Probability and Distributions
(Hogg Chapter One)

STAT 405-01: Mathematical Statistics I *

Fall Semester 2015

Contents

Tuesday 25 August 2015
– Read Sections 1.1-1.5 of Hogg

0 Preliminaries

0.1 Administrata

- Introductions!
- Mathematical Diagnostic (not graded; intended as a pedagogical guide for me).
- Syllabus
- Instructor's name (Whelan) rhymes with "wailin'".
- Text: Hogg, McKean, and Craig, *Introduction to Mathematical Statistics*, 7th edition.
- Other useful books:

  - Casella and Berger, *Statistical Inference*, 2nd edition. This is a standard first-year graduate text in statistics. It covers roughly the same material, but with a little more sophistication (more possible pathologies are mentioned) but also more of a practical philosophy.
  - Jaynes, *Probability Theory: the Logic of Science*. This is a sort of Bayesian manifesto and as such doesn't overlap much with the traditional approach, but it's got a lot of interesting bits in it, such as a demonstration that you can derive probability as an obvious extension of logic.
- Course website: `http://ccrg.rit.edu/~whelan/STAT-405/`
  - Will contain links to notes and problem sets; course calendar is probably the most useful.
  - Course calendar: *tentative* timetable for course.
- Course work:
  - Please read the relevant sections of the textbook *before* class so as to be prepared for class discussions.
  - There will be quasi-weekly homeworks. Collaboration is allowed an encouraged, but please turn in your own work, as obviously identical homeworks may not receive credit.
  - There will be two prelim exams, in class, and one cumulative final exam.
- Grading:

  5% Class Participation
  20% Problem Sets
  20% First Prelim Exam
  20% Second Prelim Exam
  35% Final Exam

  You'll get a separate grade on the "quality point" scale (e.g., 3.1667–3.5 is the B+ range) for each of these five components; course grade is weighted average.

0.2 Outline

1. Random Variables (Chapter One)
2. Multivariate Distributions (Chapter Two)
3. Specific probability distributions (binomial, Poisson, normal, χ^2 (Chapter Three)
4. Statistical Inference (Chapter Four)
5. Central limit theorem (Chapter Five)

Warning: the material in this class is rather advanced. Please make sure you're familiar with what you covered in Applied Statistics or Engineering Statistics, as well as multi-variable calculus.

1 Review of Probability Theory

Sections 1.1-1.4 of Hogg build up probability theory in a somewhat formal and axiomatic way, and in particular they develop the formalism of set theory which is used to combine events. Since you should already be familiar with these principles from Probability or Applied Statistics, and since most of this course is concerned with random variables rather than abstract probability, we'll just take a quick refresher, and make a few points by way of perspective.

1.1 Set Theory and Logic

At its heart, probability theory applies to each "event" a real number between zero and one. There are two different, math-

ematically equivalent, ways to understand what's meant by "event": one based on set theory and the other based on logic. Mathematics books tend to define things in the set theory way, in which there is a **sample space**, $\mathcal{C}$, consisting of all of the possible **outcomes** of an experiment, with an individual outcome labelled c, so that $c \in \mathcal{C}$. Then an event C is some set of outcomes, i.e., a subset of $\mathcal{C}$, $C \subseteq \mathcal{C}$.

In the application of probability theory, though, it's easier to think of events as being logical propositions, statements about the outcome of an experiment which could be true or false. The idea connecting the two is that for each outcome of the experiment, a given statement is either true or false. The set associated with an event C is the set of all outcomes in which the statement in question is true. (We'll be making these statements in the context of a repeatable experiment, but in fact the entire mathematical formalism can be extended to any true/false statements that can be made about the world.)

There are basic operations in set theory used to combine events into other events, and each one of them has an analogue in the formalism of logic:

- The **complement** C^c of the event C is the set of all outcomes which are not in C. In terms of logic, this is "not C", written as $\neg C$, C' or $\overline{C}$. It is a statement which is true if C is false and false if C is true.
- The **union** $C_1 \cup C_2$ is the set of all outcomes which are in C_1 or C_2, or both. In logic, this is "C_1 or C_2" (where we mean an "inclusive or"), written $C_1 \vee C_2$, a statement which is true if either C_1 or C_2 or both are true.
- The **intersection** $C_1 \cap C_2$ is the set of all outcomes which are in both C_1 and C_2. In logic, this is "C_1 and C_2", written $C_1 \wedge C_2$, a statement which is true if both C_1 and C_2 are true.

We can connect the two ideas using a "truth table"

Outcome	C_1	C_2	$C_1 \wedge C_2$	$C_1 \vee C_2$	$\neg C_1$
$C_1 \cap C_2$	T	T	T	T	F
$C_1 \cap C_2^c$	T	F	F	T	F
$C_1^c \cap C_2$	F	T	F	T	T
$C_1^c \cap C_2^c$	F	F	F	F	T

In set theory, two events are equal if the sets of outcomes they contain are identical; in logic, this corresponds to one event being true whenever the other is true and false whenever the other is false. Finally, it's useful to define the null event $\varnothing = \mathcal{C}^c$ which contains no outcomes and therefore is always false.

1.2 Defining and Assigning Probabilities

Mathematically speaking, probability is a number between 0 ans 1 which is assigned to each event. I.e., the event C has probability $P(C)$. If we think about the logical definition of events, then we have

- $P(C) = 1$ means the statement corresponding to C is definitely true.
- $P(C) = 0$ means the statement corresponding to C is definitely false.
- $0 < P(C) < 1$ means the statement corresponding to C could be true or false.

