Statistical Tests

What will be covering today

Statistical Tests

- Various statistical tests for hypothesis tests, mean etc
- Quantile Plots
- Testing for Normality
- Comparing two sets of samples
  - Paired data
- Analysis of Variance

What you should focus on?

- Should be able to test if a data sample has normality
- Able to test if samples are significantly different
- Should perform various statistical tests on various metrics/hypothesis

Statistical tests of differences and relations

Statistical Analysis at a very high level

- We have some data (samples)

We can do one of three things:

1. Test for **differences**

   Are two set of values really different (means)

2. Look for **associations**

   Are these two variables, columns related to each other

3. What is the **relationship** between variables?

1.Regression, fitting lines, predicting values

Statistical Analysis at a very high level

Assumptions about our data sets (columns)

1. Data is unbiased, independent
2. Normally distributed
3. Equal variances in each set

One line explanations for Statistical Terms

- **Statistical tests** let you choose between two competing hypotheses; Confidence intervals reflect the likely range of a population parameter.
- **Normality**: Is a data set normally distributed.
- **QQ plot**: A graphical method for computing two distributions by plotting their quantiles against each other
- **QQ norm**: A qqplot comparing a dataset to a theoretical normal distribution on one axis
- "**Two- sided Tests**"
  - We are testing if A is different from B, without caring about A>B or whether A<B.

