

جامعة نيويورك أبوظبي

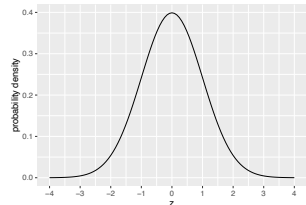



# PSYCH-UH 1004Q: Statistics for Psychology


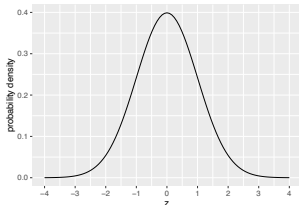


## Class 10: one-sample $t$ -test

Prof. Jon Sprouse  
Psychology

# From z-tests to *t*-tests

	<b>(one sample) z-test</b>	<b>one sample <i>t</i>-test</b>
<b>Scientific question</b>	Does our sample differ from a population with a <b>known mean</b> and <b>standard deviation</b> ?	Does our sample differ from a population with a <b>known mean</b> (but <b>unknown SD</b> )?
<b>Null Hypothesis</b>	The mean of the population that the sample comes from is equal to the mean of the known population (so, $\mu = \mu_0$ )	The mean of the population that the sample comes from is equal to the mean of the known population (so, $\mu = \mu_0$ )
<b>Equation</b>	$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$	$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$
<b>Descriptive information</b>	The <i>z</i> statistic tells us how much our sample mean differs from the population mean in terms of <b>population SE</b>	The <i>t</i> statistic tells us how much our sample mean differs from the population mean in terms of <b>sample SE</b> (as an estimate)
<b>Null distribution</b>	z-distribution 	<i>t</i> -distribution 

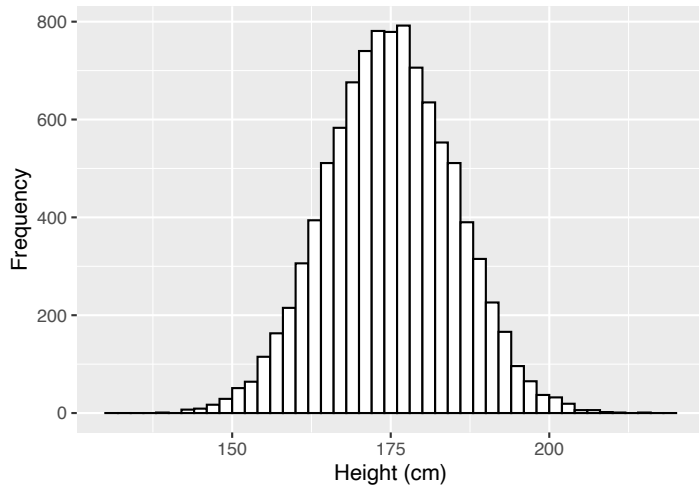
# From z-tests to *t*-tests

	(one sample) z-test	one sample <i>t</i> -test
<b>Scientific question</b>	Does our sample differ from a population with a known mean and standard deviation?	Does our sample differ from a population with a known mean (but <b>unknown SD</b> )?
<b>Null Hypothesis</b>	The mean of the population that the sample comes from is equal to the mean of the known population (so, $\mu = \mu_0$ )	The mean of the population that the sample comes from is equal to the mean of the known population (so, $\mu = \mu_0$ )
<b>Equation</b>	$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$	$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$ 
<b>Descriptive information</b>	The z statistic tells us how much our sample mean differs from the population mean in terms of population SE	The <i>t</i> statistic tells us how much our sample mean differs from the population mean in terms of sample SE (as an estimate)
<b>Null distribution</b>	z-distribution 	 <i>t</i> -distribution 

Using  $s_{\bar{x}}$  to estimate the population  $\sigma_{\bar{x}}$ .

Bessel's correction for sample variance  
(and standard deviation)

# Estimating population parameters from sample statistics



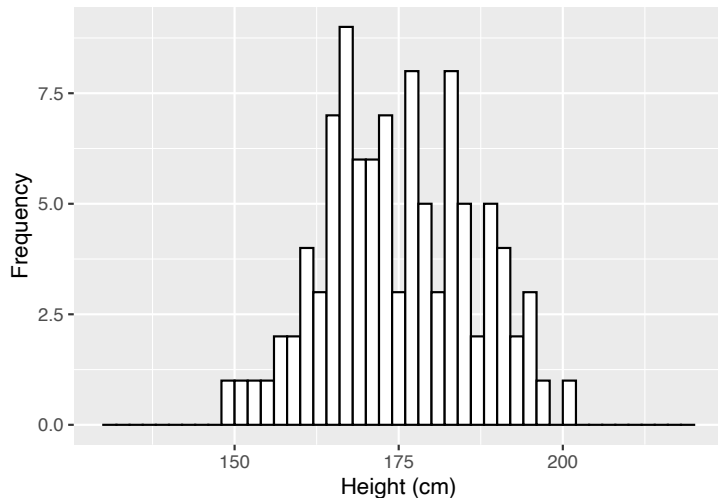
population

mean =  $\mu$  = ?

variance =  $\sigma^2$  = ?

standard deviation =  $\sigma$  = ?

When we have no way of measuring parameters directly, we can measure the sample statistics and use them as an estimate of the population parameters!



