Statistics

Lecture 8

Point and interval estimates

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

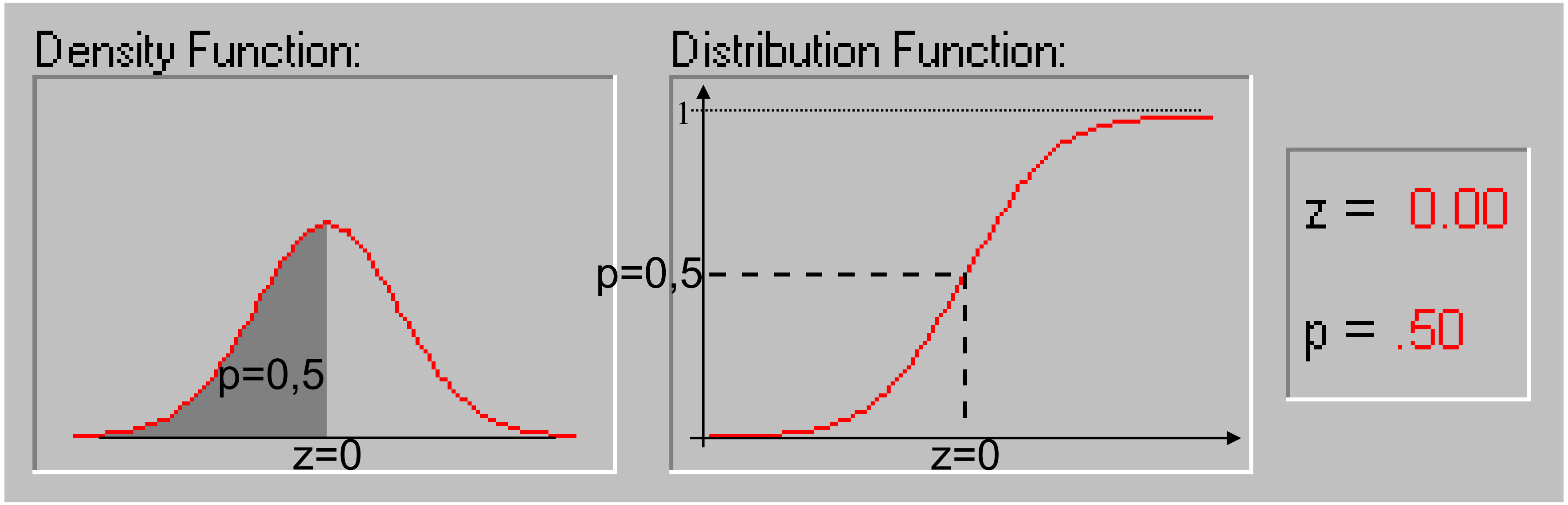
David Bartl
Statistics
INM/BASTA

Outline of the lecture

- Central Limit Theorem: the Lindenberg-Lévy Theorem
- Sampling and survey data collection
  - — sampling with replacement
  - — sampling without replacement
- Point estimates
  - — point estimates for the population mean and for the population variance
- Interval estimates
  - — interval estimates for the population mean and for the population variance

Normal distribution

The graph of the probability density function & cumulative distribution function of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$:



CLT: Lindeberg-Lévy Theorem

Let $X_1, X_2, X_3, \ldots$ be a sequence of **independent & identically distributed** random variables with finite expected value $\mathrm{E}[X_i] = \mu$ and with finite variance $\mathrm{Var}(X_i) = \sigma^2$. Then

$$P\left(\left\{\omega \in \Omega^n : \sqrt{n}\left(\frac{X_1(\omega_1) + \cdots + X_n(\omega_n)}{n} - \mu\right) < x\right\}\right) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{t^2}{2\sigma^2}} \mathrm{d}t$$

as $n \to \infty$

for every $x \in \mathbb{R}$.

