STAT540
Lecture 22b: March 30th 2015

Resampling: the bootstrap & permutation testing

Sara Mostafavi

Department of Statistics

Department of Medical Genetics

Center for Molecular Medicine and Therapeutics

(Thanks to Jenny Bryan & Dolph Schulter for slides)

Resampling methods

- Ways of performing statistical inference that are “internal to the data” under analysis: e.g., you get the necessary knowledge about sampling variability (of parameters/ estimates) from the observed data itself

- Resampling methods:
  - Bootstrap
  - Permutation testing

Parameter estimation and hypothesis testing

In statistical data analysis, we often use of the two following types of statistical inference. Each uses a different type of sampling distribution.

1. **(Parameter) estimation:** quantifying confidence in parameter estimates —sampling distribution of the estimate. Make parametric assumptions about model parameters, or use a computer intensive method for estimate the sampling distribution (Bootstrap).

2. **Hypothesis testing:** The probability distribution of the test-statistics if the null hypothesis is true, we need to estimate the null distribution. Parametric assumption, or computer intensive method (permutation testing).

Sampling distribution

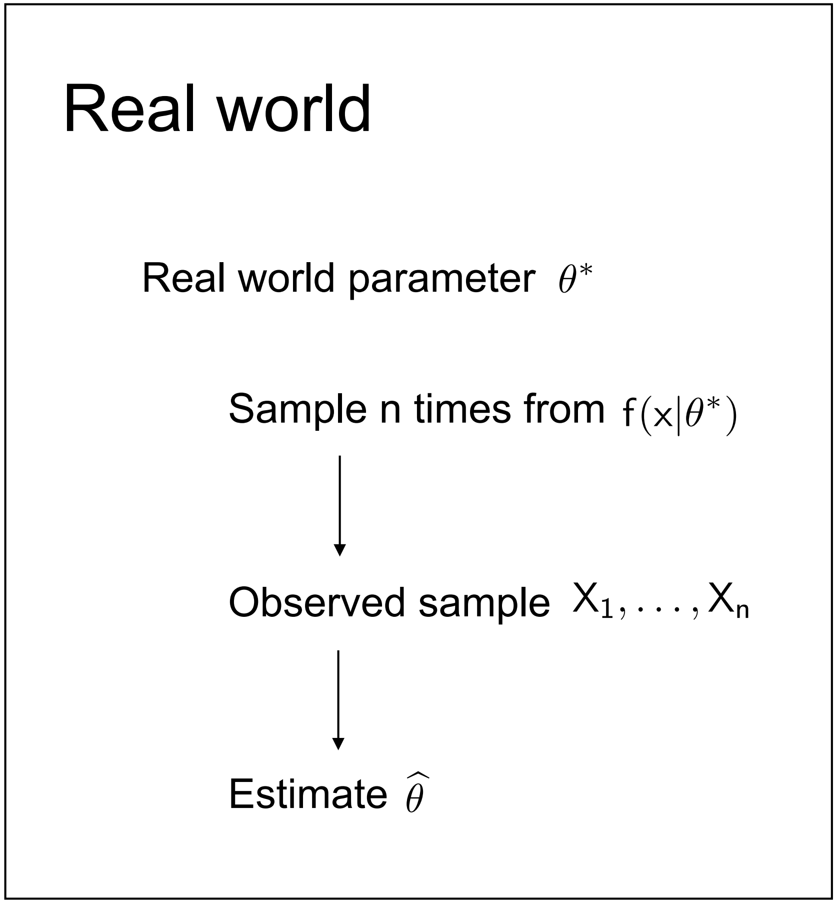
- We are given a sample (i.e., some data) $X_1,\ldots,X_n$ that are independent draws from an underlying data generating function $f(X \mid \theta^*)$

- We often want to known something about *f*, for example we want to know θ^*

- An estimate $\hat{\theta}$ is just some function of $X_1,\ldots,X_n$, for example you can think of it as $\hat{\theta} = \hat{\theta}(X_1,\ldots.,X_n)$

- If we could repeat our "experiment", we could get sampling distribution for $\hat{\theta}$

