# Essential Probability

Brad Jacobs

April 2016

# Introduction

Probability provides the mathematical tools we use to model randomness:

- Probability tells us how likely an event (Frequentist) is or how likely our beliefs are to be correct (Bayesian)
- Provides the foundation for statistics and machine learning
- Often our intuitions about randomness are incorrect because we live only one realization
- Enumerating all possible outcomes (using combinatorics) can help us compute the probability of an event

# Combinatorics

# Factorial

*Factorial* counts the number of ways of ordering or picking something when order matters:

- We write $n! = n \times (n-1) \times ... \times 2 \times 1$
- $0! = 1$ by convention
- Example: how many ways can we shuffle a deck of cards?

# Permutation

*Permutations* are the number of ways to choose a group from a larger population where order matters:

- Select $k$ representatives *in order* from a population of size $n$: $\dfrac{n!}{(n-k)!}$
- Example: In baseball a manager sets the batting order for 9 players out of a team of 25.

# Combination

*Combination* counts the number of ways of picking something when order doesn't matter:

- $\binom{n}{k} = \frac{n!}{(n-k)!k!}$
- We say '*n choose k*'
- This is the number of ways of choosing $k$ items from $n$ total items
- Example: In a class of 20 students how many pairs are there for afternoon sprints?

# Probability

# Review: sets

Some definitions:

- The *set* $S$ that consists of all possible outcomes or events is called the *sample space*
- *Union*: $A \cup B = \{x : x \in A \, \texttt{or} \, x \in B\}$
- *Intersection*: $A \cap B = \{x : x \in A, \texttt{and} \, x \in B\}$
    - *Disjoint*: $A \cap B = \emptyset$
- *Complement*: $A^c = \{x : x \notin A\}$
- *Partition*: a set of pairwise disjoint sets, $\{A_j\}$, such that $\overset{\infty}{\underset{j=1}{\cup}} A_j = S$
- *Subset*: a.k.a. "$A$ is included in $B$" is written as: $A \subset B$

# Definition of probability

Given a sample space, $S$, a *probability function*, $Pr$, has three properties:

- $\Pr[A] \geq 0, \forall A \subset S$
- $\Pr[S] = 1$
- For a set of pairwise disjoint sets $\{A_j\}$, $\Pr[\bigcup_j A_j] = \sum_j \Pr[A_j]$

Note: this means $\Pr[A] = 1 - \Pr[A^c]$

# Example: tossing a coin

Consider a coin toss:

- $S = \{H, T\}$
- $\Pr[H] = \Pr[T] = \dfrac{1}{2} > 0$
- $\Pr[S] = 1$

# Independence

Two events $A$ and $B$ are said to be *independent* if

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

or, equivalently, if

$$\Pr[B|A] = \Pr[B],$$

i.e., knowledge of $A$ provides no information about $B$

# Multiplication rule

To compute the probability that two *independent* events occur, multiply their probabilities:

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

- What is the probability that A and B happen?
  - Under independence this joint probability is easy to calculate.

# Conditional probability

We often care about whether one event provides information about another event. The *conditional probability* of $B$ given $A$ is:

$$\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]}$$

- We say this is the '*probability of $B$ conditional on $A$*'
- I.e., if $A$ has occurred, what is the probability $B$ will occur?

# Example: coin tosses

Take a moment to solve this question:

- Three types of fair coins are in an urn: HH, HT, and TT
- You pull a coin out of the urn, flip it, and it comes up H
- **Q**: what is the probability it comes up H if you flip it a second time?

# Probability chain rule

Can condition on an arbitrary number of variables:

