

Stat 411/511

# STATISTICAL TESTING

Jan 20 2012

Charlotte Wickham

[stat511.cwick.co.nz](http://stat511.cwick.co.nz)

# Announcements

**HW #1:** Submitted online, as long as it is there by 5pm today I won't consider it late.

**Lab #2:** Do on your own

**Wednesday help session** (after lecture CORD 2113)

**Extra help** (this week only) **Thursday 2-3.30pm**

Valley Library 6420

# Last time

How chance enters the study design  
determines the scope of inference

# This time

The Statistical justice system.

How chance allows us to quantify  
evidence for a hypothesis.

# Terminology

Name	Definition	Examples
<b>Parameter</b> (of a population)	An <b>unknown</b> value in a probability model	Population mean, $\mu$ (mu) Population standard deviation, $\sigma$ (sigma)
<b>Statistic</b> (of a dataset)	Something you can calculate from data	Sample average, $\bar{Y}$ (y bar) Sample standard deviation, SD or s
<b>Estimate</b> (of a parameter)	A statistic used a guess for a parameter	The sample average is an estimate of the population mean.

# Comparing two groups

**Treatment effect:** The effect of being in Group A is to change the outcome by  $\delta$  (delta), compared to Group B.

What is  $\delta$ ?

**Difference in means:** Population A has a mean of  $\mu_1$  (mu one), population B has a mean of  $\mu_2$  (mu two).

What is  $\mu_1 - \mu_2$ ?

# The statistical justice system



A **suspect** is accused of a **crime**. The suspect is declared guilty or not guilty based on a **trial**. Each trial has a **defence** and a **prosecution**. On the basis of how **evidence** compares to a **standard**, the judge makes a decision to **convict** or **acquit**.

A **dataset** is accused of coming from a probability model with a **particular parameter**. The data is declared guilty or not guilty based on the results of a **statistical test**. Each test has a **null hypothesis** and an **alternative hypothesis**. On the basis of how a **test statistic** compares to a standard **distribution**, we make the decision to **reject the null** or **fail to reject the null hypothesis**.

<p><b>Suspect</b> (the data)</p>	<p>Extrinsic Scores, <math>Y_1</math>: 5.0 5.4 6.1 10.9 11.8 12.0 12.3 14.8 15.0 16.8 17.2 17.2 17.4 17.5 18.5 18.7 18.7 19.2 19.5 20.7 21.2 22.1 24.0</p> <p>Intrinsic Scores, <math>Y_2</math>: 12.0 12.0 12.9 13.6 16.6 17.2 17.5 18.2 19.1 19.3 19.8 20.3 20.5 20.6 21.3 21.6 22.1 22.2 22.6 23.1 24.0 24.3 26.7 29.7</p>
<p><b>Crime</b> (parameter of the model)</p>	<p>The effect of getting the intrinsic questionnaire is to increase creativity compared to the extrinsic questionnaire.</p> <p><math>\delta &gt; 0</math></p>
<p><b>Evidence</b> (the test statistic)</p>	<p>The intrinsic group had an average score 4.14 higher than the extrinsic group.</p> <p><math>\bar{Y}_2 - \bar{Y}_1 = 4.14</math></p>

Compare the evidence to innocents to determine guilt

# Hypotheses

Every case has two sides:

The **null hypothesis** (the defence). This is the default, or the status quo, what you need to argue against.

The **alternative hypothesis** (the prosecution). This is the interesting case, but you need to prove it's true.

# In the creativity study

**Null hypothesis** (the defense):

The treatment effect is zero,  $\delta = 0$ . The difference in averages we saw arose purely by chance.

**Alternative hypothesis** (the prosecution):

The treatment effect is greater than zero,  $\delta > 0$ . The difference in averages we saw cannot be explained by chance.

In the statistical justice system evidence is weighed by comparing the accused to known innocents.

The population of innocents, called the **null distribution**, is generated by the combination of null hypothesis and test statistic.

This evidence is summarised with the **p-value**, the probability that a true innocent would look as (or more) guilty than the accused.

# What do innocents look like?

For **controlled experiments**, one way is to use the **randomization** distribution.

If the data are innocent, the group assignment doesn't change a subject's creativity score.

Innocents look like the **differences in averages** we could have seen with different ways of assigning the groups.

# Your turn

Under our assumption that the questionnaire has no effect, how could you choose the groups to give the **largest difference in averages** between the two groups?

<b><u>Motivation Group</u></b>			
<b><u>Intrinsic</u></b>		<b><u>Extrinsic</u></b>	
12.0	20.5	5.0	17.4
12.0	20.6	5.4	17.5
12.9	21.3	6.1	18.5
13.6	21.6	10.9	18.7
16.6	22.1	11.8	18.7
17.2	22.2	12.0	19.2
17.5	22.6	12.3	19.5
18.2	23.1	14.8	20.7
19.1	24.0	15.0	21.2
19.3	24.3	16.8	22.1
19.8	26.7	17.2	24.0
20.3	29.7	17.2	

What's the probability of that grouping occurring by chance?

