Review of Statistics

First Version - August 29, 2016

Present Version - September 19, 2020

Review of Statistics

`What is Statistics?`

Do mean earnings differ from men and women, and if so, by how much?

One way to answer this question would be to perform an exhaustive survey of the population measuring the earnings of each worker. And, next comparing the mean, for example.

Do mean earnings differ from men and women, and if so, by how much?

One way to answer this question would be to perform an exhaustive survey of the population measuring the earnings of each worker. And, next comparing the mean, for example.

Review of Statistics

The key insight of statistics is that one can learn about the population distribution of earnings simply by selecting a random sample rather than survey the entire population.

This is called "to draw statistical inferences about the population".

Review of Statistics

The key insight of statistics is that one can learn about the population distribution of earnings simply by selecting a random sample rather than survey the entire population.

This is called "to draw statistical inferences about the population".

Section 1

Estimation

Estimators

Assume that:

- Y is a random variable whose unknown mean and variance are μ_Y and σ_Y^2;
- unfortunately you do not have access to the entire population but only to a random sample of n i.i.d observations $Y_1, \ldots, Y_n$ drawn from it.

How do you exploit the information contained in the sample to guess the true unknown value of μ_Y?

Estimators

- A first "natural" way to answer this question would be to compute the sample average $\overline{Y}$.
- This is not the only way. One could simply using the first observation Y_1 or the last one Y_n. Alternatively one could take central one $Y_{\frac{n+1}{2}}$.
- In principle any function of the n components can be use to guess the true value of μ_Y.

An estimator is a function of $Y_1, \ldots, Y_n$ representing a random drawn from a population.

Terminology

To avoid confusion keep in mind that

- because of the randomness in selecting the sample an estimator is a random variable (with its proper distribution, mean, variance etc...).

- an estimate is the numerical value of the estimator when it is actually computed using data from a realized sample. An estimate is a nonrandom number.

Properties of an estimator

Since there are many possible estimators for an unknown μ_Y, how can we choose among them which are to be considered "good" or "better"?

In general we would like

- an estimator to get as close as possible to the unknown true value, at least in some average sense;
- the sampling distribution of an estimator to be as tightly centered on the unknown value as possible.

Properties of an estimator

Suppose you evaluate an estimator many times over different random samples:

It is reasonable to hope that, in expected value, you would get the correct value.

Unbiasedness. Let $\hat{\mu}_Y$ be an estimator for μ_Y, then $\hat{\mu}_Y$ is unbiased if

$$E(\hat{\mu}_Y) = \mu_Y \quad ,$$

where $E(\hat{\mu}_Y)$ is the mean of the sampling distribution of $\hat{\mu}_Y$.

example. The sample average is an unbiased estimator of μ_Y if the sample is random.

It is reasonable to hope that, in expected value, you would get the correct value.

Unbiasedness. Let $\hat{\mu}_Y$ be an estimator for μ_Y, then $\hat{\mu}_Y$ is unbiased if

$$E(\hat{\mu}_Y) = \mu_Y \quad ,$$

where $E(\hat{\mu}_Y)$ is the mean of the sampling distribution of $\hat{\mu}_Y$.

example. The sample average is an unbiased estimator of μ_Y if the sample is random.

It is desirable that when the sample size is large the uncertainty about the value of μ_Y arising from random variations in the sample becomes very small. Formally,

Consistency. Let $\hat{\mu}_Y$ be an estimator for μ_Y, then $\hat{\mu}_Y$ is consistent for μ_Y if when $n \to \infty$

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y \quad ,$$

where $\xrightarrow{p}$ means converge in probability.

example. The sample average is a consistent estimator of μ_Y if the sample is random.

It is desirable that when the sample size is large the uncertainty about the value of μ_Y arising from random variations in the sample becomes very small. Formally,

Consistency. Let $\hat{\mu}_Y$ be an estimator for μ_Y, then $\hat{\mu}_Y$ is consistent for μ_Y if when $n \to \infty$

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y \quad ,$$

where $\xrightarrow{p}$ means converge in probability.

example. The sample average is a consistent estimator of μ_Y if the sample is random.

Among unbiased estimators it is reasonable to pick the estimator with the tightest sampling distribution.

Efficiency. If $\hat{\mu}_Y$ and $\bar{\mu}_Y$ are two unbiased estimators for μ_Y, then $\hat{\mu}_Y$ is said to be more efficient than $\bar{\mu}_Y$ if

$$\text{var}(\hat{\mu}_Y) < \text{var}(\bar{\mu}_Y) \ .$$

Among unbiased estimators it is reasonable to pick the estimator with the tightest sampling distribution.

Efficiency. If $\hat{\mu}_Y$ and $\bar{\mu}_Y$ are two unbiased estimators for μ_Y, then $\hat{\mu}_Y$ is said to be more efficient than $\bar{\mu}_Y$ if

$$\mathrm{var}(\hat{\mu}_Y) < \mathrm{var}(\bar{\mu}_Y) \quad .$$

Example. Assume Y is a random variable normally distributed with the mean equal to μ_Y and the variance to σ_Y^2. We consider in turn two different estimators for μ_Y

- $\overline{Y}$, which we know is unbiased and consistent for μ_Y;
- $\overline{Y} + \frac{1}{n}$.

