Probability and Statistics

Rolling a dice — random experiment.

Its outcome is a random variable $X = \{1, 2, 3, 4, 5, 6\}$.

Similarly, tossing a coin. — r.v.

r.v. $Y = \{H, T\}$.

$P(X=1) = \frac{1}{6} = P(X=2)$

$P(X = \text{even}) = \frac{1}{2} = P(X=2) + P(X=4) + P(X=6)$.

$P(X = x_1)$ ⇝ sometimes also written as $\underline{P(x_1)}$.

X is a discrete random variable.

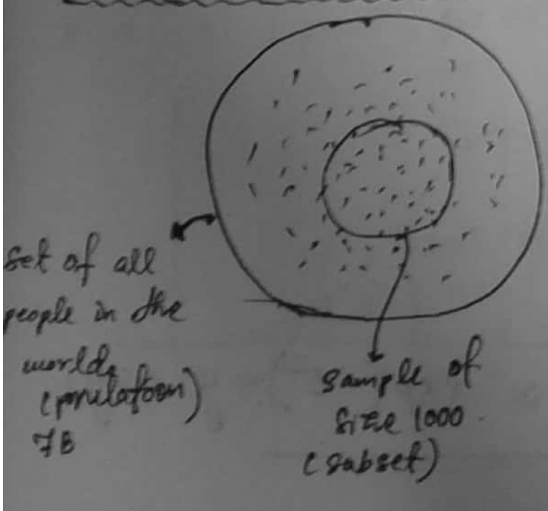
let Z is a random variable ~~of~~ storing outcome of random expt of measuring heights.

So, Z = continuous random variable.

$Z = \{122.2, 146.4, 132.5, \ldots 12.26, 156.23\}$.

$\downarrow$ outlier (maybe error or genuine)

Population & Sample



The mean height of human (μ) $= \frac{1}{7B} \sum_{i=1}^{7B} h_i$

denotes population mean

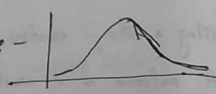
$\bar{h} = \frac{1}{1000} \sum_{i=1}^{1000} h_i$ (sample mean).

As sample size increases,

$\boxed{\bar{x} = \mu}$

Gaussian distribution (Normal distbⁿ).

PDF of a random variable having gaussian distribution :-



Graphical:

X : continuous random variable.

why bother learning about this particular kind of distributⁿ?

→ They occur in nature. Ex → the sepal length & petal length of iris flowers,

Similarly heights & weights of people also have a gauss. distbⁿ

wikis - normal-distbⁿ.

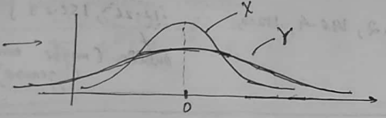
← graph with diff μ & σ.

(μ & σ^2) are parameters of Gaussian distbⁿ.

Knowing only these 2 values enough. (No need of sample datas)

$X \sim N(\mu, \sigma) \Rightarrow$ X is a random variable that follows normal distbⁿ with mean μ & std deviatⁿ σ.

Ex $X \sim N(0, 2)$.
$Y \sim N(0, 4)$ →



Mathematical:

$X \sim N(\mu, \sigma^2)$.

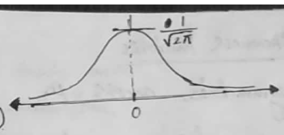
$$\boxed{P(X = x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}}$$

Simplify to analyze, $\sigma = 1$, $\mu = 0$.

$$P(x) = \underbrace{\frac{1}{\sqrt{2\pi}}}_{c_1} \exp\left\{ \underbrace{-\frac{1}{2}}_{c_2} x^2 \right\}$$

$P(x) \approx \exp(-x^2) = y$ (say).

