Lecture 4

January 15, 2019 10:25 AM

STATISTICS 475: Applied Discrete Data Analysis

Inference on Two Binomial Probabilities

(B&L Section 1.2, Appendix B.4)

1 Problem to be solved

- We now know about the binomial model for binary responses
  - PMF $f(w|\pi) = [n!/w!(n-w)!]\pi^w(1-\pi)^{(n-w)}$
  - Probability of success parameter π
  - Maximum likelihood estimate $\hat{\pi}$
  - Tests for $H_0 : \pi = \pi_0$
  - Confidence intervals for π
- We often have two separate binomial counts → 2 populations
  - Just like two sample means problem
  - Want to compare the probabilities that generated the two counts
  - What parameter are we interested in?
    - For comparing means, we used differences of means
    - For probabilities: Difference? Something else?
  - How do we test whether the probabilities are the same?
  - Can we find a confidence interval?

2 Notation and model

EXAMPLE: Salk Polio Vaccine[1] Clinical trials are performed to determine the safety and efficacy of new drugs. Frequently, the safety and efficacy responses are categorical in

[1] Description stolen straight from Bilder and Loughin (2014)

Table 1: Salk vaccine clinical trial results; see the book for source details.

	Polio	Polio free	Total
Vaccine	57	200,688	200,745
Placebo	142	201,087	201,229
Total	199	401,775	401,974

nature; for example, the efficacy response may be simply whether a drug cures or does not cure a patient of a disease. In order to ensure that a new drug is indeed better than doing nothing (patients sometimes get better without intervention), it is essential to have a control group in the trial. This is achieved in clinical trials by randomizing patients into two groups: new drug or control. The control group is often a placebo, which is administered just like the new drug but contains no medication.

One of the most famous and largest clinical trials ever performed was in 1954. Over 1.8 million children participated in the clinical trial to determine the effectiveness of the polio vaccine developed by Jonas Salk. While the actual design of the trial sparked [ethical] debate, we forgo this discussion and focus on the *data* obtained from the randomized, placebo-controlled portion of the trial. The data, given in Table 1, show that 57 out of the 200,745 children in the vaccine group developed polio during the study period, as opposed to 142 out of the 201,229 children in the placebo group. The question of interest for the clinical trial was "Does the vaccine help to prevent polio?" We will develop comparison measures in this section to answer this question.

This example is exactly what we are talking about:

- Two separate binomial responses (counts of "successes" in n independent, identical Bernoulli trials)
  - *Are* the trials independent here?
    - The actual trial went into different elementary schools and sought volunteers.
    - "Clustering" effects?
  - *Are* probabilities of polio equal in all children?
- Interest in a comparison of the probability of success

The notation is as follows:

- Label the groups "1" and "2."
  - It doesn't matter which is which: all methods are INVARIANT to this arbitrary labeling and give the same answer either way.
  - Use all the same notation as in the single binomial model, except index symbols by group number
    - Sample sizes n_1 and n_2

    - Probabilities of success π_1 and π_2
    - In group j (=1 or 2) observe n_j binary (0/1) responses $y_{j1}, y_{j2}, \ldots, y_{jn_j}$
    - Random variables W_1 and W_2 represent the populations of potential counts of successes
    - Observed counts of successes are w_1 and w_2

- We assume that in each group, the conditions for a binomial distribution have been met
    - Should be confirmed
- Furthermore, for now we assume that the *groups are independent*
    - No observation of a success or failure in one group is related to any particular observations(s) of success or failure in another group
    - Would be violated if data come from married couples, siblings, or from measurements at two times on same units, for example
- Then we can assume an INDEPENDENT BINOMIAL MODEL for the two RVs
    - $W_1 \sim Bin(n_1, \pi_1)$ and $W_2 \sim Bin(n_2, \pi_2)$
    - W_1 and W_2 are independent.

n trials, π probability of success

↳ Assume 2 separate binomial distributions.

