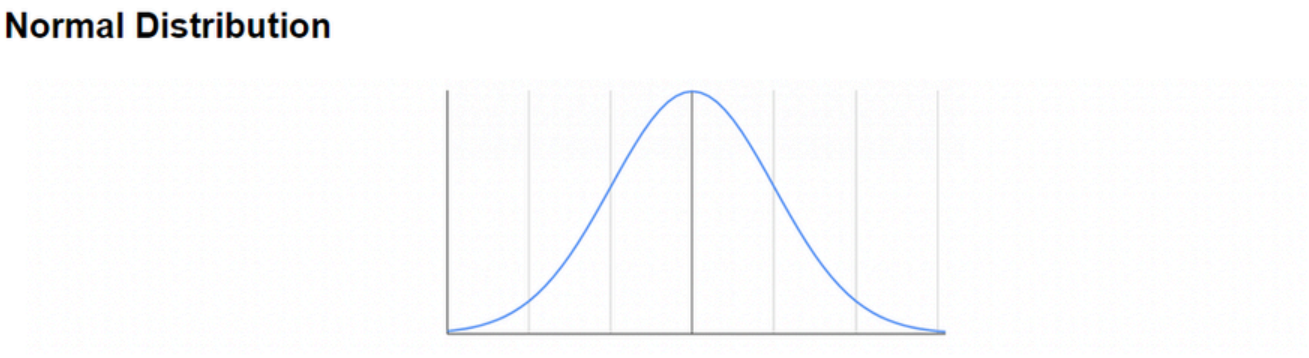
Lesson 1: Z-score



We say that a dataset has a normal distribution if its values fall into a smooth (continuous), bell-shaped curve with a symmetric pattern--each side looks the same when cut down the middle. The normal distribution is extremely important to statistics. One reason that it's so useful is that it **enables us to determine the probability of something occurring due to chance.** For example, it enables us to determine the probability that a given data point will fall a given distance from the distribution's mean.

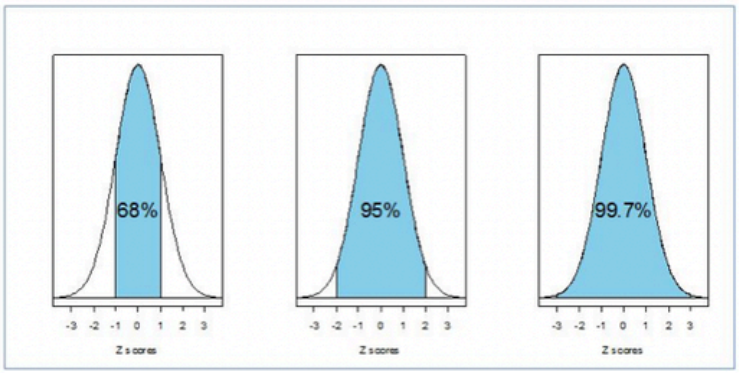
The Empirical Rule (also known as "68-95-99.7 Rule")

The standard normal (Z) distribution has a mean of zero and a standard deviation of 1. You can think of the **standard deviation** as roughly **the average distance of a data point from the mean**. A value on the Z-distribution represents the number of standard deviations the data is above or below the mean; these are called **z-scores**.

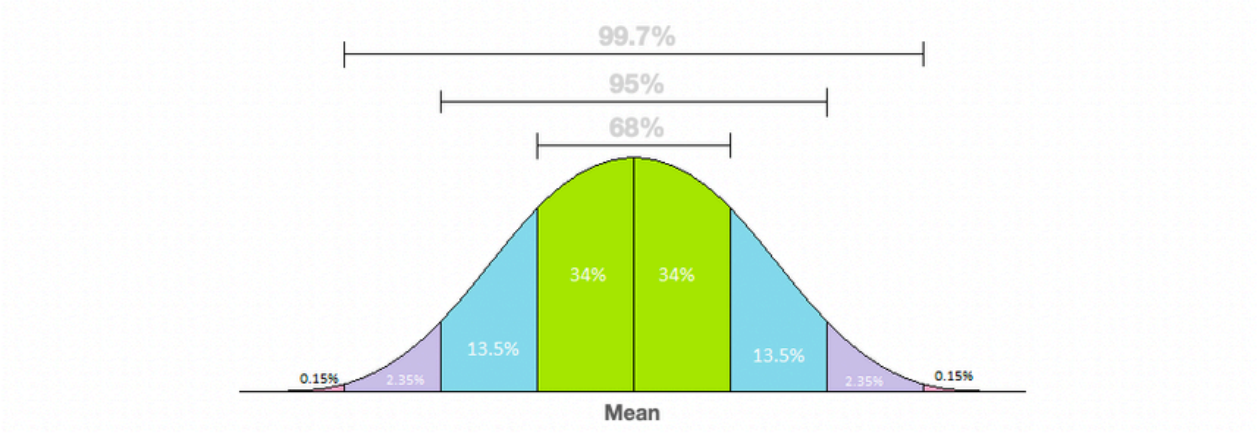
A z-score tells us how different a value is from what we would expect--how far a value is from the mean in standard deviations. For example, if a score in a distribution is equivalent to the mean score, it is zero distance from the mean, hence its z-score will be zero. Therefore, z = 1 on the Z-distribution represents a value that is 1 standard deviation above the mean. Similarly, z = -1 represents a value that is one standard deviation below the mean.

