Chapter II.2: Basic Probability Theory and Statistics

Wasserman, Ch. 1–5

What is a probability

- "If I throw a dice, I will probably get 4 or less"

- "I'll probably go running after this lecture"

- The term "probability" here means different things
  - The outcome of a repeatable experiment
  - My personal belief

Views on probability

- In **classical** definition, probability is equally shared among all outcomes, provided the outcomes are equally likely
  - "Equally likely" is decided based on physical symmetries or the like
- In **frequentism**, a probability is the frequency of which something happens over repeated experiments
  - Requires infinite number of repetitions
- In **subjectivism** (**Bayesianism**), probability refers to my subjective "degree of belief"
  - But everybody's belief is different

Axiomatic approach: sample spaces and events

- A **sample space** Ω is a set of all possible outcomes of an experiment
  - Element $e \in \Omega$ is a **sample outcome** or **realization**
- Subsets $E \subseteq \Omega$ are **events**
- Examples:
  - If we toss a coin twice, Ω = {HH, HT, TH, TT}
    - Event "Second toss is tails" is A = {HT, TT}
  - If we toss a coin until we get tails, Ω = {T, HT, HHT, HHHT, HHHHT, HHHHHT, …}
  - If we measure a temperature in Kelvins, $\Omega = \{x \in \mathbb{R}, x \geq 0\}$

Axiomatic approach: probability measures

- Collection $\mathscr{A} \subseteq 2^{\Omega}$ is a **σ-algebra** of Ω if
  - $\Omega \in \mathscr{A}$
  - If $A \in \mathscr{A}$, then $(\Omega \setminus A) \in \mathscr{A}$
  - If $A_1, A_2, A_3, \ldots \in \mathscr{A}$, then $(\cup_i A_i) \in \mathscr{A}$
- Function $\Pr: \mathscr{A} \to [0, 1]$ is a **probability measure** if
  - **Axiom 1:** $\Pr[A] \geq 0$ for every $A \in \mathscr{A}$
  - **Axiom 2:** $\Pr[\Omega] = 1$
  - **Axiom 3:** If $A_1, A_2, \ldots$ are disjoint, then $\Pr[\cup_i A_i] = \sum_i \Pr[A_i]$ (countably many A_is)

Intermission: some combinatorics

- The **power set** of a set A, 2^A (or $\mathscr{P}(A)$) is a collection of all subsets of A
  - If $A = \{1, 2, 3\}$, then $2^A = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$
  - The size of the power set is $2^{|A|}$
    - If A is finite, this is a natural number
    - If $A = \mathbb{N}$, this is the same cardinality as the real numbers
    - If $A = \mathbb{R}$, this is the next cardinal number
- The number of size-k subsets of A is

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

Axiomatic approach: probability spaces and further properties

- A **probability space** is a triple $(\Omega, \mathscr{A}, \Pr)$
  - $\mathscr{A}$ contains all the events we can assign a probability
    - If Ω is finite or countably infinite, we can have $\mathscr{A} = 2^{\Omega}$
    - If Ω is uncountable, it contains sets that cannot have probability (unmeasurable sets)
- From the axioms we can derive that
  - $\Pr[\emptyset] = 0$
  - If $A \subseteq B$, then $\Pr[A] \leq \Pr[B]$
  - $\Pr[\Omega \setminus A] = 1 - \Pr[A]$
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$

Axiomatic approach: random variables

- A **random variable** (**r.v.**) is a function $X\colon \mathscr{A} \to \mathbb{R}$

