Data displays, normal distributions and association

Original skill preparation; versioned source scope

Data Displays, Normal Distributions and Association

- Read and choose a data display: histogram, box plot, bar chart or scatterplot.
- Find the **median** 中位数, quartiles and **interquartile range** 四分位距, using a stated **quartile convention** 四分位数约定.
- Use a **normal distribution** 正态分布 and its **standard deviation** 标准差.
- Read a **frequency** 频数 table, and use a fitted line and its **residual** 残差 without claiming cause.

Choosing a display

- Histogram: one quantitative variable grouped into intervals.
- Box plot: the five-number summary (minimum, Q_1, median, Q_3, maximum).
- Bar chart: counts in categories. Scatterplot: pairs of quantitative values.
- From a histogram you can find proportions, but not the exact mean, since values within a bin are unknown.

Quartiles and spread

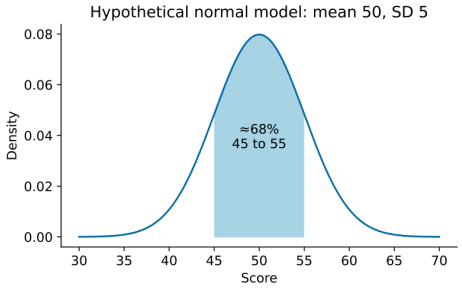
- Order the data. The median is the middle value.
- One common convention: Q_1 is the median of the lower half and Q_3 of the upper half. The ACT states its convention when it matters.
- IQR $= Q_3 - Q_1$; it is not affected by extreme values.

The normal model

- Bell-shaped and symmetric about its mean μ, with spread σ.
- About 68% lies within 1σ of the mean, 95% within 2σ, 99.7% within 3σ.
- These are model proportions; a small real sample may not match them exactly.
- If every value becomes $ax + b$ (with $a > 0$), the mean becomes $a\mu + b$ and the standard deviation $a\sigma$.

exercises.lol exercises.com.cn

Figure: The normal model



- About 68% of a normal model lies within one standard deviation of the mean
- Read the axes and units first.

Association and fitted lines

- Describe a scatterplot by direction, form and strength, and note any outliers.
- A fitted line summarises the trend; it does not join every point.
- **Residual** = observed − predicted.
- Association is not causation: another factor may explain both variables.

Worked example: a normal model

- About 68% lie between $80 - 6 = 74$ and 86; about 95% between 68 and 92.
- So about 2.5% score above 92.
- If each score becomes $2x + 5$: mean $2(80) + 5 = 165$; standard deviation $2(6) = 12$.

Worked example: a five-number summary

- Median $= \dfrac{6+9}{2} = 7.5$.
- Lower half 2, 4, 4, 6: $Q_1 = 4$. Upper half 9, 11, 12, 16: $Q_3 = 11.5$.
- Summary $(2, 4, 7.5, 11.5, 16)$; IQR $= 7.5$. A box plot shows exactly these values.
- A fitted line $y = 4x + 3$ predicts 23 at $x = 5$; if the observed value is 20, the **residual** is -3.

Common errors

- Choosing a bar chart for quantitative data.
- Calculating an "exact" mean from grouped data.
- Adding the constant b to the standard deviation.

Common errors (2)

- Treating 68% as exact for a small sample.
- Reading a positive association as proof of cause.

Key terms

- **median**: The middle value of ordered data.
- **interquartile range**: $Q_3 - Q_1$, the spread of the middle half.
- **quartile convention**: The stated rule for finding quartiles.
- **normal distribution**: A symmetric, bell-shaped model of data.

Check yourself

- In a normal model with mean 100 and standard deviation 15, about what percentage is above 130?
- A histogram has bins [0, 10), [10, 20), [20, 30] with frequencies 2, 6, 2. What proportion is below 20?
- Adding 5 to every value: what happens to the standard deviation?