Class 2: Correlation Testing

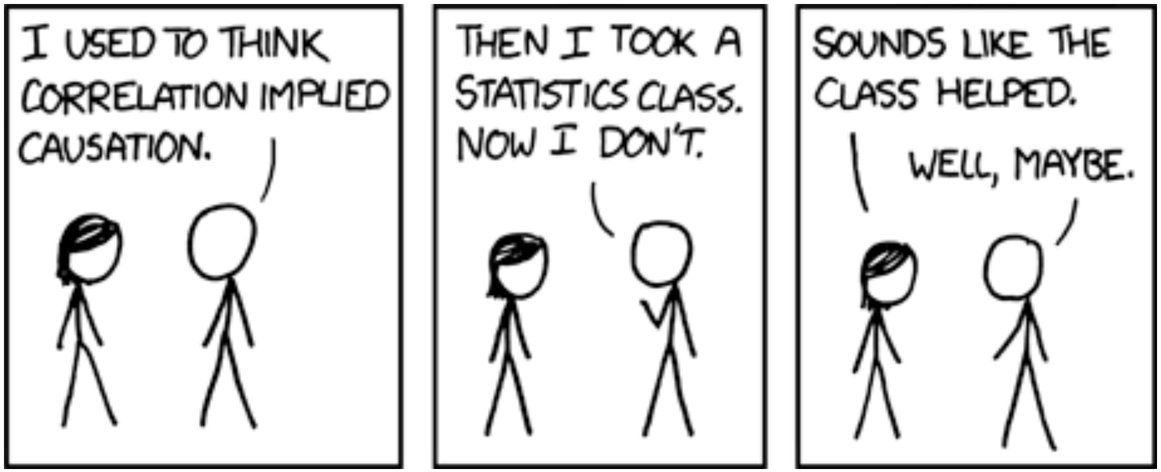
In this class we will review how to quantify correlations between variables and test for their significance, and determine whether different samples are drawn from the same underlying distributions

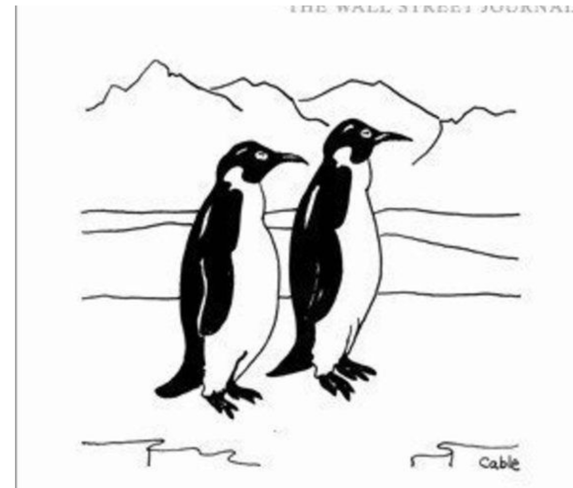
Class 2: Correlation Testing

At the end of this class you should be able to …

- … test for the degree of correlation between 2 variables, and its significance
- … implement correlation as a hypothesis test, and understand the significance of the resulting p-value
- … test if two samples are drawn from the same parent distribution
- … appreciate the pitfalls that can arise when searching for correlations

Correlation versus independence





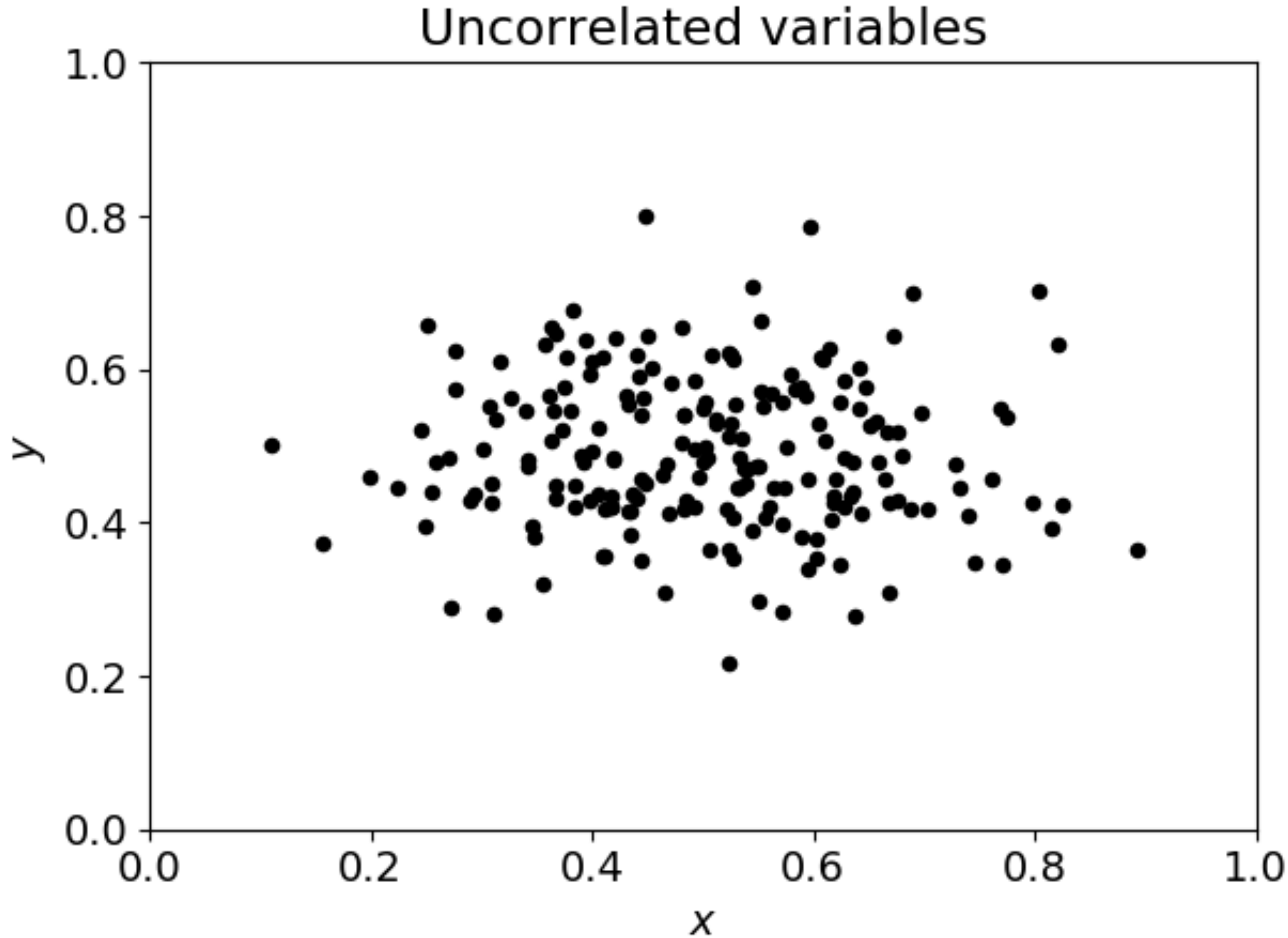
"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

- Two variables are **correlated** if they share a statistical dependence / relationship

