STATISTICS 305/605: Intro to Biostatistics

1

September 22, 2021

**Lecture 9: Measures of Association in 2x2 Tables**

**(Reading: PB Section 6.5)**

**1 Goals of Lecture**

• When we look at data in a 2 × 2 table, we often want to quantify the relationship between the variables

• In the past we have compared diﬀerences between probabilities for one variable at diﬀerent levels of the other variable

• We will look now at *ratios* between probabilities and a related quantity, odds.

– It turns out that ratios measure associations between variables better than diﬀerI ences.

• We can use these measures to quantify risks.

**2 Association in 2**×**2 tables**

• Measures of association measure the strength of a *relationship* between two variables

• We consider a situation where there are two groups on which a binary variable Y is measured

– The “groups” can be represented by another binary variable, say X

– Separate probabilities of success, p 1 and p 2 for each group

∗ These are the same as *conditional probabilities that* Y = 1 *given* X = 1 *or* X = 2*, respectively!*

– “Association” in this context means that these conditional probabilities of success are diﬀerent for the two groups.

2

∗ We want to compare p 1 and p 2

– *For example, is hospitalization with Covid associated with vaccine status?*

∗ *Vaccine status (*X*) forms several groups, but we can focus on fully vaccinated vs. non vaccinated, and measure* Y *=hospitalization status on members of each group*

∗ *Is conditional probability of hospitalization the same for each vaccine group?*

• We have previously compared two probabilities by taking diﬀerences, p 1 − p 2

• Not ideal: 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, p 1 − p 2, is practically meaningless

– With small probabilities (rare events) *all* diﬀerences are small in magnitude, even when they are extremely important

– *If Covid vaccines decrease your chance of getting Covid on an airplane with an infected passenger from .510 to .501, the* diﬀer*ence in probability is 0.009, but the utility of the vaccine is negligible*

– *If, instead, the shot reduces your chance from 0.010 to 0.001, then this* diﬀer*ence is also 0.009, but represents a 90% reduction*

• We need alternative comparisons to use, instead of diﬀerences, that provide more meaningful comparisons for small probabilities

**3 Relative Risk**

• The relative risk is deﬁned as the ratio of two probabilities, RR = p 1/p 2 (assuming )

– Measures the multiplicative change in probability of success of numerator group, *relative* to probability of success in denominator group

– Probability of success is often called “risk” when “success” is some bad thing

– 0≤ RR < ∞

– Equal probabilities means RR = 1

• In the examples above, the relative risks would be

– 0.510/0.501 = 1.02, so the risk of Covid on an airplane without vaccine would be 1.02 times as high (or .02 times higher, or 2% higher) than with it.

Table 1: Covid Hospitalizations vs. Vaccine Status in BC for Aug 7-Sept 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |  | |  |
|  |  |  |
| Vax Status | Unvax | 511 | 1,399,489 | 1,400,000 |
|  |  |  |  |
|  | Total | 588 | 4,899,412 | 4,900,000 |

Yes

No

Total

Vax

77

3,499,923

3,500,000

Hospitalization

∗ Could write this the other way: 0.501/0.510 = 0.98, so Covid risk would be reduced 2% with vaccine, relative to without.

– 0.010/0.001 = 10.0, so the risk of Covid on an airplane without the shot would be 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 would be decreased 90% by the vaccine

• We estimate relative risk using the sample proportions,= pˆ 1/pˆ2, again assuming 

• We can ﬁnd a conﬁdence interval for RR, but details are a little complicated

– Sampling distribution oftends to be very skew, but normal approximation for log() works pretty well

– Make CI in log scale and exponentiate results, like we did for mean of PSA in Lecture 3

