Recap

P-values based on multi sample splitting

avoid "double dipping" using the data twice for variable selection and using statistical inference (tests, confidence intervals) afterwards
$\leadsto$ sample splitting

multiple sample splitting is much more reliable and statistically better than splitting once

Fixed design linear model

$$Y = X\beta^0 + \varepsilon$$

split the sample into two parts I_1 and I_2 of equal size $\lfloor n/2 \rfloor$

- use (e.g.) Lasso to select variables based on I_1: $\hat{S}(I_1)$
- perform low-dimensional statistical inference on I_2 based on data $(X_{I_2}^{(\hat{S}(I_1))}, Y_{I_2})$;
  for example using the t-test for single coefficients β_j^0
  if variable $j \in \hat{S}(I_1)$: use the p-value from t-test
  if $j \notin \hat{S}(I_1)$: assign the p-value equal to 1

due to independence of I_1 and I_2, this is a “valid” strategy
(but see later)