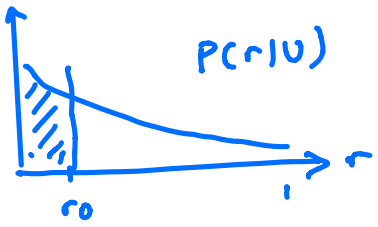
Using p-values to estimate false positives

RNA chemical modifications reactivity (SHAPE, DMS)

$0 \le r \le 1$ reactivity

H_1: $\begin{cases} r \le r_0 & \text{residue is P} \\ r > r_0 & \text{" is U} \end{cases}$

H_0: residues are Unpaired



← obtained by getting the reactivities of many residues that are U

$$\text{p-value}(r_0)_{H_0} = P(r \le r_0 | U) = CDF_{H_0}(r_0)$$

reactivity r_0	p-value		expected FP
0.0029	0.02	2% chance an Ur has reactivity ≤ 0.0029	N 0.02
0.0034	0.05		N. 0.05
0.0042	0.10		N. 0.10

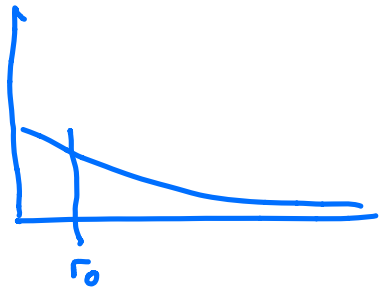
If I test N residues → max expected False Positives if all tested were derived from the null hypothesis (were unpaired)

Test N residues

F° called P

$(r \le r_0)$

$pval^\circ = P(r \le r_0 \mid U)$



* $pval^\circ$ is the fraction of N expected to be FP

$$FP^\circ = pval^\circ \cdot N$$

False Discovery Rate $\quad FDR^\circ = \frac{FP^\circ}{F^\circ}$ => fraction of F° expected to be FP

* FDR° is the fraction of F° expected to be FP

. If you are testing an RNA w/ known structure

$N = 265$

$T = 160$ (P) $\quad T$ = fraction of N that are P

$T \cap F^\circ$ = True positives with $r \le r^\circ$

$$sen^\circ = \frac{T \cap F^\circ}{T} = \text{Fraction of P that you will detect with } r \le r_0$$

r	$pval^\circ$	fdr°	sen° (%)	F^*	FP^*	T^*
0.0029	0.02	0.17	16.9	29	2	27
0.0034	0.05	0.30	23.1	42	15	37
0.0042	0.10	0.45	30.6	59	25	49
0.0063	0.28	0.67	50.0	108	73	79

detecty 50% of P, implies ~28% of 265 are FP

~67% of 108 are FP

"P-values should be treated with caution"

- no info about H_1
  rejecting H_0 does not mean H_1 is true
- p-values are not the probability of any model

  p-value = 0.02 $\not\Rightarrow$ $P(H_0 \text{ is true}) = 0.02$, $P(H_1 \text{ is true}) = 0.98$
- p-values cannot be used to compare different models
- "I run test, and the p-value is 10^{-5}"

  should you be surprised?

  $CDF_X(x)$ is a $U[0:1]$

  p-values are uniformly distributed!! $[0:1]$

  If data follows the null hypothesis

  you should be as surprised of a

  value of 0.05 as one of 0.99!
- When someone uses a p-value, request that they tell you under which null hypotheses they are operating!

How to treat p-values with caution

(1) precisely define the null hypothesis and it's pdf

$$P(x|H_0)$$

(2) when possible, estimate the $P(x|H_0)$ yourself, don't rely on test with obscure or unspecified assumptions. The all have them!

(3) P-values do not validate, nor refute any hypothesis. They give you a sense of the # of FP you should expect, based on which you may take further actions.

(4) What to do if you have more than 1 alternative hypothesis?
The bayesian way is clear:

$$P(H_1|D) = \frac{P(D|H_1)\cdot P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + P(D|H_3)P(H_3)}$$

but using p-values?