MATH 341: PROBABILITY: FALL 2009
THE POWER OF EXPECTATION

STEVEN J. MILLER (SJM1@WILLIAMS.EDU)

ABSTRACT. The purpose of these notes is to show the power of expectation. It is phenomenal how many problems can be solved by appealing to the linearity of expectation. Amazingly, it doesn't matter if the random variables are dependent or independent *if* we only care about the expected (ie, the average) value; the situation is very different if we care about the size of the fluctuations about the average value.

1. TERMINOLOGY

We began today's lecture by reviewing the definition of moments, in particular that the variance is the second centered moment, or $\sigma^2 = \mathbb{E}[(X-\mu)^2]$, with $\mu = \mathbb{E}[X]$. The standard deviation is the square-root of the variation, and has the same units as the random variable we are studying. For example, if X is the average height in the class, then the variance has units meters-squared while the standard deviation has units of meters. Thus, when studying fluctuations about the average value, it is the standard deviation (and not the variance) that gives the right scale.

If X and Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. This is a very useful relationship, and allows us to reduce complicated random variables to simpler ones. It is possible for this relation to hold without X and Y being independent (it is a nice exercise to come up with such an example); in this case we say X and Y are uncorrelated.

We proved (or discussed how one would do the algebra to prove) that

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \ = \ \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{1 \le i < j \le n} \mathrm{CoVar}(X_i, X_j),$$

with $\mathrm{CoVar}(X_i, X_j) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$. If two random variables are independent than their covariances is zero. Thus in the special case that all the X_i's are independent we have the variance of a sum is the sum of the variances. (Note, however, that $\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$, so variance is *not* linear.)

One application of our formulas for variances is portfolio theory from economics. If we have two stocks with the same expected return $\mu > 0$ and same variance σ^2, then imagine we allocate our funds as follows: if we have \$1 to spend, we spend p dollars on the first and $1-p$ on the second. Thus if X_i, $i \in \{1,2\}$ are the random variables indicating our return for stock i, our investment may be denoted by $X =$

Date: Thursday, October 8, 2009.

$pX_1 + (1-p)X_2$. Note

$$\mathbb{E}[X] \ = \ p\mathbb{E}[X_1] + (1-p)\mathbb{E}[X_2] \ = \ p\mu + (1-p)\mu \ = \ \mu$$

by linearity. Further

$$\mathrm{Var}(X) \ = \ p^2\mathrm{Var}(X_1) + (1-p)^2\mathrm{Var}(X_2) \ = \ \left(p^2 + (1-p)^2\right)\sigma^2.$$

It is a nice calculus exercise to show that the minimum value is when $p = 1/2$, which gives a variance of X of $\sigma^2/2$. In other words, we have found an investment with the same expected return as X_1 and X_2 but with less risk / uncertainty. Of course, like much of economics, there are many assumptions with this model that may not hold in the real world (the severest being that we have two independent stocks). (As a nice exercise, how should you allocate your resources if instead the stocks have two different variances, say σ_1^2 and σ_2^2?)

---

2. Double integrals

We needed to compute

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)dxdy.$$

What we actually mean by such an integral (returning to Calc III) is the following: we divide the xy-plane into small rectangles, and compute the volume of the upper and lower boxes, and then take the limit as the partition becomes finer and finer. This is the natural generalization of the Riemann sum definition from Calc I or II.

We *do not* want to evaluate the integral by working with this definition (for those who have taken analysis, by using the product measure). We want to reduce this to iterated integrals. The Fubini / Fubini-Tonelli theorems tell us when we can so evaluate multiple integrals; the Wikipedia entry

`http://en.wikipedia.org/wiki/Fubini%27s_theorem`

is a good source.

Theorem 2.1 (Fubini's Theorem)**.** *Assume f is continuous and*

$$\int_a^b\int_c^d |f(x,y)|dxdy \ < \ \infty. \tag{2.1}$$

Then

$$\int_a^b\left[\int_c^d f(x,y)dy\right]dx \ = \ \int_c^d\left[\int_a^b f(x,y)dx\right]dy. \tag{2.2}$$

Similar statements hold if we instead have

$$\sum_{n=N_0}^{N_1}\int_c^d f(x_n,y)dy, \qquad \sum_{n=N_0}^{N_1}\sum_{m=M_0}^{M_1} f(x_n,y_m). \tag{2.3}$$

For a proof in special cases, see

- P. Baxandall and H. Liebeck, *Vector Calculus*, Clarendon Press, Oxford, 1986.