The standard numerical interpretation of the probability $P(C)$ is in terms of a repeatable experiment with some random element. Imagine that we repeat the same experiment over and over again many times under identical conditions. In each iteration of the experiment (each game of craps, sequence of coin flips, opinion survey, etc), a given outcome or event will represent a statement that is either true or false. Over the long run, the fraction of experiments in which the statement is true will be approximately

given by the probability of the corresponding outcome or event. If we write the number of repetitions of the experiment as N, and the number of experiments out of those N in which C is true as $\#C$, then

$$\lim_{N\to\infty} \frac{\#C}{N} = P(C) \tag{1.1}$$

You can test this proposition on the optional numerical exercise on this week's problem set. This interpretation of probability is sometimes called the "frequentist" interpretation, since it involves the relative frequency of outcomes in repeated experiements. It's actually a somewhat more limited interpretation than the "Bayesian" interpretation, in which the probability of an event corresponds to a quantitative degree of certainty that the corresponding statement is true. (Devore somewhat pejoratively calls this "subjective probability".) These finer points are beyond the scope of this course, but if you're interested, you may want to look up e.g., *Probability Theory: The Logic of Science* by E. T. Jaynes.

1.3 Basic Rules of Probability

It's a standard approach to develop a formal theory of probability starting from a few axioms, and derives other sensible results from those. This is an interesting intellectual exercise, but for our purposes, it's enough to note certain simple properties which make sense for our understanding of probability as the likelihood that a statement is true:

1. For any event C, $0 \le P(C) \le 1$
2. $P(\mathcal{C}) = 1$ and $P(\varnothing) = 0$ (something always happens)
3. $P(C^c) = 1 - P(C)$ (the probability that a statement is false is one minus the probability that it's true.
4. If C_1 and C_2 are disjoint events, $P(C_1 \cup C_2) = P(C_1) + P(C_2)$

One useful non-trivial result concerns the probability of the union of any two events. Since $C_1 \cup C_2 = (C_1 \cap C_2^c) \cup (C_1 \cap C_2) \cup (C_1^c \cap C_2)$, the union of three disjoint events,

$$P(C_1 \cup C_2) = P(C_1 \cap C_2^c) + P(C_1 \cap C_2) + P(C_1^c \cap C_2) \tag{1.2}$$

On the other hand, $C_1 = (C_1 \cap C_2^c) \cup (C_1 \cap C_2)$ and $C_2 = (C_1 \cap C_2) \cup (C_1^c \cap C_2)$, so

$$P(C_1) = P(C_1 \cap C_2^c) + P(C_1 \cap C_2) \tag{1.3a}$$
$$P(C_2) = P(C_1 \cap C_2) + P(C_1^c \cap C_2) \tag{1.3b}$$

which means that

$$\begin{aligned} P(C_1) + P(C_2) &= P(C_1 \cap C_2^c) + 2P(C_1 \cap C_2) + P(C_1^c \cap C_2) \\ &= P(C_1 \cup C_2) + P(C_1 \cap C_2) \end{aligned} \tag{1.4}$$

so

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2) \tag{1.5}$$

Another important concept is conditional probability, but we'll postpone consideration of this until we talk about conditional distributions for random variables.

2 Random Variables

In this course we'll primarily be interested in random variables. A random variable (or rv for short) X assigns exactly one value $X(c)$ to each outcome c in the sample space $\mathcal{C}$. Thus the probability function, which assigns a number between 0 and 1 to each event $C \in \mathcal{B}$, which is a set of outcomes, can be used to assign a probability to any set of numbers D, i.e.,

$$P_X(D) = P[X \in D] = P[\{c : X(c) \in D\}] \tag{2.1}$$

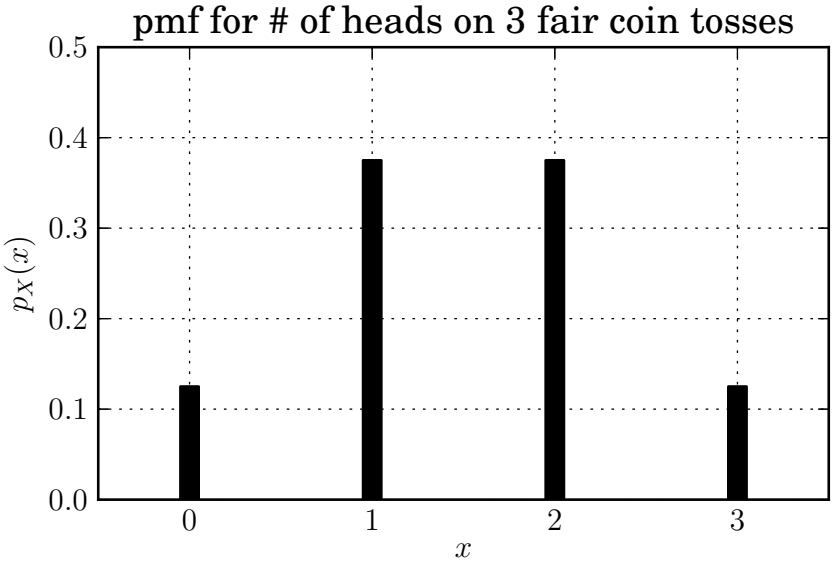
It would be kind of unwieldy if, for each random variable, we had to write down the general function which told us the probability of any set of values. Sometimes we can just write down the probability for each possible value the rv can take on; this function $p_X(x) = P[X = x]$ is called the **probability mass function** (pmf). To be acceptable as a pmf, a function has to satisfy two conditions:

1. $0 \leq p_X(x) \leq 1$ (since $p_X(x) = P[X = x]$ is a probability)
2. $\sum_{x \in \mathcal{D}} p_X(x) = 1$ where $\mathcal{D}$ is the set of allowable values of X. This condition is known as **normalization**

As an example, consider a simple random variable. Flip three coins and count the number of heads. In this case, there are only eight outcomes in the sample space: $\mathcal{C} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. If the coin is fair, the probability of each is $\frac{1}{8} = 0.125$. We can then count the heads in each outcome, defining the random variable X using the function $P(HHH) = 0.125$, $P(HHT) = 0.125$, $P(HTH) = 0.125$, etc, so that e.g., $P_X(\{1\}) = P[\{c : X(c) \in \{HTT, THT, HTT\}]$:

$$p_X(x) \equiv P(X = x) = \begin{cases} 0.125 & x = 0 \\ 0.375 & x = 1 \\ 0.375 & x = 2 \\ 0.125 & x = 3 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

