Introduction

What is statistics?

- inference from data
- building models of the world
- optimal prediction

Examples:

- mind-reading
- beating the stock market
- deciding whether a coin is fair or not

What is "mathematical" statistics (as opposed to "regular" statistics)?

The goal is to develop mathematical *theory* (theorems, approximations, etc.) for why statistical procedures work (or not)... this, in turn, leads to better procedures.

The material that we'll go over this semester mostly dates back to early part of last century. Since then, mathematical statistics has exploded, with 10^3 people working on developing theory, and many (many) more using these techniques.

Outline

1. Probability basics:

   Fundamental idea: probability distributions, parameterized by some finite number of parameters, serve as good models for observed data.

   We'll introduce a small zoo of useful probability distributions, and talk about some useful *limit theorems*, the jewels of probability theory, which permit us to make some extremely useful asymptotic simplifications of the theory, in the limit of lots of data.

2. Decision theory talks about how to behave optimally under uncertainty; this will provide us with a framework for deciding which statistical procedures are "best," or at least better than others.

3. Parameter estimation:

   Statistical inference corresponds to an inverse problem: given data, we want to answer questions about the "true" underlying state of the world, e.g., the true parameter indexing the distribution that gave rise to our observed data. Estimation is about choosing between a continuum of possible parameters.

   We'll apply ideas from decision theory to talk about how to decide between different "estimators" (that is, rules for estimating parameters from data).

   We'll also talk about the concept of "sufficiency," or "data reduction": that is, how to decide which aspects of the data matter for inference, and which aspects can be safely ignored.

   The main focus, again, will be on how to think about estimation problems *asymptotically*:

   - How do we decide if an estimator is "consistent," that is, the estimator gives you the right answer if you observe enough data?
   - How do we decide how "efficient" an estimator is? Is there a natural sense of optimal efficiency?

4. Hypothesis testing (aka classification):

   Testing and estimation are two sides of the same coin. Whereas estimation was about choosing one parameter from many, testing is about dividing the parameters into two sets, then deciding between these groups. So testing could be considered a special case of estimation, but it's special enough (and comes up often enough in practice) to warrant its own discussion and techniques.

Part I
Probability theory review

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

...The most important questions of life are indeed, for the most part, really only problems of probability.

...The theory of probabilities is at bottom nothing but common sense reduced to calculus.

Laplace, Theorie analytique des probabilites, 1820

Fate laughs at probabilities.

L.E.G. Bulwer

Statistics is the science of making inferences from probabilistic data, and as such we'll be make a great deal of use of the language of probability theory. This section will be meant as a quick and highly selective review, not as an introduction to the theory, which after centuries of work is now quite vast.

Basics: sample space, events, etc.[1]

We start with *data*. Someone (maybe you) has done an *experiment*, made observations of some stochastic process, and collected some data.

These data take values in some "sample space" Ω: this is the set of all things that can possibly happen, no matter how crazy.

We define an "event" as some set of points in sample space, i.e. a subset of what is possible. E.g.,

- this die comes up 6
- this die comes up even
- i win the lottery tomorrow
- the sky turns green a year from now

The "probability" of an event is a fairly intutive concept. We can think of probability in at least two (non-exclusive) ways:

- the average frequency of the event occurring
- one's belief that the event will happen

Let's make this more mathematically precise. We say a probability function is a scalar function on sets (i.e., you plug in a given set $A \subset \Omega$, and you get a single number out) if the function satisfies three conditions:

$$\textbf{positivity}: \ P(A) \geq 0$$

$$\textbf{normalization}: \ P(\Omega) = 1$$

$$\textbf{(sub-)additivity}: \ A_i \cap A_j = \emptyset \ \forall i \neq j \implies P(\cup_i A_i) = \sum_i P(A_i)$$

These conditions have some important implications, which you should check for yourself:

$$P(A^c) = 1 - P(A)$$

$$P(\emptyset) = 0$$

[1] See HMC 1.1-1.3. The abbreviation "HMC" refers to the text we're using, Introduction to Mathematical Statistics, by Hogg, McKean, and Craig, 6th edition.

$$P(A \cup B) = P(A) + P(B) - P(A \cup B)$$

The simplest case of this is the "flat" probability function: all events of equal "size" have equal P. (Of course, figuring out which events have equal "size" can be a matter of some subjectivity.)

The usual examples of sample spaces equipped with flat probability functions:

- fair coin flips
- fair spin of the roulette wheel
- etc.

The key point here is that computing the probability of any event you can think of here just comes down to counting:

$$P(A) = \frac{\#\omega \in A}{\#\omega \in \Omega}.$$

(Of course, life is not always fair - sometimes some equally-sized events of happen more often than others, as we'll see repeatedly...)

Conditional probability and independence[2]

What if we see that data takes value in one event, $\omega \in A$. What does this tell us about some other event, $\omega \in B$?

Assume $P(A) > 0$. Then it makes sense to define the "conditional" probability:

$$P(B|A) = P(A \cap B)/P(A).$$

This is basically just a redefinition of our original probability function, but we've now restricted our attention to the set A. (Note that if A and B are disjoint, then $P(B|A) = 0$ no matter how big $P(B)$ is.)

$P(A)$ is the probability of A *before* seeing B, so it's often called the "prior" probability of A. For similar reasons, $P(B|A)$ is called the "posterior" probability of B given A.

Another form of the above equation:

$$P(B|A)P(A) = P(A \cap B) = P(A|B)P(B).$$

This formula is important enough that people gave it a name, Bayes' rule, after Reverend Thomas Bayes (1702-1761), whose major work was discovered posthumously.

Figure 1: Bayes.

Calculations with conditional probabilities can be counterintuitive at first; hence it's important to actually get some practice on your own.

[2]HMC 1.4

Example. *Imagine a disease occurs with 0.1 % frequency in the population. Now let's say there's a blood test that comes back positive with 99 % probability if the disease is present (i.e., 99 % correct), but 2 % correct if not (i.e., 2 % false alarm rate). What is the conditional probability that someone has the disease, given a positive test result?*

Let's translate this into mathematical language.

