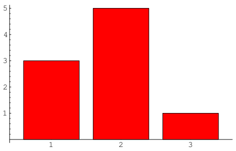
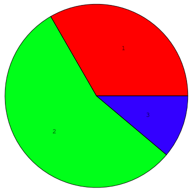
MULTIVARIABLE CALCULUS IN PROBABILITY Math21a, O. Knill

Some topics in discrete and continuous probability theory use or rely on multivariable calculus. For example, random variables in finite probability spaces are **finite dimensional vectors**. With the dot product $(f,g) = \frac{1}{n}\sum_{i=1}^n f_i g_i$, statistical quantities like expectation, variance or covariance can be reformulated geometrically:





$\mathrm{E}[f] = (1,f)$	expectation	dot product with $1 = (1,...,1)$
$\mathrm{Var}[f] = (f-\mathrm{E}[f], f-\mathrm{E}[f])$	variance	square of the length of $f-\mathrm{E}[f]$.
$\mathrm{Cov}[f,g] = (f-\mathrm{E}[f], g-\mathrm{E}[g])$	covariance	dot product of $f-\mathrm{E}[f]$ and $g-\mathrm{E}[f]$.
$\sigma[f] = \sqrt{\mathrm{Var}[f]}$	standard deviation	length of $f-\mathrm{E}[f]$.
$\mathrm{Corr}[f,g] = (f-\mathrm{E}[f], g-\mathrm{E}[g])/(\sigma[f]\sigma[g])$	correlation	$\cos(\alpha)$, angle α

For example, two random variables f and g are **uncorrelated** if and only if $f-\mathrm{E}[f], g-\mathrm{E}[g]$ are **orthogonal**. Multivariable calculus can also be used to select out some probability distributions like for example the Bernoulli distribution on the probability space of all $0-1$ sequences of length n. Multiple integrals in multivariable calculus are used when computing expectations. For random vectors, one has to integrate functions of several variables.

ABOUT PROBABILITY THEORY.

Probability emerged in the 17th century as a systematic mathematical study. While during the 18th and 19th centuries the attitudes shifted for more clarity a solid foundation of probability theory was only laid in the 20'th century, when Kolmogorov published "General Theory of Measure and Probability Theory". Published in 1929, it gave the first description of an rigorous construction of probability theory based on integration. In 1933 Kolmogorov expanded the paper into the monograph "Grundbegriffe der Wahrscheinlichkeitsrechnung", which in 1950 was translated into English under the title "Foundations of the Theory of Probability".
Nowadays, probability theory is closely connected to other fields in Mathematics like combinatorics, dynamical systems theory or the theory of partial differential equations. It appears also in seemingly unrelated fields like number theory, logic, game theory or complex analysis.

ABOUT STATISTICS.

While statistics can be considered as a part of probability theory, the approach is different. The starting point is usually a set of data. An important task is to analyze these data is to find the right mathematical model. Examples of statistical tasks are: **finding correlations** between data, to **give estimates** for the failure rate for equipment, or the **time predictions** for events. A quite recent application are **mail filters**. They use statistical analysis of mail messages to distinguish spam from legitimate mail.

PART I, DISCRETE DISTRIBUTIONS

FINITE PROBABILITY SPACES.

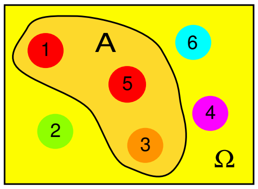
Let $\Omega = \{1,...,n\}$ be a finite set. The elements of Ω are called **experiments**. The subsets of Ω are called **events**. If we assign weights p_i to each of the experiments in Ω so that $\sum_i p_i = 1$, we have a **probability space**. The **probability** of an event A is defined as $\mathrm{P}[A] = \sum_{i \in A} p_i$, the sum of the probabilities of each experiment in A.