# Motivation Group

<u><i>Intrinsic</i></u>		<u><i>Extrinsic</i></u>	
12.0	20.5	5.0	17.4
12.0	20.6	5.4	17.5
12.9	21.3	6.1	18.5
13.6	21.6	10.9	18.7
16.6	22.1	11.8	18.7
17.2	22.2	12.0	19.2
17.5	22.6	12.3	19.5
18.2	23.1	14.8	20.7
19.1	24.0	15.0	21.2
19.3	24.3	16.8	22.1
19.8	26.7	17.2	24.0
20.3	29.7	17.2	

# Motivation Group

## Intrinsic

## Extrinsic

12.0	20.5	5.0	17.4
12.0	20.6	5.4	17.5
12.9	21.3	6.1	18.5
13.6	21.6	10.9	18.7
16.6	22.1	11.8	18.7
17.2	22.2	12.0	19.2
17.5	22.6	12.3	19.5
18.2	23.1	14.8	20.7
19.1	24.0	15.0	21.2
19.3	24.3	16.8	22.1
19.8	26.7	17.2	24.0
20.3	29.7	17.2	

# Motivation Group

## Intrinsic

## Extrinsic

12.0	20.5	5.0	17.4
12.0	20.6	5.4	17.5
12.9	21.3	6.1	18.5
13.6	21.6	10.9	18.7
16.6	22.1	11.8	18.7
17.2	22.2	12.0	19.2
17.5	22.6	12.3	19.5
18.2	23.1	14.8	20.7
19.1	24.0	15.0	21.2
19.3	24.3	16.8	22.1
19.8	26.7	17.2	24.0
20.3	29.7	17.2	

Intrinsic  
Mean = 21.72

# Motivation Group

<u><i>Intrinsic</i></u>	<u><i>Extrinsic</i></u>	
12.0	20.5	5.0
12.0	20.6	5.4
12.9	21.3	6.1
13.6	21.6	10.9
16.6	22.1	11.8
17.2	22.2	12.0
17.5	22.6	12.3
18.2	23.1	14.8
19.1	24.0	15.0
19.3	24.3	16.8
19.8	26.7	17.2
20.3	29.7	17.2

Extrinsic  
Mean = 13.82

Intrinsic  
Mean = 21.72

# Motivation Group

<u><i>Intrinsic</i></u>	<u><i>Extrinsic</i></u>	
12.0	20.5	5.0
12.0	20.6	5.4
12.9	21.3	6.1
13.6	21.6	10.9
16.6	22.1	11.8
17.2	22.2	12.0
17.5	22.6	12.3
18.2	23.1	14.8
19.1	24.0	15.0
19.3	24.3	16.8
19.8	26.7	17.2
20.3	29.7	17.2
		17.4
		17.5
		18.5
		18.7
		18.7
		19.2
		19.5
		20.7
		21.2
		22.1
		24.0

Extrinsic  
Mean = 13.82

Intrinsic  
Mean = 21.72

Diff = 7.90

# Motivation Group

<u><i>Intrinsic</i></u>	<u><i>Extrinsic</i></u>	Extrinsic
12.0	20.5	5.0
12.0	20.6	5.4
12.9	21.3	6.1
13.6	21.6	10.9
16.6	22.1	11.8
17.2	22.2	12.0
17.5	22.6	12.3
18.2	23.1	14.8
19.1	24.0	15.0
19.3	24.3	16.8
19.8	26.7	17.2
20.3	29.7	17.2
		17.4
		17.5
		18.5
		18.7
		18.7
		19.2
		19.5
		20.7
		21.2
		22.1
		24.0

Mean = 13.82

Intrinsic  
Mean = 21.72

Diff = 7.90

This is one way to choose 24 numbers out of 47.

# Motivation Group

<u><i>Intrinsic</i></u>	<u><i>Extrinsic</i></u>	Extrinsic
12.0	20.5	5.0
12.0	20.6	5.4
12.9	21.3	6.1
13.6	21.6	10.9
16.6	22.1	11.8
17.2	22.2	12.0
17.5	22.6	12.3
18.2	23.1	14.8
19.1	24.0	15.0
19.3	24.3	16.8
19.8	26.7	17.2
20.3	29.7	17.2
		17.4
		17.5
		18.5
		18.7
		18.7
		19.2
		19.5
		20.7
		21.2
		22.1
		24.0

Mean = 13.82

Intrinsic  
Mean = 21.72

Diff = 7.90

This is one way to choose 24 numbers out of 47.  
There are more than 16 trillion ways to choose a set of 24 numbers out of a list of 47 numbers.

# Motivation Group

<u><i>Intrinsic</i></u>	<u><i>Extrinsic</i></u>	Extrinsic
12.0	20.5	5.0    17.4
12.0	20.6	5.4    17.5
12.9	21.3	6.1    18.5
13.6	21.6	10.9   18.7
16.6	22.1	11.8   18.7
17.2	22.2	12.0   19.2
17.5	22.6	12.3   19.5
18.2	23.1	14.8   20.7
19.1	24.0	15.0   21.2
19.3	24.3	16.8   22.1
19.8	26.7	17.2   24.0
20.3	29.7	17.2
		<p>Mean = 13.82</p> <p><b>Intrinsic</b></p> <p><b>Mean = 21.72</b></p> <p>Diff = 7.90</p>

This is one way to choose 24 numbers out of 47.

There are more than 16 trillion ways to choose a set of 24 numbers out of a list of 47 numbers.

