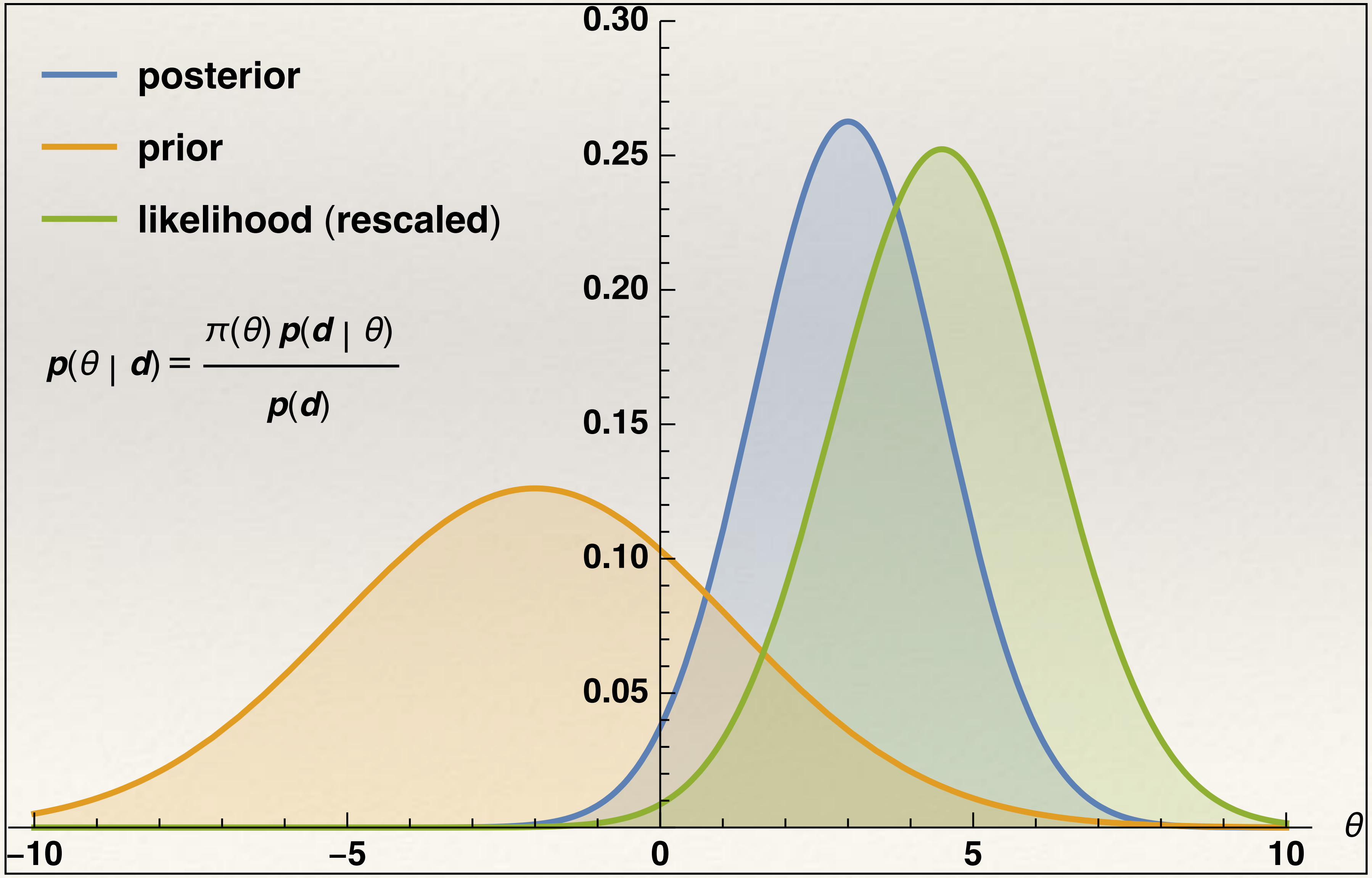
Making sense of data: introduction to statistics for gravitational wave astronomy

Lecture 1: introduction to random variables

AEI IMPRS Lecture Course

Jonathan Gair jgair@aei.mpg.de



Outline of course

- Lectures will take place at 11am Monday, Wednesday, Thursday and Friday in the weeks beginning Nov 8th, 15th, 22nd and Dec 6th. **Exceptions**: Nov. 10th (3.30pm) and Dec. 6th (3pm).
- Lectures will all take place in seminar room 0.01 at the AEI (following the 2G protocol) and will also be broadcast via Zoom
  - Meeting ID: 991 8595 0549
  - Meeting password: 490287
- Lecture recordings and other material will be made available on the course website
  - https://imprs-gw-lectures.aei.mpg.de/2021-making-sense-of-data/

Outline of course

- **Part 1 (week 1): Frequentist statistics**
    - Random variables: definition, properties, some useful probability distributions, central limit theorem.
    - Statistics: definition, estimators, likelihood, desirable properties of estimators, Cramer-Rao bound.
    - Hypothesis testing: definition, Neyman-Pearson lemma, power and size of tests, type I and type II errors, ROC curves, confidence regions, uniformly-most-powerful tests.
    - Frequentist statistics in GW astronomy: false alarm rates, Fisher Matrix, PSD estimation.

Outline of course

- **Part 2 (week 2): Bayesian statistics**
    - Bayes' theorem, conjugate priors, Jeffrey's prior.
    - Bayesian hypothesis testing, hierarchical models, posterior predictive checks.
    - Sampling methods for Bayesian inference.
    - Bayesian statistics in GW astronomy: parameter estimation, population inference, model selection.

Outline of course

- **Part 3 (week 3): Stochastic processes and sampling in python**
  - Lecture: stochastic processes, optimal filtering, signal-to-noise ratio, sensitivity curves.
  - Practicals: simulating random variables in python, sampling posterior distributions using pystan.

Outline of course

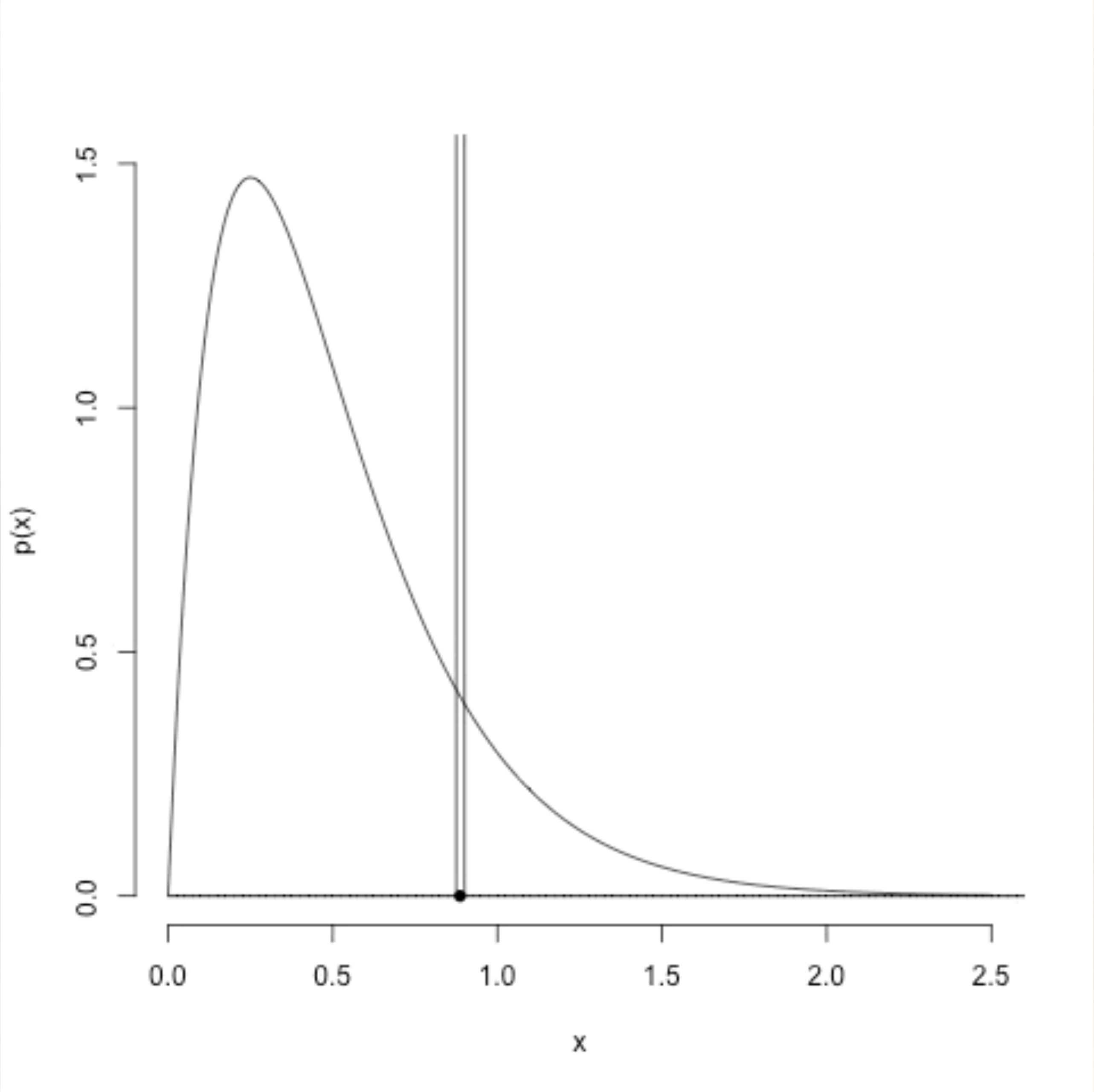
- **Part 4 (week 4): Introduction to machine learning**
  - Lectured by Stephen Green. More details at the start of that lecture block.

