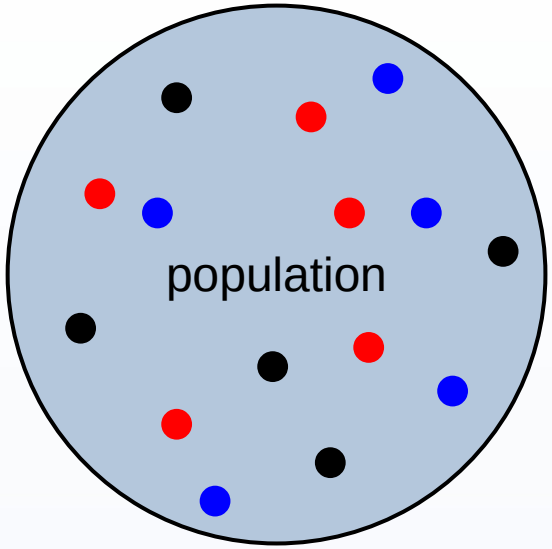
EGM101 – Skills Toolbox

Week 7, Part 6: The Central Limit Theorem

The Central Limit Theorem: Setup

- If we have:
  - Population with mean μ, finite variance σ^2
  - $S_i = \{X_1, X_2, \ldots, X_n\}$: random sample from population
  - $\bar{X}_n$: the sample mean of S_i
- What do we know about the sampling error?
  - i.e., how far away from μ is $\bar{X}_n$?
- Remember: we know from the law of large numbers that $\bar{X}_n$ gets closer and closer to μ as we increase n
- Question: what happens if we take many different random samples?



$$S_1 = \{X_1, X_2, X_3, X_4, X_5\}$$

$$S_2 = \{X_1, X_2, X_3, X_4, X_5\}$$

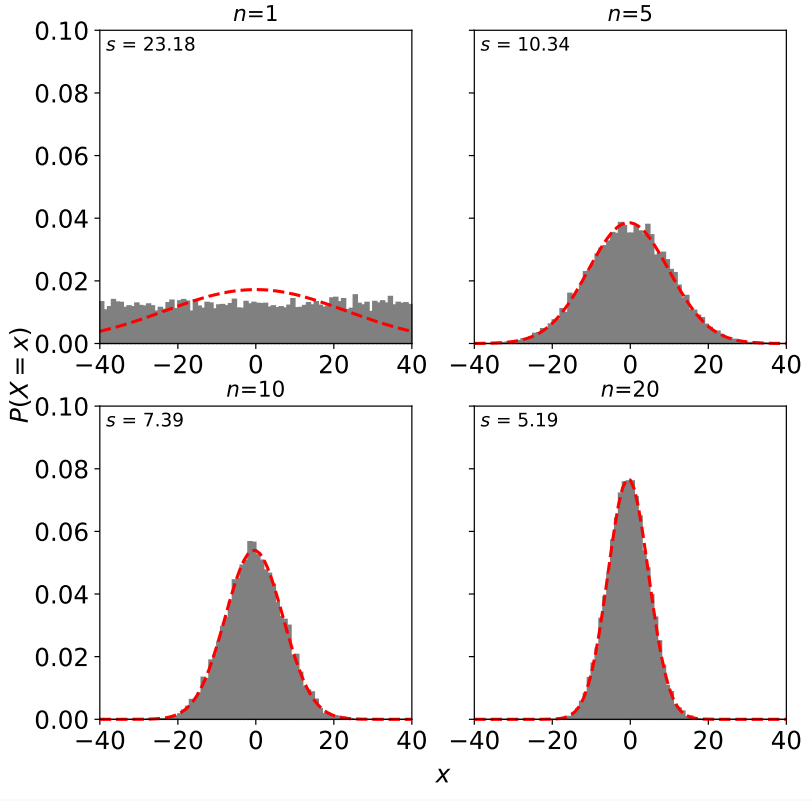
$$\vdots$$

$$S_N = \{X_1, X_2, X_3, X_4, X_5\}$$

The Central Limit Theorem (CLT)

