Review Session 2 – Random Variables and Transformations

References/suggested reading

(i) Casella & Berger's *Statistical Inference*, chapters 1–3.

1 Probability

1.1 Axiomatic Foundations

Probability theory starts with a very important notion and that is the notion of sample space.

Definition 1.1. The *sample space* of a random experiment is the set of all possible outcomes of the experiment.

Example 1.2

What is the sample space of a coin tossing experiment? Since the result will be either a heads or tails, the sample space will have two elements:

$$\mathcal{S} = \{\text{Heads}, \text{Tails}\}.$$

Definition 1.3. An *event* A is a subset of the sample space $\mathcal{S}$. After performing an experiment we say event A has occurred if the outcome of the experiment is in A. For instance, in the above example, we might take $A = \{\text{Heads}\}$.

Now, we can define a probability function:

Definition 1.4 (probability function)**.** Given a sample space S, a *probability function* is a function $\mathbb{P}$ on a space of events that satisfies

1. $\mathbb{P}(A) \geq 0$ for all events A.
2. $\mathbb{P}(\mathcal{S}) = 1$.
3. If $A_1, A_2, \ldots$ are pairwise disjoint events, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

These are also called the "axioms of probability" (or the Kolmogorov axioms). Any function $\mathbb{P}$ that satisfies the axioms of probability is called a probability function. Note here that we are being vague about the "space of events" on which a probability function is defined. We run into some tricky mathematical trouble if we try to define a probability function on the space of *all* events, so we typically consider a reasonable space of events which we care about: a so-called sigma algebra or sigma field. However, we won't dwell on this issue here, as it won't matter for most statistical purposes. You will learn about a more careful measure-theoretic formulation of probability functions in your probability theory class.

1.2 The Calculus of Probabilities

Using the axioms, we can quickly build many properties of the probability function which will be helpful in calculating more complicated probabilities. Try to prove these properties using Kolmogorov's axioms.

Theorem 1.5

If $\mathbb{P}$ is a probability function and A, B are events, then

1. $\mathbb{P}(\emptyset) = 0$ where $\emptyset$ is the empty set.
2. $\mathbb{P}(A) \leq 1$.
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
4. $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
6. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
7. $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$ (*Bonferroni's Inequality*).
8. $\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap C_i)$ for any partition $C_1, C_2, \ldots$ of $\mathcal{S}$.
9. $\mathbb{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for any sets $A_1, A_2, \ldots$ (*Boole's Inequality or union-bound*).

1.3 Conditional Probability and Independence

In many instances, we are in a position to update the sample space based on new information. This leads to the notion of a *conditional probability*.

Definition 1.6 (conditional probability)**.** If A and B are events in S, and $\mathbb{P}(B) > 0$, then the *conditional probability* of A given B, written $\mathbb{P}(A|B)$, is

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Essentially, B becomes the updated sample space so that "conditional on observing B", we can calibrate the probabilities of all further events with respect to their relation to B.

One useful way to write the conditional probability is $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$. Then, from symmetry, we immediately notice that $\mathbb{P}(A|B) = \mathbb{P}(B|A)\frac{\mathbb{P}(A)}{\mathbb{P}(B)}$, which gives us a formula for "turning around" conditional probabilities. This is often called Bayes' rule, which takes the following general form:

Theorem 1.7 (Bayes' Rule)

Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots$,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

In many important applications, it may happen that the occurrence of a particular event B has no effect on the probability of another event A. Symbolically, we are saying that $\mathbb{P}(A|B) = \mathbb{P}(A)$. If this holds, then by Bayes' rule:

$$\mathbb{P}(B|A) = \mathbb{P}(A|B)\frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B),$$

so the occurrence of A has no effect on B. Moreover, since $\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A \cap B)$, it then follows that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. This leads us to the definition of *statistical independence*.

Definition 1.8 (independence)**.** Two events, A and B, are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

To define independence for more than two events, we have to be careful. We might intuitively think that for three events A, B, C, the triple A, B, C are independent if $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. Or, we might think A, B, C are independent if any pair of them are independent. However, neither of these proposed conditions necessary implies the other (i.e., we can find counterexamples). Thus, we require a stronger definition which captures both conditions. For a general collection of events $A_1, \ldots, A_n$, another guiding intuition here is that if $A_1, \ldots, A_n$ are independent, then any subset of them should also be independent. This leads us to the notion of *mutual independence*, which captures both of the earlier intuitions:

Definition 1.9. A collection of events $A_1, \ldots, A_n$ are *mutually independent* if for any subcollection $A_{i_1}, \ldots, A_{i_k}$, we have

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

1.4 Random Variables

In many experiments, it is easier to deal with a summary variable than with the original probability structure. For example, in an opinion poll, we might decide to ask 50 people whether they agree or disagree with a certain issue. If we record a "1" for agree and "0" for disagree, the sample space for this experiment has 2^{50} elements, which is very large! It may be that the only quantity of interest is the number of people who agree out of 50 and, if we define a variable X as the number of 1's recorded out of 50, then we have captured the essence of the problem while constraining our sample space to the set of integers $\{0, 1, \ldots, 50\}$, which is much easier to deal with.

This leads us to the notion of a *random variable*, which we define simply as a mapping (or function) from the original sample space to a new sample space (which we assume is a set of real numbers).

Example 1.10
If we toss two dice, then the original sample space has size 36 where if X is the sum of the two numbers, then X has range $\{2, 3, \ldots, 12\}$.

Suppose we have a sample space $\mathcal{S}$ with a probability function $\mathbb{P}$ and we define a random variable X with range $\mathcal{X}$. Then, we can define a probability function P_X on $\mathcal{X}$ in the following way. We will observe $X = x$ if and only if the outcome of the random experiment is an $s \in \mathcal{S}$ such that $X(s) = x$. Thus,

$$P_X(X = x) = \mathbb{P}(\{s \in S : X(s) = x\}).$$

More generally, for a set $A \subset \mathcal{X}$:

$$P_X(X \in A) = \mathbb{P}(\{s \in S : X(s) \in A\}).$$

We can confirm this gives a valid probability function P_X on $\mathcal{X}$ for which the Kolmogorov axioms hold.