EQUAL PROBABILITY.
In most of the cases, an experiment has the same probability $p_i = 1/n$. This implies that the probability of an event A is the number of elements in A divided by the number of elements in Ω.

$$\boxed{\mathrm{P}[A] = \frac{|A|}{|\Omega|} = \frac{\text{Number of experiments in A}}{\text{Number of all experiments}}}$$

The task to **count** the elements in A often leads to **combinatorial problems**. Below, we will see that the probability distribution $p_i = 1/|\Omega|$ can characterized as the one which maximizes the "entropy" $-\sum_{i=1}^n p_i \log(p_i)$.

EXAMPLES.



1) With $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $p_i = 1/6$, we model a fair dice, thrown once. The event $A = \{1, 3, 5\}$ for example is the event that "the dice produces an odd number". This event has the probability 1/2.

2) What is the probability that a dice thrown twice shows a sum bigger than 10? To solve this problem, we take the set $\Omega = \{(i, j), i, j \in \{1, 2, 3, 4, 5, 6\}\}$ of all possible 36 experiments we can do. Each experiment has probability $p_{ij} = 1/36$. The event $\{(5, 6), (6, 5), (6, 6)\}$ consists of experiments which lead to a sum bigger than 10. Its probability is $3/36 = 1/12$.

3) The **Girl-Boy problem**:
"Dave has two child. One child is a boy. What is the probability that the other child is a girl"? Most people would say 1/2.

Solution: $\Omega = \{BG, GB, BB\}$, $\mathrm{P}[\{BG\}] = \mathrm{P}[\{GB\}] = \mathrm{P}[\{BB\}] = 1/3$. Now $A = \{BG, GB\}$ is the event that the other child is a girl. $\mathrm{P}[A] = 2/3$.

SET THEORETICAL NOTATIONS.

$\emptyset$ denotes the **empty set**. A^c is the **complement** of A in Ω. Example: $\Omega = \{1, 2, 3, 4, 5, 6\}, A = \{1, 3, 5\}, A^c = \{2, 4, 6\}$. $A \cap B$ is the **intersection** of A and B. $A \cup B$ is the **union** of A and B. $\mathrm{P}[A|B] = \mathrm{P}[A \cap B]/\mathrm{P}[B]$ is the probability of A **under the condition** B. It is the probability that the event A happens if we know that the event B happens. (B has to have positive probability).

Consider a general probability distribution $p = (p_1, ..., p_n)$ on $\Omega = \{1, 2, 3, 4, 5, 6\}$. For example $p = (1/7, 1/7, 1/7, 1/7, 1/7, 2/7)$, models an **unfair dice** which has been rigged in such a way that 6 appears with larger probability. Define the **entropy** of the probability distribution as $H(p) = -\sum_{i=1}^{6} p_i \log(p_i)$. For which $p = (p_1, ...p_n)$ is the entropy maximal? We have to extremize $H(p)$ under the constraint $G(p) = \sum_i p_i = 1$. This can be solved using Lagrange multipliers and leads to the solution $p_i = 1/6$.

INDEPENDENCE. Two events A, B are called **independent**, if $\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B]$.

A finite set $\{A_i\}_{i \in I}$ of events is called **independent** if for all $J \subset I$

$$\mathrm{P}[\bigcap_{i \in J} A_i] = \prod_{i \in J} \mathrm{P}[A_i] \; .$$

where $\prod i = 1^n a_i = a_1 a_2 ... a_n$ is the product of numbers.

PROPERTIES:

- $A, B \in \mathcal{A}$ are independent, if and only if either $\mathrm{P}[B] = 0$ or $\mathrm{P}[A|B] = \mathrm{P}[A]$. "Knowing B does not influence the probability of the event A".
- Every event A is independent to the empty event $B = \emptyset$.

RANDOM VARIABLE.

A real valued function X on Ω is called a **random variable**. We can look at it as a **vector** $f = (f(1), f(2), ..., f(n))$. In multivariable calculus, we mostly encountered vectors in dimensions 2 or 3. In probability theory it is natural to look at vectors with arbitrary many coordinates.

EXPECTATION. The **expectation** of a random variable f is defined as $\mathrm{E}[f] = \sum_i p_i f(i)$. It is also called the **mean** or the **average value** of f.

