Descriptive statistics

Jorge Lillo-Box

MCMC Coffee | Season 1, Episode 1

MCMC Coffee - Season 1

Introductory sessions

01/09/2016	Jorge Lillo-Box	Descriptive statistics	3.1 to 3.3
15/09/2016	Daniel Asmus	Central limit theorem + correlation coefficients	3.4 to 3.6
29/09/2016	Bruno Dias	MLE 1 (general idea, goodness of fit, confidence estimates)	4.2 to 4.5
06/10/2016		MLE 2 (hypothesis testing, model comparison, non-parametric analysis)	4.6 to 4.8
27/10/2016		Bayesian Inference 1 (Bayes theorem, priors)	5.1 to 5.2
10/11/2016		Bayesian Inference 2 (model selection)	5.3 to 5.5
24/11/2016		MCMC methods (sampling the posterior distribution)	5.8

Descriptive statistics

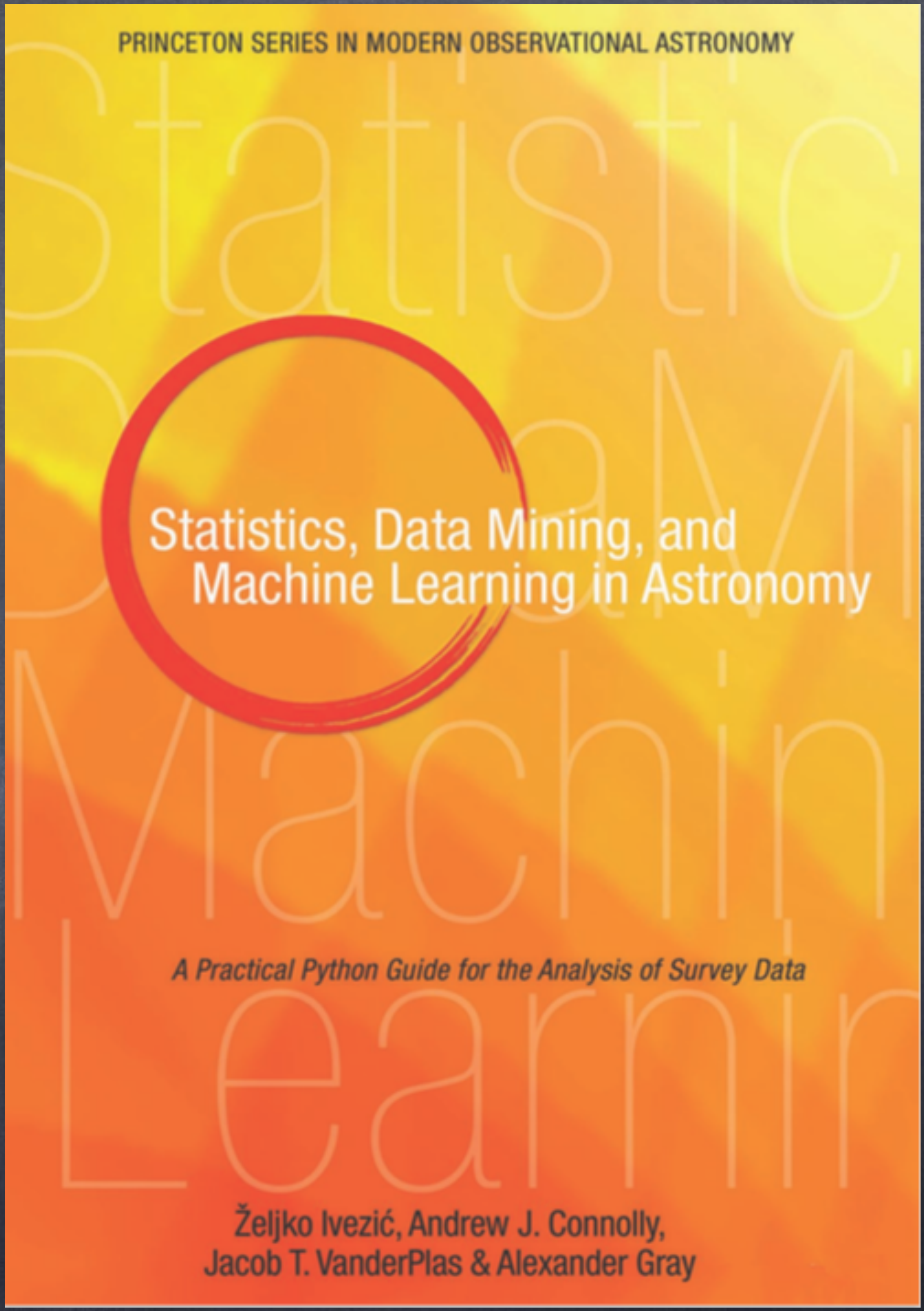
Jorge Lillo-Box

MCMC Coffee | Season 1, Episode 1

Statistics, Data Mining, and Machine Learning in Astronomy

Z. Ivezic, A.J. Conolly, J. T. Vanderplass & A. Gray

http://www.sc.eso.org/~jlillobo/mcmc_coffee/



Statistics, Data Mining, and Machine Learning

PRINCETON SERIES IN MODERN OBSERVATIONAL ASTRONOMY

MCMC Coffee

More Coffee More Confidence

HOME SCHEDULE REPOSITORY CONTACT USEFUL LINKS

Links on Astrostatistics

Name	Description
PyCoffee	Python coffee sessions taking place at ESO-Vitacura and organized by B. Dias, J. Milli, and D. Moser.
astroML	Python package to solve practical problems in Astronomy.
Astrostatistics and Astroinformatics Portal	New web site serving the cross-disciplinary communities of astronomers, statisticians and computer scientists. It is intended to foster research into advanced methodologies for astronomical research, and to promulgate such methods into the broader astronomy community

Books on Astrostatistics

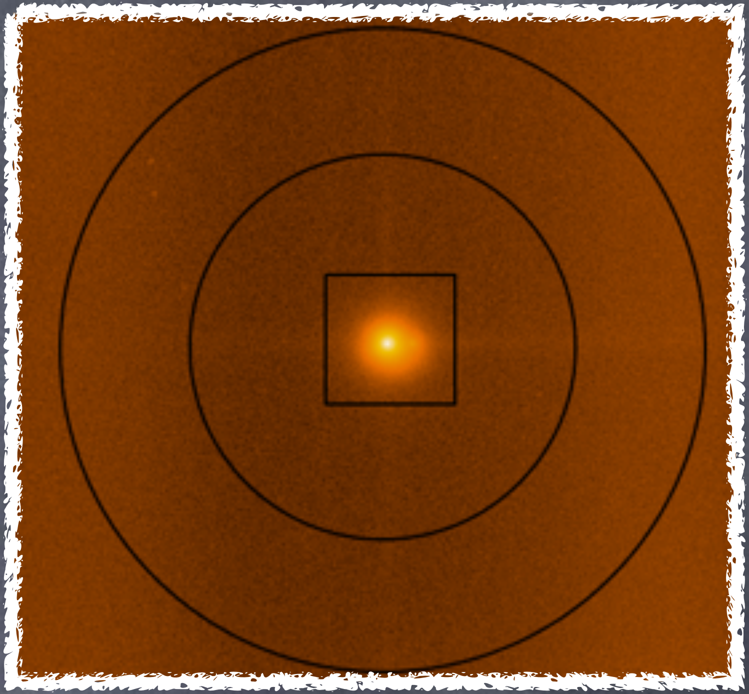
Title	Author	Abstract
Statistics, Data Mining, and Machine Learning in Astronomy	Željko Ivezić, Andrew J. Connolly, Jacob T. VanderPlas & Alexander Gray	Statistics, Data Mining, and Machine Learning in Astronomy presents a wealth of practical analysis problems, evaluates techniques for solving them, and explains how to use various approaches for different types and sizes of data sets. For all applications described in the book, Python code and example data sets are provided. The supporting data sets have been carefully selected from contemporary astronomical surveys (for example, the Sloan Digital Sky Survey) and are easy to download and use. The accompanying Python code is publicly available, well documented, and follows uniform coding standards. Together, the data sets and code enable readers to reproduce all the figures and examples, evaluate the methods, and adapt them to their own fields of interest. Modern astronomical research is beset with a vast range of statistical

Random variables

Random variable:

"A variable whose value results from the measurement of a quantity that is subject to random variations"

