Introduction to probabilities

Guillermo Cabrera-Vives
guillecabrera@udec.cl

Sample spaces and events

- **Sample space** Ω: set of possible outcomes from an experiment.
- Points ω in Ω are called sample **outcomes**, **realizations** or **elements**.
- Subsets of Ω are called **events**.

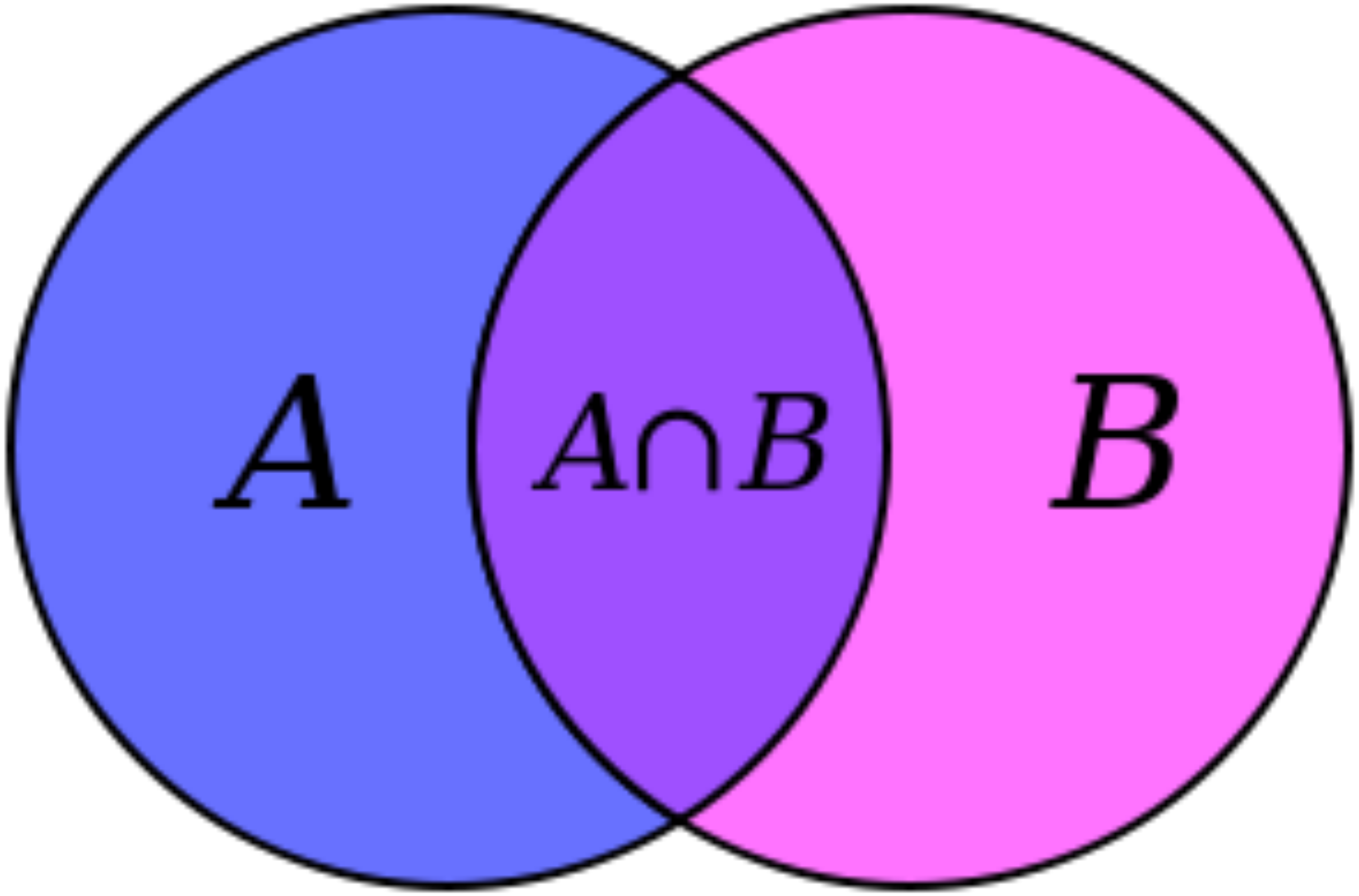
Examples

- If we toss a coin once, then $\Omega = \{H, T\}$. The event that the toss is heads is $A = \{H\}$
- If we toss a coin twice, then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$
- Let ω be the outcome of the measured temperature. Then $\Omega = (-\infty, \infty)$. The event that the temperature is larger than 10 but less or equal then 23 is $A = (10, 23]$.

Sample spaces and events

- Given an event A, let $A^c = \{\omega \text{ in } \Omega: \omega \text{ not in } A\}$ denote the complement of A.
- The complement of Ω is the empty set $\varnothing$.
- The union of events A and B is defined as
    - $A \cup B = \{\omega \text{ in } \Omega: \omega \text{ in } A \text{ or } \omega \text{ in } B \text{ or } \omega \text{ in both}\}$
- The intersection of events A and B is defined as
    - $A \cap B = \{\omega \text{ in } \Omega: \omega \text{ in } A \text{ and } \omega \text{ in } B\}$
- The difference $A - B = \{\omega \text{ in } \Omega: \omega \text{ in } A \text{ and } \omega \text{ not in } B\}$
- $|A|$ number of elements in A
- A and B are disjoint if $A \cap B = \varnothing$.

Sample spaces and events



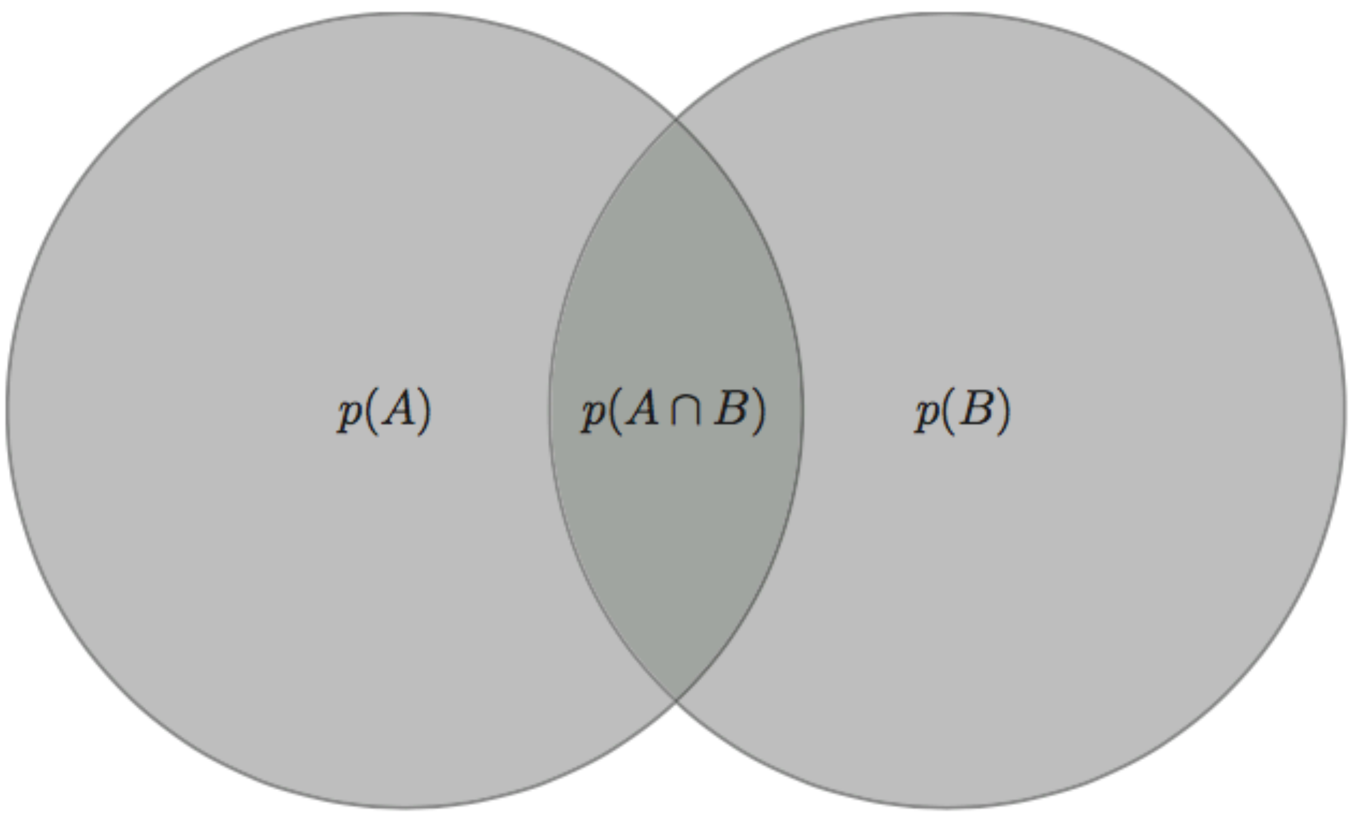
Probability axioms

- Given an event *A*, such as the outcome of a coin toss, we assign it a real number *p(A)*, called the **probability of *A*.**
- *p(A)* could also correspond to a probability that a value of *x* falls in a *dx* wide interval around *x*.
- To qualify as a probability, *p(A)* must satisfy three Kolmogorov axioms:

1. $p(A) \geq 0$ for each A.
2. $p(\Omega) = 1$, where Ω is a set of all possible outcomes.
3. If $A_1, A_2, \ldots$ are disjoint events, then $p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$, where $\bigcup$ stands for "union."

