DATE
PAGE

Ch 21 Spearman Rank Correlation Coefficient

Monotonically non decreasing function $x_1 > x_2 ; y_1 \geq y_2$.

Monotonically increasing function
$x_1 > x_2 ; y_1 > y_2$
(strictly)

It is for non-linear relationship.

	x (Height)	y (weight)	Rank$_x$	Rank$_y$
s_1	160	52	4	3
s_2	150	66	2	4
s_3	170	68	5	5
s_4	140	46	1	1
s_5	158	51	3	2

$r = f(\text{Rank}_x, \text{Rank}_y)$

smaller element gets rank 1 and so on.

DATE
PAGE

$\rho = 1$ if linear $X \uparrow$ $Y \uparrow$ linear or not $r = 1$

$\rho = -1$ if linear $X \uparrow$ $Y \downarrow$ " " " $r = -1$

Ch 22 Correlation doesn't imply causation

Ch 23 How to use Correlations?

Real life questions :-

1. Is salary correlated with square footage of your home?
2. Is no. of years of education correlated with income?
3. E-Commerce

Is time spent in 24 hours on amazon.com correlated with money spent in the next 24 hours?

The amazon can use this data to design their website for more sale.

DATE
PAGE

(4) Medicine

Dosage of a drug correlated with reduction in blood sugar.

Ch24 Confidence Interval Introduction

Instead of giving point estimate we give a population mean will lie in an interval with some probability is called Confidence Interval.

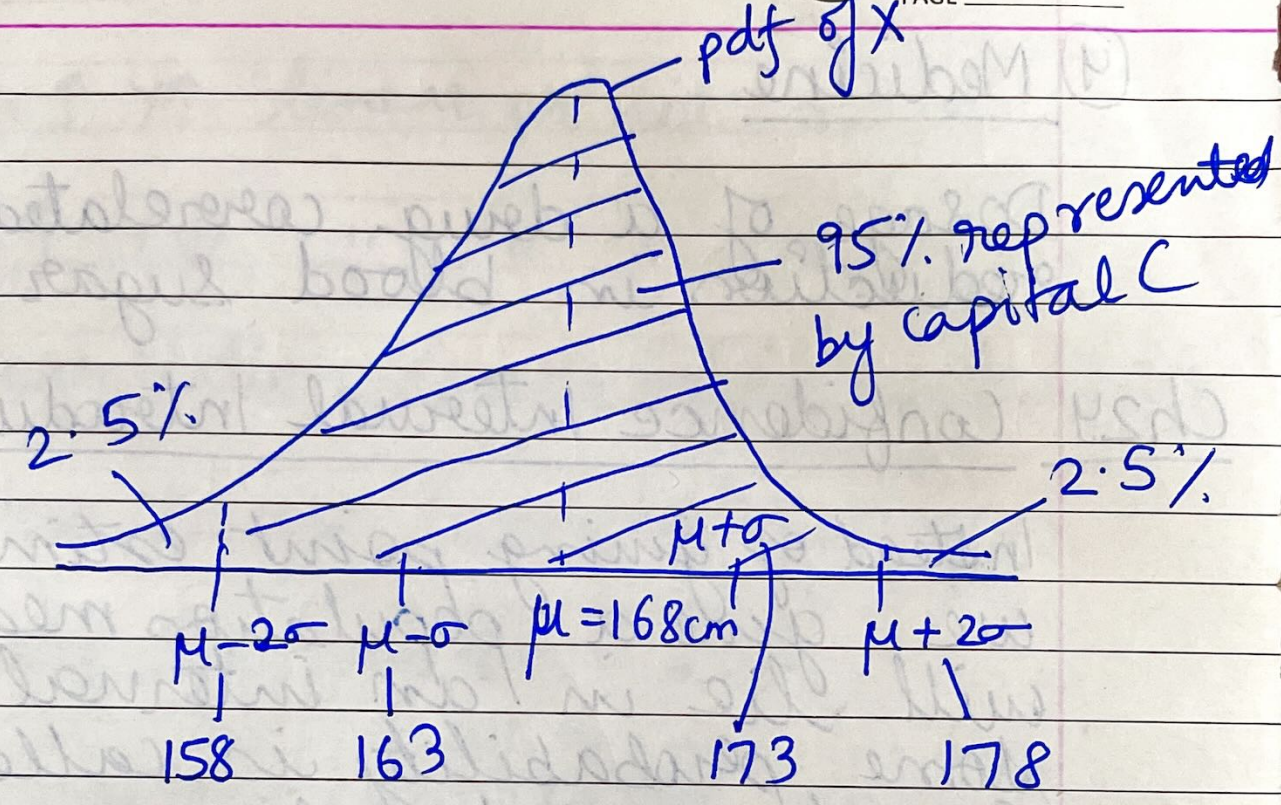
eg- $\mu \in [162.1, 174.9]$ with 95% prob.

