6.s077 — INTRODUCTION TO DATA SCIENCE
EECS, MIT, Spring 2018

Lecture 2

Estimating distributions and parameters (one-dimensional data)

Today's agenda

- Visualization

- Probability model/distribution $\mathbb{P}$
- Empirical distribution $\hat{\mathbb{P}}$

- Estimating a parameter
  - Plugin estimator
  - Feature matching (and method of moments)
  - Maximum likelihood (ML)
  - Empirical risk minimization
  - Bayesian methods

  } Later lectures

- Estimation of multiple parameters
- Representation matters

Data

- A collection of n records
- ith record: x_i
  - or (x_i, y_i, z_i)
  - e.g., x_i =height, y_i =weight, z_i =age

$$\begin{matrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{matrix}$$

- How do we...
  - visualize?
  - think about them? (Probability!)
  - get something out of them? (This class)

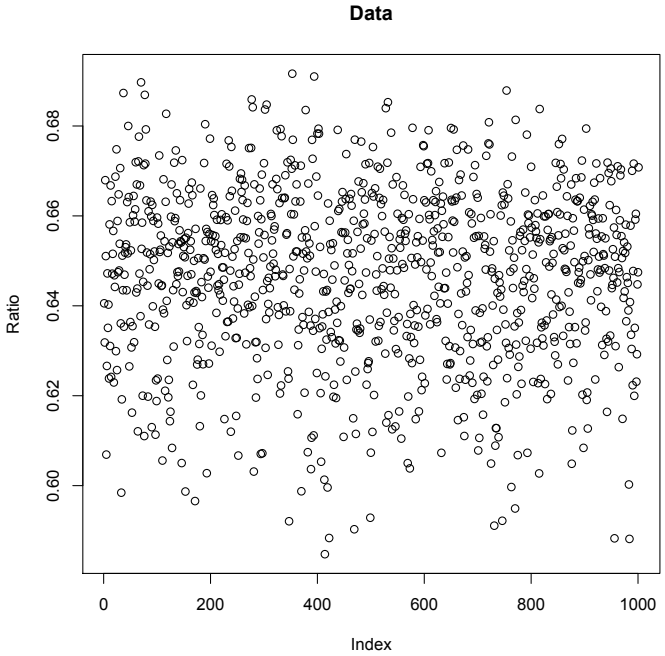
Visualizing 1-dimensional data

$$x = \frac{\text{forehead breadth}}{\text{body length}}$$

$n = 1000$

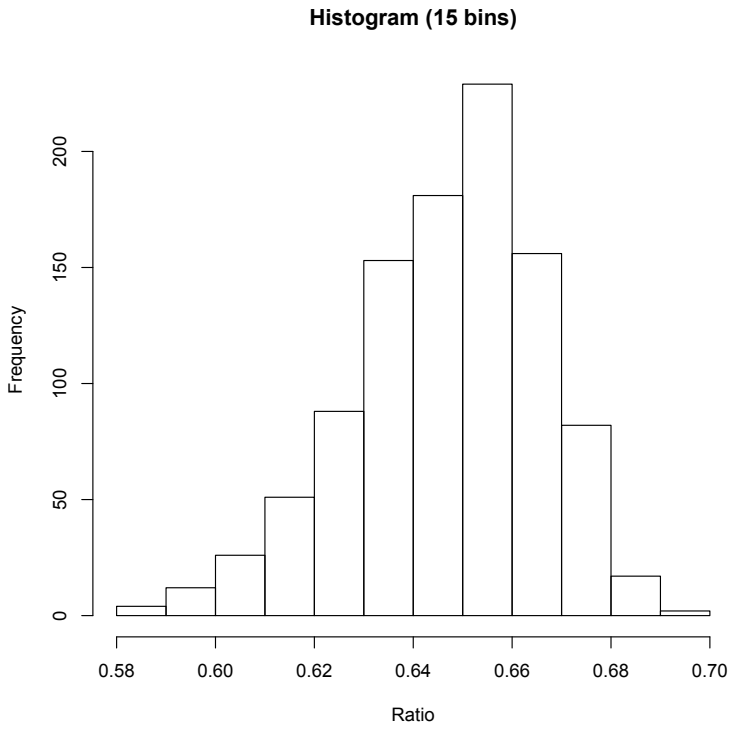
ratio	freq
0.5835	1
0.5875	3
0.5915	5
0.5955	2
0.5995	7
0.6035	10
0.6075	13
0.6115	19
0.6155	20
0.6195	25
0.6235	40
0.6275	31
0.6315	60
0.6355	62
0.6395	54
0.6435	74
0.6475	84
0.6515	86
0.6555	96
0.6595	85
0.6635	75
0.6675	47
0.6715	43
0.6755	24
0.6795	19
0.6835	9
0.6875	5
0.6915	1

