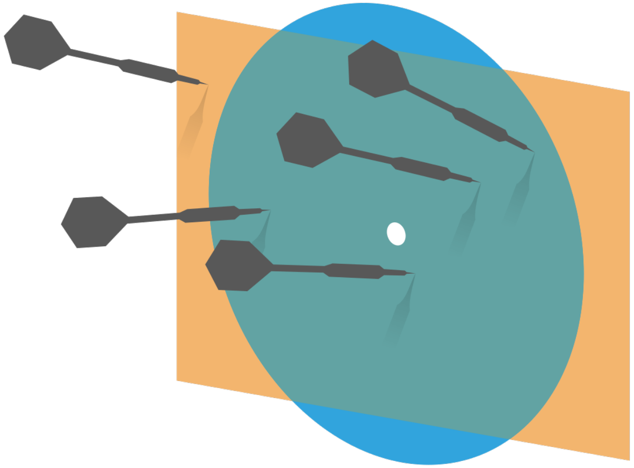
Probability and Statistics for Computer Science



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, 3.18.2021

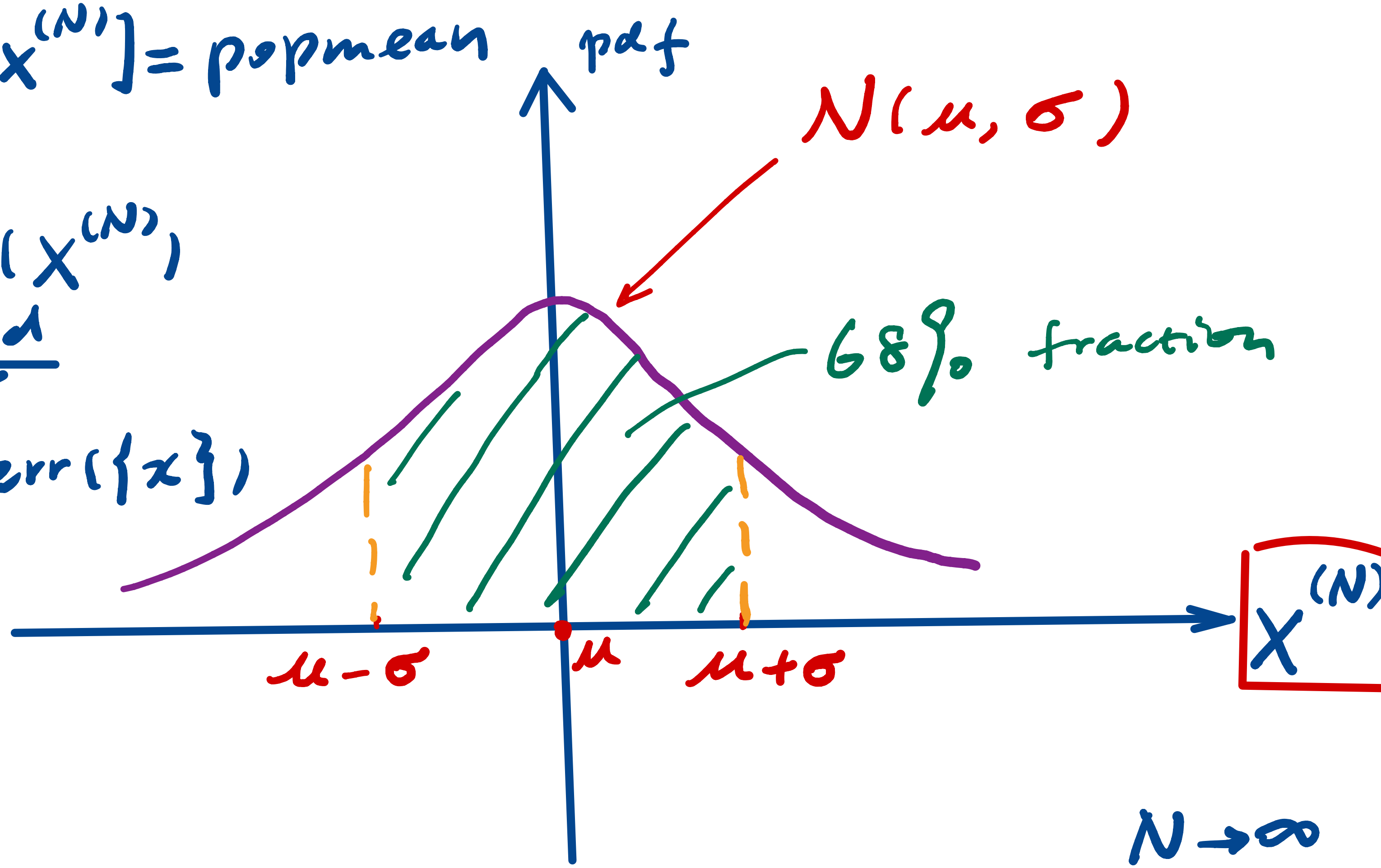
Meaning of #% Confidence Interval

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

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

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

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



