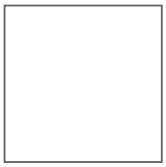
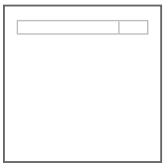

Statistics

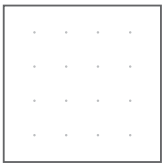
INDEX

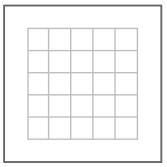
TEMPLATES *Copy and paste in the section of your choice.*


Blank


Blank + Title

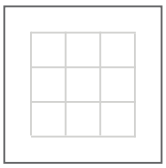

Lines


Dots

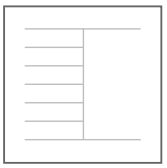

Grid


Wide Lines

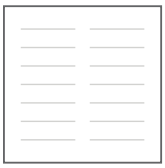

Wide Dots

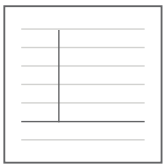

Wide Grid

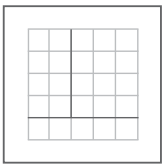

Top Blank
Bottom Lined

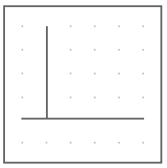

Left Lined
Right Blank

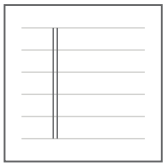

1 Column Lined

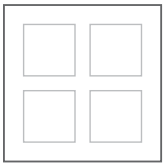

2 Column Lined

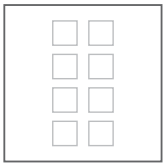

Cornell

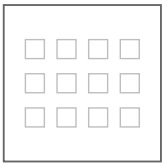

Cornell Grid

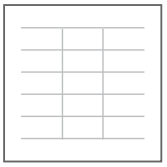

Cornell Dot

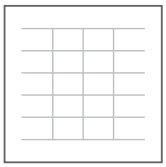

Legal

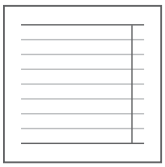

Boxes x 4

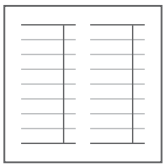

Boxes x 8

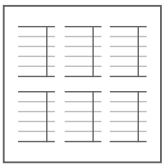

Boxes x 12

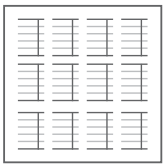

3 Column Table


4 Column Table


Lists x 1


Lists x 2

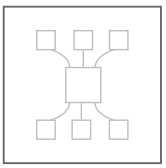

Lists x 6


Lists x 12

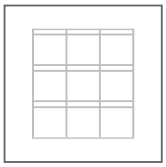

Isometric Grid 1

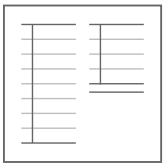

Isometric Grid 2

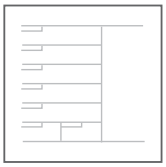

Music Paper

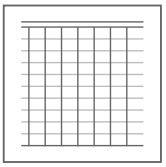

Mind Map

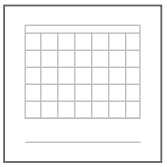

Recipe

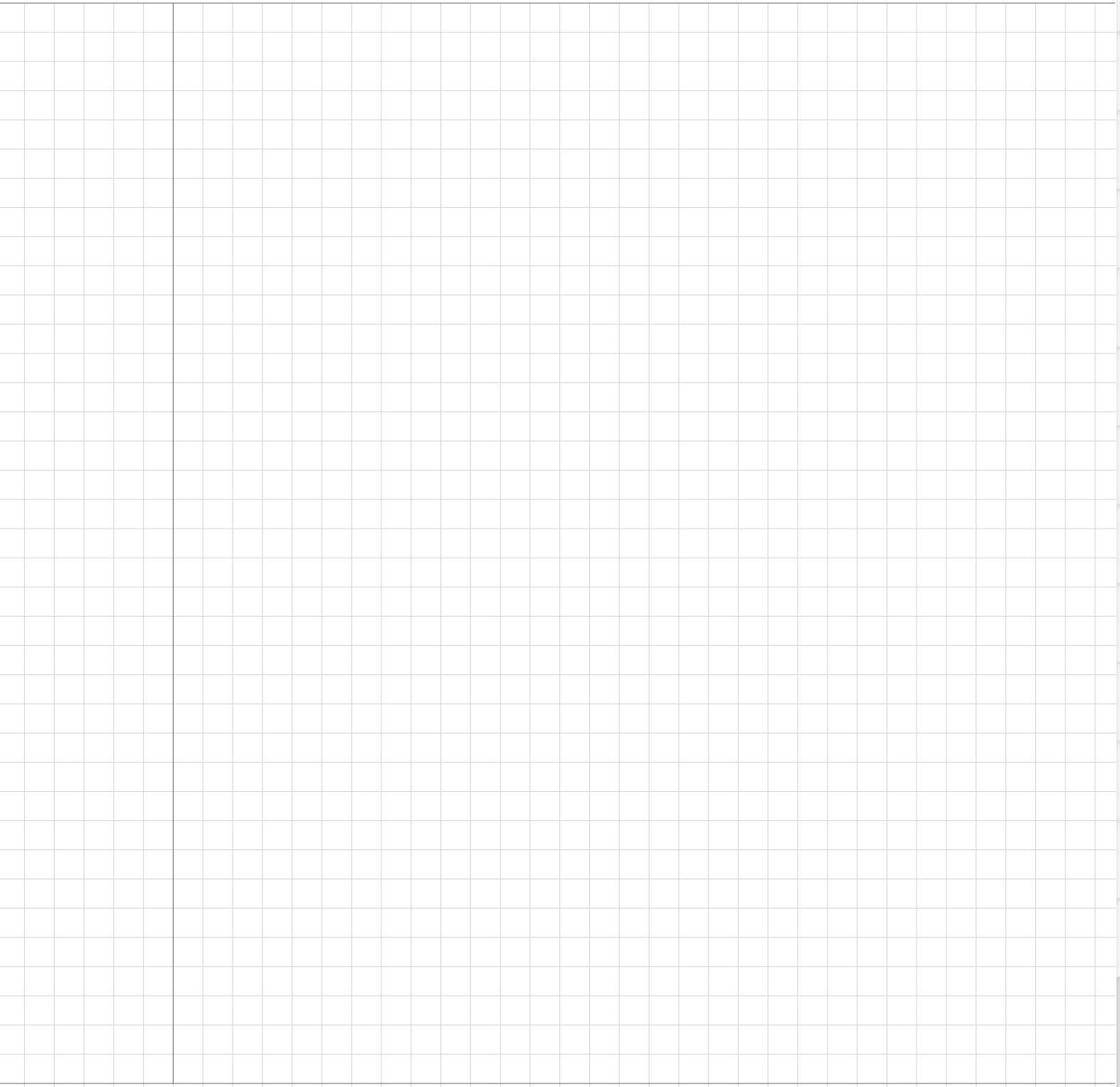

12 Month Plan

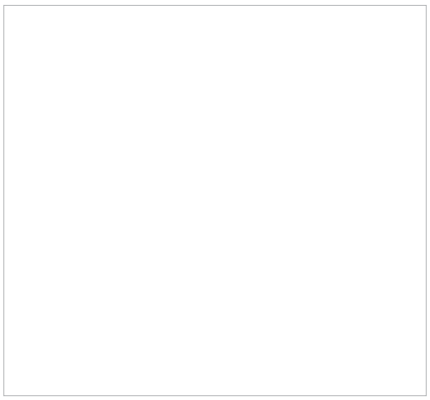
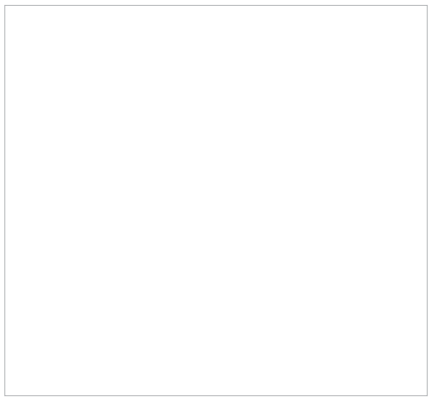
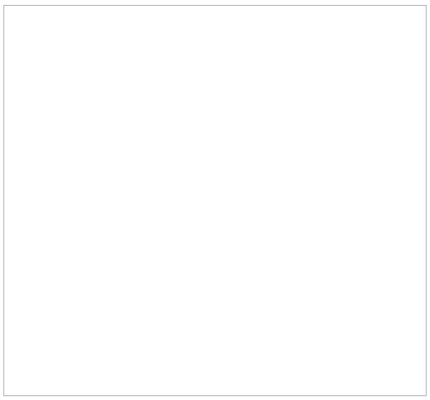

Daily Plan

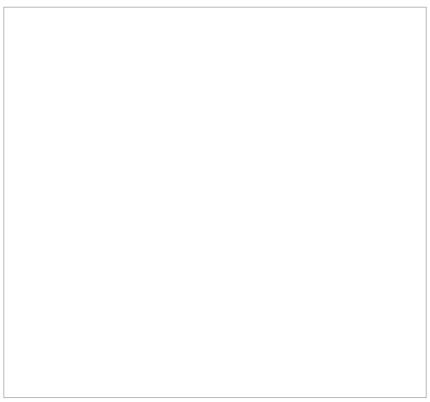
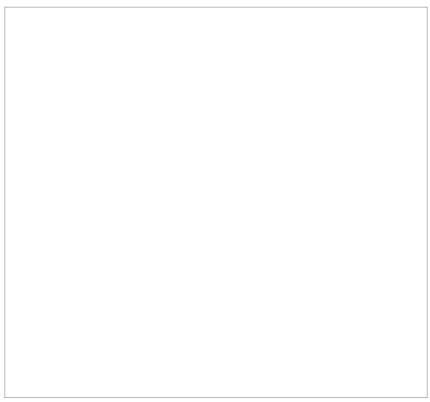
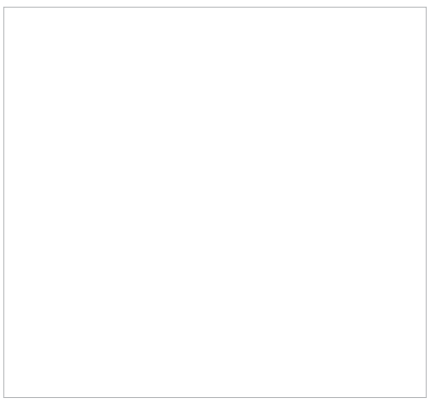

Weekly Plan


Weekly Plan 2


Monthly Plan



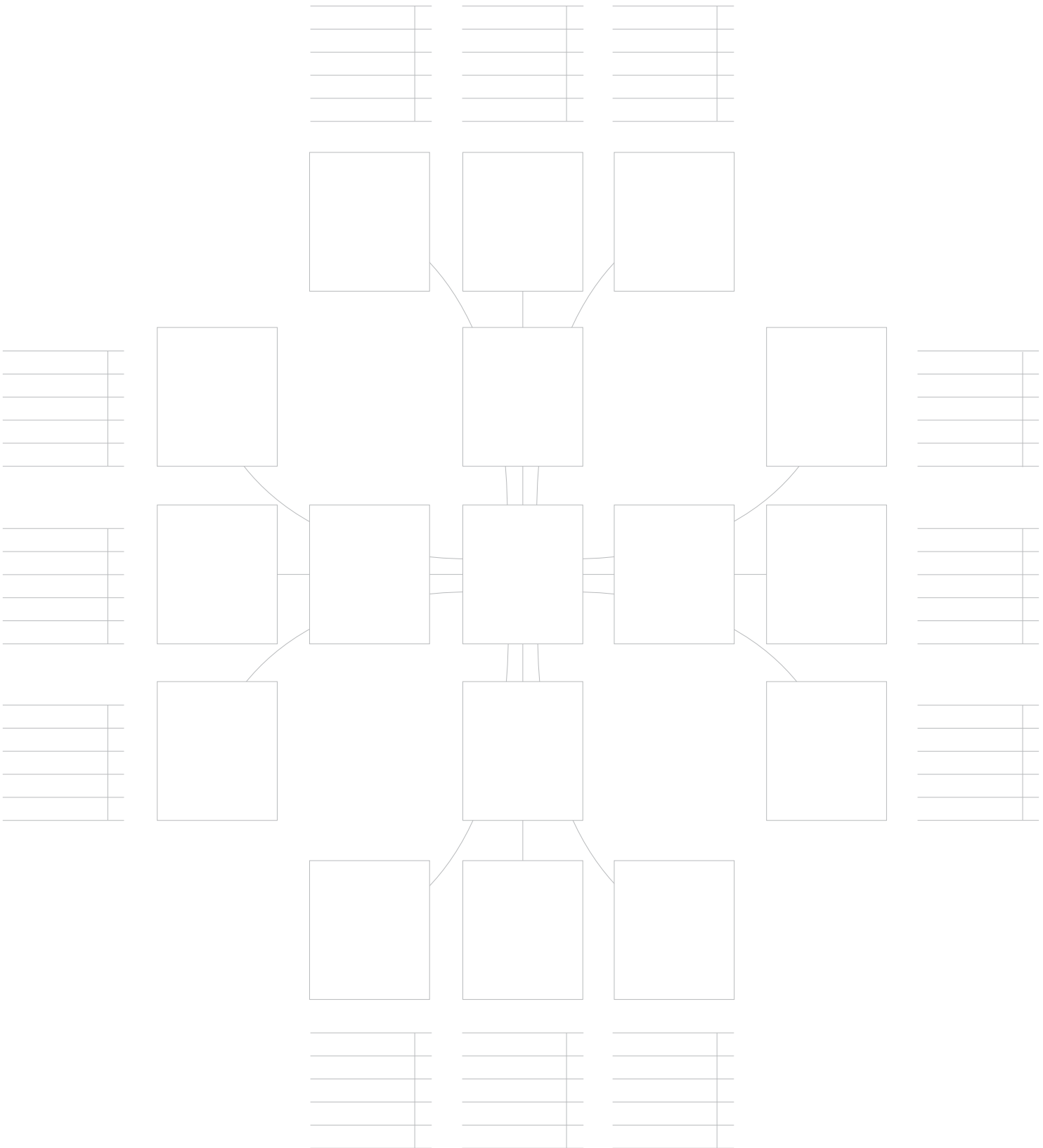




≡
+
1
2
3
4
5
6
7
8
9
10
11
12

≡
+
1
2
3
4
5
6
7
8
9
10
11
12

≡
+
1
2
3
4
5
6
7
8
9
10
11
12



RECIPE FOR **SERVES**

[Photo Here]

PREP TIME	COOK TIME	TOTAL TIME

INGREDIENTS

DIRECTIONS

NOTES

1
2
3
4
5
6
7
8
9
10
11
12

	SCHEDULE
6:00	
7:00	
8:00	
9:00	
10:00	
11:00	
12:00	
1:00	
2:00	
3:00	
4:00	
5:00	
6:00	
7:00	
8:00	
9:00	
10:00	
11:00	

TASKS

NOTES

6:00							
7:00							
8:00							
9:00							
10:00							
11:00							
12:00							
1:00							
2:00							
3:00							
4:00							
5:00							
6:00							
7:00							
8:00							
9:00							
10:00							
11:00							

1 Description

1. Description

Measure Location

Median (50%) Mode (most freq.)

