SFB-1491 Graduate School: Astrostatistics

Section 3: Significance & Modelling

Dr Angus H Wright

2023-02-02

Section 3: Introduction

Parameter Simulation, Optimisation, & Inference

(or "Applying statistics in modern scientific analyses")

We apply our understanding of Bayesian statistics to the common problems of parameter simulation, optimisation, and inference. Students will learn the fundamentals of hypothesis testing, quantifying goodness-of-fit, and parameter inference. We discuss common errors in parameter inference, including standard physical and astrophysical biases that corrupt statistical analyses.

The reason we're here

Our goal in this course is to formulate a basis for performing statistical analyses, in the natural sciences, that you can use for the rest of your academic careers.

To do that, you need to be able to do the following:

> - Be able to explore and understand complex datasets (Section 1)
> - Understand the probabilistic nature of experiments and have access to tools that allow you to estimate models from data (Section 2)
> - Understand how to interpret models/results to perform accurate **statistical inference** (Section 3).

A Significant Conundrum

Modern and future experiments will never produce data that covers the entire population Ω of possible observations.

We will always be attempting to analyses models of variables θ using samples of data, and attempting inference using estimates of θ that are random variables.

As a result, regardless of the experiment being undertaken, it is generally relevant to ask whether or not an observed relationship, parameter estimate, and/or measurement is "significantly" different from previous work and/or expectations from (e.g.) theory.

Said differently: whenever we measure a variable, it is sensible for us to ask whether or not the estimated value is consistent with our model and/or previous estimates, given the expected random fluctuations of a random variable.

A simple demonstration:

Suppose we have a theory that the true average height of all human beings is 184cm.

Measuring the height of every human being is naturally unfeasible, so we are forced to take a sample of n humans and just measure their average height.

This estimate of the average height is a random variable, as it will vary from sample-to-sample.

We require a method for determining whether or not any difference between our estimate of the average height θ and 184cm is caused by random variation due to our sampling, or whether it demonstrates that the **true** average height is unlikely to be 184cm.

A simple demonstration:

One method would be to construct some interval (given the data) within which you expect the *true* value of θ to reside with some (quite high) probability: say 95%.

If you construct this interval and find that our hypothesised value of 184cm resides outside it, then we can draw one of two conclusions:

1. the value of $\theta = 184$cm is unlikely to be correct; **or**
2. we just got very unlucky with our chosen sample.

This procedure provides us with a mechanism for determining whether the data that we have provides evidence to *contradict* a particular hypothesis.

Aside: the merits of contradiction

Why not come up with a measure of whether or not the data **agrees** with some hypothesis?

- How much evidence does it take to prove something is true?
- How much evidence does it take to prove something is false?

This is somewhat the nature of scientific inquiry:

- No amount of evidence can give absolute certainty that a hypothesis is true, it can only fail to show that it is false.
- However you only need one piece of evidence to disprove a hypothesis.

Significance

Given our observed sample of human heights, we want to assess the **significance** of the evidence against our particular hypothesis.

We can do this by calculating the fraction of samples of n humans that would produce a sample mean that is as extreme as the one we observe **if** the hypothesis is true.

Our hypothesis is that the population mean is $\hat{\mu}$, and we observe some mean $\bar{x}$ from our sample of n observations.

We approximate the variance of μ using the variance of our sample s^2, which gives us an estimate of the standard error on $\hat{\mu}$: $s/\sqrt{n}$.

We can then define a new random variable t which we call our *test statistic*:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

Significance

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

This is the **student t-statistic**. The distribution of this variable follows the student t-distribution, which is nicely analytic.

This distribution has an analytic PDF, and means that we can trivially calculate the probability of observing a sample of data that have mean $\bar{x}$ given μ, n, and s.

Recall that the PDF of the student t-distribution is:

$$p_t(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

where $\nu = N - 1$ is the degrees of freedom.

Significance

The fraction of samples that have less extreme values of $\bar{x}$ **if** the true population mean is $\mu = \hat{\mu}$ is:

$$f = \int_{-|t|}^{|t|} p_t(x;\nu)\mathrm{d}x$$

Consider a very positive (or very negative) value of t; in these circumstances, f will be close to 1 (as $-|t| \le x \le |t|$ contains all the probability mass of the distribution).

This indicates that essentially **all** random samples with ν degrees of freedom would have less extreme values of t given the hypothesis that $\mu = \hat{\mu}$.

A more convenient method of formulating this number is to look at its complement:

$$p = 1 - f$$

which is the fraction of samples that have as extreme a value of t given $\mu = \hat{\mu}$. This is known as the **p-value**.

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.
- **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.

The p-value

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

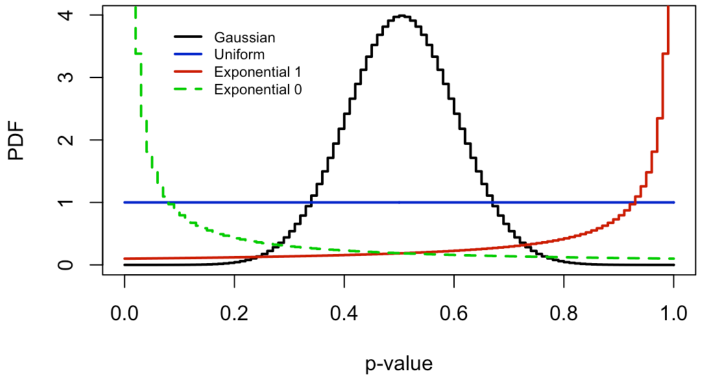
Let's assume that we are willing to believe an effect if it has a p-value of α or less (otherwise you reject the effect in favour of the null hypothesis). The probability that you accept a hypothesis that is actually *false* is the fraction of samples that would give you a `satisfactory' p-value even 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.

Using the p-value

In his original 1920 publication, Fisher used $p < 0.05$ as an example of a value that might be used to justify rejection of the null hypothesis, when taken in the context of the entire experimental landscape.

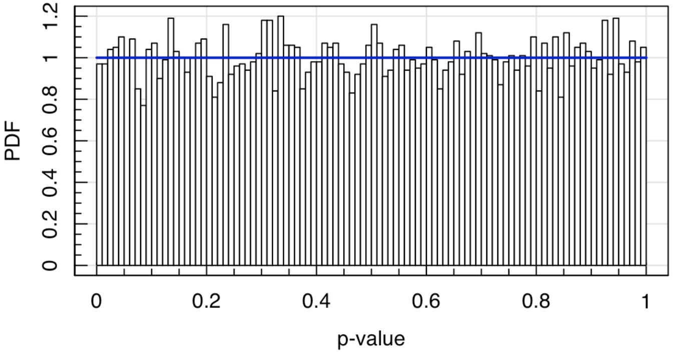
We've said already what the p-value describes. So now a question:

Given purely random measurement bias, what distribution does the p-value take under many realisations of an experiment?

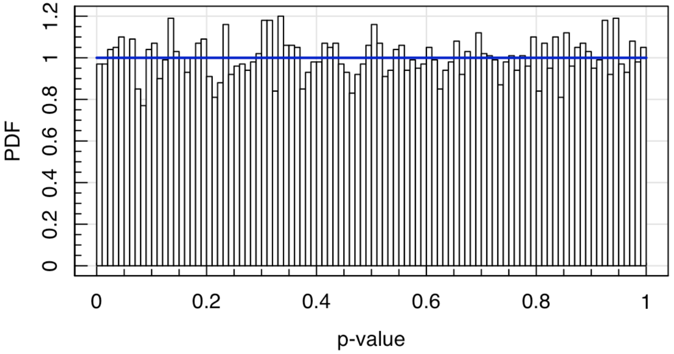


Using the p-value

$$
\begin{aligned}
p &= 1 - \int_{-|t|}^{|t|} p_t(x;\nu)\mathrm{d}x \\
&= 1 - \left[\int_{-\infty}^{|t|} p_t(x;\nu)\mathrm{d}x - \int_{-\infty}^{-|t|} p_t(x;\nu)\mathrm{d}x\right]
\end{aligned}
$$



Using the p-value



Which makes sense; the p-value describes the fraction of samples that have more extreme values than that which we observed, assuming the null hypothesis. If the null hypothesis is true, then we should see a p-value as extreme as α occur $\alpha\%$ of the time.

But this begs an important question: if every scientist were to use $p < 0.05$ as a metric for "a significant result worthy of publication", what fraction of published results ought to be false positives?

How much published research is wrong?

Let's assume that we're looking at a field of research where there are N ongoing experiments, all exploring different possible physical relationships. Of those N experiments, $f_{\rm true}$ of them are real physical relationships. Finally, each experiment has a **statistical power** of $s_{\rm p}$.

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?

How much published research is wrong?

$$
P(\mathrm{False}|\mathrm{pub}) = \frac{p \times f}{p \times f + s \times (1-f)}
$$

So if $p = 0.05$, $f = 0.9$, and $s = 0.8$:

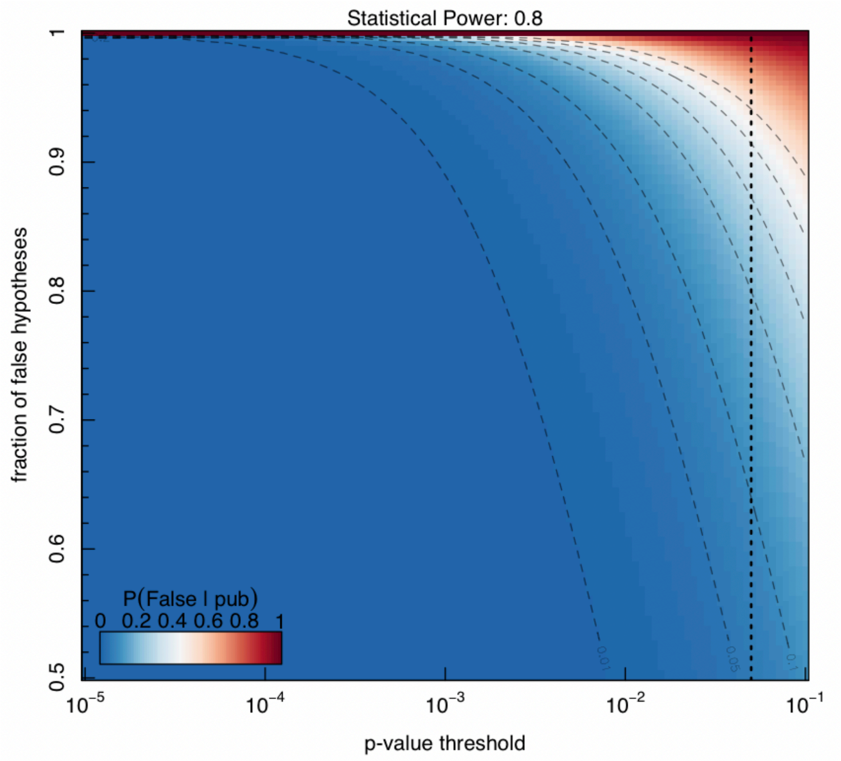
$$
\begin{aligned}
P(\mathrm{False}|\mathrm{pub}) &= \frac{0.05 \times 0.9}{0.05 \times 0.9 + 0.8 \times (1-0.9)} \\
&\approx 0.36
\end{aligned}
$$

Pratical Statistical Inference

So it's clear that the use of a standard threshold for p-values as a measure of significance can lead to problematic numbers of incorrect results being published in the literature.

However, as we just saw, this effect can be calculated simply. So why is it a problem, provided that we can easily demonstrate the effect, and so account for it?

Why not, for example, use the high-energy physics mantra of $p < 0.001$ and be done with it?



The Problem is Choice

At it's simplest level, when provided an arbitrary dataset, our statistical analyses will involve two steps:

1. **Data mining**: where we explore and summarise the data; and
2. **Data modelling**: where we extract model parameters/trends, and test hypotheses.

However in reality each of these steps involves many stages: With experimental data:

- Samples must be defined;
- Observations must be taken;
- Defective data must be identified and removed; and more.

When modelling the data:

- formulate our hypotheses;
- construct the likelihood;
- perform our inference; and more.

In this lecture, we're going to explore some of the dangers inherent to these processes. We will establish some of the fundamentals of hypothesis testing, specifically with respect to determining the significance of evidence.

However:

While we could simply go through the definitions, standards, and best practices for determining the significance of evidence, I think it is more educational (and shocking, and fun) to go about this from the *opposite* direction.

As such, now we're going to discuss…

Bad Statistics:

(Non-exhaustive) Examples of what **not** to do

We will use simulated data and real experiments to show how poor use of statistics can lead to pathologically incorrect conclusions, in a (hopefully light-hearted!) effort to demonstrate the pitfalls that careless scientists can find themselves falling into.

This discussion of bad statistics will focus on a few main areas:

> - Variable Selection
> - Sample Selection
> - Data Modification
> - Additional Observation
> - Confirmation

Importantly: for the sake of this lecture, we are going to completely ignore the concept of confounding variables (which we spoke about at the beginning of the lecture course). This effect, in reality, makes much of what we are about to discuss *much* worse.

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.

```
##              V1            V2          V3            V4           V5           
V6           V7
## 1    1.37095845  1.200965e+00 -2.00092924 -0.0046207678  1.334912585  1.0291407
19 -0.248482933
## 2   -0.56469817  1.044751e+00  0.33377720  0.7602421677 -0.869271764  0.9147748
68  0.422320386
## 3    0.36312841 -1.003209e+00  1.17132513  0.0389909129  0.055486955 -0.0024562
67  0.987653294
## 4    0.63286260  1.848482e+00  2.05953924  0.7350721416  0.049066913  0.1360095
52  0.835568172
## 5    0.40426832 -6.667734e-01 -1.37686160 -0.1464726270 -0.578355728 -0.7201535
45 -0.660521859
## 6   -0.10612452  1.055138e-01 -1.15085557 -0.0578873354 -0.998738656 -0.1981243
30  1.564069493
## 7    1.51152200 -4.222559e-01 -0.70582139  0.4823694661 -0.002432780 -1.0292088
06 -1.622975935
```

We have theoretical expectations of what the data ought to show for each of our variables, which we have already subtracted from each column.

Variable Selection

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
```

```
## 
##  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…

Variable Selection

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
```

Variable Selection

… 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
```

Variable Selection

… the fourteenth…

```
##    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
```

Variable Selection

… 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.

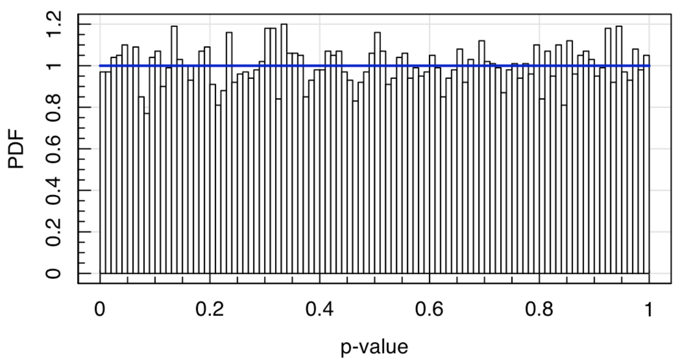
We write up our discovery, publish the result, and our discovery is enshrined in the literature forever.

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 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:



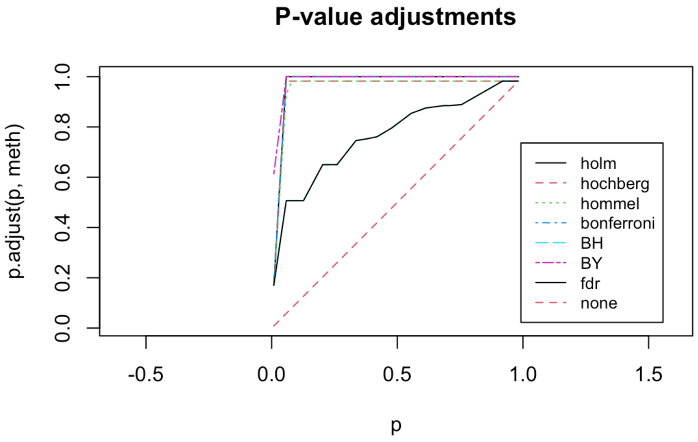
We have used in this example a threshold of $p < 0.05$. We therefore expect to find this p-value given random fluctuations in 1 out of every 20 cases. In our example we have 20 variables. So it makes sense that we found a "significant" effect for 1 variable.

How can we combat this?

We can correct for this effect by modifying the threshold that is required for determining “significance”, to account for the fact that many variables are under analysis.

The simplest example is the **Bonferroni correction**, which simply states that the threshold for significance when analysing m different variables ought to be $\alpha' = \alpha/m$.

However there are many possible corrections. In **R** there are a number of them inbuilt, which we can run over our simulated data:



Real World Example: 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 (or 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.

Real World Example: An Empathetic Fish

The researchers placed the fish in an MRI, and presented it with images of humans and other pictures.

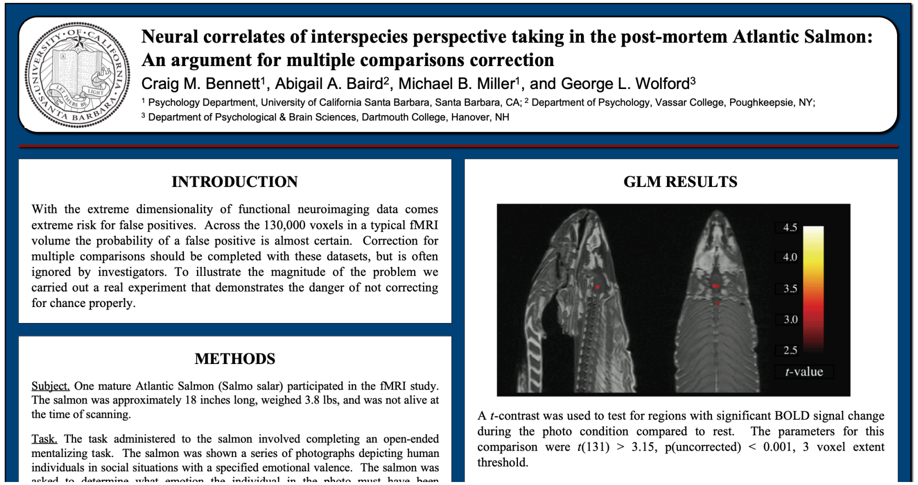
They analysed the data using standard processing tools, and found a significant discovery of activity in the brain of the salmon that correlated with the researchers presenting the fish with images of humans, as opposed to objects.

The problem?

Real World Example: An Empathetic Fish

The salmon was frozen at the time of study

It was a frozen Atlantic salmon, bought from a fish-monger.



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 asked to determine what emotion the individual in the photo must have been

GLM RESULTS

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.

Data Modification

Data modification can take a number of forms, however the most common are selecting specific subsets of data and/or rejecting certain portions of the data that are deemed to be “outliers”.

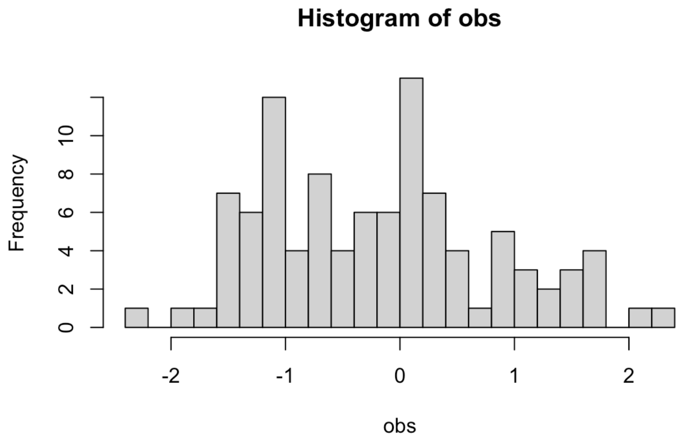
Data modification need not be malicious, or even intentional. At its weakest, we may simply discard data that we expect to be outliers.

At its most malicious, it involves hand-selecting data that suit your hypothesis. These processes are generally referred to as **cherry-picking**.

Data Modification

Suppose now that we set a more strict requirement on our p-value, $p < 0.01$, and that this is the first variable that we looked at (so no modification to our threshold is required).

We’re not quite there with our dataset:



```
##
##  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
```

But what about those two pesky data points at ~ 2?

Maybe we can convince ourselves that one of those is an error, because of something that went wrong in our experiment? We convince ourselves to drop one of those data points (after all, it's only 1% of the data!). What happens to our p-value?

```
##
##  One Sample t-test
##
## data:  obs
## t = -2.1162, df = 98, p-value = 0.03686
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.4028849 -0.0129442
## sample estimates:
##  mean of x
## -0.2079145
```

Off to the journal we go!

Data Modification

This is an example of cherry picking that is *very* easy for researchers to fall into.

This is because experimental data is messy; it's easy to fool yourself into thinking random fluctuations are bad data, and thereby justify their removal.

This has a significant influence on determinations of significance, though, as we've just seen.

Real World Example: Climate Denialism

The practice of **maliciously** cherry-picking data is also an important one to understand.

This is the realm of people who wish to use statistics to push an agenda, and one of the most common places to find examples of this practice is in climate change denialism.

In an effort to provide evidence that the globe is not warming, one climate change denier claimed in a newspaper article in 2011 that:

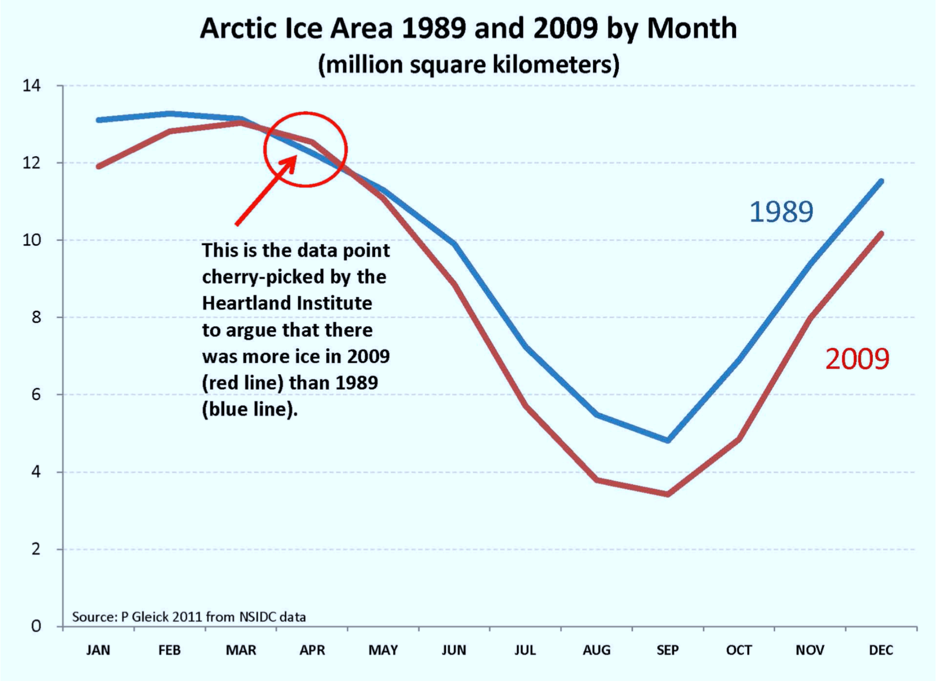
"In fact, National Snow and Ice Data Center records show conclusively that in April 2009, Arctic sea ice extent had indeed returned to and surpassed 1989 levels."

