Computer Science

CSC380: Principles of Data Science

Statistics 1

Chicheng Zhang

Announcements

- Expectations of HW code submission
  - For questions that ask ‘paste your code’, we ask you to paste your code in your solution PDF file, *in addition to* submitting to ‘HW code’

- My (Chicheng’s) office hours will be hybrid up until Feb 21; see Piazza

- HW2 due Feb 8 (Wed)

- HW3 out Feb 9 (Thu)

HW1 Problem 3

- Two players, 7 rounds. Fair game.
- W := the event of you winning the game
- S=(i,j) := you won i times and the opponent won j times.
- Goal: Compute $a_{i,j} = P(W|S = (i,j))$ for every i and j.

1) If $i = 4$ and $j < 4$, $a_{i,j} = 1$. OTOH, if $i < 4$ and $j = 4$, $a_{i,j} = 0$. Can you see why?

2) Let R_i be a random variable where $R_i = 1$ if you win round i and $R_i = 0$ if you lose that round. Note that $P(R_i = 0) = P(R_i = 1) = \frac{1}{2}$.

3) Recall that by the law of total probability $P(W \mid S = (i,j)) = P(W, R_{i+j+1} = 1 \mid S = (i,j)) + P(W, R_{i+j+1} = 0 \mid S = (i,j))$.

4) By the probability chain rule $P(W, R_{i+j+1} \mid S = (i,j)) = P(W \mid R_{i+j+1}, S = (i,j))P(R_{i+j+1} \mid S = (i,j))$.

5) Although it requires rigorous argument, for this problem, you can take it as given that $P(W \mid R_{i+j+1} = 1, S = (i,j)) = P(W \mid S = (i+1,j))$ and $P(W \mid R_{i+j+1} = 0, S = (i,j)) = P(W \mid S = (i,j+1))$. (Can you see why, intuitively?)

6) Find a way to write down $a_{i,j}$ as a function of $a_{i+1,j}$ and $a_{i,j+1}$. This will help you compute the answers in a recursive manner.

HW1 Problem 3

$$
\begin{aligned}
P(W \mid S=(i,j)) =& P(W, R_{i+j+1}=1 \mid S=(i,j)) + P(W, R_{i+j+1}=0 \mid S=(i,j)) && \text{[Hint 3]} \\
=& P(W \mid R_{i+j+1}=1, S=(i,j))P(R_{i+j+1}=1 \mid S=(i,j)) \\
& + P(W \mid R_{i+j+1}=0, S=(i,j))P(R_{i+j+1}=0 \mid S=(i,j)) && \text{[Hint 4]} \\
=& P(W \mid S=(i+1,j)) \times \frac{1}{2} + P(W \mid S=(i,j+1)) \times \frac{1}{2} && \text{[Hint 5]} \\
=& \frac{1}{2} \times \Big(P(W \mid S=(i+1,j)) + P(W \mid S=(i,j+1)) \Big)
\end{aligned}
$$

HW1 Problem 3

- Another approach for calculating $a_{i,j} = P(W|S = (i,j))$

- Let's say we play all 7 games anyways

- Conditioned on $S = (i,j)$,

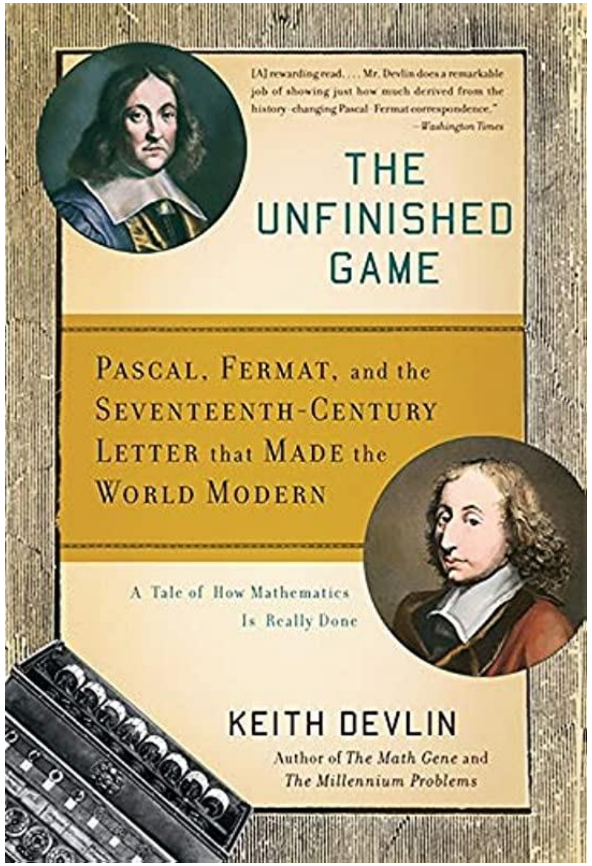
  You win $\Leftrightarrow$ You win at least $4 - i$ out of the remaining $7 - i - j$ rounds

- E.g. $a_{3,2} = P(X \geq 1)$ for $X \sim \text{Bin}(2, 0.5)$

HW1 Problem 3

- "Problem of points"
- https://en.wikipedia.org/wiki/Problem_of_points

- Motivates modern probability theory



Probability and Statistics

- Probability provides a mathematical formalism to reason about **random events**
  - Knowing the distribution, how can we compute probability of the event of interest? (e.g., two fair dice, P(sum = 3 | X_1 = 1))

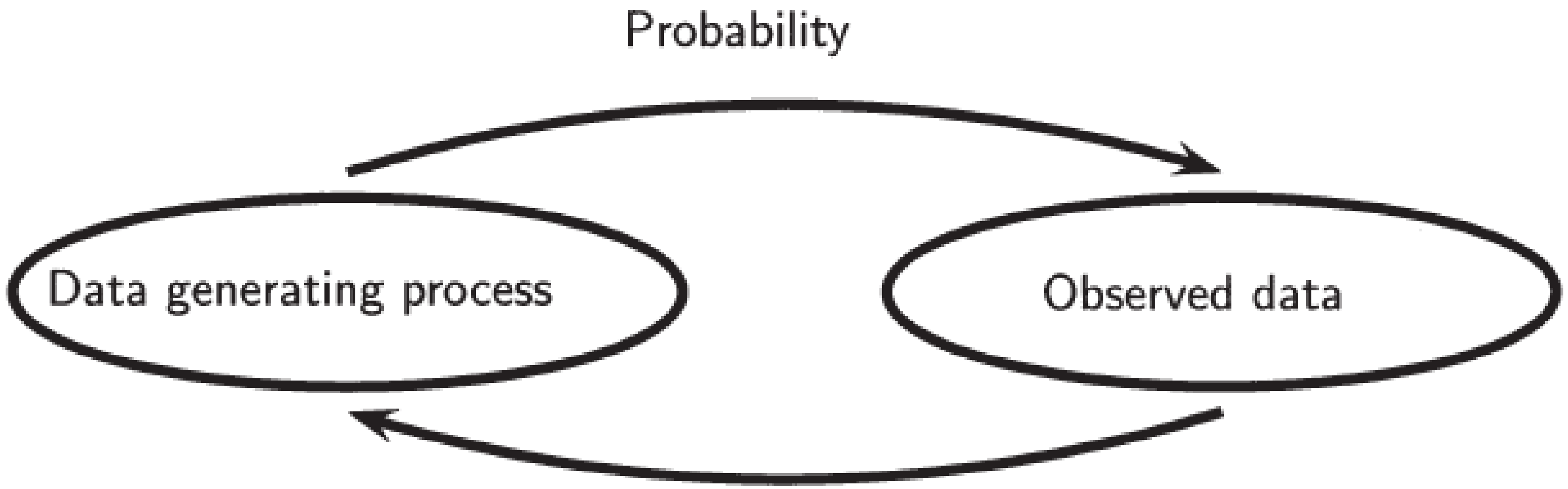
- Statistics is centered on **data**
  - Fitting models to data (estimation)
  - **E.g.,** I don't know the distribution, but I have samples drawn from it. Let's estimate what the distribution is! ⇒ **reverse engineering!**
  - Answering questions from data (statistical inference, hypothesis testing)
  - Interpretation of data

- Statistics *uses* probability to address these tasks

*Probability: **<u>Given a distribution</u>**, compute probabilities of data/events.*

e.g., data = outcome of coin flip

E.g., If $X_1, \dots, X_{10}$ ~ Bernoulli(p=.1), what is the probability of $\sum_{i=1}^{10} X_i \geq 3$?



Inference / Estimation

E.g., We observed $X_1, \dots, X_{10} \in \{0,1\}$. What is the head probability?

*Statistics: **<u>Given data</u>**, compute/infer the distribution or its properties.*

[Source: Wasserman, L. 2004]

Intuition Check

Suppose that we toss a coin 100 times. We don't know if the coin is fair or biased…

<u>Question 1</u> Suppose that we observe **<u>52</u>** heads and **<u>48</u>** tails. Is the coin fair? Why or why not?

<u>Question 2</u> Now suppose that out of 100 tosses we observed **<u>73</u>** heads and **<u>27</u>** tails. Is the coin fair? Why or why not?

<u>Question 3</u> How might we estimate the bias of the coin with **<u>73</u>** heads and **<u>27</u>** tails?

Estimating Coin Bias

We can model each coin toss as a Bernoulli random variable,

$$X \sim \text{Bernoulli}(\pi) \quad \Rightarrow \quad p(X = x) = \pi^x (1 - \pi)^{1-x}$$

Recall that π is the coin bias (probability of heads) and that,

$$\mathbf{E}[X] = \pi$$

Suppose we observe N coin flips $x_1, \ldots, x_N$, estimate π as,

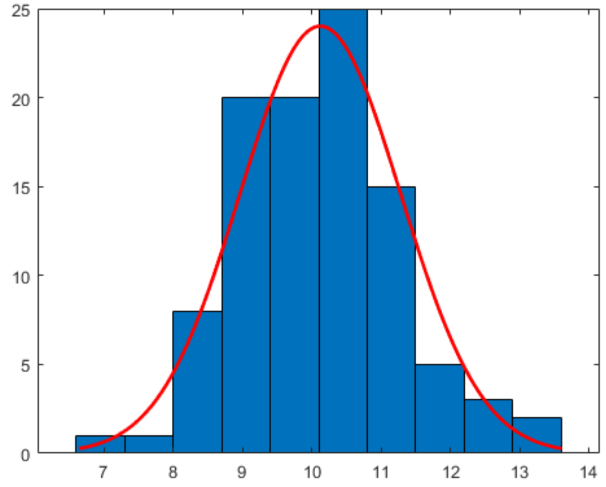
$$\hat{\pi} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

This is called <u>empirical mean</u> or <u>sample mean</u>

More generally, can use sample average of $f(x_i)$'s to estimate $\mathrm{E}[f(X)]$ (plug-in principle)

Estimating Gaussian Parameters

Suppose we observe the heights of N students at UA, and we model them as Gaussian:



(A.K.A. Normal)

$$\{x_i\}_i^N \sim \mathcal{N}(\mu, \sigma^2)$$

How can we estimate μ?

$$\mu = \mathrm{E}[X] \approx \frac{1}{N}\sum_i x_i$$

Estimate μ using sample mean

$$\hat{\mu} = \frac{1}{N}\sum_i x_i \text{ (abbrev. } \bar{x}\text{)}$$

How can we estimate σ**?**

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}[(X-\mu)^2] \approx \frac{1}{N}\sum_i (x_i - \mu)^2 \approx \frac{1}{N}\sum_i (x_i - \hat{\mu})^2$$

Estimate σ using

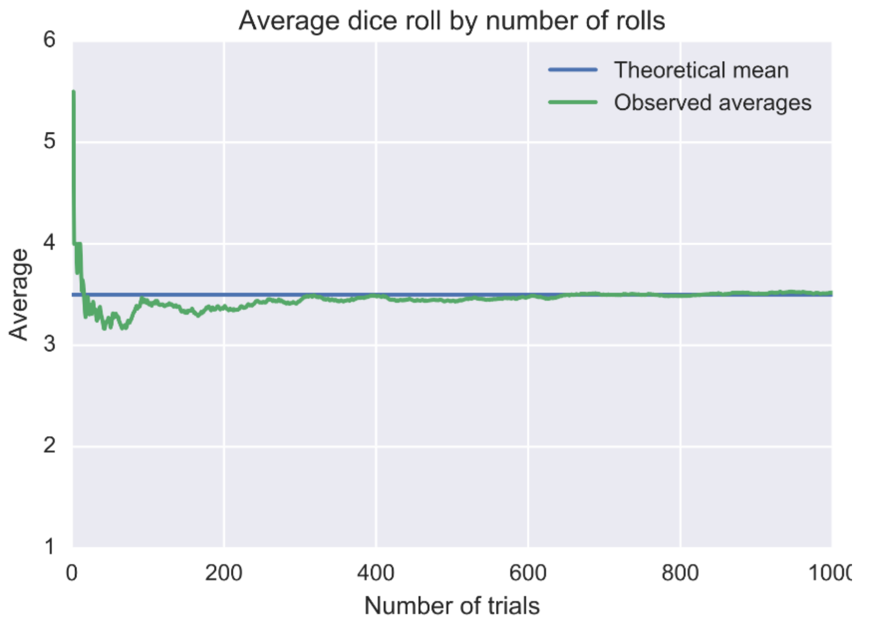
$$\hat{\sigma} = \sqrt{\frac{1}{N}\sum_i (x_i - \hat{\mu})^2}$$

Probability tool: Law of Large Numbers (LLN)

Claim: sample mean converges to the true mean.

(Theorem) Let $X_1, \ldots, X_N, \ldots$ be drawn iid from a distribution with mean μ. Let $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ be the sample mean. Then

$$\lim_{N\to\infty} \hat{\mu}_N = \mu$$



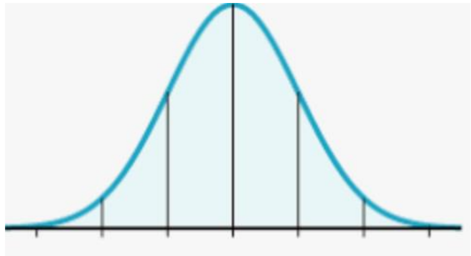
This is the **law of large numbers**

- Weak Law: Converges to mean with high probability
- Strong Law: Stronger notion of convergence; will converge at all times! (if variance is finite)

Limitation: it does not say how fast it will converge!

