SOLUTIONS
MAT 167: Statistics

Test II

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 13 questions for a total of 84 points on 12 pages.

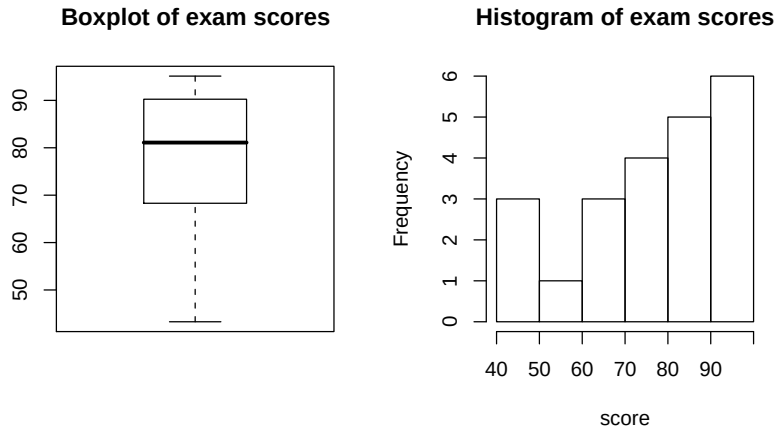
This Exam is being given under the guidelines of our institution's
Code of Academic Ethics. You are expected to respect those guidelines.

Points Earned: _____ **out of 84 total points**

Exam Score: _____

Solution: Spring 2008 results. I made the exam out of 82 points instead of 84.

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  43.29   68.60   81.10   75.75   89.79   95.12
> par(mfrow = c(1, 2))
> boxplot(score, main = "Boxplot of exam scores")
> hist(score, main = "Histogram of exam scores")
```



1. An experiment consists of randomly sampling 10 students at Pima Community College, recording their heights and computing their mean height. If we repeat the experiment over and over (assume simple random samples, no human errors, no bias) we observe that the sample mean height varies each time.

   (a) (2 points) What is the name of the error that causes the mean to vary each time?

   Solution: sampling error

   (b) (2 points) Explain how it is possible for the sample mean height to vary each time? What is going on?

   Solution: Each time we conduct a random sample of size n from a fixed population, the elements that are selected in the sample will vary. Since the samples vary then we would expect the sample mean to vary as well. This is the natural variation we would expect to see from sample to sample called sampling error.

   (c) (2 points) In words, state what the population distribution represents in this experiment. (Be specific.)

   Solution: A population distribution describes the distribution of population values. In this experiment, the population distribution describes the heights of all students at Pima Community College.

(d) (2 points) In words, state what the sampling distribution represents in this experiment. (Be specific.)

Solution: A sampling distribution describes the distribution of statistics for a random sample of size n. In this experiment, it's the distribution of all possible sample means for $n = 10$. More specifically, the sampling distribution describes all the possible sample means that could be obtained from all possible samples of size 10 at Pima Community College.

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

Solution: If the requirements for a binomial distribution are met, it can be approximated as a normal distribution when : $np, nq \geq 5$.

3. (2 points) Which distribution (normal, binomial, both, or neither) would be appropriate for describing:

The distribution for the number of people who wear glasses in a random sample of 20 people where the probability an individual person wears glasses is 0.56.

Solution: Both the normal and the binomial would be appropriate. The binomial distribution is appropriate because there is a fixed number of trials and p does not change (independent), and only two outcomes. The binomial can be approximated as a normal since $np, nq \geq 5$.

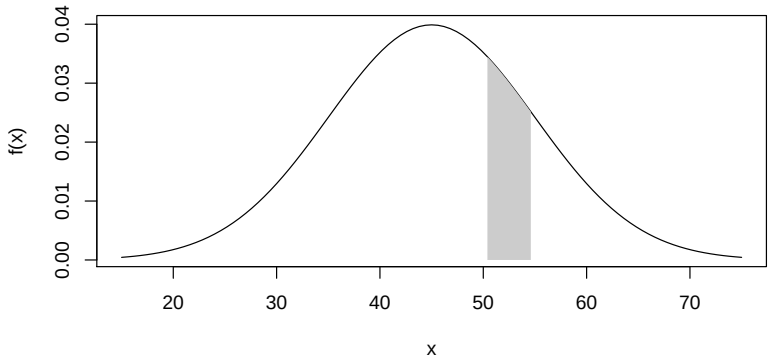
4. In regards to $\bar{x}$ and the Central Limit Theorem:
   (a) (2 points) What are the two conditions under which the CLT applies?

   Solution: Either (1) x (the population) has a normal distribution or (2) $n > 30$.

   (b) (2 points) If the conditions are met, what type of distribution will $\bar{x}$ have?

   Solution: The sampling distribution of $\bar{x}$ can be described as a normal distribution.

5. Let x be a random variable with a normal distribution where $\mu = 45$ and $\sigma = 10$.
   (a) (2 points) Make a meaningful sketch that represents $P(50 < x < 55)$.



   Solution:

   (b) (2 points) Find $P(50 < x < 55)$.

   Solution: Use normal CDF: $P(50 < x < 55) = F(55) - F(50)$