Event $A =$ *"disease positive."*

Event $B =$ *"test positive."*

We want $P(A|B)$*, and we've been given* $P(A)$ *and* $P(B|A)$*. We use Bayes:*

$$P(A|B) = P(A \cap B)/P(B) = P(B|A)P(A)/P(B)$$

We just need

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|A^c)P(A^c) \\ &= 0.99 \cdot 0.001 + 0.02 \cdot 0.999 = 0.021. \end{aligned}$$

So

$$P(A|B) = P(B|A)P(A)/P(B) = 0.99 \cdot 0.001/0.021 = 0.047,$$

i.e., less than 5 %. □

This answer seems surprisingly low given that the test is fairly reliable. The explanation of this possibly counterintutive result is that the prior probability of the disease is so small that observing the data of the blood test only perturbs this probability slightly.

We'll see another example of a possibly counterintuitive conditional probability in the homework. (**Exercise 1**: Mr. Hall's doors, exercise 1.4.30 in HMC.)

Independence

Independence may be considered the single most important concept in probability theory, demarcating the latter from measure theory and fostering and independent development. In the course of this evolution, probability theory has been fortified by its links with the real world, and indeed the definition of independence is the abstract counterpart of a highly intuitive and empirical notion.

Chow and Teicher, 1978

Now for a concept which is again very intuitive, but turns out to have some incredibly deep implications.

We say two events A and B are "independent" if seeing A tells us nothing about B, and vice versa. More mathematically, assume $P(A), P(B) > 0$. Then A and B are independent events if

$$P(A|B) = P(A)$$

thus,

$$\begin{aligned} P(A|B)P(B) &= P(A \cap B) = P(B|A)P(A) \\ &= P(A)P(B) \end{aligned}$$

so therefore

$$P(B|A) = P(B),$$

too.

Random variables[3]

What if we look at functions of the sample space?

So if the sample space is {heads, tails}, let's make a function

$$X(heads) = 1; X(tails) = 0.$$

In a sense, the function $X(\omega)$ becomes random itself (if we pretend we can't actually see the sample space). The nice thing is that we can do some math with this random object (because variables which are real valued are much easier to deal with than heads- or tails-valued).

So we'll call a function defined on the sample space a "random variable." We'll be dealing with a *lot* of random variables this year, so we'll use the abbreviation "r.v."

Each r.v. has a "probability distribution" inherited from the probability defined on the sample space,

$$P(X \leq u) = P(\omega : X(\omega) \leq u)$$

R.v.'s can come in two flavors, discrete or continuous. (There are mixed cases as well, bu we'll usually stick to one or the other.) Discrete r.v.'s take values in a discrete set; some examples include:

- values of a coin toss, as above
- dice, lotto numbers, etc.
- number of coins I have to flip before heads comes up (note that this r.v. is unbounded — it can be arbitrarily large, in principle)

We'll also define two important functions associated with a discrete r.v.:

- the probability mass function (pmf):

$$p(u) = P(X(\omega) = u)$$

- the cumulative distribution function (cdf):

$$F(u) = \sum_{i \leq u} P(i) = P(X(\omega) \leq u)$$

[3] HMC 1.5-1.7

Continuous r.v.'s, on the other hand, can take values in a continuum, and the probability of any specific outcome is taken to be zero. Examples:

- the angle at which a merry-go-round comes to rest
- the interval of time between the arrival of two buses
- the temperature at 3 pm tomorrow

The cumulative distribution function can be defined as above:

$$F(u) = P(X(\omega) \leq u).$$

Note that this is a monotonically increasing, continuous function. Also, we have the basic formula

$$P(a < X(\omega) < b) = F(b) - F(a).$$

However, the pmf makes less sense, because for continuous r.v.'s, by definition, $P(X = u) = 0 \ \forall u$. Instead we'll define the "probability density function," or pdf, as any function $p(u) \geq 0$ such that

$$\int_{-\infty}^{\infty} f(u)du = 1$$

and

$$P(X \in A) = \int_A f(u)du.$$

Transformations of random variables[4]

A function of an r.v. is another r.v. (I.e., if $X(\omega)$ is an r.v., then so is $g(X(\omega))$, for any real function $g(.)$.) In fact, as we will see soon, passing an r.v. through some function is a very convenient and general way to come up with other r.v.'s.

The natural question is, if we know the distribution of X, how do we get the distribution of $g(X)$?

In the discrete case, this is pretty straightforward. Let's say g sends X to one of m possible discrete values. (Note that $g(X)$ must be discrete whenever X is.) To get $p_g(i)$, the pmf of g at the i-th bin, we just look at the pmf at any values of X that g mapped to i. More mathematically,

$$p_g(i) = \sum_{j \in g^{-1}(i)} p_X(j).$$

Here the "inverse image" $g^{-1}(A)$ is the set of all points x satisfying $g(x) \in A$.

Things are only slightly more subtle in the continuous case. We'll talk about two methods: one based on cdf's and one based on pdf's. The cdf of $g(X)$ is just the probability that $g(X)$ is less than or equal to u. Therefore we just need to look at $P(X \in g^{-1}(y : y \leq u))$. This is often straightforward to do.

In the second case we deal with densities instead. Here there's one slight twist that's most easily explained if we think of a very simple scaling transformation: think about the density of $g(X) = cX$, where c is a constant and $X \sim p_X(x)$. Applying the logic of the discrete approach above, we might think that the density of $Y = g(X)$ would be

$$p_Y = p_X(g^{-1}(Y)) = p_X(c^{-1}Y). \qquad \textbf{But this is wrong,}$$

as we see when we try to integrate p_Y, since

$$\int_{-\infty}^{\infty} p_Y dy = |c|.$$

So the correct density is

$$p_Y = \frac{1}{|c|} p_X(c^{-1}Y).$$

[4]HMC 1.6.1, 1.7.1

This makes sense if you think about things geometrically: multiplication by c stretches out the probability mass by a factor of $|c|$, so we have to correct for this by dividing the output density by $|c|$.