CLT: Lindeberg-Lévy Theorem in brief

We have:

$$P\left(\sqrt{n}\left(\frac{X_1+\cdots+X_n}{n}-\mu\right)<x\right)\to\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{t^2}{2\sigma^2}}\,\mathrm{d}t \qquad \text{as} \quad n\to\infty$$

or

$$P\left(\left(\frac{X_1+\cdots+X_n}{n}-\mu\right)<\frac{x}{\sqrt{n}}\right)\to\int_{-\infty}^{x/\sqrt{n}}\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(t\sqrt{n})^2}{2\sigma^2}}\sqrt{n}\,\mathrm{d}t \qquad \text{as} \quad n\to\infty$$

or

$$P\left(\left(\frac{X_1+\cdots+X_n}{n}-\mu\right)<\frac{x}{\sqrt{n}}\right)\to\int_{-\infty}^{x/\sqrt{n}}\frac{1}{\sqrt{2\pi\,\sigma^2/n}}\mathrm{e}^{-\frac{t^2}{2\sigma^2/n}}\,\mathrm{d}t \qquad \text{as} \quad n\to\infty$$

or

CLT: Lindeberg-Lévy Theorem in brief

We have:

$$P\left(\sqrt{n}\left(\frac{X_1+\cdots+X_n}{n}-\mu\right)<x\right)\to\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{t^2}{2\sigma^2}}\,\mathrm{d}t \qquad \text{as} \quad n\to\infty$$

or

$$\dots$$

or

$$P\left(\left(\frac{X_1+\cdots+X_n}{n}-\mu\right)<\frac{x}{\sqrt{n}}\right)\to\int_{-\infty}^{x/\sqrt{n}}\frac{1}{\sqrt{2\pi\,\sigma^2/n}}\mathrm{e}^{-\frac{t^2}{2\sigma^2/n}}\,\mathrm{d}t \qquad \text{as} \quad n\to\infty$$

or

$$P\left(\frac{X_1+\cdots+X_n}{n}-\mu<x\right)\to\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\,\sigma^2/n}}\mathrm{e}^{-\frac{t^2}{2\sigma^2/n}}\,\mathrm{d}t \qquad \text{as} \quad n\to\infty$$

CLT: Lindeberg-Lévy Theorem in in other words

In other words, we approximately have:

$$\underbrace{\frac{X_1 + X_2 + \cdots + X_n}{n}}_{\bar{X}} - \mu \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \qquad \text{as} \quad n \to \infty$$

or

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{as} \quad n \to \infty$$

Example: Invoices

We have a population of 1250 invoices for amounts between 100 and 10000 units of money. The true population characteristics are

$$\mu = 5097 \qquad \text{and} \qquad \sigma^2 = 170156.25 \quad \text{or} \quad \sigma = 412.5$$

(We usually do not know these characteristics.)

Take a sample of 50 invoices out of the population of the 1250 invoices.

There are up to

$$\binom{1250}{50} \doteq 8.53 \times 10^{89}$$

such samples.

Example: Invoices

There are up to

$$\binom{1250}{50} \doteq 8.53 \times 10^{89}$$

of 50-element samples out of the 1250-element population.

We, actually, take 500 various 50-element samples.

The sample average amount of the samples is in the range between 3800 and 6400 units of money.

