Eco311 Optional Reading: Bayesian Statistics for Proportion

(Jing Li, Miami University)

1. Recall: according to the central limit theorem (CLT), the sample mean from a large iid sample approaches a normal random variable

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad (as\ n \to \infty) \tag{1}$$

After standardization, it follows that

$$z\ score\ \equiv \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \tag{2}$$

and

$$P\left(-1.96 < \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95 \tag{3}$$

which implies the 95 confidence interval for μ

$$P\left(\bar{y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{4}$$

Thus, the sample mean plus and minus 1.96 time standard error gives the 95 confidence interval for population mean μ.

2. For a Bernoulli distribution (like flipping a coin), Q2 in HW1 proves that population mean is equal to population proportion

$$\mu = P(y=1) \equiv p \tag{5}$$

Therefore, formula (4) also implies the 95 confidence interval for population proportion

3. Some people find formula (4) conceptually awkward: we talk about probability only for a random variable. But μ is a constant, so formula (4) makes no sense to those skeptical people.

4. Those people belong to Bayesian school, who believe it is better to treat μ as a random variable as opposed to a constant. They believe some values of μ are more likely (credible or believable) than other values if those values are supported by data or reality. Their job is to find those likely values.

5. This note uses a story to explain how to estimate population proportion with Bayesian approach.

6. Suppose Tom is a billionaire and he is interested in buying a lake from which he can catch bass fish and eat them. A fish can be bass or not bass. So we define a Bernoulli random variable as

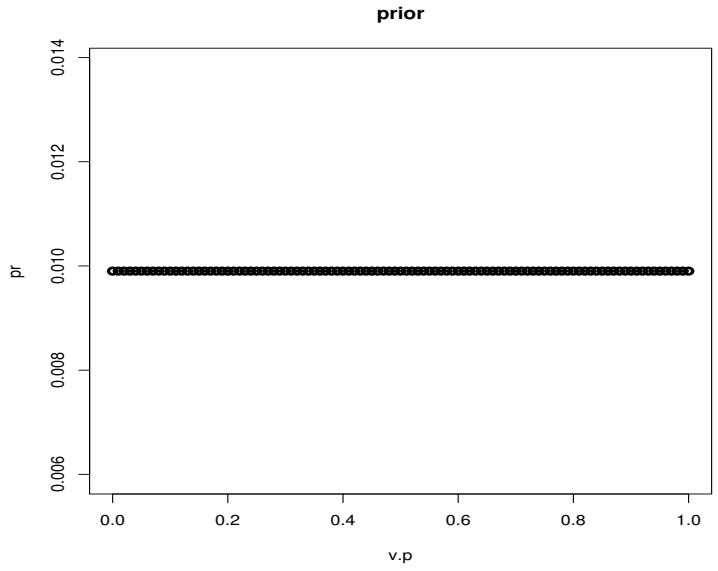
$$y = \begin{cases} 1 & \text{if a fish is bass, probability is } p \\ 0 & \text{if a fish is not bass, probability is } 1-p \end{cases} \tag{6}$$

   Tom hopes to know p is how much. He will buy the lake if proportion of bass is big enough.

7. Assume he knows nothing about the fish population in that lake. Using jargon, his prior belief is that p can take any value of $(0, 0.01, 0.02, ...1)$ with equal probability. For instance, $p = 0.05$ means out of 100 fishes, five are bass. At beginning, he assumes $p = 0.05$ is as likely as $p = 0.67$, or any other value.

8. Using jargon, his prior distribution about p is a uniform distribution. We can use following R codes to draw the prior distribution

```
> v.p = seq(0, 1, by=0.01)
> pr = rep(1/length(v.p),length(v.p))
> plot(v.p, pr, type = "b", lwd = 3, pch = 1, main="prior")
```



where we put possible values for p on horizontal axis, and their heights measure likelihood or credibility. The flat line indicates that every value in the vector $(0, 0.01, 0.02, ...1)$ is equally likely. In other words, before seeing data Tom has no reason to believe some values are more likely than others.

