The Center of Applied Data Science

python™

Statistical Data Analysis

Day 2.2

Content Outline

In [ ]:

```
import pandas as pd
import numpy as np
import seaborn as sns
from scipy.stats import norm
```

1. t-distribution

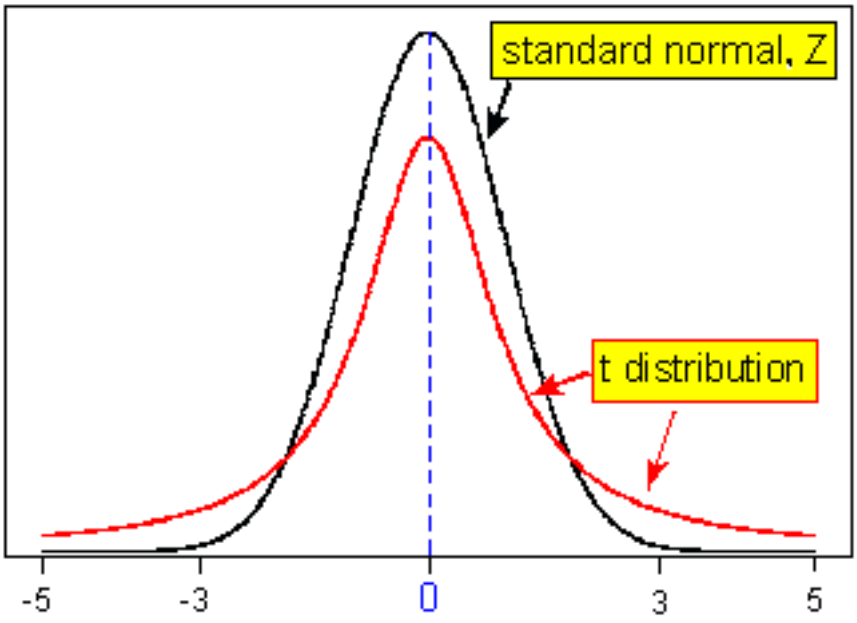
We have seen that random variables can be visually modeled by many different sorts of shapes, and we call these shapes distributions. Several distributions arise so frequently that they have been given special names, and they have been studied mathematically.

The **t-distribution** is another bell-shaped (unimodal and symmetric) distribution, like the normal distribution; and the center of the t-distribution is standardized at zero, like the center of the standard normal distribution which we call it z-distribution as well.

Like all distributions that are used as probability models, the normal and the t-distribution are both scaled, so the total area under each of them is 1.

So how is the t-distribution fundamentally different from the normal distribution? The **spread**.

The following picture illustrates the fundamental difference between the normal distribution and the t-distribution:



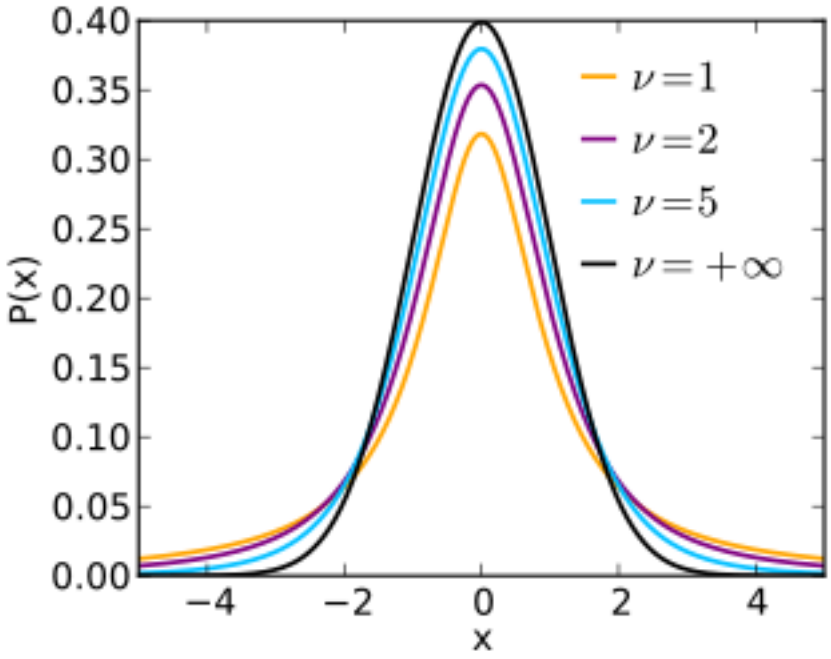
The t-distribution has slightly less area near the expected central value than the normal distribution does, and has correspondingly more area in the "tails" than the normal distribution does. (It's often said that the t-distribution has "fatter tails" or "heavier tails" than the normal distribution.)

This reflects the fact that the t-distribution has a larger spread than the normal distribution. The same total area of 1 is spread out over a slightly wider range on the t-distribution, making it a bit lower near the center compared to the normal distribution, and giving the t-distribution slightly more probability in the 'tails' compared to the normal distribution.

Therefore, the t-distribution ends up being the appropriate model in certain cases where there is more variability than would be predicted by the normal distribution.

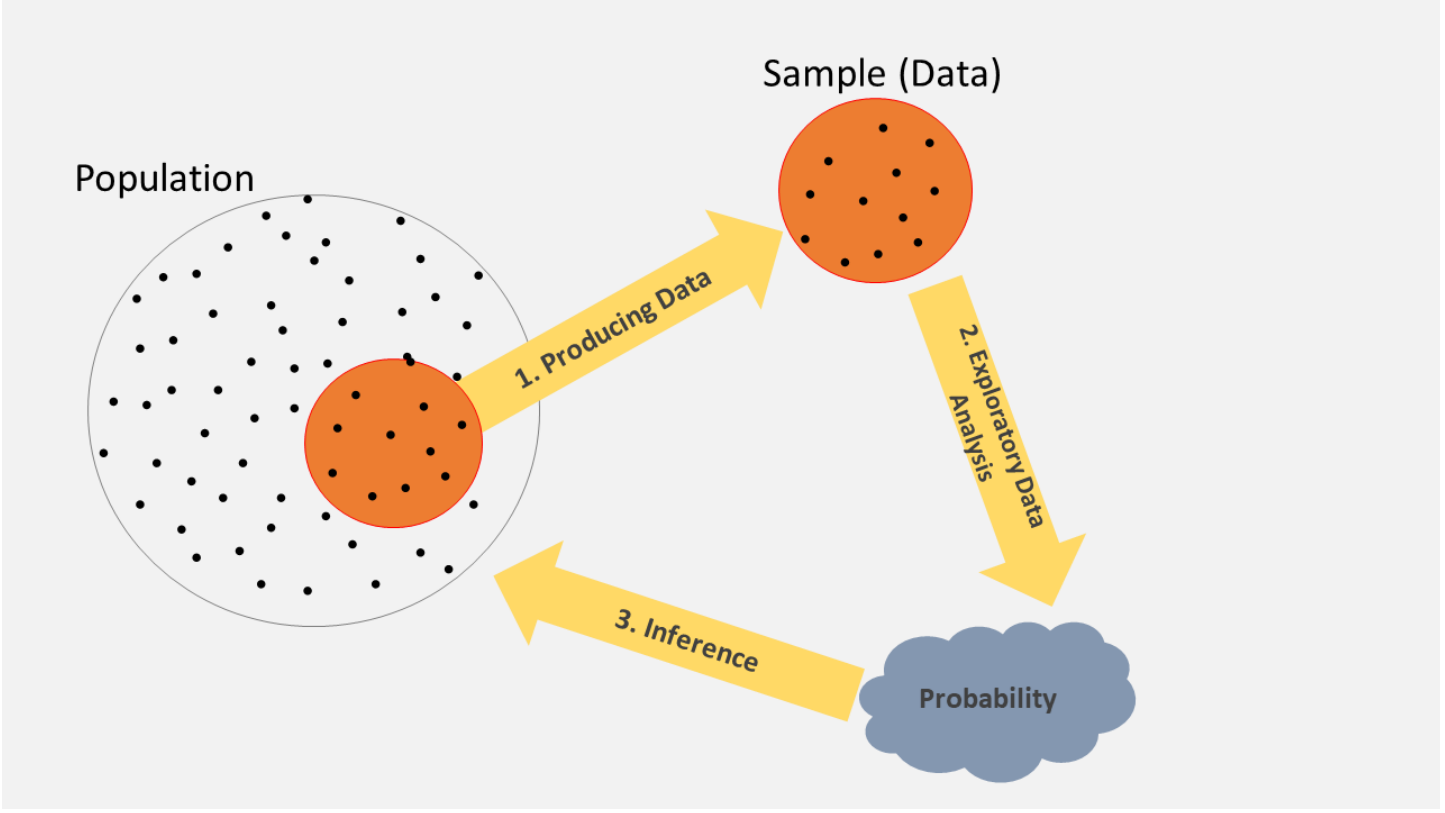
There's actually an entire family of t-distributions. They all have similar formulas (but the math is beyond the scope of this introductory course in statistics), and they all have slightly "fatter tails" than the normal distribution. But some are closer to normal than others. The t-distributions that are closer to normal are said to have higher "degrees of freedom" (that's a mathematical concept that we won't use in this course, beyond merely mentioning it here). So, there's a t-distribution "with one degree of freedom," another t-distribution "with 2 degrees of freedom" which is slightly closer to normal, another t-distribution "with 3 degrees of freedom." which is a bit closer to normal than the previous ones, and so on.

The following picture illustrates this idea with a few t-distributions (note that "degrees of freedom" is shown by ν):



2. Inference

Our ultimate goal in statistical data analysis is using a sample to make inferences or draw conclusions about the population from which it was drawn.



Our choice of the type of inference depends on the type of the variable of interest. We introduce three forms of statistical inference in this unit, each one representing a different way of using the information obtained in the sample to draw conclusions about the population. These forms are:

- Point estimation
- Interval estimation
- Hypothesis testing

A. Point Estimation

We estimate an unknown parameter using a **single number** that is calculated from the sample data.

Example

Based on sample results, we estimate that pp, the proportion of Malaysian adults who are in favor of increasing taxes on tobacco products, is 0.6.

B. Interval Estimation

We estimate an unknown parameter using an **interval of values** that is likely to contain the true value of that parameter and state how confident we are that this interval indeed captures the true value of the parameter.

Confidence intervals are not perfect. A 95% confidence interval implies that in repeated samples, 19 in 20 confidence intervals will capture the value of the population parameter.

Example

Based on sample results, we are 95% confident that pp, the proportion of Malaysian adults who are in favor of increasing taxes on tobacco products, is between 0.57 and 0.63.

B.1- Interval Estimation for Population Proportion (Categorical Variable)

Suppose we are interested in the population proportion of a categorical variable.

- **Step 1**: We collect data from a sample of our population of size nn
- **Step 2**: The values of $\hat{p}\hat{p}$ follow a normal distribution with (unknown) mean pp and standard deviation $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{\hat{p}\,(1-\hat{p}\,)}{n}}$. As we do not know the population proportion pp, we use the sample proportion $\hat{p}\hat{p}$.
- **Step 3**: According to the Standard Deviation Rule, this means that:
  - We are 95% confident that the population proportion pp falls within $2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,2 * \sqrt{\frac{\hat{p}\,(1-\hat{p}\,)}{n}}$ of our estimate $\hat{p}\hat{p}$.
  - A 95% confidence interval for the population proportion pp is:

$$\left(\hat{p} - 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$\left(\hat{p} - 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Here, then, is the general result:

Suppose a random sample of size n is taken from a population for a categorical variable whose proportion (p) is unknown. A 95% confidence interval (CI) for p is:

$$\left(\hat{p} - 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$\left(\hat{p} - 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Example

A few days before a snap election, a polling organisation would like to estimate p p, the proportion of eligible voters who support Candidate A. They choose a random sample of size 1000 and recorded their opinion. $71\%71\%$ of the sample

support this candidate. How do you estimate the proportion of the people in the constituency who will vote for this candidate?

Point estimate: $\hat{p} = 71\%$ $\hat{p} = 71\%$

Interval estimate: According to the *central limit theorem*, sample proportion, $\hat{p}\hat{p}$, follows the normal distribution

$$\hat{p} \sim Normal\left(mean = p, sd = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$\hat{p} \sim Normal\left(mean = p, sd = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

So, we can say $\bar{p}\bar{p}$ follows the normal distribution

$$Normal\left(mean = p, sd = \sqrt{\frac{0.71(1-0.71)}{1000}}\right)$$

$$Normal\left(mean = p, sd = \sqrt{\frac{0.71(1-0.71)}{1000}}\right)$$

$$Normal(mean = p, sd = 0.014)$$

$$Normal(mean = p, sd = 0.014)$$

where pp is the population proportion. Therefore, we can say that

- there is 95% chance that pp falls within $2\sigma_{\bar{p}}\sigma_{\bar{p}}$ of $\hat{p}\hat{p}$.
- Using the empirical rule, we can say the 95% confidence interval for pp is $(\hat{p} - 2\sigma_{\bar{p}}, \hat{p} + 2\sigma_{\bar{p}}) = (0.71 - 2 * 0.014, 0.71 + 2 * 0.014) = (0.68, 0.73)$ $(\hat{p} - 2\sigma_{\bar{p}}, \hat{p} + 2\sigma_{\bar{p}}) = (0.71 - 2 * 0.014, 0.71 + 2 * 0.014) = (0.68, 0.73)$, where $\sigma_{\bar{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\sigma_{\bar{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- This means that we are 95% confident that pp lies within the interval (0.68, 0.73).

Exercise

Several public health researchers conducted a study to look at the connection between watching actors smoking in movies and initialising of smoking among adolescents. In the study, 6,522 teenagers aged 10-14 who had never tried smoking were randomly selected. Of those who subsequently tried smoking for the first time, 38% were exposed to smoking in the movies.

- A. Estimate the proportion of all U.S. adolescents ages 10-14 who started smoking after seeing actors smoke in movies by constructing a 95% confidence interval.
- B. Construct a 99.7% confidence interval for pp.

In [ ]:

```
1
```

B.2- Interval Estimation for Population Mean (Numerical Variable)

Suppose we are interested in the mean of a numerical variable from a population.

- **Case1:** We assume the population standard deviation ($\sigma\sigma$) is **known**.

- **Step 1:** We collect data from a sample of our population of size nn
- **Step 2:** The values of $\bar{x}\bar{x}$ follow a normal distribution with (unknown) mean $\mu\mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (known, since both $\sigma\sigma$ and nn are known).
- **Step 3:** According to the Standard Deviation Rule, this means that:
  - There is a 95% chance that our population mean $\mu\mu$ will fall within $2 * \frac{\sigma}{\sqrt{n}}$ $2 * \frac{\sigma}{\sqrt{n}}$ of $\hat{\mu}\hat{\mu}$
  - A 95% confidence interval for the population mean

$$\mu \in \left(\bar{x} - 2 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 * \frac{\sigma}{\sqrt{n}}\right)$$

$$\mu \in \left(\bar{x} - 2 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 * \frac{\sigma}{\sqrt{n}}\right)$$

Here, then, is the general result:

Suppose a random sample of size n is taken from a normal population of values for a quantitative variable whose mean ($\mu\mu$) is unknown, when the standard deviation ($\sigma\sigma$) is given. A 95% confidence interval (CI) for

$$\mu \in \left(\bar{x} - 2 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 * \frac{\sigma}{\sqrt{n}} \right)$$

$$\mu \in \left(\bar{x} - 2 * \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 * \frac{\sigma}{\sqrt{n}} \right)$$

- Case2: We assume the population standard deviation ($\sigma\sigma$) is **unknown**. In this case, we can replace $\sigma\sigma$ with s, where s is the standard deviation of the sample however, the central limit theorem will not be valid anymore and $\bar{x}\bar{x}$ will not follow normal distribution. Instead, $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{(n)}}} t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{(n)}}}$ will follow t-distribution with degree of freedom n - 1, where n is the sample size.

$$t^* = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t^* = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t^* \sim t(n-1)$$

$$t^* \sim t(n-1)$$

Therefore, for each confidence interval $t^* t^*$ should be calculated such that:

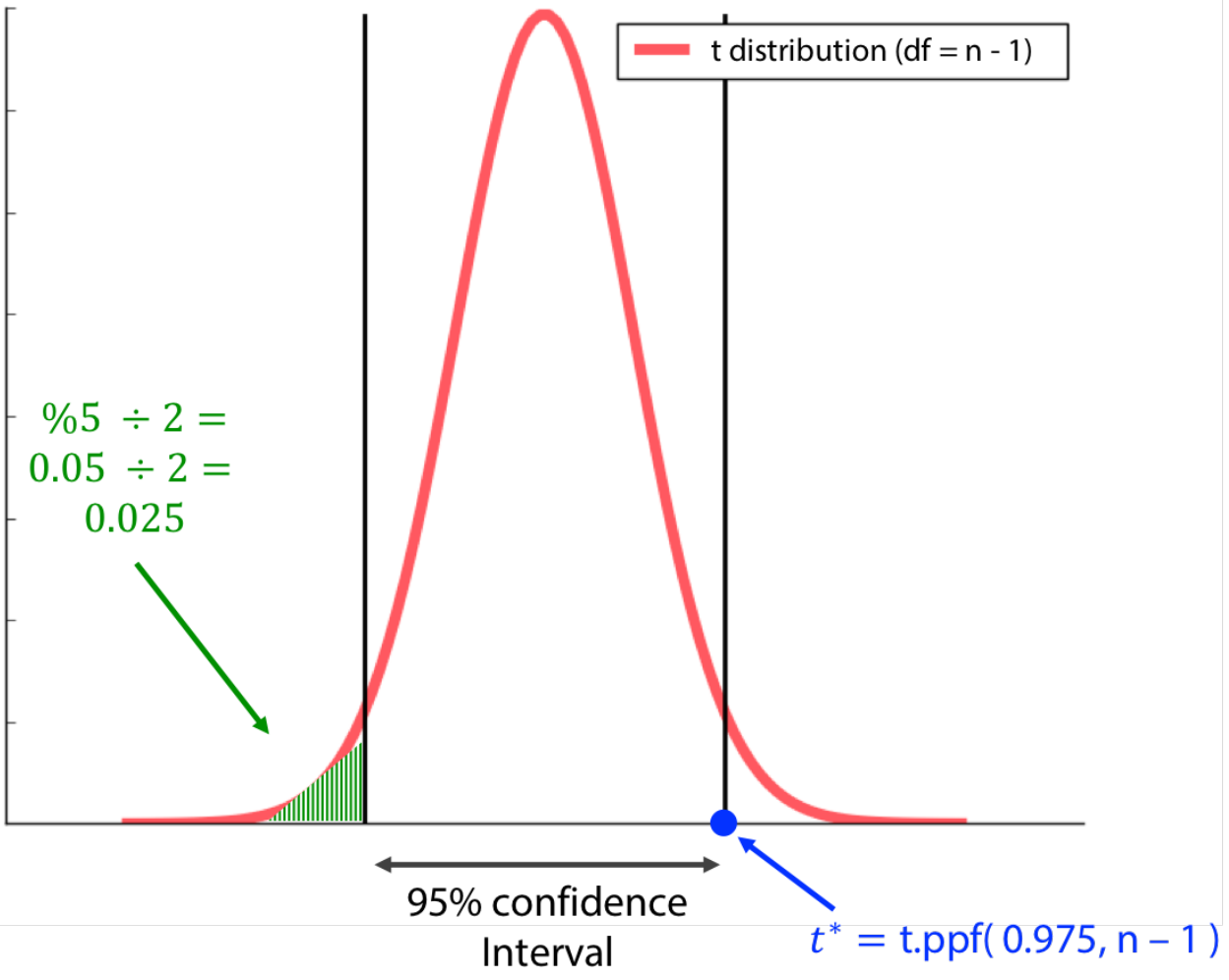
$$\mu \in \left(\bar{x} - t^* * \frac{s}{\sqrt{n}}, \bar{x} + t^* * \frac{s}{\sqrt{n}} \right)$$

$$\mu \in \left(\bar{x} - t^* * \frac{s}{\sqrt{n}}, \bar{x} + t^* * \frac{s}{\sqrt{n}} \right)$$

.

