AstroStatistics Ringvorlesung

A motivation for, and introduction to, practical statistics for Astronomy

Dr Angus H Wright

2021-11-08

Performing Modern Astronomical Research

Regardless of your field of study, your career stage, your affiliation, or any other specialisation that makes you a unique, identifiable member of the astronomical community, there are two things that can be unequivocally said about every single astronomer:

- They need to know how to write computer code
- They need to have been taught some statistics

Statistics & Computers

Modern physics and astronomy requires an understanding of programming. From theoreticians writing models to experimentalists writing analysis pipelines, most physicists and astronomers will use read, write, or use a computer program every day.

N-body Simulation

An excellent example of this is the N-body simulation. The simulation of large gravitational bodies (stars/galaxies/clusters) using point-masses. The term was coined by von Hoerner in 1960.

Astronomisches Rechen-Institut in Heidelberg
Mitteilungen Serie A Nr. 14

Die numerische Integration des *n*-Körper-Problemes für Sternhaufen I

Von

SEBASTIAN VON HOERNER

Mit 3 Textabbildungen

(Eingegangen am 10. Mai 1960)

Holmberg (1941)

However, in 1941, 20 years prior to a famous work by Sebastian von Hoerner that established the field (and name) N-body Simulations, Erik Holmberg performed the first simulations of colliding galaxies.

THE ASTROPHYSICAL JOURNAL

AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS

VOLUME 94 NOVEMBER 1941 NUMBER 3

ON THE CLUSTERING TENDENCIES AMONG THE NEBULAE

II. A STUDY OF ENCOUNTERS BETWEEN LABORATORY MODELS OF STELLAR SYSTEMS BY A NEW INTEGRATION PROCEDURE

ERIK HOLMBERG

Holmberg's work was exceptional for a number of reasons, but has become famous because of *how* it was completed. Holmberg simulated the collisions of rotating spiral galaxies:

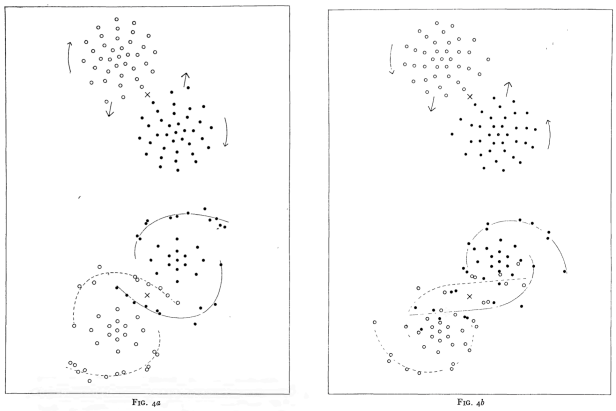


FIG. 4a FIG. 4b

FIG. 4a.—Tidal deformations corresponding to parabolic motions, clockwise rotations, and a distance of closest approach equal to the diameters of the nebulae. The spiral arms point in the direction of the rotation.

FIG. 4b.—Same as above, with the exception of counterclockwise rotations. The spiral arms point in the direction opposite to the rotation.

And generated tidal disruption features that are now seen commonly in merging spiral galaxies:

The surprise?

His work was computed entirely *by hand*. Holmberg used arrangements of lightbulbs to simulate groups of stars, and photometers to compute the gravitational pull of all mass-elements on each-other per unit time.

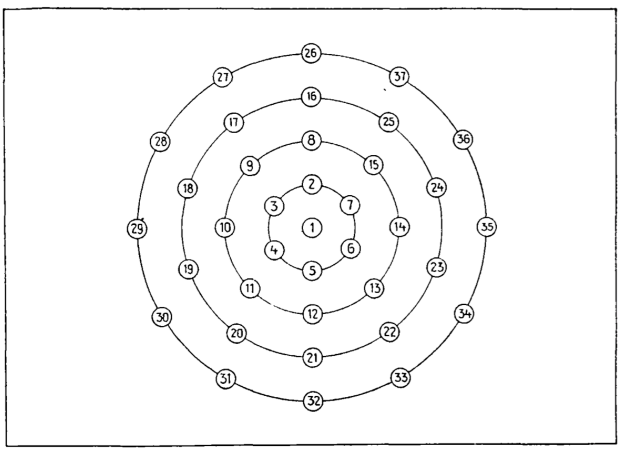


FIG. 3

Reproducing Holmberg in 2021

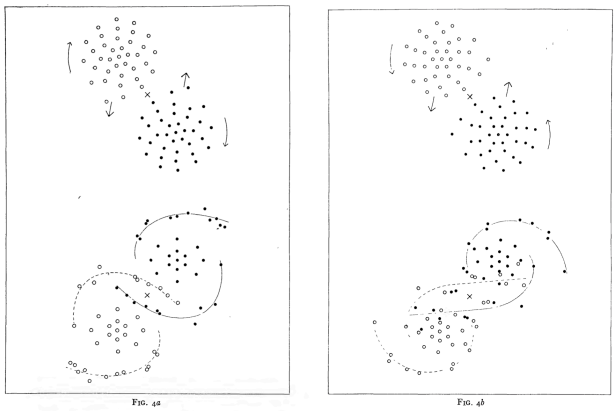
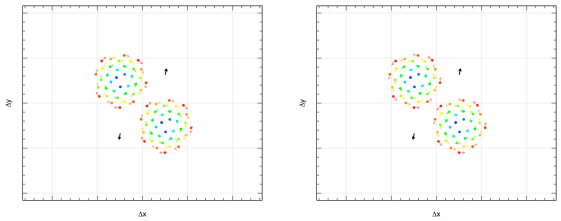
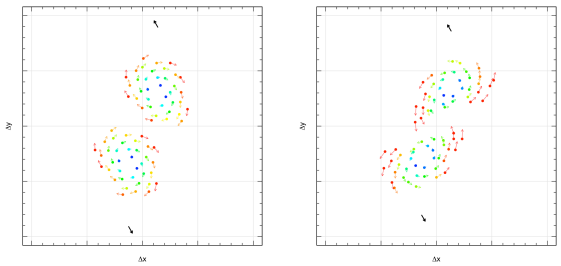


Fig. 4a Fig. 4b

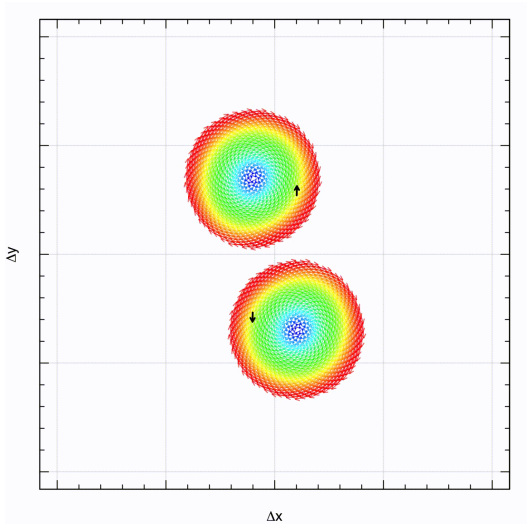
Fig. 4a.—Tidal deformations corresponding to parabolic motions, clockwise rotations, and a distance of closest approach equal to the diameters of the nebulae. The spiral arms point in the direction of the rotation.

