Quantitative Sociological Analysis

Inferential Statistics
Confidence Intervals and Hypothesis Testing

Part 5

March 6, 2025

Part 5 (*modified*)

Learning objective: begin to understand that sample data have uncertainty due to chance, which must be addressed to make generalizable statements that can be applied to the broader population

recognize how:

probability theory underlies sampling

the Central Limit Theorem (CLT) connects probability and sampling

differences between a population and sample due to chance can be addressed

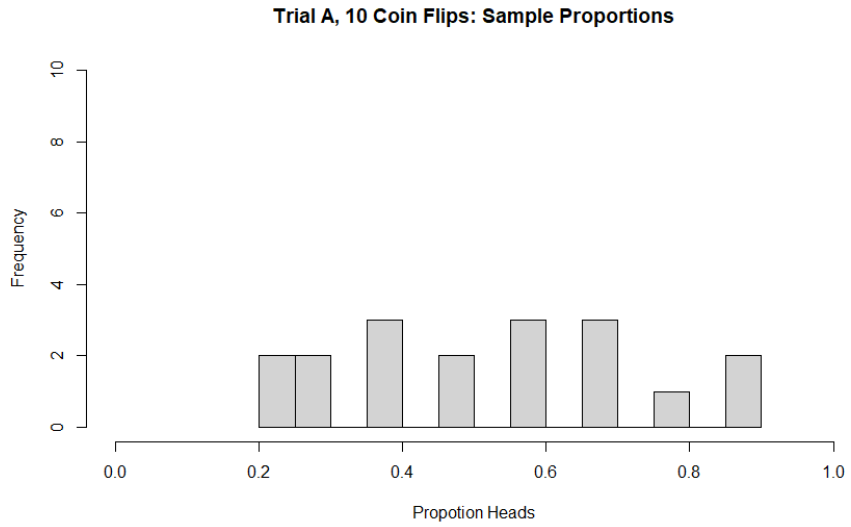
the margin of error (MoE) is critical for quantifying uncertainty in sample estimates, helping to establish confidence in their accuracy

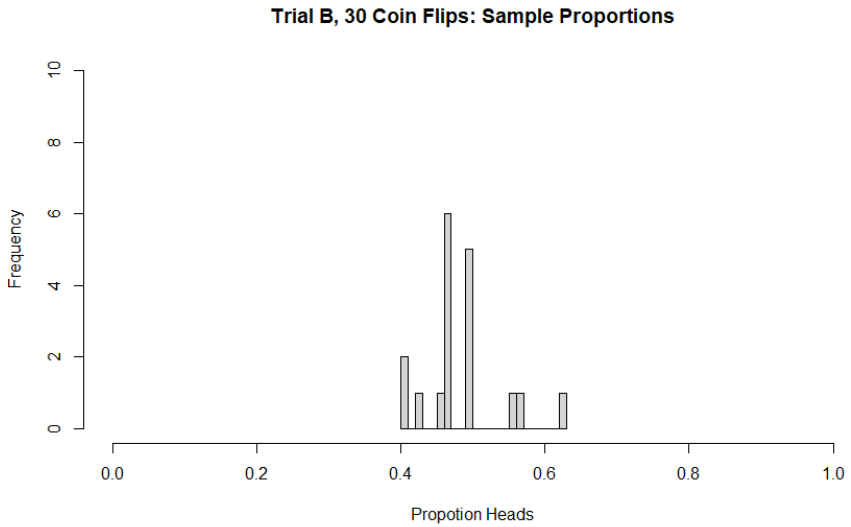
Takeaway: descriptive statistics help summarize sample data, but they cannot produce generalizable conclusions because they do not account for sampling variability

Central Limit Theorem (CLT)

- the sampling distribution of the sample mean ($\bar{X}$) or proportion ($\hat{p}$) approaches a normal distribution as the sample size increases, when
  - samples are randomly selected and independent

We saw how this worked in Exercise 5, which treated coin toss trials as a form of simple random sampling (SRS)



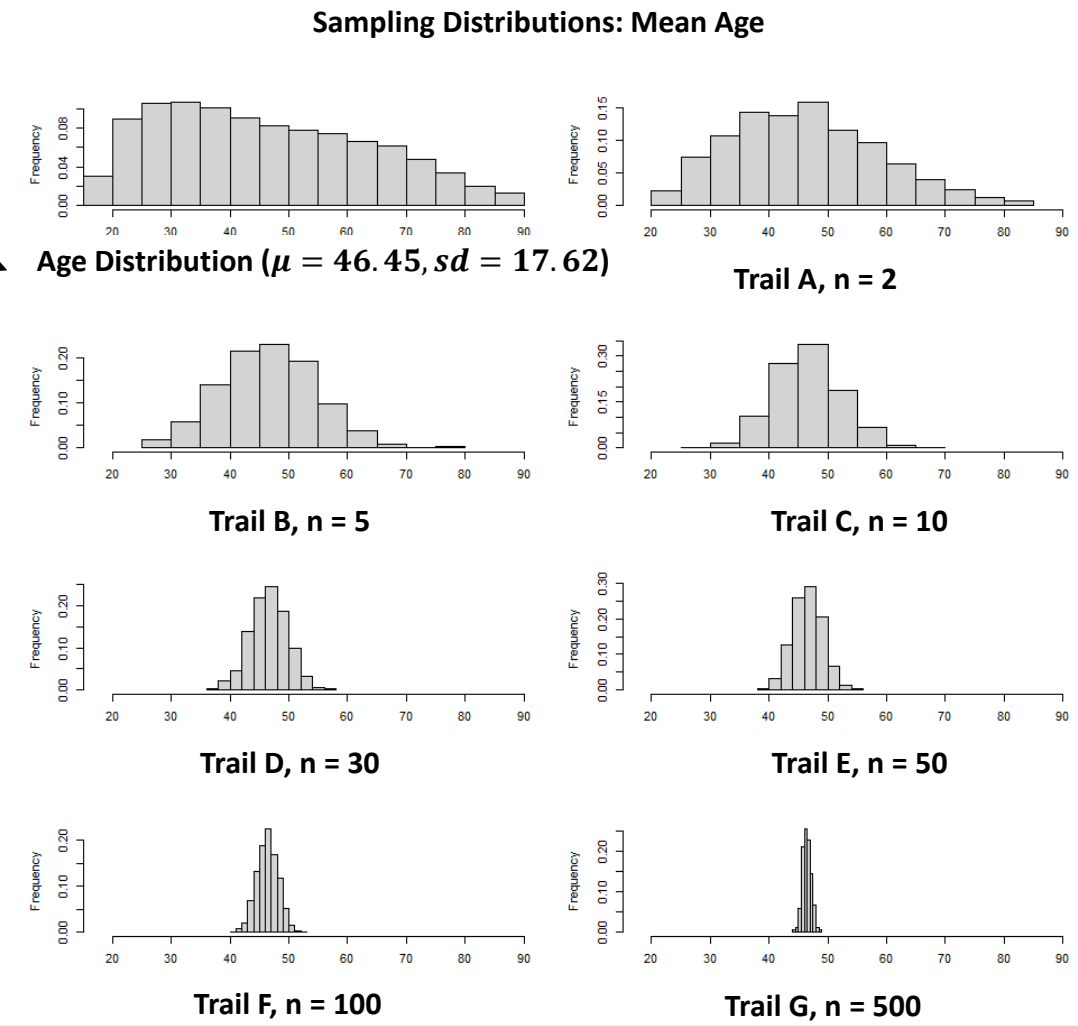


Let's further establish the importance of this point by revisiting the age example from PPT 7

