Regression and Statistics

NENS 230: Analysis Techniques in Neuroscience

Outline

Summary statistics functions

Random Variables

- Random variables, PDF, CDFs
- Estimates of central tendency and dispersion
- Standard error of the mean, confidence intervals

Statistical Hypothesis Testing

- Tests and significance
- Student’s t test walkthrough
- Other commonly used tests

Analysis of Variance

Regression

- Linear regression
- Curve fitting

Dimensionality Reduction

- PCA

Assignment 6 Overview

Outline

Summary statistics functions

Random Variables

- Random variables, PDF, CDFs
- Estimates of central tendency and dispersion
- Standard error of the mean, confidence intervals

Statistical Hypothesis Testing

- Tests and significance
- Student's t test walkthrough
- Other commonly used tests

Analysis of Variance

Regression

- Linear regression
- Curve fitting

Dimensionality Reduction

- PCA

Assignment 6 Overview

`mean()` function

`mean()` function

`mean()` computes the average (sample mean) of a vector. When dealing with matrices, you need to specify which dimension to average along.

`mean()` function

`mean()` computes the average (sample mean) of a vector. When dealing with matrices, you need to specify which dimension to average along.

`mean(X, 1)` means return the average row (average down the column, across the rows). This is the default if you only specify one argument.

`mean()` function

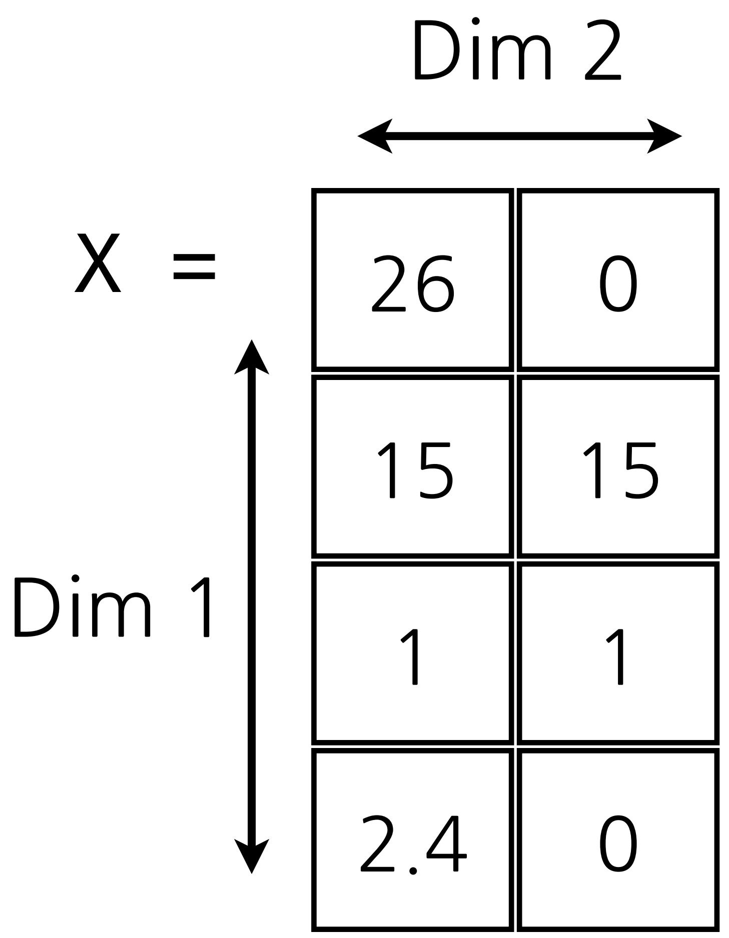
`mean()` computes the average (sample mean) of a vector. When dealing with matrices, you need to specify which dimension to average along.

`mean(X, 1)` means return the average row (average down the column, across the rows). This is the default if you only specify one argument.

`mean(X, 2)` means return the average column (average across the columns, down the row)

`mean()` function

`mean()` computes the average (sample mean) of a vector. When dealing with matrices, you need to specify which dimension to average along.



$$X = \begin{bmatrix} 26 & 0 \\ 15 & 15 \\ 1 & 1 \\ 2.4 & 0 \end{bmatrix}$$

`mean(X)`
`mean(X, 1)` evaluates to $\begin{bmatrix} 11.1 & 4 \end{bmatrix}$

`mean(X, 2)` evaluates to $\begin{bmatrix} 13 \\ 15 \\ 1 \\ 1.2 \end{bmatrix}$

`mean()` function

`mean()` function

`mean()` operates on its first argument. Be careful when averaging two things together that you pack them in a vector using `[ ]`

`mean()` function

`mean()` operates on its first argument. Be careful when averaging two things together that you pack them in a vector using `[ ]`

`mean(1, 5)` evaluates to `1`

“Take the mean of `[1]` along the 5th dimension”

`mean()` function

`mean()` operates on its first argument. Be careful when averaging two things together that you pack them in a vector using `[ ]`

~~`mean(1, 5)` evaluates to `1`~~

~~“Take the mean of `[1]` along the 5th dimension”~~

`mean()` function

`mean()` operates on its first argument. Be careful when averaging two things together that you pack them in a vector using `[ ]`

~~`mean(1, 5)` evaluates to 1~~

~~“Take the mean of `[1]` along the 5th dimension”~~

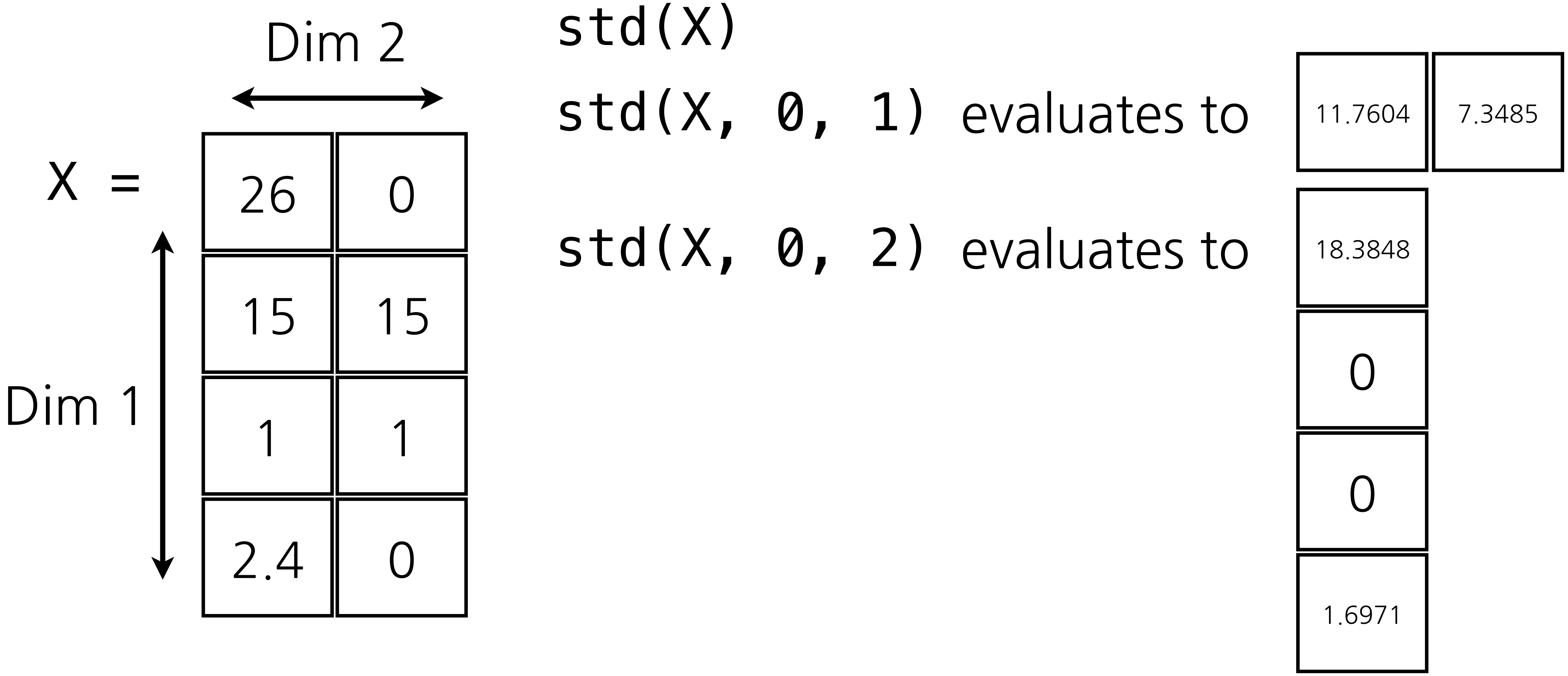
`mean([1 5])` evaluates to 3

“Take the mean value of [4 5]
(along the first non-singleton dimension)”

`std()` function

`std()` computes the standard deviation of a list of numbers

- When dealing with matrices, you need to specify which dimension to average along, **as the third argument.**
- The second argument should be `0` if you want the unbiased estimator that normalizes by `n-1`, where `n` is the number of samples



X =

26	0
15	15
1	1
2.4	0

`std(X)`

`std(X, 0, 1)` evaluates to

11.7604	7.3485

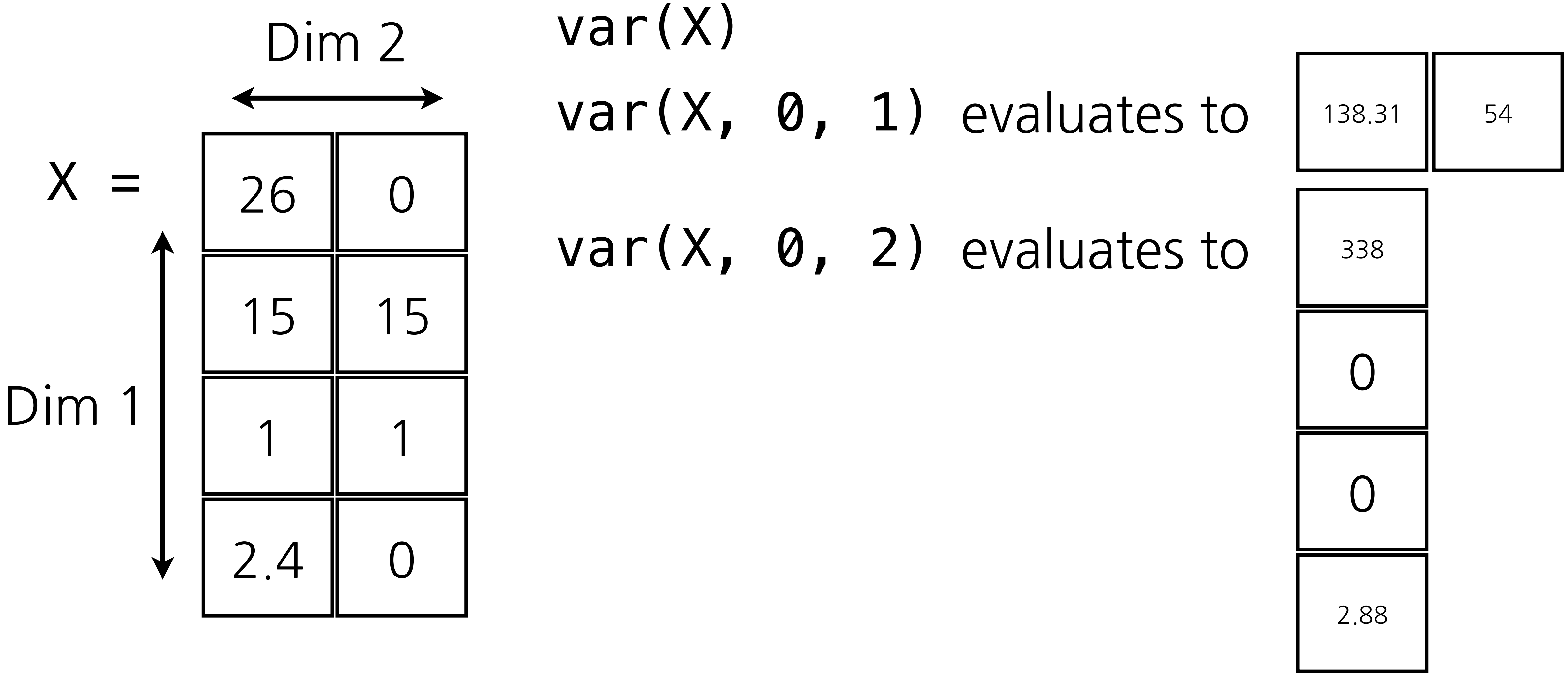
`std(X, 0, 2)` evaluates to

18.3848
0
0
1.6971

`var()` function

`var()` computes the sample variance of a list of numbers

- When dealing with matrices, you need to specify which dimension to operate along, **as the third argument.**
- The second argument should be `0` if you want the unbiased estimator that normalizes by `n-1`, where `n` is the number of samples. (This is the default)



$$X = \begin{array}{c|cc} & \multicolumn{2}{c}{\text{Dim 2}} \\ \hline \text{Dim 1} & 26 & 0 \\ & 15 & 15 \\ & 1 & 1 \\ & 2.4 & 0 \end{array}$$

```
var(X)
var(X, 0, 1)
```
evaluates to

138.31	54

```
var(X, 0, 2)
```
evaluates to

338
0
0
2.88

`sum()` function

`sum()` computes the sum of a vector. When dealing with matrices, you should specify which dimension to average along.

`sum(X, 1)` means return the sum over rows (sum over rows within each column). This is the default if you only specify one argument.

`sum(X, 2)` means return the sum over columns (sum over columns within each row)

`min()` function

`min()` computes the minimum of a vector. When dealing with matrices, you should specify which dimension to find the minimum along.

`min(X, Y)` means return an array the same size as X and Y consisting of the smaller of the elements in X and Y at each location.

`min(X, [], 1)` means return the minimum over rows (over rows within each column). This is the default if you only specify one argument.

`min(X, [], 2)` means return the minimum over columns (over columns within each row)

`max()` function

`max()` computes the maximum of a vector. When dealing with matrices, you should specify which dimension to find the maximum along.

`max(X, Y)` means return an array the same size as X and Y consisting of the larger of the elements in X and Y at each location.

`max(X, [], 1)` means return the maximum over rows (over rows within each column). This is the default if you only specify one argument.

`max(X, [], 2)` means return the maximum over columns (over columns within each row)

Outline

Summary statistics functions

Random Variables

- Random variables, PDF, CDFs
- Estimates of central tendency and dispersion
- Standard error of the mean, confidence intervals

Statistical Hypothesis Testing

- Tests and significance
- Student's t test walkthrough
- Other commonly used tests

Analysis of Variance

Regression

- Linear regression
- Curve fitting

Dimensionality Reduction

- PCA

Assignment 6 Overview

Discrete random variables

Suppose we have a random variable X.

Discrete random variables take one value within a set of k possible values.

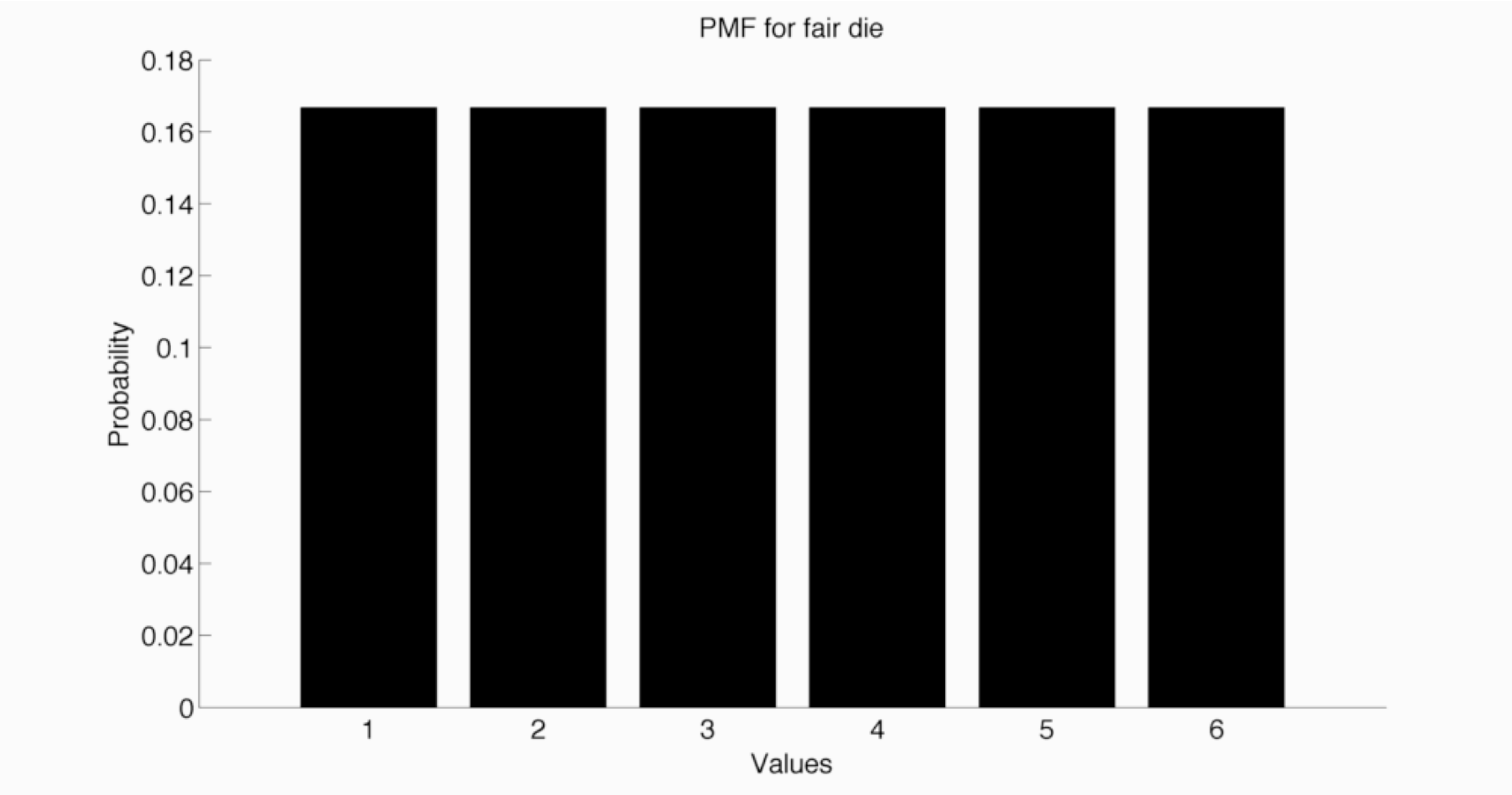
Probability mass function: For a given value x_i returns the probability p_i of X taking that value.

$$Pr[X = x_i] = p_i$$

Sum of these probabilities must be 1.

$$p_1 + p_2 + \cdots + p_k = 1$$

Probability Mass Function



Continuous random variables

Suppose we have a random variable X.

Continuous random variables take values within some continuous range of values.

Probability density function (PDF): integrating this function over some interval gives you the probability that X lies in that interval.

$$Pr[a \leq X \leq b] = \int_a^b f(x)dx$$

Therefore, the integral under this function is 1.

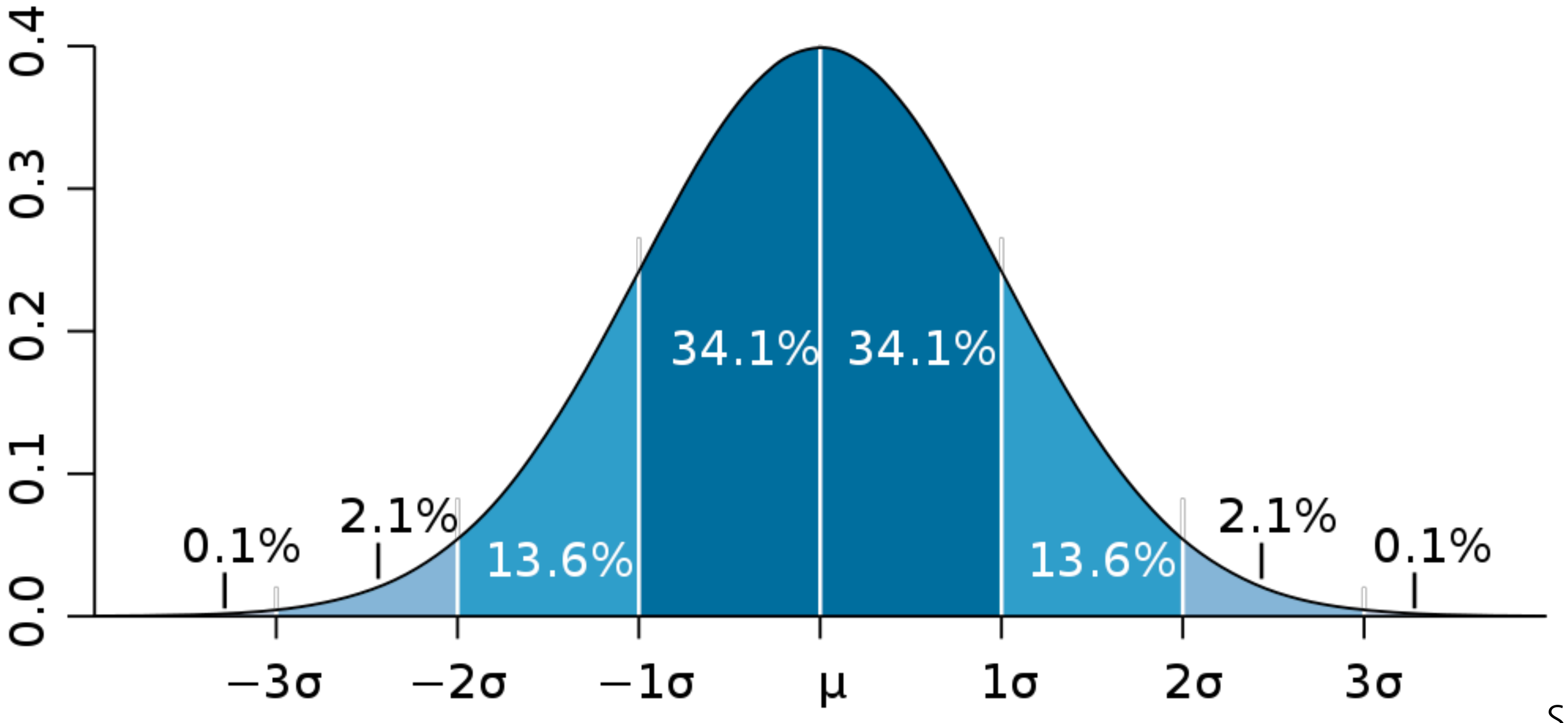
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Normal distribution

Normal or Gaussian distributions describe many naturally occurring phenomena, due to the central limit theorem.

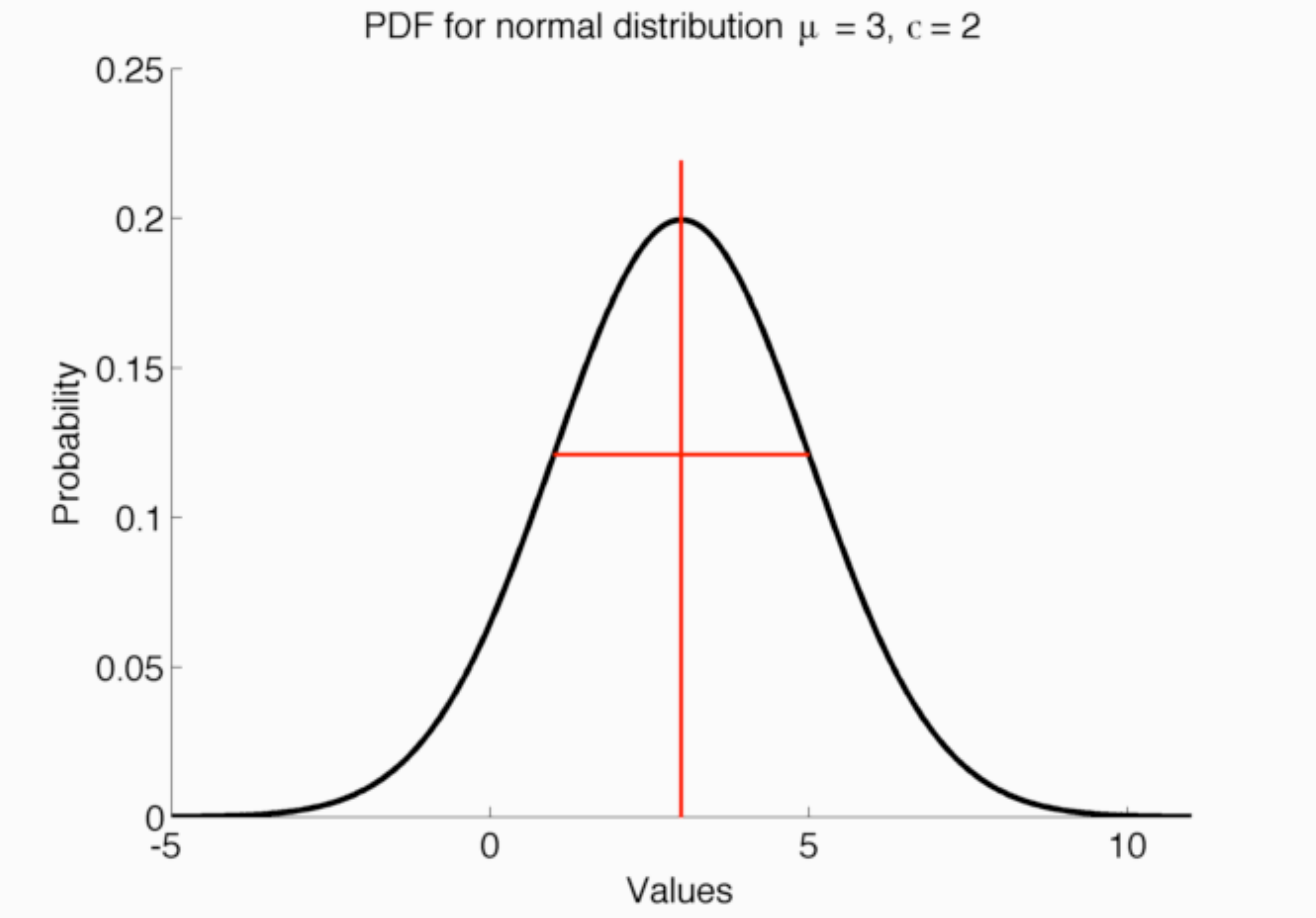
Specified by two parameters:

- **Location parameter:** the mean (μ)
- **Scale parameter:** the standard deviation (σ)

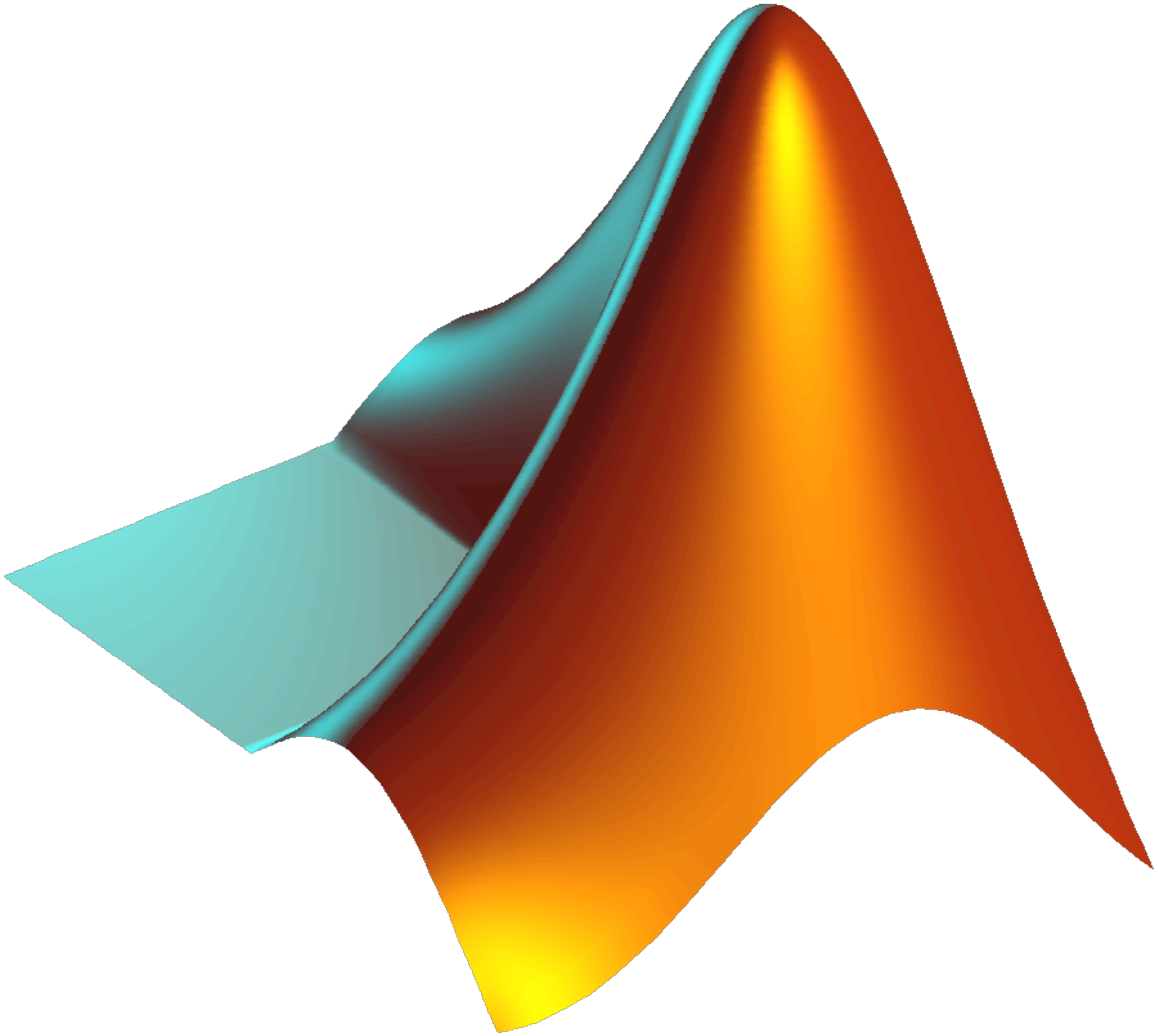


Source: wikipedia.org

PDF for normal distribution



Demo: PDF for Normal distribution



Cumulative distribution function

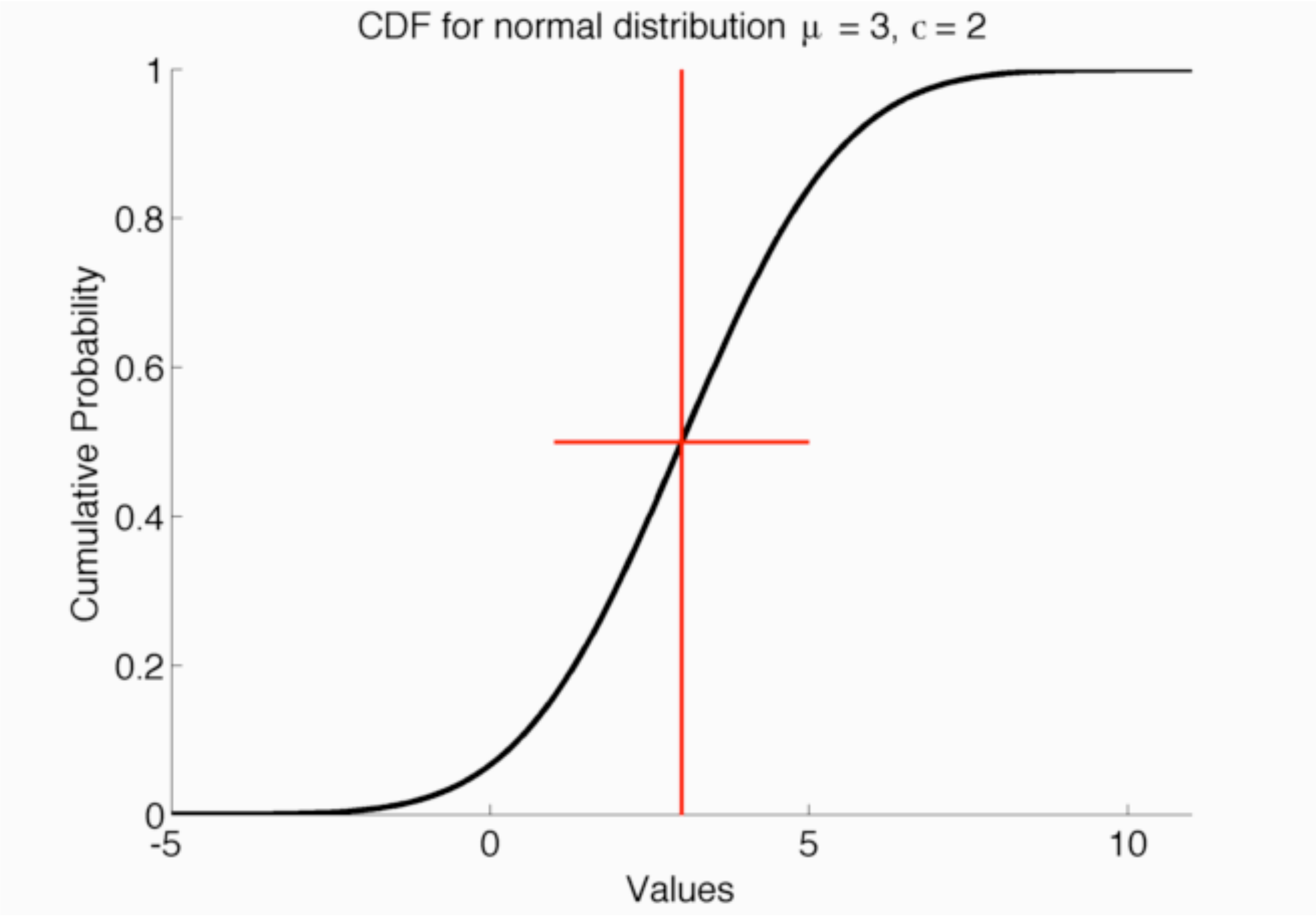
Cumulative distribution function (CDF): how likely is X less than or equal to a particular value.

$$Pr[X \leq x] = F(x)$$

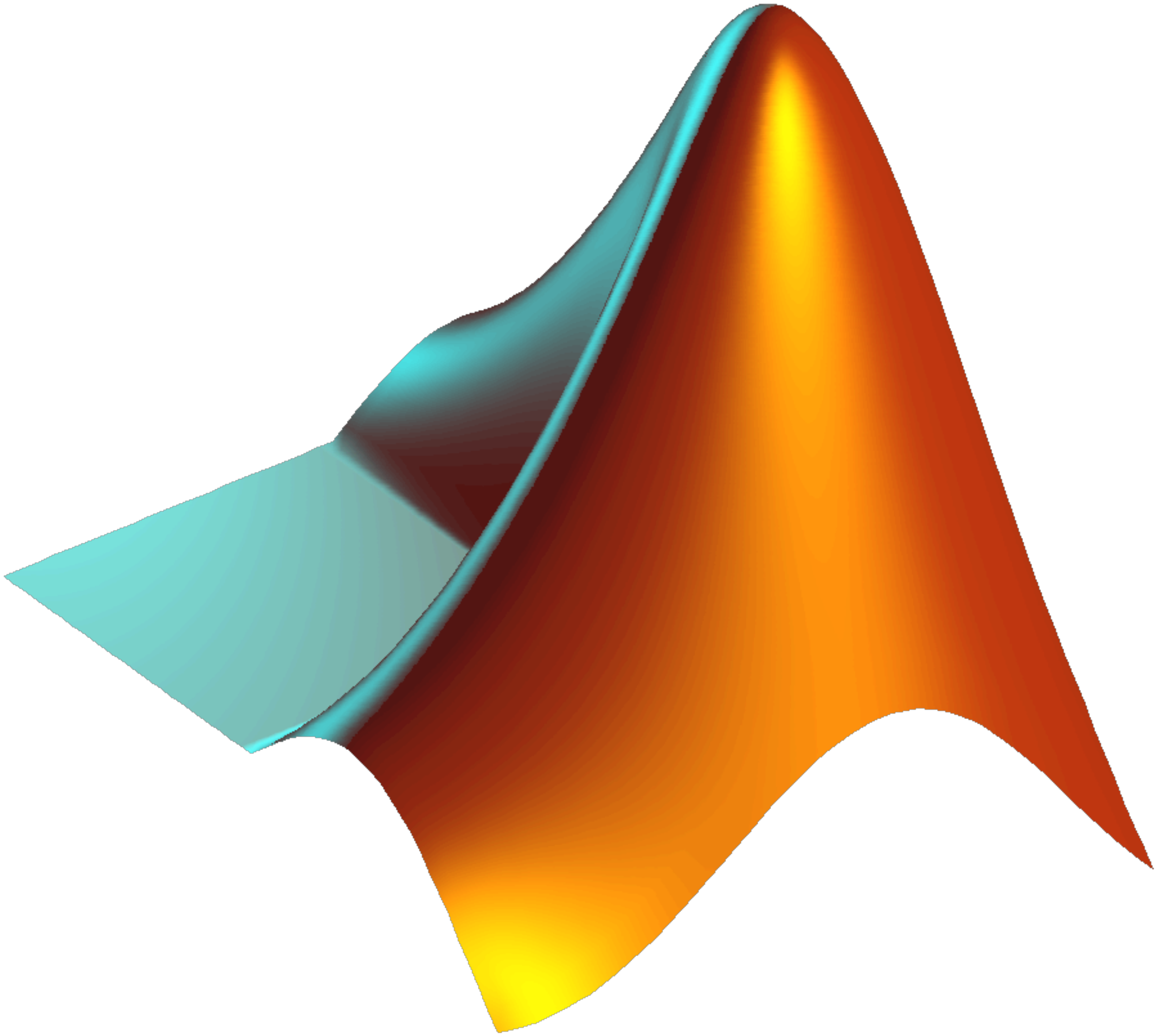
The CDF is the integral of the PDF.

