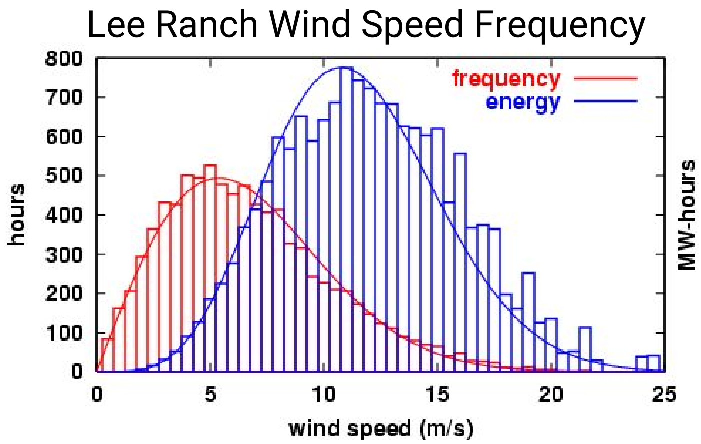
Distribution and Parameter Estimation

Objectives

- What is estimation and why do it?
- Review:
  - probability distributions
    - pmf/pdf and distribution parameters
- Parametric estimation of distribution and its parameters
  - Method of Moments (MOM)
  - Maximum Likelihood Estimation (MLE)
  - Maximum A Posteriori (MAP)
- Non-parametric estimation
  - Kernel Density Estimation

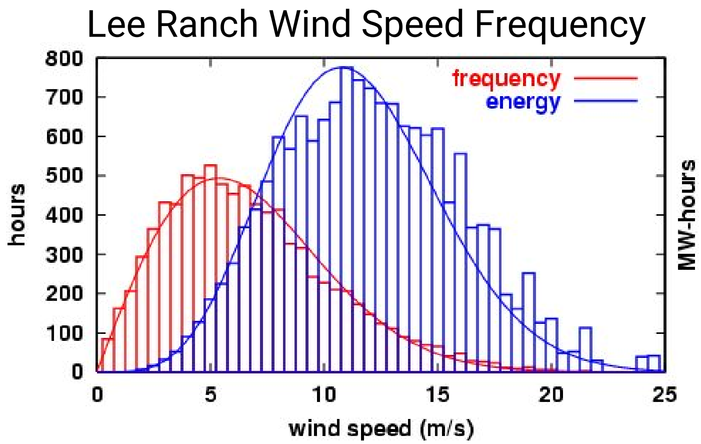
Estimation introduction

- One of the major applications of statistics is estimating *population parameters* from *sample statistics*.
  - What is the mean height of women aged 20-64 in Europe? Are you going to ask them all? Or can you work with the mean of a sample of 1000 women?
- By using probability distributions, and the few parameters needed to describe them, you can potentially make a simple model that describes the observed sample data.
  - **sample_data = model + residuals**



Estimation introduction

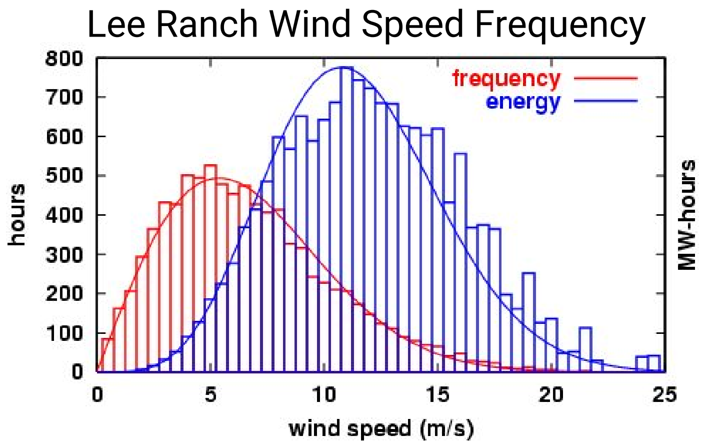
- One of the major applications of statistics is estimating *population parameters* from *sample statistics*.
  - What is the mean height of women aged 20-64 in Europe? Are you going to ask them all? Or can you work with the mean of a sample of 1000 women?
- By using probability distributions, and the few parameters needed to describe them, you can potentially make a simple model that describes the observed sample data.
  - **sample_data = model + residuals**



sample data?
model?
residuals?