A sketch of a single experiment and an estimate



Real world

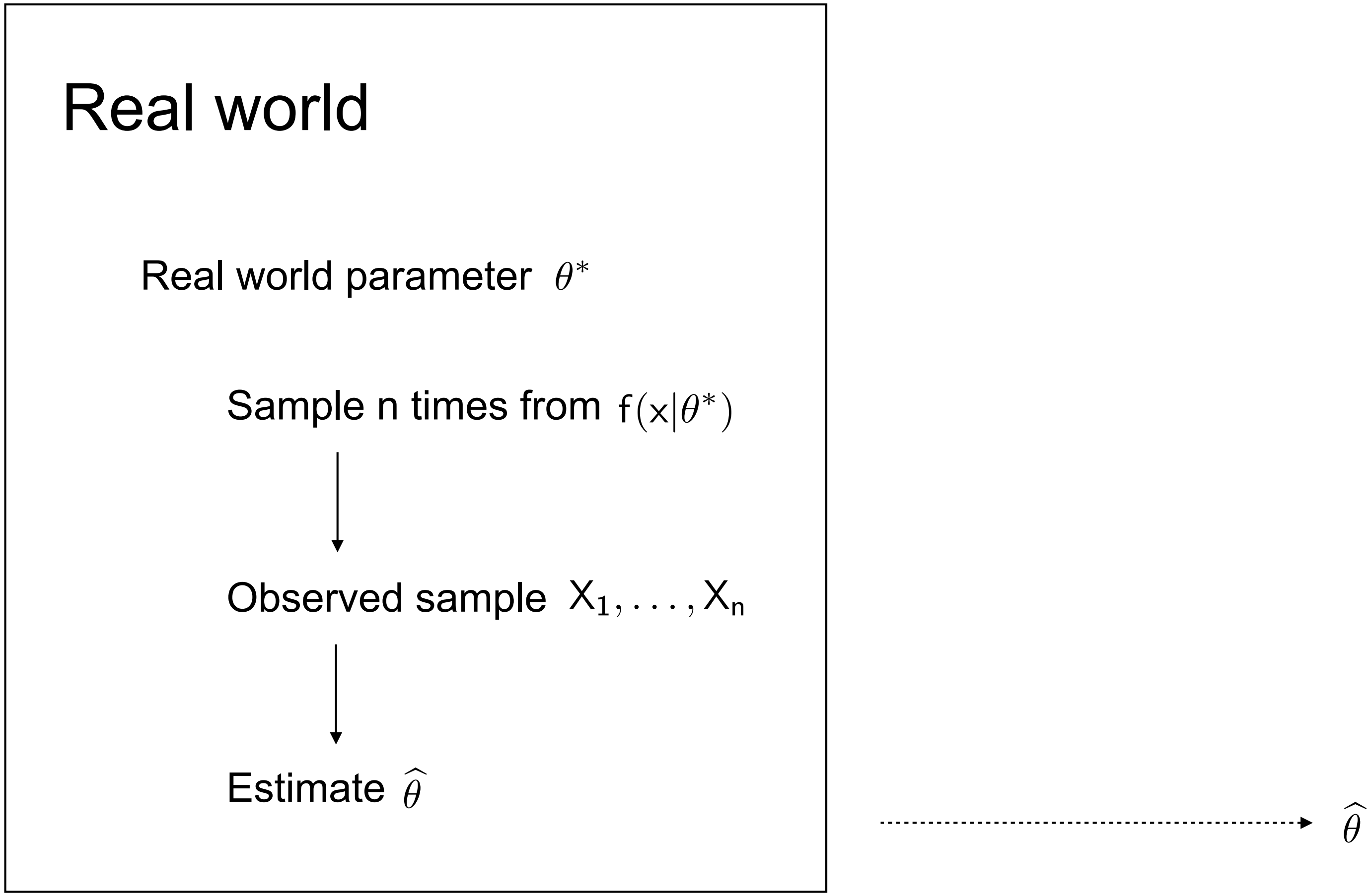
Real world parameter θ^*

Sample n times from $f(x|\theta^*)$

Observed sample $X_1, \ldots, X_n$

Estimate $\widehat{\theta}$

A sketch of a single experiment and an estimate



Real world

Real world parameter θ^*

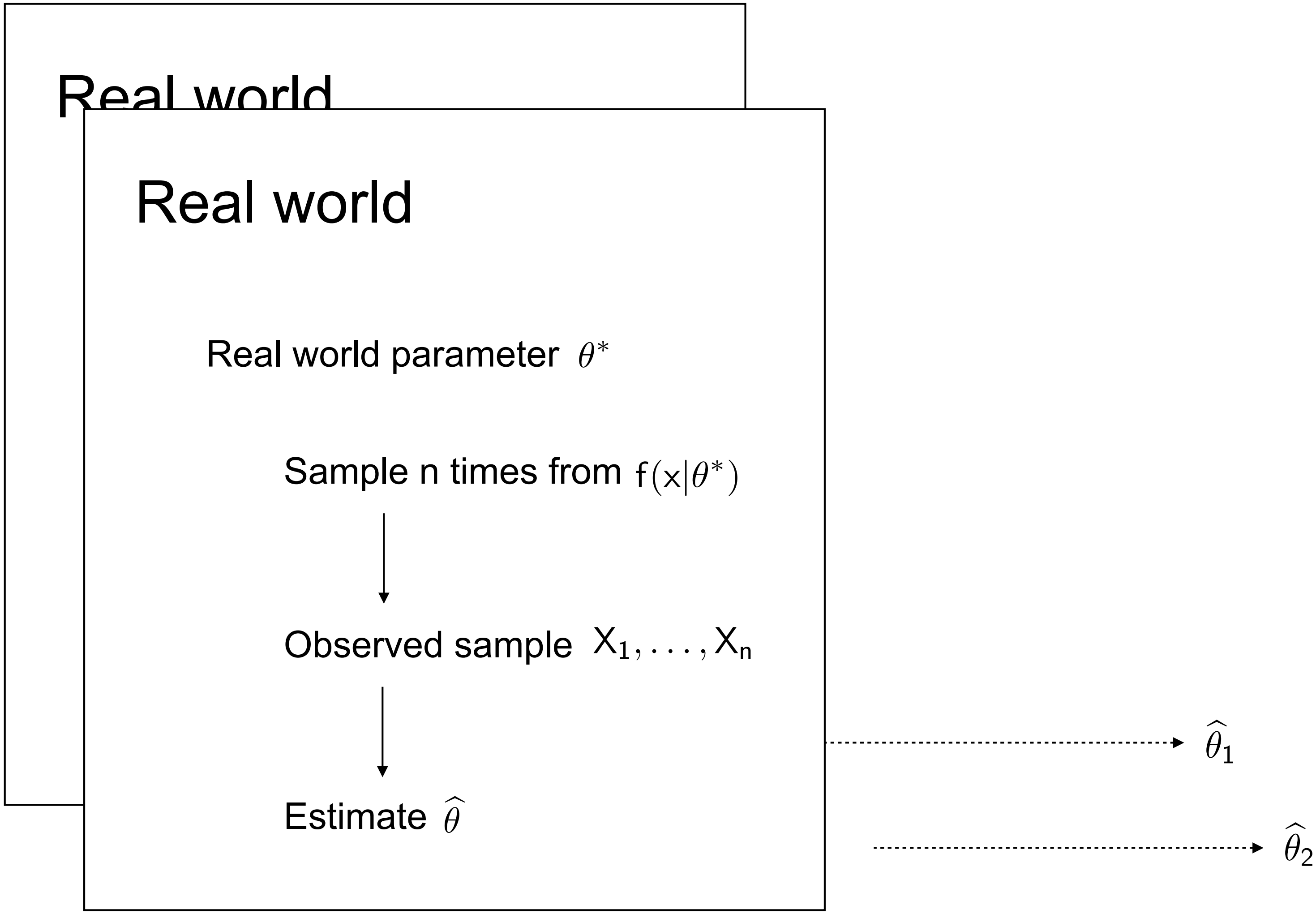
Sample n times from $f(x|\theta^*)$

Observed sample $X_1, \ldots, X_n$

Estimate $\widehat{\theta}$

$\widehat{\theta}$

Repeating the experiment produces a new estimate



Real world

Real world

Real world parameter θ^*

Sample n times from $f(x|\theta^*)$

Observed sample $X_1, \ldots, X_n$

Estimate $\widehat{\theta}$

$\widehat{\theta}_1$

$\widehat{\theta}_2$

Real world

Real world

Real world

Real world

Real world

Real world parameter θ^*

Sample n times from $f(x|\theta^*)$

Observed sample $X_1, \ldots, X_n$

Estimate $\widehat{\theta}$

$\widehat{\theta}_1$

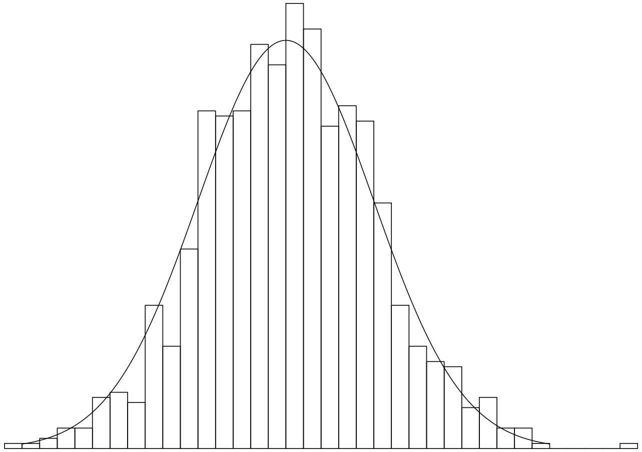
$\widehat{\theta}_2$

$\widehat{\theta}_3$

$\widehat{\theta}_4$

$\widehat{\theta}_5$ ...

Sampling distribution



- The distribution of the estimates computed from repeating the experiment multiple times: sampling distribution

- If we had it, we could assess some properties of our estimate:
  - Standard deviation (“standard error”)
  - Bias of the estimate
  - Confidence intervals
  - Assess whether asymptotic distribution has started to “kick in” at a finite sample size

Sampling distribution

- The sampling distribution, at the moment, is a theoretical construction—it consists of all possible outcomes for experiments that we could have run.

- In practice, we only ran a single experiment to get all our data. But we still want to assess the properties of our estimate. How?
  - Asymptotic theory
  - Bootstrap

The “Plug-in Principle”

- So we want to know something about the data generating distribution F
  - We are interested in parameter $\theta^* = t(F)$
  - Example: expectation $\theta^* = t(F) = E_F(X)$

- We use a random sample from F: $X_1,\ldots,X_n$
  - Get empirical distribution of $\hat{F}$
  - Then, estimate $t(F)$ by $t(\hat{F})$
  - Example: $\hat{\theta} = t(\hat{F}) = \frac{1}{n}\sum_i x_i$

The “Plug-in Principle”

- So we want to know something about the data generating distribution F
  - We are interested in parameter $\theta^* = t(F)$
  - Example: expectation $\theta^* = t(F) = E_F(X)$

- We use a random sample from F: $X_1,\ldots,X_n$
  - Get empirical distribution of $\hat{F}$
  - Then, estimate $t(F)$ by $t(\hat{F})$
  - Example: $\hat{\theta} = t(\hat{F}) = \frac{1}{n}\sum_i x_i$

- The bootstrap is an application of this plug-in principle.

