浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

Bootstrapping

ADS2, Lecture 2.14

Dr Rob Young – robert.young@ed.ac.uk

Semester 2, 2023/24

We’ve talked a lot about assumptions in ADS2…

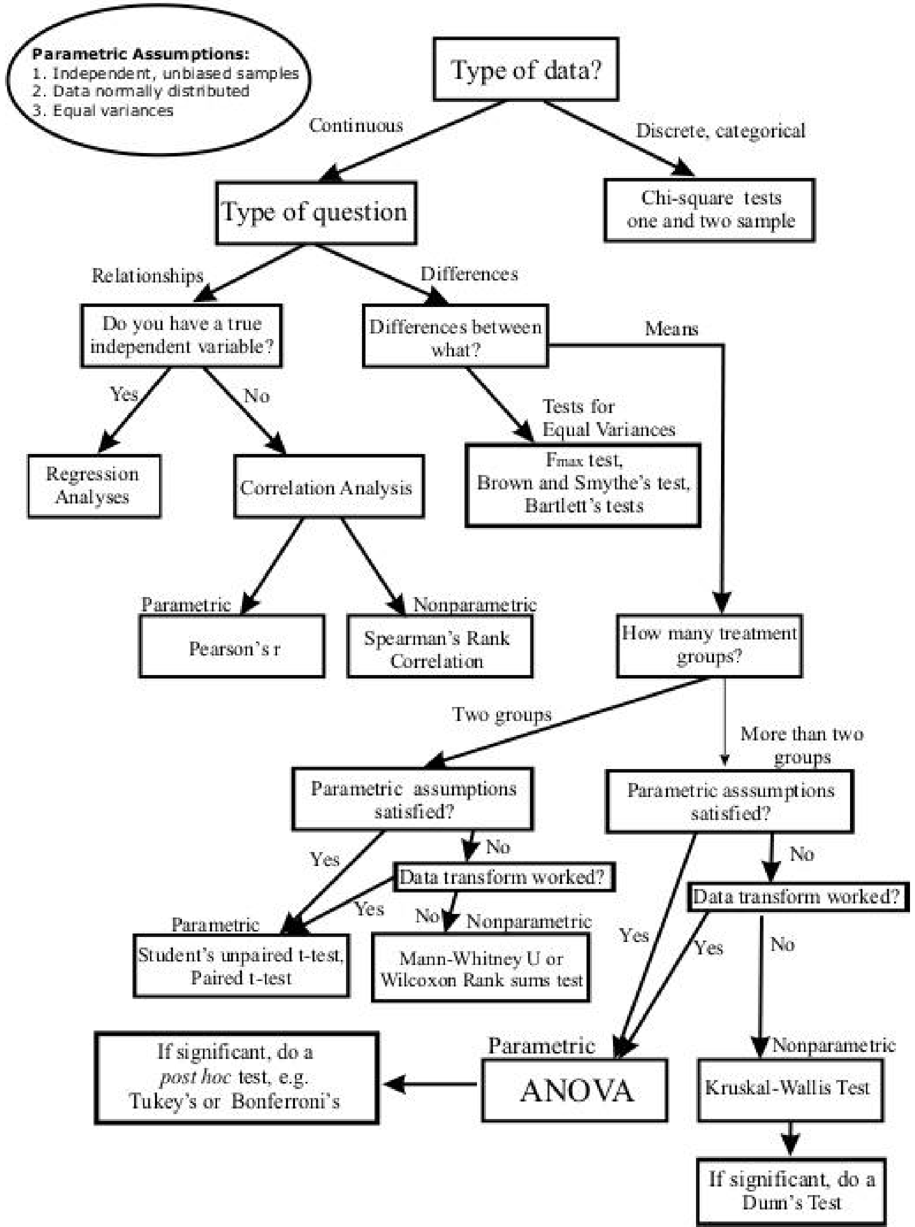
T-tests

Correlation and regression

Categorical data

ANOVA

We talk a lot about assumptions in statistics generally…



Parametric Assumptions:
1. Independent, unbiased samples
2. Data normally distributed
3. Equal variances

Type of data?

Continuous

Discrete, categorical

Chi-square tests one and two sample

Type of question

Relationships

Differences

Do you have a true independent variable?

Yes

No

Regression Analyses

Correlation Analysis

Parametric

Pearson's r

Nonparametric

Spearman's Rank Correlation

Differences between what?

Means

Tests for Equal Variances

F_{max} test, Brown and Smythe's test, Bartlett's tests

How many treatment groups?

Two groups

Parametric assumptions satisfied?

Yes

No

Data transform worked?

Yes

No

Parametric

Student's unpaired t-test, Paired t-test

Nonparametric

Mann-Whitney U or Wilcoxon Rank sums test

More than two groups

Parametric asssumptions satisfied?

Yes

No

Data transform worked?

Yes

No

Parametric

ANOVA

If significant, do a *post hoc* test, e.g. Tukey's or Bonferroni's

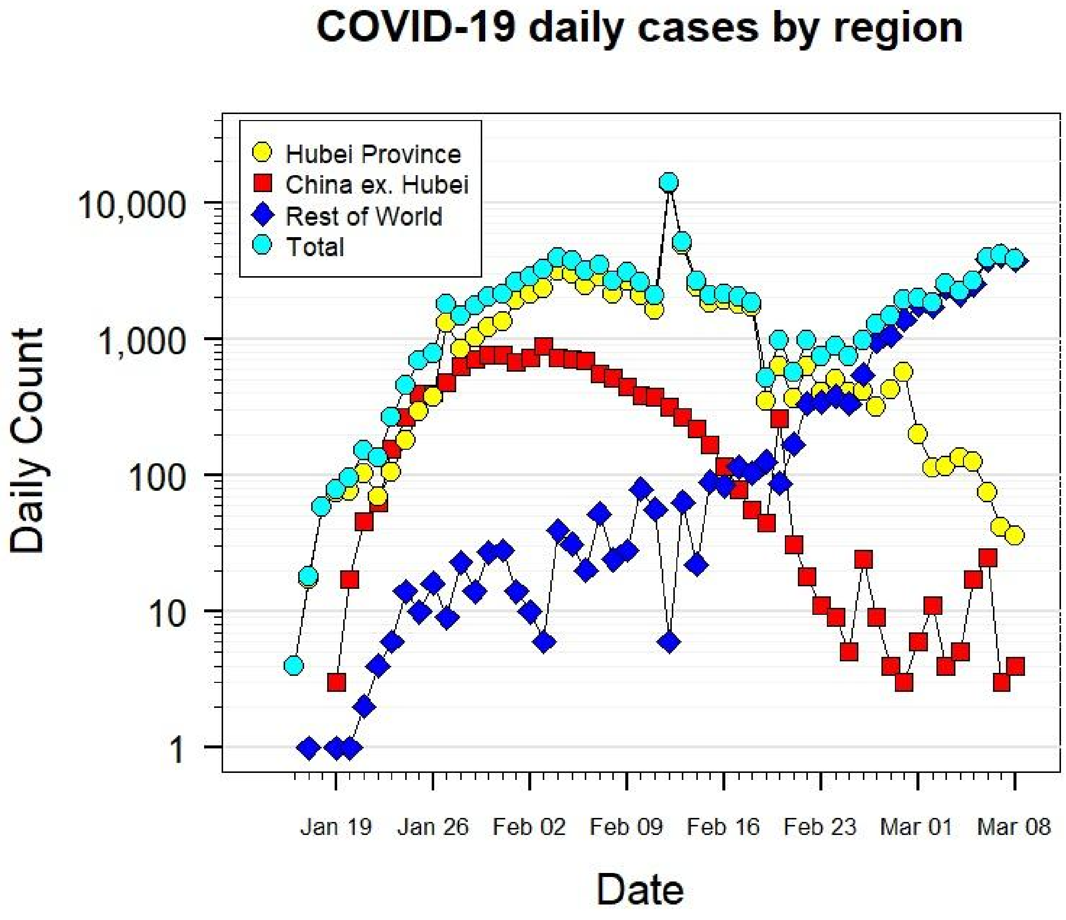
Nonparametric

Kruskal-Wallis Test

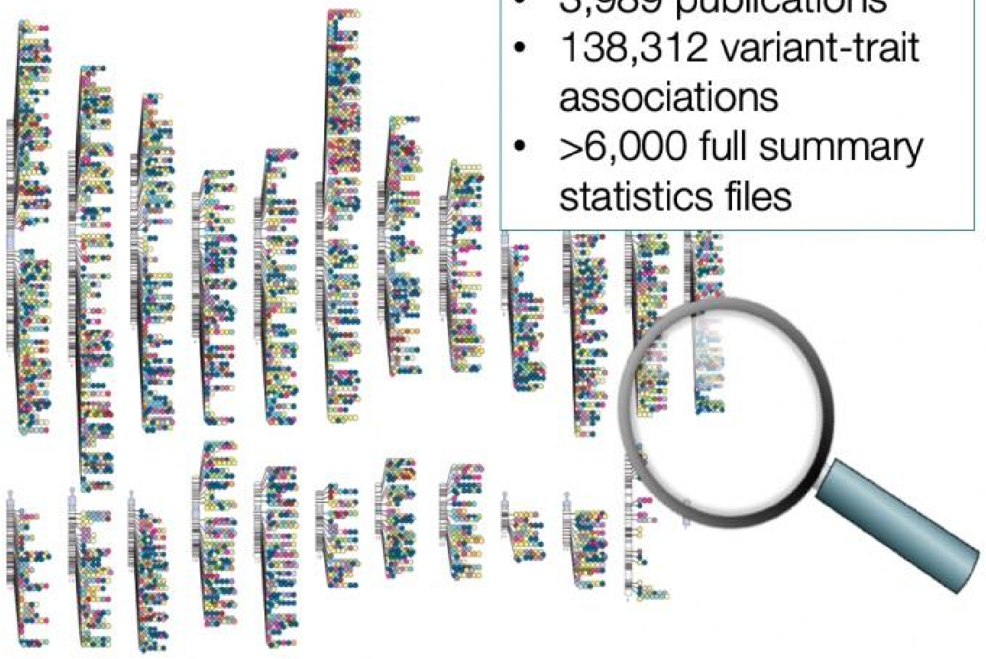
If significant, do a Dunn's Test

…but what if your data doesn't fit any of these?!

Many biomedical datasets have a complex distribution



GWAS Catalog



As of May 2019

- 3,989 publications
- 138,312 variant-trait associations
- >6,000 full summary statistics files

Learning objectives

After this lecture, you should be able to:

- Explain the concept of bootstrapping.
- Recognise situations where bootstrapping is useful.

Lecture outline

1. R function `sample`
2. Bootstrapping for hypothesis testing
3. Bootstrapping to generate confidence intervals
4. Reflection on bootstrapping

Bonus content: No equations, only a function

- Sampling from a set: `sample()`

  e.g. draw 10 samples from your dataset without replacement:

  ```
  sample(dataset, 10, replace=FALSE)
  ```

- Can you explore this function yourself in R?
- What do you think might be the most useful parameters?

Drawing random numbers in R

- Sampling from a set: `sample()`, e.g. draw 10 samples from your dataset with replacement: `sample(dataset, 10, replace=FALSE)`

If your dataset contains 50 data points, what is the maximal sample size for sampling without replacement? How about with replacement?

Drawing random numbers in R

- Sampling from a set: `sample()`, e.g. draw 10 samples from your dataset with replacement: `sample(dataset, 10, replace=FALSE)`

If your datasets contains 50 data points, what is the maximal sample size for sampling without replacement? How about with replacement?

- Normally distributed random numbers: `rnorm()`, e.g. draw 100 numbers from a normal distribution with mean 4 and standard deviation 2: `rnorm(100,4,2)`

What will the histogram look like?

Drawing random numbers in R

- Sampling from a set: `sample()`, e.g. draw 10 samples from your dataset with replacement: `sample(dataset, 10, replace=FALSE)`

If your datasets contains 50 data points, what is the maximal sample size for sampling without replacement? How about with replacement?

- Normally distributed random numbers: `rnorm()`, e.g. draw 100 numbers from a normal distribution with mean 4 and standard deviation 2: `rnorm(100,4,2)`

What will the histogram look like?

- Uniformly distributed random numbers: `runif()`, e.g. draw 100 numbers between 0 and 10: `runif(100,0,10)`

What will the histogram look like?

Lecture outline

1. R function `sample`
2. Bootstrapping for hypothesis testing
3. Bootstrapping to generate confidence intervals

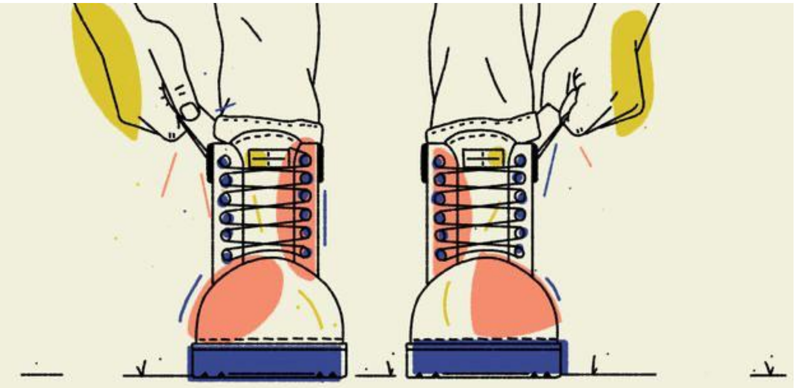
Bootstrapping

Problem:

- We do not know or fully understand the population distribution.

