# Contents

Stat 252

Introduction to Applied Statistics II

Dr. Karen Buro

webpage: academic.macewan.ca/burok/

# 1 Introduction

Organization:

- Course outline
- Contact through email
- Homework Assignments
- Lab (SPSS)
- Review (also use notes from Stat 151/161, or text old book)
  - Population versus Sample
  - Inferential Statistics
  - Parameter versus Estimate versus Estimator
  - Confidence Intervals
  - Statistical Tests
  - t-tests and t-confidence intervals
- Subjects covered in this course:
  - t-tests versus nonparametric tools
  - 1-way ANOVA versus nonparametric tool
  - 2-way ANOVA
  - Simple and Multiple Linear Regression: Model building

# 2 Review

In this section, material from the prerequisite courses will be reviewed, while at the same time adding depth to the material.
At the end of this section a full understanding of the presented material is expected, it represents the foundation on which everything that follows is built.
Attend office hours, email questions, if you are not fully understanding.

## 2.1 Random Variables and Distributions

This subsection revisits some foundational concepts from probability theory to prepare for a discussion of statistics.
In probability theory and statistics, a random variable is described informally as a variable whose values depend on outcomes of a random phenomenon, like rolling a die, or the age of a randomly chosen person, the colour of a candy randomly chosen from a bag.
A random variable's possible values represent the possible outcomes of the random event. The meaning of the probabilities assigned to the potential values of a random variable are usually considered to be the proportion of events resulting in the value considered.
Random variables are described in terms of their probability distribution.

**Definition 1**
A *probability distribution* describes a random variable by:

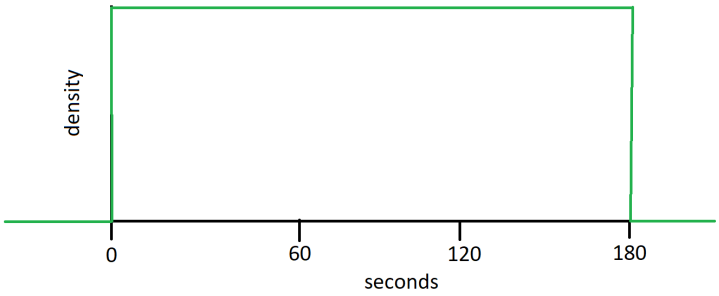
1. indicating the possible values the random variable can assume
2. how likely are those values of the random variable

**Example 1**
(discrete distribution) After rolling a fair six sided die the number showing is random, and can be described as a random variable.
The random variable: $X=$ number showing after rolling a fair six sided die
The distribution of $X$ (see definition above)

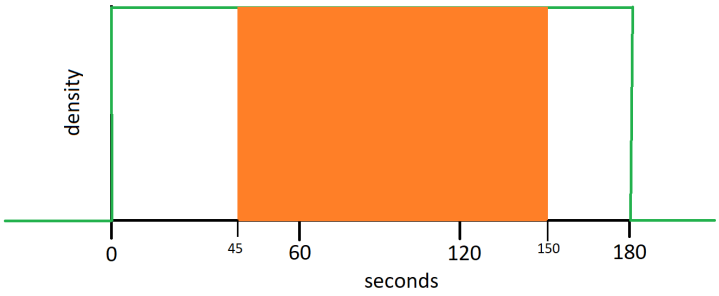
1. possible values $S = \{1, 2, 3, 4, 5, 6\}$
2. probabilities for those values

| possible values $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| their probabilities $p(x)$ | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

**Example 2**
(continuous distribution) When entering building 5 on MacEwan campus it is unknown where the elevator is, and one has to wait between 0 seconds and 3 minutes for it to arrive at the first level. Every wait time between 0 and 180 seconds is equally likely. In this case wait time can be modelled as a random variable.
The random variable: $X =$ wait time for elevator
The distribution of $X$: This is a continuous random variable since every time between 0 and 180 seconds is possible, and it is not possible any more to give a table of all possible outcomes as in the previous example. For continuous random variables density functions are used to characterize the distribution.




In this case: below 0 seconds and above 180 seconds the density equals 0, indicating that such values are impossible for the random variable. The possible values for this random variable fall between 0 and 180 seconds.
Between 0 and 180 seconds the density has the same (uniform) value indicating that every value has the same chance of occurring.
The probability to fall with in a certain interval is given through the area under the curve.
E.g. $P(45 < X < 150) =$ orange area in the diagram.




**Take away:**

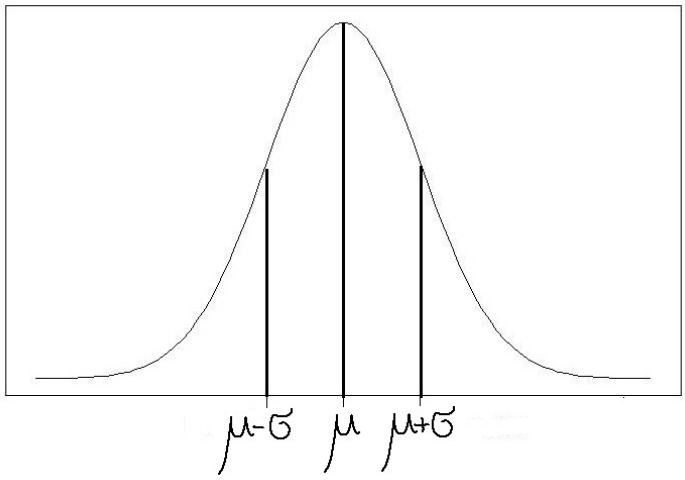
- Random variables are described by their probability distribution, which indicate their possible values and the probabilities for these to occur.
- For discrete random variables a table with this information can be provided, but for continuous random variables the distribution is given by their density.

## 2.2 Normal Distribution and Standardized Scores

The normal distribution is one example of a continuous distribution.

**Definition 2**

1. A random variable, $X$, is normally distributed, if its distribution has a bell shaped density function, and is characterized by its mean $\mu$ and standard deviation $\sigma$ $(> 0)$. Write $X \sim \mathcal{N}(\mu, \sigma)$.




2. A random variable, $Z$, is called standard normally distributed, if it is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$.

   Write $Z \sim \mathcal{N}(0, 1)$.

The z-table posted on the website provides probabilities for the standard normal distribution. The table will serve in this course for finding probabilities and critical values for the normal distribution. To use the table for any normal distribution a normally distributed random variable $X$ has to be standardized.

**Remember**:
Let $X \sim \mathcal{N}(\mu, \sigma)$, then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

By subtracting $\mu$ from $X$ the distribution is moved into 0, by dividing by $\sigma$ the width (spread) is standardized to standard deviation 1.

**Example 3**
(Example calculating normal probability and percentile)
It can be assumed that the birth weight of babies born in Canada is normally distributed with (approximately) mean $\mu = 3600g$ and standard deviation $\sigma = 500g$.

1. What is the probability that a baby has birth weight below $3000g$?

$$\begin{aligned} P(X < 3000) &= P(Z < \tfrac{3000-3600}{500}) \quad \text{(standardize)} \\ &= P(Z < -1.2) \quad \text{(use table)} \\ &= 0.1151 \end{aligned}$$

   The probability for a newborn in Canada to weigh less than 3000 g is 0.1151, or between 11 and 12% of newborns in Canada weigh less than 3000g.

