Homework 5: Normal and Binomial Distribution

Your name and student ID

March 02, 2021

```
BEGIN ASSIGNMENT
requirements: requirements.R
generate: true
```

```
library(testthat)
```

Instructions

- Solutions will be released on Tuesday, March 2
- This semester, homework assignments are for practice only and will not be turned in for marks.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration! **You must knit correctly before submitting.**
- To avoid code running off the page, have a look at your knitted PDF and ensure all the code fits in the file. If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

[7 points] Part 1: Pregnancy Length Probabilities

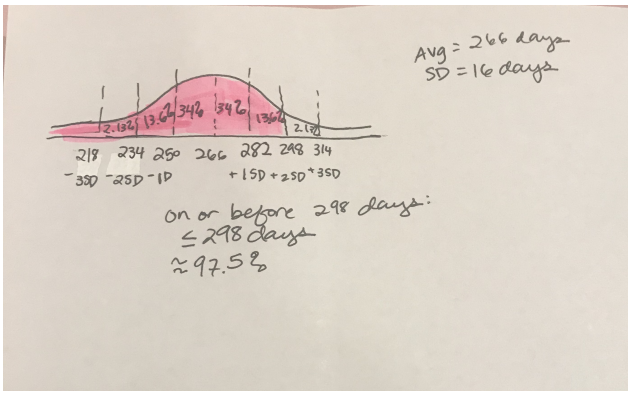
An average pregnancy for humans lasts 266 days, with a standard deviation of 16 days. Assume that human pregnancies are Normally distributed.

1. [3 marks] Approximately what proportion of births are expected to occur on or before 298 days? To aid your answer, hand-draw (or use any software) to sketch a Normal curve, and shade in the area under the Normal density curve the question represents. Add dashed lines at the mean +/- 1SD, 2SD and 3SD. Then calculate the proportion asked about in the first sentence. You shouldn't need to use R to perform any calculations for this question. Report the proportion to one decimal place.

(Use the code chunk below to include an image file of your drawing. To do so you need to delete the hashtag, upload the image to Datahub into the `src` directory and replace the file name with your file name. JPG or PNG will both work.)

```
BEGIN QUESTION
name: p1
manual: true
```

```
#knitr::include_graphics("src/Your-file-name.JPG")
# BEGIN SOLUTION NO PROMPT
knitr::include_graphics("A5_Normal-a.JPG")
```




```
# END SOLUTION
```

Students should draw the Normal density and at the days corresponding to the mean, and the mean +/- 1, 2, and 3 SD. They should notice that mean + 2SD = 298. They know that 95% of the data is between the mean +/- 2 SD, which implies that 2.5% of the data is above the mean + 2SD, or approximately 97.5% of the data is below 298 days.

2. [1 mark] Check your answer from part a) using R code. Create a vector called `p2` that stores 2 values: your answer from part a and the absolute difference between your answer from a and the exact probability that you calculated with code.

```
BEGIN QUESTION
name: p2
manual: false
points: 1
```

```
p2 <- NULL # YOUR CODE HERE
# BEGIN SOLUTION NO PROMPT
p2 <- c(pnorm(q = 298, mean = 266, sd = 16),
        abs(0.975 - pnorm(q = 298, mean = 266, sd = 16)))
# END SOLUTION
p2
```

```
## [1] 0.977249868 0.002249868
```

```
## Test ##
test_that("p2a", {
  expect_true(all.equal(p2[1], pnorm(q = 298, mean = 266, sd = 16), tol = 0.001))
  print("Checking: first value of p2")
})
```

```
## [1] "Checking: first value of p2"
## Test passed
```

```
## Test ##
test_that("p2b", {
  expect_true(all.equal(p2[2], abs(0.975 - pnorm(q = 298, mean = 266, sd = 16)), tol = 0.001))
  print("Checking: second value of p2")
})
```

