lecture 3

January 8, 2019 10:35 AM

STATISTICS 475: Applied Discrete Data Analysis

Inference on the Binomial Probability

(B&L Section 1.1.2, Appendix B.5)

1 Problem to be solved

- We have developed the ML estimate, $\hat{\pi} = w/n$ (Successes/Trials) for the probability-of-success parameter, π, in the binomial distribution.
- We now need to use it to do inference on the parameter.
- This means:
  - Perform a hypothesis test of the null hypothesis that π is equal to some particular value, say π_0.
  - Find a $100(1-\alpha)\%$ confidence interval for π
- I will also use this lecture to establish some very common techniques that we will see over and over: Wald, Score, and Likelihood Ratio (LR methods)

2 Quick review: Tests

Recall that a hypothesis test for a parameter θ consists of the following elements:

1. **Null hypothesis**: A special value for the parameter that we often wish to *disprove*. Denote this by θ_0. Then we write $H_0 : \theta = \theta_0$.
2. **Alternative (research) hypothesis**: The deviation from the null hypothesis that represents "interesting" values of θ. Depending on the context of the problem, might have
   (a) $H_a : \theta \neq \theta_0$ (2-sided, most common case)
      i. Answers the question: is θ different from θ_0?
      ii. Alternatively, *could* θ be equal to θ_0?

(b) $H_a : \theta > \theta_0$(upper tail)

i. We are interested in finding out whether the true value of the parameter is *greater* than the proposed value

(c) $H_a : \theta < \theta_0$ (lower tail)

i. We are interested in finding out whether the true value of the parameter is *less* than the proposed value

3. **Test Statistic**: A computed value comparing how far the observed value $\hat{\theta}$ is from θ. $\to \theta_0$

- In cases where $\hat{\theta}$ has a normal distribution, the form of the statistic is typically

$$Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$$

4. **Rejection region**: The values of the test statistic that would lead one to reject H_0 in favour of H_a, assuming a type I error rate of α

- In cases where $\hat{\theta}$ has a normal distribution, then Z has a standard normal distribution, $N(0, 1)$.
- Then the rejection region is based on quantiles of the standard normal, Z_q, such that $P(Z < Z_q) = q$
- In the three cases above, we reject H_0 if

(a) $|Z| > Z_{1-\alpha/2}$
(b) $Z > Z_{1-\alpha}$
(c) $Z < Z_\alpha$

- p-values can be used to represent the test statistic's position relative to the rejection region
  - p-value = probability of a test statistic at least as extreme *relative to* H_a as the one observed

5. **Decision**: A statement to to reject or not reject H_0 based on the comparison of the test stat to the rejection region.

- p-value can be compared to α instead

6. **Conclusion**: A statement *in plain language* about the implications of the test results

- *Must be written in the context of the problem*
- *May not contain symbols*
- *Cannot conclude that H_0 is true!*

3 Hypothesis tests for π

Likelihood theory is funny: there are many ways to use it to develop inference. Some details on the methods are given in Appendix B.5. Here, we just present and use the different methods, and explain when one might want to use each.

Recall the example:

EXAMPLE: Sex of newborns in Canada: Hypothesis test

Are more girls being born than boys? 1000 babies are randomly sampled from among those born in Canada in the 2000s. Their sex is recorded:

524 Females 476 Males

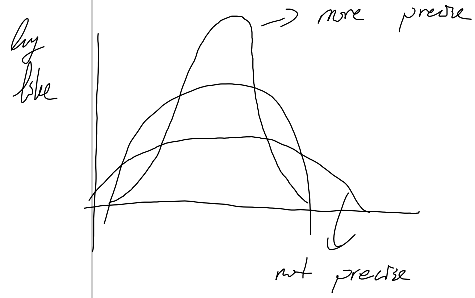
Define Success=Female, so $\pi = P(\text{Female})$.

Because the initial question asks whether there are *more* girls being born than boys, the thing we want to prove is whether $\pi > 0.5$. Thus, $\pi_0 = 0.5$ and we have $H_0 : \pi = 0.5$ and $H_a : \pi > 0.5$, a 1-sided upper-tail test. From the data, we have $\hat{\pi} = 524/1000 = 0.524$. We will assume $\alpha = 0.05$ throughout.