Now the case of a general (i.e., not just constant scaling) transformation $g(X)$ is conceptually straightforward: we just apply the same logic, but differentially (that is, piece by tiny piece). In this case, we get a similar formula:

$$p_Y(y) = \sum_{x \in g^{-1}(y)} \frac{p_X(x)}{|g'(x)|}.$$

Here $g'(x)$ denotes the derivative dg/dx, the slope of $g(x)$, evaluated at x.

Note that the derivative $|g'(x)|$ — the local stretching factor — plays the role that $|c|$ played above; indeed, this general formula reduces to the constant formula above when $g(X) = cX$. **Exercise 2**: Derive this formula using the transformation formula for cdf's and the chain rule for derivatives.

This transformation idea will come up over and over as we proceed — we'll constantly be applying functions to random data, e.g. to put things in a more convenient form — so it's a good idea to become very familiar with these transformation formulas early on.

Exercise 3: There was this statistics student who, when driving his car, would always accelerate hard before coming to any junction, whizz straight through it, then slow down again once he'd got past it. One day, he took a passenger, who was understandably unnerved by his driving style, and asked him why he went so fast over junctions. The statistics student replied, "Well, statistically speaking, you are far more likely to have an accident at a junction, so I just make sure that I spend less time there." If we agree with the student's logic for a moment, how much less likely is an accident as a function of the speed?

Joint, marginal, and conditional distributions[5]

Note that we can talk about more than one r.v. at once. Obviously we can simultaneously define as many functions on the sample space as we want. The natural question is, how are these r.v.'s related? Can one r.v. tell us anything about another?

It's helpful to define a couple new concepts here. First, the "joint distribution" of two r.v.'s X and Y is defined as

$$F(u_1, u_2) = P(X(\omega) \leq u_1 \cap X(\omega) \leq u_2).$$

This is like a pairwise cdf. (It's obvious how to generalize this to more than two r.v.'s at once.) We can also define joint pmf's and pdf's: for example, a joint pdf is a function $f(u_1, u_2) \geq 0$ such that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u_1, u_2) du_1 du_2 = 1$$

and

$$P(\{X, Y\} \in A) = \int_A f(u_1, u_2) du_1 du_2.$$

Now we can talk about how much X tells us about Y, because the independence of sets has a simple analog in the independence of r.v.'s. We say X and Y are independent if their joint pdf can be written as a product function,

$$f(u_1, u_2) = f_1(u_1) f_2(u_2).$$

More generally (since not every r.v. has a nice pdf), X and Y are independent if

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B).$$

R.v.'s which are not independent are called (naturally enough) "dependent" — but note that this is a mathematical definition, and that mathematically dependent r.v.'s may in fact not be directly coupled in any mechanistic way.

We can also talk about conditional probabilities related to r.v.'s: the conditional probability distribution of X given Y is

$$F(u|Y \in A) = P(X \leq u | Y \in A).$$

[5] HMC 2.1-2.3

Again, if

$$F(u|Y \in A) \neq F(u)$$

for some set A with positive probability $P(Y \in A)$, then Y tells us something about X. Conditional densities and pmf's may be defined in the natural way from the relevant distribution functions (although sometimes we might run into problems defining conditional densities, since the process involves a limit that does not always exist).

We can also invert the process. Let's say we're handed the joint distribution of X and Y, but we really only care about X. How do we get $P(X)$? We have to use the summation rule for probability,

$$P(X) = \sum_i P(X \cap \{Y = i\}).$$

I.e., for continuous r.v.'s, we have

$$F(u_1) = \int_{-\infty}^{\infty} du_2 \int_{-\infty}^{u} f(u_1, u_2) du_1.$$

For reasons which are not entirely clear to me, this process of integrating over the r.v. we don't care about is called "marginalization." So $F(u)$ is the "marginal" (or prior) distribution of X, to distinguish it from the conditional (or posterior) distribution given Y.

Expectations[6]

Say you were standing with one foot in the oven and one foot an ice bucket. According to the percentage people, you should be perfectly comfortable.

Bobby Bragan, 1963

You know how dumb the average guy is? Well, by definition, half of them are even dumber than that.

J.R. "Bob" Dobbs

To specify a pdf in general, you need to list an infinite number of characteristics (e.g., the pdf at every single point u). In many situations, it's very helpful to have a couple important descriptive numbers that quickly tell you the "important" things about the distribution.

Two of the most important such summary quantities are the "mean" and the "variance." These are defined in terms of "expectations" of the underlying r.v.: the expectation of an r.v. X with pdf $f(u)$ is defined as

$$E(X) = E(f) = \int_{-\infty}^{\infty} f(u)udu.$$

(In the case of discrete r.v.'s, just replace the integral with a sum and the pdf with the pmf.)

A very important property of the expectation is that it is *linear*:

$$E(aX + bY) = aE(X) + bE(Y),$$

for any constants a, b and any r.v.'s X, Y — *even if X and Y are dependent.* **Exercise 4**: Prove that $E(aX) = aE(X)$, using the rules for transformations of r.v.'s. (We'll prove that $E(X + Y) = E(X) + E(Y)$ soon.)

The i-th "moment" is defined as

$$E_i(f) = \int_{-\infty}^{\infty} f(u)u^i du.$$

Thus the first moment is the same as the expectation, which we also call the "mean." (You can tell this will be important, since it's been given so many

[6]HMC 1.8

names.) It's common to think of the mean as the "center of gravity" of a pdf.

The variance, on the other hand, is defined to measure the "spread" of the pdf, or how variable the underlying r.v. is. It's defined in terms of second moments:

$$V(f) = \int_{-\infty}^{\infty} f(u)(u - E(u))^2 = E_2(f) - E_1(f)^2.$$

Note that the variance is zero only if the r.v. is constant (i.e., not at all variable), and always nonnegative.

Also,

$$V(cX) = c^2 V(X),$$

for any constant c. For this reason, people often use the "standard deviation" instead, that is,

$$\sigma(X) \equiv \sqrt{V(X)},$$

since $\sigma(cX) = |c|\sigma(X)$ and is therefore slightly more useful as a measure of the "scale" of X.

