Topic Five: Inferences About a Mean Vector

Preview

- Motivation:
  - Statistical inference is the process of drawing confident conclusions about a population based on a sample. Example applications are confidence intervals and hypothesis tests.
  - Because *p* correlated variables should be analyzed jointly, our first step is to develop multivariate analogs to the usual one-sample confidence intervals and tests for a univariate mean.
- Goals:
  - Learn how to make *simultaneous* confidence statements about the components of a mean vector.

Hotelling's T^2 Statistic

Review: If we want to decide whether a univariate mean μ equals a particular value μ_0, we can create the two competing hypotheses $H_0 : \mu = \mu_0$ (the null hypothesis) and $H_1 : \mu \neq \mu_0$ (the (two-sided) alternative hypothesis). With a random sample $X_1, X_2, \ldots, X_n$ from a normal population, the optimal test statistic is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

When H_0 is true, $t \sim t_{n-1}$, so we can reject H_0 if t is not a likely realization from this distribution. Specifically, we reject H_0 at significance level α if $|t|$ exceeds $t_{n-1}(\alpha/2)$, where $t_{n-1}(\alpha/2)$ is the $100(1-\alpha/2)$th percentile of the t_{n-1}. Rejecting H_0 if $|t|$ is large is equivalent to rejecting if its square

$$t^2 = \frac{\left(\bar{X} - \mu_0\right)}{s^2/n}$$

is large. This is the square of the statistical distance between $\bar{X}$ and μ_0. For an observed sample, we reject H_0 at significance level α if

$$n\left(\bar{x} - \mu_0\right)\left(s^2\right)^{-1}\left(\bar{x} - \mu_0\right) > t^2_{n-1}(\alpha/2)$$

Review Continued: If we fail to reject H_0, we conclude that μ_0 is a plausible value of μ. Confidence intervals tell us a *range* of plausible values of μ at a specified level of confidence. And there is an equivalence between hypothesis tests and confidence intervals. Specifically, when testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$,

$$\{\text{Do not reject } H_0 \text{ at level } \alpha\} \quad \text{or} \quad \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| \leq t_{n-1}(\alpha/2)$$

is equivalent to

$$\left\{\mu_0 \text{ lies in the } 100(1-\alpha)\% \text{ confidence interval } \bar{x} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}\right\}$$

or

$$\bar{x} - t_{n-1}(\alpha/2)\frac{2}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}$$

Note that, before the sample is selected, confidence interval endpoints are random. If we repeatedly drew samples from the population, computing $100(1-\alpha)\%$ confidence intervals each time, we would expect $100(1-\alpha)\%$ of them to contain μ.

Hotelling's $\mathbf{T^2}$: We can generalize the squared distance univariate statistic to the multivariate case as

$$T^2 = \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \left(\frac{1}{n}\mathbf{S}\right)^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right) = n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)$$

If H_0 is true, then

$$T^2 \sim \frac{(n-1)p}{(n-p)} F_{p,n-p}$$

where $F_{p,n-p}$ is the F distribution with p and $n-p$ degrees of freedom. We can therefore test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ at significance level α by rejecting H_0 when

$$T^2 > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

where $F_{p,n-p}(\alpha)$ is the $(100(1-\alpha))$th percentile of the $F_{p,n-p}$ distribution.

In the univariate case, we can write

$$t^2 = \left(\begin{array}{c} \text{normal} \\ \text{random variable} \end{array} \right) \left(\frac{\begin{array}{c} \text{(scaled) chi-square} \\ \text{random variable} \end{array}}{\text{d.f.}} \right)^{-1} \left(\begin{array}{c} \text{normal} \\ \text{random variable} \end{array} \right)$$

and show, using the fact that $\bar{X}$ and S^2 are independent, that this has the $F_{1,n-1}$ distribution. In a similar way, we can write

$$T^2 = \left(\begin{array}{c} \text{multivariate normal} \\ \text{random variable} \end{array} \right) \left(\frac{\begin{array}{c} \text{(scaled) Wishart} \\ \text{random matrix} \end{array}}{\text{d.f.}} \right)^{-1} \left(\begin{array}{c} \text{multivariate normal} \\ \text{random variable} \end{array} \right)$$

and use this to derive the $F_{p,n-p}$ distribution from the previous slide.

Example: The `container.df` data set in the `R` package `Hotelling` contains concentration measurements of 9 elements in two container types. There are 10 containers of each type.

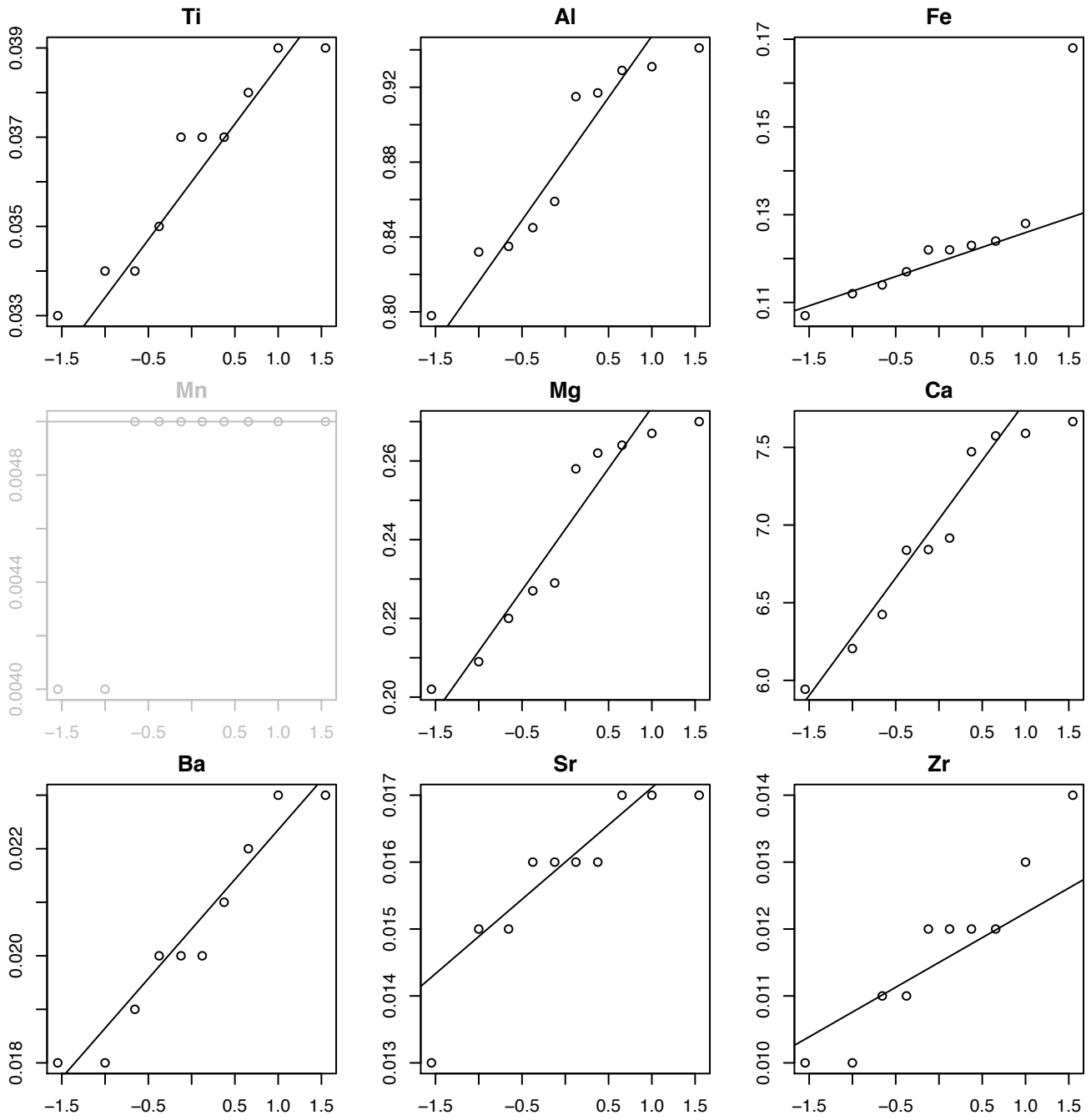
gp	Ti	Al	Fe	Mn	Mg	Ca	Ba	Sr	Zr
1	0.037	0.929	0.124	0.005	0.267	7.666	0.022	0.017	0.014
1	0.034	0.859	0.112	0.005	0.227	6.838	0.018	0.016	0.011
1	0.033	0.845	0.107	0.004	0.220	5.943	0.018	0.016	0.011
1	0.035	0.931	0.117	0.005	0.262	6.424	0.020	0.017	0.013
1	0.034	0.915	0.114	0.004	0.258	6.205	0.020	0.017	0.012
1	0.037	0.832	0.122	0.005	0.209	6.842	0.021	0.015	0.012
1	0.038	0.835	0.123	0.005	0.229	6.916	0.020	0.015	0.010
1	0.039	0.941	0.128	0.005	0.270	7.590	0.023	0.016	0.012
1	0.039	0.917	0.168	0.005	0.264	7.574	0.023	0.016	0.012
1	0.037	0.798	0.122	0.005	0.202	7.472	0.019	0.013	0.010
2	0.032	0.751	0.032	0.002	0.117	6.378	0.011	0.018	0.012
2	0.032	0.659	0.029	0.002	0.090	5.511	0.009	0.016	0.009
2	0.032	0.746	0.031	0.002	0.113	6.864	0.012	0.019	0.012
2	0.033	0.772	0.032	0.002	0.117	7.061	0.012	0.019	0.012
2	0.031	0.722	0.030	0.002	0.109	6.417	0.011	0.018	0.012
2	0.029	0.752	0.028	0.002	0.115	7.684	0.013	0.015	0.009
2	0.029	0.739	0.028	0.002	0.113	7.543	0.013	0.015	0.010
2	0.032	0.695	0.031	0.002	0.115	7.780	0.013	0.017	0.011
2	0.029	0.741	0.028	0.002	0.113	7.548	0.013	0.015	0.011
2	0.028	0.715	0.027	0.002	0.108	7.755	0.012	0.014	0.010

Example: The `container.df` data set in the `R` package `Hotelling` contains concentration measurements of 9 elements in two container types. There are 10 containers of each type.

gp	Ti	Al	Fe	Mn	Mg	Ca	Ba	Sr	Zr
1	0.037	0.929	0.124	0.005	0.267	7.666	0.022	0.017	0.014
1	0.034	0.859	0.112	0.005	0.227	6.838	0.018	0.016	0.011
1	0.033	0.845	0.107	0.004	0.220	5.943	0.018	0.016	0.011
1	0.035	0.931	0.117	0.005	0.262	6.424	0.020	0.017	0.013
1	0.034	0.915	0.114	0.004	0.258	6.205	0.020	0.017	0.012
1	0.037	0.832	0.122	0.005	0.209	6.842	0.021	0.015	0.012
1	0.038	0.835	0.123	0.005	0.229	6.916	0.020	0.015	0.010
1	0.039	0.941	0.128	0.005	0.270	7.590	0.023	0.016	0.012
1	0.039	0.917	0.168	0.005	0.264	7.574	0.023	0.016	0.012
1	0.037	0.798	0.122	0.005	0.202	7.472	0.019	0.013	0.010
2	0.032	0.751	0.032	0.002	0.117	6.378	0.011	0.018	0.012
2	0.032	0.659	0.029	0.002	0.090	5.511	0.009	0.016	0.009
2	0.032	0.746	0.031	0.002	0.113	6.864	0.012	0.019	0.012
2	0.033	0.772	0.032	0.002	0.117	7.061	0.012	0.019	0.012
2	0.031	0.722	0.030	0.002	0.109	6.417	0.011	0.018	0.012
2	0.029	0.752	0.028	0.002	0.115	7.684	0.013	0.015	0.009
2	0.029	0.739	0.028	0.002	0.113	7.543	0.013	0.015	0.010
2	0.032	0.695	0.031	0.002	0.115	7.780	0.013	0.017	0.011
2	0.029	0.741	0.028	0.002	0.113	7.548	0.013	0.015	0.011
2	0.028	0.715	0.027	0.002	0.108	7.755	0.012	0.014	0.010

Example: The `container.df` data set in the `R` package `Hotelling` contains concentration measurements of 9 elements in two container types. There are 10 containers of each type.

gp	Ti	Al	Fe	Mn	Mg	Ca	Ba	Sr	Zr
1	0.037	0.929	0.124	0.005	0.267	7.666	0.022	0.017	0.014
1	0.034	0.859	0.112	0.005	0.227	6.838	0.018	0.016	0.011
1	0.033	0.845	0.107	0.004	0.220	5.943	0.018	0.016	0.011
1	0.035	0.931	0.117	0.005	0.262	6.424	0.020	0.017	0.013
1	0.034	0.915	0.114	0.004	0.258	6.205	0.020	0.017	0.012
1	0.037	0.832	0.122	0.005	0.209	6.842	0.021	0.015	0.012
1	0.038	0.835	0.123	0.005	0.229	6.916	0.020	0.015	0.010
1	0.039	0.941	0.128	0.005	0.270	7.590	0.023	0.016	0.012
1	0.039	0.917	0.168	0.005	0.264	7.574	0.023	0.016	0.012
1	0.037	0.798	0.122	0.005	0.202	7.472	0.019	0.013	0.010
2	0.032	0.751	0.032	0.002	0.117	6.378	0.011	0.018	0.012
2	0.032	0.659	0.029	0.002	0.090	5.511	0.009	0.016	0.009
2	0.032	0.746	0.031	0.002	0.113	6.864	0.012	0.019	0.012
2	0.033	0.772	0.032	0.002	0.117	7.061	0.012	0.019	0.012
2	0.031	0.722	0.030	0.002	0.109	6.417	0.011	0.018	0.012
2	0.029	0.752	0.028	0.002	0.115	7.684	0.013	0.015	0.009
2	0.029	0.739	0.028	0.002	0.113	7.543	0.013	0.015	0.010
2	0.032	0.695	0.031	0.002	0.115	7.780	0.013	0.017	0.011
2	0.029	0.741	0.028	0.002	0.113	7.548	0.013	0.015	0.011
2	0.028	0.715	0.027	0.002	0.108	7.755	0.012	0.014	0.010



QQ plots for the nine variables in container type one.

Some evidence of non-normality of the individual variables. A check of multivariate normality also does not look very convincing.

Example Continued: Let us consider only container type one and only the elements titanium, aluminum, and iron. So, $n = 10$ and $p = 3$. We have

$$\bar{\mathbf{x}} = \begin{bmatrix} 0.0363 \\ 0.8802 \\ 0.1237 \end{bmatrix} \quad \text{and} \quad (n-1)\mathbf{S} = \begin{bmatrix} 0.000042 & 0.000148 & 0.000241 \\ 0.000148 & 0.024036 & 0.002318 \\ 0.000241 & 0.002318 & 0.002542 \end{bmatrix}$$

Then, to test $H_0 : \boldsymbol{\mu}_0' = [0.035, 0.900, 0.150]$, we compute

$$T^2 = n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right) = 93.0278$$

The critical value for significance level $\alpha = 0.05$ is

$$\frac{3(10-1)}{(10-3)} F_{3,10-3}(0.05) = 3.8571(4.3468) = 16.7660$$

Because $93.0278 > 16.7660$, we reject H_0. Equivalently, we could compute a p-value as the probability that a draw from the $F_{3,10-3}$ distribution exceeds

$$\frac{(10-3)}{3(10-1)} 93.0278 = 24.1183$$

The p-value is about 0.0005, so we reach the same conclusion as above.