sample

mean =  $\bar{x}$  = 175

variance =  $s^2$  = 126

standard deviation =  $s$  = 11.2

# The mean is an **Unbiased Estimator**

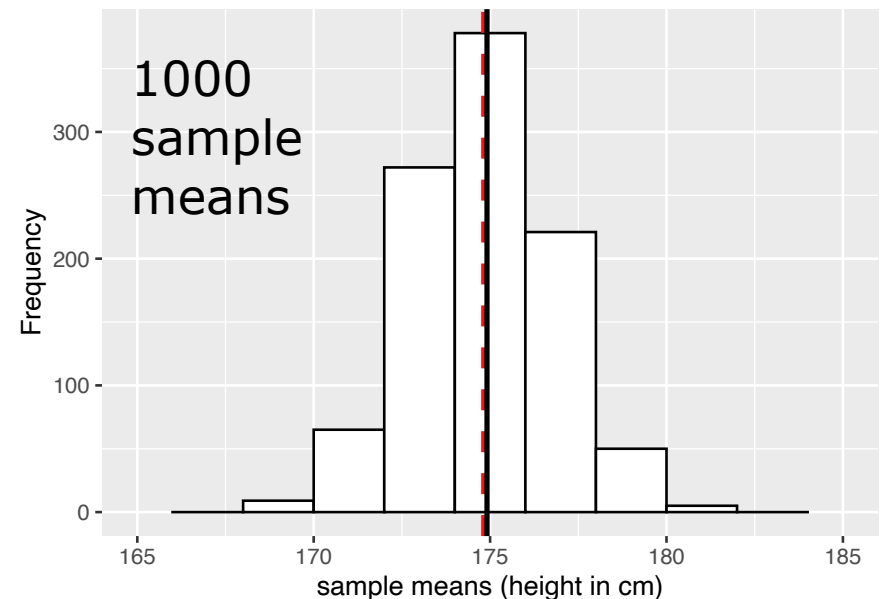
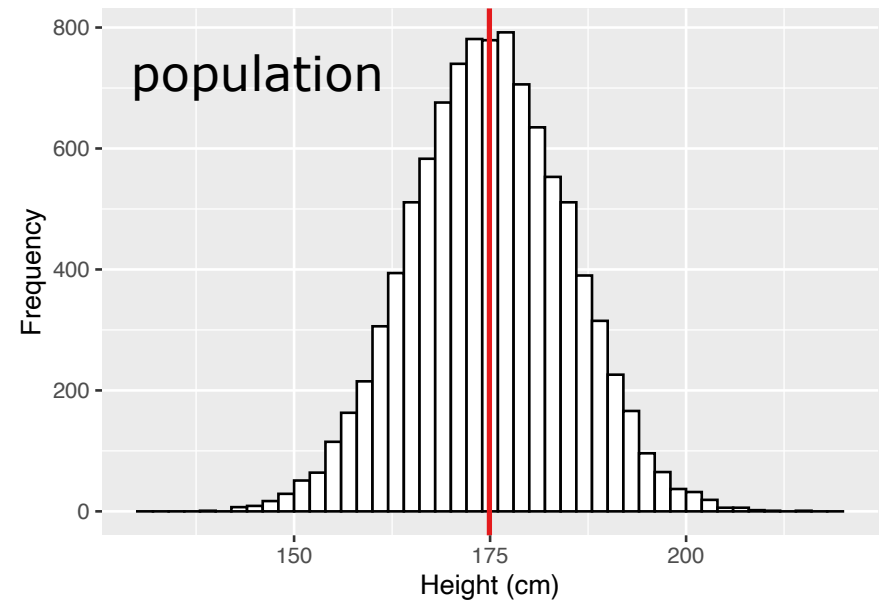
An **unbiased estimator** is a statistic that does not systematically underestimate or overestimate the population parameter.

This means the statistic for any given sample has an equal likelihood of being higher or lower than the parameter.

The **mean** is an **unbiased estimator**.

We can see this with a simulation. I selected 1000 samples of size 25 from the height population. I calculated the mean of each sample, and plotted those in this histogram.

I plot the mean of the **samples in dashed red** and the mean of the **population in black**. They are identical and the distribution is symmetric around it!



# The variance and standard deviation are **Biased Estimators**

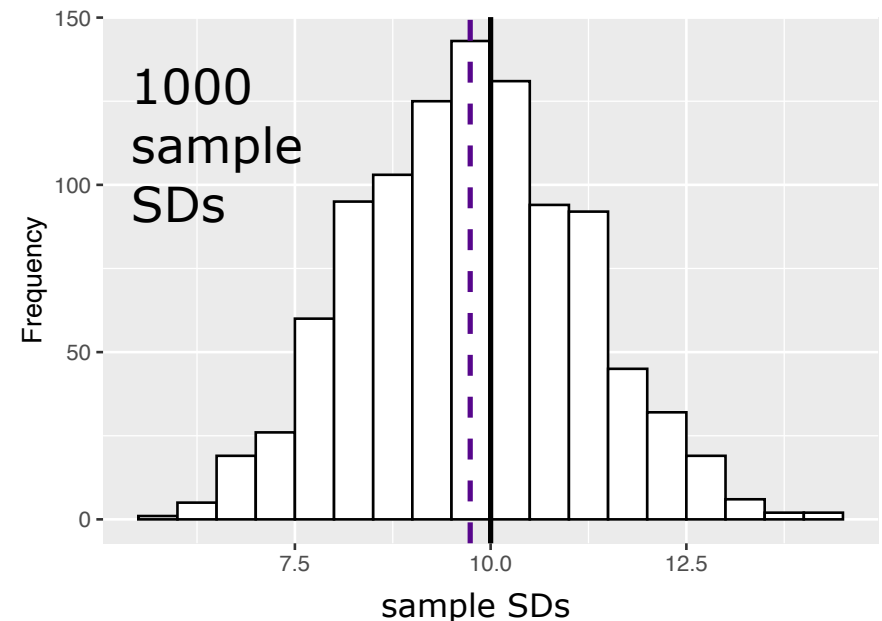
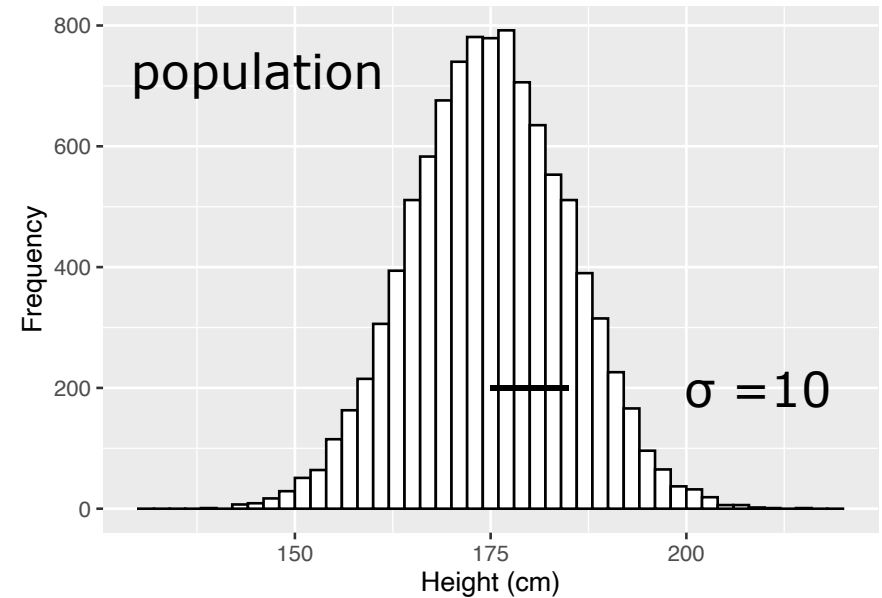
A **biased estimator** is a statistic that systematically underestimates or overestimates the population parameter.