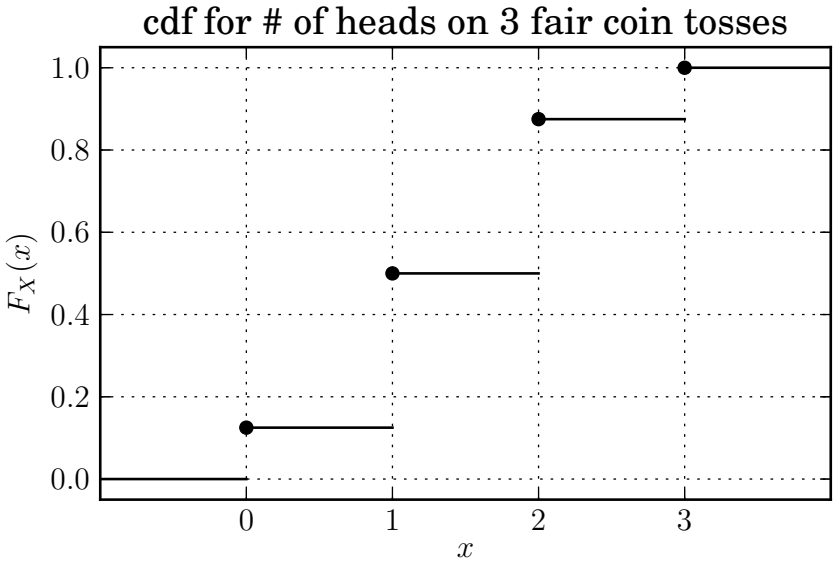
We can plot this:



As we'll see, it's not always possible to define a probability mass function, so it's useful to define a more generally applicable construct, the **cumulative distribution function**

$$F_X(x) \equiv P(X \leq x) = \begin{cases} 0 & x < 0 \\ 0.125 & 0 \leq x < 1 \\ 0.5 & 1 \leq x < 2 \\ 0.875 & 2 \leq x < 3 \\ 1 & 3 \leq x \end{cases} \tag{2.3}$$

which looks like this



Note that the cdf $F_X(x)$ is not continuous, but it is right continuous, i.e., $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$.

Thursday 27 August 2015 – Read Sections 1.6-1.7 of Hogg

2.1 The cumulative distribution function

Recall that last time, we defined the **cumulative distribution function**

$$F_X(x) = P[X \leq x] \tag{2.4}$$

Note that x is an ordinary variable, and is used to refer to a value X can take on. $F_X(x)$ need not be a continuous function, but it is always *right continuous*, i.e., $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$. This is one of the basic properties that $F_X(x)$ always has:

1. $F_X(x)$ is a non-decreasing function, i.e., for any a and b with $a < b$, $F_X(a) \leq F_X(b)$
2. $F_X(-\infty) = \lim_{x \to -\infty} F_X(x) = 0$.
3. $F_X(\infty) = \lim_{x \to \infty} F_X(x) = 1$.
4. $F_X(x)$ is right continuous, i.e., $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$.

We can use the cdf to calculate the probability of any set of values, including an interval, that X can take on. First, for a half-open interval, we note that, if $a < b$, we have $X \leq b = [X \leq a] \cup [a < X \leq b]$ and so

$$\begin{aligned} F(b) = P(X \leq b) &= P(X \leq a) + P(a < X \leq b) \\ &= F(a) + P(a < X \leq b) \end{aligned} \tag{2.5}$$

and

$$P(a < X \leq b) = F(b) - F(a) \tag{2.6}$$

On the other hand, the probability of a single value a is given by the size of the discontinuity in $F_X(x)$ at a:

$$\begin{aligned} P(X = a) &= \lim_{x \uparrow a} P(x < X \leq a) = \lim_{x \uparrow a} [F_X(x) - F_X(a)] \\ &= F_X(a) - \lim_{x \uparrow a} F_X(a) = F_X(a) - F_X(a-) \end{aligned} \tag{2.7}$$

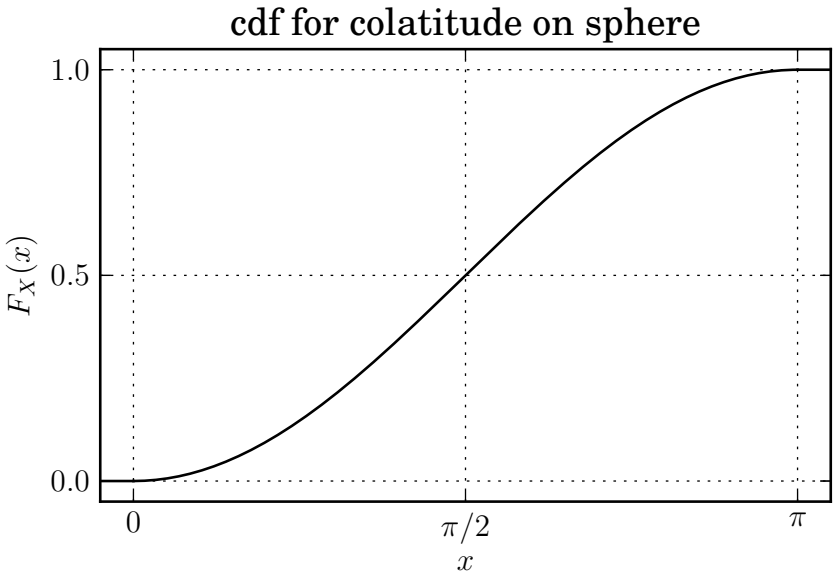
So the cdf of a random variable is equivalent to the probability function associated with it, at least if the events we're allowed to consider are limited to sets and intervals of values.