Example: Invoices

The table

of the frequencies

for the sample mean

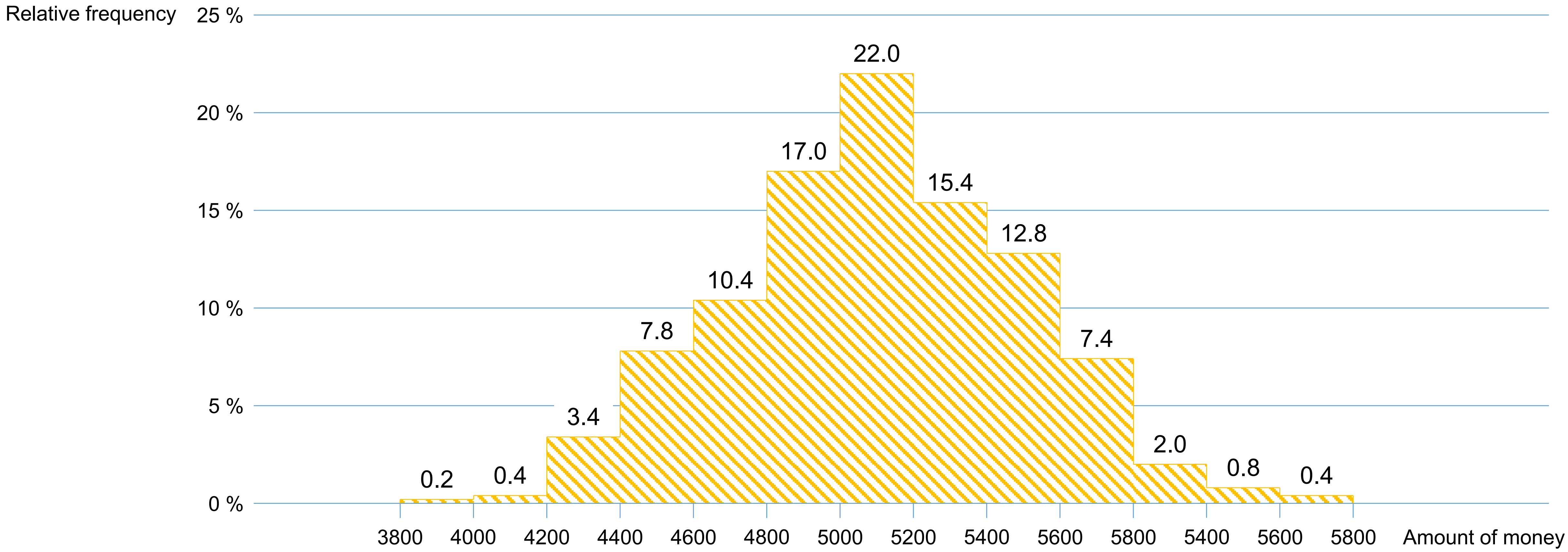
of the 50-element

samples:

Interval of the sample mean	Frequency (number)	Relative frequency
$3800 \leq \overline{x} < 4000$	1	0.2 %
$4000 \leq \overline{x} < 4200$	2	0.4 %
$4200 \leq \overline{x} < 4400$	17	3.4 %
$4400 \leq \overline{x} < 4600$	39	7.8 %
$4600 \leq \overline{x} < 4800$	52	10.4 %
$4800 \leq \overline{x} < 5000$	85	17.0 %
$5000 \leq \overline{x} < 5200$	110	22.0 %
$5200 \leq \overline{x} < 5400$	77	15.4 %
$5400 \leq \overline{x} < 5600$	64	12.8 %
$5600 \leq x < 5800$	37	7.4 %
$5800 \leq \overline{x} < 6000$	10	2.0 %
$6000 \leq \overline{x} < 6200$	4	0.8 %
$6200 \leq \overline{x} < 6400$	2	0.4 %
TOTAL:	500	100.0 %