This means the statistic cannot be used to estimate the parameter. Over the long run, it will give you a biased estimate.

The **variance** and the **standard deviation** are biased estimators.

We can see this with a simulation. I selected 1000 samples of size 25 from the height population. I calculated the standard deviation of each sample, and plotted those in this histogram.

I plot the standard deviation of the **samples in dashed purple** and the  $\sigma$  of the **population in black**. The sample s underestimates the population  $\sigma$ !



# Correcting the bias in variance (and nearly correcting it in the standard deviation)

To correct the bias in variances (and to mostly correct the bias in standard deviations), we apply something called **Bessel's correction** to the equations when we are calculating sample statistics (but not when we are calculating population parameters). We simply **divide by n-1** instead of n:

**variance**

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2}{n}$$

$\mu$  = population mean

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

$\bar{x}$  = sample mean

**standard deviation**

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2}{n}}$$

$\mu$  = population mean

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

$\bar{x}$  = sample mean

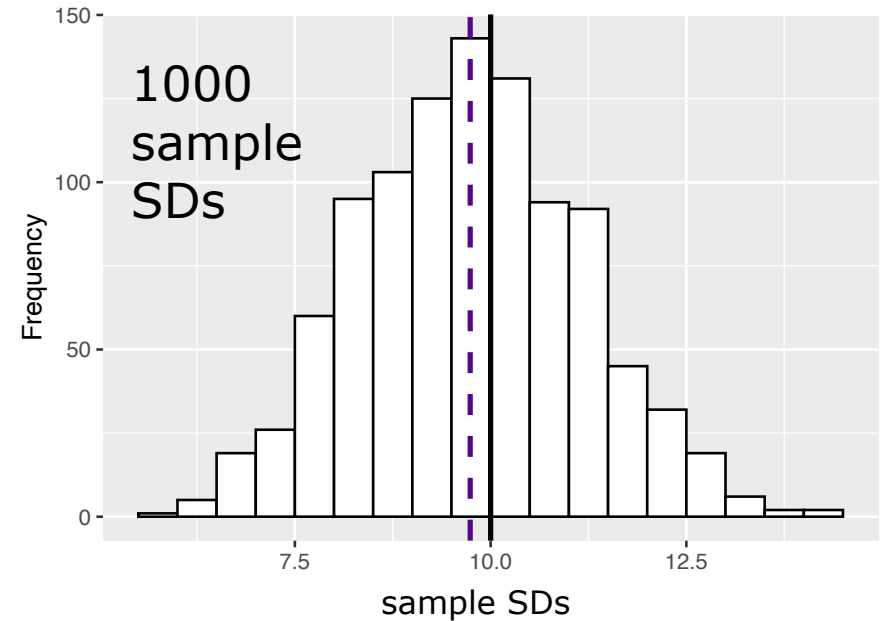


# Showing that Bessel's correction works

(perfectly for variance, close for SD)

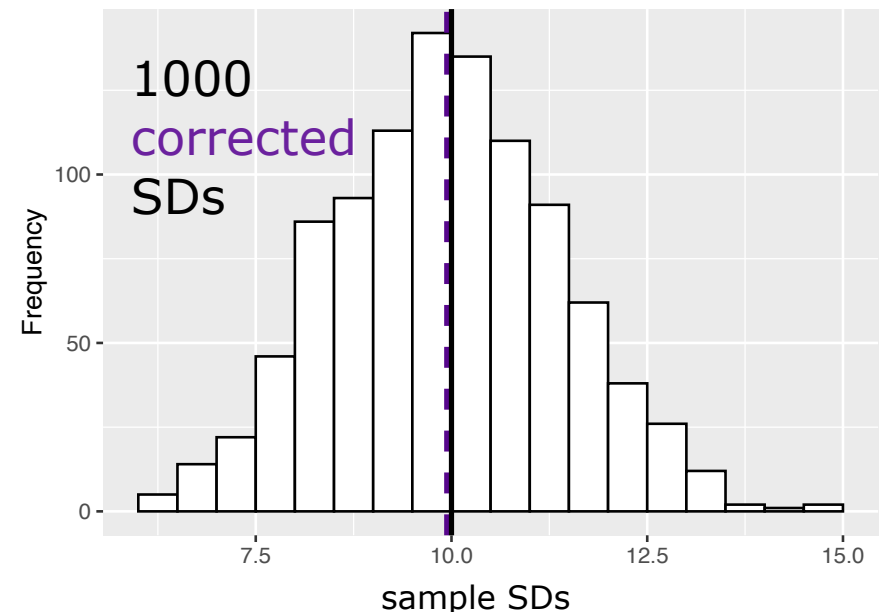
We can see that Bessel's correction works by applying it to the same simulation that we did before.

Here is the original simulation with the uncorrected equation for standard deviations.



Now we can rerun the simulation using the correction. The built-in function in R **sd()** applies the correction.

And what we see is that the **mean of the corrected SDs** is nearly identical to the **population SD**. This shows that the correction works - some sample SDs are above, some are below, but there is no bias.



