PubH 6450 Biostatistics I — Prof. Carlin

Lecture 5
Outline

Summary of Previous Lecture

Probability
  Interpretation of Probability
  Properties of Probability

Random Variables
  Classification
  Probability Distribution
Part I

Review of Previous Lecture
Review of Previous Lecture

- Think about statistics in the study design stage; don’t wait until the data have already been collected.
- Problems caused by biased sampling and unmeasured confounders cannot be fixed by “statistical magic.”
Part II

Probability
Randomness

Definition
A phenomenon or trial is *random* if its outcome is uncertain.

Example
- Tossing a coin.
- Your blood pressure at this moment.
- Tomorrow’s weather.
- Mr. Pawlenty gets re-elected.
Sources of Randomness

- Measurement Uncertainty: instrument, observer, metric.
- Intrinsic Variability.
- Randomness by Design/Sampling.
- Missing Data.
- Subjective Uncertainty (Bayesian thinking): What is the proportion of U.S. men that are HIV-positive?
Genomics Example: DNA Expression Array
Interpretation of Probability

Definition

The *probability* of any outcome of a random phenomenon or trial is the proportion of times the outcome would occur in a very long (infinite) series of repetitions.

This is known as the *frequentist* or *empirical* probability.
Example

The probability of a male live birth in the US.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Male births</th>
<th>Total births</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>1,927,054</td>
<td>3,760,358</td>
<td>.51247</td>
</tr>
<tr>
<td>1965–1969</td>
<td>9,219,302</td>
<td>17,989,361</td>
<td>.51248</td>
</tr>
<tr>
<td>1965–1974</td>
<td>17,857,857</td>
<td>34,832,051</td>
<td>.51268</td>
</tr>
</tbody>
</table>

The sample sizes are so large that the (observed, data-based) relative frequencies can be thought of as (unobserved, theoretical) probabilities!
Elements in Probability

- The trial has a set of specified outcomes. The set of all possible outcomes is *outcome space*.
- The outcome of one trial does not influence the outcome of another. That is, the trials are *statistically independent*.
- The trials are *identical* (e.g., every baby has the *same* chance of being a boy).
- Then the probability of an outcome is the *limit* of its relative frequency in an *infinite* number of repetitions.
Example
Suppose the probability of developing a certain type of stomach cancer over a one-year period among women aged 45 to 49 based on Connecticut Tumor Registry data from 1963 to 1965 is 14 per 100,000.

Now suppose we are studying this stomach cancer in women aged 45-49 in Minneapolis from 2000 to 2004. Can we compare the rates? What pitfalls would there be in doing so?
Subjective Probability

Some probabilities cannot be thought of as limits of long-run event frequencies.

*Example:* What is the probability that I get an “A” in this class?

**Definition**

*Subjective* probability is “an individual’s *degree of belief* that an event will occur.”

“My probability for the event $E$ under circumstances $H$ is the amount of money I am indifferent to betting on $E$ in an elementary gambling situation.” (I. R. Savage, 1968)
Properties of Probability

Rule 1
The probability of any event is between 0 and 1, i.e.,

$$0 \leq \Pr(A) \leq 1,$$

for any event $A$ (a subset of the sample space).
Rule 2
All possible outcomes together must have probability 1, i.e.,

$$\Pr(S) = 1,$$

where $S$ denotes the entire sample space.

Example
Consider the experiment of throwing a fair die. There are six possible outcomes, $S = \{1, 2, 3, 4, 5, 6\}$. Thus one of these outcomes will occur with probability 1. In addition, by symmetry of the die, the probability of each of the outcomes is the same, namely $1/6$. 
Rule 3
The probability of an event does not occur is 1 minus the probability that the event does occur, i.e.,

\[ \Pr(A^c) = 1 - \Pr(A), \]

where \( A^c \) denotes the event that \( A \) does not occur, or the complement of \( A \).

Example
Note the complement of the full set is the empty set, \( S^c = \emptyset \), and this rule still applies. In the die example, \( \Pr(7) = 1 - \Pr(S) = 0 \).
Rule 4: Additivity
If two events have no outcomes in common (i.e., they are disjoint), the probability that one or the other occurs is the sum of their individual probabilities, i.e.,

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B),$$

if events $A$ and $B$ are disjoint.