$mid = Q_2$

Percentiles (P_i)

$$i = |X| \cdot p\% \quad \begin{cases} \text{if } i \in \mathbb{Z}, & P_i = \frac{1}{2}(P_i + P_{i+1}) \\ \text{if } i \notin \mathbb{Z}, & P_i = P_i \end{cases}$$

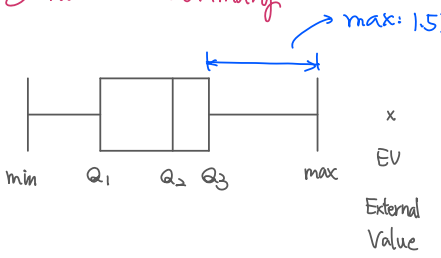
Why ? P th percentiles is mean "How many you win in 100 item"? and you wouldn't beat yourself, and the data is decrease.

Quartiles ($Q_{1..3}$, IQR)

$Q_i = P_{25 \cdot i} \quad i = 1, 2, 3$

$IQR = Q_3 - Q_1$

5 number Summary



Mean ($\bar{x}$)

$$\bar{x} = \frac{\sum x_i}{|X|}$$

1. Sum of deviation is "0" : $\sum (x_i - \bar{x}) = 0$
2. Balance point : center of mass
3. $\min \arg \left(\sum (x_i - a)^2 \right) = \bar{x}$
4. if $y_i = \alpha x_i + \beta \rightarrow \bar{y} = \alpha \bar{x} + \beta$

Measure Variability

Sample Variance & Standard Deviation

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

$$= \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

<tips>

$$\sum \bar{x} = \sum x_i = n\bar{x}$$

1. if $y = \alpha x + \beta \rightarrow S_y^2 = \alpha^2 S_x^2$

<pf>

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

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

$$= \frac{1}{n-1}\left(\sum x_i^2 - 2\sum x_i\bar{x} + \sum \bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\sum x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right)$$ ∎

Coefficient Variation

$$CV = \frac{S}{\bar{x}} \cdot 100\%$$ → Eliminate the value effect

eg. $\bar{x} = 10$, $S_x = 30$ $\Rightarrow CV_x = 300\%$ 👑

$\bar{y} = 1000$, $S_y = 300$ $\Rightarrow CV_y = 33.3\%$

Sample Covariance (with value effect)

$$Cov_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) \Rightarrow \frac{1}{n-1}\sum x_i y_i - n\bar{x}\bar{y}$$

Sample Correlation Coefficient ($-1 < r < 1$)

$$r = \frac{Cov_{xy}}{S_x S_y}$$ → Eliminate the value effect

<why>

$X = [x_1, \cdots x_n]^T$

the Euclidean Norm: $\|x\|_2 = \sum x_i^2$

and the SD = $\sqrt{E((x-\bar{x})^2)}$

and $\cos\theta_{xy} = \frac{\langle x, y\rangle}{\|x\|_2 \|y\|_2}$

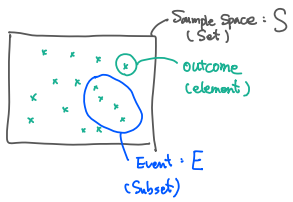
$$\Rightarrow r_{xy} = \frac{E(x-\mu_x)(y-\mu_y)}{\sqrt{E((x-\mu_x)^2)\,E((y-\mu_y)^2)}} = \frac{\langle x-\mu_x, y-\mu_y\rangle}{\|x-\mu_x\|\,\|y-\mu_y\|}$$

2 | Probability

2. Probability

Intro

Some Terms



Assigning Prob. to Event

$$P(E) = \frac{\#E}{\#S}$$ sample space

Rules of Prob.

1. if $A \in$ Event in an Exp. $\rightarrow A^c = \{x : x \notin A\}$

2. if $AB = \emptyset$ $\rightarrow$ A, B are Mutually Exclusive (Disjoint)
   - $P(A \cup B) = P(A) + P(B)$

Conditional Prob.

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$ (numerator: new event; denominator: new Sample Space)

< Dependent & Independent >

if $P(A|B) = P(A)$ $\rightarrow$ A, B are indep.
- $P(A \cap B) = P(A) \cdot P(B)$

Multiplication Rule

1. $P(B|A) = \frac{P(AB)}{P(A)} \Rightarrow P(AB) = P(B|A) \cdot P(A)$
2. $P(C|BA) = \frac{P(CBA)}{P(BA)} \Rightarrow P(ABC) = P(C|BA) \cdot P(B|A) \cdot P(A)$

2. $P(DCBA) = P(D|CBA) \cdot P(C|BA) \cdot P(B|A) \cdot P(A)$

Bayes' Theorem

if $A_1 \cup A_2 = S \Rightarrow A_1 A_2 = \emptyset$ (A_1 | A_2) (Mut. Exc.)

When A_1, $P(B)$ + when A_2, $P(B) \Rightarrow B$

$$P(A_1|B) = \frac{P(A_1)\cdot P(B|A_1)}{P(A_1)\cdot P(B|A_1) + P(A_2)\cdot P(B|A_2)}$$

$$P(B) = P(A_1)\cdot P(B|A_1) + P(A_2)\cdot P(B|A_2)$$

$$P(A|B)\cdot P(B) = P(A)\cdot P(B|A)$$

Bayes' Theorem

$P(A_1)\cdot P(A_2)$

- Often we begin probability analysis with initial or prior probabilities.
- Then, from a sample, special report, or a product test we obtain some additional information.
- Given this information, we calculate revised or posterior probabilities. $P(A_1|B)$
- Bayes' theorem provides the means for revising the prior probabilities.

Define B

Like A_1, A_2 tells How many % in B

Prior Probabilities → New Information → Application of Bayes' Theorem → Posterior Probabilities

By prompt.

< tips >

$P(A_1)\cdot P(A_2)$: prior prob. (Before)

$P(A_1|B)$: posterior prob. (After)

eg.

70% of the proposed zoning change proposals were approved

20% approve., but received Negative recommendations

10% reject, but received Positive

Q: P(reject)

step 1) $P(A_1)\cdot P(A_2)$: Prior

let A_1 : approve, A_2 : reject

70% = $P(A_1)$, 30% $P(A_2)$

$P(A_1)\cdot P(A_2) = 0.21$

step 2) New Info.:

let B = received "Negative" recommendation

step 3) Bayes' Thm

20% = $P(B|A_1)$

1-10% = $P(B|A_2)$

step 4) Posterior $P(A_1|B)$, $P(A_2|B)$

$$P(A_1|B) = \frac{P(A_1)\cdot P(B|A_1)}{P(B)} = \frac{P(A_1)\cdot P(B|A_1)}{P(A_1)\cdot P(B|A_1) + P(A_2)\cdot P(B|A_2)} = \frac{.7 \cdot .2}{.7\cdot.2 + .3\cdot.9}$$

< Tabular Approach >

Event A_i	Prior $P(A_i)$	× Condition. $P(B|A_i)$	= Joint $P(A_i \cap B)$	Posterior $P(A_i|B)$
A_1	.7	.2	.14	$\frac{P(A\cdot B)}{P(B)}$
A_2	.3	.9	.27	

$P(B) = P(A_1 B) + P(A_2 B)$

$P(B) = .41$

3 | Discrete Random Variable

3. Discrete Random Variable

Random Variable

Random Variable

a value determined by the outcome of an experiment.

Denoted : $P(X=x)$, P is Prob. func.

eg. toss a coin 3 times

let X be the number of head

then X is a RV

$P(X=0) = \frac{1}{8}$

$P(X=1) = \frac{3}{8}$

$P(X=2) = \frac{3}{8}$

$P(X=3) = \frac{1}{8}$

$\Rightarrow P(X=x) = \frac{1}{8}\binom{3}{x}$, $x=0,1,2,3$

Probability Distribution

Expectation (μ)

$E(X) = EX = \sum x \cdot P(x)$ Expectation = Expected Value = Mean

Variance & Standard Deviation (σ^2, σ)

$\sigma^2 = Var(X) = \sum (x-\mu)^2 \cdot P(x) = EX^2 - \mu^2$

< Some Properties >

1. if $y = \alpha x + \beta$, x,y are RV

   → $Ey = \alpha \cdot EX + \beta$

   → $Var\, y = \alpha^2 \cdot Var\, X$

2. Standardization :

   → $X_{STD} = \frac{X-\mu}{\sigma}$

3. $E(X+y) = EX + Ey$

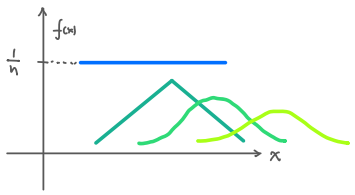
   $Var(X+y) = Var(X) + Var(y)$ ↗ if X, y are indep.

Discrete RV

Uniform Dist. $[X \sim Uni(1, n)]$

$$P(X) = \begin{cases} \frac{1}{n} & \text{for } X = 1, 2, \cdots n \\ 0 & \text{o.w.} \end{cases}$$

($\frac{1}{n}$ is a const.)



eg. Toss a coin, P(Head) = p

X = # Head in 1 toss

$\Rightarrow X \sim Uni(1, p)$ $\quad \begin{cases} \mu_X = \frac{1}{n} \cdot \sum x = 1 \cdot p = p \\ \sigma_X^2 = \frac{1}{n} \cdot (\sum x^2 - \mu^2) = \frac{1}{1} \cdot (\frac{1}{1} x^2 \cdot p - p^2) \\ \qquad = p - p^2 = p(1-p) \end{cases}$

<properties>

$\mu = n \cdot p \Rightarrow \frac{1}{n} \cdot \sum x$ (P(x))

$\sigma^2 = n \cdot pq \Rightarrow \frac{1}{n} \cdot \sum (x - \mu)^2$

$= \frac{1}{n} (E(x^2) - \mu^2)$

eg. Toss a coin 4 times, P(Head) = p

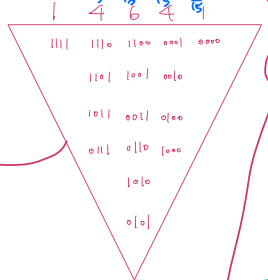
X = # Head in 4 tosses

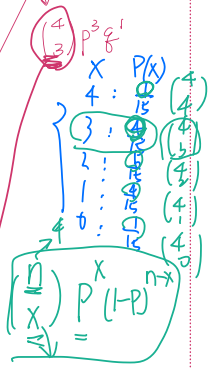
$\Rightarrow X \sim Uni(4, p)$ $\quad \begin{cases} \mu_X = 4p \\ \sigma_X^2 = 4pq \end{cases}$

But ! if I ask P(i Head in 4 tosses)

P(4 Head in 4 tosses) = p^4

P(3 Heads in 4 tosses) = $p^3 \cdot q^1$???





Binomial Dist $[X \sim B(n, p)]$

$$P(X) = \binom{n}{x} p^x q^{n-x}, \quad X = 0, 1, \cdots n$$

"1 Object do M Times"

and only 2 types outcome

and ask occur "How Many Times"

<properties>

$\mu = n \cdot p$

$\sigma^2 = n \cdot pq$

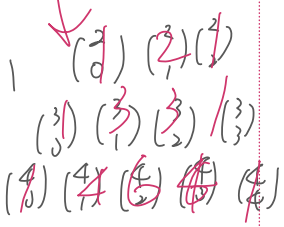
1. Bernoulli Dist $[X \sim Ber(p) = B(1, p)]$
2. let $X \sim B(n_x, p)$, $Y \sim B(n_y, p)$

$\longrightarrow X + Y \sim B(n_x + n_y, p)$

Cumulative Dist. func.

$$P(\alpha \leq X \leq \beta) = \sum_{i=\alpha}^{\beta} P(X_i)$$

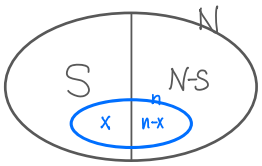
1 :
2 : 1 2 1
3 : 1 3 3 1
4 : 1 4 6 4 1
5 : 1 5 10 10 5 1



Hypergeomatric Dist. [X ~ H( )]

$$P(X) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

Obj 1 · Obj 2 · Total



"2 Objects (types) do n Time"

and only 2 types outcome

and ask occur "How Many Times"

eg. 10 red, 20 white balls, draw 5 balls at once.

X = # red balls

$$P(X) = \frac{\binom{10}{x}\binom{20}{5-x}}{\binom{30}{5}}$$

Red · White · Total

< properties >

$\mu = n \cdot p$

$\sigma^2 = C \cdot npq$ (C: correction coef.)

$p = \frac{S}{N}$ ⇒ prob. of success

$C = \frac{N-n}{N-1}$ ⇒ C=1 as n=1 or N→∞

Geometric Dist. (X ~ G(p))

$$P(X) = pq^{x-1}$$

Until Event Happen

< properties >

$\mu = \frac{1}{p}$

$\sigma^2 = \frac{q}{p^2}$

Poisson Dist. (X ~ Poi(λ))

$$P(X) = \lambda^x \frac{e^{-\lambda}}{x!}$$

Occurrence over a Specified Interval

eg. ave # accident occurring weekly on a highway is 1.2

Q: P(at least 1 accident this week)

Using B(n, p)

⇒ X ~ B(n>0, p=1.2)

of Time or Space

<properties>

$\mu = \lambda$

$\sigma^2 = \lambda$

1. $\exists\ X \sim B(n,p)$, if $n \to \infty$

$\rightarrow\ X \sim Poi(\lambda)$, $\lambda = np$

2.

very similar with Exponential Dist.

but Poi. Dist. is Discreate

$X = \sum_{n=1}^{\infty} np$

Using $Poi(\lambda)$ to approach

$\Rightarrow X \sim Poi(\lambda=1.2)$

$X = \lambda^x \frac{e^{-\lambda}}{x!}$

$P(X>0) = P(All) - P(X=0)$

$= 1 - \lambda^0 \frac{e^{-\lambda}}{0!}$

$= 1 - e^{-1.2}$ //

4 Continuous Random Variables

4. Continuous Random Variables

Continuous RV

Prob. Density Function ($1 = \int_{min}^{max} f(x)\,dx$)

$X \sim$ CRV

The pdf of X :

$$P(\alpha < X < \beta) = \int_a^b f(x)\,dx$$

Expectation

$$EX = \int_a^b x \cdot f(x)\,dx$$

Variation

$$\mathrm{Var}X = EX^2 - E^2X \quad \mu$$

Uniform Dist. [$X \sim Uni(\alpha, \beta)$]

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < x < \beta \\ 0, & \text{o.w.} \end{cases}$$

< properties >

$$\mu = \frac{\beta+\alpha}{2} \quad \text{(pf 1)}$$

$$\sigma^2 = \frac{(\beta-\alpha)^2}{12} \quad \text{(pf 2)}$$

(pf 1)

$$\mu = \int_\alpha^\beta x \cdot f(x)\,dx$$
$$= \frac{1}{2} x^2 \cdot \frac{1}{\beta-\alpha}\Big|_\alpha^\beta$$
$$= \frac{1}{2}(\beta^2-\alpha^2)\frac{1}{\beta-\alpha}$$
$$= \frac{1}{2}(\beta+\alpha)\frac{(\beta-\alpha)}{\beta-\alpha}$$
$$= \frac{1}{2}(\beta+\alpha) \quad \#$$

(pf 2)

$$\sigma^2 = EX^2 - \mu^2$$
$$EX^2 = \int_\alpha^\beta x^2 \cdot f(x)\,dx$$
$$= \frac{1}{3} x^3 \frac{1}{\beta-\alpha}\Big|_\alpha^\beta \qquad \beta^3-\alpha^3 = (\beta-\alpha)(\beta^2+\beta\alpha+\alpha^2)$$
$$= \frac{1}{3}\frac{1}{\beta-\alpha}(\beta^3-\alpha^3)$$
$$= \frac{1}{3}(\beta^2+\beta\alpha+\alpha^2)$$
$$\therefore \sigma^2 = \frac{1}{3}(\beta^2+\beta\alpha+\alpha^2) - \frac{1}{2}(\beta+\alpha)$$
$$= \frac{1}{3}(\beta+\alpha)^2 - \beta\alpha - \frac{1}{2}(\beta+\alpha)$$
$$= \frac{1}{12}(\beta-\alpha)^2 \quad \#$$

Trigonometry Dist. [$X \sim Tri(\,)$]

$$\int_{min}^{max} f(x)\,dx = \triangle = 1$$

Normal Dist. [$X \sim N(\mu, \sigma^2)$]

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

standardization (Z - score)

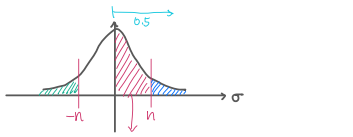
$$Z = \frac{x-\mu}{\sigma}$$

<properties>

$\mu = \mu$

$\sigma^2 = \sigma^2$

1. $Z_n = -Z_{1-n}$, $n \in (0,1)$ (reflection)



$P(0 < X < n)$ or $0.5 - P(n < X)$ or

2. let $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$

(1) $X + Y \sim N(\mu_x + \mu_y , \sigma_x^2 + \sigma_y^2)$

(2) $\alpha X + \beta Y \sim N(\alpha\mu_x + \beta\mu_y , \alpha^2\sigma_x^2 + \beta^2\sigma_y^2)$

(3) $X - Y \sim N(\mu_x - \mu_y , \sigma_x^2 + \sigma_y^2)$

percentile of $N(\lambda, \sigma)$

$Z_{0.025} = 1.96$

$Z_{0.05} = 1.645$

$Z_{0.1} = 1.28$

Exponential Dist. [$X \sim Exp(\lambda)$]

$$f(x) = \begin{cases} \lambda e^{-\lambda x} , & x \geq 0 \\ 0 , & x < 0 \end{cases}$$

<properties>

$\mu = \frac{1}{\lambda}$

$\sigma^2 = \frac{1}{\lambda^2}$

1. cdf: $F(x) = P(X \leq x) = 1 - e^{-\lambda x}$, $x > 0$

$\Rightarrow P(X > t) = e^{-\lambda t}$, $t > 0$

2. $\begin{cases} Poi. : P(\text{\# events in a time interval}) \\ Exp. : P(\text{time till 1st event}) \text{ or} \\ \quad\quad P(\text{time between events}) \end{cases}$

Relation to Poisson Dist.

If # arrivals in time interval $(0,t)$ follow a $Poi(\lambda \cdot t)$, λ is the ave # arrivals per time

$\Rightarrow \begin{cases} N_t : \text{\# arrival in time interval } (0,t) \\ X : \text{the time, to the 1st arrival} \end{cases}$

$P(X < t) = P(N_t > 0) = 1 - e^{-\lambda t}$

$P(X > t) = P(N_t = 0) = e^{-\lambda t}$

5 Sampling Distribution

5. Sampling Distribution

Simple Random Sampling

Finite Population

- Sampling with replacement
- Sampling without replacement : most often

Infinite Population

- Select from the same popu.
- Each ele. in indep.

IID (Independent & Identical Dist.)

- Drawn from the same prob. Dist.
- They constitute a sample from the Dist.

Dist of the sample mean ($\bar{X}$)

Finite popu. ($N < 50n$)

$$E\bar{X} = \mu$$

$$Var\bar{X} = \frac{N-n}{N-1}\frac{\sigma^2}{n}, \quad \sigma^2 = Var X$$

Infinite popu. ($N \geq 50n$)

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$E\bar{X} = \mu$$

$$Var\bar{X} = \frac{\sigma^2}{n}$$

CLT : Central Limit Theorem

CLT

Let $\{X_1, \cdots X_n\}$ be a sample from a (infinite) popu.

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

or

$$\sum X_i \approx N(n\mu, n\sigma^2)$$

for finite popu.

$$\bar{X} \approx N\left(\mu, C\frac{\sigma^2}{n}\right), \quad C = \frac{N-n}{N-1}$$

or

$$\sum X_i \approx N(n\mu, C\sigma^2)$$

eg. 10^4 automobile policyholders, if expected yearly claim per placeholder is \$240, SD of \$800.

Q: P(Total yearly claim exceeds \$2.8M)

X_i : Yearly Claim of policyholder i, $i = 1, \cdots 10^4$

$Y = \sum X_i$: Total yearly claim of all policyholder

CLT ↳ $Y \approx N(10^4 \cdot \mu, 10^4 \cdot \sigma^2)$

$$P(Y > 2.8M) = P\left(\frac{Y-\mu_Y}{\sigma_Y} > \frac{2.8M-\mu_Y}{\sigma_Y}\right), \text{ where } \begin{cases}\mu_Y = 10^4\mu \\ \sigma_Y = \sqrt{10^4}\sigma\end{cases}$$

$$\approx P(Z > 2.5) = 0.0062$$

Application of CLT

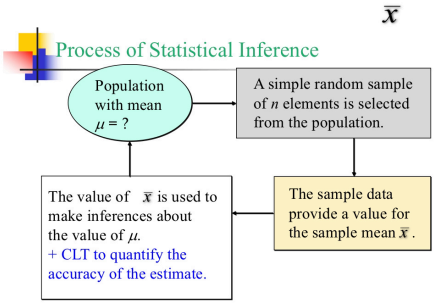
$\bar{X}$ is the sample mean, from popu with μ, σ^2

$$P(\bar{X} < \alpha) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{\alpha-\mu}{\sigma/\sqrt{n}}\right)$$

Standard Normalization

$$\approx P\left(Z \le \frac{\alpha-\mu}{\sigma/\sqrt{n}}\right)$$

$$E\bar{X} = \mu, \quad Var\bar{X} = \frac{\sigma^2}{n}, \quad SD\bar{X} = \frac{\sigma}{\sqrt{n}}$$



Sampling proportions from a Popu.

Sample Proportion

- When sampling, there are X successes from $Ber(n, p)$, $\hat{p} = \frac{X}{n}$

* $Ber(p) \Rightarrow B(1, p)$

$\mu = 1 \cdot p$, $\sigma^2 = 1 \cdot p \cdot q$

- It's a RV.

Finite vs. Infinite Popu.

When n/N is small, → Infinite

$X_1, \cdots X_n$ are approximately independent Ber(p)

$\bar{X}$ is often denoted by $\hat{p}$

When n/N is large, → Finite

$X_1, \cdots X_n$ are not independent

<properties>

(1) When n/N is Large,

$$E\hat{p} = p \quad , \quad Var\,\hat{p} \approx C \cdot \frac{1}{n}pq$$

$$\frac{\hat{p}-p}{\sqrt{C\frac{1}{n}pq}} \approx N(0,1)$$

(2) When n/N is Small,

$$E\hat{p} = p \quad , \quad Var\,\hat{p} = \frac{1}{n}pq$$

$$\frac{\hat{p}-p}{\sqrt{\frac{1}{n}pq}} \approx N(0,1)$$

(3) If $Y \sim B(n,p)$

$\longrightarrow$ $Y = X_1 + \cdots X_n$, where X_i are i.i.d. Ber(p) R.V.s

CLT $\longrightarrow$ $Y \approx N(np,\ npq)$

$$P(Y \le k) = P\left(\frac{Y-np}{\sqrt{npq}} \le \frac{k-np}{\sqrt{npq}}\right)$$

($k-np \to \mu$, $\sqrt{npq} \to \sigma$)

$$\overset{CLT}{\Rightarrow} P\left(N(0,1) \le \frac{k-np}{\sqrt{npq}}\right)$$

* $X \sim B(n,p)$, $\mu = np$, $\sigma^2 = npq$

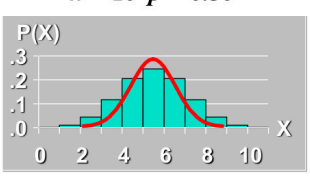
Let $Y \sim N(\mu, \sigma^2)$

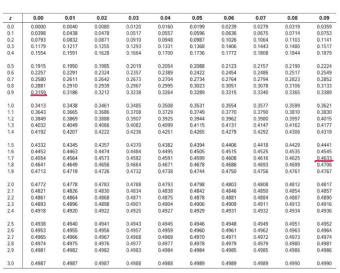
$$P(X \le x) \approx P(Y \le x + 0.5)$$

Normal Approximation of Binomial Distribution

1. Not all binomial tables exist
2. Use normal dist. to approximate
3. Requires large sample size
4. Need correction for continuity

$n = 10$ $p = 0.50$





Comparison

$X \sim B(5, \frac{1}{2})$

$P(X=1) = \binom{5}{1}\frac{1}{2}(\frac{1}{2})^4 = 0.156$

$P(X \le \frac{5}{2}) = P(X \le 2) = \sum_{i=1}^{2}\binom{5}{i}(\frac{1}{2})^5 = 0.5$

np $\sqrt{npq}$

$Y \sim N(\mu_x, \sigma_x) \Rightarrow \begin{cases} \mu_x = \frac{5}{2} \\ \sigma_x = \sqrt{\frac{5}{4}} \end{cases}$

$P(X=1) \approx P(1-0.5 \le y \le 1+0.5)$

$P(0.5 \le y \le 1.5) = P\left(\frac{0.5-\mu}{\sigma} \le y \le \frac{1.5-\mu}{\sigma}\right)$

$= P(-1.79 \le Z_y \le -0.9)$

$= P(0.9 \le Z_y \le 1.79)$

Discrete $= 0.4633 - 0.3159 = 0.15$

$P(X=\alpha) \approx P(\alpha-0.5 \le y \le \alpha+0.5)$

$P(X \le \frac{5}{2}) = P(X \le 2) = P(y \le \frac{5}{2})$ Same

Chi - Square Dist.

if $Z_i \sim N(\mu_i, \sigma_i^2)$ are indep. $i = 1, \cdots n$

$\rightarrow y = \sum (Z_i)^2 \sim \chi^2_n$ sample size - 1 = d.f.

"How many N-Dist. sum up."

< properties >

$Ey = n$ $\quad Var Z = 1$

(1) $\chi^2_n \Rightarrow \frac{\sum (X_i - \mu)^2}{\sigma^2}$

(2) $\chi^2_1 \Rightarrow \frac{n(X_i - \mu)^2}{\sigma^2} = \left(\frac{X_i - \mu}{\sigma/\sqrt{n}}\right)^2$

(3) $\chi^2_n = \chi^2_{n-1} + \chi^2_1$

$\sum (X_i - \mu)^2 = \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$

(4) if $X \sim \chi^2_n$, $y \sim \chi^2_m$

$\rightarrow X + y \sim \chi^2_{n+m}$

(5) let $X_1, \cdots X_n$ be sample from N-Dist.

popu. $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2}$ Sample

has a Chi-Square Dist. with d.f. $(n-1)$

(6) Consider $Z_i = \frac{X_i - \mu}{\sigma}$, $i = 1, \cdots n$

$\rightarrow \sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum (X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$

(7) if $X_1, \cdots X_n$ from $N(\mu, \sigma^2)$ popu.

$\rightarrow E(S^2) = \sigma^2$, S^2 is sample Variance

6 Estimation

6. Estimation

Conclusion

Goal: Get "Popu Mean"

How: By "Sample Mean($\bar{X}$)" to "Estimate"

Question: "How much" accuracy $\bar{X}$ is μ_{popu}

Tools: CLT

Transform the thinking: By "Interval"

We could say, you give me a k%,

I tell you the interval that μ_{popu} would in it with k% (probability) confidence.

k% = $1-\alpha$

Process

Sample Statistic

proportion

$\bar{x} \to np \to \hat{p}$

$\sigma_x^2 \to npq \to n\hat{p}\hat{q}$

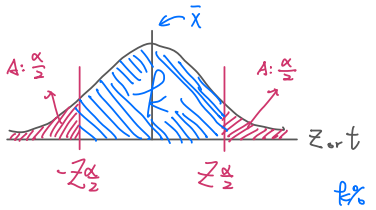
Sampling $\bar{x}, s^2$

Known σ_{popu}?