9. Of course this belief of equal likelihood is naive. It can only be used as a starting point. So next Tom tries to catch some fish in that lake (get some data), and use data to update his belief about bass proportion

10. Suppose on the first day he catches 10 fish, and three of them are bass. The key question asked by Bayesian school is, how to use this finding or information to update belief about p. Is it still plausible to believe every value is equal likely? The answer is no, since this finding supports $p = 0.3$ more than, say, $p = 0.8$. The likelihood function explained next formalizes this thinking.

11. For a given value of p, the number of bass out of 10 catches follows Binomial distribution. Another way to describe the Binomial distribution is, suppose there are 10 trails. The number of successes follows Binomial distribution.

$$P(x \texttt{ bass out of 10 catches}) = C_{10}^{x} p^{x} (1-p)^{n-x}, \quad (x = 0, 1, ...10) \tag{7}$$

For instance, if $p = 0.3$, the probability of 3 out of ten catches being bass is 0.2668279. That number is given by the following R codes

```
> dbinom(3, 10, 0.3)
[1] 0.2668279
> choose(10,3)*0.3^3*0.7^7
[1] 0.2668279
```

The last result verifies that $C_{10}^{3} 0.3^{3} (1-0.3)^{10-3} = 0.2668279$

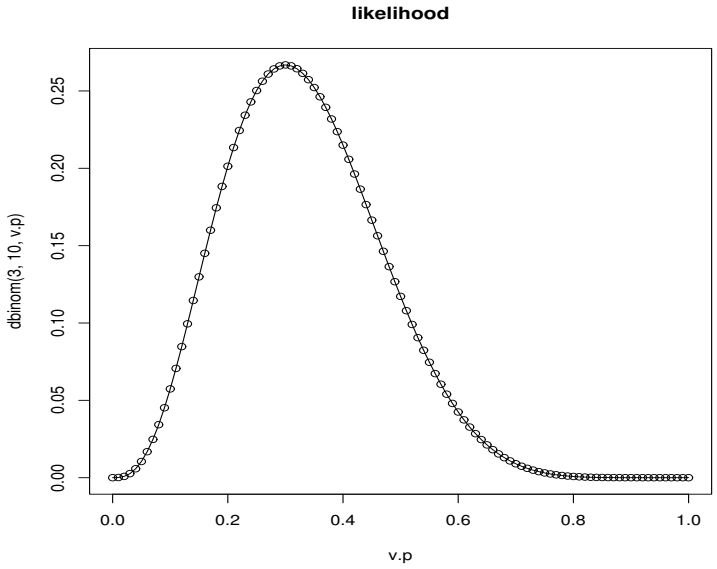
12. We can show if $p = 0.8$, the probability of 3 out of ten catches being bass is

```
> dbinom(3, 10, 0.8)
[1] 0.000786432
```

Obviously, 0.000786432 is less than 0.2668279. So having 3 out of ten catches being bass supports $p = 0.3$ more than $p = 0.8$.

13. Formally, we call $C_{10}^x p^x (1-p)^{n-x}$ the likelihood function when we hold n and x constant while let p vary. In other words, likelihood is function of p for given data. For example, we can plot the likelihood function for given $n = 10, x = 3$ as

```
> plot(v.p, dbinom(3, 10, v.p), type="o", main="likelihood")
```



Note that we put possible values of p on the horizontal axis. The height still measures likelihood or credibility. The likelihood function reaches peak at $p = 0.3$, implying that $p = 0.3$ gets more support from data than other values.

14. Next we put prior and likelihood together, and apply the Bayes rule to update belief about p

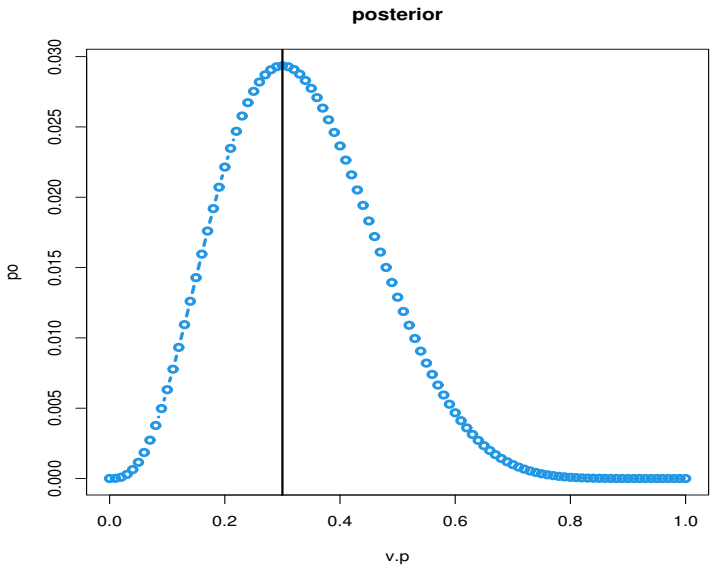
$$\texttt{posterior} \equiv P(\theta|data) = \frac{P(\theta)P(data|\theta)}{P(data)} = \frac{P(\theta)P(data|\theta)}{\sum_\theta P(\theta)P(data|\theta)} \propto \texttt{prior}\times\texttt{likelihood} \tag{8}$$

