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     −Inf 14.52226
sample estimates:
mean of the differences
             −1.111111
```

The p-value is 0.449.

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

Solution: Since p-val $\nleq \alpha$, fail to reject H_0.

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

Solution: The sample data does not support the claim that students score higher when the take the SAT a second time.

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

Weight	3175	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.**)

Solution:

```
> weight = c(3175, 3450, 3225, 3985, 2440, 2500, 2290)
> mpg = c(27, 29, 27, 24, 37, 34, 37)
> res = cor.test(weight, mpg)
> res
        Pearson's product−moment correlation

data:  weight and mpg
t = −6.393, df = 5, p−value = 0.001387
```

```
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9919064 -0.6600615
sample estimates:
      cor
-0.943927
```

Yes, there is a statistically significant linear correlation since the p-value ≤ 0.05.

(b) (2 points) What percent of a vehicle's fuel consumption can be explained by its weight?

Solution: $r^2 = 89.1\%$

(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 coefficients).

Solution:

```
> res = lm(mpg ~ weight)
> res
Call:
lm(formula = mpg ~ weight)

Coefficients:
(Intercept)        weight
  54.695038     -0.007969
```

$$\hat{y} = 54.7 + (-0.00797) \cdot x \tag{1}$$

$$(\text{MPG}) = 54.7 + (-0.00797) \cdot (\text{weight}) \tag{2}$$

(d) (2 points) What range of vehicle weights is the model valid for making predictions of fuel efficiency?

Solution:

```
> range(weight)
[1] 2290 3985
```

(e) (2 points) What is the best predicted fuel consumption for a new vehicle that weights 2800 lbs?

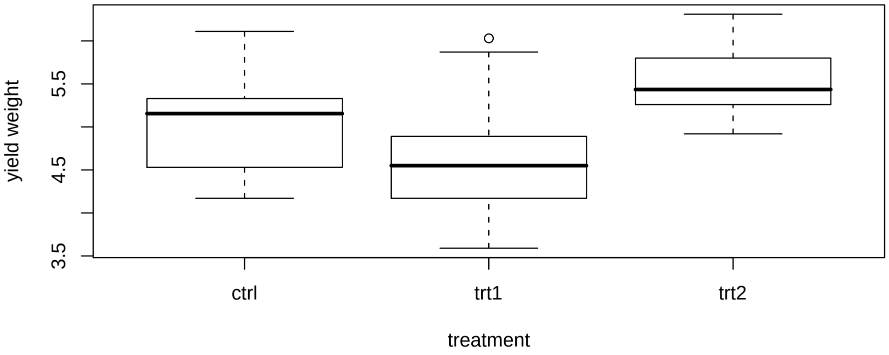
Solution: Evaluate the above equation for the given weight. The best predicted fuel consumption is 32.4 MPG.

(f) (2 points) If the liner relationship had not been statistically significant, what would the best predicted fuel consumption for a new vehicle that weights 2800 lbs be?

Solution: If the liner correlation is not statistically significant, the best prediction is $\bar{y}$

```
> y.bar = mean(mpg)
> signif(y.bar, 3)
[1] 30.7
```

7. Results from a randomized experiment to compare yields (as measured by dried weight of plants in grams) obtained under a control and two different treatment conditions are shown with a box plot of the data. The researcher who has developed the two new treatments hopes that at least one increases crop yield as compared to the control group.



(a) (1 point) From the above box plot, the researcher notices that there do appear to be differences in the crop yield depending on the treatment. The researcher concludes from the box plots that the sample data supports the claim that the treatment type a plant receives affects the crop yield.

The researcher's thought process has a serious error in forming the conclusion. What has the researcher forgotten to consider?

Solution: Sampling error. The observed differences may not be due to the treatments, they may actually be due to the random fluctuations caused by random sampling.

(b) (1 point) What type of hypothesis test should the researcher use to test her claim?

Solution: 1-WAY ANOVA (since there are 3 categories: control, treatment 1, treatment 2)

(c) (2 points) State the null and alternative hypothesis for this study **in words**.

Solution: H_0 the mean crop yield is the same for each treatment.
H_a the mean crop yield is different for at least one treatment.

(d) (1 point) If the researcher runs your suggested hypothesis test and the p-value is 0.18, what should her final conclusion be?

Solution: The sample data does not support the claim that the treatment effects the crop yield.

8. (2 points) 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?

Solution: Find n using:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 \tag{3}$$

