Introduction to Statistics

Statistics

Statistics is the science of drawing conclusions from data.

Statistics involves...

- understanding the question of interest
- designing the data collection process
- developing and applying methods for analyzing data
- assessing and reporting uncertainty associated with results

Why Study the Discipline of Statistics?

Those who ignore statistics are condemned to reinvent it.

–Brad Efron

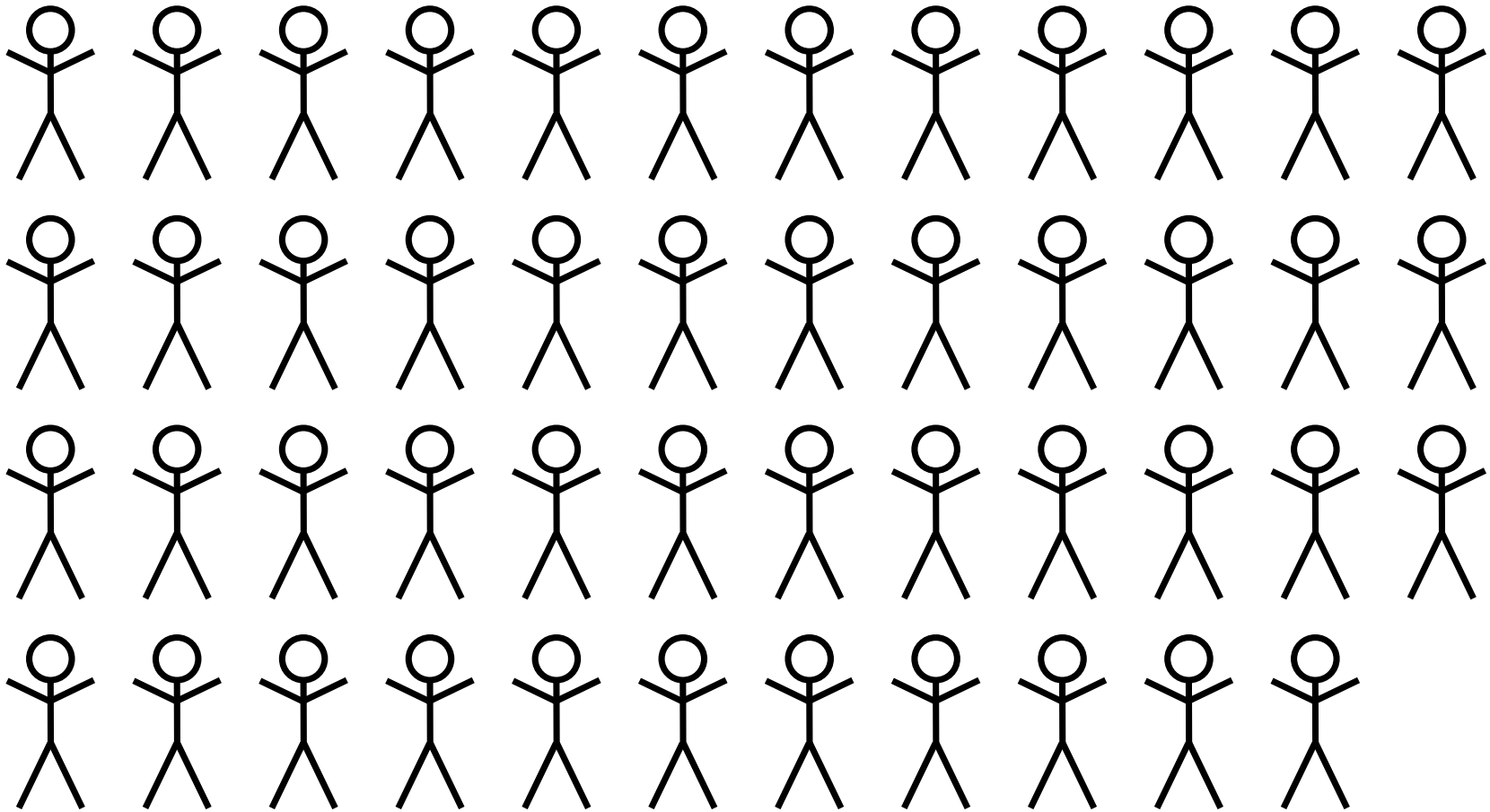
Motivation and Creativity – A Randomized Experiment

Is extrinsic or intrinsic motivation better for enhancing creativity?

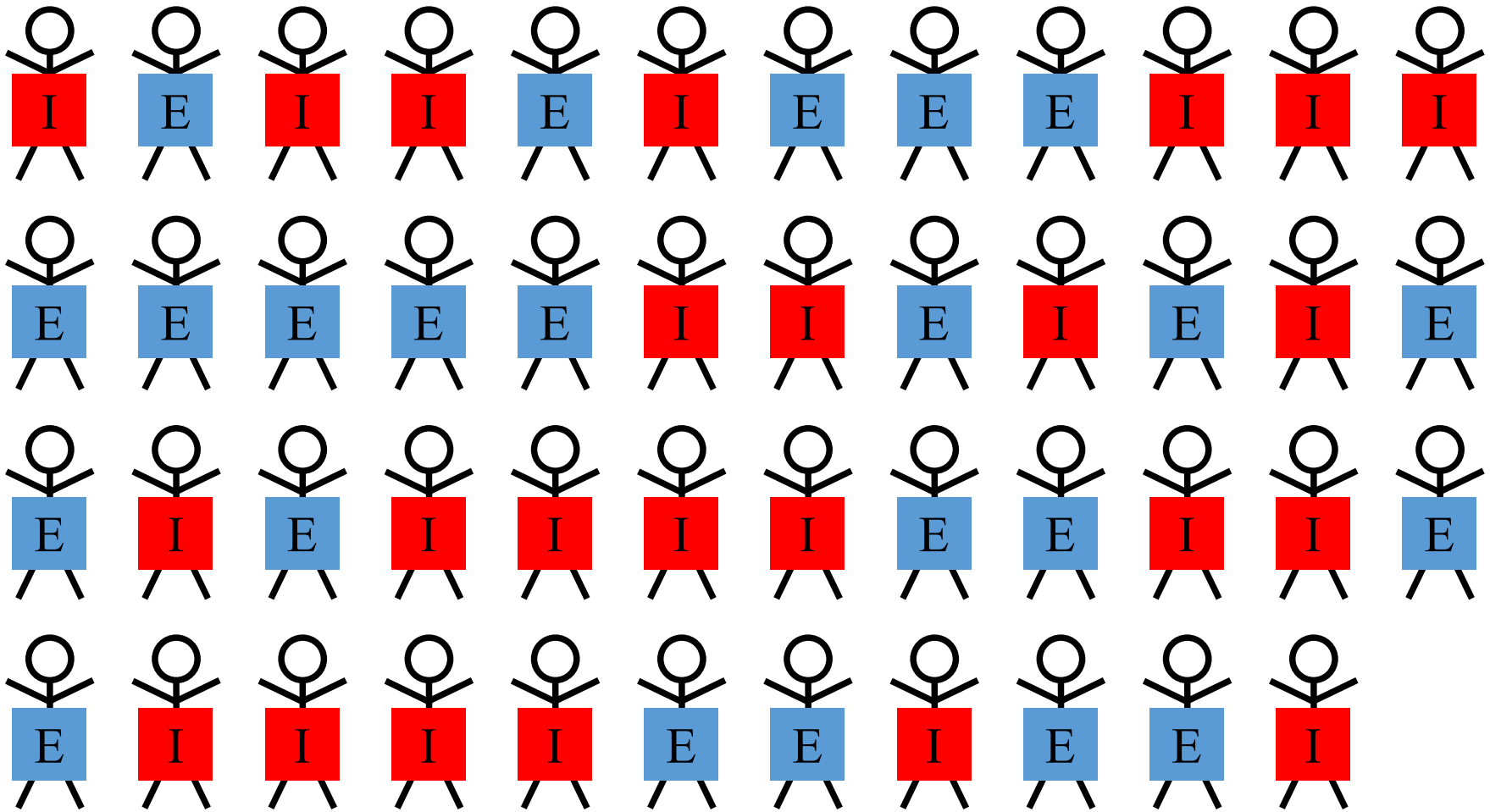
Ramsey, F., Schafer, D. (2012). *The Statistical Sleuth: a Course in Methods of Data Analysis*. Cengage Learning.

Amabile, T. M. (1985). Motivation and creativity: Effects of motivational orientation on creative writers. *Journal of Personality and Social Psychology*, 48(2), 393.

47 Subjects with Creative Writing Experience



Random Assignment of Two Treatments to Subjects



Extrinsic Treatment

Please rank the following list of reasons for writing, in order of personal importance to you.

- You realize the market for writing is constantly expanding.
- You want your writing teachers to be impressed with your writing.
- You have heard of cases where one bestselling novel or collection of poems has made the author financially secure.
- You enjoy public recognition of your work.
- You know many of the best jobs require good writing skills.
- You know that writing ability is one of the major criteria for acceptance into graduate school.
- Your teachers and parents have encouraged you to go into writing.

Intrinsic Treatment

Please rank the following list of reasons for writing, in order of personal importance to you.

- You get a lot of pleasure out of reading something good that you have written.
- You enjoy the opportunity for self expression.
- You achieve new insights through your writing.
- You derive satisfaction from expressing yourself clearly and eloquently.
- You feel relaxed when writing.
- You like to play with words.
- You enjoy becoming involved with ideas, characters, events, and images in your writing.

