Lecture 4: Probability

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

3/1/2016

Overview

Random Processes

Stochastic Process

A process with a known set of possible outcome variables, but the actual outcome that occurs is random.

- Time-independent stochastic process
  - Coin-flip
  - Roll of die
- Time-dependent stochastic process
  - Value of stock market
  - Diffusion
- Pseudo-random process
  - Some processes can be modeled as random even if they are not truly random. Animal or human behavior, for example. Even dice, coins, etc. are actually deterministic.

Terminology

Probability

The probability p_i of outcome x_i is the likelihood that it will occur.

- Frequentist interpretation: it is the proportion of occurrence of the outcome given an infinite number of repeated observations.
- Bayesian interpretation: subjective certainty of an outcome.

Examples

Coin flip

- $P(\text{heads}) = 0.5$
- $P(\text{tails}) = 0.5$

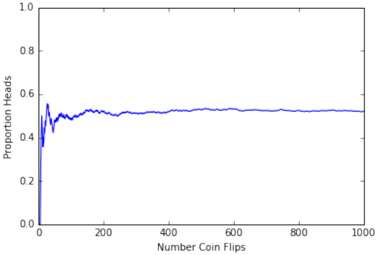
Dice roll

- $P(\text{value} = 1) = \frac{1}{6}$
- $P(\text{value} = 2) = \frac{1}{6}$
- $P(\text{value} = 3) = \frac{1}{6}$
- $P(\text{value} = 4) = \frac{1}{6}$
- $P(\text{value} = 5) = \frac{1}{6}$
- $P(\text{value} = 6) = \frac{1}{6}$

Terminology

Law of Large Numbers

As the number of observations tends to infinity, the proportion of a given outcome approaches the probability of that outcome.

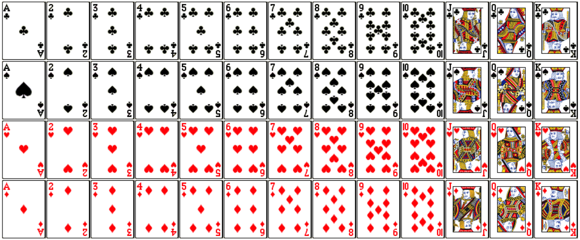


Terminology

Disjoint event

Events are disjoint if they are mutually exclusive.

- Probability of disjoint events are additive:
  $P(\text{J or Q or K}) = P(\text{J}) + P(\text{Q}) + P(\text{K}) = \frac{4}{52} + \frac{4}{52} + \frac{4}{52} = \frac{12}{52}$
- Probability of non-disjoint events:
  $P(\text{J or red}) = P(\text{J}) + P(\text{red}) - P(\text{J and red}) = \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$



Graphic from milefoot.com mathematics

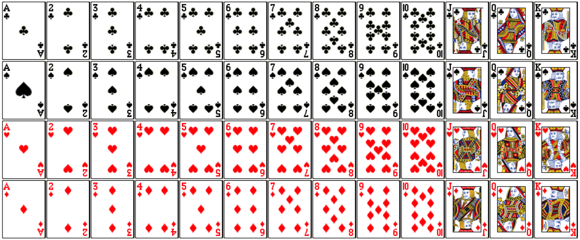
Terminology

Independent event

Knowing the outcome of one event provides no information about the outcome of the other.

- Drawing two aces in a row *with replacement* is independent.
- Drawing two aces in a row *without replacement* is dependent.

$$P(X, Y) = P(X)P(Y) \text{ if X and Y are independent events}$$



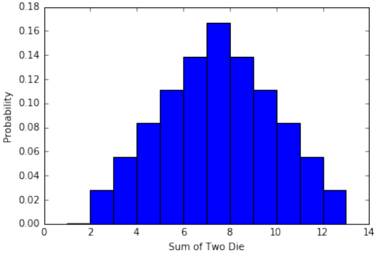
Graphic from milefoot.com mathematics

Probability Distribution

The probability p_i for each possible outcome x_i.

The probability distribution for the sum of two die is shown below in both table and graph form.

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Exercise

Law of Large Numbers and Probability Distribution exercise in IPython notebook.

Expected Value

Expected Value

The probability-weighted average of a random variable.

$$E[X] = \sum_i x_i p_i$$

Expected value is a linear operation

- If c is a constant random variable, then $E(c) = c$
- $E(X + Y) = E(X) + E(Y)$
- $E(cX) = cE(X)$

Change of Variables Theorem

- We do not need to know the distribution of a transformed variable $f(x)$ to compute its expected value, knowing the distribution of the input random variable x is enough.
- $E[f(X)] = \sum_i f(x_i)p_i$

Expected Value Calculations

- **Mean** is the expected value of the outcome. Also called the **first moment**.

$$E[X] = \sum_i x_i p_i$$

- **Variance** is the **second central moment**, defined as the expected value of the squared deviation from the mean.

