Problem Set 5: Normal and Binomial Distribution

Your name and student ID

September 30, 2024

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:testthat':
##
##     matches

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Instructions

- Solutions will be released on Sunday, October 6th.
- This semester, problem sets 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!

- 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.

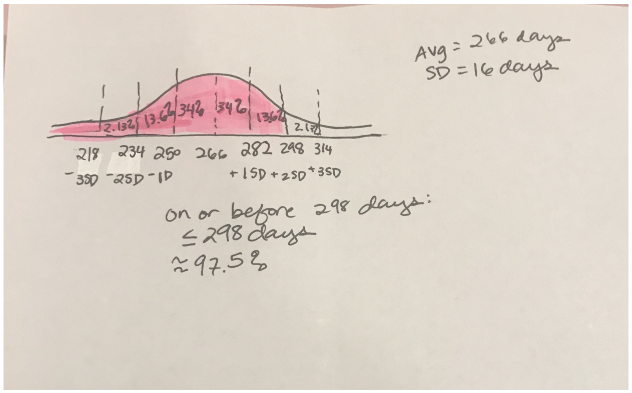
Part 1: Pregnancy Length Probabilities

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

1. [1 point] 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 add dashed lines at the mean +/- 1SD, 2SD and 3SD. Calculate the proportion of births occurring on or before 298 days by shading this region under the curve. You shouldn't need to use R to perform any calculations for this question. Round the proportion to one decimal place.

(Use the code chunk below to include an image file of your drawing. To do this 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).

```r
knitr::include_graphics("src/A5_Normal-a.JPG")
```



Students should draw the Normal density curve with the pregnancy 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 point] **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.**

```
p2 <- c(pnorm(q = 298, mean = 266, sd = 16),
       abs(0.975 - pnorm(q = 298, mean = 266, sd = 16)))
p2
```

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

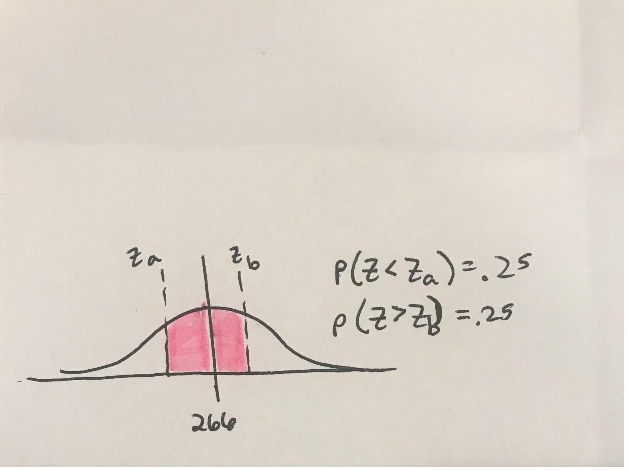
```
. = ottr::check("tests/p2.R")
```

```
##
## All tests passed!
```

3. [1 point] 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 range represents. Then use R to calculate this middle 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.)

```
knitr::include_graphics("src/A3_Normal.JPG")
```



```
# 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
```

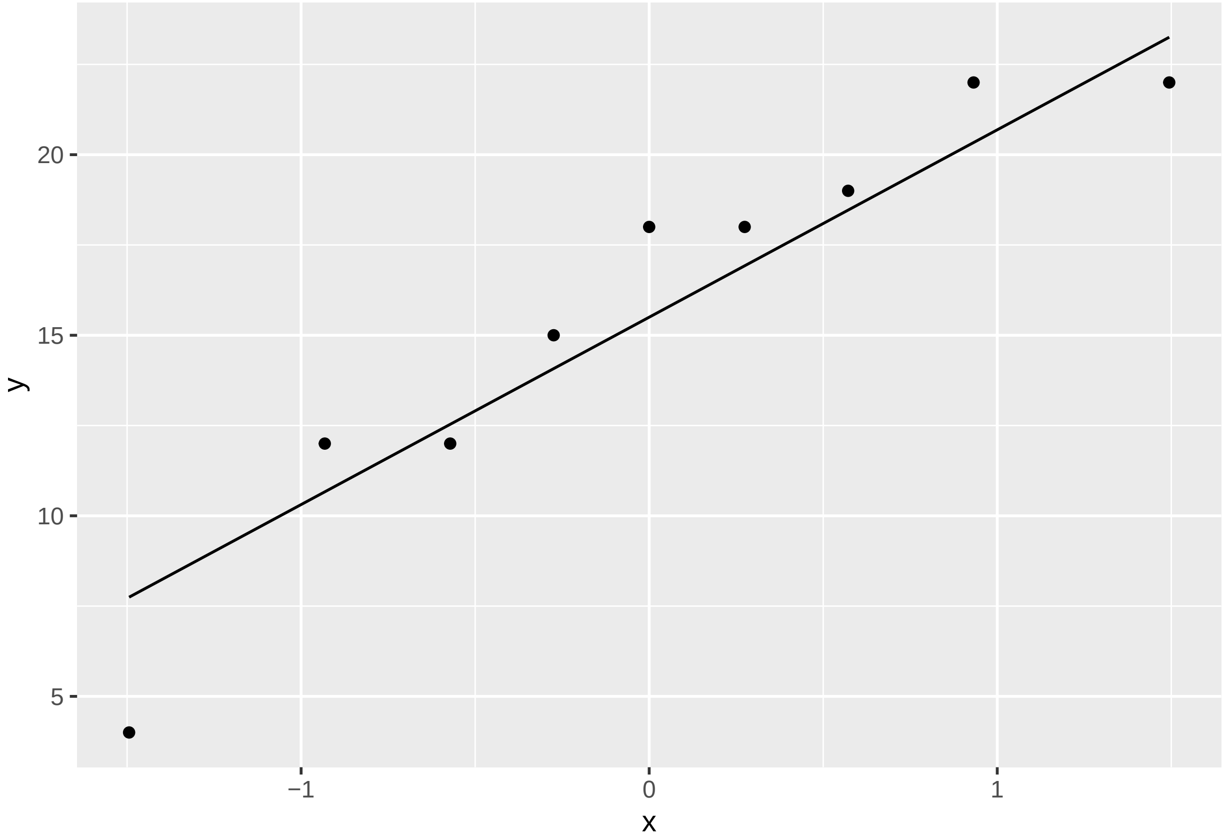
Thus, the range is from 255.21 days to 276.79 days.

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. [1 point] 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 dataframe. Example code has been provided to help get you started. After making the plot, assess whether the data appear to approximately follow a Normal distribution.

```
library(tidyverse)
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()
```



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

Part 3: Conducting a study about general anxiety disorder

Suppose that a new treatment for general anxiety disorder has undergone safety and efficacy trials and, based on these data, 30% of patients with general anxiety disorder are expected to benefit from the new treatment. You are conducting a follow-up study and have enrolled 8 participants with general anxiety disorder so far. These patients do not know each other and represent individuals who responded to a call for study participants that they saw on a flier on campus.

5. [1 point] Let X represent the number of enrolled patients who benefit from the treatment. Does X meet the assumptions of a Binomial distribution? Thoroughly explain why or why not.

Solution: Yes, because: Fixed number of obsevrations (8) 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 the same for each person

6. [1 point] Using one of the distributions whose assumptions X meets, calculate (by hand) the probability that exactly 5 participants will benefit from the treatment. Show your work.

```
# ${n \choose k}p^k(1-p)^{n-k}$
# ${8 \choose 5}0.3^5(1-0.3)^{8-5}$ = 0.04667544
p6 <- 0.04667544
p6
```

```
## [1] 0.04667544
```

```
. = ottr::check("tests/p6.R")
```

```
##
## All tests passed!
```

7. [1 point] Confirm your previous calculation using an R function and store your answer to p7.

```r
p7 <- dbinom(x = 5, size = 8, prob = 0.3)
p7
```

```
## [1] 0.04667544
```

```r
. = ottr::check("tests/p7.R")
```

```
##
## All tests passed!
```

8. [1 point] Calculate (by hand) the probability that 6 or more participants will benefit from the treatment. Show your work.

```
# ${8 \choose 6}0.3^6(1-0.3)^{8-6} +
#  {8 \choose 7}0.3^7(1-0.3)^{8-7} +
#  {8 \choose 8}0.3^8(1-0.3)^{8-8}$

p8 <- 0.01129221
```

```
. = ottr::check("tests/p8.R")
```

```
##
## All tests passed!
```

9. [1 point] Confirm your previous calculation using the function `pbinom()` and store your answer to `p9`.

```r
p9 <- 1 - pbinom(q = 5, size = 8, prob = 0.3)
p9
```

```
## [1] 0.01129221
```

```r
. = ottr::check("tests/p9.R")
```

```
##
## All tests passed!
```