In a normal distribution, there are approximately 68% of values within $\pm$ 1 standard deviation of the mean. Approximately 95% of the values fall within $\pm$ 2 standard deviations of the mean. And virtually all values (99.7%) will fall within $\pm$ 3 standard deviations of the mean.

This is known as the Empirical Rule. This characteristic of the normal distribution is what enables us to generate probabilities about the likelihood of selecting a data point a certain distance from the mean of the distribution: it's z-score.

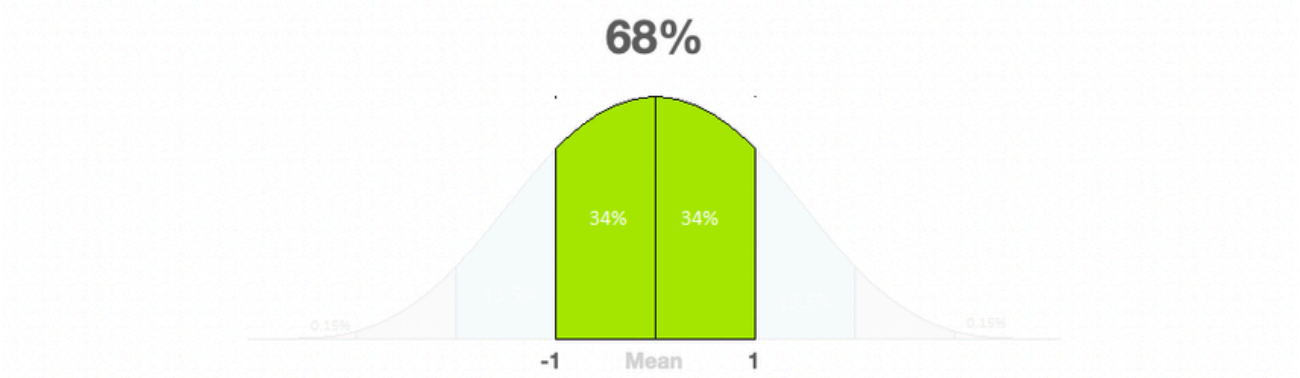


Before we go on to specifically discuss how the z-score is calculated, let's make sure that we can envision the key characteristic of a normal distribution.

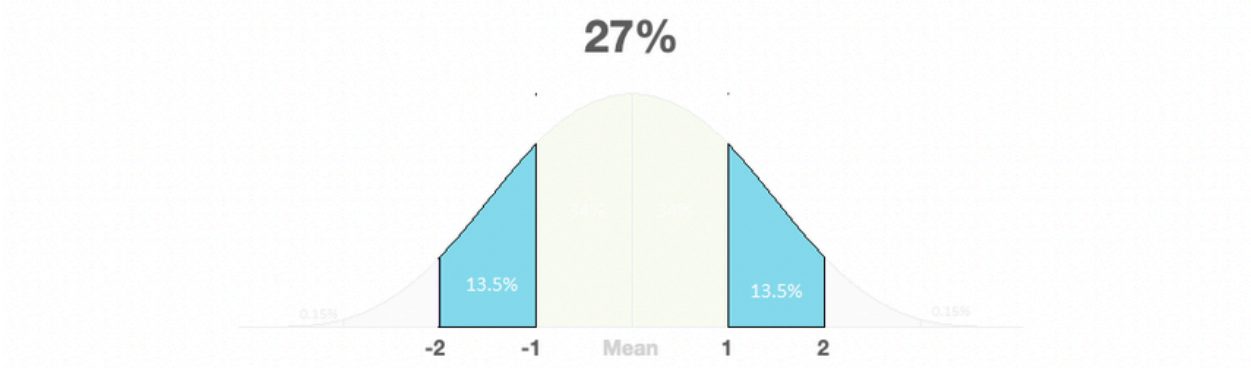


Each normal distribution has its own mean, μ, and its own standard deviation, σ, and the total area under the curve is 1.

