Student's t

Computational Social Intelligence - Lecture 06

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

http://www.dcs.gla.ac.uk/vincia
Alessandro.Vinciarelli@glasgow.ac.uk

University of Glasgow

EPSRC
Engineering and Physical Sciences Research Council

FNSNF

This lecture is based on the following text (available on Moodle):

- D.C.Howell, "Statistical Methods for Psychology", Chapter 7, Sections 7.1, 7.2, 7.3 (until page 186 included), 7.5 (until page 203 included), Cengage Learning, 2009.

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison

Carl Friedrich Gauss
1777-1855

The Gaussian (or Normal) pdf

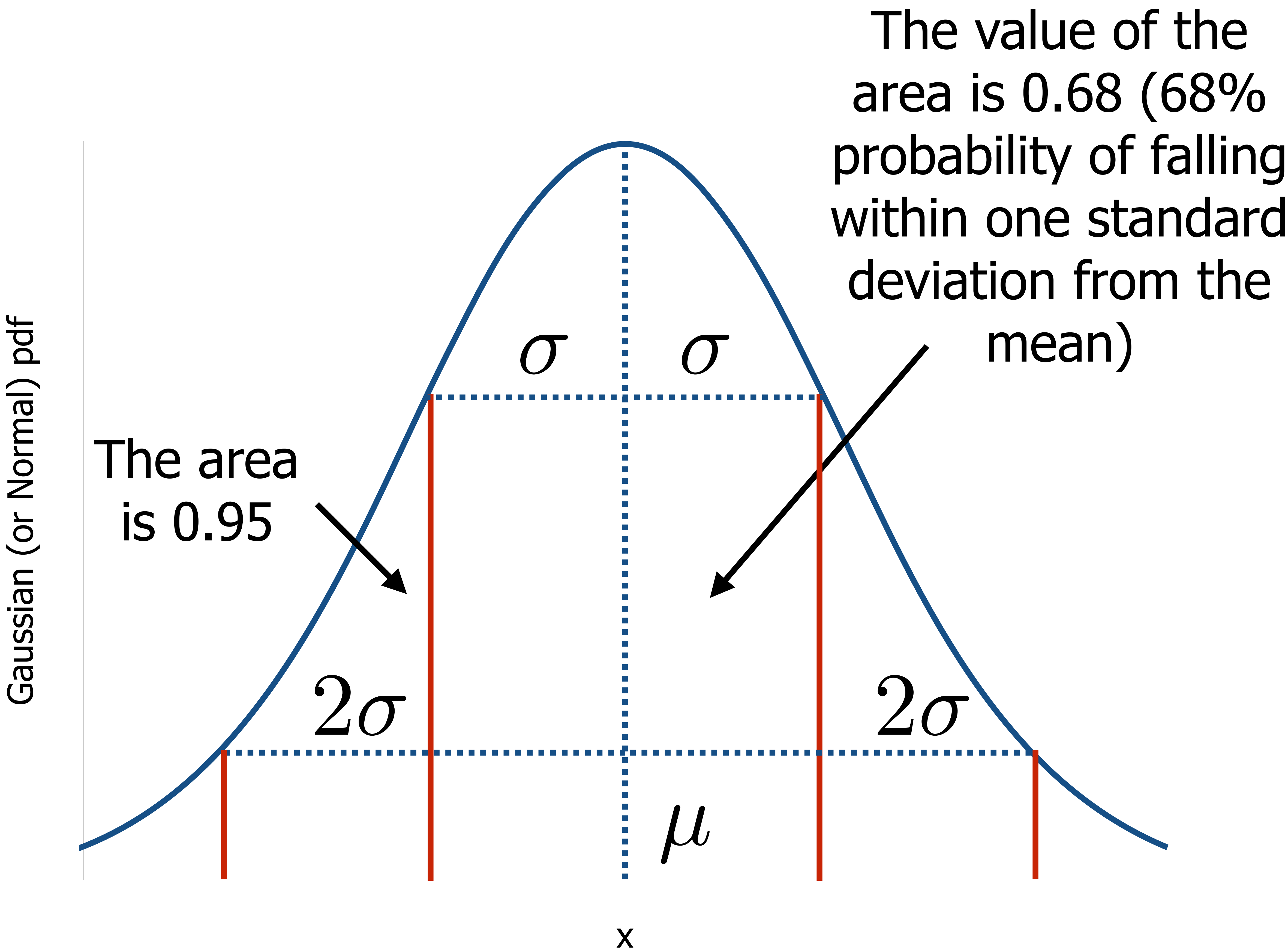
"x" is a random variable

The mean

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The variance

The standard deviation



The value of the area is 0.68 (68% probability of falling within one standard deviation from the mean)
The area is 0.95
σ
σ
2σ
2σ
μ
Gaussian (or Normal) pdf
x

