LECTURE 6: BIVARIATE CATEGORICAL DATA

Adapted from material by Martina Morris

Recap

Big picture: Bivariate descriptives

- These two weeks focus on descriptive summaries for relationships between two variables, X and Y (“bivariate data”)

X: Y:	Nominal	Ordinal	Continuous
Nominal	This Week	This Week	Last Week
Ordinal	This Week	This Week	Last Week
Continuous			Last Week

As always, the measurement scale of each variable determines the appropriate summaries

- Last week: summaries for quantitative/continuous data
  - Primary focus on summarizing linear relationships
- This week: summaries for qualitative/discrete data
  - Primary focus is on summarizing conditional probabilities

Discrete data (recap)

Measurement	Ordered?	True zero?	Type
Nominal	*No*	*No*	*Discrete*
Ordinal	Yes	*No*	*Discrete*

- All of the methods we will cover this week can be used for both nominal and ordinal data
- But they do not make use of the rank (quantitative) information in ordinal data

Cross-tabulated count data

- We classify every observation in the data by a pair of values (r,c) for two discrete variables, R and C
  - For example, marital status and home ownership
  - r and c are called “levels” of the variables R and C

- The result is a two-way table
  - Levels of R define the rows
  - Levels of C define the columns
  - Each observation falls into one cell of this table

- The cell count represents the number of observations at that joint level of R and C (r,c).

Basic table elements

Row Variable	Column Variable				
	C1	C2	C3	C4	Row Total
R1	Cell				Row Margin
R2					
R3					
Column Total	Column Margin				Table Total

Row and column variables

- Table variables R and C have discrete values

- Sometimes the variable is discrete in its original scale
  - Nominal (sex or political party affiliation)
  - Ordinal (educational degree completed)

- But sometimes it is continuous in its original scale
  - dates categorized into 5 year intervals
  - age categorized into 4 groups
  - *This is called "discretizing" the continuous variable*

Cell entries in a table

- Tables are like 2-way histograms
  - They can show frequencies, or probabilities

- There are three distributions again
  - **Marginal** distributions of the row and column variables
  - **Joint** distribution of the two variables
  - And, the **Conditional** distribution

Example: Education by age

TABLE 6.1 Years of school completed, by age (thousands of persons)

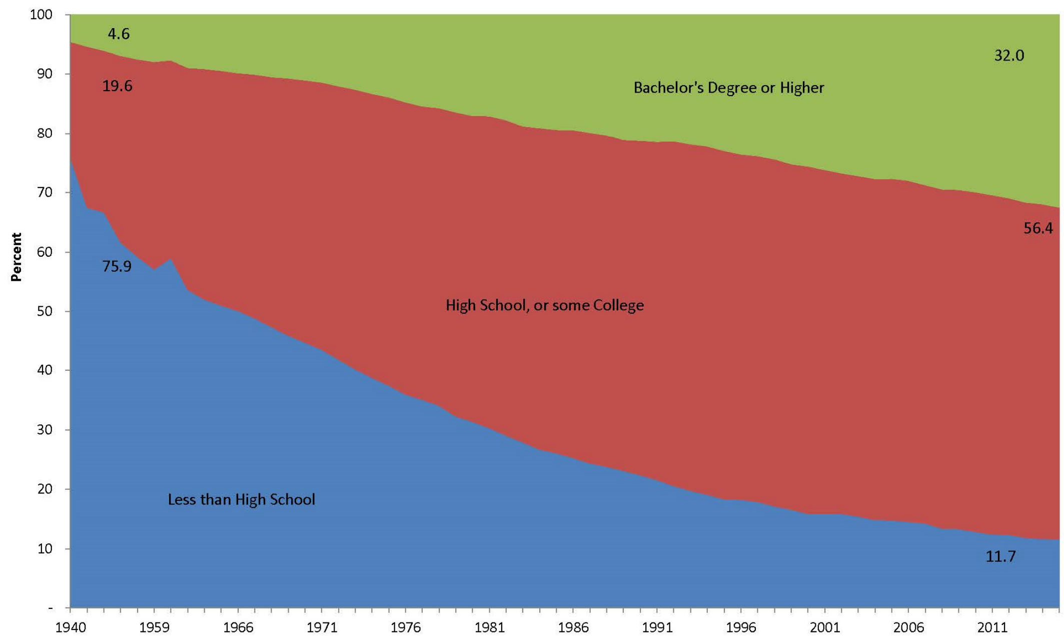
	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

Note: Both education and age have been discretized

Some context

https://www.census.gov/hhes/socdemo/education/data/cps/historical/

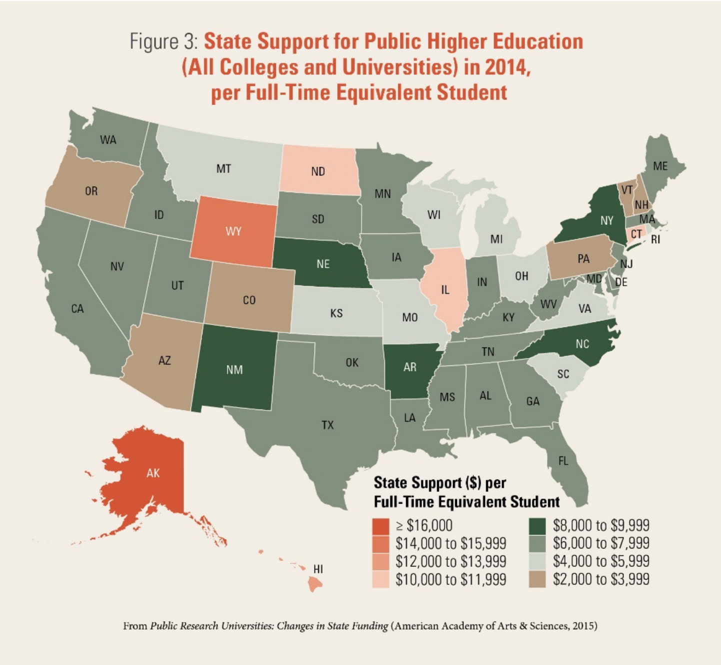
Figure 2: Percent of Population Age 25 and over by Educational Attainment: 1940-2015

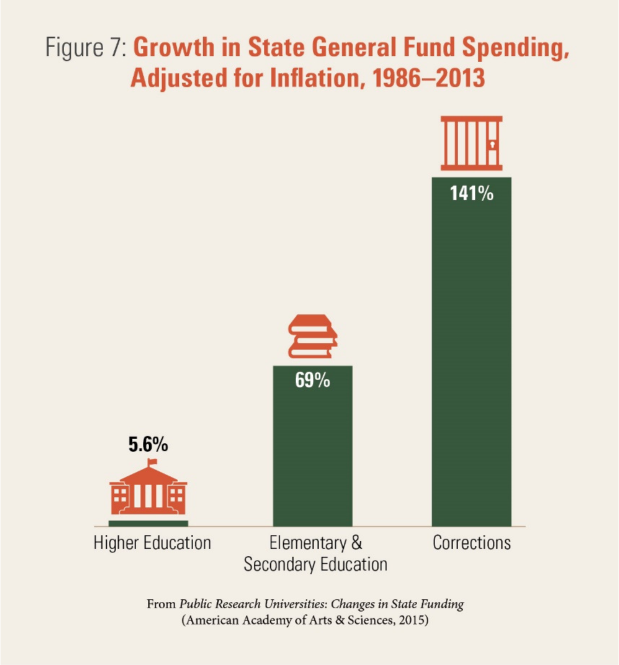


Sources: U.S. Census Bureau. 1947, 1952-2002 March Current Population Survey, 2003-2015 Annual Social and Economic Supplement to the Current Population Survey; 1940-1960 Census of Population.

Some more context…

https://www.amacad.org/content/publications/publication.aspx?d=21942





Marginal (univariate) distributions

The univariate distributions of the row and column variables are shown in the **marginal distributions** (on the margins of the table).

Marginal distribution of education

TABLE 6.1 Years of school completed, by age (thousands of persons)

	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

2000 U.S. census

Marginal distribution of age

Joint (bivariate) distribution

The cells in the table represent the **joint distribution** -- and show how the distribution of education and age co-vary.

Joint distribution of age and education

TABLE 6.1 Years of school completed, by age (thousands of persons)

Education	Age group 25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

2000 U.S. census

Frequency vs. Probability distributions

	25-34	35-54	55+	Row Total
<HSD	4,459	9,174	14,226	27,859
HSG	11,562	26,455	20,060	58,077
SC	10,693	22,647	11,125	44,465
BA	11,071	23,160	10,597	44,828
Column Total	37,785	81,436	56,008	175,229

Frequency distribution

	25-34	35-54	55+	Row %
<HSD	2.5	5.2	8.1	15.9
HSG	6.6	15.1	11.4	33.1
SC	6.1	12.9	6.3	25.4
BA	6.3	13.2	6.0	25.6
Column %	21.6	46.5	32.0	100.0

Probability distribution

Marginal (univariate) probabilities

	25-34	35-54	55+	Row Total
<HSD	4,459	9,174	14,226	27,859
HSG	11,562	26,455	20,060	58,077
SC	10,693	22,647	11,125	44,465
BA	11,071	23,160	10,597	44,828
Column Total	37,785	81,436	56,008	175,229

27,859/175,229

=15.9

	25-34	35-54	55+	Row %
<HSD	2.5	5.2	8.1	15.9
HSG	6.6	15.1	11.4	33.1
SC	6.1	12.9	6.3	25.4
BA	6.3	13.2	6.0	25.6
Column %	21.6	46.5	32.0	100.0

Marginal probabilities sum to 100%

Joint (bivariate) probabilities

	25-34	35-54	55+	Row Total
<HSD	4,459	9,174	14,226	27,859
HSG	11,562	26,455	20,060	58,077
SC	10,693	22,647	11,125	44,465
BA	11,071	23,160	10,597	44,828
Column Total	37,785	81,436	56,008	175,229

23,160/175,229

=13.2

	25-34	35-54	55+	Row %
<HSD	2.5	5.2	8.1	15.9
HSG	6.6	15.1	11.4	33.1
SC	6.1	12.9	6.3	25.4
BA	6.3	**13.2**	6.0	25.6
Column %	21.6	46.5	32.0	100.0

Joint probabilities also sum to 100%

Does education vary by age?

	25-34	35-54	55+	Row %
<HSD	2.5	5.2	8.1	15.9
HSG	6.6	15.1	11.4	33.1
SC	6.1	12.9	6.3	25.4
BA	6.3	13.2	6.0	25.6
Column %	21.6	46.5	32.0	100.0

With the joint distribution, you do see some variation in the cell proportions: some age-education combinations are more likely than others

But the joint distribution is not great for summarizing association between age and education

Does education vary by age?

	25-34	35-54	55+	Row %
<HSD	2.5	5.2	8.1	15.9
HSG	6.6	15.1	11.4	33.1
SC	6.1	12.9	6.3	25.4
BA	6.3	13.2	6.0	25.6
Column %	21.6	46.5	32.0	100.0

From the joint distribution, you might think that 25-34 year olds have the same rates of college completion as 55+ year olds.

Does education vary by age?

	25-34	35-54	55+	Row %
<HSD	2.5	5.2	8.1	15.9
HSG	6.6	15.1	11.4	33.1
SC	6.1	12.9	6.3	25.4
BA	6.3	13.2	6.0	25.6
Column %	21.6	46.5	32.0	100.0

Or that HS Graduates are more likely than HS Dropouts to be in the oldest group (55+)

But you need to be careful

	25-34	35-54	55+	Row %
<HSD	2.5	5.2	8.1	15.9
HSG	6.6	15.1	11.4	33.1
SC	6.1	12.9	6.3	25.4
BA	6.3	13.2	6.0	25.6
Column %	21.6	46.5	32.0	100.0

In the joint distribution, the cell probabilities are influenced by the marginal probabilities (look at how they vary here, and why)

So the column margins will affect comparisons across columns
And the row margins will affect comparisons down the rows.

For these comparisons, **conditional distributions are a better choice**

Conditional distributions

- Better for showing the patterns of association between the row and column variables

- The probabilities *condition* on which row or column the observation falls into
  - So the marginal totals no longer influence the distribution

Conditional Distributions: by Column

	25-34	35-54	55+	Row Total
<HSD	4,459	9,174	14,226	27,859
HSG	11,562	26,455	20,060	58,077
SC	10,693	22,647	11,125	44,465
BA	11,071	23,160	10,597	44,828
Column Total	37,785	81,436	56,008	175,229

	25-34	35-54	55+	Row %
<HSD	11.8	11.3	25.4	15.9
HSG	30.6	32.5	35.8	33.1
SC	28.3	27.8	19.9	25.4
BA	29.3	28.4	18.9	25.6
Column %	100.0	100.0	100.0	100.0

Divide the cell counts by the *column* totals to get the conditional distributions by column

All the col %s now sum to 100

Now we can see that 25-34 year olds are much more likely to have completed college than 55+.

Conditional Distributions: by Row

Divide the cell counts by the row totals to get the row conditional distributions. All the row %s now sum to 100.

	25-34	35-54	55+	Row %
<HSD	16.0	32.9	51.1	100.0
HSG	19.9	45.6	34.5	100.0
SC	24.0	50.9	25.0	100.0
BA	24.7	51.7	23.6	100.0
Column %	21.6	46.5	32.0	100.0

Now we can see that HS dropouts are much more likely to be 55+ than HS graduates.

Summary

- There are three types of probabilities in a 2-way table
- The two **marginal** probabilities show the overall distribution of education and age in the sample.
- The **joint** probabilities show the fraction of the sample at each age-education level.
- The **conditional** probabilities highlight bivariate association
  - Whether the distribution of education varies by age, or
  - Whether the distribution of age varies by education

Summarizing association

- Think back to continuous bivariate data
  - There was a whole family of association measures based on deviations from the means
  - And how these deviations co-vary for two variables
- Is there a similar set of measures here?
  - No,
  - And yes...

No… and why

- The mean doesn't make any sense for nominal variables
  - "mean" marital status?
  - "mean" religion?

- So the deviations from the mean don't make any sense either

Yes … and why

- There is still a way to think about expected values.
- Say I gave you the following marginal distributions for commuting patterns by sex:

	Bus	**Car**	***Row total***
Men			100
Women			100
Col total	40	160	200

How many men would you expect take the bus if there were no association between sex and commuting choice?

Yes … and why

- There is a way to think about expected values here.
- Say I gave you the following marginal distributions for commuting patterns by sex:

	Bus	**Car**	***Row total***
Men	20		100
Women			100
Col total	40	160	200

How many men would you expect take the bus if there were no association between sex and commuting choice?

28

Expected Values

For cross-tabulated discrete data

Expected Values: The intuition

- The expected values for the joint distribution are a function of the marginal probabilities, and the total N

	Bus	Car	Row percent
Men	$200 * 0.5 * 0.2 = 20$	$200 * 0.5 * 0.8 = 80$	$\frac{100}{200} = 0.5$
Women	$200 * 0.5 * 0.2 = 20$	$200 * 0.5 * 0.8 = 80$	$\frac{100}{200} = 0.5$
Col percent	$\frac{40}{200} = 0.2$	$\frac{160}{200} = 0.8$	200 total persons