- Three basic approaches to testing in ML: Wald, Score, and Likelihood Ratio
  - Wald uses properties of the likelihood only at the MLE $\hat{\theta}$ (often simplest, but worst)
  - Score uses properties of the likelihood only at the null-hypothesis value of the parameter, θ_0
  - Likelihood ratio compares the likelihoods at θ_0 and $\hat{\theta}$
- **Wald test for** $H_0 : \pi = \pi_0$
  - The MLE $\hat{\pi}$ is approximately normally distributed if the sample size if "large enough".
  - Therefore a test stat can be formed as

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}$$

    - * Standard error is calculated using *estimated value* of the parameter, $\hat{\pi}$
  - Rejection region is based on $Z_{1-\alpha/2}$ (2-sided), Z_α (lower tail), or $Z_{1-\alpha}$ (upper tail)
  - This is not a great test, in the sense that it tends to reject H_0 more often than it should
    - * Normal distribution is not perfect



⇒ More curvyness *

→ narrower interval

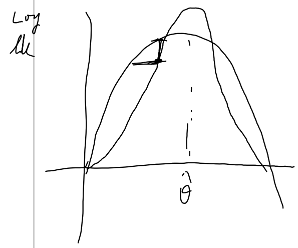
→ more precise to $\hat{\theta}$

* When the true π is close to 0 or 1, the standard error in the denominator is unstable (changes a lot with a small change in w), unless n is very large
* Some recommendations suggest not to use this unless both w and $n-w$ (counts of successes and failures) are at least 5.

- **Score test for** $H_0 {:} \pi = \pi_0$
  - Since we know what π should be when H_0 is true, use it in $SE(\hat{\pi})$
  - Test stat is

$$Z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

    * Standard error is estimated using *hypothesized value* of the parameter, π_0
  - Rejection region is based on $Z_{1-\alpha/2}$ (2-sided), Z_α (lower tail), or $Z_{1-\alpha}$ (upper tail)
  - Uses the asymptotic normality of the MLE
  - This is the test I recommend. It is a better test than the Wald.
    * Standard error less prone to big errors.
    * Still needs large sample (e.g., $n\pi_0$ and $n(1-\pi_0) > 5$), but often performs OK for slightly smaller sample sizes and is almost always better than Wald.



- **Likelihood Ratio (LR) test for** $H_0 {:} \pi = \pi_0$
  - LR test is looks at the log-likelihood curve and and compares the height at H_0 to the maximum height. See Figure 1.
    * If the difference in heights is "small," then the H_0 value is almost as good as the MLE as a model for the data.
    * If the difference in heights is "large," then the H_0 value is not nearly as good as the MLE
    * Remarkably, "small" and "large" differences are judged against the chi-squared distribution if the sample is "large enough".
  - In general, LR Test stat is $-2\log(\Lambda)$, where

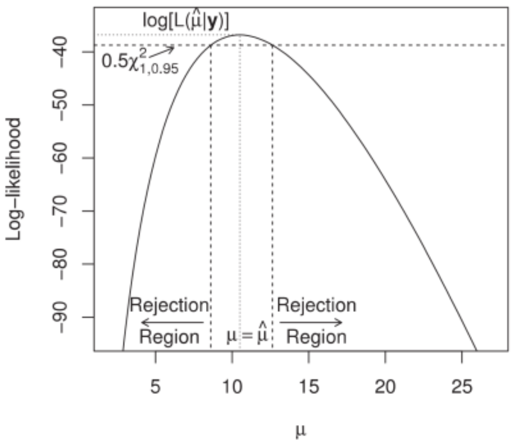
$$\Lambda = \frac{\text{Maximum of likelihood function under } H_0}{\text{Overall maximum of likelihood function}}$$

    * $= 2$[Maximized log likelihood -Best log likelihood under H_0]
    * For the probability of success from a binomial distribution, the test stat works out to be

$$-2\log(\Lambda) = -2\left\{ w\log\left(\frac{\pi_0}{\hat{\pi}}\right) + (n-w)\log\left(\frac{1-\pi_0}{1-\hat{\pi}}\right)\right\}.$$

    (If $w = 0$ or n, then corresponding term in test stat is 0, but the other term still counts.)

