Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Bayesian Analysis

Bayesian inference

- Bayes' theorem is the basis of the Bayesian Statistics.
- It describes the probability of an event, based on prior knowledge of conditions that might be related to the event
  - For example, if the risk of developing health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately (by conditioning it on their age) than simply assuming that the individual is typical of the population as a whole.
- Bayesian inference regarding the population proportion
  - an example for the application of Bayesian methods.

A Simple Case of Bayesian Analysis for Population Proportion

- Suppose that we are interested in finding the five-year survival rate (i.e., population proportion of survival) among breast cancer patients.
  - We denote this unknown population proportion μ.
  - For simplicity, we assume that the survival rate is either 0.75 or 0.85.
- Without any data, we think that both of these values are equally probable;
  - that is, $P(\mu = 0.75) = P(\mu = 0.85) = 0.5$.
  - Alternatively, we can write this as follows:

$$\frac{P(\mu = 0.85)}{P(\mu = 0.75)} = 1$$

A Simple Case of Bayesian Analysis for Population Proportion

- Now suppose that we take a random sample of $n = 20$ breast cancer patients from the population.
  - We use Y to denote the number of survivals out of 20.
- We know that Y has a Binomial(n,μ) distribution
  - assuming that the patients are selected independently and they all have the same probability of survival.
- Therefore,
  - if $\mu = 0.75$, the distribution of Y is Binomial(20, 0.75).
  - if $\mu = 0.85$, the distribution of Y is Binomial(20, 0.85).

A Simple Case of Bayesian Analysis for Population Proportion

- For our sample, we find that 18 of patients are still alive after five years: $Y = 18$.
- Our point estimate for the survival rate μ is therefore $p = 18/20 = 0.9$,
  - where p : the sample proportion.
- To see how this information changes our mind about the value of the population proportion, μ, we use Bayes' theorem.
  - Recall that Bayes' formula for two events E_1 and E_2 is

$$P(E_2|E_1) = \frac{P(E_1|E_2) \times P(E_2)}{P(E_1)}$$

A Simple Case of Bayesian Analysis for Population Proportion

- For the example, we can write Bayes' formula in terms of μ and Y, so E_1 corresponds to the event that $Y = 18$, and E_2 corresponds to the event that $\mu = 0.85$:

$$P(\mu = 0.85|Y = 18) = \frac{P(Y = 18|\mu = 0.85) \times P(\mu = 0.85)}{P(Y = 18)}$$

  – $P(\mu = 0.85|Y = 18)$ is the probability that the true value of the survival rate is 0.85 given the information that 18 people have survived (out of 20),
  – $P(Y = 18|\mu = 0.85)$ is the probability that 18 people survive assuming that the true survival rate is 0.85,
  – $P(\mu = 0.85)$ is the probability that the survival rate is in fact 0.85 (we assumed this probability is 0.5),
  – $P(Y = 18)$ is the probability that 18 people survive regardless of what the probability of survival is (0.85 or 0.75).

A Simple Case of Bayesian Analysis for Population Proportion

- In R-Commander, apply the following steps
  - to install RC: install.packages("Rcmdr", dependencies=TRUE)
  - to run RC: library(Rcmdr)

```
   Probability
0  3.325257e-17
1  3.768624e-15
2  2.028776e-13
3  6.897839e-12
4  1.661229e-10
5  3.012363e-09
6  4.267514e-08
7  4.836516e-07
8  4.453625e-06
9  3.364961e-05
10 2.097492e-04
11 1.080526e-03
12 4.592237e-03
13 1.601396e-02
14 4.537287e-02
15 1.028452e-01
16 1.821217e-01
17 2.428289e-01
18 2.293384e-01
19 1.367983e-01
20 3.875953e-02
>
```

- Click Distributions → Discrete distributions → Binomial distribution → Binomial probabilities.
- Then, set Binomial trials to 20 and Probability of success to 0.85
- The probabilities for all possible values of Y will be obtained.
- The probability for 18 survivals assuming that $\mu = 0.85$ is $P(Y = 18|\mu = 0.85) = 0.23$.

