No dumb questions!

Chapter 1 Introdution to Data

1.2.

Observation: Each row in the data set is an observation / case / unit.

Variable: Each column represents some characteristic, varies between observations.

Numerical Variables: the values are numeric, the numbers are meaningful, measuring quantities.

- Discrete Numerical :
- Continuous Numerical:

Categorical Varibles: the values are categories / types / Labels

- Ordinal Categorical: the categories are ordered.
- Nominal Catcgorical: " " unordered.

Scatter plot: Use dots to denote data from 2 variables.

Positively associated:  scatterplot shows an upward trend

Negatively associated:  " " downward trend

Response Variable (: Dependent Variable.)

Explanatary Variable: (Independent Variable)

1.3

Population: all the units of interest.

Individual Cases / Subjects : one unit in a population.

Anecdotal Evidence: (small sample size. extreme cases)

Sampling "Representative of the population"

- Simple Random Sampling — "raffle method"
- Statified Sampling — "divide and conquer"
- Clustering — "mini population"
- Multi-stage

1.4 Experiments "Apple to Apple"

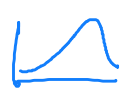
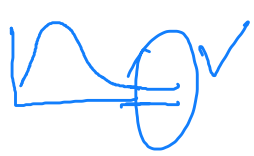
1. Controlling
2. Randomization
3. Replication
4. Blocking

Chapter 2

2.1

Scatter Plot —— "2 variables"

Dot Plot
Stacked Dot Plot
Histogram: bin size "same"
} 1 variable

Shape: { Symmetric
Left Skewed 
Right Skewed 

2.2 Center

① "Mean":

Sample mean $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Population mean μ

② Median: The number that cuts the data in half. "50th "percentil.

odd # of obs: $\frac{n+1}{2}$ position

even # of obs: $\frac{n}{2}$, $\frac{n}{2}+1$ take average

③ Mode:

Variation/Spread.

① Variance:

Sample Variance $s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Sample Standard deviation $s = \sqrt{s^2}$

Population Variance σ^2

Population St. dev. σ

② $IQR = Q_3 - Q_1$ Interquartile Range.

Robust Statistics : insensitive to outliers.

✓ IQR, Median

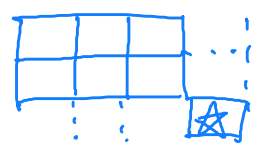
✗ Var, St. dev. mean

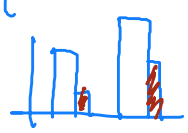
2.3 Visualization

1 Var { Bar Chart / Pie Chart

2 Categorical Variables

Contingency table (2×2, $3\times4\dots$)



Stacked Bar chart 

Mosaic Chart 

1 Categorical Variable + 1 Numerical Variable.

Side-by-Side Box plot

Stacked Histogram.

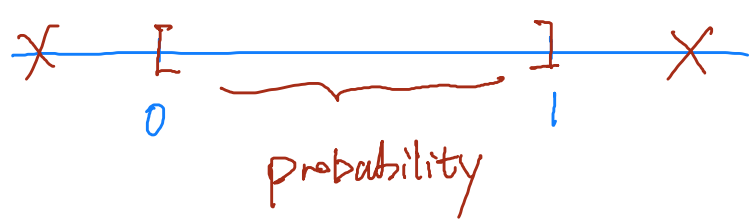
Chapter 3 Probabilities — Part 1.

3.1

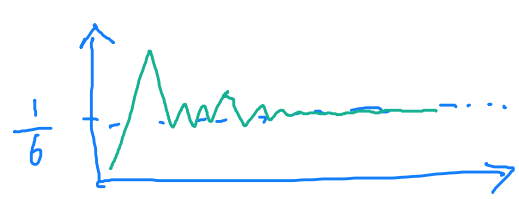
Probability. "infinite number of time" → proportion

0 → never

1 → always



Law of Large Number



Sets and Events

Event : a set outcomes we are interested in

Use A, B, C... $A = \{4, 6\}$ $P(A) = \frac{2}{6} = \frac{1}{3}$

$B = \{2, 4, 6\}$

Disjoint / Mutually Exclusive Events

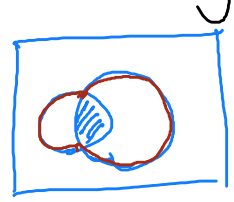
A and B X dispoint.

General Addition Rule

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$\cup$

Venn Diagram



Probability Distribution : a list of all possible results with corresponding probabilities

Price	< 300k	300≤ <600k	600k – 1000k	>1000k

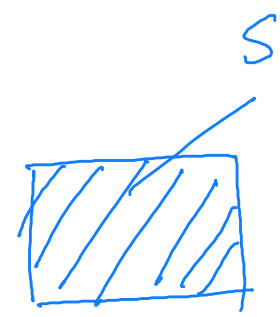
① all results/outcomes are disjoint/mutually exclusive

② all possible results/outcomes.