First,

- $E[\overline{Y} + \frac{1}{n}] = \mu_Y + \frac{1}{n}$, showing that this estimator is biased; $\frac{1}{n}$ represents the bias;
- when n grows larger $\overline{Y} + \frac{1}{n}$ tends to μ_Y since $\overline{Y}$ tends to μ_Y for the lln while $\frac{1}{n}$ to 0.

Second, $VAR[\overline{Y}] = VAR[\overline{Y} + \frac{1}{n}] = \frac{\sigma^2}{n}$.

Example. Assume Y is a random variable normally distributed with the mean equal to μ_Y and the variance to σ_Y^2. We consider in turn two different estimators for μ_Y

- $\overline{Y}$, which we know is unbiased and consistent for μ_Y;
- $\overline{Y} + \frac{1}{n}$.

First,

- $E[\overline{Y} + \frac{1}{n}] = \mu_Y + \frac{1}{n}$, showing that this estimator is biased; $\frac{1}{n}$ represents the bias;
- when n grows larger $\overline{Y} + \frac{1}{n}$ tends to μ_Y since $\overline{Y}$ tends to μ_Y for the lln while $\frac{1}{n}$ to 0.

Second, $VAR[\overline{Y}] = VAR[\overline{Y} + \frac{1}{n}] = \frac{\sigma^2}{n}$.

Example. Assume Y is a random variable normally distributed with the mean equal to μ_Y and the variance to σ_Y^2. We consider in turn two different estimators for μ_Y

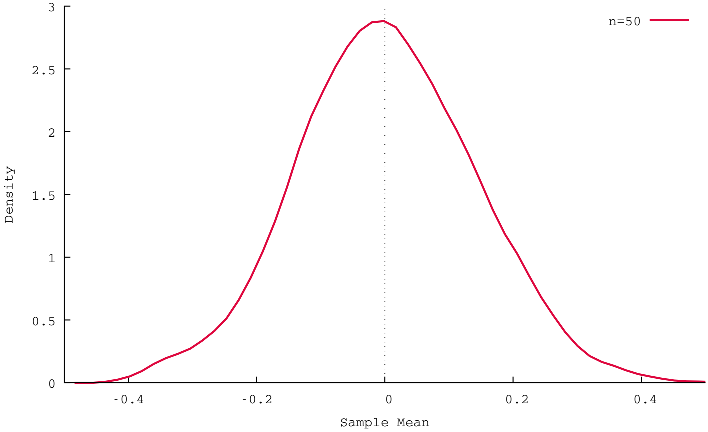
- $\overline{Y}$, which we know is unbiased and consistent for μ_Y;
- $\overline{Y} + \frac{1}{n}$.

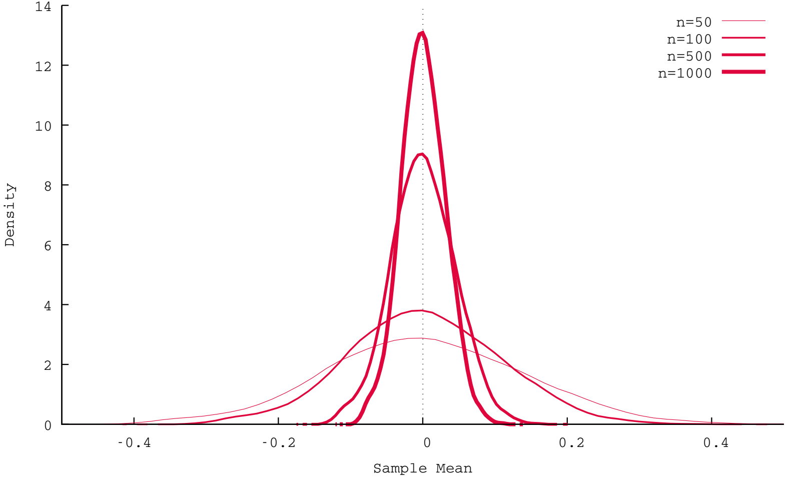
First,

- $E[\overline{Y} + \frac{1}{n}] = \mu_Y + \frac{1}{n}$, showing that this estimator is biased; $\frac{1}{n}$ represents the bias;
- when n grows larger $\overline{Y} + \frac{1}{n}$ tends to μ_Y since $\overline{Y}$ tends to μ_Y for the lln while $\frac{1}{n}$ to 0.