A Simple Case of Bayesian Analysis for Population Proportion

- To find $P(Y = 18)$, use the law of total probability:

$$P(Y = 18) = P(Y = 18|\mu = 0.85) \times P(\mu = 0.85)$$
$$+P(Y = 18|\mu = 0.75) \times P(\mu = 0.75)$$

- To find the probability of 18 survivals assuming that the true value of the survival rate is $\mu = 0.75$, repeat the same steps, setting Binomial trials to 20 and Probability of success to 0.75;

$$P(Y = 18|\mu = 0.75) = 0.07$$

- Therefore,

$$P(Y = 18) = 0.23 \times 0.5 + 0.07 \times 0.5 = 0.15$$

A Simple Case of Bayesian Analysis for Population Proportion

- Now we can find $P(\mu = 0.85|Y = 18)$:

$$P(\mu = 0.85|Y = 18) = \frac{P(Y = 18|\mu = 0.85) \times P(\mu = 0.85)}{P(Y = 18)}$$
$$= \frac{0.23 \times 0.5}{0.15} = 0.76$$

- At the beginning (before observing any data), we believed that $\mu = 0.85$ with probability of 0.5.
- Knowing that 18 out of 20 people have survived, we increase this probability to 0.76.

A Simple Case of Bayesian Analysis for Population Proportion

- Following similar steps, we find that

$$P(\mu = 0.75|Y = 18) = 0.24$$

- Given the observed data, we have reduced the probability of $\mu = 0.75$ from 0.5 to 0.24.
- Therefore, while we gave equal probabilities to both values 0.75 and 0.85 at the beginning, based on the new empirical evidence we observed, we increased the probability of $\mu = 0.85$ and decreased the probability of $\mu = 0.75$.
  – This is intuitive, because our point estimate for the survival rate is 0.9 (18 out of 20), which is closer to 0.85 than 0.75.
- We can use these updated probabilities and write

$$\frac{P(\mu = 0.85|Y = 18)}{P(\mu = 0.75|Y = 18)} = \frac{0.76}{0.24} = 3.2$$

- Therefore, given the observed data, the value 0.85 is 3.2 times more likely than 0.75

Prior and Posterior Probabilities

- In our example, $P(\mu = 0.75)$ and $P(\mu = 0.85)$ are referred to as prior probabilities for the population proportion μ.
  – These are probabilities we assign to possible values of μ before observing any data.
  – In practice, these probabilities might be obtained from previous studies.
    - For example, two other research groups might have conducted similar studies in the past;
      – one group estimated μ to be 0.75, and the other group estimated it to be 0.85, and we do not have any reason to prefer one estimate over the other.
    - In this case, we want to conduct a new study, collect new empirical evidence, and estimate μ, but we want to take the available information regarding the value of μ into account.

Prior and Posterior Probabilities

- $P(Y = 18|\mu = 0.85)$ is referred to as likelihood,
  - i.e., how likely it is to see this specific data (18 survivals out of 20) if μ is in fact 0.85.
- We can express the probability of the specific data we have observed (i.e., 18 survivals out of 20) as a function of different values of μ.
  - This function is referred to as the likelihood function.
  - For our example, the likelihood function is

$$P(Y = 18|\mu) = \begin{cases} 0.07 & \mu = 0.75 \\ 0.23 & \mu = 0.85 \end{cases}$$

Prior and Posterior Probabilities

- The updated probability of μ after we observe the data is referred to as the posterior probability of μ.
- The posterior probabilities in our example are
  - $P(\mu = 0.75|Y = 18) = 0.24$
  - $P(\mu = 0.85|Y = 18) = 0.76$
- These posterior probabilities, which are obtained after we observed 18 survivals among 20 patient, can be used to write

$$\frac{P(\mu = 0.85|Y = 18)}{P(\mu = 0.75|Y = 18)} = \frac{0.76}{0.24} = 3.2$$

- This is known as the posterior odds
  - Here, we find the odds of 0.85 over 0.75.

Prior and Posterior Probabilities

- The posterior odds can be found as follows:

$$\frac{P(\mu = 0.85|Y = 18)}{P(\mu = 0.75|Y = 18)} = \frac{P(Y = 18|\mu = 0.85) \times P(\mu = 0.85)/P(Y = 18)}{P(Y = 18|\mu = 0.75) \times P(\mu = 0.75)/P(Y = 18)}$$
$$= \frac{P(Y = 18|\mu = 0.85) \times P(\mu = 0.85)}{P(Y = 18|\mu = 0.75) \times P(\mu = 0.75)}$$
$$= \frac{P(Y = 18|\mu = 0.85)}{P(Y = 18|\mu = 0.75)} \times \frac{P(\mu = 0.85)}{P(\mu = 0.75)}$$

- The term $P(\mu = 0.85)/P(\mu = 0.75)$ on the right-hand side of the above equation is called prior odds.
  - In our example, the prior odds is 1
- The posterior odds is obtained by multiplying the prior odds by the following term:

$$\frac{P(Y = 18|\mu = 0.85)}{P(Y = 18|\mu = 0.75)} = \frac{0.23}{0.07} = 32.86$$

- This term is in fact the ratio of two possible values for the likelihood function and is known as the likelihood ratio.