Outline of course

- Lecture notes will be made available. These notes will include more material than will be covered in lectures, and these extra topics will be denoted by an asterix. A block of notes on "Advanced Topics in Statistics" will also be made available, which can be used as a reference (or not), at the discretion of the student.
- One problem set will be provided for each block of lectures. Solutions will be made available later. These sheets will have two parts, with the second part containing optional questions that are either more difficult or similar to questions in the first part.

Random variables

- *Random variables* are quantities that are not fixed, but can take new values each time they are observed (a *realisation*).
- Over many realisations the distribution of the random variable is described by a *probability distribution*.
- Random variables can be *discrete* (taken values in a countable set) or *continuous* (taking real values in some interval).



Discrete random variables

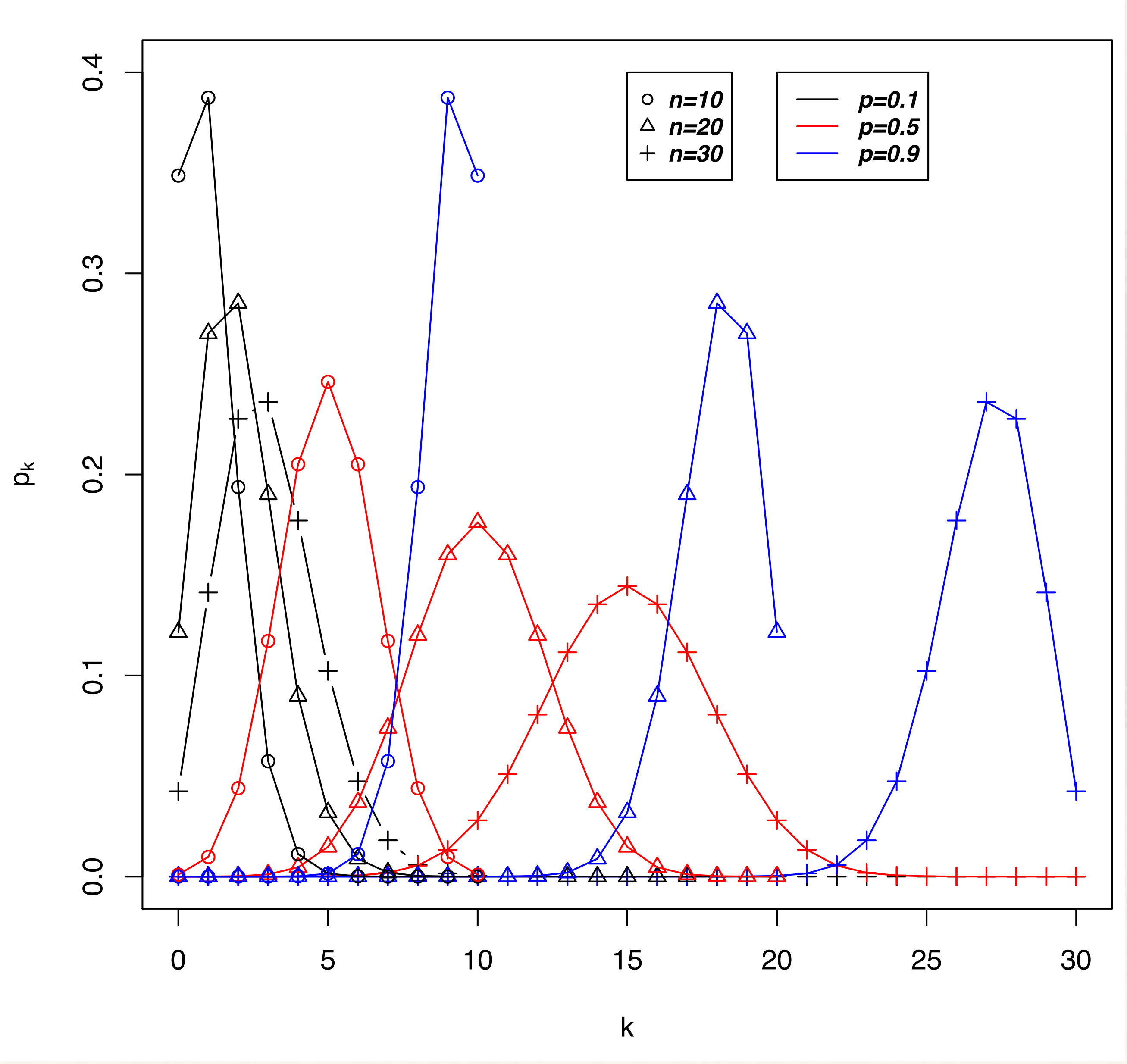
- Discrete random variables are characterised by a *probability mass function*, i.e., a set $\{p_i\}$ satisfying

$$0 \leq p_i \leq 1 \qquad \sum p_i = 1$$

- For example, **Binomial distribution**

$$P(X=k) = p_k = \begin{cases} \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} & \text{if } k \in \{1, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}$$

- Related distributions: *Bernoulli distribution, negative Binomial, geometric distribution.*

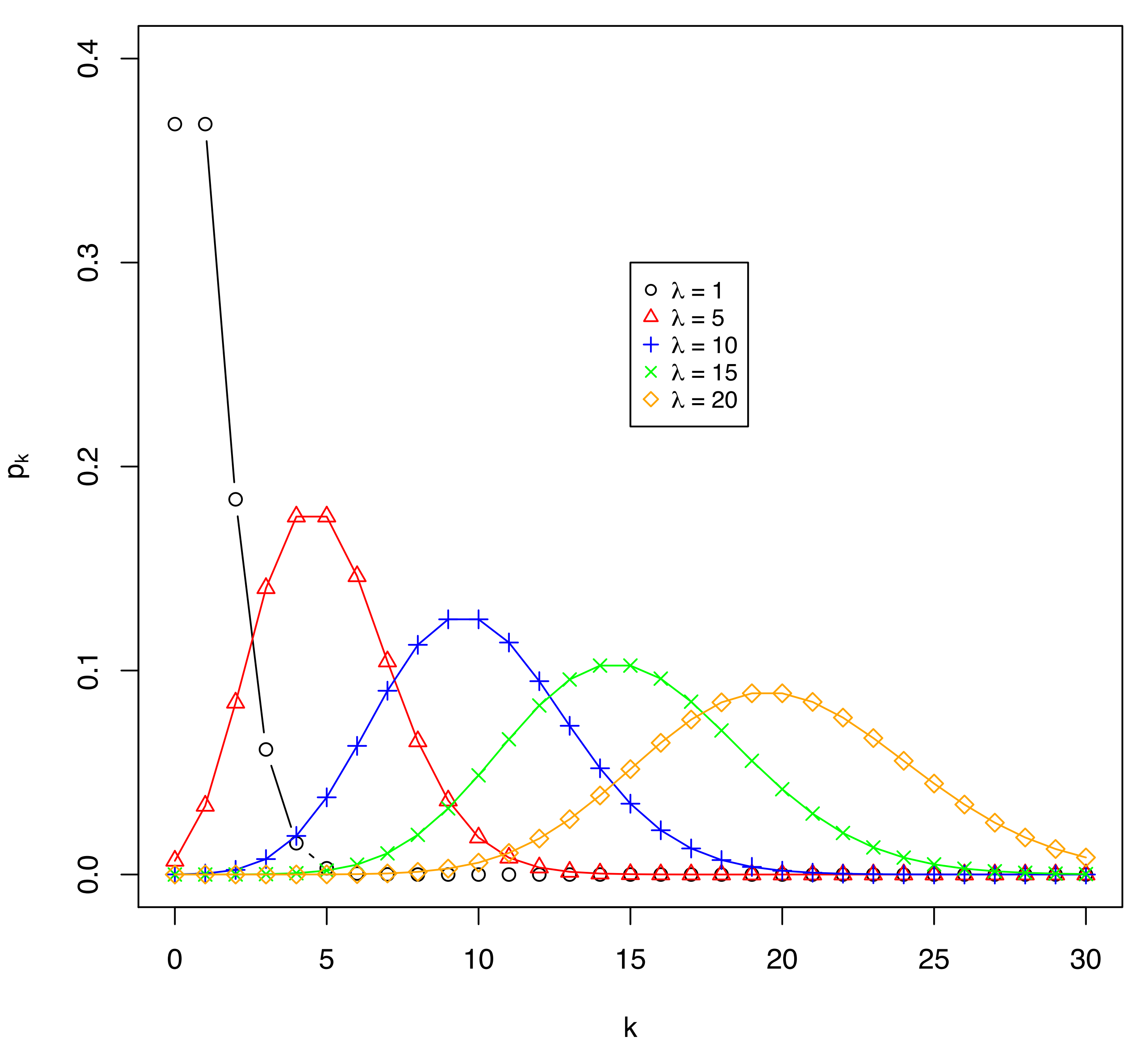


Discrete RVs: Poisson distribution

- **Poisson distribution** is defined for non-negative k by

$$P(X = k) = p_k = \begin{cases} \lambda^k e^{-\lambda}/k! \\ 0 \end{cases}$$

- Arises as the distribution of the number of counts of a process occurring in a certain period of time.



