Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Sixth lecture, November 18, 2024

Statistical testing

Statistical research is collecting, organizing, analyzing, and interpreting data.

Statistical models are mathematical and are based on probability theory.

- In probability theory, if we know the law of a random variable, then we are easily able to draw an i.i.d. sample from the distribution, compute the probabilities of different events, compute the expected value, variance, higher moments, etc.
- In statistics, we are usually given a finite sample, and we are interested in making inferences about the distribution and population parameters such as the expected value, variance, higher moments, etc. We are also interested in assessing the *uncertainty* of the population parameters (confidence interval).
  - If the data (approximately) follows a distribution which we are able to identify, we can make use of the properties of that distribution from probability theory to assess the uncertainty (parametric tests).
  - It is also important to discuss statistical methods for data which does not clearly follow a known distribution (non-parametric tests).
- Correlations between variables, regression models, . . .

Population and sample

- In statistical analysis, a population is a collection of all the people, items or events about which one wants to make inferences. (For example, university students in Germany.)
- In statistical analysis, a sample is a subset of the population (i.e., the people, items or events) that one collects and analyzes to make inferences. (For example, 200 randomly chosen university students.)
- In statistical analysis, an observation is an element of the sample. (For example Helen, a student at FU Berlin.)

In statistical research, data consists of the values of selected variables that describe the observations. The data points (the values of the selected variables) can also be called observations.

Examples:

- temperature, height, blood pressure (continuous quantitative variables)
- gender, eye color (categorical qualitative variables)
- clothing size (s,m,l) (ordinal quantitative variable)

Statistical research projects

Statistical research projects can usually be conducted in the following steps:

1. Setting of the research topic and the relevant research questions. Research questions should be defined precisely.
2. Defining of the population and interesting variables.
3. Planning of the sample collection. Collected sample must represent the population!
4. Collection of the sample.
5. Organization of the sample.
6. Description of the variables and the sample, descriptive statistics and visualization.
7. Inference based on statistical analysis. Model assumptions have to be tested separately!
8. Critical evaluation of the analysis. Possible errors and weaknesses have to be reported.
9. Communication of the research and findings.

Different statistical studies

Statistical research projects can be conducted in several different ways. Research questions, population, goals, and resources all have an effect on the choice of the methods.

- In observational research, observations are made without changing any existing conditions. For example, temperature is measured or the lung cancer risk of smokers is compared to the lung cancer risk on non-smokers.
- In controlled experiments, the effect of one variable to another is examined by controlling existing conditions. For example, the effect of allergy medicine is compared to the effect of placebo by randomizing patients to two groups.

Different statistical studies

- In simulations, mathematical modeling is used to mimic natural conditions or processes. For example, the spread of the Ebola virus is predicted by applying computer simulations or the safety of a new car model is tested using crash test dummies.
- In surveys, the goal is to find a representative sample of the population and get answers to some particular questions. For example, opinion polls are used in order to predict election results, or health related questionnaires are used to assess the health of university students.

Descriptive statistics

Descriptive statistics and inference

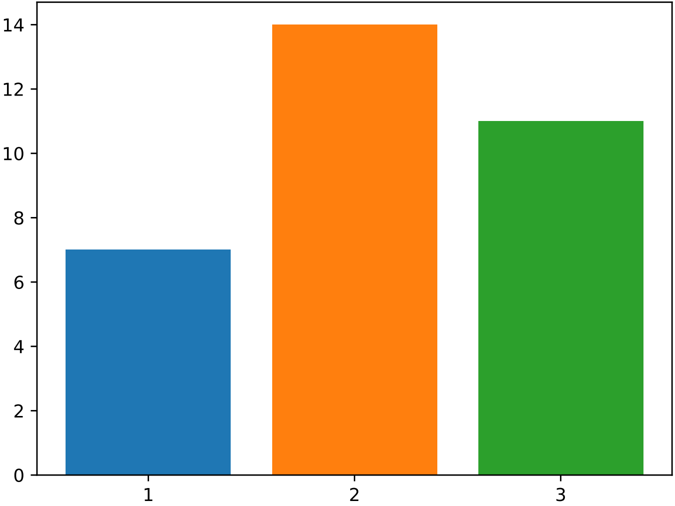
Descriptive statistics provide a concise summary of the data. The summary may either be numerical or graphical or both. Descriptive statistics may consist of, for example, numerical tables, average values, deviations, summaries and visualizations.