Here, the unknown parameter θ is the bass proportion p. $P(\theta|data)$ denotes the updated belief about θ after seeing the data, or posterior distribution; $P(\theta)$ denotes the original belief about θ prior to seeing the data; $P(data|\theta)$ denotes the likelihood function (credibility for each value of θ based on the given data).

15. In short, Bayes rule states that the posterior is proportional to the product of prior and likelihood. We can ignore the denominator $P(data) \equiv \sum_\theta P(\theta)P(data|\theta)$ since it is free of θ.

16. The R codes to obtain the posterior distribution of p is

```
> lik = choose(10,3)*v.p^3*(1-v.p)^7
> po = pr*lik
> po = po/sum(po)
> plot(v.p, po, type = "b", col = 4 , lwd = 3, pch = 1, main="posterior")
> abline(v=0.3, lwd = 2)
> v.p[po==max(po)]
[1] 0.3
> po[v.p==0.8]
[1] 8.650751e-05
> po[v.p==0.2|v.p==0.3|v.p==0.4]
[1] 0.02214592 0.02935107 0.02364899
```



Note that posterior is a function of p—the horizontal axis is about p, not x. It shows that, given the observed fact that there are 3 bass fish out of 10 catches, the most likely value (updated belief) for bass proportion is $p = 0.3$, which is denoted by a vertical line. By contrast, $p = 0.8$ is very unlikely (inconsistent with having 3 bass out of 10 catches) since its height 8.650751e-05 is almost zero. In fact, $p = 0.2$ and $p = 0.4$ are very likely as well (their heights are 0.02214592, 0.02364899).

17. You may notice that posterior looks almost the same as likelihood. Actually here likelihood is proportional to posterior since we use uniform prior in this example.

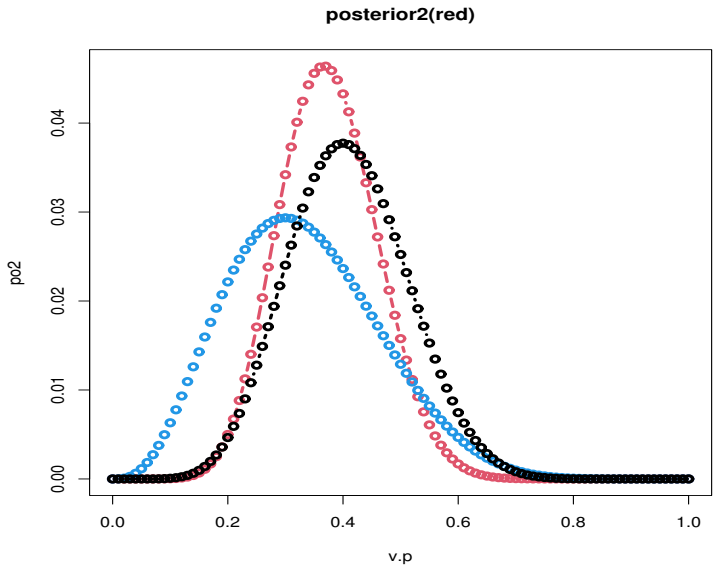
```
> lik[2]/po[2]
[1] 9.09091
> lik[v.p==0.3]/po[v.p==0.3]
[1] 9.09091
```

18. So after the first day, Tom's got this answer—most likely the bass proportion is 0.3. It is also highly possible that bass proportion is 0.2, 0.4, or any value between 0.2 and 0.4.

