[14] Probability and Statistics (3/15/18)

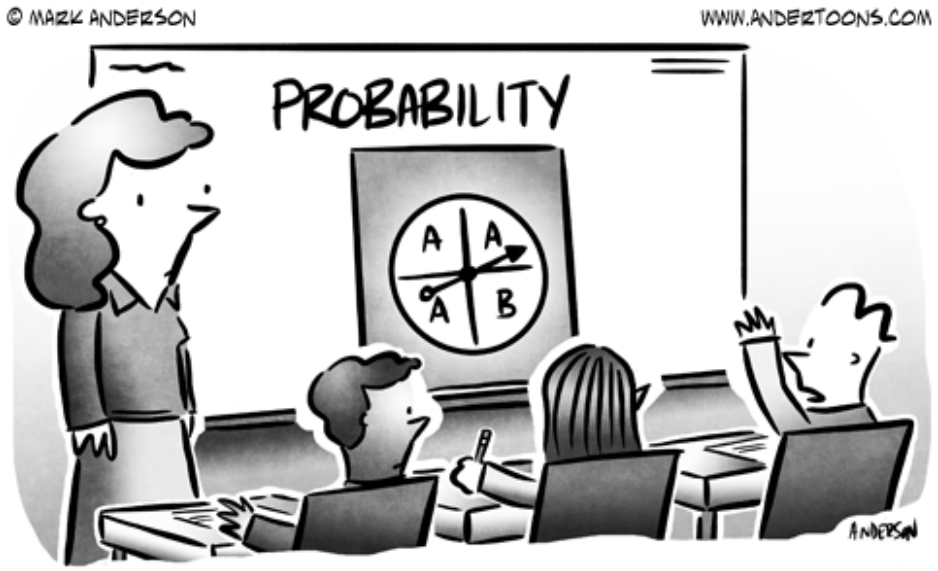
Upcoming Items

1. Read Ch. 19.4 for Tuesday, March 27 class and do the self-study quizzes.
2. Enjoy your spring break!

Learning Goals

For this class, you should be able to:

- Define ”probability”
- Compute probability using permutations and combinations
- Calculate the median, mode, and arithmetic mean
- Understand Gaussians
- Understand how to avoid statistical sins!



"I know mathematically that A is more likely, but I gotta say, I feel like B wants it more."

Thanks for your feedback!

- Many of you asked for practice problems, with solutions, related to our exams
- Okay!
- I have a new directory in Files (on the ELMS site): Files->practice
- One file in there now: practice1.pdf, related to our first midterm. I'll put another file there prior to our second midterm.
- Please read these carefully!
- Because many of you expressed concern about derivations, please also read the Files->derivations documents, and my solutions to both the homework and exam problems; those provide worked examples
- In particular, note *how* I approach these problems; I write these to help you learn how to do astrophysics

Any prob/stats questions?

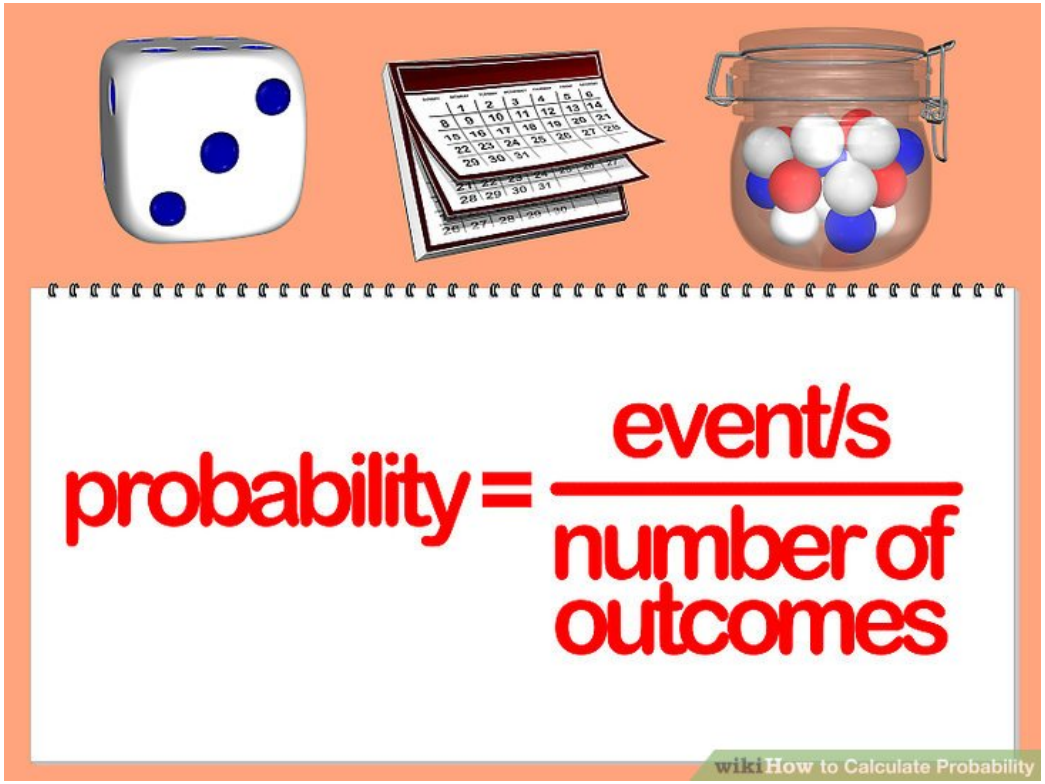
- Special deal this time; we can also address our usual general astronomy questions, but let's do prob/stats first

