1. Fundamentals

1.1 Probability Theory

2.1 Theoretical Distributions

2.1.1 Discrete Distributions

The Binomial Distribution

The Poisson Distribution

2.1.2 Continuous Distributions

The Normal Distribution

Gamma Distribution

Gumbel Distribution

The $x^2$ Distribution

The $t$ Distribution

The $F$ distribution

2.1.3 Multivariate Distributions
1.1 Theoretical Distributions

**Bernoulli Distribution**

A Bernoulli experiment is one in which there are just two outcomes of interests – event $A$ occurs or does not occur.

The indicator function of the event $A$ is called a Bernoulli random variable:

$$ X = \begin{cases} 
1, & \text{if } A \text{ occurs} \\
0, & \text{if } A^c \text{ occurs} 
\end{cases} $$

$$ f(1) = P(A) = p, \quad f(0) = P(A^c) = q = 1 - p \quad \text{or} $$

$$ f(x) = p^x q^{1-x}, \text{ for } x = 0, 1 $$

**Examples:**
- tossing a coin (heads or tails)
- testing a product (good or defective)
- precipitation (rain or non rain)

The mean and the variance are

$$ EX = 1 \cdot p + 0 \cdot q = p $$

$$ Var(X) = EX^2 - (EX)^2 = p - p^2 = pq $$
1.1 Theoretical Distributions

**Binomial Distribution** $B(n,p)$

Consider the independent and identically distributed random variables $X_1, \ldots, X_n$, which are the results of $n$ Bernoulli trials. The number of successes among $n$ trials, which is the sum of the 0’s and 1’s resulting from the individual trails

$$S_n = X_1 + \cdots + X_n$$

is described by a Binomial distribution and has the probability

$$P(S_n = k) = P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k q^{n-k}$$

$p^k q^{n-k}$: the probability of a particular pattern of $k$ successes among $n$ trials

$\binom{n}{k}$: the number of distinct patterns in which there are exactly $k$ successes

The mean and variance are

$$ES_n = EX_1 + \cdots + EX_n = np$$

$$Var(S_n) = Var(X_1) + \cdots + Var(X_n) = npq$$
1.1 Theoretical Distributions

**Approximate Binomial Probabilities**

According to the Central Limit Theorem, the distribution of the sum $S_n$ and hence the binomial distribution, is asymptotically normal for large $n$. More precisely, for fixed $p$

$$\lim_{n \to \infty} = P\left( \frac{S_n - np}{\sqrt{npq}} \leq z \right) = F_{N(0,1)}(z)$$

**Example:**

Consider a binomial distribution with $n=8$ and $p=1/2$. The binomial distribution is

$$P(X \leq x) = \sum_{k \leq x} \binom{8}{k} \left( \frac{1}{2} \right)^k \left( \frac{1}{2} \right)^{n-k}$$

and the standard normal distribution is

$$P(X \leq x) = F\left( \frac{x - 4}{\sqrt{2}} \right)$$

which approximates the binomial distribution well.
### 1.1 Theoretical Distributions

#### Example: Binomial Distribution and the Freezing of Cayuga Lake

Given the years, during which the Cayuga Lake in central New York State was observed to have frozen in the last 200 years,

- 1796
- 1816
- 1856
- 1875
- 1884
- 1904
- 1912
- 1934
- 1961
- 1979

what is the probability for the lake freezing at least once during some decade in the future?

**Model:** the number of years of lake freezing in 10 years is a binomial distributed random variable

- Lake freezing is a Bernoulli experiment
- Whether it freezes in a given winter is independent of whether it froze in recent years
- The probability that the lake will freeze in a given winter is constant

**Estimating the model parameter**

\[ \hat{p} = \frac{10}{200} = 0.05 \]

**Prediction given by the model**

\[
P(S_n = 1) = \binom{10}{1} \cdot 0.05^1 \cdot (1 - 0.05)^{10-1} = \frac{10!}{1!9!} (0.05)(0.95^9) = 0.32
\]
1.1 Theoretical Distributions

**Poisson Distribution** \( P(\lambda t) \)

The random variable

\( X = \text{number of events in an interval of width } t \)

is described by the density

\[
f(x) = P(x \text{ events in time } t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!} = \frac{e^{-m} (m)^x}{x!}
\]

