Probability & Statistics for DS & AI

Making statistical decisions

Michele Guindani

JioInstitute

Summer 2022

Testing hypotheses

- In many situations, we want to evaluate data to make informed decisions

Testing hypotheses

- In many situations, we want to evaluate data to make informed decisions

Example (Brass Alloy zinc percentage)

- A corrosion resistant brass alloy, widely used in plumbing fixtures, is composed mainly of copper, tin, lead, zinc, nickel, and iron in decreasing amounts.
- The addition of zinc to the alloy produces a jump in metal strength Interest lies in the percentage of zinc so that it can be adjusted to be within specifications.
- We have data on the zinc percentage in 12 alloy samples were tested using spectrometry by the St. Paul Brass and Aluminum Foundry in St. Paul, Minnesota, in June of 2003 .

```
import numpy as np
np.random.seed(1234)
X = np.array([4.20,4.36,4.11,3.96,
        5.63,4.50,5.64,4.38,
        4.45,3.67,5.26,4.66])
N = X.size
```

Example

- The chief operation officer reports that the zinc percentage should be centered around $m_0 = 4.75$.
- They also mentioned that - in any case - the percentage should not be less thant 4.5% or more than 5%.
- How can we decide if the zinc percentage is within specification?

Example (Passport office)

A passport office claims that the passport applications are processed within 30 days of submitting the application form and all necessary documents.

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt

df=pd.read_csv('passport.csv')
df.head()
```

	processing_time
0	16.0
1	16.0
2	30.0
3	37.0
4	25.0

Is the claim made by the passport office accurate?

A/B testing

- A large consumer bank has recently run a direct marketing campaign by creating a video ad for credit card offers to acquire more credit card applications.
- During the campaign, they also ran a split test for landing pages. The control page is their default text-based page, while the test page features a new marketing video.

A/B testing

- A large consumer bank has recently run a direct marketing campaign by creating a video ad for credit card offers to acquire more credit card applications.
- During the campaign, they also ran a split test for landing pages. The control page is their default text-based page, while the test page features a new marketing video.

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt

df=pd.read_csv('advertisement_clicks.csv')
df.head()
```

	advertisement_id	action
0	B	1
1	B	1
2	A	0
3	B	0
4	A	1

A/B testing

- A large consumer bank has recently run a direct marketing campaign by creating a video ad for credit card offers to acquire more credit card applications.
- During the campaign, they also ran a split test for landing pages. The control page is their default text-based page, while the test page features a new marketing video.

