Econometric Analysis

Review of Statistics

This chapter: Review of Statistics

Econometric analysis requires to get an idea of the **general pattern of the data**.

Since economic data is partly systematic and partly random, we pay special attention to **data obtained by random sampling** with mutually independent observations from an underlying population with fixed mean and standard deviation.

Reference: Chapter 1 of Heij et al.(2004) "Econometric Methods with Applications in Business and Econometrics" Oxford University Press, New York

Example on bank wages

Consider the following example from the field of labour economics that will be used during the following chapters:

The data set contains information on 474 employees of a US bank. Observed variables are education and salary:

EDUC	finished years of education
SALARY	yearly salary in dollars Note that we will use LOGSALARY, the natural logarithm of the salary.

Outline

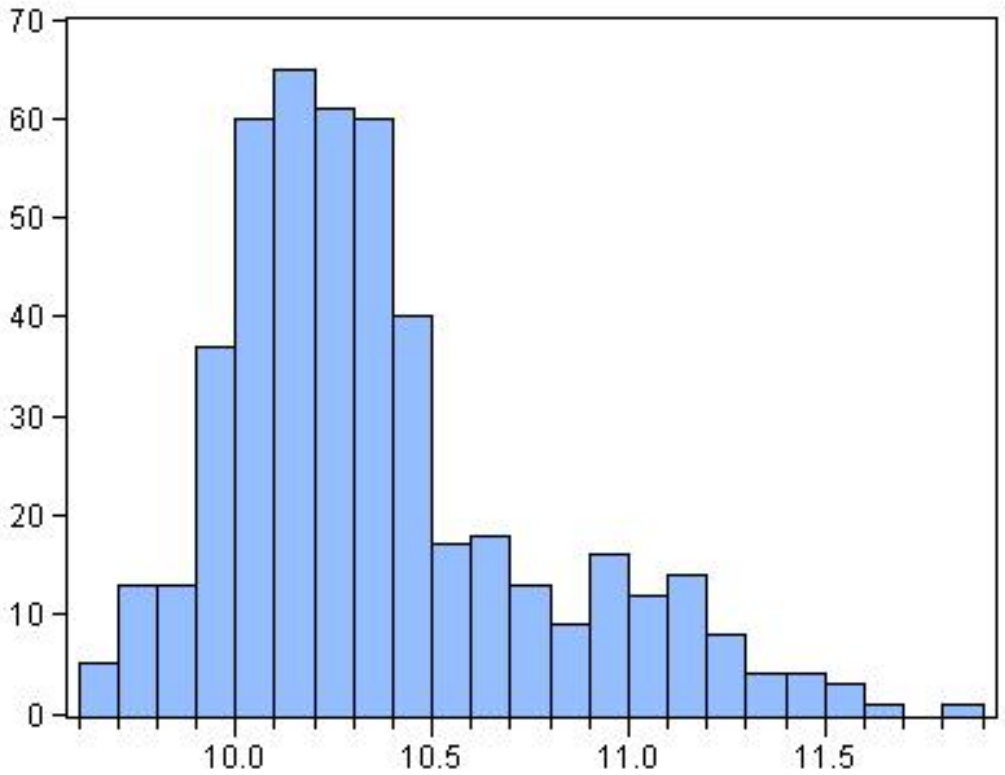
Outline

Graphs and summary statistics

In order to summarize the information given by the data, use

- simple graphical methods
  - *histograms*
  - sample *cumulative distribution function*
  - *scatter plots* $\longrightarrow$ dependencies between two variables
- summary statistics
  - *sample moments*
  - *covariance and correlation*

Histogram and summary statistics of LOGSALARY



Series: LOGSALARY
Sample 1 474
Observations 474

Mean	10.35679
Median	10.27073
Maximum	11.81303
Minimum	9.664596
Std. Dev.	0.397334
Skewness	0.998033
Kurtosis	3.662632
Jarque-Bera	87.36133
Probability	0.000000