Example Continued: Let us consider only container type one and only the elements titanium, aluminum, and iron. So, $n = 10$ and $p = 3$. We have

$$\bar{\mathbf{x}} = \begin{bmatrix} 0.0363 \\ 0.8802 \\ 0.1237 \end{bmatrix} \quad \text{and} \quad (n-1)\mathbf{S} = \begin{bmatrix} 0.000042 & 0.000148 & 0.000241 \\ 0.000148 & 0.024036 & 0.002318 \\ 0.000241 & 0.002318 & 0.002542 \end{bmatrix}$$

Then, to test $H_0 : \boldsymbol{\mu}_0' = [0.035, 0.900, 0.150]$, we compute

$$T^2 = n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right) = 93.0278$$

The critical value for significance level $\alpha = 0.05$ is

$$\frac{3(10-1)}{(10-3)} F_{3,10-3}(0.05) = 3.8571(4.3468) = 16.7660$$

Because $93.0278 > 16.7660$, we reject H_0. Equivalently, we could compute a p-value as the probability that a draw from the $F_{3,10-3}$ di[illegible]ds

$$\frac{(10-3)}{3(10-1)} 93.0278 = 24.1183$$

The p-value is about 0.0005, so we reach the same conclusion as above.

Changing the Units of Measurement: Suppose we change the units of the variables in $\mathbf{X}$ by the transformation

$$\underset{(p\times 1)}{\mathbf{Y}} = \underset{(p\times p)}{\mathbf{C}}\,\underset{(p\times 1)}{\mathbf{X}} + \underset{(p\times 1)}{\mathbf{d}}, \qquad \mathbf{C} \text{ nonsingular}$$

We might, for example, wish to convert temperatures from Fahrenheit to Celsius or weights from pounds to kilograms. We know that $E(\mathbf{Y}) = \mathbf{C}\boldsymbol{\mu} + \mathbf{d}$. Also, given a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we have

$$\bar{\mathbf{y}} = \mathbf{C}\bar{\mathbf{x}} + \mathbf{d} \quad \text{and} \quad \mathbf{S}_{\mathbf{y}} = \mathbf{C}\mathbf{S}\mathbf{C}'$$

Then, with $\boldsymbol{\mu}_{\mathbf{Y},0} = \mathbf{C}\boldsymbol{\mu}_0 + \mathbf{d}$,

$$\begin{aligned}
T^2 &= n\left(\bar{\mathbf{y}} - \boldsymbol{\mu}_{\mathbf{Y},0}\right)' \mathbf{S}_{\mathbf{Y}}^{-1} \left(\bar{\mathbf{y}} - \boldsymbol{\mu}_{\mathbf{Y},0}\right) \\
&= n\left(\mathbf{C}\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)\right)' \left(\mathbf{C}\mathbf{S}\mathbf{C}'\right)^{-1} \left(\mathbf{C}\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)\right) \\
&= n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)' \mathbf{C}' \left(\mathbf{C}\mathbf{S}\mathbf{C}'\right)^{-1} \mathbf{C}\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right) \\
&= n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)' \mathbf{C}' \left(\mathbf{C}'\right)^{-1} \mathbf{S}^{-1} \mathbf{C}^{-1} \mathbf{C}\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right) \\
&= n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{x}} - \boldsymbol{\mu}_0\right)
\end{aligned}$$

The T^2 statistic is *invariant* to such transformations.

Likelihood Ratio Tests

The Likelihood Ratio Method: Consider the n random vectors $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ drawn from a ν-dimensional joint distribution with unknown parameter vector $\boldsymbol{\theta}$. Let $L(\boldsymbol{\theta})$ be the likelihood function, the joint density evaluated at observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The parameter vector $\boldsymbol{\theta}$ takes values in the parameter set $\boldsymbol{\Theta}$. Under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\boldsymbol{\theta}$ is restricted to a subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$. A likelihood ratio test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if

$$\Lambda = \frac{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})} < c$$

for a suitably chosen threshold c. If H_0 is true, then restricting $\boldsymbol{\theta}$ to $\boldsymbol{\Theta}_0$ will not substantially change the maximized likelihood. Furthermore, since the numerator of Λ can not be greater than the denominator, evidence against H_0 looks like a small value for Λ. Likelihood ratio tests generally have very good (often, optimal) statistical power among all other tests at a fixed significance level α.

The Likelihood Ratio Method: Consider the n random vectors $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ drawn from a ν-dimensional joint distribution with unknown parameter vector $\boldsymbol{\theta}$. Let $L(\boldsymbol{\theta})$ be the likelihood function, the joint density evaluated at observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The parameter vector $\boldsymbol{\theta}$ takes values in the parameter set $\boldsymbol{\Theta}$. Under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\boldsymbol{\theta}$ is restricted to a subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$. A lik

When n is large, the sampling distribution of $-2 \ln \Lambda$ under H_0 is approximately $\chi^2_{\nu-\nu_0}$, where

$$\nu - \nu_0 = (\text{dimension of } \boldsymbol{\Theta}) - (\text{dimension of } \boldsymbol{\Theta}_0)$$

so a level α critical value can be chosen from this distribution. We will see in a moment that T^2 is equivalent to Λ in the multivariate normal case, for which the *exact* null sampling distribution is known to be $F_{p,n-p}$. And it turns out that the scaled critical values from this distribution that we use for T^2 nearly equal the corresponding critical values from the χ^2_p distribution.

The Multivariate Normal Case: In the multivariate normal case, we can define a single parameter vector $\boldsymbol{\theta}$ of dimension $p + p(p+1)/2$ as containing the p components of $\boldsymbol{\mu}$ and the $p(p+1)/2$ unique terms from $\boldsymbol{\Sigma}$. Now recall that the multivariate normal likelihood can be written as

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp\left(-\frac{1}{2} \text{tr} \left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu}) (\mathbf{x}_j - \boldsymbol{\mu})' \right) \right] \right)$$

and this is maximized by taking $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$, with maximized likelihood

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{np/2} |\hat{\boldsymbol{\Sigma}}|^{n/2}} e^{-np/2}$$

Under $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, the likelihood becomes

$$L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp\left(-\frac{1}{2} \text{tr} \left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu}_0) (\mathbf{x}_j - \boldsymbol{\mu}_0)' \right) \right] \right)$$

and, applying the same maximization technique we used to derive the unrestricted MLEs, we have

$$\max_{\boldsymbol{\Sigma}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{np/2} |\hat{\boldsymbol{\Sigma}}_0|^{n/2}} e^{-np/2}$$

where

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n} \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu}_0) (\mathbf{x}_j - \boldsymbol{\mu}_0)'$$

The Multivariate Normal Likelihood Ratio Test: In the multivariate normal case, we have

$$\Lambda = \frac{\max\limits_{\boldsymbol{\Sigma}} L\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}\right)}{\max\limits_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|}\right)^{n/2} = \left(\frac{\left|\sum_{j=1}^{n}\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \hat{\mathbf{x}}\right)'\right|}{\left|\sum_{j=1}^{n}\left(\mathbf{x}_j - \boldsymbol{\mu}_0\right)\left(\mathbf{x}_j - \boldsymbol{\mu}_0\right)'\right|}\right)^{n/2}$$

And it turns out* that

$$T^2 = (n-1)\left(\frac{1}{\Lambda^{2/n}} - 1\right) = (n-1)\left(\frac{\left|\sum_{j=1}^{n}\left(\mathbf{x}_j - \boldsymbol{\mu}_0\right)\left(\mathbf{x}_j - \boldsymbol{\mu}_0\right)'\right|}{\left|\sum_{j=1}^{n}\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)'\right|} - 1\right)$$

Note then that we can compute Λ without having to compute $\mathbf{S}^{-1}$. For reference, the related statistic *Wilk's lambda* is defined as $\Lambda^{2/n}$.

* See textbook.

Bootstrap: The null sampling distribution of the likelihood ratio statistic is known in the multivariate normal case and can be approximated generally with large n. With small n and / or if multivariate normality does not hold, these parametric inferences may not be reliable. The *bootstrap* is a resampling technique that can be used to approximate the sampling distribution of any statistic. In the context of testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ with T^2, the bootstrap is implemented as follows:

1. Compute T^2 on the observed data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Also, create transformed variables where H_0 has been forced to be true: $\mathbf{x}_{j0} = \mathbf{x}_j - \bar{\mathbf{x}} + \boldsymbol{\mu}_0$, $j = 1, 2, \ldots, n$.

2. For B iterations:

   (a) Let $j_{b1}, j_{b2}, \ldots, j_{bn}$ be a sample with replacement from the indices $1, 2, \ldots, n$. Create a "bootstrapped" sample: $\tilde{\mathbf{x}}_{bj} = \mathbf{x}_{(j_{bj})0}$, $j = 1, 2, \ldots, n$. The $\tilde{\mathbf{x}}_{bj}$ have then been sampled with replacement from the $\mathbf{x}_{j0}$ and represent a pseudosample from the null distribution.

   (b) Compute T_b^2 on the bootstrapped data $\tilde{\mathbf{x}}_{b1}, \tilde{\mathbf{x}}_{b2}, \ldots, \tilde{\mathbf{x}}_{bn}$. This represents a single draw from the null sampling distribution of T^2.

3. Compute a p-value as the proportion of the T_b^2 that equal or exceed T^2.

Bootstrap: The null sampling distribution of the likelihood ratio statistic is known in the multivariate normal case and can be approximated generally with large n. With small n and / or if multivariate normality does not hold, these parametric inferences may not be reliable. The *bootstrap* is a resampling technique that can be used to approximate the sampling distribution of any statistic. In the context of testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ with T^2, the bootstrap is implemented as follows:

1. Compute T^2 on the observed data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Also, create transformed variables where H_0 has been forced to be true: $\mathbf{x}_{j0} = \mathbf{x}_j - \bar{\mathbf{x}} + \boldsymbol{\mu}_0$, $j = 1, 2, \ldots, n$.

2. For B iterations:

   (a) Let $j_{b1}, j_{b2}, \ldots, j_{bn}$ be a sample with replacement from the indices $1, 2, \ldots, n$. Create a "bootstrapped" sample: $\tilde{\mathbf{x}}_{bj} = \mathbf{x}_{(j_{bj})0}$, $j = 1, 2, \ldots, n$. The $\tilde{\mathbf{x}}_{bj}$ have then been sampled with replacement from the $\mathbf{x}_{j0}$ and represent a pseudosample from the nu

   (b) Compute T_b^2 on the bootstrapped data $\tilde{\mathbf{x}}_{b1}, \tilde{\mathbf{x}}_{b2}, \ldots$ sents a single draw from the null sampling distribu

3. Compute a p-value as the proportion of the T_b^2 that equal or exceed T^2.

Confidence Regions and Intervals

Confidence Regions: A $100(1-\alpha)\%$ *confidence region* for a parameter vector $\boldsymbol{\theta}$ is a region within which $\boldsymbol{\theta}$ will lie with probability α. In the univariate case, for example, a $100(1-\alpha)\%$ confidence *interval* for the population mean μ is

$$\bar{X} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

where $t_{\alpha/2,n-1}$ is the $100(1-\alpha/2)$ percentile of the t distribution with $n-1$ degrees of freedom. Before the sample is obtained, we can say that this interval will contain μ with probability $1-\alpha$. Another way to say this is that we expect $100(1-\alpha)\%$ of draws from the sampling distribution of $\bar{X}$ to be within $t_{\alpha/2,n-1}(s/\sqrt{n})$ of μ. Note that once the sample is obtained and we actually compute the interval, it will either contain μ or it will not, and we will have no way of knowing which is the case.

We will now generalize univariate confidence intervals for a scalar mean to multivariate confidence regions for a mean vector.

A Confidence Region for $\boldsymbol{\mu}$: As we saw in our discussion of the T^2 statistic,

$$P\left[n\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)\leq\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\right]=1-\alpha$$

whatever the values of the unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Another way to say this is that we expect $100(1-\alpha)\%$ of draws from the sampling distribution of $\bar{\mathbf{X}}$ to be within

$$\sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}$$

of $\boldsymbol{\mu}$, where distance is defined as $\sqrt{n\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)}$. Given a sample $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ with observed mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{S}$, we therefore define a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ as the ellipsoid determined by all $\boldsymbol{\mu}$ such that

$$n\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)\leq\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)$$

Checking whether a hypothesized value $\boldsymbol{\mu}_0$ lies in the confidence region is equivalent to testing $H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0$ vs. $H_1:\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$ at significance level α.

Note that when $p = 1$,

$$n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right) = n\left(\bar{x} - \mu\right)\left(s^2\right)^{-1}\left(\bar{x} - \mu\right)$$
$$= \left(\frac{\bar{x} - \mu}{s/\sqrt{n}}\right)^2$$

So, μ will be in the confidence region if

$$\left|\frac{\bar{x} - \mu}{s/\sqrt{n}}\right| \leq \sqrt{F_{1,n-1}(\alpha)}$$

But $\sqrt{F_{1,n-1}(\alpha)} = t_{n-1}(\alpha/2)$, so we have the equivalent requirement that μ satisfy

$$\bar{x} - t_{n-1}(\alpha/2)(s/\sqrt{n}) \leq \mu \leq \bar{x} + t_{n-1}(\alpha/2)(s/\sqrt{n})$$

This defines the usual confidence interval for a univariate mean.

Also, recall that the points $\mathbf{b}$ that are a constant distance

$$(\mathbf{a}-\mathbf{b})'\,\mathbf{A}\,(\mathbf{a}-\mathbf{b}) = c^2$$

from $\mathbf{a}$ lie on an ellipse centered at $\mathbf{a}$ with axes

$$\pm\frac{c}{\sqrt{\lambda_i}}\mathbf{e}_i$$

where the λ_i and $\mathbf{e}_i$ are the eigenvalues and eigenvectors of $\mathbf{A}$, respectively. Similarly, the $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ is outlined by the ellipse centered at $\bar{\mathbf{x}}$ with axes

$$\pm\sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}\mathbf{e}_i$$

where the λ_i and $\mathbf{e}_i$ are the eigenvalues and eigenvectors of $\mathbf{S}$, respectively. We can therefore draw it in $p \leq 3$ dimensions.

Container Data: Consider the aluminum and iron concentration measurements for container type one. We have

$$\bar{\mathbf{X}} = \begin{bmatrix} 0.8802 \\ 0.1237 \end{bmatrix} \quad \text{and} \quad (n-1)\mathbf{S} = \begin{bmatrix} 0.0240 & 0.0023 \\ 0.0023 & 0.0025 \end{bmatrix}$$

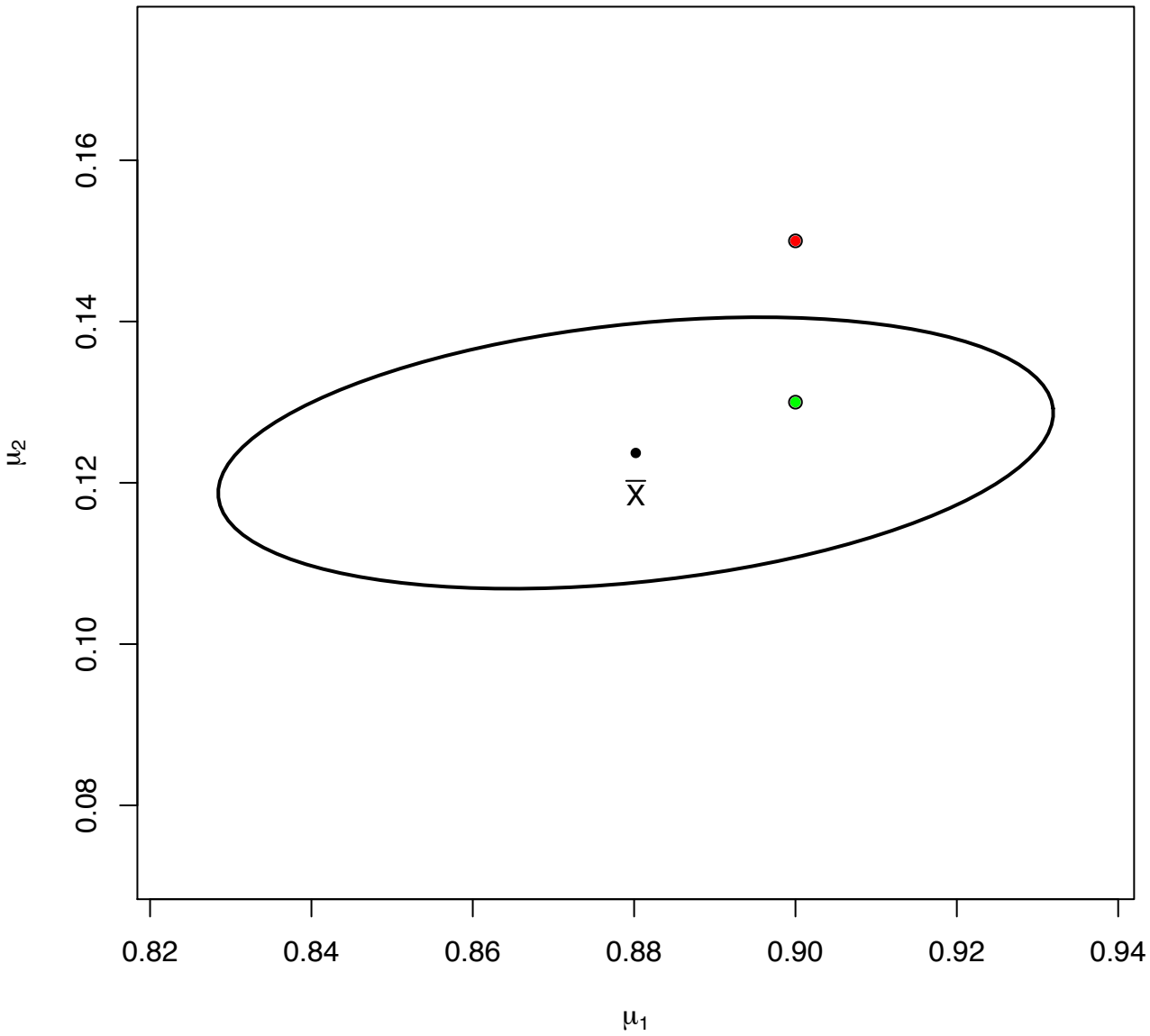
Also, for a 95% confidence region, we have

