

NAME (Please Print): _____

HONOR PLEDGE (Please Sign): _____

statistics 101

Practice Final Key

This is a multiple choice and short answer practice exam. It does not count towards your grade.

You may use the tables in your book.

1. True or False:

- (a) **FALSE** For a t-test of the slope in a regression, the degrees of freedom should be the sample size.
correct answer is $N-p-1$
- (b) **TRUE** William Gosset invented the t-test.
- (c) **TRUE** The P-value is a measure of how improbable the data are if the null is true.
p-value = probability of observing data that is as or more supportive of alternative when null is correct.
- (d) **FALSE** The Belmont Report described ethical principles for animal studies.
described ethical principles for human studies.
- (e) **TRUE** As the confidence level increases, so does the width of the confidence interval.
- (f) **TRUE** The apparent gender discrimination in the Berkeley admissions program is an example of Simpson's paradox.
- (g) **FALSE** Randomization prevents confounding in an observational study.
Helps prevent confounding in designed experiments.

2. **384** Suppose that people's heights are normally distributed with standard deviation 5 inches. How large a sample size do you need to ensure that a two-sided 95% confidence interval for the population average height has width less than 1 inch?

width of interval = $2 * se * z_{(1-C)/2} = 2 * (sd/\sqrt{n}) * z_{2.5} = 2 * (5/\sqrt{n}) * 1.96 = 1$ inch
so $\sqrt{n} = 19.6$ or $n > 384$.

3. What is the significance probability of a test?

significance probability = p-value = probability of observing data that is as or more supportive of alternative when null is correct.

4. For each of 200 M&M's in a pack, Aristides has probability 1/10 of eating it and probability 9/10 of putting it in his pocket.

What is the approximate probability that Aristides eats more than 15 M&M's?
 using CLT approximation $Z > (15 - 20)/\sqrt{200 * .1 * .9} = -5/4.24 = -1.179$ so
 $p - value = 76 + 1/2(100 - 76) = 88\%$.

5. Midterm scores from a class of 100 people have mean 85 and standard deviation 9.

What is the probability that Bucephalus got less than 88?
 $Z = (88 - 85)/9 = 0.33$ so $26 + 1/2(100 - 26) = 63\%$.

What is the probability that Cassandra got between 80 and 95?
 $P(80 < score < 95) = P((80 - 85)/9 < Z < (95 - 85)/9) = P(-0.56 < Z < 1.11) = 1/2(42.5) + 1/2(73.3) = 57.9\%$.

6. You observe the following random sample:

3, -1, 0, -1, 1, 2

What is the sample mean?
 $(3 - 1 + 0 - 1 + 1 + 2)/6 = 4/6 = 0.67$.

What is the median?
 sorted: -1, -1, 0, 1, 2, 3 so median = $(0+1)/2 = 0.5$

What is the standard deviation?

$$\begin{aligned} \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2} &= \sqrt{\frac{1}{6}(1 + 1 + 0 + 1 + 4 + 9) - (0.67)^2} \\ &= \sqrt{16/6 - 0.4489} \\ &= \sqrt{2.218} = 1.49 \end{aligned}$$

What is a 90% two-sided confidence interval on the population mean?

This is a small-sample problem ($n = 6$) where the population standard deviation is estimated by the sample SD. Thus we use the t -distribution with $6 - 1 = 5$ degrees of freedom.

$$L = -0.3332$$

$$L = \bar{X} - se * t_{5,(1-C)/2} = 0.67 - (1.49/\sqrt{5}) * 2.02 = -0.676$$

$$U = 1.6732$$

$$U = \bar{X} + se * t_{5,(1-C)/2} = 0.67 + (1.49/\sqrt{5}) * 2.02 = 2.016$$

Suppose each observation is altered by adding -2 and then multiplying that by -3. What is the new mean?

$$\bar{X}_{new} = -3 * \bar{X} - 2 = -3 * 0.67 - 2 = -4.01$$

What is the new standard deviation?

$$SD_{new} = |-3| * SD = 3 * 1.49 = 4.47$$

7. You draw a random sample of 100 students and ask each of them to secretly roll a fair die. If the result is a 1 or 2, they are supposed to answer yes to the question “Do you use marijuana?” If the result of the roll is a 3, 4, 5, or 6, they should answer honestly. Suppose 92 people answer yes. What is your estimate of the proportion of students who smoke dope?