Ch25 Computing Confidence Interval given the underlying distribution

Let $X \sim N(\mu, \sigma)$, where μ be 168 cm
σ be 5 cm

X represents heights

DATE _____
PAGE _____



Confidence interval is represented by C.

Q What is the interval with 95% confidence.

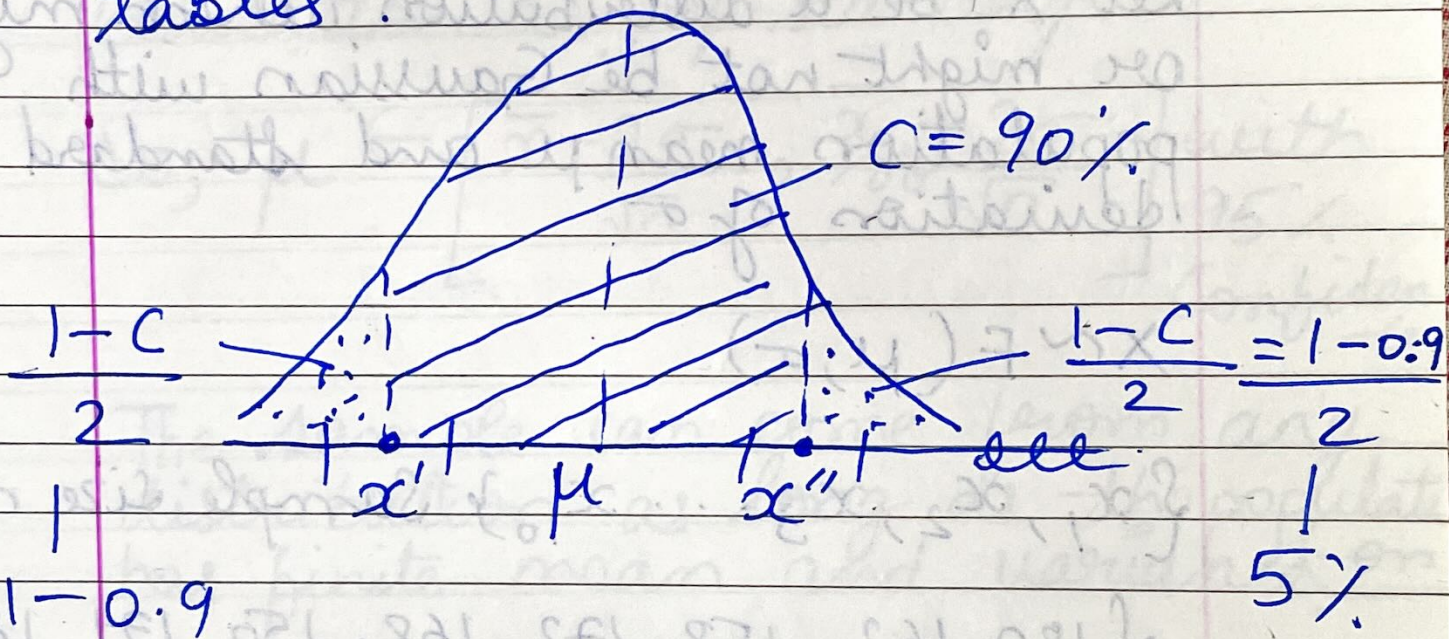
$(\mu-2\sigma, \mu+2\sigma)$ ∴ 95% of my observations lie here.

