Stat 200: Introduction to Statistical Inference **Autumn 2018/19**

Lecture 2: More probability review

Lecturer: Art B. Owen *September 27*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

This is a continuation of probability review. It should be reminding you of things you learned in probability. It is not practical do all of probability in week one of a statistics course.

2.1 Moments

We begin with expected values. For a discrete random variable (RV) X, it's expectation is

$$\mathbb{E}(X) = \sum_i p(x_i)x_i, \qquad \text{if } \sum_i p(x_i)|x_i| < \infty.$$

If X is a continuous RV then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} f(x)x\,\mathrm{d}x \qquad \text{if } \int_{-\infty}^{\infty} f(x)|x|\,\mathrm{d}x < \infty.$$

The 'if clause' is important to make the quantity a well-defined finite value. We can think of $\mathbb{E}(X)$ as one kind of 'typical' outcome for X. As sketched in class the PDF of a continuous X would 'balance' at $\mathbb{E}(X)$.

For random variables X and Y and constants a, b and c,

$$\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c.$$

In this and other identities we assume that the expecations involved all exist without always saying that.

The expected value of X is also called the first moment. The second moment is $\mathbb{E}(X^2)$, the third is $\mathbb{E}(X^3)$ and so on. When $\mathbb{E}(X) = \mu$ we define the k'th centered moment as $\mathbb{E}((X-\mu)^k)$. The variance of X is

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

so it is the second centered moment. It quantifies how spread out f is, that is, how far from $\mathbb{E}(X)$ that X may get. The standard deviation of X is $\sqrt{\mathrm{Var}(X)}$. A little algebra gives

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

A litle more algebra would give

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y), \quad \text{if } X \text{ and } Y \text{ are independent.}$$

For random variables X and Y, thier covariance is

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) = \cdots = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

where I use $\cdots$ to replace algebra that I know you can do especially when running through the algebra adds no statistical idea. [If you are not sure why one works, then plug and chug to see it work out. There's no shame in that. Besides, nobody will know. Also you might find a typo that way.] A positive covariance is one way to quantify that X and Y tend to increase together. A negative value quantifies a tendency for increases in one to accompany decreases in the other.

One of the most famous and important statistical facts is that a nonzero $\mathrm{Cov}(X,Y)$ does not necessarily mean that if you somehow caused a change in X it would bring the corresponding change in Y either at all or on average.

We can standardize X by subtracting its mean and dividing by its standard deviation, getting $(X - \mathbb{E}(X))/\sqrt{\mathrm{Var}(X)}$. The standardized random variable has mean 0 and variance 1 (work it out if you doubt). The covariance between two standardized variables is their correlation

$$\mathrm{Corr}(X,Y) \equiv \mathrm{Cov}\Big(\frac{X-\mathbb{E}(X))}{\sqrt{\mathrm{Var}(X)}}, \frac{Y-\mathbb{E}(Y))}{\sqrt{\mathrm{Var}(Y)}}\Big).$$

Here $\equiv$ means not just equal, but equal by definition of what is on the left hand side. The correlation is not well defined if either $\mathrm{Var}(X) = 0$ or $\mathrm{Var}(Y) = 0$.

2.2 Some distributions

Here we look at a few important distributions but not all of them. Appendix A of Rice has many of the most important ones. So does Chapter 6 (distributions related to the normal distribution).