Because $t^* \sim t(n-1) t^* \sim t(n-1)$, the above interval depends on both the **confidence level and the sample size n**. For instance, the 95% cut-off in the following t-distribution can be calculated using python by $t^* = t.ppf(0.975, n-1)$

$t^* = t.ppf(0.975, n-1)$



Example

Numerical Variable Mean: Suppose we are interested in studying the average IQ of students in a university. To do so, we collect a random sample of size 100 from the students in this university. Assume the mean of the IQ level of these students is 115, and its standard deviation is $\sigma = 5$ $\sigma = 5$. What is μ μ, the mean of the IQ level of the population which is the whole students at this university?

Point estimate: $\hat{\mu} = 115 \hat{\mu} = 115$
Interval estimate: $\hat{\mu} = 115, \sigma = 5, n = 100$
$\hat{\mu} = 115, \sigma = 5, n = 100.$

- Note that $\frac{5}{\sqrt{100}} \frac{5}{\sqrt{100}}$ is the standard deviation of the sampling distribution of sample estimates $\bar{x} \bar{x}$ - the *standard error*, $\sigma_{\bar{x}} \sigma_{\bar{x}}$.
- According to the *central limit theorem*, the distribution of the sample means $\bar{x} \bar{x}$(s) follows a normal distribution:

$$\bar{x} \sim Normal(mean = \mu, sd = \frac{\sigma}{\sqrt{n}})$$

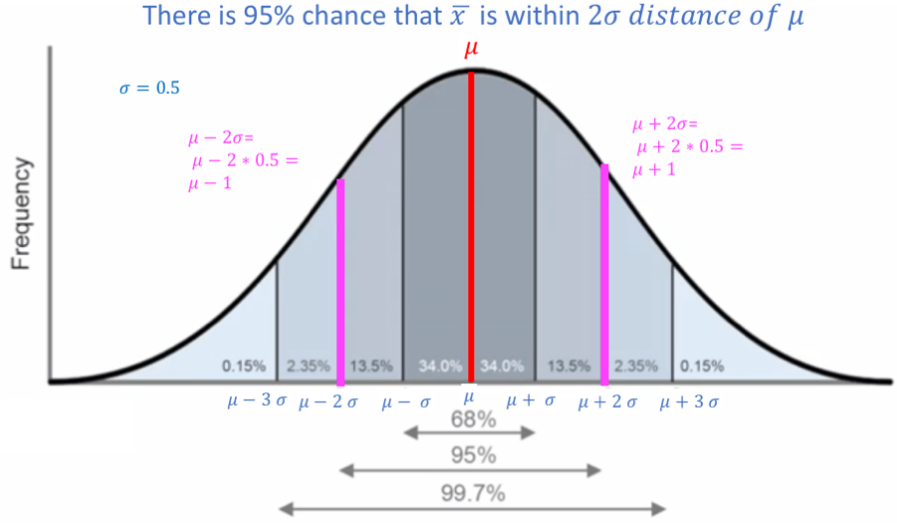
$$\bar{x} \sim Normal(mean = \mu, sd = \frac{\sigma}{\sqrt{n}})$$

- Since $\hat{\mu} \hat{\mu}$ is our estimate of $\mu \mu$, the sample means are distributed as

$$\bar{x} \sim Normal(115, 0.5)$$

$$\bar{x} \sim Normal(115, 0.5)$$

- Recall the standard deviation rule:



Since two standard errors = 1, the statement:

> "There is a **95% chance** that the sample mean $\bar{x} \bar{x}$ falls within 1 unit of $\mu \mu$".

can be rephrased as:

> "We are 95% confident that the population mean $\mu\mu$ falls within 1 units of $\bar{x}\bar{x}$".

Given a sample mean of $\bar{x} = 115\bar{x} = 115$, we can be **95% confident** that $\mu\mu$ falls within 1 unit of 115, or in other words that $\mu\mu$ is covered by the interval $(115 - 1, 115 + 1) = (114, 116)$ $(115 - 1, 115 + 1) = (114, 116)$.

Exercise

An educational researcher was interested in estimating $\mu\mu$, the mean score on the total SAT scores of all college students in a state. To this end, the researcher has chosen a random sample of 650 college students from his state, and found that their average SAT score is 1425. Based on a large body of research that was done on the SAT, it is known that the scores roughly follow a normal distribution with the standard deviation $\sigma = 300\sigma = 300$.

- A. Based on this information, construct a 95% confidence interval for $\mu\mu$.
- B. Construct a 99.7% confidence interval for $\mu\mu$.

In [ ]:

```
1
```

Exercise

What is the relationship between the level of the confidence and the length of the confidence interval?

In [ ]:

```
1
```

Example

Repeat the above example if population standard deviation is unknown but sample standard deviation is 4.5.

In [ ]:

```
from scipy.stats import t

x_bar = 115
n = 100
s = 4.5
df = n - 1
c = 0.95
beta = c + (1 - c)/2
t_star = t.ppf (beta , df)
print('Population mean estimate by 95% confidence interval: (
      round(x_bar-t_star*s/(n ** .5),4), ', ', round(x_bar+t_
```

Exercise

Repeat the above exercise it the population standard deviation is unknown, but the sample standard deviation is 298.

- A. Based on this information, construct a 95% confidence interval for $\mu\mu$.
- B. Construct a 99.7% confidence interval for $\mu\mu$.

In [ ]:

```

```

C. Hypothesis testing

The disciplinary committee of a university investigates a student suspected of cheating on an exam. There are two opposing claims in this case:

- The student claims that he did not cheat on the exam.
- The lecturer claims that the student did cheat on the exam.

The committee assumes the student to be innocent unless the lecturer can prove that the student is guilty. Therefore, the committee asks the instructor to provide evidence to support his claim. The lecturer explains that he set two versions of the exam, and on four separate exam questions, the student answered with numbers provided in the other version of the exam.

The committee agrees that it would be extremely unlikely for the lecturer to have such strong evidence if the student did not cheat. In other words, the lecturer provided strong enough evidence for the committee to **reject** the student's claim, and **conclude** that the student did cheat on the exam.

Hypothesis testing is defined as **assessing evidence provided by the data in favour of or against some claim about the population.**

Here is how the process of statistical hypothesis testing works:

- **Step 1:** We have **two claims** about what is going on in the population: claim 1 and claim 2. In the story above, where the instructor's claim challenges the student's claim, **claim 1 is challenged by claim 2**. In hypothesis testing, we usually test 'claims' (or hypotheses) about the value of population parameter(s) or about whether a relationship exists between two variables in the population.
- **Step 2:** We choose a **sample**, collect relevant data and summarize them. This is similar to the instructor collecting evidence from the student's exam.
- **Step 3:** We figure out **how likely** it is to observe data like the data we got, had claim 1 been true. (Note that the wording "how likely ..." implies that this step requires some kind of probability calculation). In our story, the committee members assessed how likely it is to observe the evidence provided by the instructor if the student's claim of not cheating was true.
- **Step 4:** Based on what we found in the previous step, we make our decision:
  - If we find that it would be extremely unlikely to observe the data that we observed if claim 1 were true, then we have strong evidence against claim 1, and we **reject** it in favour of claim 2.
  - If we find that observing the data that we observed is not very unlikely if claim 1 were true, then we do not have enough evidence against claim 1, and therefore we **cannot reject** it in favour of claim 2.

In our story, the committee decided that it would be extremely unlikely to find the evidence that the lecturer provided if the student did not cheat. In other words, the members felt that it is extremely unlikely that it is just a coincidence that the student used the numbers from the other version of the exam on four separate problems. The committee members therefore decided to reject the student's claim and concluded that the student had, indeed, cheated on the exam.

Example

A recent study estimated that 14.6% of all upper secondary school students in Malaysia smoke. (https://tobaccoinduceddiseases.biomedcentral.com/articles/10.1186/s12971-016-0108-5). The head of a district education office suspects that the proportion of smokers may be lower there. In hopes of confirming her claim, she chooses a random sample of 400 upper secondary high school students in the district, and finds that 50 of them are smokers.

Let's analyze this example using the 4 steps outlined above:

Stating the claims:

There are two claims here:

- *claim 1:* The proportion of smokers in the district is 0.146.
- *claim 2:* The proportion of smokers at Goodheart is less than 0.146.

Claim 1 basically says "nothing special goes on in this district; the proportion of smokers there is no different from the proportion in the entire country." This claim is challenged by the head of the district office, who suspects that the proportion of smokers in her district is lower.

Choosing a sample and collecting data:

A sample of $n = 400$ $n = 400$ was chosen, and summarizing the data, we find that the sample proportion of smokers is $\hat{p} = \frac{50}{400} = 0.125$ $\hat{p} = \frac{50}{400} = 0.125$

While it is true that 0.125 is less than 0.146, it is not clear whether this is strong enough evidence against claim 1.

Assessment of evidence:

To assess whether the data provide strong enough evidence against claim 1, we need to ask ourselves: How surprising is it to get a sample proportion as low as $\hat{p} = 0.125$ $\hat{p} = 0.125$ (or lower) if claim 1 is true?

In other words, we need to find how likely it is that in a random sample of size $n = 400$ $n = 400$ taken from a population where the proportion of smokers is $p = 0.146$ $p = 0.146$ we'll get a sample proportion as low as $\hat{p} = 0.125$ $\hat{p} = 0.125$ (or lower).

It turns out that the probability that we'll get a sample proportion as low as $\hat{p} = 0.125$ $\hat{p} = 0.125$ (or lower) if $p = 0.146$ $p = 0.146$ is roughly 0.117 (do not worry about how this was calculated at this point).

Conclusion:

We found that there is a probability of 0.117 of observing data like that observed if claim 1 were true.

Now you have to decide ... Do you think that a probability of 0.117 makes our data rare enough (surprising enough) under claim 1 so that the fact that we did observe it is enough evidence to reject claim 1? Or do you feel that a probability of 0.117 means that the data we observed are not very likely when claim 1 is true, but not unlikely enough to conclude that getting such data is sufficient evidence to reject claim 1?

Hypothesis testing (General Case)

- **Step 1: Stating the claims**: Our aim is to decide between two opposing points of view, *Claim 1* and *Claim 2*. In hypothesis testing, Claim 1 is called the **null hypothesis** (denoted $H_0 H_0$), and Claim 2 plays the role of the **alternative hypothesis** (denoted $H_a H_a$).
- **Step 2: Choosing a sample and collecting data**: We look at sampled data to draw conclusions about the entire population. In hypothesis testing, based on the data, you draw conclusions about whether there is enough evidence to reject H_0 H_0.
- **Step 3: Assessing the evidence**: This is the step where we calculate how likely is it to get data like that observed when $H_0 H_0$ is true. We use the **p-value** to assess the evidence. It is **the probability of observing a test statistic as extreme as (or even more extreme than) that observed assuming that the null hypothesis is true.**

p-value = The probability of observing a test statistic as extreme as (or even more extreme than) that observed assuming that the null hypothesis is true.

- **Step 4: Making conclusions**: Since our conclusion is based on how small the p-value is, it would be nice to have some kind of threshold or cutoff that will help determine how small the p-value must be, or how "rare" (unlikely) our data must be when $H_0 H_0$ is true, for us to conclude that we have enough evidence to reject $H_0 H_0$.

This cutoff has a special name. It is called the **significance level** of a test and is usually denoted by the Greek letter $\alpha \alpha$. The most commonly used significance level is $\alpha = 0.05 \alpha = 0.05$ (or 5%). We use the following decision rule:

- if the p-value < $\alpha \alpha$ (usually 0.05 or 5%), then the data we got is considered to be "rare (or surprising) enough" when $H_0 H_0$ is true, and we say that the data

provide significant evidence against H_0H_0, so we reject H_0H_0 and accept H_aH_a.

- if the p-value > $\alpha\alpha$ (usually 0.05 or 5%), then our data are not considered to be "surprising enough" when H_0H_0 is true, and we say that our data do not provide enough evidence to reject H_0H_0 (or, equivalently, that the data do not provide enough evidence to accept H_aH_a).

Linked to the concept of a *significance level* is the **confidence level** . A significance level of 0.05 or 5% corresponds with a 95% confidence level. The confidence level is associated with the confidence interval. For instance, you can construct a 95% confidence interval, where 95% refers to the confidence level. Just like p-values, confidence intervals can be used to do hypothesis testing. The decision rule for confidence intervals is as follows:

- If sample parameter falls outside the 95% confidence interval, reject H_0H_0 in favour of H_aH_a.
- If sample parameter falls inside the 95% confidence interval, do not reject H_0H_0.

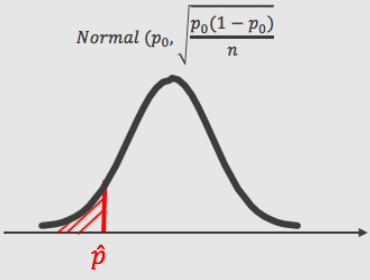
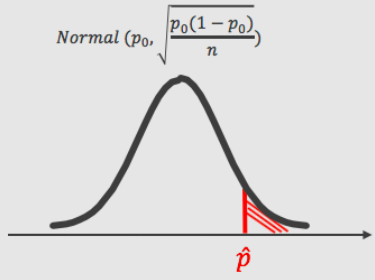
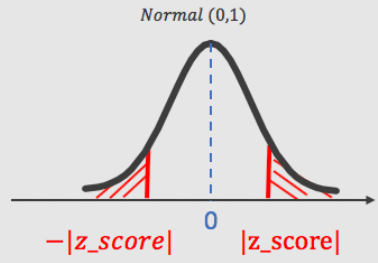
Note that the null hypothesis can never be accepted - you can only reject the null hypothesis in favour of the alternative hypothesis, or fail to reject the null hypothesis.

3. One-sample hypothesis testing

One Sample Hypothesis Testing
Population Proportion

$$p \sim Normal\ (mean = p_0,\quad sd = \sqrt{\frac{p_0 \times (1-p_0)}{n}})$$

$\hat{p}$ = Sample proportion ; p_0= Null hypothesis proportion ; n = Sample size

Case 1	Case 2	Case 3
$H0:\ p = p_0$ $Ha:\ p < p_0$	$H0:\ p = p_0$ $Ha:\ p > p_0$	$H0:\ p = p_0$ $Ha:\ p \neq p_0$
$p - value = P(p < \hat{p})$	$p - value = P(p > \hat{p}) = 1 - P(p < \hat{p})$	$z - score = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \times (1-p_0)}{n}}}$ $p - value = 2 \times P(p < -\lvert z - score\rvert)$
		

Compute p-value

Solution 1

$p - value = norm.cdf(\hat{p}, p_0, sd)$	$p - value = 1 - norm.cdf(\hat{p}, p_0, sd)$	$p - value = 2 * norm.cdf(-\lvert z - score\rvert)$

Solution 2

$p - value = proportions\_ztest\ (count = n * \hat{p}, nobs = n, value = p0, \text{prop\_var} = p0, alternative = 'smaller'\ or\ 'larger'\ or.'two - sided')$ [1]

- ✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
- ✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Proportions

Example

Our workers are known to produce 20% defective products, and are sent for retraining. After the training, 400 products produced are chosen at random and 64 are found to be defective (proportion $\hat{p} = \frac{64}{400} = 0.16\,\hat{p} = \frac{64}{400} = 0.16$). Do the data provide enough evidence that the proportion of defective products produced by our workers, pp has been reduced as a result of the training?

Based on our problem, we formulate the following hypotheses:

- $H_0 : p = 0.20\,H_0 : p = 0.20$ (No change; the training did not help, $p_0 = 0.20\,p_0 = 0.20$)
- $H_a : p < 0.20\,H_a : p < 0.20$ (The training was effective)

In [ ]:

```
x= 64
n=400
p_hat= x/n
p0=0.2
sd=(p0 * (1 - p0)/n)**.5

#Solution1
p_value = norm.cdf(p_hat, p0, sd)

alpha = .05
if p_value < alpha:
    print("\np_value = {}, Reject the null hypothesis, in fav
    that the proportion of defective products is less than 0.
    as a result of the training".format(round(p_value, 3)))
else:
    print("\np_value = {}, CANNOT Reject the null hypothesis.
    strong enough evidence to prove the proportion of defecti
```

In [ ]:

```
#Solution2
from statsmodels.stats.proportion import proportions_ztest

p_value = proportions_ztest(count=x , nobs=n, value=p0, prop_

alpha = .05
if p_value < alpha:
    print("\np_value = {}, Reject the null hypothesis, in fav
    that the proportion of defective products is less than 0.
    as a result of the training".format(round(p_value, 3)))
else:
    print("\np_value = {}, CANNOT Reject the null hypothesis.
    strong enough evidence to prove the proportion of defecti
```

Exercise

Polls on certain topics are conducted routinely to monitor changes in the public's opinions over time. One such topic is the death penalty. In 2013, a poll estimated that 91% of 1,535 Malaysian adults surveyed support the death penalty for people convicted of murder. In a more recent poll, 890 out of 1,000 Malaysian adults chosen at random were in favor of the death penalty for convicted murderers. Do the results of this poll provide evidence that the proportion of Malaysian adults who support the death penalty for convicted murderers (pp) changed between 2013 and the later poll?

In [ ]:

```
1
```

Means

We need to distinguish between two cases: where the population standard deviation ($\sigma\sigma$) is known, and the case where $\sigma\sigma$ is unknown.

- If $\sigma\sigma$ is **known**, the test is called the **z-test** for the population mean $\mu\mu$ because the sample mean follows the **normal** distribution $Normal(mean = \mu_0, std = \frac{\sigma}{\sqrt{n}})$ $Normal(mean = \mu_0, std = \frac{\sigma}{\sqrt{n}})$ where nn = sample size; $\mu_0\mu_0$ = population mean according to the null hypothesis, and $\sigma\sigma$ = population standard deviation. Therefore, the test statistic $z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ $z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$, which is the standardised sample mean, follows a **z-distribution**, or a standard normal distribution.

One Sample Hypothesis Testing
Mean; σ is KNOWN

$$\bar{x} \sim Normal\ (mean = \mu_0, \quad sd = \frac{\sigma}{\sqrt{n}})$$

$\bar{x}$ = Sample mean ; μ_0 = *Null hypothesis* mean ; σ = Population Standard Deviation; n = *Sample size*

Case 1	Case 2	Case 3
$H0$: mean = μ_0 Ha: mean < μ_0	$H0$: mean = μ_0 Ha: mean > μ_0	$H0$: mean = μ_0 Ha: mean ≠ μ_0
$p - value = P(x < \bar{x})$	$p - value = P(x > \bar{x}) = 1 - P(x < \bar{x})$	$z - score = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ $p - value = 2 \times P(x < -\lvert z - score \rvert)$
$p - value = norm.cdf(\bar{x}, \mu_0, sd)$	$p - value = 1 - norm.cdf(\bar{x}, \mu_0, sd)$	$p - value = 2 * norm.cdf(-\lvert z - score \rvert)$

- ✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
- ✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

The SAT, a standardised test for college admissions in the US, is constructed so that scores in each portion have a national average of 500 and standard deviation of 100. The distribution is close to normal. The Marketing department of your college believes that in recent years the college attracts students who are more math-inclined. A random sample of 15 students from a recent cohort at your College had an average math SAT (SAT-M) score of 550. Does this provide enough evidence for the dean to conclude that the mean SAT-M of all your college's students is higher than the national mean of 500? Assume that the standard deviation of 100 applies also to all students at your college.

