Lecture 3
Quantitative Variables

Dennis Sun
Stanford University
DATASCI / STATS 112

January 13, 2023

1. Visualizing One Quantitative Variable

2. Summarizing One Quantitative Variable

3. Recap

1 Visualizing One Quantitative Variable

2 Summarizing One Quantitative Variable

3 Recap

Visualizing One Quantitative Variable

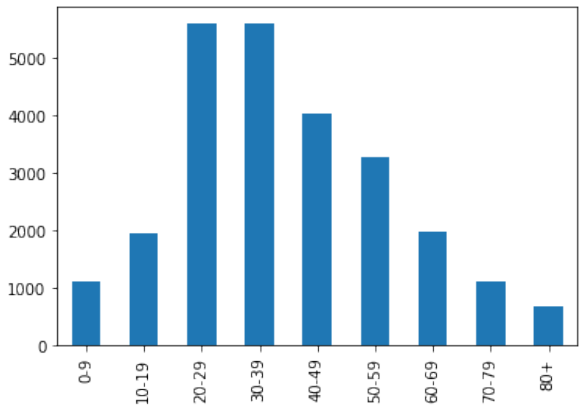
We visualize a quantitative variable using a **histogram**.

The idea of a histogram is simple:

quantitative variable → convert to categorical variable by binning → make a "bar plot"

We can make a histogram manually...

```
df_CO["age binned"] = pd.cut(
    df_CO["Edad"], bins=[0, 10, 20, 30, 40, 50, 60, 70, 80, 120],
    labels=["0-9", "10-19", "20-29", "30-39", "40-49", "50-59",
            "60-69", "70-79", "80+"], right=False)
df_CO["age binned"].value_counts().sort_index().plot.bar()
```

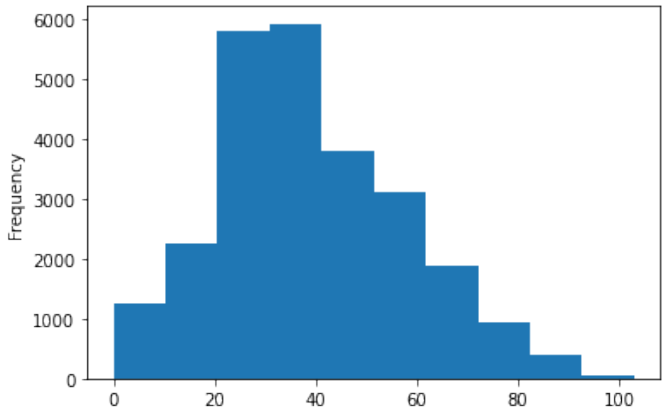


Histograms

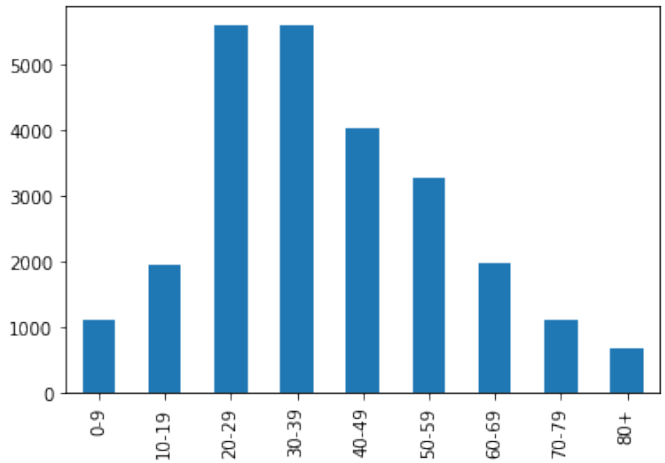
...but Pandas provides a built-in histogram method:

`Series.plot.hist()`.

```
df_CO["Edad"].plot.hist()
```



How does this differ from the manual histogram from earlier?