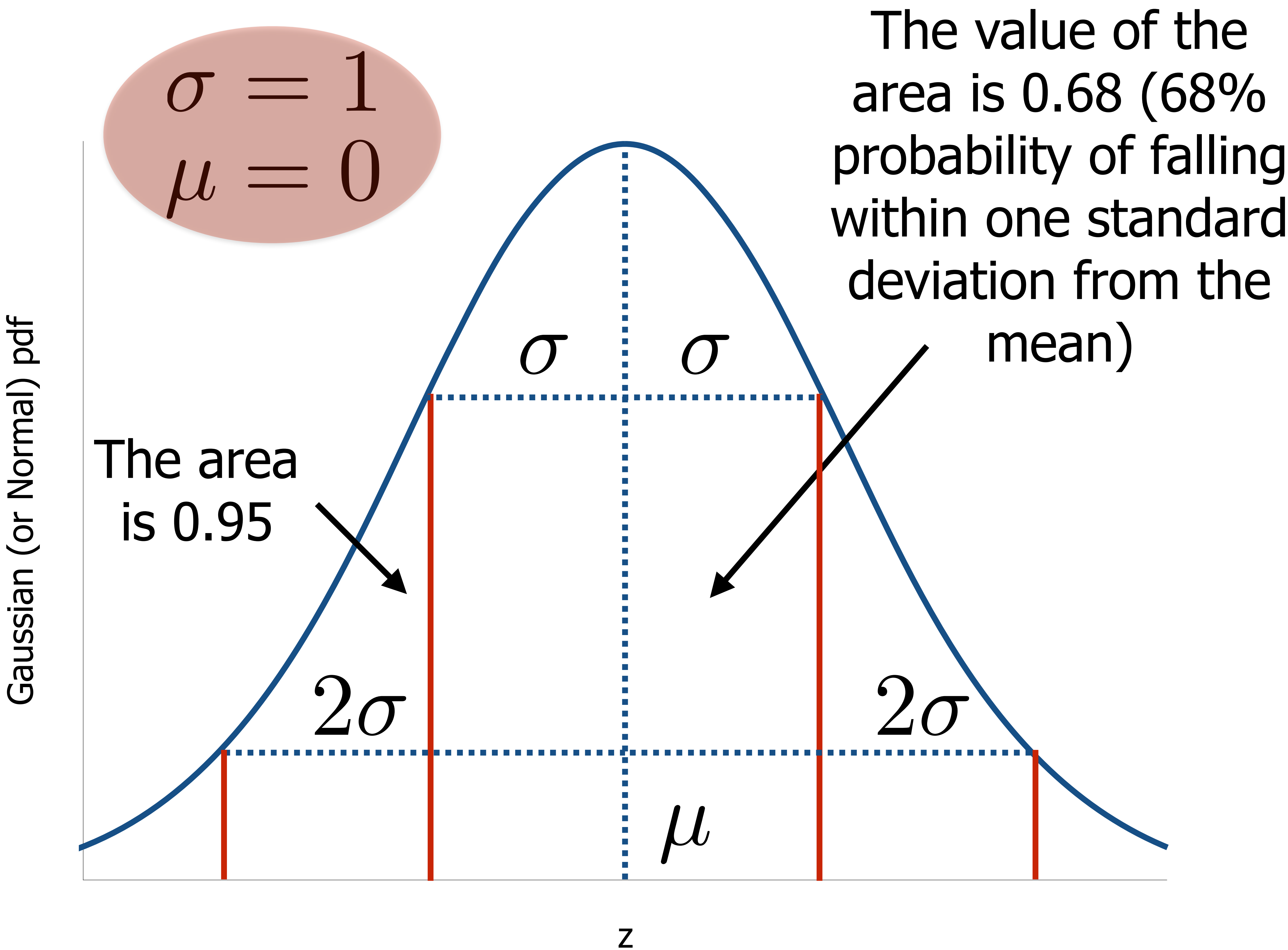
"z" is a random variable known as the z-transform of x

"x" is a random variable that follows a normal pdf

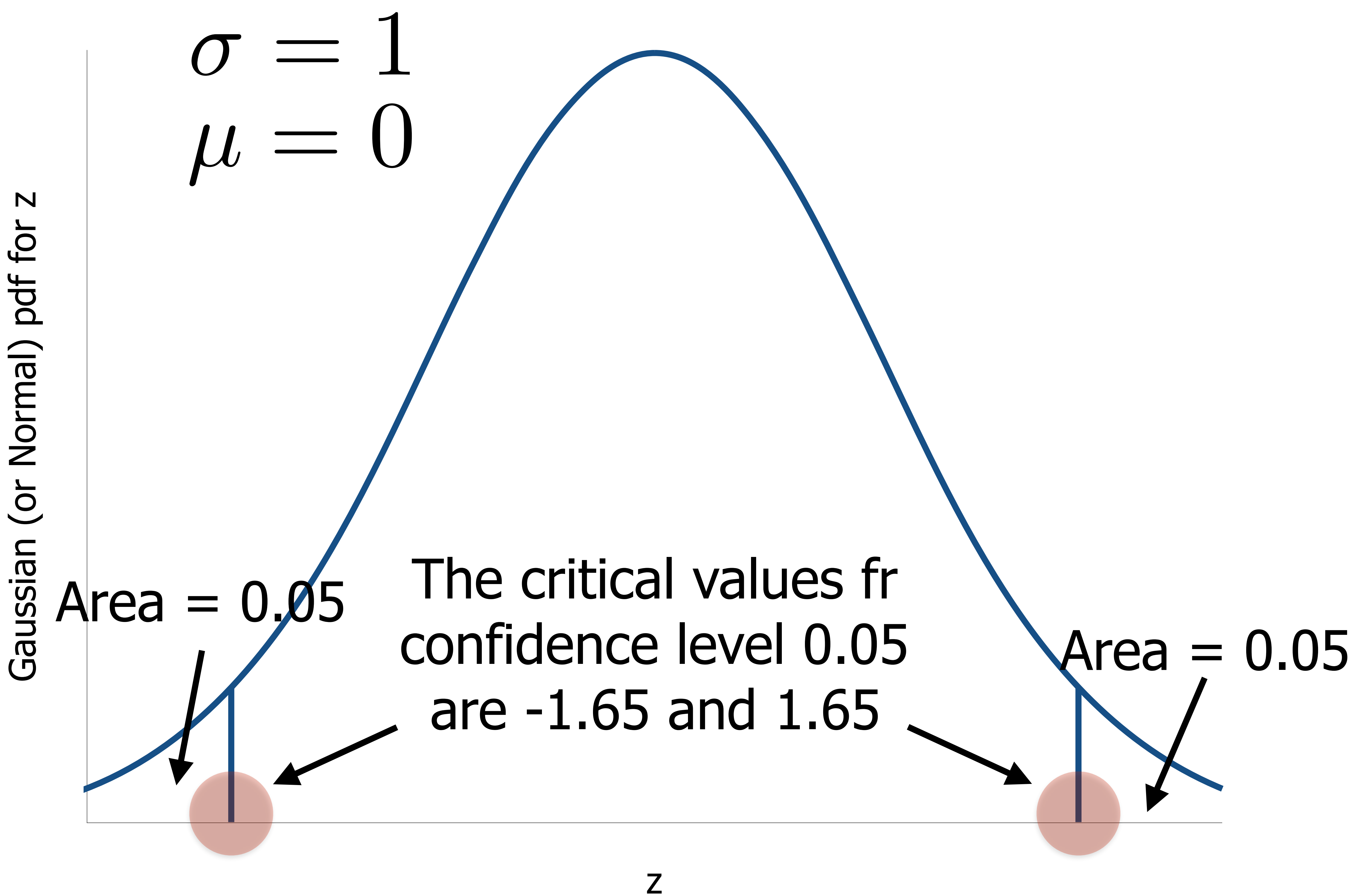
$$z = \frac{x - \mu}{\sigma}$$

The standard deviation of the normal distribution that variable "x" follows

The mean of the normal distribution that variable "x" follows



$\sigma = 1$
$\mu = 0$
The value of the area is 0.68 (68% probability of falling within one standard deviation from the mean)
The area is 0.95
σ
σ
2σ
2σ
μ
Gaussian (or Normal) pdf
z



