9. Sampling and Sampling Distributions

> *Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.*
>
> Francis Galton, 1889

Populations and Samples

We've talked about populations, samples, parameters and statistics. A **parameter** is a number that describes the population, such as its mean or standard deviation. A **statistic** is an estimate on parameter based on samples. A parameter is fixed (under the Frequentist view) but usually unknown; a statistic is known but changes from sample to sample.

Table 9.1: Notation for statistics and parameters.

Measure	Sample (Statistic)	Population (Prameter)
Number of cases	n	N
Mean	$\bar{x}$	μ
Proportion	$\hat{p}$	p
Standard deviation	s	σ
Variance	s^2	σ^2

Probability Sampling

Probability sampling is a technique of drawing a sample in which each sampling unit has a known, non-zero chance of being included in the sample.

Simple Random Sampling

In a **simple random sample**, every case in the population has the same chance of being chosen, and every combination of n cases has the same chance of being the sample.

Systematic Random Sampling

A **systematic random sample** takes every Kth case on a list, with $K = N/n$, starting from a case chosen at random among the first K.

Example 9.1. To choose 8 of the 40 parcels on a list, take $K = 40/8 = 5$ and draw a random start between 1 and 5. For example, if the start is 4, the sample is parcels 4, 9, 14, 19, 24, 29, 34 and 39.

Stratified, Cluster, and Multistage Samples

Stratified sample: divide the population into groups (**strata**), such as neighborhoods, and draw a random sample within each. Every group is represented, and the estimates are usually more precise than a simple random sample of the same size.

Cluster sample: randomly select whole groups (**clusters**), and include every unit in them. It is more convenient but the estimates are usually less precise, because units in one cluster are more likely to be similar with each other.

Multistage sample: select clusters, then units within them, and so on. Most large surveys combine several of these designs. The ACS, for example, draws addresses from the Census Bureau's Master Address File and samples small areas at higher rates than large ones, so each response carries a **weight**, representing the number of households it stands for. Check ACS Sample Design and Selection.

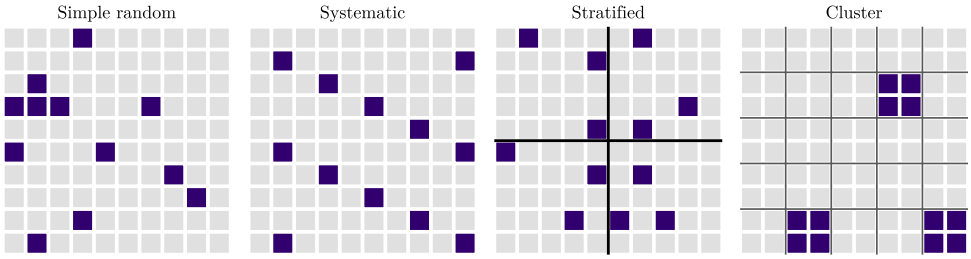


Figure 9.1: Four ways to select 12 of 100 housing units.

Non-Probability Samples

A **non-probability sample** is chosen by some other means, and its results cannot be generalized with a known margin of error. Non-probability samples are useful for discovering issues, reaching hidden populations, and piloting a questionnaire, but not for estimating the parameters.

Table 9.2: Non-probability samples.

Design	How units are chosen	Example
Convenience	Whoever is easy to reach	Surveys at one transit station
Voluntary response	Whoever chooses to respond	An online comment form, a public hearing
Snowball	Participants recruit others they know	Interviews with informal-economy workers
Quota	The researcher fills set numbers per group, without randomness	A street survey with 50 renters and 50 owners

Public Participation Is Not a Probability Sample

Planners summarize public comment all the time. The people who comment chose to comment, so comment data are a voluntary-response sample.

Example 9.2. Housing Politics Lab at Boston University collected every public comment on proposals for more than one housing unit at planning and zoning board meetings in 97 cities and towns around Boston, 2015-2017, and looked up the 3,123 speakers in the state voter file (Einstein et al. 2019, 2023).

