Chapter 23

Minimum Variance Estimators; Method of Moments

Lecture 23 | Devore 9e: 6.1–6.2

23.1 The sample variance is unbiased

We owe a proof from last lecture. First an algebraic identity.

Lemma 23.1.

$$\sum_{i=1}^{n}(X_i-\bar X)^2=\sum_{i=1}^{n}X_i^2-n\bar X^2=\sum_{i=1}^{n}X_i^2-\frac{1}{n}\Big(\sum_{i=1}^{n}X_i\Big)^2.$$

Proof. Expand the square and use $\sum_i X_i = n\bar X$:

$$\sum_i (X_i-\bar X)^2=\sum_i X_i^2-2\bar X\sum_i X_i+n\bar X^2=\sum_i X_i^2-2n\bar X^2+n\bar X^2=\sum_i X_i^2-n\bar X^2. \qquad \square$$

Theorem 23.2. *If $X_1,\dots,X_n$ are i.i.d. with mean μ and variance σ^2, then $\mathbb{E}[S^2]=\sigma^2$.*

Proof. The two ingredients are $\mathbb{E}[Z^2]=\operatorname{Var}(Z)+\mathbb{E}[Z]^2$ applied twice:

$$\mathbb{E}[X_i^2]=\sigma^2+\mu^2, \qquad \mathbb{E}\Big[\Big(\sum_i X_i\Big)^2\Big]=\operatorname{Var}\Big(\sum_i X_i\Big)+\Big(\mathbb{E}\Big[\sum_i X_i\Big]\Big)^2=n\sigma^2+(n\mu)^2,$$

the middle step using independence. Then by the lemma,

$$\begin{aligned}\mathbb{E}[S^2]&=\frac{1}{n-1}\,\mathbb{E}\Big[\sum_i X_i^2-\tfrac{1}{n}\big(\textstyle\sum_i X_i\big)^2\Big]\\&=\frac{1}{n-1}\Big\{n(\sigma^2+\mu^2)-\tfrac{1}{n}\big(n\sigma^2+n^2\mu^2\big)\Big\}\\&=\frac{1}{n-1}\Big\{n\sigma^2+n\mu^2-\sigma^2-n\mu^2\Big\}=\frac{(n-1)\sigma^2}{n-1}=\sigma^2. \qquad \square\end{aligned}$$

23.2 Which unbiased estimator is best?

Unbiasedness is a weak requirement: many estimators have it, and they can be very different. Recall the Uniform$[0, \theta]$ problem.

Example 23.3 (Two unbiased estimators of θ)**.** $X_1, \dots, X_n$ i.i.d. Uniform$[0, \theta]$.

- From last lecture, $\hat{\theta}_1 = \dfrac{n+1}{n} \max(X_1, \dots, X_n)$ is unbiased.
- Also $\hat{\theta}_2 = 2\bar{X}$ is unbiased, since $\mathbb{E}[X_i] = \theta/2$ gives $\mathbb{E}[\hat{\theta}_2] = 2 \cdot \frac{\theta}{2} = \theta$.

Both hit the target on average. Now compare their variances. Write $M = \max(X_1, \dots, X_n)$, whose density $f_M(t) = nt^{n-1}/\theta^n$ we found last lecture. Then

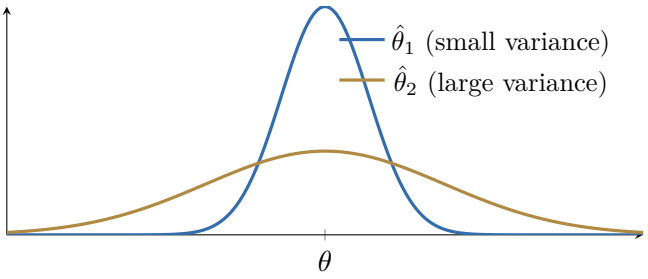
$$\mathbb{E}[M^2] = \int_0^\theta t^2 \cdot \frac{nt^{n-1}}{\theta^n}\, \mathrm{d}t = \frac{n\theta^2}{n+2}, \qquad \mathrm{Var}(M) = \frac{n\theta^2}{n+2} - \Big(\frac{n\theta}{n+1}\Big)^2 = \frac{n\theta^2}{(n+2)(n+1)^2},$$

