Statistics

Sampled from

Morris H. DeGroot & Mark J. Schervish, “Probability and Statistics”, 3rd Edition, Addison Wesley

Probability: Continuous distribution and variables

- Continuous distributions
  - Random variables
  - Probability density function
  - Uniform, normal and exponential distributions
  - Expectations and variance
  - Law of large numbers
  - Central limit theorem
  - Probability density functions of more than one variable
  - Rejection and transformation methods for sampling distributions (section)

Statistics

- Estimators: mean, standard deviation
- Maximum likelihood
- Confidence intervals
- χ^2 statistics
- Regression
- Goodness of fit

Continuous random variables

- A random variable *X* is a real value function defined on a sample space *S.*
- *X* is a continuous random variable if a non-negative function *f,* defined on the real line, exists such that an integral over the domain *A* is the probability that *X* takes a value in domain *A.* (*A* is, for example, the interval [a,b])

$$\Pr(a < X < b) = \int_a^b f(x)\,dx$$

Probability density function

- *f* is called probability density function (p.d.f.). Note that the unit of the pdf below are of 1/length, only after the multiplication with a length element we get probability
- For every p.d.f. we have

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Examples of p.d.fs

- A car is driving in a circle at a constant speed. What is the probability that it will be found in the interval between 1 and 2 radians?
- A computer is generating with equal probability density, random numbers between 0 and 1. What is the probability of obtaining 0.75?
- Protein folds at a constant rate (the probability that a protein will fold at the time interval $[t, t+dt]$ is a constant αdt). If we have at time zero N_0 protein molecules, what is the probability that all protein molecules will fold after time t' ?

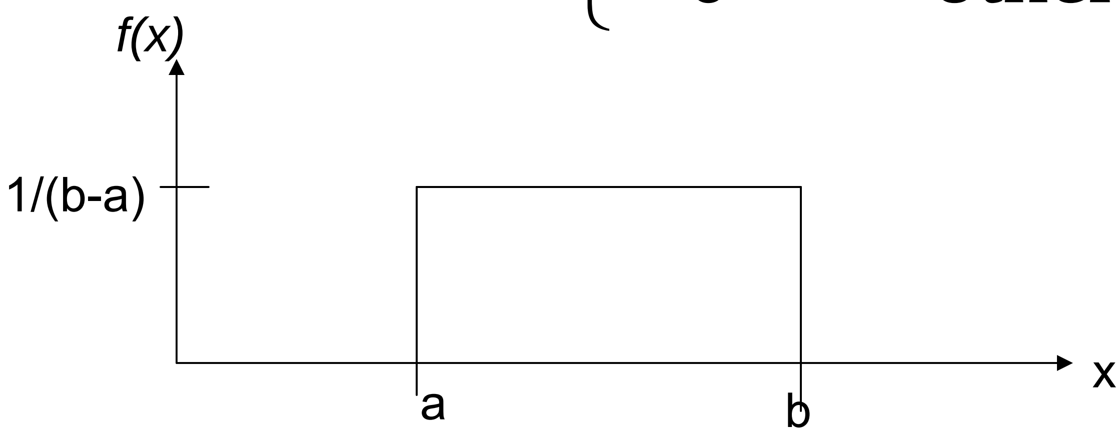
Uniform distribution on an interval

- Consider an experiment in which a point *X* is selected from an interval $S = \{x : a \le x \le b\}$ in such a way that the probability of finding *X* at a given interval is proportional to the interval length (hence the p.d.f. is a constant). This distribution is called the <u>uniform distribution</u> We must have for this distribution

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{a}^{b} f(x)\,dx = 1$$

Uniform distribution (continue)