Before we start...

- To set up a discussion later, we'll do an exercise that we can evaluate using probability
- Yesterday I went to a site that produces sun-sign horoscopes, and in particular one that indicated how yesterday should have gone, given your sun sign
- I have put all 12 possibilities on the sheets being distributed, but I have taken off their labels and have randomized their order
- Please (1) put your birthday (just month and day, not year) on the front of the sheet, and (2) circle the *single* description that best fit your day yesterday (look at both sides, please!) Even if none are close, please pick the closest one
- The randomness hypothesis is that the probability of a match is 1/12. We need to decide, as a class, how probable a match would be if astrology is correct (allowing for some errors) What value do you choose?
- We'll do a calculation related to this later!

What is probability? Discrete version...

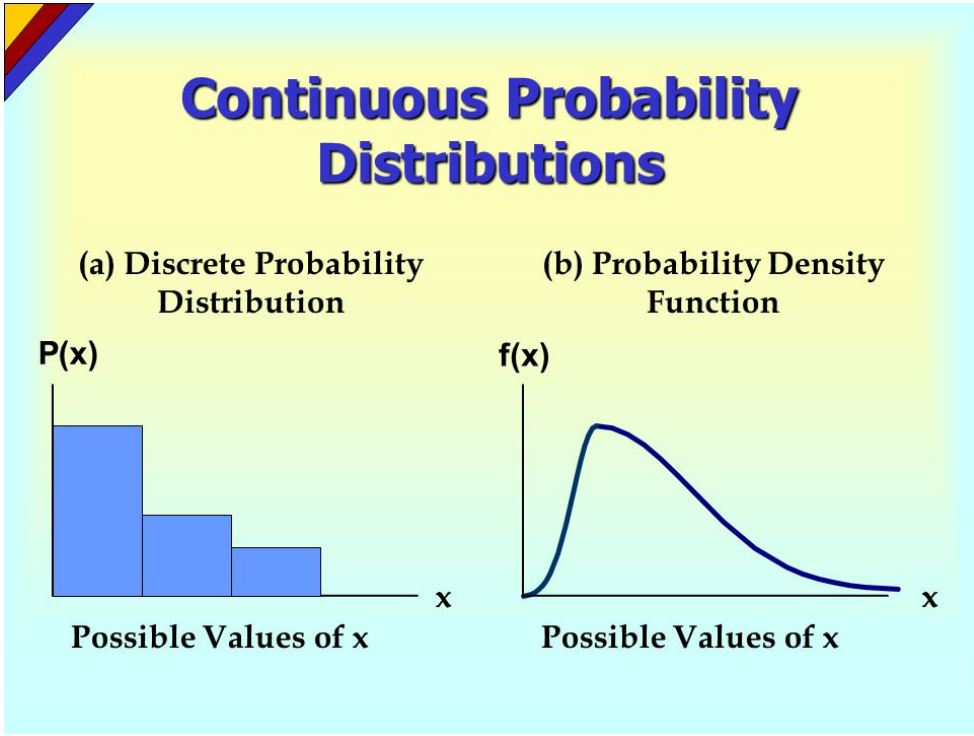
- Suppose you have a list of possible outcomes of an experiment
  E.g., you roll a die; you can get 1,2,3,4,5,6
- You are interested in a particular outcome, e.g., that you roll a 4
- One definition of prob: you roll the die many times; what is the fraction of rolls that give a 4?
- The probability of *something* happening is 1, so the sum of the probabilities of all possible outcomes is 1



