SOLUTIONS
MAT 167: Statistics

Final Exam

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 24 questions for a total of 71 points on 15 pages.

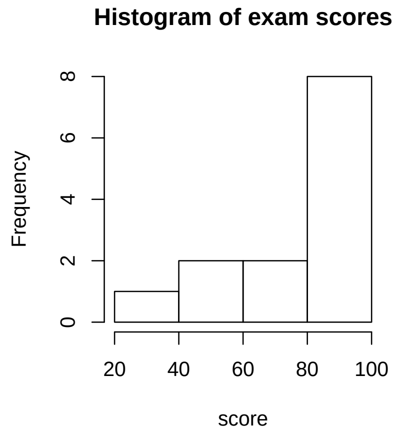
This Exam is being given under the guidelines of our institution's **Code of Academic Ethics**. You are expected to respect those guidelines.

Points Earned: __________ **out of 71 total points**

Exam Score: __________

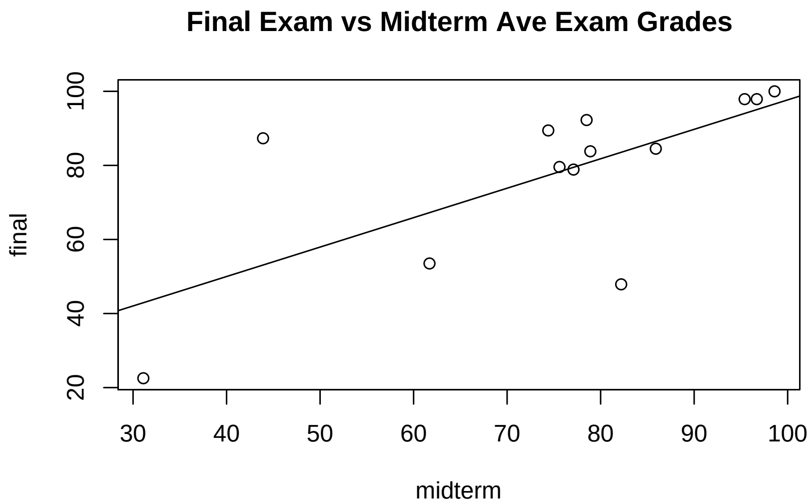
Solution: Exam Results:

```
> summary(score)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  22.54    78.87    84.51    78.11    92.25   100.00
> par(mfrow = c(1, 2))
> boxplot(score, main = "Boxplot of exam scores")
> hist(score, main = "Histogram of exam scores")
```



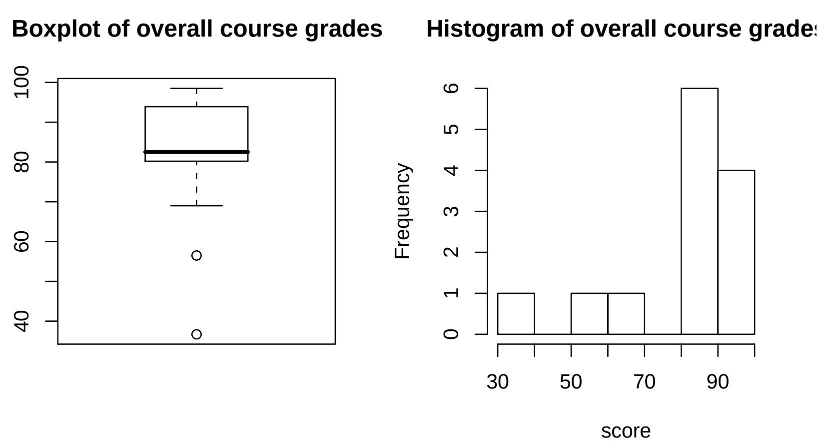
Comparison of midterm average to final exam grades (for those who took the final):

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

data:  midterm and final
t = 3.1152, df = 11, p-value = 0.009834
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2145887 0.8971790
sample estimates:
     cor
0.684627
> res = lm(final ~ midterm)
> res
Call:
lm(formula = final ~ midterm)

Coefficients:
(Intercept)      midterm
    18.1968       0.7948
> abline(res)
```



Overall Course Grades (for those who took the final):

```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  36.70   80.20   82.50   80.78   93.90   98.50
> 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. match pair test
20. one sample proportion test
21. two sample proportion test
22. test of homogeneity
23. test of independence
24. linear correlation coefficient & test
25. regression
26. 1-way ANOVA

For each situation below, which method is most applicable?

(a) (1 point) A researcher would like to estimate the mean weight of javalina.

