Empirical data analysis in accounting and finance

Advanced Financial Accounting II, School of Business and Economics at Åbo Akademi University

Some examples on quantitative empirical research problems in accounting and finance

- Pattern recognition
- Financial classification models
  - Financial distress prediction
- Causality models
  - Association between accounting data and financial market reactions
  - Causality patterns on international financial markets
- Time series modelling and prediction
- Optimization models, e.g.
  - Portfolio optimization
  - Product mix optimization

A typical process for empirical data analysis

- Define the test problem
- Collect data
  - Data bases for financial data, e.g. market data, financial statements, interest rates, exchange rates
  - Surveys for opinion data
  - To access data from the voitto database, map the network drive z:\\cdrom\voitto with your user name: AKADEMI\user_name
- Select the analysis method
- Control for the suitability of the data to the selected method
  - Different methods have different assumptions on the properties of the data, e.g. approximate normality

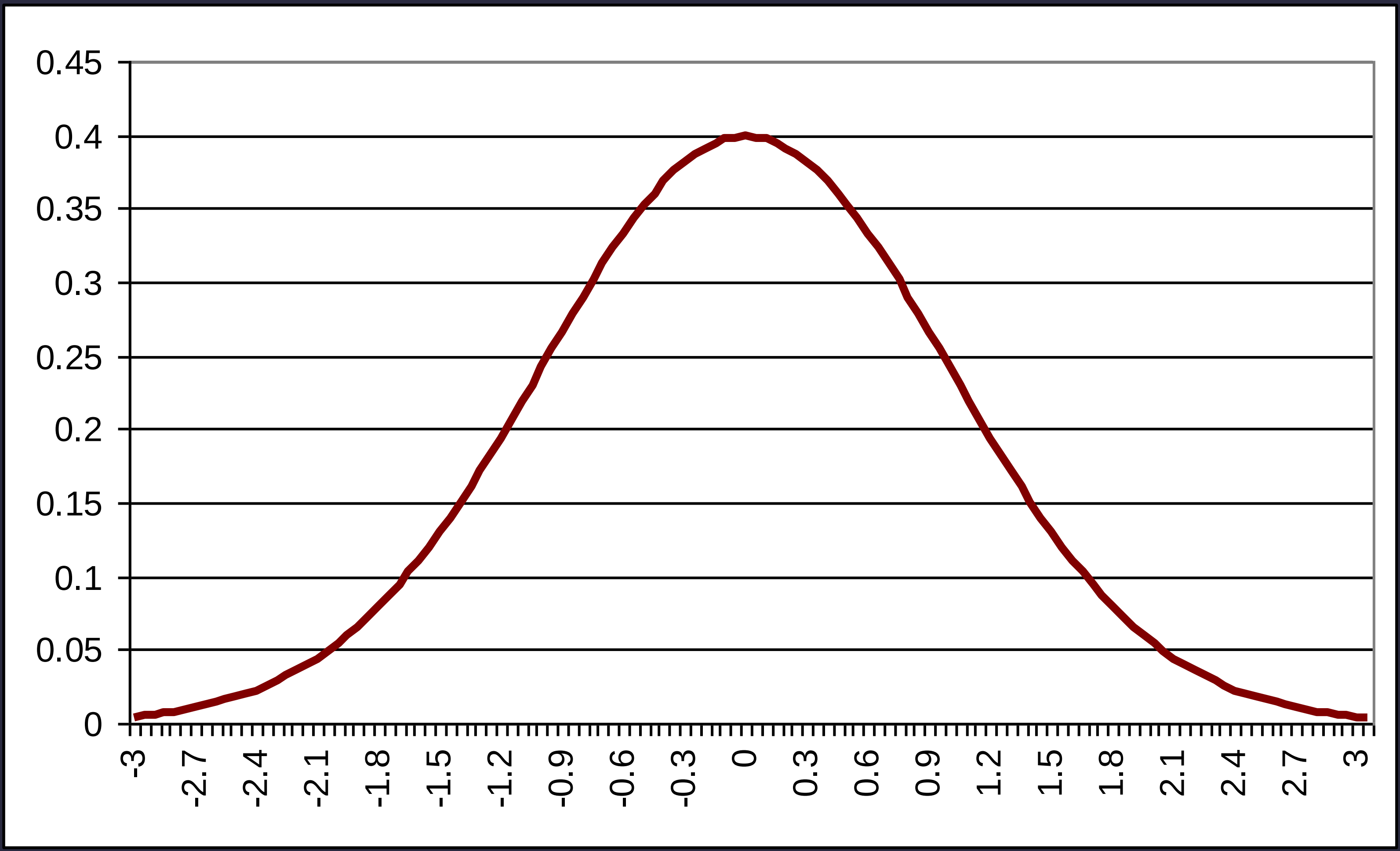
A typical process for empirical data analysis ...

- If necessary, improve the quality of the data
  - Different transformations
    - Taking logarithms of the data
    - Differencing (for time series data)
  - Removing the outliers
- Perform the data analysis by a suitable statistical program, e.g. SPSS, SAS
- Interpret the results critically

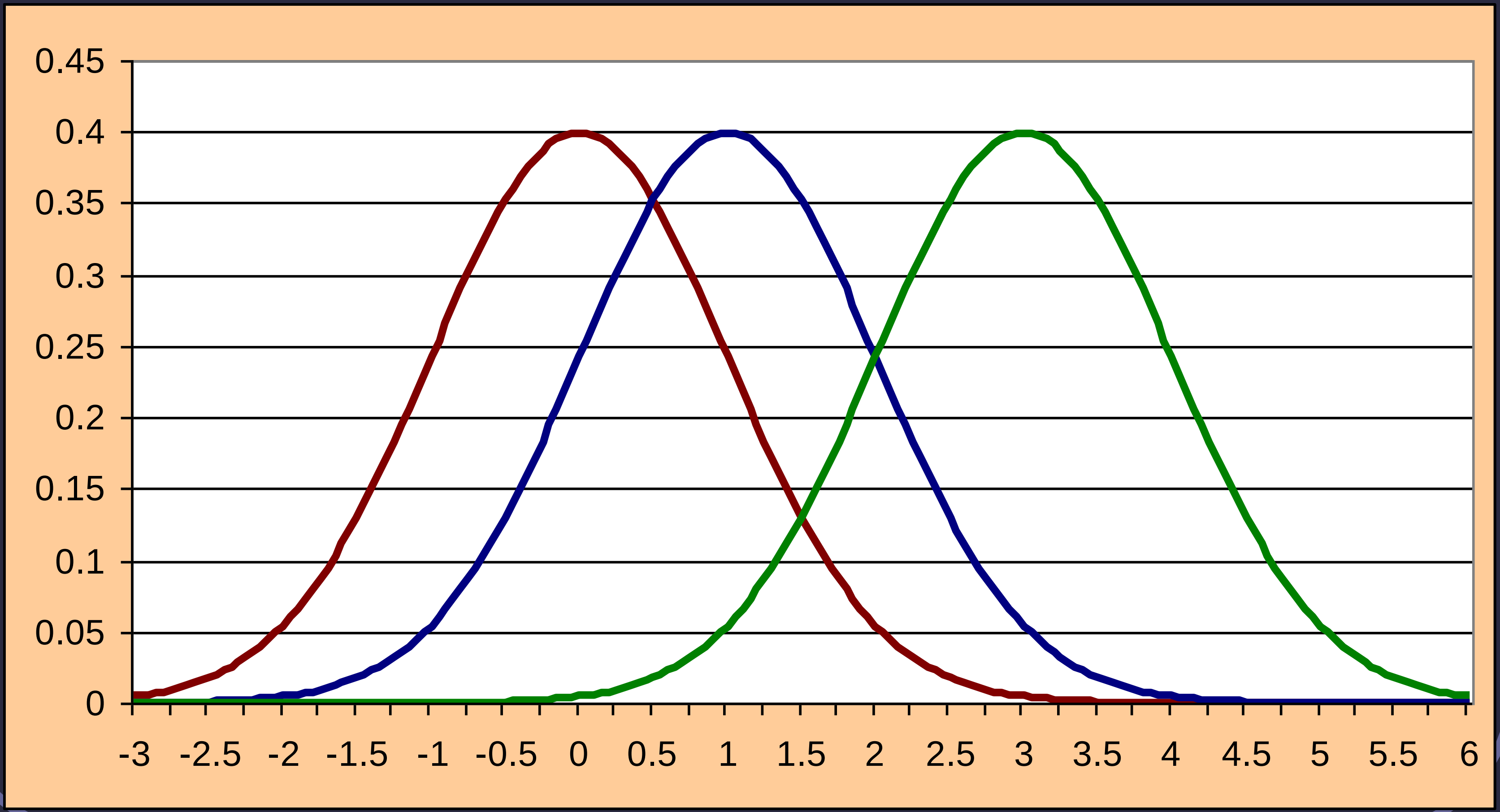
Probability distributions

- Practically all statistical methods assume the data to follow some predefined distribution (**probability density function** or **pdf**), e.g.
  - Standard normal distribution
  - Normal distribution
- Also the test statistics follow predefined distributions, e.g.
  - χ^2-distribution
  - F-distribution
  - t-distribution

The standard normal (Gaussian) distribution ($\mu = 0$, $\sigma = 1$)

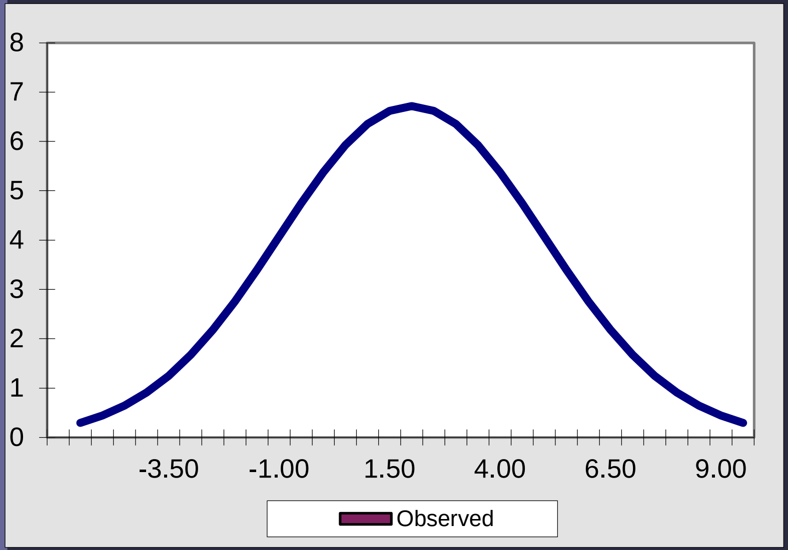


Normal distributions with different mean values ($\mu \in \{0;1;3\}$, $\sigma = 1$)

