18.05 Lecture 32
May 2, 2005

Two-sample t-test

$X_1, ..., X_m \sim N(\mu_1, \sigma^2)$
$Y_1, ..., Y_n \sim N(\mu_2, \sigma^2)$
Samples are independent.
Compare the means of the distributions.

Hypothesis Tests:
$H_1 : \mu_1 = \mu_2, \mu_1 \leq \mu_2$
$H_2 : \mu_1 \neq \mu_2, \mu_1 > \mu_2$

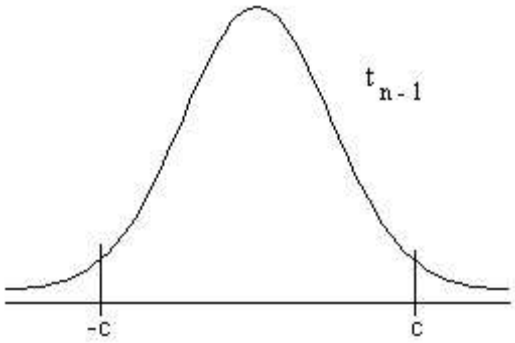
By properties of Normal distribution and Fisher's theorem:

$$\frac{\sqrt{m}(\overline{x} - \mu_1)}{\sigma}, \frac{\sqrt{n}(\overline{y} - \mu_2)}{\sigma} \sim N(0, 1)$$

$$\sigma_x^2 = \overline{x^2} - (\overline{x})^2, \sigma_y^2 = \overline{y^2} - (\overline{y})^2$$

$$\frac{m\sigma_x^2}{\sigma^2} \sim \chi^2_{m-1}, \frac{n\sigma_y^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\mathcal{T} = \frac{\overline{x} - \mu}{\sqrt{\frac{1}{n-1}(\overline{x^2} - (\overline{x})^2)}} \sim t_{n-1}$$



Calculate $\overline{x} - \overline{y}$

$$\frac{\overline{x} - \mu_1}{\sigma} \sim \frac{1}{\sqrt{m}} N(0, 1) = N(0, \frac{1}{m}), \frac{\overline{y} - \mu_2}{\sigma} \sim N(0, \frac{1}{n})$$

$$\frac{\overline{x} - \mu_1}{\sigma} - \frac{\overline{y} - \mu_2}{\sigma} = \frac{(\overline{x} - \overline{y}) - (\mu_1 - \mu_2)}{\sigma} \sim N(0, \frac{1}{m} + \frac{1}{n})$$

$$\frac{(\overline{x} - \overline{y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0, 1)$$

$$\frac{m\sigma_x^2}{\sigma^2} + \frac{n\sigma_y^2}{\sigma^2} \sim \chi^2_{m+n-2}$$

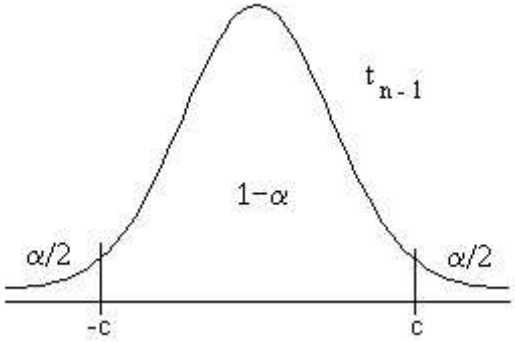
Construct the t-statistic:

$$\frac{N(0,1)}{\sqrt{\frac{1}{m+n-2}(\chi^2_{m+n-2})}} \sim t_{m+n-2}$$

$$\mathcal{T} = \frac{(\overline{x}-\overline{y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}\sqrt{\frac{1}{m+n-2}(\frac{m\sigma_x^2+n\sigma_y^2}{\sigma^2})}} = \frac{(\overline{x}-\overline{y})-(\mu_1-\mu_2)}{\sqrt{(\frac{1}{m}+\frac{1}{n})\frac{1}{m+n-2}(m\sigma_x^2+n\sigma_y^2)}} \sim t_{m+n-2}$$

Construct the test:
$H_1 : \mu_1 = \mu_2, H_2 : \mu_1 \neq \mu_2$
If H_1 is true, then:

$$\mathcal{T} = \frac{\overline{x}-\overline{y}}{\sqrt{(\frac{1}{m}+\frac{1}{n})\frac{1}{m+n-2}(m\sigma_x^2+n\sigma_y^2)}} \sim t_{m+n-2}$$

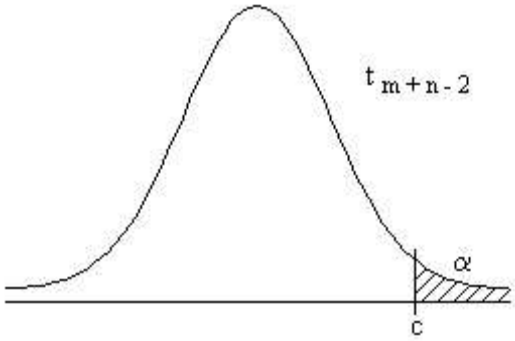


Decision Rule:

$$\delta = \{H_1 : -c \leq T \leq c, H_2 : \text{ otherwise}\}$$

where the c values come from the t distribution with m + n - 2 degrees of freedom.
c = T value where the area is equal to $\alpha/2$, as the failure is both below -c and above +c

