Chapter 20

Samples, Statistics, and Sampling Distributions

Lecture 20 | Devore 9e: 5.3

This lecture is the hinge of the course. Until now the distribution was given and we computed probabilities. From here on the distribution is *unknown* and we have only data — but the data is itself random, so probability is exactly the tool we need to describe it.

20.1 Samples and statistics

Definition 20.1 (Sample and sample data)**.** A **sample** is a collection of random variables $X_1, X_2, \ldots, X_n$. The **sample data** $x_1, x_2, \ldots, x_n \in \mathbb{R}$ is a collection of observed values of that sample.

The distinction is the whole point. Before you run the experiment, X_i is a random variable; afterwards, x_i is a number. Capital letters are random, lower case are data.

Definition 20.2 (Statistic)**.** A **statistic** is any quantity whose value can be computed from the sample data — equivalently, any random variable expressible as a function of $X_1, \ldots, X_n$.

Example 20.3. The two we care about most:

$$\bar{X} := \frac{X_1 + \cdots + X_n}{n} \qquad \text{(sample mean)},$$

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{(sample variance)}, \qquad S = \sqrt{S^2}.$$

But there are many others — $\frac{X_1}{X_1+X_2}$, $e^{X_1+X_2^2}$, $\left(\sum_i X_i^k\right)^{1/k}$ — and each is a legitimate statistic.

Remark 20.4 (Why $n-1$?)**.** Dividing by $n-1$ rather than n makes $\mathbb{E}[S^2] = \sigma^2$ exactly, for every n. We prove this in Lecture 23; the short reason is that $\bar{X}$ has already been estimated from the same data, using up one “degree of freedom”.

Remark 20.5 (The central observation).

$$\boxed{\text{A statistic is a random variable, so it has a distribution.}}$$

That distribution is called the **sampling distribution** of the statistic, and essentially all of statistics consists of working out what it is.

20.2 What the sampling distribution depends on

(1) The underlying distribution of the sample

If every X_i is identically 0, then $\bar{X} = 0$ always — a degenerate sampling distribution. If $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$, the answer is quite different.

(2) The method of sampling

Suppose you measure the heights $H_1, \ldots, H_{100}$ of 100 people and form $\bar{H}$. If you sampled 100 strangers, $\bar{H}$ behaves one way. If you sampled 100 members of the same family, the H_i are correlated and $\bar{H}$ is far more variable than it looks.

Push it to the extreme to see how badly this can go. Suppose $X_1 \sim \text{Normal}(0,1)$ but you accidentally record the *same* person n times, so

$$X_2 = X_1, \quad X_3 = X_1, \quad \ldots, \quad X_n = X_1.$$

Then

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{nX_1}{n} = X_1,$$

and therefore

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_1 - X_1)^2 = 0 \qquad \Longrightarrow \qquad S^2 = 0.$$

The sample reports zero variability, with total confidence, about a population whose variance is 1. Nothing in the formula for S^2 went wrong; the *sampling* did. No amount of data fixes this — taking $n = 10^6$ copies of the same person still gives $S^2 = 0$.

Definition 20.6 (Random sample of size n). $X_1, \ldots, X_n$ is a **random sample** (equivalently: the X_i are **i.i.d.**) if

1. $X_1, \ldots, X_n$ are independent, and
2. each X_i has the same distribution.

From now on "sample" means random sample unless stated otherwise. Both conditions are assumptions about how the data was *collected*, and no amount of clever computation can repair a violation of them.

20.3 Two ways to find a sampling distribution

1. **Closed-form derivation.** Works when the underlying distribution and the statistic are both simple.
2. **Numerical simulation.** Draw many samples, compute the statistic each time, histogram the results. This is the only option once things get complicated — and it always works.

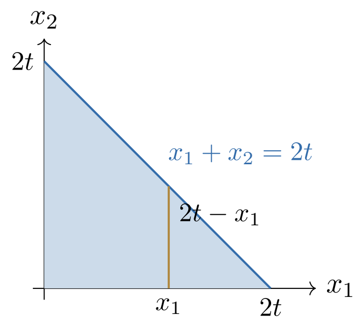
20.3.1 A closed-form example

Example 20.7. Let X_1, X_2 be i.i.d. Exponential(λ) and $\bar{X} = \frac{X_1+X_2}{2}$. Find the cdf and pdf of $\bar{X}$.

As always, go through the cdf. For $t > 0$,

$$F_{\bar{X}}(t) = \mathbb{P}\{\bar{X} \le t\} = \mathbb{P}\{X_1 + X_2 \le 2t\} = \iint_{\{x_1+x_2 \le 2t\}} f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2.$$

By independence $f(x_1, x_2) = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2}$, and the region is the triangle $\{x_1, x_2 \ge 0,\ x_1 + x_2 \le 2t\}$:



Slice vertically:

$$\begin{aligned}
F_{\bar{X}}(t) &= \int_0^{2t} \left(\int_0^{2t-x_1} \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2}\, \mathrm{d}x_2 \right) \mathrm{d}x_1 \\
&= \int_0^{2t} \lambda e^{-\lambda x_1} \left(1 - e^{-\lambda(2t-x_1)}\right) \mathrm{d}x_1 \\
&= \int_0^{2t} \left(\lambda e^{-\lambda x_1} - \lambda e^{-2\lambda t}\right) \mathrm{d}x_1 \\
&= 1 - e^{-2\lambda t} - 2\lambda t\, e^{-2\lambda t}, \qquad t \ge 0.
\end{aligned}$$

Differentiating,

$$f_{\bar{X}}(t) = F'_{\bar{X}}(t) = 4\lambda^2 t\, e^{-2\lambda t}, \qquad t \ge 0.$$

Remark 20.8. Notice what happened: the average of two exponentials is *not* exponential — its density vanishes at $t = 0$ and has an interior peak. In fact $X_1 + X_2 \sim \text{Gamma}(2, 1/\lambda)$, exactly as the additivity property of Lecture 16 predicts, and $\bar{X}$ is that divided by 2.

Also notice how much work this was for $n = 2$ with the simplest possible underlying distribution. Doing it for $n = 100$, or for a distribution without a nice closed form, is hopeless. That is why the next lecture's theorem is so valuable: it gives the approximate sampling distribution of $\bar{X}$ for *every* underlying distribution, for free.

Remark 20.9 (Sanity checks that cost nothing)**.** Whatever the underlying distribution, if $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2 then

$$\mathbb{E}[\bar{X}] = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

We prove these next lecture. Check them against the example above: Exponential(λ) has $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$, so we should find $\mathbb{E}[\bar{X}] = 1/\lambda$ and $\mathrm{Var}(\bar{X}) = 1/(2\lambda^2)$. Integrating $f_{\bar{X}}(t) = 4\lambda^2 t e^{-2\lambda t}$ against t and t^2, using $\int_0^\infty t^k e^{-2\lambda t}\, \mathrm{d}t = k!/(2\lambda)^{k+1}$,

$$\mathbb{E}[\bar{X}] = 4\lambda^2 \cdot \frac{2}{(2\lambda)^3} = \frac{1}{\lambda}, \qquad \mathbb{E}[\bar{X}^2] = 4\lambda^2 \cdot \frac{6}{(2\lambda)^4} = \frac{3}{2\lambda^2},$$

so $\mathrm{Var}(\bar{X}) = \frac{3}{2\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{2\lambda^2} = \sigma^2/2$. ✓