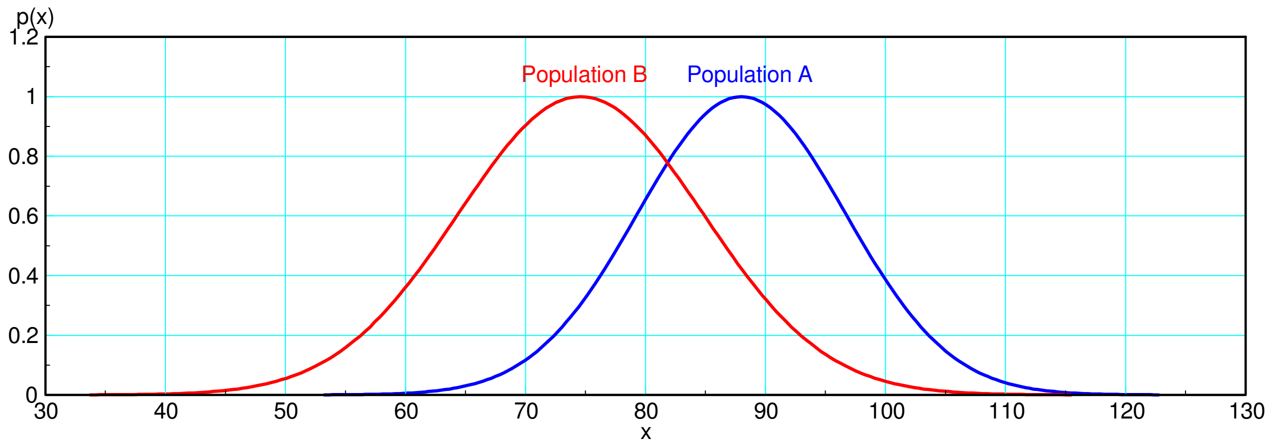
Student t Distribution with 2 Populations

Objectives:

- 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(xa > xb) = ?

Example: Two people are playing hungry-hungry hippo (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.)

In the last 5 games, the score for A and B were:

A:	78	79	95	94	94	$\bar{x}_a = 88.0$	$s_a = 8.69$
B:	77	68	77	86	75	$\bar{x}_b = 76.6$	$s_b = 6.43$

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

Solution: Create a new variable W

W = A - B

If A and B were normally distributed, then W would be normally distributed as well with

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

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

Actually, since the mean and variance are unknown but estimated from the data, A and B have a student t distribution with

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

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

for a t-score of

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

Now this is where it gets tricky: how many degrees of freedom does w have ?

If na were infinity, you would know μ_a precisely (the standard deviation drops as the square root of sample size). In that case, this reverts to the previous lecture: comparison of population B to a constant. The degrees of freedom are then

na = infinity

d.f. = nb - 1

Similarly, if nb were infinity, then

nb = infinity

d.f. = na - 1

If na = nb, you could have

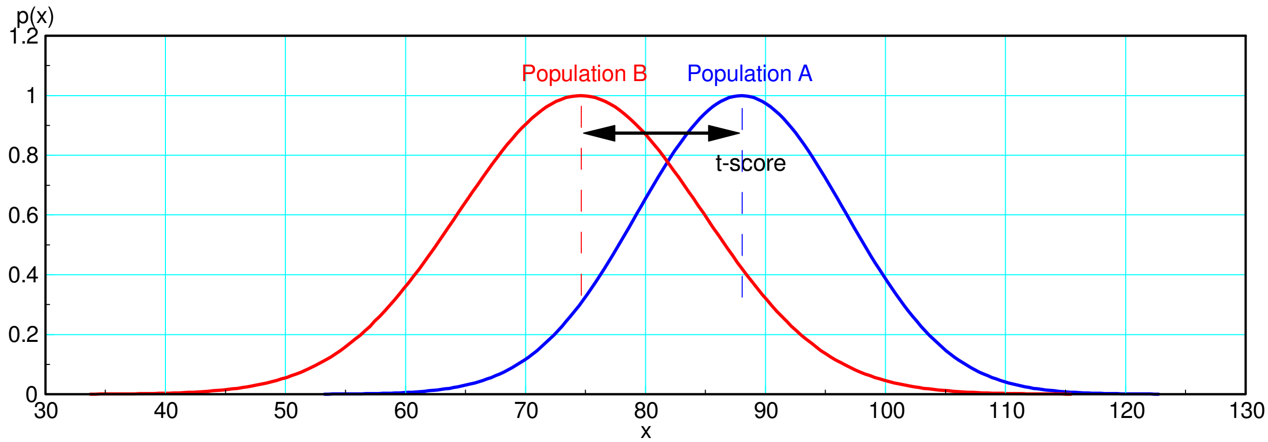
- na + nb degrees of freedom (treat this as one big data set), or
- na degrees of freedom (subtract xb from xa in a one-to-one mapping to W)

The actual degrees of freedom are (from Wikipedia)

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

With this, you can convert the t-score (1.054) to a probability using a t-table (or StatTrek) 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 is 0.8366

$$t = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_a^2 + s_b^2}} \right)$$ = 1.054 with 7 degrees of freedom

Comparison of Means (chance to win an infinite series)