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt

df=pd.read_csv('advertisement_clicks.csv')
df.head()
```

	advertisement_id	action
0	B	1
1	B	1
2	A	0
3	B	0
4	A	1

What's the advertising campaign that got more clicks?

Example (Nurses' Health Study (Colditz et al., 1990))

The study sought to estimate and compare the rates of breast cancer for 50- to 59-year-old postmenopausal women who were current users of estrogen replacement therapy (group 1) and who were not (group 2).

Group	Cases	Person-years
1-Hormone Therapy	123	46,524
None	288	145,159

- We are interested in the number of cases over the persons-year in each group. The persons-year is a measurement combining the number of persons and their time contribution in a study. It is the sum of individual units of time that the persons in the study population have been exposed or at risk to the conditions of interest

Example (Nurses' Health Study (Colditz et al., 1990))

- We can consider a Binomial distribution, but given the large number of trials, a Poisson distribution can be considered
- The rates in the two groups need to take into account the *sizes* of the two groups to be comparable:

$$\lambda_1 = \theta_1 M_1 \qquad \lambda_2 = \theta_2 M_2$$

where M_1 and M_2 denote denote the persons-year. Then θ_1 and θ_2 are called incidence rates.

- Is θ_1 (the incidence rate of breast cancer in the first group) lower than θ_2?

Hypothesis testing

- The procedure for testing whether a hypothesis should be accepted or rejected is known as hypothesis testing. In hypothesis testing, we often have two opposite hypotheses:
- H_0 : Null hypothesis. It is the "status quo", or the current status.

Hypothesis testing

- The procedure for testing whether a hypothesis should be accepted or rejected is known as hypothesis testing. In hypothesis testing, we often have two opposite hypotheses:
- H_0 : Null hypothesis. It is the "status quo", or the current status.
- H_1 : Alternative hypothesis It is the alternative to the null hypothesis.

Hypothesis testing

- The procedure for testing whether a hypothesis should be accepted or rejected is known as hypothesis testing. In hypothesis testing, we often have two opposite hypotheses:
- H_0 : Null hypothesis. It is the "status quo", or the current status.
- H_1 : Alternative hypothesis It is the alternative to the null hypothesis.
- The null hypothesis is the default assumption:
  - In a courthouse, by default, any person being prosecuted is assumed to be innocent. The police need to show sufficient evidence in order to prove the person guilty.
  - Hypothesis testing asks whether we have strong enough evidence to reject the null hypothesis. If our evidence is not strong enough, we must assume that the null hypothesis is possibly true.

Example (Coin tossing)

- $H_0 : \theta = 0.5$, and $H_1 : \theta > 0.5$. This is a one-sided alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta < 0.5$. This is another one-sided alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta \neq 0.5$. This is a two-sided alternative.

Example (Coin tossing)

- $H_0 : \theta = 0.5$, and $H_1 : \theta > 0.5$. This is a one-sided alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta < 0.5$. This is another one-sided alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta \neq 0.5$. This is a two-sided alternative.

Example

- Walmart wants to test if self-checkout is faster than using a cashier (traditional method).

Example (Coin tossing)

- $H_0 : \theta = 0.5$, and $H_1 : \theta > 0.5$. This is a one-sided alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta < 0.5$. This is another one-sided alternative.
- $H_0 : \theta = 0.5$, and $H_1 : \theta \neq 0.5$. This is a two-sided alternative.

Example

- Walmart wants to test if self-checkout is faster than using a cashier (traditional method).
- Let θ be the proportion of people that check out faster with self-checkout.

$$H_0 : \theta \leq 0.5, \text{ and } H_1 : \theta > 0.5$$

This is a one-sided alternative. Check Practice Exercise 9.4 in your textbook (which gets it wrong)

Example (Brass Alloy Zinc Percentage)

Let θ be the (unknown) true proportion of zinc in the brass alloy

$$H_0 : \theta = 4.75\% \quad \text{and} \quad H_1 : \theta \neq 4.75\%$$

Example (Brass Alloy Zinc Percentage)

Let θ be the (unknown) true proportion of zinc in the brass alloy

$$H_0 : \theta = 4.75\% \quad \text{and} \quad H_1 : \theta \neq 4.75\%$$

Example (Passport office)

Let θ be the expected number of applications processed by the passport office

$$H_0 : \theta \leq 30 \quad \text{and} \quad H_1 : \theta > 30$$

This is one sided alternative

Example (A/B testing)

Let θ_1 denote the probability that someone exposed to advertisement campaign A clicks on the ad, and θ_2 the probability that someone exposed to advertisement campaign B clicks on the ad.

Depending on the situation you may be interested in testing different hypotheses

Example (A/B testing)

Let θ_1 denote the probability that someone exposed to advertisement campaign A clicks on the ad, and θ_2 the probability that someone exposed to advertisement campaign B clicks on the ad.

Depending on the situation you may be interested in testing different hypotheses

For example, if advertisement A is a consolidated advertisement campaign

$$H_0 : \theta_1 \geq \theta_2 \quad \text{and} \quad H_1 : \theta_1 < \theta_2$$

Example (A/B testing)

Let θ_1 denote the probability that someone exposed to advertisement campaign A clicks on the ad, and θ_2 the probability that someone exposed to advertisement campaign B clicks on the ad.

Depending on the situation you may be interested in testing different hypotheses

For example, if advertisement A is a consolidated advertisement campaign

$$H_0 : \theta_1 \geq \theta_2 \quad \text{and} \quad H_1 : \theta_1 < \theta_2$$

Alternatively, if the advertisement campaigns are run concurrently for the first time, you may want first to test:

$$H_0 : \theta_1 = \theta_2 \quad \text{and} \quad H_1 : \theta_1 \neq \theta_2$$

Example (Nurses' Health Study)

- Recall, we were interested in the incidence rates:

$$\lambda_1 = \theta_1 M_1 \quad \lambda_2 = \theta_2 M_2$$

- Here, since estrogen replacement theory was an experimental treatment,

$$H_0 : \theta_1 \leq \theta_2 \quad \text{and} \quad H_1 : \theta_1 > \theta_2$$

Parametric and Nonparametric tests

- Hypothesis tests are broadly classified into parametric tests and non-parametric tests.

  Parametric tests are about population parameters of a distribution such as mean, proportion, standard deviation, etc.

  Non-parametric tests are not about parameters, but about other characteristics such as independence of events or data following certain distributions such as normal distribution.

- Here, we focus first on parametric tests.

Decision outcomes

- There are four possible outcomes when making hypothesis test decisions from sample data:

		THE TRUTH	
		The null hypothesis (H_0) is true (H_a is false)	**The null hypothesis (H_0) is not true** (H_a is true)
THE DECISION THE ANALYST MAKES	**Reject H_0** (support H_a)	**TYPE I (α) error/ Alpha Risk/ p – value** **Overreacting**	**Correct Decision** **(1 - β)** **Power of the test**
	Fail to Reject H_0 (do not support Ha)	**Correct Decision** **(1 - α) = the Confidence level of the test**	**TYPE II (β) error/ Beta Risk** **Underreacting**

Decision outcomes

- Among the possible outcomes, over-reacting (false positive decision to change the status quo) is often seen as the most troubling (often associated with increased costs).

⚠ Big issue: we can't minimize both types of errors at the same time!!

Decision outcomes

- Among the possible outcomes, over-reacting (false positive decision to change the status quo) is often seen as the most troubling (often associated with increased costs).

⚠ Big issue: we can't minimize both types of errors at the same time!!

- Frequentist statistical decision making techniques then do the following:

fix the probability of Type I error: α,

e.g., we can choose $\alpha = 0.05$ (very small)

- and then

minimize the probability of Type II error: β

