

STA130 Term Test – Section LEC0101

March 2, 2018

Last Name: SOLUTIONS

Last Name: _____

Student number: _____

Tutorial section or teaching assistant's name: _____

Instructions:

- Total marks: 45
- There are 11 pages including this page.
- Backs of pages will not be marked. Write your answers only in the space provided.
- The test is 90 minutes long.
- You are permitted a 1-sided, handwritten, 8.5x11 inch aid sheet. You must hand in your aid sheet with your test.
- Calculators are not permitted.
- Questions begin on the next page.

Question 1

[5 marks] Suppose you type commands into the R console as shown in the table below. For each of the R commands, give the output that R will produce, or an example of output that might be produced from the commands.

R command	Output
<code>5+7</code>	<i>12</i>
<code>vec <- c(2, 6, 7, 11)</code> <code>vec[2]</code>	<i>6</i>
<code>vec <- c(2, 6, 7, 11)</code> <code>quantile(vec, 0.5)</code>	<i>6.5</i>
<code>vec <- c(2, 6, 7, 11)</code> <code>sample(vec)</code>	<i>7, 2, 11, 6</i>
<code>x <- rep(NA, 3)</code> <code>for (i in 1:3)</code> <code>{</code> <code>x[i] <- i</code> <code>}</code> <code>x</code>	<i>1, 2, 3</i>

Data that will be used for the remainder of the questions

For the rest of this test, we will work with the `ncbirths` data set. This data set contains information on a random sample of 1000 births recorded in the state of North Carolina in 2004. The data set contains the following variables:

- `fage`: the age of the father in years at the time of the birth
- `mage`: the age of the mother in years at the time of the birth
- `weeks`: the length of the pregnancy in weeks
- `premie`: whether the birth was classified as premature (`premie`) or `full term`
- `gained`: weight gained by mother during the pregnancy in pounds
- `weight`: weight of the baby at birth in pounds
- `gender`: gender of the baby, `female` or `male`
- `habit`: status of the mother as a `nonsmoker` or a `smoker`

Here is a look at the data:

```
glimpse(ncbirths)
```

```
## Observations: 999
## Variables: 8
## $ fage   <int> NA, NA, 19, 21, NA, NA, 18, 17, NA, 20, 30, NA, NA, NA,...
## $ mage   <int> 13, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16,...
## $ weeks  <int> 39, 42, 37, 41, 39, 38, 37, 35, 38, 37, 45, 42, 40, 38,...
## $ premie <fctr> full term, full term, full term, full term, full term,...
## $ gained <int> 38, 20, 38, 34, 27, 22, 76, 15, NA, 52, 28, 34, 12, 30,...
## $ weight <dbl> 7.63, 7.88, 6.63, 8.00, 6.38, 5.38, 8.44, 4.69, 8.81, 6...
## $ gender <fctr> male, male, female, male, female, male, male, male, ma...
## $ habit  <fctr> nonsmoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker,...
```

```
table(ncbirths$premie)
```

```
##
## full term    premie
##      846      152
```

```
table(ncbirths$gender)
```

```
##
## female    male
##     502     497
```

```
table(ncbirths$habit)
```

```
##
## nonsmoker    smoker
##      873      126
```

Questions about this data set begin on the next page.

Question 2

[3 marks] For data wrangling, we have worked with the following functions:

- `arrange()`
- `filter()`
- `group_by()`
- `mutate()`
- `select()`
- `summarise()`

Name which of these functions is appropriate to use for each of the following tasks. You do not need to write the entire R command, just state which of the above functions you would use.

- (a) Create the variable `lowbirthweight` with is "low" for babies whose birth weight (in the variable `weight`) is 5.5 pounds or less, and "not_low" for babies whose birth weight is greater than 5.5 pounds

Function to use: *mutate*

- (b) Remove the one observation that has a missing value of `habit`

Function to use: *filter*

- (c) Create a dataset that includes only the variables that are measures on the mother and baby (and not the father)

Function to use: *select*

Question 3

- (a) [2 marks] In `ggplot` we have used the following plotting geometries:
`geom_histogram()`, `geom_point()`, `geom_bar()`, `geom_boxplot()`

Name which of these geometries would be appropriate to use to construct a plot that shows:

- i. the distribution of `gender`

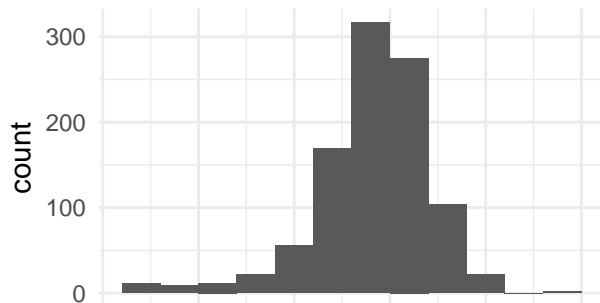
Geometry to use: `geom_bar()`

- ii. the relationship between `gained` and `weight`

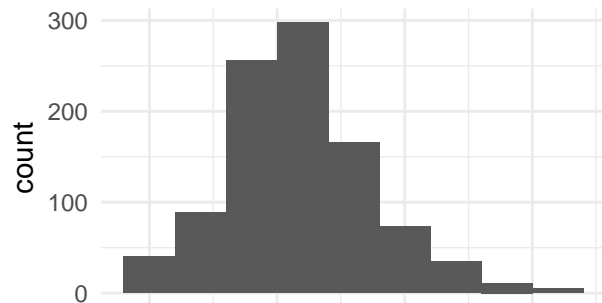
Geometry to use: `geom_point()`

- (b) [2 marks] One of the following histograms shows the distribution of `weight` and one shows the distribution of `gained`. One of the following boxplots shows the distribution of `weight` and one shows the distribution of `gained`.

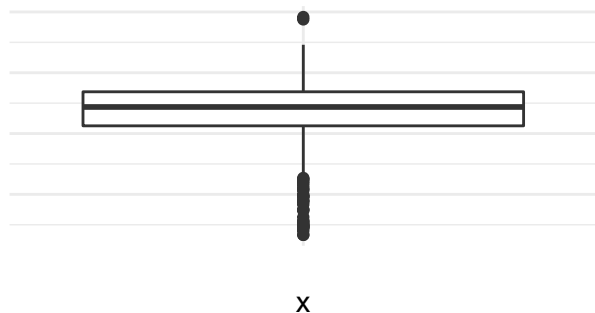
Histogram 1



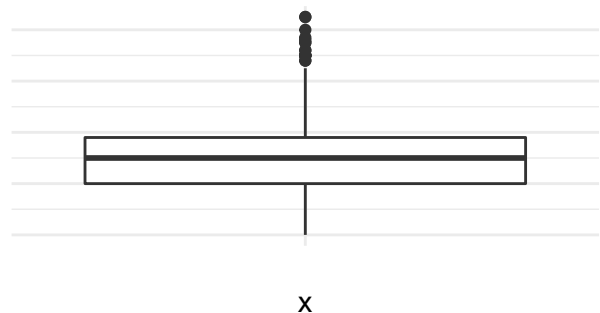
Histogram 2



Boxplot 1



Boxplot 2



Which histogram corresponds to which boxplot? Justify your answer.

Histogram 1 corresponds to boxplot 1 and histogram 2 corresponds to boxplot 2. The long left tail in histogram 1 corresponds to the dots beyond the lower whisker in boxplot 1 and the long right tail in histogram 2 corresponds to the dots beyond the upper whisker in boxplot 2.

Question 4

In this question, we will investigate whether smoking during pregnancy has an effect on the gender of the baby. Specifically, we'll investigate whether the proportion of babies born who are girls is the same for smoking and non-smoking mothers. The table below summarizes the numbers of mothers who smoked or did not smoke while pregnant and the gender of their babies.

##			
##		female	male
##	nonsmoker	445	428
##	smoker	57	69

- (a) [2 marks] Why is this table of numbers not tidy data?

*In tidy data, each observation must have its own row. For these data, observations are individual babies. In this table, the rows are aggregates of the observations, counting the number in each category of **habit**. (An alternative argument: in tidy data, each variable must have its own column, but the two columns here are values of one variable (**gender**).)*

- (b) [2 marks] We want to carry out an hypothesis test to investigate the question: Is the proportion of babies born who are girls the same for smoking and non-smoking mothers? State an appropriate null hypothesis. Define the parameter(s) you are testing.

$$H_0 : p_1 = p_2$$

where p_1 is the proportion of girls for smoking mothers and p_2 is the proportion of girls for non-smoking mothers in the population

- (c) [2 marks] In order to carry out this hypothesis test, we will carry out a simulation. Indicate (by circling) whether each of the following statements about this simulation is TRUE or FALSE.

- i. TRUE or FALSE: The distribution of the simulated values will be centred at

$$\frac{445}{445+428} - \frac{57}{57+69}.$$

This is false. (The distribution of the simulated values will be centred at 0.)

- ii. TRUE or FALSE: For each simulated value, we first shuffle the values of **habit** in the original data.

This is true.

(d) The simulation was run and the P-value for the test on the previous page is 0.27.

i. [2 marks] Explain what is wrong with the following description of the P-value:

“The probability that the null hypothesis is true is 0.27.”

The P-value does not give the probability of the null hypothesis. It gives the probability of observing the data we got (as summarized by the test statistic) or a value more extreme, under the assumption that the null hypothesis is true.

ii. [1 mark] Which one of the following statements is possibly true about this hypothesis test?
Circle the statement.

A. We conclude that there is **strong** evidence that the proportion of babies who are females is different for smoking and non-smoking mothers. It is possible that the test resulted in a **Type 1** error.

B. We conclude that there is **no** evidence that the proportion of babies who are females is different for smoking and non-smoking mothers. It is possible that the test resulted in a **Type 1** error.

C. We conclude that there is **strong** evidence that the proportion of babies who are females is different for smoking and non-smoking mothers. It is possible that the test resulted in a **Type 2** error.

D. We conclude that there is **no** evidence that the proportion of babies who are females is different for smoking and non-smoking mothers. It is possible that the test resulted in a **Type 2** error.

D is correct. (The P-value is large, indicating that there is no evidence against the null hypothesis, so no evidence of a difference. Type 2 errors are possible when we don't reject the null hypothesis.)

Question 5

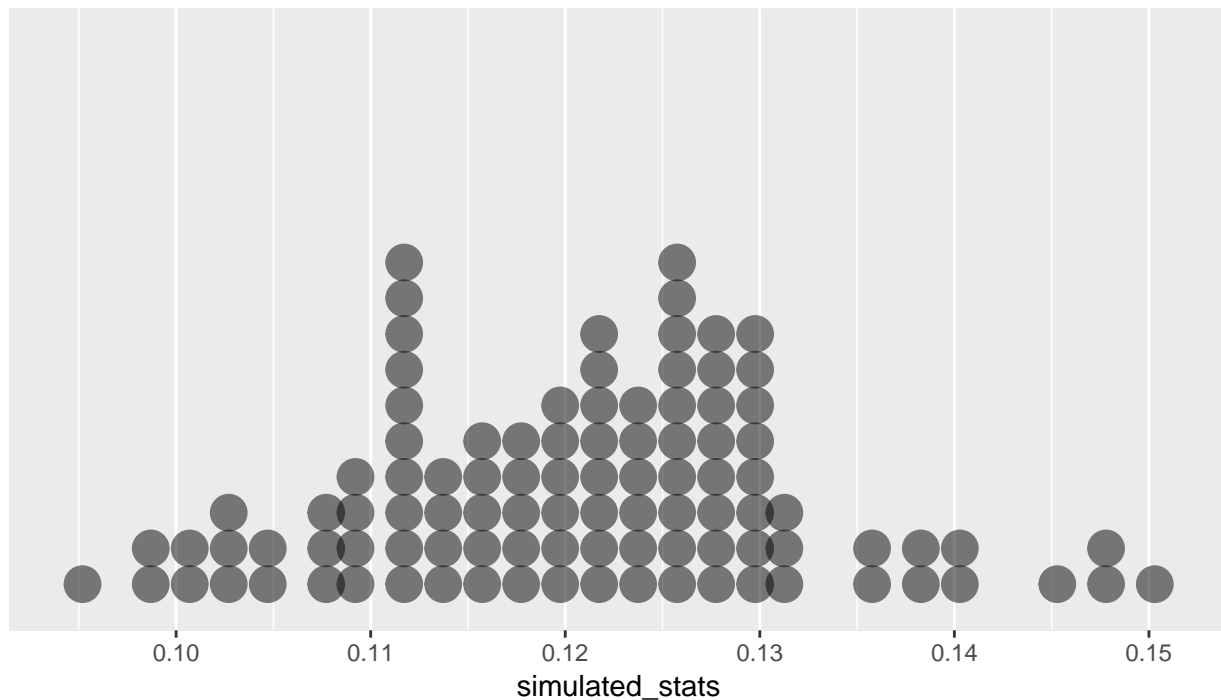
Premature births occur in 12% of all births in the United States. In this question, we will conduct an hypothesis test to test whether we have evidence that the proportion of premature births in North Carolina is statistically different than the national proportion of 12%. In our data, for the 998 observations for which we have a value of `premie`, 152 or 15.2% of babies are premature.

The R code for the simulation to carry out the hypothesis test is below. The values in `simulated_stats` are plotted below the code.

```
repetitions <- 100
simulated_stats <- rep(NA, repetitions)

for (i in 1:repetitions)
{
  new_sim <-
    sample(c("premie", "full_term"), size=n_observations,
           prob=c(0.12, 0.88), replace=TRUE)
  sim_p <- sum(new_sim == "premie") / n_observations
  simulated_stats[i] <- sim_p
}
```

Plot of the values in `simulated_stats`



The questions regarding this output begin on the next page.

- (a) i. [2 marks] If the null hypothesis changed, could you estimate the P-value for this test from this plot or would you need to run a new simulation? If yes, indicate how. If you need a new simulation, indicate which lines of code would need to change.

