Normal, Binomial, and Poisson Distribution

Nimita Gaggar

```
library(testthat)
library(dplyr)
```

```
##
## 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
```

```
library(ggplot2)
library(rlang)
```

```
##
## Attaching package: 'rlang'

## The following objects are masked from 'package:testthat':
##
##     is_false, is_null, is_true
```

More Practice with the Normal Distribution

Eating disorders affect at least 9% of the population worldwide. One such eating disorder is anorexia which affects approximately 1 in 200 American women. One study was interested in the effects of different therapies as forms of treatment for eating disorders. 72 young women were recruited and assigned to 1 of 3 different groups: control, cognitive behavioral treatment (CBT), and family therapy. Their weights (in pounds) were recorded pre-treatment and post-treatment.

```
# The data comes from an R package called MASS. Let's first load the package.
if (!require("MASS")){
  install.packages("MASS")
  libary("MASS")
}
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
# Save the dataset in an object called `anorexia`
anorexia <- MASS::anorexia

# MASS has functions with the same names as common dplyr functions
# We will detach it now so that we can continue to use our dplyr functions
detach("package:MASS", unload = T)

head(anorexia, 10) # Here are the first 10 rows of data
```

```
##      Treat Prewt Postwt
## 1    Cont  80.7   80.2
## 2    Cont  89.4   80.1
## 3    Cont  91.8   86.4
## 4    Cont  74.0   86.3
## 5    Cont  78.1   76.1
## 6    Cont  88.3   78.1
## 7    Cont  87.3   75.1
## 8    Cont  75.1   86.7
## 9    Cont  80.6   73.5
## 10   Cont  78.4   84.6
```

1. [1 point] Add a new column to anorexia called diff that is the difference between the women's weights before and after treatment. Assign this new dataset to an object called anorexia_diff.

```
anorexia_diff <- anorexia %>% mutate (diff = Postwt- Prewt)
anorexia_diff
```

```
##       Treat Prewt Postwt   diff
##  1    Cont  80.7   80.2   -0.5
##  2    Cont  89.4   80.1   -9.3
##  3    Cont  91.8   86.4   -5.4
##  4    Cont  74.0   86.3   12.3
##  5    Cont  78.1   76.1   -2.0
##  6    Cont  88.3   78.1  -10.2
##  7    Cont  87.3   75.1  -12.2
##  8    Cont  75.1   86.7   11.6
##  9    Cont  80.6   73.5   -7.1
##  10   Cont  78.4   84.6    6.2
##  11   Cont  77.6   77.4   -0.2
##  12   Cont  88.7   79.5   -9.2
##  13   Cont  81.3   89.6    8.3
##  14   Cont  78.1   81.4    3.3
##  15   Cont  70.5   81.8   11.3
##  16   Cont  77.3   77.3    0.0
##  17   Cont  85.2   84.2   -1.0
##  18   Cont  86.0   75.4  -10.6
##  19   Cont  84.1   79.5   -4.6
##  20   Cont  79.7   73.0   -6.7
##  21   Cont  85.5   88.3    2.8
##  22   Cont  84.4   84.7    0.3
##  23   Cont  79.6   81.4    1.8
##  24   Cont  77.5   81.2    3.7
##  25   Cont  72.3   88.2   15.9
##  26   Cont  89.0   78.8  -10.2
##  27    CBT  80.5   82.2    1.7
##  28    CBT  84.9   85.6    0.7
##  29    CBT  81.5   81.4   -0.1
##  30    CBT  82.6   81.9   -0.7
##  31    CBT  79.9   76.4   -3.5
##  32    CBT  88.7  103.6   14.9
##  33    CBT  94.9   98.4    3.5
##  34    CBT  76.3   93.4   17.1
##  35    CBT  81.0   73.4   -7.6
##  36    CBT  80.5   82.1    1.6
##  37    CBT  85.0   96.7   11.7
##  38    CBT  89.2   95.3    6.1
##  39    CBT  81.3   82.4    1.1
##  40    CBT  76.5   72.5   -4.0
##  41    CBT  70.0   90.9   20.9
##  42    CBT  80.4   71.3   -9.1
##  43    CBT  83.3   85.4    2.1
##  44    CBT  83.0   81.6   -1.4
##  45    CBT  87.7   89.1    1.4
##  46    CBT  84.2   83.9   -0.3
##  47    CBT  86.4   82.7   -3.7
##  48    CBT  76.5   75.7   -0.8
##  49    CBT  80.2   82.6    2.4
##  50    CBT  87.8  100.4   12.6
##  51    CBT  83.3   85.2    1.9
##  52    CBT  79.7   83.6    3.9
##  53    CBT  84.5   84.6    0.1
```

```
## 54   CBT  80.8   96.2  15.4
## 55   CBT  87.4   86.7  -0.7
## 56    FT  83.8   95.2  11.4
## 57    FT  83.3   94.3  11.0
## 58    FT  86.0   91.5   5.5
## 59    FT  82.5   91.9   9.4
## 60    FT  86.7  100.3  13.6
## 61    FT  79.6   76.7  -2.9
## 62    FT  76.9   76.8  -0.1
## 63    FT  94.2  101.6   7.4
## 64    FT  73.4   94.9  21.5
## 65    FT  80.5   75.2  -5.3
## 66    FT  81.6   77.8  -3.8
## 67    FT  82.1   95.5  13.4
## 68    FT  77.6   90.7  13.1
## 69    FT  83.5   92.5   9.0
## 70    FT  89.9   93.8   3.9
## 71    FT  86.0   91.7   5.7
## 72    FT  87.3   98.0  10.7
```

```
dim(anorexia_diff) # Uncomment this line to check the dimensions of your new dataset
```

```
## [1] 72  4
```