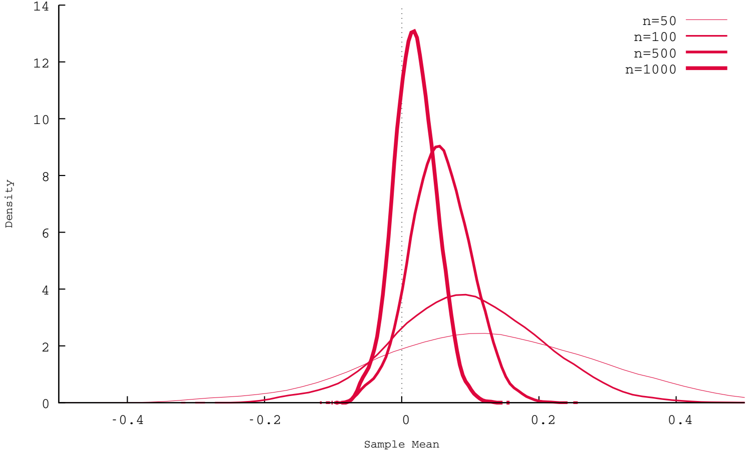
Second, $VAR[\overline{Y}]=VAR[\overline{Y} + \frac{1}{n}]=\frac{\sigma^2}{n}$.

$\overline{Y}$




n=50
n=100
n=500
n=1000
Density
Sample Mean
0
2
4
6
8
10
12
14
-0.4
-0.2
0
0.2
0.4

$$\overline{Y} + \frac{1}{n}$$



Exercise 1. Let Y be a rv with mean μ_Y and variance σ_Y^2. Consider an iid random sample $Y_1, Y_2, \dots, Y_n$.

Exercise 1. Let Y be a rv with mean μ_Y and variance σ_Y^2. Consider an iid random sample $Y_1, Y_2, \ldots, Y_n$. Prove that as an estimator of μ_Y the sample average $\overline{Y}$ is

- the most efficient among those that are weighted averages of $Y_1, \ldots, Y_n$. [To see the intuition compare $\overline{Y}$, Y_1 and $\hat{\hat{Y}} = \frac{1}{n}(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \ldots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n)$]
- the least squares estimator for μ_Y.

Estimator for μ_Y

To estimate μ_Y we use $\overline{Y}$ since it is the Best Linear Unbiased Estimator for μ_Y. It is a BLUE estimator.

Estimator for μ_Y

To estimate μ_Y we use $\overline{Y}$ since it is the Best Linear Unbiased Estimator for μ_Y. It is a BLUE estimator.

Remark. Remember that everything holds only in case of random samples. For nonrandom samples $\overline{Y}$ is typically biased.

Estimator for σ_Y^2

Let Y be a rv with mean μ_Y and variance σ_Y^2. Show that the sample variance

$$s_Y^2 = \frac{1}{n}\sum_i (Y_i - \overline{Y})^2 \quad .$$

is a biased estimator for σ_Y^2 and propose an unbiased alternative.

Estimator for σ_Y^2

The corrected sample variance s_Y^2, defined as

$$s_Y^2 = \frac{1}{n-1}\sum_i (Y_i - \overline{Y})^2$$

is an unbiased and consistent estimator of the population variance σ_Y^2. Note:

- the population mean μ_Y is replaced by the sample mean $\overline{Y}$.
- instead of n we divide by (n-1). This is due to the fact that using $\overline{Y}$ instead of μ_Y introduces a small downward bias in $(Y_i - \overline{Y})^2$ that is corrected by dividing by n-1.

Estimator for σ_Y^2

Remark. Dividing by (n-1) is called a degrees of freedom correction: estimating the mean uses up 1 degree of freedom of the data (part of the info contained in the sample) and only n-1 are left.

The standard error of $\overline{Y}$

Since

- the standard deviation of the sampling distribution of $\overline{Y}$ is $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$;
- $s_Y^2 \xrightarrow{p} \sigma_Y^2$ (consistency) ,

then one is justified using $s_Y/\sqrt{n}$ as an estimator of $\sigma_{\overline{Y}}$. $s_Y/\sqrt{n}$ is called the standard error of $\overline{Y}$ and is denoted $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$.

Exercise 2. Consider two rv X and Y with means and variance μ_X, σ_X and μ_Y, σ_Y respectively. Let σ_{XY} denote the covariance between X and Y. Show that the sample covariance

$$s_{XY} = \frac{1}{n-1} \sum (X_i - \overline{X})(Y_i - \overline{Y}) \ ,$$

is an unbiased estimator for σ_{XY}.

Estimator for σ_{XY}

The corrected sample covariance S_{XY}

$$s_{XY} = \frac{1}{n-1} \sum (X_i - \overline{X})(Y_i - \overline{Y}) ,$$

is an unbiased and consistent estimator for σ_{XY}.

Section 2

Parametric Testing

Introduction and terminology

Statistical testing provides a formal framework in which a researcher can try to answer a yes/no question based on a random sample of data. The two main building blocks of a statistical test are:

- null hypothesis H_0, the hypothesis to be tested;
- alternative hypothesis H_1, the hypothesis against which H_0 is tested.

