Mathematical Statistics I

Chapter 2: Random Variables

Jesse Wheeler

Contents

1 Introduction

Introduction

- This material is based on Chapter 2 of Rice (2007).

- Formally, a *random variable* is a function from a sample space Ω to the real numbers[1].

- That is, for any element $\omega \in \Omega$, a random variable X will map ω to a real number: $X(\omega) \in \mathbb{R}$.

- Most often people think of random variables as random numbers rather than functions; in most instances in this class, this treatment will be sufficient.

Example of a random variable
Consider the experiment of flipping three coins. The sample space is

$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}.$$

---

[1]In this class, will assume real-valued spaces, though more generally a random variable can map to any measureable space

- Some possible random variables include (1) the number of heads, (2) the number of tails, (3) the number of heads minus the number of tails.

- Importantly, a random variable must assign a value to all possible outcomes $\omega \in \Omega$.

Number of Heads

Let X be the random variable representing the number of heads. If the result of the experiment is the outcome *hth*, then $X(\{hth\}) = 2$.

A few comments on random variables

- Sometimes in this course I will use the abbreviation RV to mean "random variable", and you can do so as well.

- It is conventional to use uppercase letters (math text or italics) to denote random variables.

- While a random variable is a function, the outcome of an experiment $\omega \in \Omega$ is random (that's the point), and we only ever see a single outcome.

- Thus, the fact that X is a function is often dropped, and we just write X. The realized value of X is random, because the input is random.

2 Discrete Random Variables

Discrete Random Variables

Definition: Discrete random variable

A discrete random variable is a random variable that can take on only a finite or at most a countably infinite number of values.

- Example: The number of heads in three coin flips can only be in the set $\{0, 1, 2, 3\}$. Alternatively, consider flipping a coin indefinitely until you achieve a heads. The possible outcomes are in the set $\{1, 2, 3, \ldots\}$, which is countably infinite.

Probabilities

- The probability measure on the sample space determines the probability of the values of X.

- In our example, if a coin is fair, then we can assign a uniform probability measure on the sample set of flipping a coin three times.

- That is, all outcomes are equally likely, each with probability 1/8.

- The probability that X takes on it's potential values is easily computed, by counting the number of outcomes that result in the particular value of X:

$$P(X = 0) = \frac{1}{8}$$
$$P(X = 1) = \frac{3}{8}$$
$$P(X = 2) = \frac{3}{8}$$
$$P(X = 3) = \frac{1}{8}.$$

- These simple examples gives us some intuition to derive formulas for the more general case.

- More generally, let's assume that X is a discrete RV, and denote the possible values as $x_1, x_2, \ldots$. There exists a function p such that $p(x_i) = P(X = x_i)$ that satisfies $\sum_i p(x_i) = 1$. This function p is called the *probability mass function* (pmf) of the random variable X.

- We may also be interested in calculating for all values $x \in \mathbb{R}$, the probability $F(x) = P(X \le x)$; the function F is called the *cumulative distribution function* (cdf). The cdf plays a number of important roles in probability and statistics that we will see later on.

Some notes:

- The cdf is non-decreasing (see Theorem 1.2), and

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

- The pmf and cdf are connected: the cdf "jumps" at all values that the pdf $p(x) > 0$.

- Conventionally, the pmf is usually denoted with lower-case letters (e.g., p, f), whereas the cdf is usually denoted with upper-case letters (e.g., F).

See Figures 2.1 and 2.2 of Rice (2007) for a depiction of the pmf and cdf of the 3-coin example.

Independence

- Jumping ahead a little bit, we will define what it means for random variables to be independent (a chapter 3 topic).

Definition: Independent random variables
Let X and Y be discrete random variables defined on the same probability space, taking values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$, respectively. X and Y are said to be independent if, for all i, j,

$$P(X = x_i, Y = y_i) = P(X = x_i)P(Y = y_i).$$

- This definition follows very similarly to that of independent events. We can also extend this definition to *mutual independence* of many variables if the probabilities of all combinations of variables can be factored.

2.1 Common random variables

2.1.1 Bernoulli Random Variables

Bernoulli Random Variables

- A Bernoulli RV only takes on two values[2], 0 and 1, with probabilities $1 - p$ and p, respectively. The pmf is therefore

$$p(1) = p$$
$$p(0) = 1 - p$$
$$p(x) = 0, \quad \text{if } x \ne 0 \text{ and } x \ne 1.$$

- By using the output of 0 and 1, the pmf is usually written in a more compact form:

$$p(x) = \begin{cases} p^x(1-p)^{1-x}, & \text{if } x = 0 \text{ or } x = 1, \\ 0 & \text{otherwise} \end{cases}$$

---

[2]Sometimes you'll see the random variable take values -1 and 1.

Indicator functions

- A common instance of a Bernoulli RV is an *indicator random variable*. Let I_A be the random variable that takes on the value of 1 if the event $A \subset \Omega$ occurs, and 0 otherwise:

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

- Here, we see that $P(I_A = 1) = P(A)$.

2.1.2 Binomail Distribution

Binomial Distribution

- Using what we know about independent RVs and Bernoulli RVs, we can derive the pmf for a Binomial distribution.

- Suppose that we have n independent experiments, where n is a fixed (positive) integer. Let each experiment have two outcomes with probabilities p and $1-p$, respectively, which we call "success" or "failure". We are interested now in the random variable X, the number of "successes" in n independent trials.

- *Question:* What is the probability that $X = k$, for some $k \in \{0, 1, 2, \ldots\}$?

Solution Sketch:

- For $X = k$, we must have *exactly* k successes and $n-k$ failures. By the multiplication law, any one such sequence has probability $p^k(1-p)^{n-k}$ (for instance suppose $n = 3$. What is the probability of the event SFS? it's $p \times (1-p) \times p = p^2(1-p)^{3-2}$.)

- Because we only care about the *number* of successes, not the order, we now have a counting problem: with n total trials (positions), how many ways can we arrange the k successes? Another way of thinking is: "How many ways can we choose k out of n locations in a sequence to place the successes?".

- The answer is $\binom{n}{k}$, so in total, the probability that $X = k$ is:

$$p(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

- The function p above is the pmf for the binomial distribution.

Flipping Coins

Suppose that a coin is flipped 10 times. What is the probability that the coin lands heads heads exactly 6 times?

Here, $n = 10$, and success= Heads. Assuming the coin is fair, we have

$$P(\text{Num Heads} = 6) = \binom{10}{6}(0.5)^6(0.5)^4 \approx 210 \times 0.00098 \approx .205$$

Example: multiple dice

Suppose a five 6-sided (fair) dice are rolled simultaneously. What is the probability that at least two of the dice show the value 6? *Solution:*

- Let X denote the number of 6s in this experiment, which takes values in the set $\{0, 1, \ldots 5\}$. We want the probability that $X \geq 2$.