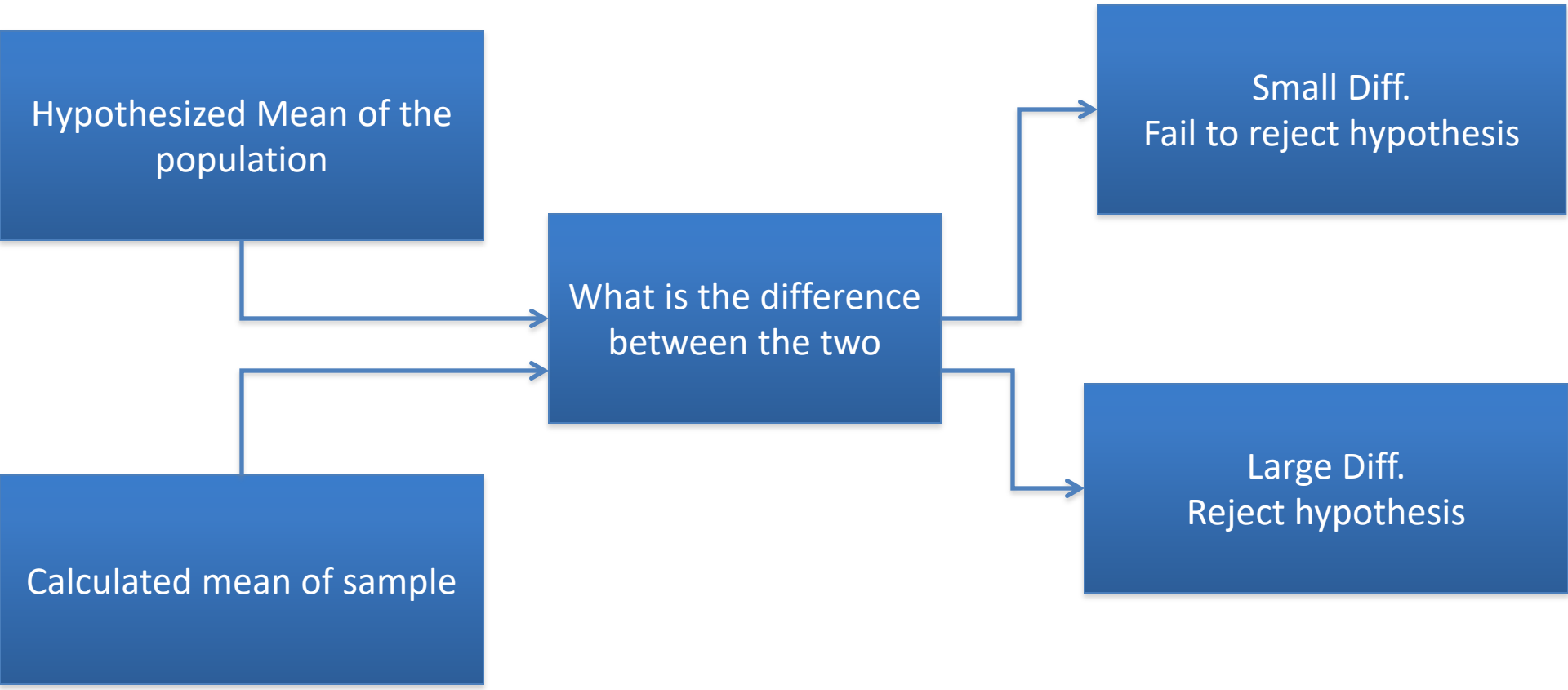
Hypothesis Testing

- Hypothesis testing is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed?
- In a formal hypothesis test, hypotheses are always statements about the population. The usual hypothesis tests involve statements about the average values (means) of some variable in the population.
- There are two hypotheses
  - The hypothesis to be tested is called the null hypothesis and given the symbol H_0. The null hypothesis states that there is no difference between a hypothesized population mean and a sample mean.

Hypothesis Testing

- we test the null hypothesis against an alternative hypothesis which is given the symbol H_a. It includes the outcomes not covered by the null hypothesis.
- The alternative hypothesis can be supported only by rejecting the null hypothesis. To reject the null hypothesis means to find a large enough difference between sample mean and the hypothesized (null) mean.

Hypothesis Testing



One line explanations for Statistical Terms

- **Null Hypotheses, Alternative Hypotheses, and p-values**
- Null Hypothesis, is that nothing happened: the mean was unchanged, the model did not improve; and so forth
- We want to determine which hypothesis is more likely in the light of data:
  - To begin, we assume that the null hypothesis is true
  - We calculate a test statistics. It could be something simple, such as mean of the sample, or it could be complex. The critical requirement is that we must know the statistic's distribution

One line explanations for Statistical Terms

- From the statistic and its distribution we can calculate a *p-value*, the probability of a test statistic value as extreme or more extreme than the one we observed, while assuming that the null hypothesis is true.
- *If the p-value* is too small, we have strong evidence against the null hypothesis. This is called rejecting the null hypothesis.

One line explanations for Statistical Terms

- If the p-value is not small then we have no such evidence. This is called failing to reject the null hypothesis.

There is one necessary decision here: when is a p-value “ too small”?

- Common convention is that we reject the null hypothesis when $p<0.05$ and fail to reject when $p>0.05$. In statistical sense, we chose a significance level of α – 0,05 to define the border between strong evidence and insufficient evidence against the null hypothesis
- But the real answer is “ it depends”

Hypothesis Testing

- Statisticians choose a level of significance or alpha level for their hypothesis test. The most frequently used values of significance are 0.05 and 0.01.
- An alpha level of 0.05 means that we will consider our sample mean to be significantly different from the hypothesized mean if the chances of observing that sample mean are less than 5%.
- If there is a significant difference ($p<0.05$), then the null hypothesis would be rejected.
- Otherwise, if no significance difference ($p>0.05$), then the null hypothesis would not be rejected

Confidence interval

- Confidence interval = 1-level of significance
- If the level of significance is 0.05, then the confidence interval is 95%
- CI = 1-0.05 = 0.95 = 95%
- If CI = 99%, then level of significance is 0.01
-

Error

- Although we have determined the level of significance and confidence interval, there is still a chance of error.

- There are two types:
  - Type I error
  - Type II error

	Actual	
Predicted	positive	negative
positive	Correct Decision	Type I error
negative	Type II error	Correct Decision

Error

Test of Significance	Correct Null Hypothesis (H0 not rejected)	Incorrect Null Hypothesis (H0 rejected)
Null Hypothesis True	Correct Conclusion	Type I error
Null Hypothesis False	Type II error	Correct Conclusion

Error

- Type I error – rejecting the null hypothesis although the null hypothesis is correct, e.g.
  - When we compare the mean/proportion of the 2 groups, the difference is small but the difference is found to be significant. Therefore the null hypothesis is rejected
  - It may occur due to inappropriate choice of alpha (level of significance)

Error

- Type II error – not rejecting the null hypothesis although the null hypothesis is wrong, e.g.
  - When we compare the mean/proportion of the 2 groups, the difference is big but the difference is found to be not significant. Therefore the null hypothesis is not rejected
  - It may occur when the sample size is too small.