Probability tool: Central Limit Theorem (CLT)

Let $X_1, \ldots, X_N$ be iid with mean μ and variance σ^2 then the sample mean $\bar{X}_N$ approaches a Normal distribution



$$\lim_{N\to\infty} \bar{X}_N \to \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

=> the convergence rate is $\frac{\sigma}{\sqrt{N}}$!!

Actually, a mathematically rigorous version is

$$\lim_{N\to\infty} \frac{\sqrt{N}}{\sigma}(\bar{X}_N - \mu) \to \mathcal{N}(0,1)$$

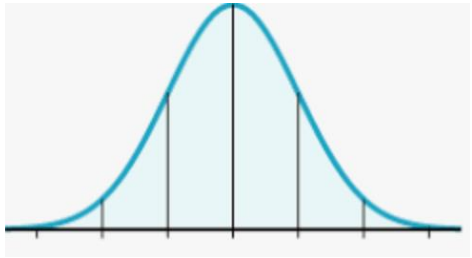
Comments

- LLN says estimates $\bar{X}_N$ "pile up" near true mean, CLT says *how* they pile up
- Very remarkable since we make **no assumption about how X_i are distributed**
- Variance of X_i **must be finite**, i.e. $\sigma^2 < \infty$ (e.g., Cauchy distribution has $\sigma^2 = \infty$)

https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/normal-distributions-library/a/normal-distributions-review

Central Limit Theorem (CLT): sanity check

- Let $X_1, \ldots, X_N$ be drawn iid from $\mathcal{N}(\mu, \sigma^2)$
- What's the distribution of $\bar{X}_N$?



$$\Rightarrow \sum_{i=1}^{N} X_i \sim \mathcal{N}\left(N\mu, N\sigma^2\right)$$

$$\Rightarrow \bar{X}_N \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

$$\Leftrightarrow \frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} \sim \mathcal{N}(0,1)$$

Recall: for normal distributions

- Closed under independent addition:

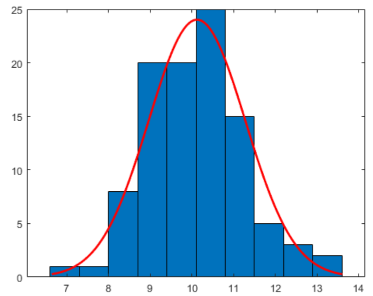
$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2) \ \ , X \perp Y$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under affine transformation (a and b constant):

$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$

Parameter Estimation



We *pose* a model in the form of a probability distribution, with unknown **parameters of interest** θ **,**

$$\mathcal{D}_\theta$$

e.g., assume Gaussian: $\theta = (\mu, \sigma^2)$

Observe data, typically *independent identically distributed (iid)*,

$$p(X_1 = x_1, \ldots, X_N = x_n) = p(X_1 = x_1) \cdots p(X_N = x_N)$$

$$x_1, \ldots, x_N \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\theta \,,$$

Compute an **estimator** to estimate parameters of interest,

$$\hat{\theta}(\{x_i\}_i^N) \approx \theta$$

Many different types of estimators, each with different properties

Examples of Non-I.I.D. data

- A and B are independent but **non**identically distributed
  - E.g., two coin flips A and B with P(A=H) = ½ and P(B=H) = 1/4

Examples of Non-I.I.D.

Dependent identical distribution

- First coin (X1): fair coin
- Second coin (X2):
  - if X1=H, throw an **unfair** coin P(H) = ¼, P(T) = ¾
  - If X1=T, throw an **unfair** coin P(H) = ¾ , P(T) = ¼

(joint probability table)

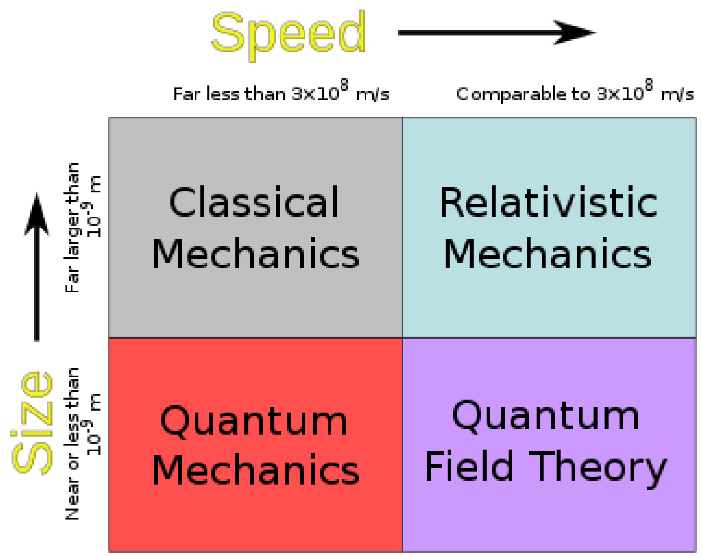
	B=H	**B=T**	
A=H	1/8	3/8	1/2
A=T	3/8	1/8	1/2
	1/2	1/2	

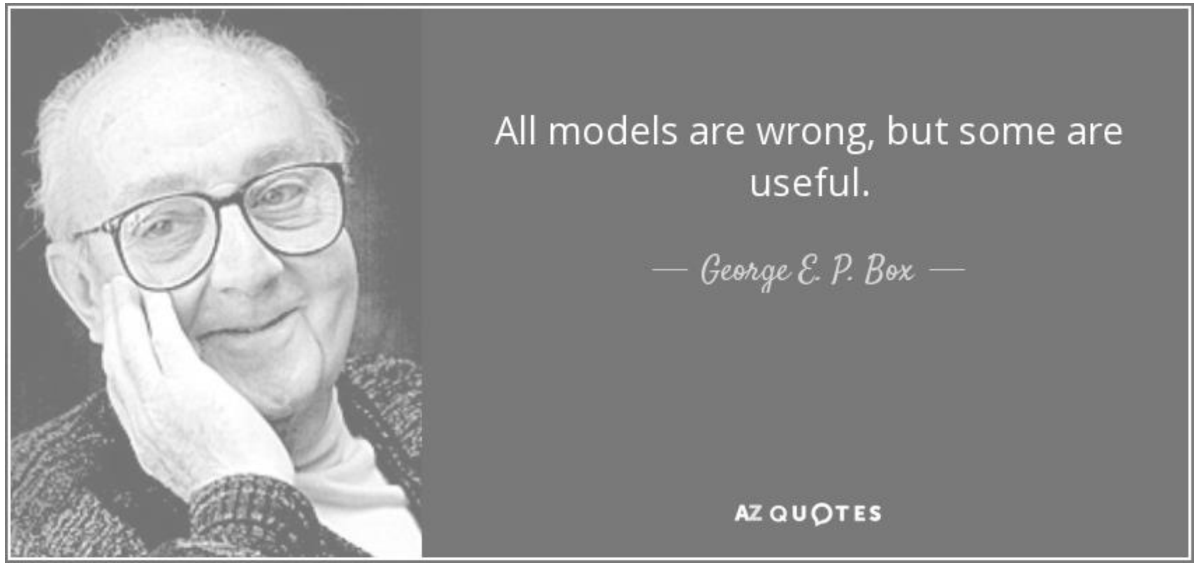
  - P(A=H)=P(B=H) but A and B are not independent (prove it!)

In general, i.i.d. is necessary to have estimators close to the true parameter

Models

- In the previous example, we assumed that the heights follow a normal distribution.
- Does it?
- Another example: Physics





- There are ways to check if one model is better than the other (will be covered much later)

Definitions

*A **statistic** is a function of the data that does not depend on any unknown parameter.*

Examples

- Sample mean $\bar{x}$
- Sample variance s^2 or $\hat{\sigma}^2$ Note: σ^2
- Sample STDEV s or $\hat{\sigma}$
- Standardized scores $(x_i - \bar{x})/s$
- Order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
- Sample (noncentral) moments $\bar{x}^m = \frac{1}{n}\sum_{i=1}^{n} x_i^m$

*An **estimator** $\hat{\theta}(x)$ is a **statistic** used to infer the unknown parameters of a statistical model.*

Q: Gaussian distribution with unknown mean and variance. Which of these are estimators?

Intuition Check

Suppose that we toss a coin 100 times. We observe 52 heads and 48 tails…

Question 1 I define an estimator that is *always* $\hat{\theta} = 0$, regardless of the observation. Is this an estimator? Why or why not?

Question 2 Is the estimator above a **good** estimator? Why or why not?

Question 3 What are some properties that could define a **good** estimator?

Two Desirable Estimator Properties

- **Consistency (asymptotic notion)** Given enough data, the estimator *converges* to the true parameter value

$$\lim_{n \to \infty} \hat{\theta}(x_1, \ldots, x_n) \to \theta$$

This convergence can be measured in a number of ways: in probability, in distribution, absolutely

A bare minimum requirement!
Otherwise, you may collect more data that will give us a worse estimator!

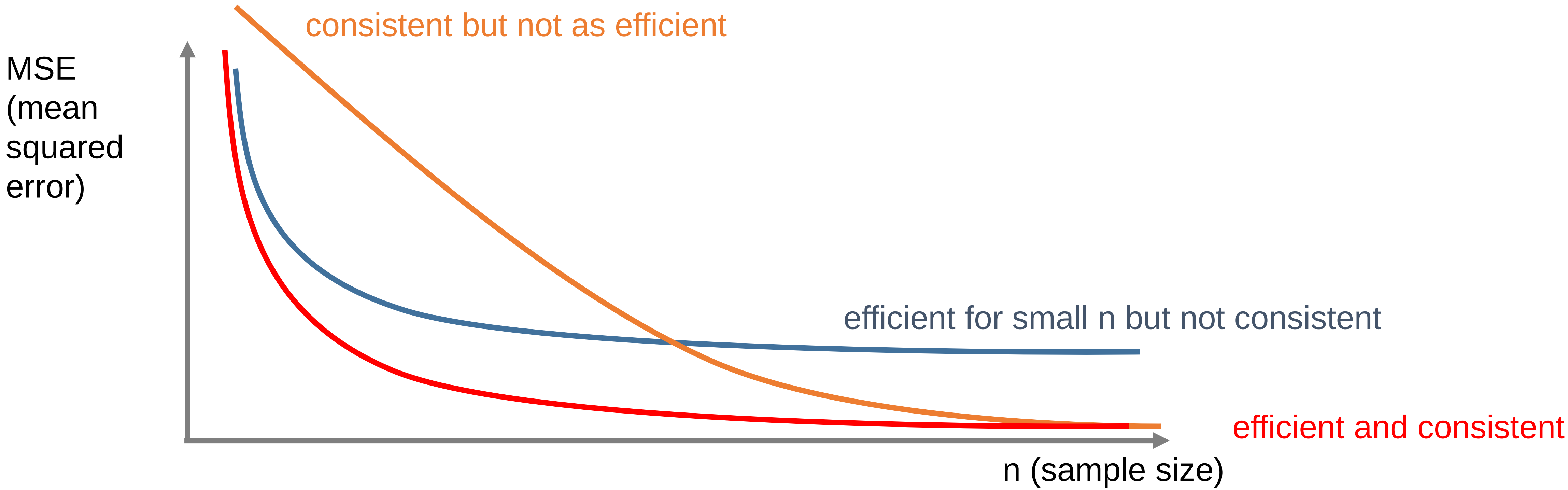
- **Efficiency (nonasymptotic notion)** It should have low error with finite n, e.g.

$$\mathrm{MSE}(\hat{\theta}_n) \coloneqq \mathrm{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]$$

Mean squared error should be small

looks like variance but it's different!
Q: spot the difference from $\mathrm{Var}(\hat{\theta}_n)$?

Two Desirable Estimator Properties



Unbiasedness - Another Desirable Property

- **Unbiasedness**: For any n, $\mathbf{E}\left[\hat{\theta}(X_1, \ldots, X_n)\right] = \theta$
  - E.g., sample mean is unbiased. If $X_1, \ldots, X_n \sim D$ with $\mathbf{E}_{X \sim D}[X] = \mu$

$$\mathbf{E}[\bar{X}_N] = \frac{1}{N} \sum_i \mathbf{E}[X_i] = \mu$$

- Traditionally, considered to be a good property.
- In modern statistics, **not a necessary condition** to be a good estimator.
  - An unbiased estimator may be **less efficient** compared to some other **biased** estimator.

- Biased estimators can still be **consistent**.
- Consistency ≈ asymptotically unbiased.

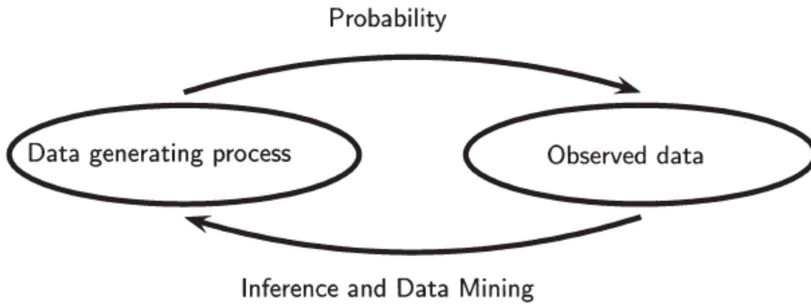
E.g., for some estimator $E\left[\hat{\theta}(X_1, \ldots, X_n)\right]$ can be $\mu + \frac{1}{n}$

Computer Science

CSC380: Principles of Data Science

Statistics 2

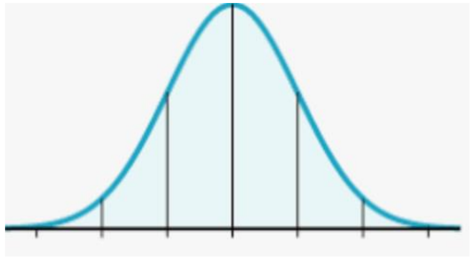
Review from last lecture



- Statistics: **<u>Given data</u>**, compute/infer the distribution or its properties.

- Probability tools: Law of Large Numbers (LLN), Central Limit Theorem (CLT)
  - Justifies the "Plug-in principle" for estimation

- Basic setup of *estimation*:
  - Data $x_1, \ldots, x_N \sim \mathcal{D}_\theta$, $\{\mathcal{D}_{\theta'}: \theta' \in \Theta\}$: class of models parameterized by $\theta' \in \Theta$
  - Estimator: $\hat{\theta}(x_1, \ldots, x_N)$



- Desirable properties of estimators: consistency, efficiency (MSE), unbiasedness

This lecture

- Maximum likelihood estimation (MLE)

- Basic statistical properties of simple estimators: sample mean and sample variance

Intuition Check

Suppose that we toss a coin 100 times. We observe 73 heads and 27 tails…

<u>Question</u> Let θ be the coin bias (probability of heads). What is a more likely estimate? What is your reasoning?

