Student t Distribution with 2 Populations

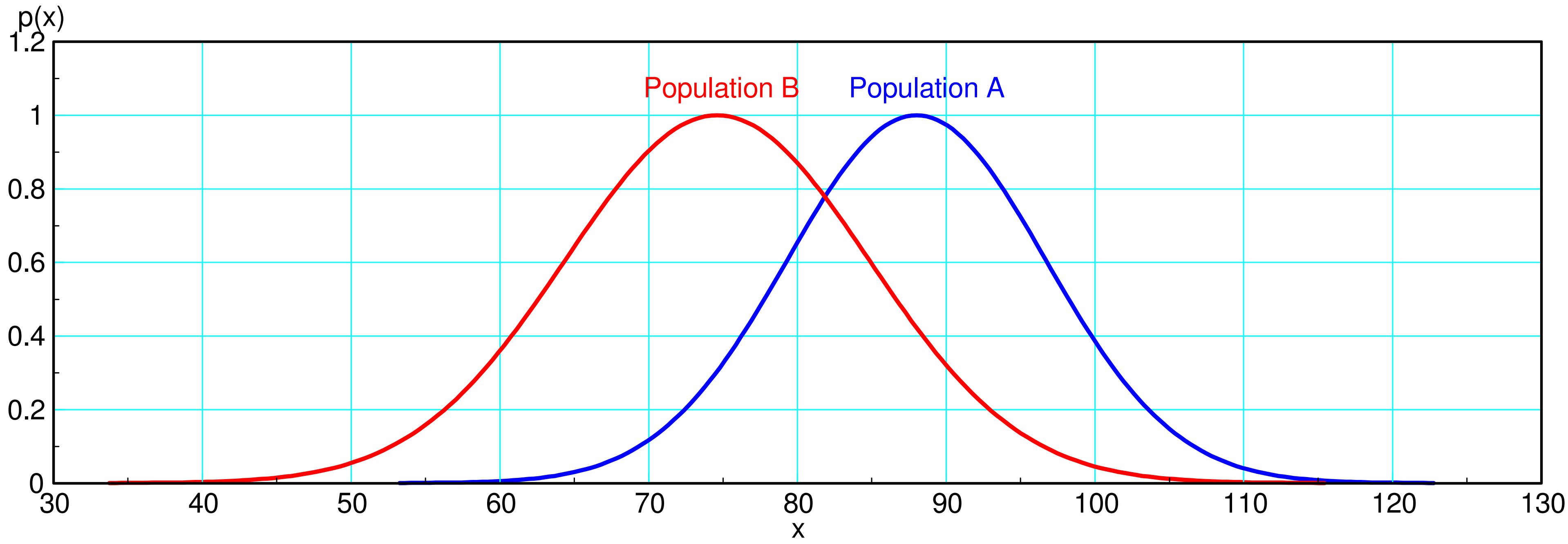
ECE 341: Random Processes

Lecture #24

note: All lecture notes, homework sets, and solutions are posted on www.BisonAcademy.com

