140.651 Problem Set 1 Solutions

Problem 1

Show the following:

a. $P(\emptyset) = 0$.

$$\begin{aligned}
\Omega &= \emptyset \cup \Omega && \text{by definition of } \Omega \text{ and } \emptyset \\
P(\Omega) &= P(\emptyset \cup \Omega) \\
P(\Omega) &= P(\emptyset) + P(\Omega) && \text{by finite additivity as } \Omega \text{ and } \emptyset \text{ are mutually exclusive} \\
1 &= P(\emptyset) + 1 && \text{since } P(\Omega) = 1 \text{ by the definition of a probability measure} \\
0 &= P(\emptyset)
\end{aligned}$$

b. $P(E) = 1 - P(E^c)$.

$$\begin{aligned}
E \cup E^c &= \Omega && \text{by definition of set complement} \\
P(E \cup E^c) &= P(\Omega) \\
P(E) + P(E^c) &= P(\Omega) && \text{by finite additivity since } E \cap E^c = \emptyset \text{ so } E \text{ and } E^c \text{ are disjoint} \\
P(E) + P(E^c) &= 1 && \text{by the definition of a probability measure } P(\Omega) = 1 \\
P(E) &= 1 - P(E^c)
\end{aligned}$$

c. If $A \subset B$ then $P(A) \leq P(B)$.

$$\begin{aligned}
&B = A \cup (A^c \cap B) && \text{since } A \subset B \\
&P(B) = P(A \cup (A^c \cap B)) \\
&P(B) = P(A) + P(A^c \cap B) && \text{by finite additivity since } A \cap (A^c \cap B) = \emptyset \text{ so } A \text{ and } (A^c \cap B) \text{ are disjoint} \\
&P(B) \geq P(A) && \text{since } P(A^c \cap B) \geq 0 \text{ by definition of a probability measure}
\end{aligned}$$

d. For any A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

We can express A and B as,

$$A = A \cap \Omega = A \cap (B \cup B^c) = (A \cap B) \cup (A \cap B^c) \tag{1}$$

$$B = B \cap \Omega = B \cap (A \cup A^c) = (B \cap A) \cup (B \cap A^c) \tag{2}$$

Furthermore, note that $(A \cap B)$, $(A \cap B^c)$, and $(B \cap A^c)$ are mutually disjoint since,

$$\begin{aligned}
(A \cap B) \cap (A \cap B^c) &= A \cap B \cap A \cap B^c = A \cap B \cap B^c = \emptyset \\
(B \cap A) \cap (B \cap A^c) &= B \cap A \cap B \cap A^c = B \cap A \cap A^c = \emptyset \\
(A \cap B^c) \cap (B \cap A^c) &= A \cap B^c \cap B \cap A^c = \emptyset
\end{aligned}$$

Consequently, by finite additivity,

$$\begin{aligned}
P(A) &= P(A \cap B) + P(A \cap B^c) \\
P(B) &= P(B \cap A) + P(B \cap A^c)
\end{aligned}$$

Implying that

$$P(A \cap B^c) = P(A) - P(A \cap B) \tag{3}$$

$$P(A^c \cap B) = P(B) - P(B \cap A) \tag{4}$$

Substituting (1) and (2) into $P(A \cup B)$, we get

$$\begin{aligned}
P(A \cup B) &= P((A \cap B) \cup (A \cap B^c) \cup (B \cap A) \cup (B \cap A^c)) && \text{by (1) and (2)} \\
&= P((A \cap B) \cup (A \cap B^c) \cup (B \cap A^c)) \\
&= P(A \cap B) + P(A \cap B^c) + P(B \cap A^c) && \text{by finite additivity} \\
&= P(A \cap B) + (P(A) - P(A \cap B)) + (P(B) - P(B \cap A)) && \text{by (3) and (4)} \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned}$$

e. $P(A \cup B) = 1 - P(A^c \cap B^c)$.

$$\begin{aligned}
\Omega &= (A \cup B) \cup (A \cup B)^c && \text{by definition of set complement} \\
P(\Omega) &= P((A \cup B) \cup (A \cup B)^c) \\
P(\Omega) &= P(A \cup B) + P((A \cup B)^c) && \text{by finite additivity since } (A \cup B) \cap (A \cup B)^c = \emptyset \\
1 &= P(A \cup B) + P((A \cup B)^c) && \text{by definition of a probability measure} \\
P(A \cup B) &= 1 - P((A \cup B)^c) \\
P(A \cup B) &= 1 - P(A^c \cap B^c) && \text{by DeMorgan's Laws}
\end{aligned}$$

f. $P(A \cap B^c) = P(A) - P(A \cap B)$.

$$\begin{aligned}
A &= A \cap \Omega \\
&= A \cap (B \cup B^c) && \text{by definition of complement} \\
&= (A \cap B) \cup (A \cap B^c) \\
P(A) &= P(A \cap B) + P(A \cap B^c) && \text{by finite additivity since } (A \cap B) \text{ and } (A \cap B^c) \text{ are disjoint} \\
P(A) - P(A \cap B) &= P(A \cap B^c)
\end{aligned}$$

g. $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$.

We will prove this statement using induction.
Base Case: Suppose $n = 2$. Then,

$$\begin{aligned}
P\left(\bigcup_{i=1}^{2} E_i\right) &= P(E_1 \cup E_2) \\
&= P(E_1) + P(E_2) - P(E_1 \cap E_2) && \text{by part 1d} \\
&\leq P(E_1) + P(E_2) && \text{by non-negativity of a probability measure} \\
&= \sum_{i=1}^{2} P(E_i)
\end{aligned}$$

Induction hypothesis: Assume this statement holds for n: $P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i)$.

Induction step: Using induction, we will show that the above statement holds for $n+1$.

$$\begin{aligned}
P\left(\bigcup_{i=1}^{n+1} E_i\right) &= P\left(\left(\bigcup_{i=1}^{n} E_i\right) \cup E_{n+1}\right) \\
&= P\left(\bigcup_{i=1}^{n} E_i\right) + P(E_{n+1}) - P\left(\left(\bigcup_{i=1}^{n} E_i\right) \cap E_{n+1}\right) && \text{by part 1d} \\
&\leq \sum_{i=1}^{n} P(E_i) + P(E_{n+1}) - P\left(\left(\bigcup_{i=1}^{n} E_i\right) \cap E_{n+1}\right) && \text{by the induction hypothesis} \\
&\leq \sum_{i=1}^{n+1} P(E_i) - P\left(\left(\bigcup_{i=1}^{n} E_i\right) \cap E_{n+1}\right) \\
&\leq \sum_{i=1}^{n+1} P(E_i) && \text{by non-negativity}
\end{aligned}$$

Thus, by induction we have shown that $P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$ holds for all natural numbers $n \geq 2$.

h. $P(\cup_{i=1}^{n} E_i) \geq \max_i P(E_i)$.

Since $E_i \in \cup_i^n E_i$, by part c, $P(E_i) \leq P\left(\cup_{i=1}^{n} E_i\right)$ for all $i = 1, 2, ..., n$. Consequently, $\max_i P(E_i) \leq \sum_{i=1}^{n} P(E_i)$.

Problem 2

Cryptosporidium is a pathogen that can cause gastrointestinal illness with diarrhea; infections can lead to death in individuals with a weakened immune system. During a recent outbreak of cryptosporidiosis in 21% of two parent families at least one of the parents has contracted the disease. In 9% of the families the father has contracted cryptosporidiosis while in 5% of the families both the mother and father have contracted cryptosporidiosis.

a. What event does the probability one minus the probability that both have contracted cryptosporidiosis represent?

This is the probability that at most one parent have contracted influenza. So the probability either the mother alone, the father alone, or neither parent has contracted the flu.

b. What's the probability that either the mother or the father has contracted cryptosporidiosis?

Let F be the event that the father contracted cryptosporidiosis and M be the event that the mother contracted cryptosporidiosis. Using this notation, we are given $P(F \cup M) = 0.21$, $P(F \cap M) = 0.05$, and $P(F) = 0.09$. Note that the event in which either the mother or the father has contracted cryptosporidiosis is denoted by $F \cup M$, or equivalently, the event in which least one parent contracted the disease. We are given $P(F \cup M)$ is 0.21.

c. What's the probability that the mother has contracted cryptosporidiosis but the father has not?