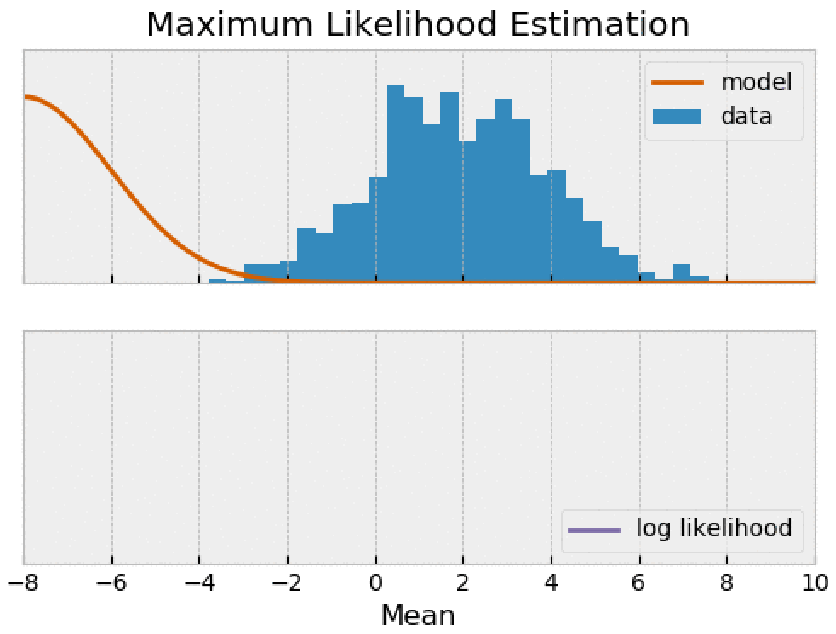
A: $\hat{\theta} = 0.73$, strong preference for heads

B: $\hat{\theta} = 0.50$, fair coin (we observed unlucky outcomes)

Likelihood (informally) Probability/density of the observed outcomes from a particular model

Likelihood (Intuitively)

Suppose we observe N data points from a Gaussian model $\mathcal{N}(\mu, \sigma^2)$ and wish to estimate its mean parameter μ



Likelihood Principle: *Given a statistical model, the <u>likelihood function</u> describes all evidence of a parameter that is contained in the data.*

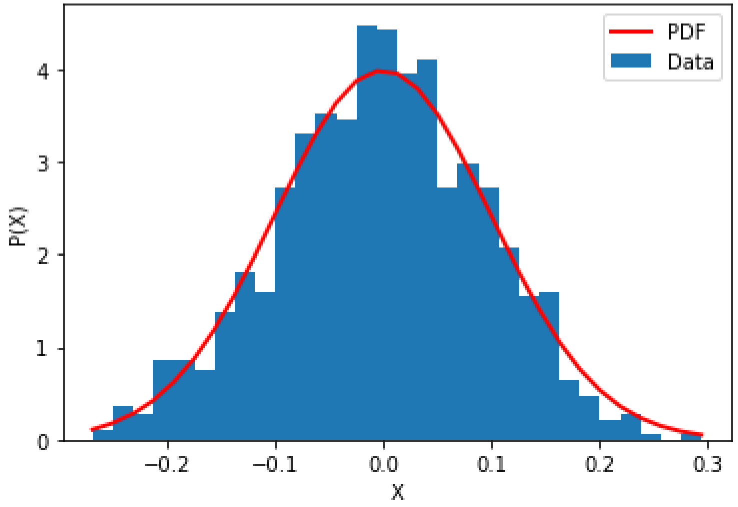
https://towardsdatascience.com/maximum-likelihood-estimation-984af2dcfcac

Likelihood: another example

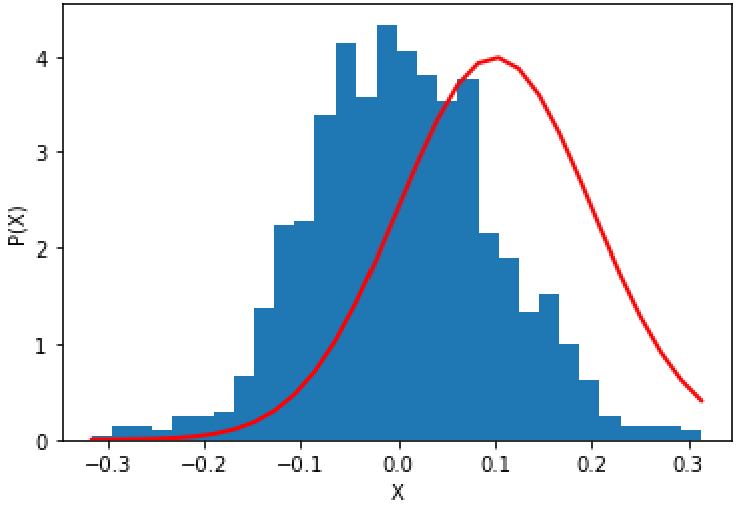
Suppose we observe N data points from a Gaussian model $\mathcal{N}(\mu, \sigma^2)$ *and wish to estimate* ***both μ and σ***

Say we only need to choose from the following three Gaussians…

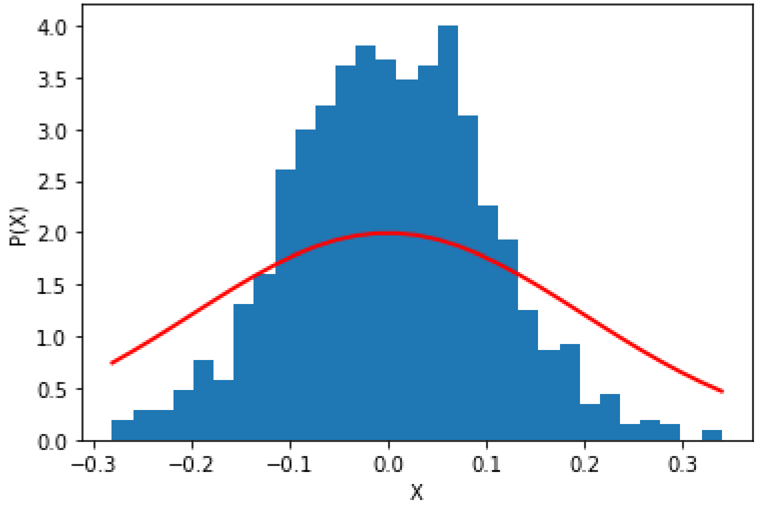
High Likelihood



Low Likelihood (mean)



Low Likelihood (variance)



Likelihood Function

Suppose $x_i \sim p(x;\theta)$, then what is the **joint probability** over N *independent identically distributed* (iid) observations $x_1, \ldots, x_N$?

$$p(x_1, \ldots, x_N; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

what appears after ';' are parameters, not random variables.

- We call this the **likelihood function**, often denoted $\mathcal{L}_N(\theta)$
- It is a function of the parameter θ, the data are fixed
- Describes how well parameter θ describes data (*goodness of fit*)

How could we use this to estimate a parameter θ*?*

Maximum Likelihood

Maximum Likelihood Estimator (MLE) as the name suggests, finds the parameter θ that maximizes the likelihood function.

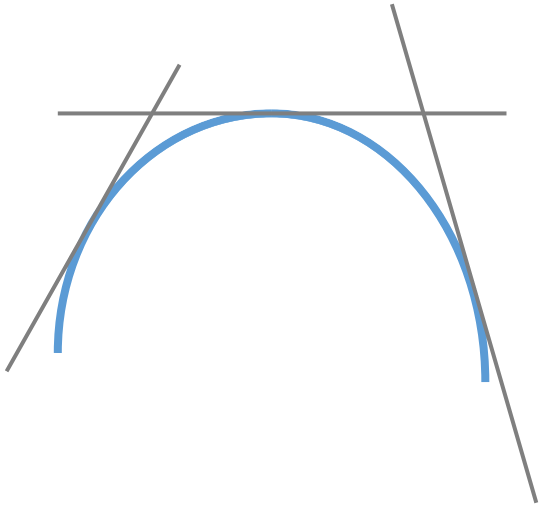
$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \mathcal{L}_N(\theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

Question How do we find the MLE?

1. closed-form
2. iterative methods

Finding the maximum/maximizer of a function

Example: Suppose $f(\theta) = -a\theta^2 + b\theta + c$ with $a > 0$

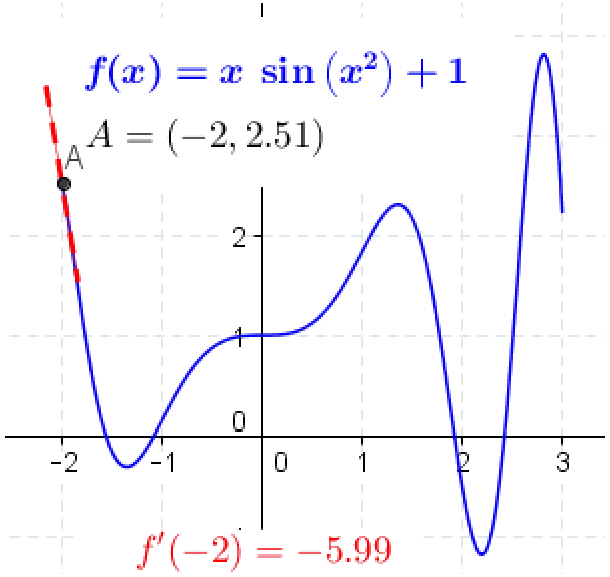


It is a quadratic function.
=> finding the 'flat' point suffices

Compute the gradient and set it equal to 0 (stationary points)

$$f'(\theta) = -2a\theta + b \quad \Rightarrow \quad \theta = \frac{b}{2a}$$

Closed form!



(gradient illustration)

Q: Does this trick of grad=0 work for other functions?
⇒ Yes for **<u>concave</u>** functions!
⇒ Roughly speaking, functions that curves down only, never upwards

Finding the maximum/maximizer of a function

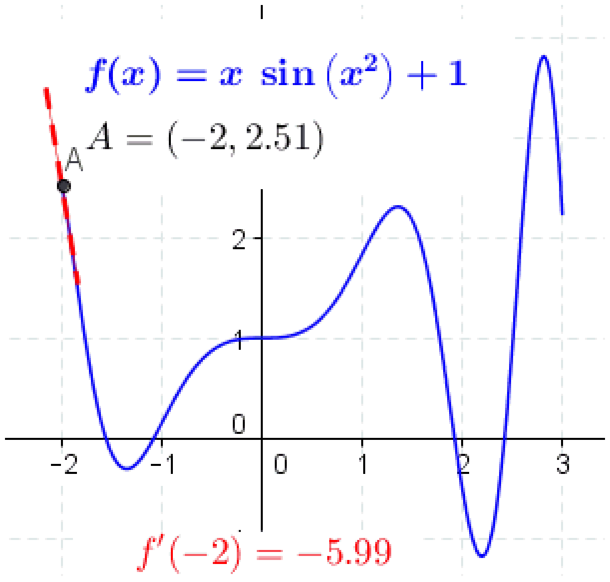
What if there is no closed form solution?

Iterative methods for optimization
=> Will find the global maximum
for **concave** functions (convex optimization)
$\Rightarrow$ More generally, finds a local maximum but could also get stuck at *stationary point*.

Example: $f(\theta) = \frac{1}{2}x(ax - 2\log(x) + 2)$

$$f'(\theta) = ax - \log(x)$$

No known closed form for $ax = \log(x)$



Iterative methods:
- Hillclimbing - gradient ascent (or *descent* if you are minimizing)
- Newton's method
- Etc. (beyond the scope of our class)

Q: find the local maxima and global maximum

Maximum Likelihood

Maximizing **log**-likelihood makes the math easier (as we will see) and doesn't change the answer (logarithm is an increasing function)

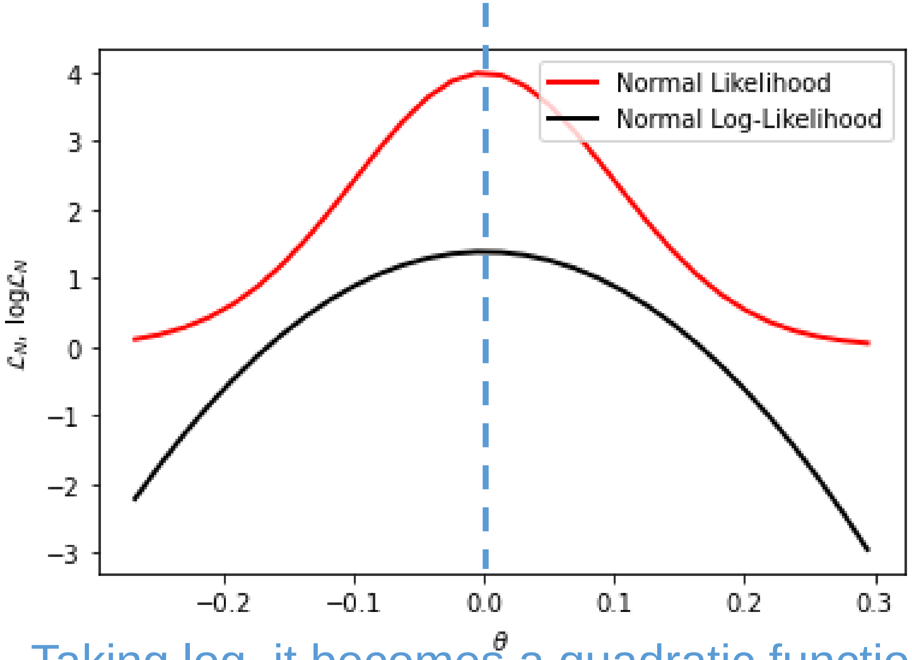
$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log \mathcal{L}_N(\theta) = \sum_{i=1}^{N} \log p(x_i;\theta)$$

Derivative is a linear operator so,

$$\frac{d}{d\theta} \log \mathcal{L}_N(\theta) = \underbrace{\sum_{i=1}^{N} \frac{d}{d\theta} \log p(x_i;\theta)}$$

One term per data point
Can be computed in parallel
(big data)

MLE



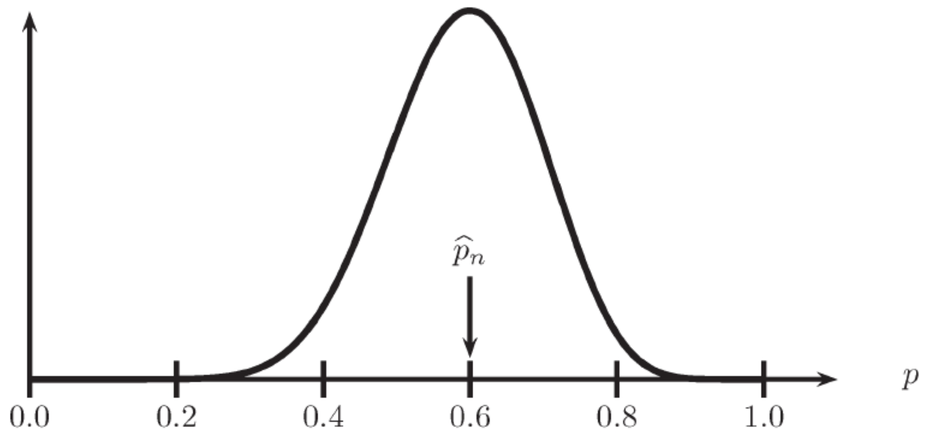
Taking log, it becomes a quadratic function!

