Lecture 6:

Covariance and Correlation

Road Map of the Statistics Part

Lecture 5

Quantification Technique	Mean, variance, skewness, & kurtosis
Uncertainty & Significance	Gaussian distribution Chi-2 distribution
Assumptions	Data is Gaussian or follows specific types of distribution Independent Sampling
Test assumptions	K-S test
Treatment	

Data is Gaussian → K-S test

What will be covered in this lecture?

1. Covariance
2. Correlation

    Cautionary notes on correlation

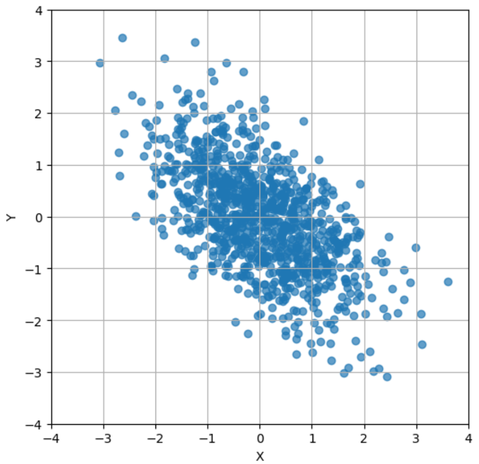
3. How to tell if a correlation is significant?

    3.1 Bootstrapping

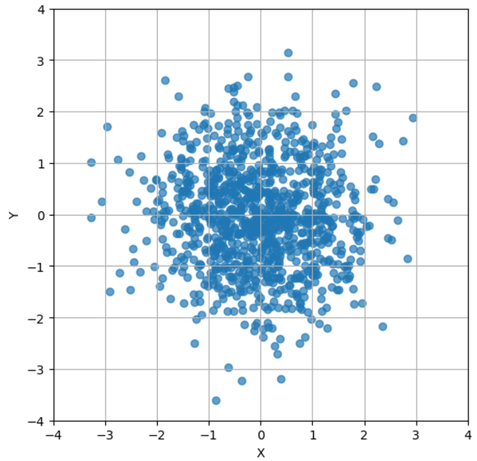
Covariance: helps to under stand if two sets of data moves together

Statistics	Meaning	How to calculate it from data	Functions to use
Covariance	Measures the linear relationship between two variables. A positive value means they tend to move in the same direction; a negative value indicates opposite directions; near zero suggests little to no linear relationship.	$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$	numpy.cov(data1, data2) **or** dataframe['x'].cov(dataframe['y'])
Variance	Indicates the spread of a single variable around its mean.	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$	numpy.var(data) **or** dataframe['column'].var()

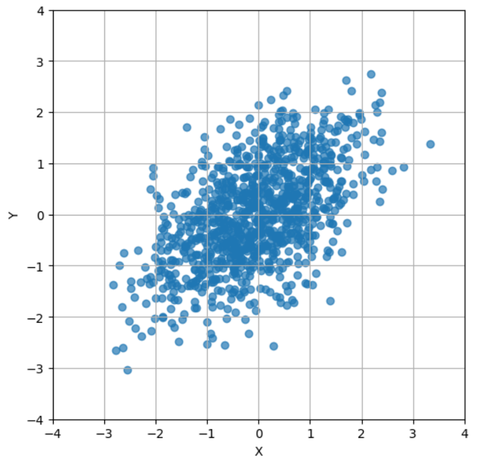
Cov < 0



Cov = 0



Cov > 0



Covariance is scale-variant, meaning its value is sensitive to the units we use.

Correlation: a scale-invariant quantification

Statistics	Meaning	How to calculate it from data	Function to use
Correlation	Measures linear association without being affected by the variables' variance.	$\rho = \mathrm{Cov}(X_{std}, Y_{std})$	scipy.stats.pearsonr(data1, data2) **or** dataframe[select_columns].corr()
Covariance	Measures linear association but is sensitive to the variables' units.	$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$	numpy.cov(data1, data2) **or** dataframe['x'].cov(dataframe['y'])

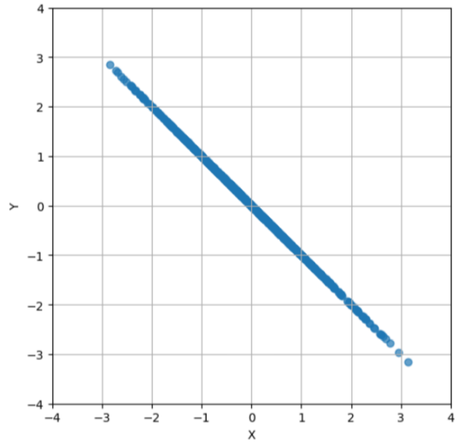
$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \mathrm{Cov}(X_{std}, Y_{std})$$

Pearson's Correlation ranges between **-1** and **1**.