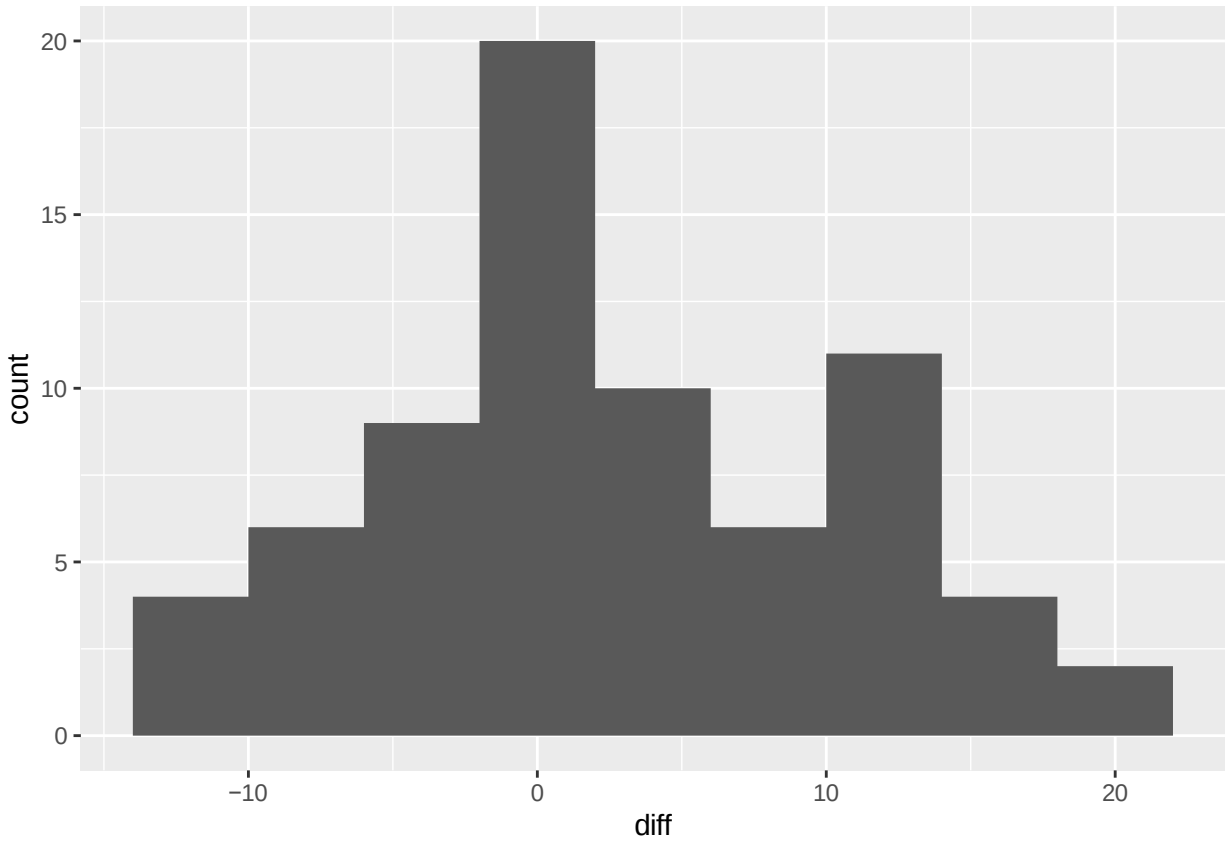
```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p5.R")
```

