Lecture 3: Binomial Distribution

Sta230/Mth230

Colin Rundel

January 22, 2014

Example

Imagine you have a bag with 6 slips of paper numbered 1 to 6. How many different pairs can you draw if you sample without replacement?

Permutations & Combinations

If we have n items and want to select k of them without replacement, then there are j possible outcomes.

Permutations - when we care about the order in which pull out the items:

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

Combination - when we *do not* care about the order in which pull out the items:

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

Permutations & Combinations - Derivation

Some properties of the Binomial coefficient

$$\binom{n}{k} = \binom{n}{n-k}$$

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}, \text{ for } 0 < k < n$$

Pascal's Triangle and the Binomial coefficient

$$\binom{0}{0}$$
$$\binom{1}{0}\ \binom{1}{1}$$
$$\binom{2}{0}\ \binom{2}{1}\ \binom{2}{2}$$
$$\binom{3}{0}\ \binom{3}{1}\ \binom{3}{2}\ \binom{3}{3}$$
$$\binom{4}{0}\ \binom{4}{1}\ \binom{4}{2}\ \binom{4}{3}\ \binom{4}{4}$$
$$\vdots$$

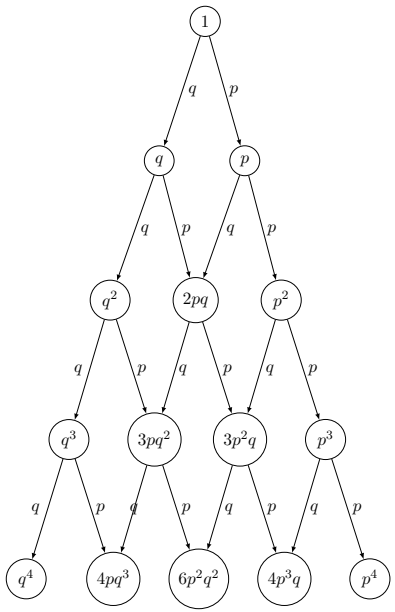
Example

A company is testing a new manufacturing process for aluminum cases, if 20% of the cases do not meet their specifications what is the probability that if the company checks the next four cases that only one of them will not meet specification?

Let the probability a test succeeds be $p = 0.8$ and the probability a test fails be $q = 0.2$ then

$$\begin{aligned} P(1 \text{ failure in } 4 \text{ tests}) &= pppq + ppqp + pqpp + qppp \\ &= 4p^3q \\ &= \binom{4}{1} p^3 q \end{aligned}$$

