DISCOVERING STATISTICS USING R

Bootstrap

1

DISCOVERING STATISTICS USING R

Chapter 2

Populations and samples

- Population
  - The collection of units (alleles, cells, organisms, groups, species, etc.) to which we want to generalize a set of findings or a statistical model
- Sample
  - A smaller (but hopefully representative) collection of units from a population used to determine truths about that population

2

DISCOVERING STATISTICS USING R

Chapter 2

From sample to population

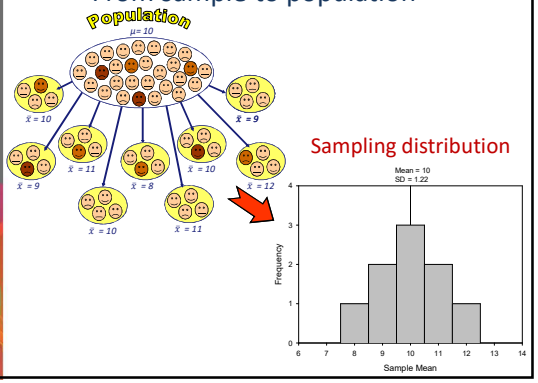
- Sample
  - Statistics we calculate (*e.g.* mean, sd, correlation coefficient, etc.) describe the sample from which they were calculated…
- Population
  - … but are intended to describe the entire population
- Sample to Population:
  - How well does a given sample represent the population?

3

DISCOVERING STATISTICS USING R

Chapter 2

From sample to population



4

DISCOVERING STATISTICS USING R

Chapter 2

From sample to population

- However, it is often very impractical (if not impossible) to collect many different samples in the real world
  - Parametric tests overcome this "fundamental unknowability" of the sampling distribution by making the **assumption** that, if sample size is big enough, then the distribution of data is approximately normal
  - The significance value of test-statistics calculated from a sample, critically depends on the validity of this, and other, assumptions

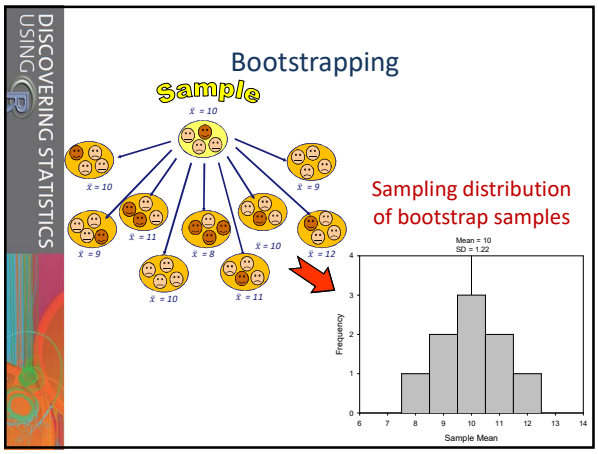
5

DISCOVERING STATISTICS USING R

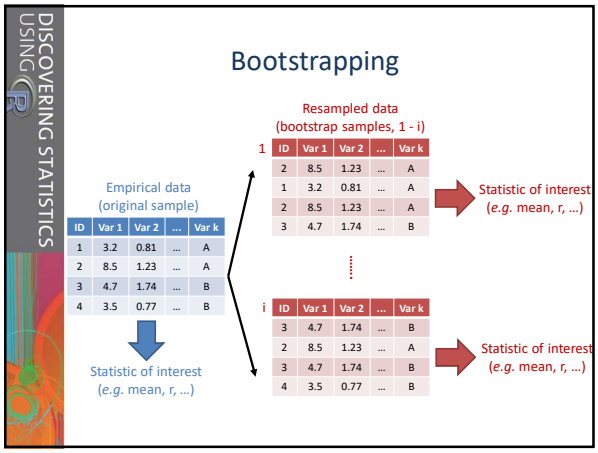
From sample to population
From re-sample to sample

- However, it is often very impractical (if not impossible) to collect many different samples in the real world
  - Bootstrapping overcomes the "fundamental unknowability" of the true sampling distribution by generating many bootstrap samples from the observed data
  - It thus makes no assumptions about the distribution of data, but reconstructs the sampling distribution using the information contained in the collected data

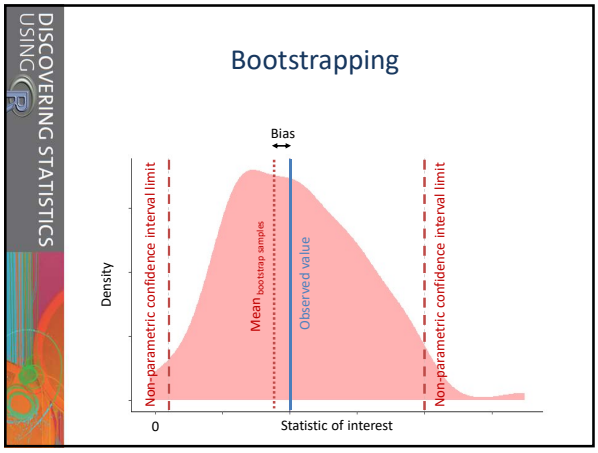
6


DISCOVERING STATISTICS USING R
Bootstrapping
Sample
Sampling distribution of bootstrap samples
Mean = 10
SD = 1.22
Frequency
Sample Mean

7


DISCOVERING STATISTICS USING R
Bootstrapping
Resampled data (bootstrap samples, 1 - i)
Empirical data (original sample)
Statistic of interest (e.g. mean, r, ...)
Statistic of interest (e.g. mean, r, ...)