③ all probabilities add up to 1.

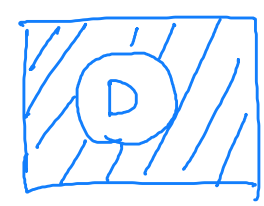
Sample Space : A set of all possible outcomes.

- Denoted by "S"
- $P(S) = 1$



Complement of event D : "not" D ,

- Denoted by D^c → not D



outside of cirde D^c

- $P(D) + P(D^c) = 1$
- D and D^c are disjoint. $P(D \text{ or } D^c) = 1$

Independent events: P(A) does not depend on whethe B happens or not.

- $P(A \text{ and } B) = P(A)P(B)$
- If A and B are disjoint , then A and B are NOT indep.

A even number {2, 4, 6} 50%

B greater than 3 {~~3~~, 4, 5. 6} 50%

A. B not disjoint.

$P(A \text{ an } B) = \frac{2}{6} = \frac{1}{3}$ $P(A) \times P(B) = \frac{1}{2} \times \frac{\cancel{4}^{2}}{6} = \frac{1}{3}$

3.2 Conditional Probabilites.

Algorithm	Bad	Good
Bad	197	22 X
Good	112 X	1491
	1513	

1. ✓ $P(\text{Algorithm say "Good"} \mid \underline{\text{Good}}) = \frac{1491}{1513} = 98.5\%$ ←

2. ✓ $P(\text{Algorithm say "Bad"} \mid \underline{\text{Bad}}) = \frac{197}{197+112} = \underline{63.7\%}$ ← 100%

$$P(\text{error rate}) = \frac{112+22}{197+112+22+1491} = 7.4\%$$

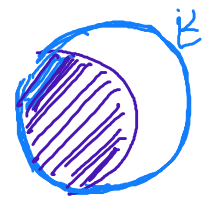
3. $P(\text{correct rate}) = 1 - P(\text{error rate}) = 93.6\%$ ←

Condition : information we know to be true. (B)

Outcome of Interest : what we want to know. (A)

$$P(\underset{\text{of interest}}{\text{outcome}} \mid \underset{\text{given}}{\text{condition}}) = P(\underset{\text{given}}{\text{Label good} \mid \text{good}})$$

Label Good



$$= \frac{P(\text{Label as good})}{P(\text{good videos})} = \frac{\text{A}}{\text{U}}$$

$P(A \cup B)$ ' U or

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

General Multiplication Rule.

$$P(A \cap B) = P(B) \cdot P(A \mid B) \checkmark$$

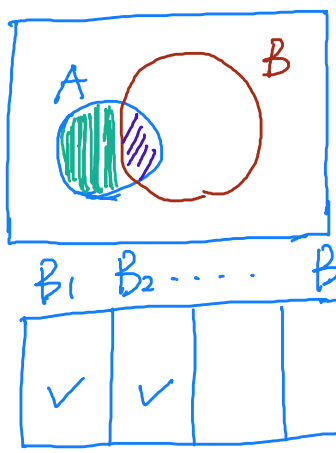
$$P(A \cap B) = P(A)\, P(B \mid A)$$

$$P(A \cap B\, C) = P(A)\,P(B|A)\cdot P(C|A\cdot B)$$

Independent

$$P(A \cap B) = P(A)P(B|A) \overset{\text{indep}}{=} P(A)P(B)$$

$$\text{☆}\quad P(B) = P(B|A) = P(B|A^c)$$



$$P(A) = \text{(purple)} + \text{(green)}$$

$$= P(B)\,P(A|B) + P(B^c)\,P(A|B^c)$$

$$= P(B_1)\,P(A|B_1) + P(B_2)\,P(A|B_2) + \cdots + P(B_k)\,P(A|B_k)$$

Chapter 3 Part 2

Random Variable:

A variable is a <u>random variable</u> if the value it takes on is a random event. (win/lose election, # of heads if flipping 5 coins).

Expected Value of a R.V. → $E(X)$. μ, mean

X, x_1 with $P(X=x_1)$, x_2 with $P(X=x_2)$ ... x_n with $P(X=x_n)$

$$E(X)=\mu = x_1 \cdot P(X=x_1) + x_2 P(X=x_2) + \cdots + x_n P(X=x_n)$$

$$= \sum_{i=1}^{n} x_i P(X=x_i)$$

Variance of a R.V.

