
Serge
Lang
Dinner
Berkeley
Mathematics

Bell-shaped curves, variance

Math 10A

November 7, 2017

Pop-in lunch tomorrow, November 8, at high noon.

Please join our group at the Faculty Club for lunch.

Means

If X is a random variable with PDF equal to $f(x)$, then we've defined:

$$\begin{aligned}\mu &= \text{mean of } X = E[X] = \text{expected value of } X \\ &= \int_{-\infty}^{\infty} x \cdot f(x)\, dx.\end{aligned}$$

Simple properties of the mean

- The mean of a sum is the sum of the means, i.e., $E[X_1 + X_2] = E[X_1] + E[X_2]$.
- The mean of a product is usually not the product of the means; for example, $E[X^2]$ and $E[X]^2$ are typically different. (The difference between the two is the variance of X, as I'll explain in a few moments.)
- If X is constant, say $X = a$, then $E[X] = a$.
- $E[X - \mu] = 0$ if $\mu = E[X]$.
- $E[47 \cdot X] = 47 \cdot E[X]$. You can replace 47 by another number if you prefer.

Here's an important question: X^2 is a random variable; what is $E[X^2]$? Can we write it as an integral?

The answer is "yes," and in fact

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx.$$

More generally,

$$E[X^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$$

for $n \geq 0$.

Here's an important question: X^2 is a random variable; what is $E[X^2]$? Can we write it as an integral?

The answer is "yes," and in fact

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx.$$

More generally,

$$E[X^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$$

for $n \geq 0$.

Want to know why?

I thought so. . . .

Want to know why?

I thought so. . . .

This won't be on the exam

We first figure out the CDF of X^2. We started with X, say with CDF equal to $F(t)$ and PDF $= f(x)$.

Let $G(t)$ and $g(x)$ be the CDF and PDF of X^2. By definition:

$$G(t) = P(X^2 \leq t).$$

For t negative, $G(t) = 0$. For $t \geq 0$,

$$G(t) = P(X^2 \leq t) = P(-\sqrt{t} \leq X \leq \sqrt{t}) = F(\sqrt{t}) - F(-\sqrt{t})$$

and thus

$$g(t) = G'(t) = f(\sqrt{t})\frac{1}{2\sqrt{t}} + f(-\sqrt{t})\frac{1}{2\sqrt{t}}.$$

Thus

$$E[X^2] = \int_0^\infty t\, g(t)\, dt = \int_0^\infty t \left(f(\sqrt{t}) + f(-\sqrt{t}) \right) \frac{1}{2\sqrt{t}}\, dt.$$

Put $x = \sqrt{t}$, $t = x^2$. We get

$$E[X^2] = \int_0^\infty x^2 \left(f(x) + f(-x) \right)\, dx$$

and we can convert this to

$$\int_{-\infty}^\infty x^2 f(x)\, dx$$

by changing the sign of the variable in $\int_0^\infty x^2 f(-x)\, dx$.

On the exam

The *variance* of a random variable X measures the extent to which X differs from its mean $\mu = E[X]$:

$$\text{Var}[X] = E[(X - \mu)^2].$$

We square $X - \mu$ in order to treat negative and positive differences the same.

Algebraically,

$$\begin{aligned}\text{Var}[X] &= E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2E[X] \cdot E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2.\end{aligned}$$

Example

We roll a fair coin once and let $X = 1$ if we get a head, $X = 0$ if we get a tail. Then $E[X] = \frac{1}{2}$. Since $0^2 = 0$ and $1^2 = 1$, $X^2 = X$. Thus

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

A biased coin

Same example, but suppose that the coin comes up heads with probabilty p, tails with probability $q = 1 - p$. Then $E[X] = p$ and

$$\text{Var}[X] = E[X^2] - E[X]^2 = p - p^2 = p(1 - p) = pq.$$

The Math 10A case

If X has mean μ, then

$$\begin{aligned}
\mathrm{Var}[X] = E[X^2] - E[X]^2 &= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 \\
&= \int_{-\infty}^{\infty} x^2 f(x)\,dx - 2\mu\cdot\mu + \mu^2 \\
&= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \int_{-\infty}^{\infty} 2\mu x f(x)\,dx + \int_{-\infty}^{\infty} \mu^2 f(x)\,dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)\,dx \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx.
\end{aligned}$$

Standard normal distribution

A change of variable ($u = \dfrac{x}{\sqrt{2}}$) yields

$$\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2}\int_{-\infty}^{\infty} e^{-u^2}\,du = \sqrt{2}\sqrt{\pi} = \sqrt{2\pi}.$$

Hence the function

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

is a PDF. It's the *gold standard normal distribution*. The statement that "X is normally distributed" most often means that $f(x)$ is its PDF.

Normal distributions (in the plural)

If X has $f(x)$ as its PDF, X has mean 0 (because of symmetry with respect to the vertical axis) and variance 1 (as we'll check). For the more general formula

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

the mean is μ and the variance is σ^2. The number σ is taken to be positive, so the *standard deviation* $\sqrt{\sigma^2}$ will be σ.

You'll find lots of pictures and a good discussion in Wikipedia.

Lognormal

If X is non-negative, it won't be associated with a normal distribution, which runs from $-\infty$ to $+\infty$. But it might be the *exponential* of a normal variable. A random variable is called *lognormal* if its natural log is normal, i.e., if it's of the form $e^{\text{normal variable}}$.

If the normal variable has parameters σ and μ, then the PDF of the lognormal variable is

$$\frac{1}{\sqrt{2\pi}\,\sigma}\,\frac{1}{x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}.$$

Warning: the mean and standard deviation of the lognormal variable are not μ and σ; those are the mean and standard deviation of the normal variable before exponentiation. The mean and variance of the lognormal variable are calculated in terms of μ and σ in problem 37 of §7.4.