Solution: Conduct a study and construct a confidence interval for a mean.

(b) (1 point) A researcher wants to determine if bear weights are normally distributed.

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 Q-Q norm plot. The data should fall close to a line on the Q-Q norm plot if it has a normal distribution.

(c) (1 point) An education researcher wants to determine if the probability a student will graduate from middle school is effected by their economic status (poor, lower middle class, middle class, . . .).

Solution: Use the test of homogeneity. The researcher needs to determine if the proportion of students graduating in the different economic classes are the same or if at least one is different.

(d) (1 point) A farmer wants to determine if the mean crop yield is the same for eight different brands of fertilizer.

Solution: Use 1-way ANOVA. H_0 : mean crop yield is equal for all eight brands. H_a : mean crop yield is different for at least one brand.

(e) (1 point) A fertility researcher wants to determine if a new drug can decrease the proportion of infertile mice. Twenty mice are randomly divided into two groups, a treatment group and a control group.

Solution: Use a two sample hypothesis test of proportions. $H_0 : p_1 = p_2$, $H_a : p_1 < p_2$. (Let group 2 be the control.)

2. (1 point) What test is a many sample generalization of the two sample t-test?

Solution: 1-Way ANOVA

3. (1 point) If the mean, median, and mode for a data set are different, what can you conclude about the data's distribution?

Solution: The distribution is not symmetrical, it is skewed.

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 : $np, nq \geq 5$. In words, there must be at least five successes and failures.

5. (1 point) What percent of data lies within one standard deviations of the mean as stated by the Empirical Rule?

Solution: 68%

6. (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.

7. (1 point) A sampling distribution characterizes what type of error?

Solution: Sampling error.

8. 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 proportion.

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

   (b) (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).

9. (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.

$$P_{\text{binom}}(x > 10) \approx$$

Solution: Use the continuity correction.

$$P_{\text{binom}}(x = 8) \approx P_{\text{norm}}(x > 10.5)$$

10. (1 point) For ANOVA, what is the distribution of the test statistic? (Give the specific name.)

Solution: F distribution

11. (2 points) A hypothesis test was conducted for $H_0 : \mu = 5$ and $H_a : \mu > 5$. The test statistic is $t = 2.2$, $n = 15$. Find the p-value.

Solution: Since this is a one tailed test. Find the upper tail area on the t distribution.

$$P(z > 2.2) = 1 - F(2.2, df = n - 1)$$

```
> p.val = 1 - pt(2.2, df = 15 - 1)
> signif(p.val, 3)
[1] 0.0226
```

12. (2 points) A histogram is a useful tool that can quickly communicate many traits about a set of data. List 4 useful pieces of information that an observer can easily assess using a histogram.

Solution: A histogram can be used to get an approximation of:

1. central tendency
2. variation in the data
3. shape of the data
4. assess if outliers exist
5. min
6. max

13. Provide **short succinct** written answers to the following conceptual questions.

(a) (1 point) Give an example of a categorical type of variable.

Solution: Hair color.

(b) (1 point) Which of the following measures of variation is least susceptible to outliers:
standard deviation, inter-quartile range, range

Solution: inter-quartile range.

(c) (1 point) What percent of data is greater than Q_3?

Solution: 25%

(d) (1 point) What does the standard deviation represent conceptually **in words**? (Be concise but don't simply state the equation in words verbatim.)

Solution: The standard deviation represents the average variation of the data from the mean.

(e) (1 point) Why would a SAT percentile be preferred over a raw SAT score for college admissions committees?

Solution: The percentile compares how the applicant did to their peers who took the test (a measure of relative standing). A raw score doesn't give information as to how this score compared to others taking the test, making it hard to determine if a 1100 is easy or hard to get.

14. (2 points) Car tires must not deform or explode when inflated up to their maximum pressure rating. Before distributing the tires, they must be tested. To test the safety of tires, an inspector randomly samples 50 tires (without replacement) from a batch of 5,000 that have been manufactured. The inspector inflates each of the fifty tires until they explode or deform to make sure they meet the minimum safety requirements. If none of the sampled tires fails the test, the tires will be distributed to dealers. If the batch contains 15 defective tires that will explode if selected, what is the probability that the batch will be rejected?

Solution: We are randomly selecting $n = 50$ tires from $n = 5000$. Since we are sampling without replacement these are dependent trials, but $n/N \leq 0.05$ so we can simplify the problem by approximating it as independent.

$$\begin{aligned} P(\text{batch rejected}) &= P(\text{at least one tire defective}) \\ &= 1 - P(\text{None defective out of 50}) \\ &= 1 - P(\text{not defective})^{50} && \text{approx. as indep} \\ &= 1 - (1 - 15/5000)^{50} \\ &= 0.139 \end{aligned}$$

15. (2 points) If a class consists of 20 males and 8 females, what is the probability of drawing 4 females without replacement?

