Bootstrap Methods in R for Real-World Data

SEPA 2018 Methodology Workshop (2)
Matthew D. Turner
mturner46@gsu.edu
Georgia State University, Atlanta, GA

Today's Workshop

- Presenters (GSU Psychology):
  - Matthew Turner, PhD, Research Scientist
  - Jessica Turner, PhD, Associate Professor of Psychology
- Teaching Assistants (GSU Neuroscience):
  - Amber Grant, B.S.
  - Kendrick King, B.S. (due spring 2018)
- All of the slides, R code, handouts, etc., are in the files you copied from the USB sticks and include web links for more information.
- For more information contact: mturner46@gsu.edu
- If you would like the materials for other uses, please contact me

Overview

1. Brief Introduction to R
2. Review of Confidence Intervals
3. Percentile Bootstraps
4. BC_a Bootstraps
5. Examples of PB/BC_a
6. Brief Overview of other Bootstrap Options
7. Where Bootstraps Succeed/Fail

Introduction to R

This material is in the script:

boot_just_enough.R

Confidence Intervals (CIs)

- To provide some structure to this workshop we will focus on:
  - Confidence Intervals
  - Implementation in R

The New Statistics: Estimation and Research Integrity

Featuring

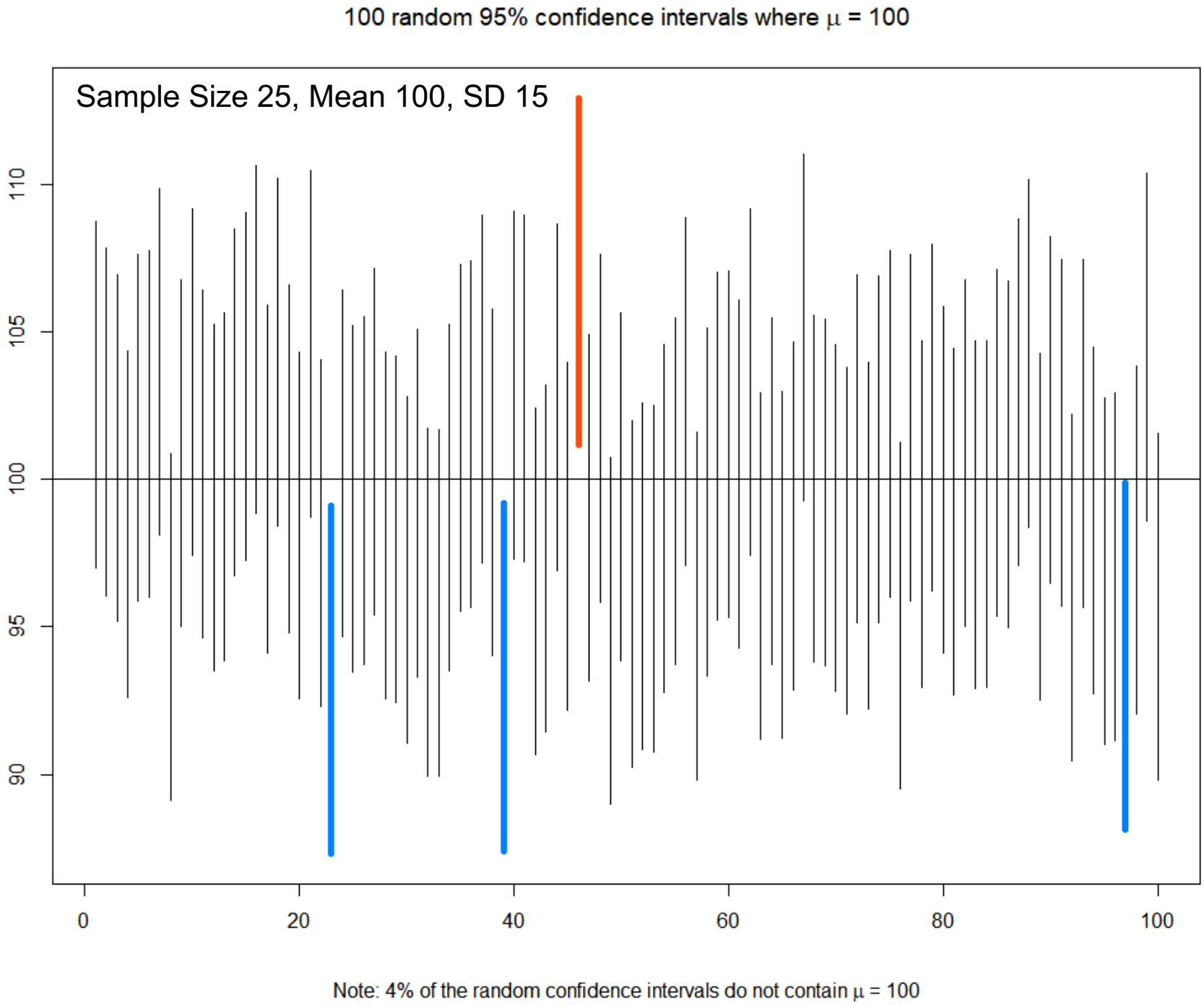
Geoff Cumming
La Trobe University, Australia

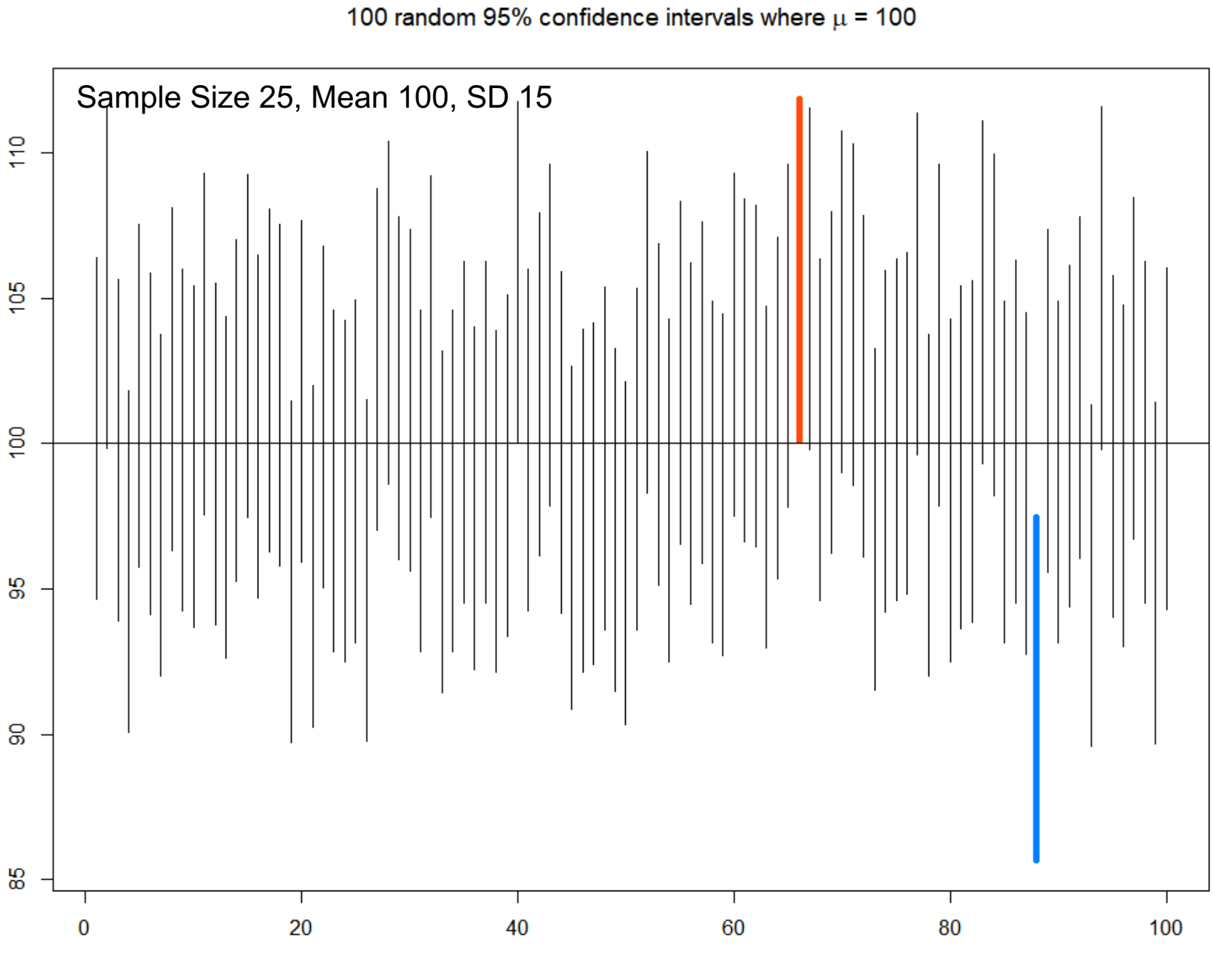
Leading scholars in psychology and other disciplines are striving to help scientists enhance the way they conduct, analyze, and report their research. They advocate the use of "the new statistics"— effect sizes, confidence intervals, and meta-analysis. APS's flagship journal, Psychological Science, has been inviting authors to use the "new statistics" as part of a comprehensive effort to enhance research methodology.

