BIOS 662 Fall 2018

Point and Interval Estimation

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- Introduction
- Confidence intervals (CIs) for the mean
  - Parametric, large sample
  - Bootstrap
- CI for quantiles
  - Exact
  - Large sample
- CI for variance

Inference

- *Inference:* Using statistics and probability theory to draw conclusions about parameters
- Two modes of inference:
  - Estimation: attempt to estimate value of parameter(s) and quantify uncertainty about these estimate(s)
  - Hypothesis testing: posit certain values for parameters and test whether the observed data are consistent with the hypothesis

Estimation

- *Estimand*: parameter of interest we are trying to estimate; a constant; e.g. μ
- *Estimator*: the statistic used to estimate the estimand; a random variable; e.g. $\bar{Y}$
- *Estimate*: a realization of an estimator from an observed dataset; e.g. $\bar{y} = 36.3$

Estimating μ

- Suppose $Y_1, \ldots, Y_n$ is a random sample from a distribution with mean μ
- The estimator $\bar{Y}$ is an *unbiased* estimator of μ, i.e.,

$$E(\bar{Y}) = \mu$$

That is, the mean of the sampling distribution of $\bar{Y}$ equals μ, the population parameter of interest

Confidence Interval for μ

- Suppose $Y_1, \ldots, Y_n$ is a random sample from a normal distribution with mean μ and variance σ^2

- Then

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

(Result 4.2 in previous set of slides)

- We can use this to derive a *confidence interval* (CI) for μ

Confidence Interval for μ

- First define z_p such that

$$\Pr[Z \leq z_p] = p$$

for $Z \sim N(0,1)$; by symmetry, $z_p = -z_{1-p}$

- z_p is the p^{th} quantile of a standard normal distribution

Confidence Interval for μ

$$
\begin{aligned}
1-\alpha &= \Pr[-z_{1-\alpha/2} < Z < z_{1-\alpha/2}] \\
&= \Pr[-z_{1-\alpha/2} < \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}] \\
&= \Pr[-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{Y}-\mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}] \\
&= \Pr[-\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}] \\
&= \Pr[\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]
\end{aligned}
$$

Confidence Interval for μ

- $100(1-\alpha)\%$ CI for μ

$$\bar{Y} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$(\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- Values of $z_{1-\alpha/2}$

$100(1-\alpha)\%$	α	$z_{1-\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

CI Interpretation, Comment

- Text (p 86): The probability is $1-\alpha$ that the interval *straddles* the population mean μ
- If we draw 100 different random samples, on average $100(1-\alpha)\%$ of them will contain μ

- To decrease the width of CI:
  - increase α, i.e., decrease confidence
  - increase sample size

CI Example

- Example 4.8 of the text: SIDS birthweights
- $n = 78,\ \ \bar{Y} = 2994g,\ \ \sigma = 800g$
- A 95% CI for the mean birthweight

$$2994 \pm 1.96\frac{800}{\sqrt{78}} = (2816, 3172)$$

- A 99% CI for the mean birthweight

$$2994 \pm 2.58\frac{800}{\sqrt{78}} = (2760, 3228)$$

Assumptions

- Ys are sampled from a normal distribution
- Variance is known

But ...

- What do we do if the variance is unknown?
- If σ^2 is not known, we can estimate it with s^2
- However, the distribution of

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

is not normal

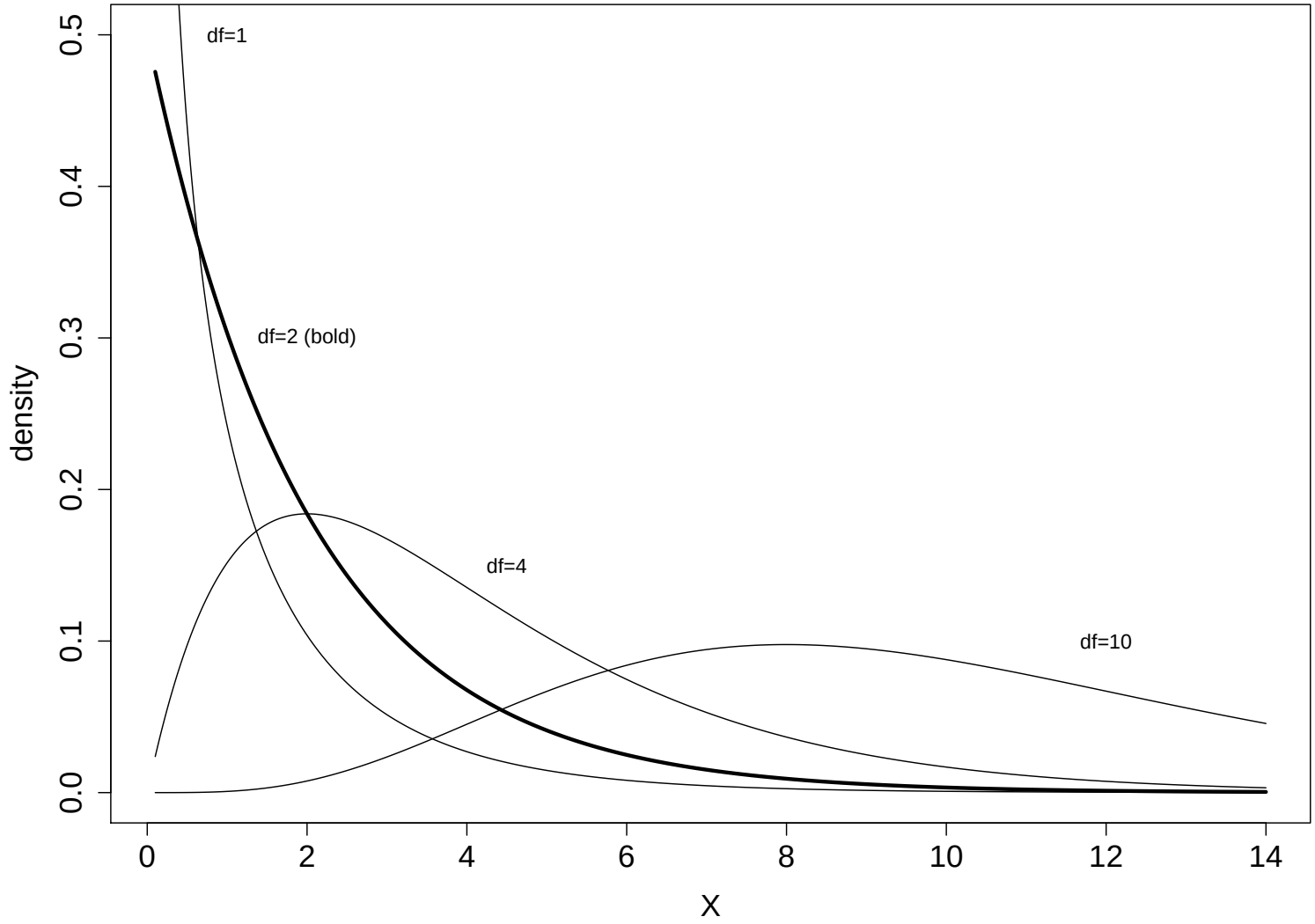
Distribution of s^2

- Result 4.4 (text, p 95): If a random variable Y is normally distributed with mean μ and variance σ^2, then for a random sample of size n, the quantity

$$\frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with $n-1$ degrees of freedom, which we denote by χ^2_{n-1}

