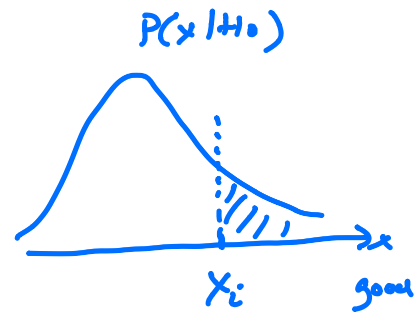
W04 - p-values, significance. The Student's t-test

Experiments - X $\longrightarrow$ Data $\{x_i\} = D$

H_1 - interesting hypothesis about D, that you want to test / validate

H_0 - null hypothesis that "could" also explain your results



H_1 is hard to formulate

H_1 is easier to formulate

take one value x_i

$$\text{p-value}(x_i) = P(x \geq x_i \mid H_0) \quad \text{(right tail)}$$

$$= 1 - P(x < x_i \mid H_0)$$

$$= 1 - CDF_{H_0}(x_i)$$

$$\text{p-value}(x_i) = P(x \leq x_i \mid H_0) \quad \text{left tail}$$

$$\text{p-value}(x_i) = P(x \geq x_i \wedge -x_i \leq x \mid H_0) \quad \text{two tailed test}$$

If p-val is small, there is little chance that H_0 describes the data.

p-value does not use any info about the other hypothesis H_1

rejecting H_0 does not mean H_1 is true

Example experiment failure rate

$H_0 : f = f_0 = 0.2$ what postdoc told you

$H_1 : f > f_0$

$D = \{ s\,s\,f\,f\,f \}$

$$\text{p-value}(ssfff)_{H_0} = P(ssfff \mid f = f_0) + P(sffff \mid f = f_0) + P(fffff \mid f = f_0)$$

$$= \frac{5!}{2!\,3!} f_0^3 (1-f_0)^2 + \frac{5!}{1!\,4!} f_0^4 (1-f_0) + \frac{5!}{0!\,5!} f_0^5$$

$$\underset{(f_0 = 0.2)}{=} 0.05792$$

What do you do with this result

$\text{p-val}(ssfff)_{H_0} = 0.05792$

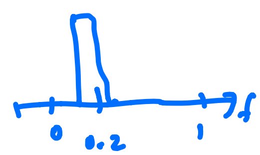
do you reject $f = f_0$ or not?

The Bayesian Approach

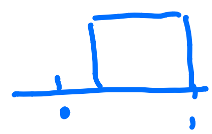
$$\frac{P(H_0 | ssfff)}{P(H_1 | ssfff)} = \frac{P(ssfff | H_0)\, P(H_0)}{P(ssfff | H_1)\, P(H_1)} \Bigg|_{P(H_0) = P(H_1)}$$

$$= \frac{\int_0^1 df\; P(ssfff | f)\, P(f | H_0)\, df}{\int_0^1 df\; P(ssfff | f)\, P(f | H_1)\, df}$$

$$P(f | H_0) = \begin{cases} \dfrac{1}{0.21 - 0.19} & 0.19 < f < 0.21 \sim |f - f_0| < 0.1 \\ 0 & \text{otherwise} \end{cases}$$



$$P(f | H_1) = \begin{cases} \dfrac{1}{1 - 0.2} & f > f_0 = 0.2 \\ 0 & \text{otherwise} \end{cases}$$



$$\frac{P(H_0 | ssfff)}{P(H_1 | ssfff)} = \frac{\frac{1}{0.02} \int_{0.19}^{0.21} P(ssfff | f)\, df}{\frac{1}{0.80} \int_{0.2}^{1} P(ssfff | f)\, df}$$

$$= \frac{\frac{1}{0.02} \int_{0.19}^{0.21} f^3 (1-f)^2\, df}{\frac{1}{0.8} \int_{0.2}^{1} f^3 (1-f)^2\, df} \simeq \frac{20}{80}$$

{ 80 : 20 chance $f > f_0$ }

Which method do you prefer?

* p-value $= 0.0579$
    - nothing about H_1
    - easier but hard to interpret

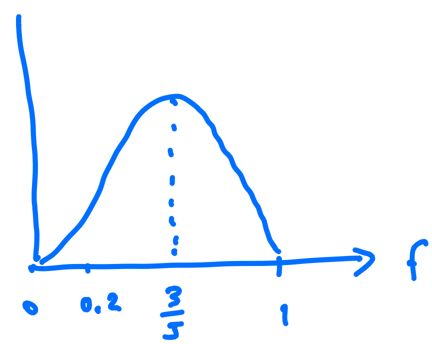
$$\text{pval} \neq P(H_0 \text{ is true})$$

$$1 - \text{pval} \neq P(H_1 \text{ is true})$$

* Bayes
    + model comparison $H_1 : H_0$ = $80 : 20$
    + posteriors

$P(f \mid ssfff)$

$f^* = \frac{n}{N}$



p-value is indirect

If I were to repeat a $N=5$ experiment many times and $f = f_0$, there will be a $\sim 6\%$ chance of obtaining at least 3 failures (ssfff or sffff or fffff)

Student's t-test   a widely used p-value

structural RNAs

RNaseP RNA
(ribozyme)

base pairs
A:U U:A
C:G G:C
G:U U:G

Given the RNaseP sequence, what is the structure?

265 residues

160 paired (80 bp)

105 unpaired

Experimental chemical modification reactivities

r = reactivity ($0 < r \simeq 1$)

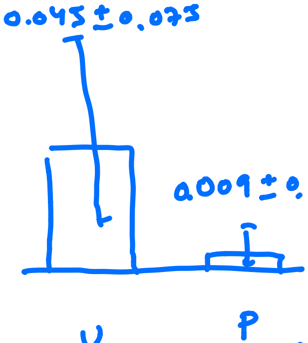
r measures "flexibility" : is this a synonym of being paired/unpaired?