Samling distribution and CLT: example

See how the frequency distribution of age is not normal, but positively (right) skewed

same age distribution, just depicted with binned ages



- Let's treat this like a population from which we will randomly select 1,000 samples of size n_i
  - number of samples like trials in coin toss exercise
  - sample size (n_i) is like number of tosses in each trial

Accounting for the expected fluctuation of a sample statistic due random sampling variability is crucial for making inferences about the population…

Sampling distribution and CLT: standard error (SE)

- measures the variability of a sample statistic across repeated samples
  - helps determine the likelihood that sample results reflect the true population parameter
    - e.g., $\bar{X} \approx \mu$ or $\hat{p} \approx p$

when parameters are known $SE = \frac{\sigma}{\sqrt{N}}$ or $SE = \frac{s}{\sqrt{N}}$ when parameters are unknown

σ is the population standard deviation

s is the sample standard deviation

N is the sample size

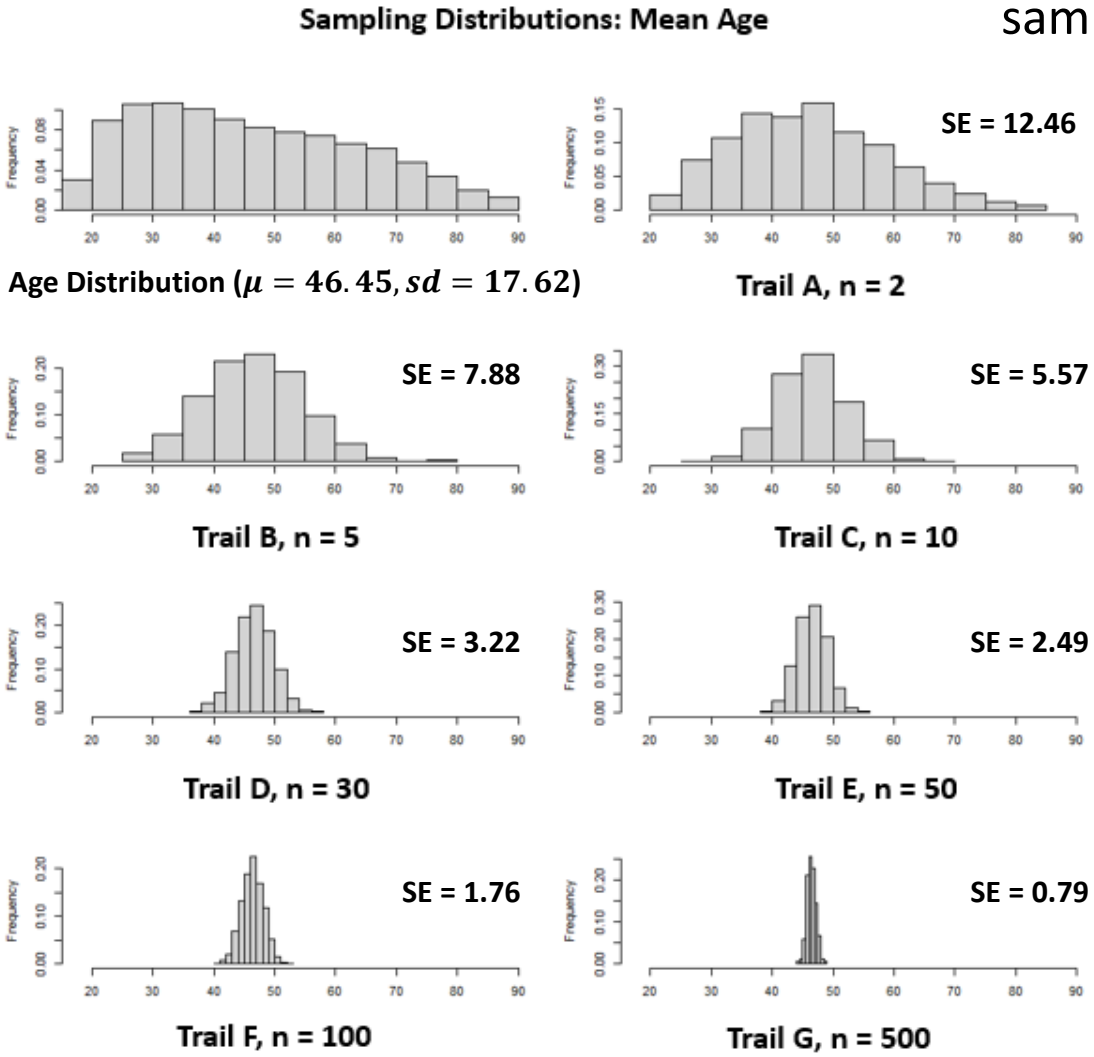
* equation for standard error specific to sample statistic type, not just whether un/known parameters

In general, as sample size increases SE deceases → more precise estimates

Useful for extending the normal distribution's properties to further understand the likelihood of sample statistics

CLT and SE: example

See how as sample size increases SE decreases and estimates become more precise, that is sample means ($\bar{X}$) more closely match the population mean (μ)



- That's because SE is a function of, depends on, sample size (n_i)