Probability of this way occurring is less than one in 16 trillion

# 1000 Innocents

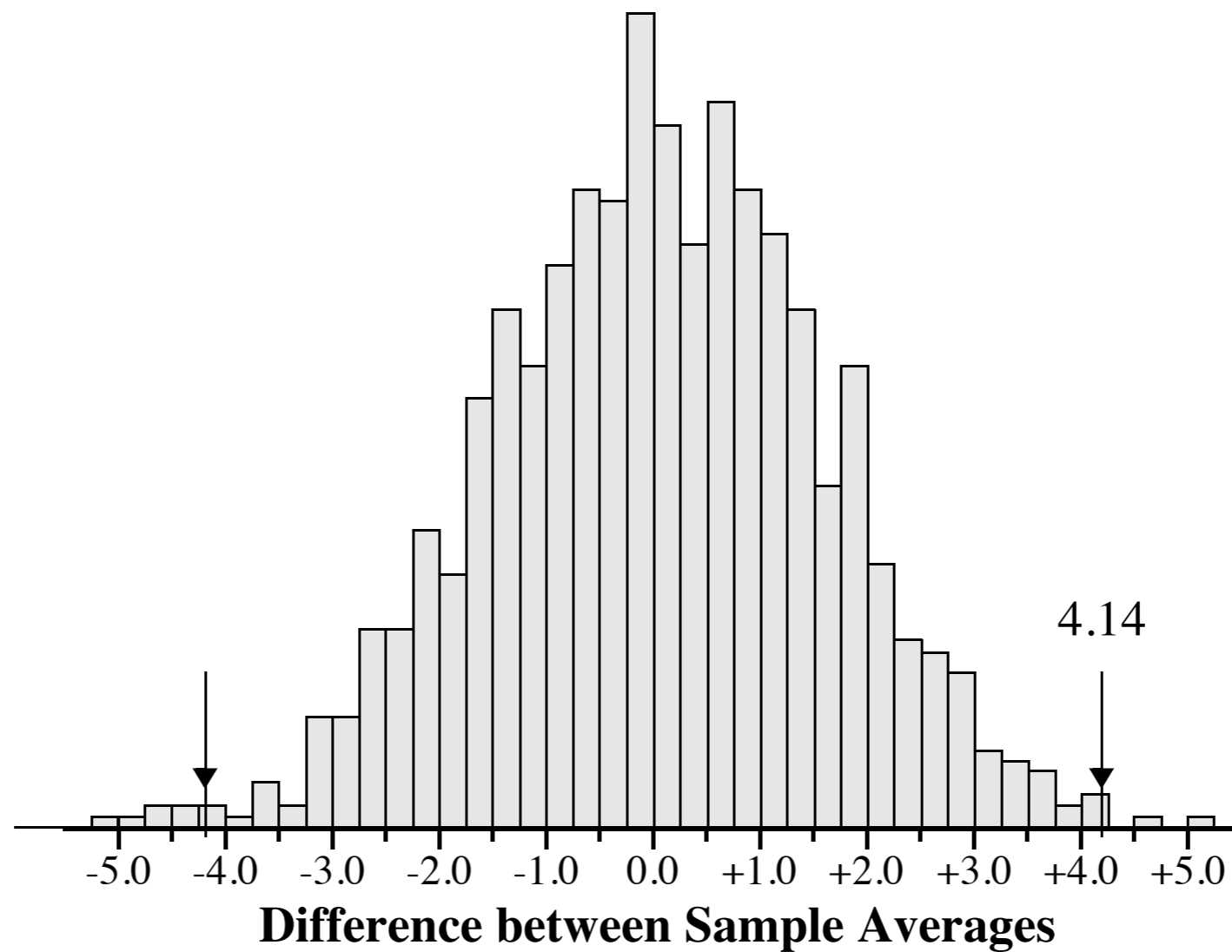
Display 1.8

p. 13

---

**A histogram of differences between group averages, from 1,000 randomizations of the creativity study data**

---