$$f(x)=\begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$



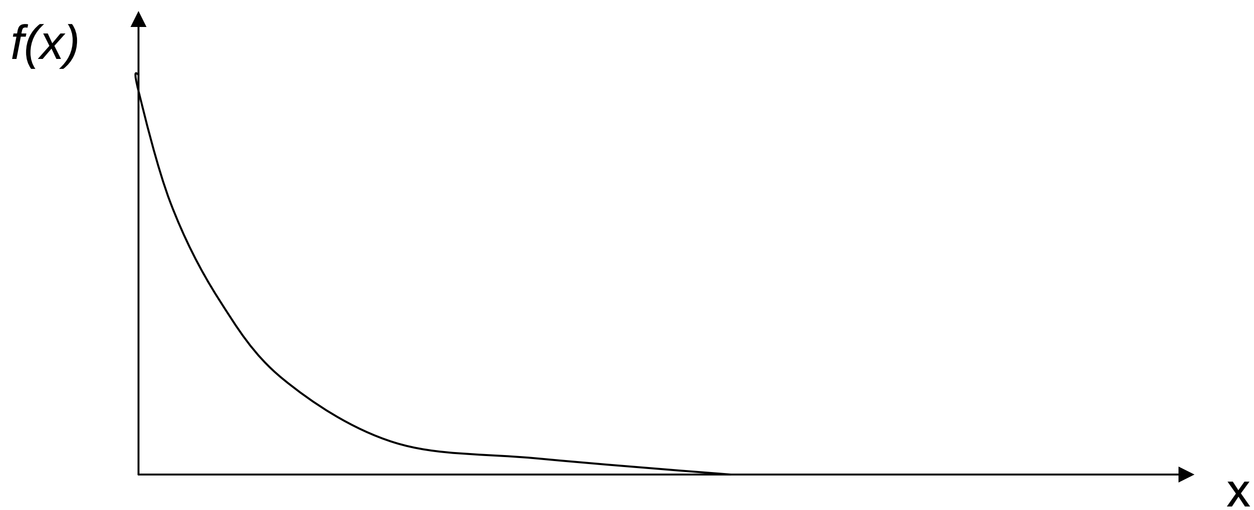
Examples of simple distributions

- X is a random variable distributed uniformly on a circle of radius a. Find $f(x)$
- Check that the following function satisfies the conditions to be a p.d.f.

$$f(x) = \begin{cases} \dfrac{2}{3} x^{-1/3} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

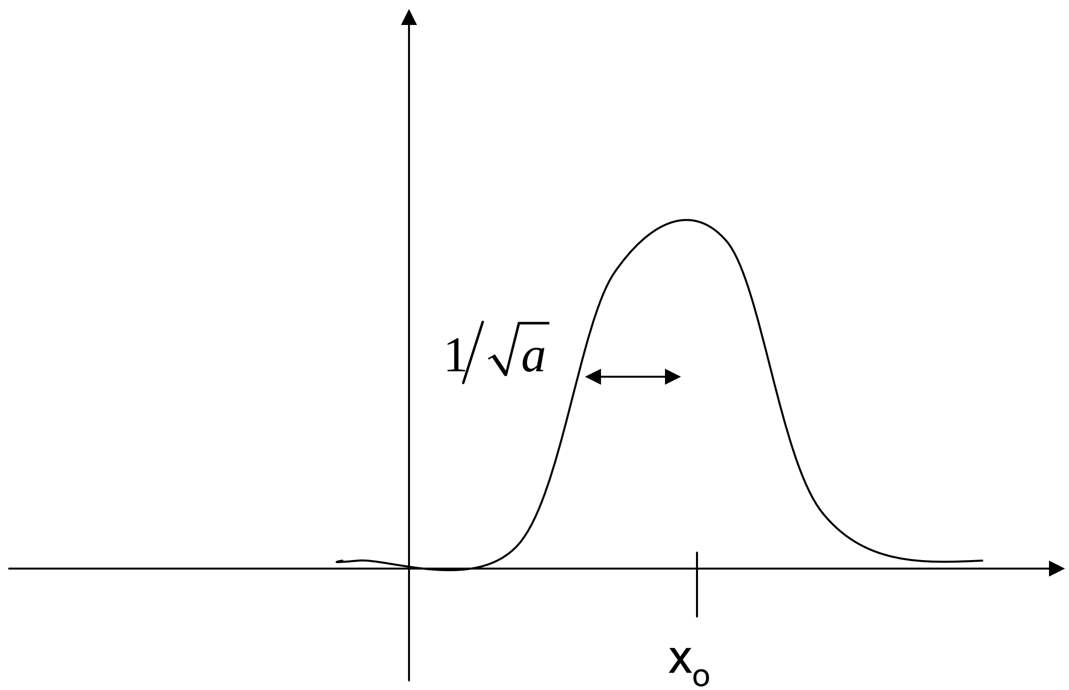
Exponential distribution

$$f(x) = \left[a\exp(-ax) \quad 0 < x < \infty \right]$$



Normal distribution

$$f(x) = \left[\left(\frac{a}{\pi} \right)^{1/2} \exp\left(-a\left(x - x_0 \right)^2 \right) \quad -\infty < x < \infty \right]$$