Response Variable Measured for Each Subject

- After treatment, each subject writes a poem in the Haiku style
- Each poem evaluated on a 40-point scale of creativity
- Evaluation blind to treatment

My Statistics Haiku

My Statistics Haiku

Statistics helps us

My Statistics Haiku

Statistics helps us

In a big uncertain world

My Statistics Haiku

Statistics helps us

In a big uncertain world

Find truth in data

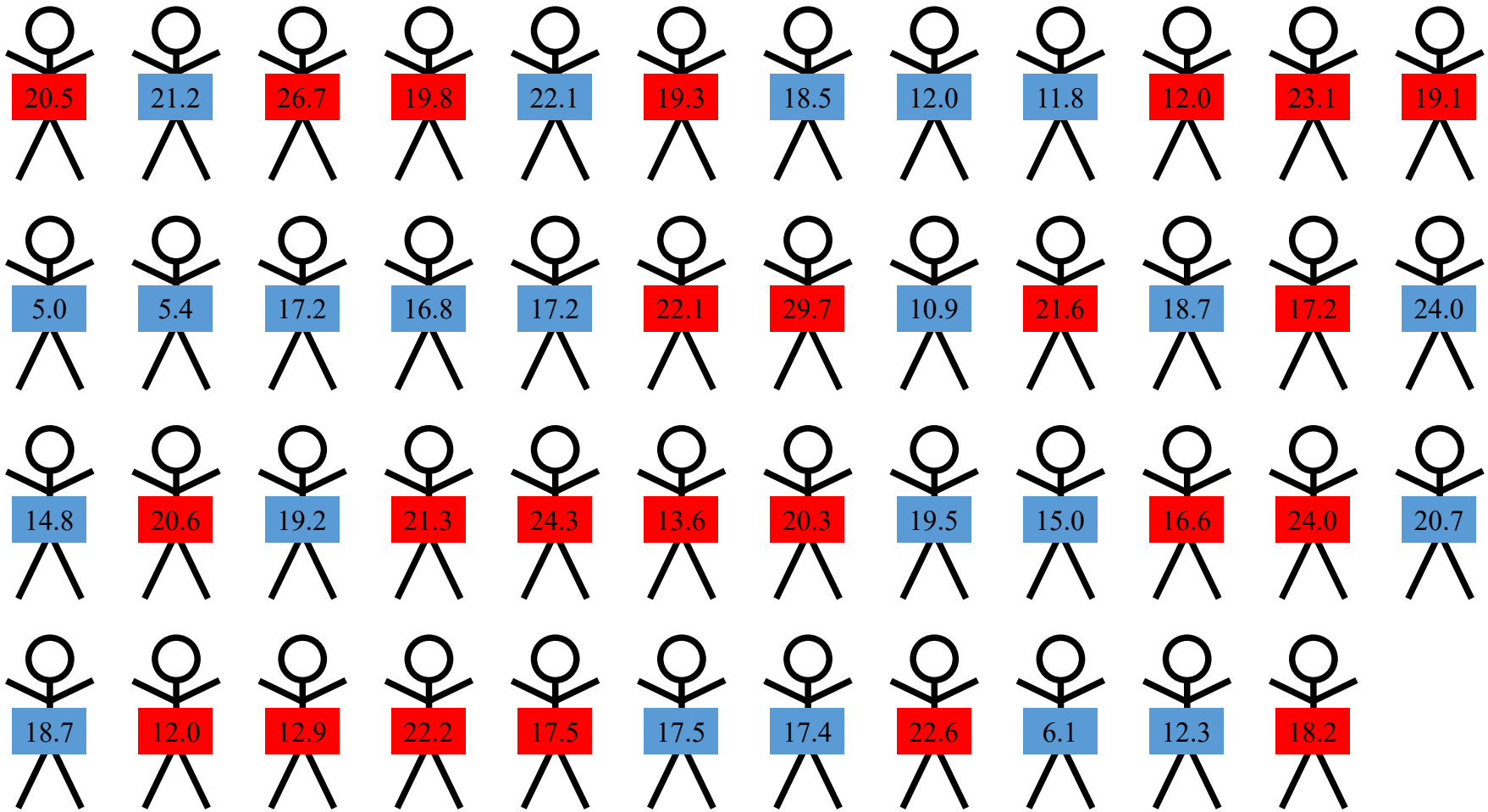
My Statistics Haiku

Statistics helps us

In a **big** uncertain world

Find truth in **data**

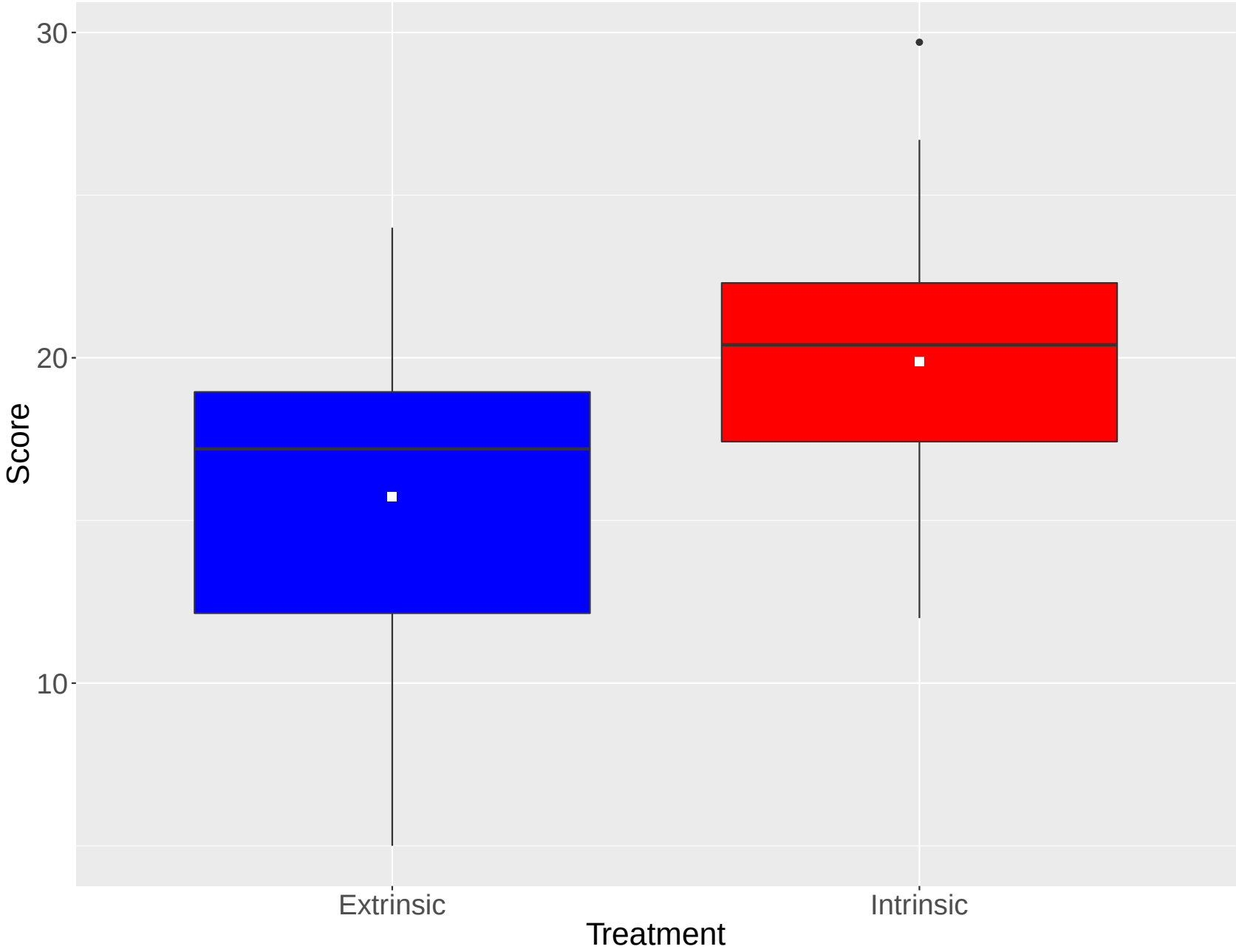
Creativity Scores for the 47 Subjects



Summary Statistics for the Creativity Scores

Statistic	Extrinsic	Intrinsic
Sample Size	23	24
Average	15.74	19.88
Standard Deviation	5.25	4.44
Maximum	24.00	29.70
Upper Quartile	18.95	22.30
Median	17.20	20.40
Lower Quartile	12.15	17.43
Minimum	5.00	12.00

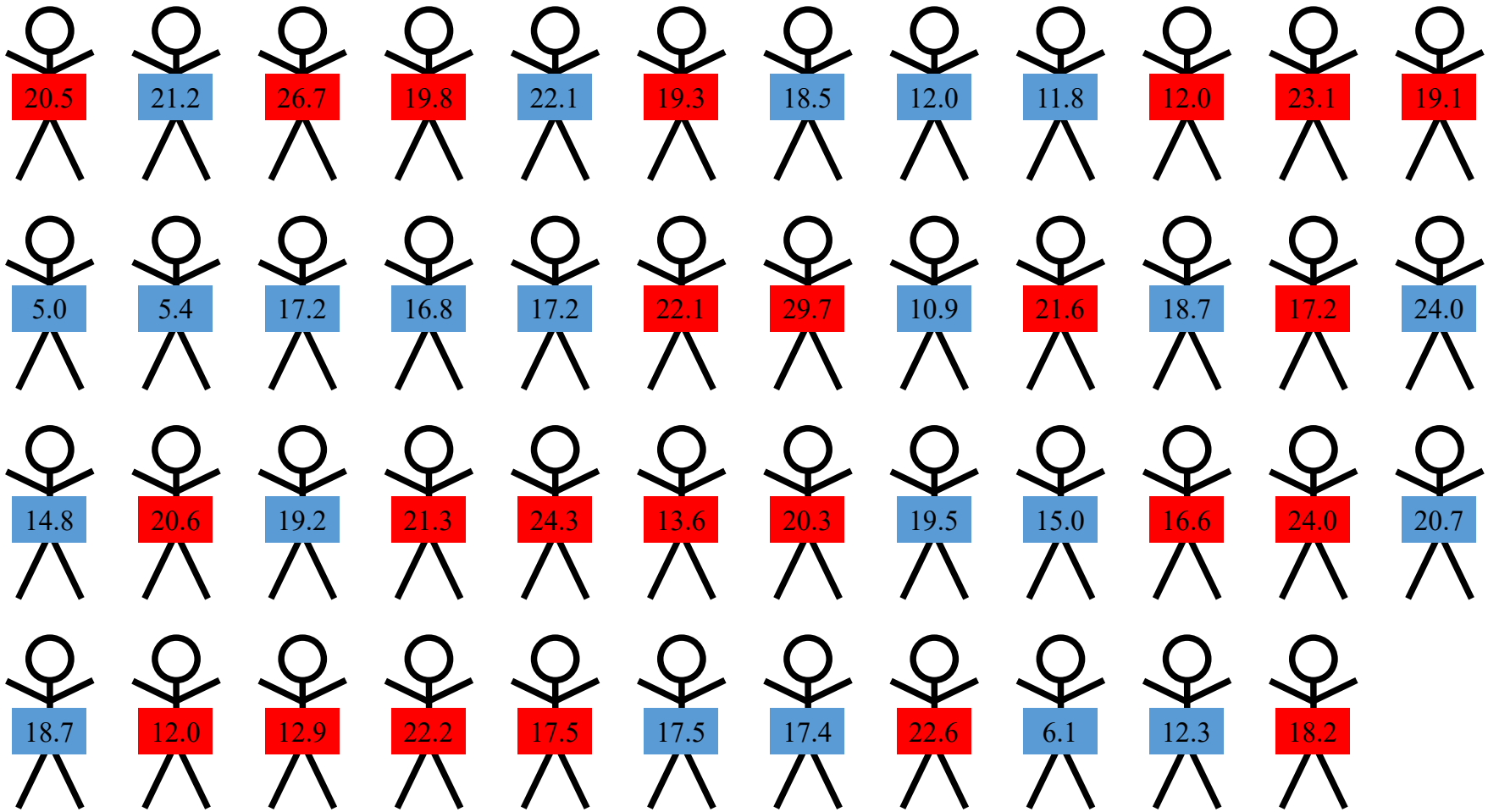
Side-by-Side Boxplots of the Creativity Scores



Is there evidence of a treatment effect on creativity?