We'll mention a couple other important expectations soon. For now, it's worth noting that the probability function associated with a random variable X, $P(X \in A)$, is itself a kind of expectation, if we define the random variable

$$I_A(X) = 1(X \in A);$$

then

$$P(X \in A) = E(I_A).$$

So we can go back and forth between probability and expectation as desired.

One last thing worth remembering: expectations don't always exist, or they can be infinite. We'll see examples of this soon, but the reason is clear: sometimes the integral in the definition of the expectation is either infinite or fails to converge. This situation actually does come up in practice: for example, the "fat-tailed" distributions that people sometimes use in financial models often have infinite moments.

Conditional expectations and independence

We can also define conditional expectations in the obvious way: the conditional expectation of X given Y is

$$E(X|Y) = \int_{-\infty}^{\infty} xp(X = x|Y)dx.$$

Note the important formula

$$E_Y(E_X(X|Y)) = E(X),$$

which follows by an interchange of expectations (as always, under the proviso that the relevant expectations exist).

If we recall the link between probabilities and expectations, we can also define independence of r.v.'s in terms of expectations. For example, X and Y are independent if

$$E(f(X)g(Y)) = E(f(X))E(g(Y)),$$

for any functions f and g s.t. the relevant expectations exist. Using the same logic, X and Y are independent if

$$E(f(X)|Y) = E(f(X)),$$

for any function f s.t. the relevant expectations exist. (Note that $E(X|Y) = E(X)$ is *insufficient* for independence. **Exercise 5**: Give an example of dependent X and Y such that $E(X|Y) = E(X)$.)

Correlation and covariance[7]

As we mentioned before, it's useful to define some simple quantities to summarize relevant properties of distributions. This holds true for joint distributions as well. Here, as we mentioned above, we're most interested in the relationships between the r.v.'s, and a very simple way to quantify whether one r.v. has anything to do with another is to compute their "covariance"

$$C(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(X,Y)(X-E(X))(Y-E(Y)) = E(XY)-E(X)E(Y).$$

Note that independent r.v.'s always have zero correlation, but *the converse is not true*; it's easy to think of uncorrelated but nonetheless dependent r.v.'s. Also, while the variance is always nonnegative, the covariance can be negative or positive.

A very important formula which you can derive yourself, using what we know about expectations of sums: **Exercise 6**:

$$V(X+Y) = V(X) + V(Y) + 2C(X,Y).$$

A very important special case: for independent X and Y,

$$V(X+Y) = V(X) + V(Y).$$

The "correlation" is a conveniently normalized version of the covariance:

$$\frac{C(X,Y)}{\sigma(X)\sigma(Y)}$$

In the case of more than two r.v.'s, say X_i, we can define the covariance matrix, with each element

$$C_{ij} = C(X_i, X_j).$$

Note that a covariance matrix is always symmetric: $C_{ij} = C_{ji}$.

Also, a covariance matrix is "positive semi-definite":

$$\sum_{ij} a_i C_{ij} a_j \geq 0,$$

where a_i is any set of real constants (a_i can be positive, negative, or zero). (If you've taken linear algebra, you'll remember that this means that the matrix C has nonnegative eigenvalues.) **Exercise 7**: Prove this using the rules of addition and scaling for variance, and the fact that $V(Z) \geq 0$ for any r.v. Z.

[7] HMC2.4, 2.6.1

Moment-generating functions; convolution[8]

Perhaps the most important transformation we'll encounter is multidimensional: if $X \sim p_X(x)$ and $Y \sim p_Y(y)$ are independent, what is the distribution of $X + Y$?

The easiest way to approach this is to think about the relevant cdf's.

$$\begin{aligned} P(X+Y \le u) &= \int_{X+Y \le u} P(X,Y)dXdY \\ &= \int_{-\infty}^{\infty} dX \int_{-\infty}^{u-X} P(X,Y)dY = \int_{-\infty}^{\infty} dY \int_{-\infty}^{u-Y} P(X,Y)dX \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{u-x} p_X(x)p_Y(y)dy \qquad \text{(by independence)} \end{aligned}$$

Exercise 8: Use the second formula above to prove, finally, that $E(X + Y) = E(X) + E(Y)$, even if X and Y are dependent. (If you don't feel like using the above formula, prove this using any method you want.)

It's worth noting that we can generalize this approach very easily, e.g., if we want to know the distribution of $X \cdot Y$, or X/Y; the only things that change in the above formula are the sets we are integrating over. Here are some examples for practice: **Exercise 9**: Assuming $X \sim p_X(x)$ and $Y \sim p_Y(y)$, with X and Y independent, what is the distribution of $X \cdot Y$? **Exercise 10**: What is the distribution of $\max X_i$, if N r.v.'s X_i are drawn i.i.d. from $p_X(x)$? **Exercise 11**: What is the distribution of $X_{(j)}$, the j-th smallest X_i? (This is called the j-th "order statistic" of the sample. Order statistics are helpful summaries of data; for example, the *median* — the $i/2$-th order sample — or the *range* — $X_{(N)} - X_{(1)}$.) **Exercise 12**: What is the distribution of the range?

If we differentiate this w.r.t. u we get the pdf:

$$p_{X+Y}(u) = \int_{-\infty}^{\infty} p_X(x)p_Y(u-x)dx.$$

This expression is common enough (not just in statistics, but also in physics, engineering, etc.) that it has its own name: we say $h(u)$ is the "convolution" of functions $f(u)$ and $g(u)$ if

$$h(u) = f * g(u) \equiv \int_{-\infty}^{\infty} f(x)g(u-x)dx = \int_{-\infty}^{\infty} f(u-x)g(x)dx.$$

[8] HMC 1.9

Note that convolution is a *smoothing* operation; roughness in f or g gets averaged over in h. In fact, you can prove that h is always at least as differentiable (in the sense of having k smooth derivatives) as the most differentiable of f and g. (**Exercise 13**: Prove this yourself.)

