Math 19b: Linear Algebra with Probability **Oliver Knill, Spring 2011**

Lecture 22: Distributions

A random variable is a function from a probability space Ω to the real line R. There are two important classes of random variables:

1) For **discrete random variables**, the random variable X takes a discrete set of values. This means that the random variable takes values x_k and the probabilities $\mathrm{P}[X = x_k] = p_k$ add up to 1.

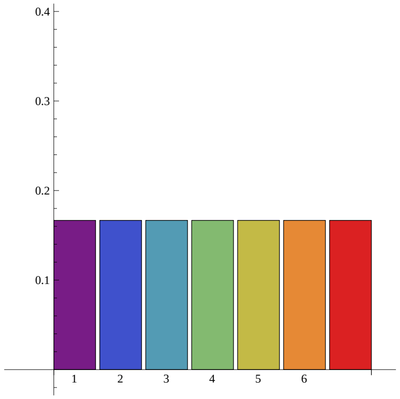
2) For **continuous random variables**, there is a probability density function $f(x)$ such that $\mathrm{P}[X \in [a,b]] = \int_a^b f(x)\, dx$ and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

Discrete distributions

1 We throw a dice and assume that each side appears with the same probability. The random variable X which gives the number of eyes satisfies

$$\mathrm{P}[X = k\ ] = \frac{1}{6}\ .$$

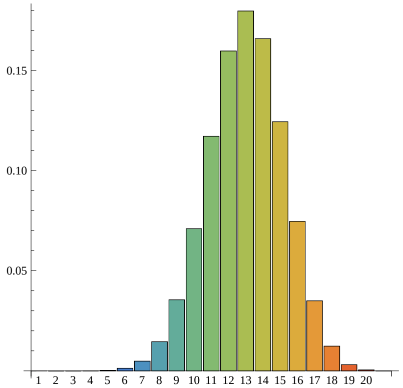
This is a discrete distribution called the **uniform distribution** on the set $\{1,2,3,4,5,6\ \}$.



2 Throw n coins for which head appears with probability p. Let X denote the number of heads. This random variable takes values $0,1,2,\dots,n$ and

$$\mathrm{P}[X = k\ ] = \binom{n}{k} p^k (1-p)^{n-k}\ .$$

This is the Binomial distribution we know.



3 The probability distribution on $N = \{0,1,2,...\ \}$

$$\mathrm{P}[X = k\ ] = \frac{\lambda^k e^{-\lambda}}{k!}$$

is called the **Poisson distribution**. It is used to describe the number of radioactive decays in a sample, the number of newborns with a certain defect. You show in the homework that the mean and standard deviation is λ Poisson distribution is the most important distribution on $\{0,1,2,3,\dots\}$. It is a limiting case of Binomial distributions

4 An epidemiology example from Cliffs notes: the UHS sees X=10 pneumonia cases each winter. Assuming independence and unchanged conditions, what is the probability of there being 20 cases of pneumonia this winter? We use the Poisson distribution with $\lambda = 10$ to see $\mathrm{P}[X = 20] = 10^{20} e^{-10}/20! = 0.0018$.

The Poisson distribution is the $n \to \infty$ limit of the binomial distribution if we chose for each n the probability p such that $\lambda = np$ is fixed.

Proof. Setting $p = \lambda/n$ gives

$$\begin{aligned}
\mathrm{P}[X = k] &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} \\
&= (\frac{n!}{n^k (n-k)!})(\frac{\lambda^k}{k!})(1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^k\ .
\end{aligned}$$

This is a product of four factors. The first factor $n(n-1)...(n-k+1)/n^k$ converges to 1. Leave the second factor a$\lambda^k/k!$ as it is. The third factor converges to $e^{-\lambda}$ by the definition of the exponential. The last factor $(1 - \frac{\lambda}{n})^k$ converges to 1 for $n \to \infty$ since k is kept fixed. We see that $\mathrm{P}[X = k] \to \lambda^k e^{-\lambda}/k!$.

