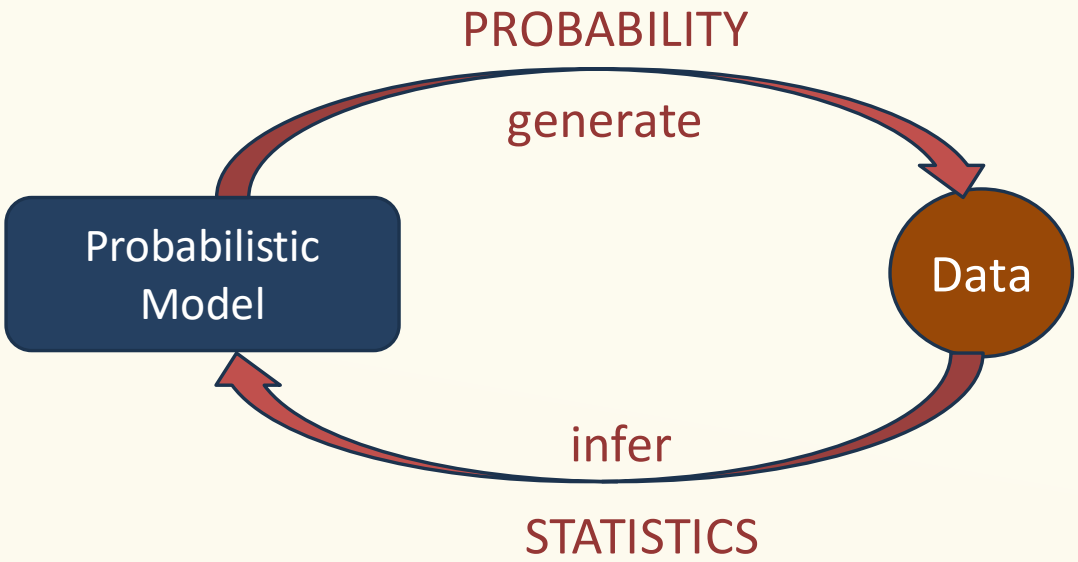
More Maximum Likelihood Estimation

CSE 312 Spring 26

Lecture 24

Several slides by Mor Harchol-Balter from
Book “Intro to Probability for Computing”

Probability vs. Statistics



Maximum Likelihood Estimation: Given some data drawn from a known distribution with unknown parameters, estimate the parameters of the distribution

Creating an maximum likelihood estimator

Goal: Estimate an unknown value θ, given sample data X from known distribution with unknown parameter(s) θ.

Step 1: Define

$$\hat{\theta}_{ML}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$$

This is the "likelihood function"

Our estimator is the value of θ that maximizes the likelihood function

Example of MLE

Goal: Estimate θ =Number of pink jelly beans

Experiment: Randomly sample $n = 20$ beans w/ replacement

X = # pink jelly beans in sample

1000 jelly beans total

$$X \sim \text{Binomial}\left(n, \frac{\theta}{1000}\right)$$

Example of MLE

Goal: Estimate θ =Number of pink jelly beans

Experiment: Randomly sample $n = 20$ beans w/ replacement

X = # pink jelly beans in sample is Binomial (n, θ /1000)

1000 jelly beans total

Suppose we observe $X = 3$ pinks in our sample with $n = 20$.

$$\boldsymbol{P}(\mathrm{X} = 3; \theta) = \binom{20}{3} \left(\frac{\theta}{1000}\right)^{3} \cdot \left(1 - \frac{\theta}{1000}\right)^{17}$$

What value of θ maximizes this?

Example of MLE

Goal: Estimate θ =Number of pink jelly beans

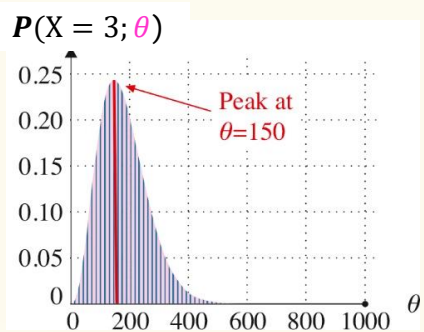
Experiment: Randomly sample $n = 20$ beans w/ replacement

X = # pink jelly beans in sample

Q: Suppose we observe $X = 3$ pinks in our sample. What is $\mathrm{P}(\mathrm{X} = 3; \theta)$?

