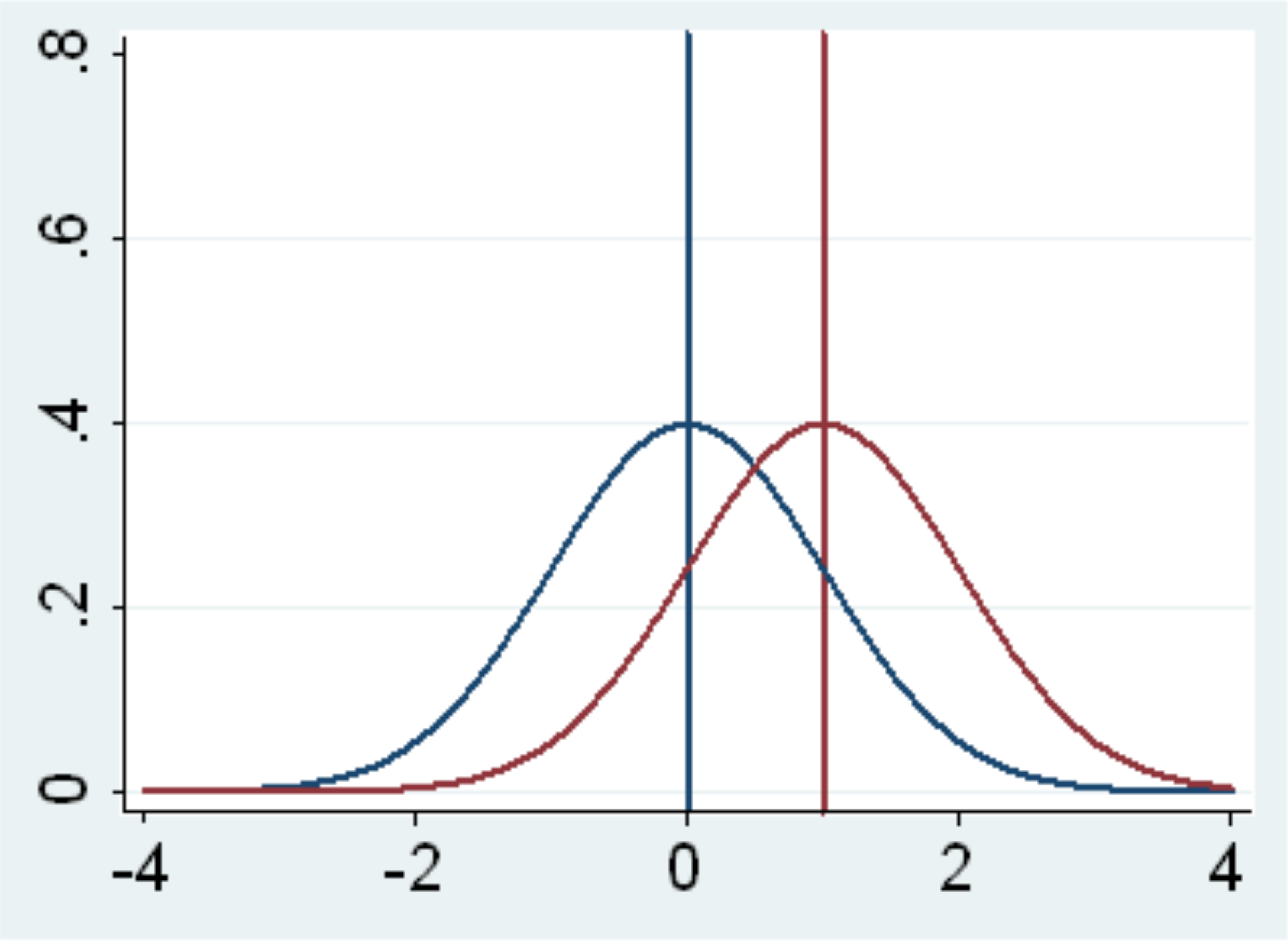
RECORD!!!!

R for bioinformatics - Lesson 4

Student t-test
Multiple testing
Analysis of Variance

Student t-test

In biological sciences Student t-test is typically used to establish whether the means of two samples are truly different (two-sample t-test)



Are these two samples coming from the same population or different ones?

Assumptions of a t-test

- Data is continuous (not discrete)
- Data is collected from a representative, randomly selected portion of the total population
- Data is sampled from a normal (gaussian) bell-shaped distribution
- The samples drawn are reasonably large (when plotted, they approach a bell-shaped distribution)
    - In practice, t-test is often times used on small samples, consisting of as few as three data points, which is inappropriate, but widely seen practice.
- The variance of two samples to be compared is approximately equal (homogeneity of variance).
    - If this assumption cannot be made, a Welch’s test should be used instead.

Variations of t-test

- Two-sample t-test
    - To establish whether the means of two samples are different
- One-sample t-test
    - To establish whether the mean of a sample is different from a pre-defined value
    - One-tailed vs. two-tailed
- Paired t-test
    - Is used in repeated measurement experiments, when, for instance, the same subject or animal is sampled over time, and we need to establish whether responses at one time point are different from responses at another.

A bit of math

The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

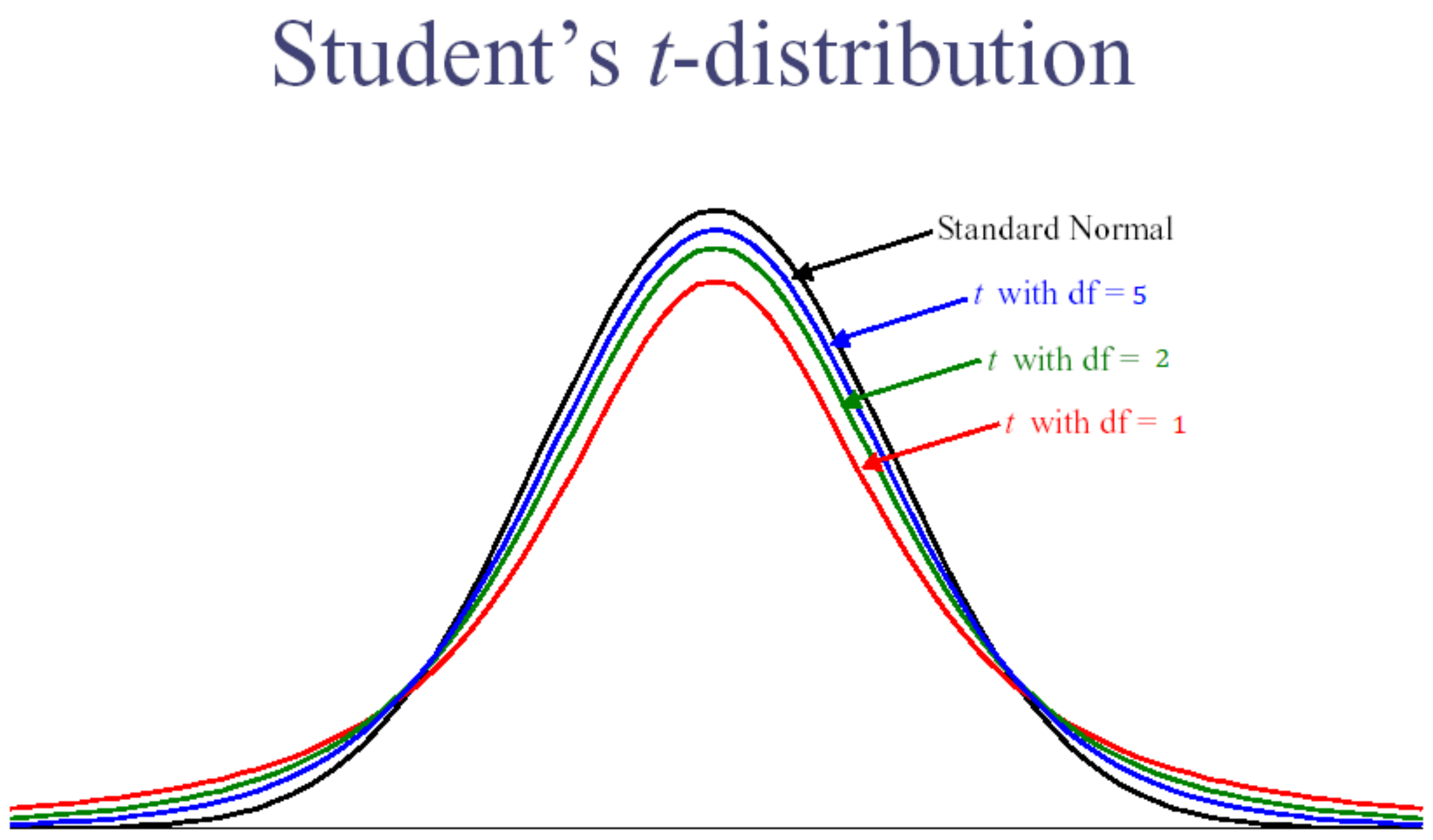
Where X1, X2 are the means of two samples, n1, n2 are the sample sizes

where

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}$$

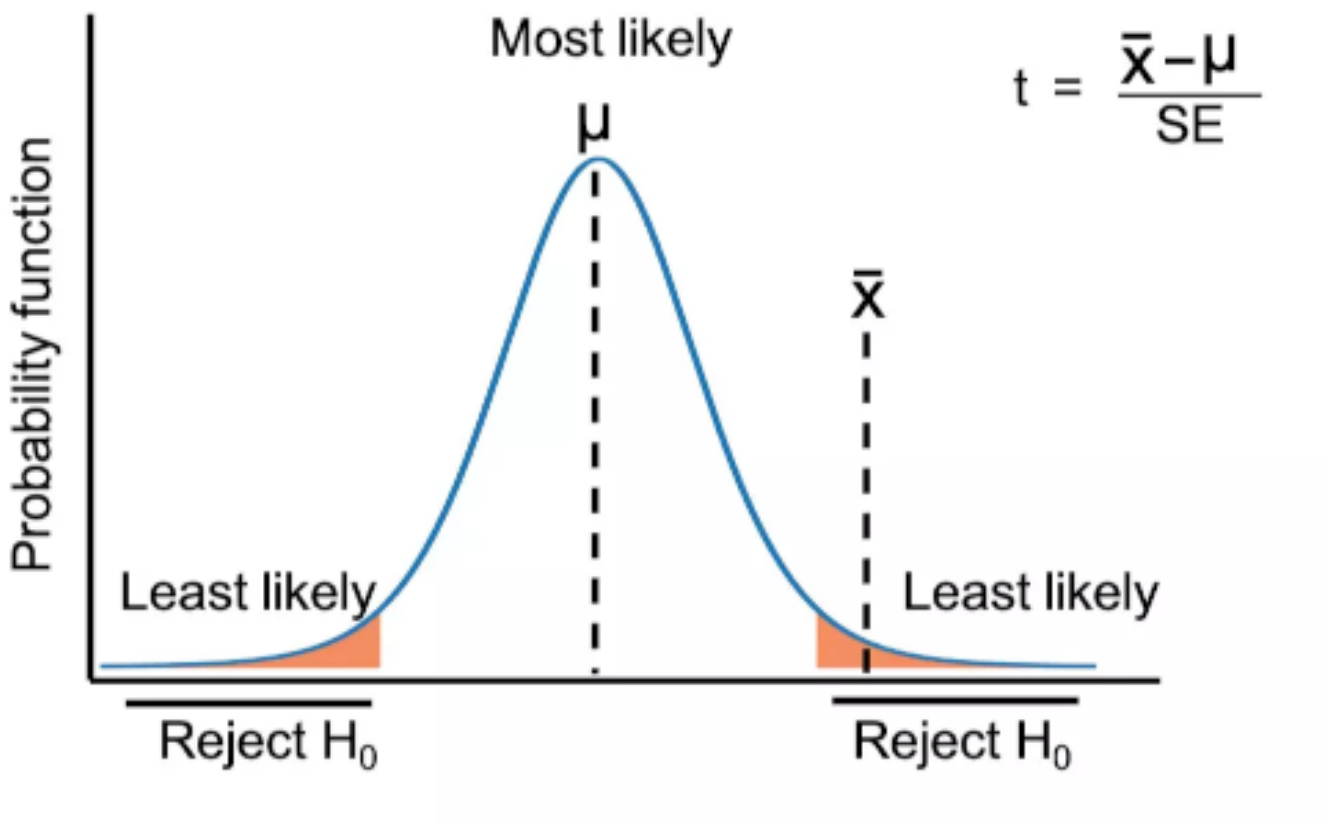
Is an estimator of a pooled standard deviation of the two samples
(sX1, sX2 are the standard deviations of the two samples)

The resulting t-statistic follows a bell-shaped distribution, with the exact form depending on the degrees of freedom (for two-sample t-test it's (n1 + n2) - 2)



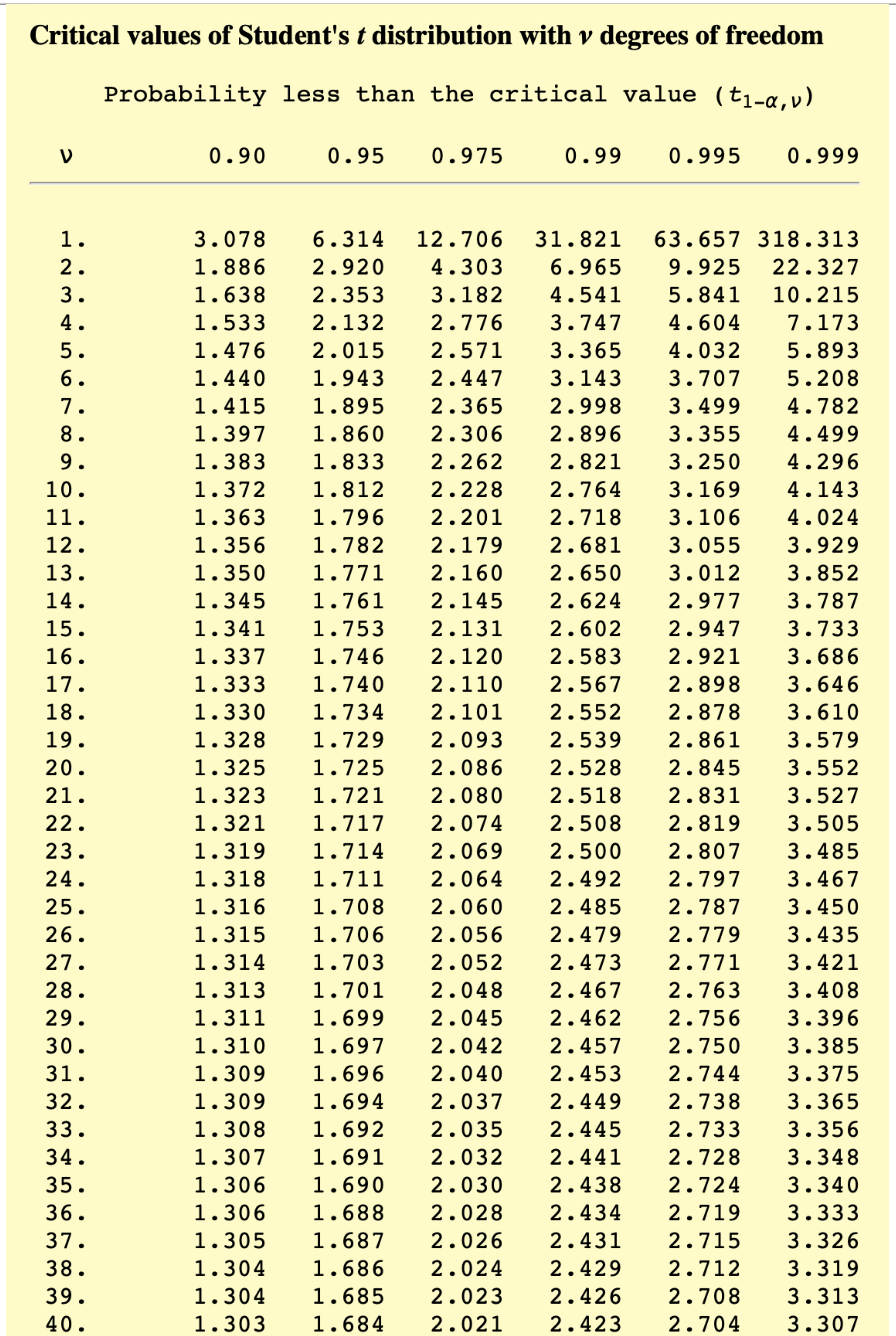
A bit of math (cont'd)

Critical values of t for different alpha (confidence) levels are found in special tables:



Based on whether or not the value of t is greater or smaller than the critical value, we can either reject the null hypothesis that there is no difference between the two means, that is, to establish that at the probability level alpha, the means of the two samples are indeed different, or keep the null hypothesis (no difference).

Typically the alpha level is established at 0.05



Critical values of Student's *t* distribution with *ν* degrees of freedom

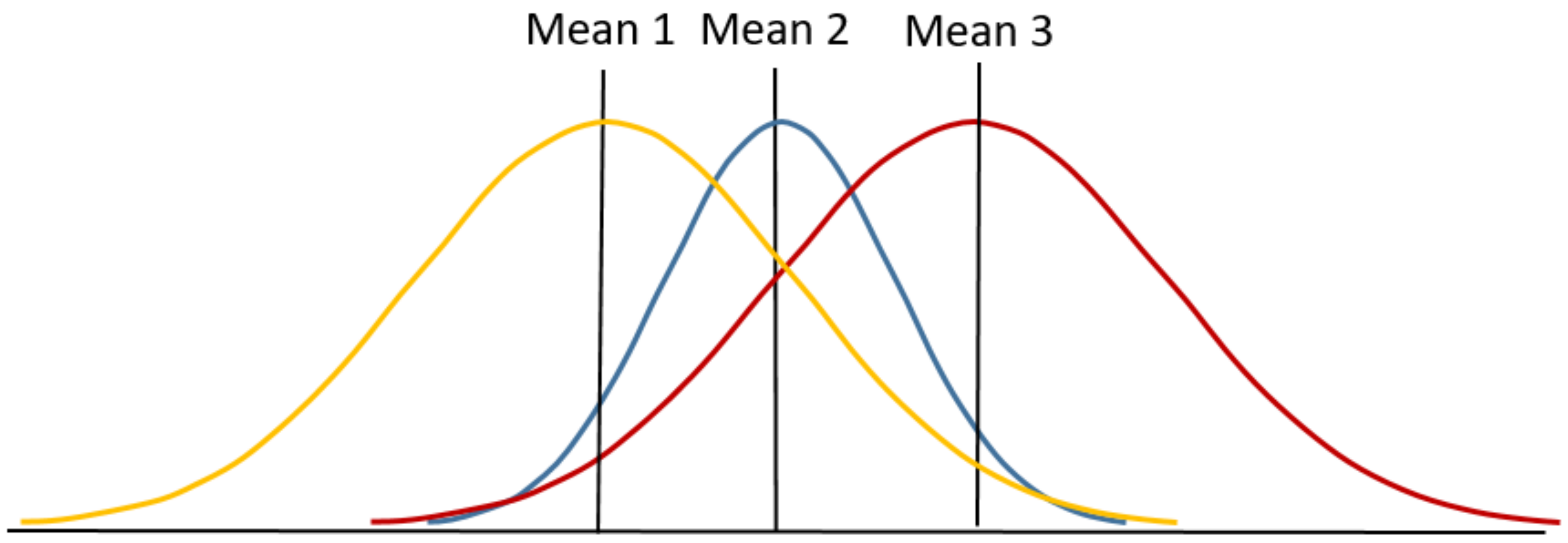
Probability less than the critical value ($t_{1-\alpha,\nu}$)