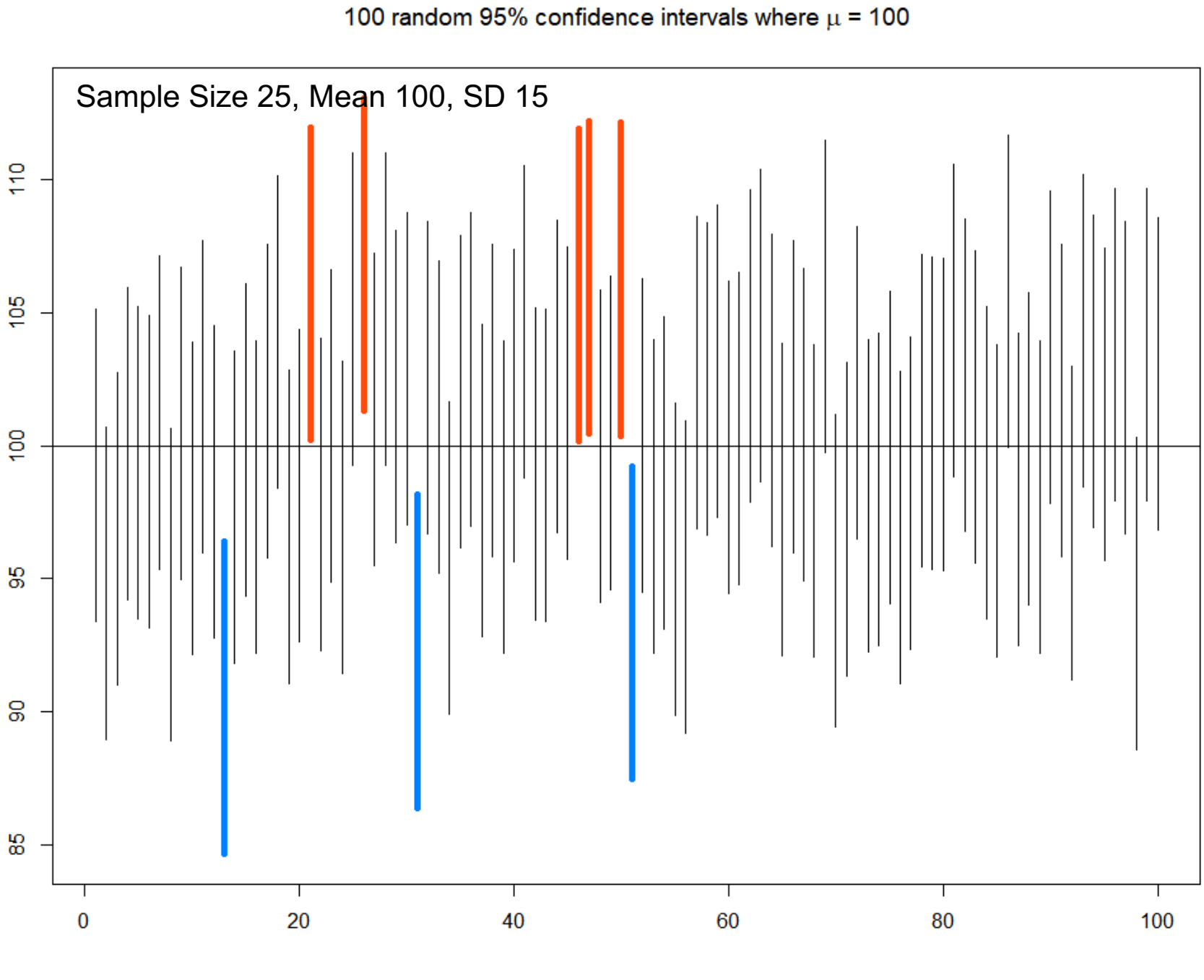
Confidence Intervals (CIs)

Most people are a little weak on CIs so let's review

- Confidence intervals place "reasonable" bounds on an estimate of a parameter
- They are a statement about procedure:
  - A 95% CI says that *"95% of confidence intervals constructed by this method will capture the true parameter within their bounds"*
  - Parameters, in classical statistics, are fixed numbers
  - **The true (value) of the parameter is either in the CI or not in it for any given case**


100 random 95% confidence intervals where μ = 100
Sample Size 25, Mean 100, SD 15
110
105
100
95
90
0
20
40
60
80
100
Note: 4% of the random confidence intervals do not contain μ = 100


100 random 95% confidence intervals where μ = 100
Sample Size 25, Mean 100, SD 15
110
105
100
95
90
85
0
20
40
60
80
100
Note: 2% of the random confidence intervals do not contain μ = 100


100 random 95% confidence intervals where μ = 100
Sample Size 25, Mean 100, SD 15
110
105
100
95
90
85
0
20
40
60
80
100
Note: 8% of the random confidence intervals do not contain μ = 100

Confidence Intervals (CIs)