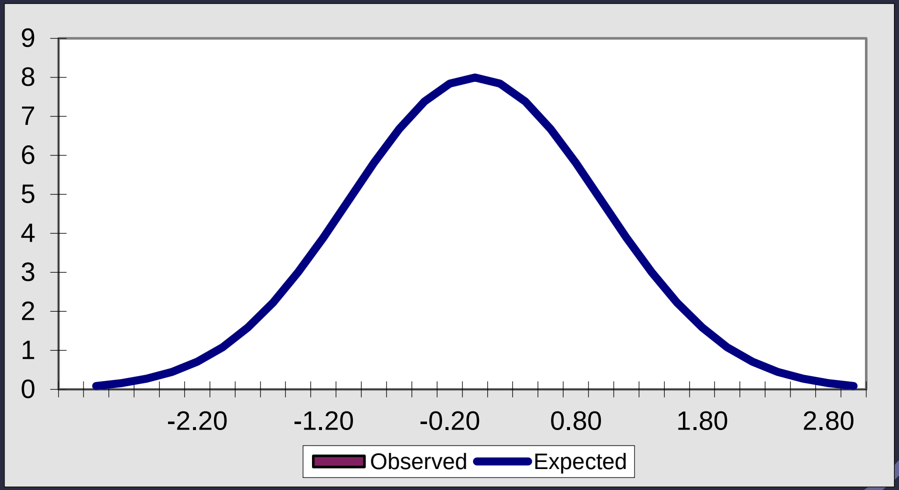


Normal distributions with different mean and standard deviation

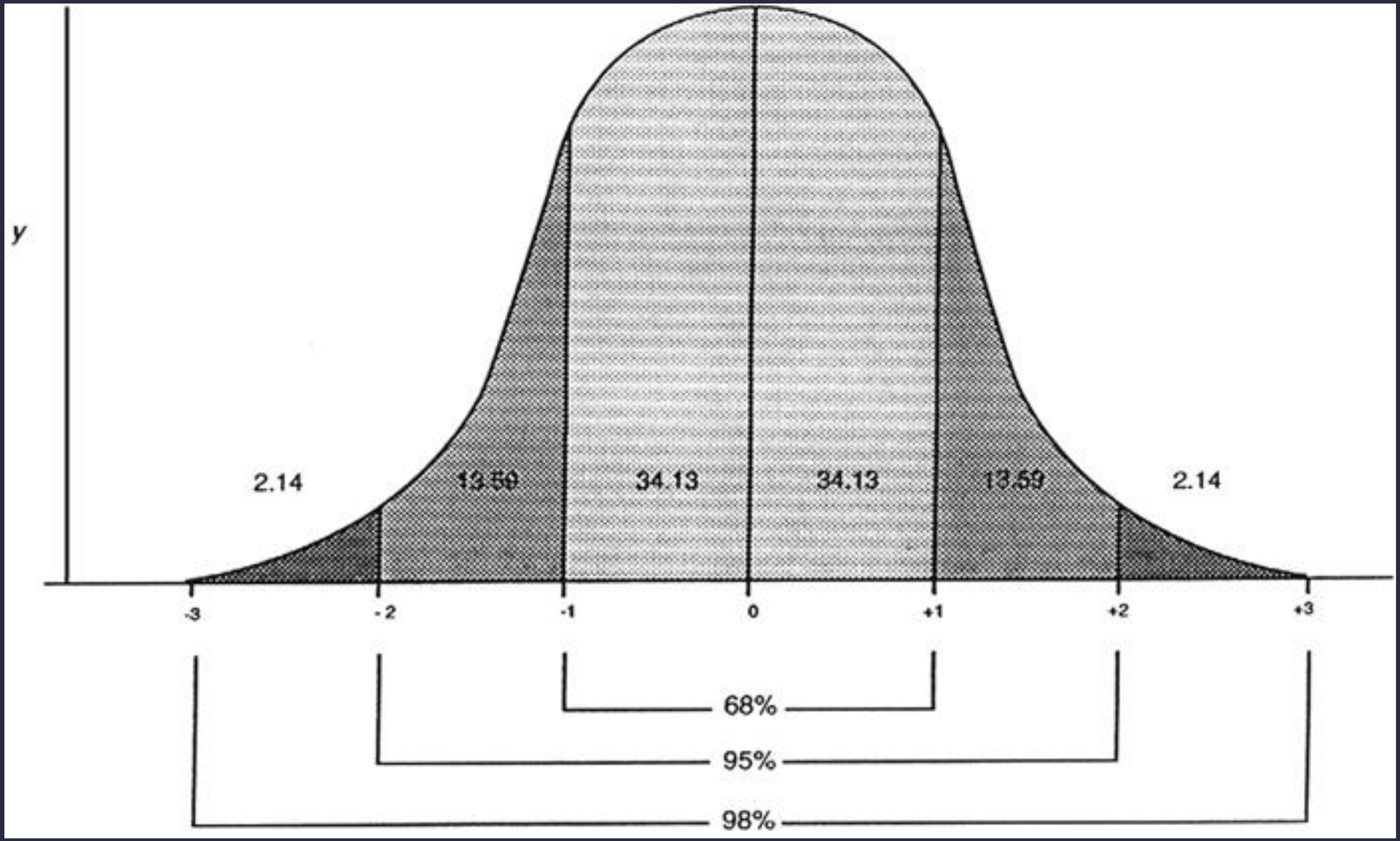
mean = 2, st.dev = 3



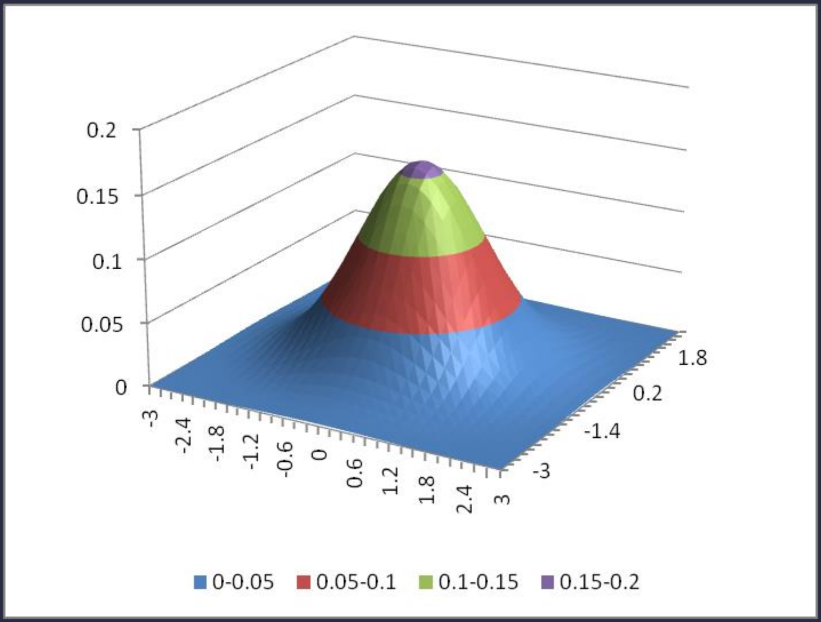
mean = 0, st.dev = 1

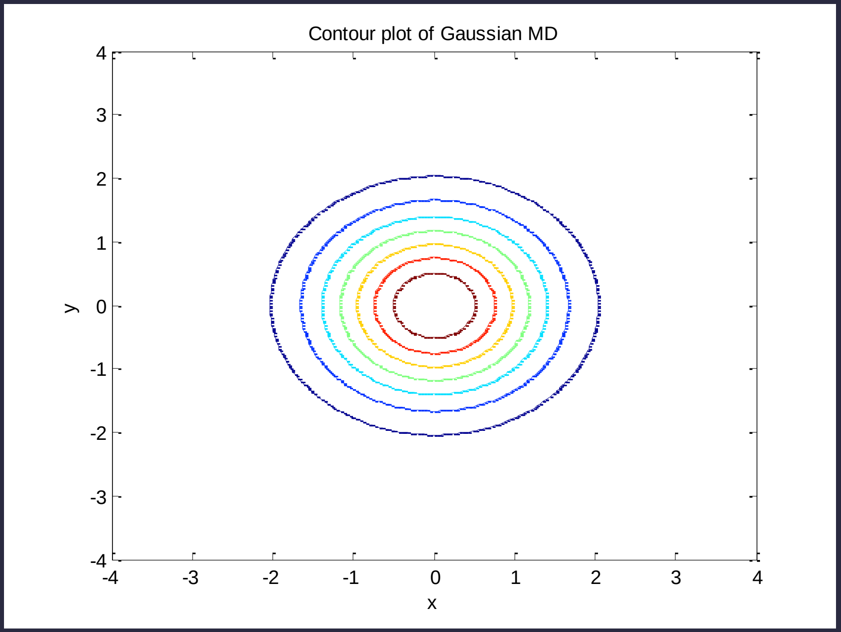


Percentage of observations within {1;2;3} standard deviations from the mean

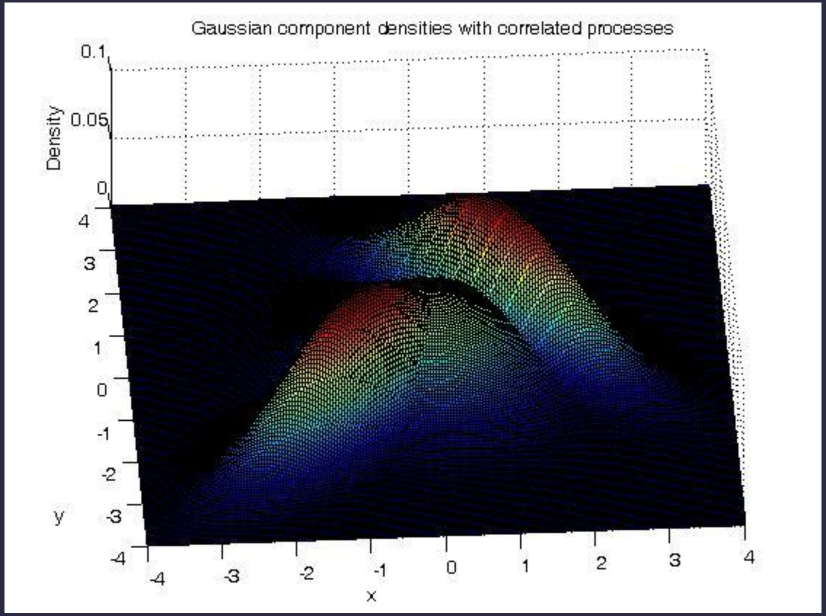


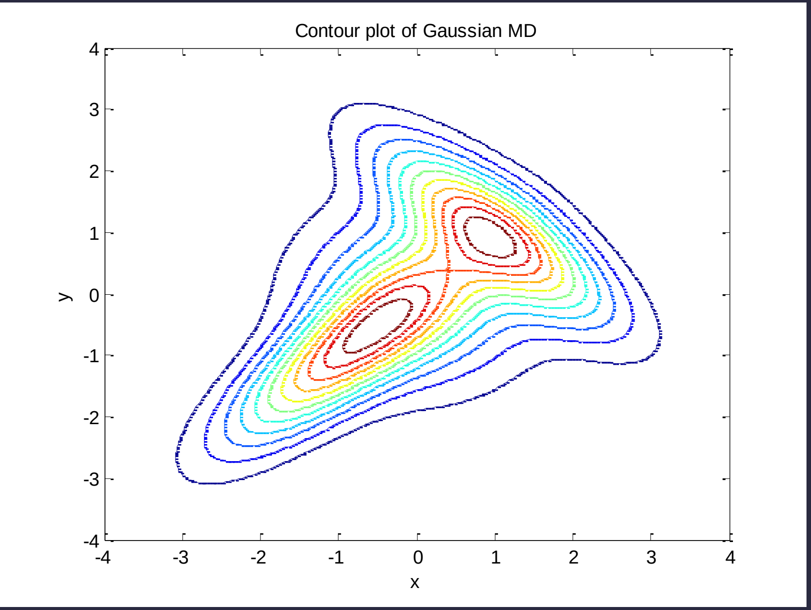
The multivariate normal distribution





The multivariate Gaussian mixture density with correlated processes



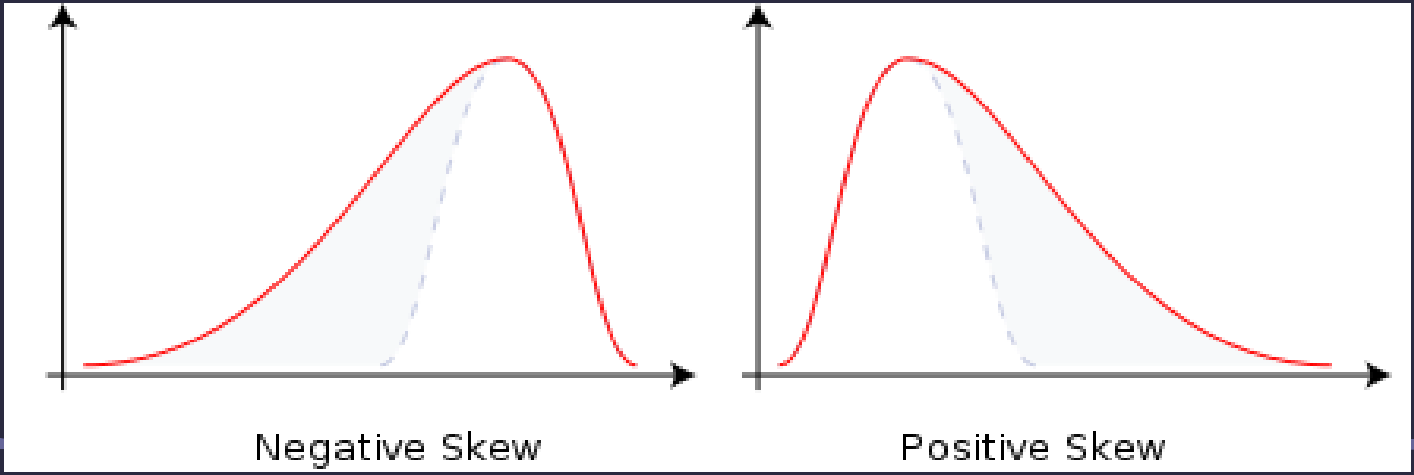


Problems with empirical data

- Most empirical data sets fail to satisfy the basic assumption on normal distribution
- Typical problems are
  - Skewness of the data
  - Leptokurtosis (thick tails)
  - Outliers
- The quality of the data may be improved for example by
  - Different transformations
    - Taking logarithms of the data
    - Taking square roots of the data
    - Differencing (for time series data)
  - Removing the outliers

Skewness

- A measure of the **asymmetry** of the probability distribution
- Skewness = 0 for a symmetric distribution
- **Negative skewness (left skewed pdf)**: The left tail is longer; the mass of the distribution is concentrated on the right of the figure
- **Positive skewness (right-skewed pdf)**: The right tail is longer; the mass of the distribution is concentrated on the left of the figure



Skewness…

- Statistically, skewness is defined as

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{E\left[(X - \mu)^3\right]}{E[(X - \mu)^2]^{3/2}}$$

- and for a sample of *n* values the *sample skewness* is