Maximum Likelihood: Bernoulli

[Source: Wasserman, L. 2004]

Example N coin tosses with $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. We don't know the coin bias p. The likelihood function is,

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^S(1-p)^{n-S}$$



Likelihood function for Bernoulli with n=20 and $\sum_i x_i = 12$ *heads*

where $S = \sum_i x_i$. The log-likelihood is,

$$\log \mathcal{L}_n(p) = S \log p + (n-S)\log(1-p)$$

Set the derivative of $\log \mathcal{L}_n(p)$ to zero and solve,

$$\hat{p}^{\text{MLE}} = S/n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Maximum likelihood is equivalent to sample mean in Bernoulli

⇒ this showcases how MLE is aligned to our intuition!

Maximum Likelihood: Gaussian

Example Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with parameters $\theta = (\mu, \sigma^2)$ and the likelihood function (ignoring some constants) is:

$$\begin{aligned}
\mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(X_i - \mu)^2\right\} \\
&= \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2\right\} \\
&= \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X} - \mu)^2}{2\sigma^2}\right\}
\end{aligned}$$

Exercise: Show that

$$\sum_i (X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$$

Where $\bar{X} = \frac{1}{n}\sum_i X_i$ and $S^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ are sample mean and sample variance, respectively.

Maximum Likelihood: Gaussian

Continuing, write log-likelihood as:

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X} - \mu)^2}{2\sigma^2}.$$

Solve zero-gradient conditions:

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0,$$

To obtain maximum likelihood estimates of mean / variance:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i} (X_i - \hat{\mu})^2$$

MLE Review: Probability/Density vs Likelihood

- The **probability/density** of data given parameter is mathematically the same object as **likelihood** of a parameter given data

- The difference is the point of view!
  - From the *probabilistic perspective*, the parameter is fixed and **PMF/PDF** is viewed as a function of the possible data

  - From the *statistical perspective*, the data is given (thus fixed) and we view **likelihood** as a function of the parameter.

- Statistics is inherently about reverse engineering.

MLE

- MLE is a very important tool.
- Usually, you write a function that computes log likelihood and then you will use existing libraries (e.g., cvxpy) to find the maximizer (next slide)
- There are efforts to develop ‘probabilistic programming’ (next slide)

CVXPY

```
# Import packages.
import cvxpy as cp
import numpy as np

# Generate data.
m = 20
n = 15
np.random.seed(1)
A = np.random.randn(m, n)
b = np.random.randn(m)

# Define and solve the CVXPY problem.
x = cp.Variable(n)
cost = cp.sum_squares(A @ x - b)
prob = cp.Problem(cp.Minimize(cost))
prob.solve()

# Print result.
print("\nThe optimal value is", prob.value)
print("The optimal x is")
print(x.value)
```

Not the usual variable, but it is an object specifically designed to work with optimization algorithms.

Cost is not an actual value; it is an object of cvxpy that encodes the operation of $\sum_i \left(\langle A_{i,\cdot}, x\rangle - b_i\right)^2$ as a tree.

Alternative: cp.Maximize

cvxpy.Problem has many numerical methods to find the optimal solution

Julia: Turing.jl package

Turing.jl

Bayesian inference with probabilistic programming.

aims to do 'declarative' programming for probabilistic models, just like SQL in databases!

```julia
using Turing
using Optim

@model function gdemo(x)
  σ² ~ InverseGamma(2, 3)
  m ~ Normal(0, sqrt(σ²))

  for i in eachindex(x)
    x[i] ~ Normal(m, sqrt(σ²))
  end
end

# Create some data to pass to the model.
data = [1.5, 2.0]
# Instantiate the gdemo model with our data.
model = gdemo(data)

# Generate a MLE estimate.
mle_estimate = optimize(model, MLE())
```

```
DynamicPPL.Model{typeof(gdemo), (:x,), (), (), Tuple{Vector{Float64}}, Tuple{}, DynamicPPL.DefaultContext}(gdemo, (x = [1.5, 2.0],), NamedTuple(), DynamicPPL.DefaultContext())
```

```
ModeResult with maximized lp of -0.07
2-element Named Vector{Float64}
A   |
----+--------
:σ² | 0.0625
:m  |   1.75

# to access the value
mle_estimate.values[:σ²]
```

Maximum Likelihood Properties

Under some mild assumptions on the model class $\{\mathcal{D}_{\theta'}: \theta' \in \Theta\}$:

1) The MLE is a **consistent** estimator:

$$\lim_{n\to\infty} \hat{\theta}_n^{\mathrm{MLE}} \xrightarrow{P} \theta_*$$

Roughly, converges to the true value.

2) The MLE is an **asymptotically efficient**: roughly, has the lowest mean squared error among all consistent estimators.

3) The MLE is an **asymptotically Normal**: roughly, the estimator (which is a <u>random variable</u>) approaches a Normal distribution (more later).

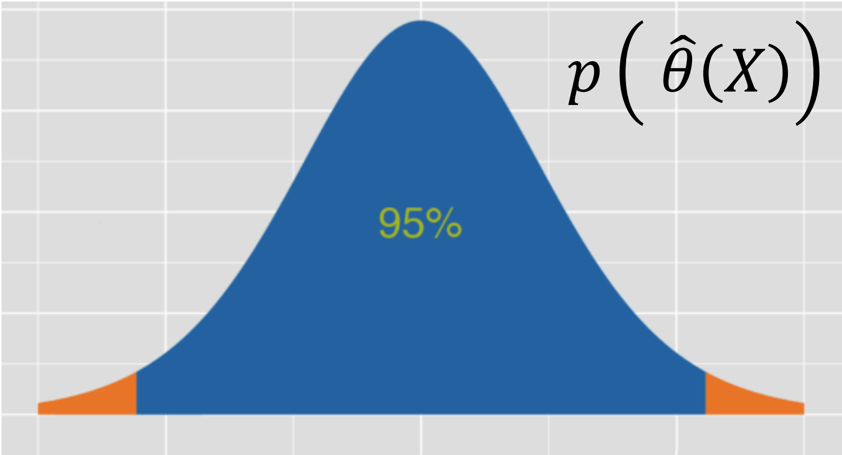
4) The MLE is **functionally invariant**: if $\hat{\theta}^{\mathrm{MLE}}$ is the MLE of θ then $g(\hat{\theta}^{\mathrm{MLE}})$ is the MLE of $g(\theta)$.

Expectation of the Sample Mean

Recall: An estimator $\hat{\theta}$ is a RV (Random Variable).

Example Let $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ and estimate $\hat{p}$ be the *sample mean*,

$$\hat{p} = \frac{1}{N} \sum_i X_i$$



Question Is $\hat{p}$ *an unbiased estimator of* p?

Notation: $X \coloneqq (X_1, \ldots, X_N)$

$$\mathbf{E}[\hat{p}(X)] = \mathbf{E}\left[\frac{1}{N}\sum_i X_i\right] \overset{\text{(a)}}{=} \frac{1}{N}\sum_i \mathbf{E}[X_i] \overset{(b)}{=} \frac{1}{N} Np = p$$

(a) Linearity of Expectation Operator

(b) Mean of Bernoulli RV = p

Conclusion *On average $\hat{p} = p$ (it is unbiased);*

In general *sample mean unbiasedly estimates mean (not nec. Bernoulli)*

Variance of the Sample Mean

Example Let $X_1, \ldots, X_N \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$ and estimate $\hat{p}$ be the *sample mean.* Calculate the variance of $\hat{p}$:

quiz candidate

$$\mathbf{Var}(\hat{p}) = \mathbf{Var}\left(\frac{1}{N}\sum_i X_i\right) \stackrel{(a)}{=} \frac{1}{N^2}\mathbf{Var}\left(\sum_i X_i\right) \stackrel{(b)}{=} \frac{1}{N^2}\sum_i \mathbf{Var}\left(X_i\right)$$

$$\stackrel{(c)}{=} \frac{1}{N^2}\sum_i p(1-p) = \frac{1}{N}p(1-p) = \frac{1}{N}\mathbf{Var}(X)$$

(a) $\mathbf{Var}(cX) = c^2\mathbf{Var}(X)$ (b) Independent RVs (c) Var(X) = p(1-p) for Bernoulli

In General Variance of sample mean $\bar{X}$ for RV with variance σ^2;

STDEV of sample mean decreases as $1/\sqrt{N}$

$$\mathbf{Var}(\bar{X}) = \frac{\sigma^2}{N}$$

Decreases linearly with number of samples N

Unbiasedness of the Sample Variance?

Recall: Sample mean is an unbiased estimator for the true mean.
How about the sample variance?

Ex. Let $X_1, \ldots, X_N$ be drawn (iid) from any distribution with $\mathbf{Var}(X) = \sigma^2$ and,

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i (X_i - \hat{\mu})^2$$

Then the sample variance is a **biased estimator**,

Source of bias:
plug-in mean estimate

$$\mathbf{E}[\hat{\sigma}^2] = \frac{1}{N} \sum_i \mathbf{E}\left[(X_i - \hat{\mu})^2\right] = \text{boring algebra} = \frac{N-1}{N}\sigma^2$$

tends to underestimate

Q: is this estimator consistent or not?

Consistent! (needs further justifications)

Correcting bias yields unbiased variance estimator:

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{N}{N-1}\hat{\sigma}^2 = \frac{1}{N-1} \sum_i (X_i - \hat{\mu})^2$$

quiz candidate:
show that $\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^N (X_i - \mu)^2\right]$ is unbiased
(note $\mu = E[X_1] = \cdots = E[X_N]$)

CSC380: Principles of Data Science

Statistics 3

Review from last lecture

- Maximum likelihood estimation

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \mathcal{L}_N(\theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

- Basic properties of sample mean, sample variance estimators
  - Sample mean -> population mean: unbiased?
  - Variance of sample mean: $\frac{\sigma^2}{N}$
  - Sample variance -> population variance: unbiased?

population variance
= true variance
population mean
= true mean

This lecture

- The bias-variance tradeoff of statistical estimation

- Interval estimation: confidence intervals
  - Cf. point estimation

Numpy Background

- Often, you have a matrix of data: e.g., movie review score

User \ Movie	Inception	Jurassic park	Batman
A	5	2	3
B	1	4	2
C	4	3	3
D	1	2	3

Numpy arrays can be 2d

```
A = np.array([[1,2,3],[4,5,6]])
A[0,1]
⟹ 2
mean(A,0)
⟹ array([2.5, 3.5, 4.5])
mean(A,1)
⟹ array([2., 5.])
```

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

access A[0,1] means 1st row, 2nd column

computes average for each column

computes average for each row

var(A,0), var(A,1) works the same way!

More on Unbiased Estimator

Task: Compare the **MSE** (mean squared error) of the two variance estimators for N=5.

```
import numpy as np
import numpy.random as ra
X = ra.randn(10_000,5)   # 10k by 5 matrix of N(0,1) => 10k random trials
```

```
np.mean((var(X,1,ddof=0) - 1)**2)
```

$\Longrightarrow$ 0.36310526687176103

X					
X[0,:]	0.35	1.45	-0.22	-2.95	-3.09
…..					
X[9999,:]	-1.78	-2.31	0.43	0.77	0.16

This estimates $\mathrm{E}\left[\left(\widehat{\sigma^2} - 1\right)^2\right]$, the MSE of estimator $\widehat{\sigma^2}$ with respect to the target $\sigma^2 = 1$

ddof=0 uses 1/N
ddof=1 uses 1/(N-1)

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (X_i - \hat{\mu})^2$$

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{N}{N-1}\hat{\sigma}^2 = \frac{1}{N-1}\sum_i (X_i - \hat{\mu})^2$$

More on Unbiased Estimator

np.mean((var(X,1,ddof=1) - 1)**2)

$\Longrightarrow$ 0.5071783438808787

Recall: np.mean((var(X,1,ddof=0) - 1)**2)

$\Longrightarrow$ 0.3631052668717610 3

- In this case, $\widehat{\sigma^2}$ (biased version) is more accurate than $\hat{\sigma}^2_{\mathrm{unbiased}}$! (but recall that it will underestimate)
- *There is a tradeoff between bias and variance!!*

Bias-Variance Tradeoff

Is an unbiased estimator "better" than a biased one? It depends…

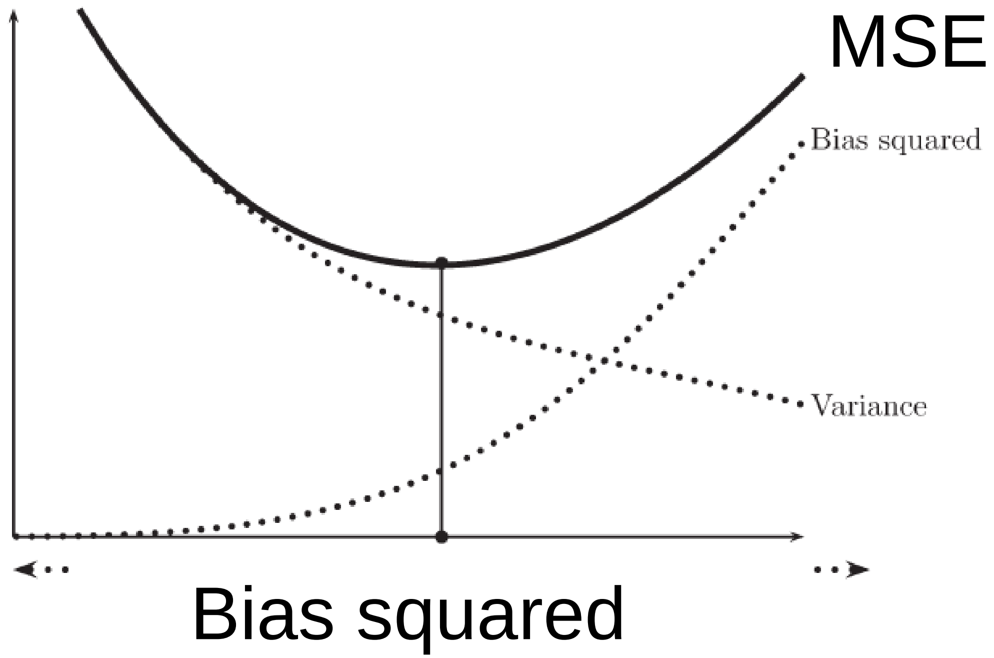
Evaluate the quality of estimate $\hat{\theta}$ using **mean squared error**,

$$\mathrm{MSE}(\hat{\theta}) = \mathbf{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{bias}^2(\hat{\theta}) + \mathbf{Var}(\hat{\theta})$$

