SOLUTIONS
MAT 167: Statistics

Test II

Instructor: Anthony Tanbakuchi

Fall 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 18 questions for a total of 70 points on 11 pages.

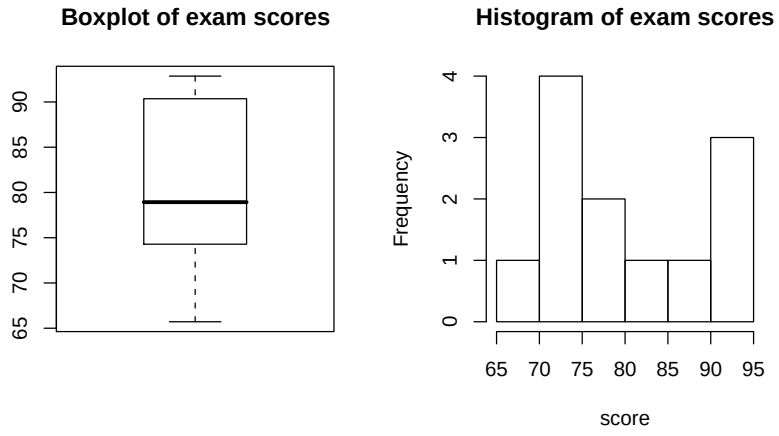
This Exam is being given under the guidelines of our institution's
Code of Academic Ethics. You are expected to respect those guidelines.

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

Exam Score: _____

Solution: Fall 2008 results.

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65.71   74.29   78.93   80.24   90.18   92.86
> par(mfrow = c(1, 2))
> boxplot(score, main = "Boxplot of exam scores")
> hist(score, main = "Histogram of exam scores")
```



1. (1 point) Why is it important to use random sampling?

Solution: To prevent bias. Most statistical methods assume random sampling therefore the results will only be reliable if we ensure the assumptions are valid.

2. For the following statements, determine if the calculation requires the use of a **population distribution** or a **sampling distribution**.

   (a) (1 point) Computing a confidence interval for a mean.

   Solution: Sampling distribution. We need to utilize the distribution of the sample means.

   (b) (1 point) Determining a p-value for a one sample proportion hypothesis test.

   Solution: Sampling distribution. We need to utilize the distribution of the sample statistic.

   (c) (1 point) Calculating the probability than an individual weights more than 100 lbs.

   Solution: Population distribution. We need to utilize the distribution of individual's weights (the population).

(d) (1 point) Computing an interval that contain 95% of individual's weights.

Solution: Population distribution. We need to utilize the distribution of individual's weights (the population).

3. (1 point) What type of error does a sampling distribution characterize?

Solution: Sampling error.

4. (1 point) 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 & $nq \geq 5$.

5. (1 point) If the normal approximation to the binomial is valid, write what the following binomial probability statement is approximately equal to in terms of the normal distribution.

$$P_{\text{binom}}(10 < x \leq 15) \approx$$

Solution: Use the continuity correction.

$$P_{\text{binom}}(10 < x \leq 15) \approx P_{\text{norm}}(10.5 < x < 15.5)$$

6. 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.

7. (1 point) Which distribution (normal, binomial, both, or neither) would be appropriate for describing:

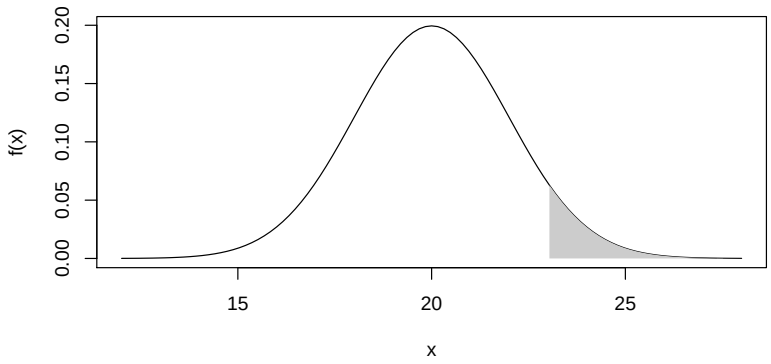
The distribution for the sample mean incomes when taking a random sample of 15 individuals and the distribution of incomes has a strong positive skew.

Solution: Neither. The CLT does not apply in this case.

8. (1 point) For the one sample proportion hypothesis test, what is the distribution of the test statistic? (Give the specific name.)

Solution: Standard normal distribution.

9. Let x be a random variable with a normal distribution where $\mu = 20$ and $\sigma = 2$.
   (a) (2 points) Make a meaningful sketch that represents $P(x > 23)$.



Solution:

   (b) (2 points) Find $P(x > 23)$.

Solution: Use normal CDF: $P(x > 23) = 1 - F(23)$

```
> p = 1 - pnorm(23, mean = 20, sd = 2)
> signif(p, 3)
[1] 0.0668
```

   (c) (1 point) Would it be unusual to observe $x > 23$?

Solution: Unusual if the probability ≤ 0.05

```
> p <= 0.05
[1] FALSE
```

10. The following questions regard hypothesis testing in general.
   (a) (1 point) 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) (1 point) 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) (1 point) You reject H_0 but H_0 is true. What type of error has occurred? (Type I or Type II)

Solution: Type I

(d) (1 point) What variable represents the actual Type I error.

Solution: p-value

(e) (1 point) 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.

(f) (1 point) A researcher takes a sample, conducts a hypothesis test, and rejects the null hypothesis since the p-value was sufficiently small. The researcher concludes that "the sample data proves that the mean height of men is greater than 5.5 feet." What is wrong with this conclusion?

Solution: We never prove a hypothesis from a sample, we can only show that the data supports a hypothesis.

11. (2 points) What is a normal Q-Q plot used for?

Solution: It is used to asses if a set of data follows a normal distribution. The data should fall close to a straight line if it has a normal distribution.

12. (2 points) Eight different second-year medical students at Bellevue Hospital measured the blood pressure of the same person and the results are shown below.

138, 130, 135, 140, 120, 125, 120, 130

Construct a 95% confidence interval estimate for the mean blood pressure assuming the data has a normal distribution.

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] 138 130 135 140 120 125 120 130
> alpha = 0.05
> n = length(x)
> n
[1] 8
> x.bar = mean(x)
> x.bar
[1] 129.75
> s = sd(x)
> s
[1] 7.685794
> std.err = s/sqrt(n)
> std.err
[1] 2.717339
> t.crit = qt(1 - alpha/2, df = n - 1)
> t.crit
[1] 2.364624
> E = t.crit * std.err
> E
[1] 6.425485
```

The confidence interval is: 130 ± 6.43 or $(123, 136)$

13. (2 points) A hypothesis test was conducted for $H_0 : p = 0.5$ and $H_a : p \neq 0.5$. The test statistic is $z = 1.8$. Find the p-value.

Solution: Since this is a two tailed test and z is positive, find the upper tail area on the standard normal distribution and double it.

```
> p.val = 2 * (1 - pnorm(1.8))
> signif(p.val, 3)
[1] 0.0719
```

14. For women aged 18-24, systolic blood pressure (in mm Hg) are normally distributed with a mean of 114.8 and a standard deviation of 13.1. (Based on data from a National Health Survey). Hypertension is commonly defined as a systolic blood pressure above 140.

Solution: Write down the given information:

```
> mu = 114.8
> sigma = 13.1
```

(a) (2 points) If a woman between the ages of 18 and 24 is randomly selected, find the probability that her systolic blood pressure is greater than 140.

Solution: Find $P(x > 140)$ using the normal distribution and the given parameters:

```
> p = 1 - pnorm(140, mean = mu, sd = sigma)
> signif(p, 3)
[1] 0.0272
```

(b) (2 points) A doctor tells a female patent who is in the age range of 18 to 24 that her systolic blood pressure is in the 25^{th} percentile. What is her blood pressure?

Solution: Solve for a in $P(x < a) = 0.25$, therefore use the inverse normal cumulative distribution using the given parameters for the population:

```
> a = qnorm(0.25, mean = mu, sd = sigma)
> signif(a, 3)
[1] 106
```

(c) (2 points) If 4 women are randomly selected and their mean blood pressure is computed, what type of distribution would the sample means have and **why**?

Solution: Normal distribution since the population has a normal distribution (CLT).

(d) (2 points) If 4 women in that age bracket are randomly selected, find the probability that their mean systolic blood pressure is greater than 140.

Solution: Find $P(\bar{x} > 140)$ using the normal distribution for the sampling distribution of $\bar{x}$ (since the CLT applies). The standard deviation will be the standard error:

```
> n = 4
> std.err = sigma/sqrt(n)
> p = 1 - pnorm(140, mean = mu, sd = std.err)
> signif(p, 3)
[1] 5.97e-05
```

15. (2 points) The music industry must adjust to the growing practice of consumers downloading songs instead of buying CDs. It therefore becomes important to estimate the proportion of songs that are currently downloaded. How many randomly selected song purchases must be

surveyed to determine the percentage that were obtained by downloading? Assume that we want to be 95% confident that the sample percentage is within one percentage point of the true population percentage of songs that are downloaded.

Solution:

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

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

Thus, we need a sample size of at least 9604.

16. You believe that the true mean head breadth for men is greater than 6.0 in. A study of 8 randomly selected men (shown below) was conducted to test this claim. Use a significance level of 0.025 and assume that men's head breadths are normally distributed.

6.2, 6.7, 5, 7.6, 7, 7, 6.8, 6.3

(a) (1 point) 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) (1 point) 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) (1 point) What α will you use?

Solution: $\alpha = 0.025$

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

Solution:

```
> x
[1] 6.2 6.7 5.0 7.6 7.0 7.0 6.8 6.3
> res = t.test(x, mu = 6, alternative = "greater")
> res

        One Sample t-test

data:  x
t = 2.1059, df = 7, p-value = 0.03662
alternative hypothesis: true mean is greater than 6
95 percent confidence interval:
 6.057695      Inf
sample estimates:
mean of x
    6.575
```

The p-value is 0.0366.

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

Solution: Since p-val $\not\leq \alpha$, fail to reject H_0.

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

Solution: The sample data does not support the claim that the mean head breadth of men is greater than 6.0 in.

(i) (1 point) If we reject H_0, what is the *actual* probability of a Type I error for this study data?

Solution: The p-value = 0.0366.

17. A pharmaceutical company has developed a new drug that they believe can be used to increase the probability a student will pass a statistics exam. To test the drug's effectiveness, they randomly select 50 students, give them the drug, and have them take the AP statistics exam of which 30 students pass. For those students who did not get the drug, only 58% passed. The drug company hopes to support the claim that the new drug increases the standard passing rate of 58%.

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

Solution: Use a one sample proportion test.

(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: $H_0 : p = 0.58$, $H_a : p > 0.58$

(d) (1 point) What α will you use?

Solution: $\alpha = 0.05$

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

Solution:

```
> res = prop.test(30, 50, p = 0.58, alternative = "greater")
> res

        1-sample proportions test with continuity correction

data:  30 out of 50, null probability 0.58
X-squared = 0.0205, df = 1, p-value = 0.443
alternative hypothesis: true p is greater than 0.58
95 percent confidence interval:
 0.4738505 1.0000000
sample estimates:
  p
0.6
```

The p-value is 0.443.

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

Solution: Since p-val $\not\leq \alpha$, fail to reject H_0.

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

Solution: The sample data does not support the claim that the drug can increase the passing rate beyond 58%.

18. Over the past 55 years, data from the National Oceanic and Atmospheric Administration (NOAA) indicates the the probability that the minimum daily temperature is at most 32 degrees fahrenheit[1] on a given day in Tucson is 0.048. (Use 365 days in a year.)

[1] Data from `http://www.wrcc.dri.edu/cgi-bin/clilcd.pl?az23160`.

(a) (1 point) Find the mean number of days per year with a minimum daily temperature of at most 32 degrees fahrenheit.

Solution: Binomial distribution: $\mu = n \cdot p$,

```
> n = 365
> p
[1] 0.04804393
> mu = n * p
```

The mean number of days with precipitation is 17.5.

(b) (1 point) Find the standard deviation for the number of days per year with a minimum daily temperature of at most 32 degrees fahrenheit.

Solution: Binomial distribution: $\sigma = \sqrt{n \cdot p \cdot q}$

```
> n = 365
> p
[1] 0.04804393
> q = 1 - p
> sigma = sqrt(n * p * q)
```

The standard deviation is 4.09.

(c) (2 points) What is the probability of 20 or more days with a minimum daily temperature of at most 32 degrees fahrenheit.

Solution: We need to find on the binomial dist:
$P(x \geq 20) = P(x = 20) + P(x = 21) + \cdots + P(x = 365)$.
Since $np, nq \geq 5$ use the normal approximation of the binomial — with the continuity correction — to find $P(x > 19.5)$.

```
> mu
[1] 17.53603
> sigma
[1] 4.085772
> P = 1 - pnorm(19.5, mean = mu, sd = sigma)
```

The probability is about 0.315.

You can also find the probability directly with the binomial by summing up all the probabilities.

```
> sum(dbinom(20:265, n, p))
[1] 0.3054070
```

As you can see, the two results are in good agreement since the normal approximation to the binomial is valid.

(d) (1 point) Would it be unusual to have 20 or more days with a minimum daily temperature of at most 32 degrees fahrenheit? (Why)

Solution: Since $P(x \geq 20) \not\leq 0.05$ it would not be unusual.

****

End of exam. Reference sheets follow.

****

Statistics Quick Reference Card & R Commands

by Anthony Tanbakuchi. Version 1.8.1
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())`

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

2 Descriptive Statistics

2.1 Numerical

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

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

2.2 Relative standing

$$z = \frac{x-\mu}{\sigma} = \frac{x-\bar{x}}{s} \quad (15)$$

Percentiles:

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

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!)

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

4 Discrete Random Variables

$$P(x_i): \text{ probability distribution} \quad (27)$$
$$E = \mu = \sum x_i \cdot P(x_i) \quad (28)$$
$$\sigma = \sqrt{\sum (x_i-\mu)^2 \cdot P(x_i)} \quad (29)$$

4.1 Binomial distribution

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

4.2 Poisson distribution

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

5 Continuous random variables

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

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

5.1 Uniform distribution

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

5.2 Normal distribution

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

5.3 t-distribution

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

5.4 χ^2-distribution

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

5.5 F-distribution

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

6 Sampling distributions

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (57)$$
$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \quad (58)$$

7 Estimation

7.1 Confidence Intervals

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

7.2 CI Critical Values (two sided)

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

7.3 Required sample size

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

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
```