19. Bayesian statistics is dynamic in the sense that we can keep updating belief after new data or evidence arrives.

20. For instance, suppose on the second day, Tom catches 20 fish, and 8 of them are bass. The R codes below aim to update the estimate of bass proportion and obtain the second round posterior distribution

```
> pr2 = po
> lik2 = choose(20,8)*v.p^8*(1-v.p)^12
> po2 = pr2*lik2
> po2 = po2/sum(po2)
> plot(v.p, po2,  type = "b", col = 2 , lwd = 3, pch = 1,, main="posterior2 (red)")
> lines(v.p, po, type = "b", col = 4 , lwd = 3, pch = 1)
> lines(v.p, lik2/sum(lik2), type = "b", lwd = 3, pch = 1)
> v.p[po2==max(po2)]
[1] 0.37
```



Since the first-day finding is informative, here we use the first round posterior distribution as the new prior distribution. The new likelihood is based on the latest finding that this time Tom catches 20 fish, and 8 of them are bass. Then the Bayes formula (8) is applied again to obtain the new posterior distribution (red color). For easy comparison, the old posterior or new prior (blue color), and normalized second-round likelihood (black color) are included in the graph.

21. The are several takeaways. First, the peak of red line is located at $p = 0.37$. This is the second-round estimate for the bass proportion. Notice that 0.37 is between the first-day sample proportion $0.3 = \frac{3}{10}$ and the second-day sample proportion $0.4 = \frac{8}{20}$. Or, the mode of red curve is a weighted average of mode of blue curve and mode of black curve.

$$0.37 = w0.4 + (1 - w)0.3 \;\Rightarrow\; w = 0.7$$

In this case, the weight for 0.4 is greater than the weight for 0.3 because the evidence supporting 0.4 (8 out of 20 being bass) is more convincing than the evidence supporting 0.3 (3 out of 10 being bass). To see this mathematically, compare the slope of likelihood based on a small-sample finding (3 out of 10 being bass) and a big-sample finding (6 out of 20 being bass)

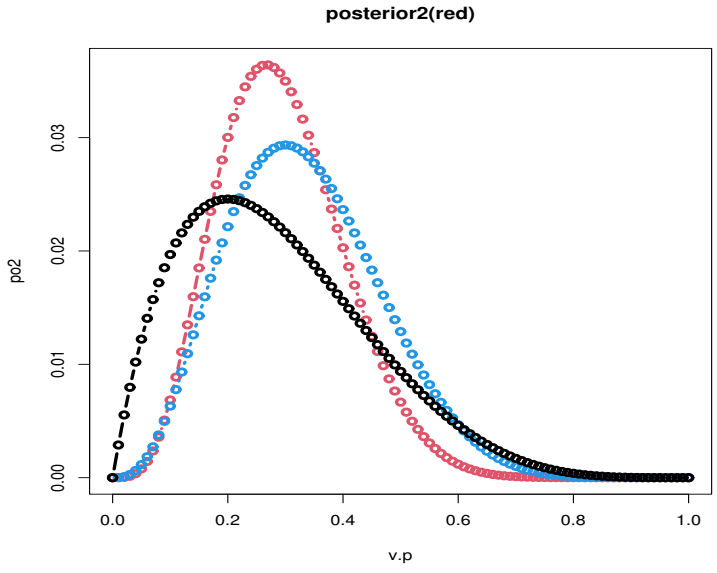
```
> dbinom(3, 10, 0.3)-dbinom(3, 10, 0.2)
[1] 0.06550134
> dbinom(6, 20, 0.3)-dbinom(6, 20, 0.2)
```

```
[1] 0.08253928
```

In words, ***Bayesian statistics automatically takes both new and old evidence into account. The weights for old and new evidence are endogenously assigned according to how convincing evidence is***.