Solution:

$$P(4 \text{ females}) = \frac{8}{28} \cdot \frac{7}{27} \cdot \frac{6}{26} \cdot \frac{5}{25} = 0.00342$$

16. (2 points) You would like to conduct a study to estimate (at the 95% confidence level) the proportion of households that own one or more encyclopedias. What sample size do you need to estimate the proportion with a margin of error of 2%.

Solution: Find n using:

$$\text{proportion: } n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2, \tag{1}$$
$$(\hat{p} = \hat{q} = 0.5 \text{ if unknown})$$

```
> E = 0.02
> alpha = 0.05
> p.hat = 0.5
> q.hat = 0.5
> z.critical = qnorm(1 - alpha/2)
> z.critical
[1] 1.959964
> n = p.hat * q.hat * (z.critical/E)^2
> n
[1] 2400.912
> ceiling(n)
[1] 2401
```

Use a sample size of 2401. (Must round up.)

17. 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 reject H_0 but H_0 is true, what type of error has occurred? (Type I or Type II)

Solution: Type I

(c) (1 point) What variable represents the actual Type I error?

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

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

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

18. 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). (3) Independent samples between groups.

(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?

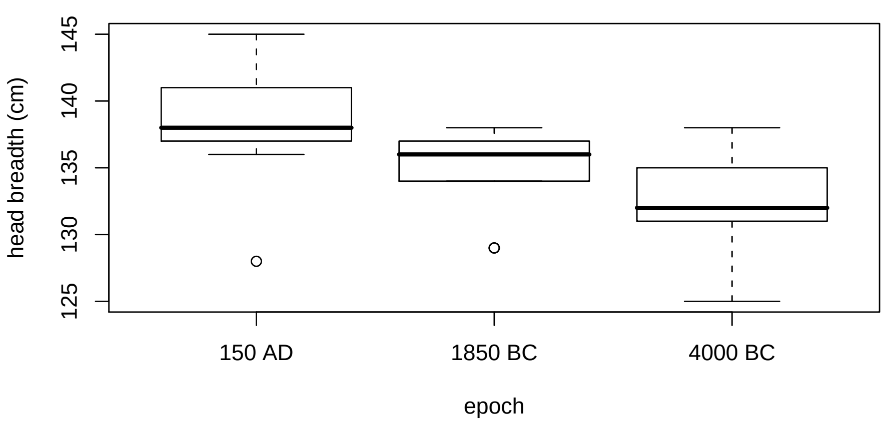
Solution: Since p-val $\not\leq \alpha$, fail to reject H_0.

(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.

19. Samples of head breadths were obtained by measuring skulls of Egyptian males from three different epochs, and the measurements are listed below (based on data from *Ancient Races of the Tebaid*, by Thomas and Randall-Maciver). Changes in head shape over time suggest that interbreeding occurred with immigrant populations. Test the claim that the different epochs do not all have the same mean head breadth.

A box plot of the data is shown below.



(a) (1 point) What type of hypothesis test (of those discussed in class) should you use?

Solution: 1-Way ANOVA

(b) (1 point) What is the alternative hypothesis for this test?

Solution: H_0 : mean head breadth is different in at least one epoch.

(c) (1 point) What alpha will you use?

Solution: $\alpha = 0.05$

(d) (1 point) What is the response variable for this study?

Solution: head breadth

(e) (1 point) What is the factor variable for this study?

Solution: epoch

(f) (1 point) The analysis of the data was run and the output is shown below: What is your

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
epoch	2	138.74	69.37	4.05	0.0305
Residuals	24	411.11	17.13		

final conclusion (not the formal decision)?

Solution: The sample data supports the claim that different epochs do not all have the same mean head breadth.

(g) (1 point) Assuming the researcher rejected the null hypothesis, what is the probability of a Type I error for this study?

Solution: The p-value = 0.0305

20. 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
```

$$\hat{y} = 54.7 + (-0.00797) \cdot x \tag{2}$$

$$(\text{MPG}) = 54.7 + (-0.00797) \cdot (\text{weight}) \tag{3}$$

(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
```

21. (2 points) A researcher is trying to determine the ideal temperature to brew coffee. A random sample of 8 of the top 100 coffee shops in New York City had their brewer temperatures (in Celsius) measured. The data is shown below.