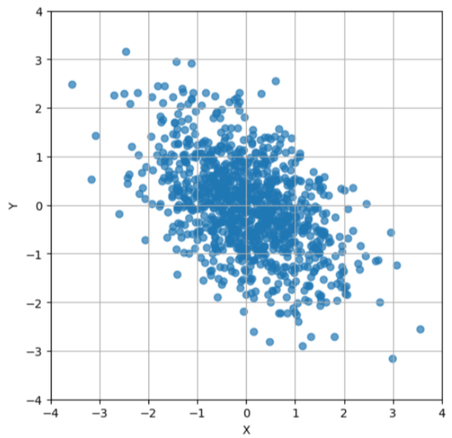
It is independent to the unit we use, which is also called **scale-invariant**.

Correlation controls the shape of the data cloud

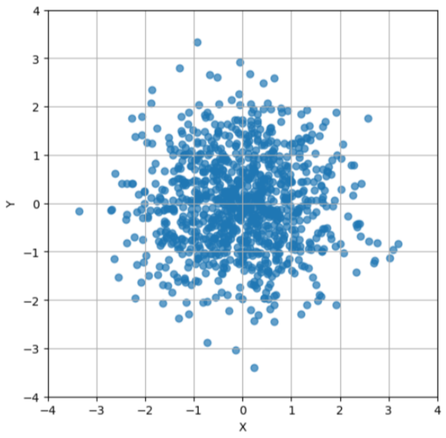
ρ = -1



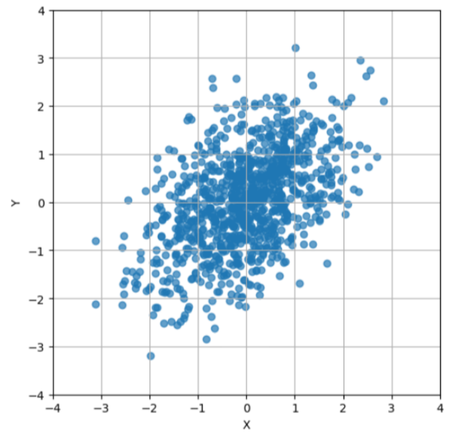
ρ = -0.5



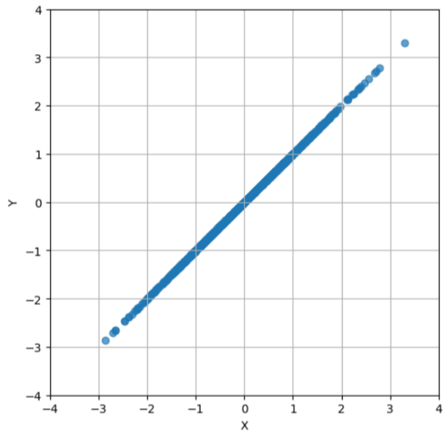
ρ = 0



ρ = 0.5



ρ = 1



Correlation

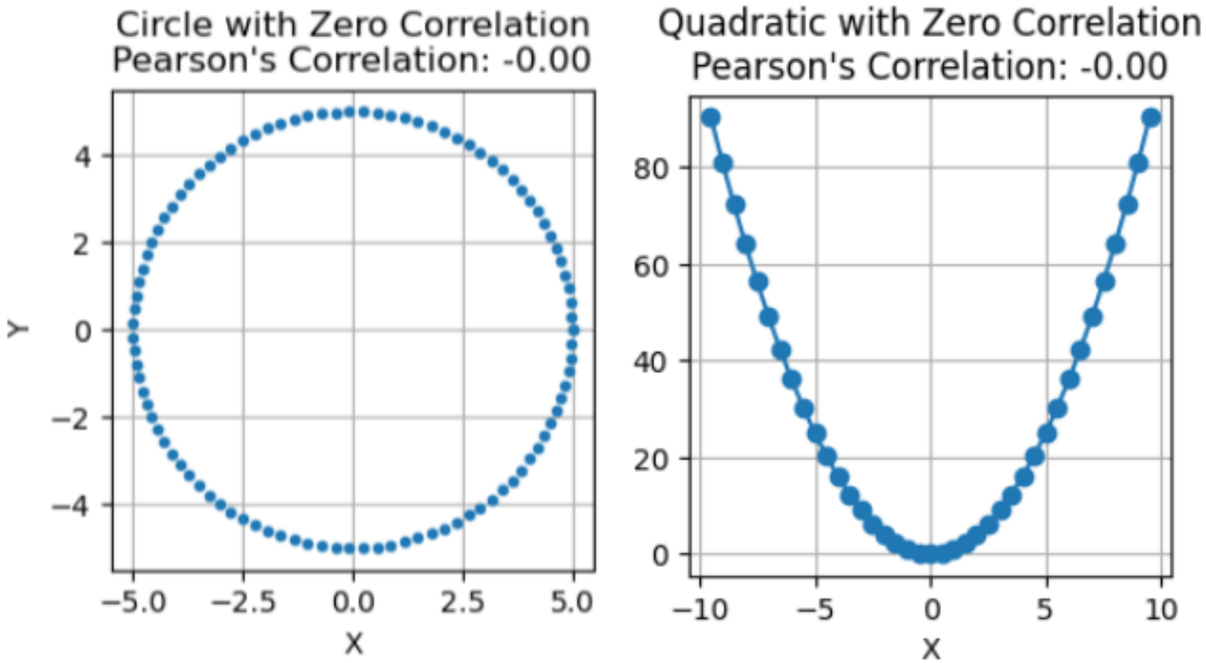
$$\mathrm{Cov}(X,Y) = \rho\sigma_x\sigma_y$$

