Table Probabilities and Independence

Dr Tom Ilvento
Department of Food and Resource Economics

Overview

- This lecture will focus on working with categorical data and building tables
- It will walk you through cross-tabulation of categorical data
- And show you how to percentage a table
- I will show some things in context of basic rules of probability – just to show you how to get around in a table
- I will also show how to build a model of independence

Basic Rules of Probability

- **Probability of a Union** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Conditional Probability** $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- **Probability of an Intersection** $P(A \cap B) = P(B)P(A \mid B)$

Cross-Tabulation of Treatment Type versus Still Smoking After 8 Weeks

		Subject Still Smoking		
		YES	**NO**	
Subject Treatment	**Nicotine Patch**	64	56	
	Placebo	96	24	

- **These are the Row Margins - they show the total for each row**
- **They are "fixed" by the design as the rows represent the Treatment**

Cross-Tabulation of Treatment Type versus Still Smoking After 8 Weeks

		Subject Still Smoking		
		YES	**NO**	**Row Margins**
Subject Treatment	**Nicotine Patch**	64	56	**120**
	Placebo	96	24	**120**

- **These are the Column Margins - they show the total for each Column**
- **They are the result of the experiment as the columns represent the outcome**

Let Event A = Received a Nicotine Patch.

		Subject Still Smoking		
		YES	**NO**	**Row Margins**
Subject Treatment	**Nicotine Patch**	64	56	**120**
	Placebo	96	24	**120**
	Column Margins	**160**	**80**	**240**

- What is the Probability of Event A? Denoted as P(A)
- **P(A) = 120/240 = .5**

Let Event B = No Longer Smoking

		Subject Still Smoking		
		YES	**NO**	**Row Margins**
Subject Treatment	**Nicotine Patch**	64	56	120
	Placebo	96	24	120
	Column Margins	160	80	240

- What is the Probability of Event B? Denoted as P(B)
- **P(B) = 80/240 = .333**

What is the Union of Events A and B?

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- What is the union of Events A (Received Nicotine Patch) and B (No Longer Smoking)
- (A ∪ B) =
- 120 Everyone who received the patch
- \+ 80 Everyone who no longer smokes
- \- 56 Everyone who is both
- **(A ∪ B) = 120 + 80 - 56 = 144**
- **P(A ∪ B) = 144/240 = .60**

Intersection of Receiving the Patch Versus No Longer Smoking

- What is the Intersection Receiving the Patch Versus No Longer Smoking?
- (A ∩ B) = ?
- This everyone who Received the Patch **AND** also is No Longer Smoking
- From the table we can see the cell that corresponds to this statement
- (A ∩ B) = **56**
- **P(A ∩ B) = 56/240 = .233**

		Subject Still Smoking		
		YES	NO	Row Margins
Subject Treatment	Nicotine Patch	64	56	120
	Placebo	96	24	120
	Column Margins	160	80	240

Probability Formulas Check

- **Probability of a Union** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- P(A) = .5
- P(B) = .333
- P(A) + P(B) = .833
- P(A ∩ B) = .233
- **P(A ∪ B) = .833 - .233 = .600**

Conditional Probability

- A **Conditional Probability** statement would be "**The probability of No Longer Smoking given you received the Nicotine Patch**" and is defined as
- **P(B|A) = .233/.50 = .467**
- I can solve for the P(B|A) directly, as long as I understand how to percentage my table

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$$

Conditional Probability

- There are 120 in total who received the Nicotine Patch in the study – see the number in the row margin
- **This is the given, as in given you received the Nicotine Patch**
- And 56 of those that received the patch were not smoking after 8 weeks
- So, **P(B|A) = 56/120 = .467**
- In a cross-tab this is called the row percentage
- It is a **conditional probability**, conditioned on the row attribute

Probability of Not Smoking Given you received the Nicotine Patch

		Subject Still Smoking		
		YES	NO	Row Margins
Subject Treatment	Nicotine Patch	64	56	120

- The new table is just the condition row - the ***given***
- **P(B|A) = 56/120 = .467**

The Compliment of A - Not Receiving the Nicotine Patch

- The Complement of A would be "Not Received the Patch" or "Received the Placebo"
- Denoted as A^c
- aka "Placebo"
- What is the $P(A^c)$ and $P(A^c \cap B)$?
  - **$P(A^c)$ = 120/240 = .50**
  - **$P(A^c \cap B)$ = 24/240 = .10**

		Subject Still Smoking		
		YES	NO	Row Margins
Subject Treatment	Nicotine Patch	64	56	120
	Placebo	96	24	120
	Column Margins	160	80	240

The Conditional Probability of Not Smoking for A^c

- The probability of No Longer Smoking given you received the Placebo
- $P(B|A^c)$
  - **$P(B|A^c)$ = .10/.50 = .20**
- **The easier way is to solve it from the table:**
  - **$P(B|A^c)$ = 24/120 = .20**

$$P(B \mid A^c) = \frac{P(A^c \cap B)}{P(A^c)}$$

		Subject Still Smoking		
		YES	NO	Row Margins
Subject Treatment				
	Placebo	96	24	120

Look at the SAS output for this data

```
TABLE OF TREATMENT BY STILL SMOKING

TREATMENT     STILL SMOKING
Frequency,
Percent  ,
Row Pct  ,
Col Pct  , YES    , NO     ,  Total
---------------------------
NICOTINE ,     64 ,     56 ,    120
         ,  26.67 ,  23.33 ,  50.00
         ,  53.33 ,  46.67 ,
         ,  40.00 ,  70.00 ,
---------------------------$
PLACEBO  ,     96 ,     24 ,    120
         ,  40.00 ,  10.00 ,  50.00
         ,  80.00 ,  20.00 ,
         ,  60.00 ,  30.00 ,
---------------------------$
Total         160      80       240
            66.67   33.33    100.00
```