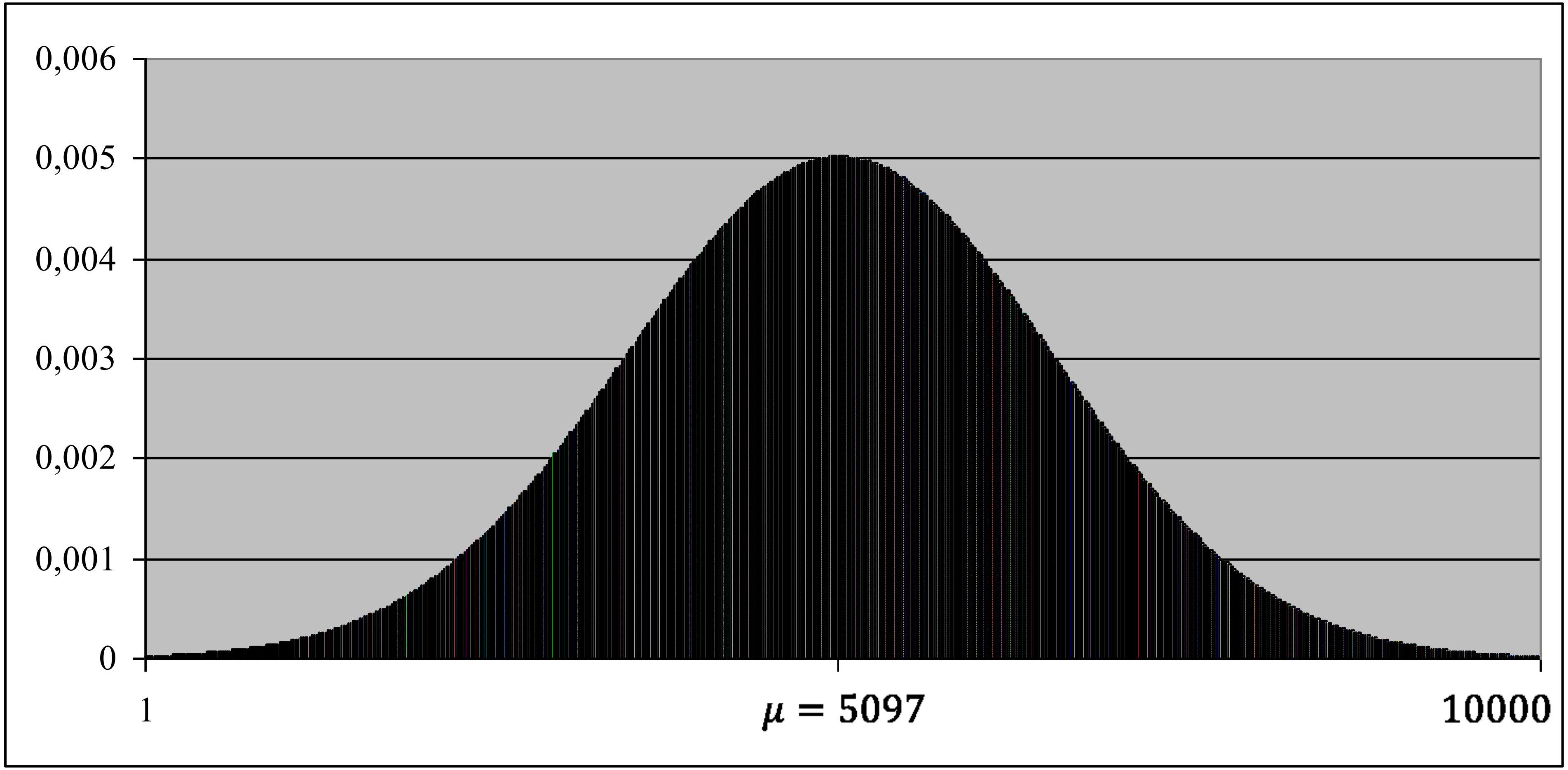
Example: Invoices

Histogram – the relative frequencies of the sample mean – 500-element sample:



Example: Invoices

Histogram – the relative frequencies of the sample mean – 5000-element sample:



Sampling and survey data collection

- The population is often very large, or the population may not be clearly specified in practice.
- **Sampling** is a method to obtain a relevant sample of the entire population.
- The sample should be representative, i.e. its structure should follow the structure of the entire population.
- Sampling plan (random sample, etc.) – see below.
- The methods of collecting the data:
  - [opinion] poll, survey
  - questionnaire
  - on-line / telephone / mail / face-to-face / …

Sampling and survey data collection

- **A random sample** is the collection of pairwise independent values of a random variable.

  [Recall: Given the probability space $(\Omega, \mathcal{F}, P)$, a random variable is any measurable function $X: \Omega \to \mathbb{R}$. Any outcome $\omega \in \Omega$ is the result of the random experiment. We then obtain the value $X(x)$ of the random variable. Recall also that we consider $\Omega = \mathbb{R}$ and $X(x) = x$ for simplicity if the random variable X is continuous.]

Sampling and survey data collection

- **Simple random sampling** – each element has equal chance of being selected.
- Systematic sampling.
- Stratified sampling.
- Cluster sampling.
- Accidental sampling.
- …

Sampling with and without replacement

Replacement of selected units:

- **sampling <u>without</u> replacement**:
  - an element can appear <u>no more than once</u> in the sample
- **sampling <u>with</u> replacement:**
  - an element <u>can appear several times</u> in the sample

Note that sampling without replacement is often the case in practice.

Point estimates & Interval estimates

There are two kinds of estimates:

- **point estimates** — we directly calculate the estimate as a single number, e.g.
  - — the sample mean $\bar{x}$ estimates the population mean μ
  - — the sample variance s^2 estimates the population variance σ^2
- **interval estimates** — the purpose is to find an interval $[a, b]$ such that the probability that the estimated value (the mean, the variance) belongs to the interval is sufficiently high, ≥ 95 %, say, $\geq 1 - \alpha$ where α is the <u>significance level</u> (most popular values are $\alpha = 5\,\%$, $\alpha = 1\,\%$, or $\alpha = 10\,\%$) and $(1 - \alpha)$ is the <u>confidence level</u>.

