Tests for Multivariate Means

Max Turgeon

STAT 7200–Multivariate Statistics

Objectives

- Construct tests for a single multivariate mean
- Discuss and compare confidence regions and confidence intervals
- Describe connection with Likelihood Ratio Test
- Construct tests for two multivariate means
- Present robust alternatives to these tests

Test for a multivariate mean: Σ known

- Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim N_p(\mu, \Sigma)$ be independent.
- We saw in a previous lecture that

$$\bar{\mathbf{Y}} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right).$$

- This means that

$$n(\bar{\mathbf{Y}} - \mu)^T \Sigma^{-1} (\bar{\mathbf{Y}} - \mu) \sim \chi^2(p).$$

- In particular, if we want to test $H_0 : \mu = \mu_0$ at level α, then we reject the null hypothesis if

$$n(\bar{\mathbf{Y}} - \mu_0)^T \Sigma^{-1} (\bar{\mathbf{Y}} - \mu_0) > \chi^2_\alpha(p).$$

Example i

```r
library(dslabs)
library(tidyverse)

dataset <- filter(gapminder, year == 2012,
                  !is.na(infant_mortality))

dataset <- dataset[,c("infant_mortality",
                      "life_expectancy",
                      "fertility")]
dataset <- as.matrix(dataset)
```

Example ii

```r
dim(dataset)
```

```
## [1] 178   3
```

```r
# Assume we know Sigma
Sigma <- matrix(c(555, -170, 30, -170, 65, -10,
                  30, -10, 2), ncol = 3)

mu_hat <- colMeans(dataset)
mu_hat
```

Example iii

```
## infant_mortality  life_expectancy        fertility
##        25.824157        71.308427         2.868933
```

```r
# Test mu = mu_0
mu_0 <- c(25, 50, 3)
test_statistic <- nrow(dataset) * t(mu_hat - mu_0) %*%
  solve(Sigma) %*% (mu_hat - mu_0)

c(drop(test_statistic), qchisq(0.95, df = 3))
```

```
## [1] 7153.275387    7.814728
```

Example iv

```r
drop(test_statistic) > qchisq(0.95, df = 3)
```

```
## [1] TRUE
```

Test for a multivariate mean: Σ unknown i

- Of course, we rarely (if ever) know Σ, and so we use its MLE

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$$

  or the sample covariance S_n.
- Therefore, to test $H_0 : \mu = \mu_0$ at level α, then we reject the null hypothesis if

$$T^2 = n(\bar{\mathbf{Y}} - \mu_0)^T S_n^{-1}(\bar{\mathbf{Y}} - \mu_0) > c,$$

  for a suitably chosen constant c that depends on α.
  - **Note**: The test statistic T^2 is known as *Hotelling's* T^2.

Test for a multivariate mean: Σ unknown ii

- We will show that (under H_0) T^2 has a simple distribution:

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

- In other words, we reject the null hypothesis at level α if

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

Example (revisited) i

```r
n <- nrow(dataset); p <- ncol(dataset)

# Test mu = mu_0
mu_0 <- c(25, 50, 3)
test_statistic <- n * t(mu_hat - mu_0) %*%
  solve(cov(dataset)) %*% (mu_hat - mu_0)

critical_val <- (n - 1)*p*qf(0.95, df1 = p,
                             df2 = n - p)/(n-p)
```

Example (revisited) ii

```r
c(drop(test_statistic), critical_val)
```

```
## [1] 5121.461370    8.059773
```

```r
drop(test_statistic) > critical_val
```

```
## [1] TRUE
```

Distribution of T^2

We will prove a more general result that we will also be useful for more than one multivariate mean.

Theorem

Let $\mathbf{Y} \sim N_p(0, \Sigma)$, let $mW \sim W_p(m, \Sigma)$, and assume $\mathbf{Y}, W$ are independent. Define

$$T^2 = m\mathbf{Y}^T W^{-1}\mathbf{Y}.$$

Then

$$\frac{m-p+1}{mp} T^2 \sim F(p, m-p+1),$$

where $F(\alpha, \beta)$ denotes the non-central F-distribution with α, β degrees of freedom.

Proof i

- First, if we write $\Sigma = LL^T$, we can replace $\mathbf{Y}$ by $L^{-1}\mathbf{Y}$ and W with $(L^{-1})^T W (L^{-1})$ without changing T^2.
    - In other words, without loss of generality, we can assume $\Sigma = I_p$.
- Now, note that since $\mathbf{Y}$ and W are independent, the conditional distribution of mW given $\mathbf{Y}$ is also $W_p(m, I_p)$.
- Consider $\mathbf{Y}$ a fixed quantity, and let H be an orthogonal matrix whose first column is $\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}$.
    - The other columns can be chosen by finding a basis for the orthogonal complement of $\mathbf{Y}$ and applying Gram-Schmidt to obtain an orthonormal basis.

