Lecture 11

Standard deviation is a gauge of closeness to the mean

Suppose $E(X^2) < \infty$. It is easy to see that if $\mathrm{Var}(X) = 0$, then X is a constant. Here is the proof:

$$0 = \mathrm{Var}(X) = \sum_k (k - EX)^2 P\{X = k\}.$$

Therefore, $P\{X = k\} = 0$ when $k \neq EX$. Because the sum of all probabilities of the form $X = k$ is one [as k varies], it follows that $P\{X = EX\} = 1$. Which is the statement that X is a constant, namely its expectation.

Based on the preceding, it stands to reason that if $\mathrm{Var}(X)$ is small, equivalently when $\mathrm{SD}(X)$ is small, then $X \approx EX$ with high probability. The following estimates that high probability:

Theorem 1 (Chebyshev's inequality)**.** *For every random variable X such that $E(X^2) < \infty$, and for all $\lambda > 0$,*

$$P\left\{|X - EX| < \lambda \mathrm{SD}(X)\right\} \geq 1 - \frac{1}{\lambda^2}.$$

Example 1. For instance, suppose $\mathrm{SD}(X) = 0.001$ [very small!]. Then we can apply Chebyshev's inequality with $\lambda := 100$ to see that

$$P\{|X - EX| < 0.1\} \geq 1 - \frac{1}{10000}.$$

Thus, $X \approx EX$ with high probability, as should be clear intuitively. Note the remarkable property that we needed only to know something about $\mathrm{SD}(X)$ in this example!

Proof of Chebyshev's inequality. We may notice that

$$\begin{aligned}\operatorname{Var}(X) &= \sum_k (k - EX)^2 P\{X = k\} \\ &\ge \sum_{k:\,|k-EX|\ge\lambda\operatorname{SD}(X)} (k - EX)^2 P\{X = k\} \\ &\ge \left[\lambda\operatorname{SD}(X)\right]^2 \cdot \sum_{k:\,|k-EX|\ge\lambda\operatorname{SD}(X)} P\{X = k\} \\ &= \lambda^2 \operatorname{Var}(X) \cdot P\left\{|X - EX| \ge \lambda\operatorname{SD}(X)\right\}.\end{aligned}$$

Therefore,

$$P\left\{|X - EX| \ge \lambda\operatorname{SD}(X)\right\} \le \frac{1}{\lambda^2}.$$

Subtract both sides from one to finish. □

Chebyshev's inequality holds quite generally. Therefore, one would expect it to be far from sharp [most of the time].

Example 2. Suppose $X = \pm 1$ with probability 1/2 each. Then it is easy to check that

$$EX = \mu = 0 \quad \text{and} \quad \operatorname{Var}(X) = \sigma^2 = 1, \quad \text{and therefore,} \quad \operatorname{SD}(X) = 1.$$

According to Chebyshev's inequality,

$$P\{|X| < \lambda\} \ge 1 - \frac{1}{\lambda^2}.$$

This is only useful for large values of λ. For instance if $0 < \lambda \le 1$, then $1 - (1/\lambda^2) \le 0$, so Chebyshev's inequality—while correct—is useless [it states that $P\{|X| < \lambda\} \ge$ a negative number!]. On the other hand, if $\lambda > 1$, then $P\{|X| < \lambda\} = 1$; in fact, $|X| = 1$ in our example; whereas Chebyshev's inequality states only that the said probability is at least $1 - \lambda^{-2}$. For instance, if $\lambda := 2$, then $P\{|X| < 2\} = 1$, but the Chebyshev lower bound is $1 - 2^{-2} = \frac{3}{4}$.

Standardization

If X is a random variable with $E(X^2) < \infty$, then we define its *standardization* X^* to be

$$X^* := \frac{X - EX}{\operatorname{SD}(X)}.$$

[This makes sense only when $\operatorname{SD}(X) > 0$; i.e., when X is not a constant, but a genuinely-random random variable].

Proposition 1. *The random variable X^* is unit free. Moreover, it is always the case that $E(X^*) = 0$ and $\operatorname{Var}(X^*) = 1$.*

The preceding is a simple fact: X^* is unit free because if for example X were measured in pounds, then so would be $EX = \sum_k kP\{X = k\}$; and $\text{Var}(X) = \sum_k (k - EX)^2 P\{X = k\}$ would be in pound-squares, therefore, SD(X) would be in pounds. The fact that X^* has mean zero is because $E(aX-b) = aEX-b$; apply the latter with $a = 1/\text{SD}(X)$ and $b = EX/\text{SD}(X)$. And it has standard deviation one because $\text{SD}(aX - b) = |a|\text{SD}(X)$.

Notice that the Chebyshev inequality states that

$$P\{|X^*| < \lambda\} \geq 1 - \frac{1}{\lambda^2} \qquad \text{for all } \lambda > 0.$$

Law of averages (a.k.a. law of large numbers)

Let $X_1, \ldots, X_n$ be independent random variables, all with common mean $\mu = EX_1 = \cdots = EX_n$ and common variance $\sigma^2 = \text{Var}(X_1) = \cdots = \text{Var}(X_n)$. We make two uses of our newly-discovered addition rules: First,

$$E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}E(X_1) + \cdots + \frac{1}{n}E(X_n) = \mu;$$

and second [because of independence],

$$\begin{aligned}\text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) &= \frac{1}{n^2}\text{Var}(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2}\left\{\text{Var}(X_1) + \cdots + \text{Var}(X_n)\right\} = \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

That is,

$$\text{SD}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma}{\sqrt{n}}.$$

This and Chebyshev's inequality together prove the following: For all $\lambda > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| < \frac{\lambda\sigma}{\sqrt{n}}\right\} \geq 1 - \frac{1}{\lambda^2}.$$

Select $\lambda := \epsilon\sqrt{n}/\sigma$ for an arbitrarily small but positive ϵ to find that

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| < \epsilon\right\} \geq 1 - \frac{\sigma^2}{n\epsilon^2} \to 1 \qquad \text{as } n \uparrow \infty.$$

We have proved the following in the special though important case where the X_i's have finite common variances:

Theorem 2 (Law of averages; Khintchine, 1932). *Let $X_1, \ldots, X_n$ be independent with finite common mean μ. Then for all $\epsilon > 0$ [however small],*

$$\lim_{n\to\infty} P\left\{\mu - \epsilon < \frac{X_1 + \cdots + X_n}{n} < \mu + \epsilon\right\} = 1.$$

The law of large numbers holds even if the X_i's do not have a finite variance. But we will not prove that refinement here.

Example 3. Consider the heart rates of a certain population; denote the possible heart rates by $r_1, \ldots, r_m$. Let $X_1, \ldots, X_n$ be an independent [i.e., with replacement] sample from those populations. Then $E(X_1) = \cdots = E(X_n) = \mu$, where

$$\mu = \frac{1}{m}\sum_{i=1}^{m} r_i = \text{ average population heart rate,}$$

and you should verify that $\mathrm{Var}(X_1) = \cdots = \mathrm{Var}(X_n) = \sigma^2$, where

$$\sigma^2 := \frac{1}{m}\sum_{i=1}^{m} r_i^2 - \mu^2.$$

[Alternatively, consult Example 1, p. 187 of your text.] According to the law of large numbers, if n is large then

$$P\left\{\frac{X_1 + \cdots + X_n}{n} \approx \mu\right\} \approx 1.$$

Here is how statisticians use this: If you wish to discover the average heart rate μ of a certain population, then you take a large independent sample $X_1, \ldots, X_n$ of heart rates. With high probability, the sample average $(X_1 + \cdots + X_n)/n$ is close to the unknown population average μ. Therefore, our estimate for μ is $(X_1 + \cdots + X_n)/n$; this is a good estimate with high probability, thanks to the law of averages.

Statisticians use the sample average often enough that they give it a special notation, viz.,

$$\bar{X}_n := \frac{X_1 + \cdots + X_n}{n}.$$

Thus, in particular, we know that

$$E(\bar{X}_n) = \mu, \qquad \mathrm{SD}(\bar{X}_n) = \sigma\sqrt{n}.$$

The latter is called a "square root law."

The central limit theorem

Theorem 3 (Central limit theorem; Kolmogorov, 1933). *Let $X_1, \ldots, X_n$ be independent random variables with a common distribution. In particular, they have a common mean $\mu := E(X_1)$ and variance $\sigma^2 := \mathrm{Var}(X_1)$. Suppose $\sigma < \infty$ and define $S_n := X_1 + \cdots + X_n$. Then, for all $-\infty \leq a \leq b \leq \infty$,*

$$P\left\{a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right\} \approx \Phi(b) - \Phi(a),$$

provided that n is large.

The preceding includes the central limit theorem for binomials. Indeed, if X has a Binomial$(n\,, p)$ distribution for a large n, then we can write $X = I_{A_1} + \cdots + I_{A_n}$ as a sum of n independent random variables, each with a "Bernoulli(p) distribution." The latter means that each I_{A_j} is one with probability p and zero with probability $q := 1 - p$.

We will prove the central limit theorem much later in this course. But for now let us note that $ES_n = n\mu$ and $\mathrm{SD}(S_n) = \sigma\sqrt{n}$. Therefore, the central limit theorem is really saying that the standardization of S_n has approximately a standard normal distribution when n is large.