Probability and Statistics

Module 1: revision

Konduri Aditya
Department of Computational and Data Sciences
Indian Institute of Science, Bengaluru

konduriadi@iisc.ac.in

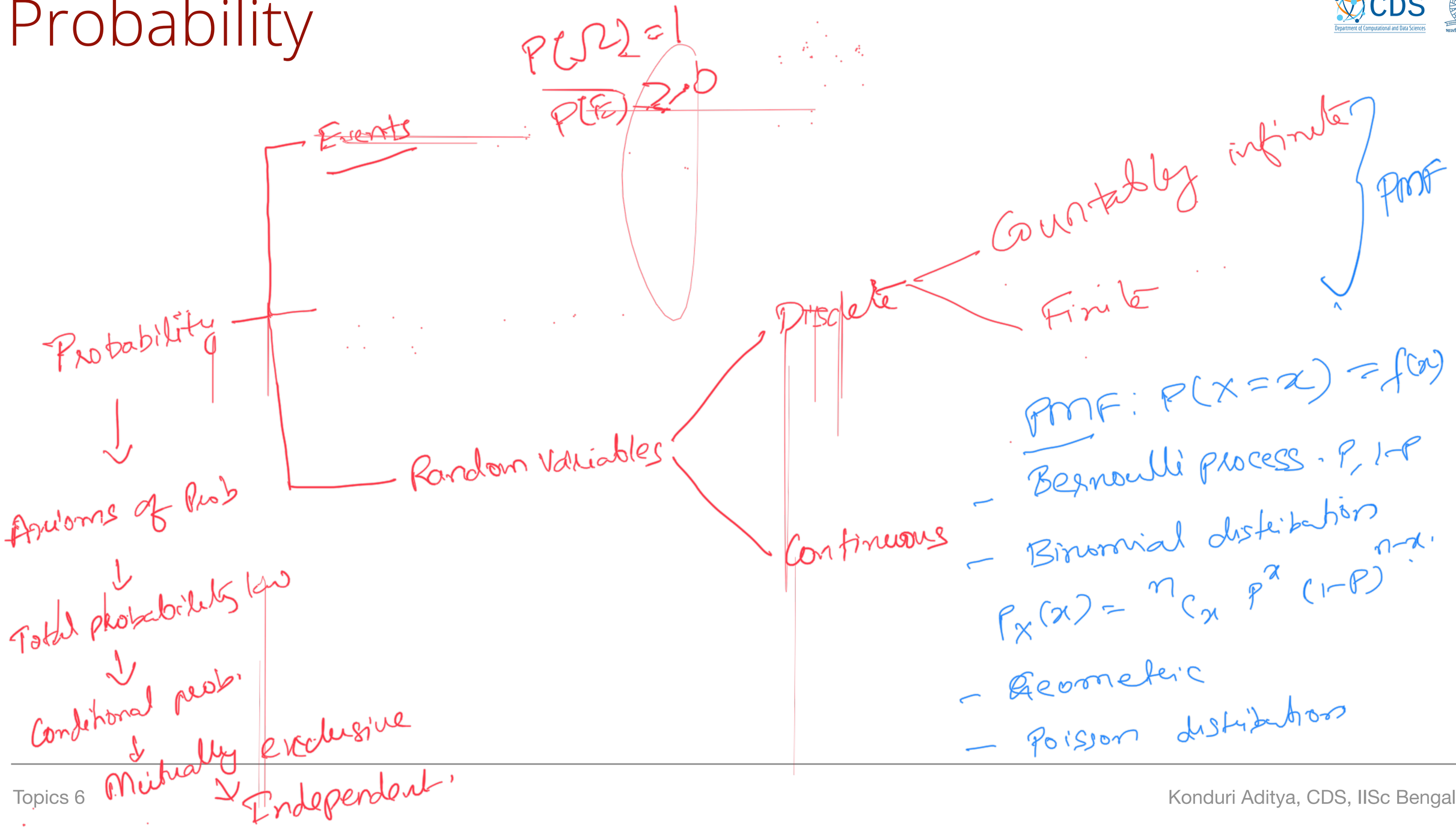
CDS
Department of Computational and Data Sciences

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Outline

- Probability
- Bayesian Inference
- Hypothesis testing
- Overall summary

Probability



Probability

Continuous R.V: $X \rightarrow$ Any real value. $(-\infty, \infty)$

PDF $f(X=x) = f_X(x) = g(x)$

- We obtain the probability associated with different interval.

