Discrete Random Variables

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.

Deriving the Binomial

Setup: how likely is it for a quarterback with a 67% completion rate to complete 0 of his first 3 passes?

Sample space for 3 fair coin flips (equally likely — this is the "counting" world we know):

H H H | T H H | H T H | T T H | H H T | T H T | H T T | T T T |

As a PMF (with K = number of heads):

$$p(K = k) = \begin{cases} \frac{1}{8} & \text{if } 0 \\ \frac{3}{8} & \text{if } 1 \\ \frac{3}{8} & \text{if } 2 \\ \frac{1}{8} & \text{if } 3 \end{cases}$$

But with an unfair coin, sequences are no longer equally likely. We need the multiplication rule *for independent events:*

$$P(A\&B) = P(A) \cdot P(B) \quad \text{when } A \perp B$$

Example — probability of two heads in three flips with a $p = 0.67$ coin:

$$P(HHT) = P(HTH) = P(THH) = 0.67 \cdot 0.67 \cdot 0.33 = 0.148$$

$$P(2 \text{ heads in } 3) = 3 \cdot 0.148 = 0.444$$

Simulation agrees:

```
[1] 0.44548
```

The binomial formula

Two ingredients: (1) number of ways to arrange k successes in n trials, (2) probability of any one such arrangement.

$$P(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

Formally Defining a Random Variable

A random variable *assigns a number and a probability to each event.*

Vocabulary note (important): random variable*,* population*, and* data generating process *(DGP) are used interchangeably for the rest of the class. All three name the same mathematical object — the thing that assigns probabilities to every possible outcome of some process. Internalize this now.*

Two types:

- **Discrete**: takes a finite (or countably infinite) number of values (e.g. coin flips, dice rolls)
- **Continuous**: takes any real value in an interval (e.g. height, income)

The Bernoulli RV

Simplest possible discrete RV: takes value 0 with probability $1-\pi$ *and value 1 with probability* π*. A single coin flip.*

$$f(x) = P(C = c) = \begin{cases} 1-\pi & \text{if } 0 \\ \pi & \text{if } 1 \end{cases}$$

PMF and CDF

The Probability Mass Function *gives* $P(X = x)$ *for each value* x*. The* Cumulative Distribution Function *gives* $P(X \leq x)$*.*

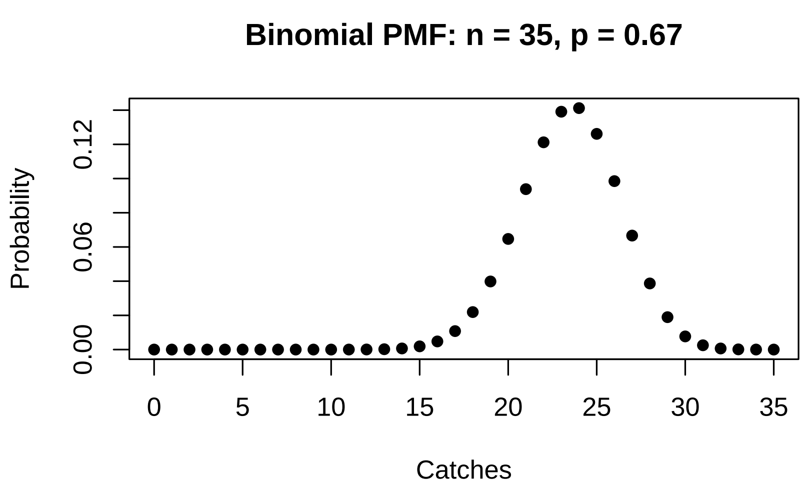
$$\text{CDF} = F(x) = P(X \leq x) = \sum_{k \leq x} f(k)$$

The two are 1:1 related — either one determines the other.

Binomial in R

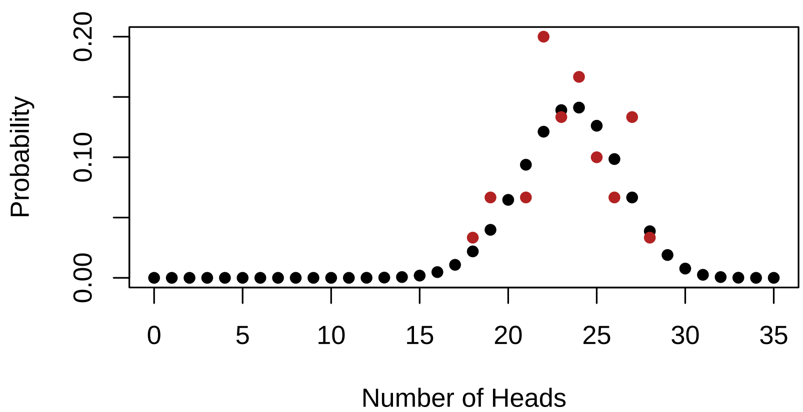
`dbinom()` *for the PMF,* `pbinom()` *for the CDF,* `rbinom()` *for draws:*

Example — Mahomes, 35 pass attempts at $\pi = 0.67$*:*



Samples from a random variable

A random variable describes *what could happen; a sample is what* does *happen when we draw from it. Below: the true PMF for a Binomial with* $n = 35$, $\pi = 0.67$ *(black) vs. 30 realizations (red).*



The red dots are related to the black dots — they came from them — but they aren't identical to them. The gap between what the RV says should happen and what a finite sample produces is exactly what we're headed toward.

Expectation and Variance

The population-level equivalents of "mean" and "standard deviation." They describe an RV, not a sample.

Expected value

$$E[X] = \sum_x x \cdot p(x)$$

For a Bernoulli: $E[X] = \pi$.

For a Binomial (shortcut, because a Binomial is n *independent Bernoullis):*

$$E[X] = n\pi$$

Variance

$$V[X] = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2$$

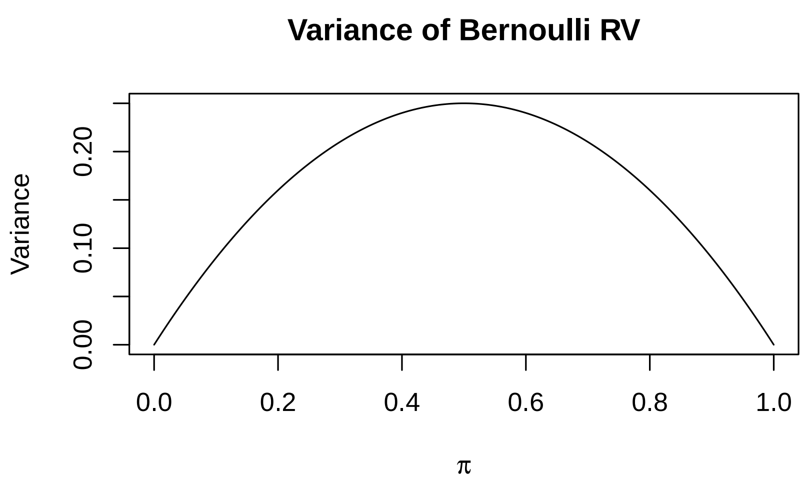
For a Bernoulli:

$$V[X] = \pi(1 - \pi)$$

For a Binomial:

$$V[X] = n\pi(1 - \pi)$$

Variance is largest at $\pi = 0.5$ *and shrinks toward zero as* π *approaches 0 or 1:*



The population standard deviation *is* $\sigma = \sqrt{V[X]}$. *Different from the* sample *standard deviation we've been computing on data.*