Non-Parametric tests :

(Parametric Tests) – Covered so far

Test-statistic

– $\dfrac{\bar{x}-c}{\sigma/\sqrt{n}}$; $\dfrac{\bar{x}-c}{s/\sqrt{n}}$; ....

– to perform the test assumption of Normal distribution or by assuming C.L.T. Holds → $\boxed{\text{Sample size being large}}$

Non-parametric tests :-

$\boxed{\text{No need to Device a test for normality}}$

– Distribution free tests

– less powerful than parametric tests

Power of test :- next statistics course

– Very helpful for a start.

Example :- – Vaccine has been developed

Q :- Is the vaccine effective or not? (for a disease)

A :- Designed an experiment

– choose n individuals from a population

- gave vaccine to n_1 of them
  placebo to $n_2 := n - n_1$

	Affected	Not affected	Total
vaccine	X_{11}	X_{12}	n_1
placebo	X_{21}	X_{22}	n_2

$\tilde{Q}$: Based on table can we answer the question Q?

Example 2 :- [Tea-Tasting Example]

A - claimed :- Can tell from tasting a cup of Tea whether :-

English preparation

- milk first and Tea next
- Tea first and milk next

Deviced an Experiment :-

- Prepared 8 cups of Tea

4 milk first

$\boxed{4}$ Tea first

- Person A tasted each one of them and gave opinion.

	Tea	Milk	Total
Tea first	3	1	4
Milk first	1	3	4

Approach 1 (Parametric) : χ^2 - test for independence.

Approach 2 :-

Experiment:

	Tea	Milk	Total
Tea first	$3 \equiv X_{11}$	1	4
Milk first	1	3	4

- Row totals are fixed by the experimenter
- Column totals are decided by person A

H_0 :- Person A has No ability
(i.e. Person A calls "tea first" by choosing a random Sample from 8)

Under H_0 :-

$\mathbb{P}(X_{11} = 3) = \mathbb{P}($ choose 4 cups from 8 cups & 3 of them are actually Tea 1st $)$

$$(\text{Hypergeometric}) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.229$$

Test p-value :- $\mathbb{P}(X_{11} \geq t_0)$ observed

Specify level of significance := $\alpha = 0.05$

Here $t_0 = 3$:- Assume H_0 is true

$$\mathbb{P}(X_{11} \geq 3) = \mathbb{P}(X_{11} = 3) + \mathbb{P}(X_{11} = 4)$$

Under H_0 :- $X_{11} \sim$ Hypergeometric $(N = 8, G = 4, n = 4)$

$$\mathbb{P}(X_{11} \geq 3) = 0.229 + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}}$$

$$= 0.229 + 0.014$$

$$= 0.243$$

$\therefore$ As $\mathbb{P}(X_{11} \geq 3) = 0.243 \gg \alpha = 0.05$

$\therefore$ the null cannot be rejected.

- there is not enough evidence to reject the null hypothesis that person A was purely guessing

---

Sign test & Signed Rank tests

Model: $X_1, \dots, X_n$ are from a random sample

$$X_i = \theta + e_i$$

Errors independent with p.d.f $f(\cdot)$ symmetric around 0.

$H_0: \theta = 0 \qquad H_A: \theta > 0$

Test statistic

$$S = \sum_{i=1}^{n} sgn(X_i)$$

where $sgn(t) = \begin{cases} -1 & t < 0 \\ 0 & t = 0 \\ 1 & t > 0 \end{cases}$

Sign test

- look at positive observations

$$S^+ = \#\{i : X_i > 0\}$$

- ignore 0's and sample is reduced.

- If we observe -1 and 1 then

X_i - can be thought as Bernoulli ($\frac{1}{2}$)

$\Rightarrow$ $S^+ \sim$ Binomial distribution $(n, \frac{1}{2})$ $\equiv$ Null distribution

<u>Test</u> :- Compare this distribution with the observed statistic

<u>Example (Sign test)</u> :-

12 people are chosen

10 prefers shorts over full pants $(X_i > 0)$
1 prefers full pants over shorts $(X_i < 0)$
1 no preference $(X_i = 0)$

- strong preference for shorts ?

How likely is such a result true if

H_0: there is no preference over shorts or full pants

is true ?