Student t Distribution with 2 Populations

- Use a student-t distribution to determine which sample has the higher mean



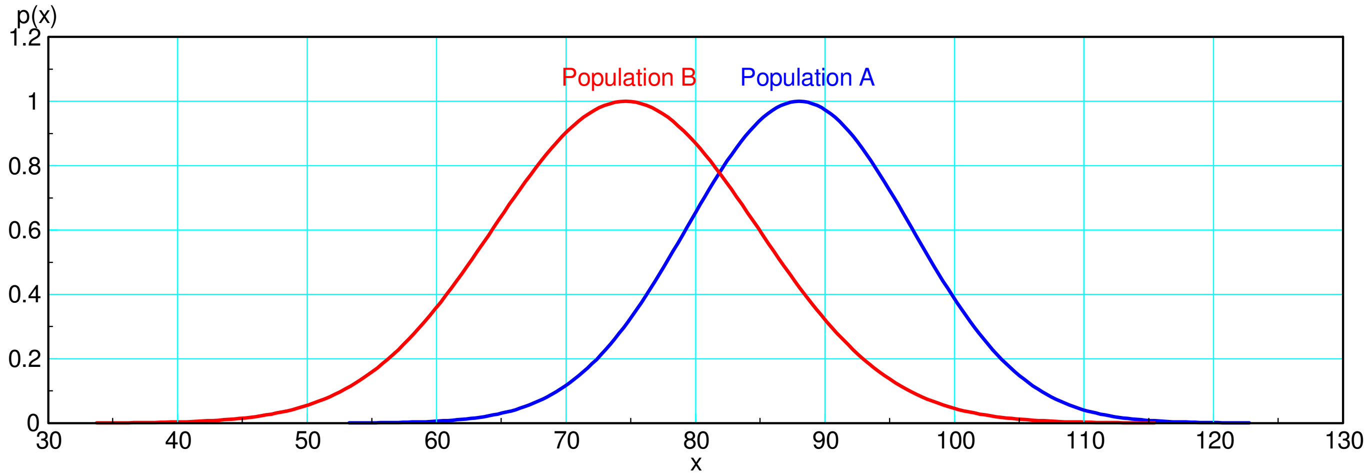
Comparison of Elements (chance to win one game)

Assume:

- A and B be normally distributed distributions with unknown means and variances.
- na and nb samples are taken from population A and B respectfully.

What is the probability that the next value from A will be larger than B?

$$p(x_a > x_b) = ?$$



Example: Hungry-Hungry-Hippo

- Two players, A and B
- Press a button as fast as you can for 5 seconds
- The winner is the person who hit their button the most number of times.

Data: Play 5 games:

A:	78	79	95	94	94
B:	77	68	77	86	75

What is the chance that A will win the next game?

Find the mean and standard deviation:

$$\bar{x}_a = 88.0 \qquad s_a = 8.69$$

$$\bar{x}_b = 76.6 \qquad s_b = 6.43$$

Create a new variable, W = A - B. If A and B were Normal distributions

$$\mu_w = \mu_a - \mu_b$$

$$\sigma_w^2 = \sigma_a^2 + \sigma_b^2$$

A and B have estimated means and variances:

$$\bar{x}_w = \bar{x}_a - \bar{x}_b = 11.40$$

$$s_w^2 = s_a^2 + s_b^2 = 10.81$$

Form the t-score

$$t = \left(\frac{\bar{x}_w}{s_w}\right) = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_a^2 + s_b^2}}\right) = 1.054$$