Using the same notation as in part b,

$$M \cap F^c = \{\text{mother contracted cryptosporidiosis, but father has not}\}$$

By part 1f, we know that

$$P(M \cap F^c) = P(M) - P(M \cap F) = 0.17 - 0.05 = 0.12$$

d. What's the probability that the mother has contracted cryptosporidiosis?

Recall the result given in 1d: $P(F \cup M) = P(F) + P(M) - P(F \cap M)$. Thus, solving for $P(M)$, we get

$$P(M) = P(F \cup M) + P(F \cap M) - P(F) = 0.21 + 0.05 - 0.09 = 0.17$$

e. What's the probability that neither the mother nor the father has contracted cryptosporidiosis?

Using the notation given in part b,

$$M^c \cup F^c = \{\text{neither the mother nor the father has contracted cryptosporidiosis}\}$$

Applying part 1e, we have $P(F \cup M) = 1 - P(F^c \cup M^c)$, or equivalently,

$$P(F^c \cap M^c) = 1 - P(F \cup M) = 1 - 0.21 = 0.79$$

f. What's the probability that the mother has contracted cryptosporidiosis but the father has not?

Using the notation given in part b,

$$M \cap F^c = \{\text{the mother has contracted cryptosporidiosis, but the father has not}\}$$

Applying part 1f and the result from 2d, we have

$$P(M \cap F^c) = P(M) - P(M \cap F) = 0.17 - 0.05 = 0.12$$

Problem 3

Suppose $h(x)$ is such that $h(x) > 0$ for $x = 1, 2, \dots, I$. Argue that $p(x) = h(x)/\sum_{i=1}^{I} h(i)$ is a valid pmf.

To show that $p(x)$ is a valid pmf, we need to show (1) $p(x) \geq 0$ for $x = 1, 2, \dots, I$ and (2) $\sum_{x=1}^{I} p(x) = 1$.

(1) Since $h(x) > 0$ for all $x = 1, 2, \dots, I$, $\sum_{i=1}^{n} h(x) > 0$ and thus $p(x) > 0$.

(2) Checking the second condition,

$$\sum_{x=1}^{I} p(x) = \sum_{x=1}^{I} \frac{h(x)}{\sum_{i=1}^{I} h(i)} = \frac{\sum_{x=1}^{I} h(x)}{\sum_{i=1}^{I} h(i)} = 1$$

Problem 4

Suppose a function h is such that $h > 0$ and $c = \int_{-\infty}^{\infty} h(x)dx < \infty$. Show that $f(x) = h(x)/c$ is a valid density.

To show that $f(x)$ is a valid pdf, we need to show (1) $f(x) \geq 0$ for all x and (2) $\int_{-\infty}^{\infty} f(x)dx = 1$.

(1) Since $h(x) > 0$ for all x, $c = \int_{-\infty}^{\infty} h(x)dx > 0$ and thus $f(x) > 0$.

(2) Checking the second condition,

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{h(x)}{c}dx = \frac{1}{c}\int_{-\infty}^{\infty} h(x)dx = \frac{c}{c} = 1$$

Problem 5

Suppose that, for a randomly drawn subject from a particular population, the proportion of a their skin that is covered in freckles follows a density that is constant on $[0, 1]$. (This is called the **uniform density** on $[0, 1]$.) That is, $f(x) = k$ for $0 \leq x \leq 1$.

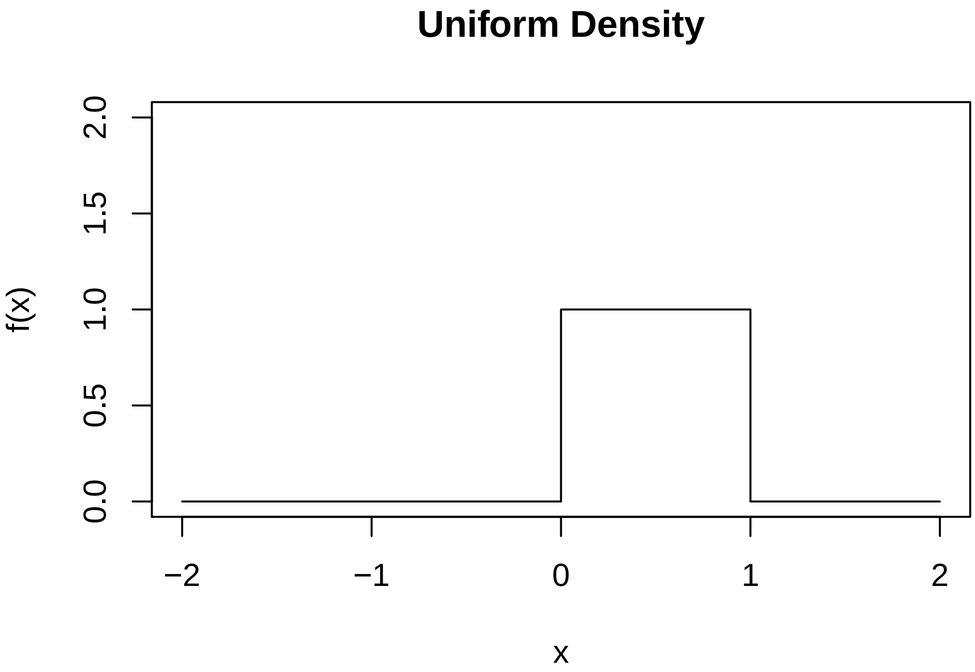
a. Draw this density. What must k be?

Recall that a valid density $f(x)$ requires (1) $f(x) \geq 0$ for all x and (2) $\int_{-\infty}^{\infty} f(x)dx = 1$. Condition (1) tells us that $k \geq 0$, but not exactly what k must be. However, from condition (2), we see that,

$$\int_0^1 f(x)dx = \int_0^1 kdx = k \cdot x\Big|_0^1 = k - 0 = k$$

Since $\int_0^1 f(x)dx = 1$ and from above we see that $\int_0^1 f(x)dx = k$, $k = 1$. We can use the following `R` code to draw the density:

```
x <- c(-2, 0, 0, 1, 1, 2)
y <- c(0, 0, 1, 1, 0, 0)
plot(x,y, "l", xlim = c(-2,2), ylim = c(0,2),
     main = "Uniform Density", xlab = "x", ylab = "f(x)")
```



b. Suppose a random variable, X, follows a uniform distribution. What is the probability that X is between .1 and .7? Interpret this probability in the context of the problem.

The probability that X is between 0.1 and 0.7 is given by,

$$P(0.1 < X < 0.7) = \int_{0.1}^{0.7} f(x)dx = \int_{0.1}^{0.7} 1dx = x\Big|_{0.1}^{0.7} = 0.7 - 0.1 = 0.6$$

In the context of this problem, $P(0.1 < X < 0.7) = 0.6$ is the probability that an individual has between 0.1 and 0.7 of their skin covered in freckles.

c. Verify the previous calculation in R. What is the probability that $a < X < b$ for generic values $0 < a < b < 1$

Recall that for a general random variable X, $P(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx$ where $F(\cdot)$ is the cumulative distribution function of X and $f(\cdot)$ is the probability density function. Thus to verify the previous calculation in `R`, we can use the `punif(x)` function which gives $F(x)$.

```
lb <- 0.1    # Lowerbound
ub <- 0.7    # Upperbound
punif(ub) - punif(lb)
```

```
[1] 0.6
```

This agrees with what we found in part (b)! Repeating the calculation in part (b) for general bounds a and b, we get

$$P(a < X < b) = \int_a^b f(x)dx = \int_a^b 1dx = x\Big|_a^b = b - a$$

d. What is the distribution function associated with this density?

The distribution function is denoted $F(x) = P(X < x)$. For the uniform density,

$$F(x) = P(X < x) = \int_0^x f(y)dy = \int_0^x 1dy = y\Big|_0^x = x - 0 = x$$

e. What is the median of this density? Interpret the median in the context of this problem.

Let m denote the median. By definition of median, $P(X < m) = F(m) = 0.5$. From part (d) above, we know that $F(x) = x$, implying that $m = 0.5$. In the context of our problem, the median of 0.5 means that the probability of individuals having less than 50% of their skin covered in freckles is 0.5.

f. What is the 95^{th} percentile? Interpret this percentile in the context of the problem.