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^{3/2}}$$

Kurtosis

- A measure of the **peakedness** of the probability distribution
- A high kurtosis distribution has a sharper peak and longer, fatter tails
- A low kurtosis distribution has a more rounded peak and shorter, thinner tails
- Statistically, kurtosis is defined as

$$\gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{E\left[(X - \mu)^4\right]}{E\left[(X - \mu)^2\right]^2}$$

Excess kurtosis

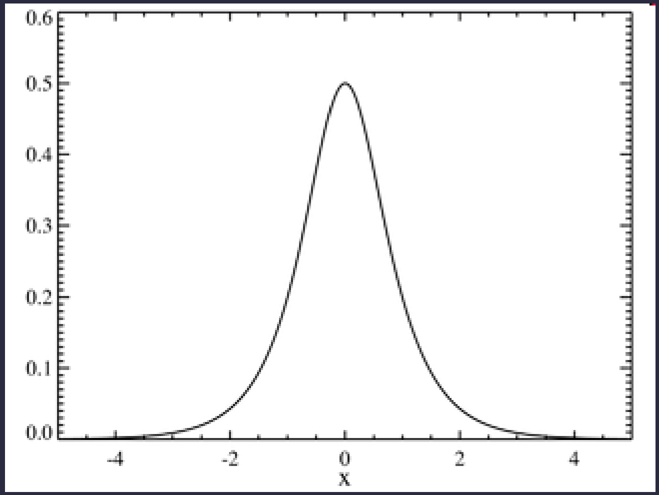
- Kurtosis for a normal distribution equals 3
- In order to have kurtosis of the normal distribution zero, kurtosis is mostly defined as excess kurtosis (also in statistical packages like SPSS or Excel)

$$\gamma_2^{\varepsilon} = \frac{\mu_4}{\sigma^4} - 3$$

- Distributions with zero excess kurtosis, e.g. normal distributions are called **mesokurtic**

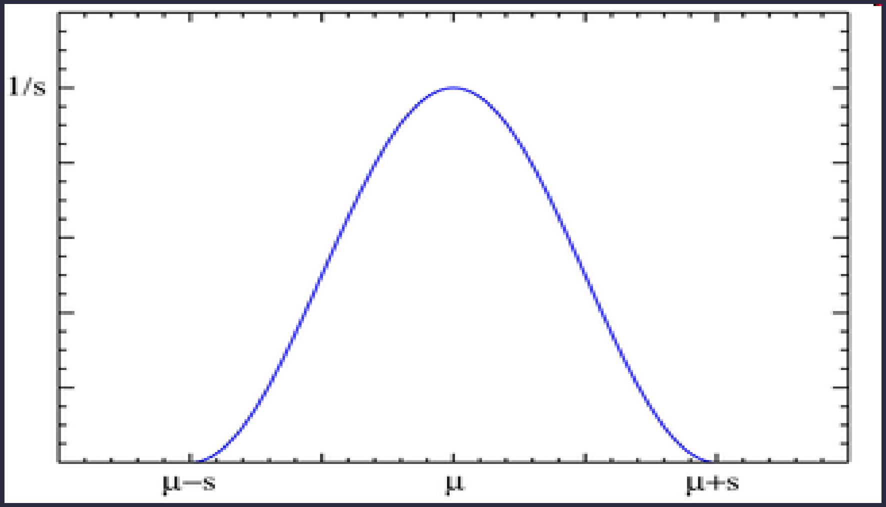
Leptokurtosis

- A distribution with positive excess kurtosis is called leptokurtic
- A leptokurtic distribution has a more acute peak around the mean and fatter tails
- Many financial data series (for example, stock returns) have leptokurtic distributions



Platykurtosis

- A distribution with negative excess kurtosis i called platykurtic
- A platykurtic distribution has a lower, wider peak and thinner tails



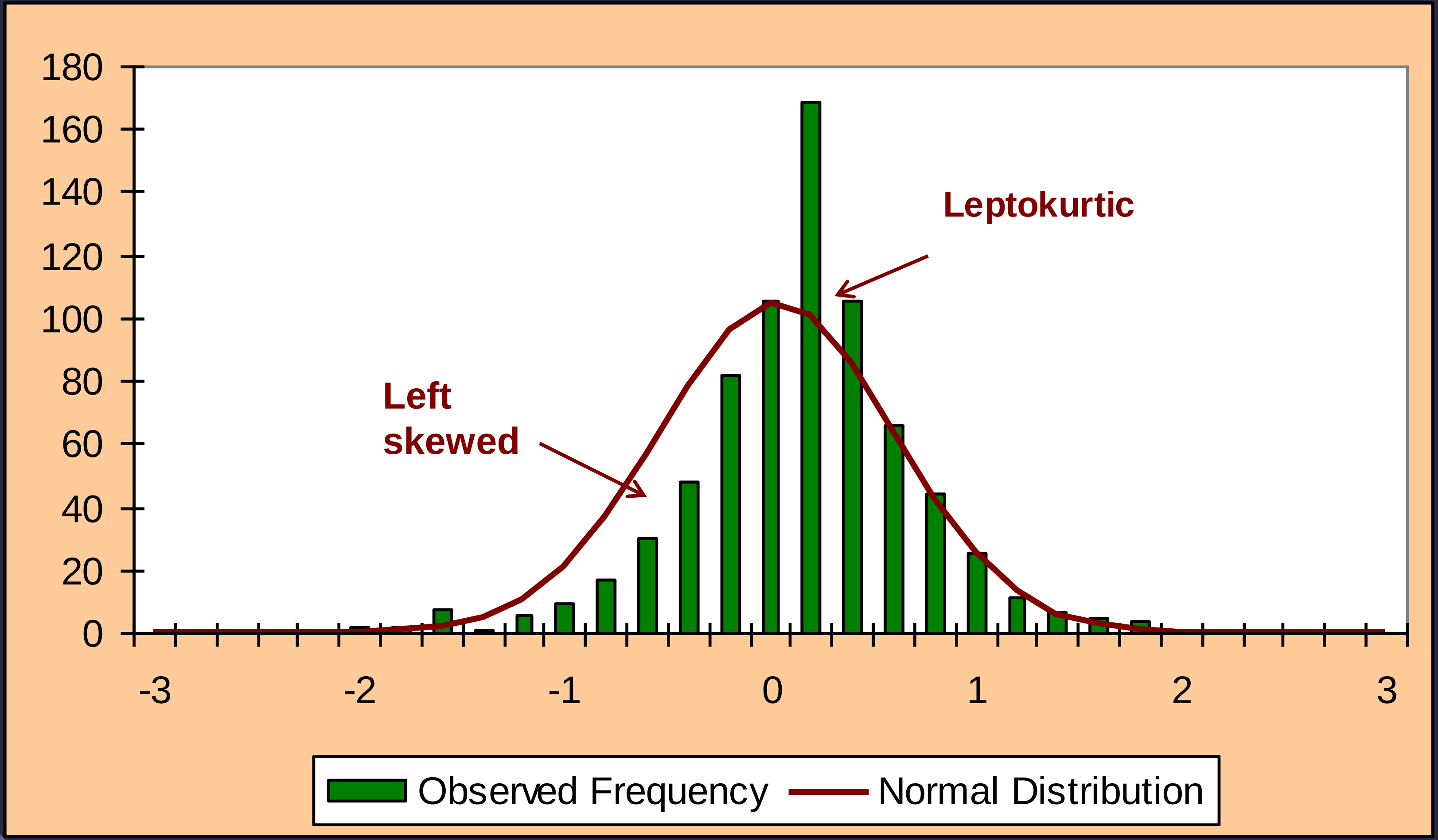
Testing for the normality of a data set

- There are several tests for measuring the normality of a data set, for example
  - Kolmogorov-Smirnov test (Massey, 1951)
  - Pearson’s chi-square test (Pearson, 1900)
  - Jarque-Bera test (Jarque & Bera, 1980)
  - Shapiro-Wilk test (Shapiro & Wilk, 1965)
  - Cramér-von-Mises test (Anderson, 1962)

