STATISTICS 305/605: Intro to Biostatistics

September 11, 2021

Lecture 6: Rates and Proportions with Categorical Variables

(Reading: PB Section 2.2, 4.1)

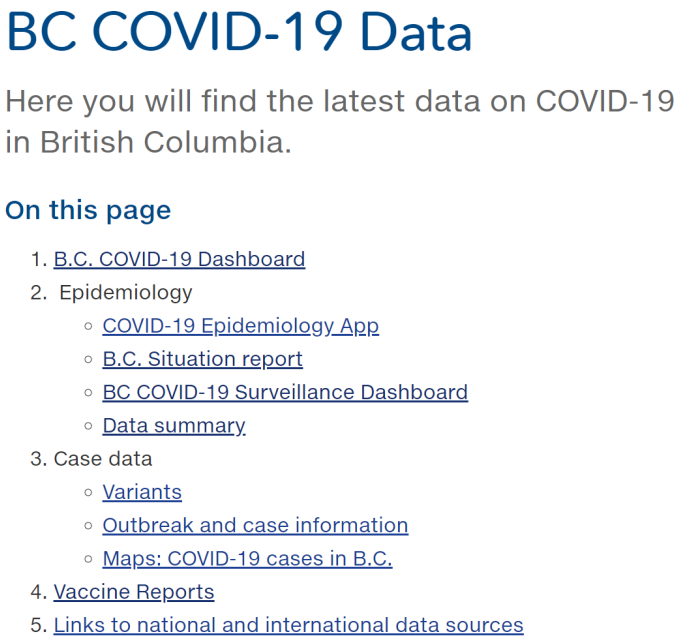
1 Goals of Lecture

- We introduce/review probability distributions and numeric (tabular) summaries for single categorical variables
- We pave the way to learn about understanding summaries and relationships for multiple categorical variables.

2 1-way tables

- VERY often the data we gather or see reported are categorical measurements on individuals
  - Essentially everything we see presented on Covid-19 is categorical
    - Tests, cases, maximum severity
    - Age groups
    - Provinces
    - Countries
- Measurements on *individuals* are categorical with one response per person
- Summaries are represented as *counts* of numbers of individuals in each category.

Figure 1: Screenshot of Resources available on BC CDC Covid Page



Example: BC Covid-19 Weekly Epidemiology Situation Reports

On its public-facing website, the BC Centre for Disease Control posts continual updates of the situation with Covid-19. The website is
http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data.
All screenshots in this lecture were taken from that page or links therein. See Figure 1 for a list of links to resources that are available from that page.

We will look at data from the weekly epidemiology updates known as "BC Situation Reports." In particular, I have posted the report from Week 34, representing data from Aug 22–28, 2021, to this module and will use information from that report in examples for this lecture.

2.1 Definitions

- For a given categorical RV, X, suppose that there are $A \geq 2$ possible outcomes for the variable
  - *For example, there are 5 Health Authorities (HAs) that administer health care within regions of BC: Fraser Health (FH), Interior Health (IH), Vancouver Island Health Authority (VIHA), Northern Health (NH), and Vancouver Coastal Health (VCH). To monitor caseload status across these 5 HAs, we can classify each case according to its HA. Then X might be the BC HA to which a a Covid-19 case is*

Table 1: Counts of Covid-19 cases by regional health authority for Aug 22–28, 2021, with cumulative totals.

Table 1. Episode-based case tallies by health authority, BC, Jan 15, 2020 – August 28, 2021 (week 34) (N= 166,287)

Case tallies by episode date	Health Authority of Residence					Outside Canada	Total
	FH	IH	VIHA	NH	VCH		
Week 34, case counts	1,222	1,490	383	309	666	0	4,070
Cumulative case counts	**90,667**	**21,650**	**6,524**	**8,736**	**38,476**	**234**	**166,287**
Week 34, cases per 100K population	62	177	44	107	54	NA	78
Cumulative cases per 100K population	**4,608**	**2,575**	**746**	**3,019**	**3,142**	**NA**	**3,195**

reported between Aug 22–28. In this case, $A = 5$, since there are 5 HAs to which cases are classified.

- "Outcomes" of a categorical variable are also called, "categories" or "classes" or "groups" or sometimes "labels"
- We can label the individual classes with the index $a = 1, 2, \ldots, A$

- We summarize counts from variable X within each category within a table, often called a 1-WAY TABLE because we are summarizing results from one variable.
  - *These counts are given in Table 1 from the Situation Report.*
  - Since we use n to represent a sample size, let's use n_a to represent the total count in category a for each $a = 1, 2, \ldots, A$
    - *For example, if we let $a = 1, 2, 3, 4, 5$ index the HAs in the order they are given in the table, then we have $n_1 = 1222, n_2 = 1490, \ldots, n_5 = 666$*
  - It doesn't matter at all whether different categories create different columns or rows of the table.
    - Choose what is best for aesthetic (visual) appeal
    - *Table 1 uses columns for categories, probably so it uses less space in the report!*

- Usually counts represent responses observed on a population of units
- In the population, each category has a probability (the proportion of units in the population that belong to that class).
  - We can denote these as p_a for each $a = 1, \ldots, A$
- If the sample is random, then we expect the counts to occur roughly in the same relative amounts as their respective probabilities.