$$SEn_2 = \frac{17.62}{\sqrt{2}} = 12.46$$

$$SEn_5 = \frac{17.62}{\sqrt{5}} = 7.88$$

$$SEn_{10} = \frac{17.62}{\sqrt{10}} = 5.77$$

$$SEn_{30} = \frac{17.62}{\sqrt{30}} = 3.22$$

$$SEn_{50} = \frac{17.62}{\sqrt{50}} = 2.49$$

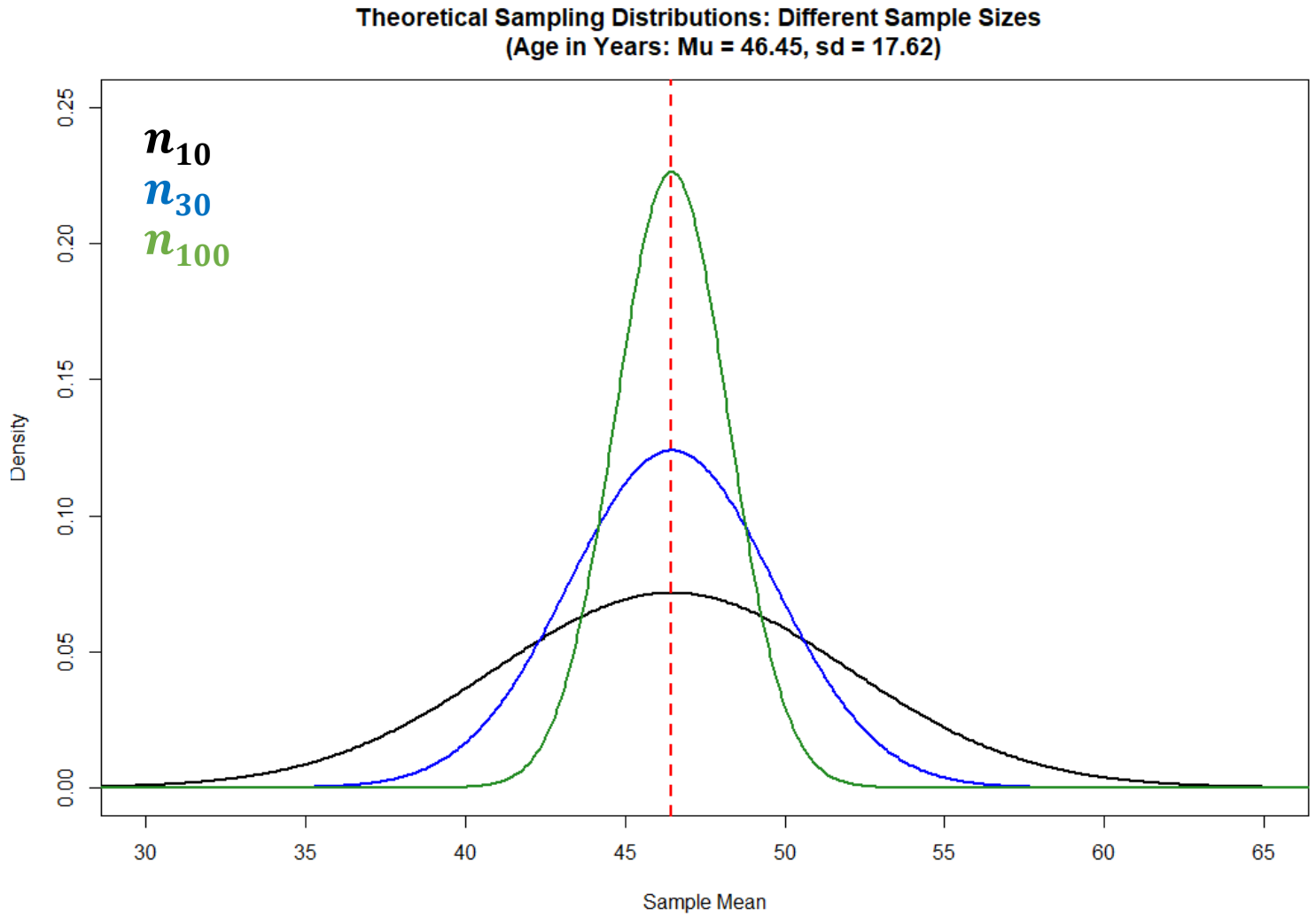
$$SEn_{100} = \frac{17.62}{\sqrt{100}} = 1.76$$

$$SEn_{500} = \frac{17.62}{\sqrt{500}} = 0.79$$

When a random sample is sufficiently large, we use probability theory and the CLT to make inferences about the broader population…

CLT and Probability: example continued

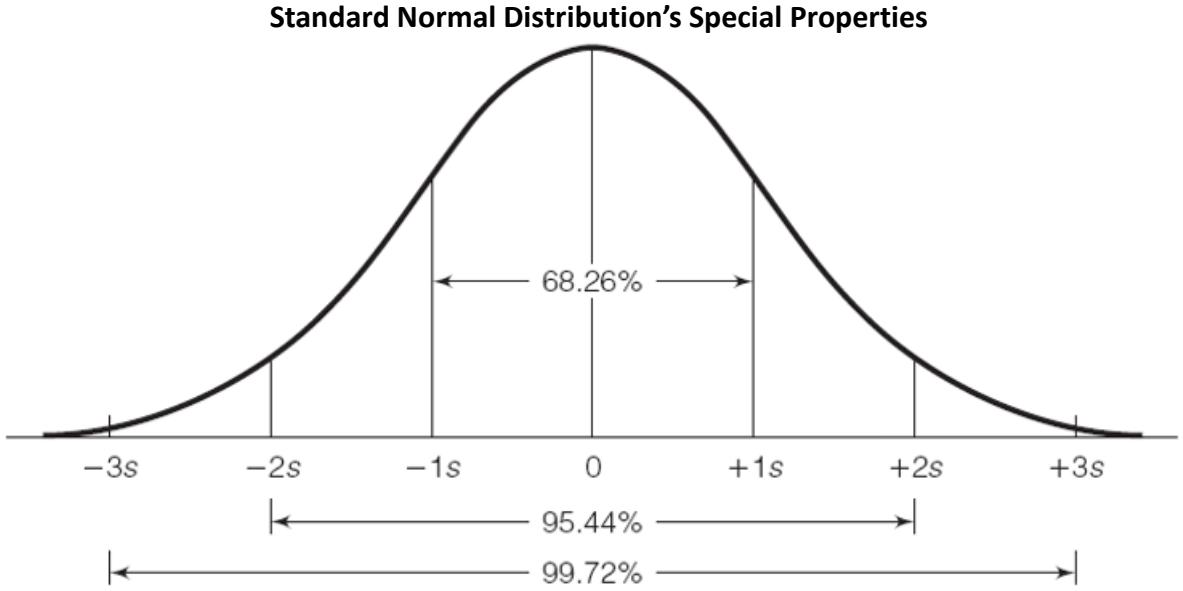
Here's another way of depicting how sample means ($\bar{X}$) more closely match the population mean (μ) as sample size increases, that is estimates become more precise