(158, 178) with 95% probability.

DATE _____

PAGE _____

To calculate confidence interval of $C = 90\%$ we use something called as Normal distribution tables.



So, They lie in $[x', x'']$ with 90% confidence.

DATE _____
PAGE _____

Ch 26 C.I. for mean of a normal rando variable

Let X be a distribution which might or might not be Gaussian with population mean μ and standard deviation of σ

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

$\{x_1, x_2, x_3 \dots x_{10}\}$ Sample size $n = 10$

{180, 162, 158, 172, 168, 150, 171, 183, 165, 176}

Q Compute the 95% Confidence Interval of μ.

Case 1 We are given $\sigma = 5$cm

According to the Central Limit Theorem,

$\bar{x}$ = Sample mean $= \frac{1}{10} \sum_{i=1}^{10} x_i$ $n = 10$

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

DATE
PAGE

In words, the sample mean follows Gaussian distribution having population mean μ and standard deviation as pop-std-dev/$\sqrt{n}$

So, $\mu \in \left[\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}\right]$ with 95% Confidence

The sample can come from any distribution as long as the population has finite mean and variance.

Case 2 When we don't know the population standard deviation.

We use student's t-distribution with n-1 degrees of freedom where n is the size of the sample.

In this as the degrees of freedom increases, the peak of pdf also gets higher and higher.

DATE _____
PAGE _____

Ch 27 Confidence Interval using bootstrapping

Suppose our random sample X has some distribution about which we don't know what type of distribution.

We have a sample of size n

$S = \{x_1, x_2, x_3, x_4 \ldots\ldots x_n\}$ Let $n = 10$

Using only this sample we will compute Confidence Interval of median of X

The sampling is done with replacement. For choosing the values we can use Uniform distribution

Note As n ↑, C.I of median of x become more clear.

So, with uniform random sample $U(1, n)$ we generate samples with repetition.

DATE _____
PAGE _____

S_1 $x_1^{(1)}, x_2^{(1)}, x_3^{(1)} \ldots x_m^{(1)}$ such that $m \leq n$.

m_1 — median

This is random sample of size m generated from S

We make many samples using the same methodology called bootstrap samples.

$S = \{x_1, x_2, \ldots x_n\}$

Now we compute median of these bootstrap samples

S_2 $x_1^{(2)}, x_2^{(2)}, x_3^{(2)} \ldots x_m^{(2)}$ — m_2

S_3 $x_1^{(3)}, x_2^{(3)}, x_3^{(3)} \ldots x_m^{(3)}$ — m_3

S_K $x_1^{(K)}, x_2^{(K)}, x_3^{(K)} \ldots x_m^{(K)}$ — m_K

For calculation of median of

DATE
PAGE

sample, first the medians of the bootstrap samples are made.

Then the medians are sorted in ascending order.

Then the confidence interval is calculated, here m'_{25} and m'_{975} are the 95 percent CI for median of the sample.

$M_1, M_2 \ldots\ldots M_{1000}$ 1000 medians generated using bootstrap samples.

$\downarrow$

$M'_1 \leq m'_2 \leq m'_3 \leq \ldots\ldots \leq m'_{1000}$ (increasing order)

$\downarrow$ CI(95%)

$-25- \leq m'_{25} \ldots -950- \ldots m_{975} \geq 25-$

Similar calculation is made for calculating the variance, mean and standard deviation

This method of calculation is called non parametric technique which does not make any assumptions about the data.

Ch 28 Hypothesis testing methodology, Null hypothesis, p-value

1) choosing a test statistic
Difference of mean between two distributions. $\mu_2 - \mu_1$

2) Define a null hypothesis (H_0) – Assumption
(Proof by contradiction.)
no diff in μ_1 & μ_2
No difference b/w heights

Alternative hypothesis (H_1) Inverse of null hypothesis!
Diff in μ_1 & μ_2

HTAB.

