Bootstrap Confidence Intervals

Sahir Bhatnagar and James Hanley

EPIB 607
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/

September 26, 2018

Review of Confidence Intervals

Sampling Distribution

Definition 1 (Sampling Distribution)

- *The sampling distribution of a statistic is the distribution of values taken by the statistic in* ***all possible samples of the same size*** *from the same population.*
- *The standard deviation of a sampling distribution is called a* ***standard error***

Sampling Distributions

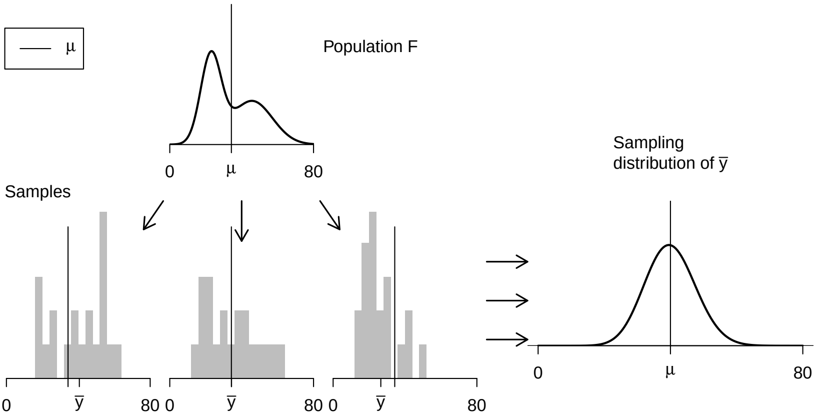


Fig.: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Sampling Distribution

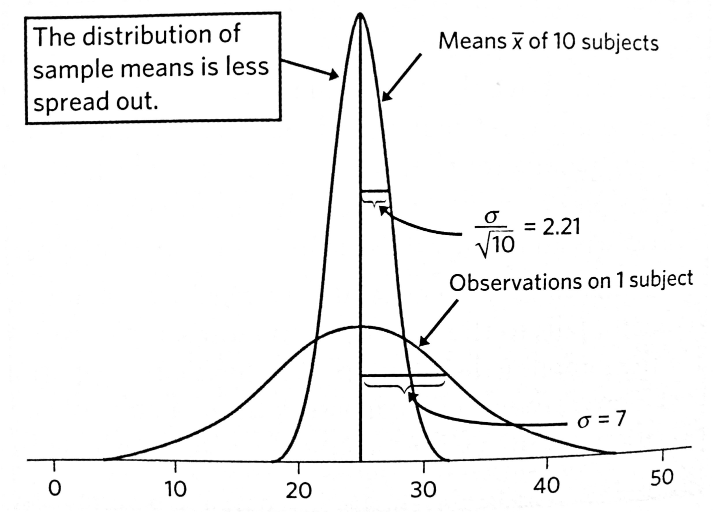


Fig.: Averages are less variable than individual observations

Traditional way to calculate CIs

How to construct a CI for the population mean?

- The **CLT** gives us that $\bar{y} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$ is approximately true when n is large.

Traditional way to calculate CIs

How to construct a CI for the population mean?

- The **CLT** gives us that $\bar{y} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$ is approximately true when n is large.
- We can standardize, to get $Z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$.

Traditional way to calculate CIs

How to construct a CI for the population mean?

- The **CLT** gives us that $\bar{y} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$ is approximately true when n is large.
- We can standardize, to get $Z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.
- To find a CI with confidence level $\mathcal{C} = 1-\alpha$, we must calculate the critical value z^* such that

$$P(-z^* < Z < z^*) = \mathcal{C} = 1 - \alpha \tag{1}$$

where α is the significance level

Traditional way to calculate CIs

How to construct a CI for the population mean?

- The **CLT** gives us that $\bar{y} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$ is approximately true when n is large.
- We can standardize, to get $Z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.
- To find a CI with confidence level $\mathcal{C} = 1 - \alpha$, we must calculate the critical value z^* such that

$$P(-z^* < Z < z^*) = \mathcal{C} = 1 - \alpha \tag{1}$$

  where α is the significance level
  - That is, we want the value z^* that gives a *lower tail probability* of $(1 - \mathcal{C})/2 = \alpha/2$.

Traditional way to calculate CIs

How to construct a CI for the population mean?