Introduction and terminology

	H_0 is True	H_0 is False
Reject H_0	Error type I (False positive)	Correct inference (True positive)
Fail to Reject H_0	Correct inference (True negative)	Error type II (False negative)

Size of the test is the probability of incorrectly rejecting H_0 when H_0 is true, that is the probability to make a type I error.

Power of the test is the probability of correctly rejecting H_0 when H_0 is false.

Subsection 1

Hypothesis tests concerning the population mean

The null hypothesis H_0 is that the population mean $E(Y) = \mu_Y$ takes on a specific value denoted μ_0

$$H_0 : \quad E(Y) = \mu_0 \quad .$$

The alternative hypothesis H_1 specifies what is true if the null hypothesis is not.

- The most general alternative hypothesis is

$$H_1 : \quad E(Y) \neq \mu_0 \quad ,$$

  known as the two-sided alternative hypothesis because it allows E(Y) to be either less or greater than μ_0;

- other specifications of the alternative hypothesis are, for example,

$$H_1 : \quad E(Y) \geq \mu_0 \quad \text{or} \quad H_1 : \quad E(Y) \leq \mu_0$$

  known as the one-sided alternative hypothesis.

The problem we face is to use the information contained in a random sample to decide if we

- reject H_0

- fail to reject H_0 since we do not have enough evidence against it. This is $\neq$ from accepting H_0.

- In any give sample $Y_1, \ldots, Y_n$ the sample average $\overline{Y}$ is in general different from the hypothesized value μ_0.

  This is caused by either the following two reasons:

  - the true $\mu_Y \neq \mu_0$ (H_0 is false);
  - because of the random sampling.

- Sadly it is impossible to distinguish between these two possibilities with certainty.

  However it is possible to do a probabilistic calculation that permits testing H_0 in a way that accounts for sampling uncertainty.

- This calculation involves using the data to compute the p-value associated with H_0.

p-value (intuition)

Let's consider a random sample of students drawn from this class.The average age $\overline{Y}$ of the sample is 23.4 Assume that the null hypothesis we would like to test is $H_0 : E(Y) = 22$.

The p-value associated with H_0 is the probability of drawing a value of $\overline{Y}$ at least as different from 22 as the observed value of 23.4 by pure random sampling variation and assuming that H_0 is true.

p-value (intuition)

Let's consider a random sample of students drawn from this class.The average age $\overline{Y}$ of the sample is 23.4 Assume that the null hypothesis we would like to test is $H_0 : E(Y) = 22$.

The p-value associated with H_0 is the probability of drawing a value of $\overline{Y}$ at least as different from 22 as the observed value of 23.4 by pure random sampling variation and assuming that H_0 is true.

p-value (intuition)

If the probability of drawing a value of Y at least as different from 22 as the observed value of 23.4 by pure random sampling variation (namely the p-value)

- is large, say 0.5, it means that under H_0 is would be likely to draw 23.4;

- is small, say 0.05, it means that under H_0 is would be very unlikely to draw 23.4;

p-value (intuition)

If the probability of drawing a value of Y at least as different from 22 as the observed value of 23.4 by pure random sampling variation (namely the p-value)

- is large, say 0.5, it means that under H_0 is would be likely to draw 23.4; [UNREASONABLE to REJECT H_0];

- is small, say 0.05, it means that under H_0 is would be very unlikely to draw 23.4; [REASONABLE to REJECT H_0].

p-value (definition)

Let

- $\overline{Y}^{act}$ be the value of the sample average actually computed with the sample at hand

- Pr_{H_0} be the probability computed under the null hypothesis (that is computed assuming that $E(Y_i) = \mu_0$).

p-Value. The p-value is defined as

$$\text{p-value} = Pr_{H_0}\left[|\overline{Y} - \mu_0| > |\overline{Y}^{act} - \mu_0|\right] .$$

p-value (definition)

Let

- $\overline{Y}^{act}$ be the value of the sample average actually computed with the sample at hand
- Pr_{H_0} be the probability computed under the null hypothesis (that is computed assuming that $E(Y_i) = \mu_0$).

p-Value. The p-value is defined as

$$p - value = Pr_{H_0}\left[|\overline{Y} - \mu_0| > |\overline{Y}^{act} - \mu_0| \right] .$$

Remark. For continuous rv this probability is the area in the tails of the distribution, under the null hypothesis, of $\overline{Y}$ beyond $\mu_0 \pm |\overline{Y}^{act} - \mu_0|$.