Taking a closer look

- The expression for the expected cell frequency :

$$200 * \frac{100}{200} * \frac{40}{200} = \frac{100 * 40}{200}$$

$$\frac{\text{Row total} \times \text{Column total}}{\text{Total } n \text{ for table}}$$

	Bus	Car	Row percent
Men	$200 * 0.5 * 0.2 = 20$	$200 * 0.5 * 0.8 = 80$	$\frac{100}{200} = 0.5$
Women	$200 * 0.5 * 0.2 = 20$	$200 * 0.5 * 0.8 = 80$	$\frac{100}{200} = 0.5$
Col percent	$\frac{40}{200} = 0.2$	$\frac{160}{200} = 0.8$	200 total

And finally, with some notation

Let the frequencies be denoted by n

	Bus	Car	Row total
Men	n_{11}	n_{12}	n_{1+}
Women	n_{21}	n_{22}	n_{2+}
Col total	n_{+1}	n_{+2}	n_{++}

We index the counts by row and column number: n_{rc}

And denote marginal totals with a "+" in the appropriate index:
n_{r+} for row totals
n_{+c} for col totals

Our expected cell count is

$$\hat{n}_{rc} \mid \text{no association} = \frac{n_{r+} n_{+c}}{n_{++}}$$

- What can we do with this?
  - Compare it to the observed frequency
- An intuitive metric: $\frac{obs}{exp}$
- A statistical metric : $\frac{(obs - exp)^2}{exp}$

This should remind you of something…

Other common summary statistics

- Risk and relative risk
  - We'll start looking at these today
- Odds and odds-ratios
  - We'll look at these on Wed

Risk measures

An example from UH Ch 4

Table 4.3 Smoking and Marital Status: Counts and Row Percentages

Smoking	Marital Status		
	Separated	Not Separated	Total
Neither smoked	41 (4.2%)	931 (95.8%)	972 (100%)
One smoked	41 (12.4%)	290 (87.6%)	331 (100%)
Both smoked	32 (16.4%)	163 (83.6%)	195 (100%)
Total	114 (7.6%)	1384 (92.4%)	1498 (100%)

Is smoking related to the risk of marital separation over 3 years?
(Data from Australia)

A “simpler” 2x2 table of these data

- Collapse the 2 bottom categories into “either partner smoked”

Either partner smoked:	Marital Status after 3 yrs		Row Total
	Separated	Not Separated	
No	41	931	972
Yes	83	453	536
Col total	124	1384	1508

Row conditional probabilities

Either partner smoked:	Marital Status after 3 yrs		Row Total
	Separated	Not Separated	
No	41	931	972
Yes	83	453	536
Col total	124	1384	1508

Either partner smoked:	Marital Status after 3 yrs		Row Total
	Separated	Not Separated	
No	4.2%	95.8%	100%
Yes	15.5%	84.5%	100%
Col total	8.2%	91.8%	100%

Divide each cell by the row total

Risk of separation:

- Unconditional (marginal) probability 8.2%
- Conditional probability
  - if neither partner smokes 4.2%
  - if either partner smokes 15.5%

Column conditional probabilities

Either partner smoked:	Marital Status after 3 yrs		Row Total
	Separated	Not Separated	
No	41	931	972
Yes	83	453	536
Col total	124	1384	1508

Either partner smoked:	Marital Status after 3 yrs		Row Total
	Separated	Not Separated	
No	33.1%	67.3%	64.5%
Yes	66.9%	32.7%	35.5%
Col total	100%	100%	100%

Divide each cell by the column total

Risk of smoking:

- Unconditional (marginal) probability 33.5%
- Conditional probability
  - if they separated 66.9%
  - if they did not 32.7%

Next: Relative risk

- Consider the following two questions:
  - Are couples who smoke more likely to get separated?
  - Are couples who separate more likely to have smoked?
- How would you summarize these 2 relative risks?
- Will that summary have the same value for both?