Confidence Intervals

Here we want to draw conclusions about the <u>pop</u> from the <u>sample</u>

<u>Unbiased Estimates:</u> $E(f_{\bar{x}}) = E(f) = \mu$ (centered at same value)
so $\bar{x}$ is an <u>unbiased</u> estimator of μ. a.k.a. $\hat{\mu} = \bar{x}$

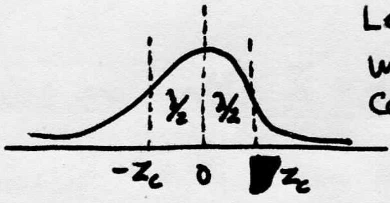
<u>Efficient Estimates:</u> If the sampling distribs of 2 statistics have the Same Expected value (center) then the one with the smaller var. is more <u>efficient</u>.

<u>Consistent Estimate:</u> As sample size increases, the statistic converges <u>in probability</u> to pop value.

<u>Confidence Interval for μ</u>

we know $f_{\bar{x}} \approx f_{N(\mu,\sigma_{\bar{x}}^2)}$ thus we do calculations with N distrib.

How do we derive a confidence interval for μ? Say we want 95% CI.
Let $\lambda = 0.95$.
we want the area of the center strip in $N(0,1)$ to be λ.
By symm, there are 2 identical areas of size $\lambda/2$.



$$Z = \frac{\bar{x} - E(\bar{x})}{\sqrt{\mathrm{var}(\bar{x})}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

[If $\lambda = 0.95$, we want area 0.4750 on each side,
so we seek $Z_c \ni Q(Z_c) = 0.4750 \Rightarrow Z_c = 1.96$
↖ value we look up in table.]

so we'd have

$$-Z_c < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < Z_c$$

$$-Z_c\sigma_{\bar{x}} < \bar{x} - \mu < Z_c\sigma_{\bar{x}}$$

$$-Z\sigma_{\bar{x}} - \bar{x} < -\mu < Z_c\sigma_{\bar{x}} - \bar{x} \Rightarrow \bar{x} - Z_c\sigma_{\bar{x}} < \mu < \bar{x} + Z_c\sigma_{\bar{x}}$$

$$\Rightarrow \boxed{\mu \in (\bar{x} - Z_c\sigma_{\bar{x}},\ \bar{x} + Z_c\sigma_{\bar{x}})}$$

$s_{\bar{x}}$

Here we can substitute $\sigma_{\bar{x}} = \frac{1}{\sqrt{N}}\sigma$ or $\sqrt{\frac{N_p - N}{N-1}}\frac{1}{\sqrt{N}}\sigma$

<u>Common values for Z_c</u>

95%	1.96
99%	2.58
99.97%	3.0

This is interpreted to mean $\mu \in (\bar{x} - 1.96\sigma,\ \bar{x} + 1.96\sigma)$ with Prob = 0.95.

<u>Objections</u>: Neyman said it was only true that if you computed many samples $\bar{x}$ and formed these intervals that 95% of the time μ would be in an interval.
He is saying that once a specific $\{x_1 \dots x_N\}$ are chosen, $\bar{x}$ is determined and there is no randomness - either $\mu \in CI$ or $\mu \notin CI$ with certainty (Prob = 1).
(This is like saying <u>before</u> I flip a coin, P(heads) = 1/2 but after I flip it, (even if I don't look) P(heads) = 1 or 0.)

E. Lehmann
'Fisher, Neyman and Creation of Classical Stats' p. 89

I don't see this as a problem. We compute $\bar{x}$ and form CI and there is a 95% chance that this is a good $\bar{x}$ and thus $\mu \in CI$.

<u>The only problem I see</u> is that this holds <u>for any</u> $\bar{x}$ I compute. $\bar{x}_1$ and $\bar{x}_2$ may have very different values, but there is no way to distinguish which is better (unless we repeatedly do it and try to compute $f_{\bar{x}}$ and see its center!)

⑧ Given $\sigma = 0.05$, how large must sample size N be to make the 95% CI bounds ≤ 0.01? How about for 99% CI?

$\bar{X} \pm z_c \frac{\sigma}{\sqrt{N}}$ so we want $z_c \frac{\sigma}{\sqrt{N}} \leq 0.01 \Rightarrow N \geq \left(\frac{z_c \sigma}{0.01}\right)^2$

For 95% $z_c = 1.96$ $N \geq \left(\frac{1.96(0.05)}{0.01}\right)^2 = 96.04 \Rightarrow N > 97$

99% 2.58 $N \geq \left(\frac{2.58(0.05)}{0.01}\right)^2 = 166.41 \Rightarrow N > 167$

⑩ In a sample of $N = 100$, 55 people said they would vote for the candidate. What is a 95% CI for the proportion of the pop who will vote for the candidate?