- Simple example:

$$\Pr[A_3, A_2, A_1] = \Pr[A_3 | A_2, A_1] \cdot \Pr[A_2 | A_1] \cdot \Pr[A_1]$$

- General case:

$$\Pr[A_n, ..., A_1] = \prod_j \Pr[A_j | A_{j-1}, ..., A_1]$$

or

$$\Pr[\bigcap_j^n A_j] = \prod_j^n \Pr[A_j | \bigcap_k^{j-1} A_k]$$

# Law of total probability

If $\{B_n\}$ is a partition of the sample space, the *Law of total probability* states:

$$\Pr[A] = \sum_j \Pr[A \cap B_j]$$

or

$$\Pr[A] = \sum_j \Pr[A|B_j] \cdot \Pr[B_j]$$

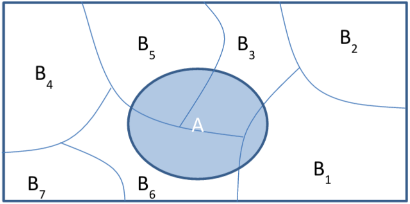



Figure 1

# Bayes's Rule

Use Bayes's Rule when you need to compute conditional probability for $B|A$ but only have probability for $A|B$:

$$\Pr[B|A] = \frac{\Pr[A|B] \cdot \Pr[B]}{\Pr[A]}$$

- Proof: use the definition of conditional probability
- For an arbitrary partition of event space, $\{A_j\}$, use the general form of Bayes's rule:

$$\Pr[A_k|B] = \frac{\Pr[B|A_k] \cdot \Pr[A_k]}{\sum_j \Pr[B|A_j] \cdot \Pr[A_j]}$$

# Example: drug testing

A test for EPO has the following properties:

| Variable | Value |
|---|---|
| $\Pr[+ \mid \textit{doped}]$ | 0.99 |
| $\Pr[+ \mid \textit{clean}]$ | 0.05 |
| $\Pr[\textit{doped}]$ | 0.005 |

**Q:** What is the probability the cyclist is using EPO if the test is positive? I.e., what is $\Pr[\textit{doped} \mid +]$?

# Solution: drug testing

1. Compute probability of being clean:

$$\Pr[clean] = 1 - \Pr[doped]$$

2. Use Bayes's Rule:

$$\Pr[doped|+] = \frac{\Pr[+|doped] \cdot \Pr[doped]}{\Pr[+|doped] \cdot \Pr[doped] + \Pr[+|clean] \cdot \Pr[clean]}$$

$$= \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.05 \cdot (1 - 0.005)}$$

$$= 0.090$$

Based on this example

# Random variables and probability distributions

# Definition: random variable

Given a sample space $S$, a *random variable*, $X$, is a function such that $X(s) : s \in S \mapsto \mathbb{R}$:

- Can think of r.v. as summary of an experiment.
  - Simplest experiment: Flip a coin $n$ times.
  - If $n = 3$ there are 8 possible outcomes
  - One way to summarize: X = number of heads seen
  - Thus for each outcome X = 0, 1, 2 or 3
- By convention, capital letters to refer to a random variable and lower case to refer to a specific realization: $X = x$
  - $\Pr[X = x] = \Pr[\{s \in S : X(s) = x\}]$

# Cumulative distribution function (CDF)

Definition: the cumulative distribution function $F_X(x) = \Pr[X \leq x]$:

- Properties:
  - $0 \leq F_X(x) \leq 1$
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to \infty} F_X(x) = 1$
  - $F_X(x)$ is monotonically increasing
- Applies to discrete and continuous random variables
- Note: $\Pr[a < X \leq b] = F_X(b) - F_X(a)$

# Discrete: probability mass function (PMF)

For a random variable, $X$, which takes discrete values $\{x_i\}$, use a PMF to determine the probability of an individual event:

- $f_X(x) = Pr[X = x], \forall x$
- We say there is *probability mass* $p_i$ on $x_i$, where $p_i = \Pr[X = x_i]$
- Example: tossing coins
  - $X \in \{H, T\}$
  - $p_H = p_T = \frac{1}{2}$

# Continuous probability density function (PDF)

For a continuous random variable, $X$, use a PDF:

- $f_X(x)dx = \Pr[x < X < x + dx]$
- Going between CDF and PDF
  - $f_X(x) = \frac{dF_x(x)}{dx}$, assuming some regularity conditions
  - $F_X(x) = \int_{-\infty}^{x} f_X(s)ds$

# Properties of distributions

Use these properties to characterize a distribution:

- Expectation/mean
- Variance/standard deviation
- Skew
- Kurtosis
- Correlation

We often compute sample analogs of these properties to compare the empirical distribution of our data to standard distributions