https://www.wikihow.com/Calculate-Probability

What is probability? Continuous version

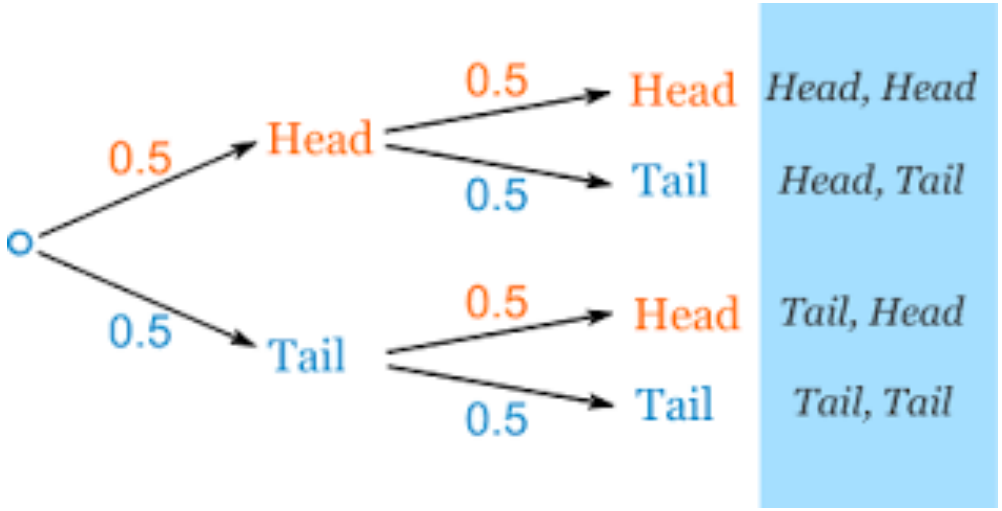
- If the thing you're measuring can take on a continuous set of values (e.g., temperature, length, mass), you need a different definition.
- Say we measure a quantity x
- The probability of measuring x between x_0 and x_0+dx is infinitesimal if dx is infinitesimal
- So we can represent that probability as $P(x_0)\,dx$
  $P(x_0)$ is ***not*** infinitesimal
- Now the ***integral*** of P(x) over all possible values of x must equal 1



http://slideplayer.com/slide/5710846/

Probability by enumeration

- Sometimes you can just count! How many possibilities are there, and of those, how many result in your specified outcome?
- In the example below, we ask: if you have a fair coin, what is the probability of getting one head and one tail in two flips? The mathematician d'Alembert got this wrong!



