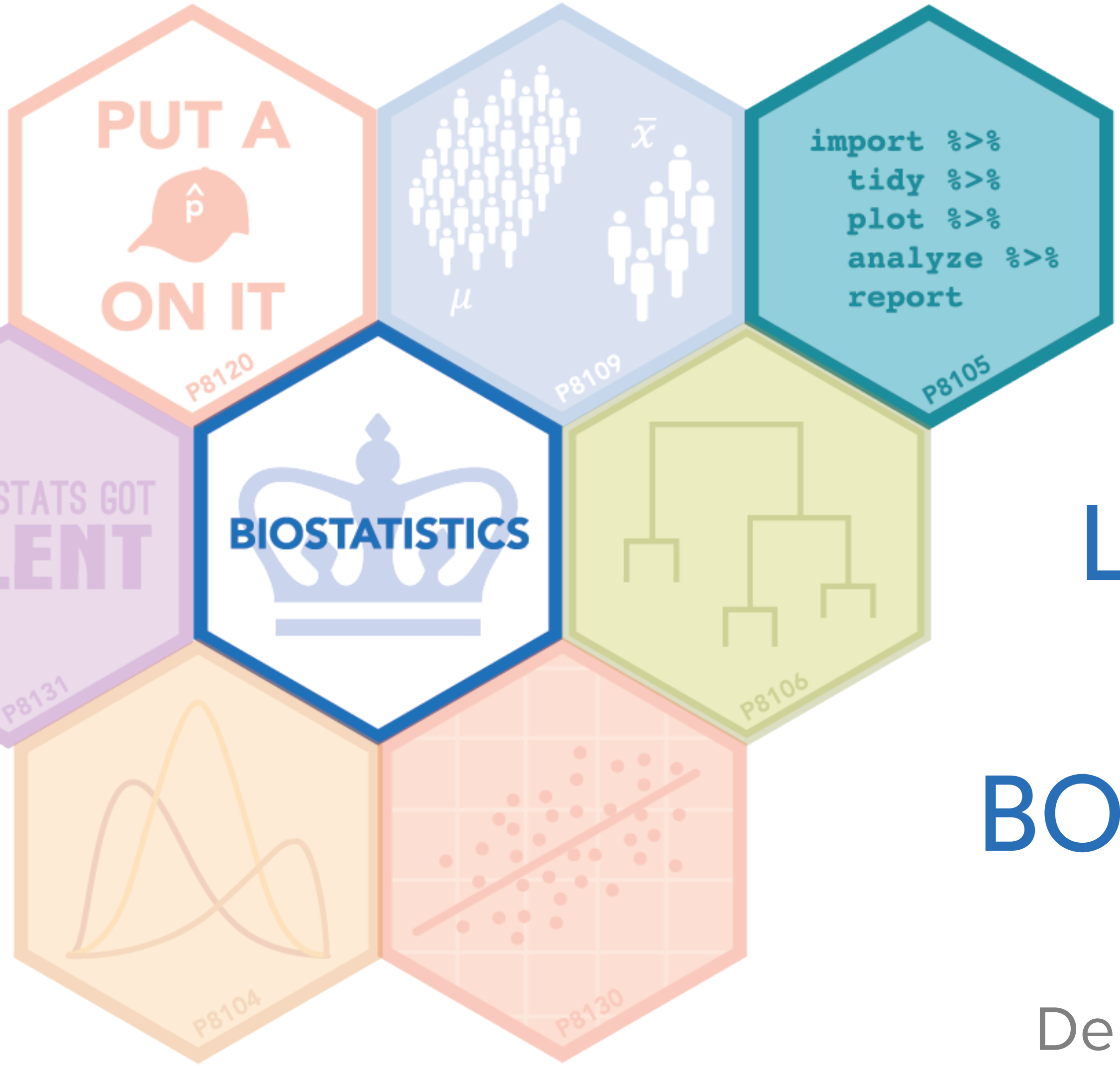


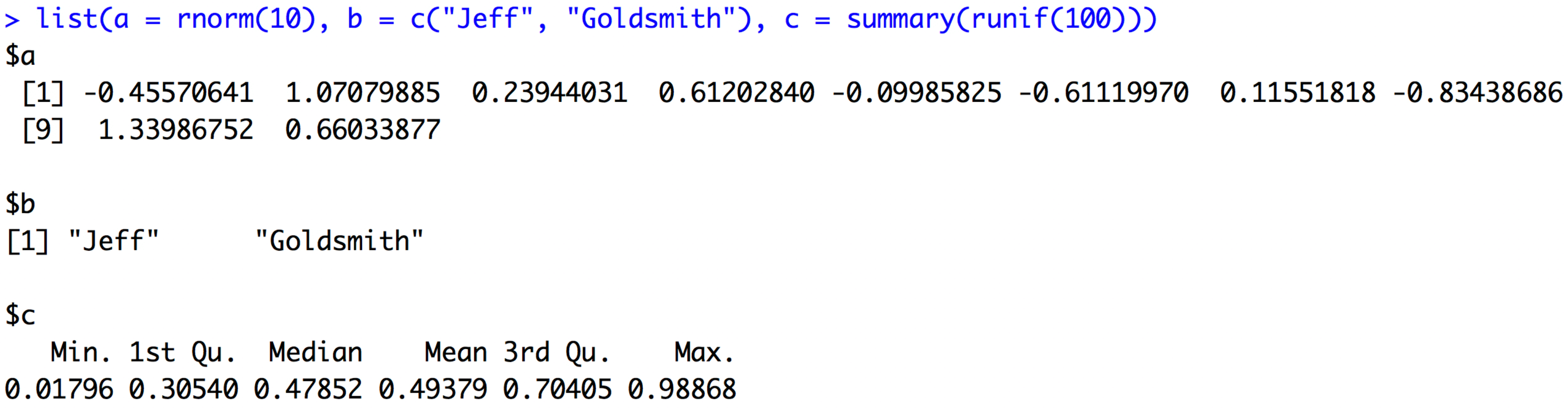
LIST COLUMNS AND BOOTSTRAPPING

Jeff Goldsmith, PhD

Department of Biostatistics

Lists

- In R, lists provide a way to store collections of arbitrary size and type
  - You can mix character vectors, numeric vectors, matrices, summaries…



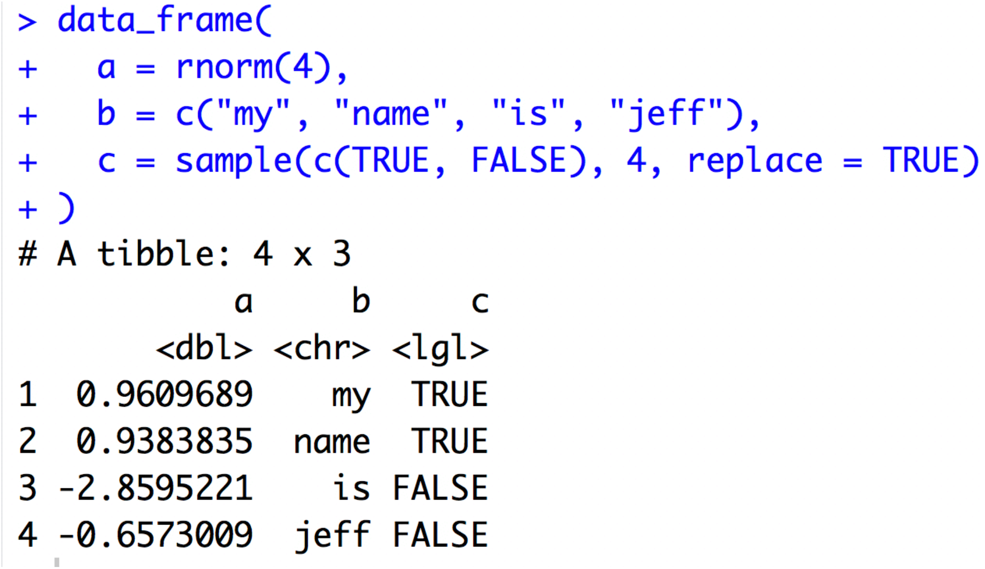
```
> list(a = rnorm(10), b = c("Jeff", "Goldsmith"), c = summary(runif(100)))
$a
 [1] -0.45570641  1.07079885  0.23944031  0.61202840 -0.09985825 -0.61119970  0.11551818 -0.83438686
 [9]  1.33986752  0.66033877

$b
[1] "Jeff"      "Goldsmith"

$c
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.01796 0.30540 0.47852 0.49379 0.70405 0.98868 
```