How counts are represented

- When the actual counts are presented in a table, they are also referred to as FREQUENCIES for each category.
- Instead, or in addition, the counts may be presented as proportions relative to the entire sample size

  - In other words, for each category $a = 1, 2, \ldots, A$ we compute the proportion $\hat{p}_a = n_a/n$.
  - These are estimates of their respective population probabilities
  - Proportions are also called RELATIVE FREQUENCIES for each category
  - *These proportions are not shown in Table 1, but we would compute them using the observed counts divided by the total, $n = 4070$.*

- When categories represent populations of different sizes, counts are sometimes scaled relative to an external measure of number of units in these populations.
  - *The table of Covid-19 case counts by regional HA has an additional row where the case counts are scaled according to the number of people living in that region, and expressed on a basis of "cases per 100K of population".*
  - Need to have the population size for category $a = 1, 2, \ldots, A$, say N_a
    * We compute a different proportion relative to this population, say $\hat{P}_a$, indicating what fraction of group a's population has been observed as outcomes
  - We can then choose to represent this in a more interpretable scale
    * It sounds weird to say that "For each person in the population, there are 0.00051 cases of Covid"
    * It is easier to understand if we multiply these proportions by a big number that we can choose to make them 1-3 digits long
      · Multiplying 0.00051 by 100,000, for example, gives is 51, so we have 51 cases per 100,000 people.
  - These scaled counts are called RATES

3 Interpreting Counts and Rates

- The choice of whether to report proportions relative to the total sample size or rates relative to external population depends on what you want table to say
  - Do you want to use counts to demonstrate the probability distribution of the categories across the whole population?
  - Do you have an external measure that you want to use to "standardize" the counts to make comparisons easier across categories?
    * ("standardize" means something different here than it did with Z-scores)
  - Maybe both?
  - *Table 1 reports only the rates relative to each HA's population size. However, in daily briefings, proportions for each HA relative to the sample size (total reported cases in the previous day) are often reported to highlight where in the province the most new cases are located.*

- Reporting the wrong proportions can result in the wrong conclusions
  - *Of the 4070 reported cases, 1222 were in FH and only 309 were in NH (proportions 0.30 and 0.08) respectively, so it looks like the outbreak is "worse" in FH. However, a larger proportion of residents in NH tested positive for Covid-19 (rates 62 and 107, respectively). Which figure should be cited depends on what conclusion you want to draw.*
  - Needless to say, unscrupulous or ignorant people will report whichever statistic suits their narrative better.
    - *Presently, in some jurisdictions there are higher counts of Covid cases among vaccinated people than among unvaccinated people. Does this mean that "vaccines don't work" as many "internet virologists" claim?*
    - *Hint: What are the *rates* of cases among vaccinated and unvaccinated people, respectively?*

Adding "total rates" to tables

- Many tables include a row or column for the total count, and then compute a total rate
- Does this make sense?
  - Yes, if all of there are no real underlying differences in population rates among categories
    - The proportion based on n/N is a better measure of the common proportions in the A groups than any one n_a/N_a
  - Yes, if administrative decisions are based on totals
  - No if the populations have meaningfully different underlying rates and separate actions can, or need to be, taken on each category
    - Province funds all HAs, so budget planning might be based on total rates (how many nurses do we need?)
    - Allocating support to HAs might need to be done based on separate rates for each HA (where do we need the nurses?)

Confounders

- Essentially *no* phenomenon can be completely "explained" by a 1-way table
  - Many factors influence human health
  - All of those factors, except 1, are glossed over by a 1-way table
- The hope is that the 1-way table captures an important pattern simply
- But there is a hidden danger: the variable X that you construct the table from might be related to some other variable(s) that also influence the counts

  - These are called CONFOUNDERS, and they are almost always present
  - *You may see reports from one place or another that claim that Covid hospitalization counts among vaccinated people are higher than among unvaccinated people. However, the age profiles of vaccinated and unvaccinated people are generally very different, and age is a vey important factor in the probability of hospitalization with Covid. (Age is confounded with vaccination status.) Unvaccinated people are mostly young, and young generally have milder cases not requiring hospitalization. Controlling for age groups, it is typically seen that vaccines are, indeed, reducing probability of hospitalization. Having more hospitalizations among older people creates the illusion that vaccines are not working*

- Advice: don't expect ANY 1-way table to give you a complete picture of any complex phenomenon.
  - 2-way tables, which we will learn about soon, are not much better...

