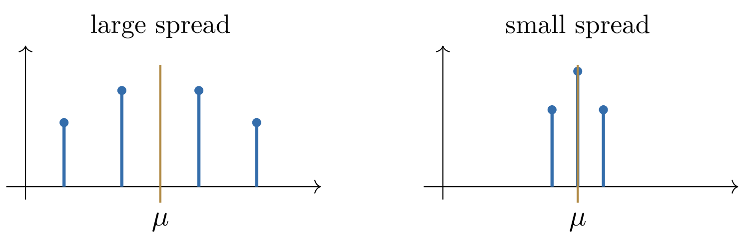
Chapter 10

Variance; Binomial, Geometric and Hypergeometric

Lecture 10 | Devore 9e: 3.3–3.5

10.1 Variance

The mean locates the distribution. It says nothing about how spread out it is: the two pmfs below have the same μ and could hardly be more different.



The natural measure of spread is the average squared distance from the mean.

Definition 10.1 (Variance and standard deviation)**.** Let X have mean $\mu = \mathbb{E}[X]$. The **variance** is

$$V(X) = \mathrm{Var}(X) = \sigma_X^2 := \mathbb{E}\big[(X-\mu)^2\big],$$

and the **standard deviation** is $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

Note that $\mathrm{Var}(X)$ is just $\mathbb{E}[h(X)]$ for $h(x) = (x-\mu)^2$, so the previous lecture already tells us how to compute it: $\mathrm{Var}(X) = \sum_{x_i \in D}(x_i - \mu)^2 \mathbb{P}\{X = x_i\}$.

Remark 10.2. Why square rather than take absolute values? Squaring is smooth (differentiable), it makes the algebra below work, and it is what turns independence into additivity later (Lecture 22). The price is that σ^2 has the *squared* units of X — which is why we also report σ, which is in the original units.

Theorem 10.3 (Computational formula)**.**

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

Proof. Expand and use linearity, remembering μ is a constant:

$$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2. \quad \square$$

So computing a variance is always the same three steps:

$$(1)\ \mathbb{E}[X], \qquad (2)\ \mathbb{E}[X^2], \qquad (3)\ \ \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Example 10.4 (Both routes on the same numbers)**.** A loaded die has pmf

x	1	2	3	4	5	6
$p(x)$	0.30	0.25	0.15	0.05	0.10	0.15

(The probabilities sum to 1. ✓) First the mean:

$$\mathbb{E}[X] = 1(0.30)+2(0.25)+3(0.15)+4(0.05)+5(0.10)+6(0.15) = 0.30+0.50+0.45+0.20+0.50+0.90 = 2.85.$$

Route 1 — straight from the definition.

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{x=1}^{6}(x-2.85)^2 p(x)\\ &= (-1.85)^2(0.30) + (-0.85)^2(0.25) + (0.15)^2(0.15)\\ &\quad + (1.15)^2(0.05) + (2.15)^2(0.10) + (3.15)^2(0.15)\\ &= 1.026750 + 0.180625 + 0.003375 + 0.066125 + 0.462250 + 1.488375\\ &= 3.2275.\end{aligned}$$

Route 2 — the computational formula.

$$\mathbb{E}[X^2] = 1(0.30)+4(0.25)+9(0.15)+16(0.05)+25(0.10)+36(0.15) = 0.30+1.00+1.35+0.80+2.50+5.40 = 11.35,$$

$$\mathrm{Var}(X) = 11.35 - (2.85)^2 = 11.35 - 8.1225 = 3.2275. \quad ✓$$

Either way $\sigma = \sqrt{3.2275} \approx 1.80$.

Remark 10.5. Route 2 is almost always less work — six squarings of small integers instead of six squarings of awkward decimals — and it is far less error-prone. Use the definition to understand what variance *means*; use $\mathbb{E}[X^2] - \mu^2$ to compute it.

Proposition 10.6. $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$ *and* $\sigma_{aX+b} = |a|\,\sigma_X$.

Proof. $\mathbb{E}[aX+b] = a\mu+b$, so $(aX+b)-(a\mu+b) = a(X-\mu)$, and squaring gives $a^2(X-\mu)^2$. Taking expectations, $\mathrm{Var}(aX+b) = a^2\mathbb{E}[(X-\mu)^2] = a^2\,\mathrm{Var}(X)$; the standard deviation is the square root, so it picks up $|a|$ rather than a. $\square$

Shifting does not change spread; rescaling scales it. That is exactly what a measure of spread should do.

Example 10.7 (Variance of the shooting variable)**.** With $\mathbb{P}\{X = n\} = (1-\beta)^{n-1}\beta$ and $\mathbb{E}[X] = 1/\beta$, we need $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \beta \sum_{n=1}^{\infty} n^2(1-\beta)^{n-1}.$$

