Central Limit Theorem

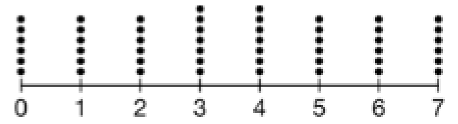
1. **Central Limit Theorem** – 中心极限定理 (Zhōngxīn jíxiàn dìnglǐ)
2. **Population** – 总体 (Zǒngtǐ)
3. **Sample** – 样本 (Yàngběn)
4. **Sampling Distribution** – 抽样分布 (Chōuyàng fēnbù)
5. **Approximate Normality** – 近似正态性 (Jìnsì zhèngtài xìng)

- The CLT describes the behavior of **sample means**, not individual data values.
- For large samples, the distribution of $\bar{x}$ is **approximately normal**, regardless of th population's shape.
- The mean of $\bar{x}$ is the population mean: $\mu_{\bar{x}} = \mu$.
- The standard deviation of $\bar{x}$ is $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.
- Larger samples give **less variability** in sample means.

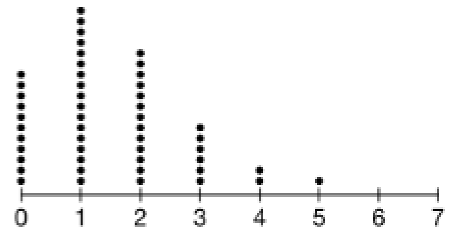
A recent study reported that 1.5 percent of flights are canceled by major air carriers. Consider a simulation with 50 trials designed to estimate the number of canceled flights from a random sample of size 100, where the probability of success, a canceled flight, is 0.015.

Of the following dotplots, which best represents the possible results from the simulation described?

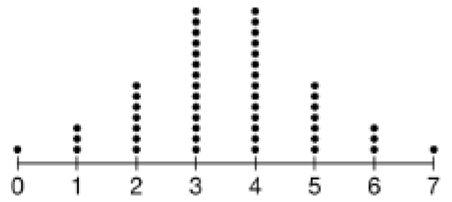
(A)



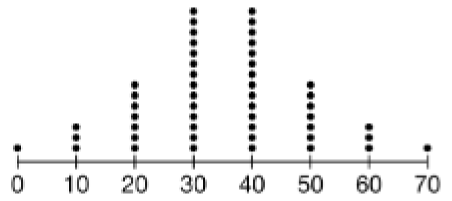
(B)



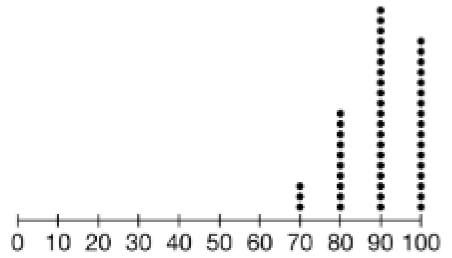
(C)



(D)



(E)



B

A population has a mean of 50 and a standard deviation of 10. A random sample of size 36 is taken.

(a) What is the mean and standard deviation of the sampling distribution of the sample mean?
(b) According to the Central Limit Theorem, what shape will the sampling distribution have?
(c) What is the probability that the sample mean is greater than 53?

(a)

- Mean of sample mean: $\mu_{\bar{x}} = \mu = 50$
- Standard deviation of sample mean: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{36}} = \frac{10}{6} = 1.667$

(b)

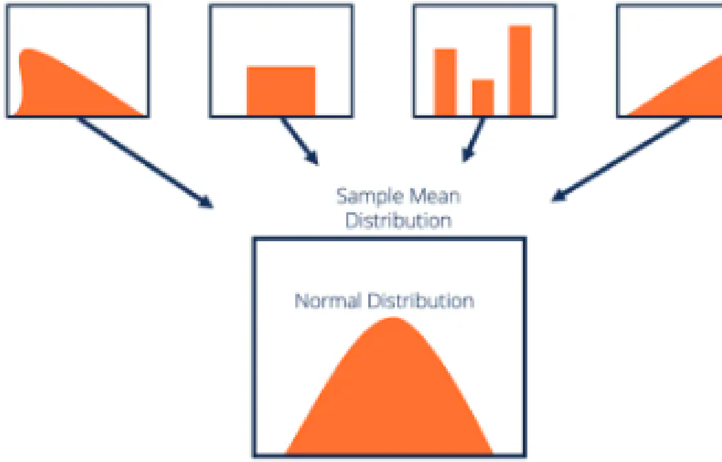
- By the Central Limit Theorem, since $n = 36$ is large, the sampling distribution of the sample mean is approximately **normal**.

(c)

- Compute z-score: $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{53-50}{1.667} \approx 1.80$
- Probability: $P(\bar{x} > 53) = P(z > 1.80) \approx 0.036$

Sampling Distribution

A sampling distribution is the probability distribution of a statistic (such as the sample mean, sample proportion, or sample variance) based on <u>all possible random samples</u> of a fixed size taken from the same population.