# 1000 Innocents

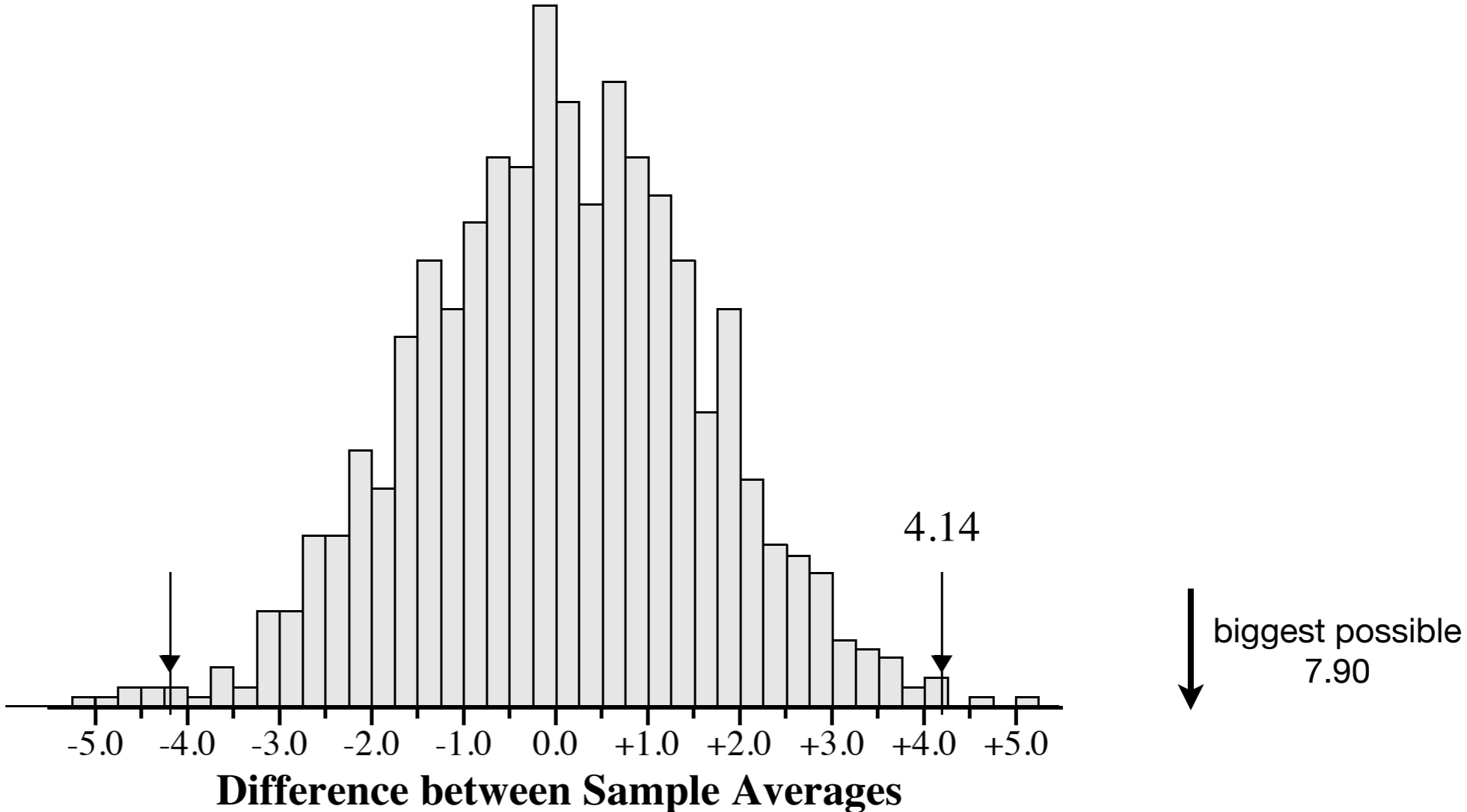
Display 1.8

p. 13

---

**A histogram of differences between group averages, from 1,000 randomizations of the creativity study data**

---



# 1000 Innocents

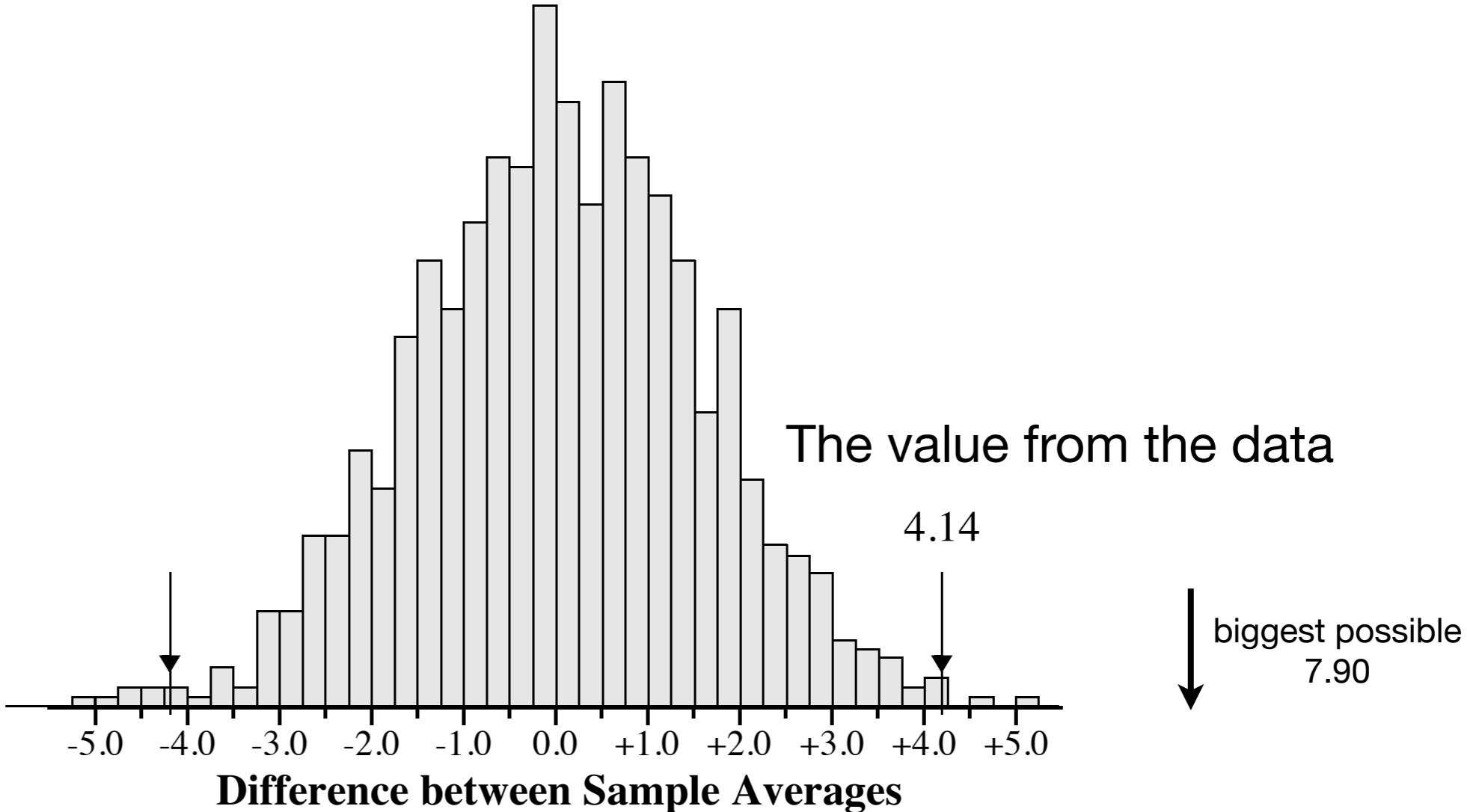
Display 1.8

p. 13

---