Fig. 4b.—Same as above, with the exception of counterclockwise rotations. The spiral arms point in the direction opposite to the rotation.

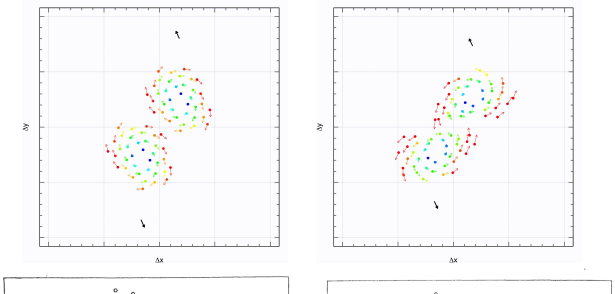




But “Scalability” is the main benefit



Astronomy and Statistics



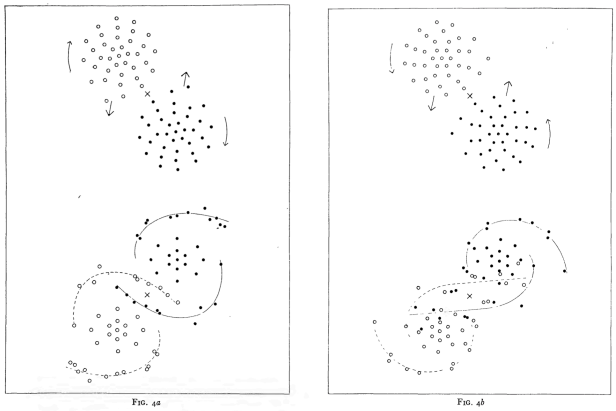


Fig. 4a Fig. 4b

Fig. 4a.—Tidal deformations corresponding to parabolic motions, clockwise rotations, and a distance of closest approach equal to the diameters of the nebulae. The spiral arms point in the direction of the rotation.

Fig. 4b.—Same as above, with the exception of counterclockwise rotations. The spiral arms point in the direction opposite to the rotation.

It is uncontroversial (I think?!) to say that having an understanding of programming is vital to modern astronomical research.

Less appreciated, and certainly less obvious to any student working their way toward becoming a future astronomer (i.e. you!), is the fact that an understanding of statistics is **just as important** a skill as programming.

“Well that’s a bit exaggerated, Angus”

Maybe…

But there is a common misconception in some astronomical fields that only people working in large surveys (and with many sources) require an understanding of statistics. This is **unequivocally** false, as you will see repeatedly throughout this lecture.

Unfortunately, though, one 90 minute lecture is an insufficient amount of time to teach (let alone learn!) the nuances of statistical analysis methods and best practices. Indeed, in a standard Astrostatistics lecture series, we would spend one 90 minute lecture on **each** the topics of:

- Introducing statistical computing
- Understanding and exploring data
- Describing and modelling data
- Understanding Probability
- Bayesian Statistics
- Modelling probabilistic events
- Complex analysis methods (MCMC)
- Machine Learning methods
- My favourite: “How to not be wrong”

and much more. Unfortunately learning these in one hour is not possible, and skimming through them typically does more harm than good.

Statistical Methods vs Statistical Awareness

This lecture will therefore not cover statistical methods. Instead we will spend the next hour trying to develop our **statistical awareness**: our ability to recognise and understand the sorts of statistical biases that can wreak havoc with otherwise solid astronomical analyses.

What we’ll cover today

In this lecture we will discuss statistics and statistical fallacies in the context of astronomy. Or, framed in a slightly more fun way:

How to **invalidate** your research in 7 easy steps!

1. Looking at your data
2. Making selections on your data
3. Using data which has uncertainties
4. Focussing on interesting sources
5. Fitting models to your data
6. Trying to find "truth" in your data
7. Reproducing previous work

Statistical Paradoxes

Many of the errors we will discuss today result from *statistical paradoxes*. These are not paradoxes in the sense that "they break physical laws and rigorous logic", but rather they are paradoxes because (for the vast majority of people) your brain instinctively draws the wrong conclusion.

A simple numerical paradox

A trivial example of a mathematical fallacy is one that is likely well known to all of us here: the fallacy of large numbers.

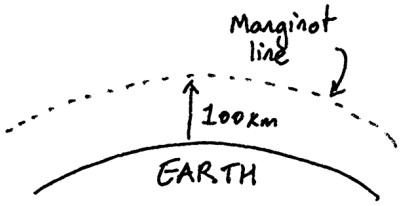
This fallacy says, quite simply, that your brain is incapable of rationalising/comprehending large numbers (or conversely, that humans generally think that numbers are large when they, mathematically speaking, are not).

This is a fallacy we encounter every day in astronomy.

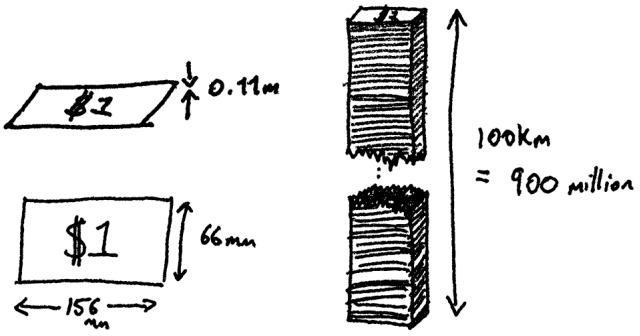
- The Andromeda galaxy is ~ 11 million light-years away
- The Universe is ~ 13 billion years old
- There are ~ 100 billion stars in the Milky Way Galaxy

The human brain is not constructed to innately understand these sorts of numbers, and this leads to bias.