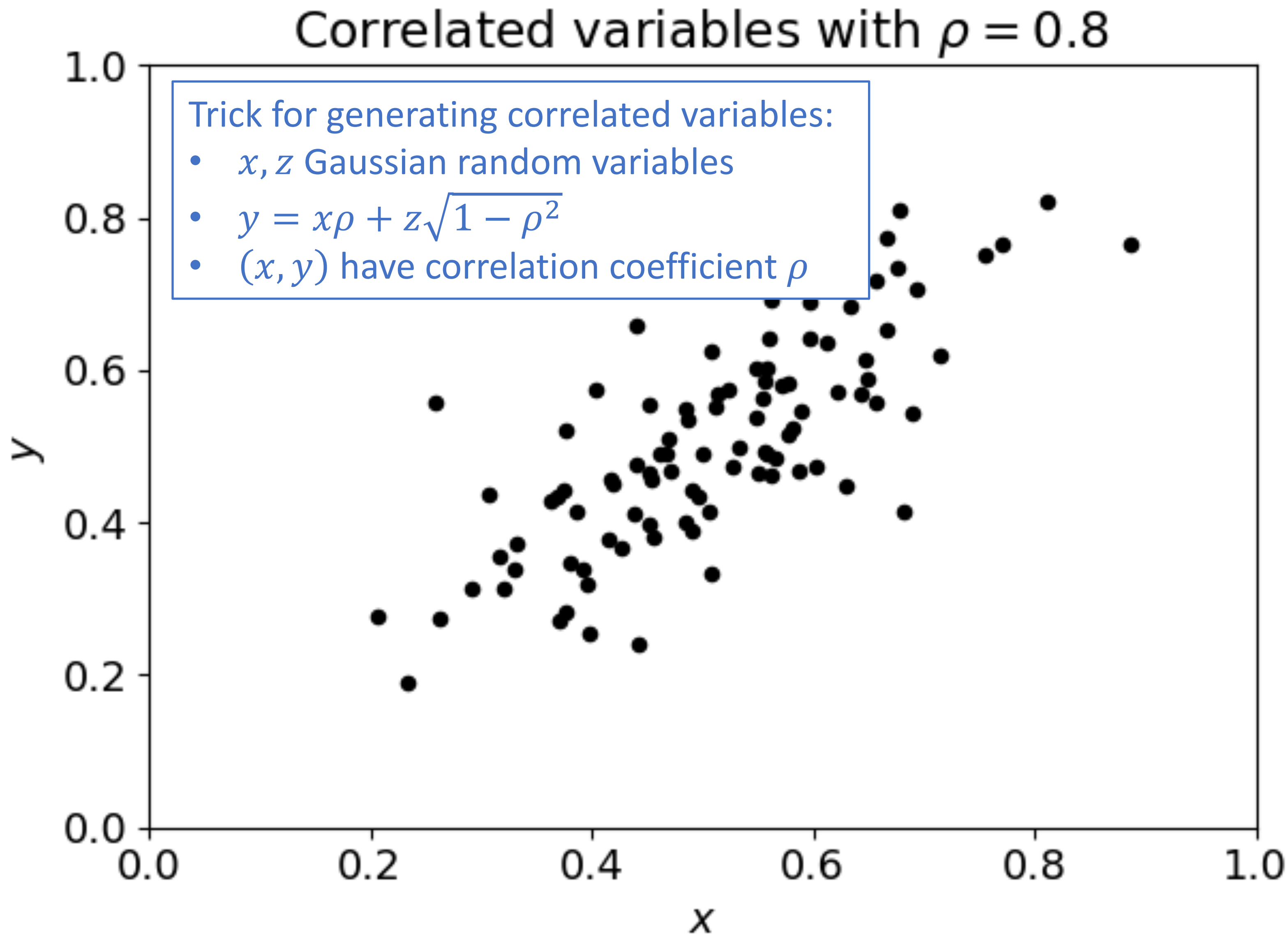
- E.g., the daily temperatures at noon and 1pm are correlated, because they both lie above the mean temperature

- Correlations between variables could indicate some **underlying physical relationship** between those variables

Correlation versus independence



Correlation versus independence



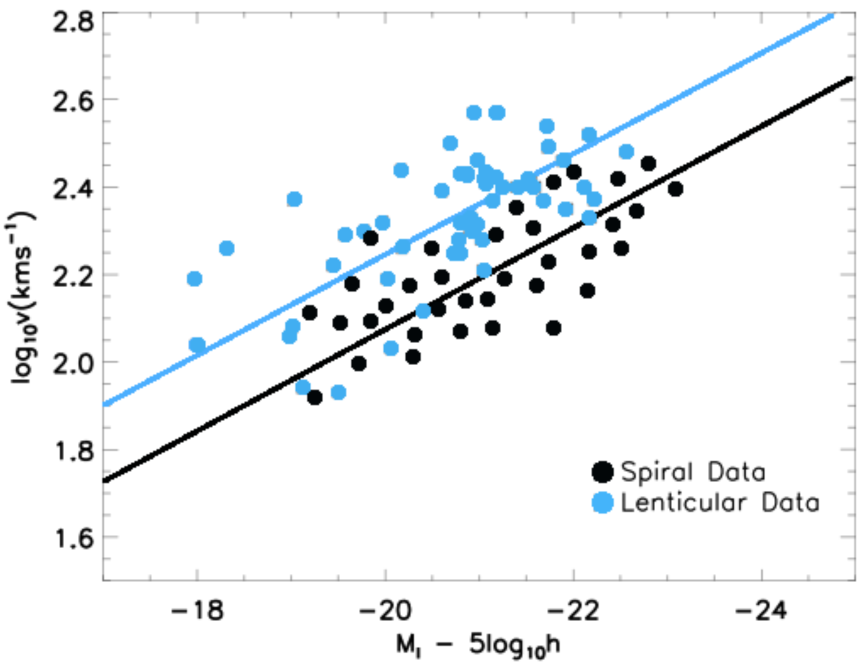
Trick for generating correlated variables:

- x, z Gaussian random variables
- $y = x\rho + z\sqrt{1-\rho^2}$
- (x, y) have correlation coefficient ρ

Correlations in astrophysics

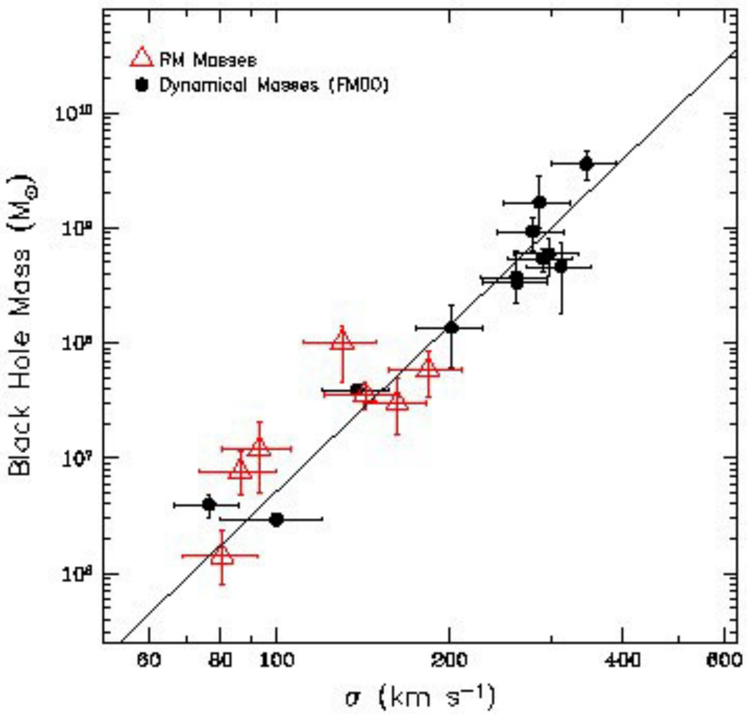
- Astrophysics contains many correlations!

Tully-Fisher relation



Credit: https://en.wikipedia.org/wiki/Tully–Fisher_relation

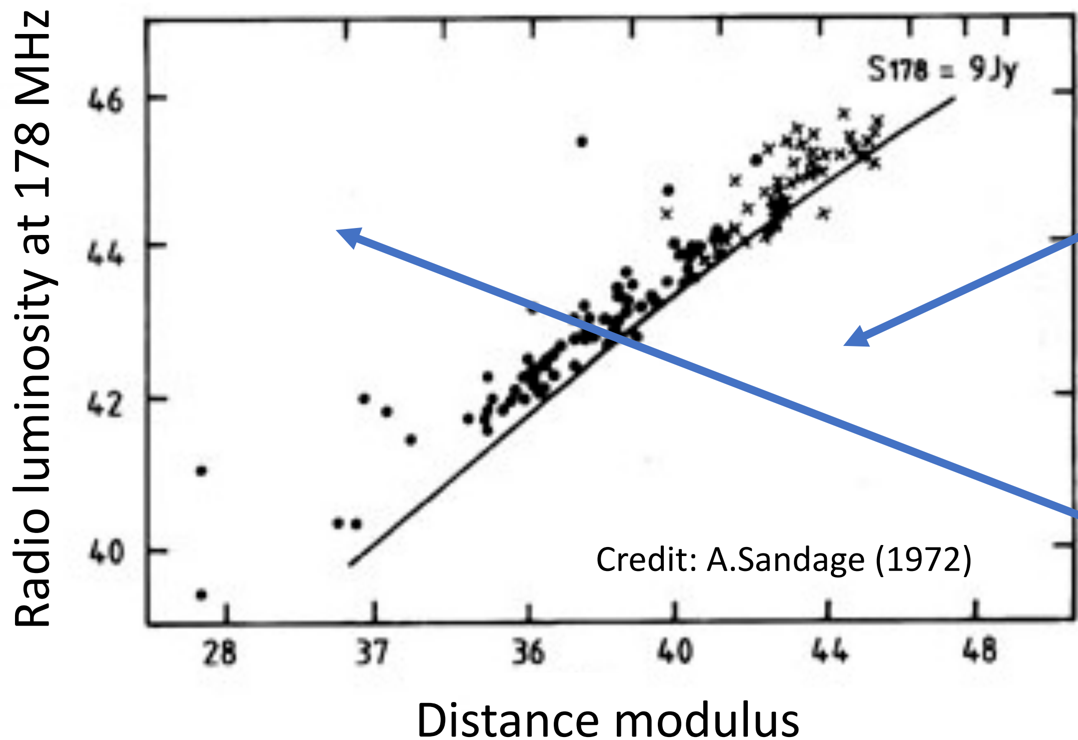
Black hole - bulge relation



Credit: ned.ipac.caltech.edu

Pitfalls when searching for correlations

- **Selection effects** can easily lead to spurious correlations
- *Here is a perfect luminosity-redshift correlation for radio galaxies in the 3CR survey:*

