Sampling variability

Data Science Team
What we have learned so far

- Often the data is a sample from a population and we want to use it to learn something about this bigger population.
- A summary of the data is a reasonable estimate for the corresponding summary of the population (as the sample size grows bigger, the sample summary becomes very close to the population value).
- When the sample size is limited, there is a considerable variability of sample summaries.
- We can use the Bootstrap to learn about the variability of sample summaries using one sample only.
Looking back on the examples we looked at

We have not commented on this before, but a number of the histograms of the estimates derived by multiple samples look like Normal distributions
As the sample size increases

Looking at the problem of estimating the mean for $U[50,100]$

- We have noted that as the sample size increases the variability of the sample estimates decreases
One more pattern

We estimate the mean of Uniform distributions centered at 75, but with different spread (different variance, see summary-lab02.Rmd)

As the population variance decreases, the sample variability decreases
Sample average

- These are not just coincidences
- In all these examples, the sample summary we looked at was an average

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

(note that the sample proportion is just an average of 0 or 1)
- There are a few things we can prove mathematically about the averages and that relate their variability to that of the population from which the sample comes from
Sample average

- We have a population with mean $\mu$ and variance $\sigma^2$
- Let $(X_1, \ldots, X_n)$ a random sample of size $n$ from that population
- Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the sample mean
- Then, the average value of $\bar{X}$ across all the possible samples we might take is equal to $\mu$
- and the variance of $\bar{X}$ across all the possible samples is equal to $\sigma^2/n$.
- For a sample size $n$ “big enough,” the histogram of $\bar{X}$ across all possible samples has the shape of a Normal distribution

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

NOTE: “all possible samples” represent our thought experiments and is an abstract population
An important characteristic

Normal distribution

95% of values

Mean

Mean−2xSD

Mean+2xSD

0.0

−4 −2 0 2 4
t

Density

Normal distribution

▶ mean and variance tell you everything
The Normal distribution

- Taking sums of many independent quantities we obtain a normal distribution (this is known at the Central Limit Theorem)
- This can be used as a definition of what the normal distribution is and where it comes from
- Ex. measurements errors are expected to be the sum of many independent processes we do not control and have a Normal distribution
- It is also known as Gaussian, as Gauss derives it in *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (1809) (a fairly impressive work)
The central limit theorem in action: Quincunx

- Obstacle arranged in a Quincunx pattern (used to plant trees)
- Each ball dropped can go right or left (+1 or -1) with equal chance, independently of which path it took before
Sampling variance I: the bootstrap

Sample: \((X_1, \ldots, X_n)\);

Sample summary: \(S(X_1, \ldots, X_n)\) which we use to estimate the corresponding value in the population \(S_{\text{pop}}\).

- Resample with replacement to obtain bootstrap samples \((X_1^b, \ldots, X_n^b), \ b = 1, \ldots, B\)
- Calculate the summaries of the bootstrap samples \(S_b = S(X_1^b, \ldots, X_n^b), \ b = 1, \ldots, B\)
- The variance of the bootstrap samples is an estimate of the variance of \(S(X_1, \ldots, X_n)\) across all possible samples

\[
\text{Average}((S(X_1, \ldots, X_n) - S_{\text{pop}})^2) \approx \sum_{b=1}^{B} \frac{(S_b - \bar{S})^2}{B}
\]

\[
\text{Standard Error}(S(X_1, \ldots, X_n)) \approx \sqrt{\sum_{b=1}^{B} \frac{(S_b - \bar{S})^2}{B}}
\]
Sampling variance II:

Sample: \((X_1, \ldots, X_n)\);

Sample summary: \(S(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i / n\), which we use to estimate the corresponding value in the population \(S_{pop} = \mu\).

- The variance of the sample gives us a way to estimate of the variance of \(S(X_1, \ldots, X_n)\) across all possible samples

\[
\text{Average}((S(X_1, \ldots, X_n) - S_{pop})^2) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}
\]

\[
\text{Standard Error}(S(X_1, \ldots, X_n)) \approx \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}}
\]

\[
= \frac{1}{\sqrt{n}} \text{Sample Standard Deviation}
\]
How do we use all of this?

![Graph showing DNA variant propensity to use left hand versus Claridge data.](image)

![Histogram showing variability of sample correlation.](image)
Report both the estimate and its standard error

**Standard error**: square root of the variance of our estimate across all the possible samples (thought experiments)

- we can estimate the standard error via Bootstrap or using the sample standard deviation and dividing it by the square root of the sample size.
- For the claridge data, we can only use Bootstrap (summary is not an average)

<table>
<thead>
<tr>
<th></th>
<th>Sample Correlation</th>
<th>Standard Error (Bootstrap)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5087758</td>
<td>0.2076155</td>
</tr>
</tbody>
</table>
Interval estimation

- The idea is to report not one number estimate, but a range of possible values that you expect to cover the true value of the population summary.
- There are multiple ways of coming up with these intervals and a proper justification of why they work and in which sense is beyond what we can cover here.
- Two rules that we can write out and guarantee that, in repeated experiments, the intervals cover the true value of the population summary 95% of the times.

For Averages

\[
(\bar{X} - 2SD(X_1, \ldots, X_n)/\sqrt{n}, \bar{X} + 2SD(X_1, \ldots, X_n)/\sqrt{n})
\]

Bootstrap

\[
(\text{Quantile}_{BootSample}(0.025), \text{Quantile}_{BootSample}(0.975))
\]