Take the humble sheet of paper



- A US$1 bill is 0.1 mm thick
- A 100km tall pile of US$1 bills is ~ 900 million bills tall



- Jeff Bezos' net worth is 177 billion USD
- Jeff Bezos could make 196 100km-tall piles of US$1 bills.
- Stacked alongside one-another…

- This is a stack of 130 sheets of standard format A4 paper.
- Without trying to crunch the numbers, ask your gut: how tall is 100 billion sheets of paper?

The humble sheet of paper

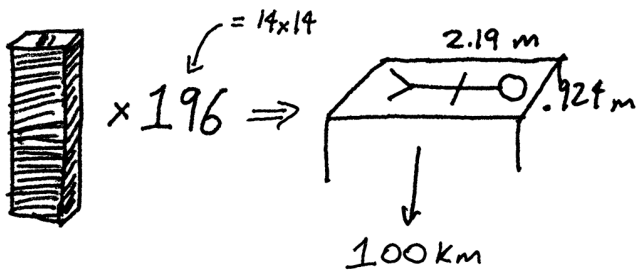
Standard format A4 paper has a thickness of 0.04mm. 100 billion sheets is therefore:

$$0.04\,\text{mm} \times 10^{-6}\,\text{km/mm} \times 100 \times 10^{9} = 4\,000\,\text{km}$$

Non-Astronomy Aside: Billionaires and the Manginot line

You've likely all seen Jeff Bezos (and other billionaires) voyages into space recently.

The "important point" is called the Manginot line, and is 100km above the earth's surface.



- Jeff Bezos could make a dollar bill tower that is 100km tall and larger than a standard-sized single bed. He could sleep above the Manginot line on a tower made of nothing but his own wealth
- With any luck he would stay there

Outline

In this lecture we will discuss statistics and statistical fallacies in the context of astronomy. Or, framed in a slightly more fun way:

How to invalidate your research in 7 easy steps!

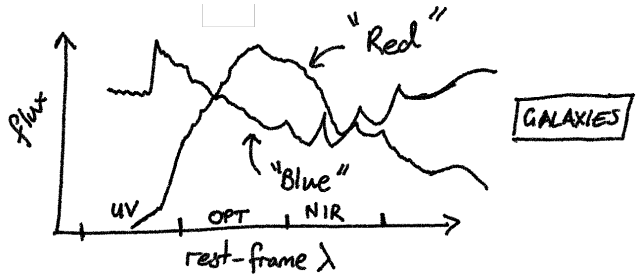
1. Looking at your data
2. Making selections on your data
3. Using data which has uncertainties
4. Focussing on interesting sources
5. Fitting models to your data
6. Trying to find "truth" in your data
7. Reproducing previous work

Method 1: Looking at your data

Let's start with a simple physical dataset in multiple dimensions that we want to investigate. For the vast majority of this lecture we're going to work with a simulated **toy universe** that I have constructed. Galaxies in this universe are simple and easy to model, the universe obeys basic cosmological principles, telescopes and detectors produce perfectly Gaussian uncertainties.

- In short, my toy universe is an astronomers idea of **heaven**.

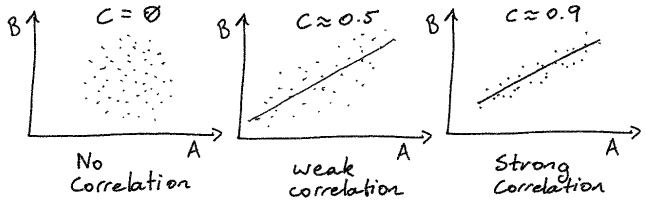
To start our exploration into my mock universe, we're going to look at the distribution of galaxies in my universe. Galaxies are observed with telescopes and fluxes are measured in various filters. We use these fluxes to measure galaxy redshifts and galaxy properties using "Spectral Energy Distribution" modelling.



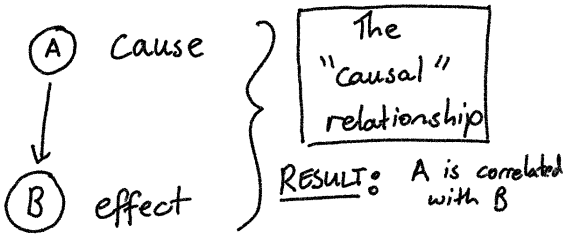
Galaxies have a number of physical parameters estimated, and a number of additional properties are included (e.g. environment).

Relationships between parameters

When plotting data in multiple dimensions, we will frequently find properties that show some sort of relationship. That is: variation in one parameter is coincident with variation in another parameter. This behaviour is called **correlation**.



If one parameter is related to another, we are generally interested in determining if that relationship is **causal**; that is, whether or not it is indicative of some underlying physical process that links the two parameters.

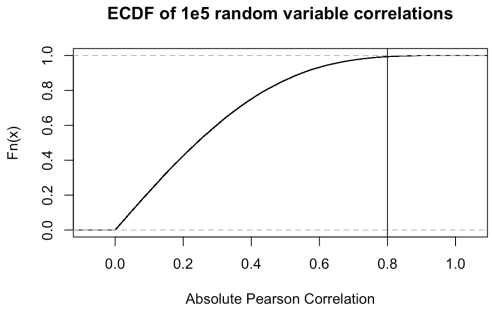


A sneak peak at probability

As a demonstration, let's say that in our toy universe we have a catalogue of 10 galaxies, with 1000 properties measured for each galaxy. A truly spectacular dataset. We decide a relationship is worth investigating if it contains an 80% correlation or more.

But, our toy universe is cruel: someone corrupted our dataset, and (unbeknownst to us) all the variable are filled with completely random values.

What fraction of our random variables do we expect to have a correlation of 80% or more?




```
## 0.65 % of variables have 80% correlation or more
```

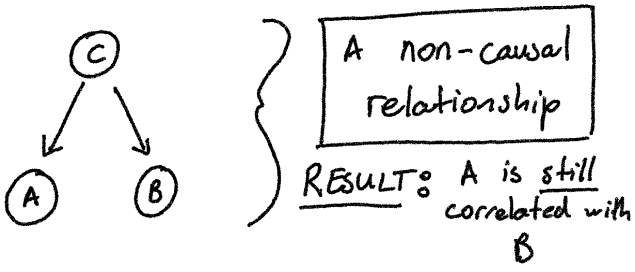
Said differently, there is a 1 in 154 chance that two totally random variables in our survey will have an absolute correlation of 0.8 or higher.

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.

