Conf intervals

for htwt data from [illegible] SAMPLE NCSU data

histogram(~weight, data=htwt)

bimodal, not normal

mean(htwt) $\overline{\text{weight}} = 139.6$ sqrt(var(htwt)) 43.1221

?sample note replace = F is default

samplemean ← function(n, data = htwt[,"weight"])
{ mean(sample(data, n)) }

samplemean(20) 139.6 = mean(htwt[,"weight"])

choose(20, 9) 167960

ns ← as.matrix(rep(9, 10000))

xbars ← apply(ns, 1, samplemean)
hist(xbars)
mean(xbars) 139.4827 ish

sqrt(var(xbars)) 10.6714 ish $\left(\underbrace{\frac{43.1221}{\sqrt{9}} \sqrt{\frac{20-9}{20-1}}}_{\text{Rice 3rd pg 214}} = 10.9370 \right)$

quantile(xbars)

50% between 25% and 75%

quantile(xbars, c(0.025, 0.975))

95% ~~between~~ 2.5% 97.5%

xbars ← apply(ns, 1, samplemean)

quantile(xbars, c(0.025, 0.975))

(118.2222, 159.8889) ish

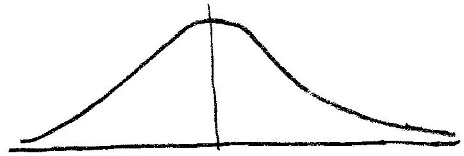
So 95% of the means from the sampling distribution of $\bar{x} = \hat{\mu}$ are between 118.2 and 159.9 ish

By CLT $\bar{X} \,\dot\sim\, N(\mu, \sigma^2/n) = N(139.6,\ 10.9370^2)$

$\sigma/\sqrt{n} = 14.3740 =$ Std error of mean $N = \infty$

$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 10.9370$ finite pop

Rice 3rd pg 214



139.6

$1.96(10.9370) = 21.4365 =$ margin of error $= ME_{\text{finite pop}}$

$\mu \pm 1.96\frac{\sigma}{\sqrt{n}} = 139.6 \pm 21.4365$ = param ± me

$(118.1635,\ 161.0365)$

Since normal $\Rightarrow P(-1.96 < z < 1.96) = .95$

pnorm(1.96) - pnorm(-1.96) .9500042

So 95% of the sample means are between

$(118.1635,\ 161.0365)$

table((xbars > 118.1635) & (xbars < 161.0365)) 4.18% 95.82%

Turn this inside out

95% of the $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ will include μ

est ± me

```
mean95ci <- function( mean = xbars, n=9, se = 43.1221/sqrt(n)*sqrt((N-n)/(N-1)), N=20)
     {  cilower <- xbars - 1.96 * se
        ciupper <- xbars + 1.96 * se
        cbind(cilower, ciupper)
     }
```

~~ci95 ← mean95ci(xbars, se)~~

~~table((ci95[,1] < 139.6) & (139.6 < ci95[,2])) 4.18% 95.82%ish~~

se ← sqrt(var(xbars)) 10.6436 ish $\approx \sqrt{\frac{43.1221}{9}}\sqrt{\frac{20-9}{20-1}} = 10.9370$

se < 14.3740 since oversampled

ci95 ← mean95ci(xbars)

table((ci95[,1] < 139.6) & (139.6 < ci95[,2])) 4.18% 95.82%ish go to CI example web page

We've found that a two sided

$100(1-\alpha)\%$ conf int for μ when σ is known

is $\bar{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$

if $n < \frac{N}{10}$

See NMS 6th p371

Approximate CI's for

p : $\frac{x}{n} \pm z_{\alpha/2} \frac{\sqrt{n(x/n)(1-x/n)}}{\sqrt{n}}$

$\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$p_1 - p_2$: $\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$ $\quad \hat{p} = \frac{x}{n}$ $\hat{q} = 1 - \frac{x}{n}$

assumes independent samples in two sample CI's

Mac 3rd pg 214

~~[illegible]~~

		mle	unbiased	
θ	sample size	$\hat{\theta}$	$E(\hat{\theta})$	$se(\hat{\theta})$
μ	n	$\bar{X}$	μ	$\sigma/\sqrt{n}$
p	n	$\frac{X}{n}$	p	$\sqrt{\frac{p(1-p)}{n}}$
$\mu_1-\mu_2$	n_1 and n_2	$\bar{X}_1-\bar{X}_2$	$\mu_1-\mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
p_1-p_2	n_1 and n_2	$\frac{X_1}{n_1}-\frac{X_2}{n_2}$	p_1-p_2	$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$

when $\sigma, \sigma_1, \sigma_2$ are known the $100(1-\alpha)\%$ CI's are

μ: $\quad \bar{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$

$\mu_1-\mu_2$: $\quad (\bar{x}_1-\bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$

if $\sigma, \sigma_1, \sigma_2$ unknown, then substitute s, s_1, s_2 and the ints are approx