If the cdf of a random variable is continuous, we call it a **continuous random variable**. If it is constant aside from discontinuities, we call it a **discrete random variable**. We'll consider some specific examples to see how each of these can be treated specially:

As an example of a continuous random variable, pick a point at random on the sphere, and note its colatitude. (This is the angle down from the north pole, which American mathematicians refer to as ϕ, and everyone else refers to as θ. We'll call it the random variable X.) Now we cannot say the probability

that X is exactly 0, or $\pi/2$, or some single value, but we can say that the probability of being within some angle x of the north pole is proportional to the area within polar cap down to colatitude x, This means that

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2}[1 - \cos(x)] & 0 \le x < \pi \\ 1 & \pi \le x \end{cases} \tag{2.8}$$



Note that in this case the cdf is continuous. By (2.7), this means that the probability of X taking on any one value like 0 or $\pi/4$ is zero. But of course we have a non-zero probability of X lying within an interval,

$$P(a < X < b) = P(a < X \le b) = F_X(b) - F_X(a) \tag{2.9}$$

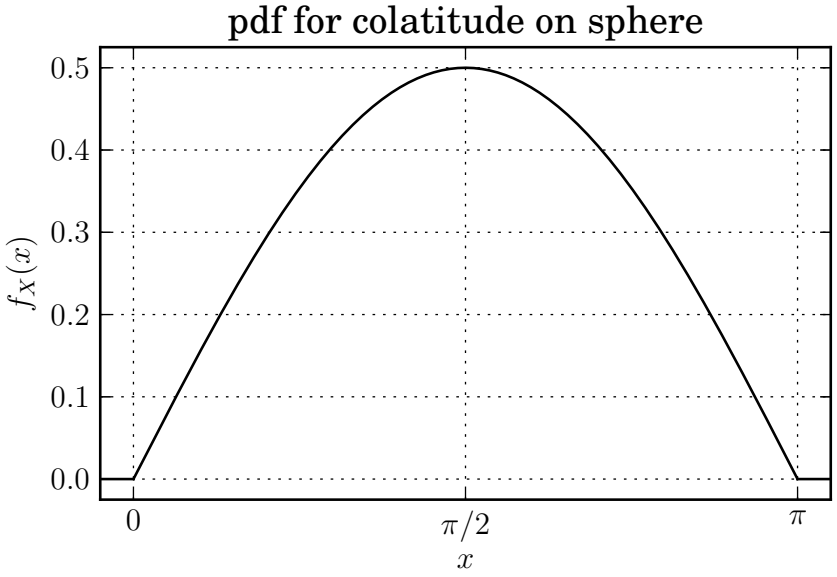
Since $F_X(x)$ is continuous, we can define its derivative $f_X(x) = F'_X(x)$, which is known as the **probability density function** for the random variable X. The fundamental theorem of calculus then tells us

$$P(a < X < b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx \tag{2.10}$$

In this case,

$$f_X(x) = \begin{cases} \frac{1}{2}\sin(x) & 0 < x < \pi \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

which looks like

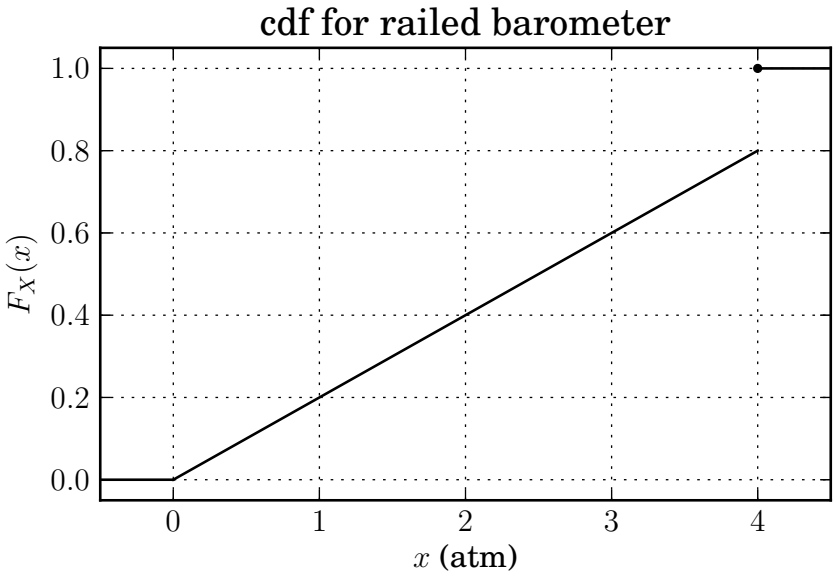


Note that some rvs are neither continuous nor discrete. Consider the case where the pressure in a chamber is a random variable equally likely to be anywhere between 0 and 5 atm, but we measure it with a gauge which rails at 4 atm. Then the cdf

of the measurement X is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.8x & 0 < x \le 4\,\text{atm} \\ 1 & 4\,\text{atm} \le x \end{cases} \tag{2.12}$$

which has a discontinuity at $x = 4\,\text{atm}$ but is not constant for $0 < x < 4\,\text{atm}$:



Incidentally, you can still write down a probability density function for this random variable in terms of the dirac delta function $\delta(x - x_0)$ which is defined by $\int_a^b g(x)\delta(x - x_0)\,dx = g(x_0)$ when $a < x_0 < b$ and $g(x)$ is sufficiently well-behaved at $x = x_0$:

$$f_X(x) = \begin{cases} 0.8x + 0.2\delta(x - 4\,\text{atm}) & 0 \le x \le 4\,\text{atm} \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

2.2 Transformations

Suppose you have a random variable X, whose properties are described by a cdf $F_X(x)$ and possibly either a pmf $p_X(x)$ or a pdf $f_X(x)$. An important question is, if you define a new rv $Y = g(X)$ using a function $g(x)$, what is its cdf $F_Y(y)$ and, if applicable, pmf $p_Y(y)$ or pdf $f_Y(y)$. We assume for simplicity that h is single-valued and invertable, so that we can define a function $x = g^{-1}(y)$.

