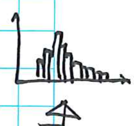
Panini & Monte Carlo tests
(Mainly R code. Read comments). Remarks:

- pvals: line 43: pval.1 $\leftarrow$ (1 + sum(~) / (nsim + 1))
  we add 1 here for the observation (else pval = 0 (in this case))

Power: unlike (simple) p-val. analysis, we need to have a specific H_a
(Fig from line 79, 80 plot: purple = Null, orange = alt. (pink = overlap)
coin package is useful.

Section

Recap Bootstrap: Goal: estimate the distribution of $\hat{\theta}$ so
eg. $\Rightarrow$ construct CI



- In non-parametric boot, we make no assumptions on the distribution of $\hat{\theta}^{**}$
  $\hookrightarrow$ to get boot. sample, we sample with replacement wrt data
- In parametric boot, we assume model P_θ (from building boot $\hat{\theta}^{*i}$; $i = 1, \dots, B$.
  $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

How was the dataset created? (ie P_θ)

The distribution of $\hat{\theta}$ is approximated by $\hat{\theta}^*$ (or $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$)

$X_1, \dots, X_n \rightsquigarrow X_1^*, \dots, X_n^* \Rightarrow \hat{\theta}^{*1}$
data $\Downarrow$ $\hat{\theta}$
$\vdots$
$X_1^*, \dots, X_n^*$ B bootstraps $\Downarrow$ $\hat{\theta}^{*B}$

Non Parametric Boot. $X_1, \dots, X_n$ data $\overset{\text{defines}}{\Longrightarrow}$ P_n.

- Bootstrap sample $X_1^*, \dots, X_n^* \overset{iid}{\sim} P_n \Longrightarrow$ sample with replacement! from $X_1, \dots, X_n$

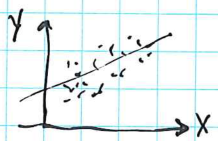
Parametric Boot $X_1, \dots, X_n \overset{iid}{\sim} P_\theta$ $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

- estimate θ and simulate from $P_{\hat{\theta}} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$

Bootstrap sample $X_1^*, \dots, X_n^* \overset{iid}{\sim} P_{\hat{\theta}}$.

Example regression $P_{\hat{\theta}}$: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

$\hat{\theta} \longrightarrow$ Note that in the bootstrapped sample, $\varepsilon_i^* \sim \mathcal{N}(0, \hat{\sigma}^2)$ $(i = 1, \dots, n)$



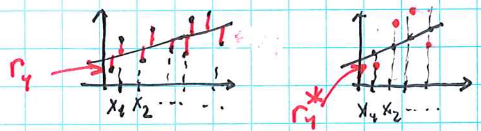
$(X_1, Y_1), \dots, (X_n, Y_n)$ and $Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i^*$

- objective: estimate $\beta_0, \beta_1, \sigma^2$. $\Longrightarrow$ Sample "new" errors
- Residuals $r_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$. Since here we make no assumptions on ε_i, we can use ~~the~~ residuals (bootstrap) sample the residuals with replacement from $r_1, \dots, r_n$

... we have $Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + r_i^*$

- Bootstrapping residuals, cont'd.

$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + r_i^*$

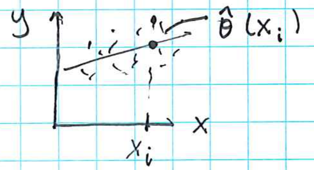


- We keep the x values and sample r_i^*, ie distance from y_i^* to the data point y_i. (regression line)
- gives us r_i (in red), and then look at difference between y_i^* and r_i^* in bootstrap.

Pros/Cons of Param. boot.

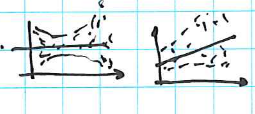
+ If P_θ suits the data, then we get a more accurate estimate (can sample from the distribution)
+ for small n, non-param. boots may be poor
− sensitive to model misspecifications.

We assume the model $\hat{\theta}$ from which we generate the data, but NOT the actual estimator $\theta = (\mu, \sigma^2)$ (model description $\to \varepsilon_i^*$)

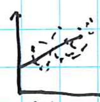


For every x_i along the axis, we sample / estimate $\hat{\theta}(x)$

ex.



- We can apply non-parametric boot even if errors are not iid. (but NOT parametric).
- assume, estimate procedure → estimate on distribution
- Need to be careful about applying the correct model boot with (ex. polynomial models)



If we know θ, we can get CI: (goal: CI with coverage $1-\alpha$)

can generate new data sets

the sample distrs that are not covered should be evenly split above/below θ line (i.e. half the reds are above, half are below).

θ $\quad \hat{\theta}_n \quad \hat{\theta}_n^* \quad \hat{\theta}^{**}$

$X_1, \dots, X_n$

$$\begin{array}{c} 1 \\ \vdots \\ B \end{array}\begin{bmatrix} X_1^*, \dots, X_n^* \\ \vdots \\ X_1^*, \dots, X_n^* \end{bmatrix} \begin{bmatrix} \hat{\theta}^{*1} \\ \vdots \\ \hat{\theta}^{*B} \end{bmatrix} \Big\} B \text{ boots}$$

$\square M \to \hat{\theta}^{**1}, \dots, \hat{\theta}^{**M} \Rightarrow \hat{sd}(\hat{\theta}^{*1})$

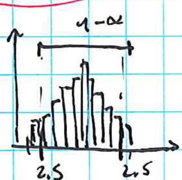
boots

$\square M \to \hat{\theta}^{**1}, \dots, \hat{\theta}^{**M} \Rightarrow \hat{sd}(\hat{\theta}^{*B})$

$\{\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}\}$

CIs

Quantile $\left[q_{\hat{\theta}^*}\left(\frac{\alpha}{2}\right), q_{\hat{\theta}^*}\left(1-\frac{\alpha}{2}\right)\right]$



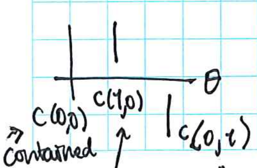
Note: $\hat{sd}$ estimate sd since sd is for infinite # samples (and B is finite)

→ Normal (we assume this is a for simplicity)

$\hat{\theta} \pm q_z\left(1-\frac{\alpha}{2}\right)\hat{sd}(\hat{\theta}), \quad Z \sim N(0,1)$

→ reversed $\hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}\left(1-\frac{\alpha}{2}\right) \hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}\left(\frac{\alpha}{2}\right)$

boots. $\left[\hat{\theta} - q_{\frac{\hat{\theta}^*-\hat{\theta}}{\hat{sd}(\hat{\theta}^*)}}\left(1-\frac{\alpha}{2}\right)\hat{sd}(\hat{\theta}),\ \hat{\theta} - q_{\frac{\hat{\theta}^*-\hat{\theta}}{\hat{sd}(\hat{\theta}^*)}}\left(\frac{\alpha}{2}\right)\hat{sd}(\hat{\theta})\right]$ $\quad \left(\frac{\hat{\theta}^{*i} - \hat{\theta}}{\hat{sd}(\hat{\theta}^{*i})}, i = 1, \dots, B\right)$



← R output if true param is ...

greater than CI (upper) bound

less than CI bound (lower)

- double boot sampled from $\{\hat{\theta}^{**}\}_M \to \hat{sd}(\hat{\theta}^{*i})$

$\{\hat{\theta}_1^{*i}, \dots, \hat{\theta}_n^{*i}\} \quad i = 1, \dots, B.$