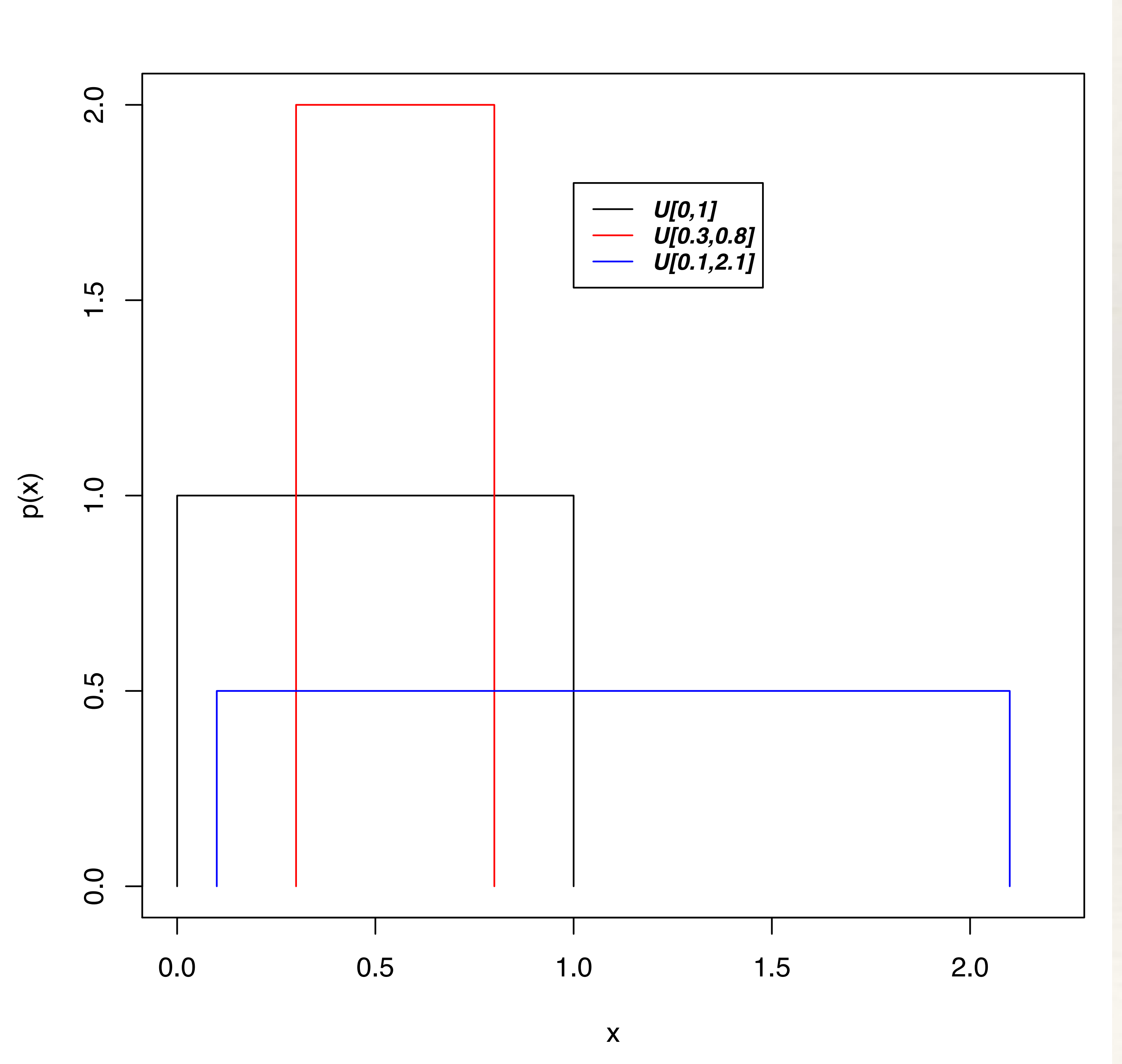
Continuous random variables

- Continuous random variables are characterised by a *probability density function*, satisfying

$$0 \leq p(x) \qquad \int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1$$

- For example, **Uniform distribution**

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



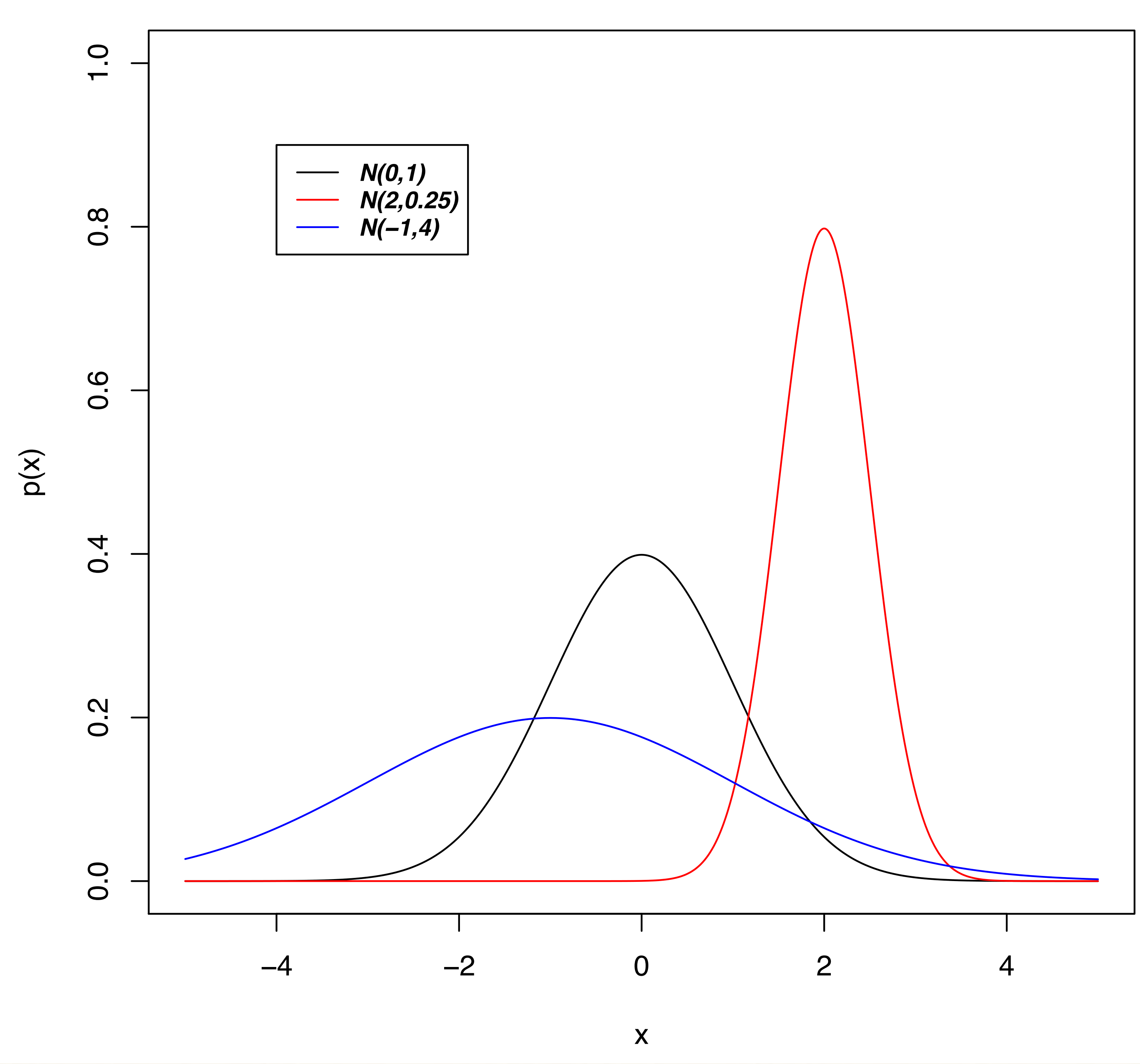
Continuous RVs: Normal distribution

- **Normal distribution** is characterised by mean μ and variance σ^2

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

- Arises as a limiting distribution and as the distribution of noise in gravitational wave detectors. Commonly used as the default distribution in parametric statistics and as a prior in Bayesian analysis.