Probability properties

- As a consequence of these axioms, several useful rules can be derived. The probability that the union of two events, *A* and *B* , will happen is given by the sum rule,
    - $p(A \cup B) = p(A) + p(B) - p(A \cap B)$



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Probability properties

- If the complement of event A is Ac, then
  - $p(A) + p(Ac) = 1$
- The probability that both A and B will happen is equal to
  - $p(A \cap B) = p(A|B)\ p(B) = p(B|A)\ p(A).$
- Here "|" is pronounced "given" and p(A|B) is the probability of event A given that (conditional on) B is true.

Example: 3 faces dice

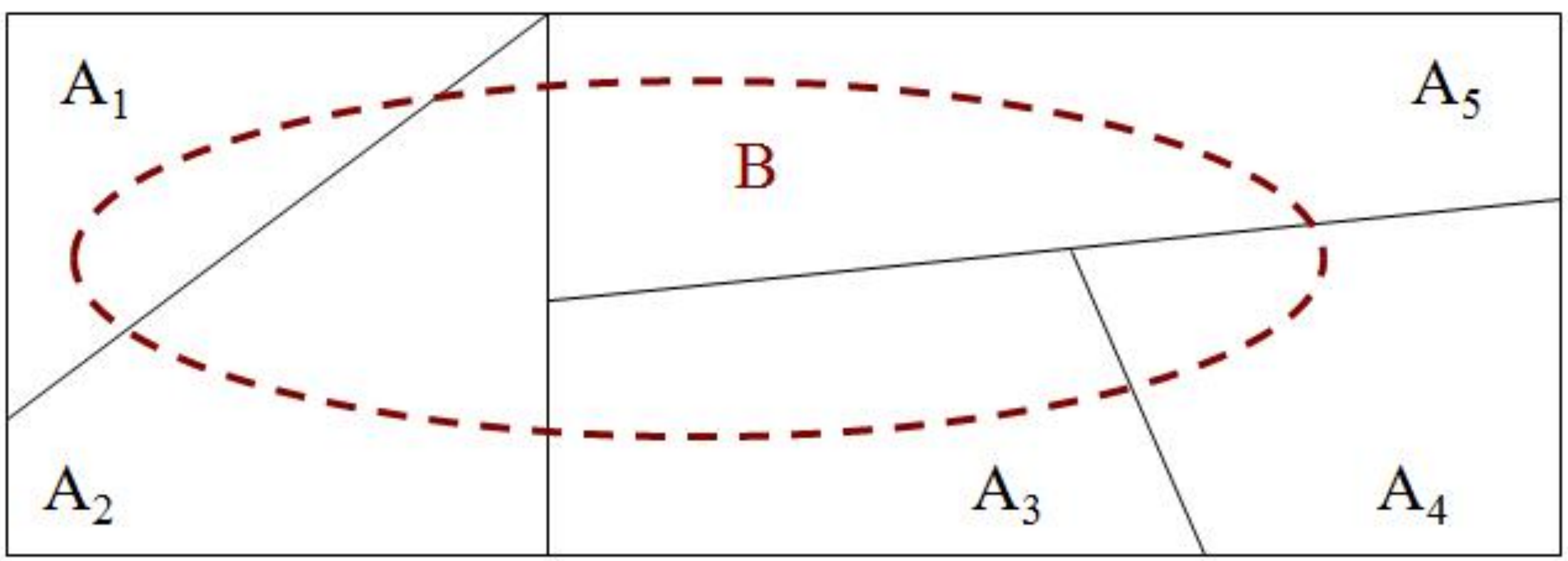
- Asume you throw two 3 faces dice
  - $\Omega = \{11, 12, 13, 21, 22, 23, 31, 32, 33\}$
  - $|\Omega| = 3 \mathrm{x} 3 = 9$
- $p(1) = p(2) = p(3) = 1/3$
- What is the probability of A = {getting a 1 in either dice}?
  - $P(A)=P(\{11, 12, 13, 21, 31\}) = 5/9$
- Another way:
  - $A = A1 \cup A2$, where A1 = {getting a 1 in first die}, A2 = {getting a 1 on second die}
  - $p(A1 \cup A2) = p(A1) + p(A2) - p(A1 \cap A2)$
  - $P(A1 \cup A2) = 1/3 + 1/3 - 1/9 = (3+3-1) / 9 = 5/9$

Example: 3 faces dice

- Note:
  - $p(A1 \cap A2) = p(A1|A2)\, p(A2)$
  - $p(A1 \cap A2) = p(A1)\, p(A2)$ independent variables!
  - $p(A1 \cap A2) = 1/3 \times 1/3 = 1/9$

Law of total probabilities

- If events A_i, $i = 1,\ldots,N$ are disjoint and their union is the set of all possible outcomes, then
- $p(B) = \Sigma_i \, p(A_i \cap B) = \Sigma_i \, p(B|A_i) \, p(A_i)$



Law of total probabilities

- Assuming that an event C is not mutually exclusive with A or any of B_i, then
  - $p(A|C) = \Sigma_i \, p(A|C \cap B_i) \, p(B_i|C)$

Bayes theorem

- recall $p(A \cap B) = p(A|B)\, p(B) = p(B|A)\, p(A)$

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- *Note:*
  - $p(B) = \Sigma_i\, p(A_i \cap B) = \Sigma_i\, p(B|A_i)\, p(A_i)$

Example: the Monty Hall problem

- There are N=3 doors, of which 2 are empty and one contains some “prize.”
- You choose a box at random; the probability that it contains the prize is 1/3. This door remains closed.
- Then the host who knows which door contains the prize opens 1 empty door chosen from the 2 remaining doors.
- You are offered to switch the door you initially chose with other unopened door.
- Would you do it?

Example: the Monty Hall problem

- Event C_i = the prize (car) is behind door i.
- Say, X_1 = you choose door 1.
- As where the car is is independent of your choice, $p(C_i \mid X_1) = 1/3$
- Say the host opens door 3 and is empty, H_3.
    - $p(H_3|C_1, X_1) = 1/2$
    - $p(H_3|C_2, X_1) = 1$
    - $p(H_3|C_3, X_1) = 0$

Example: the Monty Hall problem

- If you change door, the probability of getting the prize is
  - $p(C_2 \mid H_3, X_1) = [p(H_3 \mid C_2, X_1)\ p(C_2 \cap X_1)] / p(H_3 \cap X_1)$
  - $p(C_2 \mid H_3, X_1) = 2/3$

Continuous variables

- Let T be the outcome of the measured temperature. What is the probability of $T = 25^{\circ}$?
- What is the number of outcomes?
- ∞
- It makes no sense!!
- It makes more sense to calculate the probability of the temperature to fall within a specific range.

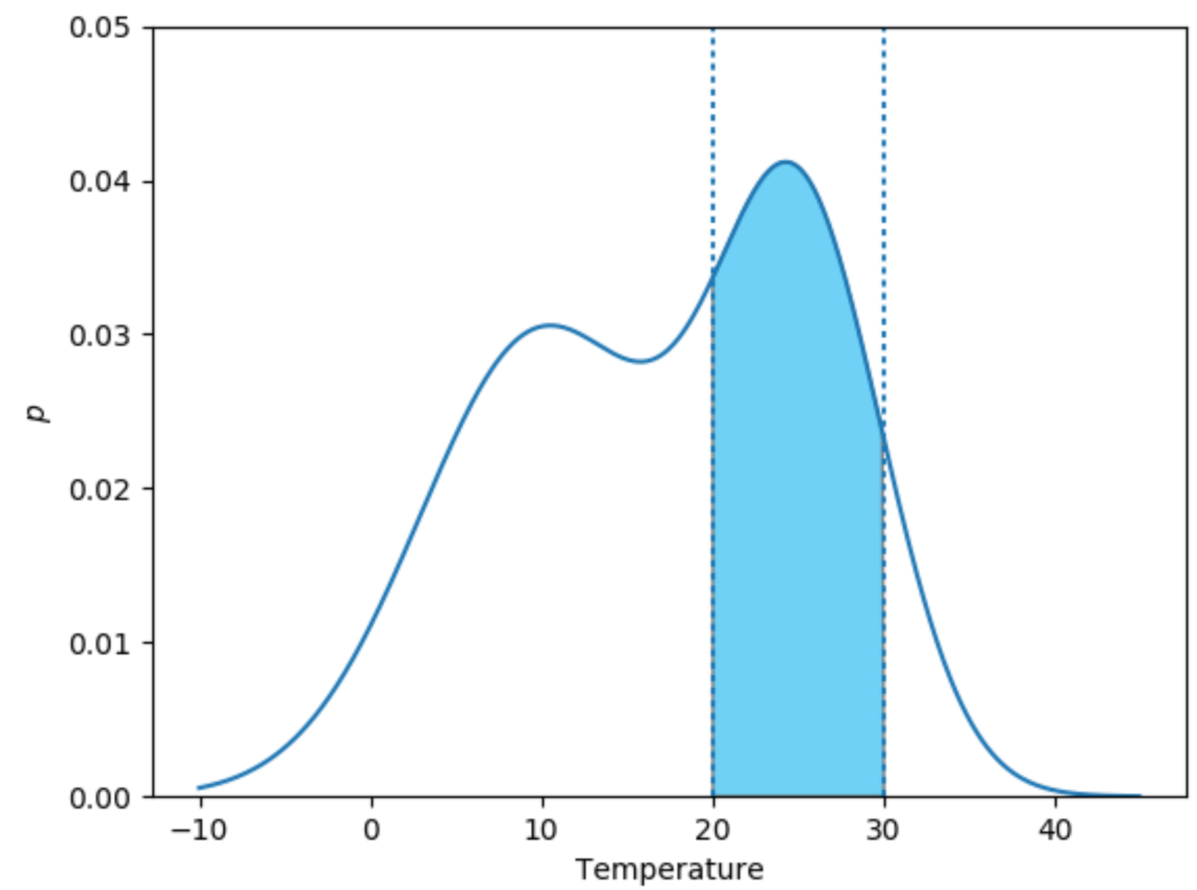
Probability density function (PDF)

- The PDF *p(x)* is used to specify the probability of the random variable falling within a particular range of values.

$$P(a \leq x \leq b) = \int_a^b p(x)dx$$

- What is the probability of 20<T<30?