So, whenever you are using **samples**, use **n-1**

$$\text{sum of squares} = (x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \dots + (x_n - \text{mean})^2$$

$$\text{variance} = \frac{(x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \dots + (x_n - \text{mean})^2}{n-1}$$

$$\text{standard deviation} = \sqrt{\frac{(x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \dots + (x_n - \text{mean})^2}{n-1}}$$

# Why does Bessel's correction work?

One basic answer is that by making the divisor smaller ( $n-1$ ), we are making the variance and standard deviation larger. Since the uncorrected versions are systematically underestimating the population parameters, we want it to be larger to overcome this, so this is a welcome result:

$$\text{standard deviation} = \sqrt{\frac{(x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \dots + (x_n - \text{mean})^2}{n-1}}$$

But this doesn't answer the question of why it is  $n-1$ . Any number that is smaller than  $n$  would increase the standard deviation:  $n-2$ ,  $n-3$ , etc.

To really answer this question we need the concept of **degrees of freedom**.

We also need degrees of freedom to understand the  $t$ -distribution, so now is a good time to look at that!

# Degrees of Freedom

# Degrees of Freedom

The idea of degrees of freedom was introduced way back in chapter 3 of our book, but we didn't need it back then. We need it now.

One definition of **degrees of freedom** is the **number of scores in a data set that can vary freely based on the amount of information you know about the data set.**

That is a strange way of talking. What do we mean by that? Let's use examples.

**Example 1:** If I tell you that we have a sample with 5 scores. But we know nothing else about the sample. How many degrees of freedom does the sample have?

— — — — —

**Answer:** It has **5 degrees of freedom** because each of the 5 scores could be anything. They are each free to vary. We have no information that would constrain them.

# Degrees of Freedom

**Example 2:** If I tell you that we have a sample with 5 scores and the mean is 3. How many degrees of freedom does the sample have?

                                              mean = 3

**Answer:** It has 4 degrees of freedom because only the first four scores can freely vary. After the first 4 vary, the 5th must be whatever number is necessary to create a mean of 3.

1 2 3 4 5                      mean = 3

2 2 2 2 7                      mean = 3

10 0 20 0 -15                      mean = 3

1 2 3 4 7                      mean = 3.4

Notice that if you try to make the 5th free to vary, you won't get the right mean!

# Degrees of Freedom

**Example 3:** If I tell you that we have a sample with 5 scores and the mean is 3 and the standard deviation is 1.41. How many degrees of freedom does the sample have?

                                              mean = 3, sd = 1.41

**Answer:** It has 3 degrees of freedom because only the first four scores can freely vary. After the first 3 vary, the 4th and 5th must be whatever numbers are necessary to create a mean of 3 and a standard deviation of 1.41.

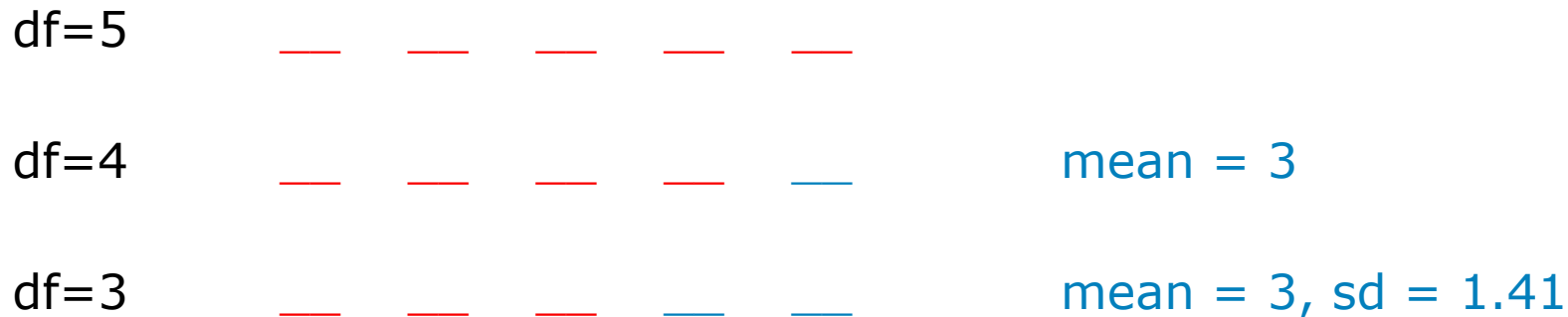
  1     2     3     4     5                        mean = 3, sd = 1.41

  2     2     2    10   -1                       mean = 3, sd = 3.69

Notice that if you let 4th vary, the 5th can be chosen to get the mean correct, but it can't be chosen to get the standard deviation correct. This means that once you know the standard deviation, you lose 2 degrees of freedom

# Degrees of Freedom

The general rule is that you lose 1 degree of freedom for every piece of information you know about a sample.



The **maximum number of degrees of freedom** is the sample size. So it is  $n$ .

When you lose 1 degree of freedom, say because you already know the mean, your degrees of freedom are  $n-1$ .

If you lose 2 degrees of freedom, say because you already know the mean and standard deviation, your degrees of freedom are  $n-2$ .

(Notice that you have to know the mean to calculate the standard deviation, so you will always know 2 pieces of information if you know the standard deviation.)



# df in Bessel's correction

Let's look again at Bessel's correction. We divide by  $n-1$ . You will now recognize that as the number of degrees of freedom when the mean is already known:


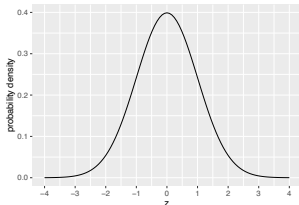


$$\text{standard deviation} = \sqrt{\frac{(x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \dots + (x_n - \text{mean})^2}{n-1}}$$

And this makes some sense, in that the mean is calculated from the sample first in order to then calculate the variance or standard deviation. So one degree of freedom has been lost when you start to calculate the standard deviation.

(But, this is not a complete answer. This just shows you why you need to make the variance larger, and why  $n-1$  is a logical choice for doing that. The mathematical proof that shows that Bessel's correction works with  $n-1$  is beyond my knowledge... but you can see it on the internet if you are curious!)

