A gentle introduction to Bayesian data analysis

Steve Simon, P.Mean Consulting

Facebook: www.facebook.com/pmean

LinkedIn: www.linkedin.com/in/pmean

Main website: www.pmean.com

Newsletter: www.pmean.com/news and
 www.facebook.com/group.php?gid=302778306676

Twitter: http://twitter.com/profmean

Webinars: www.pmean.com/webinars

2. Abstract

- P-values and confidence intervals are the fundamental tools used in most inferential data analyses. They are possibly the most commonly reported statistics in the medical literature. Unfortunately, both p-values and confidence intervals are subject to frequent misinterpretations.

3. Abstract

- In this webinar, you will review the interpretation of p-values and confidence intervals, and see an alternative approach based on Bayesian data analysis.

4. Learning objectives

- In this seminar, you will learn how to:
  - interpret p-values and confidence intervals;
  - explain the difference between informative and diffuse priors;
  - interpret statistics from a posterior distribution.

5. Outline

1. Pop quiz
2. Definitions
3. What is a p-value?
4. What is a confidence interval?
5. A simple example of Bayesian data analysis.
6. Repeat of pop quiz

6. Pop quiz

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

7. Pop quiz

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

1. The result is statistically significant and clinically important.
2. The result is not statistically significant, but is clinically important.
3. The result is statistically significant, but not clinically important.
4. The result is not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

8. Pop quiz

A Bayesian data analysis can incorporate subjective opinions through the use of

1. Bayes rule.
2. data shrinkage.
3. a prior distribution.
4. a posterior distribution.
5. p-values.
6. I do not know the answer.

9. Definitions

- Population: the group you wish to generalize your research results to.

- Usually defined in terms of
  - Demography,
  - Geography,
  - Occupation,
  - Time,
  - Care requirements,
  - Diagnosis,
  - Or some combination of the above.

10. Definitions

- An example of a population would be all infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

11. Definitions

- Sample: subset of a population.
- Random sample: every item in the population has the same probability of being in the sample.
- Biased sample: some items in the population have a decreased probability of being in the sample.

12. Definitions

- A Type I error is **rejecting the null hypothesis when the null hypothesis is true** (false positive).

- Example involving drug approval: a Type I error is allowing an ineffective drug onto the market.

13. Definitions

- A Type II error is **accepting the null hypothesis when the null hypothesis is false**.

- An example involving drug approval: a Type II error is keeping an effective drug off of the market.

14. What is a p-value?

- A p-value is a **measure of how much evidence we have against the null hypothesis**. The null hypothesis, traditionally represented by the symbol H0, represents the hypothesis of no change or no effect. The smaller the p-value, the more evidence we have against H0.

15. Definitions

- A Type II error is **accepting the null hypothesis when the null hypothesis is false**.

- An example involving drug approval: a Type II error is keeping an effective drug off of the market.

16. What is a p-value?

- The p-value is also a measure of how likely we are to get a certain sample result or a result "more extreme," assuming H0 is true. The type of hypothesis (right tailed, left tailed or two tailed) will determine what "more extreme" means.

17. **What is a confidence interval?**

- **A confidence interval is a range of values that tries to quantify uncertainty associated with the sampling process**. Consider it as a **range of plausible values**.

- There is a confidence level associated with any confidence interval, usually 95%, but sometimes 90% or 99%.

18. What is a confidence interval?

- The confidence level is related to the alpha level (probability of a Type I error).

- It also has a long range sampling interpretation.

  – If you repeatedly sampled from the same population, then 95% (or 90% or 99%) of the confidence intervals produced would contain the true value in the population.

19. **Bayesian data analysis**

- There's a wonderful example of Bayesian data analysis at work that is simple and fun. It's taken directly from an article by Jim Albert in the Journal of Statistics Education (1995, vol. 3 no. 3) which is available on the web at

    - www.amstat.org/publications/jse/v3n3/albert.html.

20. **Bayesian data analysis**

- I want to use his second example, involving a comparison of ECMO to conventional therapy in the treatment of babies with severe respiratory failure. In this study, 28 of 29 babies assigned to ECMO survived and 6 of 10 babies assigned to conventional therapy survived. Refer to the Albert article for the source of the original data.

21. **Bayesian data analysis**

- Wikipedia gives a nice general introduction to the concept of Bayesian data analysis with the following formula:

  P (H|E) = P(E|H) P(H) / P(E)

    - where H represents a particular hypothesis, and E represents evidence (data). P, of course, stands for probability.