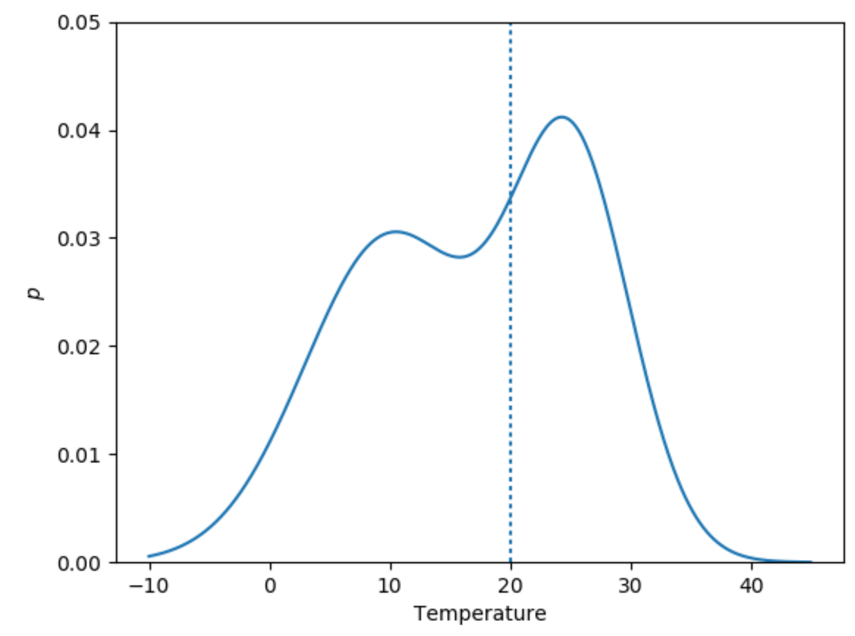
$$P(20 \leq T \leq 30) = \int_{20}^{30} p(x)dx$$



Cumulative distribution and p-value

- Cumulative distribution

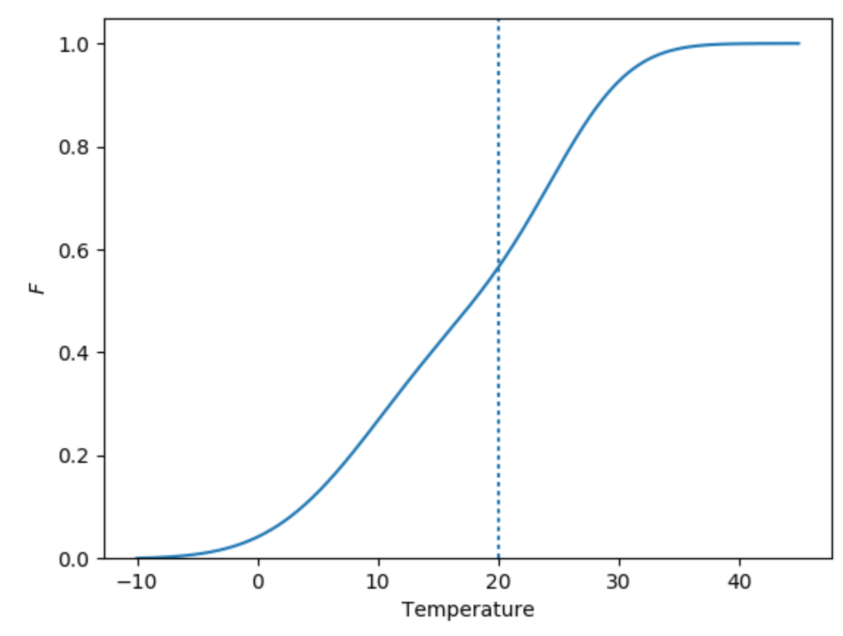
$$F(x) = P(x' \leq x) = \int_{-\infty}^{x} p(x')dx'$$



- Note: $F(\infty) = 1$

- p-value

$$P(x' > x) = \int_{x}^{\infty} p(x')dx'$$



Expectation

- Expected value (or average value) of *f*:
  - Discrete: $$\mathbb{E}[f] = \sum_x p(x) f(x)$$
  - Continuous: $$\mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x$$
  - Finite number of points: $$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$
- Conditional expectation: $$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$
- Variance: $$\begin{aligned} \mathrm{var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \end{aligned}$$

Covariance

- Covariance: expresses the extent to which x and y vary together.

$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x-\mathbb{E}[x]\}\{y-\mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy]-\mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

- For two vectors of random variables:

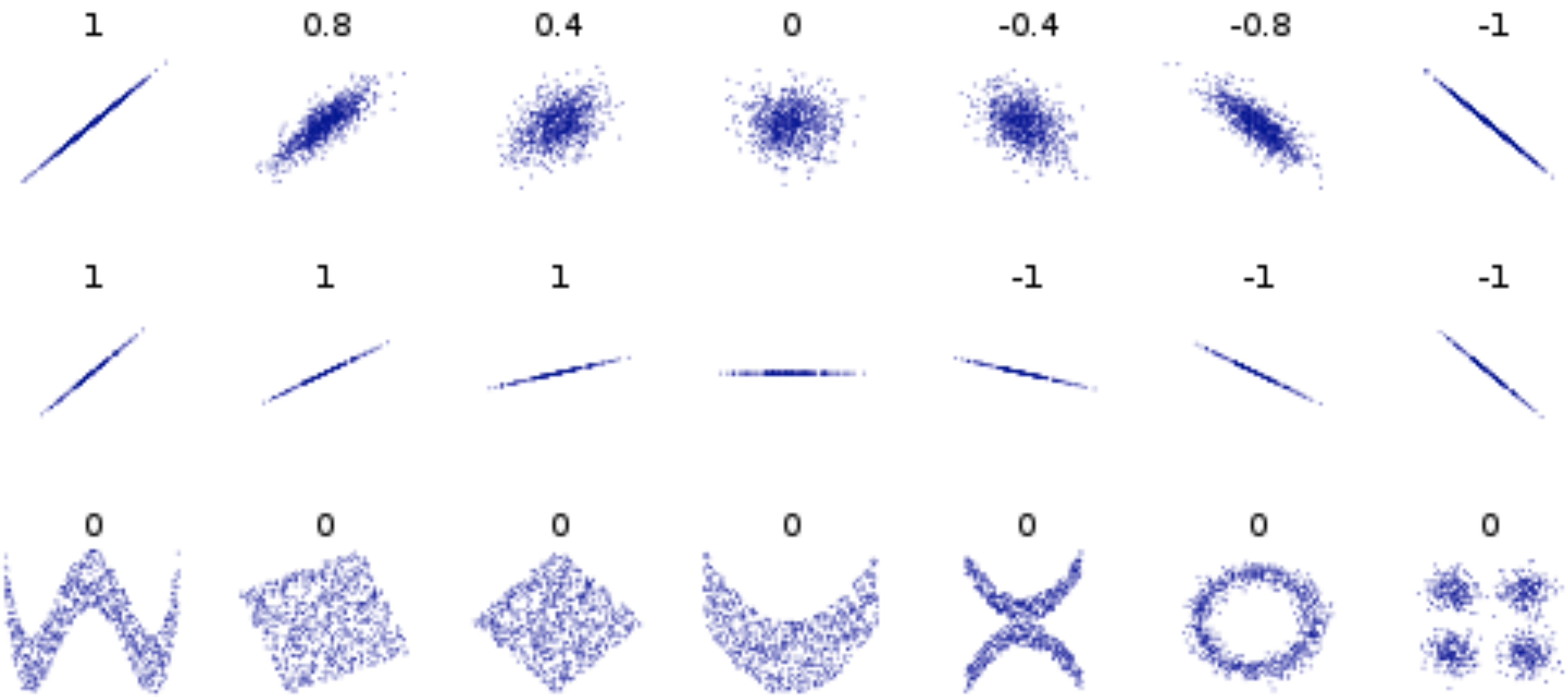
$$\begin{aligned}\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x}-\mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}}-\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}]-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}].\end{aligned}$$

$$\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x},\mathbf{x}]$$

- Correlation:

$$\rho \equiv \mathrm{corr}[x,y] = \frac{\mathrm{cov}[x,y]}{\sqrt{\mathrm{var}[x]\mathrm{var}[y]}}$$

Covariance / Correlation



Frequentist vs Bayesian

- **Frequentist:** view probabilities in terms of the frequencies of random, repeatable events.
- **Bayesian:** probabilities provide a quantification of uncertainty.
- **Example:** I have misplaced my phone somewhere in the home. I can use the phone locator on the base of the instrument to locate the phone and when I press the phone locator the phone starts beeping.Which area of my home should I search?
- **Frequentist Reasoning:** I can hear the phone beeping. I also have a mental model which helps me identify the area from which the sound is coming. Therefore, upon hearing the beep, I infer the area of my home I must search to locate the phone.
- **Bayesian Reasoning:** I can hear the phone beeping. Now, apart from a mental model which helps me identify the area from which the sound is coming from, I also know the locations where I have misplaced the phone in the past. So, I combine my inferences using the beeps and my prior information about the locations I have misplaced the phone in the past to identify an area I must search to locate the phone.

https://stats.stackexchange.com/questions/22/bayesian-and-frequentist-reasoning-in-plain-english