Solution 1:

The sampling distribution of $\bar{x}\bar{x}$ under the null hypothesis is normal:

$$\bar{x} \sim Normal(mean = \mu_0, std = \frac{\sigma}{\sqrt{n}})$$

$$\bar{x} \sim Normal(mean = \mu_0, std = \frac{\sigma}{\sqrt{n}})$$

, where $\mu_0 = 500\mu_0 = 500$ and $\sigma\sigma$ = standard deviation of population

- H0: mean = 500
- Ha: mean > 500

In [ ]:

```
mu_0 = 500
sigma = 100
x_bar = 550
n = 15
sd = sigma/n**.5
p_value = 1 - norm.cdf(x_bar, mu_0, sd)

if p_value < alpha:
    print("p_value = {}, Reject the null hypothesis in favour
    mean of SAT-M of new students at your college is higher t
else:
    print("p_value = {}, CANNOT Reject the null hypothesis. T
    there is not strong enough evidence that the mean of SAT-
    students at your college is higher than the mean of SAT-M
```

Exercise

Human pregnancy is known to have a mean of 266 days and a standard deviation of 16 days. Based on records from a large hospital, a random sample of 30 women who were smoking and/or drinking alcohol during their pregnancy and their pregnancy lengths are recorded. We calculated the average pregnancy length of these women as 258.78. Do the data provide enough evidence to support the (well-known) fact that women who smoke and/or drink alcohol during their pregnancy have shorter pregnancies than women in general (in other words, are more likely to have premature labor)?

In [ ]:

```
1
```

- If $\sigma\sigma$ is **unknown**, the test is called the **t-test** for the population mean $\mu\mu$ because the standardised sample mean follows a **t-distribution**. In other words, the test statistic $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\, t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$ follows t-distribution $t \sim t(n-1)\, t \sim t(n-1)$

  where s is the standard deviation of the sample, and nn is the sample size.

One Sample Hypothesis Testing
Mean; σ is UNKNOWN

$\bar{x} \sim t(n-1)$

$\bar{x}$ = Sample mean ; s = Sample standatd deviation ; μ_0 = *Null hypothesis* mean ; n = *Sample size*

Case 1	Case 2	Case 3
$H0$: mean $= \mu_0$ Ha: mean $< \mu_0$	$H0$: mean $= \mu_0$ Ha: mean $> \mu_0$	$H0$: mean $= \mu_0$ Ha: mean $\neq \mu_0$
$p-value = P(x < \bar{x})$	$p-value = P(x > \bar{x}) = 1 - P(x < \bar{x})$	$p-value = 2 \times P(x < -\lvert t-score \rvert)$
Compute p-value		
Solution 1: Only $\bar{x}$ and s are given		
$t-score = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$		
$p-value = t.cdf(t-score, n-1)$	$p-value = 1 - t.cdf(t-score, n-1)$	$p-value = 2 * t.cdf(-\lvert t-score \rvert, n-1)$
Solution 2: SAMPLE_DATA is given		
Calculate $\bar{x}$ and s from the SAMPLE_DATA and use solution 1. OR $p-value$ = .5 × ttest_1samp(SAMPLE_DATA , μ_0).pvalue		$p-value =$ ttest_1samp(SAMPLE_DATA , μ_0).pvalue

✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95

✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

A certain prescription medicine is supposed to contain an average of 250 parts per million (ppm) of a certain chemical. If the concentration is higher than this, the drug may cause harmful side effects; if it is lower, the drug may be ineffective. The manufacturer wants to know whether the mean concentration in a large shipment conforms to the target level of 250 ppm.

A simple random sample of 100 portions is tested, and the sample mean concentration is found to be 246 ppm with a sample standard deviation of 12 ppm.

The hypotheses are:

- $H_0 : \mu = 250 H_0 : \mu = 250$
- $H_a : \mu \neq 250 H_a : \mu \neq 250$

In [ ]:

```
mu_0, n, x_bar, s = 250, 100, 246, 12
t_score = (x_bar - mu_0)/(s/n**.5)
p_value = 2 * t.cdf(-abs(t_score), df = n - 1)

if p_value < alpha:
    print("p_value = {}, Reject the null hypothesis in favour
    concentration does NOT conform to the target level of 250
else:
    print("p_value = {}, CANNOT Reject the null hypothesis. T
    evidence that the mean concentration does NOT conform to
    250 ppm".format(round(p_value, 3)))
```

Exercise

On average, a Finnish consumes 12kg of coffee in a year, which is 5 cups a day per person. A Finnish university wants to know whether their students tend to drink more coffee than the national average. They ask 50 students how many cups of coffee they drink each day and found their average number of drinks is $\bar{x} = 5.2 \bar{x} = 5.2$, with std dev $s = 1.5 s = 1.5$. Do they have enough evidence that their students drink more than the national average?

In [ ]:

```
1
```

Example

The mean of crude birth rate has been 16.7 per 1000 population in Malaysia in 2014. The following data shows crude birth rate from January to March 2019. Does the data prove a significant difference in 2019 comparing to 2014?

- H0: mu = 16.7
- Ha: mu != 16.7

In [ ]:

```
1 sample = pd.read_csv("../data/CBR.csv")
2 sample.head()
```

In [ ]:

```
1 sample_data = sample.CBR
2 sample_data[:5]
```

In [ ]:

```
print('\n***************** Solution1 *****************\n')

mu_0=16.7
x_bar = np.mean(sample_data)
print('x_bar = ', x_bar)
s = np.std(sample_data, ddof=1)
print('s = ', s)
n = len(sample_data)
print('n = ', n)
t_score = (x_bar - mu_0)/(s/(n**.5))
print('t_score = ', t_score)

p_value = 2*t.cdf(-abs(t_score), df = n-1)

if p_value < alpha:
    print("\np_value = {}, Reject the null hypothesis in favo
    of crude birth rate is different in 2019 comparing that i
else:
    print("\np_value = {}, CANNOT Reject the null hypothesis.
    evidence that the mean of crude birth rate is different i

print('\n***************** Solution2 *****************\n')


from scipy.stats import ttest_1samp
p_value=ttest_1samp(sample_data , mu_0).pvalue

if p_value < alpha:
    print("p_value = {}, Reject the null hypothesis in favour
    of crude birth rate is different in 2019 comparing that i
else:
    print("p_value = {}, CANNOT Reject the null hypothesis. T
    evidence that the mean of crude birth rate is different i

```

Exercise

In the above example, can we say crude birth rate has been decreased in 2019?

In [ ]:

```

```

4. Two-sample hypothesis test

In the previous sections we performed inference for one variable. If this variable was categorical, we perform one-sample hypothesis test for proportions. If the variable was numerical/quantitative, we perform one-sample hypothesis test for mean.

In this section, we look at inference about relationships between two variables in a population, based on an observed relationship between variables in a sample.

Assume we are interested in studying whether a relationship exists between the variables xx and yy in a population of interest. We choose a random sample and collect data on both variables from the subjects. Our goal is to determine whether these data provide strong enough evidence for us to generalize the observed relationship in the sample and conclude (with some acceptable and agreed-upon level of uncertainty) that a relationship between xx and yy exists in the entire population.

- H_0 :H_0: There is no relationship between xx and yy
- H_a :H_a: There is a significant relationship between xx and yy

C ---> Q

We consider hypothesis testing where xx, the explanatory variable, is a **categorical** variable and yy, the response variable, is a **quantitative** variable.

Example

To investigate this relationship between year in university and GPA, we can divide the population of the university students in Malaysia into 4 sub-populations. Within each of these four groups, we are interested in the GPA.

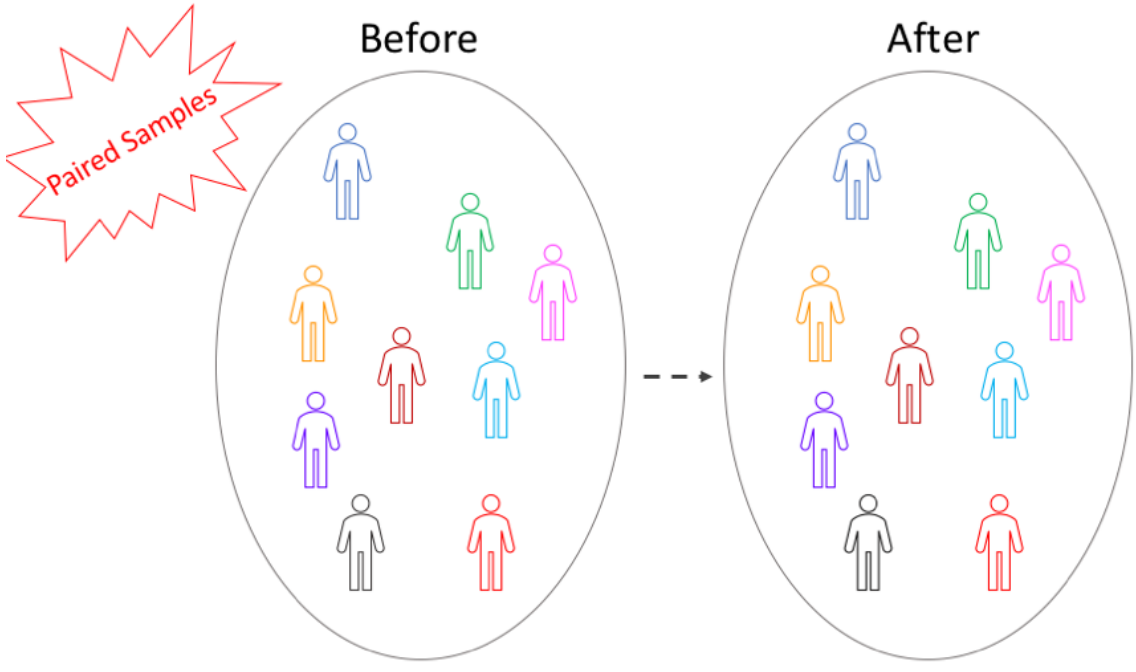
The inference must therefore involve the 4 sub-population means:

- μ_1 :μ_1: mean GPA among first year undergraduates in Malaysia
- μ_2 :μ_2: mean GPA among second year undergraduates in Malaysia
- μ_3 :μ_3: mean GPA among third year undergraduates in Malaysia
- μ_4 :μ_4: mean GPA among fourth year undergraduates in Malaysia

So, we need to compare these four means. If we infer that not all these four means are equal (i.e., that there are some differences in GPA across years in university) then that's equivalent to saying GPA is related to year in university.

Example

Assume xx is drinking/not_drinking alcohol, and yy is reaction time of the driver. We are interested to explore the impact of drinking two beers on the driver's reaction time. In this case, we measure the reaction time of 40 drivers, **before** and **after** drinking two beers.



Two-sample t-test of means for unpaired samples

Two-sample t-test of means for unpaired samples

Condition 1: The two samples are indeed independent
Condition 2: The distribution of y in both sub-populations is **NORMAL**, and both samples are random
Condition 3: The populations are **NOT NORMAL**, but the sample size of each of the random samples is **large enough ($n > 30$)**

$$t\_score \sim t(v)$$

$\overline{y_1}$ and $\overline{y_2}$ are sample means; s_1 and s_2 are sample standatd deviations; μ_1 *and* μ_2 *are population means*; n_1 and n_2 are *ample sizes*

Case 1		Case 2		Case 3	
$H0: \ \mu_1 = \mu_2$ $Ha: \ \mu_1 < \mu_2$	$H0: \ \mu_1 - \mu_2 = 0$ $Ha: \ \mu_1 - \mu_2 < 0$	$H0: \ \mu_1 = \mu_2$ $Ha: \ \mu_1 > \mu_2$	$H0: \ \mu_1 - \mu_2 = 0$ $Ha: \ \mu_1 - \mu_2 > 0$	$H0: \ \mu_1 = \mu_2$ $Ha: \ \mu_1 \neq \mu_2$	$H0: \ \mu_1 - \mu_2 = 0$ $Ha: \ \mu_1 - \mu_2 \neq 0$
Compute p-value					
Solution1: Only sample parameters are given $\overline{y_1}$, $\overline{y_2}$, s_1, s_2, n_1, n_2					
$t - score = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$; $v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2 (n_1 - 1)} + \frac{s_2^4}{n_2^2 (n_2 - 1)}}$					
$p - value = P(t - score < 0) =$ t. cdf(t − score, v)		$p - value = P(t - score > 0) =$ 1 - $P(t - score < 0)$ = 1 - t. cdf(t − score, v)		$p - value = 2 \times P(x < -\lvert t - score \rvert) =$ $2 \times$ t. cdf(−abs(t − score), v)	
Solution2: SAMPLE_DATA id given					
Calculate $\overline{y_1}$, $\overline{y_2}$, s_1, s_2 *and* n_1, n_2 from the SAMPLE_DATA and use solution 1. OR $p - value$ = .5 × ttest_ind(sample1, sample2, equal_var=False).pvalue				$p - value$ = ttest_ind(sample1, sample2, equal_var=False).pvalue	

✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

The two-sample t-test can be safely used as long as the following conditions are met:

1. **Both populations are normally distributed**, or more specifically, the distribution of yy in both sub-populations is normal, and both **samples are random** (or at least can be considered as such). In practice, checking normality in the sub-populations is done by looking at each of the samples using a histogram and checking whether there are any signs that the populations are not normal. Such signs could be extreme skewness and/or extreme outliers.
2. The populations are known or discovered not to be normal, but the **sample size of each of the random samples is large enough** (we can use the rule of thumb that $n > 30\, n > 30$ is considered large enough).

The two-sample t-test statistic is:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Where:

- $\bar{y}_1 \bar{y}_1$ and $\bar{y}_2 \bar{y}_2$ are the sample means of the samples from sub-population 1 and sub-population 2 respectively.
- $s_1 s_1$ and $s_2 s_2$ are the sample standard deviations of the samples from sub-population 1 and sub-population 2 respectively.
- $n_1 n_1$ and $n_2 n_2$ are the sample sizes of the two samples.

Attention: To understand the t-test statistic we need to know that

- $\bar{y}_1 \bar{y}_1$ estimates $\mu_1 \mu_1$ (mean of sub-population 1) and
- $\bar{y}_2 \bar{y}_2$ estimates $\mu_2 \mu_2$ (mean of sub-population 2).

Therefore, $\bar{y}_1 - \bar{y}_2 \bar{y}_1 - \bar{y}_2$ estimates $\mu_1 - \mu_2 \mu_1 - \mu_2$.

$\mu_1 - \mu_2 = 0 \mu_1 - \mu_2 = 0$ is the "null value" — what the null hypothesis, $H_0 H_0$, claims that $\mu_1 - \mu_2 \mu_1 - \mu_2$ is.

The denominator $\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}} \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$ is the standard deviation of $\bar{y}_1 - \bar{y}_2 \bar{y}_1 - \bar{y}_2$

We therefore see that our test statistic, like the previous test statistics we encountered, has the structure:

$$\frac{Sample\ Estimate - Null\ Value}{Standard\ Error}$$

$$\frac{Sample\ Estimate - Null\ Value}{Standard\ Error}$$

and therefore, like the previous test statistics, measures (in standard errors) the difference between what the data tell us about the parameter of interest $\mu_1 - \mu_2$ $\mu_1 - \mu_2$ (sample estimate) and what the null hypothesis claims the value of the parameter is (null value).

The number of degrees of freedom is $\nu \nu$ where:

$$\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

$$\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Degrees of freedom refers to the number of number of observations that are free to vary when calculating a statistic.

Example

Assume we are interested in investigating the relationship between a patient having a heart attack and the level of cholesterol. The variables we have are:

xx: patient had heart attack (yes/no) ---> Categorical

yy: patient cholesterol level (number) ---> Quantitative

We measured the cholesterol level of 38 heart attack patients (2 days after their attacks) and 40 other hospital patients who did not have a heart attack.

For the 38 heart attack patients, the mean cholesterol level was 253.9 with a standard deviation of 47.7. For the 40 other hospital patients who did not have a heart attack, the mean cholesterol level was 193.1 with a standard deviation of 22.3. Are cholesterol levels different across the different groups?

Answer:

- $H_0 : \mu_1 - \mu_2 = 0 H_0 : \mu_1 - \mu_2 = 0$
- $H_a : \mu_1 - \mu_2 \neq 0 H_a : \mu_1 - \mu_2 \neq 0$

$n_1 = 38 n_1 = 38$; $\bar{y}_1 = 253.9 \bar{y}_1 = 253.9$; $s_1 = 47.7 s_1 = 47.7$

$n_2 = 40 n_2 = 40$; $\bar{y}_2 = 193.1 \bar{y}_2 = 193.1$; $s_2 = 22.3 s_2 = 22.3$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} = \frac{253.9 - 193.1}{\sqrt{\frac{47.7^2}{38} + \frac{22.3^2}{40}}} = 7.150$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} = \frac{253.9 - 193.1}{\sqrt{\frac{47.7^2}{38} + \frac{22.3^2}{40}}} = 7.150$$

$$df = \frac{(\frac{47.7^2}{38} + \frac{22.3^2}{40})^2}{\frac{47.7^4}{38^2(38-1)} + \frac{22.3^4}{40^2(40-1)}} = 51.84$$

$$df = \frac{(\frac{47.7^2}{38} + \frac{22.3^2}{40})^2}{\frac{47.7^4}{38^2(38-1)} + \frac{22.3^4}{40^2(40-1)}} = 51.84$$

`p_value = 2*t.cdf(-abs(t), df)` = 2.89804375316508 54e-09

To make things easier, let's write a function `unpaired_t` that returns t and v:

In [ ]:

```
def unpaired_t(n_1, y_1, s_1, n_2, y_2, s_2):
    t_score = (y_1-y_2)/(((s_1**2/n_1)+(s_2**2/n_2))**.5)
    df=( (((s_1**2)/n_1)+((s_2**2)/n_2))**2 )/( (s_1**4)/((n_
    return(t_score, df)
```

In [ ]:

```
n_1, y_1, s_1 =38, 253.9, 47.7
n_2, y_2, s_2 =40, 193.1, 22.3

t_score, df = unpaired_t(n_1, y_1, s_1, n_2, y_2, s_2)
print('t_score = ', t_score)
print('df = ', df)


from scipy.stats import t
p_value = 2*t.cdf(- abs(t_score), df)

if p_value < alpha:
    print("p_value = {}, Reject the null hypothesis in favour
    a relationship between cholesterol level and heart attack
else:
    print("p_value = {}, CANNOT Reject the null hypothesis: w
    a relationship between cholesterol level and heart attack
```

Example

To check the claim that the pregnancy length of women who smoke during pregnancy is shorter, on average, than the pregnancy length of women who do not smoke, a random sample of 35 pregnant women who smoke and a random sample of 35 pregnant women who do not smoke were chosen and their pregnancy lengths were recorded.

xx: smoke (yes/no) ---> Categorical variable

yy: pregnancy length ---> Quantitative variable

Two methods can be used:

1. calculating tt and vv and then use the function `t.cdf`
2. using `scipy.stats.ttest_ind` function

The hypotheses are as follows:

- $H_0 : \mu_1 - \mu_2 = 0 H_0 : \mu_1 - \mu_2 = 0$ (There is no relationship between smoking and pregnancy length)
- $H_a : \mu_1 - \mu_2 < 0 H_a : \mu_1 - \mu_2 < 0$ (Pregnancy length of women who smoke is shorter than the pregnancy length of women who do not smoke)

In [ ]:

```
data = pd.read_csv('../data/pregnancy.csv')
data.head()
```

In [ ]:

```
data.isna().sum()
```

In [ ]:

```
sample1, sample2 = data.Smoke.dropna(), data.No_Smoke.dropna(

# Solution1:
print('\n***************** Solution1 *****************\n')



n_1, y_1, s_1 = len(sample1), np.mean(sample1), np.std(sample
n_2, y_2, s_2 = len(sample2), np.mean(sample2), np.std(sample

t_score, df = unpaired_t(n_1, y_1, s_1, n_2, y_2, s_2)
p_value = t.cdf(t_score, df)

if p_value < alpha:
    print('p_value = {}, Reject the null hypothesis in favour
      Pregnancy length of women who smoke is shorter than the
      of women who do not smoke'.format(p_value))
else:
    print('p_value = {}, CANNOT  Reject the null hypothesis:
    for a significant relationship between pregnancy length a

# Solution2:
print('\n***************** Solution2 *****************\n')


from scipy.stats import ttest_ind
p_value = .5 *  ttest_ind(sample1, sample2, equal_var=False)[

if p_value < alpha:
    print('p_value = {}, Reject the null hypothesis in favour
      Pregnancy length of women who smoke is shorter than the
      of women who do not smoke'.format(p_value))
else:
    print('p_value = {}, CANNOT  Reject the null hypothesis:
    for a significant relationship between pregnancy length a
```

Using `scipy.stats.ttest_ind` we can directly test two independent samples without calculating the means, t-statistic and degrees of freedom:

Exercise

A researcher wanted to study whether men and women watch different amounts of YouTube. A random sample of 400 adults was chosen, comprising of 191 women and 209 men. At the end of the week, each of the 400 subjects reported the total amount of time (in minutes) that he or she watched YouTube during that week.