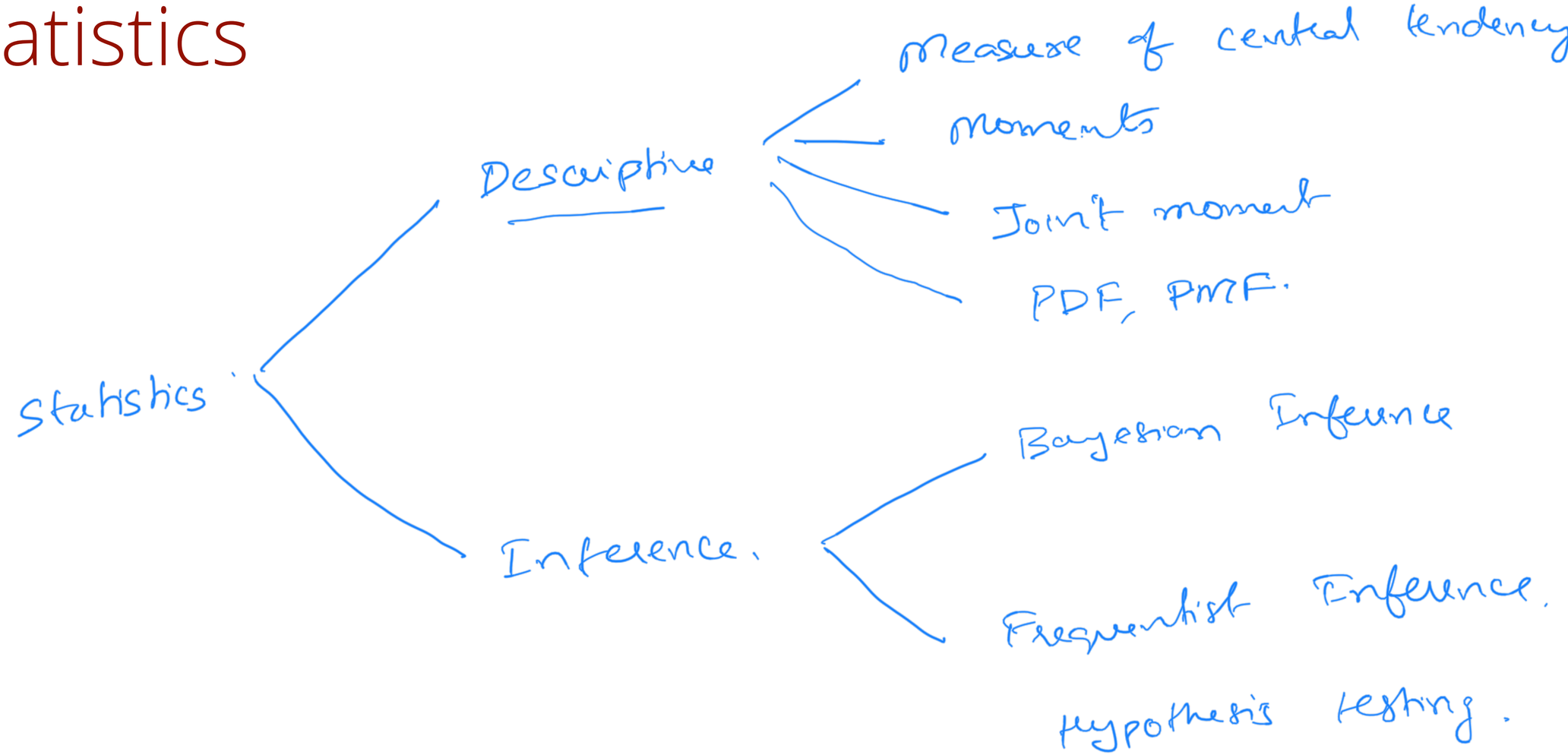
$$P_X(a \leq x \leq b) = \int_a^b f_X(x)\, dx.$$

- Cumulative distribution function.

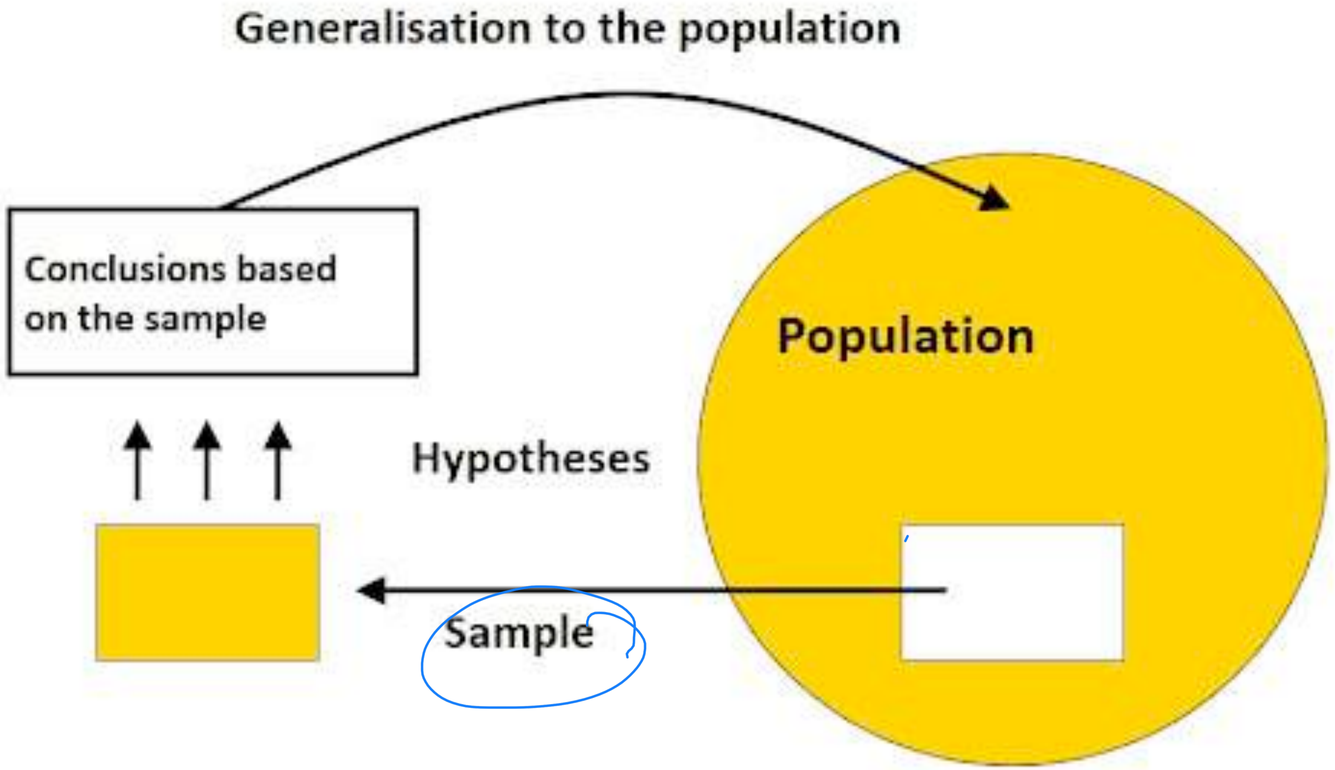
$$CDF(X) = P_X(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

- Normal distribution, exponential distribution, Beta distributions
- Mutual Information : Entropy.