68% of values, in a normal distribution, will fall within 1 standard deviation of the mean.

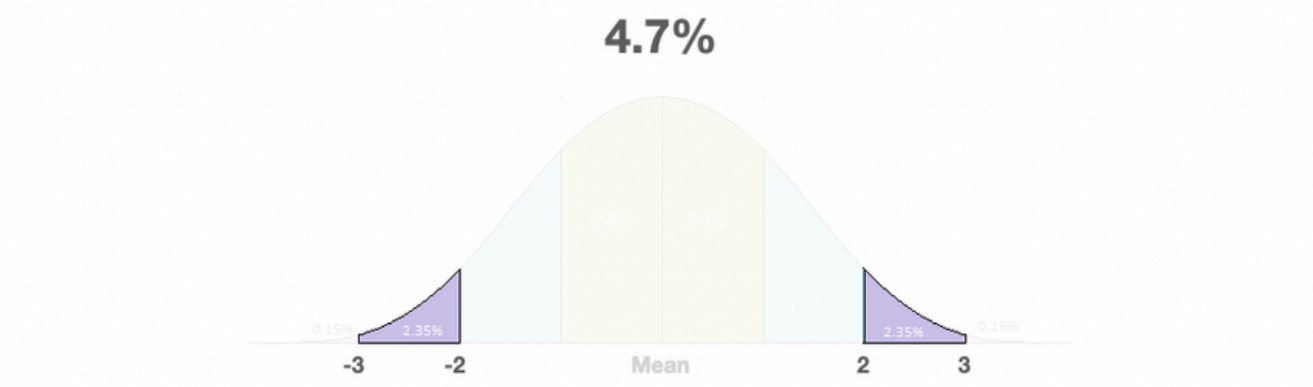


27% of values will fall between 1 and 2 standard deviations above and below the mean.

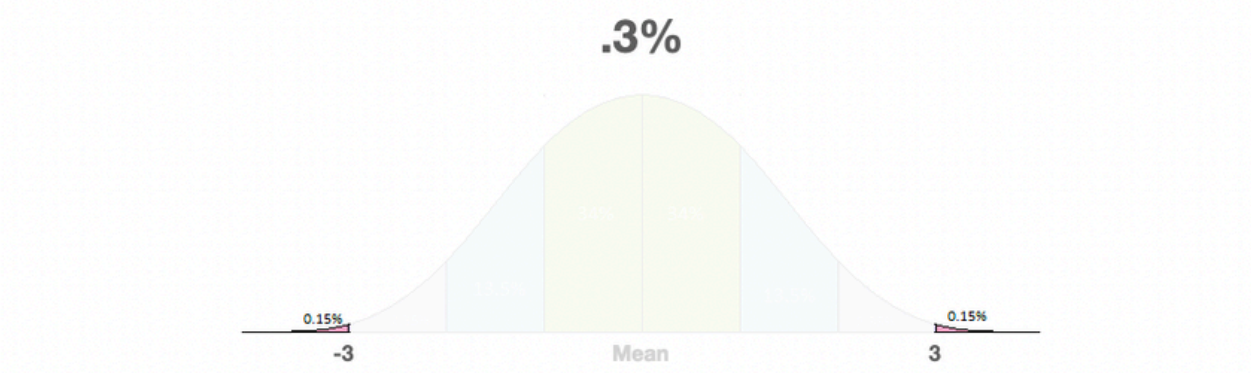


Approximately 4.7% of values will fall between 2 and 3 standard deviations above and below the mean.

Approximately 4.7% of values will fall between 2 and 3 standard deviations above and below the mean.



Only about .3% of values will fall beyond 3 standard deviations away from the mean.



We will now use the above information to understand the usefulness of the z-score.

