MAT1372, Classwork5, Fall2025

2.2 Considering Categorical Data

1. Contingency Table: A table that summarizes data for two categorical variables.

For example, with 3 types of homeownership (mortgage, rent, own) and 2 types of app_type (individual, joint), how many combinations are we have ? $3 \times 2 = 6$

	homeownership	app_type
1	MORTGAGE	individual
2	RENT	individual
3	RENT	joint
4	OWN	individual
⋮	⋮	⋮
10000	MORTGAGE	joint

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Figure 2.17: A contingency table for **app_type** and **homeownership**.

2. Finish the tables based on the information in Figure 2.17.

$\frac{3496}{8505}$

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10,000

app_type	Count
individual	8505
joint	1495
Total	10,000.

3. Row/Column proportions: Sometimes it is useful to understand the fractional breakdown of one variable in another, and we can modify the contingency table to provide such a view.

Row proportion of the table in Fig. 2.17

$\frac{362}{1495}$

	rent	mortgage	own	Total
individual	0.411	0.451	0.138	1.000
joint	0.242	0.635	0.122	1.000
Total	0.386	0.479	0.135	1.000

Column proportion of the table in Fig. 2.17

	rent	mortgage	own	Total
individual	0.906	0.802	0.865	0.851
joint	0.094	0.198	0.135	0.150
Total	1.000	1.000	1.000	1.000

(1) In the table of row proportion, what does 0.411 represent?

Under individual loaner, there are 41.1% who rent

(2) In the table of column proportion, what does 0.906 represent?

Under the renter, there are 90.6% who has individual application loan

4. Here is the result of an experiment study on a new malaria vaccine. All patients were exposed to a malaria parasite strain to test if they got infected.

		outcome		
		infection	no infection	Total
treatment	vaccine	5	9	14
	placebo	6	0	6
	Total	11	9	20

Figure 2.29: Summary results for the malaria vaccine experiment.

(1) The proportion who got infected in the treatment group is $5/14$

(2) The proportion who got infected in the control group is $6/6 = 1$

(3) The difference between the proportion of patients who got infected in the two groups is 64.3%

item(2) − item(1)

$\left(1 - \frac{5}{14} = \frac{9}{14}\right)$

(4) Could we conclude that the vaccine is effective?

Not sure. Since the sample size is very small, and it is unclear whether the difference in (3) provides convincing evidence.

2.3 Case Study: malaria vaccine

1. Independence model (H_0) and Alternative model (H_A).

When the results of a study are unclear, we label these two competing claims, H_0 (H-nought) and H_A (H-A):

H_0: Independence model. The variables `treatment` and `outcome` are independent.

H_A: Alternative model. The variables are not independent.

In the experiment study on the malaria vaccine, what are the H_0 and H_A in this study?

H_0: The treatment and infection have no relationship. the difference 63.4% was due to chance.

H_A: The difference in infection rate was not due to chance, and vaccine works.

2. Can H_0 and H_A be true at the same time? NO, only one of them could be true

3. If we believe H_0 is true, what does that mean?

It means no matter these 20 people got vaccine or not, there will be 11 people got infected

4. If we *claimed* H_0 is true, how to prove it? The simulation

The simulations where We pretend we know the vaccine being tested doesn't work

The purpose of the simulations: One wants to understand if the large difference we observed is common in these simulation.

If it is common, then maybe the difference was purely due to chanse, which means H_0 is true.

If it is very uncommon, then the possibility that vaccine was helpful seems more plausible

which means H_A is true.

5. How to implement these simulations?

(1) Prepare 20 Cards with 11 marked as "infected" and 9 "no infected"

(2) shuffle them thoroughly and deal 14 in treatment group, and 6 in control group

(3) Then we calculate the difference between the proportion of control and treatment

(4) Repeating (2) & (3) many times (100 times) and get a distribution from chance alone.

(5) What do those simulations tell us?

It apears that the large difference (64.3%) only happen twice out of 100 simulations which is very uncommon.

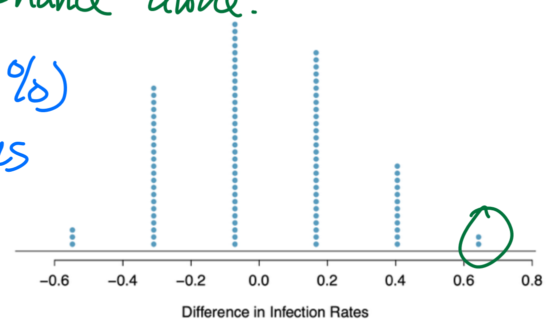

Figure 2.31: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where in these simulations infections are unaffected by the vaccine. Two of the 100 simulations had a difference of at least 64.3%, the difference observed in the study.

So, H_0 is NOT true and H_A is true