T: $Z_{\frac{\alpha}{2}} = \dfrac{\bar{x}-\mu}{\sigma_{popu}/\sqrt{n}}$

F: $t_{(n-1)}^{\frac{\alpha}{2}} = \dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}/\sqrt{n}}$

CLT



k%

$P(-Z_{\frac{\alpha}{2}} < SS < Z_{\frac{\alpha}{2}}) = 1-\alpha$

$P(-Z_{\frac{\alpha}{2}} < \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} < Z_{\frac{\alpha}{2}})$

$P(-Z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n} < \bar{x}-\mu < Z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n})$

$P(\bar{x} - Z_{\frac{\alpha}{2}}\sigma/\sqrt{n} < \mu < \bar{x} + Z_{\frac{\alpha}{2}}\sigma/\sqrt{n})$

$P(\mu \text{ in } \bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}) = k\%$

<eg>

Weights of RS of 8 people with, σ_{popu}: 22

121, 163, 144, 132,
188, 130, 128, 140

use the data to estimate popu. 95%

1. SS: $\begin{cases} \bar{x} = 145.5 \\ s = 21.27 \end{cases}$

2. Known σ_{popu}, $\frac{\alpha}{2} = 0.025$, $Z_{\frac{\alpha}{2}} = 1.96$

3. [CI: $\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}$]

$CI = 145.5 \pm 1.96 \cdot 22/\sqrt{8}$

2. If Unknown σ_{popu}

3. $CI = \bar{x} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \sigma/\sqrt{n}$

$CI = 1455 \pm t \cdot 21.27/\sqrt{8}$

<eg> RS #100, 64%

a) Entire popu. is Event: $E\bar{x} = n\hat{p} = 64$ ✗ $\bar{x} = \hat{p}$ s

b) Standard Error:

[Standard Error = Standard Deviation]

$S = \sqrt{\dfrac{n\hat{p}\hat{q}}{n}} = \sqrt{\dfrac{0.64 \cdot 0.36}{100}} = 0.048$ ✗

<eg> $s = 11.3$, $n = 81$, $\bar{x} = 74.6$, 90% CI, popu ~ N

(1) $\bar{x} = 74.6$, $s = 11.3$, $n = 81$ 0.10 $Z_{0.05}$

(2) popu ~ N-Dist, $\alpha = 0.1$, $Z_{\frac{\alpha}{2}} = 1.645$

(3) $CI = \bar{x} \pm Z_{\frac{\alpha}{2}} \cdot s/\sqrt{n}$

$= 74.6 \pm 1.645 \cdot 11.3/\sqrt{81}$

$= (72.53, 76.67)$

Point Estimator of Popu. Mean & Variance

Unbiased

$$E(\bar{X} = p) = p$$

Popu. "Mean"

$X_1, \cdots X_n$ from N-Dist, and μ is unknown

$\rightarrow E\bar{X} = \mu_{pop.}$

$$P(|\bar{x} - \mu| < 2\sigma/\sqrt{n}) \approx 0.95$$

<properties>

A bag has "N" balls, "M" are red,
a SRS of size n, let X = # red.

point estimator: $\hat{p} = \frac{X}{n}$

$B(n,p)$

also called "stand. err." $E\hat{p} = p$, $Var\hat{p} \approx \frac{pq}{n}$, $p = \frac{M}{N}$

$SD(\hat{p}) = \sqrt{\frac{pq}{n}}$, if $pq \le \frac{1}{4}$, $SD(\hat{p}) \le \frac{1}{2\sqrt{n}}$

eg. SRS n=50, 20 were in favor of proposal

(a) proportion of all students who are in favor

$E\bar{X} = p = \frac{20}{50} = 0.4$ *

(b) Stand. err. of this estimate.

$\left[SD(\hat{p}) = \frac{pq}{n} \right]$

$SD(\hat{p}) = \sqrt{\frac{0.4 \cdot 0.6}{50}} = 0.07$ *

now! How much "Confidence" you have that $SD(\hat{p})$

$$P\left(|\hat{p} - p| < 2\sqrt{\frac{pq}{n}}\right) \approx 0.95$$

$$\approx P\left(|\hat{p} - p| < 2\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

SO that $|p - 0.4| < 2.0693$ with conf. 95%

Popu. "Variance"

if $\sigma^2_{pop.}$ is Unknown, X sampled from popu.

$$S^2_{samp} = \frac{\Sigma (X - \bar{X})^2}{n-1}$$

if Known: $\mu_{popu.}$

$$\sigma^2_{popu} \approx \frac{\Sigma (X_i - \mu)^2}{n}$$

Interval Estimators of Mean & Variance

Popu. "Mean"

Interval Estimator :

an interval that is predicted to contain the params.

Margin of Err.

Interval Est. = Point Est. ± Margin of Err.

$M_{popu} = \bar{X} \pm$ Margin of Err.

Confidence Interval (CI)

An interval est. sates range within

a popu. params. prob. the specified prob.

is called (level of Conf. / Conf. Coef.)

Def : $Z \sim N(0,1)$ $\exists\ z_\alpha$ st. $P(Z \geq z_\alpha) = \alpha$ (z_α ← $z(\alpha)$)

- α : 0.1 ⇒ 90% CI ⇒ $z_{\frac{\alpha}{2}} = 1.645$
- α : 0.05 ⇒ 95% CI ⇒ $z_{\frac{\alpha}{2}} = 1.96$
- α : 0.01 ⇒ 99% CI ⇒ $z_{\frac{\alpha}{2}} = 2.58$

<pf> When n is large, by CLT

$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$

$1-\alpha = P(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\frac{\alpha}{2}})$

$= P(\bar{X} - z_{\frac{\alpha}{2}}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\sigma/\sqrt{n})$ ∎

<eg> sample $B \sim N(\mu, 3^2)$ from A

sample 10 times : 17, 21, 20, 15, 19, 23, 20, 21, 16, 19

95% CI of μ = ?

$\bar{X} = \frac{\sum x_i}{n} = 19.3$

$\bar{X} \pm 1.96\ \sigma/\sqrt{n}$

$\Rightarrow 19.3 \pm 1.96 \cdot 3/\sqrt{10} \Rightarrow 19.3 \pm 1.86$ ∎

CI of μ

a size n SRS from inf. prop. $N(\mu, \sigma^2)$

→ $1-\alpha$ CI of μ : $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ ← CI

if want make CI to 0.78

then need to increase How Much # sample

$CI = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

$0.78 = 1.96 \cdot \frac{3}{\sqrt{n}} \Rightarrow n = 56.8$ ∎

<properties>

Sometim we want make CI ≤ b

b is a specific value for limit CI.

Sample: $n \geq \left(\frac{2 z_{\frac{\alpha}{2}} \sigma}{b}\right)^2$

CI: $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

<eg> $S = 0.3$, wanted 90% certain that the est. of mean weight is correct within ±0.1 pound, How Much # sample ?

[CI: $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$]

90% ⇒ $\alpha = 0.1$ ⇒ $z_{\frac{\alpha}{2}} = 1.645$

$0.1 = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 1.645 \frac{0.3}{\sqrt{n}}$

$\therefore n = 24.35$ ∎

Small Sample Case

Popu is not N-Dist ?

Increase # sample + CLT ($\bar{X}$)

Popu $\sim N(?, \sigma^2)$

Interval-Est. + CI

Popu $\sim N(?, ?)$

Student t- Dist + CI

Student's t-Dist

when # sample is small

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

t- Dist has $\mu=0$,
but s.d. is depend on the # sample

<eg> # sample = 10, $\mu = 550$, $\sigma = 60$
95% CI est. ?

[95% CI $\Rightarrow \alpha = 0.05$]

[CI: $\bar{X} \pm t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$]

df = 10-1

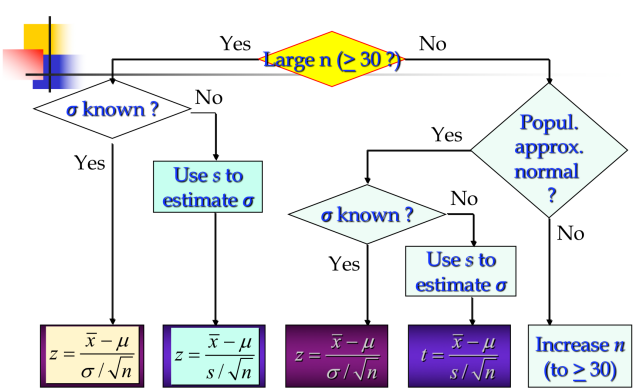
$550 \pm 2.262 \cdot \frac{60}{\sqrt{10}}$

Degree of Freedom (d.f.)

d.f. refers to the # indep. observations

$S^2_{samp.}$ has "n-1" d.f.

Summary of Interval Estimation Procedures for a Population Mean



CI of P (popu. proportion)

CI of P

sample from a popu. with prop. P

Suppose Sample proportion is $\hat{P}$

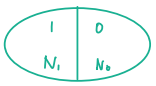
$$\sigma_{\hat{p}}^2 \approx \frac{\hat{p}\cdot\hat{q}}{n}, \quad \hat{p} = \frac{X}{n}$$

$$\hat{p} \approx N\left(p, \frac{\hat{p}\hat{q}}{n}\right)$$

<pf> $\mu_{\text{popn}} = n\hat{p} \Rightarrow \bar{X} = \frac{\mu}{n} = \hat{p}$

$$\sigma_{\text{popn}} = n\hat{p}\hat{q} \Rightarrow S = \frac{n\hat{p}\hat{q}}{n^2} = \frac{\hat{p}\hat{q}}{n}$$

popu. proportion



$N_1 + N_0 = N$

$p = \frac{N_1}{N}$

If N is large

→ $X \sim B(n,p)$

$\text{Var}\, X = npq$

$\text{Var}\, \frac{X}{n} = \frac{pq}{n}$

<eg> 220 registered voters, out of 500 contacted

95% CI Est. for the proportion ?

$\hat{p} = \frac{220}{500}$ $\left[\hat{p} \approx N\left(p, \frac{\hat{p}\hat{q}}{n}\right)\right]$

$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$ $\left[\hat{p} \pm z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$ 0.975 0.025

$0.44 \pm 1.96\sqrt{\frac{0.44\cdot 0.56}{500}}$ [95% CI $\Rightarrow \alpha = 0.05 \Rightarrow z_{\frac{\alpha}{2}} = 1.96$]

<eg> H: over $\frac{1}{2}$, S: $n = 920$, $\hat{p} = 0.52$

$\begin{cases} H_0: p \le 0.5 \\ H_1: p > 0.5 \end{cases}$

7 Testing Statistical Hypothesis

F. Testing Statistical Hypothesis

Conclusion

Goal : accept / reject assumption

How : By Sampling

Questions : when we accept / reject

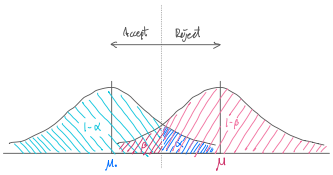
Tools : CLT, α, β, power

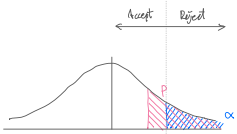
Type I error (α) = P(Rej. | correct) ← significant level

Type II error (β) = P(Acc. | incorrect)

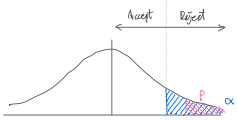
Power = $1-\beta$

p-value = $\bar{X}$ lead position





$p > \alpha$: Acc. H_0



$p < \alpha$: Rej. H_0

Process

Hypo	Select	Test Statistic	Formula Decision	Make Decision
H_0, H_1, μ_0	α	$TS = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	Critical Region	Accept / Reject

<eg> Find α, $H_0: \mu \in (998, 1002)$, $\mu_T = 1000$ → True value

SS: $n=36$, $\bar{x}=998$, $s=2$

[$\alpha = P(\text{Rej.} \mid \text{Corr.})$]

Why? μ_{pop} might be any value if in these area, then we do the type I error!

$$\alpha = 1 - P\left(\left|\frac{\bar{x}-\mu_T}{s/\sqrt{n}}\right| \le \frac{1002-\mu_T}{s/\sqrt{n}}\Big|_{\mu_T=1000}\right)$$

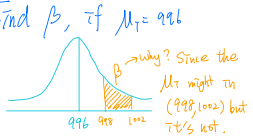
$$= 1 - P(|Z| \le 1)$$

$$= 2(1 - P(Z<1))$$

$$= 2(1-0.8413)$$

$$= 0.3174 \#$$

<eg> Find β, if $\mu_T = 996$



why? Since the μ_T might in (998, 1002) but it's not.

$$\beta = P\left(Z < \frac{1002-\mu_T}{s/\sqrt{n}}\right) - P\left(Z < \frac{998-\mu_T}{s/\sqrt{n}}\right)$$

$$= 0.4987 - 0.343 = 0.1574 \#$$

<eg> $H_0 = 250$, Sig. lev. $= 0.05$

TS: $n=81$, $\bar{x}=255$, $s=3$

Step1) $H_0: \mu = 250 = \mu_0$, $H_1: \mu \neq 250 = \mu_0$

Step2) $\alpha = 0.05$, $Z_{\frac{\alpha}{2}} = 1.96$

Step3) $TS = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{255-250}{3/9} = 15$

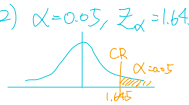
$15 > 1.96 \Rightarrow$ Reject !

<eg> $H_0: \mu \le 140$, Sign. level $= 0.05$

TS: $n=900$, $\bar{x}=146$, $s=48.2$

Step1) $H_0: \mu = 140 = \mu_0$, $H_1: \mu > 140 = \mu_0$

Step2) $\alpha = 0.05$, $Z_\alpha = 1.645$



Step3) $Z = \frac{146-140}{48.2/\sqrt{900}} = 3.726$

Step4) $Z = 3.726 > 1.645 = Z_\alpha$

Statistical Hypothesis

Hypothesis

a statement about the nature of a popu.

Hypo. Testing

By Sample Evidence to determine weather the hypo.'s approvability.

Null & Alternative Hypo.

- null hypo. H_0 is a

  Tentative Assumption about a popu. param.

- alt. hypo H_1 is a

  Opposite of what's stated in the H_0

<properties>

(1) The "$=$" always appears in H_0

(2) 3 main form

$H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$

$H_0: \mu \geq \mu_0$, $H_1: \mu < \mu_0$

$H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

Assume: the popu. mean age = 45 (H_0)

popu. → samp.: check μ_{samp}

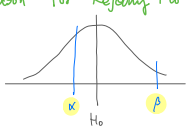
If $\mu_{samp} \approx 45$

the assume is correct!

But, How is "$\approx$",

and How about the sample bias?

Reason for Rejecting H_0



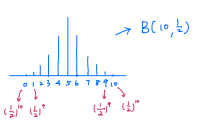
If $\mu_{\bar{x}} = \alpha \rightarrow H_0$ is correct

If $\mu_{\bar{x}} = \beta \rightarrow H_0$ is incorrect

The more $\mu_{\bar{x}}$ close to H_0,

the more prob. H_0 is correct

eg. → $B(10, \frac{1}{2})$



Type I & Type II Err.

Type I Err.

Reject H_0, when it's True

Type II Err.

Accept H_0, when it's False

α - Value

Level of significance :

max allowable prob. of making Type I err.

- $P(\text{rej. } H_0 \mid H_0 \text{ is True}) \le \alpha$

P - Value

the Smallest Significance Level at which the data lead to Rej H_0

- $P < \alpha \rightarrow$ Rej H_0
- $P > \alpha \rightarrow$ Acc. H_0

$P\text{-value} = P(Z \square Z_0)$

<eg> previous: at least 1.5 mg, current: less than 1.5 mg

sample = 20, SD = 0.7, ave = 1.42

5% level of significance? α

1. $H_0: \mu \ge 1.5$, $H_1: \mu < 1.5$

2. $Z_0 = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{1.42-1.5}{0.7/\sqrt{20}} = -0.511$

3. P-value $= P(Z \le -0.511) = |-0.695| = 0.305$ #

Five – Step Procedure

1. State H_0 and H_1
2. Select α ($\alpha = 0.05$)
3. Identify the test statistic
4. Formulate a decision rule
5. Take a sample, arrive at a decision

action

Do not reject H_0 (accept H_0) | reject H_0

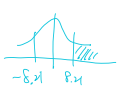
<eg> pre: #sample 30, $\mu = 8.16$, $\sigma = 0.17$,

cur: $\mu = 8.21$

1% level of significance,

H_1: ave(cur) < ave(pre)

1. $H_0: \mu \ge 8.21$, $H_1: \mu < 8.21$

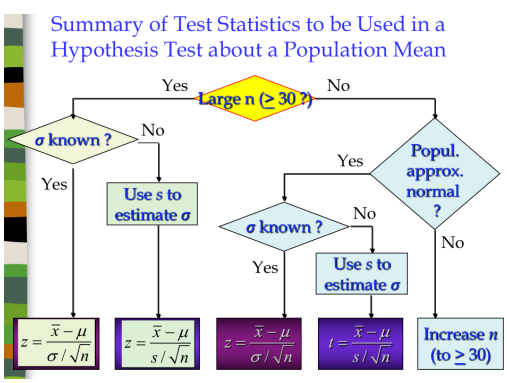


2. $\alpha = 0.01$

3. TS: $Z = \frac{\bar{x}-8.21}{s/\sqrt{n}}$

4. Rej H_0 if $Z < -2.33$

5. $Z = \frac{8.16-8.21}{0.17/\sqrt{30}} = -2.01 < -2.33 \Rightarrow$ Rej H_0!



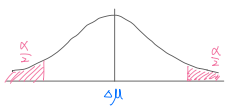
8 | About 2 Population

About 2 Population

CI

Sampling Dist of $\bar{X}-\bar{Y}$

<pf>



$$\frac{(\bar{X}-\bar{Y})-\Delta\mu}{\sigma} \sim N(0,1)$$

$$\Delta\mu = (\bar{X}-\bar{Y}) \pm \sigma \cdot \underbrace{Z_{\frac{\alpha}{2}} \cdot N(0,1)}$$

$$= \bar{X}-\bar{Y} \pm Z_{\frac{\alpha}{2}} \cdot \sigma$$

let $\bar{X}-\bar{Y}=\bar{Z}$

$$E\bar{Z} = \mu_1-\mu_2$$

$$\sigma_{\bar{Z}}^2 = \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

When σ_1, σ_2 are Known :

$$\bar{X}-\bar{Y} \pm Z_{\frac{\alpha}{2}} \cdot \underbrace{\sigma_{\bar{X}-\bar{Y}}} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Unknown σ_1, σ_2, large Sample

$$\bar{X}-\bar{Y} \pm Z_{\frac{\alpha}{2}} \cdot \underbrace{S_{\bar{X}-\bar{Y}}} = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

Unknown $\sigma=\sigma_1=\sigma_2$, Small Sample

$$\bar{X}-\bar{Y} \pm t_{(\frac{\alpha}{2}, n_1+n_2-2)} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \cdot \underbrace{S_p} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

<eg> $n_1=120, \bar{X}_1=235, S_1=15$
$n_2=80, \bar{X}_2=218, S_2=20$, 95% CI

Unknown σ , large Sample

$CI = \bar{X}-\bar{Y} \pm Z_{\frac{\alpha}{2}} \cdot S$, $S=\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$

$CI = 17 \pm 1.96 \cdot \sqrt{\frac{15^2}{120}+\frac{23^2}{80}}$

$= (11.86, 22.14)$ ※

<eg> $n_1=12, \bar{X}_1=29.8, S_1=2.56$
$n_2=8, \bar{X}_2=27.3, S_2=1.81$, 95%

Unknown σ, Small Sample

$CI = \bar{X}-\bar{Y} \pm t_{(n_1+n_2-2, 0.025)} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \cdot S_p$

$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

$CI = 29.8-27.3 \pm t_{(18, 0.025)} \cdot \sqrt{(\frac{1}{12}+\frac{1}{8}) \cdot \frac{11 \cdot 256^2+7 \cdot 1.81^2}{18}}$ (2.101)

$= (0.297, 4.703)$ ※

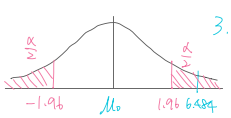
Hypothsis

<eg> Par $n_1=120, \bar{X}_1=235, S_1=15$
Rap $n_2=80, \bar{X}_2=218, S_2=20$, 0.05 sign

Is that the mean driving distance of Par Inc. golf balls is Greater than the mean driving distance of Rap. Inc. golf balls?

1. $H_0: \mu_2-\mu_1 \le 0$
   $H_1: \mu_2 > \mu_1$

2. $TS = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_2-\mu_1)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} = \frac{(235-218)-0}{\sqrt{\frac{15^2}{120}+\frac{23^2}{80}}} = 6.484$



3. $Z_{\frac{\alpha}{2}} = 1.96$

$\Rightarrow$ Rej H_0

2 - popu. Problem

Est $\mu_1 - \mu_2$

μ_1, μ_2 : popu. Mean

$\bar{X}_1, \bar{X}_2$: samp. Mean

Sample Dist:

$$\begin{cases} \dfrac{\bar{X}_1 - \bar{X}_2 - \mu_d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1), & n \text{ is large} \\ \dfrac{\bar{X}_1 - \bar{X}_2 - \mu_d}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2), & n \text{ is small} \end{cases}$$

$\mu_d = \mu_1 - \mu_2$

$$S_p: \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

d.f.

(Weighted Average)

1 - popu.

Sample Dist: $\begin{cases} \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim Z(0,1), & n \text{ is large} \\ \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1), & n \text{ is small} \end{cases}$

n is large

<pf>

CLT: $\begin{cases} \bar{X}_1 \sim N\left(\mu_1, \frac{S_1^2}{n_1}\right) \\ \bar{X}_2 \sim N\left(\mu_2, \frac{S_2^2}{n_2}\right) \end{cases}$

$\bar{X}_1 \perp \bar{X}_2$

$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)$

$\frac{\bar{X}_1 - \bar{X}_2 - \mu_d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$

<eg>

	Sample #1	#2
Inc	Par Inc.	Rap Inc.
#Sample	$n_1 = 120$	$n_2 = 80$
Mean	$\bar{X}_1 = 235$	$\bar{X}_2 = 218$
SD	$S_1 = 15$	$S_2 = 20$

$\bar{X}_1 - \bar{X}_2 \pm \cancel{1.96}^{1.96} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 17 \pm 5.14$

→ if $0 \notin CI$

$\Rightarrow \mu_1, \mu_2$: sig. diff.

↕ Reflect

Therefore we can say:

I have 95% Conf. to say that My product will better than your in test at

n is small:

① n_1, n_2 are N-Dist

② $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

$$\frac{X_1 - X_2 - \mu_d}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2)$$

∵ want to plot on the same graph but X-axis only has one dim. So need let $\sigma_1 = \sigma_2$

<eg>

	#1	#2
	M	J
#Sample	12	8
Mean	M.N	N.N
SD	2.56	1.81

$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2} = 5.28$

$\bar{X}_1 - \bar{X}_2 \pm t_{0.025} \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Hypo.

Not useful ⇒ check $\bar{X}$ in CI or not

$H_0: \mu_d \geq \mu_0$ $\quad H_0: \mu_d \leq \mu_0$ $\quad \boxed{H_0: \mu = 0}$

$H_1: \mu_d < \mu_0$ $\quad H_1: \mu_d > \mu_0$ $\quad \boxed{H_1: \mu \neq 0}$

is totally OK!

Rej. Region: $|Z| = \left| \dfrac{\bar{X}_1 - \bar{X}_2 - \mu_d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \right| > Z_{\frac{\alpha}{2}}$

$\equiv\rangle$ $3 \in CI \longleftrightarrow$ Accept H_0

<eg>

	Sample #1	#2
Inc	Par Inc.	Rap Inc.
#Sample	$n_1 = 120$	$n_2 = 80$
Mean	$\bar{X}_1 = 235$	$\bar{X}_2 = 218$

$H_0: \mu_d \le 3$ Rej Region:

$H_1: \mu_d > 3$ $Z > Z_{0.05}$

SD	$S_1 = 15$	$S_2 = 20$

$H_0: \mu_d \le 0$

$H_1: \mu_d > 0$

$\alpha = 0.05 \Rightarrow Z > 1.645$ Rej. Reg.

for $H_0: \mu_d = 0$, $H_1: \mu_d \neq 0$ ≡ 95% CI

<eg> $n = 22$, $p = \frac{10}{22}$, $\bar{X} = 7.135$, $\bar{Y}_2 = 6.45$

$S_1^2 = 0.178$, $S_2^2 = 0.391$

1. $\begin{cases} H_0: \mu_p \le \mu_c \\ H_1: \mu_p > \mu_c \end{cases}$

⇒ if $\begin{cases} H_0: \mu_p = \mu_c \\ H_1: \mu_p \neq \mu_c \end{cases}$

↳ p-value: 0.072

2. $S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2} = 0.689$

3. $TS = \frac{0.675}{\sqrt{S_p\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = 1.9$

4. p-value: $P(t(20) > 1.9)$

$$S_p = \frac{(n-1)\sigma_1^2 + (m-1)\sigma_2^2}{n+m-2}$$

$\mu_d = \mu_1 - \mu_2$

$$\frac{\bar{X} - \bar{Y}_2 - \mu_d}{S_p\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

9

10 | Analysis of Variance

Analysis of Variance

One (Factor) Way ANOVA

Overview

Goal: ~~Testing for the~~ Equality of K μ_{popu} ↗ when $K=1 \Rightarrow$ t-test; when $K=2 \Rightarrow$ 2 mean t-test; when $K \geq 3 \Rightarrow$ One way ANOVA

Structure: Predict Variable + Response Variable

Briefly: Check $\begin{cases} H_0: \mu_1=\mu_2=\cdots=\mu_K \\ H_1: \text{not all } \mu \text{ are equal} \end{cases}$ ↗ at least 2 μ are diff.

Assumptions: 1) $\forall\, \mu \in$ N-dist

2) $\forall\, \sigma^2$ are equal

3) $\forall$ obs are indep.

Data ($\bar{\bar{X}}$: overall ave.)

Treatment	Data	Sample Size	Sample Mean	Sample Var
1	$X_{11}, X_{12}, \cdots X_{1n_1}$	n_1	$\bar{X}_1$	S_1^2
2	$X_{21}, X_{22}, \cdots X_{2n_2}$	n_2	$\bar{X}_2$	S_2^2
⋮	⋮	⋮	⋮	⋮
K	$X_{K1}, X_{K2}, \cdots X_{Kn_K}$	n_K	$\bar{X}_K$	S_K^2

How & When: when Variability$_{Between}$ (By Ratio) >> Variability$_{Within}$ $\Rightarrow$ ~~Reject H_0~~

Sum of Square (SS)

Total Variability = Between group | within group

SS_{Total}	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij}-\bar{\bar{X}})^2$	Total SS
SS_R ($SS_{Treatment}$, SS_{Model})	$\sum_{i}^{K} n_i(\bar{X}_i-\bar{\bar{X}})^2$	SS between groups

Between	$\sum_{i=1}^{K} \bar{t}_{weight}$	
$SS_{Treatments}, SS_{Residuals}$ SS_{Within}	$\sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$ ($(n_i-1)S_i^2$)	SS within groups

$\sum (n_j - 1) S_j^2$

ANOVA Table

Source of Variance	SS	df.	MS	$F_{(K-1,\ n-K)}$
Between	SS_B	$k-1$	MS_B	$\frac{MS_B}{MS_W}$
Within	SS_W	$n-k$	MS_W	
Total	SS_T	$n-1$		

1) If H_0 is True, $\frac{MS_B}{MS_W}$ dist is F-dist $(K-1, n-K)$

2) MS_E ∈ Unbiased est. of σ^2

3) If H_0 is False, MSB will overest. ($\nearrow$ larger) σ^2

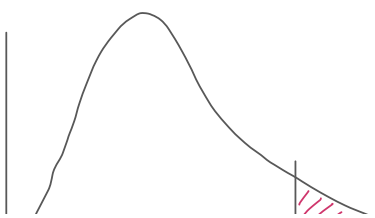
H_0 is { True : $MS_B \simeq MS_W \simeq \sigma^2$; False : $MS_B \gg MS_W$

Hypotheses

1) { H_0 : $\mu_1 = \mu_2 = \cdots \mu_K$; H_1 : at least 2 μ aren't equal

2) $TS = F_{(K-1,\ n-K)} = \frac{MS_B}{MS_W}$

3) Reject Rule : $F_t > F_\alpha$



eg. 3 Factories, 95% sign.

