Basics of Probability Theory

Georgy Gimel'farb

COMPSCI 369 Computational Science

Learning outcomes on probabilistic modelling:

Be familiar with basic probabilistic modelling techniques and tools

- Be familiar with basic probability theory notions and Markov chains
- Understand the maximum likelihood (ML) and identify problems ML can solve
- Recognise and construct Markov models and hidden Markov models (HMMs)

Recommended reading:

- C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006
- G. Strang, Computational Science and Engineering. Wellesley-Cambridge Press, 2007: Section 2.8
- L. Wasserman, All of Statistics: A Concise Course of Statistical Inference. Springer, 2004

Statistics, data mining, machine learning...

Statistician vs. computer scientist language

L. Wasserman, All of Statistics: A Concise Course of Statistical Inference. Springer, 2004, p.xi

Statistics	Computer science	Meaning
Estimation	Learning	Find unknown data from observed data
Classification	Supervised learning	Predict a discrete Y from X
Clustering	Unsupervised learning	Put data into groups
Data	Training / test sample	$(X_i, Y_i : i = 1, \ldots, n)$ / $(X_i : i = 1, \ldots, n)$
Covariates	Features, observations, signals, measurements	The X_i's
Classifier	Hypothesis	A map from features to outcomes
Hypothesis	—	A subset of a parameter space Θ
Confidence interval	—	An interval containing an unknown with a given probability
DAG: Directed acyclic graph	Bayesian network, Bayes net	Multivariate distribution with given conditional independence relations
Bayesian inference	Bayesian inference	Using data to update beliefs

Probability theory

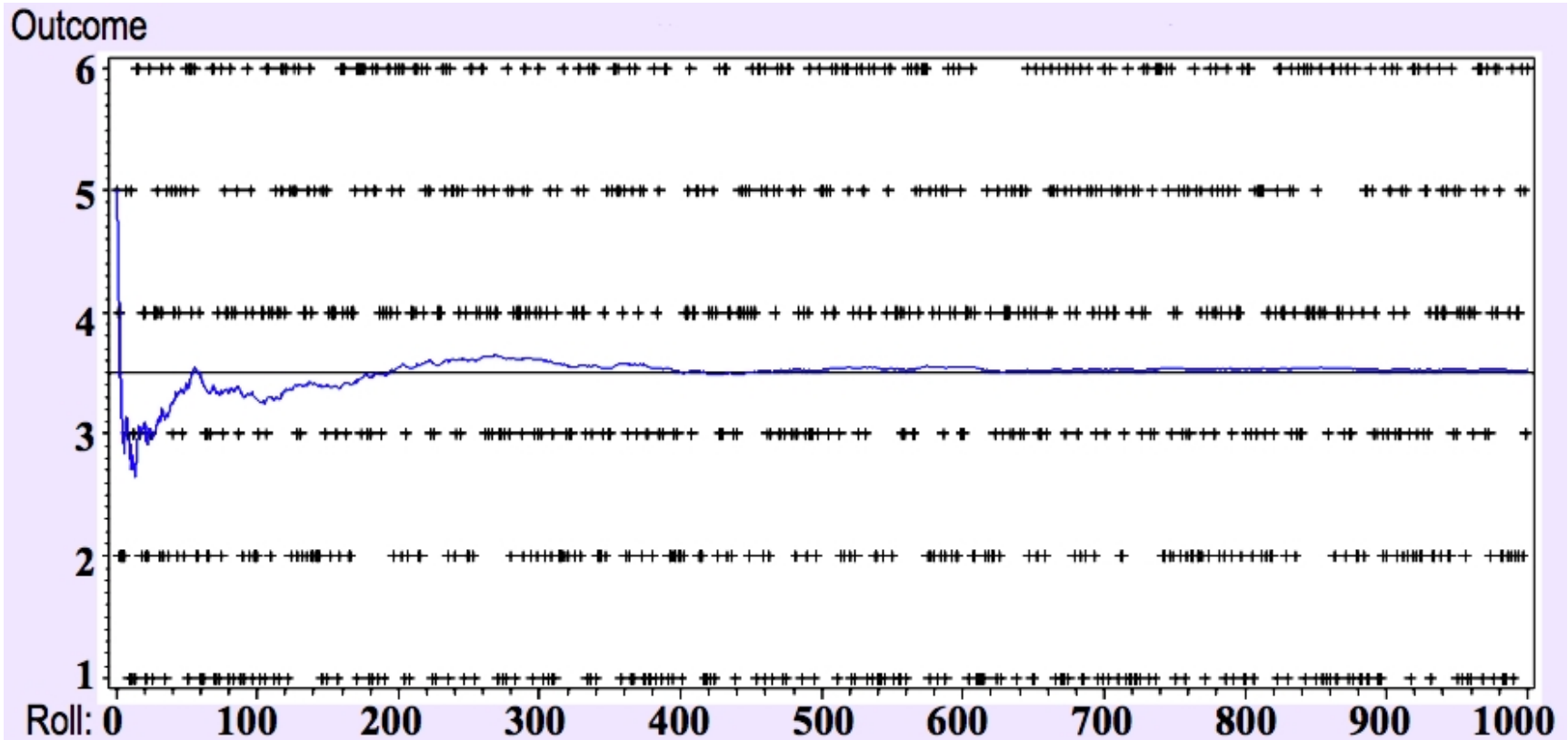
- Probability models
  - How observed data and data generating processes vary due to inherent stochasticity and measurement errors
- *Heuristic (frequentist) probability theory*
  - Probability as a limit of an empirical frequency (proportion of repetitions) of an event when the number of experiments tends to infinity
- *Modern axiomatic probability theory*
  - Based on Kolmogorov's axioms
  - Laws of Large Numbers (LLN):