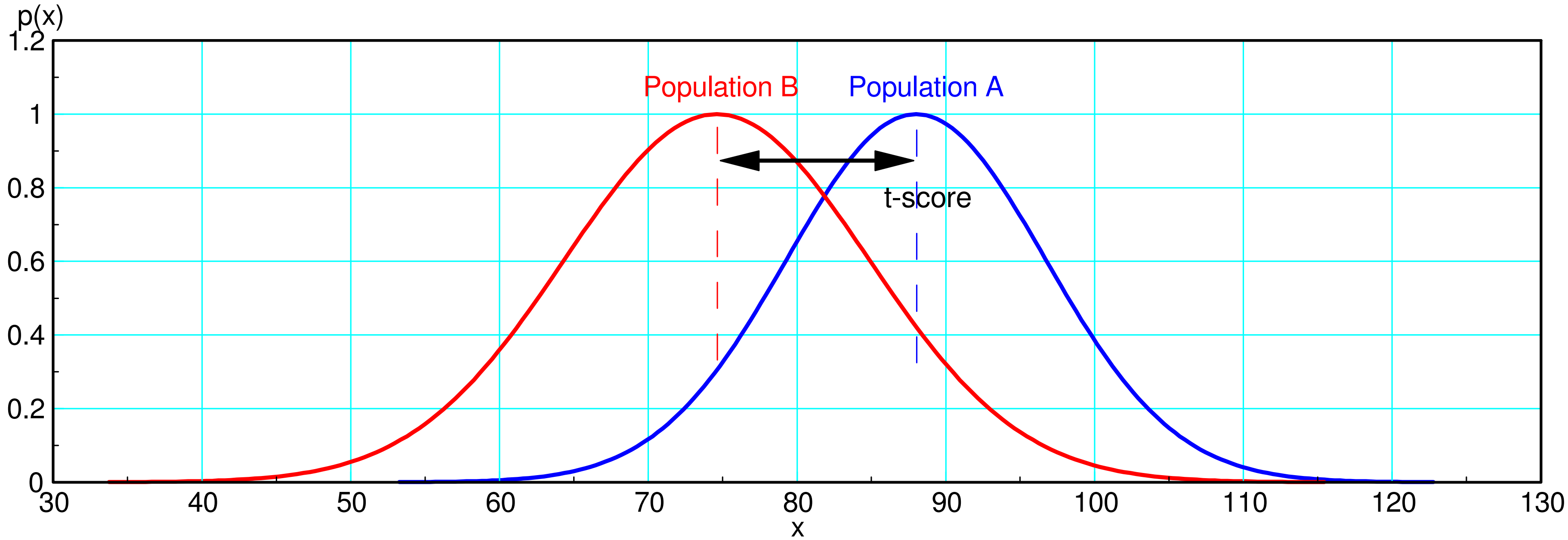
Degrees of Freedom (Wikipedia)

$$d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} = 7.37 \approx 7$$

Convert the t-score to a probability using a t-table

p = 0.8366

Team A has an 83.66% chance of winning the next game.



The probability that the next sample from A will be larger than the next sample from B

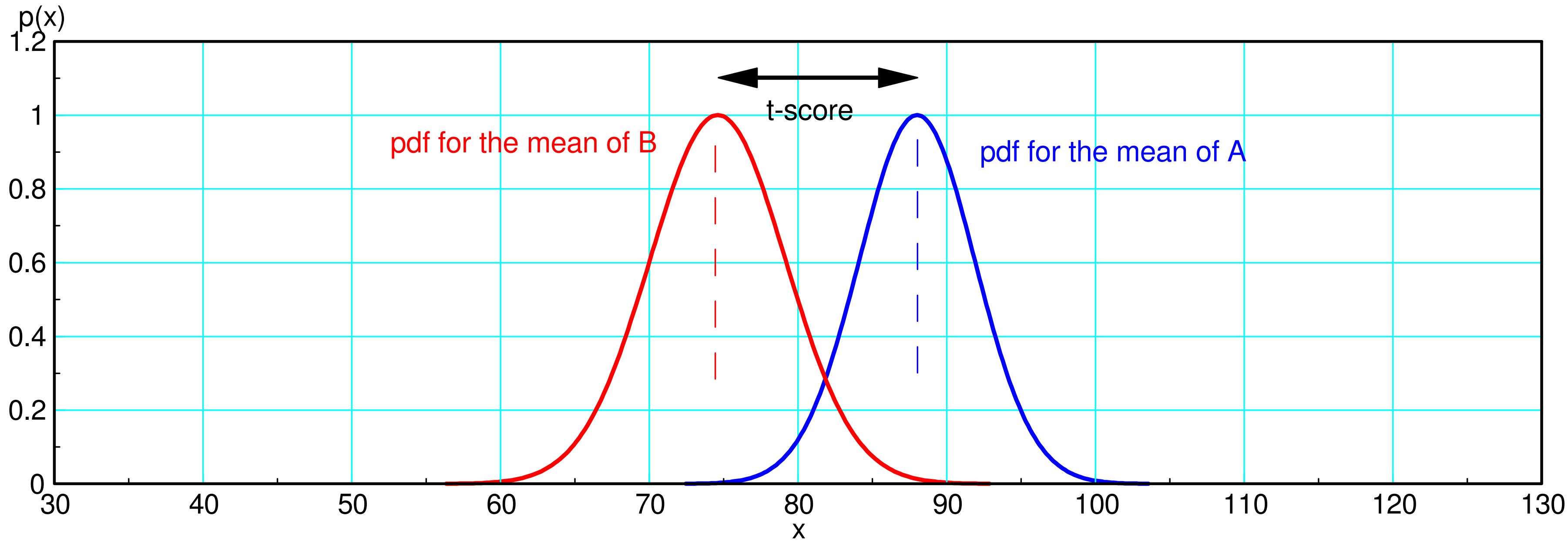
$$t = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_a^2 + s_b^2}} \right) = 0.8366$$