- Normal distribution with zero mean and unit variance is the *standard Normal distribution*.

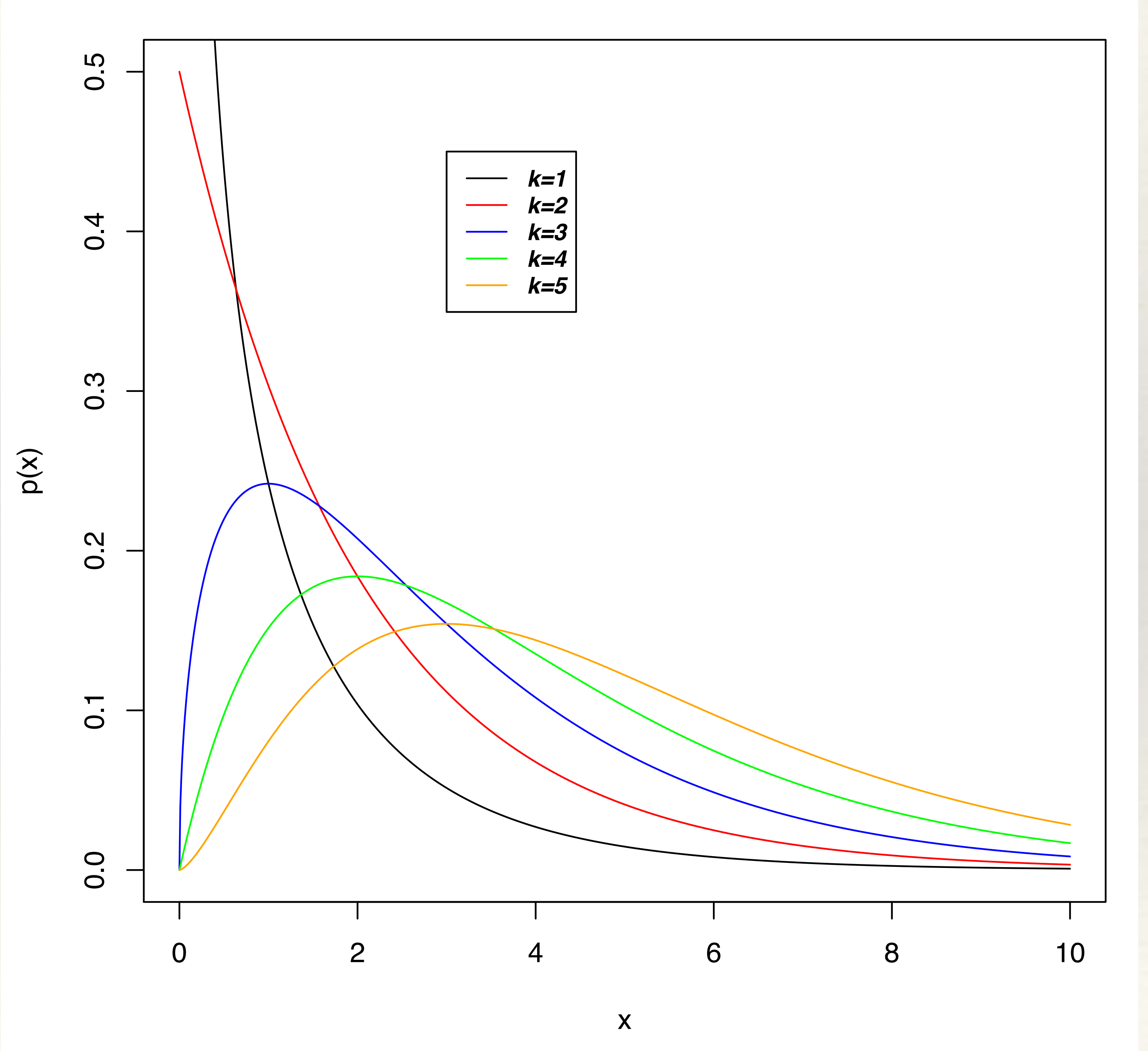


Continuous RVs: chi-squared distribution

- **Chi-squared distribution** depends on a *degrees of freedom* parameter k > *0*

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} \mathrm{e}^{-\frac{x}{2}}$$

- It is the distribution of the sum of squares of *k* standard normal random variables.
- There is also a *non-central chi-square distribution* which has also a *non-centrality parameter*.

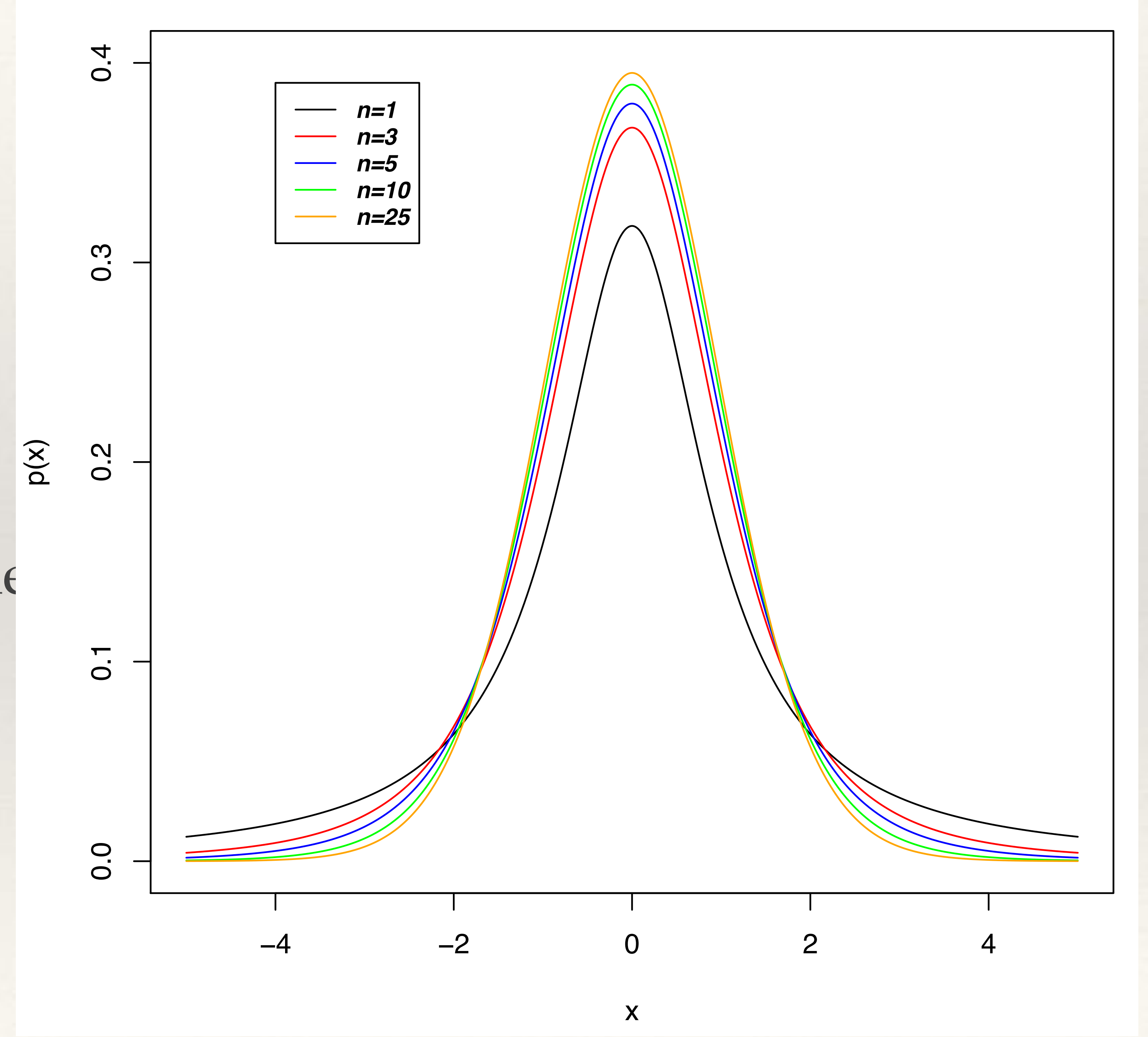


Continuous RVs: Student's t-distribution

- **Student's t-distribution** *also* depends on a degrees of freedom parameter n

$$p(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

- It arises in hypothesis testing as the ratio of a standard Normal distribution to a chi-squared distribution. It is used as a *heavy-tailed* distribution in inference.

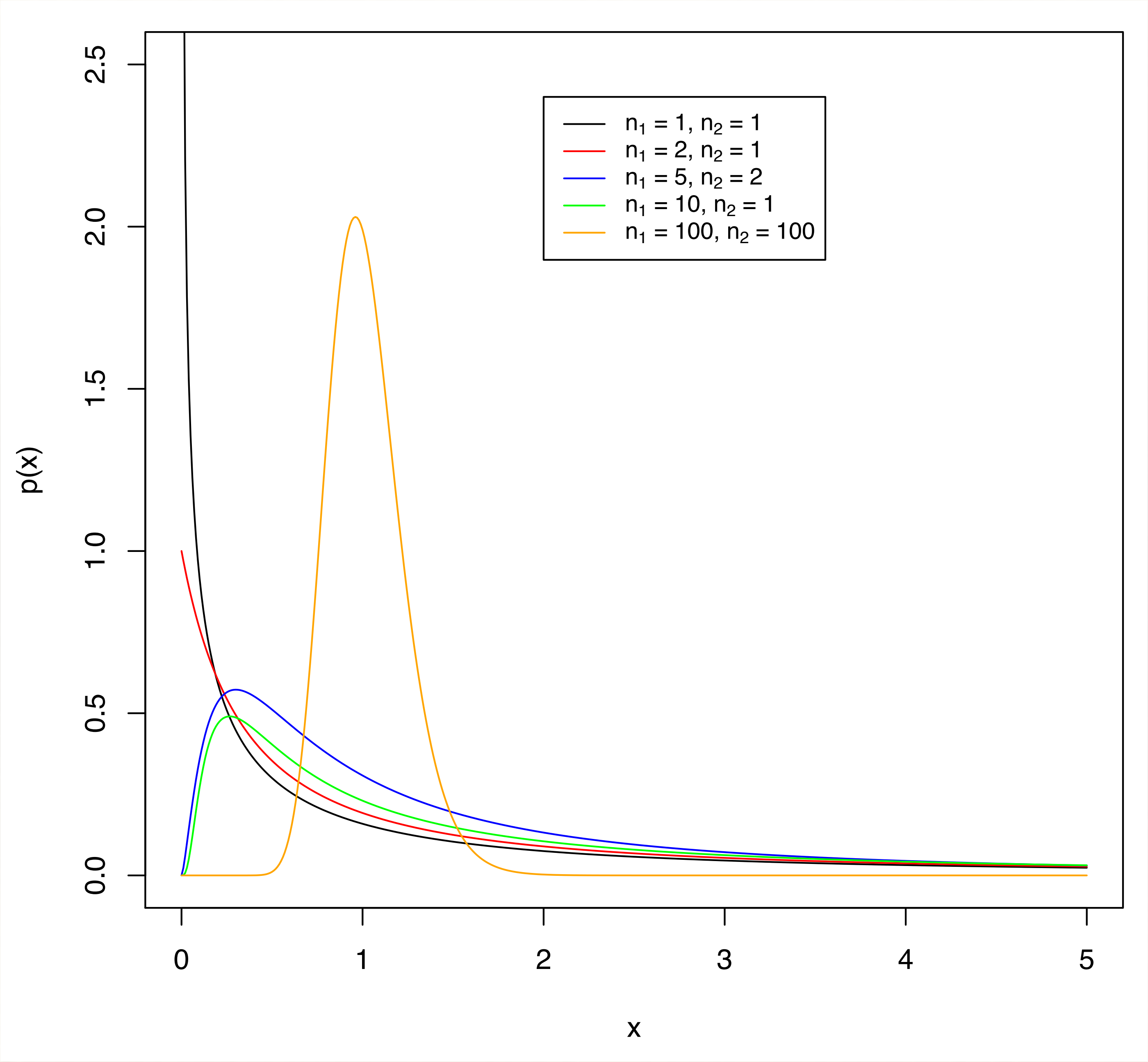


Continuous RVs: F-distribution

- The **F-distribution** depends on two degrees of freedom parameters, n_1 and n_2

$$p(x) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}$$

- This arises as the ratio of two chi-square distributions and is the basis for *analysis of variance*.

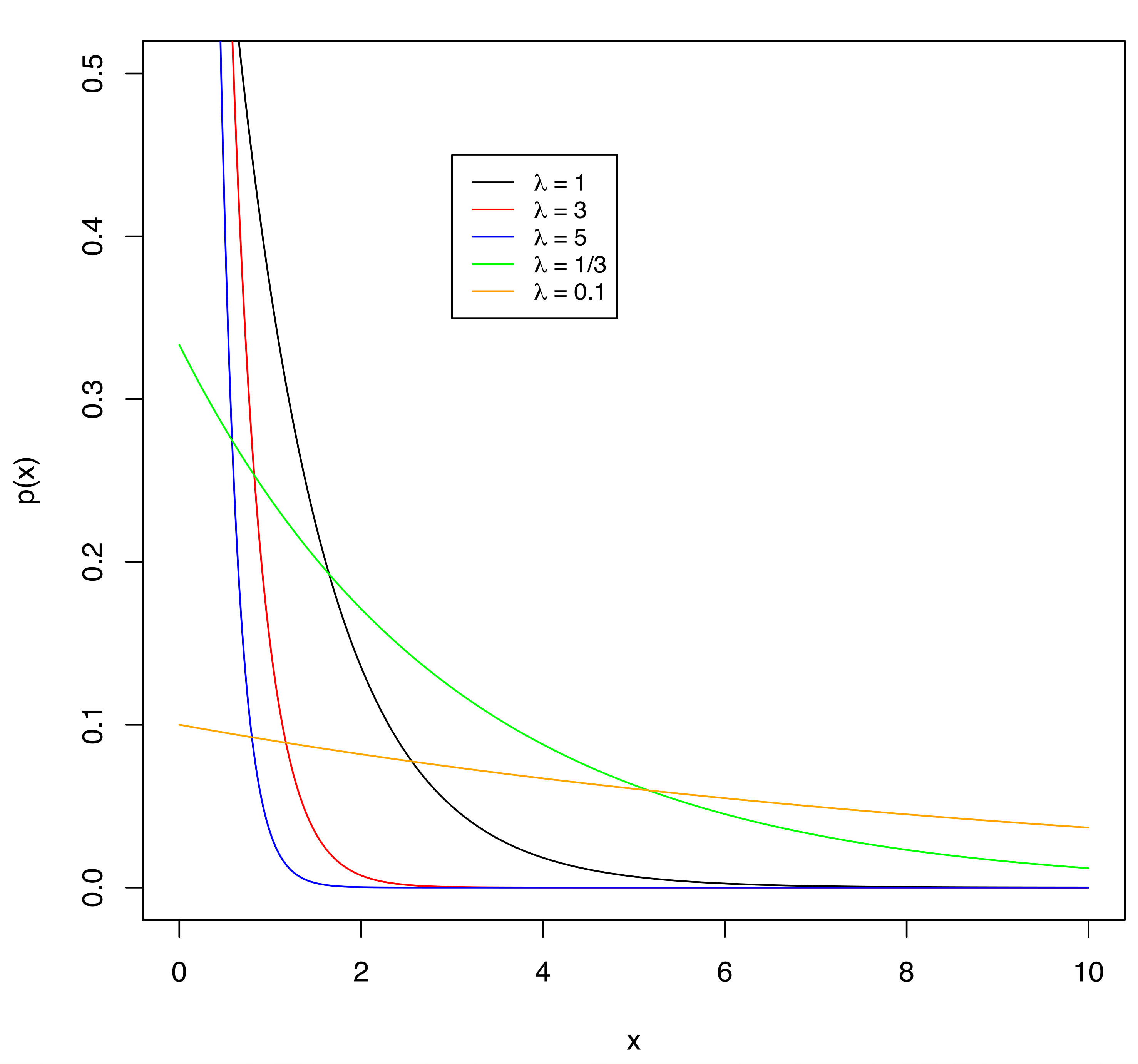


Continuous RVs: Exponential distribution

- The **Exponential distribution** depends on a *rate* parameter $\lambda > 0$

$$p(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- This arises as the distribution of the separation of events in a Poisson process.