⚠ This means that we believe that one error is more serious than the other, so we want to control it!

Steps for Hypothesis testing

- The steps for hypothesis tests are as follows:

① Define the null and alternative hypotheses ✅

Steps for Hypothesis testing

🤓 The steps for hypothesis tests are as follows:

① Define the null and alternative hypotheses ✅

② Decide the criteria for rejection and retention of null hypothesis:

- How much probability of the type I error are we OK to tolerate in this decision problem?

We often call α = Probability of Type I error significance value.
Typical value used for α is 0.05. ✅

Steps for Hypothesis testing

🤓 The steps for hypothesis tests are as follows:

① Define the null and alternative hypotheses ✅

② Decide the criteria for rejection and retention of null hypothesis:

- How much probability of the type I error are we OK to tolerate in this decision problem?

We often call α = Probability of Type I error significance value.
Typical value used for α is 0.05. ✅

③ Identify the test statistics to be used for testing the validity of the null hypothesis.

③ Identify the test statistics to be used for testing the validity of the null hypothesis.

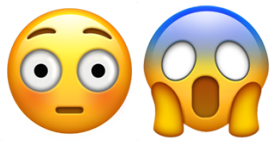
🤔 A test statistic is a random variable that is calculated from sample data (like the estimators!) but under the null hypothesis H_0, the distribution of the test statistics does not depend on the population parameters

③ Identify the test statistics to be used for testing the validity of the null hypothesis.

🤔 A test statistic is a random variable that is calculated from sample data (like the estimators!) but under the null hypothesis H_0, the distribution of the test statistics does not depend on the population parameters

③ Identify the test statistics to be used for testing the validity of the null hypothesis.

🤔 A test statistic is a random variable that is calculated from sample data (like the estimators!) but under the null hypothesis H_0, the distribution of the test statistics does not depend on the population parameters



③ Identify the test statistics to be used for testing the validity of the null hypothesis.

🤔 A test statistic is a random variable that is calculated from sample data (like the estimators!) but under the null hypothesis H_0, the distribution of the test statistics does not depend on the population parameters

③ Identify the test statistics to be used for testing the validity of the null hypothesis.

🤔 A test statistic is a random variable that is calculated from sample data (like the estimators!) but under the null hypothesis H_0, the distribution of the test statistics does not depend on the population parameters

⇨ No worries, we have already seen examples of test statistics, e.g. for testing $H_0 : \mu = \mu_0$

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

③ Identify the test statistics to be used for testing the validity of the null hypothesis.

🤔 A test statistic is a random variable that is calculated from sample data (like the estimators!) but under the null hypothesis H_0, the distribution of the test statistics does not depend on the population parameters

⇨ No worries, we have already seen examples of test statistics, e.g. for testing $H_0 : \mu = \mu_0$

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

③ Identify the test statistics to be used for testing the validity of the null hypothesis.

🤔 A test statistic is a random variable that is calculated from sample data (like the estimators!) but under the null hypothesis H_0, the distribution of the test statistics does not depend on the population parameters

⇨ No worries, we have already seen examples of test statistics, e.g. for testing $H_0 : \mu = \mu_0$

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

⚠ The choice of the test statistic depends on the problem and questions we are facing.

Test Statistics for common problems

Test For	*Null Hypothesis (H_0)*	*Test Statistic*	*Distribution*	*Use When*
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{\sigma/\sqrt{n}}$	Z	Normal distribution or $n > 30$; σ known
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{s/\sqrt{n}}$	t_{n-1}	$n < 30$, and/or σ unknown
Population proportion (p)	$p = p_0$	$\frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$	Z	$n\hat{p}$, $n(1-\hat{p}) \geq 10$
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z	Both normal distributions, or n_1, $n_2 \geq 30$; σ_1, σ_2 known
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t distribution with df = the smaller of n_1-1 and n_2-1	n_1, $n_2 < 30$; and/or σ_1, σ_2 unknown
Mean difference μ_d (paired data)	$\mu_d = 0$	$\frac{(\bar{d} - \mu_d)}{s_d/\sqrt{n}}$	t_{n-1}	$n < 30$ pairs of data and/or σ_d unknown
Difference of two proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Z	$n\hat{p}$, $n(1-\hat{p}) \geq 10$ for each group

④ Steps for Hypothesis testing: calculate the p-value

④ Steps for Hypothesis testing: calculate the p-value

The p-value (probability value) provides the answer to the following question:

If the null hypothesis is true, what is the probability that we would observe a more extreme value of the test statistic in the direction of the alternative hypothesis than we did?

④ Steps for Hypothesis testing: calculate the p-value

The p-value (probability value) provides the answer to the following question:

If the null hypothesis is true, what is the probability that we would observe a more extreme value of the test statistic in the direction of the alternative hypothesis than we did?

④ Steps for Hypothesis testing: calculate the p-value

The p-value (probability value) provides the answer to the following question:

If the null hypothesis is true, what is the probability that we would observe a more extreme value of the test statistic in the direction of the alternative hypothesis than we did?

No worries, this simply means:

If the null hypothesis is true, how likely it is that your data would have occurred by random chance?

This question is very similar to the question answered in criminal trials: "If the defendant is innocent, what is the chance that we would observe such extreme criminal evidence?"

④ Steps for Hypothesis testing: calculate the p-value

The p-value (probability value) provides the answer to the following question:

If the null hypothesis is true, what is the probability that we would observe a more extreme value of the test statistic in the direction of the alternative hypothesis than we did?

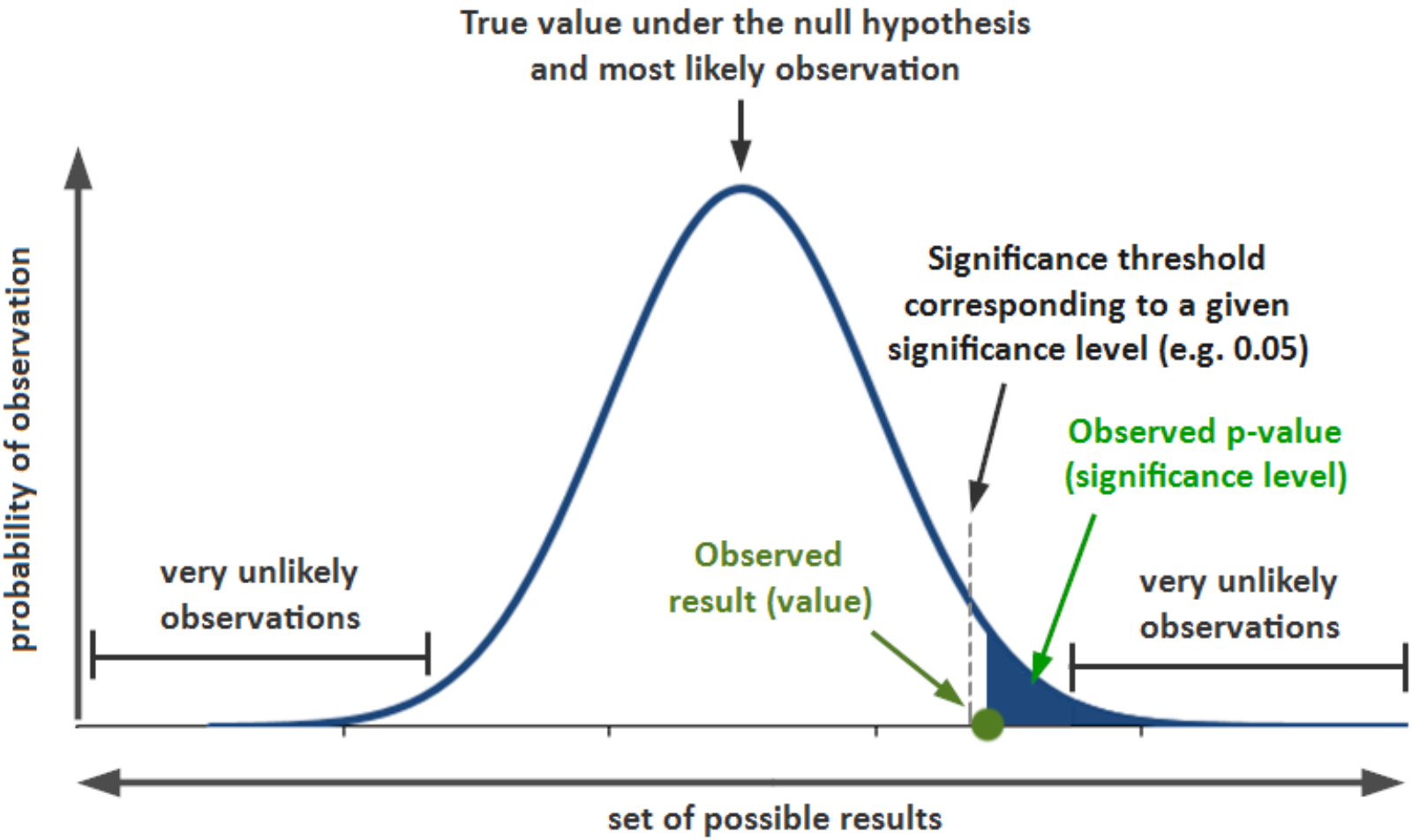
No worries, this simply means:

If the null hypothesis is true, how likely it is that your data would have occurred by random chance?

This question is very similar to the question answered in criminal trials: "If the defendant is innocent, what is the chance that we would observe such extreme criminal evidence?" ☺️

- We will use the functions provided in scipy.stats module for calculating the p-value.

P-values


True value under the null hypothesis
and most likely observation
Significance threshold
corresponding to a given
significance level (e.g. 0.05)
Observed p-value
(significance level)
Observed
result (value)
very unlikely
observations
very unlikely
observations
probability of observation
set of possible results

⑤ Take a decision

- The last step in Hypothesis testing is to take the decision to reject or fail to reject the null hypothesis based on the p-value and the significance value α:

  Decision rule:

  If p-value $\leq \alpha \Rightarrow$ Reject H_0

  If p-value $> \alpha \Rightarrow$ Fail to Reject H_0

  where α is the significance level.

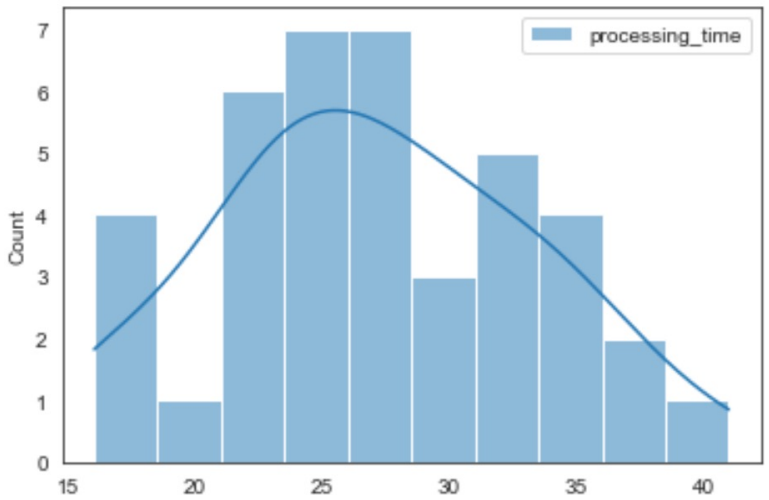
Example (One sample Z-test: Passport office)

Recall, θ is the expected number of applications processed by the passport office and we are interested in testing the one-sided hypothesis:

$$H_0 : \theta \leq 30 \text{ and } H_1 : \theta \geq 30$$

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns

df=pd.read_csv('passport.csv')
sns.histplot(data = df, kde = True, bins=10)
```

```
<AxesSubplot:ylabel='Count'>
```