    Under very general conditions, an empirical frequency converges to the probability, as the number of observations tends to infinity

    Given a random variable with a finite expected value, the mean of repeatedly observed values converges to the expected value

The Law of Large Numbers

http://en.wikipedia.org/wiki/Law_of_large_numbers



- Average of outcomes of die rolls converges to the expected value $3.5 = \frac{1+2+3+4+5+6}{6}$

Sample, or State Space (SS)


gurmeetsingh.files.wordpress.com, abnol.blogspot.com

- Set of all possible outcomes of an observation (measurement)
  - Finite sample space: if the number of outcomes is finite
- Sample point (state): an element of the SS
  1. Toss a coin – 2 outcomes: head (H) and tail (T): SS = {H,T}
  2. Toss 2 coins – 4 outcomes: SS = {HH, HT, TH, TT}
  3. Roll a die – 6 outcomes (the number of spots at top): SS={1,2,3,4,5,6}
  4. Roll 2 dice – 11 outcomes (the total number of top spots): SS={2,3,4,5,6,7,8,9,10,11,12}

Events and Probabilities

- **Event**: any subset of a sample space SS for an observation
  - Tossing 2 coins – the events of appearing 0,1, or 2 heads: $E_{\mathrm{h}:0} = \{\mathrm{TT}\}; E_{\mathrm{h}:1} = \{\mathrm{TH}, \mathrm{HT}\}; E_{\mathrm{h}:2} = \{\mathrm{HH}\}$
  - Tossing 2 coins – the event that the first coin falls tails: $E_{\mathrm{f-tails}} = \{\mathrm{TH}, \mathrm{TT}\}$
  - Rolling 2 dice – the event that the total number n of spots is equal to 11: $E_{n:11} = \{56, 65\}$
  - Rolling 2 dice – the event that the total number n of spots is equal to 4: $E_{n:4} = \{13, 22, 31\}$
- **Impossible event** if the subset is empty: e.g. $E_{n:0} = E_{n:1} = \emptyset$
- **Certain event**: the set of all sample points: $E = \mathrm{SS}$

Unions and Intersections of Events

- **Union of events** E_a OR E_b – the set of all the sample points containing in E_a or E_b or both
  - $E_{\mathrm{f-tails}}$ OR $E_{\mathrm{h:1}} = \{\mathrm{TH}, \mathrm{TT}, \mathrm{HT}\}$;
    $E_{\mathrm{h:0}}$ OR $E_{\mathrm{h:2}} = \{\mathrm{TT}, \mathrm{HH}\}$; $E_{n:2}$ OR $E_{n:11} = \{11, 56, 65\}$
- **Intersection of events** E_a AND E_b: the set of all the sample points containing simultaneously both in E_a and E_b
  - $E_{\mathrm{f-tails}}$ AND $E_{\mathrm{h:1}} = \{\mathrm{TH}\}$; $E_{n:2}$ AND $E_{n:11} = \emptyset$
- **Mutually exclusive events** E_a and E_b: if their intersection is empty
- **Complementary event** NOT E: the set of all the sample points in SS that do not belong to E
  - NOT $E_{\mathrm{f-tails}} = \{\mathrm{HT}, \mathrm{HH}\}$; NOT $E_{\mathrm{h:1}} = \{\mathrm{TT}, \mathrm{HH}\}$

Probability Axioms

Axiom 1

Probability $P(E)$ of an event E is a non-negative real number associated with each member E of a class of events being possible in the sample space for a particular experiment

Examples:

- Tossing a fair coin: $P(\text{tails}) = P(\text{heads}) = 0.5$
- Rolling a fair die:
  $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$
- Tossing an unfair coin: $P(\text{tails}) = 0.1$; $P(\text{heads}) = 0.9$

Function $P : \text{SS} \rightarrow \mathbb{R}$ is called also a probability distribution, or a probability measure

Probability Axioms

Axiom 2

The probability of the certain event is equal to 1

Examples:

- Tossing a coin: $P(\text{tails OR heads}) = 1$
- Rolling a die: $P(1 \bigcup 2 \bigcup 3 \bigcup 4 \bigcup 5 \bigcup 6) = 1$

Axiom 3

If events E_a and E_b are mutually exclusive, then

$$P(E_a \bigcup E_b) = P(E_a) + P(E_b)$$

- If the certain event can be split into K mutually exclusive events $E_1, \ldots, E_K$, then $P(E_1) + P(E_2) + \ldots + P(E_K) = 1$
  - Rolling a die: $P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$

Properties Derived from the Axioms

- If E_a and E_b are in a class of events then E_a AND $E_b \equiv E_a E_b$, E_a OR $E_b \equiv E_a \bigcup E_b$, and NOT $E_a \equiv \neg E_a$ are in that class
  - Rolling a die: if $E_a = \{1, 2, 4\}$ and $E_b = \{1, 4, 6\}$ then $E_a E_b = \{1, 4\}$; $E_a \bigcup E_b = \{1, 2, 4, 6\}$; $\neg E_a = \{3, 5, 6\}$
- If $E_1, \ldots, E_K$ are K mutually exclusive (disjoint) events, then $P(E_1 \bigcup E_2 \bigcup \ldots \bigcup E_K) = P(E_1) + P(E_2) + \ldots + P(E_K)$
  - Rolling a die: $P(1 \bigcup 2 \bigcup 4) = 1/6 + 1/6 + 1/6 = 1/2$
- Probability $P(E)$ of any event E is in the interval $[0, 1]$; $0 \leq P(E) \leq 1$, because
  1. $P(E) + P(\neg E) = P(\text{certain event}) \equiv 1$ and
  2. $P(\neg E) \geq 0$
- Probability of an impossible event: $P(\emptyset) = 0$

Properties Derived from the Axioms

Lemma

For any events A *and* B, $P(A \bigcup B) = P(A) + P(B) - P(AB)$

Proof.