Testing for the normality of a data set - an example

- Test data set: Daily returns on the Canadian stock market over the period 2.2.1994 - 31.12.1996 (752 observations)
- Descriptive statistics
  - Mean: 0.034
  - St. Dev.: 0.573
  - **Skewness: -0.617**
  - **(Excess) Kurtosis: 2.777**
  - Minimum: -2.807
  - Maximum: 2.017

Histogram of the Canadian stock market returns and a normal distribution with the observed mean and standard deviation



Kolmogorov-Smirnov –test of normality

- The One-Sample Kolmogorov-Smirnov Test procedure compares the observed **cumulative distribution function** (**cdf**) for a variable with a specified theoretical distribution
- The test statistic, Kolmogorov-Smirnov Z, is computed from the largest difference (in absolute value) between the observed and theoretical cdf:s
- In SPSS:
  - Analyze

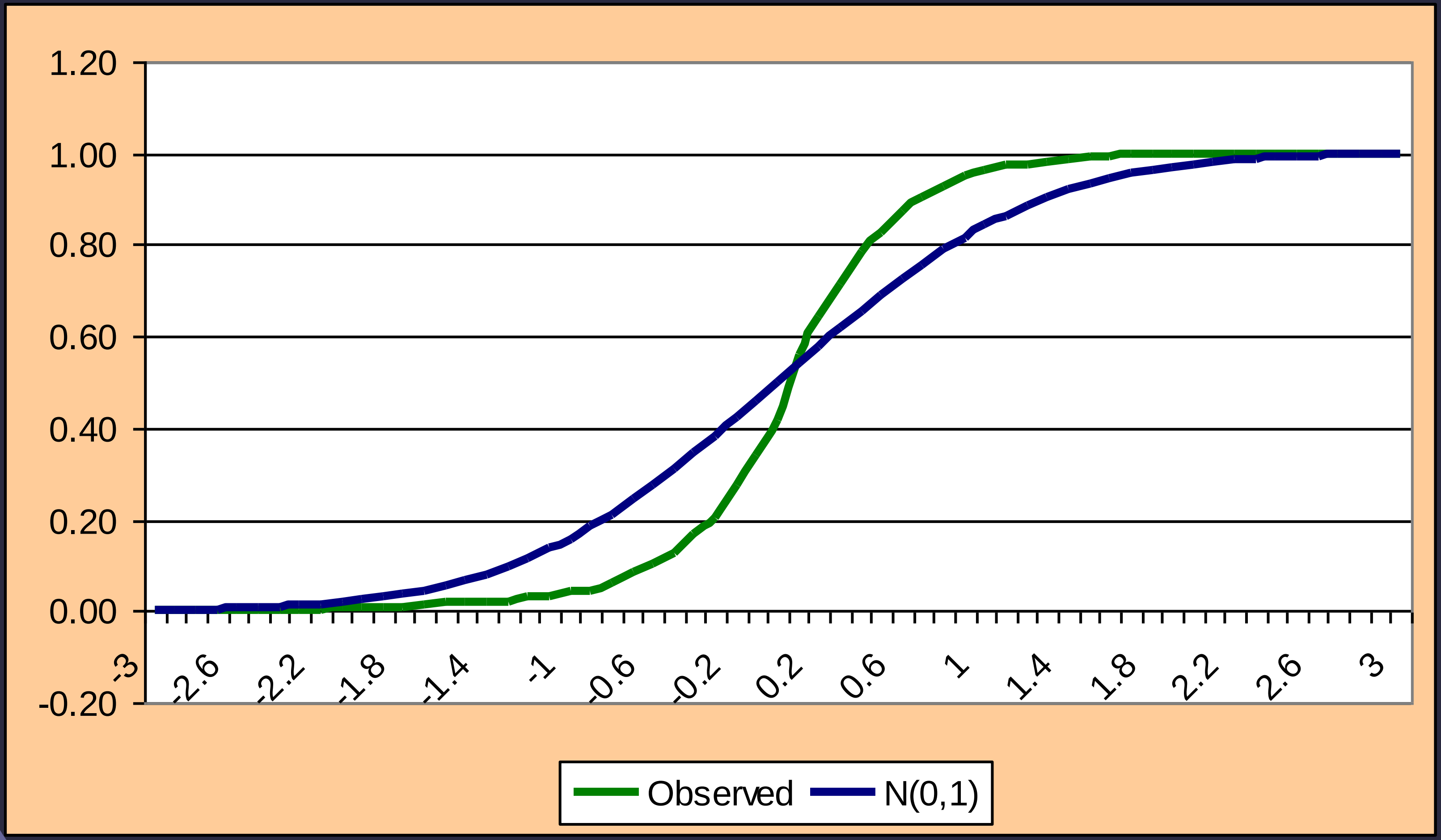
    ⇒ Nonparametric Tests

      ⇒ Legacy Dialogs

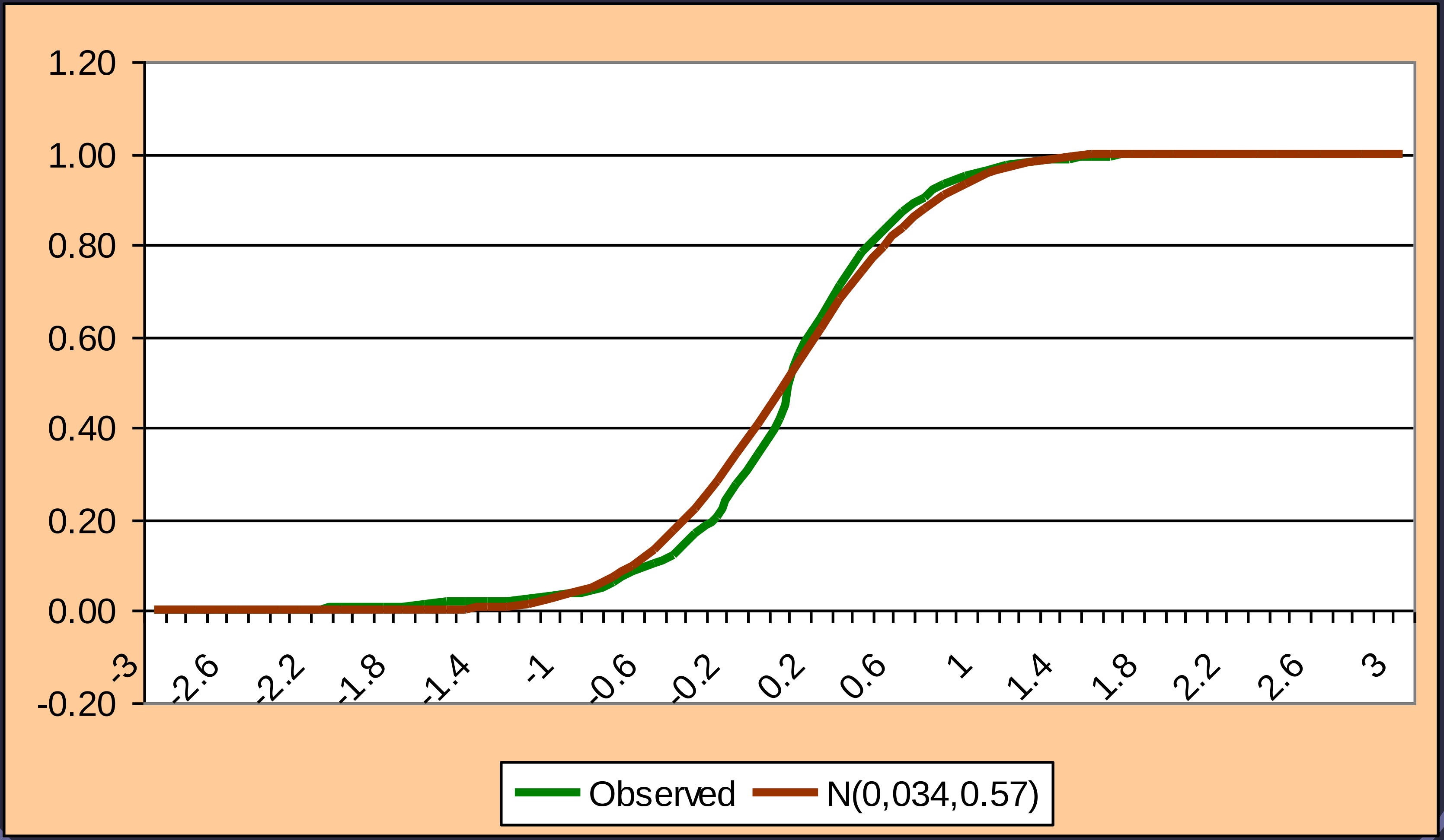
        ⇒ 1-Sample K-S

          ⇒ Test distribution: Normal

The observed Canadian cdf plotted against the standard normal cumulative distribution (Note: SPSS/K-S compares only the observed and corresponding normal distribution)



The observed cdf of the Canadian data plotted against the normal cdf with the observed μ = 0.034 and σ = 0.573 (Tested by SPSS/K-S)



The Kolmogorov-Smirnov statistic

- The empirical distribution function F_n for n iid observations X_i is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}$$

where $I_{X_i \leq x}$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise.

- The Kolmogorov–Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x | F_n(x) - F(x) |$$

where sup x is the supremum of the set of distances.

