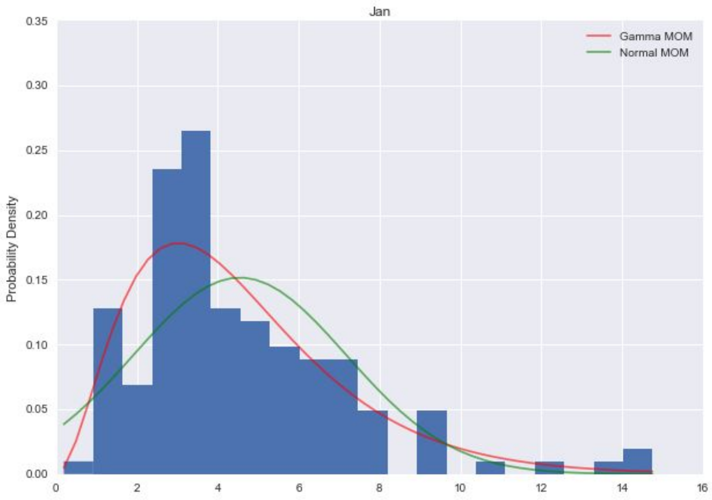
Estimation & Sampling

Estimation

DSI SEA5, jf.omhover, Sep 14 2016

Drawing on the great work of Ryan Henning



Estimation & Sampling

Estimation

DSI SEA5, jf.omhover, Sep 14 2016

Drawing on the great work of Ryan Henning

STANDARDS

- **Compute** MLE estimate for simple example (such as coin-flipping)
- **Compare** and contrast the use cases of parametric and nonparametric estimation

Estimation & Sampling

Estimation

DSI SEA5, jf.omhover, Sep 14 2016

Drawing on the great work of Ryan Henning

OBJECTIVES

- **Relate** estimation to modeling and machine learning
- **Identify** cases and conditions in which to use MOM, MLE and KDE
- **Use** MLE to estimate a parametric distribution from observed data
- **Use** the MOM to estimate a parametric distribution from observed data
- **Understand** how KDE estimates a non-parametric distribution from observed data

Required “Actionable” Concepts

Discrete / continuous

Expected Value

Moments: mean, variance...

PMF / PDF

CDF

i.i.d

What's the Big Idea ?

Reality VS Model

REALITY

MODEL

	Year	Jan	Feb	Mar
0	1871	2.76	4.58	5.01
1	1872	2.32	2.11	3.14
2	1873	2.96	7.14	4.11
3	1874	5.22	9.23	5.36
4	1875	6.15	3.06	8.14
5	1876	6.41	2.22	5.28
6	1877	4.05	1.06	4.98

data

FUNCTIONS:
descriptive
predictive
normative
...

$$f(x,\alpha,\theta) = x^{\alpha-1}\frac{e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}$$

mathematical
formulas

Relevance in a business context

Example 1 (descriptive) : You want to benchmark/compare several local branches of your company on their day to day operational indicators.

Example 2 (predictive) : You have data on the rain falling in Nashville for a century and you want to evaluate how much water you could expect in a month for your garden.

Example 3 (normative) : You have one company’s accounting/expenses data and you want to find potential data manipulation / corruption.