Binomial Distribution



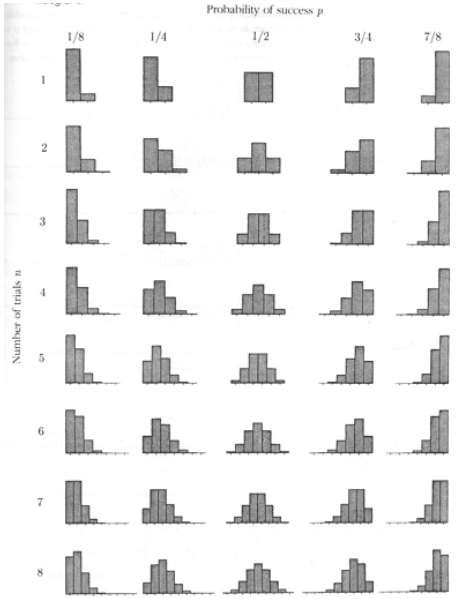
Binomial Distribution

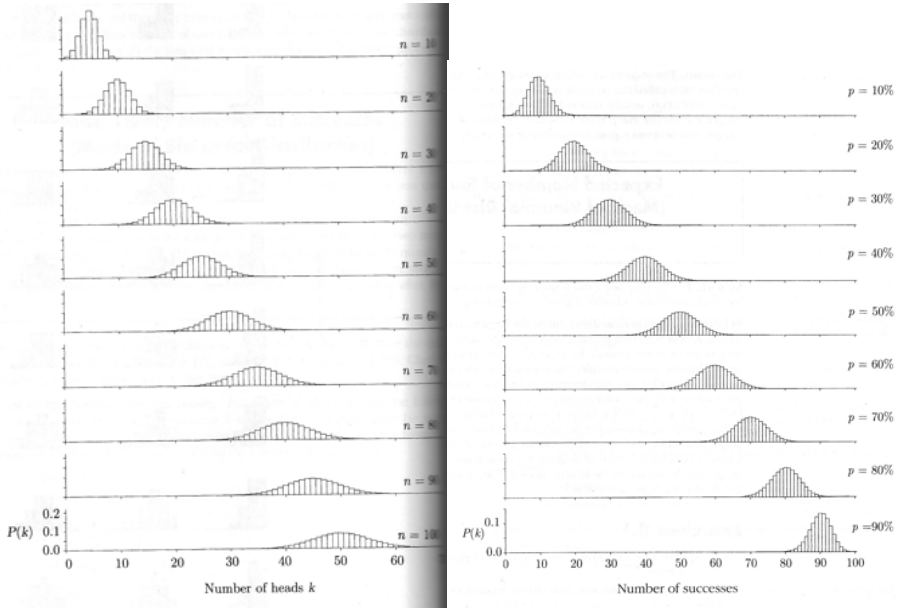
We define a random variable X that reflects the *number of successes* in a *fixed number* of *independent trials* with the *same probability of success* as having a binomial distribution.

If there are n trials then

$$X \sim \text{Binom}(n, p)$$

$$P(X = k|n, p) = f(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$





What is the most probable outcome?

Binomial distribution is unimodal which makes our life easier...

We can look at the ratio of successive outcomes,

$$r = \frac{P(X = k+1)}{P(X = k)}, \text{ for } 0 \leq k \leq n-1$$

r is largest when $k = 0$ and gets progressively smaller.

When $r > 1$ then $P(X = k+1) > P(X = k)$
When $r < 1$ then $P(X = k+1) < P(X = k)$

Maximum (mode) of the distribution occurs when r switches from being greater than 1 to less than 1.

What is the most probable outcome? cont.

$$r = \frac{P(X = k+1)}{P(X = k)} = \frac{\binom{n}{k+1} p^{k+1}(1-p)^{n-(k+1)}}{\binom{n}{k} p^{k}(1-p)^{n-k}} = \frac{n-k}{k+1}\frac{p}{1-p}$$

What value of k results in $r \leq 1$?

What is the most probable outcome? cont.

Max probability is therefore the smallest integer value of $k \geq np - q$.

We can narrow that relationship down somewhat since there must be an integer value of k between

$$\begin{aligned} np - q \leq k &\leq np - q + 1 \\ np - q \leq k &\leq np - q + (p + q) \\ np - q \leq k &\leq np + p \end{aligned}$$

Special case when $r = 1$ as it implies that $P(X = k) = P(X = k + 1)$ in which case both values are equally probable.

What is the scale of this maximum probability?

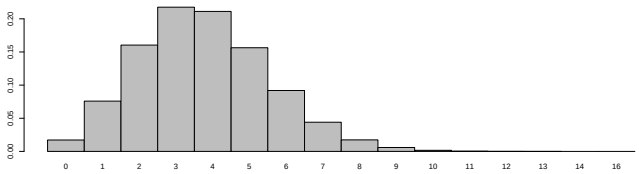
Not very large...

$P(X = k)$ maxes out at a little bit less than $1/\sqrt{n}$, therefore $P(X = k_{\text{mode}}) \to 0$ as $n \to \infty$.

Conceptually, as the number of bins increases the mass in each bin must necessarily get smaller, we are in essense moving from discrete to continuous distribution.

Some examples...

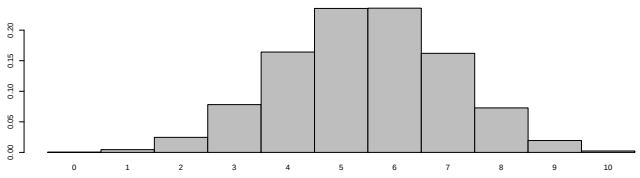
Let $X \sim \text{Binom}(25, 0.15)$ then the distribution of X looks like



$$k_{\text{mode}} = \lfloor (np - q, np + p)] = \lfloor (2.9, 3.9)] = 3$$

Some examples...

Let $X \sim \text{Binom}(10, 6/11)$ then the distribution of X looks like



$$k_{\text{mode}} = \lfloor (np - q, np + p)] = \lfloor (5, 6)] = 5, 6$$

Outcome Ranges

Often it is more interesting to talk about the probability of a range of outcomes.

