Bootstrapping

Gordon Ross

Estimation

Suppose we have some data which has been sampled from a population. We wish to estimate some property of the population (eg its mean, variance, etc)

We do this by computing an estimate from the observed data which we believe should have a value that is 'close' to the true population value

For example, we might estimate the population mean by using the sample mean.

Example

Suppose we are interested in estimating the average human weight in Scotland. We randomly sample n people from the population. Let their recorded weights (in kilograms) be $y_1, \ldots, y_n$.

Suppose these weights follow a Normal distribution $y_i \sim N(\mu, \sigma^2)$ with **known** $\sigma = 10$.

We use the sample mean $\bar{y}$ as an estimate of the population mean: $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

In practice, we will want to know how good this estimate is – i.e. how close is the sample mean to the population mean?

Confidence Interval

We can quantify this uncertainty via a confidence interval. Lets remind ourselves what these are.

In classical (frequentist) statistics we view the sample mean as being random, in the sense that if we had observed a different sample from the same population, we would get a different sample mean.

Suppose the true (unknown) population mean is $\mu = 65kg$ and we have measured (sampled) the weights of 10 people (so $n = 10$).

For the random sample we collected, suppose the sample mean was $\bar{y} = 62.5$

Sampling Distribution

But the 10 people we measured were sampled randomly from the population, i.e. the weights come from a $N(65, 10^2)$ distribution.

If we had randomly sampled a different 10 people from the same population then we would have got a slightly different sample mean, eg $\bar{y} = 66.7$

In other words, the sample mean we get depends on the random sample we collect. So the sample mean is essentially random.

We can hence talk about the **distribution** of the sample mean, i.e. the distribution of $\bar{\mu}$ if we repeatedly collect samples of $n = 10$ male weights and compute the sample mean.

Confidence Interval

We compute the confidence interval based on the **sampling distribution** of the mean.

A procedure for building 95% confidence intervals will yield an interval that contains the population mean 95% of the time – remember this is **not** the same as saying that the population mean has a 95% of lying in any given 95% confidence interval

If we know σ, then we know (from basic statistics) that the sample mean $\bar{y}$ has a $N(\mu, \sigma^2/n)$ distribution.