Kolmogorov-Smirnov test

- The goodness-of-fit test or the Kolmogorov–Smirnov test is constructed by using the critical values of the **Kolmogorov distribution**
- The null hypothesis is rejected at level α if

$$Z = \sqrt{n}\, D_n > K_\alpha,$$

where K_α is found from

$$\Pr(K \leq K_\alpha) = 1 - \alpha$$

- If F is continuous then under the null hypothesis $\sqrt{n}\, D_n$ converges to the Kolmogorov distribution, which does not depend on F

Kolmogorov distribution

- The Kolmogorov distribution is the distribution of the random variable

$$K = \sup_{t \in [0,1]} |B(t)|$$

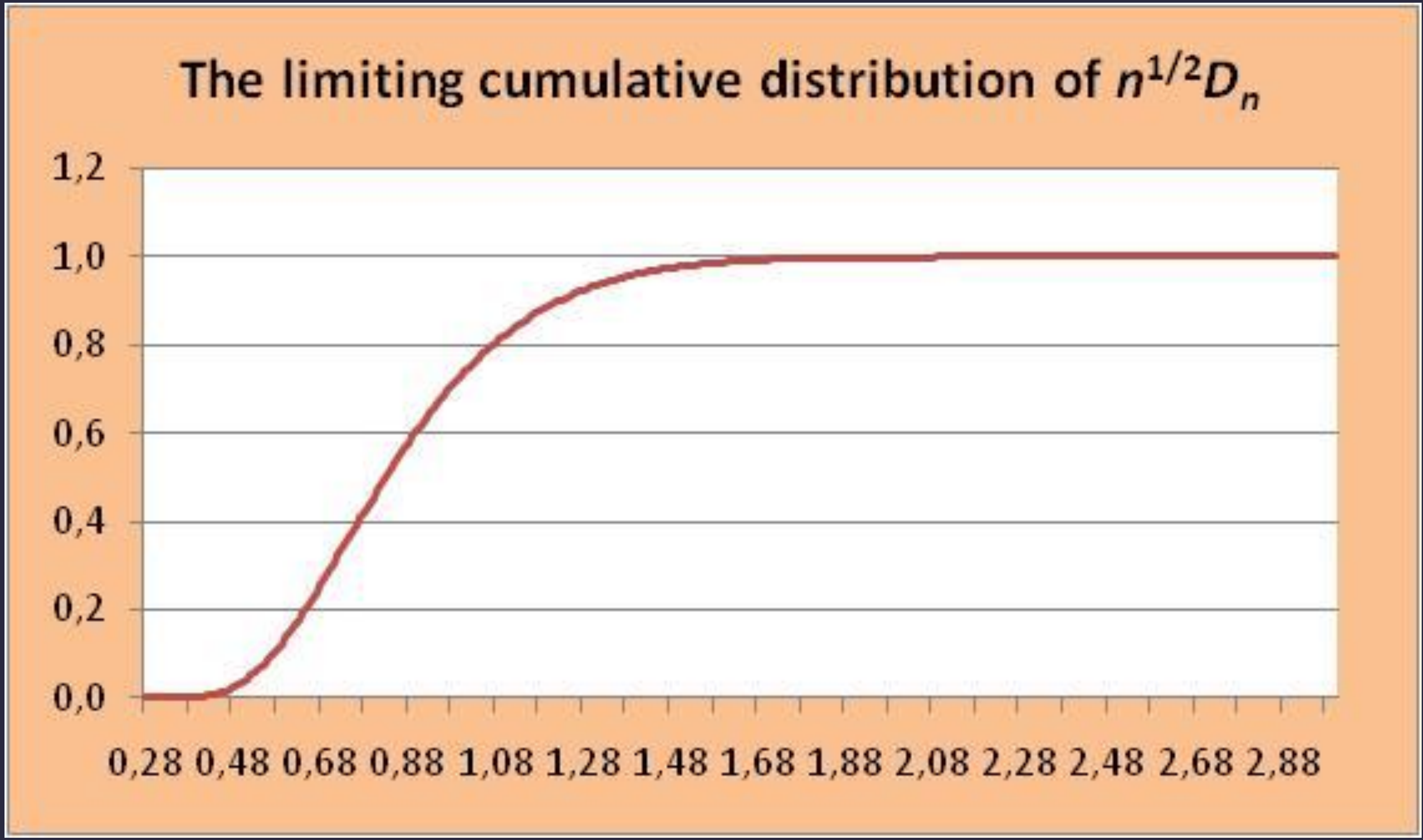
where *B*(*t*) is the Brownian bridge

- The cumulative distribution function of *K* is given by

$$Pr(K \le z) = 1 - 2\sum_{\nu=1}^{\infty}(-1)^{\nu-1} e^{-\nu^2 z^2} = \frac{\sqrt{2\pi}}{z}\sum_{\nu=1}^{\infty} e^{-(2\nu-1)^2 \pi^2/(8z^2)}$$

- A table of the distribution was published by N.V. Smirnov (1948)

The cumulative Kolmogorov distribution



SPSS-output for the K-S test with the Canadian data

One-Sample Kolmogorov-Smirnov Test

		Can
	N	752
Normal Parameters[a,,b]	Mean	,0340
	Std. Deviation	,57275
Most Extreme Differences	Absolute	,077
	Positive	,047
	Negative	-,077
	Kolmogorov-Smirnov Z	2,103
	Asymp. Sig. (2-tailed)	**,000**

a. Test distribution is Normal.

b. Calculated from data.

Data not normal

($\alpha < 0.05$)

Shapiro-Wilk test

- In order to work reliably, the Kolmogorov-Smirnov test requires a relatively large data sample (> 2 000 observations)
- For sample sizes less than 2 000, Shapiro-Wilk test should be applied instead
- The Shapiro-Wilk test statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where

- $x_{(i)}$ is the i^{th} order statistic, i.e., the i^{th}-smallest number in the sample
- $\bar{x} = (x_1 + ... + x_n) / n$ is the sample mean

Shapiro-Wilk test

- The constants a_i are given by
  $$(a_1,\dots,a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$
  where
  $$m = (m_1,\dots,m_n)^T,$$
  $m_1,\dots,m_n$ are the expected values of the order statistics of independent and identically-distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.
- The user may reject the null hypothesis if W is too small

Shapiro-Wilk test for the Canadian stock market data with SPSS

- **In SPSS the Shapiro-Wilk test statistic is achieved as a bi-product of Q-Q plots**

  Analyze ⇒ Descriptive Statistics ⇒ Explore

  Put variable into Dependent list ⇒ Plots

  Click Normality with plots

Output:

Tests of Normality

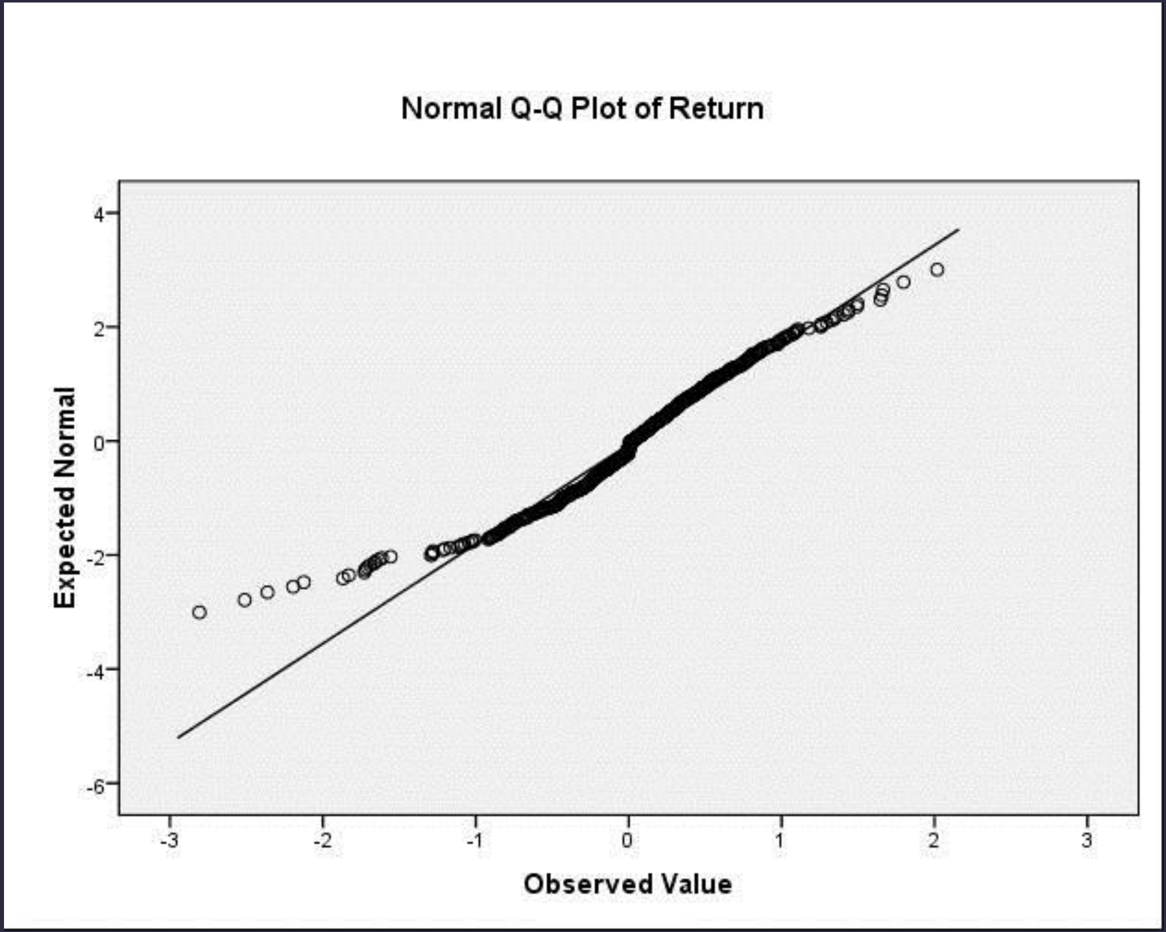
	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
	Statistic	df	Sig.	Statistic	df	Sig.
Return	,077	752	,000	,957	752	**,000**

a. Lilliefors Significance Correction

Data not normal

($\alpha < 0.05$)

Q-Q -plot for the Canadian stock market data



The Jarque-Bera test of normality

- A goodness-of-fit measure of departure from normality, based on the sample kurtosis (γ_2) and skewness (γ_1)

$$JB = \frac{n}{6}\left(\gamma_1^{\,2} + \frac{(\gamma_2 - 3)^2}{4}\right) = \frac{n}{6}\left(\gamma_1^{\,2} + \frac{(\gamma_2^{\varepsilon})^2}{4}\right)$$

- The statistic JB has an asymptotic chi-square (χ^2) distribution with two degrees of freedom and can be used to test the null hypothesis that the data are from a normal distribution (Jarque & Bera, 1980)

JB-statistic for the Canadian data

- For the Canadian stock market data

$$JB = \left(\frac{752}{6}\right)\left((-0{,}617)^2 + \frac{2{,}777^2}{4}\right) = 289.35$$