$y = \exp(-x^2)$

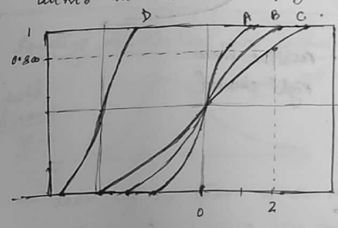


① symmetric
$\exp(-x^2) = \exp(-(-x)^2)$

② as x moves away from μ, y reduces exponentially.

CDF of gaussian distbⁿ

wikis - normal distbⁿ page.



A: $\mu = 0$ $\sigma^2 = 0.5$
B: $\mu = 0$ $\sigma^2 = 1$
C: $\mu = 0$ $\sigma^2 = 2$.
D: $\mu = -4$ $\sigma^2 = 0.4$.

$P(X \le 2) = 0.8$ for C from the CDF.

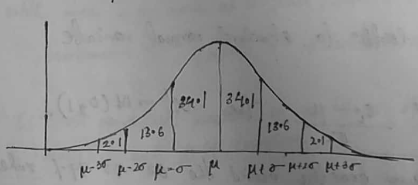
* Since gaussian dist symmetric about mean, in CDF, 0.5 for μ.

$$\boxed{P(X \le \mu) = 0.5}$$

---

68 – 95 – 99 rule

* 50% of pts lie on either side of mean.
* In range $[\mu - \sigma, \mu + \sigma]$, 68% of pts lie.
* " " $[\mu - 2\sigma, \mu + 2\sigma]$, 95% " " .
* " " $[\mu - 3\sigma, \mu + 3\sigma]$, 99% " " .

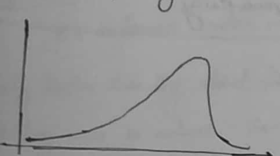


Scanned by CamScanner

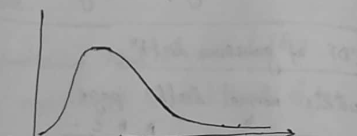
Symmetric, Skewness, Kurtosis

Gaussian is symmetric, across μ.

Measure of asymmetricness – skewness



Negative Skew
left skewed.



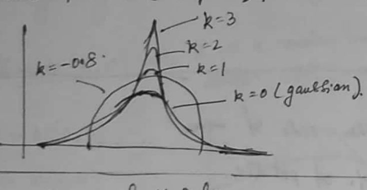
Positive skew
Right skewed

μ – location where central value lies

σ – spread.

Skewness – how assymetric is distribⁿ.

Kurtosis – peakness (sharpness). See wikipedia: kurtosis for graph.



Standard Normal Variate

$z \sim N(0,1)$ i.e. a random variable having gaussian distribution with mean $= 0$ & std-devⁿ $= 1$.

Can convert any gaussian distbⁿ to standard normal variate using standardizatⁿ.

$X \sim N(\mu, \sigma)$. $\quad x_i' = \dfrac{x_i - \mu}{\sigma}$ then $x_i' \sim N(0,1)$.

Advantage of this :– know exactly about the 68-95-99.7 rule & many more properties applicable.

Kernel density estimation

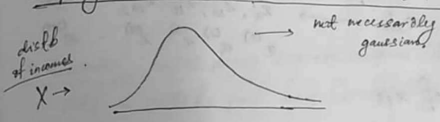
↳ converts histogram to density curve.

wiki-page: kernel_density_estimation

for each point in the sample space, a gaussian kernel is drawn (with the point being mean) & values at each point is summed up to get total height.

variance – (bandwidth) – simulated & selected a suitable value which gives a smooth curve.

Sampling distribution and Central Limit theorem



→ let's say we take m ~~sam~~ random samples, each of size n (say 30).

$S_1, S_2 \ldots \ldots S_m$ (m samples).

$\bar{x}_1, \bar{x}_2, \ldots \ldots \bar{x}_m$ are the mean of the m samples.

$\bar{x}_i \sim$ distbⁿ. called the "sampling distbⁿ of sample mean".

