Parametric (theoretical) probability distributions. (Wilks, Ch. 4)
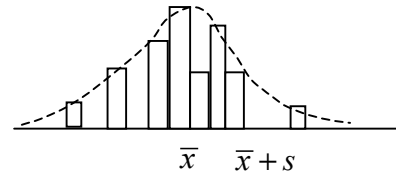
Note: parametric: assume a theoretical distribution (e.g., Gauss)
Non-parametric: no assumption made about the distribution

Advantages of assuming a parametric probability distribution:
➢ Compaction: just a few parameters
➢ Smoothing, interpolation, extrapolation

$$\overline{x} \quad \overline{x} + s$$

Parameter: e.g.: μ, σ population mean and standard deviation
Statistic: estimation of parameter from sample: $\overline{x}, s$ sample mean and standard deviation

Discrete distributions:
(e.g., yes/no; above normal, normal, below normal)

Binomial: $E_1 = 1$ (yes or success); $E_2 = 0$ (no, fail). These are MECE.

$P(E_1) = p \qquad P(E_2) = 1 - p$. Assume N independent trials.

How many "yes" we can obtain in N independent trials?

$x = (0, 1, ... N - 1, N)$, $N+1$ possibilities. Note that x is like a dummy variable.

$$P(X = x) = \binom{N}{x} p^x (1 - p)^{N-x}, \qquad \text{remember that} \binom{N}{x} = \frac{N!}{x!(N-x)!}, \quad 0! = 1$$

Bernouilli is the binomial distribution with a single trial, $N=1$:
$x = (0, 1), \quad P(X = 0) = 1 - p, \quad P(X = 1) = p$

Geometric: Number of trials until next success: i.e., x-1 fails followed by a success.
$P(X = x) = (1 - p)^{x-1} p \qquad\qquad x = 1, 2, ...$

Poisson: Approximation of binomial for small p and large N. Events occur randomly at a constant rate (per N trials) $\mu = Np$. The rate per trial p is low so that events in the same period (N trials) are approximately independent.

Example: assume the probability of a tornado in a certain county on a given day is p=1/100. Then **the average rate** per season is: $\mu = 90*1/100 = 0.9$.

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!} \qquad x = 0,1,2...$$

Question: What is the probability of having 0 tornados, 1 or 2 tornados in a season?

Expected Value: "probability weighted mean"

Example: Expected mean: $\mu = E(X) = \sum_x x.P(X = x)$

Example: Binomial distrib. mean $\mu = E(X) = \sum_{x=0}^{N} x \binom{N}{x} p^x(1-p)^{1-x} = Np$

Properties of expected value:

$$E(f(X)) = \sum_x f(x).P(X = x); \quad E(a.f(X) + b.g(X)) = a.E(f(X)) + b.E(g(X))$$

Example Variance

$$Var(X) = E((X - \mu)^2) = \sum_x (x - \mu)^2 P(X = x) =$$

$$= \sum_x x^2 P(X = x) - 2\mu \sum_x xP(X = x) + \mu^2 \sum_x P(X = x) = E(X^2) - \mu^2$$

E.g.: Binomial: $Var(X) = Np(1-p)$

Geometric: $Var(X) = \frac{(1-p)}{p^2}$

Poisson: $Var(X) = Np = \mu$

Continuous probability distributions
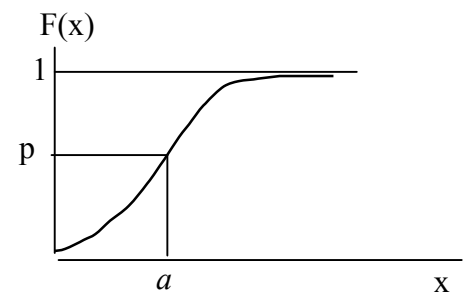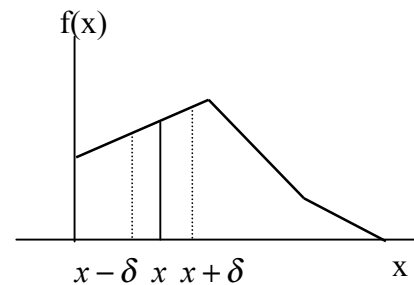
f(x)

Probability density function f(x): PDF

$$\int f(x)dx \equiv 1$$

$$\int_{x-\delta}^{x+\delta} f(x)dx = P(|X - x| < \delta)$$

$x - \delta \quad x \quad x + \delta$ x

F(x)

Cumulative prob distribution function (CDF)

1

$$F(a) = P(X \le a) = \int_{x \le a} f(x)dx$$

p

easy to invert: $a(F) = F^{-1}(P)$

a   x

Note on the use of CDF for empirical functional relationships:
The cumulative distribution function (CDF) can be easily inverted. This allows obtaining functional relationships between variables with different PDF's. For example, if we want to create random numbers with an arbitrary

PDF $p(x)$, we obtain first the corresponding CDF $F(x) = \int_{-\infty}^{x} p(u)du$

(which varies between 0 and 1). We then use computer-generated random numbers r_i uniformly distributed between 0 and 1 and invert the CDF to

obtain create random numbers x_i with the desired PDF: $F(x_i) = r_i$.

