



Taming the beast

How to assume normality
(or not) and why

The central limit theorem

- Regardless of the distribution of the population (normal or non-normal), the sampling distribution of the sample mean (SDSM) is approximately normal when n is 30 or more
- If the SDSM is normal, parametric tests can be used
- *You do not need to understand this to know that it is true, but it helps*

SDSM: an example

Player 1	Player 2	Player 3	Player 4	Player 5
76 inches	78 inches	79 inches	81 inches	86 inches

What is μ ? It = 80 inches

Let's take a sample of 2 randomly

We choose players 2 & 5. What is \bar{x} ? It = 82 inches

Can you see that \bar{x} is an estimate of μ ?

If we take all possible samples of 2 and determine \bar{x} for each one; we have the sampling distribution of the sample means.

What would happen if our sample was larger?

SDSM: example, cont'd

Sample	Heights	x-bar
1,2	76, 78	77.0
1,3	76, 79	77.5
1,4	76, 81	78.5
1,5	76, 86	81
2,3	78, 79	78.5
2,4	78, 81	79.5
2,5	78, 86	82.0
3,4	79, 81	80.0
3,5	79, 86	82.5
4,5	81, 86	83.5

The sampling distribution of the sample means

It is all possible sample means for sample size of 2

These are all estimates of μ ; only one of the x-bars is an exact estimate (sample 3,4)

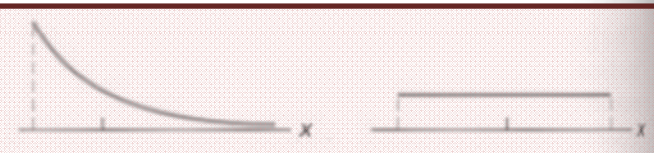
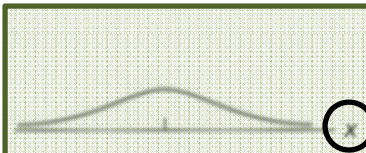
We are very interested in the distribution of these estimates because all inferential tests compare estimates (often of μ)

The Central Limit Theorem

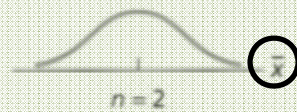
Normal Population

Non-normal populations

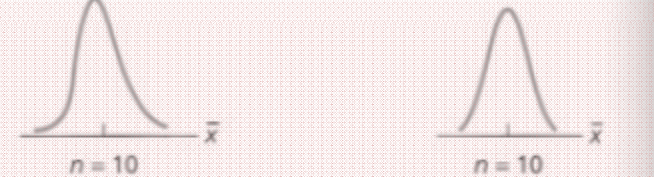
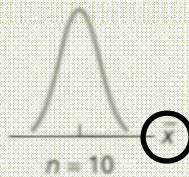
Population distribution



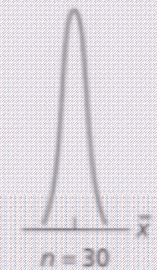
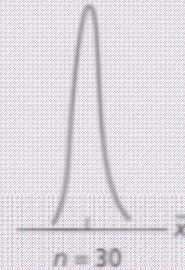
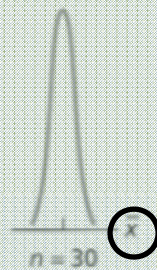
SDSM at $n = 2$



SDSM at $n = 10$



SDSM at $n = 30$



(a)

(b)

(c)

Assuming Normality w/ Samples from Normal Populations

Normality Tests

- A representative sample should pass the Shapiro Wilks W test for normality
- If it comes from a normal population, and if it is representative, it should be shaped like a normal distribution

Central Limit Theorem

- Regardless of sample size samples from normal populations always produce normal sampling distributions of the sample means
- This means that normality can be assumed from normal populations at any sample size
- The trick is to know if your sample came from a normal population

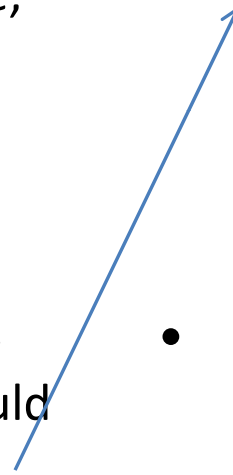
Assuming Normality w/ Samples from Non-Normal Populations

Normality Tests

- If the sample is representative, and if it is from a non-normal population it should fail a normality test
- You would choose to use non-parametric tests, and you would be WRONG to do so because...

Central Limit Theorem

- Regardless of the distribution of the population, the SDSM is approximately normal when n is 30 or more
- If the SDSM is normal, parametric tests can be used



Assuming Normality

Normality Tests

- These tests compare the z-scores for a samples to those expected from a normal curve

- In representative samples, if the sample is shaped like a normal curve, then it probably comes from a normal population
 - The first part of the Central Limit Theorem holds

The Central Limit Theorem

- In representative samples from normal populations, normality can be assumed at any sample size

- In representative samples from populations that are not known to be normal, at $n \geq 30$ normality can be assumed
 - Why? Because, regardless of the shape of the population, the sampling distribution of the sample means is normal at $n \geq 30$

How to assume normality (or not)

- 1) Know the sampling design
 - Is your sample representative?
- 2) What is the sample size?
 - If **representative** & **$n \geq 30$** assume normality
- 3) If **$n < 30$ but representative by design** do you know if your sample is from a normal population?
 - If you do not know, but you are sure the sample is representative, do a normality test
 - If the small sample passes the test, assume normality
- 4) If samples are small and do not pass, use non-parametric statistics
- 5) If you do not have representative samples, stop and start over