Sample moments

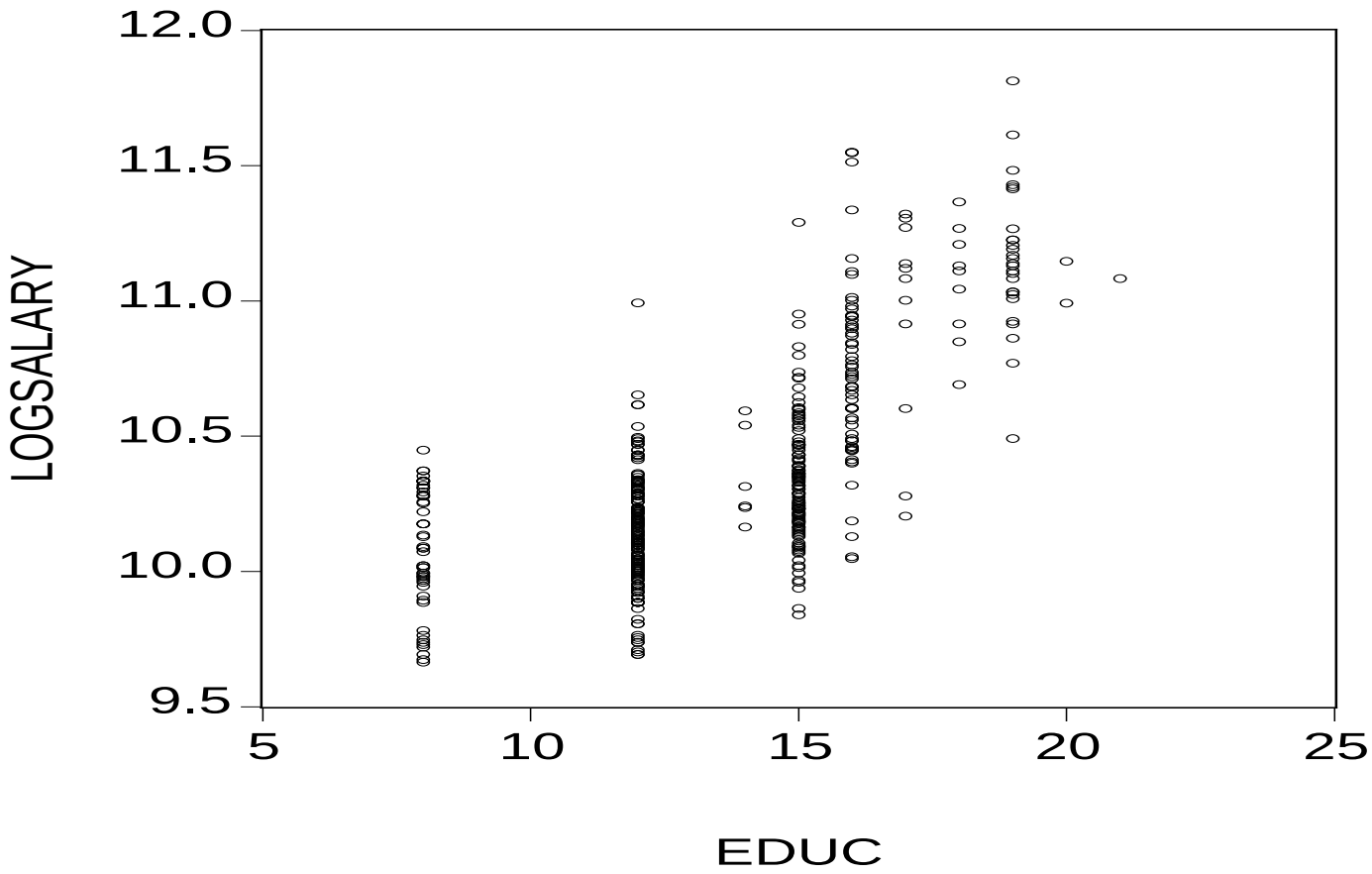
- sample **mean**: $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- **median** $y_{0.5} \longrightarrow$ *ordered observations*
  - n is odd: $y_{\frac{n+1}{2}}$
  - n is even: $\frac{1}{2}(y_{\frac{n}{2}} + y_{\frac{n+1}{2}})$
- dispersion: $m_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$.
- sample **variance**: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.
- sample **standard deviation**: $s_y = \sqrt{s_y^2}$.

Sample moments cont'd

- r^{th} centred sample moment: $m_r = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^r$.
- standardized r^{th} moment: $\frac{m_r}{s^r}$
- **skewness**: $\gamma_3 = \frac{m_3}{s^3}$.
  $\gamma_3 = 0 \rightarrow$ observations are distributed symmetrically around mean. $\rightarrow \overline{y} = y_{0.5}$.
  negative (positive) skewness: $\gamma_3 < (>) 0$.
- **kurtosis**: $\gamma_4 = \frac{m_4}{s^4}$.
  measures relative amount of observations in the tails as compared to amount of observations around mean. The fatter the tails, the larger the kurtosis.

Observations are **normally distributed** if the skewness is 0 and the kurtosis is 3. $\rightarrow$ Jarque-Bera Test

Bank wages example: Scatter Plot



Covariance and correlation

The *dependence* between two variables x and y can be measured by their common variation:

- sample **covariance** between x and y:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

- sample **correlation coefficient**:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where $r_{xy} \in [-1, 1]$.

Covariance and correlation cont'd

. . . in the case of $p \geq 2$ variables, summarize in matrices:

- the $p \times p$ sample **covariance matrix** S contains sample variances s_{jj} on the main diagonal and sample covariances s_{jk} with $j \neq k$ between each pair of the p variables:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

- the $p \times p$ sample **correlation matrix** contains the correlation coefficients $r_{jk} = s_{jk} / \sqrt{s_{jj} s_{kk}}$.
  As $r_{jj} = 1$, this matrix contains unit elements on the diagonal.

Bank wages example: Covariance and correlation

covariance matrix

	EDUC	LOGSALARY
EDUC	8.304781	0.796952
LOGSALARY	0.796952	0.157541

correlation matrix

	EDUC	LOGSALARY
EDUC	1.000000	0.696740
LOGSALARY	0.696740	1.000000

Outline

Distributions

Sampling is a reason for randomness in observations.

The uncertainty about the outcome of a random variable is described by a *probability distribution*.

- discrete outcome values v: **cumulative distribution function** (CDF) is given by

