Course 6 - Statistical Inference - Week 1 - Notes

Greg Foletta

2019-11-22

# Intro

Statistical inference is:

The generation of conclusions about a population from a noisy sample

This class is focussed on frequentist styles of statistics.

Full course information is located at [on GitHub](https://github.com/bcaffo/courses/tree/master/06_StatisticalInference). Videos are on [YouTube](https://www.youtube.com/playlist?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ).

# Probability

Given a random experiment, a probability measure is a population quantity that summarises the randomness.

Probability takes a possible outcome from an experiment and:

* Assigns it a number between 0 and 1.
* The probability that **something** occurs is 1.
* The probability of the union of any two sets of outcomes that have nothing in common is the **sum** of their respective probabilities.
* The probability that nothing occurs is 0.
* The probability that something occurs is 1.
* The probability that something is 1 minus the probability that the opposite occurs.
* The probability of at least one of two (or more) things that cannnot simulaneously occur (mutually exclusive) is the sum of their respective probabilities.
* If an event A **implies** the occurence of event B, then the probability of A occurring is less than the probability that B occurs.
* For any two events the probability that *at least* one occurs is the sum of their probabilities minus their intersection:
  + You can’t just add probabilities if they have a non-trivial interaction.
  + You’re adding the intersection twice.

## Probability Mass Functions

Densities and mass functions for random variables are the best starting point.

**Random variable**: the numerical outcome of an experiment, can be **discrete** or **continuous**. Discrete variables have values, continuous variables have ranges.

A *probability mass function* (PMF) evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid PMF:

* It must always be larger than or equal to 0.
* The sum of the possible values the random variable can take has to add up to one.

It is used for discrete values.

### Bernoulli Distribution

$ X = 0 $ represents tails and $ X = 1 $ represents heads.

This is for a for a fair coin. Let be the probability of a head, and be the probability of a tail. We now have:

## Probability Density Functions

The probabilit density function is associated with continuous variables.

TO be valid density function:

* It must be larger than or equal to zero everywhere.
* The total area under it must be one.

Areas under a PDF correspond to probabilities for that random variable. The probability of a *specific* value is zero, because the area of a single line through the density function is zero. The bell curve is initially difficult to work with, so we’ll try a simpler version:

Imagine this is the proportion of help calls that get addressed by a help line on any given day. So the probability that between 20% and 60% of calls get addressed that day is given by the area between .2 and .6 on the x-axis, bounded by the slope of the line on the y-axis.

tibble(  
 x = seq(0, 1, 0.1),  
 y = 2\*x  
) %>%  
 ggplot(aes(x,y)) +  
 geom\_line()



Is this a mathematically valid density? It’s always bigger than 0, and its area is

.5 \* 1 \* 2

## [1] 1

So it is valid.

Back to our example, what’s the probability that 75% or fewer ([0%, 75%]) of calls are answered? That’s the area under the density function between 0 and .75.

This is the area of a smaller right angle triangle, so it’s:

.5 \* .75 \* 1.5

## [1] 0.5625

This is actually a special case of a beta distribution:

pbeta(.75, 2, 1)

## [1] 0.5625

# CDF and Survival Function

The **cumulative distribution function** of a random variable, , returns the probability that the random variable is less that or equal to the value :

This applied whether the variable is discrete or continuous.

The **survival function** of a random variable is defined as the probability that the random variable is greater than the value :

Notice that:

# Quantiles

The th quantile of a distribution function is the point so that .

A percentile is simply a quantile with expressed as a percent. The median is the 50th percentile.

Let’s find the median for our previous example:

qbeta(.5, 2, 1)

## [1] 0.7071068

This tells us that on - On 50% of the days, 70% or less of phone calls are answered. - On 50% of the days, 70% or more of phone calls are answered.

It’s important to note that here we’re talking about the **population quantiles**, therefore the median being discussed is the **population median**.

The median most people think of is the **sample median**, taking a sample, ordering them and taking the middle observation.

We need to look at estimators and estimands. Sample median estimating the population median. There’s assumptions that connect the sample to the population but we’re going to formally develop them.

**This is the formal process of statistical inference**. Linking the sample to a population.

# Conditional Probability

The probability of rolling a one on a die is one in six. If you’re given extra information that someone rolled the die and it was a one, three or five, conditional on this new information you’d now say the probability is one in three.

The definition:

* Let be an event so that .
* Then the probability of A given the event B has ocurred is the probability of the intersection of A and B, divided by the probability of B

In the event that A and B are unrelated, then .

Taking our previous die roll example,

$$
P(\text{one given roll is odd}) = P(A|B) \\
= \frac{P(A \cap B)}{P(B)} \\
= \frac{P(A)}{P(B)} \\
= \frac{1/6}{1/3} \\
= \frac{1}{3}
$$

Remembering that A is containted within B, hence the intersection is simply A.

## Bayes’ Rule

Bayes’ Rule allows us to reverse the role of the conditioning set and the set we want the probability of. We want P(B|A) when we have or easily can calculate P(A|B). Bayess rule can do it:

Where is the probability of B **marginalised** over A.

### Diagnostic Tests

* Let and be the events that the results of a diagnostic test is positive or negative.
* Let and be the event that the subject of the test has or does not have the disease respectively.

This is the marker of a good test, you want the sensitivity to be high.

The specificity is the probability the test is negative given that the subject does not have the disease. Again you want the specificity to be high for the test to be good.

If you have a positive test, the number that is of most concern to you is

This is the positive predictive value.

If you have a negative test, you are interested in:

This is the negative predictive value.

In the absenve of a test

is the prevalence of the disease.

Let’s take a made up example:

* Sensitivity of a test: 99.7%
* Specificity of a test: 98.5%
* Prevalence of disease: .01%

We can plug directly into Bayes’ rule. We note that the probability of a postive result, given that the person does not have the disease, is one minus the probability of a negative result given that the person doesn’t have the disease: . Similarly

$$ P(D|+) = \ = \

$$

(.997 \* .001) / ((.997 \* .001) + (.015 \* .999))

## [1] 0.06238268

So the positive predictive value is 6% for this test. The low positive predicive value is largely due to the low prevalence of the disease.

However imagine if there were other factors about the person the test was conducted on. The relevant prevalence may then be raised or lowered based on these factors. An example would be in an HIV test that the person let it be known they were an intravenous drug user. The prevalance has changed from the entire population to intravenous drug users.

### Likelyhood Ratios

Note that in both and , the denominator is exactly the same. If we divide the two equations we get:

Whenever you take a probability and divide it by one minus that probability you get the odds (this is the first item above).

On the left most side we have the odds of disease given a positive test result. On the right we have the odds of disease in the absence of the test result.

The factor in the middle is the diagnostic likelihood ratio of a positive test result. It is *the factor by which you multiply your odds in the presence of a positive test to obtain your post-test odds.*.

# Independence