- $\mathrm{bias}(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta$
- MSE for unbiased estimators is just,

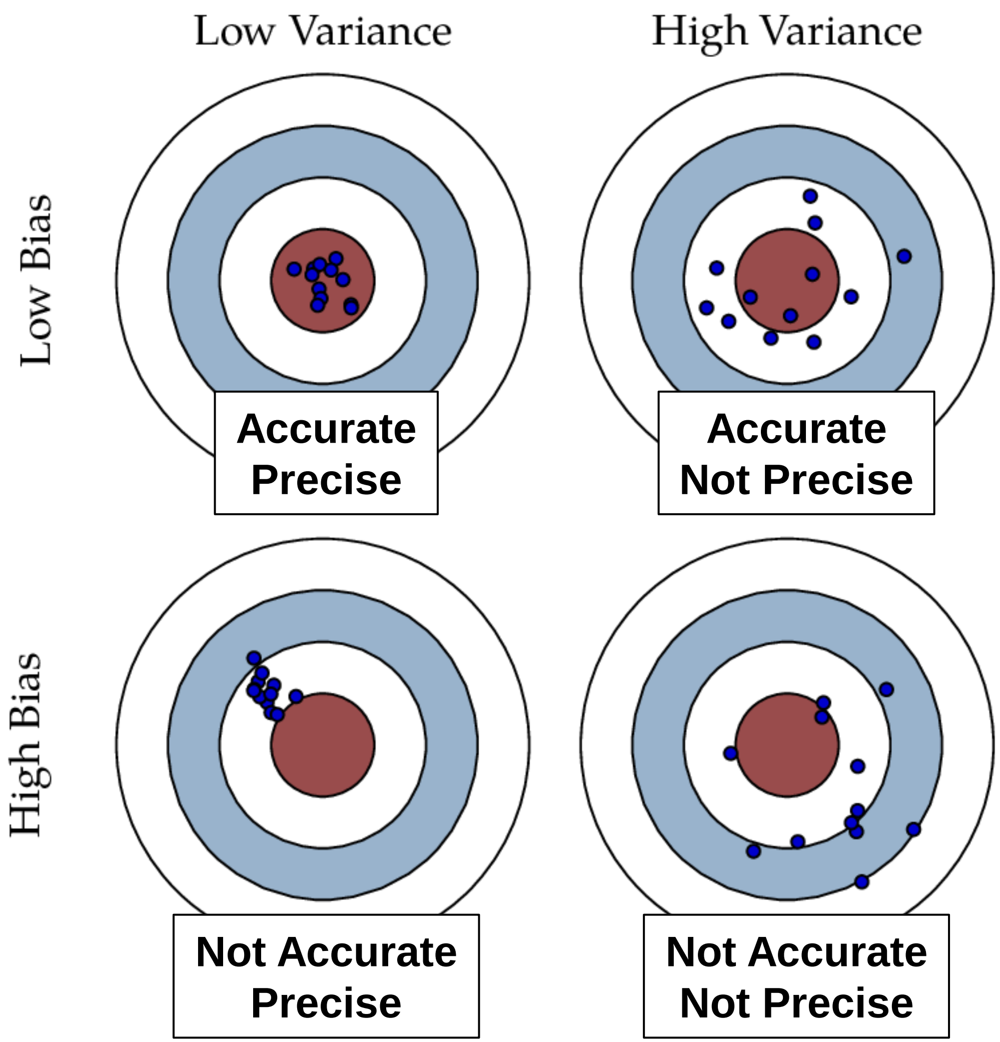
$$\mathrm{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta})$$

- Bias-variance is a fundamental tradeoff in statistical estimation
- MSE increases as **square** of bias
- Biased estimator can be more accurate than an unbiased one.



Bias-Variance Tradeoff

Suppose an archer takes multiple shots at a target…



- Target = θ
- Each shot = an estimate $\hat{\theta}$

- Bias $\approx$ systematic error
- Variance $\approx$ random error

Bias-Variance Decomposition

quiz candidate

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbf{E}\left[(\hat{\theta}(X) - \theta)^2\right] \\
&= \mathbf{E}\left[\left(\hat{\theta} - \mathbf{E}[\hat{\theta}] + \mathbf{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \mathbf{E}[(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2] + 2(\mathbf{E}[\hat{\theta}] - \theta)\mathbf{E}[\hat{\theta} - \mathbf{E}[\hat{\theta}]] + \mathbf{E}\left[\left(\mathbf{E}[\hat{\theta}] - \theta\right)^2\right] \\
&= \left(\mathbf{E}[\hat{\theta}] - \theta\right)^2 + \mathbf{E}[(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2] \\
&= \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})
\end{aligned}
$$

Intuition Check

Compare the results of two coin flip experiments…

Experiment 1 Flip 100 times and observe 73 heads, 27 tails

Experiment 2 Flip 1,000 times and observe 730 heads, 270 tails

Question The MLE estimate of coin bias for both experiments is equivalent $\hat{\theta} = 0.73$. Which should we trust more? Why?

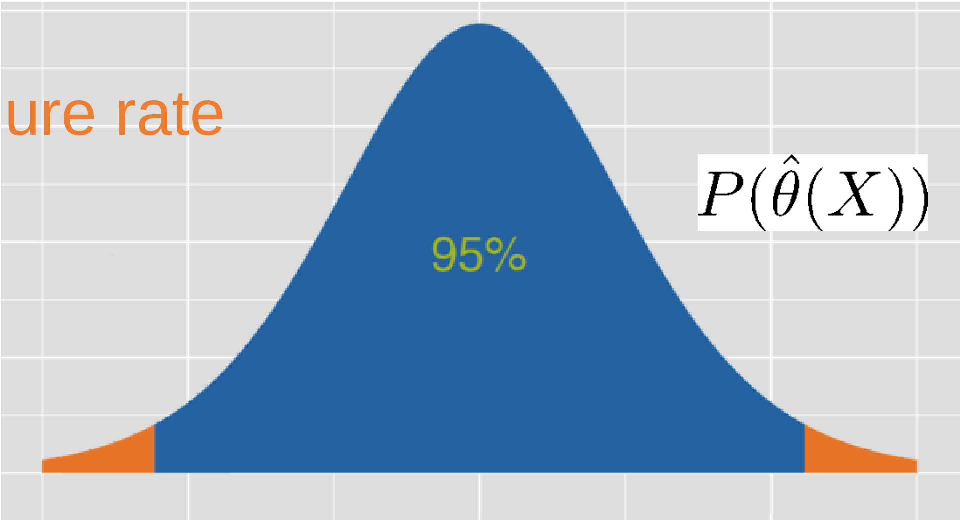
Remark The estimate $\hat{\theta}(\boldsymbol{X})$ is a function of random data. So, it is a random variable. It has a distribution.

Next lecture: confidence intervals

Confidence intervals

Confidence Intervals

Informally, find an interval such that we are *pretty sure* it encompasses the true parameter value.



significance level = failure rate

Given data $X_1, \ldots, X_n$ and ~~confidence~~ $\alpha \in (0,1)$ find interval (a,b) such that,

$$P(\theta \in (a,b)) \geq 1 - \alpha$$

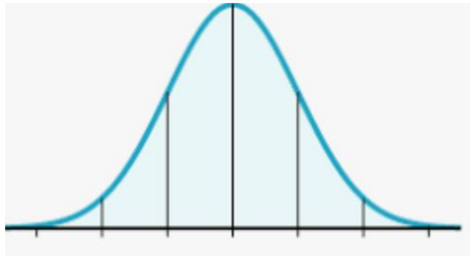
In English the interval (a,b) contains the true parameter value θ with probability **at least** $1-\alpha$ Valid confidence interval construction

- Intervals must be computed from data: i.e., $a(X_1, \ldots, X_n)$ and $b(X_1, \ldots, X_n)$
- Interval (a,b) is **random**, parameter θ is **not random** (it is fixed)
- Usually, you compute an estimator $\hat{\theta}$ and then set $a = \hat{\theta} - \epsilon_a$ and $b = \hat{\theta} + \epsilon_b$ for a carefully chosen $\epsilon_a, \epsilon_b > 0$

Recall: Central Limit Theorem

- Let $X_1, \ldots, X_N$ be iid with mean μ and variance σ^2. Then for large enough N, the sample mean $\bar{X}_N$ approaches a Normal distribution:

$$\bar{X}_N \approx \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

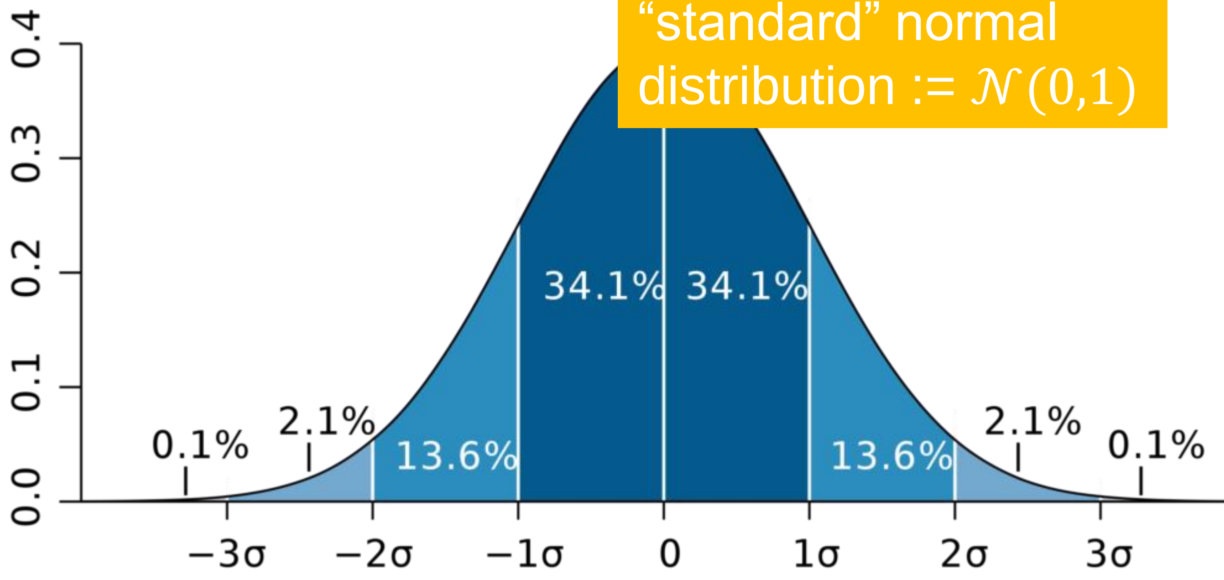


- '$\approx$' here means "approximately follows distribution"
- '$\approx$' is replaced with '$\sim$' if $X_i \sim \mathcal{N}(\mu, \sigma^2)$

Normal distribution: tail probability

Terminology: "standard" normal distribution := $\mathcal{N}(0,1)$

- $\Phi(z) \coloneqq \mathrm{P}(Z \le z)$ is the CDF of $Z \sim \mathcal{N}(0,1)$,
  Properties:
  - $\Phi(0) = 0.5$
  - $\Phi(-z) = 1 - \Phi(z)$
- e.g. $\Phi(-2) = 0.022$, $\Phi(2) = 0.978$



- **<u>(Fact 1)</u>** If $Z \sim \mathcal{N}(0,1)$,
  $$P(Z \in [-z, z]) = 1 - 2\Phi(-z) = 2\Phi(z) - 1$$
- e.g. $P(Z \in [-2,2]) = 1 - 0.022 \times 2 = 0.956$

Useful values:

- $z = 1.96 \Rightarrow$ RHS $\approx .95$, "95% confident"
- $z = 2.58 \Rightarrow$ RHS $\approx .99$

Set z such that $P(Z \in [-z, z]) = 1 - \alpha$:

```
import scipy.stats as st
alpha = 0.05
st.norm.ppf(1-alpha/2)
=> 1.959963984540054
```

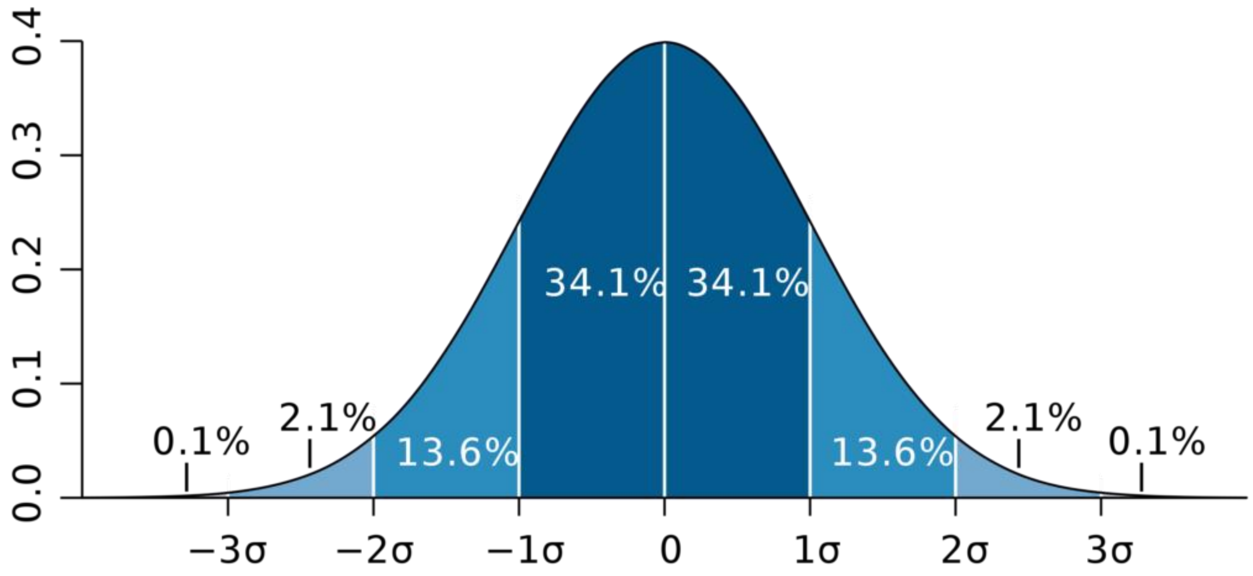
Confidence Intervals for Normal mean

Suppose $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2. Let $\hat{\mu} \coloneqq \frac{1}{n}\sum_i X_i$.