10. [1 point] Re-confirm your previous calculation, this time using the function `dbinom()`, and store your answer to `p10`.

```
p10 <- dbinom(x = 6, size = 8, prob = 0.3) +
  dbinom(x = 7, size = 8, prob = 0.3) +
  dbinom(x = 8, size = 8, prob = 0.3)
p10
```

```
## [1] 0.01129221
```

```
. = ottr::check("tests/p10.R")
```

```
##
## All tests passed!
```

11. [1 point] Interpret the binomial coefficient, $\binom{8}{7}$, in the context of this study. Write out all the possible combinations to achieve $\binom{8}{7}$.

$\binom{8}{7}$ is the number of ways to have 7 individuals benefitting out of 8 study participants. There are eight possible ways to see 7 successes across 8 individuals:

11111110 11111101 11111011 11110111 11101111 11011111 10111111 01111111

12. **[1 point] Calculate the number of patients you would expect to benefit from the treatment. Then calculate the standard deviation of this estimate. Write a sentence to interpret the mean. If the mean is not a whole number, what whole number is most probable?**

$\mu = np = 8 * 0.3 = 2.4$

$\sigma = \sqrt{np \times (1 - p)} = 1.3$

We expect 2.4 patients to benefit out of the 8. An average of 2.4 implies that seeing two patients benefit is the most probable number (because 2.4 is closer to 2 than it is to 3).

13. [1 point] Should you apply a Normal approximation to these data using the μ and σ you calculated in the last question? Why or why not?

No, because $np = 2.4$ is much smaller than 10, which is the rule of thumb threshold we used to decide whether we should apply the Normal approximation.

Late Pre-Term Birth Weights (From Baldi and Moore, 3E question 11.32, 4E question 11.34)

How much of a difference do a couple of weeks make for a baby's birthweight? Late preterm babies are born with 35 to 37 weeks of completed gestation. The distribution of birth weight (in grams) or late preterm babies is approximately normally distributed with a mean of 2750 grams and a standard deviation of 560 grams, N(2750,560).

14. [1 point] What is the 25th percentile of the birthweights for late-preterm term babies?

```
p14 <- qnorm(0.25, mean = 2750, sd = 560)
p14
```

```
## [1] 2372.286
```

```
. = ottr::check("tests/p11.R")
```

```
##
## All tests passed!
```

15. [1 point] What is the 90th percentile of the birthweights for late-preterm babies?

```
p15 <- qnorm(0.9, mean = 2750, sd = 560)
p15
```

```
## [1] 3467.669
```

```
. = ottr::check("tests/p15.R")
```

```
##
## All tests passed!
```

16. [1 point] What is the range of the middle 50% of birthweights for late-preterm babies?

```
p16 <- c(qnorm(0.25, mean = 2750, sd = 560), qnorm(0.75, mean = 2750, sd = 560))
p16
```

```
## [1] 2372.286 3127.714
```

```
. = ottr::check("tests/p16.R")
```

```
##
## All tests passed!
```

17. Think back to lab05 when we studied the distribution of full-term birthweights N(3350,440). Compare the percentiles you calculated above between full term babies and late-preterm babies. What do you notice?

Note that the larger standard deviation for the late-preterm babies, 560 g vs 440 g, makes the ranges bigger and the percentiles farther from the mean. Since the 25th percentile for the late-preterm births is 2372.3 g, well below 2500 g, we know that more than 25% of the late-preterm babies are low birth weight.

Drosophila (From Baldi and Moore, 3E questions 11.20 and 11.22, 4E question 11.23)

The common fruit fly, Drosophila melanogaster, is the most studied organism in genetic research because it is small, easy to grow, and reproduces rapidly. The length of the thorax (where the wings and legs attach) in a population of male fruit flies is approximately Normal with mean 0.800 millimeters (mm) and standard deviation 0.078 mm.

18. [1 point] Choose a male fruit fly at random. Calculate the probability that the fly you choose has a thorax longer than 1 mm (convert to a percentage and round to two decimal places).

```
p18 <- round(pnorm(1, mean = 0.8, sd = 0.078, lower.tail=FALSE)*100,2)
p18
```

```
## [1] 0.52
```

```
. = ottr::check("tests/p18.R")
```

```
##
## All tests passed!
```