Non-parametric resampling methods

1)(non-parametric) bootstrap
2) Permutation (randomization)

October 16th 2014

Goals

- What is resampling?
- The permutation/randomization test for generating distributions of test statistics.
- The (non-parametric) bootstrap for inference and constructing CI's.
- sample(), the heart of implementing resampling in R.

Readings for next week Non-parametric resampling

- Moore (this is a very introductory and gentle introduction if needed).
- Crawley (R Book): 284, 287, 418-421
- Fox_appendix, Crowley

- I have a few additional (for advanced) readings that are worth holding onto if you plan to use resampling methods in the future.

Screencasts

- https://github.com/idworkin/ZOL851/blob/master/LinksToScreencasts.md

- Updated with 4 screencasts (total time ~ 1hr).

Resources

- Good, P.I. 2006. Resampling methods. 3rd ed. Birkhauser.
- Manly, B.F.J. 2007 Randomization, Bootstrap and Monte Carlo methods in Biology. 3rd ed. Chapman Hall
- Efron, B. & Tibshriani, R.J. 1993. An introduction to the Bootstrap.

Relationship between Estimation methods

MCMC

Bayesian Estimation

$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)*p(\theta)$

Maximum Likelihood

$Max(L(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta))$

Monte Carlo Simulation

Resampling methods

Least Squares

$$Min(\sum \varepsilon_i^2)$$

Assumptions of parametric statistical approaches

- Data (in particular errors) conform to known distributions.

- Sample data is representative of population from which it is collected.

- Sampling distributions for test statistics based on these and other assumptions.

Resampling

- DISTRIBUTION FREE!!*
- Inference based on **empirically** derived distribution.
  - Computer intensive procedures
    - Permutation(randomization)/
    - Bootstrap
    - Jackknife
    - Cross validation.

* Some conditions may apply….

Resampling

- Resampling methods can be used for a broad range of problems even without any knowledge of parametric form of the variation.
- Easy to generate even "by hand" with spreadsheets.
- Can be used for "make your own statistics".
- Based on the data itself; fewer assumptions

Critiques

- Complex problems: Difficulty in determining how to perform the resampling. In particular for trying to formulate null models.
- Arguably limited in value as there is no formal logic as to how to use inductive approaches (go from your data set to the population at large).
- More generally they are a tool to help in making inferences or estimating uncertainty for **a given model**, and **are not** generally useful in thinking about model selection (with some exceptions).

Two words of caution
SAMPLE SIZE

- Resampling procedures should not be used with really small sample sizes.

- Permutation - Sampling without replacement
  # of independent ordered permutations is (N!).

But for two sample problem N!/n!m! will be far fewer (n! size of group1, m! size of group2)

- Bootstrapping - Sampling with replacement
  # of independent bootstraps(2N -1)!/(N!(N-1))

Permutation (randomization) tests

- Fisher (1935)
- Sampling **WITHOUT** replacement.
- Assumptions .. Exchangeability of observations **under the null model** (within and between treatments).   i.e *iid*
- Within - observations are *independent*.
- Between - treatments are *identically distributed* (of whatever form) and have a common variance. Or.. share same correlation structure.

How is resampling like Monte Carlo Simulations?

- You are using information directly from the sample to help generate inferences.
- Computationally intensive & iterative process.
- Can easily incorporate complexity into models (and null models) such as heterogeneity of variance across treatments.

How is resampling unlike Monte Carlo Simulations?