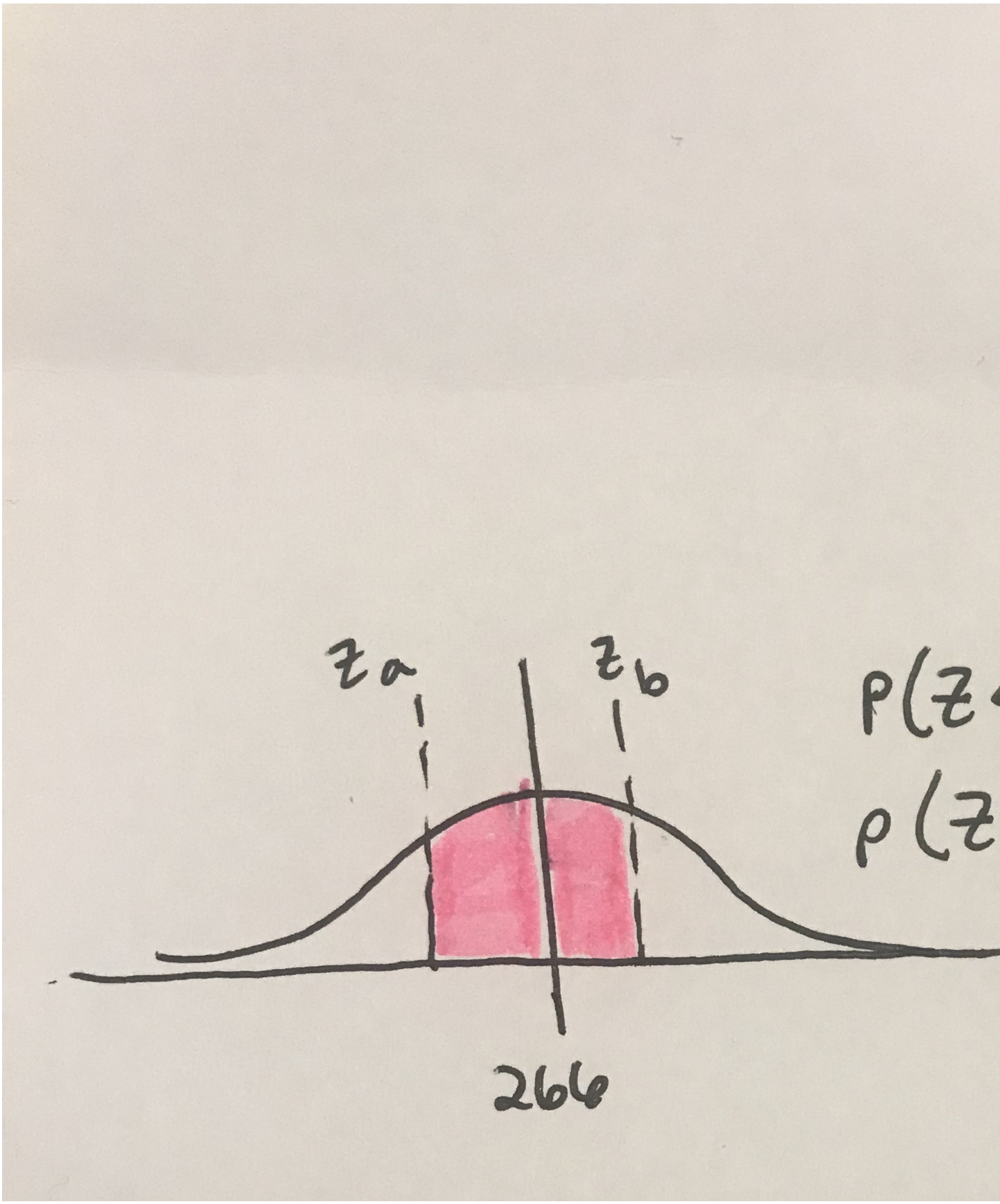
```
## [1] "Checking: second value of p2"
## Test passed
```

3. [3 marks] What is the range, in days, that the middle 50% of pregnancies last? To aid your answer, hand-draw (or use any software) to sketch a Normal curve, and shade in the area that the middle represents. Then, use R to calculate the requested range. Round the lower and upper bound of the range each to two decimal places.

(Use the code chunk below to include an image file of your drawing. To do so you need to delete the hashtag, upload the image to Datahub into the `src` directory and replace the file name with your file name. JPG or PNG will both work.)

```
BEGIN QUESTION
name: p3
manual: true
```

```
#knitr::include_graphics("src/Your-file-name.JPG")
# BEGIN SOLUTION NO PROMPT
knitr::include_graphics("A3_Normal.JPG")
```


z_a
z_b
266

```
# END SOLUTION
```

```
#Your code here
# BEGIN SOLUTION NO PROMPT
# want the quantile (aka percentile) such that 25% of the data is below it
qnorm(p = 0.25, mean = 266, sd = 16)
```

```
## [1] 255.2082
```

```
# the upper bound is the quantile (aka percentile) such that 75% of the data is
# below it
qnorm(p = 0.75, mean = 266, sd = 16)
```

```
## [1] 276.7918
```

```
# END SOLUTION
```

Thus, the range is from 255.21 days to 276.79 days.

[7 points] Part 2: Assessing Normality and Interpreting QQ Plots

The number of trees for nine plots of land, each of 0.1 hectare, have been recorded. They are: 18, 4, 22, 15, 18, 19, 22, 12, 12. Are these data Normally distributed?