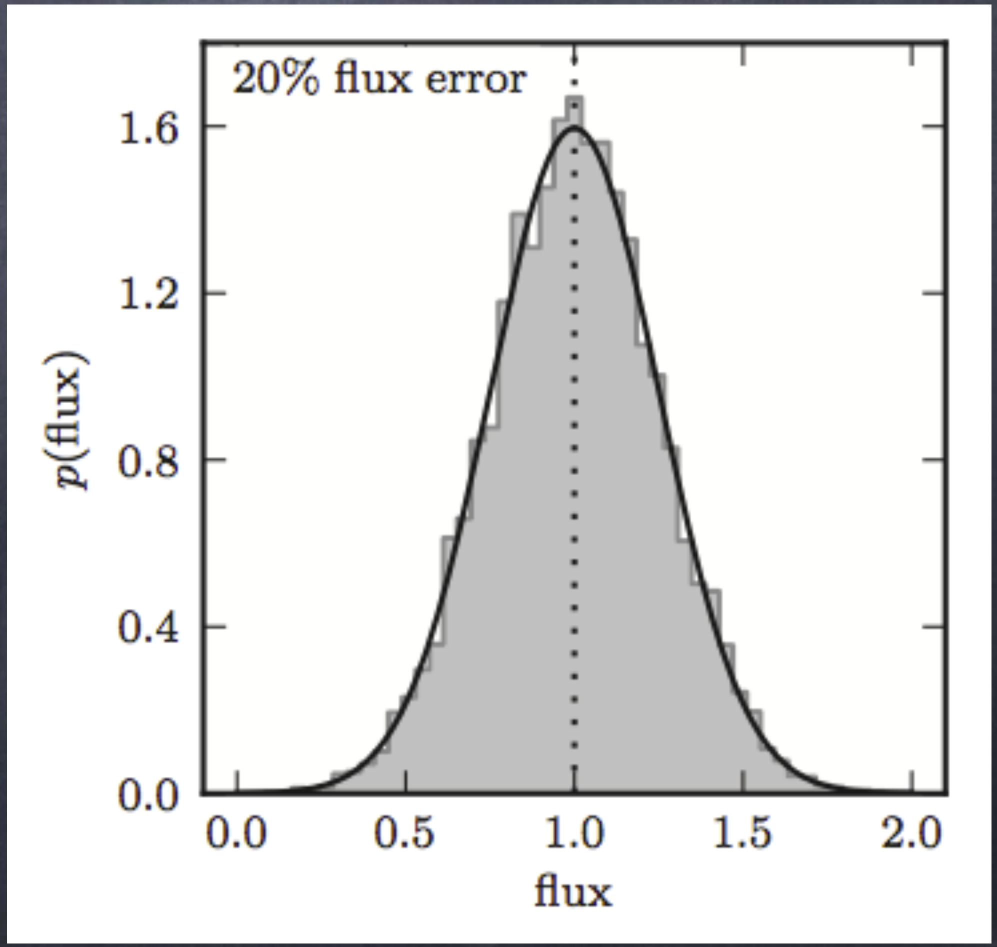
e.g.: bias level in a CCD, flux of a star, radial velocity



Probability density function (pdf):

"Probability value ascribed to each outcome of the random variable."

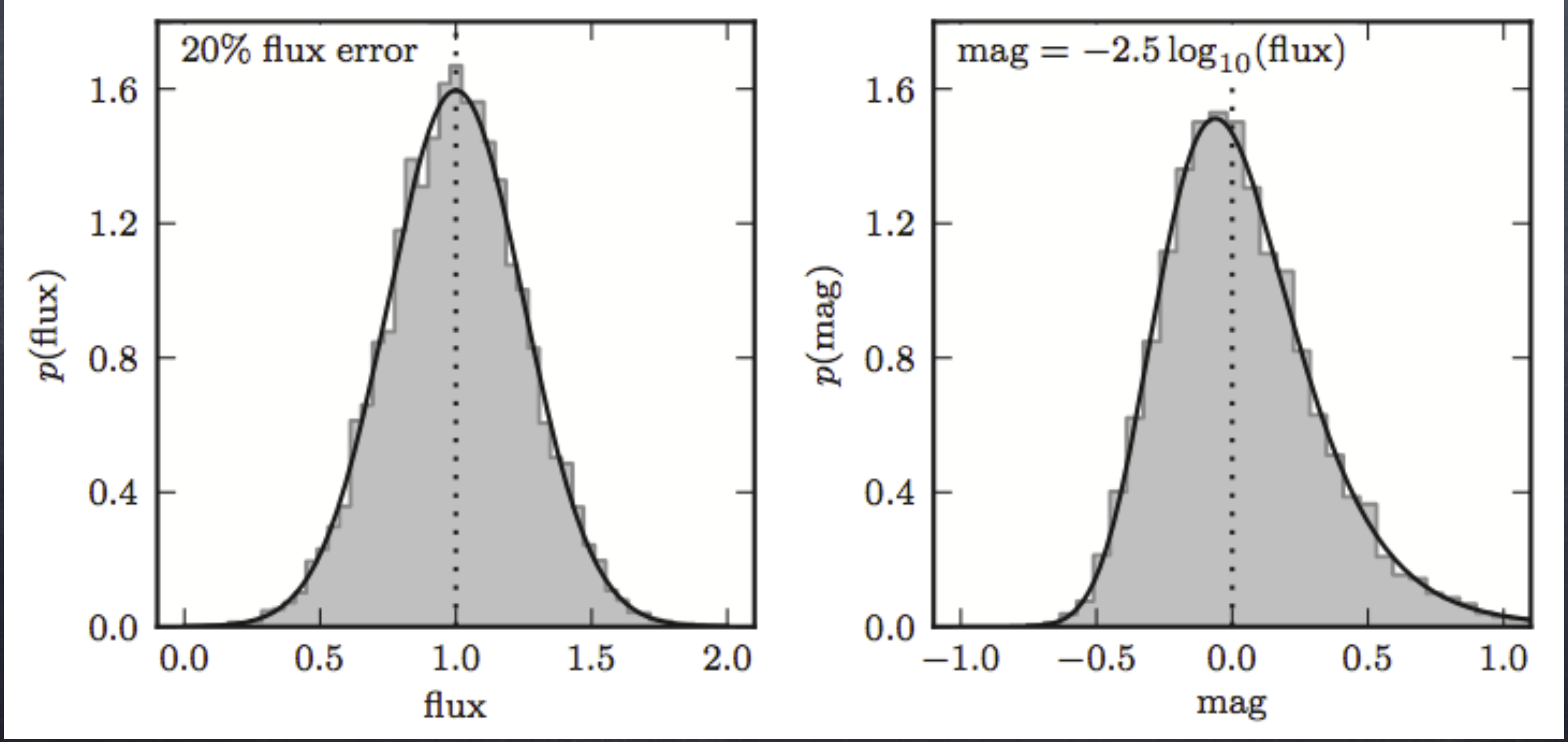
Associated with univariative distributions (uniform, Gaussian, binomial, gamma, etc.)



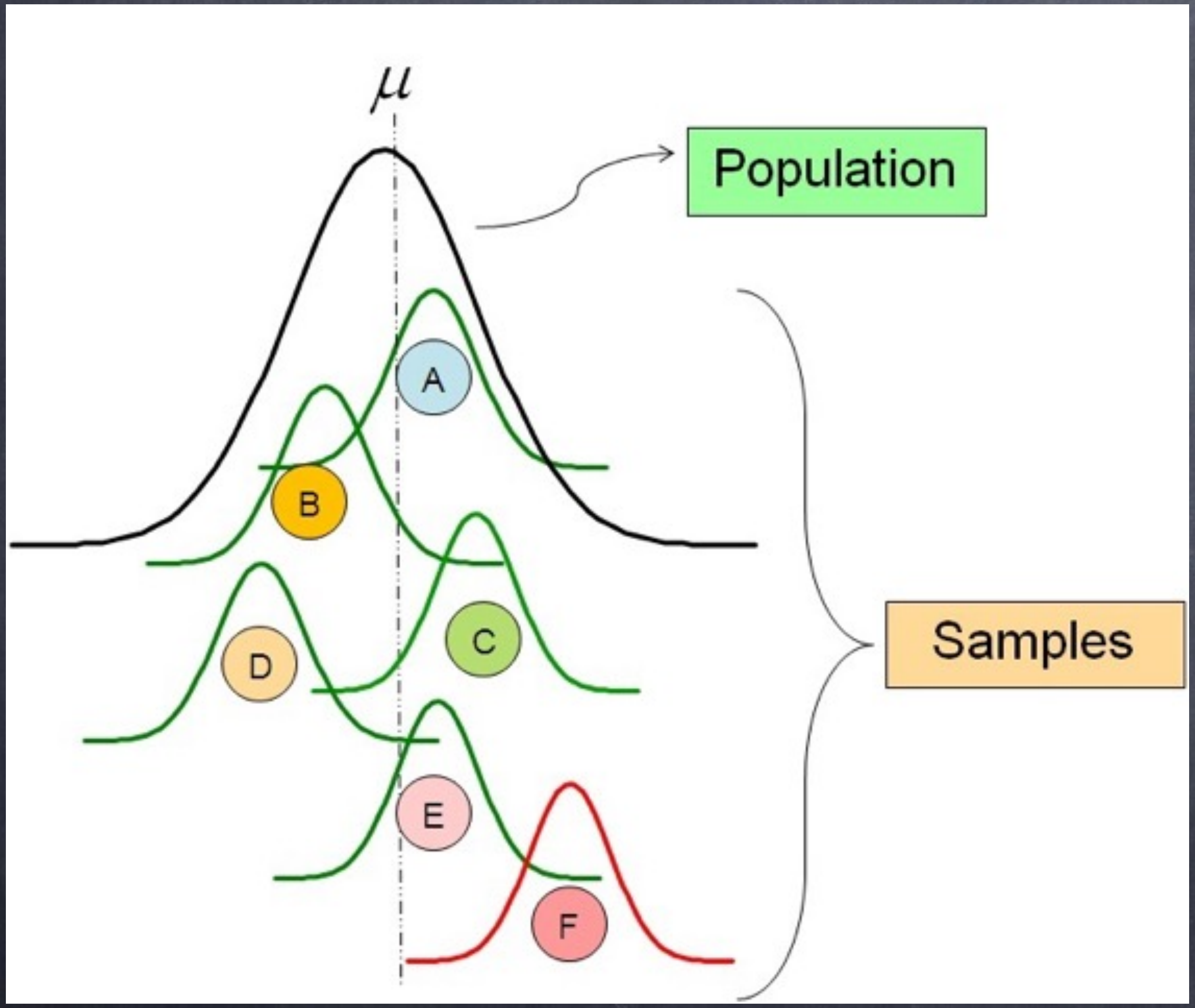
Random variables

Transformation of random variables

Be careful! Properties are not always preserved



Population vs. Sample statistics



Population: can be described by a distribution function f(x)

E.g.: the bias level of a CCD with 1024x1024 pixels

Sample: a finite number of measurements

E.g.: a subsample of the CCD of 20x20 pixels

Population vs. Sample statistics

How do you describe a population?
How do you describe a sample?

