Handout 1

Examples of Categorical Data

Let us begin with a few simple examples about variables.

Example 1. A study was done to study the distribution of educational level (years of schooling completed) of residents in a large city who are over 40 years old. A random sample of 200 individuals (over 40 years of age) was taken and their educational levels were recorded. [The variable is educational level.]

Example 2. Another analyst used the same data as in Example 1, and summarized the data into two categories - those with a high school degree (no college degree) and those with a college degree. It turns out that in the sample 71 had a college degree. [Years of schooling has been has now been binned into two categories]

Example 3. Yet another analyst binned the data in Example 1 into three groups, and the counts are given in the following table

Educational level	less than high school	high school	college
Count	38	91	71

In this case, years of schooling has been binned into three categories.

Example 4. In a large metropolitan area in California, a random sample of 400 adult residents was taken, and the counts are given below.

Race/Ethnicity	Anglo	Hispanic	Afro-American	Others
Count	87	117	101	95

Here the variable 'ethnicity' has 4 categories. We may be interested in testing if the four groups are equally present in this metropolitan area (ie, each is 25%).

Discussion

Note that in Example 1, the variable 'years of schooling' is quantitative, whereas in Examples 2 and 3 the variable 'educational level' is qualitative. In Example 1, we may use a plot such as histogram in order to examine the distribution of years of schooling. In Example 2, we have a binary variable (or a dummy or indicator variable) as it has only two categories. In Example 3, the variable 'educational level' is qualitative with three categories, and there is an ordering of the categories. In Example 4, the variable 'ethnicity' has four categories, but there is no ordering in the categories. **Categorical Data Analysis deals with data where the variables are qualitative (ie, categorical). This course is concerned with concepts, methods and analysis of those data sets where the variables are categorical.**

Here are some technical names for the qualitative variables listed above:

(i) Binary (Example 2).

(ii) Nominal (Example 4), categories are not ordered.

(iii) Ordinal (Example 3, categories are ordered). Note that binning a quantitative variable leads to an ordinal qualitative variable. Statistical methods developed for ordinal variables should not be used for nominal variables.

More Examples for future discussion

Before we get into some necessary technical details, let us look at a few more examples which we will deal with later in the course.

Example 5. In order to investigate association between smoking and lung cancer, 125 lung cancer patients (thought of as a random sample of lung cancer patients) and 150 controls (thought of as a random sample of without lung cancer) were taken. Both the samples were taken from the same large metropolitan area. The following table provides a summary of the counts. [Note that this data is from 1953]

	Smoking Habit		Total
	Smoker	Nonsmoker	
Cancer	120	5	125
Cancer-free	126	24	150
Total	246	29	275

Note we have two independent samples from the two populations (cancer patients, and cancer-free residents), and the categorical variable is smoking habit (smoker and non-smoker). We may be interested in finding out the difference in the proportion of smokers in cancer and control groups. Note that the researcher conducting the study decided what the sample sizes from the two populations would be, and thus the row totals are known in advance. [Technical note: this is a case of independent Binomial samples (also called product Binomial samples).]

Example 6. Is political ideology dependent on opinion on death penalty for murders? A random sample of 250 adults is taken in a state and the summary of counts is given below.

	Ideology		Total
Supports Death penalty	Liberal	Conservative	
Yes	37	51	88
No	76	86	162
Total	113	137	250

Note that there are two qualitative variables: ideology and opinion on death penalty. It is of interest to test if ideology is independent of opinion on death penalty. Note that in this study, each person in the sample has been asked her/his political ideology and opinion on death penalty, thus the row totals (and column totals) were not known in advance. This is in contrast with Example 5, where the row totals were known in advance. [Technical note: this can be considered a case of Multinomial sampling scheme with four categories: Yes& Liberal, Yes&Conservative, No&Liberal, and No&Conservative.]

Example 7. We have the data on the verdicts in cases of convicted murderers in a certain state since 1980.

	Death Penalty		Total
Race of Victim	Yes	No	
White	45	85	130
Black	14	218	232
Total	59	303	362

We have two qualitative variables each with two categories: Victim's race and pronouncement of death penalty. It is of interest to investigate if the rate of death penalty higher for white victims than for blacks. Note that unlike in Examples 5 and 6, neither the sample size nor the row and column totals are not decided

by the researcher and thus may be considered random. [Technical note: this can be considered a case of Poisson sampling.]