F(x)

1

r_i

0

x_i
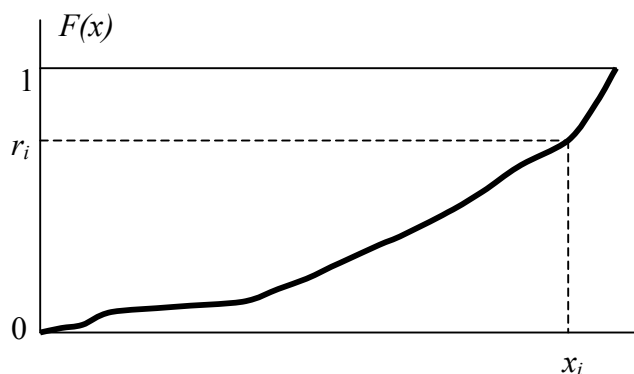
As an example, consider an empirical parameterization of the cloud cover (CC) as a function of the relative humidity (RH) at a grid point in a model. To get a realistic relationship we (scatter) plot (for different latitudes and altitudes) the observed joint distribution of RH and CC, for example:
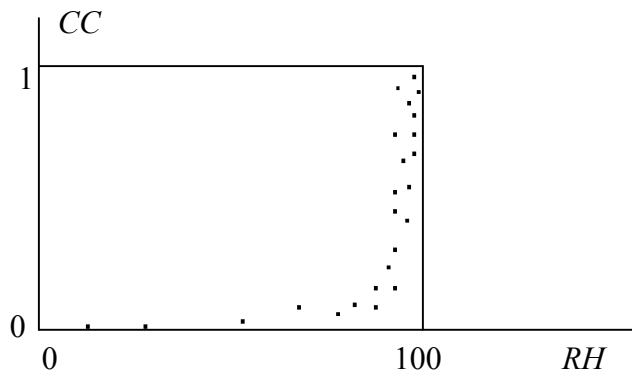
We can get an empirical relationship between RH and CC by computing their CDF(RH), CDF(CC). Then we associate to each value RH_i the corresponding value CC_i such that they have the same F: CDF(RH_i)=CDF(CC_i):
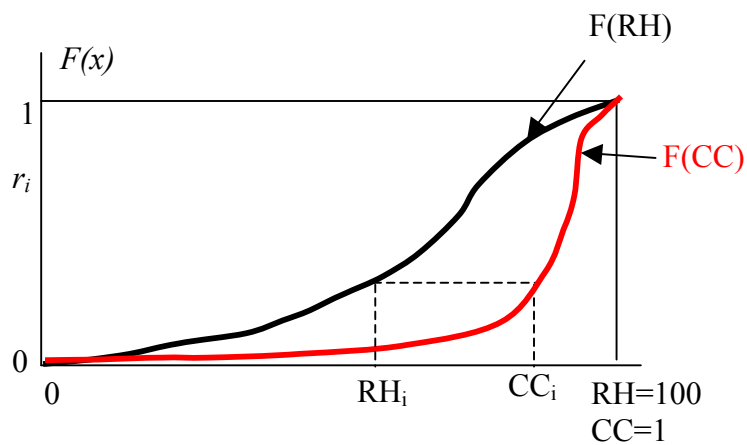
Expected value: Probability weighted average $E(g(X)) = \int_x g(x)f(x)dx$. Same

as for discrete distributions: $E = \sum_i g(x_i)p(x_i)$. For example, the

mean $\mu = E(X) = \int_x xf(x)dx$, and the variance

$$Var(X) = E((X-\mu)^2) = \int_x (x-\mu)^2 f(x)dx = \int_x (x^2 - 2\mu x + \mu^2)f(x)dx =$$

$$= \int_x x^2 f(x)dx - 2\mu \int_x xf(x)dx + \mu^2 \int_x f(x)dx = E(X^2) - \mu^2$$

An excellent resource is the NIST website
www.itl.nist.gov/div898/handbook/index.htm
and in particular the "gallery of distributions"
www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

The dataplot software seems very nice, but I have not tried it:
www.itl.nist.gov/div898/software/dataplot/

Gaussian or normal prob. distribution: $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Standard form (z is dimensionless) $z = \dfrac{x-\mu}{\sigma} \approx \dfrac{x-\bar{x}}{s};\quad f(z) = \dfrac{e^{-z^2/2}}{\sqrt{2\pi}}$

Central limit theorem: the average of a set of independent observations will have a Gaussian distribution for a large enough average.