1.5 Distribution Functions

Definition 1.11 (cdf)**.** The *cumulative distribution function* or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) := P_X(X \leq x), \text{ for all } x.$$

Example 1.12 (tossing three coins)
Consider the experiment of tossing three fair coins, and let X be the number of heads observed. The cdf of X is

$$F_X(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1/8 & 0 \leq x < 1 \\ 1/2 & 1 \leq x < 2 \\ 7/8 & 2 \leq x < 3 \\ 1 & 3 \leq x < \infty \end{cases}.$$

As we can see from the example above, F_X can be discontinuous, with jump discontinuities at certain values of x. By the way in which F_X is defined, however, at the jump points F_X takes the value at the top of the jump. Thus, $F_X(\cdot)$ is right-continuous. Along these lines, here are some properties of a cdf which also characterize it:

Theorem 1.13

The function $F(x)$ is a cdf iff the following hold

(1) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

(2) $F(x)$ is a nondecreasing function of x.

(3) $F(x)$ is right-continuous, i.e. $\forall x_0 \in \mathbb{R}$: $\lim_{x\downarrow x_0} F(x) = F(x_0)$.

The above theorem is a strong characterization of a cdf and actually states that not only does every cdf satisfy the (1)–(3) but that any function $F : \mathcal{X} \to \mathbb{R}$ satisfying (1)–(3) is the cdf of some random variable X defined on some sample space $\mathcal{S}$.

Definition 1.14 (continuous vs. discrete random variable)**.** A random variable X is *continuous* if $F_X(x)$ is a continuous function of x. A random variable X is dicrete if $F_X(x)$ is a step function of x with finite support.

In fact, the cdf F_X completely determines the probability distribution of a random variable X. What we mean here is that $\mathbb{P}(X \in A)$ is determined for each event A.

Definition 1.15. The random variables X and Y are *identically distributed* if, for every event A, $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$.

Theorem 1.16

Random variables X and Y are identically distributed iff $F_X(x) = F_Y(x)$ for every $x \in \mathbb{R}$.

1.6 Density and Mass Functions

Associated with a random variable X and its cdf F_X is another function, called either the probability density function (pdf) or probability mass function (pmf). These are concerned with "point probabilities" of random variables. For a discrete random variable, the pmf is literally a point probability, whereas for a continuous random variable, the pdf quantifies the relative probability of regions near a point.

Definition 1.17 (pmf/pdf)**.** The *probability mass function (pmf)* of a discrete random variable is given by $f_X(x) := \mathbb{P}(X = x)$ for all $x \in \mathbb{R}$. The *probability density function (pdf)* of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \iff \frac{d}{dx}F_X(x) = f_X(x), \forall x \in \mathbb{R}.$$

Like the cdf, there are intuitive conditions which characterize a function $f : \mathbb{R} \to \mathbb{R}$ as being a pdf/pmf:

Theorem 1.18

A function $f(x)$ is a pdf (or pmf) of some random variable X iff

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $\sum_x f(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f(x)\,dx = 1$ (pdf).

Note that a pdf $f_X(x)$ may not always exist as, expectedly, the cdf $F_X(x)$ may not even be differentiable (even though it is right-continuous). However, most random variables you will encounter will have an associated density or mass function.

2 Transformations and Expectations

Now, we will turn our attention to the behavior of functions of a random variable X. In particular, we will describe the cdf and pdf/pmf of a transformation of a random variable, and also discuss moments and the mgf of a random variable and its transformations.

2.1 Distributions of Functions of a Random Variable

If X is a random variable with cdf $F_X(x)$, then any function of X, say $Y := g(X)$, is also a random variable. We then have for event A:

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(X \in g^{-1}(A)),$$

where $g^{-1}(A) := \{x \in \mathcal{X} : g(x) \in A\}$ is the pre-image of A (which always exists). So, we see right away that the distribution of Y involves the preimages of the transformation g.

Let's work through a (somewhat tricky) example.

Example 2.1 (uniform transformation)
Suppose X has a uniform distribution on the interval $(0, 2\pi)$, that is

$$f_X(x) = \begin{cases} 1/(2\pi) & 0 < x < 2\pi \\ 0 & \text{otherwise} \end{cases}$$

Consider the transformation $Y := \sin^2(X)$. Then, in the domain $(0, 2\pi)$, for $y \in (0, 1)$, $y = \sin^2(x)$ has four solutions in x: x_1, x_2, x_3, x_4. Drawing the graph, we see that

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq x_1) + \mathbb{P}(x_2 \leq X \leq x_3) + \mathbb{P}(X \geq x_4).$$

From the symmetry of the function $\sin^2(x)$ and the fact that X has a uniform distribution we have that

$$\mathbb{P}(X \leq x_1) = \mathbb{P}(X \geq x_4) \text{ and } \mathbb{P}(x_2 \leq X \leq x_3) = 2 \cdot \mathbb{P}(x_2 \leq X \leq \pi).$$

Thus, we can write

$$\mathbb{P}(Y \leq y) = 2\mathbb{P}(X \leq x_1) + 2\mathbb{P}(x_2 \leq X \leq \pi)$$

where x_1, x_2 are the two solutions to $\sin^2(x) = y$ for $0 < x < \pi$.

Next, we turn our attention to *monotone* transformations g as these will be easier to work with and give a nice transformation of cdf's and pdf/pmf's. In fact, let's suppose g is strictly increasing (the case of g strictly decreasing will be similar). Using our equation $\mathbb{P}(Y \in A) = \mathbb{P}(X \in g^{-1}(A))$, we can write the cdf of Y as

$$F_Y(y) = \int_{x \in \mathcal{X} : x \leq g^{-1}(y)} f_X(x)\, dx = \int_{-\infty}^{g^{-1}(y)} f_X(x)\, dx = F_X(g^{-1}(y)).$$

This gives us a way of relating the cdf of Y to the cdf of X and the transformation g taking X to Y.

Of course, with transformations, we have to be careful about using the right sample spaces.

Theorem 2.2

Let X have cdf $F_X(x)$, let $Y = g(X)$, and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as

$$\mathcal{X} = \{x : f_X(x) > 0\}, \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}$$

1. If g is strictly increasing on $\mathcal{X}$, $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.
2. If g is strictly decreasing on $\mathcal{X}$, and X is a continuous random variable, $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.

What if g is only weakly increasing or weakly decreasing? Then, $g^{-1}(y)$ may take on multiple values and the formulas above only hold for the largest or smallest such value, depending on whether g is weakly increasing or weakly decreasing.

Example 2.3 (uniform-exponential relationship)

Suppose $X \sim f_X(x) = 1$ if $0 < x < 1$ and 0 otherwise, or the uniform$(0, 1)$ distribution. Consider the transformation $Y = g(X) = -\log X$. Since $-\log(\cdot)$ is a decreasing function, as X ranges from 0 to 1, $-\log X$ ranges from 0 to ∞, meaning $\mathcal{Y} = (0, \infty)$. We also have $g^{-1}(y) = e^{-y}$. Thus, $F_Y(y) = 1 - e^{-y}$.

If the pdf of Y is continuous, it can be obtained by differentiating the cdf. Theorem 2.2 and the chain rule then give us the familiar "inverse derivative/Jacobian" factor. Of course, now we must also assume g is differentiable.

Theorem 2.4 (pdf transformation law for monotone transformations)

Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined as in Theorem 2.2. Suppose that $f_X(x)$ is continuous on X and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then, the pdf of Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

However, we are being somewhat restrictive as we keep assuming that g is monotone. Let's demonstrate why this might be troublesome for non-monotone g:

Example 2.5 (square transformation)

Suppose X is a continuous random variable. For $y > 0$, the cdf of $Y = X^2$ is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y})$$

By continuity, we have this is equal to

$$F_Y(y) = \mathbb{P}(-\sqrt{y} < X \leq \sqrt{y}) = \mathbb{P}(X \leq \sqrt{y}) - \mathbb{P}(X \leq -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

whence differentiating and using chain rules gives:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}).$$

Note that the transformed pdf $f_Y(y)$ in this example almost obeys the prescribed equation of Theorem 2.4. The only difference is that there seem to be two versions of the transformation law added together. In fact, these correspond to the two sub-regions of the domain where the function $g(x) = x^2$ is monotone. We can determine the pdf of a general transformation $g : \mathcal{X} \to \mathcal{Y}$ by focusing on the local regions where it is monotone and patching these pieces together. Formally, we have:

Theorem 2.6 (general pdf transformation law)

Let X have pdf $f_X(x)$ and let $Y = g(X)$, and define the sample space $\mathcal{X}$ as in Theorem 2.2. Suppose there exists a partition $A_0, A_1, \ldots, A_k$ of $\mathcal{X}$ such that $\mathbb{P}(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i. Further, suppose there exist functions $g_1(x), \ldots, g_k(x)$, defined on $A_1, \ldots, A_k$, respectively, satisfying

(i) $g(x) = g_i(x)$, for $x \in A_i$.

(ii) $g_i(x)$ is monotone on A_i.

(iii) The set $\mathcal{Y} = \{y : y = g_i(x) \text{ for some } x \in A_i\}$ is the same for each $i = 1, \ldots, k$.

(iv) $g_i^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, for each $i = 1, \ldots, k$.

Then, the pdf of Y is given by

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases}$$

In this theorem, the "exceptional set" A_0 is essentially ignorable as it's just there to handle the boundary points between the sets A_i (which you might typically think of as intervals of $\mathcal{X} \subseteq \mathbb{R}$). Let's apply this to the square transformation again:

Example 2.7 (normal/chi-square relationship)

Let X have the *standard normal distribution*,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty$$

Consider $Y = X^2$. We have that the function $g(x) = x^2$ is monotone on $(-\infty, 0)$ and on $(0, \infty)$. Thus, we have $\mathcal{Y} = (0, \infty)$ and applying the previous theorem, we have the pdf of Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2} \left| -\frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{y}} e^{-y/2}, 0 < y < \infty$$

The pdf of Y here is that of a *chi squared random variable* with 1 degree of freedom.

Another very useful transformation is the probability integral transformation which involves letting g be the cdf of X itself.

Theorem 2.8 (probability integral transformation)

Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$, that is $\mathbb{P}(Y \leq y) = y$ for $0 < y < 1$.

Proof. We note that if F_X is strictly increasing, then its inverse F_X^{-1} is well-defined. If F_X is only nondecreasing, we can still define a "pseudo-inverse":

$$F_X^{-1}(y) := \inf\{x : F_X(x) \geq y\},$$

which is always well-defined. Moreover, F_X^{-1} is monotonic as well. So, we get:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F_X(X) \leq y) = \mathbb{P}(F_X^{-1}(F_X(x)) \leq F_X^{-1}(y)) = \mathbb{P}(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

Note that in the above equalities, the step $\mathbb{P}(F_X^{-1}(F_X(x)) \leq F_X^{-1}(y)) = \mathbb{P}(X \leq F_X^{-1}(y))$ is obvious if F_X is strictly increasing and thus has an inverse. If we are using the pseudo-inverse F_X^{-1}, it has to be more carefully argued. Do you see how? Hint: the events are actually not the same sets, but their probabilities are the same. ■

2.2 Expected Values and Moment Generating Functions

Definition 2.9 (expected value)**.** The *expected value or mean or expectation* of a random variable $g(X)$, denoted $\mathbb{E}[g(X)]$ is

$$\mathbb{E}[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x)\, dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) & \text{if } X \text{ is discrete.} \end{cases}$$

Of course, in the above, $f_X(x)$ is either a pdf or a pmf depending on whether X is continuous or discrete.