Lesson 7: False Discovery Rate

False Discovery Rate (FDR) is a measure of accuracy when multiple hypotheses are being tested at once.

In classical statistical testing, we begin with the **null hypothesis** as the formal basis for testing statistical significance. **The null hypothesis states that there is no association between the predictor and outcome variables in the population**. By starting with the proposition that there is no association, statistical tests can estimate the probability that an observed association could be due to chance.

After a study is completed, based on the data collected, the investigator uses statistical tests to determine whether there is sufficient evidence to reject the null hypothesis in favour of the **alternative hypothesis** that there is an association in the population.

When running a statistical test, any time a null hypothesis is rejected, it can be considered to be a "significant" finding since we can conclude that the measured difference is highly unlikely to be due to random chance alone and the treatment is likely directly influencing the metric. **Alternatively, outcomes that do not reach significance are not considered a "discovery" since we aren't able to reject the null hypothesis.**

Alert: The alternative hypothesis cannot be tested directly. It is implicitly accepted when the test of statistical significance rejects the null hypothesis.

Importantly, an investigator's conclusion may be wrong. Sometimes, by chance alone, a sample is not representative of the population. Thus, the results in the sample do not reflect reality in the population, and the random error leads to an erroneous inference. **A false positive (type I error) occurs if an investigator rejects a null hypothesis that is actually true.**

Level of statistical significance

When conducting hypothesis tests, for example to see whether two means are significantly different, we calculate a p-value, which is the probability of obtaining a test statistic that is as, or more extreme than the observed one, assuming the null hypothesis is true.

The level of statistical significance for rejecting the null hypothesis is typically set at 0.05. This states that we can reject the null hypothesis when the probability of rejecting it (p-value), when it is actually true, is less than 5%. In other words, we've set 5% as the maximum chance of **incorrectly rejecting the null hypothesis - having a false positive**.

The **False Discovery Rate** is the proportion of all outcomes deemed to be significant that are falsely significant. Only the positive outcomes matter for the false discovery rate. False negatives don't influence the false discovery rate.

The multiple testing problem

False Discovery Rate comes into play particularly when lots of hypothesis tests are being conducted. For example, when analyzing results from genome-wide studies, a typical microarray experiment might result in performing 10,000 separate hypothesis tests. If we use a p-value of 0.05 as our threshold, we'd expect 500 genes to be deemed as "significant" by chance. The implication is that if you repeat a test enough times, you're going to find an effect even though an effect may not actually exist. This is called **the multiple testing problem**.

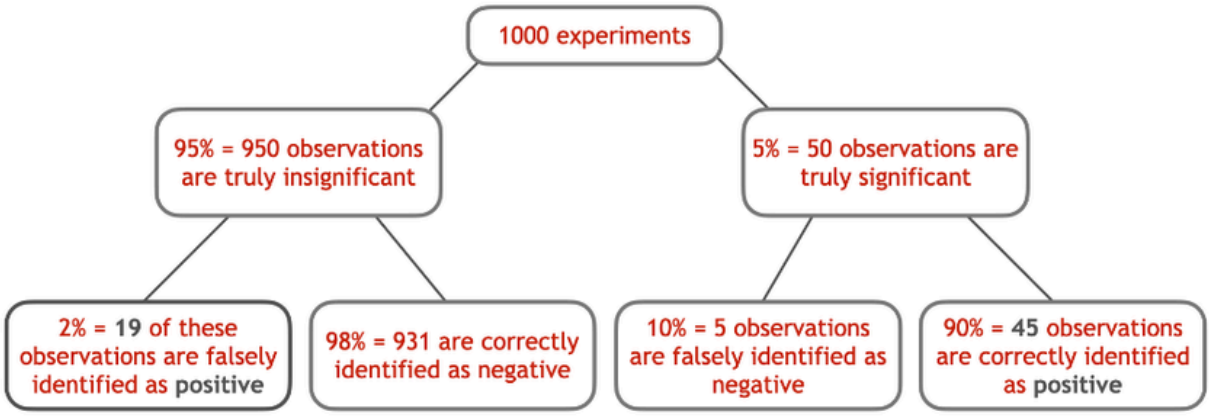
Example calculation of the false discovery rate (FDR)

$$\text{FDR} = \text{Number\_falsely\_significant} \ / \ \text{Number\_deemed\_significant}$$

Imagine that we are doing a genome-wide study looking at differential gene expression between tumor tissue and healthy tissue, and we tested 1000 genes. The image below shows that 950 (95%) of the null hypotheses are actually true, and 50 (5%) of the null hypotheses are actually false ("significant").