A Bernoulli random variable X takes only two values 0 and 1. These are sometimes called 'failure' and 'success' respectively. Examples for $X = 1$: you got the basketball through the hoop, the coin came up heads, you received lighting from the sky. We say $X \sim \mathrm{Bern}(p)$ when the Bernoulli random variable has $\Pr(X = 1) = p$. Of course $0 \leqslant p \leqslant 1$. Bernoulli random variables are commonly 'indicator' variables that indicate that some event occured. When the event is A we might use notation

$$X = 1_A \equiv \begin{cases} 1, & A \text{ occurs} \\ 0, & \text{else.} \end{cases}$$

This lets us turn events into numbers or turn logic into algebra, for instance

$$\Pr(A) = \mathbb{E}(1_A), \quad \text{and} \quad \Pr(A \cup B) = \mathbb{E}(1_{A\cup B}) = \mathbb{E}(1_A + 1_B - 1_{A\cap B}) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

It is easy to see that $\mathbb{E}(X) = p$. Also $X = X^2 = X^3 = X^4 \cdots$ so moments of Bernoulli variables are easy to work out. They all equal p.

Let $X_1, X_2, \ldots, X_n$ be independent $\mathrm{Bern}(p)$ random variables. The Binomial distribution $\mathrm{Bin}(n,p)$ is formed as the distribution of

$$Y = \sum_{i=1}^{n} X_i.$$

Perhaps this is the number of baskets you get in n tries. A direct combinatorial argument gives

$$\Pr(Y = y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y}, & y = 0, 1, 2, \ldots, n \\ 0, & \text{else.} \end{cases}$$

We can find by manipulating sums that

$$\mathbb{E}(Y) = \sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y} \times y = np.$$

Begin by removing $y = 0$ from the sum and expanding $\binom{n}{y} = n!/(y!(n-y)!)$. It is easier to notice that

$$\mathbb{E}(Y) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} p = np,$$

though you should work it both ways if the methods are new to you. Similarly from independence $\mathrm{Var}(Y) = n\mathrm{Var}(X_1) = np(1-p)$, or you can do it by manipulating the sum.

It is important to have $\Pr(X_i = 1)$ be the same for all $i = 1, \ldots, n$. Similarly the X_i should be independent.

Here we skipped: geometric distribution (number of tosses to first basket) negative binomial (number of tosses to attain some number of baskets) and the hypergeometric (all about red and black balls in an urn). Read about those in Rice.

The Poisson distribution with parameter $\lambda \geqslant 0$, written $X \sim \mathrm{Poi}(\lambda)$ has

$$\Pr(X = x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & x = 0, 1, 2, \cdots \\ 0, & \text{else.} \end{cases}$$

If we had asked for a distribution with $\Pr(X = x) = c\lambda^x/x!$ for some constant c (commonly written $\Pr(X = x) \propto \lambda^x/x!$ reading $\propto$ as 'proportional to') then we would have needed the constant c to satisfy

$$\frac{1}{c} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}.$$

We remember from calculus that the right hand side is $\exp(\lambda)$ and so we find that $c = \exp(-\lambda)$.

Now, working in slow motion, and knowing ahead of time what order to take the steps

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \times x \\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \times x \\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\ &= \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^{r+1}}{r!} \\ &= \lambda \times \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} \\ &= \lambda. \end{aligned}$$

Be sure you understand why each step is ok. When working out expected values you might have to try a few things to find the right sequence. In these problems it is very common that the fact that $\sum_{x=0}^{\infty} \Pr(X = x) = 1$

comes in handy. So you try to wrangle the expression to include that. Here the sum came through for the exact same λ we started with. In other settings you see the pattern for a different parameter value than you started with.

We saw in class that

$$\lim_{n\to\infty} \Pr(\text{Bin}(n, \lambda/n) = x) = \Pr(\text{Poi}(\lambda) = x)$$

for $x = 0, 1, \ldots$. This means that the Poisson random variable can be obtained in a setting where some very low probability Bernoulli event is tracked a large number n of times. We really needed to use $\text{Bin}(n, p)$ with $p = \lambda/n$. If we used $p = \lambda/n^r$ for any $r < 1$ then we would have a limit where $\mathbb{E}(X) = np = \lambda \times n^{1-r} \to \infty$ which is not an interesting or useful limit; it is not even a distribution. If instead $r < 1$ then the limit is $\text{Poi}(0)$ which is the same as $\text{Bern}(0)$, i.e., a random variable that is always 0, once again not very useful.

2.3 Inequalities

We proved Markov's inequality. If $\Pr(X \geqslant 0) = 1$ then

$$\Pr(X \geqslant t) \leqslant \frac{\mathbb{E}(X)}{t}.$$

The proof in class used indicator variables starting with

$$X = X \times 1_{X<t} + X \times 1_{X \geqslant t}$$

which holds because $1_{X<t} + 1_{X \geqslant t} = 1$. Then $X \times 1_{X<t} \geqslant 0$ and $X \times 1_{X \geqslant t} \geqslant t$ and a couple more steps close the argument.

As a consequence we get Chebychev's inequality: If X has mean μ and variance σ^2 then for $t > 0$

$$\Pr(|X - \mu| \geqslant t) \leqslant \sigma^2/t^2.$$

(Use Markov's on $Y = (X - \mu)^2$). If $t = k\sigma$ then

$$\Pr(|X - \mu| \geqslant k\sigma) \leqslant 1/k^2.$$

2.4 Tower property and variance decomposition

Let X and Y be random variables. **Read** how Rice defines $\mathbb{E}(Y \mid X)$ and $\text{Var}(Y \mid X)$ (around page 150). Then

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X)) \tag{2.1}$$

is called the tower property of expectation. Also

$$\text{Var}(Y) = \mathbb{E}(\text{Var}(Y \mid X)) + \text{Var}(\mathbb{E}(Y \mid X)) \tag{2.2}$$

is used a lot in statistics. An important consequence is that $\mathbb{E}(Y \mid X)$ can never have higher variance than Y has.

Suppose we had to guess at Y. If we guess m then our mean squared error is

$$\mathbb{E}((Y - m)^2) = \cdots = \text{Var}(Y) + (m - \mathbb{E}(Y))^2.$$

[You should be able to fill in the dots.] Our most accurate guess is $m = \mathbb{E}(Y)$.

If we know $X = x$ then our best guess for Y is $\mathbb{E}(Y \mid X = x)$. Now $\mathbb{E}(Y \mid X)$ is a random variable taking the value $\mathbb{E}(Y \mid X = x)$ when $X = x$. So somebody who knows X (whatever it turns out to be) would guess $\mathbb{E}(Y \mid X)$. Their expected squared error is

$$\mathbb{E}\Big((\mathbb{E}(Y \mid X) - Y)^2\Big) = \mathbb{E}\Big((\mathbb{E}(Y \mid X) - Y)\Big)^2 + \mathrm{Var}\Big(\mathbb{E}(Y \mid X) - Y\Big).$$

The first term is $0^2 = 0$ by the tower property (2.1). The second term is no larger than $\mathrm{Var}(Y)$ by (2.2). As a consequence $\mathbb{E}(Y \mid X)$ is a better guess than $\mathbb{E}(Y)$ or at least as good. Knowing X can never make you a worse guesser about Y.

2.5 Moment generating function

The moment generating function of X, denoted $M(t)$ or sometimes $M_X(t)$ is $\mathbb{E}(e^{tX})$, a function of $t \in \mathbb{R}$. Often it is infinite for some t but finite for other t. It is most useful when $M(t)$ exists for all $t \in (-h, h)$, that is for $|t| < h$ for some $h > 0$. In that case the MGF uniquely determines the distribution of X. That is, if random variables X and Y with CDFs F_X and F_Y have MGFs $M_X(t) = M_Y(t)$ for all $|t| < h > 0$ then $F_X = F_Y$. Same MGF means same distribution.

The characteristic function $\phi(t) = \mathbb{E}(e^{itX}) = M(it)$ always exists and it characterizes the distribution of X. It is useful when the MGF does not exist but it requires wrangling complex numbers.