$\sigma = 1$
$\mu = 0$
Gaussian (or Normal) pdf for z
Area = 0.05
The critical values fr
confidence level 0.05
are -1.65 and 1.65
Area = 0.05
z

(Toy) Research Hypothesis

- If the Gaussian is a sampling distribution, it can be used for hypothesis testing;
- Research Hypothesis: A person with temperature 38C has fever;
- Null Hypothesis: A person with temperature 38C does not have fever;
- Mean and variance for the people that have no fever are 37.0 and 0.01, respectively.

The values of variable, mean and standard deviation are inserted

The variable value to be tested (temperature in the toy example)

$$z = \frac{x - \mu}{\sigma} = \frac{38.0 - 37.0}{0.1} = 10.0$$

The value of z is above the critical value, the null hypothesis can be rejected

The equation can help to find the variable value (temperature) beyond which the null hypothesis can be rejected

$$z \geq 1.65 \Rightarrow \frac{x - \mu}{\sigma} \geq 1.65$$

$$x \geq \mu + 1.65 \cdot \sigma \Rightarrow x \geq 37.165$$

The (toy) temperature threshold beyond which the null hypothesis can be rejected

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison

Central Limit Theorem

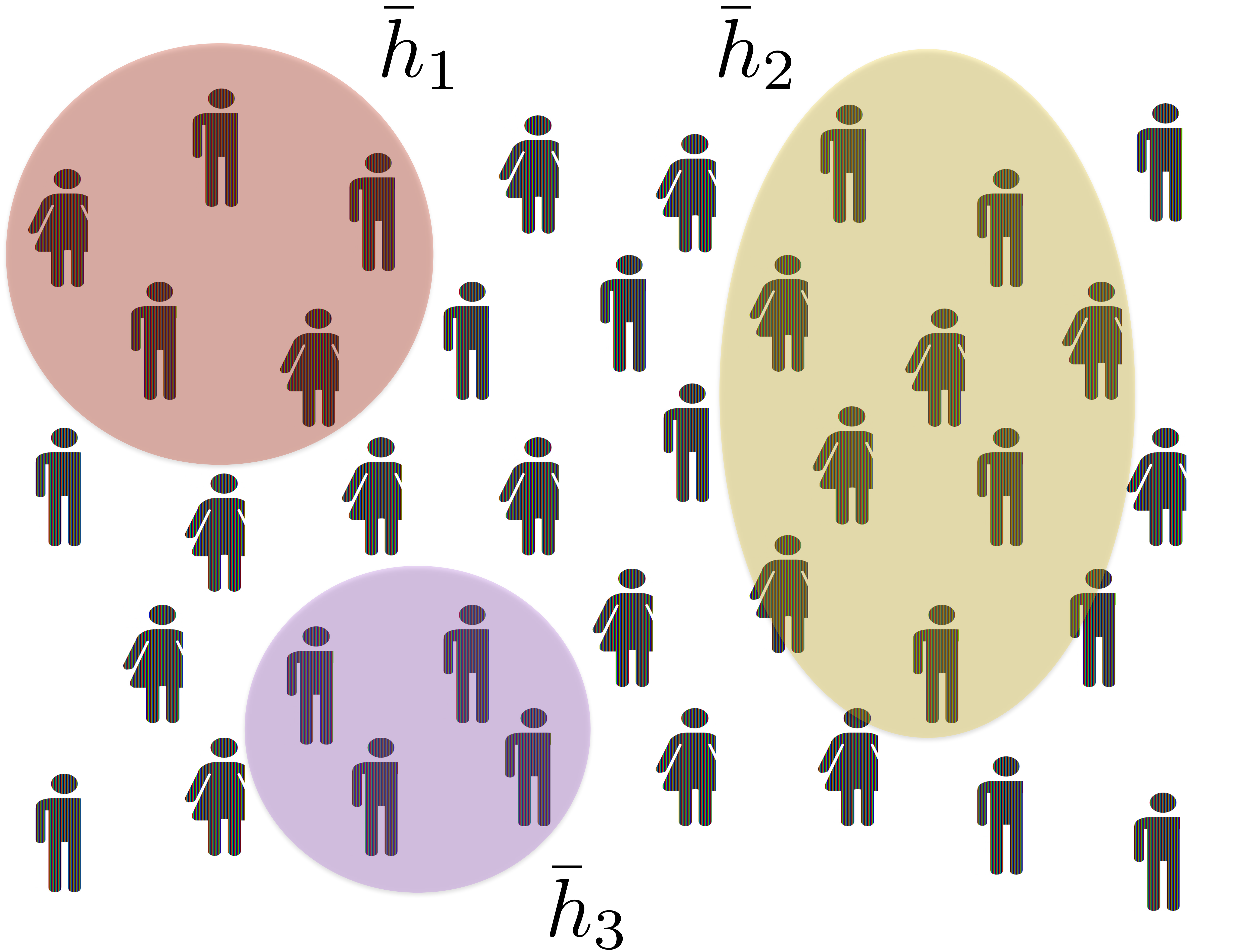
Given a population with mean μ and variance σ^2, the sampling distribution of the mean will have mean $\mu_{\bar{x}} = \mu$ and variance $\sigma_{\bar{x}}^2 = \sigma^2/n$. The distribution will approach the normal distribution as the sample size n increases.