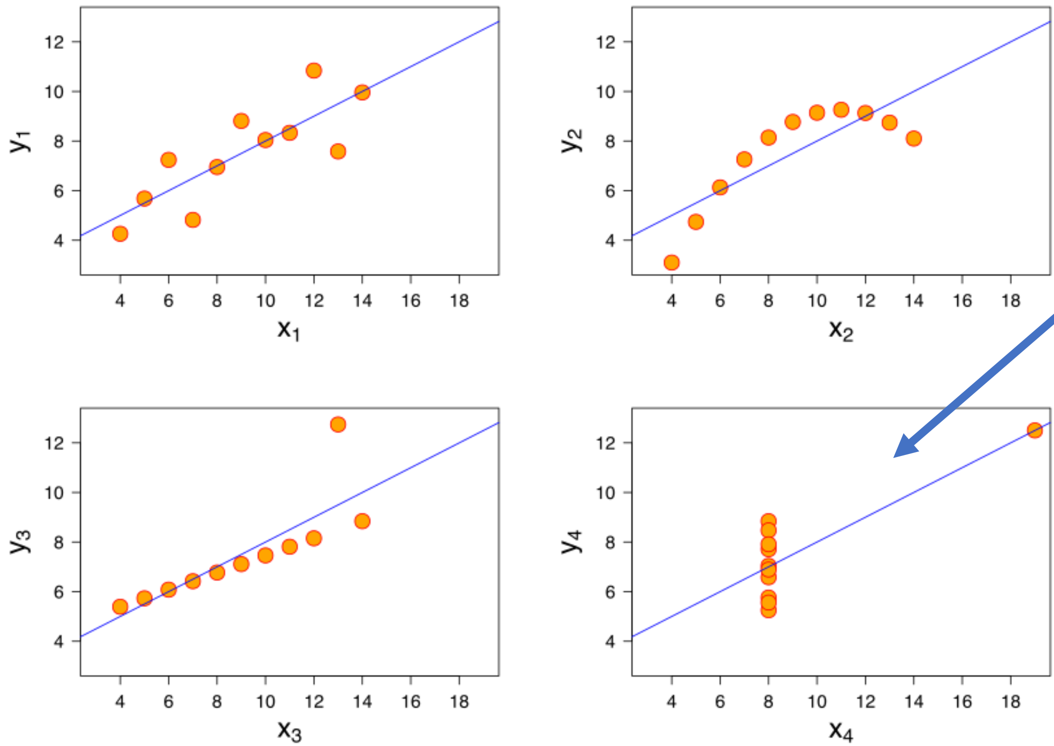


Galaxies cannot appear in this region because they are below the flux limit

Galaxies cannot appear in this region because of the steepness of the luminosity function

Pitfalls when searching for correlations

- Correlations can be driven by a **small number of outliers**
- *The following four (x, y) datasets all have the same mean, variance, correlation coefficient and regression line:*



The correlation here is completely driven by a single outlier

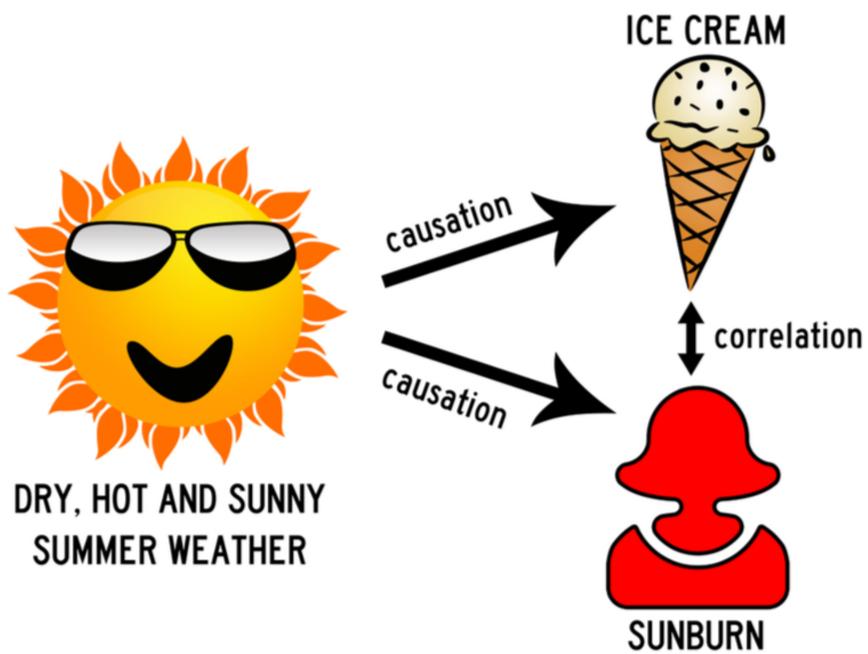
"Rule of thumb": if a correlation goes away after you cover part of the dataset with your thumb, it probably isn't real!

Credit: Anscombe's quartet (https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

Pitfalls when searching for correlations

- **Correlation is not the same as causation**
- The correlation of two variables does not necessarily imply a causal/direct connection. They might both be driven by a **“third variable”**.

Eating ice cream causes sunburn??



Credit: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Procrastinate by checking a few more examples at https://www.tylervigen.com/spurious-correlations

Correlation coefficient

- The **correlation coefficient** describes the **strength of the correlation** between two variables (x, y)

- If the variables have means (μ_x, μ_y) and standard deviations (σ_x, σ_y), then the definition of the correlation coefficient ρ is:

$$\rho = \frac{\langle (x - \mu_x)(y - \mu_y) \rangle}{\sigma_x \sigma_y} = \frac{\langle xy \rangle - \mu_x \mu_y}{\sigma_x \sigma_y}$$

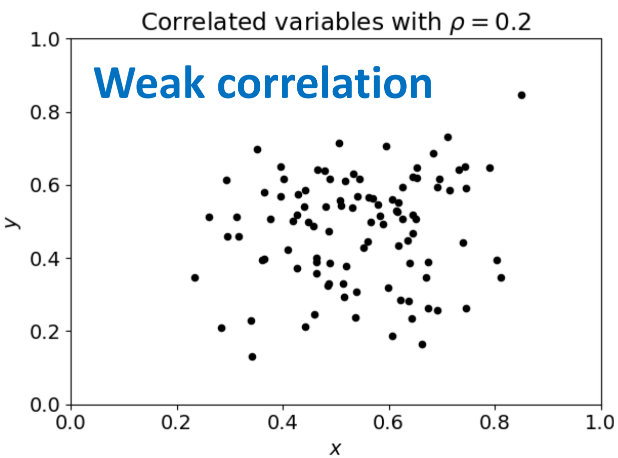
$$\langle xy \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, y\, P(x, y)\, dx\, dy$$

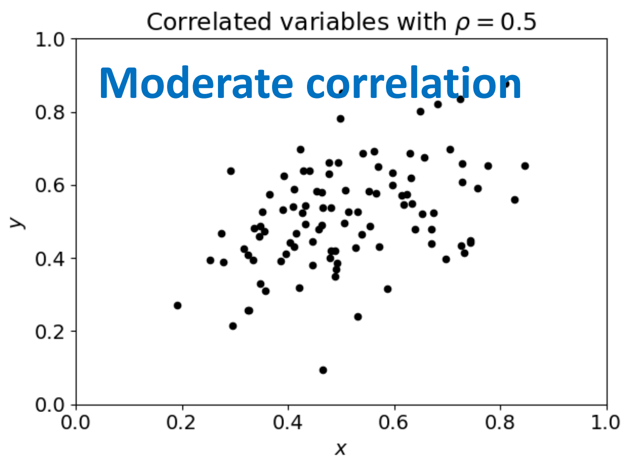
- [Small print: we'll use ρ to mean the underlying **theoretical** correlation coefficient, and r as the **value estimated from data**, i.e. $\hat{\rho} = r$]