$$Var(Y)=\sigma^2 = (x_1-\mu)^2 P(X=x_1) + \cdots + (x_n-\mu)^2 P(X=x_n)$$

$$= \sum_{i=1}^{n} (x_i-\mu)^2 P(X=x_i)$$

$$std(Y)=\sigma=\sqrt{Var(Y)}$$

Linear Transformation of a. R.V. X

$$Y = aX + b$$

$$E(Y)= aE(X) + b$$

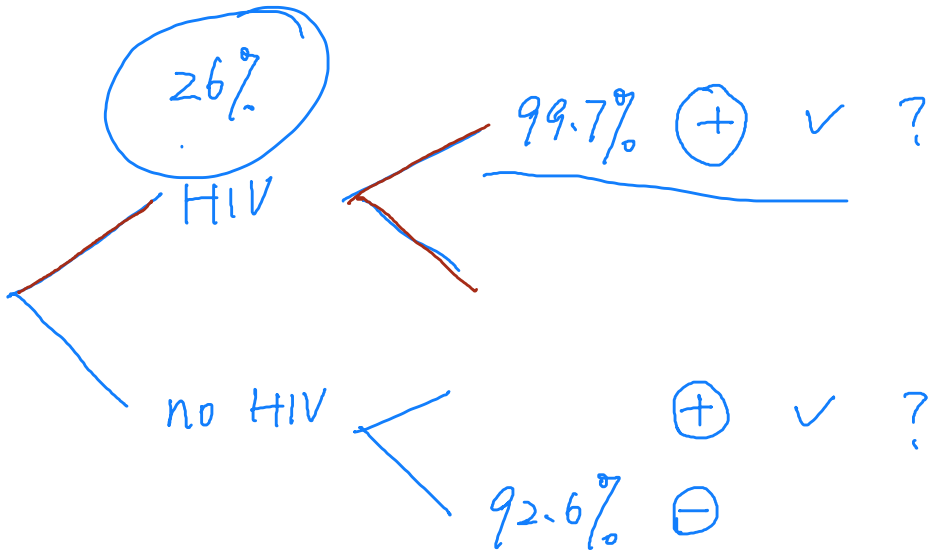
$$Var(Y)= a^2 Var(X)$$

$$Z = X_1 + X_2 + X_3$$

$$E(Z) = E(X_1) + E(X_2) + E(X_3)$$

$$Var(Z) = Var(X_1) + Var(X_2) + Var(X_3)$$

independent X_1, X_2, X_3



(a) $(26\%)(99.7\%) + (1-26\%)(1-92.6\%) = P(\oplus)$

(b) $\dfrac{(26\%)(99.7\%)}{P(\oplus)} = P(\text{HIV} \mid \oplus)$

Chapter 4 Distributions

1. Bernoulli Distribution
2. Binomial Distribution ←
3. Normal Distribution

1. Bernoulli Distribution

30% like blue cheese.

X	1 (Yes)	0 (No)
prob	0.3	1−0.3

$p = P(\text{success}) = 0.3$

X	1	0
prob	p	$1-p$

$E(X) = p$

$Var(X) = p(1-p)$

2. Binomial Distribution:

3 randomly selected student, X : # of students like blue cheese.

X	0	1	2	3
prob				

$X = X_1 + X_2 + X_3$

Andy ★ Brad ★ Cathy ★

Andy	Brad	Cathy	
✓	x	x	
x	✓	x	$p^1(1-p)^2 \cdot {}_2\binom{3}{1}$
x	x	✓	

$P(X=0) = (1-0.3)(1-0.3)(1-0.3) = (1-0.3)^3$

$P(X=3) = (0.3)(0.3)(0.3) = 0.3^3$

$P(X=k)$

$P(X=1) = (0.3)(1-0.3)(1-0.3) \times 3 = (0.3)(1-0.3)^2 \cdot 3$

$P(X=2) = (0.3)(0.3)(1-0.3) \times 3 = (0.3)^2(1-0.3)^1 \cdot 3$

$$E(X) = \sum x_i P(X = x_i) = 0(1-0.3)^3 + 3(0.3)^3 + 1(0.3)(1-0.3)^2 \cdot 3 + 2(0.3)^1(1-0.3)$$

$$= E(X_1 + X_2 + X_3)$$

$$= E(X_1) + E(X_2) + E(X_3)$$

$$= (0.3) + (0.3) + (0.3)$$

$$= (0.3) \times 3$$

$$Var(X) = Var(X_1 + X_2 + X_3) = Var(X_1) + Var(X_2) + Var(X_3)$$

$$= (0.3)(1-0.3) + 0.3(1-0.3) + 0.3(1-0.3)$$

$$= 3(0.3)(1-0.3)$$

Binomial Setting

1. X is # of successes from "n" independent trials. (V/X)
2. "independent"
3. Probability of success in each trial is constant "p"
4. "n" is fixed.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