- Because the different values of X are mutually exclusive events (i.e., $X = 2$ implies $X \neq 3$), we can calculate this as:
$$P(X \geq 2) = \sum_{i \in \{2,3,4,5\}} p(i) \approx 0.1962449,$$
where $p(i)$ is the pmf of the binomial$(5, 1/6)$ distribution.

- Alternatively, we can use the complement set, which is smaller:
$$P(X \geq 2) = 1 - P(X < 2) = 1 - \big(p(0) + p(1)\big) \approx 0.1962449$$

- *Note:* A binomial RV can be expressed as the sum of independent Bernoulli RVs. That is, let $X_1, X_2, \ldots, X_n$ be independent Bernoulli RVs, each with $P(X_i = 1) = p$. Then, $Y = X_1 + X_2 + \ldots + X_n$ is a Binomial RV, with parameters (n, p).

2.1.3 Geometric and Negative Binomial Distributions

Geometric Distribution

- We can construct a *geometric* RV in a similar way that we did with the binomial distribution.

- Suppose instead of having a fixed number of trials, we continue having a trial until our first success. That means that if $X = k$, we will have $k - 1$ failures, one success, and then stop.

- Thus, the pmf can easily be constructed to be:
$$p(k) = P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

Geometric Series
Recall from calculus the geometric series:
$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}, \quad \text{if } 0 < r < 1.$$

This identity occurs in the pmf of the geometric series. Let $0 < p < 1$, then
$$\sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{j=0}^{\infty} (1-p)^j = p \frac{1}{1 - (1-p)} = 1.$$

Negative Binomial Distribution

- The *negative binomial* (NB) distribution can be thought of as a generalization of the geometric distribution; rather than stopping when we have exactly one success, we now will stop when we have r successes.

- For any particular sequence of trials of length k that satisfy this condition, the probability is $p^r (1-p)^{k-r}$.

- The last trial must be a success (because we stopped), so we need to choose the location of the remaining $r - 1$ successes.

- Thus, if X has a negative binomial distribution, the pmf is:

$$p(k) = P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}.$$

- Another way that can be helpful for thinking about the NB-distribution is considering it as the sum of r independent geometric random variables:

- We want to represent the total number of trials until the rth success, which is the sum of the number of trials until (and including) the first success, plus the number of trials from the first success until (and including) the second success, and continued until we get r successes.

Negative Binomial Lottery
Suppose that there is a type of lottery where each purchased ticket has equal probability of winning ($p = 1/100$), and there are 3 total prizes to be won. What is the probability that exactly k tickets will be sold until all prizes have been won?

$$P(X = k) = \binom{k-1}{3-1} (0.01)^3 (0.99)^{k-3}.$$

2.1.4   Hypergeometric and Poisson Distribution

The Hypergeometric Distribution

- Suppose that there is a total population of size n, and r have some trait of interest ("success"), and $n - r$ do not ("failure").

- If we sample m items from the population, then the total number of "successes" in our sample of size m follows a hypergeometric distribution:

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}.$$

- Combinatorially, for $X = k$, we must select k successes out of the total possible r successes in the entire population; there are $\binom{r}{k}$ ways to do this.

- Since we selected m objects in our sample, and we want $m - k$ of them to be failures, we must pick $m - k$ failures from the $n - r$ failures in the population; there are $\binom{n-r}{m-k}$ ways to do this.

- Together, the multiplication principle implies there are $\binom{r}{k}\binom{n-r}{m-k}$ ways that a sample of size m contains k successes from described population.

- Finally, there are a total of $\binom{n}{m}$ ways we can pick our sample:

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}.$$

Balls in a basket
Suppose that there are n balls in a basket, and r balls are black, $n - r$ balls are some other color. If we select $1 \leq m < n$ balls randomly (without replacement), let X denote the number of black balls in our sample of size m. Then, for all $0 \leq k \leq r$,

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}.$$

The Poisson Distribution

- The Poisson distribution is used very frequently in both theory and practice, though the derivation is less intuitive than other distributions, so we will first just provide the pmf:

Definition: Poisson Distribution

The pmf of a random variable X that follows a Poisson distribution with parameter $\lambda > 0$ is

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

- Recall from calculus that $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Thus, like all pmf's, the pmf of a Poisson distributed RV sums to one:

$$\sum_{k=0}^{\infty} p(k) = \sum_{k=0} \frac{\lambda^k}{k!}e^{-\lambda}$$
$$= e^{-\lambda} \sum_{k=0} \frac{\lambda^k}{k!}$$
$$= e^{-\lambda}e^{\lambda} = 1.$$

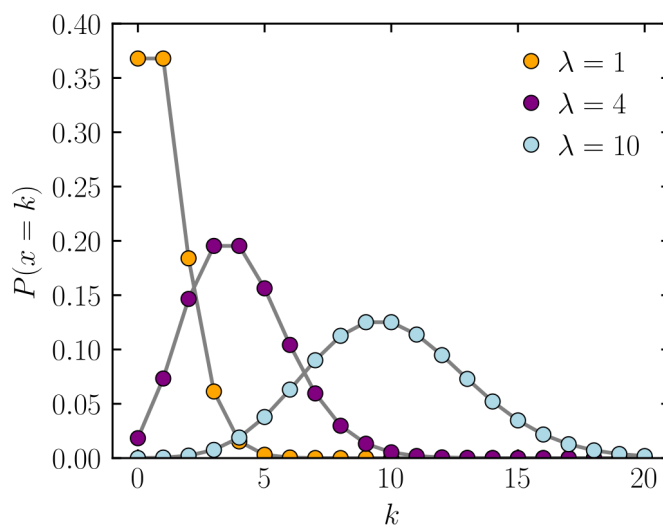- The value of λ controls the *shape* of the distribution:

Figure 1: Shape of the Poisson distribution for various values of λ (Wikipedia contributors, 2025b).

- The Poisson distribution can be derived as the limit of a binomial distribution as the number of trials $n \to \infty$, and $p \to 0$, such that $np = \lambda$.

Derivation:

Recall the pmf of the binomal distribution can be expressed as

$$p(k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}.$$

7

Setting $np = \lambda$, the expression becomes:

$$p(k) = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

Now taking the limit $n \to \infty$,

$$\frac{\lambda}{n} \to 0$$

$$\frac{n!}{(n-k)!n^k} \to 1$$

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}$$

$$\left(1 - \frac{\lambda}{n}\right)^{-k} \to 1$$

And therefore

$$p(k) \to \frac{\lambda^k e^{-\lambda}}{k!}.$$

This derivation suggests how a Poisson distribution can arise in practice.

- Let X denote the random variable representing the number of times some event occurs in a fixed time interval.

- Think of dividing the interval into very large number of small sub-intervals of equal length.

- Assume that the sub-intervals are so small that the probability of more than one event in a sub-interval is negligible relative to the probability of one event (which itself is small).

- Finally, assume that the probability of an event in a given sub-interval is identical and independent of that of other sub-intervals.

- Following this, X is nearly binomially distributed, with n being the number of sub-intervals, and $p = \lambda/n$ the probability of the event in each sub-interval.

- Taking the limit, we get something that is nearly Poisson distributed.

- This idea can actually be formalized and made rigorous; you would probably see something like this in a course on stochastic processes.

- The Poisson distribution is often used to model the number of events that occur in a fixed interval.

The Poisson distribution is often good model for the number of events in a fixed time interval if the following conditions are met:

- The occurrence of one event does not affect the occurrence of another.

- The rate at which events occur is fixed.

- Two events cannot occur at the exact same instant.

In this scenario, the random (stochastic) process that generates the data is called a *Poisson process*, which gives rise to the name *rate* for the parameter λ.

Example: Telephone calls
Suppose that an office receives telephone calls as a Poisson process with $\lambda = 0.5$ calls per minute. The number of calls in a 5-min. interval follows a Poisson distribution with parameter $5\lambda = 2.5$. Thus, the probability of no calls in a 5-min. interval is $p(0) = e^{-2.5} \approx .082$; the probability one one call is $p(1) = 2.5e^{-2.5} \approx .205$

3   Continuous Random Variables

The distribution function

- All random variables have an associated distribution function.

- We have discussed how this arises in the context of discrete random variables, now we want to discuss this for continuous random variables.

- If X is a random variable, then its cumulative distribution function is given by

$$F(x) = P(X \leq x), \quad \text{for all } x.$$

- Some general properties of F can be shown, but requires a more formal definition of probability spaces, so we will just state the result.

Theorem 2.1: The cdf
The function $F(x)$ is a cdf if and only if the following conditions hold

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

- $F(x)$ is a non-decreasing function of x.

- $F(x)$ is right-continuous; that is, for every number x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

- An interesting aspect of this theorem is that it states that every function that satisfies these conditions is a cdf for some random variable.

- The definition of cdf also leads to an alternative definition of continuous vs discrete random variables, which is that a continuous random variable has continuous cdf, and a discrete random variable does not.

- One important thing to note is that the cdf completely determines the probability distribution of a random variable, which is why it is sometimes called the distribution function.

- We will formalize this with a definition and a theorem.

Definition: Identical Distribution
Random variables X and Y are said to be *identically* distributed if for every (measurable) set $A \subset \mathbb{R}$, $P(X \in A) = P(Y \in A)$.

- If X and Y are identically distributed, we often use the shorthand(s):

$$X \overset{d}{=} Y, \quad \text{or} \quad X \sim Y.$$

- This definition is intuitive, but it can be hard to show that X and Y have the same distribution, as it would require checking all possible sets.

- The following theorem gives us an easier way to check the equivalence of distributions.

Theorem 2.2: Identical distributions
Random variables X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for every x. We will only prove the forward direction; the opposite direction requires more background on sigma-algebras than we would like to develop for this class.

Proof. If X and Y are identically distributed, then by definition, $P(X \in (-\infty, x)) = P(Y \in (-\infty, x))$, because $(-\infty, x)$ is a measurable set for all x. But $F_X(x) = P(X \in (-\infty, x)) = P(Y \in (-\infty, x)) = F_Y(x)$, proving the statement. $\square$

- Because the distribution is determined by the cdf, we often use the shorthand $X \sim F(x)$ to mean that "X has a distribution given by F(x)".

- As we have already seen for discrete RVs, the cdf is connected to the pmf in the following way:

$$F(x) = P(X \le x) = \sum_{i=-\infty}^{x} p(i).$$

- A natural extension of this idea for continuous random variables is to replace the sum with an integral.

Definition: probability density function
Let X be a random variable with distribution function F. Then the *probability density function* of X is any function $f(x)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f(t)dt, \quad \text{for all } x.$$

- By the fundamental theorem of calculus, we can see that if F is continuous, then $\frac{d}{dx}F(x) = f(x)$, or that taking the derivative of the cdf gives the pdf.

- An interesting thing to notice about the definition of a pdf is that it is not necessarily unique; it is only unique "almost everywhere", meaning the area where the functions don't match integrates to zero.

- For any given distribution, the density function may not exist, because F may exist and be continuous but not differentiable. We will avoid these pathological cases in this class.

- We also note that any non-negative function with a finite positive integral over a set A can be turned into a pdf.

- For example, suppose $h(x)$ is positive over the set A, and

$$\int_A h(t)dt = K,$$

then

$$f(x) = \begin{cases} h(x)/K & x \in A \\ 0 & x \notin A \end{cases}$$

is a valid pdf.

Theorem: properties of pmf / pdf
Given the definition of pmf and pdf, we can readily deduce the following properties. Let $f(x)$ be the pdf or pmf of a random variable X. Then

- $f(x) \ge 0$ for all x.

- $\sum_x f(x) = 1$ (if pmf)

- $\int_x f(x) = 1$ (if pdf).

- The proof to these are not so interesting or difficult, so they are left for practice.

Continuous Random Variables

- Because discrete RVs take only a finite number of possibilities, they are relatively simple to define by assigning probabilities to each possible outcome.

- In many situations, however, we are interested in random variables that can take on a continuum of values rather than a finite (or countably infinite) number.

Example: Lifetime of electronic
We might be interested in the lifetime of an electronic component; the total lifetime may be random, but may take on any positive real number.

Probabilities

- If X is a random variable with a density function f, then for any $a \leq b$, the probability that X falls in the interval (a, b) (with the treatment that if $a = b$, the interval collapses to the set $\{a\}$) is given by:
$$P(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx.$$

- An immediate consequence of this definition is that $P(X = a) = 0$ for any $a \in \mathbb{R}$.

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

- This is because the probability X being any particular value is zero; if this were not the case, then the probability of the entire set would infinite (and probabilities must sum to one).

Example: Continuous uniform random variable

- By *uniform* probability, we mean that all outcomes in the given set are equally as likely.

- Specifically, if $S \subset \mathbb{R}^p$ for some p, then for any $A \subset S$, X is uniformly distributed on S if $P(X \in A) = \text{volume}(A)/\text{volume}(S)$.

- For example, if X is a RV with uniform distribution on the interval $[0, 1]$, then any real number in this interval is equally likely, and the probability that X is in a sub-interval of length h should be equal to h.

- You can verify that the following density satisfies this condition:

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0 & x < 0 \text{ or } x > 1. \end{cases}$$

- For instance, we can pick some $c \in [0, 1]$, and $h \in (0, 1-c)$. Then the probability that $X \in (c, c+h)$ is given by:

$$P(c < X < c + h) = \int_c^{c+h} 1_{0 \leq x \leq 1} dx = \int_c^{c+h} 1 dx = (c + h) - c = h.$$

- The previous density can be generalized to any interval $[a, b]$, such that $a < b$.