Example (One sample Z-test: Passport office)

In order to take the decision, we use the function ztest in the module

`statsmodels.stats.weightstats.ztes`

of Python. See the documentation here (snippets, below).

```
from statsmodels.stats import weightstats as stests
ztest,probability_value = stests.ztest(df, x2=None, value=30, alternative="larger")
print(float(probability_value))
if probability_value<0.05:
    print("Null hyphothesis rejected , Alternative hyphothesis accepted")
else:
    print("Null hyphothesis failed to be rejected, Alternative hyphothesis rejected")
```

```
0.998589957501341
Null hyphothesis failed to be rejected, Alternative hyphothesis rejected
```

value : `float`

In the one sample case, value is the mean of x1 under the Null hypothesis. In the two sample case, value is the difference between mean of x1 and mean of x2 under the Null hypothesis. The test statistic is *x1_mean - x2_mean - value*.

alternative : `str`

The alternative hypothesis, H1, has to be one of the following

> 'two-sided': H1: difference in means not equal to value (default) 'larger' : H1: difference in means larger than value 'smaller' : H1: difference in means smaller than value

Example (One Sample T-test: Brass Alloy Zinc percentage)

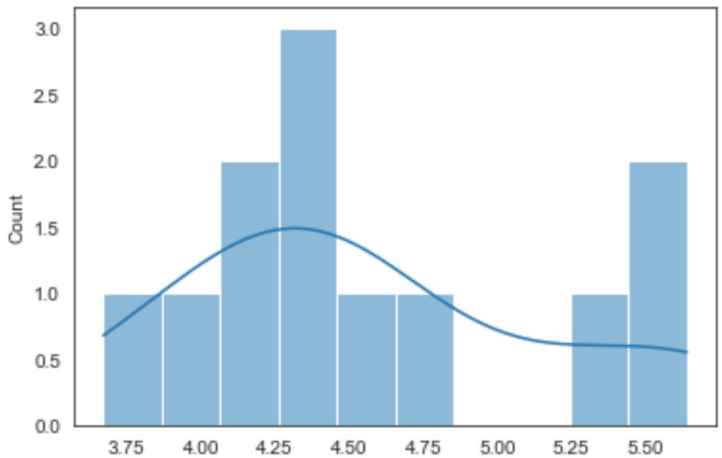
Recall, θ is the (unknown) true proportion of zinc in the brass alloy

$$H_0 : \theta = 4.75\% \quad \text{and } H_1 : \theta \neq 4.75\%$$

This is a two sided test.

```
import numpy as np
import seaborn as sns
sns.set_style("white")
np.random.seed(1234)
X = np.array([4.20,4.36,4.11,3.96,
              5.63,4.50,5.64,4.38,
              4.45,3.67,5.26,4.66])
N = X.size
sns.histplot(data = X, kde = True, bins=10)
```

```
<AxesSubplot:ylabel='Count'>
```



This may be a stretch given that we didn't collect many observations, so it is not apparent from the histogram, but
let's consider the data as approximately gaussian

Example (One Sample T-test: Brass Alloy Zinc percentage)

- We only have 12 points (and we need to estimate the variance!). So the normal Z doesn't work as our test statistic ⇨ We use a T distribution!

- Here, we use the function ttest_1samp in scipy.stats. See documentation here (under statistical tests)

```
from scipy.stats import ttest_1samp
statistic ,probability_value = ttest_1samp(X, popmean=4.75, alternative="two-sided")
print(float(probability_value))
if probability_value<0.05:
    print("Null hyphothesis rejected , Alternative hyphothesis accepted")
else:
    print("Null hyphothesis failed to be rejected, Alternative hyphothesis rejected")
```

```
0.3401986328878056
Null hyphothesis failed to be rejected, Alternative hyphothesis rejected
```

scipy.stats convention for the alternative hypothesis

alternative : *{'two-sided', 'less', 'greater'}, optional*

Defines the alternative hypothesis. The following options are available (default is 'two-sided'):

- 'two-sided': the means of the distributions underlying the samples are unequal.
- 'less': the mean of the distribution underlying the first sample is less than the mean of the distribution underlying the second sample.
- 'greater': the mean of the distribution underlying the first sample is greater than the mean of the distribution underlying the second sample.

Example (Bollywood Movie Cost of Production)

Aravind Productions (AP) is a recently formed movie production house based out of Mumbai, India.

AP was interested in understanding the production cost required for producing a Bollywood movie.

The industry believes that the production house will require INR 500 million (50 crore) on average.

