STATISTICS 305/605: Intro to Biostatistics

September 14, 2021

Lecture 7: Conditional Probability

(Reading: PB Section 6.2)

1 Goals of Lecture

- We introduce/review 2-way tables and ideas of probability for *two* categorical random variables
- We explain why conditional probability is vitally important to making sense of tables.

2 A Story

This is a true story, at least according to the person who told it to me.

- Wife of a male statistics instructor is sick
- A blood test reveals high levels of something
- The doctor tells her that, "Women at your age and with this level of test result have a 50% chance of having Disease X," a fairly rare but serious condition
- The doctor recommends an invasive medical procedure, which would definitively determine whether Disease X is present
  - The risk of a "bad outcome," including infections or, rarely, death, from the procedure is about 2%.
- The couple decide to go ahead with the procedure
- In the procedure room, while waiting for the doctor to arrive, they see the woman's file on the desk.
  - The report from the test, which is what the doctor based their recommendation on, says, "*Fifty percent of women with Disease X are in this group.*"

  - **It says nothing about the chances that a woman in that group HAS Disease X.**
  - The chances of Disease X are still very small!
    * "Women in this group" is a relatively small fraction of the population, yet half of all Disease X cases are in them.

- The stat instructor explains to the doctor what the test results really mean, and recommends a statistics course as treatment.

- *Scary realization: how many other women had been subjected to this invasive and potentially slightly risky test, all because this doctor did not understand conditional probability?*

- We will try not to let that happen to you.

3 Conditional Probability

- As I said last time, few real phenomena can be well understood by looking at count summaries on one categorical variable
  - Confounders!

- Summarizing across 2 variables is better, because we can take another potential confounder into account

- Still not great, but capable of offering better understanding than before

Example: Covid-19 Vaccine Effectiveness vs. Delta

A paper published in the *New England Journal of Medicine*[1] reports on a study done to estimate the effectiveness of the mRNA vaccines against the alpha and delta variants of Covid-19 in England. They used a test negative case-control design—which we may learn more about later in the semester—to sample people being tested for Covid-19.

> "In brief, we compared vaccination status in persons with symptomatic Covid-19 with vaccination status in persons who reported symptoms but had a negative test. This approach helps to control for biases related to health-seeking behavior, access to testing, and case ascertainment."

The results from the study are given in Table 1, which is their Table 2 from the paper. Since this table has more things in it than we need at the moment, we will focus on a reduced version of the table containing data only for those fully vaccinated with Pfizer, compared to unvaccinated, and considering only the delta variant. This subset is in our Table 2.

[1] J Lopez Bernal et al. (2021). "Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant" *N Engl J Med*, **385**, 585--594.

Table 1: Test results and vaccination status of study participants

Table 2. Vaccine Effectiveness against the Alpha Variant or *S* Target–Negative Status and the Delta Variant or *S* Target–Positive Status, According to Dose and Vaccine Type.*

Vaccination Status	Test-Negative Status		Alpha Variant or *S* Target–Negative Status			Delta Variant or *S* Target–Positive Status	
	Controls	Cases	Case:Control Ratio	Adjusted Vaccine Effectiveness (95% CI)	Cases	Case:Control Ratio	Adjusted Vaccine Effectiveness (95% CI)
	no.	*no.*		%	*no.*		%
Unvaccinated	96,371	7313	0.076	Reference	4043	0.042	Reference
Any vaccine							
Dose 1	51,470	2226	0.043	48.7 (45.5–51.7)	1493	0.029	30.7 (25.2–35.7)
Dose 2	23,993	143	0.006	87.5 (85.1–89.5)	340	0.014	79.6 (76.7–82.1)
BNT162b2 vaccine							
Dose 1	8,641	450	0.052	47.5 (41.6–52.8)	137	0.016	35.6 (22.7–46.4)
Dose 2	15,749	49	0.003	93.7 (91.6–95.3)	122	0.008	88.0 (85.3–90.1)
ChAdOx1 nCoV-19 vaccine							
Dose 1	42,829	1776	0.041	48.7 (45.2–51.9)	1356	0.032	30.0 (24.3–35.3)
Dose 2	8,244	94	0.011	74.5 (68.4–79.4)	218	0.026	67.0 (61.3–71.8)

* The adjusted analysis of vaccine effectiveness was adjusted for period (calendar week), travel history, race or ethnic group, sex, age, index of multiple deprivation, clinically extremely vulnerable group, region, history of positive test, health or social care worker, and care home residence. CI denotes confidence interval.

