Mathematical Biostatistics Boot Camp: Lecture 2, Probability

Brian Caffo

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

August 29, 2012

Table of contents

Probability measures

A **probability measure**, P, is a real valued function from the collection of possible events so that the following hold

1. For an event $E \subset \Omega$, $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1$
3. If E_1 and E_2 are mutually exclusive events $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Additivity

Part 3 of the definition implies **finite additivity**

$$P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$$

where the $\{A_i\}$ are mutually exclusive.
This is usually extended to **countable additivity**

$$P(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$$

Note

- P is defined on $\mathcal{F}$ a collection of subsets of Ω
- Example $\Omega = \{1, 2, 3\}$ then

$$\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

- When Ω is a continuous set, the definition gets much trickier. In this case we assume that $\mathcal{F}$ is sufficiently rich so that any set that we're interested in will be in it.

Consequences

You should be able to prove all of the following:

- $P(\emptyset) = 0$
- $P(E) = 1 - P(E^c)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- if $A \subset B$ then $P(A) \leq P(B)$
- $P\left(A \cup B\right) = 1 - P(A^c \cap B^c)$
- $P(A \cap B^c) = P(A) - P(A \cap B)$
- $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$
- $P(\cup_{i=1}^n E_i) \geq \max_i P(E_i)$

Example

Proof that $P(E) = 1 - P(E^c)$

$$\begin{aligned} 1 &= P(\Omega) \\ &= P(E \cup E^c) \\ &= P(E) + P(E^c) \end{aligned}$$

$\blacksquare$

Example

Proof that $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$

$$\begin{aligned} P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &\leq P(E_1) + P(E_2) \end{aligned}$$

Assume the statement is true for $n - 1$ and consider n

$$\begin{aligned} P(\cup_{i=1}^n E_i) &\leq P(E_n) + P(\cup_{i=1}^{n-1} E_i) \\ &\leq P(E_n) + \sum_{i=1}^{n-1} P(E_i) \\ &= \sum_{i=1}^{n} P(E_i) \end{aligned}$$

$\blacksquare$

Example

The National Sleep Foundation (www.sleepfoundation.org) reports that around 3% of the American population has sleep apnea. They also report that around 10% of the North American and European population has restless leg syndrome. Similarly, they report that 58% of adults in the US experience insomnia. Does this imply that 71% of people will have at least one sleep problems of these sorts?

Example continued

Answer: No, the events are not mutually exclusive. To elaborate let:

$$
\begin{aligned}
A_1 &= \{\text{Person has sleep apnea}\} \\
A_2 &= \{\text{Person has RLS}\} \\
A_3 &= \{\text{Person has insomnia}\}
\end{aligned}
$$

Then (work out the details for yourself)

$$
\begin{aligned}
P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\
&- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\
&+ P(A_1 \cap A_2 \cap A_3) \\
&= .71 + \text{Other stuff}
\end{aligned}
$$

where the "Other stuff" has to be less than 0

Random variables

- A **random variable** is a numerical outcome of an experiment.
- The random variables that we study will come in two varieties, **discrete** or **continuous**.
- Discrete random variable are random variables that take on only a countable number of possibilities.
  - $P(X = k)$
- Continuous random variable can take any value on the real line or some subset of the real line.
  - $P(X \in A)$

Examples of variables that can be thought of as random variables

- The $(0 - 1)$ outcome of the flip of a coin
- The outcome from the roll of a die
- The BMI of a subject four years after a baseline measurement
- The hypertension status of a subject randomly drawn from a population

PMF

A probability mass function evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid pmf a function, p, must satisfy

1. $p(x) \geq 0$ for all x
2. $\sum_x p(x) = 1$

The sum is taken over all of the possible values for x.

Example

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads.

$$p(x) = (1/2)^x(1/2)^{1-x} \quad \text{for} \quad x = 0, 1$$

Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head expressed as a proportion (between 0 and 1).

$$p(x) = \theta^x(1-\theta)^{1-x} \quad \text{for} \quad x = 0, 1$$

Example

For the unfair coin

$p(0) = 1 - \theta$ and $p(1) = \theta$

so

$p(x) > 0$ for $x = 0, 1$

and

$p(0) + p(1) = \theta + (1 - \theta) = 1$

PDF

A probability density function (pdf), is a function associated with a continuous random variable

Areas under pdfs correspond to probabilities for that random variable

To be a valid pdf, a function f must satisfy

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Example

Assume that the time in years from diagnosis until death of persons with a specific kind of cancer follows a density like

$$f(x) = \begin{cases} \frac{e^{-x/5}}{5} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

More compactly written: $f(x) = \frac{1}{5}e^{-x/5}$ for $x > 0$.
Is this a valid density?

1. e raised to any power is always positive
2. $$\int_0^\infty f(x)dx = \int_0^\infty e^{-x/5}/5dx = -e^{-x/5}\Big|_0^\infty = 1$$

Example continued

What's the probability that a randomly selected person from this distribution survives more than 6 years?