The implication of this statement is that there is no cause for alarm because there is no **systematic** reduction in sea ice between the two years.

The assumption being that the lack of difference in April can be used to infer systematic difference over the whole year (or longer).

Real World Example: Climate Denialism

Can you see the problem with this argument?



This is the data point cherry-picked by the Heartland Institute to argue that there was more ice in 2009 (red line) than 1989 (blue line).

Source: P Gleick 2011 from NSIDC data

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:

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
```

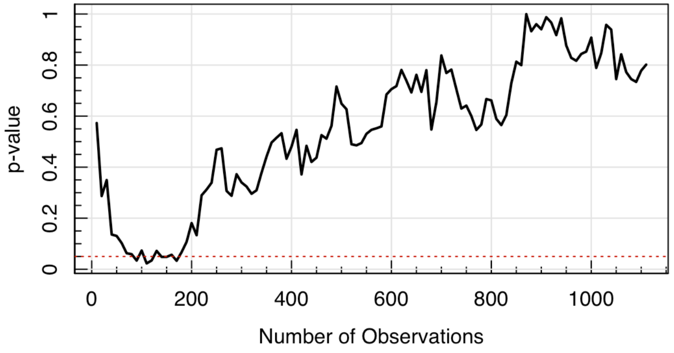
We now decide to observe more data, in a batch of 10 observations.

```
##
##  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.

Simulating the effect:

But what happens if we were to continue observing data?



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.

Simulating the effect:

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.

Confirmation

Finally, we consider the influence of conscious and subconscious human biases on measurements.

Experiments do not happen in windowless rooms in the depths of space. They are performed by human researchers who work in laboratories, and have a keen understanding of the *context* in which their experiment takes place.

In our discussion of bayesian statistics, we formulated this as a **good** thing.

The prior knowledge that we bring to an experiment can play an important role in improving our statistical inference.

However there is a dark side to prior knowledge: the (generally sub-)conscious drive to be “consistent”.

Confirmation bias

The last significant statistical fallacy that we will discuss today is one that is *extremely* important: **confirmation bias**.

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

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

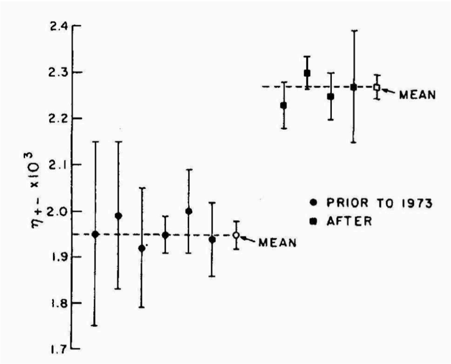


FIG. 1: Measurements of $\mid \eta_{+-} \mid$ in order of their year of publication. Reprinted with permission from A. Franklin, “Forging, cooking, trimming, and riding on the bandwagon,” Am. J. Phys. **52**, 786-793 (1984), copyright 1984, American Association of Physics Teachers.

Confirmation bias

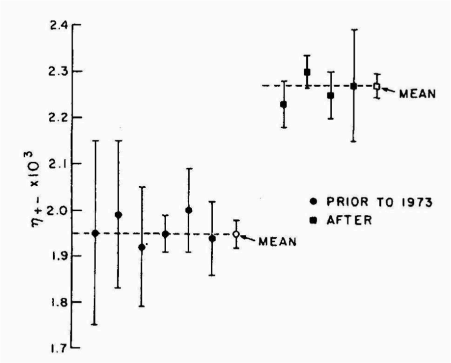


FIG. 1: Measurements of $\mid \eta_{+-} \mid$ in order of their year of publication. Reprinted with permission from A. Franklin, “Forging, cooking, trimming, and riding on the bandwagon,” Am. J. Phys. **52**, 786-793 (1984), copyright 1984, American Association of Physics Teachers.

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 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 new, different, 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 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.

That is: 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.

What have we learned

This has been an incomplete discussion of statistical fallacies. There are many more. Notable omissions include:

> - Regression to the mean
> - Spurious Correlation
> - Survivor Bias

Generally, the lesson here is to be very sceptical of using a p-value as a mechanism for determining whether or not something is “interesting”, or “significant”.

Bayesian Hypothesis Testing

As Bayesian statistics is concerned with determining estimates of underlying model parameters given the data, model comparison and hypothesis testing between different models becomes a natural extension of standard Bayesian methods.

Take the simplest possible example: > + Simple null vs Simple Alternative hypotheses: $H_0 = \theta_0$ vs $H_1 = \theta_1$.

We have two hypotheses about the model that generates our data: $H_0 = \theta_0$ vs $H_1 = \theta_1$.

These hypotheses are mutually exclusive and exhaustive (that is, $\{H_0, H_1\} = \Omega$). Next assume we have some appropriate test statistic $T = T(X_1, \dots, X_n)$.

Bayesian Hypothesis Testing

By Bayes Theorem, we have:

$$P(H_0|T) = \frac{P(T|H_0)P(H_0)}{P(T|H_0)P(H_0) + P(T|H_1)P(H_1)}$$

Given that the hypotheses are mutual exclusivity and exhaustive:

$$P(H_1|T) = 1 - P(H_0|T)$$

so

$$\frac{P(H_0|T)}{P(H_1|T)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(T|H_0)}{P(T|H_1)}$$

This is the **posterior odds ratio**, and the last ratio is known as the **Bayes factor**.

Notice that, therefore, if the prior odds ratio is unity (i.e. that $\frac{P(H_0)}{P(H_1)} = 1$), then the posterior odds equals the Bayes factor.

Jeffery’s Hypothesis Tests

“If the posterior odds ratio exceeds unity, we accept H_0. Otherwise, we reject H_0 in favour of H_1.”

The **Jeffreys Hypothesis testing criterion** above has a few important benefits over classical methods of hypothesis testing.

- There is no specification of a “significance level” that determines whether or not a hypothesis is accepted/rejected.
- It is easily generalisable to many many hypotheses: you just accept the one with the highest posterior probability.

There is one important philosophical difference as well: in accepting H_0 as the preferred model, we do not assume that it is the *true* model.

We are simply stating that, with the currently available data, H_0 is the more probable alternative.

Jeffery’s Hypothesis Tests

The “Jeffreys Scale” gives a slightly larger dynamic range to the amount of evidence that is encapsulated in the posterior odds ratio:

$\frac{P(H_0\|T)}{P(H_1\|T)}$	**Strength of evidence**

$< 10^0$	Negative (supports H_1)
10^0 to $10^{\frac{1}{2}}$	Barely worth mentioning
$10^{\frac{1}{2}}$ to 10^1	Substantial
10^1 to $10^{\frac{3}{2}}$	Strong
$10^{\frac{3}{2}}$ to 10^2	Very strong

Bayesian Model Comparison

Simple Null vs Simple Alternative

Suppose we have $X|\theta \sim N(\theta, 1)$, and $H_0 : \theta = 0$ vs $H_1 : \theta = 1$.

We observe the random sample $X_1, \ldots, X_n$, and form the sufficient test statistic $T = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

We have $T|H_0 \sim N\left(0, \frac{1}{N}\right)$ and $T|H_1 \sim N\left(1, \frac{1}{N}\right)$.

Assume a priori that we have no prior preference over the models $P(H_0) = P(H_1) = \frac{1}{2}$.

Therefore the posterior odds ratio is:

$$\frac{P(H_0|T)}{P(H_1|T)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(T|H_0)}{P(T|H_1)}$$

$$= \frac{0.5}{0.5} \times \frac{\left(\frac{N}{2\pi}\right)^{\frac{1}{2}} \exp\left(-0.5N\bar{X}^2\right)}{\left(\frac{N}{2\pi}\right)^{\frac{1}{2}} \exp\left[-0.5N(\bar{X}-1)^2\right]}$$

$$= \exp\left[-0.5N(2\bar{X}-1)\right]$$

Simple Null vs Simple Alternative

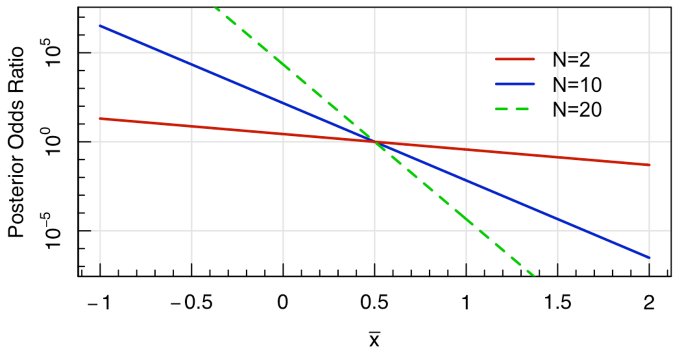
Let’s now look at some simulated data:

```
x1<-rnorm(10,mean=1.2)
x2<-rnorm(10,mean=0.2)
post_ratio<-function(x) exp(-0.5*length(x)*(2*mean(x)-1))
print(post_ratio(x1)); print(post_ratio(x2))
```

```
## [1] 0.02911078
```

```
## [1] 1.127367
```

We can now look at what the posterior odds ratio looks like, in this case, for a range of means and values of
N:



Bayesian Modelling

The previous examples of how to calculate model preferences is all well and good, but this is where the magic happens.

Because there is some uncertainty in expressing/specifying any single model:

$$f(\theta, x) = f(x|\theta)f(\theta)$$

we can instead construct a single model that we define as being the union of all alternative models that we might wish to entertain.

We will then provide a prior over the suite of encompassed models.

Bayesian Modelling

Take an example where we have two models that we think might be appropriate for our dataset, both of which fall within the general “Gamma” family of distributions.

Recall that the Gamma family of distributions all take the format:

$$f(x|\alpha, \beta, \gamma) = \frac{\gamma\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha\gamma-1} \exp(-\beta x^{\gamma})$$

Our two hypothesised models are a Weibull distribution:

$$f_1(x|\beta, \gamma) = \gamma\beta x^{\gamma-1} \exp(-\beta x^{\gamma})$$

and a two-parameter Gamma distribution:

$$f(x|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

We can analyse these two models in the same way as previously. Require that these two hypotheses be exhaustive ($P(m_1) = 1 - P(m_2)$), and formulate the values of α, β, γ.

Bayesian Modelling

However, we could alternatively specify a single **encompassing** model that is just a generalised gamma distribution.

This distribution contains both of the previous 2 models *and many many more*.

Nominally it is no more or less sensible to formulate our model comparison using priors on α, β, γ instead of on m_1, m_2, and we can construct priors that recover the behaviour of having only the two models in any case:

$$f(\alpha, \beta, \gamma) = \begin{cases} f(\alpha, \beta, \gamma) & \text{if } \alpha = 1 \\ f(\alpha, \beta, \gamma) & \text{if } \gamma = 1 \\ 0 & \text{otherwise} \end{cases}$$

What did we just do?

We just demonstrated that we can perform model comparison within the bayesian framework by specifying a generic model and providing priors on the parameters that govern that model.

In this way, the likelihood that we specified was general: we didn't pick **particular values** for the models in the likelihood, rather we specified a **distribution of possible likelihoods** and gave (possibly broad) priors on the variables that govern the distribution of possible models.

This leads us to an interesting class of models known (appropriately) as **Bayesian hierarchical models** (BHMs).

Start small

As a demonstration of the power of BHMs, we're going to take an initially simple model, and with few logical steps, construct an much more complex model that has exceptional explanatory power.

The dataset that we're going to explore today to demonstrate this process is one from the US, where 8 high-schools reported on the possible benefits of giving students additional coaching prior to the SAT-V ("scholastic aptitude test - verbal") exams.