χ^2 Distribution



t Distribution

- Let $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$
- If Z and W are independent, then

$$T = \frac{Z}{\sqrt{W/\nu}}$$

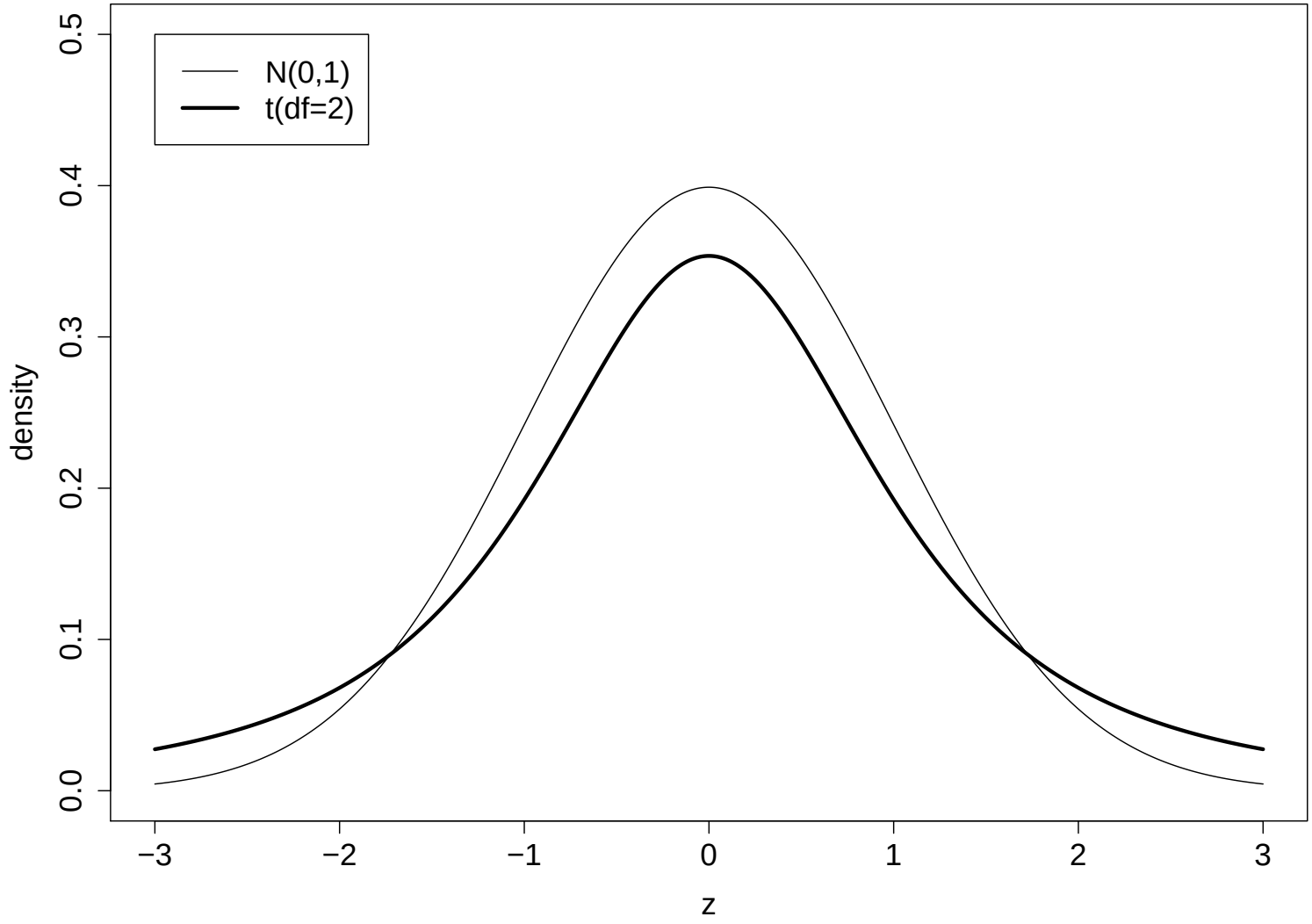
  will follow the t-distribution with ν degrees of freedom.