Meaning of #% Confidence Interval

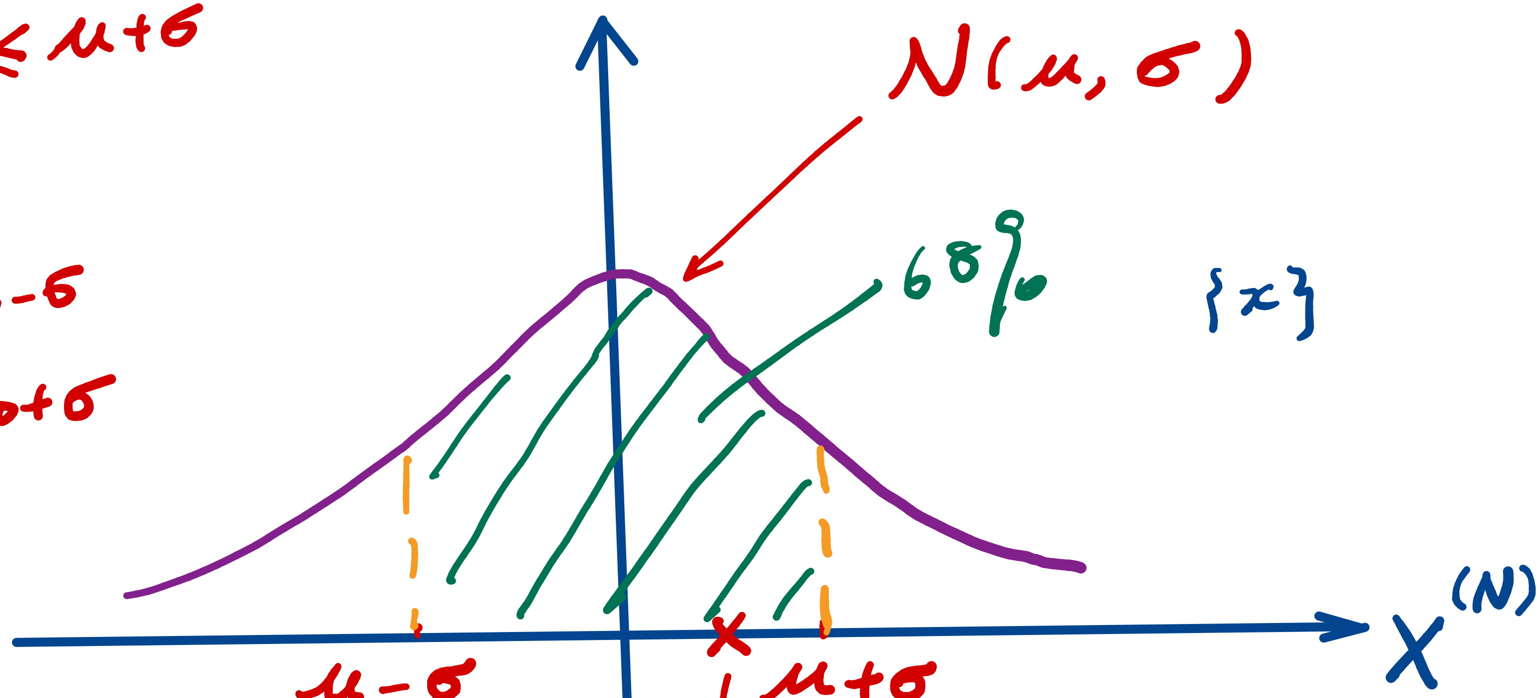
$\mu = E[X^{(N)}] = \text{popmean}$

$\mu - \sigma \le x_0 \le \mu + \sigma$

↓

$\mu \ge x_0 - \sigma$

$\mu \le x_0 + \sigma$



x_0 ~ one sample mean value

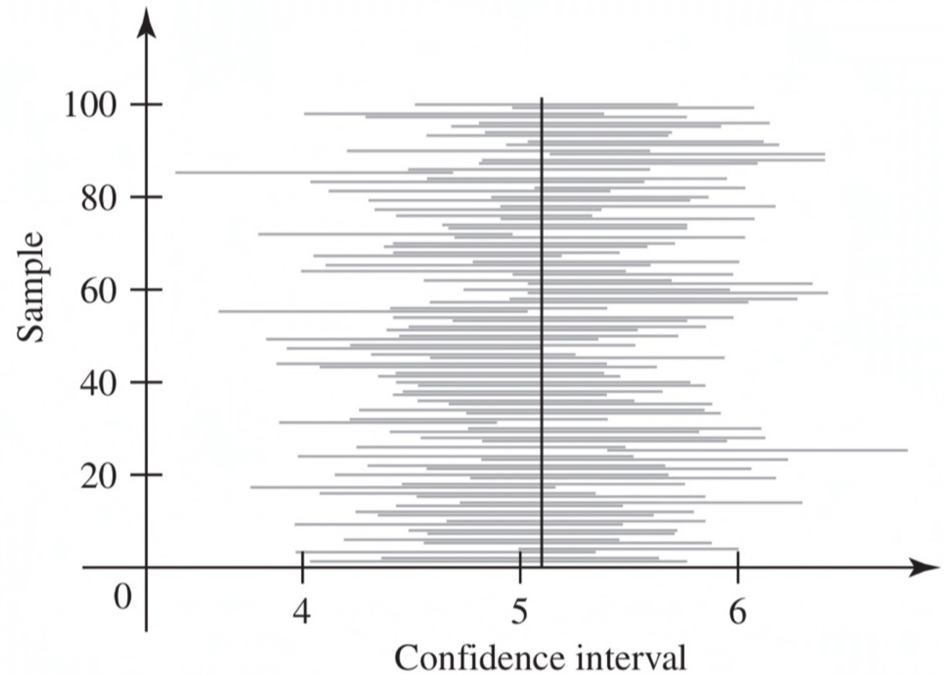
$N \to \infty$

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

↳ $\text{popmean} \in [x_0 - \sigma, x_0 + \sigma]$

Meaning of #% Confidence Inter-val

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

Objectives

- Hypothesis 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 $\neq 55\%$

		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

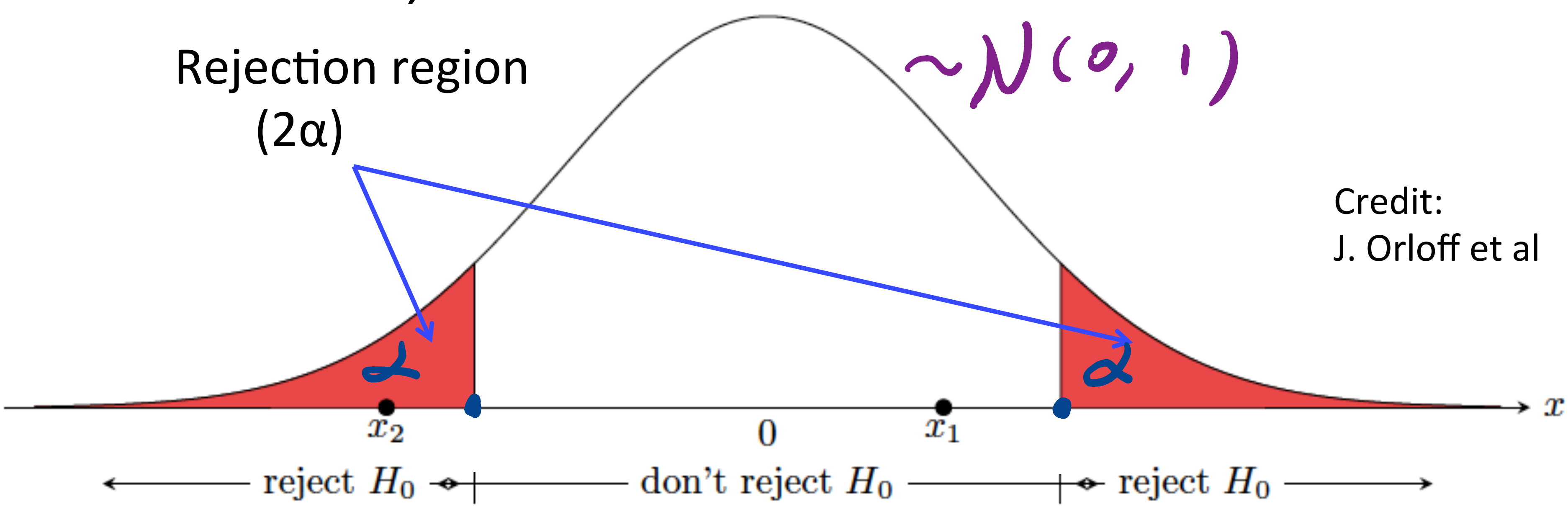
$$x = \frac{(sample\ mean) - (hypothesized\ value)}{standard\ error}$$

mean({x})

stderr({x})

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

~N(0, 1)



Credit:
J. Orloff et al

Fraction of "less extreme" statistic

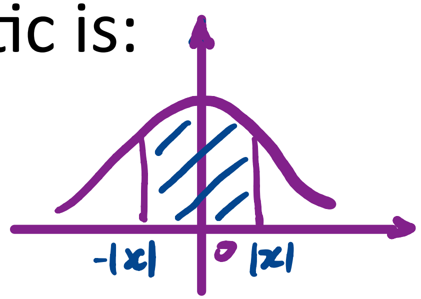