The “Plug-in Principle”

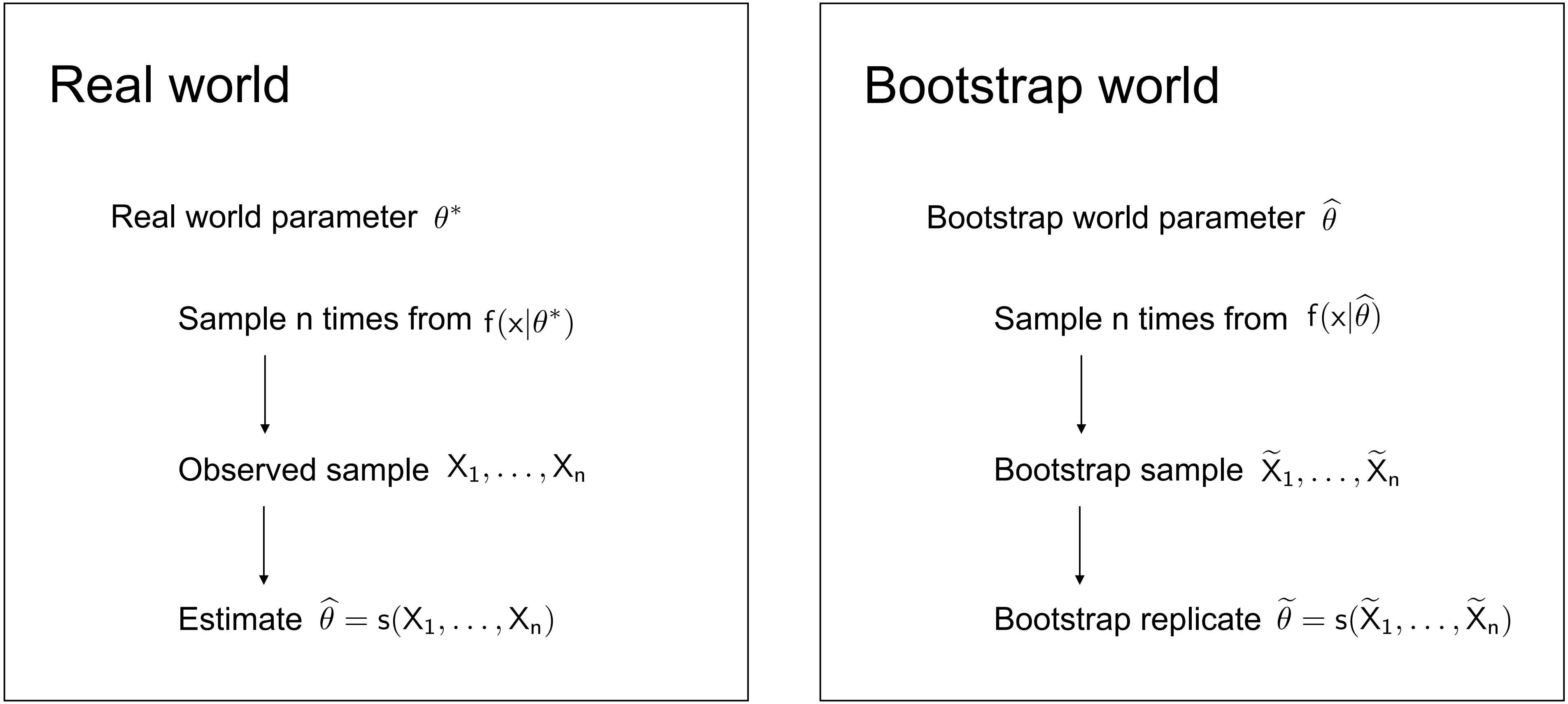
- So we want to know something about the data generating distribution F
  - We are interested in parameter $\theta^{*} = t(F)$
  - Example: expectation $\theta^{*} = t(F) = E_F(X)$

Bootstrap:

Allow us to assess $\hat{\theta}$ varies around θ^{*} by assessing how $\hat{\theta}_b$ varies around $\hat{\theta}$

-
  - Example: $\hat{\theta} = t(\hat{F}) = \frac{1}{n}\sum_i x_i$

- The bootstrap is an application of this plug-in principle.



Real world

Real world parameter θ^*

Sample n times from $f(x|\theta^*)$

Observed sample $X_1, \ldots, X_n$

Estimate $\widehat{\theta} = s(X_1, \ldots, X_n)$

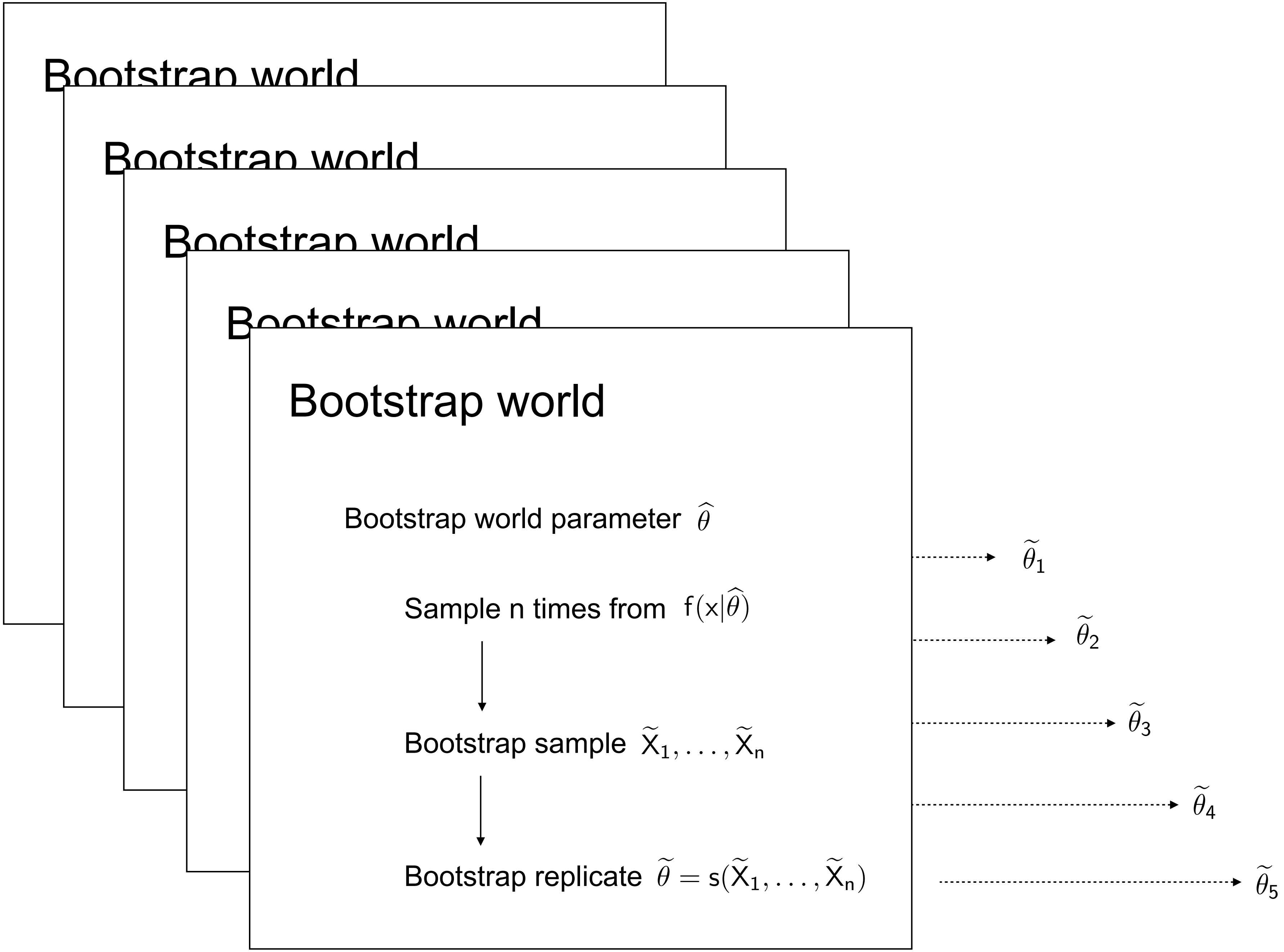
Bootstrap world

Bootstrap world parameter $\widehat{\theta}$

Sample n times from $f(x|\widehat{\theta})$

Bootstrap sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$

Bootstrap replicate $\widetilde{\theta} = s(\widetilde{X}_1, \ldots, \widetilde{X}_n)$



Bootstrap world

Bootstrap world

Bootstrap world

Bootstrap world

Bootstrap world

Bootstrap world parameter $\widehat{\theta}$

Sample n times from $f(x|\widehat{\theta})$

Bootstrap sample $\widetilde{X}_1, \dots, \widetilde{X}_n$

Bootstrap replicate $\widetilde{\theta} = s(\widetilde{X}_1, \dots, \widetilde{X}_n)$

$\widetilde{\theta}_1$

$\widetilde{\theta}_2$

$\widetilde{\theta}_3$

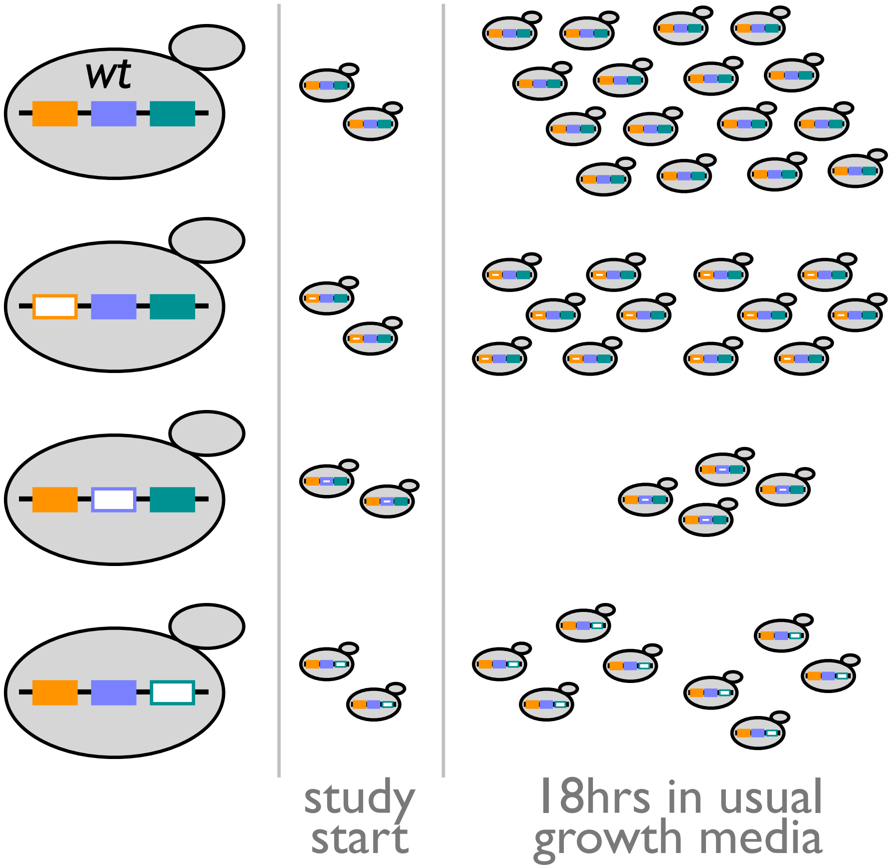
$\widetilde{\theta}_4$

$\widetilde{\theta}_5$

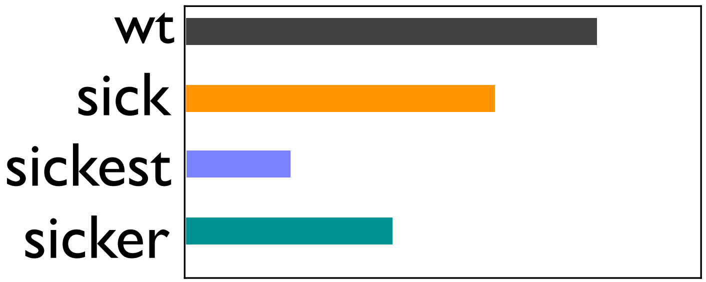
Bootstrap:

Repeat experiment B times (in the bootstrap world) to form b bootstrap replicates of your experiment, then use the B bootstraps to obtain a sampling distribution for your parameter.

Example application of the bootstrap



Study of fitness of yeast deletion mutants



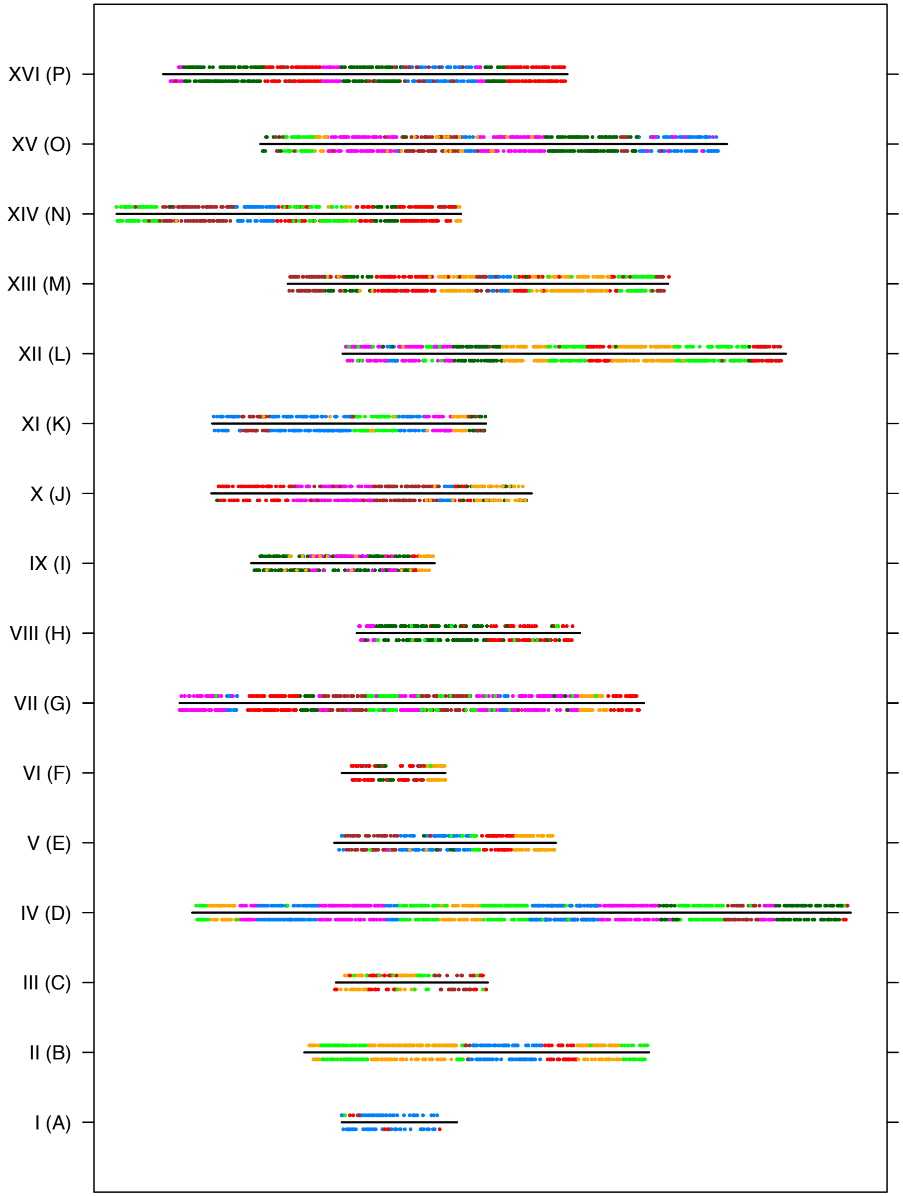
Rationale for growth studies of yeast deletion mutants

- Analogy: flipping circuit breakers in a house to determine which lights and outlets are controlled by each circuit
- If the deletion mutant for gene *g* is defective at some biological activity, that suggests that gene *g* contributes to that activity.
- Growth studies are the 'entry-level' study. In real life, we often measure more complicated phenotypes and subject the mutant to additional challenges, e.g. treatment with drugs or deletion/mutation of additional genes. Also, this type of data is often integrated with from other types of studies.

Yeast genome has 16 chromosomes.

Each gene lives somewhere on one of these chromosomes.

Therefore, each deletion mutant is also associated with one yeast chromosome.