↓

#success

$$E(X) = E(X_1 + X_2 + \cdots X_n) = p + p + p \cdots p = np$$

$$Var(X) = Var(X_1 + X_2 + \cdots X_n) = np(1-p)$$

Counting

- $\binom{n}{k}$ n choose k ✓ $\frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!\,(n-k)!}$ Combination C_n^k

$$k! = k(k-1)\cdots 1$$

rank

- order k items $\quad k(k-1)(k-2)\cdots 1 = k!$

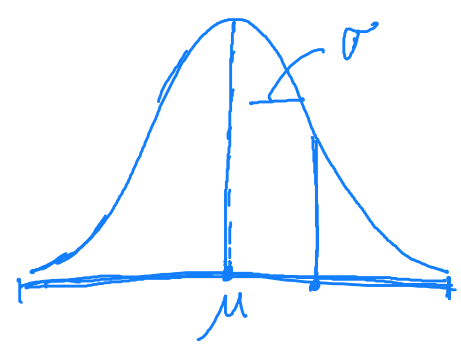
  order k items from n $\quad 100\times 99\times 98 \quad P_{100}^{3}$ $\quad$ Permutation. P_n^k

$3! = 6.$

$$\binom{5}{2} = \frac{5\times 4}{2\times 1} = 10$$

3. Normal Distribution (Continuous R.V.)

Whole area under the curve is 100%.



(X) $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$X \sim N(\mu, \sigma)$$

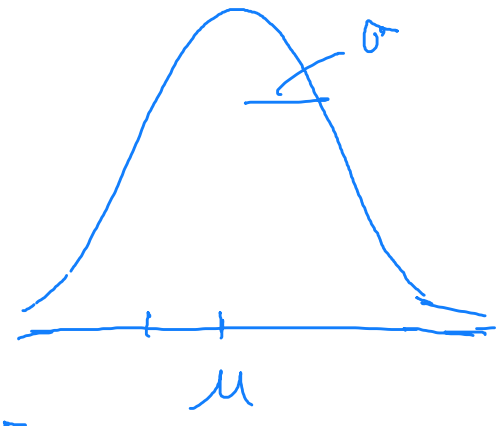
$$P(X \geq \mu) = 50\%$$

$$P(X = \mu) = 0$$

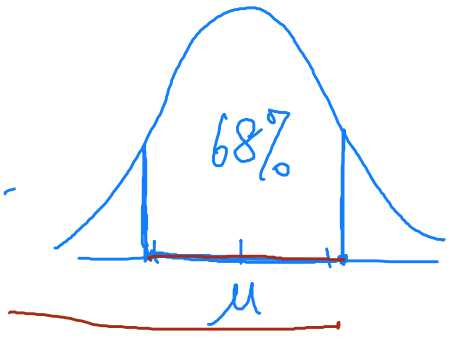
↑

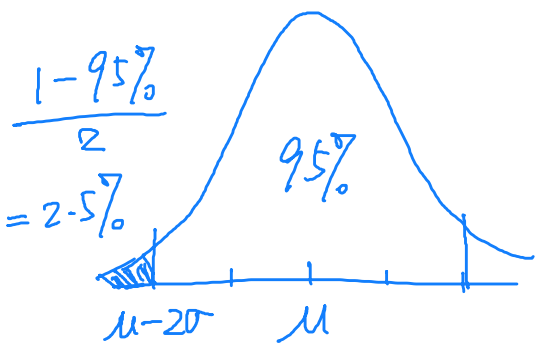
Z-score : how many st.dev is X away from the center.

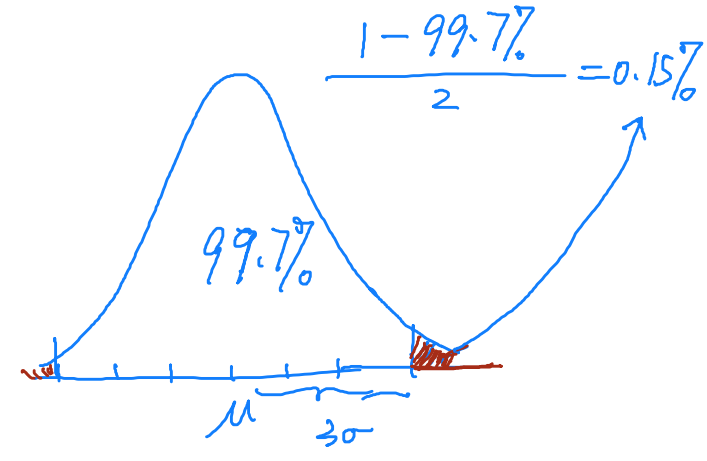
$$Z = \frac{X-\mu}{\sigma}$$



Empirical Rule 68 − 95 − 99.7%







$\mu \pm \sigma$: 1σ within μ $\quad$ $\mu \pm 2\sigma$: 2σ within μ $\quad$ $\mu \pm 3\sigma$

Chapter 5

Trump Approval Rate in 2020 June.