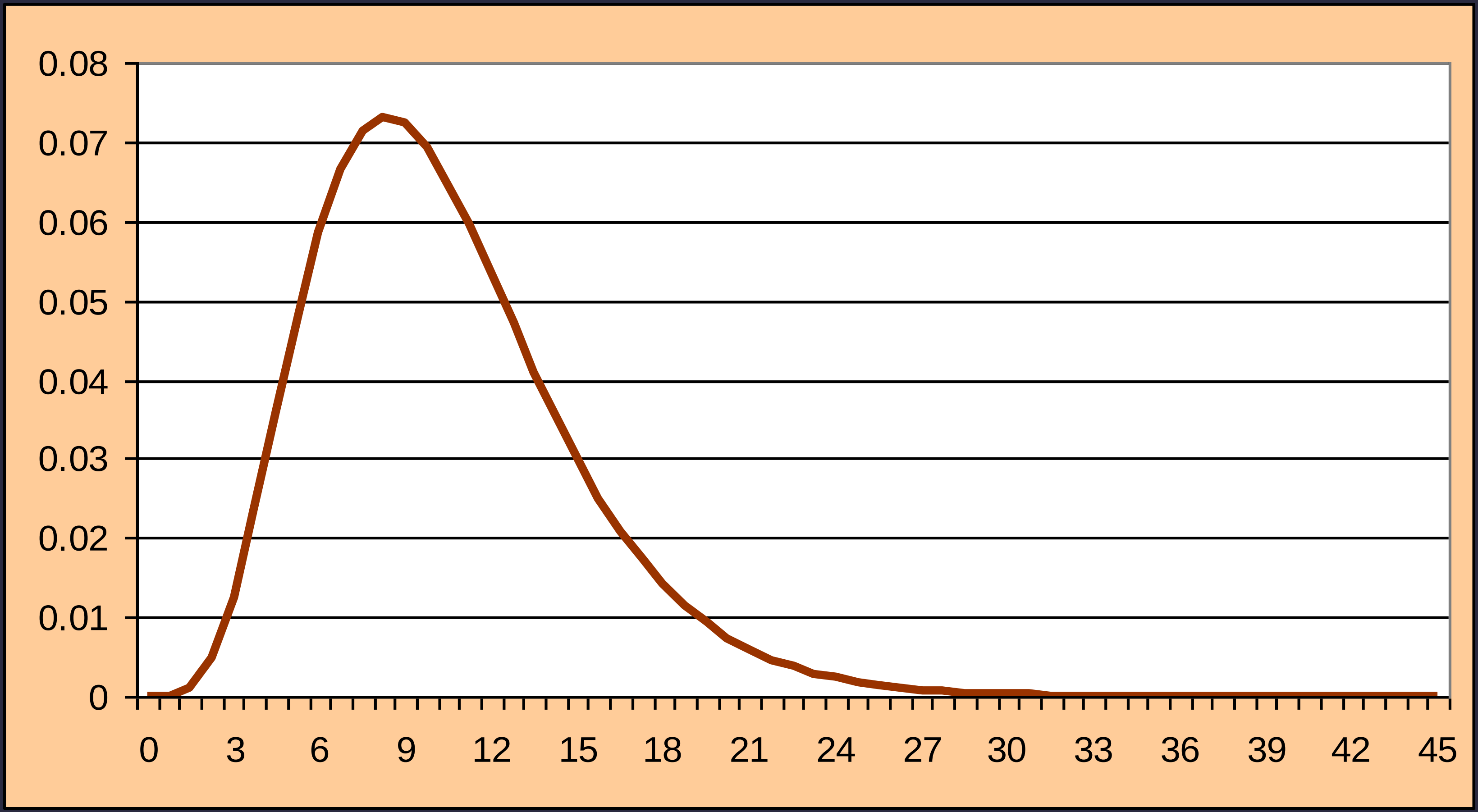
for which p = 0,000, i.e. the data is not normally distributed

- The probability level is achieved by e.g. Excels function CHIDIST() (p = CHIDIST(JB;DOF) = CHIDIST(289.35;2)
- The Jarque-Bera –test is extremely sensitive to deviations from normality

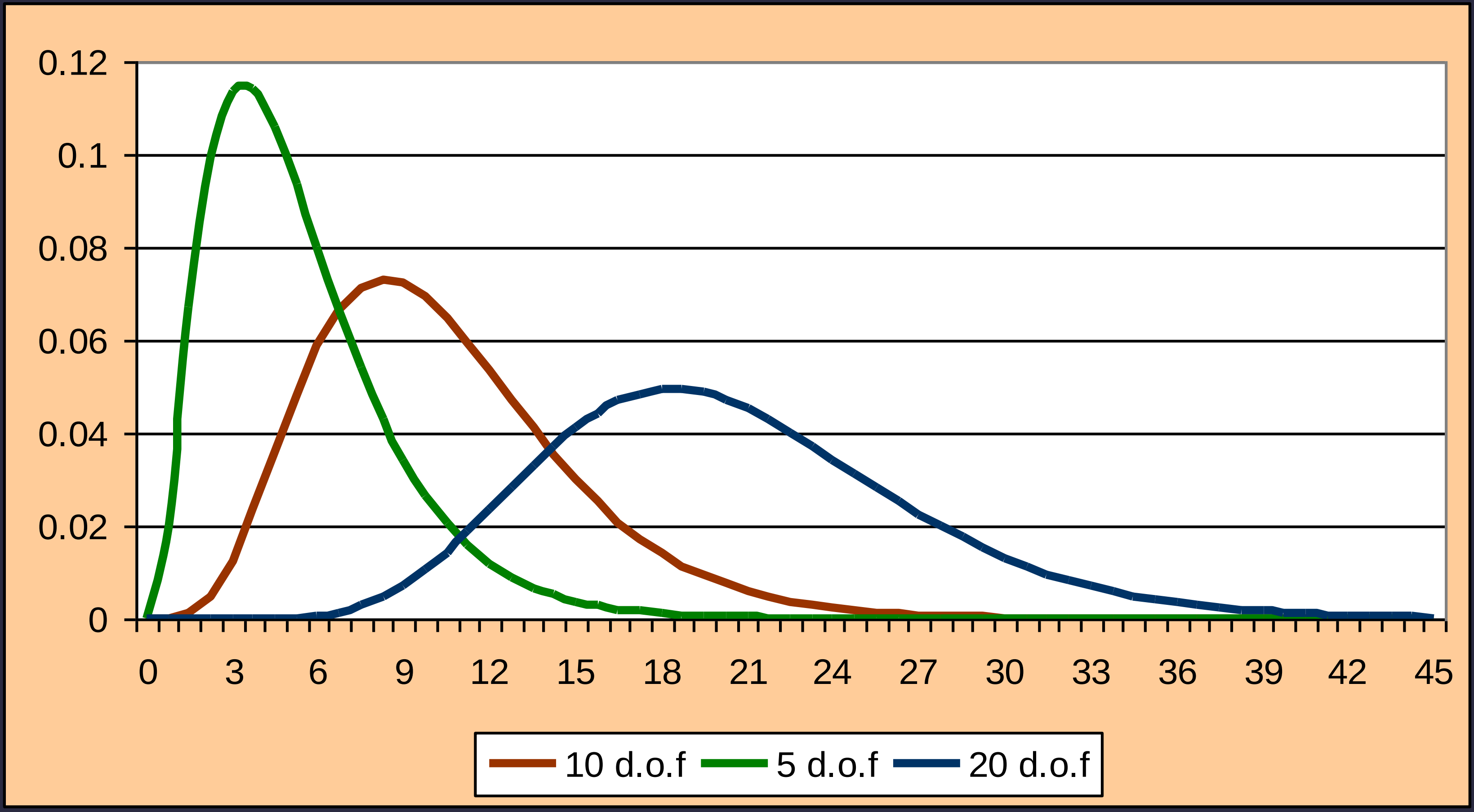
χ^2-distribution

- χ^2-distribution with k **degrees of freedom** (d.o.f), $\chi^2(k)$ is the distribution of a sum of the squares of k independent standard normal distributions
- One of the most widely used probability distributions in inferential statistics, e.g. in hypothesis testing or in construction of confidence intervals
- The best-known situations for use are
  - Tests for goodness of fit of an observed distribution to a theoretical one
  - Tests the independence of two criteria of classification of qualitative data

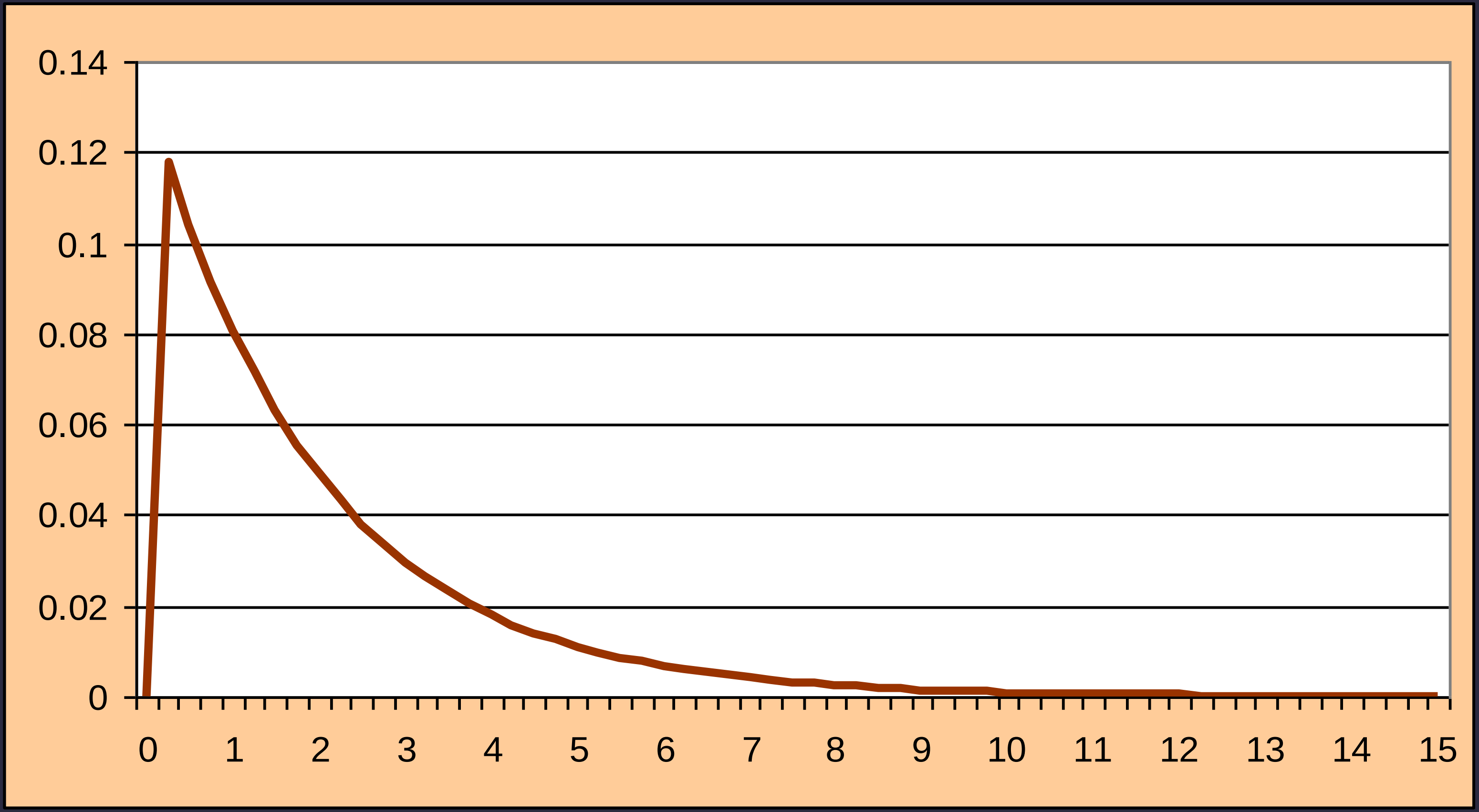
χ^2-distribution with 10 degrees of freedom