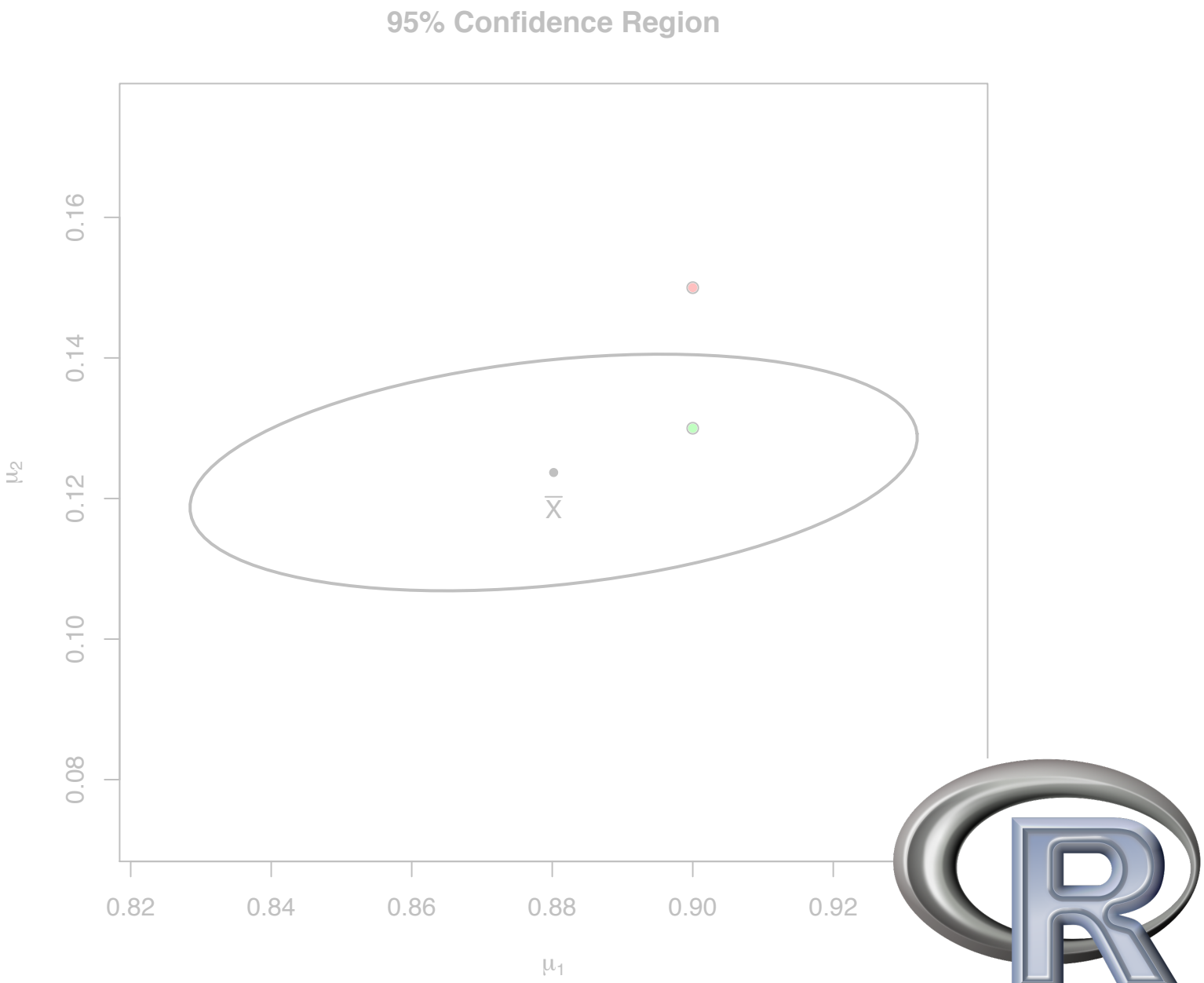
$$\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha) = \frac{18}{8} F_{2,8}(0.05) = 10.0327$$

Any value of $\boldsymbol{\mu}$ within $\sqrt{10.0327} = 3.1674$ of $\bar{\mathbf{X}}$ is inside the region, where distance is defined as

$$\sqrt{n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)}$$

For example, $\boldsymbol{\mu}' = [0.90, 0.13]$ is inside the region, because its distance to $\bar{\mathbf{X}}$ is 1.4887. But $\boldsymbol{\mu}' = [0.90, 0.15]$ is not, because its distance to $\bar{\mathbf{X}}$ is 4.9558. Equivalently, the T^2 test of $H_0 : \boldsymbol{\mu}' = [0.90, 0.13]$ fails to reject at $\alpha = 0.05$, but the test of $H_0 : \boldsymbol{\mu}' = [0.90, 0.15]$ rejects.


95% Confidence Region
0.16
0.14
0.12
0.10
0.08
μ2
X
0.82
0.84
0.86
0.88
0.90
0.92
0.94
μ1


95% Confidence Region
0.16
0.14
0.12
0.10
0.08
μ_2
$\bar{X}$
0.82
0.84
0.86
0.88
0.90
0.92
μ_1

Simultaneous Confidence Intervals: A confidence region allows us to make inference on the mean vector, but usually we will also want to make inference on the individual mean components. We might consider the usual univariate $(1-\alpha)100\%$ confidence intervals: $\bar{x}_i \pm t_{n-1}(\alpha/2)(s_i/\sqrt{n})$, $i = 1, 2, \ldots, p$. But these are designed to have $(1-\alpha)100\%$ coverage in the univariate case. Their *joint* coverage of the p means will *not* equal $(1-\alpha)100\%$. For example, consider a random sample from the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, where the variables are independent of one another ($\boldsymbol{\Sigma}$ is diagonal). Because of independence, we can factor the joint coverage probability:

$$P\left[\bigcap_{i=1}^{p}\left(\mu_i \in \bar{X}_i \pm t_{n-1}(\alpha/2)(s_i/\sqrt{n})\right)\right] = (1-\alpha)^p$$

If $1-\alpha = 0.95$ and $p = 6$, the joint coverage probability is $(0.95)^6 = 0.74$. In other words, the individual intervals are too *narrow* to maintain an overall $1-\alpha$ coverage probability. This suggests widening the intervals by an appropriate amount, so that the joint coverage probability *is* maintained at $1-\alpha$. We call such intervals *simultaneous confidence intervals*.

Simultaneous Intervals for Linear Combinations: Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $Z = \mathbf{a}'\mathbf{X}$ be a linear combination of the components of $\mathbf{X}$. We then know that $Z \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. Similarly, given a random sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we have $\bar{z} = \mathbf{a}'\bar{\mathbf{x}}$ and $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$. Thus, for a *fixed* choice of $\mathbf{a}$,

$$\bar{z} \pm t_{n-1}(\alpha/2)(s_z/\sqrt{n}) \quad \Longleftrightarrow \quad \mathbf{a}'\bar{\mathbf{x}} \pm t_{n-1}(\alpha/2)\frac{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval for $\mu_z = \mathbf{a}'\boldsymbol{\mu}$. We can re-express this to say that the interval contains all $\mathbf{a}'\boldsymbol{\mu}$ values such that

$$\frac{n\left(\mathbf{a}'\left(\bar{\mathbf{x}} - \boldsymbol{\mu}\right)\right)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} \leq t_{n-1}^2(\alpha/2)$$

Maximization Lemma: Let $\underset{(p\times p)}{\mathbf{B}}$ be positive definite and $\underset{(p\times 1)}{\mathbf{d}}$ be a given vector. Then, for an arbitrary nonzero vector $\underset{(p\times 1)}{\mathbf{x}}$,

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{(\mathbf{x}'\mathbf{d})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{d}'\mathbf{B}^{-1}\mathbf{d}$$

with the maximum attained when $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{d}$ for any constant $c \neq 0$.

Maximization of Quadratic Forms for Points on the Unit Sphere: Let $\underset{(p\times p)}{\mathbf{B}}$ be a positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and associated normalized eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$. Then

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_1 \qquad \text{and} \qquad \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_p$$

The maximum and minimum are attained when $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{x} = \mathbf{e}_p$, respectively. Also,

$$\max_{\mathbf{x}\perp\mathbf{e}_1,\cdots,\mathbf{e}_k} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_{k+1}$$

and this is attained when $\mathbf{x} = \mathbf{e}_{k+1}$, $k = 1, 2, \ldots, p-1$.

This is a slide from Topic 2.

See textbook for proofs.

Simultaneous Intervals for Linear Combinations: Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $Z = \mathbf{a}'\mathbf{X}$ be a linear combination of the components of $\mathbf{X}$. We then know that $Z \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. Similarly, given a random sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we have $\bar{z} = \mathbf{a}'\bar{\mathbf{x}}$ and $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$. Thus, for a *fixed* choice of $\mathbf{a}$,

$$\bar{z} \pm t_{n-1}(\alpha/2)(s_z/\sqrt{n}) \quad \Longleftrightarrow \quad \mathbf{a}'\bar{\mathbf{x}} \pm t_{n-1}(\alpha/2)\frac{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval for $\mu_z = \mathbf{a}'\boldsymbol{\mu}$. We can re-express this to say that the interval contains all $\mathbf{a}'\boldsymbol{\mu}$ values such that

$$\frac{n\left(\mathbf{a}'\left(\bar{\mathbf{x}} - \boldsymbol{\mu}\right)\right)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} \leq t_{n-1}^2(\alpha/2)$$

By the maximization lemma, with $\mathbf{x} = \mathbf{a}$, $\mathbf{d} = (\bar{\mathbf{x}} - \boldsymbol{\mu})$, and $\mathbf{B} = \mathbf{S}$, we have

$$\max_{\mathbf{a}} \frac{n\left(\mathbf{a}'\left(\bar{\mathbf{x}} - \boldsymbol{\mu}\right)\right)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} = n\left[\max_{\mathbf{a}} \frac{\left(\mathbf{a}'\left(\bar{\mathbf{x}} - \boldsymbol{\mu}\right)\right)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}}\right] = n\left(\bar{\mathbf{x}} - \boldsymbol{\mu}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}} - \boldsymbol{\mu}\right) = T^2$$

with equality if $\mathbf{a} \propto \mathbf{S}^{-1}\left(\bar{\mathbf{x}} - \boldsymbol{\mu}\right)$.

Simultaneous Intervals for Linear Combinations: Consider modifying the t-based confidence interval

$$\mathbf{a}'\bar{\mathbf{x}} \pm t_{n-1}(\alpha/2)\frac{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}{\sqrt{n}}$$

by replacing the $t_{n-1}(\alpha/2)$ critical value with the square root of the critical value from the T^2 test:

$$\mathbf{a}'\bar{\mathbf{x}} \pm \sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\frac{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}{\sqrt{n}}$$

This T^2-*interval* will be wider than the t-interval. Furthermore,

$$\begin{aligned}
P\left(\frac{n\left(\mathbf{a}'\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)\right)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} \leq \frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha) \quad \underline{\text{for all } \mathbf{a} \text{ simultaneously}}\right) \\
= 1 - P\left(\frac{n\left(\mathbf{a}'\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)\right)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} > \frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha) \quad \underline{\text{for at least one } \mathbf{a}}\right) \\
= 1 - P\left(n\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right) > \frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)\right) \\
= 1-\alpha
\end{aligned}$$

The T^2-intervals *simultaneously* cover $\mathbf{a}'\boldsymbol{\mu}$ for all possible $\mathbf{a}$ with probability $1-\alpha$. For example, with $\mathbf{a}' = [0,\ldots,0,1,0,\ldots,0]$, we have an interval for μ_i. And, with $\mathbf{a}' = [0,\ldots,0,1,0,\ldots,0,-1,0,\ldots,0]$, we have an interval for $\mu_i - \mu_j$.

Container Data: Taking $\mathbf{a}' = [1, 0]$ gives $\mathbf{a}'\bar{\mathbf{x}} = \bar{x}_1 = 0.8802$ and $\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}/n} = \sqrt{s_{11}/n} = 0.0163$. Similarly, $\mathbf{a}' = [0, 1]$ gives $\mathbf{a}'\bar{\mathbf{x}} = \bar{x}_2 = 0.1237$ and $\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}/n} = \sqrt{s_{22}/n} = 0.0053$. And $\mathbf{a}' = [1, -1]$ gives $\mathbf{a}'\bar{\mathbf{x}} = \bar{x}_1 - \bar{x}_2 = 0.7565$ and $\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}/n} = \sqrt{(s_{11} + s_{22} - 2s_{12})/n} = 0.0156$. The T^2 95% confidence intervals for the two mean components are therefore