88.6, 91.2, 87.5, 90.5, 85.4, 90.6, 93.5, 94.6

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

Solution:

Need to find E in

$$CI = \bar{x} \pm E \tag{4}$$
$$= \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{5}$$

```
> x
[1] 88.6 91.2 87.5 90.5 85.4 90.6 93.5 94.6
> alpha = 0.1
> n = length(x)
> x.bar = mean(x)
> x.bar
[1] 90.2375
> s = sd(x)
> s
[1] 3.03265
> std.err = s/sqrt(n)
> std.err
[1] 1.072204
> t.crit = qt(1 - alpha/2, df = n - 1)
> t.crit
[1] 1.894579
> E = t.crit * std.err
> E
[1] 2.031374
```

The confidence interval is: 90.2 ± 2.03 or $(88.2, 92.3)$

22. (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
```

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

> **Solution:** Write down the given information:
>
> ```
> > mu = 6
> > sigma = 1
> ```

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

> **Solution:** Find $P(x > 6.1)$ using the normal distribution and the given parameters:
>
> ```
> > p = 1 - pnorm(6.1, mean = mu, sd = sigma)
> > signif(p, 3)
> [1] 0.46
> ```

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

> **Solution:** Find $P(\bar{x} > 6.1)$ using the normal distribution for the sampling distribution of $\bar{x}$ (since the CLT applies). The standard deviation will be the standard error:
>
> ```
> > n = 100
> > std.err = sigma/sqrt(n)
> > p = 1 - pnorm(6.1, mean = mu, sd = std.err)
> > signif(p, 3)
> [1] 0.159
> ```

24. (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())`**

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

2 Descriptive Statistics

2.1 Numerical

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

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

2.2 Relative standing

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

Percentiles:

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

To find x_i given P_k, i is:

1. $L = (k/100\%)n$
2. if L is an integer: $i = L + 0.5$; otherwise i=L and round up.

2.3 Visual

All plots have optional arguments:

- **`main=""`** sets title
- **`xlab="", ylab=""`** sets x/y-axis label
- **`type="p"`** for **p**oint plot
- **`type="l"`** for **l**ine plot
- **`type="b"`** for **b**oth points and lines

Ex: `plot(x, y, type="b", main="My Plot")`

Plot Types:
`hist(x)` histogram
`stem(x)` stem & leaf
`boxplot(x)` box plot
`plot(T)` bar plot, `T=table(x)`
`plot(x,y)` scatter plot, `x`, `y` are ordered vectors
`plot(t,y)` time series plot, `t`, `y` are ordered vectors
`curve(expr, xmin,xmax)` plot `expr` involving x

2.4 Assessing Normality

Q-Q plot: **`qqnorm(x); qqline(x)`**

3 Probability

Number of successes x with n possible outcomes.
(Don't double count!)

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

4 Discrete Random Variables

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

4.1 Binomial distribution

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

4.2 Poisson distribution

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

5 Continuous random variables

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

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

5.1 Uniform distribution

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

5.2 Normal distribution

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

5.3 t-distribution

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

5.4 χ^2-distribution

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

5.5 F-distribution

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

6 Sampling distributions

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

7 Estimation

7.1 Confidence Intervals

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

7.2 CI Critical Values (two sided)

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

7.3 Required sample size

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

8 Hypothesis Tests

Test statistic and R function (when available) are listed for each.

Optional arguments for **hypothesis tests**:

alternative="two.sided" can be:

 "two.sided", **"less"**, **"greater"**

conf.level=0.95 constructs a 95% confidence interval. Standard CI only when `alternative="two.sided"`.

Optional arguments for **power calculations & Type II error**:

alternative="two.sided" can be:

 "two.sided" or **"one.sided"**

sig.level=0.05 sets the significance level α.

8.1 1-SAMPLE PROPORTION

$H_0: p = p_0$

prop.test(x, n, p=p_0, alternative="two.sided")

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \tag{72}$$

8.2 1-SAMPLE MEAN (σ KNOWN)

$H_0: \mu = \mu_0$

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

8.3 1-SAMPLE MEAN (σ UNKNOWN)

$H_0: \mu = \mu_0$

t.test(x, mu=μ_0, alternative="two.sided")

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