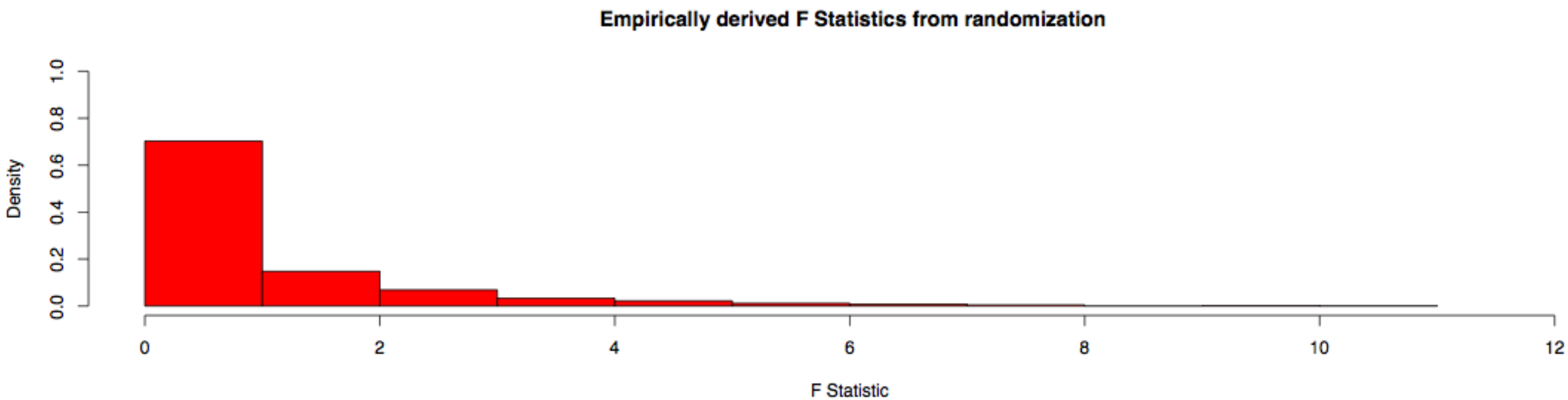
- Empirical distributions vs strong parametric assumptions.
- Resampling is about the sample, and makes no assumptions about the population.
- The downside is, that it is difficult to then relate the results back to “the population”

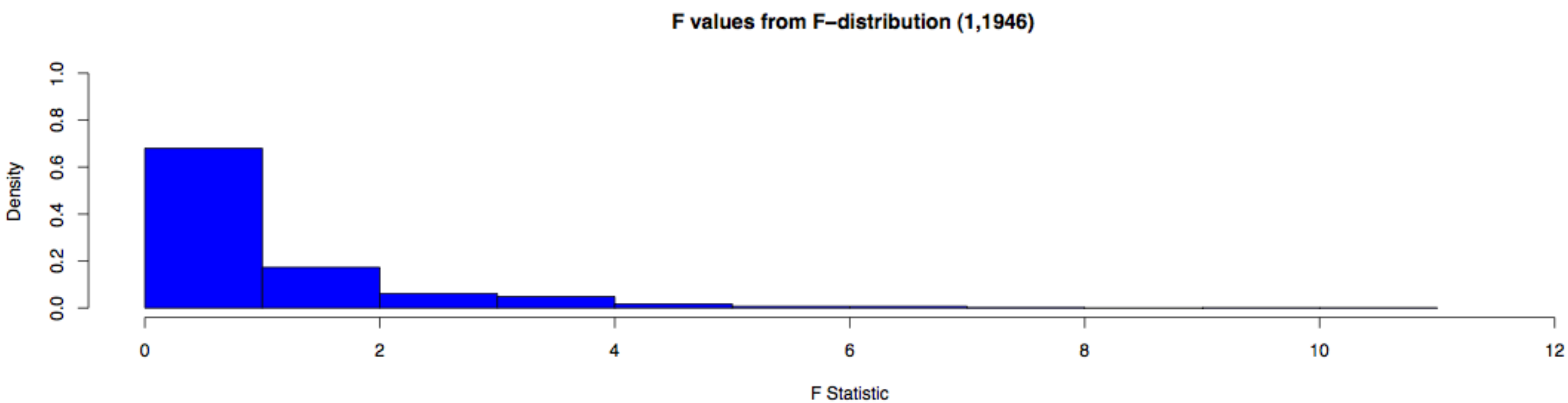
Sampling distribution of test statistics

- Sampling distributions for test statistics (student t, chi, F) were originally devised (and derived) in part, because the empirical "null" distributions could not be easily generated.
- Now with computers, this is no longer an issue.
- We can easily generate the sampling distribution of the test statistic under the null hypothesis of NO difference between treatments.