Point estimates

We already know some point estimates:

- The sample mean is an estimate of the population mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \approx \mu$$

- The sample variance is an estimate of the population variance:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \approx \sigma^2$$

Point estimates

The point estimate is an expression or formula, i.e. a statistic $f_n(x_1, \dots, x_n)$, of the sample values $x_1, \dots, x_n$ (such as $\bar{x} = f_n(x_1, \dots, x_n) = \sum_{i=1}^{n} x_n / n$ or $s^2 = f_n(x_1, \dots, x_n) = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n - 1)$). It should possess the following three properties at least:

- **Unbiasedness**
- **Consistency**
- **Efficiency**

Point estimates

Unbiasedness — the expected value of the estimator f_n should be equal to the estimated value. We already know that

$$\mathrm{E}[\bar{x}] = \mu \qquad \text{and} \qquad \mathrm{E}[s^2] = \sigma^2$$

i.e. the sample mean and the sample variance are unbiased estimators.

Point estimates

Consistency — if the estimator f_n is unbiased, then the condition sufficient for consistency is that

$$\mathrm{Var}\big(f_n(x_1, \dots, x_n)\big) \longrightarrow 0 \qquad \text{as} \quad n \to \infty$$

We already know that

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

(always)

and it holds (see below) that

$$\mathrm{Var}(s^2) = \frac{2(\sigma^2)^2}{n-1}$$

(if the sampled variables are normal)

i.e. the sample mean and the sample variance are consistent estimators too.

Point estimates

Efficiency — there are several definitions; we shall require the estimator f_n to be a "minimum variance unbiased estimator", i.e. we require that the variance $\mathrm{Var}(f_n)$ of the estimator f_n is minimal among all estimators. In other words, if $f_n(x_1, \dots, x_n)$ is an estimator of the quantity ϑ, then

$$\mathrm{Var}(g_n) \geq \mathrm{Var}(f_n) \qquad \text{for any other estimator } g_n(x_1, \dots x_n) \text{ of the quantity } \vartheta$$

It holds that the sample mean and the sample variance are efficient estimators.

An illustrative example

The goal is to estimate the average μ and the variance σ^2 of the value of a purchase (shopping) in a supermarket.

- The **population** – i.e. the collection of the data units – consists of all the customers of the supermarket in the given year.
- The data item is the value of a purchase (shopping) in the supermarket.
- We select a **random sample** of 64 customers. Collecting their data, we calculate the estimates as follows:
- **Sample mean:** $\bar{x} = 450$ units of money
- **Sample variance:** $s^2 = 16384$

Interval estimates for the mean μ

Interval estimates

Having got the point estimates $\bar{x} \approx \mu$ and $s^2 \approx \sigma^2$, we now wish to find **confidence intervals**, i.e. intervals

$$[\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}] \qquad \text{and} \qquad [s^2 - \Delta_{s^2}, s^2 + \Delta_{s^2}]$$

such that

the probability that

$$\mu \in [\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}] \qquad \text{and} \qquad \sigma^2 \in [s^2 - \Delta_{s^2}, s^2 + \Delta_{s^2}]$$

is ≥ 95 %, say,

or is $\geq 1 - \alpha$ in general,

where $\alpha = 5\,\%$ or $\alpha = 1\,\%$ is the **significance level.**

Interval estimates

First, having the sample mean $\bar{x}$, which is an estimate of the population mean μ, our goal is to find an interval

$$[\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}]$$

such that

the probability that $\mu \in [\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}]$ is ≥ 95 %, say.

The interval is the **confidence interval**, and the probability is $\geq 1 - \alpha$ in general, where $\alpha = 5\,\%$ (or $\alpha = 1\,\%$ or so) is the **significance level**.

Thus, given the α, our purpose is to find the $\Delta_{\bar{x}}$.

An illustrative example

By the Lindeberg-Lévy Central Limit Theorem, we approximately have

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{as} \quad n \to \infty$$

equivalently

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad \text{as} \quad n \to \infty$$

Thus, assuming that the number $n = 64$ of the customers is large enough, we assume roughly that the sample mean $\bar{x}$ follows the normal distribution already.

An illustrative example

Thus, assuming that the number $n = 64$ of the customers is large enough, we assume roughly that the sample mean $\bar{x}$ follows the normal distribution already ($\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$). We then have

$$P\left(-\delta < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq +\delta\right) = \int_{-\delta}^{+\delta} f(x)\,\mathrm{d}x = \Phi(\delta) - \Phi(-\delta)$$

and we wish this probability to be $\geq 1-\alpha = 95\,\%$, say, where $f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$ is the density of the normalized normal distribution.