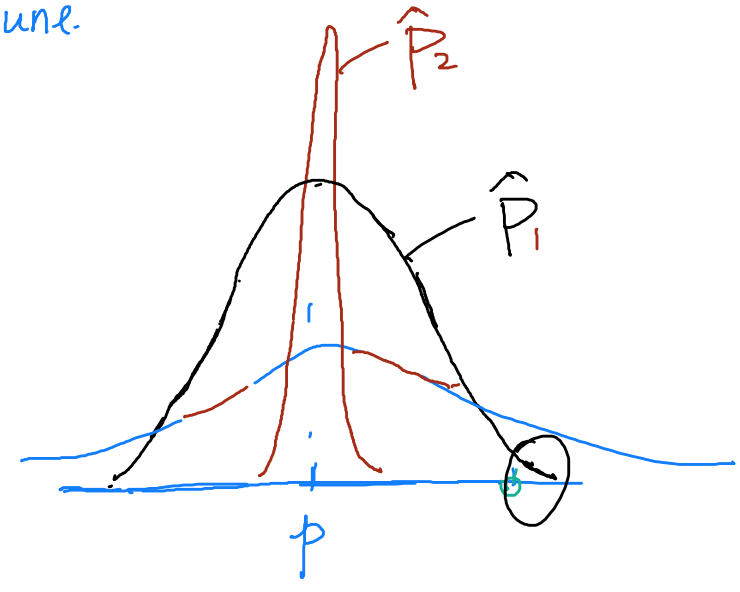
① $n=5000$ $\hat{p}_1 = 41\%$

② $n=5000$ $\hat{p}_2 = 42\%$

③ $n=5000$ $\hat{p}_3 = 38\%$

$\vdots$

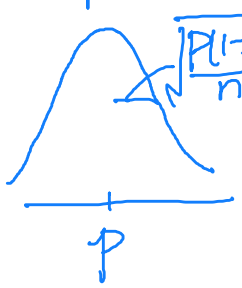
$\hat{p}_{1000} =$



- Central Limit Theorem (CLT)

$\hat{p} = \frac{x_1 + x_2 + \cdots x_n}{n}$ The sample proportion $\hat{p}$ will approximately follows a normal distribution with. p as the mean and $\sqrt{\frac{p(1-p)}{n}}$ as the standard deviation if

$np \geq 10$ and $n(1-p) \geq 10$ → success/failure condition.



$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \text{ if } np \geq 10,\ n(1-p) \geq 10$$

$n\hat{p}$ $\quad$ $n(1-\hat{p})$

- Confidence Interval

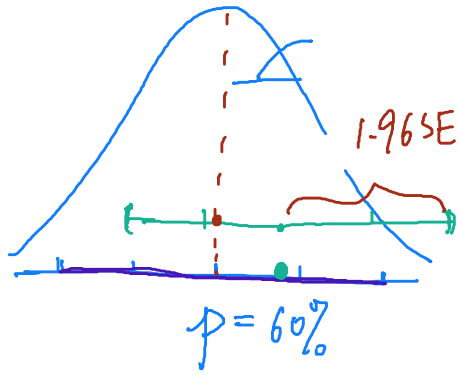
$\hat{p} \pm$ margin of error

$\hat{p} \pm z^* \cdot SE$

↑

z^*

95%	→ 1.96
90%	→ 1.645
99%	→ 2.58



sampling distribution for $\hat{p}$

$z^* = qnorm(0.995)$

$\hat{p} \pm 1.96$ S.E. $\rightarrow$ 95%

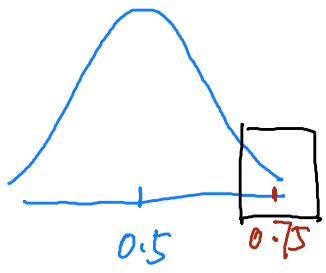
① Large n $\rightarrow$ smaller m.o.E.

② lower confidence level $\rightarrow$ smaller m.o.E.

- Hypothesis Testing

50% — overall death rate.

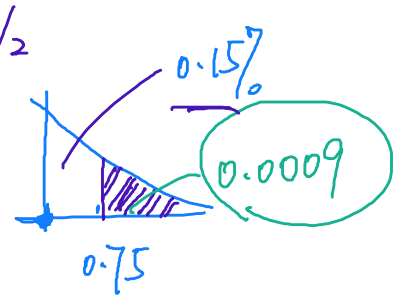
40 patients — 30 died. $\rightarrow$ 75%



$\hat{p} \sim N\left(0.5, \sqrt{\frac{0.5(1-0.5)}{40}} = 0.08\right)$

$z = \frac{0.75 - 0.5}{0.08} = 3.125$ ← pnorm(−3.125)=0.0009 ✗

$(1 - 99.7\%)/2$



H_0: $p = 0.5$ ☆ Null hypothesis

H_a: $p > 0.5$ Alternative hypothesis

	H_0 true	H_0 false.
✓ Reject H_0	X I	✓
Fail to reject H_0	✓	X II

Reject H_0

Type I : H_0 is true, reject H_0.