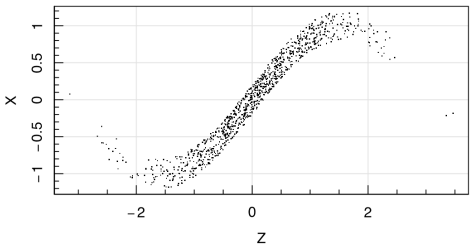
> -**Working with small samples does not mean you can ignore statistics**

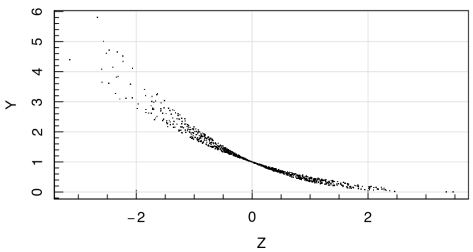
The problem is further complicated by the existence of **confounding variables**.



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.

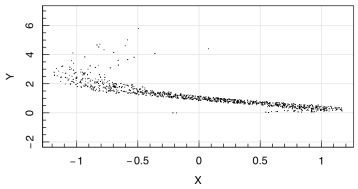
```
#A simple Gaussian dataset
obs<-data.frame(Z=rnorm(1E3,mean=0,sd=1))
#A new variable that correlates with Z
obs$X<-sin(obs$Z)+runif(1e3,min=-0.2,max=0.2)
#And another new variable that correlates with Z
obs$Y<-(1-obs$Z/4*runif(1e3,min=0.8,max=1.2))^3
#Plot them
with(obs, {
  magplot(Z,X,xlab="Z",ylab="X",pch='.')
  magplot(Z,Y,xlab="Z",ylab="Y",pch='.')
})
```





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:

```
#Plot X and Y
magplot(obs$X,obs$Y,xlab='X',ylab='Y',ylim=c(-2,7),pch='.')
```

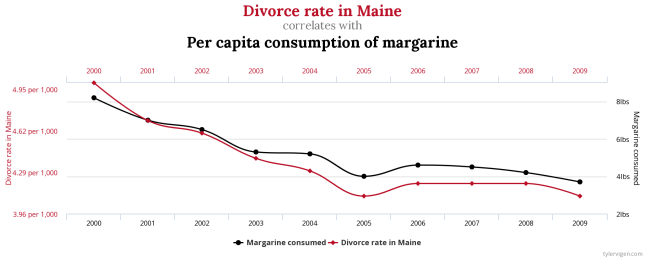



```
#Rank Correlation
cor(obs$X,obs$Y,method='spearman')
```

```
## [1] -0.9691508
```

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

A fundamental distinction



Here we can see that the number of divorces in the US state of Maine is correlated with the consumption of margarine per person in the USA.

The conclusion is clear:

- **Eating margarine in LA will invariably lead to a divorce in Maine** (you monster!).

Or more accurately:

- No, of course it won't.

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…

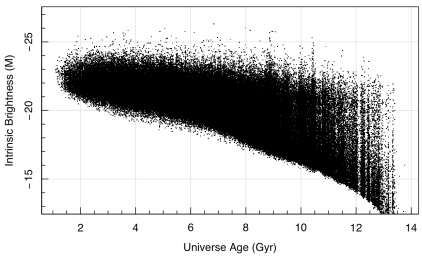
Thus have demonstrated the common but extremely important statistical fact, and Method 1 for invalidating your research:

- Looking for correlations in your data *and then assuming that these correlations are causal*.

> . Correlation does not equal Causation!

One final galaxy example

Let's go back to our toy universe and look at one last correlation: 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. Is this correlation causal? Or spurious? Or something else?

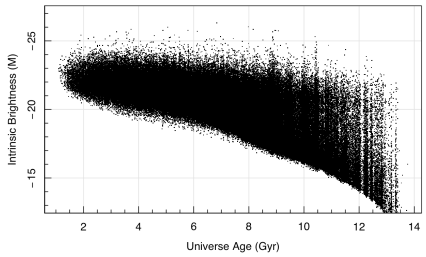
Method 2: Making Selections on your data

Prior to performing any observational astronomical analysis, we need data. However there are too many galaxies in the visible universe to observe with any one instrument, so astronomical samples are *never* complete. That is, they are always a subset of the total population of all galaxies in the universe.

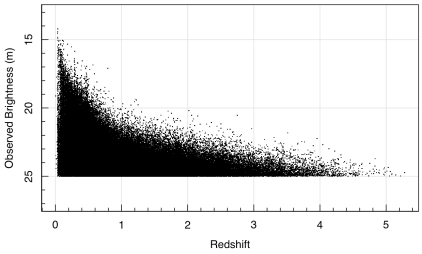
Selecting samples is therefore a fundamental part of astronomical analyses. It is unavoidable. However, whenever we perform even simple sample selections, we open the door to pathological biases in our analysis.

- **No matter what data you use, statistics governs your selection function and therefore your ~~soul~~ biases**

Let's go back to our toy universe. This is our plot of intrinsic galaxy brightness vs Universe age from the last section:



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

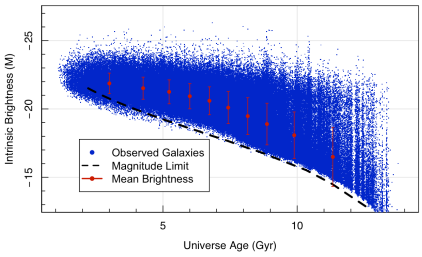


There is a clear relationship between these two properties, and an obvious (artificial) cut-off in the distribution of apparent brightness at $m = 25$.

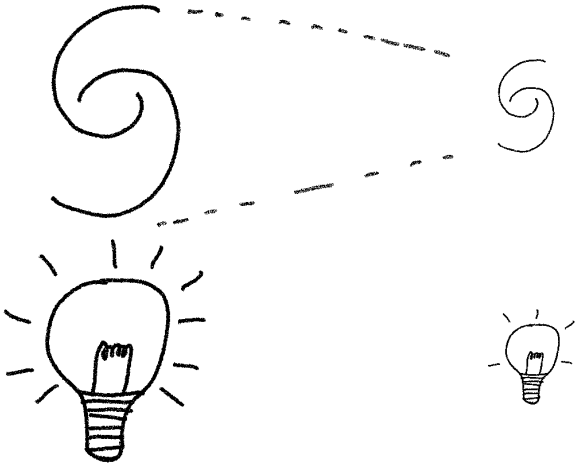
This cut-off is the **magnitude limit** of our toy survey.