- The subjects in the Intrinsic group scored 4.14 points higher, on average, than subjects in the Extrinsic group, but how do we know this difference had anything to do with the treatment?
- Maybe the poem each subject wrote was completely unaffected by the treatment.

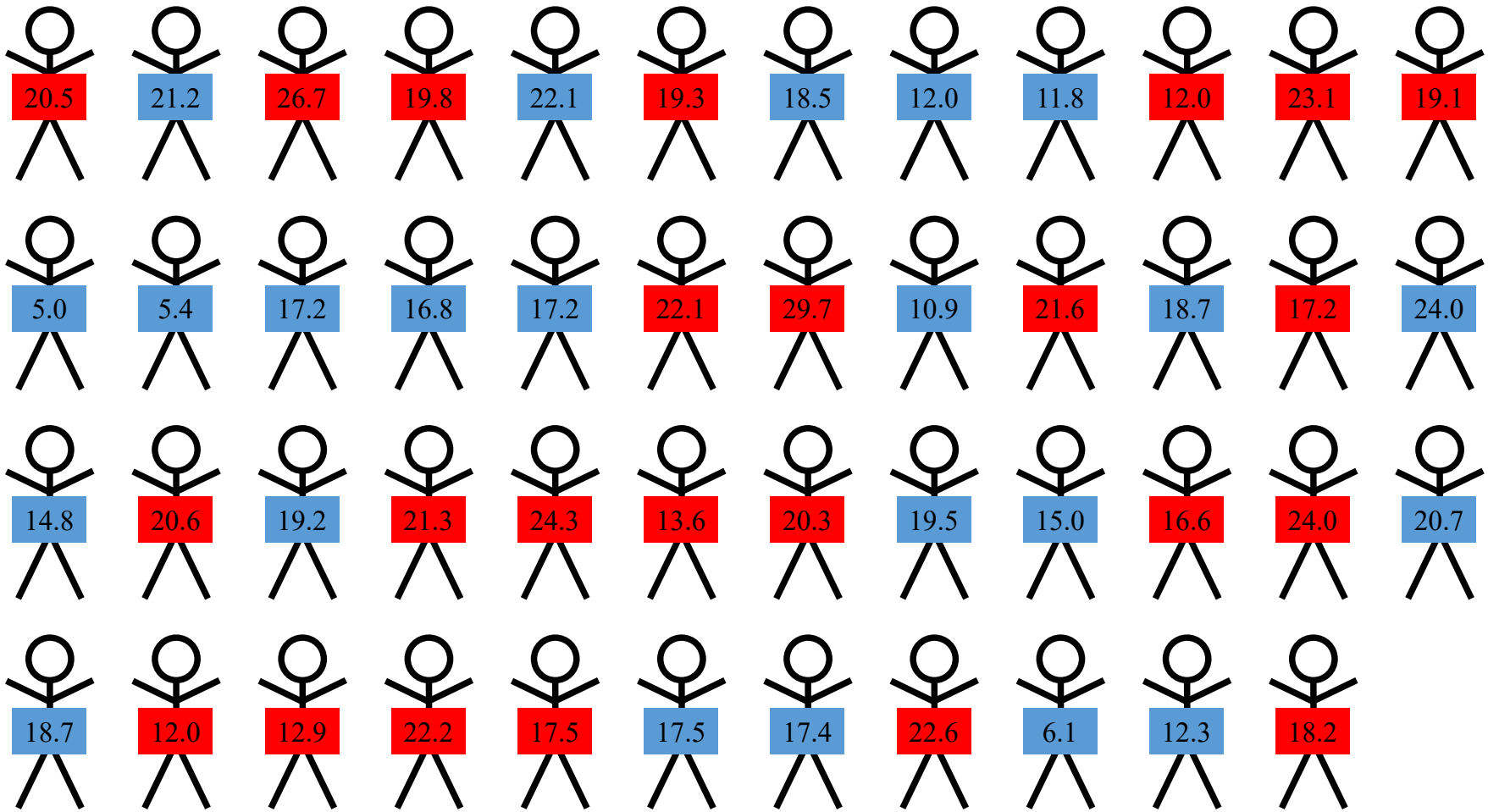
Creativity Scores for the 47 Subjects



Is there evidence of a treatment effect on creativity?

- If you randomly split a group of 47 subjects into two groups of 23 and 24 subjects, you wouldn't expect both groups to have identical average scores even if you treated them exactly alike.
- Maybe we just happened to pick, by chance, the more creative subjects for the Intrinsic group when randomly assigning the treatments, resulting in the difference of average scores equal to 4.14.
- What is the chance of a difference that big or bigger under the assumption of no treatment effect?

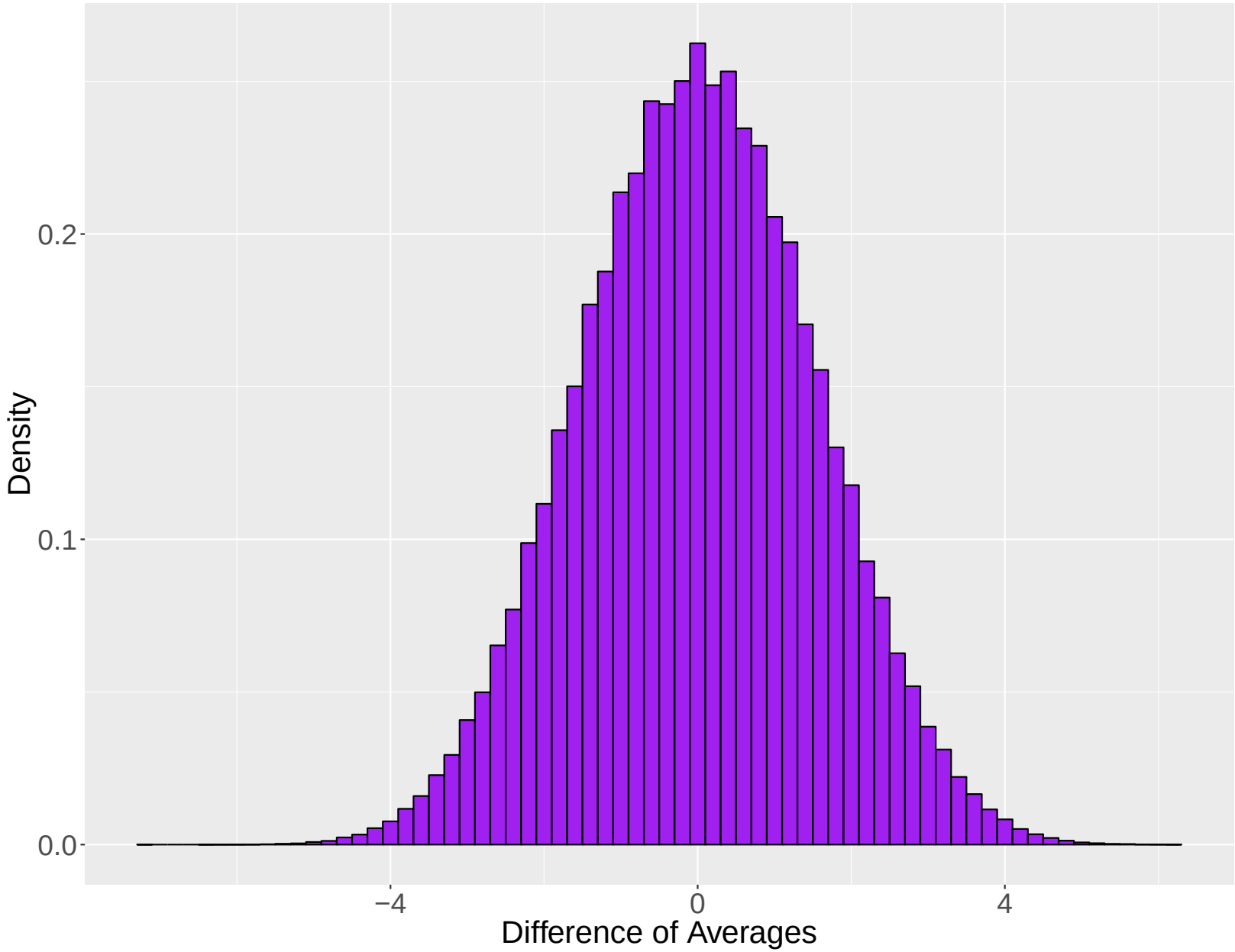
Creativity Scores for the 47 Subjects



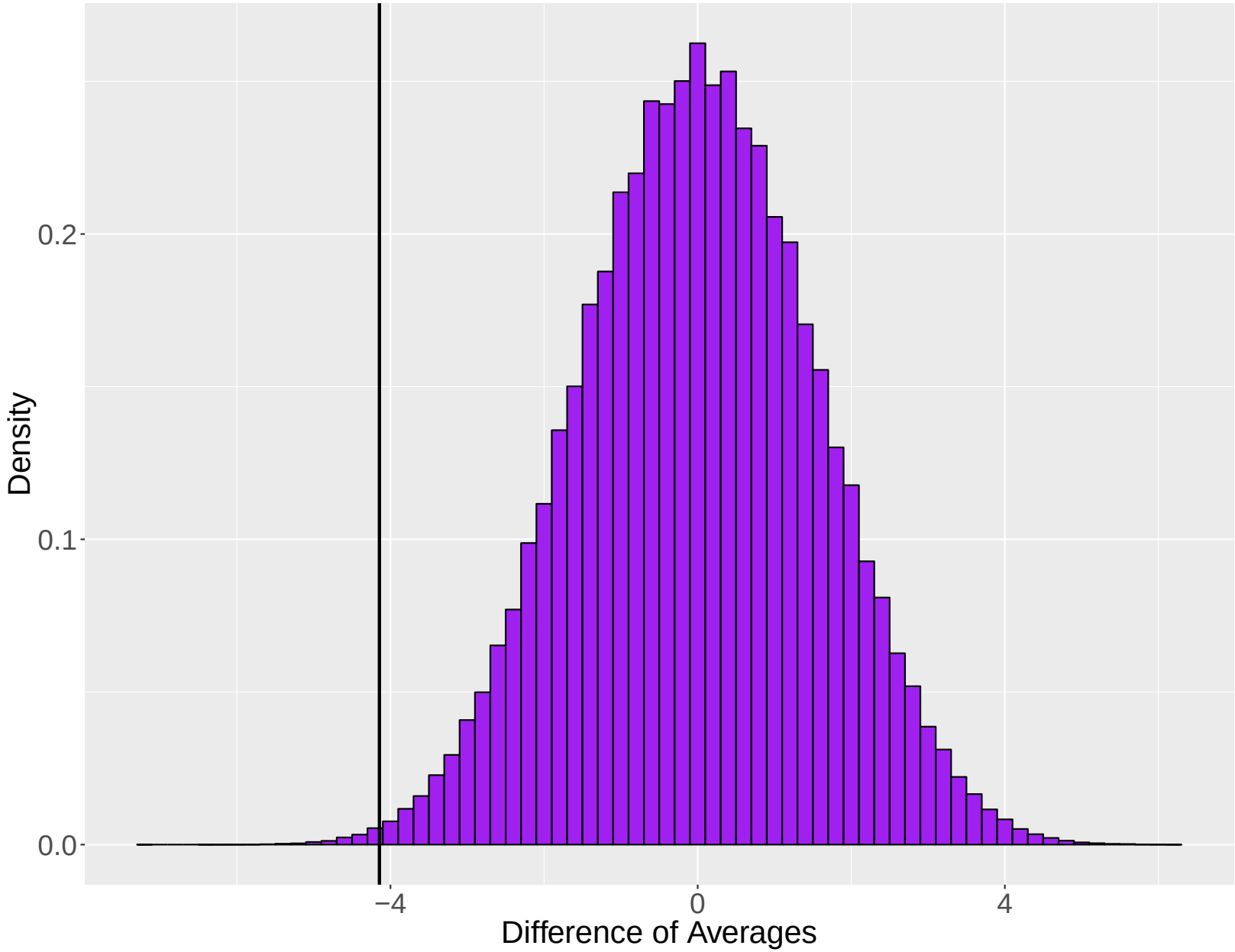
Is there evidence of a treatment effect on creativity?

