

SOLUTIONS
MAT 167: STATISTICS

TEST II

INSTRUCTOR: ANTHONY TANBAKUCHI

SPRING 2009

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. 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 19 questions for a total of 72 points on 13 pages.

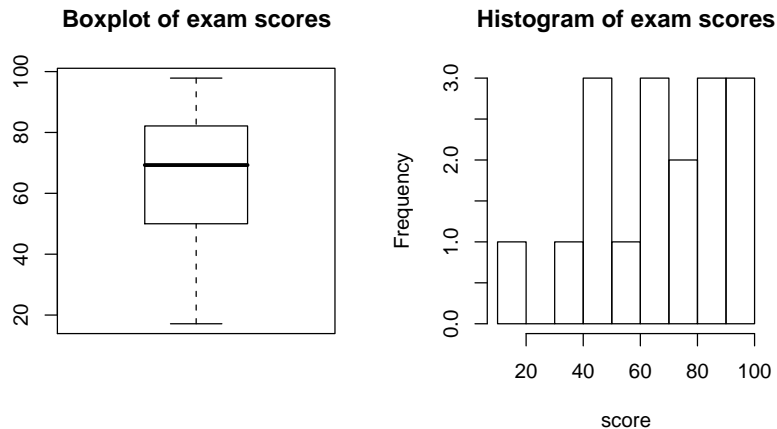
This Exam is being given under the guidelines of our institution's **Code of Academic Ethics**. You are expected to respect those guidelines.

Points Earned: _____ out of 72 total points

Exam Score: _____

Solution: Spring 2009 results.

```
> summary(score)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 17.14  50.00   69.29   67.44   82.14   97.86
> 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 | 9. histogram |
| 2. median | 10. pareto chart |
| 3. mode | 11. box plot |
| 4. standard deviation | 12. normal-quantile plot |
| 5. z-score | 13. confidence interval for a mean |
| 6. percentile | 14. confidence interval for a proportion |
| 7. coefficient of variation | 15. one sample mean test |
| 8. scatter plot | 16. one sample proportion test |

For each situation below, which method is most applicable?

- (a) (1 point) A researcher wants to estimate the mean weight loss when on the *Hot Beach Diet*. The researcher takes a random sample of 100 people's weight loss who used the diet.

Solution: Construct a confidence interval for a mean.

- (b) (1 point) A student needs to conduct a one-sample mean test on a sample of 25 measurements. Before conducting the test, the student needs determine if the sample data appears to have a normal distribution.

Solution: Plot the data with a histogram and see if it looks like a normal distribution. If there are no outliers and it does not appear skewed, then closely analyze it with a normal quantile plot. The data should fall close to a line on the Q-Q norm plot if it has a normal distribution.

- (c) (1 point) The US Bureau of Labor Statistics wants to test the claim that the unemployment rate is above 9 percent using a random sample of 500 people.

Solution: Conduct a one sample proportion hypothesis test.

2. (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.

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

- (a) (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).

- (b) (1 point) Calculating the probability the mean weight of 100 randomly selected individual is more than 100 lbs.

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

- (c) (1 point) Computing a confidence interval for a proportion.

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

- (d) (1 point) Determining a p-value for a one sample mean hypothesis test.

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

4. (1 point) What type of error does a sampling distribution characterize?

Solution: Sampling error.

5. (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 & $nq \geq 5$.

6. (1 point) If the normal approximation to the binomial is valid, express the following binomial probability statement in terms of the normal distribution.

$$P_{\text{binom}}(x = 12) \approx$$

Solution: Use the continuity correction.

$$P_{\text{binom}}(x = 12) \approx P_{\text{norm}}(11.5 < x < 12.5)$$

7. (1 point) Use the binomial distribution to find $P_{\text{binom}}(x = 12)$ assuming $n = 25$, $p = 0.4$.

Solution:

```
> p = dbinom(12, 25, 0.4)
> signif(p, 3)
[1] 0.114
```

8. (2 points) Use the normal approximation of the binomial to find $P_{\text{binom}}(x = 12)$ assuming $n = 25$, $p = 0.4$.

Solution: First find the mean and standard deviation of the binomial distribution, then use the normal approximation from the previous question. Recall that:

$$\mu = n \cdot p \tag{1}$$

$$\sigma = \sqrt{n \cdot p \cdot q} \tag{2}$$

```
> x = 12
> n = 25
> p = 0.4
> q = 1 - p
> mu = n * p
> mu
```

```

[1] 10
> sigma = sqrt(n * p * q)
> sigma
[1] 2.449490
> p.approx = pnorm(12.5, mean = mu, sd = sigma) - pnorm(11.5, mean = mu,
+      sd = sigma)
> signif(p.approx, 3)
[1] 0.116

```

9. 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) (1 point) 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.

(c) (1 point) What is the mean $\mu_{\bar{x}}$ of the sampling distribution equal to?

Solution: $\mu_{\bar{x}} = \mu$

(d) (1 point) What is the standard deviation $\sigma_{\bar{x}}$ of the sampling distribution equal to?

Solution: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

10. (1 point) Which distribution (normal, binomial, both, or neither) would be appropriate for describing:

The distribution of sample mean incomes when taking a random sample of 150 individual's incomes.

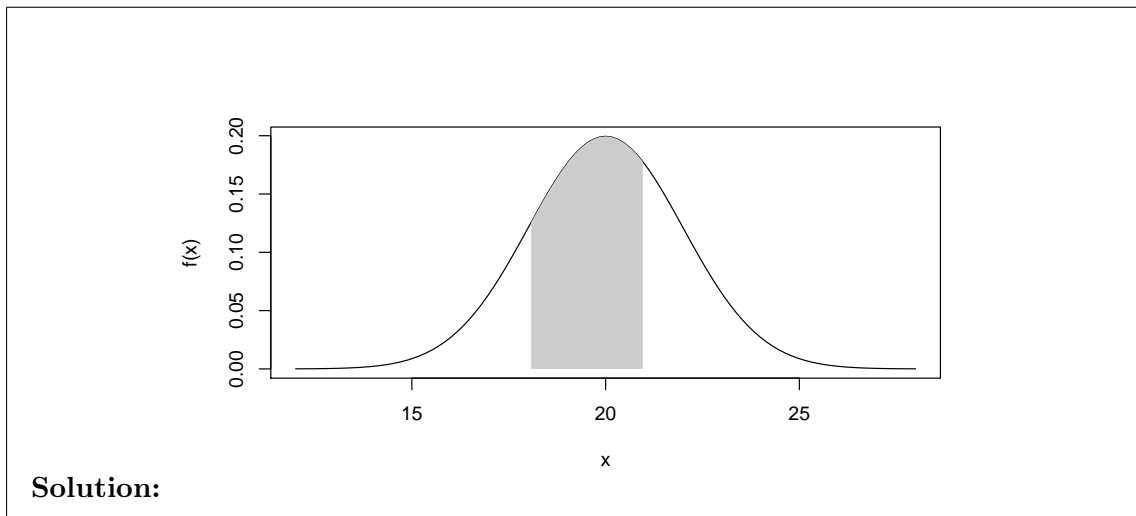
Solution: The normal distribution. The CLT applies since $n > 30$.

11. (1 point) For the one-sample mean hypothesis test, what is the distribution of the test statistic if σ is unknown? (Give the specific name.)

Solution: The t -distribution.

12. 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 $P(18 < x < 21)$.



- (b) (2 points) Find $P(18 < x < 21)$.

Solution: Use normal CDF: $P(18 < x < 21) = F(21) - F(18)$

```
> p = pnorm(21, mean = 20, sd = 2) - pnorm(18, mean = 20, sd = 2)
> signif(p, 3)
[1] 0.533
```

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

Solution: Unusual if $P(x > 20) \leq 0.05$

