The Cartoon Guide to Statistics cover

Book Review

> Cartoon is not just illustration but genuine humorous cartoons on statistical concepts
>
> Humour on statistics demands True understanding of it
>
> Covers all most content of a typical 5-10 weeks introductary statistics courses

Video Notes Structure

> first video to give a brief but intuitive overview of the entire course
>
> - understand each word, each symbol and formula
> - but their interrelationship is distance
> - a brief but full story to connect all concepts is attractive.
> - part 1, part 2, part 3, part 4, part 5, part end
>
> much short videos to explain each chapter and sections inside

Table of Content

chapter 4-5 using probability models(r.v.s) to describe world

chapter 6 Be smart & lazy study the population with samples

chapter 7-beyond how to make statistical inferences in real world applicaitons

2 Data Description (give messy data an order)

three statistical tasks (collect, describe, analysis)

Why describing data? (useful, shape, pattern)

Example on describing weight data (dot, frequency, histogram, stem-leaf, etc)

Art of representing data (science, art, politics)

Summary Statistics (shape & pattern)

Two basic stats (central tendency and spread)

Data notations

Mean (sum normalized or weighted)

Median (value at middle position)

Why both measure of central tendency? (sensitivity to outliers)

Measure of spread (how far away from mean/median)

Interquartile range (spread from median)

Standard Deviation (variance, spread from the mean)

Z-score or Standardized scores (create a unit based the dataset mean and sd, simplify each data point)

A empirical rule (1 unit cover 68%, 2units 95%, 3 units 99.7%)

3 Probability (quantify uncertainty with different data scenarios)

probability axioms (basic rules on what and how likelihood is assigned to)

outcomes and events (what likelihood is assigned to)

Probability on union, complementary (how likelihood operate or act like number)

Conditional Probability (how likelihood act given some events true)

Independence (simplify the world by assuming A has no memory or gives no info on B, rarely happen the world)

Bayes' theorem (use conditional probability + total theorem to update belief of A with new evidence B)

4 Random Variable (specify sth from the world in order to quantity its uncertainty)

Random Variables (weights of students as a r.v., assign likelihood to different weight values)

Probability distribution (probability for each value of r.v. together)

Mean and Variance (r.v. valeus with probabilities can form weighted sum, i.e. mean, and measure how far away from the mean, variance of r.v.)

Continuous Random Variable (r.v. has values, some countable, some infinite small uncountable)

Cont. r.v. Probability (each value has only prob density, range of values has probability as area)

Linear Function Expectation and Variance formula (How to calc mean and variance of Linear function of r.v.s?)

5 A tale of two distributions (Intro Bernoulli, Binomial, Normal)

Bernoulli r.v. (hep to study the type of data with structure that can be described by only two scenarios, i.e., 0, 1 success out of two trials, e.g., stock price)

Binomial r.v. (help to study the type of data with structure that can be described by 0, 1, 2, ... n success out of n trials, e.g., stock price)

binomial formula (how to calc probability of different binomial r.v. value, its mean and variance)

Normal distribution (all data with different structures have normal bell curve deep inside, standard normal structure can simplify probability calculation)

calculation the probability of normal r.v. with applet

PDF relation to CDF and formula (PDF accumulate itself to become CDF)

6 Sampling (being smart and lazy to study population with samples)

The problem of the world (can't get all of them)

Solutions: use samples to replace population

sampling design (dataset is smaller but good quality)

Simple random sample (sample elements are equally likely, unbiased, independence)

Theoretical possible but practically expensive

Stratified is more practical (group first, randomize sampe within each group)

Cluster sampling (group first, and randomize on group level)

Systematic sampling (random by selecting every k-th element)

Warnings (no random, no reliable analysis)

How to approximate sample distribution's Expectation and variance on proportion? (with population stats)

use sample mean, variance to replace population mean and variance (but how reliable?)

The **real question** (How well sample stats distributed around population stats)

think samples as random variables (to use probability distributions)

how to calc sample mean and sd (formula)?

How to approximate sampling distribution's Expectation and variance on mean? (with population stats)

sampling distribution (a collection of means of many samples)

sampling expectation (this mean approx population mean)

sampling standard deviation (this sd is population sd divided by ¯)

Sampling distribution is normal (assume sampling distribution is normal)

Central Limit Theorem (use standard normal dist with many samples' mean to approx population distribution)

definition: sample size is large, normal, mean and variance

given sample size is big enough, its sampling distribution is always normal, regardless population distribution

The t-distribution (CLT operate when sample size small and unknown)

two problems of Central Limit Theorem (sample size small, unknown)

Solution: use sample sd to replace population

Recap

7 Confidence Intervals (how confident that population mean is in the interval)

95% confidence

confidence interval (usually set to cover 80%, 95%, 99% of dataset, dive into for details)

How to make CI wider or narrower? (sample size smaller wider or bigger narrower)

empirical calculation of CI examples (dive in for examples)

How large sample size is needed if want to be 95% confident and true proportion only oscilate 1%? (see the formula inside through algebra)

Confidence Intervals for (only need to replace with for the formula above, standard error formula is different, dive in)

- Student's T and degrees of freedom (if sample size is too small, use t table to replace z table above, to calc CI)
- Example: Car crash repair cost (dive in for details)

Recap

8 Hypothesis Testing (CI and CLT help know population mean from sampling distribution mean and SE; individual sample can testify population distribution)

- Ask differently (Could these observations really have occured by chance? Or they really mean something?)

Intriguing problem on black Jurors

- Formulating the problem (select 80 people into a jury, binomial r.v.)
- Solution Attempted (probability of being chance is so rare that we should not observe it)

Four Formal Steps for Statistical Hypothesis Testing

- Step 1 Formulate all hypotheses ( )
- Step 2 The Test Statistics (define the r.v. distribution)
- Step 3 P-value (calcuate the probability of the observations to occur in null hypothesis distribution)
- Step 4 Compare the p-value to a fixed significance level (holdover artifact 0.05, 0.01)

Large Sample significance test for proportions (later for large sample mean)

- The 3 steps solution with example for details

Large Sample test for the population mean

- the problem
- solution

Small Sample Test for the population mean

- The problem of Car crash insurance
- Solution (example of acceptance sampling)

Decision Theory (apply hypothese testing to decision making, to avoid or reduce type I or II error)

- intriguing problem (no fire as null, alarm makes type 1; fire as alternative, and no alarm makes type 2 errors)
- Solution to reduce error (trade off)
- Type I II Errors with (nothing, something)
- Confidence test vs Type I II Error

Example with Type II Error Confidence Test

- Environmental Example
- Formalize the problem (likelihood of not detecting correctly and monitor effectiveness)
- Intuition on confidence test on Type II error
- what does the different monitor programs look like