Moment-generating functions

Here we're going to introduce a trick that turns out to be more helpful than it would seem to have any right to be. Let's look at the expectation of *exponentials* of a r.v. X:

$$M_X(s) = E(e^{sX}) = \int_{-\infty}^{\infty} e^{su} p_X(u) du.$$

Or for multiple r.v.'s simultaneously,

$$M_{X_1,X_2}(s_1, s_2) = E(e^{s_1 X_1 + s_2 X_2}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{s_1 u_1 + s_2 u_2} p_{X_1,X_2}(u_1, u_2) du_1 du_2.$$

These expectations are called "moment generating functions," abbreviated mgf's. Why? Because

$$\frac{\partial^i}{\partial s^i} M_X(s)\Big|_{s=0} = E_i(X),$$

which is a handy trick if the distribution p_X is complicated but M_X is simple. (Often derivatives are easier to compute than integrals.) A similar trick works in higher dimensions: **Exercise 14**: How would you compute the covariance of X and Y in terms of $M_{X,Y}$?

Another nice feature of the mgf stems from the fact that 1) it's easy to deal with products of exponentials, and 2) independent r.v.'s have product distributions. This, in turn, means that

$$M_{X_1,X_2,\ldots}(s_1, s_2, ...) = \prod_i M_{X_i}(s_i)$$

if and only if the r.v.'s X_i are independent.

This immediately tells us something interesting about convolutions and the distribution of sums of independent r.v.'s:

$$M_{X+Y}(s) = E(e^{s(X+Y)}) = E(e^{sX+sY)}) = E(e^{sX})E(e^{sY)}) = M_X(s)M_Y(s).$$

So mgf's of sums of independent r.v.'s are ridiculously easy to deal with.

So the mgf gives us a lot of useful information about the distribution. Now the natural question is: does the mgf give us *all* the information about the distribution? That is, can we invert the mgf to uniquely find the distribution? The answer to this question turns out to be yes, as the following theorem shows:

Theorem 1. *The distribution functions F_X and F_Y are equal if and only if the corresponding mgf's $M_X(s)$ and $M_X(s)$ are equal (and finite) for all $s \in (-z, z)$, for some positive constant z.*

The proof (which we'll skip) is based on ideas from Fourier analysis. A nice hand-wavey argument for why the theorem "should" be true is as follows (read along with, e.g., HMC pp. 60-61). Let's say someone hands you

$$M(s) = ae^s + be^{-3s} + ce^{6s},$$

say, for positive constants a, b, c. Now, we know, by the definition of mgf, that

$$M(s) = \int e^{su} p(u) du.$$

How many distributions $p(u)$ can you think of that satisfy

$$\int e^{su} p(u) du = ae^s + be^{-3s} + ce^{6s},$$

for all s? It's hard to think of any others besides the one that assigns mass a to the point $u = 1$, b at $u = -3$, and c on $u = 6$.

Note that not every distribution has an mgf, unfortunately: for many distributions (e.g., those with "heavy tails") the relevant integrals will be infinite. (This inconvenient fact inspired people to try complex exponentials — Fourier transforms — which turn out to work for all distributions[9]. But the humble mgf will suit our needs adequately here.)

[9] See a more advanced book on probability theory, e.g., Breiman '68, for more information.

Example distributions[10]

In this section, finally, we'll describe a bunch of useful distributions that we'll run into again and again. You'll also notice a lot of exercises, to give you some practice computing mgf's and moments (and in the process brushing up on your calculus skills).

Univariate examples

Binomial, $Bin(N, p)$

If there is a 50-50 chance that something can go wrong, then 9 times out of ten it will.

Paul Harvey News

This is the pmf of the number of heads you get if you flip N (not necessarily fair) coins independently and identically distributed, each with probablity p of heads.

$$p(n) = \binom{N}{n} p^n (1-p)^{N-n}$$

$$E(n) = Np$$

$$V(n) = Np(1-p)$$

Exercise 15: Derive the mgf in two ways. 1) directly compute it, and 2) compute the mgf for the case $N = 1$, and then use what you know about mgf's of sums of i.i.d. r.v.'s. Does your result have the correct derivatives at zero, comparing to the formula for mean and variance above?

Exercise 16: A statistics major was poorly prepared the day of his final exam. It was a True/False test, so he decided to flip a coin for the answers. The professor watched the student the entire two hours as he was flipping the coin...writing the answer...flipping the coin...writing the answer. At the end of the two hours, everyone else had left the final except for the one student. The professor walks up to his desk and interrupts the student, saying:

[10] HMC chapter 3

"Listen, I have seen that you did not study for this statistics test; you didn't even open the exam. If you are just flipping a coin for your answer, what is taking you so long?"

The student replies bitterly, as he is still flipping the coin: "Shhh! I am checking my answers!"

What is the probability that the student will get no more than two questions wrong, if the test is k questions long?

Poisson, $Poiss(\lambda)$

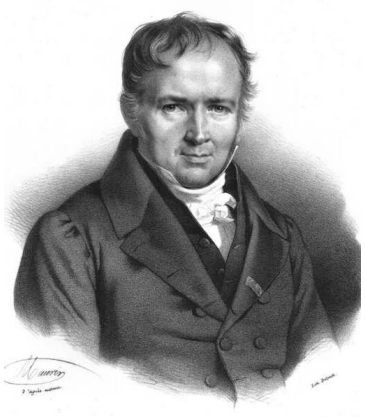

Figure 2: Poisson.

This pmf arises through a limiting process. Let's say you have N i.i.d. coins, but the more coins you flip the less likely it is that the coins come up heads. In particular, let's say that the probability of heads is λ/N. Then it turns out that the corresponding binomial distribution looks like

$$p(n) = e^{-\lambda}\frac{\lambda^n}{n!}$$

for N large enough. **Exercise 17**: Prove this.

This is our first example of a "limit theorem" — i.e., let N become large, and come up with a tractable approximation to replace the (often quite complex) large-N distribution. We'll see more examples of this.

$$E(n) = \lambda$$

$$V(n) = \lambda$$

Exercise 18: Derive the mgf. (Hint: the easiest way is through a limiting argument, using what you already know about the binomial mgf.)

