CSC380: Principles of Data Science

Classical Statistics and Estimation

Prof. Jason Pacheco

TA: Enfa Rose George

TA: Saiful Islam Salim

Probability and Statistics

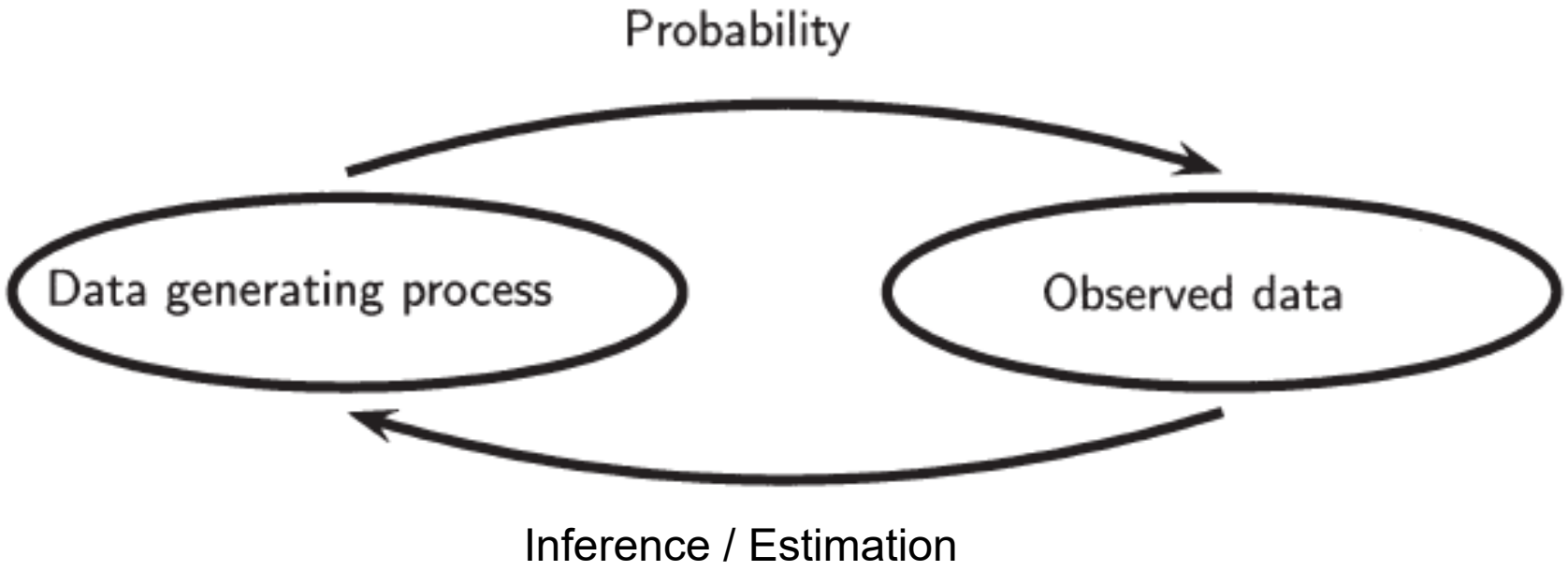
- Probability provides a mathematical formalism to reason about randomness

- Statistics deals with data and encompasses
  - Data collection / organization
  - Interpretation of data
  - Answering questions from data (statistical inference, hypothesis testing)
  - Fitting models to data (estimation)

- Statistics *uses* probability to address these tasks

Probability and Statistics

Probability describes how to generate data



Statistics describes how data were generated

[Source: Wasserman, L. 2004]

- Parameter Estimation
  - Method of Moments
  - Maximum Likelihood Estimation

- Confidence Intervals
  - Overview
  - Bootstrap confidence intervals

- Estimator Properties
  - Estimator Bias / Mean Squared Error
  - Law of Large Numbers / Central Limit Theorem

- Parameter Estimation
  - Method of Moments
  - Maximum Likelihood Estimation

- Confidence Intervals
  - Overview
  - Bootstrap confidence intervals

- Estimator Properties
  - Estimator Bias / Mean Squared Error
  - Law of Large Numbers / Central Limit Theorem

Intuition Check

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

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

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

Question 3 How might we estimate the bias of the coin with 73 heads and 27 tails?

Estimating Coin Bias

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

$$X \sim \text{Bernoulli}(\pi) = \pi^X (1-\pi)^{1-X} \quad \text{where} \quad X \in \{0, 1\}$$

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,

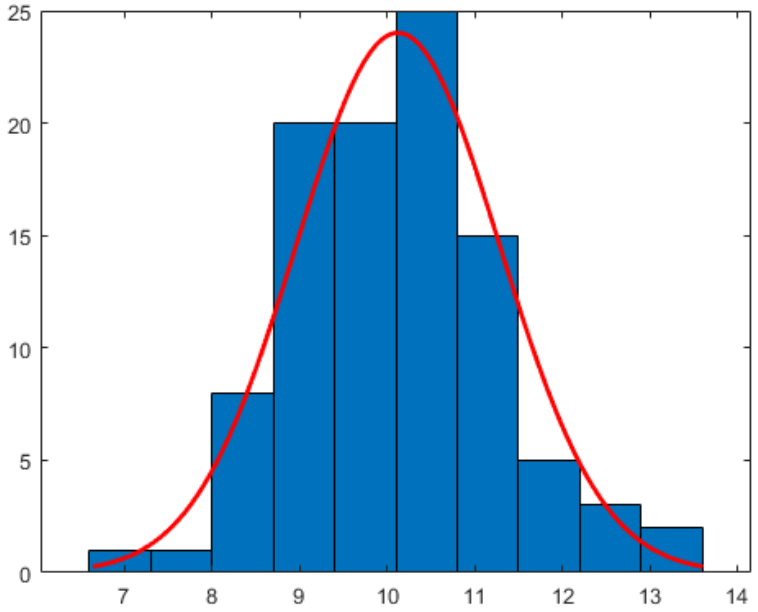
$$\hat{\pi} = \frac{1}{N} \sum_{n=1}^{N} x_n \approx \mathbf{E}[X] = \pi$$

