02403 Introduction to Mathematical Statistics

Lecture 4: Sampling distributions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Agenda

1. Simulations of experiments
2. The general framework
3. The t-distribution
4. The F-distribution
5. Sampling distributions in statistics

Overview

Example: Average and variance of normal sample

Assume that we plan a study whith 5 observations. We also assume that he mean and variance in the population is $\mu = 10$ and $\sigma^2 = 2$, what is the distribution of the average and the empirical variance uder these assumptions?
There is (at least) two ways to and the question

1: Go through the derivation and obtain the distribution functions

2: Do the experiment a large number of times (e.g. 10,000 times) on your computer and find the empirical distribution.

Do it in Python..

Overview

The framework of *statistisk inferens*

From eNote, Chapter 1:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observationsenhed**)
- The *statistical population* consists of all possible “measurements” on each *observational unit* (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Stikprøve**)

Language and concepts:

- μ and σ are parameters that describe the populationen
- $\bar{y}$ is an *estimate* of μ (an actual outcome, a number)
- $\bar{Y}$ and S^2 are *estimatorers* of μ and σ^2 (these are random variables)
- The concept *'statistic(s)'* is used for both

The aim

In lecture 1 we saw a number of summary statistics, we now assume that

$$Y_i \sim N(\mu, \sigma^2), \quad \text{and iid.}$$

In this and the next lecture we will answer the following questions

- What is the distribution of $\bar{Y}$? Lecture 3!
- What is the distribution of S^2? Lecture 3!
- What is the distrubution of $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$? Today
- If we calculated observed variances from two different groups, what is then the distribution of $\frac{S_1^2}{S_2^2}$? Today

Overview

The t-distribution

Definition

If $Z \sim N(0,1)$ and $Q \sim \chi^2(n)$ with Z and Q are independent then

$$T = \frac{Z}{\sqrt{Q/n}}$$

follows a t-distribution with n degrees of freedom.

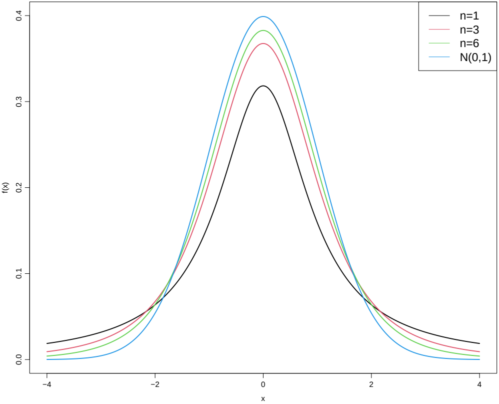
Theorem

The probability function for a t-distribution is given by

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad ; t \in \mathbb{R}$$

where n is the number of degrees of freedom and $\Gamma()$ is the Gamma function.

The t-distribution



The t-distribution as a sampling distributution

Let $Y_1, \ldots, Y_n$ be iid. $\sim N(\mu, \sigma^2)$ then

$$T = \frac{\overline{Y} - \mu}{\sqrt{S^2/n}}$$

follows a t-distribution with $n-1$ degrees of freedom.

The t-distribution as a sampling distributution - proof

We need to show that T can be written as a standard normal distribution divided the square root of by a χ^2-distributed random variable with $n-1$ degrees of freedom, and that the denominator and the enumerator are independent

1: We have already shown that $\bar{Y}$ and S^2 are independent

2: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $Q = \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$

3:

$$T = \frac{\frac{1}{\sigma/\sqrt{n}}(\overline{Y}-\mu)}{\sqrt{\frac{1}{\sigma^2/n}\frac{n-1}{n-1}S^2/n}} = \frac{Z}{\sqrt{Q/(n-1)}}$$

where $Z \sim N(0,1)$ and hence T follows a t-distribution with $n-1$ degrees of freedom.

Example: confidence interval

Let $Y_1, \ldots, Y_n$ be iid. $\sim N(\mu, \sigma^2)$, find d such that $(0 < \alpha < 0.5)$

$$1 - \alpha = P(\overline{Y} - d \cdot S < \mu < \overline{Y} + d \cdot S)$$

Example: confidence interval

Let $Y_1, \dots, Y_n$ be iid. $\sim N(\mu, \sigma^2)$, find d such that $(0 < \alpha < 0.5)$

$$1 - \alpha = P(\overline{Y} - d \cdot S < \mu < \overline{Y} + d \cdot S)$$

Answer:

$$\begin{aligned} P(\overline{Y} - d \cdot S < \mu < \overline{Y} + d \cdot S) =& P\left(-d < \frac{\overline{Y} - \mu}{S} < d\right) \\ =& P\left(-d\sqrt{n} < \frac{\overline{Y} - \mu}{S/\sqrt{n}} < d\sqrt{n}\right) \\ =& F_T(d\sqrt{n}) - F_T(-d\sqrt{n}) = 2F_T(d\sqrt{n}) - 1 \end{aligned}$$

by equating with $1 - \alpha$ and solving of d we get

$$d = \frac{1}{\sqrt{n}} F_T^{-1}\left(1 - \frac{\alpha}{2}\right) = \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

where $t_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$-quantile of a t-distribution with $n - 1$ degrees of freedom.

Confidence interval

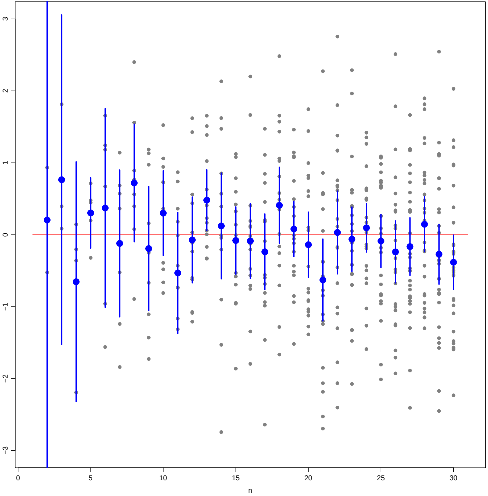
Hence we have

$$1-\alpha = P\left(\overline{Y} - \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}} \cdot S < \mu < \overline{Y} + \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n}} \cdot S\right).$$

In practice we want to make statements about unknown quantities (e.g. μ) based on realizations of the average ($\bar{y}$) and the emperical variance (s^2). As an example we could state that we are 95% confident that the true mean μ is in the interval

$$\bar{y} \pm t_{0.975} \cdot s/\sqrt{n}.$$

Confidence intervals for increasing sample size



Example[1]

The birth weight of 50 newborn girls has been recorded in an unknown country, and the sample mean and standard deviation were found to be $\bar{x}_\mathsf{p} = 3505.7$ g and $s_\mathsf{p} = 467.9$ g (use $\alpha = 0.05$).

- Calculate d.
- Find the interval $\bar{x}_\mathsf{p} \pm d \cdot s_\mathsf{p}$.
- Give an interpretation of the interval.
- If the true mean is 3300 g. is the obtained values then unusual?

[1] June 2022

The aim

In lecture 1 we saw a number of summary statistics, we now assume that

$$Y_i \sim N(\mu, \sigma^2), \quad \text{and iid.}$$

In this and the next lecture we will answer the following questions

- What is the distribution of $\bar{Y}$? Lecture 3!
- What is the distribution of S^2? Lecture 3!
- What is the distrubution of $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$? Done!
- If we calculated observed variances from two different groups, what is then the distribution of $\frac{S_1^2}{S_2^2}$? Today

Overview

The F-distribution

Definition

If $Q_1 \sim \chi^2(n_1)$, $Q_2 \sim \chi^2(n_2)$, and Q_1 and Q_2 independent then

$$F = \frac{Q_1/n_1}{Q_2/n_2}$$

follows an F-distribution with n_1 and n_2 degrees of freedom.

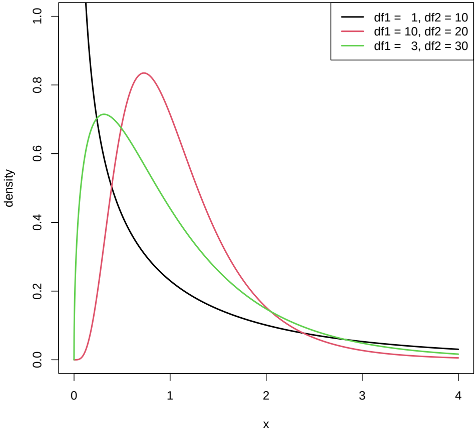
Theorem

The probability function for an F-distribution is given by

$$f_F(x) = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)\left(1+\frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}; \quad x \geq 0$$

where $B(\nu_1, \nu_2) = \frac{\Gamma(\nu_1)\Gamma(\nu_2)}{\Gamma(\nu_1+\nu_2)}$ is the Beta-funktion.

F-distribution, pdf



The F-distribution as a sampling distribution

Let $Y_{1,1},\ldots,Y_{1,n_1}$ be iid. $N(\mu_1,\sigma_1^2)$ and let $Y_{2,1},\ldots,Y_{2,n_2}$ be iid. $N(\mu_2,\sigma_2^2)$ the

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1,n_2-1)$$

where S_1^2 and S_2^2 is the sample variances for Y_1 and Y_2.

Example

Let $Y_{1,1},\ldots,Y_{1,10}$ be iid. $N(\mu_1,\sigma^2)$ and let $Y_{2,1},\ldots,Y_{2,10}$ be iid. $N(\mu_2,\sigma^2)$ find

$$P(S_1^2/S_2^2 > 2)$$

where S_1^2 and S_2^2 are sample variances for Y_1 and Y_2.

- Assume that you in a concrete study observe $s_1^2/s_2^2 = 2$, what would your assesment of the assumption equal variance in the two populations be?

Overview

Two independent samples

Assume that you are planning a study with samples from two independent populations $Y_{1,1},...Y_{1,n_1}$ and $Y_{2,1},...,Y_{2,n_2}$, further assume that $Y_{1,i} \sim N(\mu_1,\sigma^2)$ and iid., and $Y_{2,i} \sim N(\mu_2,\sigma^2)$ and iid.

- What is the best estimator S_p^2 for σ^2?

Now make the further assumption that $\mu_1 = \mu_2$ (which we will refer to as a hypothesis), what is the distribution of

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In a concrete study with $n_1 = n_2 = 10$ you observe

$$t_{obs} = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 4$$

Would that be unusual given the assumptions above?

Three independent samples

Assume that you are planning a study with samples from three independent populations $Y_{i1},...Y_{i,n_1}$, $i \in \{1,2,3\}$, further assume that $Y_{i,j} \sim N(\mu_i, \sigma^2)$ and iid.

- What is the best estimator S_p^2 for σ^2?

Now make the further assumption that $\mu_l = \mu_m = \mu$ for all (l,m) (which we will refer to as a hypothesis), and set

$$\bar{Y} = \frac{1}{n_1+n_2+n_3}\sum_{i=1}^{3}\sum_{l=1}^{n_i} Y_{il}.$$

Find the distribution of

$$F = \frac{\frac{1}{3-1}\sum_{i=1}^{3} n_i(\bar{Y}_i - \bar{Y})^2}{S_p^2}$$

In a concrete study with $n_1 = n_2 = n_3 = 10$ you observe

$$F_{obs} = \frac{\frac{1}{3-1}\sum_{i=1}^{3} n_i(\bar{y}_i - \bar{y})^2}{s_p^2} = 2.5.$$

Would that be unusual given the assumptions above?

Example[2]

The A-series of paper is defined by, long edge $= \sqrt{2}$ times the short edge. A machine is cutting an A-series of paper. Assume that the accuracy of the machine can be expressed as

$$X \sim N(k, \sigma^2)$$
$$Y \sim N(\sqrt{2}k, \sigma^2)$$

where X is the short edge and Y is the long edge, it can further be assumed the X and Y are independent.

- With X and Y as defined above, what is $E[X^2 + Y^2]$?
- Again with X and Y as defined above, what is $P\left(\frac{(X-k)^2}{(Y-\sqrt{2}k)^2} < 2\right)$?
- What is $P\left(\frac{X-k}{|Y-\sqrt{2}k|} < -1\right)$?

[2] 2021 June

Agenda