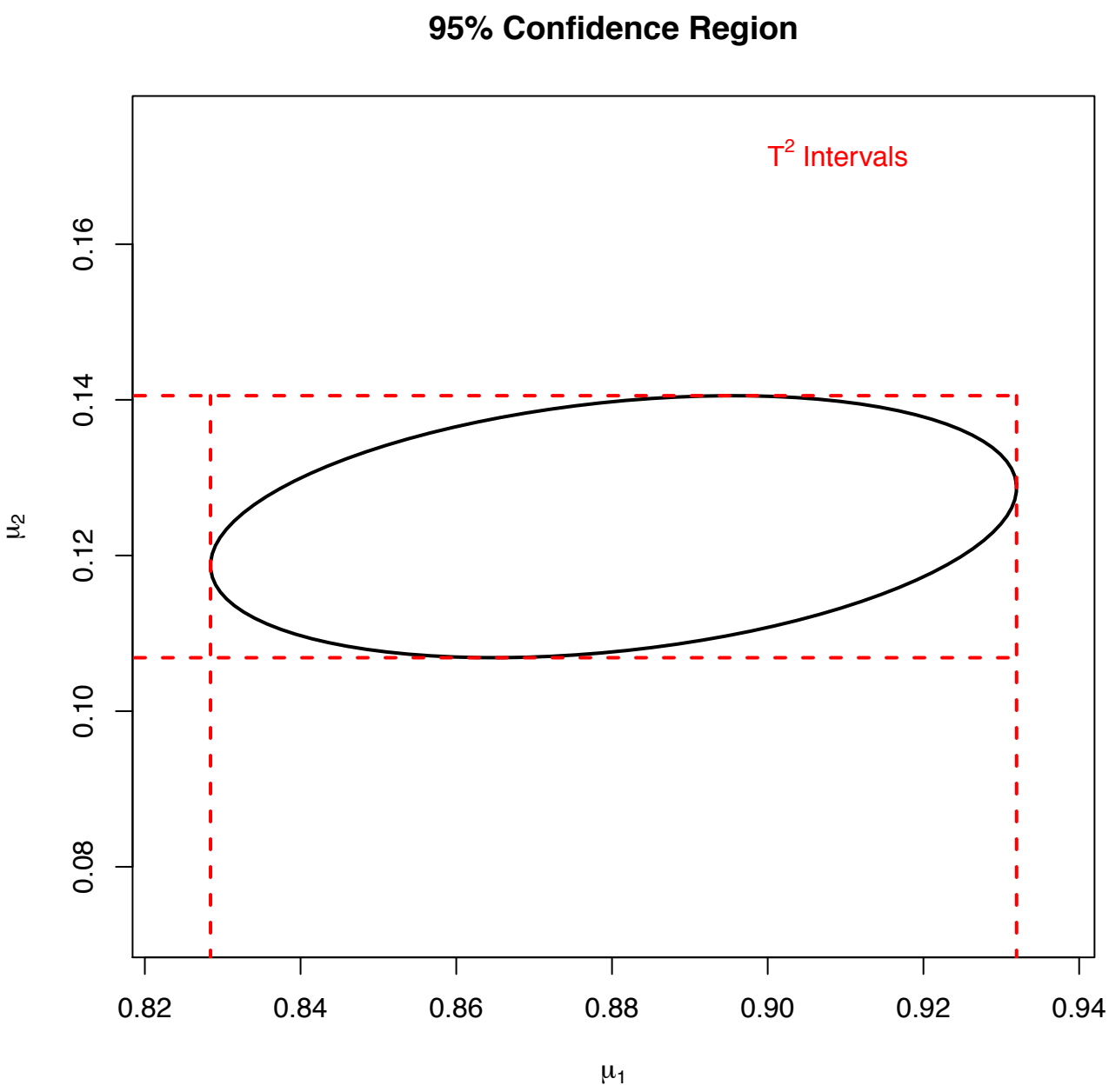
$$\text{Aluminum:} \quad 0.8802 \pm \sqrt{10.0327}(0.0163) \quad \text{or} \quad (0.8286, 0.9318)$$

$$\text{Iron:} \quad 0.1237 \pm \sqrt{10.0327}(0.0053) \quad \text{or} \quad (0.1069, 0.1405)$$

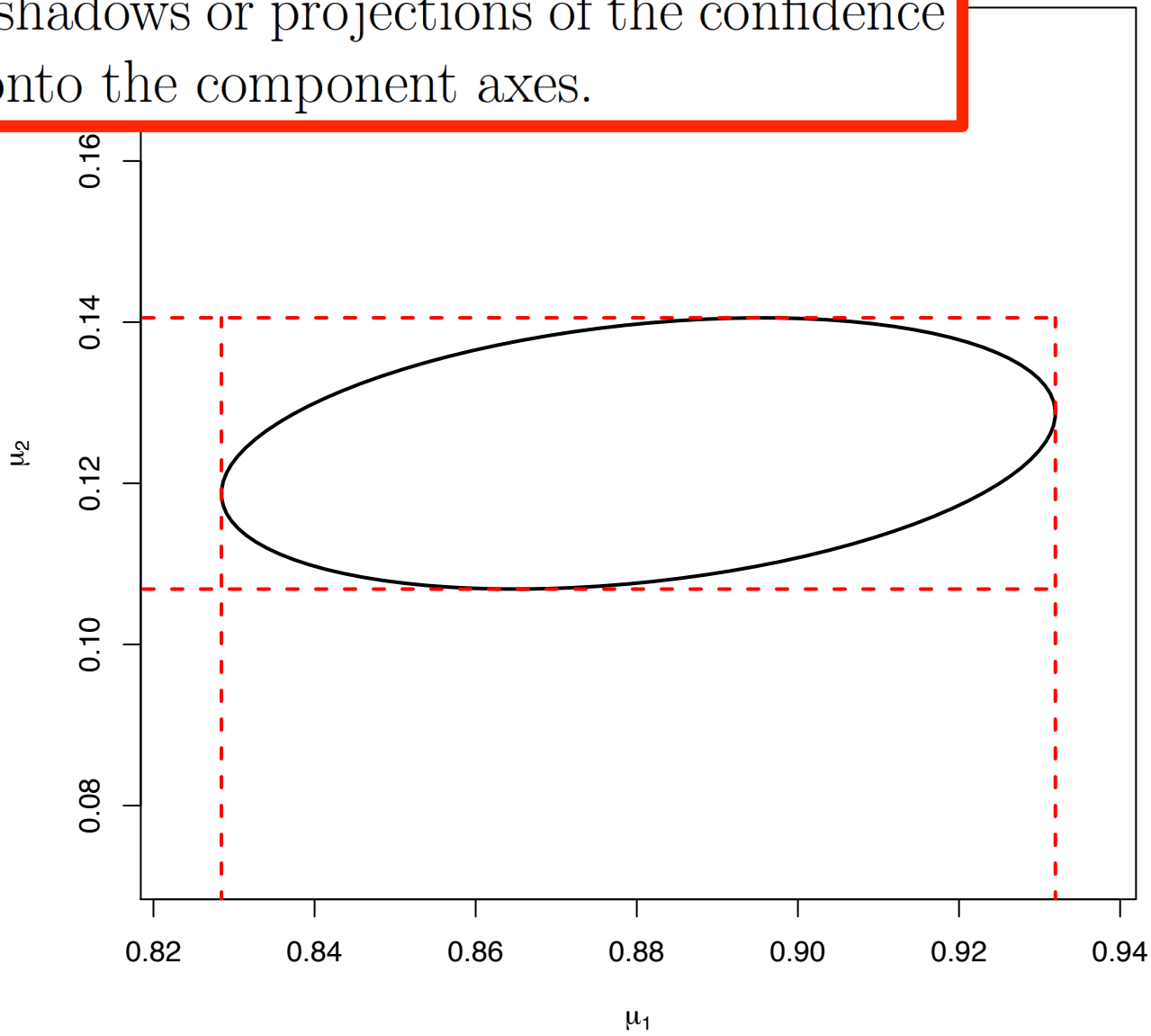
Note that the T^2 intervals are $100(\sqrt{10.0327}/2.2622 - 1) \approx 40\%$ wider than the one-at-a-time t-based 95% confidence intervals, since $t_9(0.025) = 2.2622$. Then, the T^2 95% confidence interval for the mean difference $\mu_1 - \mu_2$ is

$$0.7565 \pm \sqrt{10.0327}(0.0156) \quad \text{or} \quad (0.7070, 0.8059)$$

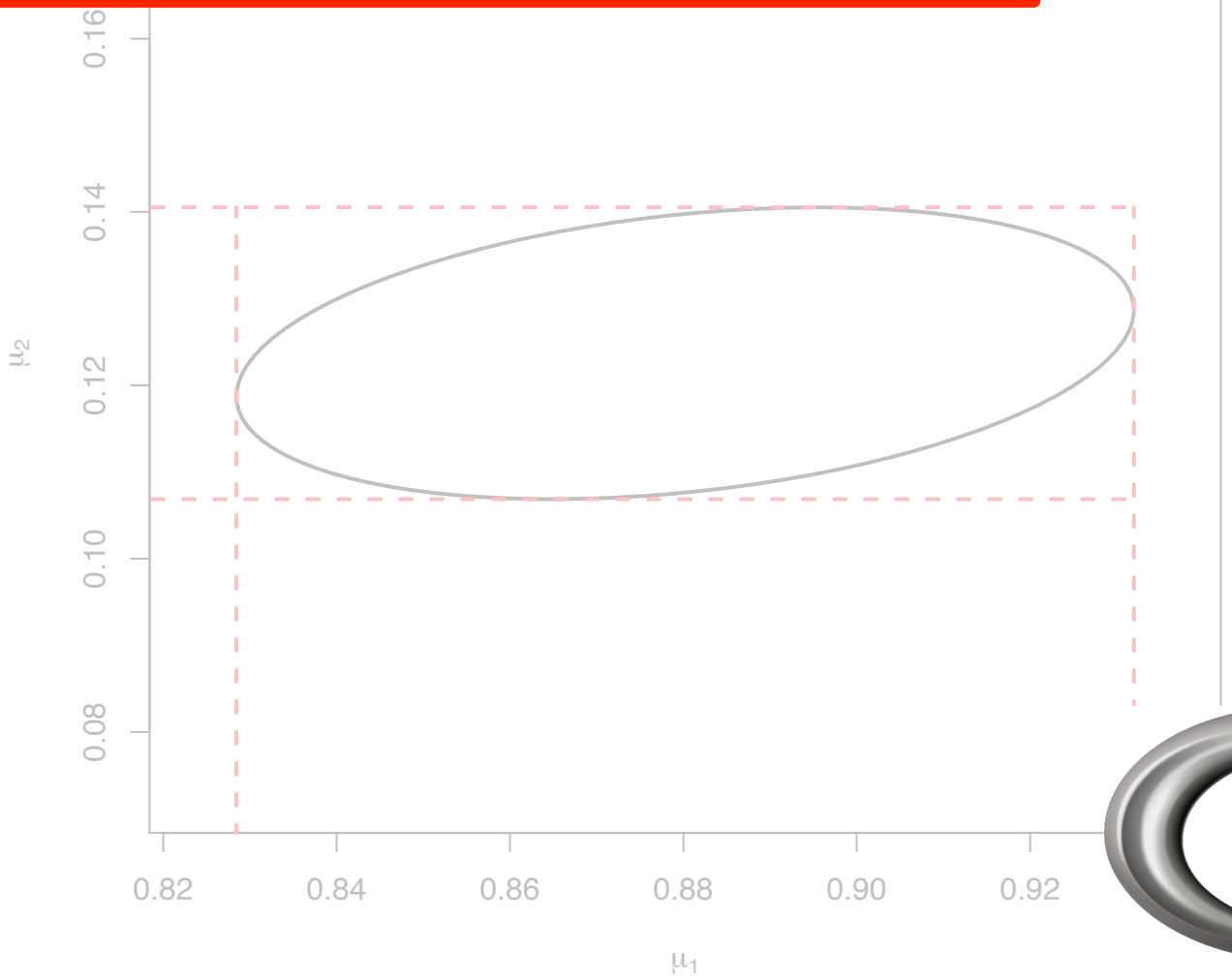
We could compute as many intervals like these as we liked, and the simultaneous coverage probability would continue to be bounded by $1 - \alpha$.


95% Confidence Region
T2 Intervals
0.16
0.14
0.12
0.10
0.08
μ2
0.82
0.84
0.86
0.88
0.90
0.92
0.94
μ1

Note that the T^2 intervals for the two means are the shadows or projections of the confidence region onto the component axes.



Note that the T^2 intervals for the two means are the shadows or projections of the confidence region onto the component axes.



Note that the T^2 intervals for the two means are the shadows or projections of the confidence region onto the component axes.

The T^2 intervals will be wider than necessary if they are only to be used for a predefined collection of inferences. We might, for example, only really want intervals for the mean components (and maybe their differences), in which case the T^2 intervals will have coverage probability greater than $1 - \alpha$. In this case, there is a better alternative.

Bonferroni Intervals: Suppose that, prior to collecting data, we require m confidence intervals for the linear combinations $\mathbf{a}_1'\boldsymbol{\mu}, \mathbf{a}_2'\boldsymbol{\mu}, \ldots, \mathbf{a}_m'\boldsymbol{\mu}$. Let C_i denote the confidence statement about the value of $\mathbf{a}_i'\boldsymbol{\mu}$ with $P(C_i \text{ true}) = 1 - \alpha_i$, $i = 1, 2, \ldots, m$. An application of the *Bonferroni Inequality* shows that

$$\begin{aligned} P(\text{all } C_i \text{ true}) &= 1 - P(\text{at least one } C_i \text{ false}) \\ &\geq 1 - \sum_{i=1}^{m} P(C_i \text{ false}) = 1 - \sum_{i=1}^{n} (1 - P(C_i \text{ true})) \\ &= 1 - (\alpha_1 + \alpha_2 + \ldots + \alpha_m) \end{aligned}$$

Thus, if we construct each of the individual m confidence statements to have coverage probability $1 - \alpha_i = 1 - \alpha/m$, we will have *simultaneous* coverage probability $\geq \alpha$. One possibility would be to just construct the $m = p$ Bonferroni-adjusted t intervals for the mean components:

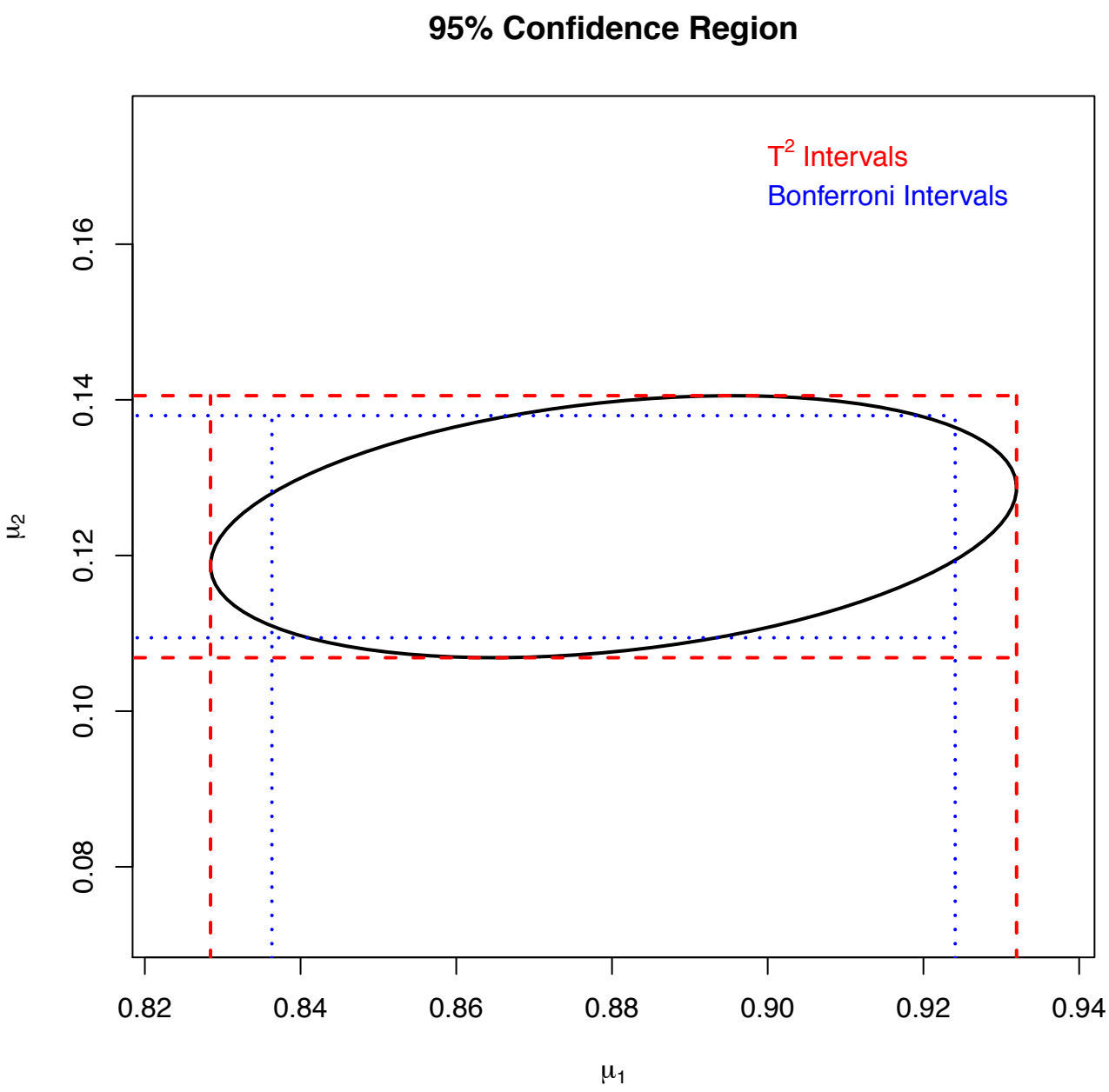
$$\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}, \quad i = 1, 2, \ldots, p$$