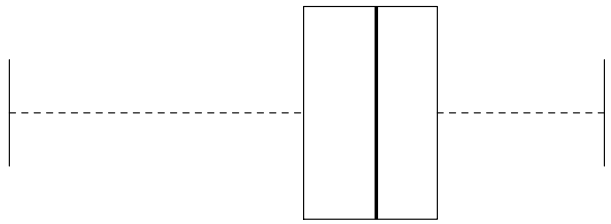
Naples crabs [W.F.R. Weldon; K. Pearson, 1894]



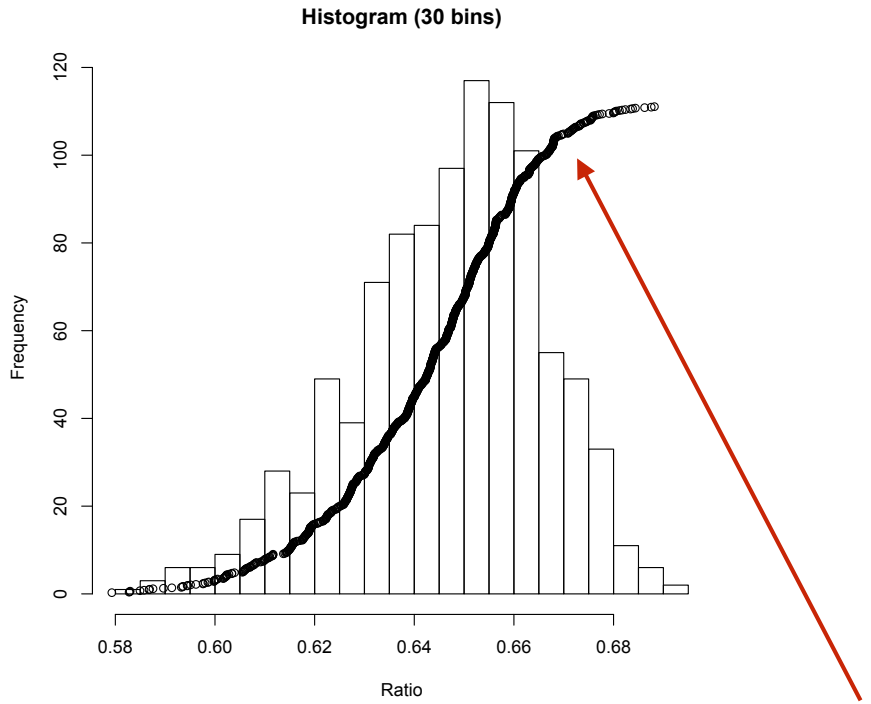
raw data ("scatter plot")

More insightful visualizations





box plot
(shows the quartiles)