What value of θ maximizes this?

$\theta = 150$



$$\hat{\theta}_{ML}(X = 3)$$

$$= \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = 3; \theta)$$

$$= 150$$

Example of MLE

Q: What is the likelihood function $\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$?

$$\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

1000 jelly beans total

X = # pink jelly beans in sample

Q: What is $\hat{\theta}_{ML}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$?

$$0 = \frac{d}{d\theta}\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \frac{d}{d\theta}\binom{n}{x}\left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

$$= \binom{n}{x}\left(\frac{\theta}{1000}\right)^{x} \cdot (n-x) \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x-1} \cdot \frac{-1}{1000} + \binom{n}{x} \cdot x \left(\frac{\theta}{1000}\right)^{x-1} \cdot \frac{1}{1000} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

$$= -\frac{n-x}{1000} \cdot \frac{\theta}{1000} + \frac{x}{1000} \cdot \left(1 - \frac{\theta}{1000}\right)$$

Solving, we get: $\theta = \frac{1000x}{n}$

2nd derivative is negative, so this is a max ✓

Maximizing Log likelihood Simpler

Goal: Estimate an unknown value θ, given sample data X

Define

$$\hat{\theta}\ (X = x) = \arg\max_{\theta}\ \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)? = \arg\max_{\theta}\ \ln\ \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)?$$

This is the likelihood function

This is the log likelihood function

equivalent
Why??

Example of MLE

$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln(a^b) = b \cdot \ln(a)$$

The likelihood function $\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$?

$$\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

1000 jelly beans total

X = # pink jelly beans in sample

$$\ln \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \ln \binom{n}{x} + \mathrm{x} \ln \left(\frac{\theta}{1000}\right) + (n - \mathrm{x}) \ln \left(1 - \frac{\theta}{1000}\right)$$

$\frac{1000-\theta}{1000}$

$$= \ln \binom{n}{x} + x \ln \theta - x \ln(1000) + (n-x) \ln(1000 - \theta) - (n-x) \ln(1000)$$

$$0 = \frac{d}{d\theta} \ln \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \frac{x}{\theta} + \frac{(n-x)(-1)}{1000 - \theta} = 0$$

Example of MLE

What is the likelihood function $\boldsymbol{P}\{X = x \mid \theta\}$?

$$\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

1000 jelly beans total

X = # pink jelly beans in sample

What is $\hat{\theta}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$ $= \arg\max_{\theta} \ln \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$?

$$\hat{\theta}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \frac{1000x}{n}$$

This holds $\forall x$

Agenda

- MLE for Continuous Distributions
- MLE for Normal Distribution
- Unbiased and Consistent Estimators

The Continuous Case

Given n (independent) samples $X_1 = x_1, \dots, X_n = x_n$ from (continuous) parametric model $f(x_i; \theta)$ which is now a family of densities

Definition. The **likelihood** of independent observations $x_1, \dots, x_n$ is

$$\mathcal{L}(X_1 = x_1, \dots, X_n = x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Replace pmf with pdf!

Why density?

$$P(X \in [x, x+dx]) \approx f(x)\,dx$$

- Density ≠ probability, but:
  - For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  - has desired property that likelihood increases with better fit to the model

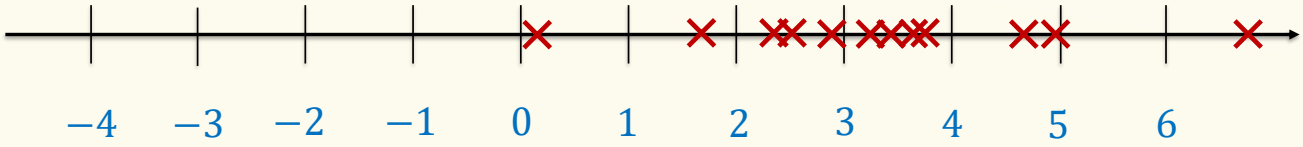
Agenda

- MLE for for Continuous Distributions
- MLE for Normal Distribution
- Unbiased and Consistent Estimators

unknown

n samples $X_1 = x_1, \ldots, X_n = x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely μ?