Covariance

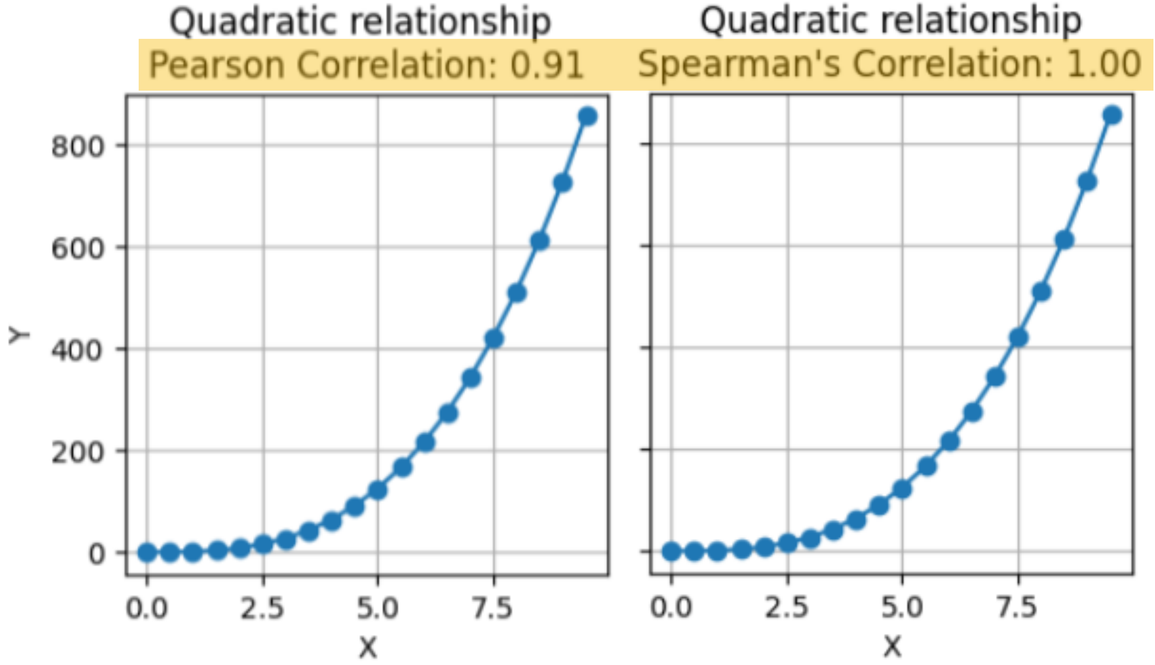
Variance of x

Variance of y

Pearson's correlation only features linear relationship



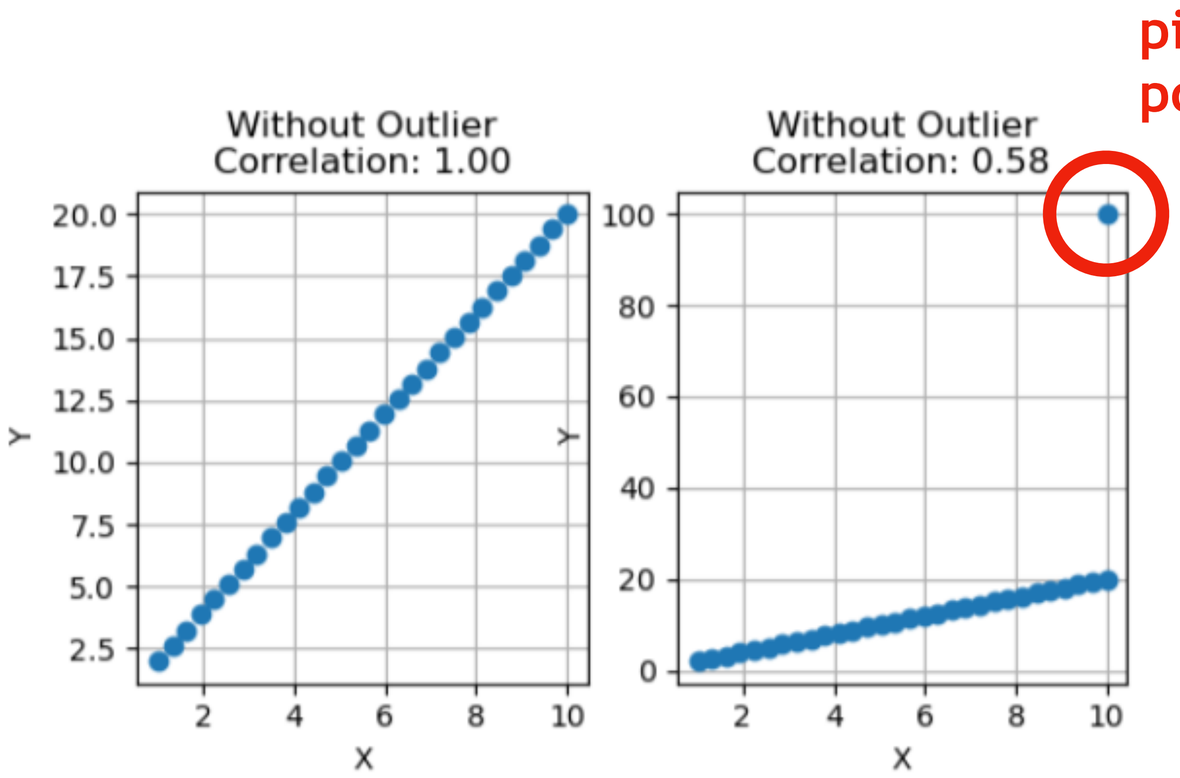