Remark. Hence to calculate the p-value we need to know what is the distribution of $\overline{Y}$ under the null hypothesis H_0. Since you master the CLT, this is not a problem anymore at least when n is large.

p-value (computation when σ_Y^2 is known)

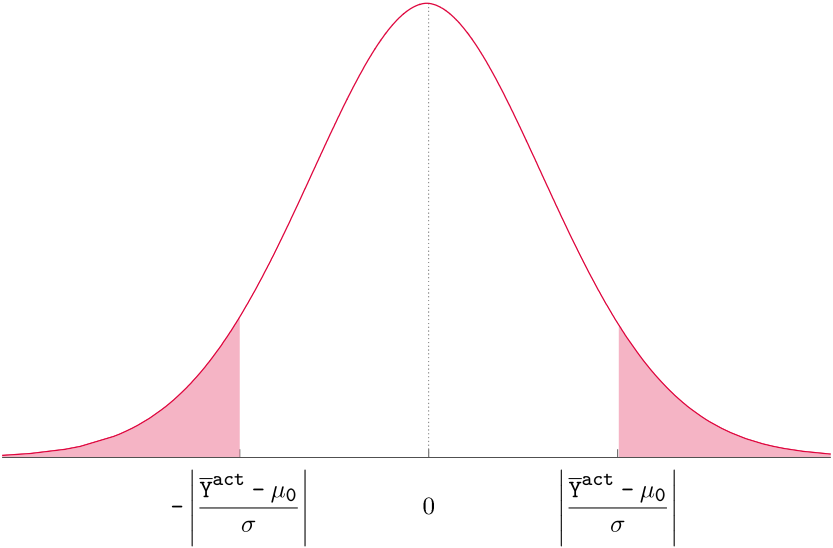
When n is large, under the null hypothesis $H_0: E(Y) = \mu_0$

$$\overline{Y} \xrightarrow{d} N(\mu_0, \frac{\sigma_Y^2}{n}) \quad ,$$

where $\sigma_{\overline{Y}}^2 = \frac{\sigma_Y^2}{n}$ is known by assumption. Then,

$$\frac{\overline{Y} - \mu_0}{\sqrt{\frac{\sigma_Y^2}{n}}} \xrightarrow{d} N(0,1) \quad .$$

So the p-value is equivalent to the probability of obtaining $(\overline{Y} - \mu_0)/\sigma_{\overline{Y}}$ greater than $(\overline{Y}^{act} - \mu_0)/\sigma_{\overline{Y}}$ in absolute value.


$-\left|\frac{\overline{Y}^{act} - \mu_0}{\sigma}\right|$
0
$\left|\frac{\overline{Y}^{act} - \mu_0}{\sigma}\right|$

p-value (computation when σ_Y^2 is unknown)

When σ_Y^2 is unknown the procedure remains essentially the same. We just need to replace σ_Y^2 with its consistent estimator s_Y^2. In this case, again for the CLT,

$$\frac{\bar{Y} - \mu_0}{\sqrt{\frac{s_Y^2}{n}}} = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} \xrightarrow{d} N(0, 1) \ ,$$

where $(\bar{Y} - \mu_0)/SE(\bar{Y})$ has a special name, the t-statistics or t-ratio.

Test procedure

In both cases the procedure to test $H_0 : \mu_Y = \mu_0$ against $H_1 : \mu_Y \neq \mu_0$ is the same. It consists in three steps:

- based on your sample and under H_0 compute the t-ratio

$$t^{act} = \frac{\overline{Y}^{act} - \mu_0}{SE(\overline{Y})} \quad ;$$

- obtain the corresponding p-value using

$$p - value = Pr_{H_0}\left[|t| > |t^{act}|\right] \quad ,$$

  where, for the CLT, t is distributed according to N(0,1);

- decide if the p-value is sufficiently small to reject H_0.

Practice

Exercise 4. Consider a random sample drawn from a Normal distribution with unknown mean μ_x and variance 1. The sample average $\bar{X}$ is found to be 5.4.

- Assume n=10 and compute the p-value associated with the test of $H_0 : \mu_x = 5$ versus $H_1 : \mu_x \neq 5$.
- Repeat the exercise for n=100, n=5. Comment.
- Assume n=100 and $\bar{X} = 7.5$ and compute the p-value associated with the test of $H_0 : \mu_x = 5$ versus $H_1 : \mu_x \neq 5$.
- Assume n=10 and $\bar{X} = 5.4$ and compute the p-value associated with the test of $H_0 : \mu_x = 5$ versus $H_1 : \mu_x < 5$.

Subsection 2

Hypothesis test with a pre-specified significance level

Typically we give a preferential treatment to the null hypothesis H_0 (Ex. with the legal system). In this case

- type I error: H_0 is true but you reject it (False Positive)

is the most dangerous.

For this reason often we set in advance the probability of making the type I error.

This probability is called significance level of the test. With a pre-specified significance level, testing H_0 does not require to explicitly calculate the p-value.

Typically we give a preferential treatment to the null hypothesis H_0 (Ex. with the legal system). In this case

- type I error: H_0 is true but you reject it (False Positive)

is the most dangerous.

For this reason often we set in advance the probability of making the type I error.

This probability is called significance level of the test. With a pre-specified significance level, testing H_0 does not require to explicitly calculate the p-value.