The above is also known as the "law of the unconscious statistician" because it is typically cited as a definition and not a theorem to be proved. In fact, even Casella & Berger and I have written it this way! What can go wrong with the definition? We know $Y := g(X)$ is another random variable with its own pdf/pmf $f_Y(y)$. The above definition then gives us two formulas (in the case of continuous X; the discrete case is similar):

$$\mathbb{E}[g(X)] = \mathbb{E}[Y] = \int_{-\infty}^{\infty} y \cdot f_Y(y)\, dy$$
$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

In order for the definition of expected value to make sense, these two formulas should be the same. If the transformation g here is differentiable and strictly monotonic, then we can do a change of variable in one integral and arrive at the other, thus concluding they are the same. However, unlike for the pdf/cdf transformation formulas in the previous section, it turns out that the law of the unconscious statistician holds in a much broader setting where g does not have to be differentiable or monotonic: instead, it is only necessary that it be measurable.

Note that the expectation may not always exist in that the associated integral/sum can be infinite or, even worse, undefined but not necessarily ∞ or $-\infty$. The latter situation can happen if $g(x)$ is a positive number for sufficiently large positive x and a negative number for sufficiently small negative x. When working with expectations, we will generally avoid these scenarios by assuming the expectations exist and are finite.

Most of the properties of the expectation follow from the familiar properties of the integral, i.e. linearity and monotonicity:

Theorem 2.10

Let X be a random variable and let $a, b, c \in \mathbb{R}$ be constants. then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist:

1. $\mathbb{E}[ag_1(X) + bg_2(X) + c] = a\mathbb{E}[g_1(X)] + b\mathbb{E}[g_2(X)] + c.$
2. $\forall x \in \mathbb{R} : g_1(x) \geq 0 \implies \mathbb{E}[g_1(X)] \geq 0.$
3. $\forall x \in \mathbb{R} : g_1(x) \geq g_2(x) \implies \mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)].$
4. $\forall x \in \mathbb{R} : a \leq g_1(x) \leq b \implies a \leq \mathbb{E}[g_1(X)] \leq b.$

Example 2.11

The expected value of a random variable has another useful property, one that we can think of as relating to the intuition that $\mathbb{E}[X]$ is a good deterministic guess at a value of X. Suppose we measure the distance between a random variable X and a constant $b \in \mathbb{R}$ by $(X-b)^2$. Then, the $b \in \mathbb{R}$ which minimizes $\mathbb{E}[(X-b)^2]$ is $b = \mathbb{E}[X]$, and hence $\mathbb{E}[X]$ is a good predictor of X. In other words, the expectation minimizes the square error or:

$$\min_b \mathbb{E}[(X-b)^2] = \mathbb{E}[(X-\mathbb{E}[X])^2].$$

How can we prove this? The standard trick is to add and subtract the term we want:

$$\begin{aligned}\mathbb{E}[(X-b)^2] &= \mathbb{E}[(X-\mathbb{E}[X]+\mathbb{E}[X]-b)^2] \\ &= \mathbb{E}[(X-\mathbb{E}[X]) + (\mathbb{E}[X]-b))^2 \\ &= \mathbb{E}[(X-\mathbb{E}[X])^2] + (\mathbb{E}[X]-b)^2 + 2\mathbb{E}[(X-\mathbb{E}[X])(\mathbb{E}[X]-b)].\end{aligned}$$

Note that the cross-term (the last term on the RHS above) vanishes by linearity of expectation for any value of b:

$$(\mathbb{E}[X]-b)^2 + 2\mathbb{E}[(X-\mathbb{E}[X])(\mathbb{E}[X]-b)] = (\mathbb{E}[X]-b)\mathbb{E}[X-\mathbb{E}[X]] = 0.$$

Thus, we are left with

$$\mathbb{E}[(X-b)^2] = \mathbb{E}[(X-\mathbb{E}[X])^2] + (\mathbb{E}[X]-b)^2.$$

However, the RHS is minimized at $b = \mathbb{E}[X]$.

Remark 2.12. This type of trick (adding and subtracting a term and then getting a sum of two squares) appears often in statistics proofs.

The expected value isn't the only number which encodes information about a random variable. The more general *moments* of a random variable also play this role:

Definition 2.13 (moment)**.** For each positive integer $n \in \mathbb{N}$, the n-th *moment* of X is $\mathbb{E}[X^n]$. The n-th *central moment* of X is $\mathbb{E}[(X-\mu)^n]$ where $\mu = \mathbb{E}[X]$.

Aside from the mean, $\mathbb{E}[X]$, of a random variable, perhaps the most important moment is the second central moment, more commonly known as the variance.

Definition 2.14 (variance)**.** The *variance* of a random variable X is its second central moment,

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

The positive square root of $\mathrm{Var}(X)$ is the *standard deviation* of X.

The variance and standard deviation both give a measure of the degree of spread of a distribution around its mean. It also behaves nicely with linearity:

$$\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X).$$

A natural question is then whether the moments characterize a distribution. The answer in general is no. Moreover, the moments can quickly become difficult to compute and so this is not a useful way to describe a distribution. So, we turn to the *moment generating function (mgf)* instead, which gives us a concise way of encoding all the moments at once. It will turn out that the mgf characterizes a distribution.

Definition 2.15 (mgf)**.** Let X be a random variable with cdf F_X. The *moment generating function (mgf)* of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) := \mathbb{E}[e^{tX}]$$

provided that the expectation exists for t in some neighborhood of 0. That is, there exists $h > 0$ such that, for all $t \in (-h, h)$, $\mathbb{E}[e^{tX}]$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

Why do we only care about the mgf $M_X(t)$ existing near $t = 0$? We shall see that plugging in $t = 0$ will give us a quick formula for the higher moments of X. First, we can series-expand the exponential e^{tX} and then use linearity of expectation:

$$\mathbb{E}[e^{tX}] = 1 + t\mathbb{E}[X] + \frac{t^2\mathbb{E}[X^2]}{2!} + \frac{t^3\mathbb{E}[X^3]}{3!} + \cdots.$$

In particular, we see that the coefficients of this expansion are the moments of X. But, this is also the Taylor series expansion of $M_X(t)$ about $t = 0$. So, we conclude that the moments of X can be obtained through differentiating $M_X(t)$ with respect to t.