Frequentist vs Bayesian

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Some known probability distributions

Most of the images borrowed from www.wikipedia.com

Uniform Distribution

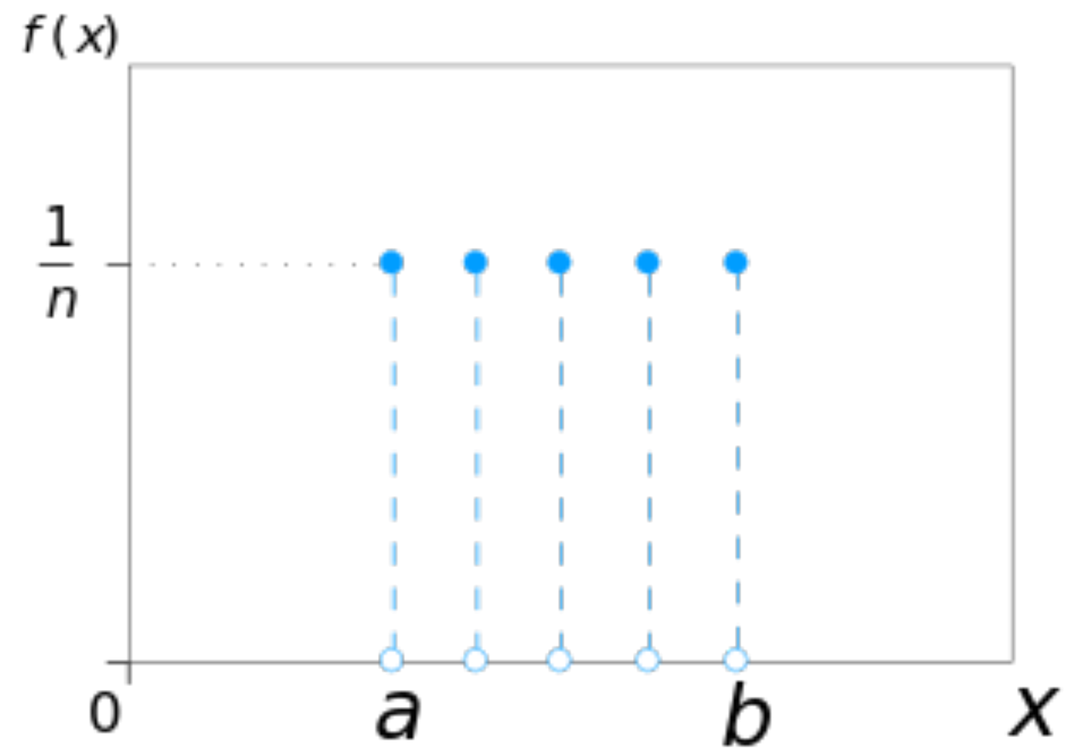
- **Discrete:** a finite number of values are equally likely to be observed.

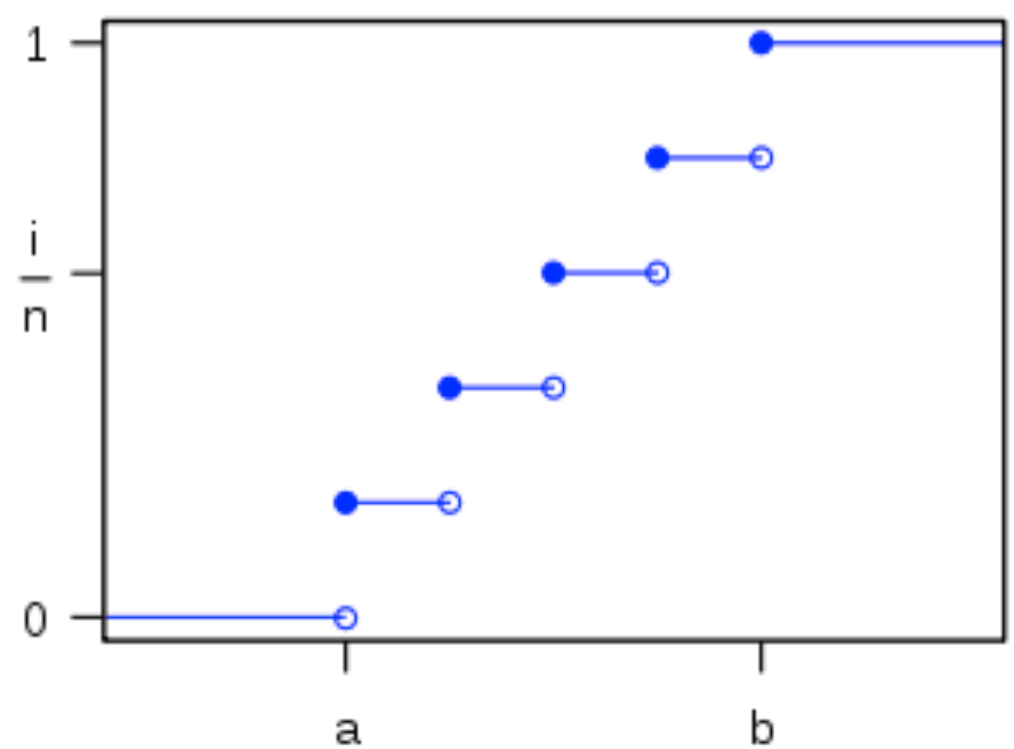
$$P(k) = \frac{1}{n}$$

- Cumulative distribution:

$$F(k;a,b) = \frac{\lfloor k \rfloor - a + 1}{b - a + 1}$$

- Mean: (a+b)/2, variance: $[(b-a+1)^2-1]/12$





Uniform Distribution

- **Continuous:** For each member of the family, all intervals of the same length on the distribution's support are equally probable.

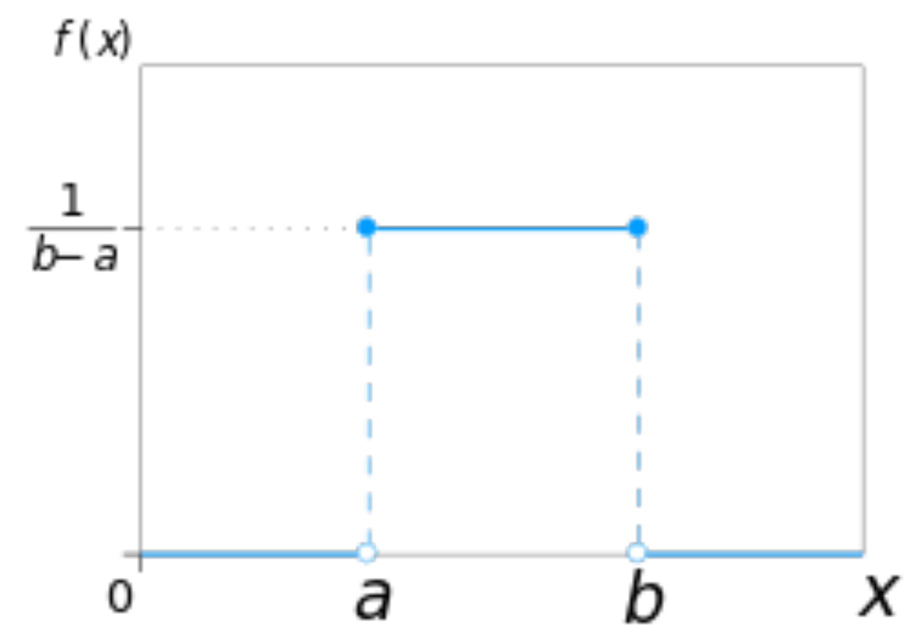
- PDF:

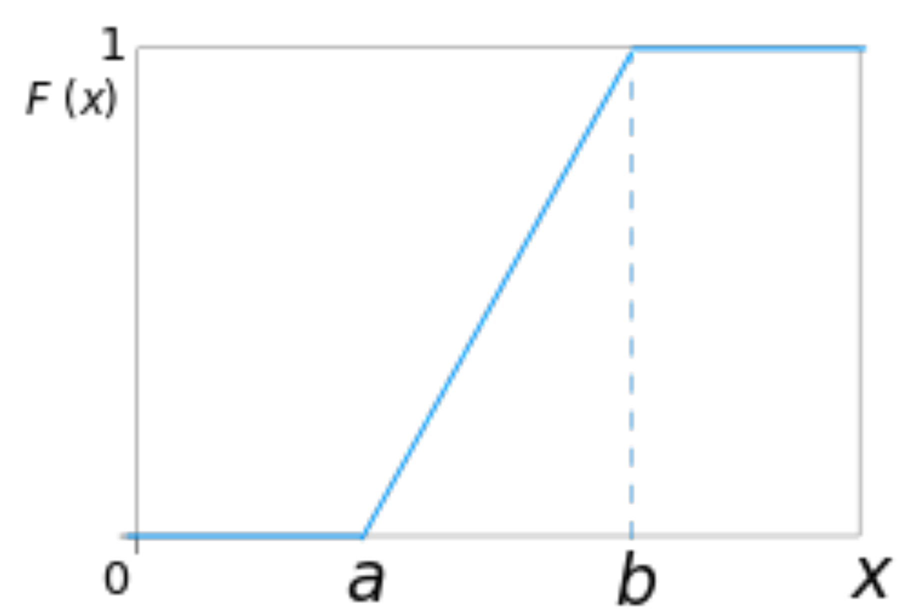
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\[8pt] 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

- Cumulative distribution:

$$F(x) = \begin{cases} 0 & \text{for } x < a \\[8pt] \frac{x-a}{b-a} & \text{for } a \le x \le b \\[8pt] 1 & \text{for } x > b \end{cases}$$