3.1 The Uniform Random Variable

Continuous uniform density
If X is a RV uniformly distributed on an interval $[a, b]$, where $a < b$, then the corresponding density function is:
$$f(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & x < a \text{ or } x > b. \end{cases}$$

cdf of continuous uniform density
From the definition, we can calculate the cdf of the continuous uniform density rather easily. Suppose that X is uniformly distributed on $[0, 1]$. Then the cdf F is:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx.$$

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(x)\,dx \\ &= \int_{-\infty}^{x} 1[0 \leq x \leq 1]\,dx \\ &= \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1. \end{cases} \end{aligned}$$

3.2 Common random variables

3.2.1 The Exponential Distribution

Exponential Distribution

- The exponential density function is:
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Like the Poisson distribution, the exponential density function depends on a single parameter λ.

- When this is the case, we refer to it as the *family* of exponential densities that is *indexed* by the parameter λ.

- The cdf is easily found via the fundamental theorem of calculus:
$$F(x) = \int_{-\infty}^{x} f(u)du = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- From this, we can easily find quantiles of the distribution, such as the median: by solving $F(x_{(.5)}) = 1/2$
$$1 - e^{-\lambda x_{(.5)}} = \frac{1}{2} \quad \implies \quad x_{(.5)} = \frac{\log 2}{\lambda}.$$

- The exponential distribution is often used to model lifetimes or waiting times (time-to-event).

- In this context, it's conventional to replace the variable x with t.

- The exponential distribution has a unique property known as the *memoryless* property.

- That is, if something follows an exponential distribution and has already lasted a time of s, then the probability that it will last another t units of time does not depend on s:

Memoryless property: Let T be an exponentially distributed RV, and $s, t > 0$. Calculate $P(T > t + s | T > s)$.

$$
\begin{aligned}
P(T > t + s | T > s) &= \frac{P(T > t + s \text{ and } T > s)}{P(T > s)} \\
&= \frac{P(T > t + s)}{P(T > s)} \\
&= \frac{1 - F(t + s)}{1 - F(s)} \\
&= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\
&= e^{-\lambda t} = P(T > t).
\end{aligned}
$$

- It can be shown that any continuous RV with the *memoryless* property must be exponentially distributed.

- Similarly, it can be shown that any discrete RV with the *memoryless* property must be geometrically distributed (maybe a HW question?)

- The exponential distribution is also related to the *Poisson process* that we have discussed.

- Consider a poisson process with rate λ over an interval $\mathcal{T} \subset \mathbb{R}$.

- While the number of events in any interval $T_0 \subset \mathcal{T}$ of length t follows a Poisson distribution, the time-to-next-event T follows an exponential distribution:

- Suppose that an event occurs at time $t_0 \in T_0$, and let T denote the time until next event. Then:

$$
\begin{aligned}
P(T > t) &= P\big(\text{no events in}(t_0, t_0 + t)\big) \\
&= P(X = 0), \quad \text{where } X \sim \text{Pois}(\lambda t). \\
&= e^{-\lambda t},
\end{aligned}
$$

- Therefore T is exponentially distributed with parameter λ.

3.2.2 The Gamma Density

The Gamma Density

- The *gamma* density function depends on two parameters $\alpha > 0$ and $\lambda > 0$.

$$
g(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, \quad t \geq 0.
$$

- For $t < 0$, we define $g(t) = 0$.

- The *gamma function*, is defined as:

$$
\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du, \quad x > 0.
$$

13

- Note that if $\alpha = 1$, then the gamma density coincides with the exponential density.

- The parameter $\alpha > 0$ in this formulation is called the *shape* parameter.

- The parameter $\lambda > 0$ is called the *scale* parameter.

- As the names suggest, α changes the *shape* of the density function, whereas λ changes the scale of the density (i.e., can be used to change from inches to feet).
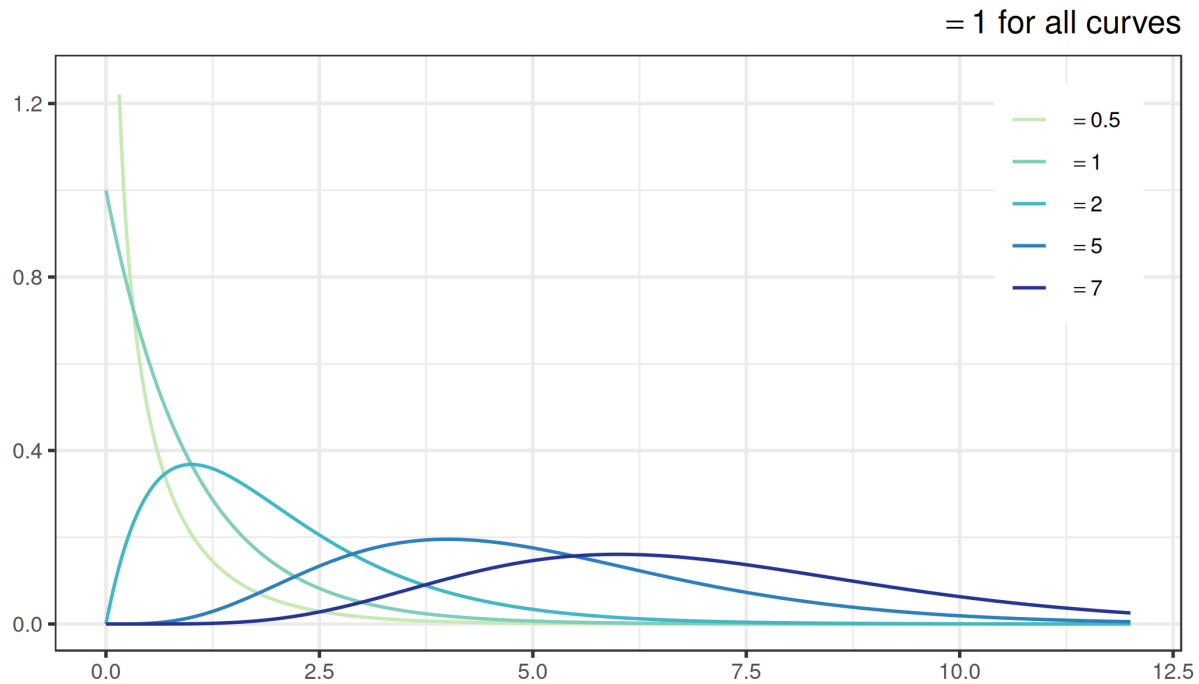

Figure 2: A few Gamma pdf functions for various levels of α.

3.2.3 The Normal Distribution

The Normal Density

- The normal distribution plays a central role in both probability and statistics.