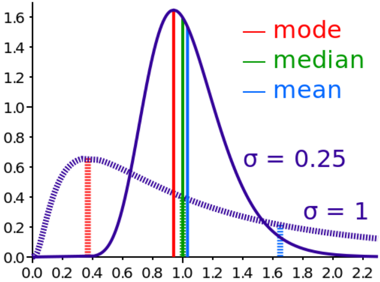
$$E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 p_i$$

- **Covariance** is a measure of the strength of correlation between random variables:

$$cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x \mu_y$$

What is "Average"?

- **Mean** - the center (of mass) of a distribution.
- **Mode** - the maximum of a distribution (i.e. the single most probable value).
- **Median** - x such that $P(X \leq x) = P(X \geq x) = \frac{1}{2}$



Graphic from Wikipedia By Cmglee - Own work, CC BY-SA 3.0

Standard Deviation

Measuring Spread of Probability Distribution

Variance:

$$var(X) = E([X - E(X)]^2)$$

Standard deviation:

$$\sigma_x = sd(X) = \sqrt{var(X)}$$

- A more computationally friendly way to calculate variance:

$$var(X) = E([X - E(X)]^2) = E(X^2) - E(X)^2$$

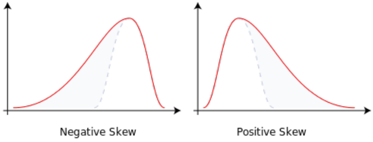
- Due to linearity of expected value:
  - $var(a + bX) = b^2 var(X)$
  - $sd(a + bX) = |b| sd(X) = |b| \sigma_x$

Long Tails are Skew

If a distribution has a long tail in one direction, it is said to be **skewed** in that direction. If the long tail is in the left direction, it is left skewed; if the long tail is in the right direction, it is right skewed.

Mathematically skew is the third moment of the Z-score, also called the Fisher-Pearson coefficient or Pearson's moment coefficient of skewness:

$$\text{skew}(X) = E\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^3\right]$$



Graphic from Wikipedia

Kurtosis

Kurtosis

Kurtosis measures the ratio "heaviness" of the tails of a unimodal, symmetric (skew=0) distribution. Higher kurtosis means more outliers.

$$\text{kurt}(X) = E\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^4\right]$$

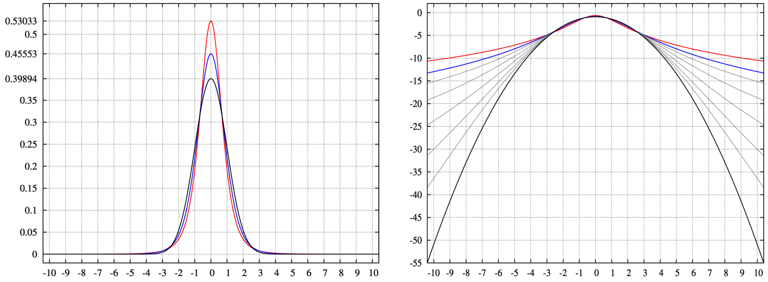
The kurtosis of the normal distribution is 3, so kurtosis is often defined as the "excess kurtosis":

$$\text{excess kurtosis} = \text{kurt}(X) - 3$$

Positive excess kurtosis means a heavy-tailed distribution and negative excess kurtosis is a light-tailed distribution.

Kurtosis

Probability distribution function (left) and log of probability distribution function (right) with excess kurtosis of infinity (red); 2 (blue); 1, 1/2, 1/4, 1/8, and 1/16 (grey); and 0 (black).

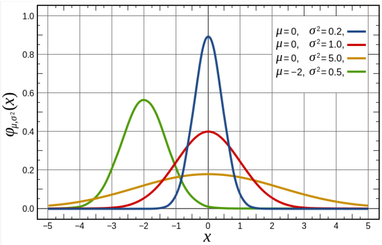


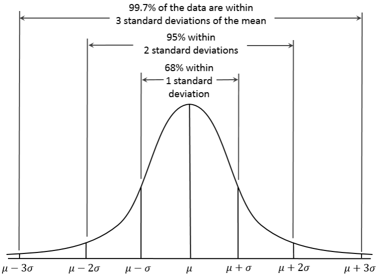
Graphic from Wikipedia

Gaussian Distribution

Also called the Normal distribution. Due to the central limit theorem, this distribution is very common in statistics.

