Lecture 02: Random Variables and Distributions

Mathematical Statistics I, MATH 60061/70061

Thursday September 2, 2021

Reference: Casella & Berger, 1.4-1.6

Example: testing for rare disease

Suppose a rare disease affects 1% of the population, and a test for the disease is 95% accurate. A patient named Fred is tested for disease, and the test result is positive. Given the evidence provided by the test result, what's the conditional probability that Fred has the disease?

Example: testing for rare disease

Suppose a rare disease affects 1% of the population, and a test for the disease is 95% accurate. A patient named Fred is tested for disease, and the test result is positive. Given the evidence provided by the test result, what's the conditional probability that Fred has the disease?

Let D be the event that Fred has the disease and T be the event that he tests positive.

The test is "95% accurate" means

- $P(T \mid D) = 0.95$ (sensitivity or true positive rate)
- $P(T^c \mid D^c) = 0.95$ (specificity or true negative rate)

Example: testing for rare disease

Suppose a rare disease affects 1% of the population, and a test for the disease is 95% accurate. A patient named Fred is tested for disease, and the test result is positive. Given the evidence provided by the test result, what's the conditional probability that Fred has the disease?

$$
\begin{aligned}
P(D \mid T) &= \frac{P(T \mid D)P(D)}{P(T)} \\
&= \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D^c)P(D^c)} \\
&= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \\
&\approx 0.16
\end{aligned}
$$

Example: testing for rare disease, continued

With a test of 95% accuracy, Fred was tested positive for a rare disease that affects 1% of the population. He decides to get tested a second time. The new test is *independent* of the original test (given his disease status) and has the same sensitivity and specificity. Unfortunately, he tests positive a second time. Given the evidence, what's the probability that Fred has the disease?

Rather than updating our beliefs in two steps, we may as well condition on both test results at once:

$$\begin{aligned} P(D \mid T_1, T_2) &= \frac{P(T_1, T_2 \mid D)P(D)}{P(T_1, T_2)} \\ &= \frac{0.95^2 \cdot 0.01}{0.95^2 \cdot 0.01 + 0.05^2 \cdot 0.99} \\ &\approx 0.78 \end{aligned}$$

The probability that Fred has the disease jumps from 0.16 to 0.78. (Getting a second opinion is a good idea!)

The example also illustrates the **coherency** property of Bayes' rule: to update the probabilities based on multiple pieces of information, it does not matter whether we take into account each piece of evidence one at a time, or all at once.

Story proofs

A *story proof* is a proof by interpretation. It goes further than an algebraic proof toward *explaining* why the result is true.

We will discuss 3 examples of story proofs:

- Choosing the complement $\binom{n}{k} = \binom{n}{n-k}$
- The committee chair $n\binom{n-1}{k-1} = k\binom{n}{k}$
- Vandermonde's identity $\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$

Choosing the complement

For any nonnegative integers n and k with $k \leq n$,

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

Choosing the complement

For any nonnegative integers n and k with $k \leq n$,

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

Consider choosing a committee of size k in a group of n people:

- $\binom{n}{k}$ ways to choose which k people on the committee (LHS)

Choosing the complement

For any nonnegative integers n and k with $k \leq n$,

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

Consider choosing a committee of size k in a group of n people:

- $\binom{n}{k}$ ways to choose which k people on the committee (LHS)
- $\binom{n}{n-k}$ ways to choose which $n-k$ people NOT on the committee (RHS)

The LHS and RHS are two ways of counting the same thing $\Rightarrow$ they are equal.

The committee chair

For any positive integers n and k with $k \leq n$,

$$n\binom{n-1}{k-1} = k\binom{n}{k}$$

The committee chair

For any positive integers n and k with $k \leq n$,

$$n\binom{n-1}{k-1} = k\binom{n}{k}$$

Consider choosing a committee of size k in a group of n people, and a chair of the committee:

The committee chair

For any positive integers n and k with $k \leq n$,

$$n\binom{n-1}{k-1} = k\binom{n}{k}$$

Consider choosing a committee of size k in a group of n people, and a chair of the committee:

- Choose one person to be the chair, then choose the other $k-1$ committee members (LHS)

The committee chair

For any positive integers n and k with $k \leq n$,

$$n\binom{n-1}{k-1} = k\binom{n}{k}$$

Consider choosing a committee of size k in a group of n people, and a chair of the committee:

- Choose one person to be the chair, then choose the other $k-1$ committee members (LHS)
- Choose k people to be on the committee, then choose one committee member as the chair (RHS)

Vandermonde's identity

For any nonnegative integers m, n and k with $k \leq m + n$,

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j} \binom{n}{k-j}$$

Vandermonde's identity

For any nonnegative integers m, n and k with $k \leq m + n$,

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j} \binom{n}{k-j}$$

Consider choosing a committee of size k in a group of m juniors and n seniors:

Vandermonde's identity

For any nonnegative integers m, n and k with $k \leq m + n$,

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j} \binom{n}{k-j}$$

Consider choosing a committee of size k in a group of m juniors and n seniors:

- $\binom{m+n}{k}$ ways to choose k people on the committee (LHS)

Vandermonde's identity

For any nonnegative integers m, n and k with $k \leq m+n$,

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$$

Consider choosing a committee of size k in a group of m juniors and n seniors:

- $\binom{m+n}{k}$ ways to choose k people on the committee (LHS)
- If there are j juniors on the committee for $0 \leq j \leq k$, then there must be $k-j$ seniors on the committee (RHS)

Drawing balls

Consider an urn with w white balls and b black balls, and draw n balls out of the urn:

Drawing balls

Consider an urn with w white balls and b black balls, and draw n balls out of the urn:

- When sampling *with replacement*, the # of white balls follows a *Binomial distribution*

$$X \sim \mathrm{Bin}(n, w/(w+b))$$

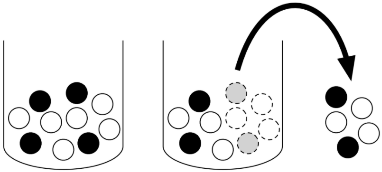
Drawing balls

Consider an urn with w white balls and b black balls, and draw n balls out of the urn:

- When sampling *with replacement*, the # of white balls follows a *Binomial distribution*

$$X \sim \mathrm{Bin}(n, w/(w+b))$$

- When sampling *without replacement*, the # of white balls follows a *Hypergeometric distribution*



Hypergeometric PMF

If $X \sim \mathrm{HGeom}(w, b, n)$, then the PMF of X is

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n - k \leq b$, and $P(X = k) = 0$ otherwise.

Hypergeometric PMF

If $X \sim \mathrm{HGeom}(w, b, n)$, then the PMF of X is

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n - k \leq b$, and $P(X = k) = 0$ otherwise.

The story of the Hypergeometric with w balls and b black balls

- $\binom{w+b}{n}$ ways to draw n balls out of $w + b$
- $\binom{w}{k}\binom{b}{n-k}$ ways to draw k white and $n - k$ black balls
- All samples are equally likely. So $P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$

Hypergeometric PMF

If $X \sim \mathrm{HGeom}(w, b, n)$, then the PMF of X is

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n - k \leq b$, and $P(X = k) = 0$ otherwise.

The story of the Hypergeometric with w balls and b black balls

- $\binom{w+b}{n}$ ways to draw n balls out of $w + b$
- $\binom{w}{k}\binom{b}{n-k}$ ways to draw k white and $n - k$ black balls
- All samples are equally likely. So $P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$

According to Vandermonde's identity, the PMF sums to 1.