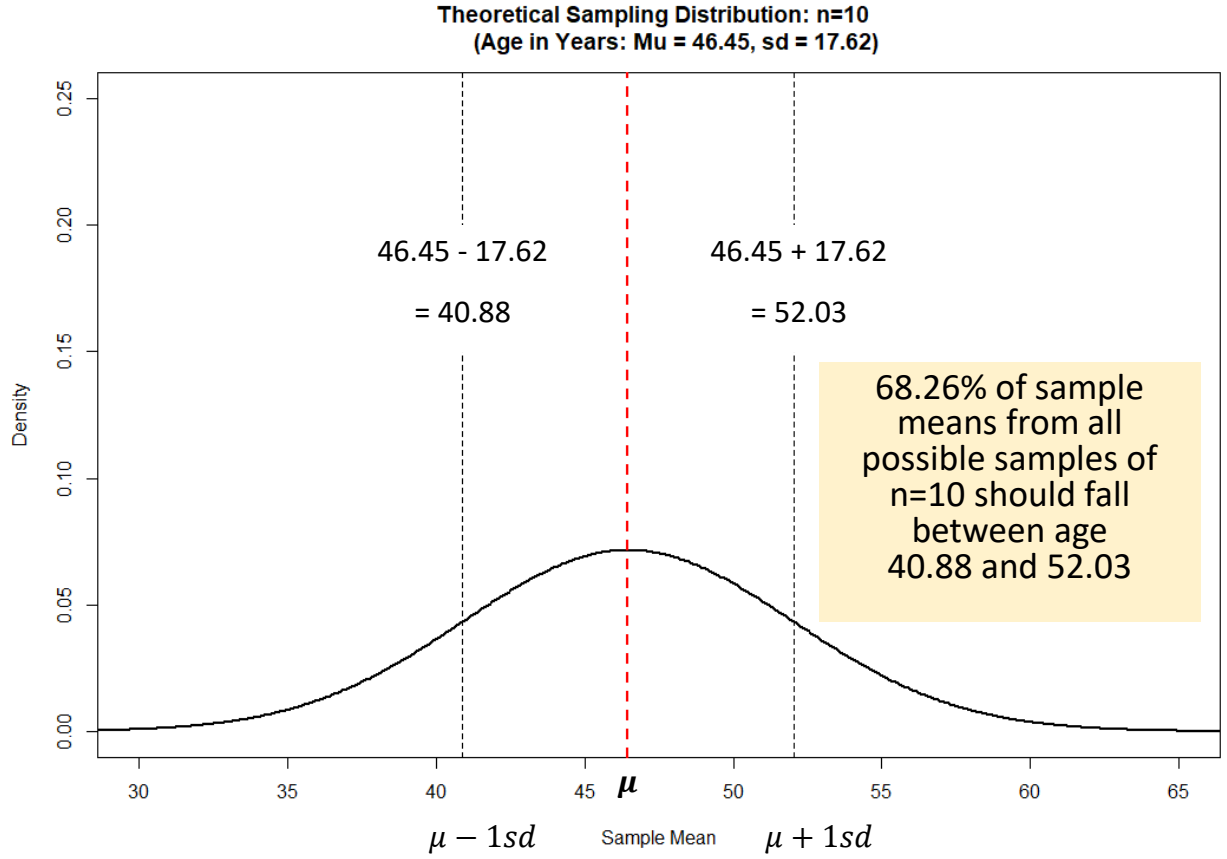


CLT and Probability: example

- Given that the sampling distribution of $\bar{X}$ or $\hat{p}$ approximates the standard normal distribution as n increases, we can use its special properties to describe a range of probabilities
  - when samples are randomly selected

First, note that in a theoretical sampling distribution 68.26% of $\bar{X}s$ or $\hat{p}s$ will fall within $\pm$ 1 sd of the population parameter (i.e., μ or p)





CLT and Probability: example continued

- Z-scores: transform data to follow properties of standard normal distribution, which
  - enables us to determine probabilities based on the special properties of this distribution

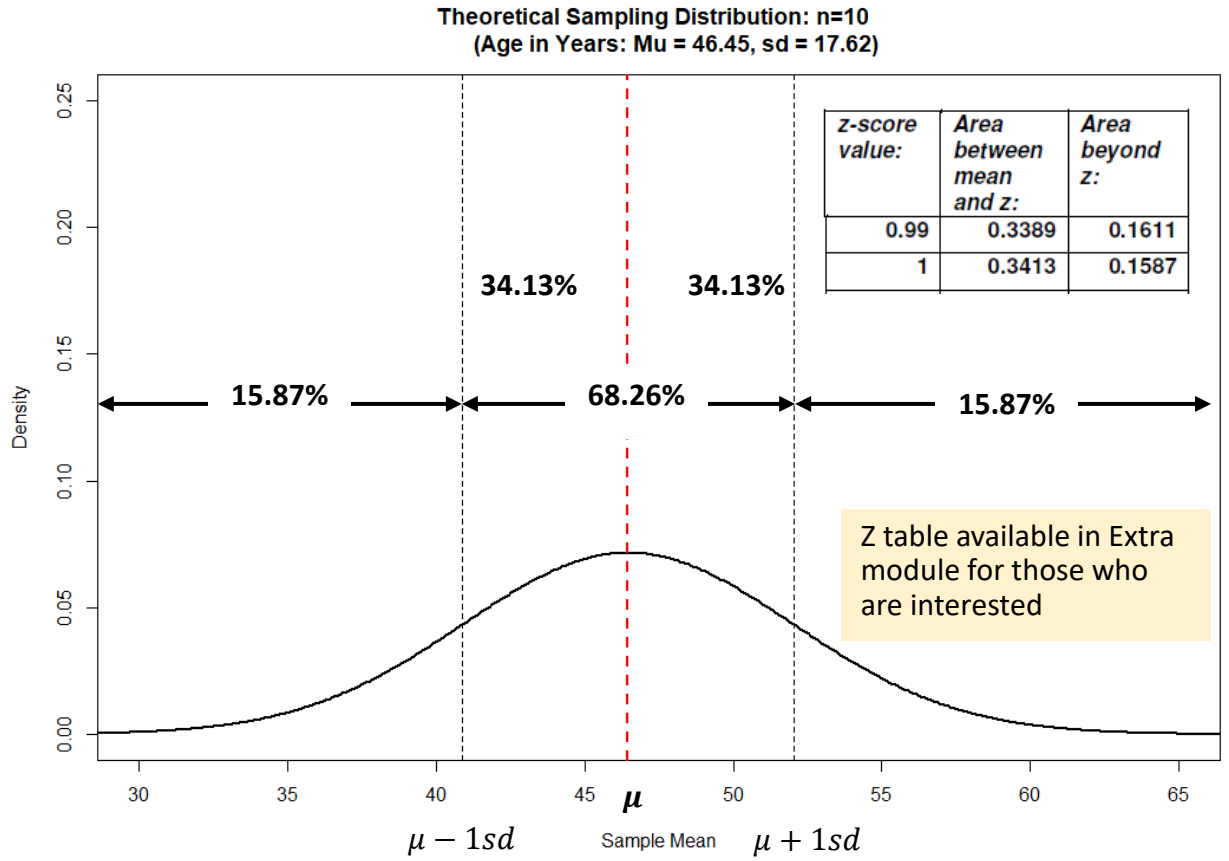
What is the probability that the mean age $(\bar{X}_{n10})$ of a random sample will fall within $\pm$ 1sd $(\sigma = 17.62)$ of the population mean $(\mu = 46.45)$?

