Random Variables and the Binomial Distribution

(Highlights from Chapters 15 and 16)

Random Variables

- A **random variable** is a variable whose possible values are numerical outcomes of a random phenomenon. Use capital letters like X, Y, Z to denote random variables. Use lower case letters x, y, z to denote specific observed values.
- **Example:** X = number of times Paul the Octopus correctly predicts the winner of a World Cup Soccer/Football game in 8 attempts. He got $x = 8$ correct!

(image credit: Wolfgang Rattay/Reuters)

- **Example:** X = number of frogs in a sample of size 30 that have a certain genetic mutation. Maybe in a particular sample I observe $x = 2$.

Binomial Distribution

The **Binomial** distribution represents the distribution of a count of the number of "successes" in n trials

- The probability of success is the same on each trial: denote by lower case p
- The trials are independent

We use a short-hand notation to describe this:

- $X \sim \text{Binomial}(n, p)$
  - "X follows a Binomial distribution with n trials and probability of success p"

Example in Detail: Paul the Octopus

Define X = the number of successful predictions in 8 attempts.

Suppose $p = 0.8$ (Paul's predictions are pretty good!)

We could use the model $X \sim \text{Binomial}(8, 0.8)$

Calculations with the Binomial

- Suppose $X \sim \text{Binomial}(8, 0.8)$

Define the following events (assumed to be independent):

A_1: Paul's 1st prediction is correct, A_2: Paul's 2nd prediction is correct, ..., A_8: Paul's 8th prediction is correct

a. What does the assumption that these events are independent actually mean?

b. What's the probability that Paul gets 8 out of 8 predictions correct?

```
dbinom(x = 8, size = 8, prob = 0.8)
```

```
## [1] 0.1677722
```

c. What's the probability that Paul gets the first 7 predictions correct and the last one wrong?

d. What's the probability that Paul gets 7 out of 8 predictions correct? (we're not specifying which one he got wrong – it could be the first, or the second, or the third, . . . , or the eighth)

```
dbinom(x = 7, size = 8, prob = 0.8)
```

```
## [1] 0.3355443
```

e. What's the probabilty that Paul gets 5 or fewer predictions correct?

Just kidding! Let's not calculate that by hand.

```
pbinom(q = 5, size = 8, prob = 0.8)
```

```
## [1] 0.2030822
```

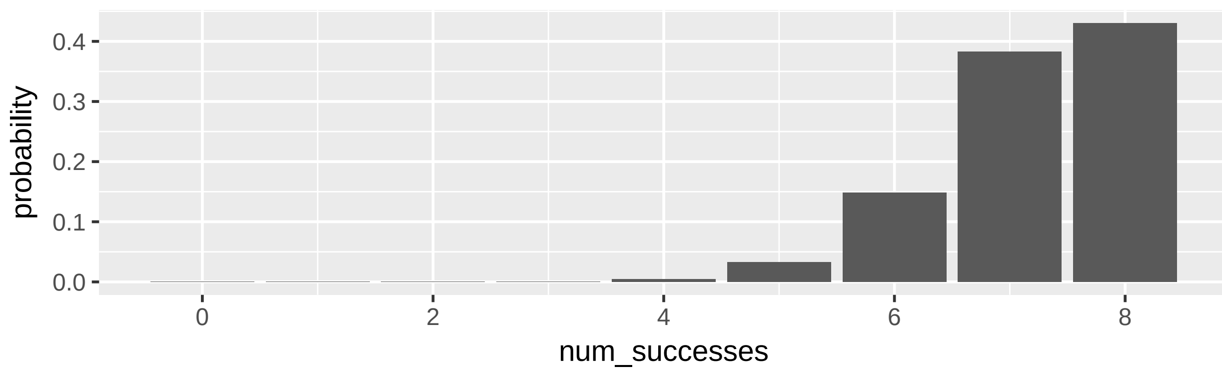
The Full Binomial Distribution

- We can use `dbinom` to calculate the probability of x successes for each possible x from 0 to 8:

```
Paul_success_probs <- data.frame(
  num_successes = seq(from = 0, to = 8),
  probability = dbinom(x = seq(from = 0, to = 8), size = 8, prob = 0.9))
Paul_success_probs
```

```
##   num_successes probability
## 1             0  0.00000001
## 2             1  0.00000072
## 3             2  0.00002268
## 4             3  0.00040824
## 5             4  0.00459270
## 6             5  0.03306744
## 7             6  0.14880348
## 8             7  0.38263752
## 9             8  0.43046721
```

```
ggplot() +
  geom_col(mapping = aes(x = num_successes, y = probability),
    data = Paul_success_probs)
```



Expected Value

- How many do we **expect** Paul to get right? (same as **average**, or **mean**)
- Weighted average of the possible values of X, weighted according to how likely they are

$$\mu = E(X) = \sum_x xP(X = x)$$

Variance and Standard Deviation

- Variance: $\sigma^2 = Var(X) = \sum_x (x - \mu)^2 P(X = x)$

- Standard Deviation: $\sigma = SD(X) = \sqrt{Var(X)}$