Quantitative Sociological Analysis

Inferential Statistics
Confidence Intervals and Hypothesis Testing

Part 6

March 25, 2025

Part 6

Learning objective: begin to understand how inferential statistics account for uncertainty in sample data due to random chance

recognize how: confidence intervals (CIs)….

account for how different samples may yield different statistics

reflect uncertainty by considering sampling variability

provide a range of plausible values for a population parameter

Takeaway: descriptive statistics like means or proportions are useful for summarizing sample data, but inferential statistics are necessary to quantify the uncertainty in these summaries and assess how well they estimate the corresponding population parameters

Confidence interval (CI)

- range of values around a sample estimate that reflects the uncertainty due to random sampling variability, with a certain level of confidence
  - that the true population parameter lies within this range

provides upper and lower bounds around an estimate in a way that expresses likelihood of where the true parameter falls

<u>CI for sample mean</u>

MoE

$$CI = \hat{\theta} \pm Z\left(\frac{\theta}{\sqrt{N}}\right)$$

when parameters are known

or

$$CI = \hat{\theta} \pm Z \times SE$$

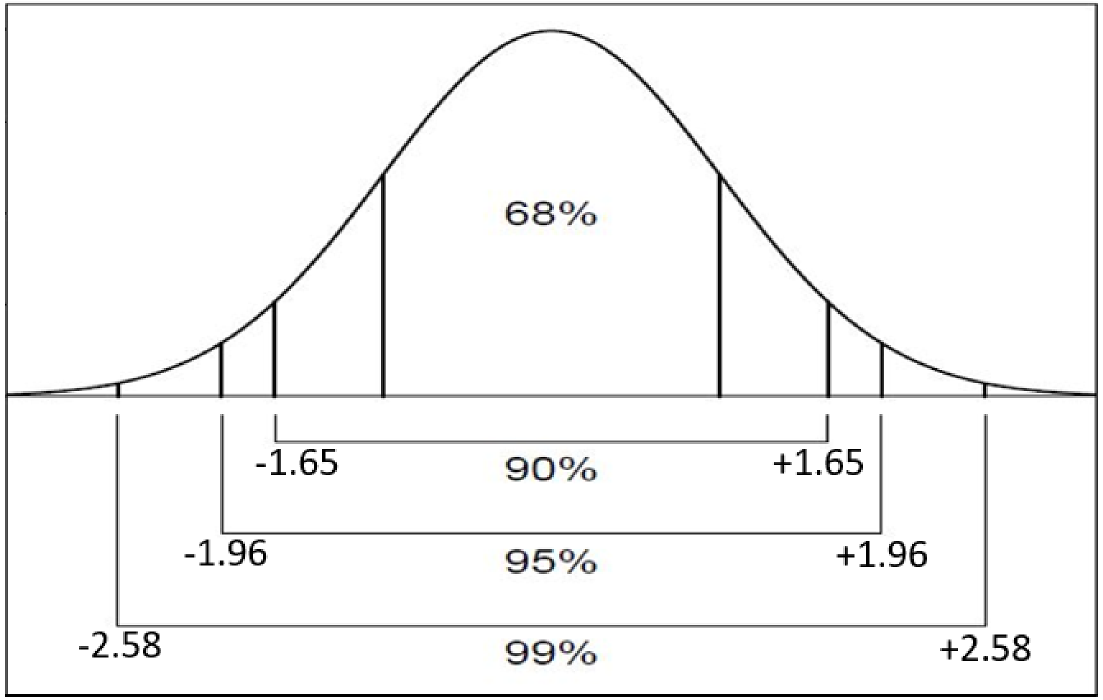
when parameters are unknown

$\hat{\theta}$ is a sample estimate (e.g., mean, proportion, variance, sd)

θ is the true population parameter

Notice how the CI around a sample estimate is a function of the margin of error (MoE)

*different equations for different statistics, but all use a MoE



lower bound critical value

Alpha

upper bound critical value

CLT and Probability: example continued

Let's consider how the CI around $\bar{X}age_{n10} \approx \mu = 46.45$ differs by level of certainty

CIs enable us to make generalizable statements about sample estimates by

providing a range of values within which the true population parameter is likely to fall,

