MTL108
Sampling Distributions-I

Rahul Singh

Sample mean

We learned how to describe a single sample using mean, variance, skewness, and kurtosis. Now, we move to statistical inference: using that single sample to make confident claims about the unknown population. The engine that makes this possible is the **sampling distribution**.

What is a Sampling Distribution?

In practice, we only take *one* sample of size n from a population. However, theoretical statistics requires us to imagine taking *every possible* sample of size n from that population.

If we calculate the sample mean ($\bar{x}$) for every single one of these possible samples, we will get a massive collection of different $\bar{x}$ values.

- **Definition:** The probability distribution of all possible values of a sample statistic (like $\bar{x}$) computed from samples of the same size n from the same population is called the **sampling distribution** of that statistic.

Properties of the Sampling Distribution of the Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean $\mathbb{E}[X_i] = \mu$ and variance $Var(X_i) = \sigma^2$. The sample mean is:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Because $\bar{X}$ is a combination of random variables, it is itself a random variable with its own mean and variance.

A. The Expected Value of $\bar{X}$

Where is the sampling distribution centered? We use the linearity of expectation:

$$\mathbb{E}[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n}(n\mu) = \mu.$$

Conclusion: The mean of all sample means is exactly equal to the population mean.

B. The Variance and Standard Error of $\bar{X}$

How spread out are the sample means? Assuming the observations X_i are independent (or the population is infinitely large):

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

The standard deviation of the sampling distribution is called the **Standard Error (SE)**:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Crucial Insight: As the sample size n increases, the standard error decreases by a factor of $\sqrt{n}$. A larger sample provides a much tighter, more precise estimate of the population mean.

Recall: The Central Limit Theorem (CLT)

We know the mean and variance of $\bar{X}$, but what is its exact shape?

Case 1: The Population is Normal If the underlying population is normally distributed, $X \sim N(\mu, \sigma^2)$, then any linear combination of normal variables is also normal. Therefore, for *any* sample size n:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Case 2: The Population is NOT Normal (The CLT) What if the population is heavily skewed, bimodal, or uniform? The Central Limit Theorem is one of the most remarkable results in probability theory:

> **Theorem:** Let $X_1, \ldots, X_n$ be a random sample from *any* distribution with a finite mean μ and finite variance σ^2. As the sample size $n \to \infty$, the sampling distribution of the sample mean $\bar{X}$ converges in distribution to a Normal distribution.

$$\lim_{n\to\infty} P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z\right) = \Phi(z)$$

For practical purposes, if $n \geq 30$, the normal approximation is generally considered sufficiently accurate regardless of the population's shape.

Practical Example: Spatial Data and the CLT

Consider an environmental study utilizing spatial statistics to measure localized PM2.5 (particulate matter) levels across thousands of 1km $\times$ 1km grid cells in Delhi.

- **The Population:** The actual distribution of PM2.5 across all grid cells is highly **right-skewed**. Most residential areas hover around a baseline level, but a few industrial zones or severe traffic bottlenecks have massive, extreme spikes in pollution. Let the population mean be $\mu = 80\,\mu g/m^3$ with a standard deviation $\sigma = 40\,\mu g/m^3$.

- **The Problem:** If we pick *one* random grid cell, its pollution level is highly unpredictable and not normally distributed.
- **The CLT Solution:** If we randomly sample $n = 35$ grid cells and calculate their *average* PM2.5 ($\bar{x}$), the CLT guarantees that the sampling distribution of this average will be approximately normal:

$$\bar{X} \approx N\left(80, \frac{40^2}{35}\right) = N(80, 45.7)$$

Because the sampling distribution is normal, we can now use standard Z-scores to calculate the probability that our sample mean falls within a certain range, completely bypassing the messy, skewed reality of the raw spatial data.

Sampling Distribution of the Sample Proportion ($\hat{p}$)

When dealing with categorical data (e.g., success/failure, defective/non-defective), we are interested in the population proportion, p. We estimate this with the sample proportion, $\hat{p} = \frac{X}{n}$, where X is the number of successes in a sample of size n.

Because X follows a Binomial distribution ($X \sim \text{Bin}(n, p)$), we can find the mean and variance of $\hat{p}$:

$$\mathbb{E}(\hat{p}) = p$$
$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

Normal Approximation: For large n, by the Central Limit Theorem, the sampling distribution of $\hat{p}$ is approximately normal, that is,

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Example 1 (Quality Control)**.** A manufacturer of cricket balls knows from historical data that 10% of their production does not meet professional weight standards ($p = 0.10$). A quality inspector takes a random sample of $n = 100$ balls. What is the probability that more than 15% of the sampled balls are defective?