2.2.1 Transformation of the cdf and pmf

Both the cumulative distribution function $F_X(x) = P(X \le x)$ and the probability mass function $p_X(x) = P(X = x)$ are probabilities, so their transformation is pretty straightforward. For a discrete rv, the pmf transforms as follows:

$$\begin{aligned} p_Y(y) &= P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) \\ &= p_X(g^{-1}(y)) \end{aligned} \tag{2.14}$$

Things are a little more complicated for the cdf, since we have to deal with the inequality $g(X) \le y$. If $g(x)$ is an invertable function, it must be either monotonically increasing or monotonically decreasing. If it's monotonically increasing, $g(X) \le y$ is equivalent to $X \le g^{-1}(y)$, and

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned} \tag{2.15}$$

If it's monotonically decreasing, $g(X) \le y$ is equivalent to $X \ge g^{-1}(y)$, and

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(g(X) \le y) = P(X \ge g^{-1}(y)) \\ &= 1 - P(X < g^{-1}(y)) = 1 - \lim_{x \uparrow g^{-1}(y)} F_X(x) \end{aligned} \tag{2.16}$$

If X is a continuous rv, this becomes $F_Y(y) = 1 - F_X(g^{-1}(y))$.

2.2.2 Transformation of the pdf

Things become a little more interesting if X is a continuous random variable described by a pdf $f_X(x)$. This is because the pdf is not a probability but a probability density, having been defined as the derivative of the cdf. An easy way to remember the answer (which we'll derive carefully in a minute) is to think of the densities as $f_X(x) \sim \frac{dP}{dx}$ and $f_Y(y) \sim \frac{dP}{dy}$. Then the chain rule looks something like this:

$$\frac{dP}{dy} \sim \frac{dP}{dx}\frac{dx}{dy} \sim \frac{dP/dx}{dy/dx} \tag{2.17}$$

The one catch is that the densities $f_X(x)$ and $f_Y(y)$ are both supposed to be positive, so the factor relating them has to be the absolute value of the derivative of the transformation, $\left|\frac{dy}{dx}\right|$.

Again, we have to look separately at the case where the transformation function $g(x)$ is monotonically increasing and monotonically decreasing. If it's monotonically increasing (so that $g'(x) > 0$ over the region of interest), we can use the chain rule to show

$$\begin{aligned} f_Y(y) = F'_Y(y) &= \frac{d}{dy}F_X(g^{-1}(y)) \\ &= F'_X(g^{-1}(y))\frac{d}{dy}(g^{-1}(y)) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} \end{aligned} \tag{2.18}$$

if it's monotonically decreasing (so that $g'(x) < 0$), the calculation becomes

$$\begin{aligned} f_Y(y) = F'_Y(y) &= \frac{d}{dy}[1 - F_X(g^{-1}(y))] \\ &= -F'_X(g^{-1}(y))\frac{d}{dy}(g^{-1}(y)) = \frac{f_X(g^{-1}(y))}{-g'(g^{-1}(y))} \end{aligned} \tag{2.19}$$

We can summarize (2.18) and (2.19) with the single equation

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|} = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| \tag{2.20}$$

This is the more precise version of the expression for which (2.17) is a good mnemonic.

To return to our colatitude example, where X has the pdf

$$f_X(x) = \begin{cases} \frac{1}{2}\sin(x) & 0 < x < \pi \\ 0 & \text{otherwise} \end{cases} \tag{2.21}$$

consider the transformation $Y = g(X) = \cos X$. The derivative is $g'(x) = -\sin x$ which is negative over the interval of interest $0 < x < \pi$. Note that if $0 < x < \pi$, then $-1 \le g(x) < 1$, so the transformed pdf vanishes unless $-1 < y < 1$, in which case it is

$$\frac{1}{2}\frac{\sin(\cos^{-1}(y))}{|-\sin(\cos^{-1}(y))|} = \frac{1}{2} \tag{2.22}$$

I.e.,

$$f_Y(y) = \begin{cases} \frac{1}{2} & -1 < y < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.23}$$

The fact that this happens to be a uniform distribution has useful consequences. It means we can generate a random point on the sphere by drawing the longitude from a uniform distribution between 0 and 2π and the cosine of the colatitude from a uniform distribution between -1 and 1.

Tuesday 1 September 2015
– Read Sections 1.8-1.9 of Hogg

3 Expectation Values

Given a random variable X, we define the expectation value of any function $g(X)$ as a weighted average of the possible values of $g(X)$, weighted by their probabilities. There are two analogous

expressions, one for in terms of the pmf $p(x)$ if X is discrete and one in terms of the pdf $f(x)$ if X is a continuous rv.[1]

$$E(g(X)) = \sum_x g(x)p(x) \qquad \text{if } \sum_x |g(x)|\, p(x) < \infty \tag{3.1a}$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx \qquad \text{if } \int_{-\infty}^{\infty} |g(x)|\, f(x)\,dx < \infty \tag{3.1b}$$

If the sum $\sum_x |g(x)|\, p(x)$ or integral $\int_{-\infty}^{\infty} |g(x)|\, f(x)\,dx$, which would define $E(|g(X)|)$, diverges, we say the expectation value does not exist. (This is probably something you didn't worry about introductory probability and statistics classes.) To see an example of what can go wrong, consider the Cauchy distribution, which you derived on the homework:

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \tag{3.2}$$

The trigonometric substitution $x = \tan\theta$ can be used to show that $\int_{-\infty}^{\infty} f(x)\,dx = 1$, however, the expectation value $E(X)$ does not exist, which we see by evaluating

$$\begin{aligned}\int_{-\infty}^{\infty} |x|\, f(x)\,dx &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{|x|}{1+x^2}\,dx \\ &= \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\,dx = \lim_{R\to\infty}\frac{2}{\pi}\int_0^R \frac{x}{1+x^2}\,dx \\ &= \lim_{R\to\infty}\frac{1}{\pi}\ln(1+x^2)\Big|_0^R = \lim_{R\to\infty}\frac{1}{\pi}\ln(1+R^2) = \infty\end{aligned} \tag{3.3}$$

You might think that because the Cauchy pdf $f(x)$ is an odd function, the integral of $xf(x)$ should be zero, because the contributions from positive and negative x cancel out, but that only works if you take limits of the integral to $-\infty$ and ∞ at the same rate.

To give an example where things work out, consider the uniform distribution $f(x) = \frac{1}{2}$, $-1 < x < 1$, and evaluate

$$E(X^2) = \int_{-1}^{1} x^2 f(x)\,dx = \int_{-1}^{1} \frac{x^2}{2}\,dx = \frac{x^3}{6}\Big|_{-1}^{1} = \frac{1-(-1)}{6} = \frac{1}{3} \tag{3.4}$$

Note that because the expectation value is defined using a sum or integral, it is a linear operation:

$$E[k_1 g_1(X) + k_2 g_2(X)] = k_1 E[g_1(X)] + k_2 E[g_2(X)] \tag{3.5}$$

3.1 Mean, Variance and Moments

The expectation value $E(X)$ is also known as the **mean** of the probability distribution and written $\mu_X = E(X)$. It is a special case of $E(X^m)$, which we call the m**th moment** of X. A related quantity is the m**th central moment** $E([X-\mu_X]^m)$; the second central moment is the **variance**

$$\mathrm{Var}(X) = E([X-\mu_X]^2) \tag{3.6}$$

The variance is *not* a linear operation, but the fact that the expectation value is means that we can write

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2 - 2\mu_X X - \mu_X^2) = E(X^2) - 2\mu E(X) + \mu_X^2 \\ &= E(X^2) - \mu_X^2\end{aligned} \tag{3.7}$$

so that the variance of a random variable is the second moment minus the square of the first moment.

Note that

$$\begin{aligned}\mathrm{Var}(aX+b) &= E(\{[aX+b] - [a\mu_X + b]\}^2) = E([aX - a\mu_X]^2) \\ &= E(a^2[X-\mu_X]^2) = a^2 E([X-\mu_X]^2) = a^2\,\mathrm{Var}(X)\,.\end{aligned} \tag{3.8}$$

[1] We'll drop the subscript X on the pmf, pdf, etc if it's obvious which random variable we're talking about.

3.2 The Moment Generating Function

We can calculate each moment of a random variable with an appropriate sum or integral of X^m, but there's a handy trick that lets us effectively calculate them all at once. Recall the McLaurin series

$$f(\alpha) = e^\alpha = \sum_{m=0}^{\infty} \frac{\alpha^m}{m!} \; ; \tag{3.9}$$

if we write the random variable

$$e^{tX} = \sum_{m=0}^{\infty} \frac{t^m}{m!} X^m \tag{3.10}$$

then its expectation value defines something called the **moment generating function** (mgf)

$$M_X(t) = E\left(e^{tX}\right) = \sum_{m=0}^{\infty} \frac{t^m}{m!} E\left(X^m\right) \tag{3.11}$$

As usual, we will drop the subscript X as long as it's apparent which random variable we're talking about. Considered as a function of t, we can think of this in terms of a McLaurin series whose mth coëfficient $M^{(m)}(0)/m!$ equals the mth moment divided by $m!$. I.e., if we take the mth derivative of the mgf, evaluated at $t = 0$, we get the mth moment:

$$M^{(m)}(0) = E\left(X^m\right) \tag{3.12}$$

For this to work, the mgf has to be defined in a neighborhood of the origin, i.e., for $-h < t < h$ where $h > 0$ is some positive number. Note that the "zeroth moment" of a distribution is $E(1) = 1$, so any mgf must have $M(0) = 1$

Moment generating functions can be used to generate moments, but they turn out to have lots of other useful properties which we will learn as the semester goes on. It is often easier to learn about a random variables via their mgfs than to work with the pdf, pmf or cdf directly.

As an example of an mgf, consider a uniformly distributed random variable with pdf

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

Its mgf is

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \frac{1}{b-a} \int_a^b e^{tx}\, dx = \frac{1}{b-a} \frac{e^{tx}}{t}\Big|_a^b \\ &= \frac{e^{tb} - e^{ta}}{(b-a)t} \end{aligned} \tag{3.14}$$

Actually, note that that only makes sense for $t \neq 0$, but we can show, using l'Hôpital's rule,

$$\lim_{t\to 0} \frac{e^{tb} - e^{ta}}{(b-a)t} = \lim_{t\to 0} \frac{be^{tb} - ae^{ta}}{(b-a)} = 1 \tag{3.15}$$

so the mgf

$$M(t) = \begin{cases} \frac{e^{tb}-e^{ta}}{(b-a)t} & t \neq 0 \\ 1 & t = 0 \end{cases} \tag{3.16}$$

We can find the various moments by differentiating the $t \neq 0$ expression and then using l'Hôpital's rule to take the limit as $t \to 0$.

As an example of a discrete rv, recall the pmf for the number of heads on three coin flips:

$$p(x) = \begin{cases} 1/8 & x = 0 \\ 3/8 & x = 1 \\ 3/8 & x = 2 \\ 1/8 & x = 3 \\ 0 & \text{otherwise} \end{cases} \tag{3.17}$$

Its mgf is

$$M(t) = \sum_{x=0}^{3} e^{tx} p(x) = \frac{1 + 3e^t + 3e^{2t} + e^{3t}}{8} = \frac{(1+e^t)^3}{8} = \left(\frac{1+e^t}{2}\right)^3 \tag{3.18}$$

Since

$$M'(t) = \frac{3}{2} e^t \left(\frac{1+e^t}{2}\right)^2 \tag{3.19}$$

and

$$M''(t) = \frac{3}{2} e^t \left(\frac{1+e^t}{2}\right)^2 + \frac{3}{2} e^{2t} \left(\frac{1+e^t}{2}\right) \tag{3.20}$$

we can show that the mean is

$$\mu = E(X) = M'(0) = \frac{3}{2} \tag{3.21}$$

the second moment is

$$E(X^2) = M''(0) = \frac{3}{2} + \frac{3}{2} = 3 \tag{3.22}$$

and the variance is

$$\sigma^2 = E(X^2) - \mu^2 = 3 - \frac{9}{4} = \frac{12-9}{4} = \frac{3}{4} \tag{3.23}$$

Finally, if we consider our random colatitude example,

$$f(x) = \begin{cases} \frac{1}{2}\sin(x) & 0 < x < \pi \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

the mgf is

$$M(t) = \int_0^\pi \frac{1}{2} e^{tx} \sin(x)\, dx \tag{3.25}$$

Now, the standard way to do the integral of $e^{tx}\sin(x)$ is to integrate by parts, but it's a lot simpler if we recall that $e^{ix} = \cos x + i \sin x$ so $\sin x = \operatorname{Im} e^{ix}$, and write[2]

$$M(t) = \operatorname{Im} \int_0^\pi \frac{1}{2} e^{tx} e^{ix}\, dx = \operatorname{Im} \int_0^\pi \frac{1}{2} e^{(t+i)x}\, dx = \operatorname{Im} \left. \frac{e^{(t+i)x}}{2(t+i)} \right|_0^\pi = \operatorname{Im} \frac{e^{i\pi} e^{t\pi} - 1}{2(t+i)} \tag{3.26}$$

If we recall that $e^{i\pi} = -1$, then we see

$$M(t) = -\operatorname{Im} \frac{e^{t\pi}+1}{2(t+i)} = -\operatorname{Im} \frac{e^{t\pi}+1}{2} \frac{t-i}{1+t^2} = \frac{e^{t\pi}+1}{2(1+t^2)} \tag{3.27}$$

Thursday 3 September 2015
– Read Section 1.10 of Hogg

4 Important Inequalities

Section 1.10 contains four inequalities related to expectation values. The barrage of theorem-proof-theorem-proof can be a bit daunting, so I'd like to break down why each of these works with a visual demonstration. In each case there's an inequality relating two expectation values, $E[g_1(X)] \geq E[g_2(X)]$. Since $E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$, and $f(x) \geq 0$, then if $g_1(x) \geq g_2(x)$ over all the range of x values with $f(x) > 0$, that implies $E[g_1(X)] \geq E[g_2(X)]$.

[2] We could also write $\sin x = \frac{e^{ix} - e^{-ix}}{2i}$ and proceed from there.

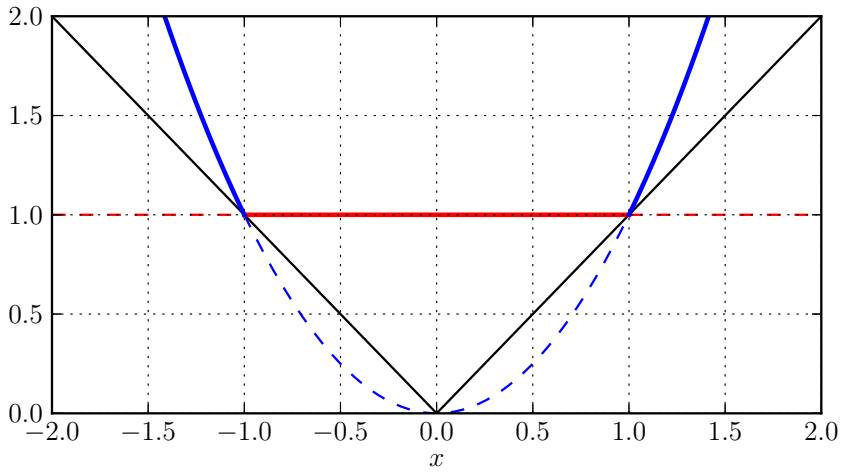
4.1 Existence of Lower Moments

If $E(X^m)$ exists for some random variable X and positive integer m, then $E(X^k)$ exists for any positive integer $k < m$.

The kth moment existing just means $\int_{-\infty}^{\infty} \left|x^k\right| \, f(x)\, dx$ is finite, so if we can show that $\left|x^k\right|$ is less than something which gives a finite result when you integrate it against $f(x)$, we're done. But we know two such quantities: 1 (because the pdf being normalized means $\int_{-\infty}^{\infty} f(x)\, dx = 1 < \infty$) and $|x^m|$ because we're assuming the mth moment exists. But $\left|x^k\right| \leq 1$ when $|x| \leq 1$ and $\left|x^k\right| \leq |x^m|$ when $|x| \geq 1$ so

$$\left|x^k\right| \leq \begin{cases} 1 & |x| \leq 1 \\ |x^m| & |x| \geq 1 \end{cases} \tag{4.1}$$

and the expression on the right-hand side gives a finite value when we integrate it against $f(x)$. We can illustrate this with $k = 1$ and $m = 2$:



4.2 Markov's Inequality

If $u(X)$ is a non-negative function of a rv X, and $E[u(X)]$ exists, then $P[u(X) > c] \leq \frac{E[u(X)]}{c}$ for any positive constant c.

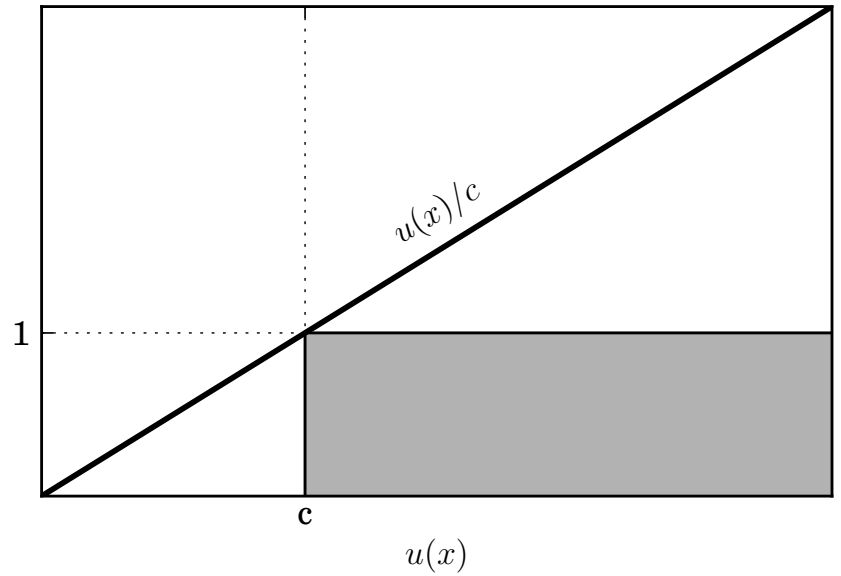
At first, this doesn't look like a statement about expectation values, but if you define a function

$$g(x) = \begin{cases} 1 & u(x) > c \\ 0 & u(x) \leq c \end{cases} \tag{4.2}$$

then

$$P[u(X) > c] = \int_{u(x)>c} f(x)\, dx = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx = E[g(X)] \tag{4.3}$$