Recall that $\Phi(\delta) = 1 - \Phi(-\delta)$ because the normal distribution is symmetric.

An illustrative example

We equivalently have

$$P\left(-\delta < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le +\delta\right) = P(-\delta\sigma/\sqrt{n} < \bar{x}-\mu \le +\delta\sigma/\sqrt{n}) =$$

$$= P(\bar{x}-\delta\sigma/\sqrt{n} \le \mu < \bar{x}+\delta\sigma/\sqrt{n}) =$$

$$= \Phi(\delta) - \Phi(-\delta) = \Phi(\delta) - 1 + \Phi(\delta) =$$

$$= 2\Phi(\delta) - 1$$

where $\delta > 0$ is such that

$$2\Phi(\delta) - 1 = 1 - \alpha$$

$$\delta = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

An illustrative example

Thus, knowing that

$$P(\bar{x} - \delta\sigma/\sqrt{n} \leq \mu < \bar{x} + \delta\sigma/\sqrt{n}) = 95\ \%$$

with $\bar{x} = 450$ units of money and $\delta \doteq 1.959963\ldots$ and $n = 64$ customers,

we conclude

$$\text{the unknown } \mu \in \left[450 - \frac{\sigma}{0.244995},\ 450 + \frac{\sigma}{0.244995}\right]$$

with the prescribed probability of about 95 %.

All right, the problem is that we do not know the standard deviation σ.

We therefore use the sample standard deviation s and another theorem.

Theorem

If $X_1, X_2, \ldots, X_n$ are independent and normally distributed random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

where

$\bar{X}$	is the sample mean	$\bar{X} = \sum_{i=1}^{n} X_i/n$
σ	is the standard deviation	$\sigma = \sqrt{\sigma^2}$
$\mathcal{N}(0,1)$	is the standard normal distribution	

Theorem

If $X_1, X_2, \dots, X_n$ are independent and normally distributed random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \dots, n$, then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

where

s^2	is the sample variance	$s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$
t_{n-1}^2	is Pearson's χ^2-distribution with $n-1$ degrees of freedom	

Theorem – Corollary

If $X_1, X_2, \dots, X_n$ are independent and normally distributed random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \dots, n,$ then

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

where

$\bar{X}$	is the sample mean	$\bar{X} = \sum_{i=1}^{n} X_i / n$
s	is the sample standard deviation	$s = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)}$
t_{n-1}	is Student's *t*-distribution with $n-1$ degrees of freedom	

Theorem – Corollary – Proof:

If $X_1, X_2, \dots, X_n$ are independent and normally distributed random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \dots, n$, then

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{s} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \times \frac{\sqrt{n-1}\,\sigma/s}{\sqrt{n-1}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}} \sim t_{n-1}$$

by the definition of Student's *t*-distribution

$$\frac{Z}{\sqrt{\dfrac{X_{n-1}^2}{n-1}}} \sim t_{n-1} \qquad \text{if} \quad Z \sim \mathcal{N}(0,1) \ \text{and} \ X_{n-1}^2 \sim \chi_{n-1}^2$$

An illustrative example

Thus, assuming that the purchase values $x_1, x_2, \dots, x_n$ are approximately normal, where $n = 64$ is the number of the customers, we obtain that

$$P\left(-\delta < \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq +\delta\right) = \int_{-\delta}^{+\delta} f(x)\,\mathrm{d}x = F(\delta) - F(-\delta)$$

where $f(x)$ is the density of Student's *t*-distribution with $n - 1$ degrees of freedom and $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ is the respective cumulative distribution function.

As above, we wish this probability to be ≥ 95 %, say.