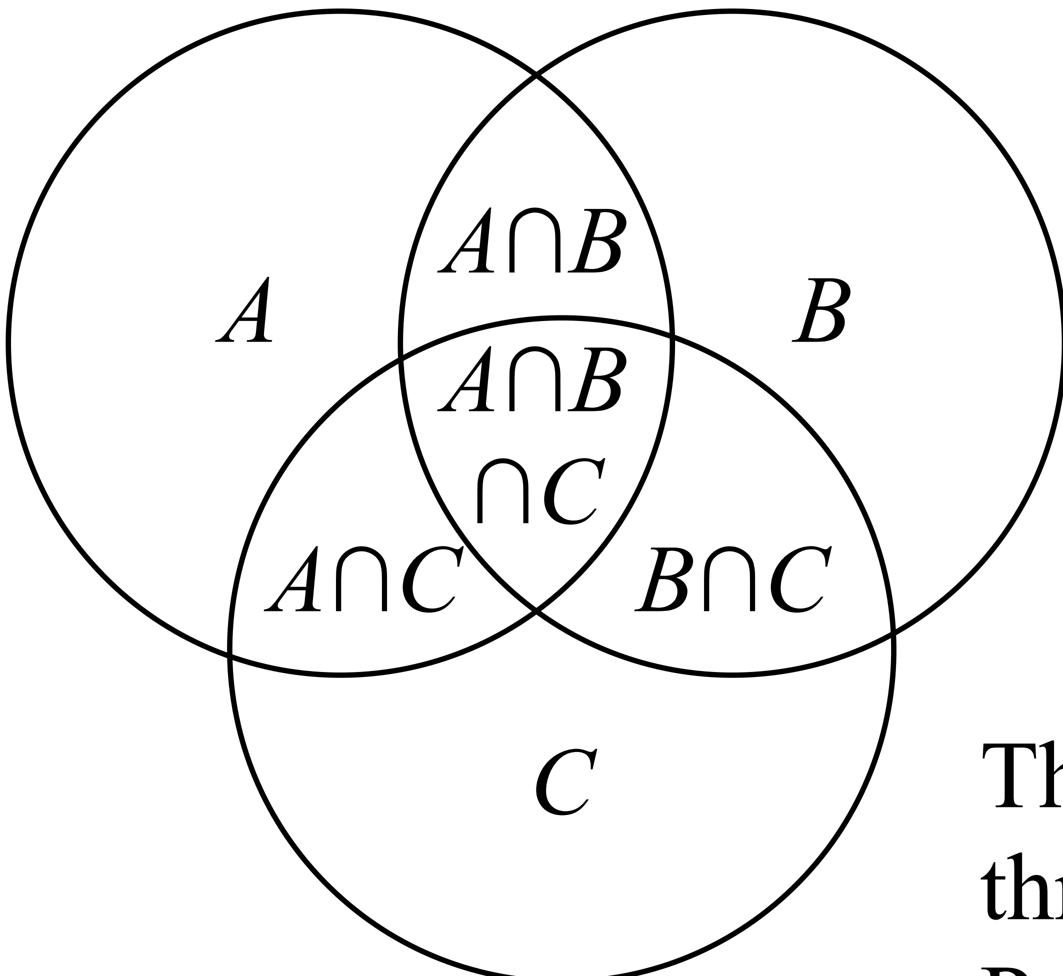
  such that $\{e \in \Omega : X(e) \leq r\} \in \mathscr{A}$ for all $r \in \mathbb{R}$
  - This is needed to define probabilities like $\Pr[a \leq X \leq b]$
  - $\Pr[X = x]$ is a shorthand for $\Pr[\{e \in \Omega : X(e) = x\}]$
- An r.v. is **discrete** if it takes at most countably infinite different discrete values
  - None of the complexities applies
- An r.v. is **continuous** if it varies continuously in one or more intervals
  - These are the ones that cause problems

Example r.v.'s

- **Indicator variable** $\mathbb{1}_E$ or χ_E for event $E \in \mathscr{A}$
  - $\mathbb{1}_E(x) = 1$ if $x \in E$ and $\mathbb{1}_E(x) = 0$ otherwise
  - $\Pr[E] = \Pr[\mathbb{1}_E = 1]$
- Let r.v. X be the number of heads in 10 coin flips
  - If e = HTTTTTHHTT, then $X(e) = 3$
  - Discrete r.v.
- Let r.v. Y be the room temperature of my kitchen (in Celsius)
  - if e = "00:22 on 22 Oct", then $X(e) = 22{,}7$
  - Continuous r.v.

Some diagrams (1)

- The **Venn diagram** is a way to visualize the combinatorial relationships of three sets



The **inclusion–exclusion principle** for three sets:

$\Pr[A \cup B \cup C] =$

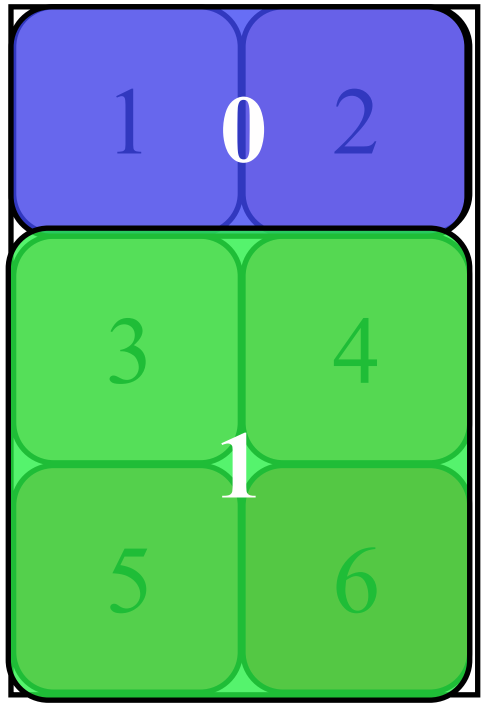
$\Pr[A] + \Pr[B] + \Pr[C]$

$- \Pr[A \cap B] - \Pr[A \cap C] - \Pr[B \cap C]$

$+ \Pr[A \cap B \cap C]$

Some diagrams (2)