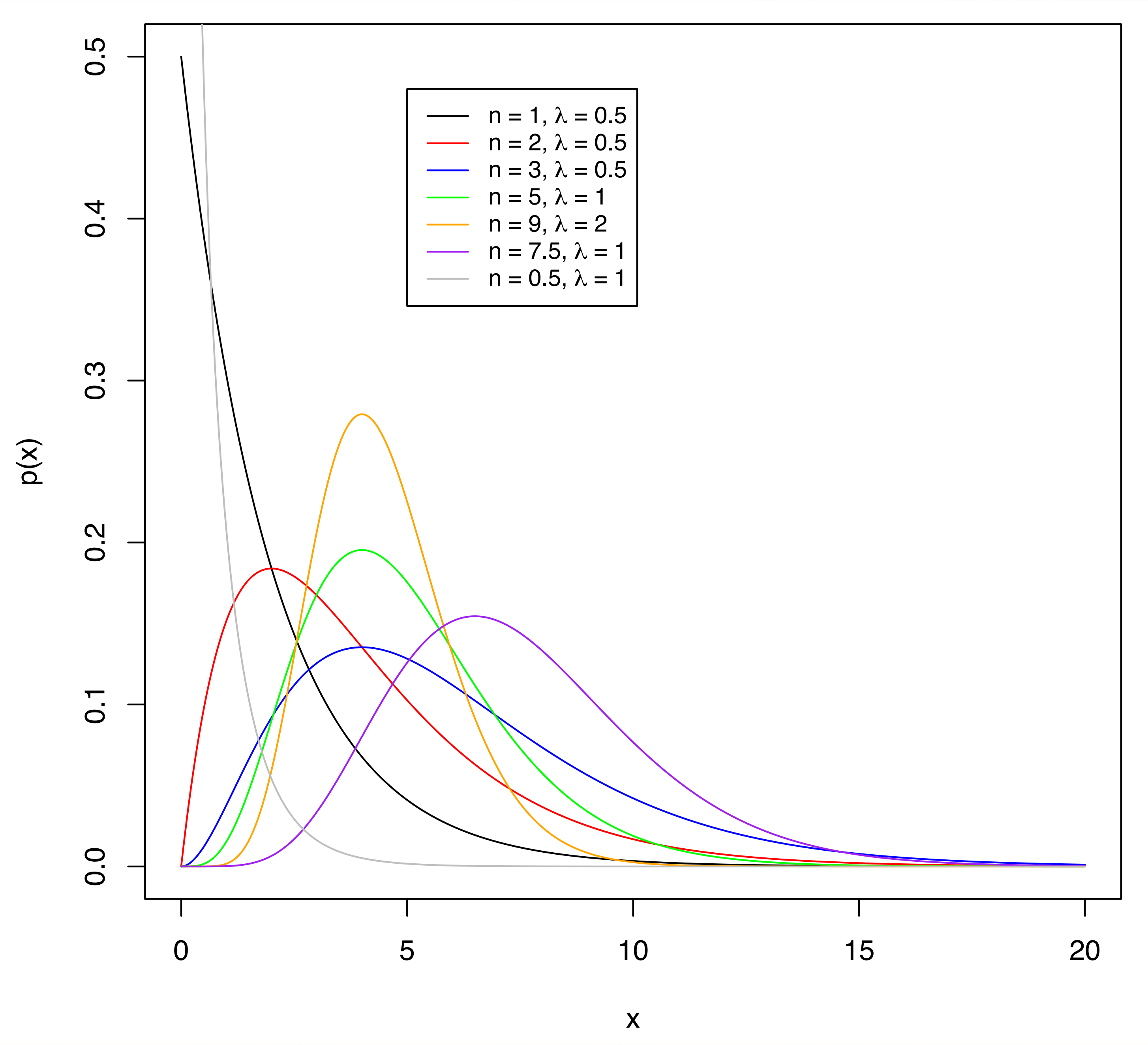


Continuous RVs: Gamma distribution

- The **Gamma distribution** depends on a *shape parameter* n > 0 and a *scale parameter* $\lambda > 0$

$$p(x) = \begin{cases} \frac{1}{\Gamma(n)}\lambda^n x^{n-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- The Gamma distribution is commonly used in Bayesian inference as a prior with support on the positive real line, and is conjugate to the Poisson distribution.



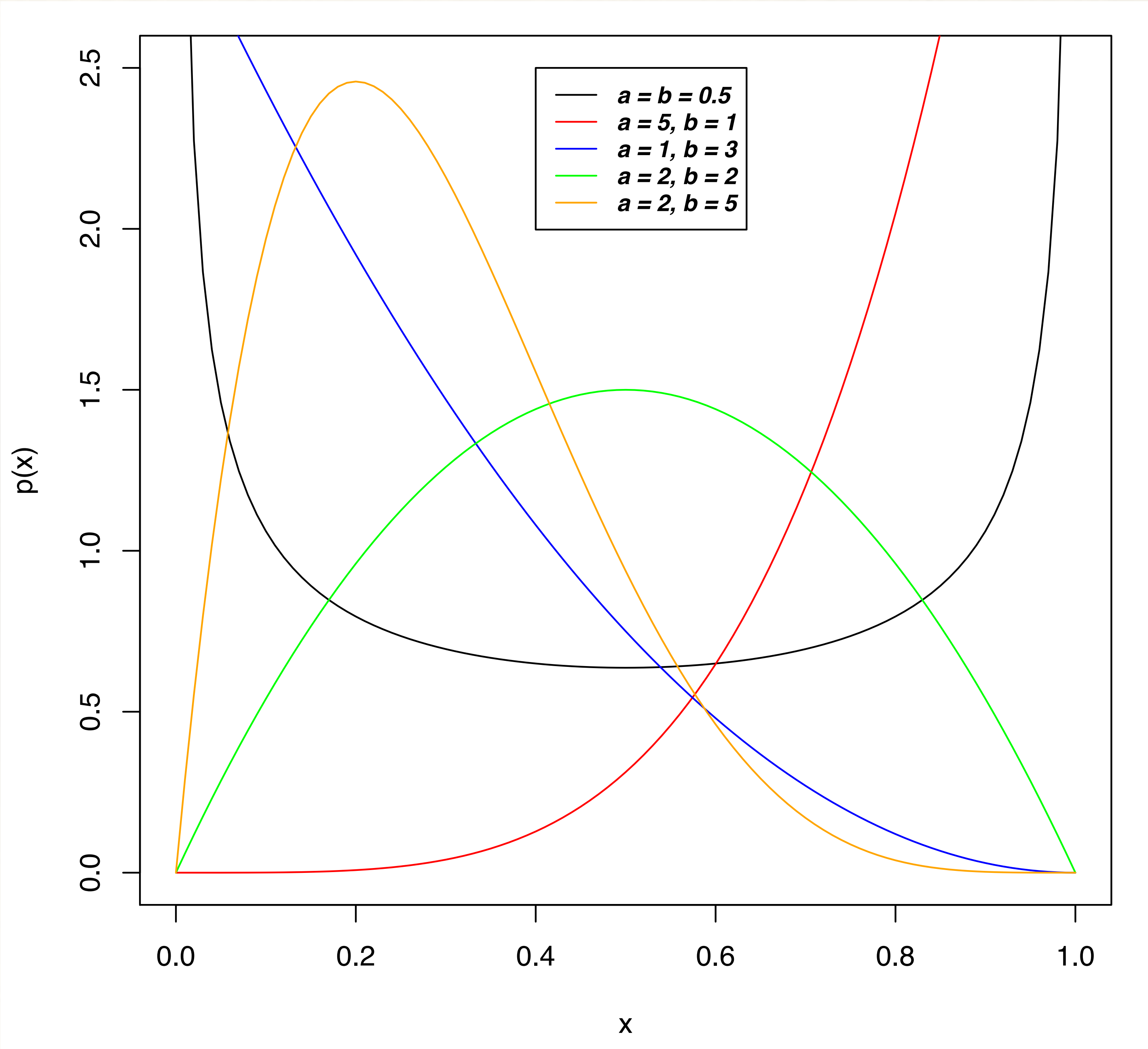
Continuous RVs: Beta distribution

- The **Beta distribution** depends on two *shape parameters* a, b > 0

$$p(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} \mathrm{d}x$$

- The Beta distribution is conjugate to the Binomial distribution and is used as a prior for parameters with support in [0,1].



Continuous RVs: Dirichlet distribution

- The **Dirichlet distribution** is a *multivariate distribution*, generating K samples $\{x_i\}$ constrained such that $0 < x_i < 1$ and

$$\sum_{i=1}^{K} x_i = 1$$

- The distribution depends on a vector of *concentration parameters*

$$\vec{\alpha} = (\alpha_1, \ldots, \alpha_K)$$

- and has pdf

$$p(x) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \qquad \text{where } B(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)}.$$