In [ ]:

```
tv = pd.read_csv('../data/tv.csv')
tv.head()
```

In [ ]:

```

```

Two-sample t-test for paired samples means

Example

Drunk driving is one the main causes of car accidents. We want to know whether drivers are impaired after drinking two beers. A sample of 30 drivers was chosen, and their reaction times in an obstacle course were measured before and after drinking two beers. The variables we have are:

xx: Drinking alcohol (yes/no) ---> Categorical

yy: Reaction time of the driver ---> Quantitative

Let $\mu_1\mu_1$ be the average of the reaction time before drinking 2 beers, and $\mu_2\mu_2$ be the average of the reaction time before drinking 2 beers.

https://www.youtube.com/watch?time_continue=1&v=URPrSH0Lg_M (https://www.youtube.com/watch?time_continue=1&v=URPrSH0Lg_M)

We reduce the two samples to only one by calculating for each pair the difference between the two observations

Pairs	1	2	3	4	...	n
Sample 1	a1	a2	a3	a4		an
Sample 2	b1	b2	b3	b4		bn
Differences	a1-b1	a2-b2	a3-b3	a4-b4		an-bn

Two sample paired test → One sample t-test

- $H_0 : \mu_d = 0 H_0 : \mu_d = 0$ OR $\mu_1 - \mu_2 = 0 \mu_1 - \mu_2 = 0$
- $H_a : \mu_d < 0 H_a : \mu_d < 0$ OR $\mu_1 - \mu_2 < 0 \mu_1 - \mu_2 < 0$ If the driver is drunk it takes more time to react $\mu_1 < \mu_2 \mu_1 < \mu_2$.

Assume

- mean of differences in the sample is -0.501,
- standard deviation of differences is 0.868.

and therefore

- $\bar{x}_d = -0.501 \bar{x}_d = -0.501,$
- $s_d = 0.868 s_d = 0.868,$
- $n = 30 n = 30,$
- $\mu_0 = 0 \mu_0 = 0$

$$t = \frac{\bar{x}_d - \mu_0}{\frac{s_d}{\sqrt{n}}} = \frac{-0.501}{\frac{0.868}{\sqrt{30}}}$$

$$t = \frac{\bar{x}_d - \mu_0}{\frac{s_d}{\sqrt{n}}} = \frac{-0.501}{\frac{0.868}{\sqrt{30}}}$$

In [ ]:

```
n, x_d_bar, s_d, mu_0 =30, -0.501, 0.868, 0
t_stat=(x_d_bar-mu_0)/(s_d/n**.5)
t_stat
df = n - 1
p_value = t.cdf(t_stat, df)

if p_value < alpha:
    print("p_value = {}, reject the null hypothesis \
    in favour of the alternative that reaction time increases
else:
    print("p_value = {}, Cannot reject the null: not enough e
    between drinking alcohol (2 beers) and reaction time of t
```

Example

Suppose we want to evaluate the effectiveness of this course on the statistics skills of the students. Assume there are 100 students in a cohort. We record their scores in a sample statistics test **before** and **after** passing this course. Did our students' statistics skills improve after taking the class?

Define $\mu_d = \mu_2 - \mu_1 \mu_d = \mu_2 - \mu_1$

- $H_0 : \mu_d = 0 H_0 : \mu_d = 0$
- $H_a : \mu_d > 0 H_a : \mu_d > 0$

In [ ]:

```
score = pd.read_csv('../data/StatScore.csv')
score.head(7)
```

In [ ]:

```
score_diff = score.Score2 - score.Score1
np.mean(score_diff)
```

There are two ways to run a paired t-test:

- treat it as a one-sample t-test where we test the differences between the two samples. Naturally, the null hypothesis is that the difference between the two samples is 0. Use `scipy.stats.ttest_1samp`
- pass the two samples directly to `scipy.stats.ttest_rel`

In [ ]:

```
from scipy.stats import ttest_1samp
ttest_result = ttest_1samp(score_diff, 0)

# for a symmetric distribution, the p-value of a one-tailed t
# the p-value for a two-tailed test

p_value = ttest_result[1]/2
alpha = .05

if p_value < alpha:
    print('p_value = {}.  Reject the null hypothesis: Therefo
    Data Analysis improved students skills'.format(round(p_va
else:
    print('p_value = {}, CANNOT Reject the null hypothesis: w
    our training in Statistical Data Analysis improved studen
```

In [ ]:

```
from scipy.stats import ttest_rel
# note: greater than not implemented for scipy. Two-sided tes
_, p = ttest_rel(score.Score2, score.Score1)
p_value = p/2

if p_value < alpha:
    print('p_value = {}.  Reject the null hypothesis: Therefo
    Data Analysis improved students skills'.format(round(p_va
else:
    print('p_value = {}, CANNOT Reject the null hypothesis: w
    our training in Statistical Data Analysis improved studen
```

5. Hypothesis testing for more than two samples

C ---> Q

Comparing More Than Two Means—ANOVA

So far, we have discussed the two samples and matched pairs designs, in which the categorical explanatory variable has two values. In these cases, examining the relationship between the explanatory and the response variables amounts to comparing the mean of the response variable yy in two populations, defined by the two values of the explanatory variable xx. The difference between the two samples and matched pairs designs is that in the former, the two samples are independent (not paired), and in the latter, the samples are dependent (paired).

We are now moving on to cases in which the categorical explanatory variable takes more than two values.

ANOVA

X: categorical variable with k categories $x_1, x_2, \ldots, x_k$

Y: numerical (quantitative) variable

sample_i = {y | where $X = x_i$}, for i = 1, 2, ..., k

x	x_i	x_i	x_i	...	x_i
y	y_{i1}	y_{i2}	y_{i3}	...	y_{iN}

← sample_i

$\mu_i = mean\ of\ y\ in\ the\ \mathbf{population}\ where\ x == i$

Question:

Is there any significant relationship between x and y in the population? Does x impact y? Is there significant difference between μ_i's?

$$H0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

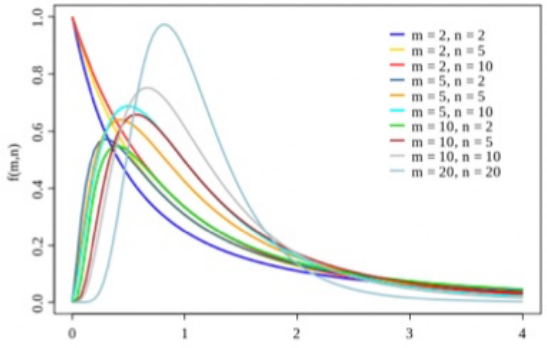
$$Ha: not\ all\ \mu's\ are\ equal$$

$$f\_stat = \frac{Variation\ Between\ Sample\ Means}{\text{Variation Within Groups}}$$

$$p\_value = P(f > f\_stat)$$

where f has f-distribution f(df_between, df_within)

$$df\_between = k - 1 \quad \& \quad df\_within = N\text{-}k \quad where \quad N = sample\ size \quad \& \quad k = number\ of\ categories\ in\ X$$



Compute p-value

f_stat, p_value = f_oneway(sample_1, sample_2, ..., sample_k)

- ✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
- ✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

A drug company tested three types of pain relief medication for migraines. For the experiment, 27 volunteers were selected and 9 were randomly assigned to each of the three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 to 10 (10 being most pain).

Groups	Collected data
Drug A	4 3 4 4 4 5 4 3 2
Drug B	4 6 5 8 6 6 8 4 5
Drug C	7 5 6 5 5 6 7 6 6

The hypotheses can be stated as follows:

- $H_0 : \mu_A = \mu_B = \mu_C$ $H_0 : \mu_A = \mu_B = \mu_C$
- H_a :H_a: not all the $\mu\mu$s are equal

In [ ]:

```
migraine = pd.DataFrame({
    'pain': [4, 3, 4, 4, 4, 5, 4, 3, 2, 4, 6, 5, 8, 6, 6, 8,
    'drug': np.repeat(["A", "B", "C"], 9)
})

migraine.head()
```

In [ ]:

```
pain_A = migraine.loc[migraine.drug == "A", 'pain']
pain_B = migraine.loc[migraine.drug == "B", 'pain']
pain_C = migraine.loc[migraine.drug == "C", 'pain']
pain_C
```

There are two ways to compute ANOVA in Python. Firstly, we can use `scipy.stats.f_oneway` , which returns the f-statistic and the p-value:

In [ ]:

```
#Solution 1:
from scipy.stats import f_oneway
f_stat, p_value = f_oneway(pain_A, pain_B, pain_C)
print('f_stat = ', f_stat)
print('p_value = ', p_value)
```

The statistic we are concerned about is the F-statistic:

$$f_{k-1,n-k} \sim \frac{Variation\ Between\ Sample\ Means}{Variation\ Within\ Groups}$$

$$f_{k-1,n-k} \sim \frac{Variation\ Between\ Sample\ Means}{Variation\ Within\ Groups}$$

Under the null hypothesis that there's no difference in group means, the F-statistic is expected to be around 1. Contrast this with our F-statistic 11.91! Should we reject our hypothesis that there's no difference in group means?

Note that you can also compute the F-statistic by taking the ratio of `mean_sq` (mean square error) of `drug` - the 'between-group variation' - to the `mean_sq` of `Residual` - the 'within groups' variation.

Together with the degrees of freedom, this gives us a p-value of 0.0003. So, we clearly reject the null hypothesis of equal means for all three drug groups.

The F-distribution has two degrees of freedom parameters, $k-1$ $k-1$ and $N-k$ $N-k$, where N N is the sample size and k k the number of groups. Knowing this, we can recompute the p-value based on the f-statistic:

In [ ]:

```
from scipy.stats import f
N, k = 27, 3
df_between, df_within = k-1, N -k
p_value = 1 - f.cdf(f_stat, dfn = df_between, dfd = df_within
print('p_value = ', p_value)
```

Draw a boxplot based on the results.

In [ ]:

```
sns.boxplot('drug', 'pain', data = migraine)
```

Multiple testing

Knowing that the means of A, B and C are not equal, we might want to know which pairs of drugs have different levels of pain. You could do a series of pairwise t-tests, i.e.

In [ ]:

```
p_AB = ttest_ind(pain_A, pain_B, equal_var=False)[1]
p_BC = ttest_ind(pain_B, pain_C, equal_var=False)[1]
p_AC = ttest_ind(pain_A, pain_C, equal_var=False)[1]
```

In [ ]:

```
ttest_ind(pain_A, pain_B)
```

However, we have to correct for multiple testing. The more tests you run, the more likely you'll incorrectly find a significant result in any one pair of results by chance alone. The Bonferroni correction simply sets the significance threshold to be

$$\alpha/m$$

$$\alpha/m$$

where mm is the number of hypotheses tested, and $\alpha\alpha$ is the level of significance.

In our case, 3 hypotheses are being tested, so divide alpha by 3 to obtain 0.167. Compare this against our array of p-values:

In [ ]:

```
np.array([p_AB, p_BC, p_AC]), np.array([p_AB, p_BC, p_AC]) <
```

which shows a significant difference between groups A and B and groups A and C but not groups A and C.

`statsmodels.sandbox.stats.multicomp.multipletests` is a convenient wrapper function for this procedure. It returns four objects, but we'll focus on three:

In [ ]:

```
from statsmodels.sandbox.stats.multicomp import multipletests
reject, pvals_corrected, _, alphaBonferroni  = multipletests(
```

The first object returned, `reject`, provides the result of the multiple testing without the p-values:

In [ ]:

```
reject
```

Should you want to compare the p-values against the original level of alpha, it is equivalent to correct the p-values by multiplying the p-values by the number of hypotheses being tested:

In [ ]:

```
pvals_corrected
```

`multipletests` also returns the Bonferroni-corrected level of alpha:

In [ ]:

```
np.array([p_AB, p_BC, p_AC]), np.array([p_AB, p_BC, p_AC]) <
```

C ---> C

The last three procedures that we studied (two-sampled t, paired t, and ANOVA) all involve the relationship between a categorical explanatory variable, x, and a quantitative response variable, y. Next, we will consider inferences about the relationship between two categorical variables.

Chi-square test for equality of proportions in two samples

For the test of proportion to be valid, we generally need the following:

- For a right- or left-tailed test, a minimum of 10 successes and 10 failures in each group are necessary.
- Two-tailed tests are more robust and require only a minimum of 5 successes and 5 failures in each group

Chi-Square

Condition: $a_{ij} > 5 \; for \; i = 1, 2, \dots, s \quad \& \quad j = 1, 2, \dots, k$

Question: Is there any significant relationship between X and Y in the population?

X: categorical variable with k categories $x_1, x_2, \dots, x_k$

Y: categorical variable with k categories $y_1, y_2, \dots, y_s$

Observed_Counts

X / Y	x_1	x_2	...	x_k
y_1	a_{11}	a_{12}	...	a_{1k}
y_2	a_{21}	a_{22}	...	a_{2k}
...				
y_s	a_{s1}	a_{s2}	...	a_{sk}

$$a'_{mn} = P(Y = y_m) \times P(X = x_n) \times \sum_{i,j=1}^{s,k} a_{ij} =$$

$$\frac{\sum_{i=1}^{s} a_{i1} \times \sum_{j=1}^{k} a_{1j}}{\sum_{i,j=1}^{s,k} a_{ij}}$$

Expected_Counts

X / Y	x_1	x_2	...	x_k
y_1	a'_{11}	a'_{12}	...	a'_{1k}
y_2	a'_{21}	a'_{22}	...	a'_{2k}
...				
y_s	a'_{s1}	a'_{s2}	...	a'_{sk}

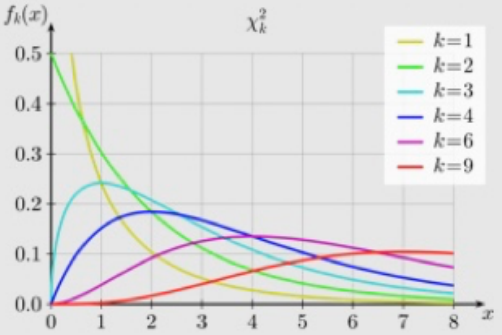
H0: X and Y are independent

Ha: X and Y are dependent

$$\chi^2 = \sum_{all\_cells} \frac{(ObservedCount - ExpectedCount)^2}{ExpectedCount}$$

$$p\_value = P(c > \chi^2)$$

where c has chi − square distribution, chi-square(df = (k-1) (s-1))



Compute p-value

Solution1:

```
chi2_stat = np.sum((((Observed_Counts - Expected_Counts)**2)/exp))
p_value = 1 - chi2.cdf(chi2_stat, df)
```

Solution2:

```
chi2_stat , p_value, df, Expected_Counts = chi2_contingency(Observed_Counts , correction=False)
```

- ✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
- ✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

A researcher wants to know if there's a relationship between gender and drunk-driving. She samples a total of 619 drivers under 20 years of age in a roadside survey.

xx: Driver gender (Male/Female)

yy: Driver alcohol (Yes/No)

Step 1: Stating the hypotheses

- H_0 :H_0: There is no relationship between the two categorical variables. (They are independent.)
- H_a :H_a: There is a relationship between the two categorical variables. (They are not independent.)

Step 2: Checking the Conditions and Calculating the Test Statistic

	Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

	Yes	No	Total
Male	$\frac{77}{481} = 16\%$	$\frac{404}{481} = 84\%$	100%
Female	$\frac{16}{138} = 11.6\%$	$\frac{112}{138} = 88.4\%$	100%

For the 619 sampled drivers, a larger percentage of males were found to be drunk than females (16.0% vs. 11.6%). Our data, in other words, provide some evidence that drunk driving is related to gender; however, this in itself is not enough to conclude that such a relationship exists in the larger population of drivers under 20. We need to further investigate the data and decide between the following two positions:

- The evidence provided by the roadside survey (16% vs 11.6%) is strong enough to conclude (beyond a reasonable doubt) that it must be due to a relationship between drunk driving and gender in the population of drivers under 20.
- The evidence provided by the roadside survey (16% vs. 11.6%) is not strong enough to make that conclusion, and could have happened just by chance, due to sampling variability, and not necessarily because a relationship exists in the population.