The PDF is the derivative of the CDF. Therefore, the parts of the CDF with the steepest slope are the highest points of the PDF, i.e. where most of the values lie.

CDF for normal distribution



Demo: CDF for Normal distribution



Expected Value

The expected value of a random variable is it's mean. You can calculate the expected value of a random variable X by taking the weighted average of all its possible values. The weights are the probability of X taking each value.

Discrete RV: $$E[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

Continuous RV: $$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

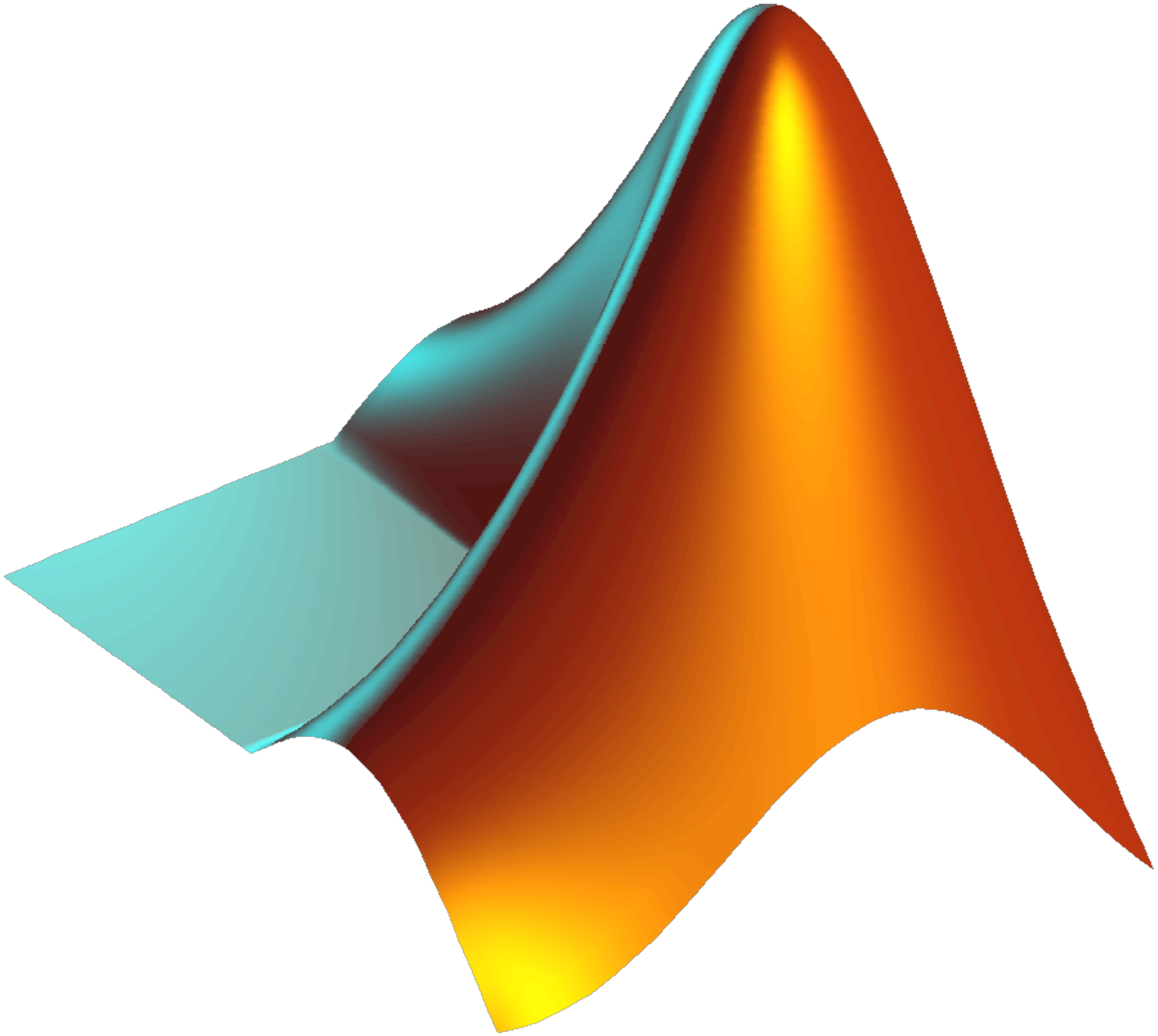
Sample mean

Sampling: When we measure some quantity in an experiment, we think of it as taking samples from a distribution.

Sample mean: By taking the average, we are estimating the mean or expected value of the underlying distribution which generated these quantities.

A central problem in statistics: How close is this estimate of the mean (the average of our samples) to the true, underlying mean?

Demo: Sampling Distribution for Sample Mean



Standard Error of the Mean

Suppose we make N measurements of X, sampling from a normal distribution with mean μ and standard deviation σ.

If we take the average of these N samples, our **estimate of the mean is a normal distribution**.

The mean of this sampling distribution is μ

The standard deviation is σ / sqrt(N).

This means that on average, our estimate will be correct. The spread around the true mean shrinks as 1/sqrt(N).

Standard Error of the Mean

Suppose we make N measurements of X which may or not be normally distributed.

If we take the average of these N samples, our estimate of the mean **approaches** a normal distribution as N gets larger (central limit theorem).

The mean of this sampling distribution is μ

The standard deviation is σ / sqrt(N).

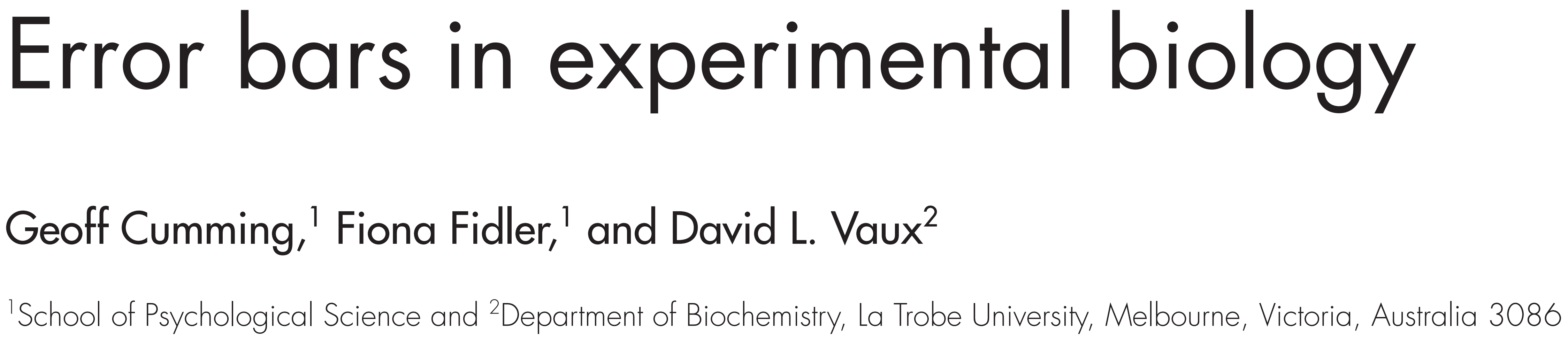
Confidence intervals

Based on the data you've collected, you can estimate the true value of some quantity, e.g. the true mean.

This estimate of the quantity isn't perfect. Confidence intervals tell you a range of values where the true value lies with some probability

95% confidence intervals are the range where the true value of the quantity will lie with 95% probability.

More information



Error bars in experimental biology

Geoff Cumming,[1] Fiona Fidler,[1] and David L. Vaux[2]

[1]School of Psychological Science and [2]Department of Biochemistry, La Trobe University, Melbourne, Victoria, Australia 3086

Concise introduction to standard deviation, standard error, and confidence intervals

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2064100/

Outline

Summary statistics functions

Random Variables

- Random variables, PDF, CDFs
- Estimates of central tendency and dispersion
- Standard error of the mean, confidence intervals

Statistical Hypothesis Testing

- Tests and significance
- Student's t test walkthrough
- Other commonly used tests

Analysis of Variance

Regression

- Linear regression
- Curve fitting

Dimensionality Reduction

- PCA

Assignment 6 Overview

Statistical hypothesis testing

Statistical hypothesis testing

The point of statistical tests is to cast doubt on the veracity of a **null hypothesis**.

Statistical hypothesis testing

The point of statistical tests is to cast doubt on the veracity of a **null hypothesis**.

We demonstrate this by stating that were the null hypothesis true, it would be very unlikely that we would observe the data we have observed.

Statistical hypothesis testing

The point of statistical tests is to cast doubt on the veracity of a **null hypothesis**.

We demonstrate this by stating that were the null hypothesis true, it would be very unlikely that we would observe the data we have observed.

Statistical tests reject the null hypothesis if the unlikelihood of the data crosses some threshold.

Statistical hypothesis testing

The point of statistical tests is to cast doubt on the veracity of a **null hypothesis**.

We demonstrate this by stating that were the null hypothesis true, it would be very unlikely that we would observe the data we have observed.

Statistical tests reject the null hypothesis if the un-likelihood of the data crosses some threshold.

This **threshold or significance level** is typically expressed as a **p-value**: the likelihood of false-rejections, i.e. the likelihood that the null hypothesis would be rejected if it were true.

Statistical hypothesis testing

Statistical hypothesis testing

State the null (and alterative) hypotheses

State the assumptions

- Independence of samples?
- Normality?

Determine an appropriate test statistic

Derive the distribution of the test statistic under the null hypothesis

Determine the critical region for the test statistic

Compute the observed value of the test statistic

Decide to fail to reject or to reject the null hypothesis

- Compute the strongest significance level at which the null hypothesis would be rejected (p-value)

Student's t-test example

Suppose we monitor scores on some behavioral assay before and after treatment. We take the difference of the scores for each subject.

Each subject's change in scores is x_i

State the null (and alterative) hypotheses:

Null hypothesis: x_i are drawn from a normal distribution with zero mean.

Alternative hypothesis: x_i are drawn from a normal distribution with non-zero mean.

Student’s t-test example

State the assumptions:

All samples are independent.

Changes in scores have a normal distribution. This follows from scores on the test before and after having normal distributions.

Test Statistics

Determine an appropriate test statistic.

A test statistic is a numerical summary of data that reduces the information needed to perform a hypothesis to a single value (or a small number of values).

The important point is that we know what this quantity's distribution would look like under the null hypothesis. If the test statistic computed from the data is very unlikely to be drawn from that distribution, we can reject the null hypothesis.

Student’s t-test example

Since the set of values are normally distributed, the Student’s t-test is appropriate. Therefore, the test statistic is:

$$T = \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}}$$

μ is zero, the mean for the null hypothesis

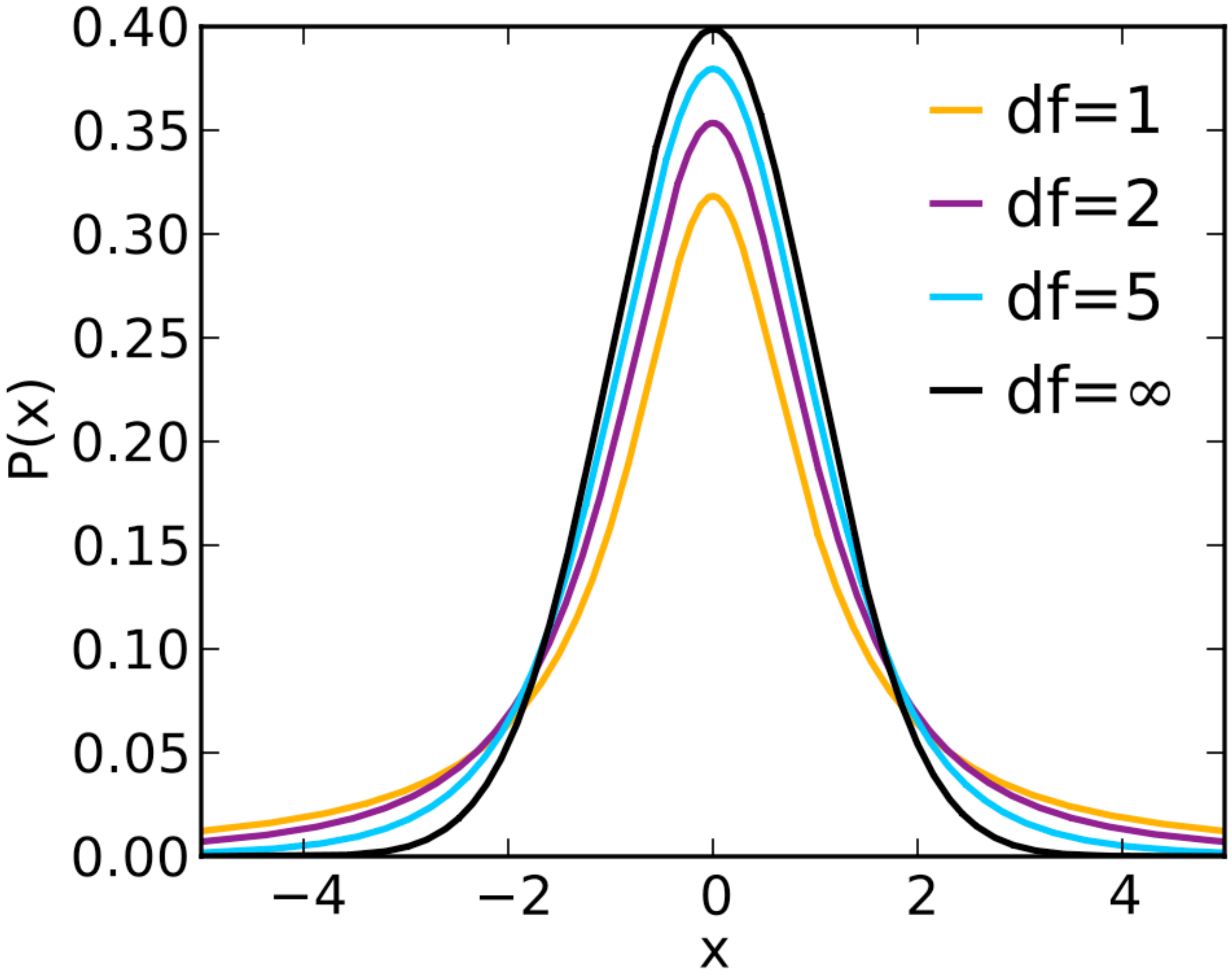
where $$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$ (sample mean)

and $$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$ (sample variance)

This T value has a Student’s t distribution with N-1 degrees of freedom.

Student’s t-test example

Derive the distribution of the test statistic under the null hypothesis: Student’s t-distribution, n-1 df



Source: wikipedia.org

Student’s t-test example

Determine the critical region for the test statistic.

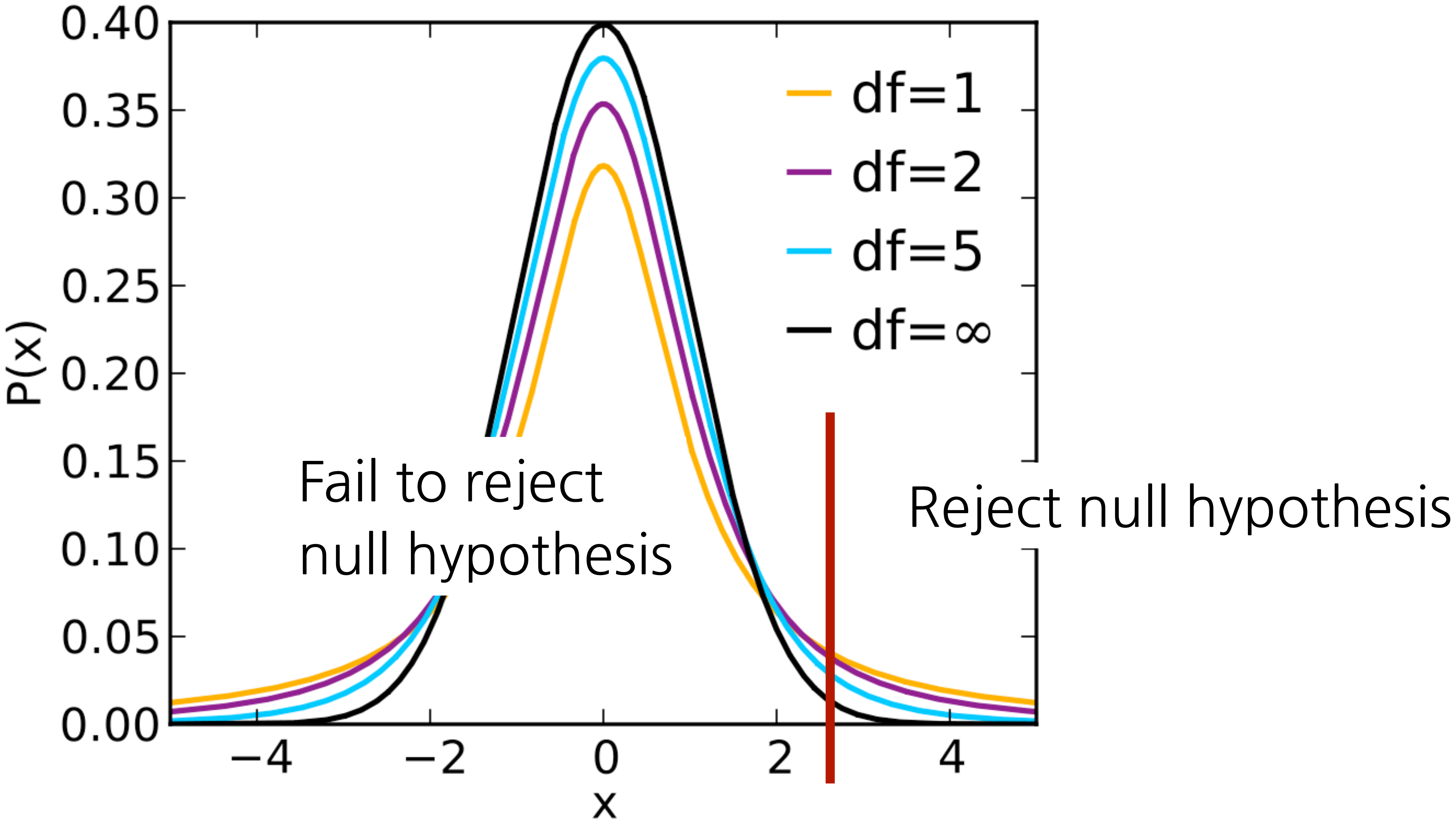
Suppose N=6. Then we have 5 degrees of freedom. At the 95% significance level, the critical value for T is **2.447**.