- To answer this question, we must account for the expected fluctuation of a sample statistic due random sampling variability
  - $SE_{n10} = 5.57$

$$z = \frac{X - \mu}{SE}$$

$$z_{lower} = \frac{(46.45 - 17.62) - 46.45}{5.57} = -1$$

$$z_{upper} = \frac{(46.45 + 17.62) - 46.45}{5.57} = 1$$



Z-scores are used to identify the relative position of a value within the standard normal distribution…

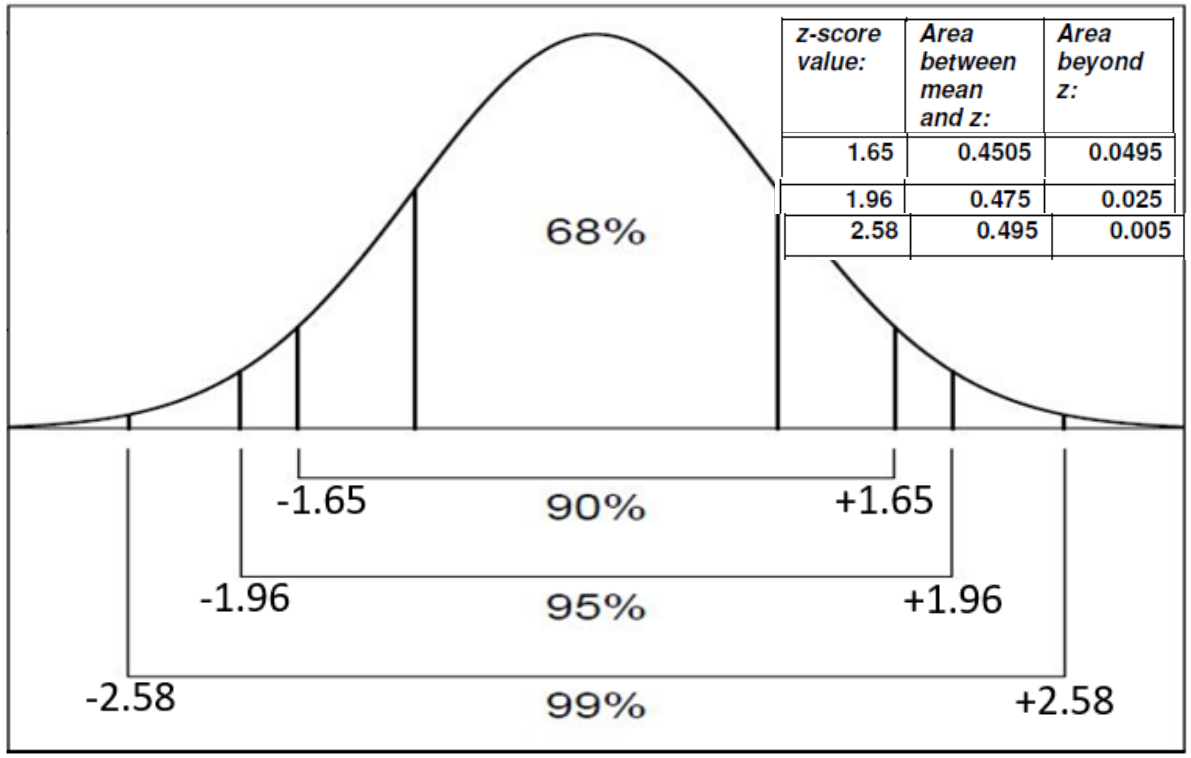
Critical values

- points on the sampling distribution that correspond to a probability threshold, which can be used to establish boundaries around an estimate
  - reflecting the expected variability due to sampling error
- Alpha (α) represents the probability threshold that determines how often an estimate, drawn from repeated random samples, would fall beyond a defined range around the true population value
  - due to random sampling variability

For example,

A lower and upper critical value of ± 1.65 corresponds with a probability threshold of 0.90 $(\alpha = 0.10)$, meaning that in 90% of all possible random samples from the population, the estimate will fall within the range set by these boundaries

In other words, there is a 10% probability that the estimate was due to random sampling variability rather than reflecting the true population value



lower bound critical value

Alpha

upper bound critical value

Critical value and SE

- Margin of error (MoE): expected range around estimate due to random sampling variability
  - boundaries determined by the probability threshold (α), while
  - random sampling variability is accounted for by the SE

$$MoE = z \times SE$$

Because of the uncertainty inherent in sample data due to random sampling variability, a range of values must be considered when determining the probability that a sample estimate reflects the true population parameter

In other words,

- MoE quantifies the uncertainty in a sample estimate and represents the range within which the true population parameter is likely to lie,
  - with a certain level of confidence

CLT and Probability: example continued

Let's consider how the margin of error around $\bar{X}age_{n10} \approx \mu = 46.45$ differs across different levels of certainty

Given the uncertainty inherent in the sampling process, the MoE suggests that if we took many random samples, the true population parameter would lie within _____ units of each sample estimate _____ percent of the time.

$\alpha = 0.10$ $MoE = 1.65 \times 5.57 = 9.20$ years 90% of the time

$\alpha = 0.05$ $MoE = 1.96 \times 5.57 = 10.93$ years 95% of the time

$\alpha = 0.01$ $MoE = 2.58 \times 5.57 = 14.38$ years 99% of the time

Notice how the MoE increases as the level of confidence (or probability threshold) increases

The MoE can be used to construct a range around a sample estimate that reflects the uncertainty due to random sampling variability…

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

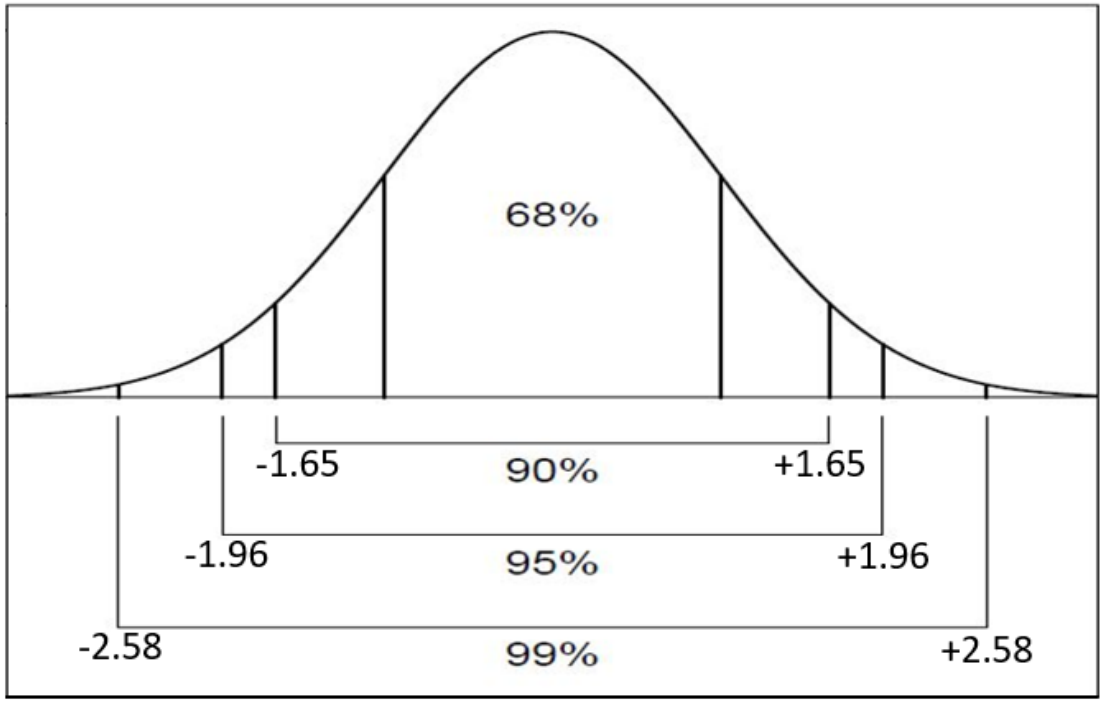
$$CI = \hat{\theta} \pm Z\left(\frac{\theta}{\sqrt{N}}\right)$$ when parameters are known

