

Learn these by heart before the exam. 考试前把这些内容背熟。

Exploring One-Variable Data

$$\bar{x} = \frac{\sum x_i}{n}, \quad s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$
 样本均值与标准差

Describe a distribution with SOCS: Shape, Outliers, Centre, Spread. In context, always.

Skewed right pulls the MEAN above the median. Use median and IQR for skewed data, mean and s for roughly symmetric.

outlier $\iff x < Q_1 - 1.5 \text{ IQR}$ or $x > Q_3 + 1.5 \text{ IQR}$ 离群值判定

$$z = \frac{x - \mu}{\sigma}$$
 how many standard deviations from the mean; unitless 标准分数

Linear transform $y = a + bx$: centre scales by b and shifts by a; spread scales by |b| only.

Exploring Two-Variable Data

$$\hat{y} = a + bx, \quad b = r \frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x}$$
 the line always passes through $(\bar{x}, \bar{y})$ 最小二乘回归线

$$r = \frac{1}{n - 1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$
 $-1 \leq r \leq 1$, unitless, and NOT resistant to outliers 相关系数
 r^2 is the proportion of variation in y explained by the linear model with x. Say it that way.

residual = $y - \hat{y}$ a curved residual plot means a linear model is wrong 残差

Correlation is not causation. Only a randomised EXPERIMENT supports a causal claim.

Influential point: far in the x direction, moves the line a lot. An outlier in y mainly inflates residuals.

Collecting Data

SRS, stratified (homogeneous groups, sample within each), cluster (sample whole groups), systematic. Say which and why.

Bias sources: voluntary response, undercoverage, nonresponse, question wording. Name the specific one.

Experiment needs: comparison, random assignment, replication, control. Random ASSIGNMENT gives causation; random SELECTION gives generalisability.

Blocking is to experiments what stratifying is to samples: group similar units, then randomise WITHIN each block.

Blind means subjects do not know the treatment. Double-blind means the assessors do not either.

Probability and Random Variables

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
 加法法则

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$
 条件概率

Independent $\iff P(A | B) = P(A) \iff P(A \cap B) = P(A)P(B)$. Mutually exclusive is NOT the same as independent.

$$\mu_X = \sum x_i p_i, \quad \sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$$
 离散型随机变量的期望与方差

$$\mu_{X \pm Y} = \mu_X \pm \mu_Y, \quad \sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2$$
 variances ADD even when subtracting, and only if X and Y are independent 随机变量的和与差

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}, \quad \mu = np, \quad \sigma = \sqrt{np(1 - p)}$$
 BINS: Binary, Independent, fixed Number, Same p 二项分布

$$P(X = k) = (1 - p)^{k - 1} p, \quad \mu = \frac{1}{p}$$
 几何分布

Sampling Distributions

$$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
 样本均值的抽样分布

$$\mu_{\hat{p}} = p, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$$
 样本比例的抽样分布

CLT: for $n \geq 30$ the sampling distribution of $\bar{x}$ is approximately normal WHATEVER the population shape.

10% condition: sampling without replacement needs $n \leq 0.10N$ for independence.

Large counts for proportions: $np \geq 10$ AND $n(1 - p) \geq 10$.

Inference - the common structure

estimate $\pm$ (critical value) $\times$ (standard error) 置信区间的通用形式

$$\frac{\text{test statistic} - \text{statistic} - \text{parameter}}{\text{standard error}}$$
 检验统计量的通用形式

Four steps every time: State (hypotheses + α), Plan (name the test + check conditions), Do (statistic + p-value), Conclude (in context).

Interpret a CI: 'We are 95% confident the true [parameter in context] is between A and B.' Never 'the probability is 95%'.

Interpret a p-value: 'the probability of a result at least this extreme IF H_0 were true'.

$p < \alpha$: reject H_0 , evidence for H_a . $p \geq \alpha$: fail to reject. NEVER 'accept H_0 '.

Type I: reject a true H_0 (probability α). Type II: fail to reject a false H_0 . Power = $1 - P(\text{Type II})$; it rises

with larger n, larger α , or a bigger true effect.

Which test to use

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \quad z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$
 the TEST uses p_0 in the standard error; the INTERVAL uses $\hat{p}$ 单比例

$$\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}, \quad t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}, \quad df = n - 1$$
 单均值

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$
 the two-proportion TEST pools: $\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2}$ 两比例之差

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$
 goodness-of-fit $df = k - 1$; independence/homogeneity $df = (r - 1)(c - 1)$; need every $E \geq 5$ 卡方统计量

$$b \pm t^* SE_b, \quad t = \frac{b}{SE_b}, \quad df = n - 2$$
 回归斜率

Paired data (before/after on the SAME unit) is a ONE-sample t on the differences, not a two-sample test.

Exam technique

Always in context. Name the variable and the population in every conclusion; a generic sentence scores E for essentially correct only if contextual.

Check and STATE conditions: random, 10%, and normal/large counts. Listing them without checking earns nothing.

Show the setup: write the formula with numbers substituted, then the answer. Calculator output alone is not enough.