Astro 426/526

Fall 2019

Prof. Darcy Barron

Lecture 15: Analyzing data

Reminders

- Mid-term grading in progress
- Initial project proposal due this Friday Oct 18 at 5pm (via Learn)
- Bring your laptop for Wednesday's class
- For next week, read Chapter 9 of *Practical Statistics for Astronomers* (Sequential Data – 1D Statistics)
- Other
  - Student Experience Project – you should have received an email to fill out a form about your experience at UNM. Please fill it out!
  - Software Carpentry workshop on the Unix shell on Friday, October 25 from 2-4pm, register here: https://ghz.unm.edu/education/workshops.html

Intro/overview of data analysis

- Reviewing and expanding on what we've covered so far in Practical Statistics for Astronomers
  - Borrowing heavily from lecture slides posted here:
  - http://www.astro.ubc.ca/people/jvw/ASTROSTATS/lectures/list.html

What are statistics?

- A statistic summarizes data (data reduction)
- Statistics are the basis for using the data to make a decision
- Example: Is the faint smudge on an image a star or a galaxy?
  - Measure FWHM of the point-spread function.
  - Measure full-width-half-maximum, the FWHM.
  - The data set, the image of the object, is now represented by a *statistic*

What is statistical analysis?

- 1. Formulate a hypothesis
- 2. Gather data to test the hypothesis (via experiment, or by finding existing datasets)
- 3. Compare with the expected probability of that result (the sampling distribution)

Problems:

We don’t know the actual underlying distribution

Small sample size

Important uses of statistics

- Statistics can create precise statements for stating the logic of what we are doing and why
- Statistics allow us to quantify uncertainty
  - Measured quantities are basically useless without some measure of the associated range/error
  - Sometimes this can be inferred, but much better to be explicit (e.g. 5 photons, 72.1 degrees)
- Statistics help us avoid pitfalls like confirmation bias
- Statistics help make decisions about data

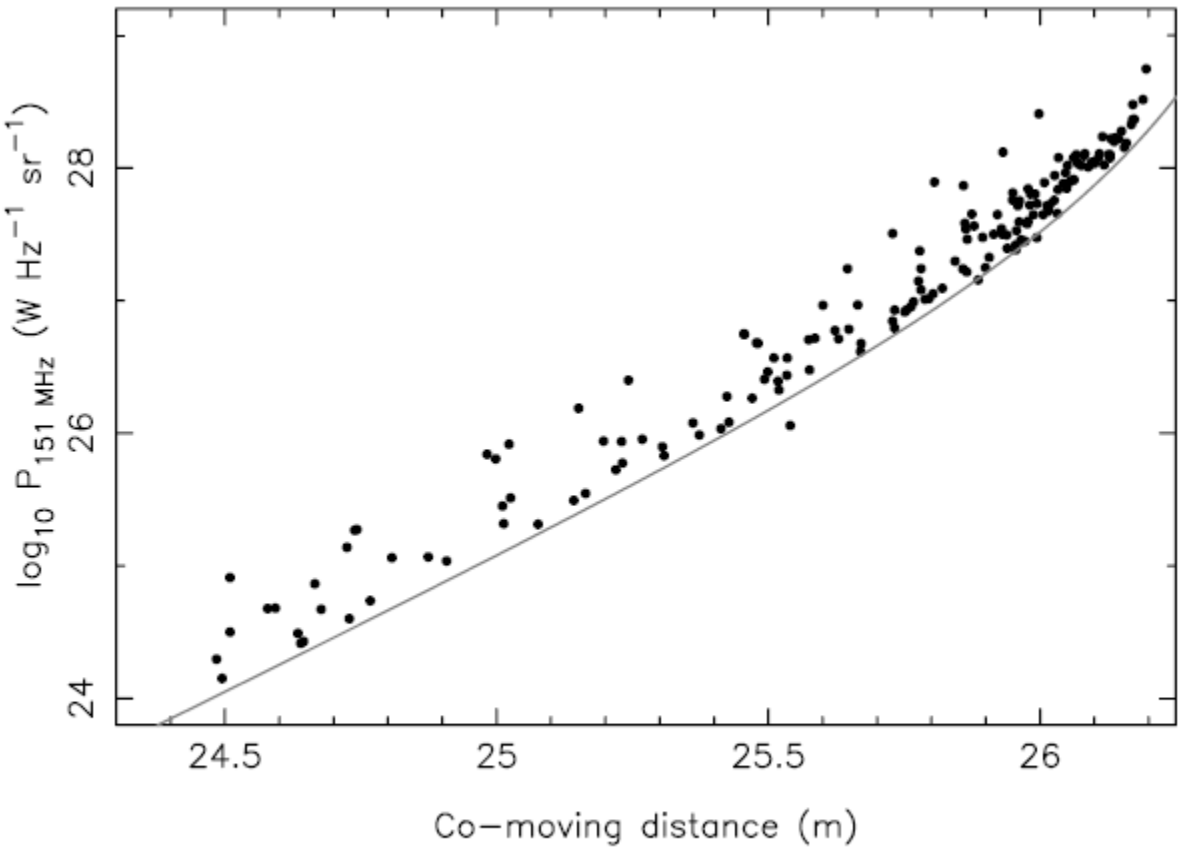
Common uses of statistics

- Measuring a quantity (parameter estimation)
  - Given the data, what is the best estimate of a particular parameter? What is the uncertainty in that estimate?
- Searching for correlations
  - Are two variables correlated, and is there an underlying physical mechanism?
- Testing a model
  - Given some data and a model, are the data consistent with the model? Which model best describes the data?

Correlating data

- When looking at new measurements, it is instinct to try to correlate it with other results
  - Checking if our measurements are reasonable
  - Checking if other results are reasonable
  - To test a hypothesis
  - Shot in the dark
- There are a few common traps to fall into when attempting to find correlations

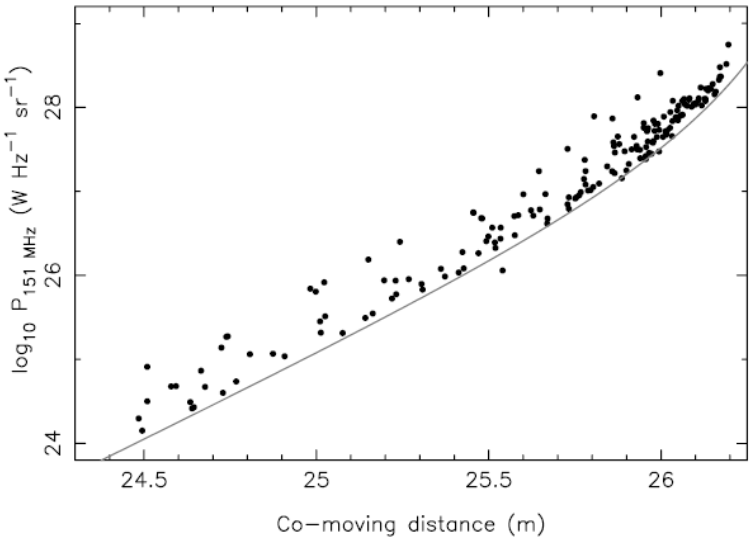
Fishing trips



Radio luminosities of 3CR radio sources versus distance modulus

Fishing trips

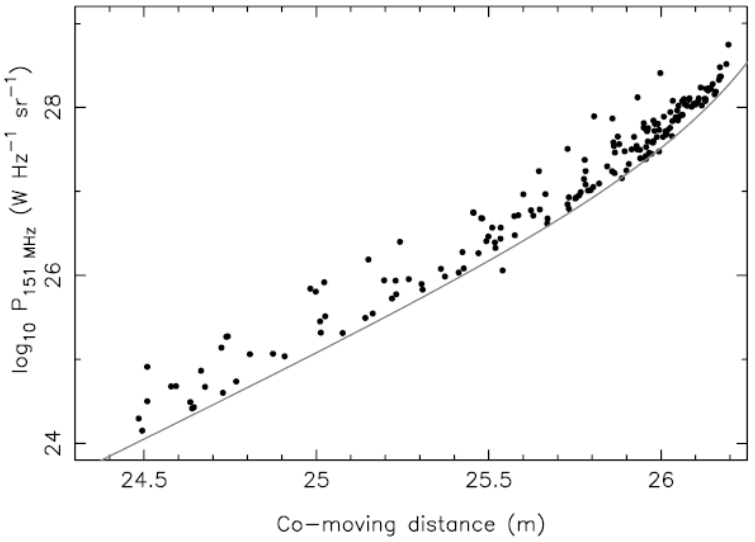
- Step 1: Is there a statistical correlation?
  - Do you see much correlation by eye?
- Step 2: Is the apparent correlation due to other effects?



Radio luminosities of 3CR radio sources versus distance modulus

Fishing trips

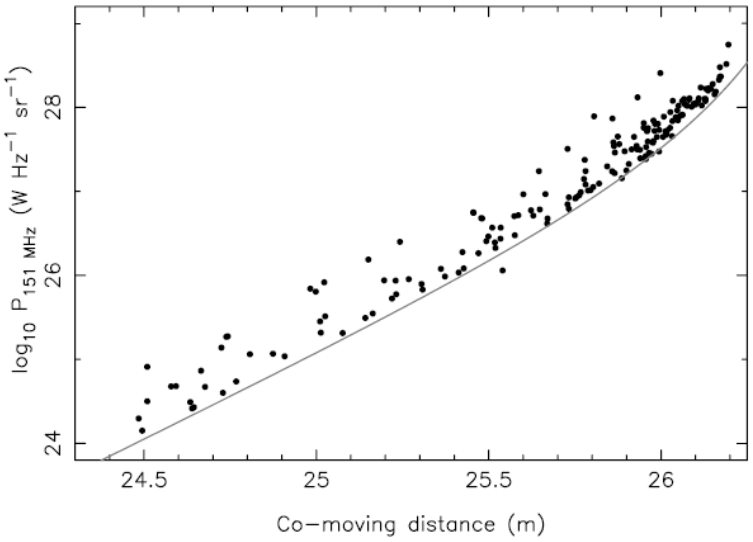
- Does this plot prove that more distant sources are more powerful?



Radio luminosities of 3CR radio sources versus distance modulus

Fishing trips

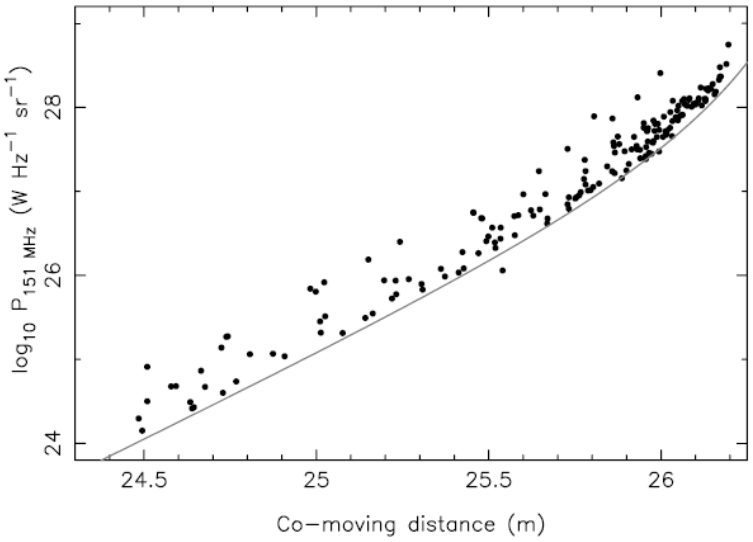
- The lower line is the flux-density limit of the dataset
  - Nothing can appear in the lower right area



Radio luminosities of 3CR radio sources versus distance modulus

Fishing trips

- Why is the upper left empty?
  - The density of brighter sources is lower, so we are biased to have more sources near the detection limit
  - The underlying probability function for luminosity is important and explains the apparent correlation



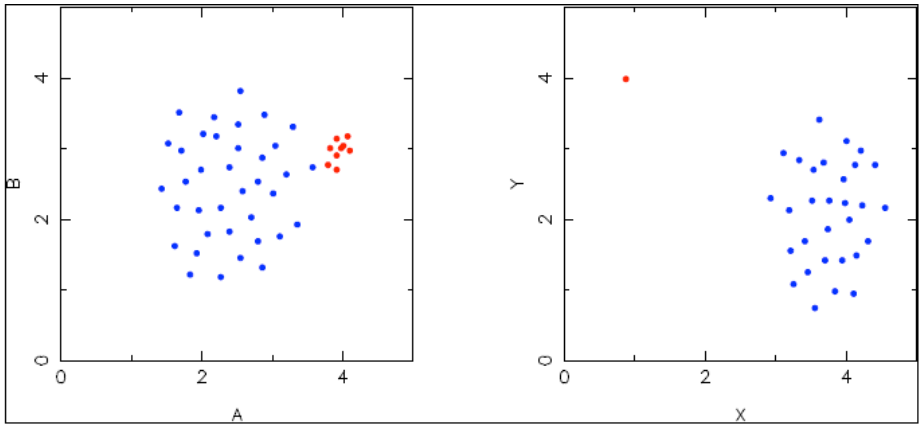
Radio luminosities of 3CR radio sources versus distance modulus

Fishing trips

- Step 3: Can calculate the significance of the correlation
  - Different from fitting to a model!

Fishing trips

- Step 4: Is the result realistic?
  - "Rule of thumb:" if 10% of the results are grouped so that covering them with your thumb destroys any apparent correlation, then go back to suspecting selection effects, data errors, or other statistical conspiracies

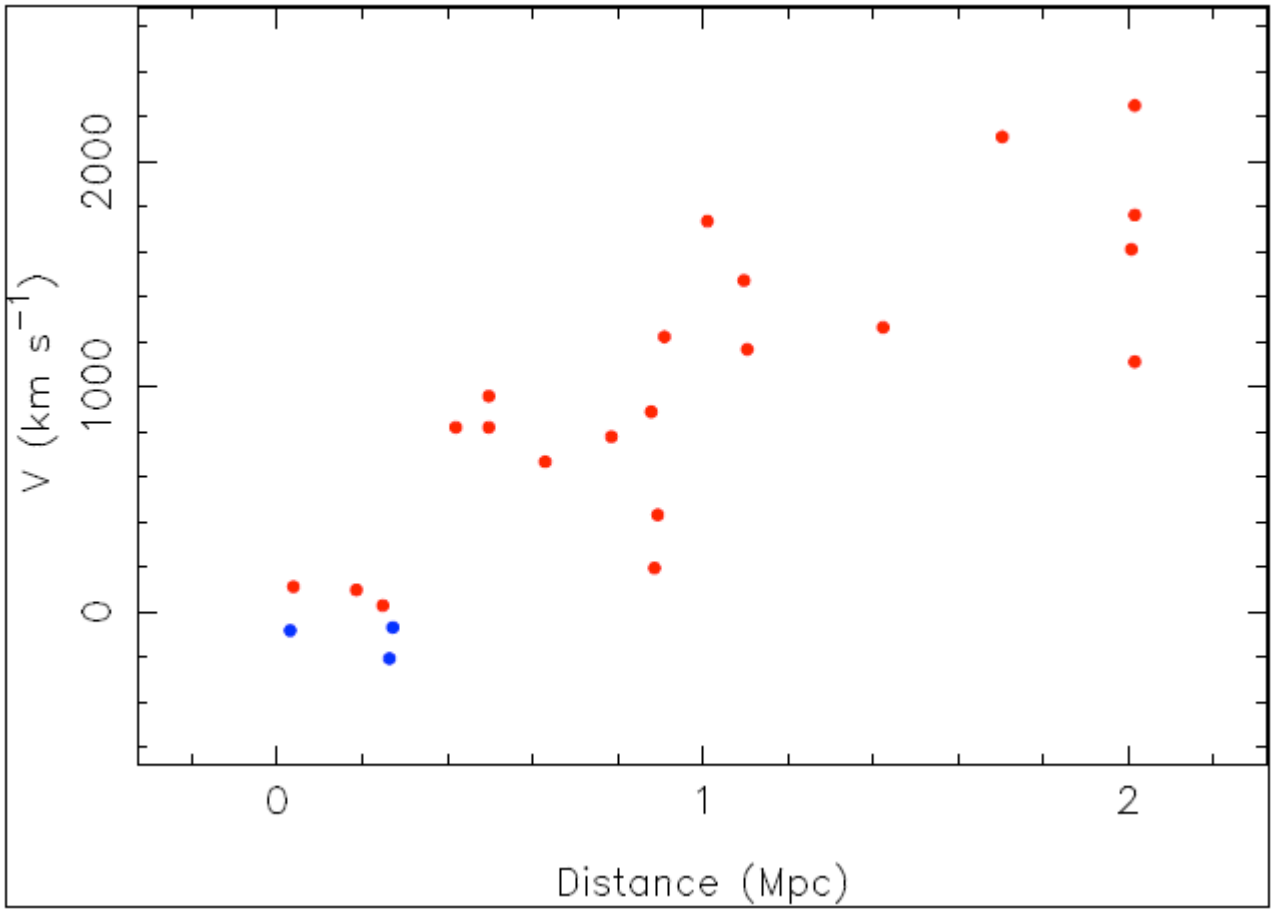


Suspect correlations: in each case formal calculation will indicate that a correlation exists to a high degree of significance.

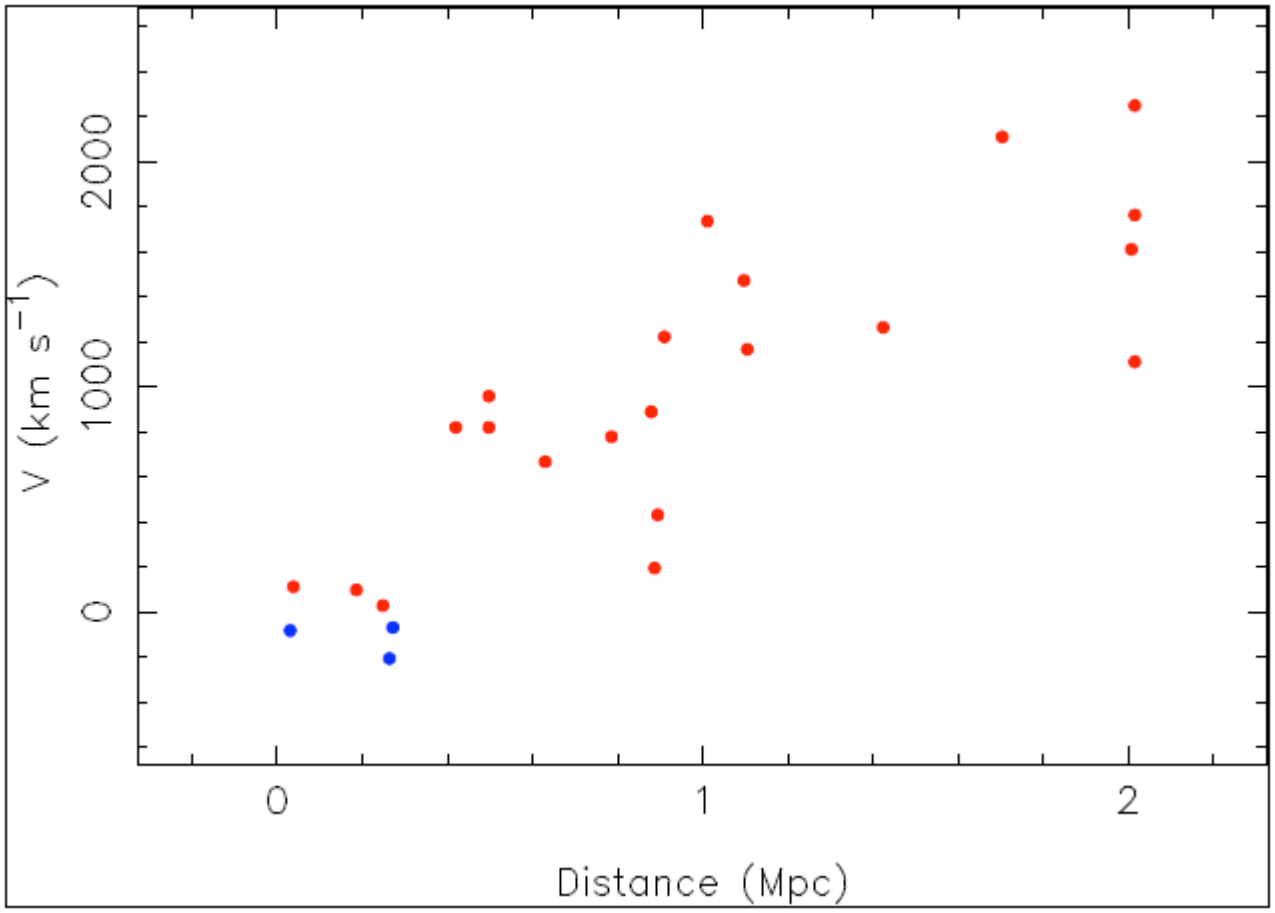
Fishing trips

- Correlation does not prove a causal connection!
  - Examples of correlations
    - Number of violent crimes in cities versus number of churches
    - The quality of student handwriting versus their height
    - Stock market prices and the sunspot cycle
    - Cigarette smoking vs lung cancer
    - Health vs alcohol intake
- Potential reasons
  - **Lurking third variables**
  - Similar time scales
  - Causal connection

Another fishing trip?



