Probability

Statistical Inference

Brian Caffo, Jeff Leek, Roger Peng

Notation

The **sample space**, , is the collection of possible outcomes of an experiment
Example: die roll
An **event**, say , is a subset of
Example: die roll is even
An **elementary** or **simple** event is a particular result of an experiment
Example: die roll is a four,
is called the **null event** or the **empty set**

Interpretation of set operations

Normal set operations have particular interpretations in this setting

implies that occurs when occurs
implies that does not occur when occurs
implies that the occurrence of implies the occurrence of
implies the event that both and occur
implies the event that at least one of or occur
means that and are **mutually exclusive**, or cannot both occur
or is the event that does not occur

Probability

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

For an event ,

If and are mutually exclusive events .

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

where the are mutually exclusive. (Note a more general version of additivity is used in advanced classes.)

Example consequences

if then

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:

Then

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.

Continuous random variable can take any value on the real line or some subset of the real line.

Examples of variables that can be thought of as random variables

The 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, , must satisfy

for all

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

Example

Let be the result of a coin flip where represents tails and represents heads.

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

PDF

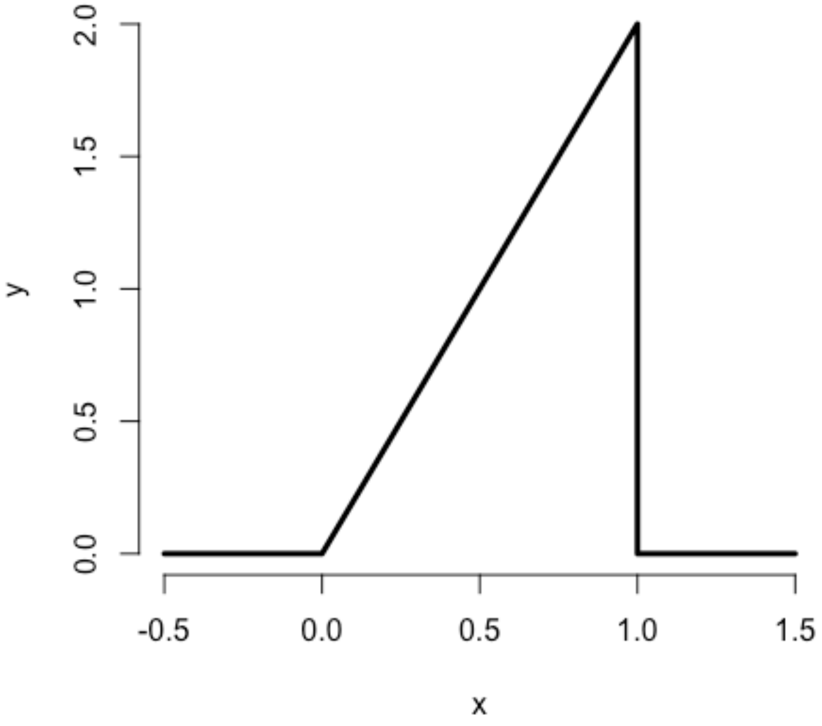
A probability density function (pdf), is a function associated with a continuous random variable

Areas under pdfs correspond to probabilities for that random variable

To be a valid pdf, a function must satisfy

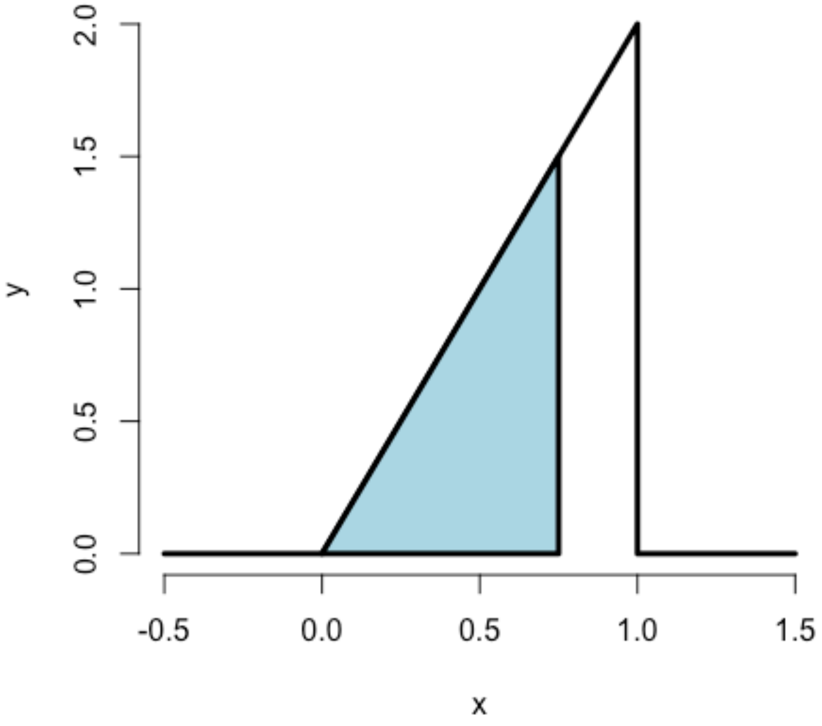
for all
The area under is one.

```
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")
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 * .75 / 2
```

```
1.5 * .75 / 2
## [1] 0.5625
pbeta(.75, 2, 1)
## [1] 0.5625
```

CDF and survival function

The **cumulative distribution function** (CDF) of a random variable is defined as the function

This definition applies regardless of whether is discrete or continuous.
The **survival function** of a random variable is defined as

Notice that
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

```
pbeta(c(0.4, 0.5, 0.6), 2, 1)
## [1] 0.16 0.25 0.36
```

Quantiles

The **quantile** of a distribution with distribution function is the point so that

A **percentile** is simply a quantile with expressed as a percent
The **median** is the percentile

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**