- R.v. X that takes finite number of values partitions the sample space into finite sets (the pre-image of X)
  - If X is a roll of dice, we have $E_1 = \{e \in \Omega : X(e) = 1\} = X^{-1}(1)$, and similarly for E_2, E_3, …, E_6
  - If Y is indicator variable for "X ≥ 2", we get

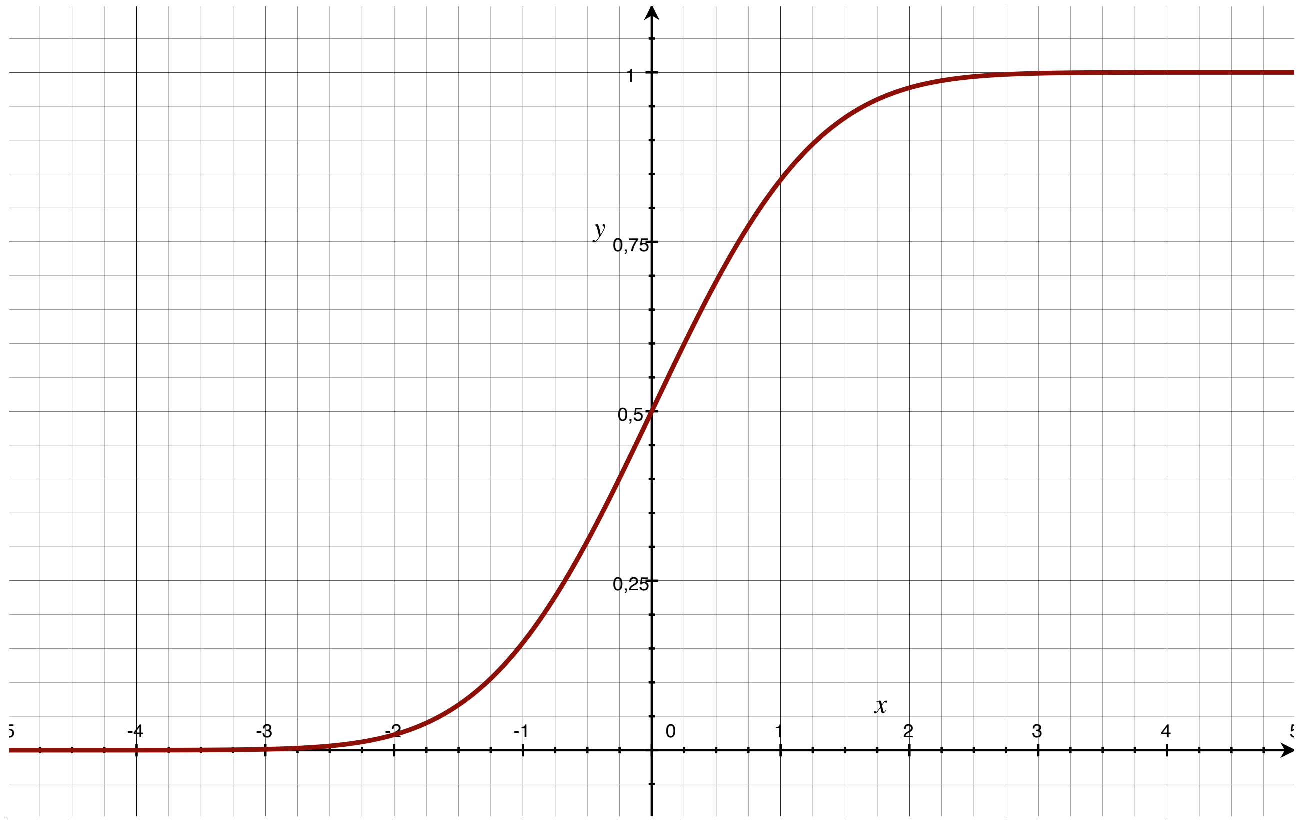


Distributions

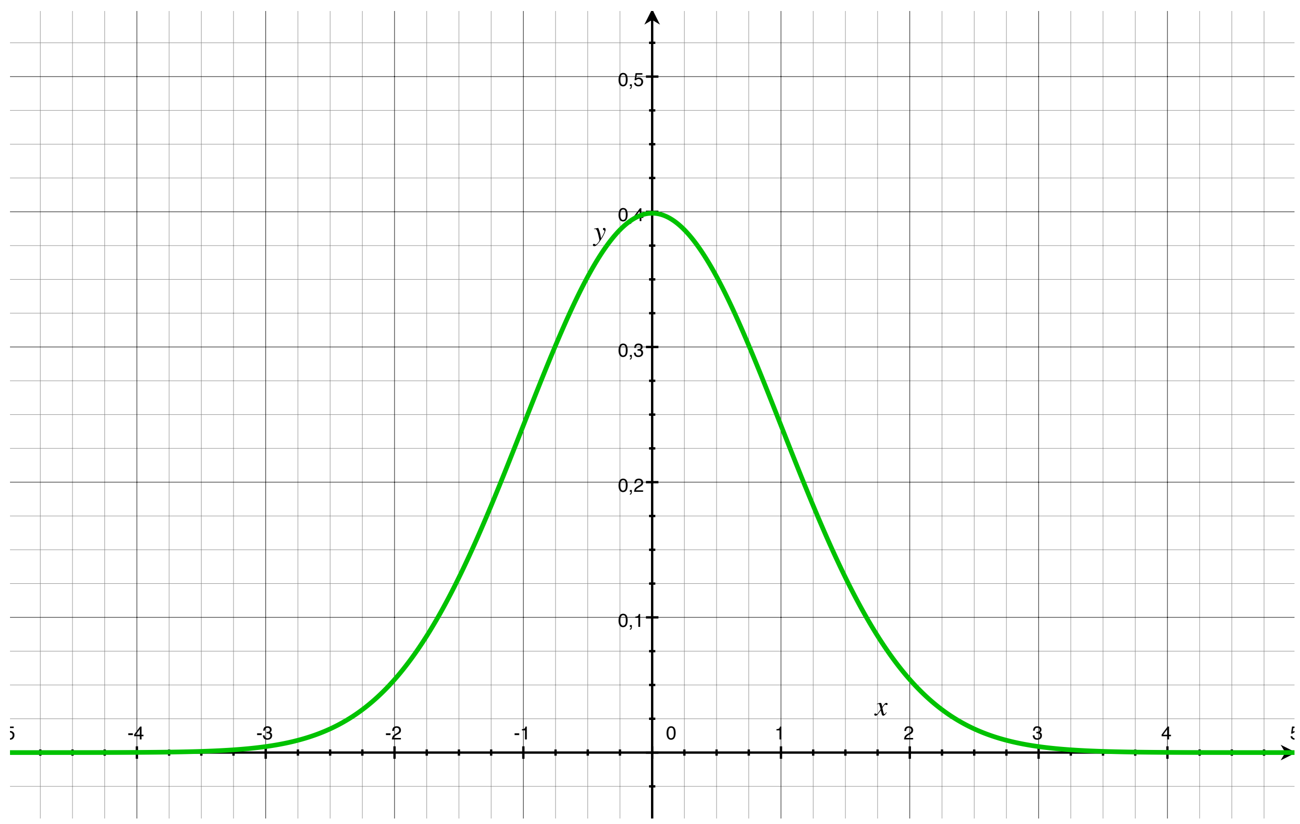
- The **cumulative distribution function** (**cdf**) of r.v. X is a function F_X: $\mathbb{R} \to [0, 1]$, $F_X(x) = \Pr[X \leq x]$
- If X is discrete, the **probability mass function** (**pmf**) of X is $f_X(x) = \Pr[X = x]$
- If X is continuous, the **probability density function** (**pdf**) of X is a function f_X for which
  - $f_X(x) \geq 0$ for all x
  - $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$
  - We have that $F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$

Example of a CDF and PDF

CDF:



PDF:



Some discrete distributions

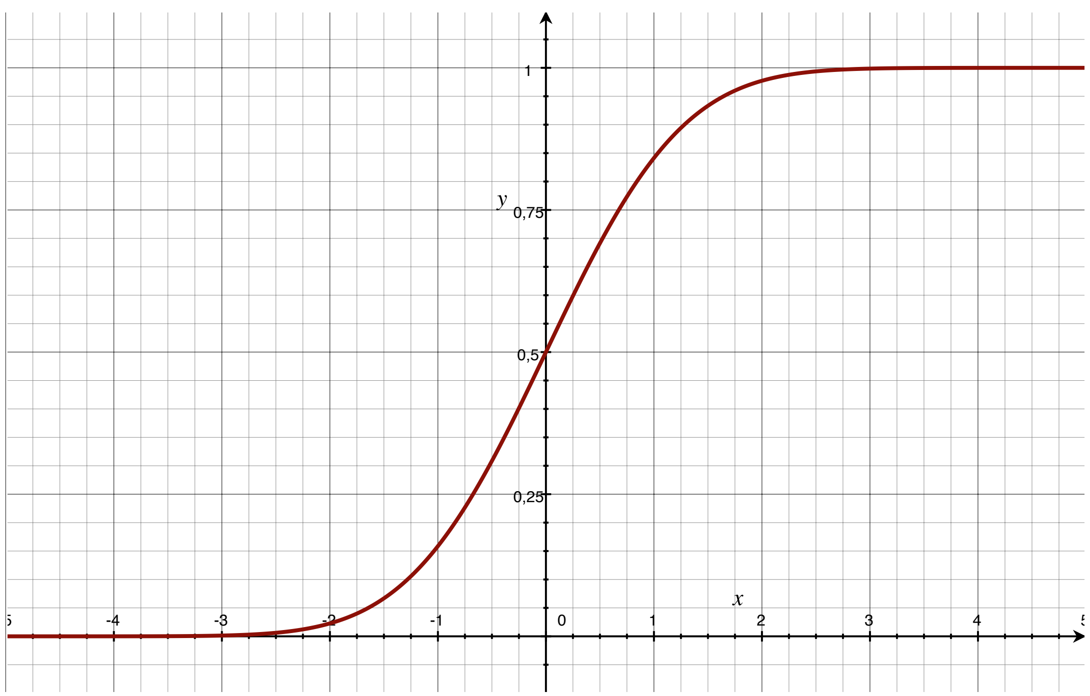
- **Uniform** distribution over $\{1, 2, \ldots, m\}$
  - $\Pr[X = k] = 1/m$ for $1 \leq k \leq m$
- **Bernoulli** distribution with parameter p
  - Binary, single coin toss
  - $\Pr[X = k] = p^k(1-p)^{1-k}$ for $k \in \{0, 1\}$
- **Binomial** distribution with parameters p and n
  - n repeated Bernoulli experiments with parameter p
  - $\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$
- **Geometric** distribution with parameter p
  - $\Pr[X = k] = (1-p)^k p$ for $k \geq 0$
- **Poisson** distribution with rate parameter λ
  - $\Pr[X = k] = e^{-\lambda} \lambda^k / k!$

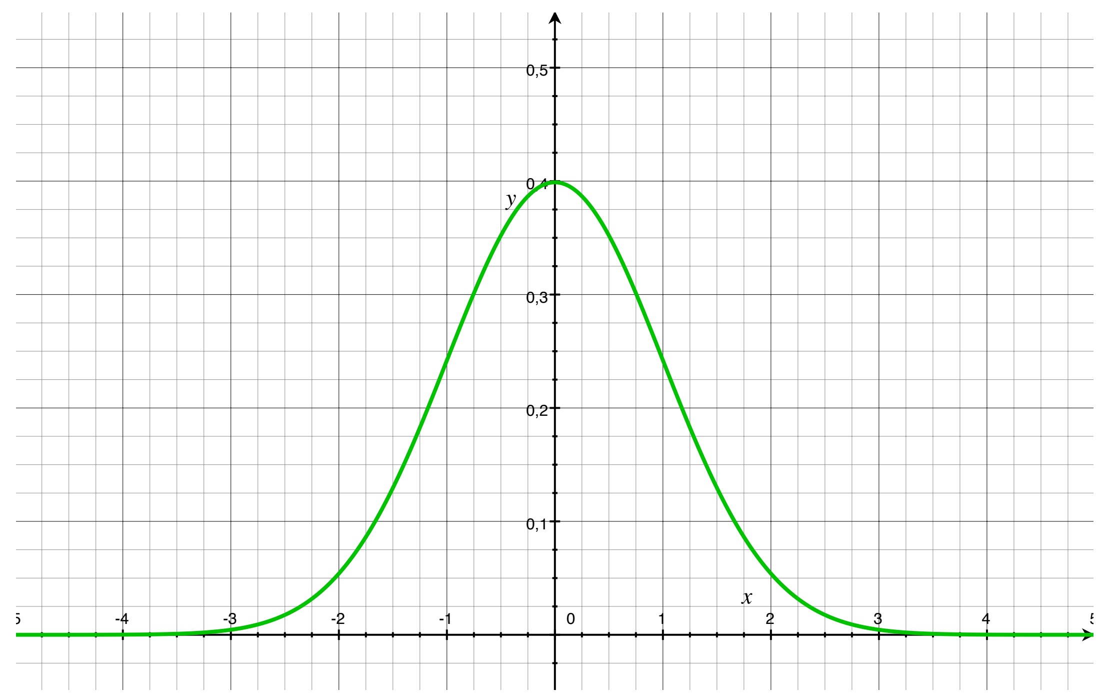
Some continuous distributions

- **Uniform** distribution in the interval $[a, b]$
  - $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$
- **Exponential** distribution with rate λ
  - Time between two events in a Poisson process
  - $f_X(x) = \lambda \mathrm{e}^{-\lambda x}$ for $x \geq 0$
- ***t*-distribution** with ν degrees of freedom
  - Typical distribution for test statistics
  - $f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
- χ^2 distribution with k degrees of freedom
  - $f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} \mathrm{e}^{-\frac{x}{2}}$

