Midterm Exam - Key

Math 530/630

Due: 11/20/2018

To turn in, please submit 2 files: one is the raw .Rmd file and the other is a single knitted html(or PDF) file. For the multiple choice, please clearly indicate the problem number and your answer on separate lines as in:

1. a
2. b
3. c

Grading:

- Free Answer questions are each worth 4 points.
- Multiple Choice are each worth 2 points.

You may include R chunks for the multiple choice problems if you like, but partial credit will not be given.

Key Terms You Should Know

Alpha-level (α): probability of making a Type I error; that is, rejecting the null hypothesis when it is in fact true (e.g., a test indicates that a man is pregnant)

Alternative Hypothesis (H_A or H_1): the prediction that there will be an effect.

Critical Value(s): the value of a given test statistic that corresponds to a rejection point- the point at which you determine to reject the null hypothesis. The critical value defines the boundary of the rejection region for H_0. This value depends on the significance level, α, and whether the test is one-sided or two-sided.

Null Hypothesis (H_0): the reverse of the alternative hypothesis that your prediction is wrong and the predicted effect does not exist.

Free Answer Questions

1. Given the probability density function (below) of the random variable X:

   a. find c
   b. find the cumulative distribution function $F(x)$.
   c. Compute $P(1 < X < 3)$

$$f(x) = \begin{cases} cx, & 0 < x < 4 \\ 0, & \text{otherwise} \end{cases}$$

Answers

a)

$$\begin{aligned} \int_0^4 cxdx &= 1 \\ 8c &= 1 \\ c &= \frac{1}{8} \end{aligned}$$

b) As $\int \frac{1}{8}xdx$ is $\frac{x^2}{16}$,

x	$F(X)$
$x < 0$	0
$0 \leq x \leq 4$	$\frac{x^2}{16}$
$x > 4$	1

c)

$$\begin{aligned} P(1 < x < 3) &= F(3) - F(1) \\ &= \frac{9}{16} - \frac{1}{16} \\ &= \frac{1}{2} \end{aligned}$$

2. A friend claims that she has drawn a random sample of size 30 from the exponential distribution with $\lambda = \frac{1}{10}$. The mean of her sample is 12.

   a. What is the expected value of a sample mean? (hint: https://en.wikipedia.org/wiki/Exponential_distribution#Mean.2C_variance.2C_moments_and_median)
   b. Run a simulation by drawing 1000 random samples, each of size 30 from Exp(1/10), and then compute the mean. What proportion of the sample means are as large or larger than 12?
   c. Is a mean of 12 unusual for a sample of size 30 from Exp(1/10)?

Answers

(a) 10.

(b) set.seed(0) my.means<-numeric(1000) for (i in 1:1000) { x<-rexp(30, 1/10) my.means[i]<-mean(x) } sum(my.means >= 12)/1000 # .132

(c) No.

3. A researcher believes that in recent years women have been getting taller. She knows that 10 years ago the average height of young adult women living in her city was 63 inches ($\sigma = 3$). She randomly samples eight young adult women currently residing in her city and measures their heights. The following data are obtained (Height, in inches): 64, 66, 68, 60, 62, 65, 66, 63