Standard Errors and Confidence intervals for rates

- *Obviously* any statistic is more interpretable if there is a measure of uncertainty attached
- But there is a difficulty here.
  - Confidence intervals for probabilities (and hence rates) assume that the counts in the table are subject to sampling variability
  - The parent distribution for the sample must exist in order for the sampling distribution for proportions to create a CI.
  - Does a parent distribution always exist?
    - *The table reports ALL cases that were recorded in each HA. Where was sampling done, and what "parent distribution" were the case counts sampled from? What were the "other people" we could have selected and measured?*
      - *The rates presented ARE the parameters if all reported cases in the population were observed.*
      - *Is it reasonable to think that the 1222 cases in FH represent a sample of the potential case counts we could have gotten if we had (hypothetically) gone back in time and rerun August 22–28? (Thinking about this makes my head hurt!)*
    - *Alternatively, if we imagine that the rates are not changing across time, then there is a parent distribution of possible counts in a given week that could be observed from now until the end of time.*
      - *We could imagine that this week was a sample from this distribution, and that past and future weeks will be randomly different.*
      - *Then we are trying to learn about the true rates in this parent distribution by creating CIs.*

- IF this is justified, a rate is just a proportion relative to the category's population size,
  - Use the same formulas for SEs and CIs as for any probability
    - Rescale probability CI to rates
  - Need to have list of counts and list of population sizes for each category.
  - Recall that I prefer the Wilson CI for a probability.
- We can also compute confidence intervals for the probability distribution of categories relative to the sample
  - Again assuming that these were sampled from a real population somehow.

Example: BC Covid-19 Weekly Epidemiology Situation Reports (Lecture 6 Scripts.R

We will **pretend** that it makes sense here to compute confidence intervals for this problem, so that I can show you the code. I don't think it makes sense, though, because we are not sampling from a population of possible weeks with the same underlying rates. ALL of these rates might change next week.

We will enter the data from Table 1 into R: `HA` for health authorities, `cases` for the week's case counts, `cumu.cases` for the cumulative cases, `rate.cases` for the rates for weekly cases given in the table, and `rate.cucases` for the rate based on cumulative cases. I don't know the HA population sizes, so the first thing I do is reverse–engineer the formula for that turns cases into rates-per-100k to estimate the population sizes. I do this in the cumulative data, because the numbers of cases are larger and will give a more stable estimate. I save these numbers as `pops` and show that the rates calculated from these estimated sizes match those in the table, up to rounding error.

Next, I use the `binon.confint()` function to compute Wilson CIs for each HA's *probability*. I can do this in one line, because I enter 5 case counts as `x=` and matching 5 population sizes as `n=`. I re-express the interval as a rate by multiplying it by 100,000 and print out the results.

```
> # Enter data, maintaining same order as in table
> HA = c("FH", "IH", "VIHA", "NH", "VCH")
> cases = c(1222,1490,383,309,666)
> cumu.cases = c(90667,21650,6524,8736,38476)
> rate.cases = c(62,177,44,107,54)
> rate.cucases = c(4608,2575,746,3019,3142)
>
> # Estimate HA pop sizes from cumulative case counts and rates
> pops = cumu.cases/rate.cucases*100000
>
> # Recalculating proportions and rates as if we didn't know
> #   them, but knew population sizes instead.
> p.hat = cases/pops
```

```
> (rates = p.hat*100000)
[1]  62.10612 177.21709  43.79491 106.78468  54.38642
>
> # Confidence intervals for proportions.
> # The binom.confint() function can take vectors of counts
> #   for both successes (x=) and trials (n=)
> library(binom)
> (all.ci = binom.confint(x=cases, n=pops,
+                         methods="wilson"))
  method    x         n          mean        lower
upper
1 wilson 1222 1967599.8 0.0006210612 0.0005872120 0.0006568605
2 wilson 1490  840776.7 0.0017721709 0.0016845154 0.0018643791
3 wilson  383  874530.8 0.0004379491 0.0003962379 0.0004840490
4 wilson  309  289367.3 0.0010678468 0.0009552872 0.0011936533
5 wilson  666 1224570.3 0.0005438642 0.0005041077 0.0005867543
>
> # Re-express confidence limits as rates,
> lower = 100000*all.ci$lower
> upper = 100000*all.ci$upper
> #  Print all together in one table.  Using data.frame to
> #    combine categorical variable (HA) with numeric.
> #   round() rounds everything in first argument to digits=
> #   cbind() binds columns together into a matrix
> data.frame(HA, round(cbind(rates, lower, upper), digits=0))
    HA rates lower upper
1   FH    62    59    66
2   IH   177   168   186
3 VIHA    44    40    48
4   NH   107    96   119
5  VCH    54    50    59
```

IF these confidence intervals are measuring uncertainty in the location of a true rate of Covid cases for this week, then we see that there are apparently real differences among the HAs' rates for the week. None of the confidence intervals overlap, which (by the way) guarantees that a hypothesis test comparing two HAs' rates would reject the null hypothesis that they are equal[1].