- Mean: (a+b)/2, variance: $(b-a)^2/12$

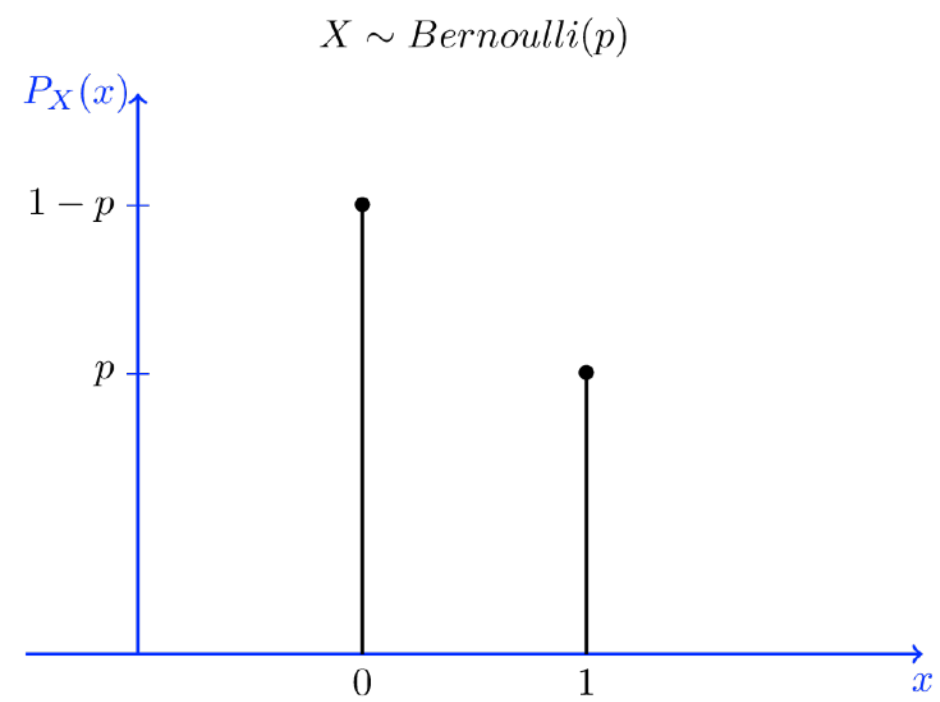




Bernoulli Distribution

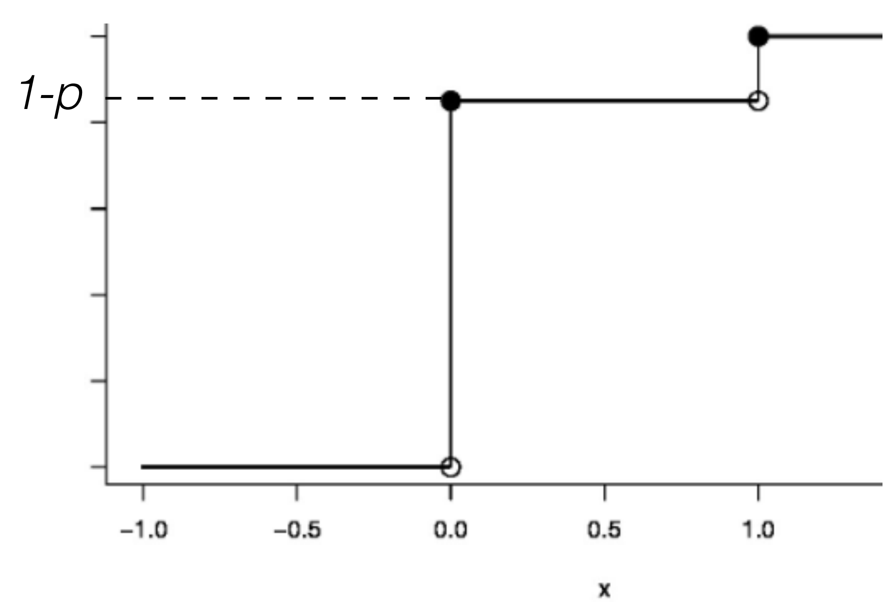
- Probability distribution of a random variable which takes the value 1 with probability *p* and the value 0 with probability *q = 1 - p*,

$$P(x;p) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \end{cases}$$



- Cumulative distribution:

F(x;p)



$$F(x;p) = \begin{cases} 0 & \text{for } x < 0 \\ 1-p & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } \leq x \geq 1 \end{cases}$$

- Mean: *p*, var: *p(1-p)*

Binomial Distribution

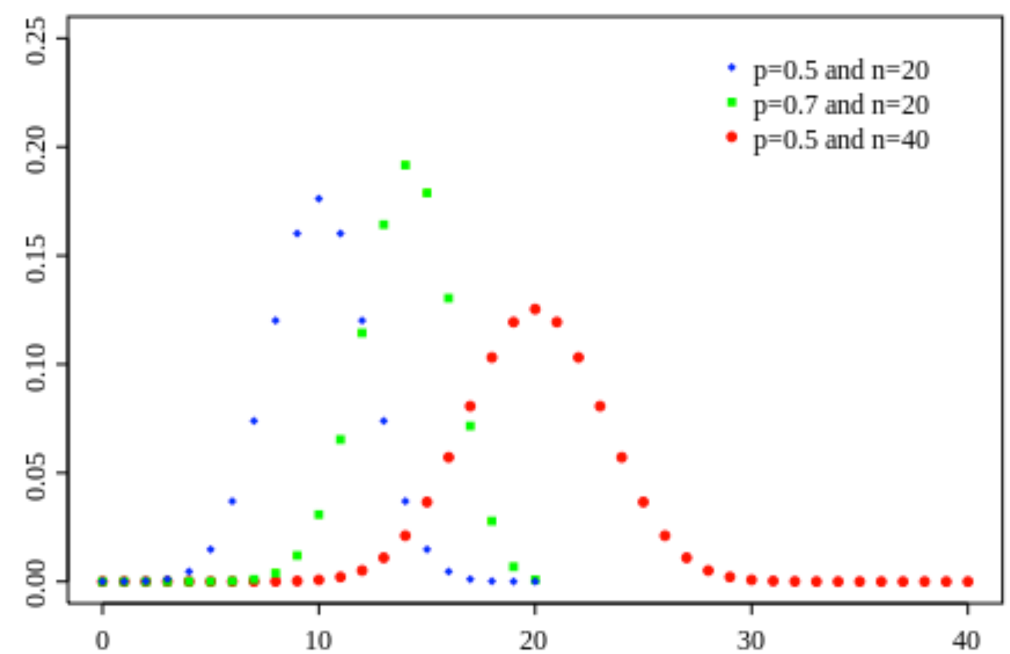
- Discrete probability distribution of the number of successes in a sequence of *n* independent experiments, each asking a yes–no question.

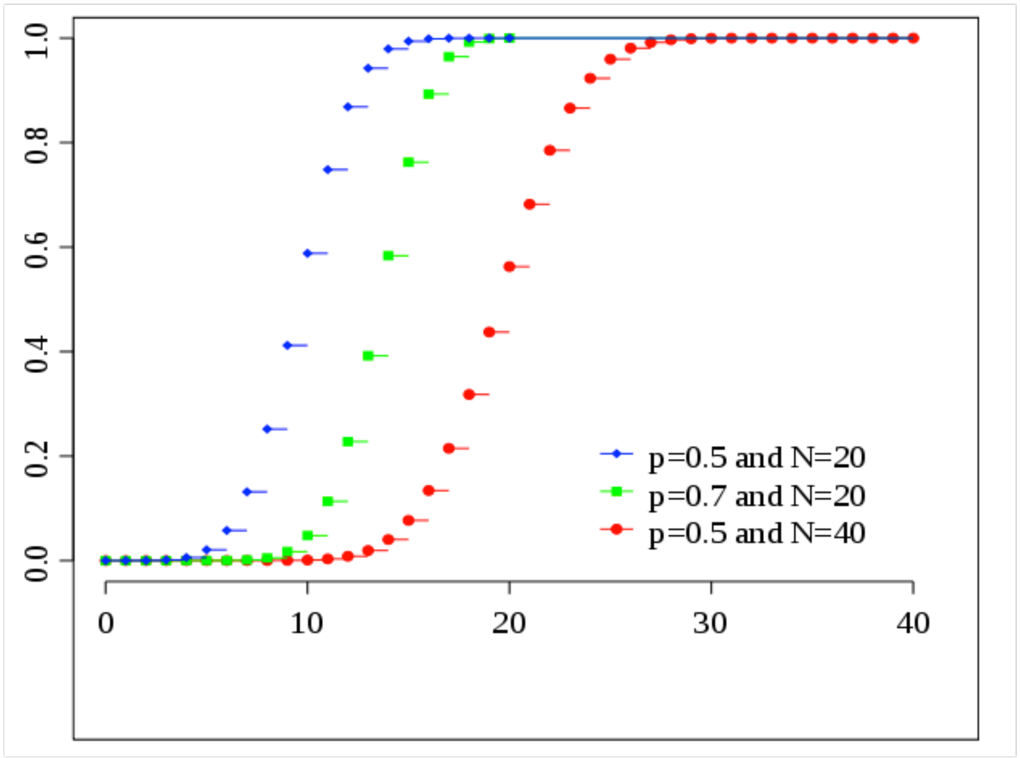
$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \qquad \binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$$