Table 2: Test results and vaccination status of study participants

	Test Results		
Vaccine Status	Negative	Positive	Total
Unvaccinated	96,371	4,043	100,414
Pfizer-2-dose	15,749	122	15,871
Total	112,120	4,165	116,285

According to the paper, the data represents **all** tests on performed on symptomatic people meeting a few other conditions between October 26, 2020, and May 16, 2021. So on the one hand, this represents *population-level* data for the study population (symptomatic tested people in England for the time period covered). Alternatively, we could consider the data to be a convenience sample relative to the entire planet, and also temporally with respect to the ongoing questions of vaccine efficacy with the delta variant.

3.1 Definitions

- Consider a situation where there are two different random variables that are measured, say X and Y
  - Each has its own set of possible outcomes
  - One can create events from each variable's outcomes
- Now we can consider occurrences of outcomes for each variable simultaneously
  - We observe an outcome of X **and** an outcome of Y on each subject
  - We therefore have *combinations* of outcomes observed on each subject.
  - We call the combination of an outcome on X and an outcome on Y a JOINT OUTCOME on (X, Y)
  - *For the example above, outcomes to a RV for "Vaccination Status" might be called "Vax" and "No Vax", and the outcomes for the RV "Test Result" are "Positive" and "Negative"*
    - *A joint outcome would a combination of Vaccination Status and Test Result*
    - *There are 4 possible joint outcomes: (Vax, Positive), (Vax, Negative), (No Vax, Positive), and (No Vax, Negative)*
- Think about sampling from the joint RV (X, Y),
  - There are probabilities on each of the joint outcomes
  - These probabilities add to 1
  - This is called the JOINT DISTRIBUTION of (X, Y)
    - Can use symbols p_{rc} to represent $P(X = r, Y = c)$ for any particular joint outcome (r, c), where r just represents an outcome from X and c represents an outcome from Y
    - *For example, we can use $r = 1, 2$ to represent Vax and No Vax respectively, and $c = 1, 2$ to represent Positive and Negative, respectively. Then we can label the four joint probabilities $p_{11}, p_{12}, p_{21}, p_{22}$. (As practice, name what these probabilities represent in terms of the example!)*

- The two variables continue to have their own probability distributions, which ignore the other variable
  - These are called the MARGINAL DISTRIBUTIONS of X and of Y.
    - $P(X = r)$ for each r
    - $P(Y = c)$ for each c
  - *We still have two separate 1-way tables we could consider on Vaccination Status and on Test Result*
- We often want to understand may want to understand how the probabilities of one variable (say Y) differ depending on the level of the other variable (X).
  - Fix X at a certain level, say r, and look at the probabilities for all the joint outcomes of (X, Y) that involve $X = r$
    - $p_{r1}, p_{r2}, \ldots$
  - These probabilities don't add to 1.
    - They add to $P(X = r)$
    - Rescale them to add to 1, by dividing by $P(X = r)$, so that it forms a complete probability distribution
  - This is called the CONDITIONAL PROBABILITY DISTRIBUTION OF Y GIVEN $X = r$, and the individual probabilities in the conditional distribution are called CONDITIONAL PROBABILITIES.
    - We write conditional probabilities using a special symbol, "|", which stands for "given"
      - "The probability that $Y = 1$ given that $X = r$" is written $P(Y = 1|X = r)$
  - *We are interested in the probability distribution of tests separately for subjects who are vaccinated and for those who aren't.*
    - *So, we set $r = 1$ for Vax and look at the probabilities of the two outcomes, (Vax, Positive) and (Vax, Negative): p_{11} and p_{12}.*
    - *These two probabilities add to $P(X = 1)$.*
    - *We rescale them into conditional probabilities by taking $P(Y = 1|X = 1) = p_{11}/P(X = 1)$ and $P(Y = 2|X = 1) = p_{12}/P(X = 1)$*
    - *These are the probabilities of a positive test given Vax and of a negative test given Vax. (As practice, write out the process for finding the probability distribution of test results given No Vax!)*

4 2-way tables

- The easiest way to display and understand joint distributions is in a table where rows represent levels of one variable and columns represent levels of the other

Table 3: Test results and vaccination status of study participants

	Y				
X	$c = 1$	$c = 2$	$\cdots$	$c = C$	Marginal X
$r = 1$	p_{11}	p_{12}	$\cdots$	p_{1C}	$P(X = 1)$
$r = 2$	p_{21}	p_{22}	$\cdots$	p_{2C}	$P(X = 2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$r = R$	p_{R1}	p_{R2}	$\cdots$	p_{RC}	$P(X = R)$
Marginal Y	$P(Y = 1)$	$P(Y = 2)$	$\cdots$	$P(Y = C)$	1

