The Bernoulli random variable

- The Bernoulli r.v. is the simplest type of random variable you a can get (apart from a constant).
- It forms a building block for many other random variables.
- A coin flip is an example of a Bernoulli trial.
- A Bernoulli rv takes on one of two values. Either a 0 or a 1.
- Sometimes we call the outcomes either a *failure* or a *success*.
- Any random variable with a dichotomous outcome, can be thought of as a Bernoulli.
- Examples
  1. Live/Die.
  2. Buy, don't buy.
  3. Market goes up/market goes down.
  4. Employee stays on the job/ employee quits.
- You have to equate one level of the outcome to the number 1, and the other to 0.

Facts about a Bernoulli

- Denote the random variable with the letter B and associate 1 with a success and 0 with a failure. Denote P(B = 1) as p.

Table 1: The probability distribution for a Bernoulli rv.

b	0	1
p(B=b)	(1 - p)	p

$$E(B) = 0 \times (1 - p) + 1 \times p = p.$$

Don't forget that p is just a constant so:

$$\begin{aligned} Var(B) &= E\{(B-p)^2\} = E(B^2 - 2p\,B + p^2) \\ &= E(B^2) - 2p\,E(B) + p^2 \\ &= p - 2p^2 + p^2 = p - p^2 = p(1-p). \end{aligned}$$

The Bernoulli random variable

Facts:

$$\begin{aligned} E(B) &= p. \\ Var(B) &= p(1-p). \\ sd(B) &= \sqrt{p(1-p)}. \end{aligned}$$

- Recall that p is a probability and therefore must lie between 0 and 1.
- When is the variance of a Bernoulli maximized?
- p(1 - p) is a quadratic (U shaped curve) so either by calculus or by plotting you can show that the maximum variance occurs at $p = 0.5$ when the variance is 0.25 and the standard deviation 0.5.
- Later, when we do *inference*, p $=$ 0.5 will create a worst case scenario.

Illegal downloading example

Background: if an unauthorized *work* is downloaded, but the download originates outside the US, it becomes a legal matter as to whether or not US law applies.

- A single download can be thought of as a Bernoulli trial.
- Equate 0 with the event that the download happens inside the US.
- Equate 1 with the event that the download happens outside the US.
- The probability that the download happens outside the US is 0.1.
- We will now consider a sequence of downloads.

A sequence of downloads

We have access to a sequence of 100 downloads. Denote the sequence as

$$\{B_1, B_2, \cdots, B_{100}\},$$

and consider the **sum** of the sequence:

$$B_1 + B_2 + \cdots + B_{100}.$$

- The **sum** counts the total number of downloads from outside the US, because each element in the sum is either a 1 or a 0.
- Now assume that the sequence is iid, that is, independent and identically distributed.
- This means that the outcome of each event has no impact on any others, and they are all Bernoulli trails with the same probability of "success", in this case 0.1.

The Binomial random variable

The number of successes in n iid Bernoulli trials is called a **Binomial** random variable.
Define

$$Y = B_1 + B_2 + \cdots + B_{100}.$$

Using the formulas from slide 6 and the facts about the mean and variance of a Bernoulli random variable it follows that

$$\begin{aligned} E(Y) &= np. \\ Var(Y) &= np(1-p). \\ sd(Y) &= \sqrt{np(1-p)}. \end{aligned}$$

- A binomial random variable is by definition the sum of **n**, iid Bernoulli trails with success probability **p**.
- We can write as a shorthand $Y \sim Bi(n, p)$.

The proportion of successes

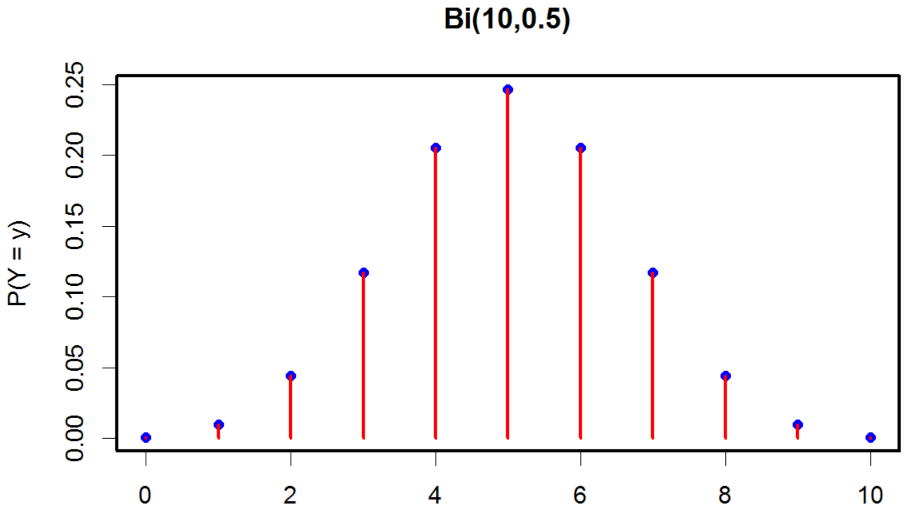
The sample mean of the B_i^s is by definition