Figure 1: Depiction of Likelihood Ratio Test



- Rejection region is based on asymptotic χ^2_ν, where ν represents the number of constraints on the parameters implied by H_0
  * ν is called the "degrees of freedom" → 1 parameter π, constrained to a specific value.
  * $\nu = 1$ here, because H_0 specifies restricts the value of only one parameter, $H_0 : \pi = \pi_0$
  * Reject H_0 if $-2\log(\Lambda) > \chi^2_{\nu,1-\alpha}$
- Technically for $H_a : \pi \neq \pi_0$ only, but for tests of a single parameter
  * Can compute 1-sided p-value as [p-value/2] if relationship between $\hat{\pi}$ agrees with the given alternative, or [1−p-value/2] if it doesn't.
- This test is also usually better than Wald, but not necessarily as good as Score.

Score > LR > Wald.

EXAMPLE: Sex of newborns in Canada: Hypothesis tests (Lecture 3 scripts.R)

We have $H_0 : \pi = 0.5$ and $H_a : \pi > 0.5$, $w = 524$, $n = 1000$, $\hat{\pi} = 524/1000 = 0.524$, $\alpha = 0.05$

- Wald test uses $Z = (0.524 - 0.5)/\sqrt{.524(.476)/1000} = 1.520$
  - p-value = 0.064
- Score test uses $Z = (0.524 - 0.5)/\sqrt{.5(.5)/1000} = 1.518$
  - p-value = 0.065
- LR Test uses $-2\log(\Lambda) = -2\left\{(524)\log\left(\frac{0.5}{0.524}\right) + (476)\log\left(\frac{0.5}{0.476}\right)\right\} = 2.304$
  - p-value = 0.13 *for a 2-sided version of the test*, =0.064 for the 1-sided version

In all three cases, we fail to reject H_0 using $\alpha = 0.05$. We conclude that there is insufficient evidence that more girls were being born than boys in Canada in the 2000's.

The scripts show how to do these computations manually in R. It also shows the `prop.test()` function for computing the Score test.

---

4 Quick review: Confidence Intervals

Recall: a $100(1-\alpha)\%$ confidence interval (CI) for a parameter θ

- Two *statistics*, L and U, such that the probability that the interval between the statistics covers the true parameter θ is $1 - \alpha$
  - $P(L < \theta < U) = 1 - \alpha$
  - $(1-\alpha)$ (or $100(1-\alpha)$) is called the (NOMINAL, STATED) CONFIDENCE LEVEL of the interval

* Typically 95%, so that $\alpha = .05$
* 90% ($\alpha = .10$) and 99% ($\alpha = .01$) are also common
* In reality, the interval may or may not cover the true parameter value with probability *exactly* $(1 - \alpha)$
    - CI's are based on sampling distributions
    - Sampling distributions are often approximate (e.g., MLEs are *approximately* normally distributed)
* The TRUE CONFIDENCE LEVEL (or COVERAGE) of a CI is the fraction of time the process *actually* covers the parameter
    - Usually needs to be determined by simulations.

- Simple formula for CIs based on statistics that have normal distributions:

$$\hat{\theta} \pm Z_{1-\alpha/2} SE(\hat{\theta})$$

- Interpretation of a CI is a little complicated, because the interval is random and the parameter is fixed
    - Fully correct: "We would expect $100(1-\alpha)\%$ of all similarly constructed intervals to contain the parameter θ"
    - Approximately correct (and easier to use): "We are $100(1 - \alpha)\%$ confident that the interval between L and U covers the parameter θ." or "We are $100(1 - \alpha)\%$ confident that the true parameter value is is covered by (is between) L and U."
    - Mostly incorrect: "The probability that the parameter θ falls between L and U is $100(1 - \alpha)$."
        * Implies that the parameter is random

5 Confidence Intervals for π

There are many methods for forming CIs for π, based on making different assumptions about what is important or using different aspects of the likelihood. I present just a few of them here.