An illustrative example

Analogously as above, we have

$$P\left(-\delta < \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq +\delta\right) = P(-\delta s/\sqrt{n} < \bar{x}-\mu \leq +\delta s/\sqrt{n}) =$$

$$= P(\bar{x}-\delta s/\sqrt{n} \leq \mu < \bar{x}+\delta s/\sqrt{n}) =$$

$$= F(\delta) - F(-\delta) = F(\delta) - 1 + F(\delta) =$$

$$= 2F(\delta) - 1$$

where $F(x)$ is the cumulative distribution function of Student's *t*-distribution with $n-1$ degrees of freedom, and $\delta > 0$ is such that

$$2F(\delta) - 1 = 1 - \alpha$$

An illustrative example

So we have

$$P(\bar{x} - \delta s/\sqrt{n} \leq \mu < \bar{x} + \delta s/\sqrt{n}) = 2F(\delta) - 1 =$$

$$= 1 - \alpha$$

and we find $\delta > 0$ so that

$$2F(\delta) - 1 = 1 - \alpha$$

$$\delta = F^{-1}\left(1 - \frac{\alpha}{2}\right)$$

If $\alpha = 5$ %, say, by using statistical tables or Excel, we find $\delta \doteq 1.99834054$ ...

Finally, recall that the sample average value of a purchase (shopping) is $\bar{x} = 450$, the sample standard deviation is $s = 128$, and the number of the customers is $n = 64$.

An illustrative example

We conclude that the probability that

$$\text{the unknown } \mu \in \left[450 - \frac{1.99834054 \times 128}{\sqrt{64}},\ 450 + \frac{1.99834054 \times 128}{\sqrt{64}}\right]$$

or (approx.)

$$\text{the unknown } \mu \in [450 - 31.973,\ 450 + 31.973]$$

is about $1 - \alpha = 95\ \%$.

¡¡¡ Notice we did several approximations in the chain of our considerations !!!

¡¡¡ Notice also that the quantities $\bar{x}$ and s are random variables !!!

The sample size for the confidence interval

We have the confidence interval:

$$\mu \in \left[\bar{x} - \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\sqrt{n}}, \bar{x} + \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\sqrt{n}}\right]$$

with the probability of $(1-\alpha)$.

The absolute error of the estimate is

$$\Delta = \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\sqrt{n}}$$

where s is the sample variance and F is the cumulative distribution function of Student's *t*-distribution with $n-1$ degrees of freedom.

The sample size for the confidence interval

We have the confidence interval:

$$\mu \in \left[\bar{x} - \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\sqrt{n}}, \bar{x} + \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\sqrt{n}}\right]$$

with the probability of $(1-\alpha)$.

The relative error of the estimate is

$$\delta = \frac{\Delta}{\bar{x}} = \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\bar{x}\sqrt{n}}$$

where s is the sample variance and F is the cumulative distribution function of Student's *t*-distribution with $n-1$ degrees of freedom.

The sample size for the confidence interval

Having the relative error

$$\delta = \frac{\Delta}{\bar{x}} = \frac{F^{-1}\left(1 - \frac{\alpha}{2}\right)s}{\bar{x}\sqrt{n}}$$

where s is the sample variance and F is the cumulative distribution function of Student's *t*-distribution with $n-1$ degrees of freedom,

→ find the sample size n so that the relative error

$$\delta \leq \text{some prescribed value}$$

$$\delta \leq 3\ \%, \quad \text{say}$$

The sample size for the confidence interval

Having the relative error

$$\delta = \frac{\Delta}{\bar{x}} = \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\bar{x}\sqrt{n}}$$

and assuming that $s \approx \text{const.}$, i.e. the sample variance s does not depend on n much, we obtain

$$\text{the new sample size } n = \left(\frac{F^{-1}\left(1-\frac{\alpha}{2}\right)s}{\bar{x}\delta}\right)^2$$

where δ is the upper bound of the relative error $(\delta = 3\ \%$, say) and F is the cumulative distribution function of Student's *t*-distribution with $n-1$ degrees of freedom.

Interval estimate for the variance σ^2

Interval estimate for the variance

Now, our purpose is to find an interval estimate for the population variance σ^2.

Given the significance level α, such as $\alpha = 5\,\%$ or $\alpha = 1\,\%$, our purpose is to find an interval

$$[s^2 - \Delta_{s^2}, s^2 + \Delta_{s^2}]$$

such that

the probability that $\sigma^2 \in [s^2 - \Delta_{s^2}, s^2 + \Delta_{s^2}]$ is $\geq 95\,\%$, say.

Given the α, our purpose is to find the Δ_{s^2}.

We use the next theorem.