$$F(v) = P[y \leq v] = \sum_{\{i; v_i \leq v\}} p_i \tag{1}$$

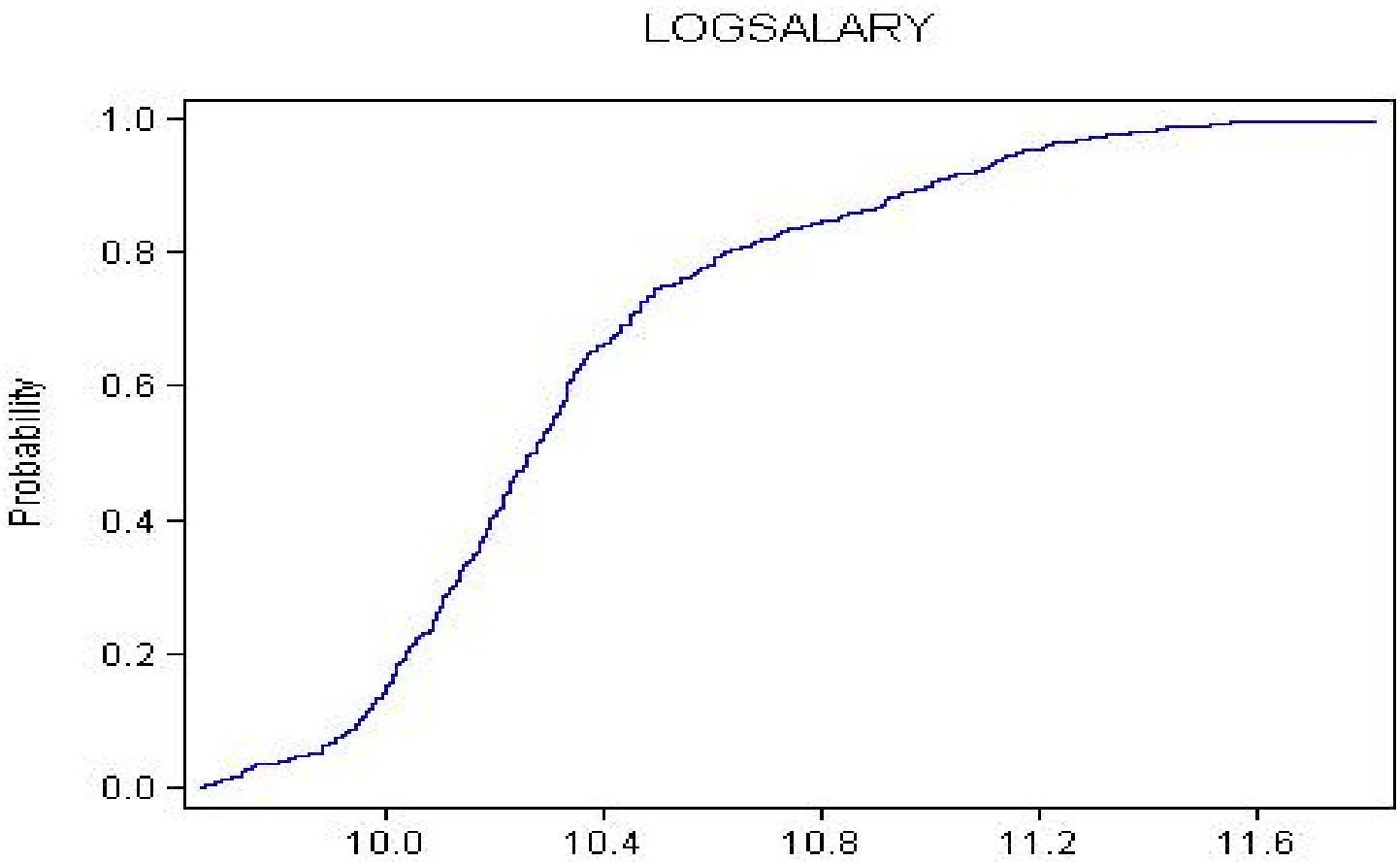
- continuous outcome values v: CDF is given as in (1).

  If CDF is differentiable, then $f(v) = \frac{dF(v)}{dv}$ is called the probability **density function**.

For observed data, the **sample CDF** is given by

$$F_s(v) = \frac{1}{n}(\text{number of } y_i \leq v).$$

CDF of LOGSALARY



First and second moments

Summarize distribution of a random variable by measures of location and dispersion:

- The **mean** μ of a random variable with outcomes v_i is determined by the expectation operator E.
  - y has a *discrete* distribution: $\mu = E[y] = \sum v_i p_i$.
  - y has a *continuous* distribution: $\mu = E[y] = \int v f(v) dv$ where f is the density function.
- The **variance** σ^2 is defined as the mean of $(y - \mu)^2$:
  - y has a *discrete* distribution: $\sigma^2 = Var[y] = E[(y - \mu)^2] = \sum (v_i - \mu)^2 p_i$.
  - y has a *continuous* distribution: $\sigma^2 = Var[y] = E[(y - \mu)^2] = \int (v - \mu)^2 f(v) dv$.

[*The* **standard variation** σ *is the square root of the variance.*]

Higher moments

The r**th centred moment** $\mu_r = E[(y - \mu)^r]$ and the **standardized** r**th moment** μ_r/σ^r are defined accordingly.

The sample moments (see above) are obtained by replacing the CDF by the sample CDF.

Although the sample moments always exist, this is not always the case for the population moments. If $E[|y - \mu|^c] < \infty$, then all the moments μ_r with $r \leqslant c$ exist.

Transformation of random variables

Suppose a **linear transformation of a random variable** y by the function $g(y) = a \cdot y + b$.
$\rightarrow$Thus $z = g(y)$ is also a random variable.

- The *mean* of z is given by $E[a \cdot y + b] = a \cdot E[y] + b$.
- The *variance* of z is given by $Var[a \cdot y + b] = a^2 \cdot Var[y]$.

The uncertainty about a pair of outcomes (x, y) can be described by a *joint probability distribution*.

The corresponding cumulative distribution function (CDF) is given by

$$F(v, w) = P[x \leq v, y \leq w] \tag{2}$$

both in the discrete and continuous case.

If the sets of possible outcomes are continuous and the second derivative of this function exists, then the corresponding density function is defined by

$$f(v, w) = \frac{\delta^2 F(v, w)}{\delta v \delta w}$$

Covariance and correlation

Suppose continuous distributions of x and y. Then **covariance** and the **correlation coefficient** are defined by:

$$cov(x,y) = E[(x-\mu_x)(y-\mu_y)] = \iint (v-\mu_x)(w-\mu_y)f(v,w)dvdw \tag{3}$$

where $f(v,w)$ is the density function.

$$\rho_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{4}$$

If $\rho_{xy} = 0$, then the two random variables are called **uncorrelated**. This is equivalent to the condition

$$E[xy] = E[x]E[y] \tag{5}$$

Conditional distribution

When the outcome $x = v_i$ is given, it follows a *conditional distribution* of y. Hence we can write the *conditional probabilities* and in the case of a continuous distribution the *conditional density*.

