Statistical Data Analysis

Victoria Blanes-Vidal
Manuella Lech Cantuaria
The Maersk Mc-Kinney Møller Institute
Applied AI and Data Science

About the lecturer…

Victoria Blanes-Vidal
Associate Professor, PhD
The Maersk Mc-Kinney Moller Institute
SDU Applied AI and Data Science

Research areas

- Data Science, Epidemiology, Prognostic models, Biostatistics, Mathematical Modelling, Population studies
- Diabetes, Cancer screening (prostate, colorectal), Cardiovascular events, Liver diseases, Respiratory health
- Screening, Early detection, Prediction
- Medical Informatics, Mobile Health
- Environmental Health, Air Quality, Noise pollution, Quality of Life

Teaching

- Statistical Data Analysis
  Victoria Blanes-Vidal & Manuella Lech Cantuaria
  09/08/2021 → …
- Data Science
  Jürgen Herp, Victoria Blanes-Vidal & Esmaeil Nadimi
  01/09/2017 → 01/01/2019
- Datamining
  Jürgen Herp, Victoria Blanes-Vidal & Esmaeil Nadimi
  01/02/2017 → 01/06/2017

Kunstig intelligens kan afsløre alvorlig leversygdom

Kunstig intelligens kan afsløre, om du lider af en livsfarlig leversygdom. Den banebrydende metode har et hold forskere fra Syddansk Universitet og Odense Universitetshospital netop offentliggjort.

Sophie Lysholt Hansen
solha@fyens.dk

Victoria Blanes-Vidal og professor Esmail Nadimi, som er teknisk chef i samarbejdet CAI-X mellem SDU og OUH, der arbejder på kunstig intelligens i sundhedsvæsenet. De har netop offentliggjort deres resultater, hvor kunstig intelligens med en blodprøve kan fortælle, om man skal undersøges nærmere for skrumpelever. Foto: Jakob Haugaard Christiansen

FYN: En computer og en blodprøve kan fremover fortælle, om du lider af arvæv på leveren, der kan føre til livsfarlige tilstande som skrumpelever og kræft.

Forskere fra Det Tekniske Fakultet ved SDU har, i samarbejde med læger fra Odense Universitetshospital, netop offentliggjort en metode, der gennem kunstig intelligens kan finde patienter med arvæv, også kaldet fibrose, på leveren tidligere - og i tide til at få behandling.

- Vores ambition var at udvikle en metode, hvor du med hjælp fra en standard blodprøve hos din praktiserende læge hurtigt og let kan afgøre, om du er i risiko for at have sygdommen. Og det er faktisk lykkedes i vores forsøg, siger Victoria Blanes-Vidal, der er lektor og specialist i biostatistik og dataanalyse ved Mærsk Mc-Kinney Møller Instituttet på SDU, der står bag resultaterne.

At det sker med kunstig intelligens betyder, at særlige matematiske algoritmer bruger resultaterne fra en blodprøve til at give svar på, om patienten har arvæv på leveren og skal i behandling.

- Helt konkret betyder det, at vi kan bruge standard-tilgængelige blodprøver på en mere intelligent måde. Den lægelige hjerne tænker meget på, om den enkelte blodprøve er normal eller for høj, men der er supermeget at hente under overfladen. Derfor har vi sammensat en algoritme, der kan bruge de eksisterende blodprøver bedre, end man gør i dag, siger Katrine Lindvig, der er læge og Ph.D.-studerende på projektet.

Fokus på de syge

300.000 danskere lider af leverfibrose, hvor der dannes arvæv på leveren og kan føre til skrumpelever. Hvis det ikke opdages i tide, kan der ske alvorlige skader på organet, og det kan øge risikoen for leverkræft med op til 40 procent. Derfor er den nye metode helt afgørende, siger overlæge og professor Aleksander Krag, der også er leder af Center for Leverforskning på OUH.

- En af udfordringerne ved leversygdomme er, at de ikke giver nogen symptomer. De fleste udvikler sig langsomt, og man kan være syg i 10-20 år, før du får symptomer, og så er det næsten for sent. Med kunstig intelligens kan vi i langt højere grad stille diagnosen i tide, forklarer han.

Med kunstig intelligens kan man nu udelukke en stor gruppe mennesker med en almindelig blodprøve i stedet for at sende dem til dyre og omfattende scanninger eller biopsier af leveren.

- Det er lykkedes os at udvikle en metode, hvor vi med næsten 99 procents sikkerhed kan udelukke personer på baggrund af en række data og brug af kunstig intelligens. Den resterende gruppe, som er i risiko, kan så blive undersøgt med enten scanning eller biopsi, siger Victoria Blanes-Vidal.

Ofte for sent til lægen

Det betyder, at man i fremtiden kan bruge ressourcerne på de patienter, der faktisk er syge, i stedet for at undersøge patienter, der slet ikke har sygdom i leveren.

Aleksander Krag er professor og overlæge på OUH og leder af Center for leverforskning. Han har været med på projektet i mange år og glæder sig over resultatet. Kunstig intelligens kommer i fremtiden til at bygge bro mellem, hvad vi ved, og hvad sundhedsvæsenet har ressourcer til, siger han. Foto: Sophie Lysholt Hansen

- Der er to ben i det her. Det ene er, at vi diagnosticerer for sent, fordi patienterne ofte først går til deres praktiserende læge, når de har symptomer på leversygdomme - og det kan godt være 15-20 år efter, de faktisk er begyndt at blive syge. Så er der ikke mange gode behandlingsmuligheder tilbage, og vi ved, at 85 procent af dem vil dø inden for fem år. Den gruppe får altså slet ikke den opmærksomhed, de burde have, selvom det er virkelig alvorligt at have skrumpelever, forklarer Katrine Lindvig og fortsætter:

- Det andet er, at alt for mange raske mennesker vader ind på hospitalerne, fordi der er en eller anden markør, der slår ud hos den praktiserende læge, som så henviser til os.

Katrine Lindvig har gennemgået samtlige lever-relaterede henvisninger til OUH i en to-årig periode, og resultaterne er tydelige.

- 54 procent af dem, der er blevet henvist med relation til leveren, er raske. Det er spild af ressourcer, og derfor er det her værktøj hovedet på sømmet, fordi det kan hjælpe os med at sortere dem fra, så vi ikke begynder at undersøge mennesker, der er raske, forklarer hun.

Første resultat

Metoden til at udelukke leverfibrose hos patienterne er det første resultat fra det nye center CAI-X, der er et samarbejde mellem Mærsk Mc-Kinney Møller Instituttet på SDU og OUH, som har til formål at udvikle teknologiske løsninger med eksempelvis kunstig intelligens og robotter til sundhedsvæsenet.

- Lige nu er der en kløft mellem, hvad man ved, kan og gerne vil, og hvad sundhedsvæsenet kan levere og har råd til. Men kunstig intelligens kan være med til at bygge bro mellem de to, fordi computer-hardware kan noget, vi mennesker ikke kan, og som ellers krævede personale og ressourcer, siger Aleksander Krag og fortsætter:

- Jeg kan godt som læge se sygdomme hos folk, hvis de har symptomer, men alle de skjulte ting, kan jeg ikke se. Det kan computer-intelligens, og de kan se mønstre, den menneskelige hjerne ikke kan.

Resultaterne er offentliggjort i tidsskriftet Scientific Reports, og snart går projektet ind i næste fase, der handler om at teste analysemodellen i udlandet.

- Vi skal have den testet og afprøvet på andre grupper i Danmark og udenfor landets grænser, for at se, om vi kan opnå lige så gode resultater der, siger Victoria Blanes-Vidal.

FAKTA

HVORDAN KAN EN COMPUTER FINDE LEVERSYGDOMME?

Målet er, at de praktiserende læger tager almindelige blodprøver fra patienterne og sender dem til analyse. En algoritme, et intelligent lag af matematik, vil støvsuge blodprøverne og sende svar tilbage til den praktiserende læge, som kan give patienten besked om, om personen er rask, skal gå til løbende kontrol eller skal på hospitalet til nærmere undersøgelser og måske behandling.

Forskerne har arbejdet på projektet i næsten 10 år. I 2016 kunne Katrine Lindvig og en række andre forskere offentliggøre algoritmen, der kunne finde tegn på leverfibrose i blodprøver, selvom blodprøverne ved almindelig analyse var normale.

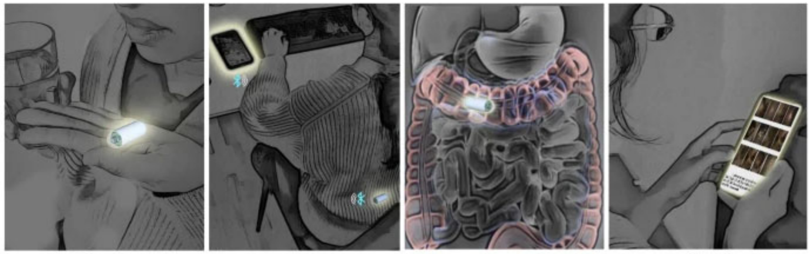
I 2019 fik forskerne flere penge til at videreudvikle på de originale opdagelser, og i 2020 indgik de et samarbejde med Victoria Blanes-Vidal og de andre forskere fra Mærsk-instituttet.

Det er blevet til en mere udvidet algoritme, der ikke bare kan finde tegn på leverfibrose hos den enkelte, men som kan blive klogere for hver gang en patient bliver analyseret.

Den originale algoritme, der kan genkende tegn på leverfibrose i en blodprøve, forventes at blive tilbudt til praktiserende læger på Fyn i juni 2023.

Den mere avancerede algoritme skal først igennem en lang godkendelsesproces, før den kommer på markedet.

CAI-X, SDU OG OUH



Stage 1 Stage 2 Stage 3 Stage 4

Naboer: Svinefarme gør os syge

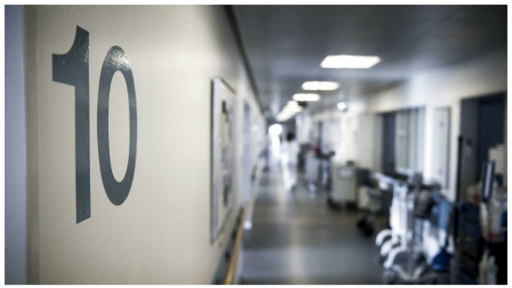
Naboer til Poldanors kæmpefarm beskylder svinefarmerne for at ødelægge deres helbred. Naboer til svinefarme er oftere syge, siger forskere.

LUFTFORURENING

VIDENSKAB ODENSE 9. apr. 2022 - 17:49

Kunstig intelligens skal spotte livsfarlig leversygdom i tide

B.T.



300.000 danskere lever med den livsfarlige sygdom leverfibrose. Ny forskning fra Syddansk Universitet kan være med til at sikre, at sygdommen opdages langt tidligere. **Foto: Asger Ladefoged**

SCIENTIFIC REPORTS
natureresearch

NATURE.COM

Facial erythema detects diabetic neuropathy using the fusion of machine learning, random matrix theory and self organized criticality

Lesson 4:
Hypothesis testing (one sample)

Victoria Blanes-Vidal
Manuella Lech Cantuaria
The Maersk Mc-Kinney Møller Institute
Applied AI and Data Science

Lesson	Week	Date	TOPICS	Teacher
1	35	1/Sep	Introduction to the course Descriptive statistics –Part I	MLC
2	36	8/sep	Descriptive statistics –Part II	MLC
3	37	15/Sep	Probability distributions	MLC
4	38	22/Sep	Hypothesis testing (one sample)	VBV
5	39	29/Sep	Hypothesis testing (two samples)	VBV
6	40	6/Oct	ANOVA one-way	VBV
7	41	13/Oct	R class (Introduction to R and descriptive statistics)	MLC
-	42	20/Oct	NO CLASS (Autum holidays)	
8	43	27/Oct	R class (hypothesis testing + ANOVA)	MLC
9	44	3/Nov	ANOVA two-way	VBV
-	45	10/Nov	NO CLASS	
10	46	17/Nov	Regression analysis	VBV
11	47	24/Nov	Multiple regression	MLC
12	48	1/Dec	Notions of experimental design and questions	VBV+MLC

VBV = Victoria Blanes-Vidal
MLC = Manuella Lech Cantuaria

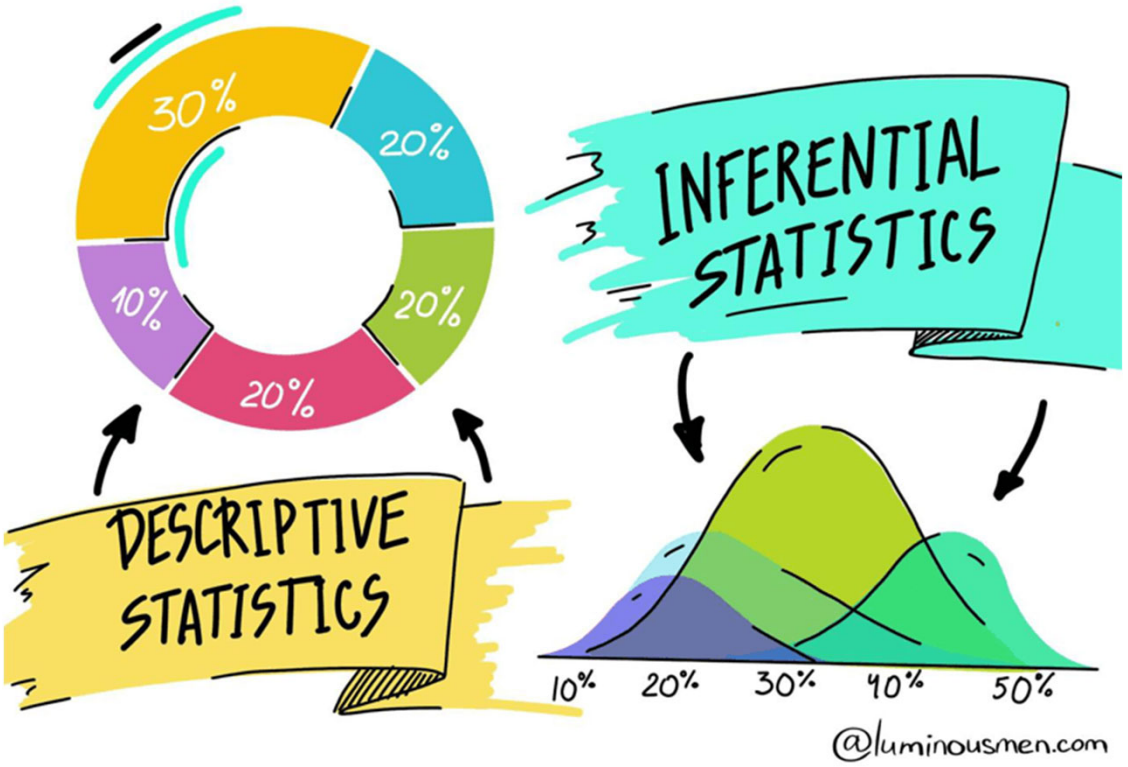
The field of statistics is divided into two major parts:

Lesson	Week	Date	TOPICS	Teacher	
1	35	1/Sep	Introduction to the course Descriptive statistics –Part I	MLC	**Descriptive Statistics**
2	36	8/sep	Descriptive statistics –Part II	MLC	
3	37	15/Sep	Probability distributions	MLC	
4	38	22/Sep	Hypothesis testing (one sample)	VBV	**Inferential Statistics**
5	39	29/Sep	Hypothesis testing (two samples)	VBV	
6	40	6/Oct	ANOVA one-way	VBV	
7	41	13/Oct	R class (Introduction to R and descriptive statistics)	MLC	
-	42	20/Oct	NO CLASS (Autum holidays)		
8	43	27/Oct	R class (hypothesis testing + ANOVA)	MLC	
9	44	3/Nov	ANOVA two-way	VBV	
-	45	10/Nov	NO CLASS		
10	46	17/Nov	Regression analysis	VBV	
11	47	24/Nov	Multiple regression	MLC	
12	48	1/Dec	Notions of experimental design and questions	VBV+MLC	

Lesson 4. Hypothesis testing (one sample)

1. **What is inferential statistics?**
2. **Hypothesis testing for the mean**
3. **Confidence Intervals for the mean**
4. **Type I error and Type II error**
5. **Two in-class exercises** 

What is inferential statistics?



Descriptive Statistics

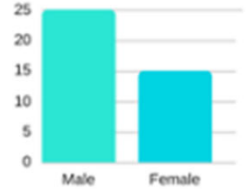
Used to describe, organize, summarize information.

	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

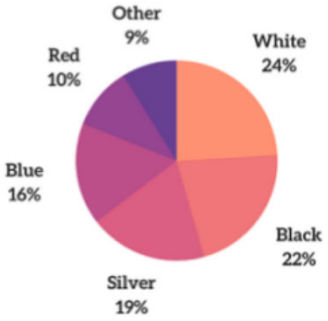
...

Raw data

What is the proportion of males vs. females of the respondents?



What is the favourite car color of the respondents?



What is the mean age of the respondents?

Mean age of the respondents = 39 yr

What is the variability of the age of the respondents?

Standard deviation of the age of the respondents = 3 yr

Inferential Statistics

Inferential statistics provides **a way to draw conclusions about a population based on a set of sample data.**

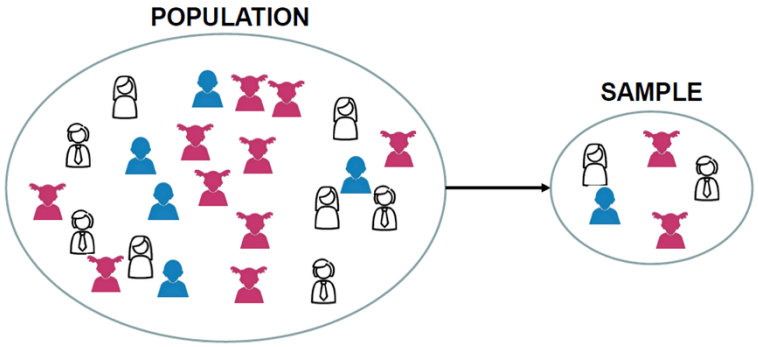
Before we proceed, let's refresh the concepts of population and sample:

Population vs. Sample

A **population** is the entire group that you want to draw conclusions about.

It is usually impossible to examine each member of the population individually.

So scientists choose a representative subset of the population, called a statistical sample.



A **sample** is the specific group that you will collect data from.

NOTE: In statistics, a "population" doesn't always refer to "people". It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, organisms, etc.

Population vs. Sample

We want to find out…

which percentage of residents in Denmark support a certain law.

What is the "Population" and what is the "Sample"?

All residents in Denmark

5000 adult residents who responded to a questionnaire to determine if he/she favors that specific law.

Population vs. Sample

We want to find out…

what is the mean weight of red squirrels living in Sjæland?

What is the "Population" and what is the "Sample"?

100 squirrels living in Sjæland, that we could catch, weight and release.

All red squirrels living in Sjæland

Population vs. Sample

We want to find out…

What is the concentration of a certain herbicide in drinking water in Odense?

What is the "Population" and what is the "Sample"?

200 water containers (of 100 ml each) taken at 200 different locations (water taps) in Odense.

All drinking water in Odense (ALL water coming out from ALL water taps in ALL households, shops, companies, public buildings, etc, in Odense)

Population vs. Sample

We want to find out…

What is the life (in hours) of a 75-watt light bulb produced by a manufacturer?

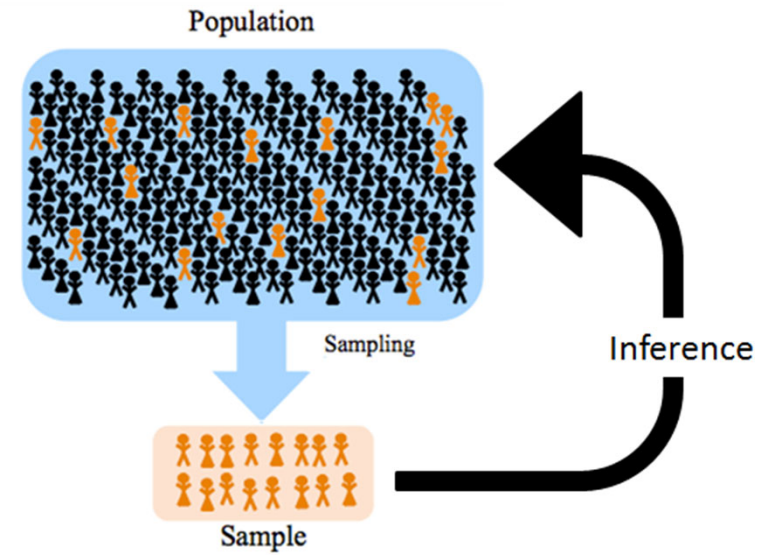
What is the "Population" and what is the "Sample"?

All 75-watt light bulbs produced by the manufacturer

50 light bulbs (75-watt) that we selected to carry out a test to find out their life in hours.

Inferential Statistics

By analyzing the statistical sample and using inferential statistics, we are able to say something about the population from which the sample came.



Inferential statistics are statistical techniques that **allow us to use samples, to draw conclusions about the populations** from which the samples were taken.

Inferential statistics is based on probability distributions theory and uses hypothesis testing.

Lesson 4. Hypothesis testing (one sample)

1. What is inferential statistics?
2. **Hypothesis testing for the mean**
3. **Confidence Intervals for the mean**

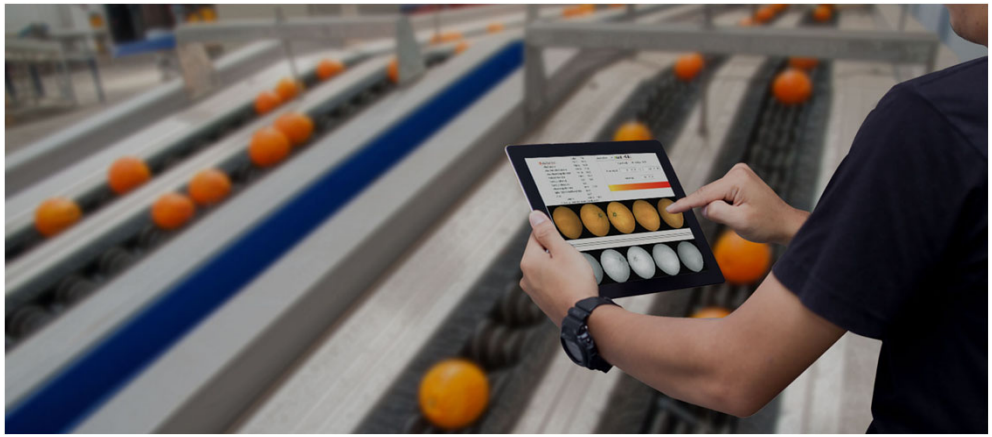
An example

4. **Type I error and Type II error**
5. **Two in-class exercises** 

An example: In-line weight and fruit packaging software




The machine that fills oranges in bags is controlled by a software.
Using this software, the machine is set to obtain a total weight per bag of 2000 g.



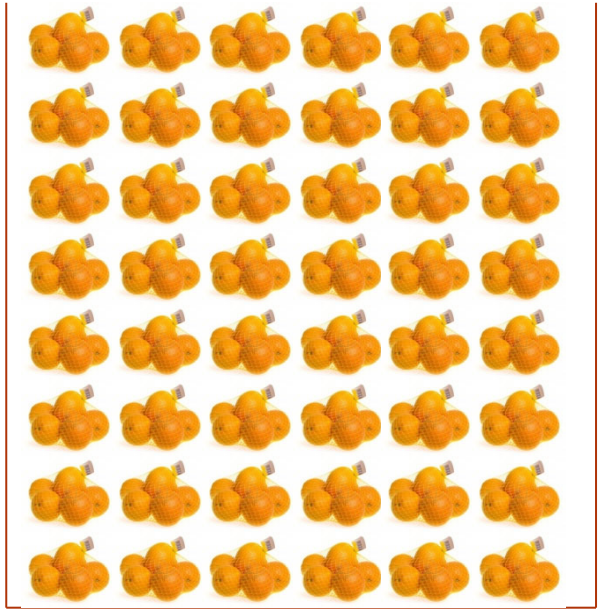
- Due to different sources of variation, it is imposible to always get bags whose weight is exactly 2000 g.
- The obtained weight is a random variable.
- It is considered that the machine is working properly, if the mean of that random variable is 2000 g.

- The Quality Control Department suspects that the mean weight of the bags is NOT 2000 g (that is, suspects that the machine is not working properly and needs to be recalibrated)…

Is that suspicion true?

- To answer this question, we take 15 bags (randomly), and we weight them.

Population: All bags



Mean = μ
Standard deviation = σ

Sampling

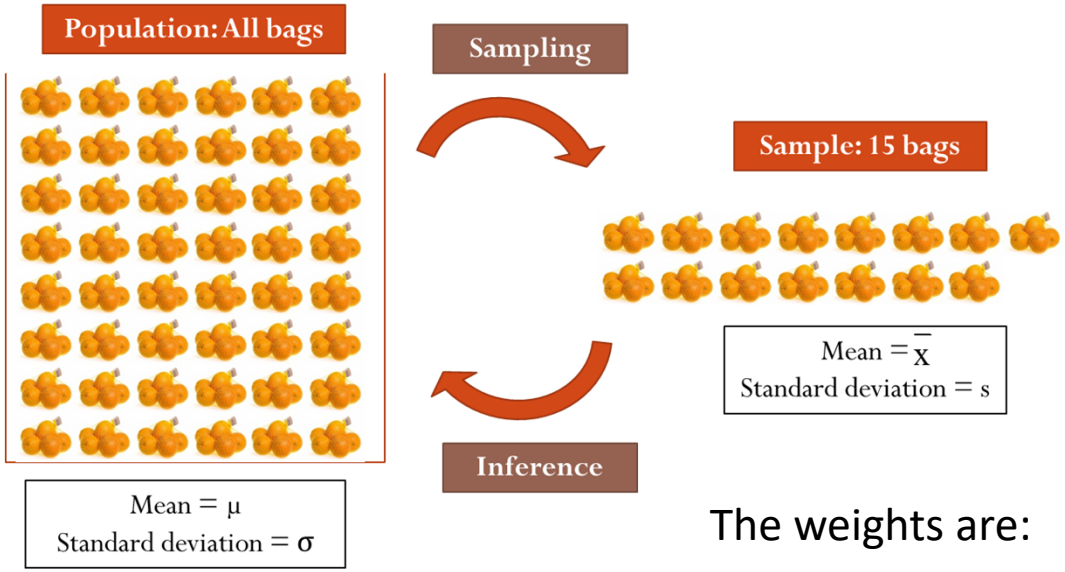
Inference

Sample: 15 bags

Mean = $\bar{x}$
Standard deviation = s

The weights are:

1989	2015	1962	2013	1983
1989	1992	2011	1958	2023
1980	1977	1994	2017	2001



The weights are:

1989	2015	1962	2013	1983
1989	1992	2011	1958	2023
1980	1977	1994	2017	2001

Calculate the mean ($\bar{x}$) and standard deviation (s) <u>of the sample</u>

$\bar{x} = 1993.6$ $\bar{x}$ is an estimator of the population mean (μ)

$s = 19.8$ s is an estimator of the population standard deviation (σ)

So, if the mean weight of those 15 bags (sample) is 1993.6 g ….

I. **Is there evidence that the population mean (μ) is NOT 2000 g, and so the machine is not working properly, and needs to be recalibrated?**

II. **What is the population mean (μ)?**

Lesson 4. Hypothesis testing (one sample)

1. What is inferential statistics?
2. **Hypothesis testing for the mean** ← I. Is there evidence that $\mu \neq 2000$ g, and so the machine is not working properly, and needs to be recalibrated?
3. **Confidence Intervals for the mean** ← II. What is the population mean μ?
4. **Type I error and Type II error**
5. **Two in-class exercises** 

An example

Hypothesis testing for the mean (of one population): 5 Steps

Step 1: Formulate the null hypothesis and the alternative hypothesis

Step 2: Select the statistical test

Step 3: Calculate the test statistic

Step 4: Find the critical value(s) of the Student-t distribution

Step 5: Make the decision to reject or accept the null hypothesis

Step 1:
Formulate the null hypothesis and the alternative hypothesis

In hypothesis testing (comparison of mean in one sample)…

The **null hypothesis (H0)** is a statistical hypothesis that states that there is no statistical difference between the population mean and a specific value.

The **alternative hypothesis (H1)**, is a statistical hypothesis that states that there is a statistical difference between the population mean and a specific value.

Null hypothesis: μ=2000

("the average weight of the bags filled by the machine is not different from 2000")

Alternative hypothesis: μ ≠2000

("the average weight of the bags filled by the machine is different from 2000")

Step 1:
Formulate the null hypothesis and the alternative hypothesis

Null hypothesis: $\mu = 2000$

("the average weight of the bags filled by the machine is not different from 2000")

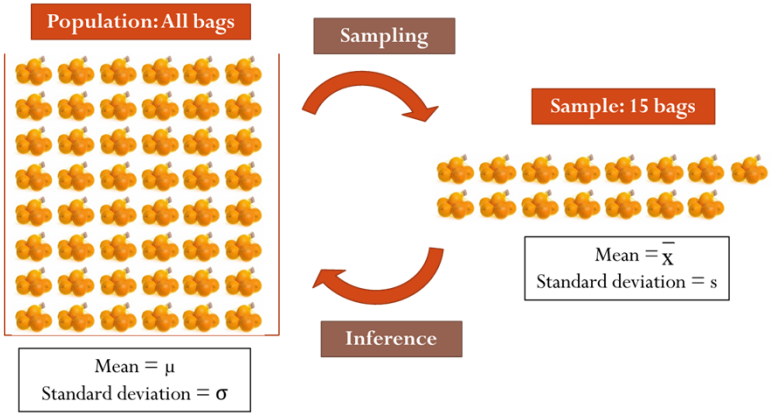
Alternative hypothesis: $\mu \neq 2000$

("the average weight of the bags filled by the machine is different from 2000")

We will perform a **statistical test, that uses the data obtained from a sample, to make a decision about** whether the null hypothesis should be rejected or accepted.

If the test shows that the null hypothesis should be rejected, then we accept the alternative hypothesis.

Step 2: Select the statistical test



Given that: $\bar{x} = 1993.6$ and so, $\bar{x} \neq 2000$

Can we conclude that $\mu \neq 2000$ (the machine is not working properly)?

Not necessarly!!

It seems reasonable to think that:

- if the sample mean ($\bar{x}$) is "quite close" to 2000 (or if: $\bar{x} - 2000$ is "quite close" to 0), then we can assume that $\mu = 2000$ (accept the null hypothesis)

- if the sample mean ($\bar{x}$) is "not that close" to 2000 (or if $\bar{x} - 2000$ is "not that close" to 0), then we can assume that $\mu \neq 2000$ (reject the null hypothesis)

Step 2: Select the statistical test

Select the statistical test: Comparison of mean (one sample Student t-test)

If the null hypothesis (H0) is true (if μ=2000), then,

the test statistic: $\frac{\overline{x} - \mu}{s/\sqrt{N}}$ "follows" a Student t distribution (t_{N-1}).

William Sealy Gosset (aka *Student*) in 1908

Student's t-distribution is a continuous probability distribution that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.

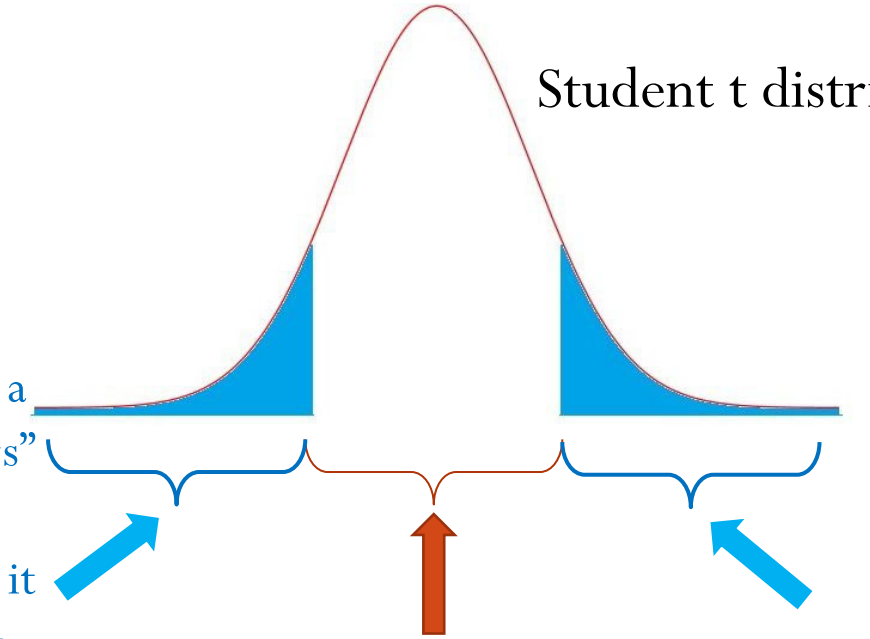
$$\frac{\overline{x} - \mu}{s/\sqrt{N}} \approx t_{N-1}$$

To make a decision, we compare the test statistic, with the Student-t probability distribution.

Step 2: Select the statistical test

What does "follows a Student t distribution" mean?

We will first calculate the test statistic: $\frac{\overline{x} - \mu}{s/\sqrt{N}}$ → We obtain a number



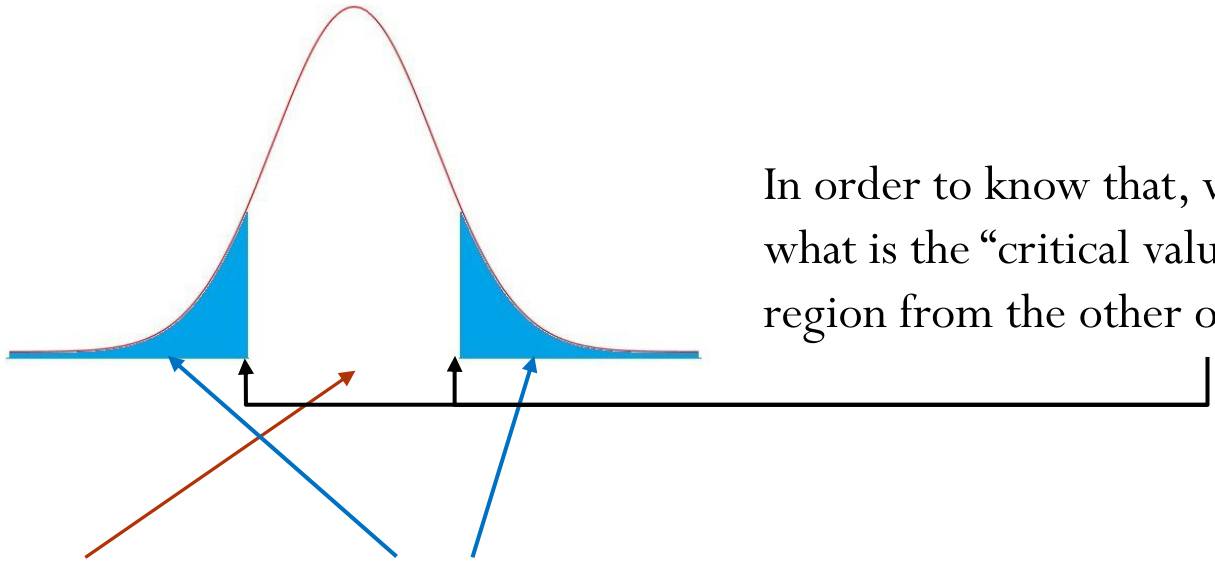
If the calculated number is a number that "rarely appears" in a Student t distribution (t_{N-1}), then we can say that it does not follow a Student t distribution, and then we REJECT the Null hypothesis ($\mu \neq 2000$)

If the calculated number is a number that "appears very often" in a Student t distribution (t_{N-1}), then we can say that it follows a Student t distribution, and then we will ACCEPT the Null hypothesis ($\mu = 2000$)

Step 3: Calculate the test statistic

Let's calculate the test statistic:

$$\frac{\overline{x} - \mu}{s/\sqrt{N}} = \frac{1993.6 - 2000}{19.8/\sqrt{15}} = -1.25$$



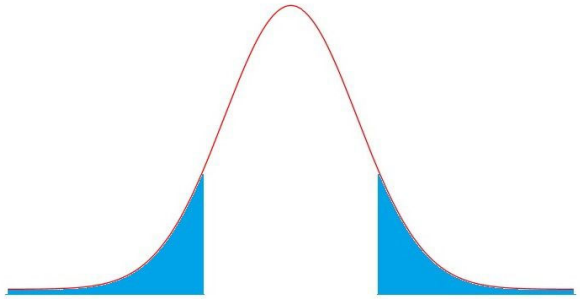
In order to know that, we need to find out what is the "critical value" that separates one region from the other one.

Is the number -1.25 here? (and then we should accept the null hypothesis)

Or is it here? (and then we should reject the null hypothesis)

Step 4: Find the critical value(s) of the Student-t distribution

We need to find the "critical value" that separates one region from the other one in the Student-t probability distribution

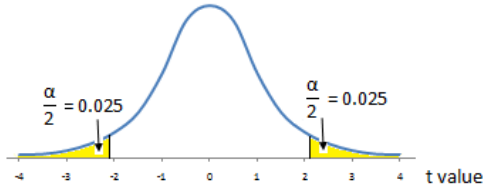


Most statistics books have look-up tables for the distributions.

You can also find online calculators: http://www.ttable.org/student-t-value-calculator.html#

Student's t Distribution Table

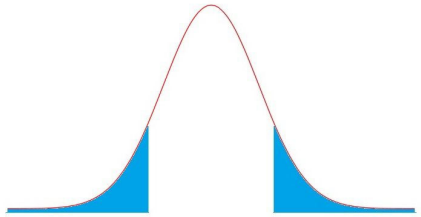
For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	**1**-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	**2**-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	**1**-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	**2**-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	

Step 4: Find the critical value(s) of the Student-t distribution

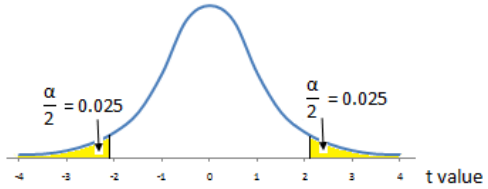
In order to find the "critical value" in the table, we need to follow 4 steps:



1. Decide on which level of significance we want to use
2. Find out whether our test is a "two-tailed test" or a "one-tailed test"
3. Find the "degrees of freedom"
4. Find the critical value of the Student t distribution

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	**1**-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	**2**-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	**1**-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	**2**-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	

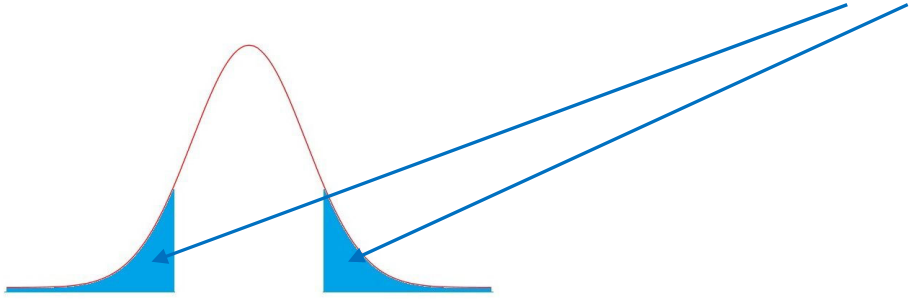
1. We have to select the **level of significance**.

The **level of significance** is the maximum probability of rejecting the null hypothesis when it is true.

We have to decide on "the risk we are willing to take" for declaring a difference, when there is not difference.

For example, if we use a significance level of 0.05 in our example, that means that, when we carry out the statistical test, there is a 5% risk of concluding that there is a difference ($\mu \neq 2000$) when there is NO actual difference.
That is, there is a 5% risk of concluding that the machine is not working properly when in reality, it is working properly.

From the graphical point of view, the level of significance is the size of this area:



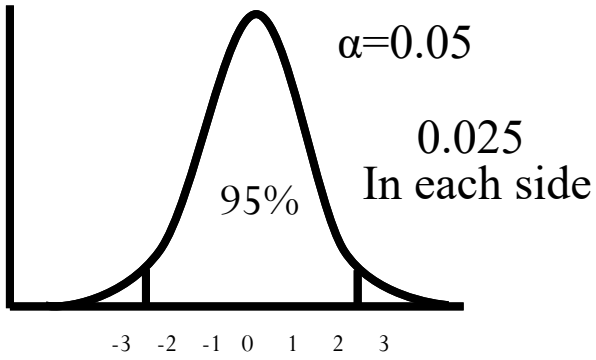
1. We have to select the **level of significance**.

Statisticians generally agree on using one of these significance levels (α): the 0.10, 0.05, and 0.01 levels.

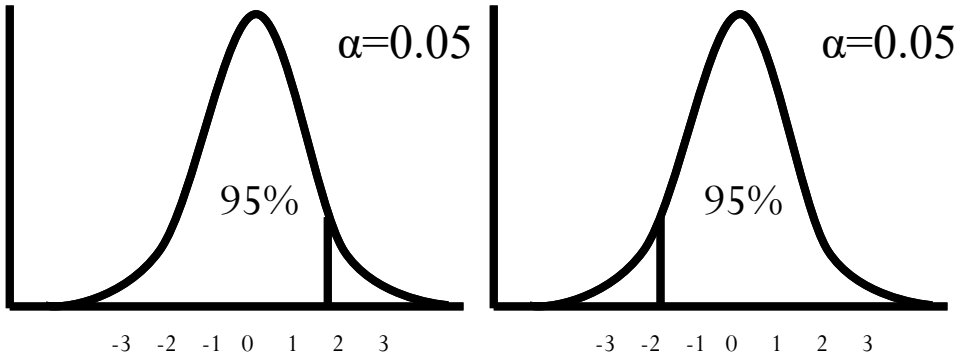
Level of significance	What does it mean?	What does it mean in our example?	Confidence level
$\alpha = 0.10$	There is a 10% probability of rejecting a true null hypothesis.	"We decide that we are willing to take a 10% risk of saying that the unknown population mean (μ) is different from 2000, when in fact it is not different from 2000"	90%
$\alpha = 0.05$	There is a 5% probability of rejecting a true null hypothesis.	"We decide that we are willing to take a 5% risk of saying that the unknown population mean (μ) is different from 2000, when in fact it is not different from 2000"	95%
$\alpha = 0.01$	There is a 1% probability of rejecting a true null hypothesis.	"We decide that we are willing to take a 1% risk of saying that the unknown population mean (μ) is different from 2000, when in fact it is not different from 2000"	99%

2. Find out if the test is **“two-tailed” or “one-tailed”**

Two-tailed test



One-tailed tests



Most tests we will perform, including this one (Hypothesis testing for the mean) will be “two-tailed” tests.

3. Find the **degrees of freedom**

We also need to know the degrees of freedom to find this "critical value" in the table (the degrees of freedom are based on the sample size: d.f. = N-1).

$$\text{d.f.} = \text{N-1} = 15 - 1 = 14$$

4. Find the **critical value** of the Student t distribution

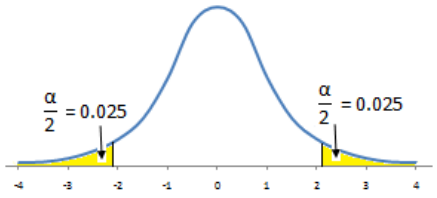
Once we know:

- the level of significance (α=0.05) (that is, confidence level = 95%)
- our test is a two-tailed test
- degrees of freedom (d.f. = 14)

we can find the critical value from the Student t distribution.

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	

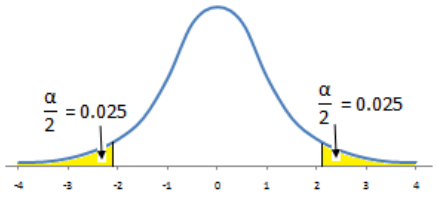
4. Find the **critical value** of the Student t distribution

Let's obtain the critical value of the t-Student distribution with degrees of freedom N-1 (15-1=14), in a two-tailed test, and α=0.05 (confidence level 95%).

$$t_{\text{table N}-1}(95\%) = t_{14}(95\%) = 2.14$$

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	

Step 4: Find the critical value(s) of the Student-t distribution

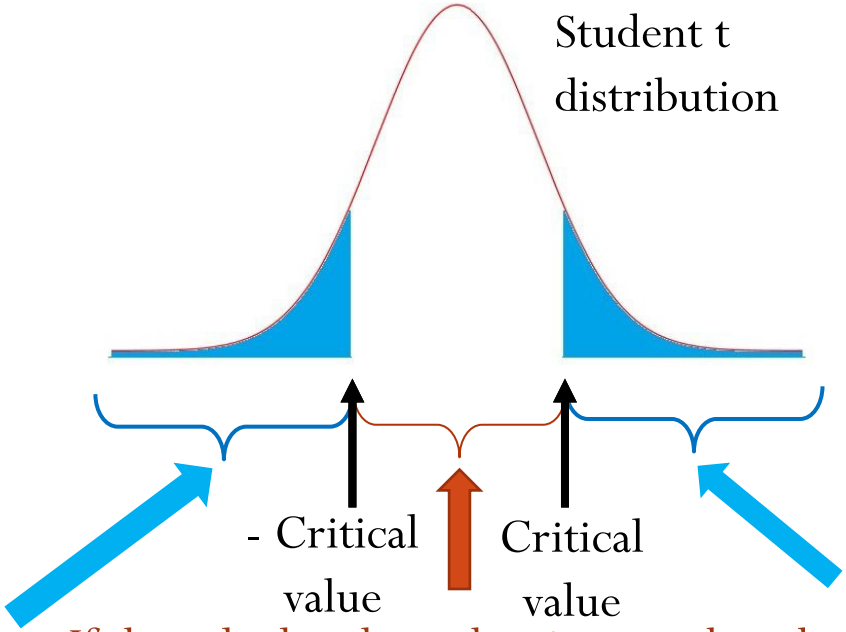
Let's go back to the statistical test (Comparison of mean, one sample t-test):

If the null hypothesis (H0) is true (if $\mu=2000$), then:

$$\frac{\overline{x}-\mu}{s/\sqrt{N}} \approx t_{N-1}$$

Test statistic = -1.25
(also called $t_{calculated}$)

Critical value = 2.14
(also called t_{table})



If the calculated number is a number that "rarely appears" in a Student t distribution, then we can say that it does not follow a Student t distribution, and then we REJECT the Null hypothesis

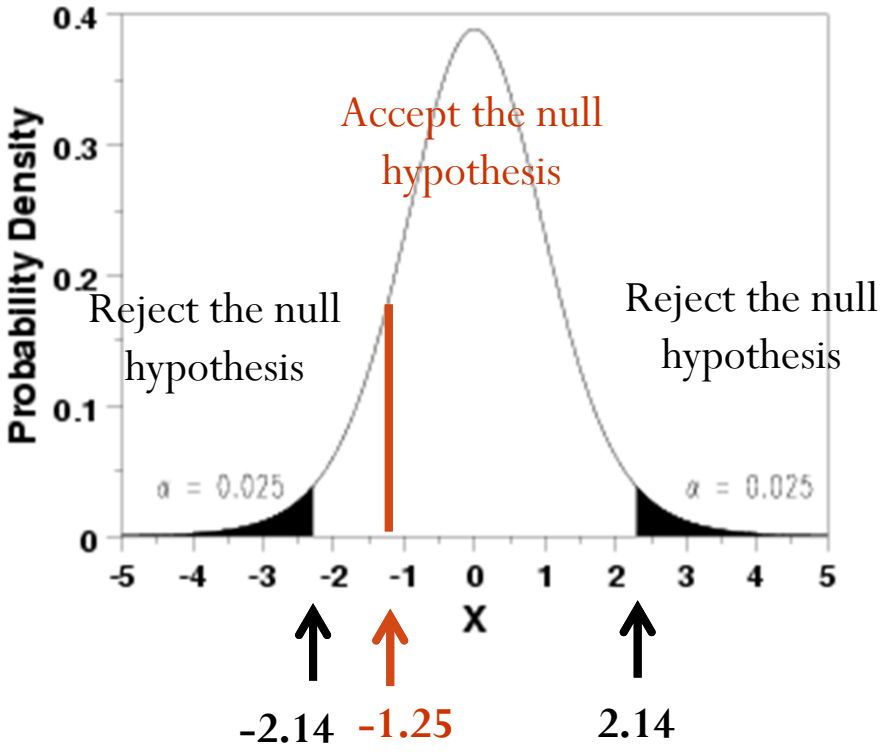
If the calculated number is a number that "appears very often" in a Student t distribution, then we can say that it follows a Student t distribution, and then we will ACCEPT the Null hypothesis

Step 5: Make the decision to reject or accept the null hypothesis

Decision:

Since: $-t_{table} < \frac{\bar{x}-\mu}{s/\sqrt{N}} < t_{table}$

$-2.14 < -1.25 < 2.14$



We accept the null hypothesis ($\mu=2000$),

and so:

We can conclude that the mean weight of the population (μ) is <u>not significantly different</u> from 2000, and that the machine that fills the bags is working properly.

Lesson 4. Hypothesis testing (one sample)

1. **What is inferential statistics?**

An example

2. **Hypothesis testing for the mean** ← I. Is there evidence that $\mu \neq 2000$ g, and so the machine is not working properly, and needs to be recalibrated?

   Comparison of mean (one sample)

3. **Confidence Intervals for the mean** ← II. What is the population mean μ?

4. **Type I error and Type II error**

5. **Two in-class exercises** 

II. What is the population mean (μ)?

- We cannot know what the exact population mean is, unless we weight ALL the bags that the machine produces.

- BUT, by knowing the weight of the bags that we sampled (15 bags), we will be able to say that, with some probability, the population mean is within a certain range.

- This range, is called the "confidence interval", and the probability is called the "confidence level".

II. What is the population mean (μ)?

Confidence interval (CI) of μ:
The range of values that we can have good confidence, that it contains the population mean μ.

Confidence level (CL) of the confidence interval of μ:
The probability that the confidence interval will contain the population mean μ.

II. What is the population mean (μ)?

- Let's first select the confidence level, usually 95% or 99%.

  We select a confidence level of 95%, which means that we estimate with 95% confidence that the true value of the population mean (μ) is within the range defined by the confidence interval.

- Let's now calculate the confidence interval.

$$\bar{x} - t_{N-1}\frac{s}{\sqrt{N}} < \mu < \bar{x} + t_{N-1}\frac{s}{\sqrt{N}}$$

s is the standard deviation of the 15 bags (sample)

19.8

$\bar{x}$ is the mean weight of the 15 bags (sample)

1993.6

N is the sample size

15

t_{N-1} is the critical value obtained from a Student t distribution with N-1 degrees of freedom and confidence level of 95% (α=0.05)

II. What is the population mean (μ)?

- Let's now calculate the confidence interval.

$$\bar{x} - t_{N-1}\frac{s}{\sqrt{N}} < \mu < \bar{x} + t_{N-1}\frac{s}{\sqrt{N}}$$

$$1993.6 - 2.14\frac{19.8}{\sqrt{15}} < \mu < 1993.6 + 2.14\frac{19.8}{\sqrt{15}}$$

$$1982 < \mu < 2004$$

What is the population mean?

"We estimate with 95% confidence that the true value of the population mean (μ) is between 1982 g and 2004 g"

II. What is the population mean (μ)?

An additional reflection about the Confidence interval of the mean:

At the beginning of this exercise, we wanted to find out whether the population mean (μ) was different from 2000 or not.

Another way to answer that, is by looking at the confidence interval:

- When the confidence interval contains the hypothesized mean, we ACCEPT the null hypothesis: $\mu=2000$
- When the confidence interval does not contain the hypothesized mean, REJECT the null hypothesis: $\mu\neq2000$

In our case, the hypothesized mean (2000) is contained in the range of the confidence interval (that is, "2000 is in between 1982 and 2004"), therefore, we ACCEPT the null hypothesis ($\mu=2000$),

as we already did! 

Lesson 4. Hypothesis testing (one sample)

1. **What is inferential statistics?**

2. **Hypothesis testing for the mean**
   Comparison of mean (one sample)

3. **Confidence Intervals for the mean**

4. **Type I error and Type II error**

5. **Two in-class exercises** 

An example

I. Is there evidence that $\mu \neq 2000$ g, and so the machine is not working properly, and needs to be recalibrated?

II. What is the population mean μ?

Error Type I and error Type II

Statisticians generally agree on using one of these three arbitrary significance levels:

$\alpha = 0.10$
$\alpha = 0.05$
$\alpha = 0.01$

α (significance level) is the probability of committing a type I error

A **type I error** occurs if one rejects the null hypothesis when it is true.

β is the probability of committing a type II error.

A **type II error** occurs if one accepts the null hypothesis when it is false.

α and β are related in that way that decreasing one increases the other.

Lesson 4. Hypothesis testing (one sample)

1. **What is inferential statistics?**

2. **Hypothesis testing for the mean**
   Comparison of mean (one sample)

3. **Confidence Intervals for the mean**

4. **Type I error and Type II error**

5. **Two in-class exercises** 

An example

I. Is there evidence that $\mu \neq 2000$ g, and so the machine is not working properly, and needs to be recalibrated?

II. What is the population mean μ?

In-class exercises

We want to carry out a one sample hypothesis test to compare the unknown population mean, with a certain value (this value is equal to 99). We decide to use a level of significance of 0.01. We obtain a statistically significant result. That means:

A. There is a 1% probability of rejecting a true null hypothesis.
B. There is a 1% probability of accepting a false null hypothesis.
C. We have decided that we are willing to take a 1% risk of saying that the unknown population mean is not different from 99, when in fact it is different from 99.
D. The confidence interval of our estimation is 99±0.01.

Question 6

We want to carry out a one sample hypothesis test to compare the unknown population mean, with a certain value (this value is equal to 99). We decide to use a level of significance of 0.01. We obtain a statistically significant result. That means:

(2 points)

Your answer:

- ○ There is a 1% probability of rejecting a true null hypothesis.
- ○ There is a 1% probability of accepting a false null hypothesis.
- ○ We have decided that we are willing to take a 1% risk of saying that the unknown population mean is not different from 99, when in fact it is different from 99.
- ○ The confidence interval of our estimation is 99±0.01.

Number of Times checking Email Every Day

According to the Harvard Business Review (in the article: "How to Spend Way Less Time on Email Every Day"), the average professional checks his/her emails **15 times per day**.

The data represent a sample of the number of times/year, that 7 employees in a company, check their emails:

5460 5900 6090 6310 7160 8440 9930

Which one of these statements is correct?

A. We can be 99% confident that the mean number of times that the employees of this company check their email each year is between 4785 and 9298.

B. We can be 99% confident that the mean number of times that the employees of this company check their email is not significantly different from that of the "average professional".

C. None of the previous responses is correct.

D. A and B are correct.