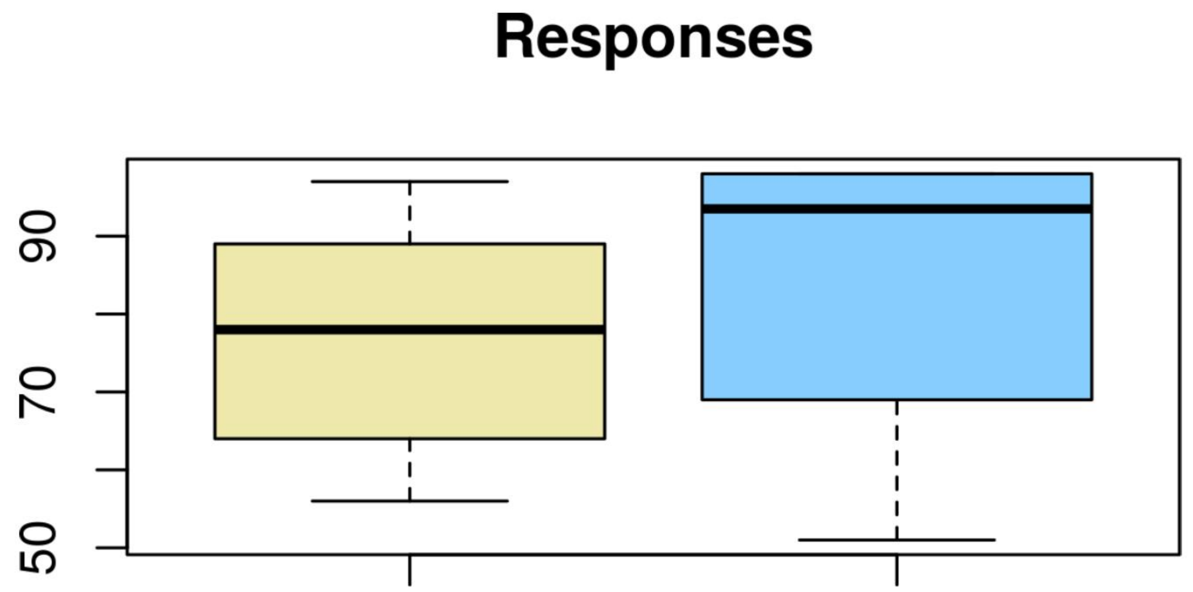
However:

We have *some* information about the underlying distribution: the data! We use the data itself as the basis for our randomisation. This is called **bootstrapping**. We create a **bootstrap sample** by sampling (with replacement) from the data, and repeat this procedure many times.



Example: How do groups respond to a drug?

A lab tests response to a cancer drug in two cell lines with different genetic backgrounds. Response is measured as survival rate (in percent) per sample. 10 samples were tested for each condition. Researchers would like to know whether the median response rate differs between both groups. What test would you use?



Example: How do groups respond to a drug?

How would you bootstrap this?

Example: How do groups respond to a drug?

How would you bootstrap this?

- If H_0 is true, then genetic background does not really matter. So, under $H_{0,}$ a given observation (from our sample) could equally well come from group 1 or from group 2.
- Pool data from group 1 and group 2.

Example: How do groups respond to a drug?

How would you bootstrap this?

- If H_0 is true, then genetic background does not really matter. So, under $H_{0,}$ a given observation (from our sample) could equally well come from group 1 or from group 2.
- Pool data from group 1 and group 2.
- For each experiment, do the following:
  - Sample (with replacement) from pool to get bootstrap sample 1.
  - Sample (with replacement) from pool to get bootstrap sample 2.
  - Compute difference in median between sample 1 and sample 2.

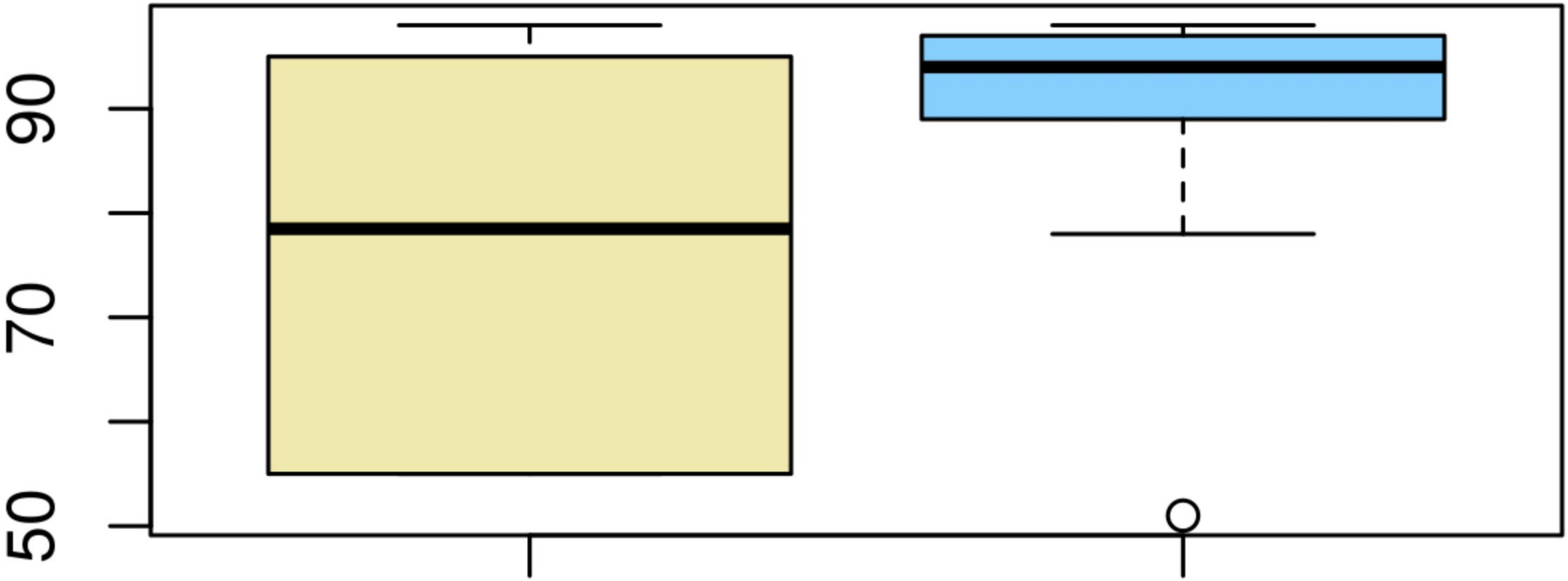
Example: How do groups respond to a drug?

How would you bootstrap this?

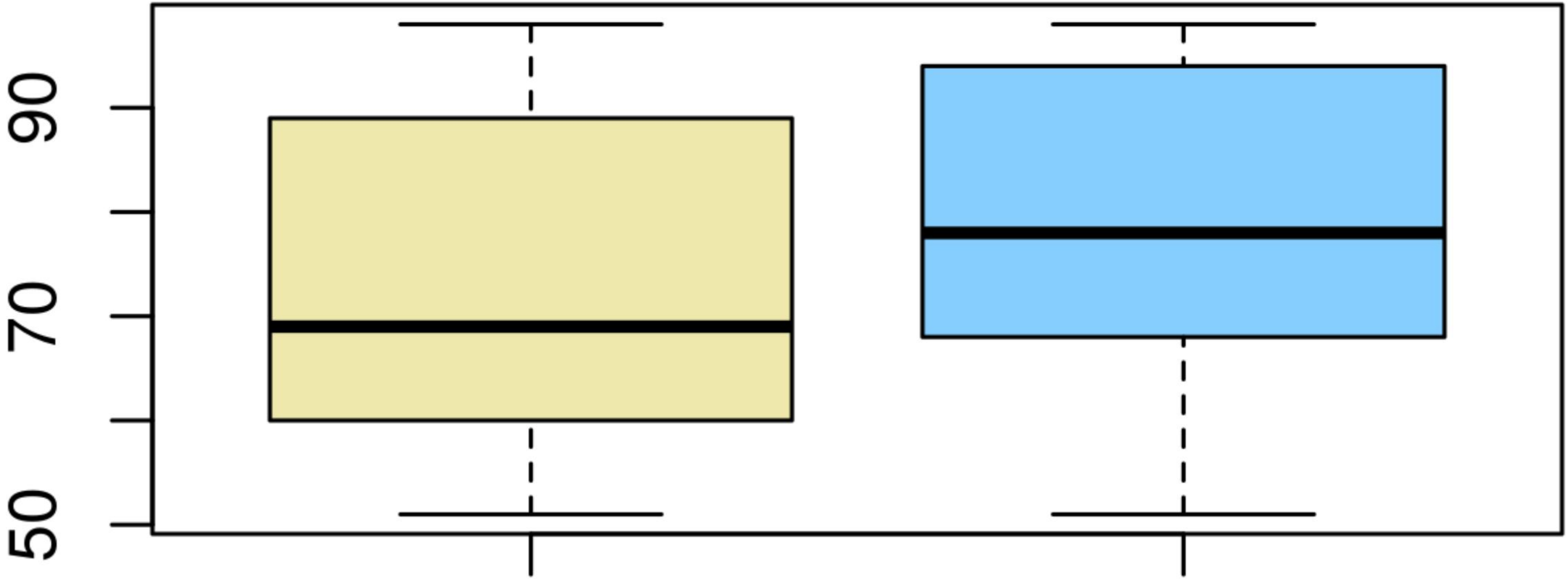
- If H_0 is true, then genetic background does not really matter. So, under $H_{0,}$ a given observation (from our sample) could equally well come from group 1 or from group 2.
- Pool data from group 1 and group 2.
- For each experiment, do the following:
  - Sample (with replacement) from pool to get bootstrap sample 1.
  - Sample (with replacement) from pool to get bootstrap sample 2.
  - Compute difference in median between sample 1 and sample 2.
- Do this many times. How often is the difference in median greater than or equal to that observed?

Why do we want to know this?