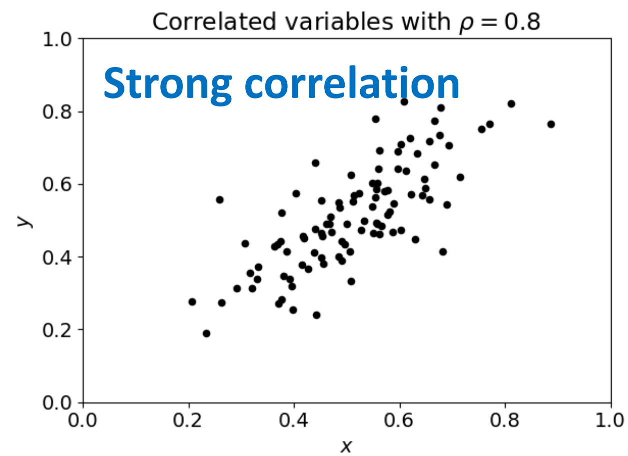
Correlation coefficient

$$\rho = \frac{\left\langle (x-\mu_x)(y-\mu_y) \right\rangle}{\sigma_x \sigma_y} = \frac{\langle xy \rangle - \mu_x \mu_y}{\sigma_x \sigma_y}$$

- For **no correlation**, $P(x, y)$ is separable into $f(x)\ g(y)$, hence $\langle xy \rangle = \langle x \rangle \langle y \rangle = \mu_x \mu_y$ and $\rho = 0$
- For **complete correlation**, $y = Cx$ and $\rho = +1$
- For **complete anti-correlation**, $y = -Cx$ and $\rho = -1$
- *The possible range is* $-1 \leq \rho \leq +1$







Pearson product-moment correlation

- We can estimate the correlation coefficient of data samples (x_i, y_i) using the **Pearson product-moment** formula:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{(N-1)\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}}$$

- Can compare this formula with the definition $\rho = \frac{\langle xy \rangle - \mu_x \mu_y}{\sigma_x \sigma_y}$ and see that $\boldsymbol{r}$ **is an estimator of** $\boldsymbol{\rho}$

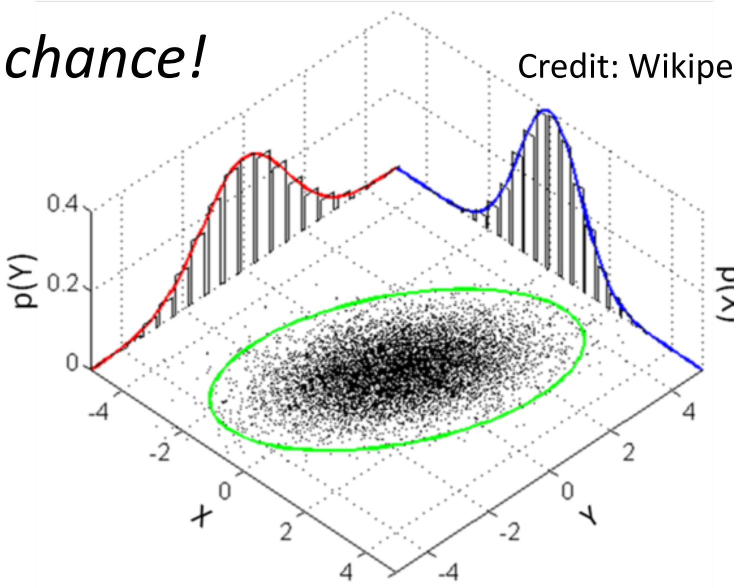
- The possible range of values is $-1 \leq r \leq +1$

Significance of correlation

- When correlation-testing, it is **not** sufficient to just measure r. We also need to check the **significance of the correlation**

- *Correlations can arise by random chance!*

Credit: Wikipedia



- Let's model the data by supposing (x, y) are drawn from a **bivariate Gaussian distribution** about an underlying relation [which often works pretty well]

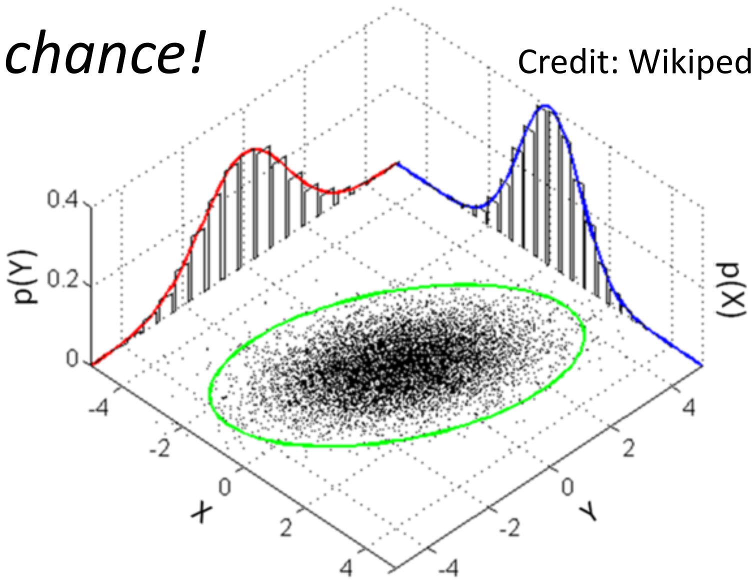
$$P(x,y) = \frac{\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x{}^2} + \frac{(y-\mu_y)^2}{\sigma_y{}^2} - \frac{2(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

Significance of correlation

- When correlation-testing, it is **not** sufficient to just measure r. We also need to check the **significance of the correlation**

- *Correlations can arise by random chance!*

Credit: Wikipedia



- Let's model the data by supposing (x, y) are drawn from a **bivariate Gaussian distribution** about an underlying relation
  [which often works pretty well]

- If this model is true, then the **uncertainty** in the measured value of r, if we have N data points, is:

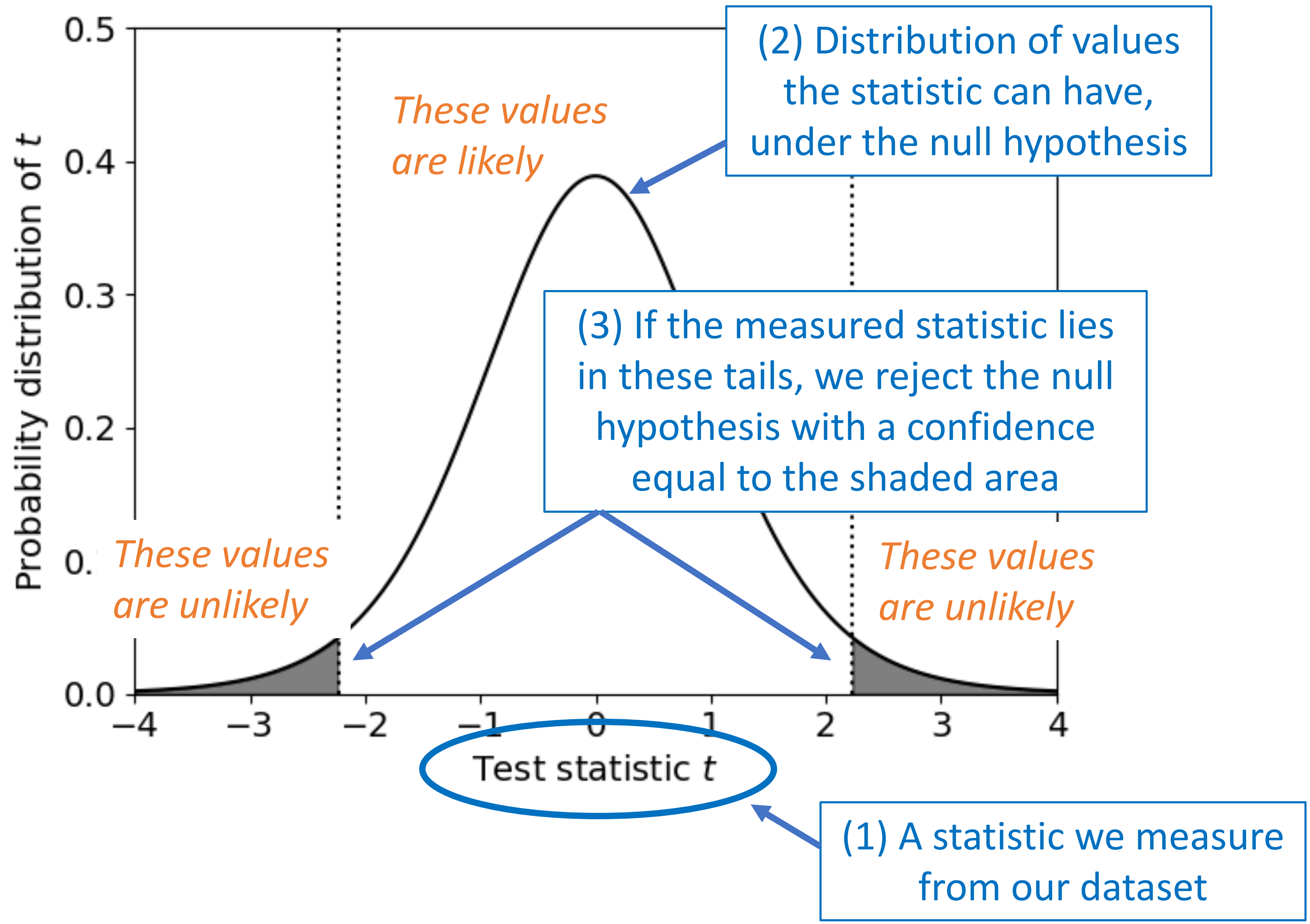
$$\boxed{\sigma(r) = \sqrt{\frac{1 - r^2}{N - 2}}}$$

