

Bayesian-statistics

Manoj

7/23/2018

Basics of Bayesian Statistics

We can calculate “What is the probability of event A given event B?” using Bayesian. This is called conditional probability

Eg: False positive or False negative rates.

Conditional Probability and Bayes’ rule

Polls : Have you even used dating websites?


2015 Gallup poll on use of online dating sites:

		Age				
		18-29	30-49 <i>B</i>	50-64	65+	Total
Used online dating site	Yes <i>A</i>	60	86	58	21	225
	No	255	426	450	382	1513
Total		315	512	508	403	1738

$$P(A | B) = \frac{P(A \& B)}{P(B)}$$

% of 30-49 year olds using online dating sites =
 $\frac{86}{512} \approx 0.17$
A & B B
P(use online dating site | 30-49 year old)

Source: <http://www.pewinternet.org/2016/02/11/15-percent-of-american-adults-have-used-online-dating-sites-or-mob>

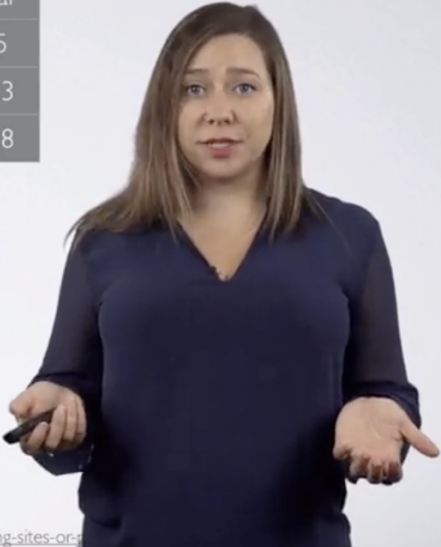


2015 Gallup poll on use of online dating sites:

		Age				
		18-29	30-49	50-64	65+	Total
Used online dating site	Yes	60	86	58	21	225
	No	255	426	450	382	1513
Total		315	512	508	403	1738

$P(\text{use online dating site} | 30-49 \text{ year old}) =$
 $= \frac{P(\text{use online dating site \& 30-49 year old})}{P(30-49 \text{ year old})}$
 $= \frac{86 / 1738}{512 / 1738} = \frac{86}{512} \approx 0.17$

Source: <http://www.pewinternet.org/2016/02/11/15-percent-of-american-adults-have-used-online-dating-sites-or-s>



Bayes Rule and Diagnostic Testing

US Military early HIV testing in military

1. Elisa screen
2. if positive, two more ELISA
3. if either positive, two western blot assays
4. if positive, HIV infection

ELISA -sensitivity (true positive): 93% $p(+/\text{HIV}) = 0.93$ -specificity (true negative) : 99% $p(-/\text{No HIV}) = 0.99$

Western blot -sensitivity (true positive): 99.9% -specificity (true negative) : 99.1%

prevalence : 1.48 / 1000 people $p(\text{HIV}) = 0.0148$ which is prior

This updating scheme we have here is general property of the Bayesian models

Bayesian and Frequentist definitions of probability

Frequentist: Its relative frequency in large number of trials Bayesian : Probability of an event happening is equated to another event

Confidence Interval : The proportion of random samples of size n from the same population that produce confidence interval that contain the true population parameter.

Credible Interval : We can express the true parameter not as a fixed value but with a probability This will let us construct something like a credible intervals expect we can make probabilistic statements about the parameter falling within that range