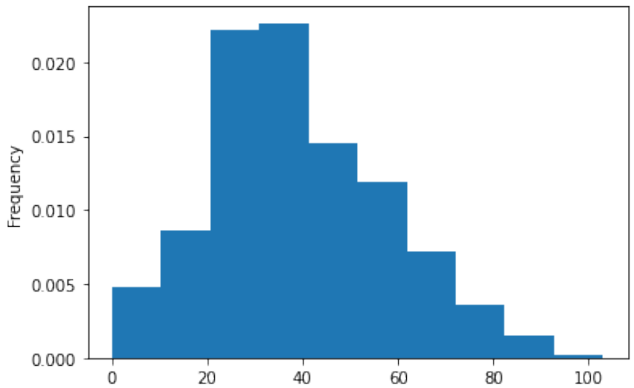
- There are no spaces between the bars.
- The x-axis is just numbers, rather than bins.

Distributions

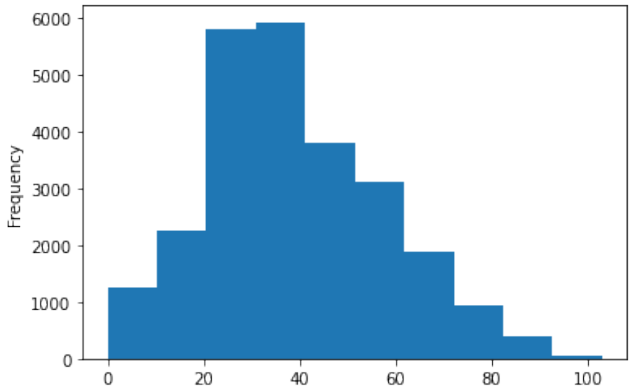
Recall how we defined the distribution of a categorical variable.

The **distribution** of a quantitative variable is similar. The counts are scaled so that the total *area* is 1.0 (or 100%).

```
df_CO["Edad"].plot.hist(density=True)
```



How does this differ from the manual histogram from earlier?



- Only the y-axis changes.
- The shape is the same!

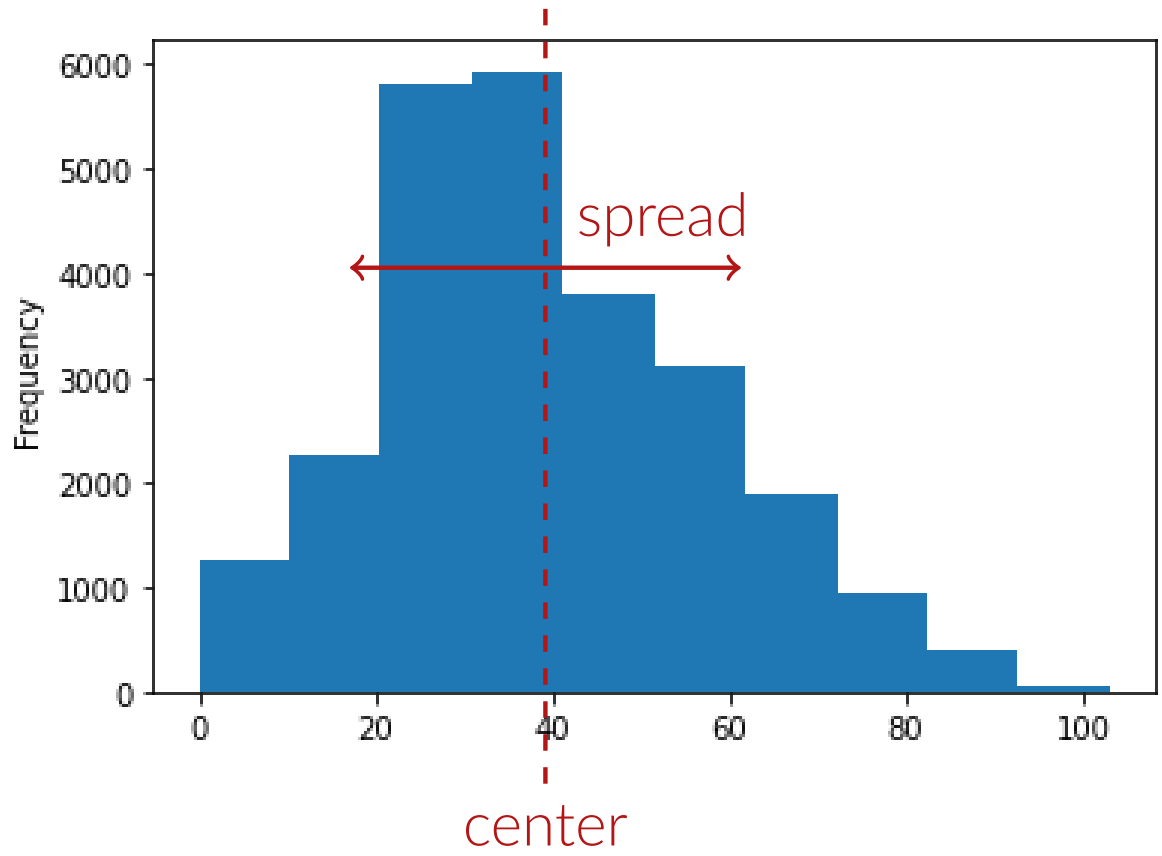
1 Visualizing One Quantitative Variable

2 Summarizing One Quantitative Variable

3 Recap

Summarizing a Quantitative Variable

The two most salient features of a quantitative variable are its **center** and its **spread**.



Summaries of Center

The **mean** of a variable X with n values is

$$\overline{\mathrm{X}} = \mathrm{mean}(\mathrm{X}) = \frac{\mathrm{sum\ of\ X}}{n}$$

You can calculate it manually...

```
df_CO["Edad"].sum() / df_CO["Edad"].count()
```

```
39.04742568792872
```

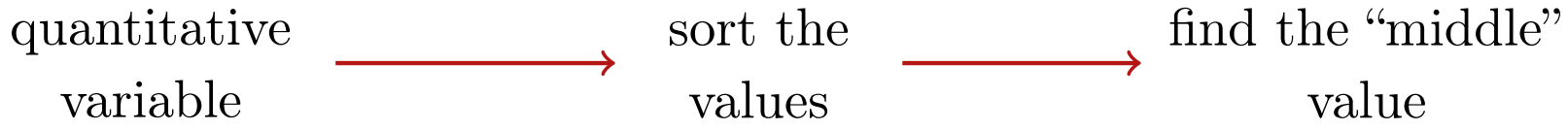
...or using a built-in Python function.

```
df_CO["Edad"].mean()
```

```
39.04742568792872
```

Summaries of Center

Another summary of center is the **median**, the "middle" of the sorted values:



You can calculate it manually...

```
def median(values):
  # assuming no missing values
  n = values.count()
  sorted_values = values.sort_values()
  if n % 2 == 1:  # if n is odd
    return sorted_values.iloc[(n + 1) // 2 - 1]
  else:  # if n is even
    return (sorted_values.iloc[n // 2 - 1] +
            sorted_values.iloc[n // 2]) / 2
```

...or using the built-in Python function.

```
df_CO["Edad"].median()
```

```
37.0
```

Summaries of Spread

The **variance** of a variable X with n values is

$$\mathrm{var}(\mathrm{X}) = \frac{\text{sum of } (\mathrm{X} - \overline{\mathrm{X}})^2}{n-1}$$

You can calculate it manually...

```
(((df_CO["Edad"] - df_CO["Edad"].mean()) ** 2).sum() /
 (df_CO["Edad"].count() - 1))
```

```
348.0870469898451
```

...or using a built-in Python function.

```
df_CO["Edad"].var()
```

```
348.0870469898451
```

