Chapter 2.1-2.3

Clarification

Midterm 1 will be on Wednesday, February 15th.

Lecture 5: Binomial Distribution

Statistics 104

Colin Rundel

January 30, 2012

Chapter 2.1-2.3

Combinations

We have already seen this in a variety of problems, if we have \( n \) items and want to select \( k \) of them how many possible groupings are there?

Given by the binomial coefficient

\[
\binom{n}{k} = \frac{n!}{(n-k)!k!}
\]

How many combinations of two numbers between 1 and 6 are there:

\[
\{1,2\} \{1,3\} \{1,4\} \{1,5\} \{1,6\} \\
\{2,3\} \{2,4\} \{2,5\} \{2,6\} \\
\{3,4\} \{3,5\} \{3,6\} \\
\{4,5\} \{4,6\} \\
\{5,6\}
\]

Chapter 2.1-2.3

Permutations

Something we haven’t seen explicitly yet, if we have \( n \) items and want to select \( k \) of them how many possible orderings are there?

Given by

\[
\frac{n!}{(n-k)!}
\]

How many permutations of two numbers between 1 and 6 are there:

\[
\{1,2\} \{1,3\} \{1,4\} \{1,5\} \{1,6\} \\
\{2,1\} \{2,3\} \{2,4\} \{2,5\} \{2,6\} \\
\{3,1\} \{3,2\} \{3,4\} \{3,5\} \{3,6\} \\
\{4,1\} \{4,2\} \{4,3\} \{4,5\} \{4,6\} \\
\{5,1\} \{5,2\} \{5,3\} \{5,4\} \{5,6\} \\
\{6,1\} \{6,2\} \{6,3\} \{6,4\} \{6,5\}
\]
Pascal’s Triangle

\[
\begin{array}{cccccc}
1 \\
1 & 1 \\
1 & 2 & 1 \\
1 & 3 & 3 & 1 \\
1 & 4 & 6 & 4 & 1 \\
1 & 5 & 10 & 10 & 5 & 1 \\
1 & 6 & 20 & 35 & 35 & 20 & 6 & 1 \\
1 & 7 & 26 & 55 & 70 & 55 & 26 & 7 & 1 \\
\cdots
\end{array}
\]

Example

Apple 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 Apple 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

\[
P(1 \text{ failure in 4 tests}) = ppq + ppqp + pqpp + qppp
\]

\[
= 4p^3q
\]

\[
= \binom{4}{1} p^3q
\]
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

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

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

What is the most probable outcome?

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

We can look at the ratio of successive outcomes,

$$r = \frac{P(X = k + 1)}{P(X = k)}$$

$r$ 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.
Chapter 2.1-2.3

What is the most probable outcome? cont.

\[ r = \frac{P(X = k + 1)}{P(X = k)} = \frac{(n_{k+1}) p^{k+1} (1-p)^{n-(k+1)}}{(n_k) p^k (1-p)^{n-k}} = \frac{n - k}{k + 1} \frac{p}{1 - p} \]

What value of \( k \) results in \( r \leq 1 \)?

\[ \frac{n - k}{k + 1} \frac{p}{1 - p} \leq 1 \]

\[ p(n - k) \leq (k + 1)(1 - p) \]

\[ np - kp \leq k - p - kp + 1 \]

\[ k \geq np - 1 - p = np - q \]

Max probability is the smallest* integer value of \( k \geq np - q \).

Chapter 2.1-2.3

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 \( P(X = k_{mode}) \to 0 \) as \( n \to \infty \).

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

Chapter 2.1-2.3

Some examples...

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

\[ k_{mode} = \text{int}([np - q, np + p]) = \text{int}([2.9, 3.9]) = 3 \]
Let $X \sim \text{Binom}(10, 6/11)$ then the distribution of $X$ looks like

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

**Outcome Ranges**

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

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

$$P(1 \text{ or fewer defective}) = P(X = 3 \text{ or } 4)$$

$$= P(X = 3) + P(X = 4)$$

$$= \binom{4}{3}(0.8)^3(0.2)^1 + \binom{4}{4}(0.8)^4(0.2)^0$$

$$= 4(0.1024) + 1(0.4096)$$

$$= 0.8192$$

**Outcome Ranges, cont.**

What if Apple 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.

$$P(np + z) = P(np + \frac{np + z - 1}{k+1} p) \prod_{k=np}^{np+z-1} \frac{n-k}{k+1} q$$

$$\log P(np + z) = \log P(np) + \sum_{k=np}^{np+z-1} \log \frac{n-k}{k+1} q$$

$$= \log P(np) + \sum_{j=0}^{z-1} \log \frac{np - nj}{npq + q(j+1)}$$

$$= \log P(np) + \sum_{j=0}^{z-1} \log \frac{npq - jnp}{npq + q(j+1)}$$

$$= \log P(np) + \sum_{j=0}^{z-1} \log \frac{npq - jn}{npq + q(j+1)}$$
log \( P(np + z) \) = log \( P(np) \) + \( \sum_{j=0}^{z-1} \log \left( \frac{np - jp}{npq} \right) \)

= log \( P(np) \) + \( \sum_{j=0}^{z-1} \log \left( \frac{1 - jp}{npq} \right) \) - log \( \left( 1 + \frac{jq + q}{npq} \right) \)

\approx log \( P(np) \) + \( \sum_{j=0}^{z-1} - \frac{jp}{npq} \) - \( \frac{jq}{npq} \) - log \( \left( 1 + \frac{jq + q}{npq} \right) \)

\approx log \( P(np) \) - \( \frac{1}{npq} \) \( \sum_{j=0}^{z-1} j + q \)

\approx log \( P(np) \) - \( \frac{z(z - 1) - 2qz}{2npq} \)

\approx log \( P(np) \) - \( \frac{z^2}{2npq} \)

---

**Normal Approximation**

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

\[
\log(P(np + z)) \approx \log(P(np)) - \frac{z^2}{2npq}
\]

\[
\log(x) \approx \log(P(np)) - \frac{(x - np)^2}{2npq}
\]

\[
P(x) \approx \exp \left( \log(P(np)) - \frac{1}{2} \frac{(x - np)^2}{npq} \right)
\]

\[
P(x) \approx c \ e^{-\frac{1}{2} \frac{(x-np)^2}{npq}}
\]

---

**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.

This is an incredibly useful trick and comes up all the time.
Normal Approximation wrap-up, cont.

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

If 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

\[ c = \frac{1}{\sqrt{2\pi npq}} \]

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

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.

Normal Distribution

If \( X \) is random variable with a normal distribution with a mean \( \mu \) and variance \( \sigma^2 \), \( X \sim N(\mu, \sigma^2) \), then

\[ P(X = x) = f(x) = \frac{1}{\sqrt{2\pi \sigma}} \, e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} \]