	$\bar{X}$	S^2
Obv1	55	26
Obv2	68	265
Obv3	57	245

1) { H_0 : $\mu_1 = \mu_2 = \mu_3$; H_1 : not all equal

2) $T = \frac{MS_B}{MS_W}$, $MS_B = \frac{SS_B}{k-1}$, $SS_B = \sum_{i=1}^{K} n_i (\bar{X}_i - \bar{\bar{X}})^2$, $MS_W = \frac{SS_W}{n-k}$, $SS_W = \sum_{i=1}^{K} (n_i - 1) S_i^2$

SS	df	MS	F
490	2	245	= 9.55
308	12	25.67	
798	14		

3) $F_{.05}(2,12) = 3.89 < F_T \Rightarrow$ Rej H_0 !

eg. Prove $SS_T = SS_B + SS_W$

eg. ...ning methods (A, B, C , D). At the level of α = 0.01, test whether differen... teaching methods affect score.

$T_j/n_j = \bar{x}_j$

Method	Score	T_j	n_j	$\sum x^2$
A	94,90,85,80	349	4	30516
B	75,68,77,83,88	391	5	30811
C	70,73,76,78,80,68,65	510	7	37338
D	68,70,72,65,74,65	414	6	28634
	Total	1664	22	127344
		Σx	n	Σx^2

$S_i^2 = \frac{\sum x_{ij}^2 - T_i/n_i}{n_i - 1}$

$SS_T = SS_B + SS_W$

$SS_T = \sum x^2 - \frac{(\sum x)^2}{n}$ (1) ...

$SS_B = \sum \frac{T_j^2}{n_j} - \frac{(\sum x)^2}{n}$

$SS_W = SS_T - SS_B$

(2) $F_T = 8.99$

(3) $F_{.01}(3,18) = 5.09 < F_T \Rightarrow$ Rej H_0

SS	df	MS	F_T
890.7	3	296.9	8.99
594.4	18	33	
1485.1	21		

ANOVA Model

eg. T-cell Count

$$X_{ij} = \underbrace{\mu}_{\text{Overall Popul. } (\mu = \frac{\sum \mu_i}{k})} + \underbrace{\alpha_i}_{\text{Effect}} + \underbrace{\epsilon_{ij}}_{\text{Error Term}}$$

Base level; Medication; Unaccounted for Variation

$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

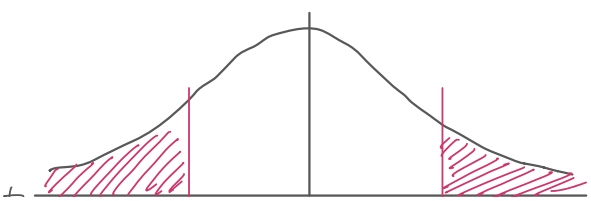
$\alpha_1 = \mu_1 - \mu$
$\alpha_2 = \mu_2 - \mu$
$\alpha_3 = \mu_3 - \mu$

"$\alpha_1 + \alpha_2 + \alpha_3 = 0$"

Fisher's LSD Procedure (Check How diff)

1) Hypo: $\begin{cases} H_0: \mu_i = \mu_j \\ H_1: \mu_i \neq \mu_j \end{cases}$

2) $TS = t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MS_W\left(\frac{1}{n_i} + \frac{1}{n_i}\right)}}$



3) Reject Rule:

$t_{(n-k)}$ not in $(-t_{\frac{\alpha}{2}}, t_{\frac{\alpha}{2}})$

$t_{(n-k)}$ $\quad -t_{\frac{\alpha}{2}}$ $\quad t_{\frac{\alpha}{2}}$

2) TS = $\bar{x}_i - \bar{x}_j$

3) Reject Rule

$|\bar{x}_i - \bar{x}_j| > LSD$

$LSD = t_{(\frac{\alpha}{2}, n-k)} \sqrt{MS_w \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

eg.

Example: Reed Manufacturing

■ Analysis of Variance

Observation	Plant 1 Buffalo	Plant 2 Pittsburgh	Plant 3 Detroit
1	48	73	51
2	54	63	63
3	57	66	61
4	54	64	54
5	62	74	56
Sample Mean	55	68	57
Sample Variance	26.0	26.5	24.5

SS	df	MS	F
490	2	245	9.3
308	12	25.67	

1) $\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$

2) TS = $|\bar{x}_1 - \bar{x}_2| = 13$

3) LSD = $t_{(12, \frac{.05}{2})} \sqrt{MS_w \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$= 2.18 \cdot \sqrt{25.67\left(\frac{1}{5} + \frac{1}{5}\right)} = 6.98$

TS $\notin (-6.98, 6.98)$

Rej H_0

factor

Two Way ANOVA

Overview

Why we need 2-way?

The Predict Variable might be 2 or More

eg. age, education level $\longrightarrow$ income

Model

$$X_{ij} = \mu + \underbrace{\alpha_i}_{\text{Blocking Factor}} + \underbrace{\beta_j}_{\text{Interested Factor}} + \underbrace{\epsilon_{ij}}_{\sim N(0,\sigma^2)}$$

$\sum \alpha_i = \sum \beta_j = 0$

Data

(Block) Factor B

(Treatment) Factor A	1	2	3	⋯	b	mean
1	X_{11}	X_{12}	X_{13}	⋯	X_{1b}	$\bar{X}_{1.}$
2	X_{21}					$\bar{X}_{2.}$
3	X_{31}					⋮
⋮	⋮					
a	X_{a1}					$\bar{X}_{a.}$
mean	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$	⋯	$\bar{X}_{.b}$	$\bar{\bar{X}}$

$$\bar{X}_{ij} - \bar{\bar{X}} = X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{\bar{X}} + (\bar{X}_{i.} - \bar{\bar{X}}) + (\bar{X}_{.j} - \bar{\bar{X}})$$

SS:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{\bar{X}})^2 = \sum\sum(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{\bar{X}})^2 + b\sum_{i=1}^{a}(\bar{X}_{i.} - \bar{\bar{X}})^2 + a\sum_{j=1}^{b}(\bar{X}_{.j} - \bar{\bar{X}})^2$$

$$SS_{Total} = SS_{Error} + SS_A + SS_B$$

if don't care about one factor (Block factor) then it will be included in SS_{Error}

df: $n-1 = (a-1)(b-1) + (a-1) + (b-1)$

Hypotheses

1) H_{0a} : all $\alpha_i = 0$

   H_{1a} : not all $\alpha_i = 0$

2) TS = $F_1(df_1, df_2) = \frac{MS_A}{MS_E}$ ($df_1 = a-1$, $df_2 = (a-1)(b-1)$)

3) Reject Rule:

$$F_1 > F_{1\alpha}(df_1, df_2)$$

1) H_{0b} : all $\beta_i = 0$

   H_{1b} : not all $\beta_i = 0$

2) TS = $F_1(df_1, df_2) = \frac{MS_A}{MS_E}$ ($df_1 = b-1$, $df_2 = (a-1)(b-1)$)

3) Reject Rule:

$$F_1 > F_{1\alpha}(df_1, df_2)$$

lable

	SS	df	MS	F
Treatment	SS_T	$(a-1)$	$\frac{SS_T}{df_T}$	$\frac{MS_T}{MS_E}$ $\frac{[1,3]}{[3,3]}$
Block	SS_B	$(b-1)$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_E}$ $\frac{[2,3]}{[3,3]}$
Error	SS_E	$(a-1)(b-1)$		
Total	Sum [:, 1] Sum(SS)	Sum [:, 2] Sum(df)		

eg.

■ **Eg.** There are four routes A, B, C, and D from the suburbs to the downtown area. Is there any difference in driving time? ($\alpha = 0.05$)

Drivers. \ Route	A	B	C	D	B_r
John	18	20	20	22	80
Marry	21	22	24	24	91
Anthony	20	23	25	23	91
Claire	25	21	28	25	99
Yvonne	26	24	28	25	103
T_c	110	110	125	119	464
Σx^2	2466	2430	3169	2839	10904

One Way

Don't Care about the Driver effect

ANOVA Table

Source of variance	Sum of square	df	Mean of square	F
Treatments	32.4	3	10.8	1.618
Error	106.8	16	6.675	
Total	139.2			

$\because F_{0.05}(3,16) = 3.24 < F_\alpha$ 19 $\Rightarrow$ Don't reject H_0

Two Way

Two-way ANOVA Table

Source of variance	Sum of square	df	Mean of square	F
A Treatments	32.4 (SSTR)	3 (a-1)	10.8 (MSTR)	4.5 MSTR/MSE
B Block (SS BL)	78.2 (SSBL)	4 (b-1)	19.6 (MSBL)	8.2 MSBL/MSE
Error (SS E)	28.6 (SSE)	12 (a-1)(b-1)	2.38 (MSE)	
Total	139.2 (SS_{total})	19 (n-1)		

Provides Acc

4.5 > 10.8/2.38

Where a : number of treatments.
b : number of blocking.

SSBL+SSE = OLD SSE (SSE in the previous ANOVA table)

Treatment Effect : $F_1 = \frac{MS_T}{MS_E}(a-1, (a-1)(b-1)) = 4.5 > 3.49 \Rightarrow$ Rej H_{a0}

Block Effect : $F_1 = \frac{MS_B}{MS_E}(b-1, (a-1)(b-1)) = 8.2 > 3.259 \Rightarrow$ Rej H_{b0}

Conclusion

Experiment

- Factor / Treatment Variable : age , education
- Level / Subcategories : $\begin{cases} 1-20 \\ 21-40 \\ 41-60 \end{cases}$, $\begin{cases} \text{Junior} \\ \text{senior} \\ \text{college} \end{cases}$

11 | Chi-Squared Goodness of Fit Test

Chi-Squared Goodness of Fit Test

Goodness of Fit Tests

A Multinomial Population

- Categorical Variable : Using Number to token the types
- Binomial - Dist :

Do n times, Succ : X, Fail : $n-X$

$X \sim B(n, p)$ $\begin{cases} EX = np \\ \mathrm{Var}X = npq \end{cases}$, $p = \frac{X}{n}$

Multinomial - Dist

Do n times, type$_1$: P_1, type$_2$: P_2 $\cdots$, type$_K$: P_k

$$f(x_1, x_2, \cdots x_k, P_1, P_2, \cdots P_k)$$

$$= Pr\left(\bigcap_{i=1}^{k} X_i = x_i\right)$$

$$= \begin{cases} \frac{n!}{x_1!\, x_2! \cdots x_k!} P_1^{x_1} \cdot P_2^{x_2} \cdots P_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \\ 0, & \text{o.w.} \end{cases}$$

when $k=2$, $f(X_1 = x_1, X_2 = x_2)$

$$= \frac{n!}{x_1!\, x_2!} P_1^{x_1} \cdot P_2^{x_2}$$

$$= \binom{n}{x_1} P_1^{x_1} (1-P_1)^{n-x_1}$$

$X_1 \sim B(n, P_1)$

Properties

- $EX_i = nP_i$
- $\mathrm{Var}X_i = nP_iq_i$
- $\mathrm{Cov}(X_i, X_j) = -nP_iP_j$

$$= E(X_iX_j) - EX_iEX_j$$

<pf>

$X_i + X_j \sim B(n, P_i + P_j)$

$\Rightarrow \mathrm{Var}(X_i + X_j) = \mathrm{Var}X_i + \mathrm{Var}X_j + 2\,\mathrm{Cov}(X_i, X_j)$

$\Rightarrow n(P_i+P_j)(1-P_i-P_j) = nP_iq_i + nP_jq_j + 2\,\mathrm{Cov}(X_i, X_j)$

$\Rightarrow \mathrm{Cov}(X_i, X_j) = -nP_iP_j$

Hypotheses

1. $\begin{cases} H_0 : \text{any} \\ H_1 : \text{not any} \end{cases}$

Why χ^2 ?

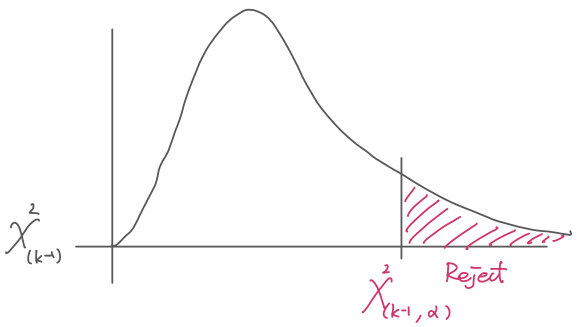
Consider $k=2 \Rightarrow \begin{cases} O_1 = X_1 \sim B(n, P_1), & EX_1 = nP \\ O_2 = X_2 = n - X_1, & EX_2 = nq \end{cases}$

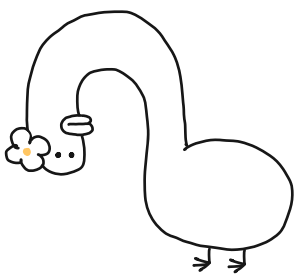
$$\frac{(O_1-E_1)^2}{} + \frac{(O_2-E_2)^2}{} = \frac{(X-np)^2}{} + \frac{(X-np)^2}{}$$

2. $TS = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)}$

3. Reject Rule:

$TS > \chi^2_{\alpha(k-1)}$





$E_1 \quad E_2 \qquad np \quad nq$

$= \frac{(X-np)^2}{npq} \cong N(0,1)^2$ (μ, S^2)

eg.