The 95th percentile is the value x such that $F(x) = 0.95$. From part (d), we know that $F(x) = x$, implying that the 95th percentile is given by $x = 0.95$. In the context of this problem, the 95th percentile means that the probability of a randomly drawn subject having less that 95% of their skin covered in freckles is 0.95.

g. Do you believe that the proportion of freckles on subjects in a given population could feasibly follow this distribution? (Why or why not.)

The uniform distribution is probably not a reasonable distribution for the proportion of freckles on subjects in a given population as it should not be equally likely to have over 50% of your skin covered in freckles as it is to have less than 50% of your skin covered in freckles.

Problem 6

Let U be a continuous random variable with a uniform density on $[0, 1]$ and $F(\cdot)$ be any strictly increasing cdf.

a. Show that $F^{-1}(U)$ is a random variable with cdf equal to F.

Let $F_U(u)$ denote the CDF of U. From part (d) of problem 5, we know that $F_U(u) = u$. Since $F(\cdot)$ is strictly increasing, F is invertible. Now, define $X = F^{-1}(U)$. Thus, the CDF of X is given by,

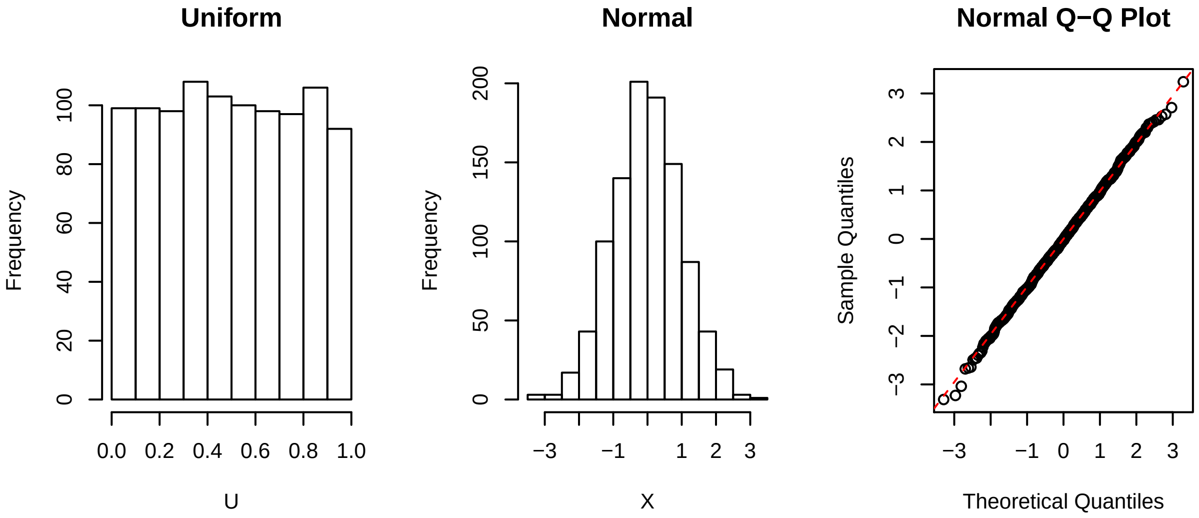
$$F_X(x) = Pr(X \leq x) = Pr(F^{-1}(U) \leq x) = Pr(U \leq F(x)) = F_U(F(x)) = F(x)$$

b. Describe a simulation procedure in R that can simulate an iid sample from a distribution with a given cdf $F(\cdot)$.

From part (a), we see that if want to sample a random variable from a distribution given only the cdf, we can first simulate an iid sample $\{U_i\}_{i=1,2,\dots n}$ from Uniform[0,1] using function `runif()`. Then, we can apply the transformation $X_i = F^{-1}(U_i)$ to get iid samples X_i from a distribution with cdf $F(\cdot)$.

c. Simulate uniform random variables and apply the normal quantile function to obtain normal variables.

```
# Simulate 1000 variables to better view the distribution.
set.seed(123)
U = runif(1000, min=0, max=1)    # Simulate 1000 Uniform[0,1] random variables
X = qnorm(U)                     # Apply the transformation X = F^{-1}(U)
par(mfrow=c(1,3))
hist(U, main="Uniform")
hist(X, main="Normal")
# Check that X follows a normal distribution using a QQ plot
qqnorm(X)
qqline(X, col=2, lty=2)
```



Problem 7

Let $0 \leq \pi \leq 1$ and f_1 and f_2 be two continuous densities with associated distribution functions F_1 and F_2 and survival functions S_1 and S_2. Let $g(x) = \pi f_1(x) + (1-\pi) f_2(x)$.

a. Show that g is a valid density.

To show that $g(x)$ is a valid density, we need to show (1) $g(x) \geq 0$ for all x and (2) $\int g(x)dx = 1$.

(1) Since f_1 and f_2 are valid densities, we know that $f_1(x) \geq 0$ and $f_2(x) \geq 0$ for all x. Furthermore, as $0 \leq \pi \leq 1$, $0 \leq 1-\pi \leq 1$, so $g(x) = \pi f_1(x) + (1-\pi) f_2(x)$ is non-negative.

(2) Checking the second condition,

$$\int g(x)dx = \int \pi f_1(x) + (1-\pi) f_2(x) dx = \pi \int f_1(x)dx + (1-\pi) \int f_2(x)dx \overset{(*)}{=} \pi + (1-\pi) = 1$$

where $(*)$ follows from the fact that f_1 and f_2 are valid densities, so $\int f_1(x)dx = \int f_2(x)dx = 1$.

b. Write the distribution function associated with g in the terms of F_1 and F_2.

Let $F_G(x)$ be the distribution function associated with g. Then,

$$\begin{aligned} F_G(x) &= \int_{-\infty}^{x} g(y)dy = \int_{-\infty}^{x} \pi f_1(y) + (1-\pi) f_2(y)dy = \pi \int_{-\infty}^{x} f_1(y)dy + (1-\pi) \int_{-\infty}^{x} f_2(y)dy \\ &= \pi F_1(x) + (1-\pi) F_2(x) \end{aligned}$$

c. Write the survival function associated with g in the terms of S_1 and S_2.

Let $S_G(x)$ be the survival function associated with g. By the definition of survival functions, we have

$$\begin{aligned} S_G(x) &= \int_{x}^{\infty} g(y)dy = \int_{x}^{\infty} \pi f_1(y) + (1-\pi) f_2(y)dy = \pi \int_{x}^{\infty} f_1(y)dy + (1-\pi) \int_{x}^{\infty} f_2(y)dy \\ &= \pi S_1(x) + (1-\pi) S_2(x) \end{aligned}$$

Problem 8

Radiologists have created cancer risk summary that, for a given population of subjects, follows (a specific instance of) the **logistic** density

$$\frac{e^{-x}}{(1+e^{-x})^2} \qquad \text{for } -\infty < x < \infty.$$

a. Show that this is a valid density.

To show that this is a valid density, we need to show that (1) the function is non-negative and (2) the function integrates to 1.

(1) $e^x < 0$ for all $x \in \mathbb{R}$, and $1 + e^x > 0$, the function is non-negative.

(2) To verify the second requirement, we can use a change in variables. Let $u = 1 + e^{-x}$. Then, $du = -e^{-x}dx$, and,

$$\int_{-\infty}^{\infty} \frac{e^{-x}}{(1+e^{-x})^2} dx = \int_{\infty}^{1} -\frac{1}{u^2} du = \frac{1}{u}\bigg|_{\infty}^{1} = 1 - 0 = 1$$

b. Calculate the distribution function associated with this density.

Again, using a change of variables where $u = 1 + e^{-y}$ and $du = -e^{-y}dy$, we have

$$F(x) = \int_{-\infty}^{x} \frac{e^{-y}}{(1+e^{-y})^2} dy = \int_{\infty}^{1+e^{-x}} -\frac{1}{u^2} du = \frac{1}{u}\bigg|_{\infty}^{1+e^{-x}} = \frac{1}{1+e^{-x}} - 0 = \frac{1}{1+e^{-x}}$$

c. What value do you get when you plug 0 into the distribution function? Interpret this result in the context of the problem.

Plugging in 0 into the distribution function, we have

$$F(0) = \frac{1}{1+e^{-0}} = \frac{1}{2}$$

So 0 is the median of the distribution. In the context of this problem, a median of 0 means that the median cancer risk is 0.

d. Define the *odds* of an event with probability p as $p/(1-p)$. Prove that the p^{th} quantile from this distribution is $\log\{p/(1-p)\}$; which is the natural log of the odds of an event with probability p.