Estimation introduction

- One of the major applications of statistics is estimating *population parameters* from *sample statistics*.
  - What is the mean height of women aged 20-64 in Europe? Are you going to ask them all? Or can you work with the mean of a sample of 1000 women?
- By using probability distributions, and the few parameters needed to describe them, you can potentially make a simple model that describes the observed sample data.
  - **sample_data = model + residuals**



sample data: the observed # of hours at the given (binned) wind speeds
model: the Rayleigh distribution with differing values of the σ parameter
residuals: the differences between the model and the measured data

Review: Distributions (partial list)

Parameters

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 0, 1, 2...$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Discrete

- X ~ Bernoulli(p) = Single coin flip turns out to be Heads
- X ~ Binomial(100, p) = # of coin flips out of 100 that turn out to be Heads
- X ~ Geometric(p) = # of Trials until coin flip turns out to be Heads
- X ~ Poisson(λ=10) = # of taxis passing a street corner in a given hour (on avg 10/hr)

Continuous

- X ~ Exponential(λ=10) = Time until taxi will pass street corner
- X ~ Uniform(0,360) = Degrees between hour hand and minute hand
- X ~ Gaussian(100, 10) = IQ Score

Review: Distributions (partial list)

Parameters

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \le k \le n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 0, 1, 2 \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \ge 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

What defines a distribution?

Discrete

- X ~ Bernoulli(p) = Single coin flip turns out to be Heads
- X ~ Binomial(100, p) = # of coin flips out of 100 that turn out to be Heads
- X ~ Geometric(p) = # of Trials until coin flip turns out to be Heads
- X ~ Poisson(λ=10) = # of taxis passing a street corner in a given hour (on avg 10/hr)

Continuous

- X ~ Exponential(λ=10) = Time until taxi will pass street corner
- X ~ Uniform(0,360) = Degrees between hour hand and minute hand
- X ~ Gaussian(100, 10) = IQ Score

Review: Distributions (partial list)

Parameters

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k(1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 0, 1, 2...$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

What defines a distribution:
the PDF/PMF (more in a bit) and the parameters

Discrete

- X ~ Bernoulli(p) = Single coin flip turns out to be Heads
- X ~ Binomial(100, p) = # of coin flips out of 100 that turn out to be Heads
- X ~ Geometric(p) = # of Trials until coin flip turns out to be Heads
- X ~ Poisson(λ=10) = # of taxis passing a street corner in a given hour (on avg 10/hr)

Continuous

- X ~ Exponential(λ=10) = Time until taxi will pass street corner
- X ~ Uniform(0,360) = Degrees between hour hand and minute hand
- X ~ Gaussian(100, 10) = IQ Score

Review: PMF - Probability Mass Function

A probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.

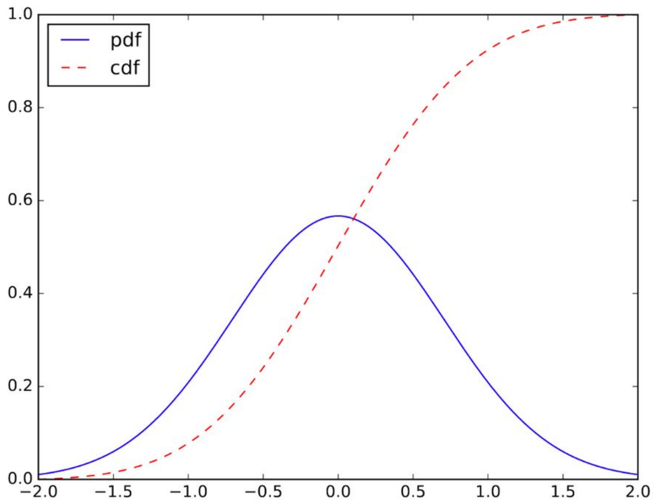
$$\sum_{s \in S} P(X = s) = 1$$

$$P(X = s)$$

Review: PDF - Probability Density Function

A probability density function can be interpreted as the relative likelihood that the value of a random variable would equal the the indicated value.

The probability is defined by integrating between the two limits of independent variable.



x

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$pdf = f(x)$$

$$P(x_1 \leq X \leq x_2) = \int_{x=x_1}^{x_2} f(x)dx$$

Distribution and parameter estimation (parametric)

After visually inspecting the data (usually a histogram) we pick a distribution. Then we can use one of three methods to estimate the parameters of the distribution:

- Method of Moments (MOM)
- Maximum Likelihood Estimation (MLE)
- Maximum a Posteriori (MAP)

Parametric Method #1: Method of Moments

1) Assume a distribution (e.g. Poisson, Bernoulli, Binomial, Gaussian)
2) Compute the relevant *moments* from the data (e.g. sample mean and variance)
3) Use the moments to calculate the appropriate parameters of the distribution, then plot it and see if it makes sense.

Method of Moments - Example #1

Your website visitor log shows the following number of visits for each of the last seven days: [6, 4, 7, 4, 9, 3, 5].

What’s the probability of zero visitors tomorrow?

What distribution should we assume?

What parameter are we estimating? How?

Method of Moments - Example #1

Your website visitor log shows the following number of visits for each of the last seven days: [6, 4, 7, 4, 9, 3, 5].

What's the probability of zero visitors tomorrow?

What distribution should we assume? Poisson

What parameter are we estimating? λ How? Use the first moment (the mean) of the data to estimate λ.

Method of Moments - Example #1

See **Estimation_MOM_Example_1.ipynb**

Method of Moments - Example #2

You flip a coin 100 times. It comes up heads 52 times.

What's the MOM estimate that in the next 100 flips the coin will be heads <= 45 times?

What's the probability of zero visitors tomorrow?

Distribution? Parameter to estimate? How?

Please write a python script and plot the distribution like Example #1.

Method of Moments - Example #2

See **Estimation_MOM_Example_2.ipynb**

Parametric Method #2: Maximum Likelihood Estimation

Likelihood of what? The sample data, given the distribution parameter(s).

Law of Likelihood:
If $P(X|\theta_1) > P(X|\theta_2)$, then the evidence supports θ_1 over θ_2.

General idea - calculate the likelihood of the data for a range of parameter value(s), and then pick the parameter that gives the maximum likelihood of the data. (I'll give a concrete example).

Parametric Method #2:
Maximum Likelihood Estimation

1) Assume a distribution (e.g. Poisson, Bernoulli, Binomial, Gaussian)
2) Define the likelihood function
3) Choose the parameter(s) that maximize the likelihood function.

MLE - in detail

Assume our data is drawn independently from some distribution, and we'd like to estimate the parameters for that distribution. The probability distribution function for the data is:

$$f(x|\theta)$$

where:

x Is the data

θ Are the parameters for a given distribution that we are trying to estimate

MLE - in detail

Since the draws are independent the *joint density function* can be defined:

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) * f(x_2|\theta) * \cdots * f(x_n|\theta)$$

Just for review:

$$f(x_1|\theta) * f(x_2|\theta) * \cdots * f(x_n|\theta) \;= \prod f(x_i|\theta)$$

Likelihood:

Likelihood ➡ $f(x_1, x_2, \ldots, x_n|\theta) = \prod f(x_i|\theta)$

Looking for theta that maximizes the likelihood of getting the data ➡ $\hat{\theta}_{\mathrm{mle}} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n|\theta)$

MLE - Revisit Example #1, but with MLE

Your website visitor log shows the following number of visits for each of the last seven days: [6, 4, 7, 4, 9, 3, 5].

What's the probability of zero visitors tomorrow?

Assume Poisson, so need to estimate λ.

Using MLE instead of MOM.

MLE - Example #1, brute force

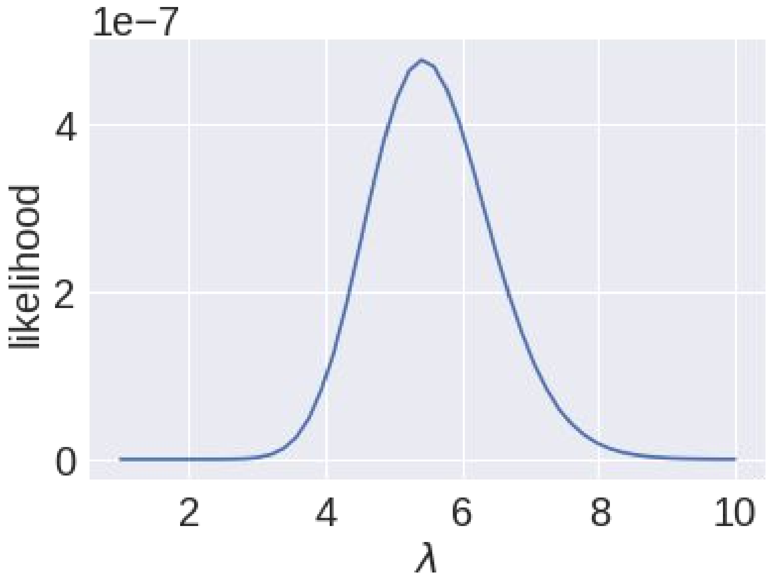
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