   a. What is the alternative hypothesis? (In evaluating this experiment, assume a non-directional hypothesis is appropriate because there are insufficient theoretical and empirical bases to warrant a directional hypothesis.)
   b. What is the null hypothesis?
   c. Using the generic test statistic formula given in class ($\frac{\hat{\theta}-\theta_0}{SE_{\theta_0}}$), what is the appropriate numerator?
   d. Using the generic test statistic formula given in class ($\frac{\hat{\theta}-\theta_0}{SE_{\theta_0}}$), what is the appropriate denominator?
   e. Using $\alpha = 0.01$, what is/are the critical value(s) of the test statistic?
   f. Using $\alpha = 0.01$, what is the conclusion? (i.e. do we reject or fail to reject the null hypothesis, and why?)
   g. Looking at the actual data, how does the sample standard deviation compare to that of the population?

```r
(mu <- 63)
```

```
## [1] 63
```

```r
(sd_mu <- sqrt(3))
```

```
## [1] 1.732051
```

```r
(heights <- c(64, 66, 68, 60, 62, 65, 66, 63))
```

```
## [1] 64 66 68 60 62 65 66 63
```

```r
(n <- length(heights))
```

```
## [1] 8
```

```r
(sd_heights =  sd(heights))
```

```
## [1] 2.54951
```

```r
(x_obt <- mean(heights))
```

```
## [1] 64.25
```

```r
(se <- sd_heights/sqrt(n))
```

```
## [1] 0.9013878
```

```
(t <- (x_obt-mu)/se)
```

```
## [1] 1.38675
```

```
(t_crits <- c(qt(.005,n-1),qt(.995,n-1)))
```

```
## [1] -3.499483  3.499483
```

```
(ci <- x_obt+se*t_crits)
```

```
## [1] 61.09561 67.40439
```

Answers

a) In recent years the height of women has been changing. Therefore the sample with $\bar{X_{obt}} = 64.25$ is a random sample from a population where $\mu \neq 63$
b) It is reasonable to consider the sample with $\bar{X_{obt}} = 64.25$ is a random sample from a population where $\mu = 63$

Use the single sample t-test on means. Not the z-test, as n<30.

c) $\bar{X_{obt}} - \mu$, 64.25-63
d) $\frac{sd_{obt}}{\sqrt{n}}$, sd(heights)/sqrt(8)
e) df = n-1 = 7, $t_{crit} = \pm 3.499$
f) SE = 2.55, $t_{obt} = 1.39$, since $\mid t_{obt} \mid < \mid t_{crit} \mid$, we retain the null hypothesis.
g) the sample standard deviation is larger than that of the population. Perhaps heights are more varied?

“‘

4. The mean height reported by men on the dating website OKCupid is approximately 5 feet 11 inches. For men living in the US, heights are normally distributed with a mean of 5 feet 9 inches, σ=3 inches. Answer the following:

    a. State the null and alternative hypotheses.
    b. What is/are the critical value(s) of that test statistic (assume $\alpha = .05$, 2-tailed)? That is, for which values of the test statistic would you reject the null hypothesis?

Answers

a)

- H0 = OKCupid and US men have the same height
- H1 = OKCupid and US men have different heights

b)

```
ll <- qnorm(.025,69,3)
ul <- qnorm(.975,69,3)
cbind(ll,ul)
```

```
##             ll       ul
## [1,] 63.12011 74.87989
```

5. What are the critical values of a t-distributed random variable (`?TDist`) for each of the following values of N and α using nondirectional hypotheses (assume testing of means)?

    a. N=12; α=.05
    b. N=20; α=.01
    c. N=2; α=.05

    What are the critical values of a t-distributed random variable for each of the following values of N and α using a directional hypothesis in the upper tail (assume testing of means)?

    d. N=8; α=.05
    e. N=15; α=.01
    f. N=51; α=.025

Answers

```
# a)
cbind(qt(.025,11), qt(.975,11))
```

```
##           [,1]     [,2]
## [1,] -2.200985 2.200985
```

```
# b)
cbind(qt(.005,19), qt(.995,19))
```

```
##           [,1]     [,2]
## [1,] -2.860935 2.860935
```

```
# c)
cbind(qt(.025,1), qt(.975,1))
```

```
##          [,1]    [,2]
## [1,] -12.7062 12.7062
```

```
# d)
qt(.95,7)
```

```
## [1] 1.894579
```

```r
# e)
qt(.99,14)
```

```
## [1] 2.624494
```

```r
# f)
qt(.975,50)
```

```
## [1] 2.008559
```

Multiple-Choice Questions

1. Suppose you have 10 numbers and have computed the mean to be 8.0. You then discover that the last number in the data was entered incorrectly. It was entered as 8.0 when it should have been 4.0. If you replace the incorrect value (8.0) with the correct one (4.0), and recompute the mean, you will obtain a new mean of:

   a. It is impossible to determine
   b. 6.6
   c. 8.6

- d. 7.6

2. You have a set of data that have a mean of 50 and a standard deviation of 12. You wish them to have a mean of 65 and a standard deviation of 10, while retaining the shape of the distribution. What values of a and b in the linear transformation formula $Y = aX + b$ will produce a new set of data with the desired mean and standard deviation?