"Well behaved" atmospheric variables (e.g., T): even a one-day average is approximately Gaussian. Multimodal or skewed variables (e.g., pp): require longer averages to look Gaussian.
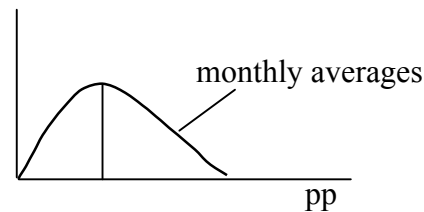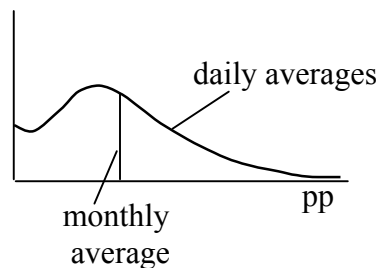
How to use a table of Gaussian probabilities: Estimate μ and σ from a sample and convert to a standard variable: $z = \dfrac{x-\mu}{\sigma} \simeq \dfrac{x-\bar{x}}{s}$. The table gives $F(z) = P(Z \le z)$.
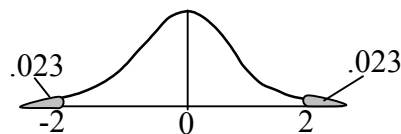
The area is $P(z_1 \le Z \le z_2) = F(z_2) - F(z_1)$
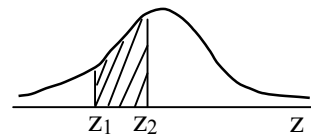
Example: What is the probability of a January average temperature of $T \le 0C$ if $\mu \simeq \bar{T}_{Jan} = 4^{\circ}C$ and $\sigma \simeq s = 2^{\circ}C$?

$z = \dfrac{T-4}{2} = \dfrac{-4}{2} = -2$ (-2 standard deviations)

$F(-2) = 0.023$. Note that

$P(|Z| \ge 2\sigma) = 2 * F(-2) = 2 * 0.023 \simeq 0.05 = 5\%$

What are the Gaussian terciles for the temperature distribution?

F(z)=0.666 →z=0.43 → $T = \mu \pm 0.43\sigma = 4^{\circ}C +/- 0.86^{\circ}C$
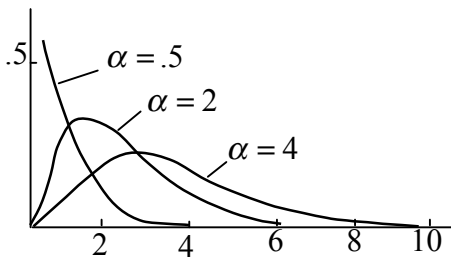
Other probability distributions:

Gamma: for data>0, positively skewed, such as precipitation.

$$f(x) = \frac{(x/\beta)^{\alpha-1} e^{-x/\beta}}{\beta \Gamma(\alpha)}$$
β: scale parameter

f(x)
α: shape parameter

.5 $\alpha = .5$
$\Gamma(\alpha) \equiv (\alpha-1)\Gamma(\alpha-1)$: gamma function

$\alpha = 2$

$\alpha = 4$
(For integers $\Gamma(n) = (n-1)!$)

2   4   6   8   10
$\mu = E(X) = \int_0^\infty x f(x) dx = \alpha\beta$,

$$\sigma^2 = E(X^2) - E(X)^2 = \alpha\beta^2$$

For $\alpha = 1$ the gamma distribution becomes the exponential distribution:

$$f(x) = \frac{e^{-x/\beta}}{\beta} \quad if \ x > 0, \ 0 \text{ otherwise. The cumulative distribution function is}$$

$$F(x) = 1 - e^{-x/\beta} \quad if \ x > 0, \ 0 \text{ otherwise.}$$

Beta: For data between 0 and 1 (e.g., RH, cloud cover, probabilities)

$$f(x) = \left[\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \right] x^{p-1}(1-x)^{q-1}, 0 \le x \le 1, p,q > 0$$

This is a very flexible function, taking on many different shapes depending on its two parameters p and q. Figure 4.11.

$$\mu = p / (p + q); \quad \sigma^2 = \frac{pq}{(p+q)^2(p+q+1)}$$

From these equations, the **moment estimators** (see parameter estimators) can be derived:

$$\hat{p} = \frac{\overline{x}^2(1-\overline{x})}{s^2} - \overline{x}; \quad \hat{q} = \frac{\hat{p}(1-\overline{x})}{\overline{x}}$$
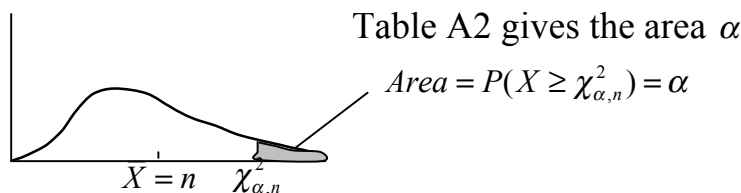
For the special case p=q=1, the PDF is f(x)=1, the uniform distribution (between 0 and 1).