Hypothesis tests

- **Hypothesis tests** are a common approach for addressing statistical questions in the frequentist framework
- They typically involve a **null hypothesis**, a **test statistic**, a **distribution of values** that statistic can take if the hypothesis is true, and a **tailed confidence limit**
- *Let's see an example …*

Credit: xkcd.com

Hypothesis tests



Significance of correlation

- *Let's apply this approach to correlation testing*

- **Null hypothesis**: there is no correlation between the variables

- **Test statistic**: $\boxed{t = r\sqrt{\dfrac{N-2}{1-r^2}}}$

  r = correlation coefficient

  N = number of data points

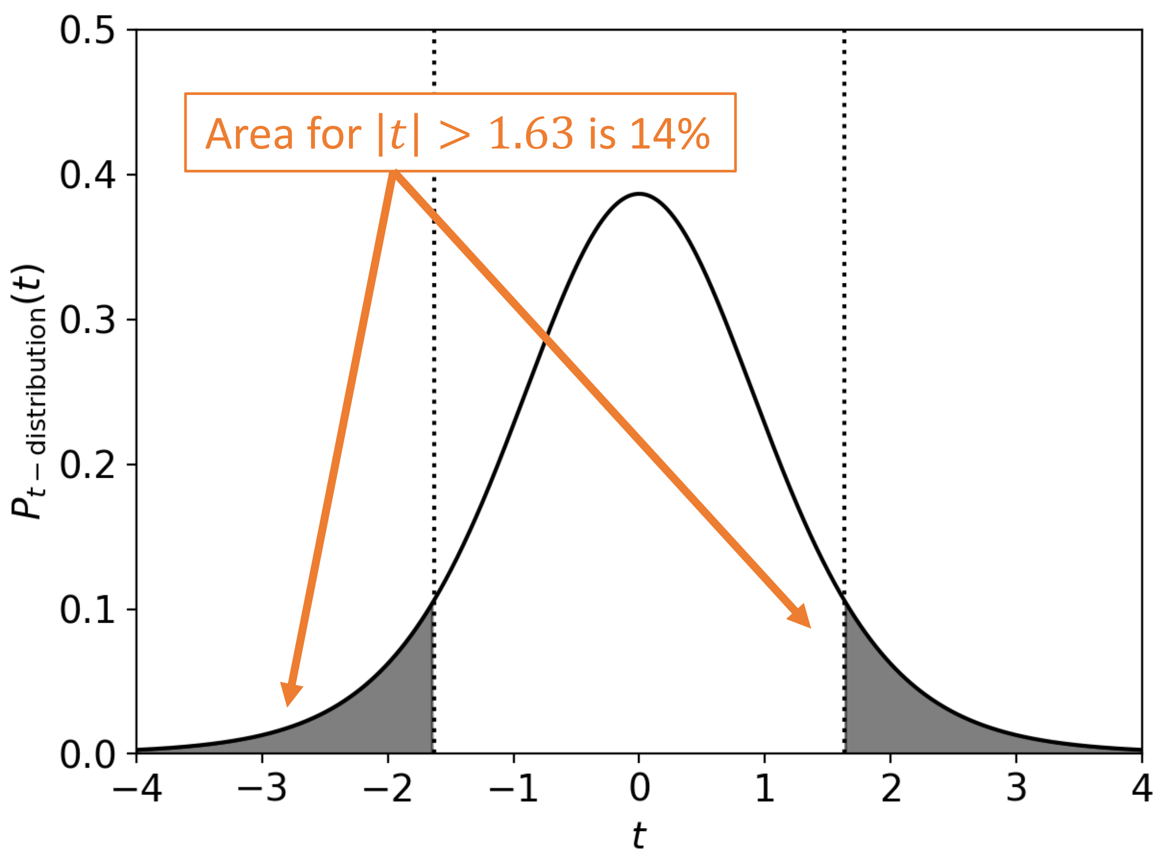
- **Distribution followed by the statistic**: the Student's t probability distribution with number of degrees of freedom $\nu = N - 2$

- **Probability of rejecting the hypothesis**: the area under the tails at higher values of $|t|$ than we have measured

Significance of correlation

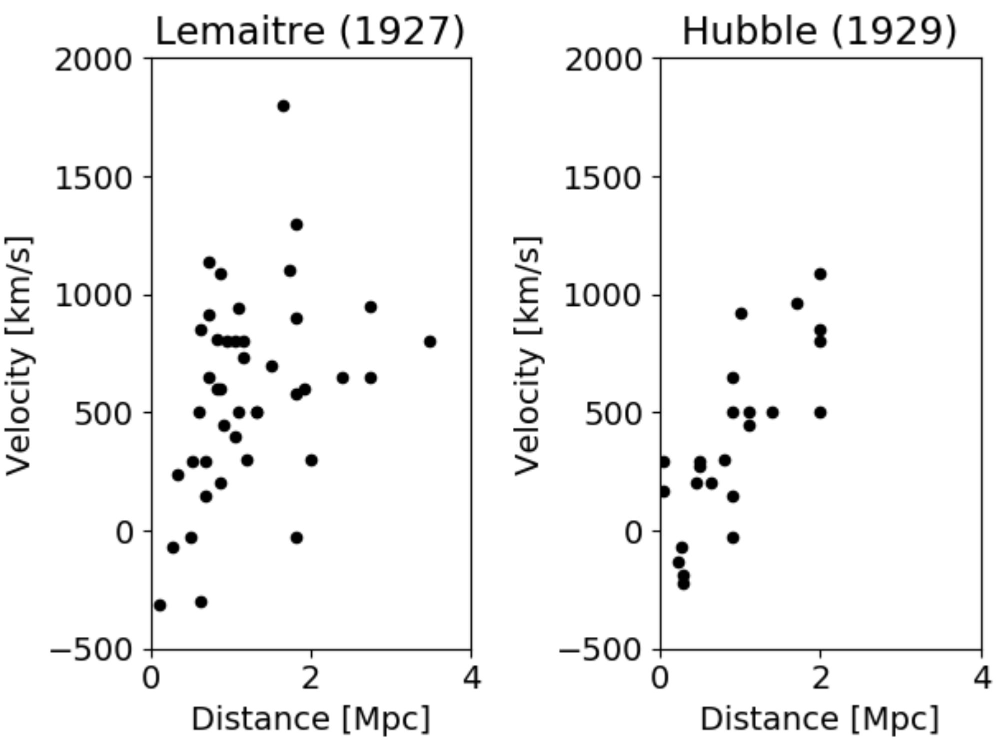
- *Example: we measure $r = 0.5$ for $N = 10$ points. Is this correlation significant?*

- We find $t = 1.63$, $\nu = 8$
- The probability of finding $|t| > 1.63$ is **14%**
- This is **not** sufficiently small to reject the hypothesis of no correlation: **this correlation is not significant**
- [we would typically reject with (e.g.) 95, 99% confidence]