Data for our analysis

response = a quantitative measure of growth

e.g. growth rate or # cells at study end

also know the specific yeast gene that was deleted

e.g. YDL133WY = a yeast ORF

and the chromosome on which the gene is found

e.g. “chromosome 4 / D”

Data for our analysis

response = a quantitative measure of growth

e.g. growth rate or # cells at study end

a

e

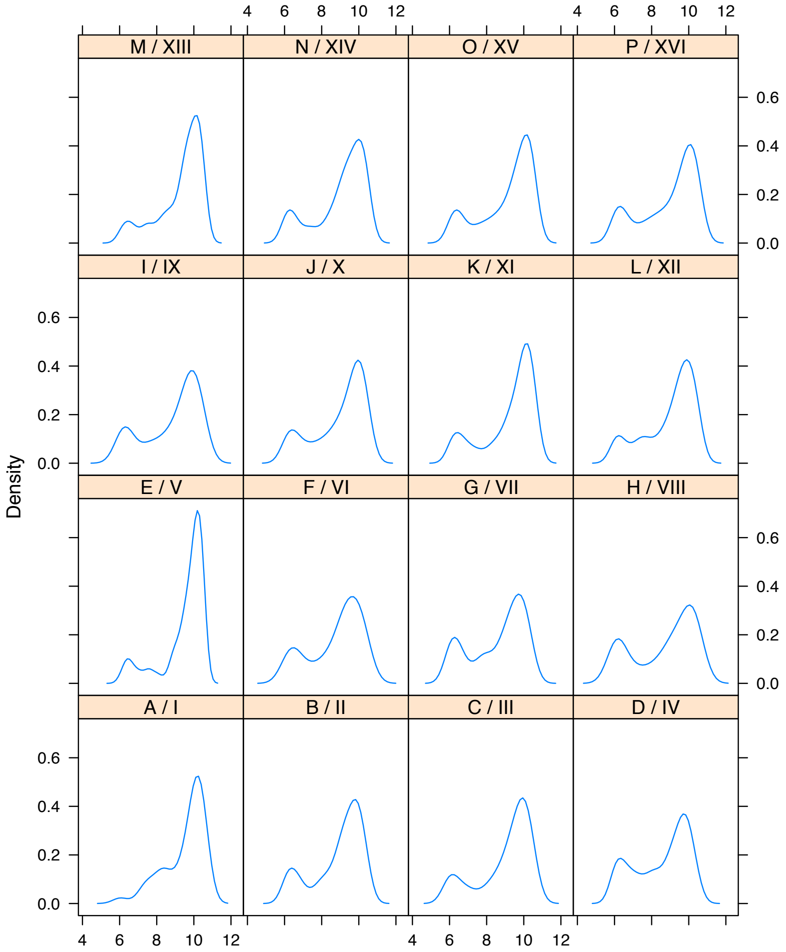
Goal: does the distribution of "growth after deletion" differs for different chromosomes?

and the chromosome on which the gene is found

e.g. "chromosome 4 / D"

quantitative growth phenotypes for gene deletion mutants

each panel = phenotypes for mutants lacking genes on that chromosome



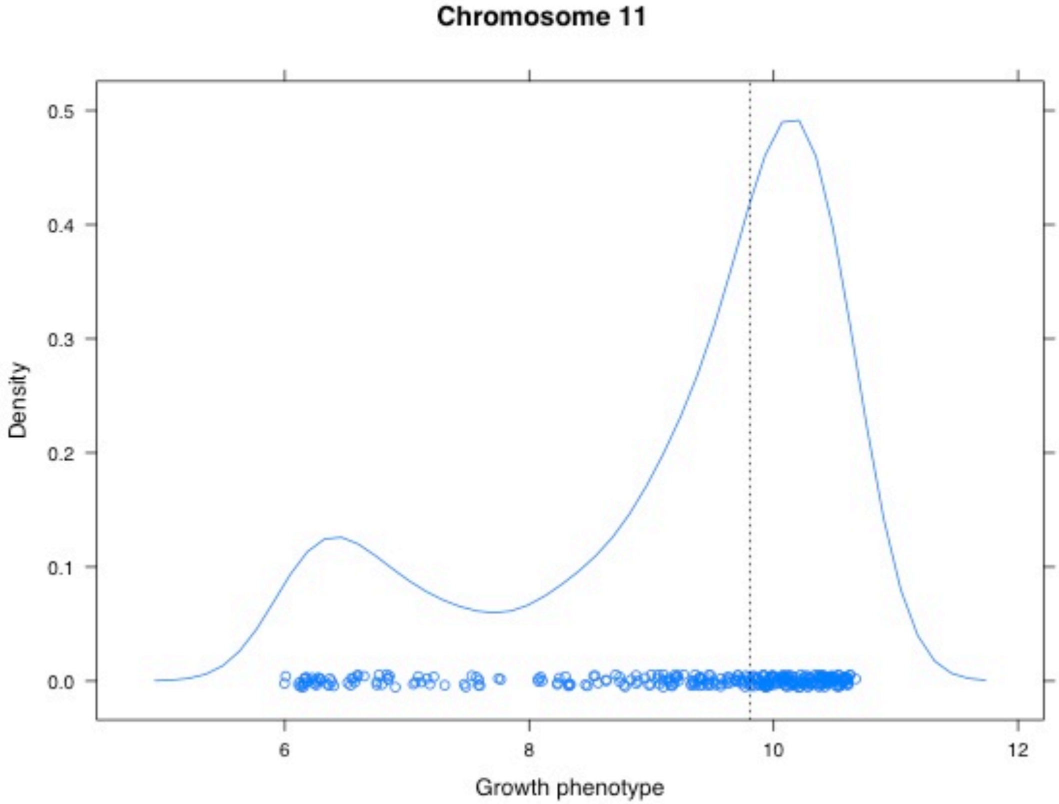
Data source: Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, et al. (2004) Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. Proc Natl Acad Sci U S A 101: 793-798. Pubmed. DOI: 10.1073/pnas.0307490100

Example: Median for chromosome 11

```
> jChromo <- 11
> x <- hDat$pheno[hDat$chromo == jChromo]

> (nx <- length(x))
[1] 302

> (jMedian <- median(x))
[1] 9.804809
```



Example: Median of phenotypes for chromosome 11

- Large-sample theory says the sample median is asymptotically normal with mean = true median = *m* and variance 1 / 4n f(*m*)^2
- Good news = asymptotic dist'n is known
- Bad news = depends on the true density at the median*
- Let's compare this theoretical, asymptotic result to the bootstrap result.

*If I knew the density, I'd know the median, wouldn't I?

Literally:
I’m going to draw tons of new samples from the empirical distribution. These are known as “bootstrap samples”.

I’m going to compute the median for each. You might call these “bootstrap statistics”.

I’m going to show you their empirical distribution -- call that the “bootstrap sampling distribution” -- and superpose the theoretical normal sampling distribution.

We’re going to see how close they are.