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



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,

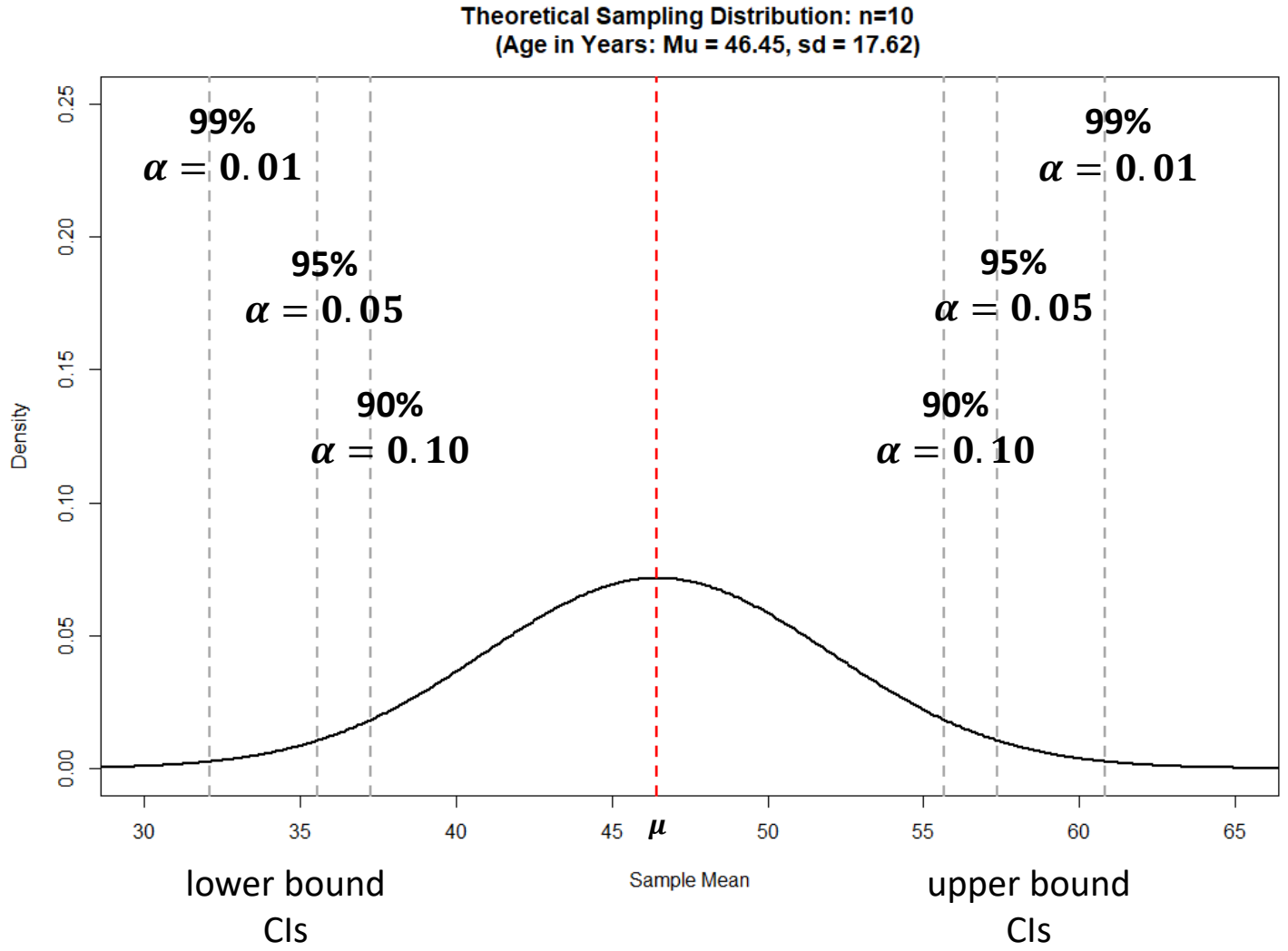
with a given level of confidence,

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)

