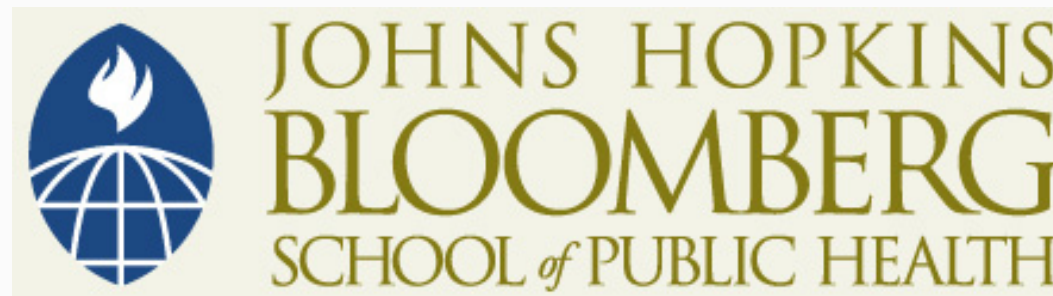


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2009, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



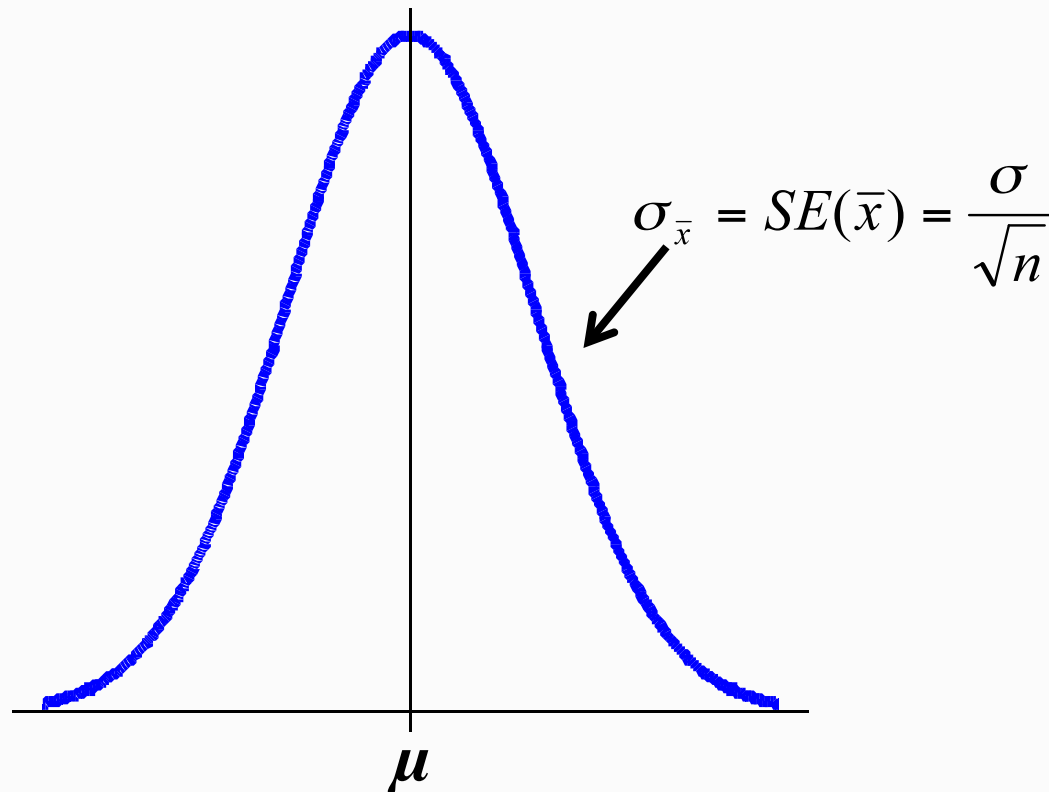
JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section D

True Confessions Biostat Style: What We Mean by
Approximately Normal and What Happens to the Sampling
Distribution of the Sample Mean with Small n