Solution: First, check the normality conditions: $np = 100(0.10) = 10 \geq 5$ and $n(1-p) = 100(0.90) = 90 \geq 5$. The normal approximation is valid. The mean and standard error of $\hat{p}$ are:

$$\mu_{\hat{p}} = 0.10$$
$$\sigma_{\hat{p}} = \sqrt{\frac{0.10(0.90)}{100}} = \sqrt{0.0009} = 0.03$$

We want to find $P(\hat{p} > 0.15)$. Converting to a standard normal Z-score:

$$Z = \frac{0.15 - 0.10}{0.03} = \frac{0.05}{0.03} \approx 1.67$$

Using a standard normal table, $P(Z > 1.67) = 1 - 0.9525 = 0.0475$. *Conclusion:* There is a 4.75% chance that the inspector will find more than 15% defective balls in this sample.

Practice Problem:

A recent survey indicates that 60% of residents in a city support a new traffic management initiative. If a random sample of $n = 150$ residents is selected, find the probability that the sample proportion of supporters will be between 0.55 and 0.65.

Sample Variance

Sample variance is defined in two ways:

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

and

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

Theorem: Let $X_1, X_2, \ldots, X_n$ be IID random variables from a population with mean μ and variance σ^2. Then $\mathbb{E}(S_{n-1}^2) = \sigma^2$.

Proof: First, let us expand the sum of squared deviations around the sample mean. We can strategically add and subtract the true population mean μ inside the square:

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}\big((X_i - \mu) - (\bar{X} - \mu)\big)^2 \\
&= \sum_{i=1}^{n}\big((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\big) \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X} - \mu)^2
\end{aligned}$$

Notice that the middle term contains the sum of deviations from μ. We can rewrite this using the definition of the sample mean ($\sum X_i = n\bar{X}$):

$$\sum_{i=1}^{n}(X_i - \mu) = \sum_{i=1}^{n} X_i - n\mu = n\bar{X} - n\mu = n(\bar{X} - \mu)$$

Substituting this back into our expansion:

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)[n(\bar{X} - \mu)] + n(\bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2
\end{aligned}$$

Now, we take the expected value of both sides using the linearity of expectation:

$$E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]-nE\left[(\bar{X}-\mu)^2\right]$$

By the fundamental definitions of variance, we know two things:

1. The variance of a single observation: $\mathbb{E}[(X_i-\mu)^2]=\mathrm{Var}(X_i)=\sigma^2$
2. The variance of the sample mean: $\mathbb{E}[(\bar{X}-\mu)^2]=\mathrm{Var}(\bar{X})=\dfrac{\sigma^2}{n}$

Substituting these known variances into our expected value equation gives:

$$\begin{aligned}E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]&=\sum_{i=1}^{n}(\sigma^2)-n\left(\frac{\sigma^2}{n}\right)\\&=n\sigma^2-\sigma^2\\&=(n-1)\sigma^2\end{aligned}$$

Finally, we apply this result to find the expected value of our sample variance estimator S^2,

$$\begin{aligned}\mathbb{E}(S_{n-1}^2)&=E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\\&=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\\&=\frac{1}{n-1}(n-1)\sigma^2\\&=\sigma^2\end{aligned}$$

The Gambler's Ruin Problem (Optional)

Two players, a gambler and a casino, repeatedly play a fair (or biased) coin-toss game. The gambler starts with $\$i$, and the casino with $\$N-i$. Each round, the gambler wins \$1 with probability p and loses \$1 with probability $q=1-p$. The game continues until one party is ruined (i.e., reaches 0 capital). Associated rules are:

1. In each play he wins \$ 1 with probability p and loses \$ 1 with probability $q=1-p$.
2. The gambler starts with \$ i (an integer, $0\le i\le N$).
3. The game stops when the gambler's fortune reaches either \$ 0 (ruin) or \$ N (target).

Winning probability

We define

$$u_i=\mathbb{P}(\text{gambler wins all money } N \text{ starting with } i),$$

where the "winning event" means reaching capital N before 0. Using the law of total probability, we have

$$u_i=p\,u_{i+1}+q\,u_{i-1},\quad \text{for } i=1,2,\dots,N-1,$$

with boundary conditions

$$u_0=0,\qquad u_N=1.$$

Case-I: Fair Game ($p = q = \frac{1}{2}$)

When the game is fair, the difference equation simplifies to

$$u_i = \frac{1}{2}u_{i+1} + \frac{1}{2}u_{i-1}.$$

Which is a condition identifying an arithmetic progression, so the solution is

$$u_i = A + Bi,$$

where $A, B \in \mathbb{R}$ are constants. Using the boundary conditions, we have

$$u_0 = 0 \implies A = 0, \qquad u_N = 1 \implies B = \frac{1}{N}.$$

Hence,

$$\boxed{u_i = \frac{i}{N}}.$$

Interpretation: The gambler's probability of eventual success equals the fraction of total capital they currently hold.

