Methods of Mathematics

Kenneth A. Ribet

UC Berkeley

Math 10B
February 30, 2016

Office hours

Monday 2:10–3:10 and Thursday 10:30–11:30 in Evans

Tuesday 10:30–noon at the SLC

Welcome to March!

Meals

- March 2, 6:30PM dinner at Chengdu Style Restaurant—send email to reserve your place
- March 3, 8AM breakfast—full
- March 4, 12:30PM pop-up Faculty Club lunch—just show up!
- March 18, 8AM breakfast—send email to reserve your place

Some variance calculations

If $\Omega = \{\, \mathrm{T}, \mathrm{H} \,\}$ and $X(\mathrm{T}) = 0$, $X(\mathrm{H}) = 1$, then $E[X] = p$, where p is the probability of a head. It follows (board or doc camera) that $\mathrm{Var}[X] = p(1 - p)$. We write σ^2 for $\mathrm{Var}[X]$, by the way.

Now imagine the binomial distribution attached to n successive coin flips, and let X be the usual variable that counts the number of heads. Trick: we think of X as $X_1 + \cdots + X_n$, where X_i is 1 or 0 according as the ith coin flip is a T or H.

Cheat: we admit (without checking the definition in detail) that the variables $X_1, X_2, \ldots, X_n$ are *independent*. We do this because the various coin flips have nothing to do with each other.

Some variance calculations

If $\Omega = \{\,\mathrm{T}, \mathrm{H}\,\}$ and $X(\mathrm{T}) = 0$, $X(\mathrm{H}) = 1$, then $E[X] = p$, where p is the probability of a head. It follows (board or doc camera) that $\mathrm{Var}[X] = p(1-p)$. We write σ^2 for $\mathrm{Var}[X]$, by the way.

Now imagine the binomial distribution attached to n successive coin flips, and let X be the usual variable that counts the number of heads. Trick: we think of X as $X_1 + \cdots + X_n$, where X_i is 1 or 0 according as the ith coin flip is a T or H.

Cheat: we admit (without checking the definition in detail) that the variables $X_1, X_2, \ldots, X_n$ are *independent*. We do this because the various coin flips have nothing to do with each other.

Some variance calculations

If $\Omega = \{\, \mathrm{T}, \mathrm{H} \,\}$ and $X(\mathrm{T}) = 0$, $X(\mathrm{H}) = 1$, then $E[X] = p$, where p is the probability of a head. It follows (board or doc camera) that $\mathrm{Var}[X] = p(1 - p)$. We write σ^2 for $\mathrm{Var}[X]$, by the way.

Now imagine the binomial distribution attached to n successive coin flips, and let X be the usual variable that counts the number of heads. Trick: we think of X as $X_1 + \cdots + X_n$, where X_i is 1 or 0 according as the ith coin flip is a T or H.

Cheat: we admit (without checking the definition in detail) that the variables X_1, X_2,..., X_n are *independent*. We do this because the various coin flips have nothing to do with each other.

It follows from the linearity of expected value that $E[X] = \sum E[X_i]$ and from the independence of the X_i that $\mathrm{Var}[X] = \sum \mathrm{Var}[X_i]$. On the other hand, each X_i is a simple Bernoulli variable with expected value p and variance $p(1-p)$. Thus:

$$E[X] = np, \quad \mathrm{Var}[X] = np(1-p).$$

Now let

$$\overline{X} := \frac{X_1 + \cdots + X_n}{n},$$

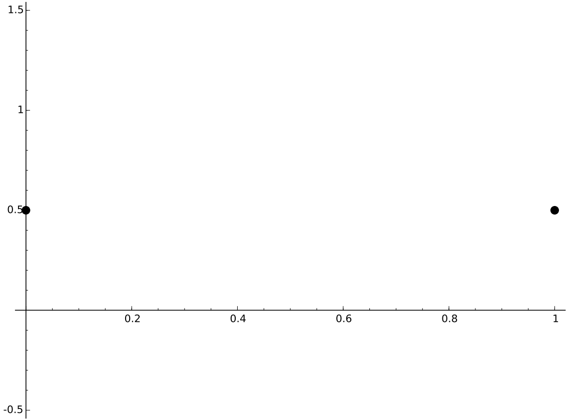
so $\overline{X}$ represents the fraction of the time that our coin landed on heads. It is immediate that