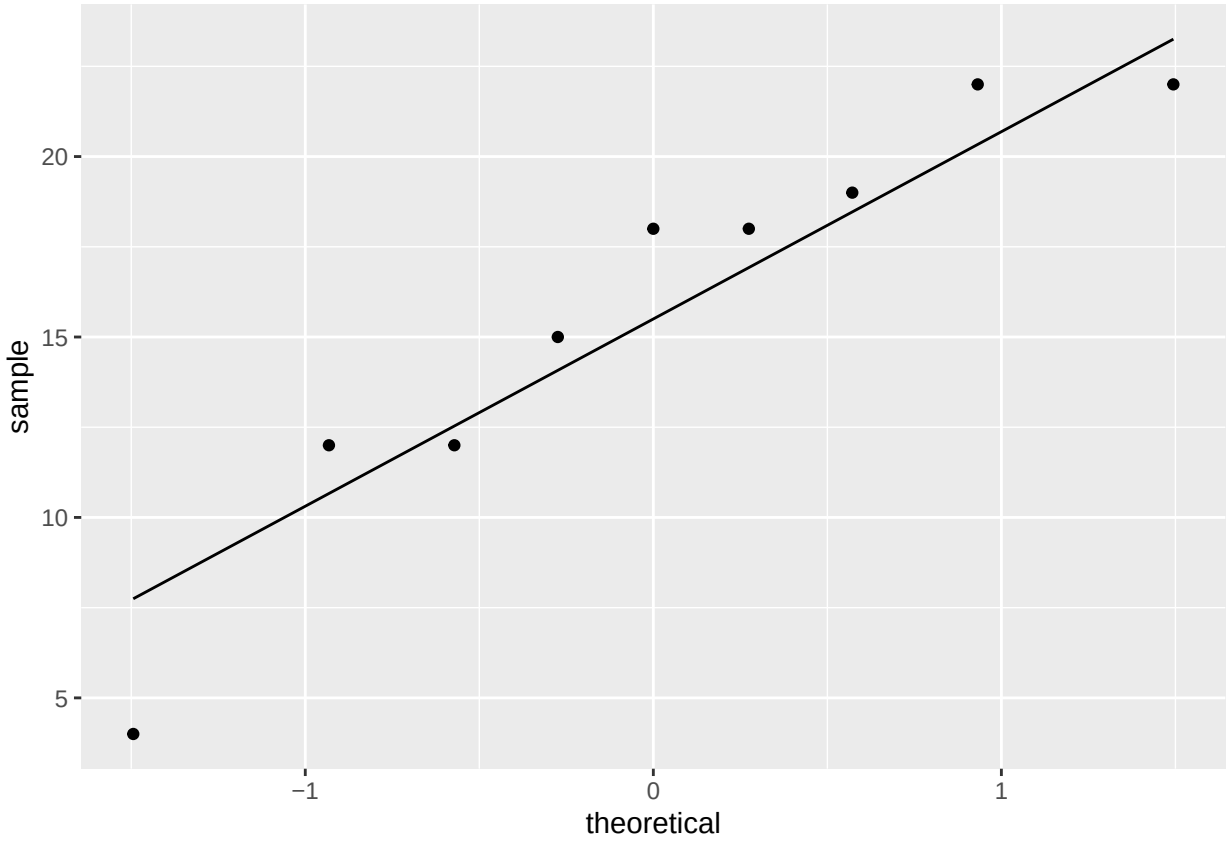
4. [3 marks] Make a Normal quantile plot for these data using R. Remember, to make a ggplot of these data, you need to first input the data as a vector and then convert that vector to a data frame. Example code has been provided to you to get you started. After making the plot, assess whether the data appear to approximately follow a Normal distribution.

```
BEGIN QUESTION
name: p4
manual: true
```

```
library(tidyverse)
# example code

counts <- c(1, 2, 3)
tree_data <- data.frame(counts)
# your code for ggplot here

# BEGIN SOLUTION NO PROMPT
counts <- c(18, 4, 22, 15, 18, 19, 22, 12, 12)
tree_data <- data.frame(counts)
ggplot(tree_data, aes(sample = counts)) + geom_qq() + geom_qq_line()
```




```
# END SOLUTION
```

The QQ Plot of the data quantiles against Normal quantiles is roughly linear, so we believe the data approximately follows a Normal distribution.

[12 points] Part 3: Conducting a general anxiety disorder study

Suppose that a new treatment for COVID-19 has undergone safety and efficacy trials and based on these data 45% of patients with COVID-19 are expected to benefit from the new treatment. You are conducting a follow-up study and so far have enrolled 14 participants with COVID-19 into your study. These patients do not know each other and represent individuals who responded to a mailed flyer.