Theorem 2.16

If X has mgf $M_X(t)$, then the moments of X are given via the derivatives of the mgf evaluated at 0:

$$\mathbb{E}[X^n] = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}$$

Example 2.17 (gamma mgf derivation)

Consider the gamma pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, 0 < x < \infty, \alpha > 0, \beta > 0$$

where $\Gamma(\alpha)$ denotes the gamma function evaluated at α. The mgf is then

$$M_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty e^{tx} x^{\alpha-1} e^{-x/\beta}\, dx = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x/\left(\frac{\beta}{1-\beta t}\right)}\, dx$$

Observe the integrand above is the kernel (i.e., the part disregarding normalizing constants) of another gamma pdf. Thus, the integral, provided it exists, evaluates to $\Gamma(\alpha)\left(\frac{\beta}{1-\beta t}\right)^\alpha$ (using the fact that the integral of the gamma pdf is 1). Thus,

$$M_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\Gamma(\alpha)\left(\frac{\beta}{1-\beta t}\right)^\alpha = \left(\frac{1}{1-\beta t}\right)^\alpha, \text{ if } t < 1/\beta$$

If $t \geq 1/\beta$, then the integral diverges meaning the mgf of the gamma distribution exists only if $t < 1/\beta$. As a further application, the mean of the gamma can now be found by

$$\mathbb{E}[X] = \frac{d}{dt} M_X(t)|_{t=0} = \frac{\alpha\beta}{(1-\beta t)^{\alpha+1}}\bigg|_{t=0} = \alpha\beta$$

However, the major usefulness of the mgf is not its ability to generate moments, but rather its ability to characterize a distribution. From Theorem 2.16, we know that if two distributions have the same mgf (in an interval about $t = 0$), then they will the same moments. Now, we turn back to the earlier question of whether a set of moments characterizes a distribution. Unfortunately, the answer in general is no. Let's work through a counterexample.

Example 2.18

Consider the two pdfs given by

$$f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}, 0 \leq x < \infty$$
$$f_2(x) = f_1(x)(1 + \sin(2\pi \log(x)), 0 \leq x < \infty.$$

Note that f_1 is a special case of a lognormal pdf or that $X_1 \sim f_1$ satisfies $\log(X_1) \sim \mathcal{N}(0,1)$. By using the formula for the mgf of a normal distribution, we have a general formula for the n-th moment:

$$\mathbb{E}[X_1^n] = e^{n^2/2}.$$

Meanwhile, the moments of $X_2 \sim f_2$ are

$$\mathbb{E}[X_2^n] = \int_0^\infty x^n f_1(x)(1 + \sin(2\pi \log(x))\, dx = \mathbb{E}[X_1^n] + \int_0^\infty x^r f_1(x) \sin(2\pi \log(x))\, dx.$$

Now, making the transformation $y = \log(x) - r$ shows that this last integral is that of an odd function over $(-\infty, \infty)$ and hence vanishes for all $n \in \mathbb{N}$. Thus, X_1, X_2 have distinct pdfs, but the same moments of all orders.

The problem of uniqueness of moments does not occur if the random variables have bounded support. In other words, distributions with bounded support are characterized by their moments. On the other hand, the mgf always characterizes the distribution, if it exists.

Theorem 2.19

Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.

1. If X, Y have bounded support, then $F_X(u) = F_Y(u)$ for all u iff $\mathbb{E}[X^n] = \mathbb{E}[Y^n]$ for all integers $n = 0, 1, 2, \ldots$ (i.e., moments characterize distributions with bounded support).
2. If the mgfs exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.

Since the mgf's characterize a distribution, they can also characterize the limit of a convergent sequence of distributions. We will study the asymptotics of random variables in much more detail in a future review session, but for now by "convergent sequence of distributions", we just mean convergence of the cdf's:

Theorem 2.20 (Lévy continuity theorem for mgf's)

Suppose $\{X_i\}_{i=1}^\infty$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that $\lim_{i\to\infty} M_{X_i}(t) = M(t)$ for all t in a neighborhood of 0 and $M(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x)$$

That is, convergence for $|t| < h$, of mgfs to an mgf implies conergence of cdfs.

Example 2.21 (Poisson approximation)

An approximation that is usually taught in elementary statistics is that binomial probabilities can be approximated by Poisson probabilities:

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

where $\lambda > 0$. In other words, if $X \sim$ binomial(n, p) and $Y \sim$ Poisson(λ) with $\lambda = np$, then

$$\mathbb{P}(X = x) \approx \mathbb{P}(Y = x)$$

for large n and small np. In fact, the mgfs converge. For a Binomial(n, p) random variable, we have

$$M_X(t) = (pe^t + (1 - p))^n$$

For Poisson(λ), we have

$$M_Y(t) = e^{\lambda(e^t-1)}$$

If we define $p = \lambda/n$, then $M_X(t) \to M_Y(t)$ as $n \to \infty$.

To show this, one only needs to consider a key lemma from real analysis on the definition of an exponential: for any real sequence $\{a_n\}$ converging to $a \in \mathbb{R}$:

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Like the mean and variance, mgf's also behave somewhat conveniently with respect to linearity:

Theorem 2.22

For any constants $a, b \in \mathbb{R}$, the mgf of the random variable $aX + b$ is given by

$$M_{aX+b}(t) = e^{bt}M_X(at)$$

3 Common Families of Distributions

Next, we will go over some of the common distributions (and their families) which appear in statistics and on the core exam.

3.1 Discrete Distributions

Definition 3.1 (discrete uniform)**.** A random variable X has a *discrete uniform* $(1, N)$ distribution if $\mathbb{P}(X = x) = \frac{1}{N}$ for $x = 1, \ldots, N$. We have

$$\mathbb{E}[X] = \frac{N + 1}{2}$$
$$\text{Var}(X) = \frac{(N + 1)(N - 1)}{12}.$$

Definition 3.2 (hypergeometric)**.** Suppose we have a large urn filled with N identical balls, M of which are red and $N - M$ of which are green. We select K balls at random (without replacement). What is the probability that exactly x of the balls are red? There are $\binom{N}{K}$ ways to choose a sample of size K, $\binom{M}{x}$ ways to choose x red balls, and $\binom{N-M}{K-x}$ ways to choose $K - x$ green balls. Thus, this is

$$\mathbb{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}.$$

If X is the number of red balls in a sample of size K, then X has a hypergeometric distribution (with parameters N, M, K).