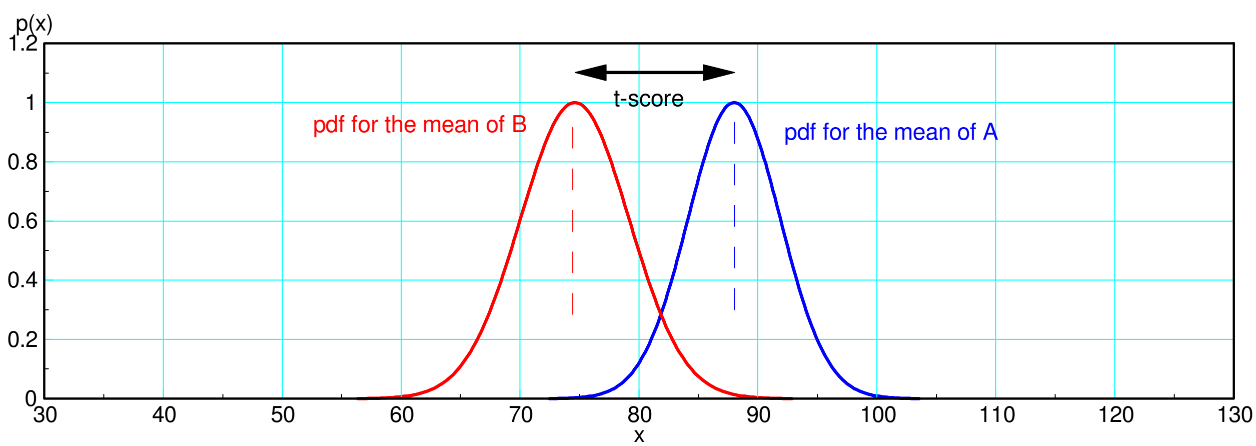
A slightly different question is

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?

With an infinite number of games, even the slightest edge would eventually be telling.



Slightly different question: Which population has the larger mean?
note that the standard deviation for the mean drops by the square root of the sample size

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:

$$W = \mu_a - \mu_b$$

W is also normally distributed

$$W \sim N(\mu_w, \sigma_w^2)$$

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

$$t = \left(\frac{\bar{x}_w}{s_w}\right) = \left(\frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}\right) = 2.358$$

Once again, the degrees of freedom are

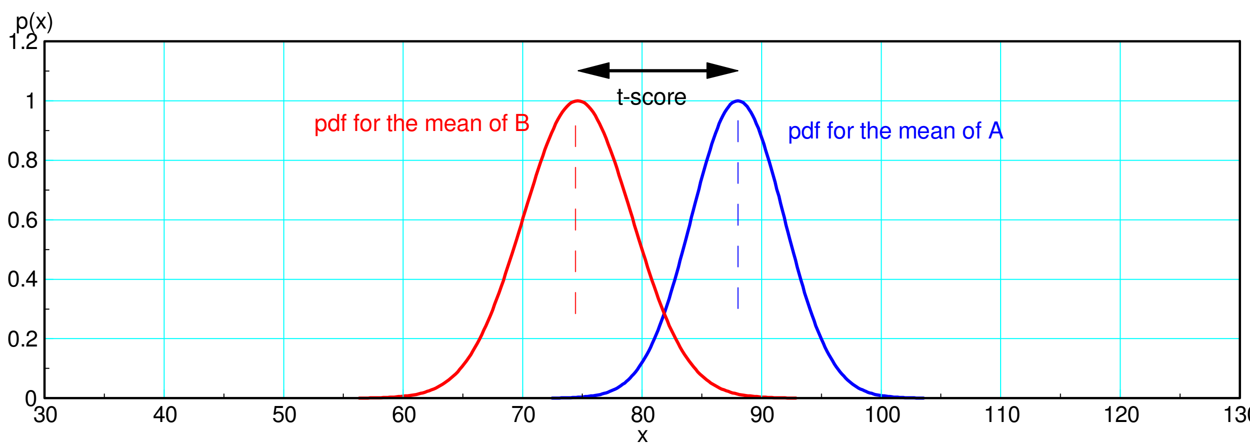
$$d.f. = \frac{\left(\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)\right)}{\left(\frac{\left(s_1^2/n_1\right)}{n_1 - 1}\right) + \left(\frac{\left(s_2^2/n_2\right)}{n_2 - 1}\right)} = 7.37 \approx 7$$

Using a t-table, a t-score of 2.3568 corresponds to a probability of 0.9747

There is a 97.47% chance that Team A has the higher mean (and would win an infinite series).

Hungry-Hungry Hippo Example:

```
Xw = mean(A) - mean(B)

Xw =   11.4000

Sw = sqrt( var(A)/5 + var(B)/5 )

Sw =    4.8343

t = Xw / Sw

t =   2.358
```



The probability that the mean of A is greater than the mean of B is 0.9747
t-score = 2.358 with 7 degrees of freedom

Note: You know more about populations than individuals:

- B has a 19% chance of winning any given game (previous calculation)
- B only has a 4.46% chance of winning a match

Also note that that you can calculate the sample size necessary to be 99% certain that the mean of A is larger than the mean of B

From StatTrek, to be 99.5% certain, the means have to be 3.499 standard deviations apart (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 to n = 12.

This is a little conservative since if you have a sample size of 12, the degrees of freedom increase, which changes the t-score needed.

Also note that as the sample size goes up, the t-score goes up as $\sqrt{n}$. In theory, you can see minute differences in the means if the sample size is large enough.