- It is also called *the Gaussian* distribution, after Carl Friedrich Gauss, who used it as a model for modeling measurement errors.

- One reason it is so important is the *Central Limit Theorem* (CLT, Chapter 6), which suggests that the normal distribution is useful in a large number of settings.

- Roughly speaking, the CLT states that large(ish) sums (or averages) of independent random variables will be approximately normally distributed.

The density function for the normal distribution depends on two parameters:

- μ: the mean of the distribution.

- σ: the standard deviation of the distribution.

The Normal Density

Let X be a random variable that is normally distributed with mean μ and standard deviation σ. The corresponding probability density function is

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

A few notes about this density / distribution:

- You'll often see the shorthand $X \sim N(\mu, \sigma^2)$ to mean "X follows a normal distribution with parameters μ and σ".

- The density is symmetric about μ, meaning $f(\mu - x) = f(\mu + x)$; $x = \mu$ is also the maximum value of the density (mode). The "spread" of the distribution is determined by σ.

- When the mean $\mu = 0$ and standard deviation $\sigma = 1$, we call this the *standard normal distribution.*

- There is no closed form expression for the cdf of the normal distribution. The cdf of the standard normal is usually denoted $\Phi(\cdot)$, and the pdf $\phi(\cdot)$.

3.2.4   The Beta Distribution

The Beta Desnity

- The *beta* distribution is useful for modeling random variables that are restricted to the interval $[0, 1]$:

- The density function depends on two parameters, $\alpha, \beta > 0$.

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1 - x)^{\beta-1}, \quad 0 \leq x \leq 1.$$

- Note that if $\alpha = \beta = 1$, the distribution becomes uniformly distributed.

- Both α and β are shape parameters, and the distribution is quite flexible.

Comments on density functions

- So far, we have been using f and F to denote the pdf and cdf of a random variable, respectively. If there is more than one random variable, say X and Y, we may want to distinguish between the functions, and we do so like: $f_X(x)$, or $F_Y(y)$.

- There exist alternative *parameterizations* of the common pmf / pdf functions we have discussed. For example, we introduced the negative binomial distribution with parameters p and r. This is known as the size-probability parameterization, but there also exists and alternative with parameters μ and k, known as the mean-ddispersion parameterization (commonly used in Ecology).

4   Functions of a random variable

Variable transformations

- Often we will be interested in function of a random variable.

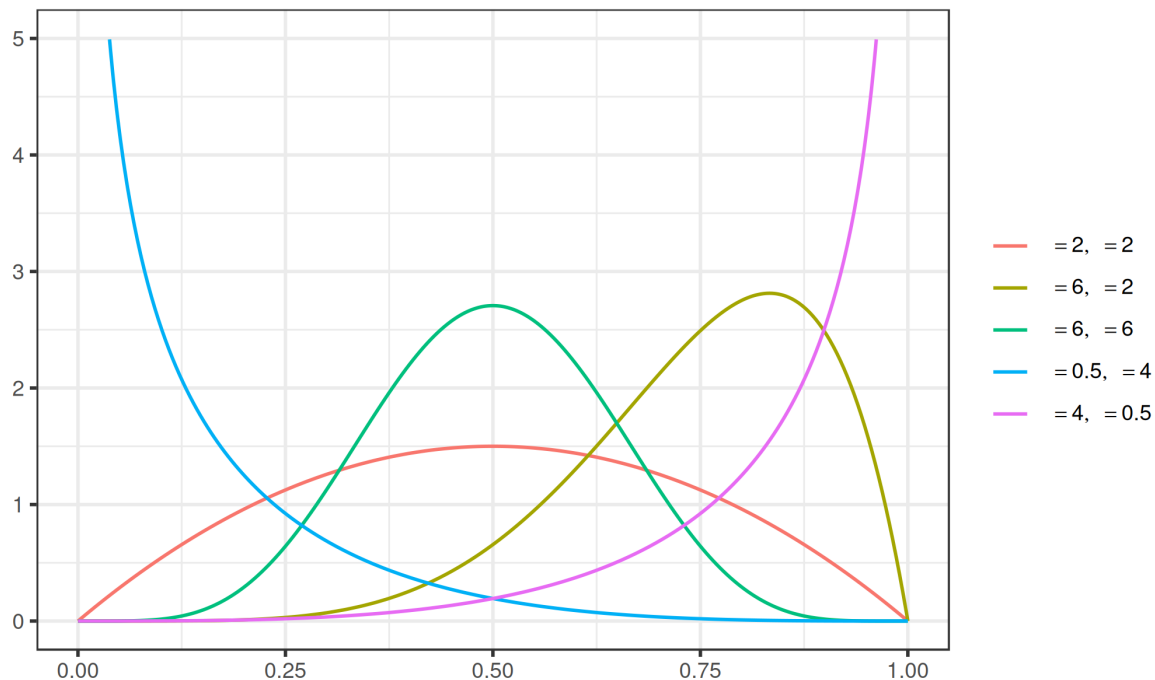- Let X be a random variable, and g an arbitrary function.

Figure 3: A few Beta pdf functions for various levels of α and β.

- Our goal is to find the distribution of the random variable $Y = g(X)$.

Kinetic energy

Let X denote a random variable representing the velocity of a particle of mass m; we might be interested in the distribution of $Y = \frac{1}{2}mX^2$, the particle's kinetic energy.

- We will eventually provide a rule for a general transformation, $g : \mathbb{R} \to \mathbb{R}$, but we will first build some intuition using simple transformations.

Example: Linear Transformations (Part I)

Let X be a random variable with pdf and cdf f_X and F_X, respectively. Find the density of $Y = aX + b$, where $a > 0$, and $b \in \mathbb{R}$.

We will use what I like to call "the cdf method". It's a simple idea that uses the connection between the cdf and pdf of a random variable, and the definition of the cdf.

$$
\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(aX + b \leq y) \\
&= P\left(X \leq \frac{y - b}{a}\right) \\
&= F_X\left(\frac{y - b}{a}\right).
\end{aligned}
$$

On the left, we have the cdf of Y, and we have equated this to the cdf of X. Now we can differentiate

16

the equation with respect to y:

$$f_Y(y) = \frac{d}{dy}F_X\left(\frac{y-b}{a}\right)$$
$$= \frac{1}{a}f_X\left(\frac{y-b}{a}\right).$$

- This same idea is how we will build a general formula for finding the pdf of a transformed random variable.

- Now let's change it slightly, and see some potential pitfalls.