No you could not estimate the P-value from this plot for a different null hypothesis. You would need to change the probability vector in this line of code that simulates what the data could be if the null hypothesis were true:

```
sample(c("premie", "full_term"), size=n_observations, prob=c(0.12, 0.88), replace=TRUE)
```

- ii. [2 marks] If the test statistic changed, could you estimate the P-value for this test from this plot or would you need to run a new simulation? If yes, indicate how. If you need a new simulation, indicate which lines of code would need to change.

Yes you could estimate the P-value from this plot for a different test statistic. The P-value is the number of simulated values that are as extreme or more extreme than the test statistic. The simulated values don't change, just the cut-off for which are extreme.

- (b) [2 marks] Estimate the P-value as accurately as you can based on the plot.

The test statistic is 0.152. The P-value is the proportion of observations in the simulation that are greater than or equal to 0.152 or less than or equal to 0.088 (just as extreme in the left tail). There are no points this extreme so the estimated P-value is 0.

- (c) [8 marks] Summarize the results of the hypothesis test in a four to six sentence paragraph, using complete English sentences. The paragraph should address what is being tested, what was observed in the data, the P-value and what it means, and an appropriate conclusion. (If you don't know the P-value, pick a value to use to answer this question.)

We are interested in whether the proportion of premature births in all North Carolina births in 2004 was different than the national value of 12%. In our sample of births, 15.2% were premature. Our hypothesis test gave an estimated P-value of 0, indicating that our observed percentage of 15.2% would be extremely unlikely if the proportion of premature births in North Carolina had been 12%. We conclude that we have strong evidence that the proportion of premature births in North Carolina in 2004 was different from the national proportion.

Question 6

In this question, we're interested in the mean weight of all baby boys in North Carolina.

- (a) [1 mark] Which R code below creates a data frame that includes only boys and then creates a data frame that includes only the mean weight of the boys? Circle the correct code.

A.

```
boys <- ncbirths %>% select(gender == "male")
boys %>% mutate(male_mean = mean(weight))
```

B.

```
boys <- ncbirths %>% select(gender == "male")
boys %>% summarize(male_mean = mean(weight))
```

C.

```
boys <- ncbirths %>% filter(gender == "male")
boys %>% summarize(male_mean = mean(weight))
```

D.

```
boys <- ncbirths %>% filter(gender == "male")
boys %>% mutate(male_mean = mean(weight))
```

E.

```
boys <- ncbirths %>% group_by(gender)
boys %>% summarize(male_mean = mean(weight))
```

F.

```
boys <- ncbirths %>% group_by(gender)
boys %>% mutate(male_mean = mean(weight))
```

C. is correct

- (b) [2 marks] Is the mean produced by the correct code in part (a) a parameter or a statistic. Why?

It is a statistic because it is the mean of a sample of data.

- (c) The data frame `boys` includes only the boys from the `ncbirths` data. The code below produces a bootstrap confidence interval for the mean weight of all baby boys in North Carolina (NC).

```
boot_means <- rep(NA, 5000)
sample_size <- as.numeric( boys %>% summarize( n() ) )
for (i in 1:5000)
{
  boot_samp <- ncbirths %>% sample_n(size = X1, replace=X2)
  boot_means[i] <- as.numeric( boot_samp %>% summarize(mean(weight)) )
}
quantile(boot_means, c(0.005, 0.995))

##      0.5%      99.5%
## 6.978888 7.222826
```

- i. [2 marks] Two parts of the code have been replaced by X1 and X2. What should replace X1 and X2 to correctly complete the code?

X1 should be 'sample_size' (or 497) ; X2 should be TRUE

- ii. [2 marks] Fill in the blanks based on the output of the code above (that is, the output that would be produced after X1 and X2 have been appropriately replaced):

A 99 % confidence interval for the mean weight of all baby boys in NC is (6.98, 7.22)

- (d) [2 marks] TRUE or FALSE. Circle the appropriate choice.

- i. TRUE or FALSE: We would get a narrower confidence interval if there were more babies in the sample data.

TRUE

- ii. TRUE or FALSE: We would get a narrower confidence interval if we simulated more bootstrap samples.

FALSE. The bootstrap sampling distribution will have more values, but not less variability.

- (e) [1 mark] Suppose that a 95% confidence interval produced by the procedure in part (c) is (7.0, 7.2). A researcher interprets this confidence interval as follows: “With 95% confidence, we can infer that a randomly chosen North Carolina baby will weigh between 7.0 and 7.2 pounds.” Why is the researcher’s interpretation of the confidence interval incorrect? Circle the best answer.

- A. Because the babies in the sample could be many possible weights, including values less than 7.0 pounds and greater than 7.2 pounds.
- B. Because the confidence interval estimates weight for the sample and not for the population.
- C. Because the confidence interval estimates the mean weight in the population and not the weight of an individual baby.
- D. Because 95% of babies born in North Carolina will weigh between 6.98 and 7.22 pounds.

C. is correct