The *conditional mean and variance* of y are the mean and variance with respect to the corresponding conditional distribution.

Independence

x and y are called **independent random variables** if (for continuous distributions) the joint distribution equals the product of the marginal distributions:

$$f(v, w) = f_x(v) f_y(w) \tag{6}$$

Independent variables are always uncorrelated, but the reverse does not hold true.

Sum of two random variables

The **sum of two random variables** x and y gives again a random variable $z = x + y$.

- The *mean* of z is given by $E[x + y] = E[x] + E[y]$.
- The *variance* of z is given by $Var[x + y] = Var[x] + Var[y]$ in case x and y are independently distributed, otherwise it is $Var[x + y] = Var[x] + Var[y] + 2 \cdot Cov[x, y]$.

Bernoulli distribution

Consider a random variable x with only two possible outcomes 0 and 1. The probability distribution is completely described by

$$p = P[y = 1] \text{ and } P[y = 0] = 1 - P[y = 1] = 1 - p.$$

This is called a **Bernoulli distribution**: $y \sim Be(p)$.

The first and second moments are given by:

$$E(y) = p \text{ and } Var(y) = p(1 - p).$$

Binomial distribution

The sum of n independently and identically distributed Bernoulli variables with probability p of success gives a **binomial distribution**: $y \sim Bi(n, p)$.

The probability distribution is

$$P[y = v] = \begin{pmatrix} n \\ v \end{pmatrix} p^v (1-p)^{n-v}$$

where y is the total number of successes.

Mean and variance are given by:

$$E(y) = np \text{ and } Var(y) = np(1-p).$$

Normal distribution

The **normal distribution** $y \sim N(\mu, \sigma^2)$ is the most widely used distribution in econometrics, since many distributions can be approximated by normal distributions if the sample size is large enough [*central limit theorem*] and besides it has several attractive properties.

A normal random variable is a continuous variable that can take on any value. Its density function is given by

$$f(v) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(v-\mu)^2}, \; -\infty < v < \infty.$$

The function is symmetric around μ and shaped like a bell.
The population moments are:

$$E(y) = \mu, \; Var(y) = \sigma^2, \; \gamma_3 = 0 \text{ and } \gamma_4 = 3.$$

Normal distribution cont'd

- The *linear transformation* of a normally distributed variable is again normally distributed:

$$a \cdot y + b \sim N(a\mu + b, a^2\sigma^2).$$

- y can be standardized by subtracting its mean and dividing by its standard deviation. It follows a **standard normal distribution**:

$$\frac{y-\mu}{\sigma} \sim N(0,1).$$

The corresponding density function ϕ and CDF Φ are given by:

$$\phi(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}, \; \Phi(v) = \int_{-\infty}^{v} \phi(u)du.$$

Normal distribution cont'd

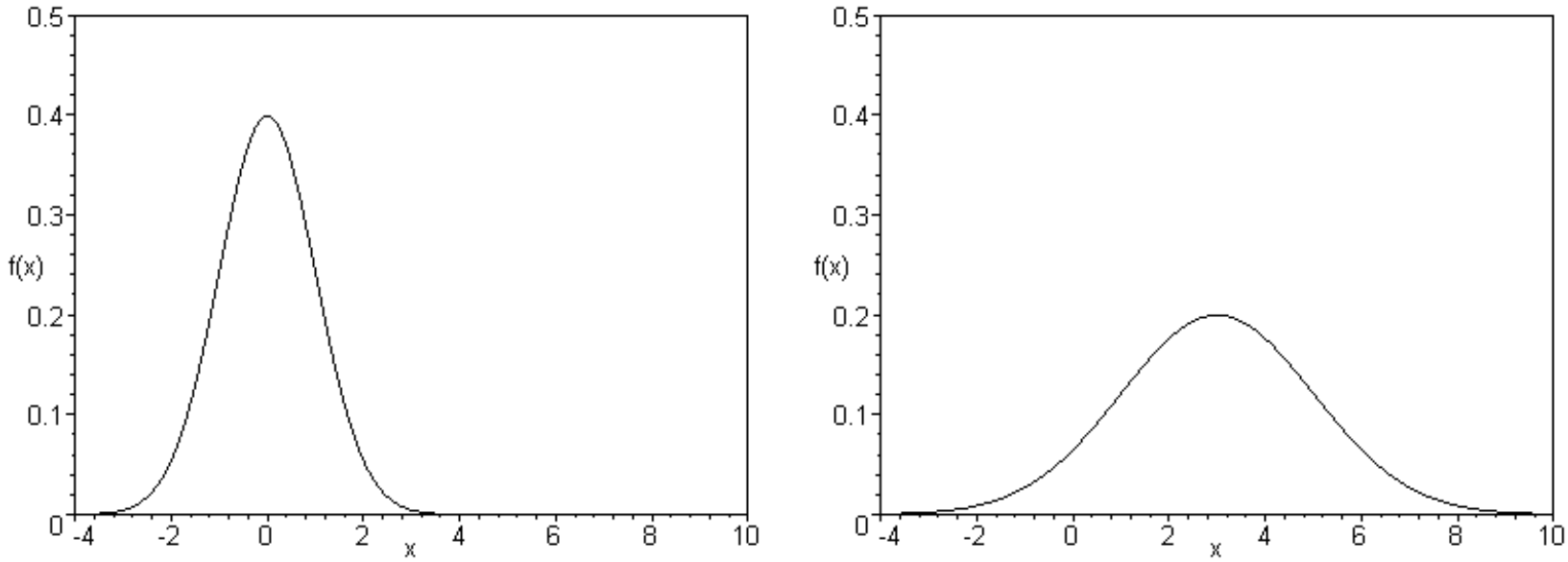


Figure: Density function of a standard normal distribution and another normal distribution with mean 3 and variance 2.