- $A \bigcup B = (A(\neg B)) \bigcup (AB) \bigcup ((\neg A)B)$
- All three right-hand events are mutually exclusive (disjoint)
- Because the probability is additive to disjoint events, it holds:

$$\begin{aligned} P(A \bigcup B) &= P(A(\neg B)) + P(AB) + P((\neg A)B) \\ &= \underbrace{P(A(\neg B)) + P(AB)}_{P((A(\neg B)) \bigcup (AB))} + \underbrace{P((\neg A)B) + P(AB)}_{P((\neg A)B) \bigcup (AB))} - P(AB) \\ &= P((A(\neg B)) \bigcup (AB)) + P((\neg A)B) \bigcup (AB)) - P(AB) \\ &= P(A) + P(B) - P(AB) \end{aligned}$$

Joint Probabilities

Sample space for outputs of several different observations

- Each event contains sample points for each observation
  - Rolling two dice – the observed spots $[1,1]$; $[1,2]$;. . . , $[1,6]$; $[2,1]$; $[2.2]$; . . . ; $[2,6]$; . . . ; $[6,1]$; . . . ; $[6,5]$; $[6,6]$
  - Probability of 2 spots on the 1^{st} and 5 spots on the 2^{nd} die: $P(2,5) = 1/36$
- If all the events $A_1, \ldots, A_K$ for the observation A are disjoint:

$$\begin{aligned} & P(A_1 \bigcup A2 \bigcup \ldots \bigcup A_K, B_m) \\ = \; & P(A_1, B_m) + P(A_2, B_m) + \ldots + P(A_K, B_m) = P(B_m) \end{aligned}$$

- If all the events $B_1, \ldots, B_M$ for the observation B are disjoint:

$$\begin{aligned} & P(A_k, B_1 \bigcup B_2 \bigcup \ldots \bigcup B_M) \\ = \; & P(A_k, B_1) + P(A_k, B_2) + \ldots + P(A_k, B_M) = P(A_k) \end{aligned}$$

Conditional and Joint Probabilities

Conditional probability $P(B|A)$ of an event B providing an event A has happened:

the ratio $P(B|A) = \frac{P(A,B)}{P(A)}$ between the probability $P(A, B)$ of the joint event (A, B) and the probability $P(A)$ of the event A

- Conditional and unconditional probabilities have the same properties
- Joint probability of A and B: $P(A, B) = P(B|A)P(A)$

Example:

- $A \in \{1_{\text{(sun today)}}, 2_{\text{(rain today)}}\}$; $B \in \{1_{\text{(sun tomorrow)}}, 2_{\text{(rain tomorrow)}}\}$
- $P(A = 1) = 0.6$; $P(A = 2) = 0.4$; $P(B = 1|A = 1) = P(B = 2|A = 2) = 0.7$; and $P(B = 2|A = 1) = P(B = 1|A = 2) = 0.3$
- $P(A = 1, B = 1) = 0.7 \cdot 0.6 = 0.42$; $P(A = 2, B = 1) = 0.3 \cdot 0.4 = 0.12$; $P(A = 1, B = 2) = 0.3 \cdot 0.6 = 0.18$; $P(A = 2, B = 2) = 0.7 \cdot 0.4 = 0.28$
- $P(B = 1) = P(A = 1, B = 1) + P(A = 2, B = 1) = 0.42 + 0.12 = 0.54$; $P(B = 2) = P(A = 1, B = 2) + P(A = 2, B = 2) = 0.18 + 0.28 = 0.46$

Statistical Independence

If $P(B|A) = P(B)$ then $P(A, B) = P(B)P(A)$ and $P(A|B) = P(A)$

- Mutual independence of each pair of events from a system of $N \geq 3$ events is insufficient to guarantee the independence of three or more events
- All N events $\{A_1, \ldots, A_N\}$ are statistically independent if for all subsets of indices $1 \leq i < j < k < \ldots < N$ the following relationships hold:

$$\begin{aligned} P(A_i, A_j) &= P(A_i)P(A_j); \\ P(A_i, A_j, A_k) &= P(A_i)P(A_j)P(A_k); \\ \ldots; & \\ P(A_1, A_2, \ldots, A_N) &= P(A_1)P(A_2)P(A_N) \end{aligned}$$

Bayes' Theorem

Let an experiment A have M mutually exclusive outcomes A_m and an experiment B have N mutually exclusive outcomes B_n

Then the conditional probability $P(B_n|A_m)$ can be represented with $P(A_m|B_n)$ and $P(B_n)$ as follows:

$$P(B_n|A_m) = \frac{P(A_m|B_n)P(B_n)}{\sum_{i=1}^{N} P(A_m|B_i)P(B_i)} \equiv \frac{P(A_m|B_n)P(B_n)}{P(A_m)} \equiv \frac{P(A_m, B_n)}{P(A_m)}$$

