Lesson 28

Inference with Simple Linear Regression

Using SLR to make predictions

CSCI 3022

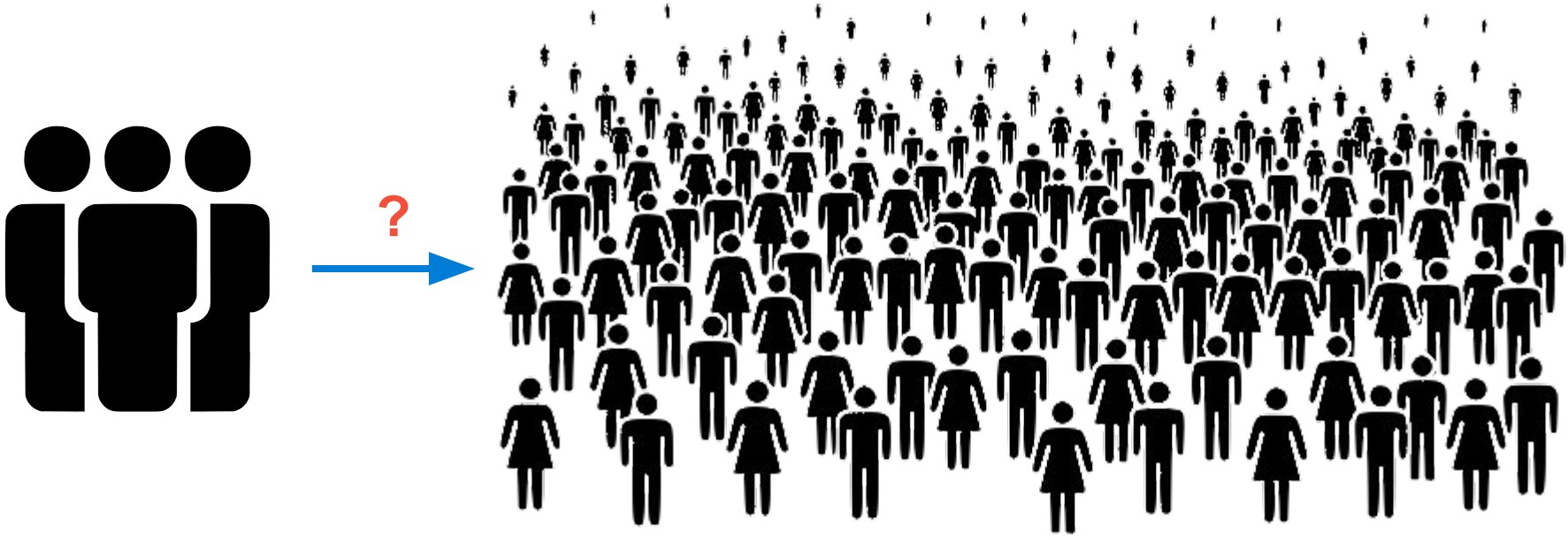
Maribeth Oscamou

Content credit: Acknowledgments

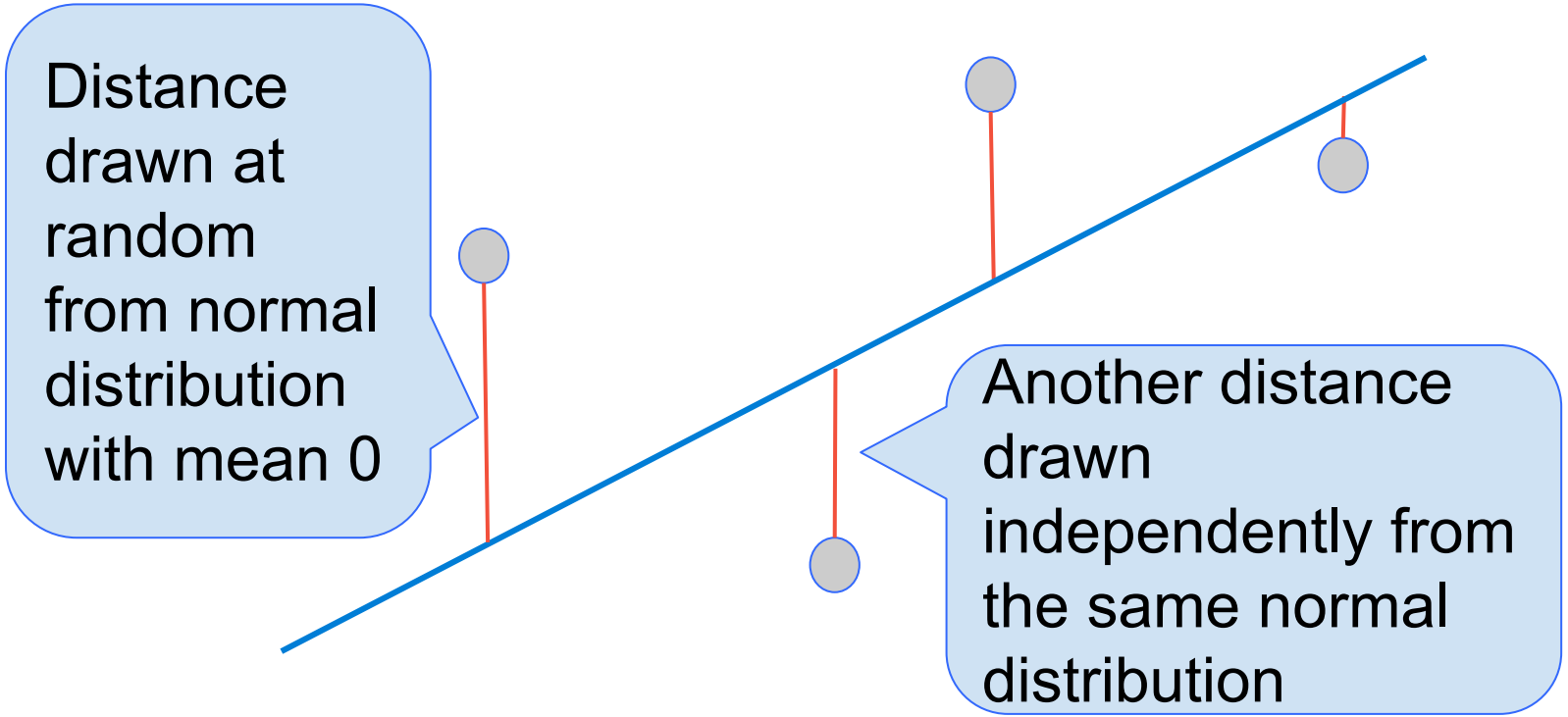