χ^2 distribution

Suppose independently distributed $y_i \sim N(0,1)$ with $i = 1, ..., n$.

The sum of the squares $\sum_{i=1}^{n} y_i^2$ is then called the **chi-square distribution** with n degrees of freedom, denoted by $\chi^2(n)$.
It has mean $E(y) = n$ and variance $var(y) = 2n$.

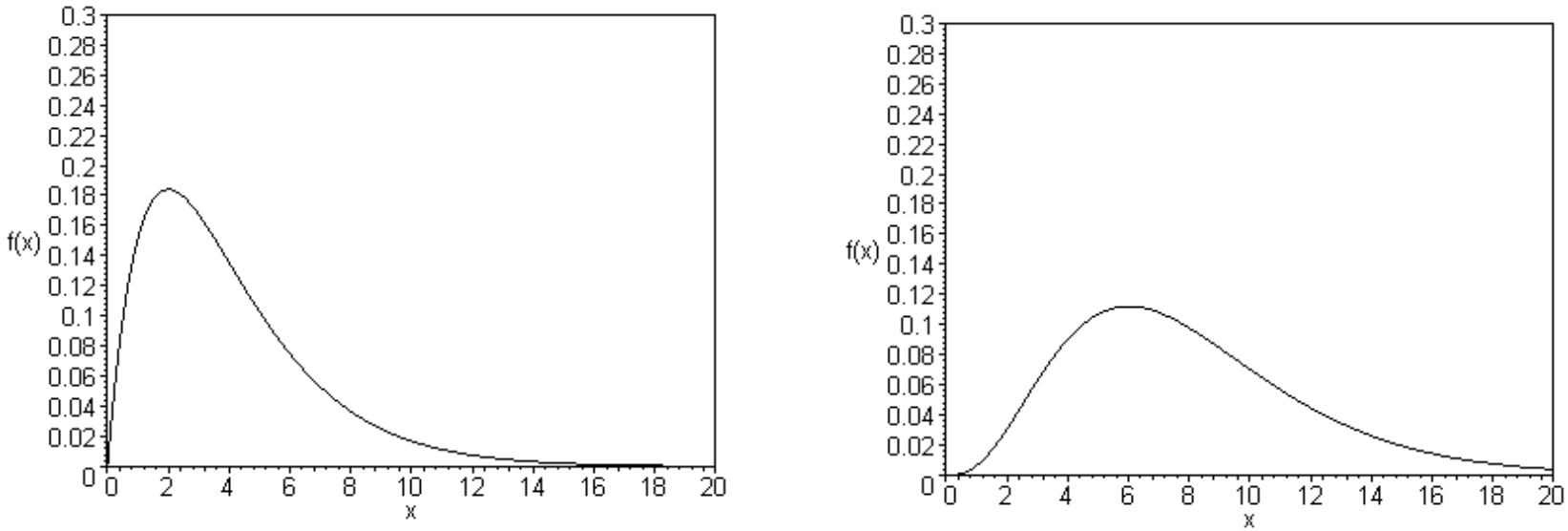


Figure: Density functions of two chi-squared distributions, with degrees of freedom equal to 4 (*left*) and 8 (*right*).

Student t-distribution

Suppose $y_1 \sim N(0,1)$ and $y_2 \sim \chi^2(r)$ are independently distributed. Then the distribution of $\frac{y_1}{\sqrt{y_2/r}}$ is called the **Student t-distribution** with r degrees of freedom:

$$\frac{y_1}{\sqrt{y_2/r}} \sim t(r).$$

The Student t-distribution is symmetric and has fat tails. Thus the kurtosis is larger than three. For $r > 1$, the mean is $E(y) = 0$ and for $r > 2$ the variance is given by $var(y) = \frac{r}{r-2}$.
If $r \to \infty$, then the $t(r)$ density converges to the standard normal density.