Test procedure

- set the significance level, say 5%;

- obtain from the statistical table the corresponding critical value;

  it is the value for which the area under the tails (left and right) is exactly 5%; in case of a significance level of 5% is |1.96|. [visualization]

- compute the actual value of the t statistics

$$t^{act} = \frac{\overline{Y}^{act} - \mu_0}{SE(\overline{Y})} \quad ,$$

  based on the available sample;

- apply the rule

$$\text{Reject } H_0 \text{ if } |t^{act}| > 1.96 \ .$$

Confidence intervals (definition)

The rejection rule in a test with 5% significance level reads

$$\text{Reject } H_0 \text{ if } |t| > t_{5\%} \;.$$

This implies that the set of values associated with non-rejection at the 5% level can be written as

$$-t_{5\%} < \frac{\overline{Y} - \mu_Y}{SE(\overline{Y})} < t_{5\%} \;.$$

As a consequence

$$\overline{Y} - SE(\overline{Y})t_{5\%} < \mu_Y < \overline{Y} + SE(\overline{Y})t_{5\%} \;.$$

The last interval represents a 95% confidence interval for the population mean.

Confidence intervals (definition)

The rejection rule in a test with 5% significance level reads

$$\text{Reject } H_0 \text{ if } |t| > t_{5\%} \ .$$

This implies that the set of values associated with non-rejection at the 5% level can be written as

$$-t_{5\%} < \frac{\overline{Y} - \mu_Y}{SE(\overline{Y})} < t_{5\%} \ .$$

As a consequence

$$\overline{Y} - SE(\overline{Y})t_{5\%} < \mu_Y < \overline{Y} + SE(\overline{Y})t_{5\%} \ .$$

The last interval represents a 95% confidence interval for the population mean.

Confidence intervals (interpretation)

The correct interpretation of a confidence interval is

- before the sample is drawn, the random interval has a 95% chance of containing the true μ_Y;

- after the sample is drawn either the unknown parameter lies in the interval or it does not! For 95% of random samples, it does.

Subsection 3

testing the difference between population means

To illustrate this testing procedure:

- let μ_W be the mean of Y_W, a rv representing the hourly earnings of a group of women recently graduated;
- let μ_M be the mean of Y_M, a rv representing the hourly earnings of a group of men recently graduated;
- assume that you have one sample of n_M men and an independent sample with n_W women.

We aim at testing the null hypothesis $H_0: \mu_M - \mu_W = 0$ against $H_1: \mu_M - \mu_W \neq 0$.

Comparing means from different populations

Since $\overline{Y}_M$ and $\overline{Y}_W$ are constructed from different random samples, they are independent. Then, when n_m and n_w are large, invoking the CLT gives

$$\overline{Y}_M - \overline{Y}_W \xrightarrow{d} N(\mu_M - \mu_W, \frac{\sigma_M^2}{n_M} + \frac{\sigma_W^2}{n_W}) \quad .$$

When σ_M^2 and σ_W^2 are unknown we can compute the t-ratio for this test as

$$t = \frac{\overline{Y}_m - \overline{Y}_w - d_0}{SE(\overline{Y}_m - \overline{Y}_w)} \xrightarrow{d} N(0, 1) \quad ,$$

where $SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$ and follow the usual procedure.

Exercise 5. Data on fifth grade (math and reading) score for 420 school districts in California yield $\bar{Y} = 646.2$ and $S_Y = 19.5$.

- Build a confidence interval at 95% level for the unknown μ_Y.
- When districts are divided into districts with large classes (more than 20 students) and districts with small classes (less than 20 students) we get

	$\bar{Y}$	S_Y	n
small	657.4	19.4	238
large	650	17.9	182

  Is there statistically significant evidence that districts with smaller classes have higher average test score?

Subsection 4

why t-ratio?

- If n is large then the CLT implies that

$$\text{t-ratio} = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{s_Y^2}{n}}} = \frac{\overline{Y} - \mu_0}{SE(\overline{Y})} \xrightarrow{d} N(0, 1) \quad ,$$

- If n is small we do not know the distribution of the t-ratio. However if we are willing to assume that Y is Normally distributed then

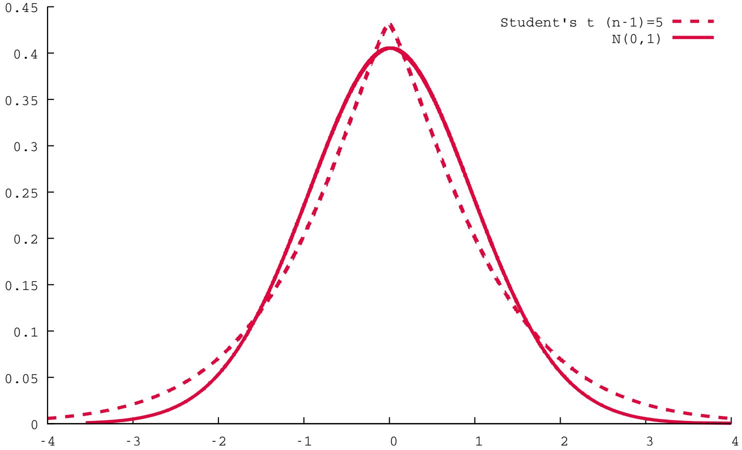
$$\text{t-ratio} = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{s_Y^2}{n}}} = \frac{\overline{Y} - \mu_0}{SE(\overline{Y})} \quad \sim \quad \text{Student's } t_{(n-1)} \quad .$$

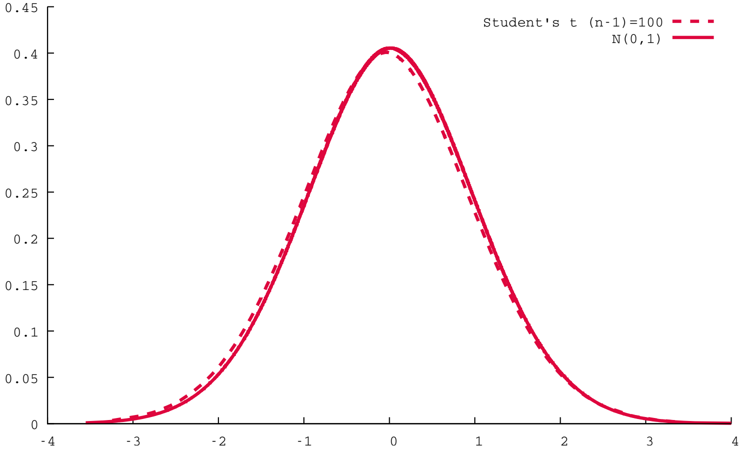
- If n is large then the CLT implies that

$$\text{t-ratio} = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{s_Y^2}{n}}} = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} \xrightarrow{d} N(0, 1) \quad ,$$

- If n is small we do not know the distribution of the t-ratio. However if we are willing to assume that Y is Normally distributed then

$$\text{t-ratio} = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{s_Y^2}{n}}} = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} \quad \sim \quad \text{Student's } t_{(n-1)} \quad .$$


Student's t (n-1)=5
N(0,1)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4



Student's t (n-1)=100
N(0,1)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4

Section 3

Distribution Free Testing

Distribution free testing

- Usually we test statistical hypothesis with respect to a random variable Y whose probability distribution p(Y) is known.

- In many situation, however, we do not know p(Y) but we need to do inference on the phenomenon summarized by Y.

- Nonparametric testing procedures fill this gap imposing only two requirements
  1. the phenomenon of interest must be described as a continuous random variable Y;
  2. the realizations of Y must be replaceable with the corresponding rank, i.e. with natural numbers 1,...,n once they have been ordered.

Distribution free test

Let's set the stage.

- Let's consider

  $(Y_1, \ldots, Y_n)$, a sample of n i.i.d. observations;
  $(R_1, \ldots, R_n)$, the corresponding ranks .

- As usual, behind a specific random sample and its ranks there are two random variables Y and R.

- In particular R is known as the random variable Rank. Note that even if $Y_1, \ldots, Y_n$ are i.i.d. $R_1, \ldots, R_n$ are not independent since

$$\sum_{i=1}^{n} r_i = \frac{n(n+1)}{2} \quad .$$

Sign Test - Fisher

Let θ_Y be the unknown [median] of a continuous rv Y and $(Y_1, \ldots, Y_n)$ a random sample drawn from the population with the aim of testing $H_0 : \theta_Y = \theta_0$ against $H_1 : \theta_Y \neq \theta_0$.

Under the null hypothesis $H_0 : \theta_Y = \theta_0$

- $Pr(Y < \theta_0) = P(Y > \theta_0) = 0.5$ implying that
  $Pr(Y_i < \theta_0) = Pr(Y_i > \theta_0) = 0.5 \quad i = 1, \ldots, n$ and
  $Pr(D_i < 0) = Pr(D_i > 0) = 0.5 \quad i = 1, \ldots, n,$ where $D_i = X_i - \theta_0$

- we can define

$$s(d_i) = \begin{cases} 1 & d_i > 0 \\ 0 & d_i < 0 \end{cases},$$

  and the associated rv $S = \sum_i s(d_i)$.

Sign Test - Fisher

Let θ_Y be the unknown [median] of a continuous rv Y and $(Y_1, \dots, Y_n)$ a random sample drawn from the population with the aim of testing $H_0 : \theta_Y = \theta_0$ against $H_1 : \theta_Y \neq \theta_0$.

Under the null hypothesis $H_0 : \theta_Y = \theta_0$

- $Pr(Y < \theta_0) = P(Y > \theta_0) = 0.5$ implying that
  $Pr(Y_i < \theta_0) = Pr(Y_i > \theta_0) = 0.5 \quad i = 1, \dots, n$ and
  $Pr(D_i < 0) = Pr(D_i > 0) = 0.5 \quad i = 1, \dots, n$, where $D_i = X_i - \theta_0$

- we can define

$$s(d_i) = \begin{cases} 1 & d_i > 0 \\ 0 & d_i < 0 \end{cases} ,$$

  and the associated rv $S = \sum_i s(d_i)$.