# The $t$ distribution

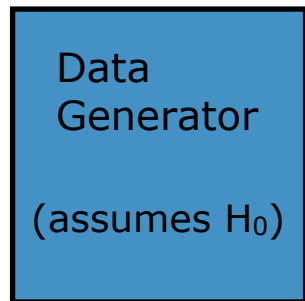
# From z-tests to *t*-tests

	(one sample) z-test	one sample <i>t</i> -test
<b>Scientific question</b>	Does our sample differ from a population with a known mean and standard deviation?	Does our sample differ from a population with a known mean (but <b>unknown SD</b> )?
<b>Null Hypothesis</b>	The mean of the population that the sample comes from is equal to the mean of the known population (so, $\mu = \mu_0$ )	The mean of the population that the sample comes from is equal to the mean of the known population (so, $\mu = \mu_0$ )
<b>Equation</b>	$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$	$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$ 
<b>Descriptive information</b>	The <i>z</i> statistic tells us how much our sample mean differs from the population mean in terms of population SE	The <i>t</i> statistic tells us how much our sample mean differs from the population mean in terms of sample SE (as an estimate)
<b>Null distribution</b>	z-distribution 	 <i>t</i> -distribution 

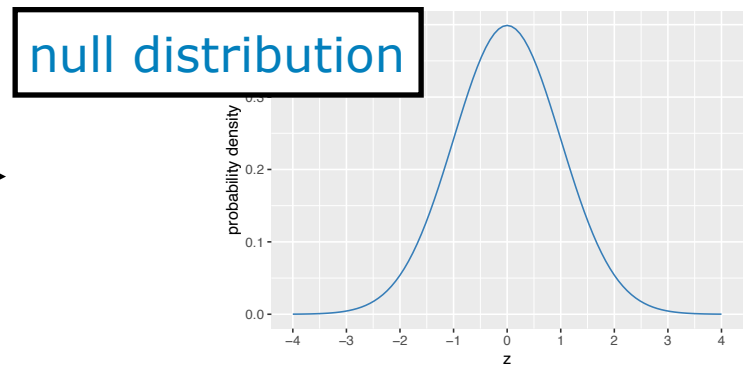
# Remember the mathematical steps of NHT

The mathematical part of NHT has three steps:

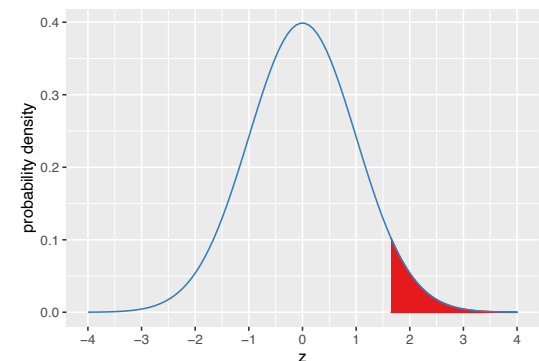
1. Run an experiment to collect the **observed data**. Calculate a statistic from it, like the mean or a z-score.
2. Assume that the null hypothesis is true, and generate **all possible data sets** that could arise (using the same sample size as your experiment). We summarize it as a distribution called the **null distribution**.



data1  
data2  
data3  
...



3. Look up the probability of the **observed data or data more extreme** in the **null distribution**. This is a conditional probability.

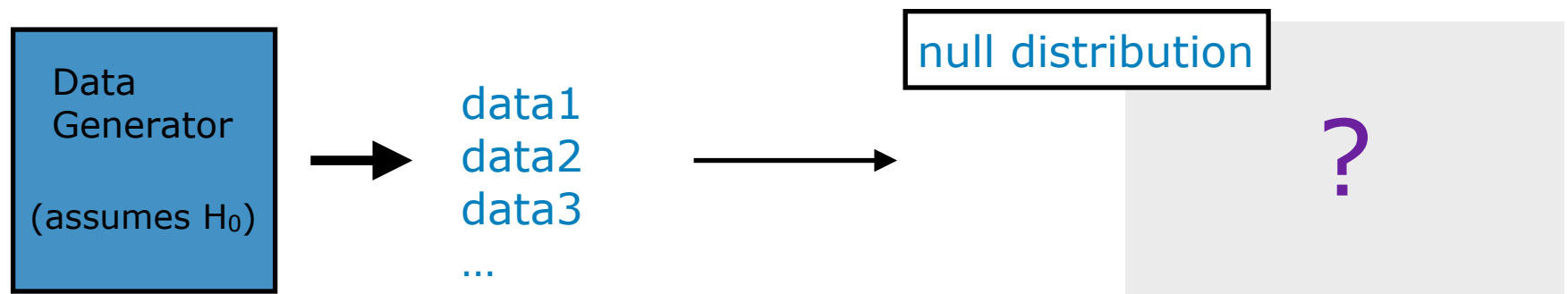


$$P(\text{data} \mid H_0) = \frac{\text{observed data}}{\text{generated data}}$$

# We need to figure this out for the $t$ -distribution

The mathematical part of NHT has three steps:

1. Run an experiment to collect the **observed data**. Calculate a statistic from it, in this case a  $t$ .
2. Assume that the null hypothesis is true, and generate **all possible data sets** that could arise (using the same sample size as your experiment). We summarize it as a distribution called the **null distribution**.



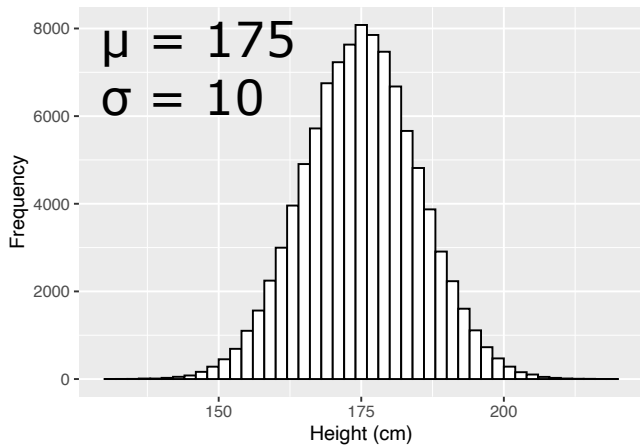
3. Look up the probability of the **observed data or data more extreme** in the **null distribution**. This is a conditional probability.