Student t-distribution cont'd

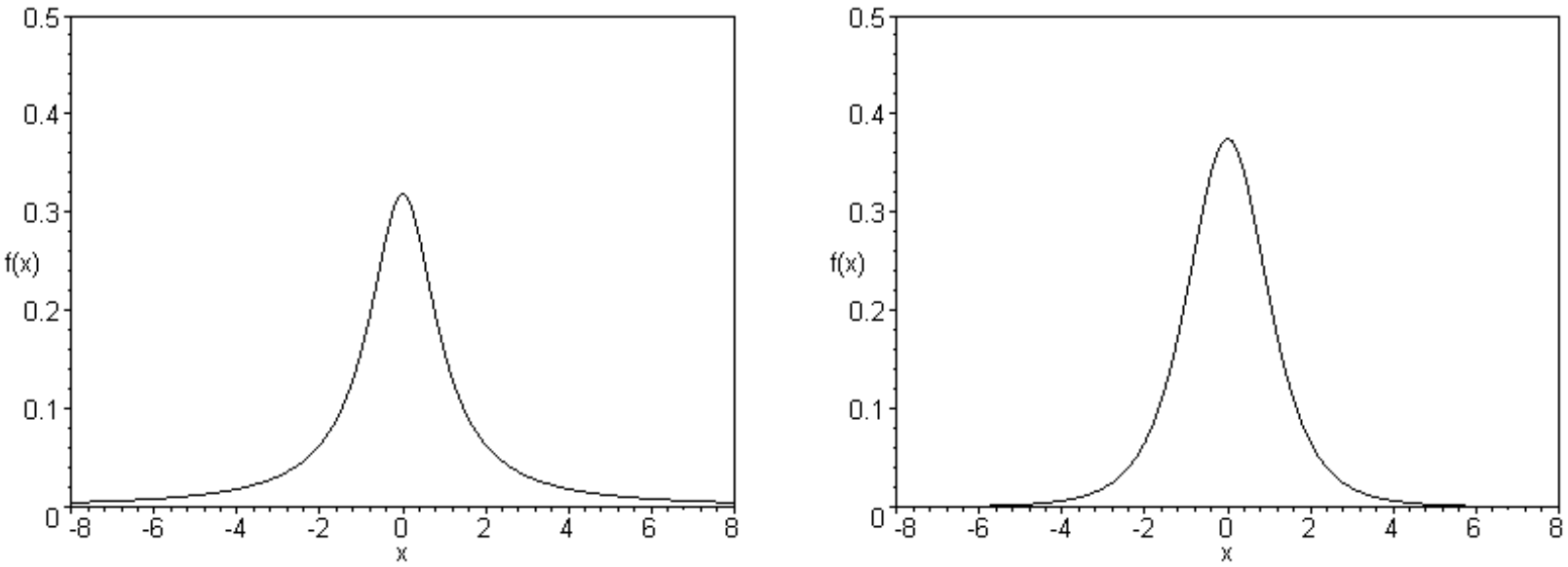


Figure: Density functions of two t-distributions, with number of degrees of freedom equal to 1 (*left*: Cauchy distribution) and 4 (*right*).

F distribution

Suppose $y_1 \sim \chi^2(r_1)$ and $y_2 \sim \chi^2(r_2)$ are independently distributed. Then the distribution of $\frac{y_1/r_1}{y_2/r_2}$ is called the **F-distribution** with r_1 and r_2 degrees of freedom:

$$\frac{y_1/r_1}{y_2/r_2} \sim F(r_1, r_2).$$

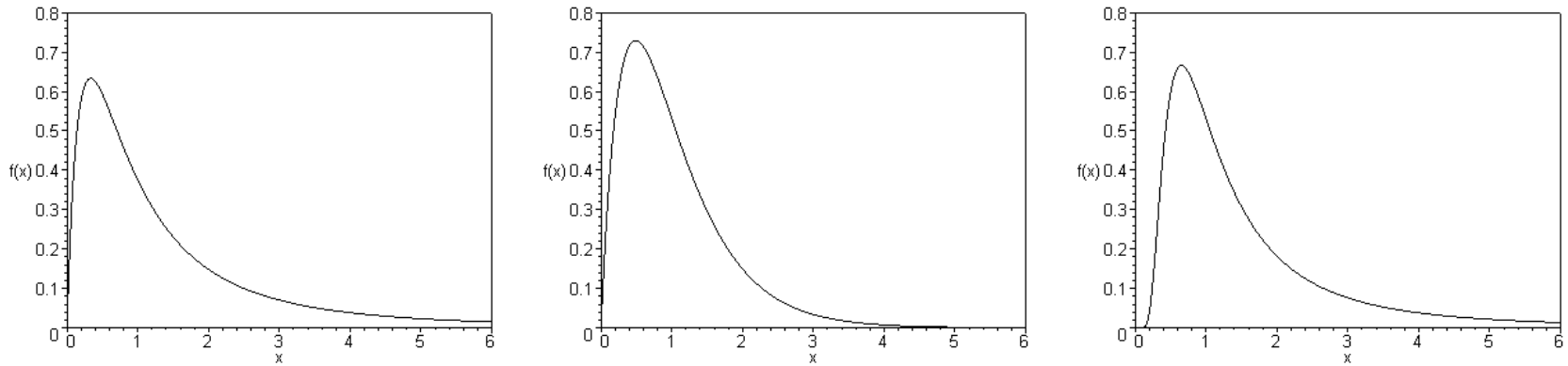


Figure: Density functions of three *F*-distributions, with number of degrees of freedom in numerator and denominator respectively (4,4) (*left*), (4,100) (*middle*) and (100,4) (*right*).

Outline

Basics

Suppose a joint probability distribution $f_\theta(y_1, \ldots, y_n)$ with observations y_i with $i = 1, \ldots, n$.

parameter θ	The general shape of the distribution f_θ is known up to one or more unknown parameters.
model $\{f_\theta; \theta \in \Theta\}$	A set of distributions that specifies the general shape of the distribution together with a set Θ of possible values for the unknown parameters.
estimator $\widehat{\theta}$	The numerical values of θ are unknown, but can be estimated from the observed data. = numerical expression in terms of random variables, as it depends on the random variables y_i
estimate	The resulting numerical value of the estimator is called the estimate of the parameter.

Estimation methods

- **Method of moments** is based on moments that are often easy to compute. However, the obtained estimates therefore depend on the chosen moments.
- **Least squares** optimizes the fit of the model with respect to the observations.

$\longrightarrow$ Both methods are based on the idea of *minimizing a distance function* where distance is measured in terms of observed data or of sample and population moments.

- **Maximum likelihood** maximizes a likelihood function which is a function of θ.

$\longrightarrow$ The idea of ML is based on the likelihood function that expresses the likelihood or 'credibility' of parameter values with respect to the observed data.

Least Squares

Suppose one intends to estimate the population mean from a random sample $y_1, \ldots, y_n$ where μ and σ^2 are unknown. Consider the respective model:

$$y_i = \mu + \varepsilon_i,\ \varepsilon \sim \text{IID}(0, \sigma^2)$$

where $\varepsilon_i = y_i - \mu$ are identically and independently distributed.