Distributions arising from the normal pdf: (used for hypothesis testing)

χ^2 chi-square: If $Z_1, ..., Z_n$ are standard normal independent variables, then

$$X = Z_1^2 + ... + Z_n^2$$

has a χ^2 distribution with n degrees of freedom.

Table A2 gives the area α

$$Area = P(X \geq \chi^2_{\alpha,n}) = \alpha$$

$X = n \quad \chi^2_{\alpha,n}$

$$\chi^2 \equiv Gamma(\beta = 2, \alpha = n/2); \quad \overline{X} = E(X) = n; \quad Var(X) = E(X^2 - \overline{X}^2) = 2n$$

For example, with n=10, a=5%, from table A2, we can find x=18.307. This means that if we have a data set which is the sum of standard normal independent variables (e.g., temperature minus the mean divided by the standard deviation) the expected value is n=10, and the probability of finding a value larger than 18.3 is less than 5%.

The exponential function is also the same as the chi-square for 2 degrees of freedom. For this reason it is appropriate for wind speed $s = (u^2 + v^2)$.

An important application of the chi-square is to test **goodness of fit**:

If you have a histogram with n bins, and a number of observations O_i and expected number of observations E_i (e.g., from a parametric distribution) in each bin, then the goodness of fit of the pdf to the data can be estimated using a chi-square test:
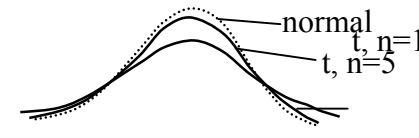
$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \text{with n-1 degrees of freedom}$$

The null hypothesis (that it is a good fit) is rejected at a 5% level of significance if $X^2 > \chi^2_{(0.05,n-1)}$.

t-distribution: If Z is normal, and $\dfrac{\chi^2}{n} = \dfrac{Z_1^2 + ... + Z_n^2}{n}$ then the random variable

$T_n = \dfrac{Z}{\sqrt{\chi^2 / n}}$ has a t-distribution with n-degrees of freedom (Table A3). If
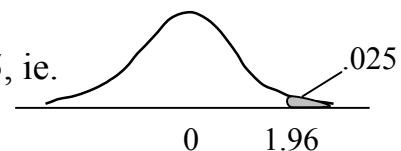
n>5, it is very close to a normal:

—normal
t, n=1
t, n=5

For example:
Normal distribution: (Table A1) $\Phi(1.96)=0.975$, ie.
T-distribution: (Table A3)   a=0.025

.025

0   1.96

With   n=10, $t_{a,n} = 2.228$
        n=20, $t_{a,n} = 2.086$
        n=∞, $t_{a,n} = 1.96$

Parameter estimation (fitting a distribution to observations)

1) Moments fitting: compute the first two moments from a sample and then use distribution: $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n};\quad s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$, and then use these values in the Gaussian or other distribution. For example, for the Gaussian distribution, simply use $\hat{\mu} = \bar{x}; \hat{\sigma}^2 = s^2$, and for the gamma distribution,

$$\bar{x} = \hat{\alpha}\hat{\beta}; s^2 = \hat{\alpha}\hat{\beta}^2 \quad \Rightarrow \hat{\beta} = \frac{s^2}{\bar{x}}; \hat{\alpha} = \frac{\bar{x}^2}{s^2}$$

The NIST web site gives the mean and the standard deviation relationship to the parameters of the probability distribution.
http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

2) Maximum likelihood method: maximize the probability of the distribution fitting all the observations $\{x_i\}$. The probability of having obtained the observations is the product of the probabilities for each observation (for a Gaussian distribution), i.e.

$$I(\mu,\sigma) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma^n \sqrt{(2\pi)^n}} e^{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}},$$

or maximizing its logarithm:

$$L(\mu,\sigma) = \ln(I) = -n\ln\sigma - n\ln\sqrt{(2\pi)} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

Then $\frac{\partial L}{\partial \mu} = 0, \frac{\partial L}{\partial \sigma} = 0$ gives the maximum likelihood parameters μ, σ. Note that for the Gaussian distribution, this gives $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (same as with the momentum fitting), but that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$. The most likely value for the standard deviation is not the unbiased estimator s.

Note: Likelihood is the probability distribution of the truth given a measurement. It is equal to the probability distribution of the measurement given the truth (Edwards, 1984).