Comparison of Means (chance to win an infinite series)

Which population has the higher mean?

Essentially,

- The previous question was who would win the next game of hungry-hungry hippo.
- This question is who would win a match with an infinite number of games?



Slightly different question: Which population has the larger mean?

Assume first that A and B are normally distributed. The statistic $\bar{x}$ then has a normal distribution

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

As the sample size goes to infinity, the estimate of the mean converges to the true mean.

If you want to compare two means, μ_a and μ_b, then create a new variable:

$$W = \mu_a - \mu_b \qquad W \sim N(\mu_w, \sigma_w^2)$$

where

$$\mu_w = \mu_a - \mu_b$$

$$\sigma_w^2 = \frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}$$

Now, suppose you estimate the variance, creating a student-t distribution. Then

$$W = \bar{x}_a - \bar{x}_b$$

will have a student t-distribution with

$$s_w^2 = \frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}$$

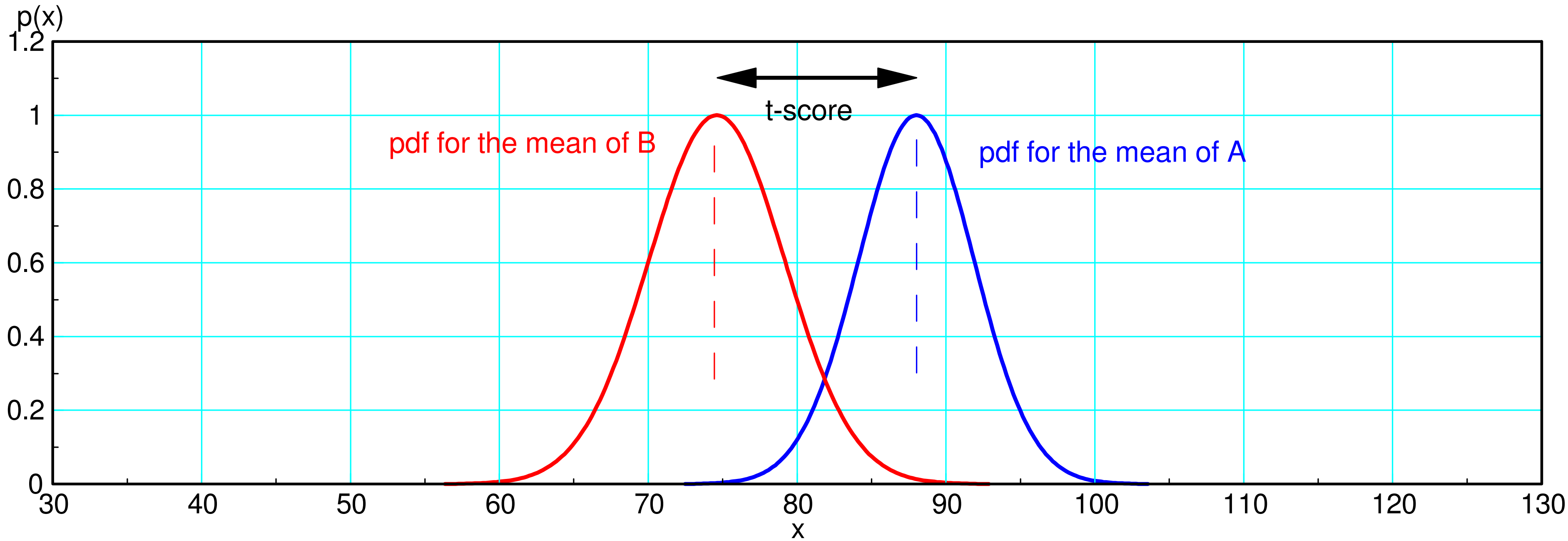
The t-score is then

$$t = \left(\frac{\bar{x}_w}{s_w}\right) = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}\right) = 2.3568$$