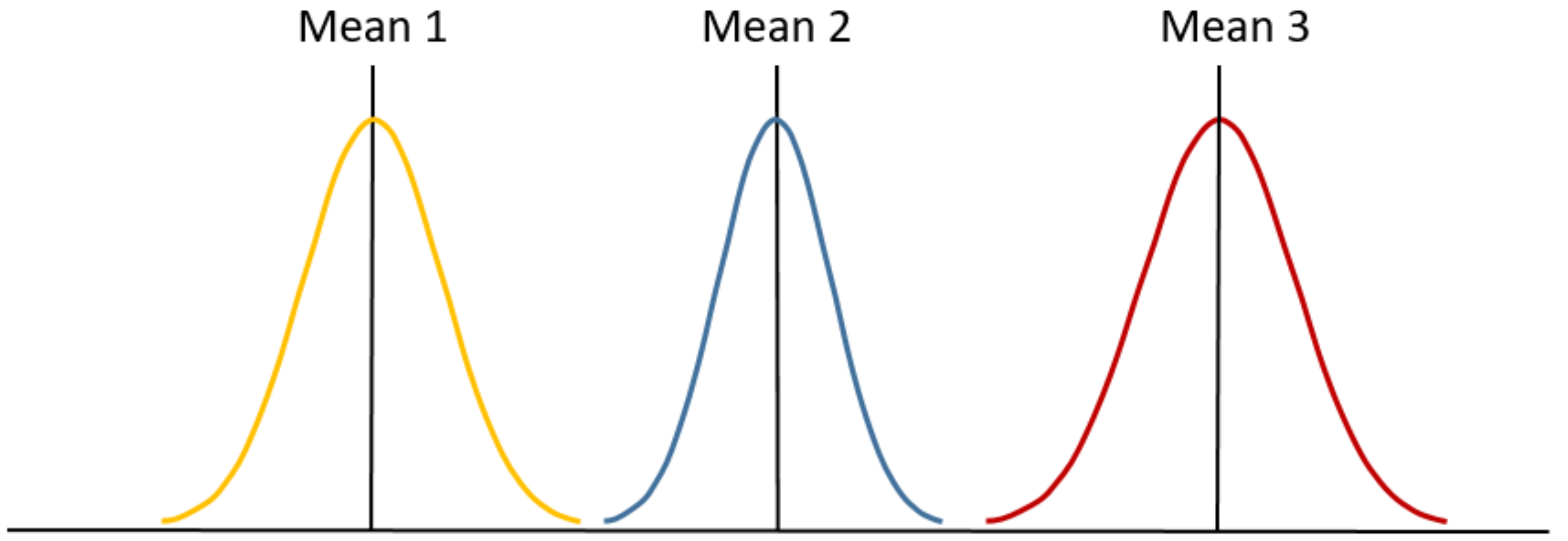
ν	0.90	0.95	0.975	0.99	0.995	0.999
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	3.552
21.	1.323	1.721	2.080	2.518	2.831	3.527
22.	1.321	1.717	2.074	2.508	2.819	3.505
23.	1.319	1.714	2.069	2.500	2.807	3.485
24.	1.318	1.711	2.064	2.492	2.797	3.467
25.	1.316	1.708	2.060	2.485	2.787	3.450
26.	1.315	1.706	2.056	2.479	2.779	3.435
27.	1.314	1.703	2.052	2.473	2.771	3.421
28.	1.313	1.701	2.048	2.467	2.763	3.408
29.	1.311	1.699	2.045	2.462	2.756	3.396
30.	1.310	1.697	2.042	2.457	2.750	3.385
31.	1.309	1.696	2.040	2.453	2.744	3.375
32.	1.309	1.694	2.037	2.449	2.738	3.365
33.	1.308	1.692	2.035	2.445	2.733	3.356
34.	1.307	1.691	2.032	2.441	2.728	3.348
35.	1.306	1.690	2.030	2.438	2.724	3.340
36.	1.306	1.688	2.028	2.434	2.719	3.333
37.	1.305	1.687	2.026	2.431	2.715	3.326
38.	1.304	1.686	2.024	2.429	2.712	3.319
39.	1.304	1.685	2.023	2.426	2.708	3.313
40.	1.303	1.684	2.021	2.423	2.704	3.307

Analysis of variance (ANOVA)

ANOVA is used to establish whether the means of several (more than two) groups are significantly different among themselves, that is, whether these samples come from different populations.

It does not provide a direct way to establish which sample of several is different. It only provides a probability measure of any of these samples being drawn from a different population.





Multiple testing

A p-value originating from a t-test simply reflects the probability that the detected difference is, in fact, not significant. For instance, a p-value of 0.01 indicates that there is 1% chance that the difference is not significant, and 99% chance that it is significant.