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

2. [1 point] Visualize the distribution of the variable diff. Choose an appropriate binwidth.

```
p6 <- ggplot(anorexia_diff, aes (x =diff)) + geom_histogram(binwidth = 4)
p6
```




```
# YOUR CODE HERE
```

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

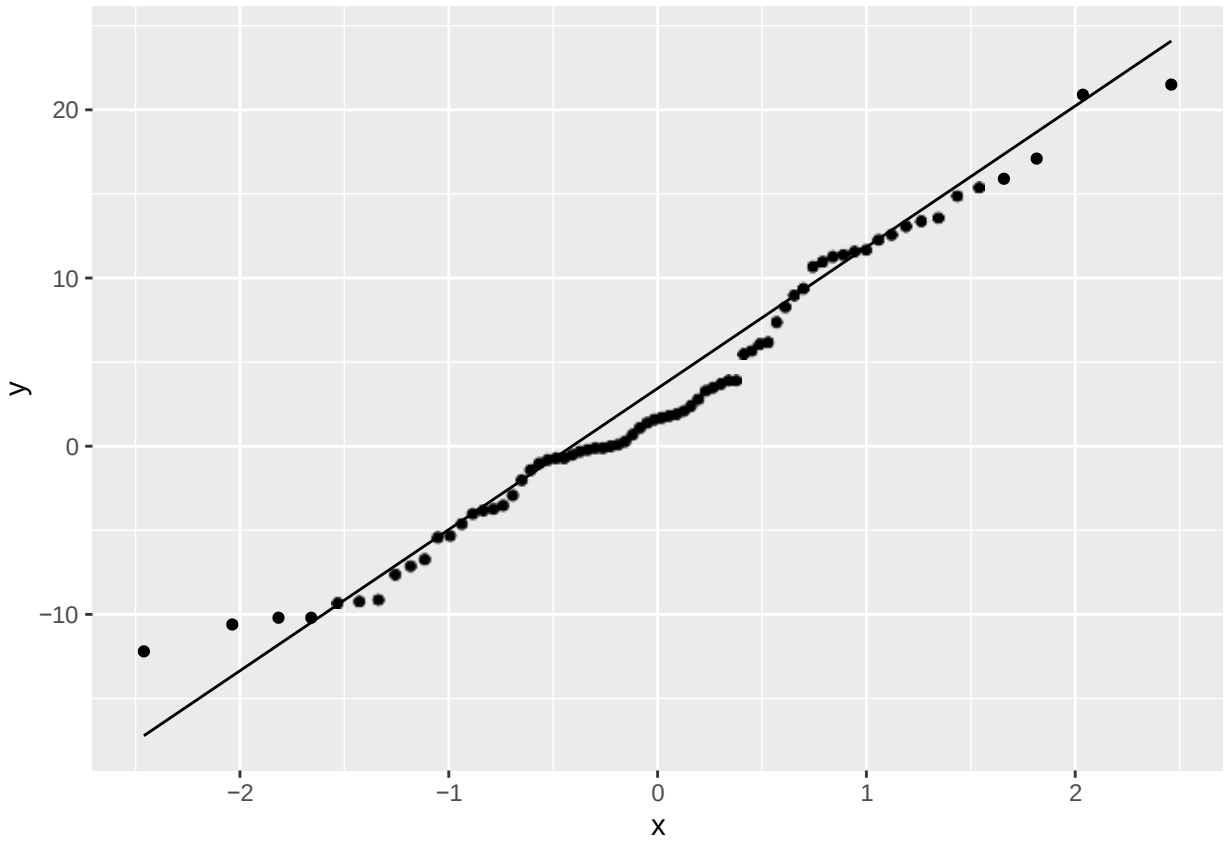
```
##
## All tests passed!
```

3. Describe the distribution of the diff variable.

We observe a bimodol distribution here with a peak around 0 and other peak around 11.

4. [1 point] Compare the diff variable from the anorexia_diff data to a Normal distribution using a different ggplot2 function. Determine if a Normal distribution is a good fit for these data.

```
p8 <-ggplot(anorexia_diff, aes(sample = diff)) + stat_qq() +      stat_qq_line()
p8
```




```
# YOUR CODE HERE
```

The QQ plot suggest that out data is closely associated with the qq line but we should beware that there is slight deviation around the tail and the center (around =0)

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

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

5. [1 point] Assume data on the difference in weights pre- and post-treatment is sampled from a population distribution that is approximately Normal with a mean of 2 pounds and standard deviation of 7 pounds. Find the probability that a randomly chosen woman suffering from anorexia gains 5 pounds or more over the course of the treatment. You may leave this as an unrounded number between 0 and 1.

```
p9 <- 1 - pnorm (5,2,7)
p9
```

```
## [1] 0.3341176
```

```
# YOUR CODE HERE
```

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

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

6. [1 point] Using the information above, find the number of pounds a randomly chosen woman would need to gain in order to be in the 90th percentile according to this probability distribution.

```
p10 <- qnorm(0.9, 2, 7)
p10
```

```
## [1] 10.97086
```

```
# YOUR CODE HERE
```

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

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

Section 3: Binomial Distribution and Normal Approximation

Example from Baldi and Moore

Antibiotic resistance occurs when disease-causing microbes no longer respond to antibiotic drug therapy. Because such resistance is typically genetic and transferred to the next generations of microbes, it is a very serious public health problem. One such disease with antibiotic resistance is Gonorrhea, the second most commonly reported notifiable disease in the USA. According to the CDC, 27% of Gonorrhea cases tested in 2010 were resistant to at least one of the three major antibiotics commonly used to treat sexually transmitted diseases. For the following examples, consider a physician who treated 20 cases of Gonorrhea in 2010.

7. Let X represent the number of patients with antibiotic resistance seen by this physician. Use notation you learned in lecture to show the distribution that X follows.

