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

<table>
<thead>
<tr>
<th>Player 1</th>
<th>Player 2</th>
<th>Player 3</th>
<th>Player 4</th>
<th>Player 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>76 inches</td>
<td>78 inches</td>
<td>79 inches</td>
<td>81 inches</td>
<td>86 inches</td>
</tr>
</tbody>
</table>

What is \( \mu \)? It = 80 inches

Let’s take a sample of 2 randomly

We choose players 2 & 5. What is \( x \)-bar? It = 82 inches

Can you see that \( x \)-bar is an estimate of \( \mu \)?

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

What would happen if our sample was larger?
SDSM: example, cont’d

The sampling distribution of the sample means

It is all possible sample means for sample size of 2

These are all estimates of $\mu$; 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 $\mu$)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Heights</th>
<th>x-bar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2</td>
<td>76, 78</td>
<td>77.0</td>
</tr>
<tr>
<td>1,3</td>
<td>76, 79</td>
<td>77.5</td>
</tr>
<tr>
<td>1,4</td>
<td>76, 81</td>
<td>78.5</td>
</tr>
<tr>
<td>1,5</td>
<td>76, 86</td>
<td>81</td>
</tr>
<tr>
<td>2,3</td>
<td>78, 79</td>
<td>78.5</td>
</tr>
<tr>
<td>2,4</td>
<td>78, 81</td>
<td>79.5</td>
</tr>
<tr>
<td>2,5</td>
<td>78, 86</td>
<td>82.0</td>
</tr>
<tr>
<td>3,4</td>
<td>79, 81</td>
<td>80.0</td>
</tr>
<tr>
<td>3,5</td>
<td>79, 86</td>
<td>82.5</td>
</tr>
<tr>
<td>4,5</td>
<td>81, 86</td>
<td>83.5</td>
</tr>
</tbody>
</table>
The Central Limit Theorem

Population distribution

SDSM at $n = 2$

SDSM at $n = 10$

SDSM at $n = 30$

Normal Population

Non-normal populations

(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 sample 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 if 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