Type II : H_0 is false, fail to reject H_0

Chapter 6 Inference for Population Proportion

- Hypothesis Test for one "p"
- Hypothesis Test for two "p's" $p_1 - p_2$

Hypothesis Test for one "p"

① $H_0: p = p_0$ X

$H_a: p \neq p_0$ — two sided test

$p > p_0$ } one sided test

$p < p_0$ }

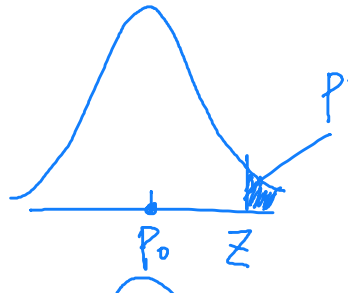
② Check CLT condition

- independent sample
- $np_0 \geq 10$ and $np_0(1-p_0) \geq 10$

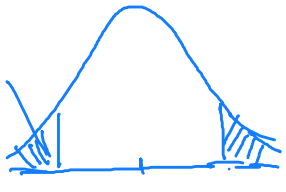
③ 

$$\hat{p} \sim N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

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

④ 

p-value : how likely to observe some data equally or more extreme than $\hat{p}$.

x2 

$H_a: p \neq p_0$

⑤ Conclusion: compare p-value with significance level (α) Type I error.

p-value $\geq \alpha$ → fail to reject H_0

p-value $< \alpha$ → reject H_0

0.01 0.05

HT for $p_1 - p_2$

trt1: $n_1 = 109$ $\hat{p}_1 = 60\%$

trt2: $n_2 = 91$ $\hat{p}_2 = 65\%$

① $H_0: p_1 = p_2$ ←

$H_a: p_2 - p_1 > 0$ trt2 is better?

② Check CLT for $\hat{P}_1, \hat{P}_2$

independence: within each group, between each group

success/failure: $n_1\hat{p}^* \geq 10$ $n_1(1-\hat{p}^*) \geq 10$

$n_2\hat{p}^* \geq 10$ $n_2(1-\hat{p}^*) \geq 10$

$$\hat{p}^* = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \frac{109(0.60)+91(0.65)}{109+91} \approx 0.62$$

Under H_0

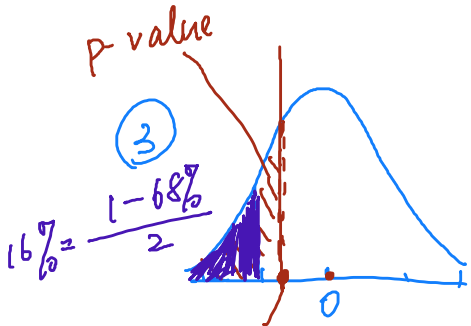
$\hat{P}_1 \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$ X

$\hat{P}_2 \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$ Y

$Var(X-Y) = Var(X) + Var(Y)$

$\hat{P}_1 - \hat{P}_2 \overset{H_0}{\sim} N\left(0, \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n_1} + \frac{\hat{p}^*(1-\hat{p}^*)}{n_2}}\right)$

$N\left(0, \sqrt{\hat{p}^*(1-\hat{p}^*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)$ $\sqrt{(0.62)(1-0.62)\left(\frac{1}{109}+\frac{1}{91}\right)} = 0.069$



$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}^*(1-\hat{p}^*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.6-0.65}{0.069} = -\frac{0.05}{0.069} \approx -0.725$$

$z = -0.725$

④ p-value ≥ 0.16 using empirical rule.

⑤ Conclusion: With a large p-value $= 16\% > \alpha = 5\%$, we have no strong evidence to claim trt 2 is better than trt 1.

Confidence Interal:

for p: $\hat{p} \pm z^* \cdot SE$ $\quad SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

for $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2 \pm z^* \cdot SE$ $\quad SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p})}{n_2}}$

z^* :	90%	1.645
	95%	1.96
	99%	2.58

Sample Size for a target M.O.E. of p

$$M.O.E = z^* \cdot \sqrt{\frac{p(1-p)}{n}} \leq \alpha$$

if prior inf p $n \geq \left(\frac{z^*}{\alpha}\right)^2 \hat{p}(1-\hat{p})$

if no prior info $n \geq \left(\frac{z^*}{\alpha}\right)^2 \cdot \left(\frac{1}{4}\right)$

always round UP

Chapter 7

① Hypothesis Test for μ (one sample mean)

CLT for $\bar{X}$

① independent data $X_1, X_2 \dots, X_n$ independently collected from a distribution.