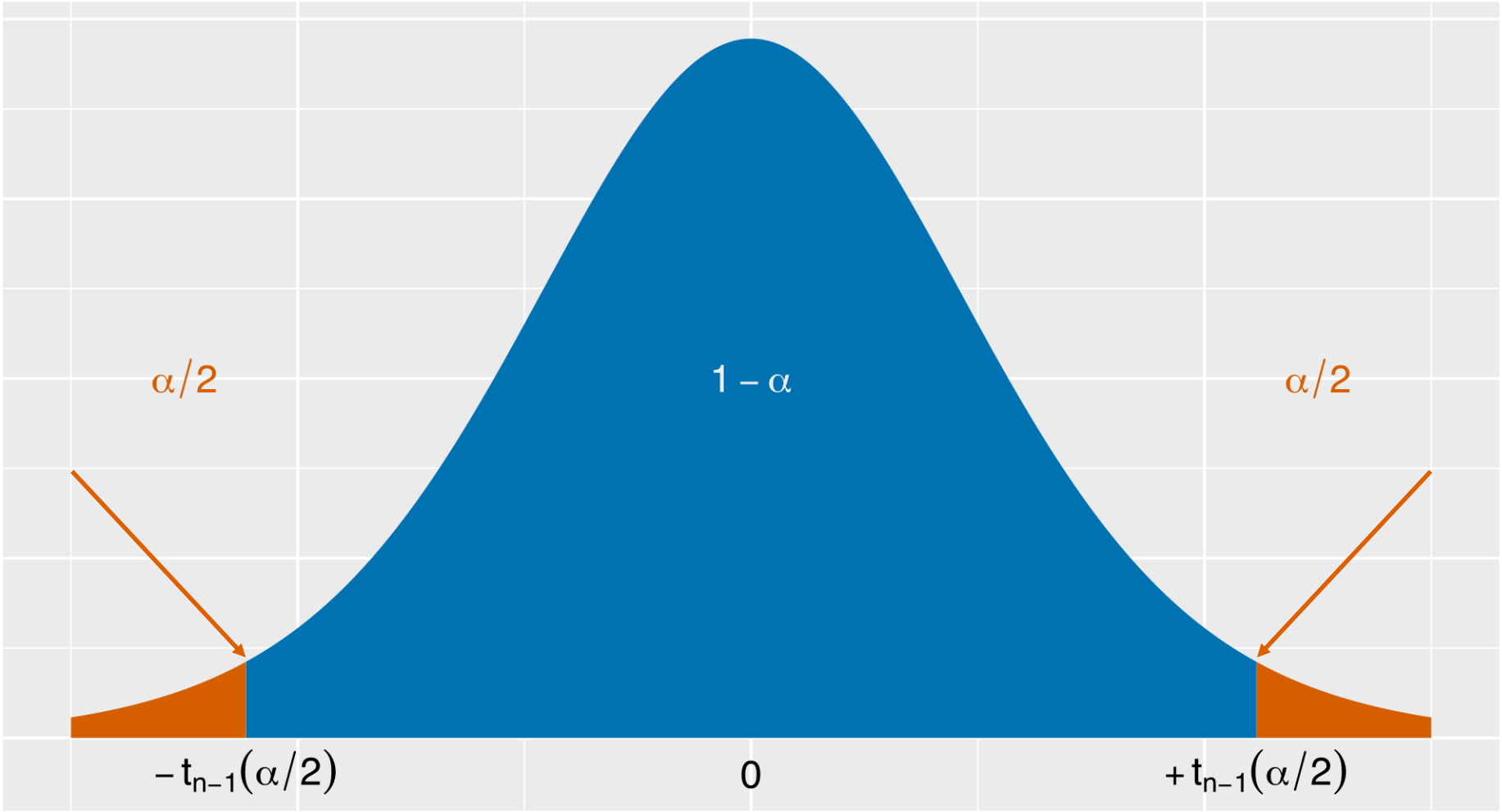
- Formulas:

$$\bar{x} \pm z_c \frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm t_c \frac{s}{\sqrt{n}}$$

- Where:
  - $\bar{x}$ is the estimate of the mean
  - s is the sample standard deviation
  - n is the sample size
  - z_c or t_c are the normal or t critical values, resp.
    - These critical values come from tables
    - They correspond to quartiles with probabilities $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$

Note:

- Standard intervals are (by definition) symmetric
- The critical values (t, z, whatever) mark out $\frac{\alpha}{2}$ area in the tails

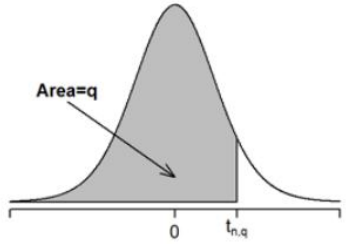


https://commons.wikimedia.org/wiki/File:Confidence-interval.svg by ARAKI Satoru

CDF for t Distribution

Quartiles of the *t* Distribution

The table gives the value if $t_{n;q}$ - the *q*th quantile of the *t* distribution for *n* degrees of freedom



	q = 0.6	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
***n* = 1**	0.3249	1.0000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.2887	0.8165	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.2767	0.7649	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.2707	0.7407	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.2672	0.7267	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.2648	0.7176	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.2632	0.7111	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.2619	0.7064	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.2610	0.7027	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.2602	0.6998	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.2596	0.6974	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.2590	0.6955	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.2586	0.6938	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.2582	0.6924	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140

CDF for the Normal

Table of Standard Normal Probabilities for Negative Z-scores

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table of Standard Normal Probabilities for Positive Z-scores

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

https://www.chegg.com/homework-help/questions-and-answers/use-standard-normal-table-find-z-score-corresponds-cumulative-area-00778-area-table-use-en-q16146181

How to do CIs in R

- There are some functions that give CIs directly, but most people just use the formula directly
  - With R's automatic tables this is easy!
  - R has simple functions for z and t

