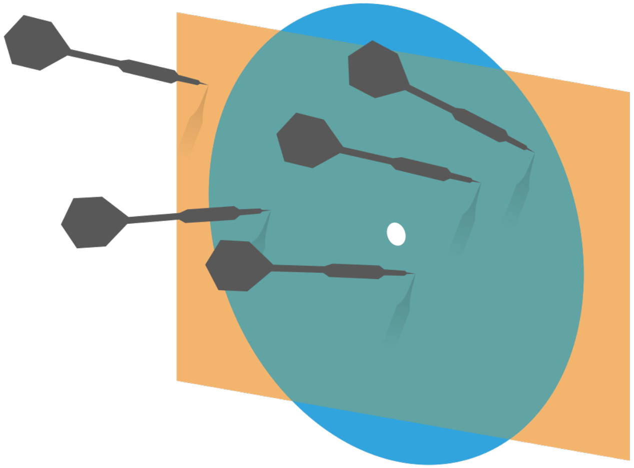
Probability and Statistics for Computer Science



Credit: wikipedia

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H. G. Wells

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.15.2021

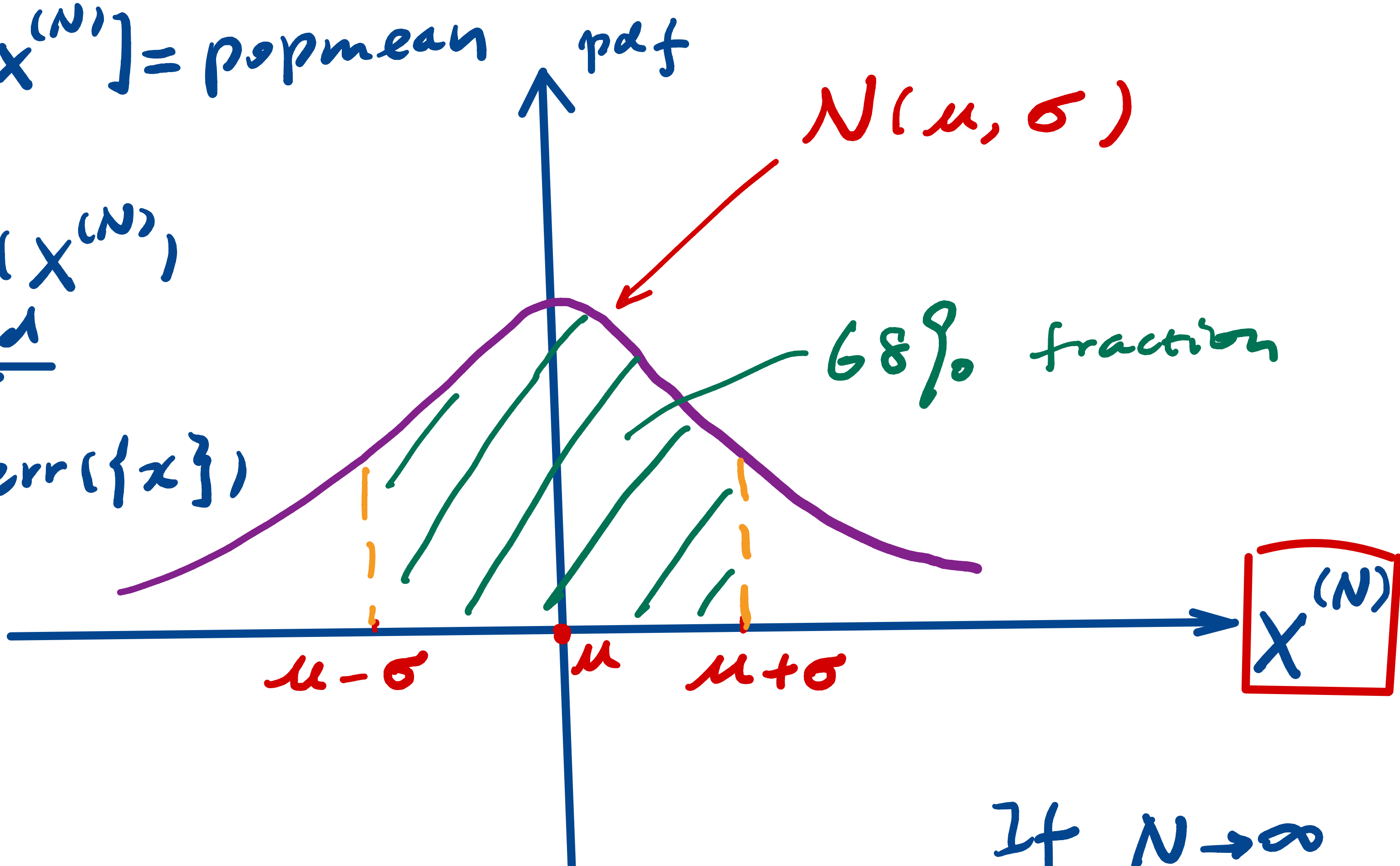
Meaning of #% Confidence Interval

$\mu = E[X^{(N)}] = \text{popmean}$

$\sigma = \text{std}(X^{(N)})$