- We know

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ and } W = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Can show that $\bar{Y}$ and s^2 are independent

t Distribution



CI for μ when σ^2 unknown

- Substituting, we get

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}(\bar{Y} - \mu)/\sigma}{\sqrt{\{(n-1)s^2/\sigma^2\}/(n-1)}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

- Thus a $100(1-\alpha)\%$ CI for μ is given by

$$\bar{Y} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

- Note 1: We are still assuming the Ys are normal
- Note 2: If $n \geq 30$, we can use z as a reasonable approximation for t

Example

- Now suppose we have the birthweights of a sample of SIDS victims but that the variance is unknown
- $\bar{Y} = 2920.0$, $s = 792.86$ and $n = 23$
- $t_{22,0.975} = 2.07$
- 95% CI for μ:

$$2920.0 \pm 2.07\left(\frac{792.86}{\sqrt{23}}\right) = 2920.0 \pm 342.9$$
$$= (2577.1, 3262.8)$$

- Note that here we multiply the (estimated) s.e. by 2.07 rather than the normal distribution's 1.96 as a penalty for not knowing σ

Quantiles of t

- How to find $t_{22,0.975} = 2.07$?
- Text, Table A.4 page 822: column 4, row 22
- R:

```
> qt(0.975,22)
[1] 2.073873
```

- SAS:

```
data;
  x=quantile('T',0.975,22);
proc print;

Obs       x

 1     2.07387
```

CIs Using Software

- R:

```
> t.test(x)$conf.int

[1] 2577.113 3262.830
attr(,"conf.level")
[1] 0.95
```

- SAS PROC TTEST (edited output):

```
proc ttest;
   var x;

The TTEST Procedure

  N        Mean     Std Dev     Std Err     Minimum     Maximum

 23      2920.0       792.9       165.3      1252.8      4369.2

    Mean        95% CL Mean        Std Dev      95% CL Std Dev

  2920.0      2577.1   3262.8        792.9       613.2   1122.2
```

Non-normal Data

- If the Ys are not normally distributed, we use the CLT:
- If $Y_1, \ldots, Y_n$ is a random sample from a distribution with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$ for $i = 1, \ldots, n$, then $\bar{Y}$ is approximately distributed as $N(\mu, \sigma^2/n)$ for large n

- The use of the CLT to construct a CI for μ requires knowledge of σ^2
- To use the CLT when σ^2 unknown requires *Slutsky's Theorem*

Slutsky's Theorem

- If X_n is a sequence of random variables that converges in distribution to X, and
- Y_n is a sequence of random variables that converges in probability to a constant c,
- Then $W_n = X_n Y_n$ converges in distribution to cX
- That is

$$\lim_{n\to\infty} \Pr[W_n \leq w] = \Pr[cX \leq w]$$

Non-normal Data

- Let

$$X_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \text{ and } Y_n = \sqrt{\frac{\sigma^2}{s^2}}$$

- We know $X_n \xrightarrow{d} Z \sim N(0,1)$ and $\sigma^2/s^2 \xrightarrow{p} 1$
- Then Slutsky's Theorem implies

$$W_n = X_n Y_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{s^2}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

  will be approximately $\sim N(0,1)$

- The approximation gets better as $n \to \infty$

Large Sample CI for μ

- If n is sufficiently large, an approximate $100(1-\alpha)\%$ CI for μ is

$$\bar{Y} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

- This is true regardless of the original distribution of the Ys

Example

- A survey was conducted to estimate the mean age that smoking was started among women who smoke.
  A random sample of 243 smoking women in NC found $\bar{y} = 16.8$ and $s = 2.36$.
- A 95% CI for the mean age of smoking onset is:

$$16.8 \pm 1.96 \left(\frac{2.36}{\sqrt{243}} \right) = (16.5, 17.1)$$

Summary of CIs for μ

Normal	σ^2 known	n large	Confidence Interval
$\surd$	$\surd$	–	$\bar{Y} \pm z_{1-\alpha/2}(\sigma/\sqrt{n})$
–	$\surd$	$\surd$	$\bar{Y} \pm z_{1-\alpha/2}(\sigma/\sqrt{n})$
$\surd$	–	–	$\bar{Y} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n})$
–	–	$\surd$	$\bar{Y} \pm z_{1-\alpha/2}(s/\sqrt{n})$
–	–	–	Transform; nonparametrics