Statistics


Measure of central tendency
Descriptive
Moments
Joint moment
PDF, PMF.
Statistics
Inference.
Bayesian Inference
Frequentist Inference.
Hypothesis testing.

The idea of statistical inference



Source: howmed.net

Bayesian statistical inference

- Bayes' rule is the key:

$$P(hypothesis\ is\ true \mid data) = \frac{P(data \mid hypothesis\ is\ true)P(hypothesis\ is\ true)}{P(data)}$$

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$$

Discrete priors

- A certain disease has a prevalence of 0.005
- A screening test has 2% false positives an 1% false negatives
- Suppose a patient is screened and has a positive test

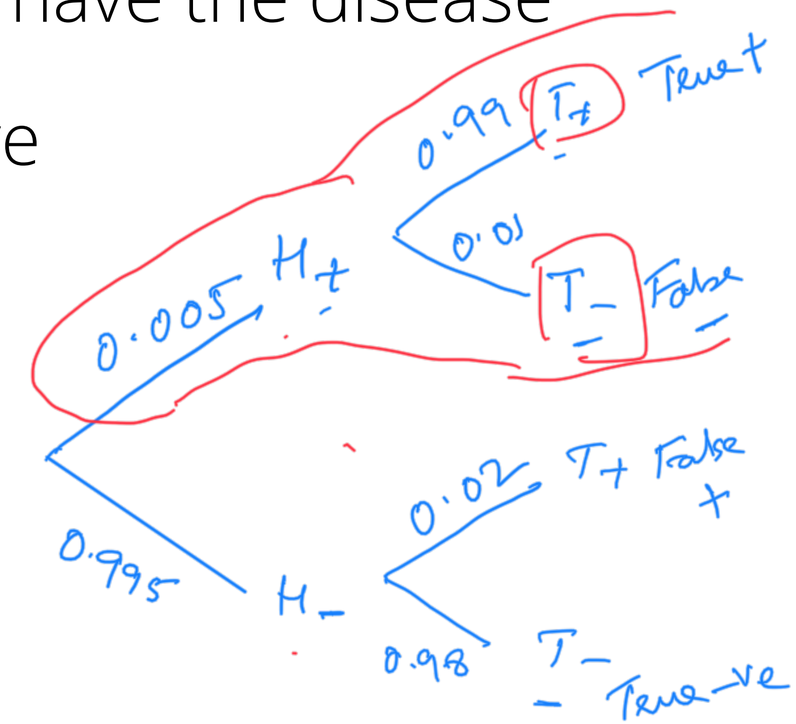
1. Represent this information with a tree and use Bayes' theorem to compute the probabilities the patient does and doesn't have the disease
2. Identify the data, hypotheses, likelihoods, prior probabilities and posterior probabilities
3. Make a full likelihood table containing all hypotheses and possible test data

Discrete priors

1.Represent this information with a tree and use Bayes' theorem to compute the probabilities the patient does and doesn't have the disease

H_+: patient has the disease, $H\_$: patient does not have the disease

T_+: patient tests positive, $T\_$: patient tests negative



Posterior

Prior

likelihood

$$P(H_+ \mid T_+) = \frac{P(H_+)P(T_+ \mid H_+)}{P(H_+)P(T_+ \mid H_+) + P(H\_)P(T_+ \mid H\_)} =$$

$$P(H\_ \mid T_+) =$$

Data

TP TN FP FN.

Discrete priors

Data: The data are the results of the experiment. In this case, the positive test

Hypotheses: The hypotheses are the possible answers to the question being asked. In this case they are H_+ the patient has the disease; H_- they don't

Likelihoods: The likelihood given a hypothesis is the probability of the data given that hypothesis. In this case there are two likelihoods, one for each hypothesis

$$P(T_+ \mid H_+) = 0.99 \quad \textit{and} \quad P(T_+ \mid H_-) = 0.02$$

Likelihood is the probability given the hypothesis

Discrete priors

Prior probabilities of the hypotheses: The priors are the probabilities of the hypotheses prior to collecting data. In this case,

$$P(H_+) = 0.005 \quad and \quad P(H\_) = 0.995$$

Posterior probabilities of the hypotheses: The posteriors are the probabilities of the hypotheses given the data. In this case

$$P(H_+ \mid T_+) = 0.199 \quad and \quad P(H\_ \mid T_+) = 0.801$$

$$P(H_+ \mid T_+) = \frac{P(H_+)P(T_+ \mid H_+)}{P(T_+)}$$

Discrete priors

3. Full likelihood table

The table holds likelihoods P(D|H) for every possible hypothesis and data combination

hypothesis $\mathcal{H}$	likelihood $P(\mathcal{D}\|\mathcal{H})$	
disease state	$P(\mathcal{T}_+\|\mathcal{H})$	$P(\mathcal{T}_-\|\mathcal{H})$
$\mathcal{H}_+$	0.99	0.01
$\mathcal{H}_-$	0.02	0.98