$= \dfrac{\text{popsd}}{\sqrt{N}}$

$\doteq \text{stderr}(\{x\})$



If $N \to \infty$

Meaning of #% Confidence Interval

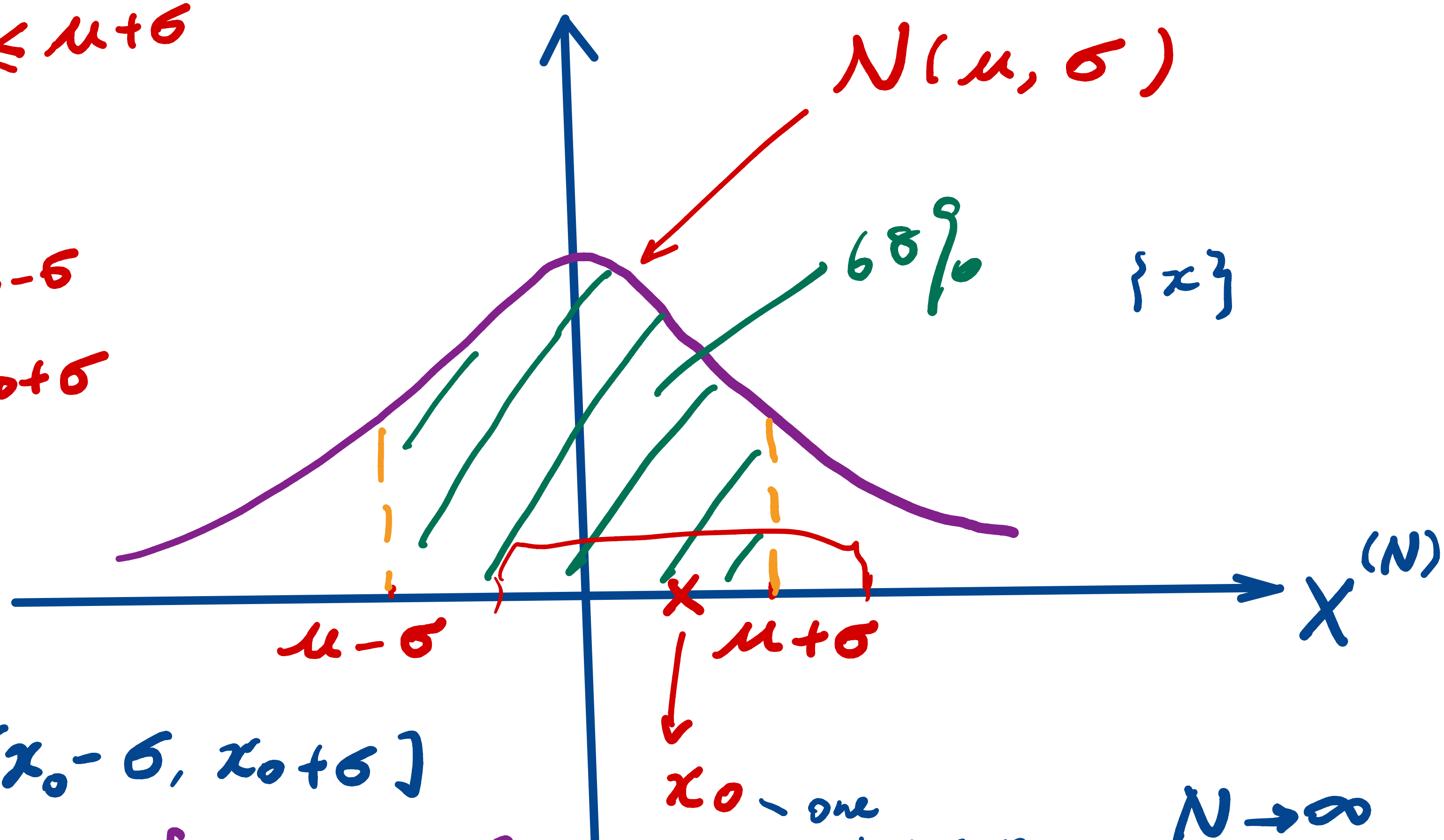
$\mu = E[X^{(N)}] = \text{popmean}$

If

$\mu - \sigma \leq x_0 \leq \mu + \sigma$

$\mu \geq x_0 - \sigma$

$\mu \leq x_0 + \sigma$



$\mu \in [x_0 - \sigma,\ x_0 + \sigma]$

popmean $\in [x_0 - \sigma,\ x_0 + \sigma]$