The **least squares estimate** is

$$\text{argmin } S(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2.$$

Procedure: 1) Take first derivative with respect to μ considering the first order condition (F.O.C.) and 2) solve for μ. The result is then $\overline{y} = \widehat{\mu}$.

Data generating process

The **data generating process**(DGP) is often used in order to evaluate the quality of estimators.

Data is then generated by a particular distribution that belongs to the specified model. Thus the DGP of $y_1, \ldots, y_n$ has a distribution f_{θ_0} where $\theta_0 \epsilon \Theta$.

The DGP will be covered in an empirical exercise.

Variance and bias

The **mean squared error (MSE)** of an estimator that consists of a single parameter is defined by

$$MSE(\widehat{\theta}) = E[(\widehat{\theta} - \theta_0)^2] = var(\widehat{\theta}) + (E[\widehat{\theta}] - \theta_0)^2. \qquad (7)$$

and provides a trade-off between the variance and the bias of an estimator.

However, in general θ_0 is unknown such that the practical use of the MSE is limited.

Unbiased and efficient estimators

- **Unbiasedness**:

$$E[\widehat{\theta}] = \theta_0$$

- Asymptotic unbiasedness:

$$lim_{n \to \infty} E[\widehat{\theta}] = \theta_0$$

- **Efficiency**: an estimator that minimizes the variance over a class of estimators.

Variance and bias cont'd

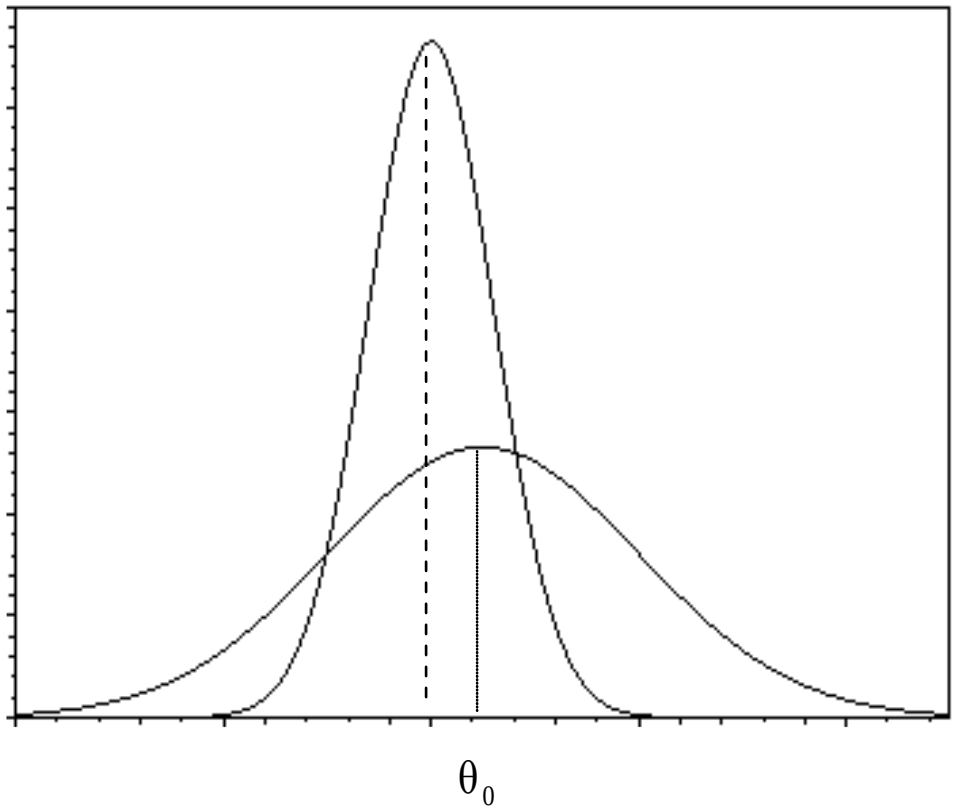


Figure: Densities of two estimators; θ_0 denotes the parameter of the DGP

Consistency

The estimator is called **consistent** if it converges in probability to θ_0 that is, if for all $\delta > 0$ there holds

$$lim_{n\to\infty} P[|\widehat{\theta}_n - \theta_0| < \delta] = 1 \tag{8}$$

where $\widehat{\theta}_n$ is an estimator of θ based on a sample of n observations and θ_0 is the **probability limit** of $\widehat{\theta}_n$:

$$plim(\widehat{\theta}_n) = \theta_0$$

A sufficient but not necessary condition for consistency is that the estimator is asymptotically unbiased and