A Poisson variable has the mean and variance

\[
EX = \lambda t, \quad Var(X) = \lambda t
\]
Relation to Binomial Distribution

- The Poisson distribution arises when concerning rare events (e.g. wind speed larger than $V_c$)

- Under the assumption of a constant $\lambda$, the base interval $t$ is divided into $n$ equal length sub-intervals with $n$ being large enough so that the likelihood of two exceedances in any one sub-interval is negligible

- The occurrence of an exceedance in any one sub-interval can be approximated as a Bernoulli trial with probability $\lambda t/n$ of success

- Assume that events in adjacent sub-intervals are independent. The number of exceedance $X$ in interval $t$ is binomially distributed with $B(n, \lambda t/n)$

- $n \to \infty$, $B(n, \lambda t/n) \to P(\lambda t)$
1.1 Theoretical Distributions

Example: Poisson Distribution and Annual Tornado Counts

Given the annual tornado counts in New York State between 1959-1988, what are the probabilities for zero tornado per year or for greater than nine tornados per year?

Model: the annual number of tornados is a Poisson distributed random variable

Estimating the model parameter $\lambda$: The rate of tornado occurrence:

$\hat{m} = \frac{138}{30} = 4.6$

Prediction of the model:

$f(0) = P_0(1) = \frac{e^{-4.6} \cdot 4.6^0}{0!} = 0.01$

$P(x > 9) = \sum_{x>9} f(x) = \left( \frac{e^{-4.6} \cdot 4.6^{10}}{10!} + \cdots \right)$

Histogram of number of tornados reported annually in New York State for 1959-1988 (dashed), and fitted Poisson distribution with $m=4.6$ tornados per year (solid)
1.1 Theoretical Distributions

**The Normal Distribution \( N(\mu, \sigma^2) \)**

\[
f_N(x) = \frac{1}{\sqrt{2\pi \sigma}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)
\]

\[
F_N(x) = \frac{1}{\sqrt{2\pi \sigma}} \int_{-\infty}^{x} \exp\left(-\frac{(t - \mu)^2}{2\sigma^2}\right) dt
\]

**The Standard Normal Distribution \( N(0,1) \)**

\[
f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)
\]

Any normal distribution can be transformed to \( N(0,1) \) using \( Z=(X-\mu)/\sigma \)

- described by two parameters \( \mu \) and \( \sigma \)
- symmetric
- for a \( N(0,1) \) random variable, probability for values larger than 1.96\( \sigma \) (2.58\( \sigma \)) relative to the mean is smaller than 5\% (1\%). Nearly all values lie within \([-3\sigma, 3\sigma]\).
- the c.d.f cannot be given explicitly
1.1 Theoretical Distributions

Normal Distribution

\[ f_N(x) \]

\[ \sigma = 1 \]

\[ \sigma = 3 \]

\[ F_N(x) \]

\[ \sigma = 3 \]

\[ \sigma = 1 \]
1.1 Theoretical Distributions

### Normal Density and Cumulative Distribution Function

Values of the standard normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu = 0$ and variance $\sigma^2 = 1$. The density function $f_X(z)$ is given by

$$f_X(z) = \frac{1}{\sqrt{2\pi}} e^{-(z-\mu)^2 / 2\sigma^2}$$

and the exact cumulative distribution function $F_X(z) = \int_{-\infty}^{z} f_X(x) \, dx$. The column labelled $F_X^*(z)$ contains the approximated cumulative distribution function.