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

Malmquist Bias

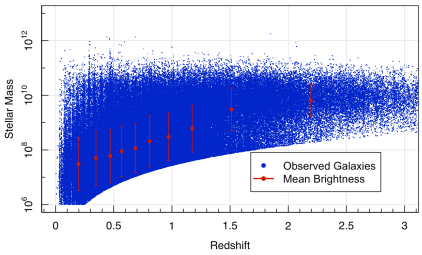


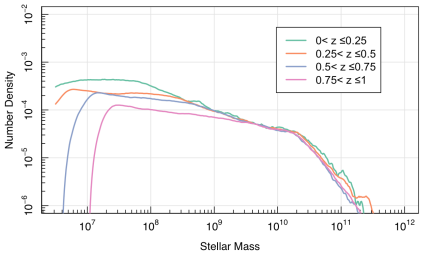
This is an observational bias called "Malmquist Bias". It refers to the bias that one observes in the mean intrinsic brightness of galaxies as a function of distance, caused by the fact observed brightness is a strong function of distance.



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. In this way, Malmquist bias is a form of **survivor 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:





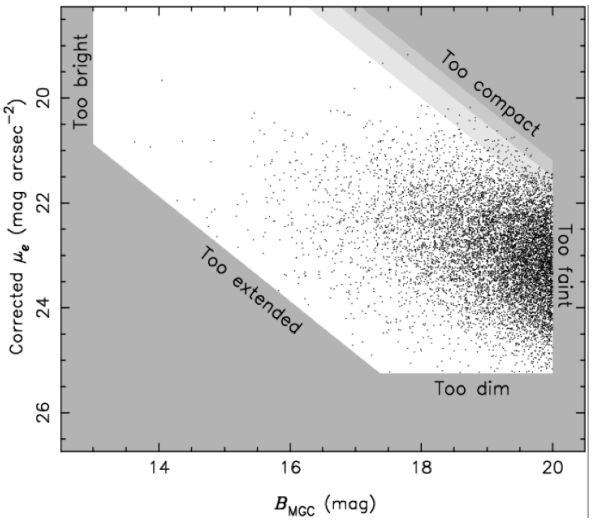
Thus we have Method 2 for invalidating your research:

- Failing to correct for observational selection functions present in **every dataset**.

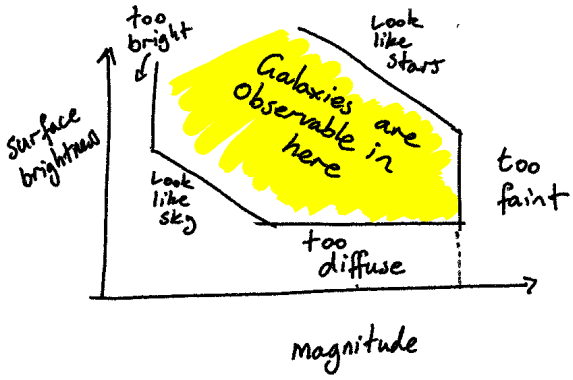
Survivor Bias

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).

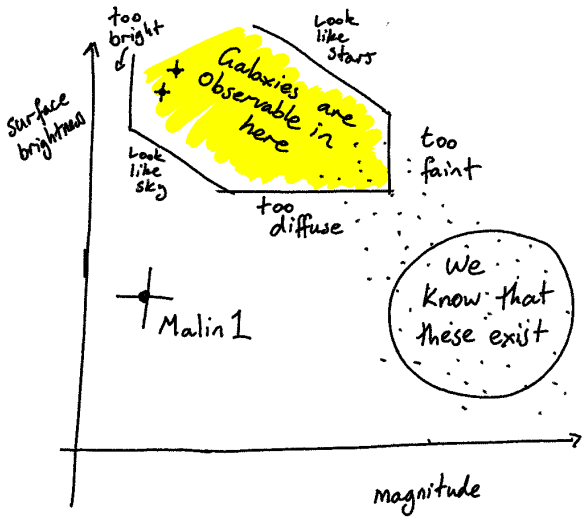
In real astronomical survey imaging, survivor biases are caused by much more than simple magnitude selections (although the magnitude limit is often the most dominant selection). A good example of this is the **bivariate brightness distribution**:

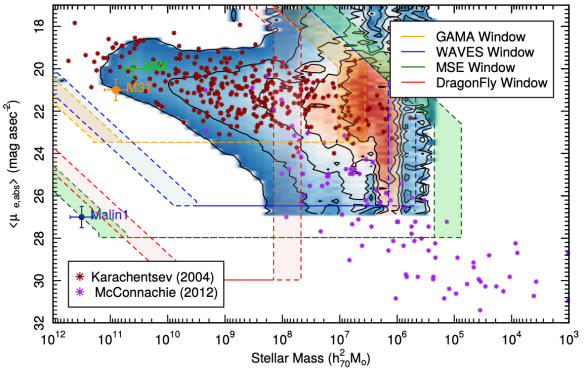


The BBD showcases the 4 major selection effects that impact survey images in astronomy:



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

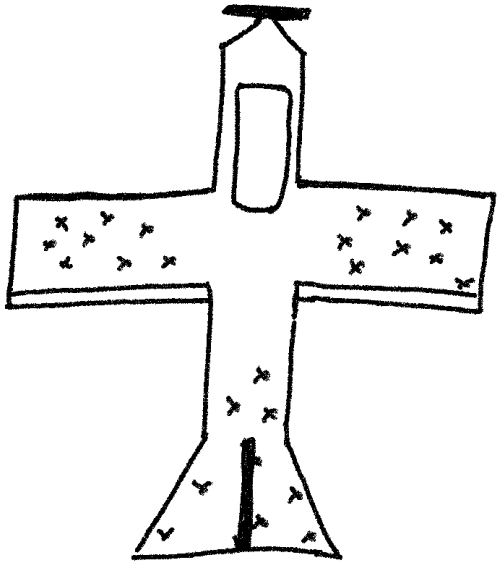




Non-Astronomy Aside: The original "survivor bias"

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!

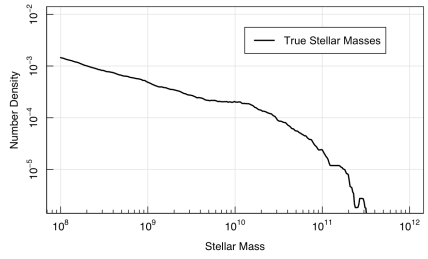
Method 3: Using data which has uncertainties

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.

