CH1 Look at data - distribution || variables

units/cases — info of characteristic - observation
what
how many

individual - variable (categorical variable
individuals — ~~numeric~~ quantitative variable) - distribution (what value, how often; tools - nature of variables)

graph.

- categorical variable — distribution.
individual → ~~X~~ group
(count / percent) individuals each cate.
— Pie chart. (distri-cate Pie, slice - count/percent)
Bar graph (Bar - cate., height - count/percent)

- Quantitative variable - distribution -
take numeric values (what value, how often.
(Histogram (Bar - class, height - # individuals)
Stemplots - original values

exam (symmetric.
right-skewed
left-skewed) - tail
outlier. - pattern.

Numbers (center - mean (outlier.)
Median.
~~outlier~~:
spread - (range.
variance) unit
Standard deviation)
Quartiles. ranked data - 4 equal parts. | 25% | 25% | 25% | 25% | Q1 Q2 Q3
Q1, Q2, Q3. Index → value.

Inter-quartile range: middle 50%
Q3 - Q1

Measure - center, spread, ~~outlier~~.
median + IQR → skew distribution
outliers.
mean + standard deviation. - (symmetric.
x outlier.
numeric summaries

5 Number summary: min Q1, Q2(median), Q3. max.

plot data. - Box plot. (min max
Q1 Q2 Q3 - IQR.
lower/upper fence.) — (spread
outliers
center.
shape - skew.

CH 2.

Relationships - 2 variables.

cases / units. - info. characteristic.

(what.
how many

↳ variable. { cate.
what. quant.
values
(Explanatory (X)
response (y)

Exam relationships. 2 cate. variables - contingency table
✓ 2 quant. variables - scatterplot.
1 cate. + 1 quant - boxplot side-by side.

Scatterplot.

same individuals -> each individual - 1 point.

which variable - which axis : { explanatory - x - { label.
response - y { scale

Interpret. scatterplot.

(patterns. - { direction. no - nonlinear - linear.
departures. { form ↗ >0 ↘ <0.
strength 0 — 0.3 — 0.5 — 0.7 — 1 (no, weak, moderate, strong)
outlier.

+1 cate. - diff. plot color / symbol

numerical supplement. :

correlation coefficient (r) - ~~calculation~~ linear relationship. ✓ cause-and-effect.

r ← lurking variable. -> y. | large.

least square regression line. ~~fit or.~~

(b₁) $\bar{x}$ $\bar{y}$ $b_0 = \bar{y} - b_1 \cdot \bar{x}$ -> $\hat{y} = b_0 + b_1 \cdot x$.

fit original data points. { extrapolating
interpolating.

Coefficient of determination (r^2)

overall pattern. - { residual $y - \hat{y}$
residual plot.

outlier - influential point.

L4.

Quiz Score: 30 out of 30

Random experiment – variable – outcomes – Event.

Event – Events
- $A \cup B$
- $A \cap B$
- $A'B'$
- $A \cap B = \phi$

Variable outcomes cause probability

1 single variable
2 cate. variable – outcomes

Count
probability

Contingency table – $A \cap B$ – counts – probability

Joint
marginal > contingency table

Additional Rule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$A \cap B = \phi$

$= P(A) + P(B)$

Joint probability $P(A \cap B)$

independent events $P(A \cap B) = P(A) \cdot P(B)$

Multiplication Rule $P(A \cap B) = P(A|B) \cdot P(B)$

$= P(B|A) \cdot P(A)$

Marginal probability $P(A)$

$P(B)$

Conditional probability $P(A|B) = \frac{P(A \cap B)}{P(B)}$

trials n	outcomes 1	2	3	Prob. calc.	Prob

independent
dependent

Conditional probability
Marginal probability) tree diagram (notation. 题干、问题 tree diagram - number p( ) p. P(B) Answer - Step.

Total Probability Rule $P(B) = P(B \cap A) + P(B \cap A')$
Bayes' Theorem
$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Discrete Random Variable

experiment - outcomes - Sample space.
events - variable - take on values : Discrete Random variable (discrete / Counted)

variable - values - probability.
(list - distribution - Probability mass function (pmf) f(x)
- $f(x) \geq 0$.
- $\sum f(x) = 1$
- $f(x) = P(X=x)$

sum - Cumulative Distribution function (CDF) $F(x) = P(X \leq x)$

f(x) F(x)
PMF - CDF - change.
F(x) - P $P(a \leq x \leq b) = F(b) - F(a-)$
X. outcomes (P 2) Probability f(x). F(x). 2.