<table>
<thead>
<tr>
<th>$z$</th>
<th>$f_X(z)$</th>
<th>$F_X(z)$</th>
<th>$F_X^*(z)$</th>
<th>$z$</th>
<th>$f_X(z)$</th>
<th>$F_X(z)$</th>
<th>$F_X^*(z)$</th>
<th>$z$</th>
<th>$f_X(z)$</th>
<th>$F_X(z)$</th>
<th>$F_X^*(z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.00</td>
<td>0.004</td>
<td>0.001</td>
<td>0.001</td>
<td>-2.00</td>
<td>0.054</td>
<td>0.023</td>
<td>0.023</td>
<td>-1.00</td>
<td>0.242</td>
<td>0.159</td>
<td>0.156</td>
</tr>
<tr>
<td>-2.95</td>
<td>0.005</td>
<td>0.002</td>
<td>0.001</td>
<td>-1.95</td>
<td>0.060</td>
<td>0.026</td>
<td>0.023</td>
<td>-0.95</td>
<td>0.254</td>
<td>0.171</td>
<td>0.169</td>
</tr>
<tr>
<td>-2.90</td>
<td>0.006</td>
<td>0.002</td>
<td>0.001</td>
<td>-1.90</td>
<td>0.066</td>
<td>0.029</td>
<td>0.025</td>
<td>-0.90</td>
<td>0.266</td>
<td>0.184</td>
<td>0.182</td>
</tr>
<tr>
<td>-2.85</td>
<td>0.007</td>
<td>0.002</td>
<td>0.001</td>
<td>-1.85</td>
<td>0.072</td>
<td>0.032</td>
<td>0.029</td>
<td>-0.85</td>
<td>0.278</td>
<td>0.198</td>
<td>0.196</td>
</tr>
<tr>
<td>-2.80</td>
<td>0.008</td>
<td>0.003</td>
<td>0.002</td>
<td>-1.80</td>
<td>0.079</td>
<td>0.036</td>
<td>0.033</td>
<td>-0.80</td>
<td>0.290</td>
<td>0.212</td>
<td>0.210</td>
</tr>
<tr>
<td>-2.75</td>
<td>0.009</td>
<td>0.003</td>
<td>0.002</td>
<td>-1.75</td>
<td>0.086</td>
<td>0.040</td>
<td>0.037</td>
<td>-0.75</td>
<td>0.301</td>
<td>0.227</td>
<td>0.225</td>
</tr>
<tr>
<td>-2.70</td>
<td>0.010</td>
<td>0.003</td>
<td>0.002</td>
<td>-1.70</td>
<td>0.094</td>
<td>0.045</td>
<td>0.041</td>
<td>-0.70</td>
<td>0.312</td>
<td>0.242</td>
<td>0.241</td>
</tr>
<tr>
<td>-2.65</td>
<td>0.012</td>
<td>0.004</td>
<td>0.003</td>
<td>-1.65</td>
<td>0.102</td>
<td>0.049</td>
<td>0.046</td>
<td>-0.65</td>
<td>0.323</td>
<td>0.258</td>
<td>0.257</td>
</tr>
<tr>
<td>-2.60</td>
<td>0.014</td>
<td>0.005</td>
<td>0.003</td>
<td>-1.60</td>
<td>0.111</td>
<td>0.055</td>
<td>0.051</td>
<td>-0.60</td>
<td>0.333</td>
<td>0.274</td>
<td>0.273</td>
</tr>
<tr>
<td>-2.55</td>
<td>0.016</td>
<td>0.005</td>
<td>0.004</td>
<td>-1.55</td>
<td>0.120</td>
<td>0.061</td>
<td>0.057</td>
<td>-0.55</td>
<td>0.343</td>
<td>0.291</td>
<td>0.290</td>
</tr>
<tr>
<td>-2.50</td>
<td>0.018</td>
<td>0.006</td>
<td>0.005</td>
<td>-1.50</td>
<td>0.130</td>
<td>0.067</td>
<td>0.063</td>
<td>-0.50</td>
<td>0.352</td>
<td>0.308</td>
<td>0.308</td>
</tr>
<tr>
<td>-2.45</td>
<td>0.020</td>
<td>0.007</td>
<td>0.005</td>
<td>-1.45</td>
<td>0.139</td>
<td>0.073</td>
<td>0.070</td>
<td>-0.45</td>
<td>0.361</td>
<td>0.326</td>
<td>0.326</td>
</tr>
<tr>
<td>-2.40</td>
<td>0.022</td>
<td>0.008</td>
<td>0.006</td>
<td>-1.40</td>
<td>0.150</td>
<td>0.081</td>
<td>0.078</td>
<td>-0.40</td>
<td>0.368</td>
<td>0.345</td>
<td>0.344</td>
</tr>
<tr>
<td>-2.35</td>
<td>0.025</td>
<td>0.009</td>
<td>0.007</td>
<td>-1.35</td>
<td>0.160</td>
<td>0.088</td>
<td>0.085</td>
<td>-0.35</td>
<td>0.375</td>
<td>0.363</td>
<td>0.363</td>
</tr>
<tr>
<td>-2.30</td>
<td>0.028</td>
<td>0.011</td>
<td>0.009</td>
<td>-1.30</td>
<td>0.171</td>
<td>0.097</td>
<td>0.094</td>
<td>-0.30</td>
<td>0.381</td>
<td>0.382</td>
<td>0.382</td>
</tr>
<tr>
<td>-2.25</td>
<td>0.032</td>
<td>0.012</td>
<td>0.010</td>
<td>-1.25</td>
<td>0.183</td>
<td>0.106</td>
<td>0.103</td>
<td>-0.25</td>
<td>0.387</td>
<td>0.401</td>
<td>0.401</td>
</tr>
<tr>
<td>-2.20</td>
<td>0.036</td>
<td>0.014</td>
<td>0.011</td>
<td>-1.20</td>
<td>0.194</td>
<td>0.115</td>
<td>0.112</td>
<td>-0.20</td>
<td>0.391</td>
<td>0.421</td>
<td>0.421</td>
</tr>
<tr>
<td>-2.15</td>
<td>0.040</td>
<td>0.016</td>
<td>0.013</td>
<td>-1.15</td>
<td>0.206</td>
<td>0.125</td>
<td>0.122</td>
<td>-0.15</td>
<td>0.395</td>
<td>0.440</td>
<td>0.440</td>
</tr>
<tr>
<td>-2.10</td>
<td>0.044</td>
<td>0.018</td>
<td>0.015</td>
<td>-1.10</td>
<td>0.218</td>
<td>0.136</td>
<td>0.133</td>
<td>-0.10</td>
<td>0.397</td>
<td>0.460</td>
<td>0.460</td>
</tr>
<tr>
<td>-2.05</td>
<td>0.049</td>
<td>0.020</td>
<td>0.017</td>
<td>-1.05</td>
<td>0.230</td>
<td>0.147</td>
<td>0.145</td>
<td>-0.05</td>
<td>0.398</td>
<td>0.480</td>
<td>0.480</td>
</tr>
<tr>
<td>-2.00</td>
<td>0.054</td>
<td>0.023</td>
<td>0.020</td>
<td>-1.00</td>
<td>0.242</td>
<td>0.159</td>
<td>0.156</td>
<td>0.00</td>
<td>0.399</td>
<td>0.500</td>
<td>0.500</td>
</tr>
</tbody>
</table>

