# Statistics MM2: Parameter estimation


**Lecturer: Israel Leyva-Mayorga**

**email: ilm@es.aau.dk**


AALBORG UNIVERSITY
DENMARK

Connectivity

# Schedule

1. Introduction to statistics
2. **Parameter estimation**
3. Confidence intervals
4. Hypothesis testing 1
5. Hypothesis testing 2
6. Regression
7. Workshop: wrap-up and exam problems

# Outline

**Recap on sampling**

**Types of estimation**

**Estimating the mean and variance**

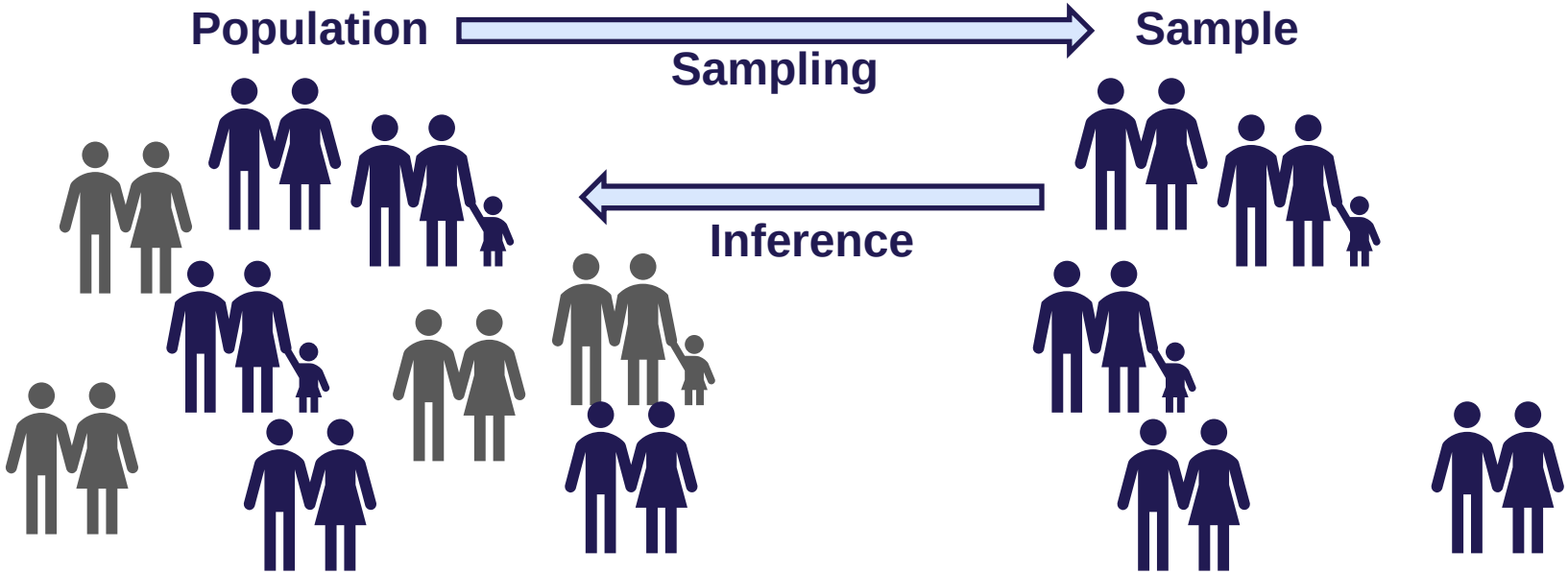
**Evaluating estimators**

**Maximum likelihood Estimation (MLE)**

# Recap on sampling

# Sampling

If we cannot measure the whole population, we use a smaller sample

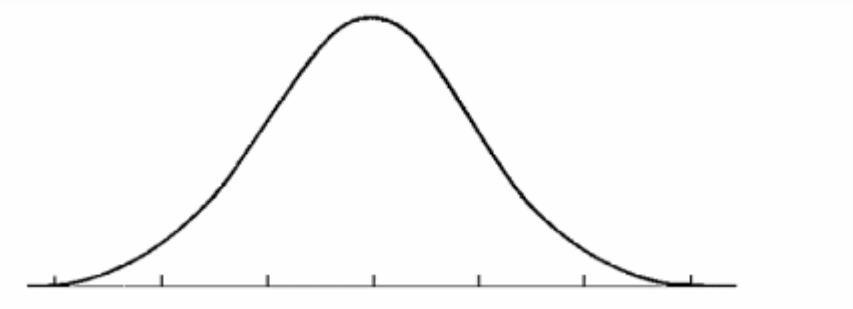


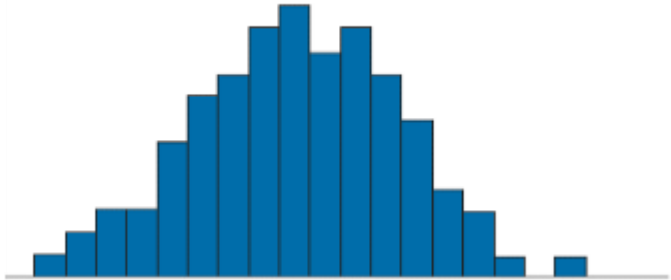
# How do we create a sample?

We randomly draw $n$ values from the population

Each value $X_i$ is a random variable with distribution $F$

We use the sample to estimate some parameter of $F$ (inference)





# Implications and assumptions during sampling

**If** we randomly draw $n$ values from the population
**And** each value $X_i$ is a random variable with distribution $F$
**Then** a sufficiently large sample will ***look like*** the whole population

**What does it mean *looking like*?**
The parameters of the population are similar to the statistics
The parameters describe the population and the statistics describe the sample

- Mean
- Variance
- Quantiles

# Types of estimation

# Parametric estimation

We observe a sample $X_1, X_2, \dots, X_n$

Each value $X_i$ is a random variable with **known distribution** $F$ with parameter $\theta$

**The parameter $\theta$ is a fixed value and not a RV**

From the sample with $n$ points, we create an estimate of $\theta$, denoted as

$$\hat{\theta}_n = h(X_1, X_2, \dots, X_n)$$

The estimate $\hat{\theta}_n$ is our statistic derived from the sample data