(Corollary)

$$P\left(\hat{\mu} \in \left[\mu - \frac{z\sigma}{\sqrt{n}}, \mu + \frac{z\sigma}{\sqrt{n}}\right]\right) = 2\Phi(z) - 1$$

hints: use the fact $\sqrt{n}\frac{\hat{\mu}-\mu}{\sigma} \sim N(0,1)$. Set $Z \coloneqq \sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}$ and use Fact 1.



Gaussians almost do not have tails!

remember: 'normal algebra' is very useful (and will appear in exams)

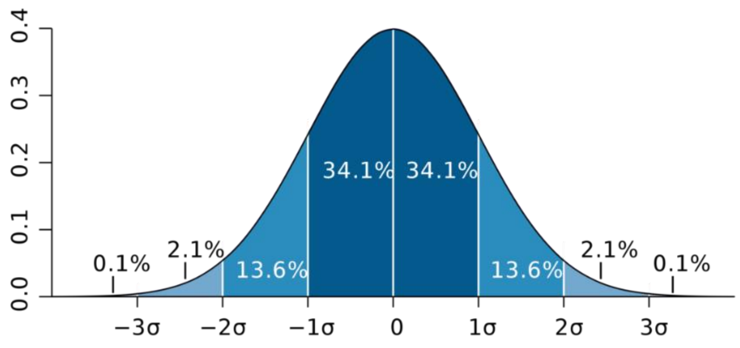
Confidence Intervals for Normal mean

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & known σ^2. Let $\hat{\mu} := \frac{1}{n}\sum_i X_i$.

Finally, by our corollary,

$$P\left(\hat{\mu} \in \left[\mu - \frac{1.96\sigma}{\sqrt{n}}, \mu + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \geq 0.95$$

$$P\left(\hat{\mu} \in \left[\mu - \frac{2.58\sigma}{\sqrt{n}}, \mu + \frac{2.58\sigma}{\sqrt{n}}\right]\right) \geq 0.99$$



note we can switch $\hat{\mu}$ and μ

$$P\left(\mu \in \left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]\right) \geq 0.95$$

This is a confidence bound for the mean μ !!

=> Compute $\left[\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}}\right]$. Done!

Q: If $X_1, \ldots, X_n$ from an arbitrary distribution, can we still use the same method?

Almost yes, if n is large enough! => central limit theorem (see later).

Caveat: interpreting confidence intervals

<u>Question</u> *How should we interpret a confidence interval (e.g. 95%)?*

$$P(\theta \in (a(X), b(X))) \geq 0.95$$

<u>Hint</u> *Think about what is random and what is not…*

This is NOT about the randomness of θ

Wrong If someone reveals θ <u>when we have exactly the same data</u>, then θ will be in the interval with probability at least 95%

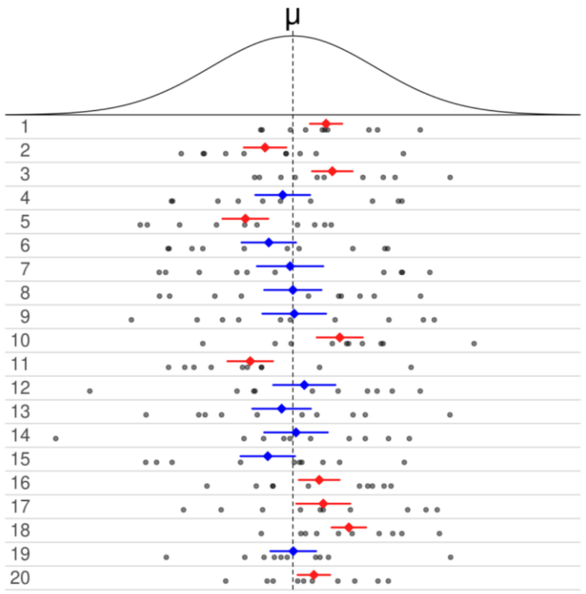
the moment you compute the interval with the data, whether or not θ is in the interval is determined.. you just don't know it!

This is commonly misinterpreted

Caveat: interpreting confidence intervals

Recommended point of view:

- Assume: Heights of UA students follow a normal distribution $\mathcal{N}(\mu, 1)$ with unknown μ
- Fork m **parallel universes**. For each universe $u \in \{1,2,\ldots,m\}$,
  - Subsample n UA students randomly, take the sample mean $\hat{\mu}^{(u)}$.
  - Compute the confidence bound $\left[\hat{\mu}^{(u)} - \frac{1.96\sigma}{\sqrt{n}}, \hat{\mu}^{(u)} + \frac{1.96\sigma}{\sqrt{n}}\right]$



- The fraction of parallel universes where the random interval includes μ is *approximately* at least 0.95 if m is large enough.
- As m goes to infinity, the fraction will become arbitrarily close to a value that is at least 0.95.

https://en.wikipedia.org/wiki/Confidence_interval

Confidence bounds for arbitrary distributions

Recall: *If* $X_1, \ldots, X_n$ *from an* ***arbitrary*** *distribution, can we still use the same method used for Gaussian to construct confidence intervals for the population mean?*
Short answer: YES, if n is large enough.

Plan
- Use central limit theorem
- 3 methods for arbitrary distributions

Confidence bounds for arbitrary distributions

But first, why do we care?

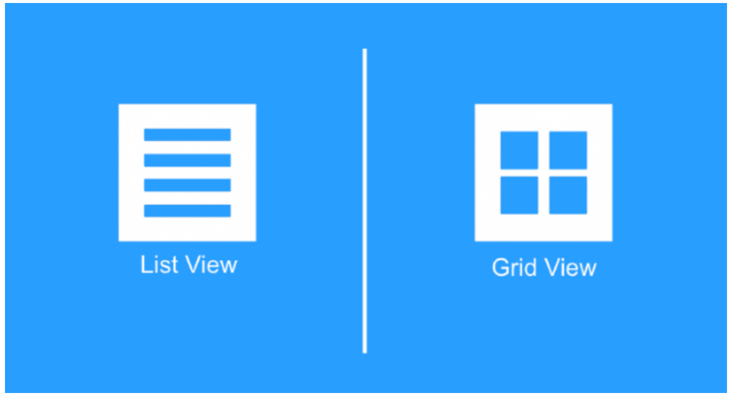
A/B testing: You are an engineer at amazon. You want to see if people buy more items if you change the search result from list view to grid view

You changed it to grid view for one day. Various metrics: click rate, purchase rate, …

You compute these, it seems to increase the click rate by 0.05%. You tell the boss about it.

Your boss: How do I know it is not a random fluctuation?

unfortunately, clicks are Bernoulli RVs, not gaussian!



Method 1: Gaussian (Naïve)

Suppose $X_1, \dots, X_n \sim \mathcal{D}$, i.i.d., but $\mathcal{D}$ is unknown. Let $\hat{\mu} \coloneqq \frac{1}{n}\sum_i X_i$.

- In light of CLT, we can pretend that $\mathcal{D} = \mathcal{N}(\mu, \sigma^2)$ with unknown μ, σ^2
- Q: Can we just use the following?

$$P\left(\hat{\mu} \in \left[\mu - \frac{z\sigma}{\sqrt{n}}, \mu + \frac{z\sigma}{\sqrt{n}}\right]\right) = 2\Phi(z) - 1$$

- No, we don't know the **population variance** σ^2.. What would you do?

- Solution: Replace σ with $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}$, the (unbiased) sample variance.
  - Q: Why not the 1/n version?

Avoid underestimating σ, so as to avoid false claims as much as possible!

Method 1: Gaussian (Naïve)

Suppose $X_1, \ldots, X_n \sim \mathcal{D}$, i.i.d., but $\mathcal{D}$ is unknown. Let $\hat{\mu} \coloneqq \frac{1}{n}\sum_i X_i$.

Summary: Let $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$

For 95% confidence:

$$\left[\hat{\mu} - \frac{1.96\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + \frac{1.96\hat{\sigma}}{\sqrt{n}}\right]$$

For 99% confidence:

$$\left[\hat{\mu} - \frac{2.58\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + \frac{2.58\hat{\sigma}}{\sqrt{n}}\right]$$

Method 1: Gaussian (Naïve)

Suppose $X_1, \ldots, X_n \sim \mathcal{D}$, i.i.d., but $\mathcal{D}$ is unknown. Let $\hat{\mu} := \frac{1}{n}\sum_i X_i$.

Summary: Let $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$

For 95% confidence:

$$\left[\hat{\mu} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

For 99% confidence:

$$\left[\hat{\mu} - 2.58\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + 2.58\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

Two assumptions:

- The data follows normal distribution.
- Sample variance is equal to the actual variance.

turns out, fixable => method 2

Review questions

all quiz candidates

- For $\hat{\sigma}^2$, why do we use the **unbiased** estimator rather than the sometimes more accurate **biased** estimator?

- What does CLT imply about the convergence rate of the sample mean $\bar{X}_n$ to the population mean μ?

- List the pros and cons of the biased variance estimator (1/n) vs unbiased variance estimator (1/(n-1)).

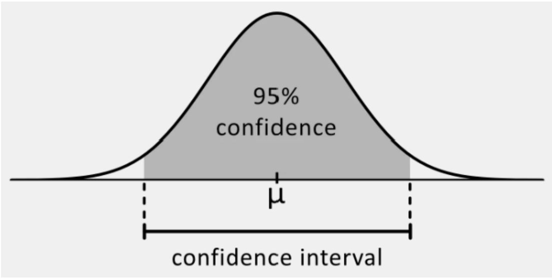
Computer Science

CSC380: Principles of Data Science

Statistics 4

Review from last lecture

- Confidence intervals
  - Given data $X_1, \ldots, X_n \sim \mathcal{D}_\theta$ with unknown θ (say, $\mathcal{D}_\theta = \mathcal{N}(\theta, 1)$)
  - α: significance level (typical values: 0.05, 0.01)
  - Construct a_n, b_n (that depends on $X_1, \ldots, X_n$), such that
$$P(\theta \in [a_n, b_n]) \geq 1 - \alpha$$



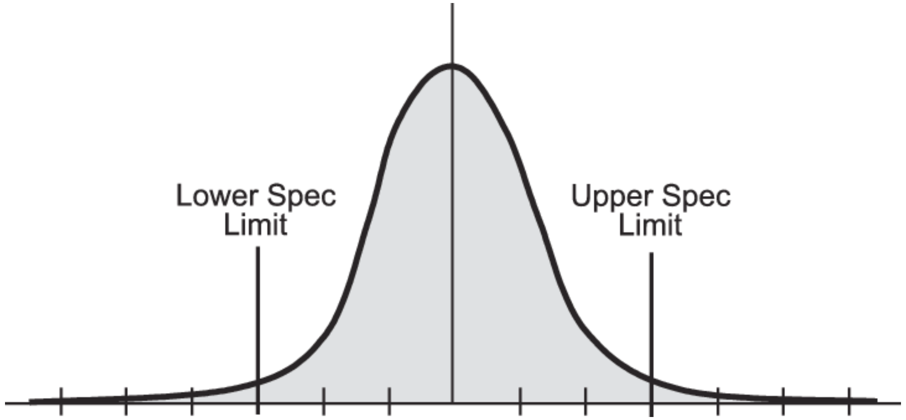
- Intuition: more samples $n \Rightarrow$ narrower confidence intervals
  - More confident about the location of θ

Review from last lecture

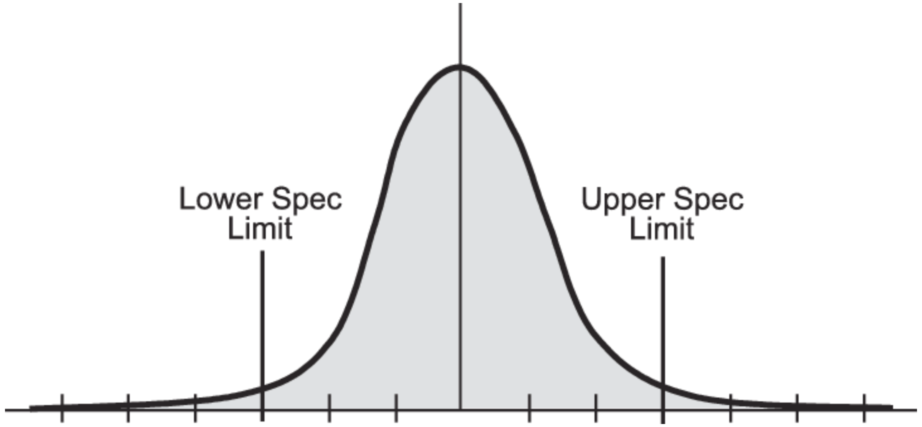
A recipe for constructing confidence intervals for θ

- Step 1: construct point estimate $\hat{\theta}_n$ (that serves as the interval's "center")
  - E.g. sample mean for estimating population mean θ

- Step 2: identify ν, the (approximate) distribution of (some function of) $\hat{\theta}_n - \theta$ – denote by Z
  - E.g. $Z = \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma} \approx \mathcal{N}(0,1)$



Review from last lecture



- Step 3: find the bottom $\alpha/2$ and top $\alpha/2$ quantile of ν, denoted by $\nu_{\alpha/2}$ and $\nu_{1-\alpha/2}$
  - $P_{Z\sim\nu}(Z \in [\nu_{\alpha/2}, \nu_{1-\alpha/2}]) \geq 1-\alpha$
  - E.g. $Z \sim \mathcal{N}(0,1), \alpha = 0.05 \Rightarrow \nu_{\alpha/2} = -1.96, \nu_{1-\alpha/2} = 1.96$

- Step 4: solve for $Z \in [\nu_{\alpha/2}, \nu_{1-\alpha/2}]$ to get a confidence interval of θ
  - E.g. $\frac{\sqrt{n}(\widehat{\theta}_n-\theta)}{\sigma} \in [-1.96, 1.96] \Leftrightarrow \theta \in \left[\hat{\theta}_n - \frac{1.96\sigma}{\sqrt{n}}, \hat{\theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right]$

Today

- More on confidence intervals
- Hypothesis testing

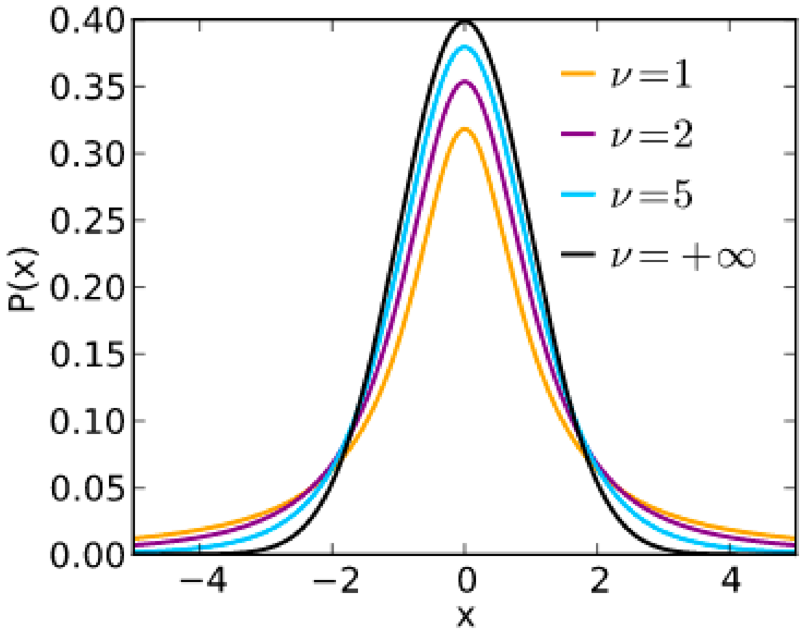
Method 2: Gaussian (Corrected)

Recall: Gaussian confidence interval with $\sqrt{n}\frac{\hat{\mu}_n-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

What if we use $\hat{\sigma}^2$ instead of σ?