Data frames

- Data frames, which we've used extensively, are a special kind of list
  - Each list entry is a vector with the same length
  - You can still mix variable classes
  - Printed as a table



```
> data_frame(
+   a = rnorm(4),
+   b = c("my", "name", "is", "jeff"),
+   c = sample(c(TRUE, FALSE), 4, replace = TRUE)
+ )
# A tibble: 4 x 3
          a     b     c
      <dbl> <chr> <lgl>
1  0.9609689    my  TRUE
2  0.9383835  name  TRUE
3 -2.8595221    is FALSE
4 -0.6573009  jeff FALSE
```

List columns

- Lists can contain almost anything
  - A list can even contain a list!

- What if an entry in your list is a list, but it has the same length as the other entries?
- Could that be a “column” in a data frame?

List columns

- Lists can contain almost anything
  - A list can even contain a list!

- What if an entry in your list is a list, but it has the same length as the other entries?
- Could that be a “column” in a data frame?

YES!!

LEONARDO DICAPRIO
KEN WATANABE JOSEPH GORDON-LEVITT MARION COTILLARD ELLEN PAGE TOM HARDY CILLIAN MURPHY TOM BERENGER and MICHAEL CAINE
A FILM BY CHRISTOPHER NOLAN
INCEPTION

Seriously? YES!!!!!!

- List columns turn out to be very useful

- Imagine you have granular data nested within large units
  - Make a list storing your granular data table
  - Add granular data table list to a data frame containing larger units

- Why stop there??
  - You can store more complex R objects, like output from regressions on each granular data table, in a list
  - You can add that list to your data frame

- Keeping everything in one data frame with list columns means there are fewer things to worry about

Repeated sampling

- "Repeated sampling" is a conceptual framework that underlies almost all of statistics
  - Repeatedly draw random samples of the same size from a population
  - For each sample, compute the mean
  - The distribution of the sample mean converges to a Normal distribution

- Repeated sampling doesn't happen in reality
  - Data are difficult and expensive to collect
  - You get your data, and that's pretty much it

- Repeated sampling can happen on a computer

Bootstrapping

- Hard to overstate how important and useful bootstrapping is in statistics

- Basic idea is to mimic repeated sampling with the one sample you have
  - That sample is draw at random from your population
  - You'd like to draw more samples, but you can't
  - So you draw a bootstrap sample from the one sample you have
  - The bootstrap sample has the same size as the original sample, and is drawn with replacement
  - Repeat

Why bootstrap?

- The repeated sampling framework often provides useful theoretical results under certain assumptions or asymptotics
  - Sample means follow a known distribution
  - Regression coefficients follow a known distribution
  - Odds ratios follow a known distribution

- If your assumptions aren’t met, or your sample isn’t large enough for asymptotics, you can’t use the “known distribution”

- Bootstrapping gets you back to repeated sampling, and uses an empirical rather than a theoretical distribution for your statistic of interest

Coding the bootstrap

- Bootstrapping is natural in the context of iteration

- Write a function (or functions) to:
  - Draw a sample with replacement
  - Analyze the sample
  - Return object of interest
- Repeat this process many times

- Keeping track of the bootstrap samples, analyses, and results in a single data frame organizes the process and prevents mistakes

- That's why you use **LIST COLUMNS!!**