visits: [6, 4, 7, 4, 9, 3, 5]

$$L(\lambda) = P(X=6) \cdot P(X=4) \cdot P(X=7) \cdot P(X=4) \cdot P(X=9) \cdot P(X=3) \cdot P(X=5)$$

$$L(\lambda) = \frac{\lambda^6 e^{-\lambda}}{6!} \cdot \frac{\lambda^4 e^{-\lambda}}{4!} \cdot \frac{\lambda^7 e^{-\lambda}}{7!} \cdot \frac{\lambda^4 e^{-\lambda}}{4!} \cdot \frac{\lambda^9 e^{-\lambda}}{9!} \cdot \frac{\lambda^3 e^{-\lambda}}{3!} \cdot \frac{\lambda^5 e^{-\lambda}}{5!}$$

$$L(\lambda) = \frac{\lambda^{38} e^{-7\lambda}}{6! \cdot 4! \cdot 7! \cdot 4! \cdot 9! \cdot 3! \cdot 5!}$$



MLE - Example #1, brute force

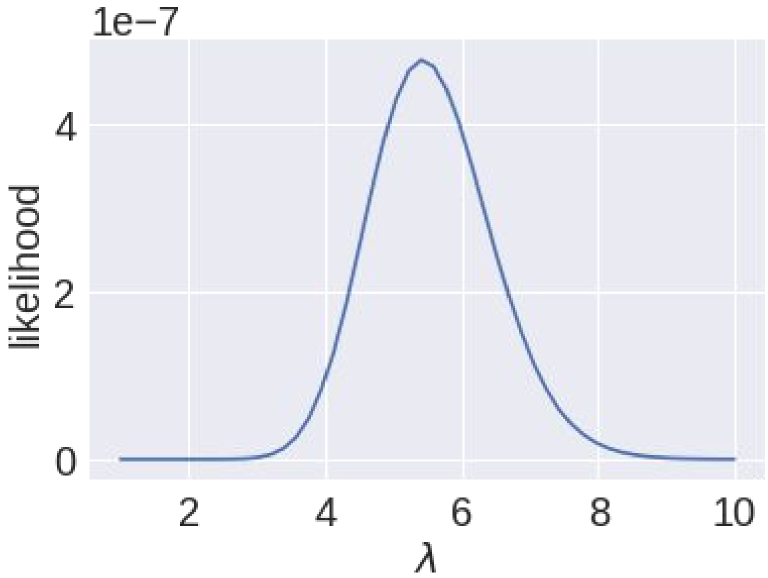
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

visits: [6, 4, 7, 4, 9, 3, 5]

$$L(\lambda) = P(X=6) \cdot P(X=4) \cdot P(X=7) \cdot P(X=4) \cdot P(X=9) \cdot P(X=3) \cdot P(X=5)$$

$$L(\lambda) = \frac{\lambda^6 e^{-\lambda}}{6!} \cdot \frac{\lambda^4 e^{-\lambda}}{4!} \cdot \frac{\lambda^7 e^{-\lambda}}{7!} \cdot \frac{\lambda^4 e^{-\lambda}}{4!} \cdot \frac{\lambda^9 e^{-\lambda}}{9!} \cdot \frac{\lambda^3 e^{-\lambda}}{3!} \cdot \frac{\lambda^5 e^{-\lambda}}{5!}$$

$$L(\lambda) = \frac{\lambda^{38} e^{-7\lambda}}{6! \cdot 4! \cdot 7! \cdot 4! \cdot 9! \cdot 3! \cdot 5!}$$



See **Brute_force_max_likelihood.ipynb**

MLE - Example #1, log & calculus are your friends

visits: [6, 4, 7, 4, 9, 3, 5]

$$L(\lambda) = \frac{\lambda^{38} e^{-7\lambda}}{6! \cdot 4! \cdot 7! \cdot 4! \cdot 9! \cdot 3! \cdot 5!}$$