CLT: If X (original population) has finite μ & σ^2, then

$$\bar{x}_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{as } n \to \infty.$$

(Pareto distbⁿ has infinite mean/var).

i.e. the $\bar{x}_i$ distbⁿ of the means of the samples is a Gaussian distbⁿ with mean $= \mu$ (mean of original population) &

variance $= \dfrac{\sigma^2}{n}$ (where σ^2 is variance of original popⁿ).

This CLT is powerful. Because it works on data having any kind of distbⁿ, not just Gaussian.

Quantile-Quantile Plot

Given $X: x_1, x_2, x_3, \ldots x_{500}$.

Q. Is X Gaussian distb? → QQ plot (graphical)
→ Statistical testing (KS, AD).

QQ:

① sort x_i's & computer percentiles.

$x_1, x_2, \ldots x_{500}$

↓ sort (asc)

$x_1', x_2', x_3', \ldots x_{500}'$ $\xrightarrow{\text{percentiles}}$ $x_5', x_{10}', x_{15}', x_{20}', \ldots x_{500}'$

(1, 2, 3, 4, … 100)

$x^{(1)}\ x^{(2)}\ x^{(3)}\ x^{(4)} - x^{(100)}$.

$x^{(i)}$ — i^{th} percentile value of x_i's

② $Y \sim N(0,1)$ — std normal distbⁿ.

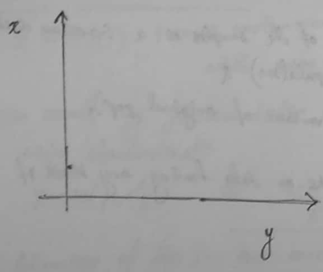
$y_1, y_2, y_3, \ldots y_{1000}$ — 1000 obsvⁿ fm $N(0,1)$.

↓ sort (asc)

$y_1', y_2', \ldots y_{1000}'$ $\xrightarrow{\text{percentiles}}$ $y^{(1)}, y^{(2)}, \ldots y^{(100)}$.

③ plot Q-Q plot using $x^{(1)}, x^{(2)}, \ldots x^{(100)}$
$y^{(1)}, y^{(2)}, \ldots y^{(100)}$.

$(y^{(i)}, x^{(i)})$ $\forall i \in [1, 100]$
ordered pairs plotted



If all these points roughly lie on a straight line, then we can say x & y have the same distbⁿ. Hence $x \in N(\mu, \sigma)$.

But we can't conclude that X also has mean=0 & variance=1.

Mean(x) = x-value corresponding to y=0 in plot.

Refer ipynb.

Q-Q plot generally answers—

Q. If X & Y come from same distribution?

Uniform distribution

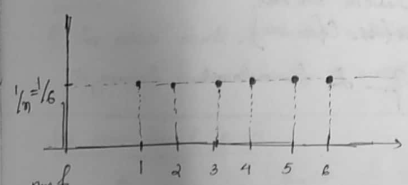
→ discrete.
→ continuous.

① discrete (wiki page: discrete-uniform-distbⁿ).

Pdf is called pmf here (probability mass function).

Ex- Throwing a dice.

All are equi-probable. That's what ~~equiprobable~~ uniform distbⁿ means.



pmf

Parameters – a & b.

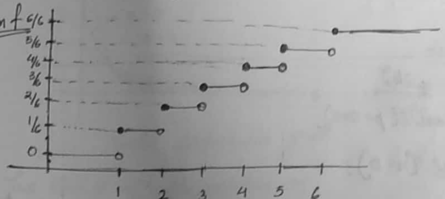
$a \in \mathbb{Z}$

$b \in \mathbb{Z}$. $b \geq a$.

$n = b - a + 1$.

parameters completely describe about the distbⁿ.

cmf



See wiki page for

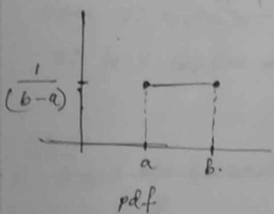
mean: $\left(\frac{a+b}{2}\right)$

median: $\frac{a+b}{2}$

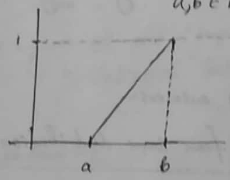
variance: $\frac{(b-a+1)^2 - 1}{12}$.

