ONE CATEGORICAL VARIABLE

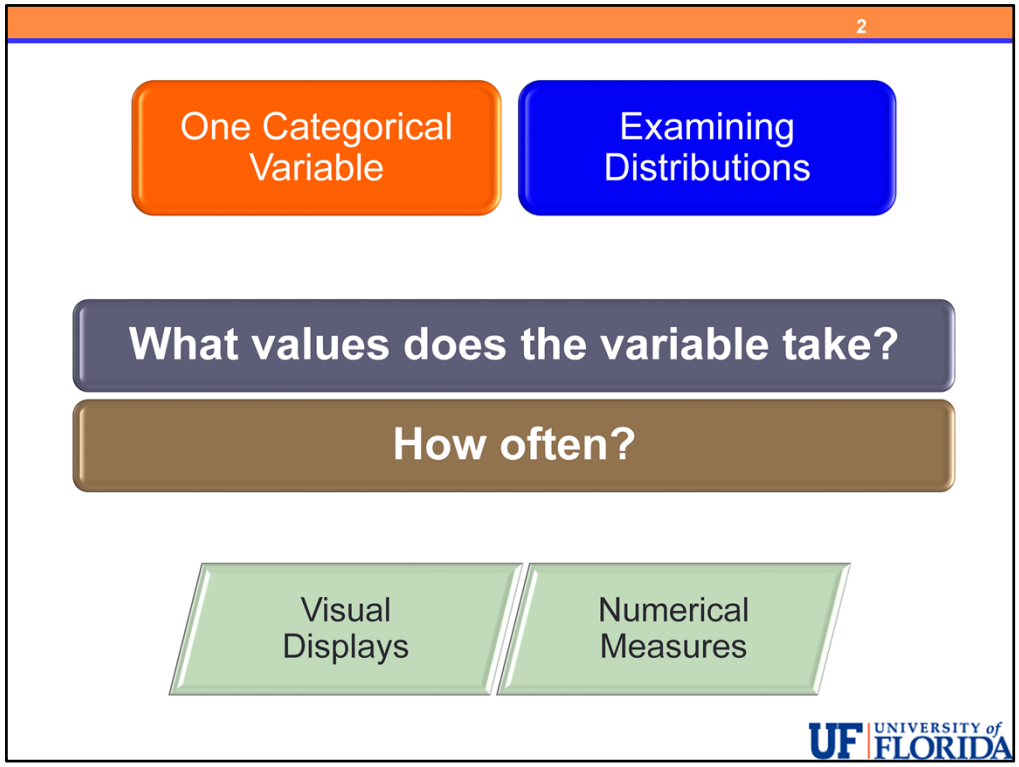
Describing Distributions

Unit 1: Exploratory Data Analysis

UF UNIVERSITY of FLORIDA

How do we summarize one categorical variable?

What visual displays and numerical measures are appropriate?



Recall that, when we say Distribution we mean

What values the variable can take

And

How often the variable takes those values.

Exploratory Data Analysis for one categorical variable is very simple since both the components are simple! And … you have most likely seen these before!!

For visual displays we typically use a bar chart or pie chart or similar variation to display the results for the variable in a graphical form.

For numerical measures we simply provide a table, called a frequency distribution, which gives the possible values along with the frequency and percentage for each value.

Example: Framingham Data

Obs	RANDID	SEX	AGE	BMI	DIABETES	bmicat
1	6238	Female	58	28.5	No	Overweight
2	11263	Female	55	31.17	Yes	Obese
3	12806	Female	57	22.02	No	Normal
4	14367	Male	64	25.72	No	Overweight
5	16365	Male	55	29.11	No	Overweight
6	23727	Female	53	26.62	No	Overweight
7	24721	Female	51	24.77	No	Normal
8	33077	Male	60	22.96	No	Normal
9	34689	Female	49	31.45	No	Obese
10	36459	Male	53	26.43	No	Overweight

http://www.framinghamheartstudy.org/

UF UNIVERSITY of FLORIDA

Here are a few variables available in a subset of the Framingham data.

We have a random id number for each individual along with the individual's
gender (categorical)
age (quantitative)
body mass index measurement (quantitative)
diabetes status – yes or no (categorical)
and
body mass index groups or categories:
underweight, normal, overweight, obese (categorical)

{Some information about the study can be found at:

http://www.framinghamheartstudy.org/

If you have trouble using the link above, copy and paste the URL above into your browser.}

Example: Framingham Data

SEX				
SEX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Male	1001	43.41	1001	43.41
Female	1305	56.59	2306	100.00

Diabetic Y/N				
DIABETES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	2142	92.89	2142	92.89
Yes	164	7.11	2306	100.00

http://www.framinghamheartstudy.org/

UF UNIVERSITY of FLORIDA

The frequency distributions (using SAS statistical software) for the two binary categorical variables gender and diabetes status are shown.

Most software packages give both the frequency and percentage. In this case we also obtain the cumulative frequency and cumulative percentage. This can be useful for ordinal categorical variables to quickly summarize the percentage greater or less than a certain category.

Notice that for these variables, providing the reader with either the frequency for each value or the total and the frequency or percentage for one value is enough to give a complete summary of the information available here. Often, variables this simple would be summarized numerically in the discussion.

For example, here we could say:

Among the 2306 individuals in the sample, 1001 (43.4%) were male and 164 (7.1%) were classified as diabetic.

We can then know that 1305 or 56.6% were female and 2142 or 92.9% were not diabetic.

Bar charts and pie charts can be displayed for these variables, however, often this would be considered a waste of valuable page space in a journal article.

{Some information about the study can be found at http://www.framinghamheartstudy.org/}

Example: Framingham Data

BMI Category				
bmicat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Underweight	33	1.43	33	1.43
Normal	1013	43.93	1046	45.36
Overweight	962	41.72	2008	87.08
Obese	298	12.92	2306	100.00

UF UNIVERSITY of FLORIDA

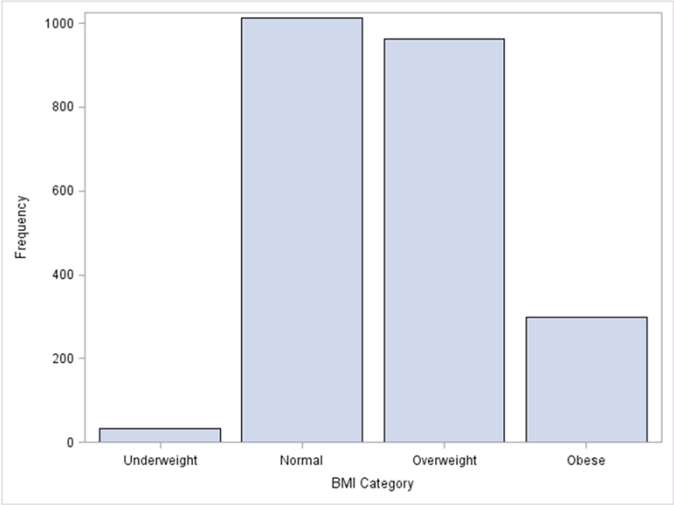
Here we have the summary of the categorized version of BMI which is an ordinal categorical variable with four levels.

To summarize the distribution of this variable we might say:

Of the 2306 individuals in the sample, individuals with normal BMI comprised the largest group with 43.9%, followed closely by the overweight group with 41.7% with obese individuals representing 12.9% of the sample. Only 33 individuals (1.4%) were classified as underweight.

Here the cumulative percent column would allow us to quickly find that approximately 87% of individuals are not obese or subtracting the cumulative percentage for normal (45.4%) from 100%, we can see that approximately 55% of individuals have a BMI above normal.