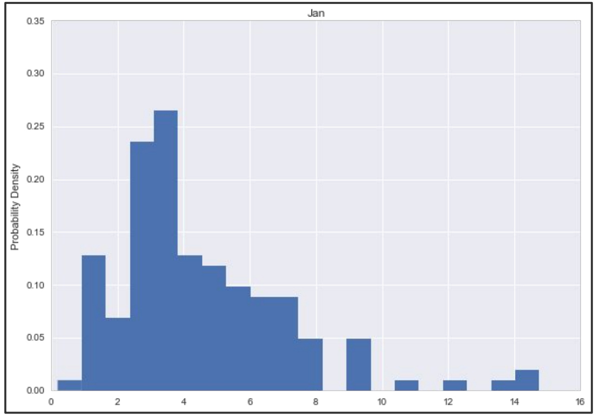
From reality to model : general process

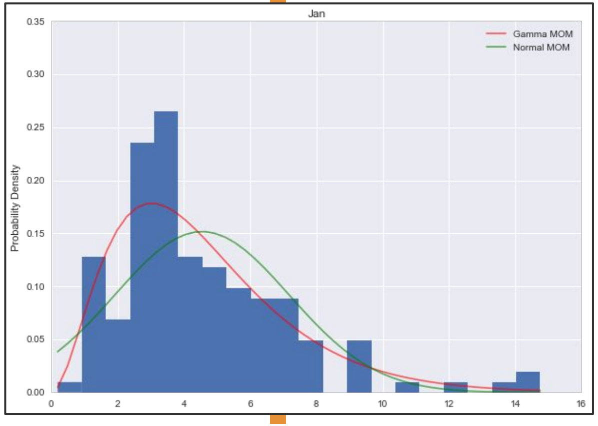
REALITY

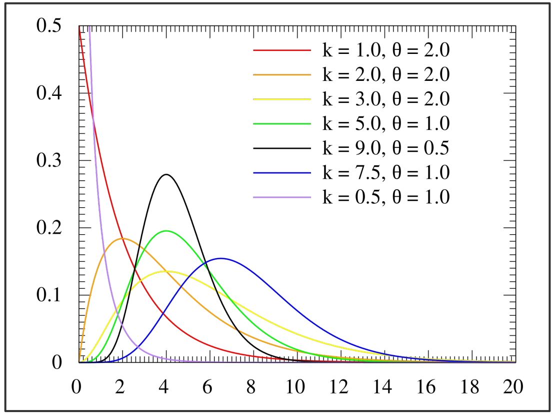
1) Having a data sample
Observing an underlying behavior

MODEL

2) Make an assumption
on the model underlying the data







3) **Find** the instance of the model
that **fits** with data sample

Gamma distribution [wikipedia]

Multiple realities / distribution models

REALITY

MODEL

	Year	Jan	Feb	Mar
0	1871	2.76	4.58	5.01
1	1872	2.32	2.11	3.14
2	1873	2.96	7.14	4.11
3	1874	5.22	9.23	5.36
4	1875	6.15	3.06	8.14
5	1876	6.41	2.22	5.28
6	1877	4.05	1.06	4.98

data

Identify
assumptions

Estimate
if assumption
is likely

Bernoulli
Binomial
Poisson
...

Normal
Gamma
Chi square
...

Data can be...
Qualitative : categories, names, semantics…
Quantitative : measures, metrics, KPI...

Models can be...
discrete / continuous
parametric / non-parametric

Framing the problem : parameter estimation

REALITY

MODEL

	Year	Jan	Feb	Mar
0	1871	2.76	4.58	5.01
1	1872	2.32	2.11	3.14
2	1873	2.96	7.14	4.11
3	1874	5.22	9.23	5.36
4	1875	6.15	3.06	8.14
5	1876	6.41	2.22	5.28
6	1877	4.05	1.06	4.98

data

SOLUTION TO PB 2:
implement a process based on data for finding the best parameter values

$$\hat{\alpha}, \hat{\beta}$$

parametric distribution model "instance"

PROBLEM 2: find actual parameter values

$$f(x, \alpha, \beta)$$

parametric distribution model "class"

PROBLEM 1: identify model class

PDF/PMF model

Methods of Moments (MOM)

Drawing on the great work of Ryan Henning

Solving the problem MOM style

REALITY

MODEL

	Year	Jan	Feb	Mar
0	1871	2.76	4.58	5.01
1	1872	2.32	2.11	3.14
2	1873	2.96	7.14	4.11
3	1874	5.22	9.23	5.36
4	1875	6.15	3.06	8.14
5	1876	6.41	2.22	5.28
6	1877	4.05	1.06	4.98

data

$$\mu, \sigma^2$$

2) COMPUTE relevant sample moments
mean, variance...

$$\hat{\alpha}, \hat{\beta}$$

3) Inject moments into the PMF/PDF of the assumed distribution

$$f(x, \alpha, \beta)$$

1) Assume an underlying distribution for your domain

Example 1: Method of Moments

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?

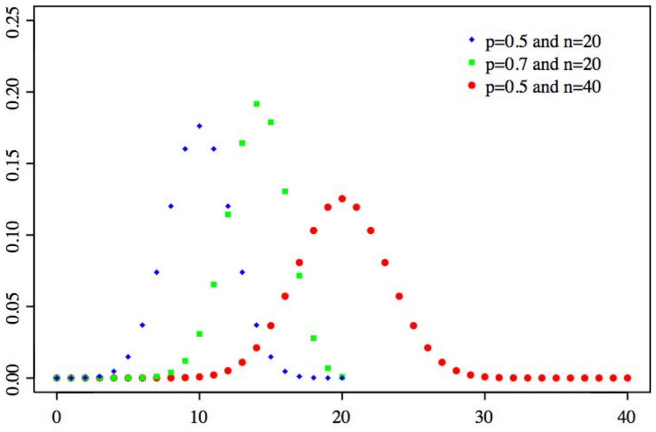
Which underlying distribution should we assume?

Let's draw the "Binomial Card" !

$$X \sim B(n, p)$$

PMF (DISCRETE)

$$f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

PARAMETERS

n: integer

p: probability

USE CASES

Drawing a coin n times, counting heads

MOM $\mathrm{E}[X] = np,$

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?

Which underlying distribution should we assume?

Binomial… note: We really only have one binomial sample here.

What moment should we estimate?

The mean. We actually only have one sample here where the result is 52. So the mean is 52.

Example 1 (con't): Method of Moments

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?

From our one binomial sample, we know:

$$\bar{x} = 52$$

The binomial distribution has mean:

$$\mu = np$$

What does MOM say to do next?

Use the sample moments to estimate the distribution parameters. In this case, the parameter we need to estimate is *p*.

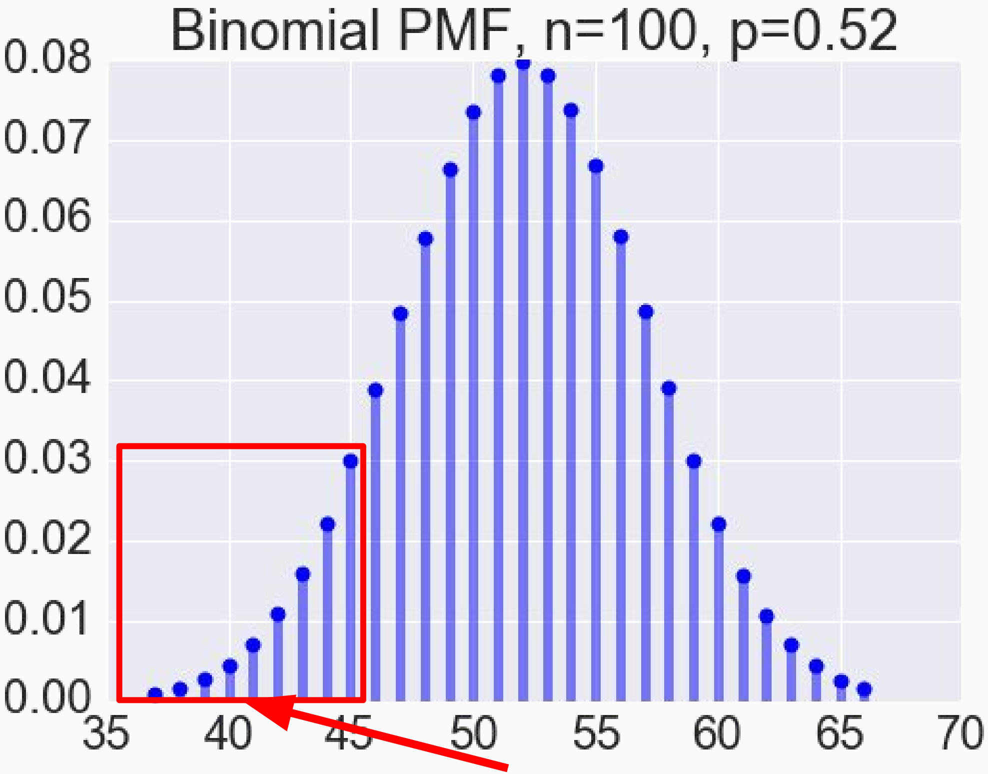
Example 1 (con't): Method of Moments

$$52 = np$$

$$\boxed{n = 100}$$

$$p = 52/100$$

$$\boxed{p = 0.52}$$



Probability in the next 100 flips the coin will be heads <= 45 times:

```
print scipy.stats.binom.cdf(45, n, p)
```

⇒ 0.096653350327

Example 2: Method of Moments

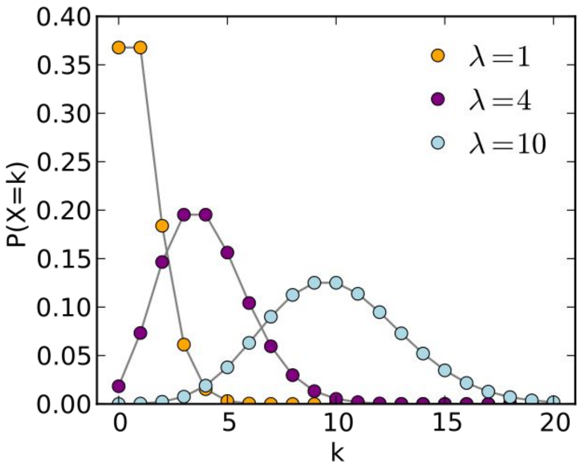
Your website visitor log shows the following number of visits for each of the last seven days: [6, 4, 7, 4, 9, 3, 5]. What’s the probability of zero visitors tomorrow?

Which underlying distribution should we assume?

Let’s draw the “Poisson Card” !

PMF (DISCRETE)

$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$



PARAMETERS

lambda: float value

USE CASES

Counting visitors of a web site

MOM $\lambda = \mathrm{E}(X) = \mathrm{Var}(X).$

Example 2: Method of Moments

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?

Which underlying distribution should we assume?

Poisson! Let’s look at Wikipedia to remind ourselves what it is. :)

What moment should we estimate?

The mean. Our mean estimate will become the estimate for the only parameter used in the Poisson distribution: λ

Example 2 (con't): Method of Moments

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?

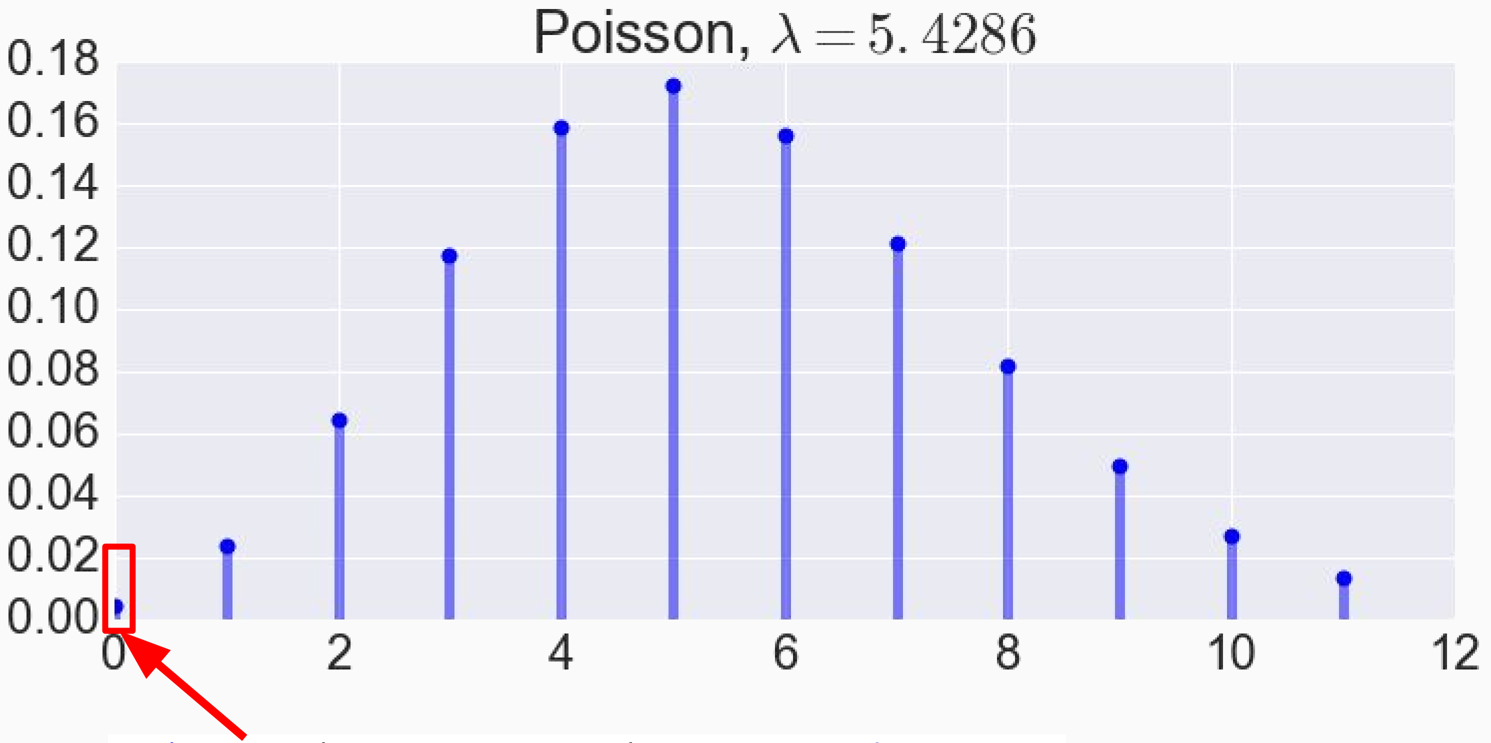
From our samples, estimate the mean:

$$\bar{x} = (6 + 4 + 7 + 4 + 9 + 3 + 5)/7 = 5.43$$

Our mean estimate is used to estimate λ:

$$\lambda = 5.43$$

Example 2 (con't): Method of Moments




```
print scipy.stats.poisson.pmf(0, lmda)
```

```
0.00438936184278
```

Solving the problem MOM style

REALITY

MODEL

	Year	Jan	Feb	Mar
0	1871	2.76	4.58	5.01
1	1872	2.32	2.11	3.14
2	1873	2.96	7.14	4.11
3	1874	5.22	9.23	5.36
4	1875	6.15	3.06	8.14
5	1876	6.41	2.22	5.28
6	1877	4.05	1.06	4.98

data

$$\rightarrow \mu, \sigma^2 \longrightarrow \hat{\alpha}, \hat{\beta} \qquad f(x, \alpha, \beta)$$

Depending on your assumption on the parametric distribution model, you'll compute specific moments to compute the parameters of the distribution model.

Maximum Likelihood Estimation (MLE)

Drawing on the great work of Ryan Henning

Solving the problem MLE style

REALITY

MODEL

DEFINE A CRITERIA
OPTIMIZE IT ON
PARAMETERS

$\hat{\alpha}, \hat{\beta}$

$f(x, \alpha, \beta)$

	Year	Jan	Feb	Mar
0	1871	2.76	4.58	5.01
1	1872	2.32	2.11	3.14
2	1873	2.96	7.14	4.11
3	1874	5.22	9.23	5.36
4	1875	6.15	3.06	8.14
5	1876	6.41	2.22	5.28
6	1877	4.05	1.06	4.98

data

$$X = \{x_1 x_2 \ldots x_n\}$$

What is the probability of observing this data, knowing that it was drawn from a distribution with known parameters ?

$$P(x_1, x_2, \ldots, x_n | \alpha, \beta)$$

likelihood function

What is the best couple or parameters for maximizing that ?

Maximum Likelihood Estimation (MLE)

Overview:

Law of Likelihood:

If *P(X|H1) > P(X|H2)*, then the evidence supports *H1* over *H2*.

Question:

Which hypothesis does the evidence most strongly support?

Answer:

The hypothesis H that maximizes *P(X|H)*, which is found via *MLE*.

Maximum Likelihood Estimation (MLE)

Overview:

1. **Assume an underlying distribution for your domain.** (just like with MOM)
   E.g. Poisson, Bernoulli, Binomial, Gaussian

2. **Define the likelihood function.**
   We want to know the likelihood of the data we observe under different distribution parameterizations.