$S^+ = \{i : X_i > 0\} \sim Bin(n=11, p=\frac{1}{2})$

$s^+ \equiv$ observed $= 10$

p-value :- $\mathbb{P}(S^+ \geq 10) \cong 0.0059$

α - level of significance :- 0.05

$\Rightarrow$ Since $0.0059 << 0.05$

H_0: No preference H_A: shorts over full

we reject null hypothesis.

Signed Rank - Wilcoxon Test

(Compare with t-test)

$H_0: \mu = 0$, $H_A: \mu > 0$

Not distribution free

<u>Test-statistic</u>. $t_0 := \frac{\bar{X}}{S/\sqrt{n}} \sim t_{n-1}$

p-value : $\mathbb{P}(t_{n-1} > t_0)$

$X_1, X_2, \ldots\ldots, X_n$ — sample $\begin{cases} X_i = \theta + e_i \\ \text{(old model)} \end{cases}$

$$W = \sum_{i=1}^{n} \operatorname{Sgn}(X_i) \operatorname{Rank}(|X_i|)$$

<u>Test-statistic</u>

$$W^+ = \sum_{i=1}^{n} \operatorname{Rank}(|X_i|) \mathbb{1}(X_i > 0)$$

W^+ symmetric around $\frac{n(n+1)}{4}$

$$\overset{Ex}{=} \frac{1}{2}\left(W + \frac{n(n+1)}{2}\right)$$

- There is no closed form for distribution of W^+

$H_0: \theta = 0$ versus $H_A: \theta > 0$

p-value: $\mathbb{P}(W^+ > t_0)$

t_0 — observed test statistic

Recall - clean up

$X_1, X_2, \dots, X_n$ are i.i.d Bernoulli (n, p)

$\Rightarrow$ $S_n \sim \text{Binomial}(n, p)$

[CLT] $$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1)$$

Approximation - Application : $a < b \in \mathbb{R}$

$$\mathbb{P}(a < S_n \leq b) = \mathbb{P}\left(\frac{a - np}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{b - np}{\sqrt{np(1-p)}}\right)$$

$$\simeq \mathbb{P}\left(\frac{a - np}{\sqrt{np(1-p)}} < Z \leq \frac{b - np}{\sqrt{np(1-p)}}\right)$$

when $Z \sim N(0,1)$

- when approximation is valid
  - $np > \frac{1}{2}$ $n(1-p) > \frac{1}{2}$
  - $n >>>$ large

To make approximation precise / more accurate

- Continuity Correction

Example :-

$n = 20, \quad p = \frac{1}{2} \qquad S_n \sim \text{Binomial}(20, \tfrac{1}{2})$

$$\mathbb{P}(8 \leq S_n \leq 10) = \mathbb{P}\left(\frac{8 - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{10 - np}{\sqrt{np(1-p)}} \right)$$

Central limit Theorem approximation

$$\simeq \mathbb{P}\left(\frac{8 - np}{\sqrt{np(1-p)}} < Z \leq \frac{10 - np}{\sqrt{np(1-p)}} \right)$$

$$\simeq \mathbb{P}\left(\frac{8 - 10}{\sqrt{5}} < Z \leq \frac{10 - 10}{\sqrt{5}} \right)$$

$$= \Phi(0) - \Phi\left(-\frac{2}{\sqrt{5}}\right)$$

$$= \frac{1}{2} - \ldots = 0.3145$$

Table or R

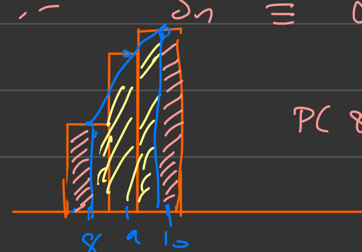
approximation is not great

$$\mathbb{P}(8 \leq S_n \leq 10) = \sum_{k=8}^{10} \binom{20}{k} \left(\frac{1}{2}\right)^{10}$$