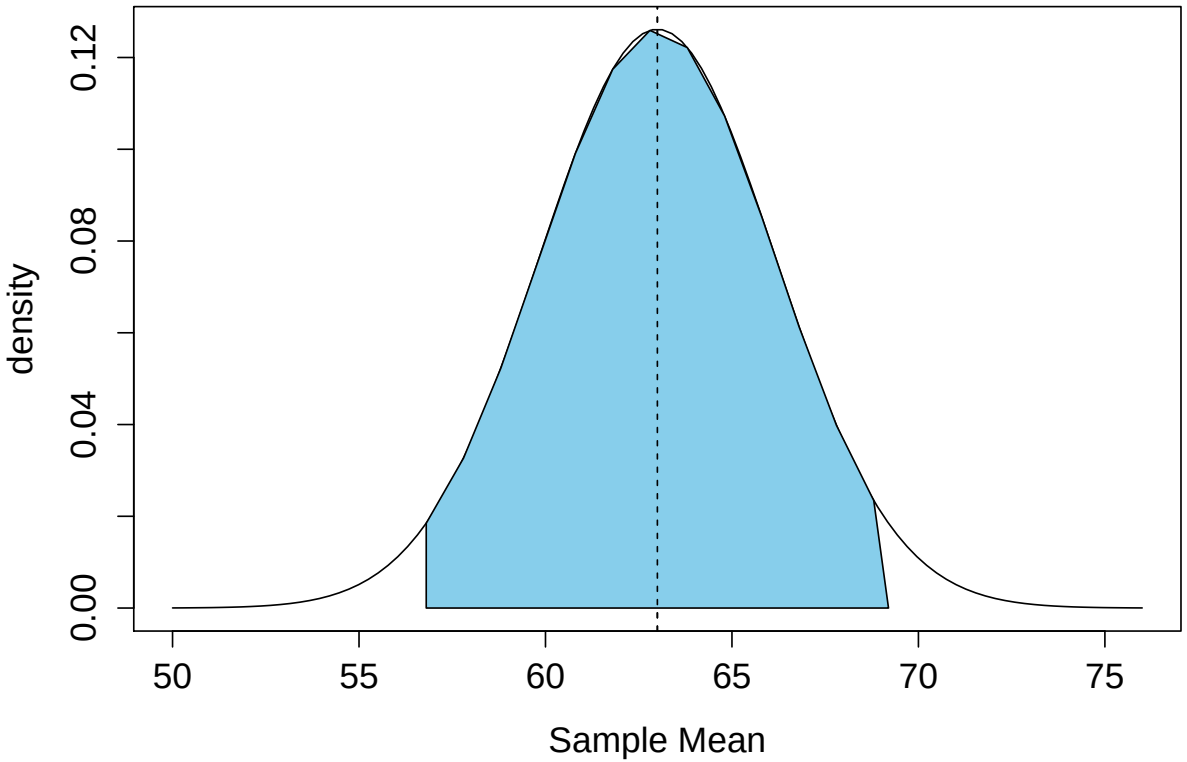
Example

Assume the distribution of weights is $N(65, 10^2)$ with unknown μ and known σ. Suppose we sample 10 people, and find an average weight of 63. We base confidence intervals on the $N(63, 10^2/10) = N(63, 10)$ distribution, which has a standard deviation of sqrt(10) = 3.16.

Confidence intervals are based on the quantiles of this sampling distribution. A 95% interval starts at the 2.5^{th} quantile, and ends at the 97.5^{th} quantile. This contains 0.95 of the total area

We can find these points in R using qnorm(0.025, 63, 3.16) and qnorm(0.0975, 63, 3.16)

Example



Confidence Intervals

So we know how to construct a confidence interval for the mean μ of a Normal distribution when the variance σ^2 is known.

This worked because we knew the sampling distribution of the mean, i.e. we can mathematically show that $\bar{y} \sim N(\mu, \sigma^2/n)$. But what if:

- We want a confidence interval for σ^2
- We want a confidence interval for the mean of some data, but the data is not Normally distributed (and we cannot appeal to the CLT because the sample size is too small)
- We want a confidence interval for ANY summary statistic (e.g. the median) of ANY data, regardless of distribution.

Bootstrapping

Bootstrapping is a very general technique which can be used to construct confidence intervals for essentially any quantity of interest, regardless of the data distribution (which may not even be known!).

Typically it is not practical to do bootstrapping by hand – it requires a computer. But it is very easy to implement on a computer.

Bootstrap estimates are not exact — they are approximations. A 95% confidence interval constructed via bootstraping will typically differ slightly from the true 95% confidencie interval. But they are usually 'close enough' to the exact answer to be useful.

In many (most?) real world situations the exact confidence intervals can't be obtained analytically, so bootstrapping is the only game in town.

Bootstrapping - Fundamental Idea

Recall what a confidence interval actually represents. Statistics like the sample mean are **random** in the sense that they are based on an observed sample from the population distribution – but we ”could” have obtained a different sample and hence a different sample mean.

We have n observations $y_1, \ldots, y_n$ from the population distribution, with sample mean $\bar{y}$. If we drew another sample of size n from the same distribution $y_1^*, \ldots, y_n^*$ we would get a different mean $\bar{y}^*$.

If we imagine many such samples being drawn, the resulting distribution of $\bar{y}^*$ is the **sampling distribution** of the mean.

This argument is the same for alll other quantities (e.g. the sample median/variance) of interest – they are all random.

Bootstrapping - Fundamental Idea

We construct a confidence interval from the **sampling distribution** by finding an interval that contains (e.g.) 0.95 of the area. To do this, we must know the sampling distribution of the quantity of interest! In the Normal case we could derive this analytically for the sample mean. But in general we cannot.

We cannot typically find an expression for (e.g.) the sampling distribution of the median, or the sampling distribution of the mean/variance of most non-Normal distributions (without appeal to the Central Limit Theorem for large data sets). Also if we do not know the data distribution, we cannot find sampling distributions at all.

The basic idea of bootstrapping is to approximate the unknown sampling distribution, and use this approximation for constructing confidence intervals.

Bootstrapping - Fundamental Idea

The bootstrap approach **estimates** the sampling distribution using nothing other than the available data.

The name bootstrapping comes from the saying ”to pull yourself up by your own bootstraps” – to essentially get something from nothing.

We construct an approximation to the true (and perhaps unknown) sampling distribution, and then find confidence intervals based on the quantiles of this approximate distribution – e.g. the 2.5^{th} and 97.5^{th} quantiles for a 95% interval.

The Nonparametric Bootstrap

The most general bootstrap is the nonparametric bootstrap. Let $y_1, \ldots, y_n$ be the observations from some distribution (known or unknown). We wish to construct a confidence interval for some quantity of this distribution - e.g. the mean, median, variance, etc.