Beta prior: problem

- Treatment has prior $f(\theta) = Beta(5,5)$

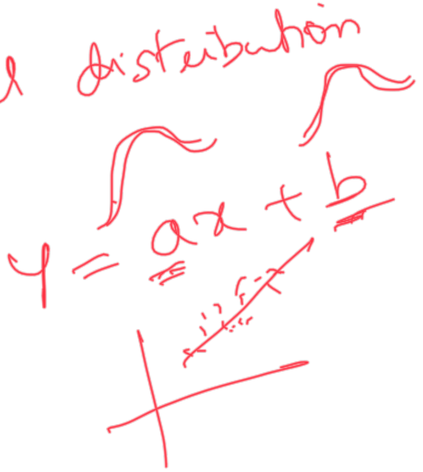
1. Suppose you test it on 10 patients and have 6 successes. Find the posterior distribution on θ. Identify the type of the posterior distribution

2. Suppose you recorded the order of the results and got SSSFFSSSFF. Find the posterior based on this data

Beta prior: solution

- Prior PDF $f(\theta) = Beta(5,5) = \frac{9!}{4!4!}\theta^4(1-\theta)^4 = c_1\theta^4(1-\theta)^4$
- Hypothesis: θ
- Likelihood: $P(Data\,|\,\theta) = {}^{10}C_6\theta^6(1-\theta)^4$
- Posterior PDF:

$$f(\theta\,|\,Data) = \frac{c_1\theta^4(1-\theta)^4 \,.\, {}^{10}C_6\theta^6(1-\theta)^4}{\int_0^1 c_1 t^4(1-t)^4 \,.\, {}^{10}C_6 t^6(1-t)^4 dt} = Beta(11,9)$$



Hypothesis testing

- H_0: the null hypothesis. This is the default assumption for the model generating the data
- H_A: the alternative hypothesis. If we reject the null hypothesis we accept this alternative as the best explanation for the data
- X: the test statistic. We compute this from the data

Ex: H_0: On an avg children watch 3 hours of TV per day.

