Data Collection

Keywords

Census: measures every member of a population
+ accurate result − expensive / testing may destroy
Sampling units: individuals of a population
Sampling frame: list of sampling units

Random Sampling

i) Simple Random Sampling

? same chance of being selected
use random number / lottery sampling
+ bias free − need sampling frame

ii) Systematic Sampling

? take every k^{th} unit, $k = \frac{pop}{sam}$
pick random number between 1 and k for starting point
+ quick to use − need sampling frame

iii) Stratified Sampling

? sample represents groups (strata) of population
$\frac{sam}{pop}$ x strata for each strata, pick randomly
+ reflects population − pop. must be classified in strata

Non Random Sampling

i) Quota Sampling

? like stratified, but strata filled by interviewer / researcher
+ no sampling frame needed − non-random, potential bias

ii) Opportunity Sampling

? quota filled by those available at the time
+ easy / cheap − unlikely to be representative

Note, there are other advantages and disadvantages

Types of Data qualitative non-numerical
quantitative numerical either discrete or continuous

Large Data Set

UK Stations | When?
Coastal | May – Oct '87 and '15
South | Only 6 months

International Stations
1) Perth, Australia
2) Beijing, China
3) Jacksonville, FL, USA — Oct '87, Oct '15

Data

Rainfall, "tr" means trace, treat as 0.025mm* in calculations
n/a means reading not available, so can't be used in sample
Cloud cover oktas, discrete integers 0 – 8
Max Gust knots, 1 kn = 1.15mph, Great Storm Oct 15th/16th '87

* updated from 0 since recording of video.

Location

KNOW YOUR CALCULATOR!

Mean

$\bar{x} = \frac{\Sigma x}{n}$ $\bar{x} = \frac{\Sigma fx}{\Sigma f}$

Listed Data

Position of $Q_1 = \frac{n}{4}$ (LQ) $Q_2 = \frac{n}{2}$ (Med) $Q_3 = \frac{3n}{4}$ (UQ)
If a decimal, round up ↑
If whole, find midpoint with next one

Grouped Data

Position of $Q_1 = \frac{n}{4}$ $Q_2 = \frac{n}{2}$ $Q_3 = \frac{3n}{4}$
Percentiles eg. 57th percentile, $P_{57} = 0.57 \times n$
Deciles, 10% chunks, $D_3 = P_{30}$, $0.3 \times n$
Do not round, use linear interpolation

Linear Interpolation

eg. weight, nearest kg | true class limits | frequency | cumulative freq.

10 – 12	9.5 ≤ w < 12.5	5	5
13 – 15	12.5 ≤ w < 15.5	8	13
16 – 18	15.5 ≤ w < 18.5	7	20

$Q_2 = 10^{th}$

5th 12.5 | 10^{th} Q_2 | 13th 15.5

$Q_2 = 12.5 + \frac{5}{8} \times 3 = 14.375$ kg

Spread

KNOW YOUR CALCULATOR!

Interquartile Range

F♥ $IQR = Q_3 - Q_1$ + ignores extremes

Interpercentile Range

eg. 10th to 90th $IPR = P_{90} - P_{10}$

Variance, σ^2 / Standard Deviation, σ

$\sigma^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$ MSMSM

$= \frac{\Sigma fx^2}{\Sigma f} - \bar{x}^2$ Note: $\Sigma fx^2 \neq (\Sigma fx)^2$

"mean of squares minus square of means"

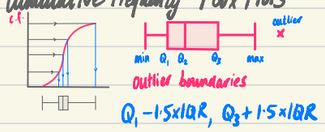
Coding

If $y = ax + b$

$\bar{y} = a\bar{x} + b$ $\sigma_y = a\sigma_x$

Representation

Cumulative Frequency / Box Plots



outlier boundaries
$Q_1 - 1.5 \times IQR$, $Q_3 + 1.5 \times IQR$

Histograms

continuous data
no gaps
frequency density = $\frac{freq}{class\ width}$
area = freq x k

Comparisons

1) location 2) spread

Regression + Correlation

KNOW YOUR CALCULATOR!