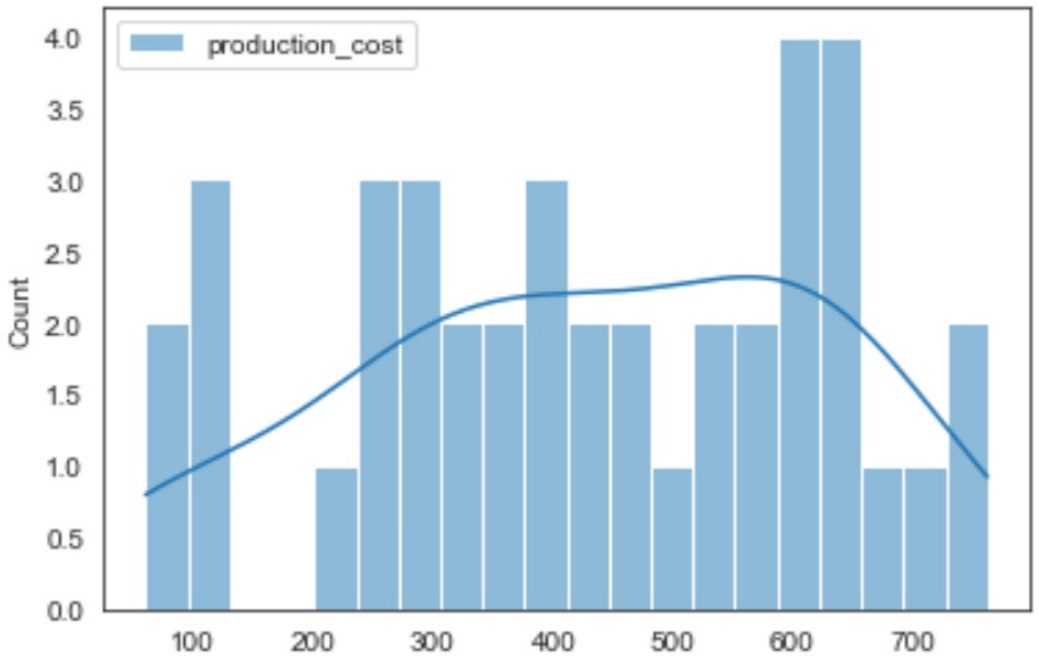
We have data on the production costs of 40 Bollywood movies.

Conduct an appropriate hypothesis test at $\alpha = 0.05$ to check whether the belief about average production cost is correct.

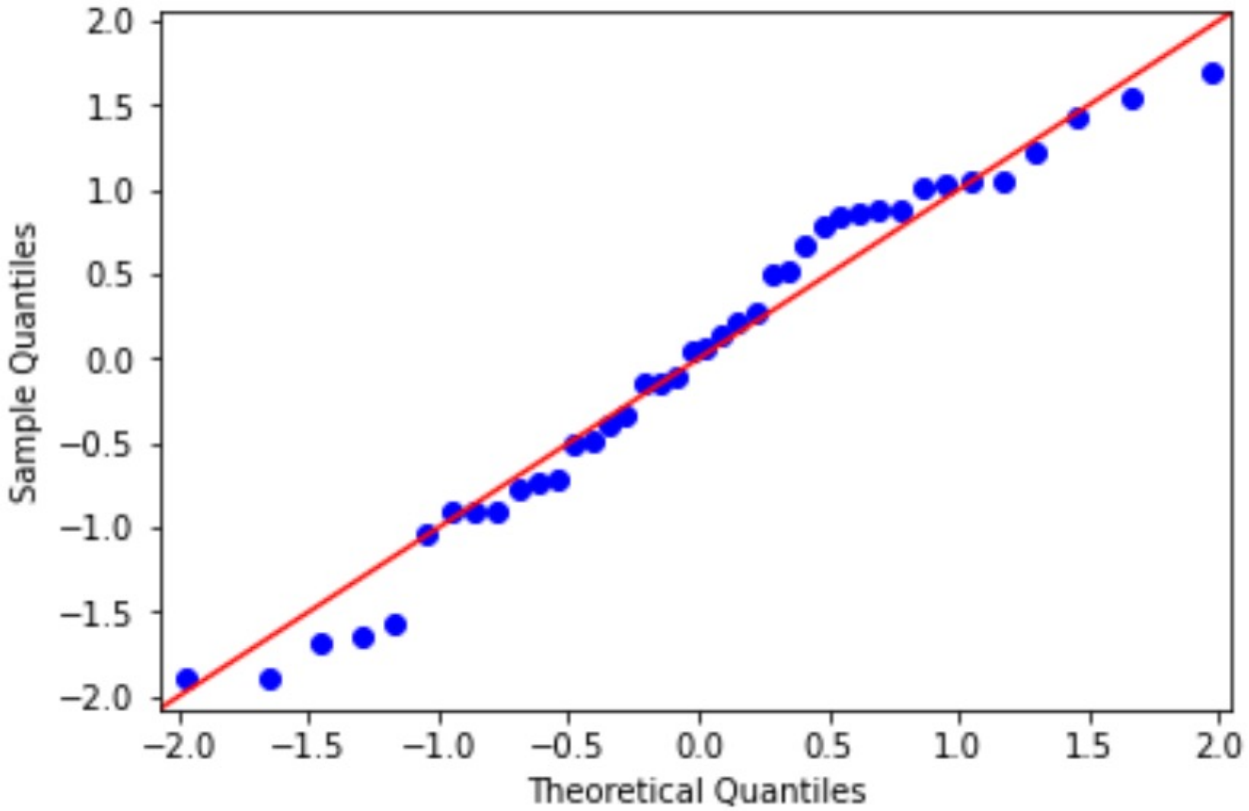
```
df=pd.read_csv('bollywoodmovies.csv')
df.head()
print(df.size)
sns.histplot(data = df, kde = True, bins=20)
```

```
40
```

```
<AxesSubplot:ylabel='Count'>
```



```
fig = sm.qqplot(df['production_cost'], line='45')
plt.show()
```



Example (Bollywood Movie Cost of Production)

We want to test the hypotheses:

$$H_0 : \mu = 500 \qquad \text{vs} \qquad H_A : \mu \neq 500$$

We assume a Gaussian distribution for these data. There are 40 samples, so we can consider either a T distribution or approximating it with a normal.

```python
from statsmodels.stats import weightstats as stests
ztest,probability_value = stests.ztest(df, x2=None, value=500, alternative="two-sided")
print(float(probability_value))
if probability_value<0.05:
    print("Null hyphothesis rejected , Alternative hyphothesis accepted")
else:
    print("Null hyphothesis failed to be rejected, Alternative hyphothesis rejected")
```

```
0.02233903464768864
Null hyphothesis rejected , Alternative hyphothesis accepted
```

```python
from scipy.stats import ttest_1samp
statistic ,probability_value = ttest_1samp(df, popmean=500, alternative="two-sided")
print(float(probability_value))
if probability_value<0.05:
    print("Null hyphothesis rejected , Alternative hyphothesis accepted")
else:
    print("Null hyphothesis failed to be rejected, Alternative hyphothesis rejected")
```

```
0.0278625564067618
Null hyphothesis rejected , Alternative hyphothesis accepted
```

The results are similar, but slightly different (why?). The decision is the same.

What p-values are not:

⚠ Incorrect interpretations of P-values are very common:

- They are NOT the probability of making a mistake by rejecting a true null hypothesis.
  - P values are calculated based on the assumptions that the null is true for the population
- The p-values are NOT continuous measures of statistical evidence. A smaller p-value doesn't mean a more "significant" evidence that the null hypothesis needs to be rejected.
  - If the null-hypothesis is true, P values are uniformly distributed, and it is just as likely to observe a P value of 0.001 as it is to observe a P value of 0.999.

See a nice discussion here.

