Numerical Summaries and Relationships

Lectures 5–6

QBA201 – Quantitative Business Analysis

Shi Qiu – Fall 2026

The mean averages the observations

$$\text{Mean} = \frac{\text{sum of observations}}{\text{number of observations}}.$$

An average divides the total equally across the observed units.

Discuss: If six students make 84 job applications in total, what is the average per student?

Population and sample notation

	Population	Sample
Number of units	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s

The notation identifies which group the quantity describes.

Read the summation as an instruction

For observations $x_1, \ldots, x_n$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}.$$

For all N population observations:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Discuss: What does the subscript i identify?

Extreme values can move the mean

An unusually large income raises the sum and therefore raises the mean income.

The median may barely change if the order of the middle observations stays the same.

Check a distribution's shape and unusual values before describing a typical observation by its mean.

The median uses the ordered middle

Order the nine values:

$$0,\ 0,\ 5,\ 7,\ \boxed{8},\ 9,\ 12,\ 14,\ 22.$$

The fifth value is the median: 8.

Add 33 to obtain ten values. The middle two are 8 and 9:

$$\text{Median} = \frac{8+9}{2} = 8.5.$$

The mode is the most frequent value

A mode identifies the value or category with the largest count.

- One dataset can have several modes.
- Some numerical datasets have no uniquely most frequent value.
- For grouped numerical data, identify the modal class.

The mean responds to a long tail

Large values in a right tail can pull the mean above the median.

Small values in a left tail can pull the mean below the median.

These are common patterns, not universal inequalities for every dataset.

Choose the center to match the question

- Mean: useful when the total and an equal-share average matter.
- Median: useful for a typical ordered value, especially with a long tail.
- Mode: useful for the most common category or repeated value.

Discuss: Which center would you report for a strongly right-skewed income distribution?

One changed observation has a predictable effect

The source reports $n = 200$ and mean 23.31.

Replace one value of 53 by 133. The total rises by $133 - 53 = 80$.

$$\text{Increase in mean} = \frac{80}{200} = 0.40.$$

The revised reported mean is approximately $23.31 + 0.40 = 23.71$.

The same mean can hide different variability

Two illustrative samples:

$$A:\ 48,\ 50,\ 52, \qquad B:\ 10,\ 50,\ 90.$$

Both have mean 50.

Sample B spreads much farther from its mean.

A center alone does not describe how reliable or predictable an individual observation is.

Range uses only the endpoints

$$\text{Range} = \max(x_i) - \min(x_i).$$

Both source datasets have range $50 - 4 = 46$:

$$4, 4, 4, 4, 50$$

$$4, 8, 15, 24, 39, 50.$$

Discuss: What differences between these datasets does the range miss?

Variance measures squared deviations

Population variance: σ^2.

Sample variance: s^2.

Squaring prevents positive and negative deviations from canceling.

Variance has squared measurement units. If x is in minutes, its variance is in minutes squared.

Population and sample variance use different divisors

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}, \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

The sample formula uses $n-1$ because the sample mean has been estimated from the same observations.

Discuss: Which formula applies to six sampled students?

An equivalent computational formula

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right].$$

Keep the two terms distinct

$\sum x_i^2$ means square first, then add.

$\left(\sum x_i\right)^2$ means add first, then square.

For hand calculations, retain precision until the final answer.

Worked example: job applications

A sample records:

$$17,\ 15,\ 23,\ 7,\ 9,\ 13.$$

1. Identify n and calculate $\bar{x}$.
2. Calculate each squared deviation.
3. Divide their sum by $n - 1$.

Discuss: Why is the denominator 5 rather than 6?

The mean and variance, step by step

$$n = 6, \quad \sum x_i = 84, \quad \bar{x} = 84/6 = 14.$$

$$\sum (x_i - \bar{x})^2 = 9 + 1 + 81 + 49 + 25 + 1 = 166.$$

$$s^2 = \frac{166}{5} = 33.2.$$

Using the shortcut, $\sum x_i^2 = 1,342$ and $[1,342 - 84^2/6]/5 = 33.2$.

Standard deviation returns to the original units

$$\sigma = \sqrt{\sigma^2}, \qquad s = \sqrt{s^2}.$$

For the applications sample:

$$s = \sqrt{33.2} = 5.76 \text{ applications.}$$

Compare a standard deviation with quantities measured in the same units, not with a variance.

The empirical rule requires a bell-shaped distribution

For an approximately bell-shaped distribution:

Interval	Approximate share
Mean ± 1 SD	68%
Mean ± 2 SD	95%
Mean ± 3 SD	99.7%

Check the shape before applying the rule.

Read the standard-deviation bands

If $\mu = 70$ and $\sigma = 5$:

$$\mu \pm \sigma: \ 65 \text{ to } 75,$$
$$\mu \pm 2\sigma: \ 60 \text{ to } 80,$$
$$\mu \pm 3\sigma: \ 55 \text{ to } 85.$$

Discuss: Which band contains approximately 95% under a bell-shaped model?

Chebyshev gives a lower bound without a bell shape

For a population distribution with finite, positive variance and $k > 1$:

$$\mathrm{P}(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

At $k = 2$, the guaranteed share is at least $1 - 1/4 = 75\%$.

At $k = 3$, it is at least $1 - 1/9 = 88.89\%$.