Example continued

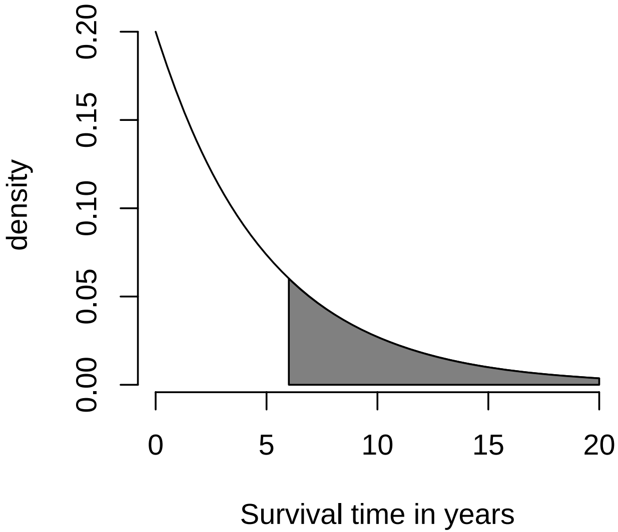
What's the probability that a randomly selected person from this distribution survives more than 6 years?

$$P(X \geq 6) = \int_6^\infty \frac{e^{-t/5}}{5} dt = -e^{-t/5}\Big|_6^\infty = e^{-6/5} \approx .301.$$

Approximation in R

```
pexp(6, 1/5, lower.tail = FALSE)
```

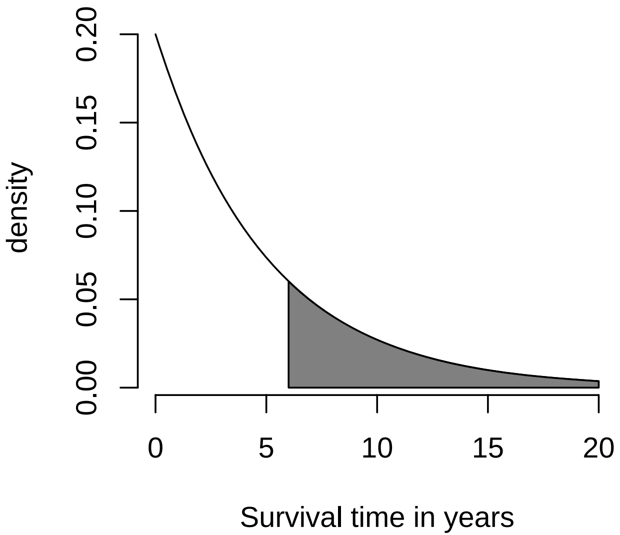
Example continued



CDF and survival function

- The **cumulative distribution function** (CDF) of a random variable X is defined as the function
$$F(x) = P(X \leq x)$$
- This definition applies regardless of whether X is discrete or continuous.
- The **survival function** of a random variable X is defined as
$$S(x) = P(X > x)$$
- Notice that $S(x) = 1 - F(x)$
- For continuous random variables, the PDF is the derivative of the CDF

Example continued



Example

What are the survival function and CDF from the exponential density considered before?

Example

What are the survival function and CDF from the exponential density considered before?

$$S(x) = \int_x^\infty \frac{e^{-t/5}}{5} dt = -e^{-t/5}\Big|_x^\infty = e^{-x/5}$$

hence we know that

$$F(x) = 1 - S(x) = 1 - e^{-x/5}$$

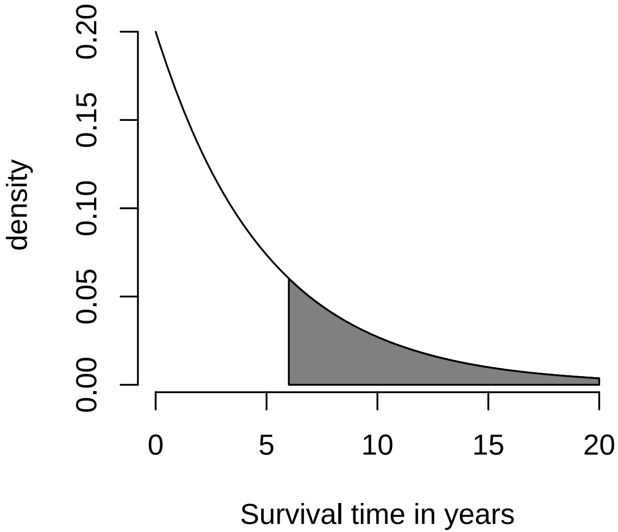
Notice that we can recover the PDF by

$$f(x) = F'(x) = \frac{d}{dx}(1 - e^{-x/5}) = e^{-x/5}/5$$

Quantiles

- The α^{th}**quantile** of a distribution with distribution function F is the point x_α so that
$$F(x_\alpha) = \alpha$$
- A **percentile** is simply a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile

Example continued



Example

- What is the 25^{th} percentile of the exponential survival distribution considered before?

Example

- What is the 25^{th} percentile of the exponential survival distribution considered before?
- We want to solve (for x)

$$
\begin{aligned}
.25 &= F(x) \\
&= 1 - e^{-x/5}
\end{aligned}
$$

resulting in the solution $x = -\log(.75) \times 5 \approx 1.44$

- Therefore, 25% of the subjects from this population live less than 1.44 years
- R can approximate exponential quantiles for you

`qexp(.25, 1/5)`

Probability models

- You might be wondering at this point "I've heard of a median before, it didn't require integration. Where's the data?"
- We're referring to are **population quantities**. Therefore, the median being discussed is the **population median**.
- A probability model connects the data to the population to the population using assumptions.
- Therefore the median we're discussing is the **estimand**, the sample median will be the **estimator**