8


DISCOVERING STATISTICS USING R
Bootstrapping
Bias
Density
Non-parametric confidence interval limit
Mean bootstrap samples
Observed value
Non-parametric confidence interval limit
0
Statistic of interest

9

DISCOVERING STATISTICS USING R

Bootstrapping

- Resampling (with replacement)
  - Useful 'trick', especially when:
    - The theoretical distribution of a statistic is unknown or complicated
    - Sample size is too small to allow parametric inference
- Confidence interval estimates
  - By default the "boot.ci()" function will try to calculate 5 non-parametric confidence intervals
  - Of these, the bias-corrected accelerated (BCa) intervals appear to be most robust across a wide range of data scenario's

10

DISCOVERING STATISTICS USING R

Bootstrapping in R

- 3 steps
  1. Write a function to calculate a statistic of interest, *e.g.* the mean

```
stat<- function(variable, i){
       mean(variable[i], na.rm= T)
       }
```

  2. Plug this function into the 'boot()' function, and set the number of bootstrap samples you want to generate

```
b.mean<- boot(dataframe$column, stat, 10000)
```

  3. Look at outcome, and confidence intervals (*e.g.* 99%)

```
b.mean
boot.ci(b.mean, 0.99)
```

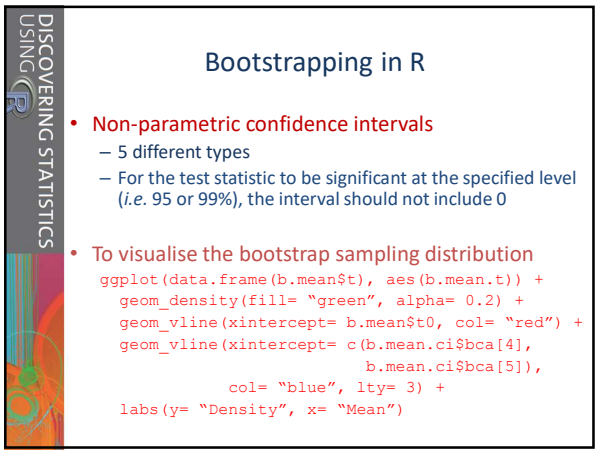
11

DISCOVERING STATISTICS USING R

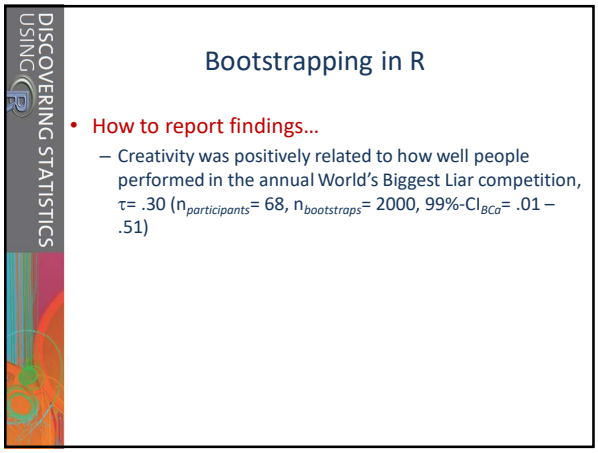
Bootstrapping in R

- The output reports three values
  - original
    - value of the test statistic in the original sample
  - bias
    - difference between the original value and the average value across all bootstrap samples
  - std. error
    - the standard deviation of the sampling distribution of values across all bootstrap samples

12


DISCOVERING STATISTICS USING R
Bootstrapping in R
• Non-parametric confidence intervals
– 5 different types
– For the test statistic to be significant at the specified level (i.e. 95 or 99%), the interval should not include 0
• To visualise the bootstrap sampling distribution
ggplot(data.frame(b.mean$t), aes(b.mean.t)) +
geom_density(fill= "green", alpha= 0.2) +
geom_vline(xintercept= b.mean$t0, col= "red") +
geom_vline(xintercept= c(b.mean.ci$bca[4],
b.mean.ci$bca[5]),
col= "blue", lty= 3) +
labs(y= "Density", x= "Mean")

13


DISCOVERING STATISTICS USING R
Bootstrapping in R
• How to report findings…
– Creativity was positively related to how well people performed in the annual World's Biggest Liar competition, τ= .30 ($n_{participants}$= 68, $n_{bootstraps}$= 2000, 99%-CI_{BCa}= .01 – .51)

14