- a. a = 0.833, b = 23.3

   b. a = 1.83, b = -23.3

   c. a = 1.67, b = 23.3

   d. a = 0.833, b = 46.7

   e. a = 1.2, b = 22.3

3. Jane had a z-score of 1.75 on her statistics midterm. If the class mean is 65.0, and the class standard deviation is 12.0, what was Jane's raw score?

   a. It is impossible to determine
   b. 66

- c. 86
   d. 76

4. IQ scores have a distribution that is approximately normal in shape, with a mean of 100 and a standard deviation of 15 in the general population. Assuming the normal distribution is a good approximation, what proportion of the general population has IQ scores between 79.0 and 109.0?

- a. 0.645

   b. 0.745

   c. 0.545

   d. 0.709

   e. None of the above answers are correct

```r
pnorm(109, 100, 15) - pnorm(79, 100, 15)
```

```
## [1] 0.6449902
```

5. You have 10 numbers with a sample mean of 9.0 and a sample variance of 11.0. You discover that the last number in the list was recorded as 8.0 when it should have been recorded as 12.0. If you correct your error and correctly recompute the sample variance, what value will you obtain?

- a. 11.71

  b. 11.0

  c. 10.54

  d. 13.01

  e. None of the above answers are correct

6. If $\alpha = 0.01$, then the probability of a correct acceptance of a true statistical null hypothesis is:

   a. 0.0001
   b. β

- c. 0.99
  d. 0
  e. It cannot be determined from the information provided

7. Given the following probability distribution for the random variable X, the variance of X is:

   a. 3.5563
   b. 3.7524
   c. 2.95
   d. 1.575

- e. 1.8475

x	$P_X(x)$
1	0.1
2	0.15
3	0.2
4	0.2
5	0.35

8. Given the following probability distribution for the random variable X, the expected value of X is:

   a. 2.48
   b. 3.68
   c. 3.58

- d. 3.48

x	$P_X(x)$
1	0.1
2	0.2
3	0.2
4	0.12
5	0.38

9. The sampling distribution of the sample mean based on N iid observations

a. converges asymptotically to a normal distribution in shape under the conditions of the Central Limit Theorem
b. has a variance of σ^2/N for any population distribution
c. is always exactly normal, for any sample size, when the population distribution is normal

- d. all of the above answers are correct

10. If one draws all possible samples for various values of N from the same population of raw scores, as N increases:

a. The standard error of the mean increases
b. The standard error of the mean stays the same

- c. The standard error of the mean decreases

d. The standard error of the mean cannot be calculated

11. If one draws all possible samples for various values of N from the same population of raw scores, as N increases:

a. The mean of the sampling distribution of the mean increases

- b. The mean of the sampling distribution of the mean stays the same

c. The mean of the sampling distribution of the mean decreases
d. None of the above

12. In cases where $N > 1$, the relationship between the raw score population standard deviation and the standard error is:

a. The standard error is greater than the population standard deviation

- b. The standard error is less than the population standard deviation

c. The standard error equals the population standard deviation
d. The standard error is the population standard deviation

13. The variance can be thought of as:

a. half the range
b. the sum of squared deviations from the mean
c. the average deviation

- d. the average squared deviation from the mean

14. What would happen to the mean of a distribution of scores if the number 10 is added to each score?

a. It would stay the same

- b. It increases by 10

c. It will become 10 times as large
d. It will increase, but the amount depends on the shape of the distribution

15. What would happen to the standard deviation of a distribution of scores if the number 10 is added to each score?

- a. It would stay the same

b. It increases by 10

c. It will become 10 times as large

d. It will increase, but the amount depends on the shape of the distribution

16. What would happen to the mean of a distribution of scores if each score is multiplied by 2?

a. It would stay the same
b. It increases by 2

- c. It will become twice as large
  d. It will increase, but the amount depends on the shape of the distribution

17. What would happen to the variance of a distribution of scores if each score is multiplied by 2?

    a. It would stay the same
    b. It will become twice as large

- c. It will become four times as large
  d. It will increase, but the amount depends on the shape of the distribution

18. A researcher has data for two variables, x and y. First, she converts both variables to z-scores with a mean of 0 and standard deviation of 1, and calls them z_x and z_y. Next, she takes the mean of both z-scores and calls that new variable ave_z. What is the mean and standard deviation of the ave_z variable?

    a. mean = 0; standard deviation = 1
    b. mean = 0; standard deviation approximately 1

- c. mean = 0; standard deviation > 1
  d. Cannot be determined from the information given

```
library(dplyr)
options(scipen=999)
x <- c(0, -1, 1, -5, 10)
y <- c(-2, -3, 5, -7, 7)
z_x <- scale(x)
z_y <- scale(y)
my_z <- data.frame(z_x, z_y)
my_z %>%
  mutate(ave_z = z_x + z_y / 2) %>%
  summarise_each(funs(mean, sd))
```

```
##                      z_x_mean z_y_mean                    ave_z_mean z_x_sd
## 1 -0.0000000000000000388578        0 -0.0000000000000000222044 6         1
##   z_y_sd ave_z_sd
## 1      1 1.458513
```

19. The sampling distribution and standard error of a statistic can be calculated by:

    a. exhaustive and exact calculations where formula solutions are possible
    b. simulation
    c. formula approximations
    d. repeatedly taking random samples of a given size from a population

- e. all of the above

20. IQ scores have a distribution that is approximately normal in shape, with a mean of 100 and a standard deviation of 15. What percentage of scores is at or above an IQ of 116?

    a. 12.464

- b. 14.306
  c. 15.737
  d. 16.355
  e. None of the above answers are correct.

```
1-pnorm(116,100,15)
```

```
## [1] 0.1430612
```

21. Suppose you want to test the null hypothesis that μ=100 with a sample size of n=25 and an α=.05 using a t-statistic, which you know follows the Student t-distribution (`?TDist`). What will the critical value(s) for the t statistic be? That is, for which values of the t statistic will you reject the null hypothesis?

- a. $\leq$ -2.06, $\geq$ 2.06

  b. $\leq$ -1.71, $\geq$ 1.71

  c. $\leq$ -2.06, $\geq$ 1.85

  d. $\leq$ -1.85, $\geq$ 2.06

  e. None of the above answers are correct.

```
ul <- qt(.975, 24)
ll <- qt(.025, 24)
cbind(ll,ul)
```

```
##             ll       ul
## [1,] -2.063899 2.063899
```

22. Could the sample X={21,21,21,20,22,20,22} reasonably have been drawn from a normal population with a mean of 20 and standard deviation of 1.5 with α=0.05 (two-tailed)?

- a. yes

  b. no

  c. cannot be tested with z test

  d. insufficient information

For this set of questions, suppose you have four children in a reading group (Beth, Marianne, Steven, Joel) and you randomly pick one child to lead the discussion in group each day of a 5-day week. Furthermore, we define the outcome of each day's selection to be binary: Steven leads the discussion or he doesn't.