Thus, if $|T| \geq 2.447$, we reject the null hypothesis at the 95% significance level.

In other words, if the mean were really zero, $|T|$ would be larger than 2.447 only 5% of the time.

Student's t-test example

Determine the critical region for the test statistic.



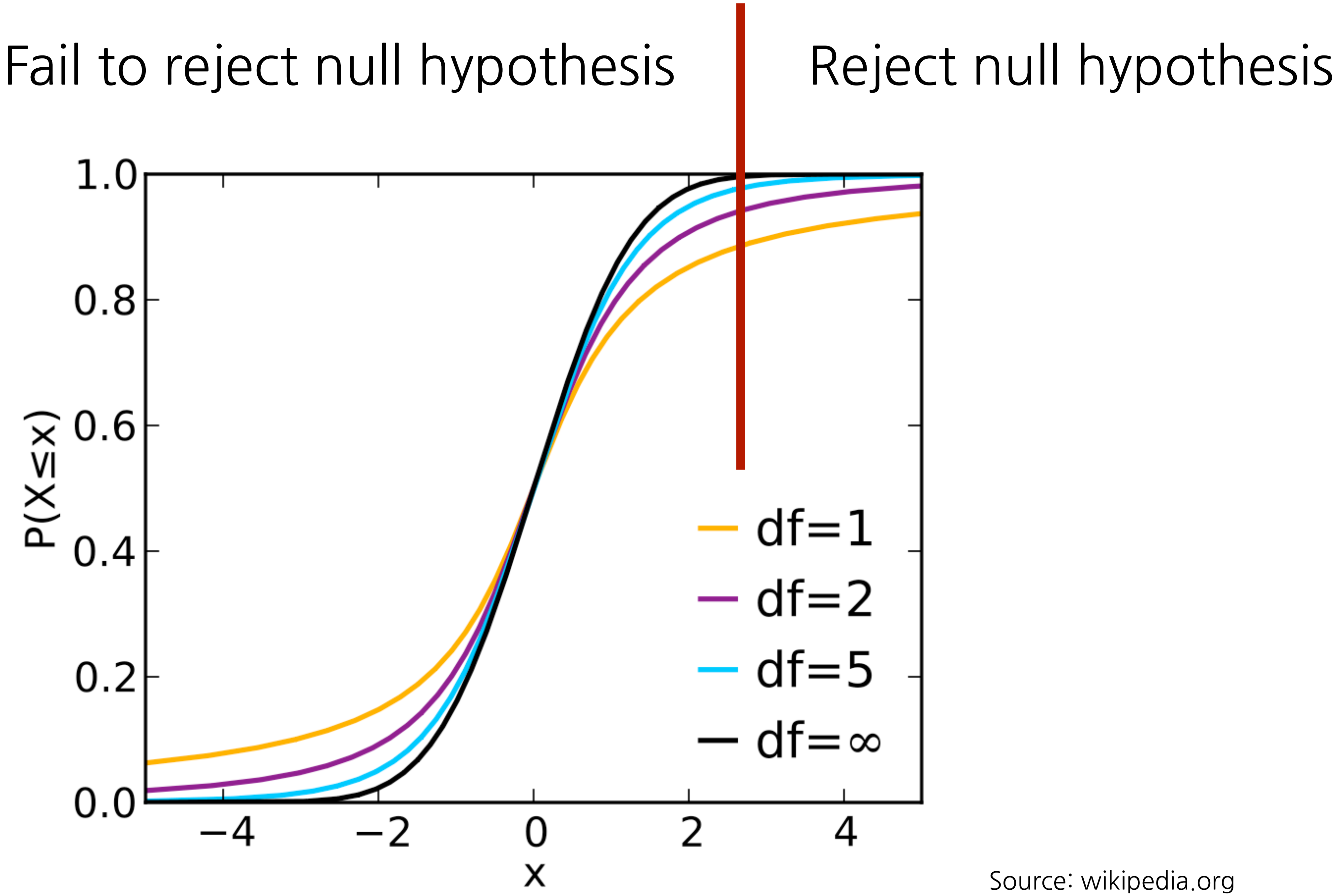
Fail to reject null hypothesis

Reject null hypothesis

Source: wikipedia.org

Student's t-test example

Determine the critical region for the test statistic.



Fail to reject null hypothesis

Reject null hypothesis

Source: wikipedia.org

Student’s t-test example

Compute the observed value of the test statistic.

Suppose we calculated T=3.1

Decide to fail to reject or to reject the null hypothesis.

98.66% of the t-distribution with 5 df lies to the left of 3.1.

Therefore we can reject the null hypothesis at the 95% level.

Our p-value is 0.0134

`ttest()` function

```
[h,p,ci,stats] = ttest(X)
```

Tests against the null hypothesis that the values in vector X are drawn from a normal distribution with zero mean.

`h`: true if the null hypothesis is rejected

`p`: p-value associated with the result

`ci`: 95% confidence intervals for the true value of the mean

`stats`: a structure that MATLAB can use for doing follow up tests, such as using `multcompare`.

ttest() function

```
[h,p,ci,stats] = ttest(X, nullMean)
```

Tests against the null hypothesis that the values in vector X are drawn from a normal distribution with mean **nullMean**.

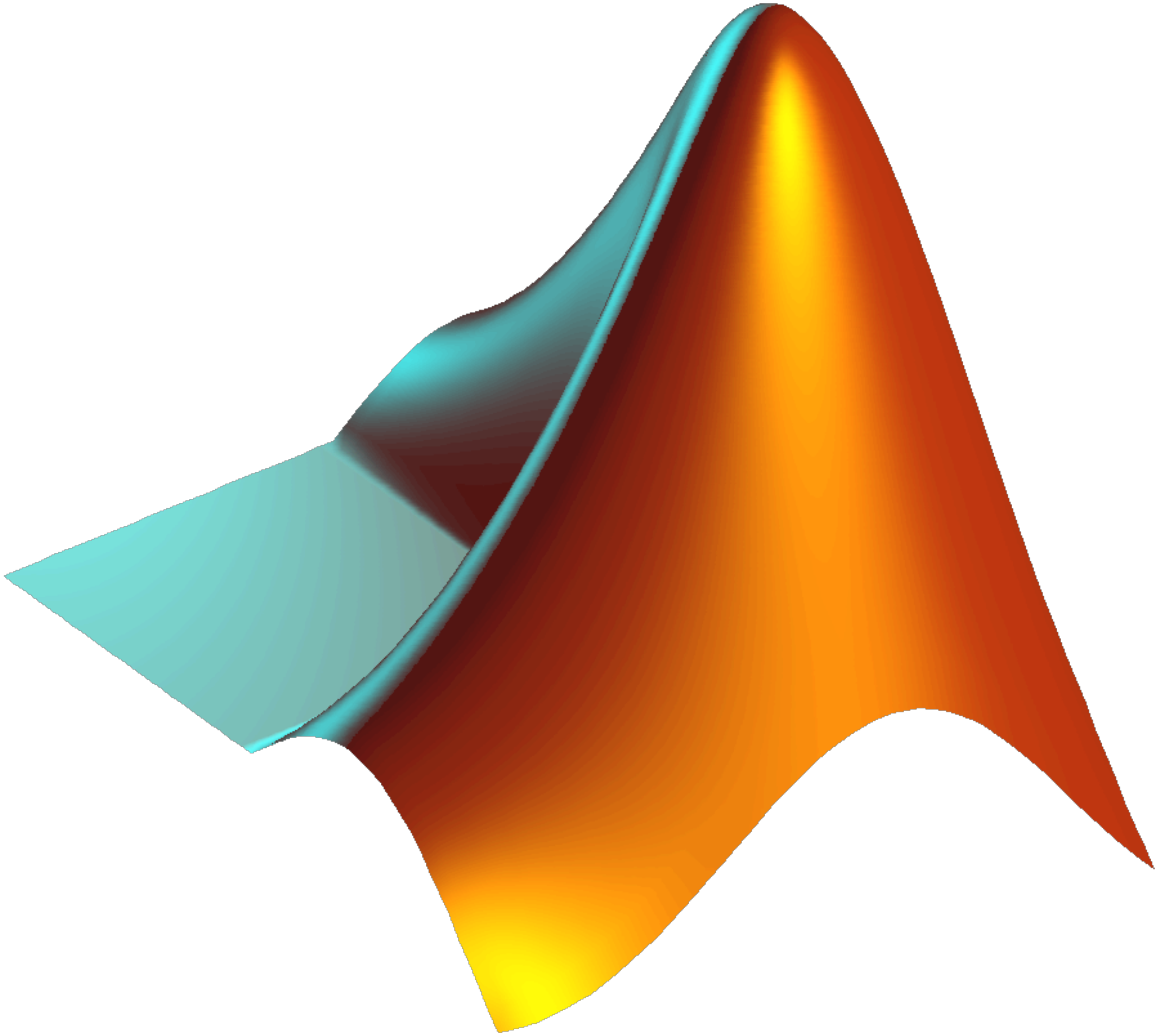
h: true if the null hypothesis is rejected

p: p-value associated with the result

ci: 95% confidence intervals for the true value of the mean

stats: a structure that MATLAB can use for doing follow up tests, such as using **multcompare**.

Demo: t-test demo



`ttest()` function

```
[h,p,ci,stats] = ttest(X, nullMean, thresh)
```

Tests against the null hypothesis that the values in vector X are drawn from a normal distribution with mean `nullMean`.

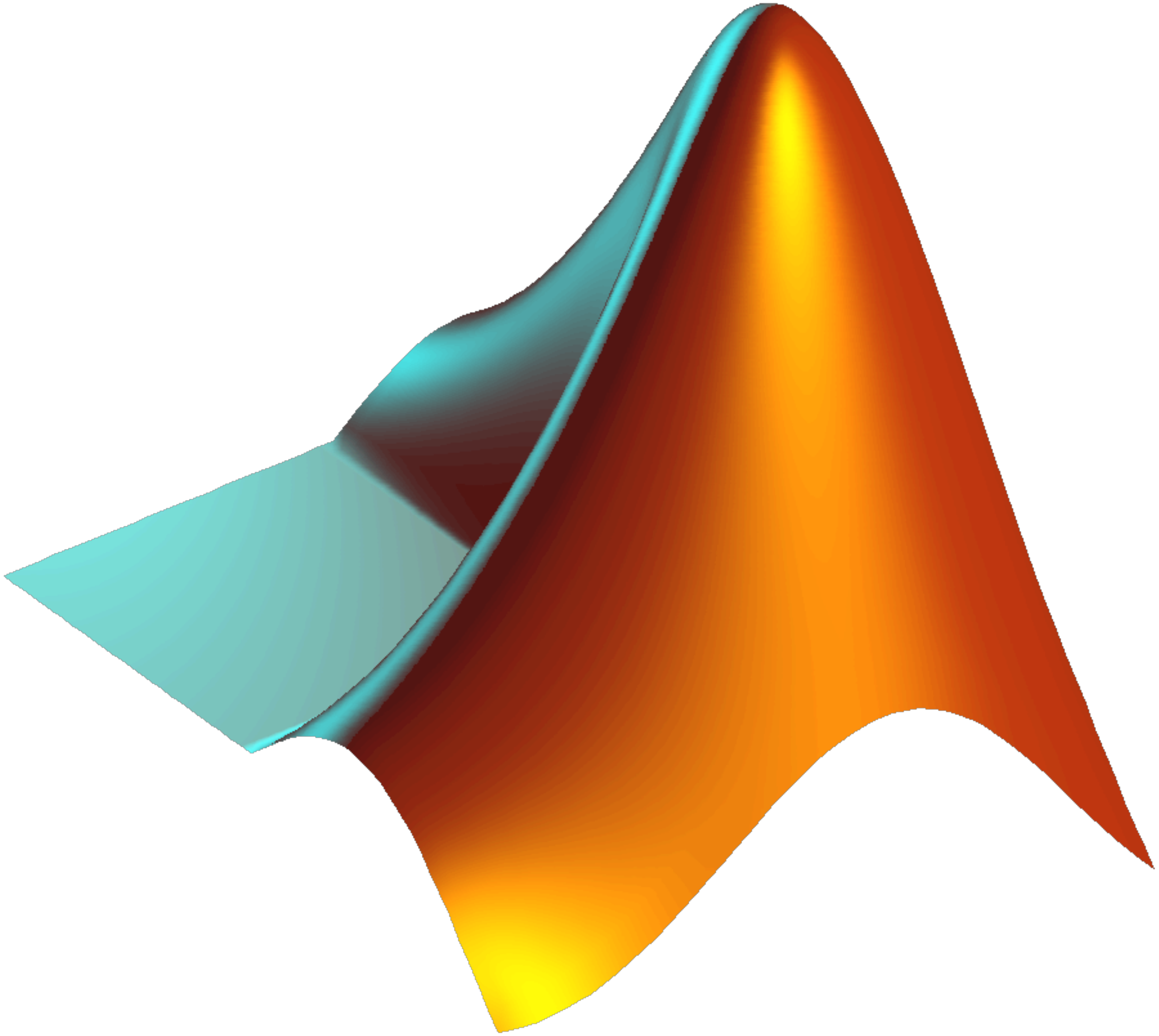
`h`: true if the null hypothesis is rejected at threshold thresh (default is 0.05)

`p`: p-value associated with the result

`ci`: 95% confidence intervals for the true value of the mean

`stats`: a structure that MATLAB can use for doing follow up tests, such as using `multcompare`.

Demo: t-test false rejections demo



Paired t-test

In a paired, t-test you make two measurements on the same subject, usually before and after. The null hypothesis is that the two measurements are drawn from distributions with equal means.

Internally, all this does is take the after-before difference for each subject and test against the null hypothesis that these differences have zero mean.

```
[h,p,ci,stats] = ttest(X, Y)
```

Two sample t-tests: `ttest2()`

You measure some quantity for two separate groups of subjects, and you want to know whether the means are different between the two groups.

Need to estimate whether the two groups of measurements have equal variances. Is one group more variable than the other?

```
[h,p,ci,stats] = ttest2(X, Y);
```

- Assumes equal variances

```
[h, p, ci, stats] = ttest2(X,Y,thresh,tail,‘unequal’);
```

- Assumes unequal variances

Two-tailed versus one-tailed

Two tailed t-test: tests the alternative hypothesis that the mean is different from zero, in either direction. You almost always want this one.

One tailed t-test: tests the alternative hypothesis that the mean is different from zero in a pariticular direction (i.e. greater than OR less than zero). This effectively **halves your p-value**, so people are often skeptical when you use this.

Tests for normality: Chi-square

```
h = chi2gof(x)
```

Performs a chi-square goodness-of-fit test of the default null hypothesis that the data in vector x are a random sample from a normal distribution with mean and variance estimated from x, against the alternative that the data are not normally distributed with the estimated mean and variance.

Source: mathworks.com

Lilliefors test for normality

```
h = lillietest(x)
```

Performs a Lilliefors test of the default null hypothesis that the sample in vector x comes from a distribution in the normal family, against the alternative that it does not come from a normal distribution.

Source: mathworks.com

Rank-sum tests

Known as Mann-Whitney U or Wilcoxon rank sum.

Assesses whether quantities in one group tend to be higher than the other.

Doesn’t require data to be normal, unlike the t-test.

```
p = ranksum(x,y)
```

Sign-rank test

Operates analogously to the t-test or paired t-test, assessing whether there a set of numbers has a median different from zero or whether there is a difference in medians between paired measurements.

Doesn't require data to be normal.

```
p = signrank(x)
p = signrank(x,y)
```

Chi-square variance test

```
[h, p] = vartest(X,v)
```

Performs a chi-square test of the null hypothesis that the samples in vector x comes from a normal distribution with variance v, against the alternative that X comes from a normal distribution with a different variance.

Data must be normal.

Source: mathworks.com

Compare variances of two groups

```
[h, p, ci] = vartest2(X,Y)
```

Performs an F test of the hypothesis that two independent samples, in the vectors X and Y, come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances.

`ci` is a 95% confidence interval for the true variance ratio `var(X) / var(Y)`

Data must be normal

Source: mathworks.com

Outline

Summary statistics functions

Random Variables

- Random variables, PDF, CDFs
- Estimates of central tendency and dispersion
- Standard error of the mean, confidence intervals

Statistical Hypothesis Testing

- Tests and significance
- Student’s t test walkthrough
- Other commonly used tests

Analysis of Variance

Regression

- Linear regression
- Curve fitting

Dimensionality Reduction

- PCA

Assignment 6 Overview

Analysis of Variance

Analysis of variance (ANOVA) is a set of statistical models and methods for partitioning variance in some quantity into components attributable to different sources.

ANOVA extends the t-test to multiple groups and allows you to test **against the null hypothesis that** some quantity measured from **all of these groups have the same mean**.

Analysis of Variance

```
[p table stats] = anova1(X, groupNames)
```

Performs a one-way balanced ANOVA comparing the means of the columns of X.

Each column is a group, each row in that column is a data point. Thus the number of subjects in each group must be equal (i.e. balanced design).

p-value is the significance threshold associating with rejecting the **null hypothesis that all means are the same.**

Analysis of Variance

The ANOVA test makes the following assumptions about the data:

All sample populations are normally distributed.

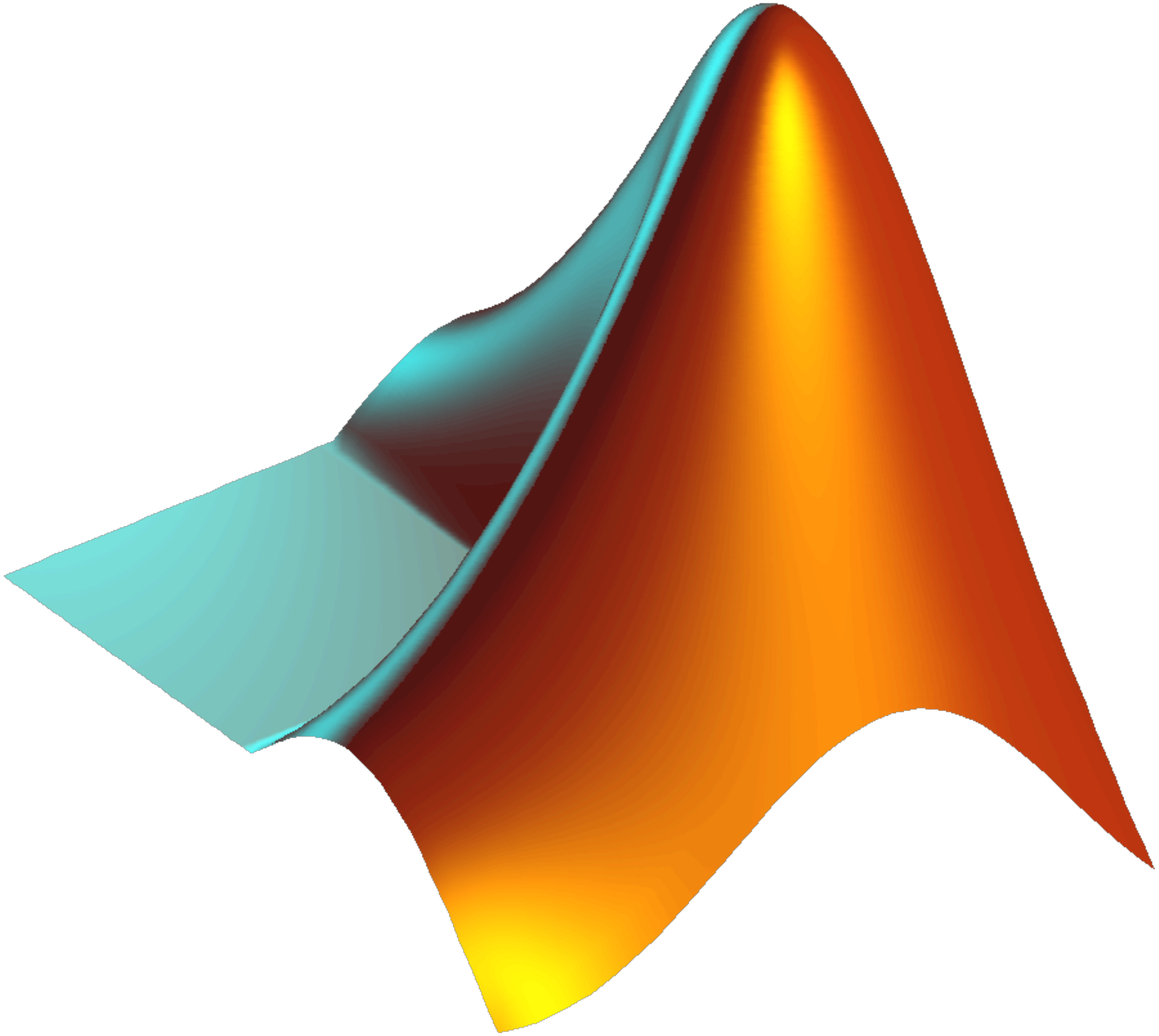
All sample populations have equal variance.

All observations are mutually independent.

The ANOVA test is known to be robust with respect to modest violations of the first two assumptions.

Source: mathworks.com

Demo: One-way anova

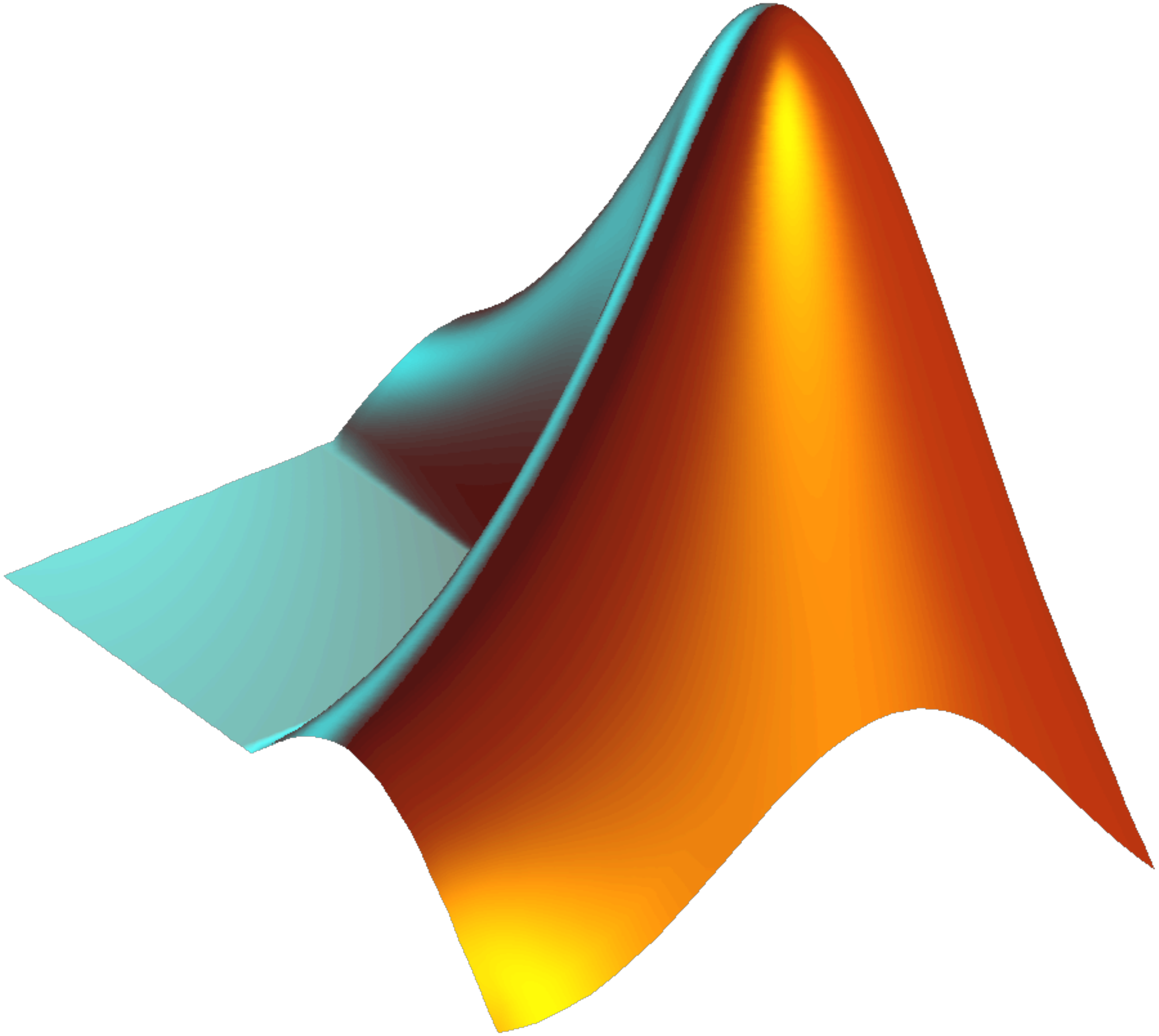


Multiple comparisons

```
[c,m] = multcompare(stats);
```

When you have many groups, allows testing for differences between pairs of groups, without the rate of false positives increasing with each comparison.

Demo: Multiple comparisons



N-way analysis of variance

In an N-way analysis of variance, you have a single output measurement for each subject, and each subject is described by N factors.

The aim is to determine which of these N factors (or interactions among these factors) affect the output quantity.

Works fine with **unbalanced designs** as well, meaning some groups can have more data points than others.

N-way analysis of variance

Each subject/trial/data point etc. is described by a single output measurement Y. It also belongs to one of several groups within each factor.

N-way analysis of variance

N-way analysis of variance

Example: Each data point represents one mouse.

- **Output quantity:** score on some behavioral assay
- **Factor 1:** Age group (Young, Old)
- **Factor 2:** Drug treatment (Control, DrugA, DrugB)

N-way analysis of variance

Example: Each data point represents one mouse.

- **Output quantity:** score on some behavioral assay
- **Factor 1:** Age group (Young, Old)
- **Factor 2:** Drug treatment (Control, DrugA, DrugB)

Main effects:

- Does the age group affect the score?
- Does the drug treatment group affect the score?

N-way analysis of variance

Example: Each data point represents one mouse.

- **Output quantity:** score on some behavioral assay
- **Factor 1:** Age group (Young, Old)
- **Factor 2:** Drug treatment (Control, DrugA, DrugB)

Main effects:

- Does the age group affect the score?
- Does the drug treatment group affect the score?

Interaction effects:

- Does the age group affect the score differently depending on what drug group a mouse is in? (Equivalently, vice versa?)

N-way analysis of variance

```
assayScore = [24 101 56  ... ]

ageGroup = {‘young’, ‘old’, ‘young’, ... }

treatmentGroup = {‘control’, ‘drugA’, ‘drugA’ ...}


[p t stats terms] = anovan(assayScore, ...
   {ageGroup treatmentGroup}, ...

   'varnames', {'Age Group','Treatment Group'}, ...

   'model', 'interaction');
```

N-way analysis of variance

```
assayScore = [24 101 56  ... ]

ageGroup = {‘young’, ‘old’, ‘young’, ... }

treatmentGroup = {‘control’, ‘drugA’, ‘drugA’ ...}


[p t stats terms] = anovan(assayScore, ...
   {ageGroup treatmentGroup}, ...

   'varnames', {'Age Group','Treatment Group'}, ...

   'model', 'interaction');
```

N-way analysis of variance

```
assayScore = [24 101 56  ... ]
ageGroup = {‘young’, ‘old’, ‘young’, ... }
treatmentGroup = {‘control’, ‘drugA’, ‘drugA’ ...}

[p t stats terms] = anovan(assayScore, ...
   {ageGroup treatmentGroup}, ...
   'varnames', {'Age Group','Treatment Group'}, ...
   'model', 'interaction');
```

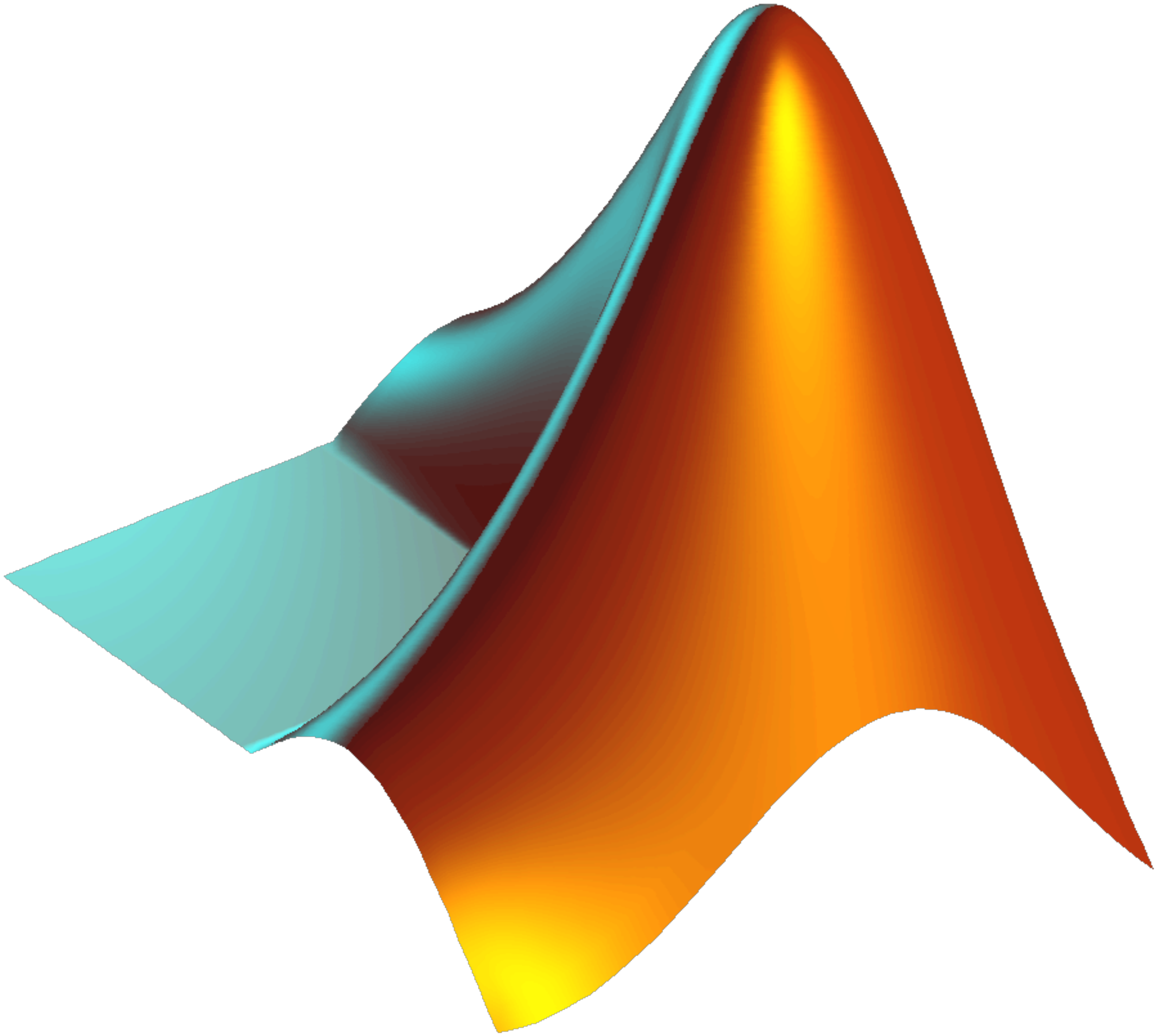
N-way analysis of variance

```
assayScore = [24 101 56  ... ]

ageGroup = {‘young’, ‘old’, ‘young’, ... }

treatmentGroup = {‘control’, ‘drugA’, ‘drugA’ ...}


[p t stats terms] = anovan(assayScore, ...
   {ageGroup treatmentGroup}, ...

   'varnames', {'Age Group','Treatment Group'}, ...

   'model', 'interaction');
```

N-way analysis of variance

```
assayScore = [24 101 56  ... ]

ageGroup = {‘young’, ‘old’, ‘young’, ... }

treatmentGroup = {‘control’, ‘drugA’, ‘drugA’ ...}


[p t stats terms] = anovan(assayScore, ...
   {ageGroup treatmentGroup}, ...

   'varnames', {'Age Group','Treatment Group'}, ...

   'model', 'interaction');
```

Demo: **anovan** example



Outline

Summary statistics functions

Random Variables

- Random variables, PDF, CDFs
- Estimates of central tendency and dispersion
- Standard error of the mean, confidence intervals

Statistical Hypothesis Testing

- Tests and significance
- Student's t test walkthrough
- Other commonly used tests

Analysis of Variance

Regression

- Linear regression
- Curve fitting

Dimensionality Reduction

- PCA

Assignment 6 Overview

Linear regression

```
  [p s] = polyfit(x, y, degree);
```

Use degree == 1 to fit a line, degree == 2 for a quadratic function. Can use polyval or polyconf to generate confidence intervals for where new points might lie.

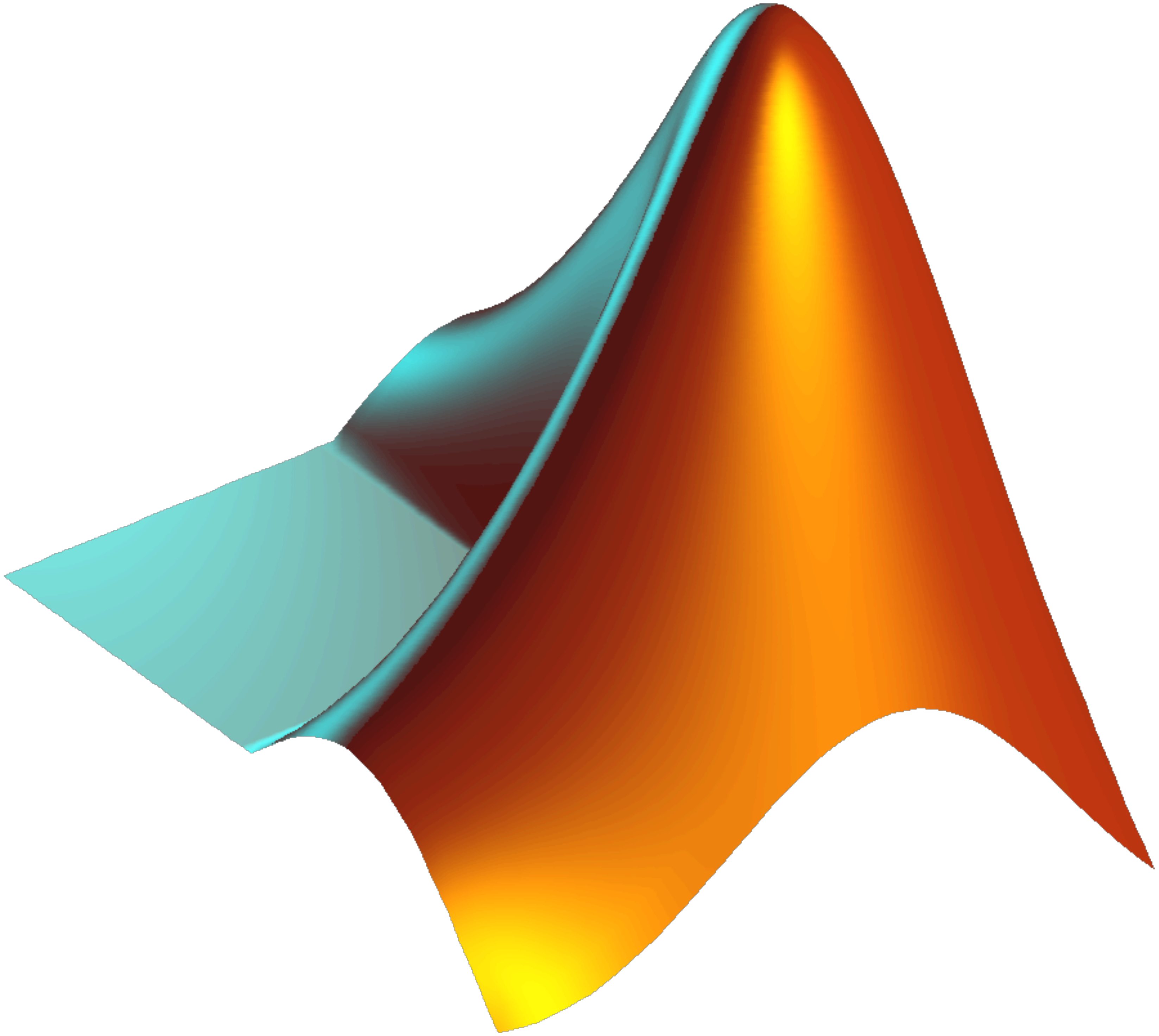
```
[coeffs ci residuals residualIntervals stats] =
    regress(y, XWithOnes);
```