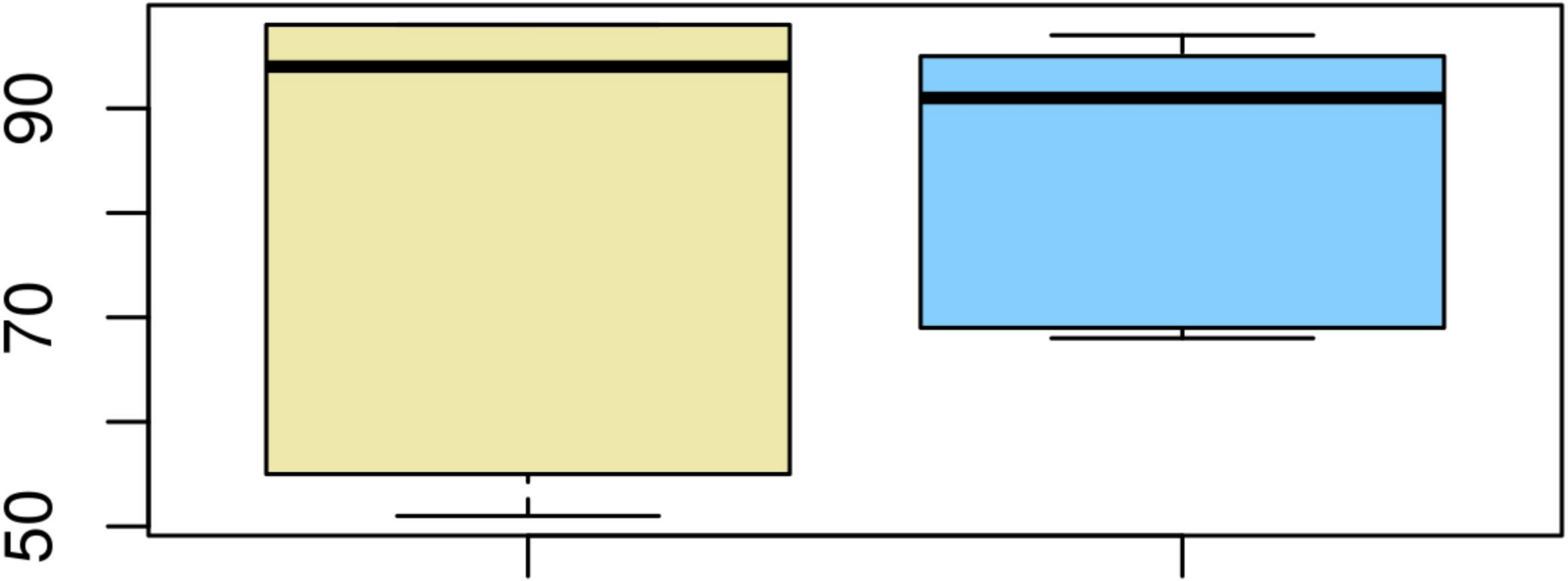
Responses



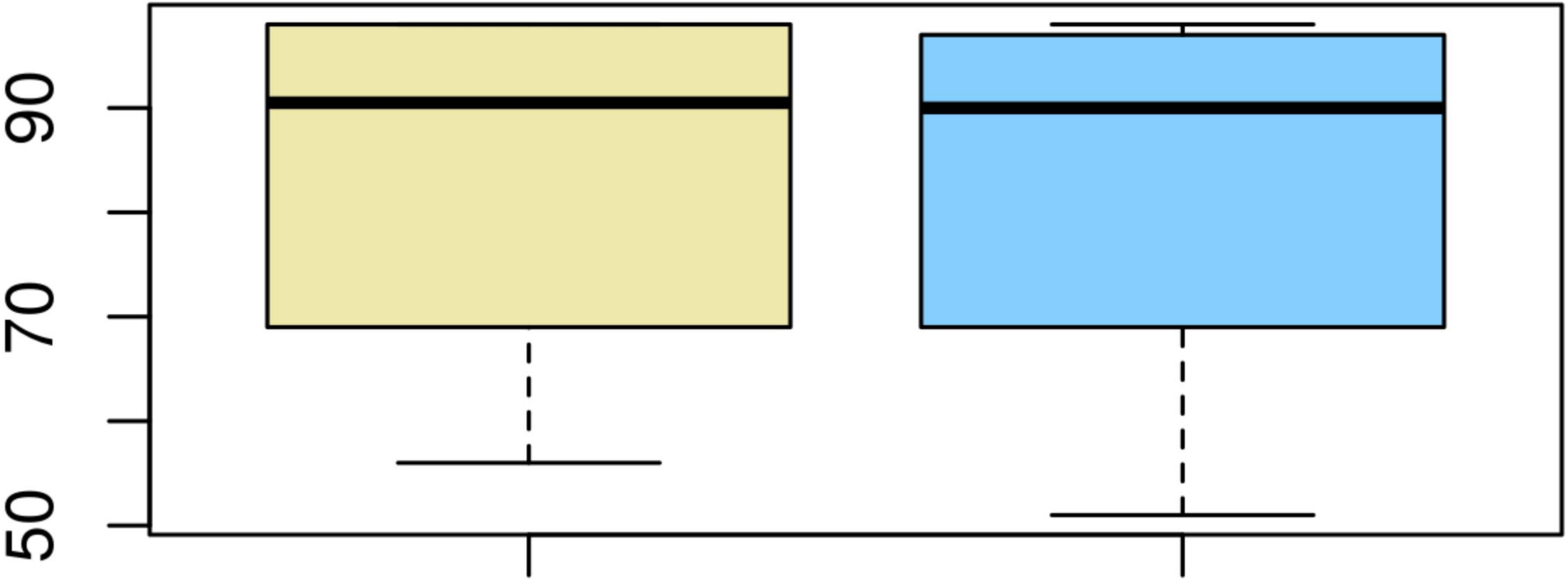
Responses



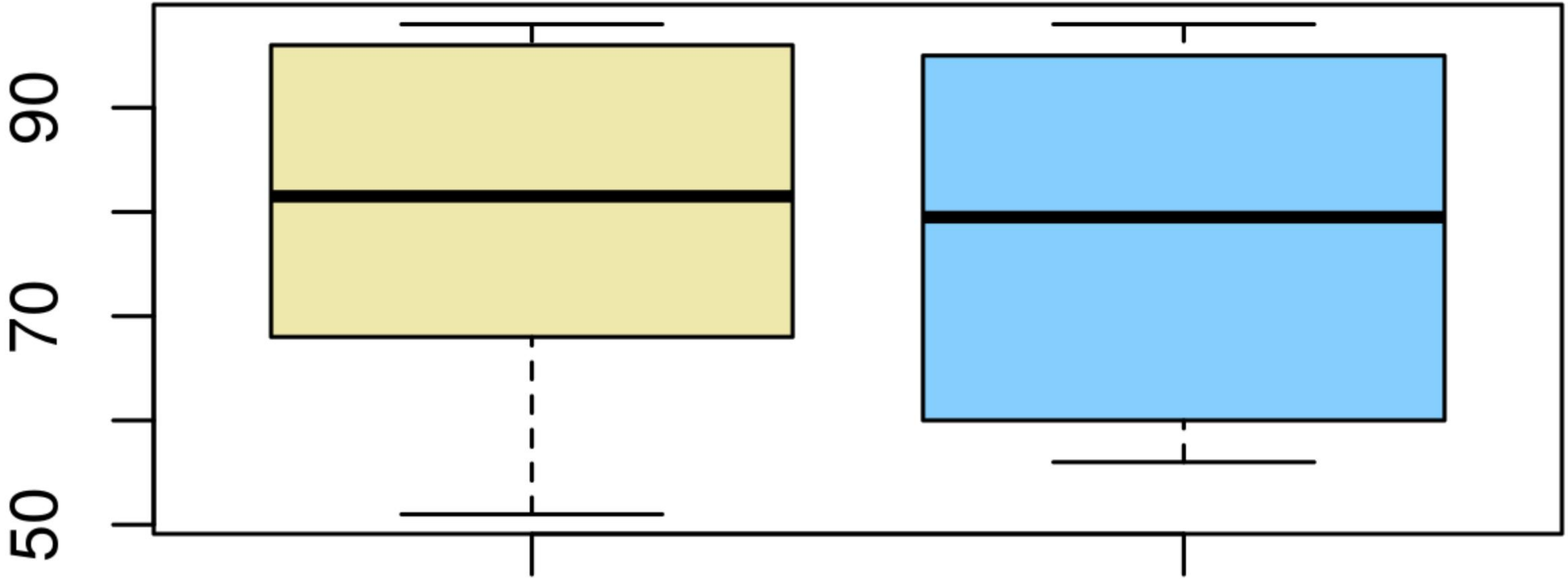
Responses



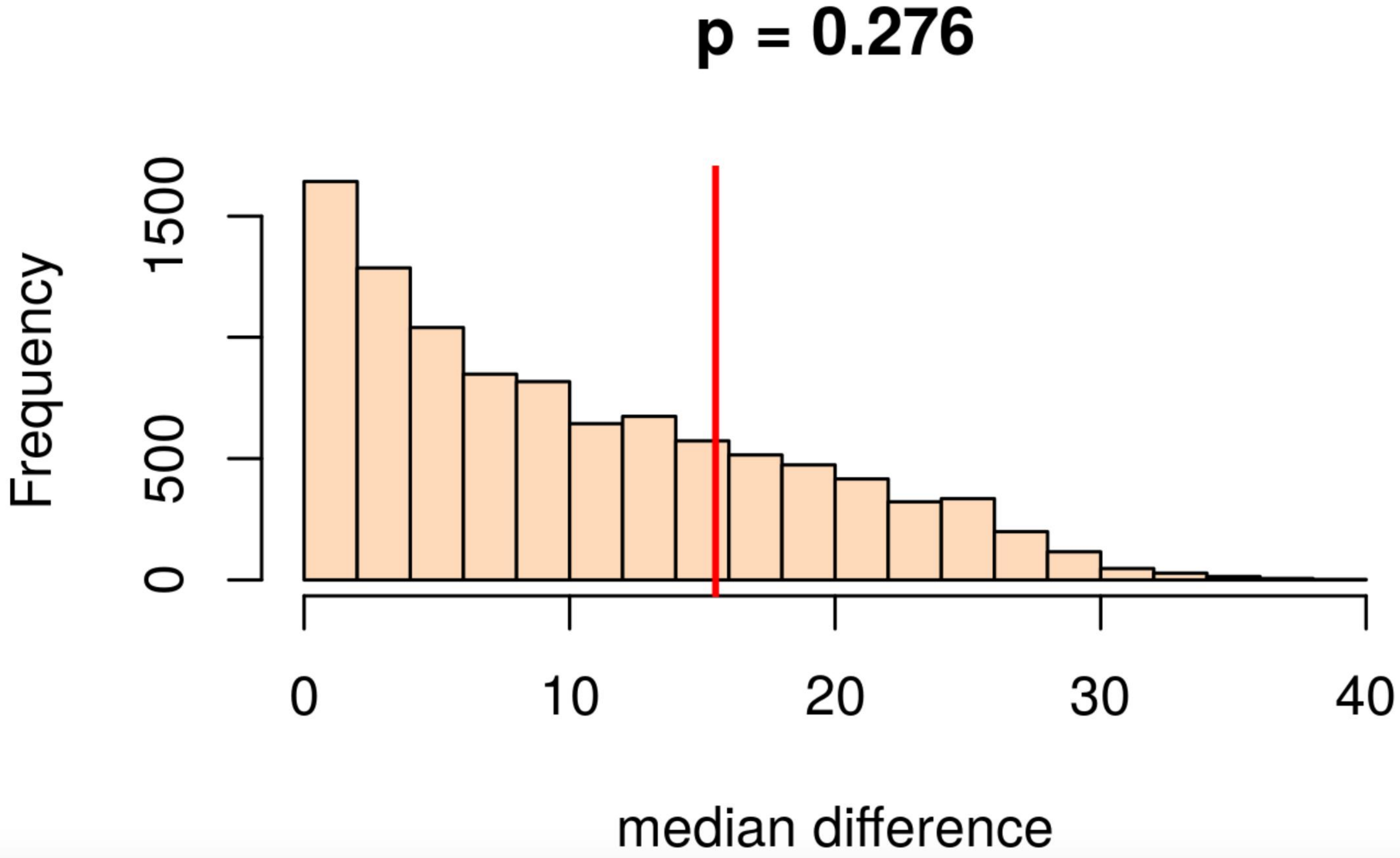
Responses



Responses



Results



Lecture outline

1. R function `sample`

2. Bootstrapping for hypothesis testing

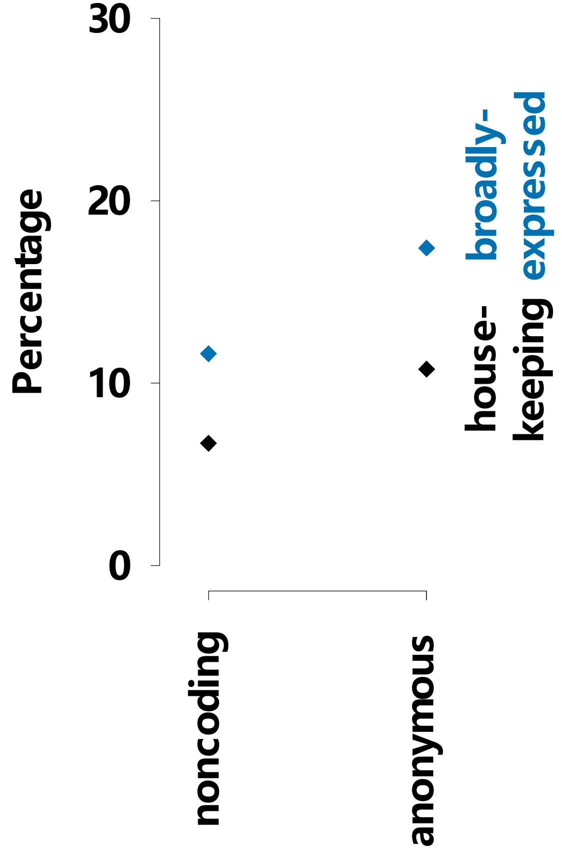
3. Bootstrapping to generate confidence intervals

4. Reflection on bootstrapping