This is the empirical mean or sample mean

Estimating Gaussian Parameters

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

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



How can we estimate the **mean**?

$$\hat{\mu} = \frac{1}{N}\sum_i x_i \approx \mu$$

Sample mean $\bar{x}$

How can we estimate the **variance?**

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

Variance estimator uses our previous mean estimate. This is a **plug-in estimator.**

Parameter Estimation

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

$$p(X;\theta)$$

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

$$\{x_i\}_i^N \overset{iid}{\sim} p(\cdot;\theta)$$

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

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

Many different types of estimators, each with different properties

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
- Sample STDEV s
- 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.*

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** 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

- **Efficiency** It should have low error with the least data, e.g.

$$\mathrm{MSE}(\hat{\theta}) = \mathbf{E}[(\hat{\theta} - \theta)^2]$$

Mean squared error should be small

Method of Moments

A simple way to estimate parameters…

Suppose we have K parameters $\theta = (\theta_1, \ldots, \theta_K)$ with j[th] **moment**,

$$\alpha_j(\theta) = \mathbf{E}_\theta[X^j]$$

and the jth **sample moment**,

$$\hat{\alpha}_j(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

…match moments to sample moments

Method of Moments

Defines a system of K equations and K unknowns

9.3 Definition. *The* **method of moments estimator** $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned} \tag{9.4}$$

[Source: Wasserman, L. 2004]

MoM Example: Estimating Coin Bias

Remember how we estimated coin bias…

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

$$X \sim \text{Bernoulli}(\pi) = \pi^X (1-\pi)^{1-X} \quad \text{where} \quad X \in \{0, 1\}$$

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,

$$\hat{\pi} = \frac{1}{N} \sum_{n=1}^{N} x_n \approx \mathbf{E}[X] = \pi$$

… this is method of moments with a change of notation

MoM Example: Estimating Coin Bias

Remember how we estimated coin bias…

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

$$X \sim \text{Bernoulli}(\theta) = \theta^X (1-\theta)^{1-X} \qquad \text{where} \qquad X \in \{0, 1\}$$

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

$$\alpha_1(\theta) = \mathbf{E}_\theta[X] = \theta$$

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

$$\hat{\alpha}_1 = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\alpha_1(\hat{\theta}) = \hat{\theta} = \hat{\alpha}_1$$

… this is method of moments with a change of notation

MoM Example: Estimating Normal Parameters

9.5 Example. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then, $\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2$. We need to solve the equations[1]

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

This is a system of 2 equations with 2 unknowns. The solution is

$$\begin{aligned} \widehat{\mu} &= \overline{X}_n \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \end{aligned}$$

[Source: Wasserman, L. 2004]

Intuition Check

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

Question 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

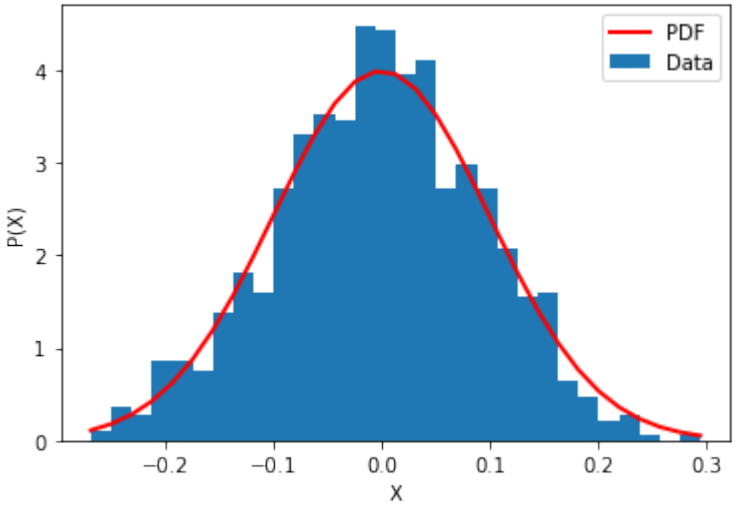
B: $\hat{\theta} = 0.50$, fair coin (we observed unlucky outcomes)

Likelihood (informally) Probability of the observed outcomes from model with parameters $\hat{\theta}$

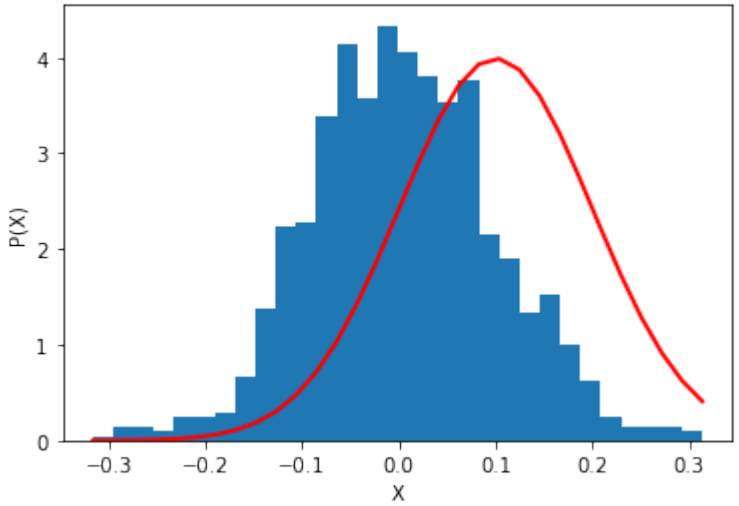
Likelihood (Intuitively)

Suppose we observe N data points from a Gaussian model and wish to estimate model parameters…

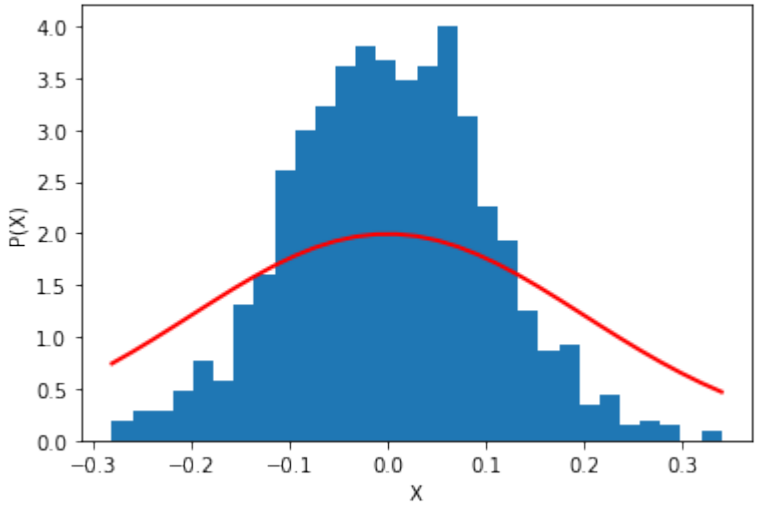
High Likelihood



Low Likelihood (mean)



Low Likelihood (variance)



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

Likelihood Function

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

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

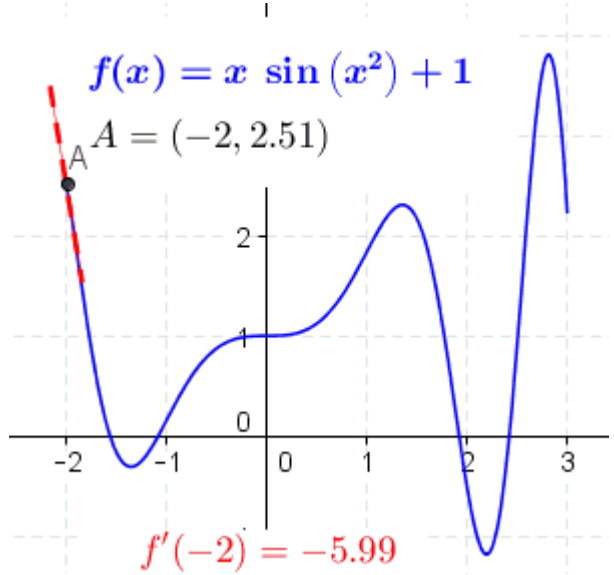
- We call this the **likelihood function**, often denoted $\mathcal{L}_N(\theta)$
- It is a function of the parameter θ, the data are fixed
- Describes how well parameter θ describes data (*goodness of fit*)

How could we use this to estimate a parameter θ?

Maximum Likelihood

Maximum Likelihood Estimator (MLE) as the name suggests, maximizes the likelihood function.

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



Question How do we find the MLE?

Answer Remember calculus…

Is $\mathcal{L}_N(\theta)$ convex?

Yes → Unique, closed-form solution

No → Gradient-based optimization

Approach

- Compute derivative $\frac{d\mathcal{L}_N}{d\theta}$
- Set to zero and solve

$$\frac{d\mathcal{L}_N}{d\theta} = 0 \Rightarrow \hat{\theta}^{\text{MLE}}$$

Still have to compute derivative…

Maximum Likelihood

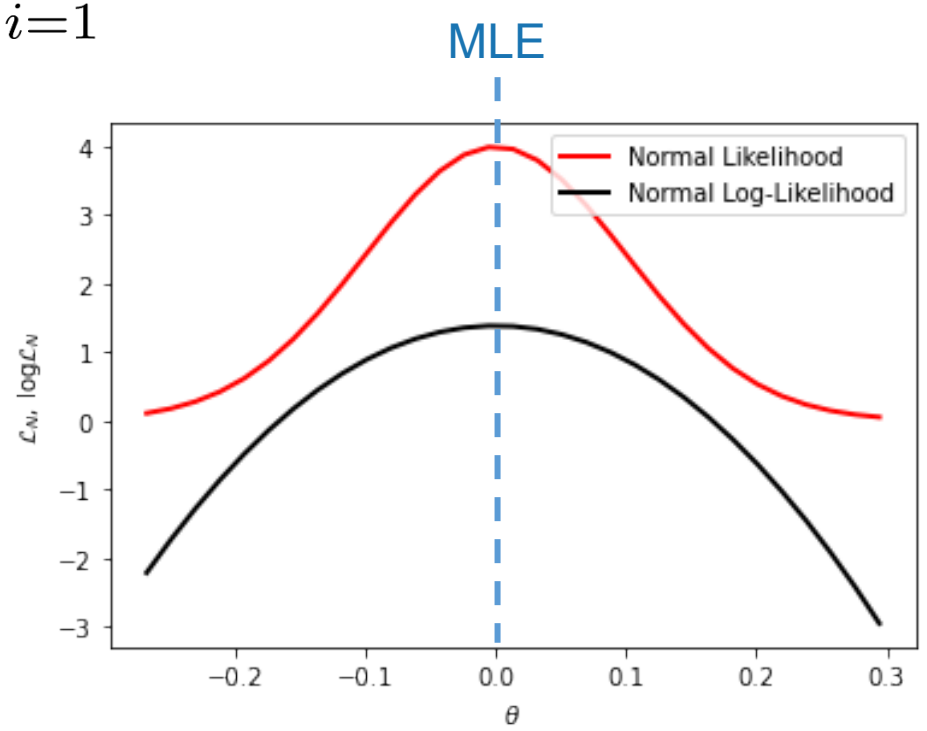
Maximizing log-likelihood makes the math easier (as we will see) and doesn't change the answer (logarithm is an increasing function)

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

Derivative is a linear operator so,

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

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

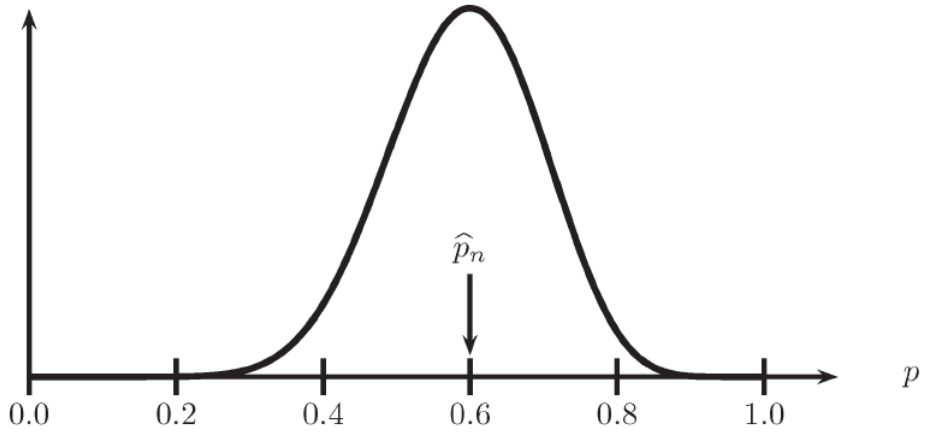


Maximum Likelihood

[Source: Wasserman, L. 2004]