23. The number of times Steven leads the discussion in a week would be the:

  a. probability distribution for this experiment
  b. probability of an outcome
  c. constant in this experiment

- d. random variable in this experiment

24. The probability that Steven leads the discussion all 5 days in a week is the:

  a. expected value
  b. probability distribution

- c. probability of a simple event
  d. random variable

25. If we could replicate this experiment many, many times, the average number of times that Steven leads the discussion in a week would be the:

- a. expected value

  b. probability distribution

  c. probability of an outcome

  d. random variable

26. If we found the probability of Steven leading the discussion zero times, one time, two times, three times, four times, and five times, the set of six probabilities would be the:

a. expected value

- b. probability distribution
  c. probability of an outcome
  d. random variable

27. Monday's selection of a discussion leader could be considered a ______, while the selections of discussion leaders for the week constitute a ______:

    a. binomial experiment, bernoulli trial

- b. bernoulli trial, binomial experiment
  c. neither

28. What is the probability that Steven would be selected all 5 days of the week?

    a. .00000000000000

- b. .0009765625
  c. .0039065500
  d. .25

```
.25^5 #or
```

```
## [1] 0.0009765625
```

```
dbinom(5, 5, .25) #or
```

```
## [1] 0.0009765625
```

```
1 - pbinom(4, 5, .25) #or
```

```
## [1] 0.0009765625
```

```
pbinom(4, 5, .25, lower.tail = FALSE)
```

```
## [1] 0.0009765625
```

29. What is the expected number of times that Steven would be selected?

    a. 0.00
    b. 1.00

- c. 1.25
  d. 2.50

```
n <- 5
p <- .25
n*p # definition of expected value for a binomial rv
```

```
## [1] 1.25
```

---

30. Using the following code, run a permutation test with the hypothesis that mean Hotwing consumption is different for men than for women (assume $\alpha = .05$, 2-tailed). Does the data support your hypothesis, and what is the associated p-value?

    a. The data supports H_A, p = 0.00078

- b. The data supports H_A, p = 0.00156
  c. We fail to reject H_0, p = 0.156
  d. We fail to reject H_0, p = 0.780

```r
#install.packages("resampledata")
library(resampledata)
library(dplyr)
```

```r
glimpse(Beerwings)
```

```
## Observations: 30
## Variables: 4
## $ ID       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ Hotwings <int> 4, 5, 5, 6, 7, 7, 7, 8, 8, 8, 9, 11, 11, 12, 12, 13, ...
## $ Beer     <int> 24, 0, 12, 12, 12, 12, 24, 24, 0, 12, 24, 24, 24, 30,...
## $ Gender   <fct> F, F, F, F, F, F, M, F, M, M, F, F, M, F, F, F, F, M,...
```

```r
# observed mean difference
Beerwings %>%
  group_by(Gender) %>%
  summarise(mean_wings = mean(Hotwings))
```

```
## # A tibble: 2 x 2
##   Gender mean_wings
##   <fct>       <dbl>
## 1 F            9.33
## 2 M           14.5
```

Run the permutation:

```r
set.seed(0)
B <- 10^5-1  #set number of times to repeat this process
result <- numeric(B) # space to save the random differences
for(i in 1:B){
  index <- sample(30, size=15, replace = FALSE) # sample of numbers from 1:30
  result[i] <- mean(Beerwings$Hotwings[index]) - mean(Beerwings$Hotwings[-index])
}
```

```r
# As calculated.
# ...and some commented out alternatives that are equivalent at 4 decimal places :)
observed <- 14.5333- 9.3333
#!! Only use the count from the smaller tail, and then double it
min_sum <- min(sum(result >= observed), sum(result <= -observed))
#min_sum <- min(sum(result > observed), sum(result < -observed))
#Compute P-value (adjusted p-value)
min_p <- sum(min_sum + 1)/(B + 1)
#Unadjusted p-value
#min_p <- min_sum/B
2*min_p
```

```
## [1] 0.00156
```

```r
#Using infer

observed <- 14.5333- 9.3333
library(infer)
null_distn_two_means <-Beerwings %>%
  specify(Hotwings ~ Gender) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100000) %>%
  calculate(stat = "diff in means",
```

```r
            order = c("M", "F"))

pvalue <- null_distn_two_means %>%
  get_pvalue(obs_stat = observed, direction = "both")
pvalue
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1 0.00147
```