BAS471 - Midterm

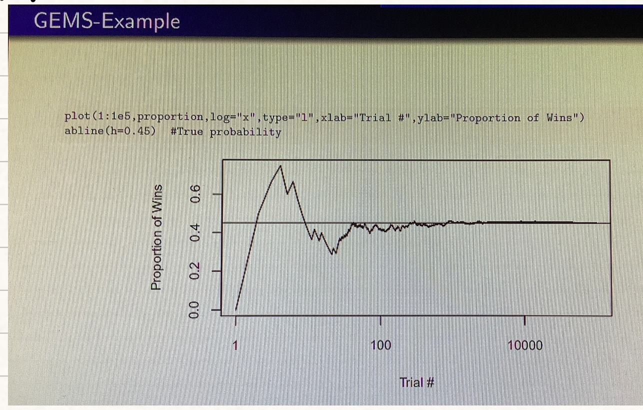
Jake Shoffner

- **A1** - Under frequentist view of probability, which events have and don't have probabilities. When does it make sense to talk about prob. of some event occurring?

  - the probability of some event of interest as the long-run fraction (or proportion) of time when that event of interest occurs.
  - We can never truly know the numerical value of a probability.
  - can't use probability to quantify our ignorance.

A2 - Probability as the long-run FRACTION of time an event occurs. Remember the behavior of what that fraction looks like as the number of trials gets big vs. what the COUNT of some event looks like vs. the "expected number of times that event should occur" as the number of trials get big (last problem on that first homework)

Response: Deviation between observed number and expected number of Norths = (n * sqrt(3)) / (2 * sqrt(n))... So as the number of picks increase, it very slowly increases at a faster rate than the deviation between the observed and expected proportion of Norths.



- A3 - Calculation of probabilities when all outcomes equally likely
  - Knowing that just because there are K possible outcomes doesn't mean each is equally likely with a probability of 1/K; this is true only if outcomes are being picked AT RANDOM
  - When all equally likely, probability of event is fraction of sample space that contains the event

- B - Addition (A or B) and Complement (not A) rules and non-Bayes theorem conditional probability. VERY useful to draw Venn diagrams.
  - If P(A) = 80%, what's the probability that A doesn't occur? 1-P(A)
  - 70% of orders have fries, 80% have burgers; 60% have both. What's the probability that an order has fries or burgers? P(fries or burgers)=? Remember that A or B means "at least one of A or B occurs"
  - Venn diagrams to aid in the calculation of probabilities (A or B, A and B, A but not B, etc.); try Q2 (defects) in Unit1B worksheet or Q5 (houses) on HW1 with Venn diagrams.
  - D - Basic Conditional: P(A|B) = P(A and B)/P(B), where P(A and B)/P(B) are given directly by the problem or can be found with a Venn diagram. E.g. Activity on Unit 3: 1500 churned, 2800 paid manually, 1200 customers both churned and paid manually. P(churn|pay manually) = P(both)/P(manually) = 1200/2800

- $P(A) = 0.8$ $\quad P(A \text{ doesn't occur}) = 1 - 0.8 = 0.2$
- $P(\text{fries}) = 0.7$, $P(\text{burgers}) = 0.8$, $P(\text{both}) = 0.6$, $P(\text{fries or burgers}) = ?$

  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = (0.7 + 0.8) - (0.6) = 0.9$

- $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$

C - Probability of A and B for *independent* events (Multiplication Rule).
- People's orders at the restaurant are independent of each other. What's the probability that the first 2 orders of the day both have fries? The probability that the first 6 orders of the day don't have burgers?
- In a sequence of 5 coin flips, what's the probability that at least one heads occurs? When you see "at least 1", it's easier to use the complement rule! P(at least 1) = 1 - P(none)

P(no heads)

$\left(\frac{1}{2}\right)^5 = \frac{1}{32}$

$\frac{32}{32} - \frac{1}{32} = \frac{31}{32} =$

P(at least one heads)

$P(\text{Disease}) = 0.04\%$, $P(+ \mid \text{Disease}) = 0.999$, $P(+ \mid ND) = 0.02$, $P(\text{Disease} \mid +) = ?$

	+	−	Total
Disease	3.996	0.004	4
No Disease	199.92	9796.08	9996
Total	203.916	9796.084	10000

$P(\text{Dis.} \mid +) = \dfrac{3.996}{203.916} = 1.959\%$

- Baye's Theorem

$P(\text{Librarian} \mid \text{description}) = P(\text{description} \mid \text{librarian})\, P(\text{librarian})$

P(description)

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

• SE = when $p=0.5$, $SE = \frac{0.5}{\sqrt{n}}$

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

$$\boxed{\hat{p} \pm 2SE}$$

"95% chance interval covers p"

$\boxed{ME = 2SE}$

$p=0.05$
no idea what $\hat{p}$ is,

$$n = \frac{4p(1-p)}{ME^2}$$

$p=0.5$
if doesn't give!

SE = typical difference b/t p and $\hat{p}$.

Syntax: `binom.test(x,n,conf.level)`

- `x` - number of times event of interest occurred
- `n` - number of trials
- `conf.levels` - desired level of confidence (a number between 0 and 1)

It might take a while for it to sink in that "the probability that p is inside the confidence interval is 95%" and "there is a 95% chance that the procedure which created the confidence interval manages to cover p" are two *very* different statements (the former being wrong, the later being correct), but take the time to do so.

Example: stockout probabilities

```
#Policy 1:  12 stockouts in 80 trials; Policy 2:  20 stockouts in 120 trials
prop.test( c(12,20), c(80,200), conf.level = 0.95 )$conf.int

## [1] -0.04735576  0.14735576
## attr(,"conf.level")
## [1] 0.95
```

Since 0 is inside the interval, 0 is a plausible value for the true difference in stockout probabilities based on the Monte Carlo simulations that have been run. Although it's doubtful that these two probabilities are *exactly* equal, the data cannot discern a difference between them.

Whenever we we want to compare probabilities among n_{groups} groups, there are a total of $0.5n_{groups}(n_{groups}-1)$ comparisons that can be made.

We need to make each confidence interval *wider* so that, as a *family*, there is a 95% chance that the *collection* only has at least one confidence interval that "misses" the true difference.

Without an adjustment, expect $0.025n_{groups}(n_{groups}-1)$ confidence intervals to miss the true difference by design!

Tukey's **Honest Significant Difference** provides a basis to adjust the collection of confidence intervals comparing two groups so that the *collection* of intervals over all pairs will cover the set of true differences in p's about 95% of the time.

The mechanics of where the adjustment comes from is well-beyond the score of this course (PhD maths!), but we'll use his procedure whenever we need to compare all pairs of groups in some statistical way (so we'll see him again when we compare multiple groups' averages).

Connecting Letters Report for Alumni

To interpret:

- Groups are ordered from largest estimated probability (left, letters earlier in alphabet) to smallest (right, letters later in alphabet).
- If two groups share a letter beneath them (e.g., Finance and Logistics), then a confidence interval for the difference in probabilities will contain 0. The data can't discern a difference in probabilities between these two groups.
- If two groups do not share a letter beneath them (e.g., Accounting and Finance), then a confidence interval for the difference in probabilities will not contain 0. The data suggests that whichever group has the letter that is first alphabetically has the higher probability.

Bayes Theorem

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$ is called the *prior* probability of A. It is also called the *marginal* or *unconditional* probability of A, and represents the overall probability that A occurs in general.
- $P(A|B)$ is called the *posterior* probability of A (given the new information that B has occurred)
- $P(A \text{ and } B)$ is called the *joint* probability of A and B while $P(B)$ is the *marginal* or *unconditional* probability of $P(B)$

In previous illustrations, finding $P(B)$ was straightforward. Often, it requires employing a "trick".