Example Suppose we have N coin tosses with $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ but 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

Maximum Likelihood

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

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

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

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 / STDEV:

$$\hat{\mu}^{\text{MLE}} = \bar{X} = \frac{1}{n} \sum_i X_i \qquad \hat{\sigma}^{\text{MLE}} = S = \sqrt{\frac{1}{n} \sum_i (X_i - \bar{X})^2}$$

Maximum Likelihood Properties

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 **with high probability**.

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

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

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

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?

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

Administrative Items

- HW2 Due tonight @ 11:59pm
- HW3 Out first thing tomorrow
- Lecture title slides had wrong class number (oops!)

- Parameter Estimation
  - Method of Moments
  - Maximum Likelihood Estimation

- Confidence Intervals
  - Overview
  - Bootstrap confidence intervals

- Estimator Properties
  - Estimator Bias / Mean Squared Error
  - Law of Large Numbers / Central Limit Theorem

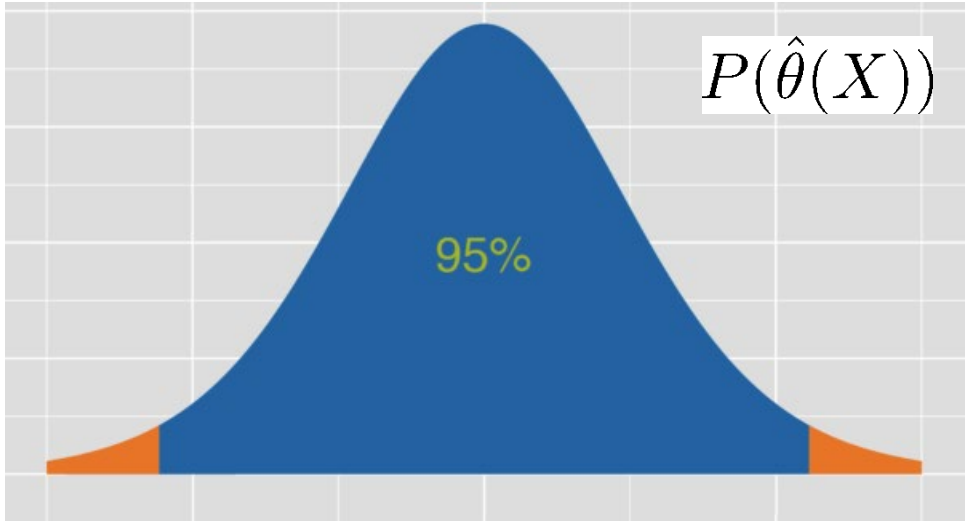
Confidence Intervals

Intuition Find an interval such that we are *pretty sure* it encompasses the true parameter value.

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

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

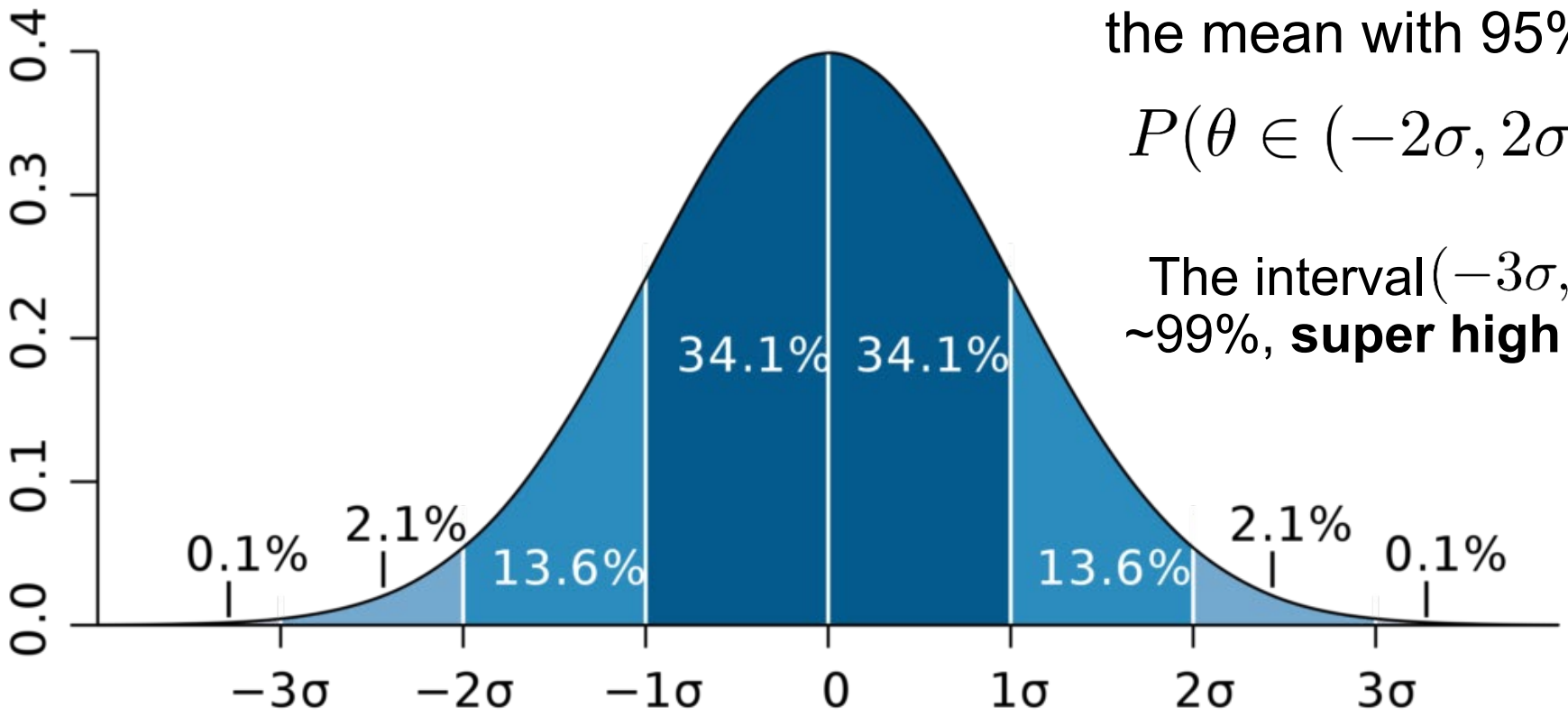
In English the interval (a, b) contains the true parameter value θ with probability **at least** $1 - \alpha$