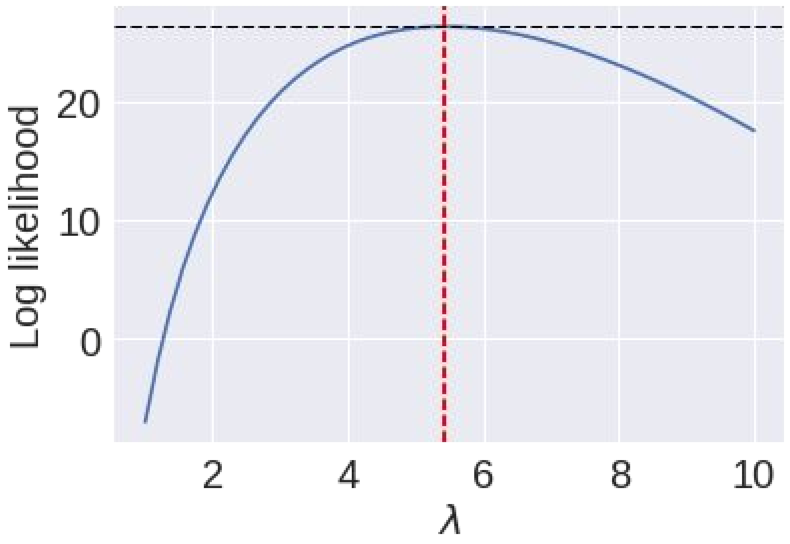
$$L(\lambda) \propto \lambda^{38} e^{-7\lambda}$$

$$lnL(\lambda) \propto 38 \cdot \ln(\lambda) - 7\lambda$$

$$\frac{d}{d\lambda}\big(lnL(\lambda)\big) = \frac{38}{\lambda} - 7 = 0$$

$$\lambda = \frac{38}{7} = 5.429$$

$$\frac{d^2}{d\lambda}\big(lnL(\lambda)\big) = -\frac{38}{\lambda^2}$$



MLE - Derivation for Binomial Distribution

$$X_i \overset{iid}{\sim} Bin(n, p) \qquad i = 1, 2, \ldots, n \qquad f(x_i|p) = \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}$$

$$log\mathcal{L}(p) = \sum_{i=1}^{n}\left[log\binom{n}{x_i} + x_i logp + (n - x_i)log(1-p)\right]$$

$$\frac{\partial log\mathcal{L}(p)}{\partial p} = \sum_{i=1}^{n}\left[\frac{x_i}{p} - \frac{n - x_i}{1-p}\right] = 0$$

$$\hat{p}_{MLE} = \boxed{\frac{\bar{x}}{n}}$$

For the Binomial distribution, MOM and MLE give the same answer!

Parametric Method #3: Maximum a Posteriori (MAP)

Similar to MLE, but reverse.

MLE finds θ to maximize:

$$f(x_1, x_2, \ldots, x_n | \theta)$$

MAP finds θ to maximize:

$$f(\theta | x_1, x_2, \ldots, x_n)$$

Parametric Method #3: Maximum a Posteriori (MAP)

MLE $\rightarrow$ MAP, just use Bayes' Theorem

$$f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_{\theta' \in \Theta} f(x|\theta')g(\theta')\,d\theta'} \propto \boxed{f(x|\theta)}g(\theta)$$

MLE is just this part.

Parametric Method #3: Maximum a Posteriori (MAP)

MLE solves:

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n | \theta)$$

MAP includes the prior belief.

MAP solves:

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \ldots, x_n | \theta) \boxed{g(\theta)}$$

What if all θ are equally likely?