Statistical inference draws conclusions about the population using data. Statistical inference is based on mathematical modeling and probabilities. Inferential statistical analysis includes, for example, estimation and statistical testing.

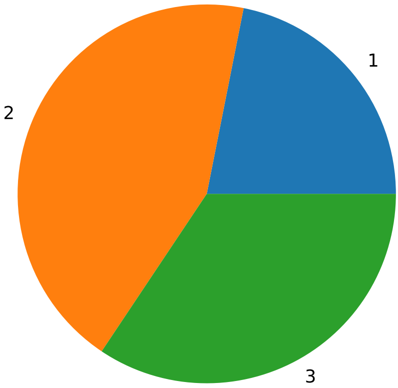
Visualization

- Discrete variable: bar plot, pie chart
- Continuous variable: box plot, histogram
- Bivariate: scatter plot

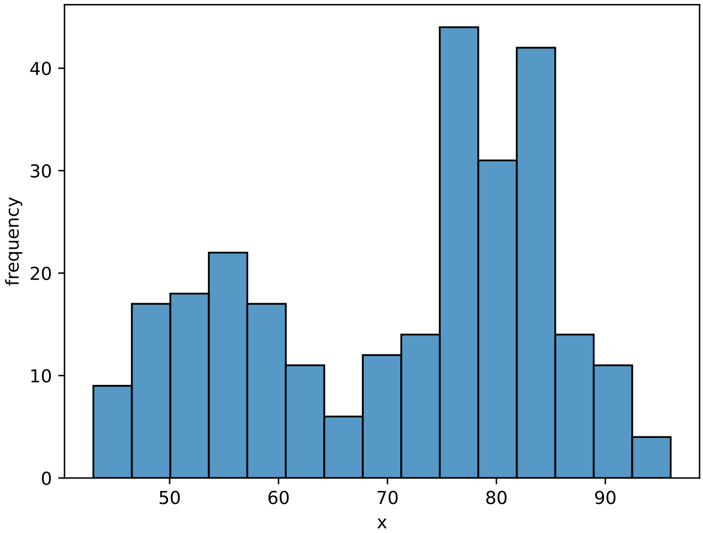
Bar plot



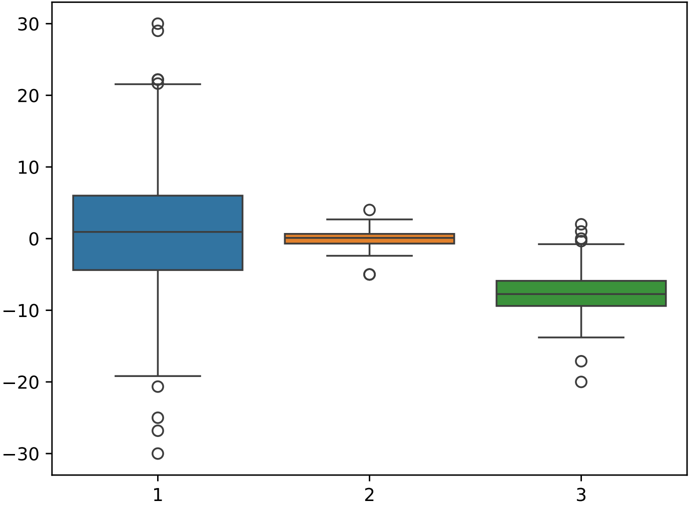
Pie chart



Histogram



Box plot



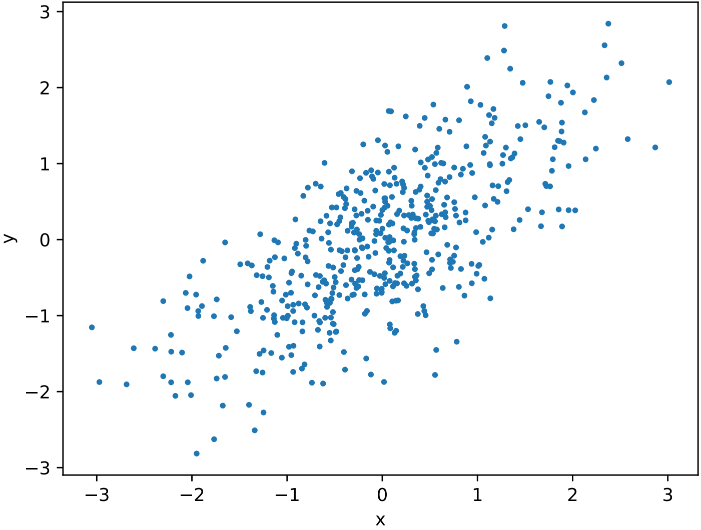
Box plot

In a box plot (sometimes also called a "box-and-whisker plot"), the box contains 50% of the data. The line in the middle is the sample median.

Let Q_1 and Q_3 denote the 25 and 75 sample percentiles. By default, the lower whisker is at the lowest data point above $Q_1 - 1.5(Q_3 - Q_1)$ and the upper whisker is at the highest data point below $Q_3 + 1.5(Q_3 - Q_1)$.

Outlying points are marked using circles.

Scatter plot



Location

Mean, median, and mode are commonly used measures of location.

Let $x_1, \dots, x_n$ be i.i.d. observations of a random variable x. Then the sample mean

$$\overline{x} = \overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

estimates the expected value $\mathbb{E}[x] = \mu$ of the variable x.

The population median m_x of a random variable x is the value with the property

$$\mathbb{P}(x < m_x) \leq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(x \leq m_x) \geq \frac{1}{2}.$$