- Assuming the hypothesis H_0 is true $\{x\} = \{10, 20, \ldots, 50\}$
- Define a statistic for the test $\{x\} = \{10, 20, 30\}$

$\text{mean}(\{x\}) = 20$ $\quad v_0 = 50$

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

- Since *N*>30, we assume x comes from a standard normal
- So, the fraction of "less extreme" statistic is:

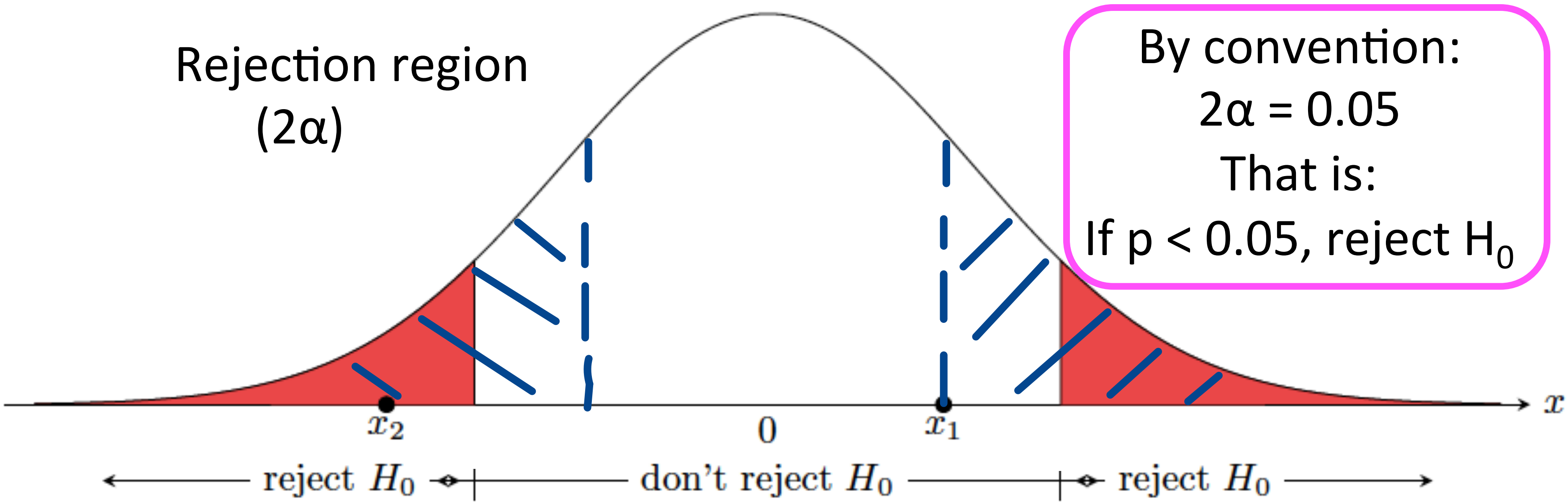
$$f = \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} exp(-\frac{u^2}{2}) du$$



