BST 140.651
Problem Set 1

Problem 1. Show the following

a. $P(\emptyset) = 0$.
b. $P(E) = 1 - P(E^c)$.
c. If $A \subset B$ then $P(A) \leq P(B)$.
d. For any A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
e. $P(A \cup B) = 1 - P(A^c \cap B^c)$.
f. $P(A \cap B^c) = P(A) - P(A \cap B)$.
g. $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$.
h. $P(\cup_{i=1}^n E_i) \geq \max_i P(E_i)$.

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?
b. What's the probability that either the mother or the father has contracted cryptosporidiosis?
c. What's the probability that the mother has contracted cryptosporidiosis but the father has not?
d. What's the probability that the mother has contracted cryptosporidiosis?
e. What's the probability that neither the mother nor the father has contracted cryptosporidiosis?
f. What's the probability that the mother has contracted cryptosporidiosis but the father has not?

Problem 3. Suppose $h(x)$ is such that $h(x) > 0$ for $x = 1, 2, \ldots, I$. Argue that $p(x) = h(x)/\sum_{i=1}^I h(i)$ is a valid pmf.

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.

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?

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.

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

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

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

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

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

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.

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

c. Simulate 100 Normal iid variables using only simulations from a uniform distribution and the normal cdf in R

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.

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

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

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.

b. Calculate the distribution function associated with this density.

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

d. Define the *odds* 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.

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 $\mathrm{Pa}(x_0, \alpha)$.

a. Derive the density of the Pareto distribution.

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

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

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.

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

Problem 10. Suppose that a density is of the form cx^k for some constant $k > 1$ and $0 < x < 1$.

a. Find c.

b. Find the cdf.

c. Derive a formula for the p^{th} quantile from f.

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

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?

b. Find the distribution function for this density.

c. Find the survival function.

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

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

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.

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

c. Let β be a known number; argue that

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

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

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

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.

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

c. Calculate the median of the Weibull density.

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

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 $\beta > 0$. This density is useful for

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

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

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

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

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

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

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

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 $|r| < 1$.

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