Lecturenotes Statistics I – Contents

Uniform and General Distributions

Uniform distribution (probability density):

$$u(x) = \begin{cases} 1 & \text{for} \quad 0 \leq x < 1; \\ 0 & \text{elsewhere.} \end{cases}$$

The corresponding (cumulative) distribution function is

$$U(x) = \int_{-\infty}^{x} u(x')\, dx' = \begin{cases} 0 & \text{for} \quad x < 0; \\ x & \text{for} \quad 0 \leq x \leq 1; \\ 1 & \text{for} \quad x > 1. \end{cases}$$

It allows for the construction of general probability distributions. Let

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

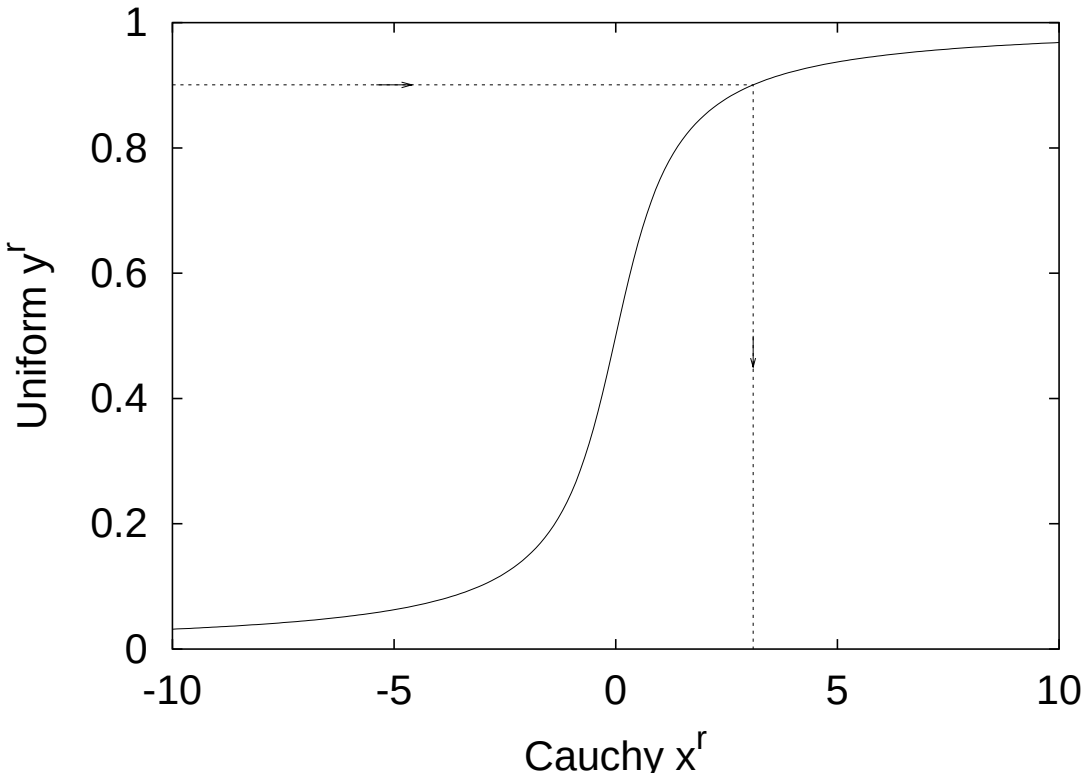
For y^r being a uniformly distributed random variable in $[0,1)$: $x^r = F^{-1}(y^r)$ is then distributed according to the probability density $f(x)$.

Example: Mapping of the uniform to the Cauchy distribution.

$$f_c(x) = \frac{\alpha}{\pi\,(\alpha^2 + x^2)} \quad \text{and} \quad F_c(x) = \int_{-\infty}^{x} f_c(x')\,dx' = \frac{1}{2} + \frac{1}{\pi}\,\tan^{-1}\left(\frac{x}{\alpha}\right),\ \alpha > 0\,.$$

The Cauchy distributed random variable x^r is generated from the uniform $y^r \in [0,1)$ through

$$\frac{x^r}{\alpha} = \tan\left(\pi y^r - \frac{\pi}{2}\right) \quad \Leftrightarrow \quad x^r = \alpha\,\tan(2\pi y^r)\,.$$



In STMC/Forlib: `rmacau.f`.

Gaussian Distribution

The Gaussian or **normal distribution** is of major importance. Its probability density is

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/(2\sigma^2)}$$

where σ^2 is the **variance** and $\sigma > 0$ the **standard deviation**. The Gaussian distribution function $G(x)$ is related to that of variance $\sigma^2 = 1$ by

$$G(x) = \int_{-\infty}^{x} g(x')\, dx' = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x/\sigma} e^{-(x'')^2/2}\, dx'' = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x}{\sigma\sqrt{2}}\right) \ .$$

In principle we could now generate **Gaussian random numbers**. However, the numerical calculation of the inverse error function is slow and makes this an impractical procedure. Much faster is to express the product probability density of

two independent Gaussian distributions in polar coordinates

$$\frac{1}{2\pi\,\sigma^2}\,e^{-x^2/(2\sigma^2)}\,e^{-y^2/(2\sigma^2)}\,dx\,dy = \frac{1}{2\pi\,\sigma^2}\,e^{-r^2/(2\sigma^2)}\,d\phi\,r dr\,,$$

and to use the relations $x^r = r^r\,\cos\phi^r$, $y^r = r^r\,\sin\phi^r$.

Assignments a0104_01 and a0104_02:

Probability densities and cumulative distribution functions for the uniform in $[-1,+1)$, a Gaussian and a Cauchy distribution.

Confidence Intervals and Sorting

One defines **q-tiles** (also **quantiles** or **fractiles**) x_q of a distribution function by

$$F(x_q) \;=\; q \;.$$

An example is the **median** $x_{\frac{1}{2}}$. The probability content of the **confidence interval**

$$[x_q, x_{1-q}] \;\text{ is }\; p = 1 - 2q \;.$$

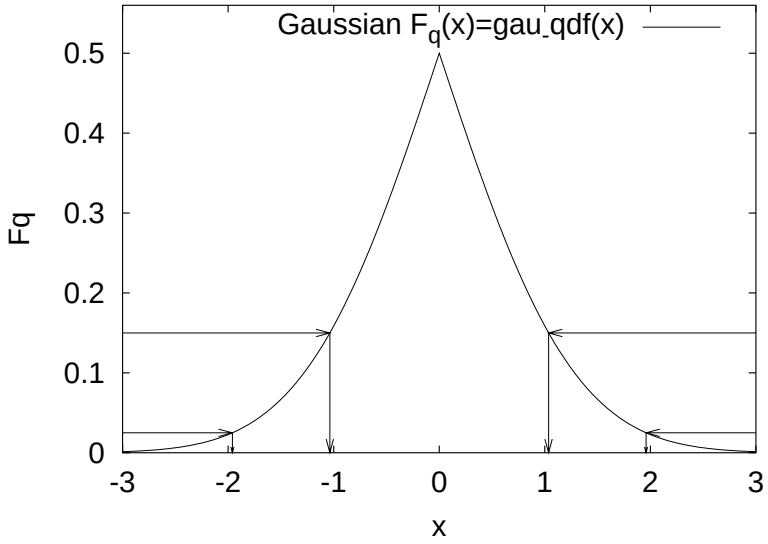
Example: Gaussian or normal distribution of variance σ^2 :

$$[-n\sigma, +n\sigma] \;\Rightarrow\; p = 0.6827 \text{ for } n = 1, \quad p = 0.9545 \text{ for } n = 2 \;.$$

The **peaked distribution function**

$$F_q(x) = \begin{cases} F(x) \text{ for } F(x) \leq \frac{1}{2}, \\ 1 - F(x) \text{ for } F(x) > \frac{1}{2}. \end{cases}$$

provides a graphical visualization of probability intervals of such a distribution:



Sorting allows for an empirical estimate. Assume we generate n random number $x_1, ..., x_n$. We may re-arrange the x_i in increasing order:

$$x_{\pi_1} \leq x_{\pi_2} \leq \ldots \leq x_{\pi_n}$$

where $\pi_1, \ldots, \pi_n$ is a permutation of $1, \ldots, n$.

In Statistics and Mathematics the random variables $x_{\pi_1}, x_{\pi_2}, \ldots\ x_{\pi_n}$ are called **order statistics**.

An estimator for the distribution function $F(x)$ is then the **empirical distribution function**

$$\overline{F}(x) = \frac{i}{n} \quad \text{for} \quad x_{\pi_i} \leq x < x_{\pi_{i+1}},\ i = 0, 1, \ldots, n-1, n$$

with the definitions $x_{\pi_0} = -\infty$ and $x_{\pi_{n+1}} = +\infty$. To calculate $\overline{F}(x)$ one needs an efficient way to **sort** n data values in ascending (or descending) order. In the STMC package this is provided by a **heapsort** routine, which arrives at the results in $O(n \log_2 n)$ steps.

Example: Gaussian distribution in assignment `a0106_02a` (200 and 20,000 data).

Download `a0106_02a` from the classwork website. Compile and run the programs. Plot with `gau_qdf.plt` (the other gnuplot drivers are for special purposes).

Statistical Bootstrap: Example Opinion polls.

We are used to statements that according to a poll an opinion has changed by a certain percentage. The statistical significance is often not addressed. One can figure that out based on multinomial distributions, but the analytical approach tends to become cumbersome. Modern PCs allow to bootstrap the confidence levels with methods we just introduced.

To be specific, we address the situation that a poll of `NDAT` people has been taken, where each person has to vote for one of `NCH` distinct candidates, or stay undecided. The outcome are estimates of the true probabilities for these choices. We estimate the confidence limits of these probabilities by using them as input to simulate the polling process `NPOLL` times, where `NPOLL` has to be a sufficiently large number. Thus we obtain an empirical peaked distribution functions for each outcome probability.

The program `polls.f` (download from the classwork page) performs the task of simulating opinion polls. It is preset to results from an opinion poll in the early stage of the 1992 US presidential election. At that time there were three candidates: George Bush (senior), Bill Clinton and Ross Perot. A poll of `NDAT = 700` people had given 42% for Bush, 38% for Clinton, 16% for Perot, and the rest undecided. The question is: To what extent is the advantage for Bush versus Clinton is statistically significant? Using the probabilities of the poll as input, we simulate the polling `NPOLL = 10,000` times.

Output are peaked distribution functions, which overlap for Bush and Clinton at a probability of 0.14. That is not yet the result for the likelihood that Clinton wins, because the distributions are correlated. The estimate for the probabilities of the election outcome is obtained by simply counting the wins of Clinton versus those of Bush and dividing the draws (possible due to the small number of peoples asked) equally between the two. In this way we get a probability of 11.8% that Clinton wins, while Bush wins with 88.2% probability.

With a 5% cut-off this is still in the dead heat.