Point and Interval Estimation
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-08-25 17:41

Outline

- Introduction
- 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; Eg μ
- *Estimator*: the statistic used to estimate the estimand; a random variable; Eg $\bar{Y}$
- *Estimate*: a realization of an estimator from an observed data set; Eg $\bar{y} = 36.3$

Estimating μ

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

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

  i.e., the mean of the sampling distribution of $\bar{Y}$ equals μ, the population parameter of interest

Confidence Interval for μ

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

- Then

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

  cf Result 4.2 for last 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 $z_{1-\alpha/2}$

α	$z_{1-\alpha/2}$
0.10	1.645
0.05	1.960
0.01	2.576

CI Interpretation, Comment

- van Belle et al (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 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

- Y's are sampled from a normal distribution
- Variance is known

- What do we do if 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

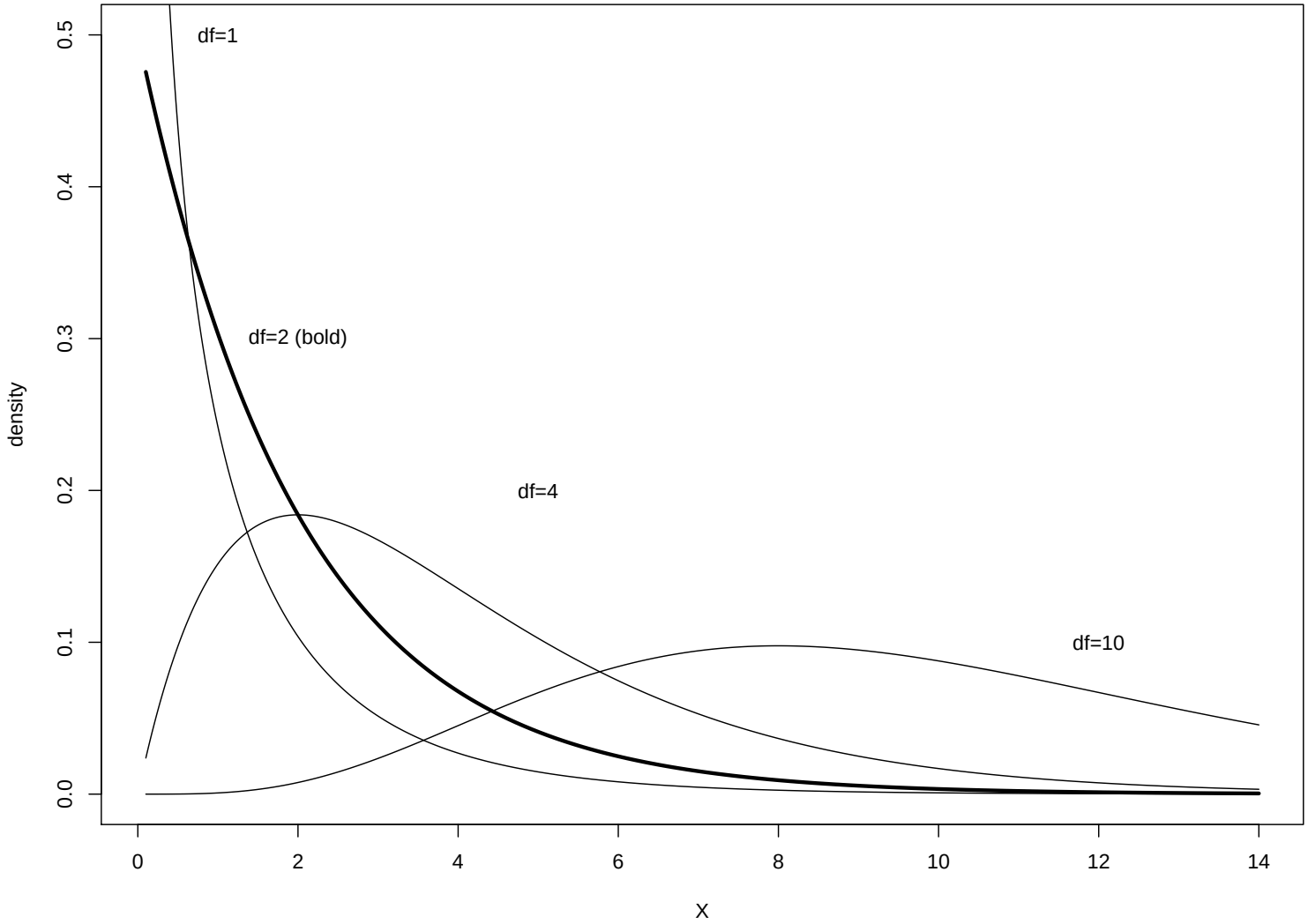
Distribution of s^2

- Result 4.4 (p 95 text): 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

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

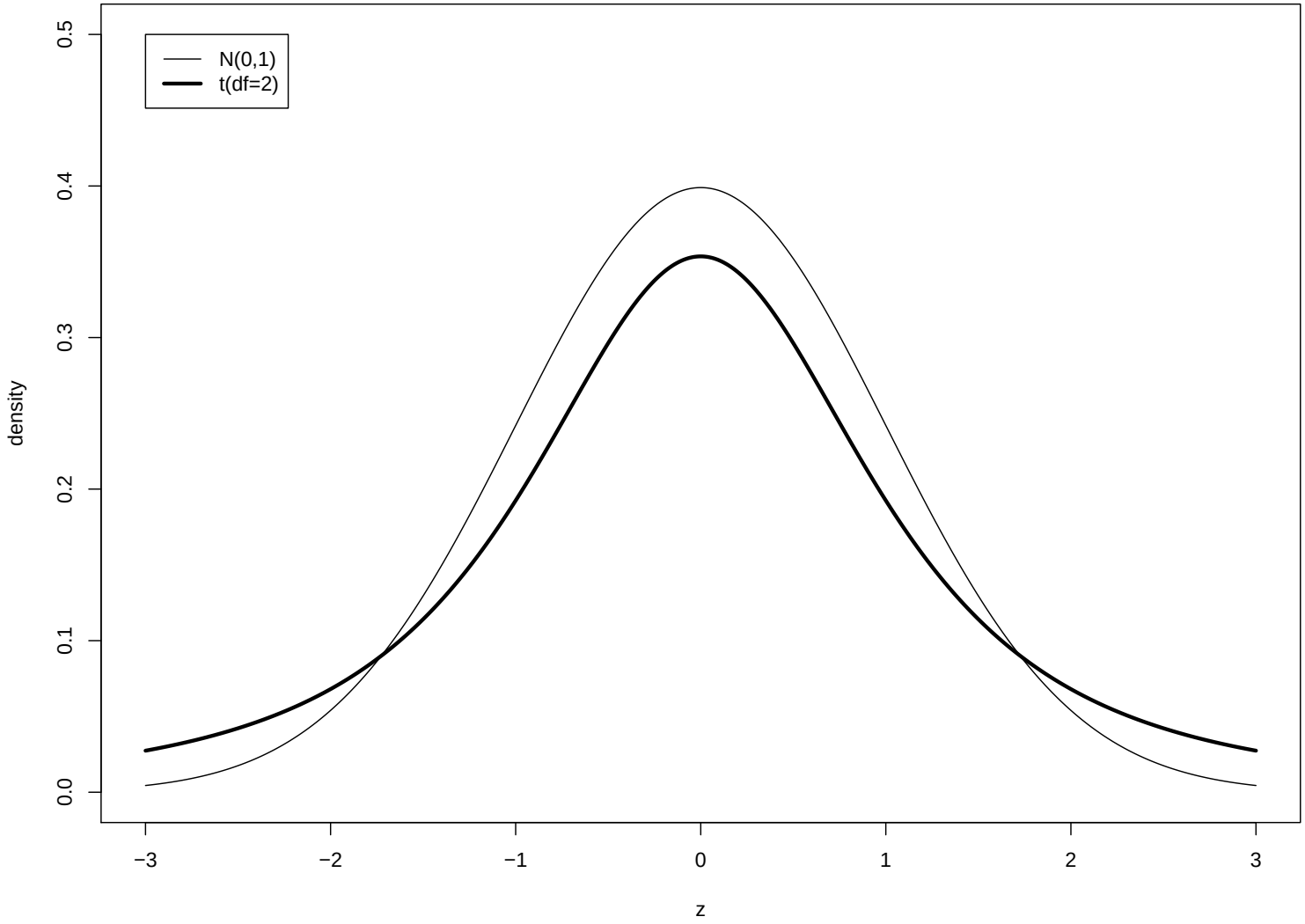
  will follow the t-distribution with ν degrees of freedom.

- 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 $\bar{Y}$ 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 $(1 - \alpha) \times 100\%$ 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 Y's are normal
- Note 2: If $n \geq 30$, can use z instead of t

Example

- $\bar{Y} = 164.6$, $s = 52.24$, $n = 23$
- $t_{22,.975} = 2.07$
- 95% CI for μ:

$$164.6 \pm 2.07 \left(\frac{52.24}{\sqrt{23}} \right) = 164.6 \pm 22.6 = (142.0, 187.2)$$

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

Quantiles of t

- How to get $t_{22,.975} = 2.07$?
- Text Table A.4 page 822: column 4, row 22
- R:

```
> qt(.975,22)
```

- SAS:

```
data;
  x=quantile('T',.975,22);
```

CIs using Software

- R:

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

[1] 141.9821 187.1631
attr(,"conf.level")
[1] 0.95
```

- SAS Proc Ttest:

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev
x	23	141.98	164.57	187.16	40.403	52.241

Non-normal data

- If the Y's 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 $Var(Y_i) = \sigma^2$ 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 r.v. that converges in distribution to X, and
- Y_n is a sequence of r.v. that converges in probability to a constant c,
- then $W_n = X_n Y_n$ converges in distribution to cX
- I.e.

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

Non-normal Data

- Let

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

- Know $X_n \rightarrow^d Z \sim N(0,1)$ and $\sigma^2/s^2 \rightarrow^p 1$
- 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 $n \rightarrow \infty$

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 Y's

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

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 w/ Small Sample size

- With 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 $F_n(x) \rightarrow_p F(x)$ where F is the population distribution function
- Bootstrap: approximate sampling distribution of statistic (in this case the sample mean) by repeatedly sampling (with replacement) from the empirical distribution function F_n

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 $\hat{t}_{(\alpha/2)}$ determined from bootstrap samples as described on next slide

Bootstrap t-intervals

1. Draw random sample of size n with replacement from $\{x_1, \ldots, x_n\}$; call this $\boldsymbol{x}^*(1)$
2. Repeat step 1 to get B bootstrap samples $\boldsymbol{x}^*(1)$,. . . ,$\boldsymbol{x}^*(B)$
3. Compute $Z^*(b)$ for each bootstrap sample as described on the next slide
4. Let $\hat{t}_{(\alpha/2)}$ be the $\alpha/2$ sample quantile of $\{Z^*(1), \ldots, Z^*(B)\}$; similarly for $\hat{t}_{(1-\alpha/2)}$

Bootstrap t-intervals

- Step 3. For each bootstrap sample compute

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

where $\bar{x}^*(b)$ is 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.,

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

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

Bootstrap t-intervals

- Simulation study; 10000 simulated data sets of size n; $B = 500$ bootstrap samples per data set
- Empirical coverage probabilities

n	Population distribution	t	Bootstrap t
20	N(1,1)	0.95	0.95
	Chi-squared (1 df)	0.89	0.94
	Exponential(1)	0.92	0.95
25	N(1,1)	0.95	0.95
	Chi-squared (1 df)	0.90	0.94
	Exponential(1)	0.92	0.94

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 $X_1, \ldots, X_n$ iid according to CDF F
- Let $\zeta_{1/2}$ be the population median
- Construct symmetric $(1 - \alpha)$ CI by finding largerst r such that

$$\Pr[X_{(r)} \leq \zeta_{1/2} \leq X_{(n-r+1)}] \geq 1 - \alpha$$

- Sufficient to find largest r such that

$$\Pr[X_{(r)} > \zeta_{1/2}] \leq \alpha/2$$

Bernoulli RV

- Let Y be a Bernoulli r.v.
- Y can take on two values, 0 or 1

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

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

Binomial RV

- Process that produces independent Bernoulli RVs with the same probability of success π
- Let Y count the number of successes in n trials
- $Y \sim$ Binomial(n,π)

$$\Pr[Y = y] = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

$$E(Y) = n\pi; Var(y) = n\pi(1-\pi)$$

Derivation of CI for Median

- CDF

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

- Therefore

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

Derivation of CI for Median

- CDF of Binomial$(n, \pi = F(x))$
- If $p = 1/2$, then $F(\zeta_p) = 1/2$
- So

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

- 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 $n = 23$
- CDF of $X \sim$Binomial(23,.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 van Belle et al.

95% CI Example

- For $n = 23$, choose $r = 7$ such that $n - r + 1 = 17$
- Therefore

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

  gives a 95% CI for the median

- This CI makes no assumptions about the distribution of the Y's
- Note:

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

```
> sum(dbinom(7:16,23,1/2))
[1] 0.9653103
```

SAS Code and Output

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

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

Large sample CI for median

- The above method of finding a $(1-\alpha)100\%$ CI for the median is *exact*, i.e., the probability the CI contains $\zeta_{.5}$ is guaranteed to be at least $(1-\alpha)$
- Now we derive a large sample CI for the median using the CLT
- This will be approx in that the probability the CI contains $\zeta_{.5}$ is $\approx (1-\alpha)$, with the approx improving as $n \to \infty$

Large sample CI for any quantile

- If general,

$$\begin{aligned}\Pr[\zeta_p \leq Z_{(r)}] &= \textstyle\sum_{i=0}^{r-1} \binom{n}{i} F(\zeta_p)^i \{1 - F(\zeta_p)\}^{n-i} \\ &= \textstyle\sum_{i=0}^{r-1} \binom{n}{i} p^i q^{n-i}\end{aligned}$$

  where $q = 1 - p$

- From CLT, if $Y \sim Bin(n, p)$, then

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

- Thus

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

Large sample CI for any quantile

- Goal is symmetric $(1-\alpha)\%$ CI, so want

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

- That is

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

- Implying

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

- For $p = 1/2$, yields

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

Large sample CI for any quantile

- Similar reasoning yields

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

- Thus $(1-\alpha)\%$ 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 yields:

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

- Can show $r = 40$ using exact method

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

CI for Variance

- Recall (result 4.4 p.95 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

- Since the χ^2 distribution is not symmetric, 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 Y's being from a normal distribution

CI for Variance Example

- $n = 23$; $s^2 = 3701.36$
- $\chi^2_{.025,22} = 10.98$; $\chi^2_{.975,22} = 36.78$
- Aside:

  R: qchisq(.025,22)

  SAS: data; x=quantile('Chisq',.025,22);

  Table A.3, page 821
- 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 - Nonnormal data

- Large sample theory

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

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, *Encyc of Stat Sci*)

CI for Variance - Nonnormal data

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