- I randomly split the 47 creativity scores into two groups of 23 and 24 scores and found the difference between the average scores of the two groups.
- I repeated this process 1,000,000 times and kept track of the 1,000,000 differences.
- 5,030 times out of 1,000,000 (i.e., 0.00503 or 0.503%) the difference in averages was as far or farther from zero than the 4.14 difference in our experiment.

Histogram of 1,000,000 Differences



Histogram of 1,000,000 Differences



Is there evidence of a treatment effect on creativity?

- If the treatment has no effect on creativity, it is quite unlikely (probability about 0.005) to see a difference in averages as far or farther from zero than the 4.14 difference in our experiment.
- Thus, it seems reasonable to believe that the treatment **caused** the difference in average creativity scores.

The p-value

- 0.005 is an example of a p-value.
- Researchers often use p-values to determine if a *null hypothesis* seems plausible based on the observed data.
- The null hypothesis is this example is

H_0 : The treatment has no effect on creativity scores.

Interpreting the p-value

- In this example, the p-value tells us that the probability of ending up with a difference in averages as large or larger than the difference we saw in the observed data (4.14) would be very unlikely (probability 0.005) if the null hypothesis were true.
- Thus, the null hypothesis does not seem plausible.
- We would reject this null hypothesis and conclude that the Intrinsic treatment raised creativity scores relative to the Extrinsic treatment.

Understanding the p-value

- Note that the p-value tells us about the probability of a data result under the assumptions that the null hypothesis is true.
- The p-value is NOT the probability that the null hypothesis is true, given the data.
- p-values are often used but have been the subject of considerable controversy recently.

Why Study the Discipline of Statistics?

Why Study the Discipline of Statistics?

You derive satisfaction from designing a study well and expertly drawing conclusions from the resulting data.

Why Study the Discipline of Statistics?

You derive satisfaction from designing a study well and expertly drawing conclusions from the resulting data.

NOT because knowing statistics will help you obtain a well-paying job.

A New Example

- In 1972, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job.
- The files were identical except that half of them indicated that the file was that of a female and the other half indicated that the file was that of a male.

A New Example (continued)

- The 24 male and 24 female files were randomly distributed to the 48 bank executives.
- The complete experiment is described in Rosen, B. and T. Jerdee (1974). Influence of sex role stereotypes on personnel decisions. *Journal of Applied Psychology*. 59, 9-14.

Is there evidence of gender discrimination?

	Promote?	
	Yes	No
Female File	14	10
Male File	21	3

Design a simulation using a deck of cards that will help determine whether the gender of the file affected the promotion decisions.

Analogous to the Creativity Study

- subjects $\Longleftrightarrow$ bank executives
- extrinsic and intrinsic $\Longleftrightarrow$ female and male
- creativity scores $\Longleftrightarrow$ promotion decisions
- difference in average creativity scores

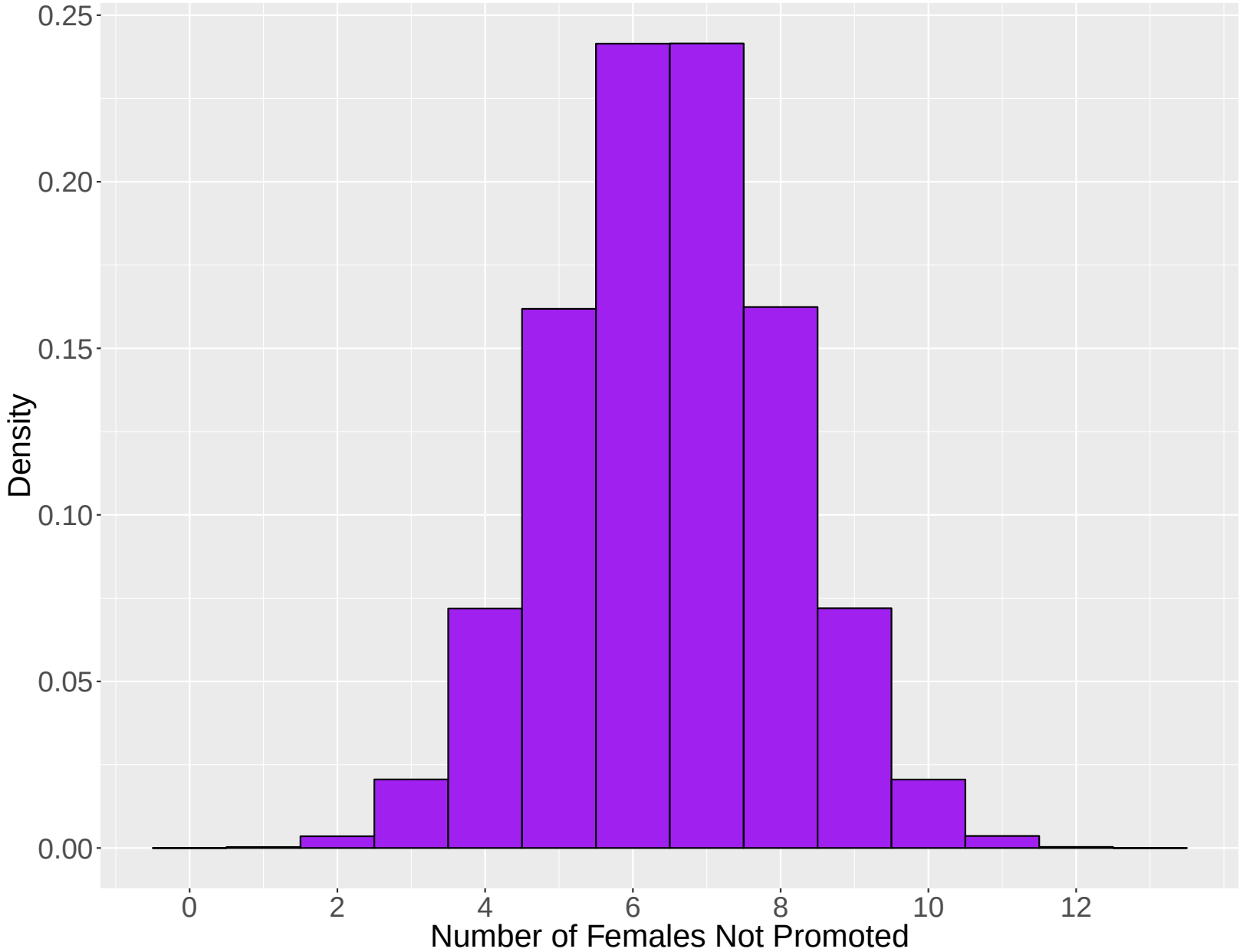
  $\Longleftrightarrow$ difference in proportion promoted

  $\Longleftrightarrow$ number of females not promoted

Evidence of Gender Discrimination?

- Suppose none of the bank executives were influenced by the gender associated with the file.
- In the experiment, we can see that 35 bank executives recommended promotion and 13 bank executives recommended against promotion.
- What is the chance that 10 or more of the female files would end up in the hands of the 13 bank executives who recommended against promotion?

Histogram of 1,000,000 Results



Evidence of Gender Discrimination?

- When randomly determining which of the 48 files goes to the 13 bank executives who recommended against promotion, the number of female files not recommended for promotion is 10 or more with probability about 0.0245.
- Thus, if there was no gender discrimination, the results in the actual experiment would be somewhat rare.

Other Important Statistical Concepts

So far, we have learned about the statistical method of *hypothesis testing*. Some other important statistical concepts include

- populations
- parameters
- samples
- statistics
- standard errors
- confidence intervals

Populations and Parameters

Population: the complete collection of all items of interest

- All the ears of corn in a field
- A collection of pennies in a bag

Parameter: a numerical characteristic of the population

- The mean number of kernels per ear
- The mean age of the pennies in a bag

Samples and Statistics

Sample: the subset of the population that is observed

- 10 ears of corn randomly selected from a field
- 10 pennies selected at random from a bag of pennies

Statistic: a numerical characteristic of the sample

- The average number of kernels on the 10 ears selected at random from the field
- The average age of the 10 pennies selected at random from the bag of pennies

We often use Greek letters to denote population parameters.

- μ usually denotes the population mean.
- σ usually denotes the population standard deviation.

We estimate a population parameter by computing a sample statistic.

- The sample average $\bar{Y}$ is used to estimate the population mean μ.
- The sample standard deviation s is used to estimate the population standard deviation σ.

Standard Errors

We can use the information in a random sample to compute

- statistics that estimate population parameters, AND
- **standard errors** that tell us how far from the population parameters the statistics are likely to be.

Standard Error Example

A **standard error** is the estimated standard deviation of a statistic.

For example, if $\bar{Y}$ is the average of a sample of n values drawn from a population with mean μ and standard deviation σ, then . . .

- The standard deviation of $\bar{Y}$ is $\sigma/\sqrt{n}$.
- The standard error of $\bar{Y}$ is $s/\sqrt{n}$.

Understanding a Statistic's Standard Deviation

- At first, it may be difficult to understand how a statistic (like $\bar{Y}$) can have a standard deviation.
- After all, standard deviation quantifies variability in a set of numbers, and the value of the statistic we compute from a sample is only one number.
- However, there are many possible samples.
- The value of the statistic that we compute from a sample is only one of many possible statistic values we could have seen.

Understanding a Statistic's Standard Deviation

- Imagine that we write down the value of the statistic for each possible sample.
- The standard deviation of these values is the standard deviation of the statistic.
- The fact that we can get an estimate of a statistic's standard deviation by seeing only one sample is really quite amazing.

Confidence Interval

- **Confidence Interval**: an interval of values computed from a random sample that will contain a population parameter with a specified level of confidence (e.g., 95%).
- A 95% confidence interval is produced using a method that – for 95% of all samples of a given size – will provide an interval that contains the true population parameter.

