Lab 2: Probability Distributions

Graded Out of 35 Points

Jennifer Bradley

2025-01-27

We will again be using our Palmer Penguins data today for some problems, so let's load that package in now:

```r
library(palmerpenguins)
```

```
Warning: package 'palmerpenguins' was built under R version 4.4.2
```

1. Working With Probability Distributions

Problem 1.1 (5 Points)

Let X be a discrete random variable whose PDF is described in the table given here:

Table 1: PDF of a discrete random variable, X

X	1	2	3	4
$f(X)$	3/8	1/8	1/8	3/8

Do the following (show work, 1 point each):

A. Calculate $E[X]$

   $\sum_{Sx} xp(x)$

   $(1 \cdot \frac{3}{8}) + (2 \cdot \frac{1}{8}) + (3 \cdot \frac{1}{8}) + (4 \cdot \frac{3}{8})$

$$\boxed{2.5}$$

B. Calculate $Var(X)$

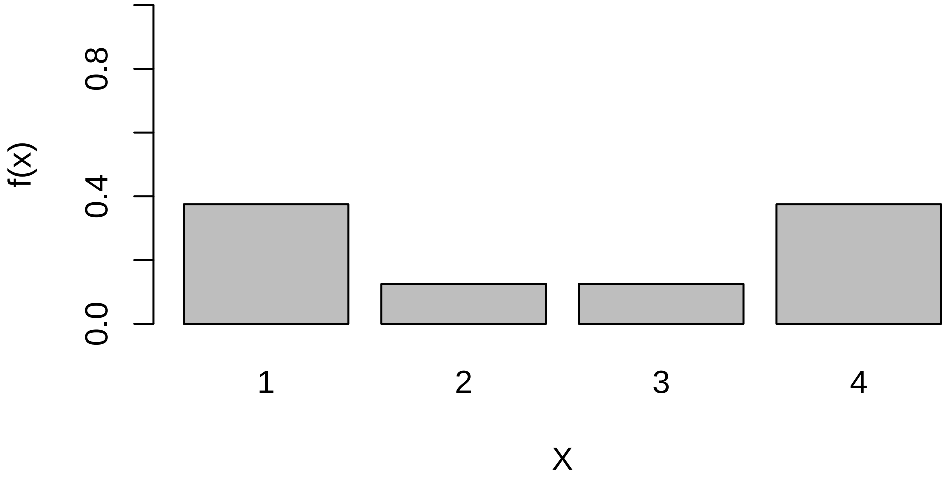
$\sum_{Sx} x^2 p(x) - (\sum_{Sx} xp(x))^2$

$((1^2 \cdot \frac{3}{8}) + (22 \cdot \frac{1}{8}) + (32 \cdot \frac{1}{8}) + (42 \cdot \frac{3}{8})) - 2.52$

$$\boxed{1.75}$$

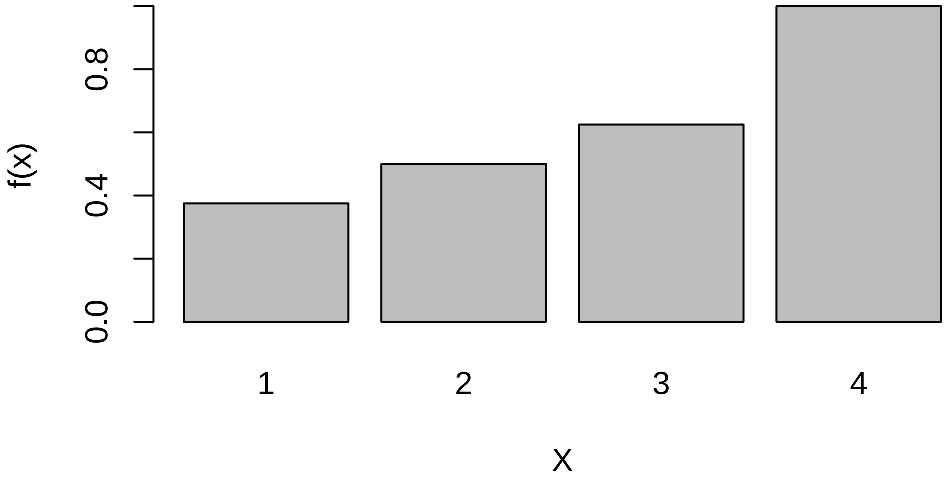
C. Plot $f(x)$. To make a barplot, use the **`barplot`** function.

```r
fx <- c(3/8,1/8,1/8,3/8)
X <- c(1,2,3,4)
# realized I do not need this: prob1C.table <- data.frame(fx,X)

barplot(fx, xlab="X", ylab="f(x)", names.arg = c("1","2","3","4"), ylim = c(0,1))
```