22. **Bayesian data analysis**

- The first step is to specify P(H), which is called the prior probability. Specifying the prior probability is probably the one aspect of Bayesian data analysis that causes the most controversy. The prior probability represents the degree of belief that you have in a particular hypothesis prior to collection of your data.

23. **Bayesian data analysis**

- The prior distribution can incorporate data from previous related studies or it can incorporate subjective impressions of the researcher. What!?! you're saying right now. Aren't statistics supposed to remove the need for subjective opinions?

  Actually, a bit of subjectivity is a good thing.

24. **Bayesian data analysis**

- First, it is impossible to totally remove subjective opinion from a data analysis. You can't do research without adopting some informal rules. These rules may be reasonable, they may be supported to some extent by empirical data, but they are still applied in a largely subjective fashion.

25. **Bayesian data analysis**

Here are some of the subjective beliefs that I use in my work:

1. you should always prefer a simple model to a complex model if both predict the data with the same level of precision.

2. you should be cautious about any subgroup finding that was not pre-specified in the research protocol.

3. if you can find a plausible biological mechanism, that adds credibility to your results.

26. **Bayesian data analysis**

- Second, the use of a range of prior distributions can help resolve controversies involving conflicting beliefs. For example, an important research question is whether a research finding should "close the book" to further research. If data indicates a negative result, and this result is negative even using an optimistic prior probability, then all researchers, even those with the most optimistic hopes for the therapy, should move on.

27. **Bayesian data analysis**

- Third, while Bayesian data analysis allows you to incorporate subjective opinions into your prior probability, it does not require you to incorporate subjectivity. Many Bayesian data analyses use what it called a diffuse or non-informative prior distribution. This is a prior distribution that is neither optimistic nor pessimistic, but spreads the probability more or less evenly across all hypotheses.

28. **Bayesian data analysis**

- Here's a simple example of a diffuse prior that Dr. Albert used for the ECMO versus conventional therapy example. Let's assume that the true survival rate could be either 0, 10%, 20%, ..., 100% in the ECMO group and similarly for the conventional therapy group. This is not an optimal assumption, but it isn't terrible either, and it allows us to see some of the calculations in action.

29. **Bayesian data analysis**

- With 11 probabilities for ECMO and 11 probabilities for conventional therapy, we have 121 possible combinations. How should we arrange those probabilities? One possibility is to assign half of the total probability to combinations where the probabilities are the same for ECMO and conventional therapy and the remaining half to combinations where the probabilities are different. Split each of these probabilities evenly over all the combinations.

30. P (H|E) = P(E|H) **P(H)** / P(E)

- This is P(H). For simplicity, we multiplied each probability by 1000 and rounded.

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	45	5	5	5	5	5	5	5	5	5	5
0.1	5	45	5	5	5	5	5	5	5	5	5
0.2	5	5	45	5	5	5	5	5	5	5	5
0.3	5	5	5	45	5	5	5	5	5	5	5
0.4	5	5	5	5	45	5	5	5	5	5	5
0.5	5	5	5	5	5	45	5	5	5	5	5
0.6	5	5	5	5	5	5	45	5	5	5	5
0.7	5	5	5	5	5	5	5	45	5	5	5
0.8	5	5	5	5	5	5	5	5	45	5	5
0.9	5	5	5	5	5	5	5	5	5	45	5
1	5	5	5	5	5	5	5	5	5	5	45

31. P (H|E) = **P(E|H)** P(H) / P(E)

- The second step in a Bayesian data analysis is to calculate P(E | H).

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	0	0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	1	0
0.3	0	0	0	0	0	0	0	0	0	6	0
0.4	0	0	0	0	0	0	0	0	1	17	0
0.5	0	0	0	0	0	0	0	0	2	31	0
0.6	0	0	0	0	0	0	0	0	3	38	0
0.7	0	0	0	0	0	0	0	0	2	30	0
0.8	0	0	0	0	0	0	0	0	1	13	0
0.9	0	0	0	0	0	0	0	0	0	2	0
1	0	0	0	0	0	0	0	0	0	0	0

32. P (H|E) = **P(E|H)** P(H) / P(E)

- P(E | H) is the probability of the observed data under each hypothesis. There are formulas for this, using the binomial distribution. I used Excel to calculate these probabilities for me. Here's an example of the formula I used.

```
=binomdist(28,29,0.9,FALSE)*binomdist(6,10,0.6,FALSE)
```

33. P (H|E) = **P(E|H) P(H)** / P(E)

- Now multiply the prior probability by the likelihood.

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	0	0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	4	0
0.3	0	0	0	0	0	0	0	0	2	25	0
0.4	0	0	0	0	0	0	0	0	6	77	0
0.5	0	0	0	0	0	0	0	0	10	141	0
0.6	0	0	0	0	0	0	0	1	13	173	0
0.7	0	0	0	0	0	0	0	4	10	138	0
0.8	0	0	0	0	0	0	0	0	45	67	0
0.9	0	0	0	0	0	0	0	0	1	77	0
1	0	0	0	0	0	0	0	0	0	0	0

34. P (H|E) = P(E|H) P(H) **/ P(E)**

- These numbers do not add up to 1, so we need to rescale them.

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	0	0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	5	0
0.3	0	0	0	0	0	0	0	0	2	32	0
0.4	0	0	0	0	0	0	0	0	7	97	0
0.5	0	0	0	0	0	0	0	0	13	178	0
0.6	0	0	0	0	0	0	0	1	16	218	0
0.7	0	0	0	0	0	0	0	5	13	174	0
0.8	0	0	0	0	0	0	0	0	57	84	0
0.9	0	0	0	0	0	0	0	0	1	97	0
1	0	0	0	0	0	0	0	0	0	0	0

35. **Bayesian data analysis**

- How likely are we to believe the hypothesis that ECMO and conventional therapy have the same survival rates? Just add the cells along the diagonal (0+0+...+5+57+97+0) to get 159 out of a thousand. Prior to collecting the data, we placed the probability that the two rates were equal at 500 out of a thousand, so the data has greatly (but not completely) dissuaded us from this belief.

36. **Bayesian data analysis**

You can calculate the probability that

- ECMO is exactly 10% better than conventional therapy
  - 0+0+...+1+13+84+0 = 98 /1000,
- ECMO is exactly 20% better
  - 0+0+...+13+218+0 = 231 / 1000,
- exactly 30% better
  - 0+0+...+7+178+0 = 185 /1000,
- and so forth.

37. Bayesian data analysis

- Here's something fun that Dr. Albert didn't show. You could take each of the cells in the table, compute a ratio of survival rates and then calculate the median of these ratios as 1.5.

Relative risk	Prob.	Cumul.	
0.8 / 0.9 = 0.89	1	1	
0.9 / 0.9 = 1.00	97	98	
0.8 / 0.8 = 1.00	57	155	
0.7 / 0.7 = 1.00	5	160	
0.9 / 0.8 = 1.12	84	244	
0.8 / 0.7 = 1.14	13	257	
0.7 / 0.6 = 1.17	1	258	
0.9 / 0.7 = 1.29	174	432	
0.8 / 0.6 = 1.33	16	448	
0.9 / 0.6 = 1.50	218	666	<< Median
0.8 / 0.5 = 1.60	13	679	
0.9 / 0.5 = 1.80	178	857	
0.8 / 0.4 = 2.00	7	864	
0.9 / 0.4 = 2.25	97	961	
0.8 / 0.3 = 2.67	2	963	
0.9 / 0.3 = 3.00	32	995	
0.9 / 0.2 = 4.50	5	1000	

38. **Bayesian data analysis**

- Dr. Albert goes on to show an informative prior distribution. There is a fair amount of data to indicate that the survival rate for the conventional therapy is somewhere between 10% and 30%, but little or no data about the survival rates under ECMO.

39. **Bayesian data analysis**

- Here's a prior distribution that utilizes this historic information.

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	Total
0	0	0	0	0	0	0	0	0	0	0	0	0
0.1	0	50	50	48	39	34	25	17	11	6	0	280
0.2	0	54	54	51	42	36	27	18	12	6	0	300
0.3	0	38	38	36	29	25	19	13	8	4	0	210
0.4	0	22	22	20	17	14	11	7	5	2	0	120
0.5	0	11	11	10	8	7	5	4	2	1	0	60
0.6	0	5	5	5	4	4	3	2	1	1	0	30
0.7	0	0	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
Total	0	180	180	170	140	120	90	60	40	20	0	

40. Repeat of pop quiz

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

41. Repeat of pop quiz

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

1. The result is statistically significant and clinically important.
2. The result is not statistically significant, but is clinically important.
3. The result is statistically significant, but not clinically important.
4. The result is not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

42. Repeat of pop quiz

A Bayesian data analysis can incorporate subjective opinions through the use of

1. Bayes rule.
2. data shrinkage.
3. a prior distribution.
4. a posterior distribution.
5. p-values.
6. I do not know the answer.