- Intervals must be computed from data $a(X_1, \ldots, X_n)$ and $b(X_1, \ldots, X_n)$
- Interval (a,b) is **random**, parameter θ is **not random** (it is fixed)
- Requires that we know the distribution of the estimator $\hat{\theta}$

Confidence Intervals of the Normal Distribution

Many estimators follow a normal distribution with enough data (central limit theorem)



A Normal RV falls within 2σ of the mean with 95% probability

$$P(\theta \in (-2\sigma, 2\sigma)) \geq 0.95$$

The interval $(-3\sigma, 3\sigma)$ covers ~99%, **super high confidence**

For various reasons, 95% has become standard confidence level

Warning

Question How should we interpret a confidence interval (e.g. 95%)?

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

Hint Think about what is random and what is not…

This is NOT a probability statement about θ.

Wrong The true parameter value lies in the interval (a,b) with probability at least 95%

Right Interval (a,b) contains the true parameter value with probability at least 95%

This is commonly misinterpreted

Warning

Question How should we interpret a confidence interval (e.g. 95%)?

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

Hint Think about what is random and what is not…

Wrong In this experiment there is a 95% chance that our interval contains the true parameter value.

Right If I repeat this experiment many times the interval will contain the true parameter value 95% of the time.

True but useless… we only have one dataset (one experiment)

This is commonly misinterpreted

Interpretation

On day 1, you collect data and construct a 95 percent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_3. You continue this way constructing confidence intervals for a sequence of unrelated parameters $\theta_1, \theta_2, \ldots$ Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

[Source: Wasserman, L. 2004]

Bootstrap Confidence Intervals

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

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

2. Compute estimate $\hat{\theta}_m(X_1^*, \ldots, X_m^*)$

2. Repeat B times to get set of estimators $\hat{\theta}_{m,1}, \hat{\theta}_{m,2}, \ldots, \hat{\theta}_{m,B}$

3. Compute sample mean and sample variance of estimators,

$$\bar{\theta}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_{m,b} \qquad \sigma^2_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} (\hat{\theta}_{m,b} - \bar{\theta}_{\text{boot}})^2$$

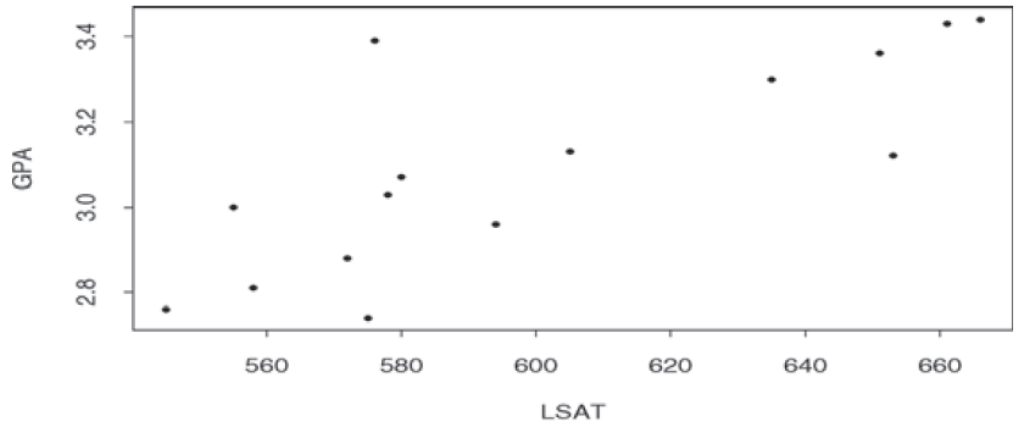
3. 95% Confidence Interval: $\bar{\theta}_{\text{boot}} \pm 2\sigma_{\text{boot}}$

Assumes Normally-distributed estimates $\hat{\theta}_m$***.***

Bootstrap Example

Example Suppose we have LSAT scores and GPA for 15 law students and wish to estimate the correlation between LSAT and GPA:

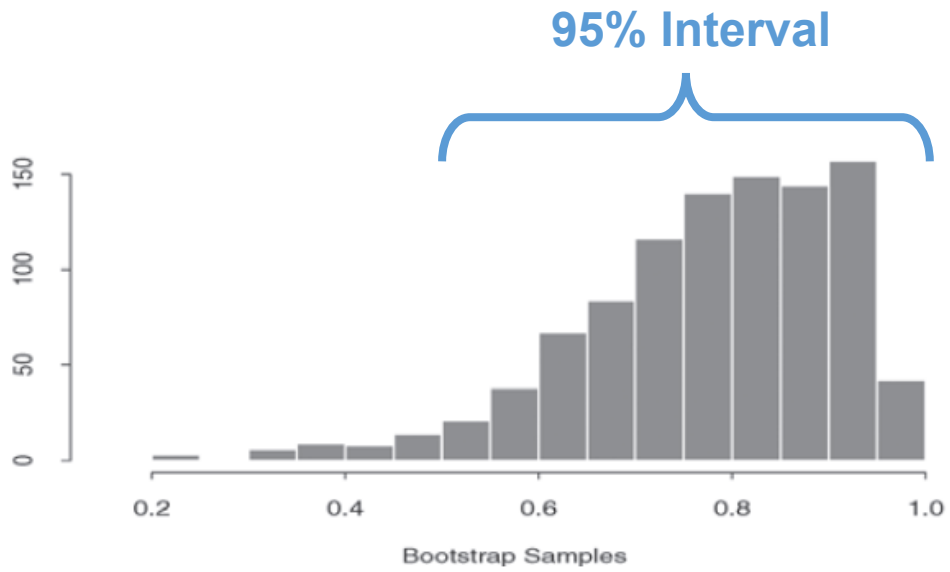
LSAT	576	635	558	578	666	580	555	661
	651	605	653	575	545	572	594	
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
	3.36	3.13	3.12	2.74	2.76	2.88	3.96	