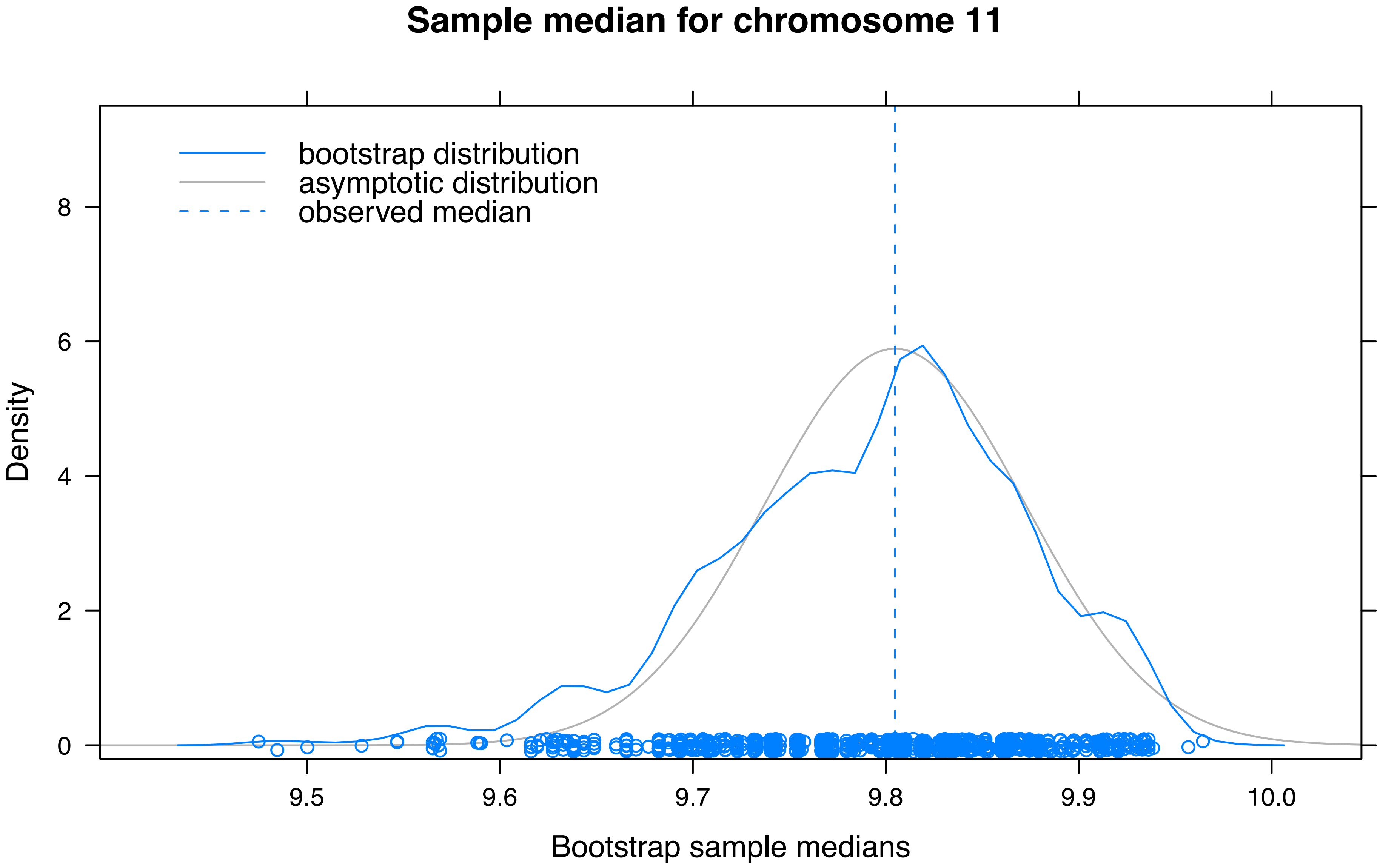
Our bootstrap samples

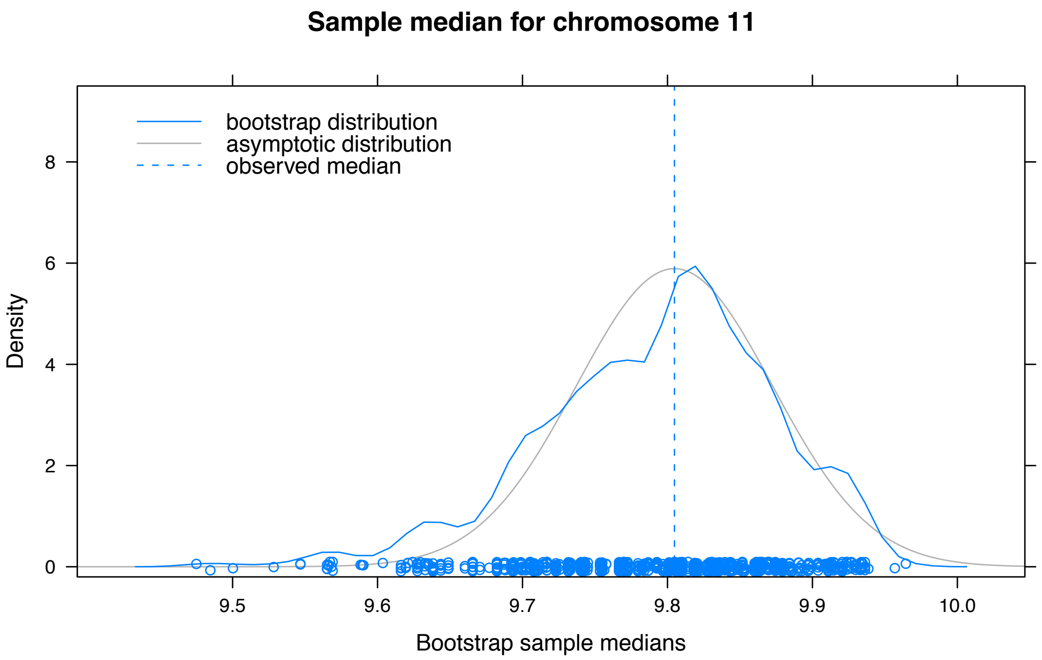
literally I draw a new sample of size n = 302 from the observed data, with replacement

some observations will re-appear ... some once, some twice, etc. .... some don’t show up in the bootstrap sample at all

I take the median of the bootstrap sample. That is a bootstrap statistic.

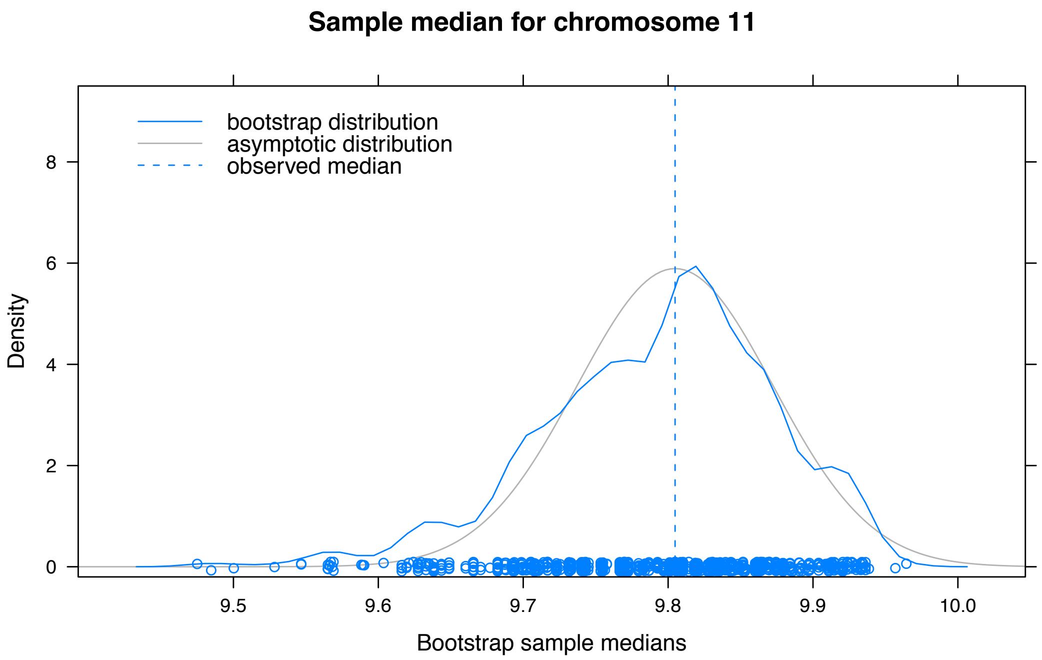
I do that B times. B is a big number.


Sample median for chromosome 11
bootstrap distribution
asymptotic distribution
observed median
Density
0
2
4
6
8
9.5
9.6
9.7
9.8
9.9
10.0
Bootstrap sample medians



Features we could foresee:

- Both dist'ns have mode @ sample median = 9.8
- Left tail of bootstrap distribution heavier than than that of asymp. norm



I conclude ... for a data-generating distribution as bimodal as this, n = 300 is close to -- but not quite in -- Asymptopia.

Features we could foresee:

- Both dist'ns have mode @ sample median = 9.8
- Left tail of bootstrap distribution heavier than than that of asymp. norm

Typical application of the bootstrap

- The standard deviation of a statistic (“standard error”)
- Confidence intervals
- The bias of an estimate
- Assess whether the asymptotic distribution has started to “kick in” at a finite sample size

Good default template for conducting a bootstrap. Can be adapted for other resampling or random data generation tasks.

```
> B <- 1000
```

make it easy to start w/ small B, then scale up

```
> bootData <-
+   matrix(sample(x, size = B * nx, replace = TRUE),
+          nrow = nx, ncol = B)
```

generate the bootstrap data all at once

```
> bootTestStat <- apply(bootData, 2, median)
```

use data aggregation techniques to compute bootstrap statistics

```
> (bootStdErr <- sqrt(var(bootTestStat)))
[1] 0.08163377

> mean(bootTestStat)
[1] 9.796118

> jMedian
[1] 9.804809

> abs(mean(bootTestStat) - jMedian)/bootStdErr
[1] 0.1094916
```

R packages for bootstrapping

- I don’t really use these ... always seems easier to just do it myself.
  - boot, a companion to another book ("Bootstrap Methods and Their Applications" by A. C. Davison and D.V. Hinkley (1997, CUP)) -- seems to be distributed with R
  - bootstrap, companion to Efron and Tibshirani’s book, seems not to be actively maintained, new work is encouraged to use boot

How large should B be?