These two different conclusions can be condensed into the two hypotheses below:

- H_0 :H_0: Drunk driving and gender are independent
- H_a :H_a: Drunk driving and gender are not independent

Algebraically, independence between gender and driving drunk is equivalent to having equal proportions who drank (or did not drink) for males vs. females. In fact, the null and alternative hypotheses could have been re-formulated as

- H_0 :H_0: proportion of male drunk drivers = proportion of female drunk drivers

- H_a :H_a: proportion of male drunk drivers ≠ proportion of female drunk drivers

Applying the rule to the first (top left) cell, if driving drunk and gender were independent then:

$$P(drunk\ and\ male) = P(drunk) * P(male) \quad P(drunk\,and\,male) = P(drunk) * P(male)$$

$$P(drunk) = 93/619 \quad P(drunk) = 93/619$$

$$P(male) = 481/619 \quad P(male) = 481/619$$

$$P(drunk\ and\ male) = (93/619) * (481/619)$$
$$P(drunk\,and\,male) = (93/619) * (481/619)$$

Therefore, since there are total of 619 drivers, if drunk driving and gender were independent, the count of drunk male drivers that I would expect to see is:

Number of drunk Men = 619* P(drunk and male) = $619 * \frac{93}{619} * \frac{481}{619}$ $619 * \frac{93}{619} * \frac{481}{619}$

Similarly:

Number of drunk Women = 619* P(drunk and female) = $619 * \frac{93}{619} * \frac{138}{619}$
$619 * \frac{93}{619} * \frac{138}{619}$

Observed Counts

	Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Expected Counts

	Yes	No	Total
Male	$\frac{93 * 481}{619} = 72.3$	$\frac{526 * 481}{619} = 408.7$	481
Female	$\frac{93 * 138}{619} = 20.7$	$\frac{526 * 138}{619} = 117.3$	138
Total	93	526	619

$$\chi^2 = \sum_{all\_cells} \frac{(ObservedCount - ExpectedCount)^2}{ExpectedCount}$$

p-value = The probability of observing χ2 at least as large as the one observed

$$\chi^2 = \Sigma_{all\_cells} \frac{(Observed\ Count - Expected\ Count)^2}{Expected\ Count}$$

$$\chi^2 = \Sigma_{all\_cells} \frac{(Observed\ Count - Expected\ Count)^2}{Expected\ Count}$$

The p-value obtained can be interpreted as the probability of observing a $\chi^2 \chi^2$ test statistic at least as large as the one observed if drunk driving and gender are independent.

Step 3: Given two categorical variables xx and yy, the p-value can be found as:

```
1 - chi2.cdf(chi2_stat, df)
```

where `chi2_stat` is the $\chi^2 \chi^2$ test statistic, and $df = (n_A - 1)(n_B - 1)$ $df = (n_A - 1)(n_B - 1)$ where `n_A` is the number of categories in xx and `n_B` is the number of categories in yy.

Example

An ice cream shop wants to know whether men and women have different preferences for eating their ice cream out of a cone or a bowl. They take a sample of 500 customers (240 men and 260 women) and ask if they prefer cones over bowls. They found that 124 men preferred cones and 90 women preferred cones. Is there a difference in preference between men and women?

In [ ]:

```
observed = pd.DataFrame({
    'IceCream': np.repeat(['cones', 'bowl', 'cones', 'bowl'],
    'Gender': np.repeat(['male', 'female'], repeats = [240, 2
})
```

`pd.crosstab` returns a contingency table:

In [ ]:

```
cont_table = pd.crosstab(observed.Gender, observed.IceCream)
cont_table
```

Let us work through this slowly to understand the concept of expected counts. Recall that the table of expected counts is what you would expect in each cell of the contingency table if each of the categorical variables of interest were independent, i.e.

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B)$$

If you were to get the number of events where $A \cap B$, multiply the number of events A and B and divide by the total number of events. (Multiply both sides of the equation by the number of elements in the contingency table, and this should become clear.)

In [ ]:

```
icecream_probs = observed.IceCream.value_counts()
gender_probs = observed.Gender.value_counts()

names = [(i, j) for j in gender_probs.index for i in icecream
exp_list = [(i * j)/500 for j in gender_probs for i in icecre

list(zip(names, exp_list))
```

This process is tedious. Fortunately, `scipy.stats.contingency` has an `expected_freq` function that simplifies this:

In [ ]:

```
obs = cont_table.values
from scipy.stats.contingency import expected_freq
exp = expected_freq(cont_table)
exp
```

Calculate the chi-square test statistic:

$$\chi^2 = \Sigma_{all\_cells} \frac{(Observed\ Count - Expected\ Count)^2}{Expected\ Count}$$

In [ ]:

```
chi2_stat = np.sum(((obs-exp)**2)/exp))
```

Notice that the chi-square test statistic quantifies how far away the observed counts are from the expected counts. This makes it similar to the 'greater than' (Case 2) hypotheses tests, and therefore makes it a one-tailed test. The chi-square test has $(r-1)(c-1)$ degrees of freedom.

In [ ]:

```
r, c = 2, 2
df = (r-1)*(c-1)
```

The p-value, under the null of independence, is calculated as

In [ ]:

```
from scipy.stats import chi2
1 - chi2.cdf(chi2_stat, df)
```

Everything we've done previously can be done in one step on our contingency table using the `scipy.stats.chi2_contingency` function.

In [ ]:

```
from scipy.stats import chi2_contingency
stat, p, dof, expected = chi2_contingency(cont_table, correct
p
```

In [ ]:

```
dof
```

In [ ]:

```
expected
```

In [ ]:

```
stat
```

Example

Risk Factors for Low Birth Weight

Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for babies with low birth weight. A woman's behavior during pregnancy (including diet, smoking habits, and obtaining prenatal care) can greatly alter her chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

In this exercise, we will use a 1986 study (Hosmer and Lemeshow (2000), Applied Logistic Regression: Second Edition) in which data were collected from 189 women (of whom 59 had low birth weight infants) at the Baystate Medical Center in Springfield, MA (an academic, research, and teaching hospital that serves as the western campus of Tufts University School of Medicine and is the only Level 1 trauma center in western Massachusetts). The goal of the study was to identify risk factors associated with giving birth to a low birth weight baby.

Variables:

- LOW: Low birth weight (0=No (birth weight >= 2500 g) 1=Yes (birth weight < 2500 g)
- AGE: Age of mother (in years)
- LWT: Weight of mother (in pounds)
- RACE: Race of mother (1=White, 2=Black, 3=Other)
- SMOKE: Smoking status during pregnancy (0=No, 1=Yes)
- PTL: History of premature labor (0=None, 1=One, etc.)
- HT: History of hypertension (0=No, 1=Yes)
- FTV: Number of physician visits during the first trimester
- BWT: The actual birth weight (in grams)

Question:

- Q1. Do the data provide evidence that the occurrence of low birth weight is significantly related to whether or not the mother smoked during pregnancy?

In [ ]:

```
1 Birth = pd.read_csv('../data/low_birth_weight.csv')
2 Birth.head()
```

In [ ]:

```
#Question1:
cont_table1 = pd.crosstab(Birth.LOW, Birth.SMOKE)
cont_table1
```

In [ ]:

```
stat1, p_value1, dof1, expected1 = chi2_contingency(cont_tabl
```

In [ ]:

```
p_value1
```

In [ ]:

```
stat1
```

In [ ]:

```
if p_value1 < alpha:
    print('p_value = {}.  There IS significant relationship b
else:
    print('p_value = {}, There IS NOT relationship between lo
```

Exercise

Answer Questions 2-4:

- Q2. Do the results of the study provide significant evidence that the race of the mother is a factor in the occurrence of low birth weight?
- Q3. Are there significant differences in age between mothers who gave birth to low weight babies and those whose baby's weight was normal?
- Q4. Are there significant relationship between the actual birth weight and the race of the mother?

In [ ]:

```

```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

In [ ]:

```
1
```

Extra

Recreating ANOVA table from scratch

In [ ]:

```
# add a column with group means of pain
migraine['group_means'] = migraine.groupby('drug').transform(
N, k = 27, 3
df_between, df_within = k-1, N-k

# sum of squares **within** treatment groups
# and total sum of squares
sum_sq_within = sum((migraine.pain - migraine.group_means)**2
sum_sq_total = sum((migraine.pain - np.mean(migraine.pain))**

# the between sum of squares is just the diff betweeen total
sum_sq_between = sum_sq_total - sum_sq_within

mean_sq_between, mean_sq_within = sum_sq_between/df_between,

F_stat = mean_sq_between/mean_sq_within

# right-tailed hypothesis test - are the means within each of
# same as the population mean? > 1 if different.
p_value = 1 - f.cdf(F_stat, dfn = df_between, dfd = df_within

# sum_sq_between - the between-group variation - is the 'expl
# this metric is known as the r-squared
sum_sq_between / sum_sq_total

1 - sum_sq_within/sum_sq_total

pd.DataFrame([[df_between, sum_sq_between, mean_sq_between, F
              [df_within, sum_sq_within, mean_sq_within, np.r
```

Chi-Square

Alternately, you can use the observed and expected values to run the chi-square test of independence using the `chisquare` function from `scipy.stats`.

In [ ]:

```
from scipy.stats import chisquare
_, p = chisquare(obs.ravel(), exp.ravel(), ddof=sum(obs.shape
p
```

The `ravel()` method is needed because without it, `chisquare` will calculate the chi-square statistic for each column.

`scipy.stats.chisquare` takes a **delta** degrees of freedom (ddof) parameter. This is a bit tricky to characterise. Recall that the degrees of freedom of the test of indepdence is

$$df = (r-1)(c-1)$$

where r is the number of rows, and c is the number of columns.

`chisquare` uses a chi-square distribution with $k - 1 - ddof$ degrees of freedom, where $k = rc$, the number of frequencies observed. With a bit of algebraic manipulation, we obtain ddof as $r + c - 2$.

$$(r-1)(c-1) = rc - r - c + 1 = rc - (r + c - 2) - 2 + 1 = rc - 1 - (r + c - 2$$

`scipy.stats.chisquare` returns the chi-square statistic and the p-value.

The Center of Applied Data Science

python™

Statistical Data Analysis

Day 2.2

Content Outline

In [ ]:

1. t-distribution

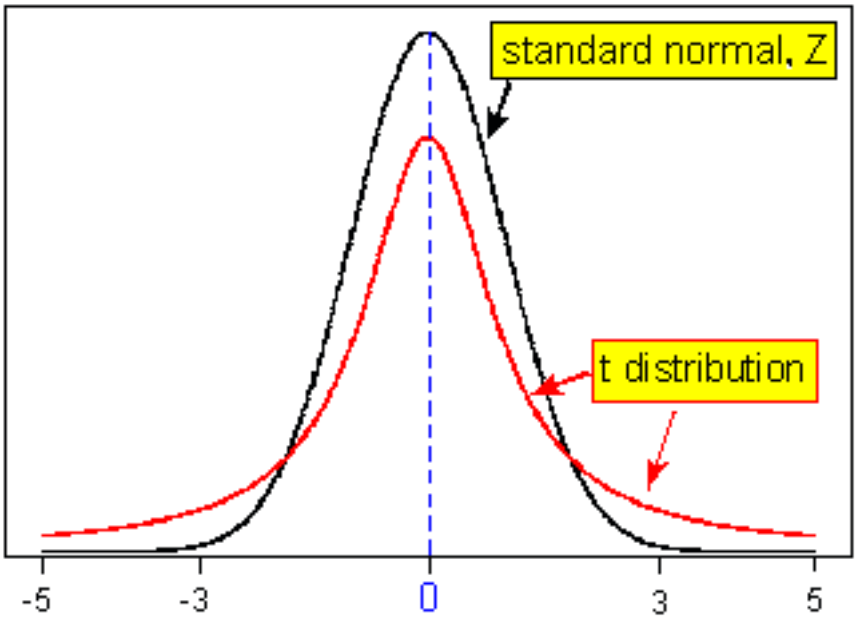
We have seen that random variables can be visually modeled by many different sorts of shapes, and we call these shapes distributions. Several distributions arise so frequently that they have been given special names, and they have been studied mathematically.

The **t-distribution** is another bell-shaped (unimodal and symmetric) distribution, like the normal distribution; and the center of the t-distribution is standardized at zero, like the center of the standard normal distribution which we call it z-distribution as well.

Like all distributions that are used as probability models, the normal and the t-distribution are both scaled, so the total area under each of them is 1.

So how is the t-distribution fundamentally different from the normal distribution? The **spread**.

The following picture illustrates the fundamental difference between the normal distribution and the t-distribution:



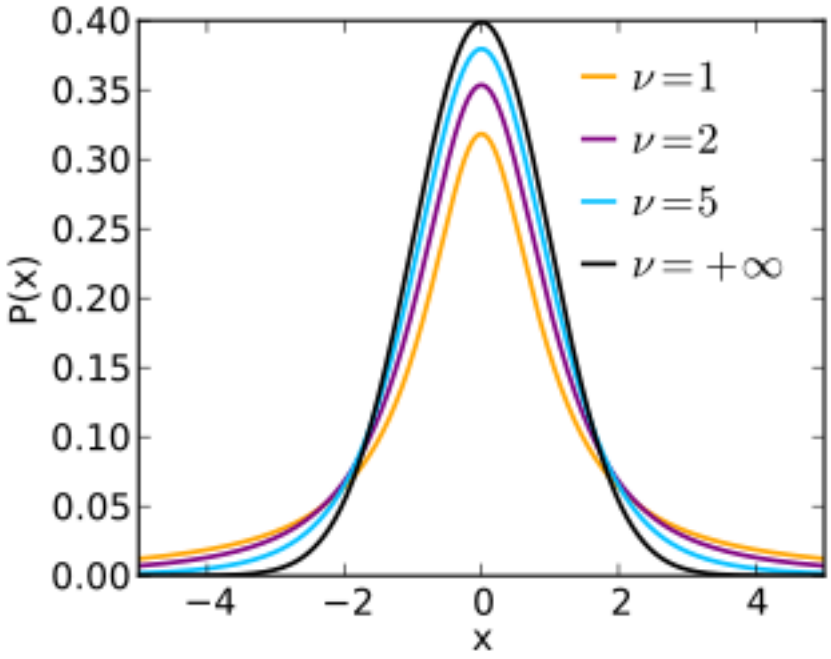
The t-distribution has slightly less area near the expected central value than the normal distribution does, and has correspondingly more area in the "tails" than the normal distribution does. (It's often said that the t-distribution has "fatter tails" or "heavier tails" than the normal distribution.)

This reflects the fact that the t-distribution has a larger spread than the normal distribution. The same total area of 1 is spread out over a slightly wider range on the t-distribution, making it a bit lower near the center compared to the normal distribution, and giving the t-distribution slightly more probability in the 'tails' compared to the normal distribution.

Therefore, the t-distribution ends up being the appropriate model in certain cases where there is more variability than would be predicted by the normal distribution.

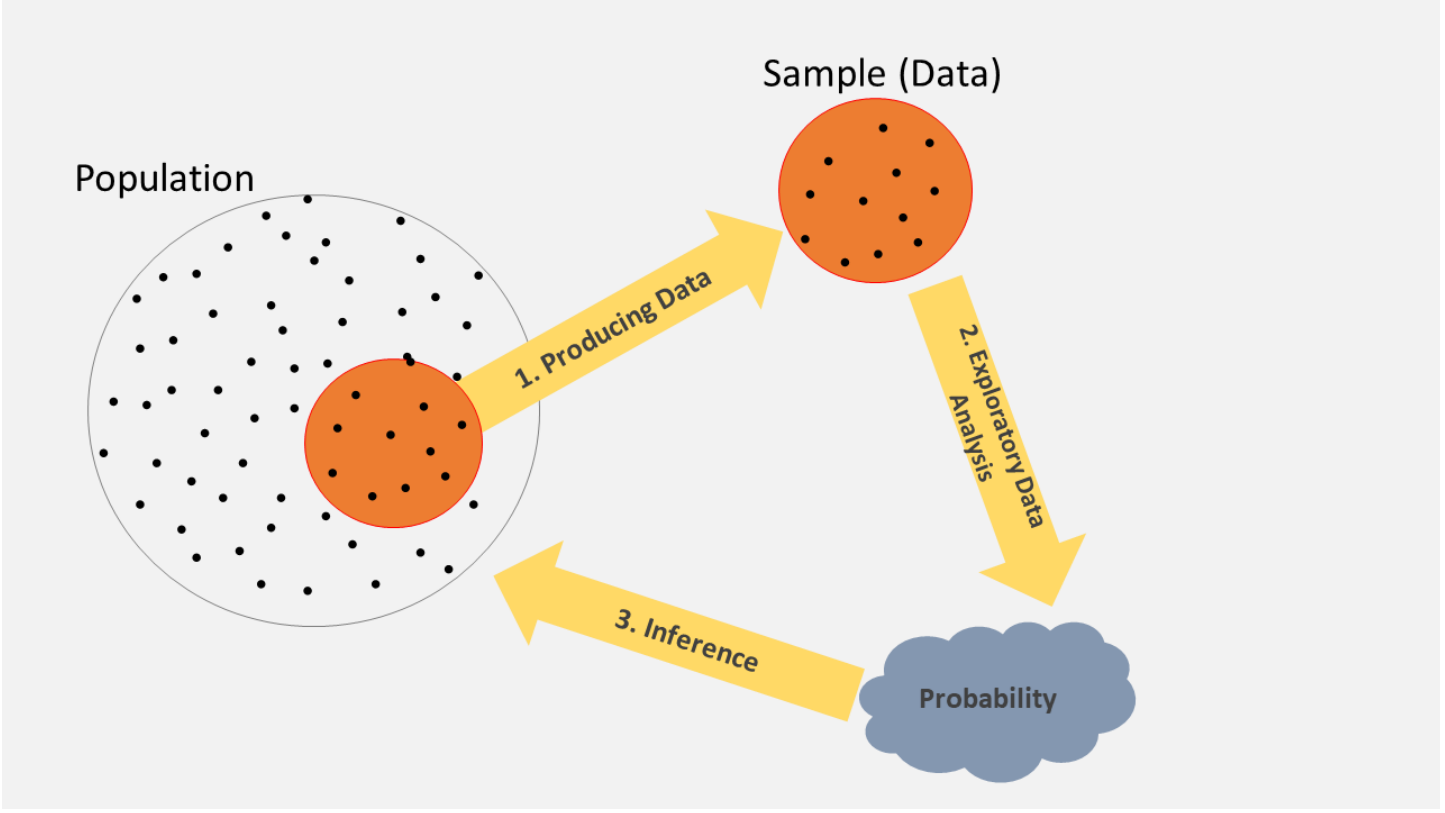
There's actually an entire family of t-distributions. They all have similar formulas (but the math is beyond the scope of this introductory course in statistics), and they all have slightly "fatter tails" than the normal distribution. But some are closer to normal than others. The t-distributions that are closer to normal are said to have higher "degrees of freedom" (that's a mathematical concept that we won't use in this course, beyond merely mentioning it here). So, there's a t-distribution "with one degree of freedom," another t-distribution "with 2 degrees of freedom" which is slightly closer to normal, another t-distribution "with 3 degrees of freedom." which is a bit closer to normal than the previous ones, and so on.

The following picture illustrates this idea with a few t-distributions (note that "degrees of freedom" is shown by \nu):



2. Inference

Our ultimate goal in statistical data analysis is using a sample to make inferences or draw conclusions about the population from which it was drawn.



Our choice of the type of inference depends on the type of the variable of interest. We introduce three forms of statistical inference in this unit, each one representing a different way of using the information obtained in the sample to draw conclusions about the population. These forms are:

- Point estimation
- Interval estimation
- Hypothesis testing

A. Point Estimation

We estimate an unknown parameter using a **single number** that is calculated from the sample data.

Example

Based on sample results, we estimate that p, the proportion of Malaysian adults who are in favor of increasing taxes on tobacco products, is 0.6.

B. Interval Estimation

We estimate an unknown parameter using an **interval of values** that is likely to contain the true value of that parameter and state how confident we are that this interval indeed captures the true value of the parameter.

Confidence intervals are not perfect. A 95% confidence interval implies that in repeated samples, 19 in 20 confidence intervals will capture the value of the population parameter.