[i.e., we are given the promise that the variance is 1]



Example – Gaussian Parameters

$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln(a^b) = b \cdot \ln(a)$$

Normal outcomes $X_1 = x_1, \dots, X_n = x_n$, known variance $\sigma^2 = 1$

Goal: estimate θ, the unknown expectation

$$\mathcal{L}(X_1 = x_1, \dots, X_n = x_n;\ \theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} \right)$$

$$= \left(\frac{1}{\sqrt{2\pi}} \right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta)^2}{2}}$$

$$\ln \mathcal{L} = n \ln\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^{n} \ln e^{-\frac{(x_i-\theta)^2}{2}}$$

$$= -\frac{n}{2}\ln(2\pi) + -\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2}$$

Example – Gaussian Parameters

Goal: estimate θ = expectation

Normal outcomes $X_1 = x_1, \dots, X_n = x_n$, known variance $\sigma^2 = 1$

$$\ln \mathcal{L}(X_1 = x_1, \dots, X_n = x_n;\ \theta) = -\,n\frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(X_1 = x_1, \dots, X_n = x_n; \theta) = -\sum_{i=1}^{n} (\theta - x_i)$$

$$n\hat{\theta} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note: $\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{(x_i-\theta)^2}{2} = \frac{1}{2} \cdot 2 \cdot (x_i - \theta) \cdot (-1) = \theta - x_i$

Example – Gaussian Parameters

Goal: estimate θ = expectation

Normal outcomes $X_1 = x_1, \ldots, X_n = x_n$, known variance $\sigma^2 = 1$

$$\ln \mathcal{L}(X_1 = x_1, \ldots, X_n = x_n;\ \theta) = -\frac{n \ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2}$$

Note: $\frac{\partial}{\partial \theta} \frac{(x_i - \theta)^2}{2} = \frac{1}{2} \cdot 2 \cdot (x_i - \theta) \cdot (-1) = \theta - x_i$

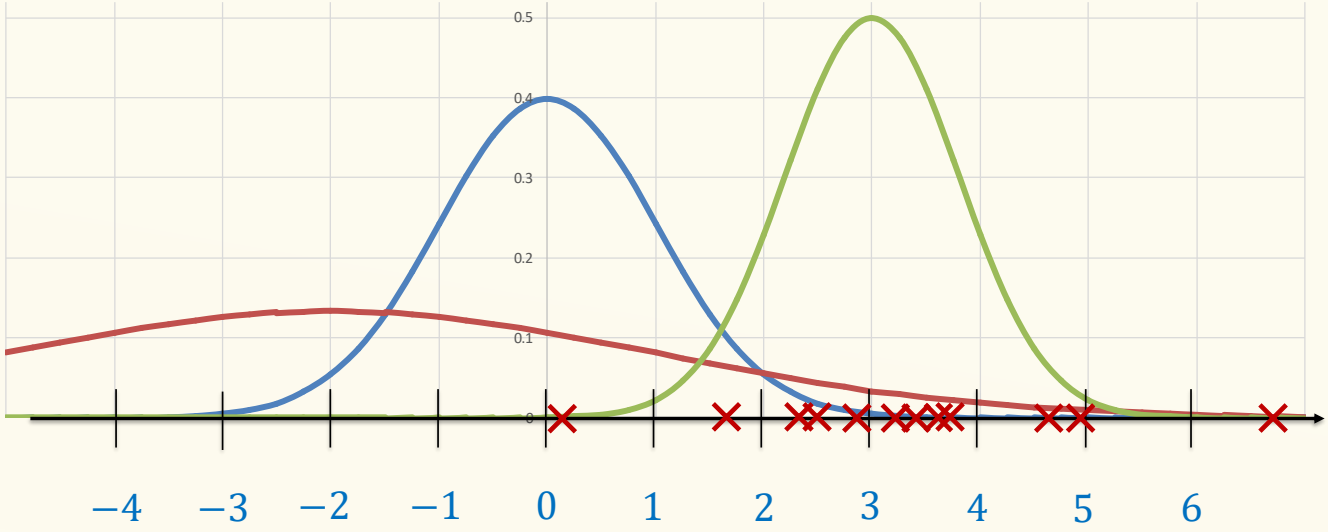
$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(X_1 = x_1, \ldots, X_n = x_n; \theta) = \sum_{i=1}^{n} (x_i - \theta) = \sum_{i=1}^{n} x_i - n\theta$$