Empirical and theoretic F distribution

For data that conforms to the standard assumptions for linear models reasonably well (iid, $\varepsilon \sim N(0,\sigma^2)$)





Permutation tests

- This is a useful property, **BUT** remember **we do not need to assume** that the data conforms to any particular distribution.
- Nor do we need to know *apriori* the distribution of the test statistic (classical or newly developed).

Sampling without replacement

- All observations from the observed data set are included for each resampling event.
- What changes is that the values for the dependent variable are randomly re-assigned to the explanatory variables.

B

Sampling without replacement

- For each resampling event, all of the observations are included. Thus all of the summary statistics (mean, var, CI) **DO NOT** change across the sample.

- Since observations are exchanged under the null hypothesis of all observations coming from the same population, resampling will produce a distribution of values of the test statistic under a **null model**.

B

Work in groups to come up with the steps for implementing the permutation test.

Permutation test - procedure

1. Determine some useful measure (test statistic) to compare populations. Calculate for observed.

2. Resample without replacement a large number of times (10000 is usually a good starting point).

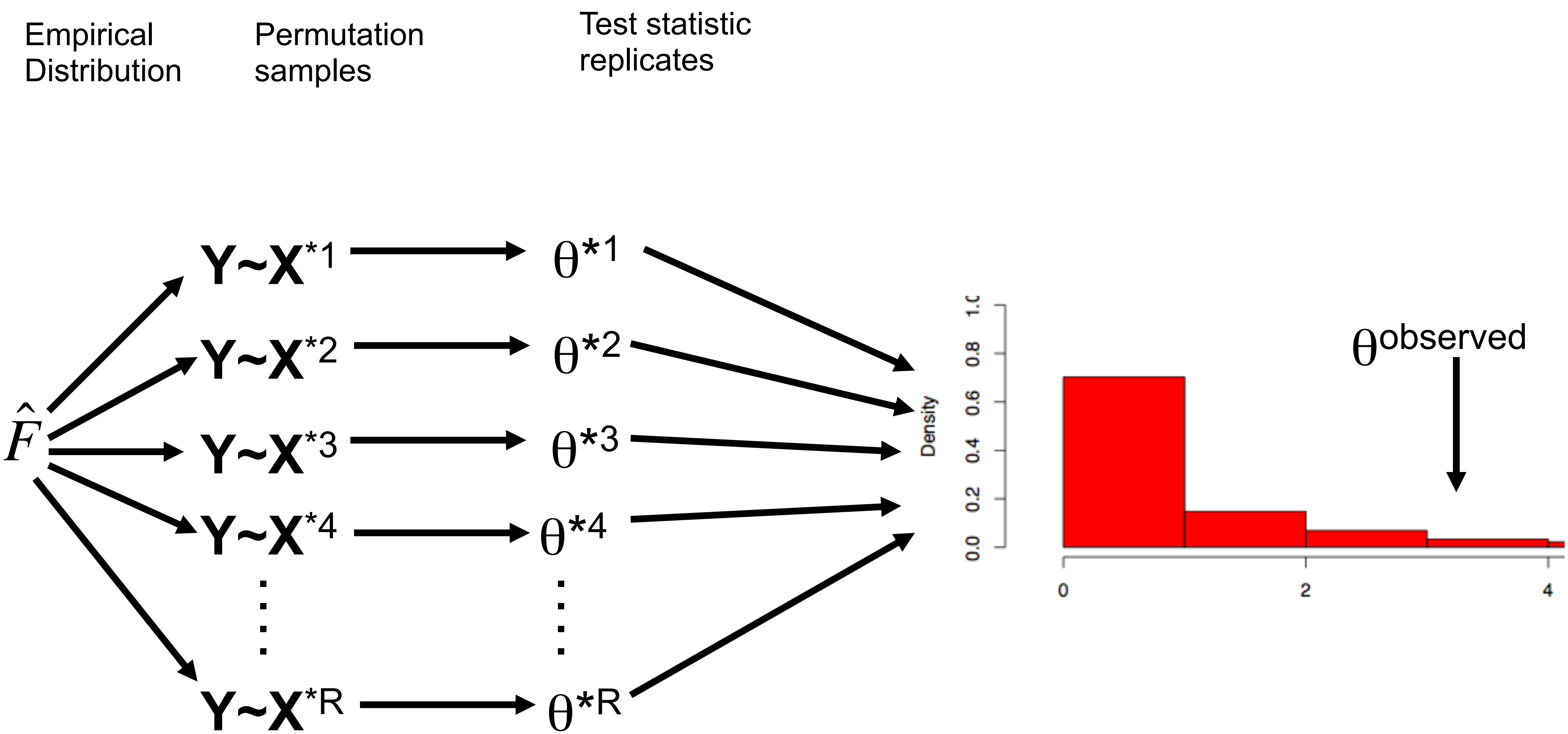
- Calculate test statistic for each resampling event. This generates the sampling distribution for the test statistic under the null hypothesis of no difference between groups.