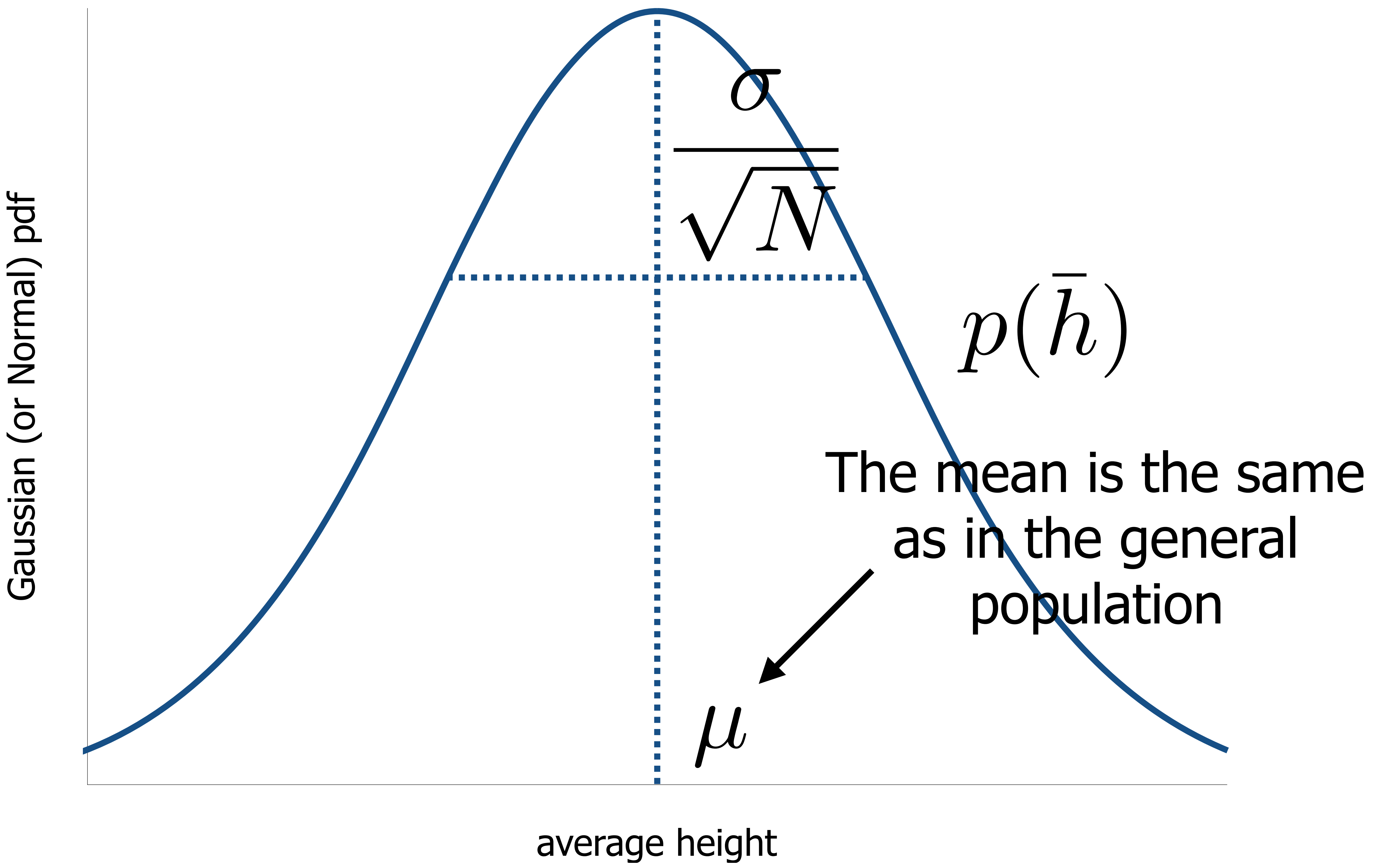
D.C.Howell, "Statistical Methods for Psychology", Chapter 7, Cengage Learning, 2009.

The height of individual i in the population

$$\bar{h} = \frac{1}{N}\sum_{i=1}^{N} h_i$$

The average height of the individuals in the population


$\bar{h}_1$
$\bar{h}_2$
$\bar{h}_3$



$\frac{\sigma}{\sqrt{N}}$
$p(\bar{h})$
The mean is the same as in the general population
μ
Gaussian (or Normal) pdf
average height

It is possible to apply the z-transform to the average age

It is still necessary to know mean and variance of the general population

$$z = \frac{\bar{h} - \mu}{\frac{\sigma}{\sqrt{N}}} \geq 1.65$$

For a given average height, the number of samples increases the value of z

Recap

- The consequence of the Central Limit Theorem is that the sampling distribution of a sample average is a Gaussian distribution;
- It is possible to apply the z-transform and test whether the z value is beyond the critical threshold (1.65 for confidence level 0.05);
- It is still necessary to know mean and variance of the general population (rare in practice).

The height of individual i in the population

$$\bar{h} = \frac{1}{N}\sum_{i=1}^{N} h_i$$

The average height of the individuals in the population

The sample variance

The N basketball players

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(h_i - \bar{h})^2$$

We still need to know the mean

The variance is replaced with the sample variance

$$z = \frac{\bar{h} - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{h} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{\bar{h} - \mu}{s}\sqrt{N}$$