- Efron & Tibshirani seem very comfortable with B = 200 for the purposes of estimating standard error
- You will need more -- generally much much more -- for confidence intervals. Depends on method. Beyond our scope today.
- I often default to B = 1000 for std err estimation or testing, but frequently use much larger B, such as 10000. Why not?

Parameter estimation and hypothesis testing

In statistical data analysis, we often use of the two following types of statistical inference. Each uses a different type of sampling distribution.

1. **(Parameter) estimation:** quantifying confidence in parameter estimates —sampling distribution of the estimate. Make parametric assumptions about model parameters, or use a computer intensive method for estimate the sampling distribution (Bootstrap).

2. **Hypothesis testing:** The probability distribution of the test-statistics if the null hypothesis is true, we need to estimate the null distribution. Parametric assumption, or computer intensive method (permutation testing).

Permutation test in hypothesis testing

Review: basics of hypothesis testing

1) Specify a null hypothesis
2) Choose a test statistic
3) Determine the distribution for the test statistic under null H_o
4) Convert the observed test statistic into a p-value:
   - "The p-value is the probability under H_o of observing a value for the test statistics the same or more extreme than what was actually observed".

     All of Statistics by Larry Wasserman. Springer 2004.

     All of nonparametric Statistics by Larry Wasserman, Springer 2006.

Simple example: differential gene expression analysis

- Suppose we find to find genes that are differentially expressed between different conditions.
- We compute the test statistic (e.g., t-statistics) for each of the g genes.
- We compute the p-value associated with each test statistic, call it p_g
  - p_g is the probability under null that the test-statistic is at least as extreme as T_g
- We correct p_g's for the number of tests (g tests)
- Declare a significant association if corrected p-values < threshold (0.05)

Standard t-test

- Assume $X_1,X_2,\ldots,X_m$ are from ~ $N(\mu_1 | \sigma^2)$
- Assume $Y_1,Y_2,\ldots,Y_n$ are from ~ $N(\mu_2 | \sigma^2)$

- Compute the pooled variance estimate:

$$s^2 = \frac{1}{m+n-2}(\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2).$$

- The t-statistic is given by

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Standard t-test

- Assume $X_1, X_2, \ldots, X_m$ are from ~ $N(\mu_1 | \sigma^2)$
- Assume $Y_1, Y_2, \ldots, Y_n$ are from ~ $N(\mu_2 | \sigma^2)$

- C

Under the null hypothesis $u_1 = u_2$ & t-statistics follows t_{m+n-2}-distribution

- The t-statistic is given by

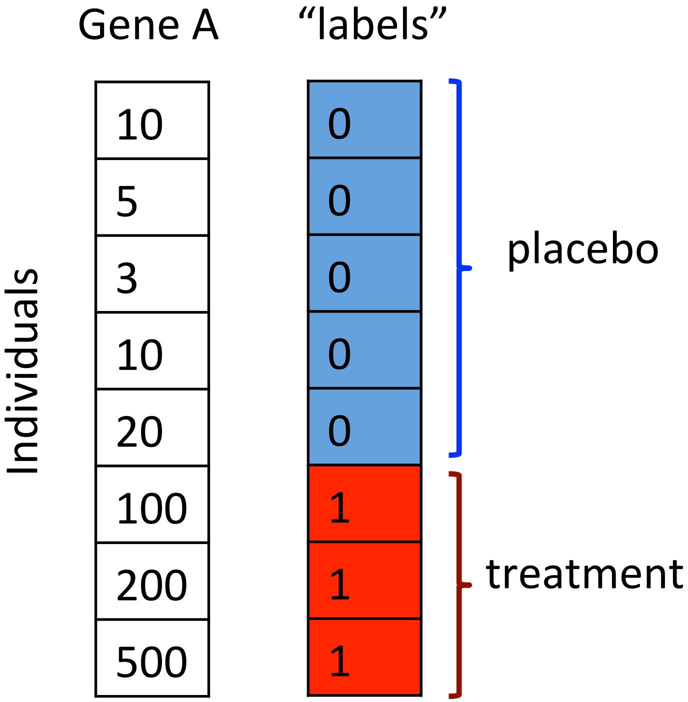
$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Permutation test

- Want to test whether observation in two groups follows the same distribution, without making assumptions about the distributions (e.g., normality)
- Generate a null distribution for the test-statistic:
  - Randomly divide individuals to ‘treatment’ groups

- For i = 1 …. p, do
  - Permute the group labels, giving new assignment of ‘group; to each individual
  - Computer the test statistic for the permutated data

“Real data”

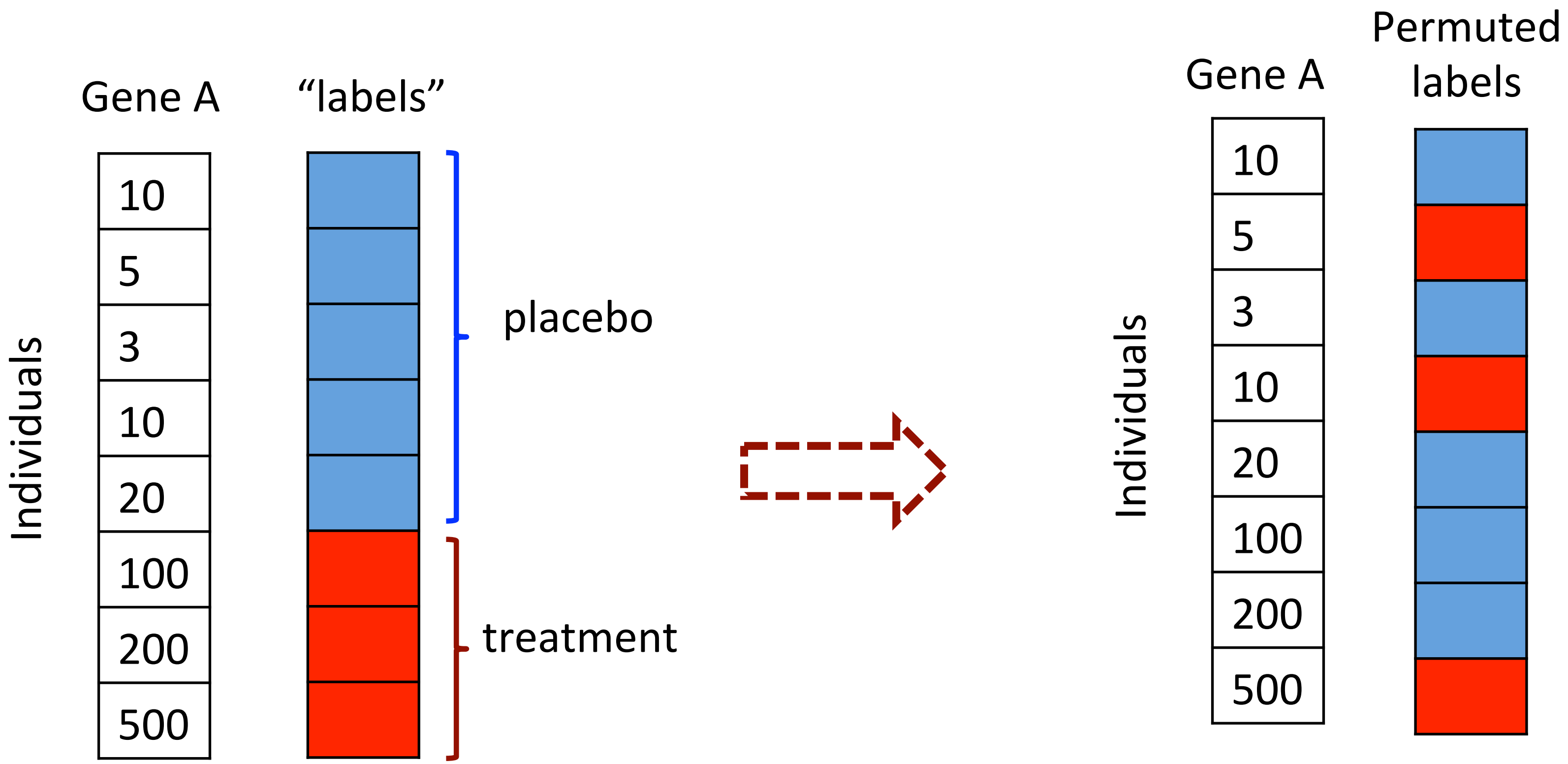


Individuals

Gene A	“labels”	
10	0	placebo
5	0	
3	0	
10	0	
20	0	
100	1	treatment
200	1	
500	1	

“Real data”