Basic interpretation:

Prior probability $P(B_n)$ of B_n and conditional probability $P(A_m|B_n)$

$\Rightarrow$ Posterior probability $P(B_n|A_m)$ of B_n given A_m was observed

Conditional Probabilities

Example of using the Bayes' Theorem

- Joint and conditional probabilities $P(A_m, B_n)$: $P(A_m|B_n)$

	A_1	A_2	A_3	A_4	A_5	$P(B_n)$
B_1	0.000 : 0.000	0.100 : 0.172	0.050: 0.086	0.400 : 0.690	0.030 : 0.052	0.580
B_2	0.150 : 0.357	0.050 : 0.119	0.200: 0.476	0.020 : 0.048	0.000 : 0.000	0.420
$P(A_m)$	0.150	0.150	0.250	0.420	0.030	

- $P(B_1|A_1) = \frac{0.000}{0.150} = 0.0$; $P(B_2|A_1) = \frac{0.150}{0.150} = 1.0$
- $P(B_1|A_4) = \frac{0.400}{0.420} = 0.952$; $P(B_2|A_4) = \frac{0.020}{0.420} = 0.048$;

σ-Algebra of Events (optional)

- Generally, it is infeasible to assign probabilities to all subsets of a sample space SS
- A set of events $\mathcal{E}$ is called a σ-algebra or a σ-field if
  1. It contains an empty set: $\emptyset \in \mathcal{E}$
  2. If $E_1, E_2, \ldots, \in \mathcal{E}$ then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{E}$
  3. $E \in \mathcal{E}$ implies $\neg E \in \mathcal{E}$
- Sets in $\mathcal{E}$ are said to be **measurable**
- $(\mathrm{SS}, \mathcal{E})$ is called a **measurable space**
- If P is a probability measure defined on $\mathcal{E}$, then $(\mathrm{SS}, \mathcal{E}, P)$ is called a probability space

Example (**Borel σ-field**): SS – the real line; $\mathcal{E}$ – the smallest σ-field containing all the open subsets (a, b) of points: $-\infty \leq a < b \leq \infty$

Distribution Functions

Sample spaces and events relate to data via the concept of

Random Variable

A **random variable** is a (measurable) mapping $X : \text{SS} \to \mathbb{R}$ that assigns a real number $X(e)$ to each outcome $e \in \text{SS}$

Examples:

- e – a sequence of 10 flipped coins, e.g. $e' = HHHTTTHTHHT$; $X(e)$ – the number of heads in e, e.g. $X(e') = 6$
- $\text{SS} = \{(x, y) : |x| \leq 1;\ |y| \leq 1\}$; drawing a point $e = (x, y)$ at random from SS; some random variables: $X(e) = x$; $Y(e) = y$; $Z(e) = x + y$, etc.

Cumulative Distribution Function (CDF)

The CDF $F_X : \mathbb{R} \to [0, 1]$ is defined by $F_X(x) = P(X \leq x)$

Typically, the CDF is written as F instead of F_X

Cumulative Distribution Functions

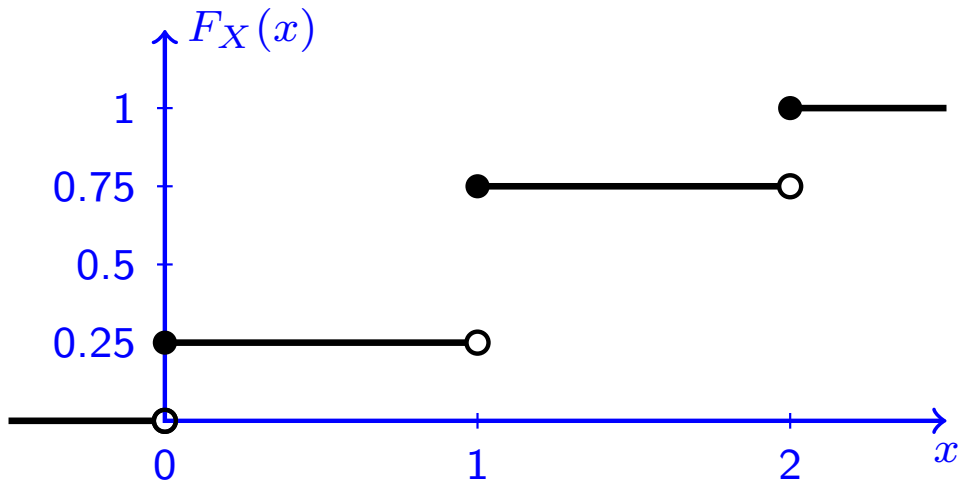
The CDF contains all the information about the random variable, i.e. completely determines its distribution

- X – the number of heads for a fair coin flipped twice:

  $P(X = 0) = P(X = 2) = \frac{1}{4}$ and $P(X = 1) = \frac{1}{2}$

- The CDF: $F_X(x) =$

$$\begin{cases} 0 & x < 0 \\ 0.25 & 0 \le x < 1 \\ 0.75 & 1 \le x < 2 \\ 1 & x \ge 2 \end{cases}$$



The CDF is right continuous, non-decreasing, and defined for all x

- Even though the above random variable X only takes values 0, 1, and 2, its CDF is defined for all $x \in \mathbb{R} = [-\infty, \infty]$

Continuous Random Variable (optional)