Probability & Statistics for DS & AI

More examples: two-sample tests

Michele Guindani

JioInstitute

Summer 2022

Two-sample t-test

A two-sample t-test is required to test difference between two population means where standard deviations are unknown. The parameters are estimated from the samples. A pooled variance between the sample is computed to estimate the variance of the test statistic.

Two-sample t-test

A two-sample t-test is required to test difference between two population means where standard deviations are unknown. The parameters are estimated from the samples. A pooled variance between the sample is computed to estimate the variance of the test statistic.

Example (Clinical trial)

An experimental treatment is given to a group of patients, whereas another group of patients receive placebo. We assume that the assignment of patients between the two groups is randomized (RCT). Suppose that, at the end of the study, a measurement (e.g. blood sugar levels) is checked on patients in both groups.

Two-sample t-test

A two-sample t-test is required to test difference between two population means where standard deviations are unknown. The parameters are estimated from the samples. A pooled variance between the sample is computed to estimate the variance of the test statistic.

Example (Clinical trial)

An experimental treatment is given to a group of patients, whereas another group of patients receive placebo. We assume that the assignment of patients between the two groups is randomized (RCT). Suppose that, at the end of the study, a measurement (e.g. blood sugar levels) is checked on patients in both groups.

A question of interest may be:

- Is the experimental treatment leading to different levels of blood sugar in the treated group (say, group 1) vs the placebo group (say, group 2)?

Example (Clinical trial)

The hypothesis test of interest then is:

$$H_0 : \theta_1 \geq \theta_2 \quad \text{and} \quad H_1 : \theta_1 < \theta_2$$

Importantly the treatment is given to two different (independent) populations.

Example (Clinical trial)

The hypothesis test of interest then is:

$$H_0 : \theta_1 \geq \theta_2 \quad \text{and} \quad H_1 : \theta_1 < \theta_2$$

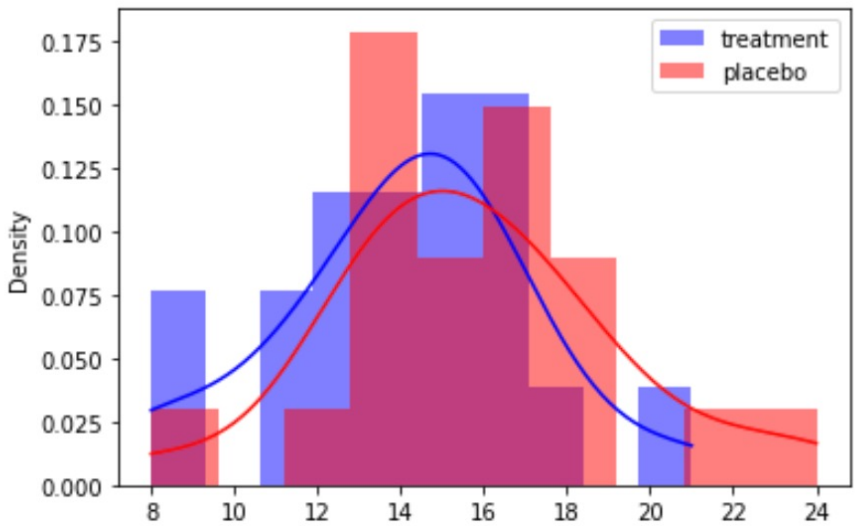
Importantly the treatment is given to two different (independent) populations.

```
data_group1 = np.array([14, 15, 15, 16, 13, 8, 14,
                        17, 8, 12, 18, 11, 21, 15,
                        15, 16, 16, 13, 14, 11 ])
data_group2 = np.array([15, 17, 14, 17, 14, 8, 12,
                        19, 19, 14, 17, 22, 24, 16,
                        13, 16, 13, 18, 15, 13, 15])

#Print the dimension of each vector
print(np.size(data_group1), np.size(data_group2))

#Print the dimension of each vector
print(np.mean(data_group1), np.mean(data_group2))

# Print the variance of both data groups
print(np.var(data_group1), np.var(data_group2))
```

```
20 21
14.1 15.761904761904763
9.290000000000001 11.705215419501135
```

Example (Clinical trial)

```
sns.histplot(data_group1, label ='treatment', kde=True,
             stat="density", linewidth=0, color="blue", bins=10)
sns.histplot(data_group2, label  ='placebo', kde=True,
             stat="density", linewidth=0, color="red", bins=10)
plt.legend();
```



Example (Clinical trial)

```
import scipy.stats as stats
statistic ,probability_value = stats.ttest_ind(data_group1,data_group2, equal_var=False, alternative="less")
print(float(probability_value))
if probability_value<0.05:
    print("Null hyphothesis rejected , Alternative hyphothesis accepted")
else:
    print("Null hyphothesis failed to be rejected, Alternative hyphothesis rejected")
```

```
0.05846609343579387
Null hyphothesis failed to be rejected, Alternative hyphothesis rejected
```

Example (A/B testing)

In this case, the experiment is a binomial experiment.

Example (A/B testing)

In this case, the experiment is a binomial experiment. Our outcome is the proportion of people that click on an ad in each group. Say, out of n_A subjects who have seen ad A, only y_A have clicked on it. So, our data are represented by the proportion y_A/n_A
Similarly, for ad B, we have y_B/n_B.

Example (A/B testing)

In this case, the experiment is a binomial experiment. Our outcome is the proportion of people that click on an ad in each group. Say, out of n_A subjects who have seen ad A, only y_A have clicked on it. So, our data are represented by the proportion y_A/n_A
Similarly, for ad B, we have y_B/n_B.

Working directly with the binomial is a bit of a hassle 😨

Example (A/B testing)

In this case, the experiment is a binomial experiment. Our outcome is the proportion of people that click on an ad in each group. Say, out of n_A subjects who have seen ad A, only y_A have clicked on it. So, our data are represented by the proportion y_A/n_A
Similarly, for ad B, we have y_B/n_B.

Working directly with the binomial is a bit of a hassle 😨

However, we know that for a large number of trials, the binomial distribution can be approximated by a normal distribution 🤩

Example (A/B testing)

In this case, the experiment is a binomial experiment. Our outcome is the proportion of people that click on an ad in each group. Say, out of n_A subjects who have seen ad A, only y_A have clicked on it. So, our data are represented by the proportion y_A/n_A
Similarly, for ad B, we have y_B/n_B.