(Theorem) $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown μ & σ^2. Let $\widehat{\mathrm{UVar}}_n \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$ (unbiased version of sample variance). Then,

$$\sqrt{n}\frac{\hat{\mu}_n-\mu}{\sqrt{\widehat{\mathrm{UVar}}_n}} \sim \text{student-t (degrees of freedom (DOF)} = n-1)$$



Note: ν = DOF for this slide only

as $\nu \to \infty$, it becomes Gaussian.

Method 2: Gaussian (Corrected)

With a similar derivation we have done before,
With at least 95% confidence:

$$\left[\hat{\mu} + t_{\alpha/2,n-1}\frac{\sqrt{\widehat{\mathrm{UVar}}_{\mathrm{n}}}}{\sqrt{n}}, \hat{\mu} + t_{1-\alpha/2,n-1}\frac{\sqrt{\widehat{\mathrm{UVar}}_{\mathrm{n}}}}{\sqrt{n}}\right]$$

Where $t_{\alpha/2,n-1}$ can be computed numerically.

<u>Key take away</u>: more conservative!
=> more likely to be valid (contain μ).

Method 2 is strongly preferred over Method 1

<u>Common practice</u>*: Apply this method even if we do not know whether true distribution is Gaussian.*

(recall: 1.96 for gaussian)

much larger number compensates for the inaccuracy of $\hat{\sigma}^2$

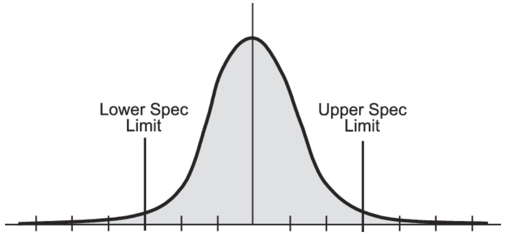
```
import scipy.stats as st
alpha = 0.05
st.t.ppf(1-alpha/2,df=2)
=> 4.302652729911275

st.t.ppf(1-alpha/2,df=5)
=> 2.5705818366147395

st.t.ppf(1-alpha/2,df=10)
=> 2.2281388519649385

st.t.ppf(1-alpha/2,df=30)
=> 2.0422724563012373

st.t.ppf(1-alpha/2,df=100)
=> 1.9839715184496334
```



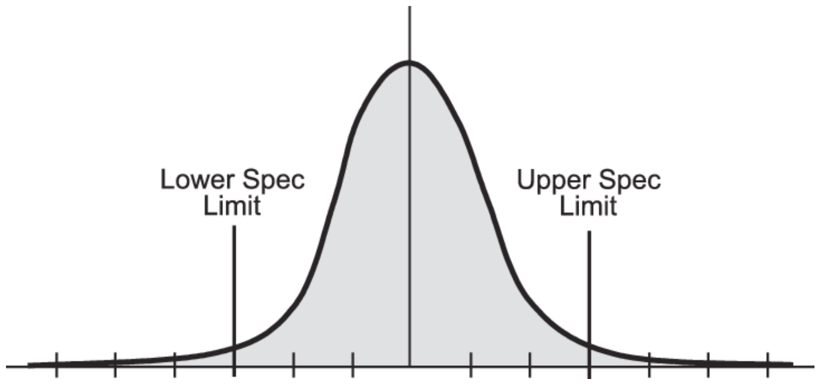
but often the data is highly nongaussian (e.g., skewed), which leads to invalidity of the confidence interval

Method 3: Bootstrap

- Suppose we observe data $S = (X_1, \ldots, X_n) \sim P(x;\theta)$
- We have some *arbitrary* estimator $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$

- Previously we know ν, the distribution of $\hat{\theta}_n - \theta$, and use that to obtain confidence intervals for θ

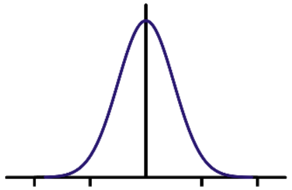
- However, the ν distribution is oftentimes unknown!
  - Recall HW3, P4(d)
  - $\hat{\rho}_n$ has a complicated formula, and has a complicated CDF..



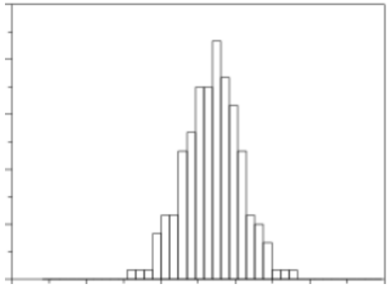
Method 3: Bootstrap

- Key idea: *approximate* ν, the distribution of $\hat{\theta}_n - \theta$
- Insight:

θ

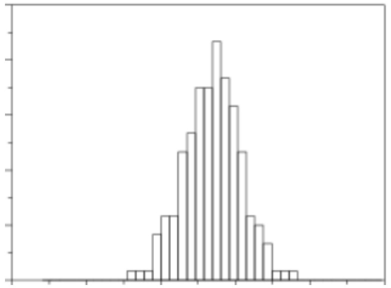


$$D_\theta \xrightarrow{n \text{ iid samples}} S$$

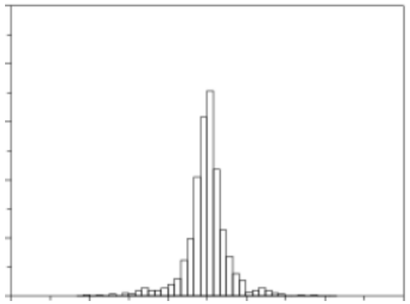


$\hat{\theta}_n$

$\hat{\theta}_n$



$$\text{Uniform}(S) \xrightarrow{n \text{ iid samples}} S_b$$



$\hat{\theta}_{n,b}$

- Use empirical distribution of $\hat{\theta}_{n,b} - \hat{\theta}_n$'s to approximate ν, obtaining approximations of $\nu_{\alpha/2}$ and $\nu_{1-\alpha/2}$
- This empirical distribution can be obtained by drawing multiple S_b's (bootstrap subsample)

Empirical Distribution (≈histogram)

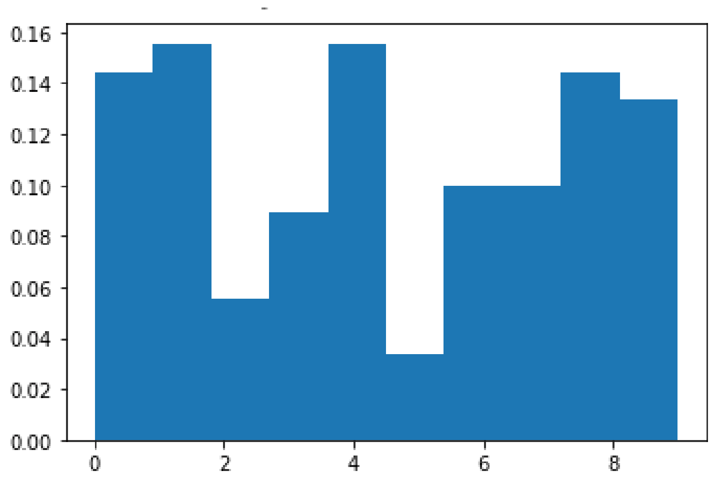
- Suppose we have $S = (X_1, \dots, X_n)$, i.i.d. sample from a distribution $\mathcal{D}$
- The empirical distribution is $\mathrm{Uniform}(S)$, (often denoted by $\widehat{\mathcal{D}}$)
- If $Y \sim \widehat{\mathcal{D}}$,

$$\forall v \in S, \qquad P(Y = v) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i = v\}$$

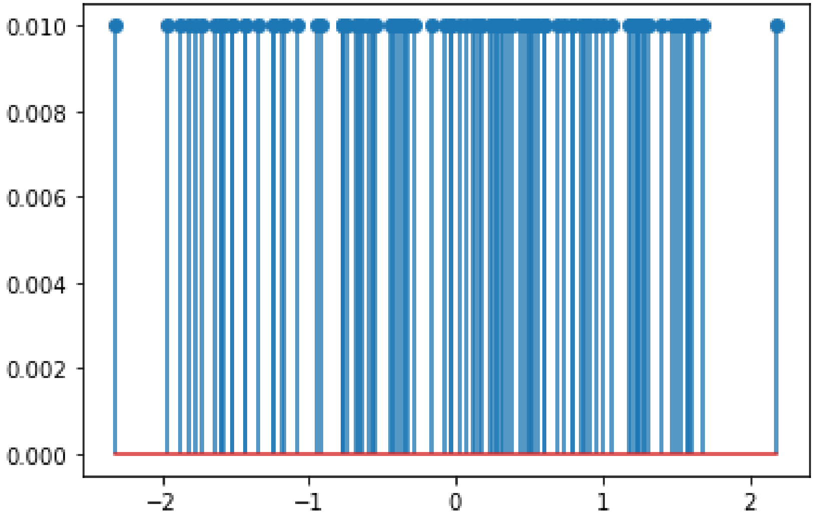
(PMF)

Empirical Distribution

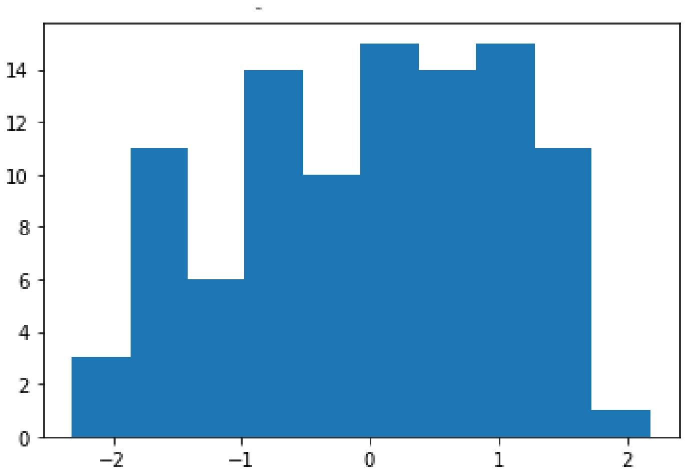

```
import numpy.random as ra
X = ra.randint(10, size=100)
plt.hist(X)
```




```
X = ra.randn(100)
plt.stem(X, 1/100*np.ones(X.shape[0]))
```




```
X = ra.randn(100)
plt.hist(X, 10)
```



not exactly what we call “empirical distribution”

Method 3: Bootstrap

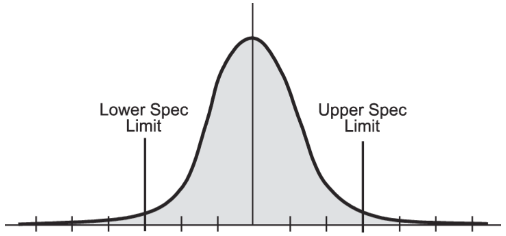
Suppose we observe data $X_1, X_2, \ldots, X_n \sim P(X;\theta)$:

1. Sample new “dataset” $X_1^*, \ldots, X_n^*$ uniformly from $X_1, \ldots, X_n$ **with replacement**

2. Compute estimate $\hat{\theta}_{n,1}$ based on $(X_1^*, \ldots, X_n^*)$

3. Repeat B times to get the estimators $\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,B}$

4. Consider the **empirical distribution** of $\{\hat{\theta}_{n,b} - \hat{\theta}_n\}_{b=1}^{B}$ and find its top $\frac{\alpha}{2}$ quantile and bottom $\frac{\alpha}{2}$ quantile (denoted by Q_U and Q_L respectively).



5. (1-α) Confidence Interval: $[\hat{\theta}_n - Q_U, \hat{\theta}_n - Q_L]$

counterintuitively, upper quantile for lower width, lower quantile for upper width. Why?

$$P(\nu_{\alpha/2} \leq \hat{\theta}_n - \theta \leq \nu_{1-\alpha/2}) \geq 1 - \alpha$$

note: there are other variations, but this version is recommended by statisticians.