Theorem

If $X_1, X_2, \ldots, X_n$ are independent and normally distributed random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$, then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

where

σ^2 is the (unknown) population variance
s^2 is the sample variance $\left(s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)\right)$
χ^2_{n-1} is Pearson's chi-squared distribution with $n-1$ degrees of freedom

An illustrative example

Thus, assuming that the purchase values $x_1, x_2, \dots, x_n$ are approximately normal, where $n = 64$ is the number of the customers, we obtain for any $b > a > 0$ that

$$P\left(a < \frac{(n-1)s^2}{\sigma^2} \le b\right) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$$

where $f(x)$ is the density of the chi-squared distribution with $n - 1$ degrees of freedom and $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ is the respective cumulative distribution function.

As above, we wish this probability to be ≥ 95 %, say.

An illustrative example

Let $0 < a < b$. Then, likewise as above, we have

$$P\left(a < \frac{(n-1)s^2}{\sigma^2} \leq b\right) = P\left(\frac{1}{b} \leq \frac{\sigma^2}{(n-1)s^2} < \frac{1}{a}\right) =$$

$$= P\left(\frac{(n-1)s^2}{b} \leq \sigma^2 < \frac{(n-1)s^2}{a}\right) =$$

$$= F(b) - F(a)$$

where $F(x)$ is the cumulative distribution function of the chi-squared distribution with $n-1$ degrees of freedom.

An illustrative example

Having

$$P\left(\frac{(n-1)s^2}{b} \leq \sigma^2 < \frac{(n-1)s^2}{a}\right) = F(b) - F(a)$$

we wish this probability to be ≥ 95 %, say, or $\geq 1-\alpha$ in general,

where $F(x)$ is the cumulative distribution function of the chi-squared distribution with $n-1$ degrees of freedom.

We then have to find the numbers $b > a > 0$ so that

$$F(b) - F(a) = 1 - \alpha$$

An illustrative example

Another natural condition is that the variance σ^2 should be in the centre of the interval $[(n-1)s^2/b, (n-1)s^2/a]$, i.e.

$$\frac{1}{2}\left(\frac{(n-1)s^2}{b}+\frac{(n-1)s^2}{a}\right)=\sigma^2$$

and

$$F(b)-F(a)=1-\alpha$$

which is a system of two equations with two unknowns $b>a>0$.

We, however, cannot solve the system because we do not know the variance σ^2.

An illustrative example

Therefore, having

$$P\left(\frac{(n-1)s^2}{b} \leq \sigma^2 < \frac{(n-1)s^2}{a}\right) = F(b) - F(a)$$

we only find the numbers $b > a > 0$ so that

$$F(b) = 1 - \frac{\alpha}{2} \qquad \text{and} \qquad F(a) = \frac{\alpha}{2}$$

(Then σ^2 may not be in the centre of the interval.)

For $\alpha = 5\ \%$, say, and $n = 64$, by using statistical tables or Excel, we find

$b \doteq 86.82959\ldots$ and $a \doteq 42.95027\ldots$

An illustrative example

Finally, recall that the sample variance of a purchase (shopping) is $s^2 = 16384$ and the number of the customers is $n = 64$.

We thus conclude that the probability that

$$\text{the unknown } \sigma^2 \in \left[\frac{(64-1)\times 16384}{86.82959}, \frac{(64-1)\times 16384}{42.95027}\right]$$

or (approx.)

$$\text{the unknown } \sigma^2 \in [11887.56, 24032.26]$$

is about $1-\alpha = 95\ \%$.

¡¡¡ Notice we did several approximations in the chain of our considerations !!!

¡¡¡ Notice also that the quantity s^2 is a random variable !!!

¿¿¿ Therefore, what does the 95 % probability mean ???

The variance of the sample variance

- The variance of the sample variance

  for normal distribution

The variance of the sample variance

Finally, we show the calculation of the variance $\mathrm{Var}(s^2)$ of the sample variance.

<u>Recall the last theorem:</u>

If $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent, then $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}$

<u>Recall also that,</u>

if $X \sim \chi_k$, then $\mathrm{Var}(X) = 2k$

Put together, we obtain:

$$\mathrm{Var}\left(\frac{(n-1)s^2}{\sigma^2}\right) = 2(n-1)$$

The variance of the sample variance

Having

$$\mathrm{Var}\left(\frac{(n-1)s^2}{\sigma^2}\right) = \frac{(n-1)^2}{(\sigma^2)^2}\mathrm{Var}(s^2) = 2(n-1)$$

we obtain

$$\mathrm{Var}(s^2) = \frac{2(\sigma^2)^2}{n-1}$$