Non-normal Data with Small Sample Size

- With a small sample size it is difficult to test for normality
- Transformation of the data
- Nonparametric methods
  - Bootstrap
  - CI for median

Bootstrap t-intervals

- Empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x]$$

- Statistical theory indicates $F_n(x) \xrightarrow{p} F(x)$ where F is the population distribution function
- Bootstrap: approximate the sampling distribution of a statistic (in this case the sample mean) by repeatedly sampling (with replacement) from the empirical distribution function F_n
- See text, section 8.10.2

Bootstrap t-intervals

- Bootstrap t-interval: an approximate $100(1-\alpha)\%$ CI for μ is

$$(\bar{Y} - \hat{t}_{(1-\alpha/2)}\frac{s}{\sqrt{n}},\ \ \bar{Y} - \hat{t}_{(\alpha/2)}\frac{s}{\sqrt{n}})$$

  where $\hat{t}_{(1-\alpha/2)}$ and $\hat{t}_{(\alpha/2)}$ are determined from bootstrap samples as described on the next slide

Bootstrap t-intervals

1. Draw a random sample of size n with replacement from $\{x_1, \ldots, x_n\}$; call this $\boldsymbol{x}^*(1)$
2. Compute $Z^*(1)$ as described on the next slide
3. Do steps 1 and 2 a total of B times, to obtain $Z^*(1), Z^*(2), \ldots, Z^*(B)$
4. Let $\hat{t}_{(\alpha/2)}$ be the $\alpha/2$ sample quantile of $\{Z^*(1), \ldots, Z^*(B)\}$; similarly for $\hat{t}_{(1-\alpha/2)}$

- The order of steps 2 and 3 can be interchanged, first obtaining B bootstrap samples $\boldsymbol{x}^*(1), \ldots, \boldsymbol{x}^*(B)$ and then computing $Z^*(b)$ for each bootstrap sample

Bootstrap t-intervals

- Step 2. For each bootstrap sample compute

$$Z^*(b) = \frac{\bar{x}^*(b) - \bar{x}}{\widehat{se}^*(b)}$$

where $\bar{x}^*(b)$ is the mean of $\boldsymbol{x}^*(b)$, $\bar{x}$ is the mean of the original sample, and $\widehat{se}^*(b)$ is the estimated standard error of $\bar{x}^*(b)$, i.e.,

$$\widehat{se}^*(b) = \sqrt{\text{Var}\{\boldsymbol{x}^*(b)\}/n}$$

where $\text{Var}\{\boldsymbol{x}^*(b)\}$ is the sample variance of the b^{th} bootstrap sample $\boldsymbol{x}^*(b)$

Bootstrap t-intervals

- Simulation study; 10,000 simulated datasets of size $n = 20$; $B = 500$ bootstrap samples per dataset
- Calculate empirical coverage probabilities for CIs using t and bootstrap t. That is, for what proportion of the 10,000 simulated datasets do the CIs contain the true mean μ
- A "good" method of generating confidence intervals should have empirical coverage close to the claimed confidence (95% in this example)
- Generate the datasets from three different population distributions, each with mean $\mu = 1$