- The **CLT** gives us that $\bar{y} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$ is approximately true when n is large.
- We can standardize, to get $Z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.
- To find a CI with confidence level $\mathcal{C} = 1 - \alpha$, we must calculate the critical value z^* such that

$$P(-z^* < Z < z^*) = \mathcal{C} = 1 - \alpha \tag{1}$$

  where α is the significance level
  - That is, we want the value z^* that gives a *lower tail probability* of $(1-\mathcal{C})/2 = \alpha/2$.
  - Often this value is denoted $z^* = z_{\alpha/2}$; thus we have

$$P(Z < -z_{\alpha/2}) = \alpha/2,$$

    and

$$P(Z > z_{\alpha/2}) = \alpha/2.$$

Traditional way to calculate CIs

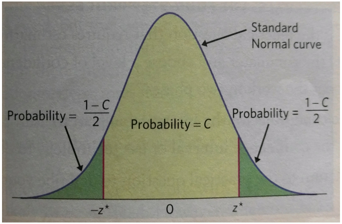


Fig.: The critical value $z^\star$ is the number that catches central probability $\mathcal{C}$ under a standard normal $\mathcal{N}(0,1)$ curve between $-z^\star$ and $z^\star$

We can use this probability statement about the standardized version of the sample mean $(\bar{y}-\mu)/\sigma/\sqrt{n}$, to place bounds on where we think the true mean lies by examining the probability that $\bar{y}$ is within $z^\star \cdot \frac{\sigma}{\sqrt{n}}$ of μ

$$
\begin{aligned}
\mathcal{C} &= P\left(-z^\star \leq \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \leq z^\star\right) \\
&= P\left(-z^\star \frac{\sigma}{\sqrt{n}} \leq \bar{y}-\mu \leq +z^\star \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(-\bar{y}-z^\star \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{y}+z^\star \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{y}+z^\star \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{y}-z^\star \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{y}-z^\star \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{y}+z^\star \frac{\sigma}{\sqrt{n}}\right) \\
&= 1-\alpha
\end{aligned}
$$

We call the interval $\left(\bar{y}-z^\star \frac{\sigma}{\sqrt{n}}, \bar{y}+z^\star \frac{\sigma}{\sqrt{n}}\right)$ a **(1-α)100% confidence interval** for μ.

Confidence intervals for depths of the ocean

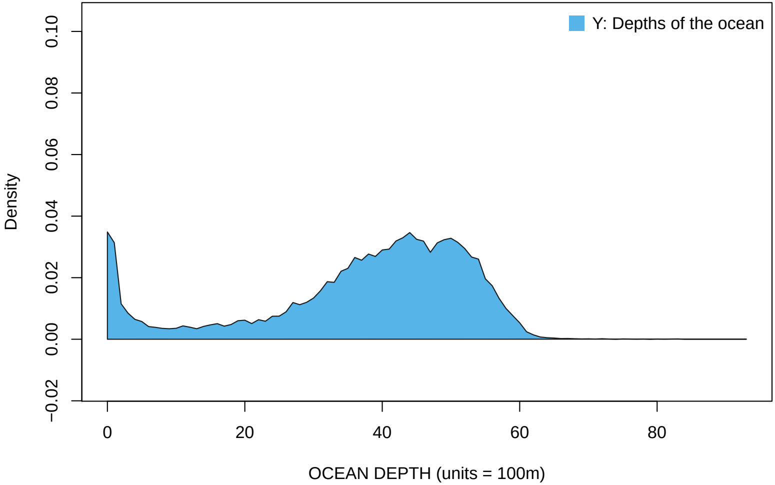


Fig.: The original data distribution of sampled depths of the ocean. Note that it has multiple modes and not Normal looking.

The CLT is 'kicking in' at $n = 16$

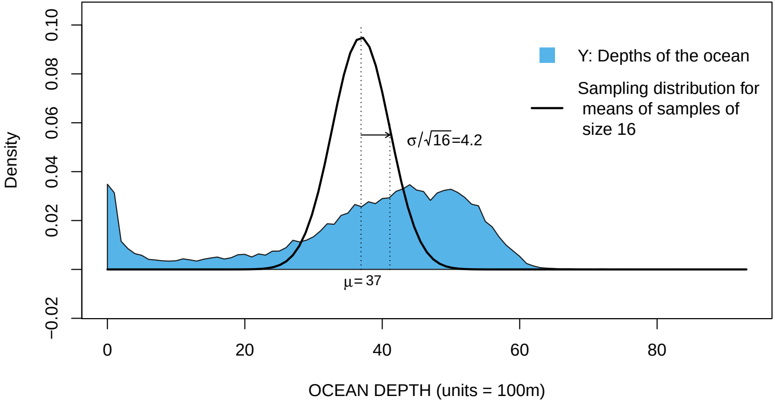


Fig.: The sampling distribution for the mean depth of the ocean with samples of size $n = 16$, looks normal (centered at $\mu = 37$ and SD equal to $\sigma/\sqrt{16}$)