2. How heavy are the heaviest 25% of newborns in Canada?

This is asking for the 75% percentile (25% to the top implies 75% to the bottom) denoted by $x_{.75}$.

$$\begin{aligned} P(X < x_{.75}) &= .75 \quad \text{(standardize)} \\ P(Z < \tfrac{x_{.75}-3600}{500}) &= 0.75 \quad \text{(use table)} \end{aligned}$$

$$\begin{aligned} \tfrac{x_{.75}-3600}{500} &= 0.67 \quad \text{(rearrange)} \\ x_{.75} &= 500 \times 0.67 + 3600 \\ &= 3935 \end{aligned}$$

The heaviest 25% of newborns weigh $3935g$ or more. Or the probability that a newborn in Canada weighs more than $3935g$ is 0.25.

Students are assumed to know how to make these kind of calculations as prerequisite coming into this course - review if necessary.

## 2.3 Inferential versus Descriptive Statistics

Statistics is the art and science of learning from data

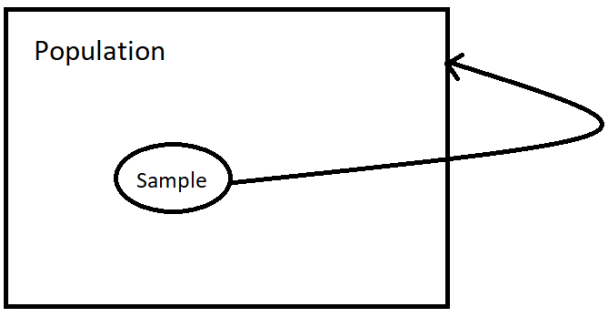
The key to the definition is "data".
Two different branches of statistics are considered.

**Descriptive Statistics**
Descriptive Statistics discusses how to summarize and describe a data set. A data set usually includes several variables an a number of individuals/participants.

**Example 4**
Consider a data set describing all animals housed in an animal shelter in Edmonton.
Potential variables: species, sex, age, weight, colour, injuries (yes/no), ready to be adopted (y/n), preferred food, etc..
Individuals/participants: all animals currently in the shelter

To describe the data set one might give the mean and standard deviation of age for each species, the percent of male and female animals of each species, percent of healthy animals, etc.

**Inferential Statistics**
The purpose of Inferential Statistics is the study of methods that allow the use of sample data to draw conclusions about entire populations.



**Definition 3**

- A population is the collection of all elements – individuals, objects – which are of interest and information about are of interest.
- A sample is a subset of the population.

The focus of this course is on the introduction of a variety of important and often applied tools from *Inferential Statistics*.

**Example 5**
The interaction of a sample of dogs and their owners are observed before and after a workshop to learn about the general effect of the workshop on obedience in dependency of age, sex and race of the dog, and personal traits of the owners.
The purpose is not so much to learn about the specific sample but to learn about the effect of the workshop in general.

## 2.4 Models in Inferential Statistics

Often in inferential statistics one has to start with a model that is supposed to describe the population of interest. The model usually depends on different parameters. E.g. one might assume that blood pressure of students after an exam is normally distributed with unknown mean and standard deviation. The statistical analysis will be based on the assumption that the model is an appropriate description of reality (population). If this assumption is violated the results become meaningless because they were obtained in the context of the model, but if the model is not describing the population they become irrelevant. To argue the same conclusion from a technical perspective, if the model is not an appropriate description of the population then the distribution of the standardized scores might be different from the one assumed to obtain P-values and critical values. In consequence the decisions hinging on these values might not hold.
This is the reason why it is so important to check assumptions for all statistical methods applied to sample data with the purpose of learning about an entire population. For each method introduced in this course the assumptions will be discussed and how they can be checked using descriptive tools.

If a model is found to be a good fit for a population, statistical tools arising from the model allow the user to estimate and answer questions concerning the parameters of the model based on sample data and deduce knowledge about the population.
This is a powerful approach because is permits to draw conclusions about entire populations by only looking at a sample from the population.

## 2.5 Parameter versus Estimator versus Estimate

- A **parameter** describes a feature of a *population*. (e.g. $\mu$, the population mean)

  In statistics one is often interested in the value of a parameter. Parameter values can only be *calculated* based on a census (a sample that includes the entire population) information. Since conducting a census is often unrealistic or not even conceivable the only information about parameters can be obtained from samples, which are relatively small in comparison to the entire population and can be used to *estimate* the value of the parameter of interest.

- The **estimator** for a parameter is a rule (*formula*) that tells how to use sample data to make an informed guess about the value of the parameter. (e.g. $\bar{X}$, the sample mean to estimate the population mean, $\mu$)

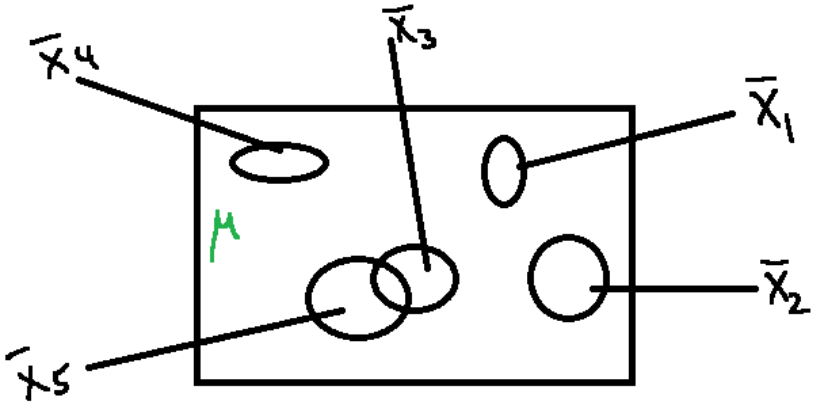
  Inferential Statistics teaches how to use sample data to learn about entire populations, therefore an estimator is a tool from Inferential Statistics.

- The parameter **estimate** is the value of the estimator *for a specific sample*. (the mean age of nurses in a sample from the U of A hospital is 39.2 years)

  A parameter estimate is obtained by applying an estimator to sample data. It has to be interpreted and filled with meaning.

## 2.6 Point estimation of model parameters

**Definition 4**
An estimator is called *unbiased* for a parameter if the mean of the estimator taken over all possible samples equals the parameter of interest.




Parameter $\mu$ is the population mean (the mean of all individuals in the population).
For every sample one might get a different sample mean. ($\bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_3 \neq \ldots$)
The sample mean is an unbiased estimator for the population mean because we can prove mathematically that the average of the sample means from all imaginable samples for a fixed sample size equals $\mu$.
**Examples:**

- The sample mean $\bar{X}$ is an *unbiased estimator* of the population mean $\mu$ based on a sample $x_1, x_2, \ldots, x_n$. ($\mu_{\bar{x}} = \mu$)
- If the population is skewed the sample median $\tilde{X}$ is NOT an unbiased *estimator* of the population mean $\mu$ based on a sample $x_1, x_2, \ldots, x_n$. ($\mu_{\tilde{X}} \neq \mu$)
- An unbiased *estimator* of the population variance $\sigma^2$ is the sample variance

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

  If one would divide by $n$ instead of $n-1$ it would make the estimator unbiased.

**Take away:** Unbiased = the mean of an estimator equals the parameter it is supposed to estimate (where the mean is taken over all imaginable samples of a given sample size)

This is the best we can hope for, because an estimator is based on sample data and will have different values depending on the sample chosen.

One big **shortcoming** of point estimators is that they will always be wrong. Point estimators will not be able to reproduce the exact value of $\mu$.
But point estimators should result in values which are "close" to the true value of the parameter of interest. That raises the question what "close" means. In order to describe how close one can expect $\mu$ to be to the sample mean of a sample, confidence intervals are given to estimate $\mu$. More on this in the next section.

When assuming that a sample has been randomly chosen from the target population, the value of the estimator is a random outcome, which makes the estimator a random variable.