Example
In the die example, $\Pr(2 \text{ or } 3) = \Pr(2) + \Pr(3) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Note that $\Pr(\{2, 3\} \text{ or } \{3, 4\}) \neq \Pr(\{2, 3\}) + \Pr(\{3, 4\})$.

(i.e., $\frac{1}{2} \neq \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$)
Example

Let $A$ be the event that a person has normotensive diastolic blood-pressure (DBP) readings ($DBP < 90$), and let $B$ denote the event that a person has borderline DBP readings ($90 \leq DBP < 95$).

- Suppose that $Pr(A) = 0.7$ and $Pr(B) = 0.1$. What is the probability of $C$, that a person has $DBP < 95$?

- Now let $D$ be the event that $DBP > 70$ and $Pr(D) = 0.9$. What is the probability of $E$, that a person has $DBP > 70$ or $DBP < 95$?

- In 2001, 22.8% percent of adults in the US were smokers. What can you say about the probability of a person having $DBP > 95$ or that he is a smoker?
Rule 5: Independence and Multiplicative Rule
Two events $A$ and $B$ are *independent* if knowing that one occurs does not change the probability that the other occurs. If $A$ and $B$ are independent,

$$\Pr(A \text{ and } B) = \Pr(A) \Pr(B).$$

**Example**
If a die is thrown twice, the probability of observing 6 twice is

$$\Pr(T_1 = 6 \text{ and } T_2 = 6) = \Pr(6) \Pr(6) = 1/36.$$ 

Note that $\Pr(T_1 = 6 \text{ and } T_2 = 5) = 0$ because these two events are *mutually exclusive* (i.e., if one happens, the other cannot).
Example
Suppose the probability of a positive EIA (enzyme immunoassay) test for a person who is not infected by HIV is 0.006. If 140 HIV-negative persons are screened, what is the probability that at least one test result is positive?

Answer: Assuming the tests are independent,

\[ P(\text{at least one positive}) = 1 - P(\text{all are negative}) \]
\[ = 1 - [P(\text{a single test is negative})]^{140} \]
\[ = 1 - [0.994]^{140} \]
\[ = 0.57. \]

So despite the very high quality of the test, a better-than-even chance of at least one false positive!
What are the probabilities that A beats B, B beats C, C beats D, and D beats A?
Part III

Random Variables
Random Variables

Definition
A random variable is a variable whose value is a numerical outcome of a random event.

A random variable defines a mapping of the sample space to a real number. It does not have to be one-to-one, and there may be more than one possible mapping.

Example
If we toss a coin twice, \( S = \{TT, TH, HT, HH\} \). We can define \( X = \) the number of heads, so that \( X = 0, 1, 2 \). Or we can define \( Y = \) the sequence of heads and tails, so that \( Y = 1, 2, 3, 4 \) (i.e., we assign numerical labels to each outcome).
Types of Random Variables

Definition
A *binary* or *Bernoulli* random variable only takes two values, usually coded as 0 or 1.

Example
Let $X$ denote diastolic blood-pressure (DBP) and define

$$Y = \begin{cases} 
1 & \text{if } X > 95 \text{ ("high") } \\
0 & \text{if } X \leq 95 \text{ ("low") }
\end{cases}$$
Nominal Variable

Definition
A *nominal* or *categorical* random variable takes only a few values among which there is *no clear ordering*.

Example
We can define races as

1. Caucasian
2. African-American
3. Asian-Pacific
4. Hispanic
5. Others
Ordinal Variable

Definition
An *ordinal* random variable takes on a few values among which there is a natural ordering, but for which distance between the categories is not defined.

Example
The Eastern Cooperative Oncology Group’s classification of patient performance status:

<table>
<thead>
<tr>
<th>Status</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Fully active, without restrictions</td>
</tr>
<tr>
<td>1</td>
<td>Restricted in physical strenuous activities</td>
</tr>
<tr>
<td>2</td>
<td>Unable to carry out any work activities</td>
</tr>
<tr>
<td>3</td>
<td>Capable only of limited self-care</td>
</tr>
<tr>
<td>4</td>
<td>Completely disabled</td>
</tr>
</tbody>
</table>
Discrete Quantitative Variable

Definition
A *discrete quantitative* random variable takes *integer* values, and thus has both ordering and distance well-defined.