3. Compare observed to the distribution of the resampled test statistics? Is it extreme for the distribution (careful for 1 vs 2 tailed hypothesis).

In general the p value will be

(# resamples > observed)/# resamples for one tailed tests.

The permutation algorithm



Sampling without replacement

In R the function we use is

sample(x, size, replace=F)

- x is a vector from which to choose, or a positive integer. This is generally your data which is going to be shuffled.
- size is a non-negative integer giving the number of items to choose.

There is also a way to do use this to sample from the index of the object (vector or matrix), instead of the data itself.

A parametric t-test

- A test of “means” across treatments
- the “t” test statistic

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \text{ where } s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2 + s_2^2}{n}}$$

Randomization Example

- Analogous to t-test
  - Continuous variable measured in two categories
  - e.g. measure # ant nests in forest vs field
- Steps
  - Randomly reassign category
  - Calculate difference in means
  - Repeat to build distribution of differences
  - If actual difference lies outside 95% interval then significant
- What is the null hypothesis?

Observed

Forest	7
Forest	5
Forest	8
Forest	8
Field	5
Field	6
Field	4

Forest = 7, Field = 5
Dif = 2

A.A

Example

Observed

Forest	7
Forest	5
Forest	8
Forest	8
Field	5
Field	6
Field	4

Forest = 7, Field = 5
Dif = 2

Random 1

Forest	7
Field	5
Forest	8