- If there is a direction to the relationship—if one variable might be thought of to influence, cause, or explain the other, but not vice-versa—then we typically put the EXPLANATORY VARIABLE (X) in the rows and the RESPONSE VARIABLE (Y) in the columns
  - If there's not a direction to the relationship, then it doesn't matter which variable is row/column, and we can label the row variable to be our "X"
  - We can use R to represent the levels of X (**R**ows) and C to represent the number of levels of Y (**C**olumns).
    - Then use $r = 1, 2, \ldots, R$ to denote individual rows and $c = 1, 2, \ldots, C$ to denote individual colums
  - *In our example, we expect that vaccination can influence whether test is positive or negative, but not the other way around. So we put vaccination status in $R = 2$ rows and test results in $C = 2$ columns.*
- We can put probabilities into a table that looks a lot like Table 2:
- When we gather data, we get counts for each joint outcome. The table of counts for all combinations of X and Y is called a TWO-WAY TABLE, or a $R \times C$ TABLE
  - Also known as a CROSS-CLASSIFICATION or CROSS-TABULATION of X and Y.
- Totals are added for convenience.
- We still refer to the total sample size as n
  - Replace all of the probability "p_{rc}" symbols with "n_{rc}" symbols to denote counts for joint outcomes

Example: Covid-19 Vaccine Effectiveness vs. Delta (Lecture 7 Scripts.R)

Since this document is about probabilities—which are population quantities—we start by treating these numbers as population totals for the study population as mentioned in the introduction to this example. However, since probabilities are just population versions of proportions, the calculations to turn the counts into probabilities are exactly the same as those we would use *estimate* probabilities with proportions in a sample.

These calculations could easily be programmed manually or punched out on a calculator. But with a larger table that would be a pain. So I show one way to enter data into a data frame with one row for each joint outcome. This means we need we need to repeat the X label C times each and the Y label R times each. I do this using two different forms of the `rep()` function. Then I combine everything into a data frame and create a table from the data frame using `xtabs()`. The `formula=` argument is new to us. We will use this a lot for modeling later. Right now, it takes the form `counts ~ rowvar + columnvar`. We finish with the `prop.table()` function, which divides counts in a table given by `x=` by the marginal totals of the variable given by `margin=` to create conditional probabilities or proportions.

```
> #   The rep() function repeats whatever is given as x=.
> #     times= is the number of times the entire x is repeated
> #     each= is used if each element of x is to be repeated
> #        before moving to the next element.
>
> (vax.status = rep(x=c("NoVax", "Vax"), each=2))
[1] "NoVax" "NoVax" "Vax"   "Vax"
> (test = rep(x=c("Neg","Pos"), times=2))
[1] "Neg" "Pos" "Neg" "Pos"
> counts = c(96371, 4043, 15749, 122)
>
> # Put everything into a data frame and add joint probs
> delta = data.frame(vax.status, test, counts,
+                    jt.prob=counts/sum(counts))
> delta
  vax.status test counts     jt.prob
1      NoVax  Neg  96371 0.828748334
2      NoVax  Pos   4043 0.034768027
3        Vax  Neg  15749 0.135434493
4        Vax  Pos    122 0.001049146
>
> # 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
> #     counts ~ row + column
> tbl = xtabs(formula=counts~vax.status + test, data=delta)
> tbl
          test
vax.status   Neg   Pos
     NoVax 96371  4043
     Vax   15749   122
>
> # Prop.table() computes conditional probabilities/proportions
> #   using a table as x= and "given" variable as margin=
> prop.table(x=tbl, margin="vax.status")
```

```
            test
vax.status          Neg          Pos
     NoVax  0.959736690  0.040263310
     Vax    0.992313024  0.007686976
```

Given that someone has *not* had a vaccine, the conditional probability of a positive test over the time period for a symptomatic person was about 0.04. On the other hand, the probability of a positive test given that the person DID have a vaccine was about 0.008. Clearly, in this population, the vaccines were associated with a substantial ($>$ 5-fold) reduction in the probability that the cause of a person's symptoms was Covid.