https://www.mathsisfun.com/data/probability-tree-diagrams.html

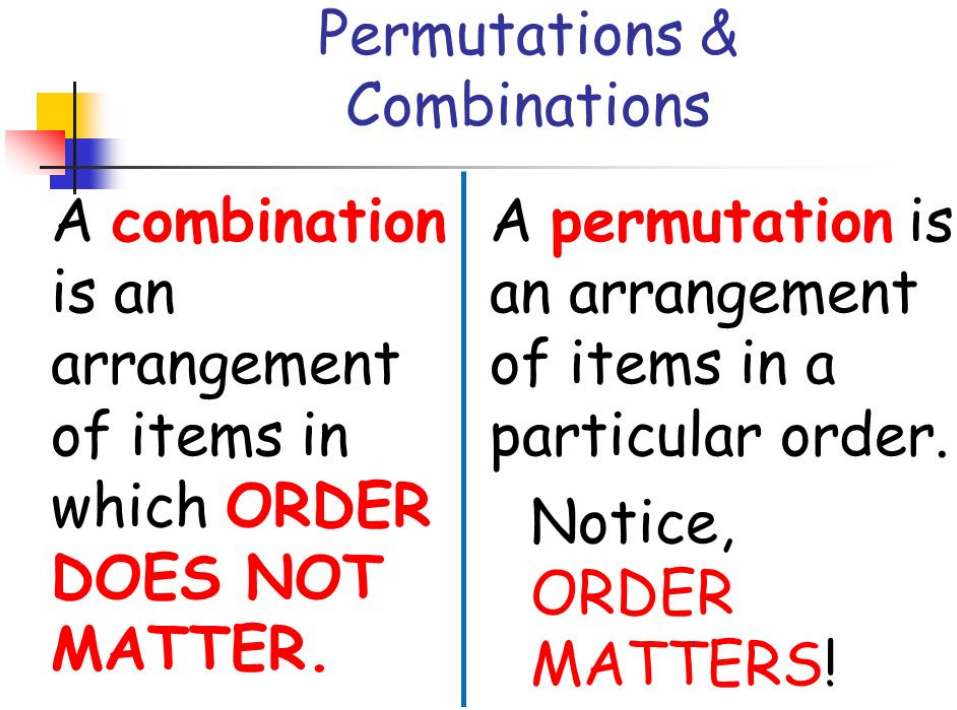
The gambler's fallacy

- If a set of events is *independent*, then past results have no influence on future results
- If I flip a fair coin ten times and get heads each time, that does ***not*** increase the probability of tails the next time!
- Confusion may come from the "law of large numbers" (relative frequency tends toward the probability with many tries)

http://factsongambling.com/gamblers-fallacy/

Permutations and Combinations

- For when you don't want to count every possibility...
- Permutations are when order matters
- Combinations are when order *doesn't* matter
- n objects; there are n*(n-1)*(n-2)*...*2*1=n! permutations
- Choosing m objects from n possibilities, order unimportant-> n!/[m!(n-m)!] combinations