Recall from part (a) the cdf is given by $F(x) = \frac{1}{1+e^{-x}}$. To find the pth quantile, we want to find the value x such that $F(x) = p$. Solving for x, we have

$$\frac{1}{1+e^{-x}} = p \implies 1 + e^{-x} = \frac{1}{p} \implies e^{-x} = \frac{1}{p} - 1 \implies e^{-x} = \frac{1-p}{p} \implies -x = \log\left(\frac{1-p}{p}\right)$$
$$\implies x = -\log\left(\frac{1-p}{p}\right) \implies x = \log\left(\frac{p}{1-p}\right)$$

Problem 9

Quality control experts estimate that the time (in years) until a specific electronic part from an assembly line fails follows (a specific instance of) the **Pareto** cdf

$$F(x) = \begin{cases} 1 - \left(\frac{x_0}{x}\right)^{\alpha} & \text{for} \quad x \geq x_0 \\ 0 & \text{for} \quad x < x_0 \end{cases}$$

The parameter x_0 is called the scale parameter, while α is the shape or tail index parameter. The distribution is often denoted by $\text{Pa}(x_0, \alpha)$.

a. Derive the density of the Pareto distribution.

Recall that we can recover the PDF from the CDF via the relationship $f(x) = F'(x)$. Thus, for $x \geq x_0$,

$$\frac{d}{dx}\left(1 - \left(\frac{x_0}{x}\right)^{\alpha}\right) = \frac{d}{dx}\left(1 - x_0^{\alpha}x^{-\alpha}\right) = \alpha x_0^{\alpha} x^{-\alpha-1}$$

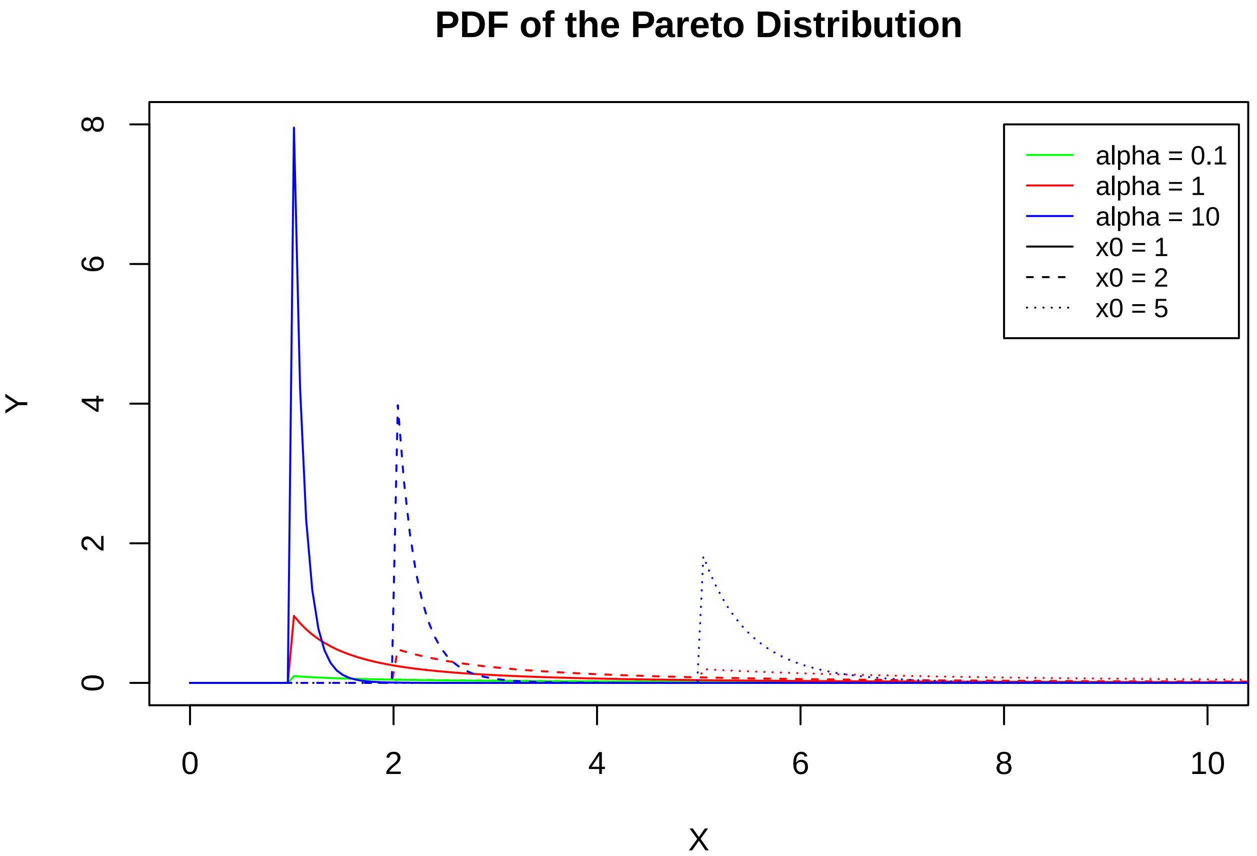
Thus, the pdf is given by,

$$f(x) = \begin{cases} \alpha x_0^{\alpha} x^{-\alpha-1} & \text{for} \quad x \geq x_0 \\ 0 & \text{for} \quad x < x_0 \end{cases}$$