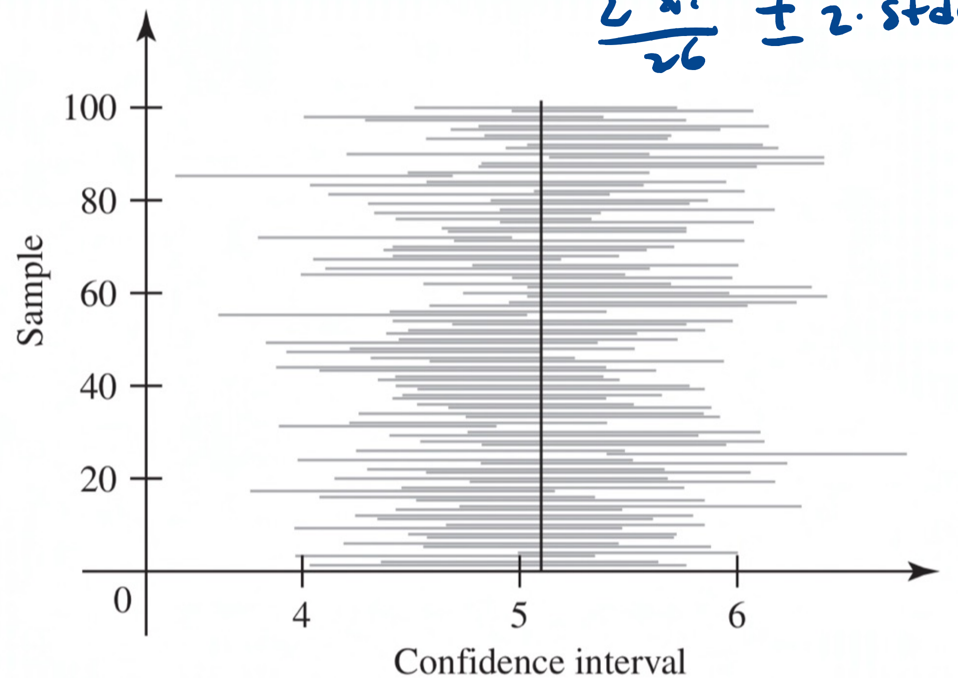
$N \to \infty$

Meaning of #% Confidence Interval

one sample $\{x_1, x_2, \ldots x_{26}\}$

$\frac{\sum^{26} x_i}{26} \pm 2 \cdot \text{stderr}$

Figure 8.5 A sample of one hundred observed 95% confidence intervals based on samples of size 26 from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this figure, 94% of the intervals contain the value of μ.