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

Required Sample size:

power.t.test(delta=h, sd =σ, sig.level=α, power=$1-\beta$, type ="one.sample", alternative="two.sided")

8.4 2-SAMPLE PROPORTION TEST

$H_0: p_1 = p_2$ or equivalently $H_0: \Delta p = 0$

prop.test(x, n, alternative="two.sided")

where: **x=c(x_1, x_2)** and **n=c(n_1, n_2)**

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

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

Required Sample size:

power.prop.test(p1=p_1, p2=p_2, power=$1-\beta$, sig.level=α, alternative="two.sided")

8.5 2-SAMPLE MEAN TEST

$H_0: \mu_1 = \mu_2$ or equivalently $H_0: \Delta\mu = 0$

t.test(x1, x2, alternative="two.sided")

where: **x1** and **x2** are vectors of sample 1 and sample 2 data.

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

Required Sample size:

power.t.test(delta=h, sd =σ, sig.level=α, power=$1-\beta$, type ="two.sample", alternative="two.sided")

8.6 2-SAMPLE MATCHED PAIRS TEST

$H_0: \mu_d = 0$

t.test(x, y, paired=TRUE, alternative="two.sided")

where: **x** and **y** are ordered vectors of sample 1 and sample 2 data.

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

Required Sample size:

power.t.test(delta=h, sd =σ, sig.level=α, power=$1-\beta$, type ="paired", alternative="two.sided")

8.7 TEST OF HOMOGENEITY, TEST OF INDEPENDENCE

$H_0: p_1 = p_2 = \cdots = p_n$ (homogeneity)

H_0: X and Y are independent (independence)

chisq.test(D)

Enter table: **D=data.frame(c1, c2, ...)**, where `c1, c2, ...` are column data vectors.

Or generate table: **D=table(x1, x2)**, where `x1, x2` are ordered vectors of raw categorical data.

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

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

For 2×2 contingency tables, you can use the Fisher Exact Test:

fisher.test(D, alternative="greater")

(must specify alternative as greater)

9 Linear Regression

9.1 LINEAR CORRELATION

$H_0: \rho = 0$

cor.test(x, y)

where: **x** and **y** are ordered vectors.

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

9.2 MODELS IN R

MODEL TYPE	EQUATION	R MODEL
linear 1 indep var	$y = b_0 + b_1 x_1$	`y~x1`
...0 intercept	$y = 0 + b_1 x_1$	`y~0+x1`
linear 2 indep vars	$y = b_0 + b_1 x_1 + b_2 x_2$	`y~x1+x2`
...inteaction	$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$	`y~x1+x2+x1*x2`
polynomial	$y = b_0 + b_1 x_1 + b_2 x_2^2$	`y~x1+I(x2^2)`

9.3 REGRESSION

Simple linear regression steps:

1. Make sure there is a significant linear correlation.
2. **results=lm(y~x)** Linear regression of `y` on x vectors
3. **results** View the results
4. **plot(x, y); abline(results)** Plot regression line on data
5. **plot(x, results$residuals)** Plot residuals

$$y = b_0 + b_1 x_1 \tag{82}$$

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

$$b_0 = \bar{y} - b_1 \bar{x} \tag{84}$$

9.4 PREDICTION INTERVALS

To predict y when $x = 5$ and show the 95% prediction interval with regression model in `results`:

predict(results, newdata=data.frame(x=5), int="pred")

10 ANOVA

10.1 ONE WAY ANOVA

1. **results=aov(depVarColName~indepVarColName, data=tableName)** Run ANOVA with data in `TableName`, factor data in `indepVarColName` column, and response data in `depVarColName` column.
2. **summary(results)** Summarize results
3. **boxplot(depVarColName~indepVarColName, data=tableName)** Boxplot of levels for factor

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

To find required sample size and power see `power.anova.test(...)`

11 Loading and using external data and tables

11.1 LOADING EXCEL DATA

1. Export your table as a CSV file (comma seperated file) from Excel.
2. Import your table into `MyTable` in R using:
   MyTable=read.csv(file.choose())

11.2 LOADING AN .RDATA FILE

You can either double click on the `.RData` file or use the menu:

- Windows: *File→Load Workspace...*
- Mac: *Workspace→Load Workspace File...*

11.3 USING TABLES OF DATA

1. To see all the available variables type: **ls()**
2. To see what's inside a variable, type its name.
3. If the variable `tableName` is a table, you can also type **names(tableName)** to see the column names or type **head(tableName)** to see the first few rows of data.
4. To access a column of data type **tableName$columnName**

An example demonstrating how to get the women's height data and find the mean:

```
> ls() # See what variables are defined
[1] "women" "x"
> head(women) #Look at the first few entries
  height weight
1     58    115
2     59    117
3     60    120
> names(women) # Just get the column names
[1] "height" "weight"
> women$height # Display the height data
 [1] 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
> mean(women$height)  # Find the mean of the heights
[1] 65
```