Case-II: Biased game ($p \neq q$)

For the biased case, the recurrence relation

$$u_i = p\,u_{i+1} + q\,u_{i-1} \quad \Rightarrow p\,u_{i+1} - u_i + q\,u_{i-1} = 0.$$

This is a difference equation (discrete analogue of differential equation). Assume solution of the form $u_i = r^i$. Plugging in gives the characteristic equation

$$pr^2 - r + q = 0.$$

Using quadratic formula, we have

$$r = \frac{1 \pm \sqrt{1-4pq}}{2p} = \frac{1 \pm (p-q)}{2p}.$$

Since $q = 1 - p$, $p - q = 2p - 1$. Hence the roots are

$$r_1 = 1, \qquad r_2 = \frac{q}{p}.$$

Thus the general solution is

$$u_i = A + B\left(\frac{q}{p}\right)^i.$$

Apply boundary conditions $u_0 = 0$ and $u_N = 1$, we get

$$\begin{cases} A + B = 0, \\ A + B\left(\dfrac{q}{p}\right)^N = 1. \end{cases}$$

Subtracting, $B\big((q/p)^N - 1\big) = 1$, so

$$B = \frac{1}{(q/p)^N - 1}, \qquad A = -B.$$

Therefore

$$u_i = \frac{1 - (q/p)^i}{1 - (q/p)^N}.$$

Equivalently, writing $\rho := q/p$,

$$\boxed{u_i = \frac{1 - \rho^i}{1 - \rho^N}, \quad \rho = \frac{q}{p},\ p \neq \tfrac{1}{2}.}$$

Remarks

- If $p > q$ (favorable game), then $\rho < 1$ and as $N \to \infty$ we get $u_i \to 1 - \rho^i$, and in particular $\lim_{N\to\infty} u_i = 1$ for fixed i (i.e. eventual success with probability 1 when the target is infinite).

- If $p < q$ (unfavorable), $\rho > 1$ and $u_i \to 0$ as $N \to \infty$ (probability to ever reach arbitrarily large target is 0).

- The fair-case formula i/N is the limit of the biased formula as $p \to \frac{1}{2}$ (L'Hôpital can be used to verify).

Expected Duration of the Fair Game

Let m_i denote the expected number of rounds before ruin or success, starting from \$$i$. For the fair case, the recurrence is

$$m_i = 1 + \tfrac{1}{2}(m_{i-1} + m_{i+1}), \quad m_0 = m_N = 0.$$

Solving, we obtain

$$\boxed{m_i = i(N - i)}.$$

Interpretation: The expected duration is maximum when $i = N/2$, i.e., both players start equally wealthy.

Example 2 (Slightly Biased Game against the Casino). Suppose $p = 0.49$ (and $q = 0.51$), with equal initial wealth $W = 100$ each. Thus, $N = 200$ and the gambler starts with $i = 100$. Here,

$$\rho = \frac{q}{p} = \frac{0.51}{0.49} \approx 1.0408.$$

The probability of gambler's eventual win is

$$u_{100} = \frac{1 - \rho^{100}}{1 - \rho^{200}} \approx 0.0196.$$

So despite starting equally, the gambler has only about a **2% chance of success** due to the tiny disadvantage.

p	Gambler's Initial Wealth (i)	Probability of Winning (u_i)
0.50	100	0.500
0.49	100	0.020
0.48	100	0.0004

Conclusion: Even a 1% disadvantage causes very fast decay in the probability of eventual success as the number of games increases.

Example 3 (Amoeba Movement on a Petri Dish)**.** Consider a tiny amoeba moving along a linear nutrient gradient divided into discrete locations $\{0, 1, 2, \ldots, N\}$. At each time step, the amoeba moves:

$$\text{Right with probability } p, \quad \text{Left with probability } q = 1 - p.$$

Suppose nutrient concentration is slightly higher to the right, giving $p = 0.52$. The boundaries 0 and N correspond to death (no nutrients) and reproduction (plentiful nutrients), respectively.
Using the same analysis as the Gambler's Ruin problem, the probability that the amoeba eventually reaches the nutrient-rich boundary starting from position i is:

$$P_i = \frac{1 - (q/p)^i}{1 - (q/p)^N}.$$

Interpretation: A small drift in movement probability ($p = 0.52$) significantly increases the organism's survival chances. This discrete random walk model helps biologists estimate the effect of chemotactic bias on microbial survival.

References

[1] Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability.* Chapman and Hall/CRC.

[2] Ross, Sheldon M. (2020). *Introduction to probability and statistics for engineers and scientists.* Academic press.

Disclaimer