DMS, SHAPE (selective 2' hydroxyl acylation analyzed by primer extension)
-seq

$D = \{r_1 \ldots r_{265}\}$

$D_u = \{105\ r\text{'s}\}$

$D_p = \{165\ r\text{'s}\}$



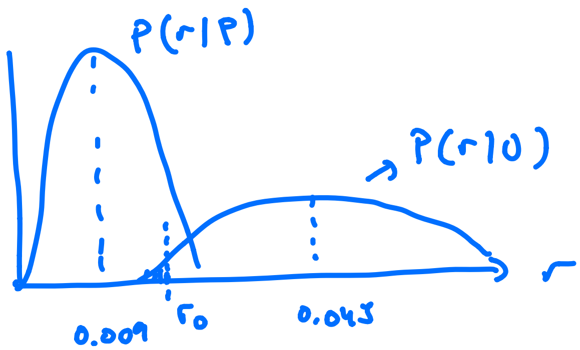
student_ttest
(D_u, D_p)

p-val $= 4.5 \cdot 10^{-9}$

Great! SHAPE should be enough to distinguish P/U

① Look at the data and assumptions

Ideal scenario



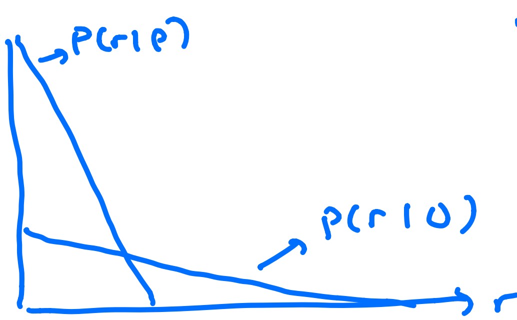
$r < r_0$ call it P

$r > r_0$ call it U

$\text{p-val}(r_0)_U = P(r < r_0 | U) \simeq$ small (False Positives)

$P(r < r_0 | P) \simeq 0.9$ (True Positives)

Real data



very non Gaussian

T-test assumes data is Gaussian!

and it uses the means only

What does the Student's distribution has to do with anything? sivia 2.3, 3.2

assume $\{x_i\}$ follows a Gaussian dist $\mathcal{N}(\mu, \sigma)$

$$P(\mu, \sigma \mid \{x_i\}) \propto P(\{x_i\} \mid \mu, \sigma)\, P(\mu, \sigma)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}} \quad P(\mu)\, P(\sigma)$$

$$P(\mu) = \begin{cases} \frac{1}{\mu_+ - \mu_-} & \mu_- < \mu < \mu_+ \\ 0 & \text{otherwise} \end{cases}$$

$$P(\sigma) = \begin{cases} c & \sigma > 0 \\ 0 & \sigma < 0 \end{cases}$$

$$P(\mu, \sigma \mid \{x_i\}) \propto \sigma^{-N} e^{-\frac{\sum_j (x_i-\mu)^2}{2\sigma^2}}$$

$$P(\mu \mid \{x_i\}) \propto \int_0^\infty d\sigma\, \sigma^{-N} e^{-\frac{\sum_j (x_i-\mu)^2}{2\sigma^2}}$$

$$t = \frac{\sqrt{\sum_i (x_i-\mu)^2}}{\sigma} \qquad d\sigma = -\frac{\sqrt{\sum_i (x_i-\mu)^2}}{t^2}\, dt$$

$$P(\mu \mid \{x_i\}) \propto \left[\sum_i (x_i-\mu)^2\right]^{\frac{-N+1}{2}} \int_0^\infty t^{N-2} e^{-t^2/2}\, dt$$

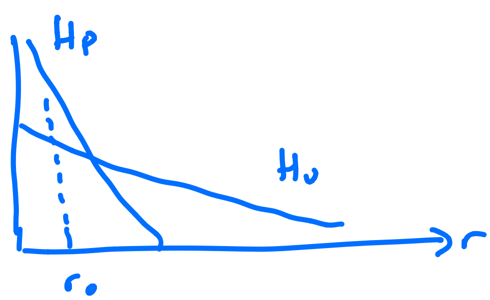
student's distribution.

Is this p-value that compare the means really what we want to know?

$\mu_1 = 0.045$ t-test $\propto (\mu_1 - \mu_2)^2$

$\mu_2 = 0.009$ ↳ follows Student's dist

What I want to know is if I make the assignment



$r < r_0$ is P

$r > r_0$ is U

What errors do I make?

$$\text{p-value}(r_0) = P(r \leq r_0 | H_U) = CDF_{H_U}(r_0)$$

prob that a unpair residue has $r \leq r_0$

If test $n \sim n \cdot \text{pval}(r_0)$ = Expected # FP if all res follow null

Then pick p-value based on how many FP you are willing to tolerate

react	p-value	
0.0029	0.02	2% FP
0.0034	0.05	5% FP
0.0042	0.10	10% FP

N = # of test

p^* = p-value at r^*

F^* = of N w/ $r \le r^*$

T^* = $T \cap F^*$

T = 165 (trues)

$$FDR = \frac{p^* \cdot N}{F^*}$$

fraction of positive calls that should be expected to be U

p^* fraction of tested (N) expected to be false calls

fdr fraction of positives (F^*) expected to be false calls

r	p^*	fdr	sen
0.0029	0.2	17.4	16.9
0.0034	0.05	30.0	23.1
0.0042	0.10	44.8	30.6
0.0063	0.28	66.5	50.0

to get 50% of real paired bases, more than half of the called paired are going to be wrong