A small or equal to zero Pearson's correlation does not mean variables have no relationships



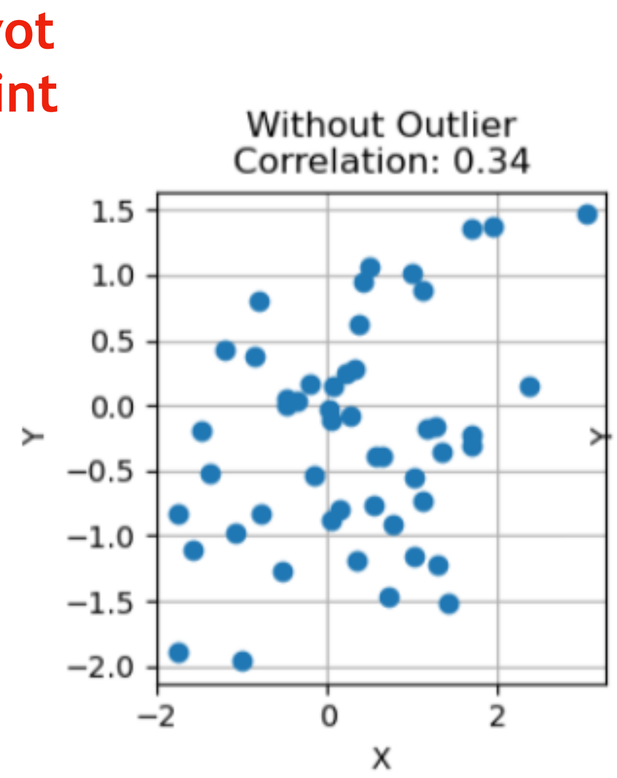
If the true relationship is non-linear, Pearson's correlation cannot fully depict that relationship

