Comparing two distributions

You measured the same thing in two groups. How do you say which runs higher and which is more spread out, and when does a difference between the samples say something about the populations?

WORDS TO KNOW

Distribution: all the values of one variable in a group, and how often each occurs, shown by a dot plot or a histogram.

Center: a typical value: the mean or the median.

Spread: how far the values stretch: the range, or the interquartile range (IQR), the width of the middle half.

Shape: symmetric, skewed right (long tail toward larger values) or skewed left, with one peak or more.

Outlier: a value far from the rest. By the 1.5 IQR rule, a value more than 1.5 × IQR above Q_3 or below Q_1.

Shift: the gap between the two groups' centers.

KEY IDEAS

1. Describe each, then write the comparison sentence

Put both plots on **one** number line, so a position means the same number in both. Describe each group by center, spread and shape, with units.

A comparison sentence names both groups, the measure, both values with units, and the direction and size of the difference.

"In these samples, the median sleep of the 11 to 13 year olds, 9 hours, is 1.5 hours more than that of the 16 to 18 year olds, 7.5 hours."

2. Choose the measures from the shape

Shape	Center	Spread
roughly symmetric, no outliers	mean	range
skewed, or with outliers	median	IQR
two clusters with a gap	describe each cluster	describe each cluster

Use the **same** measure for both groups. If either group calls for the median, compare medians.

The mean uses every value, so one far value can pull it a long way. The median moves only when the middle of the list moves.

3. Shift against spread

Random samples stand in for their populations, but another sample would give other numbers. So set the shift against the spread of each group (IQR with medians, range with means).

Shift big against the spread: the bulk of one group sits above the bulk of the other. With random samples, it is reasonable to say the populations differ.

Shift small against the spread: the plots mostly sit on top of each other. No conclusion: another pair of samples could shrink or reverse the gap.

Free sample from math.wiki for classroom use. Do not post. math.wiki

WATCH OUT

- **"Tends to" is about groups, not every member.** A younger person can still sleep less than an older one.
- **No difference shown is not "the same".** Samples that do not show a difference do not show the populations are equal.
- **A difference does not say why.** Only a designed experiment can show a cause.
- **A claim about populations needs random samples.** A convenience sample speaks only for itself.

WORKED EXAMPLES

Example 1. Random samples of 12 ticket prices, in dollars, from two theaters. Riverside: 15, 15, 18, 18, 18, 20, 20, 22, 22, 25, 25, 64. Hilltop: 10, 12, 12, 14, 14, 15, 15, 15, 16, 18, 18, 21. Which measures should a comparison use? Write the comparison.

1. Shape: the Riverside quartiles are 18 and 23.5, so the IQR is 5.5 and $23.5 + 1.5 \times 5.5 = 31.75$. The 64 is far above that: an outlier. Hilltop is roughly symmetric.
2. The outlier pulls the Riverside mean up to $282 \div 12 = 23.5$ dollars, higher than 9 of the 12 prices. So use medians and IQRs, for both theaters.
3. Medians: 20 dollars and 15 dollars. Hilltop quartiles 13 and 17, so its IQR is 4 dollars.

Answer: medians and IQRs. "In these samples, the median Riverside ticket, 20 dollars, cost 5 dollars more than the median Hilltop ticket, 15 dollars. The middle halves spread over similar widths, 5.5 and 4 dollars, and Riverside had one ticket far above the rest."

Example 2. Random samples of 20 matches from two soccer leagues. Do they support the claim that the Coastal League scores more?

Goals in a match	0	1	2	3	4	5	6
Coastal matches	1	3	5	5	3	2	1
Valley matches	2	4	5	4	3	1	1

1. Both have one peak and no far values, close enough to symmetric to use means and ranges.
2. Coastal: $(0 + 3 + 10 + 15 + 12 + 10 + 6) \div 20 = 56 \div 20 = 2.8$ goals.
3. Valley: $(0 + 4 + 10 + 12 + 12 + 5 + 6) \div 20 = 49 \div 20 = 2.45$ goals.
4. The shift is $2.8 - 2.45 = 0.35$ goals, against a range of 6 goals in each league. The two rows are nearly alike.

Answer: no. The shift is small against the spread, so the samples do not show the leagues differ. Nor do they show the leagues are the same.

Example 3. Two separate random samples of 12 adults from a walking club had their pulse taken, one sample at rest and one just after a brisk walk. At rest: median 72, quartiles 66 and 77 beats per minute. After the walk: median 87, quartiles 81 and 94. Do pulse rates after the walk tend to be higher for the whole club?

1. The shift between the medians is $87 - 72 = 15$ beats per minute.
2. The IQRs are $77 - 66 = 11$ and $94 - 81 = 13$. The shift is larger than both.
3. The middle halves, 66 to 77 and 81 to 94, do not overlap.

Answer: yes. The samples are random and the shift is big against the spread, so it is reasonable to say the club's pulse rates tend to be higher just after a walk. It says nothing about how any one person's pulse changes.

Free sample from math.wiki for classroom use. Do not post. math.wiki