Table of Contents

Resources

- Analytics Vidhya: Everything you Should Know about p-value from Scratch for Data Science

Hypothesis testing

Suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it's more than that. So you conduct a hypothesis test and randomly sample some delivery times to test the claim:

- Null hypothesis — The mean delivery time is 30 minutes or less
- Alternative hypothesis — The mean delivery time is greater than 30 minutes

We'll use one-tailed test in our case since we only care about if the mean delivery time is greater than 30 minutes.

We use z-test for hypothesis testing:

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

Normal distribution

P-value

> The lower the p-value, the more surprising the evidence is, the more ridiculous our null hypothesis looks.

> If the p-value is lower than a predetermined significance level then we reject the null hypothesis.

Now that we've collected some sampled delivery times, we perform the calculation and find that the mean delivery time is longer by 10 minutes with a p-value of 0.03.

What this means is that in a world where the pizza delivery time is 30 minutes or less (null hypothesis is true), there's a 3% chance we would see the mean delivery time is at least 10 minutes longer due to random noise.

The lower the p-value, the more meaningful the result because it is less likely to be caused by noise. There's a common misinterpretation of p-value for most people in our case: The p-value 0.03 means that there's 3% (probability in percentage) that the result is due to chance — which is not true.

When to use Z-test and T-test

When data sample is < 30 samples, use T-test otherwise use z-test. Some notes: - Use Z-test when your sample size is greater than 30. Otherwise, use a t test. - Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point. - Your data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter. - Your data should be randomly selected from a population, where each item has an equal chance of being selected. - Sample sizes should be equal if at all possible.

One sample t-test

The One Sample t Test determines whether the sample mean is statistically different from a known or hypothesised population mean. The One Sample t Test is a parametric test. Example :- you have 10 ages and you are checking whether avg age is 30 or not.

```
from scipy.stats import ttest_1samp
import numpy as np


x = np.random.randint(20,40,size=(50))
x_mean = np.mean(x)
tset, pval = ttest_1samp(x, 30)
alpha = 0.05

print('data = ', x)
```

```
print()
print('mean = ', x_mean)
print('p-value = ', pval)
print('alpha = ', alpha)
if pval < alpha:    # alpha value is 0.05 or 5%
   print(" we reject the null hypothesis")
else:
  print("we accept the null hypothesis")
```

Two sampled T-test

The Independent Samples t Test or 2-sample t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples t Test is a parametric test. This test is also known as: Independent t Test.

Example : is there any association between week1 and week2

```
import numpy as np
from scipy.stats import ttest_ind

x1 = np.random.randint(20,40,size=(50))
x2 = np.random.randint(20,40,size=(50))
alpha = 0.05

week1 = x1
week2 = x2

print(week1)
print("week2 data :-\n")
print(week2)
week1_mean = np.mean(week1)
week2_mean = np.mean(week2)
print("week1 mean value:",week1_mean)
print("week2 mean value:",week2_mean)
week1_std = np.std(week1)
week2_std = np.std(week2)
print("week1 std value:",week1_std)
print("week2 std value:",week2_std)
ttest,pval = ttest_ind(week1,week2)
print("p-value",pval)
if pval <0.05:
  print("we reject null hypothesis")
else:
```

```
print("we accept null hypothesis")
```

Paired sampled t-test

The paired sample t-test is also called dependent sample t-test. It's an uni variate test that tests for a significant difference between 2 related variables. An example of this is if you where to collect the blood pressure for an individual before and after some treatment, condition, or time point. > H0 :- means difference between two sample is 0 > H1:- mean difference between two sample is not 0

```
import numpy as np
from scipy import stats

x1 = np.random.randint(20,40,size=(50))
x2 = np.random.randint(20,40,size=(50))
alpha = 0.05
ttest,pval = stats.ttest_rel(x1, x2)
print(pval)
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

One-sampled Z-test

```
from scipy import stats
from statsmodels.stats import weightstats as wstats

ztest ,pval = wstats.ztest(x, x2=None, value=30)
print(float(pval))
if pval< alpha:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

Two-sampled Z-test

```
x1 = np.random.randint(20,40,size=(50)) # blood sugar before
x2 = np.random.randint(20,40,size=(50)) # blood sugar after
alpha = 0.05
```

```
# value is difference of mean

ztest ,pval1 = wstats.ztest(x1, x2=x2, value=0,alternative='two-sided')
print(float(pval1))
if pval< alpha:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

F-test or ANOVA

When we have more than two groups we use ANOVA test.