Summary numbers (mean = center / median $\mu = \sum x \cdot f(x)$
Deviance / standard deviation - dispersion ($\sigma^2 = \sum x^2 f(x) - \mu^2$; $\sigma = \sqrt{\sigma^2}$)

x f(x) x·f(x) x^2 $x^2 \cdot f(x)$
$\sum = 1$ $\sum = \mu$ $\sum - \mu^2 = \sigma^2$.

Binomial

Experiment: each trial - outcomes - success - Prob
(independent
Trails - outcomes - success - Prob
n no. x Distribution.

Success p
n Binomial(n, p)
x
1

$f(x) = P(X=x) = C_x^n p^x (1-p)^{n-x}$

BINOMDIST(number, trials, prob, cumulative)
BINOMDIST(x, n, p, True/false).
X ~ Binomial (n, p)
Mean: $\mu = E(X) = np$
$\sigma = \sqrt{V(X)} = \sqrt{np(1-p)}$

P(X=x).
- at least 2 - $P(X \geq 2)$
- more than 2 - $P(X > 2)$
- at most 2 - $P(X \leq 2)$
- fewer than 2 - $P(X < 2)$
- no more than 2 - $P(X \leq 2)$

$P(3 \leq X < 7)$

Quiz Score: **33** out of 33

Geometric Distribution

A trial - outcomes - success - Prob

{ independent

trials - until 1 success - Prob. p $P(X=x) = (1-p)^{x-1} \cdot p^1$

X

more than 5 $P(X>5) = 1 - P(X \le 5)$

$= P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)$

Continuous Random Variable

variable x - Probability Density function $f(x)$ - Probability distribution - curve

Normal Density — Normal distribution $X \sim N(\mu, \sigma)$

68-95-99.7

Standard Normal distribution - cumulative distribution function.

$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

$\Phi(Z) = P(Z \le x) = F(Z)$

L7 (Binomial)

Binomial → Normal (n)

$np \ge 10$, $nq \ge 10$) $\mu = np$, $\sigma = \sqrt{npq}$

Continuity correction $<$: -1, $\le$: $+0.5$ $>$: $+1$, $\ge$: -0.5 $Z = \frac{x-np}{\sqrt{np(1-p)}}$

Normal → Standard normal $Z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}}$

L9 Sampling distribution of (sample mean / sample proportion)

population - distribution: Normal μ, σ (Suppose)

size n sample

Sample mean $\bar{x}_1\ \bar{x}_2 \cdots \bar{x}_m$

Probability distribution - Sample distribution of Sample mean. $\bar{x}$

(value interval) (how often 'Probability')

$\mu_{\bar{x}} = \mu$, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ > population property

variable: population distribution Normal μ σ

Sample size n

mean $\bar{x}_1 \cdots \bar{x}_n \to \bar{X}$

Sample distribution interval ($\bar{x}$) how often-Prob

(Population → Sample)

$\bar{X} \sim$ Normal $\mu_{\bar{x}} = \mu$, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

$P(Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}})$

Normal - standard Normal

Normal distribution: population - Sample - Sample mean

Sample ($\bar{x}$, n) - $Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.   ✓ ×   ✓ $n \ge 30$ ✓ ✓.

Success

population — success p

Sample size n

success — $\hat{p}_1 \dots \hat{p}_n \rightarrow \hat{p}$

sampling distribution — interval ($\hat{p}$), Prob $P(\hat{p})$

success — class (cate.)

population — population proportion $p = \frac{\# success}{N}$

Sample — Sample proportion $= \frac{\# success}{n} = \hat{p}$

$\frac{\# success}{n}$

Sampling distribution of Sample proportion. $\hat{p} \sim$ Normal $\mu_{\hat{p}} = p$

n > large — Sample proportions $\hat{p} \sim$ Normal distribution. $\mu_{\hat{p}} = p$ $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ $P(\hat{p}) \rightarrow P(z)$

A random sample (past studies, sth.) Sample → population

L9 Confidence Interval

population parameter → μ — Standard Normal $\hat{p} \rightarrow z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ — interval probability

point estimate — 1 sample — observations — sample mean $\bar{x}$

interval estimate — interval — x how close.

(confidence interval) — point estimate $\bar{x}$

Sample standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Confidence C — contain true population mean — σ, n, z, α

point estimate $\bar{x}$

critical value CI — 2 sided — $\alpha/2$ — $Z_{\alpha/2}$

Margin of error $m = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \longrightarrow m = t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$

Margin of error $m = (Z_{\alpha/2})(\sigma_{\bar{x}})$

interval ($\bar{x} - m$, $\bar{x} + m$)

interval $\bar{x} - m$, $\bar{x} + m$

Population parameter → p

✓ point estimate, 1 sample. — sample proportion $\hat{p}$

✓ critical value

✓ Margin of error $m = (Z_{\alpha/2})\sqrt{\frac{p(1-p)}{n}}$ $p \cdot \hat{p} - 0.5$ $m = Z_{\alpha/2}\sigma_{\hat{p}} = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

✓ Confidence interval $\hat{p} - m$, $\hat{p} + m$

L10 Hypothesis test. (population mean μ, population proportion p)

Variable

- Hypotheses

population (μ, σ) p — H_0

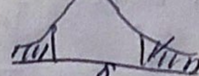
H_1: ___ — 1 or 2 sided → left or right tail

Sample n — $\bar{x}$ X → $\hat{p}$

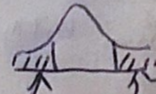
- Critical value Z_α or $Z_{\alpha/2}$
- Test statistics Z_0
- Decision (rejection region, P-value)
- Pvalue

L11

H_0			
T	F	Type I error	α = P(Reject H_0, when H_0 = T)
F	T	Type II error	β = P(Fail to reject H_0, when H_0 = F)



Sample mean — True mean.



L12: 2 sample → population proportion p_1 p_2

验证信息:

Group	Sample	variable	Proportion
1	n_1	x_1	$\hat{p}_1 = \frac{x_1}{n_1} \to p_1$
2	n_2	x_2	$\hat{p}_2 = \frac{x_2}{n_2} \to p_2$

Solution 1

Mean $\mu = np = 5(0.7) = 3.5$

Variance $\sigma^2 = np(1-p) = 5(0.7)(1-0.7) = 1.05$

Standard Deviation $\sigma = 1.02$

CI:

point estimate $\hat{p}_1 - \hat{p}_2$

critical value: $\hat{p}_1$ $\hat{p}_2$ → (Z) CI - 2 sided, $\alpha/2$, $\alpha/2$, $Z_{\alpha/2}$.

margin of error: $m = Z_{\alpha/2} \cdot \sqrt{\sigma(\hat{p}_1 - \hat{p}_2)}$ CI.

$= Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

interval $\hat{p} - m$, $\hat{p} + m$

Hypothesis test

H_0: $p_1 = p_2$

H_1: 1/2 sided left/right tail

✓ critical value, $\hat{p}_1$ $\hat{p}_2$ - (Z)

图. Z

✓ Test statistics.

$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - \mu(\hat{p}_1 - \hat{p}_2)}{\sigma(\hat{p}_1 - \hat{p}_2)} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ $p = \frac{x_1 + x_2}{n_1 + n_2}$

✓ Decision:

plot + conclusion.

✓ P-value.

2 sample - population mean ~~propo~~ μ_1 μ_2

验证信息:

Group	Sample	mean	variance.		
1	n_1	$\bar{x}_1 \to \mu_1$	σ_1^2	s^2	s_1
2	n_2	$\bar{x}_2 \to \mu_2$	σ_2^2	s_2^2	s_2

($\sigma_1^2 \neq \sigma_2^2$) ($\sigma_1^2 = \sigma_2^2$)

CI.

point estimate $\bar{x}_1 - \bar{x}_2$

critical value: σ_1^2 ✓ σ_2^2 ✓ - (Z)

CI - 2 sided - $\alpha/2$, $Z_{\alpha/2}$.

margin of error: $m = Z_{\alpha/2} \cdot \sigma(\bar{x}_1 - \bar{x}_2)$

$= Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

interval $\bar{x} - m$, $\bar{x} + m$

Critical value: s_1 s_2 - t. CI - 2 sided. $\alpha/2$, $\alpha/2$, $t_{\alpha/2, df}$.

$m = t_{\alpha/2} \cdot \sigma(\bar{x}_1 - \bar{x}_2)$

$= t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $df = n - 1$

$\bar{x} - m$, $\bar{x} + m$

Hypo. Test

Hypo.

H_0

H_1: 1/2 sided

~~left/right tail~~

Critical value σ_1^2 σ_2^2 - (Z)

图. Z.

Test statistics:

$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \mu(\bar{x}_1 - \bar{x}_2)}{\sigma(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Critical value s_1^2 s_2^2 - t.

图. $df = n - 1$) $t_{\alpha, df}$

Test statistics

$T_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \mu(\bar{x}_1 - \bar{x}_2)}{\sigma(\bar{x}_1 - \bar{x}_2)} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

critical value s_1^2, s_2^2 - t.

图. $df = n_1 + n_2 - 2$) $t_{\alpha, df}$

Test statistics

$T_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \mu(\bar{x}_1 - \bar{x}_2)}{\sigma(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Critical value s_1, s_2 - t. CI - 2 sided.

$n_1 = n_2$ s_p^2

$m = t \cdot \sigma(\bar{x}_1 - \bar{x}_2)$

$= t_{\alpha/2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $n_1 \neq n_2$ s_p^2.