Lecture 9: CLT and Confidence intervals

MVE055 / MSG810 Mathematical statistics and discrete mathematics)

Moritz Schauer

Last updated September 23, 2020, 2020

GU & Chalmers University of Technology

Central limit theorem/CLT

Recall

If $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independent, then

$$\bar{X}^{(n)} \sim N(\mu, \sigma^2/n).$$

then

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

Normal approximation of Binomial distribution

If $X_1 \dots X_n \sim \mathrm{Ber}(p)$. Then $X = \sum X_i \sim \mathrm{Bin}(n, p)$.

X is approximately normally distributed

$$X \overset{\text{approx.}}{\sim} \mathrm{N}(np, np(1-p)),$$

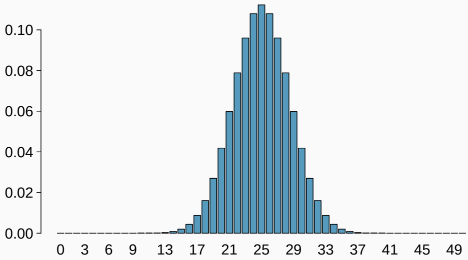
Thus again for $\bar{X}^{(n)} = \frac{1}{n} \sum X_i$,

$$\bar{X}^{(n)} \overset{\text{approx.}}{\sim} \mathrm{N}(p, p(1-p)/n),$$

or

$$\frac{\bar{X}^{(n)} - p}{\sqrt{p(1-p)/n}} \overset{\text{approx.}}{\sim} N(0, 1)$$

Normal approximation



$n = 50$, $p = 0.5$

Central limit theorem

Central limit theorem (CLT)

If $X_1, \ldots, X_n$ are independent and equally distributed random variables with expected value μ and variance $\sigma^2 < \infty$, then

$$\mathsf{P}\left(\frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \leq x\right) \to \Phi(x), \quad \text{for } n \to \infty.$$

This means,

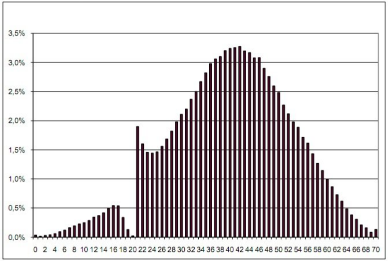
- $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ is approximatively $\mathsf{N}(\mu, \mathsf{SE}^2)$-distributed, where $\mathsf{SE} = \sigma/\sqrt{n}$ is the standard error,

for large n.

How large is large? Depends on the distribution of the X_i's.

High-school maturity exam in Poland

2.1. Poziom podstawowy



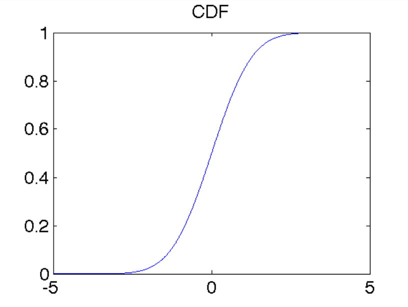
Wykres 1. Rozkład wyników na poziomie podstawowym

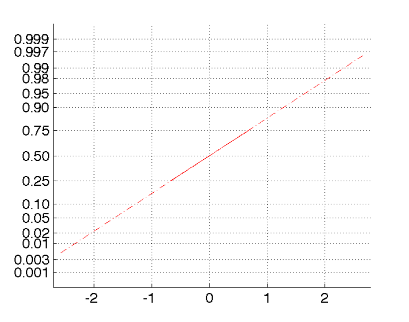
Histogram showing the distribution of scores for the obligatory Polish language test. “The dip and spike that occurs at around 21 points just happens to coincide with the cut-off score for passing the exam"

http://freakonomics.com/2011/07/07/another-case-of-teacher-cheating-or-is-it-just-altruism/

Normal probability plot

Normal probability plot





The standard normal distribution function (cdf) is

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

It is possible to transform the scaling on the y-axis so that F becomes a straight line in the plot.