- X is continuous if a **probability density function** (PDF) f_X exists such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \text{ and } f_X(x) = \frac{dF_X(x)}{dx}$$

  at all points x where $F_X(x)$ is differentiable
- $f_X(x) \geq 0$ for all x; $\int_{-\infty}^{\infty} f_X(x)dx = 1$, and for every $a \leq b$, $P(a < X < b) = \int_a^b f_X(x)dx$
- **Uniform** $(0, 1)$ distribution:

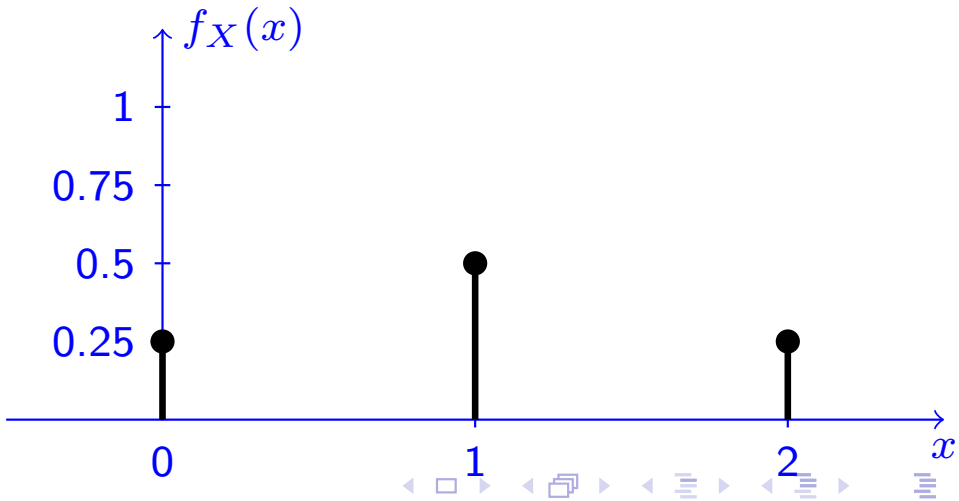
$$f_X(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{CDF } F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

Discrete Random Variable

- X is discrete if it takes countably many values $\{x_1, x_2, \ldots\}$
  - A set is countable if it is finite or can be put in a one-to-one correspondence with the integers
  - Different points x_i correspond to mutually exclusive events
- $f_X(x) = P(X = x)$ is a **probability (mass) function** for X: $f_X(x) \geq 0$ for all $x \in \mathbb{R}$; $\sum_i f_X(x_i) = 1$
- Relation to the CDF: $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$

Probability function for the number of heads after flipping a coin twice:

$$f_X(x) = \begin{cases} 0.25 & x = 0 \\ 0.5 & x = 1 \\ 0.25 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$



Discrete Random Variable: Expectation and Variance

- Expectation (mean) of X under p.d. $P(x) \equiv P(X = x)$; $x \in \{x_1, \ldots, x_K\}$; $\sum_{k=1}^{K} P(x_k) = 1$:

$$\mathbb{E}[X] = \sum_{k=1}^{K} x_k P(x_k)$$

- Expectation of a function $\varphi(x)$: $\mathbb{E}[\varphi] = \sum_{k=1}^{K} \varphi(x_k) P(x_k)$
  - Conditional expectation: $\mathbb{E}[\varphi|y] = \sum_{k=1}^{K} \varphi(x_k) P(x_k|y)$
- Variance: $\mathbb{V}[\varphi] = \mathbb{E}\left[(\varphi(x) - \mathbb{E}[\varphi])^2\right] = \mathbb{E}\left[\varphi^2\right] - (\mathbb{E}[\varphi])^2$
  - Standard deviation: $s[\varphi] = \sqrt{\mathbb{V}[\varphi]}\ \mathbb{E}[\varphi] = \sum_{k=1}^{K} \varphi(x_k) P(x_k)$

Important Discrete Random Variables

- **Point** mass distribution: $P(x = a) = 1$,
  i.e. $f(x) = 1$ for $x = a$ and 0 otherwise
  - CDF $F(x) = 0$ for $x < a$ and 1 for $x \geq a$
- **Binomial** (n, p) distribution; $0 \leq p \leq 1$:

$$f(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

  - Probability of x heads when flipping n times the coin which falls heads up with probability p
  - Binomial coefficient $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
- **Geometric** distribution with parameter p: $f(x) = p(1-p)^{x-1}$
- **Poisson** distribution with parameter λ: $f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$
  - A model for counts of rare events (like traffic accidents)

Important Continuous Random Variables (optional)

- **Uniform** (a, b) distribution: $f(x) = \frac{1}{b-a}$ if $x \in [a, b]$ and 0 otherwise
  - Distribution function: $F(x) = 0$ if $x < a$; $\frac{x-a}{b-a}$ if $a \leq x \leq b$, and 1 if $x > b$
- **Normal**, or Gaussian $N(\mu, \sigma^2)$ distribution ($\mu \in \mathbb{R}; \sigma > 0$):

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

- **Gamma** distribution with parameters $\alpha, \beta > 0$:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta); x > 0$$

  - Gamma function $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} \exp(-y) dy$
  - $\alpha = 1$: **Exponential** distribution $f(x) = \beta^{-1} \exp(-x/\beta)$

Important Continuous Random Variables (optional)

- **Beta** distribution with $\alpha, \beta > 0$:
  $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$; $0 < x < 1$
- t-**distribution** with ν degrees of freedom:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