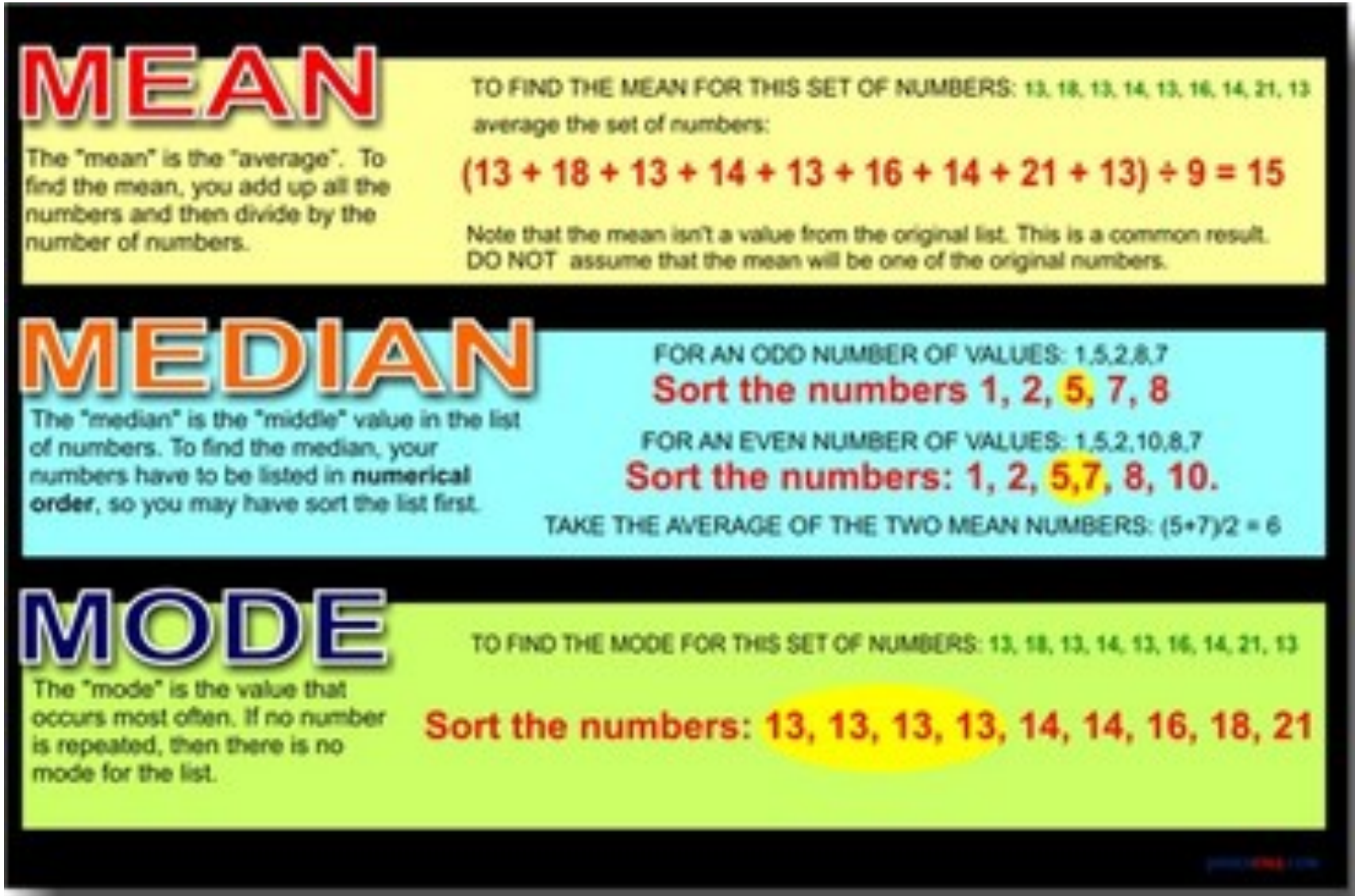
http://slideplayer.com/slide/7326242/

Group Q: Astrology vs. randomness

- We have been told the number of matches (correct astrological sign) is 3, out of 37 total students in the class today
- We have two hypotheses to compare:
  1. Random; probability of a match per selection is 1/12
  2. Astrological; probability of a match per selection is 0.8
- For each hypothesis, calculate the probability that we would have had 3 matches out of 37 total students
  Do we use permutations or combinations?
  What probabilities do we put in?
- Finally, we take the ratio of those probabilities to judge between the hypotheses
- Good luck!

Any questions about probability?

Statistics: mean, median, and mode



http://www.posterenvy.com/mean-median-mode-math-poster/

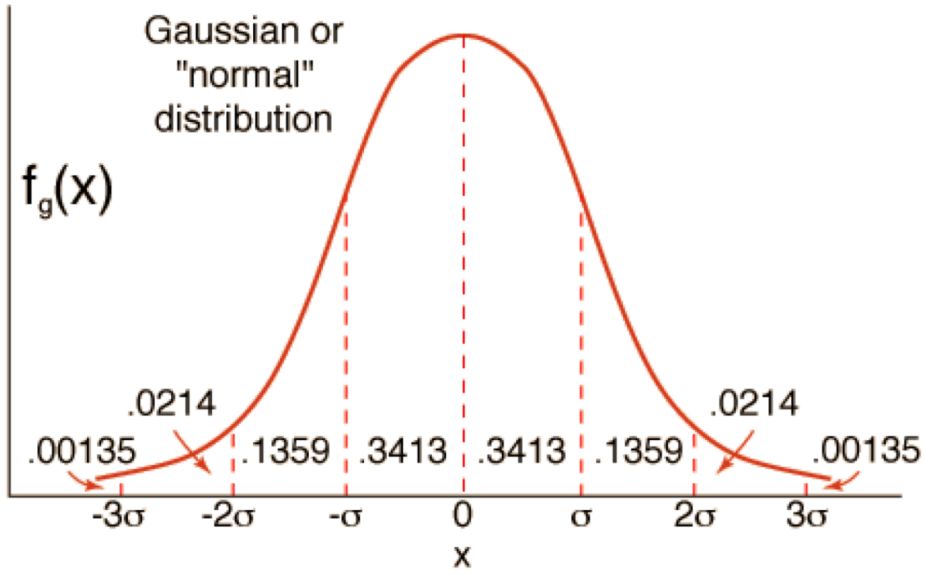
Gaussian, or "normal", distribution

- Also called a bell curve
- Very common, but not universal!
- The uncertainties you see quoted on measurements are often Gaussian
  Thus the assumption is that the probability drops off from the most probable value, in a Gaussian distribution
- The distribution of the *average* of many measurements tends to a Gaussian, almost no matter what the parent distribution is
- But if you don't have lots of measurements, this is not guaranteed...