Since CLT has 'kicked in', we use it to construct a CI

We want to construct a $\mathcal{C} = 95\%$ confidence interval for the mean.
Level of significance is $\alpha = 1 - \mathcal{C} = 0.05$

Since CLT has 'kicked in', we use it to construct a CI

We want to construct a $\mathcal{C} = 95\%$ confidence interval for the mean.
Level of significance is $\alpha = 1 - \mathcal{C} = 0.05$

1. by the CLT $\rightarrow \bar{y} \sim \mathcal{N}(mean = 37, sd = \sigma/\sqrt{16} = 4.2)$

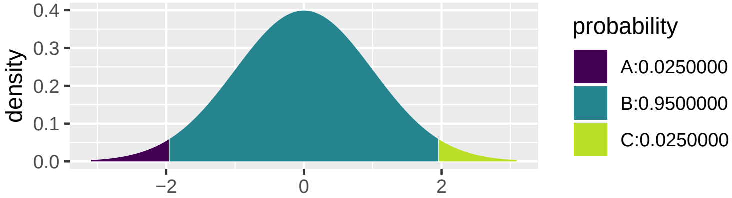
Since CLT has 'kicked in', we use it to construct a CI

We want to construct a $\mathcal{C} = 95\%$ confidence interval for the mean.
Level of significance is $\alpha = 1 - \mathcal{C} = 0.05$

1. by the CLT $\rightarrow \bar{y} \sim \mathcal{N}(mean = 37, sd = \sigma/\sqrt{16} = 4.2)$

2. The critical value $z^\star$ such that
$P(Z < -z^\star) = P(Z > z^\star) = \alpha/2 = 0.025$ is given by

```
mosaic::xqnorm(p = c(0.025, 0.975))
```




```
## [1] -1.959964  1.959964
```

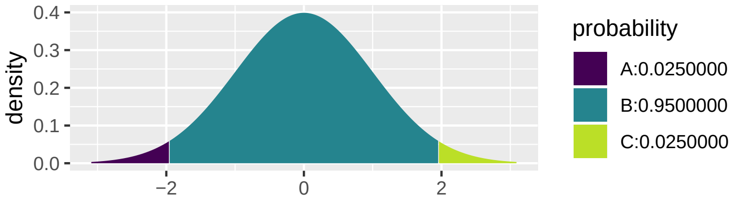
Since CLT has 'kicked in', we use it to construct a CI

We want to construct a $\mathcal{C} = 95\%$ confidence interval for the mean.
Level of significance is $\alpha = 1 - \mathcal{C} = 0.05$

1. by the CLT $\rightarrow \bar{y} \sim \mathcal{N}(mean = 37, sd = \sigma/\sqrt{16} = 4.2)$

2. The critical value $z^\star$ such that
   $P(Z < -z^\star) = P(Z > z^\star) = \alpha/2 = 0.025$ is given by

```
mosaic::xqnorm(p = c(0.025, 0.975))
```




```
## [1] -1.959964  1.959964
```

3. 95% CI for μ: $(37 - 1.96 \cdot 4.2, 37 + 1.96 \cdot 4.2) = [29, 45]$

Alternative way of calculating CI with CLT: `qnorm`

- In the previous slides we used the standard normal $\mathcal{N}(0,1)$ to calculate the critical value $z^\star$ needed for the CI
- We were able to use the $\mathcal{N}(0,1)$ for two reasons:

Alternative way of calculating CI with CLT: `qnorm`

- In the previous slides we used the standard normal $\mathcal{N}(0,1)$ to calculate the critical value $z^\star$ needed for the CI
- We were able to use the $\mathcal{N}(0,1)$ for two reasons:
  1. the CLT