95% Bootstrap confidence interval from B=1000 estimates of the **correlation**,

$$.78 \pm .274 \Rightarrow (.51, 1.00)$$

Q Should we trust this confidence interval? Why or why not?

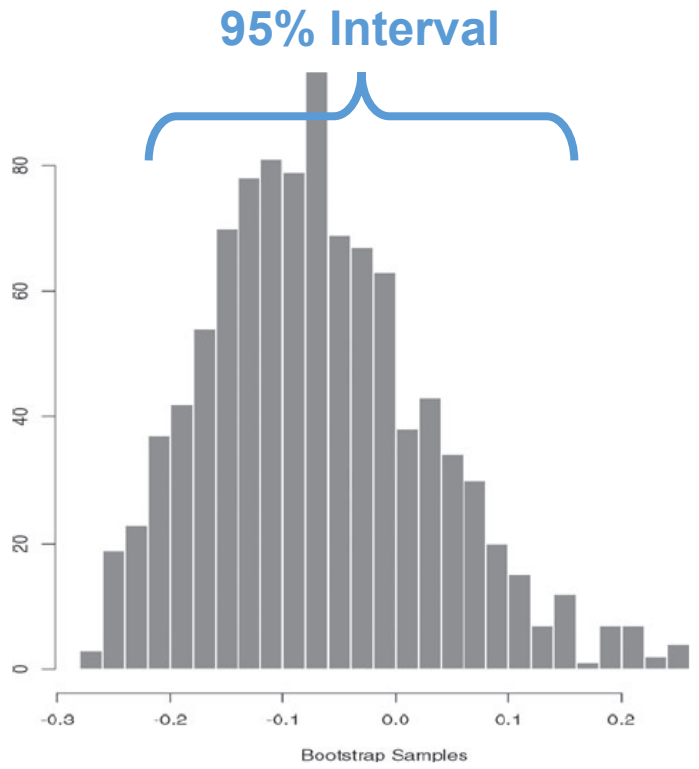


[Source: Wasserman, L. 2004]

Bootstrap Example

Eight subjects who used medical patches to infuse a hormone into the blood using three treatments: placebo, old-patch, new-patch

subject	placebo	old	new	old − placebo	new − old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719



Estimate whether relative efficacy is the same under new drug,

$$\theta = \frac{\mathbf{E}[\text{new} - \text{old}]}{\mathbf{E}[\text{old} - \text{placebo}]}$$

Bootstrap B=1,000 samples yields 95% confidence interval,

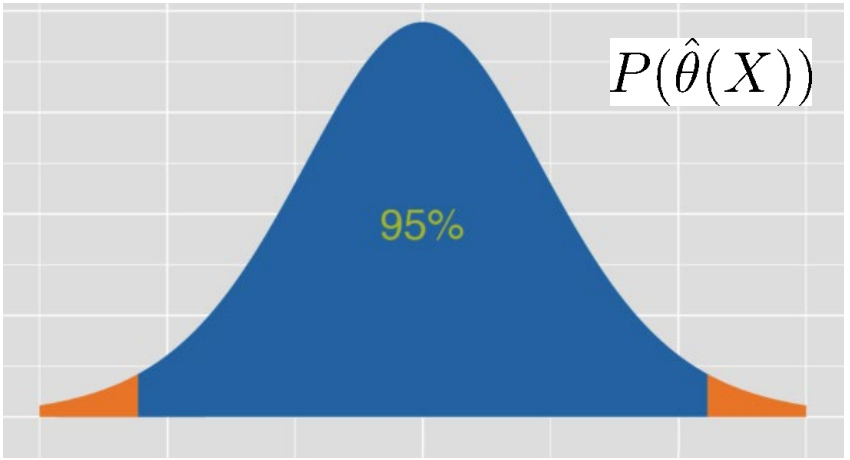
$$\theta \in (-0.24, 0.15)$$

Q Is this more trustworthy than in previous example?

- Parameter Estimation
  - Method of Moments
  - Maximum Likelihood Estimation
- Confidence Intervals
  - Overview
  - Bootstrap confidence intervals
- Estimator Properties
  - Estimator Bias / Mean Squared Error
  - Law of Large Numbers / Central Limit Theorem

Estimator Mean

An estimator $\hat{\theta}(X)$ is a RV so we can compute its moments



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 What is the expected value of $\hat{p}$?

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

(a) Linearity of Expectation Operator

(b) Mean of Bernoulli RV = p

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

Unbiased Estimator

Definition Estimator $\hat{\theta}(X)$ is an **unbiased estimator** of θ if,

$$\mathbf{E}[\hat{\theta}(X)] = \theta$$

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

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

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

Source of bias:
plug-in mean estimate

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

Correcting bias yields unbiased variance estimator:

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

Estimator Variance

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

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

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

(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,

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

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

Decreases linearly with number of samples N

Bias-Variance Tradeoff

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

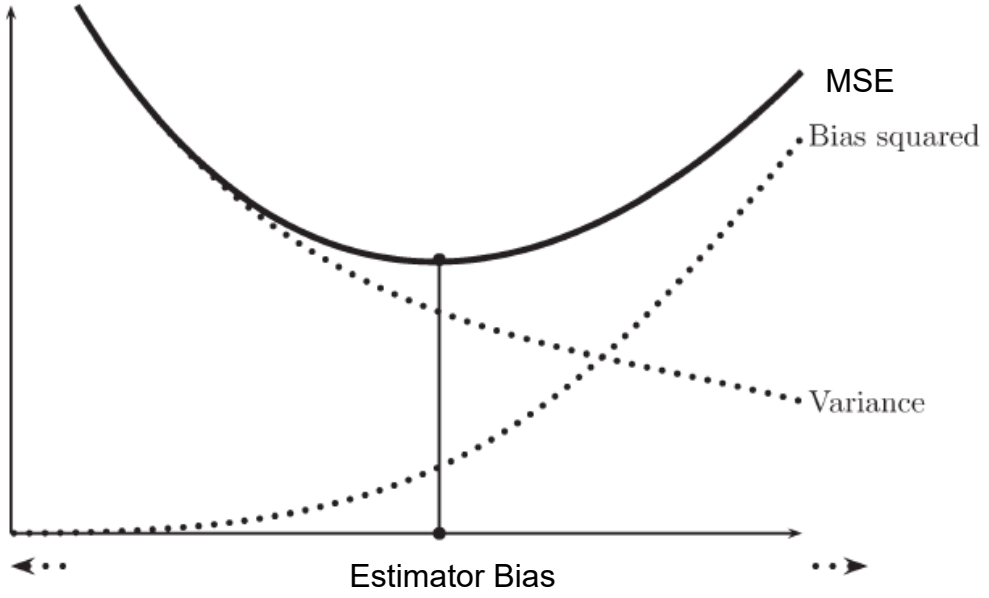
Evaluate the quality of estimate $\hat{\theta}$ using **mean squared error**,

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

- MSE for unbiased estimators is just,

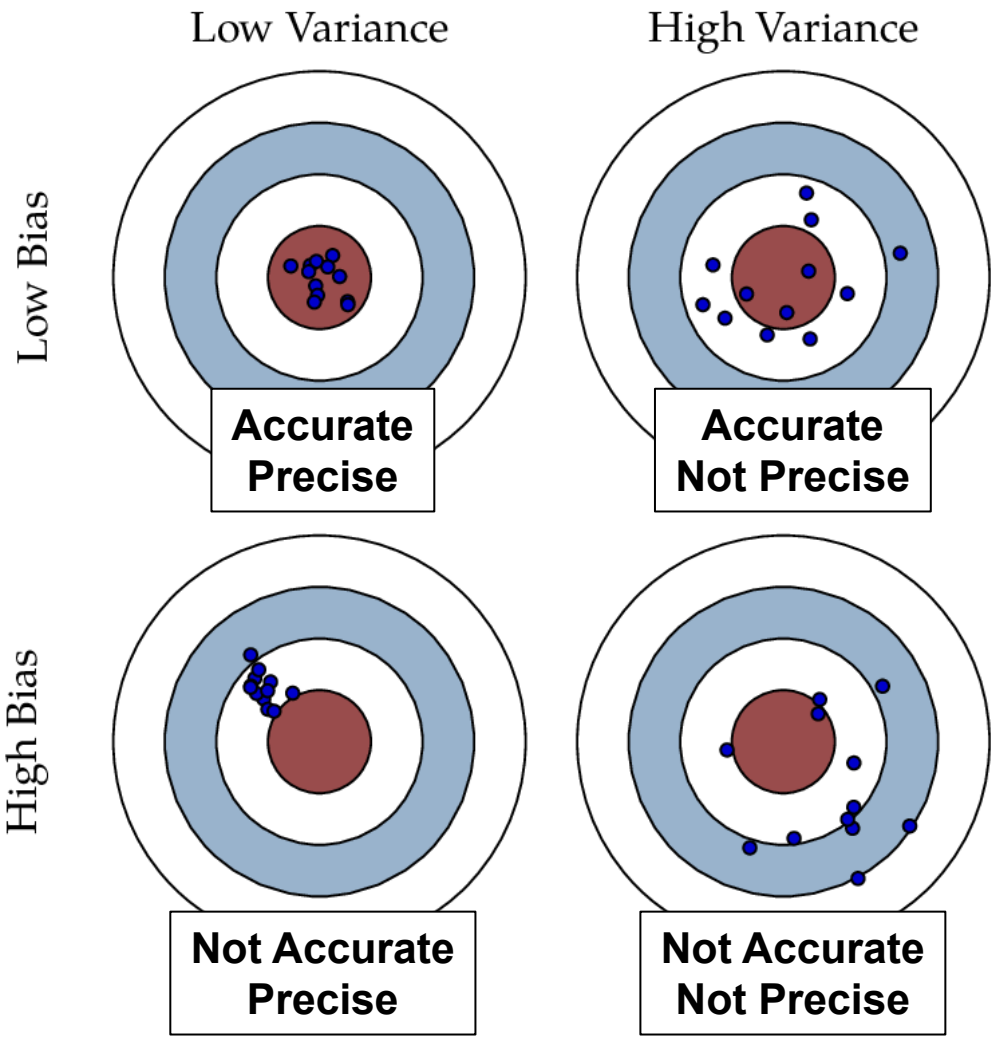
$$\mathrm{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta})$$

- Bias-variance is fundamental tradeoff in statistical estimation
- MSE increases as **square** of bias
- Estimators with small bias (but low variance) can have lower MSE than unbiased estimators



Bias-Variance Tradeoff

Suppose an archer takes multiple shots at a target…



Bias-Variance Decomposition

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

Law of Large Numbers (LLN)