Direct Computation

$$= 0.4565$$

Cause for Bad Approximation :- $S_n \equiv$ discrete and Z is continuous



$P(8 \leq S_n \leq 10)$

Approximation – Application

with Continuity Correction : $a < b \in \mathbb{R}$

$$\mathbb{P}(a < S_n \le b) = \mathbb{P}\left(\frac{a - np}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{b - np}{\sqrt{np(1-p)}}\right)$$

$$\simeq \mathbb{P}\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} < Z \le \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

where $Z \sim N(0,1)$

$$\mathbb{P}(8 < S_n \le 10) = \mathbb{P}\left(\frac{8 - np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{10 - np}{\sqrt{np(1-p)}}\right)$$

Central limit Theorem approximation with Continuity correction

$$\simeq \mathbb{P}\left(\frac{7.5 - np}{\sqrt{np(1-p)}} < Z \le \frac{10.5 - np}{\sqrt{np(1-p)}}\right)$$

$$\simeq \mathbb{P}\left(\frac{7.5 - 10}{\sqrt{5}} < Z \le \frac{10.5 - 10}{\sqrt{5}}\right)$$

$$\Phi\left(\frac{0.5}{\sqrt{5}}\right) - \Phi\left(\frac{-2.5}{\sqrt{5}}\right)$$

$$= 0.4567$$

This is indeed a better approximation!

Continuity correction

$X_i : i = 1, \dots n$ are discrete

$$S_n = X_1 + X_2 + \dots + X_n$$

Continuity correction

$$P(a \le S_n \le b) = \mathbb{P}\left(a - \tfrac{1}{2} \le S_n \le b + \tfrac{1}{2}\right)$$

Histogram
or if $a, b \in \mathbb{N}$

$\overset{C.L.T.}{=}$ ... approximation ...

Boot strap — Jack-knife :- [Re sampling techniques]

Efron -

- Parameter of interest $\theta \in \mathcal{P}$
- Estimate the parameter — Method of moments / Maximum likelihood Estimate
- Confidence interval ; - Hypothesis test

z-test, ... : parametric
(Normal approximation)

Setting :- Normal approximation does not apply

Random Sample:. $X_1, \dots, X_n \sim F$ ($F \equiv$ distribution of population)

θ - is parameter of interest

Compute :- $\hat{\theta} := g(X_1, X_2, \dots, X_n)$ is an estimate for θ.

$\hat{\theta} \sim G_n$ (Some sampling distribution)

G_n :- Unknown If Normality holds $\Rightarrow G_n \sim N(\cdot,\cdot)$

(Ideal) Find as many datasets (Samples) (B) from G_n

$\boxed{\text{WISH} \quad \text{more samples of } G_n.}$

then $\hat{\theta}_1, \dots, \hat{\theta}_B \overset{i.i.d}{\sim} G_n.$

$B >>$ large $\equiv$ not possible

Once we have B realisations from G_n

$\longrightarrow$ estimate many characteristics of G_n

Bootstrap :- - Our current data set is

$x_1, x_2, \ldots, x_n$

Make it precise

- Resample repeatedly from dataset with replacement.

(Hope...) - approximate sample from G_n. & using this we could estimate characteristics of the distribution of G_n

Non-parametric Bootstrap :-

- $\{x_1, \ldots, x_n\} \equiv S$ - Random sample from population.

$\hat{\theta} := g(x_1, \ldots, x_n)$

- Resample from S WITH replacement a sample of size n.

Repeat B time

$$X^*_{11}, \ldots, X^*_{n1} \Longrightarrow \hat{\theta}^*_1$$

$$X^*_{12}, \ldots, X^*_{n2} \Longrightarrow \hat{\theta}^*_2$$

$$\vdots$$

$$X^*_{1B}, \ldots, X^*_{nB} \Longrightarrow \hat{\theta}^*_B$$

$\boxed{\text{B Estimators}}$

(Ex.) $\cdot$ $\hat{\theta}_1^*, \dots, \hat{\theta}_B^* \approx G_n$

$\cdot$ $\widehat{\mathrm{Var}\,\hat{\theta}} = \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^* \right)^2$

Boot strap Estimate

$\overline{\hat{\theta}}_B$ (Sample mean)

$\cdot$ Boot strap Confidence intervals

(order) $\hat{\theta}_{(1)}^* < \dots < \hat{\theta}_{(k)}^* < \dots \hat{\theta}_{(B)}^*$

$L = \hat{\theta}_{\lfloor \frac{\alpha}{2} B \rfloor}$ $\quad U = \hat{\theta}_{\lfloor (1-\frac{\alpha}{2}) B \rfloor}$

$\Rightarrow$ (L, U) is $100(1-\alpha)\%$ Boot strap Confidence interval for θ

Parametric Boot strap

$X_1, \dots, X_n \sim F(\theta)$

unknown

Known distribution

$S = \{X_1, \dots, X_n\}$

set $\hat{\theta} = g(X_1, X_2, \dots, X_n)$

- Resample from $F(\hat{\theta})$ WITH replacement a sample of size n.