> **Example**
>
> Based on sample results, we are 95% confident that p, the proportion of Malaysian adults who are in favor of increasing taxes on tobacco products, is between 0.57 and 0.63.

B.1- Interval Estimation for Population Proportion (Categorical Variable)

Suppose we are interested in the population proportion of a categorical variable.

- **Step 1**: We collect data from a sample of our population of size n
- **Step 2**: The values of \hat p follow a normal distribution with (unknown) mean p and standard deviation \sqrt{\frac {{\hat p }(1-{\hat p})}{n}}. As we do not know the population proportion p, we use the sample proportion \hat p.
- **Step 3**: According to the Standard Deviation Rule, this means that:
  - We are 95% confident that the population proportion p falls within 2*\sqrt{\frac {{\hat p }(1-{\hat p})}{n}} of our estimate \hat p.
  - A 95% confidence interval for the population proportion p is: \left(\hat p - 2*\sqrt{\frac {{\hat p }(1-{\hat p})}{n}}, \hat p + 2*\sqrt{\frac {{\hat p }(1-{\hat p})}{n}}\right)

Here, then, is the general result:

Suppose a random sample of size n is taken from a population for a categorical variable whose proportion (p) is unknown. A 95% confidence interval (CI) for p is: \left(\hat p - 2*\sqrt{\frac {{\hat p }(1-{\hat p})}{n}}, \hat p + 2*\sqrt{\frac {{\hat p }(1-{\hat p})}{n}}\right)

Example

A few days before a snap election, a polling organisation would like to estimate p, the proportion of eligible voters who support Candidate A. They choose a random sample of size 1000 and recorded their opinion. 71\% of the sample support this candidate. How do you estimate the proportion of the people in the constituency who will vote for this candidate?

Point estimate: \hat p = 71\%

Interval estimate: According to the *central limit theorem*, sample proportion, \hat p, follows the normal distribution \hat p \sim Normal\left(mean=p, sd=\sqrt{\frac {{\hat p }(1-{\hat p})}{n}}\right)

So, we can say \bar p follows the normal distribution Normal\left(mean=p , sd=\sqrt{\frac {{0.71}(1-{0.71})}{1000}}\right) Normal(mean=p, sd=0.014) where p is the population proportion. Therefore, we can say that

- there is 95% chance that p falls within 2\sigma_{\bar p} of \hat p.
- Using the empirical rule, we can say the 95% confidence interval for p is ({\hat p}-2\sigma_{\bar p} , {\hat p}+2\sigma_{\bar p}) = (0.71-2*0.014, 0.71+2*0.014) = (0.68, 0.73), where \sigma_{\bar p}=\sqrt{\frac {\hat p(1-{\hat p})}{n}}.
- This means that we are 95% confident that p lies within the interval (0.68, 0.73).

Exercise

Several public health researchers conducted a study to look at the connection between watching actors smoking in movies and initialising of smoking among adolescents. In the study, 6,522 teenagers aged 10-14 who had never tried smoking were randomly selected. Of those who subsequently tried smoking for the first time, 38% were exposed to smoking in the movies.

- A. Estimate the proportion of all U.S. adolescents ages 10-14 who started smoking after seeing actors smoke in movies by constructing a 95% confidence interval.
- B. Construct a 99.7% confidence interval for p.

In [ ]:

B.2- Interval Estimation for Population Mean (Numerical Variable)

Suppose we are interested in the mean of a numerical variable from a population.

- Case1: We assume the population standard deviation (\sigma) is **known**.

- **Step 1:** We collect data from a sample of our population of size n
- **Step 2:** The values of \bar x follow a normal distribution with (unknown) mean \mu and standard deviation \sigma_{\bar x}=\frac{\sigma}{\sqrt n} (known, since both \sigma and n are known).
- **Step 3:** According to the Standard Deviation Rule, this means that:
  - There is a 95% chance that our population mean \mu will fall within 2*\frac{\sigma}{\sqrt n} of \hat \mu
  - A 95% confidence interval for the population mean \mu \in \left(\bar x - 2* \frac{\sigma}{\sqrt n}, \bar x + 2* \frac{\sigma}{\sqrt n}\right)

Here, then, is the general result:

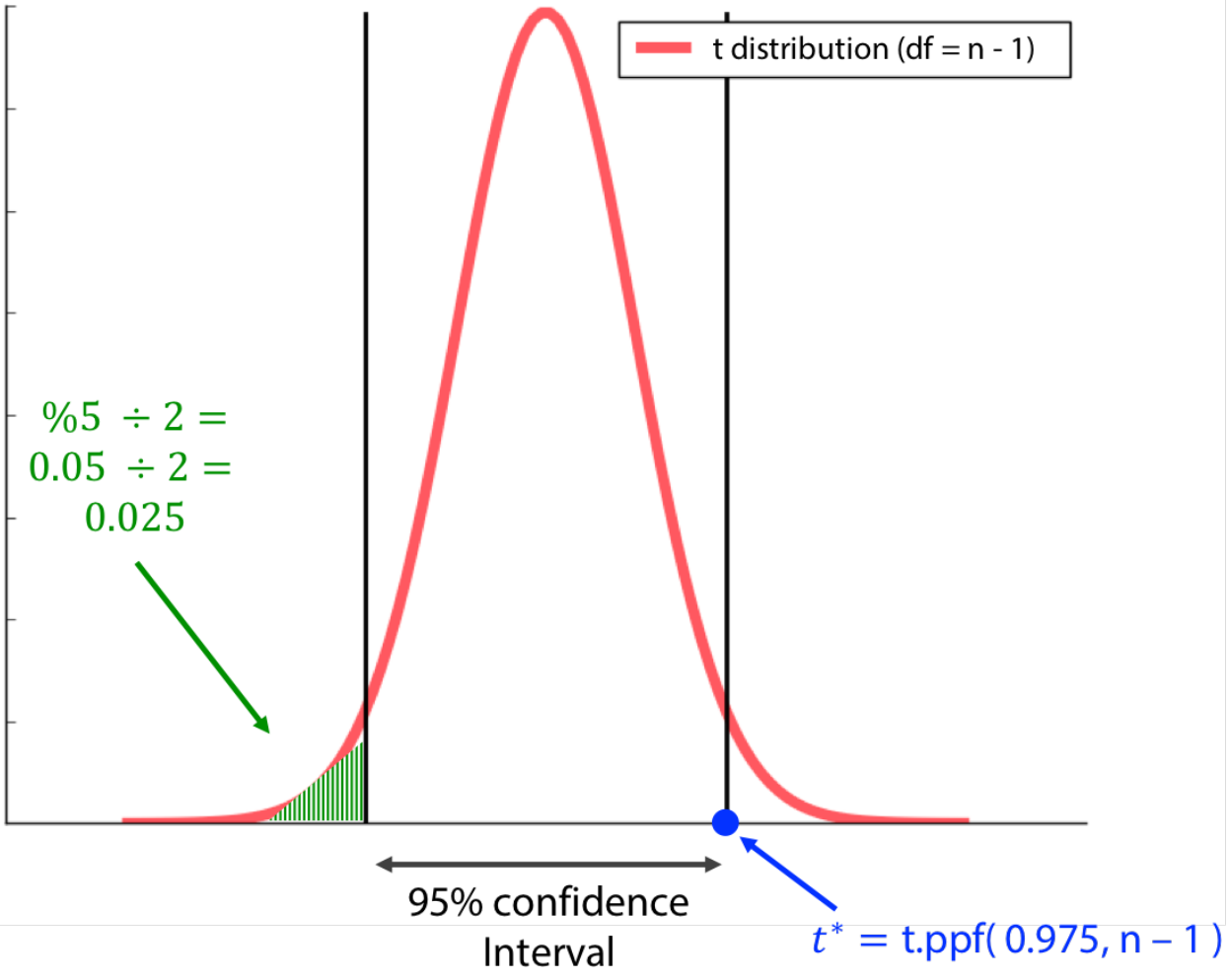
Suppose a random sample of size n is taken from a normal population of values for a quantitative variable whose mean (\mu) is unknown, when the standard deviation (\sigma) is given. A 95% confidence interval (CI) for \mu \in \left(\bar x - 2* \frac{\sigma}{\sqrt n}, \bar x + 2* \frac{\sigma}{\sqrt n}\right)

- Case2: We assume the population standard deviation (\sigma) is **unknown**. In this case, we can replace \sigma with s, where s is the standard deviation of the sample however, the central limit theorem will not be valid anymore and \bar x will not follow normal distribution. Instead, t = \frac {\bar x - \mu}{\frac {s}{\sqrt(n)}} will follow t-distribution with degree of freedom n - 1, where n is the sample size.

t^* = \frac {\bar x - \mu}{\frac {s}{\sqrt n}}

t^* \sim t(n - 1) Therefore, for each confidence interval t^* should be calculated such that: \mu \in \left(\bar x - t^* * \frac{s}{\sqrt n}, \bar x + t^* * \frac{s}{\sqrt n}\right).

Because t^* \sim t(n - 1) , the above interval depends on both the **confidence level and the sample size n**. For instance, the 95% cut-off in the following t-distribution can be calculated using python by t^* = t.ppf(0.975, n-1)

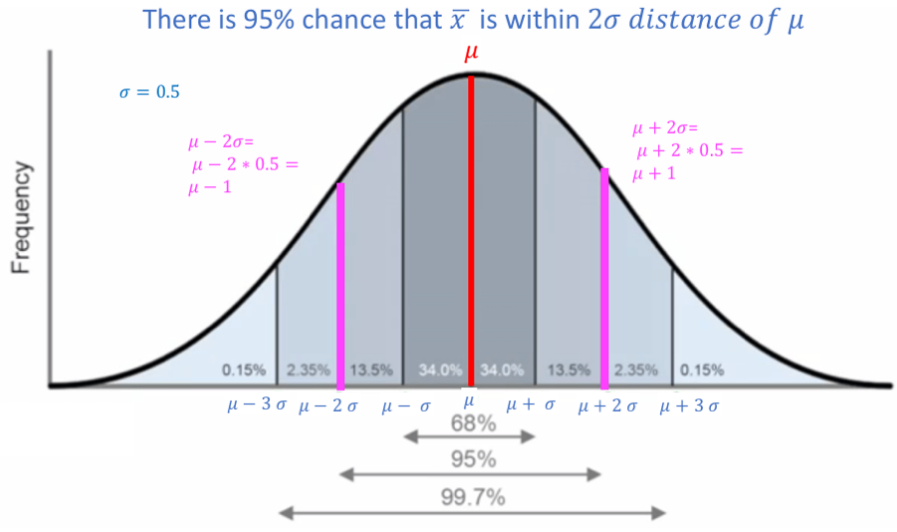


Example

Numerical Variable Mean: Suppose we are interested in studying the average IQ of students in a university. To do so, we collect a random sample of size 100 from the students in this university. Assume the mean of the IQ level of these students is 115, and its standard deviation is \sigma=5. What is \mu, the mean of the IQ level of the population which is the whole students at this university?

Point estimate: \hat \mu= 115
Interval estimate: \hat \mu=115, \sigma=5, n=100.

- Note that \frac{5}{\sqrt{100}} is the standard deviation of the sampling distribution of sample estimates \bar x - the *standard error*, \sigma_{\bar x}.
- According to the *central limit theorem*, the distribution of the sample means \bar x(s) follows a normal distribution: \bar x \sim Normal(mean=\mu, sd=\frac{\sigma}{\sqrt{n}})
- Since \hat \mu is our estimate of \mu, the sample means are distributed as \bar x\sim Normal(115, 0.5)
- Recall the standard deviation rule:



Since two standard errors = 1, the statement:

> "There is a **95% chance** that the sample mean \bar x falls within 1 unit of \mu".

can be rephrased as:

> "We are 95% confident that the population mean \mu falls within 1 units of \bar x".

Given a sample mean of \bar x=115, we can be **95% confident** that \mu falls within 1 unit of 115, or in other words that \mu is covered by the interval (115 - 1, 115 + 1) = (114,116).

Exercise

An educational researcher was interested in estimating \mu, the mean score on the total SAT scores of all college students in a state. To this end, the researcher has chosen a random sample of 650 college students from his state, and found that their average SAT score is 1425. Based on a large body of research that was done on the SAT, it is known that the scores roughly follow a normal distribution with the standard deviation \sigma=300.

- A. Based on this information, construct a 95% confidence interval for \mu.
- B. Construct a 99.7% confidence interval for \mu.

In [ ]:

Exercise

What is the relationship between the level of the confidence and the length of the confidence interval?

In [ ]:

Example

Repeat the above example if population standard deviation is unknown but sample standard deviation is 4.5.

In [ ]:

Exercise

Repeat the above exercise it the population standard deviation is unknown, but the sample standard deviation is 298.

- A. Based on this information, construct a 95% confidence interval for \mu.
- B. Construct a 99.7% confidence interval for \mu.

In [ ]:

C. Hypothesis testing

The disciplinary committee of a university investigates a student suspected of cheating on an exam. There are two opposing claims in this case:

- The student claims that he did not cheat on the exam.
- The lecturer claims that the student did cheat on the exam.

The committee assumes the student to be innocent unless the lecturer can prove that the student is guilty. Therefore, the committee asks the instructor to provide evidence to support his claim. The lecturer explains that he set two versions of the exam, and on four separate exam questions, the student answered with numbers provided in the other version of the exam.

The committee agrees that it would be extremely unlikely for the lecturer to have such strong evidence if the student did not cheat. In other words, the lecturer provided strong enough evidence for the committee to **reject** the student's claim, and **conclude** that the student did cheat on the exam.

Hypothesis testing is defined as **assessing evidence provided by the data in favour of or against some claim about the population.**

Here is how the process of statistical hypothesis testing works:

- **Step 1:** We have **two claims** about what is going on in the population: claim 1 and claim 2. In the story above, where the instructor's claim challenges the student's claim, **claim 1 is challenged by claim 2**. In hypothesis testing, we usually test 'claims' (or hypotheses) about the value of population parameter(s) or about whether a relationship exists between two variables in the population.

- **Step 2:** We choose a **sample**, collect relevant data and summarize them. This is similar to the instructor collecting evidence from the student's exam.
- **Step 3:** We figure out **how likely** it is to observe data like the data we got, had claim 1 been true. (Note that the wording "how likely ..." implies that this step requires some kind of probability calculation). In our story, the committee members assessed how likely it is to observe the evidence provided by the instructor if the student's claim of not cheating was true.
- **Step 4:** Based on what we found in the previous step, we make our decision:
  - If we find that it would be extremely unlikely to observe the data that we observed if claim 1 were true, then we have strong evidence against claim 1, and we **reject** it in favour of claim 2.
  - If we find that observing the data that we observed is not very unlikely if claim 1 were true, then we do not have enough evidence against claim 1, and therefore we **cannot reject** it in favour of claim 2.

In our story, the committee decided that it would be extremely unlikely to find the evidence that the lecturer provided if the student did not cheat. In other words, the members felt that it is extremely unlikely that it is just a coincidence that the student used the numbers from the other version of the exam on four separate problems. The committee members therefore decided to reject the student's claim and concluded that the student had, indeed, cheated on the exam.

Example

A recent study estimated that 14.6% of all upper secondary school students in Malaysia smoke. (https://tobaccoinduceddiseases.biomedcentral.com/articles/10.1186/s12971-016-0108-5). The head of a district education office suspects that the proportion of smokers may be lower there. In hopes of confirming her claim, she chooses a random sample of 400 upper secondary high school students in the district, and finds that 50 of them are smokers.

Let's analyze this example using the 4 steps outlined above:

Stating the claims:

There are two claims here:

- *claim 1:* The proportion of smokers in the district is 0.146.
- *claim 2:* The proportion of smokers at Goodheart is less than 0.146.

Claim 1 basically says "nothing special goes on in this district; the proportion of smokers there is no different from the proportion in the entire country." This claim is challenged by the head of the district office, who suspects that the proportion of smokers in her district is lower.

Choosing a sample and collecting data:

A sample of $n = 400$ was chosen, and summarizing the data, we find that the sample proportion of smokers is $\hat p = \frac {50}{400} = 0.125$

While it is true that 0.125 is less than 0.146, it is not clear whether this is strong enough evidence against claim 1.

Assessment of evidence:

To assess whether the data provide strong enough evidence against claim 1, we need to ask ourselves: How surprising is it to get a sample proportion as low as $\hat p = 0.125$ (or lower) if claim 1 is true?

In other words, we need to find how likely it is that in a random sample of size $n = 400$ taken from a population where the proportion of smokers is $p = 0.146$ we'll get a sample proportion as low as $\hat p = 0.125$ (or lower).

It turns out that the probability that we'll get a sample proportion as low as $\hat p = 0.125$ (or lower) if $p = 0.146$ is roughly 0.117 (do not worry about how this was calculated at this point).

Conclusion:

We found that there is a probability of 0.117 of observing data like that observed if claim 1 were true.

Now you have to decide ... Do you think that a probability of 0.117 makes our data rare enough (surprising enough) under claim 1 so that the fact that we did observe it is enough evidence to reject claim 1? Or do you feel that a probability of 0.117 means that the data we observed are not very likely when claim 1 is true, but not unlikely enough to conclude that getting such data is sufficient evidence to reject claim 1?

Hypothesis testing (General Case)

- **Step 1: Stating the claims**: Our aim is to decide between two opposing points of view, *Claim 1* and *Claim 2*. In hypothesis testing, Claim 1 is called the **null hypothesis** (denoted H_0), and Claim 2 plays the role of the **alternative hypothesis** (denoted H_a).

- **Step 2: Choosing a sample and collecting data**: We look at sampled data to draw conclusions about the entire population. In hypothesis testing, based on the data, you draw conclusions about whether there is enough evidence to reject H_0.
- **Step 3: Assessing the evidence**: This is the step where we calculate how likely is it to get data like that observed when H_0 is true. We use the **p-value** to assess the evidence. It is **the probability of observing a test statistic as extreme as (or even more extreme than) that observed assuming that the null hypothesis is true.**

p-value = The probability of observing a test statistic as extreme as (or even more extreme than) that observed assuming that the null hypothesis is true.

- **Step 4: Making conclusions**: Since our conclusion is based on how small the p-value is, it would be nice to have some kind of threshold or cutoff that will help determine how small the p-value must be, or how "rare" (unlikely) our data must be when H_0 is true, for us to conclude that we have enough evidence to reject H_0.

This cutoff has a special name. It is called the **significance level** of a test and is usually denoted by the Greek letter α. The most commonly used significance level is $\alpha = 0.05$ (or 5%). We use the following decision rule:

- if the p-value < α (usually 0.05 or 5%), then the data we got is considered to be "rare (or surprising) enough" when H_0 is true, and we say that the data provide significant evidence against H_0, so we reject H_0 and accept H_a.
- if the p-value > α (usually 0.05 or 5%), then our data are not considered to be "surprising enough" when H_0 is true, and we say that our data do not provide enough evidence to reject H_0 (or, equivalently, that the data do not provide enough evidence to accept H_a).

Linked to the concept of a *significance level* is the **confidence level** . A significance level of 0.05 or 5% corresponds with a 95% confidence level. The confidence level is associated with the confidence interval. For instance, you can construct a 95% confidence interval, where 95% refers to the confidence level. Just like p-values, confidence intervals can be used to do hypothesis testing. The decision rule for confidence intervals is as follows:

- If sample parameter falls outside the 95% confidence interval, reject H_0 in favour of H_a.
- If sample parameter falls inside the 95% confidence interval, do not reject H_0.