Determining the appropriate statistical test

Hypothesis Testing

- Distinguish parametric & non-parametric procedures

- Test two or more populations using parametric & non-parametric procedures
  - Means
  - Medians
  - Variances

Correlation Test

Why is it used?

To test two samples are related.

Assumptions:
- Observations in each sample are independent and identically distributed
- Observations in each sample are normally distributed
- Observations in each sample have the same variance

Interpretation:

- H0: The two samples are independent
- H1: The two samples are dependent

```
from scipy.stats import pearsonr
data1 = np.random.random_sample((100,1))
data2 = np.random.random_sample((100,1))
stat, p = pearsonr(data1, data2)

print('stat=%.3f, p=%.3f' % (stat, p))
```

Output:

Stat = 0.007, p = 0.942

Probably independent

Chi-Squared Test

Why is it used?

To test whether two categorical variables are related or independent.

Interpretation:

- H0: The two samples are independent
- H1: The two samples are dependent

```
from scipy.stats import chi2_contingency
obs = [[10, 20, 30],[6,  9,  17]]
stat, p, dof, expected = chi2_contingency(obs)
print('stat=%.3f, p=%.3f' % (stat, p))
```

Stat = 0.272, p = 0.873

One line explanations for Statistical Terms

Student's t-test

The t-test is the most commonly used method to evaluate the difference in means between two groups.

This test help determine how confident we can be that the differences between two groups as a result of the treatment is not due to chance.

The researcher calculates a t-value using the sample mean and standard deviation and compares the calculated t-value against a tabulated value. If null hypothesis is rejected, we can say that the difference between the two groups is significant.

t-test

Assumptions

•Observations in each sample are independent and identically distributed (iid).

•Observations in each sample are normally distributed.

•Observations in each sample have the same variance.

Interpretation

•H0: the means of the samples are equal.

•H1: the means of the samples are unequal.

One line explanations for Statistical Terms

Theoretically, the t-test can be used even if the sample sizes are small, as long as the variables are normally distributed within each group and variation in the two groups is not reliably different. The 2 groups are independent of each other.

One Sample t-test

Is this sample somehow different from the population at large?

- Compare the mean of a sample to a known value
  - Usually the population mean (the average for the outcome of some population of interest)
- The basis idea is to test for two things:
  - The average of the sample (observed average) and
  - The population (expected average)
- We have to adjust for the number of cases in the sample and the standard deviation of the average

One Sample t-test

- 3 steps:
  - We expect the sample mean and the pop. mean are the same (H_0)
  - Calculate the t-statistic, and therefore p-value
  - Accept or reject H_0

```
from scipy.stats import ttest_1samp
x=np.random.normal(loc = 1.0, scale = 0.005, size =100)
stat, p = ttest_1samp(x,.9)

print('stat=%.3f, p=%.3f' % (stat, p))
```

stat = 174.207, p = 0.000

Two data sets

```
from scipy.stats import ttest_ind
data1 = np.random.random_sample((100,1))
data2 = np.random.random_sample((100,1))
stat, p = ttest_ind(data1, data2)

print('stat=%.3f, p=%.3f' % (stat, p))
```

Stat = 1.207, p = 0.229

Same distribution

Testing Paired Data

“ Repeated measurement on the same individual

“paired” simply means that two different values comes from the same test subjects

Interpretation

•H0: the means of the samples are equal.

•H1: the means of the samples are unequal.

```
from scipy.stats import ttest_rel
data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
stat, p = ttest_rel(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
```

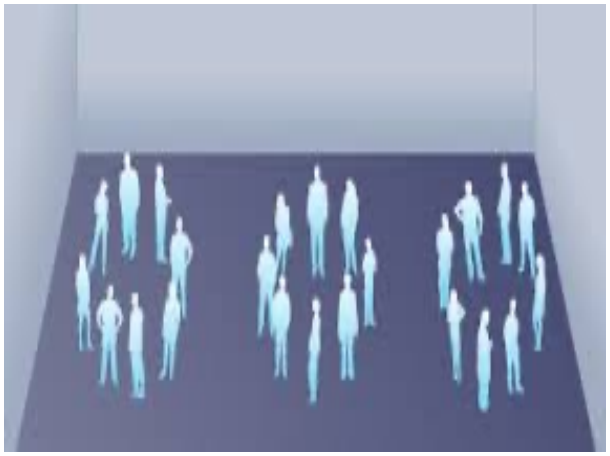
Stat = 0.334, p = 0.746

Analysis of Variance Test (ANOVA)