**A histogram of differences between group averages, from 1,000 randomizations of the creativity study data**

---



# 1000 Innocents

Display 1.8

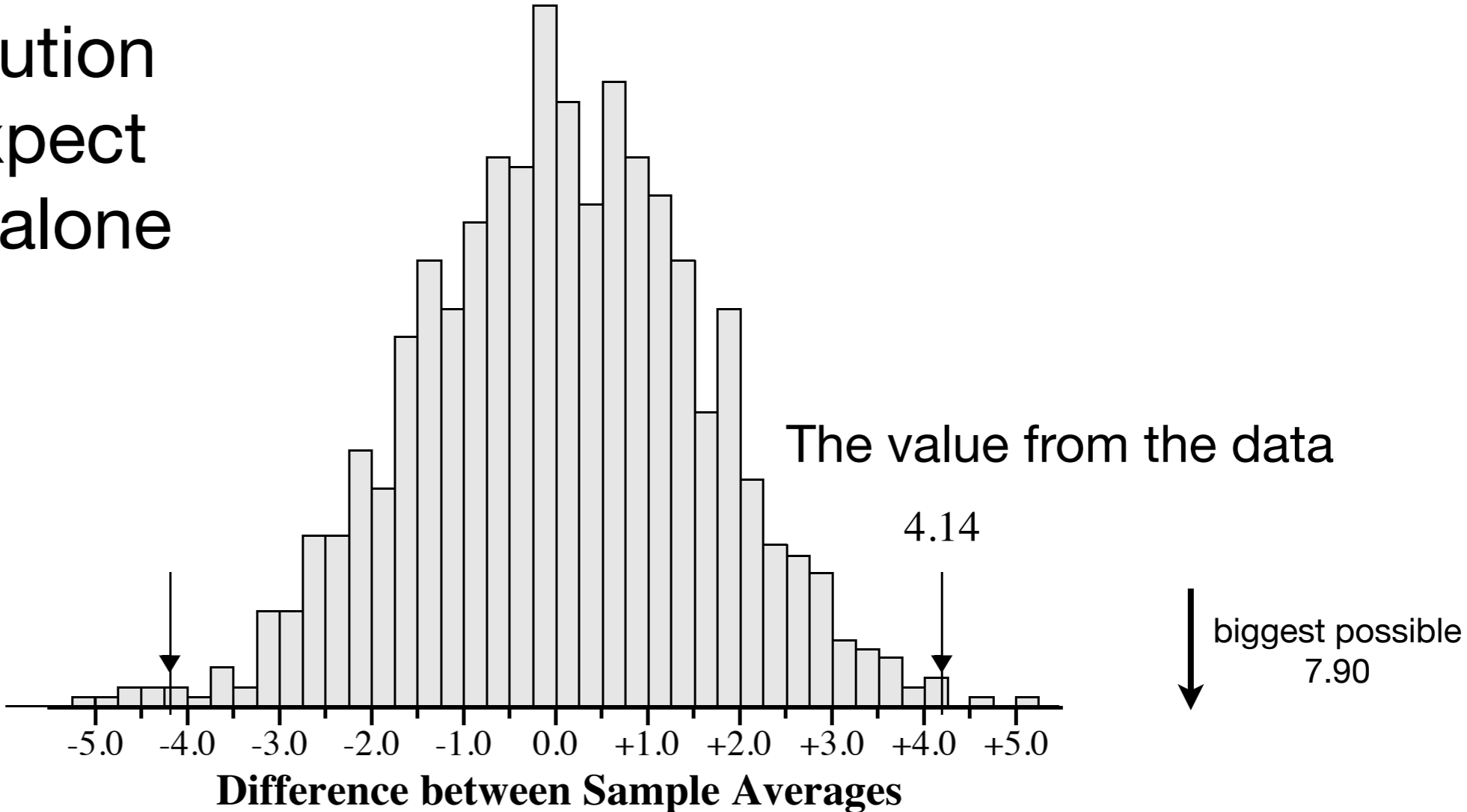
p. 13

---

A histogram of differences between group averages, from 1,000 randomizations of the creativity study data

---

An approximation to the distribution we would expect from chance alone



# 1000 Innocents

Display 1.8

p. 13

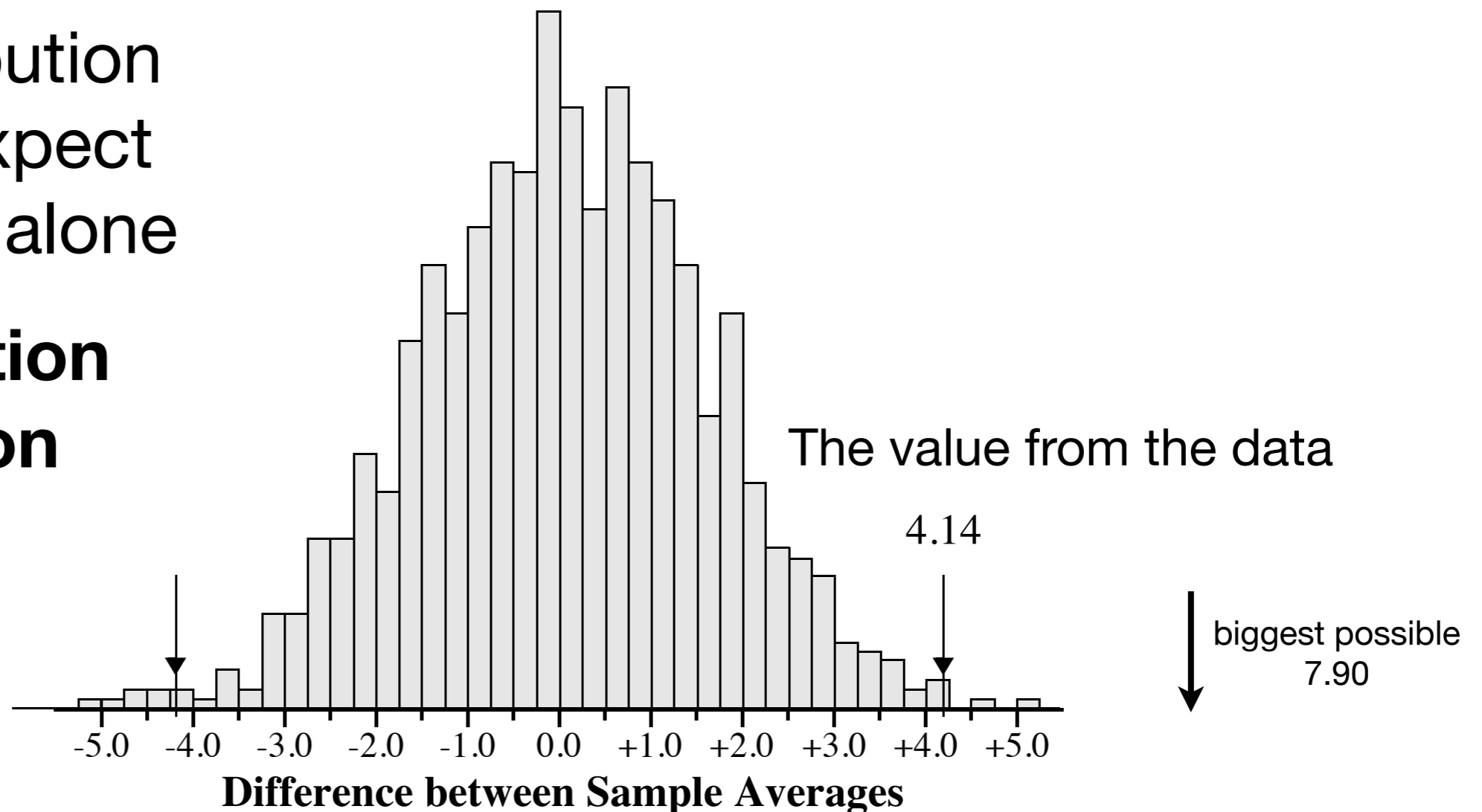
---

A histogram of differences between group averages, from 1,000 randomizations of the creativity study data

---

An approximation to the distribution we would expect from chance alone

**randomization distribution**



# 1000 Innocents

Display 1.8

p. 13

---

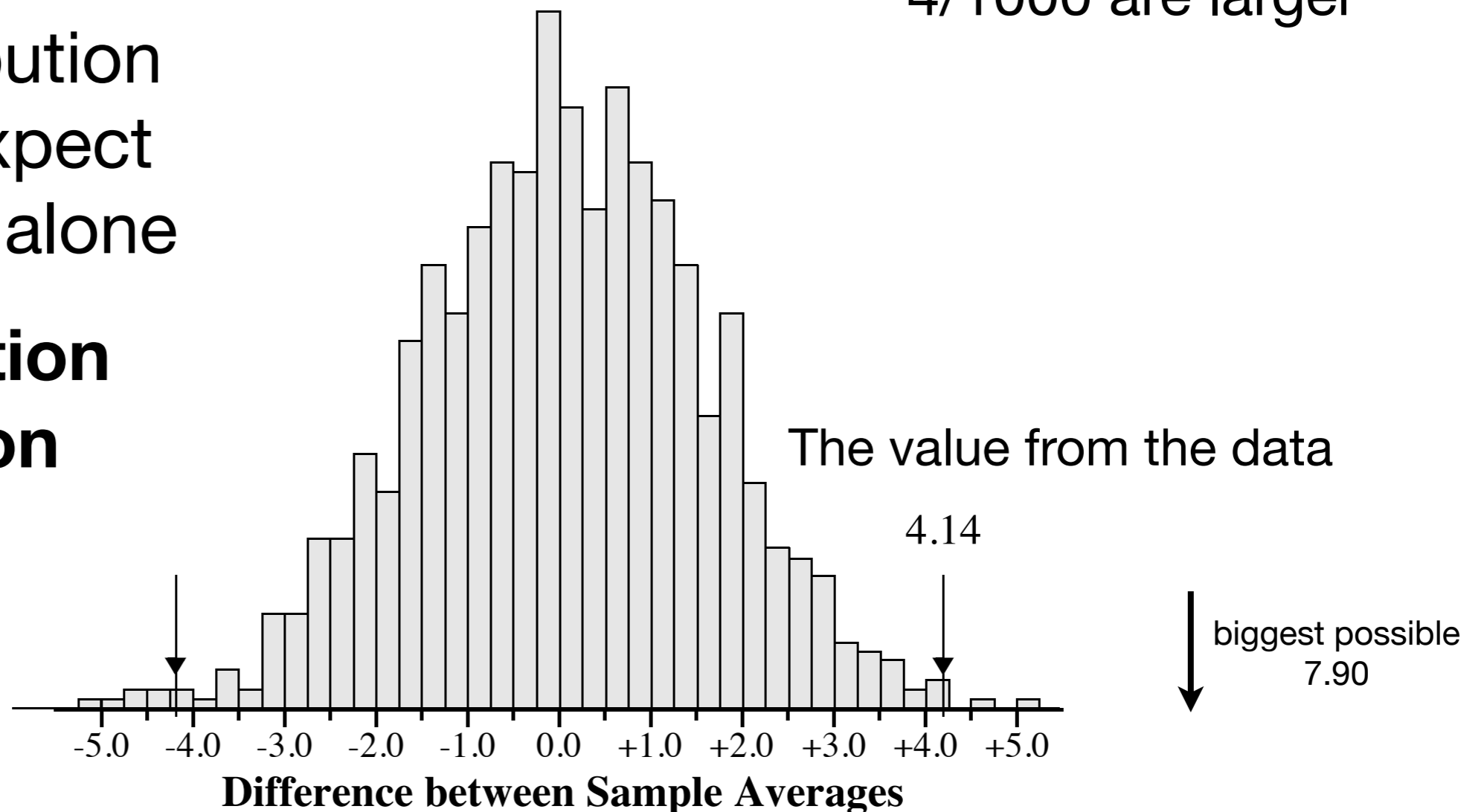
A histogram of differences between group averages, from 1,000 randomizations of the creativity study data