So… solve $\sum_{i=1}^{n} x_i - n\hat{\theta} = 0$ for $\hat{\theta}$

$$\hat{\theta}(X_1 = x_1, \ldots, X_n = x_n) = \frac{\sum_i^n x_i}{n}$$

In other words, MLE is the *sample mean* of the data.

Next: n samples $X_1 = x_1, \ldots, X_n = x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, \sigma^2)$. Most likely μ and σ^2?



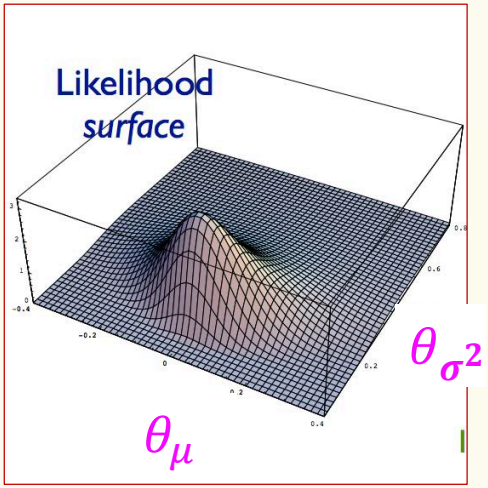
Two-parameter optimization

$\frac{d}{dx} \ln x = \frac{1}{x}$

$\ln(ab) = \ln(a) + \ln(b)$
$\ln(a/b) = \ln(a) - \ln(b)$
$\ln(a^b) = b \cdot \ln(a)$

Normal outcomes $X_1 = x_1, \ldots, X_n = x_n$

Goal: estimate θ_μ = expectation and θ_{σ^2} = variance



$$\mathcal{L}(X_1 = x_1, \ldots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}) = \prod_{i=1}^{n} f(x_i; \theta_\mu, \theta_{\sigma^2})$$

$$= \left(\frac{1}{\sqrt{2\pi\theta_{\sigma^2}}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta_\mu)^2}{2\theta_{\sigma^2}}}$$

Two-parameter estimation

$$\ln \mathcal{L}\left(X_1 = x_1, \ldots, X_n = x_n; \theta_\mu, \theta_{\sigma^2} \right) = -\frac{\ln(2\pi\, \theta_{\sigma^2})}{2} - \sum_{i=1}^{n} \frac{\left(x_i - \theta_\mu\right)^2}{2\theta_{\sigma^2}}$$

Find pair $\hat{\theta}_\mu, \hat{\theta}_{\sigma^2}$ that maximizes $\ln \mathcal{L}\left(X_1 = x_1, \ldots, X_n = x_n; \theta_\mu, \theta_{\sigma^2} \right)$

Two-parameter estimation

$$\ln \mathcal{L}\left(X_1 = x_1, \ldots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}\right) = -\frac{\ln(2\pi\,\theta_{\sigma^2})}{2} - \sum_{i=1}^{n} \frac{\left(x_i - \theta_\mu\right)^2}{2\theta_{\sigma^2}}$$

We need to find a solution $\hat{\theta}_\mu, \hat{\theta}_{\sigma^2}$ to

$$\frac{\partial}{\partial \theta_\mu} \ln \mathcal{L}\left(x_1, \ldots, x_n\,; \theta_\mu, \theta_{\sigma^2}\right) = 0$$

$$\frac{\partial}{\partial \theta_{\sigma^2}} \ln \mathcal{L}\left(x_1, \ldots, x_n\,; \theta_\mu, \theta_{\sigma^2}\right) = 0$$

And then check second order conditions.