Degrees of freedom are the same as before (7)

Use a t-table to convert this to a probability

- t = 2.3568
- 7 degrees of freedom
- p = 0.9747



The probability that the mean of A is greater than the mean of B
t-score = 2.3568, p = 0.9747

Note: You know more about populations than individuals:

- B has a 16.34% chance of winning any given game (previous calculation)
- B only has a 2.53% chance of winning a match

How large does the sample size have to be for being 99.5% certain?

- t-score for 99.5% certain = 3.499
- (assume 7 degrees of freedom)

The t-score is

$$t = 3.499 = \left(\frac{\bar{x}_w}{s_w}\right) = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}\right)$$

If we assume na = nb = n

$$3.499 = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_a^2 + s_b^2}}\right)\sqrt{n} = (1.054)\sqrt{n}$$

$$n = 11.02$$

Round up: n = 12

- A little conservative: > 7 d.o.f. now
- In theory, with a large enough sample size, you can detect minute differences

3-Player Hungry-Hyngry Hippo

- Three players, A, B, and C
- Press a button as fast as you can for 5 seconds
- The winner is the person who hit their button the most number of times.

Data: Play 5 games:

A:	78	79	95	94	94
B:	77	68	77	86	75
C:	75	88	47	79	73

- What is the chance that A will win the next game?
- What is the chance that A is the best player?

Problem: t-Test is only for two populations

Option 1: Change the data so you have two populations:

A:	78	79	95	94	94
max(B,C)	77	88	77	86	75
B:	77	68	77	86	75
C:	75	88	47	79	73

Note:

- Only works if you have access to the raw data
- Order matters
- Must have identical sample sizes for A, B, and C

Create a new variable, W

W = A - max(B, C)

$$\bar{x}_w = \bar{x}_a - \bar{x}_{bc} = 7.400$$

$$s_w = \sqrt{s_a^2 + s_{bc}^2} = 10.5262$$

For a single game

$$t = \frac{x_w}{s_w} = 0.7030$$

With 4 degrees of freedom, p = 73.96%

A has a 73.96% chance of winning any given game

For an infinite game series

$$t = \frac{x_w}{s_w/\sqrt{5}} = 1.5720$$

It is 90.45% likely that A is the best player

Option 2: Run two separate t-tests

- A vs. B
- A vs. C

Correct result if this is single-elimination playoff

- A must defeat B then
- A must defeat C

Underestiates A's chances if it's a single game with three players

- Gets even worse if there are more than 3 players

Example:

p(A) winning the match:

$$\bar{x}_a = 88.00 \qquad s_a = 8.6891$$

$$\bar{x}_b = 76.60 \qquad s_b = 6.4265$$

$$\bar{x}_c = 72.40 \qquad s_c = 15.3232$$

A vs B

$$t_{ab} = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_a^2 + s_b^2}} \right) = 1.0548$$

$$p_{ab} = 82.45\%$$

A vs. C

$$t_{ac} = \left(\frac{\bar{x}_a - \bar{x}_c}{\sqrt{s_a^2 + s_c^2}} \right) = 0.8856$$

$$p_{ac} = 78.71\%$$

$$p_{ab} \cdot p_{ac} = 64.90\% \qquad \textit{vs. 73.96\% from before}$$

Summary

With a t-test, you can compare two populations

- Create a new variable, W = A - B
- Determine the probability that W > 0

Only really works with two populations

- If you have more than two populations, you need a different tool
- ANOVA is one such tool (upcoming....)