Empirical CDF, $\widehat{\mathbb{F}}$

$$\widehat{\mathbb{F}}(\alpha) = \frac{\#i \text{ s.t. } x_i \leq \alpha}{n}$$

Numerical summaries of the data

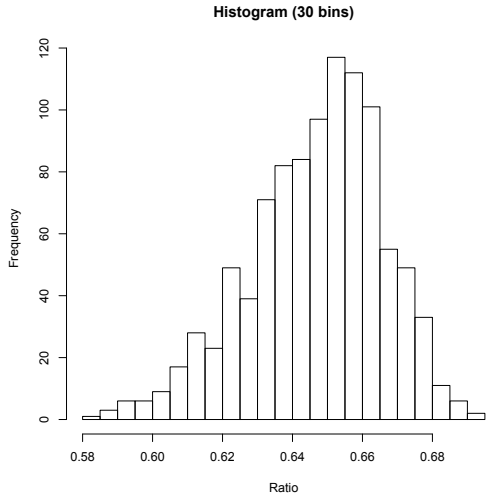
- Statistic: A real number that summarizes some aspect of the data

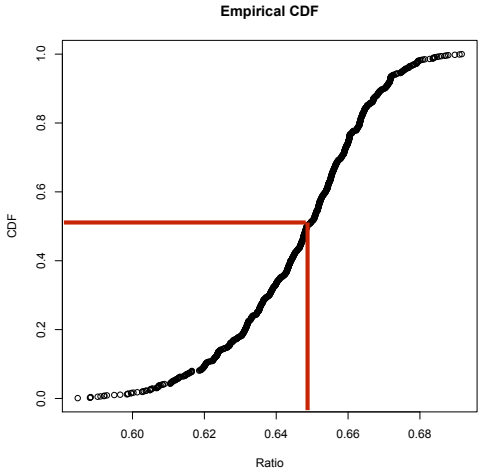
sample mean: $\overline{x} = \dfrac{x_1 + \cdots + x_n}{n}$

sample standard deviation: $\widehat{\sigma} = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

sample median:

$\widehat{\mathbb{F}}(\alpha) = \dfrac{1}{2}$





How to think about the data

- How were the data generated?

- Arbitrary/know nothing about the mechanism: can't do much

- According to a partially known mechanism
  - for example: $x_{k+1} = g(x_k, w_k)$

noise

estimate g, $\mathbb{P}$

model using probabilities $\mathbb{P}$

- Start simple: no dependence

Simplest data generation: independent random sampling

$\mathbb{P}$: discrete or continuous distribution on $\mathbb{R}$

$X, X_1, \ldots, X_n$: random variables

$X_i \sim \mathbb{P}$

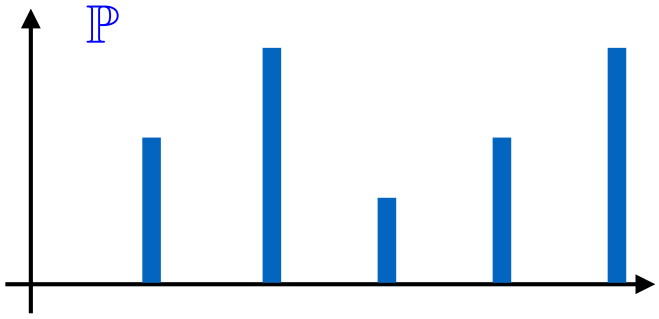
independent

} i.i.d.

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) = \prod_{i=1}^{n} \mathbb{P}(X \in A_i)$$

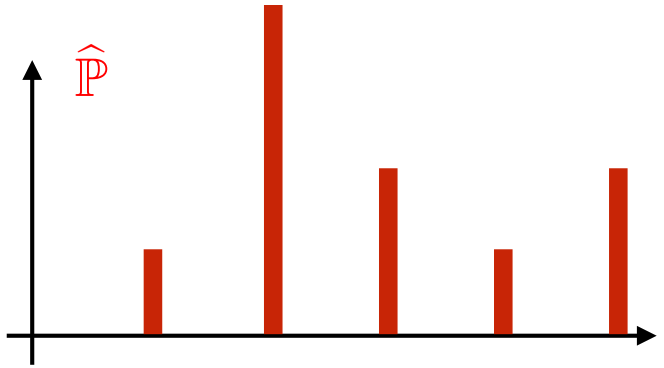
Independent random sampling

- **Example:** Finite population with 5 types of individuals
  - each person equally likely to be picked
  - independence: sampling "with replacement"



