Collecting and analyzing quantitative (survey) data with statistics

(and rest of interrater relaibility)

Mainack Mondal

CS 60081
Autumn 2020

Cohen's kappa

- Two raters classify each of N items into one of C categories
  - P0 is the observed agreement
  - PE is the expected agreement (when each rater behave randomly)
  - The kappa = (P0 – PE)/ (1- PE)
  - Max - min value?

More on Cohen's kappa

Sentences	**Label assigned by coder 1 (any of the C labels)**	**Label assigned by coder 2 (any of the C labels)**
Sentence 1	X1	X1
Sentence 2	X3	X1
…		
…		
…		
Sentence n-2	X10	X5
Sentence n-1	X11	X11
Sentence n	X4	X4

More on Cohen's kappa

Sentences	**Label assigned by coder 1 (any of the C labels)**	**Label assigned by coder 2 (any of the C labels)**
Sentence 1	X1	X1
Sentence 2	X3	X1
…		
…		
…		
Sentence n-2	X10	X5
Sentence n-1	X11	X11
Sentence n	X4	X4

More on Cohen's kappa

- Lets take the two coder example– each of the code want to label N sentences with "Yes", "NO" labels
  - So there are C = 2 labels (Yes, NO)
  - Lets assume, total rows to label, N = a + b + c + d
  - First create the confusion matrix

	Coder2_YES	Coder2_NO
Coder1_YES	a	b
Coder1_NO	c	d

More on Cohen's kappa

- Lets take the two coder example– each of the code want to label N sentences with "Yes", "NO" labels
  - So there are C = 2 labels (Yes, NO)
  - Lets assume, total rows to label, N = a + b + c + d
  - First create the confusion matrix

	Coder2_YES	Coder2_NO
Coder1_YES	a	b
Coder1_NO	c	d

  - P0 = proportion of agreement = (a + d)/(a+b+c+d)

More on Cohen's kappa

- Lets take the two coder example– each of the code want to label N sentences with "Yes", "NO" labels
  - So there are C = 2 labels (Yes, NO)
  - Lets assume, total rows to label, N = a + b + c + d
  - First create the confusion matrix

	Coder2_YES	Coder2_NO
Coder1_YES	a	b
Coder1_NO	c	d

  - P0 = proportion of agreement = (a + d)/(a+b+c+d)
  - PE = Pr (both will say YES at random) + Pr (both will say NO at random) = $\frac{a+b}{a+b+c+d} * \frac{a+c}{a+b+c+d} + \frac{c+d}{a+b+c+d} * \frac{b+d}{a+b+c+d}$

Example

	Coder2_YES	Coder2_NO
Coder1_YES	34	26
Coder1_NO	19	21

PO = (34 + 21) / (34+26+19+21) = 55/100= 0.55

PE = (34 + 26)/(100) * (34 + 19)/100 +

(19 + 21)/(100) * (26 + 21)/100 = 0.318 + 0.188 = 0.506

Kappa = (PO - PE) / (1 - PE) = (0.55 – 0.506) / (1 – 0.506) = 0.08

Interpretation

<.20	Poor		
.21-.40	Fair	.61-.80	Substantial
.41-.60	Moderate	>.81	Excellent

Other variations

- Scott's Pi
- Fleiss's Kappa (multi-rater agreemnt)
- Krippendorff's alpha (multi-rater agreement, handles missing data)

QUESTION

- The following Cohen's kappa (*k*) values strongly suggest that the instrument, the raters, the training protocol, or other aspects of the measurement situation need to be modified or there is an error in the kappa calculation (select all that apply):

A. $k = .69$

B. $k = .20$

C. $k = 3.2$

D. $k = .80$

Roadmap

- Qualitative Data Analysis
  - Selecting participants
  - Data analysis techniques
- Inter-rater agreement
- Quantitative data analysis

How to analyze data

Statistics

- In general: analyzing and interpreting data
- Statistical hypothesis testing: is it unlikely the data would look like this unless there is actually a difference in real life?
- Statistical correlations: are these things related?

Type of data

- Quantitative/numerical
  - Discrete (e.g., #emails)
  - Continuous (e.g., age)
- Categorical
  - Nominal or no order (e.g., male-female)
  - Ordinal or ordered (e.g., Ex, A, B, …, F)
- Q: Why cannot we just assign 1,2,3,… etc. ordered discrete values to the ordinal variables?