What are the typical shapes of the population distributions?

How large has to be my sample to properly represent the population?

What does "properly represent" mean?

Properties defining a population

Arithmetic mean

$$\mu = \int_{-\infty}^{\infty} x h(x) dx$$

Variance and standard deviation

$$V = \int_{-\infty}^{\infty} (x-\mu)^2 h(x) dx$$

$$\sigma = \sqrt{V}$$

Mode

$$\left(\frac{dh(x)}{x}\right)_{x_m} = 0$$

Skewness

$$\Sigma = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^3 h(x) dx$$

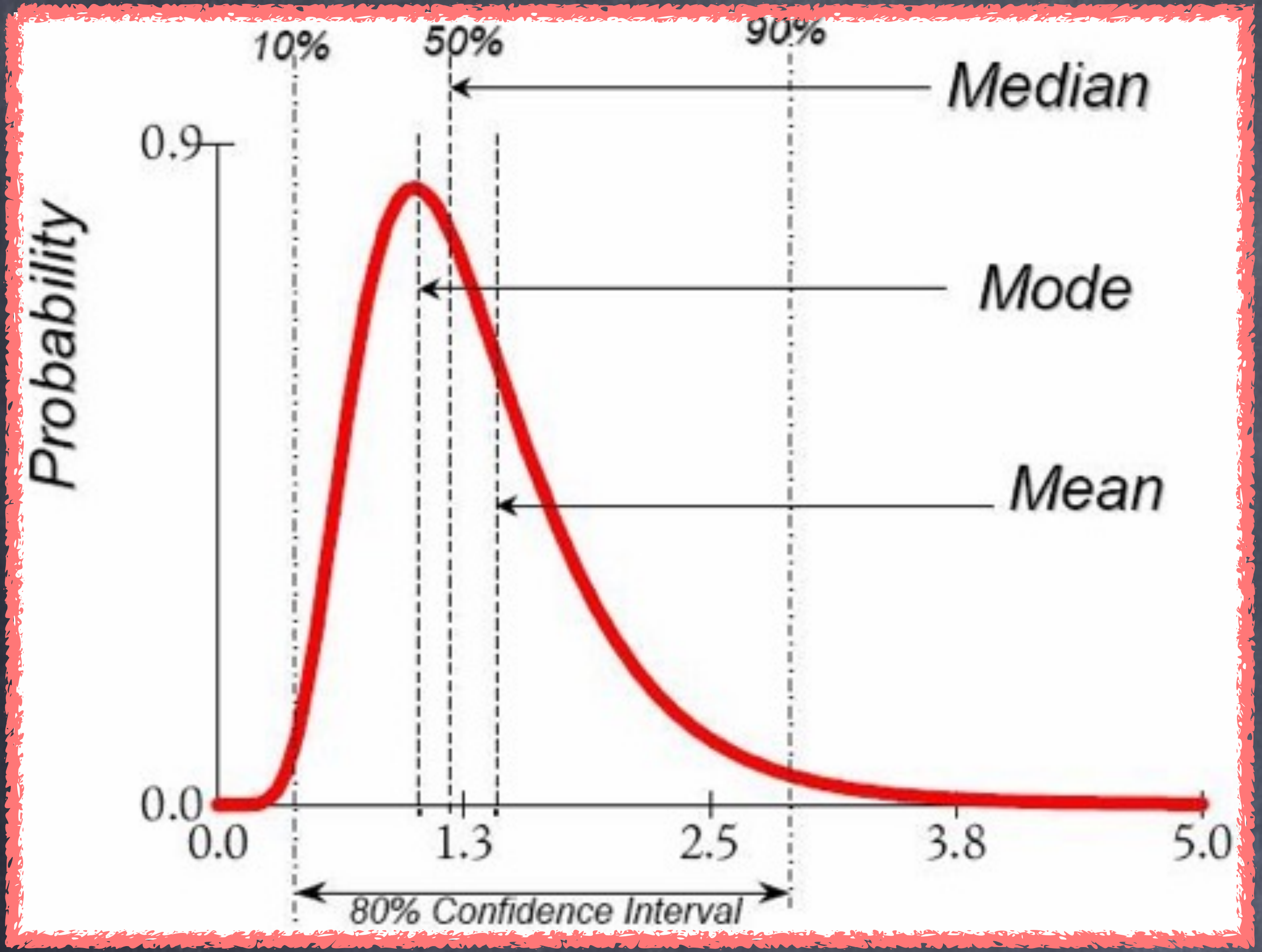
Percentile

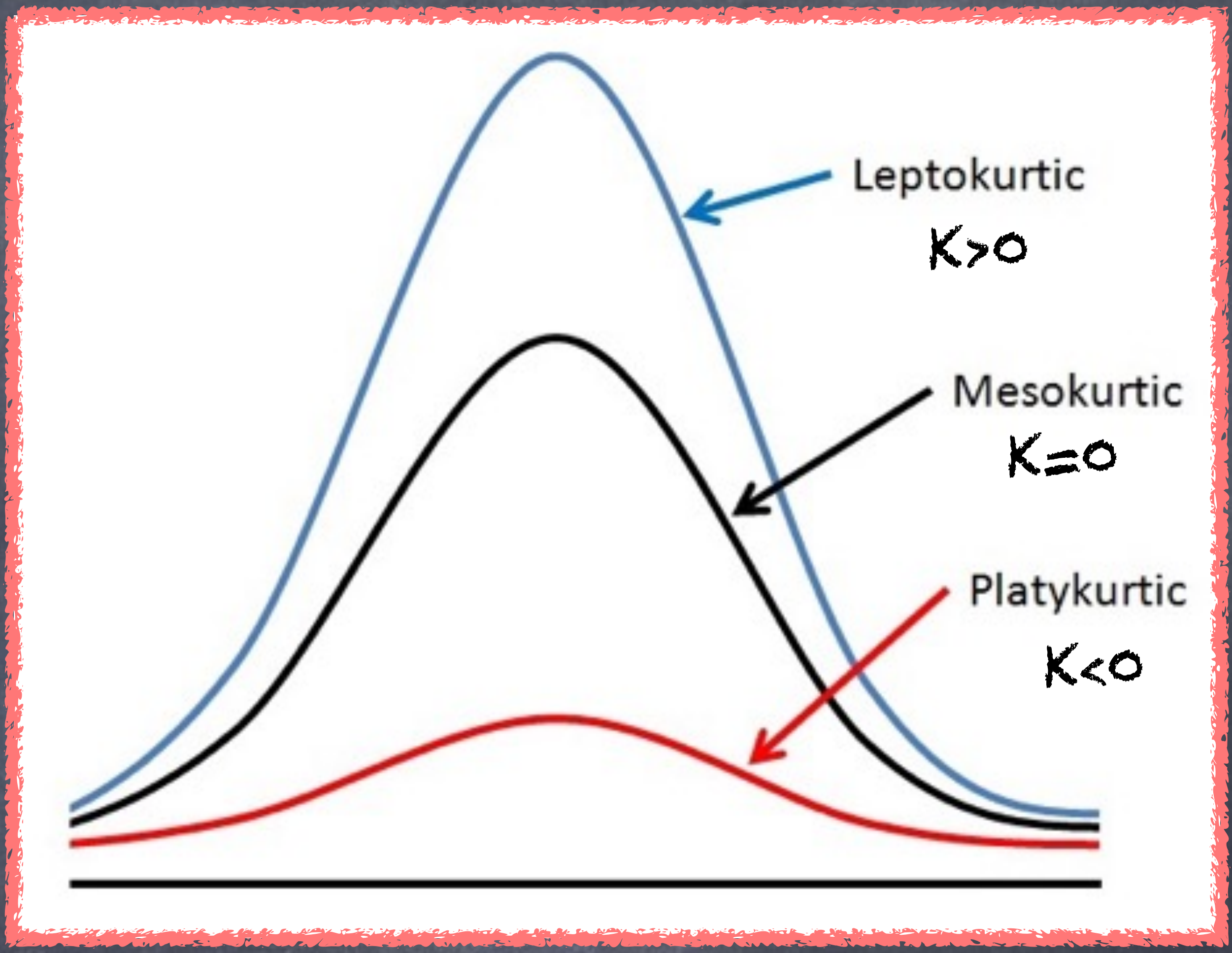
$$p = \int_{-\infty}^{q_p} h(x) dx$$

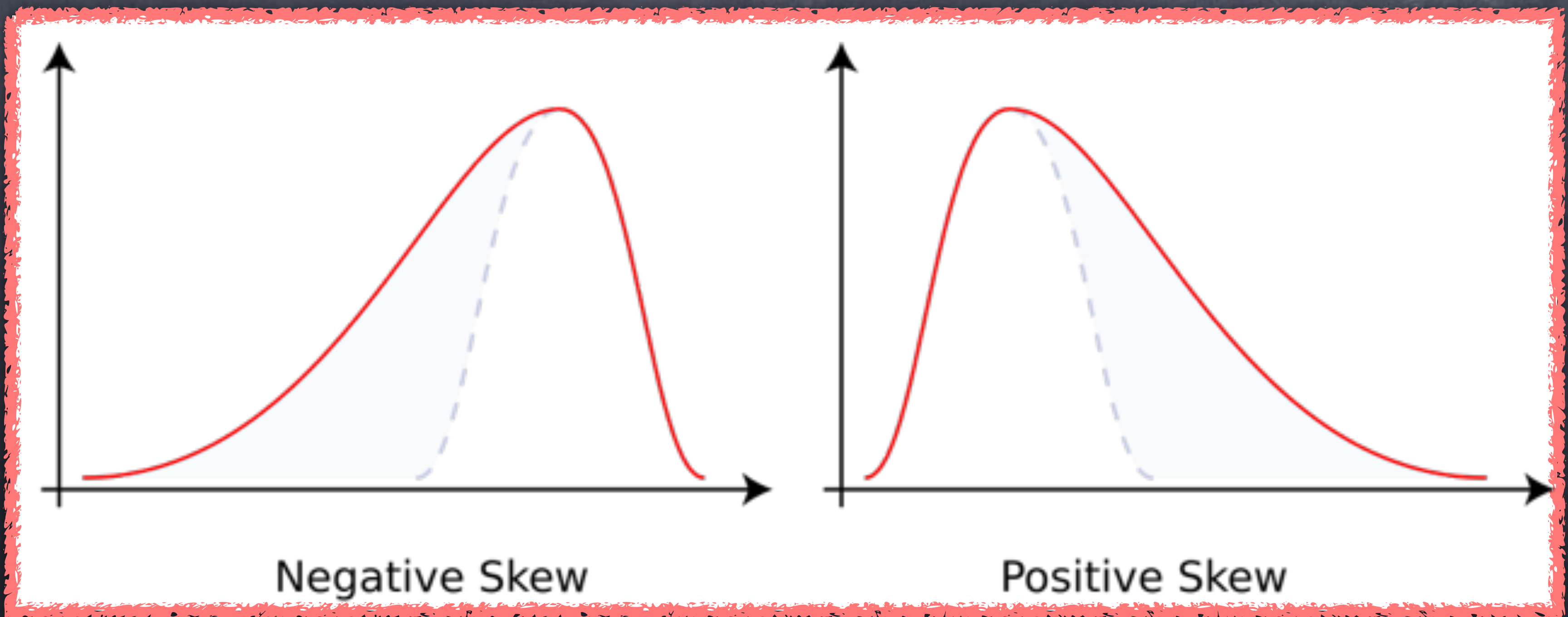
Kurtosis

$$K = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^4 h(x) dx - 3$$

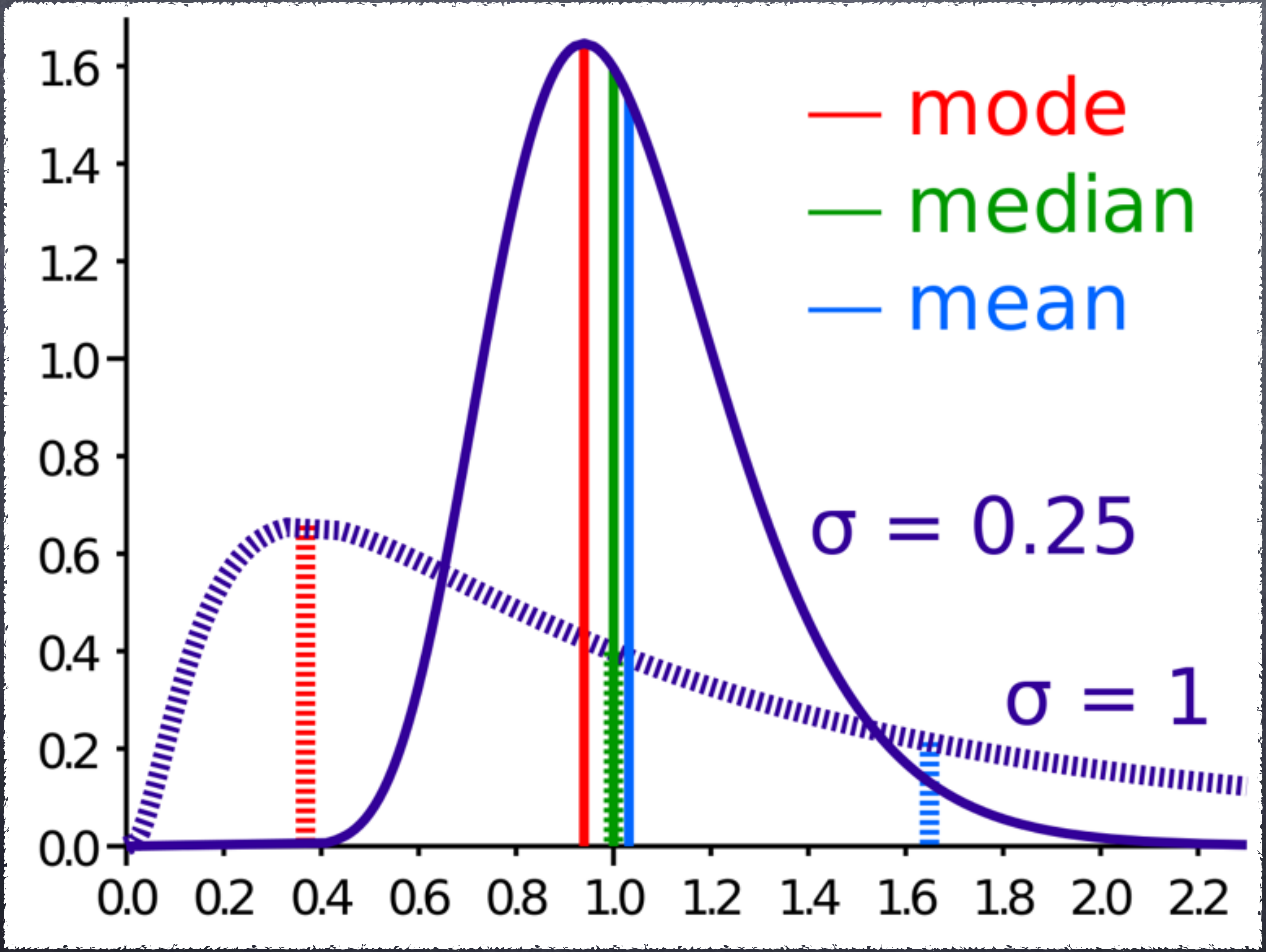
Properties defining a population



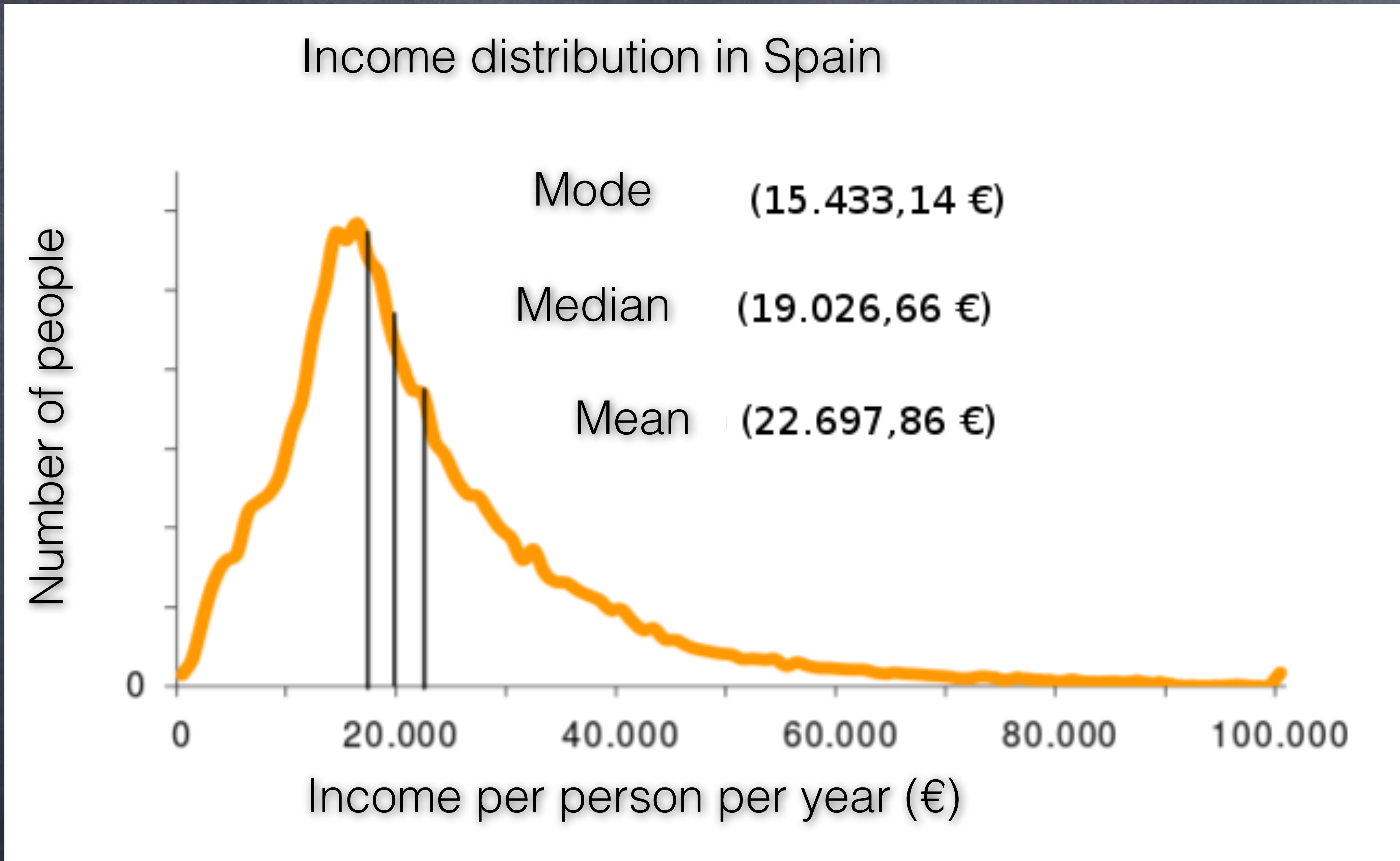




Properties defining a population



Properties defining a population



$$\int_{0.9}^{1.0} cdf(I)dI = \int_{0.0}^{0.6} cdf(I)dI$$

The 10% richest earns the same as the 60% poorest

Properties defining a population

Arithmetic mean

$$\mu = \int_{-\infty}^{\infty} x h(x) dx$$

Variance and standard deviation

$$V = \int_{-\infty}^{\infty} (x-\mu)^2 h(x) dx$$

$$\sigma = \sqrt{V}$$

Sample arithmetic mean

$$\hat{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

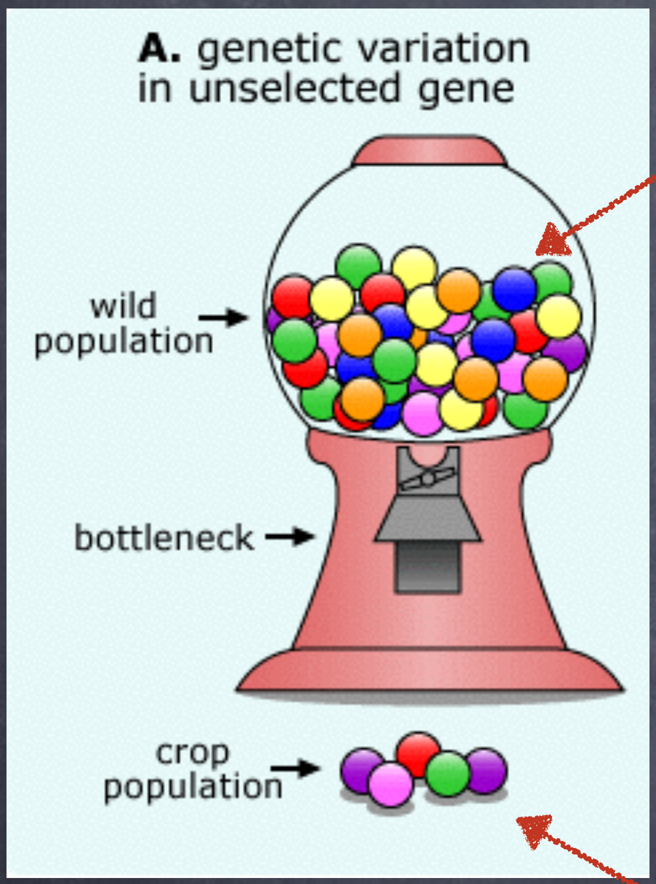
$$\sigma_{\hat{x}} = \frac{s}{\sqrt{N}}$$

Sample standard deviation

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{x})^2}$$

$$\sigma_s = \frac{1}{\sqrt{2}}\sqrt{\frac{N}{N-1}}\sigma_{\hat{x}}$$

True values vs. Estimators



population
μ, σ

sample
$\hat{x}, s$

Estimators are characterized by a bias and a variance

MSE = V + bias^2

Consistent estimator: $\text{bias}, V \xrightarrow[N \to \infty]{} 0$

True values vs. Estimators

How large a sample is required to obtain a given accuracy in our estimator?
(efficiency)

What is a good estimator?

Ideal case: subsample drawn from a Gaussian distribution

$$\sigma_{\text{median}} = 1.25\sigma_{\text{mean}}$$

The mean is more efficient than the median