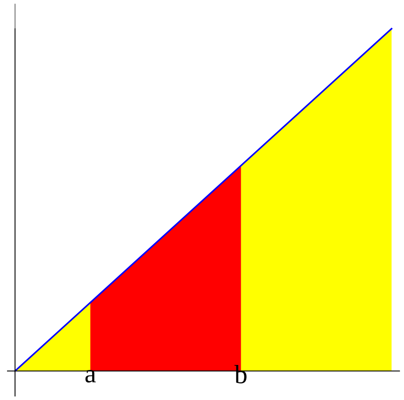
Continuous distributions

5 A random variable is called a **uniform distribution** if $\mathrm{P}[X \in [a,b]] = (b-a)$ for all $0 \leq a < b \leq b$. We can realize this random variable on the probability space $\Omega = [a,b]$ with the function $X(x) = x$, where $\mathrm{P}[I]$ is the length of an interval I. The uniform distribution is the most natural distribution on a finite interval.

6 The random variable X on $[0,1]$ where $\mathrm{P}[[a,b]] = b-a$ is given by $X(x) = \sqrt{x}$. We have

$$\mathrm{P}[X \in [a,b]] = \mathrm{P}[\sqrt{x} \in [a,b]] = \mathrm{P}[x \in [a^2,b^2]] = b^2 - a^2 \; .$$

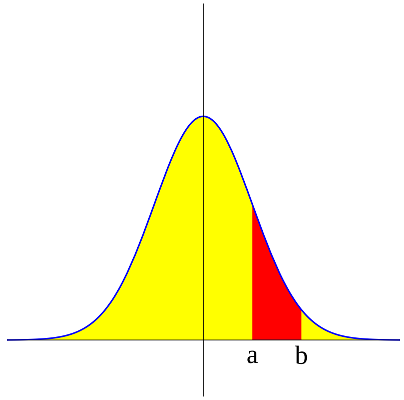
We have $f(x) = 2x$ because $\int_a^b f(x)\,dx = x^2|_a^b = b^2 - a^2$. The function $f(x)$ is the probability density function of the random variable.



7 A random variable with normal distribution with mean 1 and standard deviation 1 has the probability density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \; .$$

This is an example of a continuous distribution. The normal distribution is the most natural distribution on the real line.



8 The probability density function

$$f(x) = \lambda e^{-\lambda x} \; .$$

is called the **exponential distribution**. The exponential distribution is the most natural distribution on the positive real line.

Remark. The statements "most natural" can be made more precise. Given a subset $X \subset R$ of the real line and the mean and standard deviation we can look at the distribution f on X for which the **entropy** $-\int_X f \log(f)\,dx$ is maximal. The uniform, exponential and normal distributions extremize entropy on an interval, half line or real line. The reason why they appear so often is that adding independent random variables increases entropy. If different processes influence an experiment then the entropy becomes large. Nature tries to maximize entropy. Thats why these distributions are "natural".

9 The distribution on the positive real axis with the density function

$$f(x) = \frac{1}{\sqrt{2\pi x^2}} e^{-\frac{(\log(x)-m)^2}{2}}$$

is called the **log normal distribution** with mean m. Examples of quantities which have log normal distribution is the size of a living tissue like like length or height of a population or the size of cities. An other example is the **blood pressure** of adult humans. A quantity which has a log normal distribution is a quantity which has a logarithm which is normally distributed.

Homework due March 30, 2011

1 a) Find the mean of the exponential distribution.
b) Find the variance and standard deviation of the exponential distribution.
c) Find the **entropy** $-\int_0^\infty f(x)\log(f(x))\,dx$ in the case $\lambda = 1$.

2 Here is a special case of the **Students** t **distribution**

$$f(x) = \frac{2}{\pi}(1+x^2)^{-2} \; .$$

a) Verify that it is a probability distribution.
b) Find the mean. (No computation needed, just look at the symmetry).
c) Find the standard deviation.
To compute the integrals in a),c), you can of course use a computer algebra system if needed.

3 a) Verify that the Poisson distribution is a probability distribution: $\sum_{k=0}^\infty \mathrm{P}[X=k] = 1$.
b) Find the mean $m = \sum_{k=0}^\infty k\mathrm{P}[X=k]$.
c) Find the standard deviation $\sum_{k=0}^\infty (k-m)^2 \mathrm{P}[X=k]$.