Differentiate the geometric series twice. Since $\frac{d^2}{d\beta^2}(1-\beta)^{n+1} = (n+1)n(1-\beta)^{n-1} = n^2(1-\beta)^{n-1} + n(1-\beta)^{n-1}$, summing over $n \geq 1$ and differentiating term by term reduces everything to one series we can add up in closed form:

$$\sum_{n\geq 1}(1-\beta)^{n+1} = \frac{(1-\beta)^2}{\beta} = \frac{1}{\beta} - 2 + \beta, \qquad \frac{d^2}{d\beta^2}\left(\frac{1}{\beta} - 2 + \beta\right) = \frac{2}{\beta^3},$$

so $\sum_{n\geq 1}(n^2+n)(1-\beta)^{n-1} = \frac{2}{\beta^3}$. The first-derivative result of Lecture 9 already gave $\sum_{n\geq 1} n(1-\beta)^{n-1} = \frac{1}{\beta^2}$, and subtracting it leaves $\sum_{n\geq 1} n^2(1-\beta)^{n-1} = \frac{2}{\beta^3} - \frac{1}{\beta^2}$. Putting the pieces together,

$$\mathbb{E}[X^2] = \beta\left(\frac{2}{\beta^3} - \frac{1}{\beta^2}\right) = \frac{2}{\beta^2} - \frac{1}{\beta} = \frac{2-\beta}{\beta^2}, \qquad \text{hence} \qquad \mathrm{Var}(X) = \frac{2-\beta}{\beta^2} - \frac{1}{\beta^2} = \boxed{\frac{1}{\beta^2} - \frac{1}{\beta} = \frac{1-\beta}{\beta^2}}.$$

10.2 The binomial distribution

From here on we stop building distributions by hand and start naming the ones that occur over and over.

Definition 10.8 (Binomial experiment)**.** An experiment consisting of n *independent* trials, where

- each trial has exactly two outcomes, success (S) and failure (F);
- $\mathbb{P}(S) = p$ and $\mathbb{P}(F) = 1 - p \stackrel{\text{def}}{=} q$, the same on every trial.

Definition 10.9 (Binomial distribution)**.** In a binomial experiment let $X :=$ the number of successes among the n trials. Then $X \sim \mathrm{Binomial}(n, p)$, and its pmf is written $b(x; n, p)$.

Theorem 10.10.

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n, \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. The sample space consists of all strings of length n over $\{S, F\}$, e.g. $SSF\cdots FS$. By independence, any *particular* string with x successes has probability

$$\mathbb{P}(S)\mathbb{P}(S)\mathbb{P}(F)\cdots\mathbb{P}(F)\mathbb{P}(S) = p^x(1-p)^{n-x},$$

regardless of the order the letters appear in. The event $\{X = x\}$ is the disjoint union of all such strings, and there are $\binom{n}{x}$ of them (choose which x of the n positions hold the successes). Adding up, $\mathbb{P}\{X = x\} = \binom{n}{x}p^x(1-p)^{n-x}$. ∎

Theorem 10.11. *If* $X \sim \text{Binomial}(n, p)$ *then* $\mathbb{E}[X] = np$ *and* $\text{Var}(X) = np(1-p)$.

(The slick proof writes $X = \mathbf{1}_1 + \cdots + \mathbf{1}_n$ as a sum of indicators; we give it in Lecture 22 once we have additivity of variance.)

10.3 The geometric distribution

Definition 10.12 (Geometric distribution)**.** In a binomial experiment, let $X :=$ the number of trials up to and including the first success. Then $X \sim \text{Geometric}(p)$ with pmf

$$g(x; p) = (1-p)^{x-1} p, \qquad x = 1, 2, 3, \ldots$$

This is the shooting-at-a-target variable from Lectures 8–9, so we already know

$$\mathbb{E}[X] = \frac{1}{p}, \qquad \text{Var}(X) = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.$$

Remark 10.13. Binomial and geometric come from the *same* experiment; they differ in what is fixed and what is random. Binomial: n fixed, count the successes. Geometric: one success required, count the trials. It is named after the geometric series, which is where its normalization and its mean came from.

Remark 10.14 (Warning)**.** Some books (and software) define the geometric as the number of *failures* before the first success, i.e. shifted down by one, with mean $(1-p)/p$. We always use the "number of trials" convention, starting at $x = 1$.

10.4 The hypergeometric distribution

Binomial sampling is *with* replacement (or from an inexhaustible population), so p never changes. What if we sample without replacement?

Definition 10.15 (Hypergeometric distribution)**.** A population has N objects, of which M are "good" and $N - M$ are "bad". Select n objects without replacement (all at once). Let $X :=$ the number of good objects selected. Then $X \sim \text{Hypergeometric}(n, M, N)$ with

$$\mathbb{P}\{X = x\} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

valid for integers x with $0 \le x \le M$, $0 \le n - x \le N - M$ (and $n \le N$).

The formula is pure counting: choose which x of the M good ones you got, choose which $n - x$ of the $N - M$ bad ones you got, and divide by the total number of samples.

Theorem 10.16. *With $p := M/N$ (the population fraction of good objects),*

$$\mathbb{E}[X] = n \cdot \frac{M}{N} = np, \qquad \operatorname{Var}(X) = \underbrace{\frac{N-n}{N-1}}_{\textit{finite population correction}} \cdot\, np(1-p).$$

Remark 10.17 (Hypergeometric versus binomial)**.** The mean is *identical* to the binomial mean. The variance carries an extra factor $\frac{N-n}{N-1} \le 1$: sampling without replacement is more informative, so the count is less variable.

If $n \ll N$ then $\frac{N-n}{N-1} \approx 1$, and removing a few objects barely changes the composition of the population. In that regime

$$\mathbb{P}\{X_{\text{hyp}} = x\} \approx \mathbb{P}\{X_{\text{bin}} = x\},$$

which is why polling 1000 people out of 300 million is treated as a binomial problem. (Homework 5 asks you to make this precise.)