Degroot Pg 487

A tale of two statisticians

$\{X\} = \{1, 2, 3, \dots 12\}$ $N_p = 12$

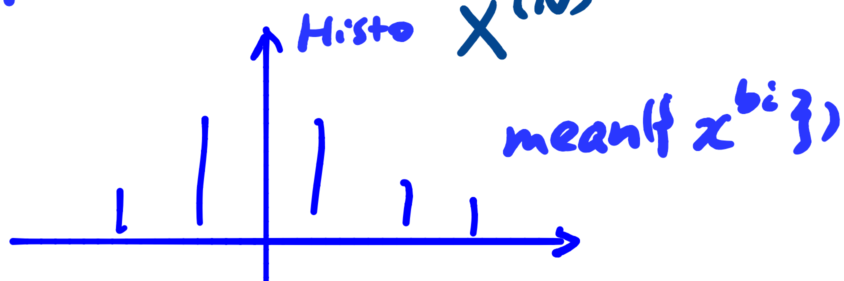
$\{x^b\} = \{1, 4, 5, 7, 11\}$

$\{x^{b1}\} = \{1, 1, 4, 5, 7\}$ $x^{b_1(n)} = \frac{18}{5}$

$\{x^{b2}\} = \{4, 5, 7, 7, 1\}$

$\vdots$

$\{x^{bn}\} = \{5, 5, 5, 5, 5\}$



i.i.d $X^{(1)}$

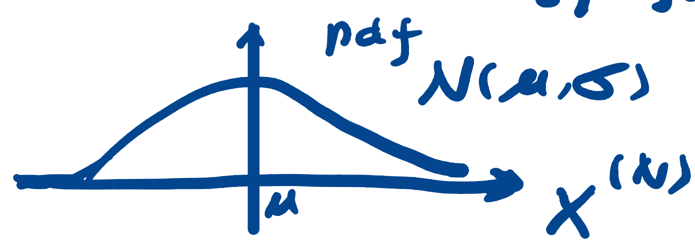
$\{x\} = \{1, 4, 5, 7, 11\}$ $N = 5$

if $N \to \infty$

$X^{(N)} \sim N(\mu, \sigma)$

$\mu = E[X^{(N)}] \doteq \text{mean}(\{x\})$

$\sigma = \text{std}[X^{(N)}] \doteq \text{stderr}(\{x\})$



Objectives

- Hypothesis test
- Chi-square test
- Maximum Likelihood Estimation

A hypothesis

- Ms. Smith's vote percentage is 55%

  simple
  $\theta = \theta_0$

  This is what we want to test, often called null hypothesis H_0

  H_1: perct ≠ 55%

		DATES	POLLSTER	SAMPLE	RESULT	NET RESULT
U.S. Senate	Miss.	NOV 25, 2018	C+ Change Research	1,211 LV	Espy 46% 51% Hyde-Smith	Hyde-Smith +5

51%

- Should we reject this hypothesis given the poll data?