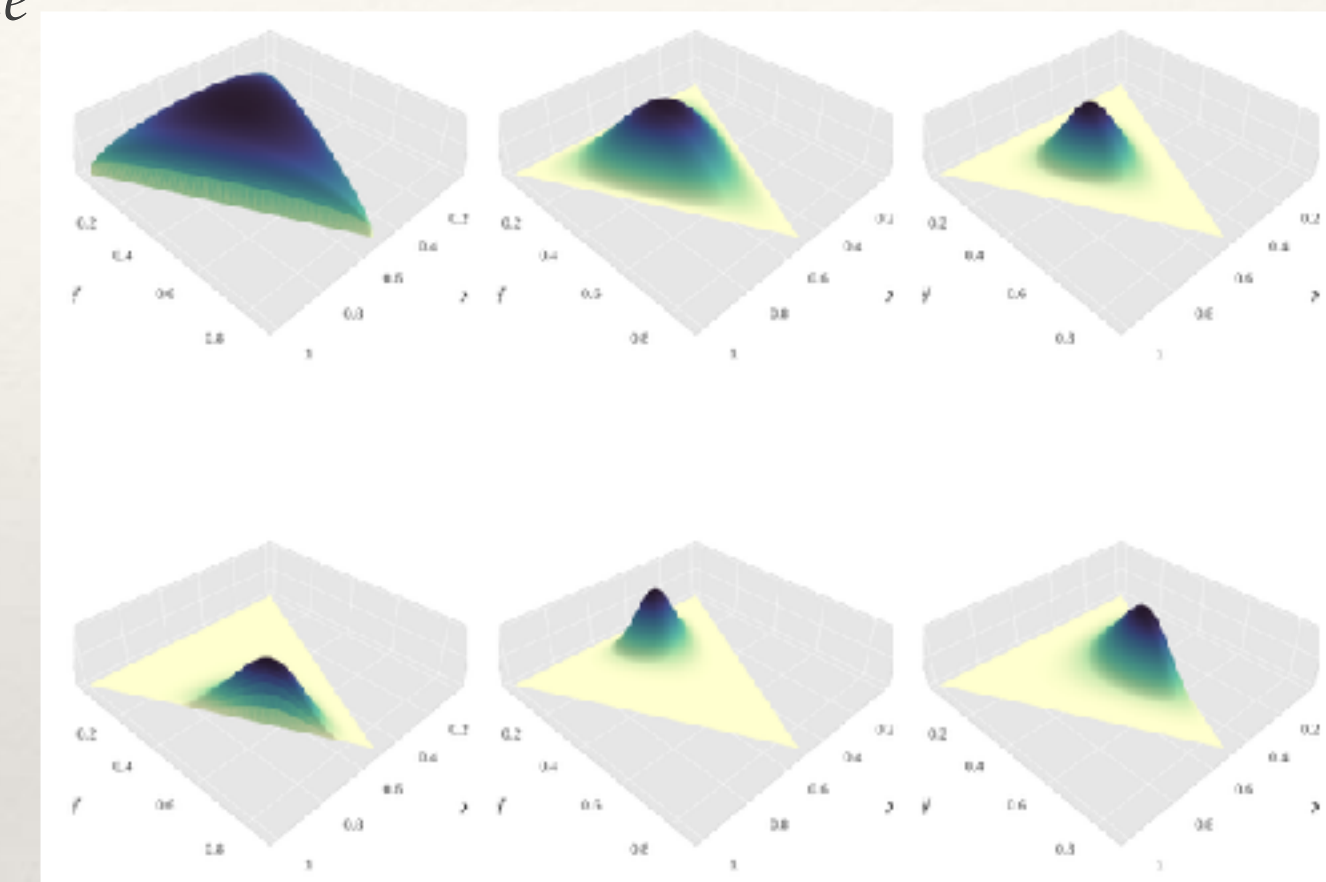


Figure from *Wikipedia*

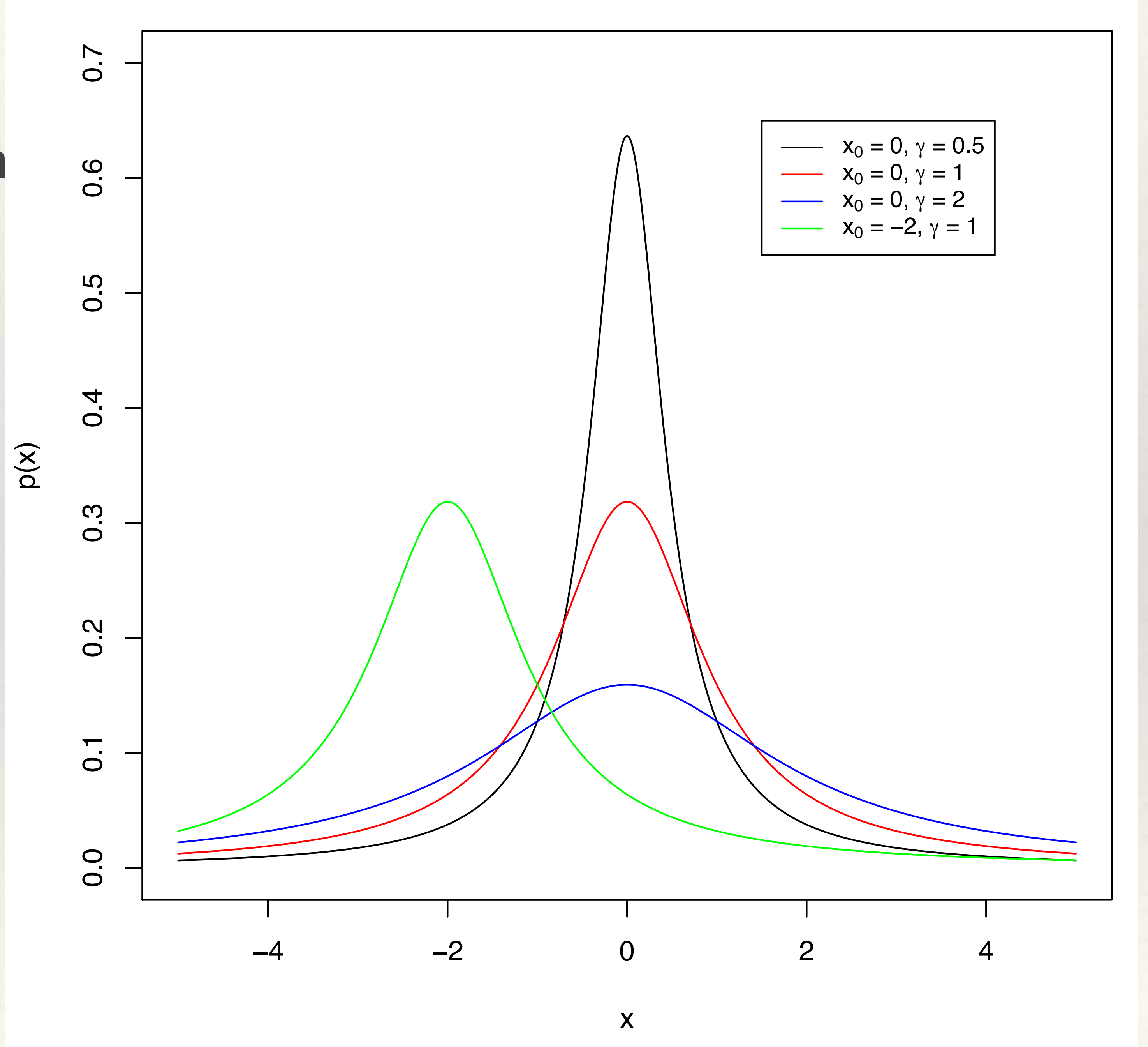
- The **Dirichlet process** is used as a prior on probability distributions in Bayesian nonparametric inference.

Continuous RVs: Cauchy distribution

- The **Cauchy distribution** (or Lorentz distribution) depends on a *location parameter*, x_0, and a *scale parameter*, $\gamma > 0$

$$p(x) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$$

- This distribution arises in optics and is used to model distributions with sharp features, e.g., spectral lines in LIGO.



Summarising random variables: average

- The pdf (or pmf) completely characterises a probability distribution, but it is often more convenient to work with summary quantities.
- These are based on *expectation values*

$$\mathbb{E}(T(X)) = \int_{-\infty}^{\infty} p(x) t(x) \mathrm{d}x$$

- There are various quantities that summarise the *average* value of a random variable
  - **Mean**

