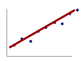
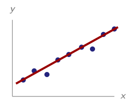
Exploring Two-Variable Data

AP Statistics ▪ Unit 2

AP Statistics



What we will cover



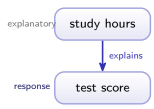
1 Related

Related variables

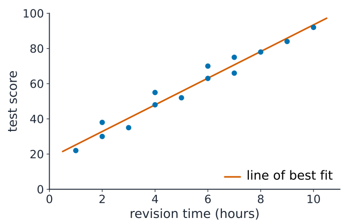
When two variables change together, they show an **association** 关联.

- **explanatory** 解释变量 variable predicts
- **response** 响应变量 variable is the outcome

> An apparent pattern may be real – or a fluke of this one sample.



Scatterplots for Two Quantitative Variables



> A line of best fit runs through the middle of the scattered points

2 Categorical

Two categorical variables

A **two-way table** 双向表 gives three relative frequencies:

- **joint** 联合相对频率: cell ÷ grand total
- **marginal** 边际相对频率: margin ÷ total
- **conditional** 条件相对频率: cell ÷ row/col

> Different **conditional distributions** 条件分布 ⇒ association – but never **causation** 因果关系.

	pet	none
boy	30	20
girl	25	25

P(pet | boy) $= 30/50 = 0.60.$

Reading a two-way table

Marginal distribution	A **marginal distribution** is one variable alone – a row total or a column total, as a percentage of the grand total.
Conditional distribution	A conditional distribution restricts to one row or column first, then takes percentages within it. That is what shows association.
Segmented bar charts	**Segmented bar charts** (or mosaic plots) draw the conditional distributions side by side: if the segments are the same height, there is no association.
Saying it	Compare conditional distributions in context – "68% of seniors, against 41% of juniors" – never raw counts.

> Marginal answers "how many overall"; conditional answers "given that". Almost every mistake in this unit is using the first when the question asked the second.

Comparing Groups with Conditional Distributions

A **conditional distribution 条件分布** is the distribution of one variable *within* a fixed category of the other (found by dividing each cell by its row or column total).

> If the conditional distributions differ across groups, the two variables are **associated**; if they are the same, there is no association.

3 Scatter

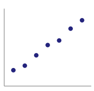
Scatterplots and correlation

Describe a **scatterplot** 散点图 by **direction** 方向, **form** 形态, **strength** 强度, outliers.

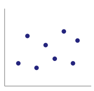
The **correlation** 相关系数 r measures *linear* strength:

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

> r near ± 1 is tight; near 0 is a cloud. r is not resistant, and never implies cause.



$r \approx 0.95$



$r \approx 0.1$

The regression line, term by term

Correlation coefficient	The **correlation coefficient** r measures the strength and direction of a *linear* relationship, between -1 and 1 – and it has no units.
Slope	The slope is the predicted change in y for each one-unit increase in x – state it in context, with the units.
y-intercept	The y-**intercept** is the predicted y when $x = 0$: meaningful only if $x = 0$ is inside the data's range.
Both from r	$b = r\frac{s_y}{s_x}$, and the line always passes through $(\bar{x}, \bar{y})$.

> r describes the scatter; the line predicts from it. A high r still says nothing about causation.

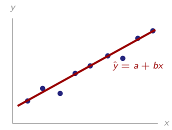
4 Regression

The regression line

The **least-squares regression line** 最小二乘回归线:

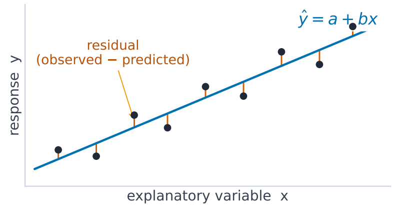
$$\hat{y} = a + bx$$

- **slope** 斜率 b: predicted change in y per unit x
- **intercept** 纵截距 a: predicted y at $x = 0$



> Use it to **predict** 预测, but never **extrapolate** 外推 far beyond the data.

Least-Squares Regression and Its Fit



residual $= y - \hat{y}$ · least squares minimises $\sum(y - \hat{y})^2$

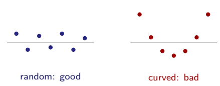
> The least-squares line minimizes the sum of squared residuals

5 Residuals

Residuals and fit

A **residual** 残差 is the miss: $y - \hat{y}$.

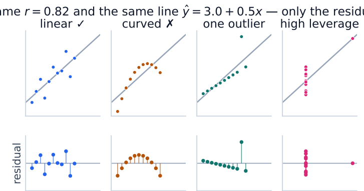
- random **residual plot** 残差图 ⇒ a line fits
- a curve ⇒ line is wrong
- s: the standard deviation of the residuals – the typical prediction error, in y's own units
- r^2: the **coefficient of determination** 决定系数



> r^2 is the fraction of y's variation the line explains – report it in context; $b = r\frac{s_y}{s_x}$.

A caution about r and the line

every panel has the same $r = 0.82$ and the same line $\hat{y} = 3.0 + 0.5x$ — only the residual plot tells them apart



Anscombe: same stats, different plots — always graph data

> A caution about r and the line: all four datasets have the same $r = 0.82$ and the same $\hat{y} = 3.0 + 0.5x$, yet only the first is genuinely linear. The scatterplots barely differ – the

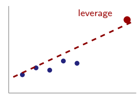
6 Departures

Departures from linearity

An **influential point** 影响点 swings the line when removed:

- a y-outlier sits far vertically
- a **high-leverage** 高杠杆 point sits at extreme x

> If the pattern curves, a **transformation** 变换 (e.g. $\log y$) can straighten it for a line.



Correlation

> Positive correlation rises together; negative correlation moves in opposite directions

Exam tips

- On a scatterplot describe **direction, form, strength, and outliers**; r ranges -1 to 1.
- **Correlation is not causation** – a lurking variable can drive both.
- Interpret the slope of the least-squares line in context ("per one unit of x, predicted y changes by b").
- Check a **residual plot**: no pattern means a line fits; a curve means it does not. Avoid extrapolation.
- r^2 is the fraction of variation in y explained by the model.

7 Recap

Worked example – using a regression line

$\hat{y} = 3 + 2x$ with $r = 0.85$. Predict y at $x = 10$; the actual value is 25. Find the residual and interpret r^2.

- $\hat{y} = 3 + 2(10) = 23$
- Residual $=$ actual $-$ predicted $= 25 - 23 = +2$ (the line under-predicted).
- $r^2 = 0.85^2 = 0.7225 \Rightarrow$ about 72% of the variation in y is explained by x.
- Slope: each 1-unit rise in x predicts a 2-unit rise in y on average.

> Residual $=$ actual $-$ predicted (positive means above the line). Interpret in context with "on average" – and never claim causation from r alone.

Unit 2 – key takeaways

1 Association

explanatory vs response; conditionals show it

2 Correlation

$-1 \leq r \leq 1$; linear only; not resistant; not cause

3 Regression

$\hat{y} = a + bx$; read slope + intercept; don't extrapolate

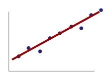
4 Fit

residual plot checks linearity; r^2 explained; leverage

Exploring Two-Variable Data
└Recap

Check yourself

1. What range must the correlation r fall in?
2. In $\hat{y} = a + bx$, what does the slope b mean?
3. What does a curved residual plot tell you?
4. Does a high r prove causation?



Answers **1.** $-1 \leq r \leq 1$ **2.** predicted change in y per unit x **3.** a line is the wrong model **4.** no