Alternative way of calculating CI with CLT: `qnorm`

- In the previous slides we used the standard normal $\mathcal{N}(0, 1)$ to calculate the critical value $z^\star$ needed for the CI
- We were able to use the $\mathcal{N}(0, 1)$ for two reasons:
  1. the CLT
  2. the formula used to calculate the CI is based on standardizing $\bar{y} \rightarrow \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$

Alternative way of calculating CI with CLT: `qnorm`

- In the previous slides we used the standard normal $\mathcal{N}(0,1)$ to calculate the critical value $z^\star$ needed for the CI
- We were able to use the $\mathcal{N}(0,1)$ for two reasons:
  1. the CLT
  2. the formula used to calculate the CI is based on standardizing $\bar{y} \to \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$
- There is an alternative, **yet equivalent**, way to calculate the CI without standardizing $\bar{y}$, and without using the $\pm$ formula

Alternative way of calculating CI with CLT: `qnorm`

- In the previous slides we used the standard normal $\mathcal{N}(0,1)$ to calculate the critical value $z^\star$ needed for the CI

- We were able to use the $\mathcal{N}(0,1)$ for two reasons:
  1. the CLT
  2. the formula used to calculate the CI is based on standardizing $\bar{y} \to \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$

- There is an alternative, **yet equivalent**, way to calculate the CI without standardizing $\bar{y}$, and without using the $\pm$ formula

- This is accomplished using `qnorm`

Alternative way of calculating CI with CLT: `qnorm`

- In the previous slides we used the standard normal $\mathcal{N}(0, 1)$ to calculate the critical value $z^\star$ needed for the CI
- We were able to use the $\mathcal{N}(0, 1)$ for two reasons:
  1. the CLT
  2. the formula used to calculate the CI is based on standardizing $\bar{y} \rightarrow \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$
- There is an alternative, **yet equivalent**, way to calculate the CI without standardizing $\bar{y}$, and without using the $\pm$ formula
- This is accomplished using `qnorm`
- Note: we **still need the CLT regardless** of whether we use the $\pm$ formula or `qnorm`

68% Confidence interval using `qnorm`

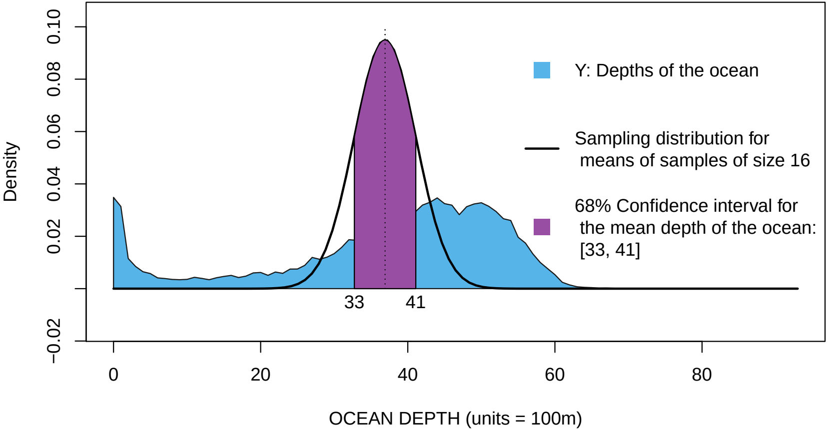


Fig.: 68% Confidence interval calculated using
`qnorm(p = c(0.16,0.84), mean = 37, sd = 4.2)`

95% Confidence interval using `qnorm`

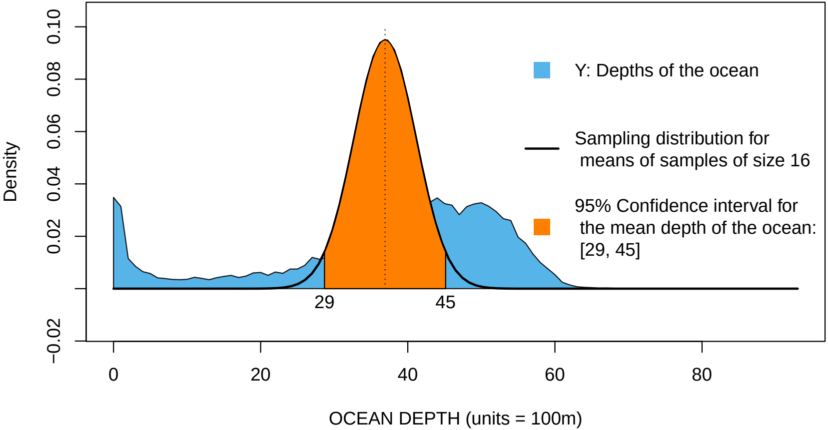


Fig.: 95% Confidence interval calculated using
`qnorm(p = c(0.025,0.975), mean = 37, sd = 4.2)`