  - Similar to a normal distribution but with thicker tails
  - The standard normal (0,1) distribution if $\nu = \infty$
  - Cauchy distribution (if $\nu = 1$): $f(x) = \frac{1}{\pi(1+x^2)}$
- χ^2 **distribution** with p degrees of freedom:

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} \exp(-x/2); \quad x > 0$$

Example: Binomial Distribution – n tossed coins

- **Bernoulli** experiment (trial): $P(H) = p$ and $P(T) = 1 - p$
  - Events H and T are mutually exclusive: $P(H) + P(T) = 1$
- n independent experiments: what's the probability of x heads?
  - Probability of a single sequence with x heads: $p^x(1-p)^{n-x}$
  - Total number of such sequences (the binomial coefficient): $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
    - $n!$ sequences for different results of each experiment
    - $x!$ and $(n-x)!$ indistinguishable duplicates of the resulting heads and tails, respectively, in these $n!$ sequences
  - Probability of x heads: $P_{\text{heads}}(x) = \binom{n}{x} p^x(1-p)^{n-x}$
- Expected number of heads: $\mathbb{E}_{\text{heads}}[n] = np$
- Variance of the number of heads: $\mathbb{V}_{\text{heads}}[n] = np(1-p)$
- Example: $n = 100$; $p = 0.5 \Rightarrow \mathbb{E}[X] = 50$; $\mathbb{V}[X] = 25$; $s[X] = 5$;
  $P_{\text{heads}}(50) \approx 0.08$; $P_{\text{heads}}(45) = P_{\text{heads}}(55) \approx 0.05$;
  $P_{\text{heads}}(40) = P_{\text{heads}}(60) \approx 0.01$; $P_{\text{heads}}(35) = P_{\text{heads}}(65) \approx 0.001$

Example: Binomial Distribution – Computing Probabilities

- Stirling's approximation of factorials: $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$
  - Constant $e \approx 2.718281828\ldots$
- Approximate probability $P_{\text{heads}}(x)$ of x heads in n tosses:

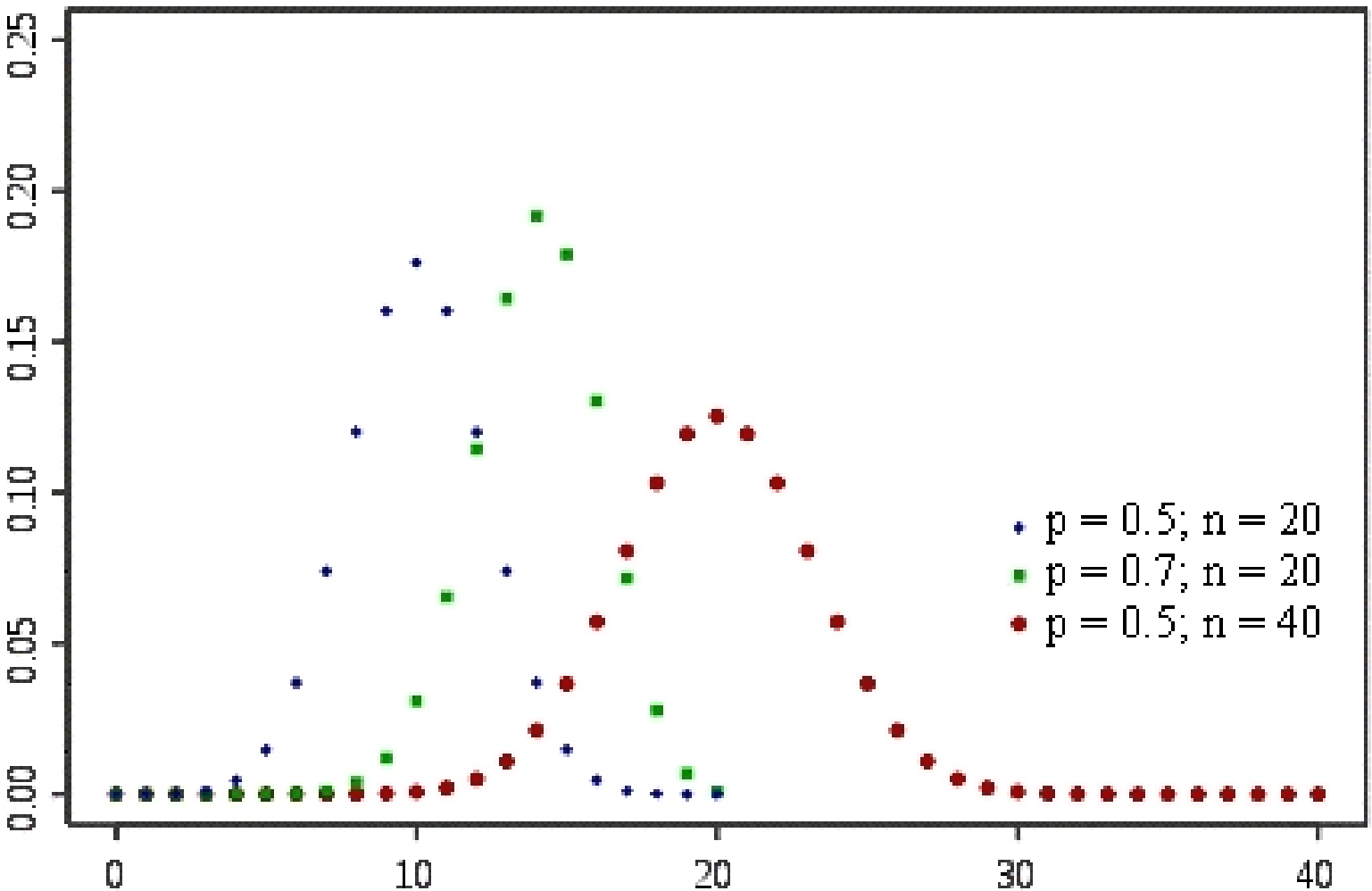
$$\begin{aligned} & \frac{\sqrt{2\pi n} n^n e^{-n}}{\sqrt{2\pi x} x^x e^{-x} \sqrt{2\pi(n-x)} (n-x)^{n-x} e^{x-n}} p^x (1-p)^{n-x} \\ = & \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{x(n-x)}} \frac{n^n}{x^x (n-x)^{n-x}} p^x (1-p)^{n-x} \\ = & \frac{1}{\sqrt{2\pi n \kappa(1-\kappa)}} \left(\frac{p}{\kappa}\right)^x \left(\frac{1-p}{1-\kappa}\right)^{n-x} \end{aligned}$$