Under randomisation, students at each school were given either extra coaching or not. SATs are designed to be resilient to short-term efforts, however all schools think that their program is useful/effective nonetheless.

There's no prior reason to expect that any program is more or less effective than the others.

The data

We have $J = 8$ independent experiments, with coaching effects θ_j being judged by n_j i.i.d. normally distributed observations y_{ij}, each with (assumed) known error variance σ^2.

That is:

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2), \quad i = 1, \dots, n_j; \quad j = 1, \dots, J$$

The estimated effect of the coaching at each school is given by the mean $\bar{y}_j$, with the standard error on the estimate σ_j^2.

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

$$\sigma_j = \sqrt{\frac{\sigma^2}{n_j}}$$

The data

The likelihood for each θ_j can be expressed in terms of the sufficient statistic:

$$\bar{y}_j|\theta_j \sim N(\theta_j, \sigma_j^2)$$

```
dat<-data.frame(Estimated_Effect=c(28,8,-3,7,-1,1,18,12),
                Standard_Error=c(15,10,16,11,9,11,10,18))
rownames(dat)<-LETTERS[1:8]
print(dat)
```

```
##   Estimated_Effect Standard_Error
## A               28             15
## B                8             10
## C               -3             16
## D                7             11
## E               -1              9
## F                1             11
## G               18             10
## H               12             18
```

Methods of analysis

```
print(dat)
```

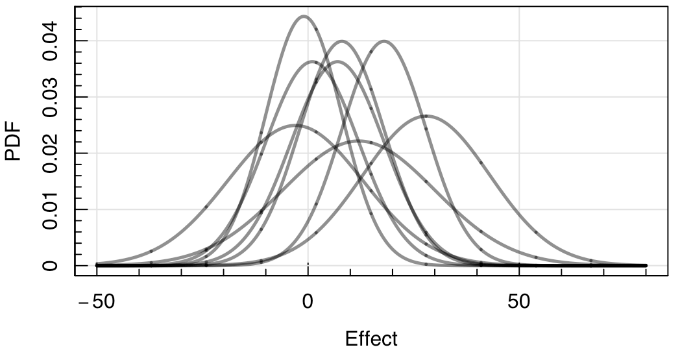
```
##   Estimated_Effect Standard_Error
## A               28             15
## B                8             10
## C               -3             16
## D                7             11
## E               -1              9
## F                1             11
## G               18             10
## H               12             18
```

Each to their own

There are multiple ways that we could approach the modelling of this dataset. The first option is to treat each experiment independently, as the data have been provided. We will call this the **separate** analysis.

At first glance, there is a mixed-bag of results. Some schools show reasonably large effects ($\theta_j \geq 18$), some show small effects ($0 \leq \theta_j \leq 12$), and some show small negative effects.

However, each estimate also has a large standard error. This makes it difficult to distinguish between the different results. The 95% posterior credibility intervals for these results all significantly overlap.



Methods of analysis

```
print(dat)
```

```
##   Estimated_Effect Standard_Error
## A               28             15
## B                8             10
## C               -3             16
## D                7             11
## E               -1              9
## F                1             11
## G               18             10
## H               12             18
```

All together now

The large overlap between the individual credible intervals might suggest that all of the experiments are trying to measure the same underlying quantity.

So we might prefer to assume that $\theta_j = \theta_0 \; \forall j \in \{1, \ldots, J\}$.

That is, that all the values of θ_j are the same. Given this hypothesis, we can estimate the value of each θ_j using the **pooled** average $\bar{y}$.

Said differently: assuming that all experiments have the same effect (and produce random independent estimates) then we can treat the 8 experiments as a i.i.d. observations from the underlying truth, with known variances.

All together now

We can estimate this quantity simply:

$$\bar{y} = \frac{\sum_{j=1}^{J} w_j \bar{y}_j}{\sum_{j=1}^{J} w_j}$$

where $w_j = 1/\sigma_j^2$.

```
ybar<-with(dat,{
  weight<-1/Standard_Error^2
  return(sum(weight*Estimated_Effect)/sum(weight))
})
print(ybar)
```

```
## [1] 7.685617
```

The variance of this estimate is the inverse of the sum of the weights:

$$\text{var}(\bar{y}) = \frac{1}{\sum_{j=1}^{J} w_j}$$

```
var_ybar<-with(dat,{
  weight<-1/Standard_Error^2
  return(1/sum(weight))
})
print(var_ybar)
```

```
## [1] 16.58053
```

All together now

So we have an estimate of $\theta_j \sim N(7.69, 16.58)$.

Does this seem reasonable?

Take the experiment at school A as a test case:

- In the independent analysis: $\hat{\theta}_1 = 28$ and $\hat{\sigma}_1 = 15$.

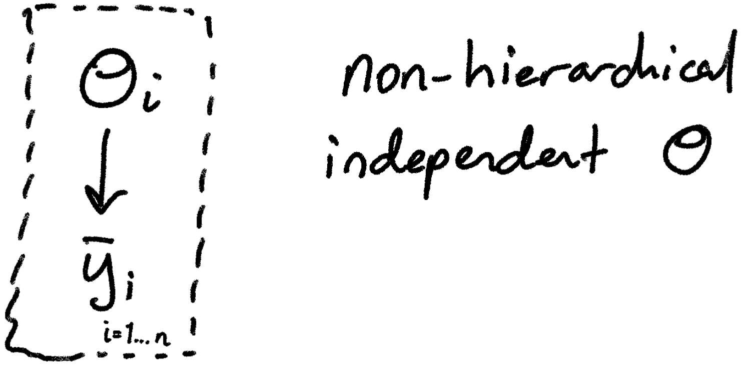
- In the pooled analysis: $\hat{\theta}_1 = 7.69$ and $\hat{\sigma}_1 = 4.07$.