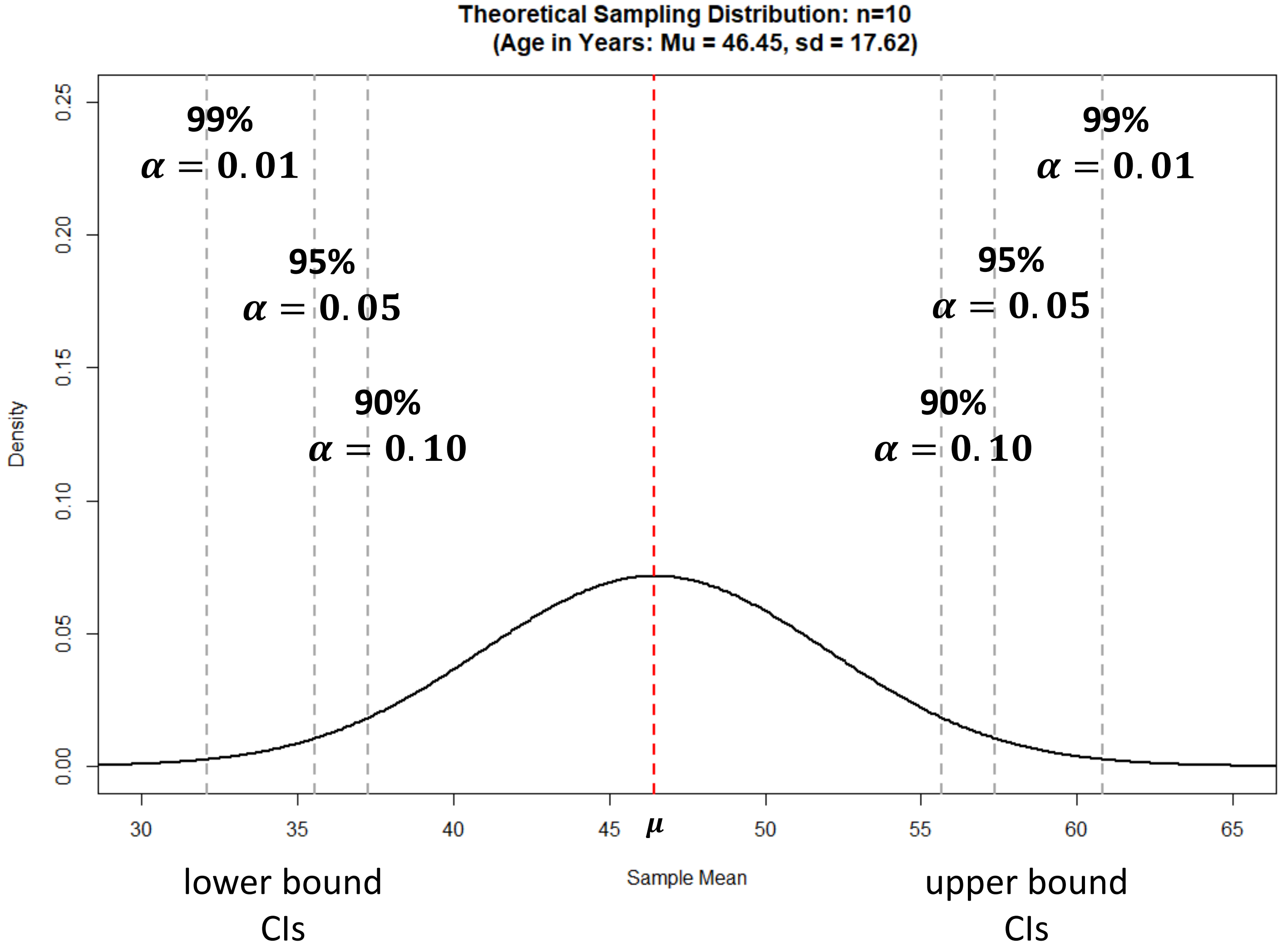
with a given level of confidence,

if we were to take many random samples from the population

$$90\%\ CI = 46.45 \pm 1.65 \times 5.57 = (37.26,55.65)$$

$$95\%\ CI = 46.45 \pm 1.96 \times 5.57 = (35.53,57.38)$$

$$99\%\ CI = 46.45 \pm 2.58 \times 5.57 = (32.07,60.84)$$



CLT and Probability: example continued

Let’s consider how the CI around $\bar{X}age_{\alpha=0.05} \approx \mu = 46.45$ differs by sample size n_i

Recall that the CI around a sample estimate is a function of the margin of error (MoE)