True population

CDF $\mathbb{F}$



Sample population of $n = 10$ people

$\widehat{\mathbb{P}}$ is a random "object"

"empirical distribution"

associated empirical CDF $\widehat{\mathbb{F}}$

- Continuous X: as if infinite underlying population

Central fact: as $n \to \infty$, we have $\widehat{\mathbb{P}} \to \mathbb{P}$
(see Lecture notes/Problem Set 1)

$\widehat{\mathbb{F}}(x) \to \mathbb{F}(x)$, for all x

Estimating parameters or numerical properties of $\mathbb{P}$

- Plugin estimator
- Feature matching (and method of moments)
- Maximum likelihood (ML)
- Empirical risk minimization
- Bayesian methods

} Later lectures

- Same/similar answers? Good
- If not, need ways to choose (performance analysis/assessment)

Plugin estimators

$\mathbb{E}$, $\widehat{\mathbb{E}}$: expectations w.r.t. $\mathbb{P}$, $\widehat{\mathbb{P}}$

$$\mu = \mathbb{E}[X] \qquad \widehat{M} = \frac{1}{n}\sum_{i=1}^{n} X_i = \widehat{\mathbb{E}}[X] \qquad \xrightarrow[n \to \infty]{} \mu \quad \text{(LLN)}$$

$$\gamma = \mathbb{E}[X^3] \qquad \widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} X_i^3 = \widehat{\mathbb{E}}[X^3]$$

$$\phi = \mathbb{E}\Big[g(X)\Big] \qquad \widehat{\Phi} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) = \widehat{\mathbb{E}}\Big[g(X)\Big]$$

$$v = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] \qquad \widehat{V} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{M})^2$$

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 \qquad \text{(if we know } \mu\text{)}$$

$a = \text{median}(\mathbb{P})$ $\qquad$ $\widehat{A} = \text{median}(\widehat{\mathbb{P}})$

$$a = h(\mathbb{P}) \qquad \widehat{A} = h(\widehat{\mathbb{P}})$$

Feature matching

- "Feature": a property of a distribution

  e.g.: mean, variance, median, 70th percentile, etc.

  $\mathbb{P} \longrightarrow h$ $\qquad\qquad$ $h = g(\mathbb{P})$

- Distribution determined by parameter θ: $\mathbb{P}^\theta$

  feature depends on the parameter: $h^\theta = g(\mathbb{P}^\theta)$

- Use data to calculate empirical value of the feature $\widehat{h} = g(\widehat{\mathbb{P}})$
- Find (analytically) h_θ as a function of θ
- Match the two: **solve, for θ, $h_\theta = \widehat{h}$**

- Example:
  - Find median $\widehat{h}$ of the data
  - Find θ such that the median of $\mathbb{P}^\theta$ agrees with $\widehat{h}$
- Food for thought:

  What would you do if you have one parameter

  but consider two (or more) features?

Maximum likelihood

- Assume X_i discrete, i.i.d.; $X = (X_1, \ldots, X_n)$

Likelihood function: $L^\theta(x_1, \ldots, x_n) = \mathbb{P}^\theta(X = x) = \prod_{i=1}^{n} \mathbb{P}^\theta(X_i = x_i)$

- Find θ under which the observed x is most likely to have been observed

having seen x: $\max_\theta L^\theta(x)$ $\max_\theta \sum_{i=1}^{n} \log\left(\mathbb{P}^\theta(X_i = x_i)\right)$

Sometimes analytically Usually numerically

- Example: $X \sim \text{Binomial}(n, \theta)$

$\max_\theta \binom{n}{x} \theta^x (1-\theta)^{1-x}$ algebra $\longrightarrow$ $\widehat{\theta} = \frac{x}{n}$

Maximum likelihood — continuous data

- X_i, independent PDF $\sim f^\theta_{X_i}(\cdot)$ $\qquad X = (X_1, \ldots, X_n)$

Likelihood function: $L^\theta(x_1, \ldots, x_n) = f^\theta_X(x) = \prod_{i=1}^n f^\theta_{X_i}(x_i)$

having seen x: $\max_\theta L^\theta(x)$ $\qquad \max_\theta \sum_{i=1}^n \log\left(f^\theta_{X_i}(x_i)\right)$

- Example: $X_i \sim \text{Exp}(\theta)$, $\quad f^\theta(x) = \theta e^{-\theta x}$, $x \geq 0$

What comes next?

- Understand the different methods
  - study their properties
  - do they have desirable properties?
  - understand their errors

- Generalize
  - Each X_i may be a vector [later]
  - Parameter θ may be a vector [now: straightforward]

The case of a vector of parameters — Plugin and ML

- Examples: $X \sim N(\mu, v)$ $\theta = (\mu, v)$

  unfair die, $X \in \{1, \ldots, 6\}$ $\theta = (p_1, \ldots, p_6)$

- **Plugin:** Apply to each component of θ separately

- **ML:** $\max_\theta L^\theta(x)$ $\nabla_\theta \log L^\theta(x) = 0$

  computation can be hard

Vector of parameters — Feature matching

- For concreteness: consider two parameters $\theta = (\theta_1, \theta_2)$
  - will use two features $h_i = g_i(\mathbb{P}), \quad i = 1, 2$

    e.g., $h_1 = \text{mean} \quad h_2 = \text{median}$

$$\left.\begin{aligned} h_1^\theta &= g_1(\mathbb{P}^\theta) \\ h_2^\theta &= g_2(\mathbb{P}^\theta) \end{aligned}\right\} \text{ find expressions, as functions of } \theta$$

$$\text{data} \longrightarrow \widehat{\mathbb{P}} \longrightarrow \begin{cases} \widehat{h}_1 = g_1(\widehat{\mathbb{P}}) \\ \widehat{h}_2 = g_2(\widehat{\mathbb{P}}) \end{cases}$$

$$\text{Solve } 2 \times 2 \text{ system of equations} \quad \begin{cases} h_1^\theta = \widehat{h}_1 \\ h_2^\theta = \widehat{h}_2 \end{cases}$$

Vector of parameters — The method of moments

- Suppose θ involves k parameters
- Use $\mathbb{E}[X], \ldots, \mathbb{E}[X^k]$ as your k features
- Find θ such that the first k moments of $\mathbb{P}^\theta$ match the first k empirical moments $\widehat{\mathbb{E}}[X], \ldots, \widehat{\mathbb{E}}[X^k]$

- Crab example: Pearson (1894)
  - fit weighted average ("mixture") of two normal PDFs
  - $k = 5$, by hand!

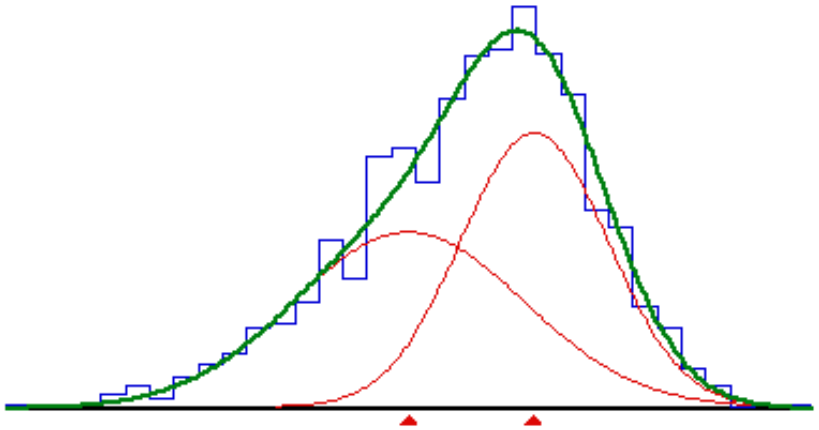


figure credit: Peter Macdonald

Representation matters

- **Data:** X has the same information as $Y = X^3$
  - Use X or Y?

- **Parameters:**

  variance v has same information as standard deviation: $v = \sigma^2$
  - Estimate v or σ?

- Does it make a difference?
  - in general, it does!
  - for ML it does not!