3. **Choose the parameter set that maximizes the likelihood function.**

Tell me more about MLE...

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

True because we assume X is i.i.d.
Recall, what does i.i.d. Mean?
What's the i. part? What's the i.d. part?

$$\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

We will find θ to maximize the log-likelihood function:

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} \log\left(\mathcal{L}(\theta|x_1, \ldots, x_n)\right)$$

Whoa whoa… what? Let's talk about it.

Example 3: Maximum Likelihood Estimation

You flip a coin 100 times. It comes up heads 52 times. What's the MLE estimate that in the next 100 flips the coin will be heads <= 45 times?

Yep, same example...

Which underlying distribution should we assume?

Binomial… (like last time)

… now we need to define our likelihood function...

Example 3 (con't): Maximum Likelihood Estimation

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

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

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

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

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

Kernel Density Estimation (KDE)

Drawing on the great work of Ryan Henning

Solving the problem : non-parametric techniques

REALITY

	Year	Jan	Feb	Mar
0	1871	2.76	4.58	5.01
1	1872	2.32	2.11	3.14
2	1873	2.96	7.14	4.11
3	1874	5.22	9.23	5.36
4	1875	6.15	3.06	8.14
5	1876	6.41	2.22	5.28
6	1877	4.05	1.06	4.98

data

MODEL

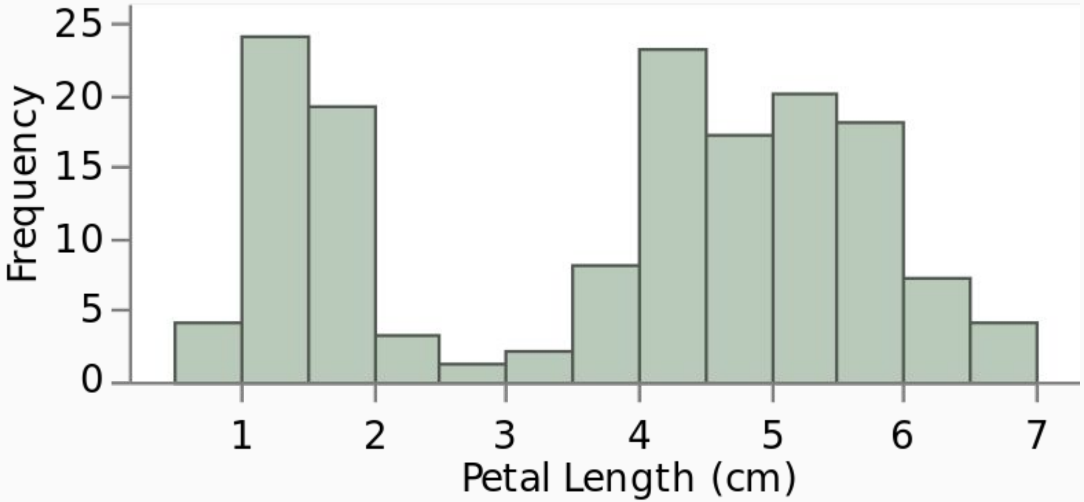
$$\cancel{f(x, \beta)}$$

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

Answer: Use a nonparametric technique.