Normal probability plot

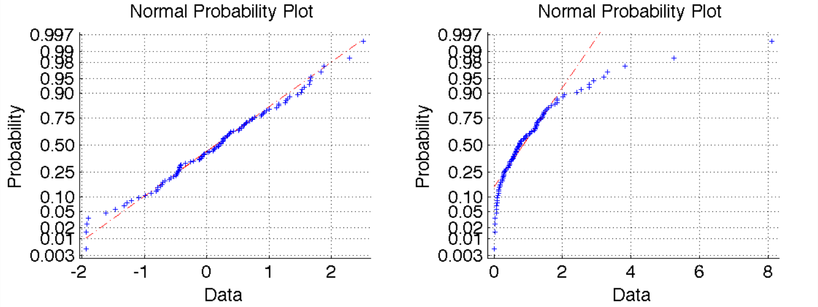
Suppose we have the data $x_1, \ldots, x_n$ and want to see if a normal distribution is a reasonable model for the data. We can use the normal probability plot for this.

First we compute the *empirical distribution function*

$$F^*(x) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \leq x)}_{\text{proportion of values smaller than } x}$$

We plot the points $F^*(x_j)$ in the normal probability diagram, and if the data is normally distributed, these points should lie along a straight line.

Normal probability plot



Example: left normally distributed data and and right exponentially distributed data in normal probability diagram. In Matlab: `normplot`.

Confidence interval

Confidence interval

Confidence interval

If $X_1, \ldots, X_n$ i.i.d random variables with distribution depending on a parameter θ, with θ_0 being the unknown value. A $100(1-\alpha)\%$ confidence interval for θ with confidence level $1-\alpha$ is an interval $I_\theta = [A, B]$ computed from the data such that

$$\mathsf{P}(A \leq \theta_0 \leq B) = 1 - \alpha.$$

Confidence interval for parameter μ of a normal distribution

Let $X_1, \ldots, X_n$ be independent $\mathsf{N}(\mu, \sigma^2)$.

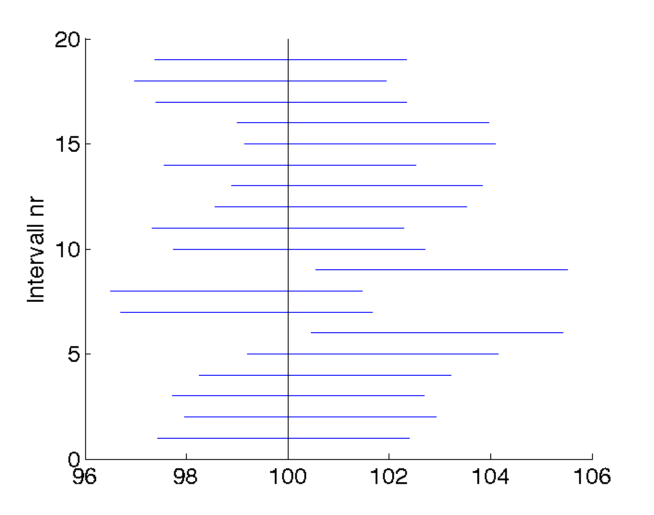
Known variance σ^2

$$I_\mu = \left(\bar{X}^{(n)} - 1.96 \frac{\sigma}{\sqrt{n}},\ \bar{X}^{(n)} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

is a confidence interval for μ with confidence level 95%.

Here 1.96 is the $0.975 = (100 - 2.5)\%$ quantile of $Z \sim N(0, 1)$:

$$\mathrm{P}(-1.96 < \frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95.$$



20 confidence intervals for μ, that where each constructed from 20 different samples of 10 N$(100, 16)$-observations.

- $[A, B]$ is a random interval, because A and B are random variables (transformations of the random variables $X_1, \ldots, X_n$).
- Interpretation. Let $\mathbf{x}_1 = (x_{11}, \ldots, x_{n1}), \mathbf{x}_2 = (x_{12}, \ldots, x_{n2}), \ldots$ be repeated measurements of $X_1, \ldots, X_n$. If we make the confidence interval for θ based on every $\mathbf{x}_i$, then $100(1-\alpha)\%$ of these intervals cover the true value θ_0.

Table 2: Quantiles of the normal distribution

Table gives $\mathsf{P}(X > \lambda_\alpha) = \alpha$ for $X \sim \mathsf{N}(0, 1)$

α	.1	.05	.025	.01	.005	.001	...	.00001
λ_α	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	...	4.2649

$t(n)$-distribution

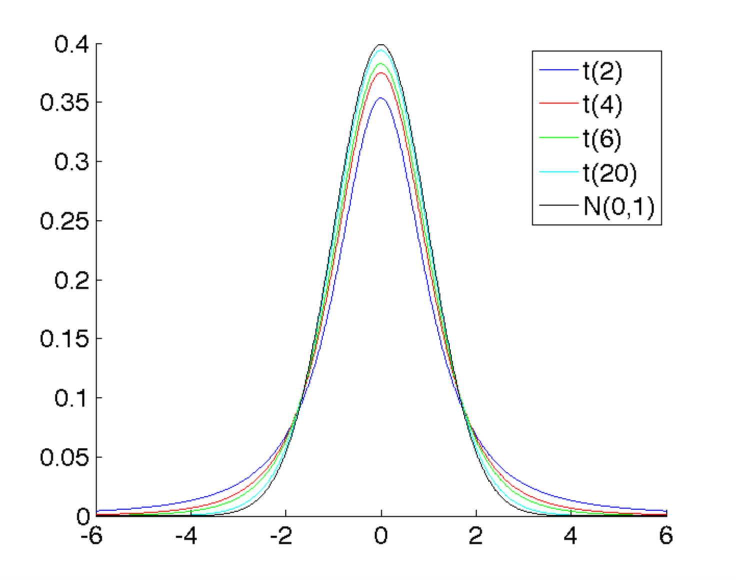


Table 3: Quantiles of the t-distribution

Table gives $\mathsf{P}(X > t_\alpha(f)) = \alpha$ for $X \sim t(f)$.

α	.1	.05	.025	.01	.001
$t_\alpha(1)$	3.0777	6.3138	12.706	31.820	318.31
$t_\alpha(2)$	1.8856	2.9200	4.3027	6.9646	22.327
$t_\alpha(3)$	1.6377	2.3534	3.1824	4.5407	10.215
$t_\alpha(4)$	1.5332	2.1318	2.7764	3.7469	7.1732
$t_\alpha(5)$	1.4759	2.0150	2.5706	3.3649	5.8934
$t_\alpha(6)$	1.4398	1.9432	2.4469	3.1427	5.2076
$t_\alpha(7)$	1.4149	1.8946	2.3646	2.9980	4.7853
$t_\alpha(8)$	1.3968	1.8595	2.3060	2.8965	4.5008
$t_\alpha(9)$	1.3830	1.8331	2.2622	2.8214	4.2968
$t_\alpha(10)$	1.3722	1.8125	2.2281	2.7638	4.1437
$t_\alpha(15)$	1.3406	1.7531	2.1314	2.6025	3.7328
$t_\alpha(20)$	1.3253	1.7247	2.0860	2.5280	3.5518
$t_\alpha(30)$	1.3104	1.6973	2.0423	2.4573	3.3852
$t_\alpha(40)$	1.3031	1.6839	2.0211	2.4233	3.3069
$t_\alpha(60)$	1.2958	1.6706	2.0003	2.3901	3.2317
$t_\alpha(\infty)$	1.2816	1.6449	1.9600	2.3263	3.0902

Confidence interval for μ of a normal distribution

Let $X_1, \ldots, X_n$ be independent $\mathsf{N}(\mu, \sigma^2)$.

Known variance σ^2

$$I_\mu = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is a confidence interval for μ with confidence level $1-\alpha$.

Unknown variance σ^2

$$I_\mu = \left(\bar{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$$

is a confidence interval for μ with confidence level $1-\alpha$. Here s^2 is the sample variance and $t_{\alpha/2}(n-1)$ are the $(1-\alpha/2)$-quantiles of the $t(n-1)$-distribution.