Rejection region of null hypothesis H_0

- Assuming the hypothesis H_0 is true

popmean = v_0

- Define a test statistic

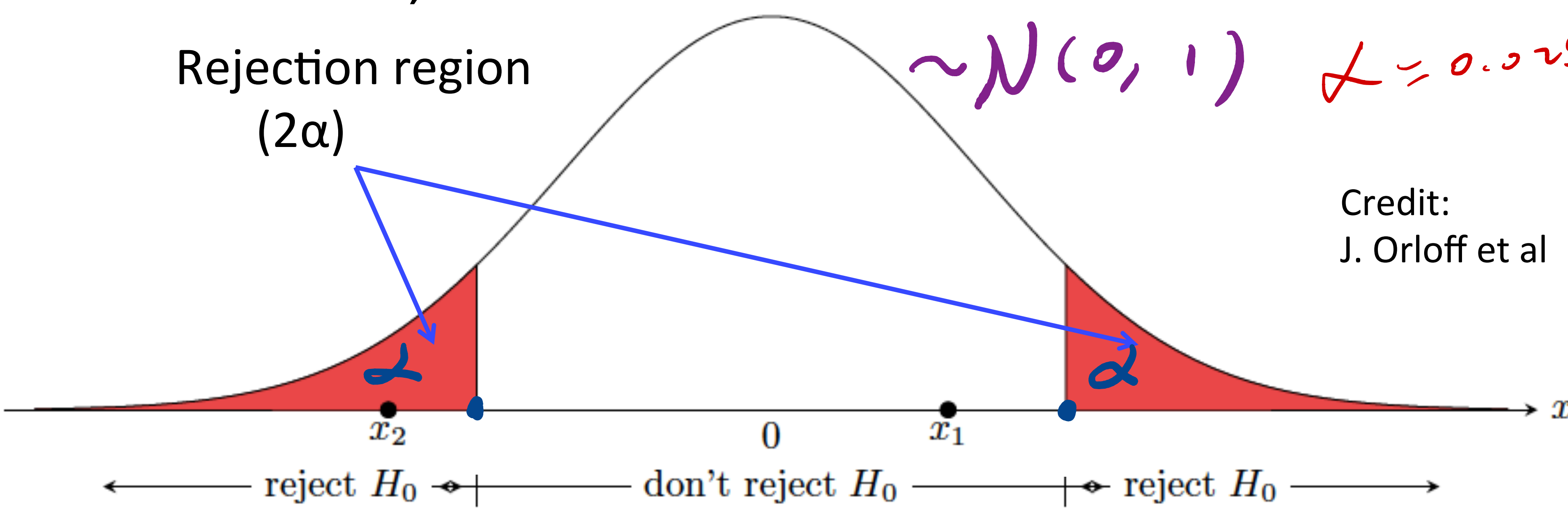
mean({x}) *

$$x = \frac{(sample\ mean) - (hypothesized\ value)}{standard\ error}$$

stderr({x})

- Since N>30, assume x comes from a standard normal

$\sim N(0, 1)$ $\alpha = 0.025$



Credit:
J. Orloff et al

Fraction of "less extreme" statistic

- Assuming the hypothesis H_0 is true $\{x\} = \{10, 20, \cdots, 50\}$

- Define a statistic for the test

$\{x\} = \{10, 20, 30\}$

$mean(\{x\}) = 20$

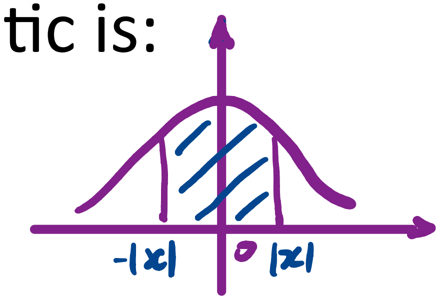
$v_0 = 50$

$$x = \frac{(sample\ mean) - (hypothesized\ value)}{standard\ error}$$

- Since *N*>30, we assume x comes from a standard normal

- So, the fraction of "less extreme" statistic is:

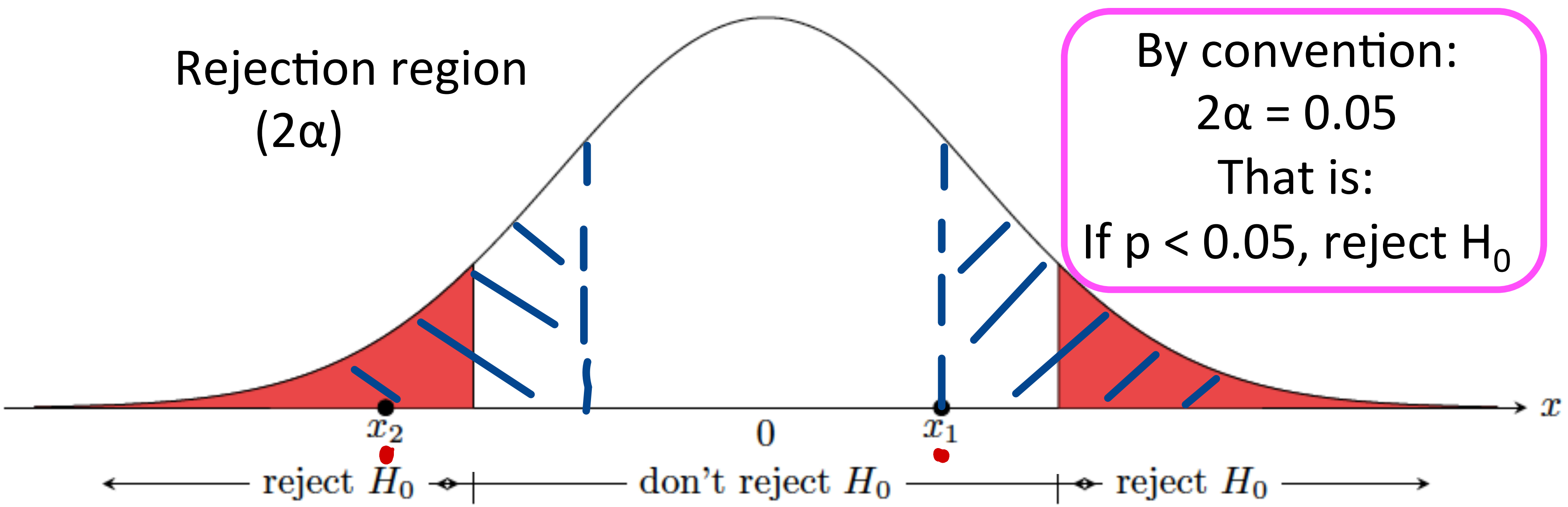
$$f = \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} exp(-\frac{u^2}{2}) du$$