PROPERTIES OF THE EXPECTATION: For random variables X, Y and a real number $\lambda \in \mathbf{R}$

$$\begin{array}{ll} \mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y] & \mathrm{E}[\lambda X] = \lambda \mathrm{E}[X] \\ X \leq Y \Rightarrow \mathrm{E}[X] \leq \mathrm{E}[Y] & \mathrm{E}[X^2] = 0 \Leftrightarrow X = 0 \\ \mathrm{E}[X] = c \text{ if } X(\omega) = c \text{ is constant} & \mathrm{E}[X - \mathrm{E}[X]] = 0 \; . \end{array}$$

PROOF OF THE ABOVE PROPERTIES:

$$\begin{aligned}
&\mathrm{E}[X+Y] = \sum_i p_i(X+Y)(i) \sum_i p_i(X(i)+Y(i)) = \mathrm{E}[X] + \mathrm{E}[Y] \\
&\mathrm{E}[\lambda X] = \sum_i p_i(\lambda X)(i) = \lambda \sum_i p_i X(i) = \lambda \mathrm{E}[X] \\
&X \leq Y \Rightarrow X(i) \leq Y(i), \text{ and } \mathrm{E}[X] \leq \mathrm{E}[Y] \\
&\mathrm{E}[X^2] = 0 \Leftrightarrow X^2(i) = 0 \Leftrightarrow X = 0 \\
&X(\omega) = c \text{ is constant} \Rightarrow \mathrm{E}[X] = c \cdot \mathrm{P}[X = c] = c \cdot 1 = c \\
&\mathrm{E}[X - \mathrm{E}[X]] = \mathrm{E}[X] - \mathrm{E}[\mathrm{E}[X]] = \mathrm{E}[X] - \mathrm{E}[X] = 0
\end{aligned}$$

VARIANCE, STANDARD DEVIATION

Variance

$$\mathrm{Var}[X] = \mathrm{E}[\ (X - \mathrm{E}[X])^2\ ] = \mathrm{E}[X^2] - E[X]^2 \ .$$

Standard deviation

$$\sigma[X] = \sqrt{\mathrm{Var}[X]} \ .$$

Covariance

$$\mathrm{Cov}[X,Y] = \mathrm{E}[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \ .$$

Correlation of $\mathrm{Var}[X] \neq 0, \mathrm{Var}[Y] \neq 0$

$$\mathrm{Corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]} \ .$$

$\mathrm{Corr}[X,Y] = 0$: **uncorrelated** X and Y.

STATISTICS. A set of data points $\{x_1, x_2, x_3, x_4, x_5\} = \{2.3, 1.4, 2.5, 0.6\}$ can be interpreted as a random variable X on $\{1, 2, \dots, 5\}$, where $X(i) = x_i$. By default we assume a uniform distribution. The expectation of X is the average $\frac{1}{5}\sum_{i=1}^5 x_i = 6.8/5 = 1.7$. The variance is $\mathrm{Var}[X] = \frac{1}{5}\sum_{i=1}^5 (x_i - 1.7)^2 = 0.575$, the standard deviation $\sigma[X] = \sqrt{0.575} = 0.758....$

Given a second set of data points $\{y_1, y_2, y_3, y_4, y_5\} = \{2.3, 1.7, 2.3, 0.7\}$, we can compute the covariance $\mathrm{Cov}[X,Y] = \frac{1}{5}\sum_{i=1}^5 (x_i - 1.7)(y_i - 1.75) = 0.485$.

REMARK. In statistics, one usually takes $s^2 = \frac{n}{(n-1)}\sigma^2$ as estimate for the standard deviation of n data x_i.

PROPERTIES of VAR, COV, and CORR:

$\mathrm{Var}[X] \geq 0$.
$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$.
$\mathrm{Var}[\lambda X] = \lambda^2 \mathrm{Var}[X]$.
$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] - 2\mathrm{Cov}[X,Y]$.