The frequent evolutionary birth and death of functional promoters in mouse and human

Genome Research

Robert S. Young,[1] Yoshihide Hayashizaki,[2] Robin Andersson,[3] Albin Sandelin,[3] Hideya Kawaji,[2,4] Masayoshi Itoh,[2,4] Timo Lassmann,[4] Piero Carninci,[4] The FANTOM Consortium, Wendy A. Bickmore,[1] Alistair R. Forrest,[4,5] and Martin S. Taylor[1]

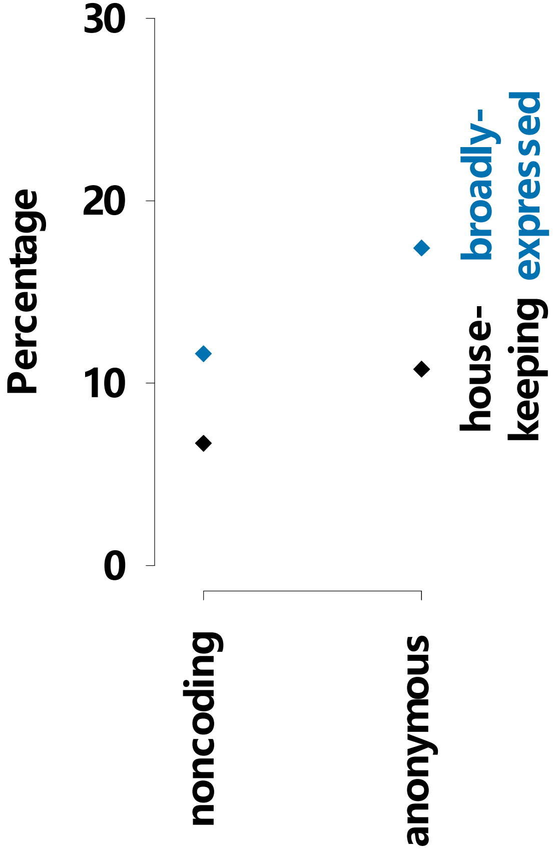


- Do the percentage of different types of noncoding and anonymous promoters differ?

The frequent evolutionary birth and death of functional promoters in mouse and human

GENOME RESEARCH

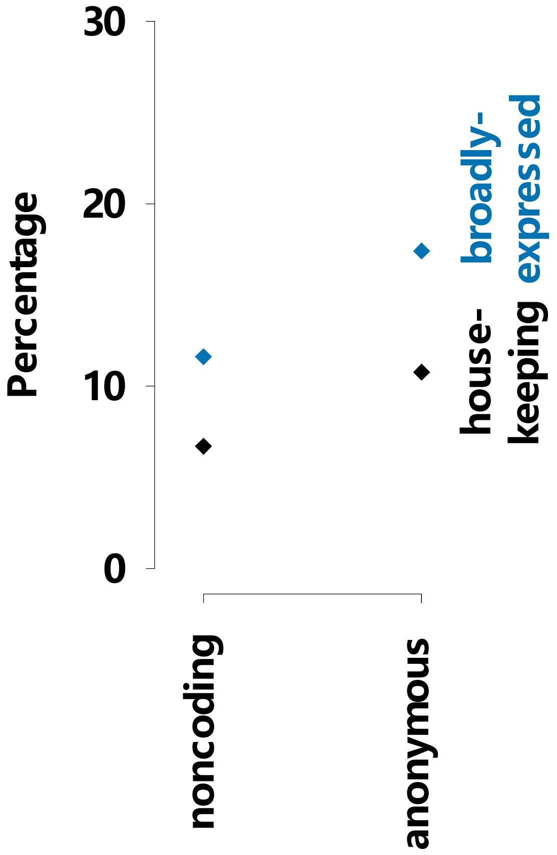
Robert S. Young,[1] Yoshihide Hayashizaki,[2] Robin Andersson,[3] Albin Sandelin,[3] Hideya Kawaji,[2,4] Masayoshi Itoh,[2,4] Timo Lassmann,[4] Piero Carninci,[4] The FANTOM Consortium, Wendy A. Bickmore,[1] Alistair R. Forrest,[4,5] and Martin S. Taylor[1]



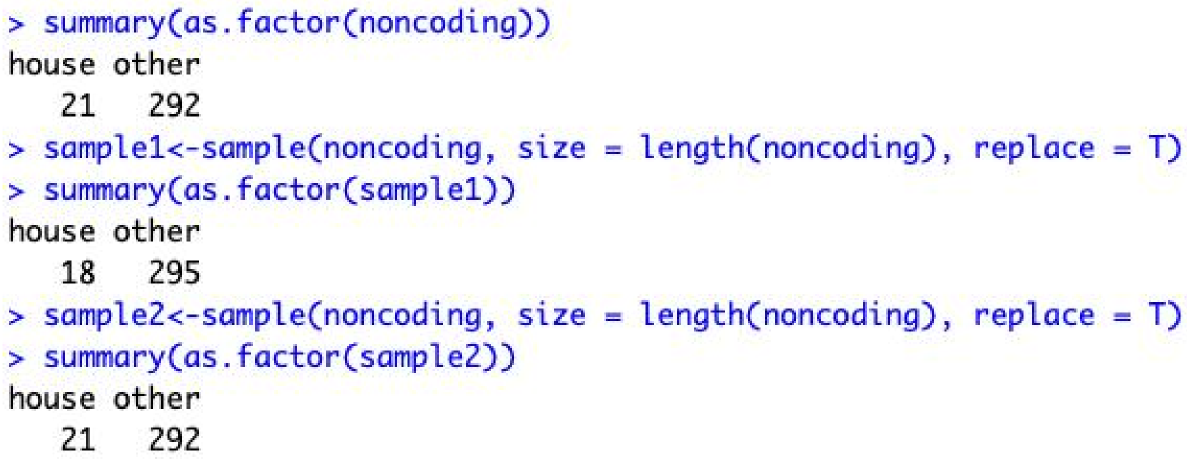
- 21/313 noncoding promoters are housekeeping (6.7%).

The frequent evolutionary birth and death of functional promoters in mouse and human

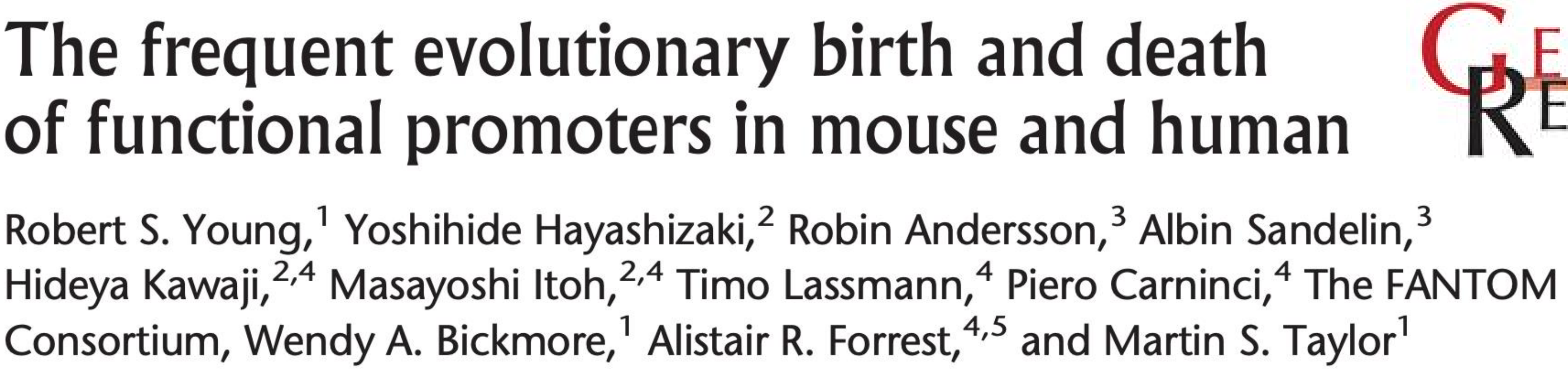
Robert S. Young,[1] Yoshihide Hayashizaki,[2] Robin Andersson,[3] Albin Sandelin,[3] Hideya Kawaji,[2,4] Masayoshi Itoh,[2,4] Timo Lassmann,[4] Piero Carninci,[4] The FANTOM Consortium, Wendy A. Bickmore,[1] Alistair R. Forrest,[4,5] and Martin S. Taylor[1]



- 21/313 noncoding promoters are housekeeping (6.7%).

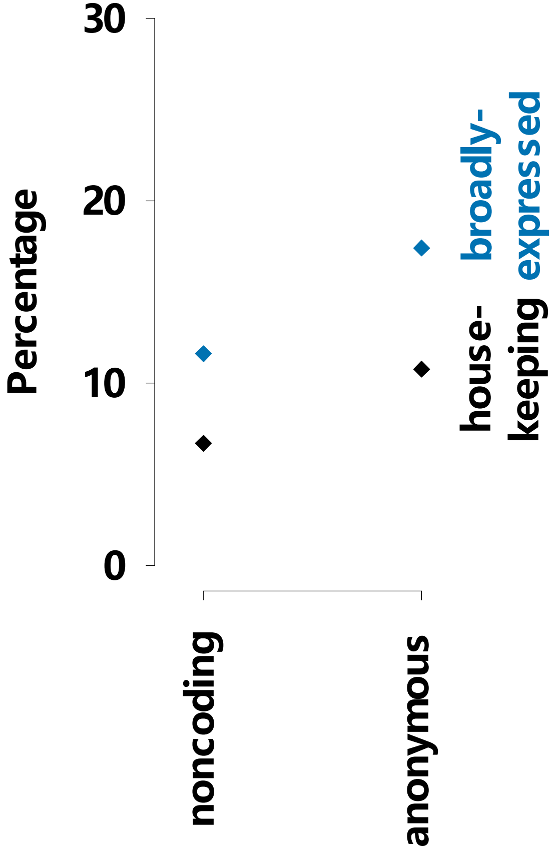


```
> summary(as.factor(noncoding))
house other
   21   292
> sample1<-sample(noncoding, size = length(noncoding), replace = T)
> summary(as.factor(sample1))
house other
   18   295
> sample2<-sample(noncoding, size = length(noncoding), replace = T)
> summary(as.factor(sample2))
house other
   21   292
```