Proof by contradiction

DATE
PAGE

1. Assume null hypothesis is true and I find it is true with very high probability and hence I accept it.

2. Assume my null hypothesis is true and I prove it is incorrect and I accept my alternative hypothesis

p-value

Probability that observations test statistic if my null hypothesis is true.

↓

it could be any entity or diff b/w two means.

eg- Diff b/w heights of students in 2 class.

if pvalue is given by someone is 0.9

* prob of 10cm diff is 0.9 if H_0 is true

So accept H_0

DATE
PAGE

* prob if pvalue is 0.05 is 5% chance that 10cm diff if H_0 is true.

pvalue is small so diff 10cm. probability is less so we reject H_0.

Ch 29 Hypothesis Testing Intuitions with coin toss example.

Eg1 Given a coin, determine if the coin is biased towards heads or not.

$$\begin{cases} \text{biased coin } P(H) > 0.5 \\ \text{non-biased coin } P(H) = 0.5 \end{cases}$$

Experiment — Flip a coin 5 times and count # heads = X ← Test statistic

↑ random variable

f_1, f_2, f_3, f_4, f_5

H H H H ✗ $X = 5$

conditional probability

DATE _____

PAGE _____

$P(X=5 \mid \text{coin is not biased towards head})$

assumption

↓

Null hypothesis (H_0)

H_0 = coin is not biased towards heads.

can also be written as $p(obs \mid H_0)$

$P(X=5 \mid H_0) = \frac{1}{2^5} \Rightarrow \frac{1}{32} \approx 0.03 = 3\%$

($X=5$: five heads in five tosses)

(H_0: coin is not biased towards heads)

$P(H) = \frac{1}{2} = 0.5$

f_1, f_2, f_3, f_4, f_5

$\frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{2^5}$

DATE
PAGE

There is a 3% chance of getting 5 heads in 5 flips if the coin is not biased towards heads

hypothesis testing =

$p(\text{obs by expt} \mid \text{assumption}) = 3\%$

If $p(\text{obs} \mid H_0) < 5\%$ then H_0 may be incorrect.

Conclusion — assumption or H_0 is not true.

↓

reject H_0 ⇒ reject coin is not biased towards heads

↓

accept coin is biased

↓

alternative hypothesis (H_1)

Note: Rejecting H_0 → Accepting H_1

" H_1 → " H_0.

DATE _____
PAGE _____

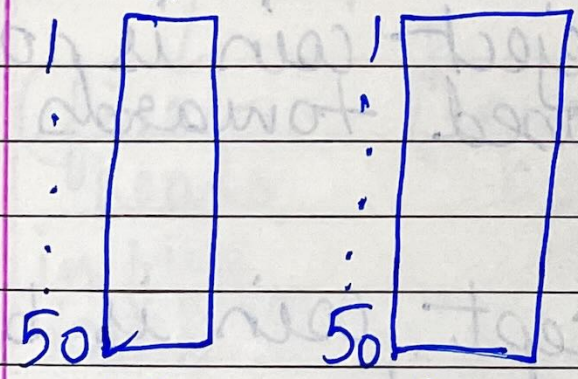
Note Change in Sample size may alter our probability so we may accept instead of rejecting my H_0.

Be careful about :—

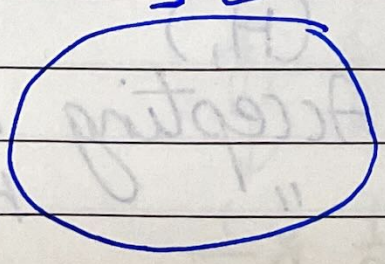
1. Design the experiment correctly.
2. Design the null hypothesis correctly.
3. Design the test statistic carefully