```
> p = pnorm(20, mean = 20, sd = 2)
> signif(p)
[1] 0.5
> p <= 0.05
[1] FALSE
```

- (d) (1 point) Find $P(x = 20)$

Solution: $P(x = 20) = 0$. Probability is area for a continuous distribution. This region has zero width.

Note that the answer is **not** `pnorm(20,mean=20,sd=2)` since this is equivalent to $P(x < 20)$.

13. 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 fail to reject H_0 but H_0 is false. What type of error has occurred? (Type I or Type II)

Solution: Type II

- (d) (1 point) What variable represents the actual Type I error for a study.

Solution: p-value

- (e) (1 point) Two studies were conducted, study A had a power of 0.3 and study B had a power of 0.8. Which study would be more likely to support a true alternative hypothesis?

Solution: Study B because it had a higher power. There is a 80% chance of supporting a true alternative hypothesis in this study.

- (f) (1 point) A researcher takes a sample, conducts a hypothesis test, and fails to reject the null hypothesis since the p-value was not small enough. The researcher concludes that “the sample data supports that the mean height of men is equal to 5.5 feet.” What is wrong with this conclusion?

Solution: We never support the null hypothesis (unless we calculate β to determine if the probability of a Type II error is sufficiently small).

- (g) (1 point) Write a correct conclusion for the research in the previous question.

Solution: “The sample data **does not contradict** the claim that the mean height of men is equal to 5.5 feet.”

14. (2 points) Ten randomly selected customers were asked their age at DangerWay Grocery Store. The ten ages are shown below.

28, 44, 39, 63, 44, 49, 26, 59, 20, 32

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

Solution:

Use:

$$CI = \bar{x} \pm E \quad (3)$$

$$= \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad (4)$$

```

> x
[1] 28 44 39 63 44 49 26 59 20 32
> alpha = 0.05
> n = length(x)
> n
[1] 10
> x.bar = mean(x)
> x.bar
[1] 40.4
> s = sd(x)
> s
[1] 14.16725
> std.err = s/sqrt(n)
> std.err
[1] 4.480079
> t.crit = qt(1 - alpha/2, df = n - 1)
> t.crit
[1] 2.262157
> E = t.crit * std.err
> E
[1] 10.13464

```

The confidence interval is: 40.4 ± 10.1 or $(30.3, 50.5)$

15. (1 point) A hypothesis test was conducted for $H_0 : \mu = 25$ and $H_a : \mu > 25$. The test statistic is $t = 1.8$ and the sample size was 15. Find the p-value.

Solution: Find the upper tail area on the t -distribution.

```

> p.val = 1 - pt(1.8, df = 15 - 1)
> signif(p.val, 3)
[1] 0.0467

```

16. For women aged 18-24, systolic blood pressures (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 less than 140.

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

```
> p = pnorm(140, mean = mu, sd = sigma)
> signif(p, 3)
[1] 0.973
```

- (b) (2 points) A doctor tells a female patient who is in the age range of 18 to 24 that her systolic blood pressure is in the 95th percentile. What is her blood pressure?

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] 136
```

- (c) (2 points) If 6 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 6 women in the age range of 18-24 years old are randomly selected, find the probability that their mean systolic blood pressure is less 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 = 6
> std.err = sigma/sqrt(n)
> p = pnorm(140, mean = mu, sd = std.err)
> signif(p, 3)
[1] 1
```

17. (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 99% confident that the sample percentage is within 2% 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 = 0.02
> alpha = 0.01
> z.crit = qnorm(1 - alpha/2)
> z.crit
[1] 2.575829
> p = 0.5
> q = 0.5
> n = p * q * (z.crit/E)^2
> n
[1] 4146.81
```

Thus, we need a sample size of at least 4147.

18. A petroleum company has developed a new type of synthetic oil that can be used to decrease the probability that a car will overheat. To test the oil's effectiveness, they randomly test 100 cars using the new oil. Out of the 100 cars only 5 overheat. In general, 8% of cars driven under the same conditions with standard oil overheat. The oil company hopes to support the claim that the oil decreases the rate of overheating to less than 8%.

- (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) Binomial distribution, (3) Normal approx. to binomial applies.

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

Solution: $H_0 : p = 0.08$, $H_a : p > 0.08$

- (d) (1 point) What α will you use?

Solution: $\alpha = 0.05$

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

Solution:

```
> res = prop.test(5, 100, p = 0.08, alternative = "less")
> res
      1-sample proportions test with continuity correction

data:  5 out of 100, null probability 0.08
X-squared = 0.8492, df = 1, p-value = 0.1784
alternative hypothesis: true p is less than 0.08
95 percent confidence interval:
 0.0000000 0.1055372
sample estimates:
      p
0.05
```

The p -value is 0.178.

Or, if you prefer to do it manually, find the p -value with the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

```
> p.hat = 5/100
> p0 = 0.08
> q0 = 1 - p0
> n = 100
> z = (p.hat - p0)/sqrt(p0 * q0/n)
> signif(z, 3)
[1] -1.11
> p.val = pnorm(z)
> signif(p.val, 3)
[1] 0.134
```

The p -value is 0.134.

Note that this p -value differs slightly from the one obtained with `prop.test` since R uses the continuity correction.

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

Solution: Since $p\text{-val} \not\leq \alpha$, 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 new oil can decrease the overheating rate below 8%.

- (h) (1 point) If the oil can't really decrease the overheating rate below 8%, what could cause us to observe only 5% overheating in the study.

Solution: Sampling error.

19. You believe that the true mean weight of statistics books is 7.0 lbs. A study of 8 randomly selected statistics books weights (shown below) was conducted to test this claim. Use a significance level of 0.025 and assume that the weights are normally distributed.

6.4, 7.1, 6.1, 6.9, 5.6, 7, 7.7, 8

- (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: $H_0 : \mu = 6.0$, $H_a : \mu \neq 6.0$

- (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] 6.4 7.1 6.1 6.9 5.6 7.0 7.7 8.0
> res = t.test(x, mu = 7, alternative = "two.sided")
> res
      One Sample t-test

data:  x
t = -0.5315, df = 7, p-value = 0.6115
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
```

```

6.182678 7.517322
sample estimates:
mean of x
      6.85

```

The p -value is 0.612.

Or, if you prefer to do it manually, find the p -value with the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1$$

