Homework 6

Autumn 2022

Jaiden Atterbury

Instructions

- This homework is due in Gradescope on Wednesday Nov 16 by midnight PST.
- Please answer the following questions in the order in which they are posed.
- Don't forget to knit the document frequently to make sure there are no compilation errors.
- When you are done, download the PDF file as instructed in section and submit it in Gradescope.

---

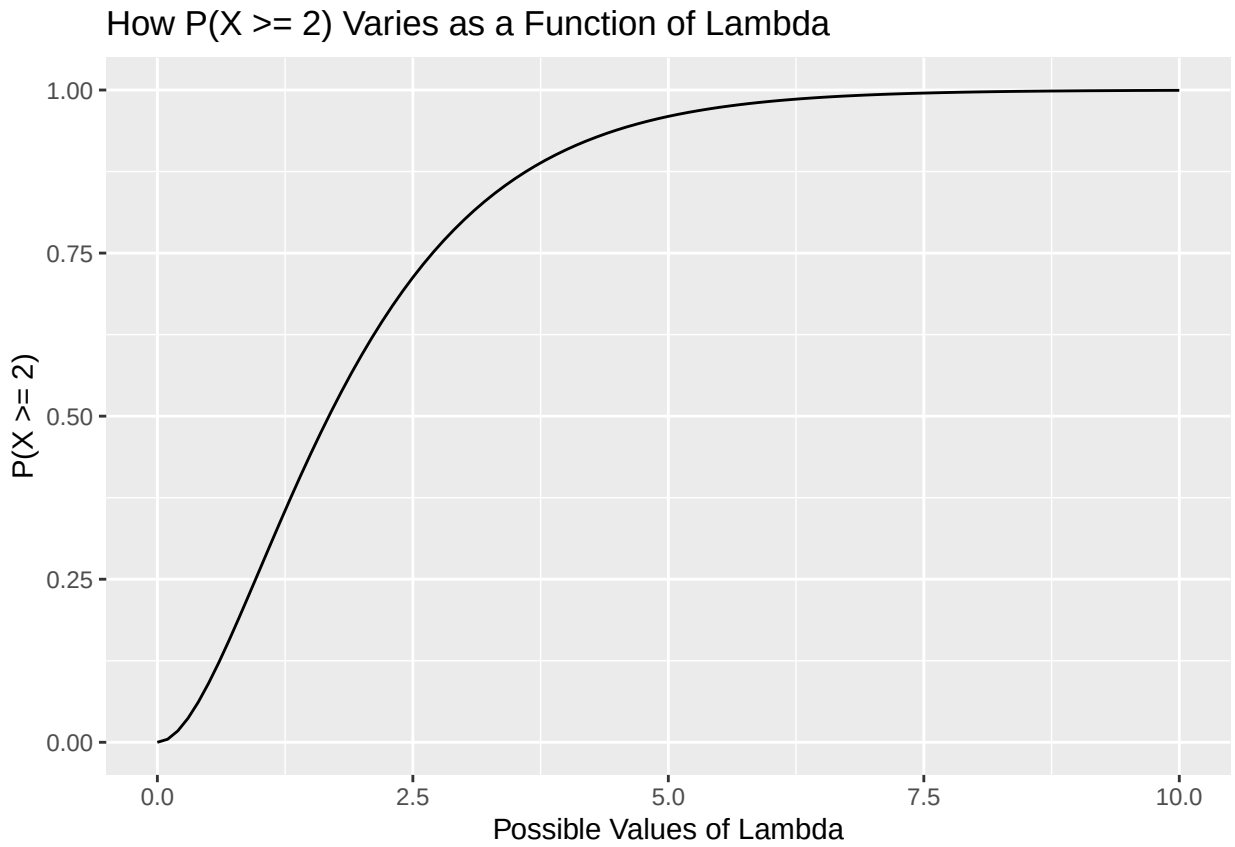
1. The number of chocolate chips in a certain type of cookie has a Poisson distribution. We want the probability that a randomly chosen cookie has at least 2 chocolate chips to be greater than 0.99.

a. Make a line plot of this probability as a function of λ.

Defining X and Finding a Function for λ:
Let X be the random variable that counts the number of chocolate chips on a randomly selected cookie. We want the $P(X \geq 2) > 0.99$. To find how this probability changes as a function of lambda, we need to find the $P(X \geq 2)$ using the PMF of $X \sim Poisson(\lambda)$. Observe that:

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - P(X \leq 1) \\ &= 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - e^{-\lambda} - e^{-\lambda}\lambda \\ &= 1 - e^{-\lambda}(1 - \lambda) \end{aligned}$$

```
ggplot() +
  geom_function(fun = function(x){(1 - (exp(-1 * x) * (1 + x)))}, xlim = c(0, 10)) +
  labs(x = "Possible Values of Lambda",
       y =  "P(X >= 2)",
       title = "How P(X >= 2) Varies as a Function of Lambda",
  )
```



b. Find the smallest value of the parameter λ of this distribution that ensures this probability.

(Modify the `uniroot` code below. You can pick 0 for a lower bound for the mean and 10 for the upper bound)

```
small <- uniroot(f = function(x){(1 - (exp(-1 * x) * (1 + x))) - 0.99}, lower= 0, upper= 10)

small_lambda <- small$root
```

Smallest value of λ:
The smallest value of λ of this distribution that ensures that the $P(X \geq 2) > 0.99$ is $\lambda = 6.6383538$.

2. (Symmetry) Suppose that $f(x)$ is the PDF corresponding to a continuous random variable which is symmetrically distributed about 0. That is, for any $a > 0$

$$f(a) = f(-a).$$

Show that

$$P(-a \leq X \leq a) = 2F(a) - 1$$

where $F(x)$ is the CDF corresponding to $f(x)$.

Hint:

$$P(-a \leq X \leq a) = P(X \leq a) - P(X < -a).$$

Write

$$P(X < -a) = \int_{-\infty}^{-a} f(x)dx.$$

Make a change of variable $u = -x$ and proceed from there.

$$
\begin{aligned}
P(-a \le X \le a) &= P(X \le a) - P(X < -a) \\
&= F(a) - \int_{-\infty}^{-a} f(x)dx \\
&= F(a) + \int_{\infty}^{a} f(-u)du \quad (\text{Let } u = -x) \\
&= F(a) - \int_{a}^{\infty} f(-u)du \\
&= F(a) - \int_{a}^{\infty} f(u)du \quad (\text{By symmetry}) \\
&= F(a) - \int_{a}^{\infty} f(x)dx \\
&= F(a) - P(X \ge a) \quad (\text{Definition 9.1}) \\
&= F(a) - (1 - P(X \le a)) \quad (\text{Theorem 2.1.a and Lemma 9.1}) \\
&= F(a) - 1 + F(a) \\
&= 2F(a) - 1
\end{aligned}
$$

3. (Flooding river) A river floods every year. Suppose that the low-water mark is set at 1 and the high-water mark X has CDF

$$
\begin{aligned}
F_X(x) &= P(X \le x) \\
&= \begin{cases} 0 & x < 1 \\ 1 - 1/x^2 & 1 \le x < \infty. \end{cases}
\end{aligned}
$$

a. Find a PDF, $f(x)$ for X.

Finding a PDF of X:
In order to find a PDF of X we must use Lemma 9.2 which states that, $f(x) = \frac{d}{dx}F(x)$. Since $F(x)$ is a multi-part function we must take the derivative of each part.

a.) Derivative of 0:
$\frac{d}{dx}(0) = 0$ with a range of $-\infty < x < 1$.

b.) Derivative of $1 - \frac{1}{x^2}$:
$\frac{d}{dx}(1 - \frac{1}{x^2}) = \frac{2}{x^3}$ with a range of $1 \le x < \infty$.

Thus, the PDF of X is:

$$
\begin{aligned}
f(x) &= \frac{d}{dx}F(x) \\
&= \begin{cases} 0 & x < 1 \\ \frac{2}{x^3} & 1 \le x < \infty. \end{cases}
\end{aligned}
$$

b. If the low-water mark is set to 0 and we use a unit of measurement that is $\frac{1}{10}$ of that given previously, the high-water mark becomes

$$Y = 10\,(X - 1).$$

Find $P(Y \le 1)$.

Finding the CDF of Y:
In order to find the $P(Y \leq 1)$, we must first find the CDF of Y. To do this, observe that:

$$\begin{aligned} G(Y) &= P(Y \leq y) \\ &= P(10(X-1) \leq y) \\ &= P(X \leq \frac{1}{10}y + 1) \\ &= F(\frac{1}{10}y + 1) \\ &= 1 - \frac{1}{(\frac{1}{10}y+1)^2}, 0 \leq y < \infty \end{aligned}$$

Finding $P(Y \leq 1)$:

$$\begin{aligned} P(Y \leq 1) &= \int_0^1 1 - \frac{1}{(\frac{1}{10}y+1)^2} dy \\ &= \int_0^1 1 - (\frac{1}{10}y+1)^{-2} dy \\ &= 10 \int_1^{\frac{11}{10}} 1 - u^{-2} du \quad (\text{Let u} = \frac{1}{10}y + 1) \\ &= 10 \left[u + \frac{1}{u} \right]_1^{\frac{11}{10}} \\ &= 10(\frac{221}{110} - \frac{200}{100}) \\ &= 10(\frac{100}{11000}) \\ &= \frac{1}{11} \end{aligned}$$