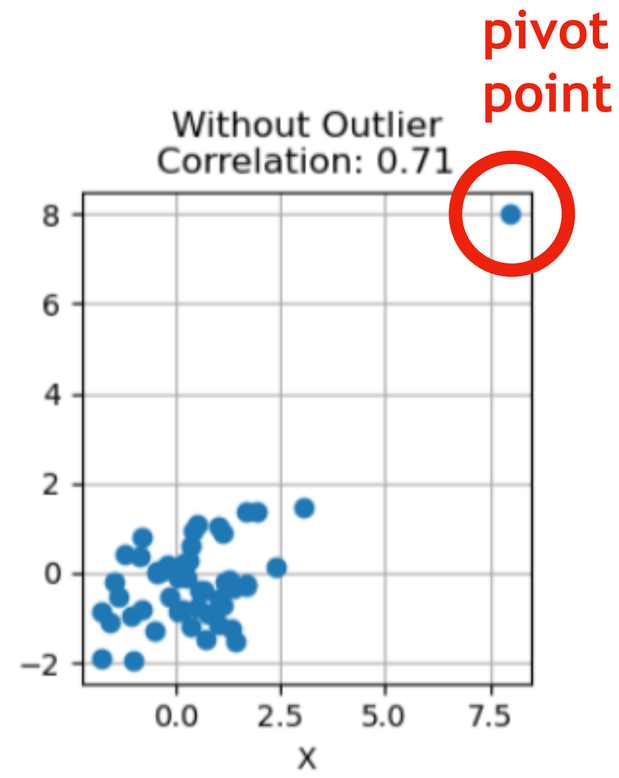
scipy.stats.spearmanr (correlation based on rank rather than data values)

Pearson's correlation is sensitive to outliers



A good correlation can be destroyed by one or several pivot point



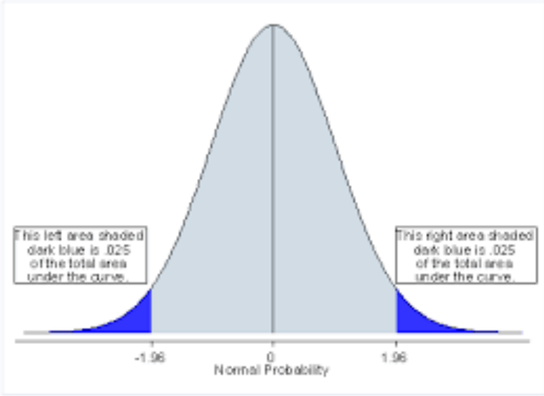


One or several pivot points give you a "fake correlation"

Whether a Correlation Coefficient is Statistically Significant?

Terminology	Meaning	When testing the significance of linear correlation
Null hypothesis (H_0)	This is our starting assumption that the effect being studied does not exist	X and Y are not linearly correlated ρ = 0
Alternative hypothesis (H_1)	This is what we might believe to be true if we find sufficient evidence against the null hypothesis.	X and Y are linearly correlated. ρ ≠ 0
Test statistics	This is a calculated value from our data that we use to test our hypothesis.	t-statistics; $\rho\sqrt{\frac{n-2}{1-\rho^2}}$
Null distribution	This represents what we would expect to see from our test statistic purely by chance if the null hypothesis were true.	Correlation coefficient that can be generated from uncorrelated X and Y.
significance level (α)	This is a threshold we set to decide when to reject the null hypothesis.	A common choice is $\alpha = 0.05$.
p-value	The probability of obtaining our data, or something more extreme, if the null hypothesis is true.	Percentage of data having a higher absolute value of correlation coefficient than your data, which is returned by **scipy.stats.pearsonr**.

For correlation, we are doing a two-sided testing because correlation can be either positive or negative.