$-\frac{n}{2}\ln(2\pi)$ $-\frac{n}{2}\ln(\theta_{\sigma^2})$

MLE for Expectation

$$\ln \mathcal{L}(X_1 = x_1, \ldots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}) = -n\frac{\ln(2\pi\,\theta_{\sigma^2})}{2} - \sum_{i=1}^{n}\frac{(x_i - \theta_\mu)^2}{2\theta_{\sigma^2}}$$

$$\frac{\partial}{\partial \theta_\mu}\ln\mathcal{L} = -\frac{1}{2\theta_{\sigma^2}}\sum_{i=1}^{n} 2(\theta_\mu - x_i) = 0$$

$$\frac{\partial}{\partial \theta_{\sigma^2}}\ln\mathcal{L} = -\frac{n}{2}\frac{1}{\theta_{\sigma^2}} - \sum_{i=1}^{n}\frac{(x_i-\theta_\mu)^2}{2}\frac{\partial}{\partial\theta_{\sigma^2}}\frac{1}{\theta_{\sigma^2}}$$

$$= -\frac{n}{2\theta_{\sigma^2}} + \sum_{i=1}^{n}\frac{(x_i-\theta_\mu)^2}{2(\theta_{\sigma^2})^2}$$

MLE for Expectation

$$\ln \mathcal{L}\left(X_1 = x_1, \dots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}\right) = -n\frac{\ln(2\pi\,\theta_{\sigma^2})}{2} - \sum_{i=1}^{n}\frac{\left(x_i - \theta_\mu\right)^2}{2\theta_{\sigma^2}}$$

$$\frac{\partial}{\partial\theta_\mu}\ln \mathcal{L}\left(X_1 = x_1, \dots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}\right) = \frac{1}{\hat{\theta}_{\sigma^2}}\sum_{i}^{n}\left(x_i - \hat{\theta}_\mu\right) = 0$$

$/\cdot\theta_{\sigma^2}$

$\sum x_i - n\hat{\theta}_\mu = 0 \Rightarrow \hat{\theta}_\mu = \frac{1}{n}\sum x_i$

$$\frac{\partial}{\partial\theta_{\sigma^2}}\ln \mathcal{L}\left(x_1, \dots, x_n\,; \hat{\theta}_\mu, \theta_{\sigma^2}\right) = -\frac{n}{2\hat{\theta}_{\sigma^2}} + \frac{1}{2\hat{\theta}_{\sigma^2}^2}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_\mu\right)^2 = 0$$

$/\cdot 2\hat{\theta}_{\sigma^2}^2$

MLE for Expectation

$$\ln \mathcal{L}\left(X_1 = x_1, \dots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}\right) = -n\frac{\ln(2\pi\,\theta_{\sigma^2})}{2} - \sum_{i=1}^{n}\frac{(x_i - \theta_\mu)^2}{2\theta_{\sigma^2}}$$

$$\frac{\partial}{\partial \theta_\mu}\ln \mathcal{L}\left(X_1 = x_1, \dots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}\right) = \frac{1}{\theta_{\sigma^2}}\sum_{i}^{n}(x_i - \theta_\mu)\,. \qquad \text{Set} = 0$$

$$\hat{\theta}_\mu(X_1 = x_1, \dots, X_n = x_n) = \frac{\sum_i^n x_i}{n}$$

In other words, MLE of expectation is (again) the *sample mean* of the data, regardless of θ_2

What about the variance?