For a given average height, the number of samples increases the value of z

The Student's t random variable

The sampling distribution of t when the null hypothesis is true is known

$$t = \frac{\bar{h} - \mu}{\sqrt{\frac{s^2}{N}}}$$

The pdf of t (the sampling distribution) when the Null Hypothesis is true

$$p(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

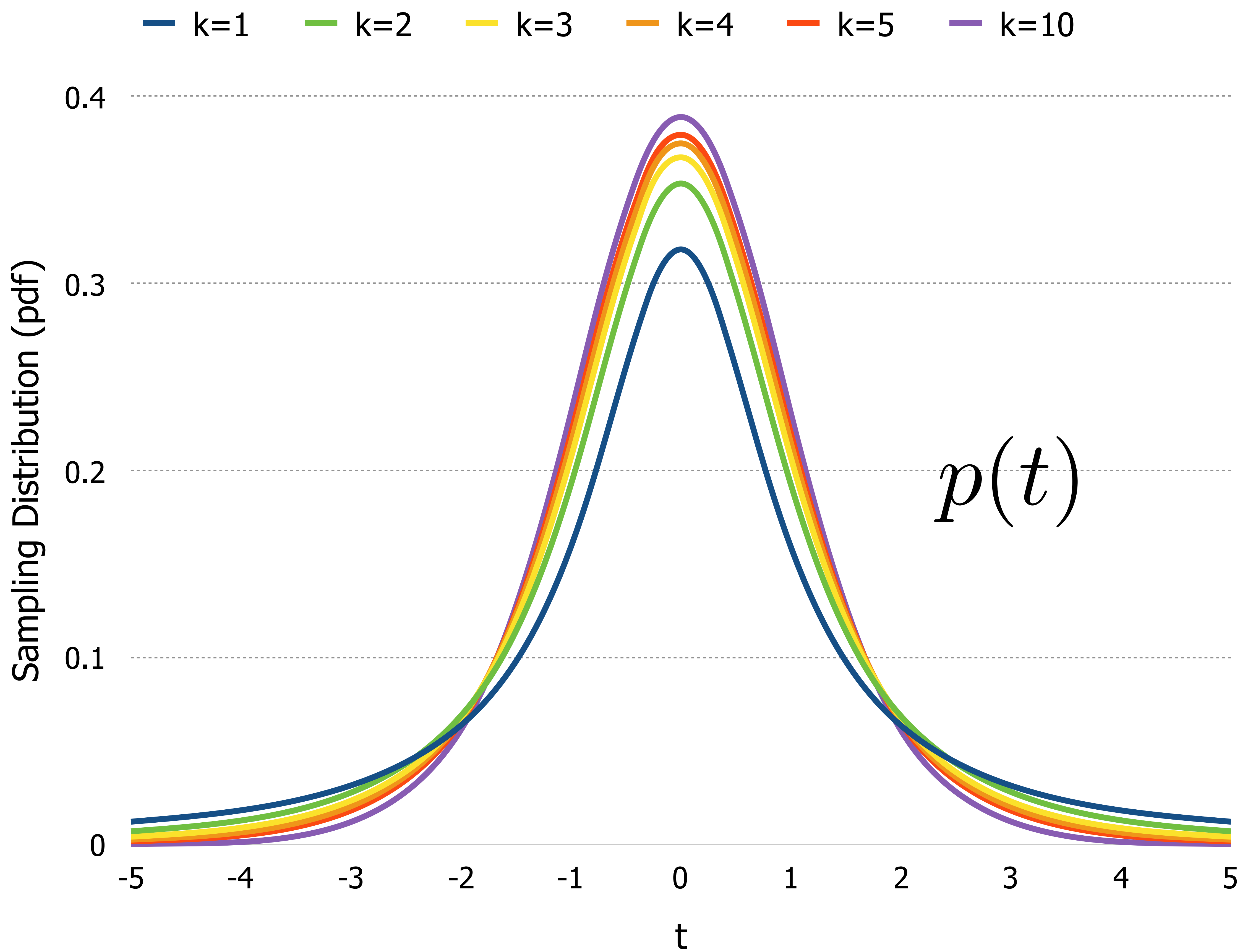
The parameter k is the number of degrees of freedom

The value of the sample variance is a constraint to be respected

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(h_i - \bar{h})^2$$

$$k = N - 1$$

Once the first N-1 values of height have been set, the last can be obtained from the sample variance


k=1
k=2
k=3
k=4
k=5
k=10
0.4
0.3
0.2
0.1
0
Sampling Distribution (pdf)
$p(t)$
-5
-4
-3
-2
-1
0
1
2
3
4
5
t

William Sealy Gosset
1876-1937

Recap

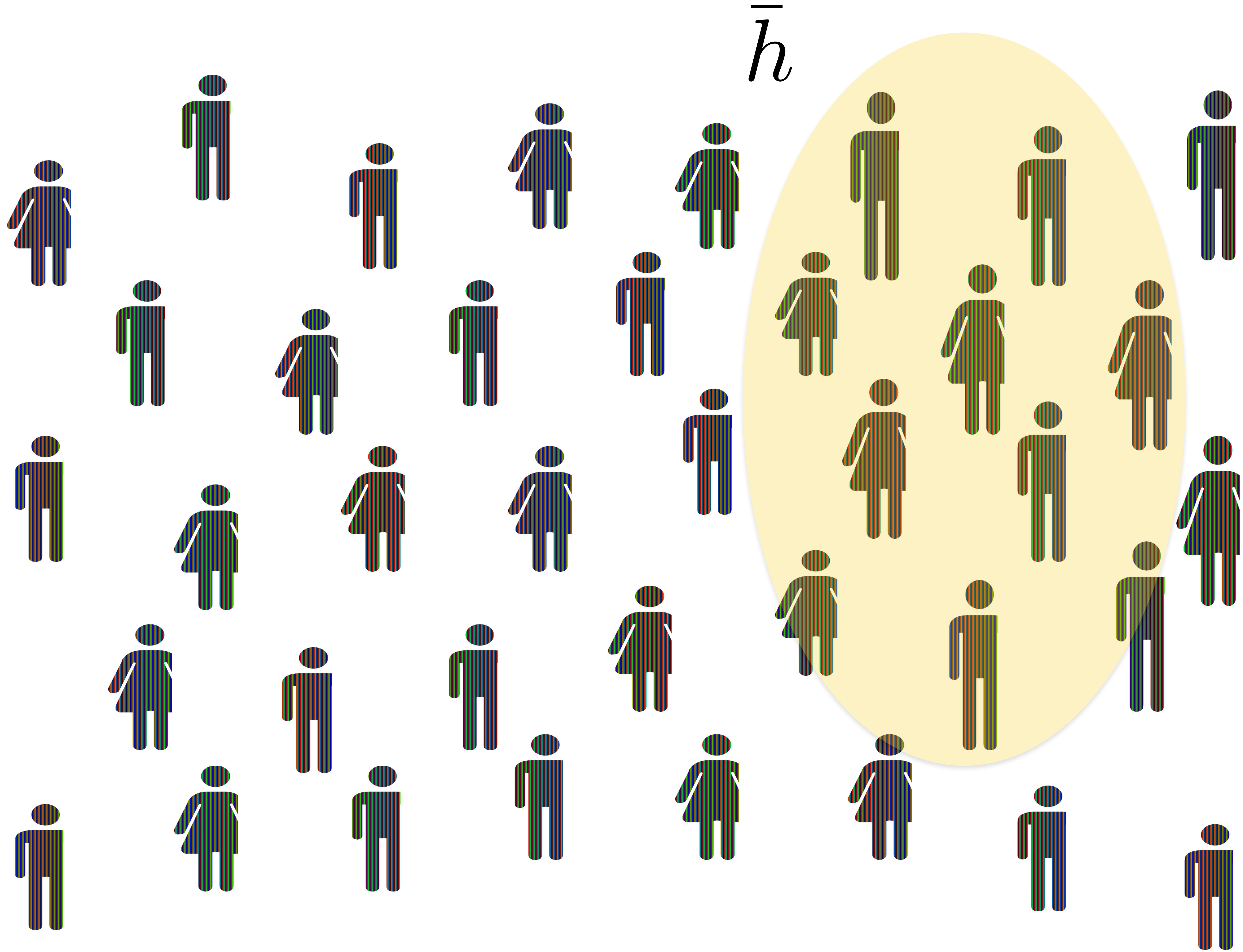
- The Student's t and its sampling distribution allow one to perform hypothesis testing when the mean of the population is known;
- When t (a function of the data) is beyond the critical value corresponding to a confidence level, it is possible to reject the NH;

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison

Research Hypothesis

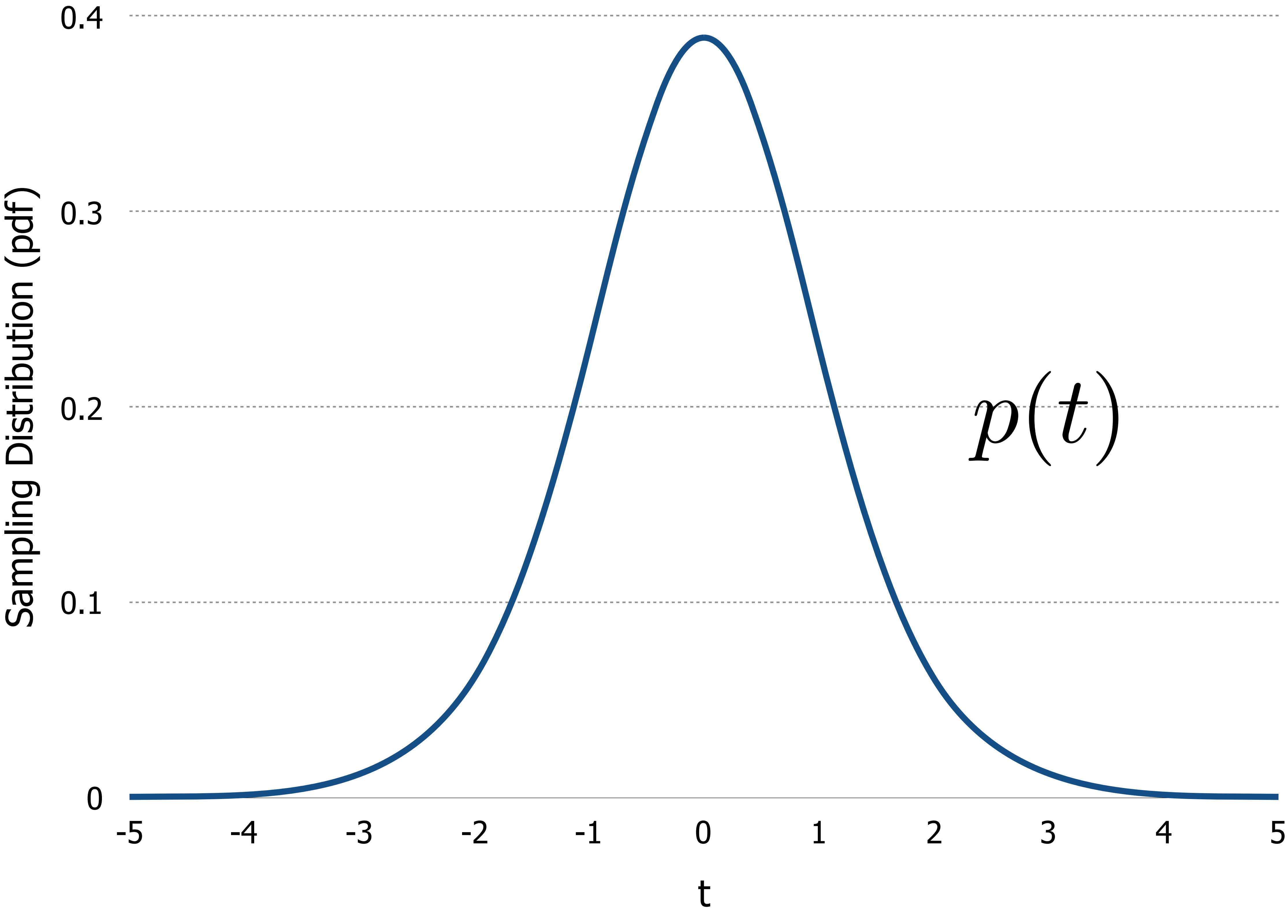
- Research Hypothesis: Basketball players are, on average, taller than the rest of the population;
- Null Hypothesis: Basketball players are, on average, as tall as the rest of the population.