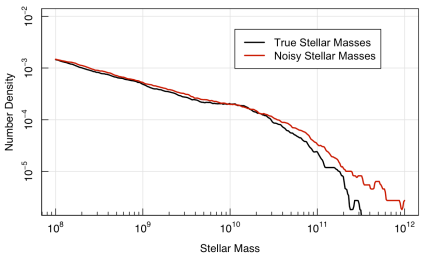
"Monte Carlo" realisations

Monte Carlo realisations refer to data that are generated by adding random perturbations to the data values. A simple example is the simulation of uncertainties given some "true" underlying distribution.

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

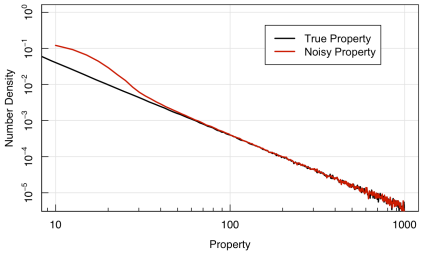


What happens if we add a small amount of noise to our estimated stellar masses?



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:



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.

Thus we have Method 3 for invalidating your research:

- Failing to account for the influence of observational uncertainties on your models and derived data distributions.

Method 4: Focussing on interesting sources

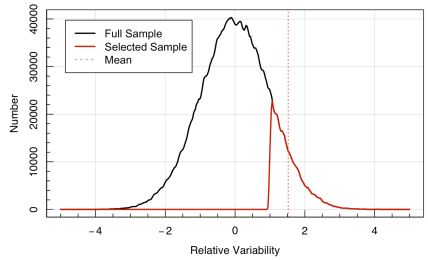
In many areas of astronomy and astrophysics, extreme and rare phenomena are a focus for study and are highly prized as "the most interesting sources". Examples include extreme stars (blue stragglers, variables, super-giants), extreme galaxies (rings, AGN, mergers), or rare objects (exotic transients, strong gravitational lenses).

Selecting and focussing on these rare and extreme objects is generally very attractive, as frequently they are sensitive laboratories for understanding underlying physical phenomena.

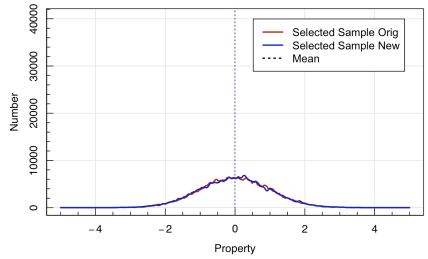
However simply focussing on extreme phenomena can in itself lead to biases in analysis. This effect is particularly dangerous in temporal studies, but also arises in other areas.

Let's say that we are working in the field of stellar transients, and we want to know if the variability of a source is impacted by the environment that the source resides in. Specifically, whether interactions between star clusters reduce the overall variability of stars.

We start by constructing a sample of sources. Using existing data, we take a population of variable stars, and select a sample which are both extremely variable and which reside in stellar clusters that are undergoing violent interaction.



We then return to these transients with our telescope at a later date time (much longer than the period of variation in the sources). And we reobserve our sample.



Thus we have Method 4 for invalidating your research:

- Focussing on extrema, failing to account for the pathological selections, and ignoring the properties global samples.

Non-astronomy Aside

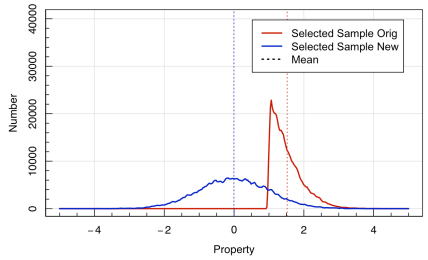
This effect is extremely common in social and medical sciences, where it goes by another name: *the placebo effect*. Contrary to popular belief, there is only very mild evidence that placebo's actually have (quantitative) pharmacological benefit. The overwhelming majority of reports claiming to show concrete increases in medical well-being after the administration of a placebo are simply finding regression to the mean.

Method 5: Fitting models to your data

The p-value

In this lecturers opinion, the **p-value** is easily the most misunderstood, misused, and/or misrepresented concept in statistics. So what is the p-value, and what is it not.

- **What does the p-value tell you**: The p-value represents the fraction of samples that would produce a test statistic that is as extreme as the one observed, given that the proposed (generally null) hypothesis is true.



We conclude that the interaction has significantly reduced the variability of our sources, and **we publish**.

What is wrong with this?

Regression to the mean

This effect is known as 'regression to the mean', It refers to the phenomenon whereby a random variable that is initially selected to be extreme will naturally tend to less extreme values under subsequent observations.

The mechanism to safeguard against this phenomenon is to ensure that your sample contains sources that are both sensitive **and insensitive** to the effect you are trying to measure. In our example, this would mean defining two samples:

- One with extreme variability **and** residing in an interacting cluster
- One with extreme variability **and not** residing in an interacting cluster

- **What is the p-value *not* tell you**: The p-value does *not* tell you anything about the probability that the proposed hypothesis is correct, *nor* about whether or not the data can be explained as being produced by random chance.

Nonetheless, the p-value is widely used in the academic literature as a tool for hypothesis testing, and/or for justification that experimental evidence is incompatible with the *null hypothesis* (i.e. that there is no underlying effect/difference).

Let's assume that we are willing to believe an effect if it has a p-value of α or less; otherwise you reject the effect in favour of the null hypothesis. The probability that you accept a hypothesis that is actually *false*, is the fraction of samples that would give you a satisfactory p-value even though the null hypothesis was true. But this value is just α. So you can consider the p-value as being the probability that you have accepted a hypothesis that is false.

Additional Observations

A significant statistical fallacy in significance estimation comes from the ability of researchers to adaptively observe more data.

Consider an experiment where we make n observations of a variable X. We compute our statistic of choice, say the t.test, and calculate a p-value.

We find that our p-value is on the cusp of being "significant". We therefore decide to perform some additional observations, and find that the p-value decreases below our required threshold. Confident that these additional data have confirmed our effect is real:

We Publish

Can you see a problem with this process?

Simulating the effect:

Again this is an effect that we can simulate easily. Let us create a dataset of n observations, and compute the p-value.