Definition 3.3 (Bernoulli/binomial)**.** A Bernoulli(p) random variable X satisfies

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases},$$

which satisfy

$$\mathbb{E}[X] = p$$
$$\mathrm{Var}(X) = p(1-p).$$

This is the simplest random experiment: tossing a coin with two possible outcomes where p is the probability of a heads.

If n identical Bernoulli trials are performed, then the total number of successes or heads in n trials is distributed as a Binomial(n,p) random variable Y which satisfies

$$\mathbb{P}(Y=y) = \binom{n}{y} p^y (1-p)^{n-y}, y = 0, 1, \ldots, n$$

By linearity of expectation and variance, we have $\mathbb{E}[Y] = np$ and $\mathrm{Var}(Y) = np(1-p)$.

Definition 3.4 (Poisson)**.** The *Poisson* distribution X has a single parameter λ, sometimes called the intensity parameter. It takes values in the nonnegative integers and:

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

We also have $\mathbb{E}[X] = \mathrm{Var}(X) = \lambda$.

Definition 3.5 (negative binomial)**.** The binomial counts the number of successes in a fixed number of Bernoulli trials. From a complementary point of view, the *negative binomial* counts the number of Bernoulli trials required to get a fixed number of successes. Let X denote the trial at which the r-th success occurs in a sequence of independent Bernoulli(p) trials where $r \in \mathbb{N}$. Then,

$$\mathbb{P}(X = x | r, p) = \binom{x-1}{r-1} \cdot p^r \cdot (1-p)^{x-r}, x = r, r+1, \ldots$$

and we say that X has a negative binomial (r,p) distribution. An alternative definition is given through the random variable Y, defined as the number of failures before the r-th success. Thus, we have $Y = X - r$. We have

$$\mathbb{E}[Y] = r\frac{(1-p)}{p} \text{ and } \mathrm{Var}(Y) = \frac{r(1-p)}{p^2}$$

Remark 3.6. The negative binomial family of distributions includes the Poisson distribution as a limiting case. If $r \to \infty$ and $p \to 1$ such that $r(1-p) \to \lambda, \lambda \in (0, \infty)$, then $\mathbb{E}[Y] \to \lambda$ and $\mathrm{Var}(Y) \to \lambda$ which agree with the Poisson mean and variance. One can further show all the probabilities converge as well.

Definition 3.7 (geometric distribution)**.** The *geometric distribution* is a special case of the negative binomial distribution. If $r = 1$ in the pmf expression for the negative binomial, we have

$$\mathbb{P}(X = x | p) = p(1-p)^{x-1}$$

which defines the pmf of a geometric random variable X with success probability p. X can be interpreted as the trial at which the first success occurs. We have

$$\mathbb{E}[X] = \frac{1}{p} \text{ and } \mathrm{Var}(X) = \frac{1-p}{p^2}$$

Remark 3.8. The geometric distribution obeys the "memoryless" property: for integers $s > t$

$$\mathbb{P}(X > s | X > t) = \mathbb{P}(X > s - t)$$

3.2 Continuous Distributions

Definition 3.9 (continuous uniform)**.** The *continuous uniform* distribution with support an interval $[a, b]$ has pdf given by

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

We also have

$$\mathbb{E}[X] = \frac{a+b}{2}$$
$$\text{Var}(X) = \frac{(b-a)^2}{12}.$$

Definition 3.10 (gamma)**.** It follows from the properties of the gamma function that

$$f(t) = \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}, 0 < t < \infty$$

is a pdf. We define the gamma(α, β) family to be given by

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}e^{-x/\beta}, 0 < x < \infty, \alpha > 0, \beta > 0$$

Here, α is the *shape parameter*, and β is the *scale parameter*. We have

$$\mathbb{E}[X] = \alpha\beta$$
$$\text{Var}(X) = \alpha\beta^2$$
$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, t < 1/\beta$$

Example 3.11 (gamma-poisson relationship)
If $X \sim$ Gamma(α, β) with $\alpha \in \mathbb{N}$, then for any x

$$\mathbb{P}(X \leq x) = \mathbb{P}(Y \geq \alpha)$$

where $Y \sim$ Poisson(x/β).

Definition 3.12 (chi squared with p degrees of freedom)**.** If we set $\alpha = p/2$ and $\beta = 2$ for $p \in \mathbb{N}$ in the gamma distribution parameters, then the pdf becomes

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1}e^{-x/2}, 0 < x < \infty$$

which is the *chi squared pdf with p degrees of freedom*.

Definition 3.13 (exponential)**.** If we set $\alpha = 1$ in the gamma pdf, then we have

$$f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}, 0 < x < \infty$$

the *exponential pdf* with scale parameter β. The exponential distribution also satisfies the memoryless property: for $s > t \geq 0$,

$$\mathbb{P}(X > s|X > t) = \mathbb{P}(X > s - t).$$

Definition 3.14 (normal)**.** The pdf of a *normal distribution* with mean μ and variance σ^2, denoted $\mathcal{N}(\mu, \sigma^2)$, is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, -\infty < x < \infty.$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable $Z := (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.