9 Comparing two populations (apply CI and HT to difference of two populations as a new r.v.)

- The real problems to deal with (Aspirin vs placebo, two pesticides, women vs men payment)

The Asprin Problem (two population or sample experiments with their own proportions of success)

The Model construction (create their difference as new r.v. and use hypothesis testing again)

- Confidence intervals for
- Hypothesis testing (null = no difference, alternative = real difference, p-value with cont r.v. range to testify)
- Assumptions for Aspirin example to work (random assigned, blind experiment, large sample size)

- Comparing the means of two populations
  - Men and Women Salary Example (Estimator)
  - Solution (— — , rest is the same)
- Comparing Small sample means
  - Car crash comparison problem (a sample with 5 vs a sample with 7)
  - t-distribution needs 2 conditions (mound shape and shared sd)
  - standard error (use t-table with their shared standard deviation , and the rest is the same)
- Paired Comparison test is better (specific paired difference is necessary, due to natural variability)
  - Gas per mile Example without paired comparison
  - Gas per mile with paired comparison (figure for details)

10 Experimental design (decisive)

- Components of Experiment design (units, treatment, assignment, control variability with randomization)
  - Replication
  - Local control
  - Randomization
  - A complete Randomized block
  - Conclusion

11 Regression (compare other linear relation or predictability between two populations)

- Regression analysis (can we predict student weight with height?)
  - present the data on graph
  - Find the best fitting line (line indicate the relationship)
  - Why called Regression? (tall short fathers and sons relation)
  - Calculation to find the best line
- ANOVA (analysis of variance, SSE vs total variance, SSR vs total variance, good vs bad fit)
  - 3 types of variability (SSR, SSE, total variance)
  - good fit vs bad fit, see the figure**
  - The squared correlation (= 1- SSE/total, goodness of fit)
  - Correlation Coefficient (measure goodness of fit + direction)
- Statistical Inference
  - Regression model (find CI for a, b, epsilon distributions)
  - To simplify the by making it standard normal distribution for all x
  - How to find the best fitting line
  - Calculate standard devation of each sample
- Confidence Intervals
  - What is the mean weight of students at height 76 inches?
  - when a new student with a new height comes in, how to predict her weight without measure it?
  - (what is the 95 confidence interval of the mean weight for student at new height?)
  - the intervals are terrible for two reasons (Height alone is not enough, 9 data points are too few)
- Hypothesis Testing (testify whether null hypo = no relation = is true or not)
- Multiple linear regression (use matrix algebra and computer)
  - Non-linear regression (use liner regression polynomial technique to deal with non-linear problem)
- Regression diagnostics (use epsilon vs to check assumption sanity)

1 What is statistics

Why Statistics

7、01为什么要学统计和概率

Life is full of uncertainties

Comfortable with uncertainty (consciously or unconsciously)

> we make decisions based on incomplete information all the time

statistics is to quantify uncertainty (consciously)

statistics is no trivial (life and death)

> Challenger Explosion vs low temperature performance analysis
>
> Salk Polio Vaccine vs strict control trials

How to do Statistics

8、02 如何做统计（三项任务）

Data analysis (display, describe data)

> gathering, displaying and summarize of data

Probability (assign likelihood/uncertainty to different data scenarios)

> assign likelihood to different scenarios of an event

Statistical inference (induce some patterns or relations to help with decision making)

> drawing conclusions with specific data and probability

Example

> how likely to get a taxi in the rainy day

Book outline

chapter 2 college students weight dataset (to describe)

chapter 3 probability in gambling den (to quantify uncertainty)

chapter 4-5 using probability models(r.v.s) to describe world

chapter 6 Be smart & lazy study the population with samples

chapter 7-beyond how to make statistical inferences in real world applicaitons

mistrust of statistics

2 Data Description (give messy data an order)

9、03 如何描述数据

> without organization, data can be messy to describe

three statistical tasks (collect, describe, analysis)

> data is raw material to statistics
>
> statistical problems
>
> - data collection
> - data description
> - data analysis

Why describing data? (useful, shape, pattern)

> to present data in **useful** way
>
> summarize data's basic **shape**
>
> useful for discovering **pattern**

Example on describing weight data (dot, frequency, histogram, stem-leaf, etc)

> collect 92 student weight data
>
> represent data with dot plot
>
> represent data with frequency table
>
> how to choose intervals as those in frequency table
>
> represent data with histogram
>
> represent data with relative frequency histogram
>
> represent data with stem-leaf diagram
>
> how to create stem-leaf diagram

Art of representing data (science, art, politics)

> part science, part art
>
> part politics
>
> statistics led to hospital improvement
>
> - Nurse Florence Nightingale compiled mortality statistics from british military hospital

Summary Statistics (shape & pattern)

10、04 如何描述数据集的中心和偏离度

Two basic stats (central tendency and spread)

> central tendency
>
> spread

Data notations

> $x_1, x_2, \ldots, x_n$ for individual observations

Mean (sum normalized or weighted)

> data distribution is like a long bread, mean is the central point where balances the weights of both side
>
> $\overline{x} = \frac{sum of data}{n} = \frac{x_1+x_2+x_3...+x_n}{n} = \sum_i^n x_i/n$

Median (value at middle position)

> ascending the data, pick the middle position value
>
> $3, 5, 7, 7, 38$ median = 7
>
> $3, 5, 7, 7,$ median = $\frac{5+7}{2} = 6$

Why both measure of central tendency? (sensitivity to outliers)

> median is robust to outliers
>
> - $3, 5, 7, 7, 200$ median=7
>
> mean is strong affected by outliers
>
> - $3, 5, 7, 7, 200$ $\overline{x} = 45.8$

Measure of spread (how far away from mean/median)

> how much observations are from the center
>
> intuitive special cases
>
> - if all observations are identical (same), then $spread = 0$, see figure
> - if some students are from football team, student weights distribution ***spread*** will be **wide**, see figure

Interquartile range (spread from median)

> **spread in terms of median**
>
> - find three median out of entire data distribution, see figure
> - IQR = 3rd Median - 1st Median = middle positioned chunk
>
> How to interpret it visually?

- central median (vertical line)
- IQR (box)
- outliers (dots) > 1.5 IQR
- whiskers (horizontal line) : data located outside box within 1.5IQR

Standard Deviation (variance, spread from the mean)

spread in terms of mean, see figure

variance = average squared distance = $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$

sample variance = $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

see example in practice

force variance to have same unit as each observation

- standard deviation = $\sqrt{\sigma^2}$, $\sqrt{s^2}$

Z-score or Standardized scores (create a unit based the dataset mean and sd, simplify each data point)

normalize observation x_i based on mean and sd into a number with a unit score

- $z_i = \frac{x_i - \overline{x}}{s}$
- observation x_i away from $\overline{x}$ measured by s
- see figure

A empirical rule (1 unit cover 68%, 2units 95%, 3 units 99.7%)

1 unit z-score cover 68% of dataset

2 units z-score cover 95% of dataset

3 Probability (quantify uncertainty with different data scenarios)

11、05 什么是概率和条件概率

> probability was populated from gambling
>
> a gambler's question see figure
>
> counter intuition outcome see figure
>
> solution
>
> - P(at least 1 six in 4 throws) = 1 - p(no six in 4 throws) = $1-(5/6)^4$ = 0.518
> - P(at least 1 double-six in 24 throws) = 1- p(no double-six in 24 throws) = $1-(35/36)^{24}$ = 0.491

probability axioms (basic rules on what and how likelihood is assigned to)

> $P(x) \geq 0$ "worse than impossible is not possible"
>
> $\sum P(x) = P(Universe) = 1$ something must occur in the universe

outcomes and events (what likelihood is assigned to)

> Universe contains all outcomes (distinct and exhaustive)
>
> one or more outcomes form an event
>
> see figure

Probability on union, complementary (how likelihood operate or act like number)

> P(E or F) = P(E) + P(F) - P(E and F)
>
> P(E or F) = P(E) + P(F), F and E are disjoint
>
> P(E) = 1- P(not E)

Conditional Probability (how likelihood act given some events true)

> example figure
>
> P(two dies added to 3) = 2/36, there are two outcomes satistify the event (sum to 3)

> P(two dies added to 3 | 1st throw = 1) = # outcomes (2nd die = 2) / # outcomes (when 1st die = 1) = 1/6
>
> P(E | F) = P(E and F) / P(F)

Independence (simplify the world by assuming A has no memory or gives no info on B, rarely happen the world)

12、06 什么是independence

> A is independent from B
>
> - event A tell me nothing of event B
> - P(B) = P(B | A)
> - P(A and B) = P(A)P(B|A) = P(A)P(B)
> - vice verse to prove independence

Bayes' theorem (use conditional probability + total theorem to update belief of A with new evidence B)

13、07 如何理解Bayes theorem

> $p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(A \cap B)}{p(A \cap B) + p(NotA \cap B)} = \frac{p(A)p(B|A)}{p(A)p(B|A) + p(NotA)p(B|NotA)}$
>
> detail explanation see Bayes' theorem visual guide

4 Random Variable (specify sth from the world in order to quantity its uncertainty)

14、08 什么是random variable 和 probability distribution

Random Variables (weights of students as a r.v., assign likelihood to different weight values)

> when we turn students into statistics (features, numbers), we create **random variables** X
>
> - heights X , x_1, x_2, x_3
> - SAT scores Y , y_1, y_2, y_3

- subjects Z , z_1, z_2, z_3
- rolling a pair of dies A, $a \in 2, 3, 4, \ldots, 12$

Probability distribution (probability for each value of r.v. together)

- $P(x_1), P(x_3), P(x_3)$ form X probability dist
- relative frequency of an event in long run $\approx$ prob distribution, see figure

Mean and Variance (r.v. valeus with probabilities can form weighted sum, i.e. mean, and measure how far away from the mean, variance of r.v.)

15、09 如何理解和计算random variable的mean和variance

Data (**sample**) properties $\overline{x}, s$ vs prob distribution (**model, population**) properties μ, σ

sample mean: weighted sum = mean (probability is weight here)

- $\overline{x} = \sum_i^n x_i \frac{n_{x_i}}{n}$

population mean:

- $\mu = \sum xp(x)$ see comparison figure

sample variance:

- $s^2 = \frac{1}{n-1} \sum_i^n (x_i - \overline{x})^2 = \sum (x - \overline{x})^2 \frac{n_x}{n-1}$ (expectation of $(x - \overline{x})^2$)

population variance:

- $\sigma^2 = \sum (x - \mu)^2 p(x)$

Continuous Random Variable (r.v. has values, some countable, some infinite small uncountable)

16、10 如何理解连续变量和它的概率

if X can take on infinite range of values between 0 and 1, then

- $p(X = x) = 0$
- but $p(0 \leq x \leq 0.25) = 0.25$
- see figure

Cont. r.v. Probability (each value has only prob density, range of values has probability as area)

> **probability at particular value is as density**, not probability itself
>
> - $f(x) \geq 0$, can even > 1
>
> **Probability at a range is as area**, real probability here
>
> - $p(x) = \int_a^b f(x)dx$ = area under density with $x \in [a, b]$
> - area is calculated with integral
> - $\int_{-\infty}^{+\infty} f(x)dx = 1$
> - see figure
>
> **mean and variance**
>
> - $\mu = \int_{-\infty}^{+\infty} xf(x)dx$, continuous r.v.
> - $\mu = \sum xp(x)$, discrete r.v.
> - $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$, continuous r.v.
> - $\sigma^2 = \sum (x-\mu)^2 p(x)$, discrete r.v.

Linear Function Expectation and Variance formula (How to calc mean and variance of Linear function of r.v.s?)

17、11 如何计算连续变量以及线性处理后的变量的均值和方差

> $E[X] = \mu$
>
> $E[aX + b] = a\mu + b$
>
> $\sigma^2(aX + b) = a^2\sigma^2(X)$
>
> $E(X_1 + X_2) = E[X_1] + E[X_2]$
>
> $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$, given X and Y are independent
>
> $E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i]$
>
> $\sigma^2(\sum_{i=1}^n X_i) = \sum_{i=1}^n \sigma^2(X_i)$, given X_i are independent

5 A tale of two distributions (Intro Bernoulli, Binomial, Normal)

Bernoulli r.v. (hep to study the type of data with structure that can be described by only two scenarios, i.e., 0, 1 success out of two trials, e.g., stock price)

18、12 什么是Bernoulli random variable

> **properties**
>
> - only two outcomes: success or failure
> - p(success), 1-p(success) = p(failure)
> - one trial is independent from another trial

Binomial r.v. (help to study the type of data with structure that can be described by 0, 1, 2, ... n success out of n trials, e.g., stock price)

19、13 什么是Binomial random variable

> **properties**
>
> - binomial experiment = multiple bernoulli trials together = it can be 2, 3, 4, ... trials together
> - X = number of heads (successes) = [0, 1,2,3,4, ...]
> - n = 2 trials, binomial r.v. X as number of heads
> -

k=# heads	0	1	2
pr(X=k)	0.25	0.5	0.25

binomial formula (how to calc probability of different binomial r.v. value, its mean and variance)

> - $p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
> - $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ all combinations of trials on k heads
> - $\mu = np$
> - $\sigma^2 = np(1-p)$

- what a four throws of a dice look like see figure

Normal distribution (all data with different structures have normal bell curve deep inside, standard normal structure can simplify probability calculation)