Working directly with the binomial is a bit of a hassle 😨

However, we know that for a large number of trials, the binomial distribution can be approximated by a normal distribution 🤩

So to compare the two proportions we can consider the statistic

$$Z = \frac{(\hat{p}_A - \hat{p}_B) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with $\hat{p}$ is the total pooled proportion calculated as:

$$\hat{p} = (p_1 n_1 + p_2 n_2) \,/\, (n_1 + n_2)$$

Example (A/B testing)

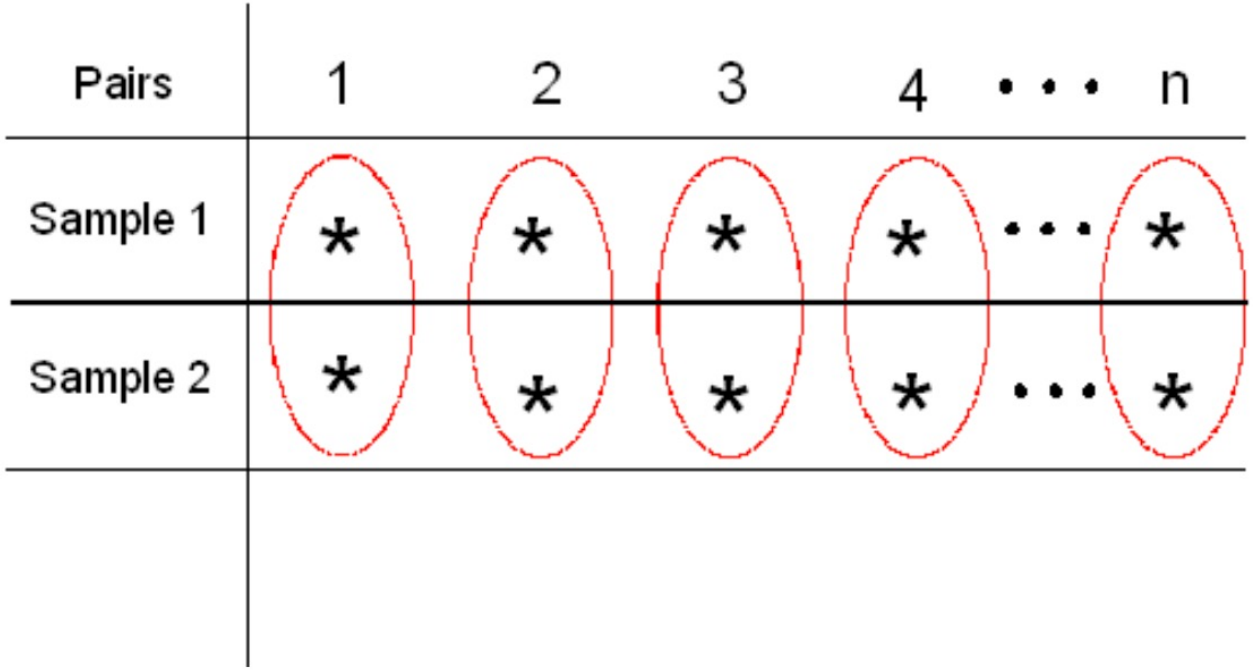
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import scipy.stats as stats

df=pd.read_csv('advertisement_clicks.csv')
df_A=df.loc[df["advertisement_id"]=="A"]
df_B=df.loc[df["advertisement_id"]=="B"]
p_hat_A=df_A['action'].sum()/df_A['action'].count()
p_hat_B=df_B['action'].sum()/df_B['action'].count()
pooled_p_hat=(df_A['action'].sum()+df_B['action'].sum())/(df_A['action'].count()+df_B['action'].count())
pooled_std_hat=math.sqrt((pooled_p_hat)*(1-pooled_p_hat)*(1/df_A['action'].count()+1/df_B['action'].count()))
mudiff=0
z = ((p_hat_A - p_hat_B) - mudiff)/pooled_std_hat
pval = 2*(stats.norm.sf(abs(z)))
print(float(pval))

from statsmodels.stats.proportion import proportions_ztest

successes = np.array([df_A['action'].sum(), df_B['action'].sum()])
samples = np.array([df_A['action'].count(), df_B['action'].count()])
stat, p_value = proportions_ztest(count=successes, nobs=samples,  alternative='two-sided')
print(p_value)
```

```
0.0013069502732125403
0.0013069502732125403
```

Paired samples t-test

- In the previous example the two groups were necessarily measuring different individuals (placebo versus a drug study).
- However sometimes we measure the same person twice (pre-test, post-test) for instance. In that case we use the "paired" (matched) samples t-test because there is a natural link between individual numbers in both groups (usually a person or some other unit of measurement).
- In this case, the number of measurements pre- and post- are the same.

Indeed, in every matched pairs problem, our data consist of 2 samples which are organized in n pairs:



- The test proceeds by reducing the two samples to only one by calculating the difference between the two observations for each pair.

Paired samples t-test

Example (Blood pressure pre- and post- intervention)

The following dataset contains blood pressure readings before and after an intervention. We are interested in knowing if there is a difference in the readings pre- and post-.

```
df = pd.read_csv("blood_pressure.csv")

df[['bp_before','bp_after']].describe()
```

	bp_before	bp_after
count	120.000000	120.000000
mean	156.450000	151.358333
std	11.389845	14.177622
min	138.000000	125.000000
25%	147.000000	140.750000
50%	154.500000	149.500000
75%	164.000000	161.000000
max	185.000000	185.000000

```
df.head()
```

	patient	sex	agegrp	bp_before	bp_after
0	1	Male	30-45	143	153
1	2	Male	30-45	163	170
2	3	Male	30-45	153	168
3	4	Male	30-45	153	142
4	5	Male	30-45	146	141

```
stat, pval = stats.ttest_rel(df['bp_before'], df['bp_after'])

print(pval)
if pval<0.05:
    print("Null hyphothesis rejected , Alternative hyphothesis accepted")
else:
    print("Null hyphothesis failed to be rejected, Alternative hyphothesis rejected")
```

```
0.0011297914644840823
Null hyphothesis rejected , Alternative hyphothesis accepted
```

Review

Test For	Null Hypothesis (H_0)	Test Statistic	Distribution	Use When
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{\sigma/\sqrt{n}}$	Z	Normal distribution or $n > 30$; σ known
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{s/\sqrt{n}}$	t_{n-1}	$n < 30$, and/or σ unknown
Population proportion (p)	$p = p_0$	$\frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$	Z	$n\hat{p},\ n(1-\hat{p}) \geq 10$
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z	Both normal distributions, or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t distribution with df = the smaller of $n_1 - 1$ and $n_2 - 1$	$n_1, n_2 < 30$; and/or σ_1, σ_2 unknown
Mean difference μ_d (paired data)	$\mu_d = 0$	$\frac{(\bar{d} - \mu_d)}{s_d/\sqrt{n}}$	t_{n-1}	$n < 30$ pairs of data and/or σ_d unknown
Difference of two proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	Z	$n\hat{p},\ n(1-\hat{p}) \geq 10$ for each group

One sample Z test
statsmodels.stats.weightstats.ztest

One Sample T test:
scypy.stats.ttest_1samp

One Sample Proportion
statsmodels.stats.proportion.proportions_ztest

Two Sample Ztest
statsmodels.stats.weightstats.ztest

Two Sample Ttest
scypy.stats.ttest_ind

Paired Sample Ttest
scipy.stats.ttest_rel

Two Sample proportion
statsmodels.stats.proportion.proportions_ztest

Probability & Statistics for DS & AI

Non-normal data

Michele Guindani

JioInstitute

Summer 2022

Wilcoxon or Mann-Whitney test

- All the tests above assume the data is normally distributed.
- In cases where there is strong suspicion your data are not normally distributed in one or the other group of a two sample test, one can use a non-parametric test known as the Wilcoxon (paired samples/within subject) or Mann-Whitney (independent samples) test.

Wilcoxon or Mann-Whitney test

- All the tests above assume the data is normally distributed.
- In cases where there is strong suspicion your data are not normally distributed in one or the other group of a two sample test, one can use a non-parametric test known as the Wilcoxon (paired samples/within subject) or Mann-Whitney (independent samples) test.