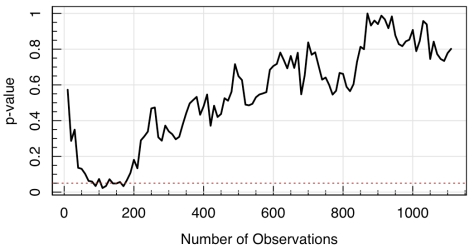
```
##
##  One Sample t-test
##
## data:  obs
## t = -1.8105, df = 99, p-value = 0.07326
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.38179663  0.01748116
## sample estimates:
##  mean of x
## -0.1821577
```

We now decide to observe more data, in a batch of 10 observations.

```
##
##  One Sample t-test
##
## data:  obs
## t = -2.3083, df = 109, p-value = 0.02287
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.40244042 -0.03061316
## sample estimates:
##  mean of x
## -0.2165268
```

Bingo! We cross the threshold of $p < 0.05$ and we rush straight to the publisher.

But what happens if we were to continue observing data?



p-Hacking

This effect is known most colloquially as **p-hacking** (although that term can be applied to many of the practices that we discuss here). Generally speaking the problem is that we can *decide* when to stop taking observations based on the significance threshold we want to achieve. This allows us to keep observing data until we work our way down to a significant result.

We can ask the question: how often can I hack my way to significance with up to 1000 observations taken 10 at a time?

```
## published
## FALSE  TRUE
##  0.71  0.29
```

So by selectively observing more data, we publish 30% of the time given a statistical significance threshold of 0.05.

```
##
##  One Sample t-test
##
## data:  obs$V2
## t = -0.96755, df = 99, p-value = 0.3356
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.26689135  0.09192394
## sample estimates:
##   mean of x
## -0.08748371
```

Also nothing… let's keep going…

The fourth variable:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.68248 -0.53272 -0.04569  0.03294  0.67478  2.42216
```

```
##
##  One Sample t-test
##
## data:  obs$V4
## t = 0.3759, df = 99, p-value = 0.7078
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1409202  0.2067931
## sample estimates:
##  mean of x
## 0.03293646
```

… the ninth…

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.55382 -0.66473  0.02398  0.06146  0.72420  3.21120
```

```
##
##  One Sample t-test
##
## data:  obs$V9
## t = 0.59221, df = 99, p-value = 0.5551
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1444566  0.2673706
## sample estimates:
##  mean of x
## 0.06145701
```

… the fourteenth…

Thus we have Method 5 for invalidating your research (a double whammy):

- Failing to understand, and therefore abusing the meaning of, p-values
- Modifying our analysis samples until we reach 'significance'

Method 6: Trying to find "truth" in your data

Variable Selection

We are scientists working to determine any interesting relationships present in our data.

Our dataset contains $n = 1e2$ observations (of galaxies, or particle collisions, etc), and we measured 20 different variables for each observation.

We have theoretical expectations of what the data ought to show for each of our variables, which we have already subtracted from each column. So the null hypothesis in these data is always $\theta_i = 0$, and we can compare how our data differs from the null hypothesis using a t-test.

So let's look at our first variable:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.99309 -0.61669  0.08980  0.03251  0.66156  2.28665
```

```
##
##  One Sample t-test
##
## data:  obs$V1
## t = 0.31224, df = 99, p-value = 0.7555
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1741130  0.2391426
## sample estimates:
##  mean of x
## 0.03251482
```

Nothing significant there… what about for our second variable?

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.02468 -0.59150 -0.06929 -0.08748  0.46179  2.70189
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.5351 -0.6938 -0.1362 -0.1486  0.6043  1.7740
```

```
##
##  One Sample t-test
##
## data:  obs$V14
## t = -1.6282, df = 99, p-value = 0.1067
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.32963497  0.03248961
## sample estimates:
##  mean of x
## -0.1485727
```

… the seventeenth…

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0985 -0.2646  0.1909  0.2741  0.8566  3.5847
```

```
##
##  One Sample t-test
##
## data:  obs$V17
## t = 2.6839, df = 99, p-value = 0.008531
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.07145775 0.47674746
## sample estimates:
## mean of x
## 0.2741026
```

Aha!! We've found a significant relationship! The 17^{th} variable is discrepant from the null hypothesis with a p-value of 0.0085308.

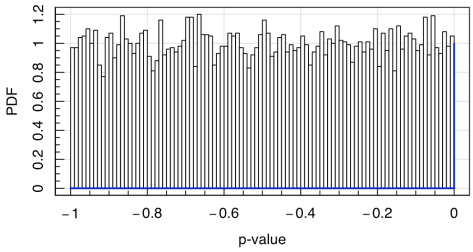
We write up our discovery, publish the result, and our discovery is enshrined in the literature forever.

What is the problem with this?

The process I've described above is known as **data-dredging**, the **look-elsewhere effect**, or the **problem of multiple comparisons**.

The core issue here is that we're looking at many different chunks of the data, any not taking that into account when we decide whether what we've found is significant.

Recall the p-value for many experiments:



We have used in this example a threshold of $p < 0.05$. We therefore expect to find this p-value given random fluctuations in 1 out of every 20 cases. In our example we have 20 variables. So it makes sense that we found a "significant" effect for 1 variable.

Thus we have Method 6 for invalidating your research:

- Data dredging, or analysing multiple hypotheses and not accounting for this in our modelling.

Non-Astronomy Aside: An Empathetic Fish

Do fish feel empathy?

This was a question posed by a group of researches working within the functional magnetic resonance imaging (fMRI) community in 2009. fMRI studies use the magnetic resonance to produce highly detailed internal images of people (and in this case, fish). The field uses analysis techniques that are designed to identify activity within (particularly) the brain that can be correlated with an external stimulus, in order to identify parts of the brain that are responsible for different things, or to just demonstrate that comprehension is occurring.

The case of this experiment was to show whether or not an Atlantic Salmon would react differently when shown images of people, rather than images of inanimate objects.

The researchers placed the fish in an MRI, and presented it with images of humans and other pictures. They analysed the data using standard processing tools, and found a significant discovery of activity in the brain of the salmon that correlated with the researchers presenting the fish with images of humans.