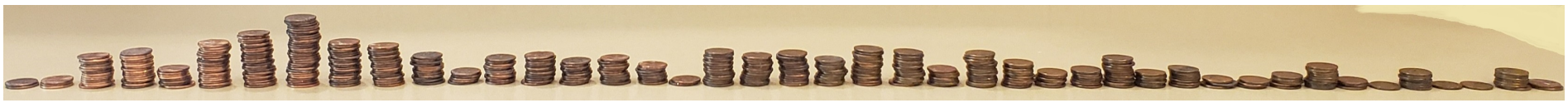
Confidence Interval for a Population Mean

An approximate 95% confidence interval
for a population mean μ is

$$\text{sample average} \pm 2 \times \text{standard error of sample average}$$

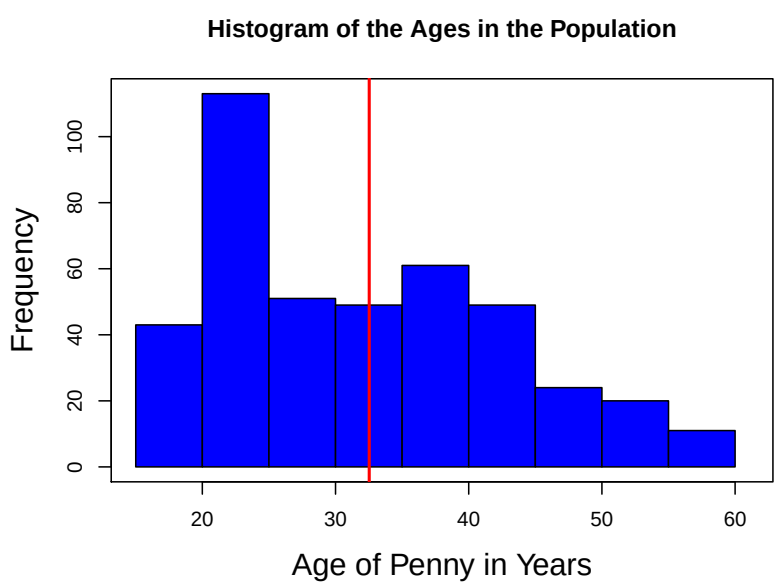
$$\bar{Y} \pm 2 \times s/\sqrt{n}$$

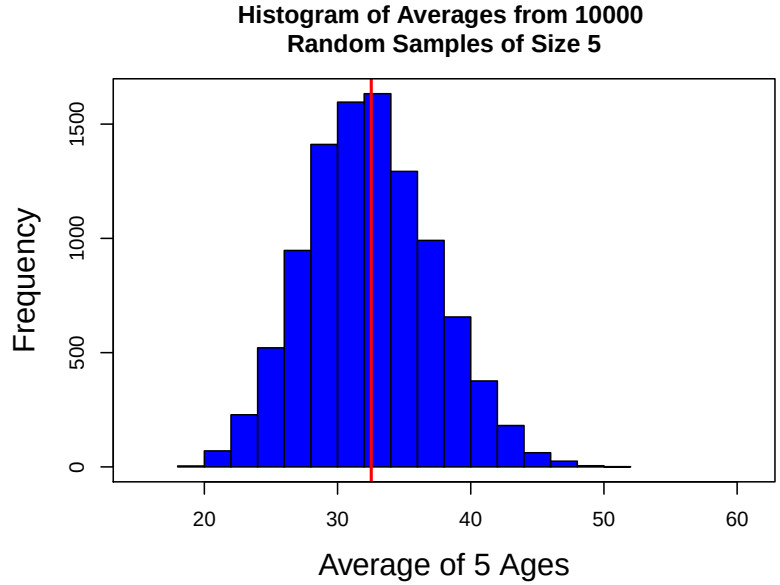
Pennies in the Bag Sorted by Age

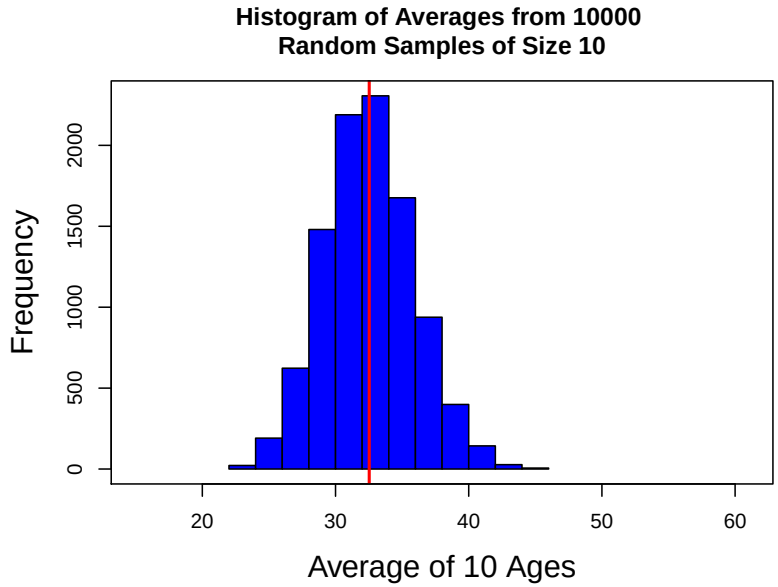


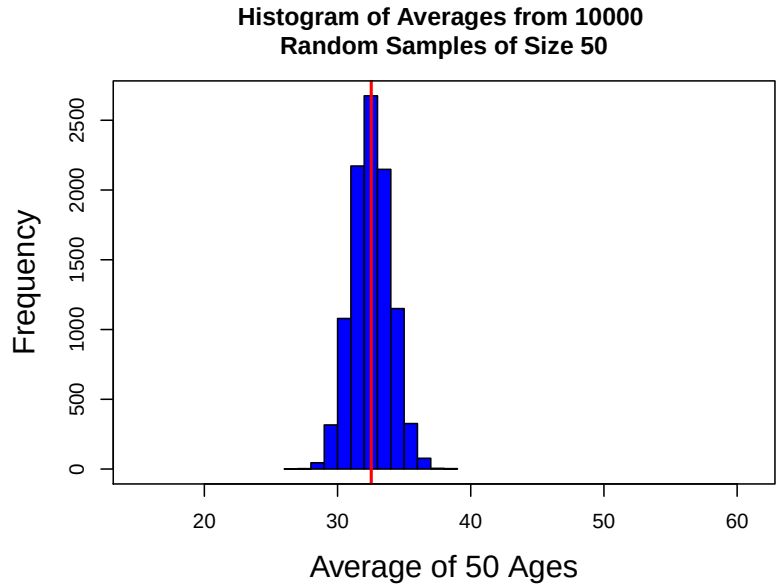
$$\mu = 32.53 \text{ years}$$

$$\sigma = 10.74 \text{ years}$$

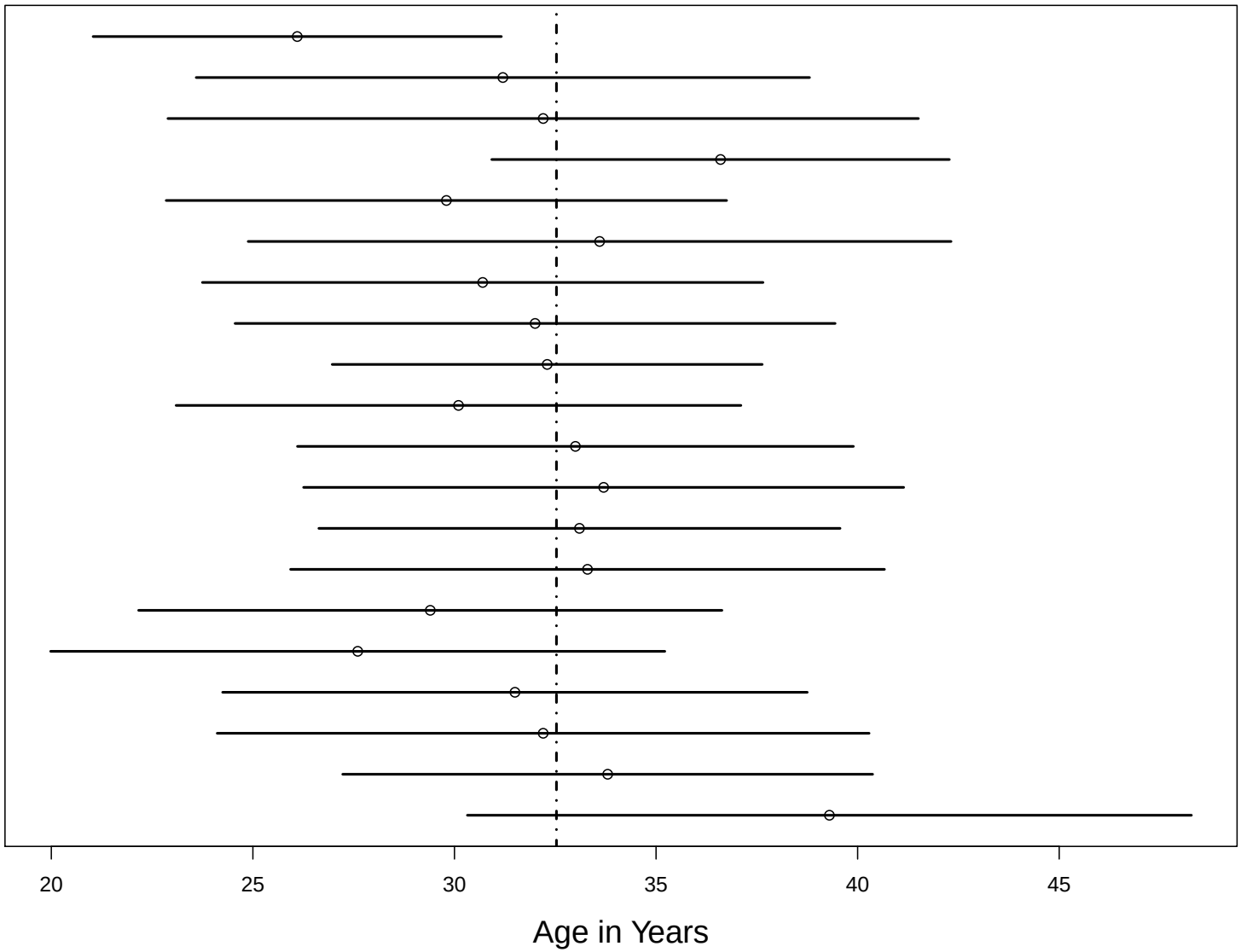

Histogram of the Ages in the Population
Frequency
0
20
40
60
80
100
20
30
40
50
60
Age of Penny in Years


Histogram of Averages from 10000
Random Samples of Size 5
Frequency
0
500
1000
1500
20
30
40
50
60
Average of 5 Ages


Histogram of Averages from 10000
Random Samples of Size 10
Frequency
0
500
1000
1500
2000
20
30
40
50
60
Average of 10 Ages

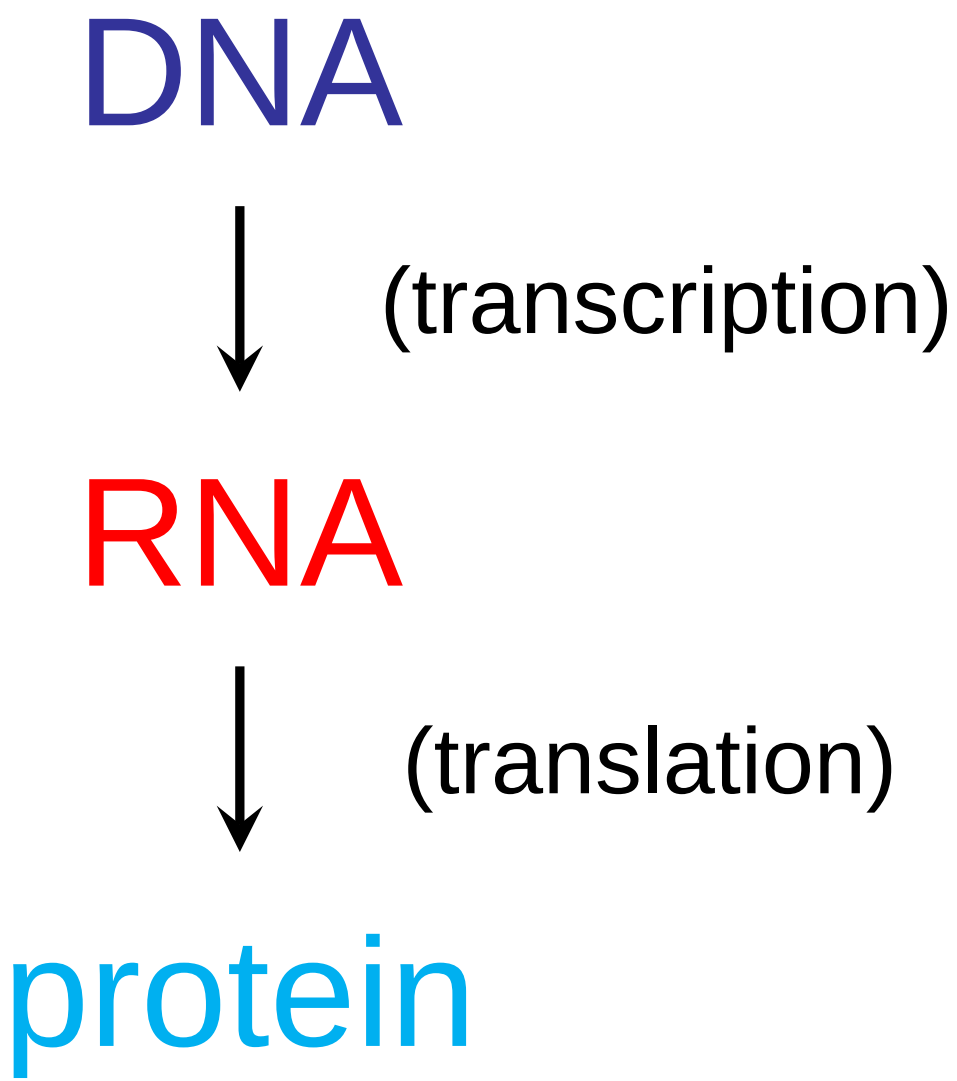

Histogram of Averages from 10000
Random Samples of Size 50
Frequency
0
500
1000
1500
2000
2500
20
30
40
50
60
Average of 50 Ages

20 Confidence Intervals Based on Samples of Size 10

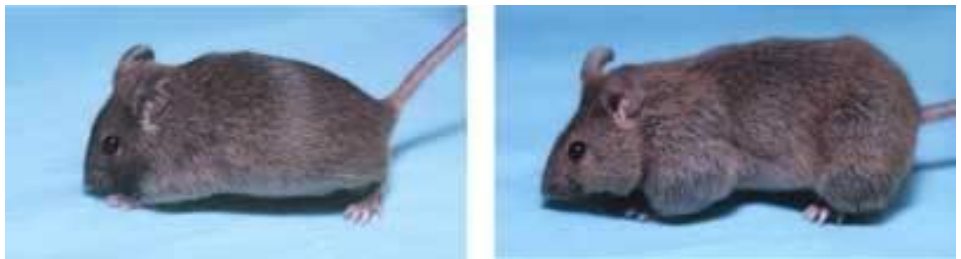


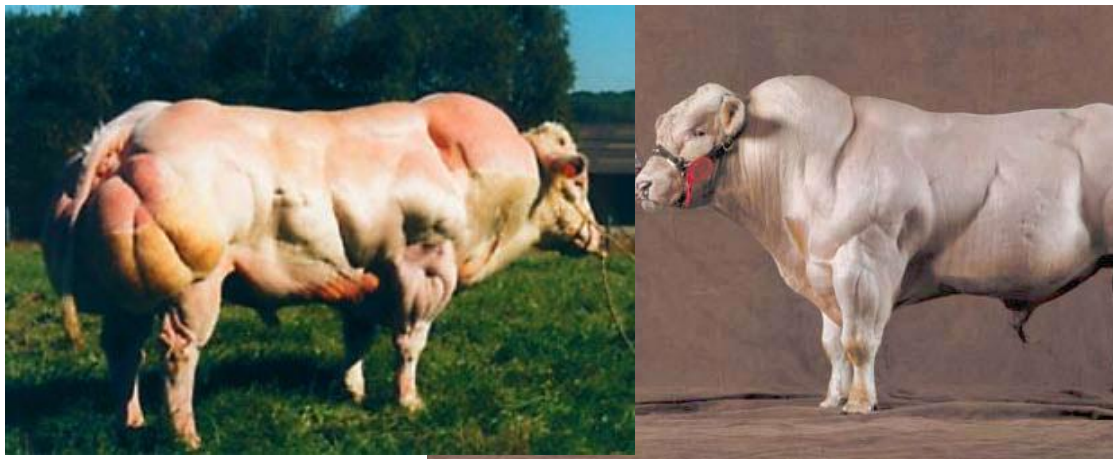
Research Areas

- Developing methods for the analysis of gene expression data
-
-

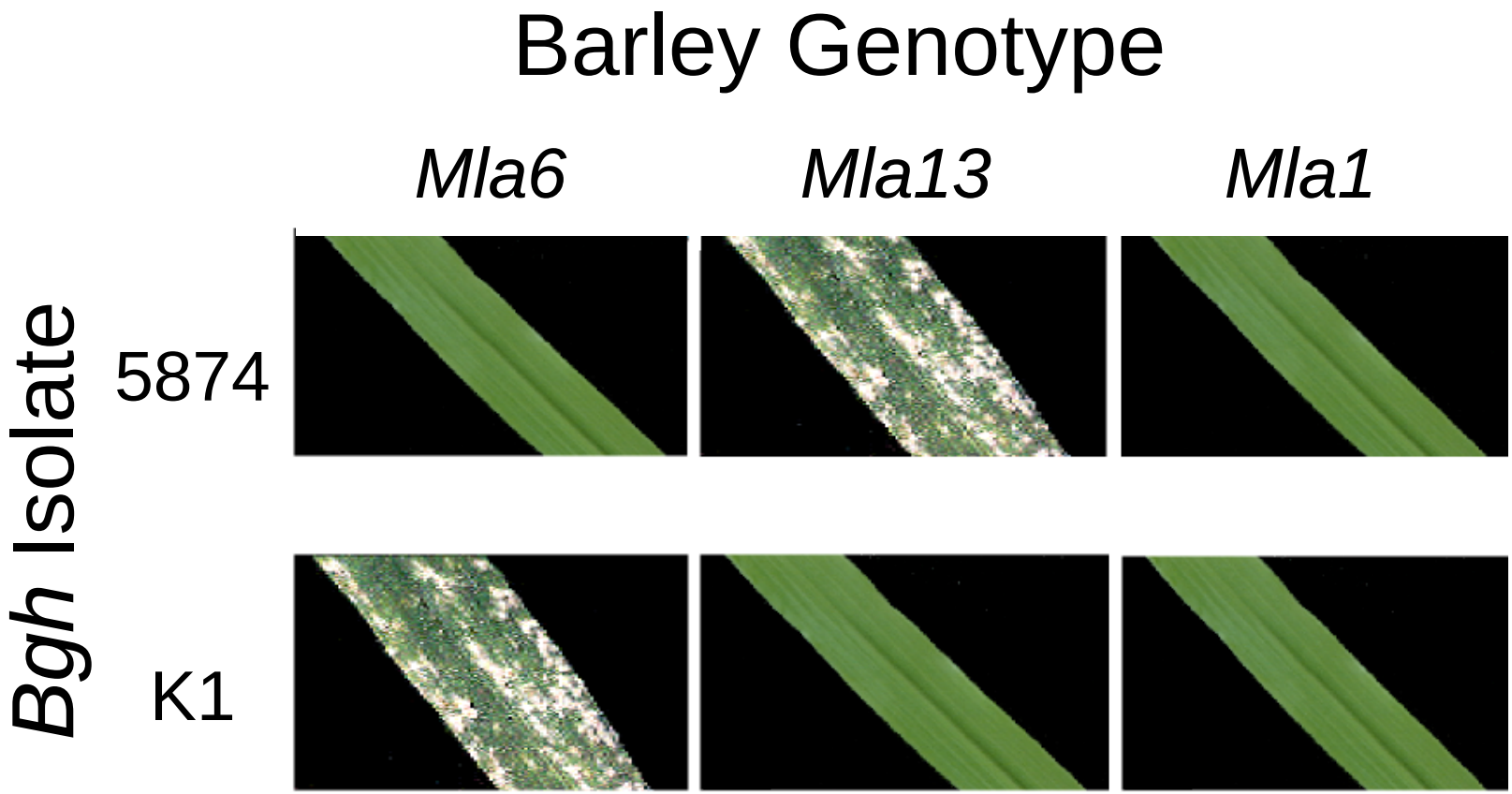

DNA
(transcription)
RNA
(translation)
protein

Myostatin Knockout Mice vs. Wild Type





Belgian Blue cattle have a mutation in the myostatin gene.