Product Moment Correlation Coefficient

PMCC, r $-1 \leq r \leq 1$
Measures strength and +/− of correlation

Regression Line

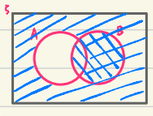
Line of best fit $y = a + bx$
a is y when $x = 0$
b is how much y changes with x
✓ interpolation: estimating inside data range
✗ extrapolation: estimating outside data range

Exponential / Non Linear Models

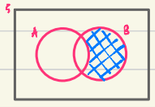
If $y = ab^x$
$\log y = \ln a + x \ln b$ Take logs
If $y = ax^n$
$\log y = \ln a + n \ln x$

Probability

Venn Diagrams

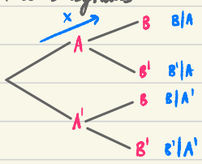


eg. A' ∪ B
Unions shade all of both



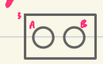
eg. A' ∩ B
Intersections shade overlap

Tree Diagrams



Mutually Exclusive

If A and B are mutually exclusive
$P(A \cap B) = 0$
$P(A \cup B) = P(A) + P(B)$

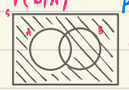


Independence

If A and B are independent
F $P(A \cap B) = P(A) \times P(B)$
F $P(A|B) = P(A)$ Can't tell from Venn

Conditional Probability

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



Reduce the whole space to A

Addition Law

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

Discrete Uniform Distribution

Probabilities of outcomes all equal

eg. X is cloud cover, measured in oktas

x	0	1	2	3	4	5	6	7	8
P(X = x)	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

Binomial Distribution

$X \sim B(n, p)$
F $P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$

When to use?

Fixed number of trials, n
Fixed probability of success, p
Independent trials
Two outcomes, success / failure

Cumulative Probabilities

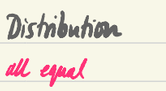
$P(X < 5) = P(X \leq 4)$ (1 2 3 4) KNOW YOUR CALCULATOR!
$P(X > 3) = 1 - P(X \leq 3)$ (4 5 6 7)
$P(6 < X \leq 10) = P(X \leq 10) - P(X \leq 6)$ (7 8 9 10)

Inverse Probabilities

Use tables!

Normal Distribution

Continuous Random Variable, Y
$Y \sim N(\mu, \sigma^2)$
Point of inflection $\mu \pm \sigma$



68% } within 1
95% } within 2
99.7% } within 3
} σ of μ

Finding Probabilities and Inverse Normal

eg. $P(Y > 122)$, $P(Y < a) = 0.2$ KNOW YOUR CALCULATOR!

Standard Normal Distribution

$Z \sim N(0, 1^2)$
Coding, $Z = \frac{Y - \mu}{\sigma}$

Missing μ, σ or both

Use coding, and sim eq. for both

Approximating Binomial as Normal

If n is large
If $p \approx 0.5$
$\mu = np$ $\sigma = \sqrt{np(1-p)}$

Continuity Correction

If approximating binomial as normal
discrete ⟶ continuous
eg. $P(X > 5) = P(Y > 5.5)$
$P(3 < X \leq 11) = P(3.5 < Y < 11.5)$

Note: the 2 underlined values are incorrect in the video. These are the correct values.

Hypothesis Testing

Definitions

Null Hypothesis: H_0, what we assume to be true
Alternative Hypothesis: H_1, what would be true if H_0 is wrong
Significance Level: α, the given threshold of likeliness
One-tailed Test: when H_1: $p > k$ or $p < k$
Two-tailed Test: when H_1: $p \neq k$ ⟶ halve α for each end

Correlation Testing

* updated from r since recording the video. ρ is PMCC for population, r is PMCC for sample.

- H_0: $\rho^* = 0$
- H_1: $\rho > 0$ or $\rho < 0$ or $\rho \neq 0$
- Compare your r with table
  - ↳ if your r is more extreme, reject H_0 REMEMBER CONTEXT!
  - ↳ if your r is less extreme, no evidence to reject H_0

Binomial Testing

Test Statistic: number of successes observed
- H_0: $p = k$ (given in context of question)
- H_1: $p > k$ or $p < k$ or $p \neq k$
- Assume H_0, i.e. $X \sim B(n, k)$
- Find $P(X \geq$ or $\leq$ value in question) and going more "extreme"
  - ↳ if $P < \alpha$, reject H_0 REMEMBER CONTEXT!
  - ↳ if $P > \alpha$, no evidence to reject H_0

Note, you should also be familiar with critical regions
Actual significance level is probability, P

Normal Testing – Sample Mean

Sample Mean, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, n is sample size
- H_0: $\mu = k$ cm or kg or mm etc.
- H_1: $\mu > k$ or $\mu < k$ or $\mu \neq k$
- Assume H_0, i.e. $\bar{X} \sim N(k, \frac{\sigma^2}{n})$
- Find $P(\bar{x} >$ or $<$ mean of sample taken)
  - ↳ if $P < \alpha$, reject H_0 REMEMBER CONTEXT!
  - ↳ if $P > \alpha$, no evidence to reject H_0