Counts of Succeses in Bernoulli Trials

Suppose we perform n "trials", each of which may be a "success" or a "failure". (These are called *Bernoulli trials*.)

If each trial has a probability π of succeeding, what is the distribution of the number of successes in the n trials?

We will see how to find this when the trials are *independent*. Will they be when:

- We test whether n machines built in some factory work correctly, picking one machine at random from among those built on each of n consecutive days?
- We check whether n apples have worms, taking 10 apples from each of $n/10$ trees?
- We randomly select n *different* people living in Canada, and ask each one whether or not they like blue cheese?

The Binomial(n,π) Distributions

The distribution of the number of successes in n independent trials, each with probability π of success, is called the binomial(n,π) distribution.

We can compute these distributions by looking at all ways we could get k successes out of n. If the distribution of X is binomial(3,0.2), then

$$\begin{aligned} &P(X=2)\\ &\quad = P(SSF) \;+\; P(SFS) \;+\; P(FSS)\\ &\quad = (0.2)(0.2)(0.8)+(0.2)(0.8)(0.2)+(0.8)(0.2)(0.2)\\ &\quad = 3(0.2)^2(0.8)\\ &\quad = 0.096 \end{aligned}$$

The book has binomial probabilities for various n and π (Appendix B.1 in the back).

Note that the binomial(n,π) probability for k equals the binomial(n,$1-\pi$) probability for $n-k$.

Formula for Binomial Probabilities

If X has the binomial(n,π) distribution, then

$$P(X=k) \;=\; \binom{n}{k}\,\pi^k\,(1-\pi)^{n-k}$$

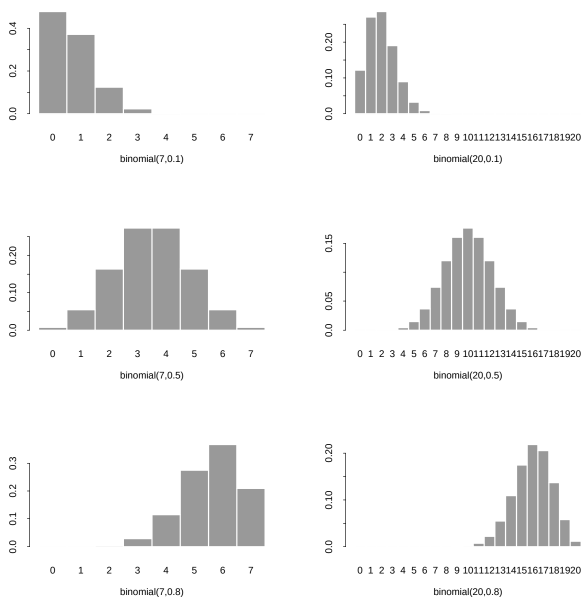
Why? Any single way of getting k successes in n trials has probability $\pi^k\,(1-\pi)^{n-k}$, because the trials are independent.

We need to multiply this by the number of (mutually exclusive) ways we can get k successes out of n. This number is called "n choose k", and is given by

$$\binom{n}{k} \;=\; \frac{n!}{k!\,(n-k)!}$$

Here, $n!$ is called "n factorial", and is defined as $1 \times 2 \times \cdots \times n$. (Note that $0! = 1$.)

Some Binomial Distributions



Example: Overbooking a Bus

Suppose you run bus tours, and have noticed that only 80% of people who say they will come actually do. The bus holds 18 people. If you book 20 people, how likely is it that more than 18 will show up?

If people show up or not independently, the number, X, who show up will have the binomial(20,0.8) distribution.

On this assumption, we can calculate the overbooking probability to be

$$\begin{aligned} P(X=19) \ + \ P(X=20) & \\ &= 0.0576 + 0.0115 \\ &= 0.0691 \end{aligned}$$

Is the independence assumption reasonable?

Continuous Random Variables

Many quantities that we might want to model as random variables have a *continuous* range of values:

- Percentage of alcohol in a batch of wine (range is 0 to 100).
- Height of a corn plant grown in a certain way (range is 0 to ∞).
- Time difference between when a potential passenger arrives at a bus stop and when the bus arrives (range is $-\infty$ to $+\infty$).

In practice, we will be able to measure these values to only a finite precision, but it may still be useful to *model* the measurements as being continuous.

Distributions for Continuous Random Variables

The distribution of a continuous random variable can't be shown in a table — it would have an infinite number of entries!

Also, the probability of any particular value is *zero* — we never get *exactly* 16.509783...% alcohol in wine.

We will assign probabilities only to events of the random variable (say X) being in some interval of numbers, such as

$$(1.25, 7.81), \quad (-\infty, -1), \quad (\sqrt{2}, \infty)$$

plus events formed from these by unions and complements, such as

(0.2, 0.3) plus everything outside (−1.1, 5.3)

The probability for this event can be written as

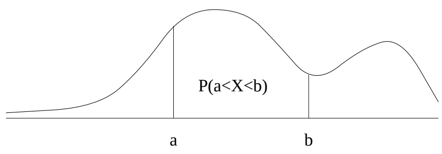
$$P(X \le -1.1 \text{ or } X \ge 5.3 \text{ or } 0.2 < X < 0.3)$$

Assigning Continuous Distributions Using Density Functions

One way to assign probabilities to events involving a continuous random variable is to specify its *probability density function*.

The density, $f(x)$, for a random variable X must be non-negative, and must have a total area under it of one — ie, $\int_{-\infty}^{+\infty} f(x)\, dx = 1$.

We find $P(a < X < b)$ by finding the area under $f(x)$ between a and b:



This can also be written as $\int_a^b f(x)\, dx$.

We can then use the rules of probability to find things like $P(1 < X < 5 \text{ or } 8 < X < 9)$.