– Best CI has complicated formula (see STAT 475) but can be computed in R

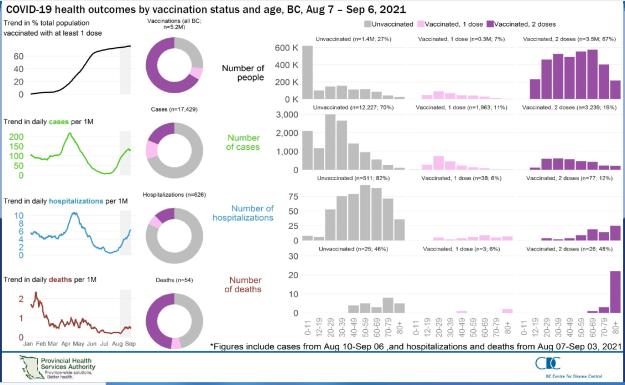
**EXAMPLE: Covid Hospitalization vs. Vaccination Status in BC (Lecture 9 scripts.R)** We once again turn to weekly data on Covid in BC, as posted on the BC CDC Website. This time we take the weekly data summary from September 9 (the most   right show total population size, cases, hospitalizations, and deaths from the past month, separated by vaccination status. The data we want are in the rows for total population and hospitalization, and in the columns for Unvaccinated and Vaccinated, 2 doses. (We ignore the middle column for now.) The numbers for population size are rounded, but they will 

Perhaps you can see what is going to happen here already, just by looking at the data. There are more hospitalizations in the Unvax group, but there are more people from the rest of the population in the Vax group. So it looks like the Vax group has a reduced risk of hospitalization. But we can quantify this more precisely by calculating the relative risk, starting from the “risk” of hospitalization during the month as the conditional probability of hospitalization, given vaccine status.

I start by entering the data as I have before, then running xtabs() to create a table and prop.table() to compute conditional probabilities for the two hospitalization levels

3

Figure 1: BC health outcomes by vaccination status from September 9 Data Summary



for each vaccine status. Letting Group **v** be the Vax group and Group **n** be Unvax, I then compute the risk diﬀerence, pˆv − pˆn and the relative risk “both ways”: 1 = pˆv/pˆnfrom which we can quantify decreased risk from the vaccine, and2 = pˆn/pˆvfrom which we can see increased risk from being unvaccinated. One is just the reciprocal of the other, so which way we compute it is just a matter of how we want to express the results.

> # Enter the data

> vax . status = rep (x=c(" NoVax ", " Vax "), each =2) > hosp . status = rep (x=c(" Yes "," No "), times =2) > counts = c(511 , 1399489 , 77, 3499923)

>

> alldat = data . frame (vax. status , hosp .status , counts )

> alldat

vax . status hosp . status counts

1 NoVax Yes 511

2 NoVax No 1399489

3 Vax Yes 77

|  |  |  |
| --- | --- | --- |
| 4 > | Vax | No 3499923 |
| > # xtabs () creates a cross - tabulation using formula = | | |
| > # | Left side of formula are the counts | |

|  |  |
| --- | --- |
| > # | Right side are the variables to form the table , in form  4 |

|  |  |  |  |
| --- | --- | --- | --- |
| > # | counts ~ row + column | | |
| > tbl = xtabs ( formula = counts ~ vax . status + hosp . status , data = alldat ) | | | |
| > tbl |  |  |  |
|  | hosp . status | | |
| vax . status | | No | Yes |
|  | NoVax 1399489 | | 511 |
|  | Vax | 3499923 | 77 |
| > | | | |
| > # Prop . table () computes conditional probabilities / proportions | | | |
| > # | using a table as x= and " given " variable as margin = | | |

> ( risks = prop . table (x=tbl , margin =" vax . status ")) hosp . status

|  |  |  |
| --- | --- | --- |
| vax . status | No | Yes |
| NoVax 0.999635 0.000365 | | |
| Vax > | 0.999978 0.000022 | |
|  |  |

> # Risk \* difference \* > risks [2 ,2] - risks [1 ,2] [1] -0.000343

>

> # Relative Risk each way > risks [2 ,2]/ risks [1 ,2] [1] 0.06027397

> risks [1 ,2]/ risks [2 ,2] [1] 16.59091

>

> # Confidence intervals using riskscoreci () from PropCIs package

>

> ### install . packages ( package = PropCIs ) > library ( package = PropCIs )