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%
- 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:

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

- 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

The parameter estimation problem

- Suppose we have a dataset that we know comes from a distribution (ie. Binomial, Geometric, or Poisson, etc.)
- What is the best estimate of the parameters (**θ** or **θ**s) of the distribution?
- Examples:
  - For binomial and geometric distribution, **θ** = *p* (probability of success)
  - For Poisson and exponential distributions, **θ** = λ (intensity)
  - For normal distributions, **θ** could be μ or σ^2.

Maximum likelihood estimation

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

p is unknown

write $P(X=k)$ $\downarrow$ write p as θ

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

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

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

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
- We toss it ***N*** times and observe ***k*** heads
- We know that this data comes from a binomial distribution
- What is the likelihood function $L(\theta) = P(D|\theta)$?

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

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

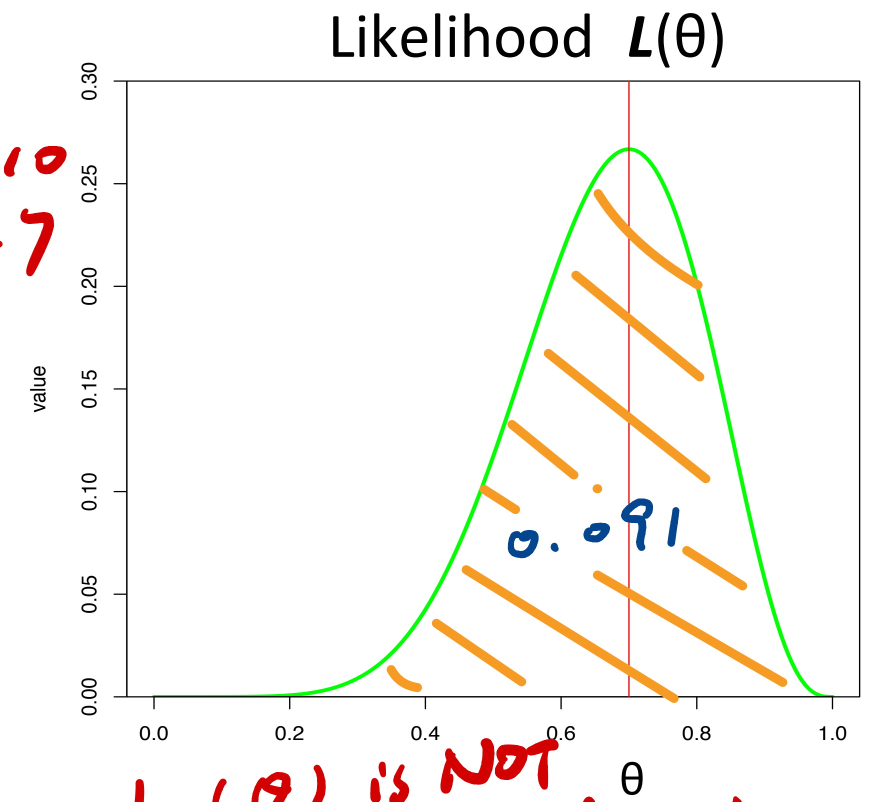
D: N=10
k=7

- 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: $\hat{\theta} = arg\ \underset{\theta}{max}\ L(\theta)$

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

MLE derivation: binomial example

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

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

$a b' + a' b$

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

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

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

p : prob of seeing H

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

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

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

$P(D|\theta)$

what is the D?