Example: Linear Transformations (Part II)
Let X be a random variable with pdf and cdf f_X and F_X, respectively. Find the density of $Y = aX + b$, where $a < 0$, and $b \in \mathbb{R}$.

$$F_Y(y) = P(Y \leq y)$$
$$= P(aX + b \leq y)$$
$$= P\left(X \geq \frac{y-b}{a}\right)$$
$$= 1 - F_X\left(\frac{y-b}{a}\right).$$

On the left, we have the cdf of Y, and we have equated this to the cdf of X. Now we can differentiate the equation with respect to y:

$$f_Y(y) = 0 - \frac{d}{dy}F_X\left(\frac{y-b}{a}\right)$$
$$= -\frac{1}{a}f_X\left(\frac{y-b}{a}\right).$$

- Note that all pdf functions have to be positive. In this example, we have $a < 0$, and therefore the negative sign in the final result helps us ensure that $f_Y(y)$ is positive.

- We can also see that $\left|\frac{1}{a}\right| = -\frac{1}{a}$, so we can write:

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right)\left|\frac{1}{a}\right|.$$

- More generally, suppose that X is a random variable, and $Y = g(X)$ is a new random variable.

- Since Y is a function of X, we can describe the probabilistic behavior of Y in terms of that of X.

- For any set $A \subset \mathbb{R}$, we have
$$P(Y \in A) = P\big(g(X) \in A\big),$$
showing that the distribution of Y depends on the functions F_X and g.

- Suppose X takes values in $\mathcal{X}$, then Y takes values in $\mathcal{Y}$, where

$$g : \mathcal{X} \to \mathcal{Y}.$$

- We will define $g^{-1}(A)$ to be the set of values in $\mathcal{X}$ that get mapped to A:

$$g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}.$$

- For ease of communicating, we'll call $g^{-1}(A)$ the *pre-image* of A.

- The mapping g^{-1} takes sets into sets.

- For point sets like $\{y\}$, we often write $g^{-1}(y)$ rather than $g^{-1}(\{y\})$.

- The quantity $g^{-1}(y)$, can still be a set, if more than one x gets mapped to y.

- If $g^{-1}(y)$ is the point set $\{x\}$, we will write $g^{-1}(y) = x$.

- Returning to probabilities, we have

$$
\begin{aligned}
P(Y \in A) &= P\big(g(X) \in A\big) \\
&= P\big(\{x \in \mathcal{X} : g(x) \in A\}\big) \\
&= P\big(X \in g^{-1}(A)\big).
\end{aligned}
$$

- This gives a formal definition for the probability distribution of Y, and one can show this satisfies the probability axioms.

- If X is a discrete random variable, then $\mathcal{X}$ is countable.

- We immediately see that $\mathcal{Y} = \{y : y = g(x), x \in \mathcal{X}\}$ is also a countable set, and therefore a discrete random variable.

- The pmf of Y is found by:

$$
\begin{aligned}
p_Y(y) &= P(Y = y) \\
&= P\big(g(X) = y\big) \\
&= P\big(X \in g^{-1}(y)\big) \\
&= \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} p_X(x).
\end{aligned}
$$

- That is, finding the pmf of Y involves identifying $g^{-1}(y)$ for all $y \in \mathcal{Y}$ and summing the appropriate probabilities.

Example: Binomial Transformation
Let X have a binomial distribution, so that

$$
p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.
$$

Find the pmf of $Y = n - X$. *Solution.* Here, $g(x) = n - x$, and $\mathcal{X} = \{0, 1, \ldots, n\}$ and $\mathcal{Y} = \{0, 1, \ldots, n\}$. For any $y \in \mathcal{Y}$, we have $n - x = y$ if and only if $x = n - y$. Therefore for all y, $g^{-1}(y)$ is just the single point $n - y$. Therefore

$$
\begin{aligned}
p_Y(y) &= \sum_{x \in g^{-1}(y)} p_X(x) \\
&= p_X(n - y) \\
&= \binom{n}{n-y} p^{n-y} (1-p)^{n-(n-y)} \\
&= \binom{n}{y} (1-p)^y p^{n-y}.
\end{aligned}
$$

This result shows that Y is also a binomial distribution, with the probability is now $1 - p$.

- If X is a continuous random variable, there are several transformation where the basic definition of $Y = g(X)$ can be used directly to find the cdf / pdf of Y, like the linear transformations we considered.

- More generally, we have:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\big(g(X) \leq y\big) \\ &= P\big(\{x \in \mathcal{X} : g(x) \leq y\}\big) \\ &= \int_{\{x \in \mathcal{X} : g(x) \leq y\}} f_X(x)dx. \end{aligned}$$

- Thus, the key is finding the set $\{x \in \mathcal{X} : g(X) \leq y\}$ and integrating over this set, which can sometimes be difficult.

Example: $g(x) = \sin^2(x)$
Suppose X is uniformly distributed on $(0, 2\pi)$, so that

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

If $Y = \sin^2(X)$, find $P(Y \leq y)$. *Solution.*

- For a fixed y, we need to find the set of X such that $g(x) \leq y$.

- The function $g(x) = \sin^2(x)$ takes values in $[0, 1]$, so we will focus on y in this set.

- To find the following information, it's helpful to plot the function: https://www.desmos.com/calculator/vvm6cm8fd3.

- for any $y \in (0, 1)$, there are four points of intersection, where $g(x) = y$.

- For $Y \leq y$, we must have $X \in (0, x_1] \cup [x_2, x_3] \cup [x_4, 2\pi)$, where $x_1, \ldots, x_4$ are the found solutions to $g(x) = y$.

- Thus,
$$P(Y \leq y) = P(X \leq x_1) + P(x_2 \leq X \leq x_3) + P(X \geq x_4).$$

- Because X is uniform and because of the symmetry of $g(x)$, we can simplify this further if we want:
$$P(Y \leq y) = 2P(X \leq x_1) + 2P(x_2 \leq X \leq \pi),$$
where x_1 and x_2 are the two solutions to $\sin^2(x) = y$, with $0 < x < \pi$.

- As we can see from the previous example, even simple transformations can result in expressions that are not very simple.

- One of the easiest places to make a mistake is keeping track of the sample space of each of the random variables.

- To simplify things as much as possible, it helps to make $\mathcal{X}$ to be the set of values of x that have positive pdf, and the let $\mathcal{Y}$ as the possible values of Y given the set $\mathcal{X}$:

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

- As defined above, the set $\mathcal{X}$ is called the *support* of X

- The easiest type of transformation are functions g that are *monotone.*

- g is a monotone, increasing function if

$$u > v \implies g(u) \geq g(v).$$

- g is a monotone, decreasing function, if

$$u < v \implies g(u) \geq g(v).$$

- We say that a function is strictly monotonic if the inequalities above are strict.

- For strictly monotonic functions, then g is a one-to-one and onto function, meaning that each $x \in \mathcal{X}$ is mapped to one and only one $y \in \mathcal{Y}$, and each $y \in \mathcal{Y}$ comes from one and only one $x \in \mathcal{X}$.