Skewness: — 0.

② continuous



pdf

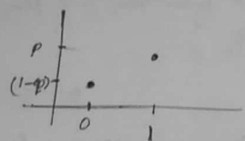


parameters:- a & b
$a, b \in R$ & $b \geq a$.

Application of uniform distbⁿ
∟ Random no. generator (uniform distbⁿ).
see ipynb for random sampling example.

Bernoulli Distribution wiki
∟ discrete distbⁿ.
has 2 outcomes. Probability – p & $(1-p)$.
X – random variable of getting head in coin toss.
$X \sim Bernoulli(p=0.5)$ ⟶ coin toss. (fair coin).

$$pmf \begin{cases} (1-p) & \text{for } k=0 \\ p & \text{for } k=1 \end{cases}$$ — the two outcomes.



Binomial distbⁿ wiki.
coin-toss: $X \sim Bernoulli(p=0.5)$.
coin tossed n-times $(n=10)$.
~~g. [illegible]~~
let Y be a random variable denoting no of heads in 10 coin tosses.

$Y \sim Bin(n, p)$ 2 parameters.
n — no. of coin toss
p — prob. of getting head.

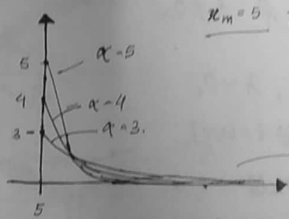
In general the parameters of binomial distbⁿ are
① no. of trials
② probability of success.

pmf: $\binom{n}{k} p^k (1-p)^{n-k}$ — $P(Y=k)$

Log Normal Distribution wiki.
$X \sim$ log-normal(μ, σ) if $\log(X) \sim$ normal distbⁿ.
wiki — for graph of pdf. & cdf.
obsvⁿ: as $\sigma^2 \uparrow$, the pdf curves become more skewed.

Applications– wiki
① Length of comments posted in internet discussions follow log normal distbⁿ.
② The users dwell on online articles follows a log normal distbⁿ.
In general, human behaviour is mostly log-normal distbⁿ.

See power law & pareto



Pareto distribution graph

$x_m = 5$ for all
$(x_m$ & $\alpha)$ — parameters.
As $\alpha \to \infty$, distbⁿ approaches $\delta(x - x_m)$ where δ is dirac delta function.
as $\alpha \downarrow$, tail's fatness $\uparrow$

This kind of relⁿ b/w variables is called power law relationship in maths

Occurrence in nature

① file size distbⁿ in internet traffic.

② hard disk drive error rates.

③ value of oil reserves in oil fields.

Q) How to check if a distbⁿ is ~~power~~ pareto?

→ draw log-log graph. — straight line

y (probability value)   x (feature value)   both's log taken & plotted

Of course we can always use Q-Q plot.

Pareto distribution to Gaussian (box-cox transformⁿ).

Pareto: ~ $X\ [x_1, x_2, x_3, \dots\ x_n]$.

Gaussian: $Y\ [y_1, y_2, y_3 \dots\ y_n]$.

① box-cox(X) = lambda (λ)
all n observatⁿs.

Some complex math done.

② $$y_i = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0. \\ \ln(x_i), & \lambda = 0. \end{cases}$$
$\forall\ i \in [1, n]$.

* $\lambda = 0$ means X is a log-normal distbⁿ.

Y = scipy.stats.boxcox(X).

Quantifying rel'ship among features

→ let say 2 features height and weight.
Are they co-related? i.e. h↑ → w↑/w↓ & vice-versa??

3 measures

① co-variance

② pearson co-relation coeff

③ spearman " " .

① co-variance

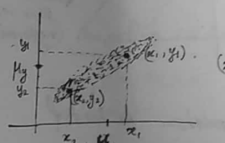
$$\text{variance}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu).$$

$$\text{co-variance}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y).$$

So, covariance(X, X) = variance of (X).