Repeat B time

$$X^*_{11}, \ldots, X^*_{n1} \sim F(\hat{\theta}) \Longrightarrow \hat{\theta}^*_1$$

$$X^*_{12}, \ldots, X^*_{n2} \sim F(\hat{\theta}) \Longrightarrow \hat{\theta}^*_2$$

$$\vdots$$

$$X^*_{1B}, \ldots, X^*_{nB} \sim F(\hat{\theta}) \Longrightarrow \hat{\theta}^*_B$$

$\boxed{B \text{ Estimators}}$

$$\boxed{\frac{1}{B}\sum_{i=1}^{B} \hat{\theta}^*_i := \overline{\hat{\theta}^*}}$$

- Estimate for $\hat{\theta}$

- Same procedure for confidence interval

Example :- Population with distribution $\sim \text{Gamma}(\alpha, 1)$

Sample :- $X_1, \ldots, X_n$ from population

$$\hat{\theta} := \overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$ is an estimate for α.

Boot strap

Non-parametric :- Resample from $S = \{X_1, \ldots, X_n\}$

Parametric :-

$$X^*_{11}, \dots, X^*_{n1} \sim \text{Gamma}(\bar{X}, 1) \longrightarrow \bar{X}^*_1$$

$$X^*_{12}, \dots, X^*_{n2} \sim \text{Gamma}(\bar{X}, 1) \longrightarrow \bar{X}^*_2$$

$$\vdots \qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots$$

$$X^*_{1B}, \dots, X^*_{nB} \sim \text{Gamma}(\bar{X}, 1) \qquad \bar{X}^*_B$$

$\frac{\alpha}{2}$, $\frac{1-\alpha}{2}$ quantiles of $\bar{X}^*_i$ $\Rightarrow$ Confidence intervals

Central limit Theorem

[check : mean & variance of Gamma$(a,1)$)

$$\frac{\sqrt{n}(\bar{X} - a)}{\sqrt{a}} \xrightarrow{d} N(0, 1)$$

Confidence interval :

$$\left(\bar{X} - z_{\frac{(1-\alpha)}{2}} \sqrt{\frac{\bar{X}}{n}}, \; \bar{X} + z_{\frac{(1-\alpha)}{2}} \sqrt{\frac{\bar{X}}{n}}\right)$$

4 Comparison :-

R code

- non parametric distribution
- Parametric distribution
- Normal distribution approximation
- True sampling distribution

Questions :

- In R-code ≡ $B = 1000$?
  - check effect of change in B.

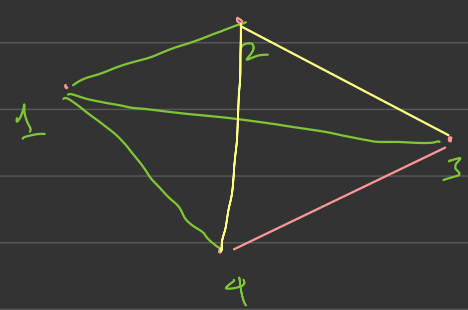
- Estimate median of Gamma ?

- Estimate median − mean of Gamma ?

HW 12

Erdös Renyi Graphs

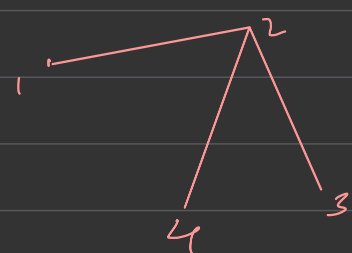
{ Probability
Computer science
Real world networks



$K_4 \equiv K_n$

w.p p − keep an edge

$1-p$ − drop an edge



$G(4, p)$

$\equiv G(n, p)$

Estimate p from the graph