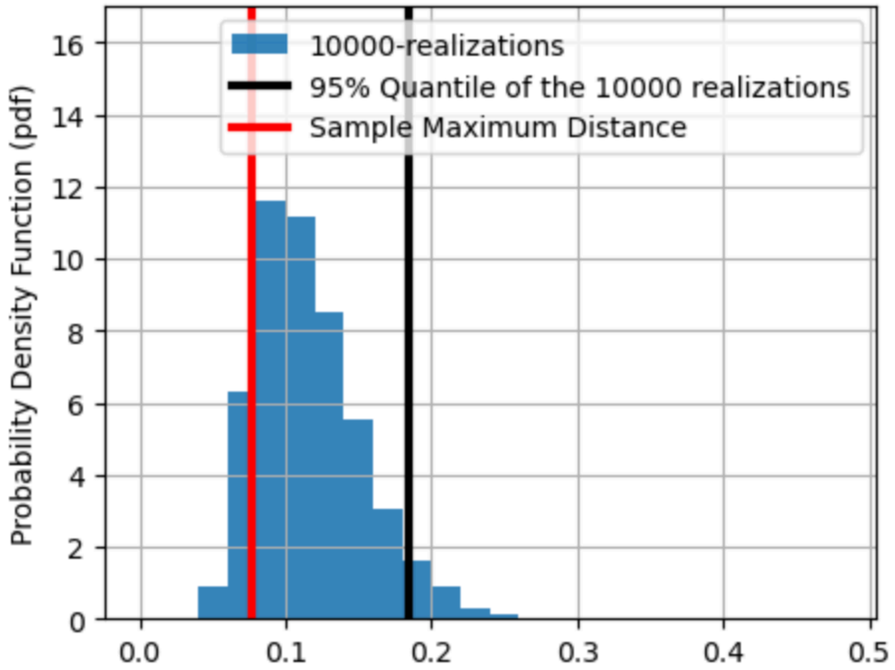


When $p < \alpha/2$, we can reject H_0, and accept H_1.

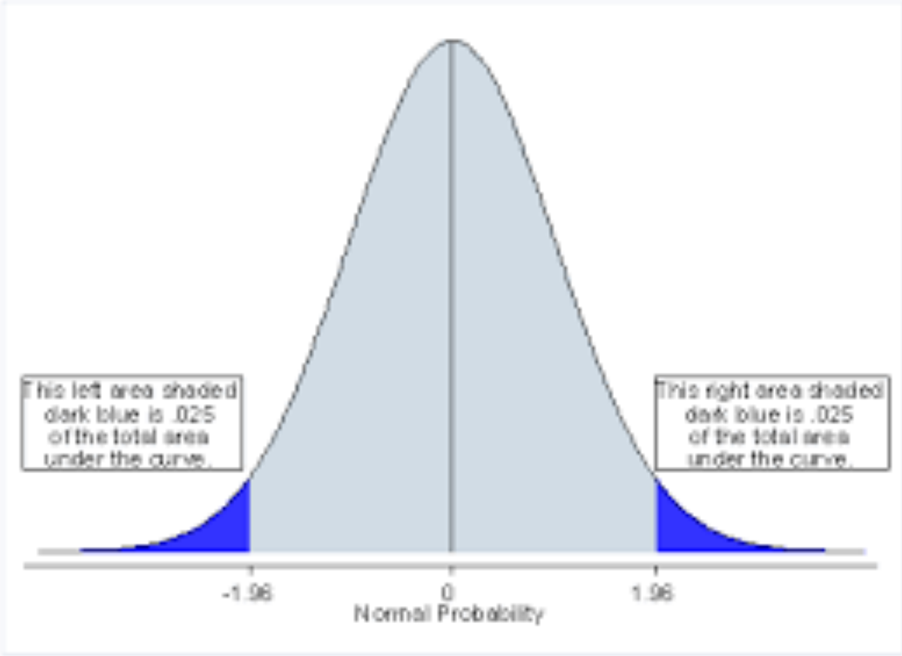
```
r, p = scipy.stats.pearsonr(x, y)
```

One-sided or Two-sided?