Let $y_1 < y_2 < \cdots < y_n$ be the *ordered* values of the data. The sample median is the middle value of the ordered values. If the number of observations is even, then the sample median is the average of the two middle elements. The sample median estimates the population median.

The sample mode is the value $x_1, \dots, x_n$ that has the highest frequency. Mode estimates a value of a qualitative variable or discrete quantitative variable that has the highest probability.

Percentiles

Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x. Let $y_1 < y_2 < \cdots < y_n$ be the *ordered* values of the data. Then the sample β percentile, $0 < \beta < 100$, is the data point y_k, where k is the closest integer that is larger than or equal to $\beta \cdot (n/100)$. The population β percentile of a random variable x is the value β_x with the property

$$\mathbb{P}(x < \beta_x) \leq \frac{\beta}{100} \quad \text{and} \quad \mathbb{P}(x \leq \beta_x) \geq \frac{\beta}{100}.$$

Numerical example

Consider the sample

$$\{3, 1, 2, 3, 7, 8, 3, 4, 4, 6\}.$$

The sample mean is

$$\overline{x} = \frac{1}{10} \cdot (3 + 1 + 2 + 3 + 7 + 8 + 3 + 4 + 4 + 6) = \frac{41}{10} = 4.1.$$

The sample median is

$$\hat{m}_x = \frac{3 + 4}{2} = \frac{7}{2} = 3.5.$$

The sample mode is 3.

Deviation/scatter

Variance, standard deviation, median absolute deviation (MAD), and range are commonly used measures of deviation/scatter.

Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x. The sample variance

$$s^2 = s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

estimates the population variance $\mathbb{E}[(x - \mathbb{E}[x])^2] = \sigma^2$.

