Bootstrap (the basics)

- Use it to get confidence intervals for a very wide variety of estimators.
- Under the bootstrap assumption:

$$\boxed{\sigma^2_{\substack{\text{sampling} \\ \text{(w/replacement)}}} \approx \sigma^2_{\text{(across samples)}}}$$

- Let B (usually in the hundreds or thousands) be the number of bootstrap replicates.
- If our original sample is $(x_i, y_i)_{i=1}^{n}$

Then, for each bootstrap replicate we'll have

$$\left(X_{J(b)}, Y_{J(b)} \right)_{J \in I}$$

where I = set of indices $= \{1, \dots, N\}$

# of bootstrap replicate for $b = 1, \dots, B$.

and $\boxed{\{J\}_{(b)}}$ is a sequence of indices which are sampled with replacement from I.

Example. we have data sample

$(x_1, x_1), (x_2, y_2), (x_3, y_3),$

$(x_4, y_4), (x_5, y_5)$

$\Rightarrow I = \{1, 2, 3, 4, 5\}$ (Set of indices)

$(b=1) \rightarrow J(1) = \{1, 1, 2, 3, 4\}$ (Sequence of indices sampled for bootstrap replicate 1).

$(b=2) \rightarrow J(2) = \{1, 2, 3, 3, 3\}$

So, for $b = 1$ ...

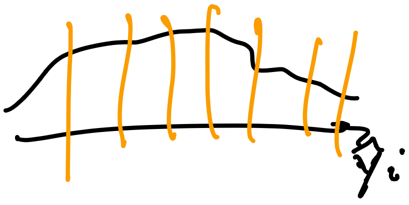
$(X_{J(1)}, Y_{J(1)})$

For each b, compute $\hat{y}_{J(b)} = E(Y_{J(b)} | \underline{X}_{J(b)}]$ using our lowess estimator. Then, save $\{J(b)\}$ and $\hat{y}(b)$

We end up with table ...

$b \rightarrow$ / $i \downarrow$	1	2	3	...	B
1	$\hat{y}_1(1)$	—			
2	$\hat{y}_2(1)$	$\hat{y}_2(2)$			
3	—	$\hat{y}_3(2)$			
4	$\hat{y}_4(1)$	—			
5	$\hat{y}_5(1)$	—			
⋮					
N	$\hat{y}_N(1)$	$\hat{y}_N(1)$			

→ For each observation i



Further, you can compute an estimate of standard error (observation-specific)

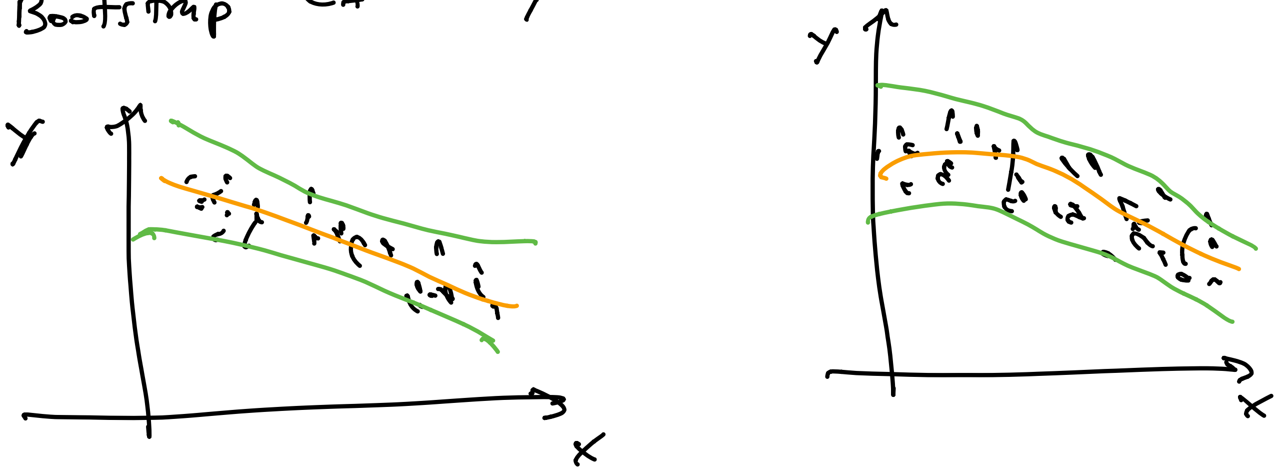
$$\sigma^{Boot}_{\hat{y}_i} = \sqrt{\frac{\sum (\hat{y}_i(b) - \bar{\hat{y}}_i)^2}{B_i - 1}}$$