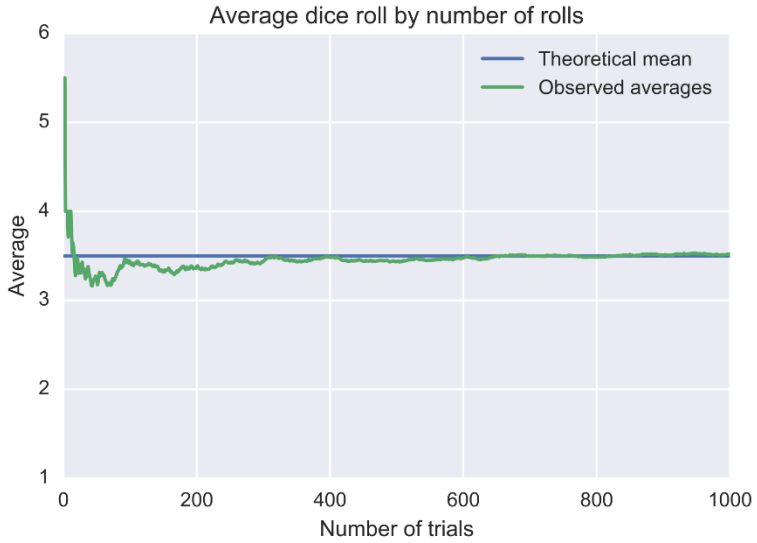
We now know the **sample mean** is an unbiased estimator, namely:

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

But, expected value is not always high probability. Will we achieve the true mean?

$$\lim_{N\to\infty} \bar{X}_N \to \mathbf{E}[X]$$

Yes, with high probability



This is the **law of large numbers**

- Weak Law: Converges to mean with high probability
- Strong Law: Stronger notion of convergence (if variance is finite)

But what is the distribution of $\bar{X}_N$?

Central Limit Theorem (CLT)

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

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

Alternatively written as,

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

Comments

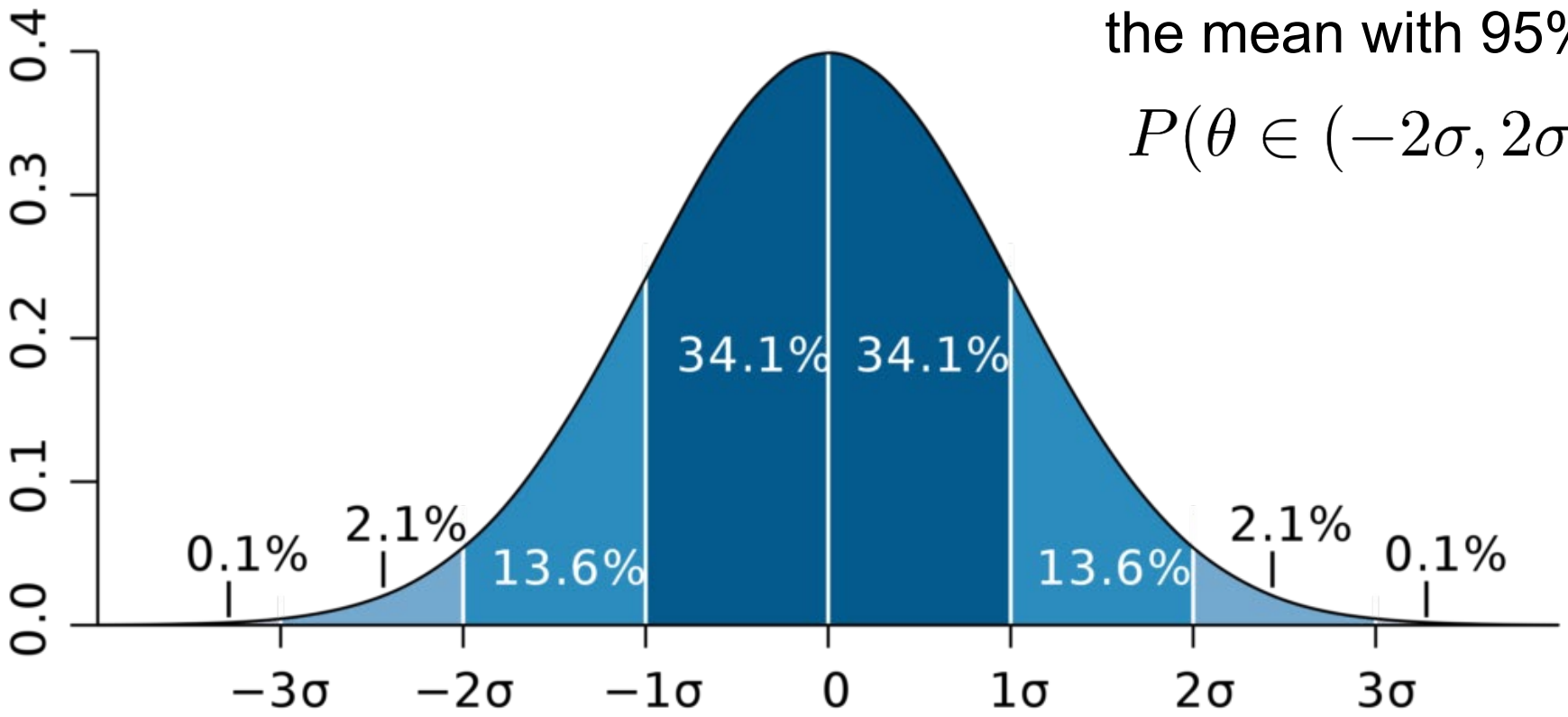
- LLN says estimates $\bar{X}_N$ "pile up" near true mean, CLT says *how* they pile up
- Pretty remarkable since we make no assumption about how X_i are distributed
- Variance of X_i **must be finite**, i.e. $\sigma^2 < \infty$

Confidence Intervals of the Normal Distribution

CLT is why we often derive confidence intervals from Normal

A Normal RV falls within 2σ of the mean with 95% probability

$$P(\theta \in (-2\sigma, 2\sigma)) \geq 0.95$$



CLT says sample mean approaches normal in the infinite limit only!

Classical Statistics Review

- **Statistical Estimation** infers unknown parameters θ of a distribution $p(X;\theta)$ from observed data $X_1, \ldots, X_n$
- There are **many** estimators $\hat{\theta}$, we have seen 3: Method of Moments, Maximum Likelihood, Sample Average (sometimes equivalent)
- 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 estimating confidence intervals

Caution

- Confidence intervals are often misinterpreted!
- Bootstrap confidence intervals we have seen **assume normal distribution**

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.
- **LLN** and **CLT** are *asymptotic statements* and do not hold for finite data