The sample standard deviation is the square root of the sample variance:

$$s = s_n = \sqrt{s_n^2}.$$

Chebyshev's inequality

Let x be a random variable with finite expected value $\mathbb{E}[x] = \mu$ and finite variance $\mathbb{E}[(x - \mathbb{E}[x])^2] = \sigma^2$. Let $k > 1$. Then

$$\mathbb{P}(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

If $k = 2$, then $1 - \frac{1}{k^2} = 75\%$.
If $k = 3$, then $1 - \frac{1}{k^3} \approx 88.9\%$.

In practice, the expected value and variance must be estimated. Chebyshev's inequality can be used to evaluate the outlyingness/rareness of a single observation:

- If an observation lies further away than two times the standard deviation of the sample mean, it is considered rare.
- If an observation lies further away than three times the standard deviation of the sample mean, it is considered very rare.

These definitions are based on Chebyshev's inequality.

Rare observation under normality

If it is known that observations follow a Gaussian distribution, then the probability for a data point lying within one standard deviation of the sample mean is $\approx 68\%$. The probability for a data point lying within two standard deviations of the sample mean is $\approx 95\%$ and for three standard deviations it is $\approx 99.7\%$.

Median absolute deviation and range

Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x and let m_x be the sample median. Then the median absolute deviation (MAD) is the median of the sample $|x_1 - m_x|$, $|x_2 - m_x|$,$\ldots$,$|x_n - m_x|$.[†]

Let Max_x be the largest data point and Min_x the smallest data point. Then the sample range is the interval $[Min_x, Max_x]$ and the length of the range is $Max_x - Min_x$.

[†] To make the MAD comparable with the standard deviation, one often multiplies the MAD with a scale factor k depending on the distribution. For example, for normally distributed data, $k = \frac{1}{\Phi^{-1}(3/4)} \approx 1.4826$. (In fact, this is the default scaling used in R, but for example the `scipy.stats.median_abs_deviation` function uses $k = 1$ by default.)

Numerical example

Consider the sample

$$\{3, 1, 2, 3, 7, 8, 3, 4, 4, 6\}.$$

The sample mean was calculated above and it was 4.1. The sample variance is

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 4.1)^2 = 4.9888\ldots$$

and the sample standard deviation is $s_n = \sqrt{s_n^2} = \sqrt{4.9888\ldots} = 2.233\ldots$ The sample median was calculated above and it was 3.5. Mean absolute deviation:

$$\begin{aligned} MAD = \mathrm{median}\{&|3-3.5|, |1-3.5|, |2-3.5|, |3-3.5|, |7-3.5|, \\ &|8-3.5|, |3-3.5|, |4-3.5|, |4-3.5|, |6-3.5|\} \\ = 1.& \end{aligned}$$

The range can be calculated from the minimum and maximum values of the sample:

$$[\min(x), \max(x)] = [1, 8].$$

The length of the range is $8 - 1 = 7$.

Skewness

Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x. Then the sample skewness coefficient is

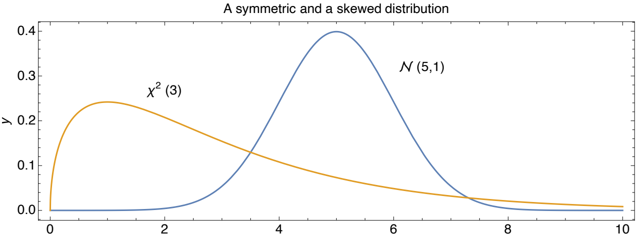
$$v = \frac{m_3}{s_n^3},$$

where

$$m_3 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^3.$$

Sample skewness coefficient estimates the population value

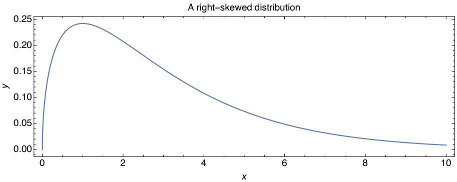
$$\mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