$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$.
$\mathrm{Cov}[X,Y] \leq \sigma[X]\sigma[Y]$ (Schwarz inequality).
$-1 \leq \mathrm{Corr}[X,Y] \leq 1$.
$\mathrm{Corr}[X,Y] = 1$ if $X - \mathrm{E}[X] = Y - \mathrm{E}[Y]$
$\mathrm{Corr}[X,Y] = -1$ if $X - \mathrm{E}[X] = -(Y - \mathrm{E}[Y])$.

FLIPPING COINS

If we flip coins n times, and $\{\mathrm{head}, \mathrm{tail}\}$ is encoded as $\{0,1\}$, we have the probability space $\Omega = \{0,1\}^n$ which contains 2^n experiments. For example, if $n = 3$, then $\Omega = \{(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\}$. If $Q[\{1\}] = p, Q[\{0\}] = q = 1 - p$ is the probability distribution on $\{0,1\}$, then then $P[\{(\omega_1, \omega_2, \dots, \omega_n)\}] = Q[\{\omega_1\}] \cdots Q[\{\omega_n\}]$ is the probability distribution on $\{0,1\}^6$.

EXAMPLE. If $p = 1/3$ is the probability that "head" happens in one flips a coin, then the probability that $(1,1,0)$ happens is $pp(1-p) = p^2(1-p) = 4/27$.

INDEPENDENT RANDOM VARIABLES:

X, Y are **independent** if for all $a, b \in \mathbf{R}$

$$\mathrm{P}[X = a; Y = b] = \mathrm{P}[X = a] \cdot \mathrm{P}[Y = b] \ .$$

A finite collection $\{X_i\}_{i \in I}$ of random variables are **independent**, if for all $J \subset I$ and $a_i \in \mathbf{R}$

$$\mathrm{P}[X_i = a_i, i \in J] = \prod_{i \in J} P[X_i = a_i] \ .$$

PROPERTIES:

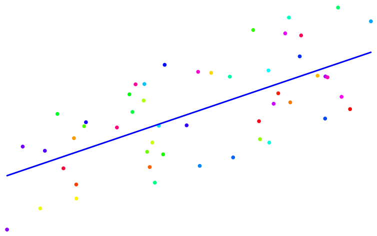
- If X and Y are independent, then $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$.
- If X_i is a set of independent random variables, then $\mathrm{E}[\prod_{i=1}^n X_i] = \prod_{i=1}^n \mathrm{E}[X_i]$.
- If X, Y are independent, then $\mathrm{Cov}[X, Y] = 0$.
- A constant random variable is independent to any other random variable.

EXAMPLE. The random variable $X[\{x, y\}] = x$ (value of first dice) and $Y[\{x, y\}] = y$ (value of second dice) on $\Omega = \{1, 2, 3, 4, 5, 6\}$ are independent: $\mathrm{P}[X = a] = 1/6$, $\mathrm{P}[Y = b] = 1/6$, $\mathrm{P}[(X, Y) = (a, b)] = 1/36$.

EXAMPLE. The random variables $X[\{x, y\}] = x + y$ (sum of two dices) and $Y[\{x, y\}] = |x - y|$ (difference of two dices) are not independent: $\mathrm{E}[X \cdot Y] = \mathrm{E}[U^2 - V^2] = \mathrm{E}[U^2] - \mathrm{E}[V^2] = 0$, where $U[\{x, y\}] = x^2, V[\{x, y\}] = y^2$. But $\mathrm{E}[X] > 0$ and $\mathrm{E}[Y] > 0$ shows that $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$ is not valid.

REGRESSION LINE: The **regression line** of two random variables X, Y is defined as $y = ax + b$, where

$$\boxed{a = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]}, \quad b = \mathrm{E}[Y] - a\mathrm{E}[X]}$$



PROPERTY: Given $X, \mathrm{Cov}[X, Y], \mathrm{E}[Y]$, and the regression line $y = ax + b$ of X, Y. The random variable $\tilde{Y} = aX + b$ minimizes $\mathrm{Var}[Y - \tilde{Y}]$ under the constraint $\mathrm{E}[Y] = \mathrm{E}[\tilde{Y}]$ and is the best guess for Y, when knowing only $\mathrm{E}[Y]$ and $\mathrm{Cov}[X, Y]$. We check $\mathrm{Cov}[X, Y] = \mathrm{Cov}[X, \tilde{Y}]$.

