Introduction to Statistical Ideas and Methods

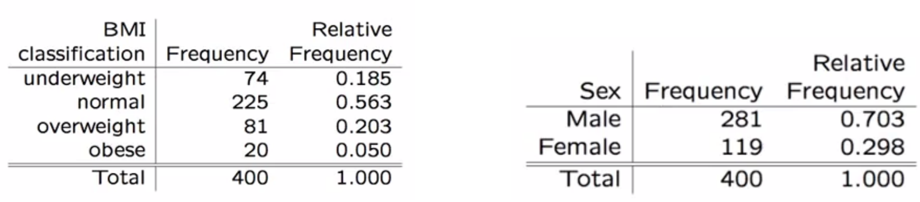
Summarizing Data: Relationships Between Variables

Examining Relationships Between Two Categorical Variables

Recall that the **distribution** of a variable is the pattern of values in the data for that variable, showing the frequency of the occurrence of the values relative to each other. For **categorical variables**, the distribution is given by the *frequencies* or *relative frequencies* of the observations for each of the categories of the variable.

Example 1

In an earlier section for the anthropology data of measurements on 400 skeletons, we saw the distribution of mass, or BMI classification and sex, shown below in Figure 1.



BMI classification	Frequency	Relative Frequency
underweight	74	0.185
normal	225	0.563
overweight	81	0.203
obese	20	0.050
Total	400	1.000

(a) BMI classification

Sex	Frequency	Relative Frequency
Male	281	0.703
Female	119	0.298
Total	400	1.000

(b) Sex

Figure 1: Distributions of categorical variables for skeleton data

Now we are interested in looking at these two categorical variables together. Our anthropologist is interested in learning about how the error in age estimation is associated with body mass index. But it is important to also consider the effect of sex here. If the error in age estimation also differs with sex, it will be important to understand if the body mass index classification differs with sex for these observations.

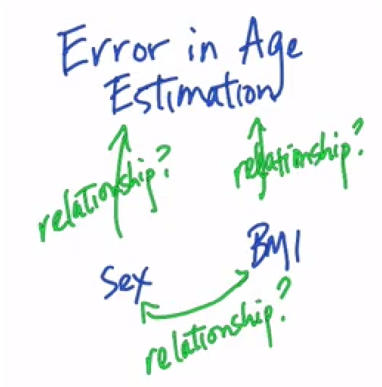
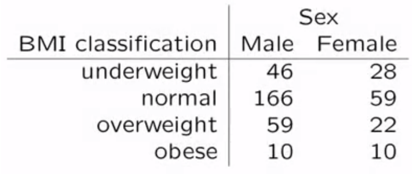


Figure 2: Possible relationships between the categorical variables in the skeleton data

By investigating the joint distribution of body mass index classification and sex, we can learn things such as:

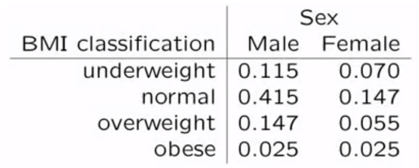
- Do we have equal numbers of females and males who are obese?
- Are there equal numbers of males and females in the underweight category?

The **joint distribution** of two categorical variables can be seen in a contingency table, sometimes called a cross tabulation, or a two way table. In the **contingency table** we classify our 400 skeletons two ways, by BMI classification and by sex. The table contains the counts (Figure 3(a)) or percentages (Figure 3(b)) of the number of observed values for males for each of the BMI classifications, and for females for each of the BMI classifications.



	Sex	
BMI classification	Male	Female
underweight	46	28
normal	166	59
overweight	59	22
obese	10	10

(a) Contingency table displaying frequencies



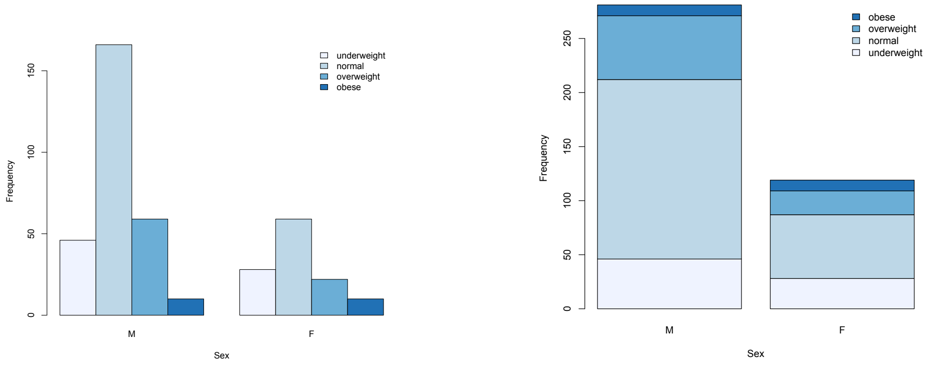
	Sex	
BMI classification	Male	Female
underweight	0.115	0.070
normal	0.415	0.147
overweight	0.147	0.055
obese	0.025	0.025

(b) Contingency table displaying relative frequencies

Figure 3: Joint distribution of BMI classification and sex for skeleton data

We can see from the tables that 46 or approximately 12% of the 400 skeletons are underweight males and 28, or about 7%, are underweight females.

For a graphical display of the joint distribution we can plot the frequencies in either side-by-side or stacked bar plots. Figure 4 has the height of each bar as the number of skeletons in each body mass classification for each sex.



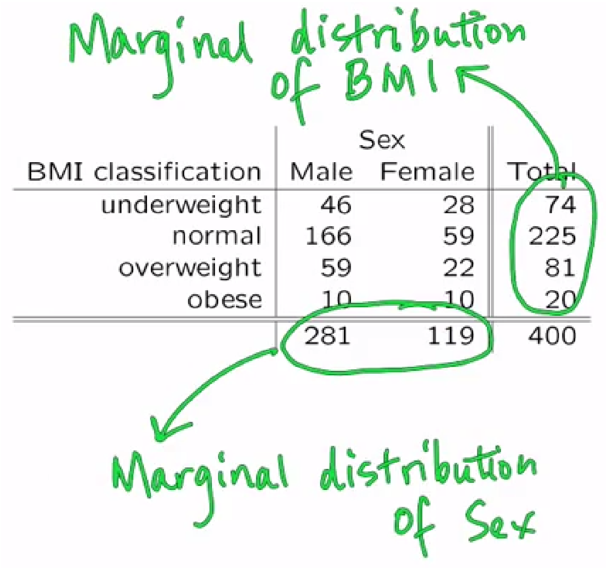
(a) Side-by-side boxplots

(b) Stacked bar plots

Figure 4: Joint distribution of BMI classification and sex for skeleton data

From the side-by-side plot it is clear that there are many more male skeletons with a normal BMI than skeletons in any other category. For female skeletons, we can also see that normal BMI is also the most common of the classifications. However, by looking at the total counts of the bars for males in the stacked bar plot, it is clear that there are more than twice as many males as females. And although the normal BMI bar is taller for males than it is for females, it is difficult to judge from these plots if a greater fraction or proportion of the males tend to have normal BMI than those of females. We need to do some more work to make a fair comparison.

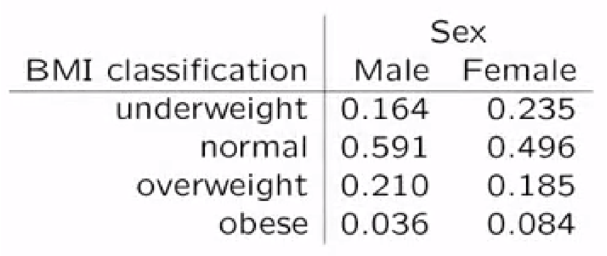
The **marginal distribution** of a categorical variable is the distribution of only one of the variables in a contingency table. We can see it in the margins of the table by taking the row or column totals. In Figure 5 below, we see the marginal distributions of BMI and sex.


Marginal distribution of BMI
Marginal distribution of Sex

	Sex		
BMI classification	Male	Female	Total
underweight	46	28	74
normal	166	59	225
overweight	59	22	81
obese	10	10	20
	281	119	400

Figure 5: Marginal distributions of BMI classification and sex

The **conditional distribution** of a categorical variable is its distribution within a fixed value of a second variable. The conditional distribution of BMI classification given sex, shown in Figure 6 below, will help us understand whether the BMI classification is the same for both sexes.



	Sex	
BMI classification	Male	Female
underweight	0.164	0.235
normal	0.591	0.496
overweight	0.210	0.185
obese	0.036	0.084

Figure 6: Conditional distributions of BMI classification given sex

Given that the skeleton is male, the conditional distribution is the distribution of BMI classification just for males, and similarly for females. The relevant quantity that we need for the conditional distribution, and we can calculate it from the contingency table of counts in Figure 5, is the column percentage.

For example, 16.4% of male skeletons are underweight

$$16.4\% = \frac{46}{281}$$

and 23.5% of female skeletons are underweight

$$23.5\% = \frac{28}{119}.$$

Note that for both males and females, the conditional distribution proportions sums to 1.

Graphically, we can compare the conditional distributions of BMI classification given sex, by plotting the column percentages in stacked bar plots like in Figure 7 below.

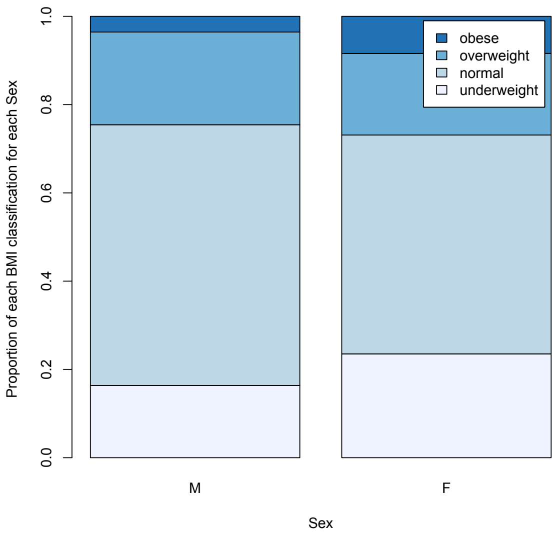


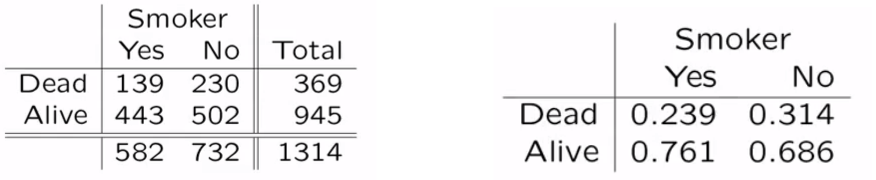
Figure 7: Stacked bar plots of the conditional distributions of BMI classification given sex. This type of bar plot is often called a segmented bar plot.

From these plots, we can see that the proportions of underweight and obese skeletons are higher in females than males. Also, the proportion of normal weight skeletons is higher for males than females.

Two variables in a contingency table are **independent** if the conditional distribution of one variable is the same for all values of the other variable. As we have noted, the distributions of BMI classifications seem to differ between males and females; it seems that BMI classification and sex are not independent for these skeletons.

Example 2

Let's look at one more example from a report on the findings from a 20 year follow-up of a large scale study of thyroid and heart disease carried on in England in the mid 1970s. We are working with a subset of the data containing measurements on 1,314 women who were classified at the beginning of the study as current smokers or having never smoked. We are interested in the 20 year survival status for these women.