Basic idea: if we could draw more samples $y_1^*, \ldots, y_n^*$ from the population distribution, we could find the sampling distribution. But we only have the observed sample!

So instead we construct many 'pseudo-samples' by **resampling** the observed data. This allows us to build an approximation to the sampling distribution of the quantity of interest

Bootstrapping -"Pseudo-Samples"

Example: suppose we have 5 observations 2.4, 3.2, 2.2, 4.6, 3.3. The sample mean is 3.14, and we want a 95% confidence interval.

We draw B pseudo samples each of size $n = 5$ by assigning each observation an equal probability of being selected (1/5), and repeatedly drawing samples of 5 observations. So we might get (e.g.):

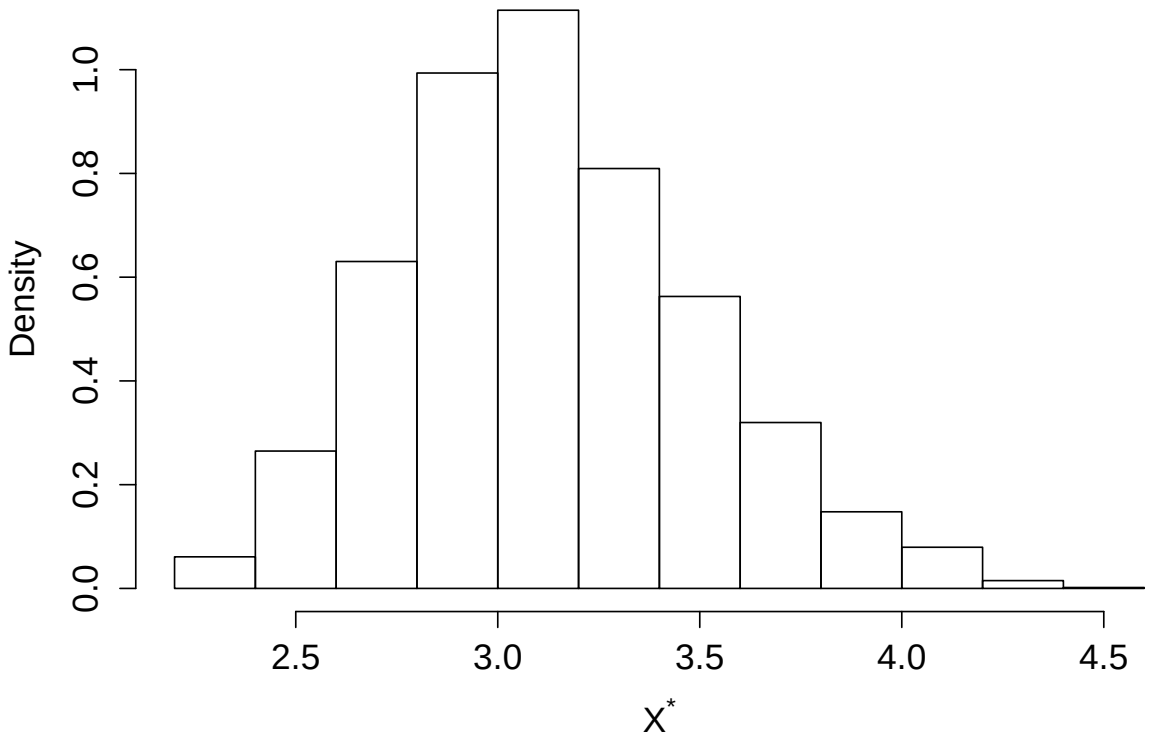
- Pseudo-sample 1: 3.3, 3.2, 2.4, 4.6, 2.2, $\quad \bar{y}_1^* = 3.14$
- Pseudo-sample 2: 2.2, 2.2, 2.4, 3.2, 2.4, $\quad \bar{y}_2^* = 2.48$
- ...
- Pseudo-sample B: 3.2, 4.6, 2.4, 4.6, 4.6, $\quad \bar{y}_B^* = 3.88$

Sampling is done **with replacement** so each observation can be selected multiple times in each pseudo-sample

Choose B to be large, e.g $B = 10000$ (exact value does not matter)

Bootstrapping - Resulting Distribution

This gives us the bootstrap distribution of $\bar{y}^*$:



Note: this distribution is necessarily discrete since there are only finitely many different pseudo-samples we could create!

