SOLUTIONS

MAT 167: Statistics

Final Exam

Instructor: Anthony Tanbakuchi

Summer 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. Show your work in the space provided. 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 and turn it in with the test. 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 16 questions for a total of 80 points on 14 pages.

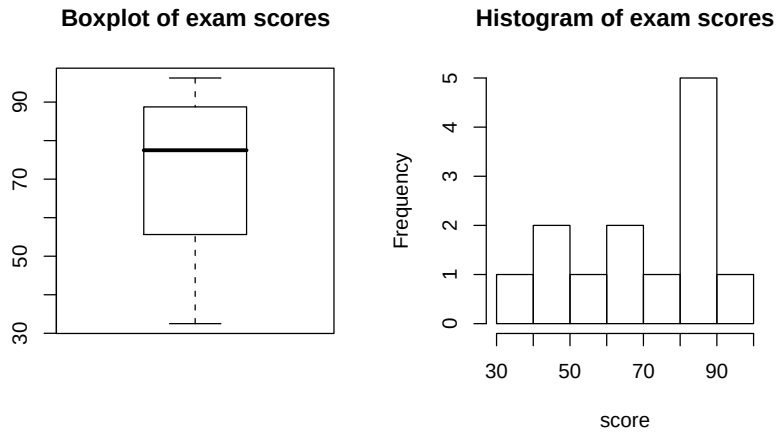
This Exam is being given under the guidelines of our institution's **Code of Academic Ethics**. You are expected to respect those guidelines.

Points Earned: _____ out of 80 total points

Exam Score: _____

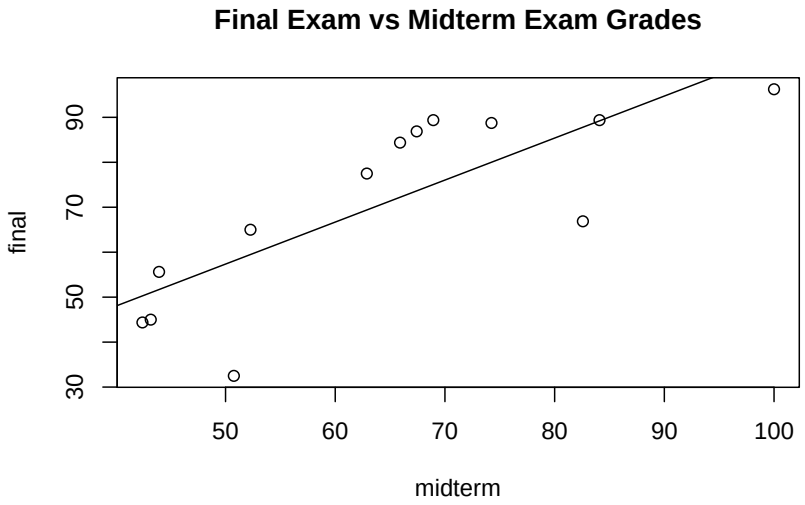
Solution: Exam Results:

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.50   55.62   77.50   70.91   88.75   96.25
> par(mfrow = c(1, 2))
> boxplot(score, main = "Boxplot of exam scores")
> hist(score, main = "Histogram of exam scores")
```



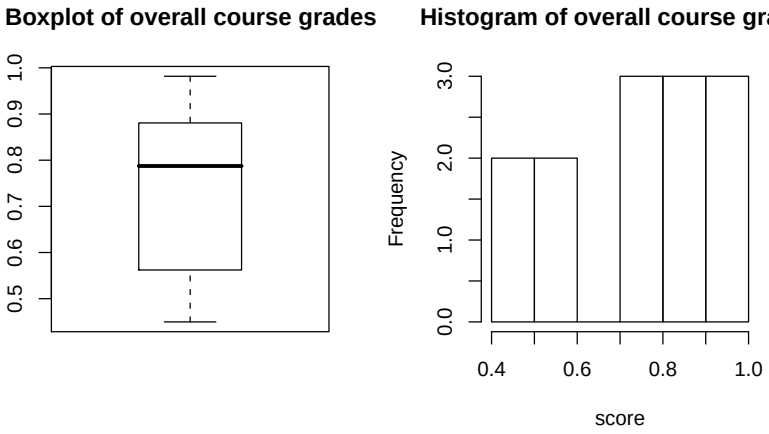
Comparison of midterm to final exam grades:

```
> plot(midterm, final, main = "Final Exam vs Midterm Exam Grades")
> cor.test(midterm, final)
        Pearson's product-moment correlation

data:  midterm and final
t = 4.3384, df = 11, p-value = 0.001178
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4329907 0.9358031
sample estimates:
      cor
0.7944442
> res = lm(final ~ midterm)
> res
Call:
lm(formula = final ~ midterm)

Coefficients:
(Intercept)      midterm
    10.6426       0.9343
> abline(res)
```



Overall Course Grades:

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4497  0.5622  0.7874  0.7444  0.8807  0.9818
> par(mfrow = c(1, 2))
> boxplot(score, main = "Boxplot of overall course grades")
> hist(score, main = "Histogram of overall course grades")
```



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. one sample proportion test
20. two sample proportion test
21. test of homogeneity
22. test of independence
23. linear correlation coefficient & test
24. regression
25. 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) A researcher at the department of labor wants to determine if the proportion of men who work in sociology, psychology, and anthropology is the same from recent study data on the subject.

Solution: Use the test of homogeneity. H_0 : proportion of men in all three fields is the same. H_a : proportion of men is different in at least one of the three fields.

(b) (2 points) A math department committee wants to award \$50 to the student who received the best score on their calculus final exam this past semester. However, the three faculty who taught calculus last semester gave different final exams. What method could help identify the top student amongst the three different exams?

Solution: Use a z score. The student with the highest z score among the three exams would be the awardee.

(c) (2 points) A drug research would like to test the claim that the mean absorption of 1 gram of vitamin E is the same for four methods of delivery: topical, intravenous, oral, and nasal spray.

Solution: Use 1-way ANOVA. H_0 : mean absorption is the same for all four methods. H_a : mean absorption is different for at least one method.

(d) (2 points) A researcher is using a statistical hypothesis test that requires the population which the sample was drawn from to have a normal distribution. How could the researcher check this assumption?

Solution: Use a normal quantile plot. Look to see if the data follows a straight line.

(e) (2 points) A drug researcher wants to determine if a new growth hormone drug can increase the mean weight of mice as compared to a control group of mice.

Solution: Use a two sample hypothesis test of means. $H_0 : \mu_1 = \mu_2$, $H_a : \mu_1 > \mu_2$. (Let group 1 be the control.)

2. (1 point) The test of homogeneity can be thought of as a generalization of what two sample test?

Solution: The two sample proportion 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. (Not necessarily a normal distribution, all we know is that it is symmetrical.)

4. (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 : $np, nq \geq 5$.

5. (2 points) Give **a clear specific example** of when you would use a population distribution.

Solution: You typically use a population distribution to find probabilities of observing an individual value. An example would be: finding the probability and individual's height is less than 6 ft.

6. (2 points) Give **a clear specific example** example of when you would use a sampling distribution.

Solution: You typically use a sampling distribution to find probabilities of observing specific value of a statistic or to make a confidence interval for a parameter from an observed statistic. An example would be: finding the probability that the mean height of students in a class is less than 6 ft.

7. (1 point) What percent of data lies within one standard deviation as stated by the Empirical Rule?

Solution: 68%

8. 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) Do we use the population distribution or the sampling distribution when calculating the p-value?

   Solution: Sampling distribution.

   (c) (1 point) If 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?

   Solution: The p-value. (α is the maximum Type I error, not the actual.)

   (e) (1 point) What variable is used to represent a Type II error?

   Solution: β

   (f) (1 point) What does the power of a hypothesis test represent?

   Solution: The power represents the probability of detecting a true alternative hypothesis of interest.

   (g) (1 point) In the one sample mean test with σ unknown, what is the distribution of the test statistic?

   Solution: The t distribution

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

   Solution: To prevent bias in the results.

9. A consumer advocate group believes that Crystal Springs Sparkling Mineral Water contains more than the advertised 35 mg of sodium per serving. They randomly sample 40 servings and measure the amount of sodium contained in each sample. The collected data has a sample mean of 37.5 mg, and a sample standard deviation of 8.2 mg.

(a) (2 points) What is the null and alternative hypothesis?

Solution: $H_0 : \mu = 35mg, H_a : \mu > 35mg$

(b) (1 point) What hypothesis test should be used?

Solution: Single sample mean test (σ unknown).

(c) (2 points) What are the requirements for the hypothesis test? Are they met?

Solution: Random sampling, σ unknown, and CLT applies. Yes, they are satisfied since $n > 30$.

(d) (1 point) What significance level will you use?