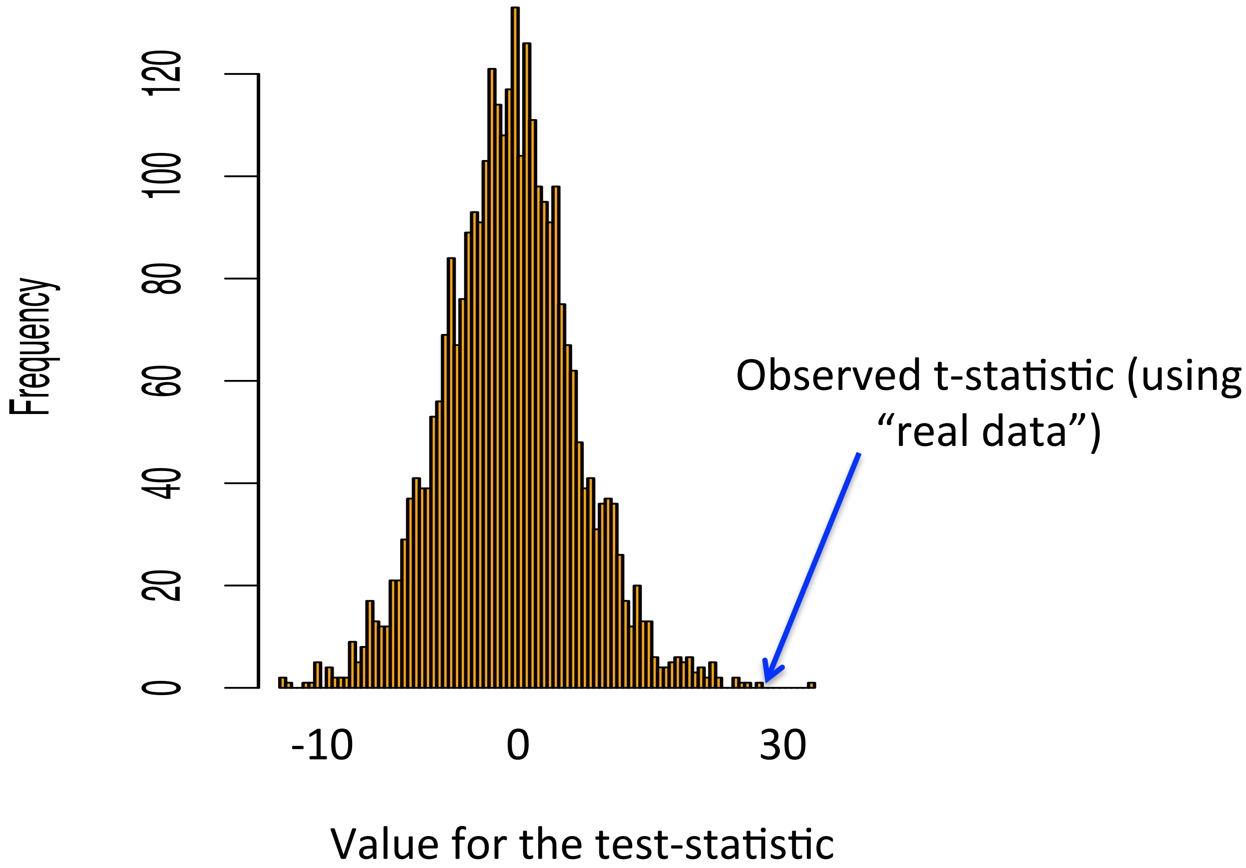
“Permutated data”



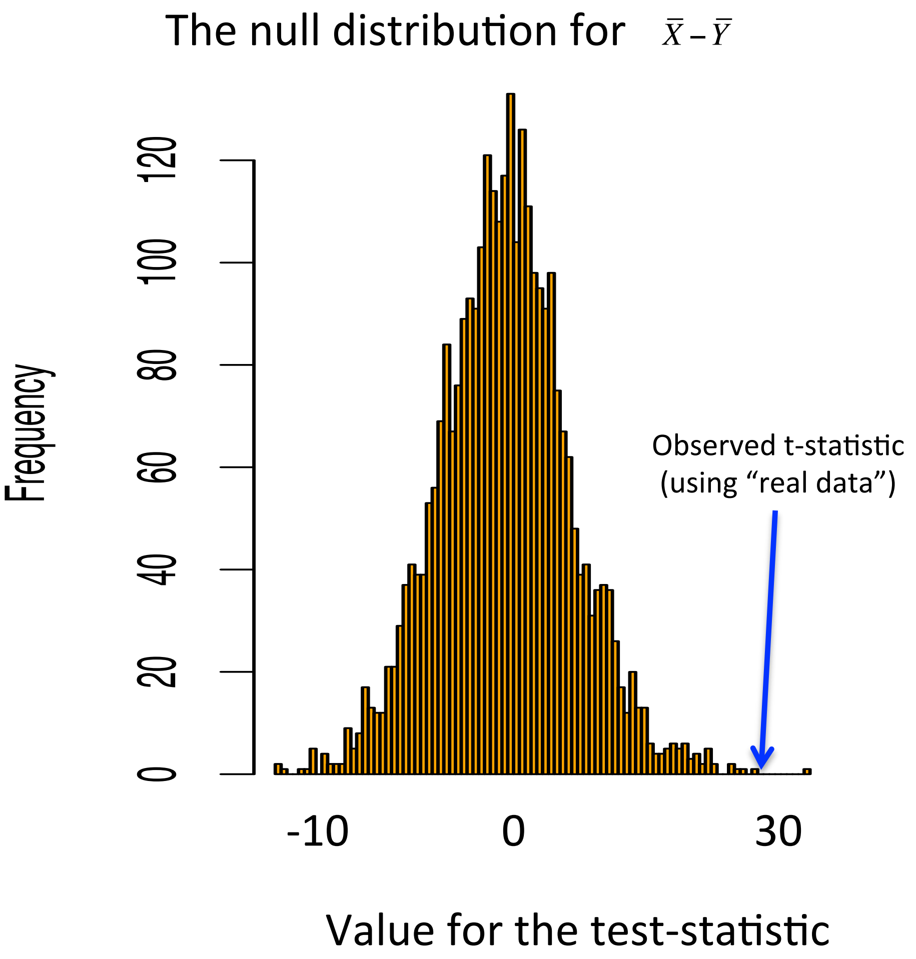
Individuals	Gene A	“labels”	
	10	🟦	placebo
	5	🟦	
	3	🟦	
	10	🟦	
	20	🟦	
	100	🟥	treatment
	200	🟥	
	500	🟥	

Individuals	Gene A	Permuted labels
	10	🟦
	5	🟥
	3	🟦
	10	🟥
	20	🟦
	100	🟦
	200	🟦
	500	🟥

Histogram of test-statistic under null (permutated data)



Histogram of test-statistic under null (permutated data)


The null distribution for $\bar{X} - \bar{Y}$

Observed t-statistic (using “real data”)

P-value: $$\frac{\#(T_p > T_r)}{\#\ permutations}$$

Resampling methods

- Ways of performing statistical inference that are “internal to the data” under analysis: e.g., you get the necessary knowledge about sampling variability (of parameters/ estimates) from the observed data itself

- Resampling methods:
  - Bootstrap
  - Permutation testing

Why the bootstrap is important

“If we choose a statistic more complicated than <sthgSimple> or a distribution less tractable than <sthgFriendly>, then no amount of mathematical cleverness will yield a simple formula.

Because of such limitations, pre-computer statistical theory focused on a small set of distributions and a limited class of statistics.

Computer-based methods like the bootstrap free the statistician from these constraints.”

p. 60-1 of Efron & Tibshirani

With power comes responsibility

“This is not all pure gain.

Theoretical formulas ... can help us understand a situation in a different way than the numerical output of a bootstrap program.

It pays to remember that methods like the bootstrap free the statistician to look *more closely* at the data, without fear of mathematical difficulties, not *less closely*.”

p. 60-1 of Efron & Tibshirani