Comparing distributions

1. Two classes took the same test. Class A has a median score of 78 and Class B has a median of 74. What can you conclude?

A. Class A is more consistent than Class B
B. Class A has higher typical performance
C. Class B has a larger range
D. Class A has lower variability

---

2. Which statistic is most useful for comparing the spread of two skewed distributions?

A. Mean
B. Standard deviation
C. Interquartile range (IQR)
D. Median

---

3. Two distributions have similar means, but one has a much larger IQR. What does this indicate?

A. One distribution is more skewed
B. One distribution has more variability
C. One distribution has a higher center
D. One distribution has no outliers

Comparing distributions

1. Two classes took the same test. Class A has a median score of 78 and Class B has a median of 74. What can you conclude?

A. Class A is more consistent than Class B
B. Class A has higher typical performance
C. Class B has a larger range
D. Class A has lower variability

---

2. Which statistic is most useful for comparing the spread of two skewed distributions?

A. Mean
B. Standard deviation
C. Interquartile range (IQR)
D. Median

---

3. Two distributions have similar means, but one has a much larger IQR. What does this indicate?

A. One distribution is more skewed
B. One distribution has more variability
C. One distribution has a higher center
D. One distribution has no outliers

1. **B** — Class A has higher typical performance (higher median).
2. **C** — Interquartile range (IQR) is best for skewed distributions.
3. **B** — One distribution has more variability.

In northwest Pennsylvania, a zoologist recorded the ages, in months, of 55 bears and whether each bear was male or female. The data are shown in the back-to-back stemplot below.

Female		Male
4	15	
	14	
2 0	13	
	12	
	11	
	10	
4	9	
2	8	
1	7	0
1 0	6	6
8 3 2	5	3 5
6 5 4 2 1	4	6
8 8 5 4 3 2	3	2 4 7 8
9 7 6 2 1 0 0	2	0 0 1 2 2 3 4 4 4 5 6 7 7 9
	1	5 7 8

7|0 represents 70 months

Based on the stemplot, which of the following statements is true?

(A) The median age and the range of ages are both greater for female bears than for male bears.

(B) The median age and the range of ages are both less for female bears than for male bears.

(C) The median age is the same for female bears and male bears, and the range of ages is the same for female bears and male bears.

(D) The median age is less for female bears than for male bears, and the range of ages is greater for female bears than for male bears.

(E) The median age is greater for female bears than for male bears, and the range of ages is less for female bears than for male bears.

A

Comparing Distributions

1. Center（中心）
2. Spread（离散程度）
3. Shape（分布形状）
4. Outliers（离群值）

📍 Measures of Center

5. Mean（平均数）
6. Median（中位数）

📏 Measures of Spread

7. Range（极差）
8. Interquartile Range (IQR)（四分位距）
9. Standard Deviation（标准差）
10. Variance（方差）

- Use **graphs and numerical summaries** to make comparisons.
    - Compare **shape** (including skewness and modality).
        - Compare **center** (mean or median) in context.
        - Compare **spread** (IQR or standard deviation).
- Note **outliers or unusual features** and make conclusions **in context**.

Feature	What to Describe	Key Questions to Answer
C — Center	Typical value of the distribution (median or mean)	Which group has a higher/lower center? By how much?
U — Unusual features	Outliers, clusters, gaps, multiple peaks	Are there outliers in one group but not the other? Any unusual patterns?
S — Shape	Symmetric, skewed left/right, uniform, bimodal	Do both groups have similar shapes, or does one skew while the other is symmetric?
S — Spread	Variability (IQR, range, standard deviation)	Is one group more spread out? Which group shows more variability?

Mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Median

Order data, select middle value (average two if even count)

IQR (Interquartile Range)

$$IQR = Q_3 - Q_1$$

Standard Deviation

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Degrees of freedom is n- 1

The sample mean is calculated from all n observations:

Once we know $\bar{x}$, the deviations from the mean must sum to zero:

- Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

- Sample variance (with $n-1$ in denominator):

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

We want to show:

$$\mathbb{E}[S^2] = \sigma^2$$

Sample variance:

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

Step 1: Express deviations from the mean:

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

Step 2: Take expectation:

$$\mathbb{E}\Big[\sum_{i=1}^{n}(X_i - \bar{X})^2\Big] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2$$

Step 3: Divide by $n-1$:

$$\mathbb{E}[S^2] = \frac{1}{n-1} \cdot (n-1)\sigma^2 = \sigma^2$$

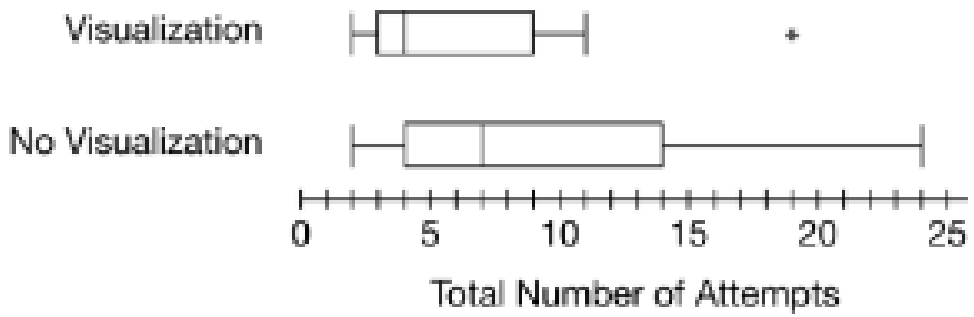
✅ **Conclusion:** S^2 is an unbiased estimator of the population variance.

Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

A team of psychologists studied the concept of visualization in basketball, where players visualize making a basket before shooting the ball. They conducted an experiment in which 20 basketball players with similar abilities were randomly assigned to two groups. The 10 players in group 1 received visualization training, and the 10 players in group 2 did not.

Each player stood 22 feet from the basket at the same location on the basketball court. Each player was then instructed to attempt to make the basket until two consecutive baskets were made. The players who received visualization training were instructed to use visualization techniques before attempting to make the basket. The total number of attempts, including the last two attempts, were recorded for each player.

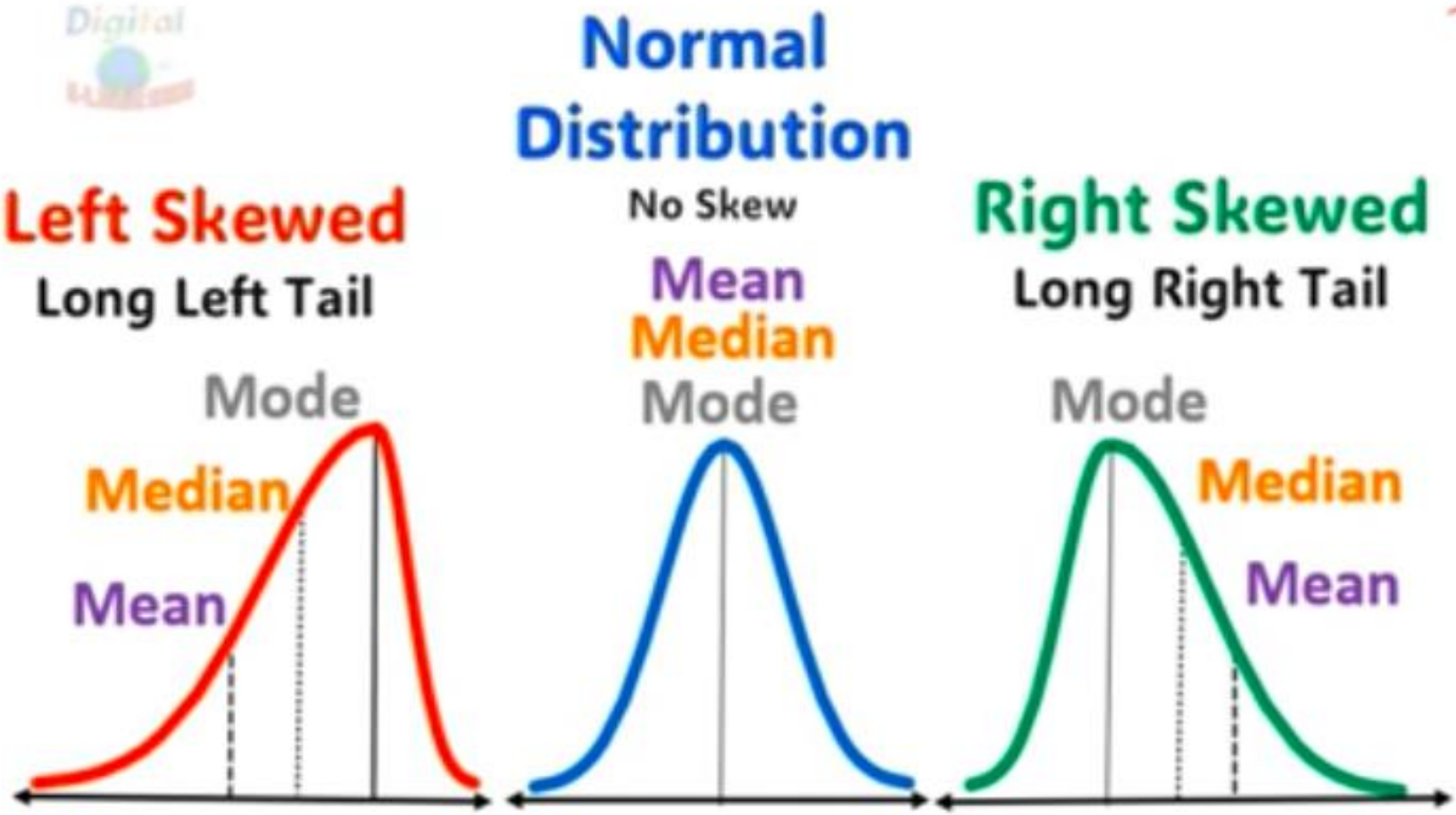
The total number of attempts for each of the 20 players are summarized in the following boxplots.