$\bar{h}$

The Student's t random variable

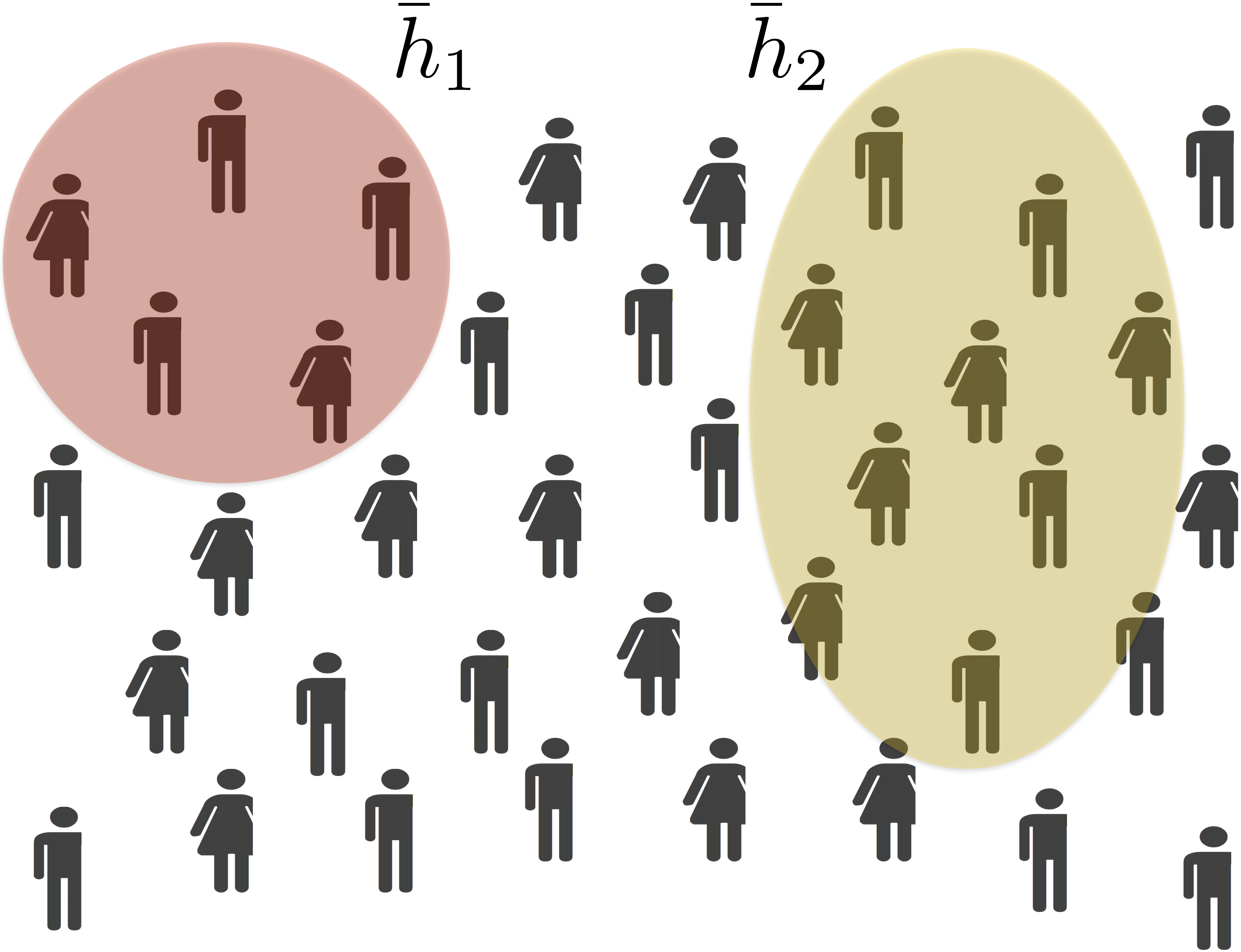
The sampling distribution of t when the null hypothesis is true is known

$$t = \frac{\bar{h} - \mu}{\sqrt{\frac{s^2}{N}}}$$


0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
t
Sampling Distribution (pdf)
$p(t)$

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison


$\bar{h}_1$
$\bar{h}_2$

Variance Sum Law

“The variance of a sum or difference of two independent variables is equal to the sum of their variances.”

D.C.Howell, “Statistical Methods for Psychology”, Chapter 7, Cengage Learning, 2009.

The random variable is the difference between the means

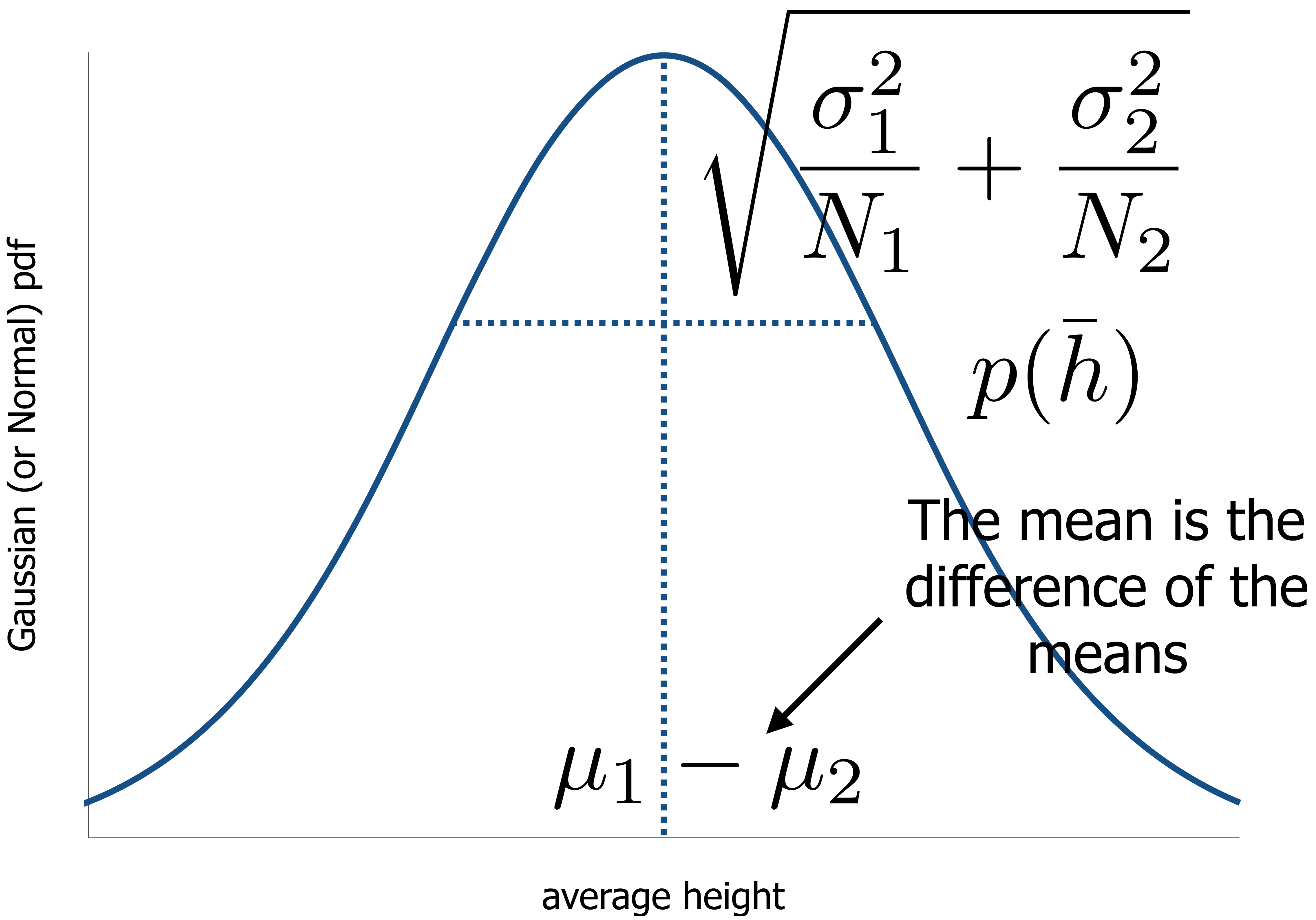
The mean of the sampling distribution (Central Limit Theorem)