Goodness of fit
Methods to test goodness of fit:

a) Plot a PDF over the histogram and check how well it fits, (Fig 4.14) or

b) check how well scatter plots of quantiles from the histogram vs quantiles from the PDF fall onto the diagonal line (Fig 4.15)
   1) A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
   2) Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2.

c) Use the chi-square test (see above): $X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

The fit is considered good (at a 5% level of significance) if

$$X^2 < \chi^2_{(0.05, n-1)}$$

Extreme events: Gumbel distribution

Examples: coldest temperature in January, maximum daily precipitation in a summer, maximum river flow in the spring. Note that there are two time scales: a short scale (e.g., day) and a long scale: a number of years.

Consider now the problem of obtaining the *maximum* (e.g., warmest temperature) extreme probability distribution:

CDF: $F(x) = \exp\left\{-\exp\left[-\dfrac{x-\xi}{\beta}\right]\right\}$: this can be derived from the

exponential distribution (von Storch and Zwiers, p49). The PDF can be obtained from the CDF:

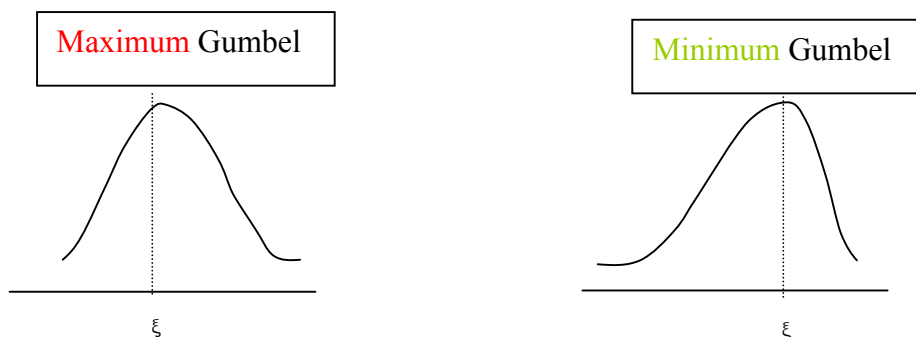$$\text{PDF: } f(x) = \frac{1}{\beta}\exp\left\{-\exp\left[-\frac{x-\xi}{\beta}\right] - \frac{x-\xi}{\beta}\right\}$$

Parameter estimation for the maximum distribution

$\hat{\beta} = \dfrac{s\sqrt{6}}{\pi}; \hat{\xi} = \bar{x} - \gamma\hat{\beta}, \quad \gamma = 0.57721...$, Euler constant.

Note that $\bar{x} = \hat{\xi} + \gamma\hat{\beta}$ indicates that for the maximum Gumbel

distribution, the mean is to the right of $\hat{\xi}$, which is the mode (value for which the pdf is maximum, or most popular value, check the pdf figure). Therefore, for the minimum distribution, since the mean is to the left of the mode (check the pdf figure), the parameters are:

$\hat{\beta} = \dfrac{s\sqrt{6}}{\pi}; \hat{\xi} = \bar{x} + \gamma\hat{\beta} \quad (i.e., \bar{x} = \hat{\xi} - \gamma\hat{\beta}), \quad \gamma = 0.57721...$