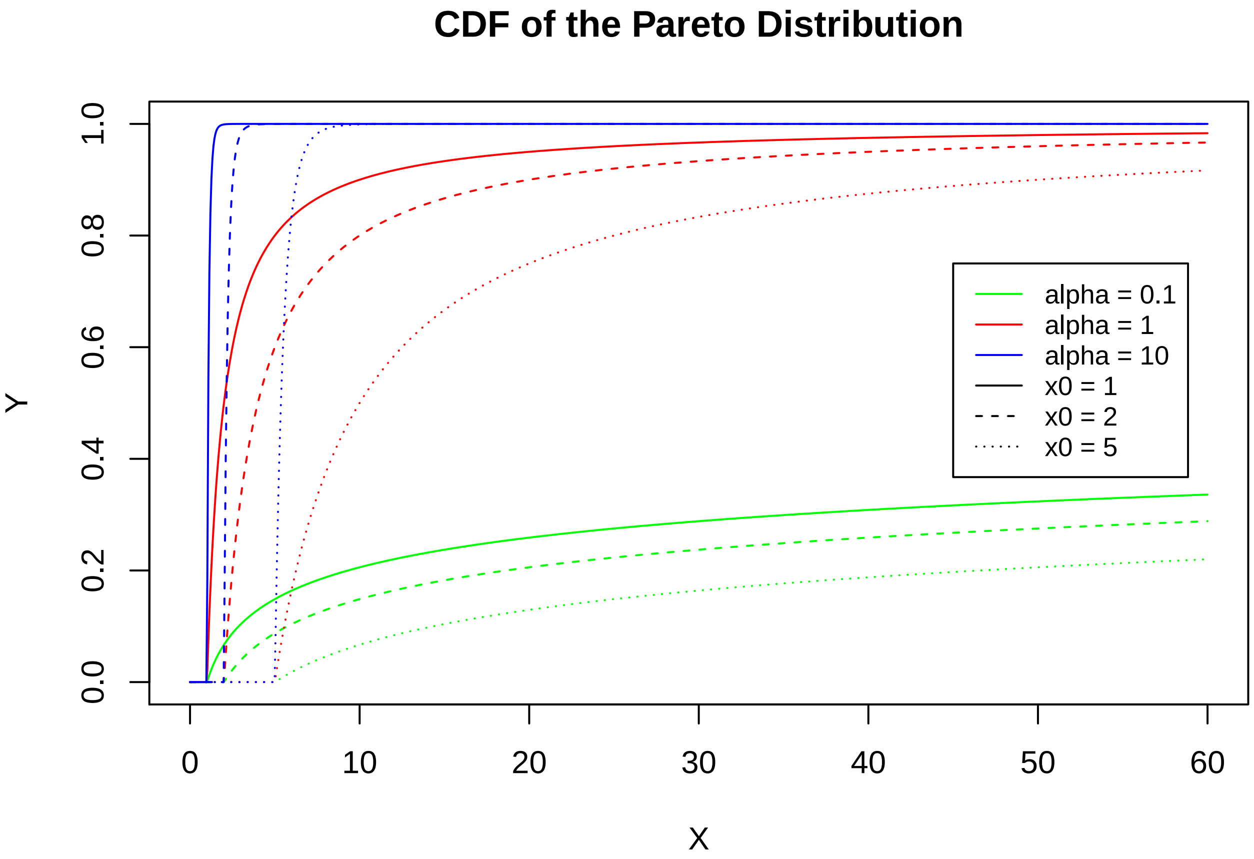
b. Plot the density and the cdf for $x_0 = 1, 2, 5$ and $\alpha = 0.1, 1, 10$. Comment on the interpretation.

```r
# Function to compute PDF of Pareto
pdf <- function(x0, xs, alph){
  pd <- alph * x0^alph*xs^(-alph - 1)
  pd[xs < x0] <- 0
  return(pd)
}

# Function to compute CDF of Pareto
cdf <- function(x0, xs, alph){
  cd <- 1 - (x0/xs)^alph   # Calculate the CDF for x >= x_0
  cd[xs < x0] <- 0         # If x < x_0, replace with 0.
  return(cd)
}

# Parameters and variables
#   x0s     Vector of all x_0
#   as      Vector of all alphas
#   xs      Sequence of x's
#   dat     Data frame of all combinations of x_0s and alphas
#   dat$lty Column representing the line types for each combination
#   dat$col Column representing the line colors for each combination
x0s <- c(1, 2, 5)
as <- c(0.1, 1, 10)
```

```
xs <- seq(0, 60, length = 1000)
dat <- expand.grid(x0 = x0s, alpha = as)
dat$lty <- rep(c("solid", "dashed", "dotted"), 3)
dat$col <- c(rep("green", 3), rep("red", 3), rep("blue", 3))
# Plot of PDF
plot(xs, rep(0, 1000), type = "n", ylim = c(0, 8), xlim = c(0, 10),
     xlab = "X", ylab = "Y", main = "PDF of the Pareto Distribution")
for (irow in 1:nrow(dat)) {
  lines(xs, pdf(dat$x0[irow], xs, dat$alpha[irow]),
        lty = dat$lty[irow], col = dat$col[irow])}
legend(x = 8, y = 8, legend = c("alpha = 0.1", "alpha = 1", "alpha = 10",
      "x0 = 1", "x0 = 2", "x0 = 5"), col = c("green", "red", "blue", "black",
      "black", "black"), lty = c("solid", "solid", "solid", "solid", "dashed",
      "dotted"), cex = 0.8)
```



```
# Plot of CDF
plot(xs, rep(0, 1000), type = "n", ylim = c(0, 1),
     xlab = "X", ylab = "Y", main = "CDF of the Pareto Distribution")
for (irow in 1:nrow(dat)) {
  lines(xs, cdf(dat$x0[irow], xs, dat$alpha[irow]),
        lty = dat$lty[irow], col = dat$col[irow])}
legend(x = 45, y = 0.75, legend = c("alpha = 0.1", "alpha = 1", "alpha = 10",
      "x0 = 1", "x0 = 2", "x0 = 5"), col = c("green", "red", "blue", "black",
      "black", "black"), lty = c("solid", "solid", "solid", "solid", "dashed",
      "dotted"), cex = 0.8)
```



From the graphs of the PDFs and CDFs, we see that x_0 gives us the smallest failure time with positive probability. Furthermore, α influences the the rate at which the probabilities decrease as time until failure increases. As α increases, the probability of a specific time until failure decreases.

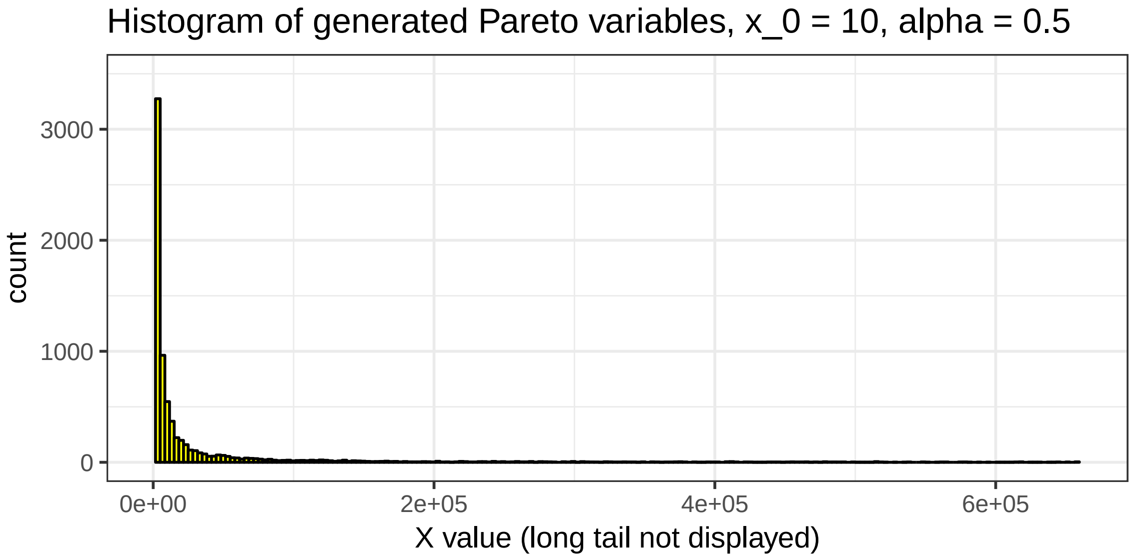
c. Generate Pareto random variables using simulated uniform random variables in R.

Recall from problem 6a that if U is a uniform random number on $[0,1]$, then $X = F^{-1}(U)$ generates a random number X from any continuous distribution with the specified cdf F. Thus, we can generate Pareto random variables by sampling from a Uniform[0,1] distribution, then transforming the result using F^{-1}. For $x \geq x_0$, $F(x) = 1 - \left(\frac{x_0}{x}\right)^{\alpha}$. Setting $F(x) := u$, we can find $F^{-1}(u)$ by solving for x. Thus,

$$u = 1 - \left(\frac{x_0}{x}\right)^{\alpha} \implies 1 - u = \left(\frac{x_0}{x}\right)^{\alpha} \implies (1-u)^{1/\alpha} = \frac{x_0}{x} \implies x = \frac{x_0}{(1-u)^{\frac{1}{\alpha}}}$$

```r
# Simulate uniform random number on [0,1]
set.seed(1)
u <- runif(100000)

# Let alpha = 0.5, x_0 = 10, and transform u using F^{-1} derived above
x_0 <- 10
alpha <- 0.5
x <- x_0/((1 - u)^(1/alpha)) # Transform U
x[x < x_0] <- NA
plt.df <- data.frame(x = x)

# Plot with ggplot2
library(ggplot2)
```

```r
ggplot(plt.df) +
   geom_histogram(aes(x = x),
                  bins = 200, fill = "yellow", color = "black") +
  scale_x_continuous(limits = c(0, 661221.1)) +   ## limit X-axis
  scale_y_continuous(limits = c(0, 3500)) +       ## limit Y-axis
  labs(x = "X value (long tail not displayed)",
       title = "Histogram of generated Pareto variables, x_0 = 10, alpha = 0.5") +
  theme_bw()
```



d. What is the survival function associated with this density? Interpret a value (say $x = 10$ years for $\alpha = 1$ and $x_0 = 2$) evaluated in the survival function in the context of the problem.

Recall that the survival function is given by $S(x) = 1 - F(x)$. Thus, the survival function associated with this density is,

$$S(x) = \begin{cases} \left(\frac{x_0}{x}\right)^\alpha & \text{for} \quad x \geq x_0 \\ 1 & \text{for} \quad x < x_0 \end{cases}$$

For $x = 10$, $\alpha = 1$, and $x_0 = 2$, $S(x) = \left(\frac{2}{10}\right)^1 = 0.2$. That is, with a Pareto distribution with parameters $\alpha = 1$ and $x_0 = 2$, we expect in the long-run for 20% of the parts to survive past 10 years.

e. Find the p^{th} quantile for this density. For $p = .8$ interpret this value in the context of the problem.

In part (c), we showed that the p^{th} quantile is given by, $x = \frac{x_0}{(1-p)^{1/\alpha}}$. Thus, for $\alpha = 1$ and $x_0 = 2$, the the 80th percentile is given by,

$$x = \frac{2}{(1 - 0.8)^{1/1}} = \frac{2}{0.2} = 10$$

indicating that the 80th percentile of survival times for parts is 10 years, which agrees with our interpretation in part (d) of the survival time.