Ch 30 Resampling and Permutation Test

eg- Cl1 Cl2



M_1 M_2 Diff → Δ

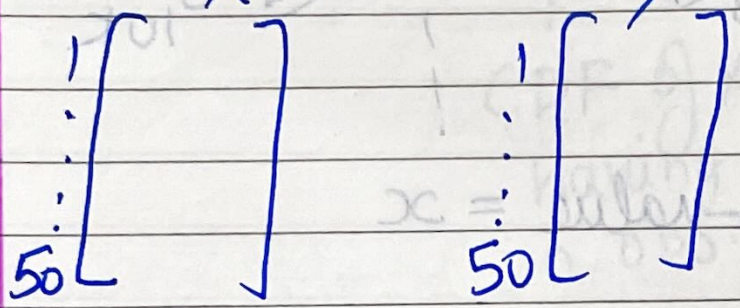


Jumble

DATE
PAGE

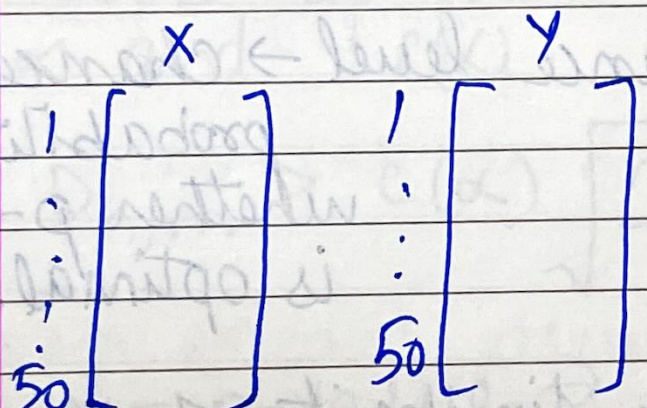
Randomly sample 50 points

① X vector Y vector



$\mu_1 - \mu_2 \Rightarrow \delta_1$

②



$\mu_1 - \mu_2 \rightarrow \delta_2$

10K 10K — δ_{10K}

$\Delta = \mu_{cl1} - \mu_{cl2}$

DATE _____
PAGE _____

Sort δ

$x\%$

$\delta_1' \leq \delta_2' \leq \delta_3' \leq \ldots \ldots \leq \delta_{10K}'$

p-value = x

By jumbling I am assuming H_0.

5% Significance level → chance/probability whether p-value is optimal.

Ch 3) K-S Test for Similarity of two distribution

First plot CDF (cumulative prob.) of X_1 and X_2

$H_0 \rightarrow X_1$ & X_2 have same distributio

$(D_{n,m})$ Test statistic → Difference of CDF of X_1 & X_2 on the plot.

Test statistic — supremum — maximal diff.

$$D_{n,m} = \sup |F_{1,n}(x) - F_{2,m}(x)|$$

$F_{1,n}(x)$: CDF of X_1 having n obs.

$F_{2,m}(x)$: CDF of X_2 having m obs.

Look up table have value of α and $c(\alpha)$.

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$$

α - p-value $\Rightarrow$ Significance level.

If $D_{n,m}$ is large than $c\alpha\sqrt{\frac{n+m}{nm}}$ then we reject null hypothesis at that Significance level.

DATE _____
PAGE _____

Ch32 Kolmogorov-Smirnov Test Code.

Ch 33 Hypothesis Testing - Another example.

Difference of means

Task. Determine if the population means of heights of people in these two cities is same or not.

City 1 μ_1 City 2 μ_2

Find μ_1 & μ_2 are same/differe

So compute with sample mean instead of population mean.

Experiment - Sample heights of 50 peopl in City 1 and City 2.

randomly C_1 $\begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{50} \end{bmatrix}$ C_2 $\begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{50} \end{bmatrix}$ — Sample he of 50 peo

DATE
PAGE

μ_1 = Sample mean of 50 people

$$\mu_1 = \frac{h_1 + h_2 + \cdots + h_{50}}{50} = \text{observed value} \; (167 \text{ cm})$$

$$\mu_2 = \frac{h'_1 + h'_2 + \cdots + h'_{50}}{50} = \text{observed value} \; (162 \text{ cm})$$

Test Statistics $= |\mu_1 - \mu_2| = x = |162 - 167| = 5\text{cm}$

Null Hypothesis (H_0) = There is no difference in population means

Compute : $p(x = 5\text{cm} \mid H_0)$

Diff in Sample means with Sample size of 50.