Of the 950 observations where the null hypothesis was actually true, 19 were incorrectly rejected, or deemed as significant (box at bottom left).

Of the 50 observations that were truly significant, 45 were correctly identified as significant (box at bottom right).



$$\text{FDR} = 19 \ / \ 64 = 30\%$$

Therefore, out of the 1000 experiments, our analysis identified 45 true positive results and 19 false positive results for a total of 64 positive results. Of these results, 19/64 are false positives so the false discovery rate is **30%**, the percentage of the rejected null hypotheses that were erroneously rejected.

Once again, the **False Discovery Rate** is the proportion of all outcomes deemed to be significant that are falsely significant.

Alert: Only the positive outcomes matter for the false discovery rate--only false positives or true positives. False negatives don't influence the false discovery rate.

Note: There are various ways to control the False Discover Rate. A common approach is known as the **Benjamini-Hochberg procedure**.

Summary

When we perform statistical tests, we are trying to see if there's a significant difference or effect in our data. But sometimes, we might get a result that looks significant purely by chance — these are called false positives.

Now, imagine you perform many tests. Some of them show significant results. But among those, some might be false positives — meaning they appear significant, but they're not truly meaningful. The False Discovery Rate (FDR) is the percentage of these significant results that are actually false positives.

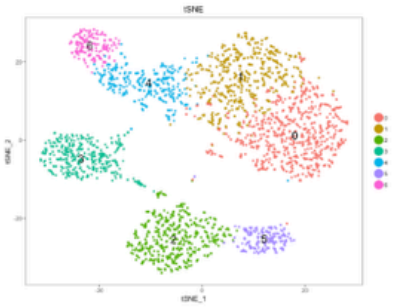
In other words, FDR helps us understand how many of our "significant findings" might just be random noise, not real effects.

Lesson 10: Dimensionality Reduction Methods: t-SNE

T-distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction algorithm that has become widely used to visualize high-dimensional genomic or proteomic data sets in a low-dimensional space (e.g., 2D or 3D). T-SNE has an incredible ability to take a set of points in a high-dimensional space (usually with hundreds or even thousands of variables) and find a faithful representation of those points in a lower-dimensional space, typically 2D, allowing the results to be plotted in a simple scatter plot.

T-SNE is similar to PCA but, unlike PCA, it allows us to separate data that cannot be separated by any straight line, known as nonlinear dimensionality reduction.

T-SNE is mostly used for visualization purposes and not for detailed quantitative analysis.



✅ Run each of the cells below:

Sample high-dimensional data set

In [11]:
```
from bioinfokit.analys import get_data
df = get_data('digits').data
df.head(2)
```

Out[11]:

	pixel_0_0	pixel_0_1	pixel_0_2	pixel_0_3	pixel_0_4	pixel_0_5	pixel_0_6	pixel_0_7	pixel_1_0	pixel_1_1	...	pixel_6_7	pixel_7_0	pixel_7_1	pixel_7_2	pix
0	0.0	0.0	5.0	13.0	9.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	6.0	
1	0.0	0.0	0.0	12.0	13.0	5.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	

2 rows × 65 columns

1797 rows, 65 columns

In [2]:
```
df.shape
```

Out[2]:
```
(1797, 65)
```

Run t-SNE

In [3]:
```
from sklearn.manifold import TSNE

tsne_em = TSNE(n_components=2, perplexity=30.0, n_iter=1000, verbose=1).fit_transform(df)
```

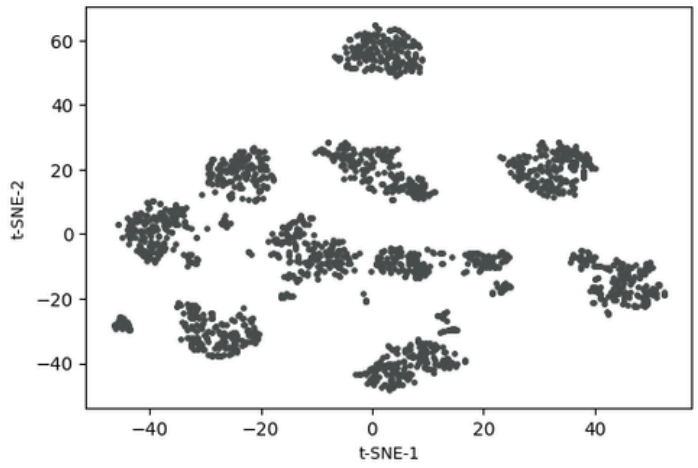
```
[t-SNE] Computing 91 nearest neighbors...
[t-SNE] Indexed 1797 samples in 0.001s...
[t-SNE] Computed neighbors for 1797 samples in 0.771s...
[t-SNE] Computed conditional probabilities for sample 1000 / 1797
[t-SNE] Computed conditional probabilities for sample 1797 / 1797
[t-SNE] Mean sigma: 11.619740
[t-SNE] KL divergence after 250 iterations with early exaggeration: 60.883026
[t-SNE] KL divergence after 1000 iterations: 0.742325
```

Plot t-SNE clusters

View high-dimensional data as 2D clusters

In [5]:
```
from bioinfokit.visuz import cluster
cluster.tsneplot(score=tsne_em, show=True)
```

```
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
```

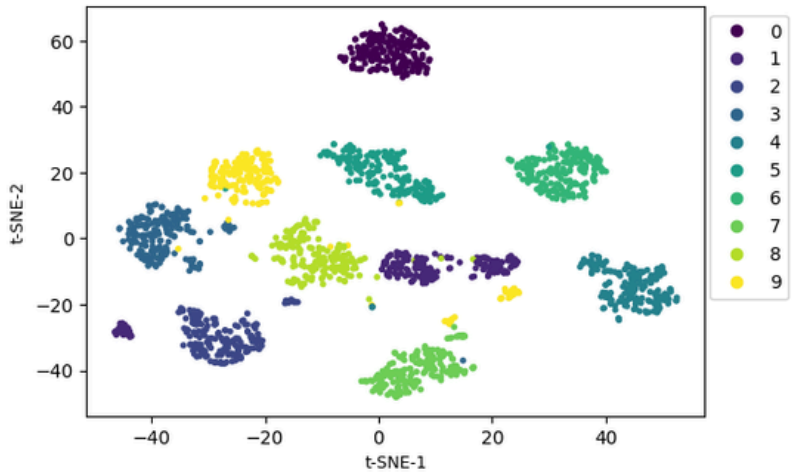


Add colors to the clusters

This will help to color and visualize clusters of similar data points. Get a list of categories.

In [6]:
```
color_class = df['class'].to_numpy()
cluster.tsneplot(score=tsne_em, colorlist=color_class, legendpos='upper right', legendanchor=(1.15, 1), show=True )
```

```
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
findfont: Font family 'Arial' not found.
```



Interpretation

The points within the individual clusters are very similar to each other and are less similar to points in other clusters. A similar pattern likely is present in the original, high-dimensional data set.

In the context of scRNA-seq, for example, these clusters would represent the cell types with similar transcriptional profiles.

NOTES:

- t-SNE is a stochastic method and produces slightly different embeddings if run multiple times. These different results could affect the numeric values on the axis but do not affect the clustering of the points.
- t-SNE has a parameter called perplexity that measures the effective number of neighbors and controls the trade-off between global high-dimensional and local low-dimensional space. We can tune the perplexity parameter to influence the structure of the clusters and how they are displayed.

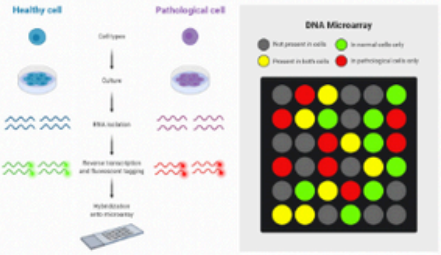
Tip: t-SNE is computationally expensive and can take several hours on large datasets. PCA is much faster and can be run before running t-SNE to reduce the number of original variables.

Lesson 4: Log2 Fold Change

Lesson 4: Log2 Fold Change

A **fold change** is a measure describing how much a quantity changes between an initial and a subsequent measurement. This is often used when comparing various measurements of a biological system taken at different times. For example, if a quantity changes from 50 to 100 over a given period of time, this is defined as a two-fold increase (i.e., a fold change of **2**). Similarly, a change from 100 to 50 would be referred to as a 0.5-fold decrease (i.e., a fold change of **.5**).