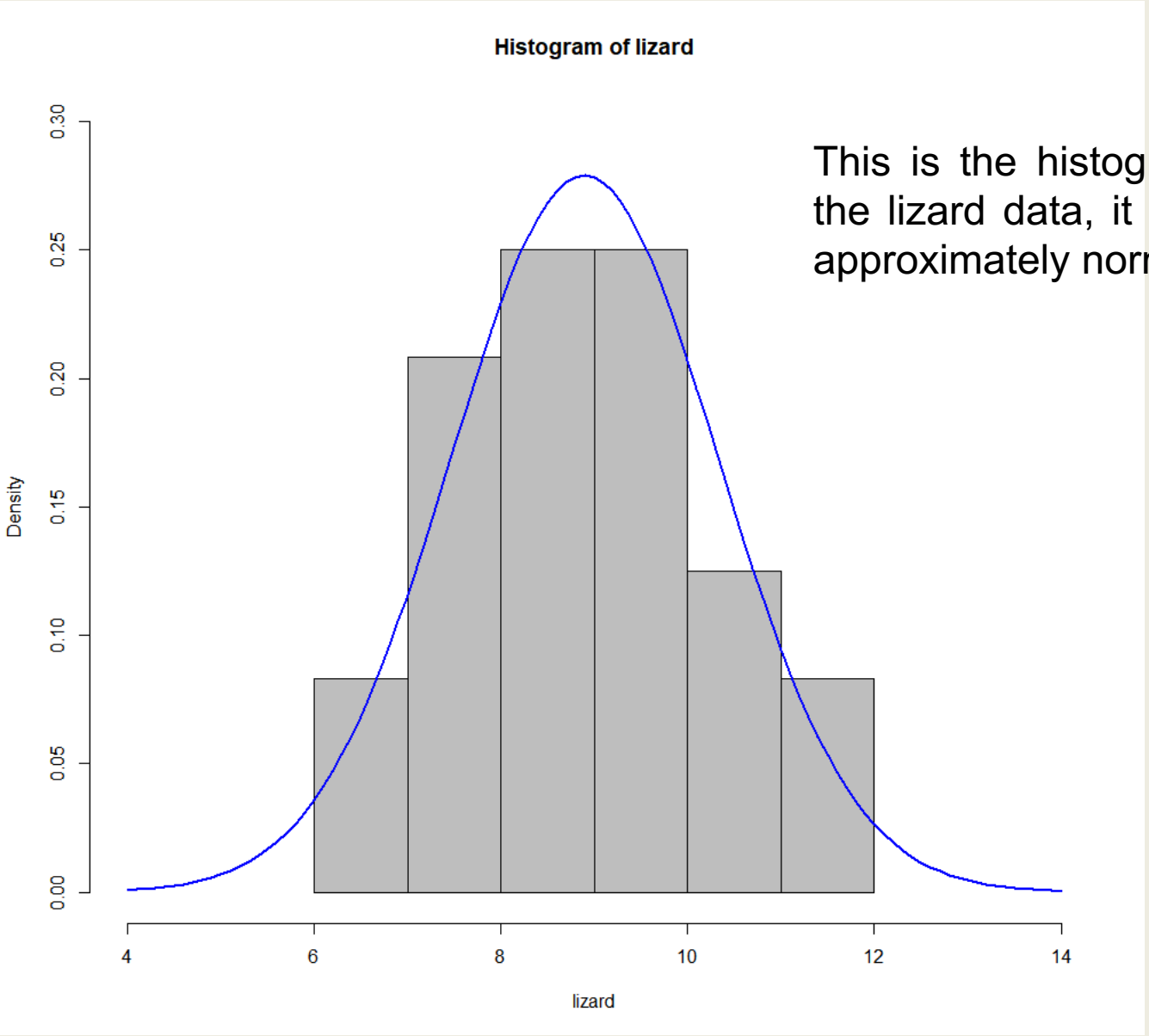
$$\bar{x} \pm z_c \frac{s}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm t_c \frac{s}{\sqrt{n}}$$

- Remember these formulas

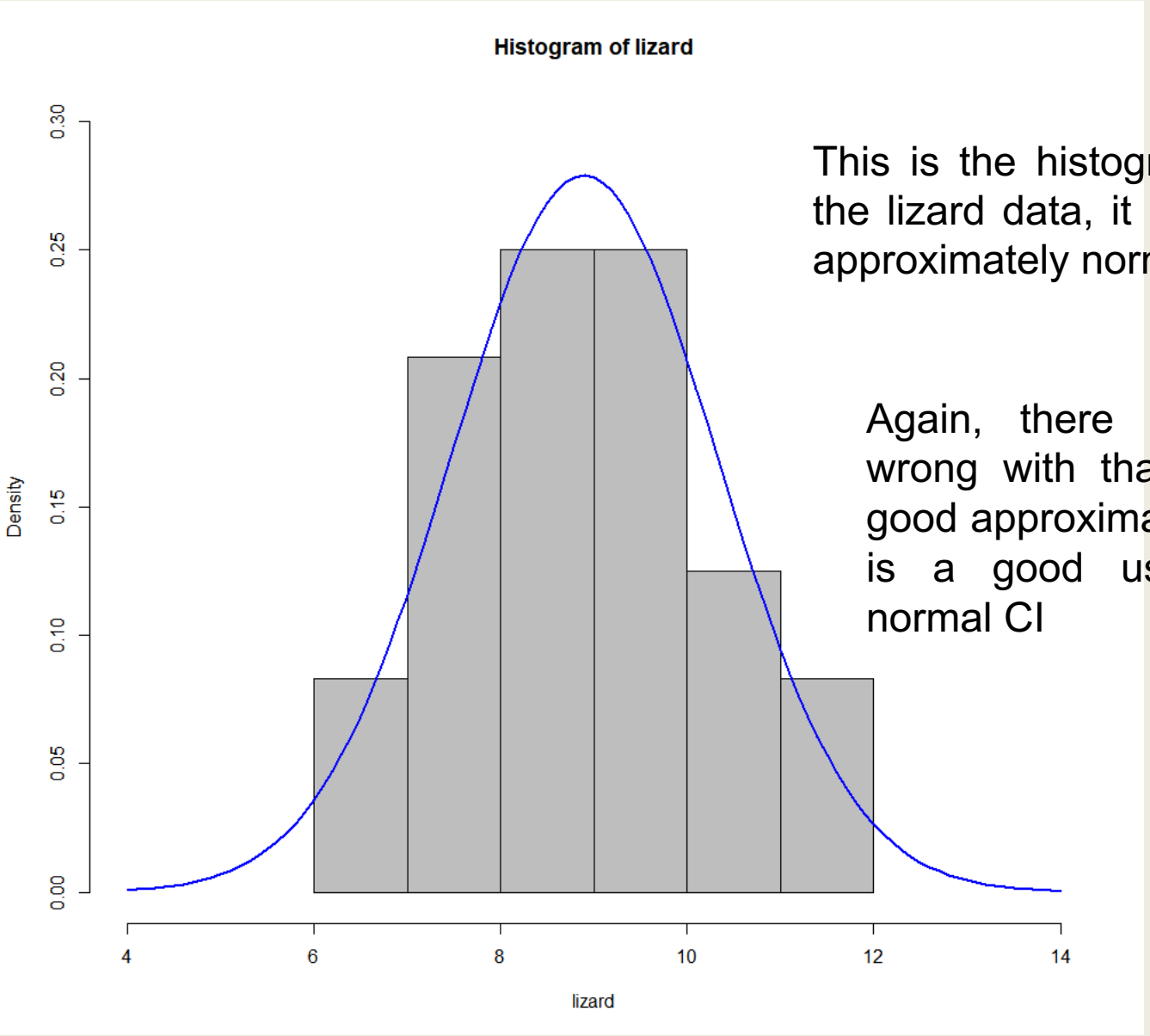
Move to section 1 on the R code file: `bootstrap_CI_Workshop_Main.R`

Normal (z or t) CIs

- People tend to think these are exact, but they are **not**
  - These Confidence Intervals are **only exact when the data is exactly normal**
  - Data is **never** exactly normal!
  - They are good approximations when the data is fairly normal
  - They are somewhat “robust” when the data is symmetric about its center
  - But they are approximations in almost all cases!!!



This is the histogram of the lizard data, it is only approximately normal!



This is the histogram of the lizard data, it is only approximately normal

Again, there is nothing wrong with that. We like good approximations! This is a good use of the normal CI

Bootstrap #1: The Percentile Bootstrap

- This is the easiest bootstrap conceptually
- It raises most of the interesting/important points
- BUT it has some problems so we only rarely use it in practice

An overview of the percentile bootstrap, with more detail, can be found at: https://garstats.wordpress.com/2016/05/27/the-percentile-bootstrap/

Two Big Ideas for the Bootstrap

1. The Plug-In Principle
2. Resampling & Sampling with Replacement

The Plug-In Principle

- Wikipedia (2018.02.24):

> In statistics, the plug-in principle is the method of estimation of functionals of a population distribution by evaluating the same functionals at the empirical distribution based on a sample.