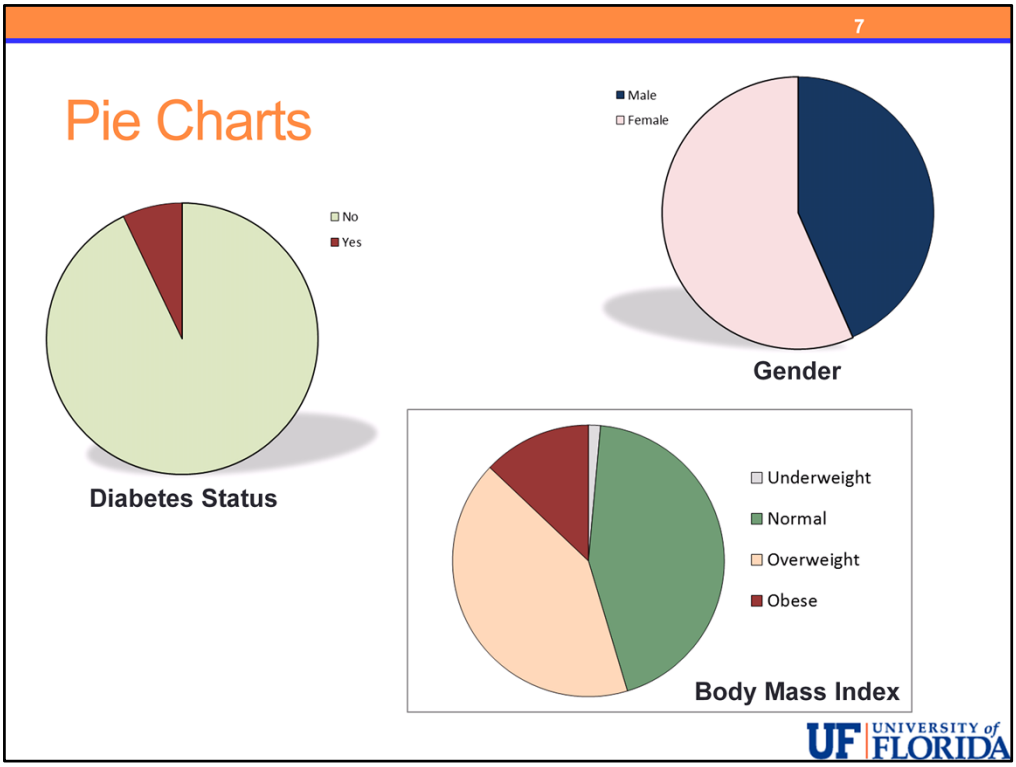
Example: Framingham Data



UF UNIVERSITY of FLORIDA

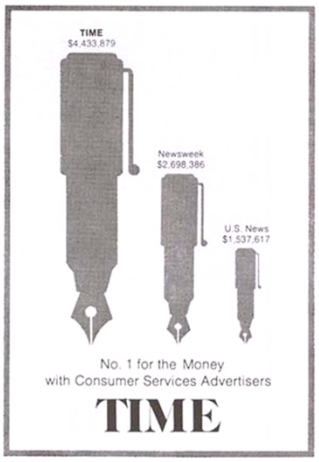
Here we have a bar chart for the BMI categories which gives the frequency on the vertical axis. This could also be provided as a percentage.

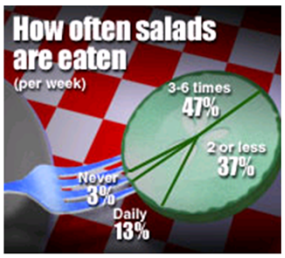
A pie chart could be used, however, as the variable is ordinal, it is best to use a display which allows the categories to be displayed in order. With a pie chart, the starting point is not clearly defined and so the ordinal nature can be lost in such a graph.



Although pie charts are often less useful than the tables used to create them, here are the pie charts for each of the categorical variables we have summarized from the Framingham data.

Pictograms





UF UNIVERSITY of FLORIDA

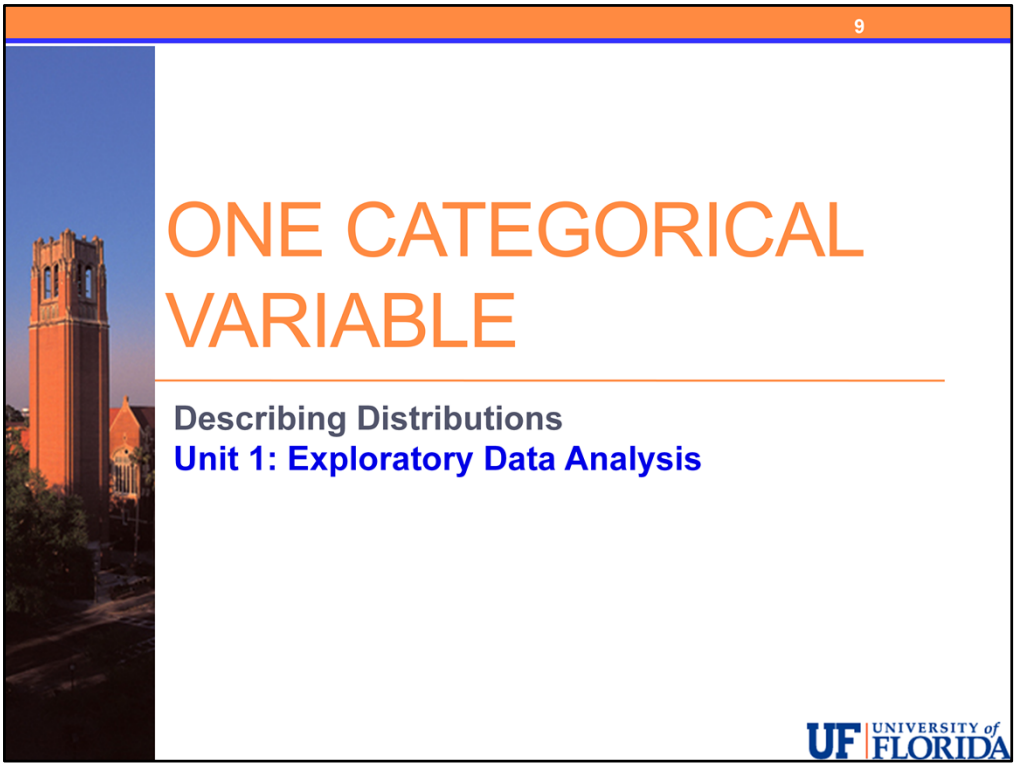
Variations on bar charts and pie charts such as these pictograms are often used.

Be very careful when using pictograms to avoid any distortion in the data.

The graphs on the right reflect the data clearly in a non distorted way.

However the one for TIME on the left overemphasizes the difference by adjusting both the height and width of the “bars” used to display the data. This tends to result in a visual comparison of the volume of the “bars” instead of only the height, which is a distortion of the data.

Think critically about any information displayed in graphs and be careful how you graphically display your own data!



Exploratory data analysis for one categorical variable includes a frequency distribution as the numerical measures and a bar chart or pie chart as a visual display (if needed).

Also remember to be able to summarize and explain the results in context.