- **What they said:** 63% of comments opposed the project and 14.6% supported it. In the 2010 referendum on Chapter 40B, the state's affordable housing law, 56% of voters in the same towns voted to keep it.

- **Who spoke:** Women were 51.3% of registered voters but 43.3% of commenters. In Arlington, renters were 39% of residents but 22% of commenters.
- **How often:** 17% of commenters spoke at more than one meeting, and the average speaker made 1.3 comments.
- **Online.:** At Zoom meetings in 2020, commenters were still more often white (by 13 percentage points), over 50 (by 22 points) and homeowners (by 25 points) than voters, and 13% of them supported the project .

Sampling Distributions

A statistic is an estimate on parameter based on samples, such as the sample mean $\bar{X}$. We care whether this statistic is good and how good it is. Before knowing the samples, **a statistic is a random variable**. Its distribution over all possible samples is the **sampling distribution** of the statistic. Three distributions are easy to confuse:

- The **population distribution** is the distribution of the variable over all cases in the population.
- The **sample distribution** is the distribution of the variable in the one sample we drew. A histogram or CDF of the sample shows it.
- The **sampling distribution** is the distribution of a statistic, such as $\bar{X}$, over repeated random samples of size n.

The following figure treats the 1,480 workers in the PSRC's 2023 Household Travel Survey who reported a trip from home to work as the population, and each worker's commute time as the variable (top, population distribution).

A random sample of n workers (middle, sample distribution) can give us one sample mean. Repeatedly, drop the sample mean onto the pile and it becomes the sampling distribution of $\bar{X}$.

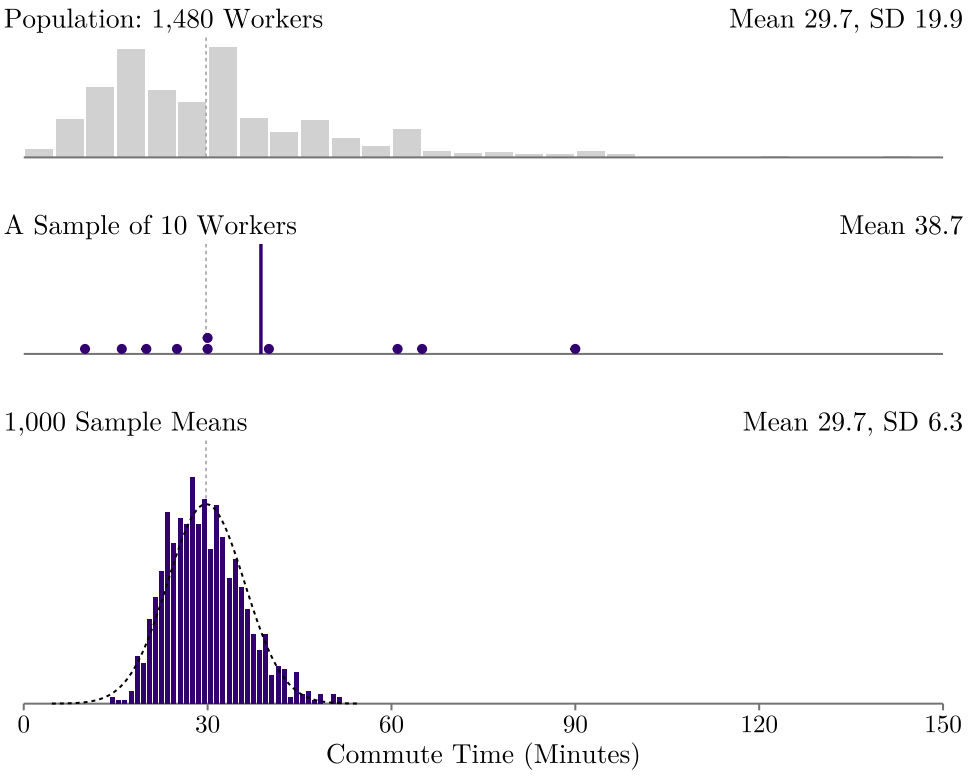


Figure 9.2: Commute times of 1,480 workers, PSRC Household Travel Survey, 2023. Top, the population; middle, one random sample of n workers and its mean; bottom, the means of many samples. The grey dashed line marks population mean μ.