20、14 什么是normal random variable

normal = bell curve = mother of all distribution

- binomial included
- Bin(n=2 or more, p = 0.5) $\approx$ normal
- Bin(n=larege, 20 or more, p $\neq 0.5$) $\approx$ normal
- try out this applet for binomial r.v.

standard normal r.v. formula

- $\mu = 0$
- $\sigma = 1$
- $e = 2.718$ (approximation)
- $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ see figure

general normal r.v. formula

- $f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, $z = \frac{x-\mu}{\sigma}$

calculation the probability of normal r.v. with applet

use this applet** to try out different normal r.v., different values, and its probability

PDF relation to CDF and formula (PDF accumulate itself to become CDF)

PDF and CDF relation, see figure1, figure2

$Pr(a \leq x \leq b) = F(\frac{b-\mu}{\sigma}) - F(\frac{a-\mu}{\sigma})$

see the gif for their relationship

6 Sampling (being smart and lazy to study population with samples)

21、15 什么是sampling随机抽样

> probability with real world, real business of statistics is to tell us how lazy we can afford to be

The problem of the world (can't get all of them)

> information in the world is too large, we don't get all information we need or want to make a decision
>
> - see the figure on info and decisions problems

Solutions: use samples to replace population

> hard way
>
> - measure and collect all population information for decision
>
> lazy way
>
> - what statistician is doing for decision
> - take a sample not total population
> - but how big should the sample be?

sampling design (dataset is smaller but good quality)

> sample size n , but everything is governed by $\frac{1}{\sqrt{n}}$

quality of sample (avoid bias)

> - to avoid bias, we need to choose sample randomly

Simple random sample (sample elements are equally likely, unbiased, independence)

> - select n objects which are equally likely

> - unbiased
>   - each unit has equal chance of being chosen from the sample
> - independence
>   - selection of one unit has no influence on the selection of other units
>
> but, in real word
>
> - **unbiased, independent samples are impossible to find**

Theoretical possible but practically expensive

> - build a sampling frame to make selection
> - use a random number generator to pick objects at random (a sampling frame)
> - but still can be very costly and controversial
>
> but we have more efficent and cost-effective methods other than simple random sample

Stratified is more practical (group first, randomize sampe within each group)

> - divided into homogeneous groups
> - then do simple random sample for each group

Cluster sampling (group first, and randomize on group level)

> - group population into different clusters
> - make a simple random sample on all clusters

Systematic sampling (random by selecting every k-th element)

> - start with a randomly chose unit, then select every k^{th} unit afterward

Warnings (no random, no reliable analysis)

> most statistical analysis depend on simple random sample with
>
> - independence
> - unbiased
>
> but other methods above need modifications

> **without randomized design, there can be no dependable statistical analysis**; so, **opportunity sample** is an example of no randomization

How to approximate sample distribution's Expectation and variance on proportion? (with population stats)

use sample mean, variance to replace population mean and variance (but how reliable?)

> when a sample or more samples can be created, we can find out sample mean s
>
> **but how good is sample mean s compared to $\mu$the population mean?**
>
> in other words, **how good is $\hat{p} = \frac{x}{n}$(n is sample size) compared with real population p ?**

The real question (How well sample stats distributed around population stats)

> (since we can't know real p):
>
> - take many samples (sample size = 1000)
> - calculate many $\hat{p}$
> - can we use many $\hat{p}$ (its distribution) to approach (estimate) p ?
> - **How would all the $\hat{p}$ be distributed around p ?**

think samples as random variables (to use probability distributions)

> $\hat{p}$ is more and more like a random variable
>
> selection of n-unit sample is a **random experiment**
>
> the observation $\hat{p}$ is a **numerical outcome**
>
> **$\hat{P}$ is random variable, $\hat{p}$ is a particular sample's value**

how to calc sample mean and sd (formula)?

> **sample expectation is used as population proportion**
>
> - $E[\hat{P}] = p$
>
> **sample standard deviation is made of sample proportion**

- $\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ (equation 1)

for large sample size, sampling distirbution $\hat{P}$ is approx normal (see applet with right skewed as $X \sim Bin(n = 4, p = 0.3)$)

- the larger sample size, smaller the standard error

How to approximate sampling distribution's Expectation and variance on mean? (with population stats)

$X_1, X_2, X_3, \ldots X_n$ are n samples

sampling distribution (a collection of means of many samples)

$\overline{X} = \frac{X_1+X_2+X_3+...+X_n}{n}$ this is mean v.r.

or , collecting mean from each sample, to form a new dataset $\overline{X}$

sampling expectation (this mean approx population mean)

$E[\overline{X}] = \mu$ (**sampling expectation** closer to population mean)

sampling standard deviation (this sd is population sd divided by $\sqrt{n}$)

$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ (**sampling standard deviation** is made of σ population standard deviation and $\sqrt{n}$)

Sampling distribution is normal (assume sampling distribution is normal)

but what is the distribution of $\overline{X}$?

- approximate normal

Central Limit Theorem (use standard normal dist with many samples' mean $\overline{X}$ to approx population distribution)

22、16 CLT for proportion and mean 和 t distribution

definition: sample size is large, normal, mean and variance

> sample size = n, take many samples from population
>
> population mean = μ , population sd = σ
>
> as n get larger, sampling distribution approaches to **normal distribution**
>
> such sampling distribution has mean approximate μ, has its standard error approach to $\sigma/\sqrt{n}$
>
> $p(a \leq \overline{X} \leq b) = p(\frac{a-\mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{b-\mu}{\sigma/\sqrt{n}})$
>
> - without calculating mean and standard error of $\overline{X}$, just know population mean and standard deviation and sample size, we can know confidence interval of sampling distribution $\overline{X}$

given sample size is big enough, its sampling distribution $\overline{X}$ is always normal, regardless population distribution

> given sample size is big enough, its sampling distribution is normal
>
> see the figure , Try out the applet** on CLT on mean
>
> more (CLT on proportion)

The t-distribution (CLT operate when sample size small and σ unknown)

two problems of Central Limit Theorem (sample size small, σ unknown)

> CLT depends on **large sample size**
>
> CLT **needs σ to** calc standard error
>
> **In reality**
>
> - sample size is often small
> - σ is often unknown
> - although sampling distirbution is normal, hard to make use of Z-score table to find out probability

Solution: use sample sd s to replace population σ

> - $s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
> - $SE(\overline{X}) = \frac{s}{\sqrt{n}}$
> - $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ (**Z-score for CLT**)
> - $t = \frac{\overline{X}-\mu}{s/\sqrt{n}}$ (**t-score for CLT**)
>
> in general
>
> - **t-distribution is sloppier than Z distribution** (see the figure**)
> - **the larger the sample size, s is closer to μ , t-distribution is closer to z-distribution, to the normal**