$$\bar{B} = \frac{\sum_{i=1}^{n} B_i}{n} = \frac{Y}{n}.$$

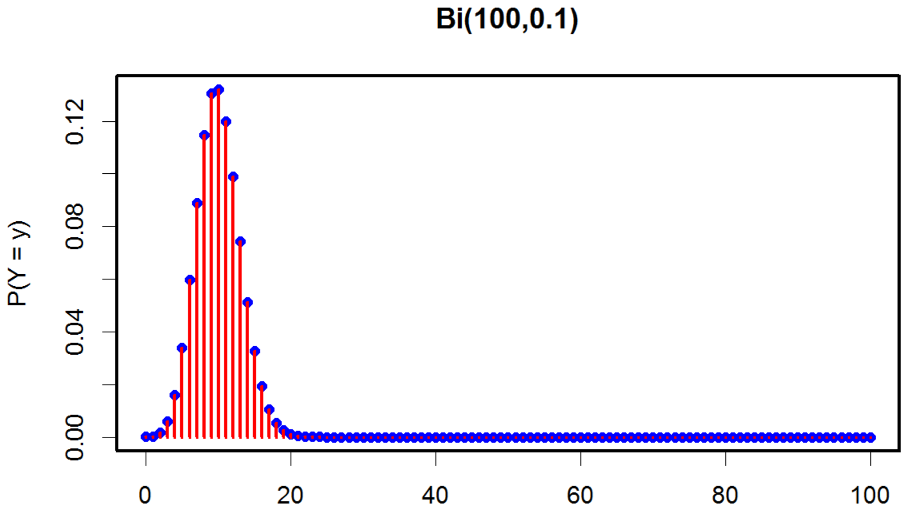
- In the special case of 0/1 outcome variables this mean has an interpretation as the proportion of 1's.
- In this example, the proportion of downloads that are from outside the US.
- It follows that:

$$\begin{aligned} E(\bar{B}) &= p. \\ Var(\bar{B}) &= \frac{p(1-p)}{n}. \\ sd(\bar{B}) &= \sqrt{\frac{p(1-p)}{n}}. \end{aligned}$$

