ST1131

Finals Cheatsheet for 23/24 S2

zaidan s.

Variables

Quantitative discrete, continuous

Categorical nominal, ordinal

Summaries of center

mean - sensitive, median - not

Summaries of variability

range - easy to compute but sensitive
variance/sd - used with mean if bell-shaped
IQR - used with median, if not bell-shaped

Histogram

overall pattern, modality, skew

Boxplot

outliers, skew, spread

Scatterplot

positive, negative, or no association

Data Collection

Lurking variable unobserved (not in dataset) influencing association between variables of primary interest.

Confounding variable included in dataset - associated with response variable, but also with each other.

Observational study no treatment

Experimental study treatment

(+) control for lurking variables
(-) long time needed
requires control comparison group, randomisation, and blinding study

Random sampling simple, stratified, cluster

F2F interview easy to get response, costly

Telephone cheaper, easy to get refused

Self-administered cheapest, easiest to get refused

Bias

Sampling introduced due to sampling design steps - sample not random, sampling frame does not represent population
(undercoverage, non-random sample)

Non-Sampling introduced not due to sampling design steps
Response *(incorrect response, misleading qns)*
Non-Response *(cannot be reached, refusal of participation*

Probability

Mutually exclusive $A \cap B = \emptyset$

Additive law $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Independent $P(A \cap B) = P(A)P(B)$

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

Law of Total Probability $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$

Bayes Theorem $P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$

Sensitivity $Sen = P(+|D)$

Specificity $Spec = P(-|D^C)$

Prevalence $Prevalence = P(D)$

Random variables

Continuous

Mean $\mu = \int x f(x) dx$

Variance $\sigma^2 = \int (x - \mu)^2 f(x) dx$

Discrete

Mean

Linear transformation $Y = bX + A$
$E(Y) = bE(X) + a = b\mu + a$
$E(a_1X_1 + a_2X_2 + ...) = a_1\mu_1 + a_2\mu_2 + ...$

Median

Linear transformation $Y = bX + A$
$Var(Y) = b^2Var(X) = b^2\sigma^2$
$Var(a_1X_1 + a_2X_2 + ...) = a_1\sigma_1^2 + a_2\sigma_2^2 + ...$

Distributions

Binomial

Conditions

n trials with two possible outcomes
same probability of success
n trials are independent

Formula $P(X = x) = C_X^n P^X (1-p)^{(n-x)}$

Notation $X \sim Bin(n, p)$, mean: np, var: $np(1-p)$

Poisson

Formula $P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$

Binomial approx. $X \sim B(n, p) \implies X \sim P(np)$ if n large, p small.

Normal

Properties

bell-shaped, symmetric, mean μ, variance σ^2

Addition of normal variables

If $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$
- $X + a \sim N(a + \mu_X, \sigma_X^2)$
- $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Constant multiplication of normal variables

If $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$
- $aX \sim N(a\mu_{X,} a\sigma_X^2)$
- $aX + bY \sim N(a\mu_X + b\mu_Y, a\sigma_X^2 + b\sigma_Y^2)$

Standardisation of normal variables

If $X \sim N(\mu, \sigma^2)$
- $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$
- if Z-score is < -3 or > 3, it is an outlier

Binomial approx.

$np(1-p) \geq 5$, $X \sim Bin(n, p) \implies X \sim N(np, np(1-p)$

Sampling distribution

Central Limit Theorem With $n >= 30$, sample mean is approximated by a normal distribution.

Sample proportion
$\hat{p} = \frac{X_1 + ... + X_n}{n}$,
if $np(1-p) \geq 5$, distribution of $\hat{p}$:
$N(p, \frac{p(1-p)}{n})$

Sample mean
When the population distribution is **normally** distributed then,
- the histogram of $\bar{X}$ has normal distribution
- variability of bell-shape decreases as n increases
- bell shapes centered at population mean μ
- sampling distribution of $\bar{X}$ depends on μ, σ^2, n

When population distribution is **not normally** distributed then,
- if $n \geq 30$, $\bar{X}$ is approximately normally distributed
- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Confidence Intervals

Proportion

$MoE = q_{\frac{1-\alpha}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$CI = \hat{p} + MoE$

Sample size
If a $(1-\alpha)100\%$ CI with length is $\leq D$, then solving for n gets:
$n \geq (\frac{2 \times q_{1-\alpha/2}}{D})^2 p(1-p)$
If p unknown, use $p = 0.5$, corresponding to the largest possible value of margin of error for a given α.

Mean

$MoE = t_{n-1, \frac{1-\alpha}{2}} \times \frac{s}{\sqrt{n}}$
$CI = \hat{p} + MoE$
Note: if n is large enough, and σ is known,
$MoE = z_{\alpha/2}(\sigma/\sqrt{n})$

Sample size
If a $(1-\alpha)100\%$ CI with length is $\leq D$, then solving for n gets:
$n \geq (\frac{2 \times q_{1-\alpha/2} s}{D})^2$

If $t-$distribution unknown
- replace with $q_{1-\alpha/2}$ from $N(0,1)$ distribution. The impact of this should be reduced by ensuring $n \geq 30$
If s unknown
- estimate using s from a similar study, or conduct a pilot study.

Hypothesis testing

1. Look at assumptions 2. State the hypothesis 3. Find the test statistic, and its null distribution 4. Find the p-value and interpret it 5. Make a conclusion

Errors

Type I: When H_0 rejected but it is true. α
Type II: When H_0 rejected but it is false. β
Power: $1 - \beta$.
(errors cannot be reduced simultaneously)

Confidence interval and significance test
when $\alpha \times 100 = x$
the test is two-sided
both CI and test have the same standard error

Categorical μ **Quantitative** $\hat{p}$

Mean

$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

Proportion

$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Two Independent, Equal Var

- must be approximately normal, quantitative, and independent.
$s_p^2 = \frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2}$
$SE = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$T = \frac{\bar{X} - \bar{Y} - 0}{SE}$

Two Independent, Unequal Var

$s_p^2 = \frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2}$
$SE = \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$
$T = \frac{\bar{X} - \bar{Y} - 0}{SE}$

Dependent

- dependent, approximately normal, quantitative μ = differences of matched subjects, run h-test for mean.

Regression

Assumptions
Randomisation - data collection
Linearity - scatterplot between response Y and regressor X, residuals
Normality - residuals
Constant variance - residuals
t-test Significance of one regressor
F-test Significance of model

Scatterplot
if linearity violated, add higher order terms
if not constant variance, transform response

Residuals
SR against $\hat{Y}$ or X: if funnel, constant variance violated
$\hat{Y}$ against X: if not linear, linearity violated
$QQ-$plot of SR: if not linear, linearity violated

Outlier $SR > 3, SR < -3$
Influential points Cook's distance > 1

R-squared Proportion of total variation of response explained by model

MLR

Use adjusted R^2 instead.
Indicator variable Takes value 1 if observed, 0 otherwise
Interaction term If two variables have interaction (response)