

1. Data: one variable

Graph

- Categorical variable
 - Bar graph
 - Pie chart
- Quantitative variable
 - Stemplot (back-to-back, splitting the stem)
 - Histogram (frequency, relative freq.)
 - Boxplot (five number summary)
- Examining distributions by graph
 - Overall pattern
 - center or midpoint
 - shape (Symmetric or skewed ?
Unimodal or multi-modes?)
 - spread
 - Any suspected outlier

Numerical measures

- Mean and standard deviation

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum x_i$$

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)}$$

- Five number summary
Min, Q1, Median, Q3, Max
- Median and quartiles measure center and spread are more resistant than mean and standard deviation do.
- Interquartile range IQR = Q3-Q1
- 1.5 X IQR Criterion for outliers.
- Effect of a linear transformation to mean, standard deviation, Q1, Q3, median.

Data: two variables (relationship)

Graph

Scatterplot

- To draw a scatter plot
- To examine a scatterplot
 - Overall pattern
 - Form
 - Direction (positive, negative)
 - Strength (weak, moderate, strong)
 - Any suspected outliers.

Residual plot

- A scatterplot of the regression residuals against the explanatory variable
- To assess the fit of a regression line
- To find out any suspected outliers or influential observations.
- To find out lurking variables.

Numerical measure

- Correlation: measures the direction and strength of the linear relationship between two quantitative variables.

$$r = \frac{1}{n-1} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{1}{n-1} \frac{\sum x_i y_i - n\bar{x}\bar{y}}{s_x s_y}$$

Properties of correlation

- Least-square regression
 $\hat{y} = a + bx$, a-intercept, b-slope
 $b = r \frac{s_y}{s_x}$, $a = \bar{y} - b\bar{x}$
 - Prediction (extrapolation)
 - Interpret the regression line
- Correlation and regression
 r^2 is the fraction of the variation in the values of y that is explained by the LS regression of y on x.
- Cautions about regression and correlation
Residuals
Outliers and influential observations
Lurking variables
Causation, common response, confounding

2. Producing Data: from population to data

•Experimental design

•Basic concepts:

Explanatory variable, response variable; observational study, experiment; experimental units (subjects); factors, level of a factor, treatment; randomization; placebo, double-blind.

•Principles of experimental design: control, randomization, replication.

•Randomized comparative experiments (completely randomized experiments)

•Block designs : matched pairs design, general block design.

•Sampling design

•Concept: population, sample

•Common sampling methods:

Simple random sample (SRS)

Systematic sampling

Stratified samples

Multistage sample

Lecture_12

3

3. Population (Probability)

•Basic tools: probability models

•Concepts: sample space, event, probability, Venn diagram, union, intersection, complement, disjoint events, independent events.

•Probability rules:

•For any event A: $0 \leq P(A) \leq 1$.

• $P(S) = 1$ if S is the sample space.

•Complement rule: $P(A^c) = 1 - P(A)$.

•Addition rule for disjoint events: $P(A \text{ or } B) = P(A) + P(B)$ if A and B are disjoint.

•Multiplication rule for independent events: $P(A \text{ and } B) = P(A)P(B)$ if A and B are independent.

•General addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ for any events A, B.

•Conditional probability: $P(A|B) = P(A \text{ and } B)/P(B)$ if $P(B) > 0$.

•General multiplication rule: $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$

•Independent: A and B are independent if $P(B|A) = P(B)$

•Tree-diagram to calculate the probability of events.

•Assigning probabilities in a finite sample space: equally likely outcomes

Lecture_12

4

Population (Random variables)

Discrete random variable

Distribution:

Value of X	x_1	x_2	x_3	...	x_k
Probability	p_1	p_2	p_3	...	p_k

• Every probability p_i is a number between 0 and 1.

• $p_1 + p_2 + \dots + p_k = 1$

• Mean: $\mu_X = \sum x_i p_i$

• Variance: $\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$

Binomial distribution B(n, p)

• Definition: 4 settings (n obs. , two outcomes, independence, P(success)=p)

• Calculation: three ways to calculate the probability of Binomial events:

• Binomial formula $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

• Binomial table

$$n! = n \times (n-1) \times (n-2) \times \dots \times 3 \times 2 \times 1 \quad 0! = 1$$

• Normal approximation

$$\mu_X = np \quad \sigma_X^2 = \sqrt{np(1-p)}$$

Lecture_12

5

Population (Random Variables) (continued)

Continuous random variables

• Distribution: density curve

• Normal distribution N(μ , σ)

• Definition

• The 68-95-99.7 rule

• Standard normal distribution N(0, 1) $Z = \frac{X - \mu}{\sigma}$ z-score

• Normal quantile plot

• Calculation: use standard normal table to do two types of calculation.

• To calculate the probabilities.

• To calculate the percentiles.

$$\mu_{a+bX} = a + b \mu_X$$

$$\mu_{X+Y} = \mu_X + \mu_Y$$

$$\sigma_{a+bX}^2 = b^2 \sigma_X^2$$

If X and Y are independent, then $\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2$

Lecture_12

6

4. Inference (from data to population)

- Basic concepts: parameter, statistic, sampling distribution, bias.
- Law of large numbers: $\bar{x} \approx \mu$ When n increases, \bar{x} eventually approaches μ .
- Sampling distribution of a count: $B(n, p)$ or approximate $B(n, p)$
- Normal approximation for counts and proportions

Draw an SRS of size n from a large population having population proportion p of successes. Let X be the count of successes in the sample and $\hat{p} = X/n$ then,

$$X \sim N(np, \sqrt{np(1-p)}) \quad \hat{p} \sim N(p, \sqrt{p(1-p)/n}) \quad \text{Approximately, when } np \geq 10 \text{ and } n(1-p) \geq 10.$$

- Mean and standard deviation of a sample mean

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \sigma / \sqrt{n} \quad \text{Where } \mu \text{ and } \sigma \text{ are population mean and std.dev.}$$

- Sampling distribution of a sample mean

- When the population has a $N(\mu, \sigma)$ distribution

$$\bar{x} \text{ is } N(\mu, \sigma / \sqrt{n})$$

- When the population is not normal distributed (central limit theorem)

$$\bar{x} \text{ is } N(\mu, \sigma / \sqrt{n}) \quad \text{Approximately, When } n \text{ is very large.}$$