5. [2 marks] Let X represent the number of patients that you have enrolled who benefit from the treatment. Does X meet the assumptions of a Binomial distribution? Thoroughly explain why or why not.

```
BEGIN QUESTION
name: p5
manual: true
```

Solution: Yes, because: Fixed number of observations (14) All the observations appear to be independent (they don't know each other) Each is either a success (benefit) or failure (no benefit) The probability of success is same for each person

6. [1 mark] Using one of the distributions learned in class that X meets the assumptions of, calculate by hand the probability that exactly 7 participants will benefit. Show your work.

```
BEGIN QUESTION
name: p6
manual: true
```

$\binom{n}{k} p^k (1-p)^{n-k}$

$\binom{14}{7} 0.45^7 (1-0.45)^{14-7} = 0.1952422$

7. [1 mark] Confirm your previous calculation using an R function. Store your answer to p7.

```
BEGIN QUESTION
name: p7
manual: false
points: 1
```

```
p7<- dbinom(x = 7, size = 14, prob = 0.45) # SOLUTION
p7
```

```
## [1] 0.1952422
```

```
## Test ##
test_that("p7a", {
  expect_true(p7 > 0 & p7 < 1)
  print("Checking: range of p7")
})
```

```
## [1] "Checking: range of p7"
## Test passed
```

```
## Test ##
test_that("p7b", {
  expect_true(all.equal(p7, dbinom(x = 7, size = 14, prob = 0.45), tol = 0.1))
  print("Checking: value of p7")
})
```

```
## [1] "Checking: value of p7"
## Test passed
```

8. [2 marks] Calculate by hand the probability that 12 or more participants will benefit. Show your work.

```
BEGIN QUESTION
name: p8
manual: true
```

$\binom{14}{12}0.45^{1}2(1-0.45)^{14-12} + \binom{14}{13}0.45^{1}3(1-0.45)^{14-13} + \binom{14}{14}0.45^{1}4(1-0.45)^{14-14}$

$= 0.002150974$

9. [1 mark] Confirm your previous calculation using code that depends on `pbinom()`. Store your answer to p9.

```
BEGIN QUESTION
name: p9
manual: false
points: 1
```

```
p9 <- 1 - pbinom(q = 11, size = 14, prob = 0.45) # SOLUTION
p9
```

```
## [1] 0.002150974
```

```
## Test ##
test_that("p9a", {
  expect_true(p9 > 0 & p9 < 1)
  print("Checking: range of p9")
})
```

```
## [1] "Checking: range of p9"
## Test passed
```

```
## Test ##
test_that("p9b", {
  expect_true(all.equal(p9, 1 - pbinom(q = 11, size = 14, prob = 0.45), tol = 0.1))
  print("Checking: value of p9")
})
```

```
## [1] "Checking: value of p9"
## Test passed
```

10. [1 mark] Re-confirm your previous calculation, this time using code that depends on `dbinom()`. Store your answer to p10.

```
BEGIN QUESTION
name: p10
manual: false
points: 1
```

```
p10 <- NULL # YOUR CODE HERE
# BEGIN SOLUTION NO PROMPT
p10 <- dbinom(x = 12, size = 14, prob = 0.45) +
  dbinom(x = 13, size = 14, prob = 0.45) +
  dbinom(x = 14, size = 14, prob = 0.45)
# END SOLUTION
p10
```

```
## [1] 0.002150974
```

```
## Test ##
test_that("p10a", {
  expect_true(p10 > 0 & p10 < 1)
  print("Checking: range of p10")
})
```

```
## [1] "Checking: range of p10"
## Test passed
```

```
## Test ##
test_that("p10b", {
  expect_true(all.equal(p10, dbinom(x = 12, size = 14, prob = 0.45) +
  dbinom(x = 13, size = 14, prob = 0.45) +
  dbinom(x = 14, size = 14, prob = 0.45), tol = 0.1))
  print("Checking: value of p10")
})
```

```
## [1] "Checking: value of p10"
## Test passed
```

11. [4 marks] Calculate the number of patients you would expect to benefit from the treatment and the standard deviation. Write a sentence to interpret the meaning of the mean. If the mean is not a whole number, what whole number is most probable?

```
BEGIN QUESTION
name: p11
manual: true
```

$\mu = np = 14 * 0.45 = 6.3$ [1 mark]

$\sigma = \sqrt{np \times (1-p)} = 1.86$ [1 mark]

[1 mark] We expect 6.3 patients to benefit out of the 14. [1 mark] An average of 6.3 implies that seeing six patients benefit is the most probable number (because 6.3 is closer to 6 than it is to 7).

12. [1 mark] Should you apply a Normal approximation to these data using the μ and σ you calculated in the last question? Why or why not?

```
BEGIN QUESTION
name: p12
manual: true
```

No, because $np = 6.3$ is much smaller than 10, which the rule of thumb threshold we used to decide whether we shouldn't apply the Normal approximation.