Solution: $\alpha = 0.05$

(e) (2 points) Manually compute the p-value.

Solution:
Find the p-value by getting the tail area from the test statistic:

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

```
> mu0 = 35
> n = 40
> x.bar = 37.5
> s = 8.2
> sigma.x.bar = s/sqrt(n)
> sigma.x.bar
[1] 1.296534
> t = (x.bar - mu0)/sigma.x.bar
> t
[1] 1.928218
> p.value = 1 - pt(t, df = n - 1)
> p.value
[1] 0.03056563
```

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

Solution: Reject H_0 since the p-value $\leq \alpha$

(g) (2 points) What is the final conclusion?

Solution: The sample data supports the claim that Crystal Springs Sparkling Mineral Water contains more than 35 mg of sodium per serving.

10. A craps table at a local casino has been losing more money than normal. It seems that bets involving a one on the face of the dice (such as "snake eyes") are appearing more than usual. The casino manager thinks that the dice have been weighted to cause the side with one to have a higher probability of occurring than a fair dice.

(a) (2 points) The casino manager takes one of the dice from the table and flips it 100 times, the side with a value of one appears 26 times. Construct a 95% confidence interval for the true probability of getting a one with this die.

Solution: Since np and $nq \geq 5$ we can use the methods discussed to construct the confidence interval using the normal approximation to the binomial:

$$\hat{p} \pm E, \quad E = z_{\alpha/2} \cdot \sigma_{\hat{p}}$$

You must use $\hat{p}$ in this calculation. Using $p = 1/6$ is wrong because that assumes the die is fair.

```
> n = 100
> x = 26
> p.hat = x/n
> p.hat
[1] 0.26
> z.crit = qnorm(1 - 0.05/2)
> z.crit
[1] 1.959964
> sigma.p.hat = sqrt(p.hat * (1 - p.hat)/n)
> sigma.p.hat
[1] 0.04386342
> E = z.crit * sigma.p.hat
> E
[1] 0.08597073
> CI = c(p.hat - E, p.hat + E)
> signif(CI, 3)
[1] 0.174 0.346
```

Thus, the 95% confidence interval for the true proportion of ones seen on the die is: $(0.174, 0.346)$

(b) (1 point) If the die is fair, what should the true probability be for getting a one?

Solution: $P(x = 1) = 1/6 = 0.167$

(c) (1 point) Why can't the casino manager conclude that the die is unfair simply based on the fact that he observed one appear more than it theoretically should have appeared?

What other explanation could account for this?

Solution: Sampling error will cause natural fluctuations in the observed sample statistic. Just because we observed a value different from the theoretical value does not prove that the die is fair. We must determine if the observed difference is large enough such that sampling error cannot account for it.

(d) (1 point) Now the casino manager would like to conduct a hypothesis test to determine if the die is unfair. What type of hypothesis test should he use?

Solution: Single sample proportion test.

(e) (2 points) What are the manager's hypothesis for the test?

Solution: $H_0 : p = 1/6$, $H_a : p > 1/6$, where p represents the probability of getting a one.

(f) (2 points) What are the test's requirements? Are they satisfied?

Solution: Simple random samples, binomial distribution satisfied, normal approximation to binomial $np, nq \geq 5$. Yes, requirements satisfied.

(g) (2 points) Conduct the hypothesis test, what is the p-value?

Solution:

```
> res = prop.test(x, n, p = 1/6, alternative = "greater")
> res

        1-sample proportions test with continuity correction

data:  x out of n, null probability 1/6
X-squared = 5.618, df = 1, p-value = 0.008888
alternative hypothesis: true p is greater than 0.1666667
95 percent confidence interval:
 0.1904180 1.0000000
sample estimates:
   p
0.26
```

The p-value is 0.00889.

(h) (1 point) What is the formal decision?

Solution: Since the p-value ≤ 0.05, reject the null hypothesis?

(i) (1 point) What is the conclusion?

Solution: The sample data supports the claim that the dies is unfair and that the value of one appears more than 1/6 of the time.

(j) (1 point) What is the probability of a Type I error for this study?

Solution: The p-value.

11. Eighteen students were randomly selected to take the SAT after having either no breakfast or a complete breakfast A researcher would like to test the claim that students who eat breakfast score higher than students who do not.

Group without breakfast: SAT Score	480	510	530	540	550	560	600	620	660
Group with breakfast: 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 two sample hypothesis test for equality of means. (The test of independents would not be appropriate since the data is not categorical. Analysis of linear correlation would also be inappropriate since the data is not paired.)