P-value: Rejection region- “The extreme fraction”

- It is conventional to report the p-value

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} exp(-\frac{u^2}{2})du$$



p-value: election polling

- $H_{0:}$ Ms. Smith's vote percentage is 55% pop mean
- The sample mean is 51% and stderr is 1.44%
- The test statistic $x = \frac{51-55}{1.44} = -2.7778$
- And the p-value for the test is:

$$p = 1 - \frac{1}{\sqrt{2\pi}} \int_{-2.7778}^{2.7778} exp(-\frac{u^2}{2}) du = 0.00547 \qquad < \underline{0.05}$$

convention

- So we reject the hypothesis

Hypothesis test if N < 30

- Q: what distribution should we use to test the hypothesis of sample mean if N<30?
  A. Normal distribution
  B. t-distribution with degree =30
  C. t-distribution with degree = N
  D. t-distribution with degree = N-1

The use and misuse of p-value

- p-value use in scientific practice
  - Usually used to reject the null hypothesis that the data is random noise
  - Common practice is $p < 0.05$ is considered significant evidence for something interesting
- Caution about p-value hacking
  - Rejecting the null hypothesis doesn’t mean the alternative is true
  - $P < 0.05$ is arbitrary and often is not enough for controlling false positive phenomenon

Chi-square distribution

- If $Z_i's$ are independent variables of standard normal distribution, $X = Z_1^2 + Z_2^2 + ... + Z_m^2 = \sum_{i=1}^{m} Z_i^2$

  has a Chi-square distribution with degree of freedom ***m*** , $X \sim \chi^2(m)$

- We can test the goodness of fit for a model using a statistic ***C*** against this distribution, where

$$C = \sum_{i=1}^{m} \frac{(f_o(\varepsilon_i) - f_t(\varepsilon_i))^2}{f_t(\varepsilon_i)}$$

$\varepsilon_i \rightarrow$ event i

Independence analysis using Chi-square

- Given the two way table, test whether the column and row are independent

indpt.

$P(A \cap B) = P(A)P(B)$

	Boy	Girl	Total
Grades	117	130	247
Popular	50	91	141
Sports	60	30	90
Total	227	251	478

$P(A/B) = P(A)$

Independence analysis using Chi-square

- The theoretical expected values if independent

$247 \times \frac{227}{478}$

	Boy	Girl	Total
Grades	117.29916	129.70084	247
Popular	66.96025	74.03975	141
Sports	42.74059	47.25941	90
Total	227	251	478

The degree of the chi-square distribution for the two way table

- The degree of freedom for the chi-square distribution for a ***r*** by ***c*** table is

$r = 3$

$$\mathbf{(r\text{-}1) \times (c\text{-}1) \quad where\ r>1\ and\ c>1}$$

$c = 2$

- Because the degree df = n-1-p

See textbook Pg 171-172

$$= rc - 1 - (r-1) - (c-1)$$

$$= (r-1) \times (c-1)$$

$$= 2$$

n is the number of cells of data;
p is the number of unknown parameters

Chi-square test for the popular kid data

- The Chi-statistic : $\boxed{21.455}$

```
chisq.test(data_BG)

        Pearson's Chi-squared test

data:  data_BG
X-squared = 21.455, df = 2, p-value = 2.193e-05
```