Examples: There are two extreme cases:
1) If X, Y are independent, then $a = 0$. It follows that $b = \mathrm{E}[Y]$. We can not guess Y better than replacing it by its mean.

2) If $X = Y$, then $a = 1$ and $b = 0$. The best guess for Y is X.

PROOF. To minimize $\mathrm{Var}[aX + b - Y]$ under the constraint $\mathrm{E}[aX + b - Y] = 0$ is equivalent to find (a, b) which minimizes $f(a, b) = \mathrm{E}[(aX + b - Y)^2]$ under the constraint $g(a, b) = \mathrm{E}[aX + b - Y] = 0$. This **least square** solution can be obtained with Lagrange or by solving $b = E[Y] - aE[X]$ and minimizing $h(a) = \mathrm{E}[(aX - Y - \mathrm{E}[aX - Y])^2] = a^2(\mathrm{E}[X^2] - E[X]^2) - 2a(E[XY] - E[X]E[Y]) = a^2\mathrm{Var}[X] - 2a\mathrm{Cov}[X, Y]$. Setting $h'(a) = 0$ gives $a = \mathrm{Cov}[X, Y]/\mathrm{Var}[X]$.

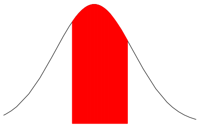
PART II, CONTINUOUS DISTRIBUTIONS

SETUP.

Let Ω be a part $\mathbf{R}^d$, where d can be any dimension. We mostly consider $d = 1, 2, 3$. An **event** is a region in Ω. A **random variable** X is a real-valued function on Ω. A **random vector** is a vector-valued function on Ω. The **probability distribution** on Ω is a piecewise smooth nonnegative function f on Ω with the property that $\int_\Omega f(x) \, dx = 1$. The probability of an event A is the integral $\mathrm{P}[A] = \int_A f(x) \, dx$. The expectation of a random variable X is $\mathrm{E}[X] = \int_\Omega X(x) f(x) \, dx$, the variance is $\mathrm{E}[(X - m)^2] = \int_\Omega (X(x) - m))^2 f(x) \, dx$.

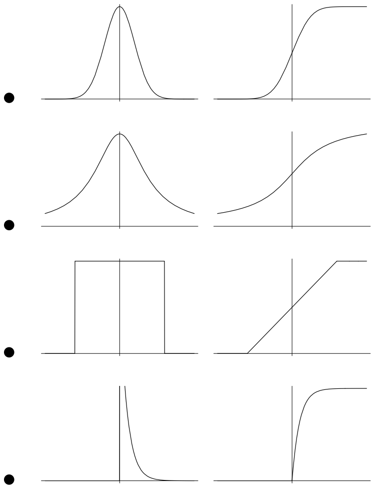
EXAMPLE. GAUSSIAN DISTRIBUTION.

Let Ω be the real line, $X(x) = x$. The probability density function $f(x) = e^{-x^2}/\sqrt{\pi}$ is called the **Gaussian distribution** or **normal distribution**. Its graph is bell shaped curve. We can look at the event $A = \{X \in [a,b]\}$ that the measurement X takes values in the interval $[a,b]$. The probability of this event is $\mathrm{P}[A] = \int_a^b f(x)\, dx$.



A more general Gaussian distribution is obtained by translation and scaling: It is custom to write it as $f(x) = e^{-(x-m)^2/\sigma^2}/\sqrt{\pi\sigma^2}$, where m is the **expectation** and σ is the standard deviation. The constants names reflect the fact that the random variable $X(x) = x$ has the mean m and standard deviation σ with this distribution.