Coefficient of variation compares relative spread

For a positive mean on a meaningful ratio scale:

$$CV = \frac{\sigma}{\mu}, \qquad cv = \frac{s}{\bar{x}}.$$

Multiply by 100 to express the result as a percentage.

Avoid CV when the mean is zero, close to zero, or on a scale with an arbitrary zero.

The same SD can mean different relative variability

Both illustrative processes have standard deviation 10.

$$A: \ \mu = 100, \quad CV = 10/100 = 10\%.$$

$$B: \ \mu = 500, \quad CV = 10/500 = 2\%.$$

Process A has more variability relative to its mean.

Discuss: Does that imply its absolute standard deviation is larger?

Percentile describes relative standing

A percentile locates a value within an ordered distribution.

Being near the 60th percentile means roughly 60% of the distribution is at or below that location.

Ties and calculation conventions affect the exact sample statement.

A percentile rank is different from the percentage score on a test.

Quartiles divide an ordered distribution

$$Q_1: \text{ 25th percentile}, \quad Q_2: \text{ 50th percentile}, \quad Q_3: \text{ 75th percentile}.$$

Q_2 is the median.

For course questions using the median-of-halves convention, sort the values and take the median of each half.

A percentile rule must specify interpolation

One textbook rule places percentile P at:

$$L_P = (n+1)\frac{P}{100}.$$

If $L_P = j + d$, interpolate between ordered observations:

$$x_{(j)} + d\Big(x_{(j+1)} - x_{(j)}\Big).$$

j is the whole-number part of L_P; d is its fractional part.

Different sample-percentile algorithms can give different answers. Use the rule stated in the question.

Worked percentile with an explicit rule

Ordered data: $0, 0, 1, 2, 3, 6, 9, 12$. Use $(n+1)P/100$ with linear interpolation.

$$L_{75} = 9(0.75) = 6.75.$$

$$P_{75} = x_{(6)} + 0.75(x_{(7)} - x_{(6)}) = 6 + 0.75(9 - 6) = 8.25.$$

Interquartile range measures the middle spread

$$IQR = Q_3 - Q_1.$$

Using median-of-halves quartiles for $0, 0, 1, 2, 3, 6, 9, 12$:

$$Q_1 = 0.5, \quad Q_3 = 7.5, \quad IQR = 7.$$

The IQR focuses on the middle half and is resistant to a few extreme observations.

Boxplot whiskers end at observations

For the usual $1.5\,IQR$ boxplot:

- The box spans Q_1 to Q_3, with a line at the median.
- Fences are $Q_1 - 1.5IQR$ and $Q_3 + 1.5IQR$.
- Whiskers end at the most extreme observations inside those fences.
- Points outside the fences are flagged for investigation.

A fence is a threshold. It need not be an observed whisker endpoint.

Reading a five-number summary

For the eight-value example, using median-of-halves:

$$\text{min} = 0, \quad Q_1 = 0.5, \quad Q_2 = 2.5, \quad Q_3 = 7.5, \quad \text{max} = 12.$$

The fences are -10 and 18, so neither endpoint is flagged.



Discuss: Which part of a boxplot would reveal the long upper half in this example?

Covariance averages paired deviations

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}.$$

A pair on the same side of both means has a positive deviation product.

Discuss: When is the product negative?

An equivalent covariance formula

$$s_{xy} = \frac{1}{n-1}\left[\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right].$$

- Multiply values from the same observation to form $x_i y_i$.
- Keep the original pairing.
- Divide by $n - 1$ for sample covariance.

The pairing determines the covariance

Use $x = (2, 6, 7)$, with $\bar{x} = 5$.

Paired y values	$\bar{y}$	s_{xy}
$(13, 20, 27)$	20	17.5
$(27, 20, 13)$	20	-17.5
$(20, 27, 13)$	20	-3.5

For the first pairing: $(21 + 0 + 14)/(3 - 1) = 17.5$.

Identical marginal values can have different relationships when the pairings change.

Correlation standardizes covariance

When both standard deviations are positive:

$$r = \frac{s_{xy}}{s_x s_y}, \qquad \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

The units cancel.

Discuss: Why is correlation undefined if every observed x value is identical?

Interpret correlation as linear association

$$-1 \leq r \leq 1.$$

- Near $+1$: strong positive linear association.
- Near -1: strong negative linear association.
- Near 0: weak linear association.

A nonlinear relationship can remain strong even when linear correlation is near zero.

Correlation does not identify a causal effect

A correlation describes how paired values vary.

It can reflect common causes, selection, reverse causality or a direct mechanism.

Discuss: Would a correlation between ice-cream sales and swimming accidents prove that ice cream causes accidents?

Compute the marginal variances

$$s_x^2 = \frac{(2-5)^2 + (6-5)^2 + (7-5)^2}{2} = \frac{14}{2} = 7.$$

$$s_y^2 = \frac{(13-20)^2 + (20-20)^2 + (27-20)^2}{2} = \frac{98}{2} = 49.$$

Keep these unrounded values for the next step.

The correlations, using full precision

Pairing	Covariance	Correlation
First	17.5	$17.5/(7\sqrt{7}) = 0.9449$
Second	-17.5	-0.9449
Third	-3.5	-0.1890

Discuss: How do the signs and magnitudes match the paired observations?