Simulating Using R

```
sim <- function(void){

   cover_tmp=c(0,0)
   n <- 20
   # y <- rnorm(n,1,1)
   # y <- rexp(n,1)  # mean of exp. 1 is 1
   y <- rchisq(n,1)  # mean of chisq k is k

   mean.y <- mean(y)
   var.y <- var(y)

   # CI using t
   lower.y <- mean.y - qt(0.975,n-1)*sqrt(var.y/n)
   upper.y <- mean.y + qt(0.975,n-1)*sqrt(var.y/n)

   if (lower.y <1 & upper.y >1) cover_tmp[1] <- 1
```

```
# bootstrap-t interval
boots <- 500
zs <- matrix(0,1,boots)
for (jj in 1:boots){
   ysamp <- sample(y,size=n,replace=T)
   zs[jj] <- (mean(ysamp)-mean.y)/sqrt(var(ysamp)/n)
   }
lower.t <- quantile(zs,0.975)
upper.t <- quantile(zs,0.025)

lower.y <- mean.y - lower.t*sqrt(var.y/n)
upper.y <- mean.y - upper.t*sqrt(var.y/n)

if (lower.y <1 & upper.y >1) cover_tmp[2] <- 1

cover_tmp
}
```

```
# run the simulation
nsims <- 10000
set.seed(43567)
cover <- matrix(0,nsims,2)
for (ii in 1:nsims){
        cover[ii,] <- sim(ii)
        }
output <-c(mean(cover[,1]),mean(cover[,2]))
print(output)
```

Bootstrap t-intervals

Empirical Coverage Probabilities

n	Population distribution	t	Bootstrap t	"Simple" bootstrap
20	N(1,1)	0.949	0.946	0.924
	Chi-squared (1 df)	0.891	0.936	0.876
	Chi-squared (1 df)*	0.905	0.942	0.875
	Exponential(1)	0.922	0.945	0.902
25	N(1,1)	0.949	0.947	0.931
	Chi-squared (1 df)	0.922	0.944	0.889
	Exponential(1)	0.902	0.939	0.889

* Using a different random number seed

Bootstrap Notes

- Many types of bootstrap CIs available
- Bootstrap CIs need not be symmetric
- Large sample theoretical justification; empirically small sample performance good
- R: library("boot")

Outline

- Introduction
- CIs for the mean
  - Parametric, large sample
  - Bootstrap
- CI for quantiles
  - Exact
  - Large sample
- CI for variance

Nonparametric CI for the Median

- Suppose $X_1, \ldots, X_n$ are iid according to CDF F
- Let $\zeta_{0.5}$ be the population median
- Construct a symmetric $100(1-\alpha)\%$ CI by finding largest r such that

$$\Pr[X_{(r)} \leq \zeta_{0.5} \leq X_{(n-r+1)}] \geq 1 - \alpha$$

- Sufficient to find largest r such that

$$\Pr[\zeta_{0.5} < X_{(r)}] \leq \alpha/2$$

Bernoulli Random Variable

- Let Y be a Bernoulli random variable
- Y can take on two values, 0 or 1

$$\Pr[Y = 1] = \pi; \ \Pr[Y = 0] = 1 - \pi$$

$$E(Y) = \pi; \ \text{Var}(Y) = \pi(1 - \pi)$$

Binomial Random Variable

- Consider a process that produces independent Bernoulli random variables with the same probability of success π
- Let Y count the number of successes in n trials
- $Y \sim \text{Binomial}(n, \pi)$

$$\Pr[Y = y] = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

$$E(Y) = n\pi$$

$$\text{Var}(y) = n\pi(1-\pi)$$

Derivation of CI for Median

- CDF

$$\Pr[X_i \leq x] = F(x)$$

- Therefore

$$\begin{aligned}
\Pr[x < X_{(r)}] &= 1 - \Pr[X_{(r)} \leq x] \\
&= 1 - \Pr[\text{at least } \ r \ \text{ of the } \ X_i \leq x] \\
&= 1 - \sum_{i=r}^{n} \binom{n}{i} F(x)^i \{1 - F(x)\}^{n-i} \\
&= \sum_{i=0}^{r-1} \binom{n}{i} F(x)^i \{1 - F(x)\}^{n-i}
\end{aligned}$$