Let B =say yes; A_1 =honest (roll 3:6); A_2 =dishonest (roll 1:2). We want

$$P(B|A_1) = P(A_1 \text{ and } B)/P(A_1).$$

$$P(B|A_1) = (P(B) - P(B|A_2) * P(A_2))/P(A_1) = (0.92 - 1 * 0.33)/0.67 = 0.88$$

8. Suppose 60% of Duke students are from North Carolina, 30% are from other states in the U.S., and the rest are from other countries. Also suppose that 80% of in-state students know how Fayetteville got its name, but only 10% of students from other states know and only 1% of non-nationals know. So if your date explains that people wanted to honor the Marquis de Lafayette’s support of the American revolution, but dropped “La” because it meant “the”, then what is the probability that your date is from a state in the U.S. but not a North Carolinian?

Let A_1 = from NC, A_2 = from other state, A_3 = other country, B = knows name. We want $P(A_2|B)$. Using Bayes’ Theorem,

$$\begin{aligned} P(A_2|B) &= P(B|A_2)P(A_2)/[P(B|A_2)P(A_2) + P(B|A_1)P(A_1) + P(B|A_3)P(A_3)] \\ &= 0.1 * 0.3 / (0.1 * 0.3 + 0.8 * 0.6 + 0.01 * 0.1) \\ &= 0.03 / 0.511 = 0.587 \end{aligned}$$

9. Suppose an urn contains 4 white balls and 2 red balls. You draw two balls without replacement. Let A be the event that the first ball is white and let B be the event that at least one ball is white.

What is the probability of A?

$$P(A) = 4/6 = \mathbf{0.67}$$

What is the probability of B?

$$\begin{aligned} P(B) &= P(W \text{ and } \bar{W}) + P(\bar{W} \text{ and } W) + P(W \text{ and } W) \\ &= (4/6 * 2/5) + (2/6 * 4/5) + (4/6 * 3/5) \\ &= 8/30 + 8/30 + 12/30 = 28/30 = 0.93 \end{aligned}$$

What is the probability of A or B?

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) = 0.67 + 0.93 - P(A \text{ and } B). \\ P(A \text{ and } B) &= P(W \text{ and } \bar{W}) + P(W \text{ and } W) = 8/30 + 12/30 = 2/3. \\ \text{So } P(A \text{ or } B) &= 0.67 + 0.93 - 0.67 = 0.93. \end{aligned}$$

What is the probability of A and B?

$$P(A \text{ and } B) = 0.67.$$

What is the probability of A given B?

$$P(A|B) = P(A \text{ and } B)/P(B) = 0.67/0.93 = 0.72$$

10. Among 50 women, 35 like statistics. Among 40 men, 20 like statistics. Is there evidence that more women than men enjoy statistics courses?

What is the null hypothesis (in words)?

$H_0 : p_w - p_m < 0$ or the proportion of women who like statistics is less than or equal to the proportion of men who like statistics.

What is the formula for your test statistic?

$$ts = (\hat{p}_w - \hat{p}_m - 0) / \sqrt{\frac{\hat{p}_w(1-\hat{p}_w)}{n_w} + \frac{\hat{p}_m(1-\hat{p}_m)}{n_m}}$$

What is the value of your test statistic?

$$ts = (0.7 - 0.5 - 0) / \sqrt{(0.7 * 0.3)/50 + (0.5 * 0.5)/40} = 0.2 / \sqrt{0.01045} = \mathbf{1.96}.$$

What kind of distribution does your test statistic have under the null?

$Z =$ standard normal

What is the significance probability?

$$P(Z > 1.96) = 1/2(100 - 95) = \mathbf{2.5\%}$$

What is your conclusion?

Since p-value is small reject the null and conclude there is evidence that the proportion of women who like statistics is greater than the proportion of men who like statistics.

11. Suppose you classify 100 random teenagers according to whether or not they have had an accident in the last year, and 200 random elderly drivers according to whether they have had an accident in the last year. You find that 40 teenagers have had accidents, and only 30 elderly drivers have had accidents.

Write the contingency table.

| Category | Teenagers | Elderly | Total |
|-------------|-----------|---------|-------|
| Accident | 40 | 30 | 70 |
| No Accident | 60 | 170 | 230 |
| Total | 100 | 200 | 300 |

What is the null hypothesis?

H_0 : The age of driver (teenager,elderly) and frequency of accidents are independent.

What is the alternative hypothesis (in words)?

H_A : The age of driver (teenager,elderly) and frequency of accidents are not independent.