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$





Left: Wikipedia, Right: By Dan Kernler - Own work, CC BY-SA 4.0

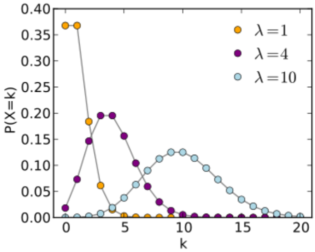
Common Probability Distributions: Poisson Distribution

Poisson Distribution

The Poisson distribution describes the likelihood of an event occurring in a fixed interval of time if the average event rate (λ) is known.

$$P(\text{observe k events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

k is a non-negative integer. Mean is $\mu = \lambda$ and standard deviation is $\sigma = \sqrt{\lambda}$



Common Probability Distributions: Binomial Distribution

Bernoulli Random Variable

- A Bernoulli random variable has two possible outcomes "success" (1) or failure (0).
- If X is a random variable with $P(X = 1) = p$ and $P(X = 0) = 1 - p$, then X is a Bernoulli random variable with mean $\mu = p$ and $\sigma = \sqrt{p(1-p)}$.

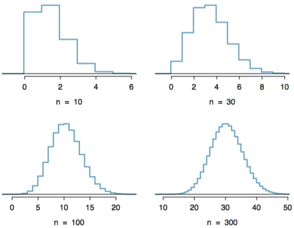
Let Y denote the number of successes in the first n trials, then the probability distribution of Y is the **binomial distribution**:

$$P(y) = \binom{n}{y} p^y (1-p)^{n-y} = \frac{n!}{k!(n-k)!} p^y (1-p)^{n-y}$$

Binomial Distribution

Normal Approximation to the Binomial Distribution

If the number of trials n is sufficiently large, then the binomial approximation is approximately equal to the normal distribution with mean $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. The condition is that $np > 10$ and $n(1-p) > 10$.



Binomial distribution with $p = 0.10$, n shown below histogram. [Diez, 2016]

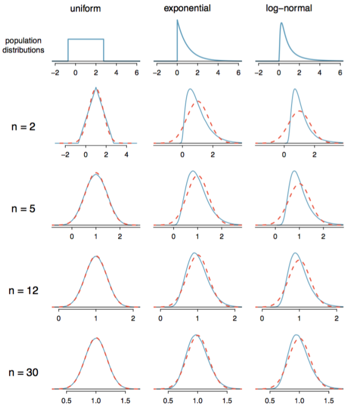
Exercise

Create a histogram of values drawn from the distributions presented using numpy.random. Vary the sample size. Add a line graph of the mathematical representation of the distribution, and vertical lines showing the population mean, population median, and mode. Add a horizontal line showing the population standard deviation. Calculate the sample mean and sample standard deviation, compare with the values for the population.

Central Limit Theorem

Central Limit Theorem

The mean of a large number of independent, identically distributed variables will be approximately normal, for all underlying distributions.



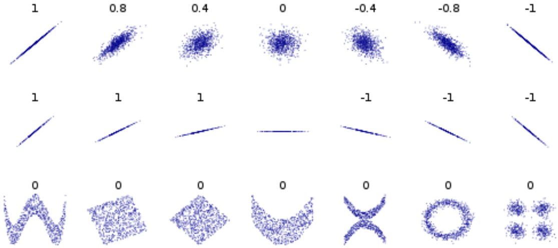
Graphic from [Diez, 2016]

Correlation

Correlation Coefficient

Also known as Pearson's [product-moment] coefficient measures the linear correlation between two random variables X and Y.

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$



Conditional Probability

Marginal Probability

Probability based on only one variable. So-called because it is calculated in the margins of a two-way probability distribution table. $P(A)$ or $P(B)$.

Joint Probability

Probability of two or more variables at the same time. $P(A \text{ and } B)$.

Conditional Probability

Probability of condition A given condition B:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Bayes' Theorem

Bayes' theorem provides a method to calculate the probability of an event (A) in a certain context (B), based on knowing the overall probability of the event, the overall probability of the context, and the probability of the context given the event:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Law of Total Probability

Bayes' Theorem can be derived from the Law of Total Probability:

$$P(E) = \sum_i g(x_i)P(E|X = x_i)$$

where g(x) is the probability distribution for x.

Exercise

Random walk exercise in IPython notebook.

References

Kyle Siegrist
Probability, Mathematical Statistics, Stochastic Processes

David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Recommended Reading

OpenIntro Statistics, Chapters 2-3
Data Science from Scratch, Chapter 6

For discussion
Video: The Best (and Worst) Ways to Shuffle Cards
Histohram intersection for change detection