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

μ = Mean

σ = Standard Deviation



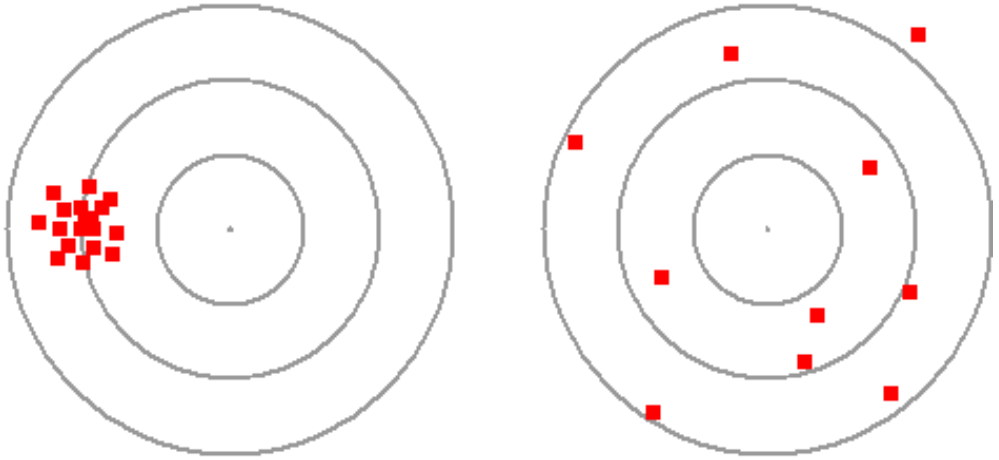
http://hyperphysics.phy-astr.gsu.edu/hbase/Math/gaufcn.html

Some statistical sins

- In many astronomy papers you'll see terrible offenses against statistics
- This actually means that (1) the authors didn't know what they were doing, and (2) neither did the referee!
- But this is an opportunity for you: it means that if you don't make these broad mistakes, you can stand out of the crowd

Sin 1: Systematic errors

- Astronomical instruments and measurements are complicated; rarely do we understand them perfectly
- What if a bump in a spectrum is the result of your lack of perfect knowledge of your instrument?
- Don't revise astrophysics unless you've done lots of tests!
  For example, do you see that bump in other observations?