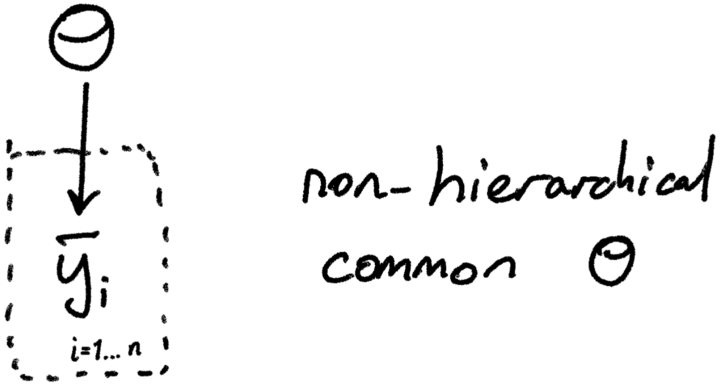
The first estimate tells us that the probability of the true θ_j being greater than 28 is $\frac{1}{2}$.

Conversely, the latter estimate tells us that probability of the true θ_1 is less than 7 is also $\frac{1}{2}$.

A Hierarchical Model

We can display our two previous models as **directed acyclic graphs**:



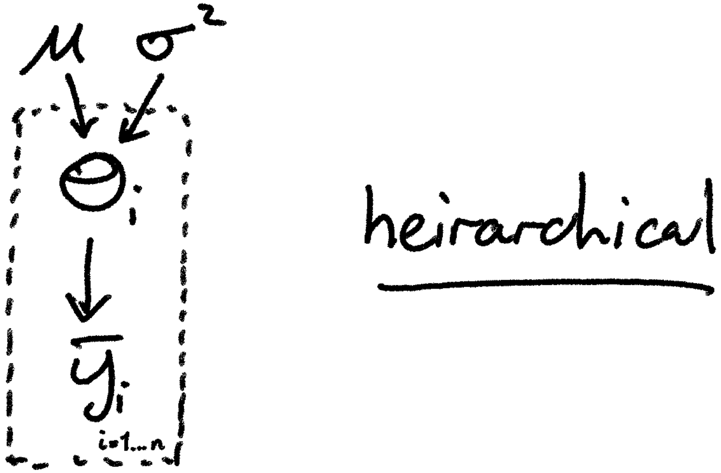


These show how the variable we observe ($\bar{y}_i$) is related to the parameter of interest θ.

In the first instance (i.e the separate estimates), we assumed that each school observed a totally independent θ_i. In the second case (i.e. the pooled estimate), however, we assumed that θ was a constant.

A Hierarchical Model

Let's now instead assume that the values of θ_j are drawn from a normal distribution. The properties of the normal distribution we will determine with two **hyper-parameters** (μ, τ).



Mathematically, we are defining the joint probability of all our θ_i as the product of the probabilities of observing the data, given that each θ_i is drawn from a parent population $N(\mu, \tau)$.

$$
\begin{aligned}
f(\theta_1, \ldots, \theta_J | \mu, \tau) &= \prod_{j=1}^{J} N(\theta_j | \mu, \tau^2) \\
&= \int \prod_{j=1}^{J} \left[N\left(\theta_j | \mu, \tau^2\right)\right] f(\mu, \tau) \mathrm{d}(\mu, \tau)
\end{aligned}
$$

A Hierarchical Model

This is a **hierarchical model**, which can interpret the θ_j's as being randomly drawn from some shared parent distribution. Why is this useful?

We initially had the problem of determining whether or not to choose the independent or pooled estimate. However in our hierarchical model:

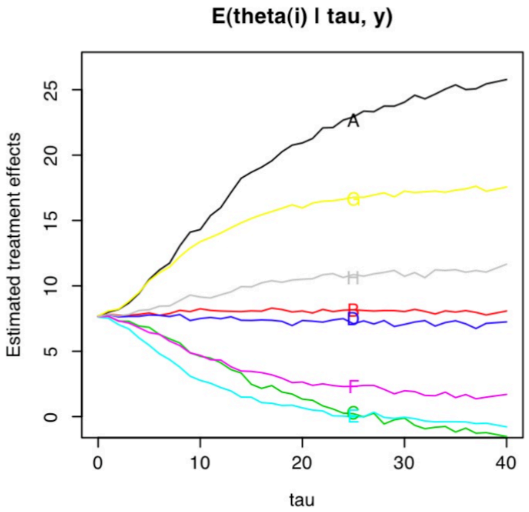
1. As $\tau \to 0$: the θ_j values are drawn from a narrower and narrower range around μ. In the limit, $\theta_j = \mu \ \forall j \in \{1, \ldots, J\}$, and so we have the **pooled estimate**.
2. As $\tau \to \infty$: the θ_j values become independent of each other. That is, if we know θ_1 with absolute certainty, this gives us no information about θ_2. This is therefore the **independent/separate estimate**.

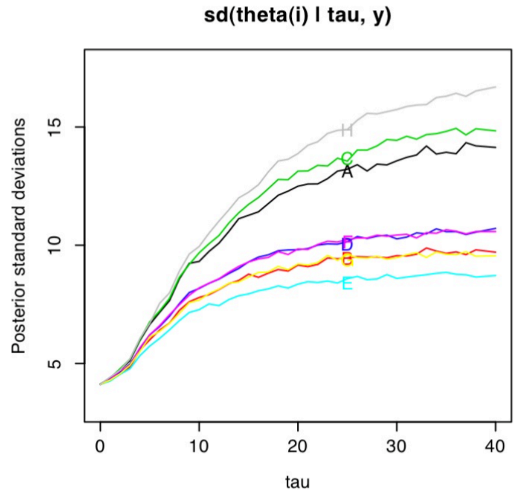
For finite, non-zero values of τ, our result will therefore be some **mixture** of the pooled and separate analyses.

Important implications

The smaller τ, the more related are the individual values of θ_j.

This means that they contribute more to the estimates of θ_j for the other experiments: the experiments "borrow strength" from one-another.


E(theta(i) I tau, y)


sd(theta(i) I tau, y)