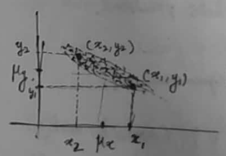
Now, co-variance (X, Y) = { (+)ve implies X↑ then Y↑ ; (−)ve implies X↑ then Y↓.

Intuition



	$(x_i - \mu_x)$	* $(y_i - \mu_y)$	
(x_1, y_1)	(+)	(+)	= (+)
(x_2, y_2)	(−)	(−)	= (+)

cov(X, Y) = (+)ve {summatⁿ over all such pts}.



(x_1, y_1)	(+)	(−)	= (−)
(x_2, y_2)	(−)	(+)	= (−)

cov(X, Y) = (−)ve.

But we don't have any idea of increase or decrease.

One major drawback of covariances – $\text{co-var}(\underset{cm}{X}, \underset{kg}{Y}) \neq \text{co-var}(\underset{ft}{X}, \underset{lbs}{Y})$

Pearson takes care of that.

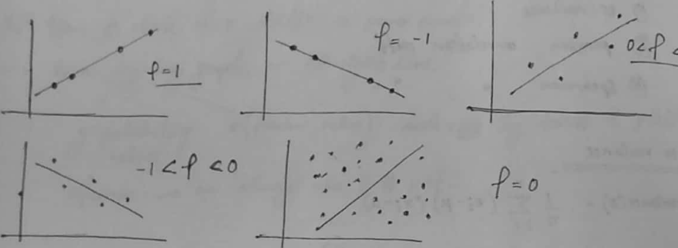
③ pearson corelation coeff (refer wiki)

$$\rho_{x,y} = \frac{covar(X,Y)}{\sigma_x \sigma_y}$$

σ_x = std-dev of x.

See graphs of diff value of ρ in wiki.

$-1 \le \rho \le +1$

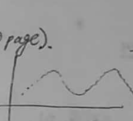


Limitations

① ∟ only +1 when linear rel^n ship b/w x & y.
so, if $y=x^2$, $\rho<1$. (even though monotonical increase).

② slope of st. line doesn't affect ρ.

③ complex relat^n ship not captured. (see wiki page).



$\rho=0$ (beautiful sine rel^n neglected)

fix - using spearman coeff.

Spearman rank - coeff

∟ sort the feature values & find the position in sorted order & find pearson coeff of these positions.

Ex →	X	Y	r_x	r_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

ρ_{r_x, r_y} is Spearman coeff.

So, if X↑ then Y↑, (doesn't matter linear rel^n) then Spearman coeff = 1.

similarly −1 if X↑ → Y↓

* Spearman more robust to outliers than pearson.

Co-relation Vs Causation

* Just because two random variables are co-related (X↑ then Y↑) doesn't mean X causes Y or vice versa.

* causal models − advanced statistics used for those purposes.

Confidence Interval

dist: X : heights.
(any dist^n) $\{x_1, x_2, \ldots x_{10}\}$ − random sample of size 10.

Ex Estimate the populat^n mean of $X = \mu$.

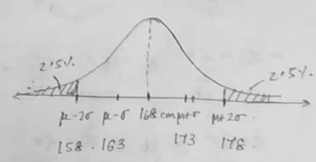
$\mu \simeq \bar{x}$ (sample mean). —— This is point estimate. Not bad, but we can do better.

If we say $\mu \in [162.1, 174.9]$ with 95%. − Interval with some confidence value. − Richer than previous in terms of informat^n.

Now, how to calculate C.I.??

↳ 1) Computing C.I. given the underlying distribution.

say $X \sim N(\mu, \sigma)$ (heights). let $\mu = 168$ cm, $\sigma = 5$cm



frm knowledge of gaussian distb^n,
$(\mu-2\sigma, \mu+2\sigma)$ contains 95% of my observat^n.