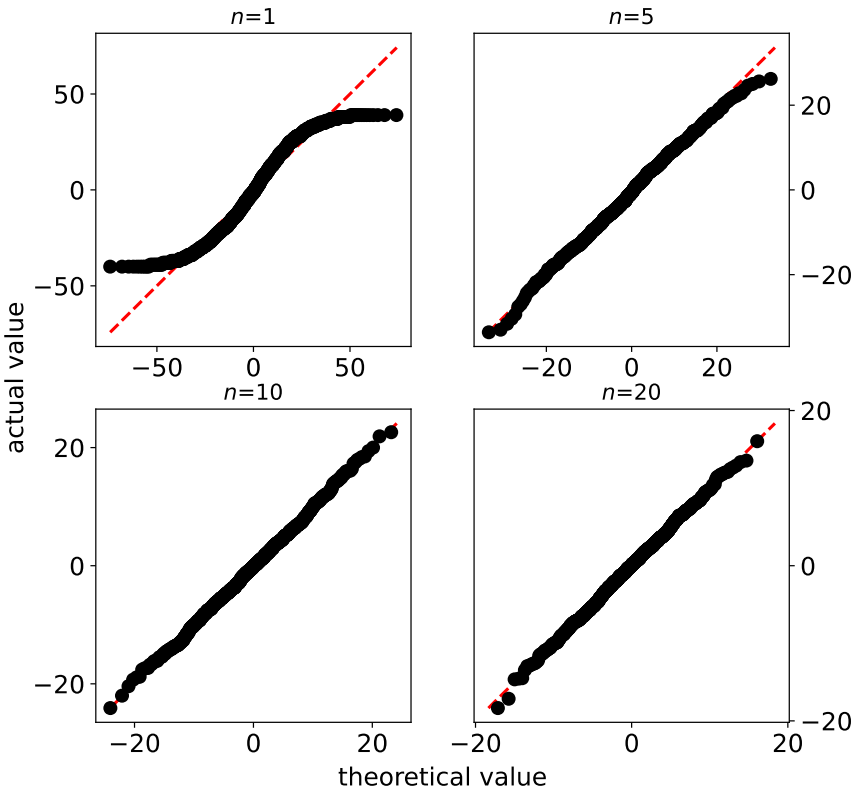
- The Central Limit Theorem:
  - As we increase n, the distribution of the sample means approximates a normal distribution with mean μ and variance σ^2
- In other words, as we take larger samples, the distribution of the *sample means*:
  - Becomes more like a normal distribution
  - Has a mean and variance that looks like the population
- Note: this does not depend on the distribution of the population!*

*some exceptions apply



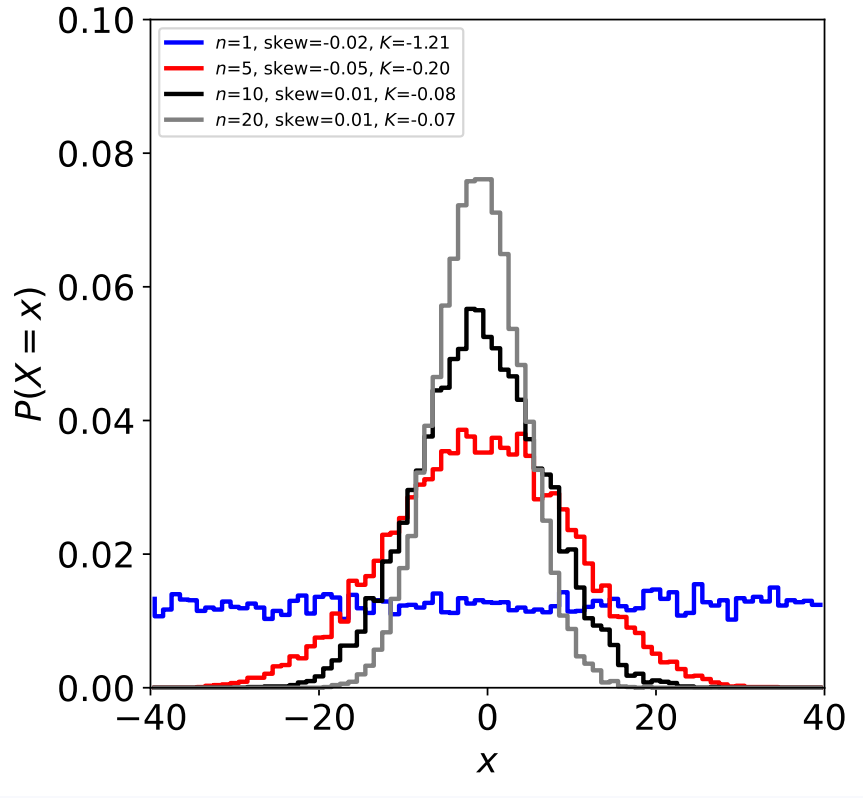
How Can We Spot Normal?