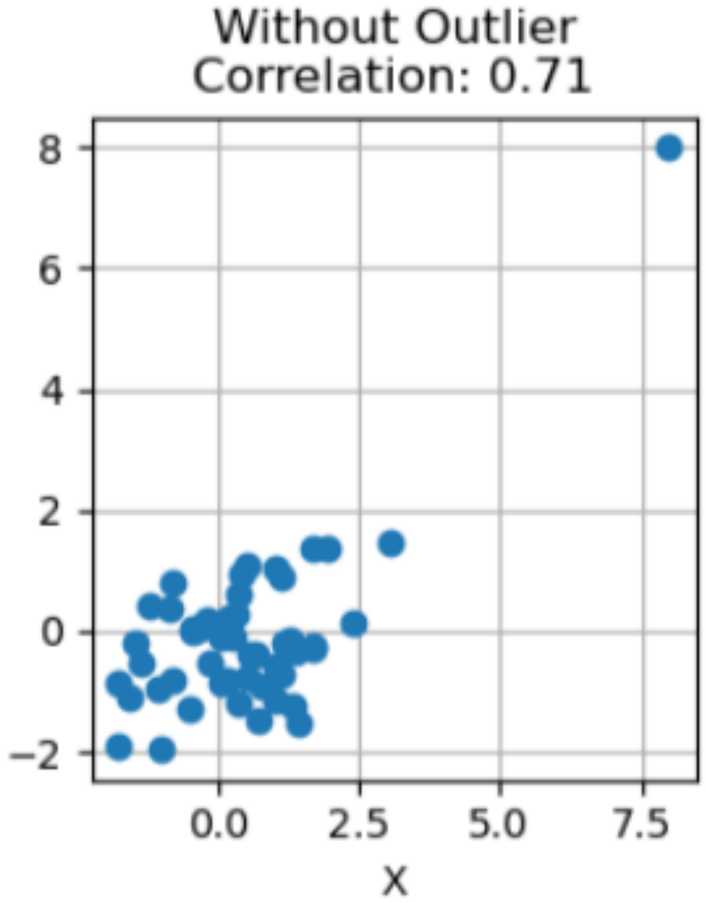
One-sided test



Two-sided test



The python package does not handle outliers



The package assumes that

Data follow a Normal distribution

```
r, p = scipy.stats.pearsonr(x, y)
```

$p < 0.01$

Bootstrapping: A Robust Alternative

Using resampling to generate slightly perturbed versions of your data and capture uncertainties arose from randomness.

(1) Resampling with replacement

(2) Calculate the target statistics on resampled data

(3) Repeat to generate a distribution

Some points can be sampled more than once, whereas some others do not appear in the resampled data.

for ct in np.arange(N_boot):

Resample your data with replacement

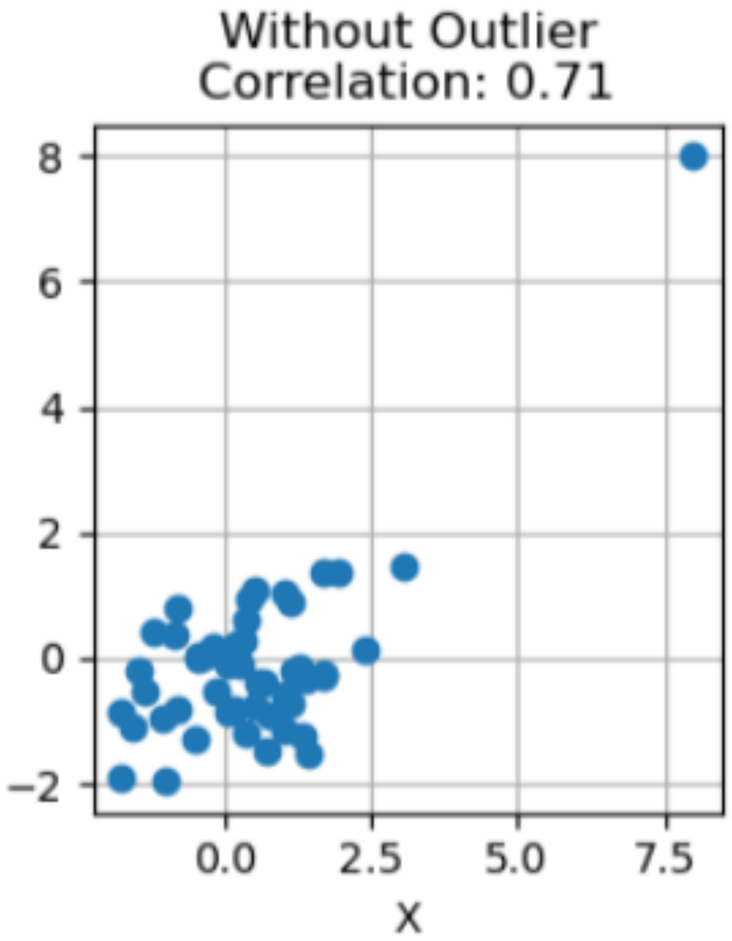
Calculate statistics using resampled data