Can handle several predictors as columns in X. Make sure that X has a column of all ones in order to include a constant offset term. Tries to fit the model:

```
y(i) = coeffs(1)*x(i,1) + coeffs(2)*x(i,2) + ...
```

Make the last column all ones

Demo: Linear regression



Curve fitting

Fitting a curve in MATLAB uses two basic tools:

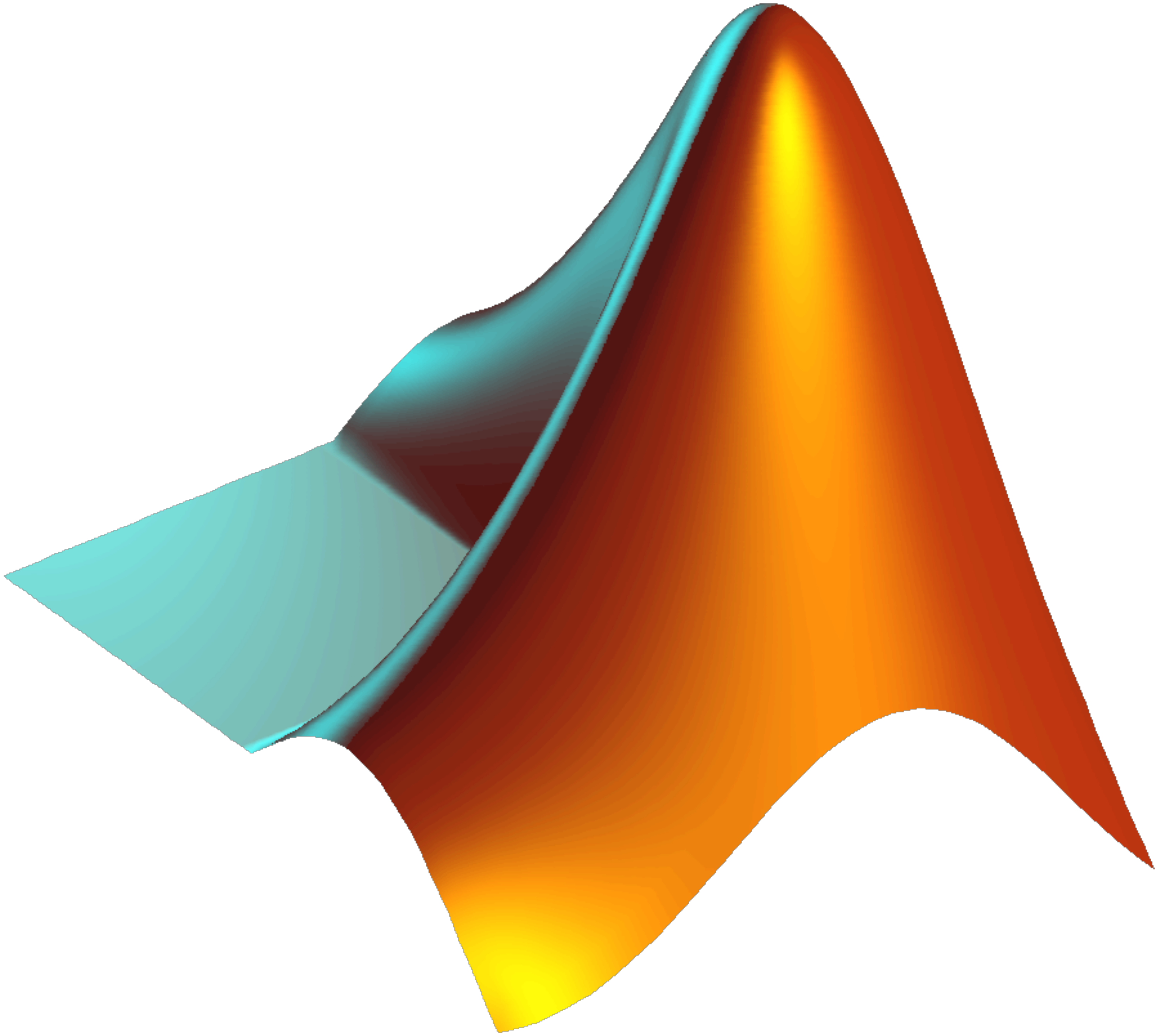
`fittype()`

- Allows you to describe the form of the function you're trying to fit to your data
- Specify what the independent variables are
- Specify what the parameters to fit are

`fit()`

- Actually does the iterative fitting
- Specify upper and lower bounds for each parameter
- Specify a starting point for each parameter
- Also returns confidence intervals for each parameter

Demo: exponential curve fitting



Principal components analysis

PCA is a dimensionality reduction technique.

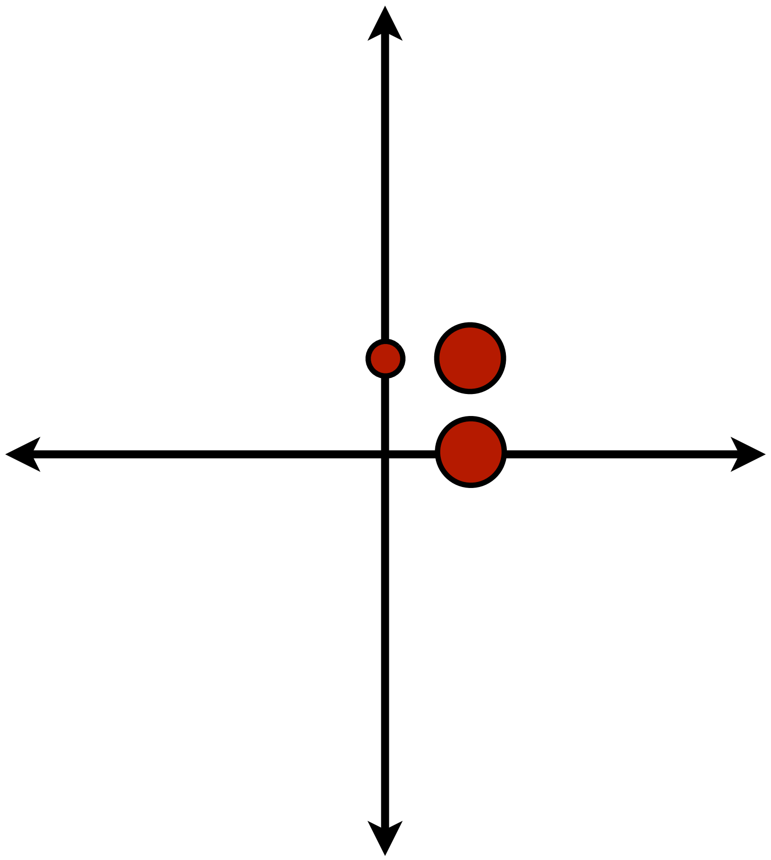
What's a dimension?

- Measurement dimensions are any quantity which is measured, e.g. height, weight, firing rate of a neuron, x coordinate of the hand, etc.
- We can also make up new dimensions by linearly combining these quantities, e.g.
  - x-coordinate + y-coordinate
  - height - 2*weight
  - firing rate 1 + firing rate 2 - 5*firing rate 10

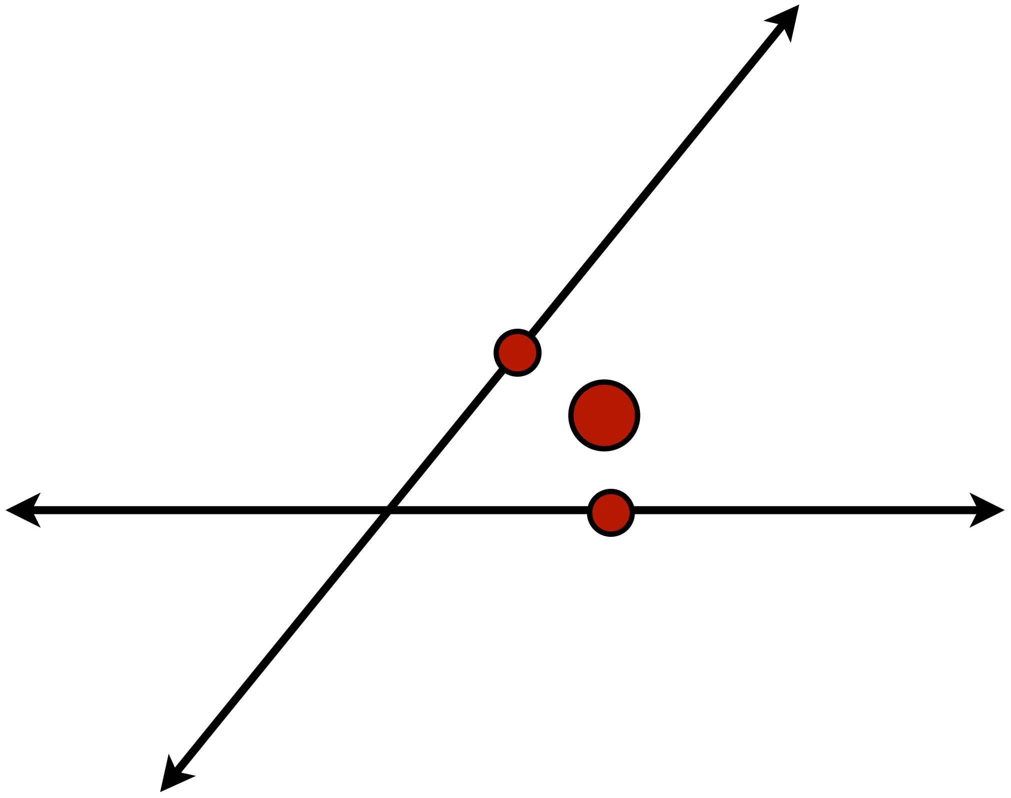
Principal components analysis

What's a space?

- Two dimensions are **orthogonal** if it you could move along one without moving along the other.



Orthogonal

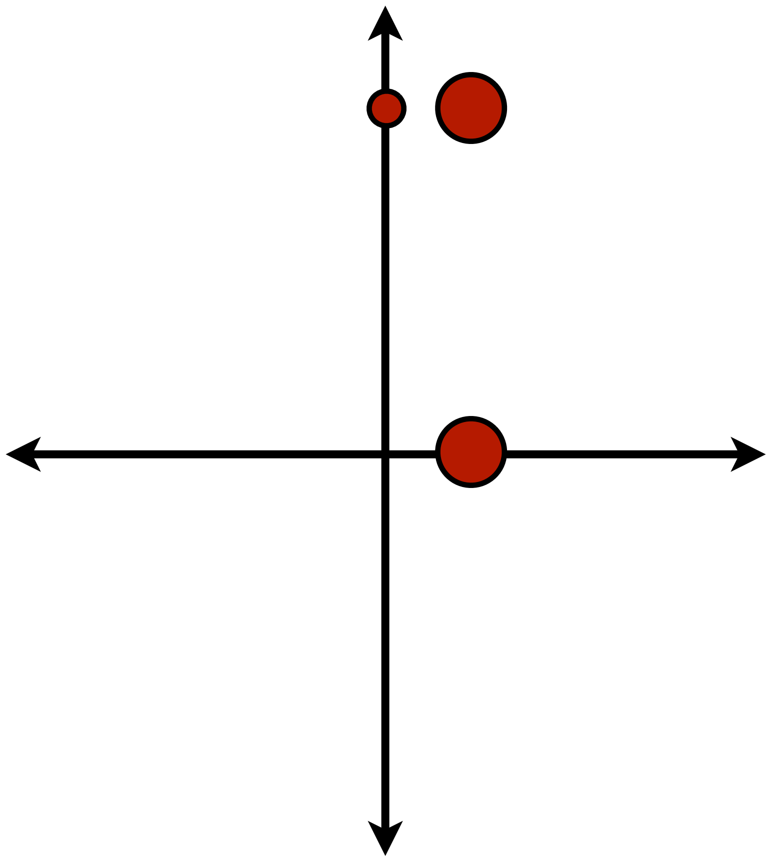


Non-orthogonal

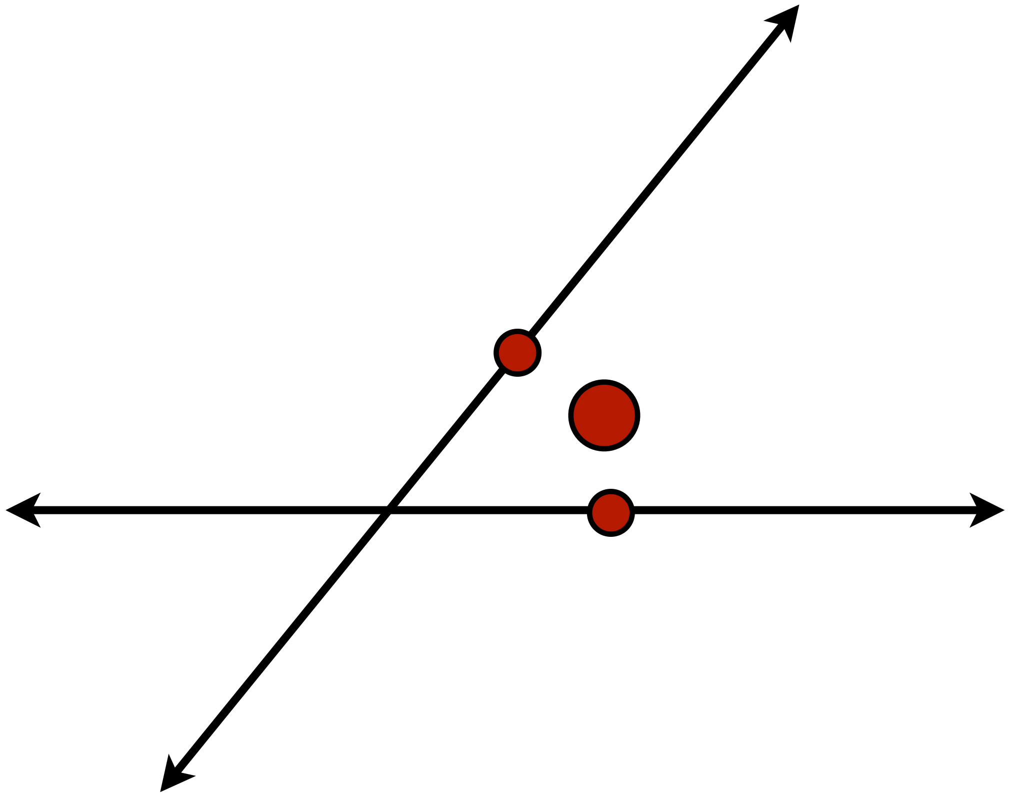
Principal components analysis

What's a space?

- Two dimensions are **orthogonal** if it you could move along one without moving along the other.



Orthogonal

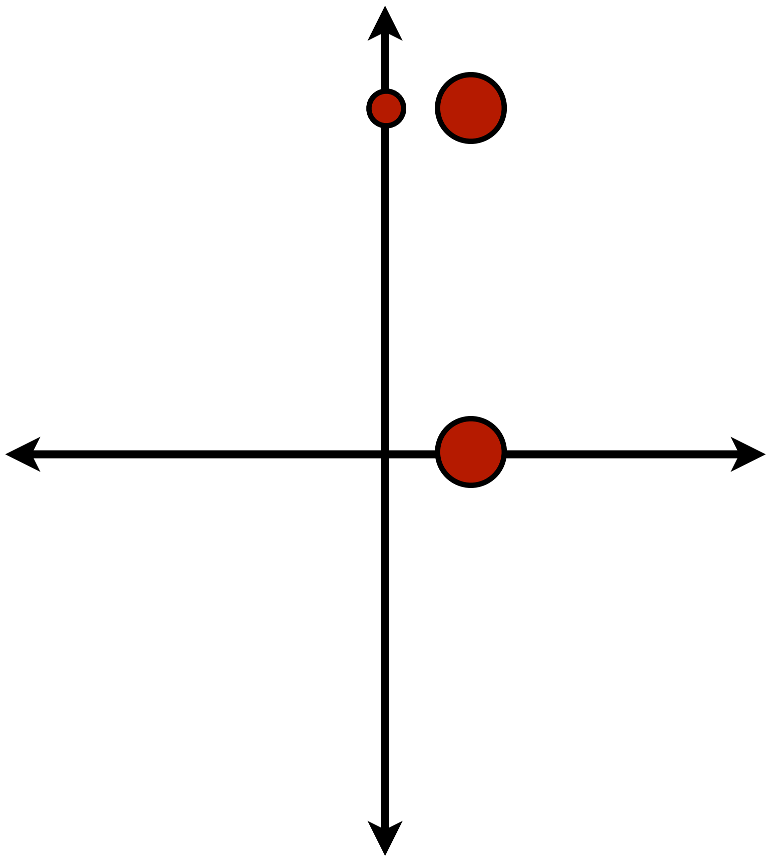


Non-orthogonal

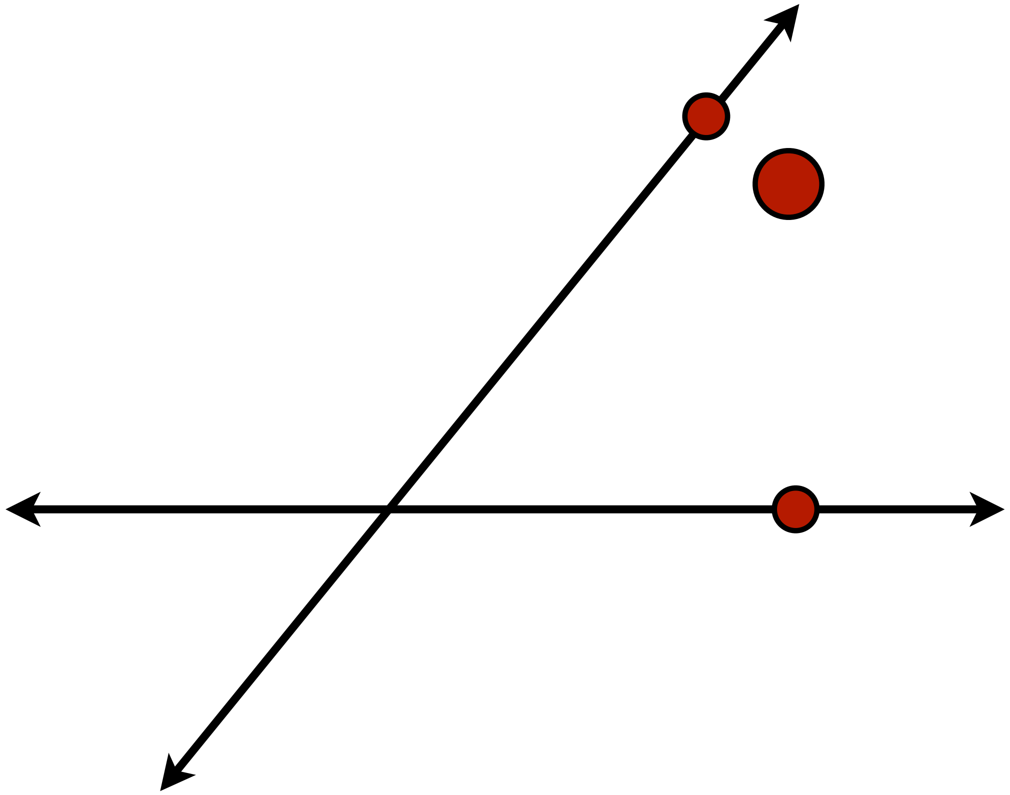
Principal components analysis

What's a space?