**Since $\hat{\theta}_n$ depends on the sampled data, it is a RV**

What are the properties of $\hat{\theta}_n$?

We hope that the estimator $\hat{\theta}_n$ is close to the real value of $\theta$

# Non-parametric estimation

There are no assumptions on the parameters for distribution $F$

**We don't need to know** $F$

When compared to parametric estimation

- 👍 Much more general
- 👎 Much more difficult to do

**Example:**

Histograms and order statistics (quantiles) are non-parametric

Can be used to show the shape of the distribution

But don't give us any mathematical description of it

# Estimating the mean and variance

# Properties of the sample mean pt. 1

If $X_1, X_2, \ldots, X_n$ are random variables, the **sample mean** is the random variable

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Assuming the samples are drawn independently:

- The expectation of the sample mean is

$$\mathbb{E}(\bar{X}_n) = \mu$$

- The variance of the sample mean is $\mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n}$

Comes from the **central limit theorem**
and the fact that the variance of each sample is $\mathrm{var}(X_i) = \sigma^2$

# Properties of the sample mean pt. 2

**The law of large numbers**

$\bar{X}_n$ **converges in probability** to $\mu = \mathbb{E}(X_i)$ as $n \to \infty$

$$\bar{X}_n \xrightarrow{P} \mu = \mathbb{E}(X_i)$$

This means that, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

So, $\bar{X}_n$ is close to $\mu$ with high probability if the sample size is large

# Properties of the sample mean pt. 3

**The central limit theorem**

If $n$ is large:

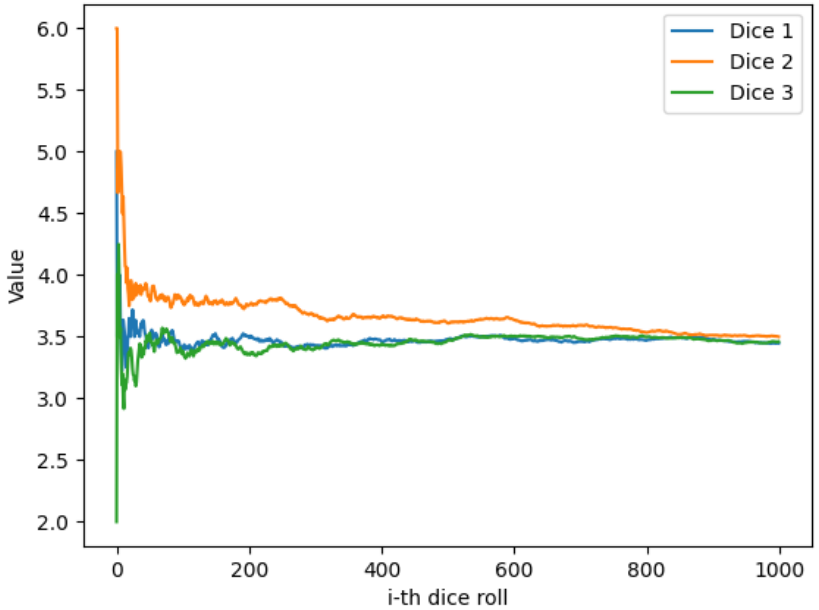
$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n \approx N(n\mu, n\sigma^2)$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Recall that $F(x; \mu, \sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