Another fishing trip?

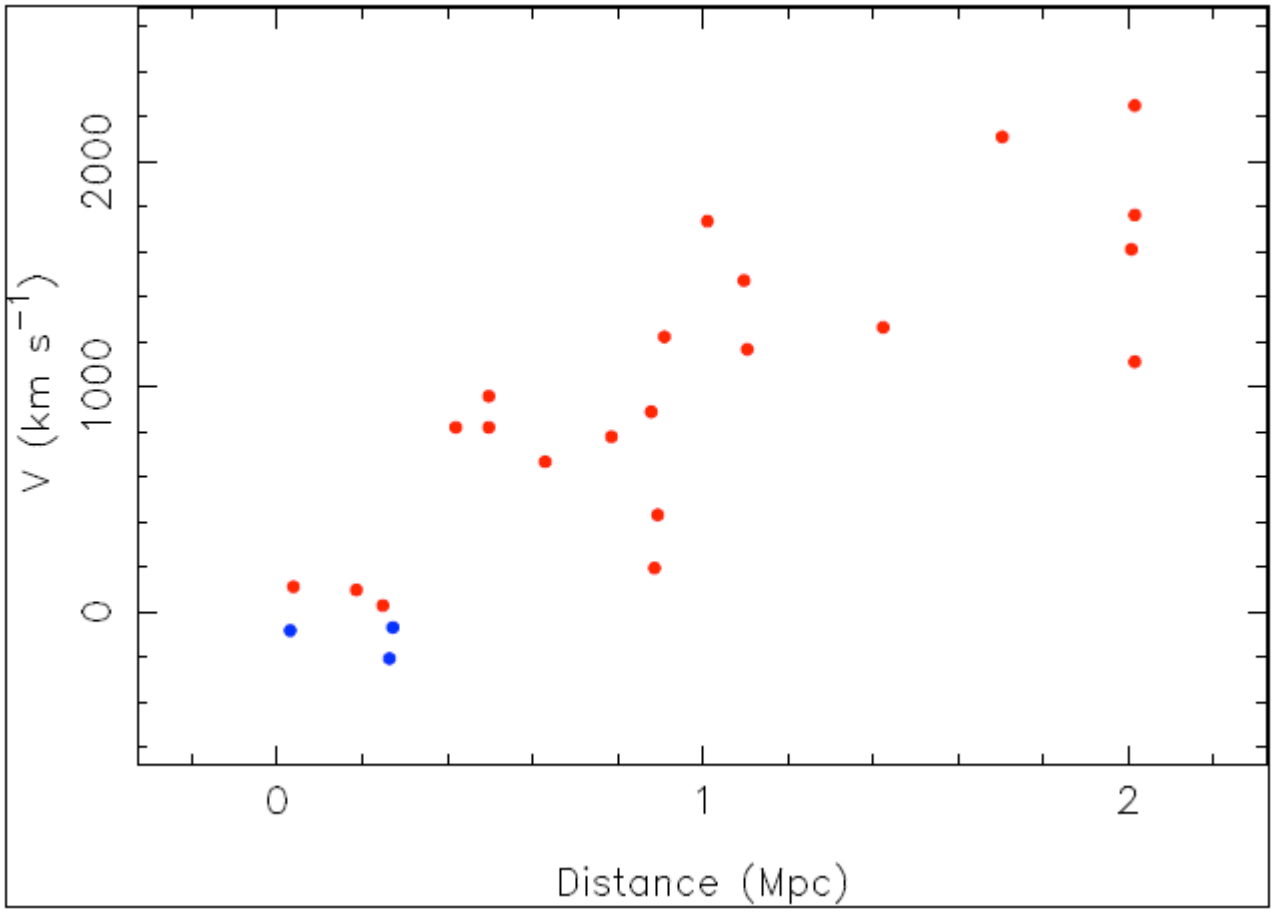


An early Hubble diagram, N=24 galaxies (1936)

So you think your data is correlated

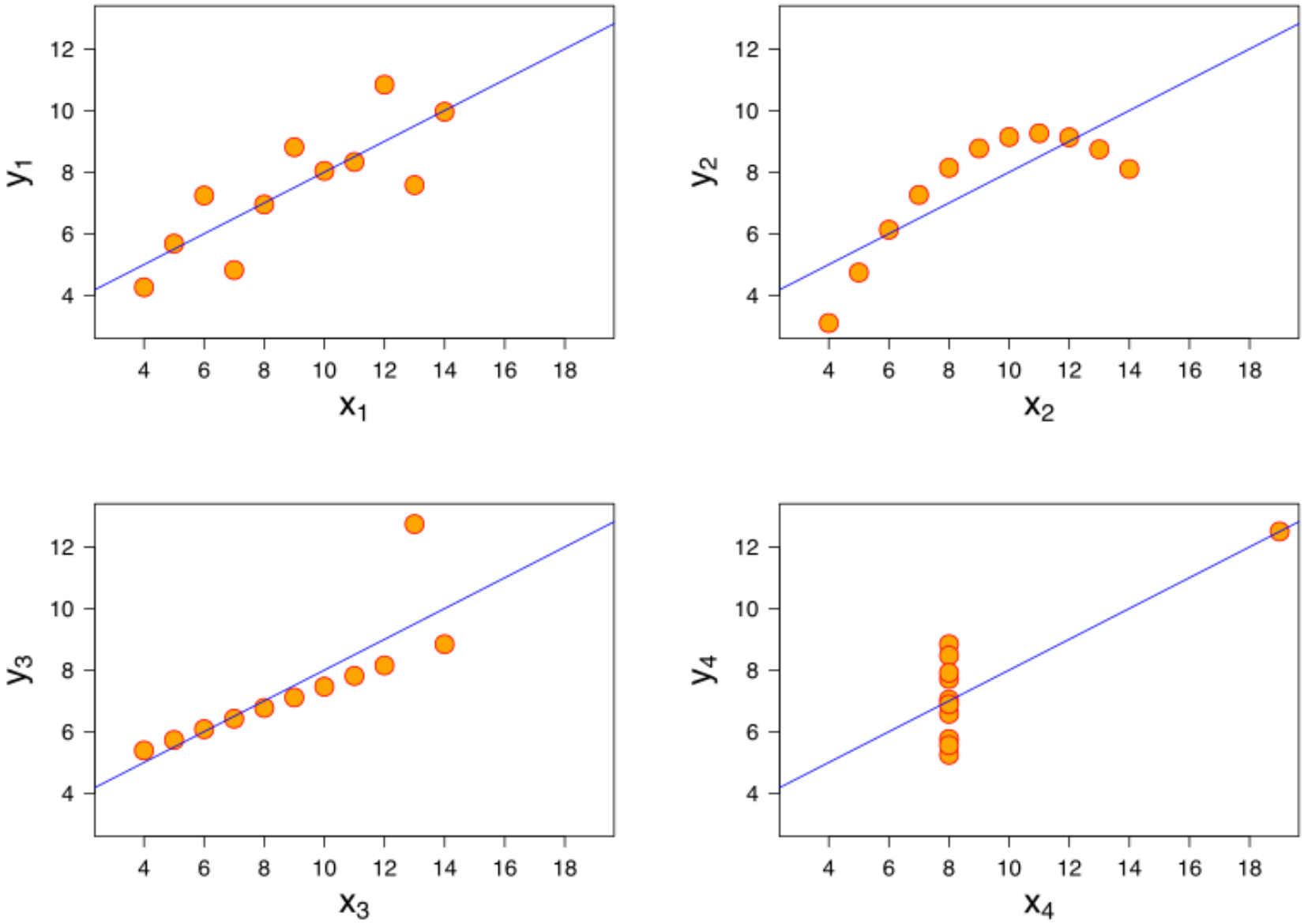
- Time to fit the data to a model!
- “All models are wrong, but some models are useful.”
  - George E. P. Box

Fitting data to a model



An early Hubble diagram, N=24 galaxies (1936)

Fitting data to a model

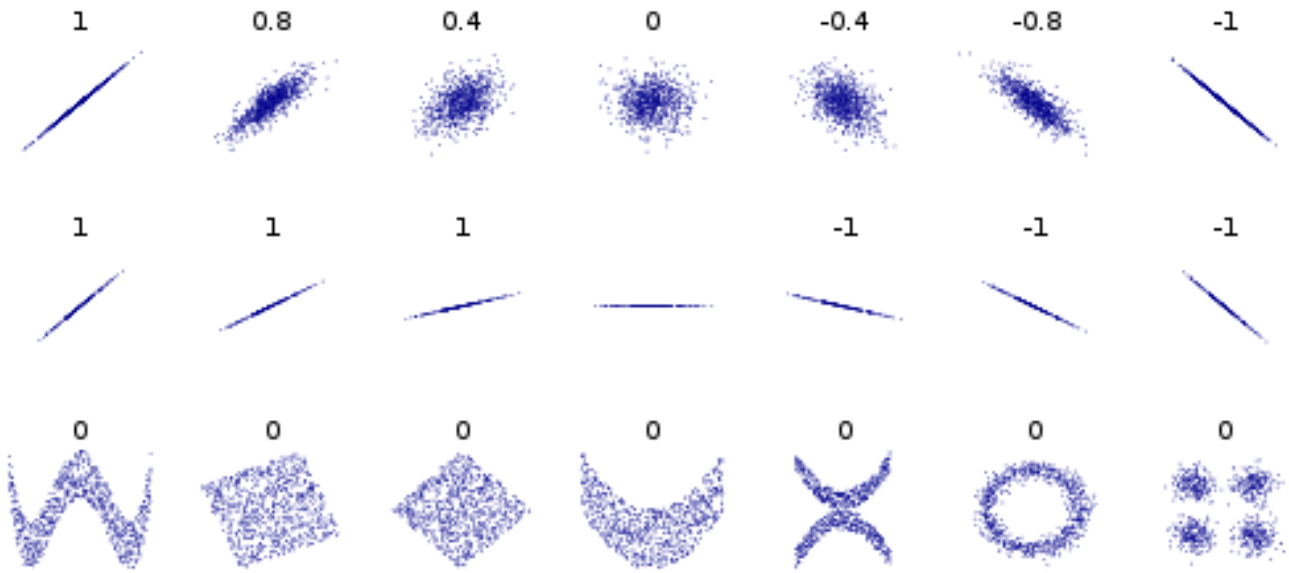


Simplifying correlations

- Linear correlation
  - y=mx+b
  - Multidimensional: z = mx + ny + b
- Linear correlations are easy to plot and examine
- Can linearize your data to make it a linear correlation
  - Example: Surface brightness of a disk
    - $I(r) = I_0 e^{-r/h}$
    - Linearized form: $\ln(I) = \ln(I_0) - r/h$
  - Also straightforward to do for power laws

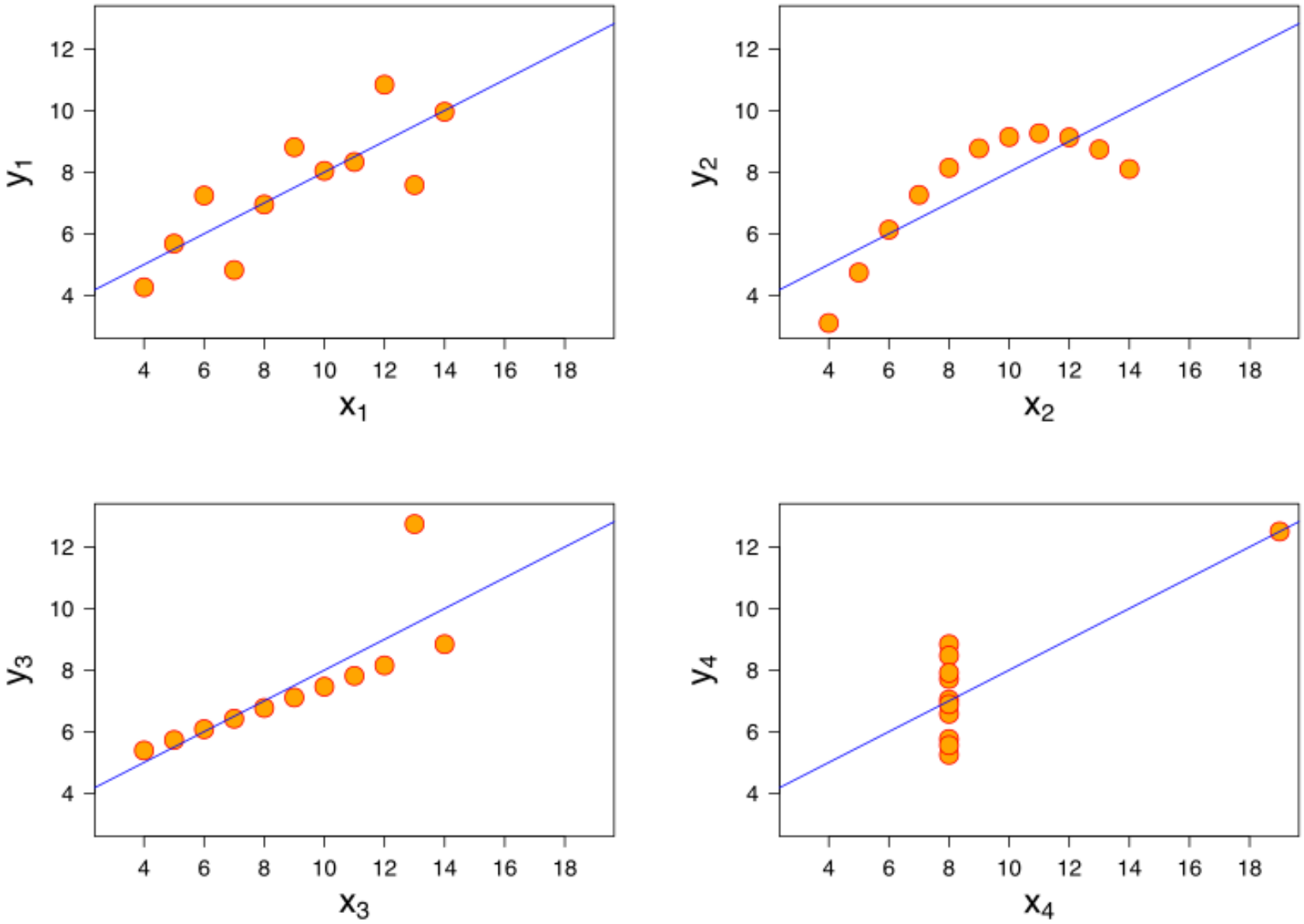
Pearson's correlation coefficient

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

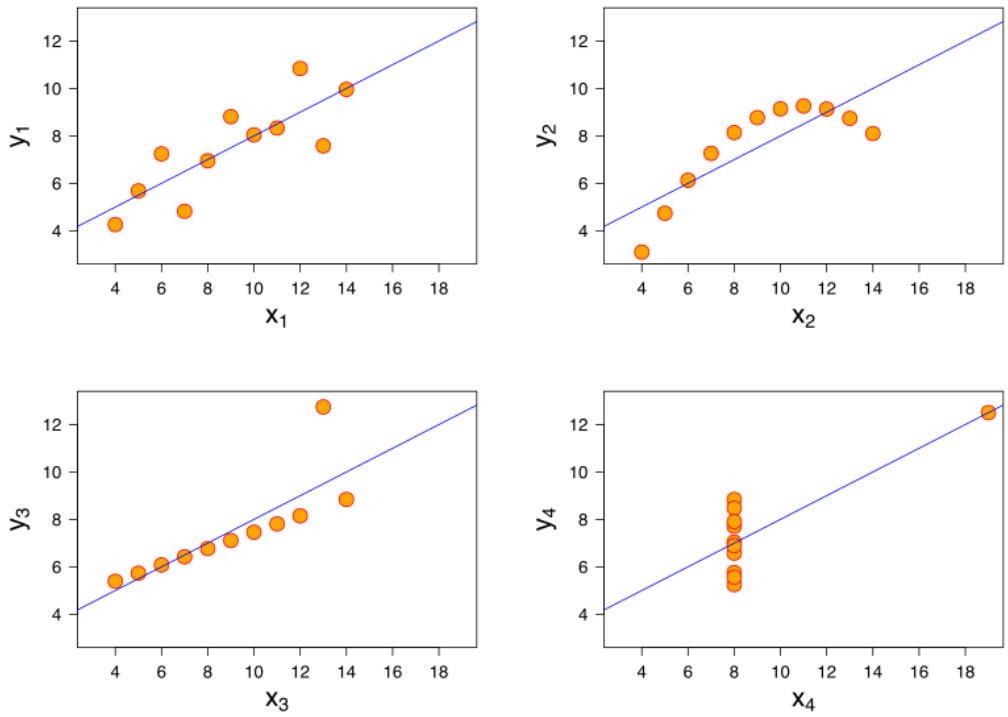


- Pearson's correlation coefficient, r, measures the linear correlation between two variables

Anscombe's quartet



Anscombe's quartet



Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

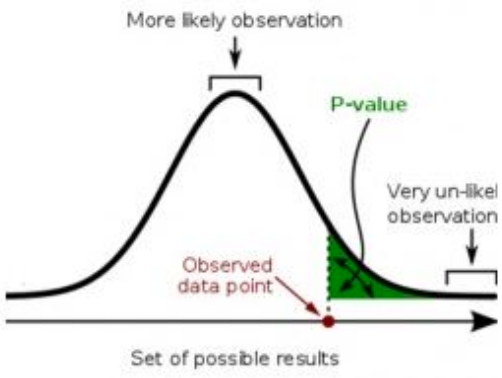
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Modeling uncertainty

- Least squares fitting
  - A straightforward output of Python/Matlab/Excel/etc
  - Assumes uncorrelated Gaussian statistics
  - Can get different results depending on the exact algorithm, especially for data with a small number of samples, or data with outliers
- Other ways to check uncertainty
  - Jack-knife
    - Go through data and toss out data points, and recalculate
    - Common to split data in half (e.g. first-half vs second-half)
  - Bootstrap
    - Go through dataset picking N points at random, recalculate and look at variation

Hypothesis testing and p-values

- **p-value** is the probability value, the probability of obtaining results as extreme as the results observed if the **null hypothesis** were true
- It does not tell you what the probability of the hypothesis is, given the data
  - "Given that someone is Catholic, what is the probability that they are the Pope?"
  - "Given that someone is the Pope, what is the probability that they are Catholic?"
- A small p-value can indicate strong evidence against the null hypothesis ($p < 0.05$)
- Can be manipulated: p-hacking, p-HARKing
- Statistics are a tool, can be mis-used for bad science!



https://www.acsh.org/news/2018/06/29/problem-p-values-13130

Bayesian estimation

- https://en.wikipedia.org/wiki/Bayes%27_theorem
- Bayes' theorem $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$
  - A = your dataset
  - B = the parameter you're trying to measure
  - P(B|A) The posterior probability
    - What is the probability of B, given that you've measured A? Your best estimate is the B that is most likely
  - P(A|B) The likelihood function
    - What is the probability of measuring A, given that model B is true?
  - P(B) The prior. What is the probability of B?
  - P(A) Normalizing factor. What is the probability that you could measure A to begin with?

Example of Bayesian estimation

- Suppose that a test for using a particular drug is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people in the general population are users of the drug. What is the probability that a randomly selected individual with a positive test is a drug user?

Example of Bayesian estimation

$$
\begin{aligned}
P(\text{User} \mid +) &= \frac{P(+ \mid \text{User})P(\text{User})}{P(+)} \\
&= \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})} \\
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\
&\approx 33.2\%
\end{aligned}
$$

Example of Bayesian estimation

- Even if an individual tests positive, it is more likely that they do not use the drug than that they do. This is because the number of non-users is large compared to the number of users. The number of false positives outweighs the number of true positives. For example, if 1000 individuals are tested, there are expected to be 995 non-users and 5 users. From the 995 non-users, $0.01 \times 995 \simeq 10$ false positives are expected. From the 5 users, $0.99 \times 5 \approx 5$ true positives are expected. Out of 15 positive results, only 5 are genuine.
- The importance of specificity in this example can be seen by calculating that even if sensitivity is raised to 100% and specificity remains at 99% then the probability of the person being a drug user only rises from 33.2% to 33.4%, but if the sensitivity is held at 99% and the specificity is increased to 99.5% then the probability of the person being a drug user rises to about 49.9%.

What to do with an astrophysics PhD

- https://www.wired.com/story/the-style-maven-astrophysicists-of-silicon-valley/