The randomness of the estimator is caused by randomly choosing a sample.

### 2.6.1 The Sampling Distribution of $\bar{X}$

The sample mean $\bar{X}$ (using capital letter here to indicate that it is a random variable) is one of the most common estimators used in inferential statistics and it is relevant to ask, how good are the estimates produced by it, or how close can one expect the sample mean $\bar{X}$ to be to the true population mean $\mu$.

To reiterate: If one were to take a random sample of size $n$ from a population then the value of $\bar{X}$ is random (because the sampling is random), therefore $\bar{X}$ is a random variable and can be described in terms of its probability distribution.

The Central Limit Theorem, maybe the most important theorem in statistics, describes the distribution of $\bar{X}$.
It provides statisticians with the mean and the standard deviation of the distribution of $\bar{X}$ relative to the mean and standard deviation of the target population, and most importantly states the (approximate) shape of the distribution if the sample size, $n$, is large enough or the target population itself is normal.

**Central Limit Theorem**
The distribution of $\bar{X}$, the sample mean of $n$ measurements randomly drawn from a population with mean $\mu$ and standard deviation $\sigma$, has the following properties

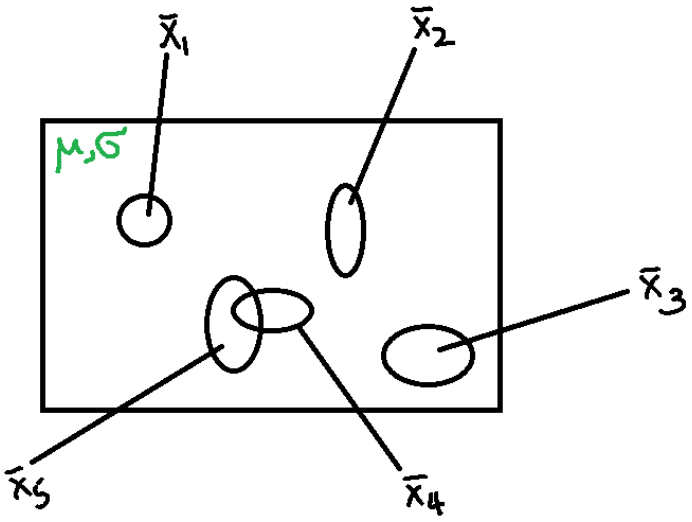
1. $\mu_{\bar{X}} = \mu$.
2. $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.
3. If the population is normalm $\bar{X}$ is normally distributed.
4. (almost) Regardless of the distribution of the population, if the sample size $n$ is *large*, $\bar{X}$ is approximately normally distributed.

This Central Limit Theorem implies:

1. The mean of all sample means equals the population mean, or **on average** $\bar{X}$ equals $\mu$.
2. The standard deviation of the sample means is smaller than the standard deviation in the population.
3. Imagine taking a sample from a population and calculating the sample mean, then do this again, draw another sample and calculate the sample mean, giving you a new sample mean, imagine doing this let's say 50 times. Now imagine drawing a histogram for those 50 sample means. The Central Limit Theorem states, that regardless of the distribution of the population the histogram will always look approximately normal if the sample means are based on *large* samples.

Consequences from the Central Limit Theorem

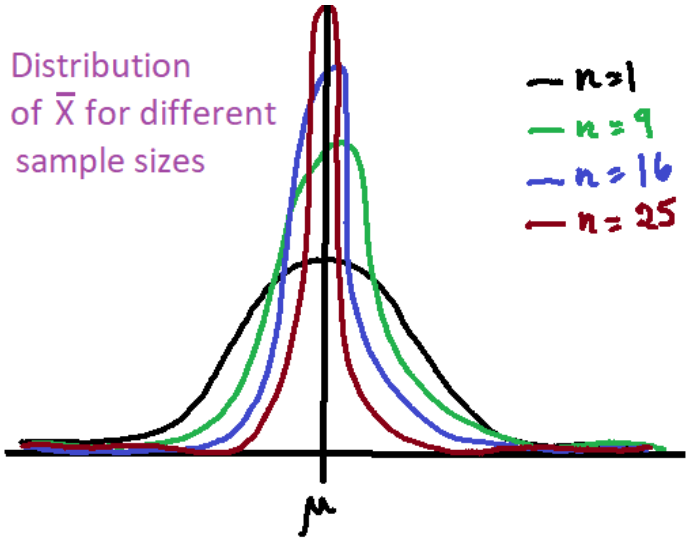
- The first result says that $\bar{X}$ is an unbiased estimator for $\mu$, i.e. the average of the sample means of all possible samples equals the population mean.




This tells us, that when we use the sample mean of a specific sample, $\bar{x}$, to estimate the unknown population mean, $\mu$, we do not make a systematic error.

- The second result says that the standard deviation of those sample means equals the standard deviation of the population divided by the square root of $n$. Which implies that the larger the sample size the smaller the variation between the sample means. I.e. if we would take really large samples then the sample means are all very close to each other.
- One and two taken together give us the law of large numbers, which tells us that statistics works.

  When using the estimator $\bar{X}$ to estimate the unknown population mean, $\mu$, we do not make a systematic error, and when we increase the sample size the probability for the sample mean to be "close" to $\mu$ is increasing.




- The fourth statement from the CLT means that statistical methods developed for normally distributed populations can be applied to a much wider range of populations, if the sample size is large and the method is based on the sample mean.
- Standardizing the random variable $\bar{X}$ for large $n$ or normal populations is done by (using that $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

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

- The Central Limit Theorem provides the distribution of $\bar{X}$, which means one is able to calculate for example how likely $\bar{X}$ will fall within 2 standard deviations of the mean, $\mu$, i.e.

$$P(\mu - 2\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 2\sigma/\sqrt{n}) \approx 0.95$$

  or equivalent

$$P(\bar{X} - 2\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 2\sigma/\sqrt{n}) \approx 0.95$$

  The last equation means that if we know $\sigma$ (population standard deviation) for approximately 95% of samples from the population the population mean $\mu$ falls into the interval $[\bar{X} - 2\sigma,\ \bar{X} + 2\sigma]$.

  Which means that $[\bar{X} - 2\sigma,\ \bar{X} + 2\sigma]$ is approximately a 95% confidence interval for $\mu$.

  More on confidence intervals in the next section.

## 2.7 Confidence Intervals

- As seen before the sample mean, $\bar{X}$, is a point-estimator for the unknown population mean, $\mu$.

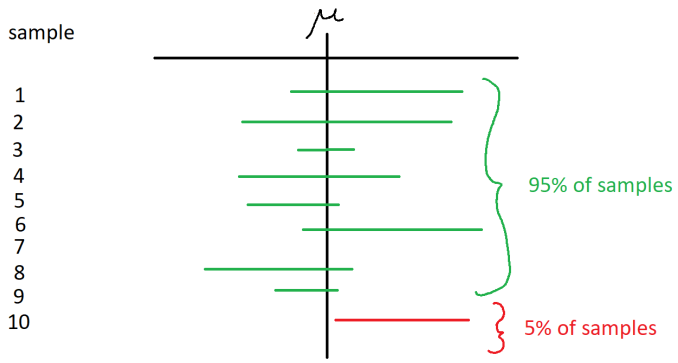
  E.g. assuming a scientist is interested in the amount of apples which can be harvested from a 10-year old apple tree of a certain cultivar.

  To estimate this amount data is collected for 120 10-year old trees of this cultivar and the sample mean of the data is used to *estimate mu* = mean amount of apples which can be harvested from a 10-year old apple tree of a certain cultivar.