  where $\kappa = \frac{x}{n}$ is the empirical probability of x heads
- $\log P_{\text{heads}}(x) = -\frac{1}{2}\left(\log(2\pi) + \log n + \log \kappa + \log(1-\kappa)\right) + x(\log p - \log \kappa) + (n-x)(\log(1-p) - \log(1-\kappa))$

$n = 100;\ p = 0.5;\ x = 35 \Rightarrow \log P_{\text{heads}}(35) = -7.0513;\ P_{\text{heads}}(35) = 0.00087$
$n = 100;\ p = 0.5;\ x = 40 \Rightarrow \log P_{\text{heads}}(40) = -4.5215;\ P_{\text{heads}}(40) = 0.01087$
$n = 100;\ p = 0.5;\ x = 50 \Rightarrow \log P_{\text{heads}}(50) = -2.5284;\ P_{\text{heads}}(50) = 0.07979$
$n = 100;\ p = 0.5;\ x = 60 \Rightarrow \log P_{\text{heads}}(60) = -4.5215;\ P_{\text{heads}}(60) = 0.01087$

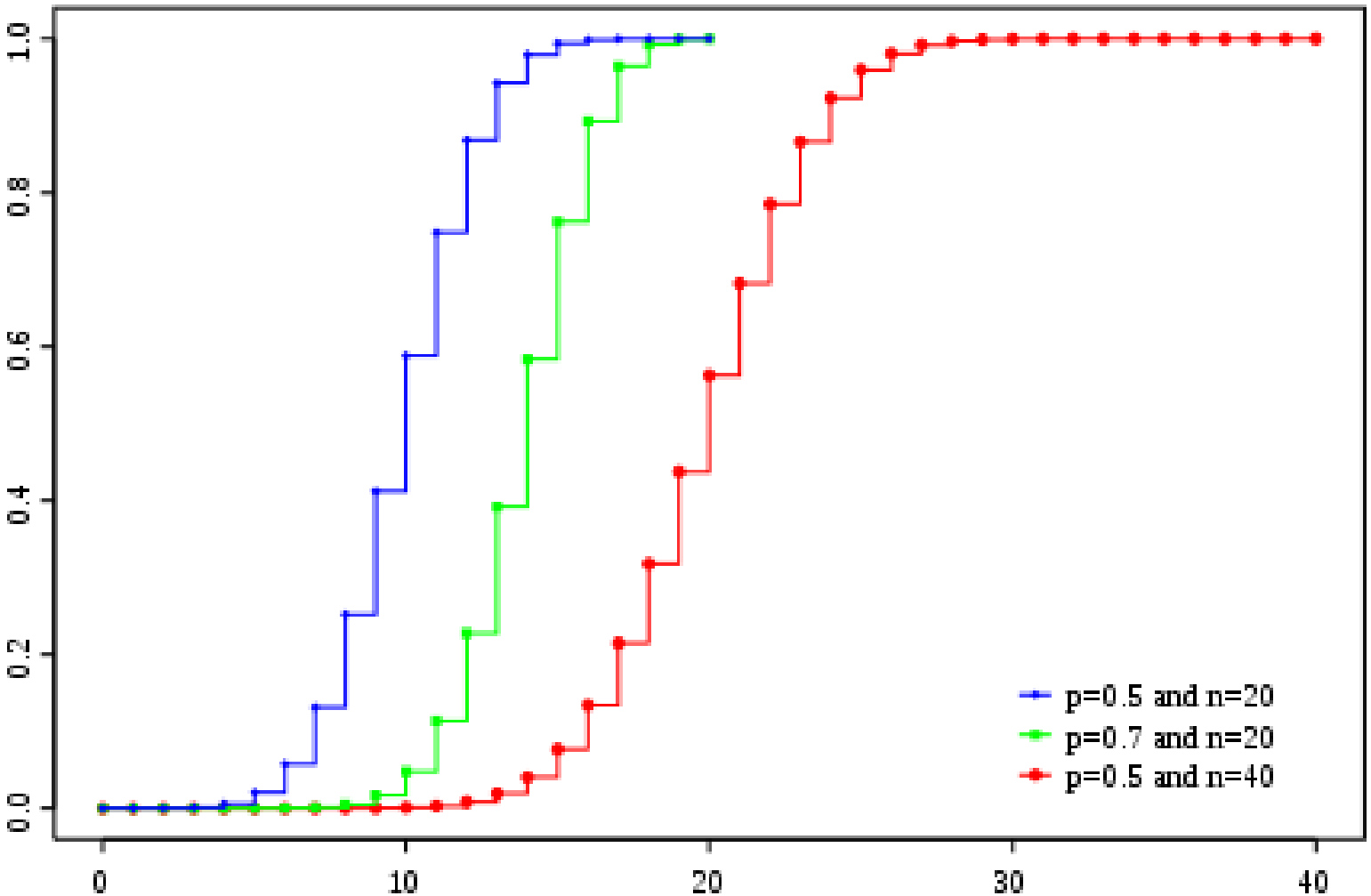
Binomial Distribution: Examples

http://en.wikipedia.org/wiki/Binomial_distribution



Cumulative Binomial Distribution: Examples

http://en.wikipedia.org/wiki/Binomial_distribution



Marginal and Conditional Distributions

- If (X, Y) have joint distribution with mass function $f_{X,Y}$, then the **marginal mass function for X** is

$$f_X(x) \equiv P(X = x) = \sum_y P(X = x, Y = y) \equiv \sum_y f(x, y)$$

  - The **marginal mass function for Y**: $f_Y(y) = \sum_x f(x, y)$
- Conditional probability mass function for discrete variables:

$$f_{X|Y}(x|y) \equiv P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \equiv \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

assuming that $f_Y(y) > 0$

Marginal and Conditional Distributions (optional)

- Probability density function for two continuous variables: $f(x, y)$
- Marginal densities for continuous variables: $f_X(x) = \int f(x, y)dy$ and $f_Y(y) = \int f(x, y)dx$
- Conditional densities: $f_{X|y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ assuming that $f_Y(y) > 0$
- Probability of X having values in A:

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$$

- Multivariate p.d.f. $f(x_1, \ldots, x_n)$
  - Marginals and conditionals: defined like as in the bivariate case

Multivariate Distributions and IID Samples

- Random vector $X = (X_1, \ldots, X_n)$
  - X_i; $i = 1, \ldots, n$ – random variables (discrete or continuous)
- Independent variables: if for every $A_1, \ldots, A_n$

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

  - It suffices that $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$
- Independent and identically distributed (i.i.d.) $X_1, \ldots, X_n$
  - Variables are independent
  - Each variable has the same marginal distribution with CDF F
  - $X_1, \ldots, X_n$ – also called a **random sample of size n from F**