Going the other way, you can write σ and μ in terms of the mean and standard deviation of the lognormal variable (problem 38).

Not on the test

Why is the lognormal PDF given by the weird formula

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

on a previous slide? We'll work this out in the simple case $\sigma = 1$, $\mu = 0$.

Say X is a variable whose ln is distributed according to the standard normal distribution. Then

$$P(a \le \ln X \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx,$$
$$P(e^a \le X \le e^b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx.$$

Hence

$$P(X \le t) = \int_{-\infty}^{\ln t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

If F is the CDF for the normal variable, then

$$P(X \le t) = F(\ln t) - F(-\infty) = F(\ln t).$$

In other words, the CDF for the lognormal variable is obtained from the CDF of the normal variable by a natural log substitution. That was probably obvious to many of you.

The PDF for the lognormal variable is then

$$\frac{d}{dt}\Big(F(\ln t)\Big) = \frac{1}{t} F'(\ln t) = \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-(\ln t)^2/2},$$

as claimed.

About that standard deviation

The standard deviation is the square root of variance:

$$\text{Standard deviation of } X = \sqrt{\text{Var}[X]}.$$

That should be it—end of story. However, it's not because people speak more frequently of standard deviations than of variances. We'll talk about Chebyshev's inequality in a bit.

About that standard deviation

The standard deviation is the square root of variance:

$$\text{Standard deviation of } X = \sqrt{\text{Var}[X]}.$$

That should be it—end of story. However, it's not because people speak more frequently of standard deviations than of variances. We'll talk about Chebyshev's inequality in a bit.

Variance of normal distribution

If $f(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the CDF of X, we will check that $\mathrm{Var}[X] = 1$, i.e., that

$$\int_{-\infty}^{\infty} x^2 f(x)\, dx = 1.$$

To compute

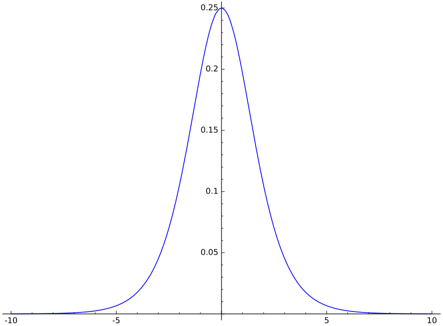
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx,$$

use integration by parts. Let $u = \dfrac{x}{\sqrt{2\pi}}$, $dv = xe^{-x^2/2}\, dx$, $v = -e^{-x^2/2}$. The term $uv\Big]_{-\infty}^{\infty}$ is 0, and we get (as desired)

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1.$$

Logistic distribution

The logistic CDF is $F(x) = \dfrac{e^x}{1 + e^x}$ and the corresponding PDF is $f(x) = F(x)(1 - F(x))$.



We have $f(x) = \dfrac{e^x}{(1 + e^x)^2}$, which we can rewrite as $\dfrac{e^{-x}}{(1 + e^{-x})^2}$ by dividing numerator and denominator by $e^{-x}e^{-x}$.

If X has PDF equal to $f(x)$, then again $E[X] = 0$ by symmetry.

Also,

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} x^2 \frac{e^x}{(1+e^x)^2}\, dx.$$

Looking at Schreiber or at Wikipedia, you'll read that

$$\mathrm{Var}[X] = \frac{\pi^2}{3}.$$

For a derivation of the formula, see this post on Mathematics Stack Exchange.

Chebyshev's inequality

The inequality in question concerns an arbitrary random variable X. Say that the mean of X is μ and that the standard deviation of X is σ. For integers $k \geq 1$, the inequality states:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

In other words: the probability of being k or more standard deviations away from the mean is at most $\frac{1}{k^2}$. For example, the probability of being two or more standard deviations away from the mean is at most $1/4$.

Why is Chebyshev's inequality true?

The explanation is provided on page 553 of the book and also (of course!) in Wikipedia. The following slides summarize the argument.

Not on the exam

For simplicity, we'll assume that the expected value of X is 0; this just means shifting the line $x = \mu$ over to the y-axis. Then

$$\begin{aligned}
\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx &\geq \int_{-\infty}^{k\sigma} x^2 f(x)\,dx + \int_{k\sigma}^{\infty} x^2 f(x)\,dx \\
&\geq \int_{-\infty}^{k\sigma} (k\sigma)^2 f(x)\,dx + \int_{k\sigma}^{\infty} (k\sigma)^2 f(x)\,dx \\
&= k^2\sigma^2 \left(\int_{-\infty}^{k\sigma} f(x)\,dx + \int_{k\sigma}^{\infty} f(x)\,dx \right).
\end{aligned}$$

Divide by σ^2 to get

$$\frac{1}{k^2} \geq \left(\int_{-\infty}^{k\sigma} f(x)\,dx + \int_{k\sigma}^{\infty} f(x)\,dx \right).$$

Not on the exam

The same inequality read differently:

$$\left(\int_{-\infty}^{k\sigma} f(x)\,dx + \int_{k\sigma}^{\infty} f(x)\,dx\right) \le \frac{1}{k^2}.$$

The left-hand sum represents the probability that X is to the right of $k\sigma$ plus the probability that X is to the left of $-k\sigma$. In other words, the left-hand term is the probability that X is k or more standard deviations from its mean.

Summary:

$$P(X \text{ is } k \text{ or more standard deviations from its mean}) \le \frac{1}{k^2}.$$

$$P(X \text{ is within } k \text{ standard deviations of its mean}) \ge 1 - \frac{1}{k^2}.$$