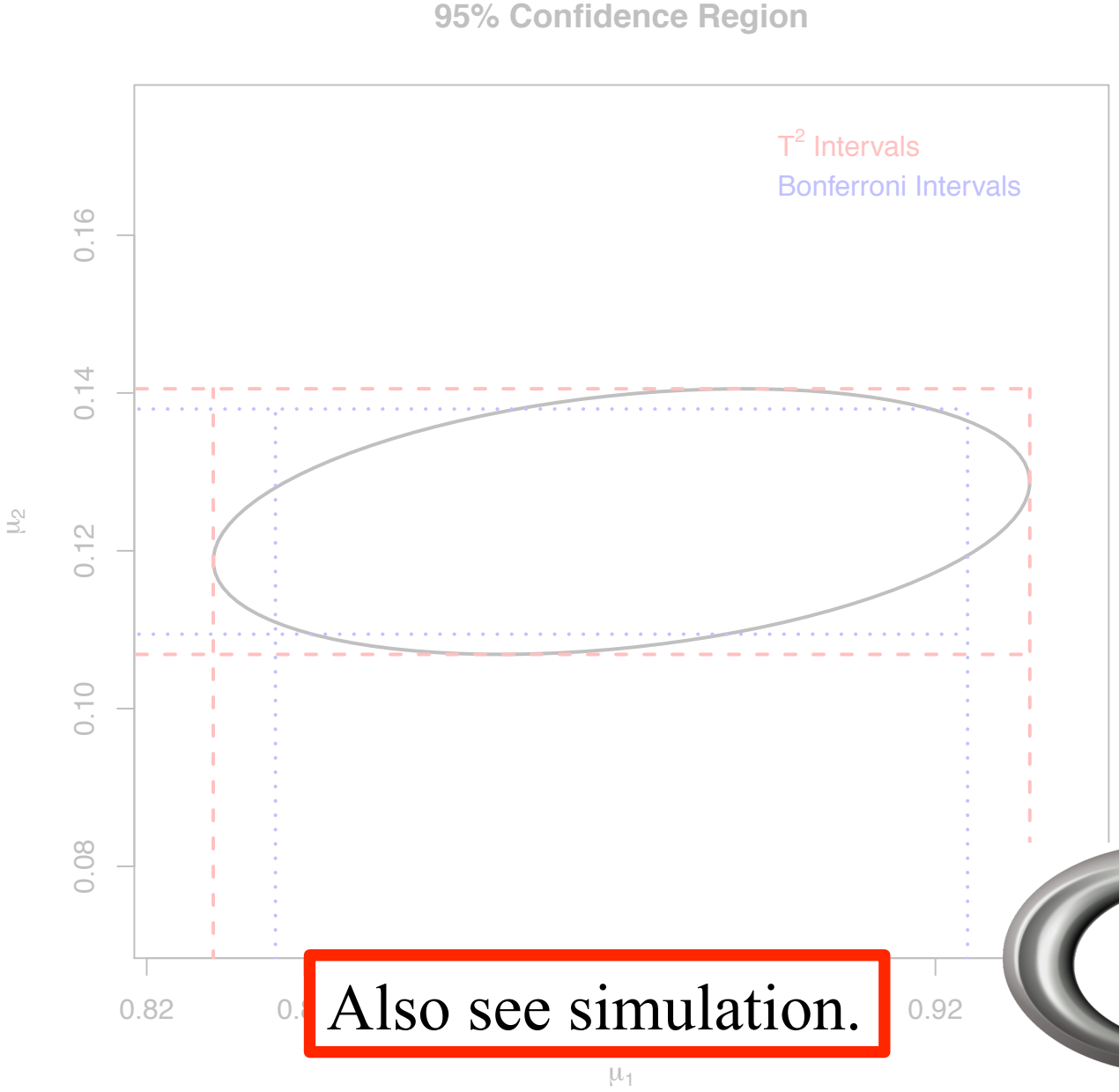
The simultaneous coverage probability is $\geq 1 - \alpha$, and these will be narrower (more precise) than the corresponding T^2 intervals.

Container Data: Bonferroni 95% confidence intervals for the $m = p = 2$ mean concentrations of aluminum and iron are t-based intervals that use the critical value $t_9\,(0.05/4) = 2.6850$ instead of $t_9(0.05/2) = 2.2622$:

$$\begin{aligned} \text{Aluminum:} \quad & 0.8802 \pm 2.6850(0.0163) \quad \text{or} \quad (0.8364, 0.9240) \\ \text{Iron:} \quad & 0.1237 \pm 2.6850(0.0053) \quad \text{or} \quad (0.1095, 0.1379) \end{aligned}$$

The corresponding T^2 intervals are $100(\sqrt{10.0327}/2.6850 - 1) \approx 18\%$ wider than the Bonferroni intervals. We could include an interval for the mean difference as well, if we instead used $t_9\,(0.05/6) = 2.9333$ in constructing the $m = 3$ intervals.


95% Confidence Region
T2 Intervals
Bonferroni Intervals
0.16
0.14
0.12
0.10
0.08
μ2
0.82
0.84
0.86
0.88
0.90
0.92
0.94
μ1



Also see simulation.

Comparison of Critical Values: Here is a table showing the critical values for 95% confidence intervals of type: (1) one-at-a-time t-based, (2) T^2, and (3) Bonferroni with $m = p$, for varying n and p. The T^2 and Bonferroni intervals have simultaneous 95% coverage, but the t intervals do not. In general, simultaneous intervals get wider for fixed n as p increases and get narrower for fixed p as n increases.

		T^2			Bonferroni		
n	t	$p = 2$	$p = 4$	$p = 10$	$p = 2$	$p = 4$	$p = 10$
15	2.14	2.86	4.13	11.51	2.51	2.86	3.33
25	2.06	2.67	3.60	6.38	2.39	2.70	3.09
50	2.01	2.55	3.31	5.04	2.31	2.59	2.94
100	1.98	2.50	3.19	4.62	2.28	2.54	2.87
∞	1.96	2.45	3.08	4.28	2.24	2.50	2.81

The last row of the table corresponds to an infinite sample size. For the t and Bonferroni intervals, the limiting critical values come from the standard normal distribution. For the T^2 intervals, as we will see shortly, the limiting critical values come from the χ^2_p distribution.

Large Sample Inference

Large Sample Inference: When n is large relative to p, we can avoid the assumption of multivariate normality. We saw in Topic 4 that

$$T^2 = n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right) \dot{\sim} \chi^2_p$$

when $n - p$ is large. We can therefore easily modify our hypothesis testing and confidence region / interval approaches by replacing

$$\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)$$

with $\chi^2_p(\alpha)$. Thus, to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ at significance level α, we reject H_0 if

$$n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right) > \chi^2_p(\alpha)$$

Similarly, simultaneous T^2-based $100(1-\alpha)\%$ confidence intervals for linear combinations $\mathbf{a}'\boldsymbol{\mu}$ can be formed as

$$\mathbf{a}'\bar{\mathbf{X}} \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}$$

Simultaneous Bonferroni intervals can be constructed as before, but with $z(\alpha/m)$ in place of $t_{n-1}(\alpha/m)$.

Large Sample Inference: When n is large relative to p, we can avoid the assumption of multivariate normality. We saw in Topic 4 that

There is no one answer for how "large" $n - p$ needs to be for the asymptotic methods to be used. For example, $n = 50$ and $p = 2$ will mostly likely be fine, while $n = 100$ and $p = 52$ may well *not* be fine, even though $n - p = 58$ in both cases. With p large, the required sample size for the asymptotic results to hold may be *very* large. It is always a good idea to explore the data with pictures and summary statistics before formally analyzing it. In the presence of outliers or extreme skewness, corrective actions, including transformations, may be appropriate.

Similarly, simultaneous T^2-based $100(1-\alpha)\%$ confidence intervals for linear combinations $\mathbf{a}'\boldsymbol{\mu}$ can be formed as

$$\mathbf{a}'\bar{\mathbf{X}} \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}$$

Simultaneous Bonferroni intervals can be constructed as before, but with $z(\alpha/m)$ in place of $t_{n-1}(\alpha/m)$.

Example: Musical Aptitude We have scores for $n = 96$ Finnish 12th graders on $p = 7$ musical aptitude variables. Summary statistics are shown below.

Variable	$\bar{x}_i$	$\sqrt{s_{ii}}$
$X_1 =$ melody	28.1	5.76
$X_2 =$ harmony	26.6	5.85
$X_3 =$ tempo	35.4	3.82
$X_4 =$ meter	34.2	5.12
$X_5 =$ phrasing	23.6	3.76
$X_6 =$ balance	22.0	3.93
$X_7 =$ style	22.7	4.03

Since $n - p$ seems reasonable large, we will construct large-sample simultaneous 95% confidence intervals for the means. Since $\chi^2_7(0.05) = 14.07$ and $z(0.05/14) = 2.69$, the T^2 and Bonferroni intervals for μ_i are $\bar{x}_i \pm \sqrt{14.07}(\sqrt{s_{ii}}/\sqrt{96})$ and $\bar{x}_i \pm 2.69(\sqrt{s_{ii}}/\sqrt{96})$, respectively.

Summary statistics are from Tables 5.5 in textbook. The authors did not provide the raw data.

	t		T^2		Bonferroni	
Variable	Lower	Upper	Lower	Upper	Lower	Upper
$X_1 =$ melody	26.96	29.25	26.52	29.68	25.90	30.30
$X_2 =$ harmony	25.43	27.77	24.99	28.21	24.36	28.84
$X_3 =$ tempo	34.64	36.16	34.35	36.45	33.94	36.86
$X_4 =$ meter	33.18	35.22	32.79	35.61	32.24	36.16
$X_5 =$ phrasing	22.85	24.35	22.57	24.63	22.16	25.04
$X_6 =$ balance	21.21	22.79	20.92	23.08	20.50	23.50
$X_7 =$ style	21.89	23.51	21.59	23.81	21.16	24.24

Large-sample 95% confidence intervals.

	t		T^2		Bonferroni	
Variable	Lower	Upper	Lower	Upper	Lower	Upper
$X_1 =$ melody	26.96	29.25	26.52	29.68	25.90	30.30
$X_2 =$ harmony	25.43	27.77	24.99	28.21	24.36	28.84
$X_3 =$ tempo	34.64	36.16	34.35	36.45	33.94	36.86
$X_4 =$ meter	33.18	35.22	32.79	35.61	32.24	36.16
$X_5 =$ phrasing	22.85	24.35	22.57	24.63	22.16	25.04
$X_6 =$ balance	21.21	22.79	20.92	23.08	20.50	23.50
$X_7 =$ style	21.89	23.51	21.59	23.81	21.16	24.24

Large-sample 95% confidence intervals.

	t		T^2		Bonferroni	
Variable	Lower	Upper	Lower	Upper	Lower	Upper
$X_1 =$ melody	26.93	29.27	26.48	29.72	25.76	30.44
$X_2 =$ harmony	25.41	27.79	24.96	28.24	24.23	28.97
$X_3 =$ tempo	34.63	36.17	34.33	36.47	33.85	36.95
$X_4 =$ meter	33.16	35.24	32.76	35.64	32.12	36.28
$X_5 =$ phrasing	22.84	24.36	22.54	24.66	22.07	25.13
$X_6 =$ balance	21.20	22.80	20.90	23.10	20.41	23.59
$X_7 =$ style	21.88	23.52	21.57	23.83	21.07	24.33

Small-sample (parametric) 95% confidence intervals.

Quality Control Charts

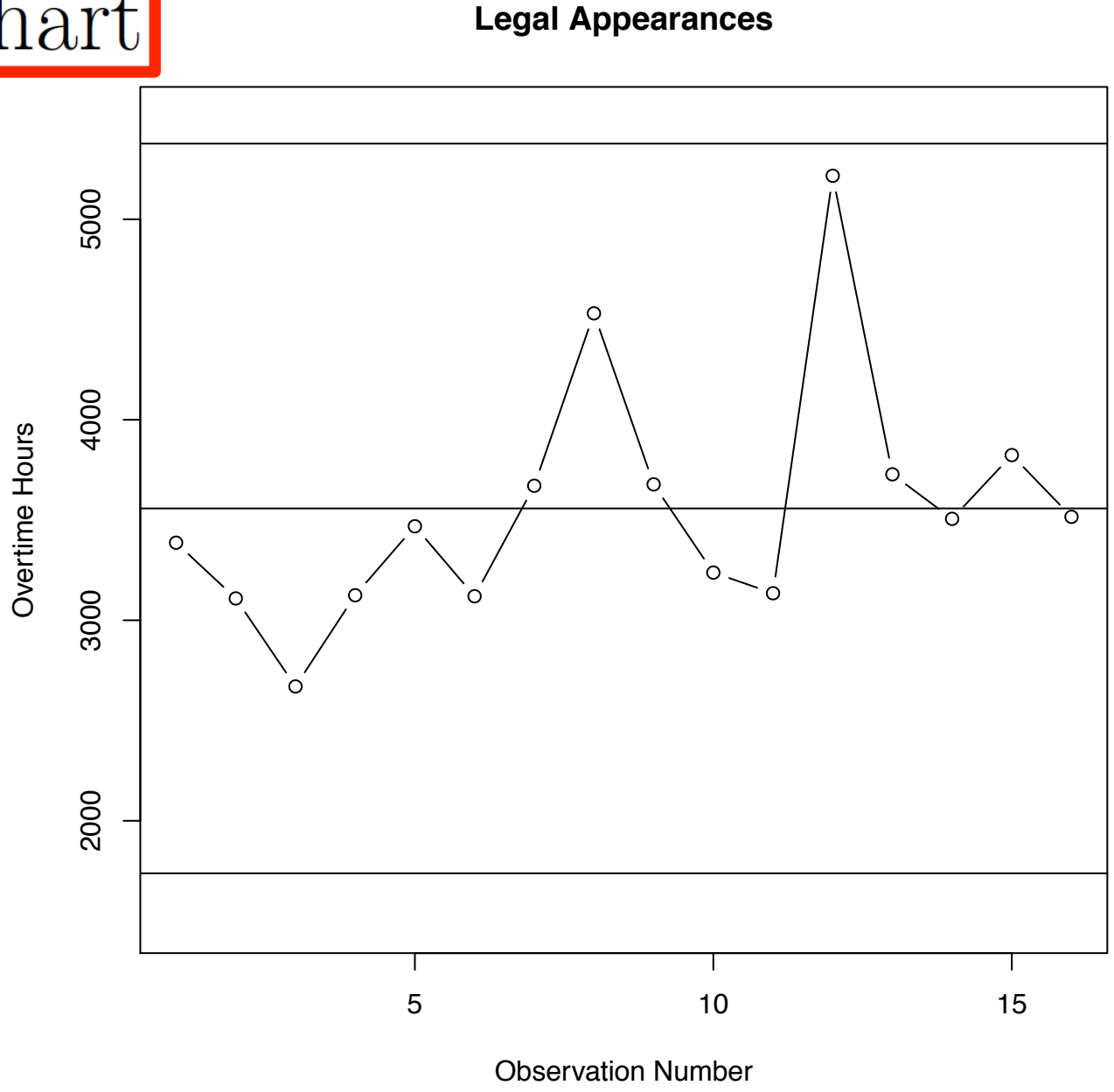
Quality Control Charts: A *quality control chart* is a picture used to assess whether processes are operating within the range of normal variability. We might, for example, be able to detect that an instrument requires repair by monitoring its outputs and inspecting when any outputs fall outside of an expected range. Essentially, quality control charts use the tools we have just learned to create confidence regions for means, triggering action if observations fall outside of the regions. A related task is the construction of *prediction* regions for monitoring quality as new observations are obtained.

Example: Overtime Hours The Madison, Wisconsin, police department monitors its number of overtime hours of five types: X_1: legal appearances, X_2: extraordinary event, X_3: holdover, X_4: compensatory overtime allowed, X_5: meetings. We have data for 16 time periods.