- Cumulative distribution:

$$F(m|N,\mu) = \sum_{i=0}^{\lfloor m \rfloor} \binom{N}{i} \mu^m (1-\mu)^{N-m}$$

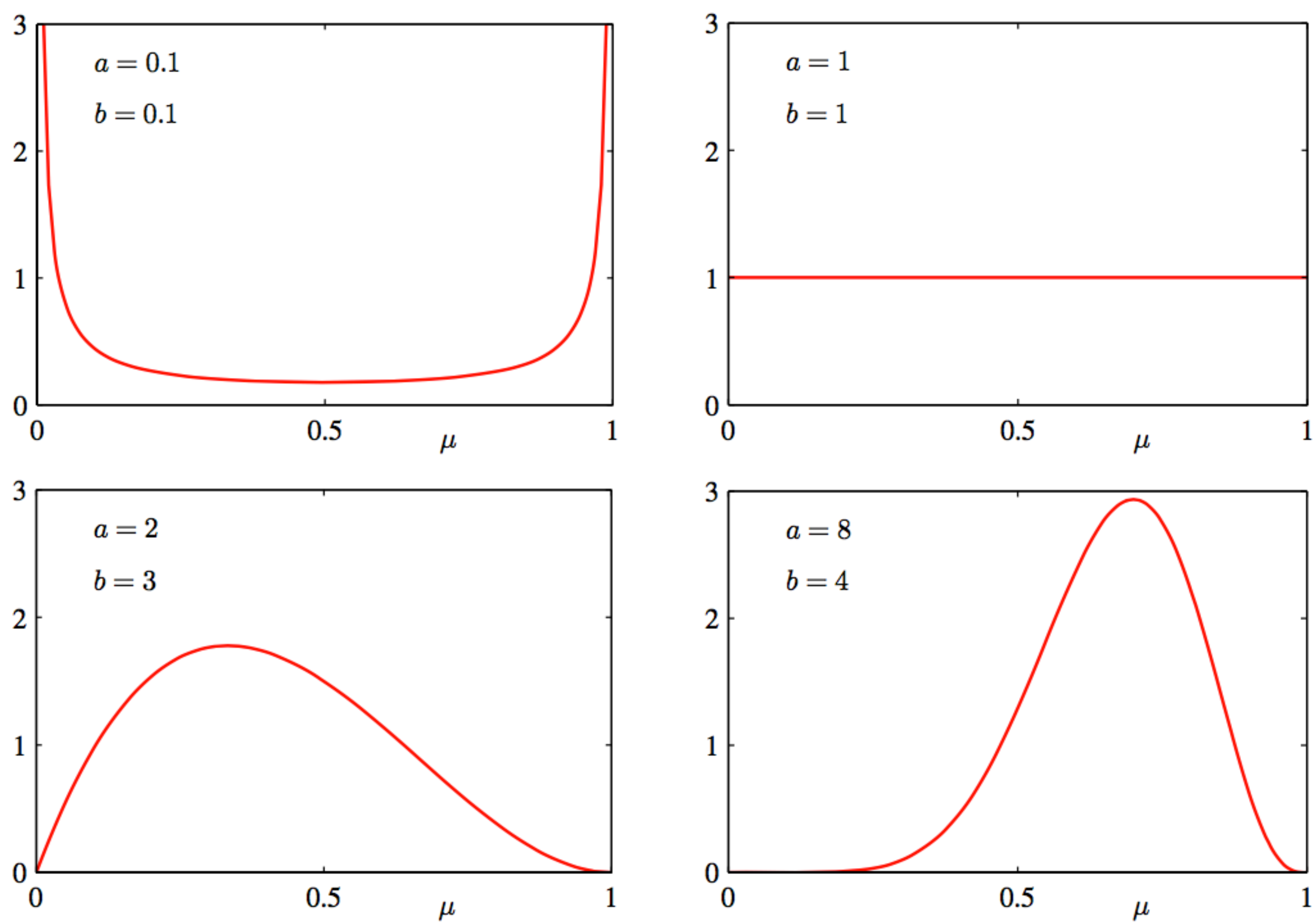
- I_{1-p}: regularized incomplete beta function
- Mean: $N\mu$, var: $N\mu(1-\mu)$





The Beta Distribution

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$



The Beta Distribution

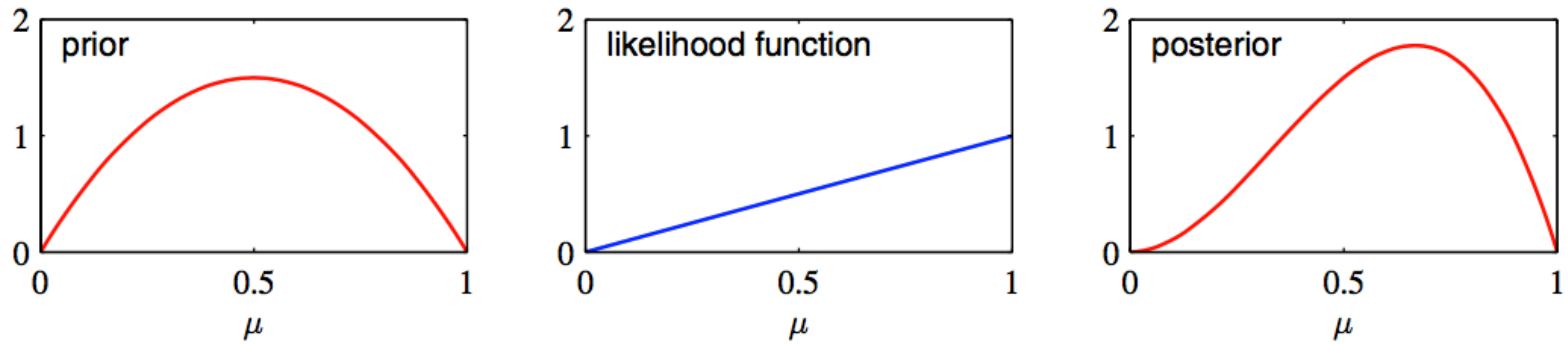
- **Conjugate prior** for the Binomial distribution: the posterior will have the same functional form as the prior.

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}.$$

- observing a data set of m observations of x = 1 and l observations of x = 0 increases the value of a by m, and the value of b by l, in going from the prior distribution to the posterior distribution.
- a and b in the prior is an effective number of observations of x = 1 and x = 0, respectively.

The Beta Distribution

- Consider a prior given by a beta distribution with parameters a = 2, b = 2, and the likelihood function, given by (2.9) with N = m = 1, corresponds to a single observation of x = 1, so that the posterior is given by a beta distribution with parameters a = 3, b = 2.



$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})\,\mathrm{d}\mu = \int_0^1 \mu p(\mu|\mathcal{D})\,\mathrm{d}\mu = \mathbb{E}[\mu|\mathcal{D}].$$

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b}$$

Geometric Distribution

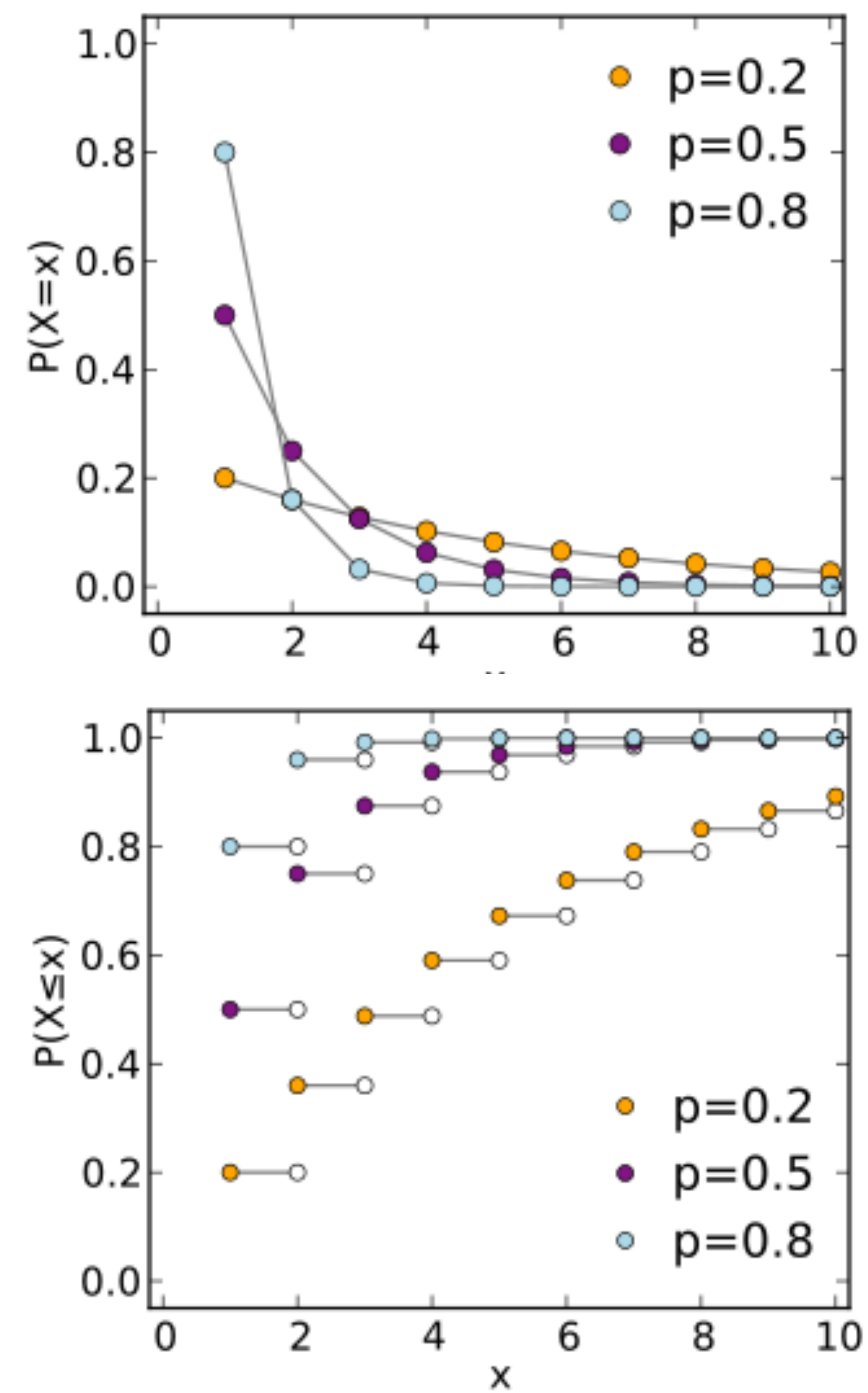
- The probability distribution of the number *x* of Bernoulli trials needed to get one success