Real case: real data

Outliers make the median a much more efficient estimator of the location

The interquartile ($q_{75}-q_{25}$) is a more robust estimator of the scale parameter

$$\sigma_G = 0.7413(q_{75} - q_{25})$$

Summary

Statistics is about dealing with random variables and describing their probability distributions

Statistics is about trying to inferr the true values of a population from estimators of a given sample

Describing a population/sample consists of providing a location (mean, median, mode) and a scale parameter (variance, skewness, kurtosis).

Samples are decribed by estimators. Their election is critical to accurately inferr the population properties

Suggested topics for the near future...

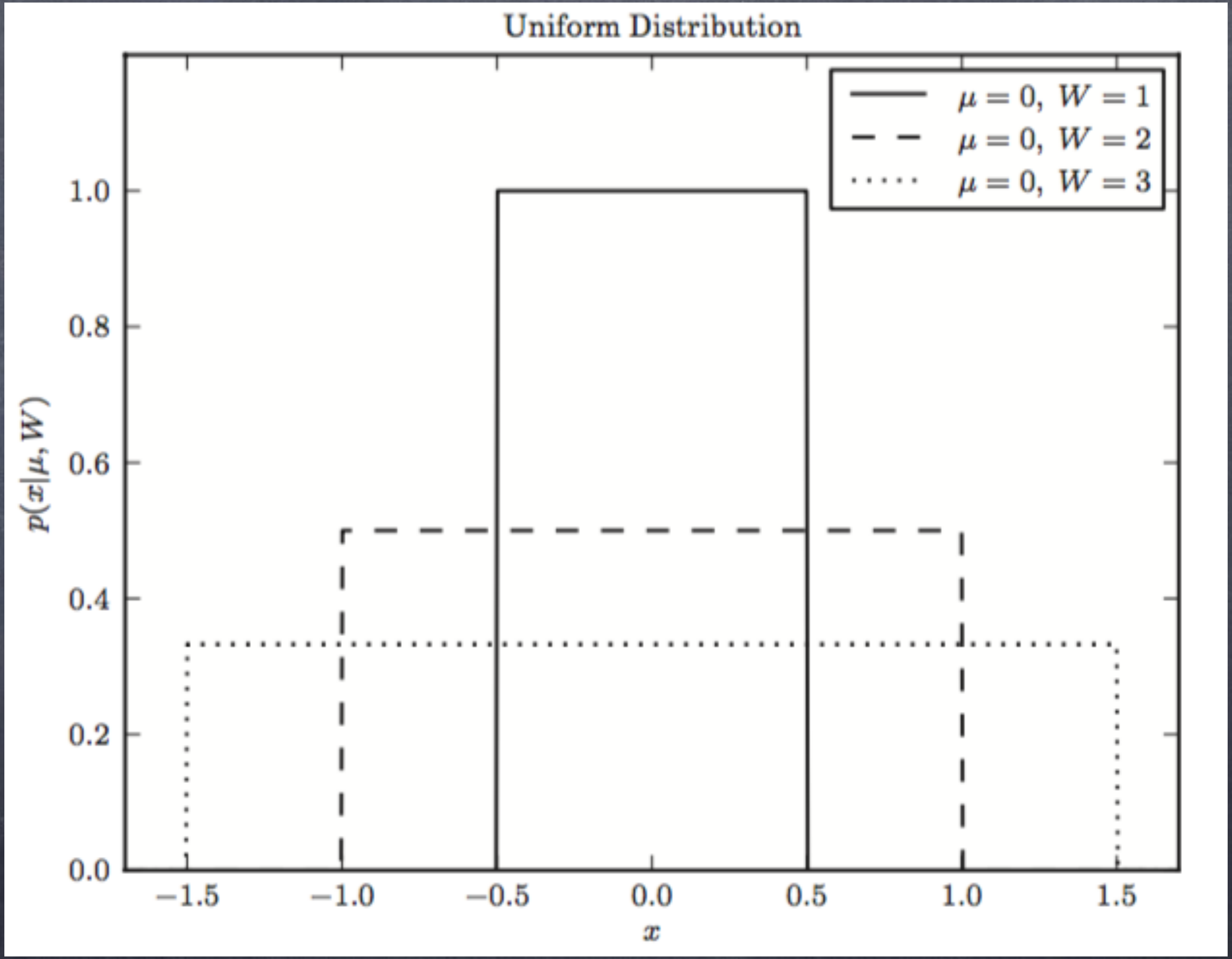
- Selection of priors
- Periodograms
- Interpreting the posterior probability
- Properly presenting your MCMC results
- Histograms (bin width selection)
- Kernel density estimators
- Computing the evidence from MCMC chains
- Model comparisson
- Noise colors (red noise, white noise, etc.)