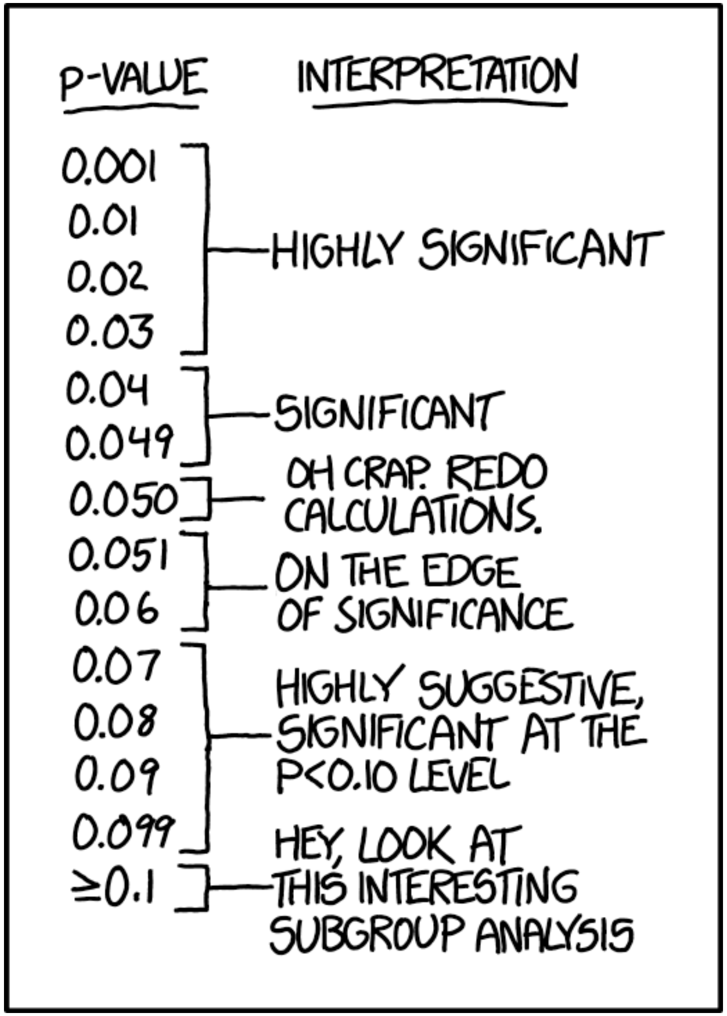
Hubble and Lemaitre's datasets

- In this Activity we will check **who discovered the expansion of the Universe**! See Hubble and Lemaitre's distance-velocity datasets. For the two datasets, determine the Pearson correlation coefficient, its error and statistical significance



We need to talk about p-values!

- The probability of rejecting a hypothesis is often known as a **“p-value”**
- It corresponds to the **“significance”** of a result
- *Let’s talk about exactly what this value means, since this can be pretty confusing*



Credit: xkcd.com

Hypothesis tests and p-values

- Suppose a (no-) correlation significance yields $p = 0.01$
- This means: **there is a 1% chance of obtaining a set of measurements at least this correlated, if the underlying data is uncorrelated**. It does not mean:
- "the probability that the points are uncorrelated is 1%"
- "the probability that the points are correlated is 99%"
- "if we claim a correlation, there is a 1% chance that we would be mistaken"
- **Frequentist statistics cannot assess the probability that the model itself is correct** (see – Bayesian statistics)

Non-parametric correlation tests

- If we do not want to assume that (x, y) are drawn from a bivariate Gaussian, we can use a **non-parametric correlation test**
- Let (X_i, Y_i) be the rank of (x_i, y_i) in the overall order, such that $1 \leq X_i \leq N$ and $1 \leq Y_i \leq N$
- Compute the **Spearman rank correlation coefficient**

$$r_s = 1 - 6\,\frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{N^3 - N}$$

- Convert the correlation coefficient into a **probability**, using the Student's t distribution as before, with number of degrees of freedom $\nu = N - 2$

Bayesian correlation methods

- To determine the significance of our correlation, we have been asking, **"what is the probability of measuring a particular value of r if there is no correlation?"** Mathematically,

$$P(r|\rho = 0)$$

- Using Bayesian statistics we can ask the opposite question: **"what is the posterior probability distribution for the correlation coefficient ρ given the measured value of r?"** Mathematically,

$$P(\rho|r)$$

- [Good example of the difference in Frequentist and Bayesian methods.]

Bayesian correlation methods

- Assuming that (x, y) data are drawn from a bivariate Gaussian distribution as before, we can use Bayes' theorem to compute $P(\rho|r)$ marginalizing over the other parameters …

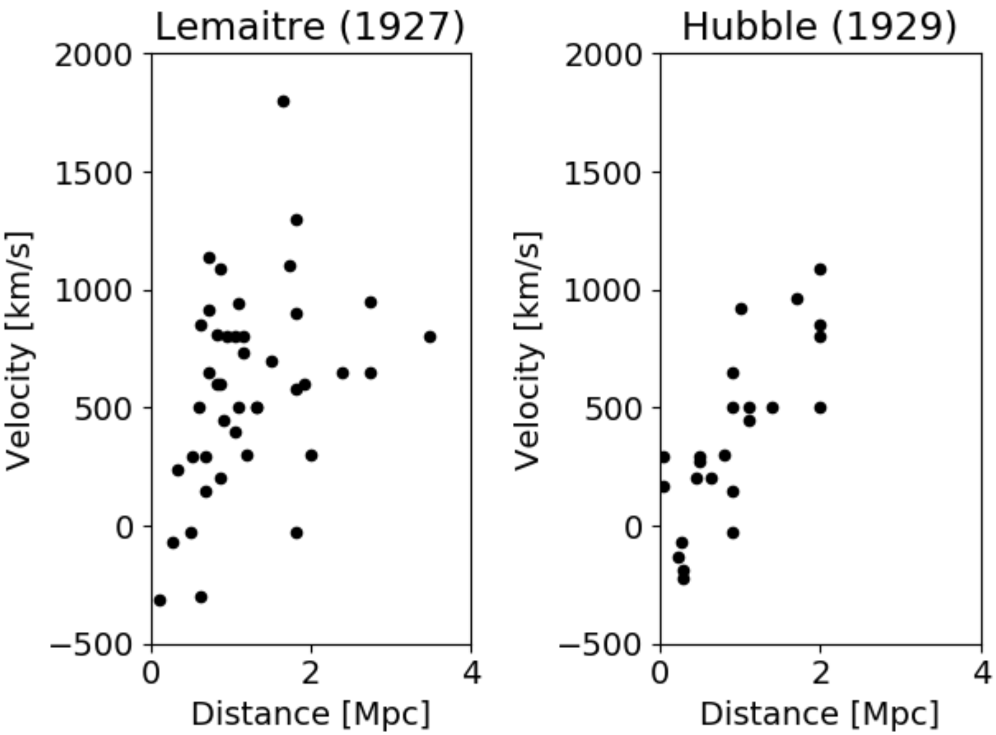
$$P(\rho|r) \propto \frac{(1-\rho^2)^{\frac{N-1}{2}}}{(1-\rho r)^{N-\frac{3}{2}}} \left(1 + \frac{1}{N-\frac{1}{2}} \frac{1+\rho r}{8} + \cdots\right)$$

- We can then substitute our values of r and N in this formula

- We obtain the **full probability distribution** of the underlying value of ρ, the correlation coefficient

