Probability

PSCI 1801 · Class Handout

Dr. Marc Trussler

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

The Basics of Probability

Probability is the expected long-term frequency of an event.

> The chance of something gives the percentage of time it is expected to happen, when the basic process is done over and over again, independently and under the same conditions.

Two key terms: an event *is a possible outcome; the* sample space *is the set of all possible outcomes.*

Sample spaces:

- Coin flip: $\{H, T\}$
- Die roll: $\{1, 2, 3, 4, 5, 6\}$

When all outcomes are equally likely, probability = (favorable outcomes) / (total outcomes).

$$P(H) = \frac{1}{2}, \qquad P(2) = \frac{1}{6}, \qquad P(\text{even}) = \frac{3}{6} = \frac{1}{2}$$

Axioms

Three rules that all probabilities obey.

$$P(A) \geq 0$$

$$P(\Omega) = 1$$

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{when } A, B \text{ mutually exclusive}$$

The general "or" rule (works whether or not events overlap):

$$P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$$

Complements

The complement of an event is everything in the sample space that isn't that event.

$$P(A) + P(A^c) = 1 \qquad \Longrightarrow \qquad P(A) = 1 - P(A^c)$$

Simulating Probability in R

Instead of computing probabilities by hand, we can estimate them by using `for()` loops to repeat a random process many times.

Example: flip 7 coins 100,000 times. How many heads on average?

```
[1] 3.50235
```

```
num.heads
    0     1     2     3     4     5     6     7
  787  5551 16051 27450 27651 16201  5540   769
```

The average is very close to $7 \times 0.5 = 3.5$, and the distribution is bell-shaped around 3–4 heads.

Permutations & Combinations

A permutation *counts ordered sequences. A* combination *counts unordered selections.*

Number of permutations of k items from n:

$$_nP_k = \frac{n!}{(n-k)!}$$

Number of combinations of k items from n:

$$_nC_k = \frac{n!}{k!(n-k)!}$$

Example: guessing a 3-digit locker code on a lock with 50 numbers (order matters, no repeats).

$$_{50}P_3 = \frac{50!}{47!} = 50 \cdot 49 \cdot 48 = 117{,}600 \qquad P(\text{guess right}) = \frac{1}{117{,}600}$$

R can confirm this by simulation:

```
results
  FALSE    TRUE
0.99999 0.00001
```

Mega Millions

Match 5 white balls (from 70) and 1 gold ball (from 25). Order doesn't matter.

$$\#\text{combos} = \binom{70}{5} \cdot 25 = 12{,}103{,}014 \cdot 25 = 302{,}575{,}350$$

```
[1] 12103014
```

Expected value of the lottery is zero when the prize P makes $p(win) \cdot P = \$2$ (the ticket cost):

$$\frac{1}{302{,}575{,}350} \cdot P = 2 \qquad \Longrightarrow \qquad P = \$605{,}150{,}700$$

The Birthday Problem

In a class of 20, what is the probability that at least two students share a birthday? It's easier to compute the complement — the probability that nobody *shares — and subtract from 1.*

$$P(\text{no shared birthdays}) = \frac{{}_{365}P_{20}}{365^{20}}$$

```
[1] 0.5885616
```

So $P(\textit{shared birthday}) \approx 1 - 0.589 = 0.411$. *About 41%.*

Confirmed by simulation:

```
result
FALSE  TRUE
 5891  4109
```

Conditional Probability

Conditional probability is a subsetting *operation: restrict the sample space to only outcomes consistent with the condition, then count within that restricted set.*

$$P(A \mid B) = \frac{P(A\&B)}{P(B)}$$

Example: roll two dice. The 36 outcomes are all equally likely. Cell values show the sum of the two dice; bold cells are sum ≥ 8.

	d2=1	2	3	4	5	6
d1=1	2	3	4	5	6	7
d1=2	3	4	5	6	7	**8**
d1=3	4	5	6	7	**8**	**9**
d1=4	5	6	7	**8**	**9**	**10**
d1=5	6	7	**8**	**9**	**10**	**11**
d1=6	7	**8**	**9**	**10**	**11**	**12**

Unconditional: 15 of 36 outcomes are bold, so $P(\text{sum} \geq 8) = \frac{15}{36} = \frac{5}{12}$.

Conditional on $d_1 = 5$: restrict to the row $d_1 = 5$. Four of six outcomes are bold.

$$P(\text{sum} \geq 8 \mid d_1 = 5) = \frac{4}{6} = \frac{2}{3}$$

Conditioning raised the probability from 42% to 67%.

Gut check: $P(A \mid B) \neq P(B \mid A)$. Different questions, different answers.

$$P(d_1 = 5 \mid \text{sum} \geq 8) = \frac{4}{15} \approx 0.27$$

Marginal, joint, and conditional in real data

ANES 2020 data. The marginal *probability of race is just the overall breakdown:*

```
Asian/Hawaiian/Pacific-Islander             Black, non-
Hispanic
                     0.03472732                       0.08877476
                       Hispanic    Multiple races, non-
Hispanic
                     0.09317682                       0.03313769
                Native American             White, non-
Hispanic
                     0.02103204                       0.72915138
```

The joint *probability is the probability of being both a certain race and a certain gender:*

```
                                 gender
race                                   Female        Male
  Asian/Hawaiian/Pacific-Islander 0.016331041 0.018541257
  Black, non-Hispanic             0.057220039 0.031311395
  Hispanic                        0.049607073 0.043835953
  Multiple races, non-Hispanic    0.019155206 0.013998035
  Native American                 0.009577603 0.011051081
  White, non-Hispanic             0.390103143 0.339268173
```

From the joint table you can recover any marginal (sum a row or column) or any conditional (subset then re-normalize).

Independence

Two events are independent when the probability of one doesn't change based on the value of the other.

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

Example: rolling a die and flipping a coin. Are they independent? Compute $P(2 \mid H)$ and compare to $P(2)$:

$$P(2 \mid H) = \frac{P(H\&2)}{P(H)} = \frac{1/12}{6/12} = \frac{1}{6} \qquad P(2) = \frac{2}{12} = \frac{1}{6} \quad \checkmark$$

The Monty Hall Problem

Three doors — one car, two goats. You pick a door; Monty (who knows) opens a different door revealing a goat; you can stay or switch.

The correct sample space has three equally likely outcomes — Monty's choice of which goat to reveal collapses out because your outcome is the same either way:

Stay	Switch
Car $\rightarrow$ Car	Car $\rightarrow$ Goat
$\text{Goat}_1 \rightarrow \text{Goat}_1$	$\text{Goat}_1 \rightarrow$ Car
$\text{Goat}_2 \rightarrow \text{Goat}_2$	$\text{Goat}_2 \rightarrow$ Car

$$P(\text{Car} \mid \text{Stay}) = \frac{1}{3} \qquad P(\text{Car} \mid \text{Switch}) = \frac{2}{3}$$

Intuition: switching wins whenever you started on a goat, which is 2/3 of the time.

```
[1] 0.3366
```

```
[1] 0.6634
```

The Bill Miller Case

In 15 straight years, Bill Miller beat the S&P 500. If success is 50/50 random, how surprising is this?

The wrong question

If we ask "how likely is one particular investor to get 15 straight successes?" the answer is very small. Below: flip 15 coins, 1,000,000 times — how many times do all 15 come up heads?

```
[1] 33
```

Only about 33 out of a million. So one specific investor doing this is genuinely rare.

The right question

But there are ~6,000 investors over ~40 years. What are the odds that any *of them get a 15-year streak of successes (heads)?*

The key idiom: `rle(s)` *returns the lengths and values of runs. We want the longest run of* `1`*s specifically, then check whether it hit 15:*

```
any.lucky
FALSE  TRUE
    6    94
```

In 94 of 100 simulations, at least one *investor got a 15-year streak of successes. Bill Miller's "unprecedented" streak was not statistically surprising once we asked the right question.*