Continuous distribution functions

defined as $F\left(x\right) = \Pr(X \leq x) \qquad \text{for } -\infty < x < \infty$

F(x) is a monotonic non decreasing function of *x (can you show it?),* that can be written in terms of its corresponding p.d.f.

$$F\left(x\right) = \Pr\left(X \leq x\right) = \int_{-\infty}^{x} f\left(x\right) dx$$

or

$$\frac{dF}{dX} = f\left(x\right)$$

Distribution function: Example

$$f(x) = \begin{bmatrix} a\exp(-ax) & \text{for } 0 < x < \infty \\ 0 & \text{otherwise} \end{bmatrix}$$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{0}^{x} a\exp(-ax)\,dx = \left[-\exp(-ax)\right]_0^x = 1 - \exp(-ax)$$

Expectation

- For a random variable *X* with a p.d.f. *f(x)* the expectation *E(X)* is defined

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

The expectation exists if and only if the integral is absolutely converged, i.e.

$$\int_{-\infty}^{\infty} |x|\, f(x)\, dx < \infty$$

Expectation (example)

$$f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x \cdot 2x \cdot dx = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

Even if the p.d.f, satisfies the requirements, it is not obvious that the expectation exists (next slide)

The Cauchy p.d.f.

$$f(x) = \left[\frac{1}{\pi\left(1+x^2\right)} \quad -\infty < x < \infty \right] \qquad \left(f(x) \geq 0 \right)$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\pi\left(1+x^2\right)} dx = \frac{1}{\pi} \arctan(x) \Big|_{-\infty}^{x} = \frac{1}{\pi}\left(\arctan(x) - \left(-\frac{\pi}{2} \right) \right)$$

$$F(\infty) = \frac{1}{\pi}\left(\frac{\pi}{2} + \frac{\pi}{2} \right) = 1 \qquad \left(\int_{-\infty}^{\infty} \frac{1}{\pi\left(1+x^2\right)} dx = 1 \right)$$

Cauchy distribution: Expectation

Test for existence of expectation

$$E\left(X\right)=\int_{-\infty}^{\infty}|x|\cdot f\left(x\right)\cdot dx=\int_{-\infty}^{\infty}|x|\frac{1}{\pi\left(1+x^2\right)}dx\to\infty$$

Expectation does not exist for the Cauchy distribution.

Some properties of expectations

- Expectation is linear

$$E(aX + bY) = aE(X) + bE(Y)$$

- If the random variables *X* and *Y* are independent $\left(f(x,y) = f(x)\,f(y)\right)$ then

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

Expectation of a function

- Is essentially the same as the expectation of a variable

$$E\left(r\left(x\right)\right)=\int_{-\infty}^{\infty} r\cdot g\left(r\right)dr=\int_{-\infty}^{\infty} r\left(x\right)\cdot f\left(x\right)\cdot dx$$

of special interest is the expectation value of moments

$$\text{variance}\equiv E\left(X^{2}\right)-\left[E\left(X\right)\right]^{2}=\int_{-\infty}^{\infty} x^{2}\cdot f\left(x\right)\cdot dx-\left[\int_{-\infty}^{\infty} x\cdot f\left(x\right)\cdot dx\right]^{2}$$

Can you show that the variance is always non-negative?