(a) Based on the boxplots, did basketball players who received visualization training tend to need fewer attempts to make two consecutive baskets from a distance of 22 feet than players who did not receive the training? Explain your reasoning.

Because the median number of attempts for players who received visualization training (4) is less than the median number of attempts for players who did not receive training (7), those who received visualization training tend to need fewer attempts to make two consecutive baskets.

Skewness


Normal Distribution
No Skew
Mean
Median
Mode
Left Skewed
Long Left Tail
Mode
Median
Mean
Right Skewed
Long Right Tail
Mode
Median
Mean

1. A dataset has a mean of 50, a median of 55, and a mode of 60. The skewness of this distribution is:

A) Positive (right-skewed)
B) Negative (left-skewed)
C) Zero (symmetrical)
D) Cannot determine

B

---

2. Which of the following statements about skewness is **true**?

A) A perfectly symmetrical distribution has skewness = 1
B) Positive skewness means the tail is longer on the left
C) Negative skewness means the tail is longer on the left
D) Skewness measures the spread of the data

C

STRETCH and CHALLENGE

8. Comparing Skewness

Two datasets have the same spread and center, but one is symmetric and the other is strongly skewed.

Question:
How could you distinguish between them using graphs? Be specific.

---

9. Real Interpretation

A company reports that salaries have increased because the mean salary rose significantly, but the median salary barely changed.

Question:
Critically evaluate this claim.

---

10. Deep Comparison (AP-style)

Distribution A: symmetric, small spread
Distribution B: right-skewed, large spread

Question:
Write a full comparison using statistical vocabulary (center, spread, shape, outliers).

8. Comparing Skewness

Use:

- **Histogram / dot plot** → shows skew directly
- **Boxplot** → skew shown by unequal whiskers and median position

---

9. Real Interpretation

The increase in mean suggests **high salaries increased**, but since the median barely changed, most workers saw **little to no increase**.
Conclusion: gains are likely concentrated among **top earners**.

---

10. Deep Comparison (AP-style)

Distribution A is symmetric with a smaller spread, indicating values are **closely clustered around the center** with no strong skew.
Distribution B is right-skewed with a larger spread, indicating **greater variability** and possible **high outliers** pulling the distribution to the right.

Comparing Two Means

1. **Mean (均值)** – average of a data set
2. **Difference in Means (均值差)** – $\bar{x}_1 - \bar{x}_2$, used to compare groups
3. **Standard Error (标准误)** – estimated variability of a sample mean
4. **Confidence Interval (置信区间)** – range of plausible values for the mean difference
5. **t-statistic (t统计量)** – value used to test hypotheses about mean differences

A researcher wants to compare the exam scores of students in two different teaching methods: traditional lecture (Group A) and online learning (Group B).

- Group A: $n_1 = 25$, $\bar{x}_1 = 78$, $s_1 = 10$
- Group B: $n_2 = 30$, $\bar{x}_2 = 85$, $s_2 = 12$

Tasks:

1. Construct a **95% confidence interval** for the difference in mean scores ($\mu_B - \mu_A$).
2. Conduct a **hypothesis test** at $\alpha = 0.05$ to determine if the online learning method leads to higher scores.
3. Interpret your results in the context of the study.

Step 1: Standard Error

$$SE = \sqrt{\frac{10^2}{25} + \frac{12^2}{30}} = \sqrt{4 + 4.8} = \sqrt{8.8} \approx 2.97$$

Step 2: 95% Confidence Interval

$$\bar{x}_B - \bar{x}_A = 7$$

$$CI = 7 \pm 2.064 \cdot 2.97 \approx (0.87, 13.13)$$

Interpretation: Online learning increases scores by 0.87–13.13 points.

Step 3: Hypothesis Test

$$H_0 : \mu_B - \mu_A = 0, \quad H_a : \mu_B - \mu_A > 0$$

$$t = \frac{7}{2.97} \approx 2.36, \quad t_{crit} = 1.711$$

Decision: t > t_crit → Reject H_0

p-value ≈ 0.013 < 0.05 → significant

Step 4: Conclusion

There is **significant evidence** that online learning results in higher scores.

TRUE or False

If the 95% confidence interval for $\mu_1 - \mu_2$ contains 0, we **reject** H_0 at the 5% significance level.

FALSE

TRUE or False

The t-statistic increases if the **difference in sample means increases** while the standard error stays the same.

TRUE

Suppose you have two samples:

- Sample 1: $\bar{X}_1$ (mean of group 1)
- Sample 2: $\bar{X}_2$ (mean of group 2)

Then the **difference in sample means** is:

$$\text{Difference} = \bar{X}_1 - \bar{X}_2$$

True or False

Outliers in one sample can **affect the mean more than the median.**

TRUE