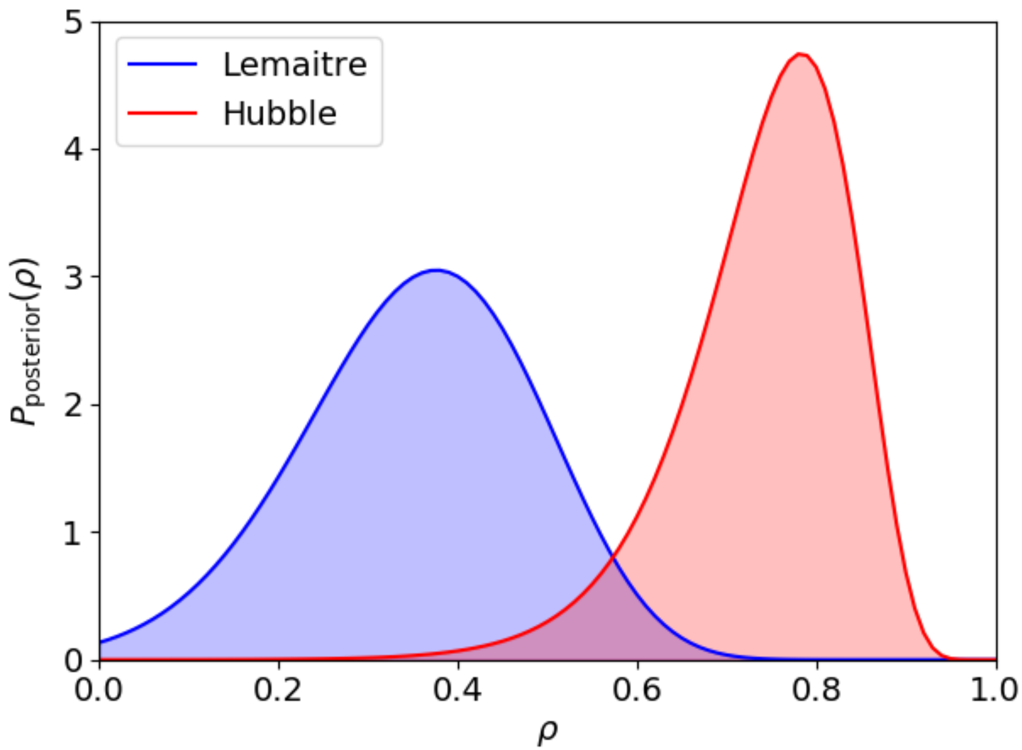
Hubble and Lemaitre's datasets

- Returning to Hubble and Lemaitre's distance-velocity datasets, now determine the Spearman rank correlation coefficient, its statistical significance, and the full probability distribution of $P(\rho|r)$ using the Bayesian formula.



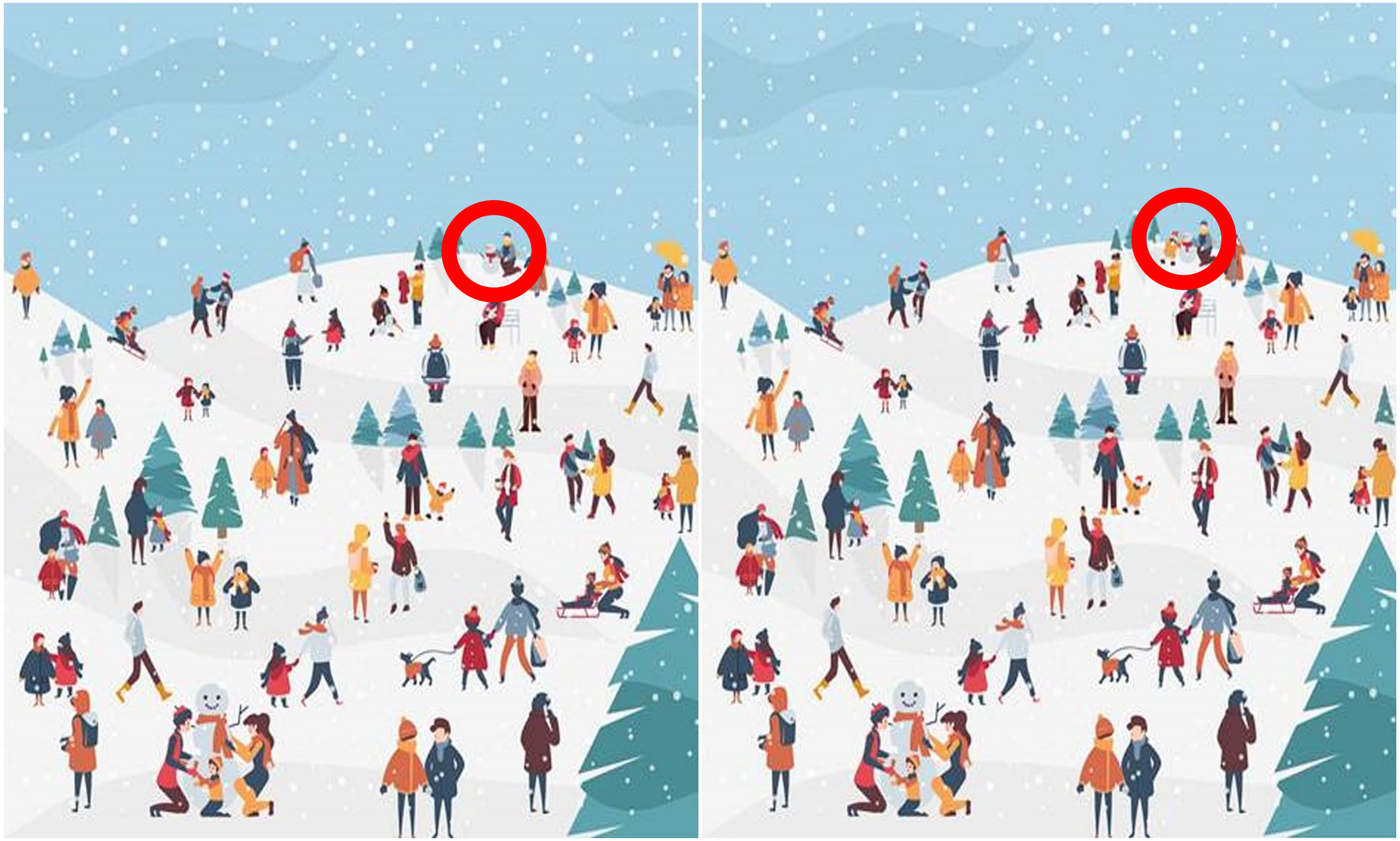
Hubble and Lemaitre’s datasets

- Returning to Hubble and Lemaitre’s distance-velocity datasets, now determine the Spearman rank correlation coefficient, its statistical significance, and the full probability distribution of $P(\rho|r)$ using the Bayesian formula.



Are two samples consistent?

- We now consider a related but different question: **testing whether two datasets are consistent**



Are the means of two samples consistent?

- Let's start with a test based on the **means and standard deviations** of 2 different samples *(this is known as a t-test)*

- Given the means (μ_x, μ_y) and standard deviations (σ_x, σ_y) of two samples of size (N_x, N_y), we can compute the $\boldsymbol{t}$ **statistic** and number of degrees of freedom ν:

$$t = \frac{|\mu_x - \mu_y|}{\sqrt{\frac{{\sigma_x}^2}{N_x} + \frac{{\sigma_y}^2}{N_y}}}$$

$$\nu = \frac{\left(\frac{{\sigma_x}^2}{N_x} + \frac{{\sigma_y}^2}{N_y}\right)^2}{\frac{{\sigma_x}^4}{{N_x}^2(N_x - 1)} + \frac{{\sigma_y}^4}{{N_y}^2(N_y - 1)}}$$

