
Case Study 10 continued

2 survival skills concerning the statistical models diagrams:

1. Which data set (population, sample, imaginary data set) is being referred to?

2. Is the question about individual elements of the data set or the summary?

(examples:

1. (True or False) About 68% of waybills in the sample were in the range $28.09 ± $0.69
   
   False because it's a statement about individual elements of the sample dataset, correct given or take for elements of sample dataset is sample SD ($31.40).
   
   To make true, should replace $0.69 by $31.40.

2. (True or False) About 68% of waybills in population were in the range of $28.09 ± $0.69
   
   Yes. No.
   
   False because the elements in the population data set differ from mean μ by an amount roughly σ in size, and our best estimate of σ is s = $31.40.

(End of Case Study 10)

Case Study 11:

Devil's advocate says: "I think the real % of all 17-year-olds who can add fractions is 90% and you really different answers by unlucky sampling."

*plausible and in need of debating
population of all 17-year olds in U.S. in 1975 can they add fractions?

sample of observed 17-year olds can they add fractions?

imaginary dataset possible p's (estimates of p)

N = very large

mean \( \mu = p = ? \) 
SD \( \sigma = \)

hypothesized

\[
\begin{bmatrix}
15 \\
0
\end{bmatrix}
\] 
\( n = 950 \)

mean \( \bar{y} = \hat{p} = \frac{617}{950} \approx 65\% \)

\[
\begin{bmatrix}
15 \\
0
\end{bmatrix}
\] 
\( n = 950 \)

mean \( \hat{p} = 62\% \)

Need to find out: long run mean? p
long run SD? 
long run histogram?

This is a statistical model because reasoning backwards from sample to the population.

### Inferential Summary

<table>
<thead>
<tr>
<th>(unknown) quantity of interest</th>
<th>( p = ) population % of all U.S. 17-year olds who can add fractions in 1975</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate</td>
<td>( \hat{p} = 65% )</td>
</tr>
<tr>
<td>what give or take?</td>
<td>( \hat{p} \pm 2 \hat{S}(\hat{p}) ) = from 62% to 68%</td>
</tr>
<tr>
<td>95% CI interval estimate</td>
<td>( \hat{p} \pm 2 \hat{S}(\hat{p}) ) = from 62% to 68%</td>
</tr>
</tbody>
</table>
Long Run mean: $E_{\hat{p}}(\hat{p}) = E_{\hat{y}}(\hat{y}) = \mu = p \rightarrow E(\hat{p}) = p$

Long Run SD: $SE_{\hat{p}}(\hat{p}) = SE_{\hat{y}}(\hat{y}) = \frac{\sigma}{\sqrt{n}} = \frac{SE(p)}{\sqrt{n}} = \sqrt{p(1-p)} \ln = \sqrt{p(1-p)} \mathbf{\left< \text{math fact: SD of a population with 0's and 1's is } \sqrt{p(1-p)} \text{ divided by square root of fraction of 0's and 1's} \right>}$

$SE_{\hat{p}}(\hat{p}) = \sqrt{p(1-p)} \rightarrow \text{not usable b/c } p \text{ is unknown}$

Fix: $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.65)(1-0.65)}{950}} \times 100 \% = 0.015 \times 100 \% = 1.5 \%$

Long Run histogram of $\hat{p}$:

$\hat{p}$ is like a $\hat{y}$ $\rightarrow$ CLT applies to $\hat{y}$ when $n$ is large, and 95% is plenty large.

They're quite confident (95%) that $\hat{p} + p$ will differ by no more than 3%, so a 95% confidence interval (CI) for $p$ looks like $\hat{p} \pm 2SE(\hat{p}) = 0.65% \pm 3%$

90% figure is completely implausible.

I'm 95% confident that $p$ is in this interval.

To be 95% confident that $\hat{p}$ and $p$ will differ by no more than 2%, $\pm 2SE \rightarrow \pm 2SE$ must solve for how big $n$ must be.
2 \times SE = 2% = 2 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}

\frac{2\%}{2} = \frac{0.02}{2} = \sqrt{\frac{0.05(1-0.05)}{n}}

0.01^2 = \left(\sqrt{\frac{0.05(0.35)}{n}}\right)^2

n = \frac{0.05(0.35)}{0.01(0.01)} = 2275 \quad \text{We would need to sample 2275 people}

\text{(end of case study 10)}

P.88 "Byrne trims Daley lead in Chicago poll" article

If this method is like SRS (which seems likely), \( SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \)

\text{worst case SE is } \underbrace{SE_{\text{worst}}(\hat{p}) = \sqrt{\frac{0.5(0.5)}{n}}}_{\text{case}}

\downarrow

\underbrace{SE_{\text{worst}}(\hat{p}) = \frac{50\%}{\sqrt{4/11}}}_{\text{well run poll}}