Statistics is Easy

Dennis Shasha

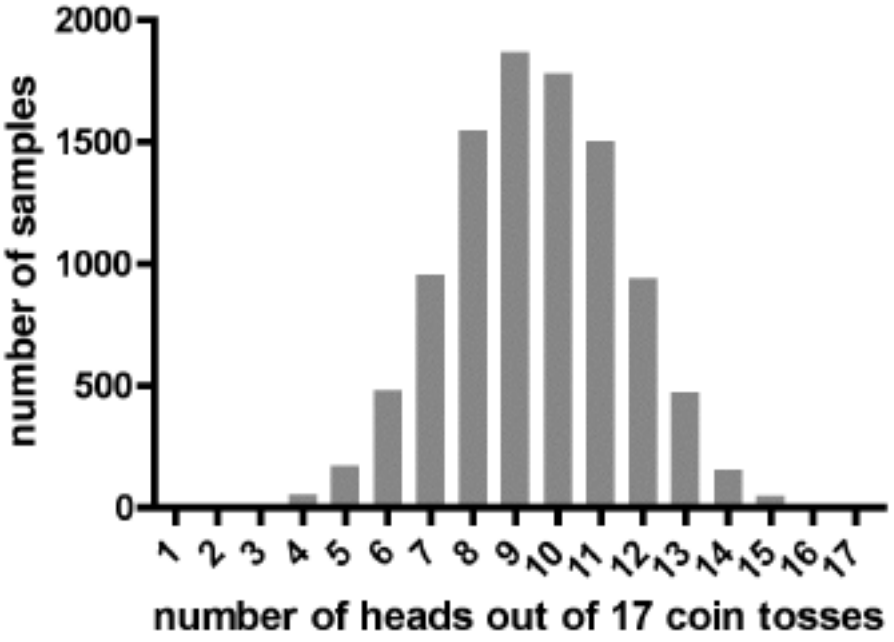
From a book co-written with Manda Wilson

Is the Coin Fair?

- You toss a coin 17 times and it comes up heads 15 out of 17 times.
- How likely is it that coin is fair?
- Could look up Gaussian approximations to Bernoulli processes. Maybe you've forgotten…
- Or…

Is the Coin Fair?

- You could do the following 10,000 times: toss a fair coin 17 times and count how many times you end up with 15 or more heads.



Is this Practical?

- Takes under a second with your computer.
- Is this cheating? No, in the spirit of our times: solve differential equations with Euler's method.
- Is it better? Yes, because more robust, easier to reason about, handles skewed distributions (e.g. average salary of $50,000 with variance of $15,000 gives non-zero probability of negative salary).

What is the Result?

- Something like 9 out of 10,000 times, get 15 heads in 17 tosses.
- This gives a "p-value" of 9/10000.
- P-value is the probability that the outcome observed would have happened by chance if the coin truly were fair.
- Smaller p-value means less likely that the "null hypothesis" (coin is fair) is true.

Is the Drug Effective?

- Random experiments are a good way to establish causality in the face of uncontrolled variation (e.g. lifestyle, wealth etc.)
- Suppose that we do an experiment in which we compare the positive effect of a drug vs. the effect of a placebo.
- Placebo: 54, 51, 58, 44, 55, 52, 42, 47, 58, 46
- Drug: 54, 73, 53, 70, 73, 68, 52, 65, 65

Is the Drug Effective?

- Average for drug is 63.7 and for placebo 50.7.
- Is that the end of the story?
- Maybe not: perhaps the drug population "just happened" to be better.
- Standard statistical approach: assume something about distributions and see how big the overlap among distributions is.

Issues with the standard approach?

- Distribution assumption may not hold.
- Must be careful about different sizes of one population vs. the other.
- Easy to get the technicalities wrong (for non-statisticians).