- P-value: $\boxed{2.193e-05}$
- It's very unlikely the two categories are independent

Q. What is the degree of freedom for this?

- The following 2-way table for chi-square test has a degree of freedom equal to:

$r = 4$

$c = 5$

Table 10.26 Data for Exercise 3

	Number of lectures attended				
	0	1	2	3	4
Freshmen	10	16	27	6	11
Sophomores	14	19	20	4	13
Juniors	15	15	17	4	9
Seniors	19	8	6	5	12

$(4-1)(5-1)$

$= 3 \times 4$

A. 20 B. 9

$\boxed{\text{C.}}$ 12 D. 4

Chi-square test is very versatile

- Chi-square test is so versatile that it can be utilized in many ways either for discrete data or continuous data via intervals
- Please check out the worked-out examples in the textbook and read more about its applications.

Maximum likelihood estimation

$P(X=k) = \binom{N}{k} p^k (1-p)^{N-k}$ $k \geq 0$

write $P(X=k)$ ↓

p is unknown

write p as θ

$L(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$

Maximize $L(\theta)$, we get $\hat{\theta}$

$\hat{\theta} = \underset{\theta}{\mathrm{Argmax}}\, L(\theta)$

$D: N, K$

Motivation: Poisson example

- Suppose we have data on the number of babies born each hour in a large hospital

λ

hour	1	2	...	N
# of babies	k_1	k_2	...	k_N

- We can assume the data comes from a Poisson distribution

- What is your best estimate of the intensity λ?

Credit: David Varodayan

Maximum likelihood estimation (MLE)

- We write the probability of seeing the data D given parameter θ

$$L(\theta) = P(D|\theta)$$

- The **likelihood function** $L(\theta)$ is **not** a probability distribution

- The **maximum likelihood estimate (MLE)** of θ is

$$\hat{\theta} = arg\ \max_{\theta}\ L(\theta)$$

Why is $L(\theta)$ not a probability distribution?

A. It doesn’t give the probability of all the possible θ values.

B. Don’t know whether the sum or integral of $L(\theta)$ for all possible θ values is one or not.

C. Both.

Likelihood function: binomial example

- Suppose we have a coin with unknown probability of coming up heads θ
- We toss it ***N*** times and observe ***k*** heads
- We know that this data comes from a binomial distribution
- What is the likelihood function $L(\theta) = P(D|\theta)$?