# Expectation/mean

The *expectation*, *mean*, or *expected value* is a measure of what is a likely value of a random variable:

- $\mu_X = \mathbb{E}[X]$ :
  - Discrete: $\mathbb{E}[X] = \sum_{s \in \{x_i\}} s f_X(s)$
  - Continuous: $\mathbb{E}[X] = \int_{-\infty}^{\infty} s f_X(s) ds$
- Expectation is a linear operator
- The sample mean is $\overline{x} = \dfrac{1}{n} \sum_{j=1}^{n} x_j$

# Variance

*Variance* measures the spread of a distribution:

- $\mathtt{Var}[X] = \mathbb{E}_X[(X - \mu_x)^2]$
- Sometimes variance is written as $\sigma^2(X) = \mathtt{Var}[X]$
- Often, we use *standard deviation*, $\sigma(X) = \sqrt{\mathtt{Var}[X]}$ which has the same dimensions as $X$
- $\mathtt{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
- Note: the sample variance is $s^2 = \dfrac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$

# Skew and kurtosis

Skew and kurtosis are higher order moments:

- Skewness:
  - $\gamma_1 = \mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^3]$
  - Measures asymmetry of a distribution
  - Sign of skewness tells whether distribution is left or right skewed
- Kurtosis:
  - $\kappa = \mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^4]$
  - Measures the 'fatness' of the tails of the distribution

# Multivariate Distributions

Often interested in a joint distribution between two (or more) random variables:

- Extend definitions of CDF, PDF/PMF, Mean
- New multivariate moments:
  - Covariance: $\texttt{Cov}[X, Y] = \mathbb{E}[(X - \mu_x) \cdot (Y - \mu_y)]$
  - Correlation: $\rho_{XY} = \dfrac{\texttt{Cov}[X, Y]}{\sigma_x \cdot \sigma_y}$

# Marginal and conditional distributions

To compute the marginal distribution from the joint (multivariate) distribution, just integrate (sum) over the other variable(s):

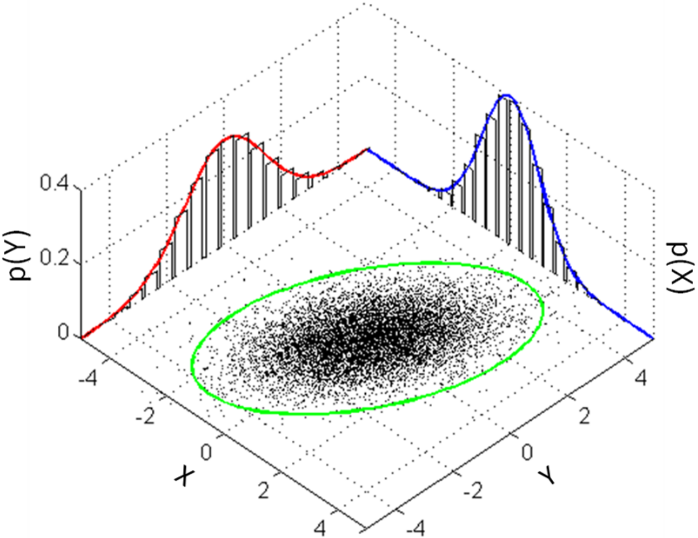
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, s) ds$$

For a bivariate distribution, conditional pdf is:

$$f(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

# Example:

*Joint Distribution*



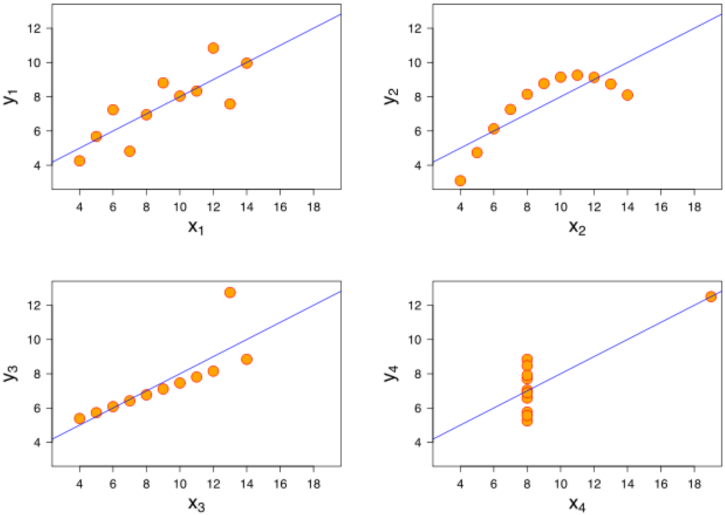
# Covariance and correlation

To explore the relationship between variables compute:

- *Covariance*:
  - $\mathtt{Cov}(x, y) = \mathbb{E}[(x - \mu_x) \cdot (y - \mu_y)]$
  - Size changes with scaling of variables
  - For random variables which are vectors, use $\mathtt{Cov}[x, y] = \mathbb{E}[(x - \mu_x) \cdot (y - \mu_y)^T]$
- *Correlation* (Pearson):
  - Dimensionless measure relationship
  - $\rho_{XY}(x, y) = \dfrac{\mathtt{Cov}(x, y)}{\sigma(x) \cdot \sigma(y)}$
  - Thus, $\rho_{XY} \in [-1, 1]$
  - Other correlation coefficients, such as Spearman, use rank and are more robust
- Correlation is not causation!

# Correlation and linearity

*Correlation and linearity:* $r = 0.816$.



# Correlation captures noisiness and direction

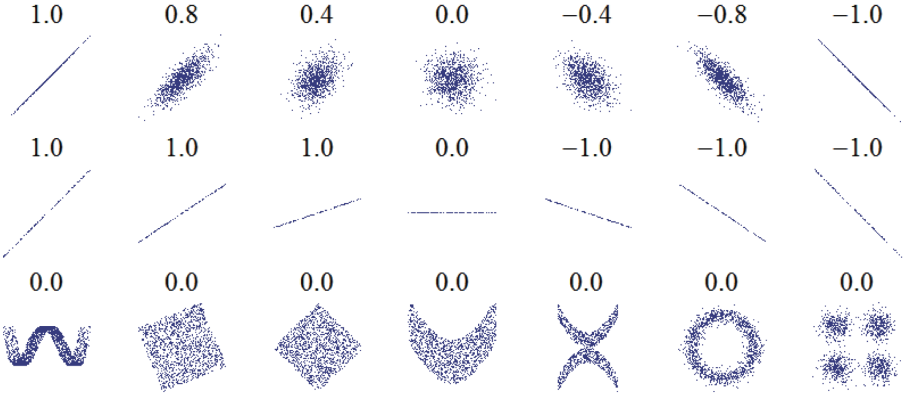



Figure 2:Correlation and non-linearity. From Wikipedia.

# Common distributions

# Overview

We now review the properties of some common distributions:

- Discrete
  - Bernoulli
  - Binomial
  - Geometric
  - Poisson
- Continuous
  - Uniform
  - Exponential
  - Gaussian a.k.a. Normal
  - $\chi^2$
  - Student's t
  - F distribution

# Distribution Notation

- We write $X \sim \texttt{XYZ}(\alpha, \beta, \ldots)$ to mean $X$ is distributed like the *XYZ* distribution with parameters $\alpha$, $\beta$, ...
- We say a series of random variables are *i.i.d.* if they are '*independent and identically distributed*'
- Example: $X \sim \texttt{N}(\mu, \sigma^2)$ or $X \sim \texttt{U}(0, 1)$

# Bernoulli

Models a toss of an unfair coin or clicking on a website:

- $X \sim \texttt{Bernoulli}(p)$
- PMF: $\Pr[H] = p$ and $\Pr[T] = 1 - p$
- Mean: $\mathbb{E}[x] = p$
- Variance: $\texttt{Var}[x] = p \cdot (1 - p)$

# Example: click through rate

Given $N$ visitors of whom $n$ click on the 'Buy' button:

- What is click through rate (CTR)?
- What is the variance of the click through rate?