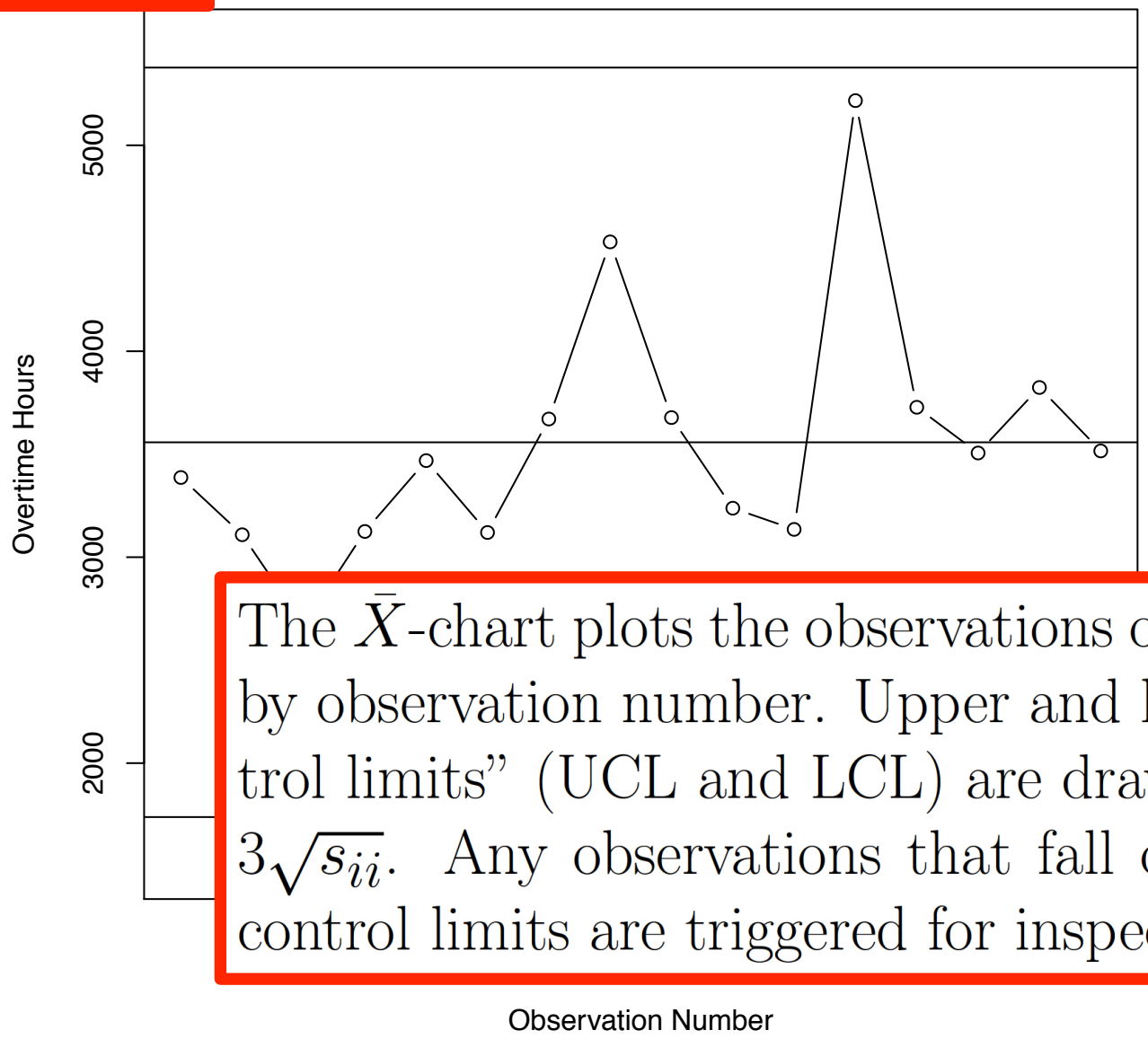
X_1	X_2	X_3	X_4	X_5
3387	2200	1181	14861	236
3109	875	3532	11367	310
2670	957	2502	13329	1182
3125	1758	4510	12328	1208
3469	868	3032	12847	1385
⋮	⋮	⋮	⋮	⋮
3516	1223	1175	15078	161

We will illustrate three different "chart" types: (1) the $\bar{X}$-chart for monitoring one variable at a time, (2) the ellipse format chart for monitoring two variables at a time, and (3) the T^2-chart for monitoring $p \geq 2$ variables at a time. In addition, we will see how these can be modified to monitor *future* observations.

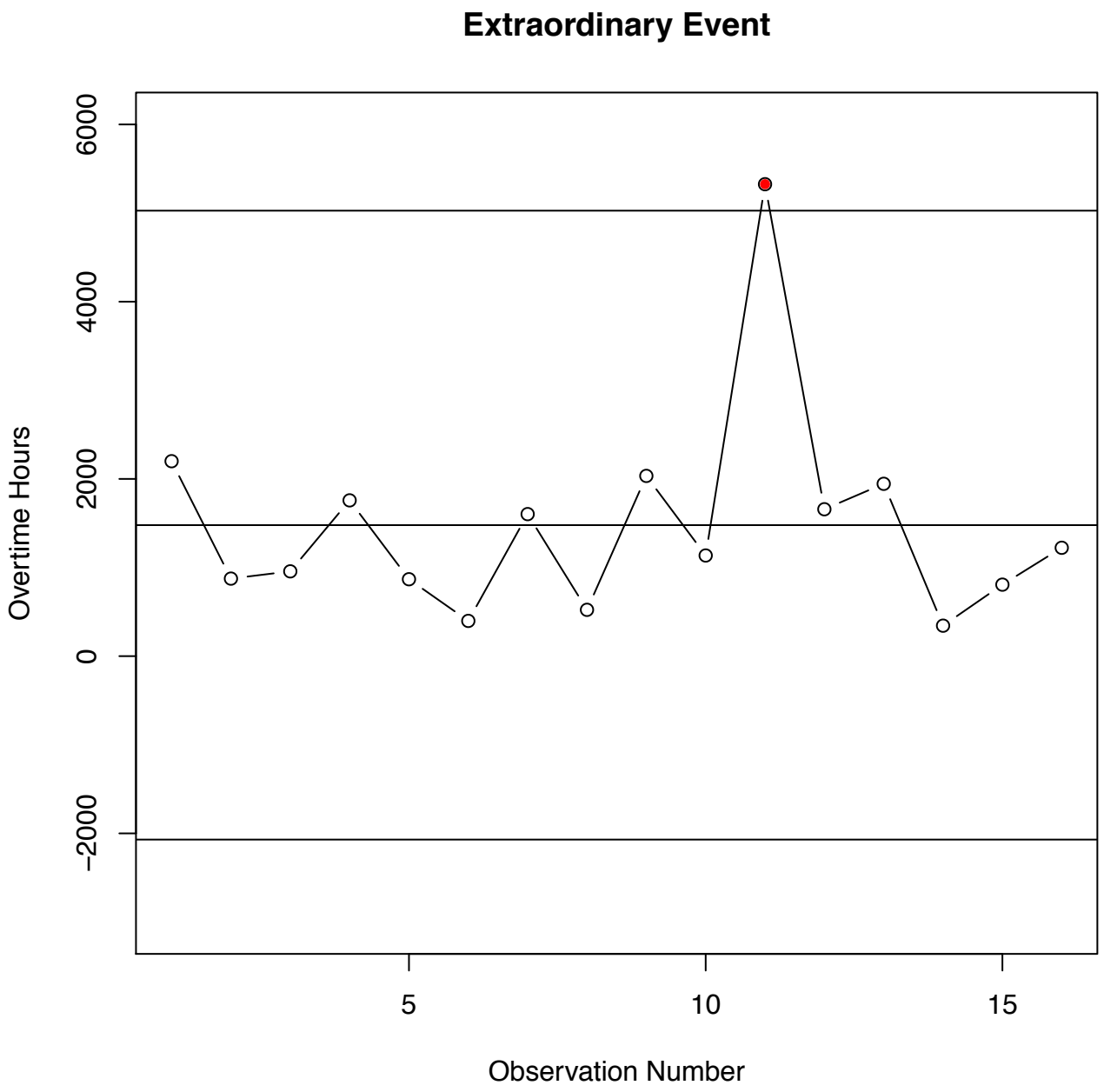
$\bar{X}$-Chart

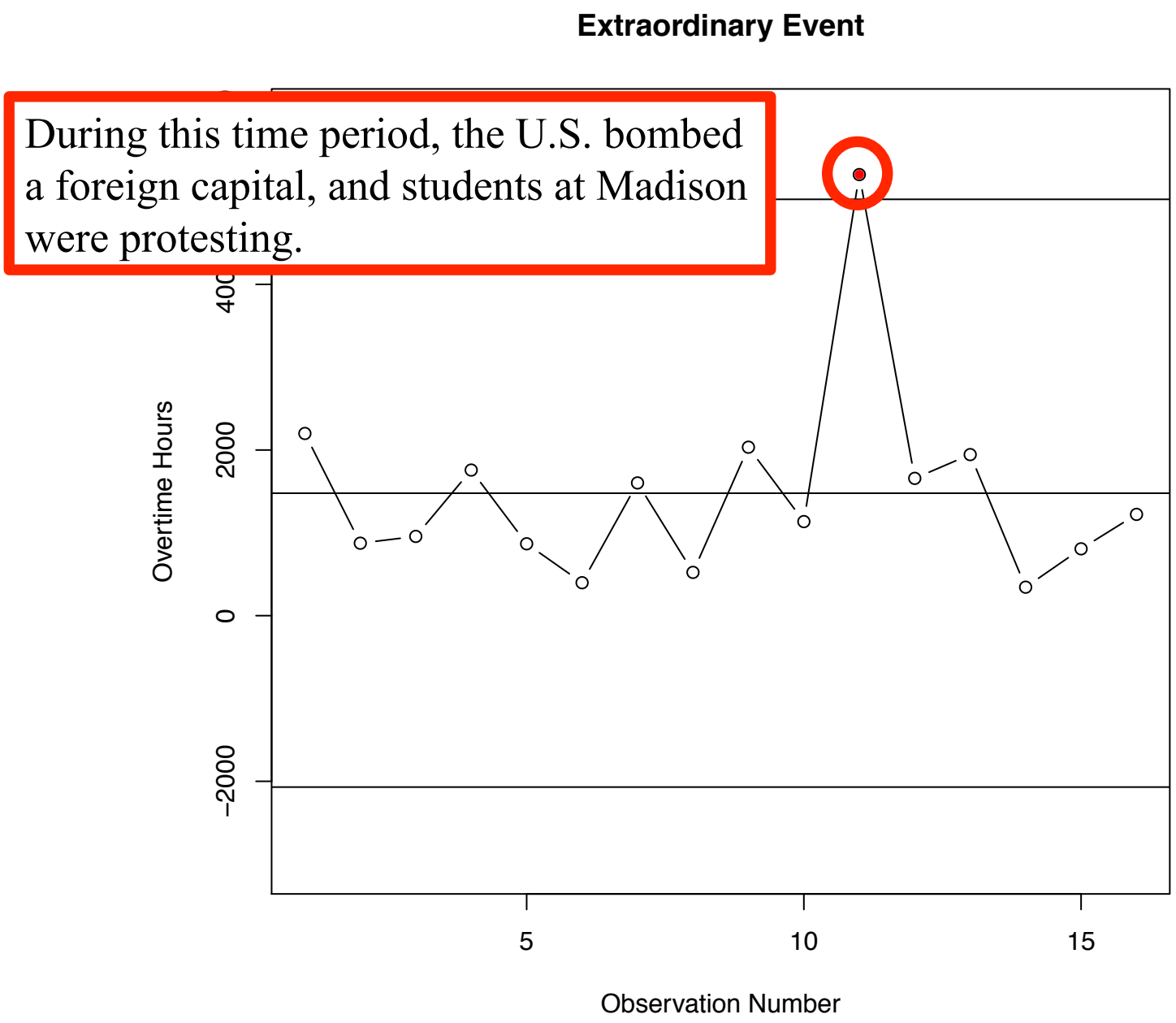

Legal Appearances
Overtime Hours
2000
3000
4000
5000
5
10
15
Observation Number

$\bar{X}$-Chart

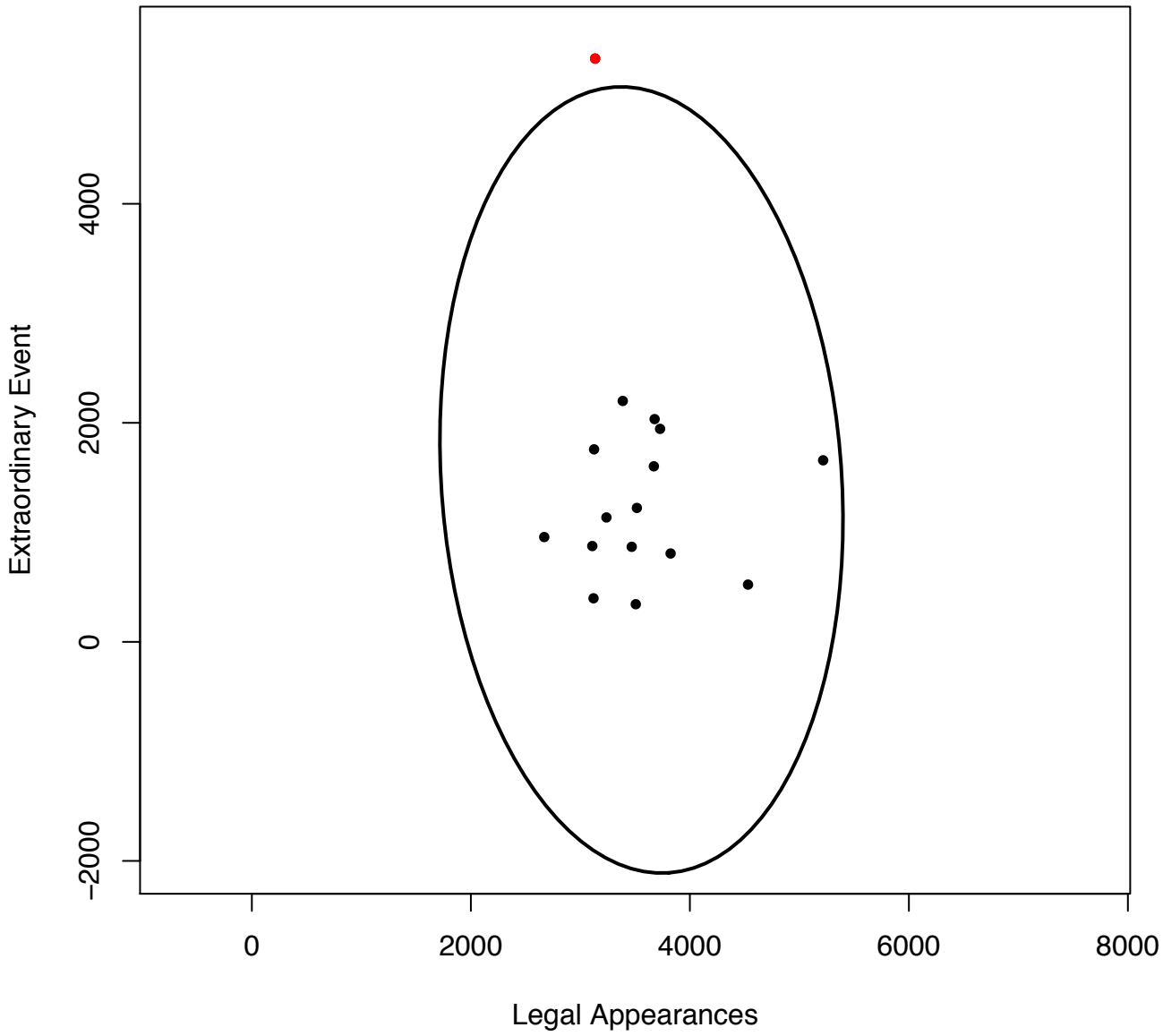


The $\bar{X}$-chart plots the observations of variable i by observation number. Upper and lower "control limits" (UCL and LCL) are drawn at $\bar{x}_i \pm 3\sqrt{s_{ii}}$. Any observations that fall outside the control limits are triggered for inspection.