Hypothesis testing

- Causation (X causes Y)
  - vs. correlation (X is related to Y)
- Develop a hypothesis (e.g., age is related to typing speed)
  - Assign to conditions (include a control)
  - Terminology: “Condition” = “Treatment”
- H0 (null hypothesis): there is no effect
- H1 (alternative hypothesis): there is an effect

Way to do the test

- You have a set of values for variable X (e.g., age)
  - x1, x2, x3, …
- You have a set of values for variable Y (e.g., typing speed)
  - y1, y2, y3, …
- Question: Is higher age affect the typing speed? Why do you need a test?

Way to do the test

- You have a set of values for variable X (e.g., age)
  - x1, x2, x3, …
- You have a set of values for variable Y (e.g., typing speed)
  - y1, y2, y3, …
- Question: Is higher age affect the typing speed? Why do you need a test?

- You chose a test H (often a python or R function)
  - Statistic, p = H ([x1, x2, x3, …], [y1, y2, y3, …])
  - p – value is essentially a probability that the statistic value occurred randomly (i.e., there is no effect aka H0 is true)
  - So if p is small (generally < 0.05, called α) you reject H0

Is P value enough?

- No! Consider:
  - Effect size (magnitude of the effect of the manipulation)
  - Power (long-term probability of rejecting H0 if there really is a difference)
- Type 1 error: wrongly reject H0 even if there is no effect (α)
- Type 2 error: wrongly fail to reject H0 even if there is an effect (β)

Type I errors

- Type I error (false positive)
  - You would expect this to happen 5% of the time if $\alpha = 0.05$

Type II errors

- Type II error (false negative)
  - There is actually a difference, but you didn't see evidence of a difference
- Statistical power is the probability of rejecting the null hypothesis (no effect) if you should → 1 – Pr(Type II Error)
  - You could do a **power analysis**,
    - Minimum sample size to achieve a given effect size
    - How many times do you have to toss a coin to know that Pr(head) = 0.7?
    - Requires that you can estimate the effect size
    - Bonferroni's correction

How to pick the right test?

- Different types of variables?
- Different data distributions? (e.g., normal vs., non-normal)
- Parametric vs. non-parametric tests

Check the variable type first

- First, types of variables
- You want to see the correlation between age and number of emails sent per day
  - age: independent variable, you can vary it by taking different users
  - #emails/day: dependent variable, you want to measure
  - Technical expertise, job, … : co-variate

Picking the right test: A limited cheat sheet

Focusing on parametric tests!

		Independent Variable	
		Categorical	**Quantitative**
Dependent Variable	**Categorical**	Chi-Squared Test Fisher's Exact Test	Logistic Regression
	Quantitative	t-Test ANOVA	Correlation Linear Regression

Picking the right test: A limited cheat sheet

Focusing on parametric tests!

		Independent Variable	
		Categorical	**Quantitative**
Dependent Variable	**Categorical**	Chi-Squared Test Fisher's Exact Test	Logistic Regression
	Quantitative	t-Test ANOVA	Correlation Linear Regression

Parametric vs non-parametric tests

When to use what?

- Finding relations between two numerical variables
  - As the age of a man increases, his/her max running speed decreases
  - Pearson's correlation / Spearman's rank correlation
- Finding relations between two categorical variables
  - People randomly assigned to exercise more than twice a week (as opposed to less than once a week) are more likely to be rated as healthy (as opposed to unhealthy)
  - $\chi 2$, Fisher's exact test

When to use what?

- Comparing a variable value between two groups (numerical)
  - People who exercise more than twice a week (as opposed to less than twice) are more likely to take a shorter time to run a race
  - ANOVA, Kruskal-Wallis, etc.