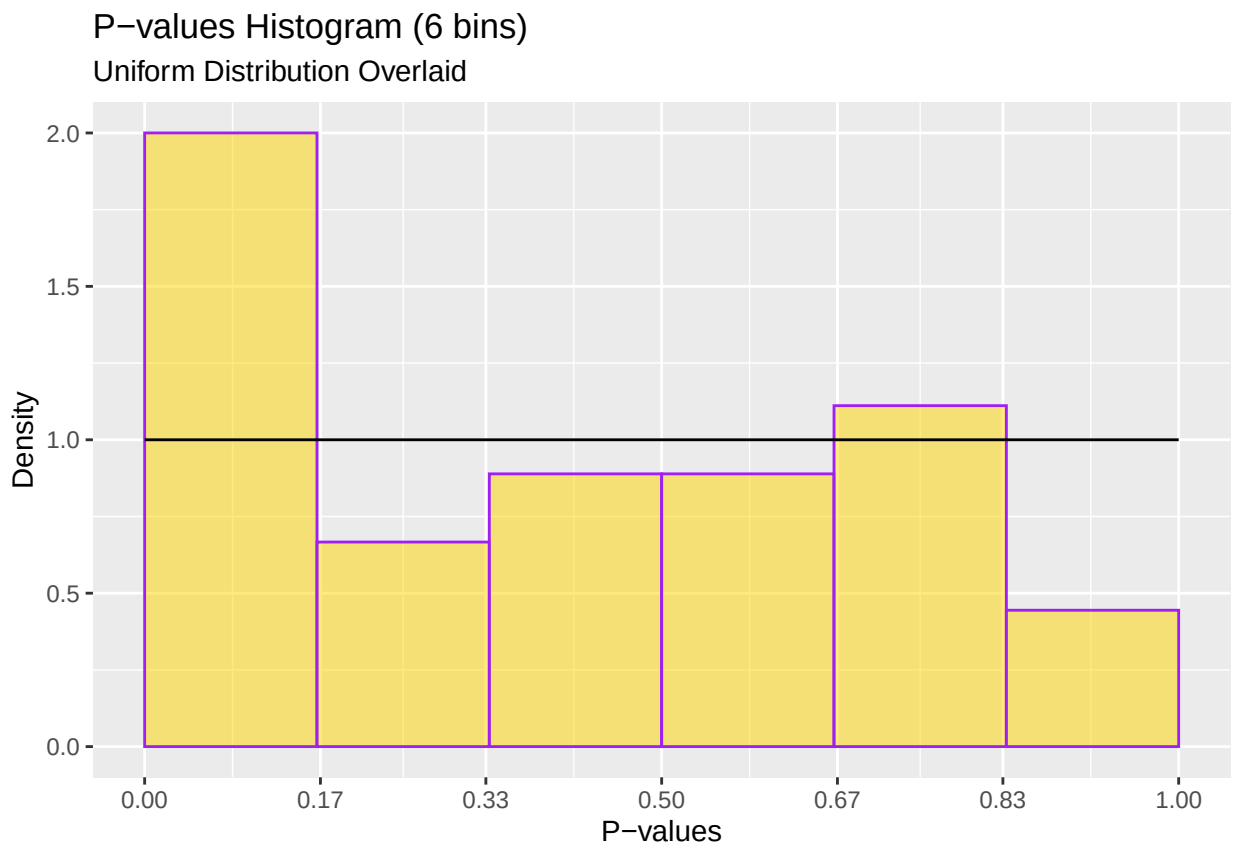
4. The following ordered set of 27 P-values (from Kaati et al., Eur. J. Hum Genetics 2007) were the result of testing many independent subgroups of a sample.

```
pvalue_df <- data.frame(
  pval = c(0.01, 0.01, 0.02, 0.04, 0.04, 0.05, 0.07, 0.07, 0.10, 0.19, 0.24, 0.27, 0.34,
0.37, 0.44, 0.50, 0.53, 0.54, 0.55, 0.61, 0.70, 0.77, 0.80, 0.80, 0.82, 0.94, 0.99)
)
```

It is hypothesized that the P-values all come from a uniform distribution on $[0, 1)$. Fit a uniform distribution to these data. How well does it fit (just visually)?

```
ggplot() +
  geom_histogram(data = pvalue_df,
                 mapping = aes(x = pval, y = ..density..),
                 binwidth = 1/6,
                 breaks = seq(0, 1.0, 1/6),
                 alpha = 0.5,
                 color = "purple",
                 fill = "gold") +
  labs(x = "P-values",
       y = "Density",
       title = "P-values Histogram (6 bins)",
       subtitle = "Uniform Distribution Overlaid") +
  geom_function(fun = dunif,
                args = list(min = 0, max = 1),
```

```
                xlim = c(0,1)) +
  scale_x_continuous(breaks = round(seq(0, 1, 1/6), 2))
```



Why Did I Choose 6 Bins?:
Using Sturges Rule $1 + \log_2 n$ we find that the optimal bin size for $n = 40$ is $1 + \log_2 n = 6.322$. Since this is closer to 6 bins than 7, I decided that 6 would be the more optimal number of bins for the histogram.

How Well Does $Unif \sim (0, 1)$ Fit?:
Other than the first and last bin, the uniform distribution on $[0, 1)$ fits the data pretty well. However, when we take into account that the first and last bin (especially the first bin) are poorly captured by the uniform distribution, it is safe to say that, overall, the uniform distribution on $[0, 1)$ doesn't fit the data very well.