$H_0: \mu = 3$

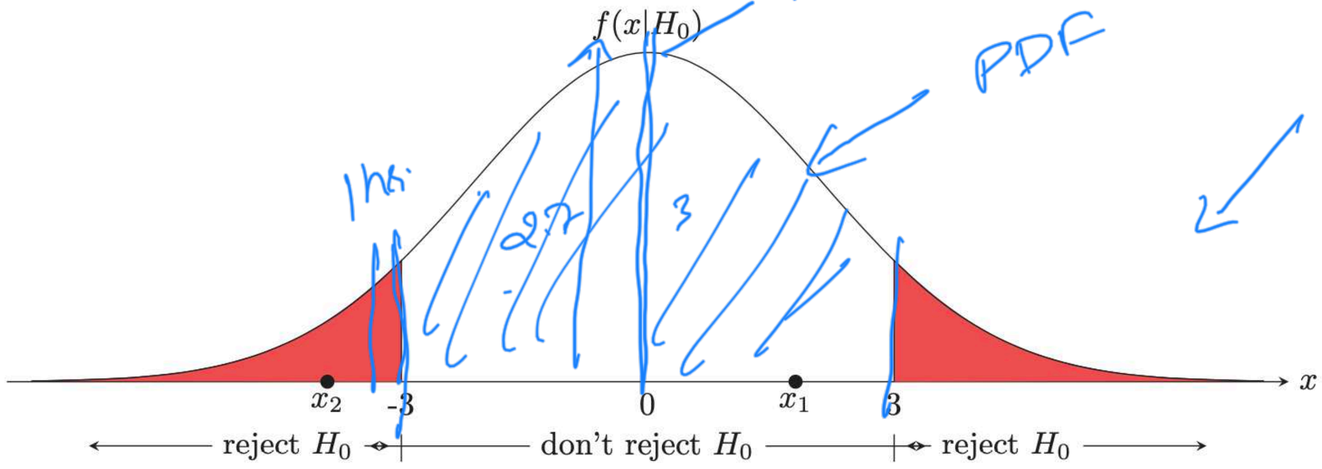
$H_A: \mu \neq 3$

$\bar{X}$ → computed from the sample : 2.9

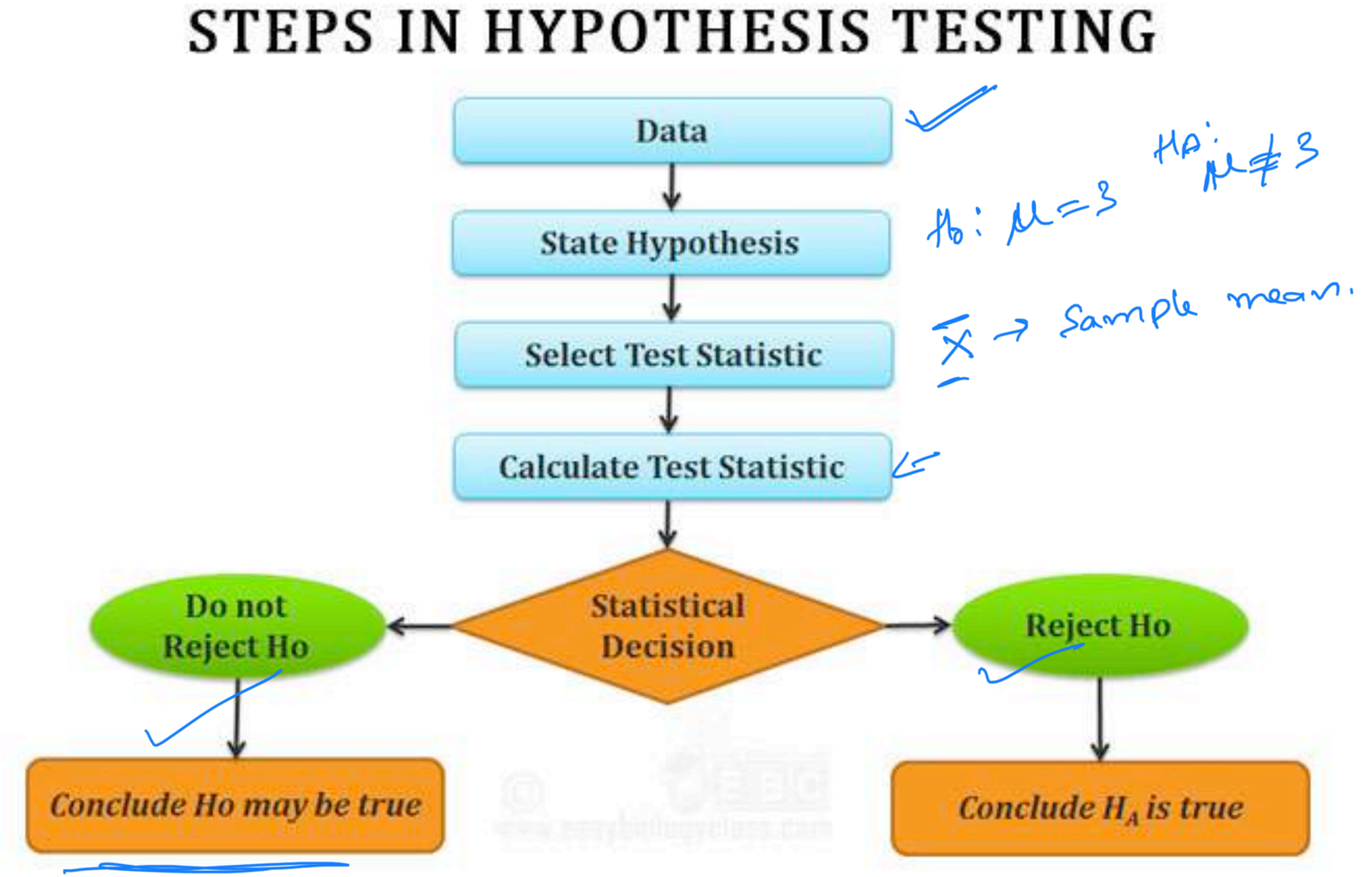
2.5

Hypothesis testing

- Null distribution: the probability distribution of X assuming H_0
- Rejection region: if X is in the rejection region we reject H_0 in favor of H_A
- Non-rejection region: the complement to the rejection region



Hypothesis testing



Source: https://medium.com/analytics-vidhya/hypothesis-testing-steps-235d2670cad4

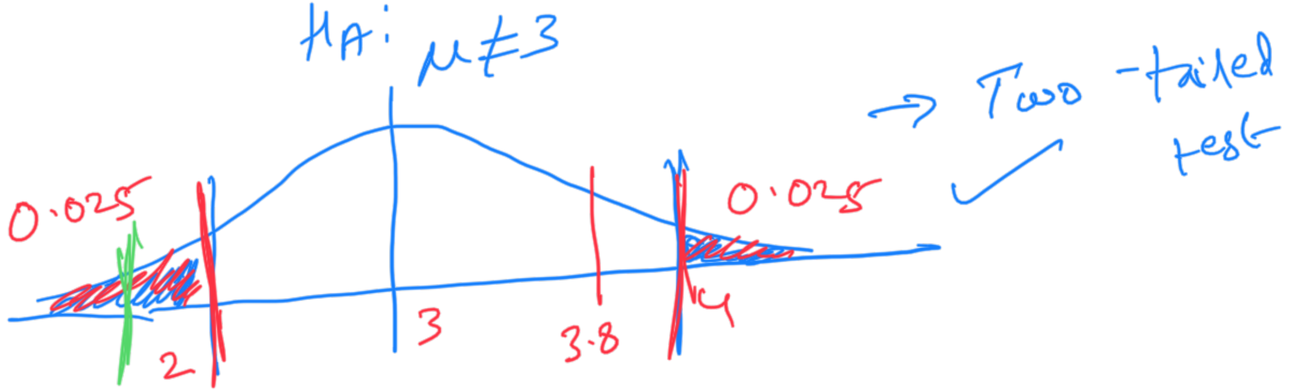
Examples of hypothesis

– tails.