So, we can say heights of people lie b/w [158, 178] with 95% probability.
∟ C (confidence)

Similarly other values like 90%, 80% can be found using Normal distb tables

C.I. for mean(μ) of r.v.

$X \sim$ Some distb^n with mean $= \mu$ & std-dev $= \sigma$. — population parameters

$\{x_1, x_2, \dots x_{10}\}$ — we are given only this sample of X

Q) what's the ~~C.I. of~~ 95 % C.I. of μ?? (given this sample)

Case I :- we are given the std-dev^n (σ) of population.

for CLT, if $\bar{x}$ is the sample mean then (analogy to petal length in iris dataset) — say in mind for $\bar{x}$

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

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

Or. i.e. $\bar{x}$'s value is b/w $\mu - \frac{2\sigma}{\sqrt{n}}$ & $\mu + \frac{2\sigma}{\sqrt{n}}$ with 95% chance

or $\mu - \frac{2\sigma}{\sqrt{n}} \le \bar{x} \le \mu + \frac{2\sigma}{\sqrt{n}}$

$-\frac{2\sigma}{\sqrt{n}} \le \bar{x} - \mu \le \frac{2\sigma}{\sqrt{n}}$

$-x - \frac{2\sigma}{\sqrt{n}} \le -\mu \le \frac{2\sigma}{\sqrt{n}} - x$.

$\bar{x} + \frac{2\sigma}{\sqrt{n}} \ge \mu \ge \bar{x} - \frac{2\sigma}{\sqrt{n}}$. Or $\mu \in \left[\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}\right]$

with 95% chance.

So, we have calculated C.I. for μ.

Case 2: we don't know σ (pop^n-std-dev)

use Student's t-distb^n

$\bar{x} \sim t(\underbrace{n-1}_{\text{degrees of freedom}})$.

$\bar{x}$ follows t-distb^n.

what about C.I. of other statistical measures like median or 90th percentile?

↳ bootstrap C.I. — using higher computational powers to do simulations.

C.I. using emperical bootstrap

task:- estimate 95% CI for median of X.
using only the given sample of X.

→ Generate k samples, each of size m ($m \le n$).
(sampling with repetition)

find medians of each one of them.
So, now we'll have k-medians.

Say $k = 1000$.
Now, sort the 1000 medians.

$m_1, m_2, \dots m_{25}, \dots\dots m_{975}, m \dots m_{1000}$ ← sorted order.

$\boxed{\text{C.I.}(95\%) \in [m_{25}, m_{975}]}$

$\frac{950}{1000} = 95\%$.

Similarly other params can also be calculated.

Also, the larger the value of n, the narrower will be interval.

Hypothesis Testing

As the name suggests, we test for the trueness of an assumed hypothesis based on the observations we've got while experimenting

Ex 1:-

Task:- Given a coin determine if the coin is biased towards head/not.

- biased towards head:- $P(H) > 0.5$.
- not biased towards head:- $P(H) = 0.5$.

design expt: flip coin 5 times & count no of heads $= X$. — Test statistic.
(r.v.)

perform expt: f, f, f, f, f
H, H, H, H, H. $X = 5$ — observation for performing expt.

$P(\underbrace{X=5}_{\text{obsr}^n} \mid \underbrace{\text{coin is not biased towards head}}_{\text{an assumption.}}) = P(obs/H_0) = \frac{1}{2^5} \simeq 0.03 = 3\%$.

an assumption. — called Null Hypothesis (H_0)

H_0: coin is not biased towards head

So, $P(X=5/H_0) = 3\%$.
There is a 3% chance of getting 5 heads in 5 flips if the coin is not biased towards head.

Probability of observation given assumption is 3%, quite low.

Since the observation is done practically, its the ground truth.
Hence, our assumption ~~must~~ may be wrong.

* $P(obs/assumption)$ — also called p-value.

Typically p-value $< 5\%$ is said to be small.

Hence H_0 may be incorrect $\Rightarrow$ we reject our null hypothesis.
the idea that coin isn't biased.

we accept the fact that coin is biased towards heads.