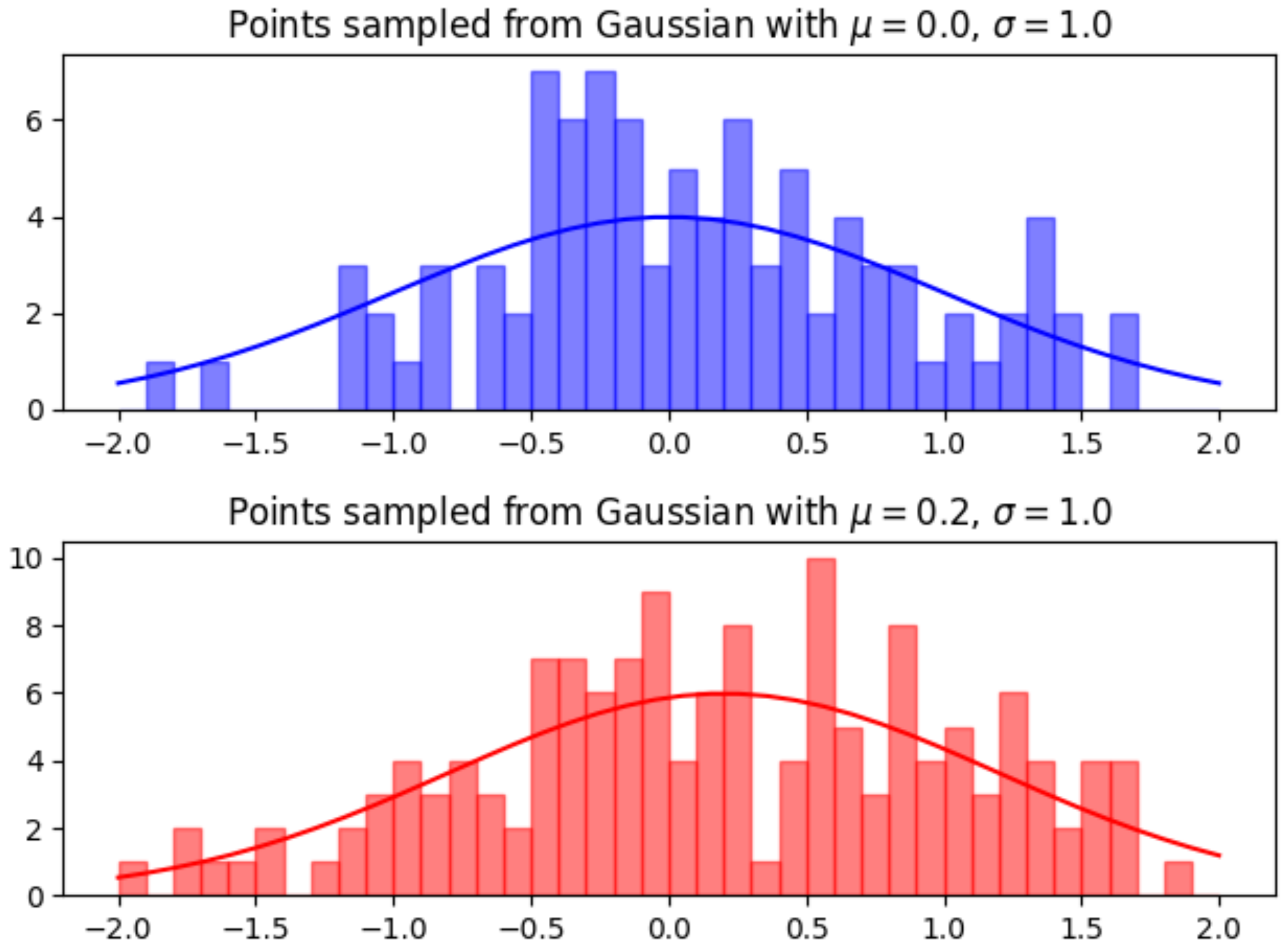
- We then compare these to **Student's $\boldsymbol{t}$ distribution** to obtain a p-value, as before

Kolmogorov-Smirnov test

- To test whether two full distributions are consistent (that is, drawn from the same parent distribution) we can use the **Kolmogorov-Smirnov (K-S) test**
- This test considers the **maximum value of the absolute difference between the two cumulative probability distributions**
- *Example: consider 2 datasets, (1)* $N = 100$ *points sampled from a Gaussian with* $\mu = 0$ *and* $\sigma = 1$*, (2)* $N = 150$ *points sampled from a Gaussian with* $\mu = 0.2$ *and* $\sigma = 1$*. Here they are …*

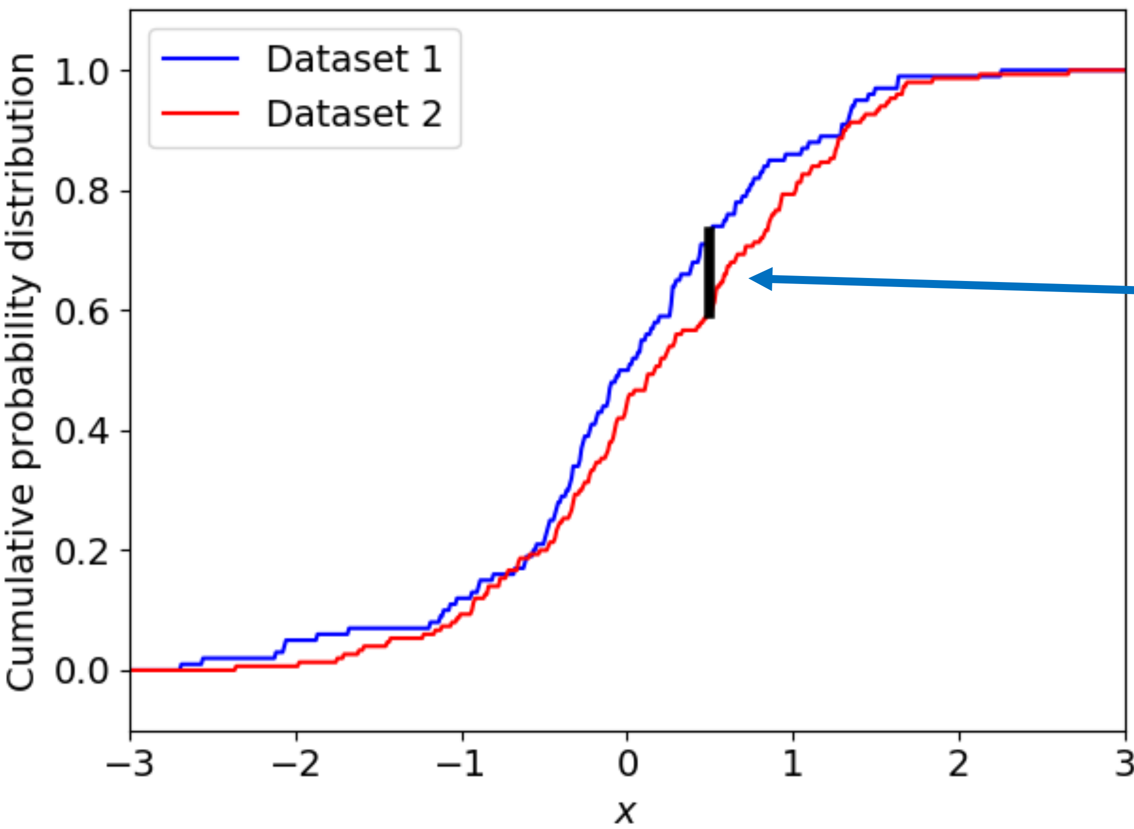
Kolmogorov-Smirnov test

- The data:



Kolmogorov-Smirnov test

- Cumulative probability distribution:

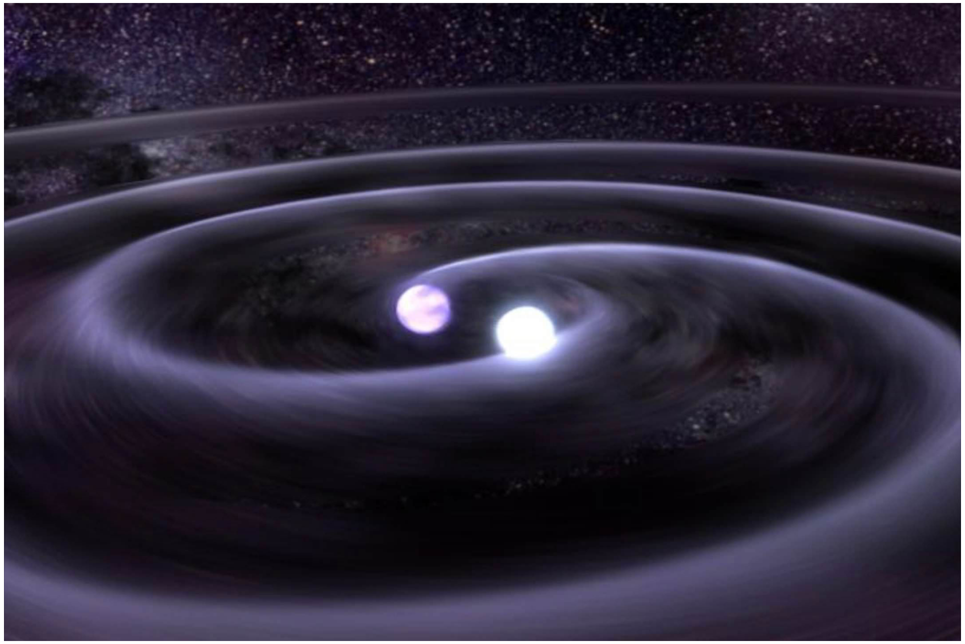


- The null hypothesis is that these datasets are drawn from the **same parent distribution**
- This is the **maximum deviation**, $d = 0.14$
- The **probability** of rejecting the null hypothesis is $p = 0.196$
- *We confirm the hypothesis!*

Kolmogorov-Smirnov test

- The provided datasets list estimated masses for neutron stars which **are** in double neutron star binaries and are **not** in double neutron star binaries

- Use the t-test to determine whether *there is any significant difference in the means of the two samples?*



- Use the K-S test to determine whether *these mass distributions are consistent?*

Summary

At the end of this class you should be able to …

- … test for the degree of correlation between 2 variables, and its significance
- … implement correlation as a hypothesis test, and understand the significance of the resulting p-value
- … test if two samples are drawn from the same parent distribution
- … appreciate the pitfalls that can arise when searching for correlations