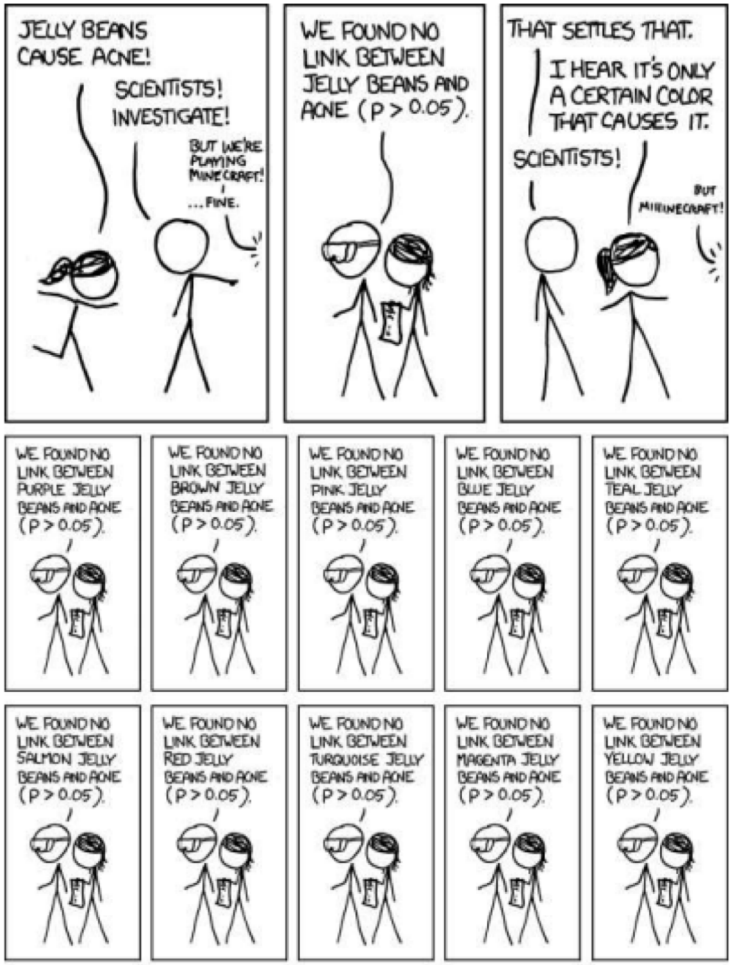
- Lots of factors has effect on the dependent variable (numerical)
  - Regression (Y = a1x1 + a2x2 etc…)
- Lots of factors has effect on the dependent variable (category)
  - Logistic regression

We talked about…

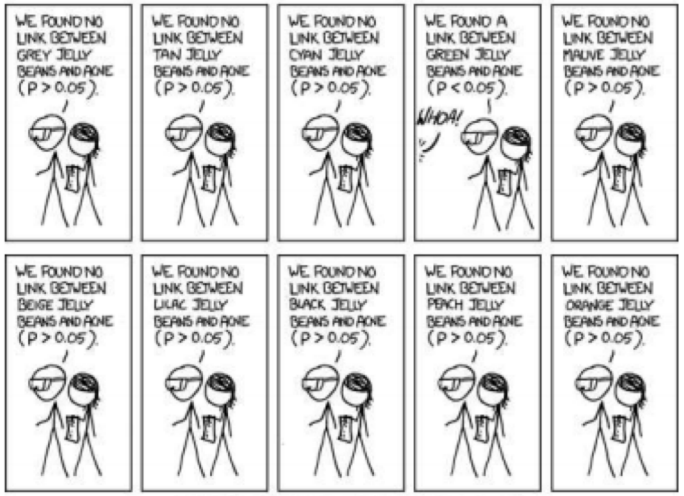
- Type I error : Wrongly reject H0 even if whatever you observed happened due to random chance
  - expect this to happen 5% of the time if $\alpha = 0.05$
- Type I error : Wrongly fail to reject H0 even if whatever you observed happened due to non-random process

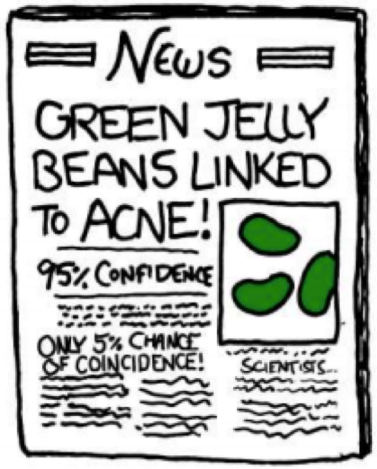
- What happens if you conduct a lot of statistical tests in one experiment?
  - In at least one case $p < 0.05$

A xkcd example



https://xkcd.com/882/





Bonferroni correction

- Divide α by #tests
  - Say you did 1000 tests
  - Previous : Likely to get $p < \alpha = 0.05$ for at least one test
  - Now: Much harder to get $p < \alpha /1000 = 0.00005$ even for one test

Case 1:
Dependent variable (DV): Categorical
Independent variable (IV): Categorical

Chi-squared (χ2) Test

- Example research questions
  - Does the gender (male, female) correlate with a user's favorite color?
  - Does the cuisines it ate this month correlate to its privacy concerns?