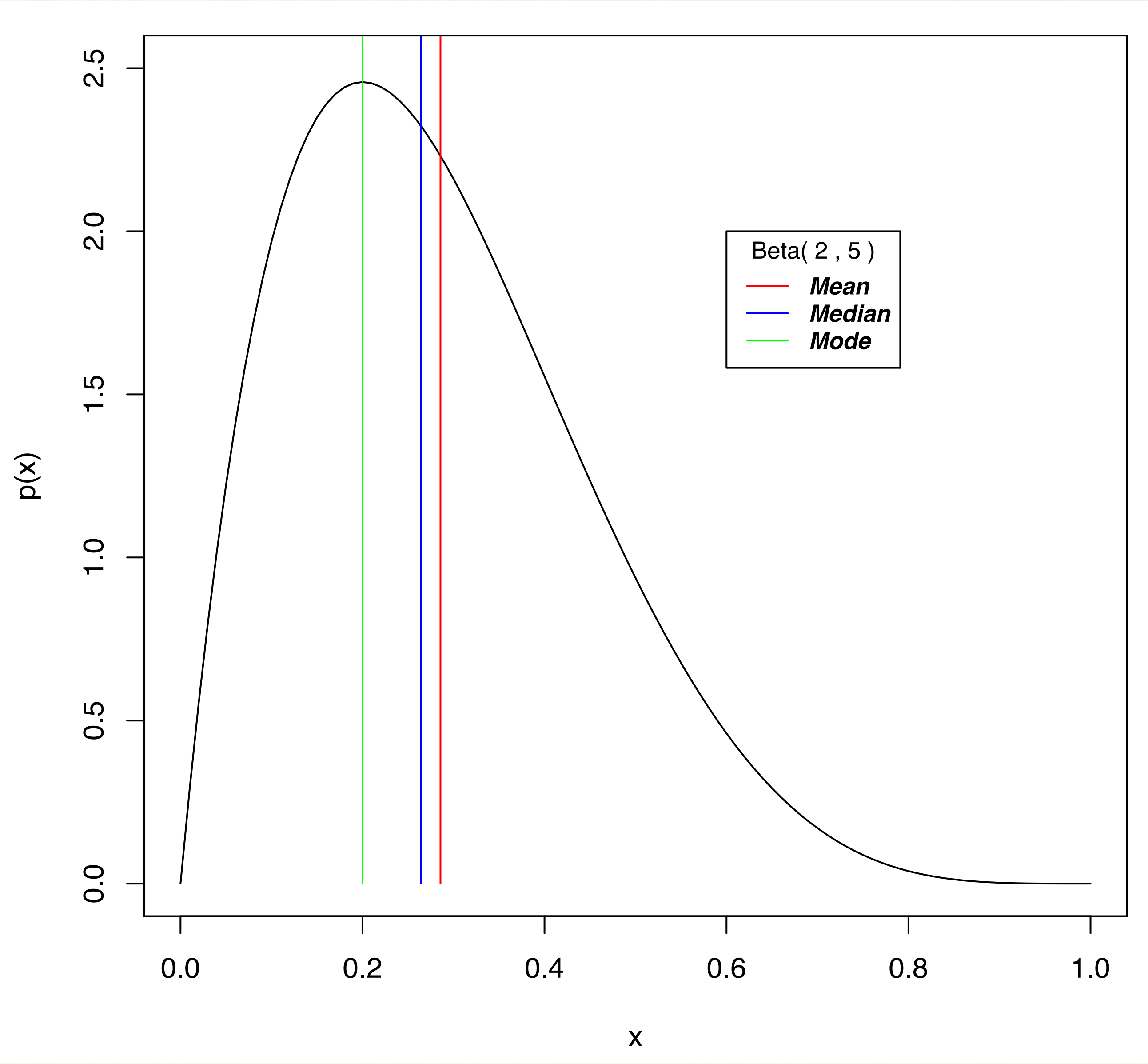
$$\mu = \mathbb{E}(X)$$

  - **Median** m satisfies

$$\int_{-\infty}^{m} p(x) \mathrm{d}x = \int_{m}^{\infty} p(x) \mathrm{d}x = \frac{1}{2}$$

  - **Mode**

$$M = \mathrm{argmax}_{x \in \mathcal{X}} p(x)$$



Summarising random variables: spread

- Other quantities summarise the spread of a RV
  - **Variance/Standard deviation**

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$$

  - **Skewness**

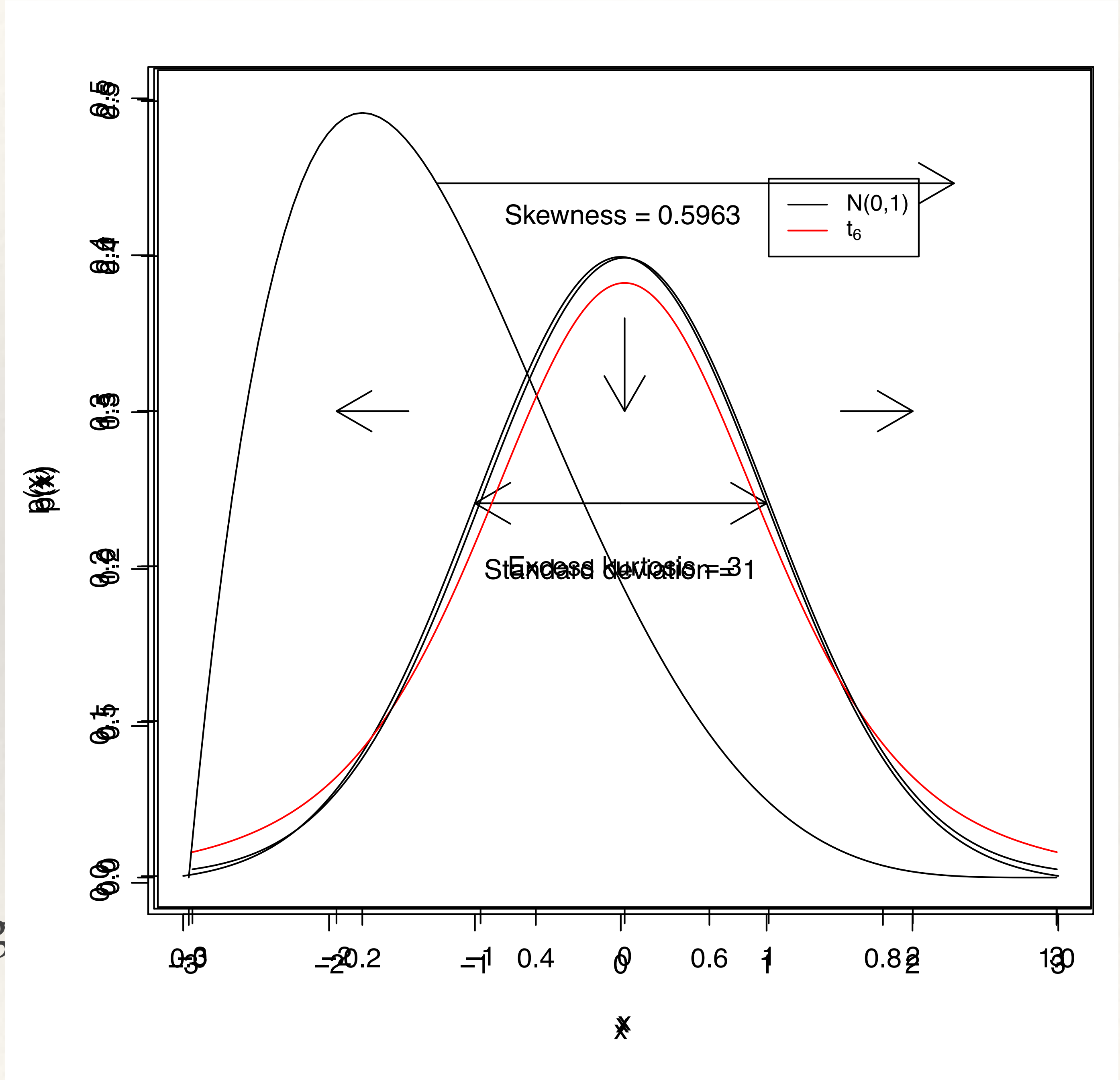
$$\gamma_1 = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

  - **Excess Kurtosis**

$$\mathrm{Kurt}(X) = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] - 3$$

  - **Higher moments**

$$\mathbb{E}\left[(X-c)^n\right]$$



- Moments can be efficiently computed using the *moment generating function*

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] \quad t \in \mathbb{R}$$

Independence

- A set of random variables $\{X_1, X_2, \ldots, X_N\}$ is *independent* if, for all choices of

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_N \leq x_N) = P(X_1 \leq x_1)P(X_1 \leq x_1) \ldots P(X_1 \leq x_1)$$

- In terms of the density function this is equivalent to

$$p(x_1, \ldots, x_N) = p_{X_1}(x_1)p_{X_2}(x_2) \ldots p_{X_N}(x_N)$$

- Two independent random variables have zero covariance

$$\mathrm{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))\,(Y - \mathbb{E}(Y))\right] = 0$$

- but the converse is not necessarily true.
- Random variables are *independent identically distributed* (IID) if they are independent and are all drawn from the same probability distribution.

Linear combinations of RVs

- Suppose $X_1, \ldots, X_N$ are random variables and consider a new RV

$$Y = \sum_{i=1}^{N} a_i X_i$$

- Y has the properties

$$\mathbb{E}(Y) = \sum_{i=1}^{N} a_i \mathbb{E}(X_i), \qquad \mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{cov}(X_i, X_j)$$

- The first equation holds for any random variables. If the RVs are *independent* then the relationships simplify

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) \qquad M_Y(t) = \prod_{i=1}^{N} M_{X_i}(a_i t)$$

- If $\{X_i\}$ are IID then the *sample mean* defined by $a_i = 1/N$ for all i has the properties

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(X_1), \qquad \mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}(X_1), \qquad M_{\hat{\mu}}(t) = \left(M_{X_1}\left(\frac{t}{N}\right)\right)^N$$

Laws of large numbers

- Averages of random variables have various nice asymptotic properties

$$S_n = \sum_{i=1}^{n} X_i \qquad \mathbb{E}(X) = \mu \qquad \mathrm{Var}(X) = \sigma^2$$

- *Weak law of large numbers*: for $\epsilon > 0$

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0, \text{ as } n \to \infty$$

- *Strong law of large numbers*

$$P\left(\frac{S_n}{n} \to \mu\right) = 1$$

- *Central Limit Theorem*: for $S_n^* = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$

$$\lim_{n\to\infty} P(a \leq S_n^* \leq b) = \Phi(b) - \Phi(a)$$