D. Plot $F(x)$. To make a barplot, use the **`barplot`** function.

```r
Fx <- c(3/8,((3/8)+(1/8)),((3/8)+(1/8)+(1/8)),((3/8)+(1/8)+(1/8)+(3/8)))
# realized I do not need this: prob1D.table <- data.frame(Fx,X)

barplot(Fx, xlab = "X", ylab = "f(x)", names.arg = c("1","2","3","4"))
```



E. Calculate $P(X \geq 4)$

4 is the only possible outcome ≤ 4 in the set

$P(4) = \frac{3}{8}$

$$\boxed{\frac{3}{8}}$$

Problem 1.2 (5 Points)

Some background: The exponential distribution is the continuous analog of the geometric distribution, which is discrete. Consider a geometric random variable X with parameter prob=0.5 and an exponential random variable Y with parameter rate=0.5. Note that R has a standard naming scheme for functions in which PDFs are named "d" + (distribution suffix) and CDFs are named "p" + (distribution suffix). The distribution suffix for the geometric distribution is `geom` (*e.g.*, `dbiom`), and for the exponential is `exp` (*e.g.* `dexp`). There is a trick in this question; think carefully!

A) (2 Points) Plot the PDF for X and Y across a range from 0 to 10. One plot should be a barplot and the other a scatter plot with lines; part of the question is figuring out which variable matches which style of plot. Some tips:

- To generate a vector of evenly spaced values over a set interval, try the `seq` function.

- If if `v` is a vector with the values you want to evaluate between 0 and 10, then you can generate a vector with all the corresponding probabilities via `dgeom(x=v, prob=0.5)` and analogously for the exponential.
- To make a line plot, use the `plot` function with the option `type="l"`

B) (2 Points) Calculate the probability that $X = 5$ and the probability that $Y = 5$

C) (1 Points) Calculate the probability that $5 < X \leq 6$ and the probability that that $5 < Y \leq 6$. Hint: Use the CDF for this.

2. Knowing Your Distributions

Problem 2.1 (5 Points)

For each of the following situations, propose a probability distribution that would suit the situation and justify your selection (1 point each).

A. We are interested in simulating the number of aphids per leaf on a tomato plant. For each leaf, there will be a discrete number of aphids, with no strict upper limit.

B. We are interested in simulating the time it takes a vulture to locate it's next meal. Assume most vultures find a meal rather quickly, but some will take a very (arbitrarily) long time.

C. We have collected 100 moths from the environment that may belong to one of two different color morphs (light and dark). We are interested in simulating the number of light moths collected in this sample.

D. We are in the field, again sampling a species of moth that may belong to one of two color morphs (light or dark). We are interested in simulating whether or not an individual moth sampled will be from the dark color morph.

E. We are interested in simulating the average number of leaves per tree in a forest across many different forests in the Northeast US.

Problem 2.2 (10 Points)

For each of the following, if necessary identify the probability distribution you are using and justify your choice, then answer the question showing your work.

A. (2 Points) What is the probability that a predator will capture it's first prey item after six failed attempts given that it's probability of capturing prey per attempt is 0.3?

B. (2 Points) If new members of a species immigrate into a defined patch of habitat at a constant rate of 0.5 per day, what is the expected wait time until the next new arrival?

C. (2 Points) For the example above (B), what is the expected number of arrivals over a one-week period?

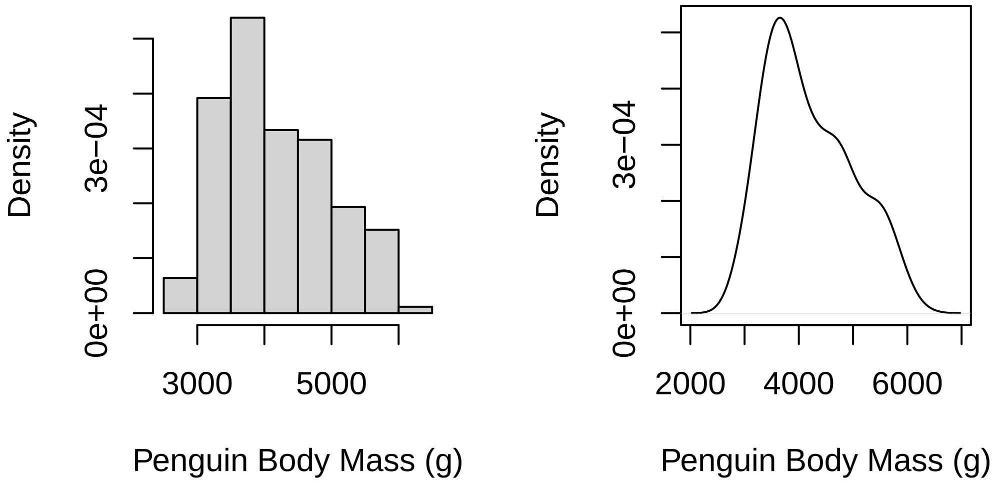
D. (3 Points) We randomly selected 100 White Ash trees from a forest in upstate NY. For the area we are interested in, we know that from previous surveys about 30% of trees are infested by Emerald Ash Borers (assume a 0.3 probability that any given tree is infested). Use R to plot the pdf of the number of infested trees in your sample. Discuss (briefly!) the limitations of our approach for modeling this probability - how might the assumptions of our model be violated?

E. (1 Point) We are studying a forest where tree heights are normally distributed with a mean height of 10 meters and a variance of 4 meters. What percent of trees do we expect to fall within a height range of 6-14 meters?

3. Checking for Normality with Quantile-Quantile Plots

Let's take a look at our penguin body mass data. We are wondering whether it would be appropriate to assume that penguin body mass is approximately normally distributed. First, let's take a look at the distribution of penguin body mass:

```r
#the par function is a handy tool for making side-by-side plots in base R
#BUT for publication-quality multi-panel figures, I recommend learning ggplot
#in combination withthe ggpubr package
#This command tells us to lay out plots in one row and two columns
par(mfrow=c(1,2))

#plot one is a histogram, use ?hist() for documentation
#try playing with the "breaks" argument
hist(penguins$body_mass_g,
     freq=FALSE, #Tells R to plot density rather than frequency on Y-axis
     main="",
     xlab="Penguin Body Mass (g)")

#plot two is a density plot, use ?density() for documentation
#try playing with the "bw" argument
plot(density(penguins$body_mass_g,na.rm=T),
     main="",
     xlab="Penguin Body Mass (g)")
```



```
#Make sure to set this back to one row and one column for your next figures
par(mfrow=c(1,1))
```

Does this look normal to you? To me, these data seem quite skewed and possibly multimodal.

One way to visually check whether data follows a particular distribution is by using something called a "Quantile-Quantile Plot". In particular, these are often used for checking normality. These plots compare two distributions (usually an empirical and a theoretical distribution - like our penguin data and a normal distribution) using their quantiles (also called percentiles). If the two distributions have the same form, then the "Q-Q Plot" should be a line.

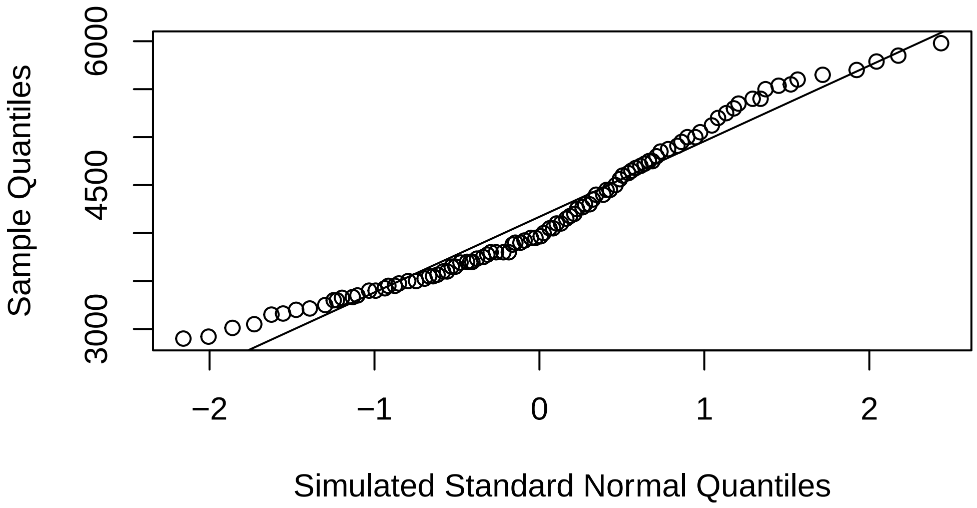
What is a quantile? They are points in your distribution below which a certain proportion of values fall. For example, the 0.90 quantile is the number such that 90% of your data falls below that value. You can generate quantiles for data easily using the `quantile` function:

```
quantile(x = penguins$body_mass_g,
         probs = 0.90,
         na.rm=T) # we need to tell this function to ignore the NA values
```

```
 90%
5400
```

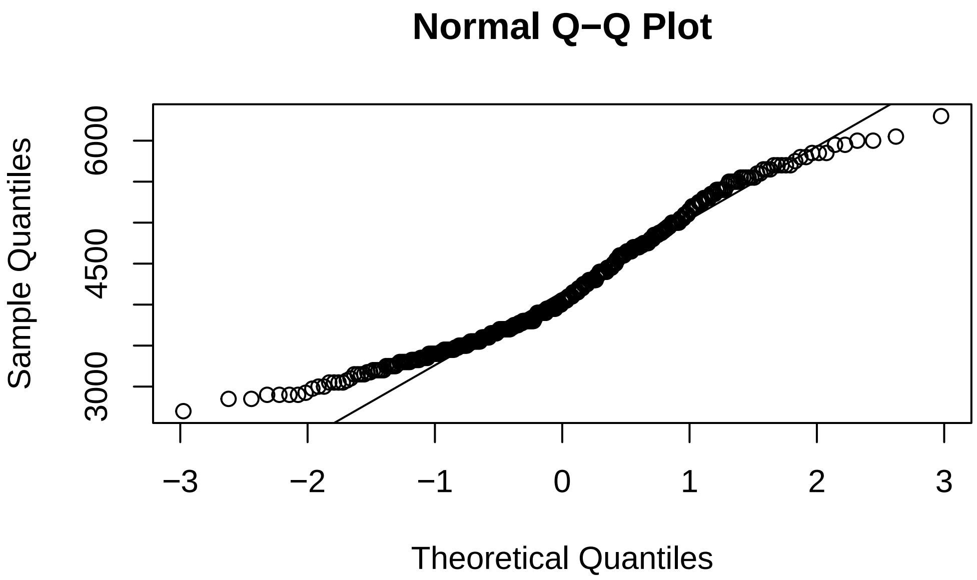
So, to compare the quantiles of our penguin data to the standard normal distribution we could write:

```r
#First let's generate some data from a standard normal distribution
norm_data <- rnorm(1000)

#Let's now calculate the quantiles for our penguins
quantiles_penguins <- quantile(x = penguins$body_mass_g,
         probs = seq(0.01,0.99,0.01),
         na.rm=T)

#And for our theoretical distribution
quantiles_norm <- quantile(x = norm_data,
         probs = seq(0.01,0.99,0.01))

#Now, let's plot them against one another
plot(quantiles_norm,quantiles_penguins,
     xlab="Simulated Standard Normal Quantiles",
     ylab="Sample Quantiles")
#For good measure, let's add a trend line to make it easier to visualize
abline(lm(quantiles_penguins~quantiles_norm))
```



Hmm, that doesn't quite look like a straight line to me… For future reference, we can generate the same plot using the functions `qqnorm` and `qqline`.

```
qqnorm(penguins$body_mass_g)
qqline(penguins$body_mass_g)
```



I will note that the interpretation of Q-Q plots is inherently subjective. Nevertheless, they serve as a pretty good check for whether there's something wonky with your data.

Problem 3.1 (5 Points)

Many biological variables follow a lognormal distribution (see Aho Ch. 3.3.2.8). That is, if you take their logarithm, the resulting values should be normally distributed. For this reason, we often apply a log-transformation to our data prior to analysis, particularly if the data is skewed. Write a function that generates a Q-Q plot (with a reference trendline) on the log-transformation of input data. Then, apply this function to your penguin mass data and provide an interpretation of what the output means.

Problem 3.2 (5 Points)

Now, try applying the function you wrote in 4.1 to the different species of penguin recorded in the dataset, the different penguin sexes recorded in the dataset, and the various species-sex combinations individually. Then, provide an interpretation for what these Q-Q plots tell you.

To get full credit for this problem use the `par` function to array these plots into three figures (so your grader maintains some semblance of sanity). For best visibility, I recommend using `par(mfrow=c(1,3)` for the three species plots, `par(mfrow=c(1,2)` for the two sex plots, and `par(mfrow=c(2,3)` for the six species-sex combinations. Be sure to label each plot with a descriptive title using the `main` option in the `plot` function to get full credit.

Hint: You may need to add a second input argument to your function from the last question to control the title of your plots.

Note: A binary view of sex, as encoded in this dataset, can often limit the scope of biological inference made. In many systems, explicit acknowledgement of sex as the complex multivariate and non-discrete phenotype it is can yield new biological insight.

5. List of All Problems (Above, 35 Points)

- 1.1(5 Points)
- 1.2 (5 Points)
- 2.1 (5 Points)
- 2.2 (10 Points)
- 3.1 (5 Points)
- 3.2 (5 Points)

6. Acknowledgements

Problem 1.2 adapted from problem sets shared by Dr. Philip L.F. Johnson. Problem 1.1 adapted from problem sets shared by Dr. Heather Lynch.