$$\bar{h} = \mu_1 - \mu_2 = \bar{h}_1 - \bar{h}_2$$

$$\sigma^2_{\bar{h}} = \sigma^2_{\bar{h}_1} + \sigma^2_{\bar{h}_2} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

The variance of the difference between the means

The Variance Sum Law



$\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$
$p(\bar{h})$
The mean is the difference of the means
$\mu_1 - \mu_2$
Gaussian (or Normal) pdf
average height

The Null Hypothesis is that such a difference is null

Student's t

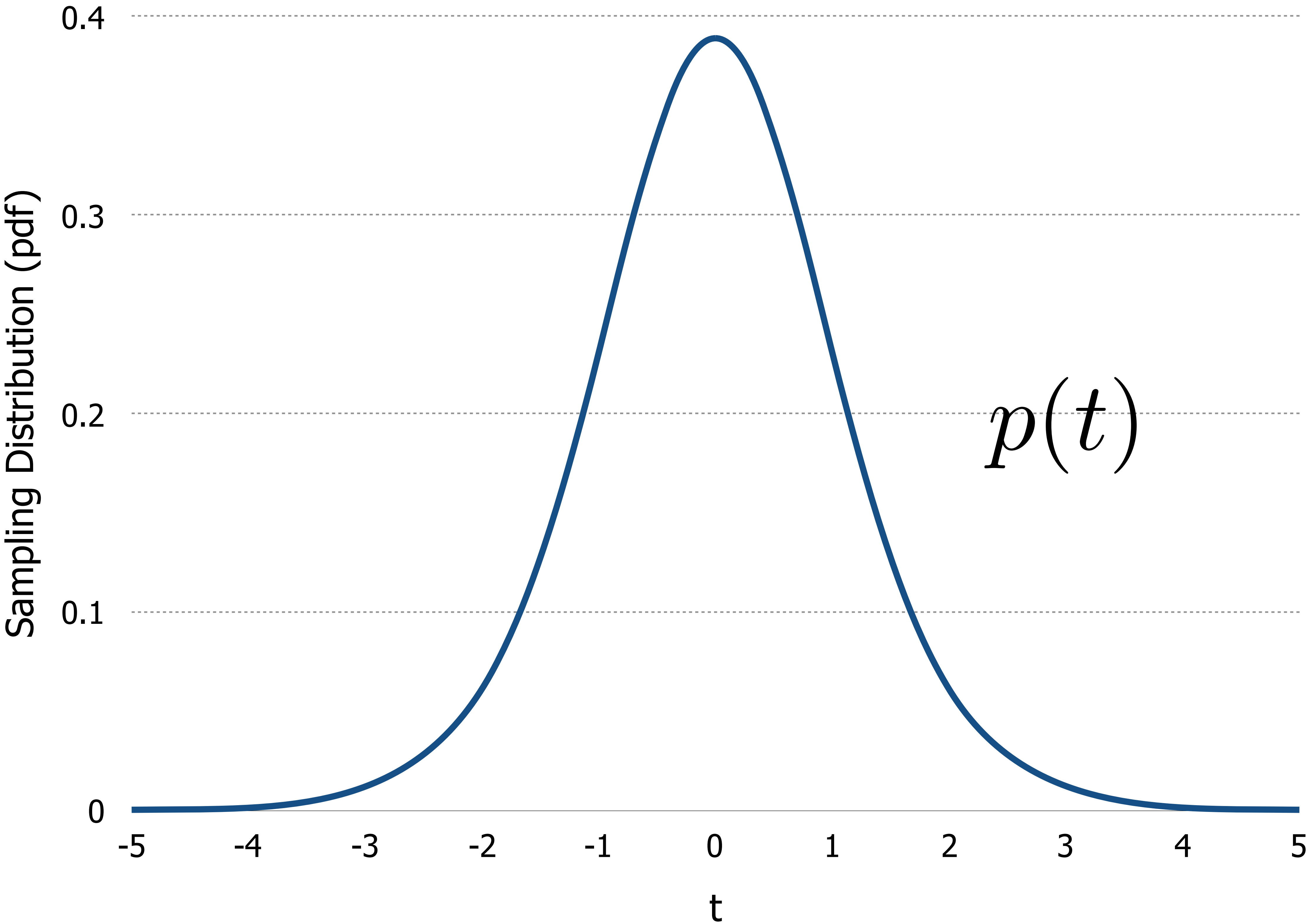
$$t = \frac{\bar{h}_1 - \bar{h}_2 - \mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Student's t

$$t = \frac{\bar{h}_1 - \bar{h}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$k = (N_1 - 1) + (N_2 - 1)$$

The degrees of freedom is the sum of the individual degrees of freedom



0.4
0.3
0.2
0.1
0
Sampling Distribution (pdf)
-5
-4
-3
-2
-1
0
1
2
3
4
5
t
$p(t)$

Research Hypothesis

- Research Hypothesis: The difference between the means is different from zero;
- Null Hypothesis: The difference between the means is null;
- The name "Null Hypothesis" comes from this case.

Recap

- The Student's t and its sampling distribution allow one to perform hypothesis testing when the mean of the population is known;

- When the statistic is the difference between the means extracted from different samples, then the mean of the population is null.

Thank You!