the last step because $(n+1)^2 - n(n+2) = 1$. Since $\hat{\theta}_1 = \frac{n+1}{n} M$,

$$\mathrm{Var}(\hat{\theta}_1) = \Big(\frac{n+1}{n}\Big)^2 \mathrm{Var}(M) = \frac{\theta^2}{n(n+2)}, \qquad \mathrm{Var}(\hat{\theta}_2) = \frac{4}{n}\,\mathrm{Var}(X_i) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Since $n(n+2) > 3n$ for $n > 1$,

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2) \qquad \Longrightarrow \qquad \hat{\theta}_1 \text{ is the better estimator.}$$



Remark 23.4. The difference is dramatic: $\mathrm{Var}(\hat{\theta}_1)$ shrinks like $1/n^2$ while $\mathrm{Var}(\hat{\theta}_2)$ shrinks only like $1/n$. Using the maximum extracts far more information about θ than the average does — which makes sense, since only the largest observation carries information about where the interval ends.

Definition 23.5 (MVUE)**.** Among all unbiased estimators of θ, the one with the smallest variance is the **minimum variance unbiased estimator** (MVUE).

Remark 23.6 (Is a biased estimator always worse?)**.** No. An estimator with a small bias and a much smaller variance can beat an unbiased one in every practical sense. The standard way to trade the two off is the **mean squared error**

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = \mathrm{Var}(\hat{\theta}) + \big(\mathrm{bias}(\hat{\theta})\big)^2,$$

which penalizes both. Unbiasedness is a convenient constraint, not a sacred one.

23.3 Constructing estimators

So far we have *checked* estimators that were pulled out of thin air. Two systematic recipes produce them:

- the **method of moments** (today);
- the **method of maximum likelihood** (Lecture 24).

Definition 23.7 (Population and sample moments)**.** The **k-th (population) moment** of a distribution is $\mathbb{E}[X^k]$; the first moment is the mean. The **k-th sample moment** of $X_1, \dots, X_n$ is

$$\overline{X^k} := \frac{X_1^k + X_2^k + \cdots + X_n^k}{n}, \qquad \text{so in particular} \qquad \overline{X^1} = \bar{X}.$$

Definition 23.8 (Method of moments)**.** Let $X_1, \dots, X_n$ be i.i.d. from a distribution with m unknown parameters $\theta_1, \dots, \theta_m$. Set the first m population moments equal to the corresponding sample moments,

$$\mathbb{E}[X^k] = \overline{X^k}, \qquad k = 1, 2, \dots, m,$$

and solve the resulting m equations for $\theta_1, \dots, \theta_m$.

The idea in one line:

$$\underbrace{\mathbb{E}[X^k]}_{\text{an expression in } \theta_1,\dots,\theta_m} = \underbrace{\overline{X^k}}_{\text{a number computed from the data}}.$$

The left side is a formula in the unknowns; the right side is something you can evaluate. Write down as many such equations as you have unknowns, then solve. The justification is the law of large numbers: $\overline{X^k} \to \mathbb{E}[X^k]$ as n grows, so matching them is matching quantities that genuinely become equal.

23.3.1 Gamma

Example 23.9. $X_1, \dots, X_n$ i.i.d. Gamma(α, β), so $m = 2$ unknowns. We know

$$\mathbb{E}[X] = \alpha\beta, \qquad \mathrm{Var}(X) = \alpha\beta^2,$$

hence $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2 = \alpha\beta^2 + \alpha^2\beta^2$. The two equations are

$$\begin{cases} \bar{X} = \alpha\beta, \\ \overline{X^2} = \alpha\beta^2 + \alpha^2\beta^2. \end{cases}$$