Recap

> see the figure

7 Confidence Intervals (how confident that population mean is in the interval)

23、17 什么是confidence interval

> **Example .55 vote A given 1000 people, infer that "I am 95% confident that the true population proportion (vote A) is within .519 and .581"**

95% confidence

> see the figure and key figure ** to **explain what does 95% confidence mean**
>
> **95% confident that population mean is within the interval**

confidence interval (usually set to cover 80%, 95%, 99% of dataset, dive into for details)

> **standard deviation of sampling distribution**
>
> - $\sigma(p) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ (standard error) (equation 1) not sure where it is from

- $z = \frac{s-\mu}{\sigma} = \frac{\hat{p}-p}{\sigma(p)}$ (z-score formula, p is population mean proportion,)
- 0.95 = P(-1.96 $\leq$ Z $\leq$ 1.96)

$0.95 \simeq P(-1.96 \leq \frac{\hat{p}-p}{\sigma(p)} \leq 1.96)$

$0.95 \simeq P(\hat{p} - 1.96\sigma(p) \leq p \leq \hat{p} + 1.96\sigma(p))$, **find interval for population proportion**

we don't have $\sigma(p)$ but $SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$ (standard error for proprotion), **use sample proportion $\hat{p}$**

$0.95 \simeq P(\hat{p} - 1.96SE(\hat{p}) \leq p \leq \hat{p} + 1.96SE(\hat{p}))$, it interprets as

- **where is the 95% confidence interval for the population true proportion to locate**

both $\hat{p}, SE(\hat{p})$ are known, so confidence interval is known

back to TOC

How to make CI wider or narrower? (sample size smaller wider or bigger narrower)

to make it wider

- according to the formula above, is to make $SE(\hat{p})$ larger,
- then it means to make n smaller

to make it narrow

- according to the formula above, is to make $SE(\hat{p})$ smaller,
- then it means to make n larger

empirical calculation of CI examples (dive in for examples)

we want to know **where is the 99% confidence interval for the population true proportion to locate**

$0.99 \simeq P(\hat{p} - 2.58SE(\hat{p}) \leq p \leq \hat{p} + 2.58SE(\hat{p}))$

$\hat{p}, SE(\hat{p})$ can be calculated with sample dataset $SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$

see the figure** on the use of α and $1-\alpha$

see the figure** on different level of confidence (**a table of Z values for 80%, 95%, 99%**)

back to TOC

How large sample size is needed if want to be 95% confident and true proportion only oscilate 1%? (see the formula inside through algebra)

How to calc the sample size if **want 99% confidence and 1% of error?**

- to let CI cover 99% data (99% confidence that population mean is within) and the CI to be very narrow
- $n = \frac{(Z_{\frac{\alpha}{2}})^2 p^*(1-p^*)}{E^2}$ (equation 3)
- see the figure** for detail calculations

Confidence Intervals for μ (only need to replace $\hat{p}$ with $\overline{x}$ for the formula above, standard error formula is different, dive in)

it is the same story with CI for p

see the figure** for **comparison between proportion CI vs mean CI**

- standard error for proportion $SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$
- standard error for mean $SE(\overline{x}) = \frac{s}{\sqrt{n}}$

back to TOC

Student's T and degrees of freedom (if sample size is too small, use t table to replace z table above, to calc CI)

24、17 什么是confidence interval part2

t-distribution and degree of freedom

- see the figure**

confidence intervals and degree of freedom

- see the figure** (**pay attention to the degree of freedom table**)

Example: Car crash repair cost (dive in for details)

sample mean = 540 and standard deviation =299 can be calculated with ease

> t-table is available (degree of freedom is 4 degree = 2.78)
>
> we are 95% confident that μ is within the interval between $\overline{x} \pm 2.78 \times \frac{s}{\sqrt{n}}$
>
> see figure part1** part2** for details of this nice and simple example
>
> back to TOC

Recap

> $z_{\frac{\alpha}{2}} SE(\hat{p}), z_{\frac{\alpha}{2}} SE(\overline{X}), t_{\frac{\alpha}{2}} SE(\overline{X})$

8 Hypothesis Testing (CI and CLT help know population mean from sampling distribution mean and SE; individual sample can testify population distribution)

25、18 什么是Hypothesis testing part1

Ask differently (Could these observations really have occured by chance? Or they really mean something?)

Intriguing problem on black Jurors

> Jury panel racial discrimination in 1960-80s
>
> **statiscal evidence**
>
> - 50% of eligible citizens were African American
> - on a 80-person panel of potential jurors, only 4 were African American
>
> **could 4 black in 80-person jury just be chance, as outlier of $Bin(n = 80, p = 0.5)$, or does it suggest a different dist $Bin(n = 80, p < 0.5)$?**

Formulating the problem (select 80 people into a jury, binomial r.v. $X \sim Bin(n = 80, p = .5)$)

> suppose the selection was random (no racial discrimination)
>
> p(black juror) = 0.5, p(non-black) = 0.5
>
> select 80 people into a jury, binomial r.v. $X \sim Bin(n = 80, p = .5)$
>
> X value can be 0, 1, 2, ... 80 black people

Solution Attempted ($P(X \leq 4) = .00000000000000000014$ probability of being chance is so rare that we should not observe it)

> $P(X \leq 4) = .00000000000000000014$
>
> **such small probability is strong evidence against** the hypothesis of random selection

Four Formal Steps for Statistical Hypothesis Testing

26、18 什么是Hypothesis testing part2 4 steps

Step 1 Formulate all hypotheses (H_0, H_A)

> H_0 null hypothesis
>
> - observations are results of pure chance of null hypothesis
>
> H_A alternative hypothesis
>
> - observations are results of real effect of a different distribution

Step 2 The Test Statistics (define the r.v. distribution)

> identity a **statistic**
>
> - can access the evidence against null hypothesis
>
> H_0 says African juror is chosen by random, so $p = 0.5$
>
> H_a says African juror is chosen not randomly, so $p < 0.5$
>
> The test **statistic** is $X \sim Bin(n = 80, p = .5)$ to be tested

Step 3 P-value (calcuate the probability of the observations to occur in null hypothesis distribution)

> **definition**: if null hypothesis is true, then what is the probability of observing the test statistic?
>
> **interpretation**: ⚛️
>
> - if null hypothesis is true,
> - then p-value is very very small, i.e., probability of observing what we observed (observation support H_a) by chance is very very small, meaning, p-value $<$ 0.05 or 0.01
> - chance is so small, but we still can easily observe it
> - something is going on, this is evidence against null hypothesis

Step 4 Compare the p-value to a fixed significance level α (holdover artifact 0.05, 0.01)

> α is a cut-off point, below it suggests that observation is not caused by chance, but by real effect
>
> $P(X \leq 4 | p = .5, n = 80) = 1.4 \times 10^{-18}$
>
> p-value $\leq \alpha$ (often to be **0.05 or 0.01**, although they are just **holdover artifact in pre-computer era**)
>
> - null hypothesis is ruled out;
> - something is going on

Large Sample significance test for proportions (later for large sample mean)

27、19 significance test for proportion mean and t distribution

> Let make the example above more general
>
> - make $p = 0.5$ into $p = p_0$
> - 1000 observation sample
> - proprotion from this sample = 0.55 = $\hat{p}$
> - how significance is this .55? or is it just by chance?