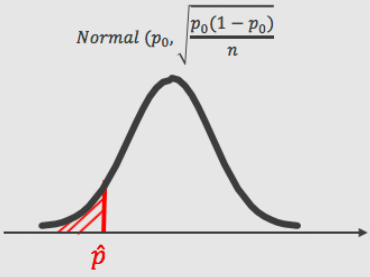
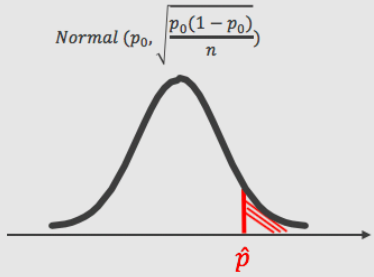
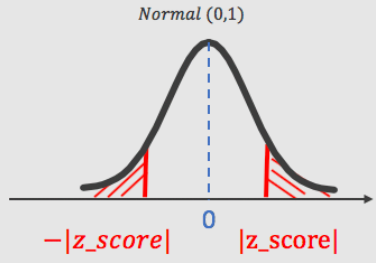
Note that the null hypothesis can never be accepted - you can only reject the null hypothesis in favour of the alternative hypothesis, or fail to reject the null hypothesis.

3. One-sample hypothesis testing

One Sample Hypothesis Testing
Population Proportion

$$p \sim Normal\ (mean = p_0,\quad sd = \sqrt{\frac{p_0 \times (1 - p_0)}{n}})$$

$\hat{p}$ = Sample proportion ; p_0 = Null hypothesis proportion ; n = Sample size

Case 1	Case 2	Case 3
$H0:\ p = p_0$ $Ha:\ p < p_0$	$H0:\ p = p_0$ $Ha:\ p > p_0$	$H0:\ p = p_0$ $Ha:\ p \neq p_0$
$p - value = P(p < \hat{p})$	$p - value = P(p > \hat{p}) = 1 - P(p < \hat{p})$	$z - score = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \times (1 - p_0)}{n}}}$ $p - value = 2 \times P(p < -\lvert z - score\rvert)$
		

Compute p-value

Solution 1

$p - value = norm.cdf(\hat{p}, p_0, sd)$	$p - value = 1 - norm.cdf(\hat{p}, p_0, sd)$	$p - value = 2 * norm.cdf(-\lvert z - score\rvert)$

Solution 2

$p - value = proportions\_ztest\ (count = n * \hat{p}, nobs = n, value = p0, \text{prop\_var} = p0, alternative = 'smaller'\ or\ 'larger'\ or.'two - sided')$ [1]

- ✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
- ✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Proportions

Example

Our workers are known to produce 20% defective products, and are sent for retraining. After the training, 400 products produced are chosen at random and 64 are found to be defective (proportion $\hat p=\frac{64}{400}=0.16$). Do the data provide enough evidence that the proportion of defective products produced by our workers, p has been reduced as a result of the training?

Based on our problem, we formulate the following hypotheses:

- H_0: $p = 0.20$ (No change; the training did not help, $p_0=0.20$)
- H_a: $p < 0.20$ (The training was effective)

In [ ]:

In [ ]:

Exercise

Polls on certain topics are conducted routinely to monitor changes in the public's opinions over time. One such topic is the death penalty. In 2013, a poll estimated that 91% of 1,535 Malaysian adults surveyed support the death penalty for people convicted of murder. In a more recent poll, 890 out of 1,000 Malaysian adults chosen at random were in favor of the death penalty for convicted murderers. Do the results of this poll provide evidence that the proportion of Malaysian adults who support the death penalty for convicted murderers (p) changed between 2013 and the later poll?

In [ ]:

Means

We need to distinguish between two cases: where the population standard deviation (\sigma) is known, and the case where \sigma is unknown.

- If \sigma is **known**, the test is called the **z-test** for the population mean \mu because the sample mean follows the **normal** distribution Normal(mean=\mu_0, std=\frac {\sigma}{\sqrt {n}}) where n = sample size; \mu_0 = population mean according to the null hypothesis, and \sigma = population standard deviation. Therefore, the test statistic z=\frac {{\bar x}-\mu_0}{\frac {\sigma}{\sqrt {n}}}, which is the standardised sample mean, follows a **z-distribution**, or a standard normal distribution.

One Sample Hypothesis Testing
Mean; σ is KNOWN

$$\bar{x} \sim Normal\left(mean = \mu_0, \quad sd = \frac{\sigma}{\sqrt{n}}\right)$$

$\bar{x}$ = Sample mean ; μ_0 = *Null hypothesis* mean ; σ = Population Standard Deviation; n = *Sample size*

Case 1	Case 2	Case 3
$H0$: mean $= \mu_0$ Ha: mean $< \mu_0$	$H0$: mean $= \mu_0$ Ha: mean $> \mu_0$	$H0$: mean $= \mu_0$ Ha: mean $\neq \mu_0$
$p - value = P(x < \bar{x})$	$p - value = P(x > \bar{x}) = 1 - P(x < \bar{x})$	$z - score = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ $p - value = 2 \times P(x < -\lvert z - score \rvert)$
$p - value = norm.cdf(\bar{x}, \mu_0, sd)$	$p - value = 1 - norm.cdf(\bar{x}, \mu_0, sd)$	$p - value = 2 * norm.cdf(-\lvert z - score \rvert)$

✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95

✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

The SAT, a standardised test for college admissions in the US, is constructed so that scores in each portion have a national average of 500 and standard deviation of 100. The distribution is close to normal. The Marketing department of your college believes that in recent years the college attracts students who are more math-inclined. A random sample of 15 students from a recent cohort at your College had an average math SAT (SAT-M) score of 550. Does this provide enough evidence for the dean to conclude that the mean SAT-M of all your college's students is higher than the national mean of 500? Assume that the standard deviation of 100 applies also to all students at your college.

Solution 1:

The sampling distribution of \bar x under the null hypothesis is normal: \bar x \sim Normal(mean=\mu_0, std=\frac {\sigma}{\sqrt {n}}), where \mu_0 = 500 and \sigma = standard deviation of population

- H0: mean = 500
- Ha: mean > 500

In [ ]:

Exercise

Human pregnancy is known to have a mean of 266 days and a standard deviation of 16 days. Based on records from a large hospital, a random sample of 30 women who were smoking and/or drinking alcohol during their pregnancy and their pregnancy lengths are recorded. We calculated the average pregnancy length of these women as 258.78. Do the data provide enough evidence to support the (well-known) fact that women who smoke and/or drink alcohol during their pregnancy have shorter pregnancies than women in general (in other words, are more likely to have premature labor)?

In [ ]:

- If \sigma is **unknown**, the test is called the **t-test** for the population mean \mu because the standardised sample mean follows a **t-distribution**. In other words, the test statistic t=\frac {{\bar x}-\mu_0}{\frac {s}{\sqrt {n}}} follows t-distribution t \sim t(n-1) where s is the standard deviation of the sample, and n is the sample size.

One Sample Hypothesis Testing
Mean; σ is UNKNOWN

$\bar{x} \sim t(n-1)$

$\bar{x}$ = Sample mean ; s = Sample standatd deviation ; μ_0 = *Null hypothesis* mean ; n = *Sample size*

Case 1	Case 2	Case 3
$H0:\ \text{mean} = \mu_0$ $Ha:\ \text{mean} < \mu_0$	$H0:\ \text{mean} = \mu_0$ $Ha:\ \text{mean} > \mu_0$	$H0:\ \text{mean} = \mu_0$ $Ha:\ \text{mean} \neq \mu_0$
$p-value = P(x < \bar{x})$	$p-value = P(x > \bar{x}) = 1 - P(x < \bar{x})$	$p-value = 2 \times P(x < -\lvert t-score \rvert)$
Compute p-value		
Solution 1: Only $\bar{x}$ and s are given		
$t-score = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$		
$p-value = t.cdf(t-score, n-1)$	$p-value = 1 - t.cdf(t-score, n-1)$	$p-value = 2 * t.cdf(-\lvert t-score \rvert, n-1)$
Solution 2: SAMPLE_DATA is given		
Calculate $\bar{x}$ and s from the SAMPLE_DATA and use solution 1. OR $p-value$ = .5 × ttest_1samp(SAMPLE_DATA , μ_0).pvalue		$p-value =$ ttest_1samp(SAMPLE_DATA , μ_0).pvalue

✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95

✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

A certain prescription medicine is supposed to contain an average of 250 parts per million (ppm) of a certain chemical. If the concentration is higher than this, the drug may cause harmful side effects; if it is lower, the drug may be ineffective. The manufacturer wants to know whether the mean concentration in a large shipment conforms to the target level of 250 ppm.

A simple random sample of 100 portions is tested, and the sample mean concentration is found to be 246 ppm with a sample standard deviation of 12 ppm.

The hypotheses are:

- H_0: \mu = 250
- H_a: \mu \neq 250

In [ ]:

Exercise

On average, a Finnish consumes 12kg of coffee in a year, which is 5 cups a day per person. A Finnish university wants to know whether their students tend to drink more coffee than the national average. They ask 50 students how many cups of coffee they drink each day and found their average number of drinks is \bar{x}=5.2, with std dev s=1.5. Do they have enough evidence that their students drink more than the national average?

In [ ]:

Example

The mean of crude birth rate has been 16.7 per 1000 population in Malaysia in 2014. The following data shows crude birth rate from January to March 2019. Does the data prove a significant difference in 2019 comparing to 2014?

- H0: mu = 16.7
- Ha: mu != 16.7

In [ ]:

In [ ]:

In [ ]:

Exercise

In the above example, can we say crude birth rate has been decreased in 2019?

In [ ]:

4. Two-sample hypothesis test

In the previous sections we performed inference for one variable. If this variable was categorical, we perform one-sample hypothesis test for proportions. If the variable was numerical/quantitative, we perform one-sample hypothesis test for mean.

In this section, we look at inference about relationships between two variables in a population, based on an observed relationship between variables in a sample.

Assume we are interested in studying whether a relationship exists between the variables x and y in a population of interest. We choose a random sample and collect data on both variables from the subjects. Our goal is to determine whether these data provide strong enough evidence for us to generalize the observed relationship in the sample and conclude (with some acceptable and agreed-upon level of uncertainty) that a relationship between x and y exists in the entire population.

- H_0: There is no relationship between x and y
- H_a: There is a significant relationship between x and y

C ---> Q

We consider hypothesis testing where x, the explanatory variable, is a **categorical** variable and y, the response variable, is a **quantitative** variable.

Example

To investigate this relationship between year in university and GPA, we can divide the population of the university students in Malaysia into 4 sub-populations. Within each of these four groups, we are interested in the GPA.

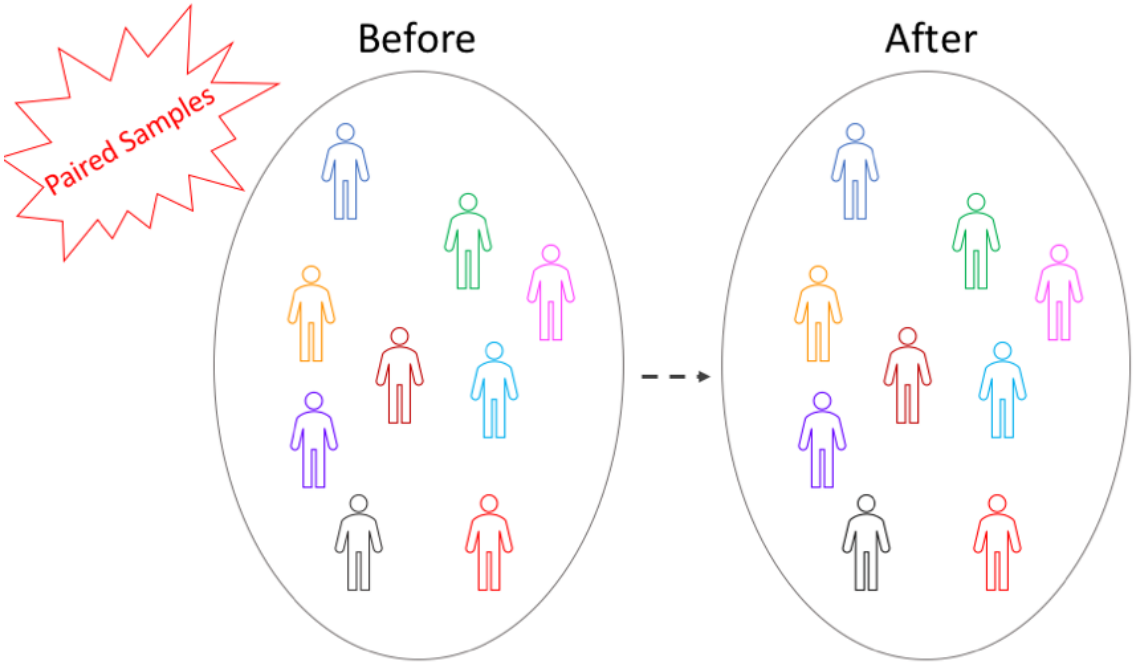
The inference must therefore involve the 4 sub-population means:

- \mu_1: mean GPA among first year undergraduates in Malaysia
- \mu_2: mean GPA among second year undergraduates in Malaysia
- \mu_3: mean GPA among third year undergraduates in Malaysia
- \mu_4: mean GPA among fourth year undergraduates in Malaysia

So, we need to compare these four means. If we infer that not all these four means are equal (i.e., that there are some differences in GPA across years in university) then that's equivalent to saying GPA is related to year in university.

Example

Assume x is drinking/not_drinking alcohol, and y is reaction time of the driver. We are interested to explore the impact of drinking two beers on the driver's reaction time. In this case, we measure the reaction time of 40 drivers, **before** and **after** drinking two beers.



Two-sample t-test of means for unpaired samples

Two-sample t-test of means for unpaired samples

Condition 1: The two samples are indeed independent
Condition 2: The distribution of y in both sub-populations is **NORMAL**, and both samples are random
Condition 3: The populations are **NOT NORMAL**, but the sample size of each of the random samples is **large enough ($n > 30$)**

$$t\_score \sim t(\nu)$$

$\overline{y_1}$ and $\overline{y_2}$ are sample means; s_1 and s_2 are sample standatd deviations; μ_1 *and* μ_2 *are population means*; n_1 and n_2 are *ample sizes*

Case 1		Case 2		Case 3	
$H0: \ \mu_1 = \mu_2$ $Ha: \ \mu_1 < \mu_2$	$H0: \ \mu_1 - \mu_2 = 0$ $Ha: \ \mu_1 - \mu_2 < 0$	$H0: \ \mu_1 = \mu_2$ $Ha: \ \mu_1 > \mu_2$	$H0: \ \mu_1 - \mu_2 = 0$ $Ha: \ \mu_1 - \mu_2 > 0$	$H0: \ \mu_1 = \mu_2$ $Ha: \ \mu_1 \neq \mu_2$	$H0: \ \mu_1 - \mu_2 = 0$ $Ha: \ \mu_1 - \mu_2 \neq 0$
Compute p-value					
Solution1: Only sample parameters are given $\overline{y_1}$, $\overline{y_2}$, s_1, s_2, n_1, n_2					
$t - score = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$; $\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2 (n_1 - 1)} + \frac{s_2^4}{n_2^2 (n_2 - 1)}}$					
$p - value = P(t - score < 0) =$ t. cdf(t − score, ν)		$p - value = P(t - score > 0) =$ 1 - $P(t - score < 0)$ = 1 - t. cdf(t − score, ν)		$p - value = 2 \times P(\mathrm{x} < -\lvert t - score \rvert) =$ $2 \times$ t. cdf(−abs(t − score), ν)	
Solution2: SAMPLE_DATA id given					
Calculate $\overline{y_1}$, $\overline{y_2}$, s_1, s_2 *and* n_1, n_2 from the SAMPLE_DATA and use solution 1. OR $p - value$ = .5 × ttest_ind(sample1, sample2, equal_var=False).pvalue				$p - value$ = ttest_ind(sample1, sample2, equal_var=False).pvalue	

✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

The two-sample t-test can be safely used as long as the following conditions are met:

1. **Both populations are normally distributed**, or more specifically, the distribution of y in both sub-populations is normal, and both **samples are random** (or at least can be considered as such). In practice, checking normality in the sub-populations is done by looking at each of the samples using a histogram and checking whether there are any signs that the populations are not normal. Such signs could be extreme skewness and/or extreme outliers.
2. The populations are known or discovered not to be normal, but the **sample size of each of the random samples is large enough** (we can use the rule of thumb that n> 30 is considered large enough).

The two-sample t-test statistic is:

t = \frac{\bar y_1-\bar y_2}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}

Where:

- \bar y_1 and \bar y_2 are the sample means of the samples from sub-population 1 and sub-population 2 respectively.

- s_1 and s_2 are the sample standard deviations of the samples from sub-population 1 and sub-population 2 respectively.

- n_1 and n_2 are the sample sizes of the two samples.

Attention: To understand the t-test statistic we need to know that

- $\bar y_1$ estimates μ_1 (mean of sub-population 1) and
- $\bar y_2$ estimates μ_2 (mean of sub-population 2).

Therefore, $\bar y_1 - \bar y_2$ estimates $\mu_1 -\mu_2$.

$\mu_1 -\mu_2 = 0$ is the "null value" — what the null hypothesis, H_0, claims that $\mu_1 -\mu_2$ is.

The denominator ${\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$ is the standard deviation of $\bar y_1 - \bar y_2$

We therefore see that our test statistic, like the previous test statistics we encountered, has the structure:

$\frac {Sample\,Estimate - Null\,Value}{Standard\,Error}$ and therefore, like the previous test statistics, measures (in standard errors) the difference between what the data tell us about the parameter of interest $\mu_1 -\mu_2$ (sample estimate) and what the null hypothesis claims the value of the parameter is (null value).

The number of degrees of freedom is ν where:

$$\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Degrees of freedom refers to the number of number of observations that are free to vary when calculating a statistic.

Example

Assume we are interested in investigating the relationship between a patient having a heart attack and the level of cholesterol. The variables we have are:

x: patient had heart attack (yes/no) ---> Categorical

y: patient cholesterol level (number) ---> Quantitative

We measured the cholesterol level of 38 heart attack patients (2 days after their attacks) and 40 other hospital patients who did not have a heart attack.

For the 38 heart attack patients, the mean cholesterol level was 253.9 with a standard deviation of 47.7. For the 40 other hospital patients who did not have a heart attack, the mean cholesterol level was 193.1 with a standard deviation of 22.3. Are cholesterol levels different across the different groups?

