IRON HACK

Data Analytics Full-Time Bootcamp

HYPOTHESIS TESTING

THE 6 STEPS OF HYPOTHESIS TESTING

1. **Set the hypothesis**
2. Choose significance / confidence level
3. Sample
4. Compute statistic
5. Get p-value
6. Decide

EXPERIMENTAL PROTOCOL

In order to do hypothesis testing we need to have an hypothesis. An assertion that we make about our population whose validity we will try to assess.

Statistical hypothesis testing is based on the premise that you cannot prove that something is universally true, but you can prove, by showing a counterexample, that something is false.

Therefore for statistical hypothesis testing we set two opposing hypothesis. One is the *null hypothesis*, denoted by H_0. The null hypothesis represents the baseline assertion that should be “the default” and it should be falsifiable. The other hypothesis, called the *alternative hypothesis*, denoted by H_1 is mutually exclusive with the null hypothesis. If H_1 is true, we should be able to gather enough evidence such that maintaining H_0 strains credibility.

In these cases we say that we *reject the null hypothesis*. Notice the careful wording.

EXPERIMENTAL PROTOCOL

What should be the null and the alternative in the following cases?

- All swans are white versus there is a swan that is black
- There is a placebo effect versus there is no placebo effect
- The mean of two distributions is the same vs the mean of two distributions is different

EXPERIMENTAL PROTOCOL

Quick validations

- Hypothesis should be about the population parameters, not statistics nor samples
- If there is a “no effect” option, it should be on the null
- “Equalities” (=,<=,>=) should be on the null

RUNNING EXAMPLE

- Let’s take our Titanic dataset. You have seen that the prices in first class were on average 85 dollars and someone told you that prices in 3rd class were usually a fifth of prices in first class. You are skeptical. Set up the hypotheses to test this.
- Now, you think the prices in third class are even cheaper than that. Set up the hypotheses to test this.

THE 6 STEPS OF HYPOTHESIS TESTING

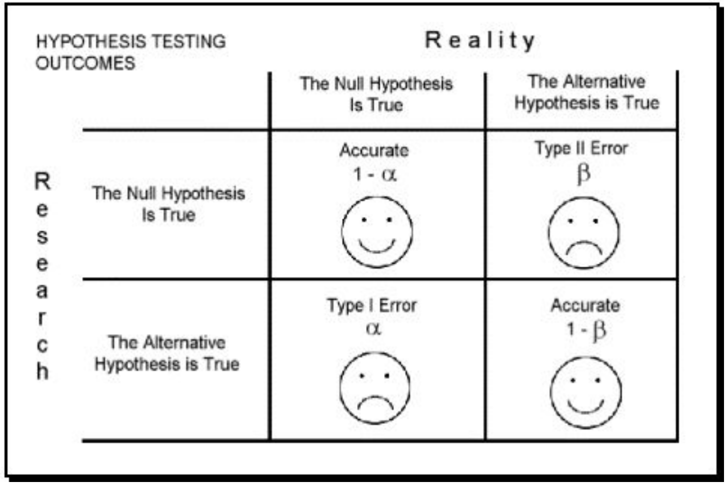
1. Set the hypothesis
2. **Choose significance / confidence level**
3. Sample
4. Compute statistic
5. Get p-value
6. Decide

SIGNIFICANCE AND TYPES OF ERRORS

We said that we would make our decision to reject the null when it strains credibility yo maintain it. But how do we measure this? We set a significance level **α** *à priori*.

Significance is the probability of rejecting the null if it happens to be true. This is the most damaging type of error so we should be demanding with our significance (**α** chosen at most 5%, usually)

We have seen that significance is the converse of confidence (1-**α**).



HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate 1 - α ☺	Type II Error β ☹
	The Alternative Hypothesis is True	Type I Error α ☹	Accurate 1 - β ☺

RUNNING EXAMPLE

- Let’s pick a significance level. Are you feeling confident about our claim or not?

THE 6 STEPS OF HYPOTHESIS TESTING

1. Set the hypothesis
2. Choose significance / confidence level
3. **Sample**
4. Compute statistic
5. Get p-value
6. Decide

RUNNING EXAMPLE

- Open the class collab.
- Let’s assume we have access to only 30 3rd class passengers (say, to get the price paid for the ticket you had to track down their families to send you a copy of the receipt).
- Sample your dataset for 30 entries

THE 6 STEPS OF HYPOTHESIS TESTING

1. Set the hypothesis
2. Choose significance / confidence level
3. Sample
4. **Compute statistic**
5. Get p-value
6. Decide

COMPUTE YOUR TEST STATISTIC

Your test statistic depends on what kind of test you are trying to make. You always compute your statistics assuming H_0, which is the assumption you may end up falsifying. To check for a mean value for the population we have seen that we can use the test statistic

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Where X is the sample mean, μ the mean value for the population under H_0, $\boldsymbol{\sigma}$ the standard deviation of the population and $\boldsymbol{n}$ the size of the sample.

This test statistic is Normally distributed (CLT) so we can check if for our particular observation of X, the value of μ given by H_0 stretches credibility (e.g., if the value of μ is correct and our sample means is, say, 5 standard deviations away from μ, that would be very very unlikely)

COMPUTE YOUR TEST STATISTIC

In our case we don't know the value of **σ**, so we instead use the t-statistic

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

Where s is the sample standard deviation.