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

$$\hat{\pi} \pm Z_{1-\alpha/2} \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$$

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

    - Derived directly from the asymptotic normality of the MLE $\hat{\pi}$
    - Easy to calculate

- Crappy interval if either n is small or π is close to 0 or 1
  - Tends to be too narrow, coverage is below $100(1-\alpha)\%$
  - Can even give endpoints beyond 0 or 1!!!
- Notice that the Wald CI is just the "inversion" of the 2-sided Wald test
  - What values of π_0 would *not* be rejected by the Wald test?

- **Wilson (Score) Interval**: Found by inverting the 2-sided score test, to find the values of π_0 that would not be rejected by this test
  - More complicated due to the use of π_0 in the standard error
  - Resulting formula is

$$\tilde{\pi} \pm \frac{Z_{1-\alpha/2}\sqrt{n}}{n + Z^2_{1-\alpha/2}} \sqrt{\hat{\pi}(1-\hat{\pi}) + \frac{Z^2_{1-\alpha/2}}{4n}},$$

new version on canvas.

where

$$\tilde{\pi} = \frac{w + Z^2_{1-\alpha/2}/2}{n + Z^2_{1-\alpha/2}}$$

n + z² tricks

  - Not something that is as friendly to hand calculation, but simple to program
  - Generally performs very well (coverage not far from nominal level) unless the true probability π is *really* close to 1, where it may be too short.
  - Interval does remain between 0 and 1
  - My preferred interval

- **Agresti-Coull Interval**: Just compute the Wald interval, but add $Z^2_{1-\alpha/2}/2$ to both the number of successes and the number of failures:

$$\tilde{\pi} \pm Z_{1-\alpha/2}\sqrt{\tilde{\pi}(1-\tilde{\pi})/n}$$

where

$$\tilde{\pi} = \frac{w + (Z^2_{1-\alpha/2}/2)}{n + (Z^2_{1-\alpha/2})}$$

Basically the $\tilde{\pi}$ on score interval.

  - Approximates Wilson interval using an easy hand-calculation
    - In particular, when $\alpha = 0.05$, $Z_{1-\alpha/2} = 1.96 \approx 2$, so use $\tilde{\pi} = (w+2)/(n+4)$ (add 2 successes and 2 failures).
  - $\tilde{\pi}$ is shifted a bit toward 0.5 compared to $\hat{\pi}$
  - Performs similarly to Wilson, but tends to be a little too long when $n\pi < 5$ or $n(1-\pi) < 5$.
    - My preferred interval for hand calculation
    - However, can fall outside of 0–1 range

- **Likelihood Ratio Interval**: Like the Wilson score interval, except based on inverting the LR test
  - Formula is complicated, must be solved using iterative numerical procedures
  - A decent interval, not worth the effort here, but in more complicated problems will be the best we can do
    - Always stays within 0–1 range
- **Clopper-Pearson Interval**: A confidence interval with true coverage probability that is *guaranteed* never to fall below the nominal level
  - Uses the binomial distribution of W directly, no asymptotic approximation
  - Complicated formula again—need tables of the beta distribution—easily done by computer
  - Interval remains within 0–1 range
  - May be wastefully wide, usually has coverage somewhat more than needed
  - Useful in a regulatory environment, where you must be *certain* that your inferences make no fewer errors than claimed.

→ eg 95% C.I
guarantee to be at least 95% confident.
Usually this C.I. is wider than necessary.

EXAMPLE: Sex of newborns in Canada: Confidence intervals (Lecture 3 scripts.R)

We will make 95% confidence intervals intervals for the true probability of a female baby. Recall that $n = 1000$, $\hat{\pi} = 0.524$, and $Z_{0.975} = 1.96$.

1. The Wald interval is

$$0.524 \pm 1.96\sqrt{0.524(0.476)/1000} = 0.524 \pm 0.031 = (0.493, 0.555).$$

   We are 95% confident that the true probability of a female baby in Canada in the 2000's is between 0.493 and 0.555. Note that this interval contains 0.5, which represents equal male-female probabilities.

2. The Agresti-Coull interval starts with $\tilde{\pi} = (w+2)/(n+4) = 526/1002 = 0.5239$. The interval works out to be

$$0.5239 \pm 1.96\sqrt{0.5239(0.4761)/1000} = 0.5239 \pm 0.031 = (0.493, 0.555).$$

   Practically the same as Wald because n and w are already pretty large.