X ~ binomial (n = 20, p = 0.27)

8. [1 point] Calculate (by hand) the probability that exactly 5 of the 20 treated people have antibiotic resistance. You can use the choose(n = , k =) function in R to help. Confirm your results using an R function.

```
p12 <-choose ( n = 20, k = 5)* (0.27^5) * ((1-0.27)^(20-5))
p12
```

```
## [1] 0.1982008
```

```
# YOUR CODE HERE
```

(choose) (success)(failure)

```
. = ottr::check("tests/p12.R")
```

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

9. [1 point] Calculate (by hand) the probability that more than 1 person has antibiotic resistance. Confirm your answer using R. Hint: work smarter not harder.

```
p13 <-1- pbinom(1, 20, 0.27)
p13
```

```
## [1] 0.9844906
```

```
# YOUR CODE HERE
```

P(x>=2) = P(X = 0) and P(X = 1)

```
. = ottr::check("tests/p13.R")
```

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

Suppose in one Western United States city there were 812 cases of gonorrhea in a population of 100,000. The probability of antibiotic resistance to at least one major antibiotic remains the same at approximately 27 percent.

10. [1 point] Calculate the expected number of antibiotic resistant cases of gonorrhea in this population. Make sure to round to the nearest whole number. Also calculate the standard deviation and round this number to two decimal places. Assign p14 to a vector of these two values.

```
p14 <-c(round(812*0.27), round(sqrt(812*0.27*0.73), 2))
p14
```

```
## [1] 219.00  12.65
```

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p14.R")
```

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

11. We learned in class that the binomial distribution can be approximated by the Normal distribution under some conditions. List the conditions below and determine whether problem 14 satisfies them.

n * p >= 10 and n* (1-P) >= 10

Let's generate some data from this distribution to check the normal approximation!

The first step is to generate the probabilities of observing each of the possible values of X ~ Binom(n = 812, p = 0.27). We will use the familiar dbinom() function to do this. However, instead of just plugging in one value, we will plug in the entire range of values (0 through 812) and save it as a vector called obs_data.

Note: You will not be tested on this use of code but you should understand what's happening at every step. It is useful to print out the object in your console to get an idea of what's happening at each stage.

```
# This is just the range of values x can take
x_vals <- 0:812

# This generates the probabilities
probs <- dbinom(0:812, size = 812, prob = 0.27)

# This combines the range of values with the probabilities as a dataframe with 2 columns:  x_vals and pr
# View(obs_data) in your console to see what this dataframe looks like
obs_data <- as.data.frame(cbind(x_vals, probs))