Maximum Gumbel   Minimum Gumbel

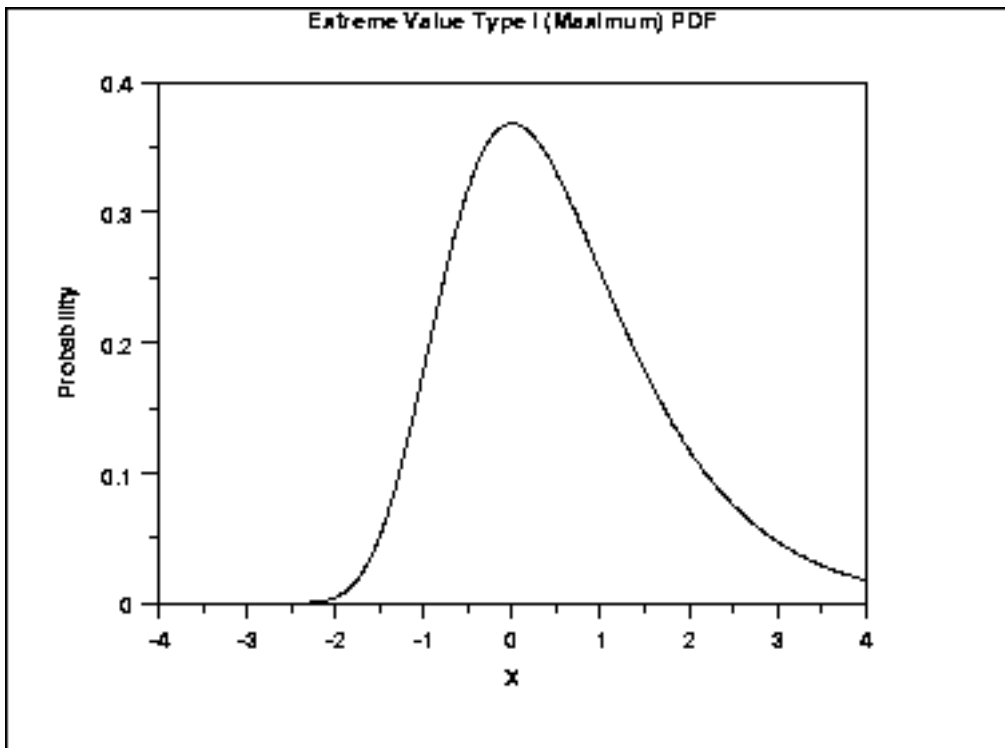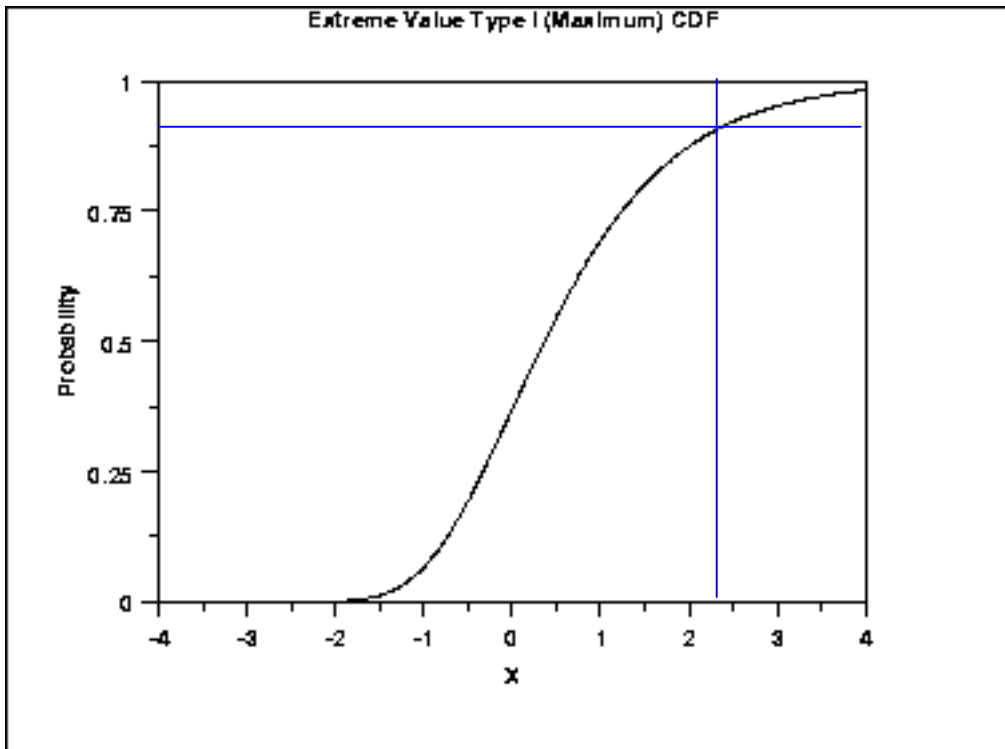If we take $\xi = 0$, and $\beta = 1$ we get the standard Gumbel maximum pdf, with the corresponding standard Gumbel maximum CDF (in parenthesis):

PDF: $f(x) = \dfrac{1}{\beta} \exp\left\{ -\exp\left[-\dfrac{x-\xi}{\beta} \right] - \dfrac{x-\xi}{\beta} \right\}$ $(= e^{-e^{-x} - x})$

CDF: $F(x) = \exp\left\{ -\exp\left[-\dfrac{x-\xi}{\beta} \right] \right\}$ $(= e^{-e^{-x}})$

The *standard* maximum PDF and CDF are plotted below:

Extreme Value Type I (Maximum) PDF

Extreme Value Type I (Maximum) CDF

The horizontal line indicates CDF=0.9, corresponding to a "return time" of 10 years, and the vertical line the "10-year return value", i.e., the value of the standardized variable (e.g., temperature) such that on the average we have to wait 10 years until we see such a value (or larger) again. A CDF=0.99 would correspond to a "100-year return value".

Once we choose the CDF for a return value, let's say

$$F(x) = \exp\left\{-\exp\left[-\frac{x-\xi}{\beta}\right]\right\} = 0.99 \quad \text{for 100 year return, we}$$

can invert it to obtain the 100-year return value itself:

$$x_{100\,years} = F^{-1}(x) = \xi - \beta\ln(-\ln F_{100\,years})$$

This can also be obtained graphically from the standard Gumbel CDF.