The 3 steps solution with example for details

> H_0 : assume proportion=0.5 is the population proportion (in fact, it can be understood as the r.v. mean value too)

> H_a : proportion > 0.5
>
> p-value = $P(H_A|H_0)$ = $P(Z > Z_{obs} = \frac{p-p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}})$ (equation to find probability area or range)
>
> sampling sd = $\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}$ (important equation)
>
> see the figure** for three steps
>
> see the figure** for the working example

Large Sample test for the population mean

the problem

> Cereal company claims to send back a new shipment if the average weight of cereal is less than 16oz
>
> Grocery store won't weight every cereal box.
>
> statistics can help here

solution

> $H_0 : \mu = 16, H_A : \mu < 16$
>
> statistic: $Z = \frac{\overline{X}-\mu_0}{SE(\overline{X})} = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$ (for large samples, it is approx normal)
>
> choose $\alpha = 0.05$
>
> take a simple random sample of 49 boxes
>
> $\overline{x} = 15.9, s = .35$ for this sample
>
> 15.9 < 16, is this difference significant?
>
> $Z_{obs} = \frac{15.9-16}{.35/\sqrt{49}} = -2$
>
> $P(H_A|H_0) = P(Z < Z_{obs} = -2) = 0.0227$
>
> see the figure for graphs

Small Sample Test for the population mean

The problem of Car crash insurance

> if crash over $1000 repair cost, insurance company won't insure with the car company

Solution (example of acceptance sampling)

> $H_0 : \mu \geq 1000$, **cost too high, insurer not insure**
>
> $H_a : \mu < 1000$, cost is ok, insurer will insure
>
> test statistic is the **t distribution**
>
> $t = \frac{\overline{X} - \mu_0}{SE(\overline{X})}$, $\mu_0 = 1000$
>
> there are 5 crashes in total, according to table of critical t values
>
> - degree of freedom is 5-1=4
> - set significance level = 0.05
> - $t_{0.05} = 2.13$
>
> for crash result to satisfy significance
>
> - $H_a < 1000$, so $-t_{0.05} = -2.13$
> - $t_{obs} < -t_{0.05} = -2.13$
>
> **calculating t to replace Z for 5 observations of this sample**
>
> - $\overline{x} = 540, s = 299$
> - $z_{obs} = t_{obs} = \frac{540 - 1000}{299/\sqrt{5}} = -3.44 < -2.13$
> - above is equivalent to **p-value < 0.05**

Decision Theory (apply hypothese testing to decision making, to avoid or reduce type I or II error)

28、20 Decision Theory and Hypothesis testing

intriguing problem (no fire as null, alarm makes type 1; fire as alternative, and no alarm makes type 2 errors)

> Household Smok-detector

> Type I Error: there is alarm but no fire
>
> Type II Error: there is fire but no alarm

Solution to reduce error (trade off)

> make alarm less sensitiveless
>
> - type I error, but more type II error
>
> make alarm more sensitive
>
> - less type II error, but more type I error

Type I II Errors with H_0, H_A (nothing, something)

	No Fire	**Fire**
No Alarm	No Error	Type II
Alarm	Type I	No Error

	H_0	H_a
Accept H_0	No Error	Type II
Reject H_0	Type I	No Error

Confidence test vs Type I II Error

> P(rejecting H_0 | H_0) = P(Type I error | H_0) = α
>
> - this is what we studied above
>
> sometimes, we want to know how sensitive is our "alarm system" when alternative hypothesis is true
>
> - P(accepting H_0 | H_a) = P(Type II error | H_a)

Example with Type II Error Confidence Test

Environmental Example

> H_0 : the dischared polluter has no damaging effect
>
> H_a : the dischared polluter has damaging effect
>
> significance level: 0.05

Formalize the problem (likelihood of not detecting correctly and monitor effectiveness)

> $\beta =$ P(accepting H_0 | H_a) = P(type II error | H_a)
>
> - given there is damaging effect, what is the probability that no damage effect detected by monitor?
>
> $1 - \beta$
>
> - given real damage is true, what is probability of detecting damage effect
> - high probability shows monitor is sensitive, otherwise bad monitor program
> - environment agency require monitor programs to show they have a high probability (sensitivity) to detecting serious pollution **power analysis**

Intuition on confidence test on Type II error

> $\beta < 0.05$
>
> given the polluter has damage effect, the chance of observing damage is very unlikely, suggest that we won't see those observations in a decade; but reality is we observe them effortlessly. so something is going on.

what does the different monitor programs look like

> see the figure

9 Comparing two populations (apply CI and HT to difference of two populations as a new r.v.)

29、21 对比2个populations中如何使用CI和Hypo Testing part1 proportion

The real problems to deal with (Aspirin vs placebo, two pesticides, women vs men payment)

> Does taking Aspirin regularly reduce the risk of heart attack?
>
> Does a particular pesticide increase the yield of corn per acre?
>
> Does men and women in the same occupation have different salaries?
>
> We need to compare **two independent random samples**

The Asprin Problem (two population or sample experiments with their own proportions of success)

> In a given year, the chance of a person having heart attack is very small
>
> simple but expensive solution:
>
> - 22071 subjects were randomly assigned ot two groups
> - a group with a placebo pill per day, another with an aspirin per day
>
> **The experiment outcome** after 5 years in table
>
> - columns: attack, no attack
> - rows: placebo, aspirin
> - see the figure for details
>
> **The result compared**
>
> - attack rate with placebo = .0217
> - attack rate with aspirin = .0126 (seeming small diff)
> - relative rate placebo/aspirin = 1.72 (big difference)

The Model construction (create their difference as new r.v. and use hypothesis testing again)

> see the figure for detail
>
> **placebo and aspirin groups**
>
> - two independent samples
> - from two binomial populations
> - r.v. value is about number of **success** (heart attack)
>
> **The objective of model**
>
> - to estimate the true difference $p_1 - p_2$
> - to estimate the true difference rate $\hat{p_1} - \hat{p_2}$

Confidence intervals for $p_1 - p_2 = \hat{p}_1 - \hat{p}_2 \pm Z_{\frac{a}{2}} SE(\hat{p}_1 - \hat{p}_2)$