$$\boxed{L(\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}}$$

θ = Prob. of head

Likelihood function: binomial example

- Suppose we have a coin with unknown probability of θ coming up heads

- We toss it ***10*** times and observe ***7*** heads

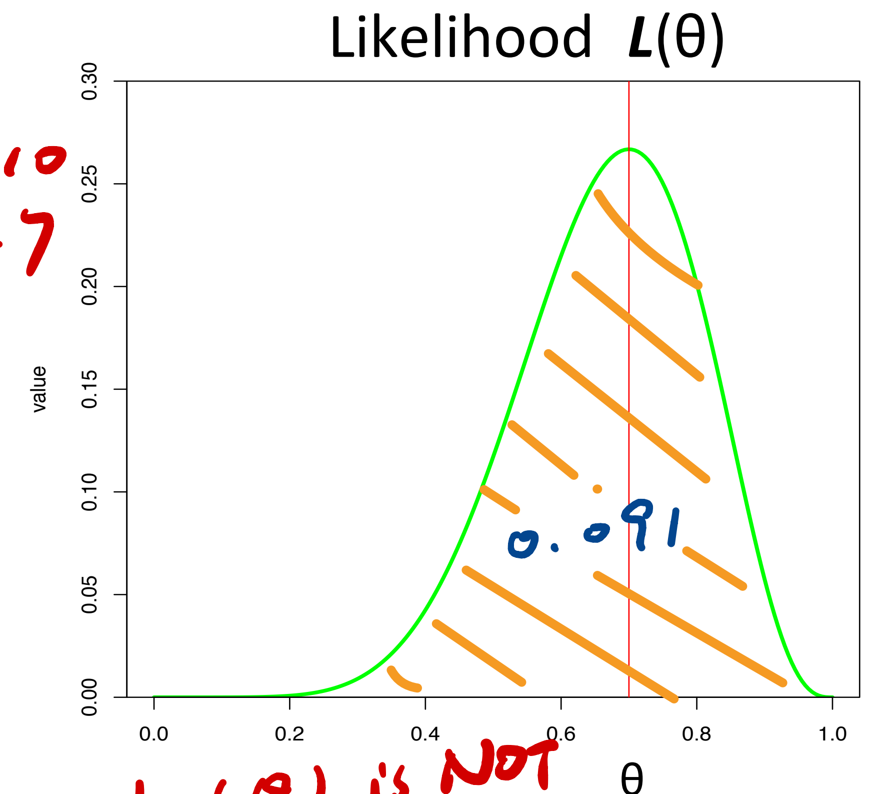
D: N=10
K=7

- The likelihood function is:

$$P(D|\theta) = \binom{10}{7} \theta^7 (1-\theta)^3$$

- The MLE is

$$\hat{\theta} = 0.7$$



L(θ) is NOT a distri.

MLE derivation: binomial example

$$L(\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$$

In order to find: $\hat{\theta} = arg\ \max_{\theta}\ L(\theta)$

We set: $\dfrac{\mathrm{d}L(\theta)}{\mathrm{d}\theta} = 0$

MLE derivation: binomial example

$$\boxed{L(\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}}$$

$(a(x)\,b(x))'$

$ab' + a'b$

$$\frac{d}{d\theta}L(\theta) = \binom{N}{k}\left(k\theta^{k-1}(1-\theta)^{N-k} - \theta^k(N-k)(1-\theta)^{N-k-1}\right) = 0$$

$$k\theta^{k-1}(1-\theta)^{N-k} = \theta^k(N-k)(1-\theta)^{N-k-1}$$

$$k - k\theta = N\theta - k\theta$$

p : prob of seeing H

$$\boxed{\hat{\theta} = \frac{k}{N}}$$

The MLE of p

Likelihood function: geometric example

- Suppose we have a die with unknown probability of coming up six
- We roll it and it comes up six for the first time on the kth roll
- We know that this data comes from a geometric distribution
- What is the likelihood function $L(\theta) = P(D|\theta)$? **Assume θ is p**.

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

$P(D|\theta)$

what is the D?

$D : k$

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Likelihood function: geometric example

- Suppose we have a die with unknown probability of coming up six
- We roll it and it comes up six for the first time on the kth roll
- We know that this data comes from a geometric distribution
- What is the likelihood function $L(\theta) = P(D|\theta)$? **Assume θ is p**.

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

$$\frac{d}{d\theta}L(\theta) = (1-\theta)^{k-1} - (k-1)(1-\theta)^{k-2}\theta = 0$$

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

$$\frac{d}{d\theta}L(\theta) = (1-\theta)^{k-1} - (k-1)(1-\theta)^{k-2}\theta = 0$$

$$(1-\theta)^{k-1} = (k-1)(1-\theta)^{k-2}\theta$$

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

$$\frac{d}{d\theta}L(\theta) = (1-\theta)^{k-1} - (k-1)(1-\theta)^{k-2}\theta = 0$$

$$(1-\theta)^{k-1} = (k-1)(1-\theta)^{k-2}\theta$$

$$1-\theta = k\theta - \theta$$

MLE derivation: geometric example

$$\boxed{L(\theta) = (1-\theta)^{k-1}\theta}$$

$$\frac{d}{d\theta}L(\theta) = (1-\theta)^{k-1} - (k-1)(1-\theta)^{k-2}\theta = 0$$

$$(1-\theta)^{k-1} = (k-1)(1-\theta)^{k-2}\theta$$

$$1-\theta = k\theta - \theta$$

$$\hat{\theta} = \frac{1}{k}$$

The MLE of p

Assignments

- Finish Chapter 7 of the textbook
- Next time: Maximum likelihood estimate, Bayesian inference

See you next time

See You!