SFB-1491 Graduate School: Astrostatistics

Section 1: Describing Data

Dr Angus H Wright

2023-02-02

Course Outline

- **Section 1**: Data Description
- **Section 2**: Probability & Bayesian Statistics
- **Section 3**: Optimisation, Complex Analysis, Machine Learning

Section 1: Data Description and Selection

Topics include:

> - Point & interval estimation
> - Correlation & covariance
> - Selection biases in an Astronomical Context

Section 2: Probability & Bayesian Statistics

Topics include:

> - Fundamentals of probability
> - Statistical distributions and their origins
> - Frequentist & Bayesian statistics
> - Bayes theory
> - Prior specification
> - Hypothesis testing

Section 3: Optimisation, Complex

Modelling, and Machine Learning

Topics include:

> - Monte Carlo Simulation
> - Markov-Chain Monte Carlo
> - Fitting high dimensional data
> - Machine Learning

Learning Objectives for Today

Measures of Central Tendency and Dispersion

Understand the motivations, limitations, and useful properties of concepts such as:

> - **Arithmetic Mean**, **Median**, **Mode**
> - **Mean Absolute Deviation**, **Mean Squared Error**,
> - **Variance**, **Standard deviation**, **Median Absolute Deviation from Median**

Understand the differences between various measures of central tendency, and when they are appropriate to use.

Understand the differences between various measures of dispersion, and when they are appropriate to use.

Selection Bias in Astronomy

Correlation & Covariance

Selection effects in astronomical surveys

Notation and Nomenclature

Let us start with some definitions and a description of the notation that we will use throughout this course:

- An **observation** ω is any individual measurement that we have made.
- A **sample** $\{\omega_1, \omega_2, \omega_3, \dots, \omega_n\}$ is any collection/set of individual measurements.
- The number of observations in a sample is $|\omega| = n$ (called ‘cardinality’ in set theory/notation).
- A **population** Ω is the collection of all measurements.

Each observation ω generally has one-or-more **variable(s)** associated with it (otherwise there was no observation!), which we define using upper-case Roman letters (X, Y, Z) or sometimes using subscripts on X (i.e. $X_1, X_2, \dots, X_p$) if there are many variables to consider.

The observations of each variable are lower-case; e.g. $\{y_1, y_2, \dots, y_n\}$ are n observations of variable Y.

If a variable is known to follow a particular probabilistic distribution (we’ll use this a lot later in the course), then the variable is defined using a tilde (‘$\sim$’).

Samples vs Populations

The distinction between a *sample* and a *population* is an important one.

In standard experimental physics and astronomy, we are almost never presented with a population, because a population is the set of *all measurements*, in the most literal sense. Instead, we almost exclusively work with samples of observations, which we use to *infer* the properties of the population.

This leads to another important notation distinction. When discussing the properties of populations and samples, statisticians generally distinguish the former by using Greek letters, while the latter are given Roman letters. For example, the population mean and standard deviation are generally given the symbols μ and σ respectively, whereas the sample mean and standard deviation are generally given the symbols $\bar{x}$ and s respectively.

Finally, an estimator for an arbitrary true parameter (e.g. θ) is denoted by placing a hat/caret over the parameter (i.e. $\hat{\theta}$).

Putting it all together

I make a sample ω of n observations for a single variable X that is follows a normal (i.e. Gaussian) distribution with population mean μ and standard deviation σ.

$$\omega \in \Omega, \quad \text{where } |\omega| = \text{n}$$

We have n observations ω from the population Ω

$$X \sim N(\mu, \sigma)$$

Variable X is drawn from a Normal μ, σ distribution

$$X : \omega \mapsto \{x_1, x_2, x_3, \dots, x_n\} \quad \text{or}$$
$$X(\omega) = \{x_1, x_2, x_3, \dots, x_n\}$$

The values of X for each observation ω are $x_1, x_2, \dots, x_n$

With this sample of n observations we now want to do our science.

For this experiment, our science goal is simply to estimate the value of μ. We decide to define our estimate of μ as simply the mean of our observations of X:

$$\hat{\mu} \equiv \bar{x}$$

We compute the mean, submit the resulting estimate to Nature, and win a Nobel Prize. Great job everyone!

Measures of Central Tendency and Dispersion

Frequently in data analysis we are interested in comparing the properties of different samples of data across a range of variables.

In these circumstances it is generally advantageous to reduce distributions of data into one-point summary statistics.

Choice of *which* summary statistic to use, however, is often important.

Point Estimation: Arithmetic Mean

The natural starting point for a discussion on point estimates for an arbitrary variable is to discuss the arithmetic mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

This is the common “average” or “mean” with which we are all familiar.

Point Estimation: Median

The median is the point that divides a dataset into two equal parts.

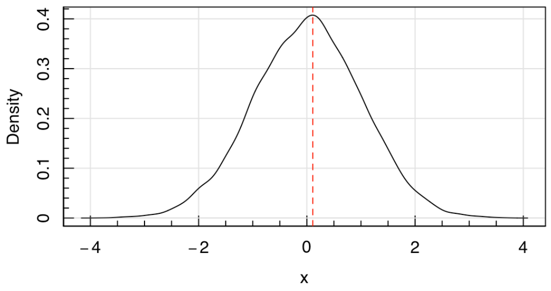
For data with an odd number of observations, this is trivially the middle (that is, the $[(n+1)/2]^{\text{th}}$) entry of the rank-ordered dataset. For even-numbered observations where there is no ‘middle’ value, the median is generally defined to be the mean of the two middle values.

$$\tilde{x}_{0.5} = \begin{cases} x_{[(n+1)/2]} & n \in 2\mathbb{Z} - 1 \\ (x_{[n/2]} + x_{[n/2+1]})/2 & n \in 2\mathbb{Z} \end{cases}$$

Point Estimation: Mode

The next frequently used point statistic is the mode, which is the most frequently observed data-point in the variable.

For continuous data, the mode is frequently estimated using a discretized or smoothed representation of the data, such as the KDE:



Dispersion Estimation

In addition to just an estimate of the central tendency of data, we often also require an estimate of the data dispersions/spread.

- Different distributions can have the same central tendency;
- When quantifying the possible range of a variable, a point estimate is obviously insufficient; and
- Even if we *do* just want a point estimate, that estimate of the central tendency will always be imperfect. Crucially, the uncertainty on it is intimately linked to the data dispersion.

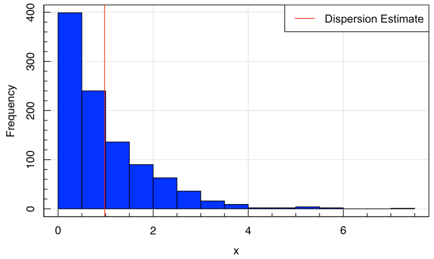
Deviation

Dispersion is a measure of deviation from a particular point. So we can construct an arbitrary dispersion metric as being, for example, the arithmetic mean of all deviations between the data and a point A:

$$D(A) = \frac{1}{n}\sum_{i=1}^{n}(x_i - A).$$

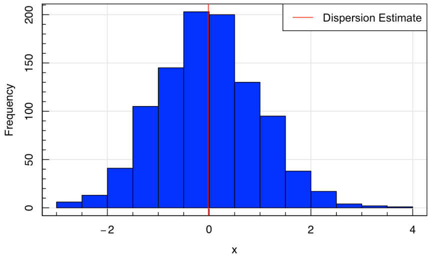
We can now run this dispersion metric for an arbitrary dataset:

```
## [1] "Summary of used sample:"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## 0.000376 0.274176 0.692038 0.977365 1.407133 7.391130 
## [1] "sd / MAD / 1-sig / 2-sig range:"
## [1] 0.9452439 0.7421323 0.8550172 1.6557027
## [1] "Using 1000 out of 1000"
```



So this dispersion measure looks sensible. Let’s try another dataset, which is Gaussian rather than Exponential:

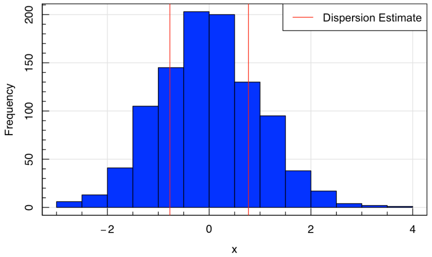
```
## [1] "Summary of used sample:"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -2.96120 -0.68077 -0.02265 -0.01725  0.60716  3.78325 
## [1] "sd / MAD / 1-sig / 2-sig range:"
## [1] 0.9800429 0.9654444 1.0009272 1.9805356
## [1] "Using 1000 out of 1000"
```



Absolute Deviation

To counter this effect we can instead use the **absolute deviation**:

$$D(A) = \frac{1}{n}\sum_{i=1}^{n} |x_i - A|.$$



This dispersion measure still carries with it the choice of A. It is intuitive to define the dispersion with respect to one of the point estimates that we've already discussed, such as the mean or median.

When we set our absolute deviation reference point to be the arithmetic mean of the distribution, $A = \bar{x}$, we recover the **absolute mean deviation**:

$$D(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|.$$

When we set reference point to the median $A = \tilde{x}_{0.5}$, we recover the **absolute median deviation**:

$$D(\tilde{x}_{0.5}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}_{0.5}|.$$

Variance & Standard Deviation

We can consider the arithmetic mean of the squares of the deviation, known as the **mean squared error (or MSE)** with respect to our reference point A:

$$s^2(A) = \frac{1}{n}\sum_{i=1}^{n} (x_i - A)^2.$$

When we set A to be the arithmetic mean $A = \bar{x}$, we recover the so-called **variance** of the sample.

$$\tilde{s}^2 \equiv s^2(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

The positive square root of the variance is called the **sample standard deviation**:

$$\tilde{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.}$$

The sample variance and standard deviation are fundamental quantities in statistics.