# Binomial

Models repeated tosses of a coin:

- $X \sim \texttt{Binomial}(n, p)$ for $n$ tosses of a coin where $\Pr[H] = p$
- PMF: $\Pr[X = k] = \binom{n}{k} p^k \cdot (1 - p)^{(n-k)}, \forall\, 0 \leq k \leq n$
- Mean: $n \cdot p$
- Variance: $n \cdot p \cdot (1 - p)$
- Approaches Gaussian for limit of large $n$

# Geometric

Models probability succeeding on the $k$-th try:

- $X \sim \texttt{Geometric}(p, k)$
- PMF: $\Pr[X = k] = p \cdot (1 - p)^{(k-1)}$
- Mean: $\dfrac{1}{p}$
- Variance: $\dfrac{1-p}{p^2}$

# Poisson

Models number of events in a period of time, such as number of visitors to website:

- $X \sim \texttt{Poisson}(\lambda)$
- PMF: $\Pr[X = k] = \exp(-\lambda) \cdot \frac{\lambda^k}{k!}, \forall\, k = 0, 1, 2, \ldots$
- Mean $=$ variance $= \lambda$
- $\lambda$ is the number of events during the interval of interest
- Note: $\Pr[X = k]$ is just one term in the Taylor's series expansion of $\exp(x)$ when suitably normalized

# Exponential

Models survival, such as the fraction of uranium which has not decayed by time $t$ or time until a bus arrives:

- $T \sim \texttt{Exp}(\lambda)$
- $1/\lambda$ is the half-life
- CDF: $\Pr[T \leq t] = 1 - \exp(-\lambda \cdot t), x \geq 0, \lambda \geq 0$
- Mean: $1/\lambda$
- Variance: $1/\lambda^2$

# Uniform

Models a process where all values in an interval are equally likely:

- $X \sim \mathtt{U}(a, b)$
- PDF: $f(x) = \dfrac{1}{b-a}, \forall x \in [a, b]$ and 0 otherwise
- Mean: $\dfrac{a+b}{2}$
- Variance: $\dfrac{(b-a)^2}{12}$

# Gaussian a.k.a. Normal

A benchmark distribution:

- $X \sim N(\mu, \sigma^2)$
- PDF: $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$
- Mean: $\mu$
- Variance: $\sigma^2$
- 'Standard normal' is $N(0, 1)$:

This is the famous 'Bell-curve' distribution and is the *most* important due to its connection with the Central Limit Theorem (tomorrow).

# Other distributions

Some other distributions:

- $\chi^2$:
  - Models sum of $k$ squared, independent, normally-distributed random variables
  - Use for goodness of fit tests
- Student's t: distribution of the *t-statistic*:
  - t-statistic: $t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$, where $s$ is the standard error
  - Perform a 't-test' to check probability of observed value
  - Has fatter tails than normal distribution
- F-distribution:
  - Distribution of the ratio of two $\chi^2$ random variables
  - Use to test restrictions and ANOVA

# PDF and CDF Plots

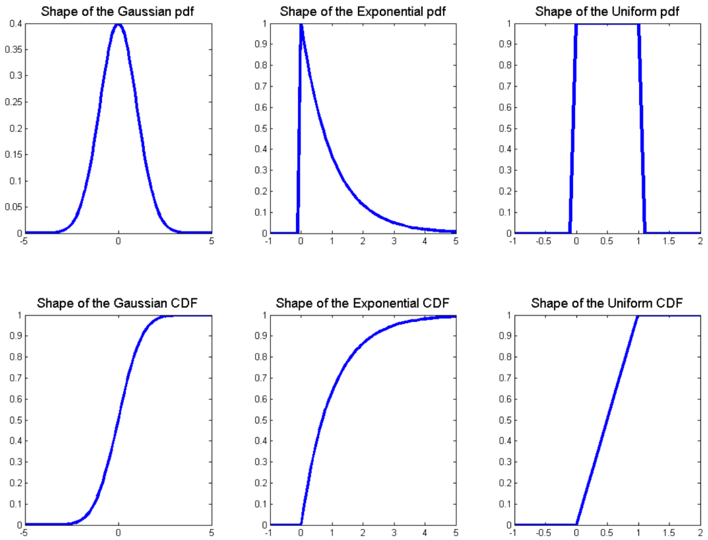



Figure 3: