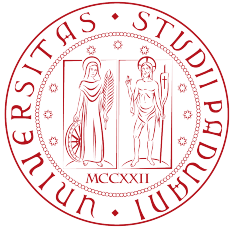
Regression and Time Series Models

Inferential Statistics



Manuela Cattelan

manuela.cattelan@unipd.it

Department of Statistical Sciences, University of Padova

Introduction

Descriptive vs inferential statistics

Descriptive statistics is used to describe a population.

However, we typically cannot observe the whole population!

This happens because of time and cost reasons, because the population is not finite, the measurement destroys the statistical unit, etc. etc.

We observe a sample and we want to draw conclusions on the population. This is what inferential statistics is about.

We consider a simple random sample (every item in the population has an even chance and likelihood of being selected) of independent and identically distributed (i.i.d.) random variables

To infer from a sample to a population, wee need tools from probability theory!

Starting point!

Some famous distributions

A random variable is a variable which associates one real number to each event of the sample space (the set of all possible outcomes).

Assume the experiment is rolling a dice once, a random variable can be the number of even values observed. If the event $e_i \in \{2, 4, 6\}$ the random variable is 1, otherwise it is 0. Hence its support is the set $\{0, 1\}$.

Random variables are

1. discrete, when the outcomes are counted
2. continuous, when the outcomes are measured

Let X be a discrete random variable, that may assume values x_i, $i = 1, \ldots, n$, with probability $p(x_i)$.
Recall that the probability distribution function is such that $0 \leqslant p(x_i) \leqslant 1$ and $\sum_{i=1}^{n} p(x_i) = 1$.
The mean of X is

$$E(X) = \mu = \sum_{i=1}^{n} x_i p(x_i),$$

and its variance is

$$\mathrm{Var}(X) = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i).$$

The r-th moment of a discrete random variable is

$$E(X^r) = \sum_{i=1}^{n} x_i^r p(x_i).$$

Let Y be a continuous random variable with probability density function $f(y)$, then the probability $P(A < X < b)$ is

$$P(A < X < b) = P(A \leqslant X \leqslant b) = \int_a^b f(y)dy.$$

The mean of Y is

$$E(Y) = \mu = \int_{-\infty}^{+\infty} y\, f(y)dy,$$

and its variance is

$$\text{Var}(Y) = \sigma^2 = \int_{-\infty}^{+\infty} (y-\mu)^2 f(y)dy = \int_{-\infty}^{+\infty} y\, f(y)dy - \mu^2$$

The r-th moment of a continuous random variable is

$$E(X^r) = \int_{-\infty}^{+\infty} y^r\, f(y)dy$$

Suppose you are a bank and, in order to decide whether to lend money to company A, you flip a coin. If the coin turns head, company A will be financed, otherwise not. The probability of a head is π, so if X denotes the variable "the company is financed", then

$$P(X = x) = \pi^x (1 - \pi)^{1-x}, \quad x = 0, 1, \quad 0 \leqslant \pi \leqslant 1,$$

where $P(X = 0)$ is the probability of not financing A. This is a Bernoulli distribution, denoted as $X \sim Ber(\pi)$.

- $E(X) = \sum_{x=\{0,1\}} x\, P(X = x) = \pi$
- $\text{Var}(X) = \sum_{x=\{0,1\}} (x - E(X))^2 P(X = x) = \pi(1 - \pi)$

Assume that n companies ask for money, and every time the bank flips a coin to decide whether to finance the company. Each company has probability π of being financed. However, this time you are interested in how many companies will be financed. The probability is

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad x = 0, 1, \ldots, n, \quad 0 \leqslant \pi \leqslant 1,$$

this is a binomial distribution, denoted as $X \sim Bin(n, \pi)$.
Important: each company is financed independently from the others and with the same probability.

- $E(X) = n\pi$
- $\text{Var}(X) = n\pi(1-\pi)$

The most well-known distribution is the normal bell-shaped distribution

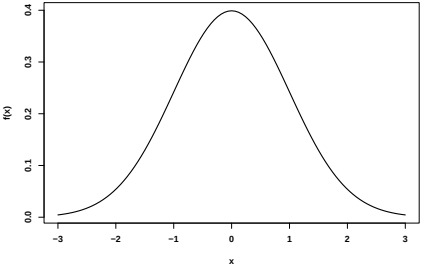
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$$

- $E(X) = \mu$
- $\text{Var}(X) = \sigma^2$

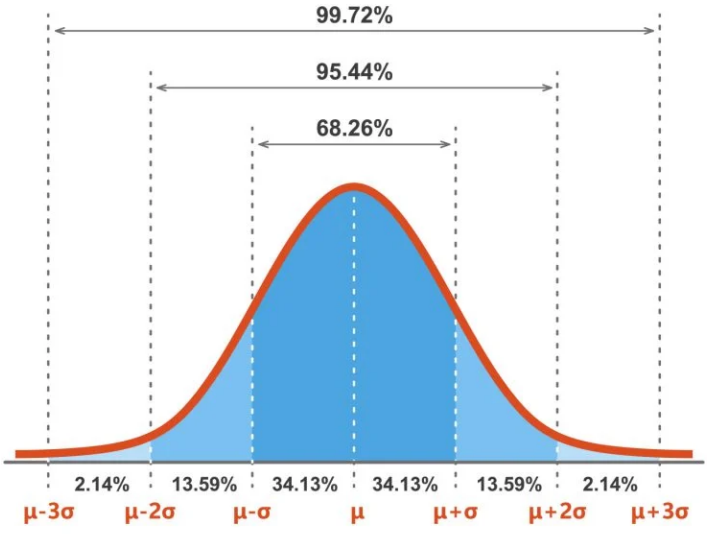
This is also denoted as $X \sim N(\mu, \sigma^2)$. The variable

$$Z = \frac{X-\mu}{\sqrt{\sigma^2}},$$

is the standard normal distribution $Z \sim N(0,1)$.



The normal distribution



- The linear transformation of a normal random variable is normally distributed: if $X \sim N(\mu, \sigma^2)$, then $V = aX + b$, where $a, b \in \mathbb{R}$, is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$.
- The sum of normally distributed random variables has a normal distribution, i.e. if $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ and $\text{Cov}(X, Y) = \sigma_{12}$, then

$$\begin{aligned} W &= X + Y \sim N\Big(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\sigma_{12}\Big) \\ Z &= X - Y \sim N\Big(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\Big) \end{aligned}$$

The mean of a sufficiently large number of identically distributed random variates will be approximately normally distributed.