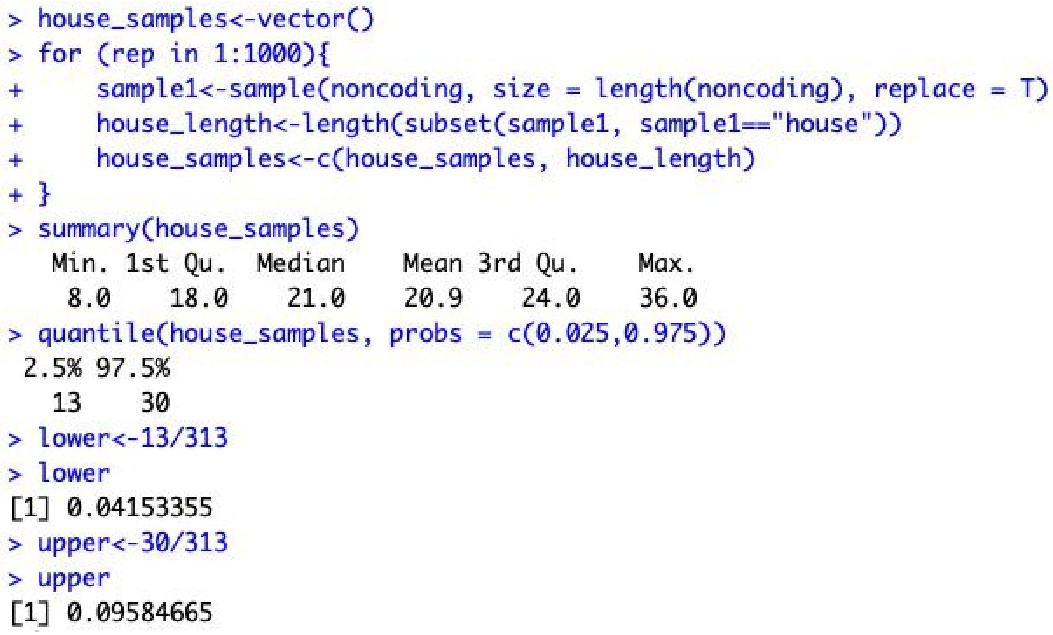


The frequent evolutionary birth and death of functional promoters in mouse and human

Robert S. Young,[1] Yoshihide Hayashizaki,[2] Robin Andersson,[3] Albin Sandelin,[3] Hideya Kawaji,[2,4] Masayoshi Itoh,[2,4] Timo Lassmann,[4] Piero Carninci,[4] The FANTOM Consortium, Wendy A. Bickmore,[1] Alistair R. Forrest,[4,5] and Martin S. Taylor[1]



- 21/313 noncoding promoters are housekeeping (6.7%).

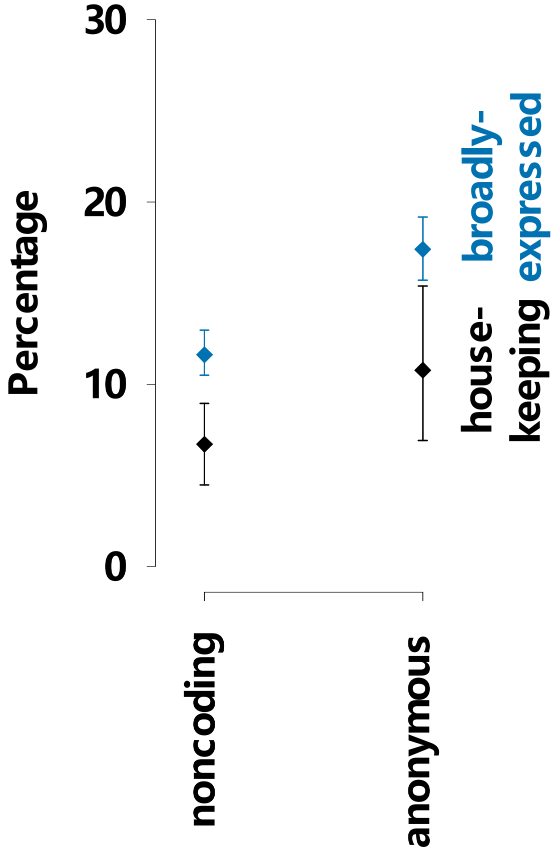


```
> house_samples<-vector()
> for (rep in 1:1000){
+     sample1<-sample(noncoding, size = length(noncoding), replace = T)
+     house_length<-length(subset(sample1, sample1=="house"))
+     house_samples<-c(house_samples, house_length)
+ }
> summary(house_samples)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    8.0    18.0    21.0    20.9    24.0    36.0 
> quantile(house_samples, probs = c(0.025,0.975))
 2.5% 97.5% 
   13    30 
> lower<-13/313
> lower
[1] 0.04153355
> upper<-30/313
> upper
[1] 0.09584665
```

The frequent evolutionary birth and death of functional promoters in mouse and human

Genome Research

Robert S. Young,[1] Yoshihide Hayashizaki,[2] Robin Andersson,[3] Albin Sandelin,[3] Hideya Kawaji,[2,4] Masayoshi Itoh,[2,4] Timo Lassmann,[4] Piero Carninci,[4] The FANTOM Consortium, Wendy A. Bickmore,[1] Alistair R. Forrest,[4,5] and Martin S. Taylor[1]



- Do the percentage of different types of noncoding and anonymous promoters differ?
- Error bars represent the 95% confidence interval of 1,000 samplings → why with replacement?
- Can we reject the H_0?

Lecture outline

1. R function `sample`
2. Bootstrapping for hypothesis testing
3. Bootstrapping to generate confidence intervals
4. Reflection on bootstrapping

Does bootstrapping make you uncomfortable?

Does bootstrapping make you uncomfortable?

- The data we have is incomplete (*but it is all we have!)*
- Fundamentally not an exact method *(but does it matter?)*

Learning objectives

After this lecture, you should be able to:

- Explain the concept of bootstrapping.
- Recognise situations where bootstrapping is useful.

浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

Bootstrapping

Any questions?

ADS2, Lecture 21

Dr Rob Young – robert.young@ed.ac.uk

Semester 2, 2023/24