Univariative distributions

- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...

Univariative distributions

- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...

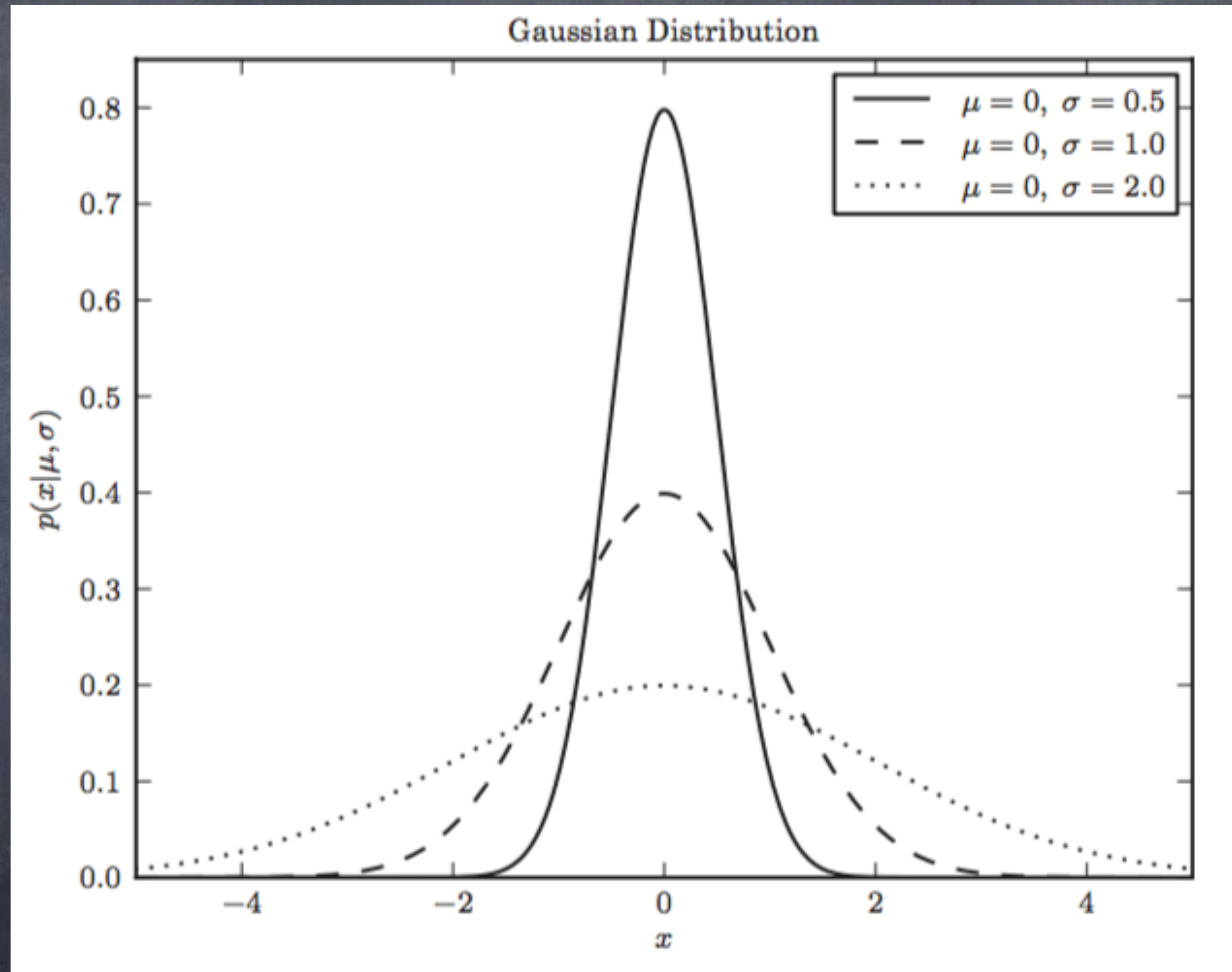


constrained values.
e.g.: date of birth {0,365}

$$p(x|\mu, W) = \frac{1}{W} \text{ for } |x - \mu| \leq \frac{W}{2},$$

Univariative distributions

- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...

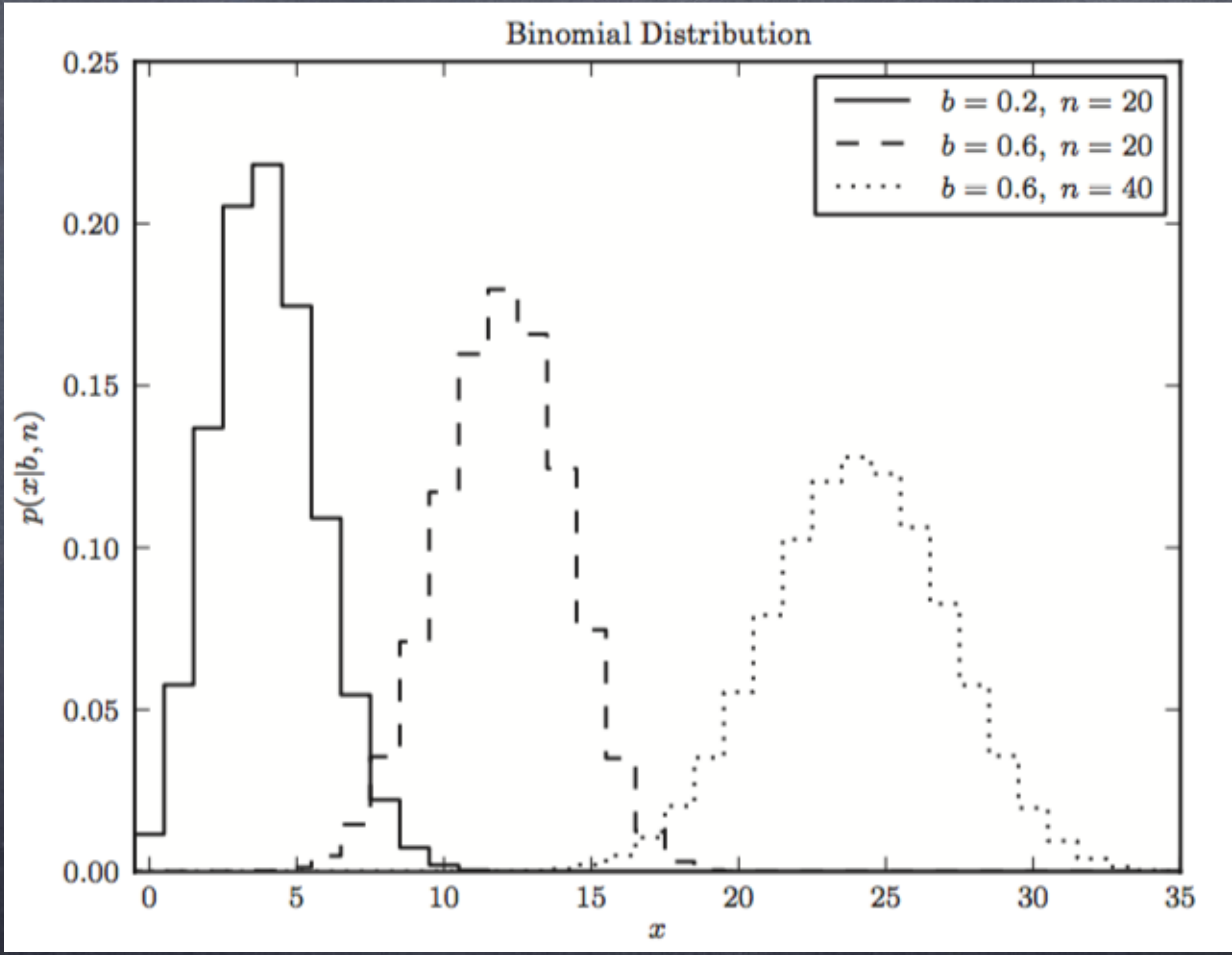


The number of particles whose velocity lies between x and x + dx is a gaussian distribution

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

Univariative distributions

- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...