Significance level α: 0.05

$H_0 : \mu = 3$

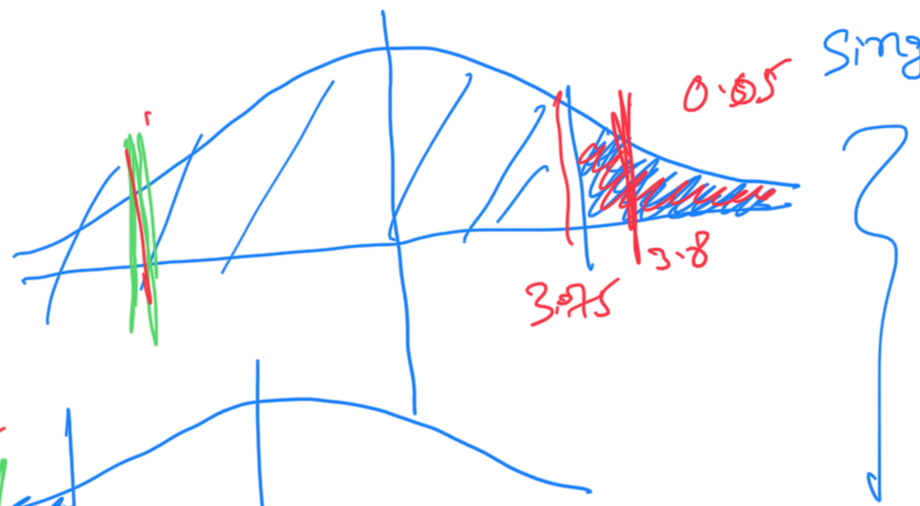
$H_A : \mu \neq 3$



→ Two-tailed test

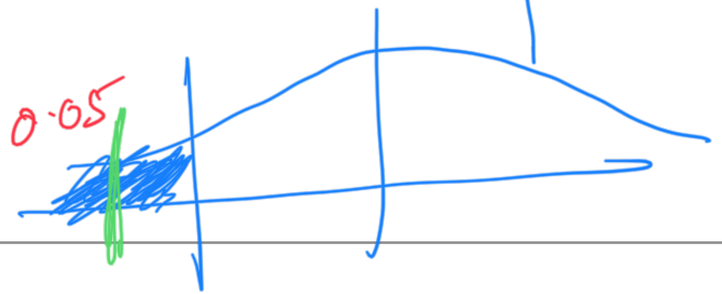
$H_0 : \mu = 3$

$H_A : \mu > 3$



Single-tailed test

$H_A : \mu < 3$



Types of errors, significance level, power

- **Type I:** false rejection of H_0
- **Type II:** false non-rejection ('acceptance') of H_0

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Don't reject' H_0	correct decision	Type II error

Significance level $= P(\text{reject } H_0 | H_0)$
$= \text{probability we incorrectly reject } H_0$
$= P(\text{type I error}).$

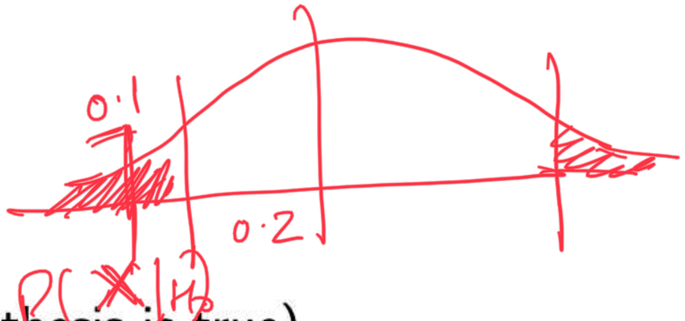
ideal case 0

Ideal values??

Power $= \text{probability we correctly reject } H_0$
$= P(\text{reject } H_0 | H_A)$
$= 1 - P(\text{type II error}).$

ideal case 1

P-value



Hypotheses: H_0, H_A.

Test statistic: value: x, random variable X.

Null distribution: $f(x|H_0)$ (assumes the null hypothesis is true)

Sides: H_A determines if the rejection region is one or two-sided.

Rejection region/Significance: $P(x \text{ in rejection region} \mid H_0) = \alpha$.

The p-value is a computational tool to check if the test statistic is in the rejection region. It is also a measure of the evidence for rejecting H_0.

p-value: $P(\textbf{data at least as extreme as } x \mid H_0)$

Data at least as extreme: Determined by the sided-ness of the rejection region.

Screenshot from MIT OCW 18.05 slides

Example

- H_0: data follows a $N(5, 10^2)$
- H_A: data follows a $N(\mu, 10^2)$ where $\mu \neq 5$.
- Test statistic: z = standardized $\bar{x}$.
- Data: 64 data points with $\bar{x} = 6.25$.
- Significance level set to $\alpha = 0.05$.

(i) Find the rejection region; draw a picture.

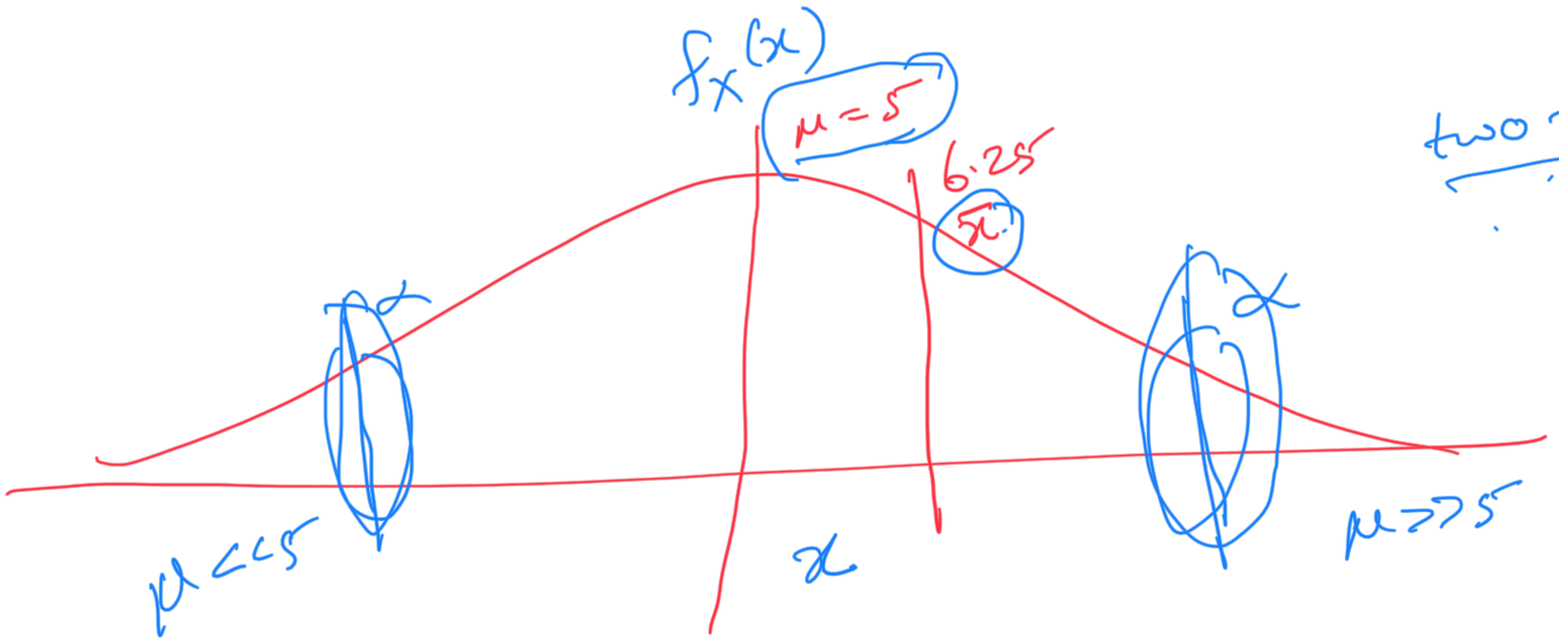
(ii) Find the z-value; add it to your picture.