H_0: coin not biased. — Null hypothesis
H_1: coin biased towards head — Alternate Hypothesis.

rejecting $H_0 \Rightarrow$ accepting H_1
rejecting $H_1 \Rightarrow$ accepting H_0.

In the expt:- flip was done 5 times — sample size
If we had flipped 3 times →

f, f, f
H, H, H — $X = 3$ $P(X=3/H_0) = \frac{1}{2^3} = 12.5\% > 5\%$

Hence H_0 can't be rejected → accepted.

So, sample size matters.

* p-value = 3%
implies $P(Obs/H_0)$ ✓ not $P(H_0)$ ✗✗

Ex 2

Task: determine if the population mean of heights of people in ~~these~~ two cities is same or not.

→ Its impossible to calculate population mean so, we use sample mean.

expt: Measure height of 50 random people for each city.
let μ_1 & μ_2 be sample means of both cities. say $(\underset{\mu_1}{162} \ \& \ \underset{\mu_2}{167})$

test statistic:- $\mu_2 - \mu_1 = 167 - 162 = 5$ cm. (X).

Null hypothesis:- There is no difference in population mean of both the cities.

compute: $P(X = 5/H_0)$.
diff in sample mean with sample size of 50

$P(X=5 \mid H_0)$

↳ probability of observing a diff of 5cm in sample mean height of sample size 50 between c_1 & c_2 if there is no population mean difference.

Case 1: $P(X=5 \mid H_0) = 0.02 = 20\%$

There is a 20% chance of observing a diff of 5cm in sample mean of height ~~with~~ of c_1 & c_2 with sample size of 50 if there is no population mean difference. So we accept the H_0.

And vice-versa.

Computing $P(X=5 \mid H_0)$ (Resampling).

① Take all heights of both cities & put them together to a new set. (S)

② Randomly select 50 pts from S to S_1 & remaining 50 to S_2. This is called resampling.

* The idea of resampling is that now since S_1 & S_2 are coming from same distbⁿ randomly, this will simulate two cities having same population means or simulate the null hypothesis.

calculate μ_1 & μ_2 and $\mu_2 - \mu_1 = \delta$

③ repeat the 2nd step k no of times.

④ sort the δ_i's in inc. order.

Case 1: $\underbrace{\delta_1 \le \delta_2 \le \cdots \le \delta^*_k}_{\text{simulated differences.}}$

Observed difference = 5cm

$P(\text{diff} \ge 5\text{cm} \mid H_0) = \underline{0.2}$ ✓ significant

Say k = 1000.
& our observed diff = δ_{801}.
So, 20% of sim. diff greater than obs. diff.

& vice versa.

KS-Test for similarity of two distribution (refer wiki)

let X_1 & X_2 be the two samples of size m & n.

let $D_{m,n}$ be the maxᵐ diff in their CDFs.

If $\boxed{D_{m,n} > c(\alpha)\sqrt{\frac{(m+n)}{mn}}}$ ✓ then the belong to diff distbⁿ, else same distbⁿ.

α & $c(\alpha)$ values are taken from table.

scipy.stats.kstest( )

Dimensionality Reduction

why? → to visualize high dimension data.

how? → PCA & t-SNE.

* By default, a vector is column vector.

$x_i \in R^d$ — column vector $d \times 1$.

* Representing dataset

$D = \{x_i, y_i\}_{i=1}^{n}$ (n — data pts.) $x_i \in R^d$; $y_i \in \{\text{setosa, virginica, versicolor}\}$.

D is usually represented in 2 matrices — X & Y.

X — rows are data pts & cols are features,

Y — rows are data pts usually 1 col. only.

Data pre-processing: Column normalization

why? — so that the data becomes nicely structured so that data modelling algos will perform well.

normalizatⁿ: — for every feature, we get

$$\boxed{a_i' = \frac{a_i - a_{min}}{a_{max} - a_{min}}}$$

$a_i' \in [0, 1]$.