The following table lists the upper tail critical values of the standard normal distribution commonly used in tests of hypothesis. These values are the solutions of $\tilde{p} = F_X(z)$. Lower tail critical values are given by $Z_{\tilde{p}} = -Z_{1-\tilde{p}}$.

<table>
<thead>
<tr>
<th>$\tilde{p}$</th>
<th>0.900</th>
<th>0.950</th>
<th>0.975</th>
<th>0.990</th>
<th>0.995</th>
<th>0.999</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_{\tilde{p}}$</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.080</td>
</tr>
</tbody>
</table>
1.1 Theoretical Distributions

Example: The normal distribution and the mean January temperature

Suppose that the mean January temperature at Ithaca (New York State) is a random variable with a Gaussian distribution with

\[ \mu = 22.2 \text{F} \quad \sigma = 4.4 \text{F} \]

What is the probability that an arbitrarily selected January will have mean temperature as cold as or colder than 21.32 F (23.08F)?

- transforming the temperature into a standard normal variable

\[ z = (21.32 \text{F} - 22.2 \text{F}) / 4.4 \text{F} = -0.2 \]
\[ z = (23.08 \text{F} - 22.2 \text{F}) / 4.4 \text{F} = +0.2 \]

- looking up the table

\[ F(z = -0.2) = 0.421 \]
\[ F(z = +0.2) = 0.579 \]

Since Gaussian distribution is symmetric, one has

\[ F(z) = P(Z \leq z) = 1 - P(Z > z) = 1 - P(Z < -z) = 1 - F(-z) \]
1.1 Theoretical Distributions

**Example: The normal distribution and the mean January temperature**

The average January temperature over the US are available for the period 1951-1980 and the year 1989. How can one assess the probabilities of the 1989 temperatures?

**Model:** the mean January temperatures at each stations are normally distributed random variables

Estimating the model parameters:

\[
\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2
\]

\(x_i\): available data
1.1 Theoretical Distributions

Probabilities of 1989 January mean temperature obtained from the model

- The probabilities are described in terms of percentiles of local Gaussian distributions

- Florida and most of the midwest were substantially warmer than usual, while only a small portion of the western United States was cooler than usual
1.1 Theoretical Distributions

**Gamma Distribution: \( G(\alpha, \beta) \)**

\[
f(x) = \frac{(x / \beta)^{\alpha-1} \exp(-x / \beta)}{\beta \Gamma(\alpha)}, \quad x, \alpha, \beta > 0
\]

\( \Gamma(\alpha + 1) = \alpha \Gamma(\alpha), \Gamma(1) = 1 \)

- defined for positive real numbers
- characterized by two parameters:
  - the shape parameter \( \alpha \) and the scale parameter \( \beta \)
- skewed
- mean and variance: \( \mu = \alpha \beta, \quad \sigma^2 = \alpha \beta^2 \)
- approaches normal distribution for large \( \alpha \)
- the c.d.f. is not explicit
- any gamma distribution can be transformed to the standard Gamma distribution using the transformation \( \xi = \frac{x}{\beta} \)

\( \alpha < 1: f(x) \to \infty \) as \( x \to 0 \)
\( \alpha = 1: f(0) = 1/\beta \)
  (exponential distribution)
\( \alpha > 1: f(0) = 0 \)
1.1 Theoretical Distributions

**Example: Gamma Distribution and Precipitation amounts**

The January precipitation amounts over the US are available for the period 1951-1980. How can one assess the probabilities of the 1989 January precipitation amounts?

**Model:** The precipitation amounts at each stations are random variables with Gamma distributions

**Estimating model parameters:**

- the method of moments (not bad for large \( \alpha \))

\[
\hat{\alpha} = \frac{\hat{\mu}}{\hat{\sigma}^2}, \quad \hat{\beta} = \frac{\hat{\sigma}^2}{\hat{\mu}}
\]

with \( \hat{\mu} \) and \( \hat{\sigma}^2 \) being estimates of the mean and variance

- maximum likelihood estimators

\[
\hat{\alpha} = \frac{1+\sqrt{1+4D/3}}{4D}, \quad \hat{\beta} = \frac{\hat{\mu}}{\hat{\alpha}}
\]

with \( D = \ln \hat{\mu} - \frac{1}{N} \sum_{i=1}^{N} \ln(x_i) \), \( x_i \) are the available data
1.1 Theoretical Distributions

Gamma distribution shape parameters for January precipitation over the conterminous United States. The distribution in the southwest are strongly skewed, while those for the most locations in the east are much more symmetrical. The distribution were fit using data from 30 years 1951-1980.
1.1 Theoretical Distributions

Answer obtained from the model:

- The probabilities for January precipitation amounts in 1989 are described in terms of percentile values of local Gamma distributions.

- Portions of the east and west were drier than usual, while parts of the central portion of the country were wetter.
1.1 Theoretical Distributions

**Gumbel Distribution** \( \Gamma(\beta, \zeta) \)

\[
f(x) = \frac{1}{\beta} \exp \left\{ - \exp \left[ - \frac{(x - \zeta)}{\beta} \right] - \frac{(x - \zeta)}{\beta} \right\}
\]

- Gumble distribution is used to describe extremes
- It has two parameters: the location parameter \( \zeta \) and the scale parameter \( \beta \)
- The density is skewed to the right and has a maximum at \( \zeta \)
- The c.d.f. is
  \[
  F(x) = \exp \left\{ - \exp \left[ - \frac{(x - \zeta)}{\beta} \right] \right\}
  \]
- The first two moments are
  \[
  \mu = \zeta + \gamma \beta, \quad \sigma^2 = \beta^2 \pi^2 / 6
  \]
  \[
  \gamma = 0.57721\ldots, \text{ Euler's constant}
  \]
1.1 Theoretical Distributions

**Return Values**

-The return values for preset periods (e.g. 10, 50, 100) years are thresholds that, according to the model, will be exceeded on average once every return period.

- Return values are the upper quantiles of the extreme value distribution.