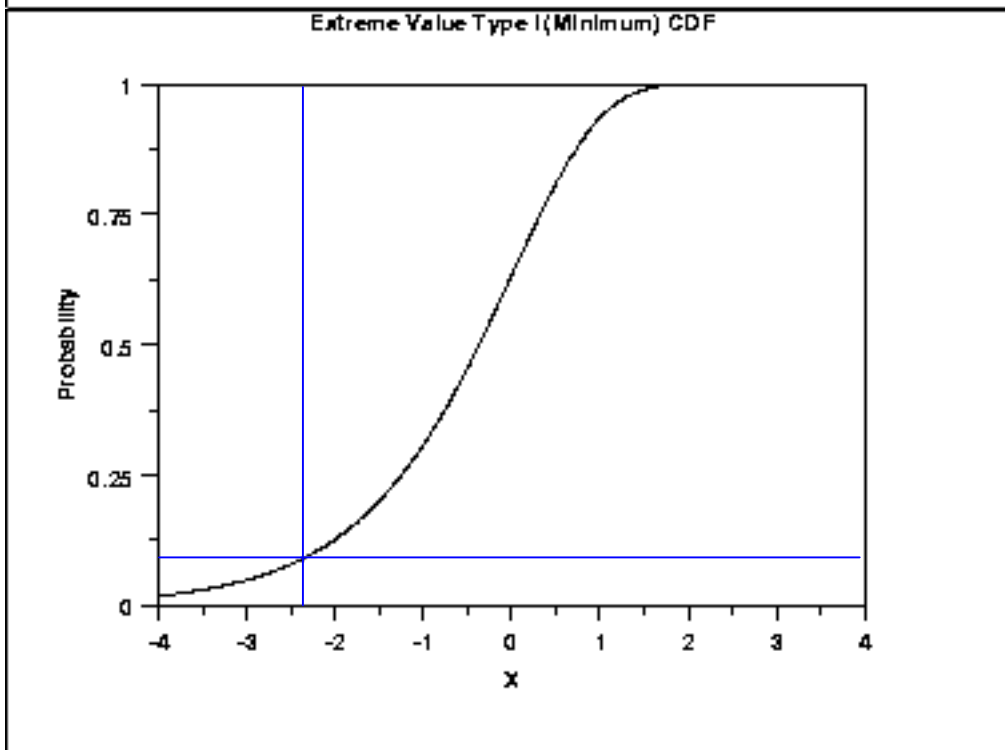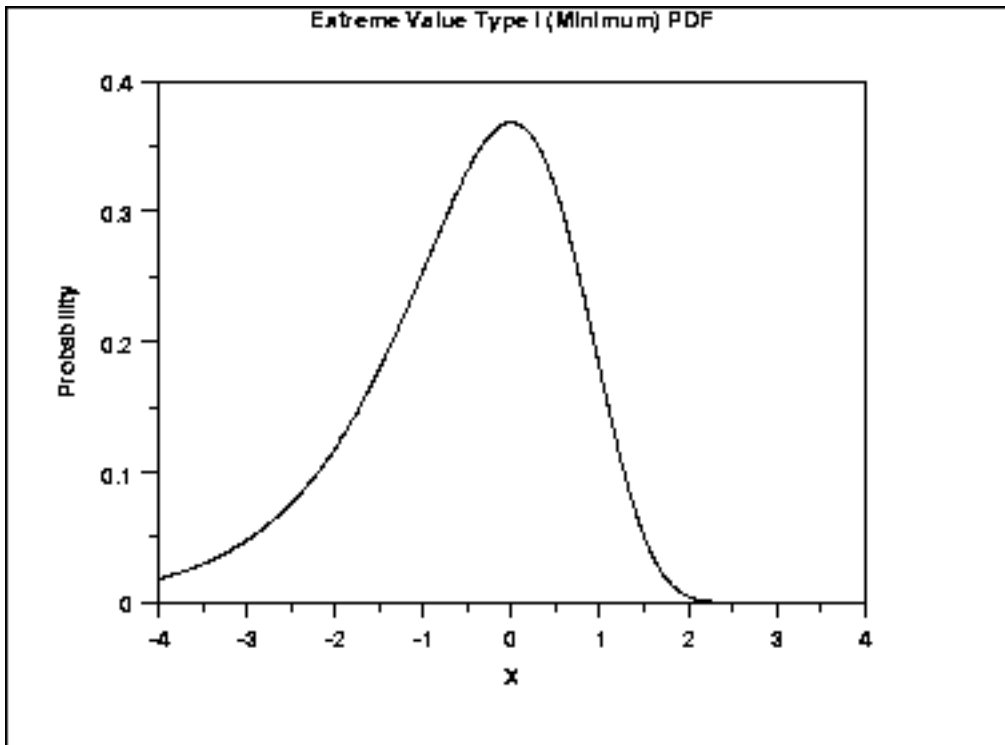If instead of looking for the maximum extreme event we are looking for the minimum (e.g., the coldest) extreme event, we have to reverse the normalized $\dfrac{x-\xi}{\beta}$ to $-\dfrac{x-\xi}{\beta}$. The Gumbel minimum distributions become (with the standard version in parenthesis)