- H0 : Variable X values are equally distributed across variable Y values (independence or no effect)
- (Not covered today) Goodness of fit: Does the distribution we observed differ from a theoretical distribution?

Contingency table

- Rows are r values of one variable, Columns are c values of other variable

CreateAnnoying Counts:

	0	1
0	161	32
1	165	33
2	168	34
3	170	30
4	164	32
5	161	35
6	167	32
7	129	60
8	128	61
9	154	40
10	153	40
11	154	38
12	142	42
13	121	67
14	124	76

Percentages:

	0	1
0	"83.42%"	"16.58%"
1	"83.33%"	"16.67%"
2	"83.17%"	"16.83%"
3	"85%"	"15%"
4	"83.67%"	"16.33%"
5	"82.14%"	"17.86%"
6	"83.92%"	"16.08%"
7	"68.25%"	"31.75%"
8	"67.72%"	"32.28%"
9	"79.38%"	"20.62%"
10	"79.27%"	"20.73%"
11	"80.21%"	"19.79%"
12	"77.17%"	"22.83%"
13	"64.36%"	"35.64%"
14	"62%"	"38%"

- $\chi 2 = 97.013$, df = 14, p = 1.767e-14

Chi-squared (χ2) usage

- Use χ2 if you are testing one categorical variable (usually a demographic factor) impacts another categorical variable
  - If you have < 5 data points in a single cell of your contingency table, use Fishher's exact test

- DO NOT use this test for numerical variables

What about Likert scale?

- Some people treat it as continuous (assign 1 to an option, 2 to another option etc.) (a controversial step)
- Others treat it as ordinal (better choice)
  - In that case, use Mann-Whitney U / Kruskal-Wallis (non-parametric)
- A simple alternative
  - Bin the data into binary agree/non-agree, or comfortable/non-comfortable categories
    - Now you can use Chi squared test (parametric)

Case 2:
Dependent variable (DV): Categorical
Independent variable (IV): Quantitative

Choosing a numerical test

- Do your data follow a normal (gaussian distribution)?
  - Use Shapiro-Wilk normality test
  - Yes → parametric test, No → non-parametric test

- Considerations
  - Is your data independent? → not from same family in case of a skin-color-based hypothesis
  - If not → repeated-measures, mixed models

Why might your data not be independent

- Reason 1: Non-independent sample (change sampling)
- Reason 2: Inherent design, e.g., within subjects design (then its ok)

Numerical data

- Popular question: Are values bigger in one group?
- Normal, continuous data (for comparing mean):
  - H0 : There are no differences in the means
  - 2 conditions: t-test (age vs. typing speed)
  - 3+ conditions: ANOVA

- Non-normal data / ordinal data:
  - H0 : No group tends to have larger values.
  - 2 conditions: Mann-Whitney U (likert scale data vs. likert scale data)
  - 3+ conditions: Kruskal-Wallis

Case 3:
Dependent variable (DV): Quantitative

Correlation

- Popular question: is X related to Y?
- less good: Pearson correlation
  - Assumes both variables as normally distributed
  - Only look for linear relationship

- Preferred: Spearman's rank correlation coefficient (Spearman's ρ)
  - Evaluates a relationship's monotonicity (always going in the same direction or staying the same)

Regressions

- What is the relationship among variables?
  - Generally one outcome (dependent variable)
  - Often multiple factors (independent variables)

- The type of regression you perform depends on the dependent variable i.e., outcome
  - Binary outcome: logistic regression
  - Ordinal outcome: ordinal / ordered regression
  - Continuous outcome: linear regression

Outcome of a regression

- Normally, outcome = ax1 + bx2 + c + …

- Interactions
  - when two variables are not simply additive. Instead, their interaction impacts the outcome
  - Then outcome = ax1 + bx2 + c + d(x1*x2) + …

Example

- Outcome: If a user can complete a task (Yes/No)
  - Logistic regression (binary outcome)

- Independent variables
  - Age
  - #prior takes completed
  - Income
  - Job
  - …

In case of non-independence?