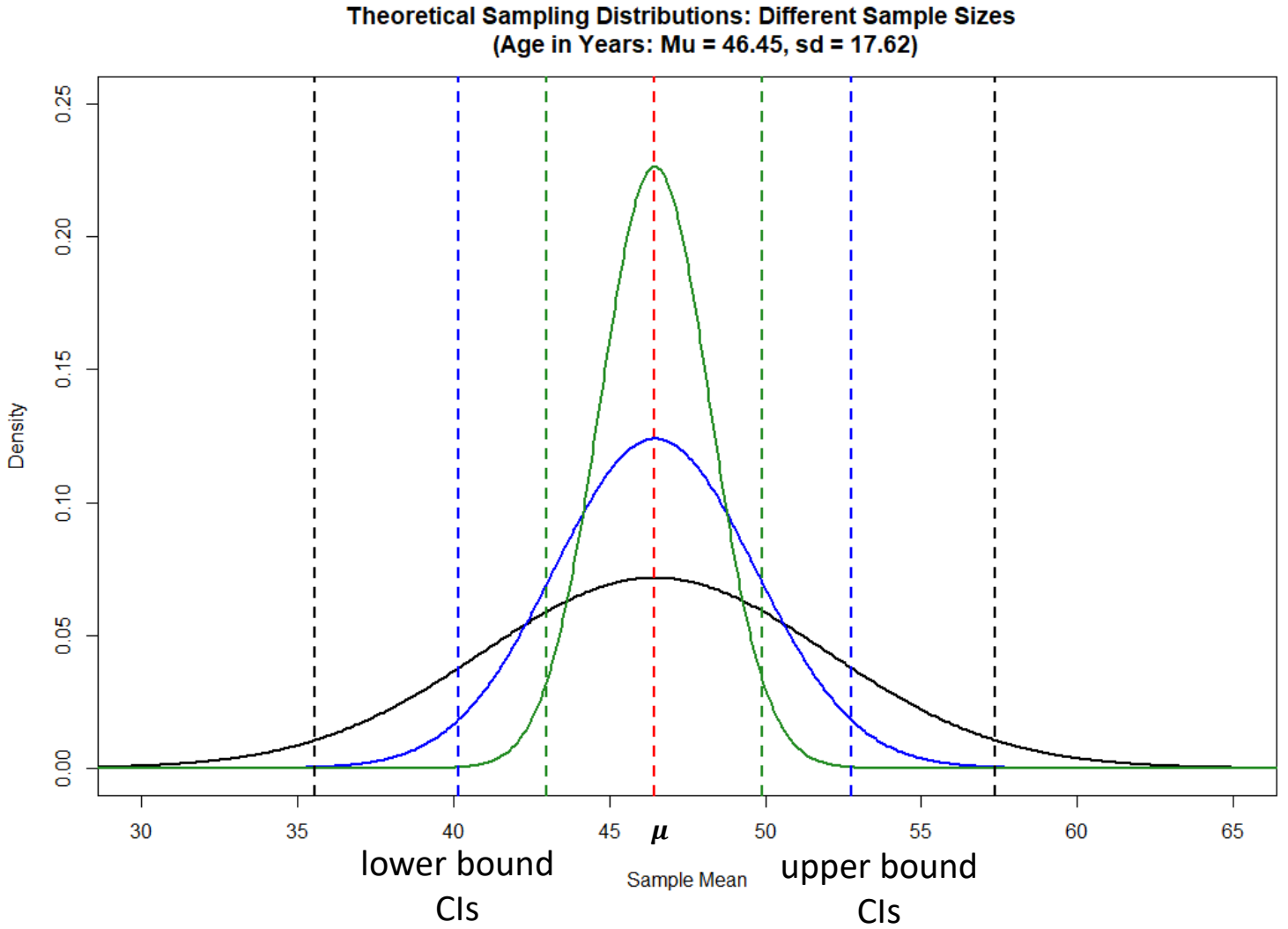
$$MoE = z \times SE$$

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



Confidence interval (CI): construction

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

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

Let's see what CIs look like using our height example…

Confidence interval (CI): example

- Let's consider theoretically randomly selecting 1,000 samples of size n_i from the US adult male population
  - $\boldsymbol{n_{10}}$, $\boldsymbol{n_{100}}$, $\boldsymbol{n_{500}}$
- How certain can we be that 95% ($\alpha = 0.05$) of the samples (n_i) contain the population mean ($\mu = 70$)?

$$\boldsymbol{CIn_{10}} = 70 \pm 1.96\left(\frac{3}{\sqrt{10}}\right) = 70 \pm 1.86$$

95% CI (68.14, 71.86)

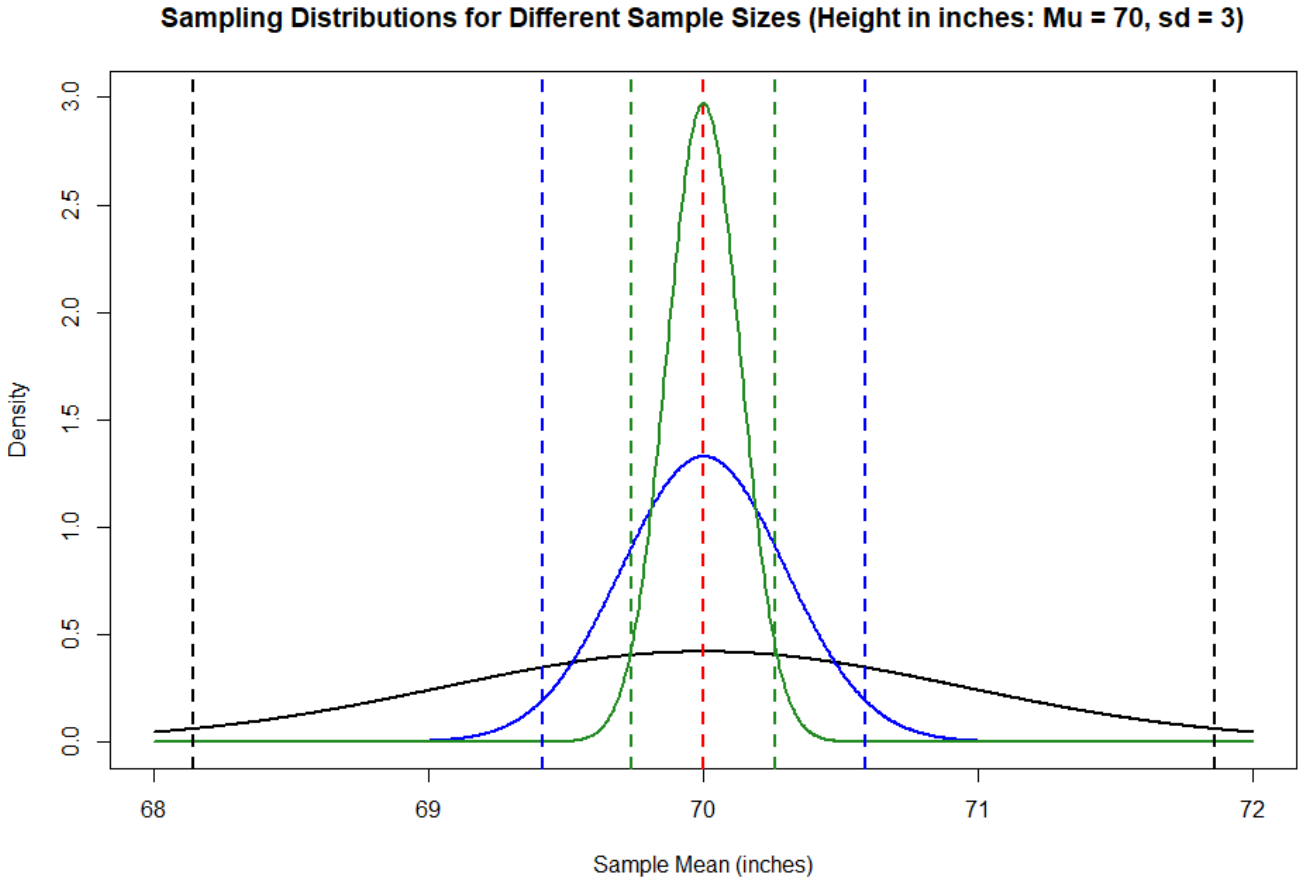
$$\boldsymbol{CIn_{100}} = 70 \pm 1.96\left(\frac{3}{\sqrt{100}}\right) = 70 \pm 0.59$$