Method 3: Bootstrap

<u>Pseudocode:</u> estimating population mean θ

Input: $X_1, \dots, X_n, B, \alpha$

- Compute $\bar{X}_n (= \hat{\theta}_n)$
- Bootstrapping B times to obtain $\{\hat{\theta}_{n,b} - \bar{X}_n\}_{b=1}^{B}$; call this array S
- Sorted S in increasing order.
- $Q_U :=$ the top $\frac{\alpha}{2}$ quantile; i.e., `S[int(np.ceil( (1-alpha/2)*(B-1) ))]`
- $Q_L :=$ the bottom $\frac{\alpha}{2}$ quantile; i.e., `S[int(np.floor((  alpha/2)*(B-1) ))]`
- Return $[\bar{X}_n - Q_U, \bar{X}_n - Q_L]$

Example: Generate 300 samples from $\mathrm{Bernoulli}(0.03)$

```
ary = ra.rand(300) < 0.03
muhat = np.mean(ary)

LB, UB = calc_ci_bootstrap(ary, 0.05, 10_000)
(LB,UB)
```

(0.006, 0.046)

```
# compute lower/upper width
(muhat-UB, muhat-LB)
```

(-0.016, 0.036)

asymmetric!! (note muhat=0.03..)

```
w = calc_confwidth_gaussian(ary, alpha)
(muhat -w, muhat + w)
```

(-0.107, 0.167)

inherently symmetric.. much looser

Confidence Intervals Comparison

good = correct
bad = incorrect

	Gaussian (corrected)	Bootstrap
small n	Bad	Bad
moderate n	Okay / Bad	Okay
large n	Good	Very Good
Computational complexity	Low	High, depends on B

bad if the estimator takes a long time to compute

Q: When could it be bad?

When the distribution is far from Gaussian

Hypothesis testing

A/B Testing

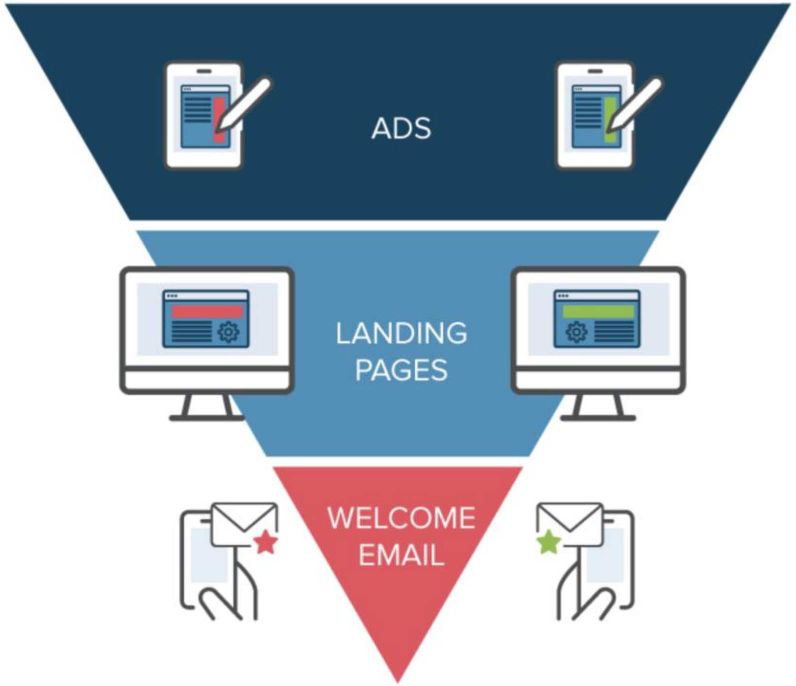
These days, webpages/apps run A/B testing extensively

Try out an alternative of webpage/interface on **randomly chosen subset of users** to gather data (and help guide the decision)

- E.g., recall the **list view** vs grid view to measure the effectiveness
- Can use **2 or more** alternatives

Review question: How do we know one is actually better than the other? (i.e., statistically significant)

A/B testing: Compute confidence intervals for alternatives, see if they overlap. If not, you have a clear winner!



(from optimizely.com)

A/B Testing

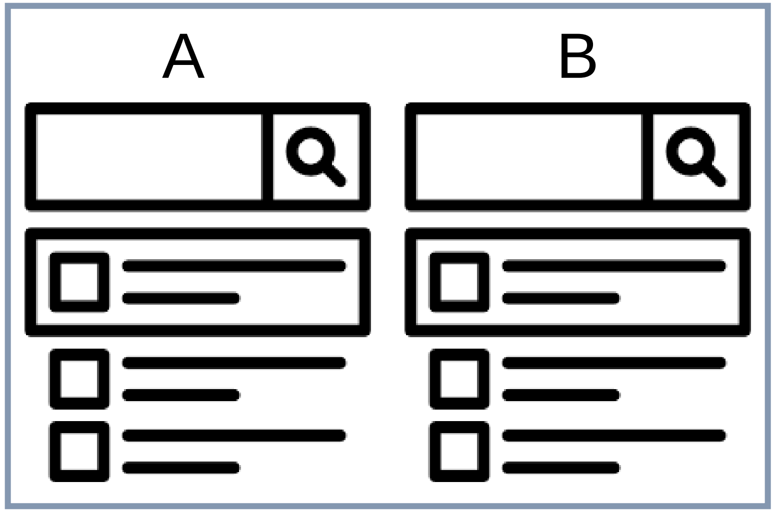
Another example: Search system evaluation (e.g., Google). Compare system A vs B.

- Each evaluator
  => a random keyword is picked, and then both systems pick top 10 relevant documents and show them.
  => the evaluator provides rating (1-5) for both lists.

Evaluator:	1	2	3	4	5	6	…
System A	5	2	2	5	4	2	…
System B	4	1	1	4	3	1	…

Methods to claim which is better.

- 1. The average — bad
- 2. Confidence intervals — okay
- 3. Paired t-test — great



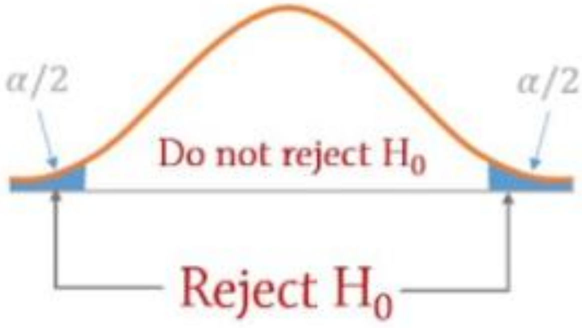
Two-sample hypothesis testing: general setup

- Samples $S = (S_A, S_B)$, $S_A = (A_1, \ldots, A_n), S_B = (B_1, \ldots, B_n)$ drawn iid from distributions with means θ_A and θ_B, respectively

- Two hypotheses:
  - H_0: $\theta_A = \theta_B$ -- the system A and B have the same performance ("null" hypothesis)
  - H_1: $\theta_A \neq \theta_B$ -- A and B have different performance

- Hypothesis test R: maps S to $\{0,1\}$
- Goal: minimize type-II error $P_{H_1}(R(S) = 0)$

  s.t. type-I error $P_{H_0}(R(S) = 1) \leq \alpha$:= significance level

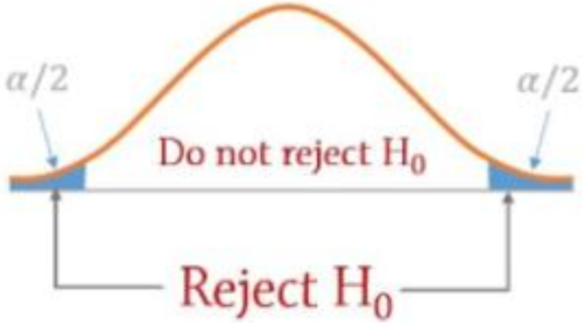


Paired t-test: intuition

- a **trial**: one evaluator's score on systems A and B
- n trials = n evaluators' score
- Let δ_i = score of (A) – score of (B) on the data point i (or, evaluator i)

- Intuition: T should output 1, if $\bar{\delta}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_i$ is large

- But how to ensure that $P_{H_0}(R(S) = 1) \leq \alpha$?



Paired t-test

We do not know what distribution δ_i follows. => again, assume normality

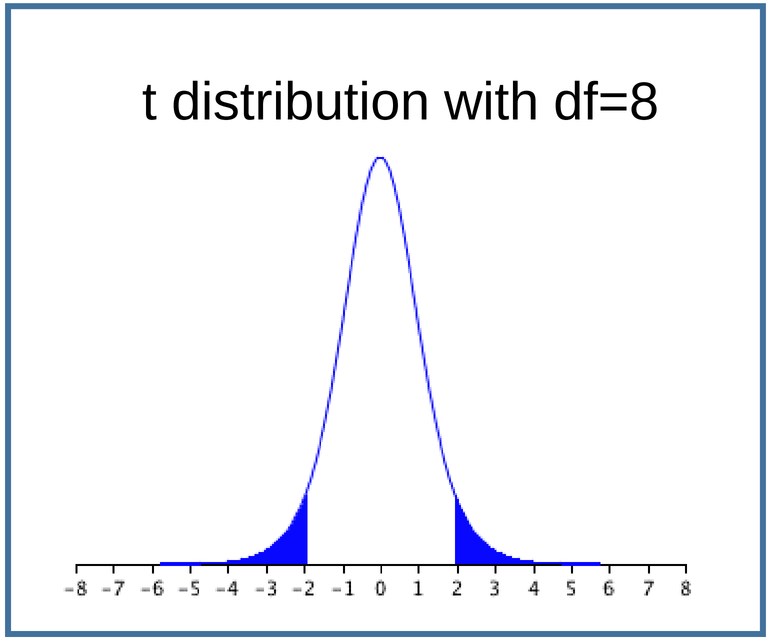
- H_0 assumes $\delta_i \sim N(0, \sigma^2), i = 1, \dots, n.$

Recall: Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, and $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$, $\widehat{UVar}_n := \frac{1}{n-1}\sum_{i=1}^{n}(\delta_i - \bar{\delta}_n)^2$

$$\left(T_n := \sqrt{n}\frac{\hat{\mu}_n - \mu}{\sqrt{\widehat{UVar}_n}}\right) \sim \text{student-t (degrees of freedom} = n - 1)$$

call it t-score

In our case, $\mu = 0$



Ask "under H_0, what is a plausible range of values that T_n lies in with probability $\geq 1 - \alpha = 0.95$?"

- Find the *quantiles* of student-t distribution!
  - $t_{\alpha/2, n-1}$ and $t_{1-\alpha/2, n-1}$
- Let $X_i \leftarrow \delta_i$ & $\mu = 0$ and see if T_n crosses the quantiles!

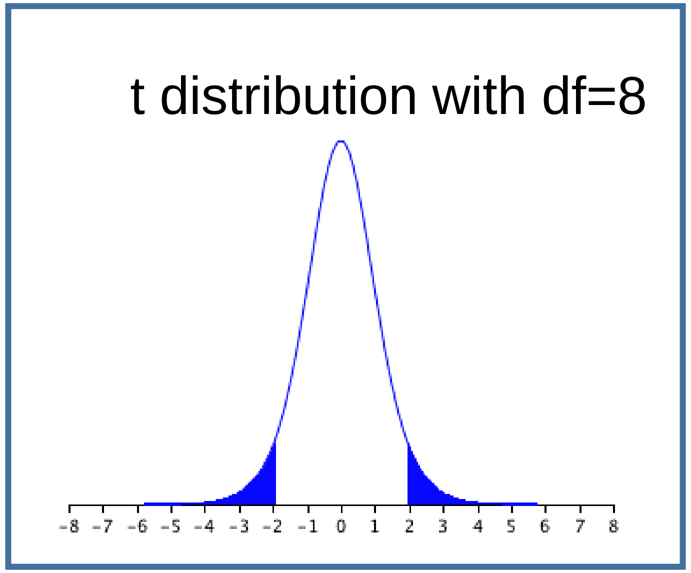
Paired t-test

Let $X_i \leftarrow \delta_i$ and see if T_n crosses the quantiles!

- Yes: Reject the null hypothesis H_0. => claim: the differences are real
- No: Accept the null hypothesis H_0. => claim: no statistically significant difference

p-value:

- a family of hypothesis tests R_α's with different significance α's
- p-value: the smallest α with which you can still reject H_0
  - Compute: Look at the CDF of t distribution, say $F(x)$, and compute $2(1 - F(|T_n|))$
- If below 0.05, reject the null hypothesis; Smaller the better.



"probability that your claim is false"

Instead of saying "our new system passed the paired t-test" people often just report the p-value. Smaller the better.

Paired t-test


```
import numpy as np
import scipy.stats as st

data = [[5,3],
        [3,1],
        [3,1],
        [5,3],
        [4,2],
        [2,1]]
data = np.array(data)

n = data.shape[0]
alpha = 0.05

def calc_width(ary, alpha):
    … (omitted)
```

Method 2: Gaussian (corrected)

```
for i in range(2):
    muhat = data[:,i].mean()
    width = calc_width(data[:,i],alpha)
    print([muhat-width,muhat+width])

diff = data[:,0] - data[:,1]
uvar = diff.var(ddof=1)
tscore = np.mean(diff)/uvar*np.sqrt(n)
threshold = st.t.ppf(1-alpha/2,n-1)  # ‘quantile’

# want tscore to be outside [-threshold,threshold]
print('tscore = %f' % tscore)
print('threshold = %f' % threshold)
```

```
output:
[2.3957369914894784, 4.9375963418438553]
[0.5624036581561451, 3.1042630085105216]
tscore = 3.061862
threshold = 2.570582
```

Confidence interval method claims no significant difference; paired t-test claims significant difference

Paired t-test

- Summary
  - More powerful than confidence bounds!
  - Reason: It focuses on the **difference**, which is what we actually care!

- Other methods
  - Permutation tests: highly recommended for A/B testing when the outcomes are not paired. (A/B testing in the content layout – no guarantee that the same user/evaluator will see both A and B)

Classical Statistics Review

- **Statistical Estimation** infers unknown parameters θ of a distribution $p(X;\theta)$ from observed data $X_1,\ldots,X_n$
- An estimator is a function of the data $\hat{\theta}(X_1,\ldots,X_n)$, it is a **random variable**, so it has a distribution
- **Confidence Intervals** measure uncertainty of an estimator, e.g.

$$P(\theta \in (a(X), b(X))) \geq 0.95$$

- **Bootstrap** A simple method for constructing confidence intervals

↑ Q: when is this good?

Caution

- Confidence intervals are often misinterpreted!
- Confidence intervals in practice may not be valid for small n

Classical Statistics Review

- **Estimator bias** describes systematic error of an estimator

- **Mean squared error (MSE)** measures estimator quality / efficiency,

$$\mathrm{MSE}(\hat{\theta}) = \mathbf{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{bias}^2(\hat{\theta}) + \mathbf{Var}(\hat{\theta})$$

- **Law of Large Numbers (LLN)** guarantees that sample mean approaches (piles up near) true mean in the limit of infinite data

- **Central Limit Theorem (CLT)** says sample mean approaches a Normal distribution with enough data. Also means $\frac{1}{\sqrt{n}}$ convergence.

- **LLN** and **CLT** are *asymptotic statements* and do not hold for small/medium data in general