Normal (Gaussian) distribution

- Two parameters, μ (mean) and σ^2 (variance)
  - $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- For **standard normal distribution** $\mu = 0$ and $\sigma^2 = 1$
- Many, many applications





- R.v. X is **log-normally** distributed if its logarithm is normally distributed

Multivariate distributions

- If X and Y are two discrete variables, their **joint mass function** is $f_{X,Y}(x, y) = \Pr[X = x, Y = y]$
  - For continuous variables it is a non-negative function s.t.
    - $f_{X,Y}(x,y) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y) \mathrm{d}x \mathrm{d}y = 1$
    - for any $A \in \mathbb{R} \times \mathbb{R}$, $\Pr[(X,Y) \in A] = \iint_A f_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y$
- The **marginal distribution** (mass function) for X is
  - $f_X(x) = \Pr[X = x] = \sum_y f_{X,Y}(x,y)$ for discrete X
  - $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y) \mathrm{d}y$ for continuous X
- All these concepts extend naturally to more than two variables

Multivariate normal distribution

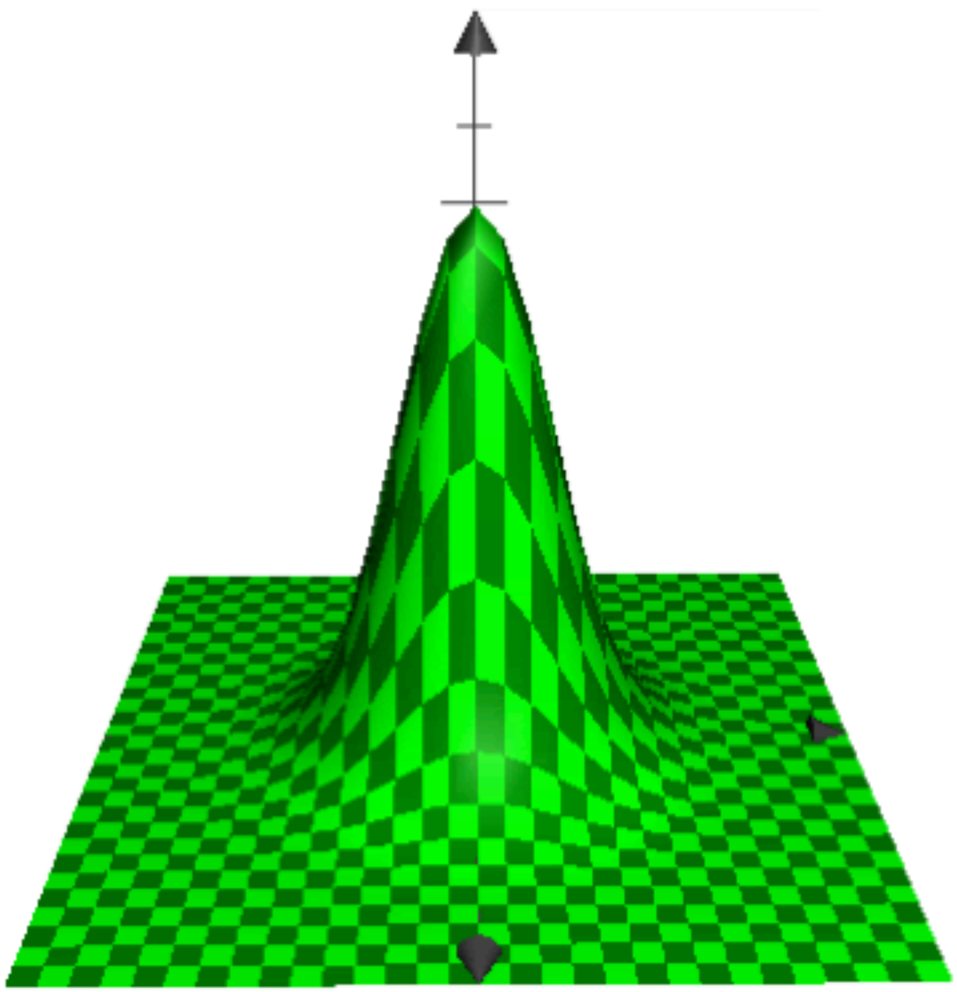
- A.k.a. multidimensional Gaussian distribution
- Two variables, vector $\boldsymbol{\mu}$ and matrix $\boldsymbol{\Sigma}$
  - For n variables, $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$
- The density function is