In case of non-independence use

- Repeated measures (multiple measurements of the same thing)
  - e.g., before and after measurements of a unicorn's time to finish a race
- Paired t-test (two samples per participant, two groups)
- Repeated measures ANOVA (more general)

Picking a test [IMPORTANT]

- http://webspace.ship.edu/pgmarr/Geo441/Statistical%20Test%20Flow%20Chart.pdf
- http://abacus.bates.edu/~ganderso/biology/resources/statistics.html
- http://med.cmb.ac.lk/SMJ/VOLUME%203%20DOWNLOADS/Page%2033-37%20-%20Choosing%20the%20correct%20statistical%20test%20made%20easy.pdf

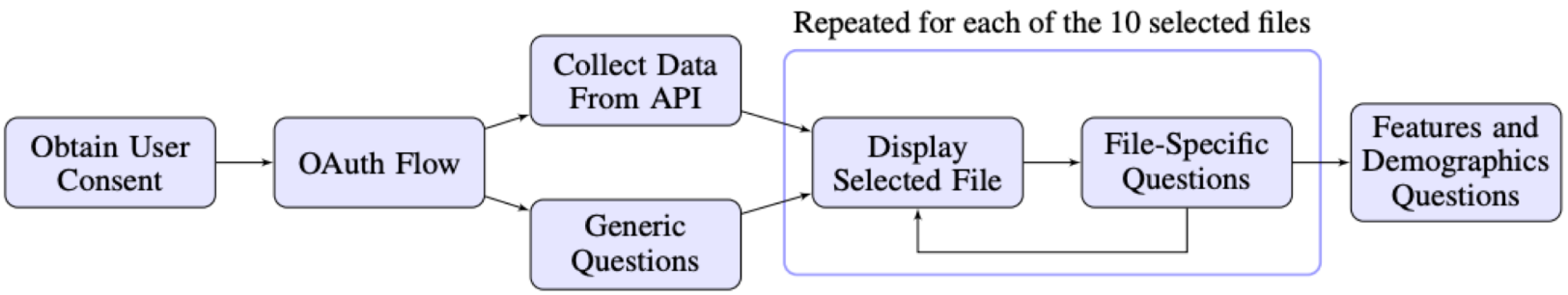
Case study: Longitudinal data management in cloud storage

Khan et. Al., CHI'18

Motivation

- People change over time
  - And so might their privacy/security requirements of their data
  - Question: Identify whether there is a need for longitudinal data management in cloud storage services

Approach



- How to find what factors does privacy decisions depend upon for 100 participants?

Steps

- First the variables:
  - Remembrance (dependent) vs. ownership (independent)
  - Remembrance: remember this file? Strongly agree to Strongly disagree
  - Ownership: owner, editor, viewer

Steps

- First the variables:
  - Remembrance (dependent) vs. ownership (independent)
  - Remembrance: remember this file? Strongly agree to Strongly disagree
  - Ownership: owner, editor, viewer
  - Both categorical

Recap: A limited cheat sheet

Focusing on parametric tests!

		Independent Variable	
		Categorical	**Quantitative**
Dependent Variable	**Categorical**	Chi-Squared Test Fisher's Exact Test	Logistic Regression
	Quantitative	t-Test ANOVA	Correlation Linear Regression

Parametric vs non-parametric tests

Steps

- First the variables:
  - Remembrance (dependent) vs. ownership (independent)
  - Remembrance: remember this file? Strongly agree to Strongly disagree
  - Ownership: owner, editor, viewer
  - Both categorical AND each combination of these values has more than 5 feedback → Chi Square