A second analysis that we could do is to find confidence intervals on the probabilities that a randomly selected person with Covid would be from a given HA. Again, we have to pretend that this would be a meaningful thing here, like if these probabilities were relatively stable across time, so that the results from 1 week could be considered a random sample from a process that creates a stable parent distribution. If that were the case, then we could

[1]When the intervals do *not* overlap, it is *not* a guarantee that a hypothesis test would *not* reject the null hypothesis of no difference between groups.

simply enter the vector of observed counts into `binom.confint()` as `x=` and use the sum of cases as a single number for `n=`, since all categories arise from the same total sample size n.

```
> # binom.confint() can compute proportions and CIs for all
> #   categories in a distribution by listing the sum of all
> #   the cases as a single n=
> (all.ci = binom.confint(x=cases, n=sum(cases),
+                         methods="wilson"))
  method    x    n       mean      lower      upper
1 wilson 1222 4070 0.30024570 0.28635751 0.31451061
2 wilson 1490 4070 0.36609337 0.35142615 0.38101312
3 wilson  383 4070 0.09410319 0.08551200 0.10345987
4 wilson  309 4070 0.07592138 0.06817784 0.08446468
5 wilson  666 4070 0.16363636 0.15258898 0.17531810
> # Grab the estimated proportion and CI from columns 4:6.
> #   Round everything to 3 digits.
> data.frame(HA, round(all.ci[4:6], digits=3))
    HA  mean lower upper
1   FH 0.300 0.286 0.315
2   IH 0.366 0.351 0.381
3 VIHA 0.094 0.086 0.103
4   NH 0.076 0.068 0.084
5  VCH 0.164 0.153 0.175
```

We see that a sampled individual has highest probability of being from IH, followed by FH. Se see lowest counts in NH, but is this just because NH is has the fewest people in it? This is exactly why rates are so useful.

---

4 What to learn from this

1. 1-way tables are commonly used to present results from a sample measuring a single categorical variable
2. For each category, proportions relative to the sample can estimate probabilities that a sampled individual would fall into each category
3. Rates can be computed for each category relative to population sizes within each category

4. Confidence intervals for probabilities and rates can/should be presented

5. Don't read too much into 1-way tables, because they hide the potential effects of many confounders.

Table 2: 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.

5 Exercises

Use R for all calculations, unless otherwise specified.

1. Refer to the data in Table 2 (Table 4 from the Aug 22–28 Situation Report). Notice that this table contains counts and the total population sizes for each listed 10-year age group, but there are no rates given. We will fix this oversight for the Hospitalizations. We will not compute confidence intervals, because I can't define the population parameter we would be studying. "Fraction of people in an age groups hospitalized with Covid so far" is not really a repeatable process.

   (a) Enter the data for Age Groups, Cases, Hospitalizations, and Populations. Check these numbers carefully.

      i. **Make a data frame out of these numbers and print out the data frame.**

   (b) Compute the rates per 100K people for each age group.

      i. **Add these as a last column to the previous data frame and print the results.**

   (c) **Plot these rates in a line plot, similar to the monthly deaths plot I made in Lecture 3.** Be sure to get the age group names on the x-axis of the plot, and include sensible labels and a plot title. You don't need a legend, because there's only one line.

   (d) Look at the plot. Report a general description of the pattern in one sentence. If you need to say more than that, you're missing the point. You don't need to report each number (that's what the tables are for!). Just focus on the pattern and summarize it.

2. Now let's address a separate question: What is the probability that a reported case becomes hospitalized in each age group? In this case, I *do* consider what has happened

so far to represent a somewhat random process. If different people had reported cases, the proportions of hospitalizations might well have changed. So we will add confidence intervals.

(a) Compute the proportions of cases that were hospitalized in each age group, along with 95% Wilson CI's for the true probabilities. **Print a data frame consisting of age group, cases, hospitalizations, the proportions, and the lower and upper confidence interval limits.**

(b) Make another line plot by age group. Put the proportions in blue. Add the lower and upper confidence limits in red. Be sure to get the age group names on the x-axis of the plot, and include sensible labels and a plot title. If you can figure out how to add a legend with one blue line for the estimates and one red line for the 95% CI, you can earn a small bonus