Proof ii

- Define $V = H^T W H$. Conditional on $\mathbf{Y}$, this is still distributed as $\frac{1}{m} W_p(m, I_p)$.
    - This distribution does not depend on $\mathbf{Y}$, and therefore V and $\mathbf{Y}$ are independent.
- Decompose V as such:

$$\begin{pmatrix} v_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

where v_{11} is a (random) scalar.

Proof iii

- By result A.2.4g of MKB (see supplementary materials), the $(1, 1)$ element of V^{-1} is given by

$$v_{11|2}^{-1} = (v_{11} - V_{12}V_{22}^{-1}V_{21})^{-1}.$$

- Moreover, note that $v_{11|2} \sim \chi^2(m - p + 1)$.
- We now have

$$\begin{aligned}
\frac{1}{m}T^2 &= \mathbf{Y}^T W^{-1}\mathbf{Y} \\
&= (H^T\mathbf{Y})^T(H^T W H)^{-1}(H^T\mathbf{Y}) \\
&= (H^T\mathbf{Y})^T(V)^{-1}(H^T\mathbf{Y}) \\
&= (\mathbf{Y}^T\mathbf{Y})^{1/2} v_{11|2}^{-1}(\mathbf{Y}^T\mathbf{Y})^{1/2} \\
&= (\mathbf{Y}^T\mathbf{Y})/v_{11|2}.
\end{aligned}$$

Proof iv

- In other words, we have expressed $\frac{1}{m}T^2$ as a ratio of independent chi-squares.
- Therefore, we have

$$\begin{aligned}\frac{m-p+1}{mp}T^2 &= \Big((\mathbf{Y}^T\mathbf{Y})/p\Big) \,/\, \Big(v_{11|2}/(m-p+1)\Big)\\ &\sim F(p, m-p+1).\end{aligned}$$

$\square$

Confidence region for μ i

- Analogously to the univariate setting, it may be more informative to look at a *confidence region*:
  - The set of values $\mu_0 \in \mathbb{R}^p$ that are supported by the data, i.e. whose corresponding null hypothesis $H_0 : \mu = \mu_0$ would be rejected at level α.
- Let $c^2 = \frac{(n-1)p}{(n-p)} F_\alpha(p, n-p)$. A $100(1-\alpha)$% confidence region for μ is given by the ellipsoid around $\bar{\mathbf{Y}}$ such that

$$n(\bar{\mathbf{Y}} - \mu)^T S_n^{-1} (\bar{\mathbf{Y}} - \mu) < c^2, \quad \mu \in \mathbb{R}^p.$$

- We can describe the confidence region in terms of the eigendecomposition of S_n: let $\lambda_1 \geq \cdots \geq \lambda_p$ be its eigenvalues, and let $v_1, \ldots, v_p$ be corresponding eigenvectors of unit length.
- The confidence region is the ellipsoid centered around $\bar{\mathbf{Y}}$ with axes

$$\pm c\sqrt{\lambda_i}v_i.$$

- When $p > 2$ we cannot easily plot the confidence regions.
    - Therefore, we first need to project onto an axis or onto the plane.
- **Theorem**: Let $c > 0$ be a constant and A a $p \times p$ positive definite matrix. For a given vector $\mathbf{u} \neq 0$, the projection of the ellipse $\{\mathbf{y}^T A^{-1}\mathbf{y} \leq c^2\}$ onto $\mathbf{u}$ is given by

$$c\frac{\sqrt{\mathbf{u}^T A\mathbf{u}}}{\mathbf{u}^T\mathbf{u}}\mathbf{u}.$$

Visualizing confidence regions when $p > 2$ ii

- If we take $\mathbf{u}$ to be the standard unit vectors, we get confidence *intervals* for each component of μ:

$$LB = \bar{\mathbf{Y}}_j - \sqrt{\frac{(n-1)p}{(n-p)}F_\alpha(p, n-p)(s^2_{jj}/n)}$$

$$UB = \bar{\mathbf{Y}}_j + \sqrt{\frac{(n-1)p}{(n-p)}F_\alpha(p, n-p)(s^2_{jj}/n)}.$$

Example i

```r
n <- nrow(dataset); p <- ncol(dataset)

critical_val <- (n - 1)*p*qf(0.95, df1 = p,
                              df2 = n - p)/(n-p)
sample_cov <- diag(cov(dataset))
```

```r
cbind(mu_hat - sqrt(critical_val*
                      sample_cov/n),
      mu_hat + sqrt(critical_val*
                      sample_cov/n))
```

Example ii

```
##                          [,1]      [,2]
## infant_mortality 20.801776 30.846538
## life_expectancy  69.561973 73.054881
## fertility         2.565608  3.172257
```

- **Theorem**: Let $c > 0$ be a constant and A a $p \times p$ positive definite matrix. For a given pair of perpendicular unit vectors $\mathbf{u}_1, \mathbf{u}_2$, the projection of the ellipse $\{\mathbf{y}^T A^{-1}\mathbf{y} \leq c^2\}$ onto the plane defined by $\mathbf{u}_1, \mathbf{u}_2$ is given by

$$\left\{(U^T\mathbf{y})^T(U^TAU)^{-1}(U^T\mathbf{y}) \leq c^2\right\},$$

where $U = (\mathbf{u}_1, \mathbf{u}_2)$.

```r
U <- matrix(c(1, 0, 0,
              0, 1, 0),
            ncol = 2)
R <- n*solve(t(U) %*% cov(dataset) %*% U)
transf <- chol(R)
```

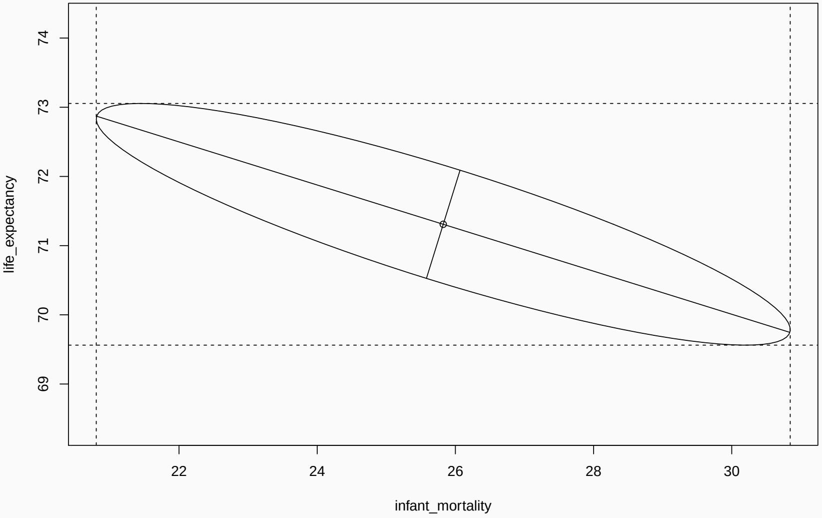
Example (cont'd) ii

```r
# First create a circle of radius c
theta_vect <- seq(0, 2*pi, length.out = 100)
circle <- sqrt(critical_val) * cbind(cos(theta_vect),
                                     sin(theta_vect))
# Then turn into ellipse
ellipse <- circle %*% t(solve(transf)) +
  matrix(mu_hat[1:2], ncol = 2,
         nrow = nrow(circle),
         byrow = TRUE)
```

Example (cont'd) iii

```
# Eigendecomposition
# To visualize the principal axes
decomp <- eigen(t(U) %*% cov(dataset) %*% U)
first <- sqrt(decomp$values[1]) *
  decomp$vectors[,1] * sqrt(critical_val)
second <- sqrt(decomp$values[2]) *
  decomp$vectors[,2] * sqrt(critical_val)
```

Example (cont'd) iv



Simultaneous Confidence Statements i

- Let $w \in \mathbb{R}^p$. We are interested in constructing confidence intervals for $w^T\mu$ that are simultaneously valid (i.e. right coverage probability) for all w.
- Note that $w^T\bar{\mathbf{Y}}$ and w^TS_nw are both scalars.
- If we were only interested in a particular w, we could use the following confidence interval:

$$\left(w^T\bar{\mathbf{Y}} \pm t_{\alpha/2,n-1}\sqrt{w^TS_nw/n}\right).$$

Simultaneous Confidence Statements ii

- Or equivalently, the confidence interval contains the set of values $w^T\mu$ for which

$$t^2(w) = \frac{n(w^T\bar{\mathbf{Y}} - w^T\mu)^2}{w^T S_n w} = \frac{n(w^T(\bar{\mathbf{Y}} - \mu))^2}{w^T S_n w} \leq F_\alpha(1, n-1).$$

- **Strategy**: Maximise over all w:

$$\max_w t^2(w) = \max_w \frac{n(w^T(\bar{\mathbf{Y}} - \mu))^2}{w^T S_n w}.$$

Simultaneous Confidence Statements iii

- Using the Cauchy-Schwarz Inequality:

$$\begin{aligned}
(w^T(\bar{\mathbf{Y}} - \mu))^2 &= (w^T S_n^{1/2} S_n^{-1/2}(\bar{\mathbf{Y}} - \mu))^2 \\
&= ((S_n^{1/2}w)^T(S_n^{-1/2}(\bar{\mathbf{Y}} - \mu)))^2 \\
&\leq (w^T S_n w)((\bar{\mathbf{Y}} - \mu)^T S_n^{-1}(\bar{\mathbf{Y}} - \mu)).
\end{aligned}$$

- Dividing both sides by $w^T S_n w/n$, we get

$$t^2(w) \leq n(\bar{\mathbf{Y}} - \mu)^T S_n^{-1}(\bar{\mathbf{Y}} - \mu).$$

Simultaneous Confidence Statements iv

- Since the Cauchy-Schwarz inequality also implies that the inequality is an *equality* if and only if w is proportional to $S_n^{-1}(\bar{\mathbf{Y}} - \mu)$, it means the upper bound is attained and therefore

$$\max_w t^2(w) = n(\bar{\mathbf{Y}} - \mu)^T S_n^{-1}(\bar{\mathbf{Y}} - \mu).$$

- The right-hand side is Hotteling's T^2, and therefore we know that

$$\max_w t^2(w) \sim \frac{(n-1)p}{(n-p)} F(p, n-p).$$

Simultaneous Confidence Statements V

- **Theorem**: Simultaneously for all $w \in \mathbb{R}^p$, the interval

$$\left(w^T\bar{\mathbf{Y}} \pm \sqrt{\frac{(n-1)p}{n(n-p)}F_\alpha(p, n-p)w^T S_n w}\right).$$

  will contain $w^T\mu$ with probability $1-\alpha$.
- **Corollary**: If we take w to be the standard basis vectors, we recover the projection results from earlier.

Further comments

- If we take $w = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)$, we can also derive confidence statements about mean differences $\mu_i - \mu_k$.
- In general, simultaneous confidence statements are good for exploratory analyses, i.e. when we test many different contrasts.
- However, this much generality comes at a cost: the resulting confidence intervals are quite large.
    - Since we typically only care about a finite number of hypotheses, there are more efficient ways to account for the exploratory nature of the tests.

Bonferroni correction i

- Assume that we are interested in m null hypotheses $H_{0i} : w_i^T \mu = \mu_{0i}$, at confidence level α_i, for $i = 1, \ldots, m$.
- We can show that

$$\begin{aligned} P(\text{none of } H_{0i} \text{ are rejected}) &= 1 - P(\text{some } H_{0i} \text{ is rejected}) \\ &\geq 1 - \sum_{i=1}^{m} P(H_{0i} \text{ is rejected}) \\ &= 1 - \sum_{i=1}^{m} \alpha_i. \end{aligned}$$

Bonferroni correction ii

- Therefore, if we want to control the overall error rate at α, we can take

$$\alpha_i = \alpha/m, \qquad \text{for all } i = 1, \ldots, m.$$

- If we take w_i to be the i-th standard basis vector, we get simultaneous confidence intervals for all p components of μ:

$$\left(\bar{\mathbf{Y}}_i \pm t_{\alpha/2p, n-1}(\sqrt{s_{ii}^2/n})\right).$$

Example i

```r
# Let's focus on only two variables
dataset <- dataset[,c("infant_mortality",
                      "life_expectancy")]

n <- nrow(dataset); p <- ncol(dataset)
```

Example ii

```
alpha <- 0.05
mu_hat <- colMeans(dataset)
sample_cov <- diag(cov(dataset))

# Simultaneous CIs
critical_val <- (n - 1)*p*qf(1-0.5*alpha, df1 = p,
                             df2 = n - p)/(n-p)

simul_ci <- cbind(mu_hat - sqrt(critical_val*
                                  sample_cov/n),
                  mu_hat + sqrt(critical_val*
                                  sample_cov/n))
```

Example iii

```r
# Univariate without correction
univ_ci <- cbind(mu_hat - qt(1-0.5*alpha, n - 1) *
                   sqrt(sample_cov/n),
                 mu_hat + qt(1-0.5*alpha, n - 1) *
                   sqrt(sample_cov/n))
```

```r
# Bonferroni adjustment
bonf_ci <- cbind(mu_hat - qt(1-0.5*alpha/p, n - 1) *
                   sqrt(sample_cov/n),
                 mu_hat + qt(1-0.5*alpha/p, n - 1) *
                   sqrt(sample_cov/n))
```

```r
simul_ci
```

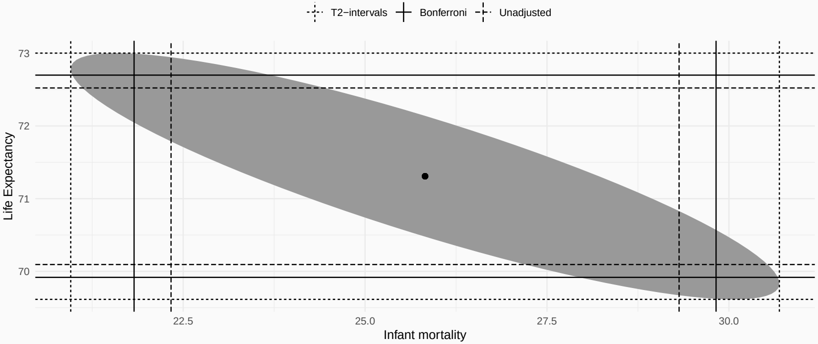
```
##                         [,1]     [,2]
## infant_mortality 20.95439 30.69392
## life_expectancy  69.61504 73.00181
```

```r
univ_ci
```

```
##                         [,1]     [,2]
## infant_mortality 22.33295 29.31537
## life_expectancy  70.09441 72.52244
```

```r
bonf_ci
```

```
##                         [,1]    [,2]
## infant_mortality 21.82491 29.8234
## life_expectancy  69.91775 72.6991
```


T2−intervals
Bonferroni
Unadjusted
73
72
71
70
Life Expectancy
22.5
25.0
27.5
30.0
Infant mortality

Summary of confidence statements

- *So which one should you use?*
  - Use the confidence region when you're interested in a single multivariate hypothesis test.
  - Use the simultaneous (i.e. T^2) intervals when testing a large number of contrasts.
  - Use the Bonferroni correction when testing a small number of contrasts (e.g. each component of μ).
  - (Almost) **never** use the unadjusted intervals.
- We can check the coverage probabilities of each approach using a simulation study:
  - `https://www.maxturgeon.ca/f19-stat4690/simulation_coverage_probability.R`

Likelihood Ratio Test i

- There is another important approach to performing hypothesis testing:
    - **Likelihood Ratio Test**
- General strategy:
    i. Maximise likelihood under the null hypothesis: L_0
    ii. Maximise likelihood over the whole parameter space: L_1
    iii. Since the value of the parameters under the null hypothesis is in the parameter space, we have $L_1 \geq L_0$.
    iv. Reject the null hypothesis if the ratio $\Lambda = L_0/L_1$ is small.

Likelihood Ratio Test ii

- In our setting, recall that the likelihood is given by

$$L(\mu, \Sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{y}_i - \mu)^T \Sigma^{-1} (\mathbf{y}_i - \mu) \right) \right).$$

- Over the whole parameter space, it is maximised at

$$\hat{\mu} = \bar{\mathbf{Y}}, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T.$$

- Under the null hypothesis $H_0 : \mu = \mu_0$, the only free parameter is Σ, and $L(\mu_0, \Sigma)$ is maximised at

$$\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Y}_i - \mu_0)(\mathbf{Y}_i - \mu_0)^T.$$

Likelihood Ratio Test iii

- With some linear algbera, you can check that

$$\begin{aligned} L(\hat{\mu}, \hat{\Sigma}) &= \frac{\exp(-np/2)}{(2\pi)^{np/2}|\hat{\Sigma}|^{n/2}} \\ L(\mu_0, \hat{\Sigma}_0) &= \frac{\exp(-np/2)}{(2\pi)^{np/2}|\hat{\Sigma}_0|^{n/2}}. \end{aligned}$$

- Therefore, the likelihood ratio is given by

$$\Lambda = \frac{L(\mu_0, \hat{\Sigma}_0)}{L(\hat{\mu}, \hat{\Sigma})} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|}\right)^{n/2}.$$

- The equivalent statistic $\Lambda^{2/n} = |\hat{\Sigma}|/|\hat{\Sigma}_0|$ is called *Wilks' lambda*.

Distribution of Wilk's Lambda i

- Let Λ be the Likelihood Ratio Test statistic, and let T^2 be Hotelling's statistic. We have

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1}\right)^{-1}.$$

  - Therefore the two tests are equivalent.
  - But note that $\Lambda^{2/n}$ involves computing two determinants, whereas T^2 involves inverting a matrix.

Proof:

- Write $V = \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$, which allows us to write $\hat{\Sigma} = n^{-1}V$.

- Using a familiar trick, we can write

$$\begin{aligned} n\hat{\Sigma}_0 &= \sum_{i=1}^n (\mathbf{Y}_i - \mu_0)(\mathbf{Y}_i - \mu_0)^T \\ &= \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}} + \bar{\mathbf{Y}} - \mu_0)(\mathbf{Y}_i - \bar{\mathbf{Y}} + \bar{\mathbf{Y}} - \mu_0)^T \\ &= V + n(\bar{\mathbf{Y}} - \mu_0)(\bar{\mathbf{Y}} - \mu_0)^T. \end{aligned}$$

Distribution of Wilk's Lambda iii

- We can now write

$$
\begin{aligned}
\frac{|n\hat{\Sigma}_0|}{|n\hat{\Sigma}|} &= \frac{|V + n(\bar{\mathbf{Y}} - \mu_0)(\bar{\mathbf{Y}} - \mu_0)^T|}{|V|} \\
&= |I_p + nV^{-1}(\bar{\mathbf{Y}} - \mu_0)(\bar{\mathbf{Y}} - \mu_0)^T| \\
&= (1 + n(\bar{\mathbf{Y}} - \mu_0)^T V^{-1}(\bar{\mathbf{Y}} - \mu_0)) \\
&= \left(1 + \frac{n}{n-1}(\bar{\mathbf{Y}} - \mu_0)^T S_n^{-1}(\bar{\mathbf{Y}} - \mu_0)\right) \\
&= \left(1 + \frac{T^2}{n-1}\right),
\end{aligned}
$$

where the third equality follows from Problem 1 of Assignment 1.

$\square$

Comparing two multivariate means

Equal covariance case i

- Now let's assume we have *two* independent multivariate samples of (potentially) different sizes:
    - $\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1} \sim N_p(\mu_1, \Sigma)$
    - $\mathbf{Y}_{21}, \ldots, \mathbf{Y}_{2n_2} \sim N_p(\mu_2, \Sigma)$
- We are interested in testing $\mu_1 = \mu_2$.
    - Note that we assume *equal covariance* for the time being.
- Let $\bar{\mathbf{Y}}_1, \bar{\mathbf{Y}}_2$ be their respective sample means, and let S_1, S_2, their respective sample covariances.
- First, note that

$$\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2 \sim N_p\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right).$$

- Second, we also have that $(n_i - 1)S_i$ is an estimator for $(n_i - 1)\Sigma$, for $i = 1, 2$.
    - Therefore, we can *pool* both $(n_1 - 1)S_1$ and $(n_2 - 1)S_2$ into a single estimator for Σ:

$$S_{pool} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2},$$

    where $(n_1 + n_2 - 2)S_{pool} \sim W_p(n_1 + n_2 - 2, \Sigma)$.

- Putting these two observations together, we get a test statistic for $H_0 : \mu_1 = \mu_2$:

$$T^2 = (\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pool}\right]^{-1} (\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2).$$

Equal covariance case iii

- Using our theorem, we can that conclude that under the null hypothesis, we get

$$T^2 \sim \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F(p, n_1 + n_2 - p - 1).$$

Example i

```
dataset1 <- filter(gapminder, year == 2012,
                   continent == "Africa",
                   !is.na(infant_mortality))

dataset1 <- dataset1[,c("life_expectancy",
                        "infant_mortality")]
dataset1 <- as.matrix(dataset1)
dim(dataset1)
```

```
## [1] 51  2
```

Example ii

```r
dataset2 <- filter(gapminder, year == 2012,
                   continent == "Asia",
                   !is.na(infant_mortality))

dataset2 <- dataset2[,c("life_expectancy",
                        "infant_mortality")]
dataset2 <- as.matrix(dataset2)
dim(dataset2)
```

```
## [1] 45  2
```

Example iii

```r
n1 <- nrow(dataset1); n2 <- nrow(dataset2)
p <- ncol(dataset1)
```

```r
(mu_hat1 <- colMeans(dataset1))
```

```
##  life_expectancy infant_mortality
##         62.14314         52.32745
```

```r
(mu_hat2 <- colMeans(dataset2))
```

Example iv

```
##  life_expectancy infant_mortality
##         73.76667         20.84000
```

```r
(S1 <- cov(dataset1))
```

```
##                  life_expectancy infant_mortality
## life_expectancy          48.7241        -107.1926
## infant_mortality       -107.1926         504.2972
```

```r
(S2 <- cov(dataset2))
```

Example v

```
##                    life_expectancy infant_mortality
## life_expectancy          26.08727        -65.19568
## infant_mortality        -65.19568        256.40655
```

```
# Even though it doesn't look reasonable
# We will assume equal covariance for now
```

Example vi

```
mu_hat_diff <- mu_hat1 - mu_hat2

S_pool <- ((n1 - 1)*S1 + (n2 - 1)*S2)/(n1+n2-2)

test_statistic <- t(mu_hat_diff) %*%
  solve((n1^-1 + n2^-1)*S_pool) %*% mu_hat_diff

const <- (n1 + n2 - 2)*p/(n1 + n2 - p - 2)
critical_val <- const * qf(0.95, df1 = p,
                           df2 = n1 + n2 - p - 2)
```

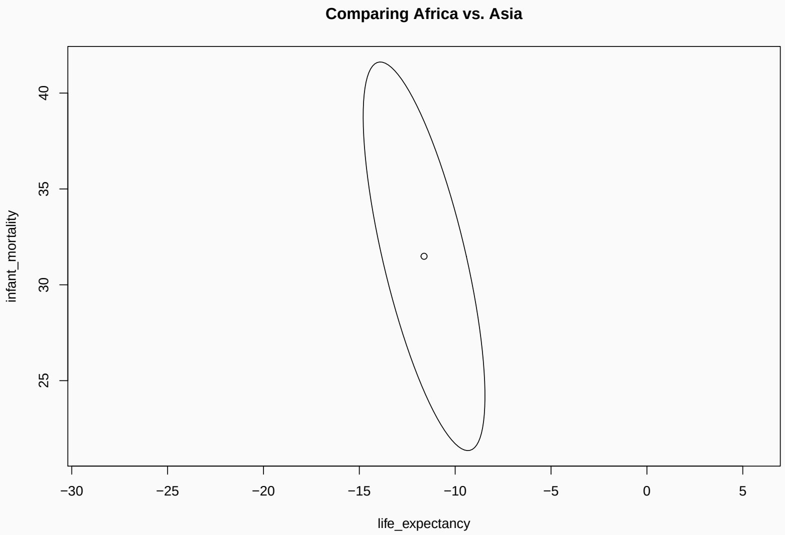
Example vii

```r
c(drop(test_statistic), critical_val)
```

```
## [1] 87.65479  6.32545
```

```r
drop(test_statistic) > critical_val
```

```
## [1] TRUE
```


Comparing Africa vs. Asia
infant_mortality
40
35
30
25
−30
−25
−20
−15
−10
−5
0
5
life_expectancy

Unequal covariance case i

- Now let's turn our attention to the case where the covariance matrices are **not** equal:
  - $\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1} \sim N_p(\mu_1, \Sigma_1)$
  - $\mathbf{Y}_{21}, \ldots, \mathbf{Y}_{2n_2} \sim N_p(\mu_2, \Sigma_2)$
- Recall that in the univariate case, the test statistic that is typically used is called *Welch's t-statistic*.
  - The general idea is to adjust the degrees of freedom of the t-distribution.
  - **Note**: This is actually the default test used by `t.test`!
- Unfortunately, there is no single best approximation in the multivariate case.

Unequal covariance case ii

- First, observe that we have

$$\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2 \sim N_p\left(\mu_1 - \mu_2, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right).$$

- Therefore, under $H_0 : \mu_1 = \mu_2$, we have

$$(\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2)^T \left(\frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right)^{-1} (\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2) \sim \chi^2(p).$$

- Since S_i converges to Σ_i as $n_i \to \infty$, we can use Slutsky's theorem to argue that if both $n_1 - p$ and $n_2 - p$ are "large", then

$$T^2 = (\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2)^T \left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)^{-1} (\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2) \approx \chi^2(p).$$

Unequal covariance case iii

- Unfortunately, the definition of “large” in this case depends on how different Σ_1 and Σ_2 are.
- Alternatives:
    - Use one of the many approximations to the null distribution of T^2 (e.g. see Timm (2002), Section 3.9; Rencher (1998), Section 3.9.2).
    - Use a resampling technique (e.g. bootstrap or permutation test).
    - Use Welch’s t-statistic for each component of $\mu_1 - \mu_2$ with a Bonferroni correction for the significance level.

Nel & van der Merwe Approximation

- First, define

$$W_i = \frac{1}{n_i} S_i \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1}.$$

- Then let

$$\nu = \frac{p + p^2}{\sum_{i=1}^{2} \frac{1}{n_i} \left(\mathrm{tr}(W_i^2) + \mathrm{tr}(W_i)^2 \right)}.$$

- One can show that $\min(n_1, n_2) \leq \nu \leq n_1 + n_2$.
- Under the null hypothesis, we approximately have

$$T^2 \approx \frac{\nu p}{\nu - p + 1} F(p, \nu - p + 1).$$

Example (cont'd) i

```r
test_statistic <- t(mu_hat_diff) %*%
  solve(n1^-1*S1 + n2^-1*S2) %*% mu_hat_diff

critical_val <- qchisq(0.95, df = p)

c(drop(test_statistic), critical_val)
```

```
## [1] 90.884961  5.991465
```

```r
drop(test_statistic) > critical_val
```

Example (cont'd) ii

```
## [1] TRUE
```

```r
W1 <- S1 %*% solve(n1^-1*S1 + n2^-1*S2)/n1
W2 <- S2 %*% solve(n1^-1*S1 + n2^-1*S2)/n2

trace_square <- sum(diag(W1%*%W1))/n1 +
  sum(diag(W2%*%W2))/n2
square_trace <- sum(diag(W1))^2/n1 +
  sum(diag(W2))^2/n2

(nu <- (p + p^2)/(trace_square + square_trace))
```