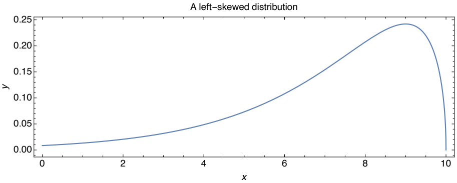


Skewness

- If the skewness coefficient $v > 0$, then the distribution is skewed to the right (positively skewed distribution).
- If the skewness coefficient $v < 0$, then the distribution is skewed to the left (negatively skewed distribution).

Usually[†], a positively (right) skewed distribution has a long right tail and the mass of the distribution is concentrated on the left. A negatively (left) skewed distribution has a long left tail and the mass of the distribution is concentrated on the right.


A right–skewed distribution


A left–skewed distribution

[†] Multimodal distributions or asymmetric distributions which have one long tail but the other tail is fat can break this rule of thumb.

Skewness

Alternative skewness coefficient v_2: Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x. Then also

$$v_2 = \frac{\overline{x} - m_x}{s_n}$$

is a measure of skewness. (Here, m_x denotes the sample median.)

For symmetric distributions, the sample mean and sample median estimate the same population value.

Kurtosis

Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x. Then the sample kurtosis coefficient is

$$k = \frac{m_4}{s_n^4} - 3,$$

where

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4.$$

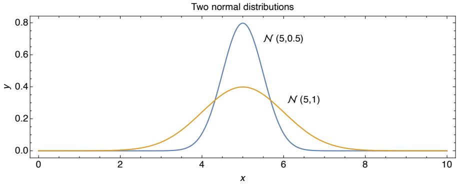
The sample kurtosis coefficient estimates the population value

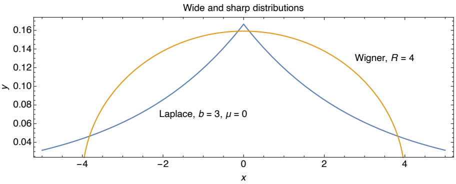
$$\mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4 - 3\right].$$

Kurtosis

A random variable with normal distribution has kurtosis value 0. If the kurtosis value is $k > 0$, then the distribution is more peaked than normal distribution. If $k < 0$, then the distribution is less peaked than normal distribution.

A distribution with large kurtosis value (leptokurtic) typically has a sharp peak and thick tails, while less peaked distributions (platykurtic) have round peaks and thin tails.


Two normal distributions


Wide and sharp distributions

Linear dependence and correlation

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y). Then the sample covariance

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

estimates the population covariance $\mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \sigma_{xy}$, and

$$\hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

estimates the Pearson correlation coefficient

$$\rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

The Pearson correlation coefficient measures numerically the linear dependence of two random variables. The coefficient is always in the interval $[-1, 1]$.

Confidence interval

Confidence interval

In statistics, we often have a sample and we estimate the value of some parameter using the observations. For example, we estimate the expected value by calculating the sample mean or we estimate the population skewness coefficient by calculating the corresponding sample estimate. The simple estimate, however, still gives us quite little information. We cannot directly evaluate how good our estimate is. It would be nice to know a bit more. That is why an estimate of a parameter is often presented with a corresponding confidence interval.

Confidence interval

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. A confidence level for a confidence interval determines the probability that the confidence interval produced will contain the true parameter value.

Confidence interval

Let x be a random variable from a distribution P_x. Let θ be a parameter of the distribution P_x and let $\hat{\theta}$ be an estimate of the parameter. (For example, θ could be the population mean, population standard deviation, population median, etc., and $\hat{\theta}$ would be the corresponding sample mean, sample standard deviation, sample median, etc.)

We say that an interval (l, u) is a confidence interval for the estimate $\hat{\theta}$ at confidence level $(1-\alpha)$ if the following holds: *before* the sample is generated, the random range (l, u) corresponding to $\hat{\theta}$ includes the true parameter value θ with probability $p = 1 - \alpha$.