Today's Roadmap

- Regression inference:
  - Assumptions about uncertainty
  - Understanding uncertainty

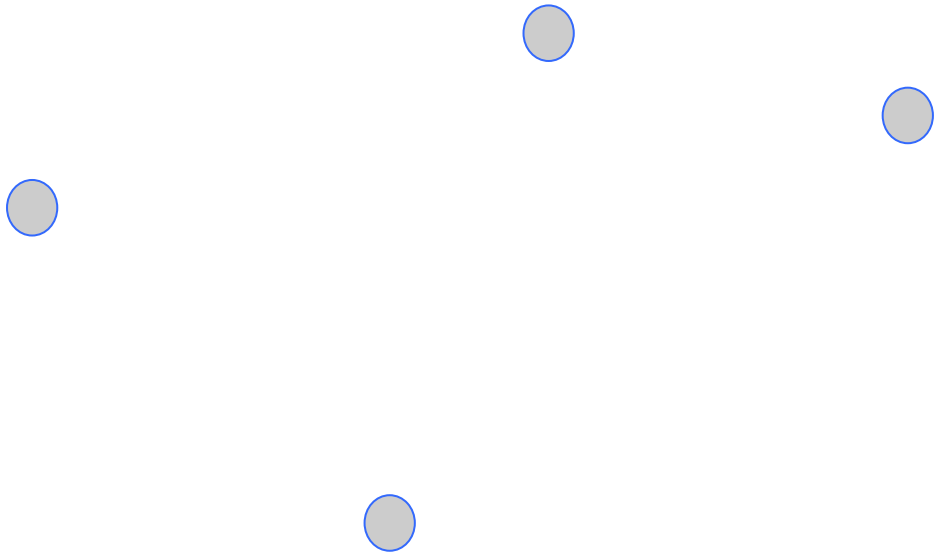
Regression Model



A “Model”: Signal + Noise



What We Get to See



(Demo)