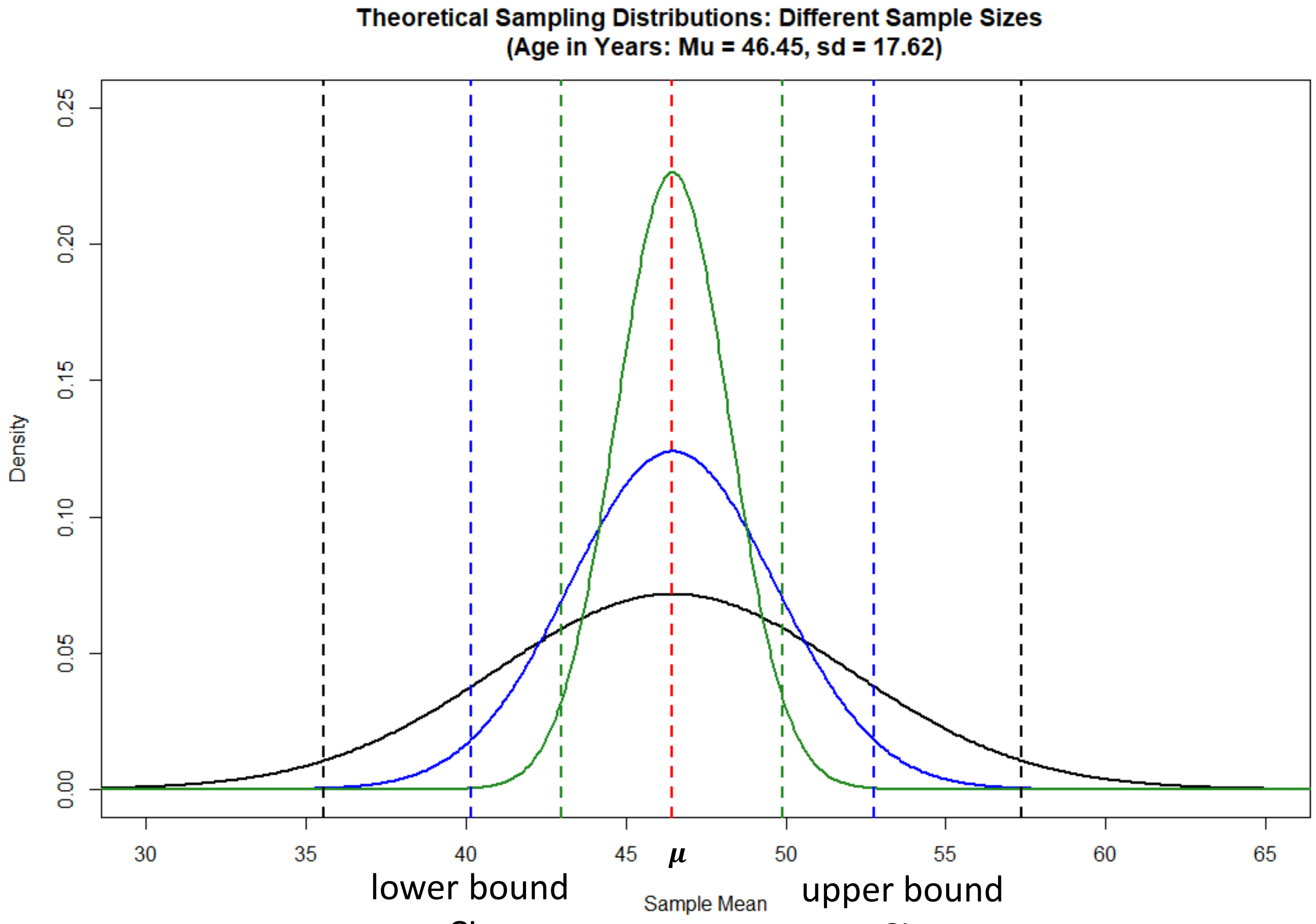
$$MoE = z \times SE$$

In the previous example we only changed the critical value z

In the present example we only changed the standard error (SE), which is a function of sample size

$$SE = \frac{\sigma}{\sqrt{N}}$$

Now, reconsider how as sample size increases SE deceases → more precise estimates



μ

lower bound CIs

upper bound CIs

Confidence interval (CI): construction

- Substitute the value of the sample statistic(s) $(\hat{\theta})$ in the corresponding CI equation
  - e.g., mean $(\bar{X})$, proportion $(\hat{p})$, differences in means $(\bar{X}_1 - \bar{X}_2)$ or proportions $(\hat{p}_1 - \hat{p}_2)$, variance (s^2), standard deviation (s)
- Set alpha (α), probability that the CI does not contain the true parameter
  - the error rate (α) represents the percentage of all possible resamples where the CI would fail to contain the true parameter
    - Note: can be found in the Z table
- Substitute the θ for the corresponding population parameter, or
  - value of the standard error (SE) for the corresponding sample statistic $(\hat{\theta})$
- Compute the margin of error (MoE)
- Construct the CI by ± the MoE to/from the sample statistic $(\hat{\theta})$, to obtain
  - lower bound and upper bound interval

MoE

<u>CI for sample mean</u>

$$CI = \hat{\theta} \pm Z\left(\frac{\theta}{\sqrt{N}}\right)$$

when parameters are known

or

$$CI = \hat{\theta} \pm Z \times SE$$

when parameters are unknown

Let's construct some CIs using our Netflix data…

CI for sample mean

$$CI = \hat{\theta} \pm \overbrace{Z\left(\frac{\theta}{\sqrt{N}}\right)}^{\text{MoE}}$$

or

$$CI = \hat{\theta} \pm Z \times SE$$

Confidence interval (CI): example

- Let’s pretend that our Netflix survey data based on simple random sample
  - from Netflix’s population of consumers
- How well do our sample statistics estimate the population parameters?
  - The parameters are unknown, so we will use
    - e.g., How well does a sample mean ($\bar{X}$) estimate the population mean (μ)?