- One way: plot a normal curve on top of the histogram
- Another way: a Q-Q (quantile-quantile) plot
  - Plots quantiles of datasets against each other
  - Here: actual values of sample means vs normal distribution with same μ, σ
  - If distributions are similar, should make a straight line
- As n increases, dots get closer to the line (becomes more normal)



Skew and Kurtosis

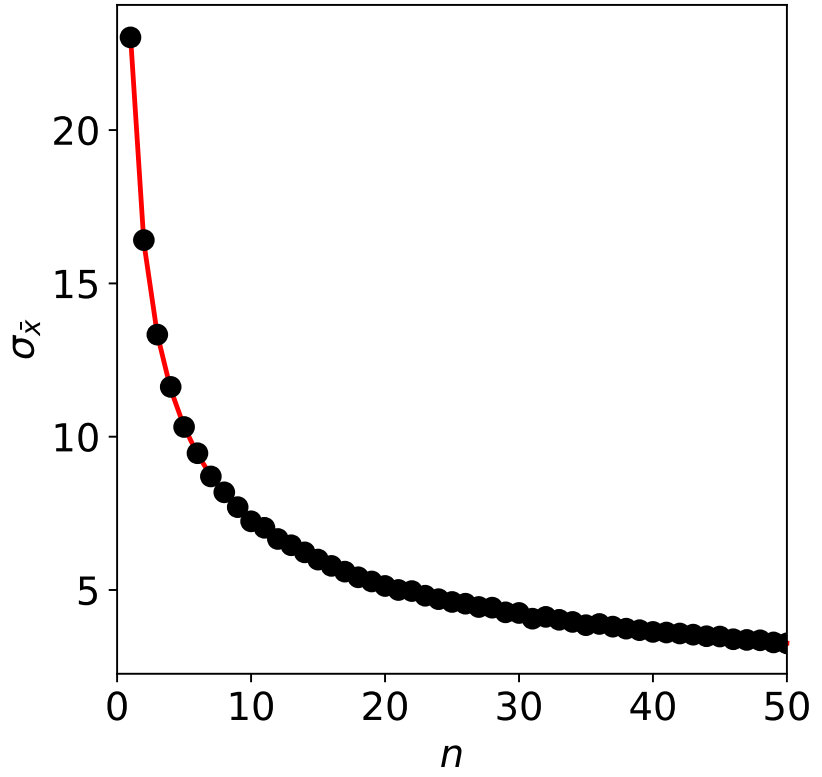
- We've seen skew before (W5, P6)
  - Measure of whether data "leans" relative to symmetrical
  - Positive: "right-skew" (leans left)
  - Negative: "left-skew" (leans right)
- Kurtosis:
  - Measure of how much of the distribution is in the "tails" vs the "peak", compared to normal distribution
  - Positive: heavier peak
  - Negative: heavier tails



Standard Error of the Mean

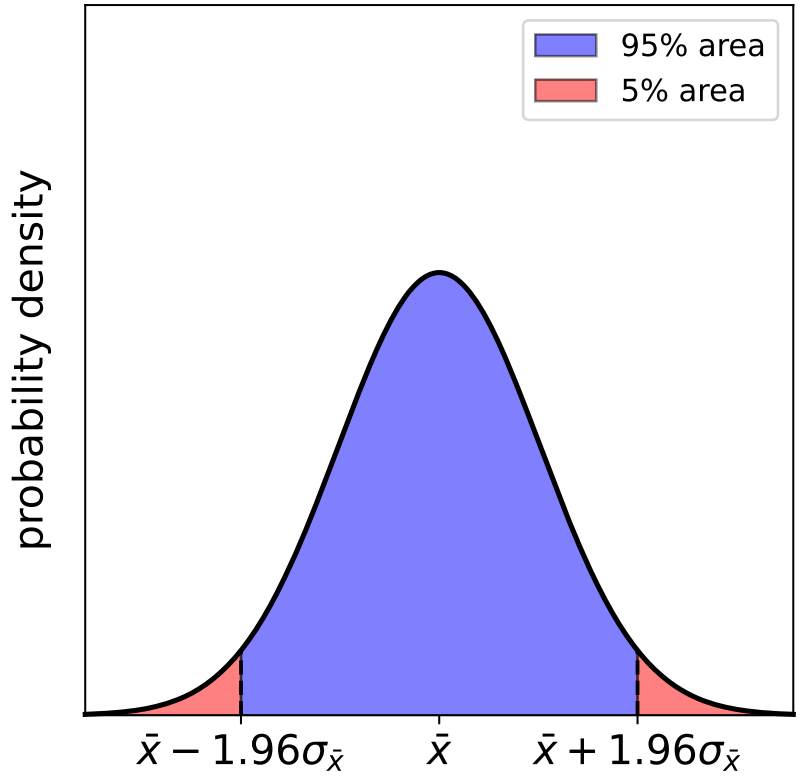
- We still have some dispersion (spread) of the mean of the sample means around μ
- But, as n increases:
  - Mean of sample means gets closer to μ
  - Dispersion decreases
- Standard error of the sample mean, $\sigma_{\bar{x}}$:

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



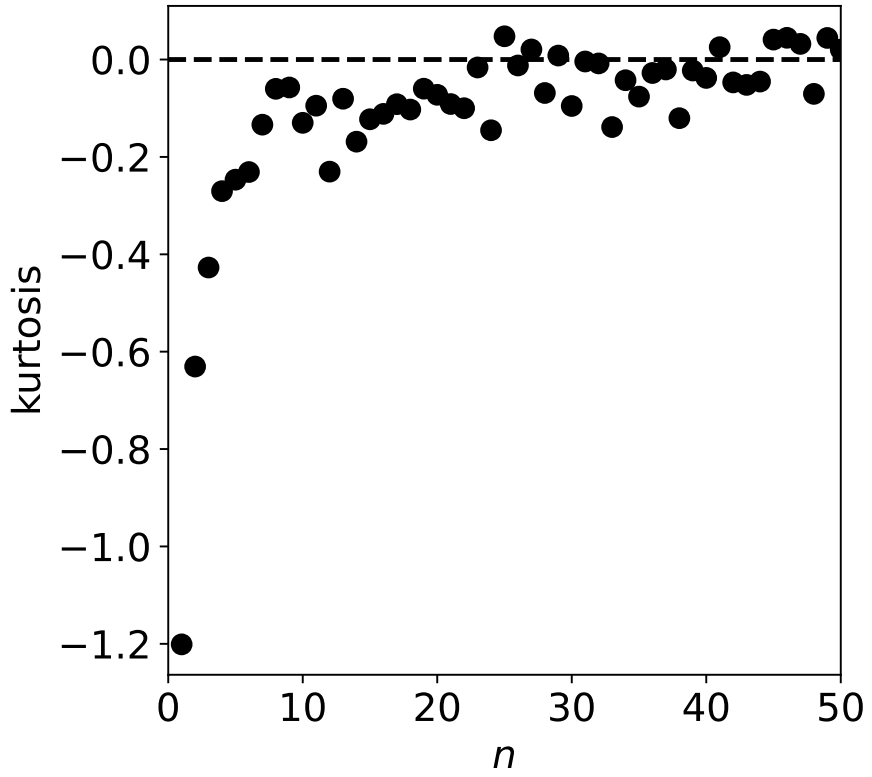
Confidence intervals

- By the empirical rule:
  - For 95.45% of samples, sample mean is within two standard errors of μ
  - For 95% of samples: 1.96 standard deviations
- In other words, there is a 95% probability that μ is within the interval (range) from $[\bar{x} - 1.96\sigma_{\bar{x}}, \bar{x} + 1.96\sigma_{\bar{x}}]$
  - Only a 5% chance that μ is outside of this range
- We can say, "with 95% confidence", that μ is within $\bar{x} \pm 1.96\sigma_{\bar{x}}$
  - Can construct similar intervals for 90%, 99%, 99.5%, 99.9%, etc.



Consequences of the CLT

- We can use hypothesis tests (next week!) with distribution of sample means of non-normal data if:
  - Sample size is “large enough”
  - “Large enough”: typically $n \geq 30$
- As n increases, sampling error decreases
  - In other words, average of many measurements is more accurate than a single measurement



Summary

- The CLT tells us that the distribution of sample means from a population approximates a normal distribution
- This means that with a “large enough” sample size, we can treat non-normal data as if it were normal
- We can also use CLT to estimate sampling error
- The average of many estimates is more accurate than a single estimate

Additional Resources

- Illowsky and Dean, Chapter 7
- Caswell, Chapter 14.1–14.2
- Weiss, Chapters 7.2–7.3
- Central Limit Theorem Explained [Jim Frost]
- Assessing Normality [Jim Frost]
- Central Limit Theorem [Khan Academy]
- Sampling distribution of the sample mean [Khan Academy]
- onlinestatbook.com sampling distribution simulator