```
> pnorm(55, mean = 45, sd = 10) - pnorm(50, mean = 45, sd = 10)
[1] 0.1498823
```

6. 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) If you are using a hypothesis test to make a decision where the effect of a Type I error may negatively effect human lives, should you increase or decrease α?

   Solution: You should **decrease** α to reduce the probability of making a Type I error.

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

   Solution: Type II

   (d) (2 points) What variable do we use to represent a Type II error with?

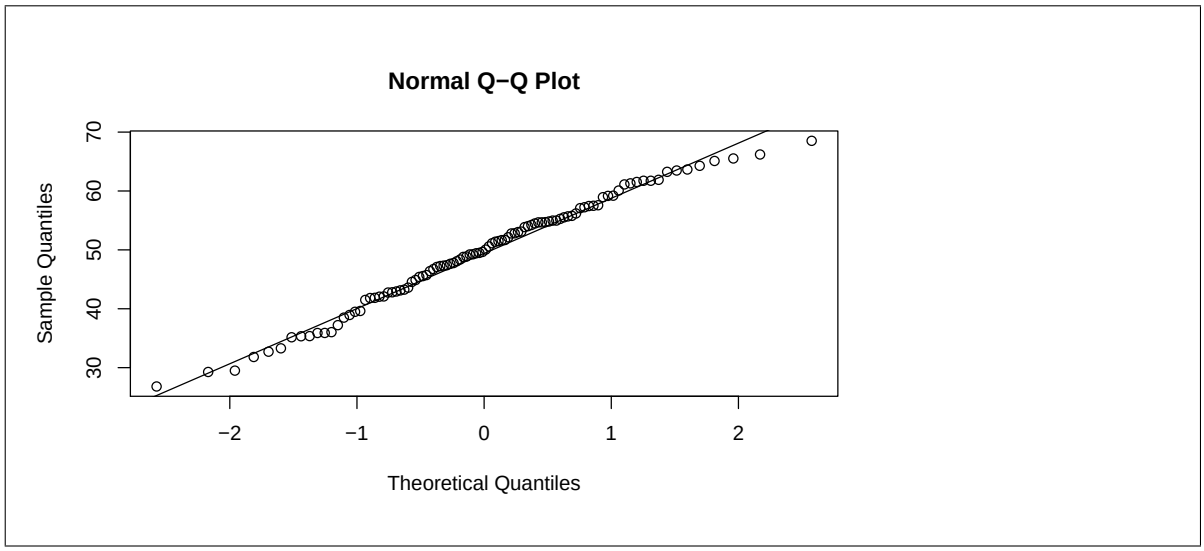
   Solution: β

   (e) (2 points) Two studies were conducted, study A had a power of 0.9 and study B had a power of 0.10. Which study would be more likely to support a true alternative hypothesis?

   Solution: Study A because it had a power of 0.90. There is a 90% chance of supporting a true alternative hypothesis in this study.

7. (2 points) Describe what a normal Q-Q plot should look like for a set of data that has a normal distribution.

   Solution: It should be a scatter plot of data points that follow a straight line with no obvious pattern. Below wold be an example.


Normal Q−Q Plot
Sample Quantiles
30
40
50
60
70
−2
−1
0
1
2
Theoretical Quantiles

8. 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 less than 6.2 in.

Solution: Find $P(x < 6.2)$ using the normal distribution and the given parameters:

```
> p = pnorm(6.2, mean = mu, sd = sigma)
> signif(p, 3)
[1] 0.579
```

(b) (2 points) If 100 men are randomly selected, find the probability that their mean head breadth is less than 6.2 in.

Solution: Find $P(\bar{x} < 6.2)$ 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 = pnorm(6.2, mean = mu, sd = std.err)
> signif(p, 3)
[1] 0.977
```

(c) (2 points) ACME motorcycle company is making a new adjustable helmet. In reality, it is not economical to make a helmet that fits everyone. You must design a helmet that will fit all but largest 5% of male head breadths. What is the largest size male head breadth that your new helmet will fit?