- Example: Suppose that the random variable $Y$ represents an annual extreme maximum and that $Y$ has the probability density distribution $f_Y(y)$. The 10-year return value for $Y$ is the value $Y_{(10)}$ such that

$$P(Y > y_{(10)}) = \int_{y_{(10)}}^{\infty} f_Y(y)dy = \frac{1}{10} $$

In general, the $T$-year return value for the annual maximum, $Y_{(T)}$, is the solution of

$$\int_{y_{(T)}}^{\infty} f_Y(y)dy = \frac{1}{T}$$

An example of Gumble distribution $f_Y$ and its c.d.f. for annual maximum with $z=\ln 6$ and $b=1$. The locations of 2, 5, 10, 100 and 1000 year return values are indicated by the vertical bars.
### 1.1 Theoretical Distributions

**Example: extreme value analysis and annual temperature maxima**

Given two GCM-integrations, one with 1XCO2 (control run) and the other with 2XCO2, what is the change induced in the 10-year return value of 2m temperature by a doubling of CO2?

<table>
<thead>
<tr>
<th>Data:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different from the previous examples, the data are required on two time scales</td>
</tr>
<tr>
<td>- daily temperature with each year</td>
</tr>
<tr>
<td>- the maximum of the 365 observations gives one realization of the random variable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual maxima of daily temperature are random variables with a Gumble distribution</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimating parameters from data:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- several methods (the method of moments, the method of maximum likelihood) are available</td>
</tr>
<tr>
<td>- the methods may produce quite different results with a small sample of extremes</td>
</tr>
</tbody>
</table>
1.1 Theoretical Distributions

Estimated 10-year return values of daily maximum temperature at 2 m height estimated from the output of model experiments with a GCM coupled to a mixed-layer ocean model and a sea-ice model. Units: °C

Top: Return values estimated from a 20-year control run
Bottom: Change of the return values as derived from above and the output of a 10-year model experiment with doubled atmospheric CO2 concentrations
1.1 Theoretical Distributions

The $\chi^2$ distribution: $\chi^2(k)$

The $\chi^2$ distribution is defined as that of the sum of $k$ independent squared $N(0,1)$-distributed random variables

$$f_X(x) = \begin{cases} 
\frac{x^{(k-2)/2}e^{-x/2}}{\Gamma(k/2)2^{k/2}} & \text{if } x > 0 \\
0 & \text{otherwise}
\end{cases}$$

- it is defined only on the positive half of the real axis
- it has a single parameter, $k$, referred to as the degrees of freedom (df)
- it is additive: if $X_1$ and $X_2$ are independent $\chi^2$ random variables with $k_1$ and $k_2$ df, then $X_1+X_2$ is a $\chi^2$ random variable with $k_1+k_2$ df
- it is skewed
- it approaches a normal distribution for large df
- it has the mean and variance, $\mu=k$, $\sigma^2=2k$ (both depend on df)
### 1.1 Theoretical Distributions

**The t distribution: \( t(k) \)**

Given \( N(0,1) \) distributed random variable \( A \) and \( \chi^2 \) distributed random variable \( B \), the variable \( T \) defined by

\[
T = \frac{A}{\sqrt{B/k}}
\]

is \( t \) distributed with the probability density function

\[
f_T = \frac{\Gamma((k+1)/2)(1+t^{2/k})^{-(k+1)/2}}{\sqrt{k\pi}\Gamma(k/2)}
\]

- the \( t \) distribution is symmetric about zero
- it has the mean and variance

\[
\mu = 0 \quad \text{for } k \geq 2
\]

\[
\sigma^2 = \frac{k}{k-2} \quad \text{for } k \geq 3
\]

\( \mu \) does not exist for \( k = 1 \), \( \sigma^2 \) does not exist for \( k = 1,2 \)
1.1 Theoretical Distributions

**The F distribution: F(k,l)**

If X and Y are independent \( \chi^2 \) distributed random variables, then is the random variable

\[
\frac{X/k}{Y/l}
\]

F distributed with k and l degrees of freedom. The probability density function is given by

\[
f_F = \frac{(k/l)^{k/2} \Gamma((k+l)/2)}{\Gamma(k/2)\Gamma(l/2)} x^{(k-2)/2} \left(1 + \frac{k}{l} x\right)^{-(k+l)/2}
\]

- The first two central moments are

\[
\mu = \frac{l}{l-2} \quad \text{for } l > 2
\]

\[
\sigma^2 = \frac{2l^2(k+l-2)}{k(l-2)^2(l-4)} \quad \text{for } l > 4
\]

As for the t distribution, not all moments of the F distribution exist

- skewed for all values of l and k. For a fixed k, the skewness decreases slightly with increasing l
1.1 Theoretical Distributions