We use the MGF to generate moments. As we saw in class

$$M'(0) = \mathbb{E}(X).$$

Similarly $M''(0) = \mathbb{E}(X^2)$ and for integers $r \geqslant 1$ taking the r'th derivative we get $M^{(r)}(0) = \mathbb{E}(X^r)$. For continuous random variables these findings require us to be able to interchange differentation and integration.

Our main use of MGFs is to find the distribution of the sum of two independent random variables with known distributions.

We saw that a Gamma random variable with shape α has MGF $M(t) = (1-t)^{-\alpha}$ for all $|t| < h = 1$. From the definitions of MGF we see that for independent X and Y, $M_{X+Y}(t) = M_X(t) \times M_Y(t)$. If they're $\mathrm{Gam}(\alpha)$ then $M_{X+Y}(t) = (1-t)^{-2\alpha}$ which we recognize as the MGF of the $\mathrm{Gam}(2\alpha)$ distribution. If we sum n independent $\mathrm{Gam}(\alpha)$ random variables we get a $\mathrm{Gam}(n\alpha)$ random variable.

The Gamma distribution has two parameters. In addition to the shape α there is a rate $\lambda > 0$. If $X \sim \mathrm{Gam}(\alpha)$ and $Y = X/\lambda$ then $Y \sim \mathrm{Gam}(\alpha, \lambda)$. Now $M_Y(t) = (1 - t/\lambda)^{-\alpha}$ for $|t| < \lambda$.

For us the most important Gamma distributions will be χ^2. The $\chi^2_{(n)}$ distribution is $\mathrm{Gam}(n/2, 1/2)$.

2.6 Convergence

You should know Rice chapter 5 on convergence of distributions. Here are the key ideas.

First, the law of large numbers. Let X_i be IID with mean μ and let $\bar{X}_n \equiv (1/n)\sum_{i=1}^n X_i$ be their average. The law of large numbers states that for any $\epsilon > 0$

$$\lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \epsilon) = 0.$$

In this sense averages settle down to their corresponding expecations. Rice proves this via Chebychev assuming also $\mathrm{Var}(X_i) < \infty$, but it holds without that assumption. This is not the only law of large numbers, but it is the one that we will use.

The random variables $\bar{X}_n$ "converge in probability" to the value μ.

Sometimes a sequence of random variables does not converge to a constant but has instead a limit that is also a random variable. Let X_n be a sequence of random variable with CDFs F_n and let X be a random variable with CDF F. If $\lim_{n\to\infty} F_n(x) = F(x)$ holds at all x where F is continuous then we say that the sequence X_n converges in distribution to X.

To understand the exception 'where F is continuous' think about $F_n(x) = 1_{x\geqslant 1/n}$ and $F(x) = 1_{x\geqslant 0}$. What kind of random variable is X_n, what kind is X and do we think these X_n are converging to X?

Now let X_n have MGF M_n, let X have MGF M and suppose that $M_n(t) \to M(t)$ as $n \to \infty$ holds for all $|t| < h$ where $h > 0$. Then X_n converges in distribution to X. This follows from the continuity theorem in Rice.

The central limit theorem is about averages having nearly the normal distribution. See Rice p 184. Let

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

be the $N(0,1)$ PDF and

$$\Phi(x) = \int_{-\infty}^{x} \varphi(z)\,\mathrm{d}z$$

be the corresponding CDF. Then for IID variables X_i with mean μ and variance $\sigma^2 > 0$

$$\lim_{n\to\infty} \Pr\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\Big) = \Phi(x)$$

for all $x \in \mathbb{R}$. This means that $Z_n \equiv (\bar{X} - \mu)/(\sigma/\sqrt{n})$ converges in distribution to $N(0,1)$. We can tell from the formula that Z_n has mean 0 and variance 1. The CLT adds the normality (in the limit).

Rice does the CLT a little differently. Read his treatment carefully to be sure it is the same. He proves it via the MGF.

We can rearrange the CLT result to find that $\bar{X}_n$ has distribution close to $N(\mu, \sigma^2/n)$.