	Smoker		
	Yes	No	Total
Dead	139	230	369
Alive	443	502	945
	582	732	1314

(a) Joint distribution of survival status and smoking status

	Smoker	
	Yes	No
Dead	0.239	0.314
Alive	0.761	0.686

(b) Conditional distribution of survival status given smoking status

Figure 8: Tables for thyroid and heart disease 20 year study

Looking at the contingency table for these data, the column proportions in Figure 8 tell an interesting story. Of the smokers, only 24% had died but of the non-smokers 31% had died. Does this study show that smoking might lead to a greater chance of surviving 20 years? Of course there's a twist here!

Let's look at the column proportions for the tables of smoking and survival status broken down by age grouping. Although age is a quantitative variable, it is sometimes given in groups to illustrate a point. As we can see from the side-by-side bar chart in Figure 9, for all age groups except the 25 to 34 year olds, the death rate is higher in the group of smokers than in the group of non-smokers. How did this happen?

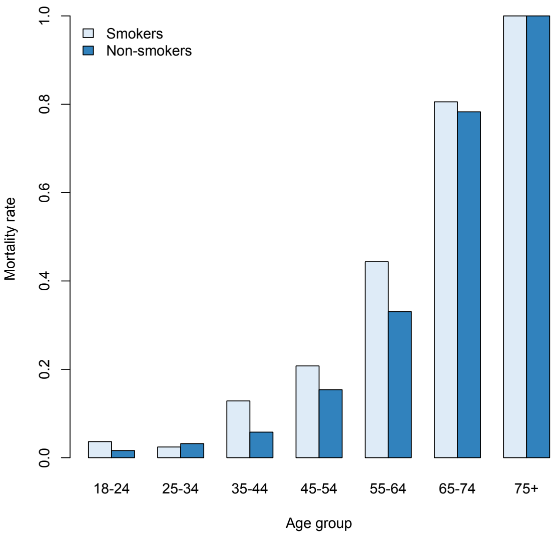


Figure 9: Side-by-side bar charts of mortality rate by smoking status, for each age group

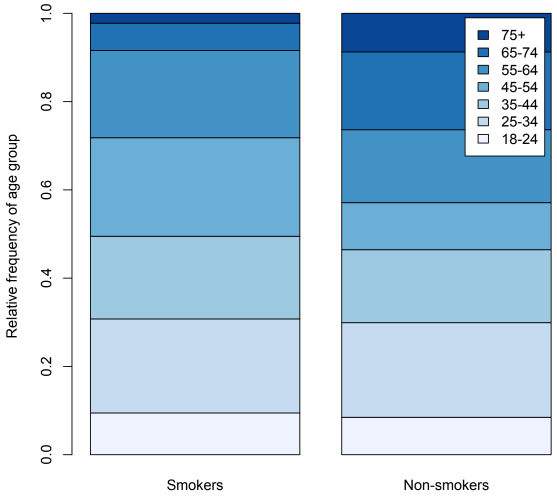


Figure 10: Stacked bar charts of age group by smoking status

Age is related to both smoking status and survival. The stacked bar chart in Figure **??** shows the distributions for smokers and non-smokers. The non-smoking population includes more older women; when the study started, few of the women over age 65 were smokers. But, of course many of them, since they were at least 65 at the start, had passed away by the end of the 20 year follow-up period. Moreover, this study could potentially underestimate the harmful effects of smoking, since the observed small percentage of older smokers could have happened because smokers tend not to survive to age 65.

This is an example of **Simpson's Paradox**, in which conditional distributions within subgroups can differ from conditional distributions for combined observations. Age here is a *lurking* variable. We need to always watch for lurking variables which, if taken into account in our analyses, might affect our conclusions. In some upcoming lectures, we will talk about data collection and how to design a study to mitigate the effects of lurking variables.