Why the Mean

The rest of this lecture largely follows sample mean, for some reasons. Two of them are:

It is the quantity we care about. The mean is a parameter of the common distributions: μ of the normal, λ of the Poisson, p of a 0/1 variable. A mean times the number of cases is also a total; medians do not add up this way.

It has worked-out theory. It is unbiased at every n, its standard error is $\sigma/\sqrt{n}$, and its shape follows the CLT, all shown below. A median has no simple standard error formula.

Center

Like any distribution, a sampling distribution has a center, a spread and a shape. Its center is the expected value of $\bar{X}$:

$$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot n\mu = \mu$$

The sample mean is an **unbiased estimator** because its sampling distribution centers exactly on the population mean. Even though one sample's mean still misses μ, unbiased means the misses have no preferred direction.

Spread

The spread needs independent draws, each with SD σ.

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \qquad \mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

The SD of a sampling distribution is the **standard error** of the statistic. Intuitively, the more samples we have, the more precise the estimate should be. Also, the precision depends on how much the population itself varies. If the residents of an area are all about the same age, a small sample already tells us their average age; if their ages run from infants to retirees, the same sample can land far from the average.

Example 9.3. A survey samples 25 residents in each of two neighborhoods (made up). In the first, a retirement community, residents' ages have an SD of $\sigma = 5$ years. In the second, a neighborhood of young families and retirees, the SD is $\sigma = 20$ years.

$$\mathrm{SE}_1 = \frac{5}{\sqrt{25}} = 1 \text{ year}, \qquad \mathrm{SE}_2 = \frac{20}{\sqrt{25}} = 4 \text{ years}$$

Example 9.4. Howard Wainer calls $\mathrm{SE} = \sigma/\sqrt{n}$ "the most dangerous equation" (Wainer 2007). In the 1990s small schools were promoted because they were overrepresented among the top-scoring schools, and the Gates Foundation paid to split large high schools. Wainer and Harris Zwerling ranked Pennsylvania's 1,662 schools by mean 5th-grade reading score and counted the smallest 3 percent of schools at each end. If size did not matter, each share would be about 3 percent.

$$\text{top 50: } \frac{6}{50} = 12\%, \qquad \text{bottom 50: } \frac{9}{50} = 18\%$$

Small schools are not better; their mean scores are noisier. With few students, $\sigma/\sqrt{n}$ is large, so small schools land at both ends of a ranking.

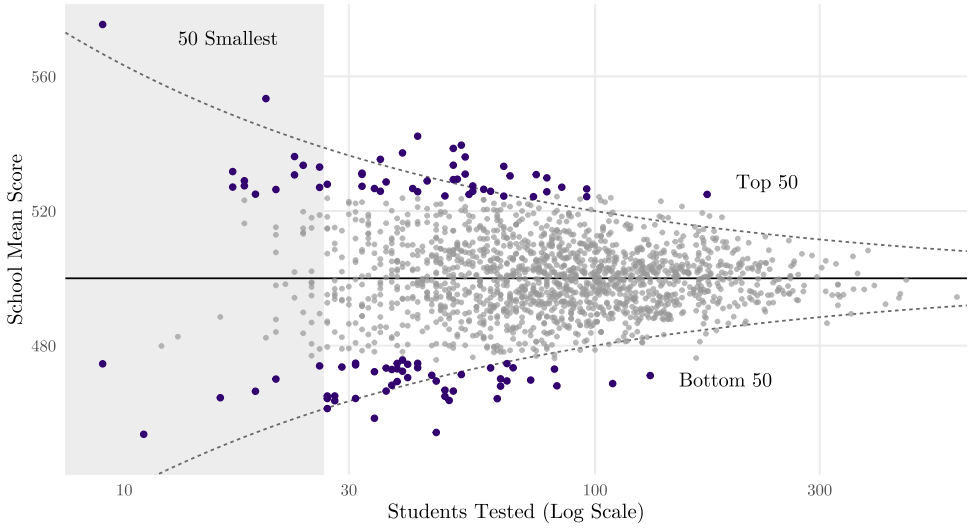


Figure 9.3: 1,662 simulated schools that share one true mean score (solid line). The dashed curves mark 2 standard errors, $2\sigma/\sqrt{n}$, either side.

The Law of Large Numbers

The following table repeats the draw of Figure 9.2 4,000 times at each sample size. The means center on the population mean at every n, and their SD follows $\sigma/\sqrt{n}$.

Sample size n	Mean of the sample means	SD of the sample means	$\sigma/\sqrt{n}$	Skewness
1	29.7	19.9	19.9	1.6
2	29.6	13.8	14.1	1.0
5	29.8	9.0	8.9	0.8
10	29.7	6.3	6.3	0.5
30	29.8	3.6	3.6	0.2
60	29.7	2.5	2.6	0.2

Table 9.3: Means of 4,000 random samples of the 1,480 workers' commute times (minutes) at each sample size. The first row is the population itself (one worker at a time). The samples are drawn without replacement, so at $n = 60$ their SD falls slightly below $\sigma/\sqrt{n}$.

As n grows, the standard error shrinks toward zero, so the sample mean settles on μ. This is the **law of large numbers**. The law of large numbers is also what makes the frequentist reading of a probability: "the long-run relative frequency" is the limit this law guarantees.

Central Limit Theorem

The center and the spread say where the sample mean lands and how closely. The **central limit theorem (CLT)** gives the third property of its sampling distribution, the shape: for large n, the sum or average of independent draws is approximately normal, whatever the shape of the distribution the draws come from.

A Galton board shows the CLT with coin flips. At each row of pegs a ball goes right (1) or left (0) with probability 1/2, so after n rows its bin is the sum of n coin flips, $\text{Bin}(n, 0.5)$.

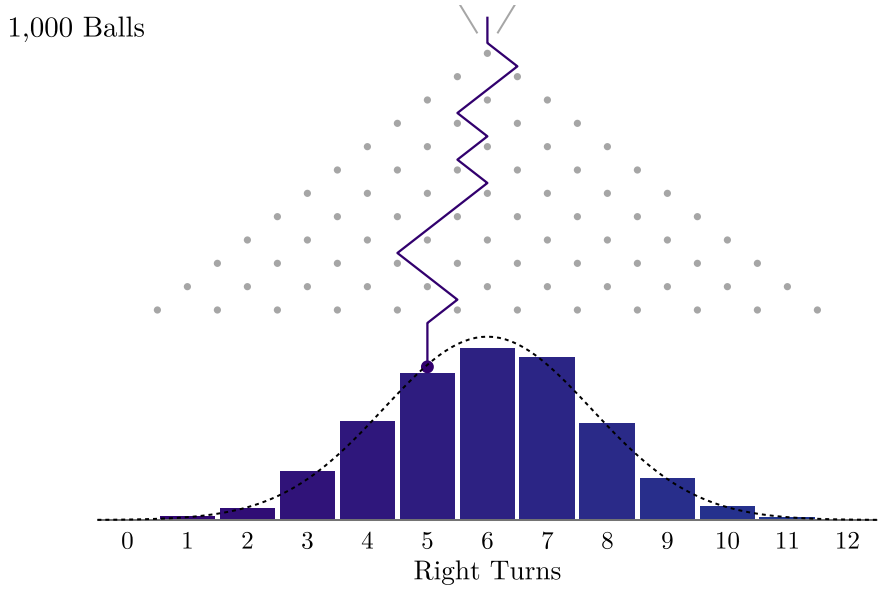


Figure 9.4: A Galton board. At each of n rows a ball goes left or right with probability 1/2, so its bin is $\text{Bin}(n, 0.5)$.