Field	8
Forest	5
Forest	6
Field	4

Forest = 6.5, Grass = 5.67
Dif = 0.833

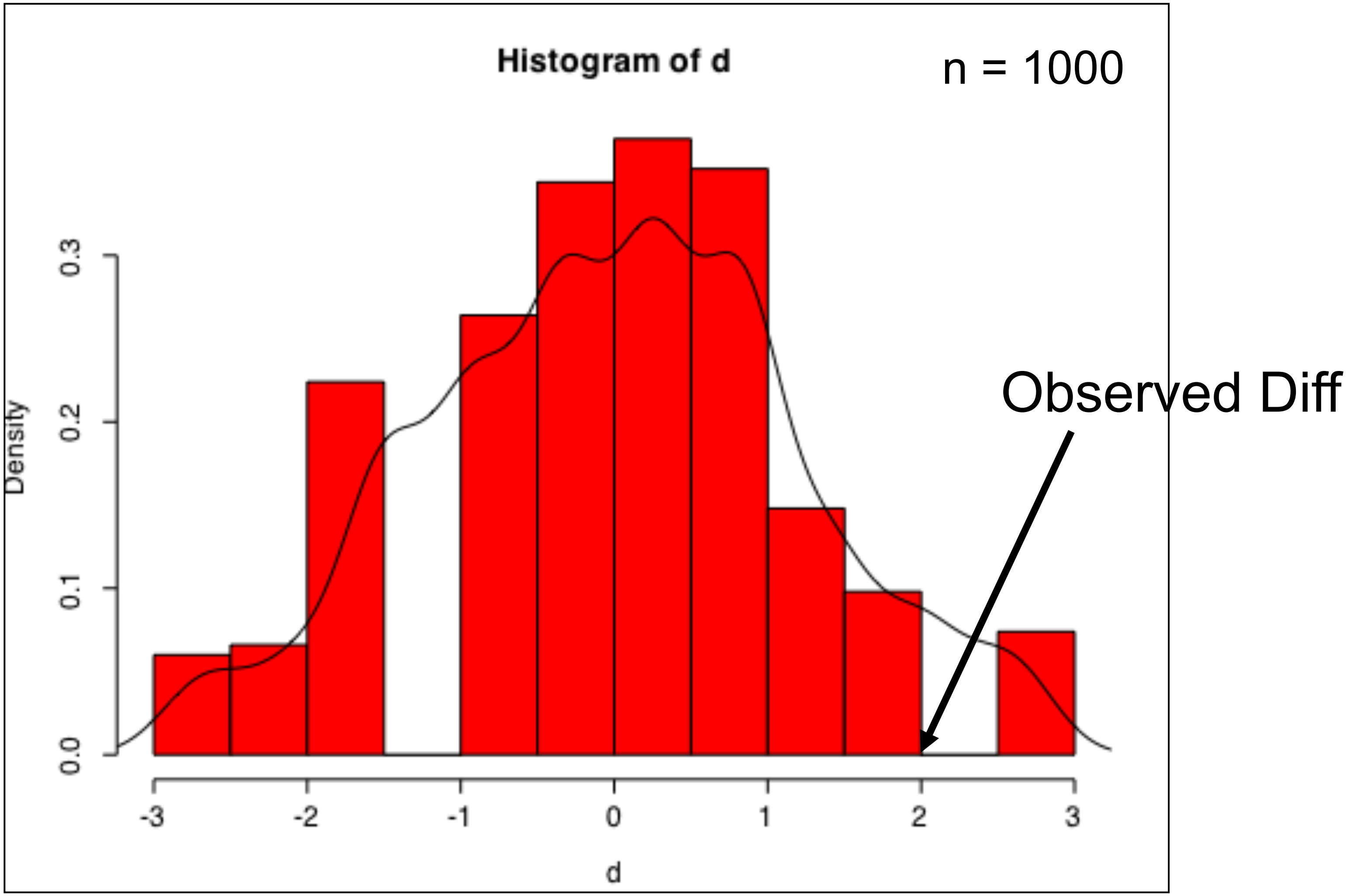
Random 2

Field	7
Forest	5
Field	8
Forest	8
Field	5
Forest	6
Forest	4

Forest = 5.75, Field = 6.67
Dif = -0.917

A.A

Randomized Diff's



86 Diff' s ≥ 2; P = 86/1000; P = 0.086

A.A

Advantages of Randomization

- Makes underlying assumptions and null hypothesis explicit
- Does not assume that data come from a specified distribution
- Can be tailored to your particular question of interest rather than being shoved into the box of an existing test

Disadvantages of Permutation

- Some dislike randomization because repeated analysis of the same dataset will yield different results (it is random!; P = 0.086, 0.085, 0.081). However large enough resampling sets (>10000) tend to converge enough for most purposes.

- Conclusions from randomization tests are restricted to the specific domain of the collected data.

- If **ALL** of the assumptions of a parametric model are met, then the equivalent parametric test will be more powerful than the resampling bases test.

Uses of the Permutation test

- data violated model assumptions.
- Test statistic distributions are difficult to determine (i.e. mixed model) or are unknown.
- Permutation tests can also be used for controlling for multiple testing issues (QTL mapping and association mapping).
- If you data conforms to the assumption of exchangeability, there is no reason not to perform it!!!! (go get a coffee while the computer does the work).