Uniform, $U(a, b)$

This is the simplest continuous distribution. The density is zero outside the interval $[a, b]$, and constant ($= 1/|b - a|$) on it.

Exercise 19: Compute the mean, variance, and mgf.

Exercise 20: Compute the distribution of X^α, with α a constant and $X \sim U[a, b]$.

Gaussian (normal), $\mathcal{N}(\mu, \sigma^2)$

Steinhaus, with his predilection for metaphors, used to quote a Polish proverb, 'Forturny kolem sie tocza' [Luck runs in circles], to explain why Pi, so intimately connected with circles, keeps cropping up in probability theory and statistics, the two disciplines which deal with randomness and luck.

Mark Kac

This is the classic bell-shaped curve:

$$p(x) = (\sigma\sqrt{2\pi})^{-1} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E(x) = \mu$$

$$V(x) = \sigma^2$$

Exercise 21: Compute the mgf. Notice any similarities to the pdf?

We'll have much much more to say about the Gaussian distribution in a little bit: the Gaussian is the star of the "central limit theorem."

Figure 3: Gauss.

Exponential, $exp(\lambda)$, and geometric

This is the classic "waiting time" density (exponential) or pmf (geometric).

$$p(x) = \lambda e^{-\lambda x}$$

$$E(x) = 1/\lambda$$

Exercise 22: You know what to do: compute the variance and the mgf. Does the mgf exist for all points s?

The exponential has the interesting property that

$$p(x > a) = p(x > b + a | x > b); \tag{1}$$

in other words, the fact that you've waited longer than b minutes for the bus has absolutely no impact on how long you'll have to continue waiting. This is why the exponential distribution is often called "memoryless." (Luckily, this is a better model for radioactive decay than for buses.)

Exercise 23: Is this the only class of densities that has this property? I.e., if p satisfies property (1), then is p automatically exponential? Can you a) prove this conjecture or b) provide a counterexample (and therefore disprove the conjecture)?

Gamma

This what you get if you add i.i.d. exponential r.v.'s together.

Exercise 24: Compute the pdf. Compute the mgf. Then compute the mean and variance.

Chi-square

This is what you get when you add together squared normal distributions.

Exercise 25: Compute the pdf (in the case of one squared gaussian) by the transformation method. Then compute the mgf. Then compute the general pdf and mgf. Then compute the mean and variance.

Cauchy

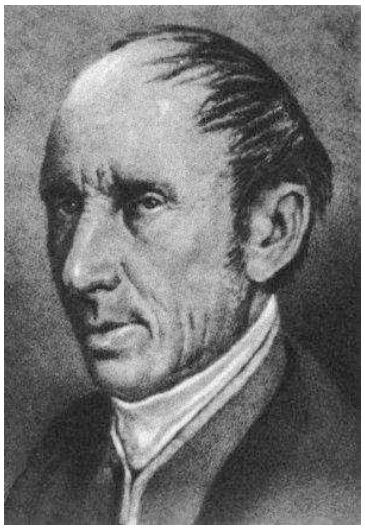

Figure 4: Cauchy.

This is what you get when you divide two i.i.d. standard Gaussian r.v.'s.

$$p(x) = \frac{1}{\pi(1+x^2)}$$

This is interesting from a mathematical point of view as a simple density with infinite variance (in fact, the mean does not exist, although clearly we can define the "center of gravity" by symmetry nonetheless).

Exercise 26: Try computing the mgf. Any problems?

Exercise 27: Try taking the average of N i.i.d. Cauchy variables. What is the resulting distribution? Interpret your result.

Multivariate examples

Multinomial

Like the binomial, except now toss N m-sided dice instead of N 2-sided coins.

$$p(n_1, n_2, ..., n_m) = \frac{N!}{\prod_{i=1}^{m} n_i!} \prod_{i=1}^{m} p_i^{n_i}$$

Note that this reduces to the binomial when $m = 2$.

Exercise 28: Compute the covariance $C(n_i, n_j)$ and the corresponding correlation coefficient.

Multivariate normal

This is the fundamental multivariate distribution, on which a large part of the theory of analysis of multivariate data is based.

$$p(\vec{x}) = (\sqrt{2\pi|C|})^{-1} e^{-\frac{1}{2}(x-\vec{\mu})^t C^{-1}(x-\vec{\mu})}$$

Here $\vec{\mu}$ is the "mean vector," and C is the covariance matrix.

Probability inequalities[11]

There is an adage in probability that says that behind every limit theorem lies a probability inequality (i.e., a bound on the probability of some undesired event happening). Since a large part of probability theory is about proving limit theorems, people have developed a bewildering number of inequalities.

Luckily, we'll only need a few key inequalities. Even better, three of them are really just versions of one another. **Exercise 29**: Can you think of example distributions for which each of the following inequalities are tight (that is, the inequalities may be replaced by equalities)? Are these "extremal" distributions unique?

Markov's inequality

Figure 5: Markov.

For a nonnegative r.v. X,

$$P(X > u) \leq \frac{E(X)}{u}.$$

So if $E(X)$ is small and we know $X \geq 0$, then X must be near zero with high probability. (Note that the inequality is *not* true if X can be negative.)

[11] HMC 1.10

The proof is really simple:

$$uP(X > u) \leq \int_u^\infty t p_X(t) dt \leq \int_0^\infty t p_X(t) dt = E(X). \quad \square$$

Chebyshev's inequality

Figure 6: Chebyshev.

$$P(|X - E(X)| > u) \leq \frac{V(X)}{u^2},$$

aka

$$P\left(\frac{|X - E(X)|}{\sigma(X)} > u\right) \leq \frac{1}{u^2}.$$

Proof: just look at the (nonnegative) r.v. $(X - E(X))^2$, and apply Markov.

So if the variance of X is really small, X is close to its mean with high probability.

Chernoff's inequality

$$P(X > u) = P(e^{sX} > e^{su}) \leq e^{-su} M(s).$$

So the mgf controls the size of the tail of the distribution — yet another surprising application of the mgf idea.