(iii) Decide whether or not to reject H_0 in favor of H_A.

(iv) Find the p-value for this data; add to your picture.

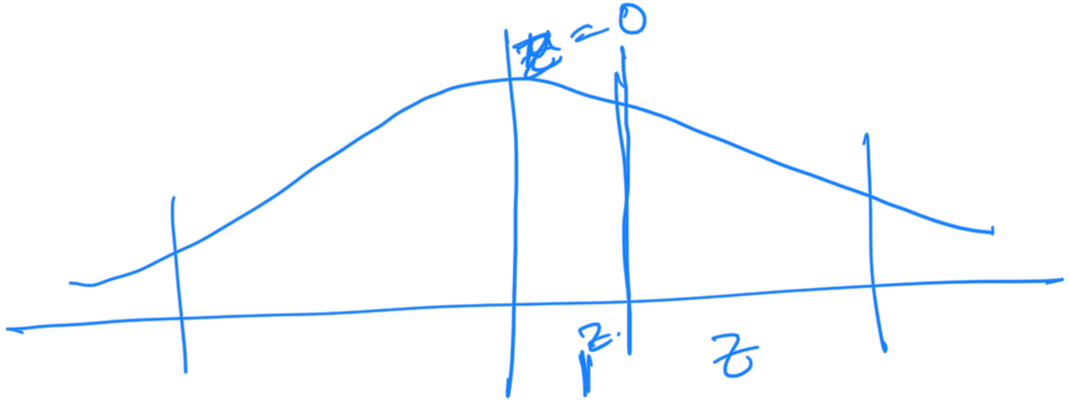
(v) What's the connection between the answers to (ii), (iii) and (iv).

Screenshot from MIT OCW 18.05 slides



two-tail. test

$P(|Z| \geq 1)$



$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{5 - 5}{10/8} = 0$$

$$Z = \frac{6.25 - 5}{5/4}$$

Z-statistic value for the sample.

Example: solution

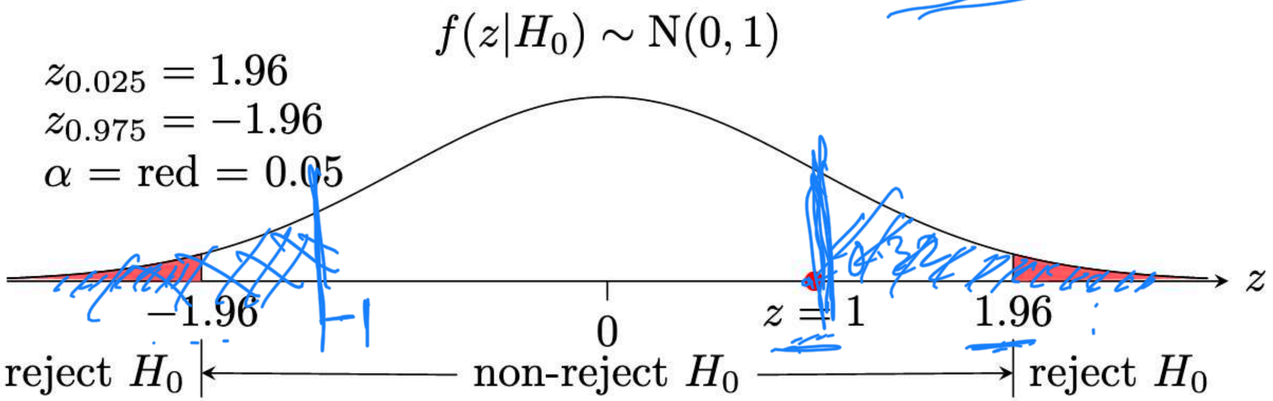
The null distribution $f(z \mid H_0) \sim N(0,1)$

(i) The rejection region is $|z| > 1.96$, i.e. 1.96 or more standard deviations from the mean.

(ii) Standardizing $z = \dfrac{\overline{x} - 5}{5/4} = \dfrac{1.25}{1.25} = 1.$

(iii) Do not reject since z is not in the rejection region

(iv) Use a two-sided p-value $p = P(|Z| > 1) = .32$



(v) The z-value is not in the rejection region tells us exactly the same thing as the p-value being greater than the significance, i.e. don't reject the null hypothesis H_0.

Screenshot from MIT OCW 18.05 slides

Z test - sample. Hypothesis

$$\bar{Z} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

σ : Population variance. (Known)

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

$N(\mu, \sigma^2)$ → $N(0,1)$ for large n

value R.V.

$\frac{\bar{x} - \mu}{\sigma}$.