After the sample has been generated and the estimate $\hat{\theta}$ and the corresponding confidence interval (l, u) has been calculated, the confidence interval either includes or does not include the true parameter value θ. If 100 samples are generated, the corresponding 100 estimates $\hat{\theta}$ and the corresponding 100 confidence intervals are calculated, then $\approx (1-\alpha) \cdot 100$ of the confidence intervals include the true parameter value and $\approx \alpha \cdot 100$ do not include it.

Bootstrap confidence intervals

Let $\{x_1, \ldots, x_n\}$ denote i.i.d. observations from the distribution P_x. Let θ be a parameter of the distribution P_x. (For example, θ could be the population mean, population standard deviation, population median, etc.) Let $\hat{\theta}$ be an estimate of the parameter θ calculated from the sample $\{x_1, \ldots, x_n\}$. (For example, $\hat{\theta}$ would be the sample mean, sample standard deviation, sample median, etc., corresponding to θ.)

An estimate for the confidence interval (l, u) can now be obtained by resampling as follows:

1. Select n data points randomly with replacement from the original sample $x_1, \ldots, x_n$. Each data point can be selected once, multiple times, or not at all. (Note that the sample size of the new sample is the same as the sample size of the original sample.)
2. Calculate a new estimate for the parameter θ from the new sample formed in the previous step.

(Continued on the next slide.)

3. Repeat the steps 1–2 k times and order the obtained estimates from the smallest to the largest. Include also the original estimate $\hat{\theta}$.
4. Calculate an estimate for a $(1-\alpha)\cdot 100\%$ confidence interval by selecting a lower bound l that is smaller than (or equal to) $(1-\frac{\alpha}{2})\cdot 100\%$ of the ordered estimates and an upper bound u that is larger than (or equal to) $(1-\frac{\alpha}{2})\cdot 100\%$ of the estimates.

Example

Assume that we compute 999 bootstrap estimates. Then, in total, there are 1000 estimates – the original one and the 999 new ones. Now, an estimated 90% confidence interval (l, u) is obtained by choosing the 50^{th} ordered estimate as l and the 951^{st} estimate as u.

An estimate for the 95% confidence interval (l, u) is obtained by choosing the 25^{th} estimate as l and the 976^{th} estimate as u.

On the accuracy of the bootstrap confidence interval:

- The larger the original sample size, the better the confidence interval.
- The larger the number k of bootstrap samples, the better the confidence interval.

Exact confidence intervals

Bootstrap confidence intervals are nowadays easy to calculate and they have the advantage of being distribution free.

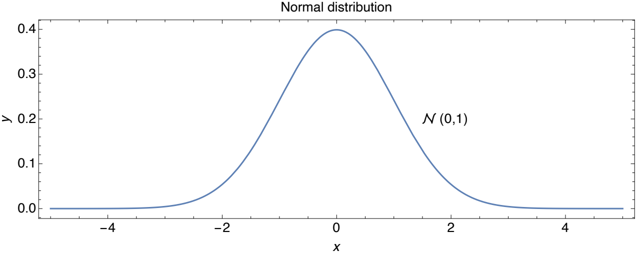
However, when the type of distribution is known, also exact confidence intervals can be calculated. It is possible to obtain exact confidence intervals for the parameters of the normal distribution or for the parameter of the Bernoulli distribution, for example.

Confidence interval, normal distribution

A random variable with normal distribution has a probability density function (PDF) of the form

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right).$$

The normal distribution has two parameters: the mean μ and the variance σ^2.