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

- In the **standard multivariate normal** distribution, $\boldsymbol{\mu}$ is all-zeros and $\boldsymbol{\Sigma}$ is the identity, giving

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}} \exp\left\{\tfrac{1}{2}\boldsymbol{x}^T \boldsymbol{x}\right\}$$

Bivariate normal distribution



Independence, moments & Bayes'

- Two events A and B are **independent** if $\Pr[A \cap B] = \Pr[A]\Pr[B]$
- Two r.v.'s X and Y are independent if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y
- The **conditional probability** of A given B is $\Pr[A \mid B] = \Pr[A \cap B]/\Pr[B]$
  - Assumes $\Pr[B] > 0$
  - If A and B are independent, $\Pr[A \mid B] = \Pr[A]$
- The **conditional pmf/pdf** is $f_{X|Y}(x \mid y) = f_{X,Y}(x, y)/f_Y(y)$
  - For independent X and Y, $f_{X|Y}(x \mid y) = f_X(x)$
- A and B are **conditionally independent** given C if $\Pr[A \cap B \mid C] = \Pr[A \mid C]\Pr[B \mid C]$

Example

- Test for sickness with outcomes + and –

	sick	healthy
+	0.009	0.099
–	0.001	0.891

- Test seems to work:
  - $\Pr[+ \mid \text{sick}] = \Pr[+ \cap \text{sick}]/\Pr[\text{sick}] = 0.9$
  - $\Pr[- \mid \text{healthy}] \approx 0.9$
- But what is the probability that you are sick if you get +?
  - $\Pr[\text{sick} \mid +] = \Pr[+ \cap \text{sick}]/\Pr[+] \approx 0.08$

Bayes' theorem and total probability

- The **law of total probability** states that if $A_1, A_2, \ldots, A_k$ partition Ω, then for any event B

$$\Pr[B] = \sum_{i=1}^{k} \Pr[B \mid A_i] \Pr[A_i]$$

  - Sum B piece-wise over A_i's

- The **Bayes' theorem** states that if $A_1, A_2, \ldots, A_k$ is partition of Ω s.t. $\Pr[A_i] > 0$ for all i, then for any B s.t. $\Pr[B] > 0$ and for each $i = 1, \ldots, k$

$$\Pr[A_i \mid B] = \frac{\Pr[B \mid A_i] \Pr[A_i]}{\sum_{j=1}^{k} \Pr[B \mid A_j] \Pr[A_j]}$$

  - $\Pr[A_i]$ is the **prior probability** and $\Pr[A_i \mid B]$ the **posterior probability**

Expectation and variance

- The **expected value** or r.v. X is
  - $E[X] = \sum_k k f_X(k)$ for discrete X
  - $E[X] = \int_{\mathbb{R}} x f_X(x) \mathrm{d}x$ for continuous X
    - Exists only if $\int |x| f_X(x) \mathrm{d}x < \infty$
- The ***i*-th moment** is $E[X^i] = \int_{\mathbb{R}} x^i f_X(x) \mathrm{d}x$
  - Assuming that $\int \left|x^i\right| f_X(x) \mathrm{d}x < \infty$
- The **variance** of X is $V[X] = E[(X - E[X])^2]$ $= E[X^2] - E[X]^2$
  - Also denoted by σ^2
  - **Standard deviation** sd(X) is $\sqrt{V[X]}$

Properties of expectation and variance

- $E[aX + b] = aE[X] + b$ for constants a and b
- $E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]$
  - **Linearity of expectation**
  - Works for any X_i's (e.g. don't have to be independent)
- $E[XY] = E[X]E[Y]$ for *independent* X and Y
- $V[aX + b] = a^2V[X]$ for constants a and b
- $V[X_1 + X_2 + \ldots + X_n] = V[X_1] + V[X_2] + \ldots + V[X_n]$
  - For *independent* X_i's

Correlation and covariance

- The **covariance** between r.v.'s X and Y is $\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$
  - $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$
    - $\mathrm{Cov}(X, X) = V[X]$
  - If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$
    - The converse is not generally true
- The **correlation** between X and Y is $\rho_{X,Y} = \mathrm{Cov}(X, Y)/(\mathrm{sd}(X)\times\mathrm{sd}(Y))$
  - We have $-1 \le \rho_{X,Y} \le 1$
  - If $Y = aX + b$ for some constants a and b, then $\rho_{X,Y} = \mathrm{sign}(a)$ (i.e. either -1 or 1)