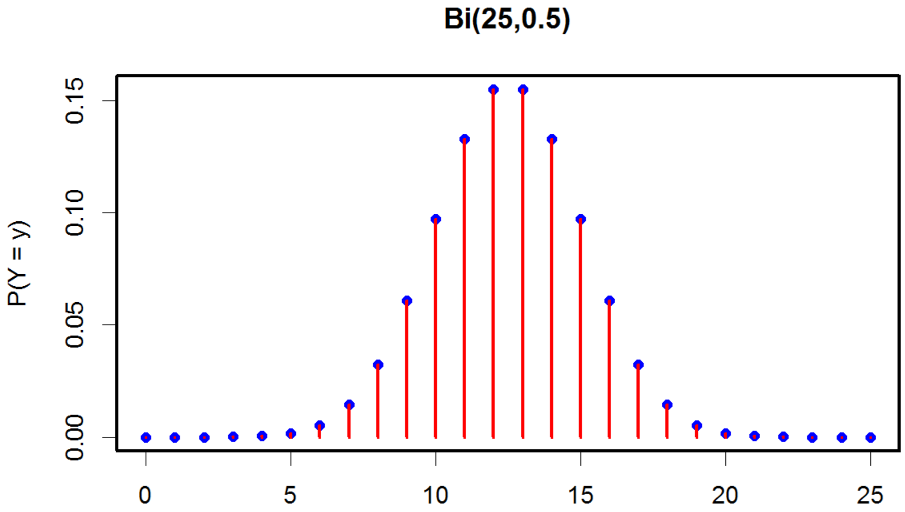
Graphing Binomial distributions



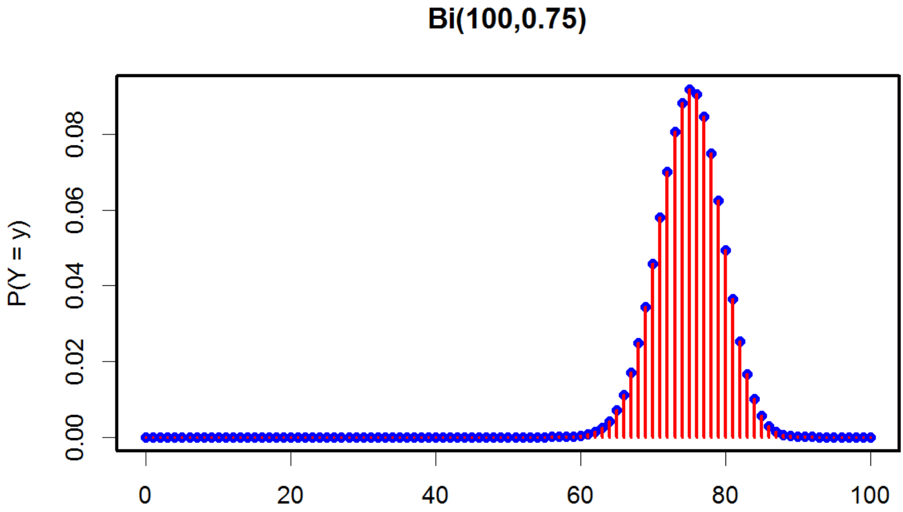
Graphing Binomial distributions



Graphing Binomial distributions



Graphing Binomial distributions



Comments on the probability distributions

- By construction, the binomial random variable can only take on whole number values in the range 0 to n.
- The probability distribution is centered around its mean.
- The probability distribution is very close to bell shaped.

Calculating probabilities

- In the downloads example we can easily calculate the expected number of downloads that were not from the US. Because $p = 0.1$, it is $0.1 \times 100 = 10$.
- But what is the probability all of the downloads were non-US based?
- That only happens if $Y = 100$. Recall, Y is the number of non-US downloads.
- What's the probability that $Y = 100$?
- This means that $B_1 = 1$ and $B_2 = 1$ and .. and $B_{100} = 1$.
- By independence we have

$$P(Y = 100) = 0.1^{100},$$

an event with essentially zero probability.

Calculating probabilities more generally

- Toss a coin 5 times. Assume it is fair (p $= 0.5$).
- What is the probability that you get exactly 2 heads?
- Note that $Y \sim Bi(5, 0.5)$.
- The probability of the first two tosses being heads, and then getting 3 tails is $0.5^2 \times 0.5^3 = 0.5^5 = 0.03125$.
- But, you might have got the heads in different places:

$\{H, H, T, T, T\}$ $\{H, T, H, T, T\}$ $\{H, T, T, H, T\}$ $\{H, T, T, T, H\}$
$\{T, H, H, T, T\}$ $\{T, H, T, H, T\}$ $\{T, H, T, T, H\}$
$\{T, T, H, H, T\}$ $\{T, T, H, T, H\}$
$\{T, T, T, H, H\}$
There are in fact 10 ways of getting 2 heads and three tails. Each possible way is equally likely so $P(Y = 2) = 10 \times 0.03125 = 0.3125$.

The Binomial Coefficient

- The *Binomial coefficient* tells you in how many ways y successes can happen in n trials.
- It has two different notations and is sometimes articulated as *"n choose y"*:

$$\binom{n}{y} = {}_nC_y = \frac{n!}{y!(n-y)!},$$

  where $n! = n \times (n-1) \times (n-2) \cdots \times 1$ and $0! = 1$.
- Example: 2 heads in 5 tosses.

$$\binom{5}{2} = {}_5C_2 = \frac{5!}{2!(5-2)!} = \frac{120}{2 \times 6} = 10.$$

The general formula for the Binomial probability function

If Y is a Binomial random variable that counts the number of successes in n trials, each with probability of success p, then:

$$p(Y = y) = {}_nC_y p^y (1-p)^{n-y}, \text{ where } y = 0, 1, \cdots, n.$$

In the downloads example what are the chances of seeing exactly 5 non-US downloads from the 100 observations?

$$\begin{aligned} p(Y = 5) &= {}_{100}C_5\, 0.1^5 (1-0.1)^{100-5} \\ &= \frac{100!}{5! \times 95!} 0.1^5 0.9^{95} \\ &= \frac{100 \times 99 \times \cdots \times 96}{120} \times 0.00001 \times 0.00004498196 \\ &= 0.0338658. \end{aligned}$$