---

An approximation to the distribution we would expect from chance alone

**randomization distribution**

4/1000 are larger



# 1000 Innocents

Display 1.8

p. 13

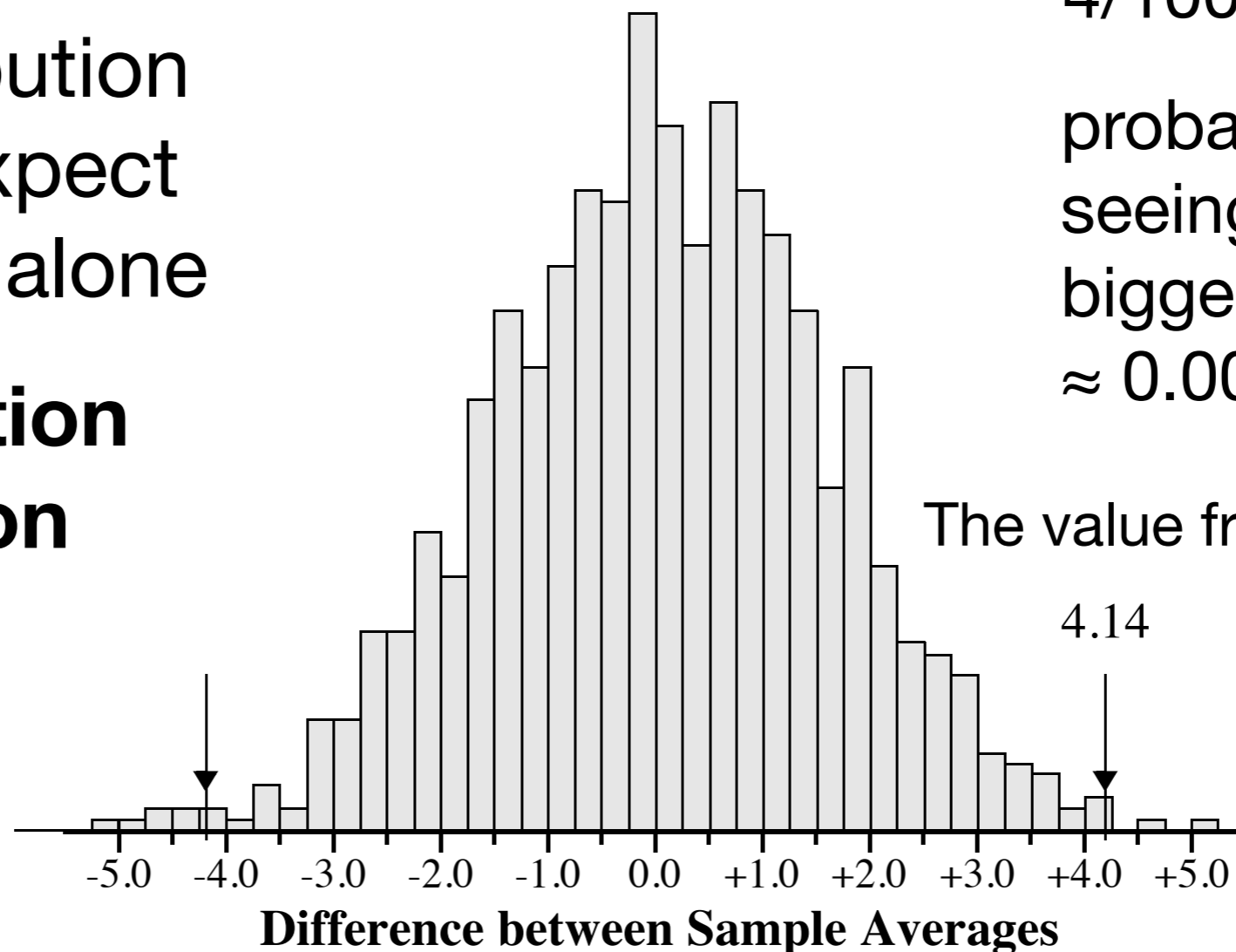
---

A histogram of differences between group averages, from 1,000 randomizations of the creativity study data

