Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Notation

- The **sample space**, Ω, is the collection of possible outcomes of an experiment
  - Example: die roll $\Omega = \{1, 2, 3, 4, 5, 6\}$
- An **event**, say E, is a subset of Ω
  - Example: die roll is even $E = \{2, 4, 6\}$
- An **elementary** or **simple** event is a particular result of an experiment
  - Example: die roll is a four, $\omega = 4$
- $\emptyset$ is called the **null event** or the **empty set**

Interpretation of set operations

Normal set operations have particular interpretations in this setting

1. $\omega \in E$ implies that E occurs when ω occurs
2. $\omega \notin E$ implies that E does not occur when ω occurs
3. $E \subset F$ implies that the occurrence of E implies the occurrence of F
4. $E \cap F$ implies the event that both E and F occur
5. $E \cup F$ implies the event that at least one of E or F occur
6. $E \cap F = \emptyset$ means that E and F are **mutually exclusive**, or cannot both occur
7. E^c or $\bar{E}$ is the event that E does not occur

Probability

A **probability measure**, P, is a 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 $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

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. (Note a more general version of additivity is used in advanced classes.)

Example consequences

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

Example

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. Does this imply that 13% 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:

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

Then

$$
\begin{aligned}
P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\
&= 0.13 - \text{Probability of having both}
\end{aligned}
$$

Likely, some fraction of the population has both.

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).

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

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. The area under $f(x)$ is one.

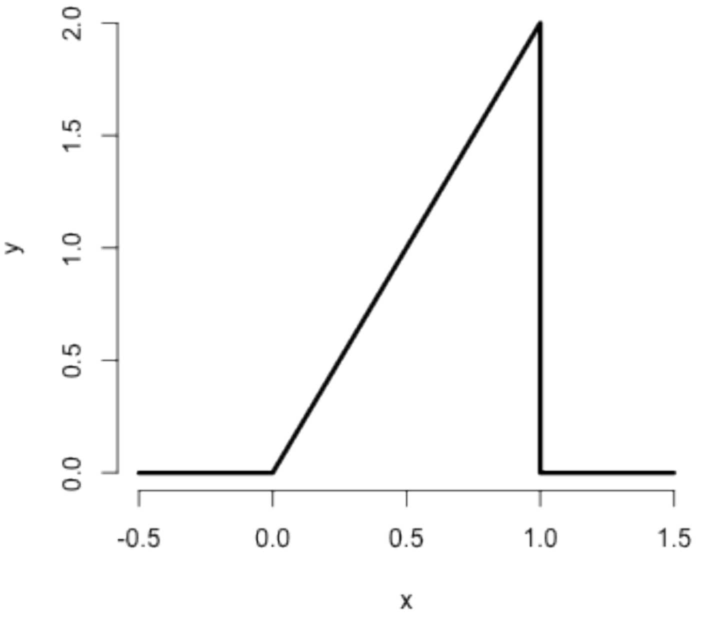
Example

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by

$$f(x) = \begin{cases} 2x & \text{for } 1 > x > 0 \\ 0 & \text{otherwise} \end{cases}$$

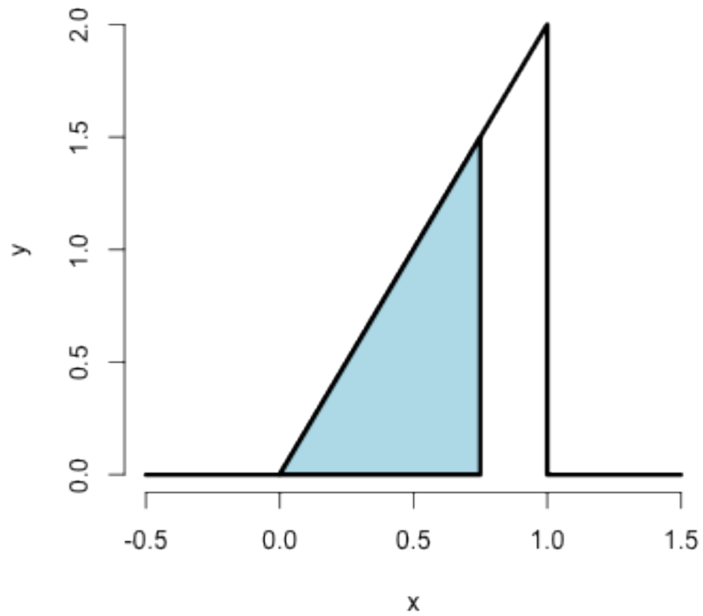
Is this a mathematically valid density?

```r
x <- c(-0.5, 0, 1, 1, 1.5)
y <- c(0, 0, 2, 0, 0)
plot(x, y, lwd = 3, frame = FALSE, type = "l")
```



Example continued

What is the probability that 75% or fewer of calls get addressed?



```
1.5 * 0.75/2
```

```
## [1] 0.5625
```

```
pbeta(0.75, 2, 1)
```

```
## [1] 0.5625
```

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

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

For $1 \geq x \geq 0$

$$F(x) = P(X \leq x) = \frac{1}{2} Base \times Height = \frac{1}{2}(x) \times (2x) = x^2$$

$$S(x) = 1 - x^2$$

```
pbeta(c(0.4, 0.5, 0.6), 2, 1)
```

```
## [1] 0.16 0.25 0.36
```

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

- We want to solve $0.5 = F(x) = x^2$
- Resulting in the solution

```
sqrt(0.5)
```

```
## [1] 0.7071
```

- Therefore, about 0.7071 of calls being answered on a random day is the median.
- R can approximate quantiles for you for common distributions

```
qbeta(0.5, 2, 1)
```

```
## [1] 0.7071
```

Summary

- 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 using assumptions.
- Therefore the median we're discussing is the **estimand**, the sample median will be the **estimator**