PDF: $f(x) = \dfrac{1}{\beta}\exp\left\{-\exp\left[\dfrac{x-\xi}{\beta}\right] + \dfrac{x-\xi}{\beta}\right\}$ $(= e^{-e^{x}+x})$

The integral of $e^{-e^{x}+x}$ is $-e^{-e^{x}} + const = 1 - e^{-e^{x}}$ so that

CDF: $F(x) = 1 - \exp\left\{-\exp\left[\dfrac{x-\xi}{\beta}\right]\right\}$ $(= 1 - e^{-e^{x}})$

The PDF and CDF plots follow:

Extreme Value Type I (Minimum) PDF

Extreme Value Type I (Minimum) CDF

Return year:
For Minimum Gumbel distribution
How many years we need wait to see x<=X happen again?

P(x<= -2.3)
=CDF(x=-2.3)=0.1

return year= 1/CDF
=10YEARS

The return time of 10 years (marked in blue) is the extreme value that has a cumulative probability of 0.1 (for the minimum) or 0.9 (for the maximum).

Example of PDF and CDF for Gumbel (maximum) distribution: $\beta = 1, \quad \xi = 1.79 = \ln 6$.

Note: The return time is computed from the CDF. The CDF probability 0.5 (which means that on the average it happens every other year) corresponds to a return time of 2 years, 0.9 to 10 years, etc. The PDF is only used to compare with a histogram.
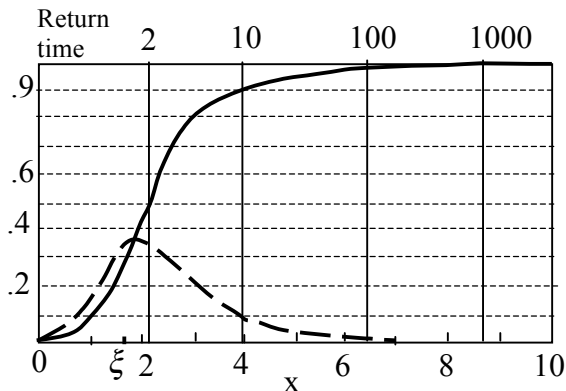
Multivariate Gaussian distributions

For multivariate problems, Gaussian distributions are in practice used almost exclusively…

For one (scalar) variable, the Gaussian distribution can be written as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)\sigma^{-2}(x-\mu)}{2}} \quad \text{or} \quad f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)(z)}{2}}$$

For two variables

$$f(x_1, x_2) = \frac{1}{\left(\sqrt{2\pi}\right)^2 \sqrt{\begin{vmatrix} \sigma_1^2 & \overline{x_1'x_2'} \\ \overline{x_1'x_2'} & \sigma_2^2 \end{vmatrix}}} e^{-\frac{\left[\binom{x_1}{x_2} - \binom{\mu_1}{\mu_2}\right]^T \Sigma^{-1} \left[\binom{x_1}{x_2} - \binom{\mu_1}{\mu_2}\right]}{2}}$$

where $\Sigma = \begin{bmatrix} \sigma_1^2 & \overline{x_1'x_2'} \\ \overline{x_1'x_2'} & \sigma_2^2 \end{bmatrix}$ is the covariance matrix.

With two standardized variables

$$f(z_1, z_2) = \frac{1}{\left(\sqrt{2\pi}\right)^2 \sqrt{\begin{vmatrix} 1 & \rho \\ \rho & 1 \end{vmatrix}}} e^{-\frac{\left[\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right]^T R^{-1} \left[\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right]}{2}}$$

where $R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ is the correlation matrix.

For k variables, we define a vector $\mathbf{x} = \begin{pmatrix} x_1 \\ \dots \\ x_k \end{pmatrix}$

and $f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^k \sqrt{|\Sigma|}} e^{-\frac{[\mathbf{x}-\mu]^T \Sigma^{-1} [\mathbf{x}-\mu]}{2}}$.

For standardized variables,

$f(\mathbf{z}) = \frac{1}{\left(\sqrt{2\pi}\right)^k \sqrt{|R|}} e^{-\frac{[\mathbf{z}]^T R^{-1} [\mathbf{z}]}{2}}$ where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & \overline{x_1'x_k'} \\ \dots & \dots & \dots \\ \overline{x_1'x_k'} & \dots & \sigma_k^2 \end{bmatrix} \text{ and } R = \begin{bmatrix} 1 & \dots & \rho_{1k} \\ \dots & \dots & \dots \\ \rho_{k1} & \dots & 1 \end{bmatrix}$$

are the covariance and correlation matrices respectively.
See figure 4.2/4.5 of Wilks showing a bivariate distribution.