Bootstrapping - Constructing Confidence Intervalls

We now have the samples means $\bar{y}_1^*, \bar{y}_2^*, \ldots, \bar{y}_B^*$, each one corresponding to one pseudo-sample of size n ($n = 5$ in this case)

We can then construct a confidence interval for the mean directly – for a 95% confidence interval, we take the 2.5^{th} and 97.5^{th} quantiles of the bootstrap estimate of the sampling distribution.

Eg if $B = 10000$ then sort the $\bar{y}_i^*$'s in order from smallest to largest, then the $10000 * 0.025 = 250^{th}$ smallest one would be the bottom point of the 95% confidence interval, and the $10000 * 0.975 = 9750^{th}$ smallest would be the upper point

In this example we end up with a 95% confidence interval of $[2.64, 3.64]$.

Bootstrapping - Rcode

Example R code to implement this:

```
y <-c(2.4, 3.2, 2.2, 4.6, 3.3)
sims <- 10000
means <- numeric(sims)
for (s in 1:sims) {
  bs <- sample(y,n,replace=TRUE)
  means[s] <- mean(bs)
}
means <- sort(means) #vital step!!!
means[0.025*sims]
means[0.975*sims]
```

Bootstrapping - Accuracy

The bootstrap **approximates** the sampling distribution using the distribution of the statistics obtained on the pseudo-samples. As such, **bootstrap confidence intervals are not exact**.

However for small samples such as $n = 10$ the bootstrap distribution is close to the true sampling distribution, and so the bootstrap confidence intervals are close to the true confidence intervals.

Note: the accuracy also depends on the number of pseudo-samples B. This should be chosen to be quite large, say $B > 1000$. The exact number doesn't matter, and computers can usually do bootstrapping fast.

Increasing B beyond a certain point (say $B = 1000$ or so) **does not improve the accuracy of the bootstrap interval**. The deviation of the bootstrap interval from the true confidence interval is inherent in using an approximation – we cannot remove it by just taking more bootstrap samples! However using too few will add to the error.

Bootstrapping - Example

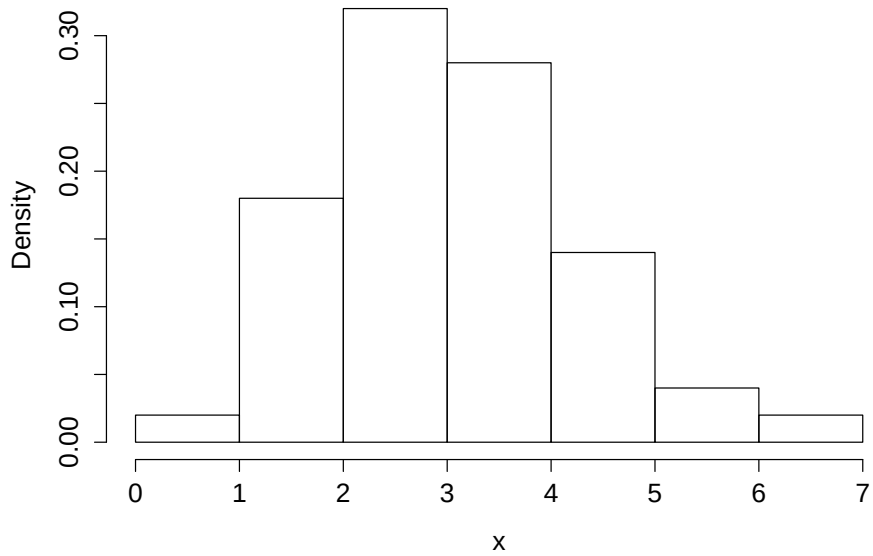
Bootstrapping key point: **We do not need to know the underlying distribution of the data**. We can also use it to compute confidence intervals for quantities other than the mean.

Example: A company that manufactures bottled water wishes to test how much chlorine is present in the bottles made by one of its factories. It selects 100 bottles at random, and measures the chlorine content (measurements in mg/L)

It is interested in the mean, median, and standard deviation of chlorine content. A 95% confidence interval for each is desired.

Bootstrapping - Example

The data has the following distribution - clearly not Normal due to the skew.