Subtract the square of the first equation from the second. Since $\bar{X}^2 = (\alpha\beta)^2 = \alpha^2\beta^2$, the $\alpha^2\beta^2$ terms cancel:

$$\overline{X^2} - \bar{X}^2 = \alpha\beta^2.$$

Now divide this by the first equation to eliminate α:

$$\frac{\overline{X^2} - \bar{X}^2}{\bar{X}} = \frac{\alpha\beta^2}{\alpha\beta} = \beta,$$

and substitute back into $\bar{X} = \alpha\beta$ to recover α:

$$\boxed{\hat{\beta} = \frac{\overline{X^2} - \bar{X}^2}{\bar{X}}, \qquad \hat{\alpha} = \frac{\bar{X}}{\hat{\beta}} = \frac{\bar{X}^2}{\overline{X^2} - \bar{X}^2}}$$

Remark 23.10. Sanity check on the structure: $\overline{X^2} - \bar{X}^2$ is the (biased) sample variance, so the estimators read $\hat{\beta} = \widehat{\mathrm{Var}}/\widehat{\mathbb{E}}$ and $\hat{\alpha} = \widehat{\mathbb{E}}^2/\widehat{\mathrm{Var}}$ — exactly the inversion of $\mathbb{E} = \alpha\beta$, $\mathrm{Var} = \alpha\beta^2$.

With the ten observations $152, 115, 109, 94, 88, 125, 40, 128, 123, 136$ you get $\bar{x} = 1110/10 = 111$ and $\overline{x^2} = 132024/10 = 13202.4$, so $\overline{x^2} - \bar{x}^2 = 13202.4 - 12321 = 881.4$ and

$$\hat{\beta} = \frac{881.4}{111} \approx 7.94, \qquad \hat{\alpha} = \frac{12321}{881.4} \approx 13.98.$$

23.3.2 Negative binomial

Example 23.11. $X_1, \ldots, X_n$ i.i.d. NegBinomial(r, p), where *both* r and p are unknown. Here

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

Set these equal to $\bar{X}$ and $\overline{X^2}$. As with the gamma, the system is easier after subtracting the square of the first equation from the second, which replaces the second moment by the variance:

$$\bar{X} = \frac{r(1-p)}{p}, \qquad \overline{X^2} - \bar{X}^2 = \mathrm{Var}(X) = \frac{r(1-p)}{p^2}.$$

Dividing the first by the second cancels r and $1 - p$ outright:

$$\frac{\bar{X}}{\overline{X^2} - \bar{X}^2} = \frac{r(1-p)/p}{r(1-p)/p^2} = p \qquad \Longrightarrow \qquad \hat{p} = \frac{\bar{X}}{\overline{X^2} - \bar{X}^2}.$$

Solving the first equation for r gives $r = \bar{X}\, p/(1-p)$; substituting $1 - \hat{p} = \dfrac{\overline{X^2} - \bar{X}^2 - \bar{X}}{\overline{X^2} - \bar{X}^2}$ yields

$$\hat{r} = \frac{\bar{X}^2}{\overline{X^2} - \bar{X}^2 - \bar{X}}.$$

Remark 23.12. Note that $\hat{p}$ and $\hat{r}$ are random variables (functions of $X_1, \ldots, X_n$); plugging in the observed data turns them into numbers.

23.4 Where the method of moments runs out

1. **Computing $\mathbb{E}[X^k]$ can be hard.** The method needs closed-form population moments. For the negative binomial above, the second moment was already unpleasant; for many distributions there is no elementary formula at all.
2. **Solving the system can be hard.** With $m \geq 2$ parameters the moment equations are non-linear and need not have a closed-form solution — or any solution.
3. **The answer can fall outside the parameter space.** In the negative binomial, if the sample happens to have variance smaller than its mean, then $\overline{X^2} - \bar{X}^2 - \bar{X} < 0$ and $\hat{r}$ comes out *negative* — impossible, since r counts successes. Nothing in the method prevents this.

Remark 23.13. Method-of-moments estimators are cheap to compute and are a good way to get a starting value, but they carry no guarantee: not of landing in the parameter space, and not of being efficient. Maximum likelihood, next lecture, is built to fix both.