Remembrance vs. ownership

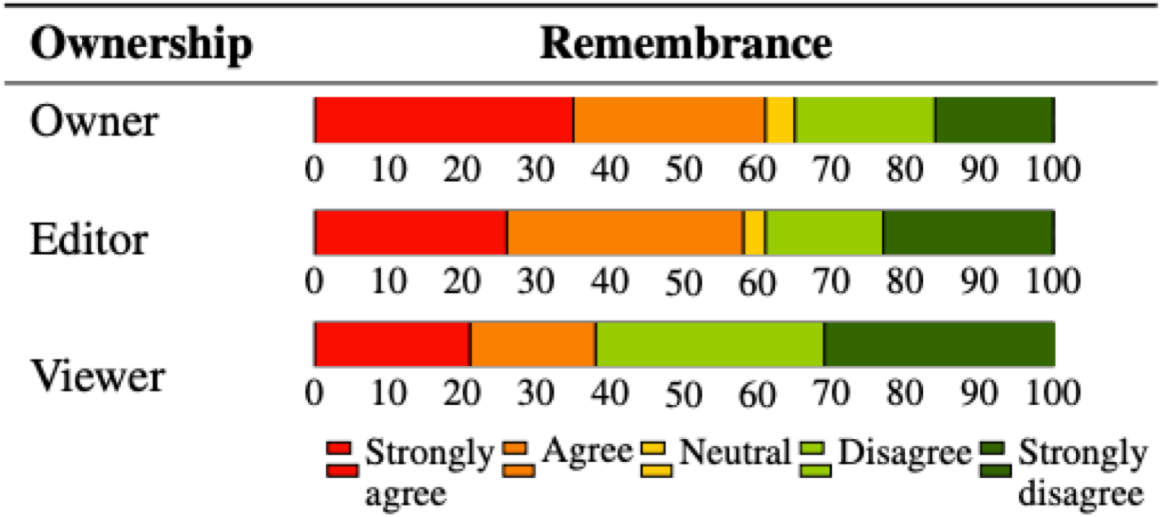


Figure 3: Comparison of file ownership and remembrance (agreement or disagreement that they remembered the file was stored in their cloud account). File ownership had a significant positive correlation with remembering the file was stored in the cloud ($\chi^2(8, N = 862) = 32.244$, $p < .001$).

Remembrance vs. ownership

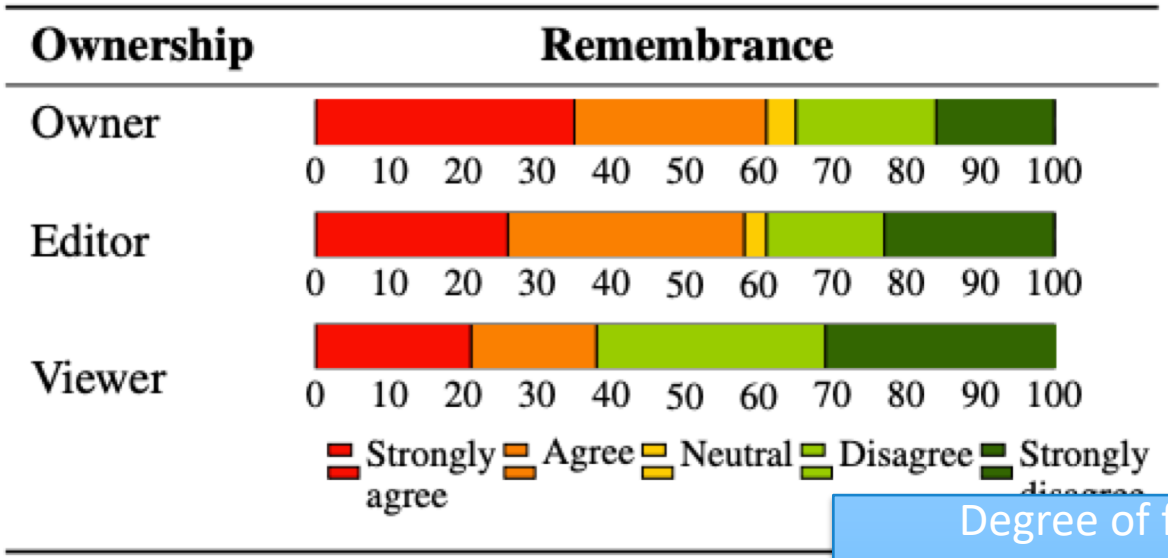


Figure 3: Comparison of file ownership a
(agreement or disagreement that they remem
stored in their cloud account). File ownershi
positive correlation with remembering the
the cloud ($\chi^2(8, N = 862) = 32.244, p < .001$).

Degree of freedom = 8
#observations = 862
Pr(this value of Chi happened due to random chance) < 0.001

Other questions

- Recognition vs. ownership
- Deletion decision vs. ownership
- Participant background (technical/non-technical) vs. ownership
- Keep-sharing decision vs. ownership

- All Chi-square
  - Then answer *why* with qual coding