$$P(x;p) = (1-p)^{x-1}p$$

- Cumulative distribution:

$$F(x;p) = 1 - (1-p)^{x}$$

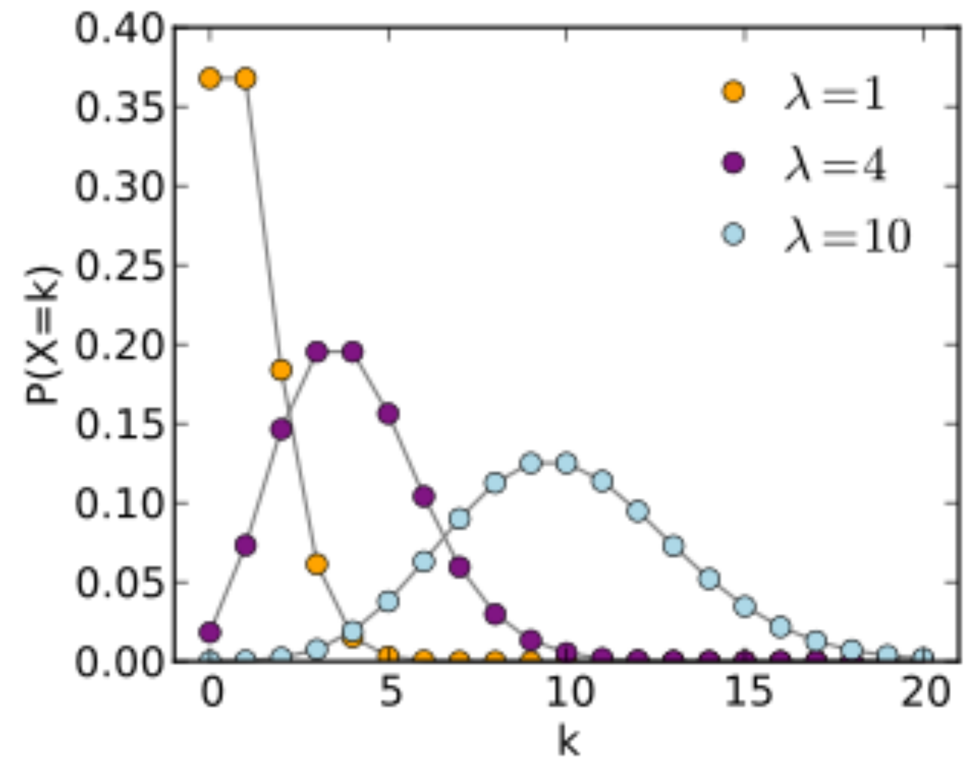
- Mean: $1/p$, var: $(1-p)/p^2$



Poisson Distribution

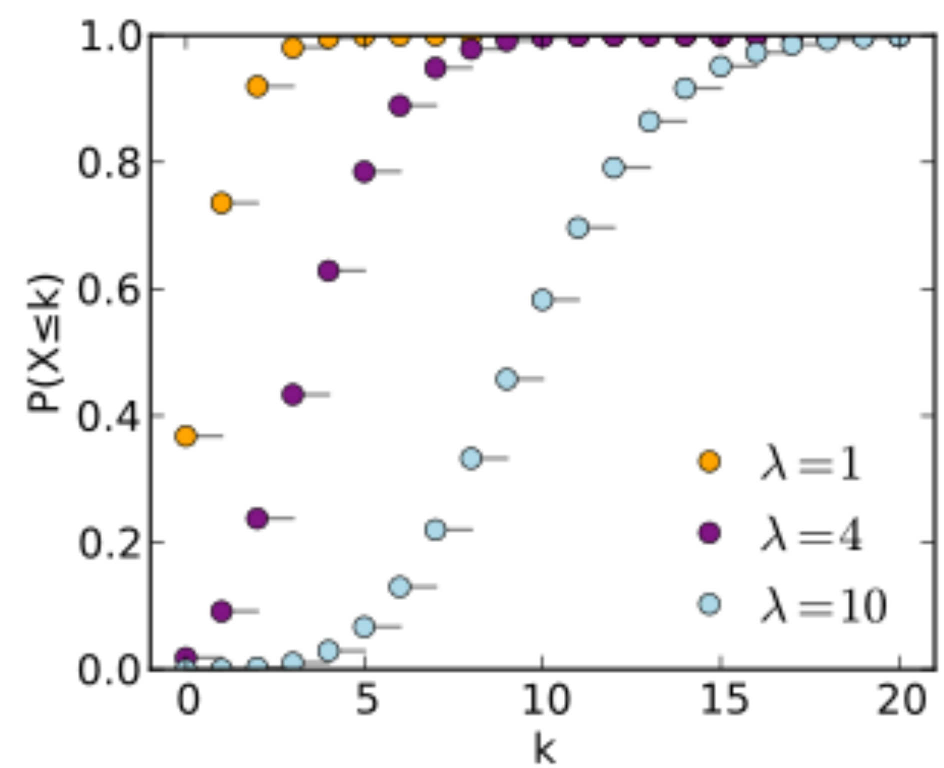
- Discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval if these events occur with a known average rate λ and independently of the last event.

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



- Cumulative distribution:

$$F(x;\lambda) = e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!} = \frac{\Gamma(\lfloor x+1 \rfloor, \lambda)}{\lfloor x \rfloor !}$$



- Γ: incomplete gamma function
- Mean: λ, var: λ

Exponential Distribution

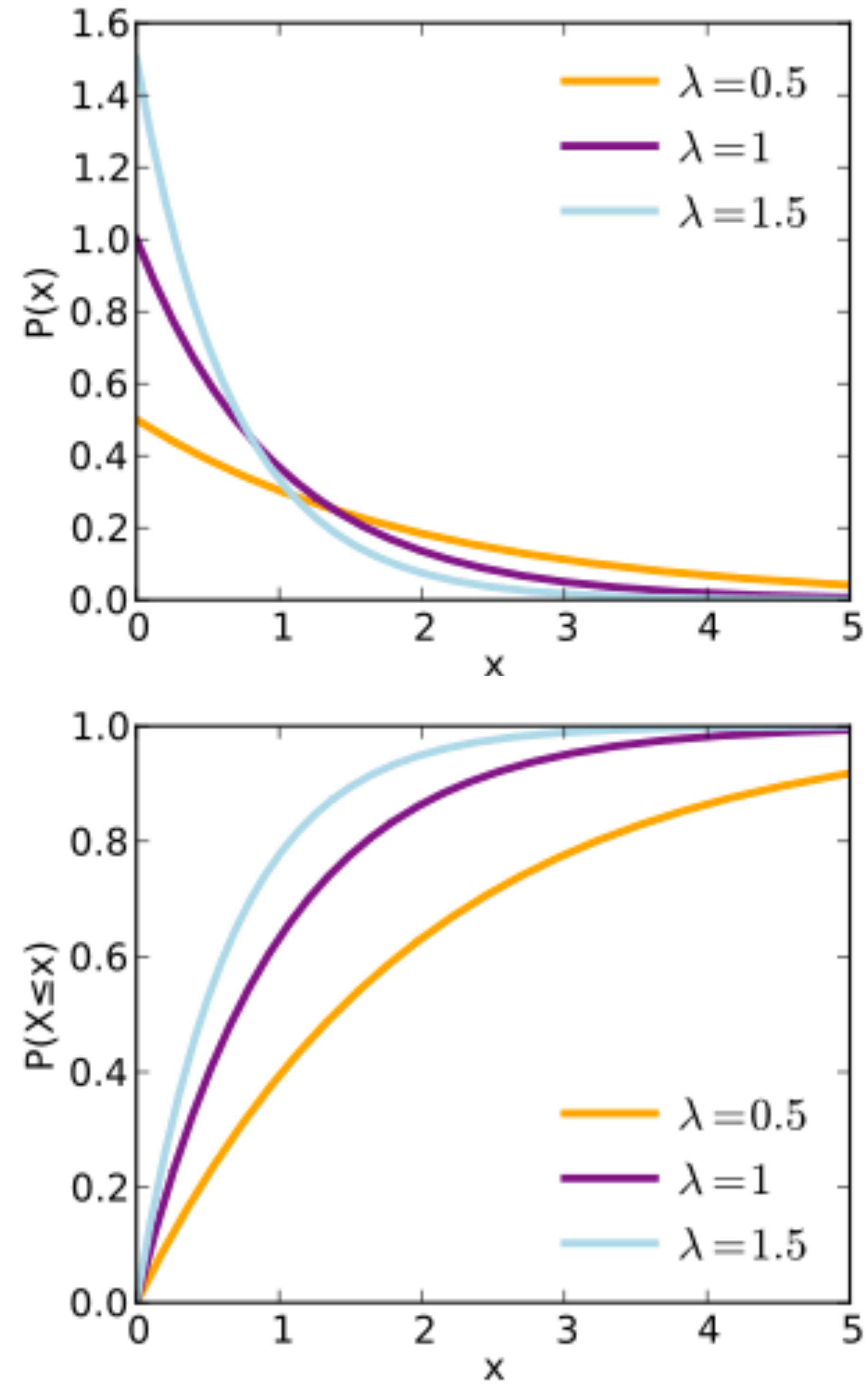
- Describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate.

$$P(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

- Cumulative distribution:

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

- Mean: $1/\lambda$, var: $1/\lambda^2$



Multinomial Variables

- Some discrete variables can take one of K possible mutually exclusive states.
- A convenient representation is the 1-of-K scheme in which the variable is represented by a K-dimensional vector **x** in which one of the elements x_k equals 1, and all remaining elements equal 0.