22. Second, the red distribution is narrower than the blue distribution. Therefore, we have more confidence for the new estimate $p = 0.37$ than the old estimate $p = 0.3$. The posterior distribution gets narrower simply because new information has been added to old information. More information leads to preciser estimate (narrower posterior). As another example, suppose Tom only catches 5 fish on the second day, and only one is bass. Now the second-round posterior (red) looks like

```
> pr2 = po
> lik2 = choose(5,1)*v.p^1*(1-v.p)^4
> po2 = pr2*lik2
> po2 = po2/sum(po2)
> plot(v.p, po2,  type = "b", col = 2 , lwd = 3, pch = 1, main="posterior2(red)")
> lines(v.p, po, type = "b", col = 4 , lwd = 3, pch = 1)
> lines(v.p, lik2/sum(lik2), type = "b", lwd = 3, pch = 1)
> v.p[po2==max(po2)]
[1] 0.27
```



The most likely bass proportion now is 0.27, and the posterior still gets narrower

relative to the blue prior or old posterior. This second-day finding (1 out of 5 being bass) is from a smaller sample, so is less convincing than the first-day finding (3 out of 10 being bass). That is why 0.27 is closer to 0.3 than 0.2. Despite that the second sample is smaller, it still enlarges total sample ($5+10 = 15 > 10$) and helps narrow the posterior. ***Bayesian statistics automatically accounts for uncertainty associated with sampling. That uncertainty is indicated by the narrowness of posterior distribution. A narrow posterior is preciser than a wide posterior. There is no need to compute standard error if Bayesian approach is used. More information can narrow posterior.***.

23. The next but not the last advantage of Bayesian approach is that there is no need to mathematically derive the sampling distribution. ***The posterior distribution is the sampling distribution. It is mechanically obtained by applying the Bayes rule. It can be mechanically updated once new information arrives.***

24. The only disadvantage of Bayesian statistics I am aware of is that it is hard to teach and learn, because of heavy math and confusing concepts. The good news is, in many cases coding can replace math.

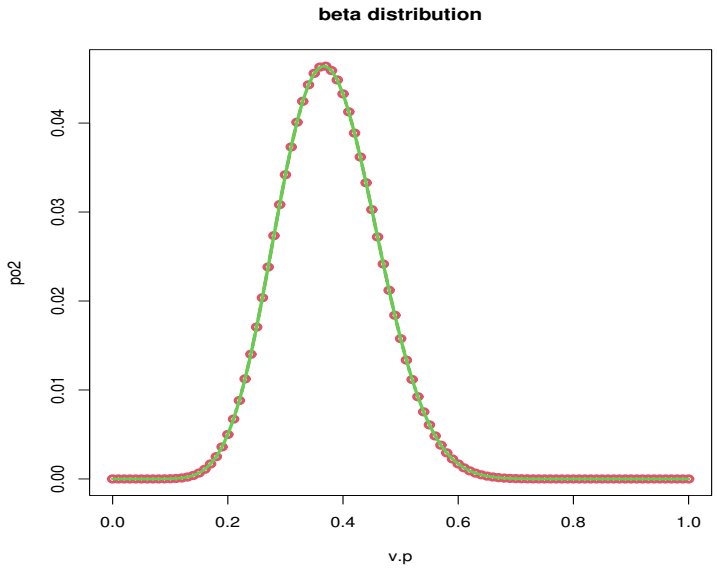
25. For instance, we can use math to show the distribution of most recent estimate of p is a beta distribution:

$$kp^{3+8}(1-p)^{7+12} = beta(12, 20), \quad (0 < p < 1)$$

where k is a constant. Please google "beta distribution" to learn more. Do not confuse beta distribution with binomial distribution—the former is about p for given data, while the latter is concerned with the number of bass for given p.

26. We use following R codes to compare the second-round posterior distribution (red color) and beta(12,20) distribution (green color).

```
plot(v.p, po2,  type = "b", col = 2 , lwd = 3, pch = 1)
lines(v.p, dbeta(v.p, 12, 20)/sum(dbeta(v.p, 12, 20)), type = "l", col = 3 , lwd =
```



As expected, they are the same!

27. Bayesian statistics is quickly gaining popularity in the era of powerful computer and software. For instance, it is used to find a missing aircraft—Google "Bayes, find wreckage". It is also used to design self-driving car—Google "self-driving car, bayes". ***Basically, Bayesian statistics tells us how to digest information efficiently***.

28. To summarize, you need to figure out prior, likelihood, and posterior distribution in order to conduct Bayesian analysis. The confidence interval is readily available from the posterior distribution. Keep applying Bayes rule to update the posterior after new information arrives. More information helps!