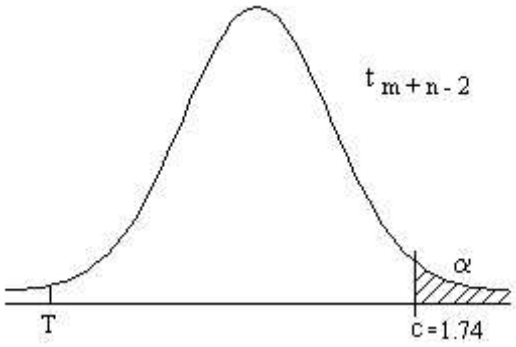
If the test were: $H_1 : \mu_1 \leq \mu_2, H_2 : \mu_1 > \mu_2$,
then the $\mathcal{T}$ value would correspond to an area in one tail, as the failure is only above +c.



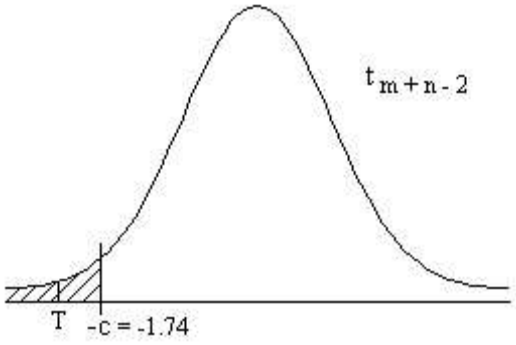
There are different functions you can construct to approach the problem,
based on different combinations of the data.
This is why statistics is entirely based on your assumptions and the resulting

distribution function!

Example: Testing soil types in different locations by amount of aluminum oxide present.
$m = 14, \overline{x} = 12.56 \sim N(\mu_1, \sigma^2); n = 5, \overline{y} = 17.32 \sim N(\mu_2, \sigma^2)$
$H_1 : \mu_1 \leq \mu_2; H_2 : \mu_1 > \mu_2 \rightarrow \mathcal{T} = -6.3 \sim t_{14+5-2=17}$



c-value is 1.74, however this is a one-sided test. T is very negative, but we still accept H_1
If the hypotheses were: $H_1 : \mu_1 \geq \mu_2; H_2 : \mu_1 < \mu_2$,
Then the T value of -6.3 is way to the left of the c-value of -1.74. Reject H_1



Goodness-of-fit tests.
Setup: Consider r different categories for the random variable.
The probability that a data point takes value B_i is p_i
$\sum p_i = p_1 + ... + p_r = 1$
Hypotheses: $H_1 : p_i = p_i^0$ for all i = 1, ..., r; H_2 : otherwise.

Example: (9.1.1)
3 categories exist, regarding a family's financial situation.
They are either worse, better, or the same this year as last year.
Data: Worse = 58, Same = 64, Better = 67 (n = 189)
Hypothesis: $H_1 : p_1 = p_2 = p_3 = \frac{1}{3}, H_2$: otherwise.

N_i = number of observations in each category.
You would expect, under H_1, that $N_1 = np_1, N_2 = np_2, N_3 = np_3$
Measure using the central limit theorem:

$$\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \rightarrow N(0,1)$$

However, keep in mind that the N_i values are not independent!! (they sum to 1)
Ignore part of the scaling to account for this (proof beyond scope):

$$\frac{N_1 - np_1}{\sqrt{np_1}} \to \sqrt{1-p_1}N(0,1) = N(0, 1-p_1)$$

Pearson's Theorem:

$$\mathcal{T} = \frac{(N_1 - np_1)^2}{np_1} + ... + \frac{(N_r - np_r)^2}{np_r} \to \chi^2_{r-1}$$

If H_1 is true, then:

$$\mathcal{T} = \sum_{i=1}^{r} \frac{(N_i - np_i^0)^2}{np_i^0} \to \chi^2_{r-1}$$

If H_1 is not true, then:

$$\mathcal{T} \to +\infty$$

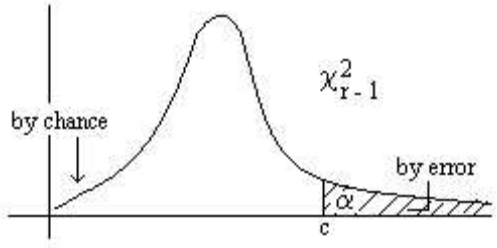
Proof:

$$\text{if } p_1 \neq p_1^0, \frac{N_1 - np_i^0}{\sqrt{np_i^0}} = \frac{N_1 - np_1}{\sqrt{np_1^0}} + \frac{n(p_1 - p_1^0)}{\sqrt{np_i^0}} \to N(0, \sigma^2) + (\pm\infty)$$

However, is squared $\to +\infty$

Decision Rule:

$$\delta = \{H_1 : T \leq c, H_2 : T > c\}$$



The example yields a T value of 0.666, from the $\chi^2_{r-1=3-1=2} = \chi^2_2$
c is much larger, therefore accept H_1.
The difference among the categories is not significant.

** End of Lecture 32