- Two dimensions are **orthogonal** if it you could move along one without moving along the other.



Orthogonal



Non-orthogonal

Principal components analysis

What's a space?

- Two dimensions are **orthogonal** if it you could move along one without moving along the other.
- The original measurement directions are orthogonal, even though quantities along these dimensions may be correlated.
- A space is simply a set of orthogonal dimensions.

Principal components analysis

Why reduce the number of dimensions?

- **Visualization:** We can only plot 3 dimensions, and really we're only looking at 2
- **Intuition:** Low dimensional visualizations are easier to understand than high dimensional data
- **Noise:** often, when we make lots of measurements, each dimension captures a little bit of meaningful information and a bit of noise.

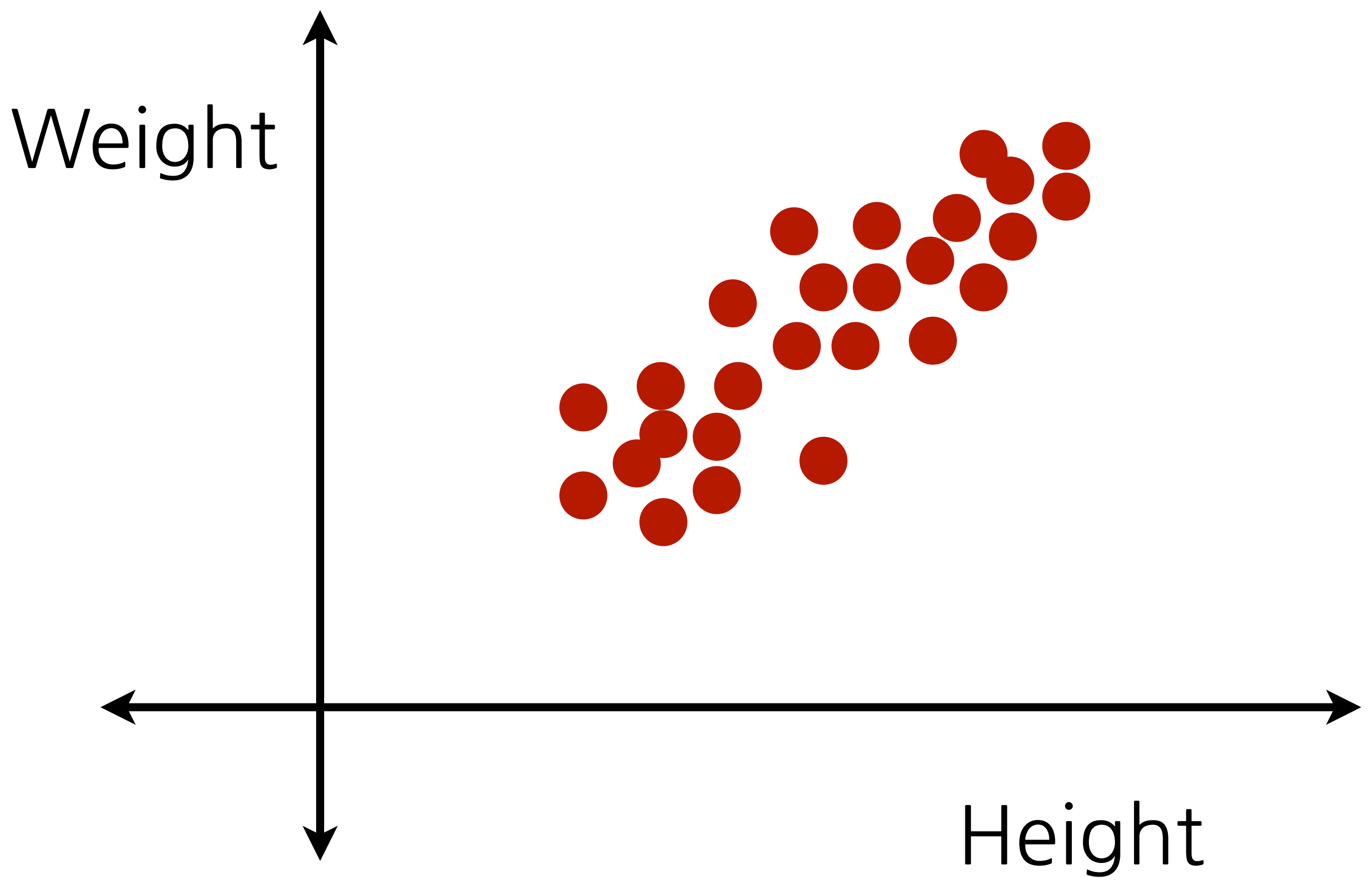
Principal components analysis

How can we find a reduced set of dimensions?

- We need a guiding principle for figuring out which dimensions matter and which don't.
- PCA operates on the assumption that dimensions in the data where the data points spread out most (have the highest variance) are the most important dimensions.
- This amounts to saying the "signal" we're interested in has higher variance than the noise.

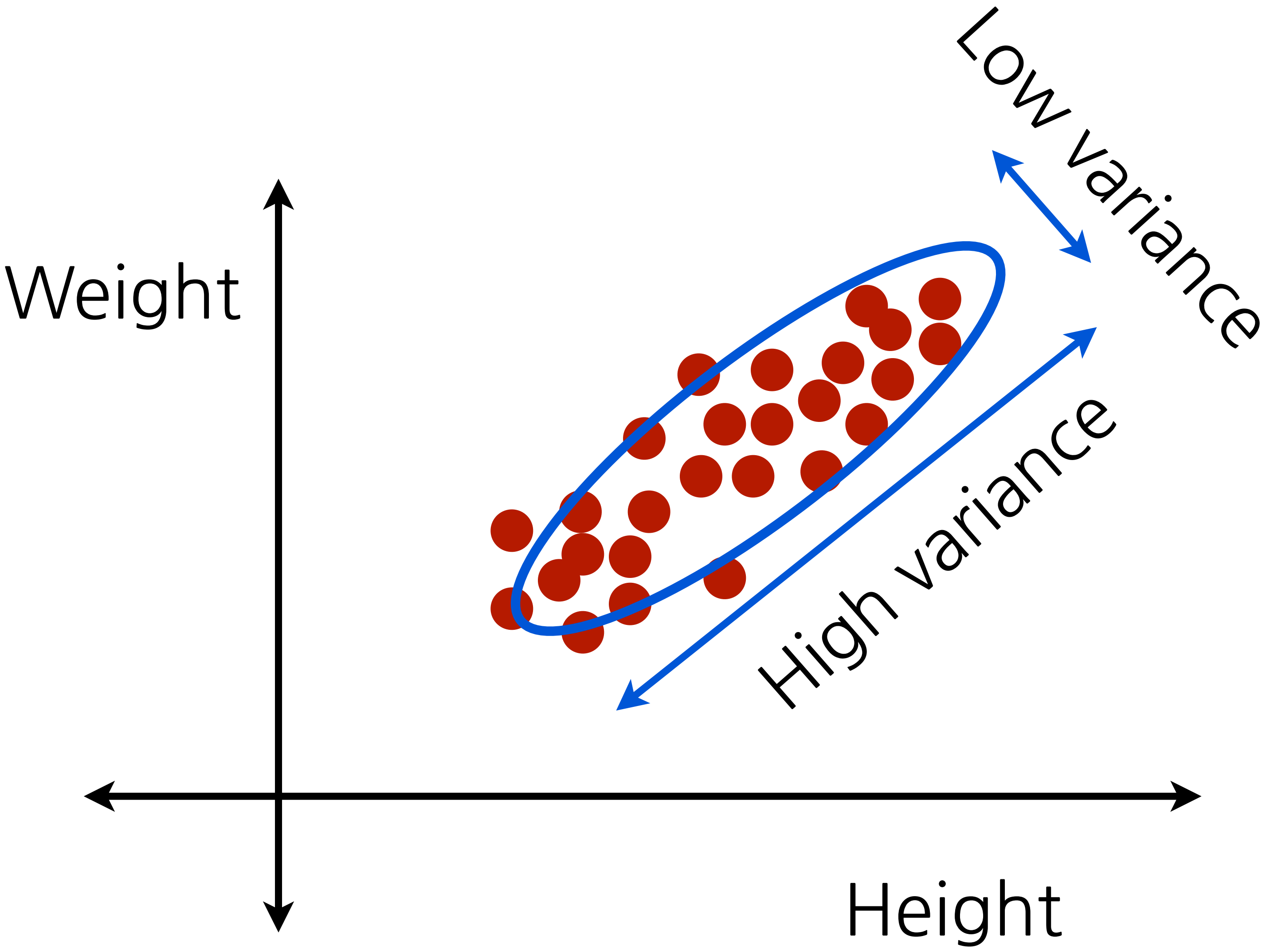
Principal components analysis

A graphical example of what this means:



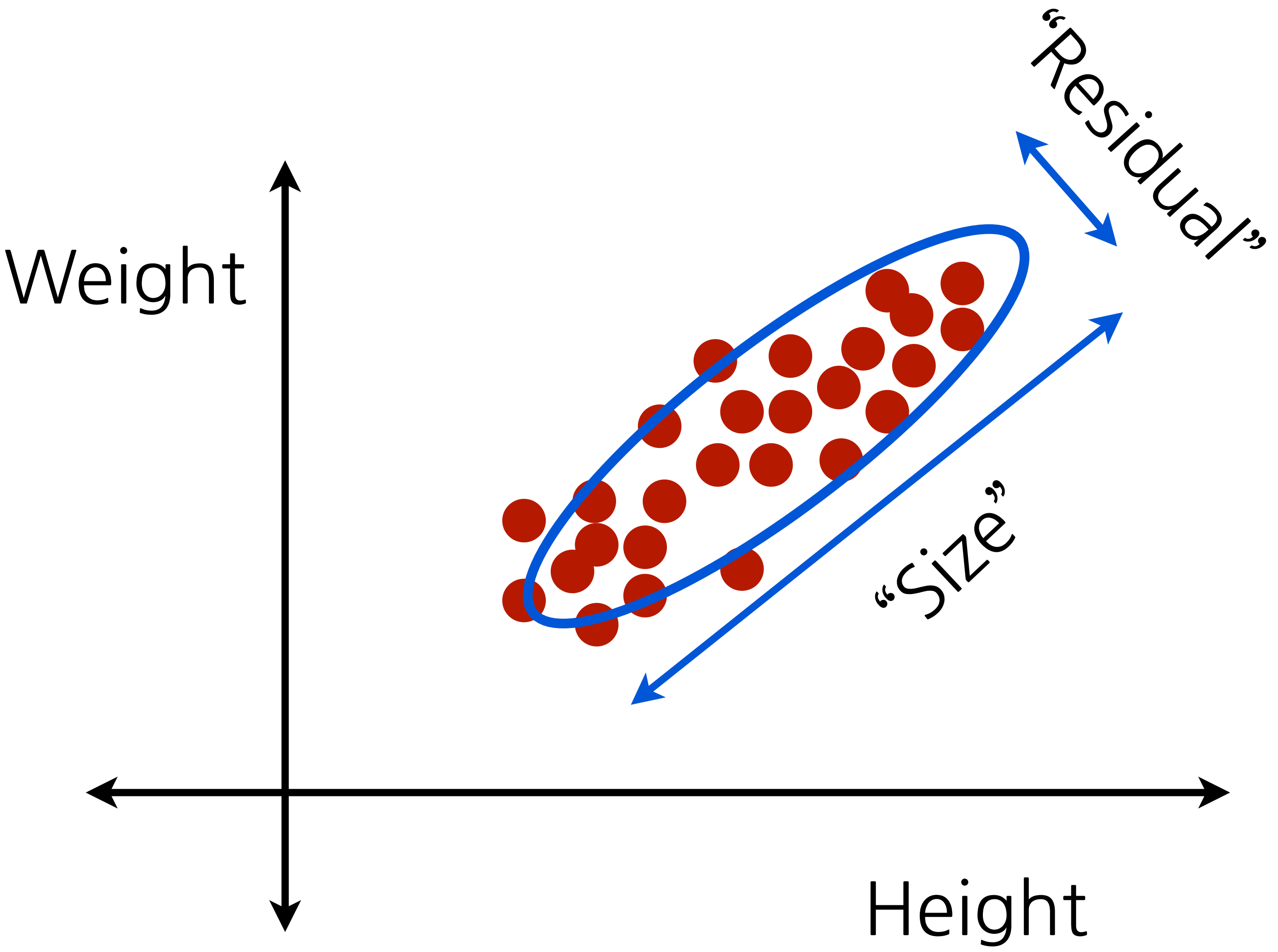
Principal components analysis

A graphical example of what this means:



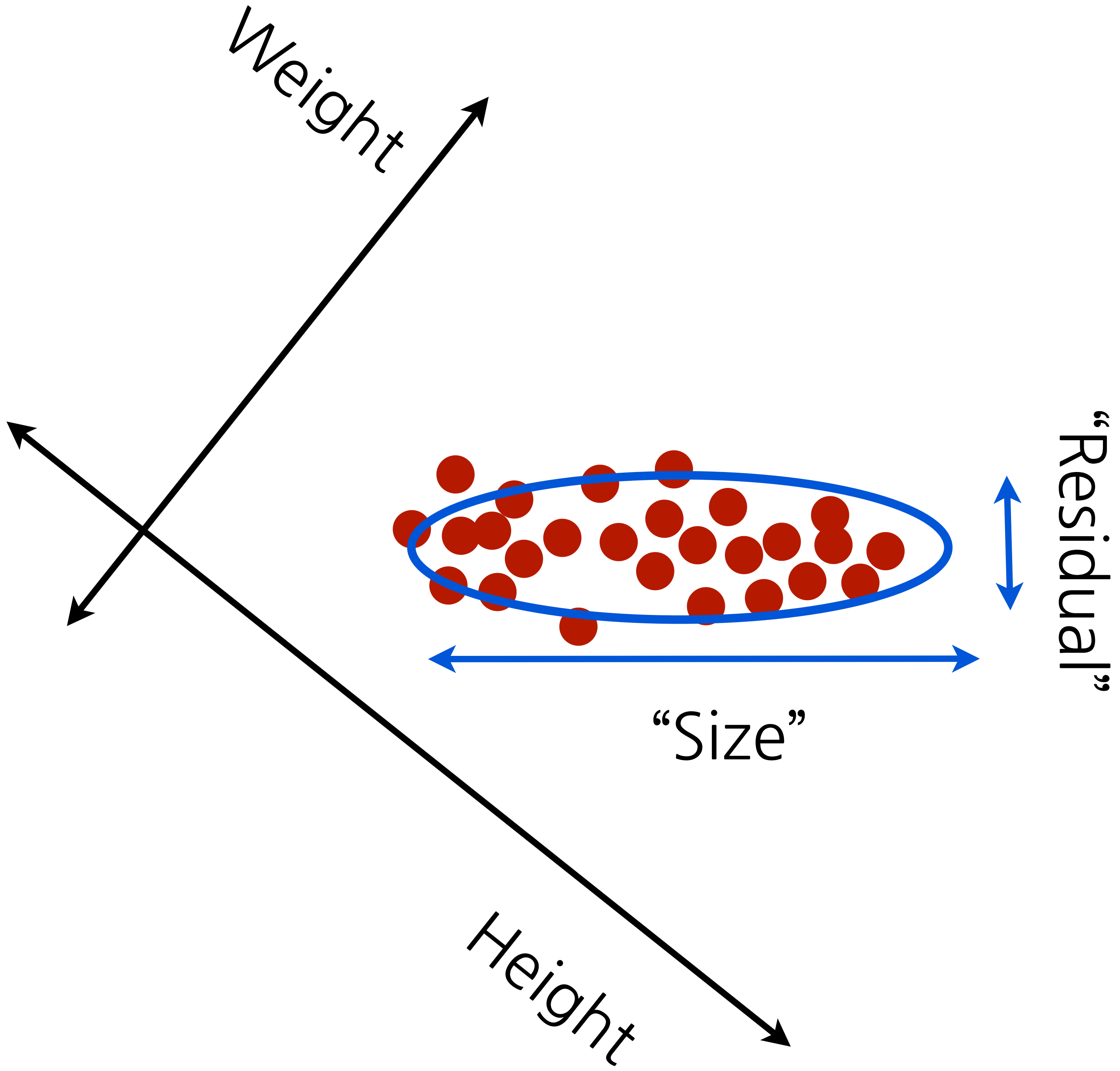
Principal components analysis

A graphical example of what this means:



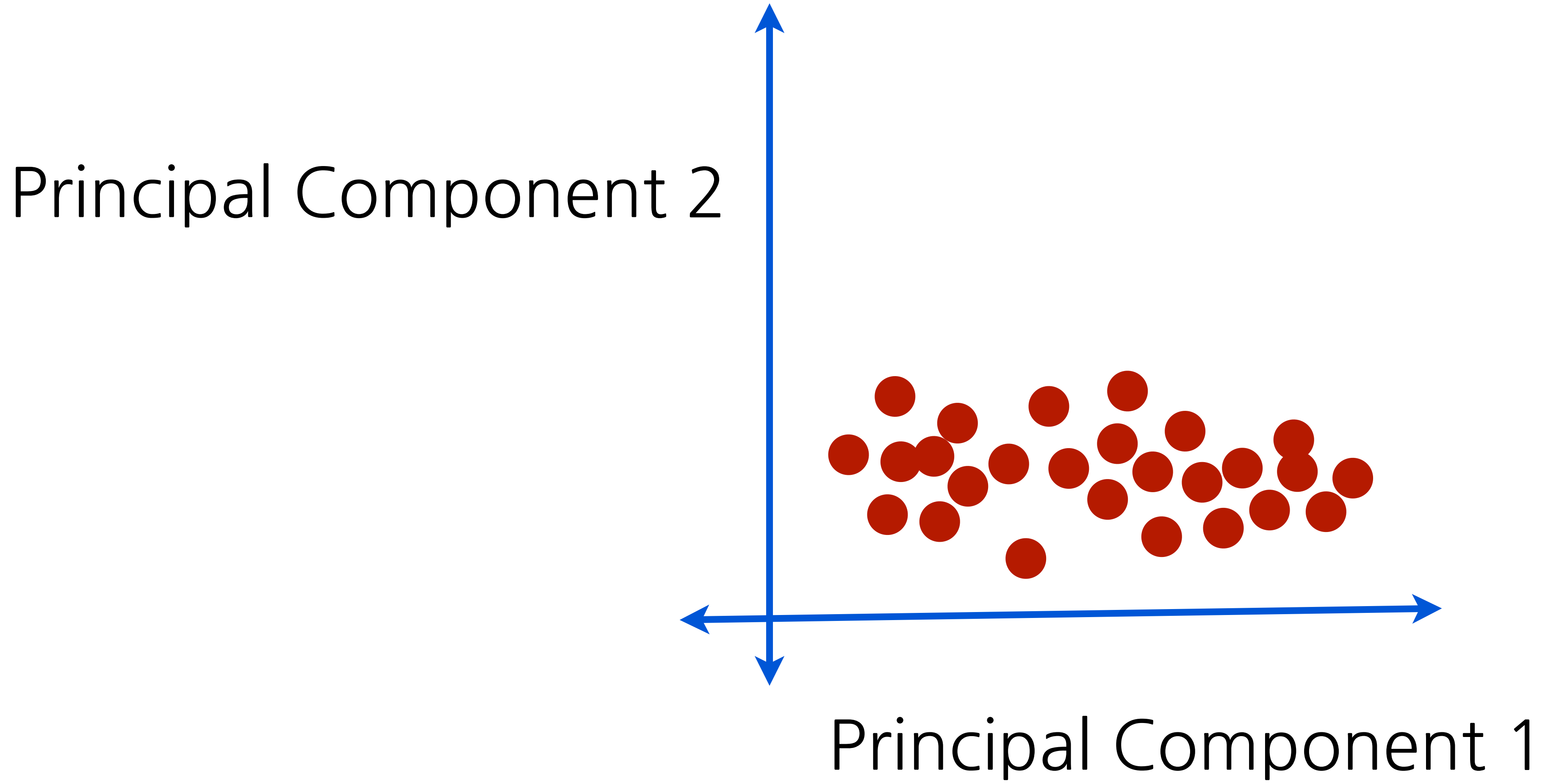
Principal components analysis

A graphical example of what this means:

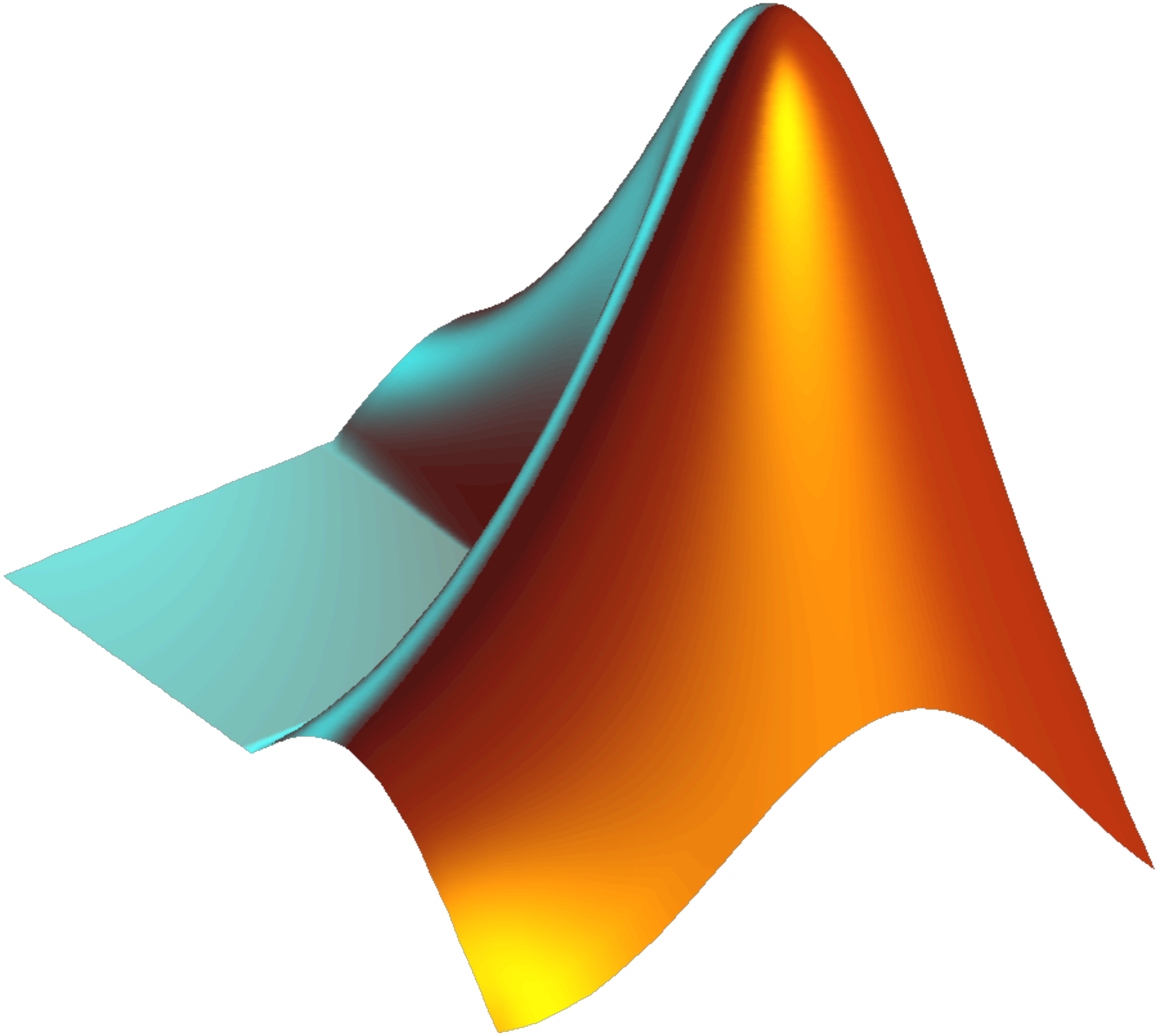


Principal components analysis

A graphical example of what this means:



Demo: PCA



Outline

Summary statistics functions

Random Variables

- Random variables, PDF, CDFs
- Estimates of central tendency and dispersion
- Standard error of the mean, confidence intervals

Statistical Hypothesis Testing

- Tests and significance
- Student’s t test walkthrough
- Other commonly used tests

Analysis of Variance

Regression

- Linear regression
- Curve fitting

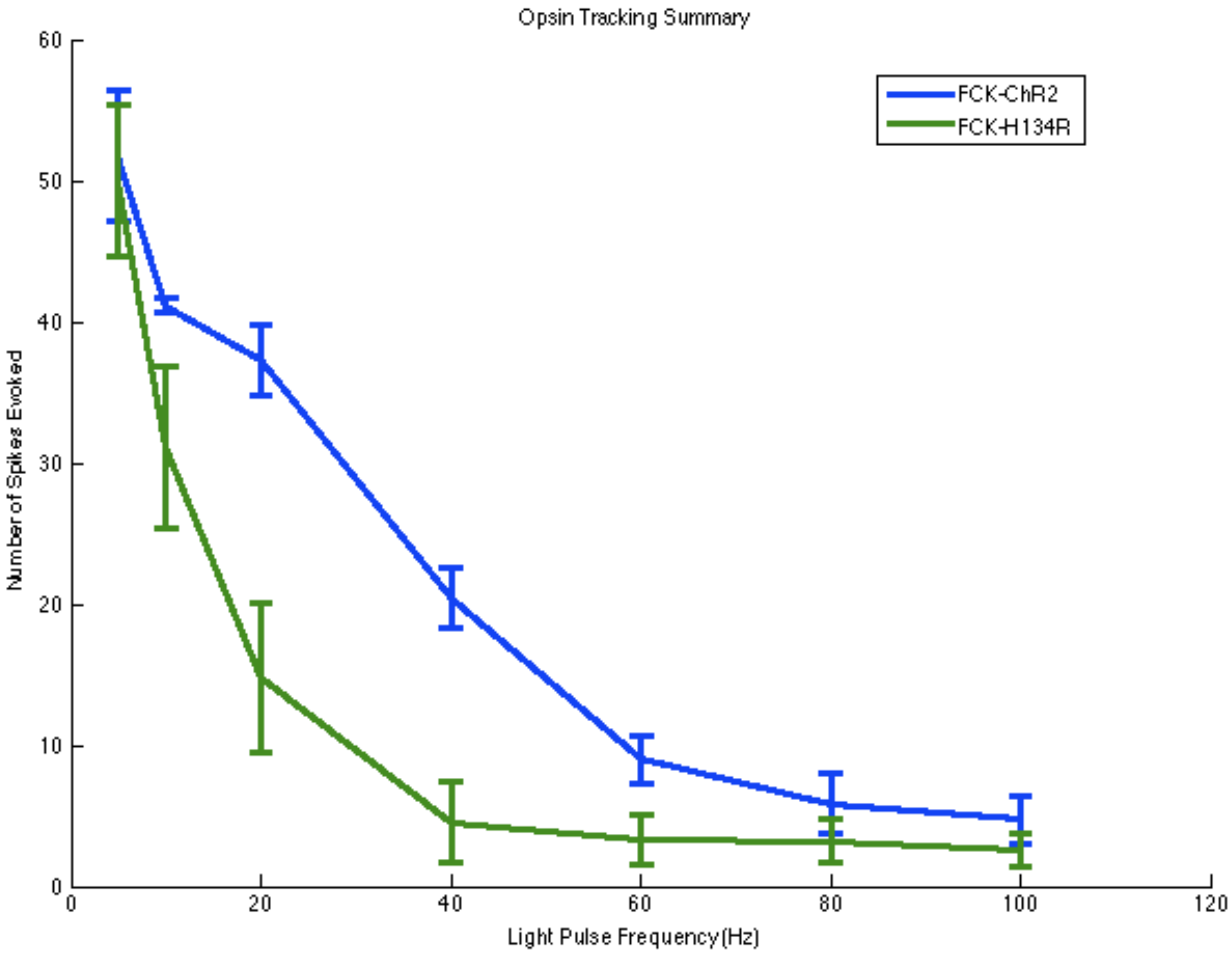
Dimensionality Reduction

- PCA

Assignment 6 Overview

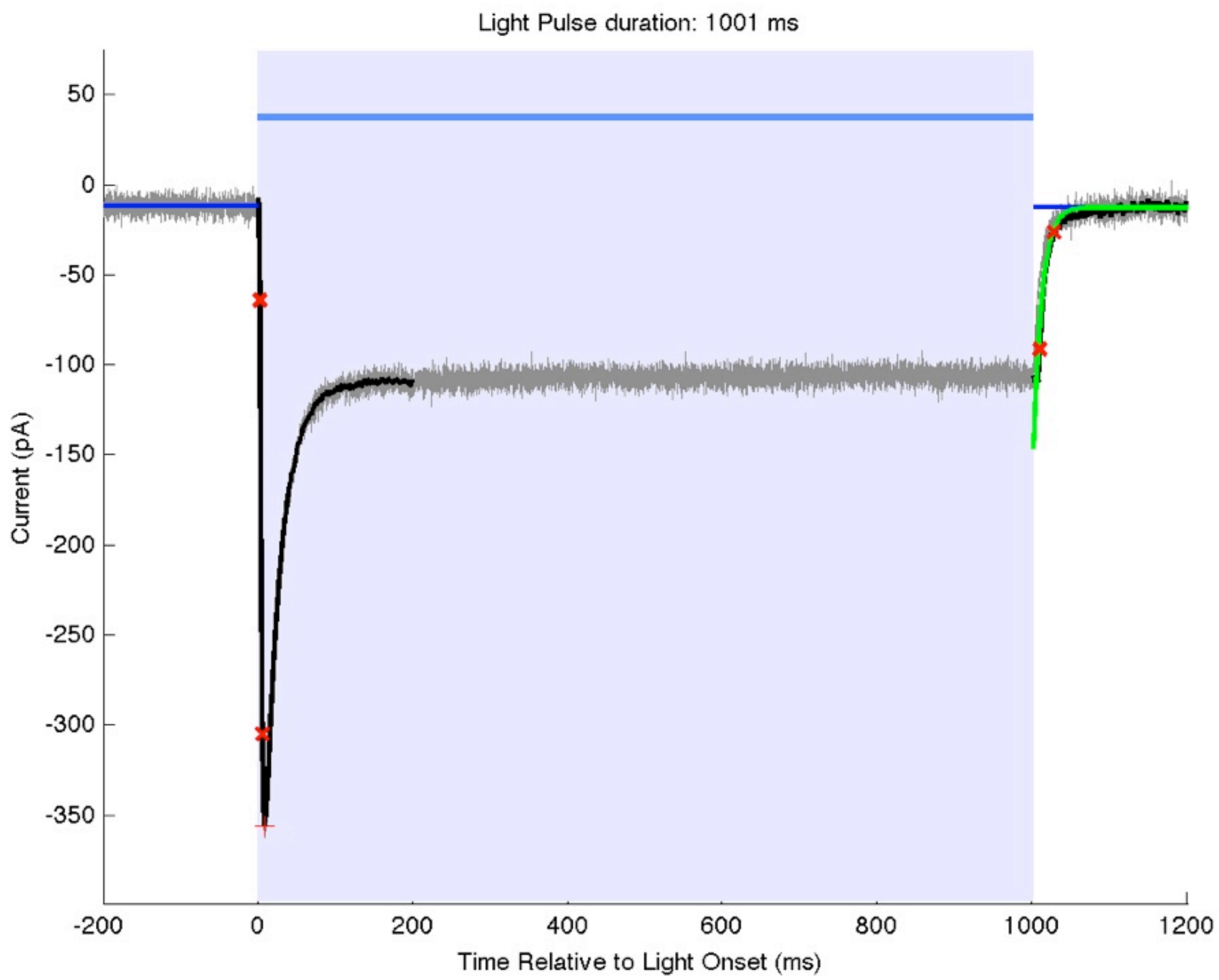
Significance Markers

Add asterisks above comparisons that are significant.



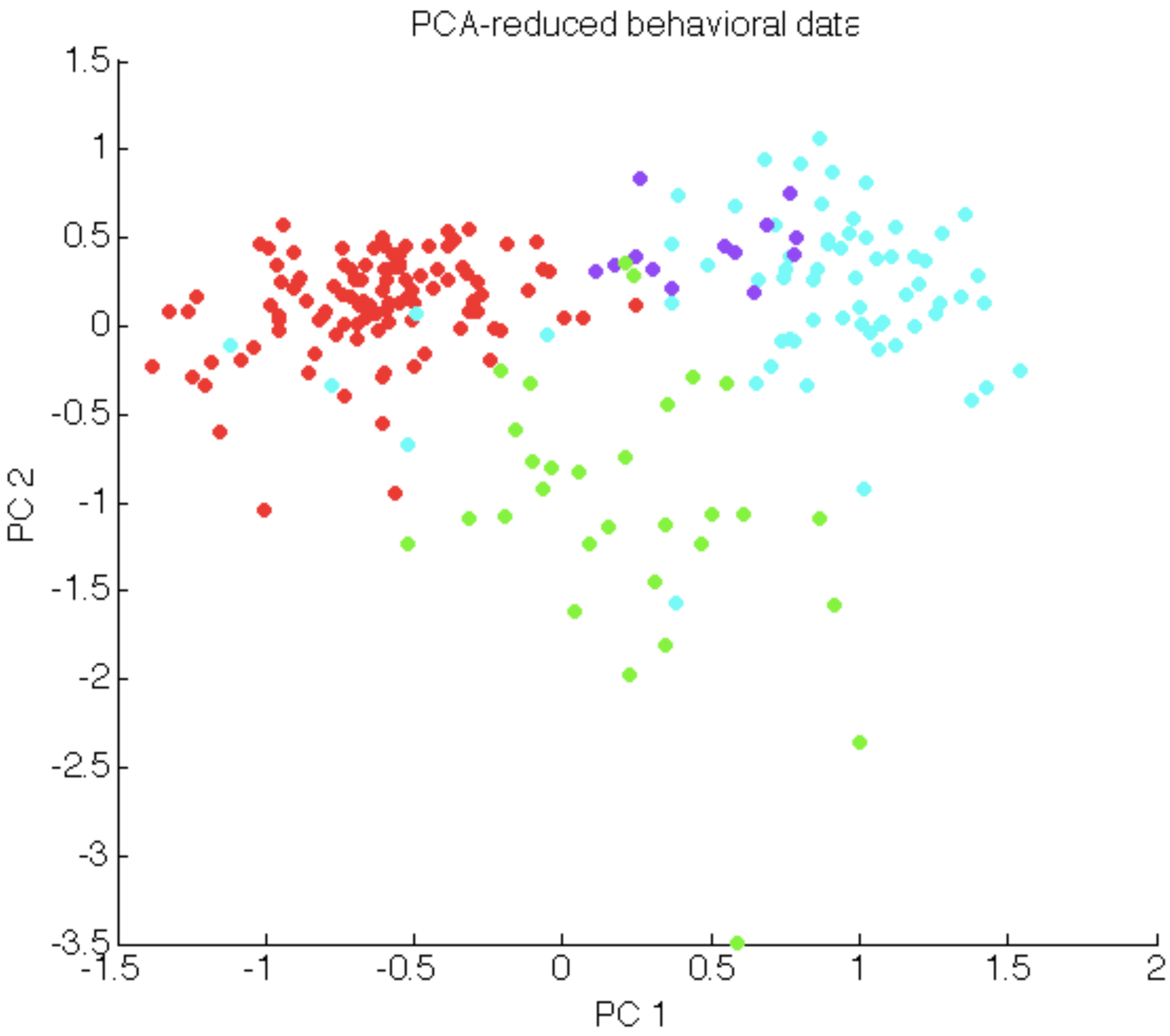
Curve fitting

Fit an exponential function to find the inactivation kinetics of of ChR2.



PCA on fly behavioral data

Exploratory data visualization using PCA.


PCA-reduced behavioral data
PC 2
PC 1