Solution: Solve for a in $P(x < a) = 0.95$, therefore use the inverse normal cumulative distribution using the given parameters for the population:

```
> a = qnorm(0.95, mean = mu, sd = sigma)
> signif(a, 3)
[1] 7.64
```

Thus, the largest size male head breath the helmet will fit is 7.64 in.

9. (2 points) ACME helmet company needs to know the mean head breadth of women for a new helmet design. You conduct a study of 8 randomly selected women (via a simple random sample). Below is the data from the study.

   5.1, 5.7, 5.5, 6.4, 5.7, 5.9, 5.1, 6.2

   Construct a 95% confidence interval for the mean head breadth size for women (**Assume σ is unknown.**)

Solution:

Need to find E in

$$CI = \bar{x} \pm E \tag{1}$$

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

```
> x
 [1] 5.1 5.7 5.5 6.4 5.7 5.9 5.1 6.2
> alpha = 0.05
> n = length(x)
> x.bar = mean(x)
> x.bar
 [1] 5.7
> s = sd(x)
> s
 [1] 0.4690416
> std.err = s/sqrt(n)
> std.err
 [1] 0.1658312
> t.crit = qt(1 - alpha/2, df = n - 1)
> t.crit
 [1] 2.364624
> E = t.crit * std.err
> E
 [1] 0.3921286
```

The confidence interval is: 5.7 ± 0.392 or $(5.31, 6.09)$

10. You believe that the true mean head breadth for women is less than that of men (6.0 in). Using the same study data from the previous question of 8 randomly women (shown below again), conduct a hypothesis test to test your claim. Use a significance level of 0.1 and **assume σ is unknown** and women's head breadths are normally distributed.

    5.1, 5.7, 5.5, 6.4, 5.7, 5.9, 5.1, 6.2

(a) (2 points) What type of hypothesis test will you use?

Solution: Use a one sample mean test with σ unknown.

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

Solution: (1) Simple random samples, (2) CLT applies.

(c) (2 points) Are the requirements satisfied? **State how they are satisfied.**

Solution: Yes. Simple random samples used, and population was normally distributed.

(d) (2 points) What are the hypothesis?

Solution: $H_0 : \mu = 6.0$, $H_a : \mu < 6.0$

(e) (2 points) What α will you use?

Solution: $\alpha = 0.1$

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

Solution:

```
> x
[1] 5.1 5.7 5.5 6.4 5.7 5.9 5.1 6.2
> res = t.test(x, mu = 6, alternative = "less")
> res

        One Sample t-test

data:  x
t = -1.8091, df = 7, p-value = 0.05668
alternative hypothesis: true mean is less than 6
95 percent confidence interval:
    -Inf 6.01418
sample estimates:
mean of x
      5.7
```

The p-value is 0.0567.

(g) (2 points) What is your formal decision?

Solution: Since p-val $\leq \alpha$, reject H_0.

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

Solution: The sample data support the claim that the mean head breadth size of women is less than 6.0 in. [Exact meaning matters.]

(i) (2 points) What is the *actual* probability of a Type I error for this study data?

Solution: The p-value = 0.0567.

(j) (2 points) If the researcher had an α of 0.005 and failed reject H_0, have we proven that the mean head breadth size of women is 6.0in?

Solution: No.

11. Over the past 55 years, data from the National Oceanic and Atmospheric Administration (NOAA) indicates the the probability of precipitation[1] on a given day in Tucson is 0.146. (Use 365 days in a year.)

(a) (2 points) Find the mean and standard deviation for the number of days per year with precipitation in Tucson.