- Extension of independent sample t test
- Compares the means of groups of independent observations
  - Don’t be fooled by the name. ANOVA does not compare variances
- Can compare more than two groups
- Tests the equality of 2 or more population means
- Used to analyze completely randomized experimental designs
- Comparing the blood cholesterol levels between the students, teachers and Adm. staff

ANOVA explained pictorially

- Is there difference in the means of 3 or more groups?
- One dependent (response) variables
- One or more categorical variables – called “factors”
- Each level of a factor is called a “treatment”



Treament A

Treatment B

Treatment C

Question: Are the mean scores from A, B and C really different?

ANOVA can answer that

ANOVA

```
from scipy.stats import f_oneway
data1 = np.random.random_sample((100,1))
data2 = np.random.random_sample((100,1))
data3 = np.random.random_sample((100,1))

stat, p = f_oneway(data1, data2, data3)
print('stat=%.3f, p=%.3f' % (stat, p))
```

stat = 0.381, p = 0.683

Example

Wilcoxon Signed Rank Test

Why / When is it used?

Two paired samples have same distribution

To test the mean of a sample when normal distribution is not assumed.

Wilcoxon signed rank test can be an alternative to t-Test, especially when the data sample is not assumed to follow a normal distribution.

Observations across each sample are paired

Interpretation

- H0: the distributions of both samples are equal.

- H1: the distributions of both samples are not equal.

Example

```
from scipy.stats import wilcoxon
data1 = [0.725, 1.725, 0.245, -0.895, -1.566, -1.256, 1.678, -2.867, 1.345, 2.345]
data2 = [2.567, -0.567, -0.825, -0.987, 0.256, 1.256, 0.800, 2.831, 1.075, 0.981]
stat, p = wilcoxon(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
```

```
stat = 23.000, p = 0.646
```

Kruskal- Wallis H test

Tests whether the distributions of two or more independent samples are equal or not.

Interpretation

- H0: the distributions of all samples are equal.

- H1: the distributions of one or more samples are not equal.

Example

```
from scipy.stats import kruskal
data1 = [0.725, 1.725, 0.245, -0.895, -1.566, -1.256, 1.678, -2.867, 1.345, 2.345]
data2 = [2.567, -0.567, -0.825, -0.987, 0.256, 1.256, 0.800, 2.831, 1.075, 0.981]
stat, p = kruskal(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
```

stat = 0.571, p = 0.450

Friedman Test

Tests whether the distributions of two or more paired samples are equal or not.

Interpretation

- H0: the distributions of all samples are equal.
- H1: the distributions of one or more samples are not equal.

Example

```
from scipy.stats import friedmanchisquare
data1 = np.random.random_sample((50,1))
data2 = np.random.random_sample((50,1))
data3 = np.random.random_sample((50,1))

stat, p = friedmanchisquare(data1, data2, data3)
print('stat=%.3f, p=%.3f' % (stat, p))
```

stat = 22125.500, p = 0.000

Testing Differences

We want a statistical test to determine whether our data sample is from a normally distributed population.

- One Set : Testing for normality qqplot, qqnorm, Shapiro test
- Two sets Independent: T-test, Wilcoxon Test
- Two sets Paired: Paired T-test
- More than two sets: Anova Models

Testing for Normality

An important decision point when working with a sample of data is whether to use parametric or nonparametric statistical methods.

Parametric statistical methods assume that the data has a known and specific distribution, often a Gaussian distribution. If a data sample is not Gaussian, then the assumptions of parametric statistical tests are violated and nonparametric statistical methods must be used.

There are a range of techniques that you can use to check if your data sample deviates from a Gaussian distribution, called normality tests.

Testing for Normality

you will know:

•How whether a sample is normal dictates the types of statistical methods to use with a data sample.

•Graphical methods for qualifying deviations from normal, such as histograms and the Q-Q plot.

•Statistical normality tests for quantifying deviations from normal.

Testing for Normality

We can summarize the decision as follows:

If Data Is Gaussian:

Use Parametric Statistical Methods

Else:

Use Nonparametric Statistical Methods

There is also some middle ground where we can assume that the data is Gaussian-enough to use parametric methods or that we can use data preparation techniques to transform the data to be sufficiently Gaussian to use the parametric methods.

