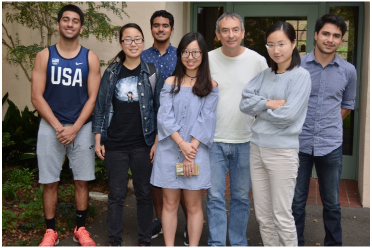
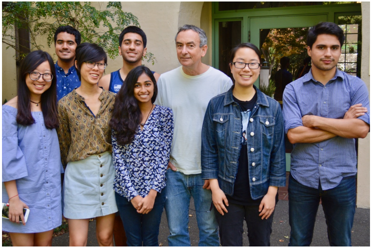
Probability

Kenneth A. Ribet

Math 10A
After Election Day, 2016





Today's breakfast

Breakfast next Monday (November 17) at 9AM.

Friday is a UC holiday but I'll be at Blue Bottle Coffee at 8AM.

Both remaining quiz dates are Wednesdays: November 16 and November 30.

A Facebook post by an old friend who was a CNN foreign correspondent:

> *Today I feel tremendous disillusionment in all the supposedly sophisticated political polling models, which apparently are incapable of discovering and transmitting truth about public opinion. Polling, like "news" media, has taken a huge hit in my book. They have both become veneers without substance. All the "big data" being collected and analyzed did not reveal the truth, or even close.*

If you really like probability and statistics, please help out!

One more variance calculation

Suppose that X is "uniform" on the interval $[a, b]$ and 0 outside of this interval. (Take $a < b$.) Then the PDF of X is $\dfrac{1}{b-a}$ on the interval and 0 outside the interval. The mean of X is then the midpoint of the interval, i.e., $\dfrac{b+a}{2}$ and

$$\mathrm{Var}[X] = \frac{1}{b-a}\int_a^b (x - \frac{a+b}{2})^2\,dx.$$

One way to calculate the integral without a lot of chalk is to set $c = \dfrac{b+a}{2}$ and make a change of variable in the integral, introducing $t = x - c$ as the new variable (so $x = t + c$, $dx = dt$). Then

$$\mathrm{Var}[X] = \frac{1}{b-a}\int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} t^2\,dt = \frac{1}{b-a}\cdot 2\frac{(b-a)^3}{24} = \frac{(b-a)^2}{12}.$$

We next have a brief discussion about the method of least squares. This method is usually credited to Gauss, but I learned from Wikipedia that Legendre found it first (who knew?!). The idea is that you are given a cloud of points in the plane, say $(1, 6)$, $(2, 5)$ and $(3, 7)$. (This is the baby example in Wikipedia.) You want to find the line $y = \beta_1 + \beta_2 x$ that "fits" these points most closely. You do this by finding that β_1 and β_2 that minimize the

sum of the squares of the differences

between $\beta_1 + \beta_2 x$ and y for each of the points (x, y).

The Wikipedia page in question does out the calculation and finds that the best-fitting line in this situation is $y = 3.5 + 1.4x$.

In 10B, you will learn the formula for β_1 and β_2 in terms of the coordinates of the given points. It is obtained by setting derivatives equal to 0. Schreiber (page 563): "Using technology to find the best-fitting line…"; he must think calculators have the relevant formula built into them. Is he right—I don't have a calculator?

The slope of the best-fit line is given by the formula

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

in which the given points are various (x_i, y_i) and the numbers $\bar{x}$ and $\bar{y}$ are the averages of the x- and y- coordinates of the given points. (If the denominator is 0, then all the x_i are equal, so the line should be vertical—it'll have infinite slope.) For more details, see `apr26.pdf` in the Lectures portion of `bCourses`.

Logistic functions

We consider the differential equation $y' = ry(1-y)$ where r is a positive constant (e.g., $r = 1$) and an initial value $y(0)$ with $0 < y(0) < 1$. The solution is

$$y(t) = \frac{1}{1+e^{a-rt}},$$

where a is a constant. The relation between $y(0)$ and a is given by $y(0) = \dfrac{1}{1+e^a}$; if I calculated correctly, $a = \ln\left(\dfrac{1-y(0)}{y(0)}\right)$.

The main point here is that $y(t)$ is a CDF: its derivative

$$f(t) = \frac{re^{a-rt}}{(1+e^{a-rt})^2}$$

is positive (and is the associated PDF). When $t \to -\infty$, $y(t) \to 0$; when $t \to \infty$, $y(t) \to 1$.

Thus we have a nice stock of probability distributions floating around for examples. With a and r as on the previous slide, Schreiber gives:

$$E[X] = \frac{a}{r}, \quad \mathrm{Var}[X] = \frac{1}{3}\left(\frac{\pi}{r}\right)^2.$$

The mean (expected value) calculation is easy, but the variance calculation seems to be annoying. (Schreiber says it's "challenging.") For the mean, let's assume first that $a = 0$ and see why the mean is 0:

$$f(t) = \frac{re^{-rt}}{(1 + e^{-rt})^2} = \frac{re^{+rt}}{(e^{+rt} + 1)^2} = f(-t).$$

Thus the mean is 0, by symmetry. (The PDF is even.) The case when a is not necessaryly 0 is similar because we can check that there's a symmetry about the point $t = a/r$.

After class, I tried to calculate the “challenging” integral and ended up reading the stackexchange page
http://math.stackexchange.com/questions/1267635/compute-variance-of-logistic-distribution
on this subject. The main question for me was how the π^2 could possibly emerge in the answer. The main fact that seems to be needed is that

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots = \frac{\pi^2}{6}.$$

Logistic regression

This is Example 6 on page 562. The technique appears also on a homework problem.

In this story, we have a table of pairs (t, p), where t is typically positive and p is between 0 and 1. What's a possible relation between t and p? We might believe for some reason that

$$p(t) \approx \frac{1}{1 + e^{a-rt}},$$

with a and r to be found. In other words, we get smitten with the idea that p is approximated by a logistic function of t; we have to figure out which one.

If $p(t) = \dfrac{1}{1+e^{a-rt}}$, then

$$1 - p(t) = \frac{e^{a-rt}}{1+e^{a-rt}}, \quad \frac{p(t)}{1-p(t)} = \frac{1}{e^{a-rt}}$$

and so

$$\ln\left(\frac{p(t)}{1-p(t)}\right) = rt - a$$

is a linear function of t.

To find the best values of a and r, we use the method of Gauss (least squares).

For example, suppose for $t = 1$, 2 and 3, the values of p are:

$$0.9957, \quad 0.9933, \quad 0.9990.$$

Then the corresponding values of $\ln\left(\frac{p}{1-p}\right)$ are respectively 6, 5 and 7 (to lots of decimal places).

The points $\left(t, \ln\left(\frac{p}{1-p}\right)\right)$ are approximated by the graph of the function $1.4t + 3.5$ (as we saw before!). Hence

$$p(t) \approx \frac{1}{1+e^{-3.5-1.4t}}$$

because $r = 1.4$ and $-a = 3.5$ with our notations.