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables having a distribution with expected value μ and finite variance σ^2. Set

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

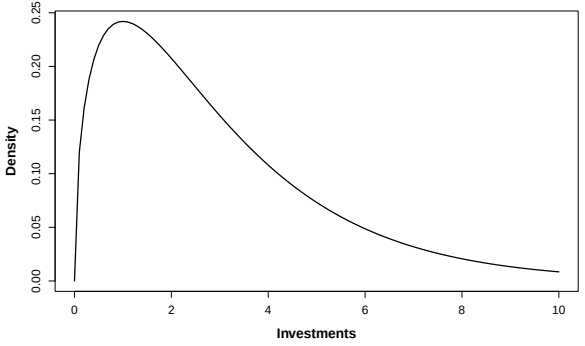
then, the random variable

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{(\bar{X}_n - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

converges in distribution to the standard normal random variable when $n \to \infty$.

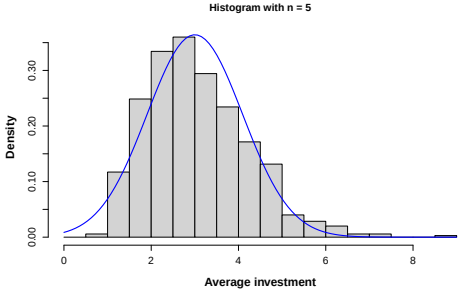
Example

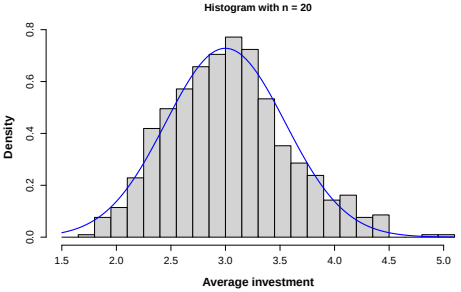
Assume that small investors invest an amount of money drawn from a distribution $X \sim \chi^2_3$, which states that on average they invest 3 (in hundreds of dollars). This is the distribution



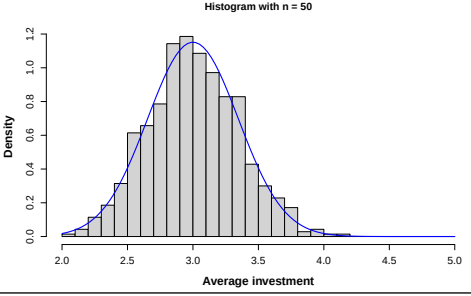
Assume that each day n small investors enter the market, what is the distribution of the mean amount invested per day? We can try empirically and record the average investment per day during, for example, 700 days, and plot its histogram. The CLT tells us the form of the distribution of this mean.

Example





X has mean 3 and variance $2 \times 3 = 6$
$n = 5 \rightarrow \bar{X}_n$ mean 3 and variance $\frac{6}{5} = 1.2$
$n = 20 \rightarrow \bar{X}_n$ mean 3 and variance $\frac{6}{20} = 0.3$
$n = 50 \rightarrow \bar{X}_n$ mean 3 and variance $\frac{6}{50} = 0.12$



- The CLT says that for a series of standardised r.v. with n large enough, we may assume the approximation

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Hence, for X from any distribution, with minimum assumptions, if X can be written as a sum of n random independent phenomena, $X = \sum_{i=1}^{n} X_i$, with the same distribution, for n large enough

$$X \sim N(n\mu, n\sigma^2).$$

- The *Binomial* r.v. can be seen as the sum of n Bernoulli *iid* r.v., hence, for n large enough, its distribution is close to the r.v. $N(n\pi, n\pi(1-\pi))$

Inference

Assume you are the banker that tosses the coin for deciding about lending money to a company.

You do not know the probability of a head, π, and you want to guess it.

To this purpose, you register whether the loan is granted or not in 10 different cases

$$1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0$$

and from the observed data you want to find a reasonable value for π.

What is the probability of observing that particular sample? If the observations are **independent**, then this probability is the product of the probabilities of single observations, that is (sample: 1, 1, 1, 0, 1, 0, 1, 1, 1, 0)

$$\begin{aligned} P(X_1, \ldots, X_{10}; \pi) &= P(X_1 = x_1; \pi) P(X_2 = x_2; \pi) \cdots P(X_{10} = x_{10}; \pi) \\ &= \pi^{x_1}(1-\pi)^{1-x_1} \pi^{x_2}(1-\pi)^{1-x_2} \cdots \pi^{x_{10}}(1-\pi)^{1-x_{10}} \\ &= \pi^1(1-\pi)^{1-1} \pi^1(1-\pi)^{1-1} \cdots \pi^0(1-\pi)^{1-0} \\ &= \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i}, \end{aligned}$$

where n is the total number of observations, in this example 10.
We can see the above quantity as a function of the unknown parameter π, $L(\pi; X_1, \ldots, X_n)$ and find the value π that maximises this probability

$$l(\pi; X_1, \ldots, X_n) = \log L(\pi; X_1, \ldots, X_n) = \sum_{i=1}^n x_i \log \pi + \left(n - \sum_{i=1}^n x_i\right) \log(1-\pi)$$

which is maximised for $\hat{\pi} = \frac{\sum_{i=1}^n x_i}{n}$. In the example $\hat{\pi} = \frac{7}{10} = 0.7$.

You collect data on the average revenue per user (which is the total monthly revenue made divided by the total number of users making purchases) of 10 clothing companies.
Assume this variable follows a normal distribution, with unknown mean and variance.
What are reasonable values for these parameters?

The data are

177 128 137 93 99 159 303 95 100 146

and you want to find a reasonable value for μ and σ.

More formally, let $X_1, X_2, \ldots, X_n$ be a random sample of independent and identically distributed random variables observed on n units, with values $x_1, x_2, \ldots, x_n$. The aim is to obtain an estimate of the parameter using a function of the sample

$$t = t(x_1, \ldots, x_n),$$

which is itself a random variable, called estimator $T = t(X_1, \ldots, X_n)$.

If the observations are **independent**, the likelihood function for the example in the previous slide is

$$L(\mu, \sigma; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\}$$

and the maximum likelihood estimator is found by maximising the above function:

$$\begin{aligned} \hat{\mu} &= \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \\ \hat{\sigma}^2 &= S_n^2 = \frac{\sum_{i=1}^{n} (X_i - \hat{\mu})^2}{n} \end{aligned}$$

In the revenues example, we can compute

$$\begin{aligned} \bar{x} &= \frac{177 + \cdots + 146}{10} = \frac{1437}{10} = 143.7 \\ s_n^2 &= \frac{(177 - 143.7)^2 + \cdots + (146 - 143.7)^2}{10} = \frac{35866.1}{10} = 3586.61 \end{aligned}$$

Another approach for estimating the parameters is based on the moments of the distribution, and consists in setting the theoretical moments equal to the empirical ones.

So, if μ denotes the mean of the distribution, then

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

and when there is more than one parameter, you consider successive moments, as for example,

$$E(X^2) = \frac{\sum_{i=1}^{n} X_i^2}{n}$$

Consider $\bar{X}$, since it is a linear combination of random variables, it is a random variable, with its own distribution. This is true for any estimator.

Some desirable properties of the estimators are

- Unbiasedness, when the mean of the estimator T is equal to the parameter θ it estimates

$$\mathrm{E}(T) = \theta, \qquad \forall\, \theta$$

- Consistency.
  Let $T_1, T_2, \ldots, T_n$ be a sequence of estimators of a parameter θ, where $T_n = T_n(X_1, \ldots, X_n)$ is a function of $X_1, \ldots, X_n$. The sequence $\{T_n\}$ is a consistent sequence of estimators for θ if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|T_n - \theta| < \epsilon) = 1$$

  This definition says that as the sample size n increases, the probability that T_n is getting closer to θ is approaching 1.

- It is easy to show that $\mathrm{E}(\bar{X}) = \mu$, so $\bar{X}$ is an unbiased (and consistent) estimator of the mean of the population.
- On the contrary, $\mathrm{E}(S_n^2) = \frac{n}{n-1}\sigma^2$, is a biased estimate of the population variance, but $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is unbiased.

Is a point estimator enough?

- The sample we observe is only one of all the possible samples one could have observed.
- Can we consider this uncertainty when estimating a parameter?
- Instead of one value (a point estimator), can we find a set of plausible values?
- Yes, we can find an interval estimator.
- We need to define a confidence level.

Consider the example of the average revenue per user, in which each observation is drawn from a normal distribution, i.e. $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Assume the variance σ^2 is known, then the point estimator of the mean is

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

which is a (weighted) sum of normally distributed random variables.

Do you remember which is the distribution of $\hat{\mu}$?

Consider the example of the average revenue per user, in which each observation is drawn from a normal distribution, i.e. $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Assume the variance σ^2 is known, then the point estimator of the mean is

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

which is a (weighted) sum of normally distributed random variables.

Do you remember which is the distribution of $\hat{\mu}$?

It is a combination of normal r.v., hence $\hat{\mu} \sim N\left(\frac{1}{n} \sum_{i=1}^{n} \mu, \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2\right)$, that is $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

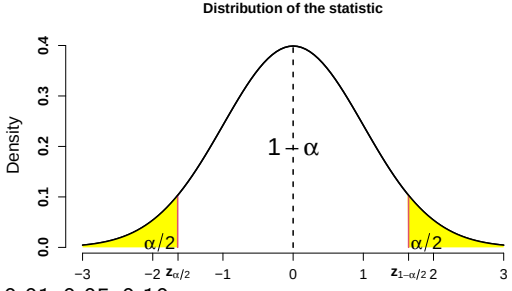
Can we use this information to find a set of plausible values for μ?

Since $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, we can write

$$P\left(z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

where z_α indicates the α-th quantile of the standard normal distribution, that is $P(Z < z_\alpha) = \alpha$. Then

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$



Typical values for α are $0.01, 0.05, 0.10$.

- The quantity $(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ is a random interval, because $\bar{X}$ is a random variable.
- However, after collecting the data, we can compute the observed interval with associated confidence level (not probability) $1 - \alpha$.
- Example. Suppose the variance of the distribution of the average revenue per user is known and it is $\sigma^2 = 3721$, then we can compute an interval estimate for μ as

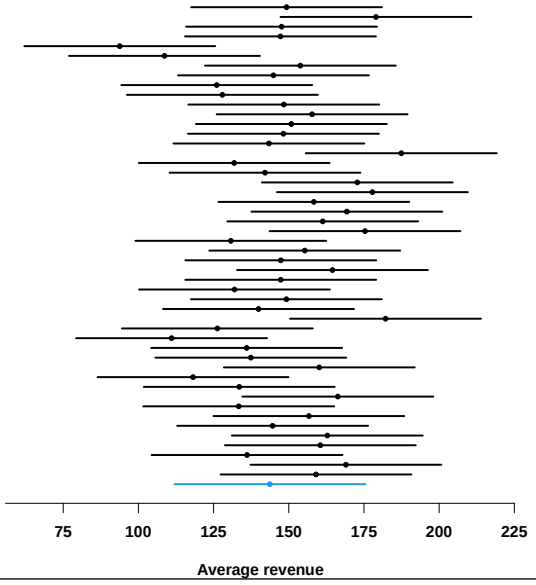
$$\left(143.7 - z_{1-\alpha/2}\frac{\sqrt{3721}}{\sqrt{10}}, 143.7 - z_{\alpha/2}\frac{\sqrt{3721}}{\sqrt{10}}\right)$$

We need to set the confidence level in order to compute it. If $\alpha = 0.1$, then $z_{1-\alpha/2} = z_{0.95} = 1.64$, so a 90% confidence interval for the mean is (the standard normal is symmetric around 0, so $z_\alpha = -z_{1-\alpha}$)

$$\left(143.7 - 1.64\frac{61}{\sqrt{10}}, 143.7 + 1.64\frac{61}{\sqrt{10}}\right) = (112.06, 175.34)$$

Compute the 95% confidence interval using the information $z_{0.975} = 1.96$.

Why confidence and not probability?



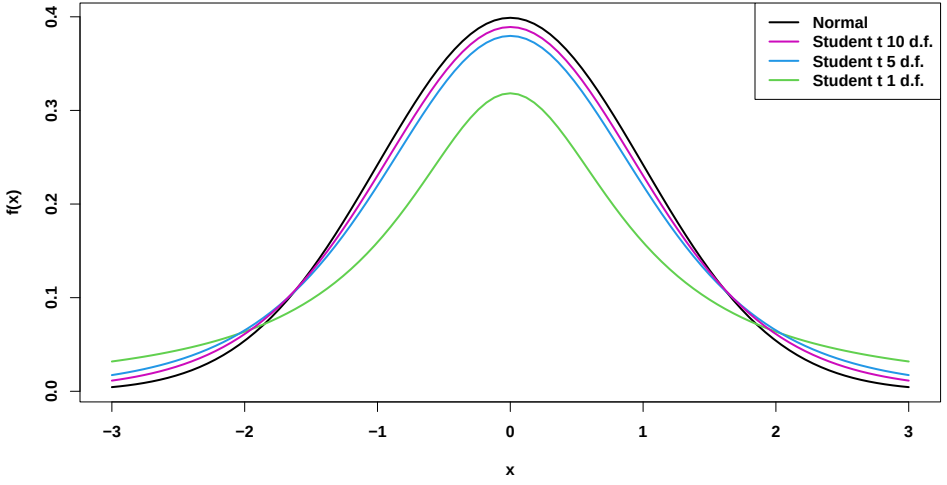
In blue the observed confidence interval

- When you want to estimate the mean, it is unlikely that you know the variance.
- Can we determine a confidence interval for the mean even though we don't know the variance?
- We need to find a function of the data whose distribution does not depend on unknown parameters, just like $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a known distribution.
- Luckily, somebody studied the distribution of $\frac{\bar{X}-\mu}{S_{n-1}/\sqrt{n}}$ and found that its distribution is Student's-t with $n-1$ degrees of freedom

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2},$$

for $x \in \mathbb{R}$, $\nu > 0$ degrees of freedom and Γ is the Gamma function ($\Gamma(y) = \int_0^\infty t^{y-1}e^{-t}dt,\ y > 0$).
- Hence, if you know the degrees of freedom, you can find the quantiles!

Student's-t distribution



Consider the example on the average revenue per user, in which each observation is drawn from a normal distribution, i.e. $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. The parameters μ and σ^2 are both unknown, but the point estimator of the mean remains

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

which is a (weighted) sum of normally distributed random variables, whose distribution depends on the unknown parameter σ^2.

But we have just learnt that $\frac{\bar{X}-\mu}{S_{n-1}/\sqrt{n}}$ has a distribution that depends only on quantity n, so we can write

$$P\Big(t_{\alpha/2;n-1} < \frac{\bar{X} - \mu}{S_{n-1}/\sqrt{n}} < t_{1-\alpha/2;n-1}\Big) = 1 - \alpha$$

where $t_{\alpha;n-1}$ denotes the α-th quantile of the Student's-t distribution with $n-1$ degrees of freedom. Then

$$P\Big(\bar{X} - t_{1-\alpha/2;n-1}\frac{S_{n-1}}{\sqrt{n}} < \mu < \bar{X} - t_{\alpha/2;n-1}\frac{S_{n-1}}{\sqrt{n}}\Big) = 1 - \alpha$$

Typical values for α are $0.01, 0.05, 0.10$.

The quantity $(\bar{X} - t_{1-\alpha/2;n-1}\frac{S_{n-1}}{\sqrt{n}}, \bar{X} - t_{\alpha/2;n-1}\frac{S_{n-1}}{\sqrt{n}})$ is again a random interval.
The variance of the data on the average revenue per user is $s^2_{n-1} = 3985.122$, then an interval estimate for μ is

$$\left(143.7 - t_{1-\alpha/2;n-1}\frac{\sqrt{3985.122}}{\sqrt{10}}, 143.7 - t_{\alpha/2;n-1}\frac{\sqrt{3985.122}}{\sqrt{10}}\right)$$

We need to set the confidence level in order to compute it. If $\alpha = 0.1$, then $t_{1-\alpha/2;9} = t_{0.95;9} = 1.833$, so a 90% confidence interval for the mean is (the Student's-t distribution is symmetric around 0, so $t_{\alpha;n-1} = -t_{1-\alpha;n-1}$)

$$\left(143.7 - 1.833\frac{\sqrt{3985.122}}{\sqrt{10}}, 143.7 + 1.833\frac{\sqrt{3985.122}}{\sqrt{10}}\right) = (107.11, 180.29)$$

Compute the 95% confidence interval using the information $z_{0.975} = 2.26$.

- What can we do if the sample is not from a normal random variable?
- If the sample is sufficiently large, we can consider an approximation of the confidence interval for the mean through the central limit theorem.
- The CLT states that $\bar{X} \dot{\sim} N(\mu, \sigma^2/n)$, so $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- Since the variance is not known, we use its estimator S^2_{n-1}
- Hence, the random variable $\frac{\bar{X}-\mu}{S_{n-1}/\sqrt{n}} \sim t_{n-1}$, but when n is large $t_{n-1} \to N(0,1)$.
- We can thus use the approximate $1-\alpha$ confidence interval

$$\left[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

- Consider the amount of money invested by small investors, which we assumed to follow a chi-square distribution.
- Suppose that we collect the amount invested by 80 people, can we estimate the mean amount invested?
- The sample mean is $\bar{x} = 2.74$ and the (unbiased) sample variance is $s^2 = 4.40$, hence, an approximate 95% confidence interval for the mean amount invested is

$$\left[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

$$\left[2.74 - 1.96\frac{\sqrt{4.4}}{\sqrt{80}}; 2.74 + 1.96\frac{\sqrt{4.4}}{\sqrt{80}}\right]$$

$$[2.28; 3.20]$$

We are interested in the proportion of loans given by a bank, and to this purpose we collect information on a sample of people.

If X denotes the result of loan requests, then $X \sim Bin(n, \pi)$, and an estimator of π is the sample mean, for which

$$E(\bar{X}) = \pi \text{ and } Var(\bar{X}) = \frac{\pi(1-\pi)}{n}$$

The exact distribution of π is not straightforward, but if n is large, the CLT says that the distribution $\bar{X}$ tends to a Normal, so as $n \to \infty$

$$Z = \frac{\bar{X} - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0, 1)$$

- However, the variance depends on the unknown parameter π

$$1-\alpha \cong P\left(-z_{\alpha/2} \leqslant \frac{\bar{X}-\pi}{\sqrt{\pi(1-\pi)/n}} \leqslant z_{\alpha/2}\right)$$
$$= P\left(\bar{X}-z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leqslant \pi \leqslant \bar{X}+z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

- The solution is to use an estimator of the variance and, as $\bar{X}$ is a consistent estimator of π, the estimator $\bar{X}(1-\bar{X})$ will tend to $\pi(1-\pi)$.
- So, if n is sufficiently large, an **approximate** $1-\alpha$ confidence interval for the proportion π is:

$$\left(\bar{X}-z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leqslant \pi \leqslant \bar{X}+z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right)$$

- Typically sufficiently large means both $n\pi \geqslant 5$ and $n(1-\pi) \geqslant 5$, and we evaluate these quantities using the estimate of π.

The observation on 200 people that requested a loan to a bank reveals that 135 were given the loan. Find a 95% confidence interval for π, the probability that the bank gives a loan.

The random variable X_i is 1 if the loan is given and 0 otherwise. The loan is given with probability π, so

$$X_i \sim Ber(\pi) \qquad i = 1, 2, \ldots, 200$$

If $X = \sum_{i=1}^{200} X_i$ is the number of loans given, then $X \sim Bin(200, \pi)$. As n is large, we can use the normal approximation.

- Thanks to the CLT, $\bar{X} = \frac{1}{200}\sum_{i=1}^{200} X_i$,

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

- A point estimate of π is the proportion of loans given

$$\hat{\pi} = \bar{x} = \frac{135}{200} = 0.675$$

- The estimate of the variance is

$$\frac{\hat{\pi}(1-\hat{\pi})}{n} = \frac{0.675(1-0.675)}{200} = 0.0011$$

- And the approximate 95% confidence interval is

$$\left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}; \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right)$$
$$= \left(0.675 - 1.96 \cdot \sqrt{\frac{0.675 \cdot 0.325}{200}}; 0.675 + 1.96 \cdot \sqrt{\frac{0.675 \cdot 0.325}{200}}\right)$$
$$= [0.61;\ 0.74]$$

Hypothesis testing

- Assume you are a portfolio manager and you are interested in testing whether the daily returns of the S&P500 index have zero mean or not.
- Then, you need a rigorous procedure to perform such a test on the basis of an observed sample.
- How can we test this hypothesis?
- The theory of hypothesis testing tell us how to do it
- Note that this hypothesis concerns the value of the parameter of a distribution.

The approach proposed by J. Neyman and E. S. Pearson distinguishes two hypothesis

- the null hypothesis, denoted by H_0
- the alternative hypothesis, typically denoted by H_1 or H_a

The null hypothesis is considered true unless you prove it is not.

In the case of the portfolio manager, the null hypothesis states that the mean of the distribution of the returns is 0, i.e. $H_0 : \mu = 0$.

In hypothesis testing, we collect a sample and evaluate whether there is enough proof in the sample as to reject the null hypothesis.

If you are the portfolio manager: collect a sample of returns and see whether its mean is sufficiently different from the one hypothesised, in this case it is an indication towards the rejection of the null hypothesis.

Let Θ be the parametric space, the set of all possible values of the parameter θ, and let Θ_0 and Θ_1 be a partition of the parametric space.
An hypothesis system is the partition in two of the parametric space

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

in the example

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

There are one-tailed tests and two-tailed tests, depending on the alternative.

One tailed-tests when the alternative has "one direction"

The most frequent situations are

- $$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$
- $$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$
- $$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$
- $$\begin{cases} H_0 : \theta \geqslant \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

where θ_0 represents a fixed value of the parameter.

- How can we determine whether the sample shows evidence against the null hypothesis?
- We need a statistical test that allows to discriminate whether the sample conducts to the rejection of the null hypothesis
- The test statistic is a sample statistic whose distribution is completely known under the null hypothesis H_0.
- There is a set of values of the test statistic for which we accept the null hypothesis, this is the *acceptance region*
- There is a set of values of the test statistic for which we reject the null hypothesis, this is the *rejection region*

The portfolio manager that wants to test whether the mean of daily returns of the S&P500 index is null, considers the hypothesis

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

for a fixed μ_0. This is a two-tailed test.

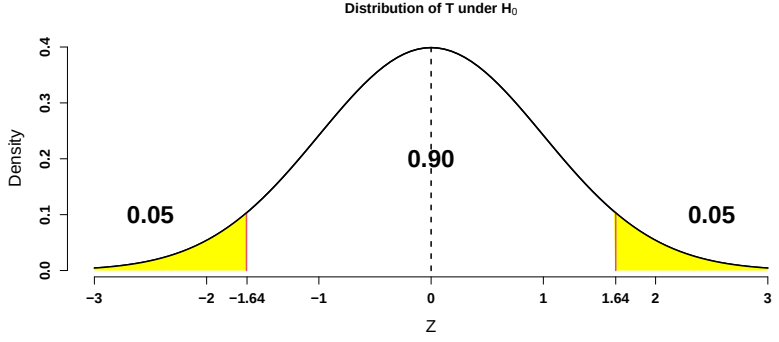
A random sample, $X_1, X_2, \ldots, X_n$, with $X_i \sim N(\mu, \sigma^2)$ of daily returns is collected, and, when H_0 is true, $X_i \sim N(\mu_0, \sigma^2)$ and $\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$. Assume σ^2 is known, then, if H_0 is true

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

If H_0 is true

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

So, if the observed value of T is likely for a standard normal distribution, we don't reject the null hypothesis H_0, otherwise we reject it. How much likely? It depends on the size of the test α.

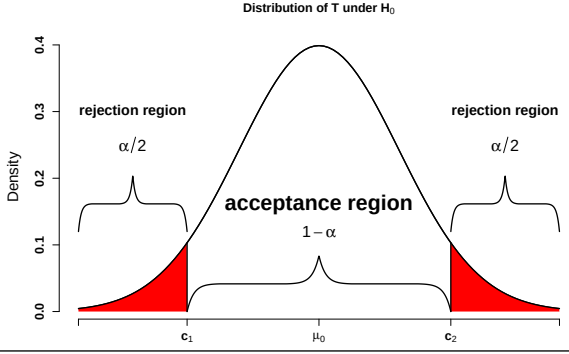


- Critical values depend on the distribution of the test statistic under H_0 and on the level of the test α
- Let T be the test statistic, if the test is two-tailed ($H_1 : \theta \neq \theta_0$) we need two critical values, c_1 and c_2, such that
$$P(c_1 \leqslant T \leqslant c_2) = 1 - \alpha$$
- if the test has one tail ($H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$), we have to find a critical value c_1 such that
$$P(T \geqslant c_1) = \alpha$$
or
$$P(T \leqslant c_1) = \alpha$$

Hence, we define the acceptance and rejection regions according to the distribution of T when H_0 is true.

We define the size α of the test (typically 5% or 1%) and define the rejection region accordingly

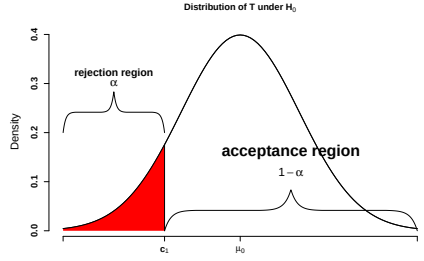
$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

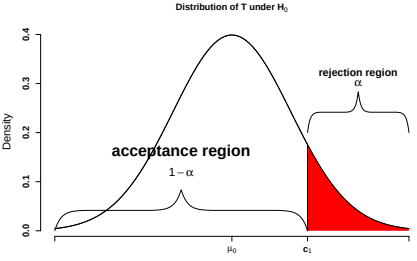


Acceptance and rejection regions

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$





You observe the returns of the General Electric stocks for 40 days. We can assume that the returns are normally distributed with variance $\sigma^2 = 0.00031$ and the interest lies in testing whether the mean of the daily returns is zero.

Different hypothesis systems can be considered.
Let the mean of the sample be $\bar{x} = 0.00029$, and assume you are interested in

$$\begin{cases} \mu = 0 \\ \mu \neq 0 \end{cases}$$

Then, if $\mu = 0$, we know that

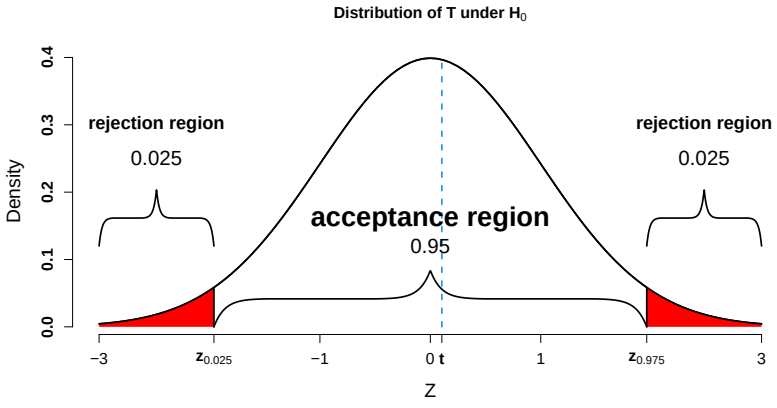
$$T = \frac{\bar{X} - 0}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

so

$$t = \frac{0.00029}{\sqrt{0.00031/40}} = 0.104$$

Is it a plausible value?

Let us find the critical values. If the level of the test is 5%, then



and $z_{0.975} = 1.96$.
What if we set the level of the test to 10% or to 1%? ($z_{0.95} = 1.64$, $z_{0.995} = 2.58$)

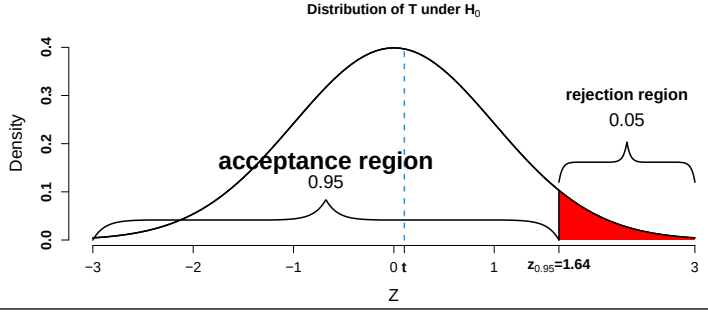
Now assume that we are interested in testing the following hypothesis

$$\begin{cases} \mu = 0 \\ \mu > 0 \end{cases}$$

Then, if $\mu = 0$, we know that

$$T = \frac{\bar{X} - 0}{\sqrt{\sigma^2/n}} \sim N(0,1) \text{ so } t = \frac{0.00029}{\sqrt{0.00031/40}} = 0.104$$

Is it a plausible value?



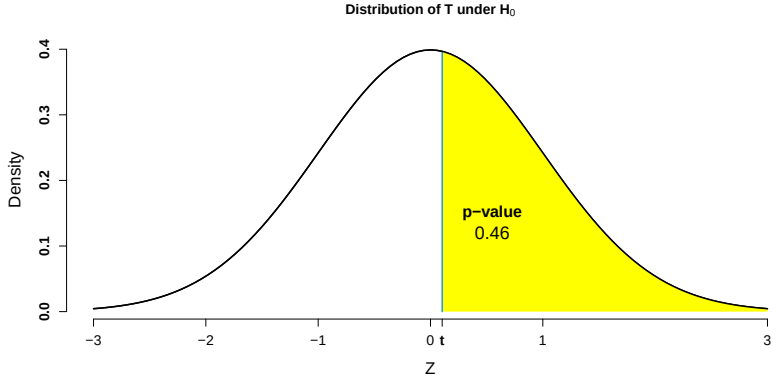
What if we set the level of the test to 10% or to 1%? ($z_{0.90} = 1.28$, $z_{0.99} = 2.32$)

- Often, instead of deciding the level of the test, it is more informative to report the p-value of the test.
- p-value = probability of observing a value of the test statistic more extreme than the one observed under the null hypothesis.
- The p-value is a measure of the evidence against H_0, hence

$$\text{if } p - \text{value} < \alpha \rightarrow \text{reject } H_0$$

Consider the previous example

$$\begin{cases} \mu = 0 \\ \mu > 0 \end{cases}$$



$1 - \Phi(0.104) = 0.4585$, where $\Phi(x)$ denotes the cumulative distribution function of a standard normal random variable computed in x.

And what happens with two-tailed hypothesis?

$$\begin{cases} \mu = 0 \\ \mu \neq 0 \end{cases}$$

In this case both extremities should be condsidered, so the *p*-value is

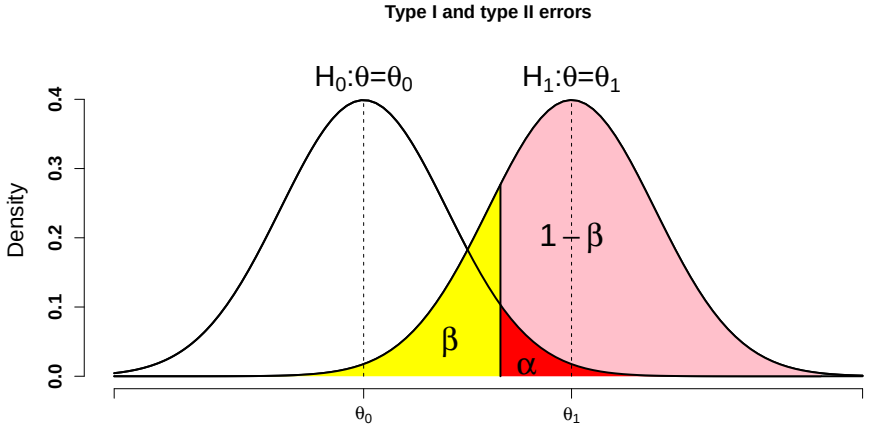
$$2\min(\Phi(t), 1 - \Phi(t)) = 2\min(0.5414, 0.4585) = 0.917$$

Type I and type II errors

- The level of the test α **is the probability of values of the test statistic in the rejection region when H_0 is true**.
- In hypothesis testing there are two types of errors:
  - **Type I error**: reject H_0 when it is true
  - **Type II error**: not rejecting H_0 when it is false

	Decision	
	Accept H_0	**Reject** H_0
H_0 **is true**	*Correct*	Type I error
	$1-\alpha$	α
H_0 **is not true**	Type II error	*Correct*
	β	$1-\beta$

- the level of the test α is the probability of a type I error.
- β is the probability of a **type II error**.
- $1-\beta$ is the **power of the test**, which is the probability of rejecting the null hypothesis when it is false.


Type I and type II errors
H0:θ=θ0
H1:θ=θ1
Density
0.0
0.1
0.2
0.3
0.4
β
α
1 − β
θ0
θ1

Consider the returns of the General Electric stocks for 40 days. Suppose you can assume they are normally distributed, but the variance is unknown. The sample statistics are $\bar{x} = 0.00029$ and $s^2 = 0.00031$ and you are interested in testing whether the return is null or positive. Then

$$\begin{cases} \mu = 0 \\ \mu > 0 \end{cases}$$

If $\mu = 0$, we know that

$$T = \frac{\bar{X} - 0}{\sqrt{s^2/n}} \sim t_{n-1}$$

so in order to evaluate whether

$$t = \frac{0.00029}{\sqrt{0.00031/40}} = 0.104$$

is a plausible value, we refer to the t_{39} distribution.

$t_{39;0.90}$	$t_{39;0.95}$	$t_{39;0.975}$	$t_{39;0.99}$	$t_{39;0.995}$
1.304	1.685	2.023	2.426	2.708

Suppose an investor is interested in testing whether the mean of the S&P500 return is zero, or different from zero, then

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

The investor collected a sample of 7300 observations, whose sample mean is $\bar{x} = 0.00027$ and sample variance is $s^2 = 0.00015$.

Even making no assumptions on the distribution of the returns, since the sample is large, we know that, if H_0 is true

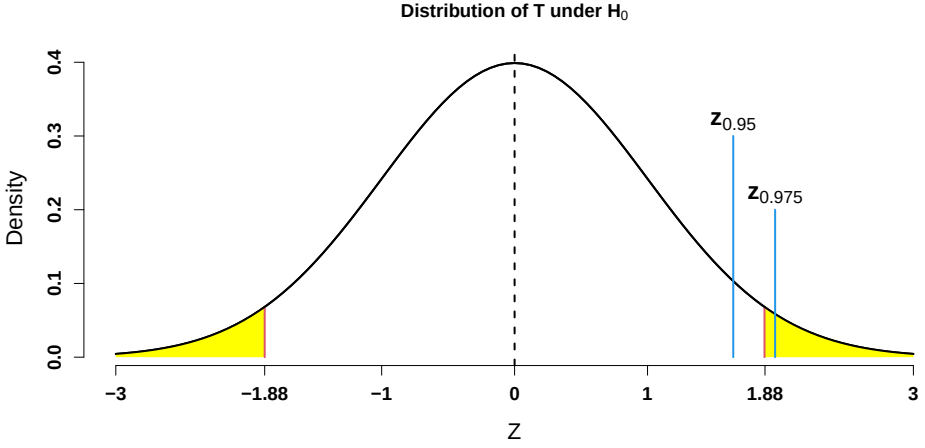
$$T = \frac{\bar{X} - 0}{\sqrt{s^2/n}} \dot{\sim} N(0, 1)$$

So

$$t = \frac{0.00027}{0.0122/85.44} = 1.884$$

this value can be compared to the quantiles of a standard normal distribution ($z_{0.95} = 1.645$ and $z_{0.975} = 1.96$). What can we conclude?

Otherwise, even better, we can compute its p-value:
$2\min(\Phi(t), 1-\Phi(t)) = 2\min(0.970, 0.030) = 0.060$.



You own a stock and, each day, you consider whether the price goes up or down.

If you are interested in the probability of the price going up, then you can model this random variable as a Bernoulli distribution with probability π of the price increasing.

In order to test whether the probability of the price going up is at least one-half, you perform the test

$$\begin{cases} H_0 : \pi = \pi_0 \\ H_1 : \pi < \pi_0 \end{cases}$$

and in this case $\pi_0 = 0.5$. You know that, if H_0 is true, and if the sample is large, then

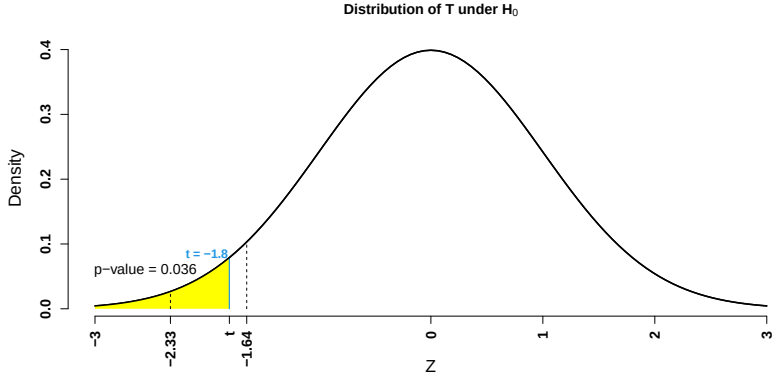
$$T = \frac{\bar{X} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \dot{\sim} N(0,1)$$

Assume that you observe that in 41 out of 100 days, the price goes up.

The observed value of the test statistic is

$$t = \frac{0.41 - 0.5}{\sqrt{0.5(1-0.5)/100}} = -1.8$$

Which can be compared to the quantiles of the normal distribution, specifically $z_{0.01} = -2.326$ or $z_{0.05} = -1.645$



Tests for independent samples from two populations

- Assume there are two populations and the variable of interest in the first population has distribution $X_1 \sim N(\mu_1, \sigma_1^2)$ while in the second has distribution $X_2 \sim N(\mu_2, \sigma_2^2)$.
- Assume variances are known σ_1^2 and σ_2^2.
- Interest lies in testing the equality of the two means

$$H_0 : \mu_1 = \mu_2$$

- The alternative hypothesis can be

$$H_1 : \mu_1 > \mu_2 \qquad H_1 : \mu_1 < \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$

- Consider a sample of size n_1 from the first population and a sample of size n_2 from the second population
- Then

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

- If H_0, $\mu_1 = \mu_2$ is true, then

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

that is (under H_0):

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

which may be used as test statistic.

- If we set $\mu_D = \mu_1 - \mu_2$ then

$$H_0 : \mu_D = 0$$

and the alternative hypothesis may be

$$H_1 : \mu_D > 0 \qquad H_1 : \mu_D < 0 \qquad H_1 : \mu_D \neq 0$$

- If the variances are not known, we use the sample statistics
- Assume that the unknown variances are the same, i.e. $\sigma_1^2 = \sigma_2^2$
- The variance can be estimated through a pooled estimator

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

  where S_1^2 and S_2^2 are the unbiased variance estimators in each sample
- Hence S_p^2 is a weighted average of the two unbiased estimators of σ_1^2 e σ_2^2 with weights proportional to the sizes of the samples.

- If the means are equal, we can employ the statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$

  which, when $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_D = 0$ is true, follows a Student's-t distribution with $(n_1 + n_2 - 2)$ dof

- The rejection regions, depending on the alternative hypothesis are

Alternative hypothesis	**Rejection region**
$H_1 : \mu_D > 0$	$T \geqslant t_{1-\alpha}$
$H_1 : \mu_D < 0$	$T \leqslant t_{\alpha}$
$H_1 : \mu_D \neq 0$	$T \leqslant t_{\alpha/2}$ and $T \geqslant t_{1-\alpha/2}$

- Assume the distribution of the variable under consideration in the two populations is not known, nor its variance
- The variances, σ_1^2 and σ_2^2, can be estimated through the usual unbiased estimators S_1^2 and S_2^2 which can be plugged in the test statistic used for normal populations with known variances

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

which, when H_0 is true, follows a $N(0, 1)$ distribution.

Suppose a portfolio manager is interested in whether the daily stock returns of General Electric (GE) and International Business Machines (IBM) had identical means. Let μ_1 denote the mean of the distribution of the returns of GE and μ_2 be the one for IBM, then

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

The manager collects a sample of 15 daily returns of GE, for which $\bar{x}_1 = -0.00092$ and a sample of 25 daily return of IBM, for which $\bar{x}_2 = -0.0034$. Assume the variance of the daily return of GE is $\sigma_1^2 = 0.000268$ while the variance of the daily returns of IBM is $\sigma_2^2 = 0.000485$. Then

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \overset{H_0}{\sim} N(0,1) \qquad \text{so} \qquad t = \frac{-0.00092 + 0.0034}{\sqrt{\frac{0.000268}{15} + \frac{0.00048}{25}}} = 0.406$$

What can the portfolio manager conclude?

$z_{0.90}$	$z_{0.95}$	$z_{0.975}$	$z_{0.99}$	$z_{0.995}$
1.282	1.645	1.96	2.326	2.576

Assume you want to test whether the mean of the percentage change of the exchange rate between euro and dollar and the exchange rate between euro and pound is equal or the first one is lower, then

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases}$$

You have a sample of 8 exchange rates EUR/USD with $\bar{x}_1 = -0.00225$ and $s_1^2 = 0.00002535$ and 5 exchange rates EUR/GBP with $\bar{x}_2 = -0.00137$ and $s_2^2 = 0.0000123$. If we can assume that the variances are equal, then

$$S_p^2 = \frac{7 \cdot 0.00002535 + 4 \cdot 0.0000123}{8 + 5 - 2} = 0.00006264$$

and

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \overset{H_0}{\sim} t_{n_1+n_2-2} \qquad \text{so} \qquad t = \frac{-0.00225 - (-0.00137)}{\sqrt{0.00006264(1/8 + 1/5)}} = -0.196.$$

Conclusion, considering that

$t_{11;0.005}$	$t_{11;0.01}$	$t_{11;0.025}$	$t_{11;0.05}$	$t_{11;0.10}$
-3.106	-2.718	-2.201	-1.796	-1.363

Suppose a portfolio manager is interested in whether the daily stock returns of General Electric (GE) and International Business Machines (IBM) had identical means. In this case s/he does not make any assumption, but collects, for both stocks a large sample, $n_1 = n_2 = 7300$, for which $\bar{x}_1 = 0.00029$ and $s_1^2 = 0.00027$ and $\bar{x}_2 = 0.00027$, $s_2^2 = 0.00032$, then, for testing

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

s/he uses the test statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \overset{H_0}{\sim} N(0,1)$$

The observed statistic is

$$t = \frac{0.00029 - 0.00027}{\sqrt{(0.00027/7300) + (0.00032/7300)}} = 0.07035$$

What can the portfolio manager conclude?

$z_{0.90}$	$z_{0.95}$	$z_{0.975}$	$z_{0.99}$	$z_{0.995}$
1.282	1.645	1.96	2.326	2.576

- Assume a variable in two independent populations follows a Bernoulli distribution with parameters π_1 and π_2, respectively, and interest lies in testing whether the proportion of successes in the two populations is equal

$$H_0 : \pi_1 = \pi_2$$

- The alternatives may be

$$H_1 : \pi_1 > \pi_2 \qquad H_1 : \pi_1 < \pi_2 \qquad H_1 : \pi_1 \neq \pi_2$$

- Considering the difference $\pi_D = \pi_1 - \pi_2$, the null hypothesis becomes:

$$H_0 : \pi_D = 0$$

and the alternative hypothesis:

$$H_1 : \pi_D > 0 \qquad H_1 : \pi_D < 0 \qquad H_1 : \pi_D \neq 0$$

- When the sample sizes, n_1 and n_2, are large enough to estimate π_1 and π_2 we can use the sample means as, thanks to the CLT, they follow a normal distribution. Specifically:

$$\hat{\pi}_1 = \bar{X}_1 \sim N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \qquad \hat{\pi}_2 = \bar{X}_2 \sim N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

- When the sample sizes are big enough, the test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\bar{X}_p(1-\bar{X}_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where $\bar{X}_p$ is the joint estimator of π:

$$\bar{X}_p = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} X_{2i} + \sum_{i=1}^{n_1} X_{1i}\right) = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1+n_2}$$

Assume an holding has two branches both giving personal loans and is interested in checking whether the proportion of loans given is equal in the two branches or the first brach concedes more loans.

A sample of 125 requests is examined from the first branch, finding that 35 where given the loan, and from the second brach a sample of 150 requests were examined, finding that 32 obtained the loan.The hypothesis of interest is

$$\begin{cases} H_0 : \pi_1 - \pi_2 = 0 \\ H_1 : \pi_1 - \pi_2 > 0 \end{cases}$$

The test statistic is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\bar{X}_p(1-\bar{X}_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \overset{H_0}{\sim} N(0,1)$$

and $\bar{x}_1 = \frac{35}{125} = 0.28$, $\bar{x}_2 = \frac{32}{150} = 0.213$, $\bar{x}_p = \frac{35+32}{125+150} = 0.244$.

The observed value of the test statistic is

$$t = \frac{0.28 - 0.213}{\sqrt{0.244(1 - 0.244(1/125 + 1/150)}} = 1.28235$$

What can one conclude?

$z_{0.90}$	$z_{0.95}$	$z_{0.975}$	$z_{0.99}$	$z_{0.995}$
1.2815	1.645	1.96	2.326	2.576

(p-value$= 0.09986$)