Example (Confidence interval for population mean μ of a normal i.i.d. sample with *known variance* σ^2)

Let $x_1, \ldots, x_n$ be i.i.d. copies of $x \sim \mathcal{N}(\mu, \sigma^2)$. Suppose that we are interested in finding a level $(1-\alpha)$ confidence interval for the population mean μ given the sample $x_1, \ldots, x_n$. If we know the population variance σ^2, then we can use the whitening transform

$$Z = \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{1}$$

and deduce that the $(1-\alpha)$ confidence interval for the population mean is given by

$$\left(\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \tag{2}$$

where $z_{\alpha/2} = \Phi^{-1}(1-\frac{\alpha}{2})$ is the $(1-\alpha/2)\cdot 100$ percentile of the standard normal distribution. E.g., if $\alpha = 0.05$, then $z_{0.025} = \Phi^{-1}(0.975) \approx 1.96$.

In practice, the population standard derivation must be *approximated* by the sample standard deviation s_n. If n is large (e.g., $n > 30$), then simply approximating $\sigma \approx s_n$ in (1)–(2) may lead to a reasonable approximation of the CI. However, simply replacing σ by s_n makes the test statistic (1) non-Gaussian in general. A better method is to note that $\frac{\overline{x}_n - \mu}{s_n/\sqrt{n}}$ follows *Student's t-distribution*.

Confidence interval, mean of normal distribution

Let $x_1, \ldots, x_n$ be i.i.d. copies of $x \sim \mathcal{N}(\mu, \sigma^2)$. We are interested in finding a level $(1-\alpha)$ confidence interval for the population mean μ given the sample $x_1, \ldots, x_n$. In practice, the population standard deviation σ must be approximated by the sample standard deviation s_n. Substituting the population standard deviation σ by the sample standard deviation s_n in (1) yields the t-statistic

$$t_{n-1} := \frac{\overline{x}_n - \mu}{s_n/\sqrt{n}}$$

and we say that t_{n-1} has Student's t-distribution with $n-1$ degrees of freedom. Then the $(1-\alpha)$ confidence interval for the population mean μ is given by

$$\left(\overline{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \overline{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right),$$

where $t_{n-1,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$ percentile of the t_{n-1} distribution. E.g., if $n = 10$ and $\alpha = 0.05$, then $t_{9,0.025} = F_{t_9}^{-1}(0.975) = 2.262$, where $F_{t_9}^{-1}$ is the quantile function of t_9.

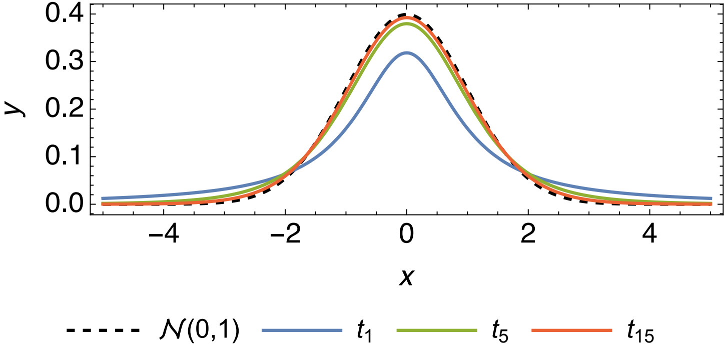


Figure: Student's t-distributions with different degrees of freedom. The t-distribution has heavier tails than the standard Gaussian distribution. As the degrees of freedom increase, the t-distributions tend to the standard Gaussian distribution.

Confidence interval, variance of normal distribution

Let $x_1, \ldots, x_n$ be i.i.d. copies of $x \sim \mathcal{N}(\mu, \sigma^2)$. We are interested in finding a level $(1-\alpha)$ confidence interval for the population variance given the sample $x_1, \ldots, x_n$. It is assumed that the population mean μ is also unknown. The statistic

$$Q = \frac{(n-1)s_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

has the χ^2 distribution with $n-1$ degrees of freedom, i.e., $Q \sim \chi^2(n-1)$. Then the level $(1-\alpha)$ confidence interval for the variance of a normal distribution can be given as