②
- * if $X_1, X_2 \dots X_n$ are from a Normal distribution ✓
- or
- * $n \geq 30$, no extreme outliers ☆
- or
- * $n < 30$, no outlier

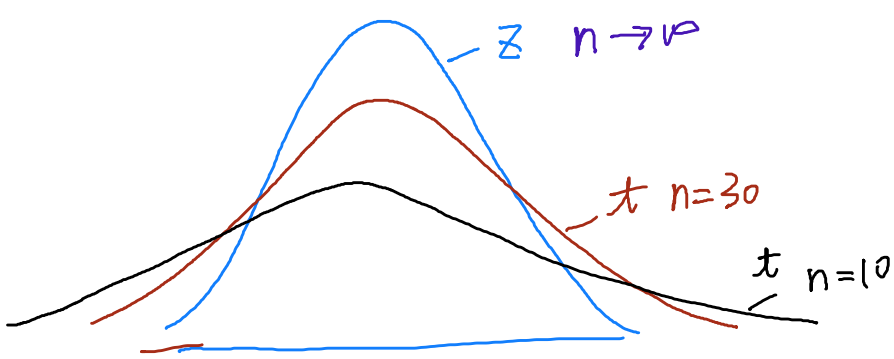
$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$$

$$E(\bar{X}) = \left(E(X_1) + E(X_2) + \cdots E(X_n)\right)\frac{1}{n} = \frac{1}{n}(\mu + \mu + \cdots \mu) = \mu$$

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 \left(Var(X_1) + Var(X_2) + \cdots Var(X_n)\right) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots \sigma^2) = \frac{\sigma^2}{n}$$

($Var(X_1) = \sigma^2$, $Var(X_2) = \sigma^2$, $Var(X_n) = \sigma^2$)



T distribution $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

- center 0
- symmetric
- bell shape
- big tail on each side.
- degrees of freedom $n-1$

$n > 30$

HT for one sample mean (one sample t test)

one population

① $H_0: \mu = \mu_0$ (μ: $\bar{x}$)

$H_a: \mu \neq \mu_0$

$\mu > \mu_0$

$\mu < \mu_0$

② Check CLT

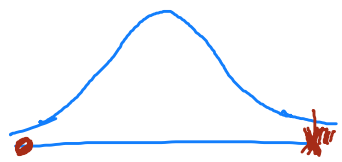
* independent data
* normality {1, 2, 3}

③ $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ $df = n-1$

④ p-value or rejection region

Ⓘ Conclusion:

Reject H_0 with rare observation



Confidence Interval for μ

$$\bar{x} \pm t^*(df=n-1) \cdot \frac{s}{\sqrt{n}}$$

$H_0: \mu = 93.3$ μ (95.3, 99.3) 95% C.I. → $\alpha = 5\%$

μ_0 ?

99% C.I. $\alpha = 1\%$

② HT for paired data (dependent sample)

$H_0: \mu_{diff} = 0$

$H_a: \mu_{diff} = 0$ one sample t test

$\mu_{diff} < 0$

$\mu_{diff} > 0$

③ HT for two population mean

(1) $H_0: \mu_1 = \mu_2$ $\mu_s = \mu_{ns}$

$H_a: \mu_1 = \mu_2$ $\mu_s \neq \mu_{ns}$

(2) CLT $\bar{X}_1 - \bar{X}_2$ point est for $\mu_1 - \mu_2$

✓ CLT $\bar{X}_1$
- independent data
- $n \geq 30$ or $n < 30$ Normal.

✓ CLT $\bar{X}_2$
- inde
- $n \geq 30$ or $n < 30$ Normal.

$\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1}{\sqrt{n_1}})$

$\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2}{\sqrt{n_2}})$

✓ independent between sample 1 and sample 2

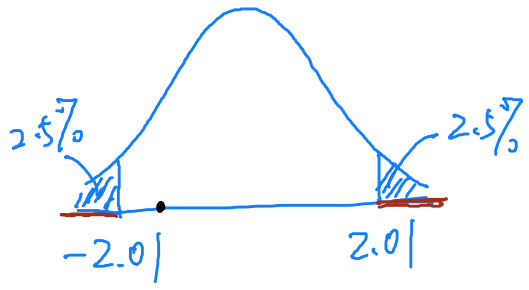
$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}})$

(3) $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(6.78 - 7.18) - 0}{\sqrt{\frac{1.43^2}{50} + \frac{1.6^2}{100}}} = \boxed{-1.55}$

$df = \min(n_1 - 1, n_2 - 1) = 50 - 1 = 49$

↓ smaller

④ Rejection Region ($\alpha = 5\%$)