---

An approximation to the distribution we would expect from chance alone

**randomization distribution**



4/1000 are larger probability of seeing as big or bigger difference  $\approx 0.004$

The value from the data

4.14

biggest possible  
7.90

# 1000 Innocents

Display 1.8

p. 13

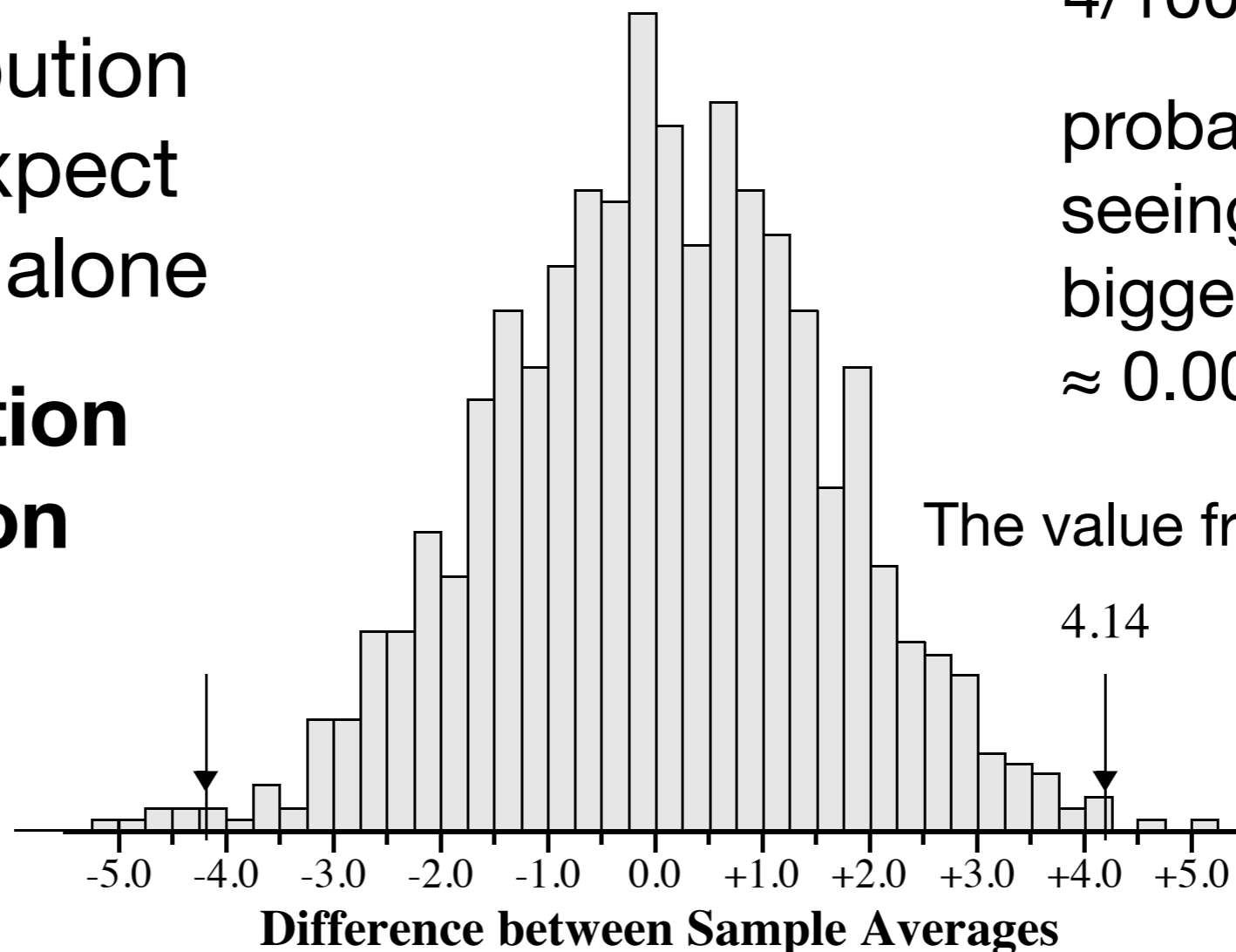
---

A histogram of differences between group averages, from 1,000 randomizations of the creativity study data

---

An approximation to the distribution we would expect from chance alone

**randomization distribution**



4/1000 are larger  
probability of  
seeing as big or  
bigger difference  
 $\approx 0.004$  **p-value**

The value from the data

4.14

biggest possible  
7.90

# Definitions

**test statistic:** numerical summary used to collapse sample into a single number

**null distribution:** distribution of test-statistic of innocents

**p-value:** probability that a true innocent would look as guilty as the suspect

# Rejecting the null

- To make a concrete decision, we would decide on a threshold of the p-value: the **significance level**
- Never declare the suspect innocent: only not guilty
- We never **accept the null hypothesis**, we **fail to reject the null hypothesis**.
- Best practice: **report the p-value** (read 2.5.1 & 2.5.3)

# Back to the creativity example

There is strong evidence (randomization test,  $p$ -value = 0.004) that the intrinsic questionnaire increased creativity compared to the extrinsic questionnaire in this set of subjects. The intrinsic questionnaire is estimated to increase creativity by 4.14 points more than the extrinsic questionnaire.