It can be shown that the sample variance is not an unbiased estimate of the population variance. Rather, an unbiased estimate of the population variance is:

$$\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Median Absolute Deviation from Median

As the standard deviation contains the arithmetic mean of the data, it can be sensitive to outlier values. As such it is sometimes preferable to use a dispersion estimator that has less sensitivity to outliers, such as the **median absolute deviation from the median (or MAD)**:

$$\mathrm{MAD}(x) = \mathrm{median}(|x_i - \tilde{x}_{0.5}|)$$

The MAD estimate is not a simple replacement for the standard deviation, though, because the two statistics aren't equivalent:

```
#Generate some Gaussian data
x<-rnorm(1e3)
#Create our MAD function
mad_disp<-function(x) {
  return(median(abs(x-median(x))))
}
sd(x); mad_disp(x);
```

```
## [1] 1.015025
```

```
## [1] 0.660692
```

However, it can be shown that the difference between the MAD and true standard deviation (for normally distributed data) is simply a multiplicative factor.

Therefore we can define the **normalised MAD (or nMAD)**:

$$\mathrm{nMAD}(x) = 1.4826 \times \mathrm{median}(|x_i - \tilde{x}_{0.5}|).$$

Useful Properties of Point and Dispersion Estimates

When using point estimates as summaries of data, it is useful to understand some fundamental properties of each statistic.

In each of the subcategories below we detail useful properties of each estimator, and important conceptual details about them.

The Mean

Recall that, for an arbitrary dataset of variables $\mathbf{x} = x_1, x_2, \ldots, x_n$, the mean is defined as:

$$\mathrm{mean}(x) \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Important Properties of the Mean

- Scaling the data scales the mean:

$$\mathrm{mean}(k\mathbf{x}) = k \times \mathrm{mean}(\mathbf{x})$$

- Translating the data also translates the mean:

$$\mathrm{mean}(\mathbf{x} + c) = \mathrm{mean}(\mathbf{x}) + c$$

- The sum of signed differences from the mean is zero:

$$\sum_{i=1}^{n}(x_i - \mathrm{mean}(\mathbf{x})) = 0$$

- The average squared distance between all data x_i and a single point μ is minimised at the mean:

$$\underset{\mu}{\mathrm{argmin}} \sum_{i=1}^{n}(x_i - \mu)^2 = \bar{x}$$

Proof

We can prove that the mean minimises the mean square distance to all data by finding the minima of the function:

$$\begin{aligned}\frac{\delta}{\delta\mu}\sum_{i=1}^{n}(x_i - \mu)^2 &= -2\sum_{i=1}^{n}(x_i - \mu)\\ &= 0\end{aligned}$$

so:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \mu) &= 0\\ \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu &= 0\\ \therefore \mu &= \frac{1}{n}\sum_{i=1}^{n} x_i\\ &\equiv \bar{x}\end{aligned}$$

The Standard Deviation

Recall that, for an arbitrary dataset of variables $\mathbf{x} = x_1, x_2, \ldots, x_n$, the sample standard deviation is defined as:

$$\begin{aligned}\mathrm{std}(\mathbf{x}) \equiv s &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\\ &= \sqrt{\mathrm{mean}(x_i - \bar{x})^2}.\end{aligned}$$

The unbiased estimator of the population standard deviation is:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Important Properties of the Standard Deviation

- Translating the data does not change the standard deviation:

$$\text{std}(\mathbf{x} + C) = \text{std}(\mathbf{x})$$

> - Scaling the data scales the standard deviation:

$$\text{std}(k\mathbf{x}) = k \times \text{std}(\mathbf{x})$$

> - For n observations of an arbitrary variable x, whose standard deviation is s, there are at most $\frac{n}{k^2}$ data points lying k or more standard deviations away from the mean.

Assume we construct a length n dataset of variable y with m data that are k standard deviations from the mean. The fraction of data beyond k standard deviations is $r = m/n$. Furthermore, let's assume $\bar{y} = 0$ (which is fine, because of the translation point above). Therefore:

$$\text{std}(\mathbf{y}) \equiv s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}$$

Let's now make our dataset as pathological as possible. To start, we'll assign $n - m$ data points to have $y_0 = 0$, because these contribute 0 to the standard deviation. We'll then place the other m elements at exactly k standard deviations from 0; $|y_1| = ks_y$,. For this very strange dataset, the standard deviation becomes:

$$\begin{aligned} s_y &= \sqrt{\frac{1}{n}\sum_{i=1}^{m} y_1^2 + \sum_{i=m+1}^{n} y_0^2} \\ &= \sqrt{\frac{mk^2 s_y^2}{n}} \\ &= \sqrt{rk^2 s_y^2} \end{aligned}$$

so:

$$\begin{aligned} s_y^2 &= rk^2 s_y^2 \\ \therefore r &= \frac{1}{k^2} \end{aligned}$$

As this was the most pathological dataset possible, we therefore conclude that *for any dataset*, the maximal fraction of data that can sit k standard deviations away from the mean is $r = k^{-2}$.

> - For any dataset, there must be at least one data point more than one standard deviation from the mean.

Given the formula for the standard deviation:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where again we can generalise to an arbitrary dataset with $\bar{x} = 0$:

$$s_0 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

So

$$n \times s^2 = \sum_{i=1}^{n} x_i^2$$

The right hand side here is the sum of all squared deviations from the mean. However:

$$\sum_{i=1}^{n} x_i^2 \leq n \times \max(x_i^2).$$

That is, the sum of all deviations must be less than or equal to n times the maximal squared deviation. Therefore:

$$\begin{aligned} n \times s^2 &\leq n \times \max(x_i^2). \\ s^2 &\leq \max(x_i^2). \end{aligned}$$

So there must be at least 1 data value that is greater than or equal to the standard deviation.

Comment: Usefulness of the variance

If

$$Z = X + Y$$

then

$$s_Z^2 = s_X^2 + s_Y^2$$

```
## [1] 5.002786
```

The Median

Recall that, for an arbitrary dataset of variables $\mathbf{x} = x_1, x_2, \ldots, x_n$, the median is defined as:

$$\text{med}(\mathbf{x}) \equiv \tilde{x}_{0.5} = \begin{cases} x_{[(n+1)/2]} & n \in 2\mathbb{Z} - 1 \\ (x_{[n/2]} + x_{[n/2+1]})/2 & n \in 2\mathbb{Z} \end{cases}$$

Important Properties of the Median

- Scaling the data scales the median:

$$\text{med}(k\mathbf{x}) = k \times \text{med}(\mathbf{x})$$

- Translating the data also translates the median:

$$\text{med}(\mathbf{x} + c) = \text{med}(\mathbf{x}) + c$$

The nMAD

Recall that, for an arbitrary dataset of variables $\mathbf{x} = x_1, x_2, \dots, x_n$, the normalised median absolute deviation from median (nMAD) is defined as:

$$\text{nMAD}(x) = 1.4826 \times \text{med}(|x_i - \tilde{x}_{0.5}|).$$

Important Properties of the nMAD

- Translating the data does not change the nMAD:

$$\text{nMAD}(\mathbf{x} + C) = \text{nMAD}(\mathbf{x})$$

- Scaling the data scales the nMAD:

$$\text{nMAD}(k\mathbf{x}) = k \times \text{nMAD}(\mathbf{x})$$

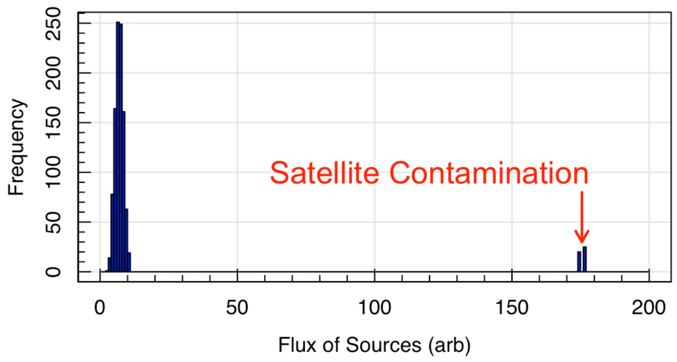
Comparing the statistics

How do these different statistics compare? What makes one more useful than another for a particular dataset?

Central Tendency

The primary pitfall with measures of central tendency come from the presence of outlier data. Given that the mean minimises the average distance to *all* data, the presence of outliers in a dataset can catastrophically bias the statistic.

Consider the case of galaxy images contaminated by satellite trails.



If we compute the point statistics of central tendency for this dataset:

```
## [1] 14.26484
```

```
## [1] 7.058896
```

The presence of the satellite contamination completely ruins our mean estimate of the galaxy fluxes. However, the median statistic doesn't fall into the same trap.

Dispersion

The story gets even worse when we want to calculate the dispersion statistics:

```
#Dispersion for gals
sd(gals_second); mad(gals_second); IQR(gals_second)
```

```
## [1] 34.37872
```

```
## [1] 1.563354
```

```
## [1] 2.102616
```

The catastrophic failure of the standard deviation here is a combination of the inclusion of the problematic mean estimate and the requirement that standard deviations have few data-points at many standard deviations from the mean.

Recall our formula for the fraction of data that can reside k standard deviations from the mean: $r \leq \frac{1}{k^2}$. In this dataset, the outliers make up 4.31$\%$ of the dataset. Therefore, they can reside ***at most*** 4.82 standard deviations from the mean.

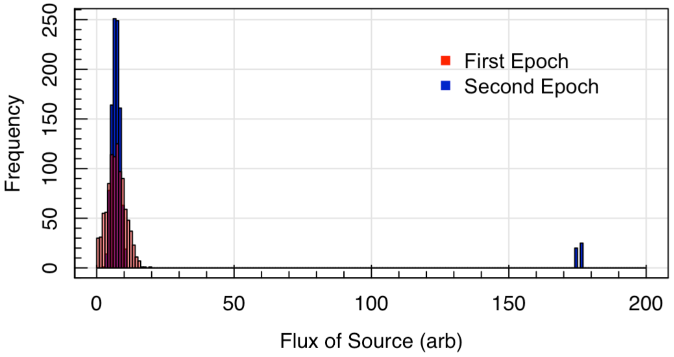
In reality, though, the outliers here *aren't* drawn from the same Gaussian distribution as the rest of the data, and in truth reside 112.67 standard deviations from the mean.

Crucially, the nMAD statistic is robust to the outliers, as it uses median statistics in its computation.

Comparing multiple datasets

Archive imaging of the same galaxy population exists, and was taken 15 years prior to our contaminated imaging. Satellites were less common then, and none of the images were affected. But the image quality is generally worse.

The Archive results are below.



The results of these two surveys, ignoring the outliers, are identical. However if we were to summarise these data only using mean statistics…

```
#Central Tendency statistics for two gals studies
mean(gals_first); mean(gals_second);
```

```
## [1] 6.989466
```

```
## [1] 14.26484
```

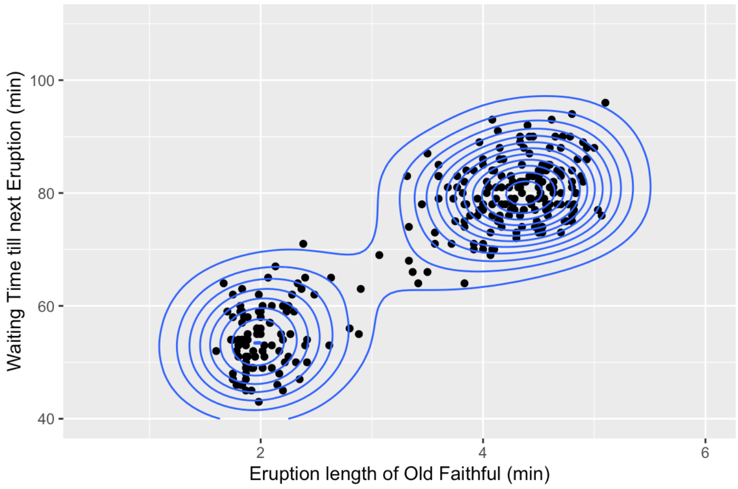
…then we would be forced to draw the conclusion that the samples have physically brightened by a factor of two between the two observations!

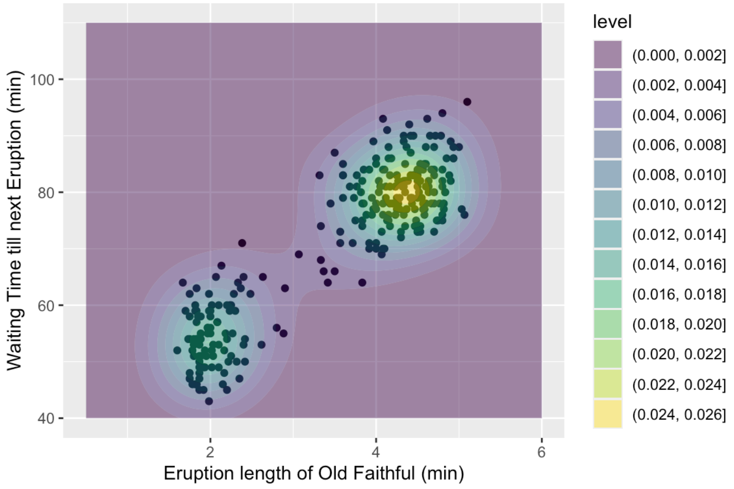
Summarising relationships in 2D

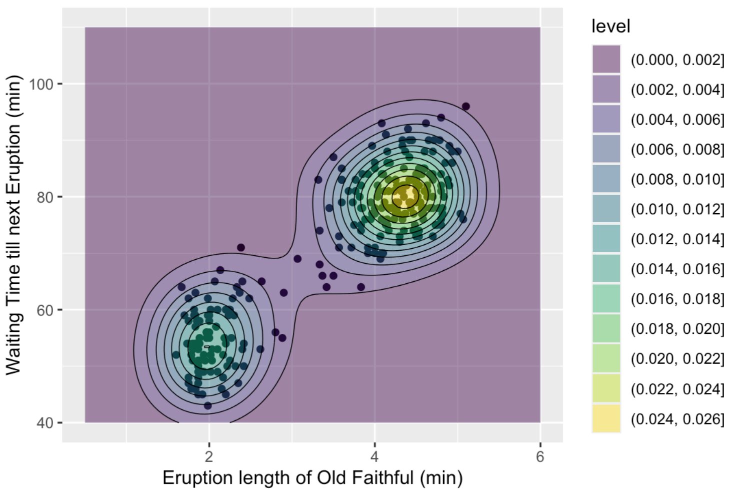
Until now, we have essentially explored datasets with only one variable (and how to compare different sets of observations of this one variable).

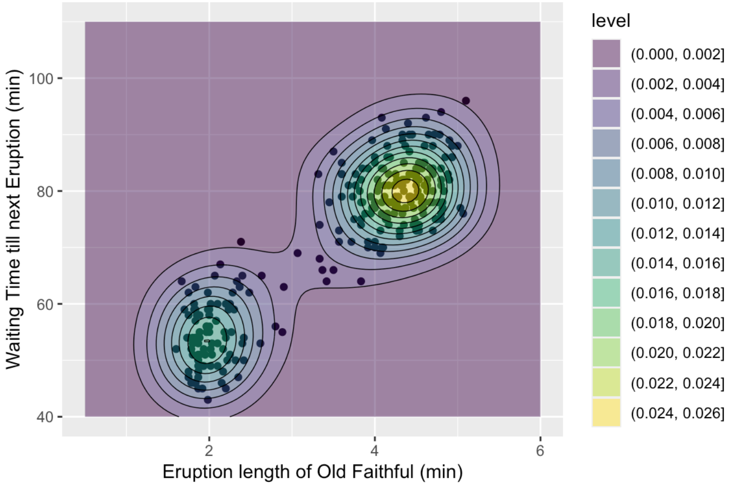
We now want to extend our analysis to datasets that contain two (or more) variables.

When provided with datasets containing multiple dimensions, we are frequently interested in determining relationships between variables.









There is a clear relationship between the duration of the eruption and the time until the next eruption. This relationship may have an underlying physical cause that we are interested in, or it may be coincidental. Determining the relationship between variables, and their significance, is therefore an important topic in statistics.

Covariance and Correlation

We've previously explored the concept of variance and standard deviation. For a single variable, recall that the variance was defined as:

$$\begin{aligned}\tilde{s}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \text{mean}[x_i - \text{mean}(x_i)]^2\end{aligned}$$

The **covariance** of two variables is then defined as the joint variance between each variable:

$$\text{cov}(X, Y) = \text{mean}[(x_i - \text{mean}(x_i)) \times (y_i - \text{mean}(y_i))]$$

We will discuss the covariance formula more later in the course, but for now you can see that the definition formally makes sense if you compute the covariance of a variable with itself:

$$\begin{aligned}\text{cov}(X, X) &= \text{mean}[(x_i - \text{mean}(x_i)) \times (x_i - \text{mean}(x_i))] \\ &= \text{mean}[(x_i - \text{mean}(x_i))^2] \\ &\equiv \tilde{s}^2\end{aligned}$$

The covariance of two variables describes the degree of joint variation that exists between two variables.

For our "faithful" dataset, we find that the covariance is 13.9778078. This value, though, is dependent on the absolute dispersion of the dataset.

That is, if we were to convert the faithful dataset into standard coordinates, the covariance changes: 0.9008112.

Pearson Correlation

It is therefore often useful to compute the amount of **correlation** between variables, that is invariate under scaling of the variables. For this we can compute the so-called **Pearson correlation coefficient**:

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{s(X)s(Y)}$$

The correlation coefficient varies between -1 (for perfectly negatively correlated data), and 1 (for perfectly positively correlated data). For our faithful dataset, we have:

```
#Compute ourselves
with(faithful,
  cov(eruptions,waiting)/(sd(eruptions)*sd(waiting))
)
```

```
## [1] 0.9008112
```

```
#Use the internal function
cor(faithful$eruptions, faithful$waiting)
```

```
## [1] 0.9008112
```

The covariance and correlation values are useful for computing the relationships between any two variables.

Covariance & Correlation Matricies

For datasets with two or more variables, the covariance can be computed for all combinations of different variable combinations, to create the **covariance matrix** and **correlation matrix**:

$$\mathrm{cov}(\Omega) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \dots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \dots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \dots & \mathrm{var}(X_n) \end{pmatrix}$$

$$\mathrm{cor}(\Omega) = \begin{pmatrix} 1 & \mathrm{cor}(X_1, X_2) & \dots & \mathrm{cor}(X_1, X_n) \\ \mathrm{cor}(X_2, X_1) & 1 & \dots & \mathrm{cor}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cor}(X_n, X_1) & \mathrm{cor}(X_n, X_2) & \dots & 1 \end{pmatrix}$$

Covariance & Correlation Matricies

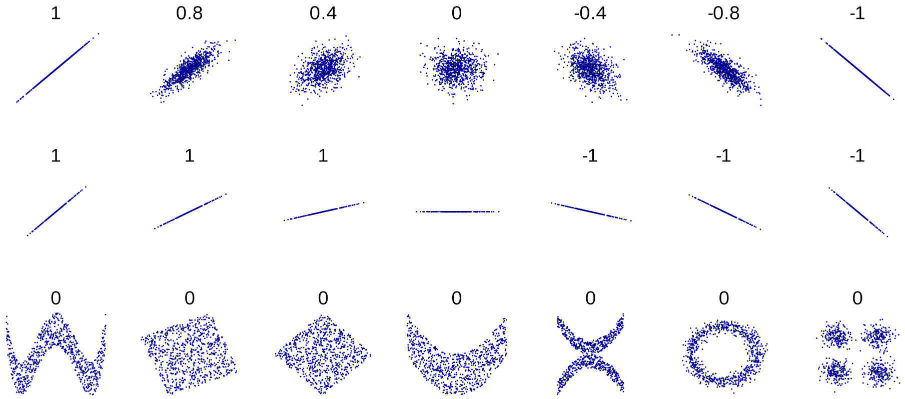
We can compute these matrices for our faithful dataset:

```
cov(faithful); cor(faithful)
```

```
##           eruptions   waiting
## eruptions  1.302728  13.97781
## waiting   13.977808 184.82331
```

```
##           eruptions   waiting
## eruptions 1.0000000 0.9008112
## waiting   0.9008112 1.0000000
```

Pearson correlation, however, should be used with caution. For linear data, the coefficient is sensible. However for strongly non-linear data the coefficient is less interpretable:

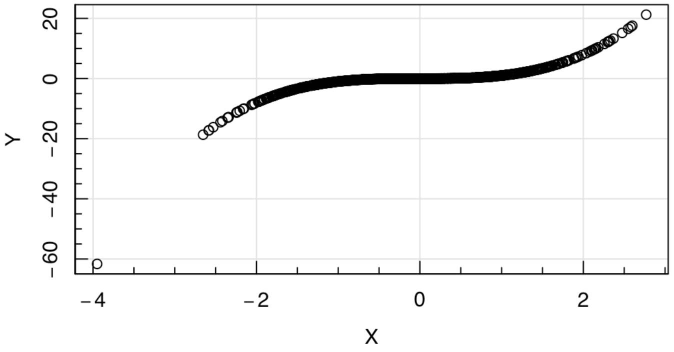


Spearman Correlation

There are other correlation formalisms that attempt to circumvent the problems of the Pearson correlation coefficient is to utilise an associated measure called the **Spearman Rank Correlation**.

The Spearman Rank Correlation is defined as the Pearson correlation between the *rank-orders of the variables*.

As a demonstration, we'll construct a dataset with one non-linear variable, but which is exactly correlated to another (in the intuitive sense: knowing one perfectly informs the other).



```
##           X         Y
## X 1.0000000 0.7647762
## Y 0.7647762 1.0000000
```

So Pearson tells us that the variables are correlated at the $\sim 70\%$ level, while a quick look at our figure shows us that this is clearly an underestimate.

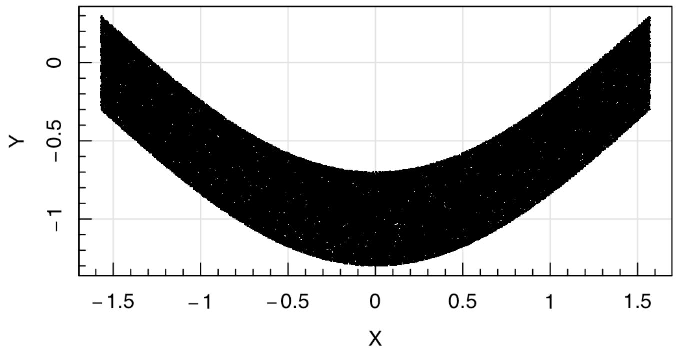
If we compute the pearson correlation of the rank-ordered variables, though:

```
## [1] 1
```

This makes sense intuitively, because the function Y is monotonically increasing and perfectly correlated to X (albeit non-linearly). This is the Spearman Rank correlation:

```
##   X Y
## X 1 1
## Y 1 1
```

Let's now see if the Spearman correlation can recover the correlation of one of our We can now look at one of the strange Pearson results from our figure:



```
##               X             Y
## X  1.0000000000 -0.0001742265
## Y -0.0001742265  1.0000000000
```

```
##             X           Y
## X 1.000000000 0.001351952
## Y 0.001351952 1.000000000
```

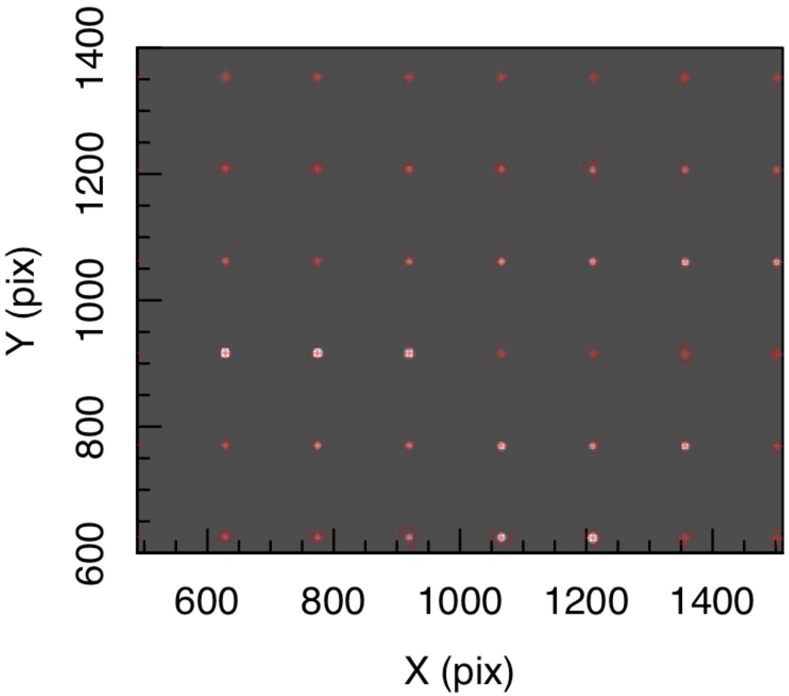
Notice that the rank correlation is unable to recover the correlation between non-monotonically increasing variables!

In this way, the correlation coefficients are describing the joint information between two variables.

Interpretting Correlation

For parts of this lecture we're going to work with various simulated **toy universes** that I have constructed.

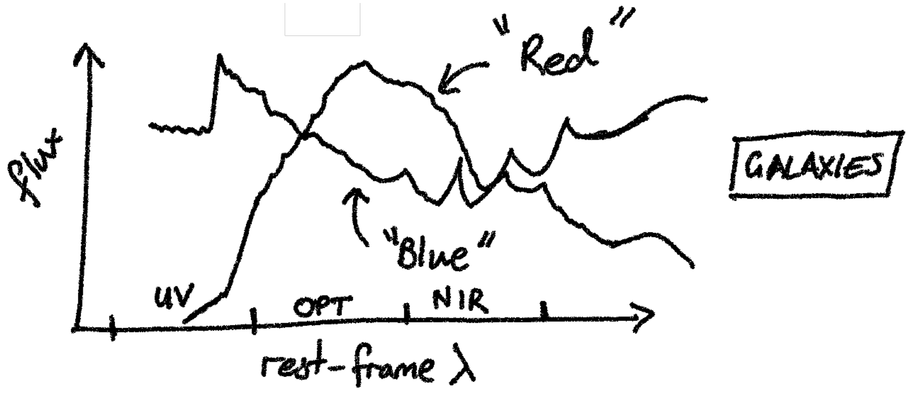
Galaxies in the first toy universe are simple and easy to model, the universe is a perfect expanding lattice, and telescopes and detectors have infinite sensitivity



> “Theorist_Universe_1”

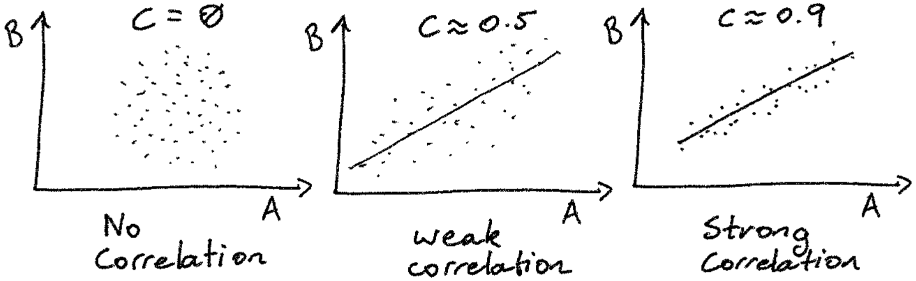
To start our exploration into my mock universe, we're going to look at the distribution of galaxies in my universe.

Galaxies are observed with telescopes and fluxes are measured in various filters. We use these fluxes to measure galaxy redshifts and galaxy properties using “Spectral Energy Distribution” modelling.

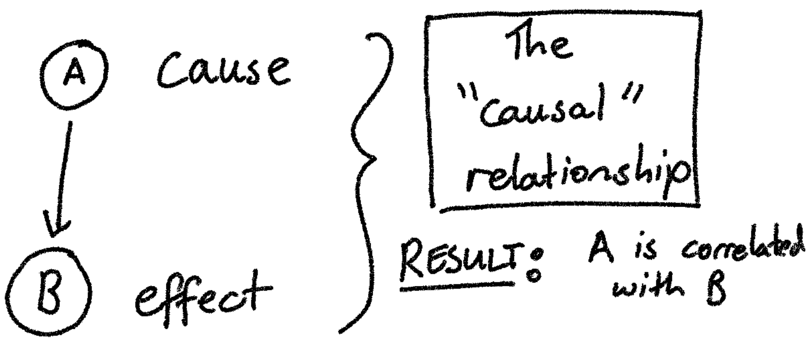


Galaxies have a number of physical parameters estimated, and a number of additional properties are included (e.g. environment).

Correlation 1: Relationships between parameters



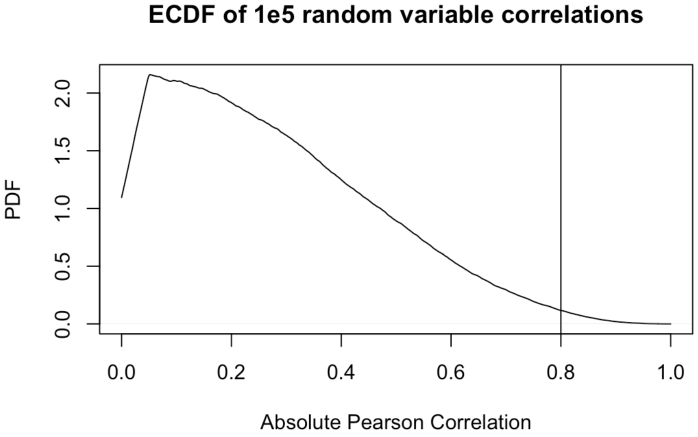
If one parameter is related to another, we are generally interested in determining if that relationship is **causal**.



A sneak peak at probability

Let's say that in our toy universe we have a catalogue of 10 galaxies, with 200 properties measured for each galaxy. We decide a relationship is worth investigating if it contains an 80% correlation or more.

What fraction of our properties do we expect to have a correlation of 80% or more, *assuming* that the data are all truly random?



```
## 0.64 % of variables have 80% correlation or more
```

Said differently, there is a 1 in 156 chance that two totally random variables in our survey will have an absolute correlation of 0.8 or higher.

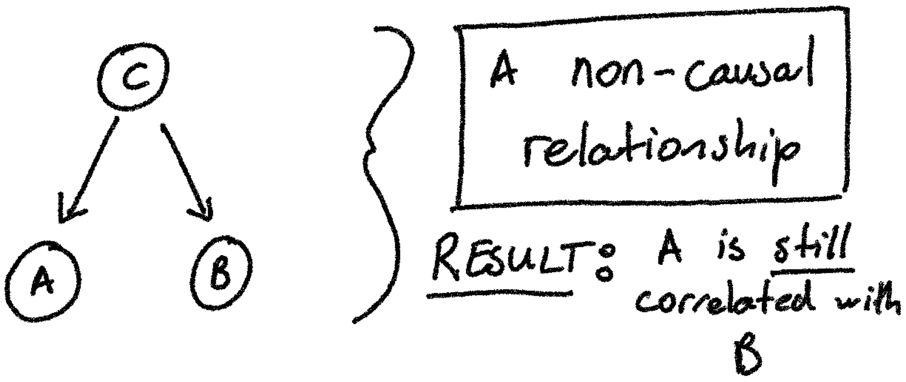
What does this mean?

The likelihood of finding “significant” correlations between truly random data is non-zero, and grows with decreasing numbers of observations and increasing numbers of observed variables.

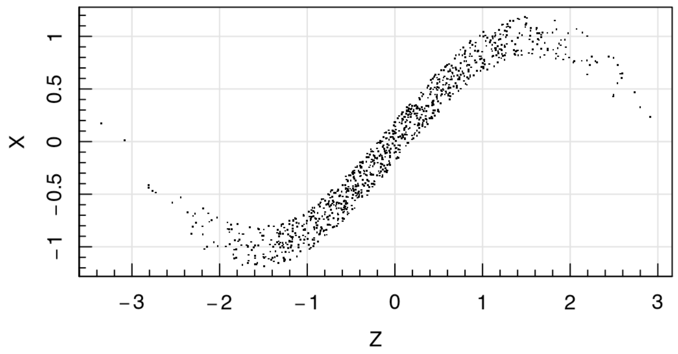
> **Working with small samples does not mean you can ignore statistics**

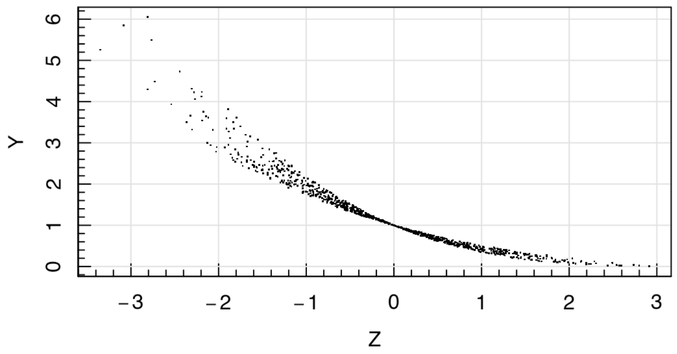
Making matters worse

The problem is further complicated by the existence of **confounding variables**.

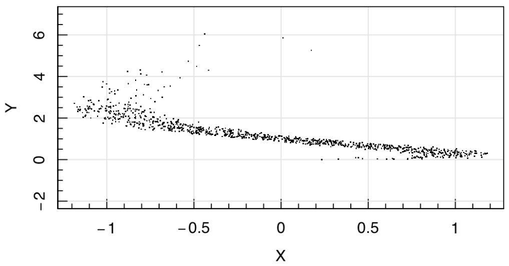

C
A
B
A non-causal relationship
RESULT: A is still correlated with B

A confounding variable is one that acts upon both the dependent and independent variables in a measurement of correlation, and thereby creates a spurious correlation between the two.





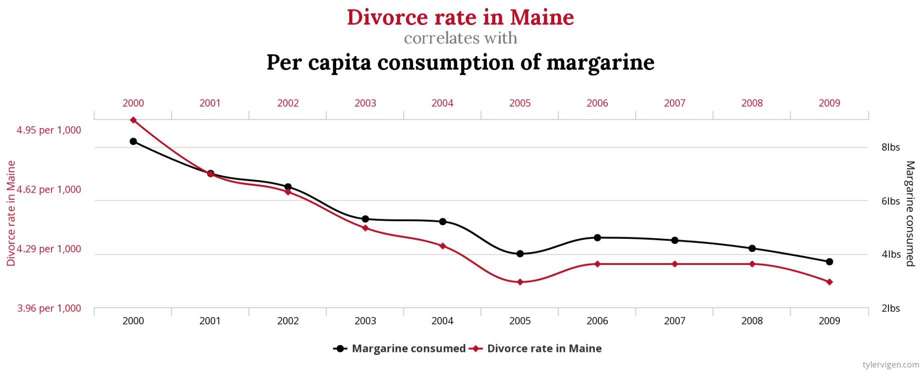
We’ve created two variables that correlate with Z. But what if we never actually *observed* the variable Z… We would instead plot X and Y:



```
## [1] -0.9645978
```

And be tempted to decide that there is a **causal** relationship between these two parameters, when in fact none exists.

A fundamental distinction



This is an example of a spurious correlation. Such correlations are possible (and indeed likely!) when you have few observations of many variables (more on this later).

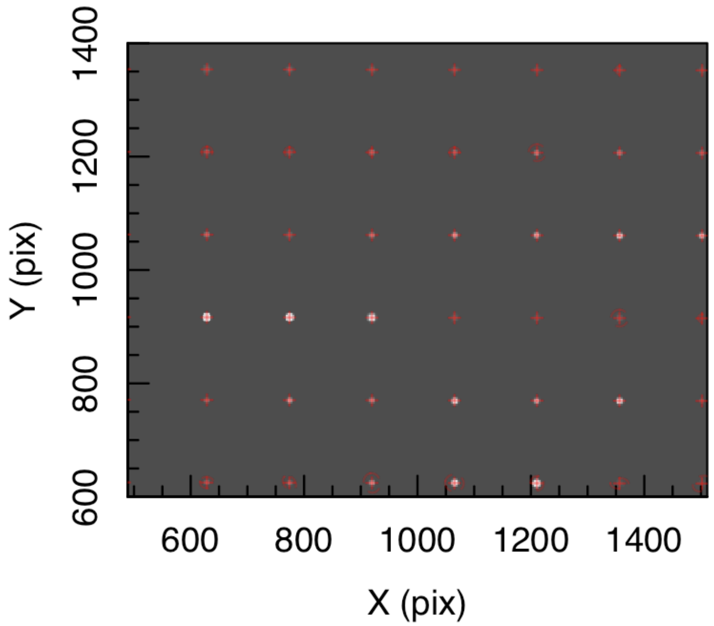
If you only remeber **one thing** from this lecture...

> Correlation does not equal Causation!