"functionals" = fancy math speak for things like the mean, variance, correlations, etc. Basically most of our statistics of interest are functionals or based on them

https://en.wikipedia.org/wiki/Plug-in_principle

The Plug-In Principle

Roughly speaking, the plug-in principle says that a feature of a given distribution can be approximated by the same feature of the empirical distribution of a sample of observations drawn from the given distribution. The feature of the empirical distribution is called a plug-in estimate of the feature of the given distribution. For example, a quantile of a given distribution can be approximated by the analogous quantile of the empirical distribution of a sample of draws from the given distribution.

-- Marco Taboga, economist at the Bank of Italy (link)

link

0.1 The plug-in principle for finding estimators

Under a parametric model $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ (or a non-parametric $\mathcal{P} = \{P_F; F \in \mathcal{F}\}$), any real-valued characteristic τ of a particular member P_θ (or P_F) can be written as a mapping from the parameter-space Θ, i.e. $\tau : \Theta \mapsto \mathbf{R}$. If your observations y comes from P_{θ_0} and you have derived an estimate $\hat{\theta} \in \Theta$ of θ_0 (for example by Maximum-Likelihood), it is natural to use $\tau(\hat{\theta})$ as an estimate of $\tau(\theta_0)$. This method for constructing estimates is commonly reffered to as *the plug-in principle*, since we "plug" the estimate $\hat{\theta}$ into the mapping $\tau(\cdot)$.

All of this is a lot of words to say something so simple that you probably did not even realize it had a name…

Plug-In Principle

Do the exact same calculation on the sample that you would do on the population if you could.

The Plug-In Principle

- The basic idea:
  - In absence of other information about a population, the sample is the best estimate of the population distribution
- This means:
  - The **sample histogram** is an estimator of the population histogram
  - The **sample mean** is an estimator of the population mean
  - The **sample variance** is an estimator of the population variance

And so on.

The Plug-In Principle

- The basic idea:
  - In absence of other information about a population, the sample is the best estimate of the population distribution
- This means:
  - The **sample histogram** is an estimator of the population histogram
  - The **sample mean** is an estimator of the population mean
  - The **sample variance** is an estimator of the population variance

And so on.

Two Big Ideas for the Bootstrap

1. The Plug-In Principle
2. Resampling & Sampling with Replacement

Resampling

- If our sample data (sample histogram) is the best guess we have for the shape of the population, then use it as a stand-in for the population
- Define a probability distribution on the sample data, and use this to simulate many samples from the population
- How do we get many samples from one?

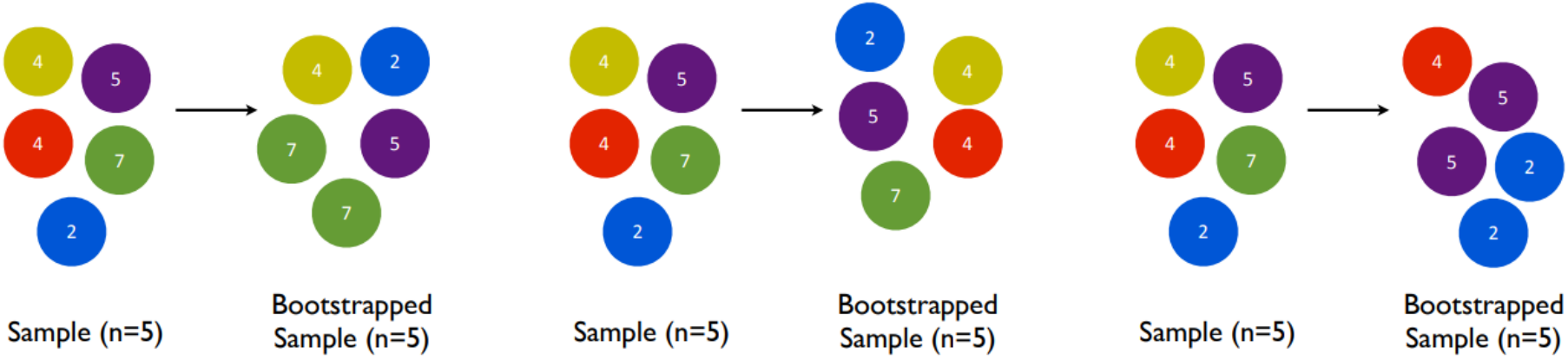
Sampling with Replacement

Throw all of your sample data into a bin, then:

1. Mix the data items thoroughly
2. Draw one item out
3. Write down what the item is
4. **Put the item back into the bin!**
5. Repeat 1-4 until you have a new “sample” the same size as the original

The critical idea here is that we use replacement so that data can come up more than once during the draws

https://web.stanford.edu/class/psych252/tutorials/doBootstrapPrimer.pdf



"Notice that the middle bootstrapped sample reproduces the original sample exactly. Counter to a naive intuition, resampling the original sample exactly is NOT more likely (and in fact, is just as likely) as drawing a "skewed" sample such as the one on the right."

Example & text from: Ong (2014) *A primer to bootstrapping and an overview of doBootstrap*.
Link

The idea in resampling is that the variations that occur due to sampling with replacement represent analogous variations in the population

Aside: Formal Definition

- For a sample of size, n:
  - Assign each sample data point a probability of $\frac{1}{n}$ of being drawn
  - This defines a probability distribution

  - The machinery for this is the ECDF, the **Empirical Cumulative Distribution Function**, but we can neglect the details

So what is the Bootstrap?

- It is a procedure that resamples, with replacement, your original data to make very many different copies of the data
  - For each resample, we compute the statistic of interest
  - These copies can be used to approximate the sampling distribution of the statistic
  - This approximate sampling distribution is called the **bootstrap distribution**
  - The Bootstrap distribution stands in for the true sampling distribution
  - It eliminates the need to assume a mathematical form for this sampling distribution

Aside: Sampling Distributions

- The sampling distribution is the **distribution of the statistic (not the data!)**
- In traditional statistics, you assume that the sampling distribution is known and has an exact mathematical form
- Example: the sampling distribution of the **mean**
  - Center = population mean
  - Spread = standard error of the mean, $\frac{s}{\sqrt{n}}$
  - Shape = normal (due to the central limit theorem)
  - **This last item is an assumption, and technically only valid for large sample sizes**