From ch 8, we know proportions are really $\bar{X} = \frac{1}{N}\sum^{N} X_i$ $X_i = \begin{cases} 1 \\ 0 \end{cases}$

Estimate of population prob from sample: $\hat{p} = 0.55$ we have $\hat{p} = \bar{X}$

$\mu_{\hat{p}} = \mu_{\bar{X}} = \mu$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}}$

$$-z_c \leq \frac{\hat{p} - \mu_p}{\sigma_p} \leq z_c$$

$\Rightarrow \hat{p} - z_c \sigma_p < \mu_p < \hat{p} + z_c \sigma_p$

$\mu_p \in \left(0.55 - \underbrace{1.96(0.0497)}_{0.1},\ 0.55 + \underbrace{1.96(0.0497)}_{0.1}\right)$

we have to use $\hat{p}$ to estimate σ_p

$\sigma_p = \sqrt{\frac{(0.55)(0.45)}{100}} = 0.0497$

For 99% CI,

$\mu_p \in (0.55 - 2.58(0.0497),\ 0.55 + 2.58(0.0497))$

⑪ For prob ⑩, how large a sample size N should we take so that we have $\begin{cases} 95\% \\ 99\% \end{cases}$ CI that the candidate will get elected?

To win, the candidate must recieve $> 50\%$ of vote in pop.

$\hat{p} - z_c \sigma_{\hat{p}} = \hat{p} - z_c \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{N}} \overset{!}{>} 0.50$ $\sqrt{p(1-p)} = \sqrt{.55(.45)} = 0.4794$

$= 0.55 - z_c \frac{(0.4794)}{\sqrt{N}} > 0.5$

$0.05 > \frac{z_c(0.4794)}{\sqrt{N}} \Rightarrow \sqrt{N} > z_c \frac{(0.4794)}{0.05} \Rightarrow N > z_c^2 \cdot 99$

plug in $1.96 \rightarrow N > 380.3$

$2.58 \Rightarrow N > 659$

[This is a bit unrealistic though, because it says $\hat{p} = 0.55$ remains fixed even as we change the sample size N]

(12) If we have binomial form $\sigma_p = \sqrt{\frac{p(1-p)}{N}}$ we can give a more elaborate formula for the CI, which reduces to the usual one for large N.

we know $-z_c \le \frac{\hat{p}-\mu}{\sigma_p} \le z_c$ Rename $\hat{p} \leadsto Y$, $\mu \leadsto p$ $-z_c \le \frac{Y-p}{\sqrt{\frac{p(1-p)}{N}}} \le z_c$

$$\Rightarrow -z_c\frac{\sqrt{p(1-p)}}{\sqrt{N}} \le Y-p \le z_c\frac{\sqrt{p(1-p)}}{\sqrt{N}} \xrightarrow{\text{square everything}} z^2\frac{p(1-p)}{N} \le (Y-p)^2 \le z^2\frac{p(1-p)}{N}$$

$a<b<c \Rightarrow a^2<b^2<c^2$

so equality: $z^2\frac{p(1-p)}{N} = Y^2-2Yp+p^2$

$$\Rightarrow z^2p - z^2p^2 = NY^2 - 2NYp + Np^2$$

$$0 = \underbrace{NY^2}_{c} - \underbrace{(z^2+2NY)}_{b}p + \underbrace{(N+z^2)}_{a}p^2 \quad \text{Quadratic in } p$$

$$p = \frac{-b\pm\sqrt{b^2-4ac}}{2a} = \frac{(z^2+2NY)}{2(N+z^2)} \pm \frac{\sqrt{(z^2+2NY)^2-4(N+z^2)NY^2}}{2(N+z^2)}$$

$z^4 + 4NYz^2 + \cancel{4N^2Y^2} - \cancel{4N^2Y^2} - 4NY^2z^2$

$z^2(4NY-4NY^2+z^2)$

$$= \frac{\frac{1}{2N}}{\frac{1}{2N}}\;\frac{(2NY+z^2)\pm z\sqrt{4NY(1-Y)+z^2}}{2N+2z^2}$$

$$= \frac{Y+\frac{z^2}{2N}\pm\frac{z}{2N}\sqrt{\frac{4N}{4N^2}Y(1-Y)+\frac{z^2}{4N^2}}}{1+\frac{z^2}{N}}$$

Now rename Y back to $\hat{p}$, p back to μ_p

$$\mu_p = \frac{\hat{p}+\frac{z^2}{2N}\pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{N}+\frac{z^2}{4N^2}}}{1+\frac{z^2}{N}} \approx \hat{p}\pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Does this even have any usefulness?