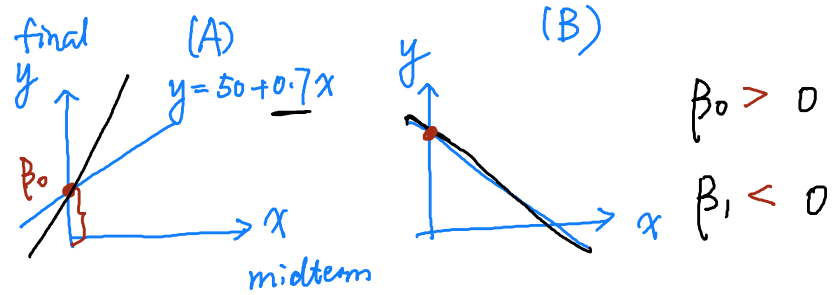


⑤ Conclusion: With test statistic −1.55 not in the rejection region of 5% significance level, we do not have strong evidence to claim the newborns from mothers who smoke have different average birth weights than those from mothers who don't smoke.

Chapter 8 Linear Regression

Simple Linear Regression mode

$$y = \beta_0 + \beta_1 x + \varepsilon$$



x: predictor / explanatory variable / independent variable

y: outcome / response variable / dependent variable

β_0: intercept (y value when x is zero)

β_1: slope (change in y for a unit change in x)

Estimated Model / Fitted Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \hat{y} = b_0 + b_1 x$$

Residuals: around zero, no special pattern

Avoid giving prediction outside available data (x) range.

Correlation Coefficient

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} \right)$$

only measures linear relationship!

$R \in [-1, 1]$

if $|R|$ is close to 1 → strong linear relationship

0 → little to no linear relationship.

Final Review

Bernoulli (p)

Binomial (n, p)

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Normal (μ, σ)

$$z = \frac{x - \mu}{\sigma} \rightarrow \text{standard normal } (0, 1)$$

p or μ ?

one p "rate" percentage / proportion "p_0"

$p_1 - p_2$ " " two independent groups

one μ time / money / amount of rain ... numerical data "μ_0"

$\mu_1 - \mu_2$ " " two independent groups

μ_{diff} difference paired data. "0"

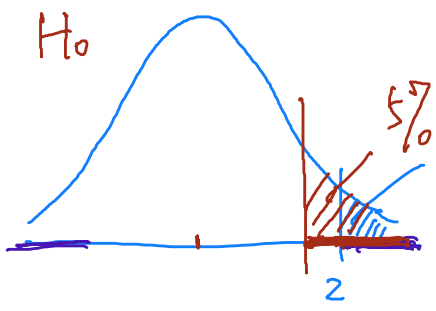
Type I or Type II error

↑

(significance level)

$H_0: p = 20\%$ Youtube music subscription

$H_a: p > 20\%$ $\hat{p} = 40\%$



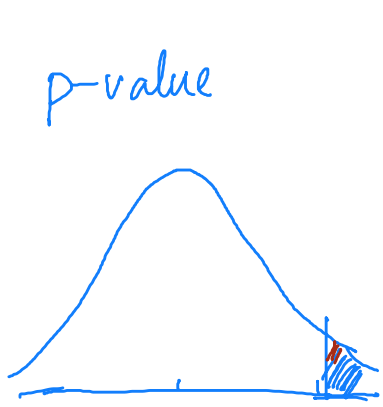
p-value = 0.025

$\alpha = 5\%$

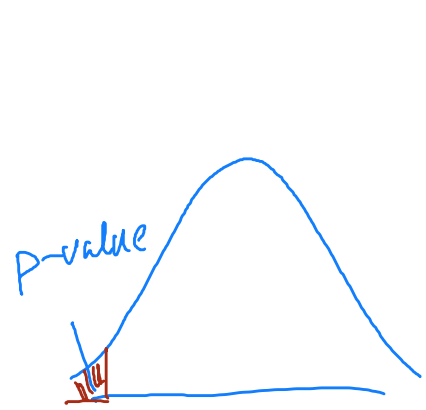
Reject H_0. — Type I H_0 true, reject

Type II H_0 false fail to reject H_0

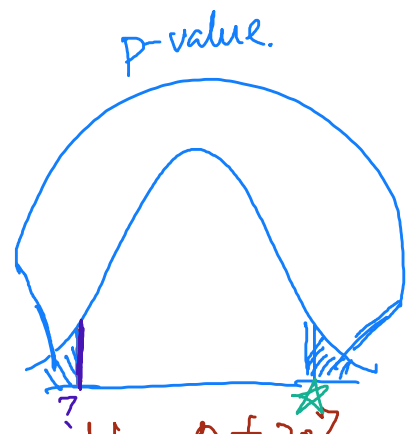
p-value



$H_a: p > 20\%$



$H_a: p < 20\%$



$H_a: p \neq 20\%$

p-value: how often to observe data equally or more extreme than sample statistic.