**Multivariate Normal Distribution**

\[ f_{\tilde{x}}(\tilde{x}) = \frac{1}{(2\pi |\Sigma|)^{1/2}} \exp\left(-\frac{(\tilde{x} - \bar{\mu})^T \Sigma^{-1} (\tilde{x} - \bar{\mu})}{2}\right) \]

- symmetric across all planes which past through the mean
- the spread is determined by the covariance matrix \( \Sigma \)
- linear combinations of normal random variables are again normally distributed, in particular

  let \( A \) be a full rank \( m' \times m \) matrix of constants with \( m' < m \)

  then \( \tilde{Y} = A\tilde{X} \) is a \( m' \)-dimensional normal distributed random vector with mean vector \( \tilde{\mu} \) and covariance matrix \( E \):

  \[ \tilde{\mu} = A\bar{\mu}, \quad E = A\Sigma A^T \]

- All marginal distributions of a multivariate normal distribution are also normal
1.1 Theoretical Distributions

A bivariate normal probability density function with variances $\sigma_1^2 = \sigma_2^2 = 1$ and covariances $\sigma_{12} = \sigma_{21} = 0.5$

Top: Contours of constant density
Bottom three-dimensional representation
### 1.1 Theoretical Distributions

**Non-parametric approaches**

For the examples considered, one may suggest to directly estimate the distributions, rather than the parameters of the distributions, i.e. to use non-parametric approaches.

<table>
<thead>
<tr>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>- there is no need to make specific distribution assumptions</td>
</tr>
<tr>
<td>- the non-parametric approaches are often only slightly less efficient than methods that use correct parametric model, and generally more efficient compared with methods that use the incorrect parametric model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>One may face the problem of overfitting (i.e. the number of parameters is, relative to the sample size, too large). In the distribution examples, the ‘parameters’ to be estimated in an non-parametric approach are values of $f(x)$ for all possible $x$.</td>
</tr>
</tbody>
</table>
1.1 Theoretical Distributions

An example demonstrating the danger of overfitting

The table shows the total winter snowfall at Ithaca in inches for the seven winters beginning 1980 through 1986, and four potential ‘predictors’ arbitrarily taken from an almanac:

1. federal deficit (in billions of dollars)
2. the number of personal in the US Air Force
3. the number of sheep in the US (in thousands)
4. the average Scholastic Aptitude Test (SAT) scores

<table>
<thead>
<tr>
<th>Winter beginning</th>
<th>Ithaca snowfall (inches)</th>
<th>U.S. federal deficit ($ \times 10^9$)</th>
<th>U.S. Air Force personnel</th>
<th>U.S. sheep (\times 10^3)</th>
<th>Average SAT scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>52.3</td>
<td>59.6</td>
<td>557,969</td>
<td>12,699</td>
<td>992</td>
</tr>
<tr>
<td>1981</td>
<td>64.9</td>
<td>57.9</td>
<td>570,302</td>
<td>12,947</td>
<td>994</td>
</tr>
<tr>
<td>1982</td>
<td>50.2</td>
<td>110.6</td>
<td>582,845</td>
<td>12,997</td>
<td>989</td>
</tr>
<tr>
<td>1983</td>
<td>74.2</td>
<td>196.4</td>
<td>592,044</td>
<td>12,140</td>
<td>963</td>
</tr>
<tr>
<td>1984</td>
<td>49.5</td>
<td>175.3</td>
<td>597,125</td>
<td>11,487</td>
<td>965</td>
</tr>
<tr>
<td>1985</td>
<td>64.7</td>
<td>211.9</td>
<td>601,515</td>
<td>10,443</td>
<td>977</td>
</tr>
<tr>
<td>1986</td>
<td>65.6</td>
<td>220.7</td>
<td>606,500</td>
<td>9,932</td>
<td>1001</td>
</tr>
</tbody>
</table>

A regression model with 6 parameters

\[
\text{SNOW} = a_0 + a_1(\text{year}) + a_2(\text{deficit}) + a_3(\text{AFpers}) + a_4(\text{sheep}) + a_5(\text{SAT})
\]

is fitted into the data with sample size being 6

The figure shows the snowfall amounts given by the model (line) and observations (circles). One has a perfect description of data, but a disastrous prediction for any independent data.