Remark 1: In Example 5, we may want to test if the proportions of smokers in the cancer and cancer-free groups are the same (this is known as a test for homogeneity). In Example 6, we may want to test if political ideolgy is independent of opinion on death penalty (known as a test of independence). In Example 7, we may want to test if the rate of death penalty verdict the same across victim's racial status. Though the sampling schemes are different in these three examples, the procedure for testing these hypotheses turn out to be the same. We will discuss the inferential issues later in the course.

Example 8. Is the highest level of schooling (graduated from university, graduated from high school or neither) of young adults the same for parents' of different socioeconomic status? In order to investigae this, an educationist took three random samples: 50 students of wealthy parents, 75 students of middle-class parents and 50 students of poor parents.

	Level of Schooling			Total
Parents' status	University	High School	None	
Wealthy	22	17	11	50
Middle-class	35	22	18	75
Poor	9	16	25	50
Total	66	55	54	175

The variable here is level of schooling with three categories. Thus we have three populations (corresponding to different socioeconomic status of parents), and we have independent samples from these three populations. The educationist wanted to know if the level of schooling the same for the three populations. Here there is one qualitative variable with three categories: level of schooling. Note that the row totals are determined by the educationist, and thus are known in advance. [Technical note: this is a case of three independent multinomial samples (also called product multinomial sampling).]

Example 9. A survey is conducted of 175 young adults whose parents are classified either as wealthy, middle class or poor to determine their highest level of schooling (graduated from university, graduated from high school or neither). Based on the data collected, can we state that students' level of schooling is independent of their parents' wealth?

	Level of Schooling			Total
Parents' status	University	High School	None	
Wealthy	20	15	10	45
Middle-class	40	25	20	85
Poor	8	14	23	45
Total	68	54	53	175

Note that there are two qualitative variables (parents' status and schooling level of young adults) each with three categories. Note that neither the row totals, nor the column totals are known in advances. [Technical note: this is a case of Multinomial sampling scheme with 9 categories Wealthy&University, Wealthy&High School,...,Poor&None.]

Remark 2: In Example 8, the educationist wanted to test if the level of schooling the same for the three populations (also known as a test for homegeneity). In Example 9, we may want to test if the level of

schooling of young adults independent of parents' status (also known as test of independence). Though the sampling scemes in Examples 8 and 9 are different, the procedure for testing these hypotheses testing turn out to be the same.

Example 10. Suppose that we are working with some doctors on heart attack patients. The dependent variable is whether the patient has had a second heart attack within 1 year (yes = 1). We have two independent variables, one is whether the patient completed a treatment consistent of anger control practices (yes=1). The other is a score on a trait anxiety scale (a higher score means more anxious).

Person	2^{nd} heart attack	Treatment of anger	Trait Anxiety
1	1	1	70
2	1	1	80
3	1	1	50
4	1	0	60
5	1	0	40
6	1	0	65
7	1	0	75
8	1	0	80
9	1	0	70
10	1	0	60
11	0	1	65
12	0	1	50
13	0	1	45
14	0	1	35
15	0	1	40
16	0	1	50
17	0	0	55
18	0	0	45
19	0	0	50
20	0	0	60

A goal of the analysis would be to investigate (and model) how the probability of a second heart attack within a year of the first attack depends on trait anxiety and treatment of anger. Note that the response (dependent) variable is qualitative (binary), and the independent variables are treatment of anger (binary) and trait anxiety (quantitative). We will discuss this using what is known as a logistic regression model.

Example 11. (Effect of anti-epilepsy drug) In order to study the anti-epilepsy drug progabide, researchers randomly assigned 59 patients suffering from epileptic seizures to receive either progabide or a placebo, in addition to standard chemotherapy. Data below lists the number of epileptic seizures in the 8 weeks prior to administration of the treatment, the number of seizures in 8 weeks after the start of the treatment, (control=0, progabide=1), and the age (in years) of each patient.

Treatment	Age	Pretreatment count	Posttreatment count
0	31	11	14
0	30	11	14
0	25	6	11
0	36	8	13
⋮	⋮	⋮	⋮
1	26	22	51
1	21	25	6
1	36	13	0
1	37	12	10

Here the main goal is to check if progabide is effective in reducing the incidence of epileptic seizures. This data will be analyzed using a Poisson regression model.