Testing for Normality

There are three main areas where you may need to make this evaluation of a data sample in a machine learning project; they are:

- Input data to the model in the case of fitting models.
- Model evaluation results in the case of model selection.
- Residual errors from model predictions in the case of regression.

Testing for Normality

we will look at two classes of techniques for checking whether a sample of data is Gaussian:

- **Graphical Methods**. These are methods for plotting the data and qualitatively evaluating whether the data looks Gaussian.

- **Statistical Tests**. These are methods that calculate statistics on the data and quantify how likely it is that the data was drawn from a Gaussian distribution.

Testing for Normality - Graphically

Histogram Plot

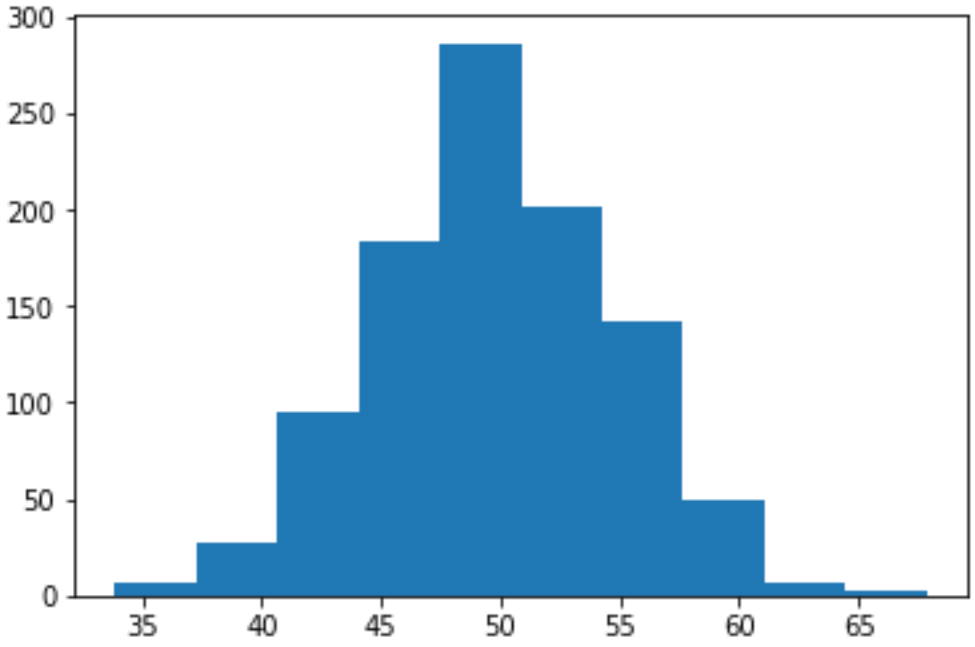
A simple and commonly used plot to quickly check the distribution of a sample of data is the histogram.

In the histogram, the data is divided into a pre-specified number of groups called bins. The data is then sorted into each bin and the count of the number of observations in each bin is retained.

A sample of data has a Gaussian distribution of the histogram plot, showing the familiar bell shape.

Testing for Normality - Graphically

matplotlib.pyplot.hist(data)



Testing for Normality - Graphically

Quantile-Quantile Plot (QQ plot)

This plot generates its own sample of the idealized distribution that we are comparing with, in this case the Gaussian distribution. The idealized samples are divided into groups (e.g. 5), called quantiles. Each data point in the sample is paired with a similar member from the idealized distribution at the same cumulative distribution.

