Data 8R Summer 2017 — Tables and more Visualizations

Discussion 5: July 11, 2017

1 Review

Write down the definitions of the following terms in your own words:

RCT (Randomized Control Trial)

Observational Study

Confounding factor

Observational Study or RCT:

A researcher at a hospital decides to study a group of students from Berkeley. He/she decides to record data for each subject on the amount of exercise they do for an entire year. At the same time, the researcher also records the number of colds each volunteer gets.

Is this an observational study or an RCT?

If the researcher finds that people who exercise more get fewer colds, what can we say from this?

Is it a valid conclusion to claim that there is an association between exercise and fewer colds in elderly people?

Write all the subexpressions for the following expressions:

```
>>> 5 + (15 * (6 / 2))
```

```
>>> make_array(10, 15, 20).item(0)
```

Write down the outputs of the following code:

Assume that the following lines of code have already been executed:

```
>>> array1 = np.arange(1, 11)
>>> array2 = make_array(3, 5, 9, 10)
```

Write down the outputs of following line:

```
>>> np.diff(array2)
```

```
>>> array1 / sum(array1)
```

2 Bar Charts

A **frequency/probability** distribution is a distribution whose amounts have been normalized to add up to 1. In other words, a frequency distribution describes the proportions of some data. On the other hand, when a distribution does not describe the proportions, it is called a **count** distribution. In other words, this means that a **count** distribution contains data about a *count* of some things.

Answer the following questions:

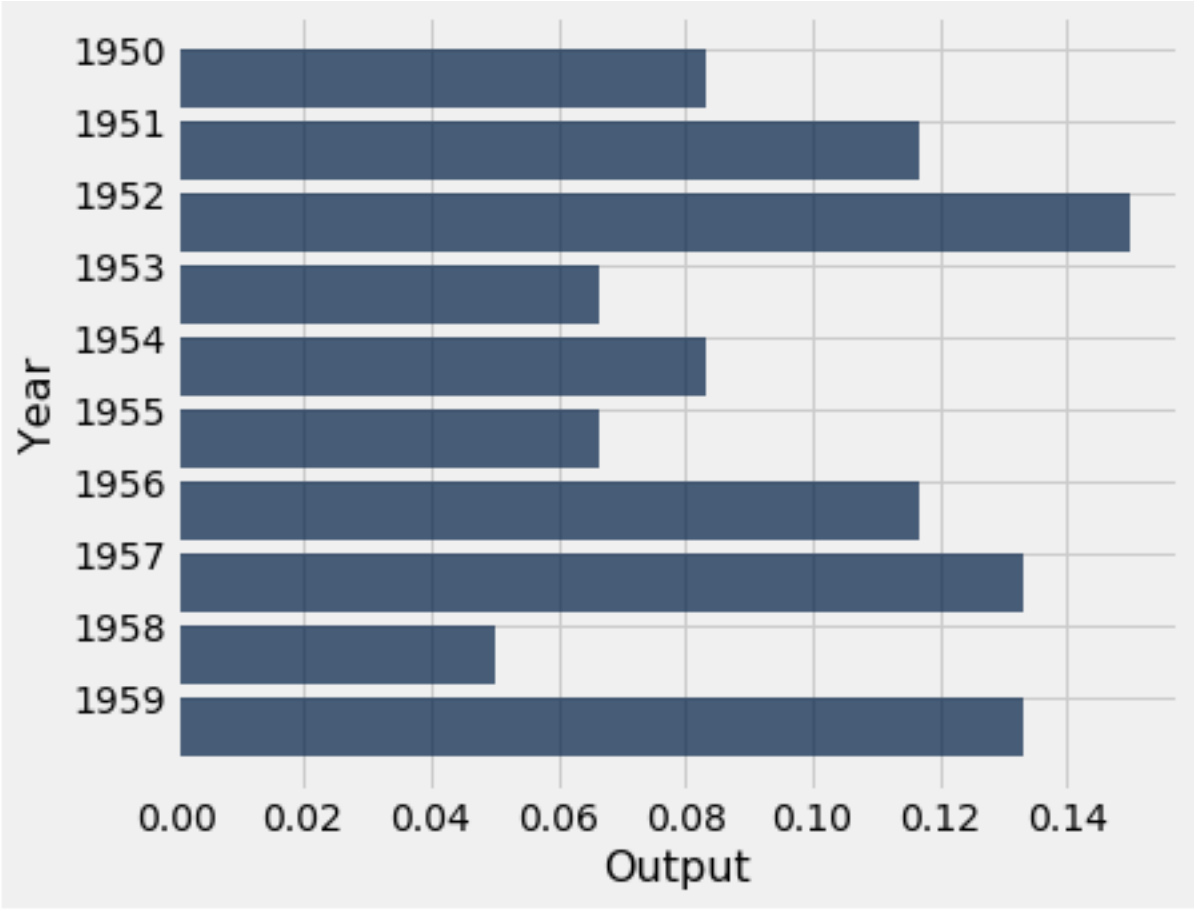
What is the difference between categorical and numerical variables?

Which variable should you use a bar chart to visualize and why?

Can bar charts be used to graph proportions?

Examine the following bar chart and answer some questions about it.

A business has graphed the proportion of outputs in each year as a bar chart.



If the business wanted to compare outputs from year to year, does this bar chart serve its purpose?

If the business wanted to compare outputs for each 2-year period, does the bar chart serve its purpose?

3 Histograms

Should you use a histogram to graph categorical or numerical variables and why?

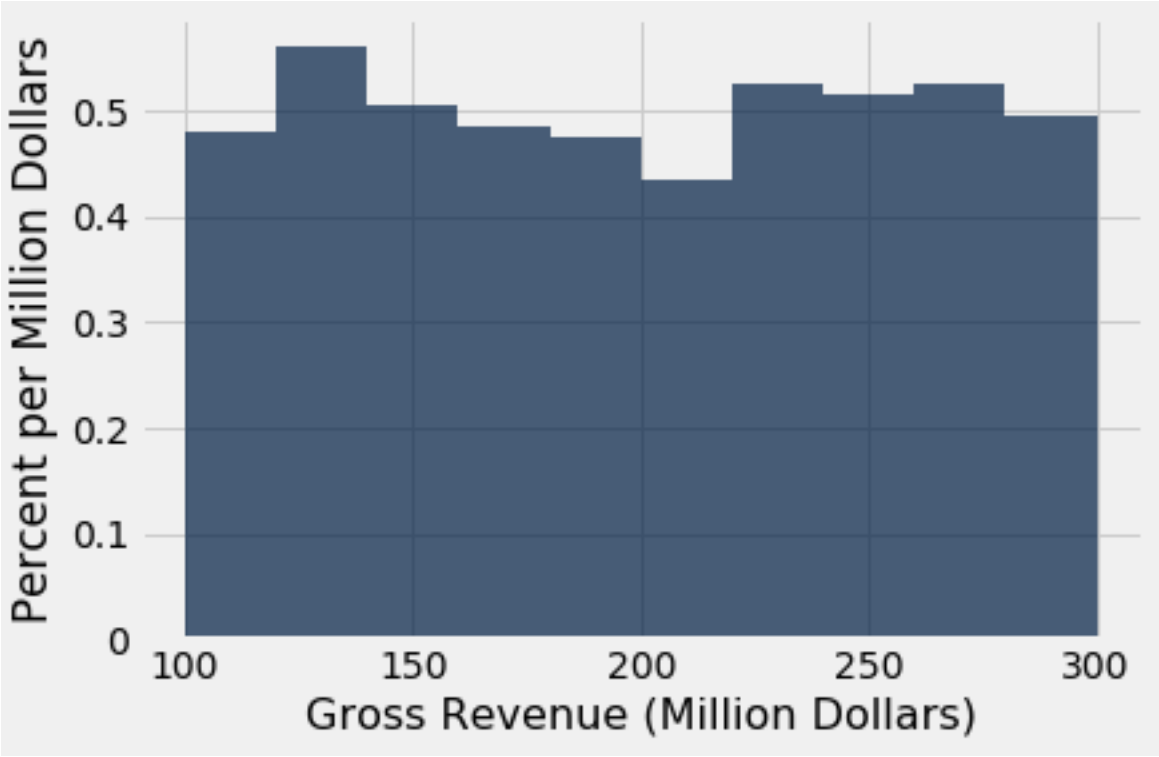
What does the width of a histogram bar represent?

What does the height of a histogram bar represent?

What does the area of a histogram bar represent?

What should the entire area of a histogram sum to (if we're using a frequency distribution)?

Suppose the same business has now made a histogram of their gross revenues:



What was the most common revenue range?

What does the height of each bar of this histogram represent?

What does the area of each bar in this histogram represent?