Motivation for the Bootstrap

- The $\pm$ and `qnorm` methods to calculate a CI both require the CLT

Motivation for the Bootstrap

- The $\pm$ and `qnorm` methods to calculate a CI both require the CLT

Q: What happens if the CLT hasn't 'kicked in'?
Or you don't believe the CLT?

Motivation for the Bootstrap

- The $\pm$ and `qnorm` methods to calculate a CI both require the CLT

Q: What happens if the CLT hasn't 'kicked in'?
Or you don't believe the CLT?

A: Bootstrap

The Bootstrap

Ideal world: known sampling distribution

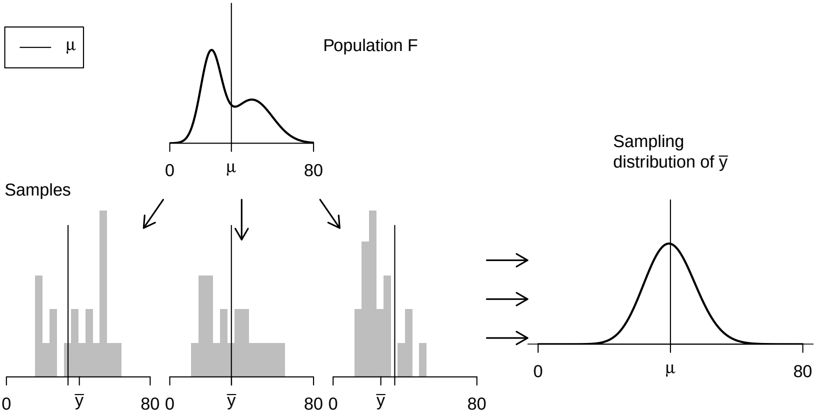


Fig.: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Reality: use the bootstrap distribution instead

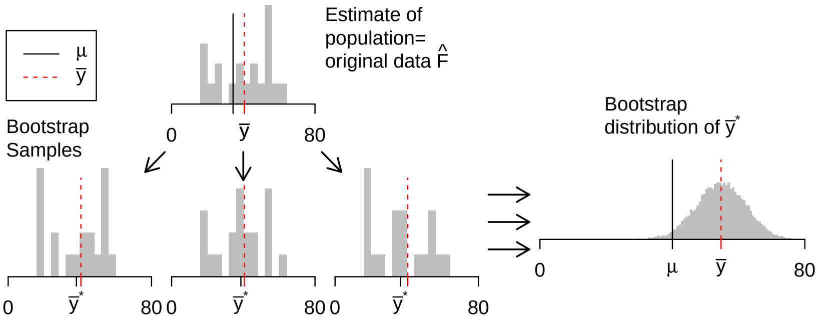


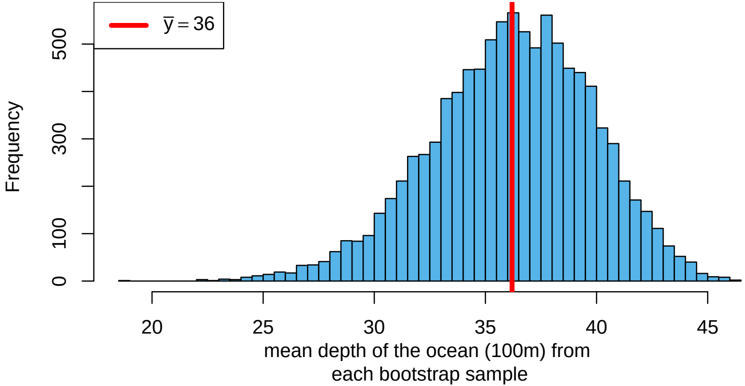
Fig.: Bootstrap world. The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{y}$), not the parameter (μ).

Main idea: simulate your own sampling distribution

```
mean(~ alt, data = depths.n.20)
```

```
## [1] 36.2
```

```
library(mosaic)
s_dist <- do(10000) * mean( ~ alt, data = resample(depths.n.20))
```



Main idea: simulate your own sampling distribution