Functions of several random variables

- We consider a p.d.f.

$$f\left(x_1,...,x_n\right)$$

of several random variables

$$X_1,...,X_n$$

The p.d.f. satisfies (of course)

$$f\left(x_1,...,x_n\right)\geq 0$$

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2}...\int_{a_n}^{b_n} f\left(x_1,...,x_n\right)dx_1...dx_n = 1$$

Expectation of function of several variables

- Similarly to one variable case, expectations of functions with several variables are computes

$$E\left(Y = r\left(x_1, \ldots, x_n\right)\right) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} r\left(x_1, \ldots, x_n\right) \cdot f\left(x_1, \ldots, x_n\right) dx_1 \ldots dx_n$$

Example: expectation of more than one variable

$$f\left(x,y\right)=\begin{cases}1 & \text{for } \left(x,y\right)\in S \\ 0 & \text{otherwise}\end{cases}$$

S is a square: $0<x<1$ $0<y<1$

$$E\left(X^2+Y^2\right)=\int_0^1\int_0^1\left(x^2+y^2\right)f\left(x,y\right)\cdot dx\cdot dy$$

$$=\int_0^1\int_0^1\left(x^2+y^2\right)dx\cdot dy=\frac{2}{3}$$

Markov Inequality

- *X* is a random variable such that

$$\Pr\left(X \geq 0\right) = 1$$

- For every *t>0*

$$\Pr\left(X \geq t\right) \leq \frac{E\left(X\right)}{t}$$

- Prove it
- Why *E(X)>t* is not interesting?

Chebyshev Inequality is a special case of the Markov inequality

- X is a random variable for which the variance exists. For $t>0$

$$\Pr\left(\left[X - E(X)\right]^2 \geq t^2\right) \leq \frac{\mathrm{var}(X)}{t^2}$$

- Substitute

$$Y = \left[X - E(X)\right]^2 \quad \rightarrow E(Y) = \mathrm{var}(X) \quad \text{and } t^2 \text{ by } t$$

to obtain the Markov inequality

The law of large numbers I

- Consider a set of *N* random variables $X_1,...X_n$ i.i.d. Each of the random variables has mean (expectation value) μ and variance σ^2
- The arithmetic average of *n* samples is defined $\bar{X}_n = \frac{1}{n}\left(X_1 + ... + X_n\right)$. It defines a new random variable that we call the sample mean
- The expectation value of the sample mean

$$E\left(\bar{X}_n\right) = \frac{1}{n}\sum_i E\left(X_i\right) = \frac{1}{n}\cdot n\mu = \mu$$

The Law of Large Numbers II

- The variance of $\bar{X}_n$

$$\operatorname{var}\left(\bar{X}_n\right) = E\left(\bar{X}_n^2 - E^2\left(\bar{X}_n\right)\right) = \frac{1}{n^2}\sum_{i,j} E\left(X_i X_j\right) - \frac{1}{n^2}\left[\sum_i E\left(X_i\right)\right]^2$$

Since X_i and X_j are independent for $i \neq j$ $E\left(X_i X_j\right) = E\left(X_i\right)E\left(X_j\right)$

$$\operatorname{var}\left(\bar{X}_n\right) = \frac{1}{n^2}\left[n \cdot E\left(X^2\right) + \left(n^2 - n\right) \cdot E^2\left(X\right) - n^2 E^2\left(X\right)\right]$$

$$\operatorname{var}\left(\bar{X}_n\right) = \left(E\left(X^2\right) - E^2\left(X\right)\right)\Big/ n = \operatorname{var}\left(X\right)/n$$

Which means that the variance is decreasing linearly with the number of sampled points

Law of Large numbers III

Chebyshev Inequality:

$$1-\Pr\left(\left(\bar{X}_n-\mu\right)^2 \geq \varepsilon^2\right)=\Pr\left(\left(\bar{X}_n-\mu\right)^2 < \varepsilon^2\right)\geq 1-\frac{\operatorname{var}\left(X\right)}{n\varepsilon^2} \quad \text{for } \varepsilon \geq 0$$

$$\Rightarrow \bar{X}_n \to \mu$$

Central Limit Theorem

- Statement without proof:
- Given a set of random variables $X_1, \ldots, X_n$ with mean μ_i and variance σ^2_i we define a new random variable

$$Y_n = \frac{\sum_{i=1,\ldots,n} X_i}{\left(\sum_{i=1,\ldots,n} \sigma_i^2 \right)^{1/2}}$$

- For very large *n*, the distribution of $\sum_{i=1,\ldots,n} X_i$ is normal with mean $\sum_{i=1,..,n} \mu_i$ and variance $\sum_{i=1,\ldots,n} \sigma_i^2$