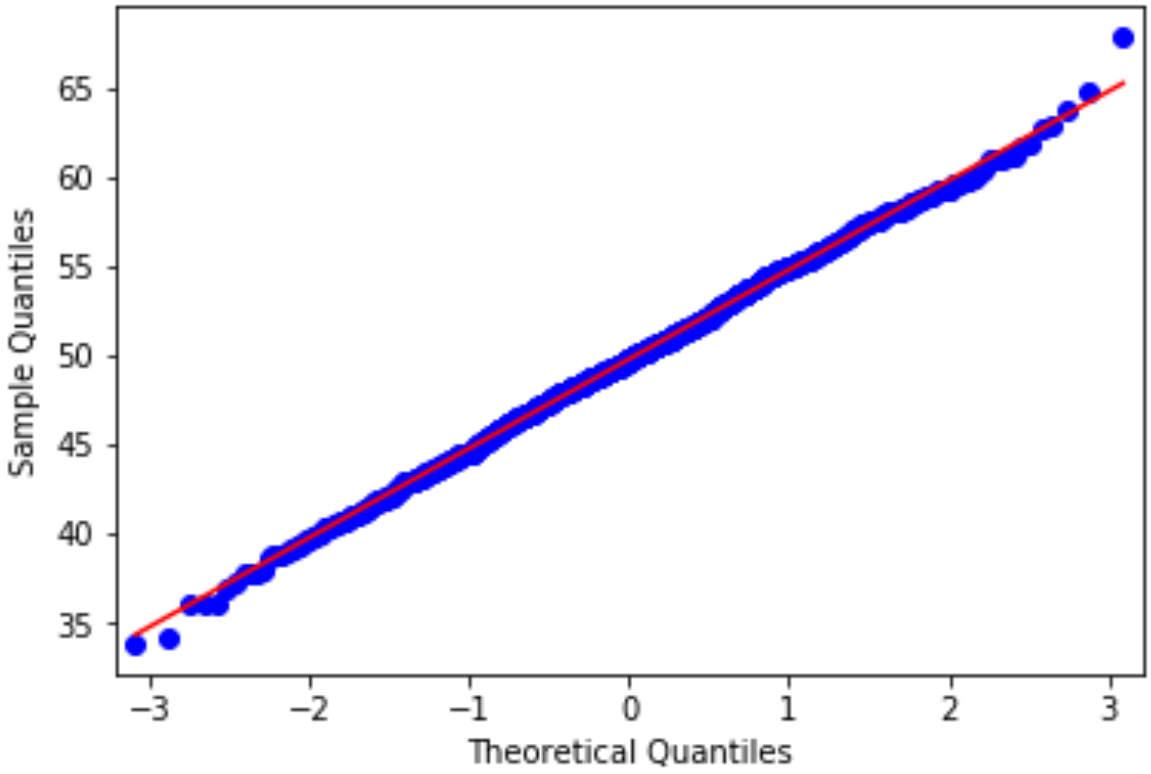
The resulting points are plotted as a scatter plot with the idealized value on the x-axis and the data sample on the y-axis.

A perfect match for the distribution will be shown by a line of dots on a 45-degree angle from the bottom left of the plot to the top right. Often a line is drawn on the plot to help make this expectation clear.

Deviations by the dots from the line shows a deviation from the expected distribution.

Testing for Normality - Graphically

```
from statsmodels.graphics.gofplots import qqplot
qqplot(data, line = 's')
```



Testing for Normality - Statistical Test

Shapiro Test

Why is it used?

To test if a sample follows a normal distribution.

Interpretation:

- H0: The sample has a Gaussian distribution
- H1: The sample does not have a Gaussian distribution

Lets see how to do the test on a sample from a normal distribution.

```
from scipy.stats import shapiro

np.random.seed(123456)
mu,sigma = 0,0.01
x = np.random.normal(mu,sigma, 1000)
stat,p = shapiro(x)
print(p)

0.0897568
```

Take another example from a uniform distribution

```
np.random.seed(123456)
x1 = np.random.uniform(-1,0,100)
stat,p = shapiro(x1)
print(p)

0.0013424
```

Wilcox Test

- You have samples from two populations. You don't know the distribution of the populations, but you know they have similar shapes. You want to know: Is one population shifted to the left or right compared with the other?

- You can use a nonparametric test, the Wilcoxon–Mann–Whitney test, which is implemented by the wilcox.test function.

- The test output includes a p-value. Conventionally, a p-value of less than 0.05 indicates that the second population is likely shifted left or right with respect to the first population whereas a p-value exceeding 0.05 provides no such evidence.

Wilcox Test

- When we stop making assumptions regarding the distributions of populations, we enter the world of nonparametric statistics. The Wilcoxon–Mann–Whitney test is non parametric and so can be applied to more datasets than the t test, which requires that the data be normally distributed (for small samples). This test's only assumption is that the two populations have the same shape.
- This is similar to asking whether the average of the second population is smaller or larger than the first.