Determine if a six-sided die is fair, 10% sign level, $n = 60$

Die Value	1	2	3	4	5	6
Obs. Freq.	9	15	9	8	6	13

1. $\begin{cases} H_0: P_1 = P_2 = \cdots = P_6 = \frac{1}{6} \\ H_1: P_1, \cdots P_6 \text{ are not equal to } \frac{1}{6} \end{cases}$

2. $TS = \chi^2_5 \left(\sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \right)$

$= \chi^2_5 \left(\frac{(9-10)^2}{10} + \frac{(15-10)^2}{10} + \cdots + \frac{(13-10)^2}{10} \right) = 5.6$

3. $TS < \chi^2_{5,0.1} = 9.24 \Rightarrow$ do not rej H_0

eg.

Side 2,5, is twice that of other point, 95%

Die Value	1	2	3	4	5	6	Total
Obs. Freq.	13	33	14	7	36	17	120
Exp. Freq.	15	30	15	15	30	15	120

1. $\begin{cases} H_0: P_2 = P_5 = \frac{2}{8}, \; P_1 = P_3 = P_4 = P_6 = \frac{1}{8} \\ H_1: \text{not } H_0 \end{cases}$

2. $TS = \chi^2 = \frac{1}{30}\left((33-30)^2 + (36-30)^2 \right) + \frac{1}{15}\left((13-15)^2 + \cdots + (17-15)^2 \right)$

$= 6.4$

3. Rej Rule: $TS = 6.4 < 11.07 = \chi^2_{(5,.05)} \Rightarrow$ Do not Reject H_0

Contingency Table Analysis (Test Independence)

Independence vs. Non-Independence

	D	¬D	Total
M	16	64	80
F	24	96	120
Total	40	160	200

	D	¬D	Total
M	24	56	80
F	6	114	120
Total	30	170	200

$P(D|M) = 20\%$

$P(D|\bar{M}) = 20\%$

Disease $\perp$ Sex

$P(D|M) = 30\%$

$P(D|\bar{M}) = 5\%$

Disease $\not\perp$ Sex

Contingency Table

Joint probability

Marginal probability

	D	D′	Total
M	16 (P_{11}, $\frac{16}{200}$)	64 (P_{21}, $\frac{64}{200}$)	80 ($P_{1\cdot}$, $\frac{120}{200}$)
F	24 (P_{21}, $\frac{24}{200}$)	96 (P_{22}, $\frac{96}{200}$)	120 ($P_{2\cdot}$, $\frac{120}{200}$)
Total	40 ($P_{\cdot 1}$, $\frac{40}{200}$)	160 ($P_{\cdot 2}$, $\frac{40}{200}$)	200

$\perp$: Joint prob. = product of marginal prob.

$$\begin{cases} (M,D) = (M,T) * (T,D) \\ (M,D') = (M,T) * (T,D') \\ (F,D) = (F,T) * (T,D) \\ (F,D') = (F,T) * (T,D') \end{cases}$$

Test of Independence

Obs. Freq.

	D	¬D	Total
M	O_{11}	O_{12}	R_1
F	O_{21}	O_{22}	R_2
Total	C_1	C_2	n

Exp. Freq. (Indep.)

	D	¬D	Total
M	E_{11}	E_{12}	R_1
F	E_{21}	E_{22}	R_2
Total	C_1	C_2	n

total	C_1	C_2	⋯

obs. freq. : O_{ij}

total	C_1	C_2	⋯

exp. freq. : $E_{ij} = \frac{1}{n} R_i C_j$

Hypotheses

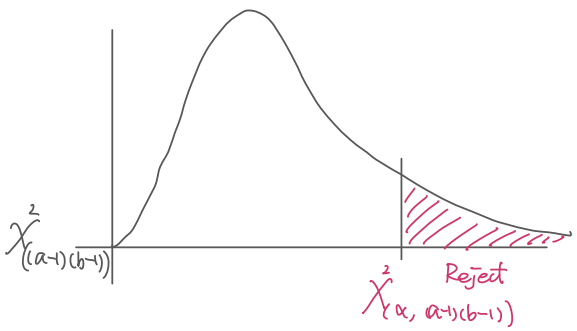
1. $\begin{cases} H_0: \text{all var. are indep.} \\ H_1: \text{at least 2 var are depen.} \end{cases}$

2. $E_{ij} = \frac{1}{n} R_i C_j = \frac{(\text{Total Row}_i \times \text{Total Col}_j)}{\text{Sample size}}$

3. $TS = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2$

4. Reject Rule :

$$TS > \chi^2_{(\alpha, (a-1)(b-1))}$$



eg.

Obs Freq. 95%

	C	NC	Sum(Row)
M	24	56	80
F	16	104	120
Sum(Col)	40	160	200

2.

Exp Freq.

	C	NC	% (Row)
M	$40 \cdot 0.4 = 16$	$160 \cdot 0.4 = 64$	0.4
F	$40 \cdot 0.6 = 24$	$160 \cdot 0.6 = 96$	0.6
% (Col)	0.2	0.8	1

3. $TS = \frac{(24-16)^2}{16} + \frac{(56-64)^2}{64} + \frac{(16-24)^2}{24} + \frac{(104-96)^2}{96} = 8.33$

4. $TS = 8.33 > \chi^2_{(\alpha, (2-1)(2-1))} = 3.8$

$\Rightarrow$ Reject H_0

Does the Data Come From Poisson-Dist ?

Hypotheses (Poisson)

1. $\begin{cases} H_0: \text{data come from Poisson-Dist} \\ H_1: \text{data do not} \quad : \end{cases}$

2. select RS and Record Obs. freq. (f_i)

3. Compute Exp freq. (E_i) merge cell

eg.

A RS: n=100, 1-min time intervals resulted in the customer arrivals listed below. Is $X \sim$ Poi ?

Arrivals	0	1	2	3	4	5	6	7	8	9	10	11	12
freq.	0	1	4	10	14	20	12	12	9	8	6	3	1

$ X = "# customers arrival in 1-min"

↗ Unknown

when $E_i < 5$

4. $TS = \sum_{i=1}^{k} \frac{(f_i - E_i)^2}{E_i} \sim \chi^2_{(k-2)}$

5. Rejeit Rule :

$TS > \chi^2_{(\alpha, k-2)}$

!!! Generally ,

$df = k - p - 1$ { K : # categries ; p : # popu. params. est.

1. { H_0 : $X \sim Poi(\underline{\mu})$; H_1 : $X \not\sim Poi(\mu)$

[let $\mu = \hat{\mu}$: sample Mean]

2. $\mu = \frac{\text{Total Arrivals}}{\text{Total time}} = \frac{\sum Arr_i * freq_i}{n} = \frac{600}{100}$

$\Rightarrow f(x) = \frac{6^x e^{-6}}{X!}$

3. Exp <5, merge <5, merge

Arrivals	0	1	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	.0025	.0149	.0446	.0892	.1339	.1620	.1606	.1389	.1041	.0694	.0417	.0227	.0155
$100 \cdot f(x)$	.25	1.49	4.46	8.92	13.39	16.2	16.06	13.89	10.41	6.94	4.17	2.27	1.55

4. TS

Arrivals	0 1 2	3 4 5 6 7 8 9	10 11 12
f_i	5	10 ~ 8	10
E_i	6.2	8.92 ~ 6.94	7.99
$f_i - E_i$	-1.2	1.08 ~ 1.06	2.01

K = 9 !! $TS = \frac{(-1.2)^2}{6.2} + \cdots + \frac{(2.01)^2}{7.99} = 3.42 \sim \chi^2_9$

5. Rej Rule :

$TS = 3.42 < \chi^2_{(0.05,\ 9-2)} = 14.07$

$\Rightarrow$ Do not reject H_0

Hypotheses (Normal - Dist)

1. { H_0 : data come from Normal-Dist $(\underline{\mu}, \underline{\sigma}^2)$ ($\bar{X}$, S^2) ; H_1 : data do not :

2. select RS and Record Obs. freq. (f_i)

3. Compute Exp freq. (E_i), merge cell when $E_i < 5$

4. $TS = \sum_{i=1}^{k} \frac{(f_i - E_i)^2}{E_i} \sim \chi^2_{(k-3)}$

5. Rejeit Rule :

$TS > \chi^2_{(\alpha, k-3)}$

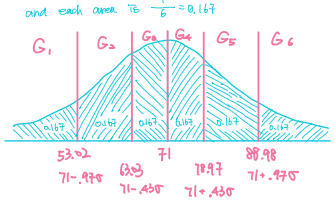
!!! Generally , p=2 since, we have μ, σ^2 to est.

eg. A RS: n=30, $\bar{X}$=71, S=18.54, N~Dist ?

33 43 44 45 52 52 56 58 63 64
64 65 66 68 70 72 73 73 74 75
83 84 85 86 91 92 94 98 102 105

1. { H_0 : $X \sim N(\mu, \sigma^2)$; H_1 : $X \not\sim N(\mu, \sigma^2)$

2. we have n=30, merge 5 to 1 → K=6
therefore, we have 6 Interval,
and each area is $\frac{1}{6} = 0.167$



{ group 1 : $(-\infty, 53.02)$; group 2 : $(53.02, 63.03)$

$df = k - p - 1$ { K : # categories; P : # popu. params. est.

group6: $(88.98, +\infty)$

3. k=6

group i	f_i	E_i	$f_i - E_i$
< 53.02	6	↓ 5 ↑	1
$53.02 \sim 63.03$	3		-2
$63.03 \sim 71$	6		1
$71 \sim 78.97$	5		0
$78.97 \sim 88.98$	4		-1
> 88.98	6		1
Total	30	30	

4. $TS = \sum \frac{(f_i - E_i)^2}{E_i} = 1.6$

5. Reject Rule: $TS = 1.6 < \chi^2_{(0.05,\,6)} = 7.81$

$\Rightarrow$ Do not Reject H_0

12 Simple Linear Regression

Simple Linear Regression

Least Square Procedure

Target

Examine the relationship between

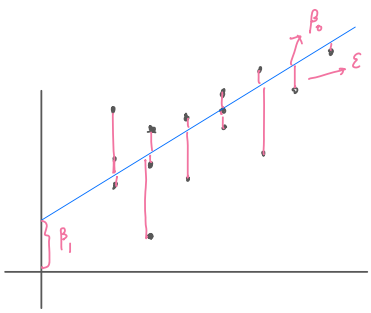
explanatory var.(x) and response var. (y)

Model

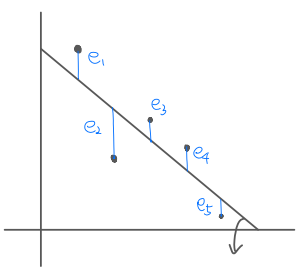
$\varepsilon \sim N(0, \sigma^2)$

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{effect}} + \underbrace{\varepsilon}_{\text{Error term}}$$

- y = response
- x = explanatory
- β_0 = y- intercept
- β_1 = slope of the line
- ε = random error



Procedure



$$y_i = \beta_0 + \beta_1 x + e_i$$

1) Assume $e_i \sim N(0, \sigma^2)$
$\forall e_i$ are indep.

2) def. SSE (Sum of Squared Errors)

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

3) Find β_0, β_1 s.t. SSE has min.

For minimize SSE

$$\$ \begin{cases} \frac{\partial}{\partial \beta_0} SSE = 0 \\ \frac{\partial}{\partial \beta_1} SSE = 0 \end{cases}$$

we have when $\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \end{cases}$ have min SSE

where $\begin{cases} S_{xx} = \sum_i x_i^2 - \frac{\sum x_i^2}{n} = \sum (x_i - \bar{x})^2 \\ S_{xy} = \sum_i x_i y_i - \frac{\sum x_i \sum y_i}{n} = \sum (x_i - \bar{x})(y_i - \bar{y}) \\ \qquad = \sum (x_i - \bar{x}) y_i \end{cases}$

<pf>

$$\min \sum_i (y_i - \beta_0 - \beta_1 x_i)^2 \triangleq f(\beta_0, \beta_1)$$

$$\$ \begin{cases} \frac{\partial}{\partial \beta_0} f = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial}{\partial \beta_1} f = -2 \sum_i x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

$$\Rightarrow \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

4) $y = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\begin{cases} \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r_{xy} \cdot \frac{S_y}{S_x} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$

$\Rightarrow y = \bar{y} + \hat{\beta}_1 (x - \bar{x})$ → therefore the C.M. is $(\bar{x}, \bar{y})$ since if $x = \bar{x}$, then $y = \bar{y}$

Example: The data y has been observed for various values of x, as follows:

y	240	181	193	155	172	110	113	75	94
x	1.6	9.4	15.5	20.0	22.0	35.5	43.0	40.5	33.0

Fit the simple linear regression model using least squares.

eg.

$\hat{\beta}_0$ S_{xx} $S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

1. $\begin{cases} \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$ $\begin{cases} S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\ S_{xy} = \sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n} \end{cases}$

2.

n	$\sum x_i$	$\sum x_i^2$	$\sum y_i$	$\sum y_i^2$	$\sum x_i y_i$
9	2205	7033.7	1330	220549	248644

3. $\begin{cases} S_{xx} = 7033.7 - \frac{1}{9}(2205)^2 = 1651.45 \\ S_{xy} = 26864.4 - \frac{1}{9}(2205 \cdot 1330) = -5794.1 \end{cases}$

4. $\begin{cases} \hat{\beta}_1 = \frac{1651.45}{-5794.1} = -3.5085 \\ \hat{\beta}_0 = \frac{1}{9}(1330) - (-3.5085) \times \frac{1}{9}(220.5) = 234.1 \end{cases}$

5 $\Rightarrow$ $y = 234.1 - 3.509\,x$ ※

$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i - n\bar{x}^2$

$= \frac{1}{S_{xx}} \begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$

$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i \sum y_i + \sum x_i^2 \sum x_i y_i}{\sum x_i y_i - n\bar{x}\bar{y}} = \frac{S_{xy}}{S_{xx}}$

$\Rightarrow \hat{\beta}_0 = \frac{1}{S_{xx}}\left(\sum x_i^2 \bar{y} - (\sum x_i y_i)\bar{x}\right)$

$= \bar{y} - \hat{\beta}_1 \bar{x}$

Quantifying the Goodness of the fit

residuals

Est. error $= \hat{e}_i = y_i - \hat{y}_i$ $\quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

usage residuals to Est. error

therefore the error $e_i \approx \frac{1}{n-1}\sum \hat{e}_i^2$ residuals

$\hat{\sigma}^2 = \frac{1}{n-2}\sum \hat{e}_i^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2$

$= \frac{S_{xx} - \hat{\beta}_1 S_{xy}}{n-2}$ $\left(\frac{\text{Residual Sum of Squares}}{df}\right)$

β_0, β_1 have been est.

stick-up.