Solution: Binomial distribution: $\mu = n \cdot p$, $\sigma = \sqrt{(n \cdot p \cdot q)}$

```
> n = 365
> p
[1] 0.1460535
> q = 1 - p
> mu = n * p
> mu
[1] 53.30954
> sigma = sqrt(n * p * q)
```

The mean number of days with precipitation is 53.3. The standard deviation is 6.75.

(b) (2 points) What is the probability of 40 or fewer days of precipitation in Tucson annually?

Solution: We need to find on the binomial dist:
$P(x \leq 40) = P(x = 40) + P(x = 39) + \cdots + P(x = 1) + P(x = 0)$.
Since $np, nq \geq 5$ use the normal approximation of the binomial — with the continuity correction — to find P(x<40.5).

```
> mu
[1] 53.30954
> sigma
[1] 6.74711
> P = pnorm(40.5, mean = mu, sd = sigma)
```

[1] Data from `http://www.wrcc.dri.edu/cgi-bin/clilcd.pl?az23160`. Precipitation defined as 0.01 inches or more.

The probability is about 0.0288.

Although it's easiest to use the normal approximation, you can find the probability directly with the binomial by summing up all the probabilities (just don't forget 0):

```
> sum(dbinom(0:40, n, p))
[1] 0.02554455
```

(c) (2 points) Would it be unusual to have 40 or fewer days of precipitation annually in Tucson? (Why.)

Solution: Since $P(x < 40.5) \leq 0.05$ it would be unusual.

12. (2 points) You have been hired by the Tucson Weekly to estimate the proportion people who are in support of John McCain in Tucson. What random sample size should you use to estimate the proportion with a margin of error of 1%?

Solution: Need to find n in:

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 \tag{3}$$

since no information is given, assume $p = q = 0.5$. and $\alpha = 0.05$

```
> p = 0.5
> q = 0.5
> E = 0.01
> alpha = 0.05
> z.crit = qnorm(1 - alpha/2)
> z.crit
[1] 1.959964
> n = p * q * (z.crit/E)^2
> n
[1] 9603.647
```

Thus, you should randomly sample 9604 people to attain the desired margin of error at the 95% confidence level.

13. The Tucson Republican party claims that more than half the people who live in Tucson support John McCain.[2] They conduct a random sample of 100 people at the annual Pima Country Fair. You are hired by them to analyze the data and test their hypothesis.

(a) (2 points) What type of hypothesis test will you use?

Solution: Use a one sample proportion test.

[2]The data in this question is fictitious and is not an endorsement of any candidate or party.

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

Solution: (1) Simple random samples, (2) Bimomial distribution, (3) Normal approx to binomial applies.

(c) (2 points) What are the hypothesis?

Solution: $H_0 : p = 0.5$, $H_a : p > 0.5$

Using the data from the study, you run the analysis in R. Below is the output.

```
        1-sample proportions test with continuity correction

data:  78 out of 100, null probability 0.5
X-squared = 30.25, df = 1, p-value = 1.899e-08
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.6995942 1.0000000
sample estimates:
   p
0.78
```

(d) (2 points) What is the point estimate from the study for the proportion of people who support John McCain?

Solution: $\hat{p} = 0.78$.

(e) (2 points) What is the p-value.

Solution:
The p-value is 1.9e-08.

(f) (2 points) What is your formal decision?

Solution: Since p-val $\leq \alpha$, reject H_0.

(g) (2 points) State your final conclusion in words based upon the analysis above.

Solution: The sample data support the claim that the proportion of people who support John McCain is greater than 50%.

(h) (2 points) What is wrong with this study.

Solution: The random sample of people at the Pima County Fair is not necessarily representative of people who live in Tucson. This is most likely a biased sample. Not all people who live in Tucson go to the fair. The conclusion should be clear and state that **the proportion of people who go to the Pima County Fair and support Jon McCain** is greater than 50%.

****

End of exam. Reference sheets follow.

****

Basic Statistics: Quick Reference & R Commands

by Anthony Tanbakuchi. Version 1.8
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, ...)`

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

2.2 Relative standing

$$z = \frac{x-\mu}{\sigma} = \frac{x-\bar{x}}{s} \tag{15}$$

Percentiles:

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

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

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

7.2 CI Critical Values (two sided)

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

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**`)`**

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

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

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.

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

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

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 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
```