Now, $p(x = 5\text{cm} \mid H_0)$ is interpreted as:-
Probability of observing a difference of 5cm in Sample mean heights of Sample

Sample mean ~~of~~ probability 20% ~~of~~ ~~that~~ ~~it~~ is given H_0 which is → "There is no difference in population mean of 2 cities". So $P(x = 5cm | H_0) = 20\%$. ~~5%~~, ~~Accept H_0~~

DATE
PAGE

Size 50 between C_1 & C_2 if there is no population difference in mean heights.

Case 1 $p(x = 5cm | H_0) = 0.2 = 20\%$

There is a 20% chance of observing a difference of 5cm in sample mean heights of C_1 & C_2 with sample of 50 if there is no population mean difference.

20% is a significant amount of probability

Assumption must be true.

Accept H_0

Case 2 $p(x = 5 | H_0) = 0.03 = 3\%$

$p(\text{obs} | \text{assumption}) = 3\%$ — Small (< 5%)

⇒ Assumption must be incorrect. Reject H_0. Accept H_1. Population mean are not same.

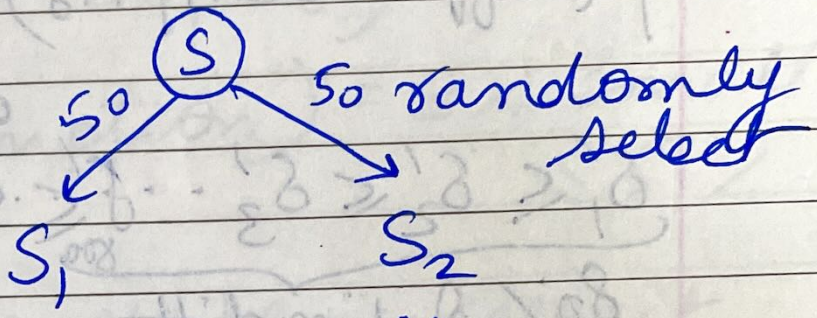
DATE
PAGE

Ch 34 Resampling and Permutation Test: Another Example

Step 1 ⇒ $S = \{h_1, h_2, h_3 \ldots h_{50}, h_1', h_2' \ldots h_{50}'\}$

$C_1 \cup C_2$

Step 2 ⇒



Resampling is to simulate the H_0.

(1) $\mu_2 - \mu_1 = 3cm$ δ_1

repeat (2) $\mu_2 - \mu_1 = -2cm$ δ_2

Let $k = 1000$

(k) times $\mu_2 - \mu_1 \longrightarrow 6cm$ δ_k

DATE _____
PAGE _____

Step 3 ⇒ Sort δ_i:

$\delta_1' \leq \delta_2' \leq \delta_3' \cdots\cdots \leq \delta_k'$

Sorted in increasing order.

Case 1 observed differences $x = 5cm$ $(167 - 162)$ cm.

$p(\text{diff} \geq 5cm \mid H_0)$

observed diff (x)

$\delta_1' \leq \delta_2' \leq \delta_3' \cdots \delta_{800}' \leq 5cm < \delta_{801}' \leq \delta_{1000}'$

80% of sim diff $\leq 5cm$ (Observed diff)

20% of simulated diff $> 5cm$

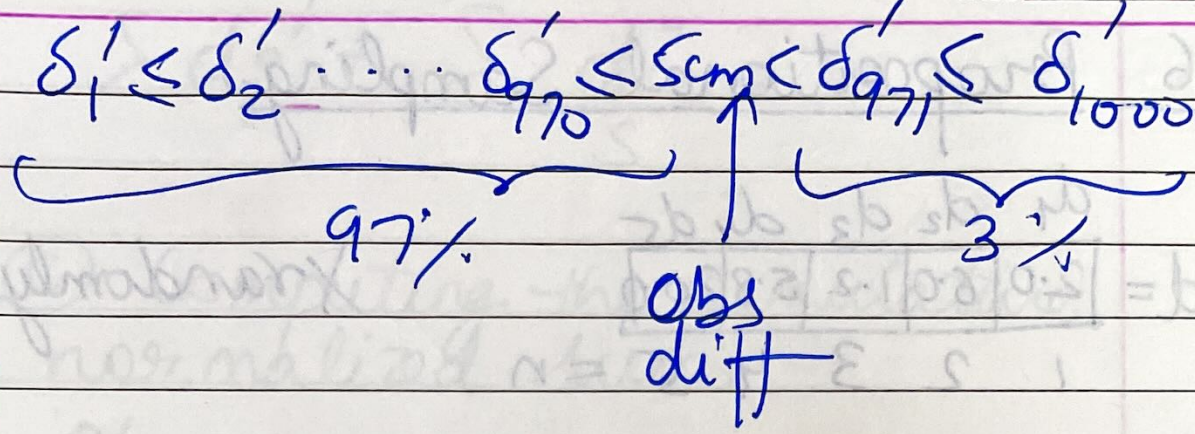
p-value$(\text{obs diff} \mid \text{assumption}) = 20\%$ - Significant

$x \geq 5cm$ H_0 → 75%

Assumption must be true. So accept H_0.
Observation cannot be False.

DATE _____
PAGE _____

Case 2



$p(\text{obs-diff} \geq 5cm \mid H_0) = 3\%$

$p(\text{obs} \mid \text{assumption}) = 3\%$ Small $\leq 5\%$

assumption must be incorrect.
Reject H_0. Accept H_1.

Ch 35 How to use Hypothesis Testing?

α value in medical Hypothesis testing is 1% or even 0.1%

Drug Designers use Hypothesis Testing like this.

DATE
PAGE

Ch 36 Proportional Sampling

	d_1	d_2	d_3	d_4	d_5
$d =$	2.0	6.0	1.2	5.8	20.0
	1	2	3	4	5 = n

X randomly

Task Pick an element amongst the n element s.t. prob. of picking an element is proportional to d_i's.

Step 1 ⓐ $S = \sum_{i=1}^{n} d_i = 35$ ← Compute the sum

ⓑ $d_i' = d_i / S$

Compute d_i'. Normalizing using the sum.

$d_1' = 0.0571$

$d_2' = 0.171428$

$d_3' = 0.0343$

$d_4' = 0.1657$

$d_5' = 0.5714$

b/w 0 to 1.

all sum to 1.

$$\sum d_i' = \sum \frac{d_i}{S} = 1$$

(c) Computing the cumulative normalized sum.

$\tilde{d}_1 = d_1' = 0.0571$

$\tilde{d}_2 = \tilde{d}_1 + d_2' = 0.228528$

$\tilde{d}_3 = \tilde{d}_2 + d_3' = 0.262828$

$\tilde{d}_4 = \tilde{d}_3 + d_4' = 0.428528$

$\tilde{d}_5 = \tilde{d}_4 + d_5' = 1.00$

$\tilde{d}_i \Rightarrow$ Cumulative normalized sum

Step 2 Sample one value from Unif(0,1)

r = numpy.random.uniform(0.0, 1.0)

let $r = 0.6$

DATE
PAGE

Step 3 Proportional Sampling.

if $r \leq \tilde{d}_1$

return 1

else if $r \leq \tilde{d}_2$

return 2

else if $r \leq \tilde{d}_3$

return 3

else if $r \leq \tilde{d}_4$

return 4

else if $r \leq \tilde{d}_5$

return 5.

