Comparing 2 Samples Cont.

Review case study 15 model for tribe 1:

- Population: all tribe 1 females at relevant time
  - Height
  - Mean $\mu_1 = \bar{\mu}$
  - SD $\sigma_1 = \sigma$
  - Pop. histogram

- Sample: the observed skeletons
  - Height
  - $\bar{y}_1 = 59.4$ in.
  - $s_1 = 1.8$ in.

- Imaginary data set: possible $y_1$'s
  - $\bar{y}_1 = 59.4$ in.
  - $s_1 = 1.8$ in.

Long run mean: $E(\bar{y}_1) = \mu_1$
Est. long run SD: $\text{SE}(\bar{y}_1) = 0.36$ in.

Inferential Summary: Tribe 1

<table>
<thead>
<tr>
<th>Quantity of Interest</th>
<th>Estimate</th>
<th>Give-or-take</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$ = mean height of all adult females in tribe 1</td>
<td>$\bar{y}_1 = 59.4$ in</td>
<td>$s_{\bar{y}_1} = \frac{1.8}{\sqrt{25}} = 0.36$ in</td>
</tr>
</tbody>
</table>

Model for Tribe 2:

- Population: same (Tribe 2)
  - Height
  - Mean $\mu_2 = \bar{\mu}$
  - SD $\sigma_2 = \sigma$
  - Pop. histogram

- Sample: same as above
  - Height
  - $\bar{y}_2 = 61.3$ in
  - $s_2 = 2.4$ in

- Imaginary data set: possible $y_2$'s
  - $\bar{y}_2 = 61.3$ in

Long run mean: $E(\bar{y}_2) = \mu_2$
Est. long run SD: $\text{SE}(\bar{y}_2) = 0.46$ in

19 May 2005
Inferential summary: Tribe 2

<table>
<thead>
<tr>
<th>Quantity of Interest</th>
<th>Estimate</th>
<th>Give-or-take</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_2 )</td>
<td>Same as above (Tribe 2)</td>
<td>( \bar{y}_2 = 61.3 \text{ in.} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( SE_{\bar{y}}(\bar{y}_2) = \frac{s_2}{\sqrt{n_2}} = 2.74 \sqrt{\frac{1}{27}} = 0.40 \text{ in.} )</td>
</tr>
</tbody>
</table>

Revised inferential summary

<table>
<thead>
<tr>
<th>Quantity of Interest</th>
<th>Estimate</th>
<th>Give-or-take</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (M_2 - M_1) )</td>
<td>Mean diff. between adult female height in tribe 2 vs. tribe 1</td>
<td>( (\bar{y}_2 - \bar{y}_1) = 61.3\text{ in.} - 59.4\text{ in.} = 1.9 \text{ in.} )</td>
</tr>
<tr>
<td>Is diff. pract. sig.?</td>
<td>Yes, most of tribe 2 women would be taller than tribe 1 women.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( SE(\bar{y}_2 - \bar{y}_1) = 0.40 \text{ in.} )</td>
<td></td>
</tr>
<tr>
<td>95% C.I. for ( (M_2 - M_1) )</td>
<td>( (\bar{y}_2 - \bar{y}_1) \pm 2SE = 1.9 \pm 2(0.4) = (0.7, 3.1 \text{ in.}) )</td>
<td></td>
</tr>
</tbody>
</table>

Math fact:

Uncertainty combines with 2 independent samples like legs of a right triangle.

\[ SE(\bar{y}_1 - \bar{y}_2) = \sqrt{(SE(\bar{y}_1))^2 + (SE(\bar{y}_2))^2} = \sqrt{(0.30 \text{ in.})^2 + (0.40 \text{ in.})^2} = 0.58 \text{ in.} \approx 0.6 \text{ in.} \]

\[ SE(\bar{y}_2 - \bar{y}_1) = SE(\bar{y}_1 - \bar{y}_2) = \sqrt{SE(\bar{y}_1)^2 + SE(\bar{y}_2)^2} \]

for 2 independent samples

Long run histogram of \( \bar{y}_2 - \bar{y}_1 \):

95% C.I. for \( M_2 - M_1 \): (0.7, 3.1 in.)
$H_0$: (no real diff.) $\mu_2 - \mu_1 = 0$

$\Rightarrow \emptyset$ is not in the 95\% C.I. so $H_0$ is wrong = Stat. Sig.

**BUT** is her data gathering method like SRS or might it instead be biased?

$\Rightarrow$ size-biased (length-biased) sampling: big things are easier to find than small things

$\Rightarrow$ so her estimate $\bar{y}_1$ of $\mu_1$ is likely to have been biased on the high side, as well as for $\bar{y}_2$ as an estimate of $\mu_2$; but this bias should (largely) cancel in looking at the difference $(\bar{y}_2 - \bar{y}_1)$ as an estimate of $(\mu_2 - \mu_1)$, so it is okay.

* Biased measurement methods are bad for 1 sample at a time but may well be okay for comparing 2 Samples.

**Case Study 16**

model - 2 independent samples w/ 1's and 0's

"black" "white"

$\begin{bmatrix} 1's \\ 0's \end{bmatrix} n_1 = 100$

$\begin{bmatrix} 1's \\ 0's \end{bmatrix} n_2 = 900$

Black model

$I = yes, O = no$

pop.

all blacks in US in 1977

favor $\frac{2}{3}$

mean $\hat{p}_i = $ ?

pop hist.

$\begin{bmatrix} 1's \\ 0's \end{bmatrix} n_1 = 100$

favor $\frac{2}{3}$

mean $\hat{p}_i = 27\%$

imaginary data set

possible $\hat{p}_i$'s

[27\%] $\uparrow$

[30\%] $\downarrow$

long run mean $E(\hat{p}) = P_i$

est. long run SD $\hat{SE}(\hat{p}) = 4.4\%$

long run hist. 8

CLT

$P_1$
\[ * \hat{SE}_{\text{IID}} (\hat{p}_2) = \sqrt{\frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} = \sqrt{\frac{0.08(1 - 0.08)}{900}} = 0.009 \text{ or } 0.9\% \]

**Inferential Summary (real)**

**Quantity of Interest**: \( (\hat{p}_1 - \hat{p}_2) = \text{diff. in prop. btwn "black" and "white" on opinion of pref. treatment} \)

**Estimate**: \( (\hat{p}_1 - \hat{p}_2) = 27\% - 8\% = 19\% \)

**Pract. Sig.?**: Yes, \( \frac{27}{8} = 3.4 \) times greater \% for "black"

**Give-or-take**: \( SE(\hat{p}_1 - \hat{p}_2) = 4.5\% \)

**95\% CI**: \( (\hat{p}_1 - \hat{p}_2) \pm 2SE = 19\% \pm 2(4.5\%) = 19\% \pm 9\% \)

**\( SE_{\text{IID}} (\hat{p}_1 - \hat{p}_2) = \)**

\[ = \frac{(4.4)^2 + (0.9)^2}{100} = 4.5\% \]

\( H_0: \theta \text{ not near} \)

**Stat. Sig. interval so**

\[ = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \]