When to use "standard" hyp test vs. perm test

- standard hyp test
  - may just need one column of data
  - ex.) jury Alameda Cty (TVD)
    (only need ethnicity)
  - ex.) fair coin
    (only need H or T on each flip)
    [0.5, 0.5]
    [how frequently something happens
    [0.5, 0.5]

Null: sell equal amount of fiction + nonfiction

Alt: sell more fiction than non

perm tests

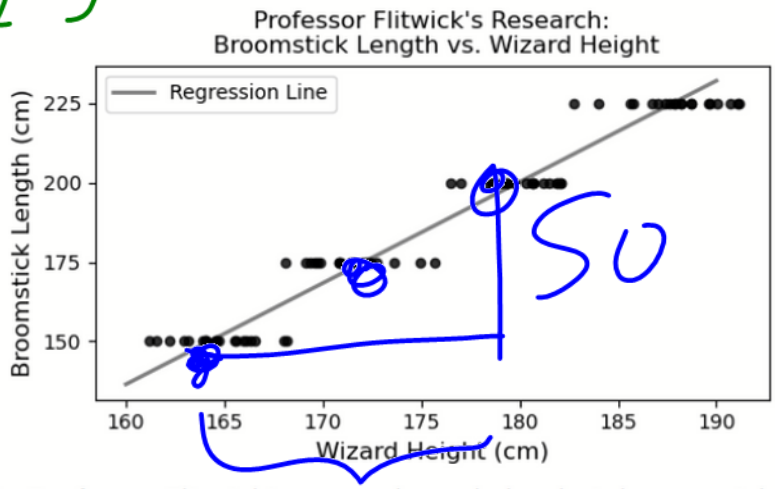
need 2 columns of data,

1 numerical	1 categorical
- baby weight	smoking/non
- pressure drops	Patriots/colts

[comparing # across categories

See winter 2021 final (last few problems)

Winter 25
#8



If we group the wizards in Professor Flitwick's research study by their broomstick length, and average the heights of the wizards in each group, we get the following results.

Broomstick Length (cm)	Average Wizard Height (cm)
150	165.0
175	172.5
200	180.0
225	187.5

It turns out that the regression line that predicts broomstick length (y) based on wizard height (x) passes through the four points representing the means of each group. For example, the first row of the DataFrame above means that $(165, 150)$ is a point on the regression line, as you can see in the scatterplot.

$(165, 150)$

$$m = \frac{50}{15} = \frac{10}{3}$$

$$m = r\frac{SD_y}{SD_x} = \frac{10}{3}$$

Problem 8.1

Based only on the fact that the regression line goes through these points, which of the following *could* represent the relationship between the standard deviation of broomstick length (y) and wizard height (x)? Select all that apply.

- [ ] $SD(y) = SD(x)$
- [ ] $SD(y) = 2 \cdot SD(x)$
- [ ] $SD(y) = 3 \cdot SD(x)$
- [ ] $SD(y) = 4 \cdot SD(x)$
- [ ] $SD(y) = 5 \cdot SD(x)$

$\rightarrow \frac{SDy}{SDx} = 1$

$\rightarrow \frac{SDy}{SDx} = 2$

$r\left(\frac{SDy}{SDx}\right) = \frac{10}{3}$

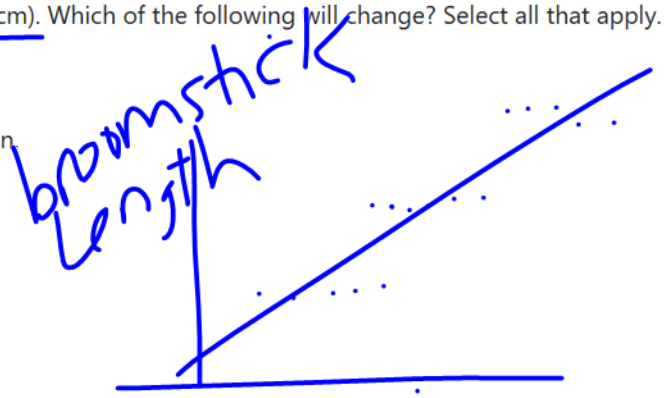
When $r = 1 \Rightarrow \frac{SDy}{SDx} = \frac{10}{3}$

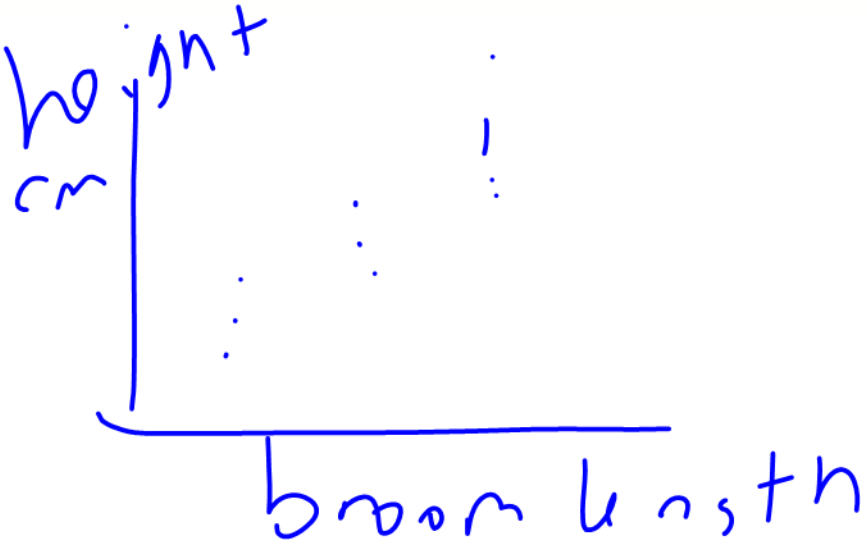
When $r < 1 \Rightarrow \frac{SDy}{SDx} > \frac{10}{3}$

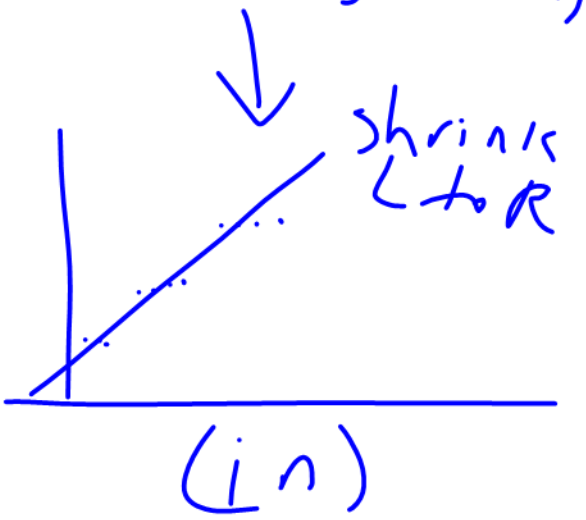
Problem 8.3

Suppose we convert **all wizard heights** from centimeters to inches (1 inch = 2.54 cm). Which of the following will change? Select all that apply.

- [ ] The standard deviation of wizard heights.
- [ ] The proportion of wizard heights within three standard deviations of the mean.
- [ ] The correlation between wizard height and broom length.
- [ ] The slope of the regression line predicting broom length from wizard height.
- [ ] The slope of the regression line predicting wizard height from broom length.
- [ ] None of the above.







$$m = \frac{r\,SD_y}{SD_x}$$

Question 5 (21 pts)

The night before the Hunger Games begins, each tribute is interviewed in front of a live audience. During this interview, the host asks each tribute a few personal questions and reveals their overall score from the training camp. These interviews are broadcast across the country, so that the residents of Panem can get to know the tributes better and form opinions about who they want to win.

The Capitol wants to understand public perceptions of the tributes after the interviews for the 74th Hunger Games. They conduct a survey of a sample of residents from all 12 districts, asking them two questions:

1. "What district do you live in?"
2. "Who do you think will win this year's Hunger Games?"

The survey results are in the DataFrame `survey`, with columns `"District"` and `"Tribute"` which contain each person's answers to the two questions above. The first few rows of `survey` are shown below.

	District	Tribute
0	7	Glimmer
1	3	Clove
2	1	Katniss
3	2	Clove
4	12	Peeta
5	10	Thresh

In this problem, we will try to estimate the proportion of residents from a given district who think a certain tribute will win the Hunger Games.

a) (4 pts) What proportion of residents in District 11 think Peeta will win? Write **one line of code** that evaluates to this proportion **in our sample**, based on the data in `survey`.

b) (5 pts) Next, we want to create a 95% confidence interval for the proportion of **all** residents from a given district who think a certain tribute will win. Fill in the blanks in the function `win_CI` below. This function takes the name of a tribute and the number of a district and returns the endpoints of a 95% bootstrapped confidence interval for the proportion of all residents of that district who think that tribute will win, based on the data in `survey`.

For example `win_ci("Peeta", 11)` returns the endpoints of a 95% confidence interval for the proportion of all residents from District 11 who think Peeta will win.

```
def win_ci(tribute, district):
    only_district = survey[survey.get("District") == district]
    props = np.array([])
    for i in np.arange(10000):
        resample = __(a)__
        tribute_count = __(b)__
        boot_prop = tribute_count / __(c)__
        props = np.append(props, boot_prop)
    return [np.percentile(props, 2.5), np.percentile(props, 97.5)]
```

(a):

(b):

(c):

c) (4 pts) Suppose we were to plot a histogram of `props` within the function `win_CI`. Which of the following best describes this histogram?

- ◯ The histogram reflects the shape of the population.
- ◯ The histogram reflects the shape of the data in `survey`.
- ◯ The histogram reflects the shape of the data in `survey` which corresponds to the given district.
- ◯ The histogram is roughly normal because of the Central Limit Theorem (CLT).
- ◯ The histogram is roughly normal, but not because of the CLT.