For example, going back to the manufacturing example (where p=0.8, n=4) what is the probability that there are 1 or fewer defective cases?

$$\begin{aligned}
P(\text{1 or fewer defective}) &= P(\text{3 or more successes}) \\
&= P(X = 3 \text{ or } 4) \\
&= P(X = 3) + P(X = 4) \\
&= \binom{4}{3}(0.8)^3(0.2)^1 + \binom{4}{4}(0.8)^4(0.2)^0 \\
&= 4(0.1024) + 1(0.4096) \\
&= 0.8192
\end{aligned}$$

Outcome Ranges, cont.

What if the company manufactured 1,000 cases and wants to know the probability of at least 850 cases being within specification.

$$P(X \geq 850) = \sum_{k=850}^{1000} \binom{1000}{k}(0.8)^k(0.2)^{1000-k}$$

We can obviously calculate this, but it is a pain to calculate all 251 terms.

For large enough values of n we can approximate this discrete distribution with a continuous distribution.

Normal Approximation

Let $x = z + np$ and $c = P(np)$ then $z = x - np$ and

$$\begin{aligned}
\log P(np + z) &\approx \log P(np) - \frac{z^2}{2npq} \\
\log P(x) &\approx \log P(np) - \frac{(x - np)^2}{2npq} \\
P(x) &\approx \exp\left(\log P(np) - \frac{1}{2}\frac{(x - np)^2}{npq}\right) \\
P(x) &\approx c\, e^{-\frac{1}{2}\frac{(x-np)^2}{npq}}
\end{aligned}$$

de Moivre-Laplace Limit Theorem

When n is large enough the Binomial distribution will always have this bell-curve shape.

- Approximation is usually considered reasonable when $np \geq 10$ and $nq \geq 10$

Shape of the curve given by $c\ e^{-b(x-a)^2}$

de Moivre and Laplace where the first to identify this pattern and characterize the shape of the curve (by finding a, b, c).

This is a special case of a more general result known as the Central Limit Theorem. (More on this later)

Normal Approximation wrap-up

We have shown that

$$f(x) = c\ e^{-\frac{1}{2}\frac{(x-np)^2}{npq}}$$

but what is the value of c?

Since $f(x)$ is describing a probability density function then

$$\int_{-\infty}^{\infty} f(x)\ dx = 1$$

which we can use to calculate the value of c that makes this relationship hold. c in this case is called the normalizing constant.

Normal Approximation wrap-up, cont.

$$\int_{-\infty}^{\infty} c\ e^{-\frac{1}{2}\frac{(x-np)^2}{npq}}\ dx = 1$$

Let $z = (x - np)/\sqrt{npq}$ then $dx = \sqrt{npq}\ dz$ and we change the variables inside the integral such that we get

$$c \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} \sqrt{npq}\ dz = 1$$

Take as a given that $\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = \sqrt{2\pi}$ then

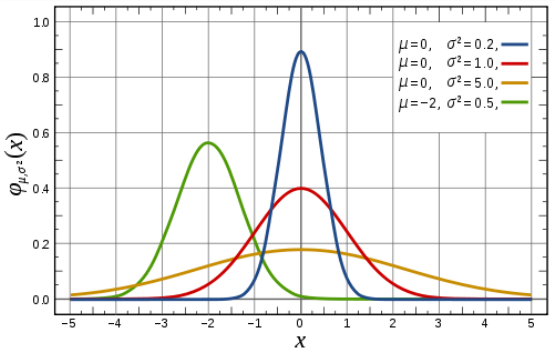
$$c = \frac{1}{\sqrt{2\pi npq}}$$

$$f(x) = \frac{1}{\sqrt{2\pi npq}}\ e^{-\frac{1}{2}\frac{(x-np)^2}{npq}}$$

Normal Distribution

If X is random variable with a normal distribution with a mean μ and variance σ^2, $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}\ e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \qquad P(a \leq x \leq b) = \int_a^b f(x)\ dx$$



Connections

We can see the connection between the approximation and the normal distribution if we set

$$\mu = np$$
$$\sigma^2 = npq$$

We will talk more about the mean and variance of a random variable in the next chapter.