Correlation 2: Noise & Detection Effects

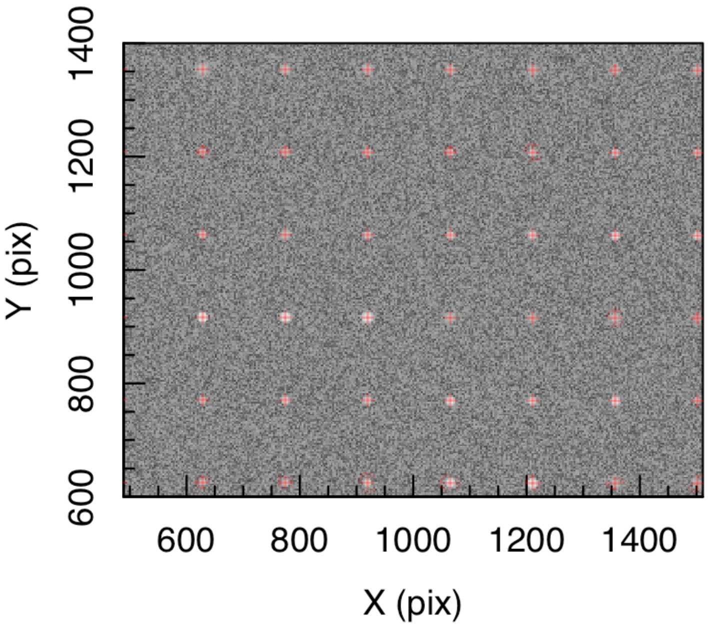
Correlation plays a significant role in astronomical research in particular due to **correlated noise**.

This is an image taken in our “Theorist_Universe_1”.



Let’s now assume that we don’t have perfect detectors, but ones that produce perfectly Gaussian noise

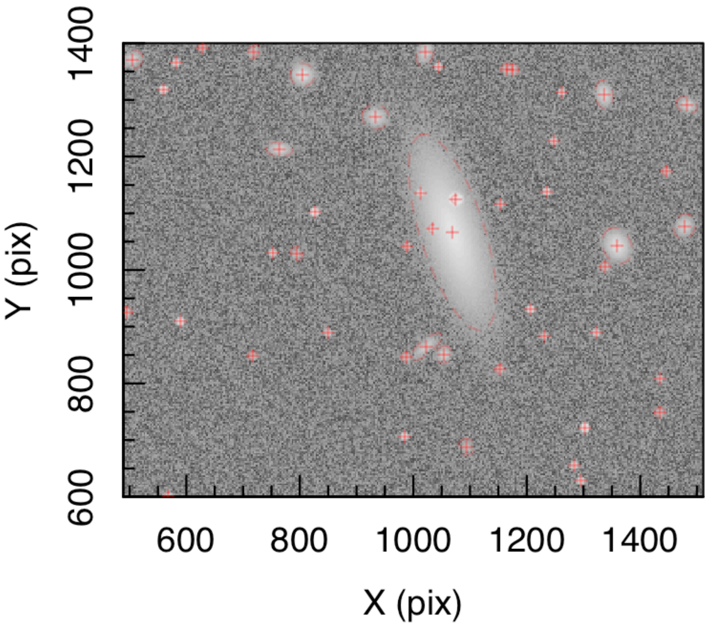
> “Theorist Universe 2: Equally Implausible Boogaloo”



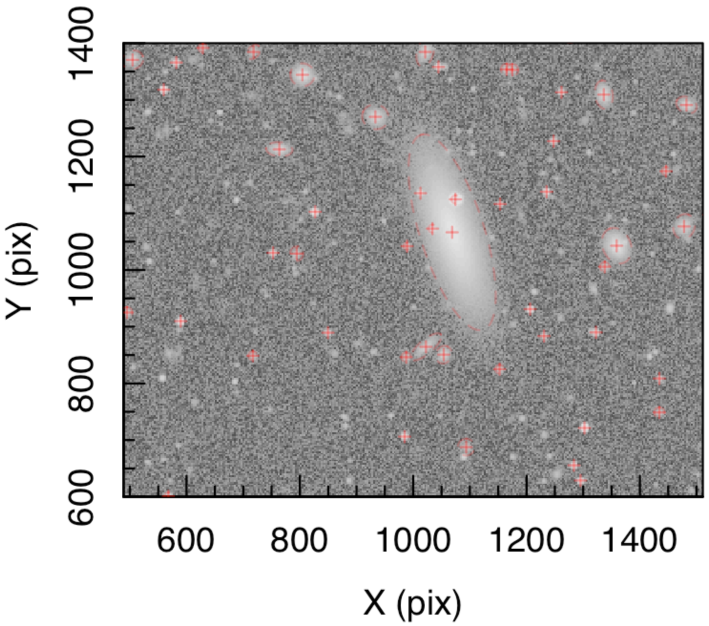
Correlation 2: Noise & Detection Effects

We can make this more complex by adding in:

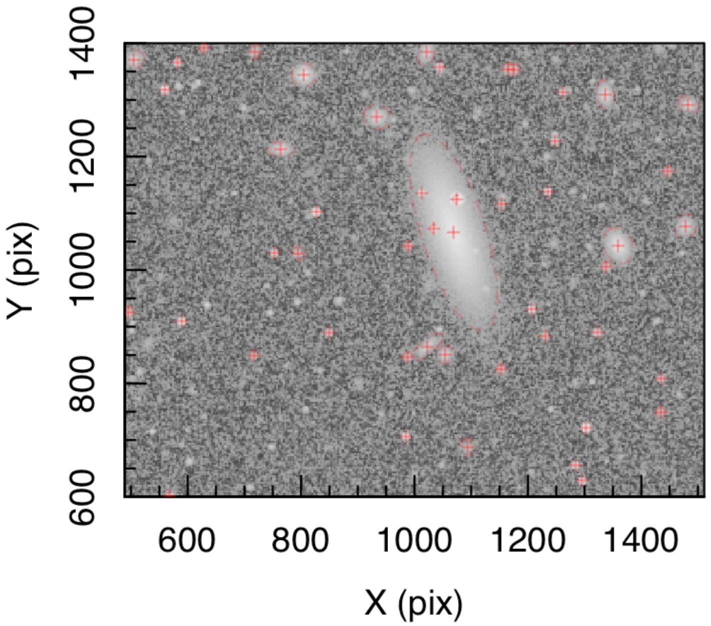
> realistic galaxy distributions & blending



> realistic blending of sources below the detection limit

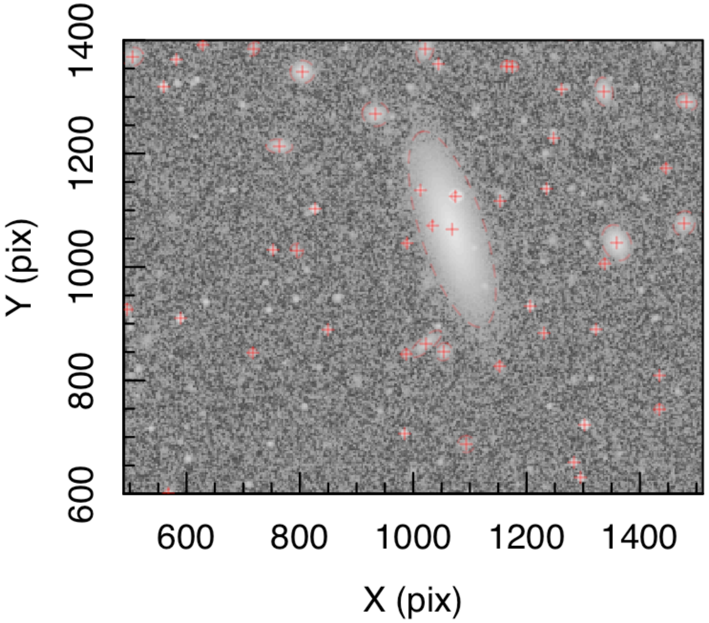


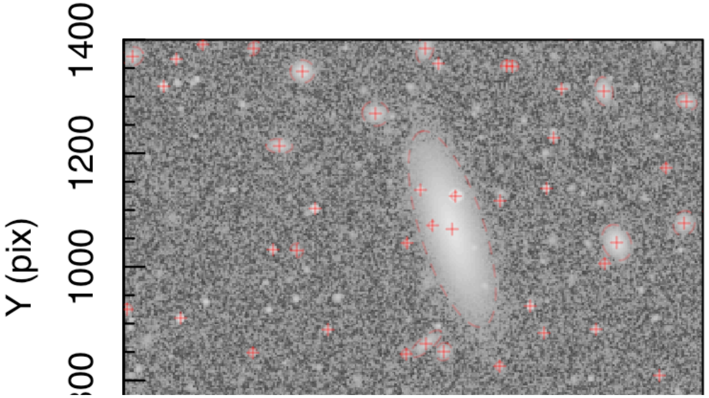
> realistic correlations in the noise profile

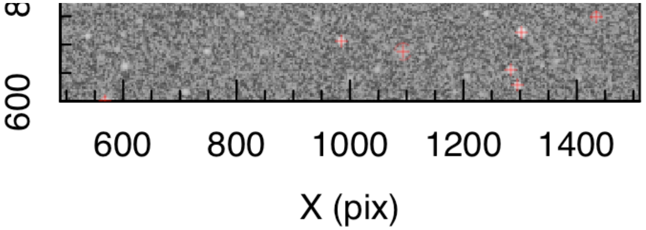


Correlation 2: Noise & Detection Effects

This final product is not dissimilar to actual data:

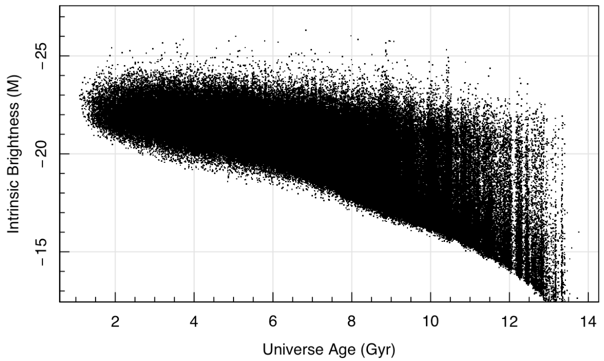






Detection Effects & Bias

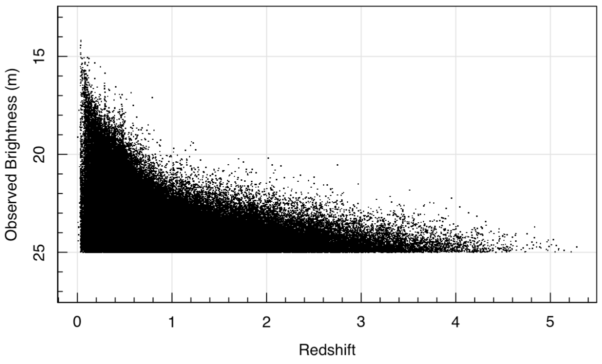
Using our simulated universe, that looks very much like reality, let's look at the distribution of **intrinsic galaxy brightness** vs the age of the universe:



This correlation suggests that galaxies were, on the whole, brighter at earlier times in the universe, and that they are dimming over time.

Detection Bias

Now let's plot the relationship between *observed brightness* and distance:

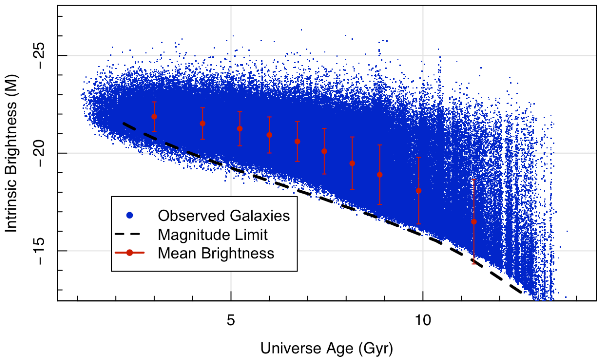


There is a clear relationship between these two properties, and an obvious (artificial) cut-off in the distribution of apparent brightness at $m = 25$.

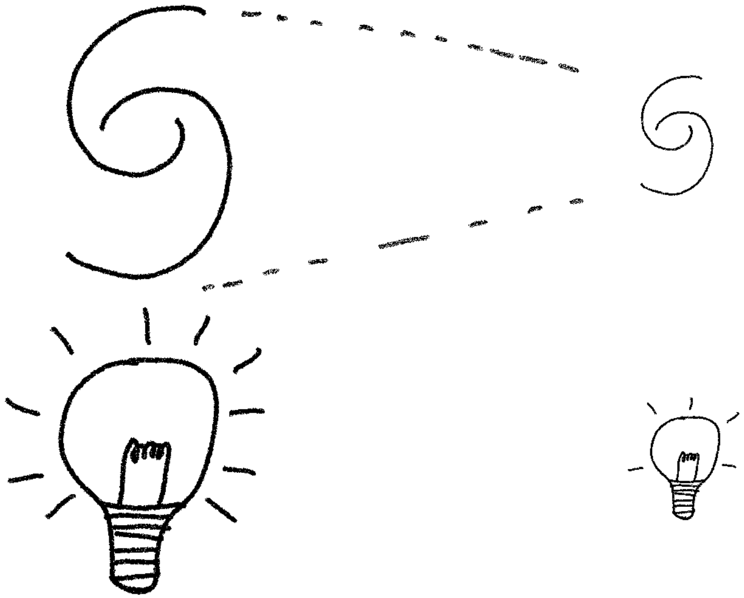
This cut-off is the **magnitude limit** of our toy survey.

What influence does the magnitude limit of our survey have on the distribution of intrinsic brightness?

Malmquist Bias



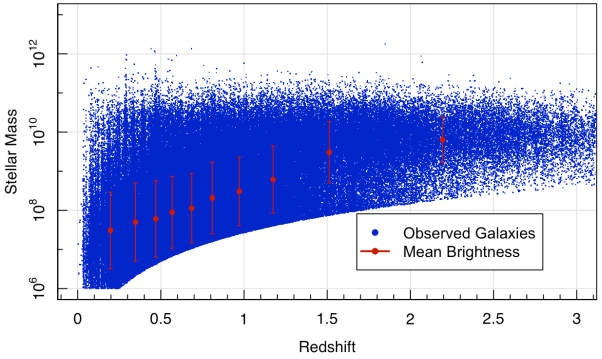
This is an observational bias called “Malmquist Bias”.

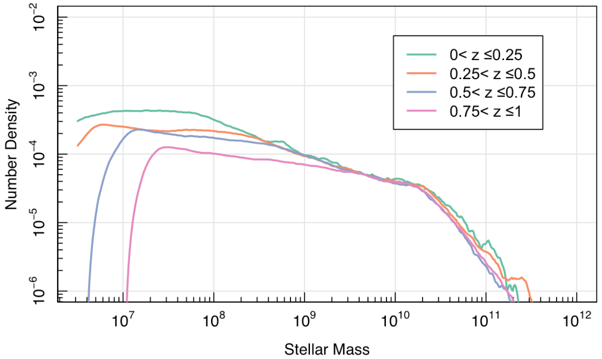


This effect means that galaxy properties, measured as a function of redshift or in wide chunks of redshift, must account for the changing galaxy population.

Malmquist Bias

A common place that Malmquist bias occurs is in the modelling of galaxy distribution functions, such as the galaxy stellar mass function (GMSF) or galaxy Luminosity function (GLF). Failing to account for Malmquist bias in these measurements leads to catastrophic errors:

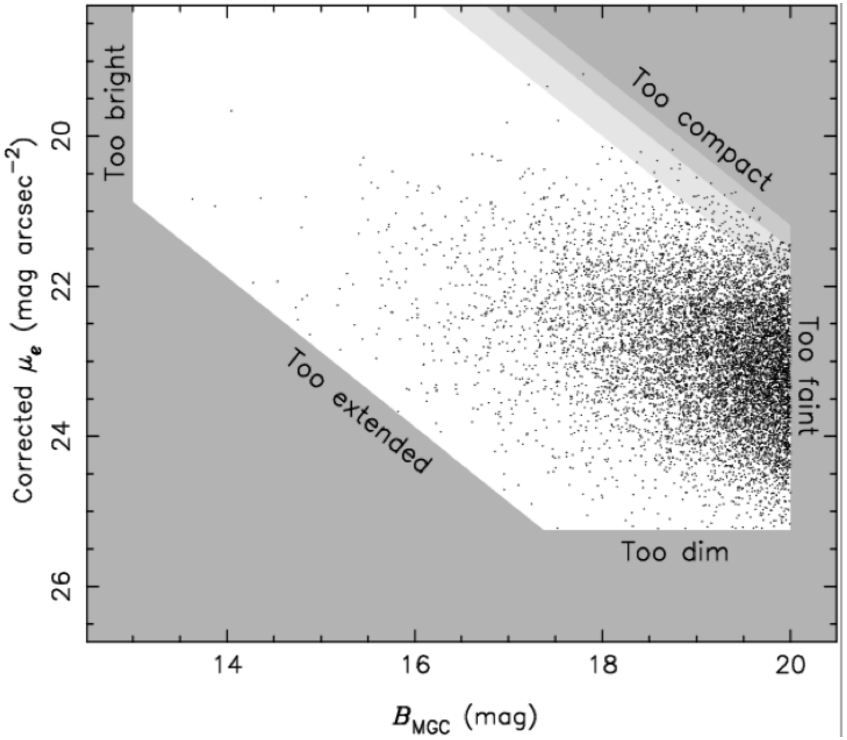




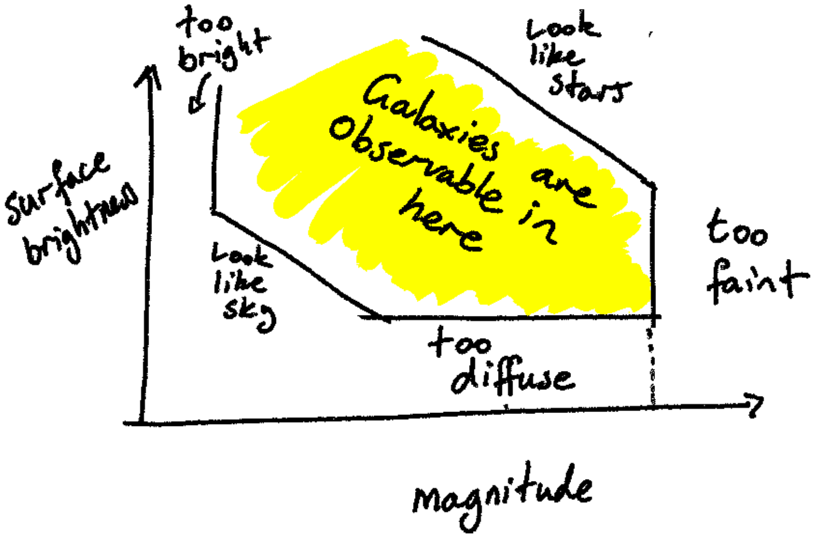
Survivor Bias

The general term for a bias that originates because a studies sample is not representative of the general population because of some selection effect is known as **survivor bias** (in the sense that the remaining sample has "survived" some test).

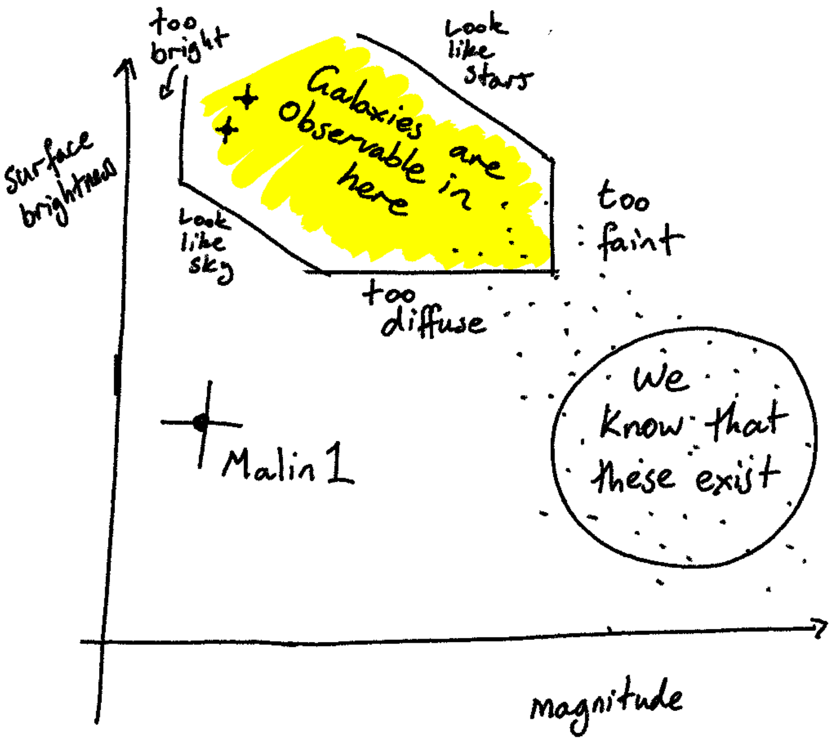
In real astronomical survey imaging, survivor biases are caused by much more than simple magnitude selections (although the magnitude limit is often the most dominant selection). A good example of this is the **bivariate brightness distribution**:

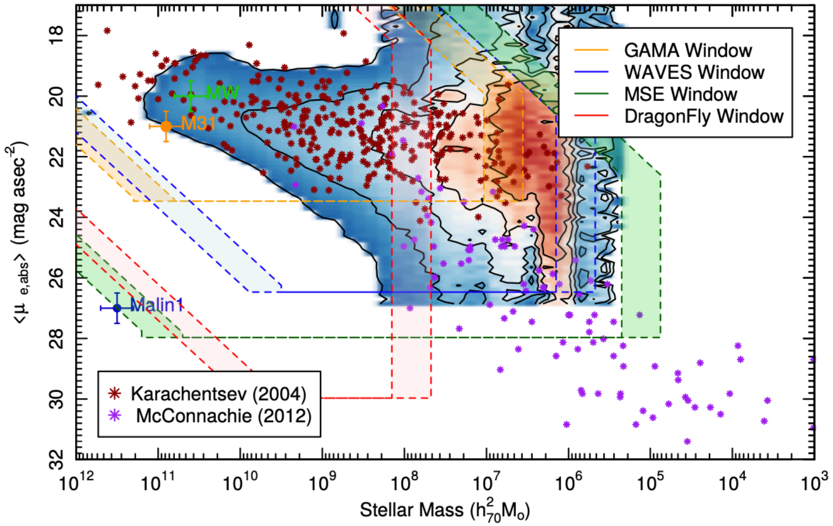


The BBD showcases the 4 major selection effects that impact survey images in astronomy:



The particular problem with these selections is that we *know* galaxies exist outside these limits, because we have (often by accident) discovered them.

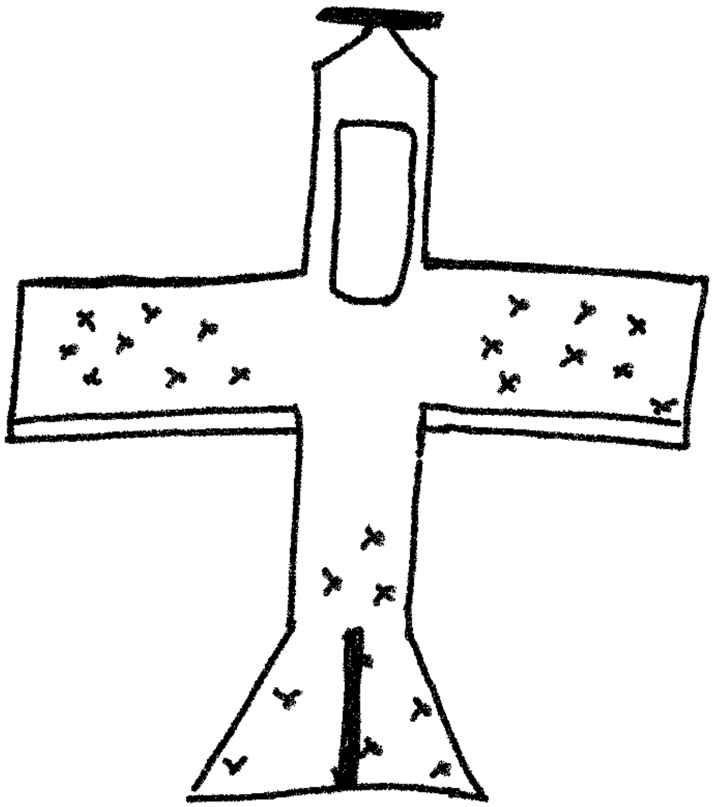




Non-Astronomy Aside: The original "survivor bias"

The name originates from the studies of military aircraft during the dawn of airborne warfare. In an effort to protect aircraft from destruction, the planes ought to be shielded with armour.

Military analysts noted that planes were generally most heavily damaged on their wings and tails when they returned from battle. And so they decided that it was necessary to fortify these areas. However armour is heavy, and reduces the efficiency of the aircraft. So they contracted statistician Abraham Wald to optimise the placement of armour on the aircraft.



Wald returned to the analysts with a recommendation that was somewhat unexpected: Armour the parts of the plane that **don’t** have any bullet holes.

The reason was simple: bullet’s do not preferentially strike wings and tails. They should be randomly distributed over the fuselage, but they are not.

Therefore, there must be some selection effect that means planes with bullet holes on the wings and tails *preferentially* return home. Why? Because planes that get shot in the engine crash!

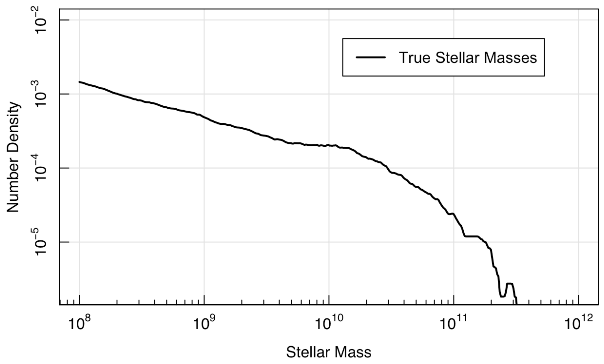
More Biases in Imaging: More Noise, Mo’ problems

> 1. All observational data has uncertainties, because no instrument is perfect.
> 2. All models have uncertainties, because no model looks exactly like the real universe.
> 3. Ergo: you will always be working with uncertain quantities.

Noise effects

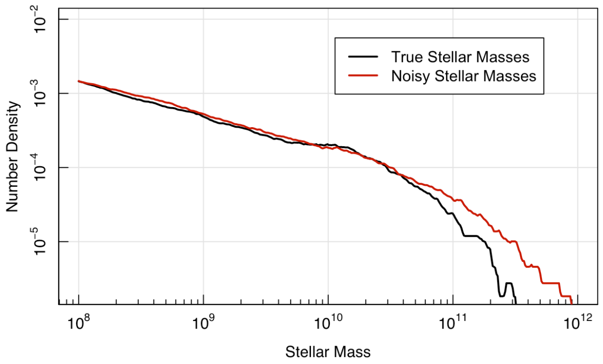
Back to the toy universe: our observed galaxies have some distribution of “true” stellar mass (now corrected for Malmquist Bias!):

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 8 y values <= 0 omitted from l
ogarithmic plot
```



What happens if we add a small amount of noise to our estimated stellar masses?

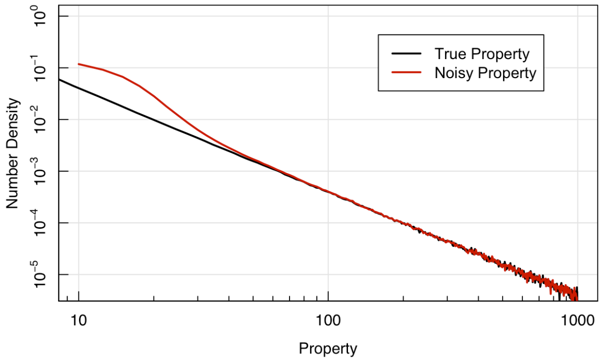
```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 8 y values <= 0 omitted from l
ogarithmic plot
```



Eddington Bias

This effect known as “Eddington Bias”, and is most apparent when modelling a property that follows a simple power law distribution, and a constant Gaussian uncertainty on the parameter:

```
## Warning in samp + rnorm(length(ind), sd = 8): longer object length is not a mul
tiple of shorter
## object length
```



It is caused by the distribution of sources being highly asymmetric. You can think of this probabilistically:

- The probability that any one source scatters by ± 10 is very small.
- At any one point on the x-axis, there are more sources to the left than the right
- Ergo: there is a greater absolute chance that sources from the left will scatter rightward, than vice versa.

Statistical Biases & Astronomical Analyses

Failing to understand the properties of the statistics that are used in an analysis, or how to interpret correlations, or how to account for selection effects, or how to account for noise, will all lead to errors in an analysis.

But: simply **being aware** of these effects already puts you at an advantage. Great Job!