Sign Test - Fisher

The intuition behind the test is very simple:

- under H_0, S^{act} should not be too far away from the mean of the (so far unknown) distribution of S.
- Then the test procedure is standard and depends on the specification of the alternative hypothesis H_1, if it is one-sided or two-sided.

Problem. It remains to establish what is the distribution of S. Does it require to specify a distribution for Y? Or is it free from the distribution of Y?

Sign Test - Fisher

Under H_0 the distribution of S is given by

$$P(S = 0) = Pr\left[\sum_i s(D_i) = 0\right] = 0.5^n$$

$$P(S = n) = Pr\left[\sum_i s(D_i) = n\right] = 0.5^n$$

$$P(S = s) = \binom{n}{s} 0.5^n$$

that is S is a Binomial rv with parameters $(n, 0.5)$ and so it is free from the distribution of X.

Remark. If the median of Y is not θ_0 but another value θ_1 then it is not possible to evaluate $P[(X - \theta_0) > 0]$ and the distribution free property disappears.

Signed Rank Test - Wilcoxon

Procedure to test $H_0 : \theta = \theta_0$, that is the median of a symmetric continuous rv X is equal to θ_0. Let

$(x_1, \ldots, x_n)$ [sample of n independent Bernoulli trials]

$(d_1, \ldots, d_n)$ [$d_i = x_i - \theta_0$]

$(|d_1|, \ldots, |d_n|)$ [absolute values of d_i]

$(r_1, \ldots, r_n)$ [ranks of $|d_i|$] ,

the Wilcoxon test statistics reads

$$t = \sum_{i=1}^{n} r_i s(d_i) ,$$

where $s(d_i) = 1$ if $d_i > 0$ and $s(d_i) = 0$ if $d_i < 0$.

Practice

Exercise 6. Consider the following random sample with n=4: $x_1 = 9$, $x_2 = 0$, $x_3 = -3$ and $x_4 = 3$. Assume $\theta_0 = 5$. Compute the Wilcoxon test statistics for this sample.

Distribution of the Wilcoxon statistics

The distribution of T is in general unknown. But,

Distribution of the Wilcoxon statistics

Consider a small sample composed by 3 observations (x_1, x_2, x_3). Then all the possible combinations of the $s(d_i)$ values with the corresponding t can be summarized as follows

Ranks			t
1	2	3	
1	1	1	(1+2+3)=6
0	1	1	(2+3)=5
1	0	1	(1+3)=4
1	1	0	(1+2)=3
0	0	1	(3)=3
0	1	0	(2)=2
1	0	0	(1)=1
0	0	0	0(0)=0

Then the distribution of T is

t	0	1	2	3	4	5	6
P(T=t)	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Signed Rank Test - Wilcoxon

In general,

- T assumes values in the range $[0, \frac{n(n+1)}{2}]$;

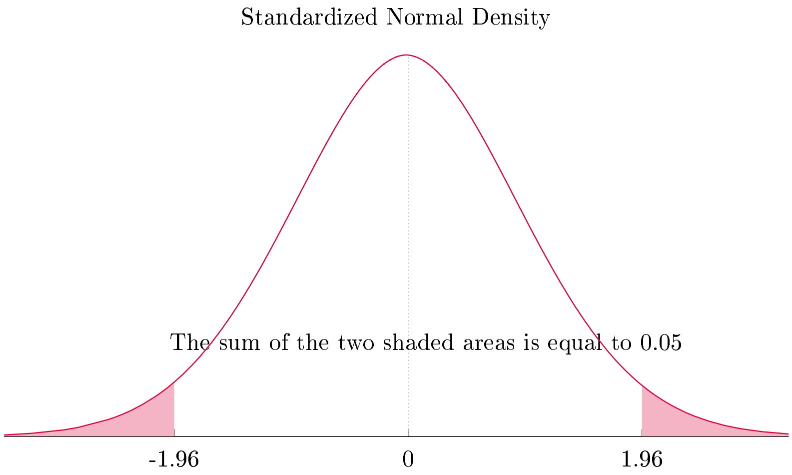
- T is symmetric around its mean and it is free from the distribution of X;

- $E[T] = \sum_{i=1}^{n} r_i E[s(d_i)] = \sum_{i=1}^{n} r_i (1\frac{1}{2} + 0\frac{1}{2}) = \frac{n(n+1)}{4}$

- $Var[T] = \sum_{i=1}^{n} r_i^2 Var[s(d_i)] = \sum_{i=1}^{n} r_i^2 (\frac{1}{2} - \frac{1}{4}) = \frac{n(n+1)(2n+1)}{24}$

The intuition behind this test is simple: under H_0 t should be close to E[T] the mean of T which under symmetry is also the median. The test procedure is then standard.

Hyper-references


Standardized Normal Density
The sum of the two shaded areas is equal to 0.05
-1.96
0
1.96

[back]

Why the median of Y and not the mean?

Because when the distribution of Y is unknown it is always possible to assign the probability to the event $Y - \theta > 0$, that is by definition 0.5.

This is not the case for the event $Y - \mu_Y$, except when Y is symmetric.

[back]