For large n

t - test

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

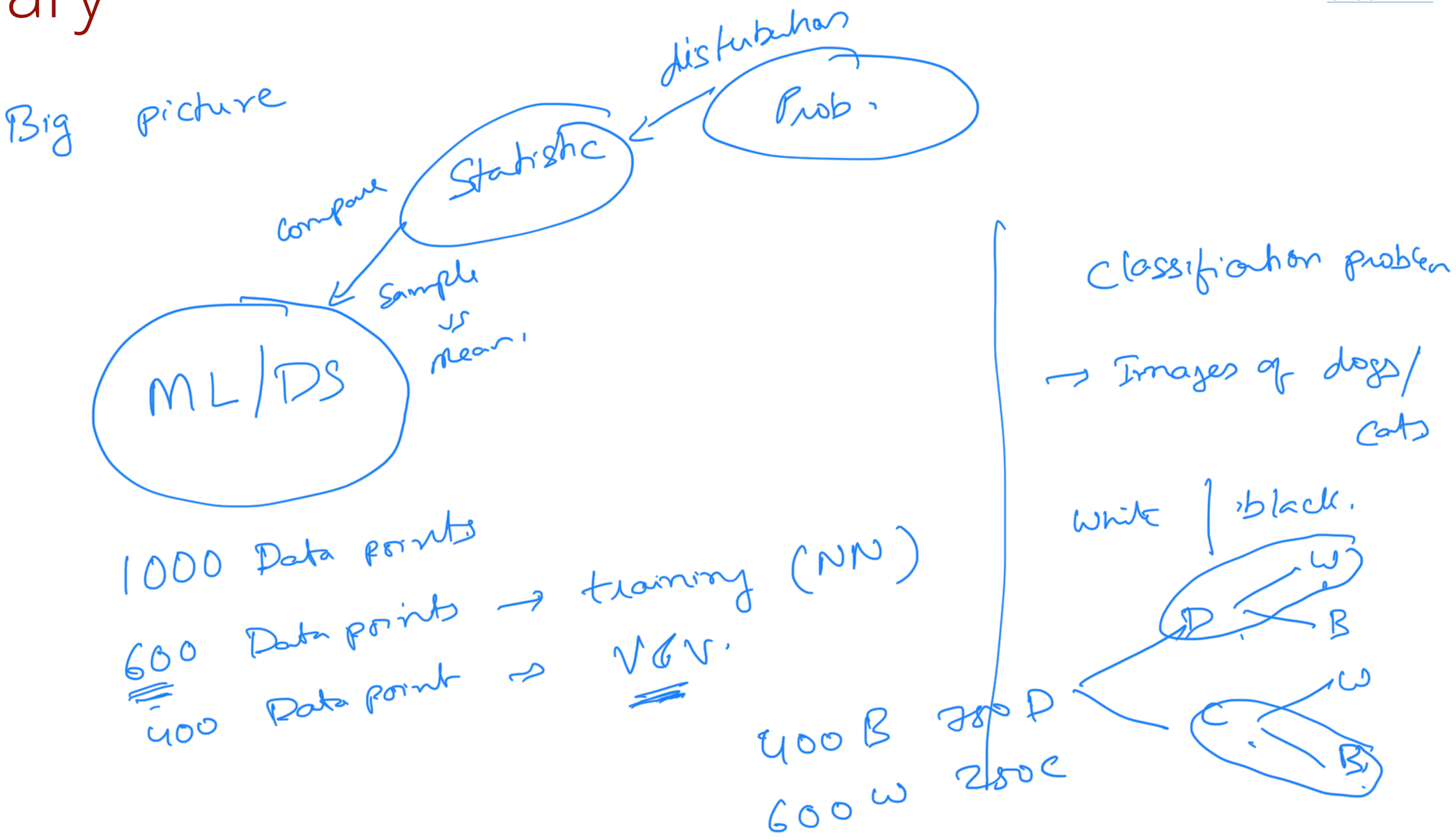
S → Standard deviation of the sample.

→ both Z and t converge.

t - student t distribution

Small n.

Summary


Big picture
Prob.
distribution
Statistic
compare
ML/DS
Sample vs pop.
1000 Data points
600 Data points → training (NN)
400 Data point → V&V.
Classification problem
→ Images of dogs/cats
White | black.
400 B
600 W
300 D
200 C
W
B
W
B

Training : 600 WD + BC

BD / WC

Safely → BC

Testing : BD

↓

Input image.

Compare the diversity between the sample and the population

→ Use hypothesis test.

↓

Based on a particular parameter.

↓

Distribution → Null distribution

$y = \hat{a}x + \hat{b}$

→ Ex: Bayesian linear regression.

→ Bayesian NN.

∴

Question Paper

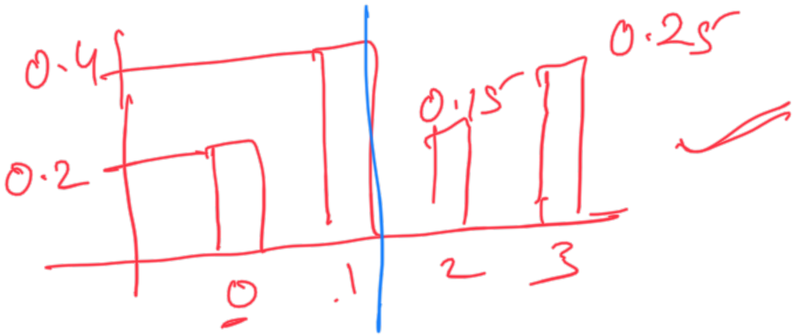
(1) T/F:

(2) MCQ

(3) Descriptive Questions

Moment Generating function

Discrete R.V. $X \rightarrow$ PMF $P_X(x) = g(x)$



~~0~~ + Mean $= \sum_x x \cdot P_X(x)$

Second moment $= \sum_x x^2 P_X(x)$ MGF

n^{th} moment $= \sum_x x^n P_X(x)$