```
## All Results:
##
##            NoEffect+NoResult   NoEffect+SignificantResult          TrueEffect+N
oResult
##                        0.852                        0.048
0.023
## TrueEffect+SignificantResult
##                        0.077
## Published Results:
##
##   NoEffect+SignificantResult TrueEffect+SignificantResult
##                        0.384                        0.616
```

Standard scientific practice is that only measured relationships get published, so while we have 85% insignificant findings, these generally never see the light of day. Instead we only look at the significant results.

So: If *everything is working as it should*, roughly one-third of all published papers that use $p \leq 0.05$ as a threshold for publication ought to be false-positives. This fraction is determined by the following important numbers:

- The **Statistical Power**: how probable an experiment is to find a true relationship when one does exist
- The **Ratio of true-to-false hypotheses**: if we have many many more false hypotheses than true ones, this effect is exacerbated
- The **Significance Threshold**: what magic number people select for determining whether or not an effect is significant (i.e. the p-value threshold).

If we have particularly insightful and disciplined researchers:

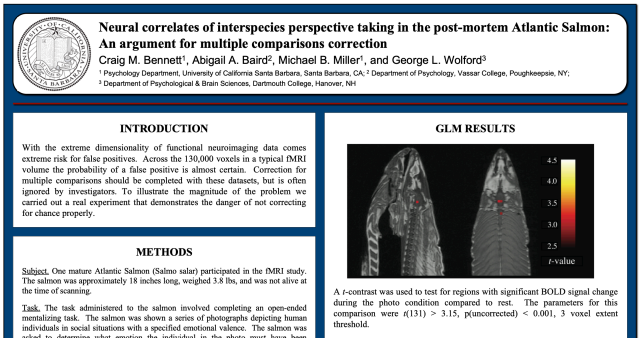
```
## Published Results:
##
##   NoEffect+SignificantResult TrueEffect+SignificantResult
##                   0.05870841                   0.94129159
```

But if, for example, researchers are incentivised to explore greater numbers of exotic hypotheses with less and less prior justification:

```
## Published Results:
##
##   NoEffect+SignificantResult TrueEffect+SignificantResult
##                    0.8787879                    0.1212121
```

The problem?

- **The salmon was frozen at the time of study**



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was

GLM RESULTS

4.5
4.0
3.5
3.0
2.5
t-value

A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

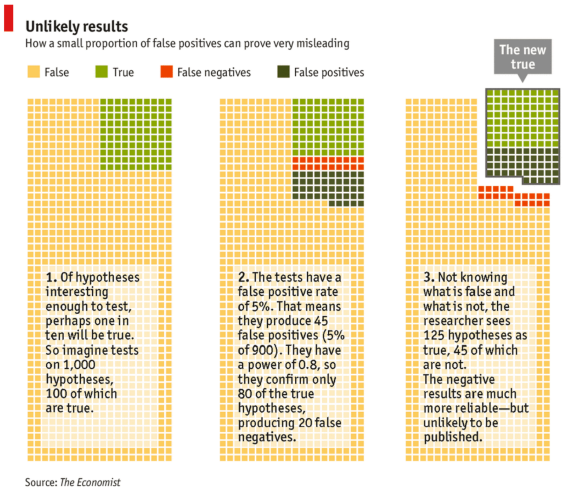
A Disturbing Aside: How much published research is wrong?

Let's assume that we're looking at a field of research where there are 500 ongoing experiments, all exploring different possible physical relationships. Of those 500 experiments, 50 of them are real physical relationships.

If all researchers use a metric of $p < 0.05$ as their determination for whether an effect is real or not, *and researchers only publish when they find a significant result*, what will the fraction of published results that are wrong?

A simple simulation

We can construct a simulation to demonstrate this situation. We simply simulate 500 draws from a Gaussian, where 50 draws have $\mu \neq 0$ and the rest have $\mu = 0$. We can then compute the p-value for each of these experiments, and "publish" those with 'significant' findings. Assume that we know the true parameter variance $s^2 = 1$, for simplicity.



Method 7: Reproducing previous work

Generally speaking, the reproduction of previous work is an extremely valuable task. Here we show the dark side of being aware of the literature.

Take, as an example, measurements of the coefficient of charge-parity violation:

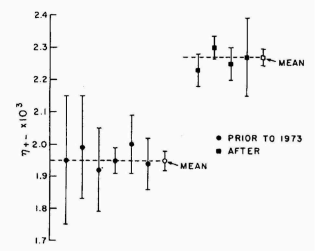


FIG. 1: Measurements of $|\eta_{+-}|$ in order of their year of publication. Reprinted with permission from A. Franklin, "Forging, cooking, trimming, and riding on the bandwagon," Am. J. Phys. **52**, 786-793 (1984), copyright 1984, American Association of Physics Teachers.

The figure above was taken from Jeng (2005), and was originally printed in Franklin (1984): "Forging, cooking, trimming, and riding on the bandwagon".

The figure shows **Confirmation Bias**.

Confirmation bias is the tendency for researchers to continue adapting their results until they agree with some prior belief.

The figure demonstrates the problem nicely. Prior to 1973, there was a consensus on the value that $|\eta_{\pm}|$ ought to hold. However in the early seventies, there was a shift in the consensus: and all observations began to cluster around that particular value.

The pre- and post-1973 distributions of $|\eta_{\pm}|$ are catastrophically inconsistent with one-another. The cause: confirmation bias. Similar effects have been seen in measurements of the speed of light, and in the build-up to the discovery of the Higgs Boson.

Real World Example: the θ^+ penta-quark

Confirmation bias, however, need not require previous measurements. Humans can have a prior belief about a particular result, and simply analyse their data until that result is observed.

Such was the case with the discovery of the θ^+ penta-quark.

In 2002, a japanese lab published the discovery of the θ^+ penta-quark at greater than 5σ significance (a false positive rate of 1 in ~20 million). Subsequently over the next 4 years 11 other research groups searched for and found high-significance detections of the same penta-quark.

However, subsequent searches with more sensitive equipment failed to find any evidence for the penta-quark. In the same year, one group quoted an 8σ detection of the pentaquark, while another group performing the exact same experiment at a different lab with comparable statistical power found *nothing*.

The problem here is that researchers were not **blinded** to their data. They knew the signal that they were trying to detect, and they found it.

As such **blind** analyses are now a staple in many fields within the natural sciences, including cosmology and high-energy particle physics.

Thus we have Method 7 for invalidating your research:

> - Not blinding your analyses, and thereby (sub)consciously modifying your data to reproduce your expectations.

Summary

- Looking for correlations in your data *and then assuming that these correlations are causal*.
- Failing to correct for observational selection functions present in **every dataset**.
- Failing to account for the influence of observational uncertainties on your models and derived data distributions.
- Focussing on extrema, failing to account for the pathological selections, and ignoring the properties global samples.
- Failing to understand, and therefore abusing the meaning of, p-values
- Modifying our analysis samples until we reach 'significance'
- Data dredging, or analysing multiple hypotheses and not accounting for this in our modelling.
- Not blinding your analyses, and thereby (sub)consciously modifying your data to reproduce your expectations.