

Is the cereal maker cheating?

- In the coupon collector problem, consider the problem of trying to get the last coupon/toy. The probability of success is $\frac{1}{k}$ to get it (or any particular coupon).
- We saw last time you expect to wait about k “turns” on the average before getting the last toy, assuming the manufacturer distributes the toys uniformly?
- What if the manufacturer is cheating? How could you tell?
- This simple-appearing question leads to a completely different (but related) field.

Statistics

- Before we computed the consequences of a simple model (i.e., toys distributed uniformly at random). The calculations were always sound.
- Could you ever be *sure* the manufacturer was cheating? Perhaps we were just unlucky?
- Let's consider a simpler version of this question: how can we tell that a coin isn't fair?
- Suppose we see a certain set of data (coin flips). Could these have been generated by a fair coin? Of course!

- What is the expected number of heads we would see from a fair coin in n trials? Obviously, $\frac{n}{2}$.
- The probability of seeing exactly k heads will be the number of outcomes with k heads over the total number of outcomes, which is:

$$p_k = \frac{\binom{n}{k}}{2^n}. \quad (1)$$

- If you want to play with this in Excel, the relevant function is BINOMDIST (example will be on the CS280 web site).
- Unsurprisingly, this peaks at $k = \frac{n}{2}$ and falls away fairly quickly, in a shape that looks like a “bell curve”.

- Intuitively, we want to say that a fair coin would not generate a number of heads that's "very far" from $\frac{n}{2}$.
- Of course, it could do that, especially for small values of n . So we want our notion of "very far" to somehow change with n , and our certainty that the coin is biased to increase.
- However, note that as n gets larger the chance of any particular proportion of coins gets smaller. For instance, the probability of observing $\frac{n}{2}$ heads at $n = 100$ is 8%, but only 2.5% at $n = 1000$.
- So we can't argue that the particular thing we saw was unlikely, because it's *always* unlikely.

- Intuition: from a fair coin, most of the time we would see a number of heads k that is near $\frac{n}{2}$.
- We will declare certain ranges of heads to be “unreasonable”, and if k falls into that range we claim the coin is biased.
- The observed number of heads k is called a *statistic*: it is something we can compute from the data, and we can reason about how it would behave if the coin were fair.
- In fact, k is actually a random variable.

- What is the unreasonable range for k ? There is no range of k with zero probability, so we are going to declare unreasonable some values of k that we might actually see from a fair coin.
- We want to make the probability of this small, say $\alpha = 5\%$. Furthermore, the problem is symmetric, so we want to find both the top part of the range of k that has probability 2.5%.
- What is the probability we observe at most j heads from a fair coin? It's the sum $p_0 + p_1 + \dots + p_j$ (recall our definition of p_k).
- So we need to solve for j such that

$$p_0 + p_1 + \dots + p_j = 2.5\%.$$

Terminology

- Old-fashioned statistics books (before spreadsheets) had big tables solving this. It's called the *critical region* of the statistic k . For $n = 1000$, the critical region is $j = 469$.
- In other words with probability $1 - \alpha = \%95$ a fair coin would generate 500 ± 30 heads. This is called the *%95 confidence interval* for the statistic.
- What if we have less data? For a fair coin tossed 100 times, we would observe 50 ± 10 heads. This makes sense; with less data we have more uncertainty.
- The width of the confidence interval is proportional to \sqrt{n} , so it shrinks slowly as a percentage of the range.

Statistical reasoning

- Such reasoning is vital to science, but you have to be precise about what it says. We do not *know* that the coin is unfair; we can merely state that if it were fair our observation was very unlucky.
- Another example: suppose that %50 of people with bronchitis get better the next day. You are testing a new drug. How many people need to get better for you to claim it's not due to chance?
- If you enroll 100 subjects, if the new drug has no effect you will see 50 ± 10 patients get better the next day.

- This is a large part of why medicine is so incredibly expensive. If your new drug is vastly better than chance, you will see an effect with small numbers, but this is extremely rare.
- Many interesting related questions, extensively studied.
- If your new drug cures 55% of patients, how many patients do you need to enroll to see a difference? What is the probability that you would miss a difference if it existed? What is the probability that you would hallucinate an improvement, if it didn't exist?

- How can you estimate the bias of a coin from data? Intuitively, if you see 51 heads on 100 flips, guessing $p = .5$ is good and $p = .9$ is bad.
- Generally this is called *model selection*. There are several models that might produce your data (as usual, any of them *could* have). How do you pick the best one?
- In this specific case it's called *parameter estimation*. One model is "fair coin with $p = .5$ ", another model is "fair coin with $p = .9$." Same model except for the parameter p .
- What principle allows you to claim that the $p = .5$ model is better than the $p = .9$ model?

Maximum likelihood estimation

- Intuitively, a good model would not make what you saw a “fluke”.
- Any model has some probability of having generated what you actually observed (51/100 heads). This is an important conditional probability called the *likelihood*.
- ML parameter estimation: choose the model with highest likelihood. Extremely important, but not necessarily the right answer.
- In our example, a model is a bell curve centered on $100 \cdot p$. The max in a bell curve is at its center, so the ML estimate is $p = .51$.

What else might tell you about bias?

- Consider this coin: HTHTHTHTTH
- Seems pretty fair, about as many heads as tails. Still, something seems fishy. What about a coin whose outcome is HHHHHH-HHTTTTTTTTTT?
- Can we tell from a streak (“run”) that the trials are NOT independent? If so, how?
- Clearly, even a fair coin will generate some runs. What do we expect to be the length of the longest run of heads in 256 tosses?
- Answer: 87% of the time you see at least 6, 63% of the time at least 7, 38% of the time at least 8. Can easily tell students pretending to flip coins.

What is the longest run of heads for a fair coin?

- There are $\frac{n}{2}$ tails, each of which can be the start of a run of heads
- Half of these are followed by a head, so there are $\frac{n}{4}$ runs that start TH
- Half of these are followed by a head, so there are $\frac{n}{8}$ runs that start THH, etc.
- More generally, there are $\frac{n}{2^{r+1}}$ runs of length at least r
- This goes up with n linearly (sanity check), down with r exponentially (ditto)

- As r gets large, at some point the number of runs we expect goes down to 1. This is a good guess as to the length of the longest run.
- Solve for this:

$$\frac{n}{2^{r+1}} = 1 \rightarrow r + 1 = \log_2 n$$

- With some additional effort one can show that the 95% confidence interval for the length of the longest run of heads is

$$[\log_2\left(\frac{n}{2}\right) - 1.9, \log_2\left(\frac{n}{2}\right) + 5.3].$$

Do winning streaks exist?

- In roulette, on August 18, 1913 in Monte Carlo, black came up 26 times in a row.
- The house made a ton of money that night.
- Assuming the odds are of black or red are .48 and there are 500 million spins of the roulette wheel in history, what do we expect to be the length of the longest streak of the same color?
- Answer: 27. 95% confidence interval is [26,32].
- Implications for Wall Street (mutual funds managers) are quite amusing.

- In fact, there is very little statistical evidence in favor of the “hot hand” .
- Almost all famous streaks in sports lie within the 95% confidence interval of what you would expect due to chance.
- Basketball examples: consecutive wins/losses (33, LA Lakers/24, Cleveland Cavaliers); consecutive free throws (97, M. Williams).

- Joe Dimaggio hit safely in 56 consecutive baseball games in 1941, a major league record (many believe it to be the most impressive record in baseball).
- Since 1900, there have been about 500,000 player-game combinations by the top 20% of hitters.
- Assume these hitters bat .300, and get 4 at bats per game. The chance of getting at least one hit in a given game is

$$1 - (1 - .300)^4 = .76$$

- Consider a biased coin with probability of heads $p = .76$, which you toss 500,000 times. What do you expect to be the length of the longest run of heads?
- Answer: 43. However, the 95% confidence interval is [38,56]

Chance versus skill

- The fact that a streak occurred as often as we would expect it to, just by chance, doesn't mean there is no skill involved.
- The minor-league record for hitting in consecutive games is 61 games. It was set in 1933 by a young right-fielder for the San Francisco Seals.
- His name? Joe Dimaggio.