Average across bootstrap replicates

# of bootstrap replicates containing observation i.

B is fixed , B_i is variable

Comparing the width and shape of CI for OLS vs. Bootstrap CI's, we observe:



OLS

① OLS CI's are tighter (considers an infinite # of samples).

② On the other hand, our bootstrap CI's for Loess just consider B samples.

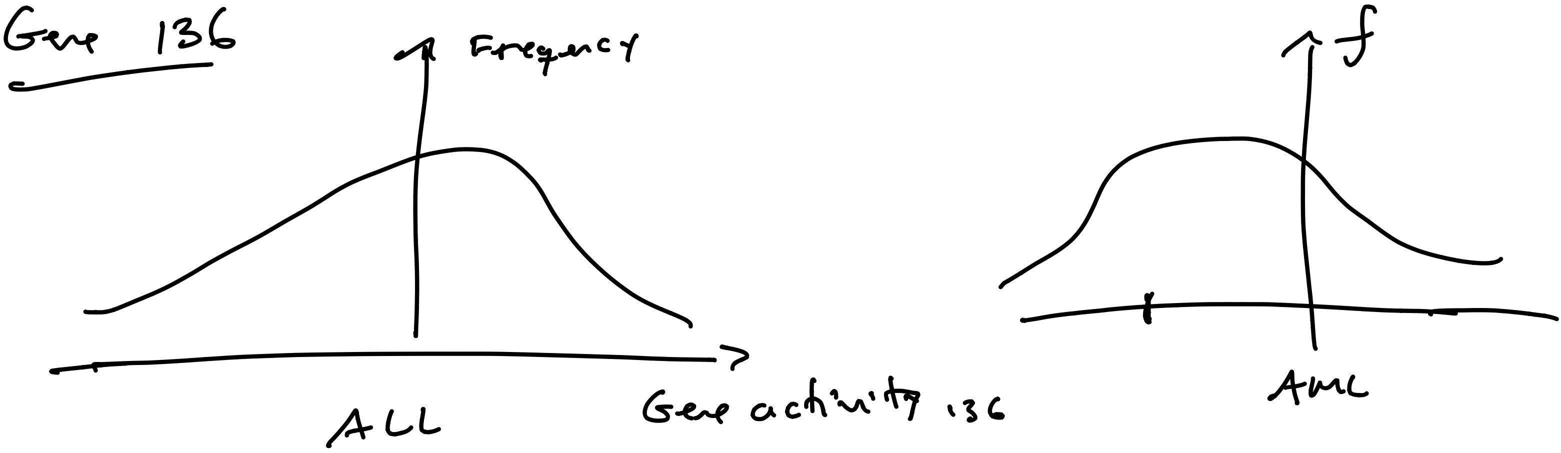
Big data implications on hypothesis testing.

- 72 Leukemia Patients
  - . 47 with ALL (Acute Lymphoblastic Leukemia)
  - . 25 with AML (Acute Myeloid Leukemia)

- Measurements available for 7128 genes for each patient!

→ Let's focus on 1 of them (Gene 136)

Gene 136



- Classic way to check if means are equal is using a t-test for hypothesis

$$H_0 : \boxed{\mu_{AML}} = \mu_{ALL}$$

$$t = \frac{\boxed{\bar{Y}^{AML}_{136}} - \boxed{\bar{Y}^{ALL}_{136}}}{\hat{se}}$$

- $\hat{se}$ = standard error estimate.

- For simplicity, let's assume equal variances.

$$\hat{se} = \hat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}$$

where $n_2 = 47$
$n_1 = 25$

- $\nu = n_1 + n_2 - \boxed{2} = 70$ (degrees of freedom)