Variance

Robert Y. Lewis

CS 0220 2024

April 22, 2024

Overview

Beyond mean and median

The expectation (mean) of a random variable tells us something about how values of that variable are distributed over our sample space.

The median tells us something similar, but with a different edge to it.

One (or two) measures aren't enough to fully summarize a data set. Compare median and mean for:

$$10, 10, 10, 10, 10, 10, 10$$

$$0, 1, 2, 10, 18, 19, 20$$

Markov's inequality

Markov's inequality gives a generally coarse estimate of the probability that a random variable takes a value much larger than its mean.

Theorem. If R is a nonnegative random variable, then for all $x > 0$,

$$\Pr[R \geq x] \leq \frac{\mathbb{E}[R]}{x}.$$

Markov's inequality

Theorem. If R is a nonnegative random variable, then for all $x > 0$,

$$\Pr[R \geq x] \leq \frac{\mathbb{E}[R]}{x}.$$

Proof.

$$\begin{aligned}
\mathbb{E}[R] &= \sum_{y \in \text{range}(R)} y \cdot \Pr[R = y] \\
&\geq \sum_{y \in \text{range}(R), y \geq x} y \cdot \Pr[R = y] \\
&\geq \sum_{y \in \text{range}(R), y \geq x} x \cdot \Pr[R = y] \ = \ x \sum_{y \in \text{range}(R), y \geq x} \Pr[R = y] \\
&= x \Pr[R \geq x]
\end{aligned}$$

Markov's inequality, rephrased

Corollary. If R is a nonnegative random variable, then for all $c \geq 1$,

$$\Pr\left[R \geq c\mathbb{E}\left[R\right]\right] \leq \frac{1}{c}.$$

"No more than $1/c$ of the population can be c-times outliers."

No more than 10% of the population earns more than 10x the average income (assuming incomes are nonnegative).

Changing variables

Fix some random variable R. $|R|^z$ is also a nonnegative random variable. And $|R|^z \geq x^z \leftrightarrow |R| \geq x$.

$$\begin{aligned}\Pr[|R| \geq x] &= \Pr[|R|^z \geq x^z] \\ &\leq \frac{\mathbb{E}[|R|^z]}{x^z}\end{aligned}$$

$R - \mathbb{E}[R]$ is *also* a random variable. Plug this in:

$$\Pr[|R - \mathbb{E}[R]| \geq x] \leq \frac{\mathbb{E}[|R - \mathbb{E}[R]|^z]}{x^z}$$

(Hold onto this one a sec.)

Defining variance

The *variance* of a random variable R is defined to be

$$\text{Var}[R] = \mathbb{E}\left[(R - \mathbb{E}[R])^2\right].$$

Unpacking: at each outcome, measure the distance between R and its mean. Square this. Average this square over all outcomes.

If R is always close to its mean: variance is small. If R wanders away from its average a lot: variance is high.

Chebyshev's Inequality

Rephrasing our calculation from before:

$$\Pr[|R - \mathbb{E}[R]| \geq x] \leq \frac{\mathrm{Var}[R]}{x^2}$$

Variance lets us bound the probability that a variable is far from its mean.

Gambling example

Bet 1: Win \$2 with probability 2/3, lose \$1 with probability 1/3.
Bet 2: Win \$1002 with probability 2/3, lose \$2001 with probability 1/3.

$\mathbb{E}[B_1] = 2 \cdot 2/3 - 1 \cdot 1/3 = 1$
$\mathbb{E}[B_2] = 1002 \cdot 2/3 - 2001 \cdot 1/3 = 1$

$\text{Var}[B_1] = 2$
$\text{Var}[B_2] = 2{,}004{,}002$
(standard deviation = sq rt of variance sometimes more intuitive)

Computing variance

Theorem. $\text{Var}[R] = \mathbb{E}[R^2] - (\mathbb{E}[R])^2$.
Proof. Algebra and linearity (see the book!).

Particularly handy for indicator (Bernoulli) variables taking values in $\{0, 1\}$:

Corollary. If B is a Bernoulli random variable with $\Pr[B = 1] = p$, then $\text{Var}[B] = p - p^2 = p(1 - p)$.

Mean time to failure

Something disappointing about our mean time to failure analysis: no bound/restriction on the “long tail” of successes.

Reminder: if failure occurs independently with probability p, the expected number of successes before failure is $1/p$.

MTtF variance

Let C be the random variable measuring the number of successes before failure.

$$\begin{aligned}\mathrm{Var}[C] &= \mathbb{E}[C^2] - (\mathbb{E}[C])^2 \\ &= \mathbb{E}[C^2] - \frac{1}{p^2}\end{aligned}$$

Need to get a grasp on $\mathbb{E}[C^2]$. Reason about conditional expectations again:

$$\begin{aligned}\mathbb{E}[C^2] &= \mathbb{E}[C^2|\text{failure first}] \cdot \Pr[\text{failure first}] + \mathbb{E}[C^2|\text{success first}] \cdot \Pr[\text{success first}] \\ &= 1^2 \cdot p + (\mathbb{E}[(1+C)^2]) \cdot (1-p)\end{aligned}$$

MTtF variance continued

$$\begin{aligned}
\mathbb{E}[C^2] &= 1^2 \cdot p + (\mathbb{E}[(1+C)^2]) \cdot (1-p) \\
&= p + (1-p)(\mathbb{E}[C^2 + 2C + 1]) \\
&= p + (1-p)(\mathbb{E}[C^2] + 2\mathbb{E}[C] + 1) \\
&= p + (1-p)\mathbb{E}[C^2] + (1-p)(\frac{2}{p} + 1)
\end{aligned}$$

Solving for $\mathbb{E}[C^2]$ gives

$$\mathbb{E}[C^2] = \frac{2-p}{p^2}$$

so

$$\mathrm{Var}[C] = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$