Example 3.15 (normal approximation to binomial)
Let $X \sim$ binomial$(25, .6)$. We can approximate X with a normal random variable, Y with the same mean and standard deviation as X: $\mu = 25(.6) = 15$ and $\sigma = ((25)(.6)(.4))^{1/2} = 2.45$. Thus,

$$\mathbb{P}(X \leq 13) \approx \mathbb{P}(Y \leq 13) = \mathbb{P}\left(Z \leq \frac{13-15}{2.45}\right) = \mathbb{P}(Z \leq -.82) = .206$$

while the exact binomial calculation gives

$$\mathbb{P}(X \leq 13) = \sum_{x=0}^{13} \binom{25}{x} (.6)^x (.4)^{25-x} = .267$$

This is a good, but not great approximation. We can correct this by doing a "continuity correction": instead of approximating $\mathbb{P}(X \leq 13)$, we approximate $\mathbb{P}(X \leq 13.5)$ (so the Gaussian more closely matches the discrete pdf), and obtain

$$\mathbb{P}(X \leq 13) = \mathbb{P}(X \leq 13.5) \approx \mathbb{P}(Y \leq 13.5) = \mathbb{P}(Z \leq -.61) = .271$$

In general, if $X \sim$ binomial(n, p) and $Y \sim \mathcal{N}(np, np(1-p))$ then we approximate

$$\mathbb{P}(X \leq x) \approx \mathbb{P}(Y \leq x + 1/2)$$
$$\mathbb{P}(X \geq x) \approx \mathbb{P}(Y \geq x - 1/2)$$

Definition 3.16 (beta)**.** The *beta* family of distributions is a continuous family with support $(0, 1)$ indexed by two parameters $\alpha, \beta > 0$. The beta(α, β) pdf is

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1,$$

where $B(\alpha, \beta)$ denotes the beta function:

$$B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx.$$

The beta function is related to the gamma function through the following identity:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

We also have

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta},$$
$$\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

3.3 Inequalities and Identities

Inequalities are the main ingredients of many results and proofs in statistical theory. These are some of the inequalities which tend to appear most often in probability theory and statistics.

Theorem 3.17 (Markov's Inequality)
Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$:

$$\mathbb{P}(g(X) \geq r) \leq \frac{\mathbb{E}[g(X)]}{r}$$

Theorem 3.18 (Jensen's Inequality)

Let X be a random variable and let $g(x)$ is a convex function. Then,

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Example 3.19 (Gaussian tail bound)

If $Z \sim \mathcal{N}(0,1)$ is standard normal, then

$$\mathbb{P}(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \cdot \frac{e^{-t^2/2}}{t}, \text{ for all } t > 0$$

4 Problems

4.1 Previous Core Competency Problems

Problem 1 (2018 Summer Practice, # 4)**.** Consider a routine screening test for a disease. Suppose the frequency of the disease in the population (base rate) is 0.5%. The test is highly accurate with a 5% false positive rate and a 10% false negative rate [a false positive happens when a test result indicates that the disease is present (the result is positive), but it is, in fact, not present. Similarly, a false negative happens when a test result indicates that the disease is not present (the result is negative), but it is, in fact, present].

You take the test and it comes back positive. What is the probability that you have the disease?

Problem 2 (2018 Summer Practice, # 4)**.** Suppose that the radius of a circle is a random variable having the following probability density function:

$$f(x) = \frac{1}{8}(3x+1), \qquad 0 < x < 2$$

and 0 otherwise. Determine the probability density function of the area of the circle.

Problem 3 (2018 Summer Practice, # 9)**.** Suppose X_1, X_2 are i.i.d. random variables from a distribution F with mean 0 and variance 1.

(a) If $F = N(0,1)$, show that $\frac{X_1+X_2}{\sqrt{2}} \stackrel{\mathrm{d}}{=} X_1$.

(b) Now suppose that $\frac{X_1+X_2}{\sqrt{2}} \stackrel{\mathrm{d}}{=} X_1$. Show that $F = N(0,1)$.

Problem 4 (2018 Summer Practice, # 14)**.** Suppose $f : [0,\infty) \to \mathbb{R}$ is a function such that $f(x+y) = f(x)f(y)$.

(a) Show that $f(x) \geq 0$ for all real $x \geq 0$.

(b) Show that $f(0) \in \{0,1\}$.

(c) Show that for any nonnegative rational number r one has $f(r) = c^r$, where $c \in [0,\infty)$.

(d) If f is assumed to be continuous, show that $f(x) = c^x$ for all real $x \geq 0$.

(e) Suppose X is a nonnegative random variable such that

$$\mathbb{P}(X > s+t) = \mathbb{P}(X > s)\mathbb{P}(X > t)$$

for every $s, t \geq 0$. If X has a continuous distribution function, name the distribution of X.

Problem 5 (2018 Summer Practice, # 15)**.** Suppose X is a random variable taking values in $[0,1]$.

(a) Show that $\mathrm{Var}(X) \leq 1/4$.

(b) Find a random variable X for which equality holds in part (a).

Problem 6 (2018 September, # 2)**.** We consider balls of random radius R.

(i) Suppose that R is uniformly distributed on $[1,10]$. Find the probability density function of the volume V of a ball (Recall that $V = \frac{4}{3}\pi R^3$).

(ii) Suppose that R has a log-normal distribution, meaning that $\log(R) \sim \mathcal{N}(\mu, \sigma^2)$ for some parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Show that V also has a log-normal distribution and find its parameters.

Problem 7 (2018 September, # 4)**.** (i) Let X be a random variable and $a \in \mathbb{R}$. Show that (using Markov's inequality or otherwise):

$$\mathbb{P}[X \geq a] \leq \inf_{s \geq 0} e^{-sa}\mathbb{E}[e^{sX}].$$

(ii) Let N be a Poisson random variable with parameter $\lambda > 0$; i.e.,

$$\mathbb{P}[N = n] = e^{-\lambda}\frac{\lambda^n}{n}, \qquad n \geq 0.$$

Show that $\mathbb{E}[e^{sN}] = e^{\lambda(e^s-1)}$ for all $s \in \mathbb{R}$.

(iii) Let N be as in (ii) and let $m \geq \lambda$ be an integer. Use (i) and (ii) to show that

$$\mathbb{P}[N \geq m] \leq \left(\frac{\lambda}{m}\right)^m e^{m-\lambda}.$$

Problem 8 (2019 May, # 4)**.** We model the lifetime of a device as a random variable $T \geq 0$ with c.d.f. $F(t)$ and density $f(t)$. Suppose that $f(t)$ is continuous for $t \geq 0$ and define the intensity of failure as

$$\lambda(t) = \lim_{h \downarrow 0} \frac{P[t \leq T \leq t+h | T \geq t]}{h} \qquad \text{for } t \geq 0.$$

(a) Express $\lambda(t)$ through $f(t)$ and $F(t)$.