Sample statistics (based on 100 observations): Mean = 3.17, Median = 2.88, Standard Deviation = 1.47

Bootstrapping - Example

Example R code to implement this:

```
x <- #type data in here
sims <- 10000
means <- numeric(sims)
medians <- numeric(sims)
sds <- numeric(sims)
for (s in 1:sims) {
  bs <- sample(x,n,replace=TRUE)
  means[s] <- mean(bs)
  medians[s] <- median(bs)
  sds[s] <- sd(bs)
}
```

Bootstrapping - Example

```
means <- sort(means)
medians <- sort(medians)
sds <- sort(ads)

c(means[0.025*sims], means[0.975*sims])
c(medians[0.025*sims], medians[0.975*sims])
c(sds[0.025*sims], sds[0.975*sims])
```

Bootstrapping - Example 2

This gives the following 95% intervals:

- Mean: $[3.39, 3.14]$
- Median: $[2.99, 3.24]$
- Standard Deviation: $[1.16, 1.77]$

Note – in practice we cannot compare bootstrap intervals to the true intervals, because we do not know the true intervals. That is the entire point of using the bootstrap!

We trust bootstrap estimates because they tend to perform well in situations where we do know the truth (see this week's lab)

Mathematical theory can make more precise statements about the accuracy of bootstrapping, but we will not explore this here.

Bootstrapping - Example 2

We can even use bootstrapping to find confidence intervals for probability forecasts.

Suppose our data is the past history of time-between-earthquakes on some fault. For example, we have collected $n = 10$ data points (years between earthquakes):

$$7.6, 11.8, 1.5, 1.4, 4.4, 28.9, 12.3, 5.4, 9.6, 1.5$$

We assume these follow an Exponential distribution and would like to predict the time of the next earthquake, based on how much time typically passes between earthquakes.

Bootstrapping - Example 2

So the data is $y_i \sim Exponential(\lambda)$. If we knew the true value of λ then we could answer questions like "what is the probability of the next earthquake occuring within the next 10 years?" as:

$$\int_{-\infty}^{10} \lambda e^{-\lambda y} dy$$

However in practice we dont know λ so we must estimate. It is easy to show that the MLE of λ is given by

$$\hat{\lambda} = 1/\bar{y} = 0.12$$

Bootstrapping - Example 2

So using this estimate, the probability of the next earthquake occurring 10 years is estimated as:

$$\int_{-\infty}^{10} \hat{\lambda} e^{-\hat{\lambda} x} dx = \text{pexp}(10, 0.12) = 0.70$$

But we know that our estimate of λ will not be equal to the true value. So to incorporate this uncertainty, we can use bootstrapping.

Bootstrapping - Example 2

```
y<- c(7.6, 11.8, 1.5, 1.4, 4.4, 28.9, 12.3, 5.4, 9.6, 1.5)
sims <- 10000
lambdas <- numeric(sims)
forecasts <- numeric(sims)
for (s in 1:sims) {
  bs <- sample(y,n,replace=TRUE)
  lambdas[s] <- 1/mean(bs)
  forecasts[s] <- pexp(10, 1/mean(bs))
}
lambdas <- sort(lambdas)
forecasts <- sort(forecasts)
```

Bootstrapping - Example 3

Our 95% confidence interval for λ is $[0.07, 0.23]$

Our 95% confidence interval for the probability of the next earthquake happening within 10 years is: $[0.52, 0.90]$

So "Our confidence interval for the probability of an earthquake happening within the next 10 years is $[0.52, 0.90]$" – this sounds a bit funny! But remember what confidence intervals mean. The data we observed was a sample from a population. We originally gave a 0.7 probability to an earthquake occurring within 10 years, but if we had observed different data from the same population distribution, we would have given a different prediction. Our prediction itself is random (i.e. it is dependent on the random data we observed).

Important Point

Every prediction you make has some degree of randomness/uncertainty in this sense – so rather than giving a point forecast ("this event has an x% chance of happening") you should always give some measure of the uncertainty ("a 95% confidence interval for the chance of this event happening is $[a, b]$").

Without this, you have no way of knowing how much confidence to have in the forecast – i.e. how accurate it is.

But when the media reports on scientific predictions – and sometimes even when scientists discuss things – how often do you see uncertainty being reported?

Society (and science) has a general problem with misplaced certainty in forecasts, due to not expressing (or measuring) the inherent uncertainty in the forecasting procedure.

Bootstrapping - Summary

Bootstrapping is a very useful too which allows us to construct confidence intervals for essentially any quantity we like, regardless of whether we know the true data distribution. Examples:

- The variance parameter σ^2 of the Normal distribution
- The λ parameter of the Exponential distribution
- The median (not mean!) of a random process
- The probability of an earthquake occurring within the next 10 years

However the confidence intervals we find using bootstrapping are only **approximate** and may deviate from the true intervals, particularly in small samples