Derivation of CI for Median

- CDF of Binomial(n, $\pi = F(x)$)
- If $p = 0.5$, then $F(\zeta_p) = 0.5$
- So

$$\Pr[\zeta_{0.5} < X_{(r)}] = \frac{1}{2^n} \sum_{i=0}^{r-1} \binom{n}{i}$$

- Choose largest r such that

$$\frac{1}{2^n} \sum_{i=0}^{r-1} \binom{n}{i} \leq \alpha/2$$

Derivation of CI for Median: Example

- Using $n = 23$
- CDF of $X \sim \text{Binomial}(23, 0.5)$

x	$\Pr[X \leq x]$
0	1.192093e-07
1	2.861023e-06
2	3.302097e-05
3	2.441406e-04
4	1.299739e-03
5	5.311012e-03
6	1.734483e-02
7	4.656982e-02
⋮	

- Pick $r = 7$

Derivation of CI for Median

- Values of r for 95% CI for Median

n	r
1-5	0
6-8	1
9-11	2
12-14	3
15-16	4
17-19	5
20-22	6
23-24	7
25-27	8
28-29	9
30-32	10
33-34	11

- Cf. page 269-270 of the text

95% CI Example

- For $n = 23$, choose $r = 7$ and then $n - r + 1 = 17$
- Therefore

$$(x_{(7)}, x_{(17)})$$

  gives a 95% CI for the median

- This CI makes no assumptions about the distribution of the Ys
- Note:

$$\frac{1}{2^{23}} \sum_{i=7}^{23-7} \binom{23}{i} = 0.9653 \geq 1 - \alpha$$

```
> sum(dbinom(7:16,23,0.5))
[1] 0.9653103
```

SAS Code and Output

```
proc univariate data=beta cipctldf;
  var base1;
run;
```

Quantiles (Definition 5)

		95% Confidence Limits		-------Order Statistics-------		
Quantile	Estimate	Distribution Free		LCL Rank	UCL Rank	Coverage
100% Max	298					
99%	298	.	.	.	.	.
95%	252	212	298	21	23	58.75
90%	212	202	298	19	23	83.83
75% Q3	192	162	252	13	22	97.35
50% Median	152	106	186	7	17	96.53
25% Q1	100	74	124	2	11	97.35
10%	80	68	92	1	5	83.83
5%	74	68	80	1	3	58.75
1%	68	.	.	.	.	.
0% Min	68					

Large Sample CI for the Median

- The above method of finding a $(1-\alpha)100\%$ CI for the median is *exact*, i.e., the probability the CI contains $\zeta_{0.5}$ is guaranteed to be at least $(1-\alpha)$
- Now we derive a large sample CI for the median using the CLT
- This will be approximate in that the probability the CI contains $\zeta_{0.5}$ is approximately $(1-\alpha)$, with the approximation improving as $n \to \infty$

Large Sample CI for Any Quantile

- In general,

$$\Pr[\zeta_p < X_{(r)}] = \sum_{i=0}^{r-1} \binom{n}{i} F(\zeta_p)^i \{1 - F(\zeta_p)\}^{n-i}$$

$$= \sum_{i=0}^{r-1} \binom{n}{i} p^i q^{n-i}$$

  where $q = 1 - p$

- From the CLT, if $Y \sim \text{Binomial}(n, p)$, then

$$\frac{Y - np + 1/2}{\sqrt{npq}} \sim N(0, 1)$$

- The $1/2$ is a *continuity correction* (see text p. 156)

Large Sample CI for Any Quantile

- Thus

$$\begin{aligned}\Pr[\zeta_p < X_{(r)}] &= \Pr[Y \leq r-1] \\ &\approx \Pr[Z \leq \frac{(r-1) - np + 1/2}{\sqrt{npq}}] \\ &= \Phi(\frac{r - np - 1/2}{\sqrt{npq}})\end{aligned}$$