3 Estimation

- The model has two parameters: the probabilities of success in each group, π_1 and π_2
- We can use ML to estimate these
    - Use $j = 1, 2$ to index the two groups
    - We know that the PMF for W_j is

$$f(w_j|\pi_j) = P(W_j = w_j) = \frac{n_j!}{w_j!(n_j - w_j)!}\pi_j^{w_j}(1-\pi_j)^{n_j - w_j},\ w_j = 0, 1, \ldots, n_j,\ j = 1, 2$$

    - It turns out that when RVs are independent, then the likelihood function for any parameter estimated using responses from these RVs is just the product of their respective likelihood functions
        - i.e., $L(\pi_1, \pi_2|w_1, w_2) = L(\pi_1|w_1) \times L(\pi_2|w_2) = f(w_1|\pi_1) \times f(w_2|\pi_2)$
    - It can then be shown that the MLEs are the "obvious" results: the two respective sample proportions
        - $\hat{\pi}_1 = w_1/n_1$ and $\hat{\pi}_2 = w_2/n_2$

4 Approach #1: Inference on $\pi_1 - \pi_2$

- One way to compare 2 probabilities is through their difference
  - Just like comparing two means
- So we have a new parameter, $\pi_1 - \pi_2$
  - Need to estimate this parameter: let's use ML!
  - INVARIANCE PROPERTY OF MLEs: The MLE for any function of parameters is the same function applied to the MLEs of the parameters (See App B.4)
    - So the MLE of $\pi_1 - \pi_2$ is $\hat{\pi}_1 - \hat{\pi}_2$
  - $\hat{\pi}_1 - \hat{\pi}_2$ has the properties of an MLE:
    - *Asymptotic normality*
    - Asymptotic accuracy ("consistency")
    - Asymptotic optimality
    - A variance estimate that can be computed automatically
- The sampling distribution of $\hat{\pi}_1 - \hat{\pi}_2$ is approximately normal with mean $\pi_1 - \pi_2$ and variance

$$Var(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

adding 2 binomial variances together.
↳ because 2 groups are independent

4.1 Hypothesis Tests for $\pi_1 - \pi_2$

- Most often want to test $H_0 : \pi_1 = \pi_2$, so use $H_0 : \pi_1 - \pi_2 = 0$
  - Same three alternatives as usual: $\neq, >, <$
  - In very rare circumstances, might ask $H_0 : \pi_1 - \pi_2 = c$ for some c other than 0.
- **Wald Test**: Make a Z statistic with the standard error estimated using the MLEs:

$$Z_W = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$$

If $H_0 : \pi_1 - \pi_2 = c$ → Z_W $\frac{\hat{\pi}_1 - \hat{\pi}_2 - c}{\text{----}}$

  - Uses MLEs for parameters in SE, ignoring H_0
  - Usual standard normal rejection regions based on $Z_{1-\alpha/2}, Z_{1-\alpha}$, or Z_α
  - Not a great test in small samples, but easy
- **Score test**: Make a Z statistic using the standard error estimated assuming that H_0 is true
  - i.e. assuming that $\pi_1 = \pi_2$

* If this is true, then we have $n = n_1 + n_2$ trials from a *single* binomial with probability $\pi (= \pi_1 = \pi_2)$
* Have $w = w_1 + w_2$ successes
* MLE for π is $\hat{\pi} = w/n = (w_1 + w_2)/(n_1 + n_2)$

– Then $\widehat{Var}(\hat{\pi}_1 - \hat{\pi}_2) = \hat{\pi}(1 - \hat{\pi})(1/n_1 + 1/n_2)$

– And the test statistic is

$$Z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})(1/n_1 + 1/n_2)}}$$

* Again, compare to a standard normal distribution

– Preferred to Wald test in small samples, not really any more difficult

- **Likelihood Ratio Test**: Has a test statistic that is similar to the LR test stat for one sample (see the book, p. 35)

– Compare test stat, $-2 \log \Lambda$ to χ^2_1 → 1 df. because treating 2 parameters as 1 $[\pi_1 - \pi_2]$

* 1 df because H_0 implies one constraint on the two probabilities (they must be equal)
* Another way to look at it is that you have to estimate two probabilities, $\hat{\pi}_1$ and $\hat{\pi}_2$, under H_a but only one $\hat{\pi}$ under H_0, for a difference of $2 - 1 = 1$

– Again, better than Wald, not as good as Score in small samples.