> # Putting unvax in numerator (row 1 of tbl) > riskscoreci (x1= tbl [1 ,2] , n1= sum ( tbl [1 ,1:2]) ,

|  |  |
| --- | --- |
| + data : | x2= tbl [2 ,2] , n2= sum( tbl [2 ,1:2]) , conf . level =.95) |

95 percent confidence interval : 13.06368 21.07051

> # Putting vax in numerator ( row 2 of tbl )

> riskscoreci (x1= tbl [2 ,2] , n1= sum ( tbl [2 ,1:2]) ,

|  |  |
| --- | --- |
| + data : | x2= tbl [1 ,2] , n2= sum( tbl [1 ,1:2]) , conf . level =.95) |

95 percent confidence interval : 0.04745969 0.07654812

5

From the table of conditional probabilities of hospitalization given vaccine status, we see that the probability of hospitalization in the Aug 7-Sept 6 period was is very, very low as measured across the total population of BC, but we also see that the conditional probability is rather higher for the unvaccinated group. The ﬁrst relative risk is 1 = pˆv/pˆn = 0.060, which tells us that vaccinated people have 0.06 times the probability of hospitalization that unvaccinated people do. We can interpret this to mean that vaccine group had a 100∗ (1− 0.060) = 94% lower risk of hospitalization. Next, we ﬁnd2 = pˆn/pˆv = 16.6, meaning that unvaccinated people are 16.6 times as likely to be hospitalized with Covid as vaccinated people are. Conﬁdence intervals for these two RR’s are (0.047, 0.077) for RR 1 (equivalently, (92.3, 95.3) expressed as percent reduction), and (13.1, 21.2) for RR 2. The tightness of these intervals clearly eliminates the possibility of equal probabilities, RR = 1. Either way, there is substantial evidence of a reduced probability of hospitalization among vaccinated people.



OR =p 1/(1− p 1) p 2/(1− p 2)=

p 1(1− p 2)

p 2(1− p 1)

**4 Odds Ratios**

• Relative risks are *very* useful for comparing small probabilities, like in our example.

• Not as useful for large probabilities, because the range of possible values is very limited

– If p 1 = 0.75, then the minimum possible RR is 0.75/1 = 0.75

– if p 2 = 0.75, then the maximum possible RR is 1/0.75 = 1.33

• Also, the sampling distributions of  and log() have asymmetry that persists unless samples are moderately large, especially when probabilities are small

• The odds ratio (OR) is an alternative comparison that has some better properties for inference

• Odds of success is deﬁned as P(Success)/P(Failure) = p/(1− p)

• Odds ratio is the ratio of the odds of success in group 1 to the odds of success in

group 2:

– 0 < OR < ∞as long as no probabilities are zero

– If p 1 = p 2 then OR = 1

– Interpretation like RR except referring to *odds* instead of conditional probabilities

∗ “The odds of <success> in <group 1> are <OR> times as high as they are in <group 2>”

• As with RR, denominator group is the “baseline” against which the numerator group is compared.

6

• Note that OR is also the ratio of [relative risk of success] to the [relative risk of failure]:

=n 11n 22

n 21n 12

|  |  |
| --- | --- |
| OR = | p 1/p 2  (1− p 1)/(1− p 2) |

– If success is relatively rare, then OR ≈ RR

• Estimated from data using counts from a table, which simpliﬁes to

– Multiply counts for success in Group 1 (n 11) and failures Group 2 (n 22) ∗ i.e., counts from opposite corners of table

– Multiply counts in other opposite corners in denominator • Conﬁdence interval for OR found similar to that for RR

**EXAMPLE: Covid Hospitalization vs. Vaccination Status in BC (Lecture 9 scripts.R)** Continuing with the hospitalization data from Table 1, we now compute odds ratios. Speciﬁcally, we compute two ORs similar to the two RRs we computed before, where the preference of which one to present would depend on whether you want to use vaccinated or unvaccinated status as the baseline. Then1 =the odds of hospitalization for vaccinated people relative to those odds for unvaccinated, and 2 =the odds of hospitalization for unvaccinated people relative to those odds for vaccinated.