$$E[\overline{X}] = p, \quad \mathrm{Var}[\overline{X}] = \frac{p(1-p)}{n} = \frac{\sigma^2}{n}.$$

When we flip a coin n times, the number of heads divided by n is "expected" to be p; as $n \to \infty$, the fraction in question is close to p with high probability.

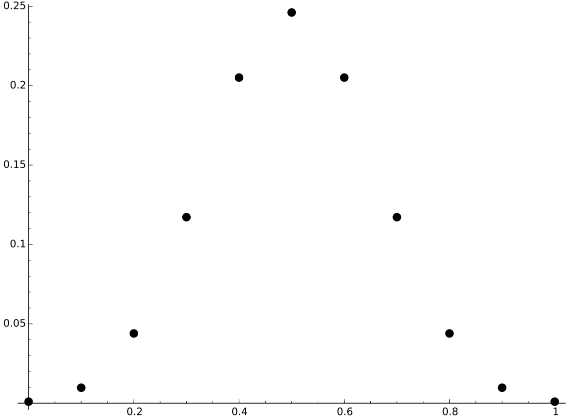
As an example, let's flip a fair coin ($p = 1/2$) n times. The probability that there are k heads is $\frac{1}{2^n}\binom{n}{k}$. If we plot this number as a function of k, the graph looks more and more spiked as n gets big.

The distribution of $\overline{X}$ when $n = 1$



Curved

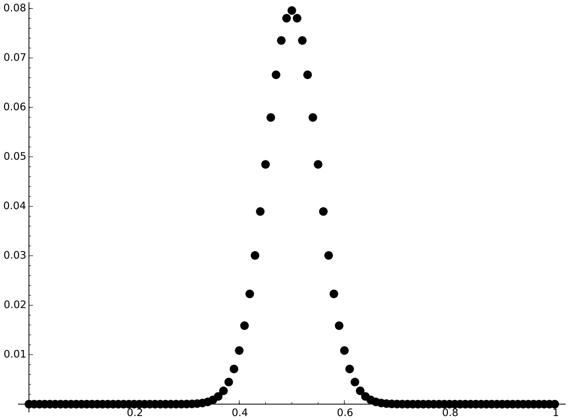
The distribution of $\overline{X}$ when $n = 10$



Note that $\binom{10}{5} = 252$, so that $\binom{10}{5}/2^{10}$ is about 0.246.

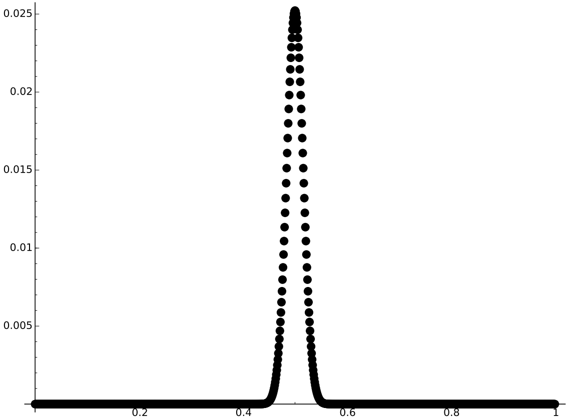
Spiking

The distribution of $\overline{X}$ when $n = 100$



Spiked

The distribution of $\overline{X}$ when $n = 1000$



In this story, we started with a Bernoulli random variable X ("heads or tails?") and considered the average of a large number of copies. We could re-do the story with *any* random variable X as long as the X_i continue to be independent copies of X. In stat lingo, the X_i are independent, identically distributed random variables.

The **Law of Large Numbers** states roughly that $\overline{X}$ approaches the expected value of X (written μ, typically) as $n \to \infty$.

The correct way to state the Law is to note that the probability space Ω is growing as $n \to \infty$. In our coin-flipping example, it has 2^n elements when there are n flips. In the limit, Ω acquires a probability structure that is built from the structures on its finite pieces. The law states that the set of $\omega \in \Omega$ for which $\overline{X}(\omega) \to \mu$ is an event whose probability is 1.

Central Limit Theorem

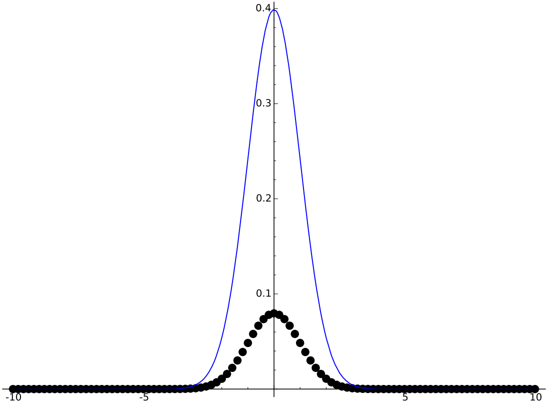
According to the "textbook," Math 10A veterans will not be surprised by the introduction of

$$Z := (\overline{X} - \mu) \cdot \frac{\sqrt{n}}{\sigma}.$$

Subtracting μ from $\overline{X}$ gives you a random variable with mean 0; multiplying by $\frac{\sqrt{n}}{\sigma}$ scales the variable so its variance is 1.

In the examples that we've done pictorially, $p = 1/2$, $\sigma^2 = 1/4$, so $\sigma = 1/2$. We are taking the values of $\overline{X}$, which ranged from 0 to 1 and shifting them by subtracting 1/2, thereby getting numbers between $-1/2$ and $+1/2$. We are then multiplying by $2\sqrt{n}$, so the values range between $\pm\sqrt{n}$.

The Central Limit Theorem states that Z is approximately normal for large n. The "textbook" refers to outside sources, and I'll do the same. There's something that I need to explain, at least to myself. If you plot together the "bell curve" $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and the probability distribution Z, you'll see that Z looks miuch less tall than the bell curve (= normal curve).



This needs some explanation. We will focus on the case of n flips of a fair coin:

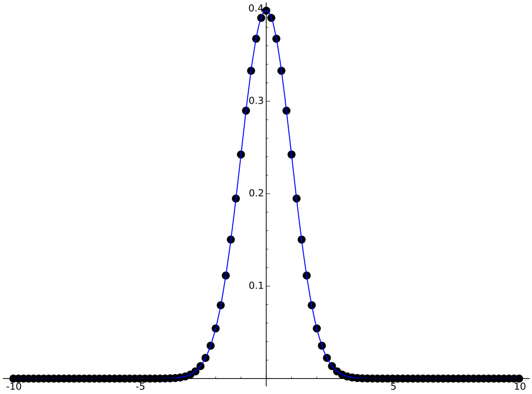
The plot of Z runs horizontally from $-\sqrt{n}$ to $\sqrt{n}$ and includes $n+1$ points. If we were to estimate the area under the plot, we'd add together the areas of rectangles whose widths would be $2\sqrt{n} \cdot \frac{1}{n}$, in view of the fact that the $n+1$ points divide an interval of length $2\sqrt{n}$ into n sub-intervals. The heights of the rectangles would be the various probabilities associated with the distribution of Z; these probabilities sum to 1. Thus our estimate for the area would be $2\sqrt{n} \cdot \frac{1}{n} \cdot 1 = \frac{2}{\sqrt{n}}$.

We want to compare the plot of Z with the bell curve; the area under the bell curve is 1. Accordingly, we expect the plot of Z to be roughly $\frac{2}{\sqrt{n}}$ as tall as the bell curve. In particular, the maximum height of the Z plot should be around $\frac{2}{\sqrt{n}} \cdot \frac{1}{\sqrt{2\pi}}$.

For example, when $n = 10$, the maximum height of the Z-plot is around 0.246, as we saw earlier. According to Sage, the value of $\frac{2}{\sqrt{n}} \cdot \frac{1}{\sqrt{2\pi}}$ when $n = 10$ is 0.252313252202016.

For $n = 100$, $\frac{2}{\sqrt{n}} \cdot \frac{1}{\sqrt{2\pi}} \approx 0.08$. That looks pretty much like the height of the dotted curve that we saw two slides back.

As I explained in class, it was a failure on my part not to include a graph showing the exponention curve together with the discrete plot that has been scaled up so that the y-axis is stretched by a factor of $\frac{\sqrt{n}}{2}$. Here is that happens when $n = 100$:



This is a pretty good fit!! The results for $n = 1000$ and $n = 10000$ are similar and perhaps even more dramatic.

We now have a complete attitude adjustment where we imagine trying to learn about X through sampling. For example, we might know that X corresponds to the flip of a biased coin and would like to know p, the probability of a head. We do lots of coin flips and compile data. The X_i are the same as before (so that X_i refers to the ith coin flip). The actual flips of our coin generate *values* of the functions X_i, and we write x_i for these actual values. Thus the x_i are numbers, whereas the X_i are functions on the probability space.

Jargon: a *statistic* is a function g of n variables. Then $g(X_1, \ldots, X_n)$ is a function on the probability space; it's a random variable. The quantity $g(x_1, \ldots, x_n)$ is a number. More jargon: a *point statistic* is a function g that can be used to estimate the mean, variance or standard deviation of X.

Example: the function $\overline{X}$ is a statistic that estimates the mean of X. If flip a coin 1000 times and observe 678 heads, we would estimate that the coin is biased with $p = 0.678$.

This blows my mind: the statistic

$$\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2$$

estimates $\operatorname{Var}[X]$. This is upsetting since $\operatorname{Var}[X] = E[(X-\mu)^2]$ and since expected values are estimated by taking averages with n in the denominator. So why do we have $n-1$? It's because

$$E[\sum_{k=1}^{n}(X_k - \overline{X})^2] = (n-1)\operatorname{Var}[X],$$

as we're about to see.

One point is that $\overline{X}$ is not $E[X] = \mu$, but only an estimate for μ. Moreover, $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ involves the various X_k in its definition. As a result, the difference $X_k - X$ is a combination of the X_i for which the coefficient of X_k is $\frac{n-1}{n}$ and the coefficients of the other X_i are all $-\frac{1}{n}$. This proves nothing–so far–but presages a somewhat lengthy computation in which $n - 1$s are likely to pop up.

Following the “textbook," we will subtract μ from X, $\overline{X}$ and the X_i. This makes their expected values all equal to 0 instead of μ and does not change the differences $X_k - X$ or the variance of X. Also, we note for distinct j and k that $E[X_j X_k] = E[X_j]E[X_k]$ by independence. The right-hand expected values are both 0, so $E[X_j X_k] = 0$. The takeaway is that squares of sums will be sums of squares, when taking expected values—cross terms won't contribute.

The next comment is that $E[X_j^2] = \text{Var}[X]$ for all j. That's because the variables X_j are all distributed like X. The upshot is that the quantity to be computed, $E[\sum_{k=1}^n (X_k - \overline{X})^2]$, is just $C \cdot \text{Var}[X]$, where C is the sum of the coefficients of the various X_j^2 that appear when you write out $\sum_{k=1}^n (X_k - \overline{X})^2$. It's easy to make mistakes calculating (as you'll see if I try to do this in front of you with the document camera), but you should get $C = n - 1$ if you persevere and don't get spooked by the subscripts.