For N large (they drop $\frac{1}{N}, \frac{1}{N^2}$ terms but keep $1/\sqrt{N}$)

▷ Confidence Intervals for Sums and Differences

(14) Brand A light bulbs: batch size $N_A = 150$, mean lifetime $\bar{X}_A = 1400$, sample std dev $S_A = 120$

Brand B: $N_B = 200$, $\bar{X}_B = 1200$, $S_B = 80$

Find {95%, 99%} CIs for the difference in lifetimes between the populations

General formula $\bar{X} \pm z_c\frac{\sigma}{\sqrt{N}}$

Following ch 8 (see prob 12) $\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_A^2}{N_A}+\frac{\sigma_B^2}{N_B}}$

So here: $(\bar{X}_A-\bar{X}_B) \pm z_c\sqrt{\frac{S_A^2}{N_A}+\frac{S_B^2}{N_B}} = \underbrace{(\bar{X}_A-\bar{X}_B)}_{1400-1200} \pm 1.96\sqrt{\frac{120^2}{150}+\frac{80^2}{200}}$

$= 200 \pm 22.2$

plug in 2.58 for 99%.

Thus we have $\mu_A - \mu_B \in (200-22.2,\ 200+22.2)$ with prob 95%.

□

(15) A random sample of 400 adults and another sample of 600 teens were asked if they liked a tv show. Construct a CI of 95% for the difference in proportion of the populations of adults and teens.

adults $N_A = 400$
liked show: $X_A = 100$
$\hat{p}_A = \frac{100}{400} = \frac{1}{4}$

Teens $N_T = 600$
liked: $X_T = 300$
$\hat{p}_T = \frac{300}{600} = \frac{1}{2}$

If we do $\hat{p}_A - \hat{p}_T = -\frac{1}{4}$

$$(\hat{p}_T - \hat{p}_A) \pm z_c \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{N_A} + \frac{\hat{p}_T(1-\hat{p}_T)}{N_T}}$$ From ch8 results

$$\left(\tfrac{1}{2} - \tfrac{1}{4}\right) \pm 1.96 \sqrt{\frac{\frac{1}{4}\cdot\frac{3}{4}}{400} + \frac{\frac{1}{2}\cdot\frac{1}{2}}{600}} \Rightarrow \tfrac{1}{4} \pm 0.0583$$ or

$\mu_T - \mu_A \in (0.19, 0.31)$

(16) EMF of batteries $\mu = 45.1$, $\sigma = 0.04$

Connect 4 batteries in Series

1 2 3 4

Let X_i = emf of battery i

Find a
(a) 95%
(b) 99%
(c) 99.73%
(d) 50%

CIs for total voltage output: $T = \sum^{n=4} X_i$

Referring to ch8 prob #5, we know the first step is to convert to a statement about $\bar{x}$

Here we know $T = n\bar{x}$ and we have our one trick $f_{\bar{x}} \approx f_{\mathcal{N}(\mu, \sigma_{\bar{x}})}$

All the prev problems have $-z_c \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le z_c$

but $$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{\frac{1}{n}\sum^n X_i - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\cancel{\frac{1}{n}}\left(\sum^n X_i - n\mu\right)}{\cancel{\frac{1}{n}}\left(\sqrt{n}\sigma\right)} \Rightarrow -z_c \le \frac{\sum X_i - n\mu}{\sqrt{n}\sigma} \le z_c$$

$$\Rightarrow \underbrace{n\mu}_{\substack{4\cdot 45.1 \\ 180.4}} - z_c \underbrace{\sqrt{n}\sigma}_{2\cdot 0.04} \le \sum X_i \le n\mu + z_c\sqrt{n}\sigma$$

95% CI $\Rightarrow 180.4 - 1.96(0.08) \le \sum^4 X_i \le 180.4 + 1.96(0.08)$

99% CI $\Rightarrow 180.4 \pm 2.58(0.08)$

99.73% $\quad 180.4 \pm 3\,(0.08)$

50% $\quad 180.4 \pm 0.6745(0.08)$

This is called 'Probable error'

□