> see the figure** for calculation detail
>
> Get **Z score values**
>
> - sample size is large, we approx it to be normal
> - we can choose z to select confidence level
> - since we don't have $p_1 - p_2$, we can't calc z directly with $z = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sigma(\hat{p_1} - \hat{p_2})}$
> - we can use z-score table
>
> Get **Standard error**
>
> - $\sigma^2(\hat{p_1} - \hat{p_2}) = \sigma^2(\hat{p_1}) + \sigma^2(\hat{p_2})$
> - $SE(\hat{p_1} - \hat{p_2}) = \sigma(\hat{p_1} - \hat{p_2}) = \sqrt{\sigma^2(\hat{p_1}) + \sigma^2(\hat{p_2})}$
> - $\sigma^2(\hat{p_1}) = \hat{p}_1(1 - \hat{p}_1)/n_1, \sigma^2(\hat{p_2}) = \hat{p}_2(1 - \hat{p}_2)/n_2$

Hypothesis testing (null = no difference, alternative = real difference, p-value with cont r.v. range to testify)

> If Aspirin has no effect, what is the probability of this observation to occur by chance?
>
> $H_0 : p_1 = p_2$
>
> - Aspirin has no effect
>
> $H_a : p_1 > p_2$
>
> - Aspirin can reduce heart attack

> Under H_0
>
> - $p_1 - p_2 = 0$
> - $z = \frac{\hat{p_1}-\hat{p_2}-(p_1-p_2)}{\sigma(\hat{p_1}-\hat{p_2})} = z = \frac{\hat{p_1}-\hat{p_2}-(0)}{SE(\hat{p_1}-\hat{p_2})} = 5.2$ (calculated ok)
> - we expect aspirin to reduce heart attack, so $\hat{p_1} - \hat{p_2} > 0$
> - p-value = $P(Z_{obs} > 5.2 | H_0)$, this is **cont r.v. must be a range**
>
> see the figure part1** part2** for details

Assumptions for Aspirin example to work (random assigned, blind experiment, large sample size)

> subjects are randomly assigned
>
> experiment was blind: subjects don't know which they are taking, placebo or aspirin
>
> the sample size was large enough to assume approx normal

Comparing the means of two populations

30、22 comparing mean of two samples with large and small sample size

Men and Women Salary Example (Estimator $\mu_1 - \mu_2 = \overline{X}_1 - \overline{X}_2$)

> women
>
> - population one, with μ_1, σ_1
> - get a random sample n_1, mean and sd $\overline{x}_1, S_1$
>
> men
>
> - population two, with μ_2, σ_2
> - get a random sample n_2, mean and sd $\overline{x}_2, S_2$
>
> Estimator $\mu_1 - \mu_2 = \overline{X}_1 - \overline{X}_2$

Solution ($SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, rest is the same)

> for large sample size, the estimator is approx normal
>
> **its standard error**

- $SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- due to samples are independent, the variances add

confidence intervals

- $\mu_1 - \mu_2 = \overline{x}_1 - \overline{x}_2 \pm Z_{\frac{\alpha}{2}} SE(\overline{X}_1 - \overline{X}_2)$

Hypothesis testing

- $H_0 : \mu_1 - \mu_2$
- $Z_{obs} = \frac{\overline{X}_1 - \overline{X}_2}{SE(\overline{X}_1 - \overline{X}_2)}$ test statistic
- from Z we can calc probability p-value

Comparing Small sample means

Car crash comparison problem (a sample with 5 vs a sample with 7)

Car company 1 crash 5 cars

Car company 2 crash 7 cars

see their table on sample size, mean and variance

n_1	**5**	n_2	**7**
$\overline{x}_1$	540	$\overline{x}_2$	300
s_1	299	s_2	238

degree of freedom = $n_1 + n_2 - 1 - 1 = 2.23$

t-distribution needs 2 conditions (mound shape and shared sd)

both population are mound shaped

both standard deviations are the same

standard error (use t-table with their shared standard deviation s^2_{pool}, and the rest is the same)

$s^2_{pool} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 264$

> $SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_{pool}^2}{n_1} + \frac{s_{pool}^2}{n_2}} = s_{pool}\sqrt{1/n_1 + 1/n_2} = 154$
>
> $\alpha = 0.05, 1 - \alpha = 0.95$
>
> (1-a)100% confidence interval =
> $\mu_1 - \mu_2 = \overline{x}_1 - \overline{x}_2 \pm t_{a/2} SE(\overline{X}_1 - \overline{X}_2) = 540 - 300 \pm t_{0.025}(154) = 240 \pm 340$
>
> see the figure** for calculation detail

conclusion

> since 0 is included inside confidence interval
>
> we can't say one group is better than the other

Paired Comparison test is better (specific paired difference is necessary, due to natural variability)

31、23 paired comparison on car gas mileage difference

> nature variability is a problem

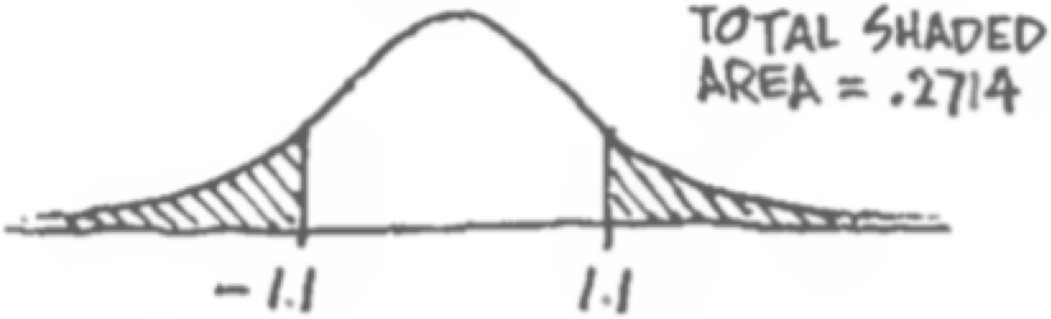
Gas per mile Example without paired comparison

> samples: large size, but variability is large too
>
> see the figure for detailed calculation
>
> Cab group A: $\overline{x}_1 = 25, s_1 = 5, n_1 = 50$
>
> Cab group B: $\overline{x}_2 = 26, s_2 = 4, n_2 = 50$
>
> $\overline{x}_1 - \overline{x}_2 = -1$
>
> $SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.905$
>
> $\mu_1 - \mu_2 = \overline{x}_1 - \overline{x}_2 \pm z_{0.025}(0.905)$ = -1 $\pm(1.96)(0.905)$
>
> $z_{obs} = \frac{\overline{x}_1 - \overline{x}_2 - 0}{SE(\overline{X}_1 - \overline{X}_2)} = -1/0.905$
>
> but H_a : $\mu_1 \neq \mu_2$, we need both parts of probability

$$Pr(|z| \geq |z_{OBS}|) = Pr(|z| \geq \frac{1}{.905})$$

$$= Pr(|z| \geq 1.1) = 2(.1357)$$

$$= .2714$$



- part** 1 and part** 2

Gas per mile with paired comparison (figure for details)

what does paired difference mean?