It's pretty easy to see that $\frac{u(x)}{c} \geq g(x) \geq 0$ for all x:



so

$$E\left[\frac{u(X)}{c}\right] = \frac{1}{c} E\left[u(X)\right] \geq E[g(X)] = P[u(X) > c] \tag{4.4}$$

4.3 Chebyshev's Inequality

If X has a finite variance σ (and therefore a finite mean μ), then $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$ for any $k > 0$ (not necessarily an integer).

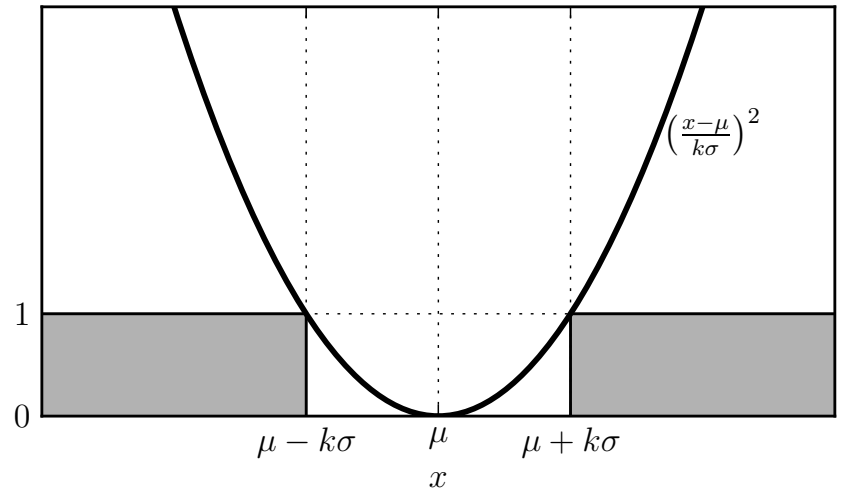
This is actually a special case of Markov's inequality, using $u(x) = (x-\mu)^2$ and $c = k^2\sigma^2$ to show that

$$\frac{1}{k^2} = \frac{E[(X-\mu)^2]}{k^2\sigma^2} \geq P[(X-\mu)^2 \geq k^2\sigma^2] = P(|X-\mu| \geq k\sigma) \tag{4.5}$$

but since Chebyshev's inequality is an important result, it's worth keeping a separate picture in mind for it. Define

$$g(x) = \begin{cases} 1 & |x-\mu| > k\sigma \\ 0 & |x-\mu| \leq k\sigma \end{cases} \tag{4.6}$$

and we can see $\left(\frac{x-\mu}{k\sigma}\right)^2 \geq g(x)$:



so that

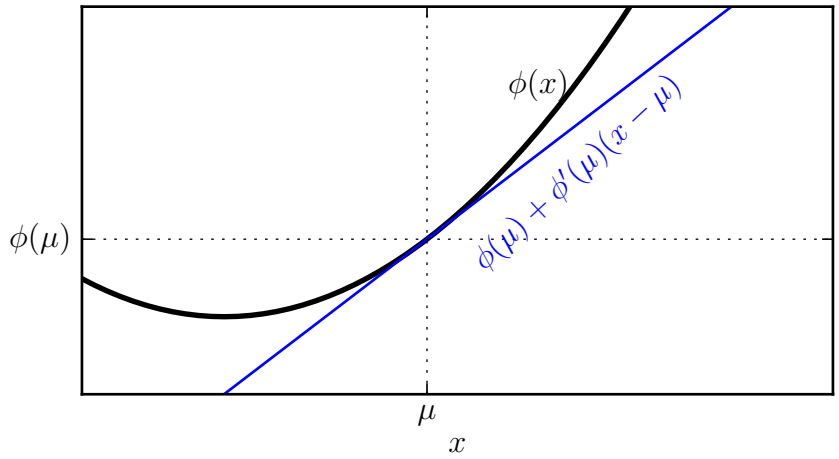
$$\frac{1}{k^2} = \frac{E[(X-\mu)^2]}{k^2\sigma^2} \geq E[g(X)] = P(|X-\mu| \geq k\sigma) \tag{4.7}$$

4.4 Jensen's Inequality

If $\phi(x)$ is a convex function over the set of possible values of X, which has a finite mean $E(X)$, then $\phi(E[X]) \leq E[\phi(X)]$.

The hardest part about this is keeping straight the definition of a convex function, which is a function such that if you connect any two points on the graph, the line will lie above the graph, or equivalently, if you make a tangent at any point, the line with that slope will lie above the graph. If the function is differentiable, this is equivalent to saying that $\phi'(x)$ never decreases as x increases, or, if it's twice differentiable, that $\phi''(x) \geq 0$.

The geometric construction is to make a tangent line to the graph of $\phi(x)$ at $x = \mu = E(X)$:



Since $\phi(x) \geq \phi(\mu) + \phi'(\mu)(x-\mu)$, we have

$$E[\phi(X)] \geq E[\phi(\mu) + \phi'(\mu)(X-\mu)] = E[\phi(\mu)] \tag{4.8}$$

where we have used the fact that $E(X) = \mu$.

A familiar manifestation of Jensen's inequality is the fact that the variance $\sigma^2 = E[X] - \mu^2 \geq 0$.