Resampling/Shuffle Approach

- Create a table that associates each outcome with the label D (drug) or P (placebo). Here is the beginning of such a table:

D	D	D	P	P
54	73	53	54	51

Using the Table

- If we take the table as given, then of course the entries associated with D have an average value of 63.7 and the placebo entries have average 50.7.
- But now consider shuffling the labels among the entries. That causes a total loss of the connection between treatment and improvement.

Establishing Significance

- If, after shuffling, the average of the values of the D entries is often (say more than 15% of the time) greater than the average of the values of the P values by 63.7-50.7, then the apparent effectiveness of the drug in the experiment could be entirely due to chance.
- If very uncommon, then the drug may be doing something.
- In this case p-value is 0.1%.

Different numbers give different results

- Placebo: 56, 348, 162, 420, 440, 250, 389, 476, 288, 456.
- Drug: 69, 361, 175, 433, 453, 263, 402, 489, 301, 469.
- So difference in averages is still 13, but now the p-value is about 40%. Could easily be due to chance.

Significance vs. Importance

- Suppose that we try a different drug/placebo experiment on 1 million patients and the drug increases life by 5 years and 3 days whereas the placebo increases life by 5 years alone.
- This might, because of the large sample size, give a low p-value (thus statistically significant).
- But is it important? Do we care? Please ask this question.

Confidence Interval

- Confidence interval is the range of values a measurement is likely to take.
- In the case of our first drug/placebo experiment, the difference of the average effect was about 13 (63.7 – 50.7).
- Can we say what this average difference will be in 90% of the measurements we are likely to make? (90% “confidence interval”)

Bootstrapping Procedure

- 10,000 times, create a sample uniformly randomly with replacement of the drug values and the placebo values (keeping the labels) and then evaluate the difference of the averages.
- Sort these differences. The 90% confidence interval falls between the 500^{th} difference and the $9{,}500^{th}$ difference in the sorted list.

Example of Uniform Random with Replacement

- Original placebo values were:
  54, 51, 58, 44, 55, 52, 42, 47, 58, 46
- So, one uniform random sample with replacement might be:
  55, 54, 51, 47, 55, 47, 54, 46, 54, 54
- Note that some values are repeated and some are missing.
- Difference in the averages yields a range of 7.81 to 18.11 for 90% confidence interval.

Social vs. Natural

- Confidence intervals tend to vary more for social/ cultural phenomena than natural ones: people's weights vary by a factor of maybe 20, but incomes can vary by a factor of 1000 or more.
- More important: in human affairs, past behavior is a bad predictor of future behavior (e.g. German Mark vs. Dollar in the early 1920s).
- See Nassim Taleb's book: *The Black Swan*

Questions for You

- If you know the confidence interval, does significance bring anything to the party?
- Can bootstrapping be used to find the maximum value you are likely to see in an underlying population?
- How does bootstapping/shuffling help when a sample is not representative?

Answer to 1

- If you know the confidence interval, does significance bring anything to the party?
- Consider single drug value 66 and single placebo value 53. What will bootstrapping do? What will shuffling do?
- In fact, first check whether the p-value is small, before measuring the confidence interval.

Answer to 2

- Can bootstrapping be used to find the maximum value you are likely to see in an underlying population?
- Definitely not. Take 1000 people at random. Very unlikely that Warren Buffet or Bill Gates is among them. Would never be able to find maximum wealth from those 1000 using bootstrapping.

Answer to 3

- How does bootstapping/shuffling help when a sample is not representative?
- It doesn't. If there is any selection bias, a study's conclusion may be totally wrong. Example: age of death is highest in Monaco and Andorra. Is the diet so much better? Are people so much more fit?

Further Reading

- Our book *Statistics is Easy* goes on to discuss the most important statistical topics: mean, difference between means, chi-squared tests, statistical power, fisher exact test, anova, regression, correlation, and multiple testing, all from the point of view of resampling.
- Without the case study, it's only 53 pages.
- More extensive books are given as references.