Starting with the table of counts, I mimic the  formula using counts as given above. Then I calculate 95% CIs for each OR using the orscoreci() function, which takes the same arguments as the riskscoreci() function we used for RR.

> # Recall data table

> tbl

hosp . status

|  |  |  |  |
| --- | --- | --- | --- |
| vax . status | | No | Yes |
|  | NoVax 1399489 | | 511 |
|  | Vax | 3499923 | 77 |
| > | | | |
| > # Relative Risk each way : | | | |
| > # > # > # > # > # > # | NOTE ORDER OF COUNTS IN TABLE TO GET THE INDEXES RIGHT !  1 is odds of hosp for vax relative to novax (Yes , Vax )(No , Novax ) / (Yes , Novax )(No , Vax)  2 is odds of hosp for novax relative to vax (Yes , Novax )(No , Vax ) / (Yes , Vax )(No , Novax ) | | |
|  |  |  |

> ( OR1 = ( tbl [2 ,2]\* tbl [1 ,1])/( tbl [2 ,1]\* tbl [1 ,2]))

[1] 0.0602533

7

> ( OR2 = ( tbl [1 ,2]\* tbl [2 ,1])/( tbl [2 ,2]\* tbl [1 ,1]))



8

[1] 16.5966 > # ( OR2 = 1/ OR1 )

>

> # Confidence intervals using riskscoreci () from PropCIs package

>

> ### install . packages ( package = PropCIs ) > library ( package = PropCIs )

> # Putting vax in numerator ( row 2 of tbl ) > orscoreci (x1= tbl [2 ,2] , n1= sum ( tbl [2 ,1:2]) ,

|  |  |
| --- | --- |
| + data : | x2= tbl [1 ,2] , n2= sum ( tbl [1 ,1:2]) , conf . level =.95) |

95 percent confidence interval : 0.04743304 0.07653863

> # Putting unvax in numerator (row 1 of tbl)

> orscoreci (x1= tbl [1 ,2] , n1= sum ( tbl [1 ,1:2]) ,

|  |  |
| --- | --- |
| + data : | x2= tbl [2 ,2] , n2= sum( tbl [2 ,1:2]) , conf . level =.95) |

95 percent confidence interval : 13.06530 21.08235

Because the probabilities of “success” (hospitalization) are so low, we ﬁnd thatin each case. The odds of hospitalization are 0.060 times as high with vaccination as without, and the reverse of this is that the odds of hospitalization are 16.6 times as high without vaccination as with. The 95% CIs are (0.047, 0.077) for OR 1 and (13.1, 21.1) for OR 2. Again, we see that the possibility of equal probabilities, OR = 1 is eliminated.

**5 What to learn from this**

1. Relative risks and odds ratios are two popular measures of association between binary RVs

2. RR is directly interpreted as a ratio of probabilities, which is useful. (a) We prefer this measure

3. OR is interpreted as ratio of *odds* which are related to probabilities, but not exactly the same

(a) But this measure may provide more accurate inferences when the four table counts are not all fairly large (more than 10 or so?)

9

4. The two measures are very similar when the “success” they are based on is relatively rare.

Table 2: Salk vaccine clinical trial results; see the book for source details.

Polio Polio free

Total

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

**6 Exercises**

**Use R for all calculations, unless otherwise sp**eciﬁed.

1. 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  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 

(a) Use relative risk to estimate how much higher the probability of polio is in children who are not vaccinated, compared to those with vaccine. Also ﬁnd a conﬁdence interval for this. **Report the R code and output, and write a sentence interpreting the results.**

(b) Repeat this, but arrange the relative risk as the reduction in probability of polio that is caused by the vaccine. **Report the R code and output, and write a sentence interpreting the results.**

(c) Find a 95% conﬁdence interval for the population RR in (b). **Report the code and output, and interpret the results, especially addressing whether there is evidence that the vaccine is** eﬀ**ective in reducing the chance that a child contracts polio.**

(d) I will not ask you to compute the ORs for this problem, but if I did, would you expect them to be similar to the RRs? **Explain in 1 sentence why or why not?**

1 Description stolen straight from Bilder and Loughin (2014)

10