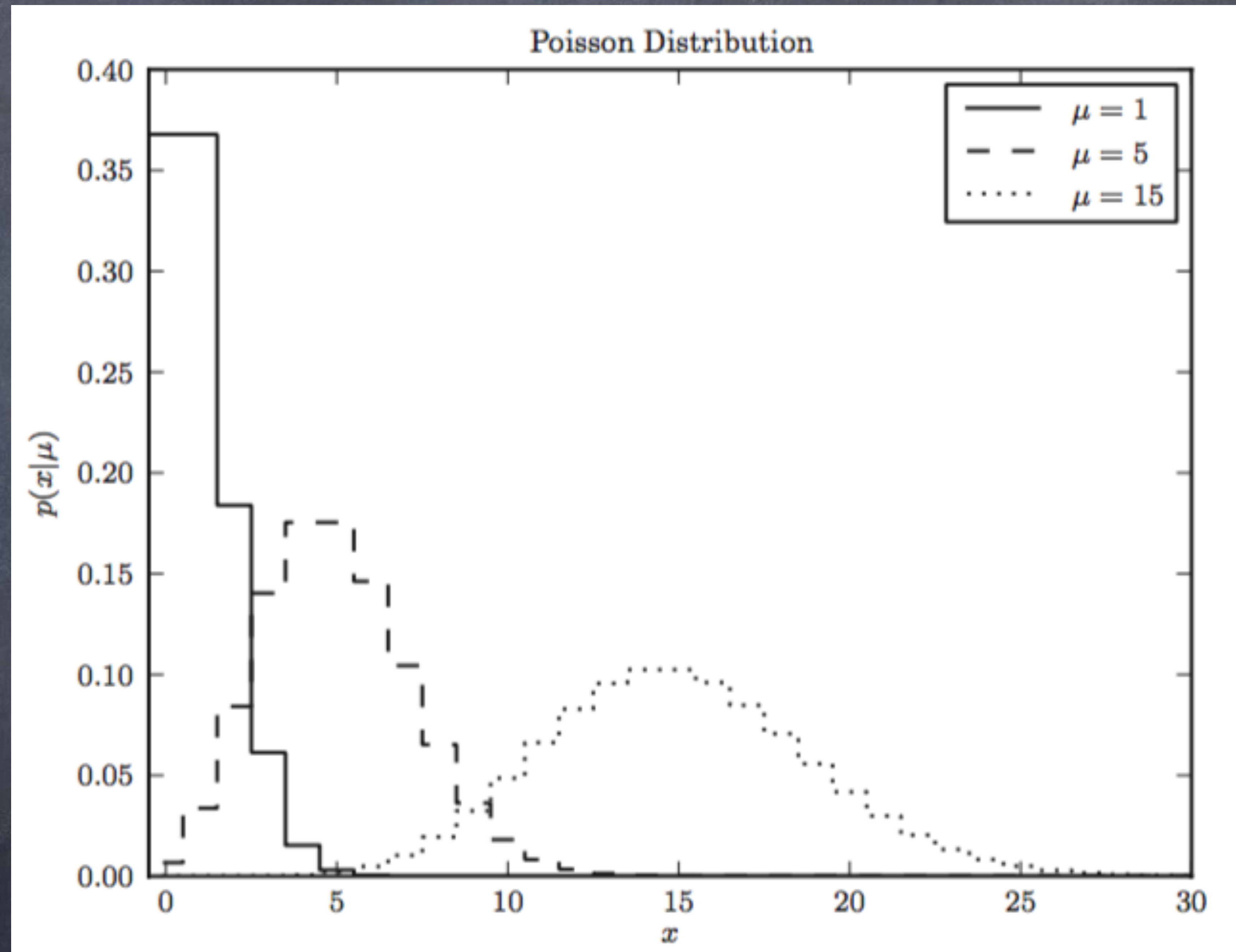


x can just take discrete values (integers). e.g.: flipping a coin {heads,tails}

$$p(k|b, N) = \frac{N!}{k!(N-k)!} b^k (1-b)^{N-k}$$

Univariative distributions

- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...

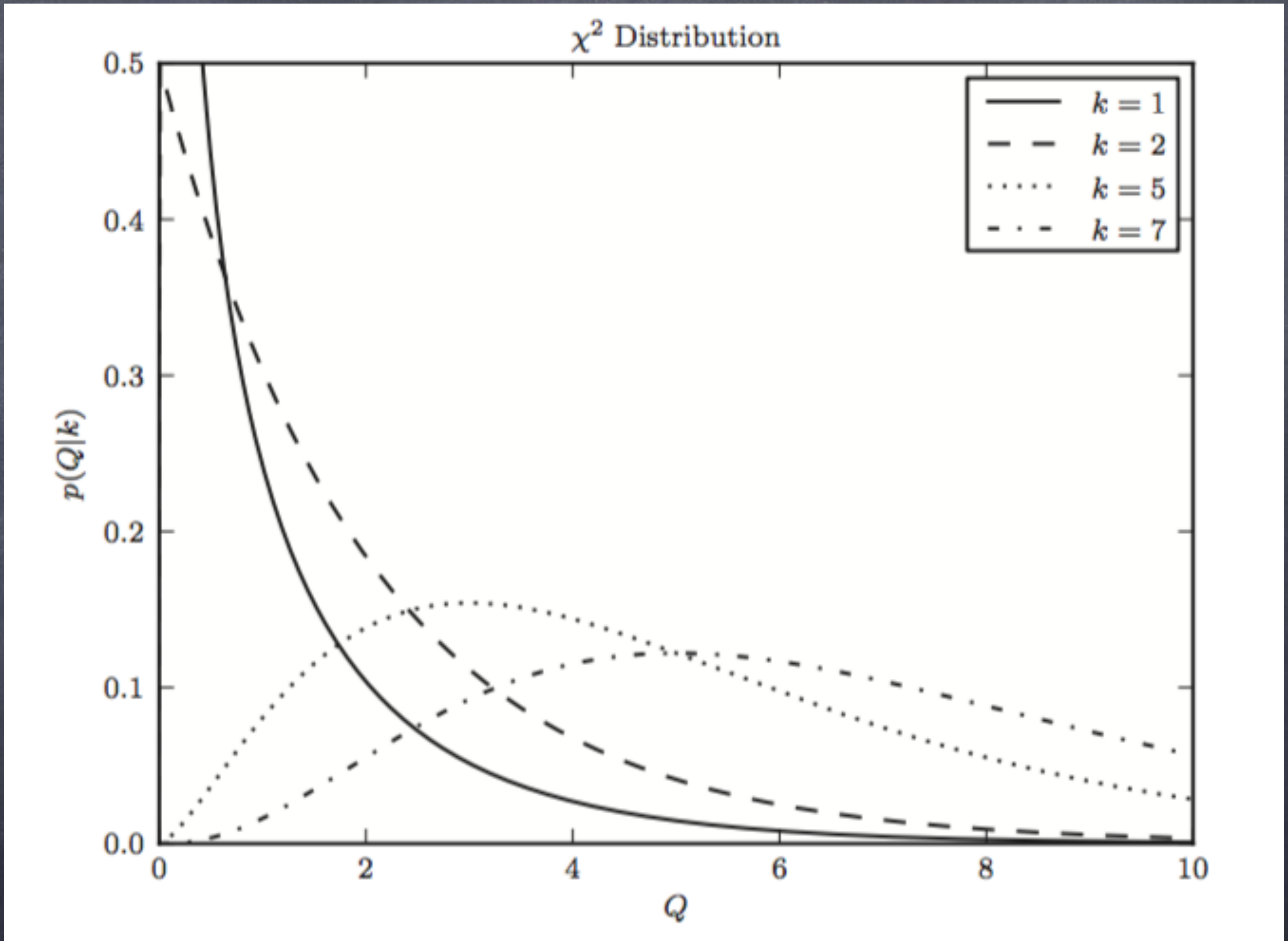


e.g.: the distribution of the number of photons counted in a time interval.

$$p(k|\mu) = \frac{\mu^k \exp(-\mu)}{k!}$$

Univariative distributions

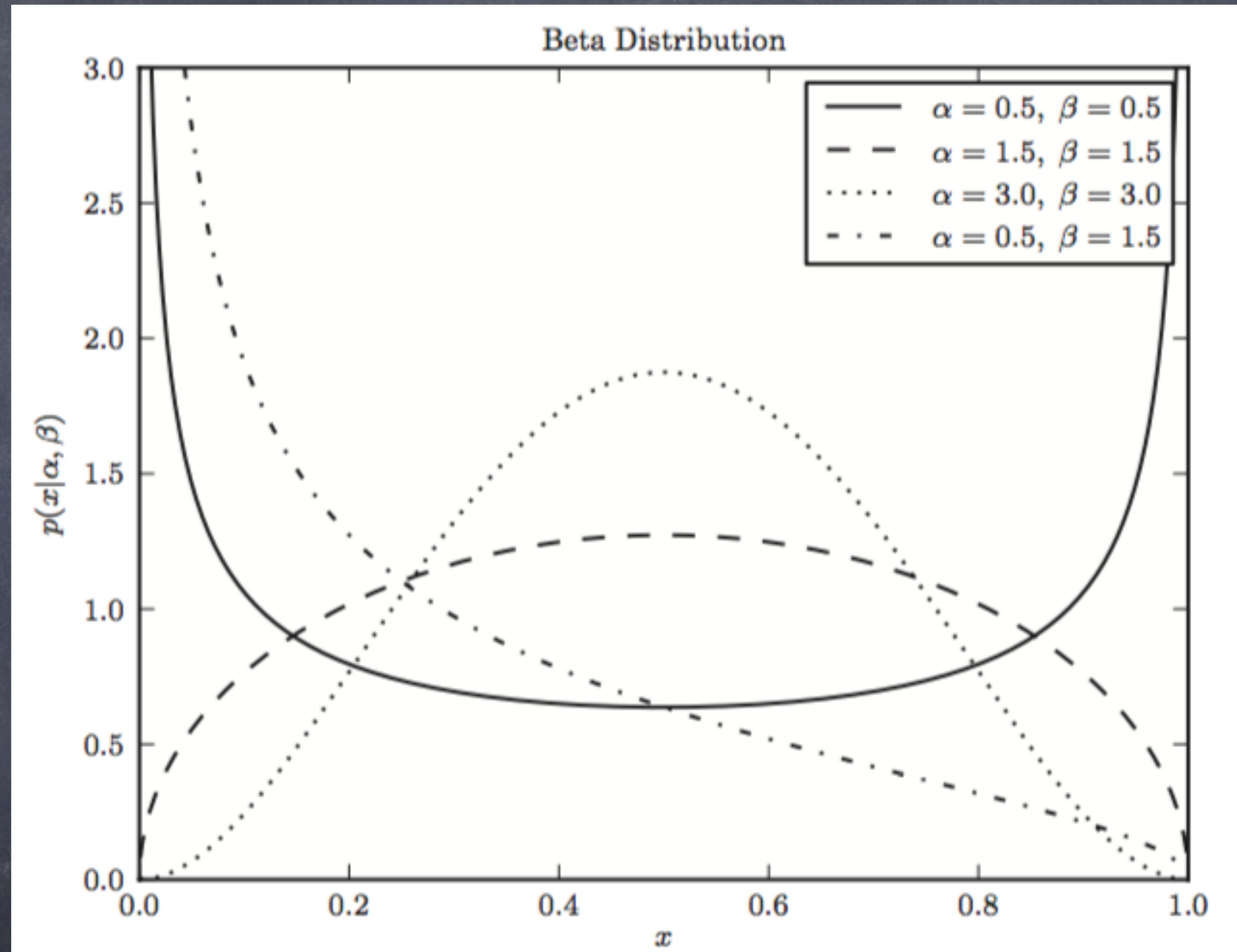
- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...