Conditional expectation

- The **conditional expectation** of X given $Y = y$ is
  - $E[X \mid Y = y] = \sum x f_{X|Y}(x \mid y)$ for discrete X
  - $E[X \mid Y = y] = \int x f_{X|Y}(x \mid y)\mathrm{d}x$ for continuous X
- The conditional expectation $\mathrm{E}[X \mid Y]$ is a r.v. of Y
  - It only becomes a number when we observe $Y = y$
  - If X is a roll of dice and Y is an indicator variable for event "$X \geq 5$", then $\mathrm{E}[X \mid Y]$ is
    - $(1 + 2 + 3 + 4)\times(1/6)/(4/6) = 2.5$ if $Y = 0$
    - $(5 + 6)\times(1/6)/(2/6) = 5.5$ if $Y = 1$

Bounds and convergence

- Sometimes we don't know everything about a r.v., but we want to still study its behaviour
  - E.g. we want to bound the "tail probability"
- Trivial bound: If $E[X]$ exists, then $\Pr[X \leq E[X]] > 0$
  - Also $\Pr[X \geq E[X]] > 0$
- **Markov's inequality**: $\Pr[X \geq t] \leq E[X]/t$
  - Assumes X is nonnegative and $t > 0$
- **Chebyshev's inequality**: $\Pr[|X - E[X]| \geq t] \leq V[X]/t^2$
  - Any X, $t > 0$
  - Corollary of Markov's with $(X - E[X])^2$ as the r.v.

More bounds

- **Chernoff–Hoeffding**: If $X_1, \ldots X_n \sim$ Bernoulli(p), then for any $\varepsilon > 0$, $\Pr[|\bar{X}_n - p| > \varepsilon] \leq 2e^{-2n\varepsilon^2}$
  - $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$
  - A large family of inequalities for different settings
- **Mill's inequality**: $\Pr[|Z| > t] \leq \sqrt{\frac{2}{\pi}} \frac{\exp\{-t^2/2\}}{t}$ for $Z \sim N(0, 1)$ and $t > 0$
- **Cauchy–Schwartz**: $|E[XY]|^2 \leq E[X^2]E[Y^2]$
  - Assumes finite variances
- **Jensen's inequality**: $E[g(X)] \geq g(E[X])$ for convex g and $E[g(X)] \leq g(E[X])$ for concave g

Convergence

- A sequence $X_1, X_2, \ldots$ of r.v.'s can **converge** to r.v. X in the following senses
  - X_n converges to X **almost surely**, $X_n \rightarrow_{\text{a.s.}} X$, if $\Pr[\lim_{n\to\infty} X_n = X] = 1$
  - X_n converges to X in **probability**, $X_n \rightarrow_{\text{P}} X$, if for every $\varepsilon > 0$, $\Pr[|X_n - X| > \varepsilon] \rightarrow 0$ as $n \rightarrow \infty$
  - X_n converges to X in **distribution**, $X_n \rightarrow_{\text{D}} X$, if $\lim_{n\to\infty} F_n(x) = F(x)$ at all points where $F(x)$ is continuous
    - F_n is the cdf of X_n and F the cdf of X
- Almost sure convergence implies convergence in probability implies convergence in distribution

Laws of large numbers

- The **weak law of large numbers** states that if X_1, X_2, …, X_n are independent and identically distributed (i.i.d.) r.v.'s with mean μ, then
$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i \to_{\mathrm{P}} \mu \ .$$
- The **strong law of large numbers** replaces the convergence in probability with almost sure convergence
- The laws of large numbers show that the expected value is the average value over infinite number of repetitions

Central limit theorem

- If $X_1, X_2, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, and if $X \sim N(\mu, \sigma^2/n)$, then per the **central limit theorem**, $\bar{X}_n \to_{\mathrm{D}} X$.
  - Does not depend on distributions of X_i
    - Except that they must have mean and variance
  - One main reason why normal distribution is ubiquitous

Statistical inference

- A **statistical model** M is a set of distributions
  - All smooth distributions, all unimodal distributions, all discrete distributions with mean 1, …
- M is **parametric model** if it can be completely described with a finite number of parameters
  - E.g. the family of Normal distributions with parameters μ and σ^2

    $$M = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$$

Statistical inference

- Given a parametric model M and a sample $X_1, \ldots, X_n$, how do we infer the parameters of M?
- The **sample mean** is $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$
- The **sample variance** is

$$S^2_{X_n} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

- The **bias** of the estimator $\hat{\theta}$ for parameter θ is $E[\hat{\theta}] - \theta$
  - The estimator is **unbiased** if it has bias 0

Summary

- What "probability" means is debatable
  - Axiomatic approach side-steps interpretation issues
- With discrete r.v.'s, most of prob. theory is simple combinatorics
  - Continuous variables are more problematic
- Conditional expectation is a random variable!