Lecture 5

January 17, 2019 10:12 AM

STATISTICS 475: Applied Discrete Data Analysis

Alternative Comparisons of Two Probabilities

(B&L Section 1.2)

1 Problem to be solved

- The interpretation of a difference in probability depends on the magnitude of the two probabilities
  - With moderate or large probabilities, a small difference in probability, $\pi_1 - \pi_2$, is practically meaningless
  - With small probabilities (rare events) *all* differences are small in magnitude, even when they are extremely important
- For example
  - If a vaccine decreases your chance of a disease from .510 to .501, the difference in probability is 0.009, but the utility of the vaccine is negligible
  - If a different shot reduces your chance of a disease from 0.010 to 0.001, then this difference is also 0.009
- We need alternative comparisons to use, instead of differences, that provide more meaningful comparisons for small probabilities

2 Relative Risk

- The RELATIVE RISK is defined as the ratio of two probabilities, $RR = \pi_1/\pi_2$ (assuming $\pi_2 \neq 0$)
- Measures the multiplicative change in probability of success
  - Probability of success is often called "risk" when "success" is some rare, bad thing

  - $0 \leq RR < \infty$
  - Equal probabilities means $RR = 1$

- In the examples above, the relative risks would be
  - $0.510/0.501 = 1.02$, so the risk of disease without the shot is 1.02 times as high (or .02 times higher, or 2% higher) than with it.
    - Could write this the other way: $0.501/0.510 = 0.98$, so disease risk is reduced 2% by the vaccine
  - $0.010/0.001 = 10.0$, so the risk of disease without the shot is 10 times as high (or 9 times higher or 900% higher) than with it.
    - Could write this as $0.001/0.010 = 0.1$, so disease risk is decreased 90% by the vaccine

- Estimate using invariance property MLEs: $\widehat{RR} = \hat{\pi}_1/\hat{\pi}_2$, again assuming $\hat{\pi}_2 \neq 0$

2.1 Inference using relative risk

- It turns out that the distribution of $\log(\widehat{RR})$ has nicer math/stat properties than the distribution of $\widehat{RR}$ (all logs are natural logs)

- Variance of $\log(\widehat{RR})$ is estimated by

$$\widehat{Var}(\log(\widehat{RR})) = \frac{1}{w_1} - \frac{1}{n_1} + \frac{1}{w_2} - \frac{1}{n_2}$$

  assuming w_1 and w_2 both > 0

- Wald confidence interval is the one that is mathematically most feasible:

$$\log \frac{\hat{\pi}_1}{\hat{\pi}_2} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{w_1} - \frac{1}{n_1} + \frac{1}{w_2} - \frac{1}{n_2}}$$

  - This leads to two numbers, L and U, such that $P(L < \log(RR) < U) = 1 - \alpha$
  - To get $100(1-\alpha)\%$ confidence interval for RR, simply exponentiate the interval: $e^L < RR < e^U$.

- Note that when either $w_1 = 0$ or $w_2 = 0$, this cannot be calculated.
  - Common to add .5 to each w_j and 1 to each n_j in that case.

- Of course, a score interval would be better, but it is again complicated to find.
  - Need to find all r for which $H_0 : RR = r$ would not be rejected by a score test.
  - The `riskscoreci()` function of the `PropCIs` can find it.

EXAMPLE: Salk Polio Vaccine: Relative Risk (Lecture 5 scripts.R) Recall that $w_V = 57, n_V = 200745, w_P = 142, n_P = 201229$, leading to $\hat{\pi}_V = 0.00028$ and $\hat{\pi}_P = 0.00071$. Use a confidence interval on relative risk to estimate the percentage reduction in chance of polio with the vaccine compared to without the vaccine. That is, we want to estimate π_V/π_P.

- $\widehat{RR} = 0.00028/0.00071 = 0.40$, so the estimated risk of polio is .4 times as high (or 40% as high or 60% lower) with the vaccine than without it.
- 95% confidence interval is based on

$$\log 0.40 \pm Z_{1-\alpha/2}\sqrt{\frac{1}{57} - \frac{1}{200745} + \frac{1}{142} - \frac{1}{201229}} = (-1.22, -0.60)$$

  - $e^{-1.22} = 0.30$, $e^{-0.60} = 0.55$, so with 95% confidence we believe that the true relative risk is between 0.30 and 0.55.
  - This implies an estimated 60% reduction in risk of polio, with confidence interval 45%–70% reduction.
- Score interval is shown in the program: $0.30 < RR < 0.55$

---

3 Odds Ratios

- Relative risks are *very* useful for comparing small probabilities.
- Not as useful for large probabilities, because probabilities are limited above by 1
  - If $\pi_1 = 0.75$, then the minimum possible RR is $0.75/1 = 0.75$
  - if $\pi_2 = 0.75$, then the maximum possible RR is $1/0.75 = 1.33$
- Also, the mathematical properties of $\widehat{RR}$ and $\log(\widehat{RR})$ are not particularly excellent.
  - Sampling distributions have asymmetry that persists unless samples are moderately large
- The ODDS RATIO (OR) is an alternative comparison that has some better mathematical properties
  (HUGELY important! Learn this bullet!)
  - ODDS OF SUCCESS is defined as $P(\text{Success})/P(\text{Failure}) = \pi/(1-\pi)$
  - Odds ratio is the ratio of the odds of success in group 1 to the odds of success in group 2:

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

- $0 < OR < \infty$ as long as there is at least one success and one failure in each group
- If $\pi_1 = \pi_2$ then $OR = 1$
- Slightly complicated interpretation: "The odds of <success> in <group 1> are <OR> times as high as they are in <group 2>"
  - *Get good at this!*
- As with RR, should choose which group makes best sense in numerator
- Note that OR is also the ratio of [relative risk of success] to the [relative risk of failure]:

$$OR = \frac{\pi_1/\pi_2}{(1-\pi_1)/(1-\pi_2)}$$

  - If success is relatively rare, then $OR \approx RR$

Inference using odds ratio

- Again MLE is found by plugging in MLEs for π_1 and π_2.
- Result simplifies to

$$\widehat{OR} = \frac{w_1(n_2 - w_2)}{w_2(n_1 - w_1)}$$

- It turns out again that the distribution of $\log(\widehat{OR})$ has nicer properties that the distribution of $\widehat{OR}$
- Variance of $\log(\widehat{OR})$ is estimated by

$$\widehat{Var}(\log(\widehat{OR})) = \frac{1}{w_1} + \frac{1}{n_1 - w_1} + \frac{1}{w_2} + \frac{1}{n_2 - w_2}$$

  assuming w_1 and w_2 both > 0
- Wald confidence interval is the one that is mathematically most feasible:

$$\log \widehat{OR} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{w_1} + \frac{1}{n_1 - w_1} + \frac{1}{w_2} + \frac{1}{n_2 - w_2}}$$

  - This leads to two numbers, L and U, such that $P(L < \log(OR) < U) = 1 - \alpha$
  - To get $100(1-\alpha)\%$ confidence interval for OR, simply exponentiate the interval: $e^L < OR < e^U$.
- Of course, a score interval would be better, but it is again complicated to find.
  - Need to find all r for which $H_0 : OR = r$ would not be rejected by a score test.
  - The `orscoreci()` function of the `PropCIs` can find it.
- Note that when any of the counts of successes or failures is 0, this cannot be calculated.
  - Common to add .5 to each w_j and each $n_j - w_j$ in that case.

EXAMPLE: Salk Polio Vaccine: Odds Ratio (Lecture 5 scripts.R) Recall that $w_V = 57, n_V = 200745, w_P = 142, n_P = 201229$, leading to $\hat{\pi}_V = 0.00028$ and $\hat{\pi}_P = 0.00071$. Use a confidence interval on the odds ratio to estimate the change in odds of polio with the vaccine compared to without the vaccine. That is, we want to estimate

$$OR = \frac{\pi_V/(1-\pi_v)}{\pi_P/(1-\pi_P)}.$$

- We find that
$$\widehat{OR} = \frac{57 * (201129 - 142)}{142 * (200475 - 57)} = 0.40,$$
so the estimated odds of polio are .4 times as high (or 40% as high or 60% lower) with the vaccine than without it.

- 95% confidence interval is based on
$$\log 0.40 \pm Z_{1-\alpha/2}\sqrt{\frac{1}{57} + \frac{1}{200745-57} + \frac{1}{142} + \frac{1}{201229-142}} = (-1.22, -0.60)$$
  - $e^{-1.22} = 0.30$, $e^{-0.60} = 0.55$, so with 95% confidence we believe that the true odds ratio is between 0.30 and 0.55.
  - This implies an estimated 60% reduction in odds of polio, with confidence interval 45%–70% reduction.

- The score confidence interval given in the program gives essentially the same endpoints
  - This is because the sample size is large enough that all groups are well represented with counts (all counts are much bigger than 5).

- Note that the results of this analysis with OR are essentially the same as those with RR
  - This is because both probabilities of failure are nearly 1, so that their ratio is nearly 1 and $OR \approx RR$

---

4 Notes

1. Relative risk is often a very interpretable quantity and would be the main parameter of interest, except that it is mathematically less convenient to work with than the odds ratio

(a) The important term in the binomial PMF, $\pi^w(1-\pi)^{n-w}$, factors into a combined function of the odds and the observed data,

$$(1-\pi)^n \left(\frac{\pi}{1-\pi}\right)^w,$$

which has useful theoretical implications.

(b) Therefore, people historically used OR as a surrogate or approximation to RR.

(c) Less necessary to do that when we have computers that can do tricky calculations.

2. On the other hand, OR turns out to occur very naturally as a parameter in logistic regression (Chapter 2), whereas it is difficult to construct RR.

   (a) Therefore OR is *hugely* important in categorical data analysis

3. Interpreting OR can be a pain when the definitions of "success" and the groups are not simple as they were in this example.

   (a) The book uses an example of Larry Bird's free throw history where

      i. Group 1 is "the first free throw attempt out of two is made"
      ii. Group 2 is "the first free throw attempt out of two is missed"
      iii. A trial is "the second free throw attempt"
      iv. Success is "Making the second free throw"

   Then if the odds ratio is, say 1.5, it means that "the odds that the second free throw is made given that the first one is made are 1.5 times as high as the odds that the second free throw is made given that the first one is missed." That's a lot to try to take in.

5 Exercises (due on date to be announced)

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

1. Exercise 17 (c)–(e). In (c) and (d), compute both the Wald and score confidence intervals.
2. Exercise 18 (c)–(d). Use Score only.

In addition, you can do Exercises 26-28 for practice. These will not be marked.