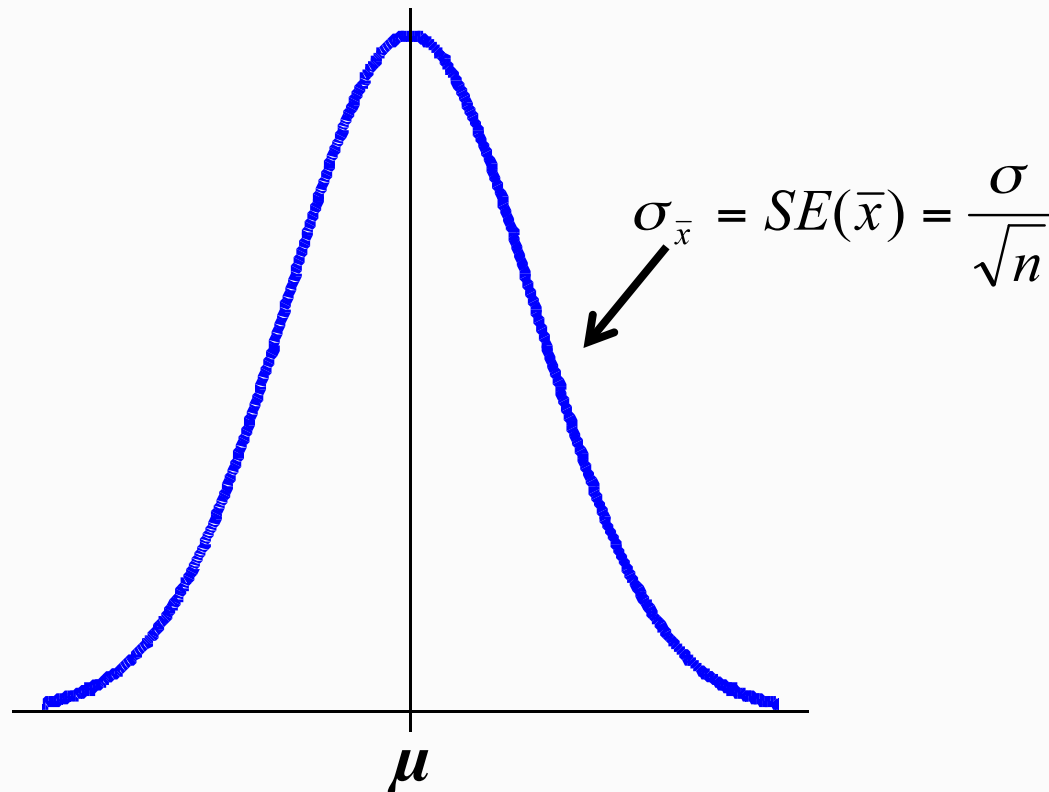
Recap: CLT

- So the CLT tells us the following: when taking a random sample of continuous measures of size n from a population with true mean μ and true sd σ the theoretical sampling distribution of sample means from all possible random samples of size n is:



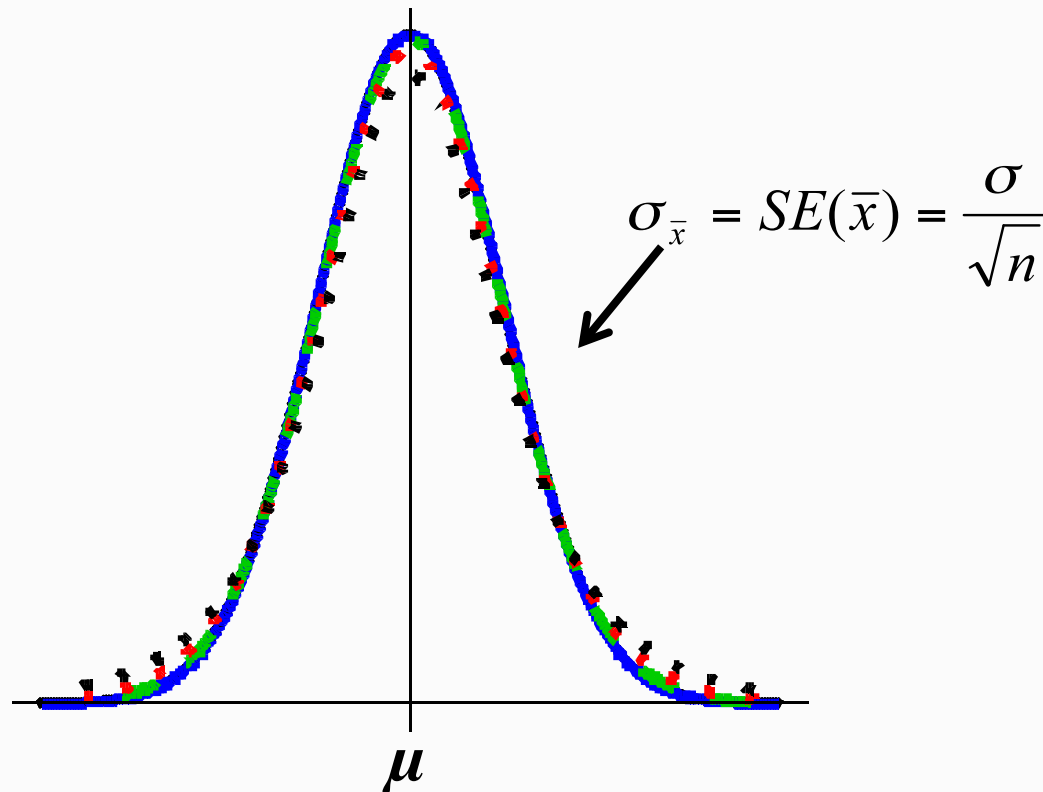
Recap: CLT

- Technically this is true for “large n ”: for this course, we’ll say $n > 60$; but when n is smaller, sampling distribution is not quite normal, but follows a *t-distribution*



t-distributions

- The t-distribution is the “fatter, flatter cousin” of the normal: t-distribution is uniquely defined by degrees of freedom



Why the t?

- Basic idea: remember, the true $SE(\bar{x})$ is given by the formula

$$\sigma_{\bar{x}} = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

- But of course we don't know σ , and replace with s to estimate

$$\hat{SE}(\bar{x}) = \frac{s}{\sqrt{n}}$$

- In small samples, there is a lot of sampling variability in s as well: so this estimate is less precise
- To account for this additional uncertainty, we have to go slightly more than $\pm 2 \times \hat{SE}(\bar{x})$ to get 95% coverage under the sampling distribution

Underlying Assumptions

- How much bigger the 2 needs to be depends on the sample size
- You can look up the correct number in a “t-table” or “t-distribution” with $n-1$ degrees of freedom

The t-distribution

- So if we have a smaller sample size, we will have to go out more than 2 SEs to achieve 95% confidence
- How many standard errors we need to go depends on the degrees of freedom—this is linked to sample size
- The appropriate degrees of freedom are $n - 1$
- One option: you can look up the correct number in a “t-table” or “t-distribution” with $n - 1$ degrees of freedom

$$\bar{x} \pm t_{.95, n-1} \times \hat{SE}(\bar{x}) \Rightarrow$$

$$\bar{x} \pm t_{.95, n-1} \times \frac{s}{\sqrt{n}}$$

Notes on the t-Correction

- The particular t-table gives the number of SEs needed to cut off 95% under the sampling distribution

df	t	df	t
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.360	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	∞	1.960

Notes on the t-Correction

- You can easily find a t-table for other cutoffs (90%, 99%) in any stats text or by searching the internet
- Also, using the *cii* command takes care of this little detail
- The point is not to spend a lot of time looking up t-values: more important is a basic understanding of why slightly more needs to be added to the sample mean in smaller samples to get a valid 95% CI
- The interpretation of the 95% CI (or any other level) is the same as discussed before

Example

- Small study on response to treatment among 12 patients with hyperlipidemia (high LDL cholesterol) given a treatment
- Change in cholesterol post-pre treatment computed for each of the 12 patients
- Results: $\bar{x}_{change} = -1.4 \text{ mmol/L}$
 $s_{change} = 0.55 \text{ mmol/L}$

Example

- 95% confidence interval for true mean change

$$\bar{x} \pm t_{.95,11} \times \hat{SE}(\bar{x}) \Rightarrow$$

$$\bar{x} \pm 2.2 \times \hat{SE}(\bar{x}) \Rightarrow$$

$$-1.4 \pm 2.2 \times \frac{0.55}{\sqrt{12}} \Rightarrow$$

$$(-1.75, \text{mmol/L}, -1.05 \text{ mmol/L})$$

Using Stata to Create Other CIs for a Mean

- The “cii” command,

```
. cii 12 -1.4 .55
```

Variable	Obs	Mean	Std	. Err.	[95% Conf. Interval]
	12	-1.4	.1587713		-1.749453 -1.050547