χ^2-distributions with {5;10;20} degrees of freedom



χ^2-distribution with 2 degrees of freedom (e.g. Jarque-Bera test)



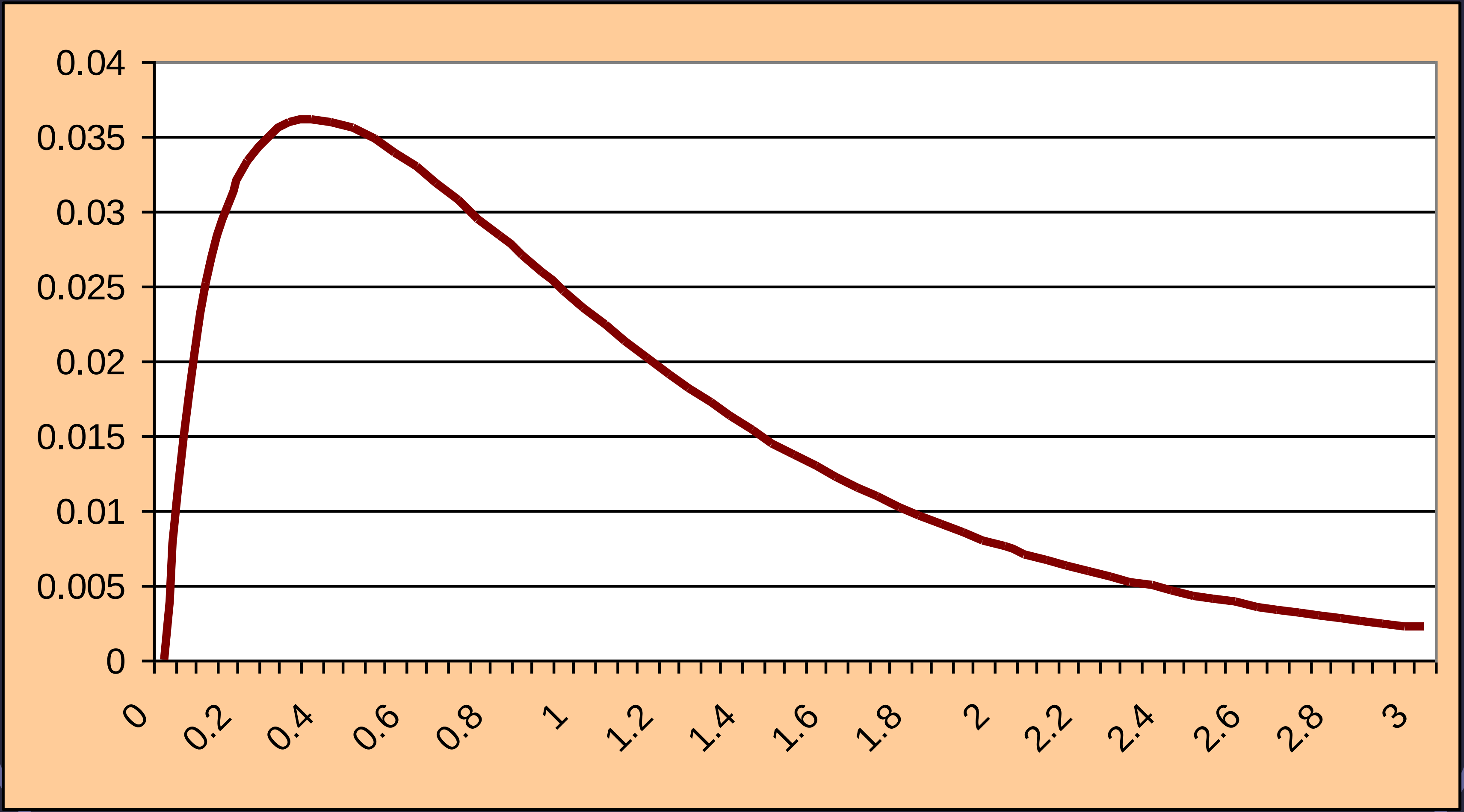
F-distribution

- A ratio of two χ^2-distributions
- Characterized by two d.o.f. –parameters

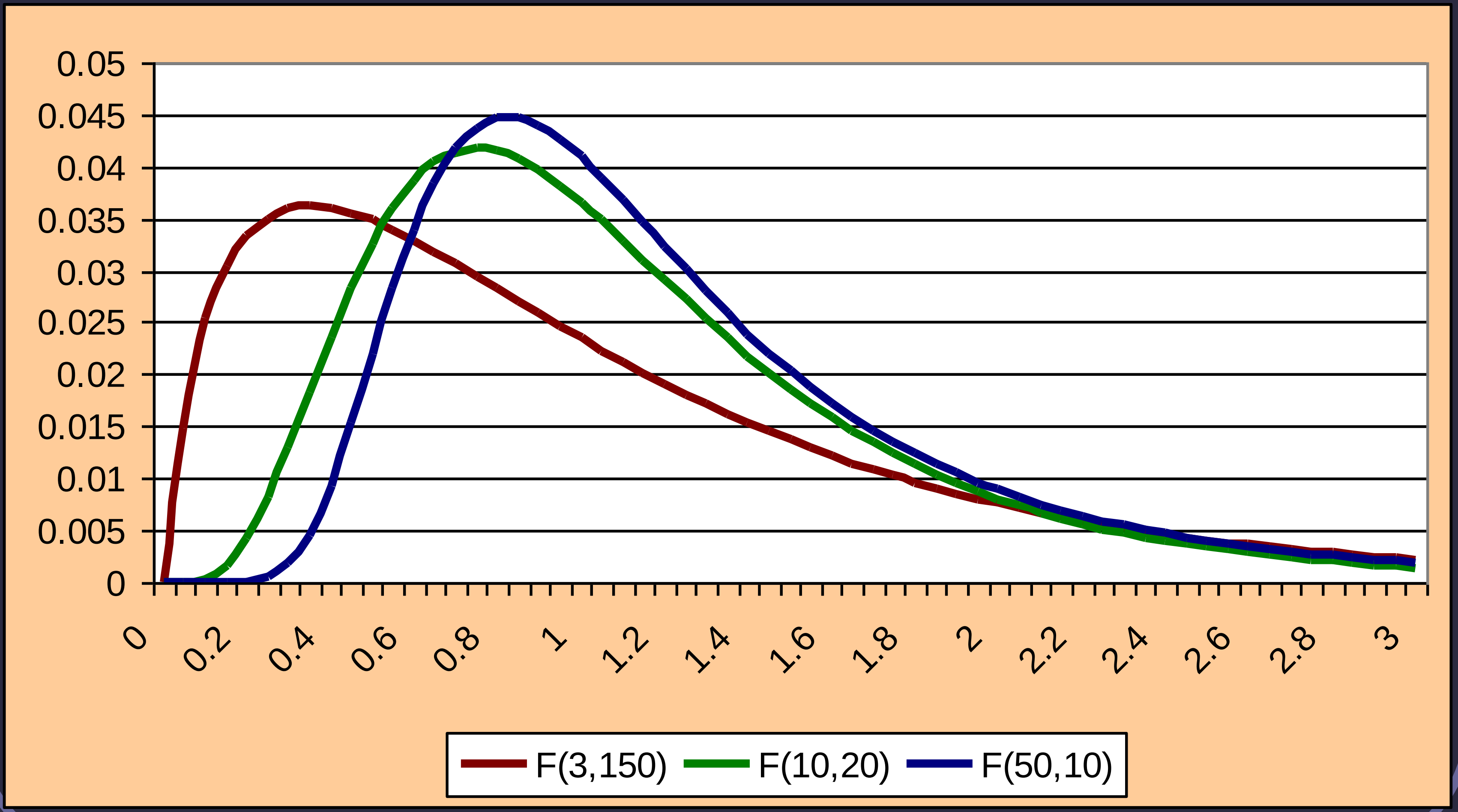
$$F\ (d_1>0,\ d_2>0)$$

- Arises as the null distribution of test statistics
  - Likelihood ratio tests
  - Analysis of variance (ANOVA)

F-distribution with 3 and 150 degrees of freedom



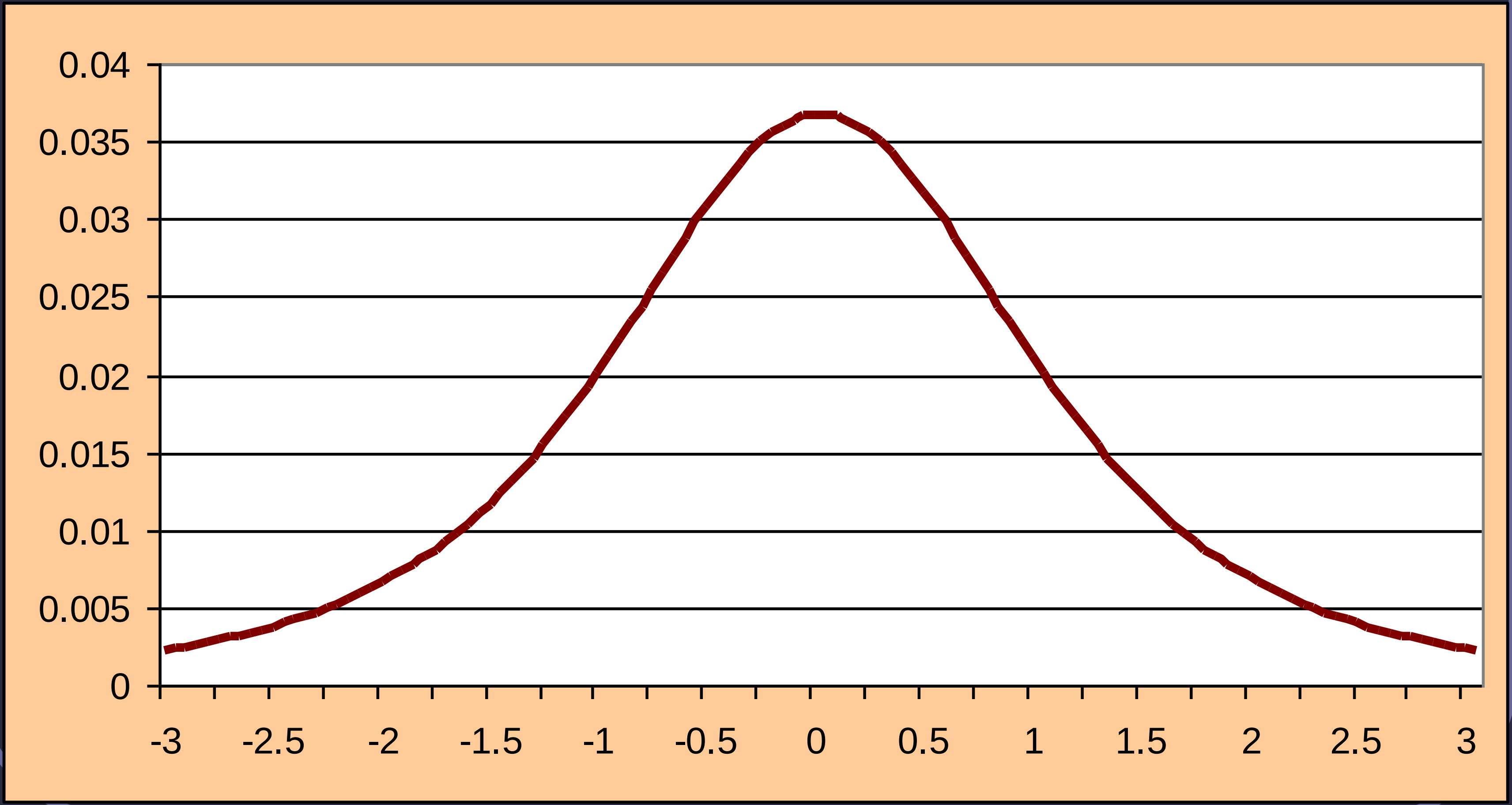
F-distributions with different degrees of freedom