- W. Voxman and R. Goetschel, Jr., *Advanced Calculus*, Mercer Dekker, New York, 1981.

An advanced, complete proof is given in

- G. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd edition, Pure and Applied Mathematics, Wiley-Interscience, New York, 1999.

The exercise below gives an example where we cannot change the order of summation (by smoothing things out, we could make this a counterexample for integrals).

Exercise 2.2. *One cannot always interchange orders of integration. For simplicity, we give a sequence a_{mn} such that $\sum_m(\sum_n a_{m,n}) \neq \sum_n(\sum_m a_{m,n})$. For $m, n \geq 0$ let*

$$a_{m,n} \;=\; \begin{cases} 1 & \text{if } n = m \\ -1 & \text{if } n = m+1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Show that the two different orders of summation yield different answers (the reason for this is that the sum of the absolute value of the terms diverges).

We will prove later that if X and Y are independent random variables with marginals f_X and f_Y and joint distribution $f_{X,Y}$ that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Let's recall what all this means:

$$\begin{aligned} \mathbb{P}(a \leq X \leq b, c \leq Y \leq d) &= \int_a^b \int_c^d f_{X,Y}(x,y)dxdy \\ \mathbb{P}(a \leq X \leq b) &= \int_a^b f_X(x)dx \\ \mathbb{P}(c \leq Y \leq d) &= \int_c^d f_Y(y)dy. \end{aligned}$$

Assuming this fact for now, we analyzed the double integrals and proved $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if they are independent.

While we will prove the claim later, we give the key insight. Assume not, so there is some point such that $f_{X,Y}(x_0,y_0) \neq f_X(x_0)f_Y(y_0)$. Without loss of generality assume $f_{X,Y}(x_0,y_0) - f_X(x_0)f_Y(y_0) > 0$; let $\epsilon = |f_{X,Y}(x_0,y_0) - f_X(x_0)f_Y(y_0)|/2009$. By continuity, we can find a small square centered at (x_0,y_0) such that $f_{X,Y}(x,y)$ is within ϵ of $f_{X,Y}(x_0,y_0)$, and similar statements hold for f_X and f_Y. This violates

$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) \;=\; \mathbb{P}(a \leq X \leq b, c \leq Y \leq d)\mathbb{P}(c \leq Y \leq d).$$

---

3. Modeling deterministic systems randomly

As we've stated numerous times, a given integer is either divisible by 7 or it is not; what does it mean to say it has a 1 in 7 chance of being divisible by 7? What we mean is that if we consider a large number of consecutive integers, then roughly 1 in 7 will be multiples of 7.

We discuss a model for counting the number of prime divisors of n. Let p be a prime. For each prime at most n, we flip a coin with probability $1/p$ of heads. If a head comes up, we say n is p-good; else n is not p-good. For a given n, on average how many primes will it be considered p-good?

Clearly this models how many distinct prime divisors a number has. There are, of course, some differences. First, n cannot be divisible by a prime between $n/2$ and $n-1$, while this is not the case in our random model (though we will see that the contribution from such terms is small).

Fix n. Let X_p be the random variable that equals 1 with probability $1/p$ and 0 otherwise. Then

$$\mathbb{E}[X_p] \ = \ 1 \cdot \frac{1}{p} + 0 \cdot \left(1 - \frac{1}{p}\right) \ = \ \frac{1}{p}.$$

Similarly we find

$$\mathrm{Var}(X_p) \ = \ \left(1 - \frac{1}{p}\right)^2 \cdot \frac{1}{p} + \left(0 - \frac{1}{p}\right)^2 \cdot \left(1 - \frac{1}{p}\right) \ = \ \frac{1}{p} - \frac{1}{p^2},$$

where the last equality follows from elementary algebra. Let X be the random variable equalling the number of primes for which n is p-good; thus $X = \sum_{p \le n} X_p$. By linearity of expectation, we have

$$\mathbb{E}[X] \ = \ \sum_{p \le n} \mathbb{E}[X_p] \ = \ \sum_{p \le n} \frac{1}{p}.$$