Example
Many discrete random variables are in the form of *counts*, i.e., the number of children in a household, or the number of heart attacks in a particular county in a year.
Continuous Variable

Definition
A *continuous* random variable takes values over a *range* of numbers on a continuum.

Example
Examples of continuous random variables are everywhere: height, blood pressure, BMI, etc.

Note that due to rounding, all data are “discrete” in some sense (e.g., height is typically measured only to the nearest inch). But many variables are “continuous enough” (many different values, few ties) to be thought of as continuous.
Different types of random variables usually call for different analytic approaches.

There is often more than one way to define a random variable for a quantity of interest; e.g., age could be thought of as either continuous (32.3) or discrete (32). It could also be dichotomized into “young” (< 40) or “old” (> 40), or perhaps reclassified into even more categories.

Note: There will be a *loss of information* when converting a continuous variable (like age) into discrete, ordinal, or binary one.
How to define random variables depends on, among other things,

- the scientific question of interest
- available analytical tools
- practical constraints in data collection
- convention
Probability Distribution

Definition
The probability distribution of a random variable gives the probability that it takes any particular value or range of values. The numeric characteristics that distinguish different distributions are called the parameters.

Examples
**Bernoulli:** The distribution of a Bernoulli (or binary) random variable $X$ is defined by a single parameter $p = \Pr(X = 1)$. 
Distribution function for Discrete Variables

The distribution of a discrete random variable $X$ is often described by a *probability mass function* (pmf), which assigns probability to each potential observed data value $x$. (Note use of upper case $X$ for random variable, lower case $x$ for fixed data value).

**Example**
For a Bernoulli variable $X$, its pmf is

$$
Pr(X = x) = \begin{cases} 
p & \text{if } x = 1, \\
1 - p & \text{if } x = 0.
\end{cases}
$$

Or more concisely, it can be written as:

$$
Pr(X = x) = p^x(1 - p)^{1-x}, \quad x = 0, 1.
$$
Distribution function for Continuous Variables

For a continuous variable $X$, the probability of $X = x$ for any particular $x$ is 0. So its distribution is described by a *probability density function* (pdf) which assigns probability to *intervals*.

![Diagram showing the probability density function and the event A]
Uniform Distribution

Example
The pdf of a Uniform(0, 1) random variable:

\[ P(0.3 \leq X \leq 0.7) \]
\[ P(X \leq 0.5 \text{ or } X > 0.8) \]
Uniform Distribution (cont)

Definition
In general, a Uniform(\(a, b\)) random variable \(X\) has pdf:

\[
f(x) = \frac{1}{b - a}, \quad \text{for any } x \text{ in the interval } (a, b).
\]

Thus the area under the curve is the area of a rectangle,

\[
Area = \text{width of interval} \times \text{height of density}
\]

\[
= (b - a) \times \frac{1}{b - a}
\]

\[
= 1
\]

as required for a sensible probability density!
Probability distributions have a central place in statistics. Having collected some data for some random variable(s), statistical analyses involve the following activities:

- *summarizing* and *presenting* the data so that it is easier to understand its empirical distribution
- *estimating* the parameters or functional forms of the distributions from the data
- *testing* to see if a theoretical distribution is a good fit to the data, or whether there is a difference in the distributions of two or more groups — the "analysis of variance" (ANOVA) problem
Thought Experiment: A Noncubic Die

- Can you guess the relation between $\Pr(6)$ and $r$? (Hint: what happens when $r = 1$, $r \to 0$ or $r \to \infty$?)
- How do you design an experiment to test your thesis?
  - Which values of $r$ to use?
  - How many times to roll each die? Should it depend on $r$?
  - What other factors that might influence the answer?
  - How should the data be reported?
  - What steps should be taken to conduct the experiment?
  - How should the die be rolled?