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}} \qquad \textstyle\sum_{k=1}^{K} x_k = 1$$

- We denote $\mu_k \equiv p(x_k = 1)$
- The distribution of **x** is $p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}, \qquad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}, \qquad \sum_k \mu_k = 1$
- $\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_M)^{\mathrm{T}} = \boldsymbol{\mu}$

Multinomial Variables

- Consider a data set D of N independent observations: $\mathbf{x}_1, \ldots, \mathbf{x}_N$
- Likelihood: $p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}$
- The number ob observations of $x_k = 1$ are $m_k = \sum_n x_{nk}$
- These are called the the **sufficient statistics** for this distribution: *"no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter"*

- *I*n order to find the maximum likelihood solution for μ, we need to maximize ln $p(D|\mu)$ with respect to μ_k taking account of the constraint that the μ_k must sum to one.
- max $\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$ $\boxed{\mu_k^{\mathrm{ML}} = \frac{m_k}{N}}$

Multinomial Distribution

- Probability of any particular combination of numbers of successes for the various categories.

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!} \qquad \sum_{k=1}^{K} m_k = N.$$

- The conjugate prior for the Multinomial Distribution is the Dirichlet Distribution:

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

Normal Distribution

- Normal distributions are important in statistics and are often used in science to represent real-valued random variables whose distributions are not known.

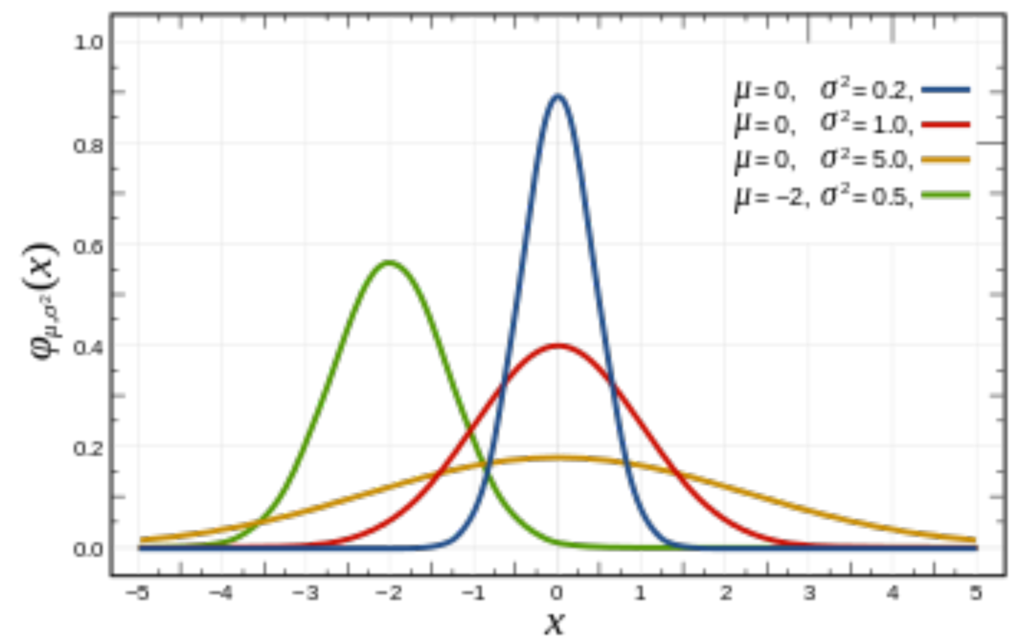
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

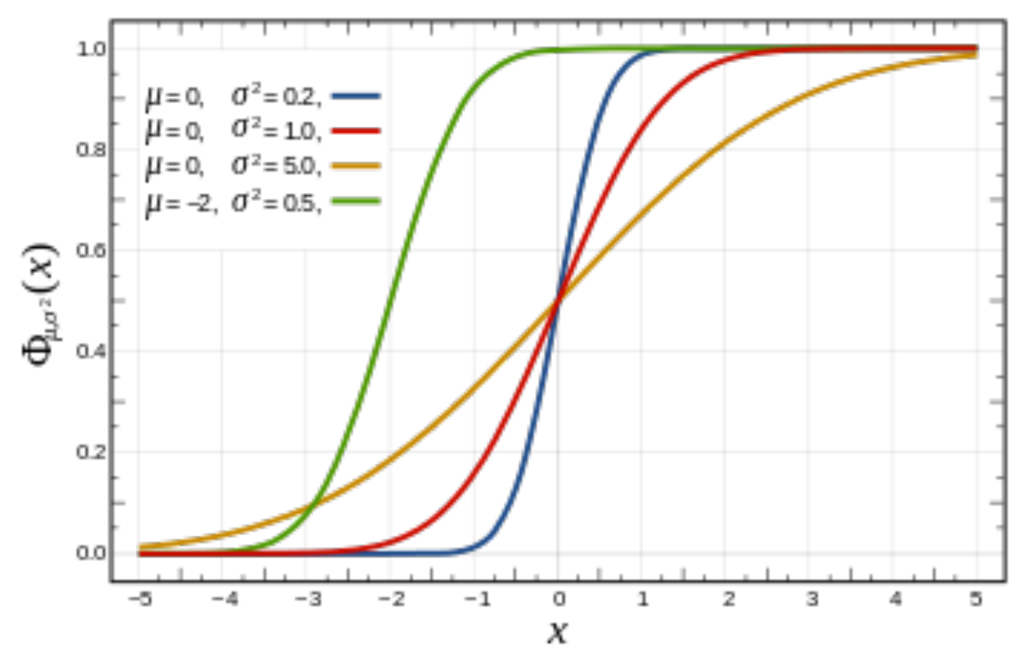
- Cumulative distribution:

$$F(x;\mu,\sigma) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{(x-\mu)}{\sqrt{2}\sigma}\right)\right]$$

- erf: error function, defined as the probability of a random variable with normal distribution of mean 0 and variance 1/2 falling in the range [-x, x]

- Mean: μ, var: σ^2

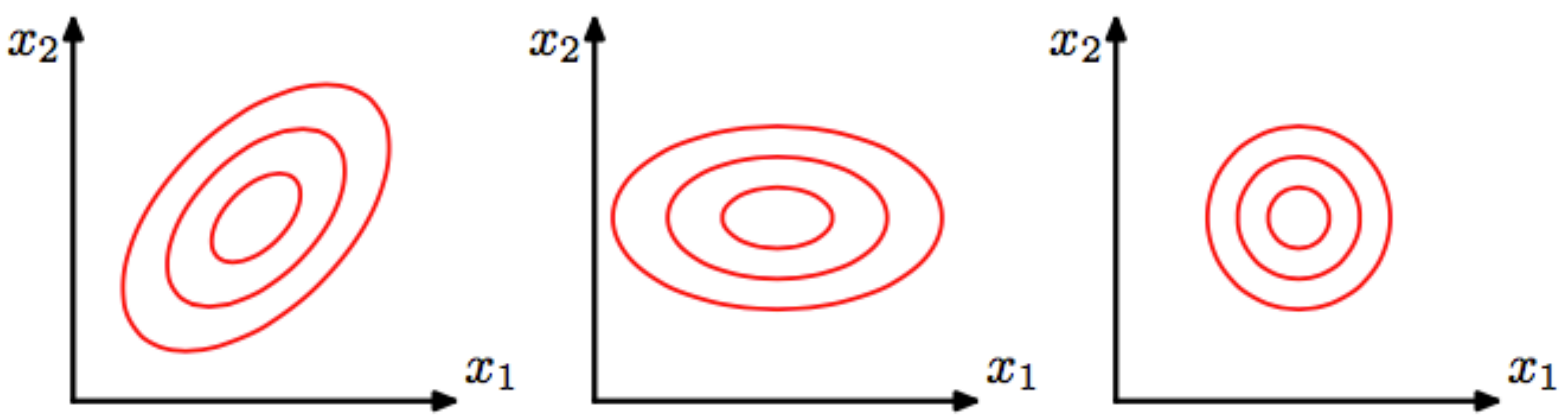




Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Where $\boldsymbol{\mu}$ is a D-dimensional mean vector
- $\mathbf{\Sigma}$ is a $D \times D$ covariance matrix
- $|\mathbf{\Sigma}|$ denotes the determinant of $\mathbf{\Sigma}$.



(a) (b) (c)

Central Limit Theorem

- When independent random variables are added, their sum tends toward a normal distribution even if the original variables themselves are not normally distributed.
- Let {X1, …, Xn} be a set of independent random variable of size n drawn from the same distribution with expected values given by μ and finite variances given by σ^2.
- We are interested in the sample average

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

- The central limit theorem states that as n gets larger, the distribution of the difference between the sample average Sn and its limit μ, when multiplied by the factor √n (that is √n(Sn – μ)), approximates the normal distribution with mean 0 and variance σ^2.

Demo

For next class…

- Model selection, hypothesis testing
- Bishop, *Pattern Recognition and Machine Learning*:
  - 1.1, 1.2.5, 1.2.6: re-visit
  - 1.3 Model Selection
  - 1.5 Decision Theory
- Hypothesis testing: I’ll send material.

Time for a quiz!!