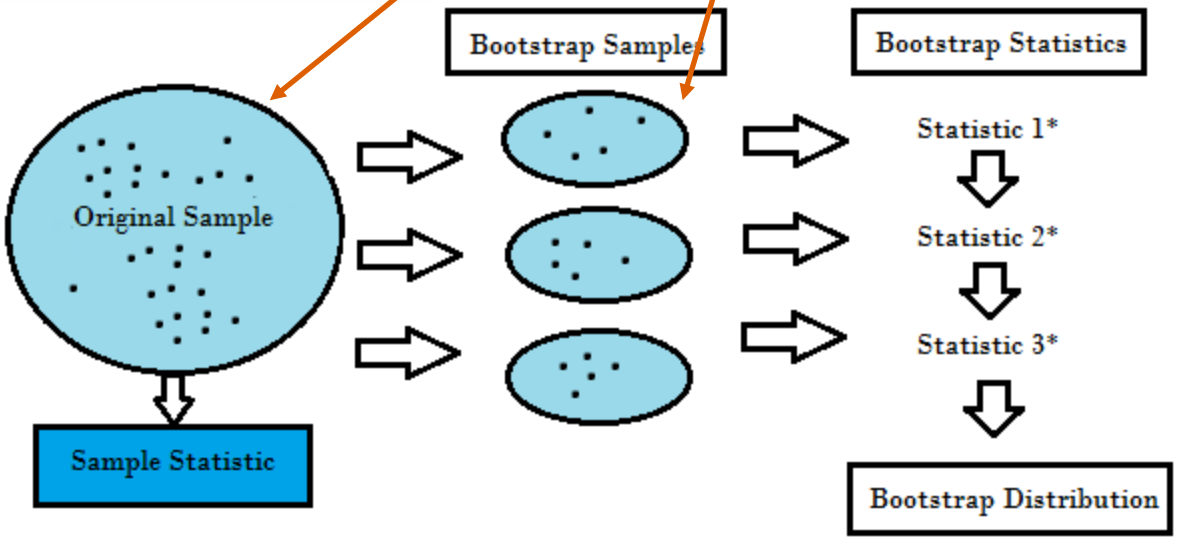
www.arb.ca.gov/research/weekendeffect/CARB041300/

Bootstrap Sampling

The population distribution of any statistic can be approximated by the distribution of that statistic derived from the bootstrap samples

Note: bootstrap samples are the **same size** as the original sample

So imagine more dots here!



www.statisticshowto.com/bootstrap-sample/

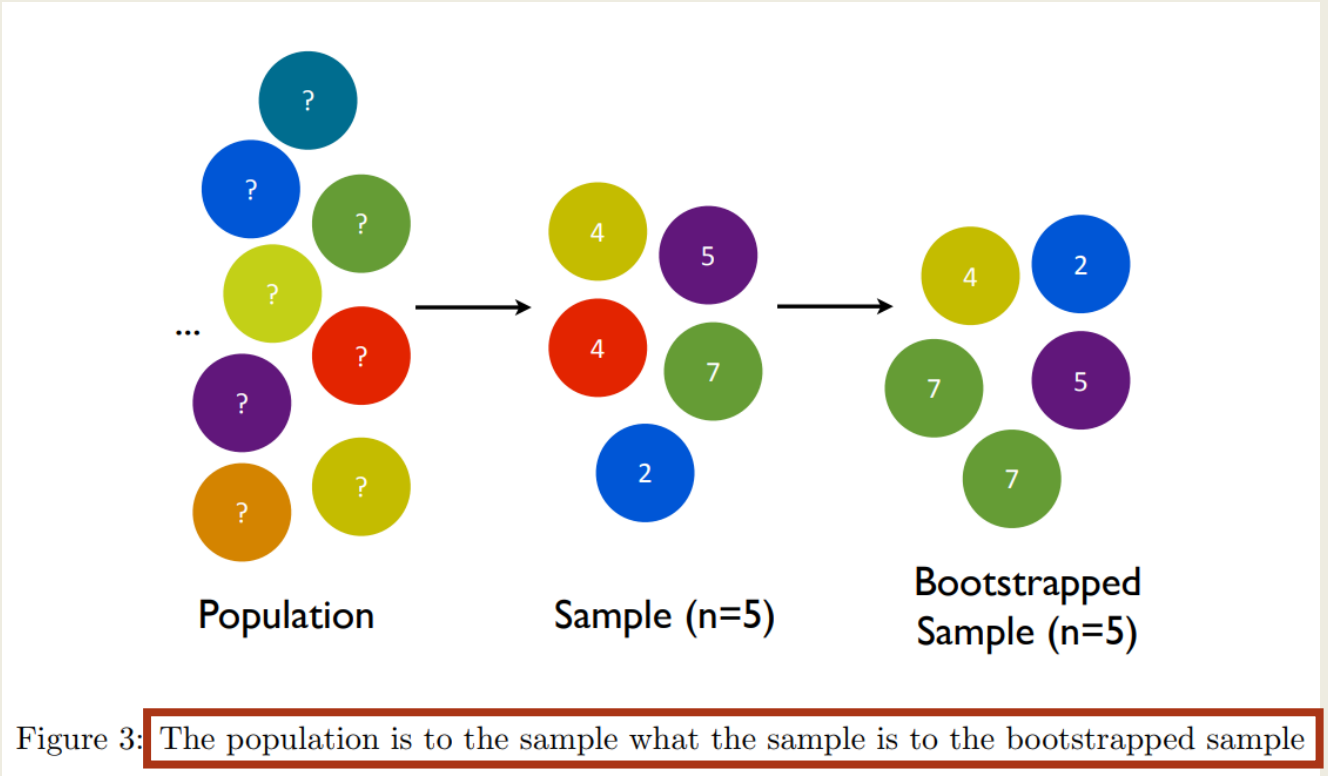


Figure 3: The population is to the sample what the sample is to the bootstrapped sample

Ong (2014) *A primer to bootstrapping and an overview of doBootstrap*.

web.stanford.edu/class/psych252/tutorials/doBootstrapPrimer.pdf

Bootstrap #1: Percentile

Algorithm:

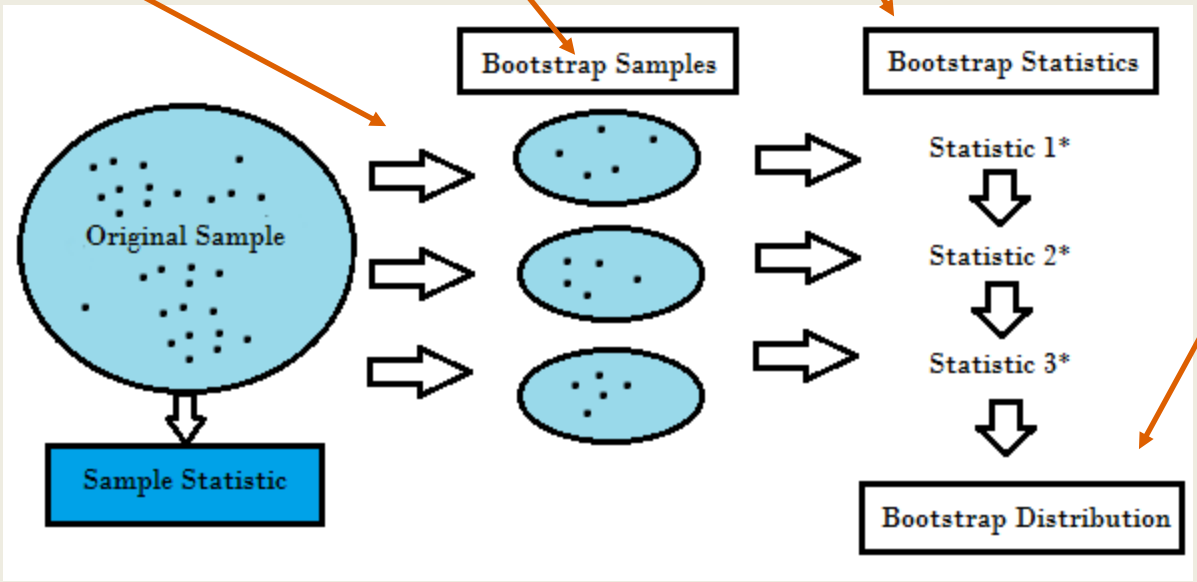
- Select, with replacement, B bootstrap samples of size n, called $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_B^*$
  - B is usually large like 1000, 2000, or 10000
- Compute the statistic of interest for each of these bootstrap samples: $\hat{\theta}_1, \ldots, \hat{\theta}_B$
  - Together these statistics are the bootstrap distribution of $\hat{\theta}$ called $\mathrm{boot}(\hat{\theta})$
- The confidence interval of interest is the interval defined by the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of $\mathrm{boot}(\hat{\theta})$

$\hat{\theta}_1, \ldots, \hat{\theta}_B$

$\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_B^*$

B of these!