Figure 7: Chernoff.

The really nice thing about this bound is that it is easy to deal with sums of independent r.v.'s (recall our discussion above of mgf's for sums of independent r.v.'s). **Exercise 30**: Derive Chernoff's bound for sums of independent r.v.'s (i.e., derive an upper bound for the probability that $\sum_i X_i$ is greater than u, in terms of $M(X_i)$).

The other nice thing is that the bound is exponentially decreasing in u, which is much stronger than Chebyshev. (On the other hand, since not all r.v.'s have mgf's, Chernoff's bound can be applied less generally than Chebyshev.)

The other other nice thing is that the bound holds for all s simultaneously, so if we need as tight a bound as possible, we can use

$$P(X > u) \leq \inf_s e^{-su} M(s),$$

i.e., we can minimize over s.

Jensen's inequality

This one is more geometric. Think about a function $g(u)$ which is curved upward, that is, $g''(u) \geq 0$, for all u. Such a $g(u)$ is called "convex." (Downward-curving functions are called "concave." More generally, a convex

Figure 8: Jensen.

function is bounded above by its chords:

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y),$$

while a concave function is bounded below.)

Then if you draw yourself a picture, it's easy to see that

$$E(g(X)) \geq g(E(X)).$$

That is, the average of $g(X)$ is always greater than or equal to g evaluated at the average of X. **Exercise 31**: Prove this. (Hint: try subtracting off $f(X)$, where f is a linear function of X such that $g(X) - f(X)$ reaches a minimum at $E(X)$.)

Exercise 32: What does this inequality tell you about the means of $1/X$? of $-X \log X$? About $E_i(X)$ vs. $E_j(X)$, where $i > j$?

Cauchy-Schwarz inequality

$$|C(X,Y)| \leq \sigma(X)\sigma(Y),$$

that is, the correlation coefficient is bounded between -1 (X and Y are anti-correlated) and 1 (correlated).

The proof of this one is based on our rules for adding variance:

$$C(X,Y) = \frac{1}{2}[V(X+Y) - V(X) - V(Y)]$$

(assuming $E(X) = E(Y) = 0$). **Exercise 33**: Complete the proof. (Hint: try looking at X and $-X$, using the fact that $C(-X,Y) = -C(X,Y)$.)

Exercise for the people who have taken linear algebra: interpret the Cauchy-Schwarz inequality in terms of the angle between the vectors X and Y (where we think of functions — that is, r.v.'s — as vectors, and define the dot product as $E(XY)$ and the length of a vector as $\sqrt{E(X^2)}$). Thus this inequality is really geometric in nature.

Limit theorems[12]

An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.

The first golden rule of applied mathematics, sometimes attributed to John Tukey

(Weak) law of large numbers

Chebyshev's simple inequality is enough to prove perhaps the fundamental result in probability theory: the law of averages. This says that if we take the sample average of a bunch of i.i.d. r.v.'s, the sample average will be close to the true average. More precisely, under the assumption that $V(X) < \infty$, then

$$P\left(|E(X) - \frac{1}{N}\sum_{i=1}^{N} X_i| > \epsilon\right) \to 0$$

as $N \to \infty$, no matter how small ϵ is.

The proof:

$$E\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = E(X),$$

by the linearity of the expectation.

$$V\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{V(X)}{N},$$

by the rules for adding variance and the fact that X_i are independent.

Now just look at Chebyshev. □

Remember, the LLN does not hold for all r.v.'s: remember what happened when you took averages of i.i.d. Cauchy r.v.'s? **Exercise 34**: What goes wrong in the Cauchy case?

Stochastic convergence concepts

In the above, we say that the sample mean $\frac{1}{N}\sum_{i=1}^{N} X_i$ "converges in probability" to the true mean. More generally, we say r.v.'s Z_N converge to Z in

[12] HMC chapter 4

probability, $Z_N \to_P Z$, if

$$P(|Z_N - Z| > \epsilon) \to 0$$

as $N \to \infty$. (The weak LLN is called "weak" because it asserts convergence in probability, which turns out to be a somewhat "weak" sense of stochastic convergence, in the mathematical sense that there are "stronger" forms of convergence — that is, it's possible to find sequences of r.v.'s which converge in probability but not in these stronger senses. In addition, it's possible to prove the LLN without assuming that the variance exists; existence of the mean turns out to be sufficient. But discussing these stronger concepts of convergence would take us too far afield[13]; convergence in probability will be plenty strong enough for our purposes.)

We discussed convergence of r.v.'s above; it's often also useful to think about convergence of distributions. We say a sequence of r.v.'s with cdf's $F_N(u)$ "converge in distribution" if

$$\lim_{N\to\infty} F_N(u) \to F(u)$$

for all u such that F is continuous at u (here F is itself a cdf). **Exercise 35**: Explain why do we need to restrict our attention to continuity points of F. (Hint: think of the following sequence of distributions: $F_N(u) = 1(u < 1/N)$, where the "indicator" function of a set A is one if $x \in A$ and zero otherwise.)

It's worth emphasizing that convergence in distribution — because it only looks at the cdf — is in fact weaker than convergence in probability. For example, if p_X is symmetric, then the sequence $X, -X, X, -X, ...$ trivially converges in distribution to X, but obviously doesn't converge in probability. **Exercise 36**: Prove that convergence in probability actually is stronger, that is, implies convergence in distribution.

Central limit theorem

The second fundamental result in probability theory, after the LLN, is the CLT: if X_i are i.i.d. with mean zero and variance 1, then

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \to_D \mathcal{N}(0,1),$$

[13] Again, see e.g. Breiman '68 for more information.


Figure 9: De Moivre and Laplace.

where $\mathcal{N}(0,1)$ is the standard normal density. More generally, the usual rescalings tell us that

$$\frac{1}{\sigma(X)\sqrt{N}} \sum_{i=1}^{N} (X_i - E(X)) \rightarrow_D \mathcal{N}(0,1).$$

Thus we know not only that (from the LLN) the distribution of the sample mean approaches the degenerate distribution on $E(X)$, but moreover (from the CLT) we know exactly what this distribution looks like, asymptotically, if we take out our magnifying glass and zoom in on $E(X)$, to a scale of $N^{-1/2}$. In this sense the CLT is a stronger result than the WLLN: it gives more details about what the asymptotic distribution actually looks like.

One thing worth noting: keep in mind that the CLT really only tells us what's going on in the local neighborhood $(E(X) - N^{-1/2}c, E(X) + N^{-1/2}c)$ — think of this as the mean plus or minus a few standard deviations. But this does *not* imply that, say,

$$P(\frac{1}{N}\sum_{i=1}^{N} X_i \leq -\epsilon) \sim \int_{\infty}^{-\epsilon} \mathcal{N}(0, \frac{1}{N})(x)dx = \int_{\infty}^{-\sqrt{N}\epsilon} \mathcal{N}(0,1)(x)dx \quad \textbf{not true};$$

a different asymptotic approximation typically holds for the "large devia-

tions," the tails of the sample mean distribution[14].

More on stochastic convergence

So, as emphasized above, convergence in distribution can drastically simplify our lives, if we can find a simple approximate (limit) distribution to substitute for our original complicated distribution. The CLT is the canonical example of this; the Poisson theorem is another. What are some general methods to prove convergence in distribution?

Delta method

The first thing to note is that if X_N converge in distribution or probability to a constant c, then $g(X_N) \to_D g(c)$ for any continuous function $g(.)$. **Exercise 37**: Prove this, using the definition of continuity of a function: a function $g(u)$ is continuous at u if for any possible fixed $\epsilon > 0$, there is some (possibly very small) δ such that $|g(u+v) - g(u)| < \epsilon$, for all v such that $-\delta < v < \delta$. (If you're having trouble, just try proving this for convergence in probability.)

So the LLN for sample means immediately implies an LLN for a bunch of functions of the sample mean, e.g., if X_i are i.i.d. with $V(X) < \infty$, then

$$\left(\prod_{i=1}^{N} e^{X_i} \right)^{1/N} = e^{\frac{1}{N}\sum_{i=1}^{N} X_i} \to_P e^{E(X)},$$

(which of course should not be confused with $E(e^X)$; in fact, **Exercise 38**: Which is greater, $E(e^X)$ or $e^{E(X)}$? Give an example where one of $E(e^X)$ or $e^{E(X)}$ is infinite, but the other is finite).

We can also "zoom in" to look at the asymptotic distribution (not just the limit point) of $g(Z)$, whenever g is sufficiently smooth. For example, let's say $g(.)$ has a Taylor expansion at u,

$$g(z) = g(u) + g'(u)(z-u) + o(|z-u|), \ |z-u| \to 0,$$

where $|g'(u)| > 0$ and $z = o(y)$ means $z/y \to 0$. Then if

$$a_N(z_N - u) \to_D q,$$

[14]See e.g., Large deviations techniques and applications, Dembo and Zeitouni '93, for more information.

for some limit distribution q and a sequence of constants $a_N \to \infty$ (think $a_N = N^{1/2}$, if Z_N is the sample mean), then

$$a_N \frac{g(Z_N) - g(u)}{g'(u)} \to_D q,$$

since

$$a_N \frac{g(Z_N) - g(u)}{g'(u)} = a_N(Z_N - u) + o\left(a_N \frac{|Z_N - u|}{g'(u)}\right);$$

the first term converges in distribution to q (by our assumption) and the second one converges to zero in probability (**Exercise 39**: Prove this; i.e., prove that the remainder term

$$a_N \frac{g(Z_N) - g(u)}{g'(u)} - a_N(Z_N - u)$$

converges to zero in probability, by using the Taylor expansion formula). In other words, limit distributions are passed through functions in a pretty simple way. This is called the "delta method" (I suppose because of the deltas and epsilons involved in this kind of limiting argument), and we'll be using it a lot. The main application is when we've already proven a CLT for Z_N,

$$\sqrt{N} \frac{Z_N - \mu}{\sigma} \to_D N(0, 1),$$

in which case

$$\sqrt{N}(g(Z_N) - g(\mu)) \to_D N(0, \sigma^2 (g'(\mu))^2).$$

Exercise 40: Assume $N^{1/2} Z_N \to_D \mathcal{N}(0, 1)$. Then what is the asymptotic distribution of 1) $g(Z_N) = (Z_N - 1)^2$? 2) what about $g(Z_N) = Z_N^2$? Does anything go wrong when applying the delta method in this case? Can you fix this problem?

Mgf method

What if the r.v. we're interested in, Y_N, can't be written as $g(X_N)$, i.e., a nice function of an r.v. we already know converges? Are there methods to prove limit theorems directly?

Here we turn to our old friend the mgf. It turns out that the following generalization of the mgf invertibility theorem we quoted above is true:

Theorem 2. *The distribution functions F_N converge to F if:*

- *the corresponding mgf's $M_{X_N}(s)$ and $M_X(s)$ exist (and are finite) for all $s \in (-z, z)$, for all N, for some positive constant z.*
- *$M_{X_N}(s) \to M_X(s)$ for all $s \in (-z, z)$.*

So, once again, if we have a good handle on the mgf's of X_N, we can learn a lot about the limit distribution. In fact, this idea provides the simplest way to prove the CLT.

Proof: assume X_i has mean zero and unit variance; the general case follows easily, by the usual rescalings.

Now let's look at $M_N(s)$, the mgf of $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i$. If X_i has mgf $M(s)$, then $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i$ has mgf

$$M(s/\sqrt{N})^N.$$

Now let's make a Taylor expansion. We know that $M(0) = 1, M'(0) = 0$, and $M''(0) = 1$. (Why?) So we can write

$$M(s) = 1 + s^2/2 + o(s^2).$$

Now we just note that $M_N(s)$ converges to $e^{s^2/2}$, recall the mgf of a standard normal r.v., and then appeal to our general convergence-in-distribution theorem for mgf's. □