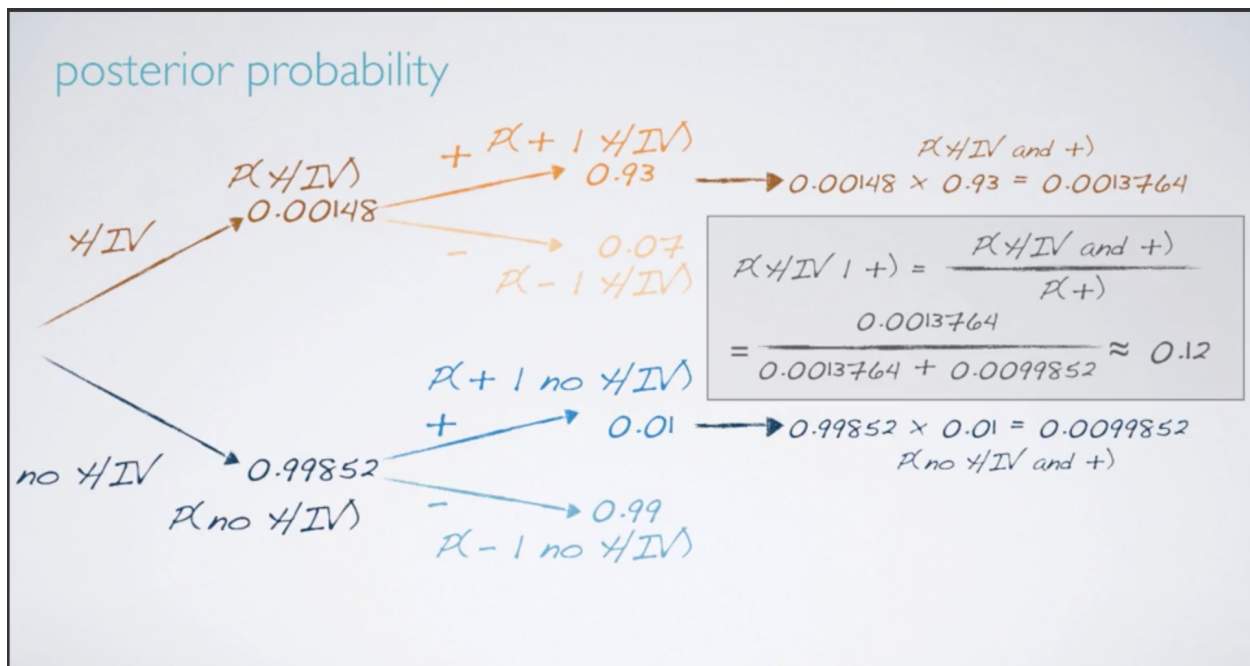


Figure 1: Bayes-tree

Inference for proportions :

RU-486 effective or not? 40 women are divided into two groups with one taking RU 486 and the other standard therapy. 4/20 got pregnant where RU 486 used and 16/20 got pregnant where standard methods are used. How strongly does the data indicate RU 486 is effective? This is a two proportion problem but we can frame it as one proportion test as following: How likely that 4 pregnancies occur in treatment group?

Frequentist Approach :

p = probability that a given pregnancy comes from a treatment group

$H_0 : p = 0.5$ - No difference, pregnancy is equally likely to come from the treatment or control group

$H_A : p > 0.5$ - treatment is more effective, a pregnancy is less likely to come from the treatment group

Calculating P value:

$k = 4$, $n = 20$ $p = 0.5$ assuming H_0 is true

We need to calculate the p-value as obtaining 4 or fewer success in 20 trials where probability of success is 0.5. The number of success in a fixed number of independent trials for a categorical variables with two levels follows a binomial distribution with two parameters.

p value = $P(k \leq 4)$

```
sum(dbinom(0:4, 20, 0.5))
```

```
## [1] 0.005908966
```

With such a small probability, we reject the null hypothesis

Bayesian Approach :

Step 1: We have to create hypothesis or models. Let's assume p could be ranging from 10% to 90%. $p = 20\%$
: Given a pregnancy occurs there is a 1:4 chance that it will occur in the treatment group

Step 2: Set Priors. Use experience, previous experiments, research etc.,

Step 3: Calculate Likelihood, $P(\text{data}/\text{model}) = P(k = 4 \mid n = 20, p)$

```
p <- seq(from = 0.1, to = 0.9, by = 0.1)
prior <- c(rep(0.06, 4), 0.52, rep(0.06, 4))
likelihood <- dbinom(4, size = 20, prob = p)
```

Step 4: Calculate Posterior,

$p(\text{model} \mid \text{data}) = p(\text{model} \ \& \ \text{data}) / p(\text{data})$

$= p(\text{data} \mid \text{model}) * p(\text{model}) / p(\text{data})$

```
numerator <- prior * likelihood
denominator <- sum(numerator)
posterior <- numerator / denominator
```

Effect of sample size on the posterior

As we increase the sample size, there will be less variability in the likelihood obtained from binomial distribution and hence posterior probability also becomes more certain

Frequentist vs Bayesian Inference:

M & Ms are either 10% or 20% in the population?

Frequentist Inference:

H_0 : 10% Yellow M&Ms H_A : > 10% Yellow M&Ms

We cannot set the value of alternate hypothesis equal to something

Significance level = 0.05, probability of rejecting a null hypothesis Sample : $K = 1, n = 5$

```
1 - dbinom(0, 5, prob = 0.10)
```

```
## [1] 0.40951
```

Fail to reject null hypothesis

Bayesian Inference:

H_1 : 10% Yellow M&Ms H_2 : > 10% Yellow M&Ms

priors : $P(H_1) = 0.5, p(H_2) = 0.5$ obs. data : $k = 1, n = 5$

```
p1 <- dbinom(1, 5, p = 0.1)
p2 <- dbinom(1, 5, p = 0.2)
```

```
prior <- 0.5  
Ph1_data <- prior * p1 / ( prior * p1 + prior * p2)  
Ph2_data <- prior * p2 / ( prior * p1 + prior * p2)
```

We would pick 20 % in this case which is contradictory.

Doubts

What is likelihood in the Bayes tree? I think it is the sensitivity and specificity of the test.