MLE for Variance

$$\ln \mathcal{L}(X_1 = x_1, \dots, X_n = x_n; \theta_\mu, \theta_{\sigma^2}) = -n\frac{\ln(2\pi\,\theta_{\sigma^2})}{2} - \sum_{i=1}^{n}\frac{(x_i - \theta_\mu)^2}{2\theta_{\sigma^2}}$$

$$= -n\frac{\ln 2\pi}{2} - n\frac{\ln \theta_{\sigma^2}}{2} - \frac{1}{2\theta_{\sigma^2}}\sum_{i=1}^{n}(x_i - \hat{\theta}_\mu)^2$$

$$\frac{\partial}{\partial \theta_{\sigma^2}} \ln \mathcal{L}(x_1, \dots, x_n\,; \hat{\theta}_\mu, \theta_{\sigma^2}) = -\frac{n}{2\theta_{\sigma^2}} + \frac{1}{2\theta_{\sigma^2}^2}\sum_{i=1}^{n}(x_i - \hat{\theta}_\mu)^2 \quad \text{Set } = 0$$

to solve $= 0$ cross multiply by $2\theta_{\sigma^2}^2$ and solve for θ_{σ^2}

$$\hat{\theta}_{\sigma^2}\,(X_1 = x_1, \dots, X_n = x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_\mu)^2$$

$\hat{\theta}_\mu = \frac{1}{n}\sum x_i$

$Y = \{ x_i \text{ with prob } \frac{1}{n} \quad i = 1, \dots, n$

In other words, MLE of variance is the *population variance* of the data.

$$Y = \begin{cases} x_1 & \frac{1}{n} \\ x_2 & \frac{1}{n} \\ \vdots & \\ x_n & \frac{1}{n} \end{cases}$$

$$E(Y) = x_1\frac{1}{n} + x_2\frac{1}{n} + \cdots + x_n\frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Var(Y) = E\left((Y - E(Y))^2\right)$$

$$E(g(Y)) = \sum_{y \in \Omega_Y} g(y)P(Y=y)$$

$$= \sum_{i=1}^{n} \left(x_i - E(Y)\right)^2 \cdot \frac{1}{n}$$

$$= \sum_{i=1}^{n} \left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \cdot \frac{1}{n}.$$

Likelihood – Continuous Case

Definition. The **likelihood** of independent observations $x_1, \dots, x_n$ is

$$\mathcal{L}(X_1 = x_1, \dots, X_n = x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Normal outcomes $(X_1 = x_1, \dots, X_n = x_n)$

$$\hat{\theta}_{\mu}\ (X_1 = x_1, \dots, X_n = x_n) = \frac{\sum_i^n x_i}{n}$$

MLE estimator for
expectation

$$\hat{\theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_{\mu}\right)^2$$

MLE estimator for
variance

General Recipe (single parameter)

1. **Input** Given n i.i.d. samples $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ from parametric model with parameter θ.
2. **Likelihood** Define your likelihood $\mathcal{L}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta\right)$.
   - For discrete $\mathcal{L}(x_1, \ldots, x_n ; \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \ldots, x_n ; \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots, x_n ; \theta)$
4. **Differentiate** Compute $\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta)$
5. **Solve for** $\hat{\theta}$ by setting derivative to 0 and solving for max.

Do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.

General Recipe (multiple parrameters)

1. **Input** Given n i.i.d. samples $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ from parametric model with parameters $\vec{\theta} = (\theta_1, \ldots, \theta_k)$.
2. **Likelihood** Define your likelihood $\mathcal{L}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \vec{\theta}\right)$.
   - For discrete $\mathcal{L}(x_1, \ldots, x_n ; \vec{\theta}) = \prod_{i=1}^{n} P\left(x_i ; \vec{\theta}\right)$
   - For continuous $\mathcal{L}(x_1, \ldots, x_n ; \vec{\theta}) = \prod_{i=1}^{n} f\left(x_i ; \vec{\theta}\right)$
3. **Log** Compute $\ln\mathcal{L}\left(x_1, \ldots, x_n ; \vec{\theta}\right)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta_i} \ln \mathcal{L}\left(x_1, \ldots, x_n ; \vec{\theta}\right)$
5. **Solve for** $\hat{\theta}$ by setting all the derivatives to 0 and solving.

Check second order conditions, but we won't ask you to do that in CSE 312.