Code

DATE
PAGE

1) Random number generator for Uniform Distribution

```
import random
print(random.random())
```

#Loading Iris dataset with 150 points

```
from sklearn import datasets
iris = datasets.load_iris()
d = iris.data
d.shape
```

#Sample 30 points randomly from the 150 point dataset

```
n=150
m=30
p= m/n
print(p)
```

DATE
PAGE

```
sampled_data = [ ]
for i in range(0, n):
    a = random.random( )
    if a <= p:
        sampled_data.append(d[i,:]
len(sampled_data)
```

2) Emperical bootstrap based Confidence Interval

```
import numpy
from sklearn.utils import resample
from matplotlib import pyplot
# load dataset
x = numpy.array([180, 162, 158,
    172, 168, 150, 171, 183, 165, 176]
# Configure bootstrap
n_iterations = 1000
n_size = int( len(x))
```

DATE _____
PAGE _____

```
# run bootstrap
medians = list()
for i in range(n_iterations):
    # prepare train and test sets
    s = resample(x, n_samples=n_size)
    m = numpy.median(s)
    medians.append(m)
# plot scores
pyplot.hist(medians)
pyplot.show()
# Computing Confidence Interval
alpha = 0.95
p = ((1.0-alpha)/2.0) * 100
lower = numpy.percentile(medians, p)
p = (alpha+((1.0-alpha)/2.0)) * 100
upper = numpy.percentile(medians, p)
```

print('%.3f Confidence interval %.1f and %.1f' % (alpha*, lower, upper))

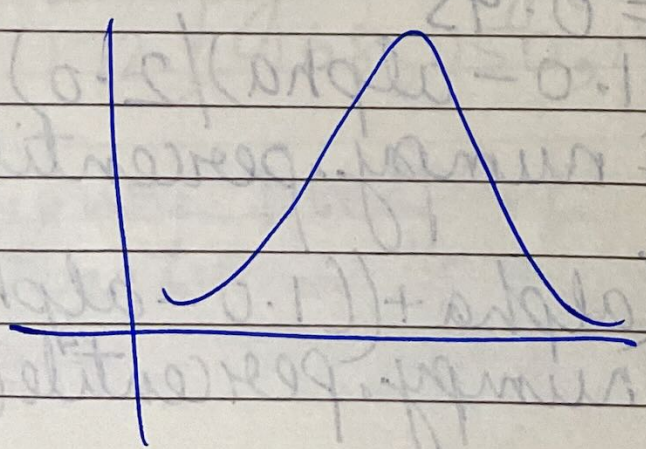
3 KS Test Code

```
import numpy as np
import seaborn as sns
from scipy import stats
import matplotlib.pyplot as plt

# generate a Gaussian r.v. X
x = stats.norm.rvs(size=1000)
sns.set_style('whitegrid')
sns.kdeplot(np.array(x), bw=0.5)
```

Kernel Density Estimation

```
plt.show
```



Stats

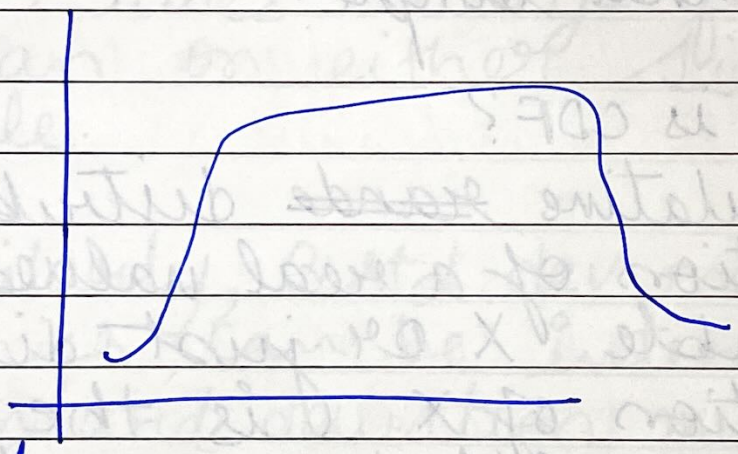
~~sts~~.kstest(x, 'norm')

100,

Continuous Uniform Distributio
(0,1) as Y.

y = np.random.uniform(0,1, 10000)

~~Sns~~.kdeplot(np.array(y), bw=0.1)
plt.show()



~~stats~~.
stats.kstest(y, 'norm')