What is the formula for your test statistic?

$$\text{Chi-Square Test of Independence} = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

What is the value of your test statistic?

$$E_{ij} = \frac{(i\text{th row sum}) * (j\text{th column sum})}{\text{total}} \text{ so we get:}$$

$$E_{11} = 70 * 100/300 = 23.33$$

$$E_{12} = 70 * 200/300 = 46.67$$

$$E_{21} = 230 * 100/300 = 76.67$$

$$E_{22} = 230 * 200/300 = 153.33$$

so $ts = (40 - 23.33)^2/23.33 + \dots + (170 - 153.33)^2/153.33 = 23.30$.

What kind of distribution does your test statistic have under the null?

Chi-Square on $k = (nrows - 1) * (ncolns - 1) = (2 - 1) * (2 - 1) = 1$ df.

What is the significance probability?

For a chi-squared random variable with 1 degree of freedom, the 1% value is 6.64. So

$$.01 = P[W > 6.64] > P[W > 23.30] = P\text{-value.}$$

What is your conclusion?

p-value is small so we reject the null and conclude that age of driver and frequency of accidents are not independent.

What confounding factor could explain this relationship in a way favorable to the teenagers?

amount of time spent driving, time of day when frequently drive

12. You own a casino. To test a new supplier of dice, you choose one at random and roll it 100 times. You get 12 ones, 16 twos, 20 threes, 14 fours, 20 fives, and the rest are sixes.

What is the null hypothesis (in words)?

H_0 : The die is fair, i.e., the proportion of 1:6 is 1/6 each.

What is the alternative hypothesis (in words)?

H_a : The die is not fair, i.e., the proportions of 1:6 is different from 1/6 each.

What is the formula for your test statistic?

$$ts = \sum \frac{(O_i - E_i)^2}{E_i}$$

What is the value of your test statistic?

$E_i = 16.67$ for all i so

$$ts = ((12 - 16.67)^2)/16.67 + ((16 - 16.67)^2)/16.67 + \dots + ((18 - 16.67)^2)/16.67 = 3.199$$

What kind of distribution does your test statistic have under the null?

Chi-Square on $k = 5$ df

What is the significance probability?

On $k = 5$, $P(ts > 11.07) = 0.05 < P(ts > 3.199)$

What is your conclusion?

fail to reject; not enough evidence to conclude that distribution different from 1/6 each, i.e., that die not fair

13. You draw a sample of size 100 without replacement from a class of 120 students and you measure their IQs. You find that the sample mean is 115 and the standard deviation is 10.

What is the value of the finite population correction factor?

$$FPCF = \sqrt{(B - n)/(B - 1)} = \sqrt{(120 - 100)/(120 - 1)} = \sqrt{20/119} = 0.41$$

Use the fpcf in setting a 90% confidence interval on the mean IQ in the class.

$$L = \mathbf{114.3235}$$

$$L = \bar{X} - (FPCF * SD/\sqrt{n}) * z_{0.05} = 115 - (0.41 * 10/\sqrt{100}) * 1.65 \\ = 115 - 0.6765 = 114.3235$$

$$U = \mathbf{115.6765}$$

$$U = \bar{X} + (FPCF * SD/\sqrt{n}) * z_{0.05} = 115 + (0.41 * 10/\sqrt{100}) * 1.65 \\ = 115 + 0.6765 = 115.6765$$

14. Name the three principles laid out in the Belmont Report:

Respect for individuals

Beneficence

Justice

15. You do a regression analysis on 120 students that attempts to predict the score on a person's exam from the amount of time (in hours) spent studying the night before. You find that the intercept is 80, the slope is -7, the correlation coefficient is -.8, and the standard deviation of the residuals is 3.

What is your estimate of the grade for someone who spends 10 hours studying?

$$\hat{y} = 80 - 7(10) = \mathbf{10}$$

Set a 90% upper confidence interval on the grade of someone who spent 10 hours studying.

$$\hat{y} + SD * z_{0.05} = 10 + 3 * 1.65 = \mathbf{14.95}$$

What proportion of the variance in grade is explained by knowing how many hours were spent studying?

$$R^2 = \text{proportion of variance in grade explained by hours spent studying} = (-0.8)^2 = \mathbf{0.64}$$

If you were doing multiple regression, what additional variable would help predict the response?

grade on previous exam, amount of sleep, etc.

16. Each day my son has probability .4 of doing something crazy. His behavior from one day to the next is independent. What is the probability that he is crazy on exactly 6 of the next 10 days?

binomial with $n = 10$, $r = 6$, $p = 0.4$ so

$$P(\text{exactly 6 success in 10 trials}) = \binom{10}{6} (0.4)^6 (0.6)^4 = 210 * 0.004096 * 0.1296 = \mathbf{0.1115}$$

17. Who was the greatest statistician ever?

Fisher