Problem 10

Suppose that a density is of the form cx^k for some constant $k > 1$ and $0 < x < 1$.

a. Find c.

To find c, we can use the fact that a valid density must integrate to 1:

$$1 = \int_0^1 cx^k dx = \frac{c}{k+1} x^{k+1}\Big|_0^1 = \frac{c}{k+1}$$

Thus, $c = k + 1$.

b. Find the cdf.

By the definition of a cdf, we have for $0 < x < 1$,

$$F(x) = P(X \leq x) = \int_0^x (k+1)y^k dy = y^{k+1}\Big|_0^x = x^{k+1}$$

c. Derive a formula for the pth quantile from f.

To get the pth quantile, we want the value of x such that $F(x) = p$. Solving for x,

$$p = x^{k+1} \implies x = p^{1/(k+1)}$$

Thus, the pth quantile is given by $p^{1/(k+1)}$.

d. Let $0 \leq a \leq b \leq 1$. Derive a formula for $P(a < X < b)$.

For $0 \leq a \leq b \leq 1$,

$$P(a < X < b) = \int_a^b (k+1)x^k dx = x^{k+1}\Big|_a^b = b^{k+1} - a^{k+1}$$

Problem 11

Suppose that the time in days until hospital discharge for a certain patient population follows a density $f(x) = c\exp(-x/2.5)$ for $x > 0$

a. What value of c makes this a valid density.

Since the density must integrate to 1, we can determine the value of c by setting the integral of $f(x)$ equal to 1.

$$1 = \int_0^\infty c\exp\left(-\frac{x}{2.5}\right) dx = -2.5c\exp\left(-\frac{x}{2.5}\right)\Big|_0^\infty = 2.5c$$

implying that $c = 1/2.5 = 2/5$.

b. Find the distribution function for this density.

The distribution function for this density is given by,

$$F(x) = P(X \leq x) = \int_0^x \frac{1}{2.5}\exp\left(-\frac{y}{2.5}\right) dy = -\exp\left(-\frac{y}{2.5}\right)\Big|_0^x = -\exp\left(-\frac{x}{2.5}\right) + 1$$

Thus, $F(x) = 1 - \exp(-x/2.5)$.

c. Find the survival function.

The survival function is given by,

$$S(x) = 1 - F(x) = 1 - \left(1 - \exp\left(-\frac{x}{2.5}\right)\right) = \exp\left(-\frac{x}{2.5}\right)$$

d. Calculate the probability that a person takes longer than 11 days to be discharged.

The probability that a person takes longer than 11 days to be discharged is $P(X > 11) = S(11)$. Using our answer from part (c),

$$S(11) = \exp\left(-\frac{11}{2.5}\right) = 0.012$$

e. What is the median number of days until discharge?

The median number of days until discharge is the value of x that satisfies $S(x) = F(x) = 0.5$. Solving for x, we have

$$F(x) = 0.5 \implies 0.5 = 1 - \exp\left(-\frac{x}{2.5}\right) \implies 0.5 = \exp\left(-\frac{x}{2.5}\right) \implies \log 0.5 = -\frac{x}{2.5} \implies x = -2.5 \log 0.5$$

Thus, the median number of days until discharge is given by $x = -2.5log(0.5) = 1.73$.

Problem 12

The (lower) incomplete gamma function is defined as $\Gamma(k, c) = \int_0^c x^{k-1} \exp(-x)dx$. By convention $\Gamma(k, \infty)$, the complete gamma function, is written $\Gamma(k)$. Consider a density

$$\frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x) \quad \text{for} \quad x > 0$$

where α is a known number.

a. Argue that this is a valid density.

To show that the above function is a valid density, we need to show (1) it is non-negative for all x in its domain and (2) the density integrates to 1.

(1) Since $\Gamma(\alpha) \geq 0$, $x > 0$ so $x^{\alpha-1} > 0$ for all values of α, and the exponential function is always positive, the above function will always be non-negative.

(2) Checking the second condition,

$$\int_0^\infty \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x)dx = \frac{1}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} \exp(-x)dx \overset{(*)}{=} \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1$$

where $(*)$ follows directly from the definition of $\Gamma(\cdot)$.

b. Write out the survival function associated with this density using gamma functions.

The survival function is given by $S(x) = 1 - F(x)$, so

$$F(x) = \int_0^x \frac{1}{\Gamma(\alpha)} y^{\alpha-1} \exp(-y)dy = \frac{1}{\Gamma(\alpha)} \int_0^x y^{\alpha-1} \exp(-y)dy = \frac{\Gamma(\alpha, x)}{\Gamma(\alpha)}$$

Consequently, the survival function is given by, $S(x) = 1 - \frac{\Gamma(\alpha,x)}{\Gamma(\alpha)}$.

c. Let β be a known number; argue that

$$\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta) \quad \text{for} \quad x > 0$$

is a valid density. This is known as the **gamma density**.

To show that the above function is a valid density, we need to show (1) it is non-negative for all x in its domain and (2) the density integrates to 1.

(1) In the gamma density, $\beta, \alpha > 0$. Since $\Gamma(\alpha) \geq 0$, $x > 0$ so $x^{\alpha-1} > 0$, and the exponential function is always positive, the above function will always be non-negative.

(2) Checking the second condition,

$$\int_0^\infty \frac{1}{\beta^\alpha\Gamma(\alpha)} x^{\alpha-1}\exp(-x/\beta)dx = \int_0^\infty \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp(-x/\beta)dx$$

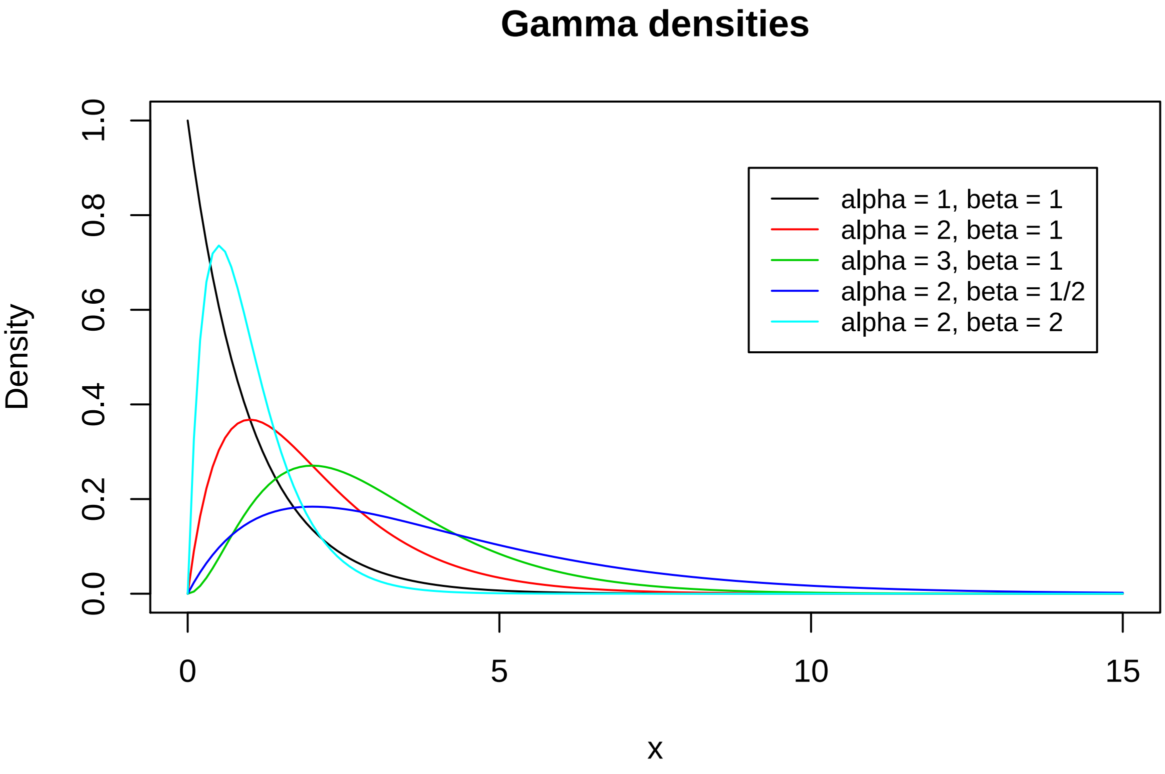
$$\overset{(1)}{=} \int_0^\infty \frac{1}{\Gamma(\alpha)} u^{\alpha-1}\exp(-u)du \overset{(2)}{=} 1$$

where (1) follows from a change in variables using $u = x/\beta$, $du = \frac{1}{\beta}dx$, and (2) follows from noting that the integrand is the density in (a) which you showed was a valid density, so it integrates to 1.

d. Plot the Gamma density for different values of α and β.

We can use the built-in function in R to plot the Gamma density. Specifically, the `dgamma()` function returns the density of Gamma distribution.

```
x = seq(0,15,0.1)
plot(x,dgamma(x,shape=1,scale=1), main = "Gamma densities",
     xlab = "x", ylab = "Density", type = "l")
lines(x,dgamma(x,shape=2,scale=1), col=2)
lines(x,dgamma(x,shape=3,scale=1), col=3)
lines(x,dgamma(x,shape=2,scale=2), col=4)
lines(x,dgamma(x,shape=2,scale=0.5), col=5)
legend(x = 9, y = 0.9, legend = c("alpha = 1, beta = 1", "alpha = 2, beta = 1",
                                   "alpha = 3, beta = 1", "alpha = 2, beta = 1/2",
                                   "alpha = 2, beta = 2"),
       col = c(1, 2, 3, 4, 5), lty = c("solid", "solid", "solid", "solid", "solid"),
       cex = 0.8)
```



Problem 13

The **Weibull density** is useful in survival analysis. Its form is given by

$$\frac{\gamma}{\beta}x^{\gamma-1}\exp\left(-x^{\gamma}/\beta\right)$$

for $x > 0$ and γ and β are fixed known numbers.

a. Demonstrate that the Weibull density is a valid density.

To show that the above function is a valid density, we need to show (1) it is non-negative for all x in its domain and (2) the density integrates to 1.

(1) In the Weibull density, $\gamma, \beta > 0$. Since $\frac{\gamma}{\beta} > 0$, $x > 0$ so $x^{\gamma-1} > 0$, and the exponential function is always positive, the above function will always be non-negative.

(2) Checking the second condition, ✓

$$\underline{\int_0^\infty \frac{\gamma}{\beta}x^{\gamma-1}}\exp(-x^{\gamma}/\beta)dx = \underline{\int_0^\infty \frac{1}{\beta}(\gamma x^{\gamma-1})}\exp(-x^{\gamma}/\beta)dx$$

$$\overset{(*)}{=} \int_0^\infty \frac{1}{\beta}\exp(-u/\beta)du = -\exp(-u/\beta)\Big|_0^\infty = 0-(-1) = 1$$

where $(*)$ follows from a change in variables using $u = x^{\gamma}$, $du = \gamma x^{\gamma-1}dx$.

b. Calculate the survival function associated with the Weibull density.

Using the same change of variables as in part (a) where $u = y^{\gamma}$ and $du = \gamma y^{\gamma-1}dx$, the survival function is given by,

$$\begin{aligned} S(x) = P(X > x) &= \int_x^\infty \frac{\gamma}{\beta}y^{\gamma-1}\exp(-y^{\gamma}/\beta)dy \\ &= \int_x^\infty \frac{1}{\beta}(\gamma y^{\gamma-1})\exp(-y^{\gamma}/\beta)dy \\ &= \int_{x^\gamma}^\infty \frac{1}{\beta}\exp(-u/\beta)du \\ &= -\exp(-u/\beta)\Big|_{x^\gamma}^\infty = 0-(-\exp(-x^{\gamma}/\beta)) = \exp(-x^{\gamma}/\beta) \end{aligned}$$

c. Calculate the median of the Weibull density.

The median of the Weibull density is the value x such that $S(x) = F(x) = 0.5$. From part (b), we have that $S(x) = \exp(-x^{\gamma}/\beta)$ for the Weibull density. Solving for x, we have,

$$S(x) = 0.5 \implies \exp\left(-\frac{x^{\gamma}}{\beta}\right) = 0.5 \implies -\frac{x^{\gamma}}{\beta} = \log 0.5 \implies x = (-\beta\log 0.5)^{1/\gamma}$$

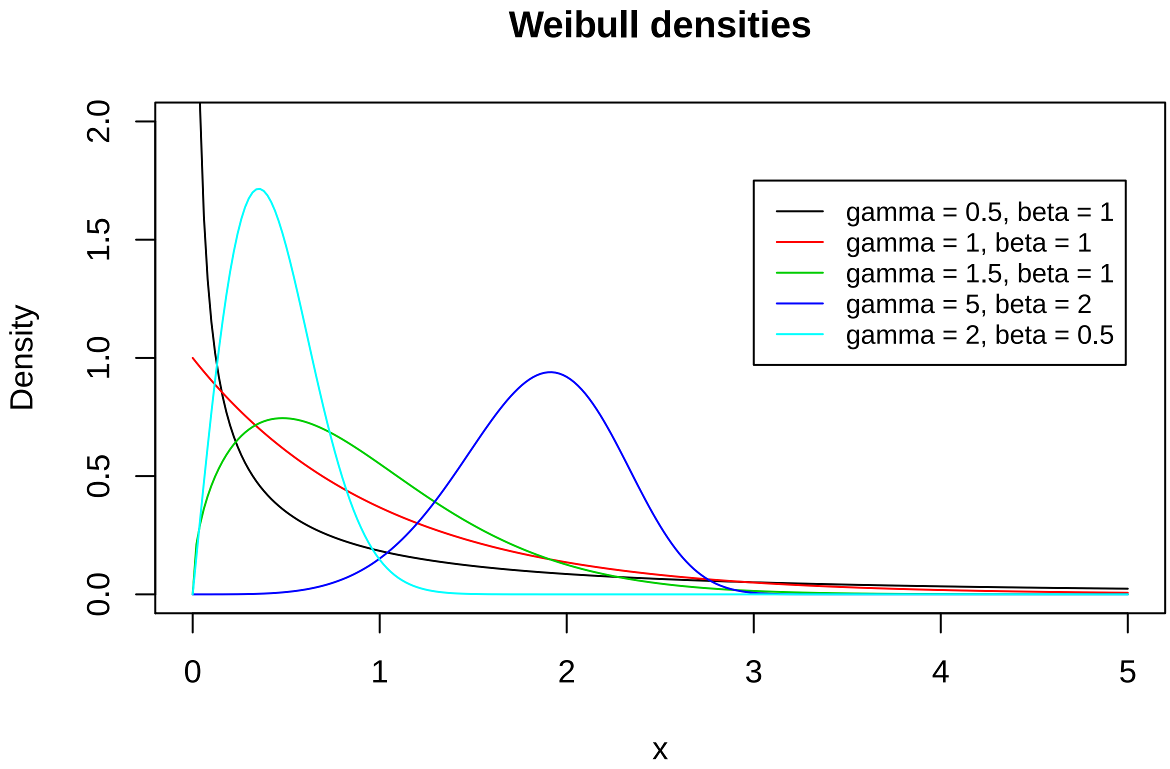
Thus, the median of the Weibull density is given by $x = (-\beta\log 0.5)^{1/\gamma}$.

d. Plot the Weibull density for different values of γ and β.

In R, the `dweibull()` function returns the density of Weibull distribution.

```
x = seq(0,5,0.02)
plot(x,dweibull(x,shape=0.5,scale=1), main = "Weibull densities",
     xlab = "x", ylab = "Density", type = "l",ylim=c(0,2))
lines(x,dweibull(x,shape=1,scale=1), col=2)
lines(x,dweibull(x,shape=1.5,scale=1), col=3)
```

```r
lines(x,dweibull(x,shape=5,scale=2), col=4)
lines(x,dweibull(x,shape=2,scale=0.5), col=5)
legend(x = 3, y = 1.75, legend = c("gamma = 0.5, beta = 1", "gamma = 1, beta = 1",
                                    "gamma = 1.5, beta = 1", "gamma = 5, beta = 2",
                                    "gamma = 2, beta = 0.5"),
       col = c(1, 2, 3, 4, 5), lty = c("solid", "solid", "solid", "solid", "solid"),
       cex = 0.8)
```



Problem 14

The Beta function is given by $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}$ for $\alpha > 0$ and $\beta > 0$. It turns out that

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta).$$

The **Beta density** is given by $\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for fixed $\alpha > 0$ and $\beta > 0$.

a. Argue that the Beta density is a valid density.

To show that the Beta density is a valid density, we need to show (1) it is non-negative for all x in its domain and (2) the density integrates to 1.

(1) In the Beta density, $\alpha, \beta > 0$ and $0 < x < 1$ Consequently, $B(\alpha, \beta) > 0$, $x^{\alpha-1} > 0$, and $(1-x)^{\beta-1} > 0$, so the Beta density will always be non-negative.

(2) Checking the second condition,

$$\int_0^1 \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \overset{(*)}{=} \frac{B(\alpha,\beta)}{B(\alpha,\beta)} = 1$$

where $(*)$ follows from the definition of $B(\alpha, \beta)$.

b. Argue that the uniform density is a special case of the Beta density.

Suppose $\alpha = \beta = 1$. Then,

$$B(1,1) = \int_0^1 x^{1-1}(1-x)^{1-1}dx = \int_0^1 1dx = 1$$

and the Beta(1,1) density is given by,

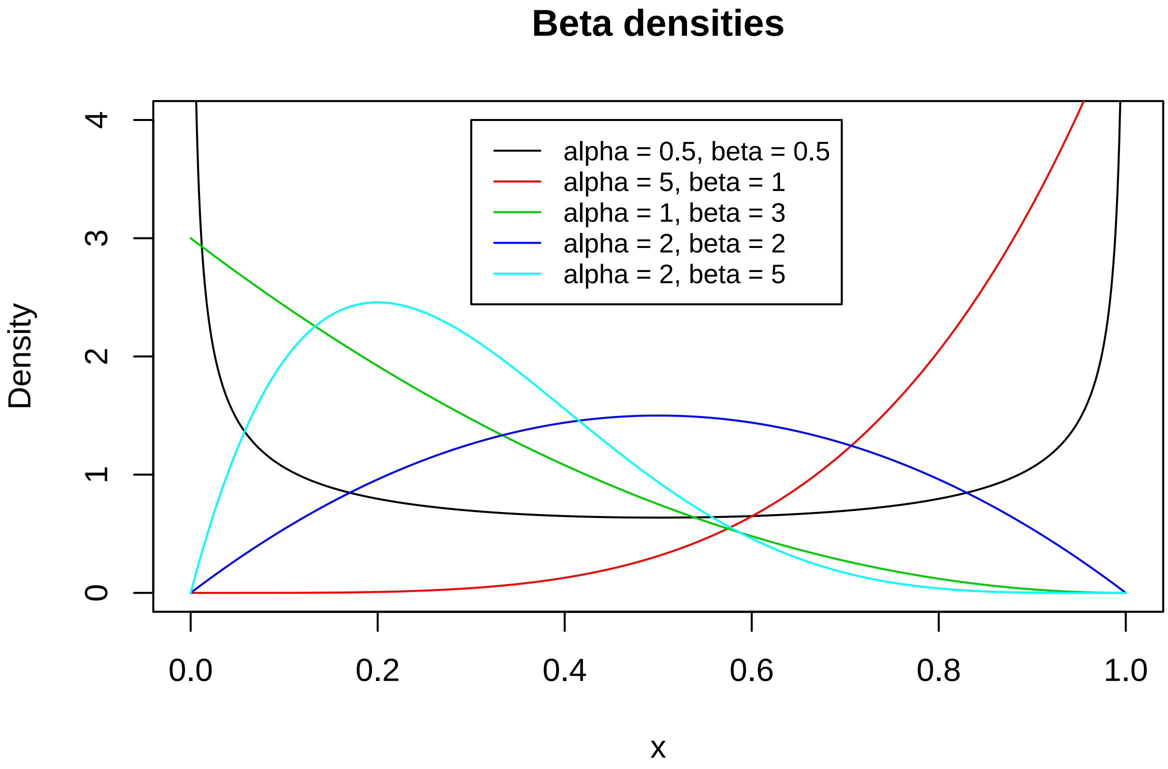
$$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} = \frac{1}{1}x^{1-1}(1-x)^{1-1} = 1$$

which is equivalent to the density of the uniform distribution.

c. Plot the Beta density for different values of α and β.

In `R`, the `dbeta()` function returns the density of Beta distribution.

```
x = seq(0,1,0.001)
plot(x,dbeta(x,shape1=0.5,shape2=0.5), main = "Beta densities",
     xlab = "x", ylab = "Density", type = "l",ylim=c(0,4))
lines(x,dbeta(x,shape1=5,shape2=1), col=2)
lines(x,dbeta(x,shape1=1,shape2=3), col=3)
lines(x,dbeta(x,shape1=2,shape2=2), col=4)
lines(x,dbeta(x,shape1=2,shape2=5), col=5)
legend(x = 0.3, y = 4.0, legend = c("alpha = 0.5, beta = 0.5", "alpha = 5, beta = 1",
                                   "alpha = 1, beta = 3", "alpha = 2, beta = 2",
                                   "alpha = 2, beta = 5"),
       col = c(1, 2, 3, 4, 5), lty = c("solid", "solid", "solid", "solid", "solid"),
       cex = 0.8)
```



Problem 15

A famous formula is $e^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$ for any value of λ. Assume that the count of the number of people infected with a particular disease per year follows a mass function given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad x = 0, 1, 2, 3, \ldots$$

where λ is a fixed known number. (This is know as the **Poisson mass function**.)

a. Argue that $\sum_{x=0}^{\infty} P(X = x) = 1$.

Recall the Taylor expansion for e^x is $\sum_{n=0}^{\infty} \frac{x^n}{n!}$. Consequently,

$$\sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \overset{(*)}{=} e^{-\lambda}e^{\lambda} = e^0 = 1$$

where (*) follows directly from the Taylor expansion of e^λ.

Problem 16

Consider counting the number of coin flips from an unfair coin with success probability p until a head is obtained, say X. The mass function for this process is given by $P(X = x) = p(1-p)^{x-1}$ for $x = 1, 2, 3, \ldots$. This is called the **geometric mass function**.

a. Argue mathematically that this is a valid probability mass function. Hint, the geometric series is given by $\frac{1}{1-r} = \sum_{k=0}^{\infty} r^k$ for $|r| < 1$.

A valid probability mass function satisfies (1) for all x, $P(X = x) \geq 0$ and (2) $\sum_{x=1}^{\infty} P(X = x) = 1$.

(1) Since $0 < p < 1$, $p(1-p)^{x-1} > 0$ for $x = 1, 2, 3, \cdots$.

(2) Verifying the second condition,

$$\sum_{x=1}^{\infty} P(X = x) = \sum_{x=1}^{\infty} p(1-p)^{x-1} = p \sum_{x=1}^{\infty} (1-p)^{x-1} \overset{(1)}{=} p \sum_{k=0}^{\infty} (1-p)^k \overset{(2)}{=} p \frac{1}{1-(1-p)} = \frac{p}{p} = 1$$

where (1) follows from changing the index of summation (by letting $k = x - 1$) and (2) follows from the hint.

b. Calculate the survival distribution $P(X > x)$ for the geometric distribution for integer values of x.

First, recall that a finite geometric sum is given by,

$$\sum_{k=0}^{n} r^k = \frac{1 - r^{n+1}}{1 - r}$$

$$1 - P(X \leq x)$$
$$= 1 - \sum_{k=1}^{x} p(1-p)^{k-1}$$
$$= 1 - p \frac{1(1-(1-p)^x)}{1-(1-p)}$$
$$= 1 - (1-(1-p)^x)$$
$$= (1-p)^x$$

Let $S(x) = P(X > x) = 1 - P(X \leq x)$ denote the survival distribution. Then,

$$S(x) = 1 - P(X \leq x) = 1 - \sum_{k=1}^{x} P(X = x) = 1 - \sum_{k=1}^{x} p(1-p)^{k-1} = 1 - p \sum_{k=1}^{x} (1-p)^{k-1}$$
$$\overset{(1)}{=} 1 - p \sum_{n=0}^{x-1} (1-p)^n \overset{(2)}{=} 1 - p \frac{1 - (1-p)^x}{1 - (1-p)} = (1-p)^x$$

where (1) follows from changing the index of summation (by letting $n = k - 1$) and (2) follows from the formula for a finite geometric sum. Thus, $S(x) = (1-p)^x$.