Important Multivariate Distributions

- **Multinomial** (n, p) distribution: $f(x) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$
  - k items; $p_j \geq 0$ – the probability of drawing an item j: $\sum_{j=1}^{k} p_j = 1$
  - n independent draws with replacement from a box of items
  - X_j – the number of times that item j appears; $\sum_{j=1}^{k} X_j = n$
  - Marginal distribution of X_j is binomial (n, p_j)
- Multivariate **normal** (μ, Σ) distribution:

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \mu)^{\mathsf{T}} \Sigma^{-1} (x - \mu) \right\}$$

  - μ – a vector of length k
  - Σ – a $k \times k$ symmetric, positive definite matrix
  - Standard normal pdf: $\mu = 0$; $\Sigma = I$

Statistical Characteristics

- 1D discrete variable:
  p.d.f. $\{f(x_i) \equiv P(X_i = x_i) : i = 1, \ldots, n\}$; $\sum_{i=1}^{n} f(x_i) = 1$
- **Expectation** of a function $\varphi(x)$: $\mathbb{E}[\varphi] = \sum_i \varphi(x_i) \cdot f(x_i)$
  - Sample mean for i.i.d. $X_1, \ldots, X_n$: $m_n = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - If $\mathbb{E}[X_i] = \mu$, then $\mathbb{E}[m_n] = \mu$
- **Variance** ("spread"): $\mathbb{V}[g] = \mathbb{E}[(\varphi(x) - \mathbb{E}[\varphi])^2]$
  - Sample variance for i.i.d. $X_1, \ldots, X_n$:
    $s_n = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m_n)^2$
  - If $\mathbb{V}[X_i] = \sigma^2$, then $\mathbb{V}[s_n] = \sigma^2$
- 2D random variables X and Y
  - Expectations μ_X,μ_Y and variances σ_X^2,σ_Y^2
  - **Covariance**: $\mathrm{Cov}(X, Y) \equiv \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$
  - **Correlation**:

$$\rho \equiv \rho_{X,Y} \equiv \rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1]$$

Covariance Matrix

Running n different experiments at once

- Each observation – an n-component vector $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$
- Random deviations $e_i = x_i - \mathbb{E}[x_i]$ around expectations
  - Deviations could be correlated, i.e. interdependent

Covariance $\sigma_{ij} = \sigma_{ji} = \mathbb{E}[e_i e_j]$ (expected product value)

- $P_{ij}(e_i, e_j)$ – joint distribution of two deviations e_i and e_j
- $\sigma_{ii} \equiv \sigma_i^2 = \mathbb{E}[e_i^2]$ – variance of the deviation e_i
- For independent deviations: $P_{ij}(e_i, e_j) = P_i(e_i)P_j(e_j)$; $\sigma_{ij} = 0$

Covariances σ_{ij} are entries of the **covariance matrix $\mathbf{\Sigma}$**

- Variances σ_i^2 are the diagonal entries of $\mathbf{\Sigma}$
- When deviations are independent, $\mathbf{\Sigma}$ is a diagonal matrix