In this case, this test statistic is T-distributed with n-1 degrees of freedom.

RUNNING EXAMPLE

- Compute your sample mean and standard deviation
- Compute your test statistic under H_0

THE 6 STEPS OF HYPOTHESIS TESTING

1. Set the hypothesis
2. Choose significance / confidence level
3. Sample
4. Compute statistic
5. **Get p-value**
6. Decide

GET THE P-VALUE

Now that we have a test statistic and we know how our sampling distribution is distributed (Gaussian in the case where we know the **σ**, T-distributed otherwise) we can check **how likely is it to get a sample mean as extreme as the one we actually sampled.**

This quantity is called the **p-value**

If the p-value is very small, it means that it is extremely unlikely, under H_0, to get a result like the one we did in our sample and thus we should reject H_0, because staying with it strains credibility.

We compute the p-value from a table and the value of our statistic or, because it’s 2020, from a scipy call.

one-tailed

```
st.norm.sf(abs(stat))
st.t.sf(abs(stat),n-1)
```

two-tailed

```
st.norm.sf(abs(stat))*2
st.t.sf(abs(stat),n-1)*2
```

RUNNING EXAMPLE

- Get the p-value for your test statistic
- Watch out, the way to compute p-values for each of our hypotheses is actually different

THE 6 STEPS OF HYPOTHESIS TESTING

1. Set the hypothesis
2. Choose significance / confidence level
3. Sample
4. Compute statistic
5. Get p-value
6. **Decide**

DECISION CRITERIA

We now compare our obtained p-value (chance to see an observation at least as extreme as the one we saw) with our significance level .

- If $p<\alpha$, we have just witnessed an event that, if H_0 is true, happens less than a fraction α of the times. This strains credibility and we therefore reject H_0
- If $p\geq\alpha$, we have witnessed an event that, if H_0 is true, happens more than a fraction α of the times. This is not enough to convince us to change our minds about H_0 and thus we do not reject it

Word of warning: in single sided tests, your test statistic needs to “go against” H_0 for you to reject it. E.g. If H_0 posits that average weights are lower than 50 and your observation is below 50kg, you can’t reject, even if the p value is smaller than α

RUNNING EXAMPLE

Make a decision on whether you reject that

- prices in 3rd class were usually a fifth of prices in first class
- prices in third class are a fifth of prices in first class or more expensive

HYPOTHESIS TESTING WITH PYTHON

Kindly, for the vast majority of tests, Python takes steps 4 and 5 of the process and arranges them in a neat little package

1. Set the hypothesis
2. Choose significance / confidence level
3. Sample
4. **Compute statistic**
5. **Get p-value**
6. Decide

HYPOTHESIS TESTING WITH PYTHON

In the case of our particular Titanic problem, we call the function

`ttest_1samp`

On the sample and on the posited μ.

The default test is double sided, so if we are doing a uni-sided test we need to

- **Halve** the p-value
- Check if the test statistic is falling on the “right” side of the T-distribution

TESTING - FIXED VALUE

Tests

- $H_0: \mu=k$ vs $H_1: \mu \neq k$
- $H_0: \mu \geq k$ vs $H_1: \mu < k$
- $H_0: \mu \leq k$ vs $H_1: \mu > k$

Function

```
ttest_1samp(sample,k)
```

The default test is double sided, so if we are doing a uni-sided test we need to

- **Halve** the p-value
- Check if the test statistic is falling on the “right” side of the distribution

TESTING - MATCHED PAIR

Tests

- H_0: $\mu_1=\mu_2$ vs H_1: $\mu_1\neq\mu_2$
- H_0: $\mu_1\geq\mu_2$ vs H_1: $\mu_1<\mu_2$
- H_0: $\mu_1\leq\mu_2$ vs H_1: $\mu_1>\mu_2$

Data in the two samples is dependent. E.g. heart rate before and after taking a pill, measured for the same individuals

Function

```
ttest_rel(sample1, sample2)
```

The default test is double sided, so if we are doing a uni-sided test we need to

- **Halve** the p-value
- Check if the test statistic is falling on the “right” side of the distribution

TESTING - INDEPENDENT SAMPLES

Tests

- $H_0: \mu_1=\mu_2$ vs $H_1: \mu_1\neq\mu_2$
- $H_0: \mu_1\geq\mu_2$ vs $H_1: \mu_1<\mu_2$
- $H_0: \mu_1\leq\mu_2$ vs $H_1: \mu_1>\mu_2$

Data in the two samples is independent. E.g. effect of pill in men and women

Function

```
ttest_ind(ab_test.a, ab_test.b, equal_var=False)
```

The default test is double sided, so if we are doing a uni-sided test we need to

- **Halve** the p-value
- Check if the test statistic is falling on the “right” side of the distribution

In cases when we happen to know the variance is equal in both populations, we can set the third parameter to True

OTHER TYPES OF TESTING - ANOVA AND F-TESTS

Anova is the most well known of a series of complex tests that test for many parameters at once. In the case of Anova we are trying to understand if the mean of several samples can reasonably said to be the same, e.g., the income of multiple ethnic groups, the CO2 emitted by the citizens of multiple countries, etc.

Tests

- H_0: $\mu_1=\mu_2=\ldots=\mu\square$ vs H_1: $\mu\square\neq\mu\square$ for some s,t

Function

```
f_oneway(sample1,sample2,...,samplen)
```