- It shows how the statistic would vary if we repeatedly took samples of the same size.
- The variability of a sampling distribution decreases as sample size increases because larger samples reduce sampling error.

Example: If we take many samples of size $n = 50$ from a population with mean $\mu = 100$, the average values from each sample will not be identical, their distribution forms the sampling distribution of the sample mean.

Central Limit Theorem (CLT)

The CLT states that if the sample size is sufficiently large, the sampling distribution of the sample mean $\bar{X}$ will be approximately normal, regardless of the population's distribution.

Parameters of the Sampling Distribution of the Mean:

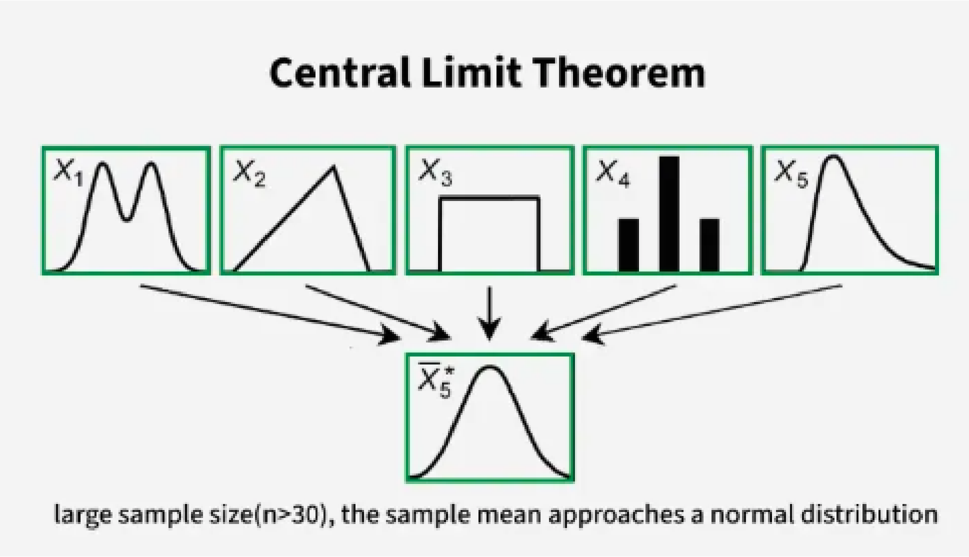
- Mean: $\mu_{\bar{X}} = \mu$ (the mean of the sample means equals the population mean).
- Standard Deviation (Standard Error): $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and n is the sample size.

- As n increases, the sampling distribution becomes more tightly clustered around the population mean.

Example: Even if exam scores in a class are skewed, the distribution of average scores from many random samples of 40 students will look approximately normal.

Conditions for the CLT

- **Independence:** The sampled values must be independent. This is usually satisfied if sampling is random and the population is large relative to the sample size (rule of thumb: population at least 10 times the sample size).
- **Sample Size:** The sample size must be sufficiently large.
  - If the population distribution is normal, any sample size works.
  - If the population is not normal, $n \geq 30$ is usually considered large enough.



Randomization Distribution

A randomization distribution is created by simulating the process of random assignment under the assumption that the null hypothesis is true.

- For experiments, this involves repeatedly and randomly reallocating the observed responses to treatment groups to see what kinds of differences could occur just by chance.
- This forms the basis for significance testing: if the observed statistic is very unlikely compared to this randomization distribution, we conclude that the treatment had an effect.

Example: In a drug experiment, we could shuffle “improvement” responses among treatment and placebo groups thousands of times to build a randomization distribution of differences in means.

Example

A population has mean $\mu = 50$ and standard deviation $\sigma = 12$. Suppose we take random samples of size $n = 36$.

Find the mean and standard deviation of the sampling distribution of the sample mean $\bar{X}$.

Step 1: Recall the formulas for the sampling distribution:

- Mean: $\mu_{\bar{X}} = \mu$
- Standard deviation: $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

Step 2: Substitute values:

- $\mu_{\bar{X}} = 50$
- $\sigma_{\bar{X}} = \dfrac{12}{\sqrt{36}} = \dfrac{12}{6} = 2$

Example

Heights of adult women follow a distribution with mean $\mu = 64$ inches and standard deviation $\sigma = 3$ inches. A random sample of $n = 36$ women is taken.

What is the probability that the sample mean height $\bar{X}$ is greater than 65 inches?

Step 1: Parameters of the sampling distribution:

- $\mu_{\bar{X}} = 64$
- $\sigma_{\bar{X}} = \dfrac{3}{\sqrt{36}} = 0.5$

Step 2: Standardize 65:

$$z = \frac{65–64}{0.5} = \frac{1}{0.5} = 2.$$

Step 3: Find probability:

$$P(\bar{X} > 65) = P(Z > 2) = 0.0228.$$

Example

A study compares average test scores between students who used a new study app and those who did not. The observed difference in means is 4 points (app group higher). To test whether this could happen by chance, researchers repeatedly shuffled the scores between the two groups and calculated new differences in means.

If only 2% of the simulated differences are greater than or equal to 4, what conclusion should be made?

Step 1: This is a randomization distribution under the null hypothesis (no treatment effect).

Step 2: The observed difference of 4 is rare compared to the simulated differences (only 2% as extreme).

Step 3: Since 2% is less than the usual significance level (5%), the result is statistically significant.

Answer: There is strong evidence that the study app improves test scores.

True or False

- As the sample size increases, the standard deviation of the sampling distribution of the sample mean decreases.

True

True or False

- The Central Limit Theorem can be used to approximate probabilities about the sample mean even when the population distribution is skewed, provided the sample size is large.

True

True or False

- A sample size of 15 is always large enough for the CLT to make the sampling distribution approximately normal.

False

Stretch and Challenge

A company knows that the **time to assemble a product** follows a highly **skewed distribution** with a mean of **50 minutes** and a standard deviation of **15 minutes**.

a) If a random sample of **n = 5 products** is selected, describe the shape of the sampling distribution of the **sample mean**.

b) If a random sample of **n = 100 products** is selected, describe the shape of the sampling distribution of the sample mean. Explain why it differs from part (a).

c) Using the **Central Limit Theorem**, calculate the **standard error** of the sample mean for n = 100.

d) A sample mean of **47 minutes** is observed for n = 100. Compute the **z-score** for this sample mean in the sampling distribution and interpret its meaning.

e) Discuss how the CLT allows us to make **probability statements** about the sample mean even though the original distribution is skewed.

Stretch and Challenge

a) $n = 5$ → sampling distribution skewed, like population.

b) $n = 100$ → approximately normal (CLT), shape symmetric.

c) $SE = 15/\sqrt{100} = 1.5$

d) $z = (47-50)/1.5 = -2$ → moderately unusual.

e) CLT: large n → sample mean ≈ normal → can calculate probabilities even if population is skewed.

Sample Proportion

Population	总体
Sample	样本
Population Proportion (p)	总体比例
Sample Proportion ($\hat{p}$)	样本比例
Sampling Distribution	抽样分布
Parameter	参数
Statistic	统计量
Standard Error	标准误
Random Sample	随机样本
Central Limit Theorem (CLT)	中心极限定理

1. **A sampling distribution describes a statistic, not individual data values.**
   - It shows how the sample proportion ($\hat{p}$) varies from sample to sample.
2. **The center of the sampling distribution is the population proportion.**
   - The mean of $\hat{p}$ is equal to p.
3. **Larger sample sizes reduce variability.**
   - As n increases, the standard deviation of $\hat{p}$ decreases.
4. **Random sampling is required.**
   - The sample should be selected randomly to avoid bias.
5. **Independence must be satisfied.**
   - When sampling without replacement, the sample size should be less than 10% of the population.
6. **The Large Counts Condition must be checked.**
   - $np \geq 10$ and $n(1-p) \geq 10$.
7. **The sampling distribution of $\hat{p}$ is approximately normal when conditions are met.**
   - This result comes from the Central Limit Theorem.
8. **Standard error measures sampling variability.**
   - It describes how much sample proportions vary from one sample to another.
9. **Sampling distributions are the foundation of statistical inference.**
   - They allow us to estimate population parameters and test hypotheses.
10. **Unusual sample results can be evaluated using the sampling distribution.**
    - We can determine whether an observed sample proportion is likely or unlikely under a given population proportion.

A school claims that **60% of its students participate in at least one extracurricular activity**. A random sample of **150 students** is selected.

(a) Define the random variable $\hat{p}$ in this context.

(b) Calculate the mean and standard deviation of the sampling distribution of $\hat{p}$.

(c) Determine whether the sampling distribution of $\hat{p}$ can be reasonably modeled by a normal distribution. Justify your answer.

(d) What is the probability that the sample proportion is **less than 0.55**?

(e) The school takes a second random sample of **600 students**. Explain how the sampling distribution of $\hat{p}$ will change compared with the original sample of 150 students.

(a) $\hat{p}$ = proportion of the 150 sampled students who participate in an extracurricular activity.

(b) Mean = **0.60**; SD ≈ **0.040**.

(c) Yes. $np = 90 \geq 10$ and $n(1-p) = 60 \geq 10$.

(d) $P(\hat{p} < 0.55) \approx \mathbf{0.106}$.

(e) The mean remains **0.60**, but the standard deviation decreases; therefore, the sampling distribution becomes **less variable**.