Answer:

- H_0: \mu_1-\mu_2 = 0
- H_a: \mu_1-\mu_2 \neq 0

n_1 = 38; \bar y_1 = 253.9; s_1 = 47.7

n_2 = 40; \bar y_2 = 193.1; s_2 = 22.3

t = \frac{\bar y_1-\bar y_2}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}} = \frac{253.9-193.1}{\sqrt{\frac{{47.7}^2}{38}+\frac{{22.3}^2}{40}}} = 7.150

df = \frac{(\frac{47.7^2}{38} + \frac{22.3^2}{40})^2}{\frac{47.7^4}{38^2(38-1)} + \frac{22.3^4}{40^2(40-1)}} = 51.84

`p_value = 2*t.cdf(-abs(t), df)` = 2.8980437531650854e-09

To make things easier, let's write a function `unpaired_t` that returns t and \nu:

In [ ]:

```

```

In [ ]:

```

```

Example

To check the claim that the pregnancy length of women who smoke during pregnancy is shorter, on average, than the pregnancy length of women who do not smoke, a random sample of 35 pregnant women who smoke and a random sample of 35 pregnant women who do not smoke were chosen and their pregnancy lengths were recorded.

x: smoke (yes/no) ---> Categorical variable

y: pregnancy length ---> Quantitative variable

Two methods can be used:

1. calculating t and \nu and then use the function `t.cdf`
2. using `scipy.stats.ttest_ind` function

The hypotheses are as follows:

- H_0: \mu_1 - \mu_2 = 0 (There is no relationship between smoking and pregnancy length)
- H_a: \mu_1 - \mu_2 < 0 (Pregnancy length of women who smoke is shorter than the pregnancy length of women who do not smoke)

In [ ]:

In [ ]:

In [ ]:

Using `scipy.stats.ttest_ind` we can directly test two independent samples without calculating the means, t-statistic and degrees of freedom:

Exercise

A researcher wanted to study whether men and women watch different amounts of YouTube. A random sample of 400 adults was chosen, comprising of 191 women and 209 men. At the end of the week, each of the 400 subjects reported the total amount of time (in minutes) that he or she watched YouTube during that week.

In [ ]:

In [ ]:

Two-sample t-test for paired samples means

Example

Drunk driving is one the main causes of car accidents. We want to know whether drivers are impaired after drinking two beers. A sample of 30 drivers was chosen, and their reaction times in an obstacle course were measured before and after drinking two beers. The variables we have are:

x: Drinking alcohol (yes/no) ---> Categorical

y: Reaction time of the driver ---> Quantitative

Let \mu_1 be the average of the reaction time before drinking 2 beers, and \mu_2 be the average of the reaction time before drinking 2 beers.

https://www.youtube.com/watch?time_continue=1&v=URPrSH0Lg_M (https://www.youtube.com/watch?time_continue=1&v=URPrSH0Lg_M)

We reduce the two samples to only one by calculating for each pair the difference between the two observations

Pairs	1	2	3	4	...	n
Sample 1	a1	a2	a3	a4		an
Sample 2	b1	b2	b3	b4		bn
Differences	a1-b1	a2-b2	a3-b3	a4-b4		an-bn

Two sample paired test → One sample t-test

- H_0: $\mu_d = 0$ OR $\mu_1 - \mu_2 = 0$
- H_a:$\mu_d < 0$ OR $\mu_1 - \mu_2 < 0$ If the driver is drunk it takes more time to react $\mu_1 < \mu_2$.

Assume

- mean of differences in the sample is -0.501,
- standard deviation of differences is 0.868.

and therefore

- $\bar x_d = -0.501$,
- $s_d = 0.868$,
- $n=30$,
- $\mu_0 = 0$

$t = \frac {{\bar x_d}-\mu_0}{\frac {s_d}{\sqrt {n}}} = \frac {-0.501}{\frac {0.868}{\sqrt {30}} }$

In [ ]:

Example

Suppose we want to evaluate the effectiveness of this course on the statistics skills of the students. Assume there are 100 students in a cohort. We record their scores in a sample statistics test **before** and **after** passing this course. Did our students' statistics skills improve after taking the class?

Define \mu_d=\mu_2-\mu_1

- H_0: \mu_d = 0
- H_a: \mu_d > 0

In [ ]:

In [ ]:

There are two ways to run a paired t-test:

- treat it as a one-sample t-test where we test the differences between the two samples. Naturally, the null hypothesis is that the difference between the two samples is 0. Use `scipy.stats.ttest_1samp`
- pass the two samples directly to `scipy.stats.ttest_rel`

In [ ]:

In [ ]:

5. Hypothesis testing for more than two samples

C ---> Q

Comparing More Than Two Means—ANOVA

So far, we have discussed the two samples and matched pairs designs, in which the categorical explanatory variable has two values. In these cases, examining the relationship between the explanatory and the response variables amounts to comparing the mean of the response variable y in two populations, defined by the two values of the explanatory variable x. The difference between the two samples and matched pairs designs is that in the former, the two samples are independent (not paired), and in the latter, the samples are dependent (paired).

We are now moving on to cases in which the categorical explanatory variable takes more than two values.

ANOVA

X: categorical variable with k categories $x_1, x_2, \dots, x_k$

Y: numerical (quantitative) variable

sample_i = {y | where $X = x_i$}, for i = 1, 2, ..., k

x	x_i	x_i	x_i	...	x_i
y	y_{i1}	y_{i2}	y_{i3}	⋯	y_{iN}

← sample_i

$\mu_i = mean\ of\ y\ in\ the\ \mathbf{population}\ where\ x == i$

Question:

Is there any significant relationship between x and y in the population? Does x impact y? Is there significant difference between μ_i's?

$$H0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

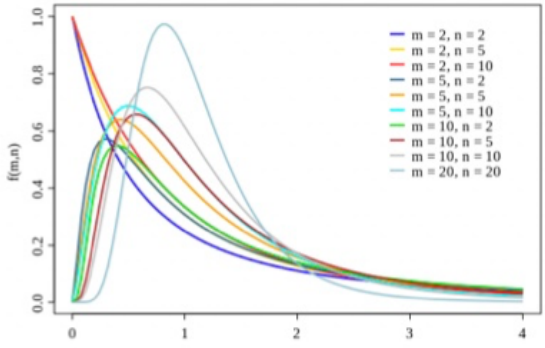
$$Ha: not\ all\ \mu's\ are\ equal$$

$$f\_stat = \frac{Variation\ Between\ Sample\ Means}{\text{Variation Within Groups}}$$

$$p\_value = P(f > f\_stat)$$

where f has f-distribution f(df_between, df_within)

$$df\_between = k - 1 \quad \& \quad df\_within = N - k \quad where \quad N = sample\ size \quad \& \quad k = number\ of\ categories\ in\ X$$



Compute p-value

f_stat, p_value = f_oneway(sample_1, sample_2, ..., sample_k)

- ✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
- ✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

A drug company tested three types of pain relief medication for migraines. For the experiment, 27 volunteers were selected and 9 were randomly assigned to each of the three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 to 10 (10 being most pain).

Groups	Collected data
Drug A	4 3 4 4 4 5 4 3 2
Drug B	4 6 5 8 6 6 8 4 5
Drug C	7 5 6 5 5 6 7 6 6

The hypotheses can be stated as follows:

- H_0: \mu_A = \mu_B = \mu_C
- H_a: not all the \mus are equal

In [ ]:

In [ ]:

There are two ways to compute ANOVA in Python. Firstly, we can use `scipy.stats.f_oneway` , which returns the f-statistic and the p-value:

In [ ]:

The statistic we are concerned about is the F-statistic: f_{k-1,n-k} \sim \frac {Variation\,Between\,Sample\,Means}{Variation\,Within\,Groups}

Under the null hypothesis that there's no difference in group means, the F-statistic is expected to be around 1. Contrast this with our F-statistic 11.91! Should we reject our hypothesis that there's no difference in group means?

Note that you can also compute the F-statistic by taking the ratio of `mean_sq` (mean square error) of `drug` - the 'between-group variation' - to the `mean_sq` of `Residual` - the 'within groups' variation.

Together with the degrees of freedom, this gives us a p-value of 0.0003. So, we clearly reject the null hypothesis of equal means for all three drug groups.

The F-distribution has two degrees of freedom parameters, k-1 and N-k, where N is the sample size and k the number of groups. Knowing this, we can recompute the p-value based on the f-statistic:

In [ ]:

Draw a boxplot based on the results.

In [ ]:

Multiple testing

Knowing that the means of A, B and C are not equal, we might want to know which pairs of drugs have different levels of pain. You could do a series of pairwise t-tests, i.e.

In [ ]:

In [ ]:

However, we have to correct for multiple testing. The more tests you run, the more likely you'll incorrectly find a significant result in any one pair of results by chance alone. The Bonferroni correction simply sets the significance threshold to be \alpha/m where m is the number of hypotheses tested, and \alpha is the level of significance.

In our case, 3 hypotheses are being tested, so divide alpha by 3 to obtain 0.167. Compare this against our array of p-values:

In [ ]:

which shows a significant difference between groups A and B and groups A and C but not groups A and C.

`statsmodels.sandbox.stats.multicomp.multipletests` is a convenient wrapper function for this procedure. It returns four objects, but we'll focus on three:

In [ ]:

The first object returned, `reject`, provides the result of the multiple testing without the p-values:

In [ ]:

Should you want to compare the p-values against the original level of alpha, it is equivalent to correct the p-values by multiplying the p-values by the number of hypotheses being tested:

In [ ]:

`multipletests` also returns the Bonferroni-corrected level of alpha:

In [ ]:

C ---> C

The last three procedures that we studied (two-sampled t, paired t, and ANOVA) all involve the relationship between a categorical explanatory variable, x, and a quantitative response variable, y. Next, we will consider inferences about the relationship between two categorical variables.

Chi-square test for equality of proportions in two samples

For the test of proportion to be valid, we generally need the following:

- For a right- or left-tailed test, a minimum of 10 successes and 10 failures in each group are necessary.
- Two-tailed tests are more robust and require only a minimum of 5 successes and 5 failures in each group

Chi-Square

Condition: $a_{ij} > 5 \; for \; i = 1, 2, \dots, s \quad \& \quad j = 1, 2, \dots, k$

Question: Is there any significant relationship between X and Y in the population?

X: categorical variable with k categories $x_1, x_2, \dots, x_k$

Y: categorical variable with k categories $y_1, y_2, \dots, y_s$

Observed_Counts

X / Y	x_1	x_2	...	x_k
y_1	a_{11}	a_{12}	...	a_{1k}
y_2	a_{21}	a_{22}	...	a_{2k}
...				
y_s	a_{s1}	a_{s2}	...	a_{sk}

$$a'_{mn} = P(Y = y_m) \times P(X = x_n) \times \sum_{i,j=1}^{s,k} a_{ij} =$$

$$\frac{\sum_{i=1}^{s} a_{i1} \times \sum_{j=1}^{k} a_{1j}}{\sum_{i,j=1}^{s,k} a_{ij}}$$

Expected_Counts

X / Y	x_1	x_2	...	x_k
y_1	a'_{11}	a'_{12}	...	a'_{1k}
y_2	a'_{21}	a'_{22}	...	a'_{2k}
...				
y_s	a'_{s1}	a'_{s2}	...	a'_{sk}

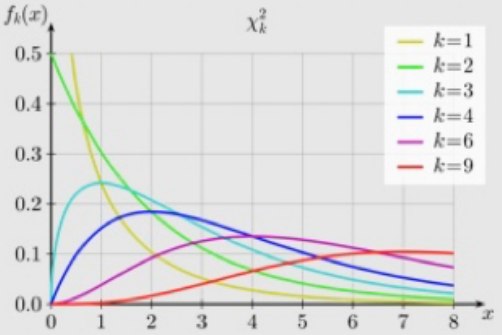
H0: X and Y are independent

Ha: X and Y are dependent

$$\chi^2 = \sum_{all\_cells} \frac{(ObservedCount - ExpectedCount)^2}{ExpectedCount}$$

$$p\_value = P(c > \chi^2)$$

where c has chi − square distribution, chi-square(df = (k-1) (s-1))



Compute p-value

Solution1:

```
chi2_stat = np.sum((((Observed_Counts - Expected_Counts)**2)/exp))
p_value = 1 - chi2.cdf(chi2_stat, df)
```

Solution2:

```
chi2_stat , p_value, df, Expected_Counts = chi2_contingency(Observed_Counts , correction=False)
```

- ✓ If p-value < 0.05, then we CAN reject the null hypothesis and accept the alternative hypothesis with significant level 0.05 od confidence level 0.95
- ✓ If p-value > 0.05, then we CANNOT reject the null hypothesis with significant level 0.05 od confidence level 0.95

Example

A researcher wants to know if there's a relationship between gender and drunk-driving. She samples a total of 619 drivers under 20 years of age in a roadside survey.

x: Driver gender (Male/Female)

y: Driver alcohol (Yes/No)

Step 1: Stating the hypotheses

- H_0: There is no relationship between the two categorical variables. (They are independent.)
- H_a: There is a relationship between the two categorical variables. (They are not independent.)

Step 2: Checking the Conditions and Calculating the Test Statistic

	Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

	Yes	No	Total
Male	$\frac{77}{481} = 16\%$	$\frac{404}{481} = 84\%$	100%
Female	$\frac{16}{138} = 11.6\%$	$\frac{112}{138} = 88.4\%$	100%

For the 619 sampled drivers, a larger percentage of males were found to be drunk than females (16.0% vs. 11.6%). Our data, in other words, provide some evidence that drunk driving is related to gender; however, this in itself is not enough to conclude that such a relationship exists in the larger population of drivers under 20. We need to further investigate the data and decide between the following two positions:

- The evidence provided by the roadside survey (16% vs 11.6%) is strong enough to conclude (beyond a reasonable doubt) that it must be due to a relationship between drunk driving and gender in the population of drivers under 20.
- The evidence provided by the roadside survey (16% vs. 11.6%) is not strong enough to make that conclusion, and could have happened just by chance, due to sampling variability, and not necessarily because a relationship exists in the population.

These two different conclusions can be condensed into the two hypotheses below:

- H_0: Drunk driving and gender are independent
- H_a: Drunk driving and gender are not independent

Algebraically, independence between gender and driving drunk is equivalent to having equal proportions who drank (or did not drink) for males vs. females. In fact, the null and alternative hypotheses could have been re-formulated as

- H_0: proportion of male drunk drivers = proportion of female drunk drivers
- H_a: proportion of male drunk drivers ≠ proportion of female drunk drivers

Applying the rule to the first (top left) cell, if driving drunk and gender were independent then:

$$P(drunk\, and\, male) = P(drunk) * P(male)$$

$$P(drunk) = 93 / 619$$

$$P(male) = 481 / 619$$

$$P(drunk\, and\, male) = (93 / 619)*(481 / 619)$$

Therefore, since there are total of 619 drivers, if drunk driving and gender were independent, the count of drunk male drivers that I would expect to see is:

Number of drunk Men = 619* P(drunk and male) = $619 * \frac {93}{619} * \frac {481}{619}$

Similarly:

Number of drunk Women = 619* P(drunk and female) = $619 * \frac {93}{619} * \frac {138}{619}$

Observed Counts

	Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Expected Counts

	Yes	No	Total
Male	$\frac{93 * 481}{619} = 72.3$	$\frac{526 * 481}{619} = 408.7$	481
Female	$\frac{93 * 138}{619} = 20.7$	$\frac{526 * 138}{619} = 117.3$	138
Total	93	526	619

$$\chi^2 = \sum_{all\_cells} \frac{(ObservedCount - ExpectedCount)^2}{ExpectedCount}$$

p-value = The probability of observing χ2 at least as large as the one observed

\chi^2=\Sigma_{all_cells}\frac{(Observed\,Count-Expected\,Count)^2}{Expected\,Count}

The p-value obtained can be interpreted as the probability of observing a \chi^2 test statistic at least as large as the one observed if drunk driving and gender are independent.

Step 3: Given two categorical variables x and y, the p-value can be found as:

```
1 - chi2.cdf(chi2_stat, df)
```

where `chi2_stat` is the \chi^2 test statistic, and df = (n_A-1)(n_B-1) where `n_A` is the number of categories in x and `n_B` is the number of categories in y.

Example

An ice cream shop wants to know whether men and women have different preferences for eating their ice cream out of a cone or a bowl. They take a sample of 500 customers (240 men and 260 women) and ask if they prefer cones over bowls. They found that 124 men preferred cones and 90 women preferred cones. Is there a difference in preference between men and women?

In [ ]:

`pd.crosstab` returns a contingency table:

In [ ]:

Let us work through this slowly to understand the concept of expected counts. Recall that the table of expected counts is what you would expect in each cell of the contingency table if each of the categorical variables of interest were independent, i.e. $\Pr(A \cap B)=\Pr(A)\times \Pr(B)$

If you were to get the number of events where $A \cap B$, multiply the number of events A and B and divide by the total number of events. (Multiply both sides of the equation by the number of elements in the contingency table, and this should become clear.)

In [ ]:

This process is tedious. Fortunately, `scipy.stats.contingency` has an `expected_freq` function that simplifies this:

In [ ]:

Calculate the chi-square test statistic:
$\chi^2=\Sigma_{all\_cells}\frac{(Observed\,Count-Expected\,Count)^2}{Expected\,Count}$

In [ ]:

Notice that the chi-square test statistic quantifies how far away the observed counts are from the expected counts. This makes it similar to the 'greater than' (Case 2) hypotheses tests, and therefore makes it a one-tailed test. The chi-square test has $(r-1)(c-1)$ degrees of freedom.

In [ ]:

The p-value, under the null of independence, is calculated as

In [ ]:

Everything we've done previously can be done in one step on our contingency table using the `scipy.stats.chi2_contingency` function.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Example

Risk Factors for Low Birth Weight

Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for babies with low birth weight. A woman's behavior during pregnancy (including diet, smoking habits, and obtaining prenatal care) can greatly alter her chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

In this exercise, we will use a 1986 study (Hosmer and Lemeshow (2000), Applied Logistic Regression: Second Edition) in which data were collected from 189 women (of whom 59 had low birth weight infants) at the Baystate Medical Center in Springfield, MA (an academic, research, and teaching hospital that serves as the western campus of Tufts University School of Medicine and is the only Level 1 trauma center in western Massachusetts). The goal of the study was to identify risk factors associated with giving birth to a low birth weight baby.

Variables:

- LOW: Low birth weight (0=No (birth weight >= 2500 g) 1=Yes (birth weight < 2500 g)
- AGE: Age of mother (in years)
- LWT: Weight of mother (in pounds)
- RACE: Race of mother (1=White, 2=Black, 3=Other)
- SMOKE: Smoking status during pregnancy (0=No, 1=Yes)
- PTL: History of premature labor (0=None, 1=One, etc.)
- HT: History of hypertension (0=No, 1=Yes)
- FTV: Number of physician visits during the first trimester
- BWT: The actual birth weight (in grams)

Question:

- Q1. Do the data provide evidence that the occurrence of low birth weight is significantly related to whether or not the mother smoked during pregnancy?

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Exercise

Answer Questions 2-4:

- Q2. Do the results of the study provide significant evidence that the race of the mother is a factor in the occurrence of low birth weight?
- Q3. Are there significant differences in age between mothers who gave birth to low weight babies and those whose baby's weight was normal?
- Q4. Are there significant relationship between the actual birth weight and the race of the mother?

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Extra

Recreating ANOVA table from scratch

In [ ]:

Chi-Square

Alternately, you can use the observed and expected values to run the chi-square test of independence using the `chisquare` function from `scipy.stats`.

In [ ]:

The `ravel()` method is needed because without it, `chisquare` will calculate the chi-square statistic for each column.

`scipy.stats.chisquare` takes a **delta** degrees of freedom (ddof) parameter. This is a bit tricky to characterise. Recall that the degrees of freedom of the test of indepdence is df = (r-1)(c-1) where r is the number of rows, and c is the number of columns.

`chisquare` uses a chi-square distribution with k-1-ddof degrees of freedom, where k = rc, the number of frequencies observed. With a bit of algebraic manipulation, we obtain ddof as r+c-2.

(r-1)(c-1)=rc-r-c+1= rc-(r+c-2)-2+1=rc-1-(r+c-2)

`scipy.stats.chisquare` returns the chi-square statistic and the p-value.