$\text{boot}(\hat{\theta})$



Note: bootstrap samples are usually the **same size** as the original sample, so imagine more dots in the bootstrap samples

www.statisticshowto.com/bootstrap-sample/

Let's do it!

- Example data from Wright, London, & Field (2011)
- Section 2 of the R code file:

`bootstrap_CI_Workshop_Main.R`

Journal of Experimental Psychopathology
JEP Volume 2 (2011), Issue 2, 252–270
ISSN 2043-8087 / DOI: 10.5127/jep.013611

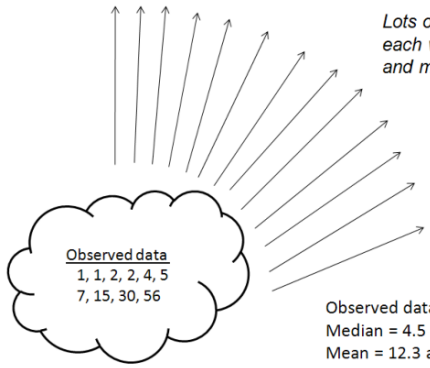
Using Bootstrap Estimation and the Plug-in Principle for Clinical Psychology Data

Daniel B. Wright[a], Kamala London[b], Andy P. Field[c]

[a] Psychology Department, Florida International University

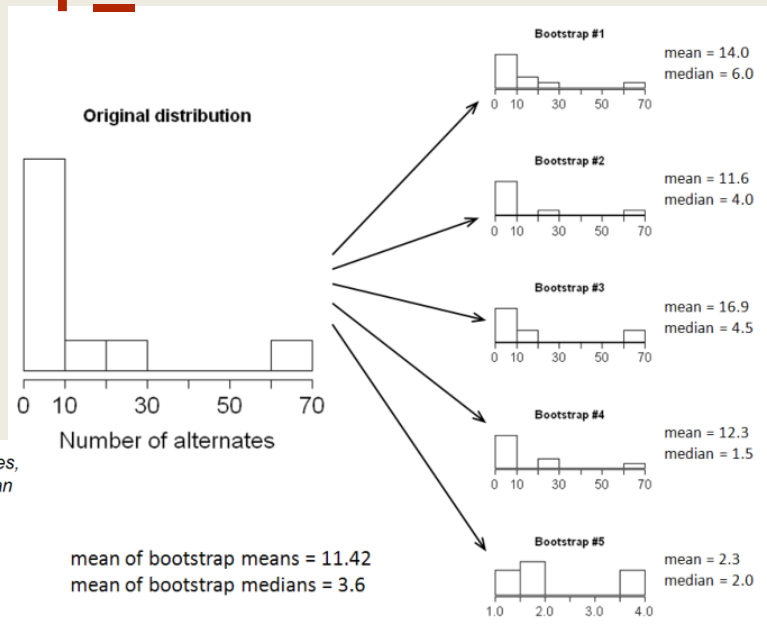
[b] Department of Psychology, University of Toledo

[c] School of Psychology, University of Sussex



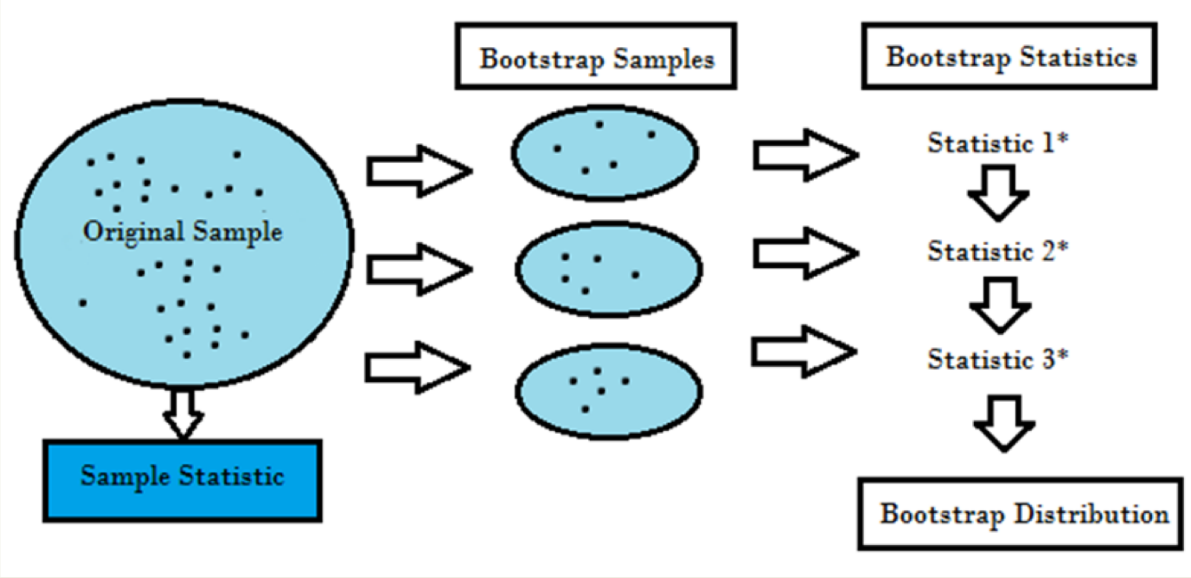
Lots of bootstrap samples, each with their own mean and median.

Observed data
Median = 4.5 alternative personalities
Mean = 12.3 alternative personalities



Pause and Review

- Now is the time to raise issues with the method
  - Does it seem like a magic trick? Why? Why not?
  - Please ask anything you want now



www.statisticshowto.com/bootstrap-sample/

Why we like Bootstrap Percentile CIs

- Simple to make
- Simple to understand (I hope!)
- Better than the *t* confidence interval when the data are somewhat skewed

See: users.stat.umn.edu/~helwig/notes/bootci-Notes.pdf for more details

So what is the problem?

- The percentile bootstrap tends to have “coverage error”
  - Intervals too narrow for small sample sizes
  - More like a z-statistic rather than a t
  - So, for instance, a 90% CI might only provide 80% percent coverage
- Partially corrects for skew in the data
  - But does so in a way that adds random variability
  - When data are symmetric, the percentile bootstrap works reasonably

See: users.stat.umn.edu/~helwig/notes/bootci-Notes.pdf for more details

So what do we do?

