Chapter 5
Confidence Regions and the Bootstrap

This chapter follows Olive (2014, ch. 9; 2017b, ∮ 5.3) closely. Also see Olive (2022abcd). Sections 5.1–5.3 consider confidence intervals from asymptotic pivots while Section 5.4 covers bootstrap confidence regions. Closed regions are better than open regions. Again, $0 < \delta < 1$.

Notation: As in Chapter 4, $P(A_n)$ is "eventually bounded below" by $1-\delta$ if $P(A_n)$ gets arbitrarily close to or higher than $1-\delta$ as $n \to \infty$. Hence $P(A_n) > 1 - \delta - \epsilon$ for any $\epsilon > 0$ if n is large enough. If $P(A_n) \to 1-\delta$ as $n \to \infty$, then $P(A_n)$ is eventually bounded below by $1-\delta$. The actual coverage is $1 - \gamma_n = P(\theta \in [L_n, U_n])$, the nominal coverage is $1-\delta$ where $0 < \delta < 1$. The 90% and 95% large sample confidence intervals and confidence regions are common.

5.1 Confidence Intervals

Definition 5.1. Let the data $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ have joint pdf or pmf $f(\boldsymbol{y}|\theta)$ with parameter space Θ and support $\mathcal{Y}$. Let $L_n(\boldsymbol{Y})$ and $U_n(\boldsymbol{Y})$ be statistics such that $L_n(\boldsymbol{y}) \leq U_n(\boldsymbol{y})$, $\forall \boldsymbol{y} \in \mathcal{Y}$. Then $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a 100 $(1-\delta)$ % **confidence interval** (CI) for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y})) = 1 - \delta$$

for all $\theta \in \Theta$. The interval $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a large sample $100(1-\delta)$ % CI for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y}))$$

is eventually bounded below by $1-\delta$ for all $\theta \in \Theta$ as the sample size $n \to \infty$.

Pivots and asymptotic pivots are used to make CIs. An asymptotic pivot is a random quantity that is not a statistic since the asymptotic pivot depends on the unknown parameters $\boldsymbol{\theta}$.

Definition 5.2. Let the data $Y_1, ..., Y_n$ have joint pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ with parameter space Θ and support $\mathcal{Y}$. The quantity $R(\boldsymbol{Y}|\boldsymbol{\theta})$ is a **pivot** or pivotal quantity if the distribution of $R(\boldsymbol{Y}|\boldsymbol{\theta})$ is independent $\boldsymbol{\theta}$. The quantity $R(\boldsymbol{Y}, \boldsymbol{\theta})$ is an **asymptotic pivot** or asymptotic pivotal quantity if the limiting distribution of $R(\boldsymbol{Y}, \boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$.

The first CI in Definition 5.1 is sometimes called an exact CI. The words "exact" and "large sample" are often omitted. In the following definition, the scaled asymptotic length is closely related to asymptotic relative efficiency of an estimator and high power of a test of hypotheses.

Definition 5.3. Let $[L_n, U_n]$ be a 100 $(1-\delta)$% CI or large sample CI for θ. If

$$n^{\tau}(U_n - L_n) \xrightarrow{P} A_{\delta}$$

where $0 < \tau \leq 1$, then A_{δ} is the *scaled asymptotic length* of the CI. Typically $\tau = 0.5$ but superefficient CIs have $\tau = 1$. For fixed τ and fixed coverage $1-\delta$, a CI with smaller A_{δ} is "better" than a CI with larger A_{δ}. If $A_{1,\delta}$ and $A_{2,\delta}$ are for two competing CIs with the same τ, then $(A_{2,\delta}/A_{1,\delta})^{1/\tau}$ is a measure of "asymptotic relative efficiency."

Definition 5.4. Suppose a nominal $100(1-\delta)$% CI for θ has actual coverage $1-\gamma$, so that $P_{\theta}(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y})) = 1-\gamma$ for all $\theta \in \Theta$. If $1-\gamma > 1-\delta$, then the CI is *conservative*. If $1-\gamma < 1-\delta$, then the CI is *liberal*. Conservative CIs are generally considered better than liberal CIs. Suppose a nominal $100(1-\delta)$% large sample CI for θ has actual coverage $1-\gamma_n$ where $\gamma_n \to \gamma$ as $n \to \infty$ for all $\theta \in \Theta$. If $1-\gamma > 1-\delta$, then the CI is *asymptotically conservative*. If $1-\gamma < 1-\delta$, then the CI is *asymptotically liberal*. It is possible that $\gamma \equiv \gamma(\theta)$ depends on θ, and that the CI is (asymptotically) conservative or liberal for different values of θ, in that the (asymptotic) coverage is higher or lower than the nominal coverage, depending on θ.

Example 5.1. a) Let $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$ where $\sigma^2 > 0$. Then

$$R(\boldsymbol{Y}|\mu, \sigma^2) = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

is a pivot. A statistic does not depend on any unknown parameters. Hence the above pivot is not a statistic if μ is unknown.

To use this pivot to find a CI for μ, let $t_{p,\delta}$ be the δ percentile of the t_p distribution. Hence $P(T \leq t_{p,\delta}) = \delta$ if $T \sim t_p$. Using $t_{p,\delta} = -t_{p,1-\delta}$ for $0 < \delta < 0.5$, note that

$$1-\delta = P(-t_{n-1,1-\delta/2} \le \frac{\overline{Y}-\mu}{S/\sqrt{n}} \le t_{n-1,1-\delta/2}) =$$

$$P(-t_{n-1,1-\delta/2} \ \ S/\sqrt{n} \le \overline{Y}-\mu \le t_{n-1,1-\delta/2} \ \ S/\sqrt{n}) =$$
$$P(-\overline{Y} - t_{n-1,1-\delta/2} \ \ S/\sqrt{n} \le -\mu \le -\overline{Y} + t_{n-1,1-\delta/2} \ \ S/\sqrt{n}) =$$
$$P(\overline{Y} - t_{n-1,1-\delta/2} \ \ S/\sqrt{n} \le \mu \le \overline{Y} + t_{n-1,1-\delta/2} \ \ S/\sqrt{n}).$$

Thus

$$\overline{Y} \pm t_{n-1,1-\delta/2} \ \ S/\sqrt{n}$$

is a $100(1-\delta)\%$ CI for μ.

b) If $Y_1, ..., Y_n$ are iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2 > 0$, then, by the CLT and Slutsky's Theorem,

$$R(\boldsymbol{Y}, \mu, \sigma^2) = \frac{\overline{Y}-\mu}{S/\sqrt{n}} = \frac{\sigma}{S} \ \ \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$

is an asymptotic pivot.

To use this asymptotic pivot to find a large sample CI for μ, let z_δ be the δ percentile of the $N(0,1)$ distribution. Hence $P(Z \le z_\delta) = \delta$ if $Z \sim N(0,1)$. Using $z_\delta = -z_{1-\delta}$ for $0 < \delta < 0.5$, note that for large n,

$$1-\delta \approx P(-z_{1-\delta/2} \le \frac{\overline{Y}-\mu}{S/\sqrt{n}} \le z_{1-\delta/2}) =$$

$$P(-z_{1-\delta/2} \ \ S/\sqrt{n} \le \overline{Y}-\mu \le z_{1-\delta/2} \ \ S/\sqrt{n}) =$$
$$P(-\overline{Y} - z_{1-\delta/2} \ \ S/\sqrt{n} \le -\mu \le -\overline{Y} + z_{1-\delta/2} \ \ S/\sqrt{n}) =$$
$$P(\overline{Y} - z_{1-\delta/2} \ \ S/\sqrt{n} \le \mu \le \overline{Y} + z_{1-\delta/2} \ \ S/\sqrt{n}).$$

Thus

$$\overline{Y} \pm z_{1-\delta/2} \ \ S/\sqrt{n} \tag{5.1}$$

is a large sample $100(1-\delta)\%$ CI for μ.

Since $t_{n-1,1-\delta/2} > z_{1-\delta/2}$ but $t_{n-1,1-\delta/2} \to z_{1-\delta/2}$ as $n \to \infty$,

$$\overline{Y} \pm t_{n-1,1-\delta/2} \ \ S/\sqrt{n} \tag{5.2}$$

is also a large sample $100(1-\delta)\%$ CI for μ. This t interval is the same as that in a), and is likely the most widely used confidence interval in statistics. Replacing $z_{1-\delta/2}$ by $t_{n-1,1-\delta/2}$ makes the CI longer and hence less likely to be liberal.

Remark 5.1.

$$\overline{Y} \pm t_{n-1,1-\delta/2} \ \ S/\sqrt{n} = \overline{Y} \pm \frac{t_{n-1,1-\delta/2}}{z_{1-\delta/2}} \ \ z_{1-\delta/2} S/\sqrt{n}$$

where

$$\frac{t_{n-1,1-\delta/2}}{z_{1-\delta/2}} \to 1$$

as $n \to \infty$ is a small sample correction factor. The CI (5.2) should be used instead of the CI (5.1). If a large sample $100(1-\delta)\%$ CI for θ is $\hat{\theta} \pm z_{1-\delta/2} SE(\hat{\theta})$, then the large sample $100(1-\delta)\%$ CI $\quad \hat{\theta} \pm t_{d_n,1-\delta/2} SE(\hat{\theta})$ where $d_n \to \infty$ as $n \to \infty$ tends to perform better for small sample sizes. Typically the actual distribution of the asymptotic pivot has heavier tails than the N(0,1) distribution for moderate sample sizes, and using a correction factor improves performance.

5.2 Large Sample CIs and Tests

Large sample theory can be used to construct *confidence intervals* and *hypothesis tests*. Suppose that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and that $W_n \equiv W_n(\boldsymbol{Y})$ is an estimator of some parameter μ_W such that

$$\sqrt{n}(W_n - \mu_W) \xrightarrow{D} N(0, \sigma_W^2)$$

where σ_W^2/n is the asymptotic variance of the estimator W_n. The above notation means that if n is large, then for probability calculations

$$W_n - \mu_W \approx N(0, \sigma_W^2/n).$$

Suppose that S_W^2 is a consistent estimator of σ_W^2 so that the (asymptotic) *standard error* of W_n is $\mathrm{SE}(W_n) = S_W/\sqrt{n}$. Using the notation of Example 5.1,

$$P\left(-z_{1-\delta/2} \leq \frac{W_n - \mu_W}{SE(W_n)} \leq z_{1-\delta/2}\right) \to 1 - \delta$$

and a large sample $100(1-\delta)\%$ CI for μ_W is given by

$$[W_n - z_{1-\delta/2} SE(W_n), W_n + z_{1-\delta/2} SE(W_n)]. \tag{5.3}$$

Three common approximate level δ tests of hypotheses all use the *null hypothesis* $H_o : \mu_W = \mu_o$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W > \mu_o$, a left tailed test uses $H_A : \mu_W < \mu_o$, and a two tail test uses $H_A : \mu_W \neq \mu_o$. The test statistic is

$$t_o = \frac{W_n - \mu_o}{SE(W_n)},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \delta$.

Remark 5.2. Frequently the large sample CIs and tests can be improved for smaller samples by substituting a t distribution with d_n degrees of freedom for the standard normal distribution Z where d_n is some increasing function of the sample size n. Then the $100(1-\delta)\%$ CI for μ_W is given by

$$[W_n - t_{d_n,1-\delta/2} SE(W_n), W_n + t_{d_n,1-\delta/2} SE(W_n)]. \tag{5.4}$$

The test statistic rarely has an exact t_{d_n} distribution, but CI (5.6) often performs better than the CI (5.5) in small samples. The CI (5.6) is longer than the CI (5.5), and H_0 is less likely to be rejected. Hence the CI (5.6) is more *conservative* than the CI (5.5). This book will typically use very simple rules for d_n and not investigate the exact distribution of the test statistic. Note that the small sample correction factor

$$\frac{t_{d_n,1-\delta/2}}{z_{1-\delta/2}} \to 1$$

if $d_n \equiv p_n \to \infty$ as $n \to \infty$.

Paired and two sample procedures can be obtained directly from the one sample procedures. Suppose there are two samples $Y_1, ..., Y_n$ and $X_1, ..., X_m$. If $n = m$ and it is known that (Y_i, X_i) match up in correlated pairs, then *paired* CIs and tests apply the one sample procedures to the differences $D_i = Y_i - X_i$. Otherwise, assume the two samples are independent, that n and m are large, and that

$$\begin{pmatrix} \sqrt{n}(W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ \sqrt{m}(W_m(\boldsymbol{X}) - \mu_W(X)) \end{pmatrix} \xrightarrow{D} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_W(Y) & 0 \\ 0 & \sigma^2_W(X) \end{pmatrix} \right).$$

Then

$$\begin{pmatrix} (W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ (W_m(\boldsymbol{X}) - \mu_W(X)) \end{pmatrix} \approx N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_W(Y)/n & 0 \\ 0 & \sigma^2_W(X)/m \end{pmatrix} \right),$$

and

$$W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}) - (\mu_W(Y) - \mu_W(X)) \approx N\left(0, \frac{\sigma^2_W(Y)}{n} + \frac{\sigma^2_W(X)}{m}\right).$$

Hence $SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) =$

$$\sqrt{\frac{S^2_W(\boldsymbol{Y})}{n} + \frac{S^2_W(\boldsymbol{X})}{m}} = \sqrt{[SE(W_n(\boldsymbol{Y}))]^2 + [SE(W_m(\boldsymbol{X}))]^2},$$

and the large sample $100(1-\delta)\%$ CI for $\mu_W(Y) - \mu_W(X)$ is given by

$$(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) \pm z_{1-\delta/2} SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})).$$

Often approximate level δ tests of hypotheses use the *null hypothesis* H_o : $\mu_W(Y) = \mu_W(X)$. A right tailed test uses the *alternative hypothesis* H_A : $\mu_W(Y) > \mu_W(X)$, a left tailed test uses $H_A : \mu_W(Y) < \mu_W(X)$, and a two tail test uses $H_A : \mu_W(Y) \neq \mu_W(X)$. The test statistic is

$$t_o = \frac{W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})}{SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}))},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \delta$.

Remark 5.3. Again a t_{p_n} cutoff will often be used instead of the z cutoff. If d_n is the degrees of freedom used for a single sample procedure when the sample size is n, use $d_{n,m} = \min(d_n, d_m)$ for the two sample procedure if a better formula is not given. Then the large sample $100(1-\delta)\%$ CI for $\mu_W(Y) - \mu_W(X)$ is

$$(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) \pm t_{d_{n,m},1-\delta/2} SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})). \tag{5.5}$$

These CIs are known as *Welch intervals*. See Welch (1937) and Yuen (1974).

Example 5.2. Consider the single sample procedures where $W_n = \overline{Y}_n$. Then $\mu_W = E(Y)$, $\sigma_W^2 = \mathrm{VAR}(Y)$, $S_W = S_n$, and $d_n = n - 1$. Then the classical *t-interval* for $\mu \equiv E(Y)$ is

$$\overline{Y}_n \pm t_{n-1,1-\delta/2} \frac{S_n}{\sqrt{n}}$$

and the *t-test statistic* is

$$t_o = \frac{\overline{Y} - \mu_o}{S_n/\sqrt{n}}.$$

The right tailed p-value is given by $P(t_{n-1} > t_o)$.

Now suppose that there are two samples where $W_n(\boldsymbol{Y}) = \overline{Y}_n$ and $W_m(\boldsymbol{X}) = \overline{X}_m$. Then $\mu_W(Y) = E(Y) \equiv \mu_Y$, $\mu_W(X) = E(X) \equiv \mu_X$, $\sigma_W^2(Y) = \mathrm{VAR}(Y) \equiv \sigma_Y^2$, $\sigma_W^2(X) = \mathrm{VAR}(X) \equiv \sigma_X^2$, and $d_n = n - 1$. Let $d_{n,m} = \min(n-1, m-1)$. Since

$$SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) = \sqrt{\frac{S_n^2(\boldsymbol{Y})}{n} + \frac{S_m^2(\boldsymbol{X})}{m}},$$

the *two sample t-interval* for $\mu_Y - \mu_X$ is

$$(\overline{Y}_n - \overline{X}_m) \pm t_{d_{n,m},1-\delta/2}\sqrt{\frac{S_n^2(\boldsymbol{Y})}{n} + \frac{S_m^2(\boldsymbol{X})}{m}}$$

and *two sample t-test statistic* is

$$t_o = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\frac{S_n^2(\boldsymbol{Y})}{n} + \frac{S_m^2(\boldsymbol{X})}{m}}}.$$

The right tailed p-value is given by $P(t_{d_n,m} > t_o)$. For sample means, values of the degrees of freedom that are more accurate than $d_{n,m} = \min(n-1, m-1)$ can be computed. See Moore (2007, p. 474).

5.3 Some CI Examples

Example 5.3. Suppose that $Y_1, ..., Y_n$ are iid from a one parameter exponential family with parameter τ. Assume that $T_n = \sum_{i=1}^n t(Y_i)$ is a complete sufficient statistic. Then Olive (2014, pp. 92-93), often $T_n \sim G(na, 2b\ \tau)$ where a and b are known positive constants. Then

$$\hat{\tau} = \frac{T_n}{2nab}$$

is the UMVUE and often the MLE of τ. Since $T_n/(b\ \tau) \sim G(na, 2)$, a $100(1-\delta)\%$ confidence interval for τ is

$$\left[\frac{T_n/b}{G(na, 2, 1-\delta/2)}, \frac{T_n/b}{G(na, 2, \delta/2)}\right] \approx \left[\frac{T_n/b}{\chi^2_d(1-\delta/2)}, \frac{T_n/b}{\chi^2_d(\delta/2)}\right] \tag{5.6}$$

where $d = \lfloor 2na \rfloor$, $\lfloor x \rfloor$ is the greatest integer function (e.g. $\lfloor 7.7 \rfloor = \lfloor 7 \rfloor = 7$), $P[G \leq G(\nu, \lambda, \delta)] = \delta$ if $G \sim G(\nu, \lambda)$, and $P[X \leq \chi^2_d(\delta)] = \delta$ if X has a chi-square χ^2_d distribution with d degrees of freedom.

This confidence interval can be inverted to perform two tail tests of hypotheses. By Olive (2014, p. 186: Theorem 7.3), if $w(\theta)$ is increasing, then the uniformly most powerful (UMP) test of $H_o : \tau \leq \tau_o$ versus $H_A : \tau > \tau_o$ rejects H_0 if and only if $T_n > k$ where $P[G > k] = \delta$ when $G \sim G(na, 2b\ \tau_o)$. Hence

$$k = G(na, 2b\ \tau_o, 1-\delta). \tag{5.7}$$

A good approximation to this test rejects H_0 if and only if

$$T_n > b\ \tau_o \chi^2_d(1-\delta)$$

where $d = \lfloor 2na \rfloor$.

Example 5.4. Olive (2014, pp. 264-266): If Y is half normal HN(μ, σ) then the pdf of Y is

$$f(y) = \frac{2}{\sqrt{2\pi}\ \sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $y > \mu$ and μ is real. Since

$$f(y) = \frac{2}{\sqrt{2\pi}\ \sigma} I[y > \mu] \exp\left[\left(\frac{-1}{2\sigma^2}\right)(y-\mu)^2\right],$$

Y is a 1P–REF if μ is known.

Since $T_n = \sum(Y_i - \mu)^2 \sim G(n/2, 2\sigma^2)$, in Example 5.3 take $a = 1/2$, $b = 1$, $d = n$ and $\tau = \sigma^2$. Then a $100(1-\delta)\%$ confidence interval for σ^2 is

$$\left[\frac{T_n}{\chi^2_n(1-\delta/2)}, \frac{T_n}{\chi^2_n(\delta/2)}\right]. \tag{5.8}$$

The UMP test of $H_0 : \sigma^2 \leq \sigma^2_o$ versus $H_A : \sigma^2 > \sigma^2_o$ rejects H_o if and only if

$$T_n/\sigma^2_o > \chi^2_n(1-\delta).$$

Now consider inference when both μ and σ are unknown. Then the family is no longer an exponential family since the support depends on μ. Let

$$D_n = \sum_{i=1}^{n} (Y_i - Y_{1:n})^2. \tag{5.9}$$

Pewsey (2002) showed that $(\hat{\mu}, \hat{\sigma}^2) = (Y_{1:n}, \frac{1}{n} D_n)$ is the MLE of (μ, σ^2), and that

$$\frac{Y_{1:n} - \mu}{\sigma \Phi^{-1}(\frac{1}{2} + \frac{1}{2n})} \xrightarrow{D} EXP(1)$$

where $Y_{1:n} = Y_{(1)} = \min(Y_1, ..., Y_n)$ is the first order statistic. Since $(\sqrt{\pi/2})/n$ is an approximation to $\Phi^{-1}(\frac{1}{2} + \frac{1}{2n})$ based on a first order Taylor series expansion such that

$$\frac{\Phi^{-1}(\frac{1}{2} + \frac{1}{2n})}{(\sqrt{\pi/2})/n} \to 1,$$

it follows that

$$\frac{n(Y_{1:n} - \mu)}{\sigma\sqrt{\frac{\pi}{2}}} \xrightarrow{D} EXP(1). \tag{5.10}$$

Using this fact, it can be shown that a large sample $100(1-\delta)\%$ CI for μ is

$$[\hat{\mu} + \hat{\sigma} \log(\delta)\ \Phi^{-1}(\frac{1}{2} + \frac{1}{2n})\ (1 + 13/n^2),\quad \hat{\mu}] \tag{5.11}$$

where the term $(1 + 13/n^2)$ is a small sample correction factor.

Note that

$$D_n = \sum_{i=1}^{n}(Y_i - Y_{1:n})^2 = \sum_{i=1}^{n}(Y_i - \mu + \mu - Y_{1:n})^2 =$$

$$\sum_{i=1}^{n}(Y_i - \mu)^2 + n(\mu - Y_{1:n})^2 + 2(\mu - Y_{1:n})\sum_{i=1}^{n}(Y_i - \mu).$$

Hence

$$D_n = T_n + \frac{1}{n}[n(Y_{1:n} - \mu)]^2 - 2[n(Y_{1:n} - \mu)]\frac{\sum_{i=1}^{n}(Y_i - \mu)}{n},$$

or

$$\frac{D_n}{\sigma^2} = \frac{T_n}{\sigma^2} + \frac{1}{n}\frac{1}{\sigma^2}[n(Y_{1:n} - \mu)]^2 - 2[\frac{n(Y_{1:n} - \mu)}{\sigma}]\frac{\sum_{i=1}^{n}(Y_i - \mu)}{n\sigma}. \quad (5.12)$$

Consider the three terms on the right hand side of (5.12). The middle term converges to 0 in distribution while the third term converges in distribution to a $-2EXP(1)$ or $-\chi^2_2$ distribution since $\sum_{i=1}^{n}(Y_i - \mu)/(\sigma n)$ is the sample mean of HN(0,1) random variables and $E(X) = \sqrt{2/\pi}$ when $X \sim HN(0, 1)$.

Let $T_{n-p} = \sum_{i=1}^{n-p}(Y_i - \mu)^2$. Then

$$D_n = T_{n-p} + \sum_{i=n-p+1}^{n}(Y_i - \mu)^2 - V_n \quad (5.13)$$

where

$$\frac{V_n}{\sigma^2} \xrightarrow{D} \chi^2_2.$$

Hence

$$\frac{D_n}{T_{n-p}} \xrightarrow{D} 1$$

and D_n/σ^2 is asymptotically equivalent to a χ^2_{n-p} random variable where p is an arbitrary nonnegative integer. Pewsey (2002) used $p = 1$.

Thus when both μ and σ^2 are unknown, a large sample $100(1 - \delta)\%$ confidence interval for σ^2 is

$$\left[\frac{D_n}{\chi^2_{n-1}(1 - \delta/2)}, \frac{D_n}{\chi^2_{n-1}(\delta/2)}\right]. \quad (5.14)$$

It can be shown that $\sqrt{n}$ CI length converges in probability to $\sigma^2\sqrt{2}(z_{1-\delta/2} - z_{\delta/2})$ for CIs (5.8) and (5.14) while n length CI (5.11) converges in probability to $-\sigma \log(\delta)\sqrt{\pi/2}$.

When μ and σ^2 are unknown, an approximate δ level test of $H_o : \sigma^2 \leq \sigma_o^2$ versus $H_A : \sigma^2 > \sigma_o^2$ that rejects H_o if and only if

$$D_n/\sigma_o^2 > \chi^2_{n-1}(1-\delta) \tag{5.15}$$

has nearly as much power as the δ level UMP test when μ is known if n is large.

Example 5.5. Let $X_1, ..., X_n$ be iid Poisson(θ) random variables. The classical large sample 100 $(1-\delta)$% CI for θ is

$$\overline{X} \pm z_{1-\delta/2}\sqrt{\overline{X}/n}$$

where $P(Z \leq z_{1-\delta/2}) = 1 - \delta/2$ if $Z \sim N(0,1)$.

Following Byrne and Kabaila (2005), a modified large sample 100 $(1-\delta)$% CI for θ is $[L_n, U_n]$ where

$$L_n = \frac{1}{n}\left(\sum_{i=1}^{n} X_i - 0.5 + 0.5z^2_{1-\delta/2} - z_{1-\delta/2}\sqrt{\sum_{i=1}^{n} X_i - 0.5 + 0.25z^2_{1-\delta/2}}\right)$$

and

$$U_n = \frac{1}{n}\left(\sum_{i=1}^{n} X_i + 0.5 + 0.5z^2_{1-\delta/2} + z_{1-\delta/2}\sqrt{\sum_{i=1}^{n} X_i + 0.5 + 0.25z^2_{1-\delta/2}}\right).$$

Following Grosh (1989, p. 59, 197–200), let $W = \sum_{i=1}^{n} X_i$ and suppose that $W = w$ is observed. Let $P(T < \chi^2_d(\delta)) = \delta$ if $T \sim \chi^2_d$. Then an "exact" 100 $(1-\delta)$% CI for θ is

$$\left[\frac{\chi^2_{2w}(\frac{\delta}{2})}{2n}, \frac{\chi^2_{2w+2}(1-\frac{\delta}{2})}{2n}\right]$$

for $w \neq 0$ and

$$\left[0, \frac{\chi^2_2(1-\delta)}{2n}\right]$$

for $w = 0$.

The "exact" CI is conservative: the actual coverage $(1-\delta_n) \geq 1-\delta =$ the nominal coverage. This interval performs well if θ is very close to 0. See Problem 5.2.

Example 5.6. Let $Y_1, ..., Y_n$ be iid bin$(1, \rho)$. Let $\hat{\rho} = \sum_{i=1}^{n} Y_i/n =$ number of "successes"$/n$. The classical large sample 100 $(1-\delta)$% CI for ρ is

$$\hat{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n}}$$

where $P(Z \leq z_{1-\delta/2}) = 1 - \delta/2$ if $Z \sim N(0,1)$.

The Agresti Coull CI takes $\tilde{n} = n + z^2_{1-\delta/2}$ and

$$\tilde{\rho} = \frac{n\hat{\rho} + 0.5z^2_{1-\delta/2}}{n + z^2_{1-\delta/2}}.$$

(The method "adds" $0.5z^2_{1-\delta/2}$ "0's" and $0.5z^2_{1-\delta/2}$ "1's" to the sample, so the "sample size" increases by $z^2_{1-\delta/2}$.) Then the large sample 100 $(1-\delta)$% Agresti Coull CI for ρ is

$$\tilde{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\tilde{\rho}(1-\tilde{\rho})}{\tilde{n}}}.$$

Now let $Y_1, ..., Y_n$ be independent bin(m_i, ρ) random variables, let $W = \sum_{i=1}^n Y_i \sim \text{bin}(\sum_{\text{i=1}}^{\text{n}} \text{m}_\text{i}, \rho)$ and let $n_w = \sum_{i=1}^n m_i$. Often $m_i \equiv 1$ and then $n_w = n$. Let $P(F_{d_1,d_2} \leq F_{d_1,d_2}(\delta)) = \delta$ where F_{d_1,d_2} has an F distribution with d_1 and d_2 degrees of freedom. Assume $W = w$ is observed. Then the Clopper Pearson "exact" 100 $(1-\delta)$% CI for ρ is

$$\left[0, \frac{1}{1 \ + \ n_w \ F_{2n_w,2}(\delta)}\right] \ \text{ for } \ \text{w} = 0,$$

$$\left[\frac{n_w}{n_w \ + \ F_{2,2n_w}(1-\delta)}, 1\right] \ \text{ for } \ \text{w} = \text{n}_\text{w},$$

and $[\rho_L, \rho_U]$ for $0 < w < n_w$ with

$$\rho_L = \frac{w}{w + (n_w - w + 1)F_{2(n_w-w+1),2w}(1-\delta/2)}$$

and

$$\rho_U = \frac{w+1}{w + 1 + (n_w - w)F_{2(n_w-w),2(w+1)}(\delta/2)}.$$

The "exact" CI is conservative: the actual coverage $(1-\delta_n) \geq 1-\delta =$ the nominal coverage. This interval performs well if ρ is very close to 0 or 1. The classical interval should only be used if it agrees with the Agresti Coull interval. See Problem 5.3.

Example 5.7. Let $\hat{\rho}$ = number of "successes"$/n$. Consider a taking a simple random sample of size n from a finite population of known size N. Then the classical finite population large sample 100 $(1-\delta)$% CI for ρ is

$$\hat{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n-1}\left(\frac{N-n}{N}\right)} = \hat{\rho} \pm z_{1-\delta/2}SE(\hat{\rho}) \tag{5.16}$$

where $P(Z \leq z_{1-\delta/2}) = 1 - \delta/2$ if $Z \sim N(0,1)$.

Following DasGupta (2008, p. 121), suppose the number of successes Y has a hypergeometric $(C, N-C, n)$ where $p = C/N$. If $n/N \approx \lambda \in (0,1)$ where n and N are both large, then

$$\hat{\rho} \approx N\left(\rho, \frac{\rho(1-\rho)(1-\lambda)}{n}\right).$$

Hence CI (5.16) should be good if the above normal approximation is good.

Let $\tilde{n} = n + z^2_{1-\delta/2}$ and

$$\tilde{\rho} = \frac{n\hat{\rho} + 0.5z^2_{1-\delta/2}}{n + z^2_{1-\delta/2}}.$$

(Heuristically, the method adds $0.5z^2_{1-\delta/2}$ "0's" and $0.5z^2_{1-\delta/2}$ "1's" to the sample, so the "sample size" increases by $z^2_{1-\delta/2}$.) Then a large sample 100 $(1-\delta)$% Agresti Coull type (ACT) finite population CI for ρ is

$$\tilde{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\tilde{\rho}(1-\tilde{\rho})}{\tilde{n}}\left(\frac{N-n}{N}\right)} = \tilde{\rho} \pm z_{1-\delta/2}SE(\tilde{\rho}). \tag{5.17}$$

Notice that a 95% CI uses $z_{1-\delta/2} = 1.96 \approx 2$.

For data from a finite population, large sample theory gives useful approximations as N and $n \to \infty$ and $n/N \to 0$. Hence theory suggests that the ACT CI should have better coverage than the classical CI if the p is near 0 or 1, if the sample size n is moderate, and if n is small compared to the population size N. The coverage of the classical and ACT CIs should be very similar if n is large enough but small compared to N (which may only be possible if N is enormous). As n increases to N, $\hat{\rho}$ goes to p, SE$(\hat{\rho})$ goes to 0, and the classical CI may perform well. SE$(\tilde{\rho})$ also goes to 0, but $\tilde{\rho}$ is a biased estimator of ρ and the ACT CI will not perform well if n/N is too large.

Want an interval that gives good coverage even if ρ is near 0 or 1 or if n/N is large. A simple method is to combine the two intervals. Let $[L_C, U_C]$ and $[L_A, U_A]$ be the classical and ACT $100(1-\delta)$% intervals. Let the modified $100(1-\delta)$% interval be

$$[\max[0, \min(L_C, L_U)], \min[1, \max(U_C, U_A)]]. \tag{5.18}$$

The modified interval seems to perform well. See Problem 5.4.

Example 5.8. Assume $Y_1, ..., Y_n$ are iid with mean μ and variance σ^2. Bickel and Doksum (2007, p. 279) suggest that

$$W_n = n^{-1/2}\left[\frac{(n-1)S^2}{\sigma^2} - n\right]$$

can be used as an asymptotic pivot for σ^2 if $E(Y^4) < \infty$. Notice that $W_n =$

$$n^{-1/2}\left[\frac{\sum(Y_i-\mu)^2}{\sigma^2} \; - \; \frac{n(\overline{Y}-\mu)^2}{\sigma^2} \; - \; n\right] =$$

$$\sqrt{n}\left[\frac{\sum\left(\frac{Y_i-\mu}{\sigma}\right)^2}{n} \; - \; 1\right] \; - \; \frac{1}{\sqrt{n}}\; n\; \left(\frac{\overline{Y}-\mu}{\sigma}\right)^2 = X_n - Z_n.$$

Since $\sqrt{n}Z_n \xrightarrow{D} \chi^2_1$, the term $Z_n \xrightarrow{D} 0$. Now $X_n = \sqrt{n}(\overline{U} - 1) \xrightarrow{D} N(0, \tau)$ by the CLT since $U_i = [(Y_i - \mu)/\sigma]^2$ has mean $E(U_i) = 1$ and variance

$$V(U_i) = \tau = E(U_i^2) - (E(U_i))^2 = \frac{E[(Y_i-\mu)^4]}{\sigma^4} - 1 = \kappa + 2$$

where κ is the kurtosis of Y_i. Thus $W_n \xrightarrow{D} N(0, \tau)$.

Hence

$$1 - \alpha \approx P(-z_{1-\alpha/2} < \frac{W_n}{\sqrt{\tau}} < z_{1-\alpha/2}) = P(-z_{1-\alpha/2}\sqrt{\tau} < W_n < z_{1-\alpha/2}\sqrt{\tau})$$

$$= P(-z_{1-\alpha/2}\sqrt{n\tau} < \frac{(n-1)S^2}{\sigma^2} - n < z_{1-\alpha/2}\sqrt{n\tau})$$

$$= P(n - z_{1-\alpha/2}\sqrt{n\tau} < \frac{(n-1)S^2}{\sigma^2} < n + z_{1-\alpha/2}\sqrt{n\tau}).$$

Hence a large sample $100(1-\alpha)\%$ CI for σ^2 is

$$\left[\frac{(n-1)S^2}{n + z_{1-\alpha/2}\sqrt{n\hat{\tau}}}, \; \frac{(n-1)S^2}{n - z_{1-\alpha/2}\sqrt{n\hat{\tau}}}\right]$$

where

$$\hat{\tau} = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^4}{S^4} - 1.$$

Notice that this CI needs $n > z_{1-\alpha/2}\sqrt{n\hat{\tau}}$ for the right endpoint to be positive. It can be shown that $\sqrt{n}$ (length CI) converges to $2\sigma^2 z_{1-\alpha/2}\sqrt{\tau}$ in probability.

Problem 5.7 uses an asymptotically equivalent $100(1-\alpha)\%$ CI of the form

$$\left[\frac{(n-a)S^2}{n + t_{n-1,1-\alpha/2}\sqrt{n\hat{\tau}}}, \ \frac{(n+b)S^2}{n - t_{n-1,1-\alpha/2}\sqrt{n\hat{\tau}}} \right]$$

where a and b depend on $\hat{\tau}$. The goal was to make a 95% CI with good coverage for a wide variety of distributions (with 4th moments) for $n \geq 100$. The price is that the CI is too long for some of the distributions with small kurtosis. The $N(\mu, \sigma^2)$ distribution has $\tau = 2$, while the EXP(λ) distribution has $\sigma^2 = \lambda^2$ and $\tau = 8$. The quantity τ is small for the uniform distribution but large for the lognormal LN(0,1) distribution.

By the binomial theorem, if $E(Y^4)$ exists and $E(Y) = \mu$ then

$$E(Y-\mu)^4 = \sum_{j=0}^{4} \binom{4}{j} E[Y^j](-\mu)^{4-j} =$$

$$\mu^4 - 4\mu^3 E(Y) + 6\mu^2(V(Y) + [E(Y)]^2) - 4\mu E(Y^3) + E(Y^4).$$

This fact can be useful for computing

$$\tau = \frac{E[(Y_i - \mu)^4]}{\sigma^4} - 1 = \kappa + 2.$$

Example 5.9. Following DasGupta (2008, p. 402-404), consider the pooled t CI for $\mu_1 - \mu_2$. Let $X_1, ..., X_{n_1}$ be iid with mean μ_1 and variance σ_1^2. Let $Y_1, ..., Y_{n_2}$ be iid with mean μ_2 and variance σ_2^2. Assume that the two samples are independent and that $n_i \to \infty$ for $i = 1, 2$ in such a way that $\hat{\rho} = \frac{n_1}{n_1+n_2} \to \rho \in (0, 1)$. Let $\theta = \sigma_2^2/\sigma_1^2$, and let the pooled sample variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Then

$$\begin{pmatrix} \sqrt{n_1}(\overline{X} - \mu_1) \\ \sqrt{n_2}(\overline{Y} - \mu_2) \end{pmatrix} \xrightarrow{D} N_2(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2)$. Hence

$$\sqrt{n}[(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)] \xrightarrow{D} N(0, \frac{\sigma_1^2}{\pi_1} + \frac{\sigma_2^2}{\pi_2}).$$

So

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \xrightarrow{D} N(0, 1).$$

Thus

$$\frac{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\ \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\ =\ \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\xrightarrow{D}N(0,\tau^2)$$

where

$$\frac{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}{(\frac{1}{n_1}+\frac{1}{n_2})\frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2}}=\frac{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}{\hat{\rho}\sigma_1^2+(1-\hat{\rho})\sigma_2^2}\ \frac{1/\sigma_1^2}{1/\sigma_1^2}\ \frac{n_1n_2}{n_1+n_2}$$

$$=\frac{\frac{1}{n_1}+\frac{\theta}{n_2}}{\hat{\rho}+(1-\hat{\rho})\theta}\ \frac{n_1n_2}{n_1+n_2}\xrightarrow{D}\frac{1-\rho+\rho\theta}{\rho+(1-\rho)\theta}=\tau^2.$$

Now let $\hat{\theta}=S_2^2/S_1^2$ and

$$\hat{\tau}^2=\frac{1-\hat{\rho}+\hat{\rho}\ \hat{\theta}}{\hat{\rho}+(1-\hat{\rho})\ \hat{\theta}}.$$

Notice that $\hat{\tau}=1$ if $\hat{\rho}=1/2$, and $\hat{\tau}=1$ if $\hat{\theta}=1$.

The usual large sample $(1-\alpha)100\%$ pooled t CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ t_{n_1+n_2-2,1-\alpha/2}\ \ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}. \tag{5.19}$$

The large sample theory says that this CI is valid if $\tau=1$, and that

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\hat{\tau}\ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\xrightarrow{D}N(0,1).$$

Hence a large sample $(1-\alpha)100\%$ CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ z_{1-\alpha/2}\ \hat{\tau}\ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Then the large sample $(1-\alpha)100\%$ modified pooled t CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ t_{n_1+n_2-4,1-\alpha/2}\ \hat{\tau}\ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}. \tag{5.20}$$

The large sample $(1-\alpha)100\%$ Welch CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ t_{d,1-\alpha/2}\ \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{5.21}$$

where $d=\max(1,[d_0])$, and

$$d_0 = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{S_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{S_2^2}{n_2})^2}.$$

Suppose $n_1/(n_1+n_2) \to \rho$. It can be shown that if the CI length is multiplied by $\sqrt{n_1}$, then the scaled length of the pooled t CI converges in probability to $2z_{1-\alpha/2}\sqrt{\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2}$ while the scaled lengths of the modified pooled t CI and Welch CI both converge in probability to $2z_{1-\alpha/2}\sqrt{\sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2}$.

Example 5.10. Hesterberg (2014) gives the following two competitors of the t interval given by Equation (5.2): the skewness adjusted t interval is

$$[\,\overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(1+2t_{n-1,1-\alpha/2}^2)-t_{n-1,1-\alpha/2}],\ \overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(1+2t_{n-1,1-\alpha/2}^2)+t_{n-1,1-\alpha/2}]\,], \tag{5.22}$$

and the asymptotic percentile t CI is

$$[\,\overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(t_{n-1,1-\alpha/2}-1)^2-t_{n-1,1-\alpha/2}],\ \overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(t_{n-1,1-\alpha/2}-1)^2+t_{n-1,1-\alpha/2}] \tag{5.23}$$

where

$$\hat{\kappa} = \frac{\hat{\gamma}}{6\sqrt{n}} \quad \text{with} \quad \hat{\gamma} = \frac{1}{\mathrm{nS}^3}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i} - \overline{\mathrm{Y}})^3.$$

Another competitor is the Johnson (1978) CI is

$$[\,\overline{Y} + \frac{\hat{\mu}_3}{6S^2n} - t_{n-1,1-\alpha/2}\ \ S/\sqrt{n},\ \ \overline{Y} + \frac{\hat{\mu}_3}{6S^2n} + t_{n-1,1-\alpha/2}\ \ S/\sqrt{n}\,] \tag{5.24}$$

where $\mu_3 = E[(Y-\mu)^3]$ and

$$\hat{\mu}_3 = S^3\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^3.$$

The t–interval (5.2) may perform better if the distribution has second moments but does not have third or fourth moments. McKinney (2021) gave some more competitors. The Johnson (1978) CI (5.24) appeared to be best, but only very slightly better than the usual t– interval (5.2).

5.4 Bootstrap Confidence Regions and Hypothesis Tests

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 4.2 to a bootstrap sample results in a confidence region. The volume of a confidence region $\to 0$ as $n \to 0$, while the volume of a prediction region goes to that of a population region that

would contain a new $\boldsymbol{x}_f$ with probability $1-\delta$. The nominal coverage is $100(1-\delta)$. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, then the region is *conservative*. If $100(1-\delta_n) < 100(1-\delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal.

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of $\theta = \mu$ is about 95%. Hence the probability that θ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\theta - 1.96S/\sqrt{n}, \theta + 1.96S/\sqrt{n}]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if θ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}]$ is a large sample 95% confidence interval for the population mean θ. Note that the lengths of the two intervals are the same. Where the interval is centered, at the parameter θ or the statistic $\overline{Y}_n$, determines whether the interval is a prediction or a confidence interval. See Theorem 5.2 for a similar relationship between confidence regions and prediction regions. Let $\boldsymbol{\theta}$ be a $g \times 1$ vector of parameters.

Definition 5.5. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

If $\mathcal{A}_n$ is based on a squared Mahalanobis distance D^2 with a limiting distribution that has a pdf, we often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$.

There are several methods for obtaining a bootstrap sample $T_1^*, ..., T_B^*$ where the sample size n is suppressed: $T_i^* = T_{in}^*$. The parametric bootstrap, nonparametric bootstrap, and residual bootstrap will be used in this text. Applying nonparametric prediction region (4.11) to the bootstrap sample will result in a confidence region for $\boldsymbol{\theta}$. When $g = 1$, applying the percentile PI (4.1) or the shorth PI (4.4) to the bootstrap sample results in a confidence interval for θ. Section 5.4.2 will help clarify ideas.

When $g = 1$, a confidence interval is a special case of a confidence region. One sided confidence intervals give a lower or upper confidence bound for θ. A large sample $100(1-\delta)\%$ lower confidence interval $(-\infty, U_n]$ uses an upper confidence bound U_n and is in the lower tail of the distribution of $\hat{\theta}$. A large sample $100(1-\delta)\%$ upper confidence interval $[L_n, \infty)$ uses a lower confidence bound L_n and is in the upper tail of the distribution of $\hat{\theta}$. These CIs can be useful if $\theta \in [a, b]$ and $\theta = a$ or $\theta = b$ is of interest for a hypothesis test. For example, $[a, b] = [0, 1]$ if $\theta = \rho^2$, the squared population correlation. Then use $[0, U_n]$ and $[L_n, 1]$ as CIs, e.g. if we expect $\theta = 0$ we might test $H_0 : \theta \leq 0.05$ versus $H_0 : \theta > 0.05$, and fail to reject H_0 if $U_n < 0.05$. Again we often want the probability to converge to $1-\delta$ if the confidence interval is based on a statistic with an asymptotic distribution that has a pdf.

Definition 5.6. The interval $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ *confidence interval* for θ if $P(L_n \leq \theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $(-\infty, U_n]$ is a large sample $100(1-\delta)\%$ *lower confidence interval* for θ if $P(\theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $[L_n, \infty)$ is large sample $100(1-\delta)\%$ *upper confidence interval* for θ if $P(\theta \geq L_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

Next we discuss bootstrap confidence intervals that are obtained by applying prediction intervals (4.1) and (4.4) to the bootstrap sample. Some additional bootstrap CIs are obtained from bootstrap confidence regions from Definition 5.12 when $g = 1$. See Efron (1982) and Chen (2016) for the percentile method CI. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^{n} Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample. The CI (5.25) is obtained by applying PI (4.1) to the bootstrap sample with B used instead of n. Hence (5.25) is also a large sample prediction interval for a future value of T_f^* if the T_i^* are iid from the empirical distribution discussed in Section 5.4.1.

Definition 5.7. The bootstrap percentile method large sample $100(1-\delta)\%$ confidence interval for θ is an interval $[T_{(k_L)}^*, T_{(K_U)}^*]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T_{(k_1)}^*, T_{(k_2)}^*]. \tag{5.25}$$

The large sample $100(1-\delta)\%$ *lower percentile method* CI for θ is $(-\infty, T_{(\lceil B(1-\delta) \rceil)}^*]$. The large sample $100(1-\delta)\%$ *upper percentile method* CI for θ is $[T_{(\lceil B\delta \rceil)}^*, \infty)$.

In the next defiinition, the large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length, denoted by $[T_{(s)}^*, T_{(s+c-1)}^*]$.

Definition 5.8. The large sample $100(1-\delta)\%$ *lower shorth* CI for θ is $(-\infty, T_{(c)}^*]$, while the large sample $100(1-\delta)\%$ *upper shorth* CI for θ is $[T_{(B-c+1)}^*, \infty)$. The large sample $100(1-\delta)\%$ *shorth*(c) *CI*

$$[T_{(s)}^*, T_{(s+c-1)}^*] \text{ where } c = \min(B, \lceil B[1-\delta+1.12\sqrt{\delta/B}\,] \rceil). \tag{5.26}$$

Applied to a bootstrap sample, the shorth CI can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. See Remark 5.8 for some theory for bootstrap CIs such as (5.25) and (5.26).

5.4.1 The Bootstrap

This subsection illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 5.9. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 5.11. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 5.9. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \frac{n-1}{n}\boldsymbol{S} \equiv \frac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \quad \square$$

Example 5.12. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 5.13. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $\mathrm{Cov}(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T_{1n}^*, ..., T_{Bn}^*$ is formed where $T_{jn}^* = t(\boldsymbol{w}_{j1}^*, ..., \boldsymbol{w}_{jn}^*)$. Theorem 5.1 will show that $\sqrt{B}(T_{1n}^* - T_n), ..., \sqrt{B}(T_{Bn}^* - T_n)$ is pseudodata for $\sqrt{n}(T_{1n} - \boldsymbol{\theta}), ..., \sqrt{n}(T_{Bn} - \boldsymbol{\theta})$ when n and B are large in that $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{B}(T^* - T_n) \xrightarrow{D} \boldsymbol{u}$.

Example 5.14. Suppose there is training data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ for the model $\boldsymbol{y}_i = m(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$, and it is desired to predict a future test value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and the training data. The model can be fit and the residual vectors formed. One method for obtaining a prediction region for $\boldsymbol{y}_f$ is to form the pseudodata $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$, and apply the nonparametric prediction region (4.11) to the pseudodata. See Olive (2017b, 2018). The residual

bootstrap could also be used to make a bootstrap sample $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_1^*, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_B^*$ where the $\hat{\boldsymbol{\epsilon}}_j^*$ are selected with replacement from the residual vectors for $j = 1, ..., B$. As $B \to \infty$, the bootstrap sample will take on the n values $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ (the pseudodata) with probabilities converging to $1/n$ for $i = 1, ..., n$.

Suppose there is a statistic T_n that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* \quad \text{and} \quad \boldsymbol{S}_T^* = \frac{1}{B-1}\sum_{i=1}^{B}(T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \tag{5.27}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = T_{i,n}^*$. Fix n, and let $E(T_{i,n}^*) = \boldsymbol{\theta}_n$ and $\mathrm{Cov}(T_{i,n}^*) = \boldsymbol{\Sigma}_n$.

We will often assume that $\mathrm{Cov}(T_n) = \boldsymbol{\Sigma}_T$, and $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ where $\boldsymbol{\Sigma}_A > 0$ is positive definite and nonsingular. Often $n\hat{\boldsymbol{\Sigma}}_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. For example, using least squares and the residual bootstrap for the multiple linear regression model, $\boldsymbol{\Sigma}_n = \dfrac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $T_n = \boldsymbol{\theta}_n = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\hat{\boldsymbol{\Sigma}}_T = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\boldsymbol{\Sigma}_A = \sigma^2 \lim_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}$.

Suppose the $T_i^* = T_{i,n}^*$ are iid from some distribution with cdf $\tilde{F}_n$. For example, if $T_{i,n}^* = t(F_n^*)$ where iid samples from F_n are used, then $\tilde{F}_n$ is the cdf of $t(F_n^*)$. With respect to $\tilde{F}_n$, both $\boldsymbol{\theta}_n$ and $\boldsymbol{\Sigma}_n$ are parameters, but with respect to F, $\boldsymbol{\theta}_n$ is a random vector and $\boldsymbol{\Sigma}_n$ is a random matrix. For fixed n, by the multivariate central limit theorem,

$$\sqrt{B}(\overline{T}^* - \boldsymbol{\theta}_n) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_n) \quad \text{and} \quad \mathrm{B}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n})^\mathrm{T}[\boldsymbol{S}_\mathrm{T}^*]^{-1}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n}) \xrightarrow{\mathrm{D}} \chi^2_\mathrm{r}$$

as $B \to \infty$.

Remark 5.4. For Examples 5.11 and 5.14, the bootstrap works but is expensive compared to alternative methods. For Example 5.11, fix n, then $\overline{T}^* \xrightarrow{P} \boldsymbol{\theta}_n = \overline{\boldsymbol{x}}$ and $\boldsymbol{S}_T^* \xrightarrow{P} (n-1)\boldsymbol{S}/n$ as $B \to \infty$, but using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ makes more sense. For Example 5.14, use the pseudodata instead of the residual bootstrap. For these two examples, it is known how the bootstrap sample behaves as $B \to \infty$. The bootstrap can be very useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, but it not known how to estimate $\boldsymbol{\Sigma}_A$ without using a resampling method like the bootstrap. The bootstrap may be useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, but the limiting distribution (the distribution of $\boldsymbol{u}$) is unknown.

The following theorem shows that $\sqrt{m}(T_{1,n}^* - T_n), ..., \sqrt{m}(T_{B,n}^* - T_n)$ are pseudodata for $\sqrt{n}(T_{1,n} - \boldsymbol{\theta}), ..., \sqrt{n}(T_{B,n} - \boldsymbol{\theta})$. Here $T_i^* = T_{i,m}^*$ with n suppressed or $T_{i,n}^* = T_{i,n,m}^*$ where m is the sample size of the bootstrap data set used to compute T_i^*. Often $m = n$ for the nonparametric bootstrap. The first two convergence assumptions are with respect to the data distribution, while the third convergence assumption is with respect to the bootstrap dis-

tribution. The technique is similar to using a triangular array, except both $n \to \infty$ and $m \to \infty$. Note that for large n, $N_g(\mathbf{0}, \boldsymbol{\Sigma}_n) \approx N_g(\mathbf{0}, \boldsymbol{\Sigma})$, and often the $N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ approximation is used to produce output since $\boldsymbol{\Sigma}$ is unknown. Typically large sample theory is used to prove the three assumptions of the following theorem.

Theorem 5.1, Bootstrap Proof Technique: Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_n \xrightarrow{P} \boldsymbol{\Sigma}$ as $n \to \infty$, and for fixed n, $\sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ as $m \to \infty$. Then a) $\sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $m, n \to \infty$. Also b) $\sqrt{n}(T^*_n - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$ where $T^*_n = T^*_{n,n}$ has $m = n$.

Proof: By the three assumptions, $\boldsymbol{u}_n = \sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$, $\boldsymbol{w}^*_{n,m} = \sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} \boldsymbol{w}_n \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ as $m \to \infty$ for fixed n, and $\boldsymbol{w}_n \xrightarrow{D} \boldsymbol{u}$ as $n \to \infty$. Hence $\boldsymbol{w}^*_{n,m} = \sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} \boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $m, n \to \infty$. Since this result does not depend on m as long as $m \to \infty$, b) follows. □

Example 5.15. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors with $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = \boldsymbol{\Sigma}$. a) For the parametric bootstrap, let $\boldsymbol{x}^*_1, ..., \boldsymbol{x}^*_m$ be iid $N_p(\overline{\boldsymbol{x}}_n, \boldsymbol{S}_n)$ where $\boldsymbol{S}_n \xrightarrow{P} \boldsymbol{\Sigma}$ as $n \to \infty$. By the multivariate central limit theorem $\sqrt{n}(\overline{\boldsymbol{x}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and for fixed n, $\sqrt{m}(\overline{\boldsymbol{x}}^*_{n,m} - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{S}_n)$ where $\overline{\boldsymbol{x}}^*_{n,m} = \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}^*_i$ is the sample mean of the bootstrap data set $\boldsymbol{x}^*_1, ..., \boldsymbol{x}^*_m$. Hence $\sqrt{m}(\overline{\boldsymbol{x}}^*_{n,m} - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ as $n, m \to \infty$ by Theorem 5.1. Note that $m = n$ can be used by Theorem 5.1 b).

b) For the nonparametric bootstrap, $E(\overline{\boldsymbol{x}}^*_n) = E(\boldsymbol{w}_n) = \overline{\boldsymbol{x}}_n$, and $\mathrm{Cov}(\overline{\boldsymbol{x}}^*_n) = \mathrm{Cov}(\boldsymbol{w}_n)/n = (n-1)\boldsymbol{S}_n/n^2$ by Example 5.11 where $\boldsymbol{w} = \boldsymbol{w}_n$. The $\boldsymbol{x}^*_i$ are iid with respect to the bootstrap distribution. If the sample mean $\overline{\boldsymbol{x}}^*_{n,m}$ is computed from m $\boldsymbol{x}^*_i$ selected with replacement from the $\boldsymbol{x}_i$, then $\sqrt{m}(\overline{\boldsymbol{x}}^*_{n,m} - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \frac{n-1}{n}\boldsymbol{S}_n)$ for fixed n by the multivariate CLT. Then by Theorem 5.1 b) with $m = n$, $\sqrt{n}(\overline{\boldsymbol{x}}^*_n - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$.

5.4.2 Bootstrap Confidence Regions for Hypothesis Testing

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1 - \delta$ as $n, B \to \infty$. The B is often suppressed. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T^*_1, ..., T^*_B$ be the bootstrap sample

for T_n. Let $\boldsymbol{A}$ be a full rank $g \times p$ constant matrix. For variable selection using notation from Chapter 6, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \boldsymbol{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The statistic $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the variable selection estimator padded with zeroes.

Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. See Equation (5.27). Let $k_B = \lceil B(1-\delta) \rceil$.

Definition 5.10. a) The standard bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{5.28}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$. b) The Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(k_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \leq D^2_{(k_B,T)}\} \tag{5.29}$$

where the cutoff $D^2_{(k_B,T)}$ is the $100k_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (T_i^* - T_n) = n(T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A]^{-1} (T_i^* - T_n)$.

Confidence region (5.28) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption. Bickel and Ren (2001) have interesting sufficient conditions for (5.29) to be a confidence region when $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of positive definite $\boldsymbol{\Sigma}_A$. Let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and the bootstrapped statistic $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from $F_n =$ "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{u}$.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method applies prediction region (4.11) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2021a). Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{5.30}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let $D_{(U_B)}$ be the $100q_B$th sample quantile of the D_i. Use (5.30) as a correction factor for finite $B \geq 50g$.

Definition 5.11. a) The prediction region method large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{5.31}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. (This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{g,1-\delta}$ cutoff by $D^2_{(U_B)}$.) b) The modified Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\} \tag{5.32}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$. c) Shift region (5.31) to have center T_n, or equivalently, change the cutoff of region (5.32) to $D^2_{(U_B)}$ to get the hybrid large sample $100(1 - \delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{5.33}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Hyperellipsoids (5.31) and (5.33) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (5.31) and (5.32) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{5.34}$$

The volume of confidence region (5.32) tends to be greater than that of (5.31) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

If $g = 1$, then a hyperellipsoid is an interval, and confidence intervals are special cases of confidence regions. Suppose the parameter of interest is θ, and

there is a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^*. Let $a_i = |T_i^* - \overline{T}^*|$. Let $\overline{T}^*$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D_{(U_B)}^2$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*. Hence the prediction region method CI is a special case of the percentile method CI if $g = 1$. See Definition 5.4. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI $[T_n - a_{(U_B,T)}, T_n + a_{(U_B,T)}]$ corresponding to (5.32) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (5.33). Note that the three CIs corresponding to (5.31)–(5.33) can be computed without finding S_T^* or $D_{(U_B)}$ even if $S_T^* = 0$. The shorth(c) CI (5.26) computed from the T_i^* can be much shorter than the Efron (2014) or prediction region method confidence intervals. See Remark 5.8 for some theory for bootstrap CIs.

In the following definition, let U_B and U_B, T be as in Definition 5.10. Let a_i be as in the above paragraph.

Definition 5.12. a) The large sample $100(1-\delta)\%$ prediction region method CI is $[\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$.

b) The large sample $100(1-\delta)\%$ Bickel and Ren CI is $[T_n - a_{(U_B,T)}, T_n + a_{(U_B,T)}]$.

c) The large sample $100(1-\delta)\%$ hybrid CI is $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$.

Remark 5.5. From Chapter 6, $\text{Cov}(\hat{\boldsymbol{\beta}}^*) = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n} \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ where $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ starts to give good estimates of $\text{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_T$ for many error distributions if $n \geq 10p$ and $T = \hat{\boldsymbol{\beta}}$. For the residual bootstrap with large B, note that $\boldsymbol{S}_T^* \approx 0.95 \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 20p$ and $\boldsymbol{S}_T^* \approx 0.99 \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 100p$. Hence we may need $n >> p$ before the $\boldsymbol{S}_T^*$ is a good estimator of $\text{Cov}(T) = \boldsymbol{\Sigma}_T$. The distribution of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$, but n may need to be large before the approximation is good.

Suppose the bootstrap sample mean $\overline{T}^*$ estimates $\boldsymbol{\theta}$, and the bootstrap sample covariance matrix $\boldsymbol{S}_T^*$ estimates $c_n \widehat{\text{Cov}}(T_n) \approx c_n \boldsymbol{\Sigma}_T$ where c_n increases to 1 as $n \to \infty$. Then $\boldsymbol{S}_T^*$ is not a good estimator of $\widehat{\text{Cov}}(T_n)$ until $c_n \approx 1$ ($n \geq 100p$ for OLS $\hat{\boldsymbol{\beta}}$), but the squared Mahalanobis distance $D_{\boldsymbol{w}}^{2*}(\overline{T}^*, \boldsymbol{S}_T^*) \approx D_{\boldsymbol{w}}^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)/c_n$ and $D_{(U_B)}^{2*} \approx D_{1-\delta}^2/c_n$. Hence the prediction region method has a cutoff $D_{(U_B)}^{2*}$ that estimates the cutoff $D_{1-\delta}^2/c_n$. Thus the prediction region method may give good results for much smaller n than

a bootstrap method that uses a $\chi^2_{g,1-\delta}$ cutoff when a cutoff $\chi^2_{g,1-\delta}/c_n$ should be used for moderate n.

Remark 5.6. For bootstrapping the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, we will often want $n \geq 20p$ and $B \geq \max(100, n, 50p)$. If T_n is $g \times 1$, we might replace p by g or replace p by d if d is the model degrees of freedom. Sometimes much larger n is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}^*_T$ is a good estimator of $Cov(T^*_n)$. Prediction region theory uses correction factors like (4.10) and (4.4) to compensate for finite n. The bootstrap confidence regions (5.31)–(5.33) and the shorth CI use the correction factors (5.30) and (5.26) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as B decreases. Hence a test with larger B will have more power.

5.4.3 Theory for Bootstrap Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator.

Remark 5.7. Some regularity conditions used for bootstrap confidence regions are i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T^*_i - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T^*_i - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}^*_T \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$. Regularity condition v) is rather strong by Machado and Parente (2005). Regularity conditions i) and ii) are often shown using large sample theory. Since (5.32) is a large sample confidence region by Bickel and Ren (2001), (5.31) and (5.33) are too, provided $vi) \sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$. Also note that (5.32) is a large sample confidence region if the standard confidence region (5.28) is a large sample confidence region.

Olive (2017b: ∮ 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under v) from Remark 5.7 and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma_u})$, but the following Pelawa Watagoda and Olive (2021a) theorem and proof is simpler. Since iii) and iv) hold by Theorem 5.2, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma_u}$ is nonsingular.

Theorem 5.2. a) Suppose i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Then iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and vi) $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$.

b) Then the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$.

Proof. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Fix B. Then

$$\begin{bmatrix} \sqrt{n}(T_1^* - T_n) \\ \vdots \\ \sqrt{n}(T_B^* - T_n) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 3.7 and 3.8, and see Example 3.2.) For fixed B, the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B} \sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

by Theorem 3.12 where $\boldsymbol{z} \sim AN_g(\mathbf{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$, and iii), iv), and vi) hold. Hence b) follows. □

Remark 5.8. Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions (5.31)–(5.33), the shorth confidence interval (5.26), and the "usual" percentile method confidence interval (5.25) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $g \times g$ matrices, and T_n is an estimator of $\boldsymbol{\theta}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{5.35}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2} \boldsymbol{u} = \boldsymbol{z},$$

$$n\ (T_n - \boldsymbol{\theta})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{z}^T \boldsymbol{z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\theta})^T \ [\boldsymbol{S}_T^*]^{-1} \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2 \tag{5.36}$$

as $n, B \to \infty$. Assume the cumulative distribution function of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then we could use the large sample confidence region (5.28) $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\}$. Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_g$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$. The following remark gives a simple technical explanation for why bootstrap confidence regions and tests work.

Remark 5.9. a) Under reasonable conditions, i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D^2_{T_i^*}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - T_n),$$

are well behaved. If $(n\boldsymbol{S}_T^*)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_T^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T \boldsymbol{\Sigma}_T^{-1} \boldsymbol{u}$. If $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned" then $D_j^2 \approx \boldsymbol{u}^T (n\boldsymbol{S}_T^*)^{-1} \boldsymbol{u}$ for large n, and the confidence regions (5.31), (5.32), and (5.33) will have coverage near $1 - \delta$. The regularity conditions for (5.31)–(5.33) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

b) Both I) $\sqrt{n}(T_{1n}^* - T_n), ..., \sqrt{n}(T_{Bn}^* - T_n)$ and II) $\sqrt{n}(T_{1n}^* - \overline{T}^*), ..., \sqrt{n}(T_{Bn}^* - \overline{T}^*)$ can be used as pseudodata for III) $\sqrt{n}(T_{1n} - \boldsymbol{\theta}), ..., \sqrt{n}(T_{Bn} - \boldsymbol{\theta})$ when n is large since i), ii) and iv) hold. We can't get the random quantities in III) since $\boldsymbol{\theta}$ is unknown, and we only have $B = 1$ value of the statistic T_n. Note that i) would give an asymptotic pivot if the distribution of $\boldsymbol{u}$ was known.

The following Pelawa Watagoda and Olive (2021a) theorem is very useful. Let $D^2_{(U_B)}$ be the cutoff for the nonparametric prediction region (4.11) computed from the $D_i^2(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. Hence n is replaced by B. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved ("not too ill conditioned") for each $n \geq 20g$, say. This condition is weaker than $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Note that $T_i = T_{in}$.

Theorem 5.3: Geometric Argument. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region

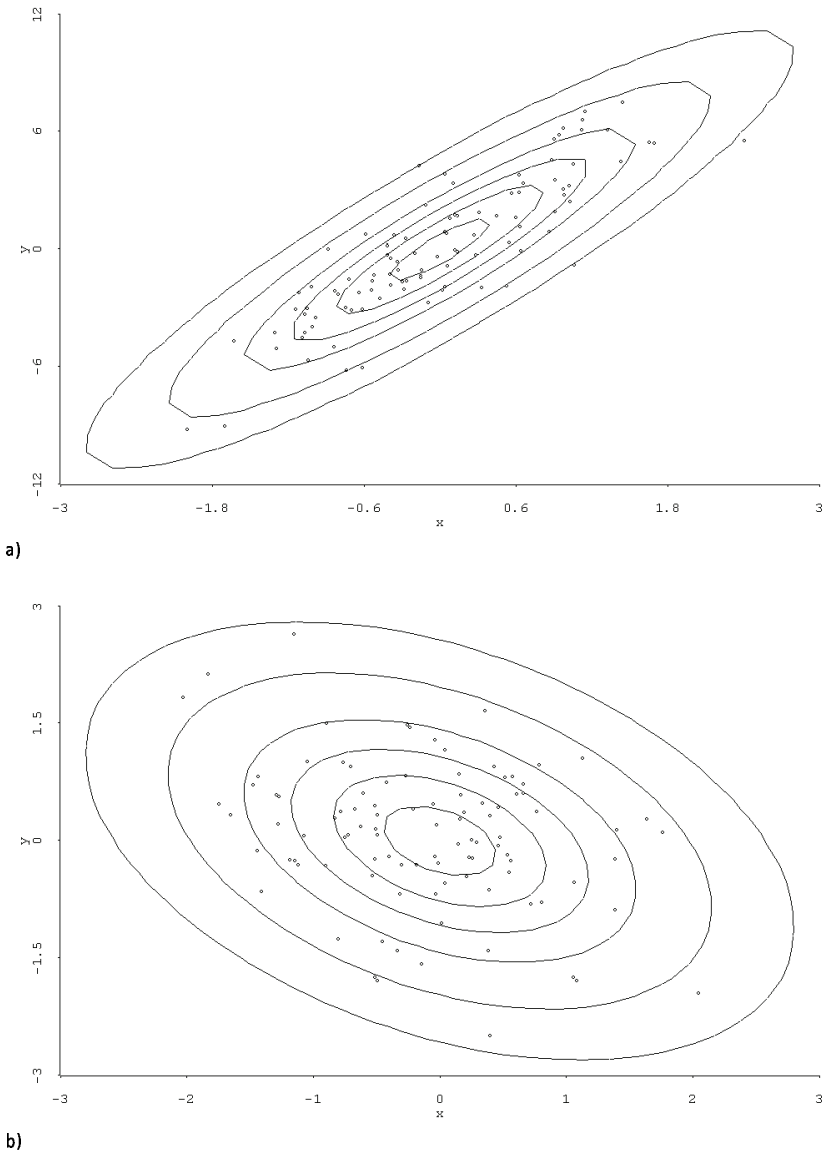

Fig. 5.1 Confidence Regions for 2 Statistics with MVN Distributions

$R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ where T_n is a randomly selected T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 3.7 and 3.8, and see Example 3.2.) For fixed B, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma_u}}{B}\right)$$

by Theorem 3.12. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. □

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u}$, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. The nonparametric confidence region (5.31) applies the prediction region to the bootstrap. Then the hybrid region (5.33) centers that region at T_n. Hence (5.33) is a confidence region by the geometric argument, and (5.31) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$. Since the T_i^* are closer to $\overline{T}^*$ than T_n on average, $D^2_{(U_B,T)}$ tends to be greater than $D^2_{(U_B)}$. Hence the coverage and volume of (5.32) tend to be at least as large as the coverage and volume of (5.32).

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1}(\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1}(T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$.

The prediction region method will often simulate well even if B is rather small. If the ellipses are centered at T_n or $\overline{T}^*$, Figure 4.3 shows confidence regions if the plotted points are $T_1^*, ..., T_B^*$ where the T_i^* are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 5.1 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$.

If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 5.3, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 5.1. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx \text{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jg$ where $J = 20$ or 50. For small g, using $B = 1000$ often led to good simulations, but $B = \max(50g, 100)$ may work well.

Remark 5.10. Remark 5.5 suggests that even if the statistic T_n is asymptotically normal so the Mahalanobis distances are asymptotically χ^2_g, the prediction region method can give better results for moderate n by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{g,1-\delta}$. Theorem 5.3 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \geq Jg$ where $J = 20$ or 50. See Remark 5.9. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jg$. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

5.5 Summary

5.6 Complements

Confidence Intervals

Guenther (1969) is a useful reference for confidence intervals. Agresti and Coull (1998), Brown, Cai and DasGupta (2001, 2002) and Pires and Amado (2008) discuss CIs for a binomial proportion. Agresti and Caffo (2000) discuss CIs for the difference of two binomial proportions $\rho_1 - \rho_2$ obtained from 2

independent samples. Barker (2002), Byrne and Kabaila (2005), Garwood (1936) and Swift (2009) discuss CIs for Poisson (θ) data. Abuhassan and Olive (2008) and Olive (2014) consider CIs for some transformed random variables. Also see Brownstein and Pensky (2008).

The Bootstrap

Rajapaksha and Olive (2022) has two more bootstrap confidence regions which have simple large sample theory and which are quick to compute.

Good references for the bootstrap include Efron (1979, 1982), Efron and Hastie (2016, ch. 10–11), and Efron and Tibshirani (1993). Also see Chen (2016) and Hesterberg (2014). One of the sufficient conditions for the bootstrap confidence region is that T has a well behaved Hadamard derivative. Fréchet differentiability implies Hadamard differentiability, and many statistics are shown to be Hadamard differentiable in Bickel and Ren (2001), Clarke (1986, 2000), Fernholtz (1983), Gill (1989), Ren (1991), and Ren and Sen (1995). Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

The double bootstrap technique may be useful. See Hall (1986) and Chang and Hall (2015) for references. The double bootstrap for $\overline{T}^* = \overline{T}^*_B$ says that $T_n = \overline{T}^*$ is a statistic that can be bootstrapped. Let $B_d \geq 50 g_{max}$ where $1 \leq g_{max} \leq p$ is the largest dimension of $\boldsymbol{\theta}$ to be tested with the double bootstrap. Draw a bootstrap sample of size B and compute $\overline{T}^* = T^*_1$. Repeat for a total of B_d times. Apply the confidence region (5.31), (5.32), or (5.33) to the double bootstrap sample $T^*_1, ..., T^*_{B_d}$. If $D_{(U_{B_d})} \approx D_{(U_{B_d},T)} \approx \sqrt{\chi^2_{g,1-\delta}}$, then $\overline{T}^*$ may be approximately multivariate normal. The CI (5.31) applied to the double bootstrap sample could be regarded as a modified Frey CI without delta method techniques. Of course the double bootstrap tends to be too computationally expensive to simulate.

Warning: Much of the bootstrap theory in the literature is for when all possible bootstrap samples are taken (the population bootstrap quantities). This theory does not apply when B is fixed, e.g. $B = 1000$, and may not apply if $B = \max(1000, n) \to \infty$ as $n \to \infty$.

5.7 Problems

5.1^Q**.** Suppose that $X_1, ..., X_n$ are iid with the Weibull distribution, that is the common pdf is

$$f(x) = \begin{cases} \frac{b}{a} x^{b-1} e^{-\frac{x^b}{a}} & 0 < x \\ 0 & \text{elsewhere} \end{cases}$$

where a is the unknown parameter, but $b(> 0)$ is assumed known.

a) Find a minimal sufficient statistic for a.

b) Assume $n = 10$. Use the Chi-Square Table and the minimal sufficient statistic to find a 95% two sided confidence interval for a.

R Problems

Use a command like *source("G:/sipack.txt")* **to download the functions**. See the Preface. Typing the name of the `sipack` function, e.g. *accisimf*, will display the code for the function. Use the `args` command, e.g. *args(accisimf)*, to display the needed arguments for the function.

5.2. Let $X_1, ..., X_n$ be iid Poisson(θ) random variables.

From the website (http://parker.ad.siu.edu/Olive/lspack.txt), enter the R function `poiscisim` into R. This function simulates the 3 CIs (classical, modified and exact) from Example 5.5. To run the function for $n = 100$ and $\theta = 5$, enter the R command `poiscisim(theta=5)`. Make a table with header "theta ccov clen mcov mlen ecov elen." Fill the table for *theta* = 0.001, 0.1, 1.0, and 5.

The "cov" is the proportion of 500 runs where the CI contained θ and the nominal coverage is 0.95. A coverage between 0.92 and 0.98 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.98 suggests that the CI is conservative while a coverage less than 0.92 suggests that the CI is liberal (too short). Typically want the true coverage $\geq$ the nominal coverage, so conservative intervals are better than liberal CIs. The "len" is the average scaled length of the CI and for large $n\theta$ should be near $2(1.96)\sqrt{\theta}$ for the classical and modified CIs.

From your table, is the classical CI or the modified CI or the "exact" CI better? Explain briefly. (Warning: in a 1999 version of R, there was a bug for the Poisson random number generator for $\theta \geq 10$. The 2011 version of R seems to work.)

5.3. Let $Y_1, ..., Y_n$ be iid binomial($1, \rho$) random variables.

From the website (http://parker.ad.edu/Olive/lspack.txt), enter the R function `bcisim` into R. This function simulates the 3 CIs (classical, Agresti Coull and exact) from Example 5.6, but changes the CI (L,U) to (max(0,L),min(1,U)) to get shorter lengths.

To run the function for $n = 10$ and $\rho \equiv p = 0.001$, enter the R command `bcisim(n=10,p=0.001)`. Make a table with header "n p ccov clen accov aclen ecov elen." Fill the table for $n = 10$ and $p = 0.001, 0.01, 0.5, 0.99, 0.999$ and then repeat for $n = 100$. The "cov" is the proportion of 500 runs where the CI contained p and the nominal coverage is 0.95. A coverage between 0.92 and 0.98 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.98 suggests that the CI is conservative while a coverage less than 0.92 suggests that the CI is liberal. Typically want the true coverage $\geq$ the nominal coverage, so conservative intervals are better than liberal CIs. The "len" is the average scaled length of the CI and for large n should be near $2(1.96)\sqrt{p(1-p)}$.

From your table, is the classical estimator or the Agresti Coull CI better? When is the "exact" interval good? Explain briefly.

5.4. This problem simulates the CIs from Example 5.7.

a) Download the function `accisimf` into *R*.

b) The function will be used to compare the classical, ACT and modified 95% CIs when the population size $N = 500$ and p is close to 0.01. The function generates such a population, then selects 5000 independent simple random samples from the population. The 5000 CIs are made for both types of intervals, and the number of times the true population p is in the ith CI is counted. The simulated coverage is this count divided by 5000 (the number of CIs). The nominal coverage is 0.95. To run the function for $n = 50$ and $p \approx 0.01$, enter the command `accisimf(n=50,p=0.01)`. Make a table with header "n p ccov clen accov aclen mcov mlen." Fill the table for $n = 50$ and then repeat for $n = 100, 150, 200, 250, 300, 350, 400$ and 450. The "len" is $\sqrt{n}$ times the mean length from the 5000 runs. The "cov" is the proportion of 5000 runs where the CI contained p and the nominal coverage is 0.95. For 5000 runs, an observed coverage between 0.94 and 0.96 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.96 suggests that the CI is conservative while a coverage less than 0.94 suggests that the CI is liberal. Typically want the true coverage $\geq$ the nominal coverage, so conservative intervals are better than liberal CIs. The "ccov" is for the classical CI, "accov" is for the Agresti Coull type (ACT) CI and "mcov" is for the modified interval. Given good coverage > 0.94, want short length.

c) First compare the classical and ACT intervals. From your table, for what values of n is the ACT CI better, for what values of n are the 3 intervals about the same, and for what values of n is the classical CI better?

d) Was the modified CI ever good?

5.6. This problem simulates the CIs from Example 5.1.

a) Download the function `hnsim` into *R*.

The output from this function are the coverages scov, lcov and ccov of the CI for σ^2, μ and of σ^2 if μ is known. The scaled average lengths of the CIs are also given. The lengths of the CIs for σ^2 are multiplied by $\sqrt{n}$ while the length of the CI for μ is multiplied by n.

b) The 5000 CIs are made for 3 intervals, and the number of times the true population parameter $\theta = \mu$ or σ^2 is in the ith CI is counted. The simulated coverage is this count divided by 5000 (the number of CIs). The nominal coverage is 0.95. To run the function for $n = 5$, $\mu = 0$ and $\sigma^2 = 1$ enter the command `hnsim(n=5)`. Make a table with header
"CI for σ^2 CI for μ CI for σ^2, μ known."
Then make a second header "n cov slen cov slen cov slen" where "cov slen" is below each of the three CI headers. Fill the table for $n = 5$ and then repeat

for $n = 10, 20, 50, 100$ and 1000. The "cov" is the proportion of 5000 runs where the CI contained θ and the nominal coverage is 0.95. For 5000 runs, an observed coverage between 0.94 and 0.96 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.96 suggests that the CI is conservative while a coverage less than 0.94 suggests that the CI is liberal. As n gets large, the values of slen should get closer to 5.5437, 3.7546 and 5.5437.

5.7. a) Download the function `varcisim` into R to simulate a modified version of the CI of Example 5.8.

b) Type the command `varcisim(n = 100, nruns = 1000, type = 1)` to simulate the 95% CI for the variance for iid N(0,1) data. Is the coverage *vcov* close to or higher than 0.95? Is the scaled length *vlen* = $\sqrt{n}$ (CI length) = $2(1.96)\sigma^2\sqrt{\tau} = 5.554\sigma^2$ close to 5.554?

c) Type the command `varcisim(n = 100, nruns = 1000, type = 2)` to simulate the 95% CI for the variance for iid EXP(1) data. Is the coverage *vcov* close to or higher than 0.95? Is the scaled length *vlen* = $\sqrt{n}$ (CI length) = $2(1.96)\sigma^2\sqrt{\tau} = 2(1.96)\lambda^2\sqrt{8} = 11.087\lambda^2$ close to 11.087?

d) Type the command `varcisim(n = 100, nruns = 1000, type = 3)` to simulate the 95% CI for the variance for iid LN(0,1) data. Is the coverage *vcov* close to or higher than 0.95? Is the scaled length *vlen* long?

5.8. a) Download the function `pcisim` into R to simulate the three CIs of Example 5.9. The modified pooled t CI is almost the same as the Welch CI, but uses degrees of freedom $= n_1 + n_2 - 4$ instead of the more complicated formula for the Welch CI. The pooled t CI should have coverage that is too low if

$$\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2 \ < \ \sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2.$$

b) Type the command `pcisim(n1=100,n2=200,var1=10,var2=1)` to simulate the CIs for $N(\mu_i, \sigma_i^2)$ data for $i = 1, 2$. The terms *pcov*, *mpcov* and *wcov* are the simulated coverages for the pooled, modified pooled and Welch 95% CIs. Record these quantities. Are they near 0.95?

Problems from old qualifying exams are marked with a Q.

5.9^Q. Let $X_1, ..., X_n$ be a random sample from a uniform$(0, \theta)$ distribution. Let $Y = \max(X_1, X_2, ..., X_n)$.

a) Find the pdf of Y/θ.

b) To find a confidence interval for θ, can Y/θ be used as a pivot?

c) Find the shortest $(1-\alpha)$% confidence interval for θ.

5.10. Let $Y_1, ..., Y_n$ be iid from a distribution with fourth moments and let S_n^2 be the sample variance. Then

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{D} N(0, M_4 - \sigma^4)$$

where M_4 is the fourth central moment $E[(Y-\mu)^4]$. Let $\hat{M}_{4,n} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^4$.

a) Use the asymptotic pivot

$$\frac{\sqrt{n}(S_n^2 - \sigma^2)}{\sqrt{\hat{M}_{4,n} - S_n^4}} \xrightarrow{D} N(0,1)$$

to find a large sample $100(1-\alpha)\%$ CI for σ^2.

b) Use equation (5.4) to find a large sample $100(1-\alpha)\%$ CI for $\sigma_1^2 - \sigma_2^2$.