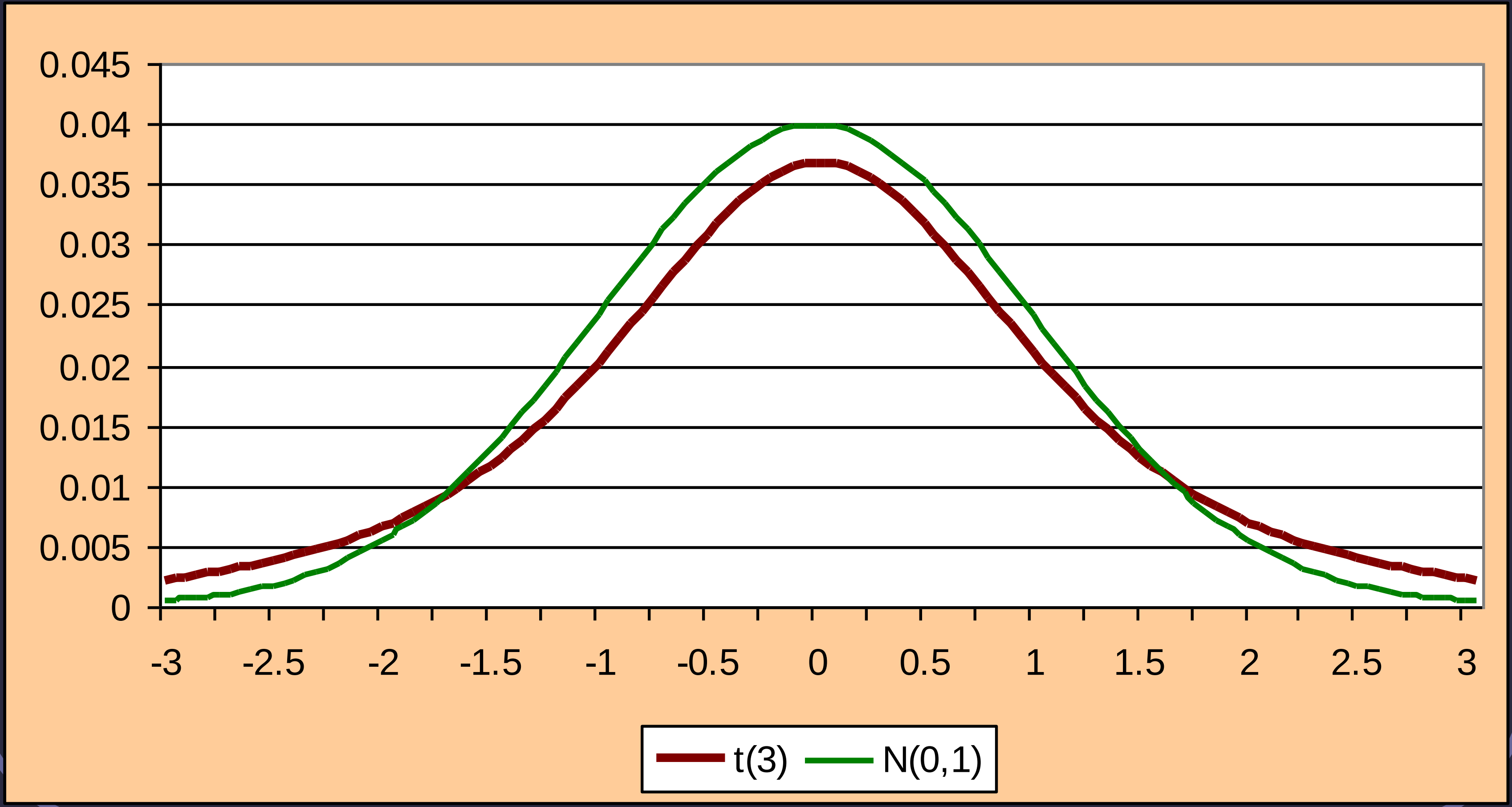
Student's *t*-distribution

- A probability distribution that arises in the problem of estimation the mean of a normally distributed population **when the sample size is small**
- The basis of the popular Student's t-tests for
  - The statistical significance of the difference between two sample means
  - Confidence intervals for the difference between two population means

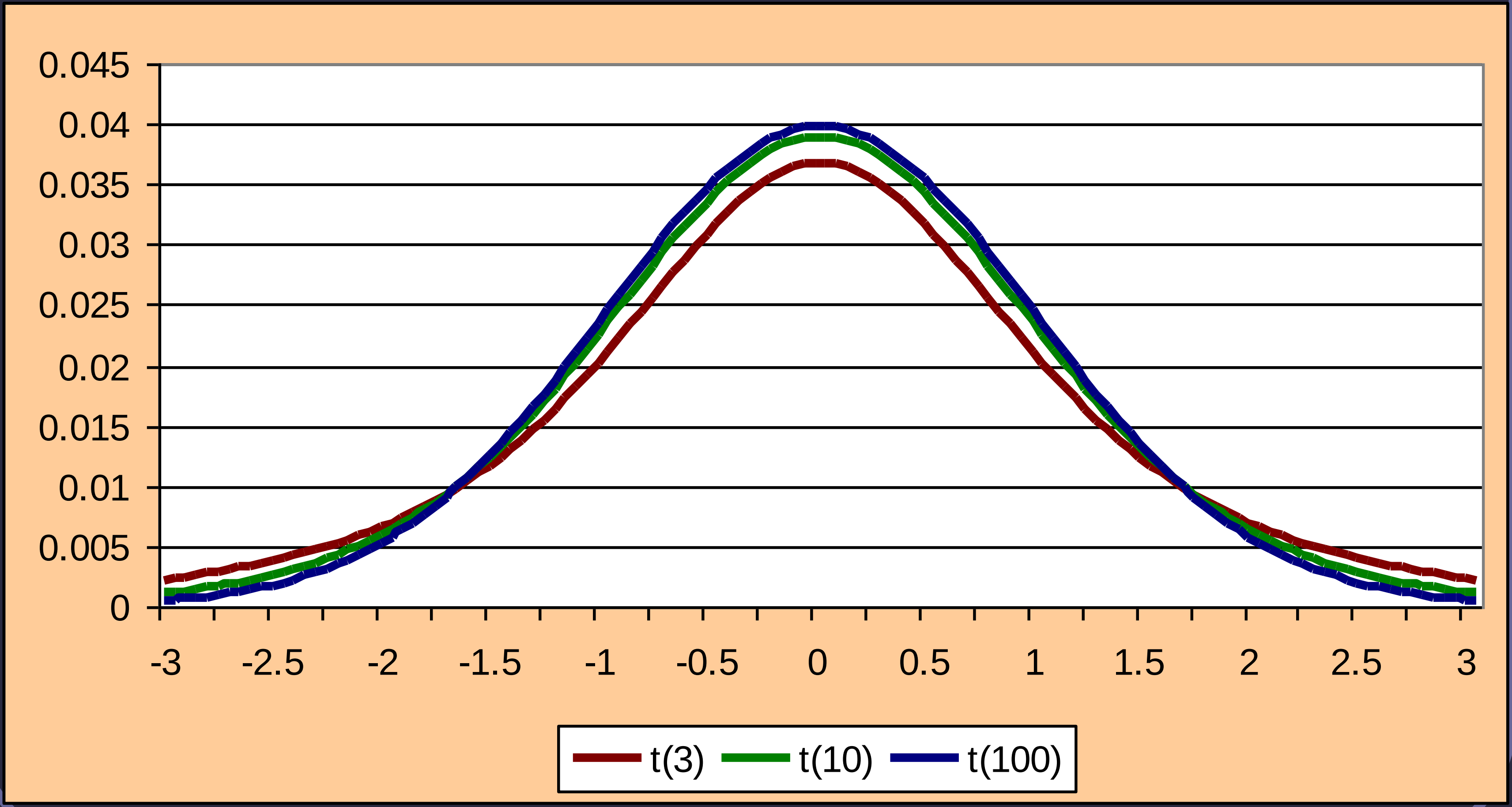
t-distribution with 3 degrees of freedom



t-distribution with 3 degrees of freedom compared to a standard normal distribution



t-distributions with {3;10;100} degrees of freedom



References

- Jarque, Carlos M. & Anil K. Bera (1980) “Efficient tests for normality, homoscedasticity and serial independence of regression residuals”. *Economics Letters* **6**(3): 255–259.
- Massey, Frank J. Jr. (1951) “The Kolmogorov-Smirnov test for goodness of fit”, *Journal of the American Statistical Association* **46**(253): 68-78.
- Pearson, Karl (1900) “On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”, *Philosophical Magazine* **50**: 157-175.

References…

- Shapiro, S. S. & M. B. Wilk (1965) “An analysis of variance test for normality”, *Biometrika* **52**(3): 591-599.
- Smirnov, N. V. (1948) “Table for estimating the goodness of fit of empirical functions”, *Annals of Mathematical Statistics* **19**: 279-281.
- Östermark R (2010) “Concurrent processing of heteroskedastic vector-valued mixture density models”, Forthcoming in *Journal of Applied Statistics.*