- The goal is a symmetric $(1-\alpha)\%$ CI, so we want

$$\alpha/2 = \Pr[\zeta_p < X_{(r)}] = \Phi(\frac{r - np - 1/2}{\sqrt{npq}})$$

- That is

$$-z_{1-\alpha/2} = \frac{r - np - 1/2}{\sqrt{npq}}$$

Large Sample CI for Any Quantile

- This implies

$$r = np + \frac{1}{2} - z_{1-\alpha/2}\sqrt{npq}$$

- Similar reasoning yields

$$s = np + \frac{1}{2} + z_{1-\alpha/2}\sqrt{npq}$$

- For $p = 1/2$:

$$r = \frac{n+1}{2} - z_{1-\alpha/2}\frac{\sqrt{n}}{2}$$

$$s = \frac{n+1}{2} + z_{1-\alpha/2}\frac{\sqrt{n}}{2}$$

Large Sample CI for Any Quantile

- Thus a $100(1-\alpha)\%$ CI for ζ_p is given by

$$(X_{(\lfloor r \rfloor)}, X_{(\lceil s \rceil)})$$

- Note: n large enough ensures $\lfloor r \rfloor, \lceil s \rceil \in \{1, \ldots, n\}$

Large Sample CI for Median: Example

- Suppose $n = 100$ and $\alpha = 0.05$
- Then

$$z_{1-\alpha/2}\frac{\sqrt{n}}{2} = 5(1.96) = 9.8$$

- Rounding (using the floor and ceiling functions) yields:

$$50.5 \pm 9.8 \Rightarrow (x_{(40)}, x_{(61)})$$

- Can show $r = 40$ using the exact method

```
> sum(dbinom(40:60,100,1/2))
[1] 0.9647998
> sum(dbinom(41:59,100,1/2))
[1] 0.943112
>
> 2*sum(dbinom(0:39,100,1/2))
[1] 0.0352002
```

CI for Variance

- Suppose $Y_1, \ldots, Y_n$ is a random sample from a normal distribution with mean μ and variance σ^2
- Recall (result 4.4 on p. 95 of the text)

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Therefore

$$1 - \alpha = \Pr[\chi^2_{\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}]$$

- Implying

$$1 - \alpha = \Pr\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$$

CI for Variance

- Because the χ^2 distribution is not symmetric, we need to look up both $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$
- This CI is dependent on the Ys being from a normal distribution

CI for Variance Example

- Using the same data as for the example of the CI for the median (slides on pp. 44, 46 & 47)
- $n = 23$; $s^2 = 3701.36$
- R: qchisq(0.025,22)

  SAS: data; x=quantile('Chisq',0.025,22);

  Table A.3, page 821
- $\chi^2_{0.025,22} = 10.98$; $\chi^2_{0.975,22} = 36.78$
- Therefore, 95% CI for σ^2:

$$(22(3701.36)/36.78, 22(3701.36)/10.98)$$
$$= (2213.973, 7416.203)$$

- 95% CI for $\sigma = (47.05, 86.12)$

SAS Code and Output

```
proc univariate data=beta cibasic;
  var base1;
run;
```

Basic Confidence Limits Assuming Normality

Parameter	Estimate	95% Confidence Limits	
Mean	150.78261	124.47394	177.09128
Std Deviation	60.83880	47.05242	86.10828
Variance	3701	2214	7415

CI for Variance – Non-normal Data

- Large sample theory

$$\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} N(0, (\alpha_4 - 1)\sigma^4)$$

where $\alpha_4 = E(X - \mu)^4/\sigma^4$ is the *kurtosis*
(cf. Dudewicz and Mishra, *Modern Mathematical Statistics*, p. 325)

- "Crude approximation": replace usual CI with

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}(1 + g_2/n)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}(1 + g_2/n)} \right)$$

where $g_2 = a_4 - 3$ and a_4 is an estimate of α_4
(cf. Solomon and Stephens, *Encyclopedia of Stat Sci*)

CI for Variance – Non-normal Data

- Nonparametric approach such as bootstrap (cf. Efron and Tibshirani, *An Introduction to the Bootstrap*, Ch. 14)
- Software?