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

Solution: (1) Simple random samples, (2) the sampling distribution of for both groups is normally distributed (CLT must apply to both samples).

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

Solution: $H_0 : \mu_1 = \mu_2$, $H_a : \mu_1 < \mu_2$ (Let group 1 be the group without breakfast.)

(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:

```
> g1 = c(480, 510, 530, 540, 550, 560, 600, 620, 660)
> g2 = c(460, 500, 530, 520, 580, 580, 560, 640, 690)
> res = t.test(g1, g2, alternative = "less")
> res

        Welch Two Sample t-test

data:  g1 and g2
t = -0.0368, df = 15.239, p-value = 0.4856
alternative hypothesis: true difference in means is less than 0
```

```
95 percent confidence interval:
     -Inf 51.76744
sample estimates:
mean of x mean of y
 561.1111  562.2222
```

The p-value is 0.486.

(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 students score higher on the SAT when they have breakfast.

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

Weight	3180	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(3180, 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 = -6.431, df = 5, p-value = 0.001351
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9919960 -0.6632053
sample estimates:
       cor
-0.9445332
```

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

(b) (1 point) What percent of a vehicle's fuel consumption can be explained by its weight?

Solution: $r^2 = 89.2\%$

(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
  54.707462    -0.007971
```

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

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

(d) (1 point) What range of vehicle weights is the model valid for making predictions of fuel efficiency?

Solution:

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

(e) (1 point) What is the best predicted fuel consumption for a new vehicle that weights 3200 lbs?

Solution: Evaluate the above equation for the given weight. The best predicted fuel consumption is 29.2 MPG.

(f) (1 point) If the liner relationship had not been statistically significant, what is the best predicted fuel consumption for a new vehicle that weights 3200 lbs?

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
```

13. (2 points) A random sample of 5 chihuahuas was conducted to determine the mean tong length of the breed. Below is the study data in inches.

1.6, 2.1, 1.5, 1.9, 2

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

Solution:

Need to find E in

$$CI = \bar{x} \pm E \tag{3}$$

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

```
> x
[1] 1.6 2.1 1.5 1.9 2.0
> alpha = 0.1
> n = length(x)
> x.bar = mean(x)
> x.bar
[1] 1.82
> s = sd(x)
> s
[1] 0.2588436
> std.err = s/sqrt(n)
> std.err
[1] 0.1157584
> t.crit = qt(1 - alpha/2, df = n - 1)
> t.crit
[1] 2.131847
> E = t.crit * std.err
> E
[1] 0.2467791
```

The confidence interval is: 1.82 ± 0.247 or $(1.57, 2.07)$

14. (2 points) A ski resort is designing a new super tram to carry 40 people. If the mean weight of humans is approximately 165 lbs with a standard deviation of 25 lbs, what should the tram's maximum weight limit be so that it can carry the desired capacity 95% of the time?

Solution: The total maximum weight limit $= \bar{x}_{95} \cdot n$. Where $\bar{x}_{95}$ is the sample mean found using the *sampling distribution* of $\bar{x}$ that has an area of 0.95 to the left.

```
> n = 40
> mu = 165
> sigma = 25
> std.err = sigma/sqrt(n)
> std.err
[1] 3.952847
> x.bar = qnorm(0.95, mean = mu, sd = std.err)
> x.bar
[1] 171.5019
> weight.limit = n * x.bar
> weight.limit
[1] 6860.074
```

15. (2 points) You would like to conduct a study to estimate (at the 90% confidence level) the mean weight of brown bears with a margin of error of 5 lbs. A preliminary study indicates that bear weights are normally distributed with a standard deviation of 22 lbs, what sample size should you use for this study?

Solution: Find n using:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 \tag{5}$$

```
> E = 5
> sigma = 22
> alpha = 0.1
> z.critical = qnorm(1 - alpha/2)
> z.critical
[1] 1.644854
> n = (z.critical * sigma/E)^2
> n
[1] 52.37932
> ceiling(n)
[1] 53
```

Use a sample size of 53. (Must round up.)

16. (2 points) Given $y = \{a, -2a, 4a\}$, where a is a constant, completely simplify the following expression:

$$\left(\sum y_i\right)^2 - 2$$

Solution: $9a^2 - 2$

****

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 / 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**`)`**

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

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

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.

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

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

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

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

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
```