# Dice example

We roll a dice $i = 1, 2, \ldots, n$ times and compute the sample mean $X_i$



# Estimating the variance

Sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

The term $n-1$ is needed to have an unbiased estimate

$$\mathbb{E}[S_n^2] = \sigma^2$$

# Evaluating estimators

# Metrics to evaluate estimators

An estimator should be as close as possible to the true value
This should happen as frequently as possible

What does this mean?

**Metrics for evaluation:**

- **Bias:** measures accuracy
- **Variance:** measures precision
- **Mean square error:** measures both accuracy and precision

# Bias

**Measure of how close the estimate is to the true value**

Let $\hat{\theta}_n = h(X_1, X_2, \ldots, X_n)$ be the estimator for parameter $\theta$

**The bias is defined as**

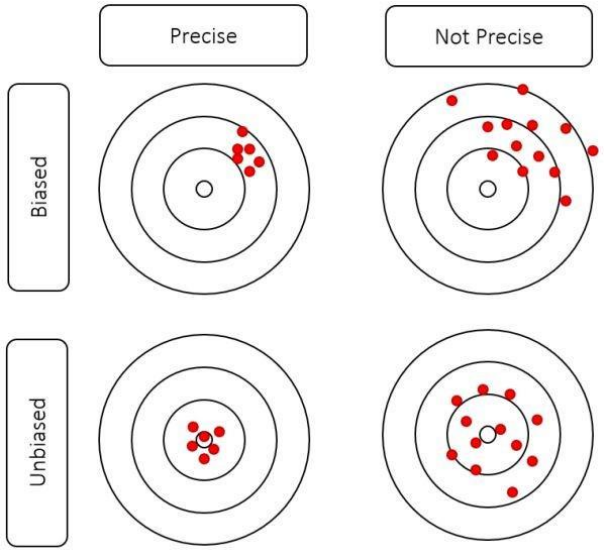
$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

**Unbiased estimator:**

An estimator with $\text{bias}(\hat{\theta}_n) = 0$

This is what we hope for

There are tricks to correct the bias



# Variance and standard error

**Measure of how precise the estimator is for different samples**

Let $\hat{\theta}_n = h(X_1, X_2, \ldots, X_n)$ be the estimator for parameter $\theta$

**The variance is**

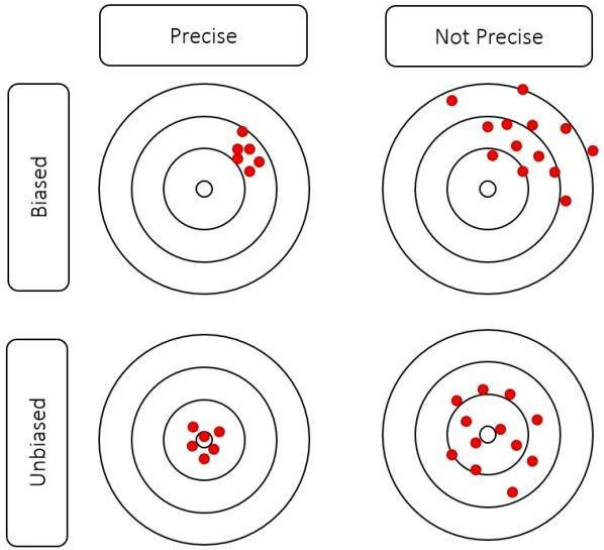
$$\mathrm{var}(\hat{\theta}_n) = \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)^2\right]$$

**The standard error is**

$$\mathrm{se}(\hat{\theta}_n) = \sqrt{\mathrm{var}(\hat{\theta}_n)}$$

but $\mathrm{se}(\hat{\theta}_n)$ depends on $F$ and might be unknown

**The estimated standard error is** $\widehat{se}(\hat{\theta}_n)$



# Example: Point estimator for Bernoulli RVs pt. 1

Let $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$. We don't know the real value of $p$
We have to create an estimate for $p$, defined as

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then

$$\mathbb{E}(\hat{p}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = p$$

Is this estimator biased or not?