$\hat{\beta}_1 = r_{xy} \cdot \frac{S_y}{S_x}$

<pf>

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})\sum (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$S_{xx} = (n-1)S_x^2$ $= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_x^2 S_y^2}} \cdot \frac{S_y}{S_x}$ r_{xy}

$\begin{cases} S_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 \\ S_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 \end{cases}$

$S_x^2 = S_{xx}$

$S_y^2 = S_{yy}$

Sum of Error of Square (SSE)

SSE

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$= S_{yy} - \hat{\beta}_1 \cdot S_{xy}$$

$$= S_{yy} - \hat{\beta}_1^2 \cdot S_{xx}$$

$$\hat{\sigma}^2 = S_\varepsilon^2 = \frac{SSE}{n-2}$$

<pf>

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \sum (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))^2$$

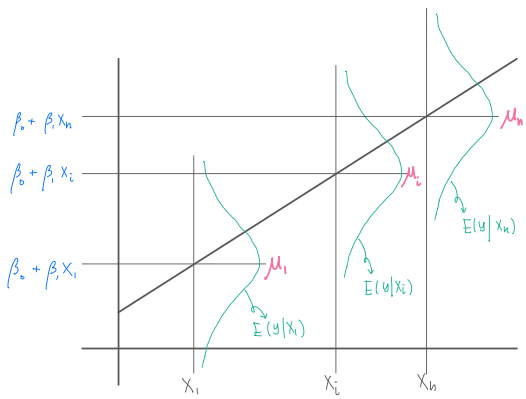
$$= \sum (y_i - \bar{y})^2 - 2\hat{\beta}_1 \cdot \sum (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2 \cdot \sum (x_i - \bar{x})^2$$

$$= S_{yy} - \hat{\beta}_1 S_{xy}$$

$$2\hat{\beta}_1 \cdot \sum (y_i - \bar{y})(x_i - \bar{x}) = 2\hat{\beta}_1 \cdot S_{xy} = 2\frac{S_{xy}^2}{S_{xx}}$$

$$\hat{\beta}_1^2 \cdot \sum (x_i - \bar{x})^2 = \left(\frac{S_{xy}}{S_{xx}}\right)^2 \cdot S_{xx} = \frac{(S_{xy})^2}{S_{xx}}$$

The Normality of ε



Hypotheses (β_1 : slope)

1) $\begin{cases} H_0: \beta_1 = \beta_1' \\ H_1: \beta_1 \neq \beta_1' \end{cases}$ or $\begin{cases} H_0: \beta_1 = \beta_1' \\ H_1: \beta_1 > \beta_1' \end{cases}$ ...

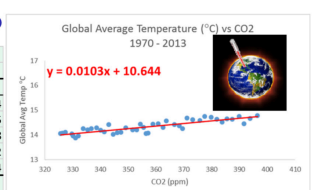
2) check : r^2

3) est. Residuals : $\hat{e}_i = y_i - \hat{y}_i$

eg.

Global Temperature (cont.)

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.857739864
R Square	0.735717675
Adjusted R Square	0.729425238
Standard Error	0.131613902
Observations	44



ANOVA						
			MS	*F*	*Significance F*	
Regression		1	2.025330433	116.9209567	1.03354E-13	
Residual		42	0.727533203	0.017322219		
Total		43	2.752863636			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	10.6441999	0.341822184	31.1395819	1.19922E-30	9.954374809	11.334025
CO2	0.010324616	0.000954834	10.81299943	1.03354E-13	0.008397684	0.012251548

Results of hypothesis test
$H_0: \beta_1 = 0$
$H_A: \beta_1 \neq 0$
Reject H_0
Confidence interval for β_1

Test the Slope (X, y relationship)

if Slope ($\hat{\beta}_1$) ≈ 0, then X, y have no "Linear Relationship."

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad \begin{cases} E\, y_i = \beta_0 + \beta_1 x_i \\ Var\, y_i = Var(\varepsilon_i) = \sigma^2 \end{cases}$$

- $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})\, y_i$
- $Var\, S_{xy} = \sum_i (x_i - \bar{x})^2 \cdot Var(y_i)$

1. $\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \,|\, > 0 \,|\, < 0 \end{cases}$

2. $TS \sim t\text{-Dist} = \dfrac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$

$S_{\hat{\beta}_1} = \dfrac{S_\varepsilon}{\sqrt{n-1} \cdot S_x}$

3. Reject Rule:

2-T : $|t| > t_{(n-2, \frac{\alpha}{2})}$

L-T : $t < t_{(n-2, \alpha)}$

R-T : $t > t_{(n-2, \alpha)}$

* $CI = \hat{\beta}_1 \pm t_{(n-2, \frac{\alpha}{2})} \, S_{\hat{\beta}_1}$

$= S_{xx} \cdot \sigma^2$

- $Var \dfrac{S_{xy}}{S_{xx}} = \dfrac{Var\, S_{xy}}{S_{xx}^2} = \dfrac{\sigma^2}{S_{xx}}$
- $S_{\hat{\beta}_1} = \sqrt{Var\, \hat{\beta}_1} = \sqrt{Var \dfrac{S_{xy}}{S_{xx}}} = \dfrac{\sigma}{\sqrt{S_{xx}}}$
- $\hat{\sigma}^2 = \dfrac{SSE}{n-2}$, $\dfrac{S_{xx}}{n-1} = S_{\hat{\beta}_1}^2 = S_x^2$

Coefficient of Determination ($r^2 = \dfrac{SS_R}{SS_T}$)

if $r^2 \uparrow$, then Model fixed well!

- Case I) { Ignored X / use $\bar{y}$ to predict y

$SS_{Total} = \sum (obs - pred)^2$

S_{yy}

$= \sum (y_i - \bar{y})^2$

- $SS_T = SS_E + SS_{Regression} = \hat{\beta}_1 S_{xy}$ (better / better)

<pf>

$SS_T = \sum (y_i - \bar{y})^2$

$= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$

$= \underbrace{\sum (y_i - \hat{y}_i)^2}_{SS_E} + \underbrace{\sum (y_i - \bar{y})^2}_{SS_R} + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ (= 0)

- Case II) { use X / use $\hat{y} = \beta_0 + \beta_1 X$

$SS_{Error} = \sum (obs - pred)^2$

$S_{yy} - \hat{\beta}_1 S_{xy}$

$= \sum (y_i - \hat{y}_i)^2$

- $r^2 = \dfrac{SS_R}{SS_T}$, if $r^2 \approx 1$, then better

$= (r_{xy})^2$

<pf>

$\dfrac{SS_R}{SS_T} = 1 - \dfrac{SS_E}{SS_T} = 1 - \dfrac{S_{yy} - \hat{\beta}_1 S_{xy}}{S_{yy}}$

$= \dfrac{\hat{\beta}_1 S_{xy}}{S_{yy}} = \dfrac{\frac{S_{xy}}{S_{xx}} S_{xy}}{S_{yy}}$

$= \dfrac{S_{xy}^2}{S_{xx} \cdot S_{yy}} = (r_{xy})^2$

eg.

$t^2 = \left(\dfrac{\hat{\beta}_1}{S_{\hat{\beta}_1}}\right)^2$

	SS	df	MS	F	p-value
Reg.	1065.8	1	1065.5	10.87	0.01
Err.	784.2	8	98		
Total	1850	9			

(+ = =)

F-dist (1, 8) = 10.87

<pf>

$\Rightarrow F = \dfrac{MS_R}{MS_E} = \dfrac{\frac{SS_R}{df_R}}{\frac{SS_E}{n-2}} = \dfrac{SS_R}{\hat{\sigma}^2}$ (1)

$\therefore F = \dfrac{1}{\hat{\sigma}^2} \dfrac{S_{xy}^2}{S_{xx}}$

$\Rightarrow t = \dfrac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$, $t^2 = \dfrac{\hat{\beta}_1^2}{S_{\hat{\beta}_1}^2} = \dfrac{\left(\frac{S_{xy}}{S_{xx}}\right)^2}{\frac{1}{S_{xx}}\hat{\sigma}^2}$

$\therefore t^2 = \dfrac{1}{\hat{\sigma}^2} \dfrac{S_{xy}}{S_{xx}} = F$ #

Using the Regression Equation

Point Estimation → Interval Estimation

when we have Reg Eq ($\hat{y} = \beta_0 + \beta_1 x$)

Just input x, we could give $\hat{y}$ (point)

But "How Close" to the real point?

Confidence Interval (for "average" of all 95% CI)

Est. the "Average y" for given X_v

$$\hat{y} \pm t_{\frac{\alpha}{2}} \cdot \sqrt{SE^2(\hat{\beta}_1)\cdot(X_v-\bar{x})^2 + \frac{S_\varepsilon^2}{n}}$$

df : $n-2$

C.I.

$$y \mid X_v \sim N(\beta_0+\beta_1 X_v, \hat{\sigma}^2) \quad (MS_E)$$

$$\Rightarrow \hat{E}(y \mid X_v) = \hat{\beta}_0 + \hat{\beta}_1 X_v = \bar{y} + \hat{\beta}_1(X_v - \bar{x})$$

$$\Rightarrow Var\left(\hat{E}(y \mid X_v)\right) = Var(\bar{y}) + (X_v-\bar{x})^2 \cdot Var\,\hat{\beta}_1 + 2\,Cov(\bar{y}, \hat{\beta}_1(X_v-\bar{x})) \quad (Cov \to 0)$$

$$= \frac{\hat{\sigma}^2}{n} + (X_v-\bar{x})^2 \cdot \frac{\hat{\sigma}^2}{S_{xx}}$$

Predict Interval (for "this" X_v's prediction)

"Predict y" for given X_v

$$\hat{y} \pm t_{\frac{\alpha}{2}} \cdot \sqrt{SE^2(\hat{\beta}_1)\cdot(X_v-\bar{x})^2 + \frac{S_\varepsilon^2}{n} + S_\varepsilon^2}, \quad S_\varepsilon^2 = \hat{\sigma}^2$$

df : $n-2$

P.I.

$$y^* = \beta_0 + \beta_1 X_v + \varepsilon = E(y \mid X_v) + \varepsilon \quad (\varepsilon \text{ let} = 0)$$

(though $\varepsilon = 0$, but we still should consider the effect to Variance of Error (ε))

⇒ Predict y^*: C.I.

$$Var\,y^* = Var(\hat{E}(y \mid x_{n+1})) + Var(\varepsilon)$$

$$= \frac{\hat{\sigma}^2}{n} + (X_v-\bar{x})^2 \cdot \frac{\hat{\sigma}^2}{S_{xx}} + S_\varepsilon^2$$

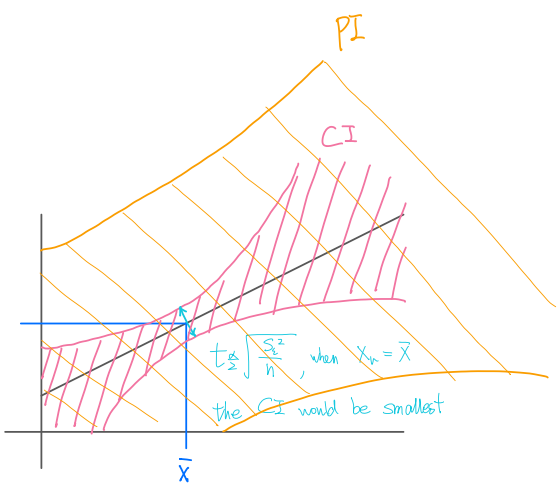
X_v's effection to the Interval

if $|X_v - \bar{x}|$ ↘ then CI ↖

- $CI = \hat{y} \pm t_{\frac{\alpha}{2}} \cdot \sqrt{SE^2(\hat{\beta}_1)\cdot(X_v-\bar{x})^2 + \frac{S_\varepsilon^2}{n}}$

  if $X_v = \bar{x}$, then $CI = \hat{y} \pm t_{\frac{\alpha}{2}} \sqrt{0 + \frac{S_\varepsilon^2}{n}}$

- $PI = \hat{y} \pm t_{\frac{\alpha}{2}} \sqrt{SE^2(\hat{\beta}_1)\cdot(X_v-\bar{x})^2 + \frac{S_\varepsilon^2}{n} + S_\varepsilon^2}$



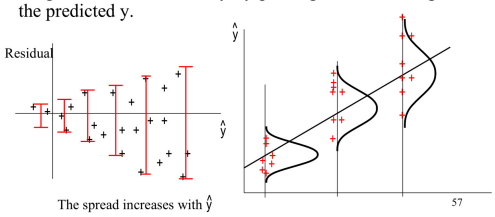
Regression Diagnostics

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ ← is correct?

- ε_i is N-Dist ? ($\varepsilon_i \sim N()$?)
- $\forall x_i$, $\varepsilon_i \in$ const ? ($x_i \perp \varepsilon_i$?)
- $\forall \varepsilon_i$ are indep. ? ($\perp_i \varepsilon_i$?)

Residual Analysis

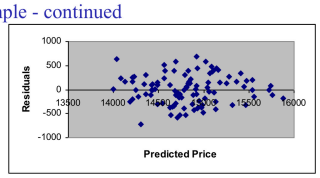
- When the requirement that SD(ε) is the same for all x is violated, we have a condition called heteroscedasticity.
- Diagnose heteroscedasticity by plotting the residual against the predicted y.



57

Homoscedasticity

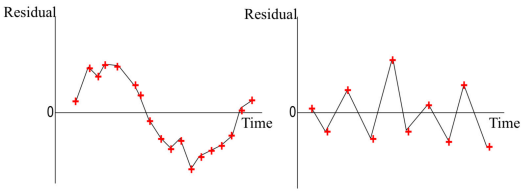
- When the requirement that SD(ε) is the same for all x is not violated we have a condition of homoscedasticity.
- Example - continued



Non-Independence of Error Variables

- **A time series** is constituted if data were collected over time.
- Examining the residuals over time, no pattern should be observed if the errors are independent.
- When a pattern is detected, the errors are said to be autocorrelated.
- Autocorrelation can be detected by graphing the residuals against time.

Patterns in the appearance of the residuals over time indicates that autocorrelation exists.



Note the runs of positive residuals, replaced by runs of negative residuals

Note the oscillating behavior of the residuals around zero.