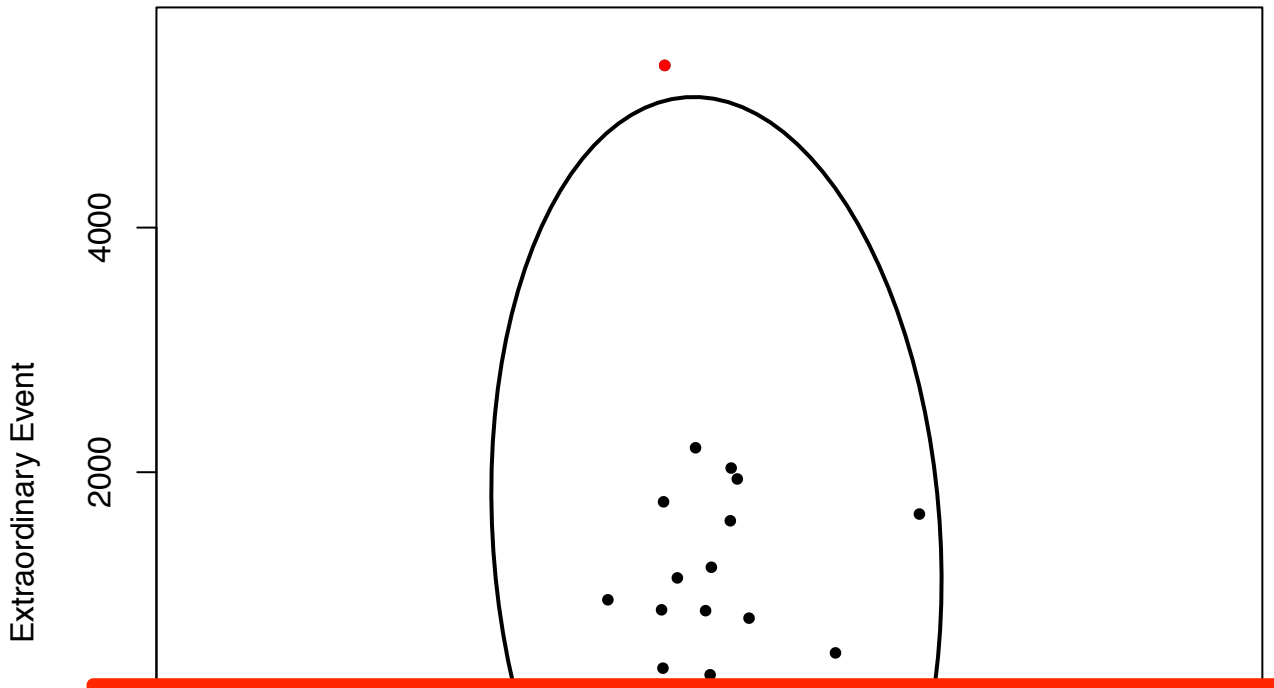

Extraordinary Event
Overtime Hours
6000
4000
2000
0
−2000
5
10
15
Observation Number


Extraordinary Event
During this time period, the U.S. bombed a foreign capital, and students at Madison were protesting.
Overtime Hours
400
2000
0
−2000
5
10
15
Observation Number

Ellipse Format Chart

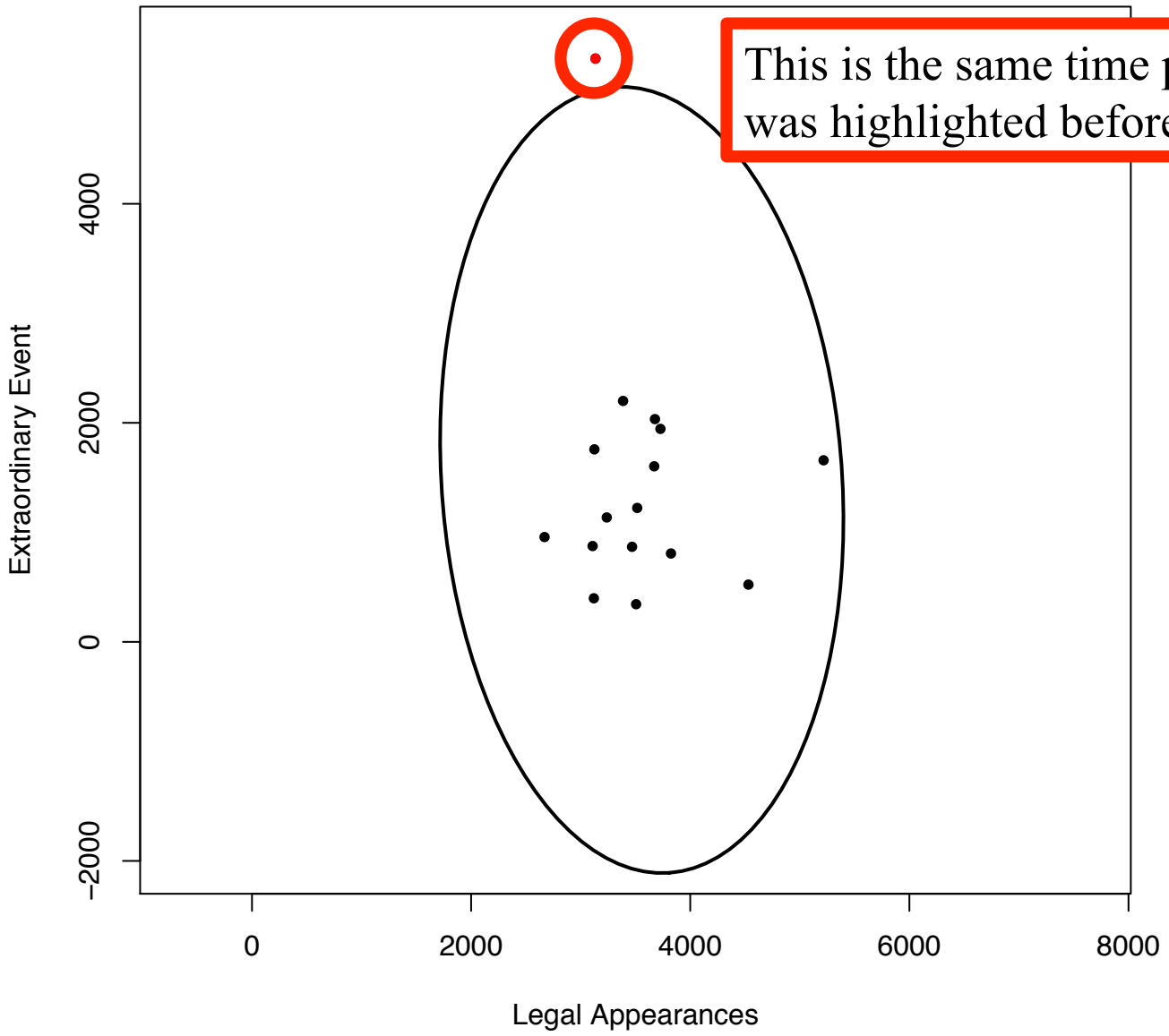


Ellipse Format Chart

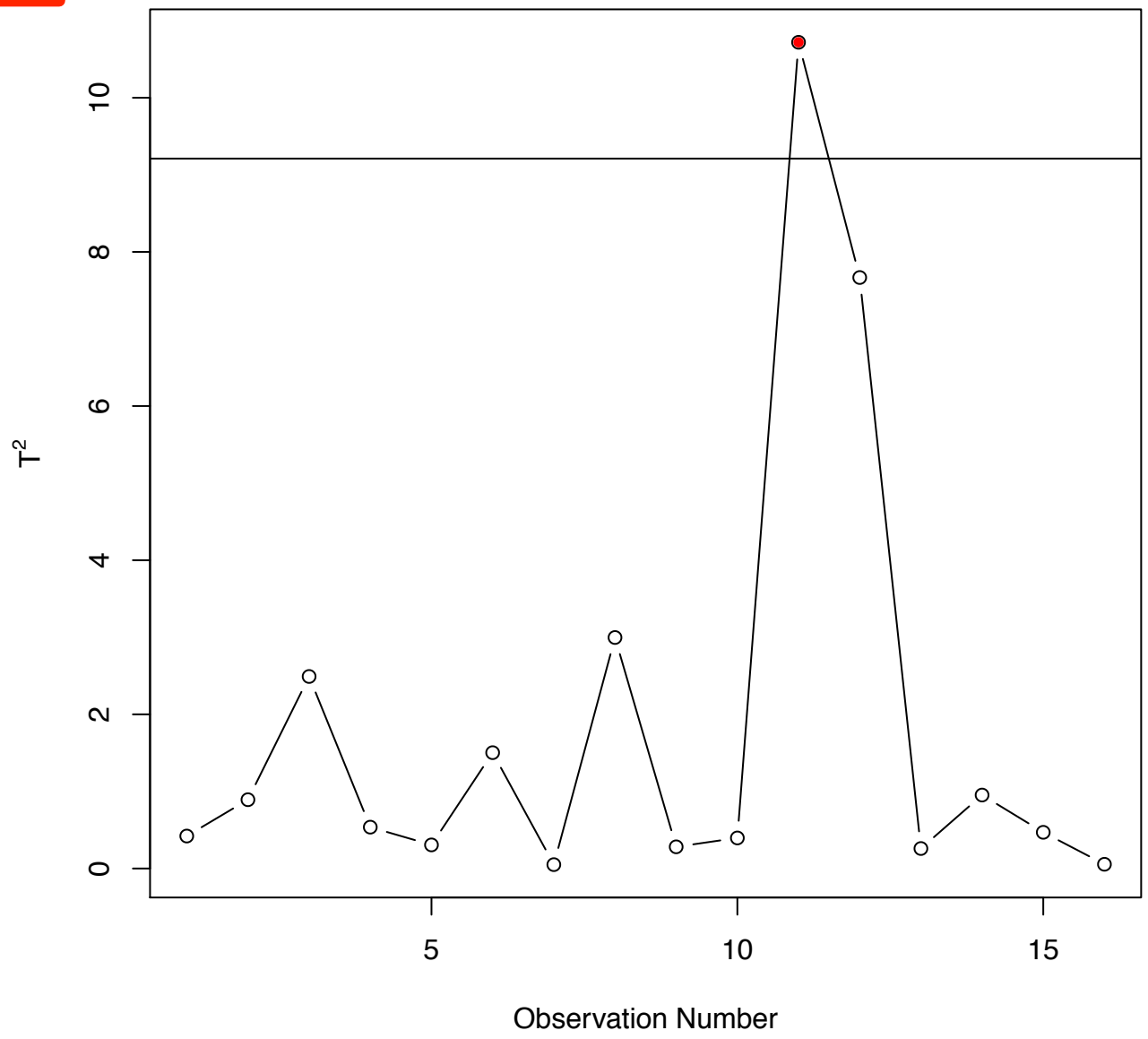


With $\mathbf{x}$ consisting of just two variables, the ellipse format chart assumes that the $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$, $j = 1, 2, \ldots, n$, are independent and approximately χ_2^2-distributed. A $(1-\alpha)100\%$ probability ellipse is drawn for this distribution, and any observations that fall outside the ellipse are triggered for inspection. Note that this is the same rationale behind the "χ^2 plot" from Topic 4 for assessing bivariate normality.