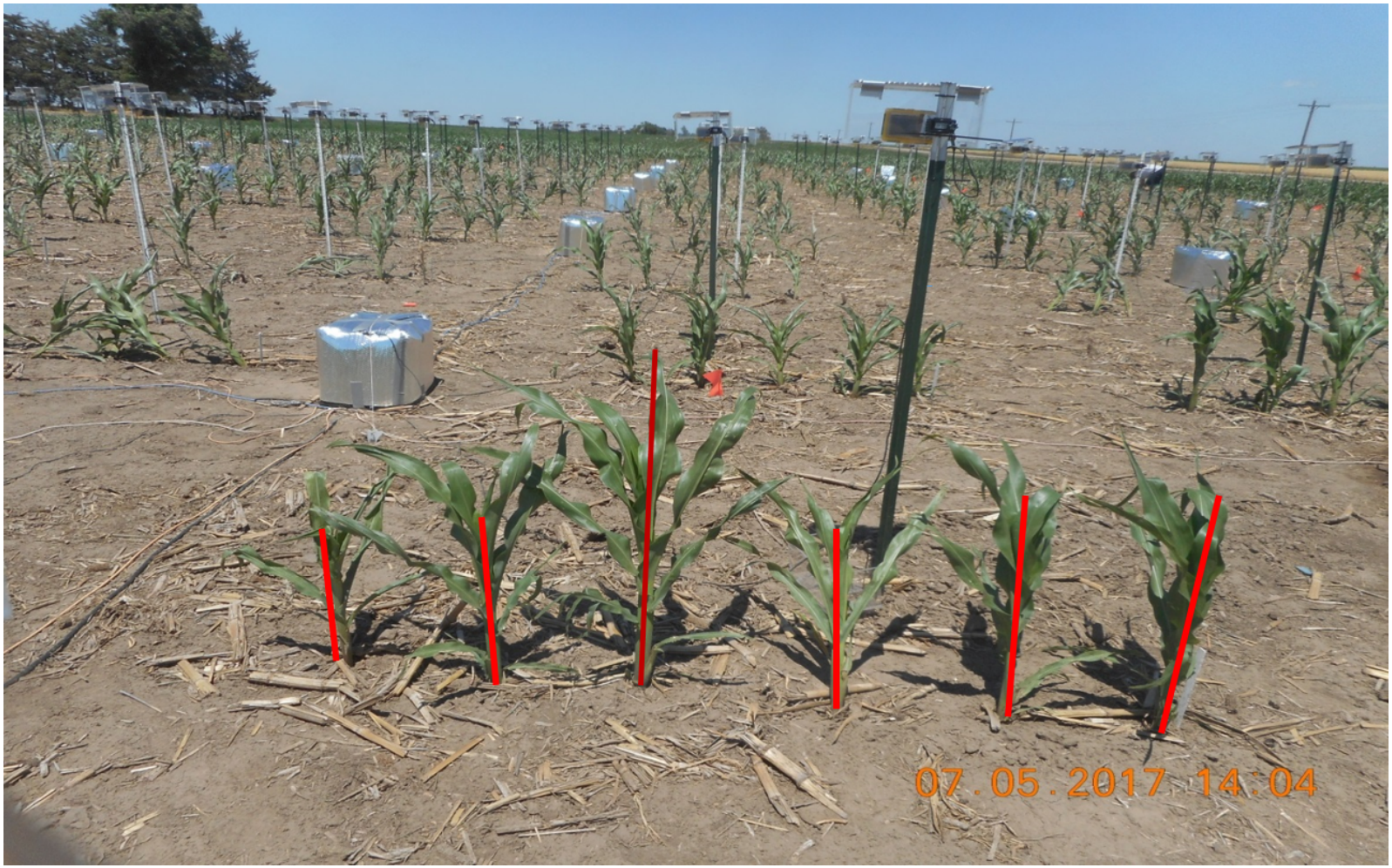
B73 F1 Mo17 B73 F1 Mo17

Swanson-Wagner, R., Jia, Y., DeCook, R., Borsuk, L. Nettleton, D., Schnable, P.S. (2006) *PNAS*. **103**, 6805-6810.

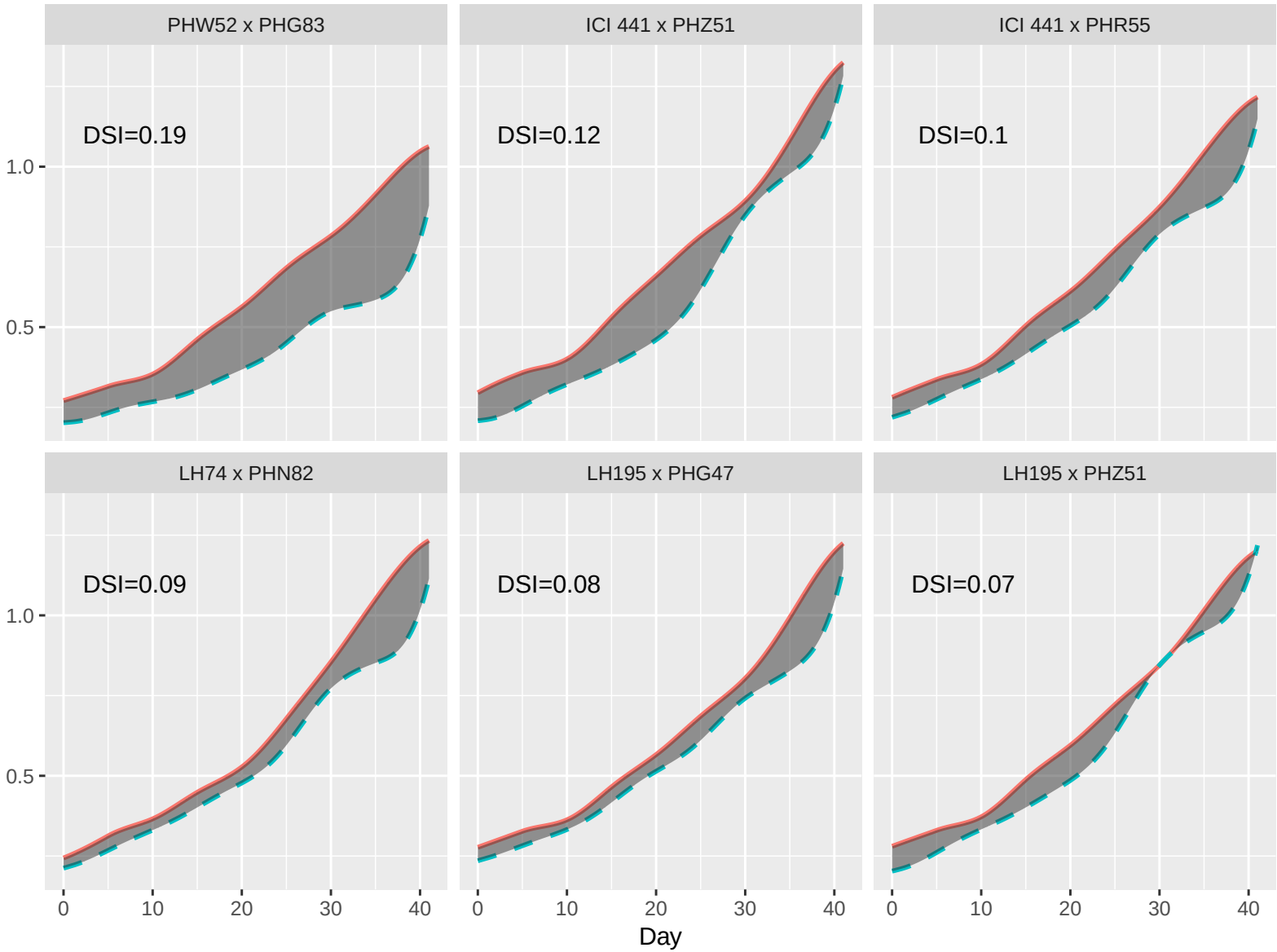
Research Areas

- Developing methods for the analysis of gene expression data
- Developing methods for predictive phenomics
-

Measuring Plant Phenotypes from Images



Drought-Sensitivity Index



Research Areas

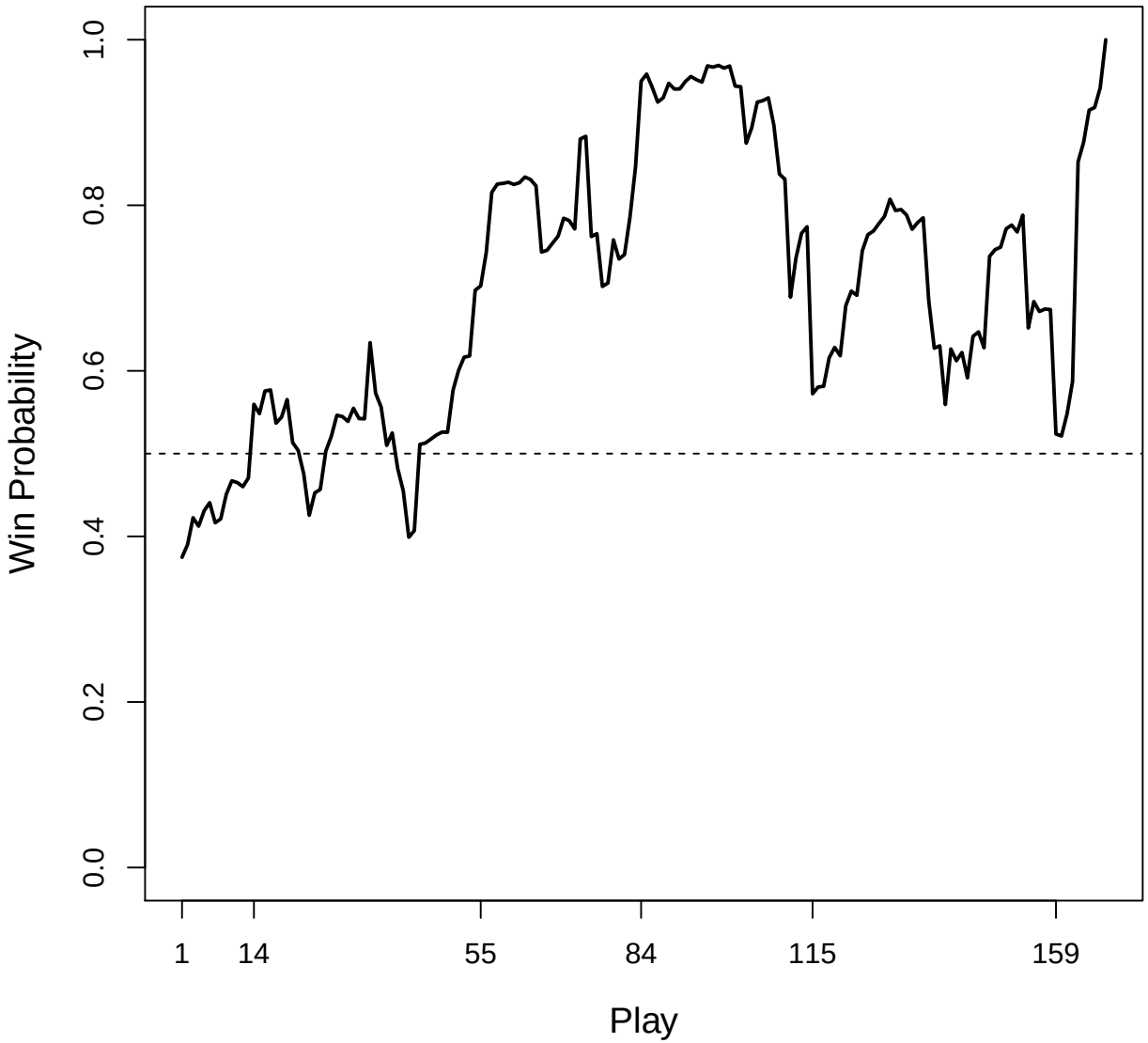
- Developing methods for the analysis of gene expression data
- Developing methods for predictive phenomics
- Improving random forest methodology