- A solution is to do a higher-order bootstrap that adjusts for these problems
- The most common, and generally recommended one is the BC_a – the bias-corrected and accelerated bootstrap
- This introduces two correction terms
  - One for bias and one for skew
  - These additional parameters are automatically estimated using a “jackknife” (*hold-out* / *leave one out* estimation procedure)

See: Chernick and LaBudde (2011) for details, or, for technical explanations, see: blogs.sas.com/content/iml/2017/07/12/bootstrap-bca-interval.html (SAS implementation example)

The Bc_a Bootstrap

- Work well for a wide variety of estimation situations
- "Second-order" accurate, so less influenced by skew
  - That is, less coverage error for many common situations
- Unfortunately, not as intuitive as the percentile bootstrap, but is still similar
- Note: if the data distribution is symmetric, the percentile and BC_a should be close

See: users.stat.umn.edu/~helwig/notes/bootci-Notes.pdf for more details

Further Examples

- Examples in R Script:

`additional_bca_examples.R`

General Advice

- More data = Better bootstraps
  - As the bootstrap assumes the data **is** the population, more data is better
  - Some statisticians have concerns when $n << 100$
  - If n low, data should be "well-behaved" – the same as for normal-theory methods
- Bootstraps work for things like averages
  - "Measures of central tendency" are good
  - Quartiles are ok
  - **Do NOT use bootstraps for things like maxima, minima, or other extremes!**

General Advice

- A good general-purpose nonparametric bootstrap CI to use is the BC_a
  - It deals with moderate bias and skew in the data
  - It works reasonably well at moderate sample sizes
  - An alternative is the ABC interval (not discussed here)
- When you know more about the distribution the data comes from, you can use “parametric” bootstraps to good effect
  - These assume that you know the population distribution’s shape, but not its parameter values
  - See Carpenter & Bithell (2000) and Chernick & LaBudde (2011)

Table III. Summary of properties of bootstrap confidence interval methods. For further details on the categories, see Section 3.8.

Method	Theoretical coverage error	Transformation respecting	Use with parametric simulation	Use with non-parametric simulation	$\hat{\sigma}, \hat{\sigma}^*$ required	Analytic constant or variance stabilizing tranformation required	Use for functions of parameters
Non-Studentized pivotal	$O(n^{-1/2})$	×	✓	✓	×	×	✓
Bootstrap-t	$O(n^{-1})$	×	✓	✓	✓	✓	✓
Percentile	$O(n^{-1/2})$	✓	✓	✓	×	×	✓
BC percentile	$O(n^{-1/2})$	✓	✓	✓	×	×	✓
BCa percentile	$O(n^{-1})$	✓	✓	✓	×	✓	✓
Test-inversion	$O(n^{-1/2})$	✓	✓	×	×	×	×
Studentized test-inversion	$O(n^{-1})$	×	✓	×	✓	×	×

From Carpenter & Bithell (2000) Bootstrap Confidence Intervals: when, which, what? A Practical Guide for Medical Statisticians. *Statistics in Medicine* **19**:1141-1164.

Another bootstrap that deserves mention is the “Bootstrap-t”

Table III. Summary of properties of bootstrap confidence interval methods. For further details on the categories, see Section 3.8.

Method	Theoretical coverage error	Transformation respecting	Use with parametric simulation	Use with non-parametric simulation	$\hat{\sigma}, \hat{\sigma}^*$ required	Analytic constant or variance stabilizing tranformation required	Use for functions of parameters
Non-Studentized pivotal	$O(n^{-1/2})$	×	✓	✓	×	×	✓
Bootstrap-*t*	$O(n^{-1})$	×	✓	✓	✓	✓	✓
Percentile	$O(n^{-1/2})$	✓	✓	✓	×	×	✓
BC percentile	$O(n^{-1/2})$	✓	✓	✓	×	×	✓
BCa percentile	$O(n^{-1})$	✓	✓	✓	×	✓	✓
Test-inversion	$O(n^{-1/2})$	✓	✓	×	×	×	×
Studentized test-inversion	$O(n^{-1})$	×	✓	×	✓	×	×

From Carpenter & Bithell (2000) Bootstrap Confidence Intervals: when, which, what? A Practical Guide for Medical Statisticians. *Statistics in Medicine* **19**:1141-1164.

Bootstrap t

- The bootstrap-t is another bootstrap that attempts to deal with issues raised by the percentile bootstrap
  - Essentially this version makes a T-statistic for the problem
    - Same as the t-statistic, but we do not assume it follows the t-distribution
    - Then it **empirically determines** the distribution of this T and builds an empirical T interval
    - The math involved allows this interval to have properties similar to regular normal-theory t intervals
  - However, it is not transformation respecting so you cannot transform data, use this interval to get the CI, and then undo the transformation on the CI
  - See example: **`bootstrap_t_example.R`**

General Advice

- Regression Problems
  - Bootstrap cases (AKA bootstrap vectors)
    - Resample the cases in the data, build many models
  - Bootstrap residuals
    - Fix the model, resample just the residuals
  - General Advice: if the parametric form is not known, use the bootstrap cases method (Chernick & LaBudde, 2011, p. 58 and elsewhere)
  - See Carpenter & Bithell (2000); and Bootstrapping Regression Models in R, Fox & Weisberg (2017) →

Bootstrapping Regression Models in R

An Appendix to *An R Companion to Applied Regression, Second Edition*

John Fox & Sanford Weisberg

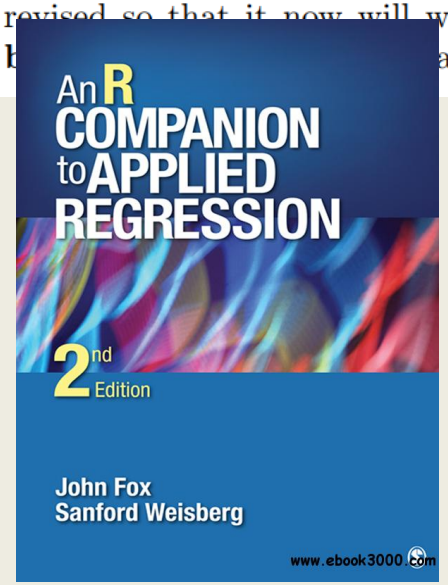
last revision: 10 October 2017

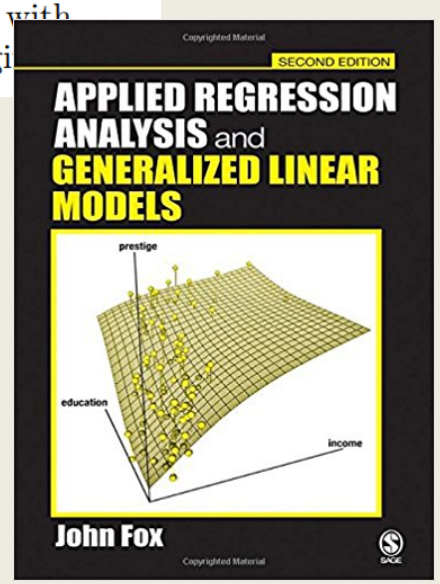
Abstract

The *bootstrap* is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand. This appendix to Fox and Weisberg (2011) briefly describes the rationale for the bootstrap and explains how to bootstrap regression models using the `Boot` function, which was added to the **car** package in 2012, and therefore is not described in Fox and Weisberg (2011). This function provides a simple way to access the power of the `boot` function (lower-case "b") in the **boot** package.

In 2017, the generic `Boot` function was extensively revised so that it now will work with many regression problems. Use of the function with the **b** are gi

The online supplement is usable on its own, but it is connected to two standard stats books for the social sciences





The End