Prediction Variability

Regression Prediction

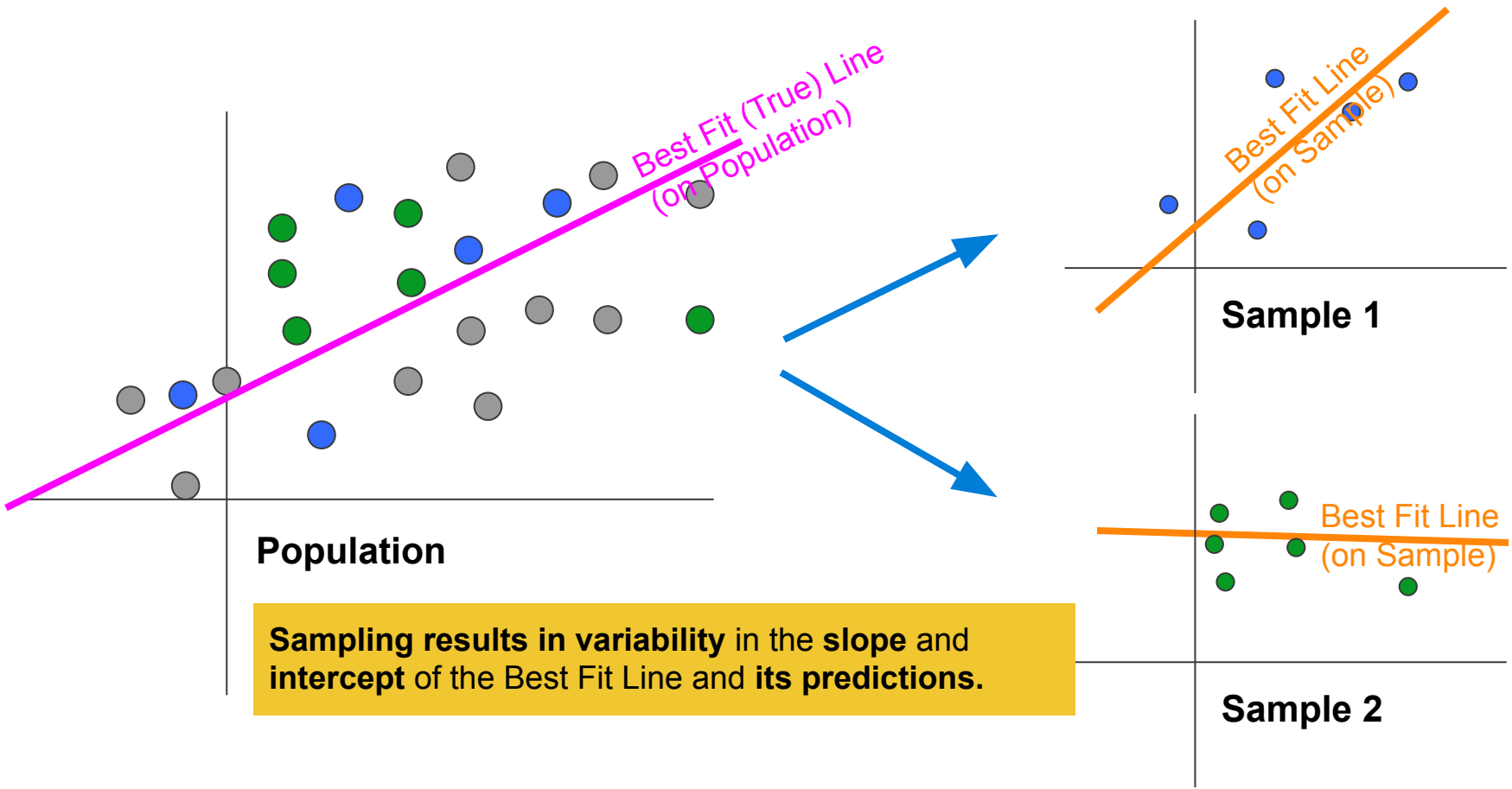
If the data come from the regression model,

- The “true value” of the response *y* at a given value of *x* is the **height of the true line** at *x*
- We can’t see the true line, so we have to estimate this height
- The regression line is most likely close to true line
- Given a new value of *x*, predict y by finding the point on the regression line at that *x*

(Demo)