$$\left(\frac{(n-1)s_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s_n^2}{\chi^2_{n-1,1-\alpha/2}}\right),$$

where $\chi^2_{n-1,\alpha/2}$ is the $(1-\alpha/2)\cdot 100$ percentile of the $\chi^2(n-1)$ distribution. Similarly, $\chi^2_{n-1,1-\alpha/2}$ is the $(\alpha/2)\cdot 100$ percentile of the $\chi^2(n-1)$ distribution. E.g., if $n = 10$ and $\alpha = 0.05$, then $\chi^2_{9,0.025} = F^{-1}_{\chi^2(9)}(0.975) \approx 19.02$ and $\chi^2_{9,0.975} = F^{-1}_{\chi^2(9)}(0.025) \approx 2.70$.

χ^2 distribution

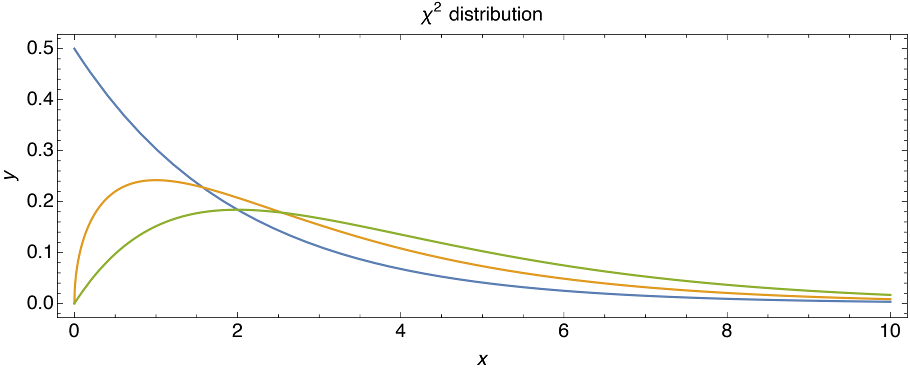


Figure: χ^2 distribution with different degrees of freedom.

Confidence interval, parameter p of Bernoulli distribution

Let $\{x_1, \ldots, x_n\}$ denote i.i.d. observations of a random variable x. Assume that $\mathbb{P}(x_i = 1) = p$ and $\mathbb{P}(x_i = 0) = 1 - p$. Then $x \sim \mathrm{Ber}(p)$, with expected value $\mathbb{E}[x] = p$ and $\mathbb{E}[(x - \mathbb{E}[x])^2] = p(1-p)$. An unbiased estimate of the expected the expected value p is the sample mean

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

If n is large, the level $(1 - \alpha)$ confidence interval for the mean p of the Bernoulli distribution can be given as

$$\left(\hat{p} - z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right),$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2) \cdot 100$ percentile of the standard normal distribution $\mathcal{N}(0,1)$.

There exist several alternative estimates for the confidence interval for the mean of the Bernoulli distribution. If the sample size is small, one can try the Wilson score interval, for example.

Numerical example, confidence intervals

The masses of Brand X cookie packages are approximately normally distributed with expected value μ. The randomly chosen packages were weighted and the following data (measured in grams) was obtained: 397.3, 399.6, 401.0, 392.9, 396.8, 400.0, 397.6, 392.1, 400.8, 400.6.

The mean of the masses is $397.87g$ and the sample standard deviation is

$$s = \sqrt{\frac{1}{10-1}\sum_{i=1}^{10}(x_i - 397.87)^2} \approx 3.2128.$$

As we saw above, the 97.5% percentile of the Student's t-distribution with $10 - 1 = 9$ degrees of freedom is $t = 2.262$. The 95% confidence interval for the mean masses of the cookie packages is

$$\left(\overline{x} \pm t\frac{s}{\sqrt{n}}\right) = \left(397.87g \pm 2.262 \cdot \frac{3.2128g}{\sqrt{10}}\right) = (395.6g, 400.3g).$$