# Example: Point estimator for Bernoulli RVs pt. 2

The **variance of a RV** is $\text{var}(X_i) = \mathbb{E}[(X_i - \mu_i)^2] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$

$$\text{var}\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} a_i^2 \text{var}(X_i)$$

The **variance** for our estimator is

$$\text{var}(\hat{p}_n) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$$

Then, the **estimated standard error** is $\widehat{se} = \sqrt{\text{var}(\hat{p}_n)} = \sqrt{p(1-p)/n}$

# The Mean Squared Error (MSE)

We define the **Mean Squared Error (MSE)** of an estimator as

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

Combines the bias and variance

**An estimator is consistent** if $\mathrm{bias}\left(\hat{\theta}_n\right) \to 0$ and $\mathrm{var}\left(\hat{\theta}_n\right) \to 0$ as $\mathrm{n} \to \infty$

This means that $\hat{\theta}_n \xrightarrow{P} \theta$

Is our estimator for parameter $p$ of a Bernoulli RV, namely $\hat{p}_n$, consistent?
Recall that $\mathbb{E}(\hat{p}_n) = p$, $\mathrm{bias}(\hat{p}_n) = \mathbb{E}(\hat{p}_n) - p$, and $\mathrm{var}(\hat{p}_n) = p(1-p)/n$

# Example

Let $X_1, X_2, \dots, X_n$ be i.i.d. RVs with mean $\mathbb{E}[X_i] = \theta$ and variance $\text{var}(X_i) = \sigma^2$

Consider the following two estimators for $\theta$

1. $\hat{\theta}_1 = X_1$
2. $\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Which one is a better estimator?

For $\hat{\theta}_1 = X_1$: $\text{bias}(\hat{\theta}_1) = \mathbb{E}[X_1] - \theta = 0$ and $\text{var}(X_1) = \sigma^2$

For $\hat{\theta}_n = \bar{X}_n$: $\text{bias}(\hat{\theta}_n) = \mathbb{E}[\bar{X}_n] - \theta = 0$ and $\text{var}(\bar{X}_n) = \sigma^2/n$

Since $\text{var}(\bar{X}_n) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$, **the estimator $\hat{\theta}_n = \bar{X}_n$ is consistent**

The estimator $\hat{\theta}_1 = X_1$ **is not consistent**

# The bias-variance trade-off

Present in many cases in statistics and machine learning

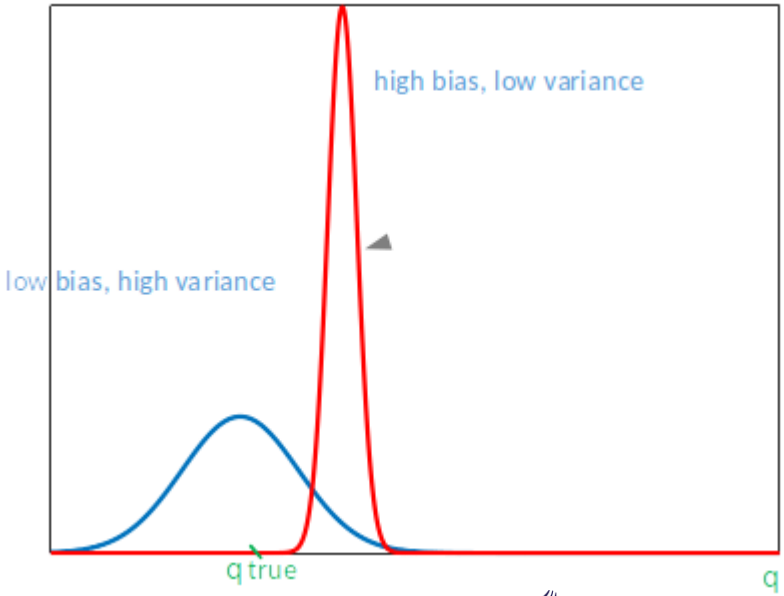
The variance of a parameter estimation can be decreased by increasing the bias

**Increased complexity**

High accuracy but sensitive to variations

**Reduced complexity**

Low accuracy but resilient to variations



# Maximum Likelihood Estimation (MLE)

# Maximum Likelihood Estimation (MLE)

**The most common method for parametric estimation**

The **likelihood function for** $X_1, X_2, \dots, X_n$ i.i.d RVs with PDF/pmf $f(x;\theta)$ is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Usually, we use the **log-likelihood function** is $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$