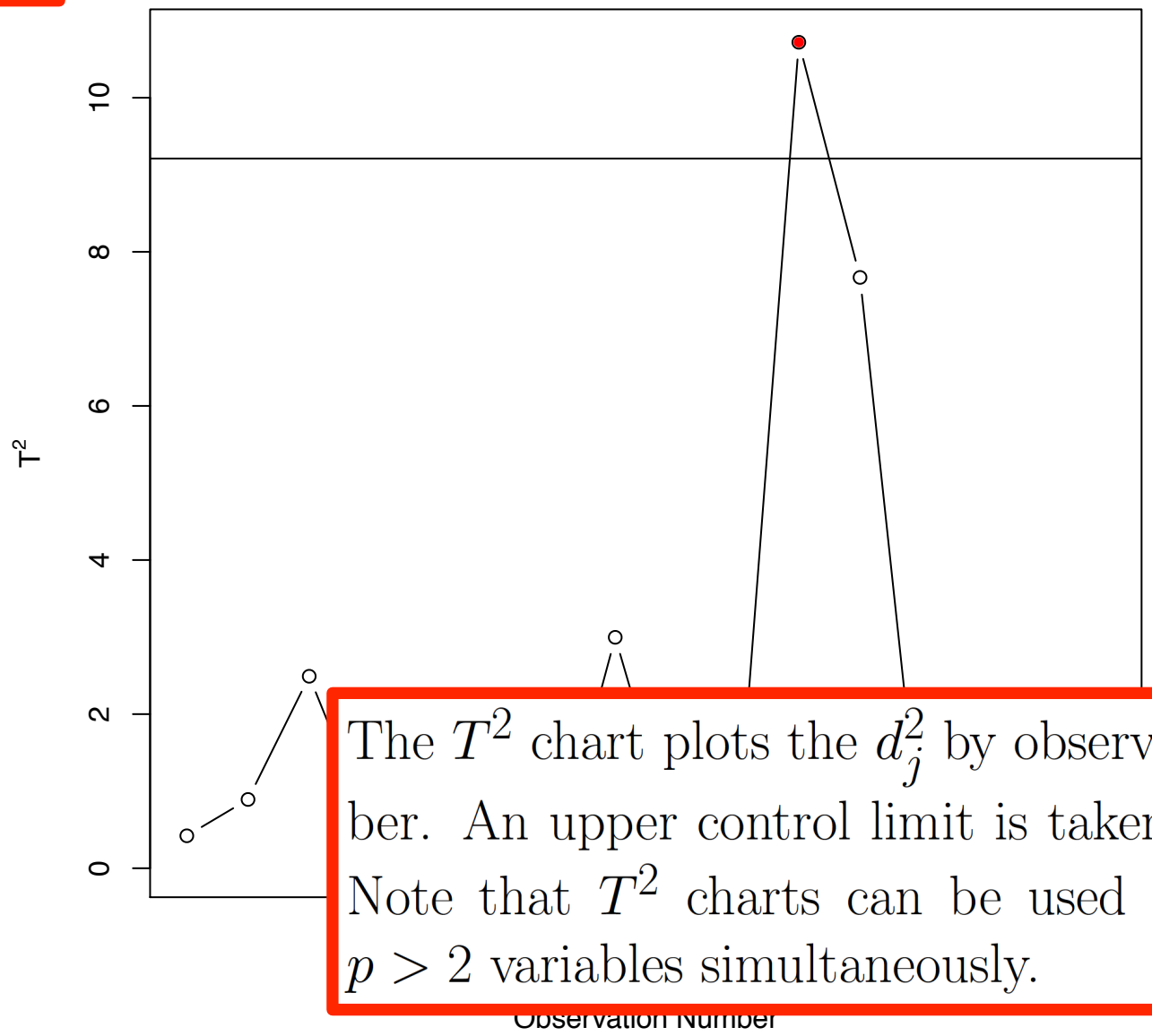
Ellipse Format Chart



T^2 Plot

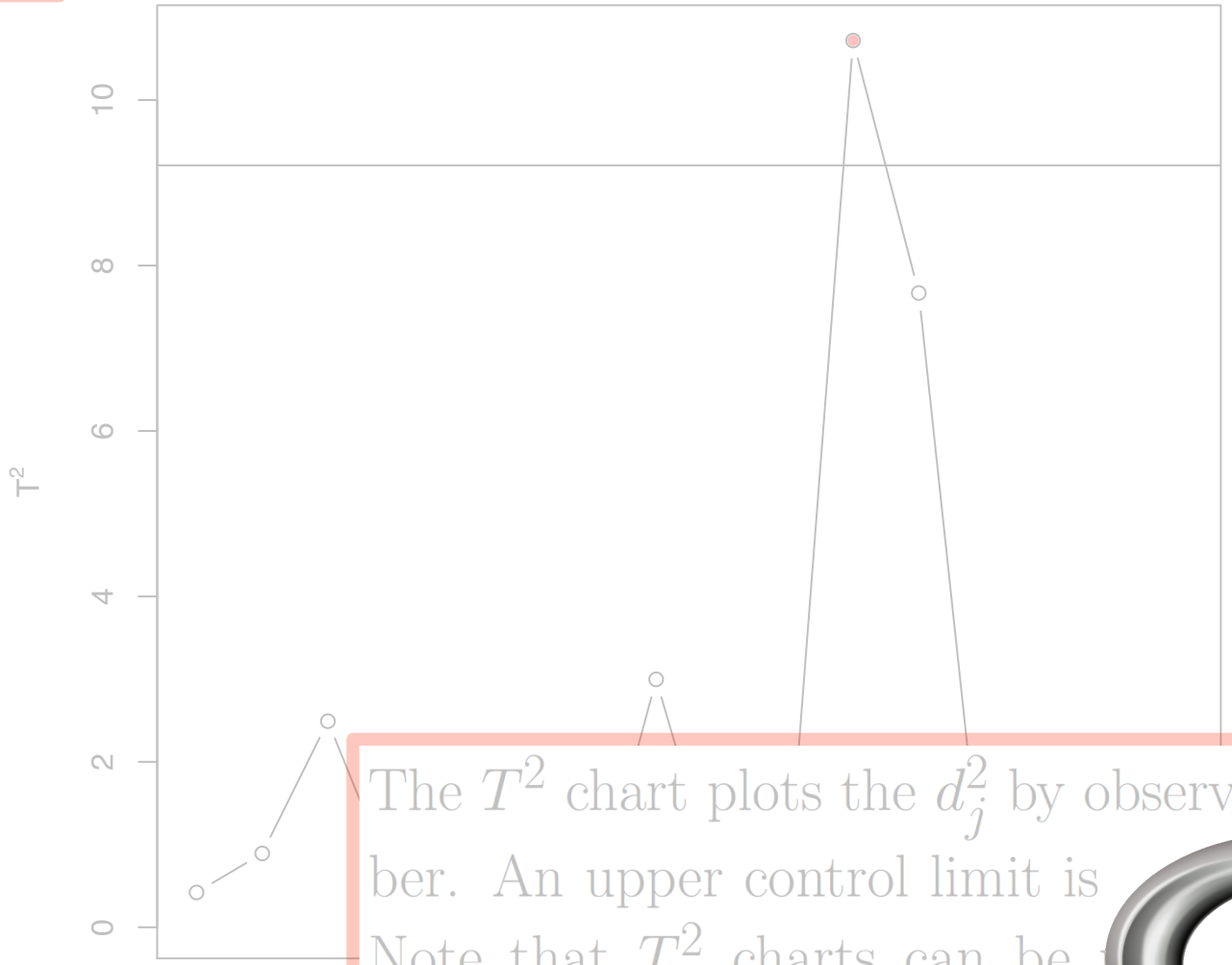


T^2 Plot



The T^2 chart plots the d_j^2 by observation number. An upper control limit is taken as $\chi_p^2(\alpha)$. Note that T^2 charts can be used to monitor $p > 2$ variables simultaneously.

T^2 Plot



The T^2 chart plots the d_j^2 by observation number. An upper control limit is .
Note that T^2 charts can be r
$p > 2$ variables simultaneously.

Control Regions for Future Individual Observations: In order to monitor *future* observations, we require *prediction* regions instead of confidence regions. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Let $\bar{\mathbf{X}}$ and $\mathbf{S}$ be the sample mean vector and covariance matrix, computed with these n observations. Now let $\mathbf{X}$ be a new independent observation from the same distribution. Then*

$$P\left((\mathbf{x}-\bar{\mathbf{x}})'\,\mathbf{S}^{-1}\,(\mathbf{x}-\bar{\mathbf{x}}) \leq \frac{(n^2-1)\,p}{n(n-p)}F_{p,n-p}(\alpha)\right) = 1-\alpha$$

Control ellipses and T^2 charts can then be drawn for new observations by replacing χ^2 critical values with the above F-based ones. Note that, because of the additional variability involved in monitoring single observations, prediction regions will always be larger than their corresponding confidence regions.

*See textbook for proof.

Missing Data

Missing Data: Missing observations are common and can happen due to a variety of reasons. An observation is *missing completely at random* (MCAR) if the probability of missingness does not depend on observable variables or unobservable parameters of interest. *Missing at random* (MAR) means that the probability of missingness depends only on observable variables. For example, if males are less likely to fill out a depression survey but their decision has nothing to do with their depression level, then missingness is MAR. *Not Missing at Random* (NMAR) means that the probability of missingness depends on unobservable variables and / or parameters. For example, if a male's decision to not fill out a depression survey *does* depend on his depression level, then missingness is NMAR. Data with MCAR missingness can be easily handled by only including the complete data in the analysis. By default, this is what most software will do. But MCAR rarely truly holds. NMAR is very difficult to handle and is beyond the scope of this course. MAR is typically handled with either maximum likelihood or multiple imputation. A general technique for maximum likelihood estimation with MAR data is the *Expectation-Maximization* (EM) algorithm. The EM algorithm has a straightforward implementation with multivariate normal data.

Expectation-Maximization (EM) Algorithm: Consider the likelihood $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$. Suppose we don't get to observe values for $\mathbf{Z}$. In order to obtain the maximum likelihood estimate of $\boldsymbol{\theta}$, we can do the following. First, construct an initial estimate $\boldsymbol{\theta}_0$. Then iterate through the following steps until convergence:

- **Expectation (E) Step:** Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ and using the current estimate of the parameters $\boldsymbol{\theta}_t$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = E_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}_t}[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$

- **Maximization (M) Step:** Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ to obtain a revised parameter estimate:

$$\boldsymbol{\theta}_{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$$

Reminder About Conditional Distributions: Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Then the conditional distribution of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is normal with

$$\text{Mean } = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

and

$$\text{Covariance } = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

The covariance result means that

$$\begin{aligned} E\left((\mathbf{X}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_1 - \boldsymbol{\mu}_1)' | \mathbf{x}_2\right) &= E(\mathbf{X}_1\mathbf{X}_1' | \mathbf{x}_2) - 2E(\mathbf{X}_1 | \mathbf{x}_2)\boldsymbol{\mu}_1' + \boldsymbol{\mu}_1\boldsymbol{\mu}_1' \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

so that

$$\begin{aligned} E(\mathbf{X}_1\mathbf{X}_1' | \mathbf{x}_2) &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} + 2E(\mathbf{X}_1 | \mathbf{x}_2)\boldsymbol{\mu}_1' - \boldsymbol{\mu}_1\boldsymbol{\mu}_1' \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} + 2\left(\boldsymbol{\mu}_1\boldsymbol{\mu}_1' + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\boldsymbol{\mu}_1'\right) - \boldsymbol{\mu}_1\boldsymbol{\mu}_1' \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} + \boldsymbol{\mu}_1\boldsymbol{\mu}_1' + 2\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\boldsymbol{\mu}_1' \end{aligned}$$

EM for Multivariate Normal Data: In the multivariate normal case, the EM algorithm need only be concerned with the sufficient statistics

$$\mathbf{T}_1 = \sum_{j=1}^{n} \mathbf{X}_j = n\bar{\mathbf{X}} \quad \text{and} \quad \mathbf{T}_2 = \sum_{j=1}^{n} \mathbf{X}_j \mathbf{X}_j' = (n-1)\mathbf{S} + n\bar{\mathbf{X}}\bar{\mathbf{X}}'$$

For each vector $\mathbf{x}_j$ with missing values, let $\mathbf{x}_j^{(1)}$ denote the missing components and $\mathbf{x}_j^{(2)}$ denote those components which are available. Given current parameter estimates $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$, the **E Step** proceeds as follows. Replace the missing components in $\mathbf{T}_1$ with the mean of the conditional distribution of $\mathbf{x}_j^{(1)}$ given $\mathbf{x}_j^{(2)}$:

$$\widetilde{\mathbf{x}}_j^{(1)} = E\left(\mathbf{X}_j^{(1)} | \mathbf{x}_j^{(2)}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right) = \widetilde{\boldsymbol{\mu}}^{(1)} + \widetilde{\boldsymbol{\Sigma}}_{12}\widetilde{\boldsymbol{\Sigma}}_{22}^{-1}\left(\mathbf{x}_j^{(2)} - \widetilde{\boldsymbol{\mu}}^{(2)}\right)$$

Use the $\widetilde{\mathbf{x}}_j^{(1)}$ to compute $\widetilde{\mathbf{T}}_1$. Similarly, replace the missing components in $\mathbf{T}_2$ with conditional means:

$$\begin{aligned}\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(1)\prime}} &= E\left(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(1)\prime} | \mathbf{x}_j^{(2)}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right) \\ &= \widetilde{\boldsymbol{\Sigma}}_{11} - \widetilde{\boldsymbol{\Sigma}}_{12}\widetilde{\boldsymbol{\Sigma}}_{22}^{-1}\widetilde{\boldsymbol{\Sigma}}_{21} + \widetilde{\boldsymbol{\mu}}^{(1)}\widetilde{\boldsymbol{\mu}}^{(1)\prime} + 2\widetilde{\boldsymbol{\Sigma}}_{12}\widetilde{\boldsymbol{\Sigma}}_{22}^{-1}\left(\mathbf{x}_j^{(2)} - \widetilde{\boldsymbol{\mu}}^{(2)}\right)\widetilde{\boldsymbol{\mu}}^{(1)\prime}\end{aligned}$$

and

$$\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(2)\prime}} = E\left(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(2)\prime} | \mathbf{x}_j^{(2)}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right) = \widetilde{\boldsymbol{\mu}}^{(1)}\mathbf{x}_j^{(2)\prime}$$

Use the $\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(1)\prime}}$ and $\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(2)\prime}}$ to compute $\widetilde{\mathbf{T}}_2$.

EM for Multivariate Normal Data: In the multivariate normal case, the EM algorithm need only be concerned with the sufficient statistics

The **M Step** then uses the $\widetilde{\mathbf{T}}_1$ and $\widetilde{\mathbf{T}}_2$ to update the parameter estimates:

$$\widetilde{\boldsymbol{\mu}} = \frac{1}{n}\widetilde{\mathbf{T}}_1 \text{ and } \widetilde{\boldsymbol{\Sigma}} = \frac{1}{n}\widetilde{\mathbf{T}}_2 - \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}'$$

Use the $\widetilde{\mathbf{x}}_j^{(1)}$ to compute $\widetilde{\mathbf{T}}_1$. Similarly, replace the missing components in $\mathbf{T}_2$ with conditional means:

$$\begin{aligned}\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(1)\prime}} &= E\left(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(1)\prime}|\mathbf{x}_j^{(2)};\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}}\right)\\ &= \widetilde{\boldsymbol{\Sigma}}_{11} - \widetilde{\boldsymbol{\Sigma}}_{12}\widetilde{\boldsymbol{\Sigma}}_{22}^{-1}\widetilde{\boldsymbol{\Sigma}}_{21} + \widetilde{\boldsymbol{\mu}}^{(1)}\widetilde{\boldsymbol{\mu}}^{(1)\prime} + 2\widetilde{\boldsymbol{\Sigma}}_{12}\widetilde{\boldsymbol{\Sigma}}_{22}^{-1}\left(\mathbf{x}_j^{(2)} - \widetilde{\boldsymbol{\mu}}^{(2)}\right)\widetilde{\boldsymbol{\mu}}^{(1)\prime}\end{aligned}$$

and

$$\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(2)}} = E\left(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(2)}|\mathbf{x}_j^{(2)};\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}}\right) = \widetilde{\boldsymbol{\mu}}^{(1)}\mathbf{x}_j^{(2)}$$

Use the $\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(1)\prime}}$ and $\widetilde{\mathbf{x}_j^{(1)}\mathbf{x}_j^{(2)}}$ to compute $\widetilde{\mathbf{T}}_2$.

Example: Consider the incomplete data set

$$\mathbf{X} = \begin{bmatrix} \cdot & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ \cdot & \cdot & 5 \end{bmatrix}$$

Here, $n = 4$, $p = 3$, and parts of observation vectors $\mathbf{x}_1$ and $\mathbf{x}_4$ are missing. Computing initial values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as sample means and variances involving only the observed values, and defining $\boldsymbol{\theta}' = [\mu_1, \mu_2, \mu_3, \sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}]$:

Iter.	$\widetilde{\mu}_1$	$\widetilde{\mu}_2$	$\widetilde{\mu}_3$	$\widetilde{\sigma}_{11}$	$\widetilde{\sigma}_{22}$	$\widetilde{\sigma}_{33}$	$\widetilde{\sigma}_{12}$	$\widetilde{\sigma}_{13}$	$\widetilde{\sigma}_{23}$
1	6.031818	1.075000	4.000000	0.546715	0.563125	2.500000	-0.046705	0.872727	0.450000
2	6.110489	1.063750	4.000000	0.667781	0.624623	2.500000	-0.007706	0.621681	0.588750
3	6.114341	1.074812	4.000000	0.790618	0.624419	2.500000	-0.099175	0.763612	0.530438
4	6.167215	1.071747	4.000000	0.840933	0.632157	2.500000	-0.117902	0.559821	0.556528
5	6.184901	1.073590	4.000000	0.921787	0.630924	2.500000	-0.178255	0.594827	0.545325
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
192	6.813042	1.073171	4.000000	1.696675	0.631866	2.500000	-1.111584	-0.626092	0.548781
193	6.813046	1.073171	4.000000	1.696678	0.631866	2.500000	-1.111590	-0.626100	0.548781
194	6.813050	1.073171	4.000000	1.696680	0.631866	2.500000	-1.111596	-0.626107	0.548781
195	6.813054	1.073171	4.000000	1.696683	0.631866	2.500000	-1.111602	-0.626115	0.548781
196	6.813057	1.073171	4.000000	1.696685	0.631866	2.500000	-1.111607	-0.626122	0.548781

Example: Consider the incomplete data set

$$\mathbf{X} = \begin{bmatrix} \cdot & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ \cdot & \cdot & 5 \end{bmatrix}$$

Here, $n = 4$, $p = 3$, and parts of observation vectors $\mathbf{x}_1$ and $\mathbf{x}_4$ are missing. Computing initial values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as sample means and variances involving only the observed values, and defining $\boldsymbol{\theta}' = [\mu_1, \mu_2, \mu_3, \sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}]$:

Iter.	$\widetilde{\mu}_1$	$\widetilde{\mu}_2$	$\widetilde{\mu}_3$	$\widetilde{\sigma}_{11}$	$\widetilde{\sigma}_{22}$	$\widetilde{\sigma}_{33}$	$\widetilde{\sigma}_{12}$	$\widetilde{\sigma}_{13}$	$\widetilde{\sigma}_{23}$
1	6.031818	1.075000	4.000000	0.546715	0.563125	2.500000	-0.046705	0.872727	0.450000
2	6.110480	1.063750	4.000000	0.667781	0.624623	2.500000	-0.007706	0.621681	0.588750
3	6.1143								
4	6.1672								
5	6.1849								
⋮	⋮								
192	6.8130								
193	6.813040	1.073171	4.000000	1.696678	0.631866	2.500000	-1.111590	-0.626100	0.548781
194	6.813050	1.073171	4.000000	1.696680	0.631866	2.500000	-1.111596	-0.626107	0.548781
195	6.813054	1.073171	4.000000	1.696683	0.631866	2.500000	-1.111602	-0.626115	0.548781
196	6.813057	1.073171	4.000000	1.696685	0.631866	2.500000	-1.111607	-0.626122	0.548781

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 6.8131 \\ 1.0732 \\ 4.0000 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.6967 & -1.1116 & -0.6261 \\ -1.1116 & 0.6319 & 0.5488 \\ -0.6261 & 0.5488 & 2.5000 \end{bmatrix}$$

Correlation Between Observation Vectors

Correlated Observations: We have assumed thus far that $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ constitutes a random sample. This means that the $\mathbf{X}_j$ are mutually independent draws from the same population distribution. This assumption will not necessarily hold in all situations. For example, if the $\mathbf{X}_j$ are repeat draws on the same individual over time, then there will be correlation between the observation vectors. In such cases, the inferential methods that we have learned in this topic do not apply. Such methods are unable to adequately capture the sampling variability of our data. This will result in hypothesis tests that do not have the claimed Type I error rate and confidence regions / intervals that do not have the claimed coverage probability.