Random Forests: a Versatile Tool for Prediction

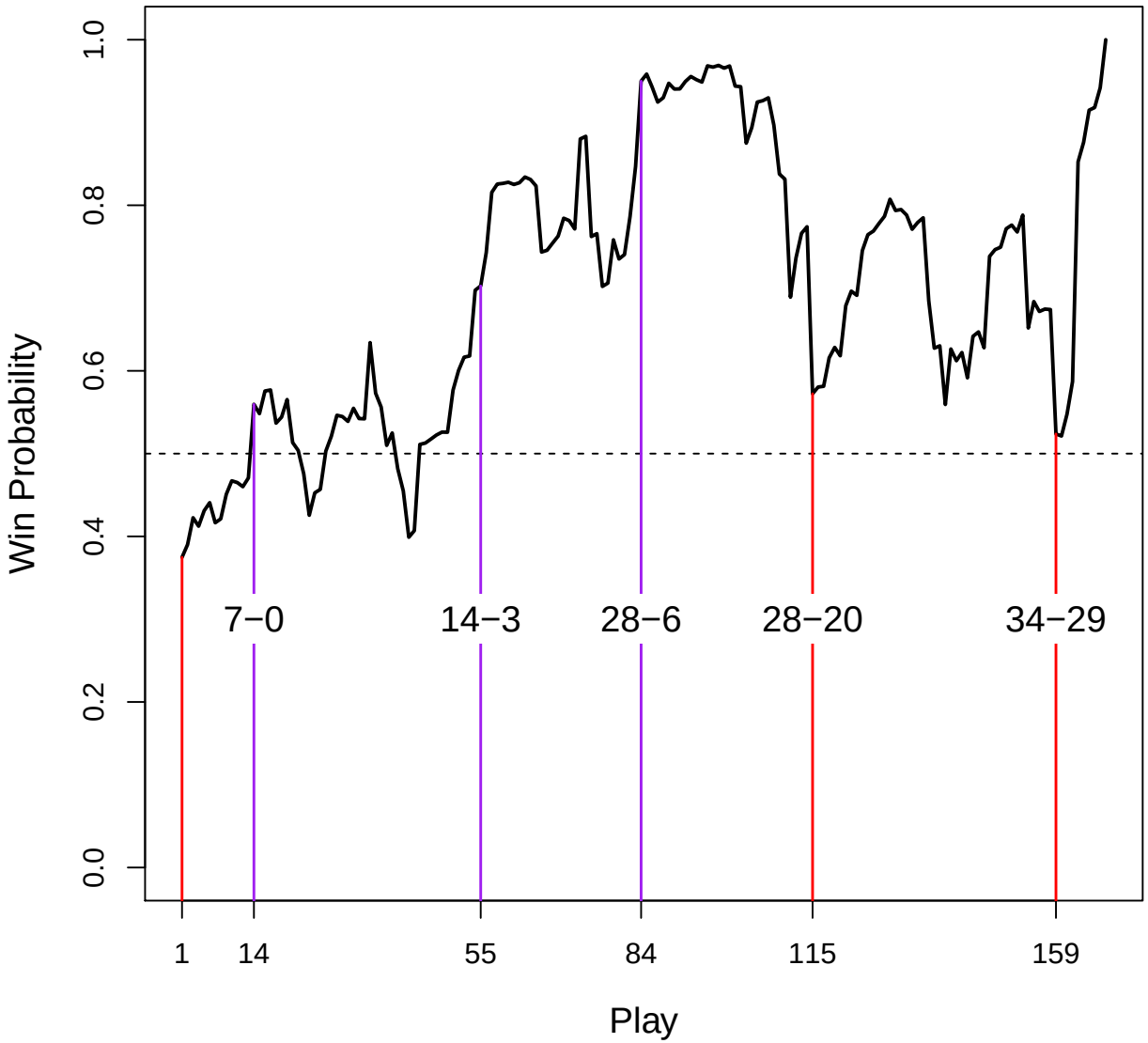
We have used random forests (Breiman, 2001) to predict

- house selling price,
- protein separation characteristics,
- lifestyle of a fungus from protein sequence data,
- mining site potential from measures of surface elements,
- student risk for leaving STEM majors at ISU,
- corn yield from environment and genotype, and
- win probability during the course of NFL games.

Super Bowl XLVII: Baltimore 34 vs. San Francisco 31



Super Bowl XLVII: Baltimore 34 vs. San Francisco 31



Variables from NFL Play-by-Play Data

$X_1 =$ *spread* (Las Vegas pre-game point spread)

$X_2 =$ *score* (score difference: offense - defense)

$X_3 =$ *seconds* (seconds remaining in the game)

$X_4 =$ *adjusted score* ($score/\sqrt{seconds + 1}$)

$X_5 =$ *totp* (total points scored)

$X_6 =$ *yardline* (yards from own goal line)

$X_7 =$ *down* (1st, 2nd, 3rd, or 4th down)

$X_8 =$ *ytd* (yards to go for a 1st down)

$X_9 =$ *timo* (time outs remaining for offense)

$X_{10} =$ *timd* (time outs remaining for defense)

$Y =$ 0-1 indicator of whether the team on offense won

ESPN Stats & Info Plot for Super Bowl LI

Some Responses on Twitter

Bobby Manning @RealBobManning @ESPNStatsInfo

> This is why win probability is nonsense. I'd need 10 hands to count the # of times Ive seen a $95+$% team lose.

Some Responses on Twitter

Bobby Manning @RealBobManning @ESPNStatsInfo

This is why win probability is nonsense. I'd need 10 hands to count the # of times Ive seen a $95+$% team lose.

Cody Taylor @CTaylor412 @RealBobManning @ESPNStatsInfo

and probably 200 hands to count the number of times the 95% team wins

Some Responses on Twitter

Bobby Manning @RealBobManning @ESPNStatsInfo

This is why win probability is nonsense. I'd need 10 hands to count the # of times Ive seen a $95+$% team lose.

Cody Taylor @CTaylor412 @RealBobManning @ESPNStatsInfo

and probably 200 hands to count the number of times the 95% team wins

racquetman @racquetman75 @ESPNStatsInfo

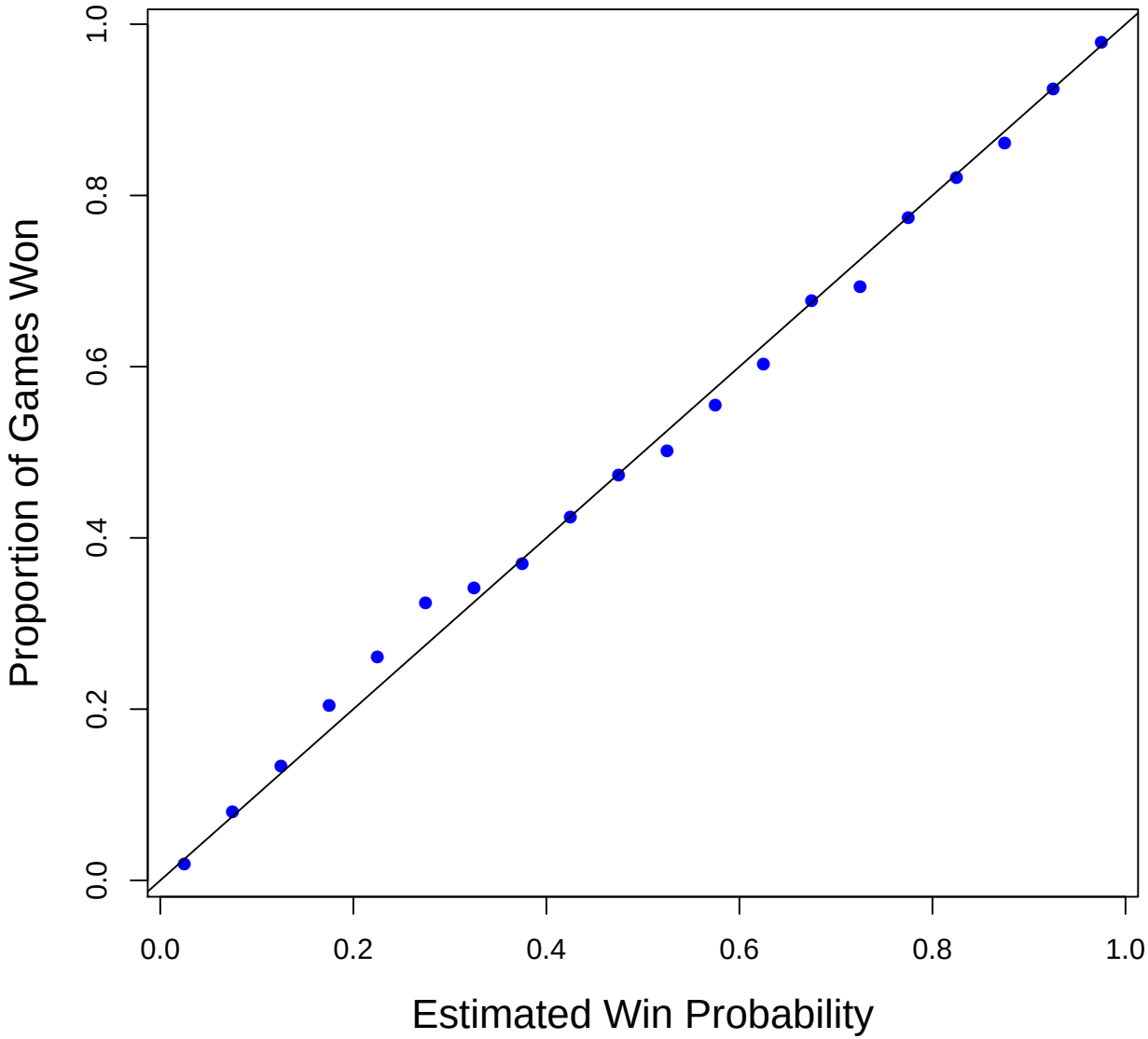
There may be nothing dumber than these stupid win probabilities.

Evaluation of Random Forest Win Probabilities

The training dataset consists of $N = 430168$ plays from NFL seasons 2001–2011.

The test dataset consists of all play from the 2012 NFL season.

Accuracy of Random Forest WP Estimates



Dennis Lock Featured in *Significance*

Dennis Lock Featured in *Significance*

Work Experience of Former Students

- Entertainment Companies: NFL, Disney
- Tech Companies: Google, Amazon, PayPal, eBay, Microsoft Research
- Pharmaceutical Industry: Eli Lilly
- Government: National Center for Toxicological Research, National Institute of Standards and Technology
- Medical Device Industry: Gore
- Medical Research: Children's Hospital of Philadelphia, Mayo Clinic, OptumInsight, St. Jude Children's Research Hospital
- Academia: Fudan, Illinois-Chicago, Iowa, Lawrence, Melbourne, Missouri-Columbia, Nebraska, North Dakota State, Old Dominion, Waterloo, Wright State