We treat the likelihood function as a function of parameter $\theta$

**Maximum Likelihood Estimator (MLE)** $\hat{\theta}_n$: the value of $\theta$ that maximizes $\mathcal{L}_n(\theta)$

**The MLE is consistent**

# MLE for discrete and continuous RVs

Suppose we observe the outcomes $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$

If $X_1, X_2, \dots, X_n$ are **discrete RVs**, the likelihood function is the joint pmf

$$\mathcal{L}_n(\theta) = \mathcal{L}(x_1, x_2, \dots, x_n; \theta) = P_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)$$

If $X_1, X_2, \dots, X_n$ are **continuous RVs**, the likelihood function is the joint PDF

$$\mathcal{L}_n(\theta) = \mathcal{L}(x_1, x_2, \dots, x_n; \theta) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)$$

# Example: MLE for Bernoulli RV pt. 1

Let $X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p)$ with pmf

$$f(x; p) = p^x (1-p)^{1-x},$$

The likelihood function is

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

If we define $X = \sum_{i=1}^{n} X_i$ we get the likelihood and log-likelihood functions:

$$\mathcal{L}_n(p) = p^X (1-p)^{n-X}$$

$$\ell_n(p) = \log(p^X (1-p)^{n-X}) = X \log p + (n-X)\log(1-p)$$

Which one should we work with?

# Example: MLE for Bernoulli RV pt. 2

How do we find the value of $p$ that maximizes the likelihood $\mathcal{L}_n(p)$ or $\ell_n(p)$?

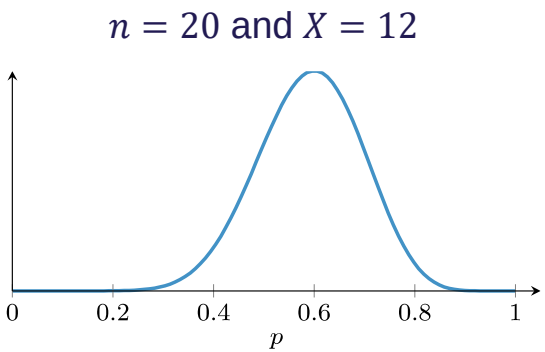
1. Take the derivative with respect to $p$ and make it equal to 0

$$\frac{d\mathcal{L}_n(p)}{dp} = Xp^{X-1}(1-p)^{n-X} - (n-X)(1-p)^{n-X-1}p^X = 0$$

$$\frac{d\ell_n(p)}{dp} = \frac{X}{p} - \frac{n-X}{(1-p)} = 0$$

2. Solve for $p$

$$\hat{p}_n = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$



# Finally, an MLE example with numbers pt. 1

There's a bag with 3 balls. Each ball is either red or blue.

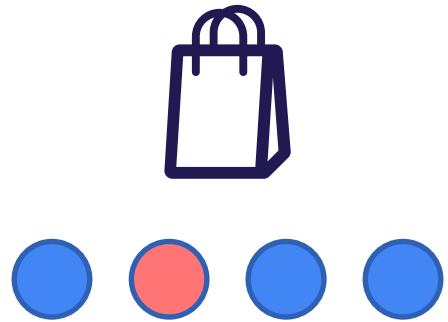
We denote the number of blue balls as $\theta$, whose value can be 0, 1, 2, or 3.

We estimate $\theta$ by grabbing and putting back a ball 4 times

This is a process of selection with replacement

Let $X_1, X_2, X_3, X_4$ be the RVs of the color of the $i$-th ball

$$X_i = \begin{cases} 1, & \text{if the ball is blue} \\ 0, & \text{if the ball is red} \end{cases}$$



Then, $X_i \sim \text{Bernoulli}\left(\frac{\theta}{3}\right)$

The outcomes are $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1$

That is, 3 out of 4 of the balls are blue

# Finally, an MLE example with numbers pt. 2

1. For each possible value of $\theta$, find $P_{X_1,X_2,X_3,X_4}\ (1,0,1,1;\theta)$

$$P_{X_1,X_2,X_3,X_4}\ (1,0,1,1;\theta=0)=\prod_{i=1}^{4}P(X_i=x_i;\ \theta=0)=0\times1\times0\times0=0$$