  The limitation of point estimators is that they will always be wrong, since they will not be able to capture the exact value of $\mu$. In order to address this shortcoming confidence intervals are introduced for estimating the values of parameters describing populations.

- Confidence intervals provide interval estimates for parameter values, in this case the population mean, by providing an interval that most likely captures the true value.

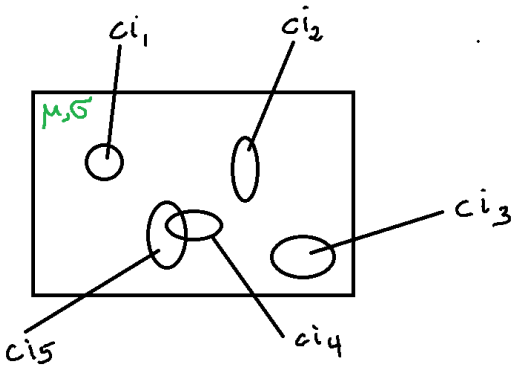
  How likely is indicated by the confidence level of the interval. When calculating one hundred 95% confidence intervals, based on different samples, one should expect that 95 of the intervals capture the true mean, but 5 won't. 95% is the percentage of samples for which a confidence interval captures $\mu$.




  To be more descriptive: Imagine drawing one sample of size $n = 25$, based on this sample a 95% confidence interval for $\mu$ is calculated.

  Then another sample of size 25 is drawn from the population, again a 95% confidence interval for $\mu$ is calculated.

  And then again and again. Imagine repeating this process 100 times, such that by the end 100 confidence intervals have been calculated for 100 different samples of size 25.



We should expect that 95 of the 100 calculated confidence intervals capture the true mean $\mu$.

**Important Concept:** What is the meaning of being e.g. 95% confident?
If such a confidence interval would be calculated for many different independent samples, then 95% of the confidence intervals would truly capture the mean $\mu$ (in this case), the other 5% would not!
Usually we only have ONE sample and get ONE confidence interval for the parameter of interest. We can only hope that the confidence interval we calculated (based on the sample data) is one of those that captures the parameter, we are only 95% confident it does.

**Take away:** The purpose of calculating a confidence interval is to find a range of possible values for the unknown value of the parameter of interest. The confidence interval reveals given the level of confidence how close the estimate is to the true value of the parameter. A true gain over the point estimator.

## 2.8 Statistical Tests

**Review (this section is very important to be fully understood, since almost all tools introduced in this course are statistical tests):**

When *estimating* the scientist tries to obtain information on the values of parameters through sample data. Confidence intervals are used for estimation.

When *testing* in statistics the scientist attempts to answer questions regarding one or more parameters of interest using only the information available from a sample.

E.g.:

1. Is the mean survival timer greater for the new therapy than the mean survival time for the best known therapy yet?
2. Is the proportion of voters supporting candidate A less than the proportion of voters supporting candidate B?
3. Is the mean amount of beer filled in each bottle equal to 500ml?
4. Is sorting algorithm 1 on average faster than sorting algorithm 2?
5. Is the mean inhaled amount of anaesthetic gases by nurses exceeding the value recommended by the government?

How is it possible to make decisions about entire populations by only considering information from a (small) sample?

Magic? No, probability theory!

The downside of making decision on sample information only (not conducting a census) is the possibility of making wrong decisions (errors).
Statistical tests are constructed and are conducted (process) in such a manner so they "control" the probability for making errors.

For statistical tests the question of interest is expressed in two competing hypotheses relating to parameters from the model.
The hypotheses are usually chosen such that they are the logical complement of each other enforcing that one of them has to be true.

$\boldsymbol{H_0}$ is called the null hypothesis

$\boldsymbol{H_a}$ is called the alternative hypothesis. Whenever possible the alternative hypothesis captures what the researcher tries to show, which is due to the way the probabilities for making errors are controlled (see further down).

**The process of deciding between $\boldsymbol{H_0}$ and $\boldsymbol{H_a}$:**

1. Hypotheses: Word the hypotheses: $H_0$ (null hypothesis, H-null) and $H_a$ (alternative hypothesis, Ha), expressing the research question in terms of parameters from the model.

   Choose the significance level to use in this decision (Error probability of type I).

2. Assumptions: Decide which test to use. Check assumptions, if necessary.

3. Test Statistics: Find test statistic – the test statistic measures how "consistent" the sample data is with the statement in $H_0$, usually test statistic values close to zero indicate consistency.

4. P-value: Find p-value – the p-value measures how likely the sample data summarized by the test statistic (or one more extreme) would occur if $H_0$ were true.

5. Decision: Make a decision concerning the hypotheses. (reject $H_0$/do not reject $H_0$) – might want to include the error probability for the decision

6. Conclusion: Answer the question you started of with.

Remember these steps and always apply them when conducting statistical tests.

**Possible errors when conducting statistical tests:**

type I error – the error of rejecting $H_0$ even though $H_0$ is true

type II error – the error of failing to reject $H_0$ even though $H_0$ is false

| | | Truth/Reality/Population | |
|---|---|---|---|
| | | $H_0$ is true | $H_0$ is false |
| Test: | do not reject $H_0$ | OK | type II error |
| | reject $H_0$ | type I error | OK |

If it would be possible we would avoid making any of the errors or at least make both error probabilities really small. Unfortunately the probabilities for the errors of type I and type II behave like being on a seesaw.

**Seesaw effect of probabilities for errors of type I and II:**
Given a population and a fixed sample size: When choosing to reduce the probability for one error, this will in turn increase the other error probability. Push one error probability down will make the other error probability go up, like on a seesaw.
The only way to bring both error probabilities down is by increasing the sample size.

To avoid that both error probabilities are too high to be comfortable with any decision, the science community decided to make sure one error probability is small and accepting to let the other go (uncontrolled).

Statistical tests are constructed in such a way that the probability for an error of type I is controlled, i.e. the experimenter chooses a significance level, $\alpha$, and then the test procedure (see process earlier) assures that for the decision arising out of the process
$P(\text{error of type I}) = P(\text{reject } H_0 \mid H_0 \text{ is true}) \leq \alpha$.
This is at the cost of the probability of an error of type II ($\beta$) being unknown: When a statistical test results in not rejecting $H_0$, this might be an error of type II, but it is unknown how likely such an error occurs, so better not commit to this decision.
**Say**: "At significance level of $\alpha$ the data do not provide sufficient evidence that $H_0$ is wrong."
But when a test leads to the rejection of $H_0$, this might constitute an error of type I, for which it is known that the probability is less than $\alpha$ (an upper bound chosen to represent an acceptable risk), so with some confidence in the decision, the wording is:
"At significance level of $\alpha$ the data provide sufficient evidence that $H_0$ is wrong (or $H_a$ is correct)" (still not 100% certain).

**Catalog of questions the reader should be able to answer now:**

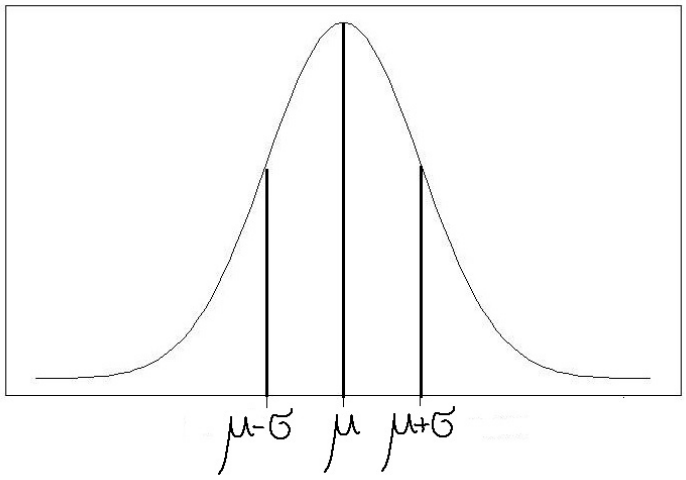
1. What is the purpose of a statistical test?
2. What is reflected by the significance level?
3. How should one choose the hypotheses (which one should include what the researcher wants to confirm)?
4. What is the seesaw effect of the error probabilities
5. How to word conclusions differently depending on if the decision indicated that $H_0$ could be rejected or not.
6. If $H_0$ could be rejected, what is known about the error probability?
7. If $H_0$ could not be rejected, what is known about the error probability?

## 2.9 T-tests and T-confidence intervals

In this section confidence intervals and statistical tests are applied to different models, illustrating how the question guides the choice of model and statistical tool.
The different scenarios considered are, drawing conclusions about one mean, or comparing two means either based on paired or independent samples.
In each case first the model will be introduced and then the methods/tools/formulas the model leads to.

### 2.9.1 One Mean – Single Sample

**The model describing the population:**
The variable of interest follows a normal distribution with (unknown) mean $\mu$ and (unknown) standard deviation $\sigma$.
That is a histogram displaying data from this population should resemble a bell curve




The parameters in the model are the mean $\mu$ and the standard deviation $\sigma$, whose values are unknown, and statistics shall help to find information about.

**Finding a Confidence Interval for the population mean $\mu$:**
From the Central Limit Theorem it can be deduced that under this model for the population (or if $n$ is large) $\bar{X}$ follows a normal distribution with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Therefore the standardized score (Z-score) of the mean

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is standard normally distributed (mean is 0, standard deviation is 1).
Unfortunately this score does not much help in inferential statistics for obtaining information about the unknown mean $\mu$, because it includes also the unknown standard deviation $\sigma$.
Replacing $\sigma$ by its estimator $s = \sqrt{s^2}$ leads to the t-score

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Student showed that the t-score is t-distributed. T-distributions look similar to a standard normal distribution but with thicker tails to accommodate outliers, which occur because of the division by an **estimate** for $\sigma$, which depends on the sample data.
The distribution of the $t$-score depends on the sample size, recognized through the degrees of freedom, $df = n - 1$. The larger the $df$ the closer the t-distribution to the standard normal distribution.

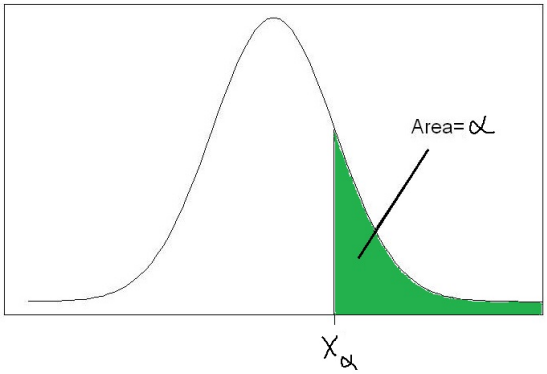
**Calculating confidence intervals for a population mean $\mu$**

For $0 \leq \alpha \leq 1$ a $(1 - \alpha) \times 100\%$ confidence interval for $\mu$ is given by:

$$\bar{x} \pm t_{\alpha/2}\ s/\sqrt{n}$$

Where $t_{\alpha/2}$ is the $\alpha/2$ critical value of the t-distribution with $df = n - 1$ (t-table).

**Definition 5**
For $0 \leq \alpha \leq 1$
The $\alpha$ critical value, $x_\alpha$, of a distribution of a random variable $X$ is the $x$-value such that $P(X \geq x_\alpha) = \alpha$ (Area to the right=$\alpha$).




In the section on statistical test we introduced, that tests can be conducted in 6 steps. Next we show how these 6 steps look specifically for the one-sample t-test.

**Hypothesis test for a mean $\mu$: The one sample t-test**

1. Hypotheses: Parameter of interest $\mu$.

   Choose one set of hypotheses and specify $\mu_0$, the value the population mean shall be compared with.

   | test type | hypotheses |
   |---|---|
   | upper tail | $H_0 : \mu \leq \mu_0$ vs. $H_a : \mu > \mu_0$ |
   | lower tail | $H_0 : \mu \geq \mu_0$ vs. $H_a : \mu < \mu_0$ |
   | two tail | $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ |

   Choose $\alpha$.

2. Assumptions: The sample is a random sample from the population.

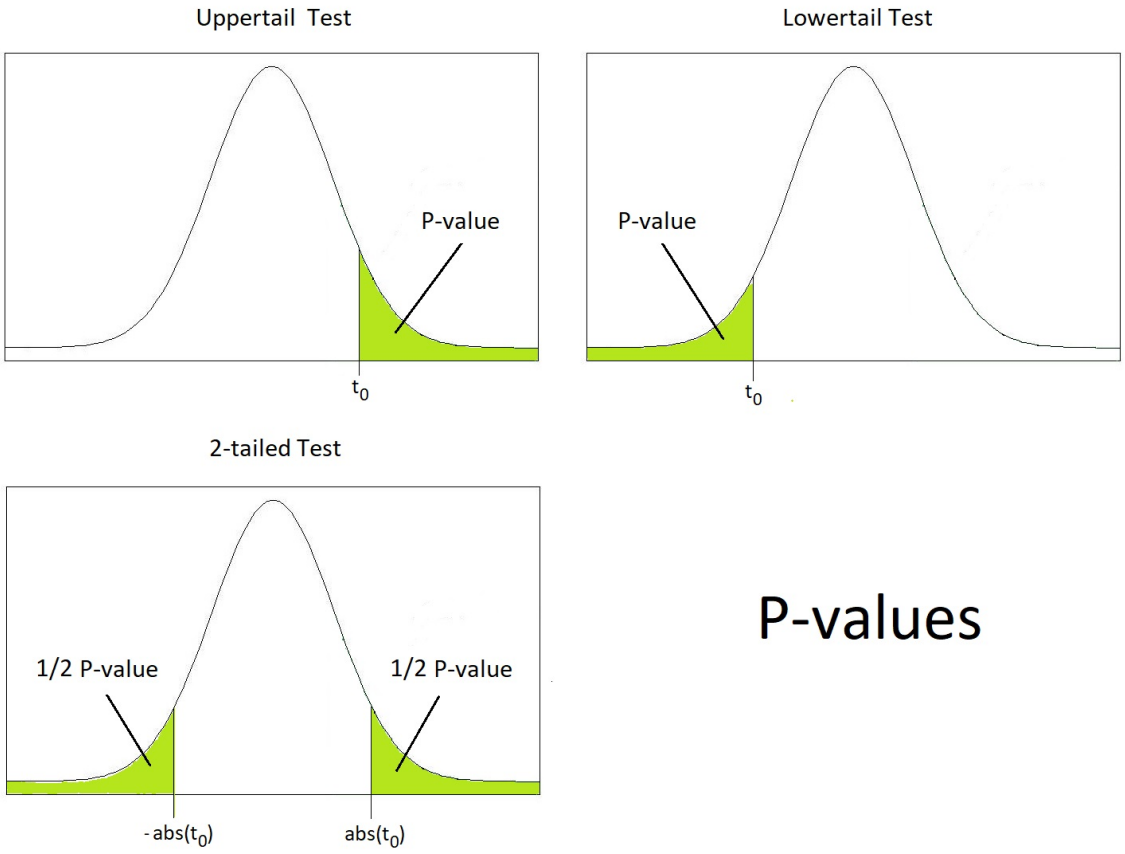
   Population is normally distributed, or the sample size is large.

3. Test statistic:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1$$

4. P-value:

| test type | P-value | |
|---|---|---|
| upper tail | $P(t > t_0)$ | P-value in upper tail |
| lower tail | $P(t < t_0)$ | P-value in lower tail |
| two tail | $2P(t > abs(t_0))$ | P-value in both tails |




5. Decision: If P-value $< \alpha$ reject $H_0$, otherwise do not reject $H_0$.

6. Conclusion: Put into context

**Example 6**
How much TV do students watch? A random sample of students at Penn State University (cp. Hetts, Heckard, Mind on Statistic) were asked : "How many hours TV do you watch a day in average?"
The sample data resulted in the following summary statistics:
$\bar{x} = 2.09, s = 1.644$ and $n = 175$, the median $M = 2$.
Is this data giving sufficient evidence that students at Penn State watch in average more than 2 hours TV per day?
In order to answer this question a one sided test should be conducted.
Let $\mu$ = mean number of hours students at Penn State watch TV

1. Hypotheses:
   $H_0 : \mu \leq 2$ vs. $H_a : \mu > 2$, use $\alpha = 0.05$

   *the question matches the expression in the alternative hypothesis*

2. Check assumptions:

   Is this a random sample? It's stated that the sample was chosen randomly.

   Is the variable $x$=number of hours students at Penn State watch TV normally distributed? It might not be exactly normally distributed, but close to. High outliers are possible (12 hours), but smaller outliers (below zero) are impossible.

3. Test statistic – one sample t

$$t_0 = \frac{2.09 - 2}{1.644/\sqrt{175}} = 0.7242, \quad df = 175 - 1 = 174$$

   *always include df s with test statistic, the value of the test score is only meaningful in context of the df*

4. P-value:
   *See the alternative hypothesis to observe that this is an upper tail test*
   P-value $= P(t > t_0) = P(t > 0.7242)$, is greater than 0.1 (t-table)

5. Decision:
   Since the P-value is greater than $\alpha = 0.05$ $H_0$ is NOT rejected and the test is not significant.

   *This means with this decision we might conduct an error of type II, for which we do not know the error probability, be careful not to commit with the decision when wording the conclusion.*

6. The data do not provide sufficient evidence that students at Penn State watch on average more than 2 hours TV per day.

   *For the students in the sample it was true, but the evidence from the sample is not enough that it is true for students in general at Penn State.*

**How much do the students watch in average? Estimate** the mean time students watch TV at Penn State using a 95% confidence interval.
$t_0.025$ from t-table is the 0.025 critical value. For 174 degrees of freedom use 1.984.

$$2.09 \pm 1.984 \times \frac{s}{\sqrt{n}} \qquad 2.09 \pm 0.2435$$

The CI is [1.864 ; 2.334].
We are 95% confident that the mean numbers of hours students at Penn State watch TV fall between 1.85 and 2.33 hours.
*The confidence example is consistent with the test result, according to the confidence interval the mean might be below 2 hours.*

**Example 7**
A used car dealer says that the mean price of a 2002 Ford F-150 is at least \$18800 (this is an example from 2012).
You suspect that this claim is incorrect and find that a random sample of 14 similar vehicles has a mean price of \$18000 with a standard deviation of \$1250. Is this sufficient evidence to disprove the dealer's claim?

1. Hypotheses:
   $H_0 : \mu \geq 18800$ vs. $H_a : \mu < 18800$, use $\alpha = 0.05$

2. Check assumptions:
   Is this a random sample? This depends on how the data was collected? Since you took STAT 252 you know to choose randomly the dealerships and randomly choose one of their vehicles that qualify.

   Is the variable $x$=price of Ford F-150 from 2002 normally distributed?

   It might not be exactly normally distributed, but close to.

3. Test statistic:
   $$t_0 = \frac{18000 - 18800}{1250/\sqrt{14}} = -2.39, \quad df = 14 - 1 = 13$$

4. P-value:
   According to the alternative hypothesis this is a lower tail test
   P-value $= P(t < t_0) = P(t < -2.39) = P(t > 2.39)$, According to t-table $0.01 < P - value < 0.025$

5. Decision:
   Since the P-value is less than $\alpha = 0.05$ $H_0$ is rejected.

   *Since $H_0$ is rejected one might conduct an error of type I, we know that the probability for this error is at most 5%, which we chose to represent an acceptable risk, we commit to the decision acknowledging the error probability*

6. At significance level of 5% the data do provide sufficient evidence that the mean price of a 2002 Ford F-150 is less than \$18800.

### 2.9.2 Descriptive Methods for Assessing Normality

Often we can not argue, why it should be reasonable to assume that the target population is normally distributed, and have to use the sample to assess if it is reasonable to make the assumption. Features of the distribution that would invalidate the normality assumption are: strong skew, outliers, and non-unimodal distribution.

Approaches for using sample data for assessing if it is reasonable to assume normality for the population

1. Construct a histogram or stem-and-leaf plot and note the shape of the graph. If the graph resembles a normal (bell- shaped) curve, the population could be normally distributed. Big deviations from a normal curve will leads us to reject the assumption (check for outliers, skew, and multiple peaks).

2. Compute the intervals $\bar{x} \pm s$, $\bar{x} \pm 2 \cdot s$, and $\bar{x} \pm 3 \cdot s$. If the empirical rule holds approximately, the data might be normally distributed. If big deviations occur, the assumption might be unreasonable.

   *Remember that according to the empirical rule form normally distributed populations 68% of measurements fall within 1 standard deviation of the mean, 95% within 2 standard deviations, and 99.7% within 3 standard deviations*

3. Construct a normal probability plot (normal Q-Q Plot) for the data. If the data are approximately normal, the points will fall (approximately) on a straight line.

Construction of normal probability plot (according to Weiss, Introductory Statistics):

1. Sort the data from smallest to largest and assign the rank, $k$ (smallest gets 1, second 2, etc.)
2. Find the relative rank $(k - 0.375)/(n + 0.25)$ for each measurement.
3. Find the z-scores (percentiles) for the relative rank (Table II).
4. Construct a scatter plot of the measurements and and the expected normal scores.

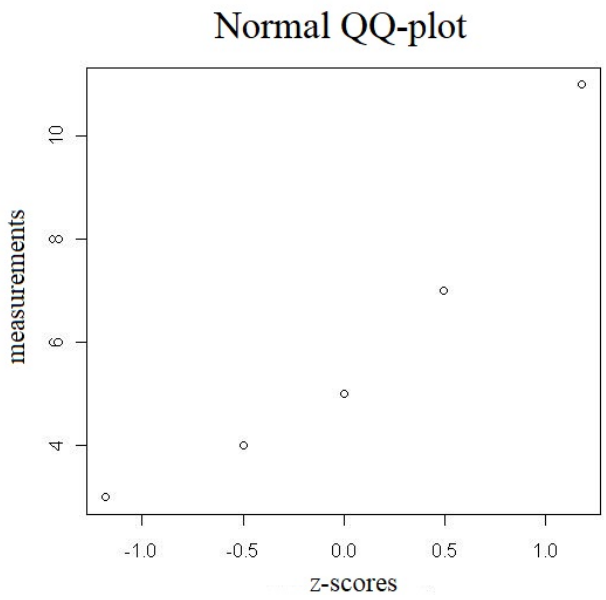
**Example 8**
Assume we have the data set 3,4,5,7,11
This is not enough data to use either method 1 or 2 to check for normality. Use a normal probability plot to investigate.
To create the plot, first setup a table with the following columns:

1. $x_i$, sorted measurements (from smallest to largest)
2. $k$, rank of the measurement, 1 means smallest
3. $(k-0.375)/(n+0.25)$ = relative rank, a number between 0 and 1 (excluding 0 and 1) indicating the relative standing of the measurement within the sample.
4. z-score, the z-score associated with a lower tail area equal to the relative rank (from z-table).

E.g. the middle value would have relative rank 0.5, with a z-score of 0.

| $x_i$ | $k$ | $(k - 0.375)/(n + 0.25)$ | $z$-score |
|---|---|---|---|
| 3 | 1 | (1-0.375)/(5.25)=0.119 | -1.18 |
| 4 | 2 | (2-0.375)/5.25=0.309 | -0.50 |
| 5 | 3 | (3-0.375)/5.25=0.5 | 0 |
| 7 | 4 | 0.691 | 0.50 |
| 11 | 5 | 0.881 | 1.18 |

Plot the measurements against the z-scores:




The points do not seem to fall on a straight line, but is this small deviation from a straight line enough evidence that the population is not normal?
In the lab you will look at a large number of QQ-plots suc that by the end of the term you will feel comfortable to assess QQ-plots.
If this would be a graph I would have to assess, I would conclude that the evidence is not very strong to reject that the sample originates from a normal population.

### 2.9.3 Comparing two Means – Paired Samples

Many studies are set up to compare two population means, $\mu_1$ and $\mu_2$. In this case two principle differences have to be considered.

1. The analysis is based on paired samples.

2. The analysis is based on independent samples.

Two samples are considered paired, if each measurement in the first sample directly relates to a measurement in the second sample.

For the analysis of paired samples we only need the sample of the differences in those paired observations

| sample 1, $x_1$ | sample 2, $x_2$ | differences, $x_1 - x_2$ |
|---|---|---|
| $x_{11}$ | $x_{21}$ | $x_{11} - x_{21}$ |
| $x_{12}$ | $x_{22}$ | $x_{12} - x_{22}$ |
| $x_{13}$ | $x_{23}$ | $x_{13} - x_{23}$ |
| $\vdots$ | $\vdots$ | |
| $x_{1n}$ | $x_{2n}$ | $x_{1n} - x_{2n}$ |

**Example 9**
Examples for paired samples are:

1. blood pressure before (sample 1) and after (sample 2) therapy (same patients)
2. pain score before surgery (sample 1) and after surgery (sample 2) (same patients)
3. hight of father (sample 1), hight of son (sample 2)
4. run-time using algorithm 1 (sample 1) and algorithm 2 (sample 2) for same problem
5. stock values in January (sample 1) and stock value of the same stocks in July (sample 2)

**t-score for standardizing the mean differences**
For paired random samples let $\bar{X}_d$ be the sample mean of the differences and $s_d$ the sample standard deviation of the differences, $\mu_d$ the population mean of the differences, and $\sigma_d$ the population standard deviation of the differences.
Then

$$t = \frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}}$$

is t-distributed with degrees of freedom $df = n - 1$.

This is virtually the same result as stated for a single sample, only that the formula is applied to the sample of differences, which means that 1-sample t-confidence interval and t-test can be used for paired samples, once the sample of differences has been found.

**Calculating confidence intervals for the difference between two population means, $\mu_1 - \mu_2$, using paired samples**
$(1 - \alpha) \times 100\%$ confidence interval for the difference in the population means, $\mu_d = \mu_1 - \mu_2$, is given by

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$$

with $t^* = \alpha/2$ critical value of the t-distribution with $df = n - 1$.

The **Paired t-test** is the same as the 1-sample t test applied to the sample of differences.
Instead of $\mu$ being the parameter of interest it is now $\mu_d = \mu_1 - \mu_2$. Make sure to use $x_d = x_1 - x_2$ (not $x_2 - x_1$) matching the direction in the difference in $\mu_d$.

**Example 10**
According to research the mere belief in receiving effective treatment for pain can reduce the pain you actually feel (Science Feb.20, 2004). Researchers tested this on 24 volunteers.
Each volunteer was put into an MRI for two consecutive sessions. During the first electric shocks were applied to their arms and the blood oxygen level dependent (BOLD) signal was recorded during pain. This signal is related to neural activity in the brain.
The second session was identical, except that prior researchers applied a lotion to the volunteer's arms, telling them that the lotion would block the pain. The lotion was a regular skin lotion i.e. a placebo.
The recordings showed for the $n = 24$ volunteers a mean difference (first - second session's BOLD) in the BOLD signal of $\bar{x}_d = 0.21$ with a standard deviation of $s_d = 0.47$. Does this indicate that the

placebo decreased the neural activity in the brain, i.e. the pain felt by the volunteers? If the placebo is effective the measurement in the first session should be higher than in the second session.

The samples are paired, since the same patients were tested, so a paired t-test should be conducted for comparing the mean BOLD in the two sessions. Let $\mu_d$ = mean of the difference in the BOLD signal for the two sessions, i.e. the decrease in the mean BOLD from session1 to session 2.

1. Hypotheses:
   $H_0 : \mu_d \leq 0$ vs. $H_a : \mu_d > 0$, use $\alpha = 0.05$

2. Check assumptions: Is this a random sample? Are volunteers a random sample? What do you think? What does that mean for the interpretation of the result?

   Is the variable $X$=difference in the recordings for session one and session 2 coming from a normal distribution? It is reasonable to assume that they are normally distributed. The difference is a numerical variable which will be centered around a mean, there is no reason to assume that one type of outlier is more likely than the other. A histogram and a QQ-plot of the data should be analyzed, but we do not have the original data :-(

3. Test statistic – one-sample t applied to paired data
   $$t_0 = \frac{0.21 - 0}{0.47/\sqrt{24}} = 2.188, \quad df = 24 - 1 = 23$$

4. P-value – upper tail test (see alternative hypothesis)
   P-value $= P(t > t_0) = P(t > 2.188)$, falls between 0.010 and 0.025 (t-table).

5. Decision: Since the P-value is smaller than $\alpha = 0.05$ $H_0$ is rejected (the test is significant).

6. At significance level of 5% the data provide sufficient evidence that the placebo was effective in reducing the mean BOLD, indicating a reduction in felt pain.

   *What are the limitations of this result arising out of the selection of the sample?*

To estimate the difference in the mean BOLD a 95% confidence interval shall be used

$$\bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}$$
$$0.21 \pm 2.069 \frac{0.47}{\sqrt{24}}$$
$$0.21 \pm 0.1985$$
$$[0.0115, 0.4085]$$

With 95% confidence the data indicates that the mean difference in BOLD between two sessions as described above falls between 0.0115 and 0.4085.
Since 0 does not fall with in the confidence interval there is at significance level of (100-95)=5% evidence that the mean difference is not 0. Since the interval falls entirely above 0,we can be 95% confident that the mean difference is greater than 0, indicating that the felt pain decreased.

### 2.9.4 Comparing Two Means – Independent Samples

The other situation to consider when comparing two population means is having two independent samples. Two samples are considered independent, if the measurement in the two samples were obtained from different, unrelated objects/individuals.
When analyzing this type of samples, we need the sample means, sample standard deviations, and sample sizes for both samples.

| | sample 1 | sample 2 | population 1 | population 2 |
|---|---|---|---|---|
| mean | $\bar{x_1}$ | $\bar{x}_2$ | $\mu_1$ | $\mu_2$ |
| standard deviation | $s_1$ | $s_2$ | $\sigma_1$ | $\sigma_2$ |
| sample size | $n_1$ | $n_2$ | | |

For paired samples the sample sizes for the two samples have to be identical since for every measurement in sample 1 is a paired observation in sample 2. For independent samples the sample sizes can be very different, but are ideally (almost) the same.

**t-score for standardizing the differences between the sample means from two independent samples**

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu 2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

is t-distributed with degrees of freedom $df \approx \min(n_1 - 1, n_2 - 1)$.
There exist more sophisticated formulas for the approximate degrees of freedom for this t-score, but for hand calculations this formula will serve us well enough.

**Calculating confidence intervals for the difference between two population means, $\mu_1 - \mu_2$, using independent samples – two sample t confidence interval**
A $(1 - \alpha) \times 100\%$ confidence interval for the difference in the population means, $\mu_1 - \mu_2$, is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

with $t^* = (\alpha/2)$ critical value of the t-distribution with $df \approx \min(n_1 - 1, n_2 - 1)$.

**Two sample t-test for comparing two means**

1. Hypotheses: (parameter of interest $\mu_1 - \mu_2$)

   | test type | hypotheses |
   |---|---|
   | upper tail | $H_0 : \mu_1 - \mu_2 \leq d_0$ vs. $H_a : \mu_1 - \mu_2 > d_0$ |
   | lower tail | $H_0 : \mu_1 - \mu_2 \geq d_0$ vs. $H_a : \mu_1 - \mu_2 < d_0$ |
   | two tail | $H_0 : \mu_1 - \mu_2 = d_0$ vs. $H_a : \mu_1 - \mu_2 \neq d_0$ |

   $d_0$ is the value the difference is compared against, often $d_0 = 0$.

   Choose $\alpha$.

2. Assumptions:
   Independent random samples.

   Both populations are normally distributed, or both sample sizes are large.

3. Test statistic:
$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad df = \min(n_1 - 1, n_2 - 1)$$

4. P-value: (exactly the same in 1-sample t-test)

| test type | P-value |
|---|---|
| upper tail | $P(t > t_0)$ |
| lower tail | $P(t < t_0)$ |
| two tail | $2P(t > abs(t_0))$ |

5. Decision: If P-value $< \alpha$ reject $H_0$, otherwise do not reject $H_0$.

6. Conclusion: Put into context

**Pooled t-test for comparing two population means**

In text books and other courses sometimes the Pooled t-test and t-confidence interval for comparing two means are introduced as an alternative to the two-sample t-test and confidence interval.
For the pooled test and confidence interval an additional assumption has to be met:
the standard deviation in the two populations are the same ($\sigma_1 = \sigma_2$)

The pooled test and confidence interval should be avoided. Its advantage is that the type 2 error is smaller than for the two sample t-test, if the extra assumption is correct, but on the downside if the assumption is not correct, the result becomes invalid. Since using data there is no way beyond a census to check if the assumption is met, and a two sample t-test should be used.
Only if it can be argued through theory that the standard deviations are the same the pooled test may be used. In my many years of applied statistics no researcher could provide such theoretical argument.

**Example 11**

In the paper "Migratory Dark-Eyed Juncos, Junco Hyemalis, Have Better Spatial Memory and Denser Hippocampal Neurons than Nonmigratory Conspecifics" (animal Behaviour, Vol. 66, pp. 317-328) by D. Cristol et al. different features of two subspecies of dark-eyed juncos were compared. One feature of interest was the wing lengths, which is of interest because one subspecies was migratory and the other was nonmigratory.
Migratory black-eyed juncos can be observed in Edmonton between spring and fall.
The wing lengths of 14 birds of each sub species were measured, the results are summarized in the following table.

| | sample size | mean | standard deviation |
|---|---|---|---|
| migratory | 14 | 82.1 mm | 1.501 mm |
| nonmigratory | 14 | 84.9 mm | 1.698 mm |

At significance level of 5% do these data indicate a difference in the mean wing length of the two subspecies?

Let $\mu_m$ = mean wing length of the migratory subspecies and $\mu_{nm}$ = mean wing length of the nonmigratory subspecies. Then the parameter of interest is $\mu_m - \mu_{nm}$.

1. Hypotheses:
   $H_0 : \mu_m - \mu_{nm} = 0$ vs. $H_a : \mu_m - \mu_{nm} \neq 0$, use $\alpha = 0.05$

2. Check assumptions: Are the samples random? That depends how they collected the sample, one should check the method section in the paper. For the moment let's assume the samples represent random samples of the subspecies.

   Are the variables $X_m$=wing length of the migratory subspecies and $X_{nm}$=wing length of the nonmigratory subspecies normally distributed?

   Measurements of length weight, etc., can usually be considered normally distributed.

3. Test statistic – two sample t

$$t_0 = \frac{82.1 - 84.9}{\sqrt{\frac{1.501^2}{14} + \frac{1.698^2}{14}}} = \frac{-2.8}{0.6057} = -4.623, \quad df = 13$$

4. P-value – 2-tailed test
   P-value $= 2 \times P(t > abs(t_0)) = 2 \times P(t > 4.623)$, is smaller than (t-table ) $2 \times 0.0005 = 0.001$.

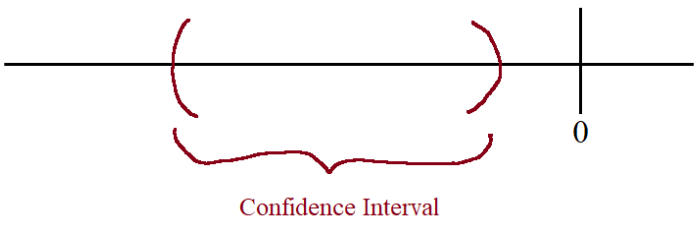
5. Decision:
   Since the P-value is smaller than $\alpha = 0.05$ $H_0$ is rejected (the test is significant).

6. At significance level of 5% the data provide sufficient evidence that the mean wing length of the two subspecies under investigation differ from each other.

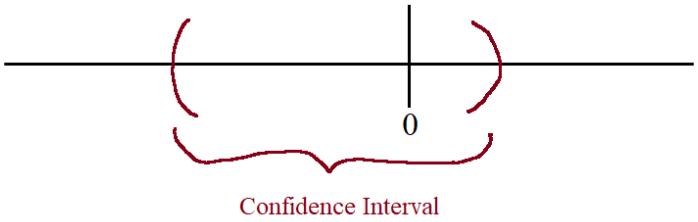
To **estimate** the difference in the mean wing length of migratory and nonmigratory black-eyed juncos use a 95% 2-sample t confidence interval.

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
$$-2.8 \pm 2.160(0.6057)$$
$$-2.8 \pm 1.308$$
$$[-4.108, -1.492]$$

We are 95% confident that the difference in mean winglength of migratory black-eyed between 1.5mm and 4.1mm shorter than for nonmigratory black-eyed juncos.
Since 0 does not fall with in the confidence interval the data provide at significance level of (100-95)=5% evidence that the mean difference is not 0. Since the interval falls entirely below 0, we can be 95% confident that the mean difference is lower than 0, indicating that the mean winglength migratory birds is smaller than for nonmigratory birds. The reason might be that migratory birds are of ten smaller than the nonmigratory birds, ask your bio prof)



If zero falls within a confidence interval, we conclude that at the appropriate confidence level the mean difference might be less than, equal to, or greater than 0. No direction can be concluded, and it also does not mean that the means are the same, they might be, but also might be different.




If 0 falls within a confidence interval the matching two-tailed test is always not significant and vice versa.

## 2.10 Conclusion

This concludes the review of the material covered in the pre-requisite courses for STAT 252. If some of these concepts are still very hazy, please make sure to attend office hours or ask for clarifications by email. All subsequent material will be based on the concepts discussed in this section.