R

Bonferroni: divide alpha value (0.05) by number of replicates/samples

- assumes independence, if you have correlation (which you will in genomics) then this test is farrr too conservative
- How do we deal with this?
  - set up matrix with response and matrix of explanatory variables (when there's correlation)
  - shuffle the response across explanatory
  - take the smallest p-value from each test
    - this will create a distribution of p-values
  - this allows you to empirically determine the necessary threshold for a p-value
  - called the Churchill + Doerge permutation test

Holmes sequential: divide alpha by number of replicates/samples, divide alpha by n - 1, etc.

(non-parametric) Bootstrap

- Sampling **WITH** replacement
- Assumption: Exchangeability of observations within treatments (*independence*).
- No (general) assumption that treatments share the same distributional form (or common variance).
- Ho $\mu_a = \mu_b$

Bootstrap

- Bootstrap allows us to generate a sampling distribution for quantities we are interested in (mean, SD, CI).
- We are in effect generating random samples from the population of interest.
- It is about estimating how the quantity of interest varies due to random sampling.
- Also useful for examining *bias* (difference between estimated means for population and sample).

Sampling with replacement

- Each observation always has the same (and equal) probability of being “chosen” for each resampling event.

Sampling with replacement

- Observed data (5, 7, 6, 8, 10), mean = 7.2

- “bootstrap”1 (7, 7, 6, 10, 6), mean = 7.2
- “bootstrap”2 (5, 6, 8, 10 ,10), mean = 7.8
- “bootstrap”3 (6, 8, 5, 10 , 8), mean = 7.4
- “bootstrap”4 (7, 5, 7, 10, 5), mean = 6.8
- “bootstrap”5 (6, 5, 7 , 8, 10), mean = 7.2
- bootstrapped mean 7.28

In this instance this is highly biased since we have such a small sample size!!!