$$P_{X_1,X_2,X_3,X_4}\ (1,0,1,1;\theta=1)=\prod_{i=1}^{4}P(X_i=x_i;\ \theta=1)=\frac{1}{3}\times\frac{2}{3}\times\frac{1}{3}\times\frac{1}{3}=\frac{2}{81}=0.0247$$

$$P_{X_1,X_2,X_3,X_4}\ (1,0,1,1;\theta=2)=\prod_{i=1}^{4}P(X_i=x_i;\ \theta=1)=\frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3}=\frac{8}{81}=0.0987$$

$$P_{X_1,X_2,X_3,X_4}\ (1,0,1,1;\theta=3)=1\times0\times1\times1=0$$

# Finally, an MLE example with numbers pt. 3

2. Use the definition of MLE to estimate $\theta$ with $n$ experiments given $p = \theta/3$

Previously, we defined $\ell_n(p) = \log(p^X(1-p)^{n-X}) = X\log p + (n-X)\log(1-p)$

$$\ell_n(\theta) = \log\left(\left(\frac{\theta}{3}\right)^X\left(1-\frac{\theta}{3}\right)^{n-X}\right) = X(\log\theta - \log 3) + (n-X)\log\left(1-\frac{\theta}{3}\right)$$

$$\frac{d\ell_n(\theta)}{d\theta} = \frac{X}{\theta} - \frac{n-X}{3\left(1-\frac{\theta}{3}\right)} = 0$$

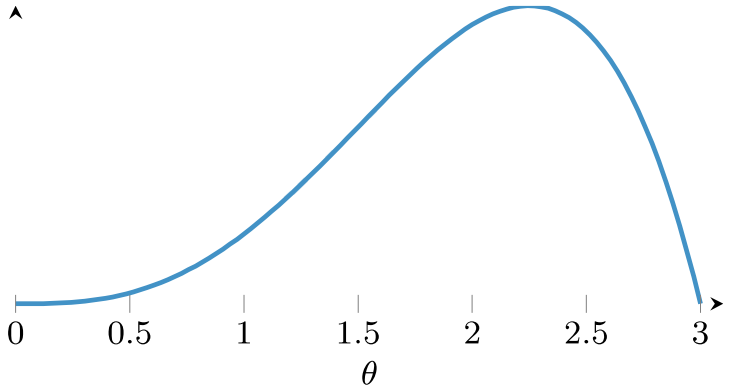
$$X(3-\theta) = \theta(n-X) \Longrightarrow 3X - \theta X = n\theta - \theta X \Longrightarrow \theta = \frac{3X}{n} = 3\bar{X}_n$$

# Finally, an MLE example with numbers pt. 4

For $n = 4$ and $X = 3$ we have $\bar{X}_4 = 3/4$ and the MLE becomes

$$\hat{\theta}_4 = 3\bar{X}_4 = 3 \times 3/4 = 9/4 = 2.25$$

Since $\hat{\theta}_n$ must be an integer, we take the closest integer value and $\hat{\theta}_4 = 2$



# Summary

# Summary

Sampling is essential for estimation

In **parametric estimation**, we know the distribution but not the parameter

- The estimator is a RV that takes a value based on the sample

In **non-parametric estimation**, we don’t need to know the distribution

There can be an infinite number of estimators

We find the best ones using **metrics: bias, variance, and MSE**

If the MSE goes to 0, the estimator is **consistent**

**The Maximum Likelihood Estimator (MLE) is consistent**

# Appendix

# Useful derivative formulas and rules

$$\frac{da^x}{dx} = \log(a)a^x$$

$$\frac{d\log_a(x)}{dx} = \frac{1}{x\log(a)}$$

$$\frac{d\log(x)}{dx} = \frac{1}{x}$$

**Product rule:** $h(x) = f(x)g(x)$

$$h'(x) = f'(x)g(x) + f(x)g'(x)$$

**Quotient rule:** $h(x) = f(x)/g(x)$

$$h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

**Chain rule:** $h(x) = f\big(g(x)\big)$

$$h'(x) = f'\big(g(x)\big)g'(x)$$

# Derivations for Bernoulli MLE (log-likelihood)

$$\frac{d\ell_n(p)}{dp} = \frac{X}{p} + \frac{n-X}{(1-p)}(-1) = \frac{X}{p} - \frac{n-X}{(1-p)} = 0$$

$$\frac{X}{p} = \frac{n-X}{(1-p)}$$

$$X(1-p) = p(n-X) \Rightarrow X - Xp = np - Xp$$

$$X = np \Rightarrow p = \hat{p}_n = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$