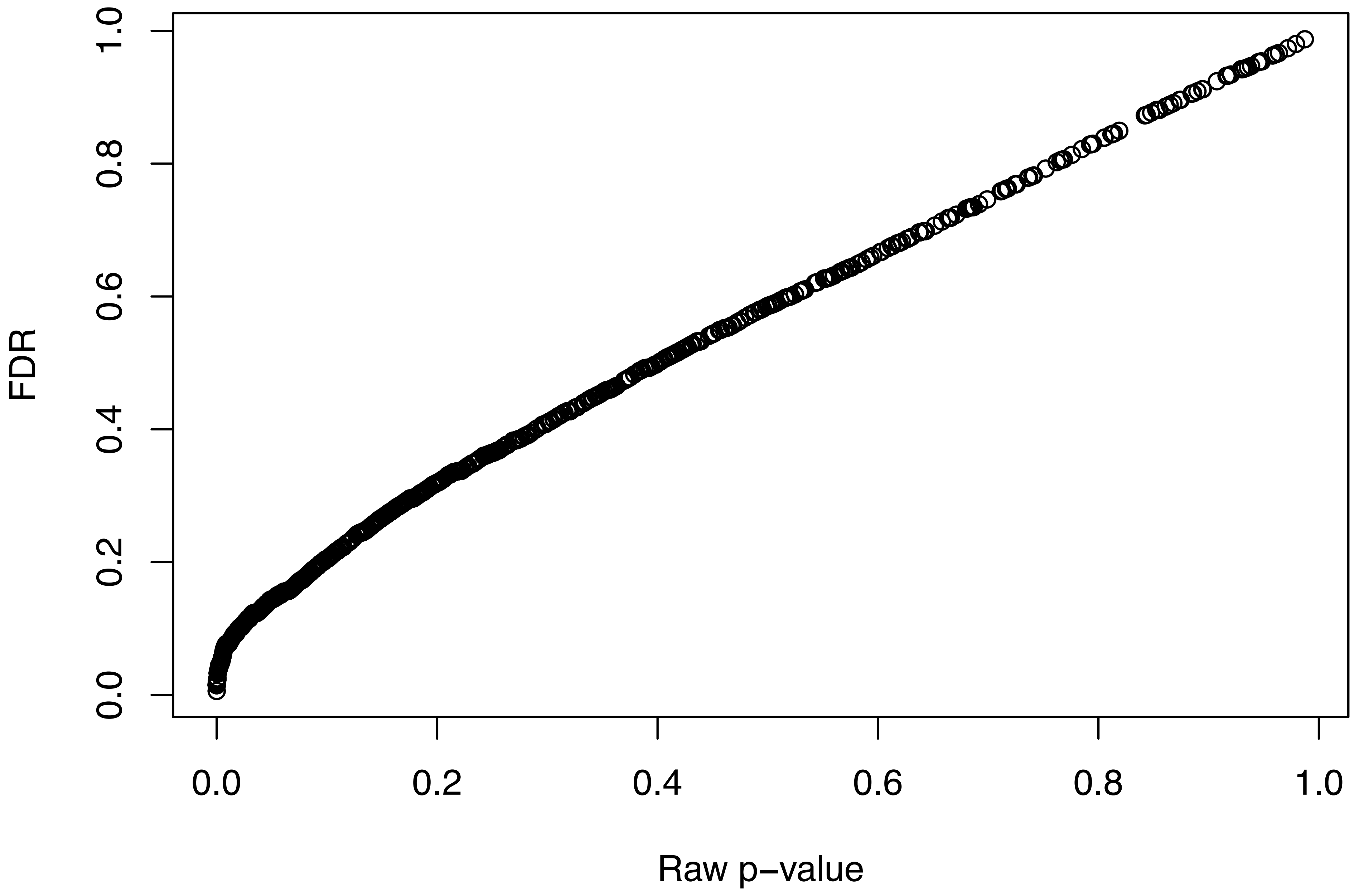
While we can be content with this assumption for a single test, if we run many, many tests, we will start accumulating false positives. For instance, if we run 1,000 tests, at alpha level 0.05, we will get 5% positive tests, or 50 tests, even if there is no real difference.

Bottom line: for a completely random set of numbers, if you run enough t-tests, you will get some significant p-values by random chance.

Remedy:

Correction for multiple testing (also called adjustment of p-values for multiple tests)

- Many different methods
- In biological sciences, the most common is a False Discovery Rate calculation, which, based on the distribution of raw p-values, provides a matching probability of making a Type I error (not rejecting the null hypothesis, or detecting a false positive). This statistic, sometimes referred to as a q-value, represents a False Discover Rate, or FDR.
- We typically try to keep the False Discovery Rate at the 0.05 level, which means that we agree that among the observations with q = 0.05 or less, 5% will be in fact false positives.


FDR
Raw p−value