Quiz

$x_1, \ldots, x_n$ are a sample of i.i.d observations with distribution depending on a parameter θ.

Winnie computes a $95\,\%$ confidence interval for θ.

Piglet computes a $90\,\%$ confidence interval for θ using the same data.

Which interval is smallest? Piglet's $90\,\%$ confidence interval.

Confidence interval for μ from central limit theorem

- By the CLT the sample mean $\bar{X}^{(n)}$ is approximatively $\mathsf{N}(\mu, \sigma^2/n)$-distributed for large n.
- If we have a sample with known variance σ^2,

$$I_\mu = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

  is a confidence interval for the mean μ with confidence level $1 - \alpha$.
- If σ is not known we can estimate it by S. For the estimate to be good, it is important that n is large and the distribution for X_i is not too heavy tailed.
- Since n is big, we use $t_{\alpha/2}(n-1) \approx z_{\alpha/2}$, so if σ is unknown, we use

$$I_\mu = \left(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right).$$

Confidence interval for σ^2 for the normal distribution

Confidence interval for σ

If $X_1, \ldots, X_n$ are independent $\mathrm{N}(\mu, \sigma^2)$ then a confidence interval with confidence level $1-\alpha$ for σ is

$$I_\sigma = \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}} \right).$$

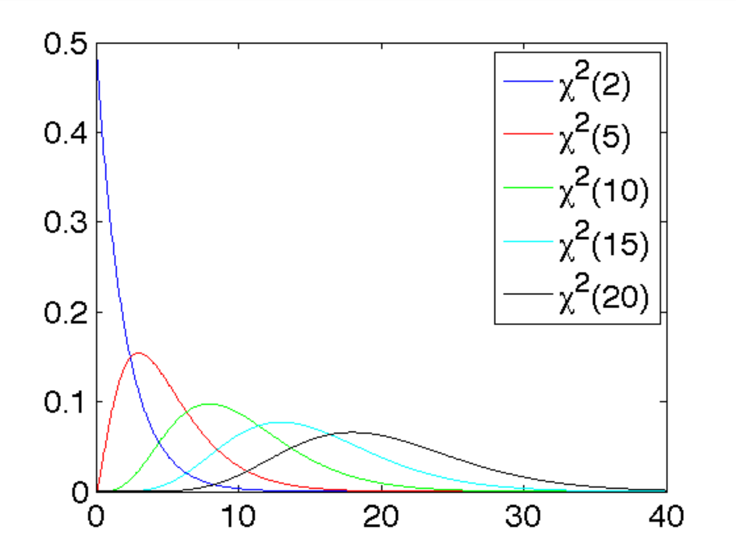
Here $\chi^2_{\alpha/2}(n-1)$ are the $(1-\alpha/2)$-quantiles of the $\chi^2(n-1)$ distribution.

If Z_i are independent $\mathrm{N}(0,1)$, it holds

$$\sum_{i=1}^{n} Z_i^2$$

is $\chi^2(n)$-distributed

$\chi^2(n)$-distribution



Confidence interval for σ^2 for the normal distribution

Confidence interval for σ

If $X_1, \ldots, X_n$ are independent $\mathrm{N}(\mu, \sigma^2)$ then a confidence interval with confidence level $1 - \alpha$ for σ is

$$I_\sigma = \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}} \right).$$

Important: In contrast to the confidence interval for the expected value, the confidence interval for the variance is very sensitive to deviations from the normal distribution.

Summary

For a confidence interval

- for the expected value μ
  - of the normal distribution: Slide: confidence interval for μ of a normal distribution
    - Known σ or large n: use confidence interval based on normal quantiles.
    - Small n and unknown σ: use quantiles based on t-distribution.
  - of a general distribution
    - Large n: use confidence interval based on normal quantiles (valid approximation by CLT). Slide: Confidence interval for μ from central limit theorem.
- for the variance σ^2
  - of the normal distribution: Slide: Confidence interval for σ^2 for the normal distribution.