$$CI = \hat{\theta} \pm Z \times SE$$

when parameters are unknown

- Inferential statistics, like CIs, work even when parameters are unknown because
  - Central Limit Theorem: holds that samples behave predictably
  - Standard Error: accounts for sampling variability
  - Theoretical probability distributions: adjust for sample uncertainty
  - Probability theory: provides level of certainty given repeated sampling

Let’s consider how well our sample mean age ($\bar{X}_{age}$) estimates the population mean age (μ_{age})…

CI for sample mean

$$CI = \hat{\theta} \pm \overbrace{Z\left(\frac{\theta}{\sqrt{N}}\right)}^{\text{MoE}}$$

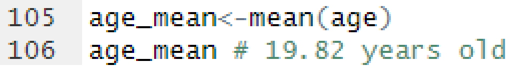
or

$$CI = \hat{\theta} \pm \overbrace{Z \times SE}^{\text{MoE}}$$

Confidence interval (CI): example cont.

- Substitute the value of the sample statistic(s) $(\hat{\theta})$ in the CI equation
  - e.g., mean $(\bar{X})$, proportion $(\hat{p})$, differences in means $(\bar{X}_1 - \bar{X}_2)$ or proportions $(\hat{p}_1 - \hat{p}_2)$, variance (s^2), standard deviation (s)

$$\bar{X}_{age} = 19.82\ years\ old$$



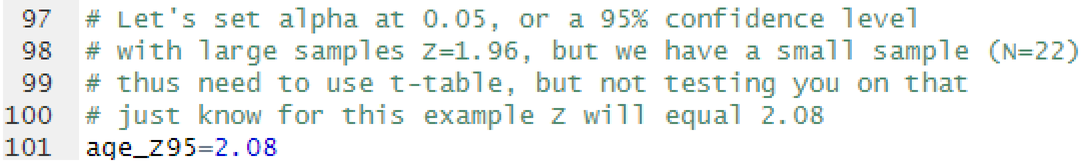
```
age_mean<-mean(age)
age_mean # 19.82 years old
```

RScript_Netflix_v2 in Netflix Data module on Canvas

- Set alpha (α), probability that the CI does not contain the true parameter
  - the error rate (α) represents the percentage of all possible resamples where the CI would fail to contain the true parameter
    - Note: can be found in the Z table, but we will use other theoretical probability distribution t-table b/c small sample (N<100)

$\alpha = 0.05$, which corresponds to a 95% confidence level

- Thus, 95% of all possible resamples should have a mean within the interval we will estimate
  - 95% of the time a sample of N=22 is drawn from this population it will contain a mean age within this range



```
# Let's set alpha at 0.05, or a 95% confidence level
# with large samples Z=1.96, but we have a small sample (N=22)
# thus need to use t-table, but not testing you on that
# just know for this example Z will equal 2.08
age_Z95=2.08
```

So far, we have this much of the equation

$$CI = 19.82 \pm 2.08 \times SE$$

Let's compute the standard error (SE)....

CI for sample mean

MoE

$$CI = \hat{\theta} \pm Z\left(\frac{\theta}{\sqrt{N}}\right)$$

or

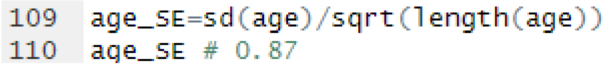
$$CI = \hat{\theta} \pm Z \times SE$$

Confidence interval (CI): example cont.

- Substitute the θ for the corresponding population parameter, or
  - value of the standard error (SE) for the corresponding sample statistic $(\hat{\theta})$

$$SE = \frac{s}{\sqrt{N}}$$

$$SE_{age} = \frac{sd_{age}}{\sqrt{22}} = 0.87$$

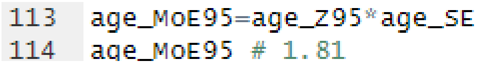


```
age_SE=sd(age)/sqrt(length(age))
age_SE # 0.87
```

- Compute the margin of error (MoE)

$$MoE = Z \times SE$$

$$MoE_{age} = 2.08 \times 0.87 = 1.81$$



```
age_MoE95=age_Z95*age_SE
age_MoE95 # 1.81
```

- Construct the CI by $\pm$ the MoE to/from the sample statistic $(\hat{\theta})$, to obtain
  - lower bound and upper bound interval

$$95\% \; CI \; (18.00{,}19.82)$$

Anyone want to take a shot at interpreting this?

Exercise 6: CIs for sample mean

- Paper handout
  - Also available on Canvas “Exercise 6.pdf”