$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{k/2-1} \exp(-Q/2) \text{ for } Q > 0,$$

Univariative distributions

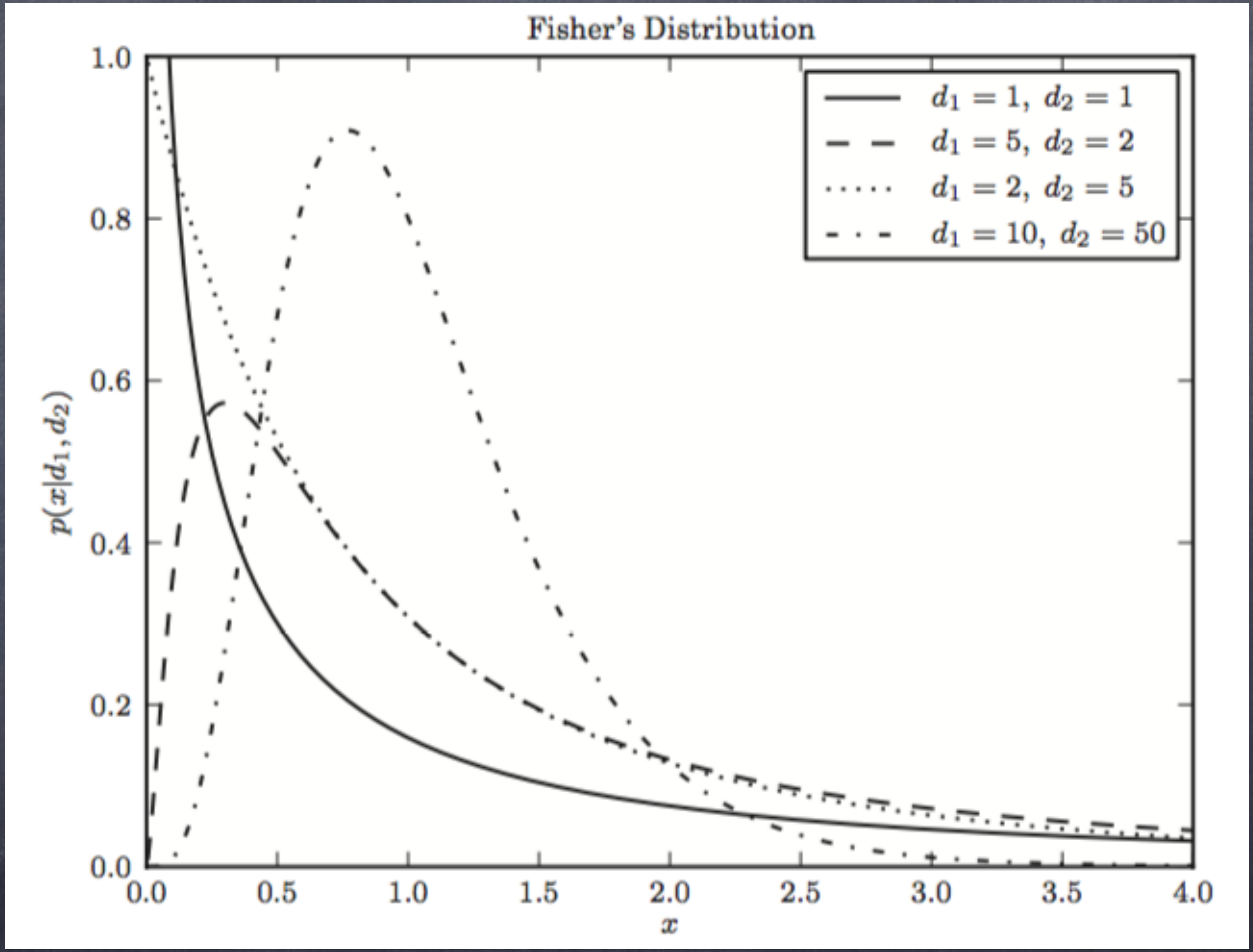
- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...



$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1},$$

Univariative distributions

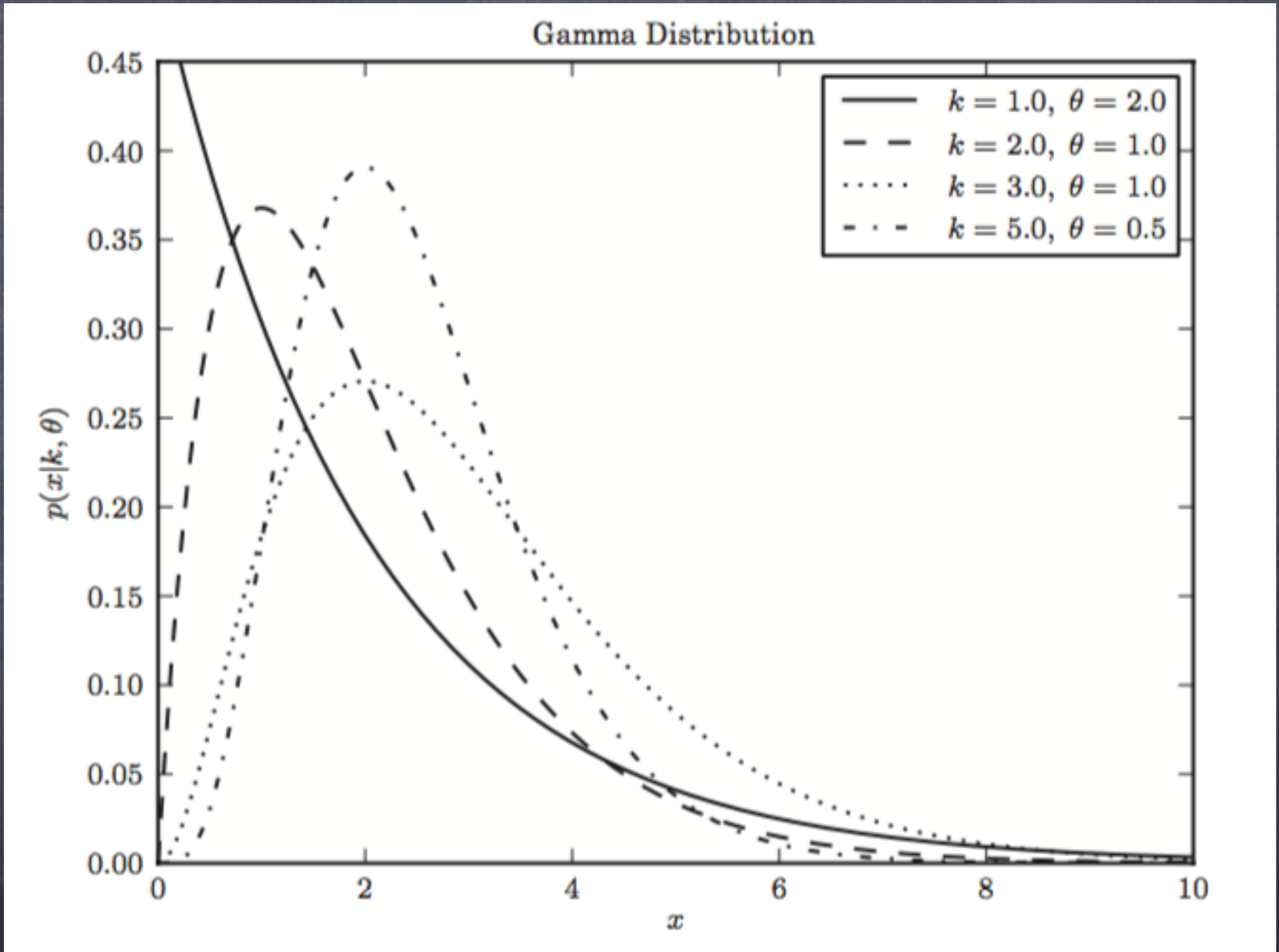
- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...



$$p(x|d_1, d_2) = C\left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}} x^{\frac{d_1}{2}-1},$$

Univariative distributions

- Uniform
- Gaussian
- Binomial
- Poisson
- Chi-square
- Beta
- Fischer
- Gamma
- ...



$$p(x|k,\theta) = \frac{1}{\theta^k}\frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)},$$