$$P(\text{data} \mid H_0) = \frac{\text{observed data}}{\text{generated data}}$$



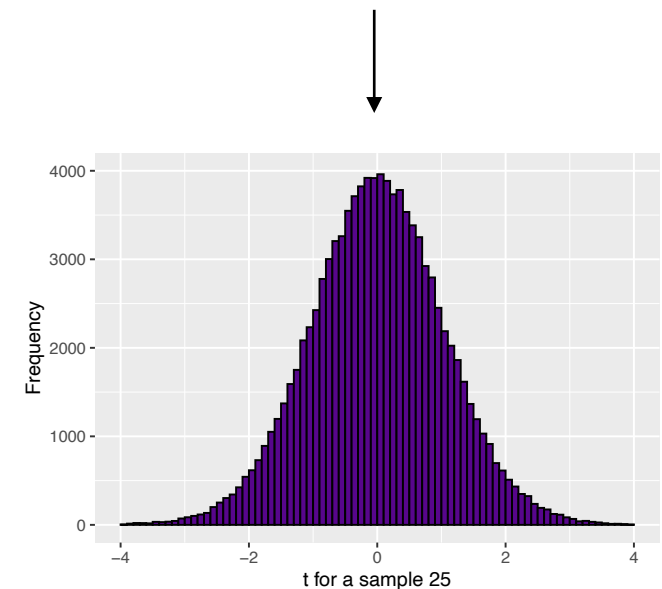
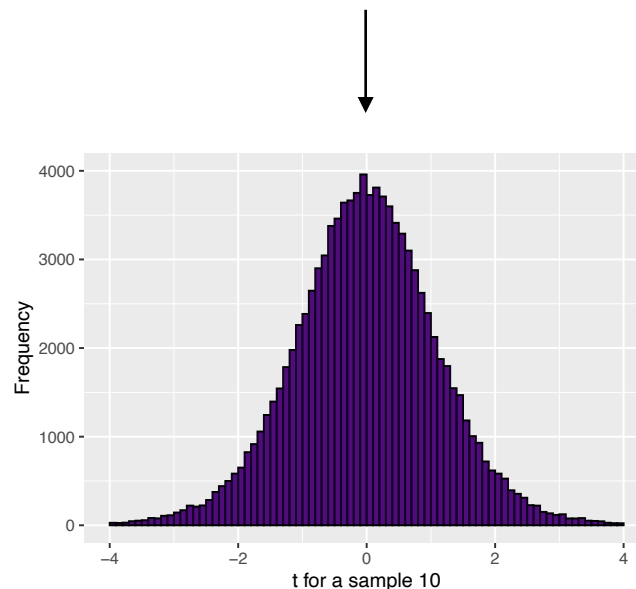
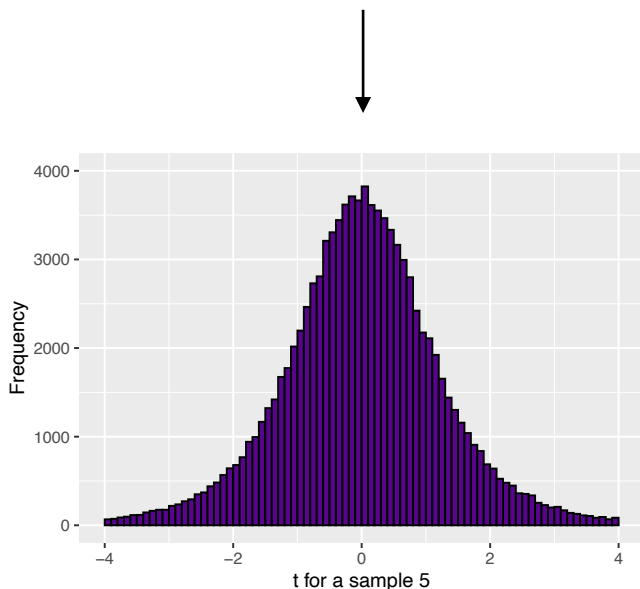
# Let's empirically simulate the $t$ distribution for experiments of different sample sizes

William Sealy Gossett developed the  $t$ -distribution by going out and running experiments of different sizes and empirically determining the distribution.



We can do it faster with a simulation. Here is a population of heights. We can repeatedly sample from it 100,000 times for three sample sizes: 5, 10, and 25.

We can calculate a  $t$  for each of the 100,000 samples, and plot the distribution of those  $t$  values.



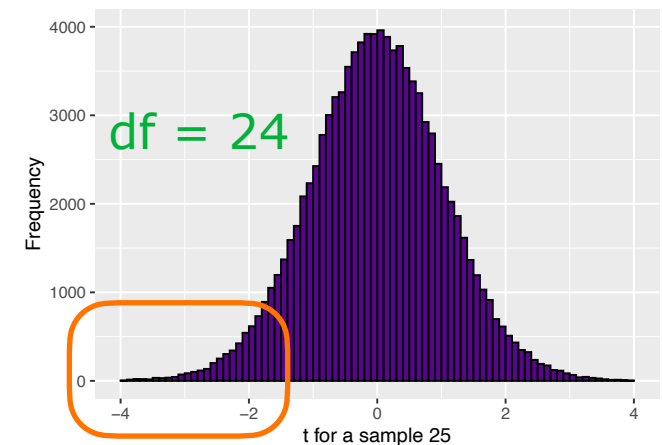
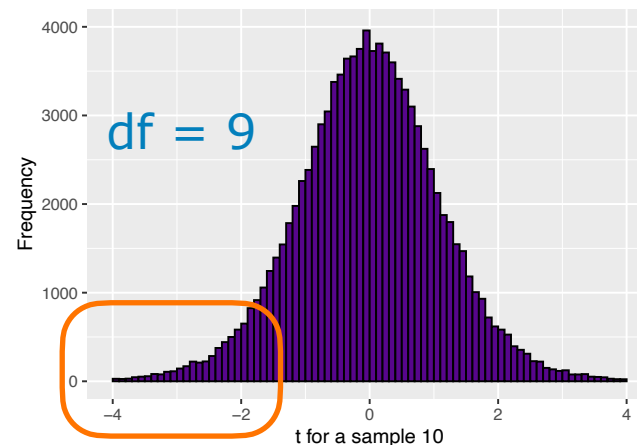
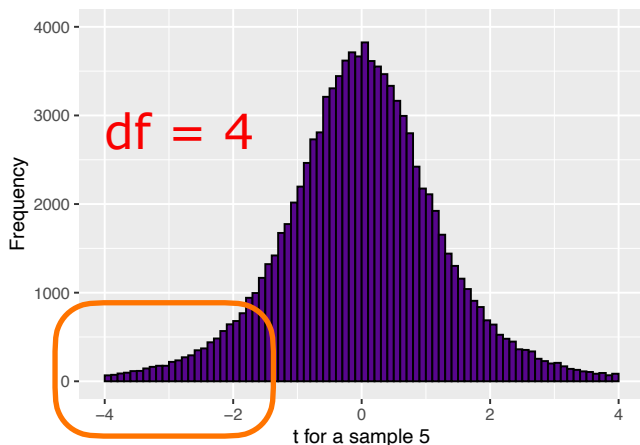
# We see small differences based on df