Systematic Error Statistical Uncertainty

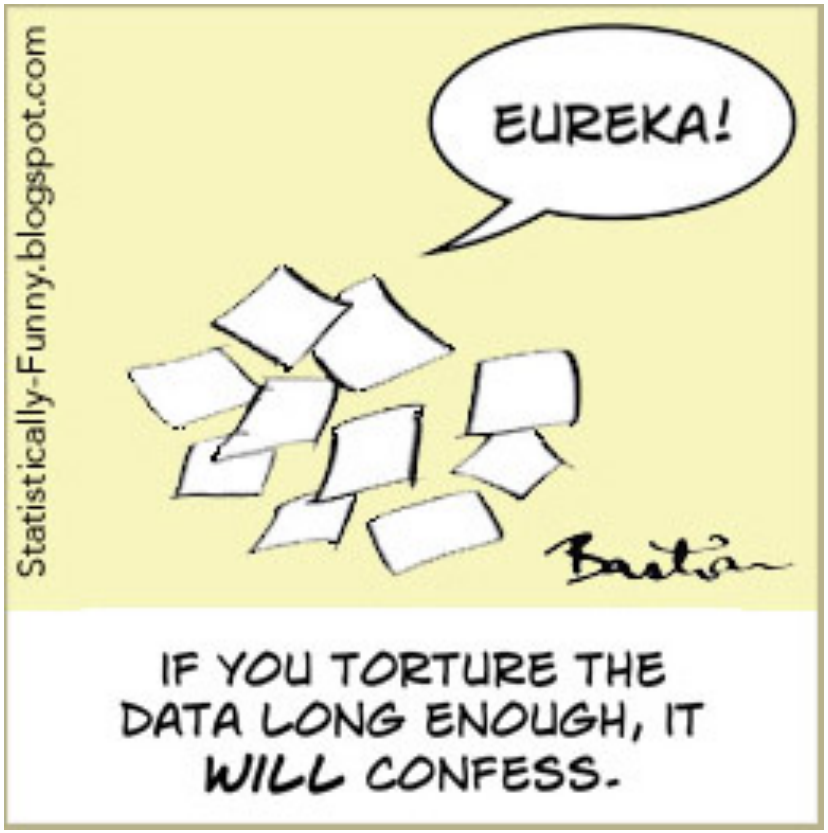
http://j-dm.org/archives/1773

Sin 2: Removal of outliers

- Sometimes, in a set of observations, there is a good trend except for a few points
- Again, the observations are complicated; you can probably find a reason that those points are suspect
- But beware! By doing this *after* you see the results, you are susceptible to **confirmation bias**; then your estimates of significance are compromised

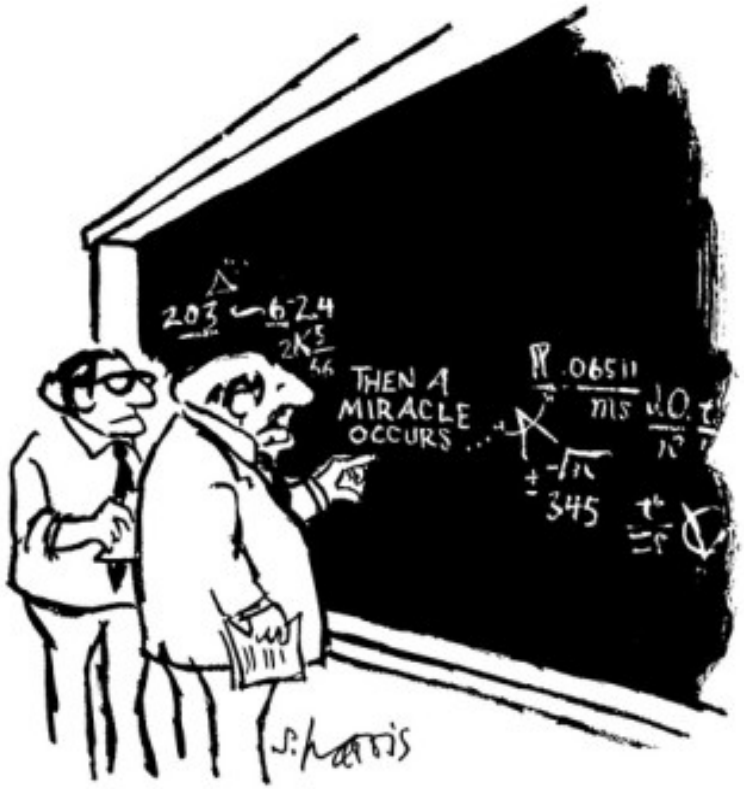
Sin 3: Not estimating # of trials correctly

- With enough data, there are plenty of things that might seem improbable *after the fact*
  Example: I just flipped a coin 10 times and got THTTHHHTHT; only one chance in 1024 for that exact sequence. Wow, amazing!
- But that's why you need to specify your procedure in advance
- Many astronomers will see "something weird" and compute after the fact how odd that exact thing is
- If you're going fishing, you should designate a small fraction of your data (10%?) as a playground; look for anything there, decide on your procedure for the other 90%



Sin 4: Using a black box code

- The temptation is strong: just use something put together by someone else!
- Advantage: reads in data properly, relieves you of effort
- Disadvantage: you have no idea what is going on inside the code!
- It could easily be making assumptions that don't apply to your data
- At *least* understand those assumptions!



"I think you should be more explicit here in step two."

Sin 5: Not thinking about your results

- This is true about everything, not just statistics
- Is your answer reasonable? Does it pass gut check tests?
- Example: you do an analysis and find that the average human height is 50.735 meters, with a standard deviation of 0.13 cm
- Does that make sense?
- ***Look*** at your data! Think in advance about checks you can do. It will avoid embarrassment...

EXPERIENCE CAN JUST MEAN MAKING THE SAME MISTAKE WITH INCREASING CONFIDENCE.

Any questions about statistics?