- In other words, g uniquely pairs each x and y together, and we can find an inverse function such that $g^{-1}(y) = x$ if and only if $g(x) = y$.

- If g is strictly monotone and *increasing*, then the set of all points x such that $g(x) \leq y$ is given by the set of all points x smaller than $g^{-1}(y)$:

$$\{x \in \mathcal{X} : g(x) \leq y\} = \{x \in \mathcal{X} : g^{-1}\big((g(x)) \leq g^{-1}(y)\big)\}$$
$$= \{x \in \mathcal{X} : x \leq g^{-1}(y)\}.$$

- The exact opposite is true if g is a strictly monotone *decreasing* function.

- For the set of x such that $g(x) \leq y$ is the same as the set of x such that $x \geq g^{-1}(y)$:

$$\{x \in \mathcal{X} : g(x) \leq y\} = \{x \in \mathcal{X} : g^{-1}\big((g(x)) \geq g^{-1}(y)\big)\}$$
$$= \{x \in \mathcal{X} : x \geq g^{-1}(y)\}.$$

- These results will lead us to our next theorem

Theorem 2.3: cdf of monotone transformations
Let X have a cdf of $F_X(x)$, and let $Y = g(X)$. Then

- If g is a strictly monotone increasing function on $\mathcal{X}$, then

$$F_Y(y) = F_X\big(g^{-1}(y)\big), \quad \text{for } y \in \mathcal{Y}.$$

- If g is a strictly monotone decreasing function on $\mathcal{Y}$, then

$$F_Y(y) = 1 - F_X\big(g^{-1}(y)\big), \quad \text{for } y \in \mathcal{Y}.$$

Proof.   - We will show this for the case that g is decreasing; the proof when g is increasing is similar.

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P\big(g(X) \leq y\big) \\
&= P\big(\{x \in \mathcal{X} : g(x) \leq y\}\big) \\
&= P\big(\{x \in \mathcal{X} : x \geq g^{-1}(y)\}\big) \\
&= \int_{g^{-1}(y)}^{\infty} f_X(x)\, dx = 1 - F_X\big(g^{-1}(x)\big).
\end{aligned}$$

$\square$

- This theorem gives the immediate consequence for random variables that have pdfs.

Theorem 2.4: pdf of monotonic transformations

Let X be a random variable with continuous pdf $f_X(x)$ on $\mathcal{X}$, and let $Y = g(X)$. If g is a strictly monotonic transformation such that g^{-1} has a continuous derivative on $\mathcal{Y}$, then the pdf of Y is given by

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

Proof. The proof is a direct consequence of the previous theorem and the chain rule. Specifically,

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\frac{d}{dy}g^{-1}(y) & \text{if } g \text{ is increasing.} \\ -f_X\left(g^{-1}(y)\right)\frac{d}{dy}g^{-1}(y) & \text{if } g \text{ is decreasing.} \end{cases}$$

Because the derivative of a decreasing function is always negative, we can concisely express both cases by using the absolute value. □

Example: Inverted Gamma Distribution

Let X be a gamma(λ, α) distributed random variable, which has pdf

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \quad x \geq 0$$

Find the pdf of $Y = g(X) = 1/X$.

- First, we will find the support sets $\mathcal{X}$, $\mathcal{Y}$.

- $f(x)$ is positive on the interval $(0, \infty)$, (we will avoid dividing by zero, which doesn't change the distribution because it's only a single point), giving $\mathcal{X} = (0, \infty)$.

- Over this set, the support of Y is also $\mathcal{Y} = (0, \infty)$.

- On these sets, g is monotone, and if $g(x) = y$, then $g^{-1}(y) = x = 1/y$.

- Over the set $\mathcal{Y}$, we can evaluate the derivative as

$$\frac{d}{dy}g^{-1}(y) = -\frac{1}{y^2}.$$

- Applying the change of variable theorem,

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\left(1/y\right)^{\alpha-1}e^{-\lambda/y}\frac{1}{y^2} \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\left(1/y\right)^{\alpha+1}e^{-\lambda/y}. \end{aligned}$$

- This pdf is a special case of the inverted gamma pdf.

- Theorem 2.4 is useful to have / know, but it's often easier to just work from scratch (we'll see a few examples of this).

Example: Square Transformation

Suppose X is a continuous random variable with support $\mathcal{X} = \mathbb{R}$, and let $Y = X^2$. Find the pdf of Y.

- In this case, the function $g(x) = x^2$ is *not* monotonic on the support $\mathcal{X}$, and we can't apply the theorem.

- However, it's relatively easy to work from first-principles in this case.

- First, we can readily see that $\mathcal{Y} = [0, \infty)$.

- Consider the interval $A_y = (-\infty, y]$, for some $y \in \mathcal{Y}$. Then $Y \in A_y$ if and only if $X^2 < y$, or $-\sqrt{y} \leq X \leq \sqrt{y}$, so
$$g^{-1}(A_y) = \{x \in \mathcal{X} : -\sqrt{y} \leq x \leq \sqrt{y}\}.$$

- Thus, we're looking for the probability that X lies in the interval $(-\sqrt{y}, \sqrt{y})$.

- We want to express this in terms of the pdf of X (which we assume exists), in which case we want to express this first in terms of the cdf of X and then take a derivative.

- We can short-hand this calculation as:
$$\begin{aligned}
P(Y \in A_y) = P(Y \leq y)\big(= F_Y(y)\big) \\
= P(X^2 \leq y) \\
= P(\sqrt{y} \leq X \leq \sqrt{y}) \\
= P(X \leq \sqrt{y}) - P(X \leq -\sqrt{y}) \\
= F_X(\sqrt{y}) - F_X(-\sqrt{y}).
\end{aligned}$$

- Now taking derivatives, we have
$$\begin{aligned}
f_Y(y) &= \frac{d}{dy}\big(F_X(\sqrt{y}) - F_X(-\sqrt{y})\big) \\
&= \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}) \\
&= \frac{1}{2\sqrt{y}}\big[f_X(\sqrt{y}) + f_X(-\sqrt{y})\big]
\end{aligned}$$

- Notice in the pdf of $Y = X^2$, we have expressed the pdf as the sum of two pieces. These two pieces represent the two different sets where $g(x)$ is monotone (decreasing on $(-\infty, 0)$, and increasing on $[0, \infty)$).

- This principle can be used to extend Theorem 2.4 to general, non-monotonic functions g. The idea is that you break $g : \mathcal{X} \to \mathcal{Y}$ into a series of functions $g_i : \mathcal{X}_i \to \mathcal{Y}$ where $\mathcal{X}_i \subset \mathcal{X}$ is a set such that g_i is monotonic. Then, the final density f_Y can be found by summing these functions:
$$f_Y(y) = \sum_i f_X\big(g_i^{-1}(y)\big)\Big|\frac{d}{dy}g_i^{-1}(y)\Big|.$$

See Casella and Berger (2024, Theorem 2.1.8) for more details.

- Here are some common RV transformations that really come in handy.

Example: Normal Distribution I
Let $X \sim N(\mu, \sigma^2)$. Then if $Y = aX + b$, $Y \sim N(a\mu + b, a^2\sigma^2)$.
proof: direct consequence of Linear Transformation examples (parts I and II)

Proposition 2.2: Uniform CDF
Let X be a random variable with cdf F. Then $Z = F(X)$ has a uniform distribution on $[0, 1]$.
proof: $P(Z \leq z) = P\big(F(X) \leq x\big) = P\big(X \leq F^{-1}(z)\big) = F\big(F^{-1}(z)\big) = z$

Proposition 2.3: Inverse Uniform CDF
Let U be uniform on $[0, 1]$, and let $X = F^{-1}(U)$. Then the cdf of X is F.
proof: $P(X \leq x) = P\big(F^{-1}(U) \leq x\big) = P\big(U \leq F(x)\big) = F(x)$

- This last proposition is very useful. We can use it to *generate random numbers* (pseudorandom).

- Many computer packages have ways of generating numbers uniformly on $U[0, 1]$; the proposition implies that to generate from any arbitrary distribution with cdf F, all we need to do is apply F_{-1} to uniform $[0, 1]$ random numbers.

Chi-square distribution

- If $Z \sim N(0, 1)$, find the density of $X = Z^2$.

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\ &= \Phi(\sqrt{x}) - \Phi(\sqrt{x}). \end{aligned}$$

- We now find the density by differentiation the cdf with respect to x.

$$\begin{aligned} f_X(x) &= \frac{1}{2}x^{-1/2}\phi(\sqrt{x}) + \frac{1}{2}x^{-1/2}\phi(-\sqrt{x}) \\ &= x^{-1/2}\phi(\sqrt{x}) \quad \text{since } \phi \text{ is symmetric} \\ &= \frac{x^{-1/2}}{\sqrt{2\pi}}e^{-x/2}, \quad x \geq 0. \end{aligned}$$

- If you recall that $\Gamma(1/2) = \sqrt{\pi}$, you may recognize that this is a particular instance of the gamma density, with $\alpha = \lambda = 1/2$.

- We call this density the *chi-square density* with 1 degree of freedom.

5 Miscellaneous

Percentiles

- You're probably familiar with the term *median*. For any given sample, the median defines the "mid-point", meaning that half of the values are larger, half are smaller.

- This same concept applies to distribution functions.

- That is, the median of a distribution F is defined to be that value $x_{.5}$ such that $P(X < x_{.5}) = 0.5$.

- Formally, the sample median is the same as the definition above, using the *empirical* distribution function (Wikipedia contributors, 2025a).

It is important to note that, as defined, the median value may not be unique!

Definition: Percentile

Let F be the cdf of a continuous random variable. The pth quantile of the distribution F is defined to be any value x_p such that $F(x_p) = P(X \leq x_p) = p$. If F is strictly increasing, then x_p is unique and we say that $F^{-1}(p) = x_p$.

If F is not strictly increasing, then x_p may not be unique; in this case, all such values are considered percentiles. If an inverse function is needed in this case, we will define $F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$.

- The last bit of the definition is just some important book keeping to ensure the inverse function exists in odd examples, though I don't think it comes up in this course.

Some important percentiles have their own names, including:

- Median: $p = 1/2$.

- Quartiles (lower and upper): $p = 1/4$, and $p = 3/4$, resp.

- Min: $p = 0$.

- Max: $p = 1$.

Note that the inverse cdf is sometimes called the *quantile function.*

Calculating the inverse cdf
Suppose that
$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 < x < 1 \\ 1 & x > 1 \end{cases}$$

for $0 \leq x \leq 1$. Find the inverse distribution function F^{-1}.

Solution:

First, let's check that this is a valid distribution function. First, $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$.

Now we note that it is trivially monotonically increasing from $(-\infty, 0)$ and $[1, \infty)$. Now on $[0, 1)$, x^2 is also increasing, and hence we have $F(x)$ is a monotonically increasing function.

Finally, we need to check that it is right-continuous. In this case it is trivial, because the only points of potential discontinuity are $x = 0$ and $x = 1$, but the limit clearly exists and equals the function value at both of these points.

Now to find the inverse function, we will focus on the more interesting part of the function, and solve $y = F(x) = x^2$ for x, obtaining $x = F^{-1}(y) = \sqrt{y}$. This provides the inverse function for all points in $(0, 1)$, but what about the endpoints? These are not unique, so we take:

$$F^{-1}(0) = \inf\{x \in \mathbb{R} : F(x) \geq 0\} = \inf_x [0, \infty) = 0,$$

and

$$F^{-1}(1) = \inf\{x \in \mathbb{R} : F(x) \geq 1\} = \inf_x [1, \infty) = 1.$$

Fortunately, these points already match what we found in the mid-point with the square-root function: $\sqrt{0} = 0$ and $\sqrt{1} = 1$ (you could have guessed this would happen since the function is always continuous), so we get

$$F^{-1}(p) = \sqrt{p}.$$

Thus, the inverse function defined on the interval $[0, 1]$ is

Final remarks

- From Theorem 2.2, we have the equivalency between cdf's and distributions. Naturally, we want to extend this to pdf's as well, but there is a subtle distinction that must be made.

- Namely, if two pdf's are the same, then the random variables have the same distribution. However, two random variables can have the same distribution but different pdf's, but the difference can only occur on sets with probability 0.

- This is a consequence of the definition of a pdf, which implies that a pdf may not be unique.

- We have introduce some concepts of random variables, but a full rigorous discussion about random variables requires background in measure theory. If you are interested in learning more, a good textbook for a statistics student is Resnick (2019); this is considered a graduate level text, and some background in analysis will be beneficial.

- We have discussed only discrete and continuous random variables. In practice, we often run into random variables that have both a discrete and continuous component. For instance, consider a zero-inflated continuous random variable X, where $X = 0$ with probability p (discrete component), but $X \sim N(0, 1)$ with probability $1 - p$ (continuous component).

Acknowledgments

- Compiled on September 8, 2025 using R version 4.5.1.

- Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin.

- We acknowledge students and instructors for previous versions of this course / slides.

References

Casella G, Berger R (2024). *Statistical inference*. Chapman and Hall/CRC. 23

Resnick S (2019). *A probability path*. Springer. 26

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 1, 5

Wikipedia contributors (2025a). "Empirical distribution function — Wikipedia, The Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Empirical_distribution_function&oldid=1300940691. [Online; accessed 14-August-2025]. 25

Wikipedia contributors (2025b). "Poisson distribution." Accessed 14 August 2025, URL https://en.wikipedia.org/wiki/Poisson_distribution. 1