95% CI (69.41, 70.59)

$$\boldsymbol{CIn_{500}} = 70 \pm 1.96\left(\frac{3}{\sqrt{500}}\right) = 70 \pm 0.26$$

95% CI (69.74, 70.26)

see how CI becomes narrower as sample size increases



Confidence interval (CI): in R w/o the math

- All the math examples available in the RScript for those interested
  - will be tested only on conceptual meaning and practical interpretation
    - relatively simple command to construct CI for $\bar{X}$

```
189  t.test(age, conf.level=(0.90))
190  t.test(age, conf.level=(0.95))
191  t.test(age, conf.level=(0.99))
```

```
data:  age
t = 669.51, df = 64554, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 46.34064 46.56891
sample estimates:
mean of x
 46.45477
```

- We can even construct CIs for nominal and ordinal variables, but
  - this gets a little more complicated, because it requires different equations that account for an approximation to the normal distribution

```
260  # binary (0,1): nominal or ordinal variables
261  # first, identify number of successes
262  table(female)
263  # prop.test(number of success,sample size)
264  prop.test(35973,64555, conf.level=(0.95))
```

```
data:  35973 out of 64555, null probability 0.5
X-squared = 845.98, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5534030 0.5610816
```

```
266  # this gets more complicated when more than two categories, here is one way...
267  # first, identify number of responses in each category
268  # happiness: (1) not too happy, (2) pretty happy, (3) very happy: ordinal
269  table(happy)
270  # generate a variable to save number of responses in each category
271  responses <- c(8751,36159,19645)  # Frequency of responses in each category
272  # generate a variable to save sample size
273  n<-sum(responses) #one way to do this, but could just type in the sample size
274  # compute confidence intervals for each category
275  ci_list <- lapply(responses, function(x) prop.test(x, n, conf.level = 0.95)$conf.int)
276  # see results
277  ci_list
```

```
[[1]]
[1] 0.1329321 0.1382290
attr(,"conf.level")
[1] 0.95
```

```
[[2]]
[1] 0.5562868 0.5639601
attr(,"conf.level")
[1] 0.95
```

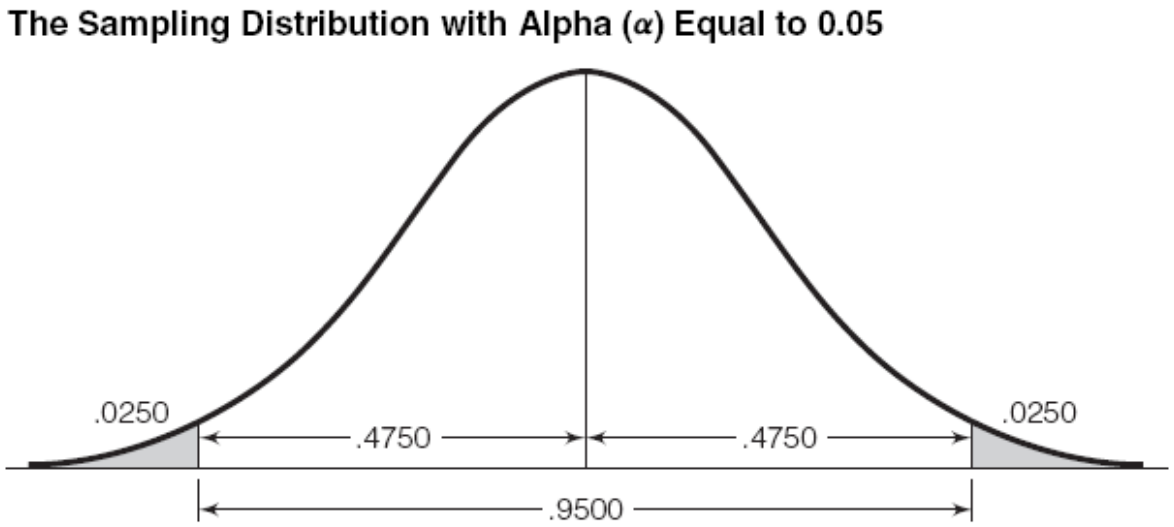
```
[[3]]
[1] 0.3007688 0.3078828
attr(,"conf.level")
[1] 0.95
```

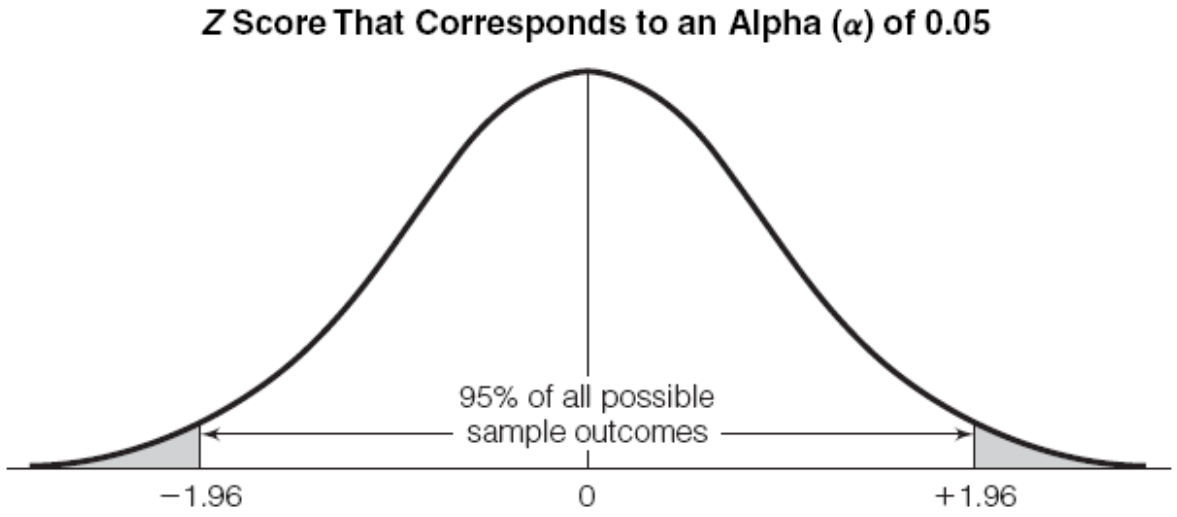
Note: this treats the GSS as sample data, whereas previous examples pretended that this was a population

Confidence interval (CI): key takeaway

- built upon probability theory and the CLT
  - relies on theoretical resampling of all possible samples, and
    - even though we often only work with one sample
  - draws heavily on the properties of the standard normal distribution

Useful for drawing generalizable conclusions about population parameters from a sample, even when the true parameters are unknown. Consider how this inferential technique extends descriptive statistics so they can be applied to the broader population.





Let’s cover one more univariate technique, which will hopefully help better establish your foundation in inferential statistics...