What are the units? years^2

Summaries of Spread

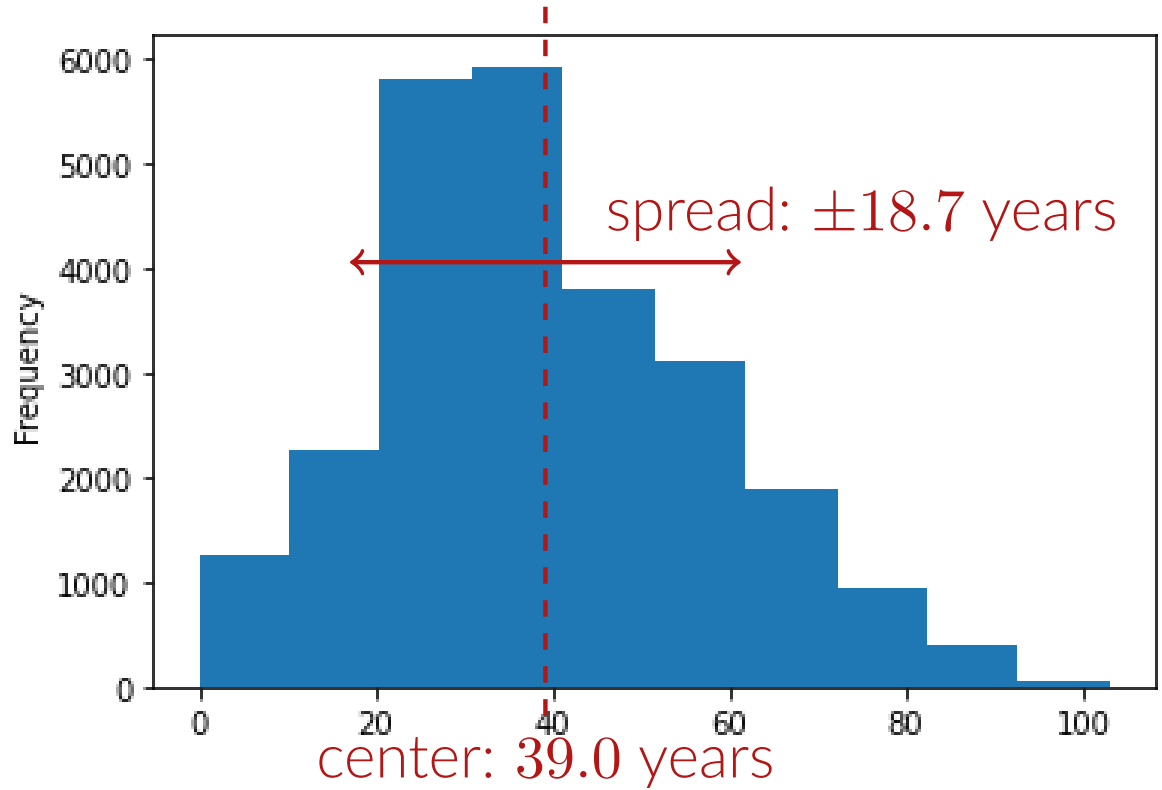
To fix the units, we take the square root to get the **standard deviation**:

$$\underset{\text{years}}{\text{sd}(X)} = \sqrt{\underset{\text{years}^2}{\text{var}(X)}}$$

You can calculate it using the built-in Pandas method `Series.std`:

```
df_CO["Edad"].std()
```

```
18.65709106452142
```



1 Visualizing One Quantitative Variable

2 Summarizing One Quantitative Variable

3 Recap

What We Learned Today

- visualizing a quantitative variable using a histogram
  - We've now seen two plots that can be made within Pandas: `.plot.bar()` and `.plot.hist()`.
- summarizing a quantitative variable
  - summarizing the center By the mean or median
  - summarizing the spread By the standard deviation
- some new Python tricks
  - `.iloc[...]` allows you to index a `Series` or `DataFrame` by position (instead of by name).
  - `//` is division but always returns an integer

Reminders

- Assignment 1 is now posted. It is due next Friday.
- I am working on Colab for Tuesday's section, should be ready by the end of the day.
- As usual, post on the Ed Discussion board if you have any questions!

No class on Monday (MLK Day).
Enjoy the long weekend, and see you on Wednesday!