CAB	GAS A	GAS B	DIFFERENCE
1	27.01	26.95	0.06
2	20.00	20.44	−0.44
3	23.41	25.05	−1.64
4	25.22	26.32	−1.10
5	30.11	29.56	0.55
6	25.55	26.60	−1.05
7	22.23	22.93	−0.70
8	19.78	20.23	−0.45
9	33.45	33.95	−0.50
10	25.22	26.01	−0.79
MEAN	25.20	25.80	−0.60
STANDARD DEVIATION	4.27	4.10	0.61

part** 1, part** 2 and part** 3

10 Experimental design (decisive)

> design of experiment often spells success or failure in the paired comparison example

Components of Experiment design (units, treatment, assignment, control variability with randomization)

> **components**
>
> - experimental units
> - treatments: assigned to units
> - objective: to compare treatments
>
> Medical Example
>
> - patients are units
> - drugs are treatments
>
> Gas per mile Example
>
> - cabs are units
> - treatments to be compared: Gas A and Gas B
>
> Agriculture Example
>
> - crops are units
> - crop varieties, pesticide etc are treatments

Replication

> definition
>
> - assign the same treatment to different experimental units
>
> significance

- without it, impossible to assess natural variability and measurement error

Local control

definition

- any method that account for and reduces natural variability

methods

- group similar experimental units into blocks
- each cab is a block for gasolines

Randomization

- randomize the day for cab to run
- so no effect on the difference day the cab run

A complete Randomized block

compare the effect of two brands on tires and gas

- see the figure for detail

compare the effect of different 4 days

- see the figure for details

Conclusion

experimental design is to allocating total variability among different sources

variability sources

- cab, tire, gas type, day, random error

Analysis of variance, ANOVA

- partitions the total variability
- allocating portions to each source

11 Regression (compare other linear relation or predictability between two populations)

32、24 什么是 regression

33、25 什么是 regression

> previously, dealing with single variable
>
> **examples**
>
> - weights vs heights
> - blood pressure vs life expectancy
> - SAT scores vs college performance
> - Reading statistics vs making a better person
>
> **relations in Graphs**
>
> - display variable relation nicely like a straight line
> - but data is never clear for displaying
> - see heights vs weights scatterplot

Regression analysis (can we predict student weight with height?)

> x is **predictor** variable, y is response variable
>
> $y = b \times x + a$ is **regression**

present the data on graph

> data on table to data on graph, see figure

Find the best fitting line (line indicate the relationship)

> **How to find the best line that fit all the data?**
>
> **criterion**

- the best line has the smallest distance to all the data point
- SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ to be the minimum
- find the best line by trying out all values of a, b

Why called Regression? (tall short fathers and sons relation)

Francis Galton on laws of inheritance

regression toward the mean

- taller fathers have shorter sons (to meet population mean)
- shorter fathers to have taller sons (to meet population mean)
- **line is following the regression rule, staying the middle**
- think of the graph above

Calculation to find the best line

formula to find b and a, see the figure

calculation process, see the figure

ANOVA (analysis of variance, SSE vs total variance, SSR vs total variance, good vs bad fit)

how good is the best fit line ?

3 types of variability (SSR, SSE, total variance)

variability by regression vs error vs total

- SSR vs SSE vs SS_{yy}
- see the table for formula of the 3 variability

good fit vs bad fit, see the figure**

good fit

- small SSE even compared with small total spread (top left)
- moderate SEE but large total spread

bad fit

- big SSE given small total spread
- large SSE given large total spread

The squared correlation (R^2 = 1- SSE/total, goodness of fit)

$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$

$R^2 = 0$, SSE is very large, error is very large, so fit very badly, see the figure

$R^2 = 1$, SSE is very small, error is little, so fit perfectly

$R^2 = 0.58$:

- 58% variation of weight is explained by height
- 42% is error

Correlation Coefficient ($r = \sqrt{R^2}$ measure goodness of fit + direction)

r = (sign of b) $\sqrt{R^2}$

$+$: line toward right up

$-$: line toward right down

r measures both tightness of fit and direction of line

- close to $|1|$, fitting perfectly
- $\pm$ refers to direction
- see the figure

Statistical Inference

Regression model (find CI for a, b, epsilon distributions)

to describe the whole population with a linear relation

$y = \alpha + \beta x + \epsilon$

α, β are unknown parameters to map from x to y

ϵ makes the random distribution component of weight y for each height x

see the graph** detail

To simplify the ϵ by making it standard normal distribution for all x

> assume all values of X, the ϵ are
>
> - independent
> - normal with same mean 0 and sd 15 (e.g.)
> - see the figure** for detail

How to find the best fitting line

> each sample of data is different
>
> each sample generate a regression line
>
> many regression lines distribute around the model line
>
> how a, b distributed around α, β?
>
> how to construct confidence interval and hypothesis testing

Calculate standard devation of each sample

> how to calculate s for each sample dataset?
>
> $s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$
>
> - why n-2 (degree of freedom)? (figure)
>   - due to a, b take two parameters

Confidence Intervals $y = a + bx + \epsilon$

> $SE(b) = \frac{s}{\sqrt{SSxx}}$, $SE(a) = s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SSxx}}$
>
> $\beta = b \pm t_{0.25} SE(b)$
>
> $\alpha = a \pm t_{0.25} SE(a)$

What is the mean weight of students at height 76 inches?

> - $\alpha + \beta x_0 = a + bx_0 \pm t_{0.25} SE(\hat{y})$, $SE(\hat{y}) = s\sqrt{1/n + \frac{(x_0 - \overline{x})^2}{SSxx}}$

when a new student with a new height comes in, how to predict her weight without measure it? (what is the 95 confidence interval of the mean weight for student at new height?)

> - $Y_{new} = a + bx_{new} \pm t_{0.25} SE(Y_{new})$, $SE(Y_new) = s\sqrt{1/n + \frac{(x_{new} - \overline{x})^2}{SSxx}}$

the intervals are terrible for two reasons (Height alone is not enough, 9 data points are too few)

> detail info see part** 1, part2, part3**, part4**

Hypothesis Testing (testify whether null hypo = no relation = $\beta = 0$ is true or not)

> the calc detail process** is the same

Multiple linear regression (use matrix algebra and computer)

> matrix algebra + computer can help a lot

Non-linear regression (use liner regression polynomial technique to deal with non-linear problem)

> use linear regression technique to solve non-linear problem
>
> see the figure**

Regression diagnostics (use epsilon vs $\hat{y}$ to check assumption sanity)

> fitting a complex model to data can obscure many difficulties
>
> a simplest procedure is to graph the relation between epsilon and y
>
> see the figure for detail