R

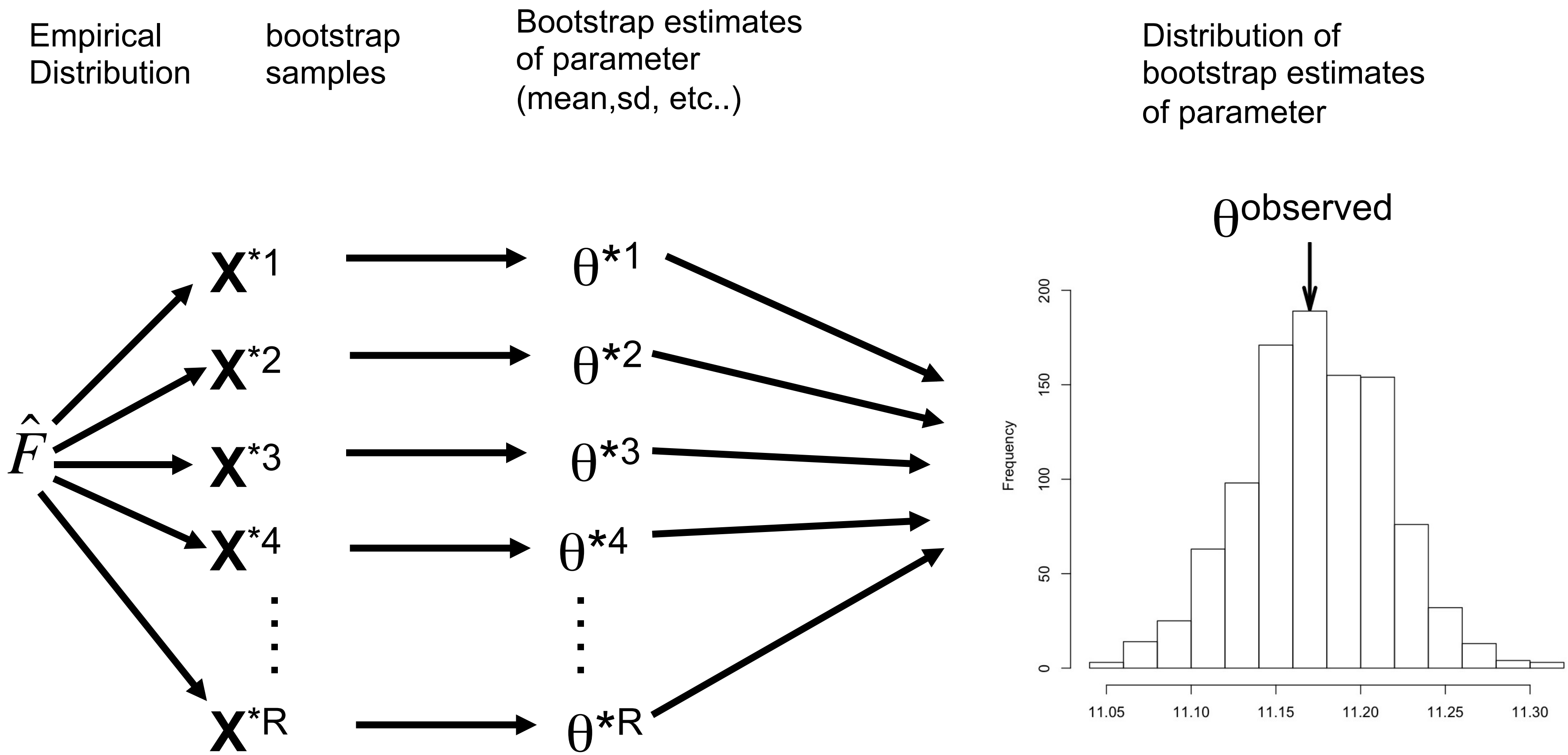
The bootstrap procedure

1. Calculate quantities of interest from the observed data.

2. Resample with replacement lots of times (>1000 is generally good). Calculate quantities of interest.

3. Look at empirically determined sampling distribution of quantities of interest. Calculate SE, CI’s etc..

The bootstrap algorithm



How would you use the bootstrap to construct confidence intervals?

Bootstrap confidence interval

- basic (percentile) bootstrap 95% CI are just the 2.5%, and 97.5% percentiles from the empirical distribution.
- Bias-corrected and accelerated confidence intervals (BC_a CI) are preferred (but the algorithm is complicated, so we will use a pre-built function).

Useful tidbits

- The bootstrap variance for any estimator ..

$$\hat{V}^*(T^*) = \frac{\sum_{b=1}^{R} (T^* - \overline{T^*})^2}{R-1}$$

where R is # of bootstrap replicates (independent of number of observations)..

• Bootstrap S.E = bootstrap S.D

•Bootstrap estimate of bias = mean of bootstrapped distribution - statistic for original data

Using the bootstrap to make statistical inference

- In addition to providing measures of uncertainty for our point estimates, bootstrapping can also be used for statistical inference.
- Ho $\mu_a = \mu_b$

A parametric t-test

- A test of “means” across treatments
- the “t” test statistic

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \text{ where } s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2 + s_2^2}{n}}$$

A bootstrapped based “t-test”

- Test statistic is still the difference in treatment means ($\mu_{a} - \mu_{b}$).
- However it is the uncertainty in the estimates that is treated differently.
- By sampling with replacement from the observations under each treatment, we are empirically estimating this uncertainty in the means, and thus the uncertainty in the difference ($\mu_{a} - \mu_{b}$).

t-test via bootstrap procedure

1. For each treatment (a & b) calculate test statistic of interest
   (i.e. mean(a) - mean(b))
2. Perform bootstrap resampling events as before for each group (a & b) separately. For each resampling event calculate test statistic.
3. Compare sampling distribution of the test statistic to a pre-determined null hypothesis.
   i.e. mean(a) - mean(b) = 0
4. P-values calculated as
   (# resampled test statistics > HO)/ # resampling events
   (or < depending on hypothesis)

Bootstrap

- We have only scratched the surface.
- Thursday we will examine different methods for making statistical inferences using the bootstrap.

(case bootstrapping vs. residuals).