$$lim_{n\to\infty} Var(\widehat{\theta}) = 0$$

Consistency cont'd

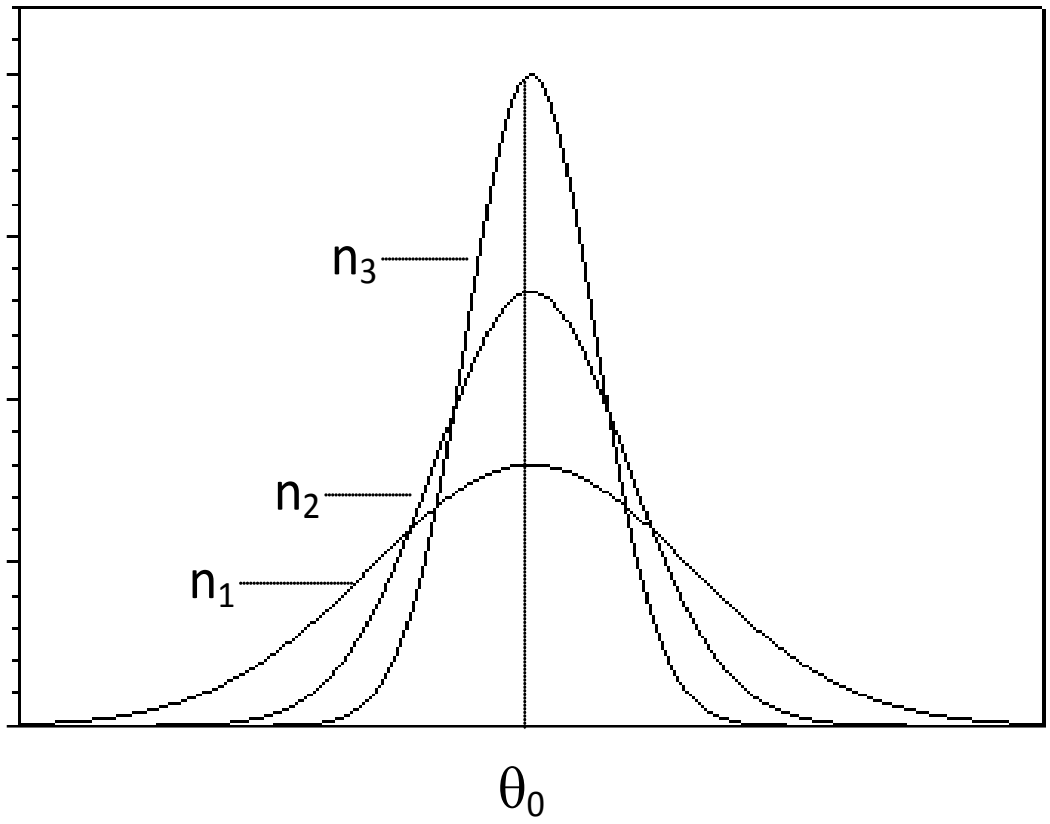


Figure: Distribution of a consistent estimator for three sample sizes, with $n_1 < n_2 < n_3$.

Law of large numbers

When data consist of a random sample from a population, sample moments provide *consistent estimators* of the population moments $\rightarrow$ **law of large numbers**

If $y_i \sim IID$ with finite population mean $E[y_i] = \mu$, then

$$plim(\frac{1}{n}\sum_{i=1}^{n} y_i) = \mu \tag{9}$$

Similarly, if $y_i \sim IID$ *and* $\mu_r = E[(y_i - \mu)^r] < \infty$, *then*

$plim(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^r)) = \mu_r$.

Central limit theorem

Let y_i, $i = 1, \ldots n$ be independently and identically distributed random variables with mean μ and finite variance σ^2. Then

$$z_n = \sqrt{n}\frac{\overline{y}_n - \mu}{\sigma} \xrightarrow{d} z \sim N(0,1) \tag{10}$$

For large enough sample sizes, the finite sample distribution of z_n can be approximated by a standard normal distribution $N(0,1)$. It follows that

$$\overline{y}_n \approx N(\mu, \frac{\sigma^2}{n})$$

The above central limit theorem for the IID case can be generalized in several directions.

Outline

Null hypothesis and alternative hypothesis

We concentrate on parametric hypotheses where the **null hypothesis** H_0 is tested against a **alternative hypothesis** H_1. Consider for instance a test for the mean of a population.

The observed data $(y_1, \ldots, y_n)$ are used to decide which of the hypotheses seems to be most appropriate.

This decision is made by means of a *test statistic* t that can be computed from the observed data.

Test statistic and critical region

The possible outcomes of this **test statistic t** are divided into two regions, the **critical region C** and the respective complement.

$\rightarrow$ Then, H_0 is rejected if $t \in C$ and not rejected if $t \notin C$.

In order to test a hypothesis, one has to decide about t and C in such a way that one can discriminate well between H_0 and H_1.

Size and power

Error of first type α: if H_0 is valid but the observed data lead to rejection of H_0

significance level / size: the probabilty of this α-error

Error of second type β: if H_0 is false but the observed data do not lead to rejection of H_0

The rejection probability of a false H_0 is called the **power** of the test. A test is called **consistent** if the power converges to 1 for all cases in which H_0 is false if $n \to \infty$. *What would be a perfect test?*

Significance level

In practice one often fixes *a maximally tolerated size* to control for errors of the first type, for instance 5%.
Therefore tests should be formulated in such a way that errors of the first type are more serious than errors of the second type.

One should distinguish *statistical significance* from *practical significance*.

Two-sided test for the mean

Let $y_i \sim N(\mu, \sigma^2)$ and both mean μ and variance σ^2 are unknown.

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0$$

where μ_0 is a given / hypothesized value.

As test statistic we consider the sample mean $\overline{y}$ and reject H_0 if

$$|\overline{y} - \mu_0| > c$$

where c determines the significance level.

Under H_0,

$$\frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Two-sided test for the mean cont'd

Since σ^2 is unknown, we replace by the unbiased estimator s^2 and get the *test statistic*:

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} \sim t(n-1) \tag{11}$$

H_0 is rejected if $|t| > c$ / if $\overline{y}$ falls in the critical region

$$\overline{y} < \mu_0 - c\frac{s}{\sqrt{n}} \quad or \quad \overline{y} > \mu_0 + c\frac{s}{\sqrt{n}} \tag{12}$$

One-sided test for the mean

In some cases it might be of interest to test

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

The test statistic is again as given in (11) and H_0 is rejected if $t > c$.

Probability value (P-value)

Instead of fixing the size / significance level, one can leave the size unspecified. One can then ask for which sizes the test outcome would lead to rejection of H_0.

The minimal value of the size for which H_0 is rejected is called the **probability value** or P-**value** of the test outcome.

H_0 should be rejected for all sizes larger than P and should not be rejected for all sizes smaller than P.

P-value