```

> x
[1] 6.4 7.1 6.1 6.9 5.6 7.0 7.7 8.0
> x.bar = mean(x)
> x.bar
[1] 6.85
> s = sd(x)
> s
[1] 0.7982123
> mu0 = 7
> n = length(x)
> n
[1] 8
> t = (x.bar - mu0)/(s/sqrt(n))
> signif(t, 3)
[1] -0.532
> p.val = 2 * pt(t, df = n - 1)
> signif(p.val, 3)
[1] 0.612

```

The p -value is 0.612.

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

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

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

Solution: The sample data does not contradict the claim that statistics text book weight is equal to 7.0 lbs.

- (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 = 0.612.

End of exam. Reference sheets follow.

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 v store data: $x=c(x1, x2, \dots)$
 Help: general `RSiteSearch("Search Phrase")`
 Get: function `?functionName`
 Get column of data from table:
`tableName$columnName`
 List all variables: `ls()`
 Delete all variables: `rm(list=ls())`

$$\sqrt{x} = \text{sqrt}(x) \quad (1)$$

$$x^n = x^n \quad (2)$$

$$n = \text{length}(x) \quad (3)$$

$$T = \text{table}(x) \quad (4)$$

2 Descriptive Statistics

2.1 NUMERICAL

Let $x=c(x1, x2, x3, \dots)$

$$\text{total} = \sum_{i=1}^n x_i = \text{sum}(x) \quad (5)$$

$$\text{min} = \text{min}(x) \quad (6)$$

$$\text{max} = \text{max}(x) \quad (7)$$

$$\text{six number summary} = \text{summary}(x) \quad (8)$$

$$\mu = \frac{\sum x_i}{N} = \text{mean}(x) \quad (9)$$

$$\bar{x} = \frac{\sum x_i}{N} = \text{mean}(x) \quad (10)$$

$$\bar{x} = P_{50} = \text{median}(x) \quad (11)$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \quad (12)$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \text{sd}(x) \quad (13)$$

$$CV = \frac{\sigma}{\mu} = \frac{s}{\bar{x}} \quad (14)$$

2.2 RELATIVE STANDING

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

Percentiles:

$$P_k = x_i, \text{ (sorted } x) \quad (16)$$

$$k = \frac{i-0.5}{n} \cdot 100\%$$

To find x_i given P_k , i is:

- $L = (k/100)n$
- 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 point plot
- `type="l"` for line plot
- `type="b"` for both 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!)

$$P(A) = \frac{x}{n} \quad (17)$$

$$P(\bar{A}) = 1 - P(A) \quad (18)$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad (19)$$

$$P(A \text{ or } B) = P(A) + P(B) \text{ if } A, B \text{ mut. excl.} \quad (20)$$

$$P(A \text{ and } B) = P(A) \cdot P(B|A) \quad (21)$$

$$P(A \text{ and } B) = P(A) \cdot P(B) \text{ if } A, B \text{ independent} \quad (22)$$

$$n! = n(n-1) \cdots 1 = \text{factorial}(n) \quad (23)$$

$${}_n P_k = \frac{n!}{(n-k)!} \text{ Perm. no elem. alike} \quad (24)$$

$${}_n C_k = \frac{n!}{n_1! n_2! \cdots n_k!} \text{ Perm. } n_1 \text{ alike, } \dots \quad (25)$$

$${}_n C_k = \frac{n!}{(n-k)! k!} = \text{choose}(n, k) \quad (26)$$

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

$$\mu = n \cdot p \quad (30)$$

$$\sigma = \sqrt{n \cdot p \cdot q} \quad (31)$$

$$P(x) = {}_n C_x p^x q^{n-x} = \text{dbinom}(x, n, p) \quad (32)$$

4.2 POISSON DISTRIBUTION

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

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

$$p = P(a < u') = F(u') \quad (44)$$

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

5.2 NORMAL DISTRIBUTION

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

$$p = P(z < z') = F(z') = \text{pnorm}(z') \quad (47)$$

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

$$p = P(x < x') = F(x') \quad (49)$$

$$= \text{pnorm}(x', \text{mean}=\mu, \text{sd}=\sigma) \quad (49)$$

$$x' = F^{-1}(p) \quad (50)$$

$$= \text{qnorm}(p, \text{mean}=\mu, \text{sd}=\sigma) \quad (50)$$

5.3 t-DISTRIBUTION

$$p = P(t < t') = F(t') = \text{pt}(t', \text{df}) \quad (51)$$

$$t' = F^{-1}(p) = \text{qt}(p, \text{df}) \quad (52)$$

5.4 χ^2 -DISTRIBUTION

$$p = P(\chi^2 < \chi'^2) = F(\chi'^2) \quad (53)$$

$$= \text{pchisq}(\chi'^2, \text{df}) \quad (53)$$

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

5.5 F-DISTRIBUTION

$$p = P(F < F') = F(F') \quad (55)$$

$$= \text{pf}(F', \text{df1}, \text{df2}) \quad (55)$$

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

6 Sampling distributions

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (57)$$

$$\mu_{\bar{p}} = p \quad \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} \quad (58)$$

7 Estimation

7.1 CONFIDENCE INTERVALS

$$\text{proportion: } \hat{p} \pm E, \quad E = z_{\alpha/2} \cdot \sigma_{\bar{p}} \quad (59)$$

$$\text{mean } (\sigma \text{ known): } \bar{x} \pm E, \quad E = z_{\alpha/2} \cdot \sigma_{\bar{x}} \quad (60)$$

$$\text{mean } (\sigma \text{ unknown, use } s): \bar{x} \pm E, \quad E = t_{\alpha/2} \cdot \sigma_{\bar{x}} \quad (61)$$

$$df = n - 1$$

$$\text{variance: } \frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L} \quad (62)$$

$$df = n - 1$$

$$2 \text{ proportions: } \Delta \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \quad (63)$$

$$2 \text{ means (indep): } \Delta \bar{x} \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (64)$$

$$df \approx \text{min}(n_1 - 1, n_2 - 1)$$

$$\text{matched pairs: } \bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}, \quad d_i = x_i - y_i, \quad (65)$$

$$df = n - 1$$

7.2 CI CRITICAL VALUES (TWO SIDED)

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

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

$$\chi^2_L = F_{\chi^2}^{-1}(\alpha/2) = \text{qchisq}(\alpha/2, \text{df}) \quad (68)$$

$$\chi^2_R = F_{\chi^2}^{-1}(1 - \alpha/2) = \text{qchisq}(1 - \alpha/2, \text{df}) \quad (69)$$

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=p0, alternative="two.sided")

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}} \quad (72)$$

8.2 1-SAMPLE MEAN (σ KNOWN)

$H_0: \mu = \mu_0$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad (73)$$

8.3 1-SAMPLE MEAN (σ UNKNOWN)

$H_0: \mu = \mu_0$

t.test(x, mu=mu0, alternative="two.sided")

Where **x** is a vector of sample data.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1 \quad (74)$$

Required Sample size:

power.t.test(delta=h, sd=s, sig.level=alpha, 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(x1, x2)** and **n=c(n1, n2)**

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

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}, \quad \hat{q} = 1 - \hat{p} \quad (76)$$

Required Sample size:

power.prop.test(p1=p1, p2=p2, power=1 - beta, sig.level=alpha, 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.

$$t = \frac{\Delta \bar{x} - \Delta \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad df \approx \min(n_1 - 1, n_2 - 1), \quad \Delta \bar{x} = \bar{x}_1 - \bar{x}_2 \quad (77)$$

Required Sample size:

power.t.test(delta=h, sd=s, sig.level=alpha, 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.

$$t = \frac{\bar{d} - \mu_{d0}}{s_d/\sqrt{n}}, \quad d_i = x_i - y_i, \quad df = n - 1 \quad (78)$$

Required Sample size:

power.t.test(delta=h, sd=s, sig.level=alpha, power=1 - beta, type="paired", alternative="two.sided")

8.7 TEST OF HOMOGENEITY, TEST OF INDEPENDENCE

$H_0: p_1 = p_2 = \dots = p_m$ (homogeneity)

$H_0: X$ and Y are independent (independence)

chisq.test(D)

Enter table: **D=matrix(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.

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

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

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.

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

9.2 MODELS IN R

MODEL TYPE	EQUATION	R MODEL
linear 1 indep var	$y = b_0 + b_1 x_1$	$y \sim x_1$
... 0 intercept	$y = 0 + b_1 x_1$	$y \sim 0 + x_1$
linear 2 indep vars	$y = b_0 + b_1 x_1 + b_2 x_2$	$y \sim x_1 + x_2$
... interaction	$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$	$y \sim x_1 + x_2 + x_1 * x_2$
polynomial	$y = b_0 + b_1 x_1 + b_2 x_1^2$	$y \sim x_1 + 1(x_1^2)^2$

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

$$y = b_0 + b_1 x_1 \quad (82)$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (83)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (84)$$

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

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

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 separated file) from Excel.
2. Import your table into 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
```