The  $t$  distribution is a family of distributions. The family is determined by the **degrees of freedom**, here  $n-1$  (because the mean was used in calculating  $t$ ).

There is a difference in the tails of the distribution: lower df leads to a fatter tail; higher df leads to a thinner tail.

There is also a difference in the peaks, but it is hard to see here: as df grows, the peak gets higher (the fat tails shift to a more peaky peak).

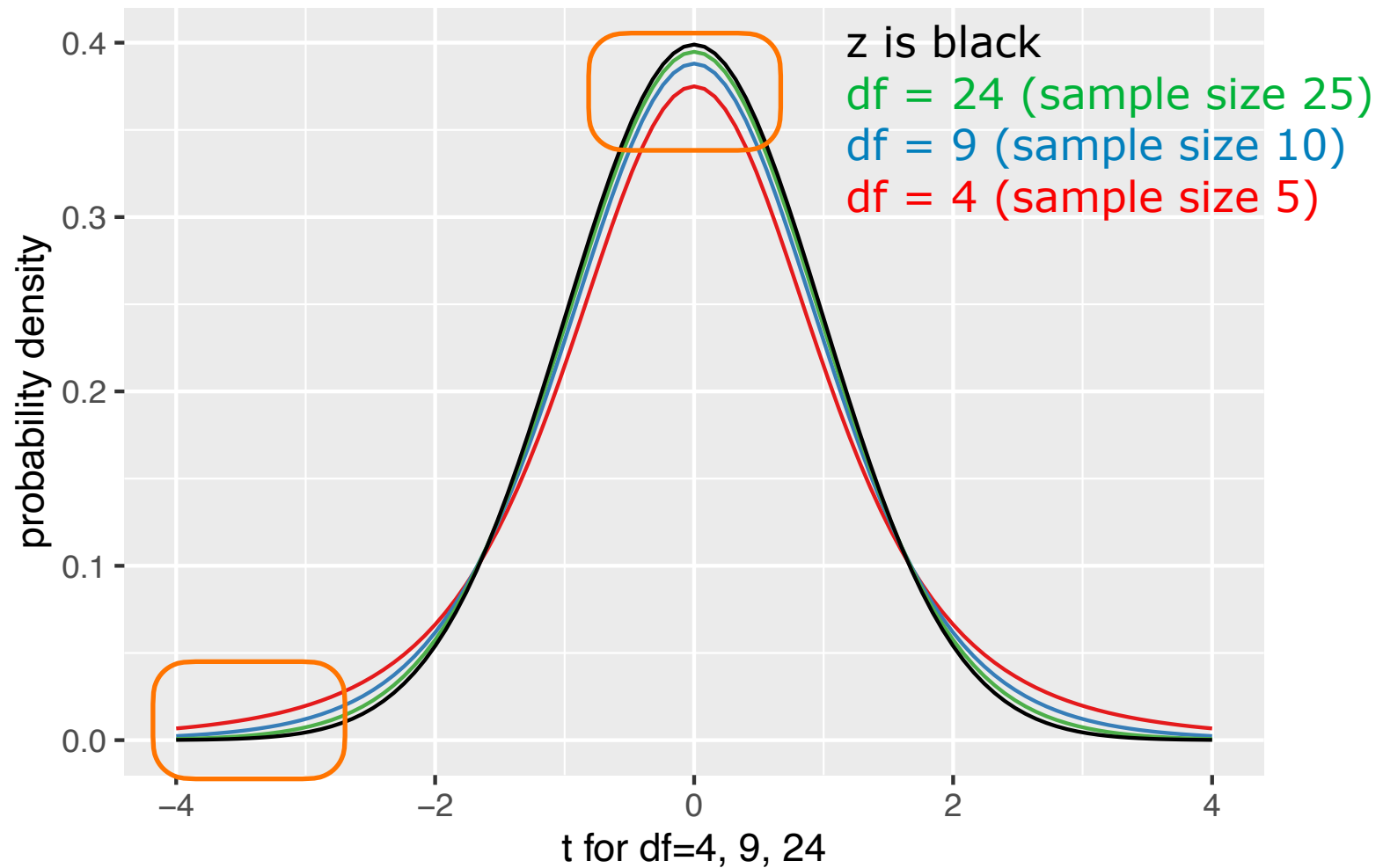
As df grows (so, as the sample size gets bigger) the  $t$  distribution approaches normal — it approaches the  $z$  distribution.



# The analytic $t$ distribution

Fisher helped Gossett calculate an analytic  $t$  distribution. Using the analytic form can help us see the differences in the tails a bit more clearly.

I'll also plot them on top of each other to highlight the differences: as  $df$  increases, the distribution approaches a  $z$  distribution.