Look at the first cell - Nicotine Patch who are Still Smoking

Frequency Percent Row Pct Col Pct	YES
Nicotine	**64** **26.67** **53.33** **40.00**

Percent	The cell value over the total	**64/240*100 = 26.67**
Row Pct	The cell value over the row margin on the right	**64/120*100 = 53.33**
Col Pct	The cell value over the column margin on the bottom	**64/160*100 = 40.00**

Look at the second cell - Nicotine Patch who are No Longer Smoking

Frequency Percent Row Pct Col Pct	No
Nicotine	**56** **23.33** **46.67** **70.00**

Percent	The cell value over the total	**56/240*100 = 23.33**
Row Pct	The cell value over the row margin on the right	**56/120*100 = 46.67**
Col Pct	The cell value over the column margin on the bottom	**56/80*100 = 70.00**

Now answer me this....

- The P(A|B) for our table = ?
- This is the Probability of receiving a Nicotine Patch given you are No Longer Smoking
- We can solve this using the probability formula
  - **P(A|B) = P(A∩B)/P(B) = .233/.333 = .70**
- Or we can simply calculate a column percentage
  - **P(A|B) = 56/80 = .70**

		Subject Still Smoking	
		NO	
Subject Treatment		56	
		24	
		80	

Does this make any sense??

How to percentage a table

- If you can specify a conditional probability
- Or if you can specify that one variable causes or influences a second variable
  - The first variable is called an **independent variable** (this is the given)
  - The second is the **dependent variable**
- Percentage in the direction of the independent variable
  - If the independent variable is at the top, use column percentages
  - If the independent variable is on the side, use row percentages

What is the best way to percentage the smoking data?

- It seems to me that:
  - Given the analysis fits a designed experiment
  - And subjects were randomly assigned to a treatment (Nicotine Patch) and control group (Placebo)
  - And there is a time lag between when the patch is first administered and when the recording of "still smoking" occurred (8 weeks)
  - And the interest of the experiment is whether the patch helped keep people from smoking
- The direction of the conditional probability is expected to be, **given that you received a patch, what is the probability that you are no longer smoking?**

Independence

- Events A and B are independent events if the occurrence of B does not alter the probability that A has occurred.
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$
- Events that are not independent are dependent

Independence

- Furthermore, if Events A and B are independent, then the probability of their intersection simplifies to:
  - $P(A \cap B) = P(A)P(B)$
- Why???
  - $P(A \cap B) = P(A)P(B|A)$
  - And if A and B are independent then, $P(B|A) = P(B)$
- So, with independence, $P(A \cap B) = P(A)P(B)$

What would our data look like if it were independent?

- One strategy in statistics is to propose a hypothesized distribution and then compare what we observe to our model of independence
- We could propose a model of independence.
- If our variables were independent of each other, then the data would be based on the marginal distributions
- Our model of independence is based on row and column marginals

Observed versus Expected Data

- this is the data we **observe** based on the results of the experiment
- and this is the data we **"expect"** based on a **model of independence**
- Notice in the model of independence the row and column marginals are the same, but the cell frequencies changed.
- Next, how to generate **expected frequencies**

		Subject Still Smoking		
		YES	NO	Row Margins
Subject Treatment	Nicotine Patch	64	56	120
	Placebo	96	24	120
	Column Margins	160	80	240

Solving for Expected Frequencies

- Remember, I wanted a model of independence, which means
  - P(B|A) = P(A∩B)/P(A) = P(B)
  - P(A|B) = P(A∩B)/P(B) = P(A)
- A simple way to make this happen is make the expected frequencies a function of the row and column marginals

Solving for Expected Frequencies

- For the second cell, I want the expected frequency e_{12} to equal the following:
  - $e_{12}/80 = 120/240$
  - $e_{12} = (80*120)/240 = 40$
- If this cell is 40, then
  - P(B|A) = P(B)
  - The probability of Not Smoking given the Nicotine Patch = the probability of Not Smoking
  - 40/120 =
  - = 80/240 = .333

		Subject Still Smoking		
		YES	NO	Row Margins
Subject Treatment	Nicotine Patch	64	?	120
	Placebo	96	24	120
	Column Margins	160	80	240

Solving for Expected Frequencies

- Patch, Yes
  - = (160*120)/240 = 19,200/240 **= 80**
- Patch, No
  - = (80*120)/240 = 9,600/240 **= 40**
- Placebo, Yes
  - = (160*120)/240 = 19,200/240 **= 80**
- Placebo, No
  - = (80*120)/240 = 9,600/240 **= 40**

		Subject Still Smoking		
		YES	NO	Row Margins
Subject Treatment	Nicotine Patch	80	40	120
	Placebo	80	40	120
	Column Margins	160	80	240

Model of Independence

- Generating expected frequencies under a model of independence can be very useful
- We can compare our model to the data to see how well the data fits the expected frequencies – how we do this will come later!
- Depending upon our model, we may or may not want to see a good fit.
  - With a Model of Independence, we often don't want a good fit!
  - Because a bad fit means there is a relationship between the two variables – **using a patch influences whether a subject stops smoking.**

Summary

- Let me simplify – know how to percentage a table!!!!
  - Decide on total, row or column percentages
  - Can be based on assuming one variable to be dependent and another independent
- The concept of independence is very important in statistics!
  - We can fit a model of independence based on row and column margins
  - We can see how our model compares with the actual data