names(obs_data)
```

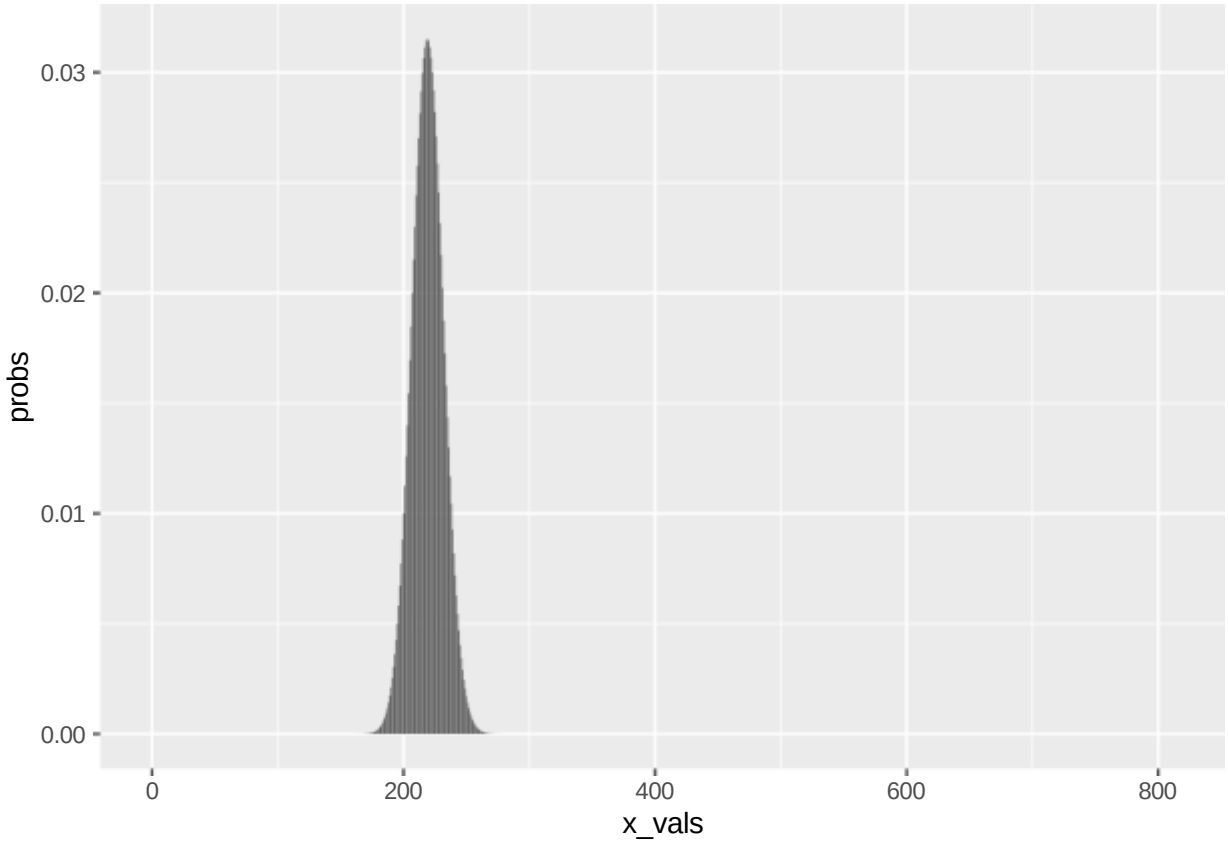
```
## [1] "x_vals" "probs"
```

12. [1 point] Now use ggplot2 to plot a histogram of obs_data with x_vals on the x-axis and the respective probabilities on the y-axis.

```
p16 <- ggplot(obs_data, aes(x = x_vals, y = probs)) + geom_histogram(stat = "identity")
```

```
## Warning in geom_histogram(stat = "identity"): Ignoring unknown parameters:
## 'binwidth', 'bins', and 'pad'
```

```
p16
```




```
# YOUR CODE HERE
```

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

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

Section 4: Poisson Distribution

Example from Baldi and Moore

Between 2006 and 2010, the state of New York reported 1484 live births of infants with Down Syndrome, which averages to about 5.7 cases per week. While the causes of Down Syndrome are not fully understood, it is reasonable to assume that live births are independent and the weekly rate is constant. Let X represent the count of babies born with Down Syndrome in New York in a given week.

13. What distribution does X approximately follow? Write it using notation learned in lecture. What are the possible values X can take?

X~Poisson (Lambda = 5.7), X can take any integer greater than or equal to 0.

14. [1 point] What are the mean and standard deviation of X? Assign these values to a vector called p18.

```
p18 <- c(5.7, sqrt(5.7))
p18
```

```
## [1] 5.700000 2.387467
```

```
# YOUR CODE HERE
```

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

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

15. [1 point] What is the probability that no child will be born with Down Syndrome in a given week in New York? Calculate the probability by hand and confirm your answer using a function in R.

```
p19 <-dpois(0, 5.7)
p19
```

```
## [1] 0.003345965
```

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p19.R")
```

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

16. [1 point] What is the probability that 2 or more children will be born with Down Syndrome in a given week in New York? Calculate the probability by hand and confirm your answer in R.

```
p20 <- 1- (dpois(0, 5.7) + dpois(1,5.7))
p20
```

```
## [1] 0.977582
```

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p20.R")
```

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

17. [1 point] Use R to calculate the probability that more than 12 children are born with Down Syndrome.

```
p21 <- 1 - ppois(12, 5.7)
p21
```

```
## [1] 0.005921731
```

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p21.R")
```

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