In Figure 9.5 each draw is a roll of a die with different probability for each number. We sum all n rolls (sampling distribution) and draw the histogram.

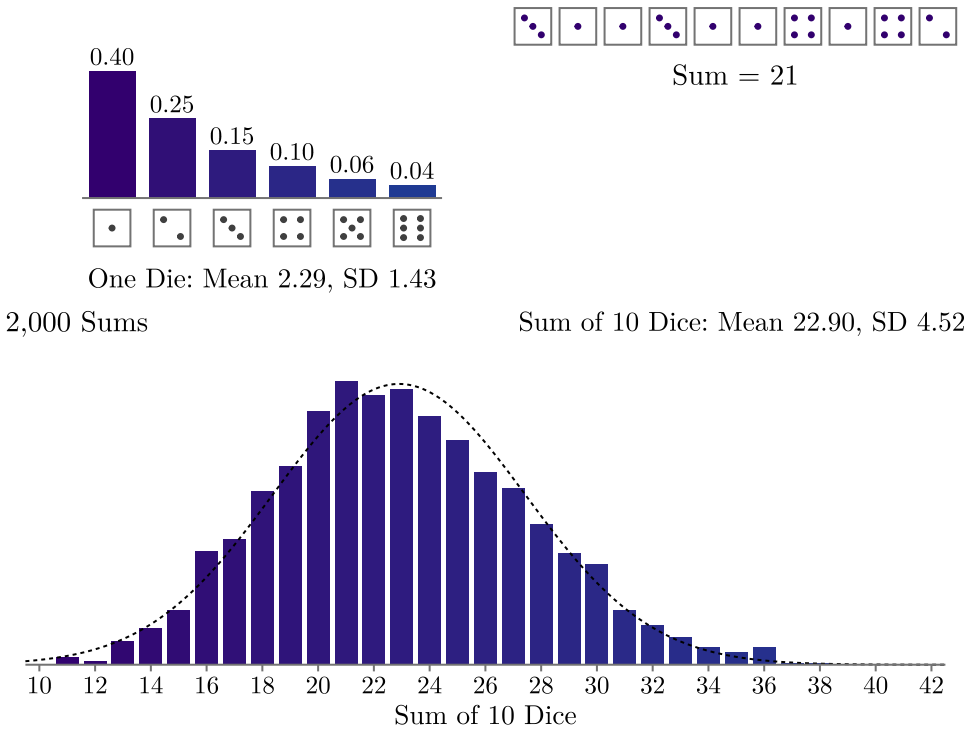


Figure 9.5: Sums of a loaded die. Top left, the probability of each face; top right, one roll of n dice and its sum; below, the sums of many rolls.

Conditions

> *Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.*
>
> Francis Galton, 1889

The cost of using The CLT needs some conditions:

Independent draws. A simple random sample can work. Drawing without replacement is close enough when the sample is a small part of the population.

A finite SD. The population's σ must exist. Real data always have one, but a few extreme values (a handful of 50-million-dollar sales among ordinary home sales) make a population very skewed, which raises the n the next condition needs.

A large enough n. How large depends on how skewed the population is. In Table 9.3 the commute times have skewness 1.6; means of 10 workers have 0.5, and means of 30 have 0.2, close to the normal's 0. The usual rule of thumb is $n \geq 30$: a nearly symmetric population needs fewer, a very skewed one more.

Example 9.5. Back to the 1,480 workers of Figure 9.2: $\mu = 29.7$ minutes, $\sigma = 19.9$ minutes, and a right-skewed population. Take a random sample of $n = 30$, 2 percent of the population. The center and the spread give the mean and standard error of $\bar{X}$. The CLT adds that its sampling distribution is close to normal, so by the empirical rule about 95 percent of sample means fall within 2 standard errors of μ:

$$\text{SE} = \frac{19.9}{\sqrt{30}} \approx 3.6, \qquad 29.7 \pm 2 \times 3.6 = 22.5 \text{ to } 36.9 \text{ minutes}$$