EXAMPLE: Salk Polio Vaccine: Hypothesis tests (Lecture 4 scripts.R)

The question of interest for the clinical trial was "Does the vaccine help to prevent polio?" Recall the data:

	Polio	Polio free	Total
Vaccine	57	200,688	200,745
Placebo	142	201,087	201,229
Total	199	401,775	401,974

For convenience, instead of groups "1" and "2", let's use "V" for vaccine group and "P" for placebo group, and again, success = Polio. Then

$H_a: \pi_V - \pi_P < 0$, cuz we care about whether the vaccine ↓ or ↓ prob. of polio

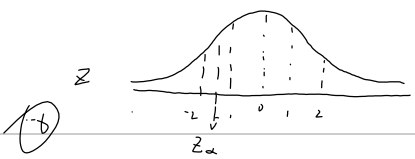
$n_V =$ 200,745 $n_P =$ 201,229, $w_V =$ 57, $w_P =$ 142, $H_0:$ $\pi_V - \pi_P = 0$
$H_a:$ $\pi_V - \pi_P < 0$

Also, $\hat{\pi}_V =$ 0.00028 and $\hat{\pi}_P =$ 0.00071

- Wald test uses $Z_w = (0.00028 - .00071)/\sqrt{.00028(.99972)/200745 + .00071(.99929)/201229} = -6.0$

-6 is a huge test stat.

$Z_w < Z_\alpha$



(-6)

  - p-value = 0.0000000009

- Score test uses $\bar{\pi} = 0.00050$, which leads to
  $Z_0 = (0.00028 - .00071)/\sqrt{.00050(.99950)/(1/200745 + 1/201087} = -6.0$

  - p-value = 0.0000000009
  - Can also get the score test from `prop.test()`, but it comes out squared as a 2 sided test.

- LR Test uses calculations shown in the program. The test stat is $-2\log(\Lambda) = 37.3$

  - p-value = 0.000000001 *for a 2-sided version of the test*, =0.0000000005 for the 1-sided version

In all three cases, we clearly reject H_0 using $\alpha = 0.05$. We conclude that the probability of polio *is lower* for children given vaccine versus placebo.

---

5 Confidence Intervals for $\pi_1 - \pi_2$

Again with this problem, there are several methods of making confidence intervals.

- **Wald Interval**: The simplest confidence interval for π—and the one you probably learned in your first STAT course:

  $$\hat{\pi}_1 - \hat{\pi}_2 \pm Z_{1-\alpha/2}\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2}$$

  where $Z_{1-\alpha/2}$ is the standard normal reference value corresponding to cumulative probability $1 - \alpha/2$

  - Easy to calculate
  - Bad interval if either n_j is small or either π_j is close to 0 or 1
    - Tends to be too narrow, coverage is below $100(1-\alpha)$

- **Agresti-Caffo Interval**: Just compute the Wald interval, but add 1 to *all four* counts:

  $$\tilde{\pi}_1 - \tilde{\pi}_2 \pm Z_{1-\alpha/2}\sqrt{\tilde{\pi}_1(1-\tilde{\pi}_1)/n_1 + \tilde{\pi}_2(1-\tilde{\pi}_2)/n_2},$$

  where $\tilde{\pi}_j = (w_j + 1)/(n_j + 2)$.

  - Very similar to Wald in larger samples, but much better in small samples (e.g. when any expected count is <5)

- **Score Interval**: Found by inverting score test: identifying the values d for which $H_0 : \pi_1 - \pi_2 = d$ would not be rejected at the α level.

  - Computationally complicated; no simple formula but available in special functions in R.

EXAMPLE: Salk Polio Vaccine: Hypothesis tests (Lecture 4 scripts.R)

Recall the data:

	Polio	Polio free	Total
Vaccine	57	200,688	200,745
Placebo	142	201,087	201,229
Total	199	401,775	401,974

We have $\hat{\pi}_V = 0.000284$ and $\hat{\pi}_P = 0.000706$

- Wald interval

  $$(.000284 - .000706) \pm 1.96\sqrt{.000284(.999716)/200745 + .000706(.999294)/201229}$$
  $$= (-0.00056, -0.00028)$$