Since we are estimating p, p_1, p_2 and their variances are based upon p, p_1, p_2 we get approximate $100(1-\alpha)\%$ CI's

p: $\quad \frac{x}{n} \pm z_{\alpha/2}\sqrt{\frac{(x/n)(1-x/n)}{n}}$

p_1-p_2: $\quad \left(\frac{x_1}{n_1}-\frac{x_2}{n_2}\right) \pm z_{\alpha/2}\sqrt{\frac{(x_1/n_1)(1-x_1/n_1)}{n_1}+\frac{(x_2/n_2)(1-x_2/n_2)}{n_2}}$

The $z_{\alpha/2} \cdot se(\hat{\theta})$ is called the <u>margin of error</u>.

Generalized approach for a $100(1-\alpha)\%$ CI

For the population mean

let $z_\alpha = 1 - \Phi^{-1}(\alpha)$ be a value such that $P(Z > z_\alpha) = \alpha$

$[Z \sim N(0,1)]$

Then $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$

By <u>CLT</u> $P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \approx 1-\alpha$

$\Rightarrow P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha$

For the population median, η

Form an interval $(X_{(k)}, X_{(n-k+1)})$ with $1-\alpha$ coverage

$P(X_{(k)} \le \eta \le X_{(n-k+1)}) = 1 - P(\eta < X_{(k)} \cup \eta > X_{(n-k+1)})$

$= 1 - [P(\eta < X_{(k)}) + P(\eta > X_{(n-k+1)})]$

Now $P(\eta > X_{(n-k+1)}) = \sum_{j=0}^{k-1} P(j \text{ obs are greater than } \eta)$

$n-k+1$ $k-1$
$n-1$ 1
n 0
X_i's $< \eta <$ X_i's
$X_i =$

$P(\eta < X_{(k)}) = \sum_{j=0}^{k-1} P(j \text{ obs are less than } \eta)$

$k-1$ $n-k+1$
0 n
X_i's $< \eta <$ X_i's
$X_i =$

By def $P(X_i > \eta) = P(X_i < \eta) = \frac{1}{2}$

Assume X_i's are iid

Then the number of obs greater than η is

$B(n, \frac{1}{2})$

Thus

$P(\eta > X_{(n-k+1)}) = \sum_{j=0}^{k-1} \overbrace{\binom{n}{j}\left(\frac{1}{2}\right)^j\left(\frac{1}{2}\right)^{n-j}}^{\sim}$

$= \binom{n}{j} 2^{-n}$

By Symmetry

$$P(X_{(k)} \le \eta \le X_{(n-k+1)}) = 1 - \sum_{j=0}^{k-1} \binom{n}{j} 2^{-n} - \sum_{j=0}^{k-1} \binom{n}{j} 2^{-n}$$

$$= 1 - 2\,(2^{-n}) \sum_{j=0}^{k-1} \binom{n}{j}$$

$$= 1 - 2^{-(n-1)} \sum_{j=0}^{k-1} \binom{n}{j}$$

$$= \sum_{j=0}^{k-1} \binom{n}{j} \frac{1}{2^n}$$

$$= P(Y \le k-1)$$

where $Y \sim B(n, \frac{1}{2})$

Ex: If $n = 9$ then for $Y \sim B(9, \frac{1}{2})$

k	$P(Y \le k)$		
0	$.0019 = P(Y<1)$	$(x_{(1)}, x_{(9)})$	is a $1-2(.0019) \times 100 = 99.61\%$ CI
1	$.01953 = P(Y<2)$	$(x_{(2)}, x_{(8)})$	96.09% CI
2	$.0898 = P(Y<3)$	$(x_{(3)}, x_{(7)})$	82.03% CI
3	$.2539 = P(Y<4)$	$(x_{(4)}, x_{(6)})$	49.22% CI

HTWT data (weight)

sort(sample(x,9))[c(2,8)] (119,199), (103,195), (87,191), (87,159)

median(x) 123.5