```
> E = 1
> sigma = 2.3
> alpha = 0.05
> z.critical = qnorm(1 - alpha/2)
> z.critical
[1] 1.959964
> n = (z.critical * sigma/E)^2
> n
[1] 20.32132
> ceiling(n)
[1] 21
```

Use a sample size of 21. (Must round up.)

9. (2 points) A random sample of 5 people was conducted to determine the mean length of index fingers. Below is the study data in inches.

3.2, 3.9, 3, 3.7, 3.7

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

Solution:

Need to find E in

$$CI = \bar{x} \pm E \tag{4}$$

$$= \bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{5}$$

```
> x
[1] 3.2 3.9 3.0 3.7 3.7
> alpha = 0.1
> n = length(x)
> x.bar = mean(x)
> x.bar
[1] 3.5
> s = sd(x)
> s
[1] 0.3807887
> std.err = s/sqrt(n)
> std.err
[1] 0.1702939
> t.crit = qt(1 - alpha/2, df = n - 1)
> t.crit
[1] 2.131847
> E = t.crit * std.err
> E
[1] 0.3630404
```

The confidence interval is: 3.5 ± 0.363 or $(3.14, 3.86)$

10. A bag of M&M's contains 18 red, 12 blue, 8 green, and 7 brown candies.

    (a) (2 points) What is the probability of randomly selecting a red or brown M&M?

Solution: $P(\text{red or brown}) = P(\text{red}) + (\text{brown})$

```
> total = 18 + 12 + 8 + 7
> P = (18/total) + (7/total)
> signif(P, 3)
[1] 0.556
```

    (b) (2 points) If 10 M&M's are randomly selected with replacement, what is the probability of getting exactly 4 green M&M's?

Solution: Use the binomial distribution.

```
> x = 4
> n = 10
> p = 8/total
> P = dbinom(x, n, p)
> signif(P, 3)
[1] 0.0648
```

11. Engineers must consider the breadths of male heads when designing motorcycle helmets. Men have head breadths that are normally distributed with a mean of 6.0 in and a standard deviation of 1.0 in (based on anthropometric survey data from Gordon, Churchill, et al.).

Solution: Write down the given information:

```
> mu = 6
> sigma = 1
```

(a) (2 points) If 1 man is randomly selected, find the probability that his head breadth is greater than 6.1 in.

Solution: Find $P(x > 6.1)$ using the normal distribution and the given parameters:

```
> p = 1 - pnorm(6.1, mean = mu, sd = sigma)
> signif(p, 3)
[1] 0.46
```

(b) (2 points) If 100 men are randomly selected, find the probability that their mean head breadth is greater than 6.1 in.

Solution: Find $P(\bar{x} > 6.1)$ using the normal distribution for the sampling distribution of $\bar{x}$ (since the CLT applies). The standard deviation will be the standard error:

```
> n = 100
> std.err = sigma/sqrt(n)
> p = 1 - pnorm(6.1, mean = mu, sd = std.err)
> signif(p, 3)
[1] 0.159
```

12. (2 points) Given $x = \{4c, 2c, -2c\}$, where c is a constant, completely simplify the following expression:

$$\sqrt{\frac{\sum(x_i^2 - 2c)}{6c}}$$

Solution: $\sqrt{4c-1}$

****

End of exam. Reference sheets follow.

Points earned: ______ / 2 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://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())**

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

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:

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

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
```