- Agresti-Caffo interval: $\tilde{\pi}_1 = 58/200747 = 0.000289,\ \tilde{\pi}_2 = 143/201231 = 0.000711$

  $$(.000289 - .000711) \pm 1.96\sqrt{.000289(.999711)/200745 + .000711(.999289)/201229}$$

- Agresti-Caffo interval: $\tilde{\pi}_1 = 58/200747 = 0.000289,\ \tilde{\pi}_2 = 143/201231 = 0.000711$

$$(.000289 - .000711) \pm 1.96\sqrt{.000289(.999711)/200745 + .000711(.999289)/201229}$$

$$= (-0.00056, -0.00028)$$

- Score interval from `diffscoreci()` function of the `PropCIs` package gives

$$-0.00056 < \pi_1 - \pi_2 < -0.00029$$

In all three intervals are very similar because of the large sample size. Clearly, the probability of polio is smaller in vaccinated children who receive vaccine than in those who receive placebo.

---

6 Notes

1. There is actually a 4th test that is popular in 2×2 contingency tables, the Pearson chi-squared test of independence. It turns out that
   (a) The null hypothesis of "independence" in the contingency table is identical to our null hypothesis of equal probabilities of success
   (b) The Pearson test statistic is algebraically identical to the Score test!
   (c) However, Pearson test can do only 2-sided alternatives.
      i. On the other hand, it can extend to larger tables where our Z_0 statistic does not.
      ii. We will learn more about this later

the chance of polio is independent of vaccine injection.

2. The tests and confidence intervals are very similar in this example, but that's because we had large samples and both successes and failures were plenty in both groups.

3. The size of the difference in probabilities in the example is very small! Should we get excited about something that reduces a probability of a disease by only 0.0004?

   (a) Notice that if the vaccine were 100% effective—which we would all say is a *great* thing—the estimated difference in probabilities would have been only 0.00071...again *really small*

4. This actually points to a flaw in using differences of probabilities: very small probabilities have little room for change, but reducing risks to much smaller probabilities is a *huge* thing!

   (a) What can we do about this?

7 Conclusions: What to learn from this

1. Comparisons of probabilities between 2 samples can be based on difference parameter, $\pi_1 - \pi_2$

2. MLE is $\hat{\pi}_1 - \hat{\pi}_2$, has usual MLE properties

3. Score test is easy to compute

4. Various confidence intervals available

   (a) Score is still best but no longer a formula

   (b) Agresti-Caffo (add 1 to counts) is a hack that works OK.

8 Exercises (due on date to be announced)

Compete the following exercises from B&L, Chapter 1:

1. Exercise 17 (a), (b):
   (a) In (a), also find the Wald, Agresti-Caffo, *and Score* confidence intervals.
   (b) In (b) compute the LRT test rather than the Pearson test, which is the same as the Score test, after all.
2. Exercise 22 (This is an open-ended problem. It asks a question without telling you how to answer it. Part of the problem is figuring out what to do to answer the question. I *love* these questions!)

In addition, here are exercises that will not be marked, but that you can do for practice:

1. Exercise 25: Stat students should be able to show prove that the Pearson test statistic is the square of the Score test statistic.
2. **Why did the Salk trial involve so many children?** We can answer that by understanding the nature of these tests and confidence intervals when the event being studied is relatively rare.

   In the 1940s, the rate of occurrence of polio was about is about .05 percent of children.

   (a) Suppose for a moment that .0005 is the true probability of polio in a randomly selected child.
      i. If you do a trial with 100 children in each group, how many polio cases would you expect in the control group?
      ii. How many children need to be in the trial in order to have an expectation that about 5 children in the control group would contract polio? (Note that expected counts of 5 are what is often recommended to ensure that the sample is large enough for the normal and chi-squared approximations to hold reasonably well).
   (b) Suppose further that the vaccine is 50% effective; that is, it reduces the probability of polio by about 50%.
      i. If you ran 20000 children in each group, then what would be the expected counts of polio cases in each group? (These will be whole numbers)
      ii. Suppose that they had run this trial and that exactly as many children as expected had contracted polio in the two groups. Conduct a score test for the difference in probability and draw conclusions.
      iii. If a 50% reduction in probability were considered useful, would it be wise to use 20000 children per group? Explain.