May 3, 2005 Lecture Notes Building a Statistical Model, Point and Interval Estimates

Case Study 10:

statistical model

population: all 22,984 waybills

amount owed

$N = 22,984$

\$14.15
\$100
⋮

mean $\mu = ?$

SD $\sigma = ?$

SRS

sample: the observed waybills

amount owed

$n = 2,072$

mean $\bar{y} = \$28.09$

SD $s = \$31.40$

hypothetical SRS

$n = 2,072$

mean $\bar{y} = ?$ (ex: \$28.42)

imaginary data set: possible $\bar{y}$'s

\$28.09 (a $\bar{y}$)
\$28.42

∞

Need to find:

Long run mean $\bar{y}$:

Long run mean SD $\bar{y}$

Long run histogram of $\bar{y}$

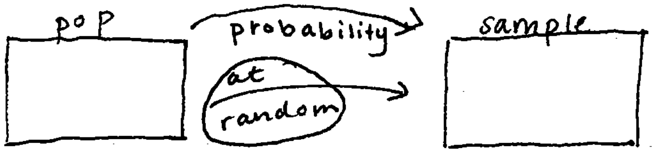
population histogram:

(could possibly look like this)

Probability model vs. Statistical model:

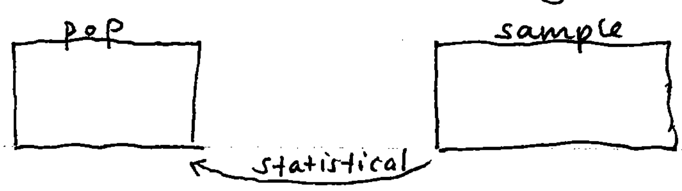
If the population is known, we reason from the population to the sample (whole → part) deductive reasoning

If the sample is drawn at random, this is probability in action



If instead the sample is known and chosen randomly, we reason from the sample to the population

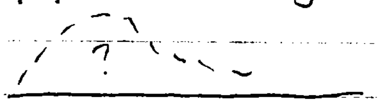
(part → whole) inductive reasoning (Statistical Inference)



sample SD ($s = \$31.40$) is bigger than sample mean $\bar{y} = \$28.09$, and this variable can't go negative

↳ population histogram has a long right hand tail

population histogram:



(as shown before)

a good single # guess (point estimate) for μ based on sample is $\bar{y} = \$28.09$, but give or take how much?

Inferential Summary (table)

quantity of interest (unknown)	μ = mean amount of money owed in population waybills
estimate	$\bar{y} = \$28.09$
give or take	$\widehat{SE}(\bar{y}) = \$0.69$
95% Confidence Interval (CI) for μ	$\bar{y} \pm 2\widehat{SE}(\bar{y})$ = from \$26.71 to \$29.47 (\$26.71, \$29.47)

long run mean $\bar{y}$ in imaginary data set:

expected value of $\bar{y}$ = $EV = E(\bar{y}) = \boxed{E_{SRS}(\bar{y}) = \mu}$ ← math fact ($\bar{y}$ is an unbiased estimate of μ)

long run SD of $\bar{y}$'s in imaginary data set:

standard error of $\bar{y}$ = $SE = SE(\bar{y}) = SE_{SRS}(\bar{y}) = ?$

$SE_{IID}(\bar{y}) = \frac{\sigma}{\sqrt{n}}$

$SE_{SRS}(\bar{y}) = [SE_{IID}(\bar{y})] \cdot \underbrace{\left[\sqrt{\frac{N-n}{N-1}}\right]}_{\text{Correction Factor (CF) for SRS}}$

↑ more info than IID, thus requires more

<u>Correction Factor</u> is revelant for HW4 #1, not relevant for any other problems in this class. In other words, in all other problems, $n << N$ and therefore $CF = 1 \rightarrow SRS = IID$

$SE_{IID}(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ (unusable when sample is known and population is unknown)

<u>to fix</u>: s is a good estimate of σ, so use estimated SE of $\bar{y}$ =

$$\hat{SE}_{IID}(\bar{y}) = \frac{\cancel{\sigma}\, s}{\sqrt{n}} = \boxed{\hat{SE}_{IID}(\bar{y}) = \frac{s}{\sqrt{n}}} \leftarrow \text{important formula}$$

$$\hat{SE}_{IID}(\bar{y}) = \frac{\$31.40}{\sqrt{2072}} = \$0.69$$

$$SE_{SRS} = SE_{IID}(\bar{y}) \cdot (CF)$$

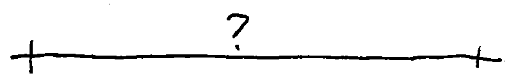
$$\hat{SE}_{SRS}(\bar{y}) = \frac{s}{\sqrt{n}} \cdot \left(\sqrt{\frac{N-n}{N-1}}\right) = \left(\frac{\$31.40}{\sqrt{2072}}\right) \cdot \left(\sqrt{\frac{22984-2072}{22984-1}}\right)$$

$$= \$0.69 \cdot (0.95) = \$0.66$$

However, because the CF (Correction Factor) is close enough to one, ~~and~~ we don't need to calculate it and will use \$0.69 for $SE(\bar{y})$

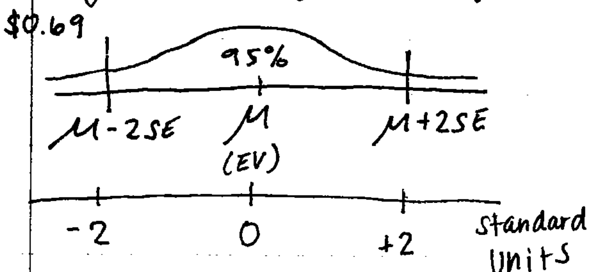
<u>meaning</u>: I think that μ is around \$28.09 ($\bar{y}$) give or take about \$0.69 ($\hat{SE}(\bar{y})$)

we want an interval along the number line in which μ can probably be found.



<u>Long Run Histogram of $\bar{y}$:</u>

SE = \$0.69



95% likely that $\bar{y}$ and μ will differ by no more than $2SE = \$1.38$

$$\bar{y} \pm 2SE(\bar{y})$$

$$\$28.09 \pm 2(\$0.69) = \$28.09 \pm \$1.39$$

$$(\$26.71, \$29.47)$$