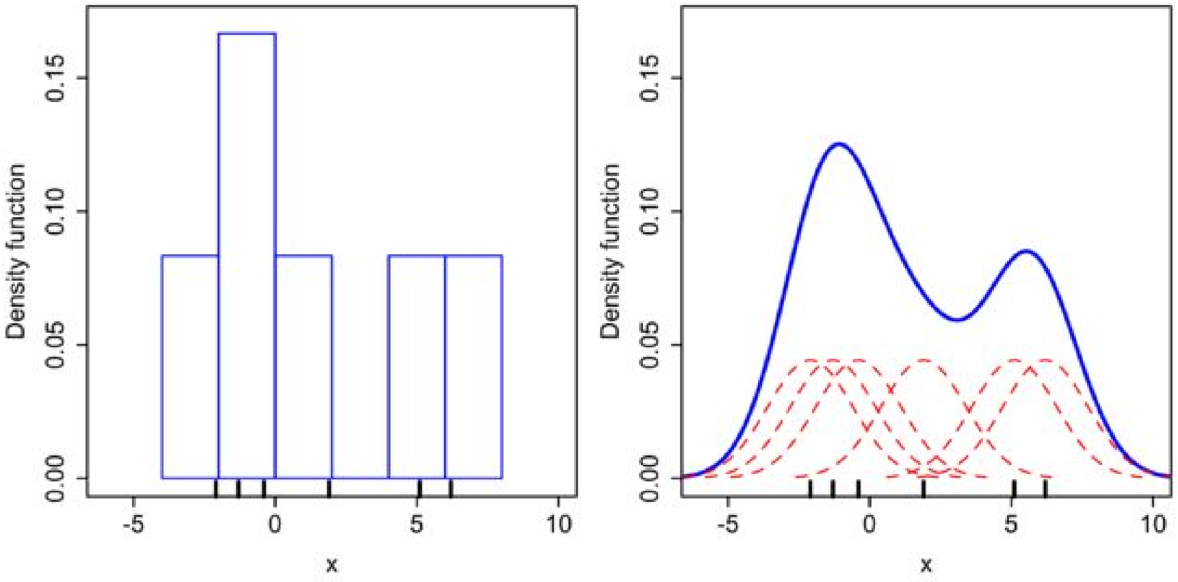
Non-parametric method:
Kernel Density Estimation (KDE)

Question: How can we model data that does not follow a known distribution?

Answer: Use a nonparametric technique.

Non-parametric method:
Kernel Density Estimation (KDE)

KDE is a nonparametric way to estimate the PDF of a random variable. KDE smooths the histogram by summing “kernel functions” (usually Gaussians) instead of binning into rectangles.

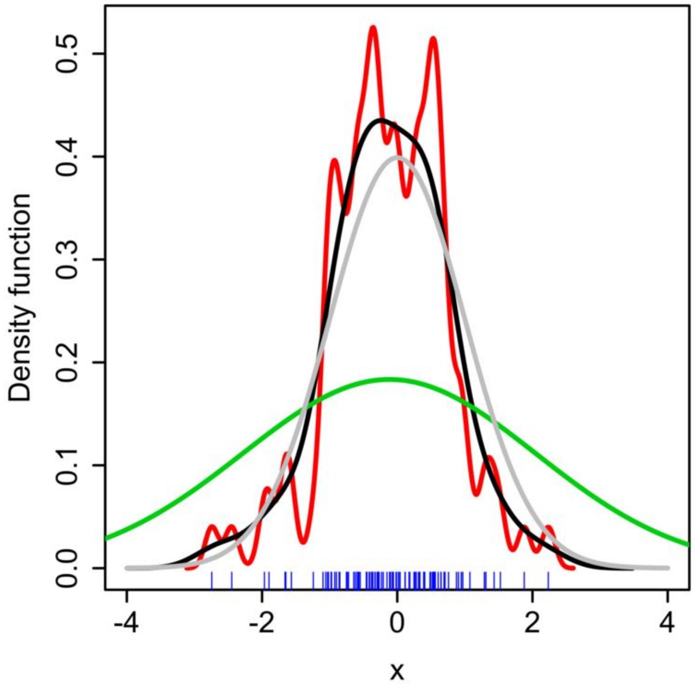


Non-parametric method:
Kernel Density Estimation (KDE)

Kernel functions have a *bandwidth* parameter to control under- and over-fitting.

Each curve on the right shows an estimated PDF with different bandwidths.

See Adam's **sampling-estimation.ipynb** in the repo.



Parametric vs. Non-parametric

Parametric methods:

- Based on assumptions about the distribution of the underlying population and the parameters from which the sample was taken.
- If the data deviates strongly from the assumptions, could lead to incorrect conclusions.

Nonparametric methods:

- NOT based on assumptions about the distribution of the underlying population.
- Generally not as powerful -- less inference can be drawn.
- Interpretation can be difficult… what does the wiggly curve mean?