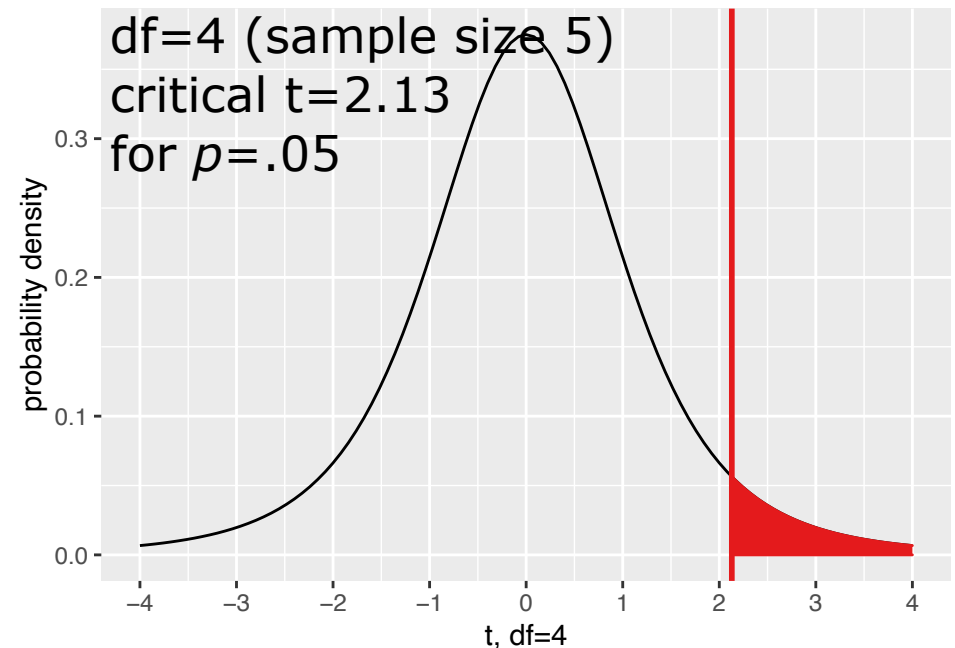
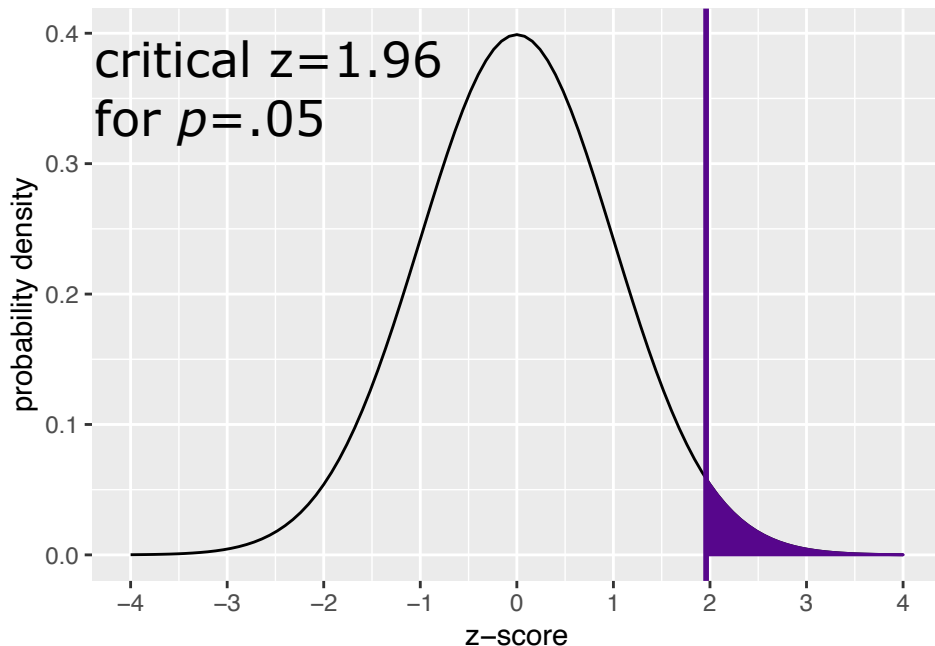


# What is the practical difference between a $t$ distribution and $z$ distribution?

Remember the difference between the two is in the tails. The  $t$  distribution has fatter tails.

The practical consequence of this is that **you will need a larger critical  $t$  value than a critical  $z$  value to reach a  $p$ -value of .05.**

This is because that fatter tail means more extreme observations happen with a  $t$  distribution. This makes logical sense - these are small sample experiments, so more extreme things can happen by chance.



# Writing up $t$ -test results

When you write up the results, you need to tell readers the mean, standard deviation,  $t$ -statistic, the degrees of freedom, and of course the  $p$ -value:

“Although the mean hourly fee for our sample of current psychotherapists was considerably higher ( $M = \$72$ ,  $SD = 22.5$ ) than the 1960 population mean ( $\mu = \$63$ , in current dollars), this difference only approached statistical significance,  $t(24) = 2.00$ ,  $p = .06$ .”

# If we saw a one-sample $t$ -test on homework or an exam

Imagine that we have a new intelligence test. We don't know the standard deviation of the test, but we want to know if our sample scored higher than the expected mean of 100.

We'd first remember our formula:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \dots \text{ and remember: } s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Then plug in our numbers:

$$t = \frac{106 - 100}{2.5} \quad \dots \text{ and remember: } s_{\bar{x}} = \frac{12.5}{\sqrt{25}}$$

$$t = 2.4$$

## **scores:**

72 93 93 96 98  
99 100 101 101 102  
103 103 104 105  
107 109 110 113  
115 118 119 122  
125 126 127

$$\bar{x} = 106$$

$$s = 12.5$$

# If we saw this on homework or an exam

Imagine that we have a new intelligence test. We don't know the standard deviation of the test, but we want to know if our sample scored higher than the expected mean of 100.

Then we'd look up our  $t$  in Table A2 in the book, or use `pt()` in R:

$$t = 2.4$$

$$p = .012$$

## scores:

72 93 93 96 98  
99 100 101 101 102  
103 103 104 105  
107 109 110 113  
115 118 119 122  
125 126 127

$$\bar{x} = 106$$

$$s = 12.5$$

Table A2 **only tells you the critical t** for the df:

## df            t for $p < .05$

20	1.325	1.725
21	1.323	1.721
22	1.321	1.717
23	1.319	1.714
24	1.318	1.711
25	1.316	1.708
26	1.315	1.706
27	1.314	1.703

For the `pt()` function, you need to specify the **degrees of freedom**:

```
> pt(2.4, df=24)
[1] 0.9877451
> 1-pt(2.4, df=24)
[1] 0.01225495
>
```