$D : k$

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

MLE derivation: geometric example

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

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

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

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

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

The MLE of p

MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

- Each x_i is one observed result from an IID trial

MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

log a×b
= log a + log b

- The likelihood function is hard to differentiate in general, except for the binomial and geometric cases.

- Clever trick: take the (natural) log

Log-likelihood function

- Since log is a strictly increasing function

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

- So we can aim to maximize the **log-likelihood function**

$$logL(\theta) = logP(D|\theta) = log \prod_{x_i \in D} P(x_i|\theta) = \sum_{x_i \in D} logP(x_i|\theta)$$

- The log-likelihood function is usually much easier to differentiate

Log-likelihood function: Poisson example

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

hour	1	2	…	*N*
# of babies	k_1	k_2	…	k_N

- We can assume the data comes from a Poisson distribution λ

- What is the log likelihood function $LogL(\theta)$?

Log-likelihood function: Poisson example

ith hour, k_i

D: N, k_i

$$L(\theta) = \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{k_i}}{k_i!}$$

$$log\ L(\theta) = log\ (\prod_{i=1}^{N} \frac{e^{-\theta}\theta^{k_i}}{k_i!}) = \sum_{i=1}^{N} log(\frac{e^{-\theta}\theta^{k_i}}{k_i!})$$

$$= \sum_{i=1}^{N} (-\theta + k_i\ log\theta - log\ k_i!)$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}(-\theta + k_i \; log\theta - log \; k_i!)$$

$$\frac{d}{d\theta} log \; L(\theta) = 0 \; \Rightarrow \sum_{i=1}^{N}(-1 + \frac{k_i}{\theta} - 0) = 0$$

$$-N + \frac{\sum_i^N k_i}{\theta} = 0$$

$$\hat{\theta} = \frac{\sum_i^N k_i}{N}$$

The MLE of λ

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed $\mathcal{N}(\mu, \sigma)$
- What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution?

  A. Yes B. No

x_i : height of one student

$\{x_i\}$

i.i.d trials

$P(D \mid \theta(\mu, \sigma))$

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed

- What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution? **Yes and No**. The idea is similar but the normal distribution is continuous, we need to use the **probability density** instead.

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed
- The likelihood function of a normal distribution:

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(x_i - \mu)^2}{2\sigma^2})$$

$$\text{Log } L(\theta) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed
- There are two parameters to estimate: $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$
  - If we fix $\boldsymbol{\sigma}$ and set θ= $\boldsymbol{\mu}$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - If we fix $\boldsymbol{\mu}$ and set θ= $\boldsymbol{\sigma}$

$$\hat{\theta} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Drawbacks of MLE

- Maximizing some likelihood or log-likelihood function is mathematically hard
- If there are very few data items, the MLE estimate maybe very unreliable
  - If we observe 3 heads in 10 coin tosses, should we accept that p(heads)= 0.3 ?
  - If we observe 0 heads in 2 coin tosses, should we accept that p(heads)= 0 ?

Confidence intervals for MLE estimates

- An MLE parameter estimate $\hat{\theta}$ depends on the data that was observed

- We can construct a confidence interval for $\hat{\theta}$ using the parametric bootstrap
  - Use the distribution with parameter $\hat{\theta}$ to generate a large number of bootstrap samples
  - From each "synthetic" dataset, re-estimate the parameter using MLE
  - Use the histogram of these re-estimates to construct a confidence interval

Assignments

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

Additional References

- Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. “Probability and Statistical Inference”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

See you next time

See You!