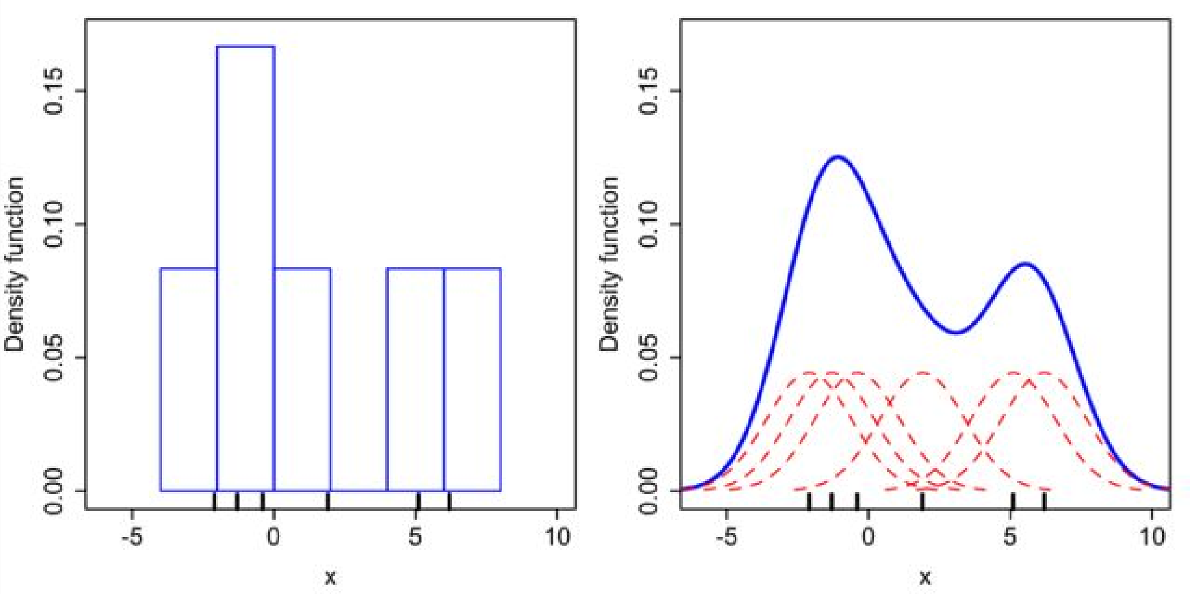
Histograms

A histogram groups continuous data into discrete intervals and displays relative frequencies. But it's not a smooth distribution. :(



Kernel Density Estimation (KDE)

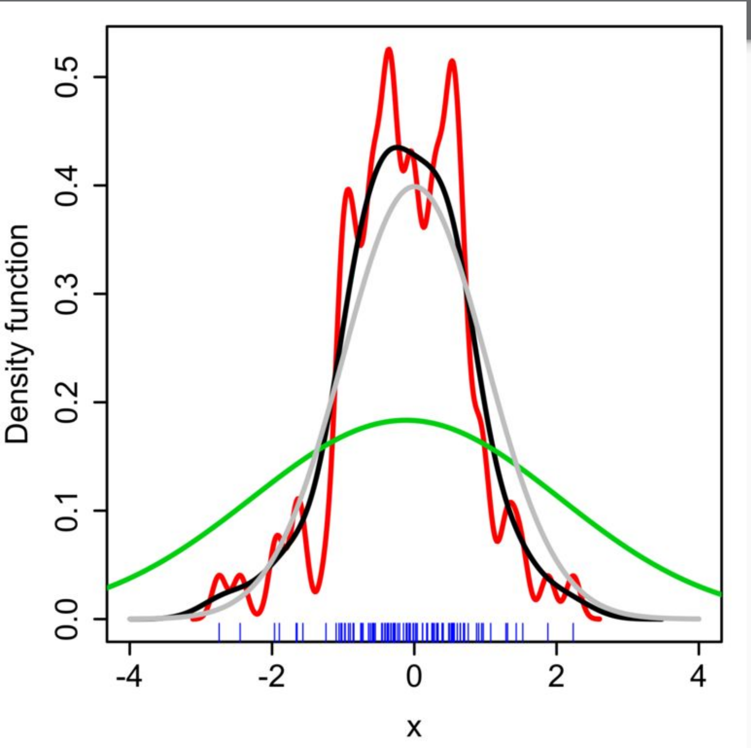
KDE is a nonparametric way to estimate the PDF of a random variable. KDE smooths the histogram by summing "kernel functions" (usually Gaussians) instead of binning into rectangles.



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.



Parametric vs Nonparametric Methods

Drawing on the great work of Ryan Henning

Estimating Distributions
Parametric vs Nonparametric Methods

Parametric: We assume an underlying distribution, then we use our data to estimate the parameters of that underlying distribution. E.g. Using:

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

Nonparametric: We don't assume any *single* underlying distribution, but instead we fit a combination of distributions to the observed data. E.g. Using:

- *Kernel Density Estimation* (KDE)

When to use each method? (Parametric vs Nonparametric)

Parametric methods:

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

Nonparametric methods:

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