Commonly, fold change is used in the analysis of gene expression data from microarray experiments for measuring a change in the expression level of a gene.

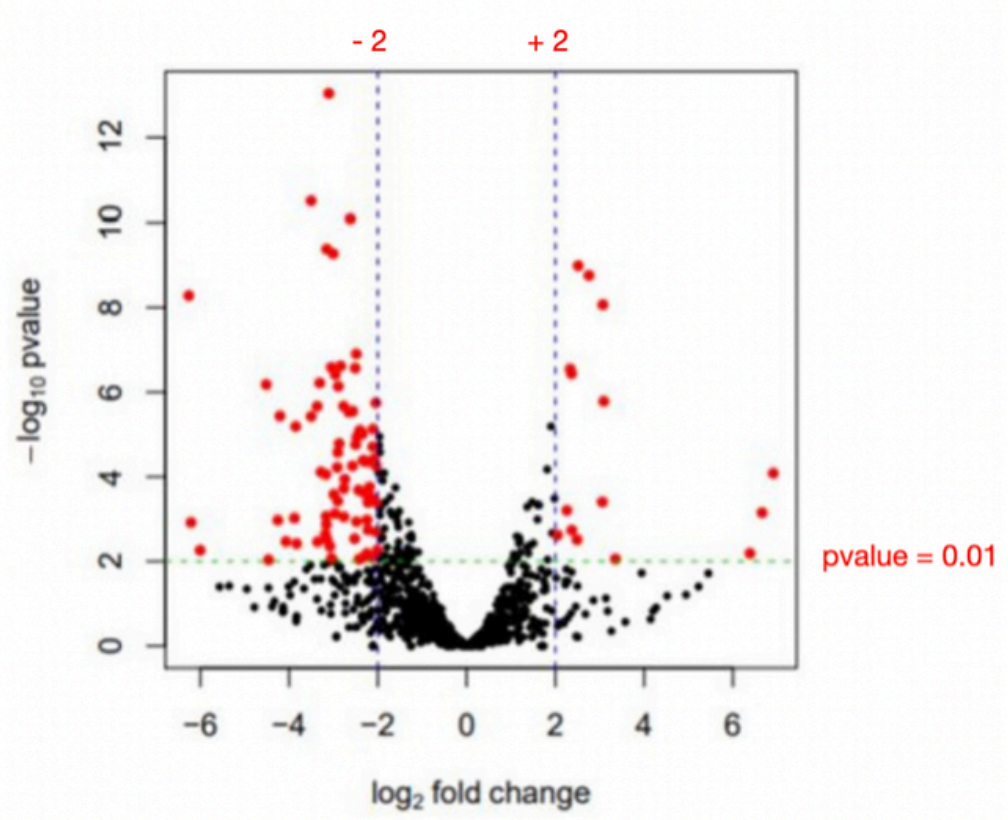


Log_2-Fold Change

This is the effect size estimate. This value indicates, for example, how much the gene or transcript's expression seems to have changed between the comparison and control groups. For example, suppose there are two gene expression values: A for the initial measurement, and B for the treatment. If A = 50 and B = 75, then the **fold change** is B/A (i.e., 1.5). The **log_2-fold change** would be $log_2(1.5) = .58$.

Why Log_2-Fold Change:

When analyzing and visualzing fold changes, this value is reported on a logarithmic scale to base 2 (i.e., log_2). This is because it is easy to interpret. For example, doubling (2) the initial quantity is equal to a log_2 fold change of 1 (i.e., $log_2(2) = 1$). And quadrupling an initial quantity is equal to a log_2 fold change of 2 (i.e., $log_2(4) = 2$). Further, a nice property of log_2 is that it is symmetric for reciprocals. For example, conversely, when the initial quantity is decreased by half, this is equivalent to a log_2 fold change of −1, and quartering an initial quantity is equivalent to a log_2 fold change of −2 and so on. This leads to more aesthetically pleasing plots as exponential changes are displayed as linear, so the dynamic range is increased. For example, on a plot axis showing log2 fold changes, an 8-fold increase will be displayed at an axis value of 3 since $log_2(8) = 3$.