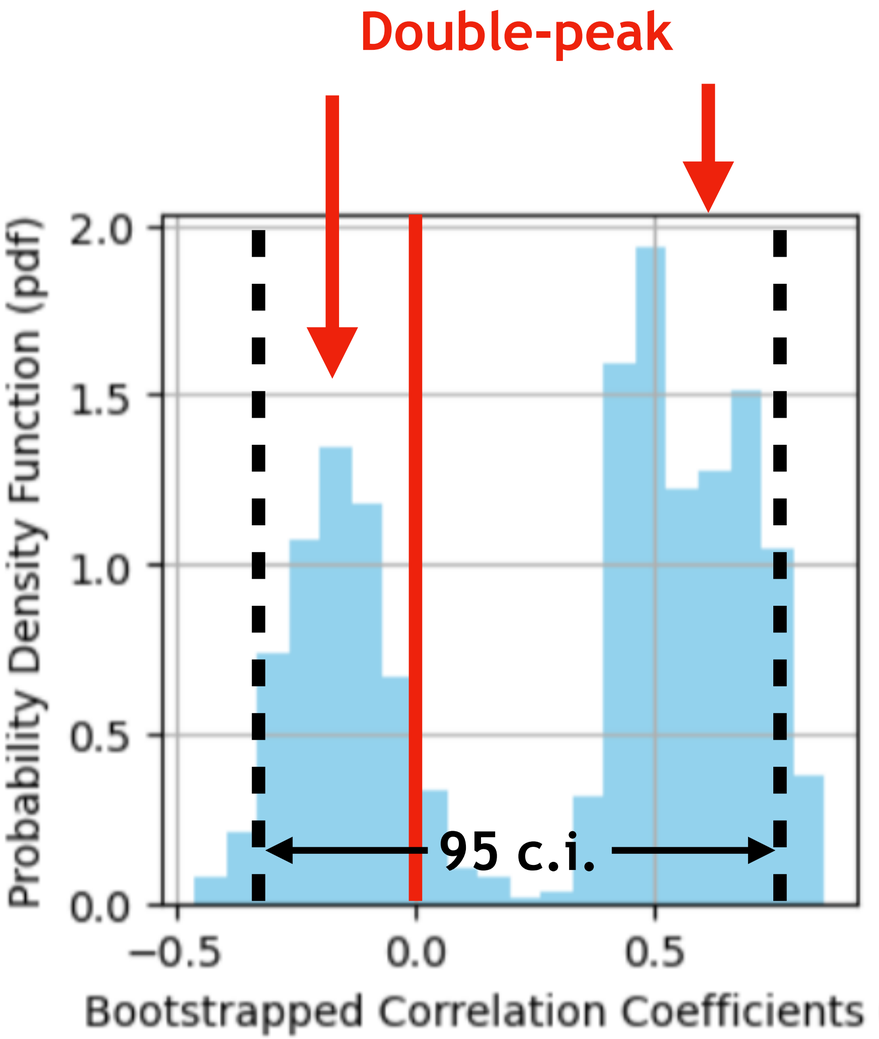
Save calculated statistics in an array

Evaluate the confidence interval of statistics

np.random.randint()

Bootstrapping: A Robust Alternative





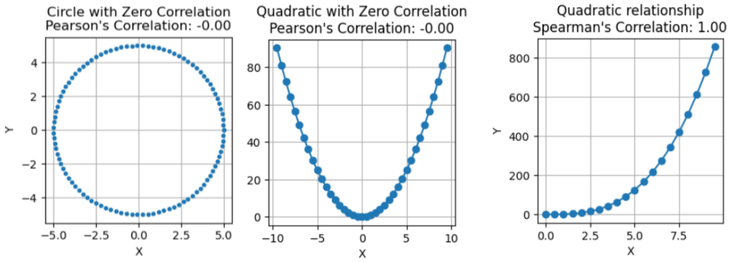
Road Map of the Statistics Part

	Lecture 5	Lecture 6
Quantification Technique	Mean, variance, skewness, & kurtosis	Pearson's Correlation (Linear relationship)
Uncertainty & Significance	Gaussian distribution Chi-2 distribution	r, p = scipy.stats. pearsonr(x, y)
Assumptions	Data is Gaussian or follows specific types of distribution Independent Sampling	Data is Gaussian Independent Sampling
Test assumptions	K-S test	
Treatment		Bootstrapping

Prepare

Remove mean / seasonal cycle...

Plot and visualise your data



Are we interested in a linear relationship?

No → Use other methods
e.g., Spearman's rank correlation, or regression

Yes

Analysis

Calculate Pearson's Correlation → r, p = scipy.stats.pearsonr(x, y)

Significance

Is data Gaussian?

K-S test

No → Bootstrapping

Yes → use scipy's p value to evaluate significance