There are many ways to evaluate this sum, some of which are discussed in the additional comments from Tuesday, October 6th's lecture. One way is to use the Prime Number Theorem and partial summation. Another is to use the Riemann zeta function $\zeta(s)$ and some truncation. For $\mathfrak{Re}(s) > 1$, set

$$\zeta(s) \ = \ \sum_{n=1}^{\infty} \frac{1}{n^s} \ = \ \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}.$$

We argue informally to give the general flavor (one needs to justify that the two cutoffs can be chosen as we do below):

$$\begin{aligned} \sum_{n \le x} \frac{1}{n^s} \ &\sim \ \prod_{p \le x} \left(1 - \frac{1}{p^s}\right)^{-1} \\ \log \sum_{n \le x} \frac{1}{n^s} \ &\sim \ -\sum_{p \le x} \log\left(1 - \frac{1}{p^s}\right) \\ \log \sum_{n \le x} \frac{1}{n^s} \ &\sim \ \sum_{p \le x} \left(\frac{1}{p^s} + \frac{1}{2p^{2s}} + \frac{1}{3p^{3s}} + \cdots\right), \end{aligned}$$

where the last follows from the Taylor series expansion of $\log(1-u)$. Taking $s = 1$ and noting the left hand side is the harmonic sum ($\sum_{n \le x} \frac{1}{n} \sim \log x$), and the sum over the prime squares and higher is bounded, we find

$$\log \log x \ \sim \ \sum_{p \le x} \frac{1}{p}.$$

Thus the expected number of prime divisors of n should be about $\log \log n$. For n enormous, the bounded constant doesn't really matter, though for 'small' n it will be

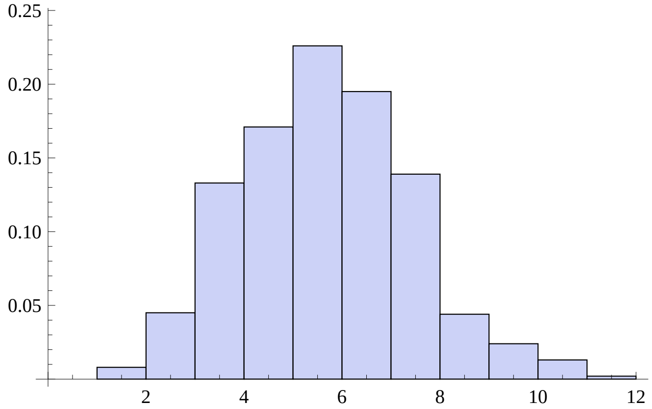

FIGURE 1. Distribution of the number of prime factors for n (1000 consecutive values starting at 5487525252462375634352364513298043621345687989991218989811).

noticeable. What is its size? Remember that $\log$ grows slowly, and $\log\log$ even slower! For example, $\log\log(10^{100})$ is about 5.4, while $\log\log(10^{1000})$ is only 7.7 (if we go up to the astronomically large 10^{10000} it only increases to about 10).

What is the scale of the fluctuations? To understand this we need to know the variance of X. Fortunately the X_i's are independent. This is clear in our model, as they are chosen independently from each other. For the actual primes, this is a reasonable assumption – whether or not a generic number is divisible by one prime is independent of whether or not it is divisible by another. For example, one-third of all integers are divisible by 3, one-fifth by 5, and one-fifteenth by 3 and 5.

Thus all the covariance terms are zero, and

$$\text{Var}(X) \ = \ \sum_{p\le n} \text{Var}(X_p) \ = \ \sum_{p \le x} \left(\frac{1}{p} - \frac{1}{p^2}\right).$$

The sum of $1/p^2$ converges (it is a p-series in the lingo of Calc II, though here 'p' refers to the exponent 2), and we've discussed that the sum of $1/p$ is of size $\log\log n$. As the standard deviation is the square-root of the variance, we see that the fluctuations about the mean of $\log\log n$ are quite small in the limit, typically of size $\sqrt{\log\log n}$.

We plot the distribution of the actual number of distinct prime divisors for 1000 values of n starting at 5487525252462375634352364513298043621345687989991218989811. The Erdos-Kac theorem, which is linked in the additional comments, describes a true gem of number theory, namely that the number of prime divisors is normally distributed.

4. DIFFERENTIATING IDENTITIES

I have written a handout on this when I was at Brown; you should look at pages 2 through 5 of the handout online at

```
http://www.williams.edu/go/math/sjmiller/public_html
/341/handouts/DifferentiatingIdentities.pdf
```