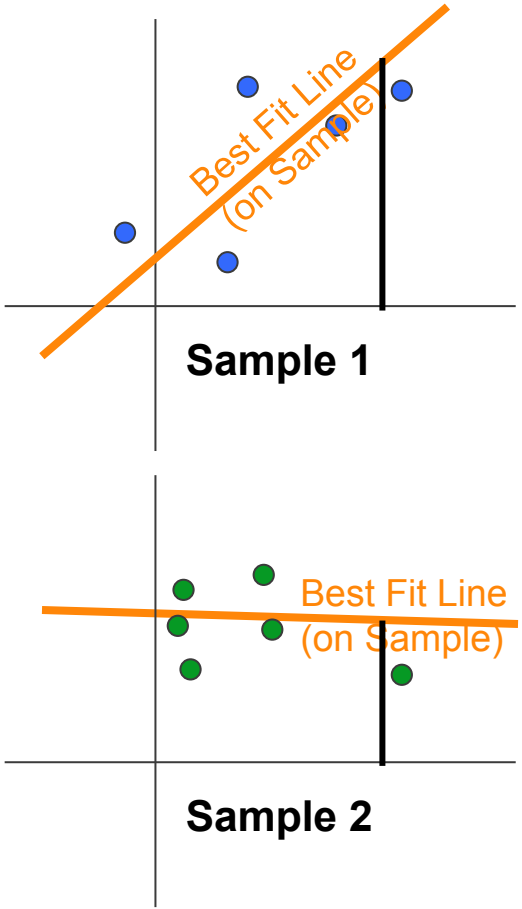
Variability Due to Sampling

Variability in the Line and Prediction due to Sampling



Sampling results in variability in the **slope** and **intercept** of the Best Fit Line and **its predictions.**

Prediction Variability



Prediction Interval

- **Bootstrap the scatter plot**
- **Get a prediction for *y* using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the “middle 95%” interval. This is our range of predictions of *y*.
- It is an approximate **95% confidence interval for the height of the true line** at *x*.

(Demo)

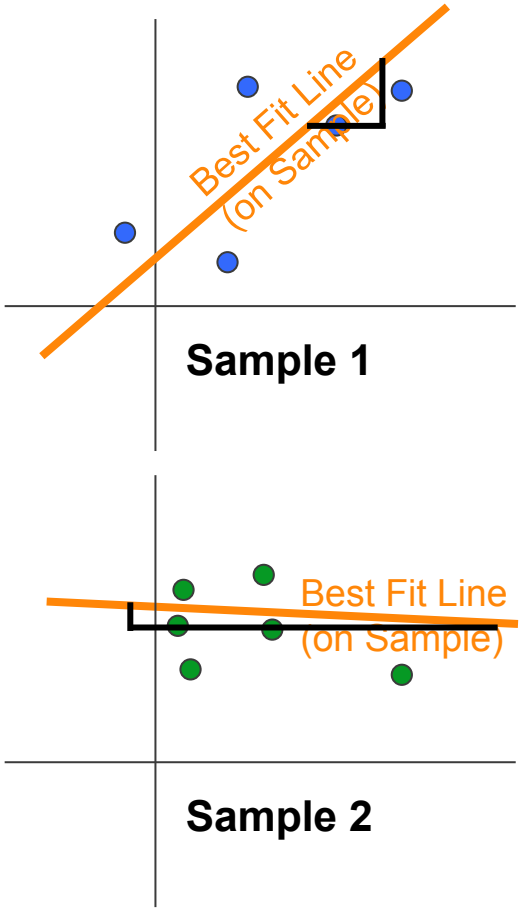
Predictions at Different Values of *x*

- Since *y* is correlated with *x*, the predicted values of *y* depend on the value of *x*.

- The width of the prediction interval also depends on *x*.
  - Typically, intervals are wider for values of *x* that are further away from the mean of *x*.

(Demo)

Slope Variability



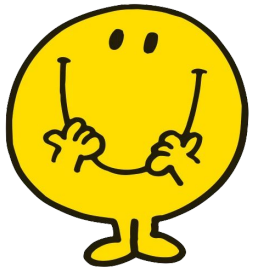
Confidence Interval for True Slope

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- **Repeat.**
- Draw the empirical histogram of all the generated slopes.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the slope of the true line.

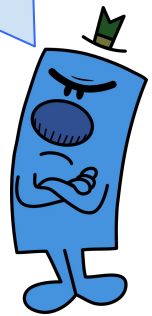
(Demo)

Rain on the Regression Parade

We observed a slope based on our sample of points.



But what if the sample scatter plot got its slope just by chance?



What if the true line is actually FLAT?!

(Demo)

Test Whether The Slope has a Trend

- **Null hypothesis:** The slope of the true line is 0.
- **Alternative hypothesis:** No, it's not.
- Method:
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval
    - **doesn't contain 0**, reject the null hypothesis.
    - **does contain 0**, there isn't enough evidence to reject the null hypothesis (fail to reject the null)

Test Whether The Slope has a Trend

- **Null hypothesis:** The slope of the true line is 0.
- **Alternative hypothesis:** No, it's not.
- Method:
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval
    - **doesn't contain 0**, reject the null hypothesis.
    - **does contain 0**, there isn't enough evidence to reject the null hypothesis (fail to reject the null)

Food for Thought

- It is important to take into account the processes that people take when constructing opinions.
- Based on just one sample of something, we cannot assume that the trend we see is true.
- Bootstrapping with regression inference introduce a way that we can quantify how certain we are of our conclusions.