General Form of Bayesian Analysis for Population Proportion

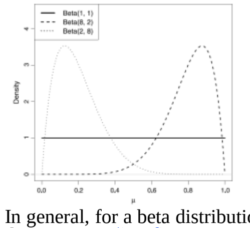
- In general, the population proportion could take values from 0 to 1.
  - Therefore, we need a continuous prior distribution whose range is from 0 to 1.
- The beta distribution, whose range is from 0 to 1, is commonly used as the prior distribution for the population proportion μ.
  - The beta distribution is specified by two parameters, α and β, and is denoted as Beta(α,β).
  - We refer to α and β as shape 1 and shape 2, respectively.
    - Both parameters must be positive numbers.

General Form of Bayesian Analysis for Population Proportion

- In R-Commander, we can plot different beta distributions by setting α and β to different values.
- For example, suppose that we want to plot Beta(8, 2).
- In R-Commander, click Distributions → Continuous distributions → Beta distribution → Plot beta distribution and set Shape 1 and Shape 2 to 8 and 2, respectively.
- Make sure the option Plot density function is checked and press OK.

General Form of Bayesian Analysis for Population Proportion

- Comparing the plots of the probability density function for a beta distribution with different parameter values.



- The *solid line* represents the pdf of Beta(1,1).
- This distribution is known as the Uniform(0,1) distribution.
- The *dashed line* represents the pdf of Beta(8, 2), and the *dotted line* represents the pdf of Beta(2, 8)
- In general, for a beta distribution with parameters α and β, the mean is $\alpha/(\alpha+\beta)$.
  - For example, the mean of the Beta(2, 8) is 2/(2 +8) = 0.2.

General Form of Bayesian Analysis for Population Proportion

- Reconsider the breast cancer survival example.
  - Instead of assuming that only two values are possible, assume that the true population proportion could be any value from 0 to 1
- In general, it is recommended to avoid making overly restrictive assumptions such as the one we used for illustrative purposes earlier.
  - That is, even if previous studies estimated the population proportion to be either 0.75 and 0.85, we still should consider all other feasible values.

General Form of Bayesian Analysis for Population Proportion

- We could of course use the results from previous studies and assume that while the survival rate could be any value from 0 to 1, it is more likely to be around 0.8 .
- When specifying the prior distribution, we can use a beta distribution that reflects this assumption.
  - For the Beta(8,2) distribution (dashed curve in the figure in slide 18), the probability (i.e., the area under the density curve) is high for values around 0.8, whereas the probability is almost zero for values around 0.2.
- Therefore, we use Beta(8,2) as the prior distribution for the survival rate of breast cancer patients.

General Form of Bayesian Analysis for Population Proportion

- Note that this prior probability distribution reflects our knowledge (based on previous studies) regarding the possible values of survival rate before we obtain new data.
  - We update our knowledge after we observe new empirical evidence.
- Our updated knowledge is expressed as the posterior probability distribution, which could be drastically different from the prior probability distribution.
  - Therefore, even though we believe in prior that the survival rate is around 0.8, a new empirical evidence could overwhelmingly change this belief.
  - We might be even convinced that values around 0.2 are more probable than values around 0.8 if the observed data strongly suggest that.

General Form of Bayesian Analysis for Population Proportion

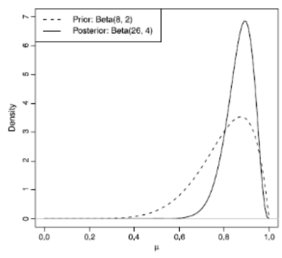
- To find the posterior probability (PP) distribution, we use Bayes' theorem as before.
  - PP Distribution is a beta distribution with updated parameters
- If we assume that the prior knowledge of the population proportion μ, can be expressed using a Beta(α,β) distribution, then the posterior distribution of μ is Beta($\alpha + y$,$\beta + n - y$),
  - where n is the sample size, and y is the number of times the event of interest has been observed.

General Form of Bayesian Analysis for Population Proportion

- In our example, we obtained a sample of 20 patients from the population and found that 18 of them survived after 5 years.
- Assuming that the prior probability distribution for the breast cancer survival rate is Beta(8,2), the posterior probability distribution for the survival rate is Beta(8+18, 2+20−18).
- We can use R-Commander to plot the probability density function for this distribution by following the steps described earlier, but this time we set Shape 1 and Shape 2 to 26 and 4, respectively.

General Form of Bayesian Analysis for Population Proportion

- The density curve for the posterior probability distribution, Beta(26, 4)



- The prior probability distribution (*dashed curve*) for breast cancer survival rate and the resulting posterior probability distribution (*solid curve*) after observing 18 survivals among 20 patients