---

5 What to learn from this

1. Conditional probabilities in a 2-way table focus on looking at probabilities from one variable, *given* (or conditioned on) a specific value for the other variable
2. They are particularly useful for examining the effect of one variable on the probability of different outcomes for another variable
3. Results can be displayed in a 2-way $R \times C$ table.
4. Same caveats apply as with 1-way tables.
   (a) Conditioning on another variable can give a refined understanding of probabilities
   (b) But confounders can still lurk behind the scenes!

Table 4: Covid Outcomes by age distribution, through Aug 28, 2021

Table 4: Age distribution: COVID-19 cases, hospitalizations, ICU admissions, deaths, and BC population by age group January 15, 2020 (week 3) – August 28, 2021 (week 34) (N= 166,262)[a]

Age group (years)	Cases n (%)	Hospitalizations n (%) [b]	ICU n (%)	Deaths n (%)	General BC population n (%)
<10	9,903 (6)	104 (1)	8 (<1)	2 (<1)	470,017 (9)
10-19	18,235 (11)	78 (<1)	18 (1)	0 (<1)	529,387 (10)
20-29	38,543 (23)	481 (6)	58 (3)	2 (<1)	699,476 (13)
30-39	31,179 (19)	893 (10)	173 (9)	16 (1)	750,054 (14)
40-49	23,903 (14)	981 (11)	226 (11)	31 (2)	648,377 (12)
50-59	19,959 (12)	1,354 (16)	389 (19)	78 (4)	711,930 (14)
60-69	12,862 (8)	1,629 (19)	495 (25)	179 (10)	686,889 (13)
70-79	6,549 (4)	1,602 (18)	456 (23)	386 (21)	454,855 (9)
80-89	3,525 (2)	1,161 (13)	171 (8)	633 (35)	193,351 (4)
90+	1,604 (1)	408 (5)	18 (1)	495 (27)	52,885 (1)
Total	**166,262**	**8,691**	**2,012**	**1,822**	**5,197,221**
Median age [c]	**34**	**62**	**62**	**84**	**41**

a. Among those with available age information only.
b. Data sources: health authority case line lists and a subset of PHSA Provincial COVID19 Monitoring Solution (PCMS) data for children <20 years of age. PCMS data were included as of June 8 2021. Due to this change in data source, additional admissions that occurred since the start of the pandemic are now included in age groups 0-9 and 10-19 years.
c. Median ages calculated are based on health authority case line lists only.

6 Exercises

Use R for all calculations, unless otherwise specified.

1. Recall the data in Table 4 from the Aug 22–28 Situation Report described last time. Notice that, ignoring the last "population" column, this table contains data resembling a cross-tabulation of Covid outcomes (Y) and age groups (X). I say "resembling", because I think that a single person could actually show up multiple times if they are a case who is hospitalized, and ends up dying after a visit to the ICU. But we can look again at hospitalizations and create a variable that classifies all cases as either hospitalized or not. Then we can study the conditional probability that a reported case becomes hospitalized, given their age group.

   (a) The data frame you made for Lecture 6 exercises has Cases and Hospitalizations. Each of these vectors should have $R = 10$ elements in it, corresponding to the number of age groups. We need to create a variable for hospitalization status with $C = 2$ levels ("yes" and "no" or something like that), count the number of non-hospitalized cases, and create a data frame where each row is one joint outcome between age group and hospitalization status. So our final data frame should have 20 rows and 3 columns: age group, hospitalization status, and count.

      i. Make a new age group variable, say `age2` or something, where you repeat the current age group vector twice
      ii. Make a new variable for `hosp.status` containing two levels, where each one is repeated 10 times.
      iii. Make a new variable for non-hospitalized cases by subtracting hospitalizations from cases.
      iv. Create a vector of 20 `counts` by using `c()` to combine the hospitalized cases and non-hospitalized cases
      v. Create a data.frame using `data.frame()`.

vi. Print out your code and the data frame. Check to make sure that everything has lined up properly, so that one row contains one age group, one hospital status, and the *right count*! If there's anything wrong, fix any problems. **Only turn in the final working code and data frame.**

(b) Use this data frame to create a cross-tabulation of hospitalization status (Y) against age group (X).

i. **Print the `xtab()` results.**

ii. **Why did we set X to be age group and Y to be hospitalization status, and not the other way around?**

(c) Compute the conditional probability distribution of hospitalization status given age group for each age group. **Print out the code and table. Note that you can apply `round()` to the results of `prop.table()`. Present the probabilities rounded to 3 decimal places.**

(d) Look at the table. **Report in a sentence or two what the apparent trend seems to be in the conditional probability of hospitalization, given age group.** You don't need to report each number (that's what the tables are for!). Just focus on the pattern and summarize it.