```
df = pd.read_csv("blood_pressure.csv")

stats.wilcoxon(x=df['bp_before'], y=df['bp_after'],
                                  alternative='two-sided')
```

```
WilcoxonResult(statistic=2234.5, pvalue=0.0014107333565442858)
```

Wilcoxon or Mann-Whitney test

- All the tests above assume the data is normally distributed.
- In cases where there is strong suspicion your data are not normally distributed in one or the other group of a two sample test, one can use a non-parametric test known as the Wilcoxon (paired samples/within subject) or Mann-Whitney (independent samples) test.

```
df = pd.read_csv("blood_pressure.csv")

stats.wilcoxon(x=df['bp_before'], y=df['bp_after'],
                                    alternative='two-sided')
```

```
WilcoxonResult(statistic=2234.5, pvalue=0.0014107333565442858)
```

```
stats.mannwhitneyu(data_group1,data_group2, alternative="two-sided")
```

```
MannwhitneyuResult(statistic=155.0, pvalue=0.15249157824623777)
```

Wilcoxon or Mann-Whitney test

- All the tests above assume the data is normally distributed.
- In cases where there is strong suspicion your data are not normally distributed in one or the other group of a two sample test, one can use a non-parametric test known as the Wilcoxon (paired samples/within subject) or Mann-Whitney (independent samples) test.

```
df = pd.read_csv("blood_pressure.csv")

stats.wilcoxon(x=df['bp_before'], y=df['bp_after'],
                                  alternative='two-sided')
```

```
WilcoxonResult(statistic=2234.5, pvalue=0.0014107333565442858)
```

```
stats.mannwhitneyu(data_group1,data_group2, alternative="two-sided")
```

```
MannwhitneyuResult(statistic=155.0, pvalue=0.15249157824623777)
```

Compare the results with those of the parametric tests.

The nonparametric tests are in general less powerful (need more evidence to reject the null)

Probability & Statistics for DS & AI

Analysis of Variance

Michele Guindani

JioInstitute

Summer 2022

- Sometimes it may be necessary to conduct a hypothesis test to compare mean values simultaneously for more than two groups (samples) created using a factor (or factors).
- For example, a marketer may like to understand the impact of three different discount values (such as 0%, 10%, and 20% discount) on the average sales.
- One-way ANOVA can be used to study the impact of a single treatment (also known as factor) at different levels (thus forming different groups) on a continuous response variable (or outcome variable).
- Then the null and alternative hypotheses for one-way ANOVA for comparing 3 groups are given by

  $H_0 : \mu_1 = \mu_2 = \mu_3$

  H_A : Not all μ values are equal

  where μ_1, μ_2, μ_3 are mean values of each group (Omnibus test)

Example (Witty Bazaar & Discounts)

The brand manager of the "Witty Bazaar" store wants to understanding whether the price discounts have any impact on the sales quantity of a product. To test whether the price discounts had any impact, price discounts of 0% (no discount), 10%, and 20% were given on randomly selected days.

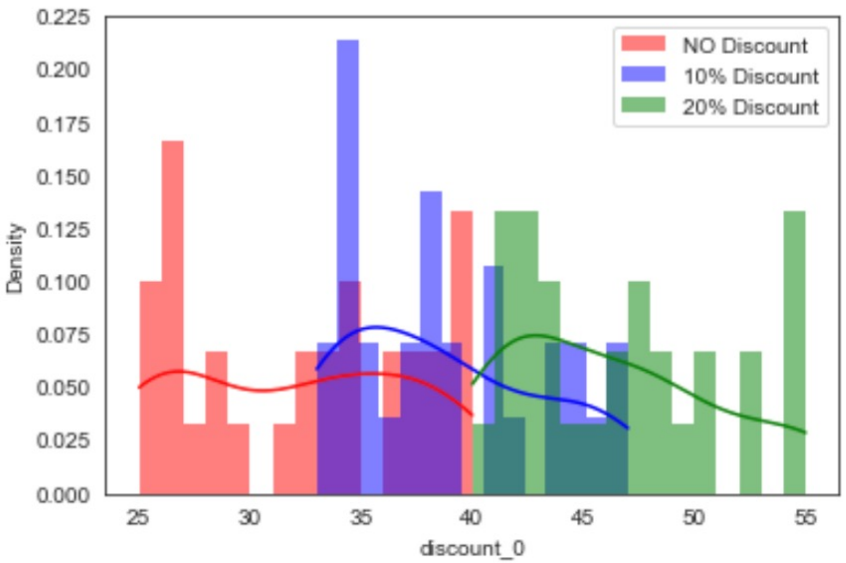
We have recorded the quantity (in kilograms) of the product sold in a day under different discount levels.

We conduct a one-way ANOVA to check whether discount had any significant impact on the average sales quantity at $\alpha = 0.05$.

```
onestop_df=pd.read_csv('onestop.csv')
```

```
sns.histplot(onestop_df['discount_0'], label = 'NO Discount', kde=True,
             stat="density", linewidth=0, color="red", bins=15)
sns.histplot(onestop_df['discount_10'], label = '10% Discount', kde=True,
             stat="density", linewidth=0, color="blue", bins=15)
sns.histplot(onestop_df['discount_20'], label = '20% Discount', kde=True,
             stat="density", linewidth=0, color="green", bins=15)
plt. legend()
```

```
<matplotlib.legend.Legend at 0x7fb5a1618070>
```




```
onestop_df.head(5)
```

	discount_0	discount_10	discount_20
0	39	34	42
1	32	41	43
2	25	45	44
3	25	39	46
4	37	38	41

```
from scipy.stats import f_oneway
statistic, pvalue=f_oneway (onestop_df ['discount_0'],
onestop_df ['discount_10'],
onestop_df ['discount_20'])
print(pvalue)
if pval<0.05:
    print("Null hyphothesis rejected , Alternative hyphothesis accepted")
else:
    print("Null hyphothesis failed to be rejected, Alternative hyphothesis rejected")
```

```
3.821500669725641e-18
Null hyphothesis rejected , Alternative hyphothesis accepted
```