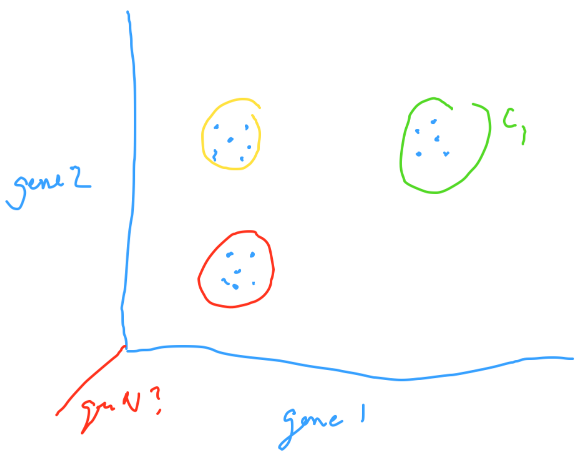
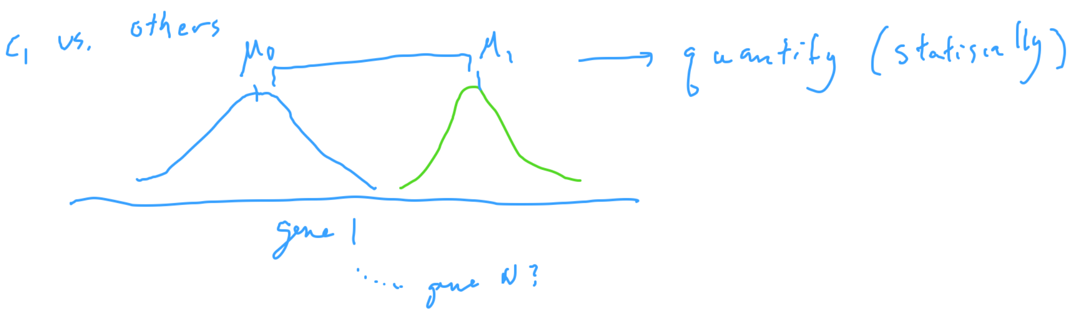
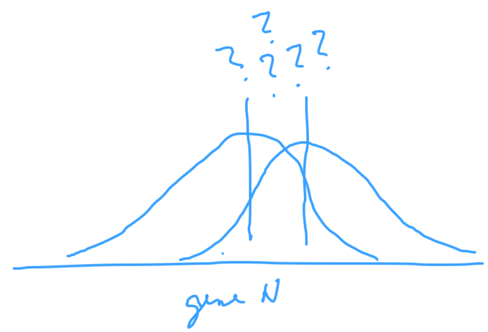
Differential Expression Analysis

Goal: identify cell-types
Last class: identify clusters of cells presumably transcriptionally similar, visualize them in a 2D



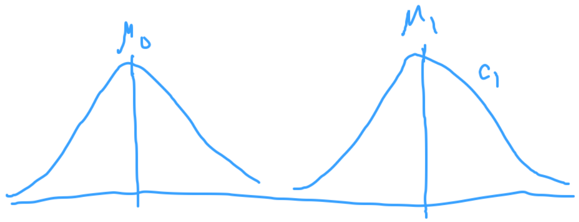
Taking a statistical approach to sift through all the genes and ask questions like:
- which genes are upregulated in our cluster of cells of interest

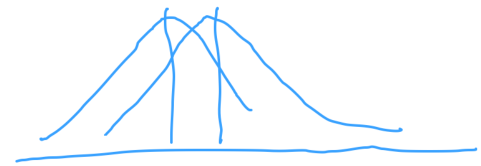




T-test

assumes normal distribution -> we can test whether there's a difference between the means





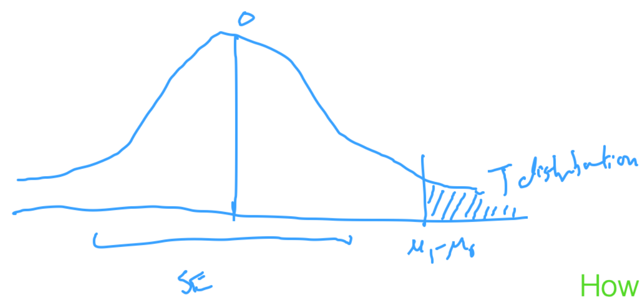
H_0: $\mu_0 - \mu_1 = 0$ there's no difference

H_A: one sided $\mu_0 > \mu_1$ (1) differentially downregulated

$\mu_0 < \mu_1$ (2) ✓ differentially upregulated

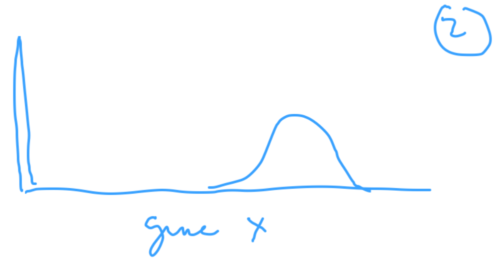
two-sided $\mu_0 \neq \mu_1$ (3) differentially expressed

T-statistic: $\frac{\mu_1 - \mu_0}{SE}$ →



How likely am I to observe such a high m1-m0 if the H0 is true?
-> very unlikely

↳ p-value low < 0.05



to keep in mind: look at how your data looks and pick the appropriate statistical test

At least in the spatial transcriptomics gene expression data that we've been looking, data is not normally distributed -> t test is not the most appropriate

Mann Whitney U test (Wilcoxon test -> there is another wilcoxon test)

"non-parametric" test -> no assumption of normality

Rather than asking: is there a difference between the means
It's asking instead, is there a difference between the ranks

gene 1 expression

c_1		other	
20	(7)	6	(6)
30	(9)	1	(1)
50	(10)	2	(2)
20	(8)	3	(3)
		5	(4)
		6	(5)

$n_1 = 4$ $n_2 = 6$

rank sum: 7+9+10+8

T_1 T_2

$r(c_1) > r(\text{other})$

↳ statistically?

H_0: no group differences between the rank sums

H_A: one-sided or two-sided version

U-statistic

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2$$

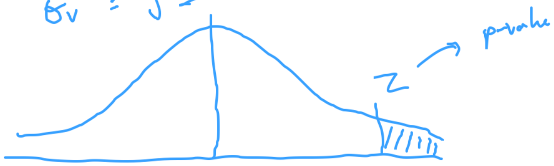
$U \propto -T$

$$U = \min(U_1, U_2)$$

↳ distribution is known → gaussian

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{2}}$$



$$Z = \frac{U - \mu_u}{\sigma_U}$$

There are other statistical tests that are more tailored to our data
-> you could also develop other statistical tests
-> improve our power in detecting true positives and minimizing false negatives