These and other intervals are available from the `binom.confint()` in the `binom` package:

```
> library(package = binom)
> binom.confint(x = w, n = n, conf.level = 1-alpha,
                methods = "all")
          method   x    n      mean     lower     upper
1  agresti-coull 524 1000 0.524000 0.4930131 0.5548032
2     asymptotic 524 1000 0.524000 0.4930460 0.5549540
3          bayes 524 1000 0.523976 0.4930460 0.5548792
4        cloglog 524 1000 0.524000 0.4925693 0.5544324
5          exact 524 1000 0.524000 0.4925140 0.5553444
6          logit 524 1000 0.524000 0.4929934 0.5548227
7         probit 524 1000 0.524000 0.4930047 0.5548507
8        profile 524 1000 0.524000 0.4930143 0.5548629
9            lrt 524 1000 0.524000 0.4930221 0.5548537
10     prop.test 524 1000 0.524000 0.4925137 0.5552996
11        wilson 524 1000 0.524000 0.4930133 0.5548030
```

= wald

= clop pearson

The methods include Wald (#2, "`asymptotic`"), Wilson (#11), Agresti-Coull (#1), LR (#9, "`lrt`") and Clopper-Pearson (#5, "`exact`"). Each one returns an interval $0.493 < \pi < 0.555$. We are 95% confident that this interval has covered the true proportion of female babies in Canada in the 2000s.

---

6 Notes

1. In general, a confidence interval tells you more about a parameter than a test does
   (a) If a CI is based on an inverted test, then it contains a hypothesized parameter value if and only if the same test at the same error level would not reject that value.
   (b) Merely rejecting a null hypothesis tells you nothing about what values of a parameter might alternatively be plausible. A CI tells you this and gives you a form of hypothesis test.
2. These CIs were all very similar in this example, but that's because we had large samples and both successes and failures were common.
3. *Think about how many digits you need to present!!!*
   (a) The standard error of a statistic tells you which digits it believes are well known and which ones are not.
   (b) One suggestion is to present digits down to 1/3 of the SE.

7 Conclusions: What to learn from this

1. Remember general construction of tests and CIs.
   (a) CIs often found by inverting tests: Which H_0 values would *not* be rejected?
2. Three methods—Wald, Score, and LR—will appear again and again with the same basic properties
   (a) All are accurate in large samples (hundreds of successes and failures)
   (b) In smaller samples, usually Score is better than LR, which is better than Wald
      i. "Better" mean retains rejection/coverage rates closer to nominal level α
   (c) For this simple, one-parameter problem, additional CIs are available
      i. Wilson Score is probably best, but Agresti-Coull is good and easy.

8 Exercises (due on date to be announced)

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

1. For the data collection described in Exercise 1d:
   (a) Discuss the 5 conditions for using a binomial distributional model for this problem. Is each one obviously satisfied, obviously not satisfied, or possibly satisfied under certain assumptions (and what are the assumptions?)
   (b) Presuming the assumptions in part (a) are satisfied, compute the Wald, Agresti-Coull, Wilson, and Clopper-Pearson intervals to estimate the probability that a car passing through the intersection uses alternative fuel.
   (c) Suppose that, nationwide, 8% of cars use alternative fuels. Do the cars using this intersection during this time appear to have a similar propbability of alternative fuel use? Explain.
2. Exercise 9
3. Exercise 12: Leave $\alpha = 0.05$, but use $n = 10$ and 1000. These represent "small" and "large" samples. Comment on
   (a) how the 4 intervals compare to one another at each sample size
   (b) how all 4 intervals' coverage patterns change across as sample sizes are increased or decreased.

Wherever a confidence interval interpretation is requested, pretend that someone has come to you with this problem and is asking you to tell them what the results mean. *Explain it in a sentence in the context of the problem.*

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

1. Exercise 1(a-c): Practice identifying whether it would be appropriate to use a binomial distribution as a model for these examples. This is important to learn, because inappropriately applying binomial inference methods leads to poor results.
2. Exercise 10, for those who like solving algebra problems.