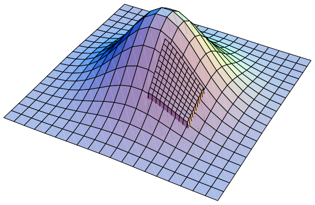
- The **normal distribution** $N(m, \sigma^2)$ is given by the density function
$$f(x) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-m)^2}{2\sigma^2}} .$$
- The **Cauchy distribution** is given by the density function
$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2} .$$
- The **uniform distribution** has the density function
$$f(x) = \frac{1}{(b-a)} 1_{[a,b]} .$$
- The **exponential distribution** $\lambda > 0$ has the density function
$$f(x) = 1_{[0,\infty)}(x) \lambda e^{-\lambda x} .$$



EXAMPLE. SHOOTING DARTS.

Let Ω be the plane and let $f(x,y) = \exp(-|x|^2/\sigma^2)/(\pi\sigma^2)$. This function $f(x,y)$ is the product of two one-dimensional Gaussian distributions $f(x), f(y)$. The function f gives the probability density to hit a certain spot (x,y) in the plane, if we shoot with a dart and aim at the point $(0,0)$.

The probability of an event A is the double integral $\int \int_A f(x,y)\, dxdy$. Let us assume for example, that we gain 20 points in a dart game when hitting into the disc $x^2 + y^2 \leq 1/5$. The standard deviation σ of the probability space is a measure for the accuracy of the dart player. Let us assume that $\sigma = 1$. What is the probability to hit 20 points?



Answer:
$2\pi \int_0^{1/2} (e^{-r^2}/\pi) r \, dr = -e^{-r^2}/2|_0^{1/2} = 1 - e^{-1/4}$.

CONDITIONAL PROBABILITY. Let $f(x,y)$ be the probability density on the plane $\mathbf{R}^2$. Consider a random vector (X,Y) with this density. Assume we know that $Y(x,y)$ is in the interval $[a,b]$. What is the expectation of X under this condition?

The answer is called the **conditional probability** $\mathrm{E}[X|Y \in [a,b]] = \int_{-\infty}^{\infty} \int_a^b x f(x,y)\, dydx / \mathrm{P}[Y \in [a,b]\ ]$. A special case is $\mathrm{E}[X|Y = a] = \int_{-\infty}^{\infty} x f(x,a)\, dx / \int_{-\infty}^{\infty} f(x,a)\, dx$.

More generally, we could have the situation that the random variables X and Y are linked: what is the expectation of X under the condition that $X^2 + Y^2 = 1$? The answer is again given by a **conditional probability**. It is now expressed as a **line integral**: $\mathrm{E}[X|X^2+Y^2 = 1] = \int_0^{2\pi} x f(\cos(\theta), \sin(\theta))\, d\theta / \int_0^{2\pi} f(\cos(\theta), \sin(\theta))\, d\theta$. We see that both iterated integrals as well as line integrals can occur in probability theory.

RANDOM VARIABLES WITH GIVEN PROBABILITY DISTRIBUTION. Computers usually produce random variables X which are random in $[0,1]$. How would we produce new random variables Y with a given distribution?

If ϕ is an invertible function and $Y = \phi(X), X = \psi(Y)$, $y = \phi(x)$. Then $F_X(y) = \mathrm{P}[Y \leq y] = \mathrm{P}[\phi(X) \leq y] = \mathrm{P}[X \leq \psi(y)] = F_X[\psi(y)]$. By the chain rule $f_Y(y) = F'_Y(y) = F'_X(\psi(y))\psi'(y) = f_X(\psi(y))\psi'(y)$.

A special case is if X is the uniform distribution and $Y = \phi(X)$. Then $f_X(x)$ is constant 1 on the unit interval and $\boxed{f_Y(y) = \psi'(y)}$.

EXAMPLE: the map $\phi(x) = y = \tan(x\pi)$ maps $[0,1]$ into the real line. The inverse is $\psi(y) = \tan^{-1}(y)/\pi$ and $\psi'(y) = \frac{1}{\pi}\frac{1}{1+y^2}$, which is the density of the Cauchy distribution. So, to generate Cauchy distributed random variables, just take uniform distributed random variables X and put $Y = \tan(\pi X)$.

In Mathematica Tan[Pi Random[]] produces Cauchy distributed random variables.