(b) Compute the intensity of failure when $T \sim \text{Exp}(\alpha)$, $\alpha > 0$.

(c) Show that $F(t) = 1 - \exp\{-\int_0^t \lambda(s)\, ds\}$ for $t \geq 0$.

(d) Determine $F(t)$ and $f(t)$ in the case that $\lambda(t) = \alpha t^\gamma$ for some $\alpha > 0$ and $\gamma > 0$.

Problem 9 (2019 May, # 5)**.** You are working as a TA in the help room for a duration t. The number of students arriving during that period is Poisson distributed with parameter $t\lambda$. For each student, the time T to answer their questions is exponentially distributed with parameter α and this time is independent of all other students. Prove that the distribution of the number X of students that arrive while you are busy with one fixed (randomly chosen) student is geometric with some parameter p and determine p in terms of α and λ.

Hint: The formula $\int_0^\infty s^k e^{-s}\, ds = k!$ for $k = 0, 1, 2, \ldots$ can be used without proof.

Problem 10 (2019 May, # 7)**.** Suppose you have n red balls and one blue ball. We will do two experiments.

(a) In the first experiment, you first drop the n red balls uniformly on the interval $[0, 1]$, independent of each other. Having done this, now you drop the blue ball uniformly in the interval $[0, 1]$, independent of previous ball drops. Let X denote the number of red balls to the left of the blue ball. Find $\mathbb{P}(X = k)$, for $k = 0, \ldots, n$.

(b) In the second experiment, you drop all the $(n + 1)$ balls uniformly on $[0, 1]$, independent of each other. Let Y denote the number of red balls to the left of the blue ball as before. Find $\mathbb{P}(Y = k)$, for $k = 0, \ldots, n$.

Problem 11 (2019 September, # 1)**.** (i) Suppose that X is a nonnegative random variable. Show that

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}X}{t}, \qquad \text{for any } t > 0.$$

(ii) Suppose that $Y \sim N(0, 1)$. Show that

$$\mathbb{P}(Y > t) \leq e^{-t^2/2}, \qquad \text{for any } t > 0.$$

Hint: you can assume without proof that $\mathbb{E}[e^{\lambda Y}] = e^{\lambda^2/2}$ for all $\lambda \in \mathbb{R}$.

Problem 12 (2019 September, # 2)**.** (i) Let X be a random variable distributed according to an exponential law with expectation $1/\lambda$, where λ is a positive constant. Given the real positive random variable X, let us define the discrete variable $Y = \lceil X \rceil$, where $\lceil \cdot \rceil$ is the ceiling function, i.e., the function which rounds any real number upwards to the closest integer. For instance, $\lceil 14.3 \rceil = 15$ and $\lceil 14.8 \rceil = 15$. Show that the random variable Y follows a geometric distribution and identify the parameter.

(ii) Show that for any continuous random variable W with a strictly increasing cumulative distribution function F, we have that $F(W) \sim \text{Uniform}[0, 1]$.

(iii) Using the results of (i) and (ii), propose an algorithm to simulate a realization of the geometric random variable in (i), from $U \sim \text{Uniform}([0, 1])$.

Problem 13 (2020 May, # 3)**.** Suppose $N, \{X_i\}_{i \geq 1}$ are i.i.d. Poisson random variables with mean 1. Let $T = \sum_{i=1}^N X_i$.

(i) Compute expectation $\mathbb{E}[T]$.

(ii) Compute variance $\text{Var}(T)$.

(iii) Find $\mathbb{P}(T = 1)$ as explicitly as possible.

Problem 14 (2021 May, # 8)**.** Daniel and Ann alternatively toss a fair coin. Daniel tosses the coin first, then Ann tosses the coin, then it is again Daniel's turn and so on. We record the sequence. If there is a head followed by a tail the game ends and the person who tosses the tail wins. What is the probability that Daniel wins the game?

Hint: Call the event of Daniel winning the game A, and let $x = P(A)$. Also, let B denote the event that Daniel sees a Head in the first toss. Use the law of total probability to write down

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

Problem 15 (2021 Sept, # 3)**.** Let X and Y be two jointly distributed random variables with finite expectations and variances. Show that $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$.

4.2 Additional Practice

Problem 16 (Casella & Berger, Exercise 2.7)**.** Let X have pdf $f_X(x) = \frac{2}{9}(x + 1)$ for $-1 \leq x \leq 2$. Find the pdf of $Y = X^2$.

Problem 17 (Casella & Berger, Exercise 2.14)**.** Let X be a continuous, nonnegative random variable. Show that

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\, dx,$$

where $F_X(x)$ is the cdf of X.

Problem 18 (Casella & Berger, Exercise 2.18)**.** Show that if X is a continuous random variable, then

$$\min_a \mathbb{E}|X - a| = \mathbb{E}[X - m|,$$

where m is the median of X.

Problem 19 (Casella & Berger, Exercise 2.31)**.** Does a distribution exist for which $M_X(t) = \frac{t}{1-t}$ for $|t| < 1$? If yes, find it. If no, prove it.

Problem 20 (Casella & Berger, Exercise 3.18)**.** Let Y be a negative binomial random variable with parameters r and p, where p is the success probability. Show that as $p \to \infty$, the mgf of the random variable pY converges to that of a gamma distribution with parameters r and 1 (r is the shape parameter). This is known as *convergence in distribution*, which we will discuss in Review Session 5.

Problem 21 (Casella & Berger, Exercise 3.44)**.** For any random variable X for which $\mathbb{E}[X^2]$ and $\mathbb{E}[|X|]$ exist, show that $\mathbb{P}(|X| \geq b)$ does not exceed either $\mathbb{E}[X^2]/b^2$ or $\mathbb{E}[|X|]/b$, where b is a positive constant. If X has pdf $f_X(x) = e^{-x}$ for $x > 0$ show that one bound is better when $b = 3$ and the other when $b = \sqrt{2}$.