In the volcano plot shown above, the red points indicate genes that display both large-magnitude fold changes (x-axis) as well as high statistical significance (-$\log_{10}$ p-value, y-axis). The dashed green line shows the p-vaule cutoff of 0.01 (10^{-2}) with points above the line having a p-value < 0.01 and points below the line having a p-value > 0.01. The vertical dashed blue lines indicate $\log_2$-fold changes of 2. **Therefore, all red dots exhibit $\log_2$-fold changes beyond ±2 (four-fold change) and statistical significance less than 0.01.**

Tip: The formula for the **$\log_2$-fold change** is: $\log_2(B) - \log_2(A)$
The fold change = $2^{\log_2 FC}$

Note: $\log_2(x) = \log_{10}(x)/\log_{10}(2)$

Lesson 2: P-value

The term **p-value** is used when you want to test a hypothesis. Let's look at an example.

Let's say that you believe that your food delivery service has gotten slower recently. You determine that previous delivery times averaged right at about 30 minutes, but you believe that it has slowed significantly over the last month or so.

You decide to test your hypothesis. Over the next month, you decide to time each of your food deliveries. You are able to collect about 20 samples, and it turns out that the average delivery time is 37 minutes. Can you say that the previous average of 30 minutes and the average of 37 minutes that you recently collected are significantly different?

In other words, are the delivery times actually slower, or did you just happen to get a few slower delivery persons by chance while, on the whole, the delivery times actually remain at the 30-minute average?

Null Hypothesis versus Alternative Hypothesis

When you establish a hypothesis, it's typically stated in the form of the status quo. For example, in this case, you would state that there is no difference in average delivery times between the current deliveries and previous deliveries (known as the **null hypothesis**).

You will try to gather evidence against this hypothesis in order to support your belief (**alternative hypothesis**) that there is a significant difference between the time it takes to have food delivered currently versus a few months ago.

Null Hypothesis:	Current food delivery times are no different than a few months ago.
Alternative Hypothesis:	Current food delivery times are significantly slower than a few months ago.

Hypothesis Testing

You would like to test the significance of your observed difference between previous and current delivery times. Is the 7-minute difference (30 minutes previously versus your recently obtained average of 37 minutes) significant, or is it just **due to randomness in your 20 samples**?

Significance levels

The level of statistical significance is often expressed as the so-called **p-value** (probability value). You will calculate the probability of observing your sample results (or more extreme), given that the null hypothesis is true. In other words, if there really is no difference between previous and current delivery times, how likely would it be to see a difference as large as (or larger than) that which you observed in your samples?

p-value

Let's say that, after evaluating your results, you get a p-value of 0.04 (p = .04). This means that there is a 4% chance of finding a difference as large as (or larger than) the one that you obtained, given that the null hypothesis is true. In terms of significance, typically a p-value of 0.05 is used as the threshold. This is interpreted as, assuming that the null hypothesis is true, if there is a 5% chance or less of observing a difference as extreme (or larger) as you observed, you would reject the null hypothesis, and accept the alternative hypothesis. Alternately, if the probability was greater than 0.05, you would fail to reject the null hypothesis. This is because the result that you obtained could happen too frequently by chance to be confident that the current delivery times are truly different from previous deliveries.

Alert: Note that you cannot accept the null hypothesis, but only reject, and find evidence against it.

In our example, where p = .04, you would **reject the null hypothesis, and accept the alternative hypothesis** that current delivery times are significantly slower than previous deliveries.

Tip: A p-value of 0.05 is commonly used as the threshold for significance. However, when increased confidence is desired, a more stringent p-value of 0.01 may be used.

Understanding the p-value

- The p-value or calculated probability provides a universal language to interpret test results.
- The p-value is a number between 0 and 1 that provides the statistical significance of hypothesis testing.
- The p-value tests whether there is enough evidence to reject the null hypothesis.

Interpreting significance

p-value less than 0.05

- A small p-value (< 0.05) indicates that the result is possible, but not very likely under the null hypothesis.
- Thus, for a hypothesis with a p-value less than 0.05, the null hypothesis is rejected, and the alternative hypothesis is accepted.
- This suggests that the results of the study are statistically significant.

p-value greater than 0.05

- If the p-value is large (> 0.05), it indicates weak evidence against the null hypothesis.
- Thus, for a hypothesis with a p-value greater than 0.05, the null hypothesis is not rejected, and the alternative hypothesis is not accepted.
- This indicates that the results of the study are not statistically significant.