Wilcox Test

```
from scipy.stats import wilcoxon
data1 = [0.725, 1.725, 0.245, -0.895, -1.566, -1.256, 1.678, -2.867, 1.345, 2.345]
data2 = [2.567, -0.567, -0.825, -0.987, 0.256, 1.256, 0.800, 2.831, 1.075, 0.981]
stat, p = wilcoxon(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
```

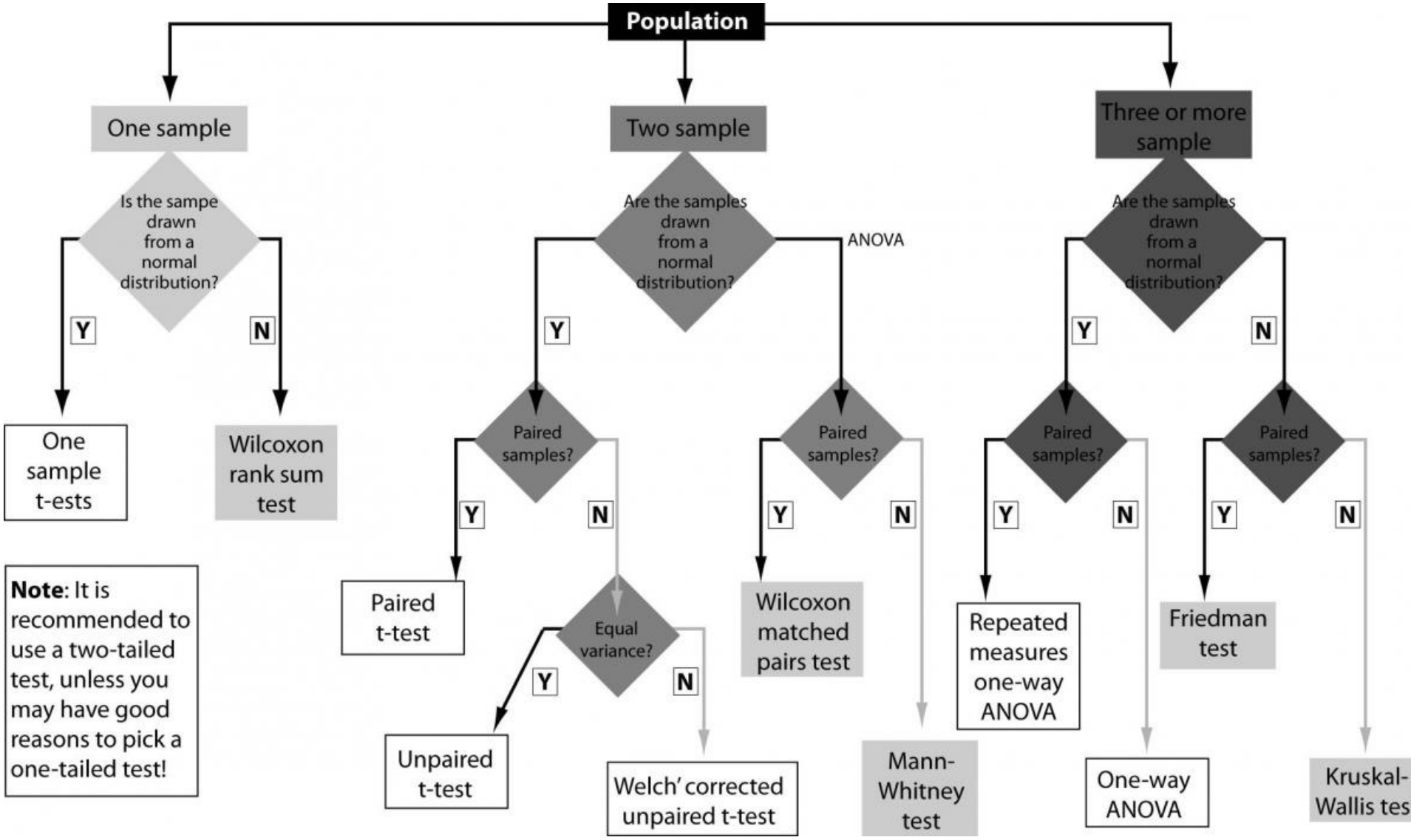
Stat = 23.00 p = 0.646

Fisher's F-test

It can be used to check if two samples have same variance

var.test(x,y)

How to choose the right statistical test?



http://www.biochemia-medica.com/content/comparing-groups-statistical-differences-how-choose-right-statistical-test