Example (cont'd) iii

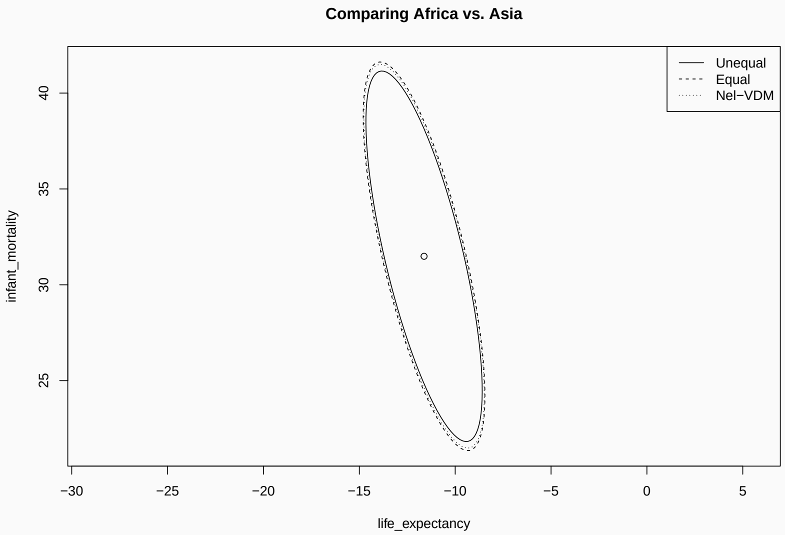
```
## [1] 88.85241
```

```r
const <- nu*p/(nu - p - 1)
critical_val <- const * qf(0.95, df1 = p,
                           df2 = nu - p - 1)

c(drop(test_statistic), critical_val)
```

```
## [1] 90.884961  6.422322
```

```r
drop(test_statistic) > critical_val
```

```
## [1] TRUE
```


Comparing Africa vs. Asia
Unequal
Equal
Nel−VDM
infant_mortality
40
35
30
25
−30
−25
−20
−15
−10
−5
0
5
life_expectancy

Robustness

- To perform the tests on means, we made two main assumptions (listed in order of **importance**):
    1. Independence of the observations;
    2. Normality of the observations.
- Independence is the most important assumption:
    - Departure from independence can introduce significant bias and will impact the coverage probability.
- Normality is not as important:
    - Both tests for one or two means are relatively robust to heavy tail distributions.
    - Test for one mean can be sensitive to skewed distributions; test for two means is more robust.

Simulation i

```r
library(mvtnorm)
set.seed(7200)

n <- 50; p <- 10
B <- 1000

# Simulate under the null
mu <- mu_0 <- rep(0, p)
# Cov: diag = 1; off-diag = 0.5
Sigma <- matrix(0.5, ncol = p, nrow = p)
diag(Sigma) <- 1
```

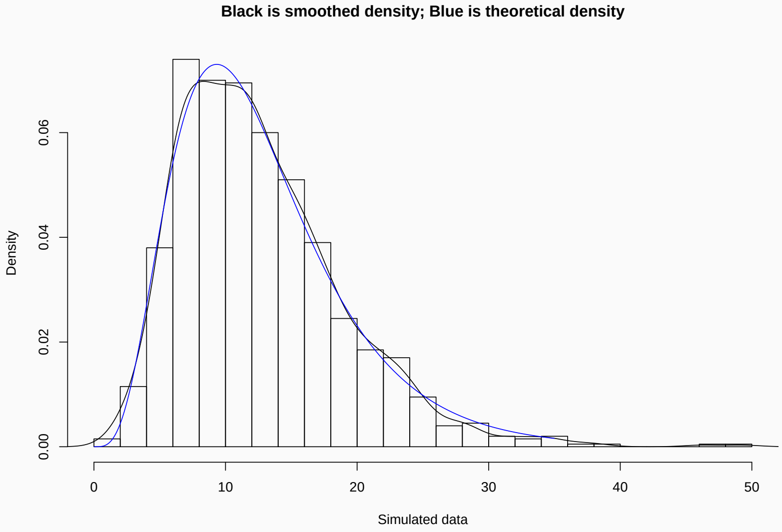
Simulation ii

```r
critical_val <- (n - 1)*p*qf(0.95, df1 = p,
                               df2 = n - p)/(n-p)

null_dist <- replicate(B, {
  Y_norm <- rmvnorm(n, mean = mu, sigma = Sigma)
  mu_hat <- colMeans(Y_norm)
  # Test mu = mu_0
  test_statistic <- n * t(mu_hat - mu_0) %*%
    solve(cov(Y_norm)) %*% (mu_hat - mu_0)
})
```

```r
# Type I error
mean(null_dist > critical_val)
```

```
## [1] 0.035
```

Simulation iv



Simulation v

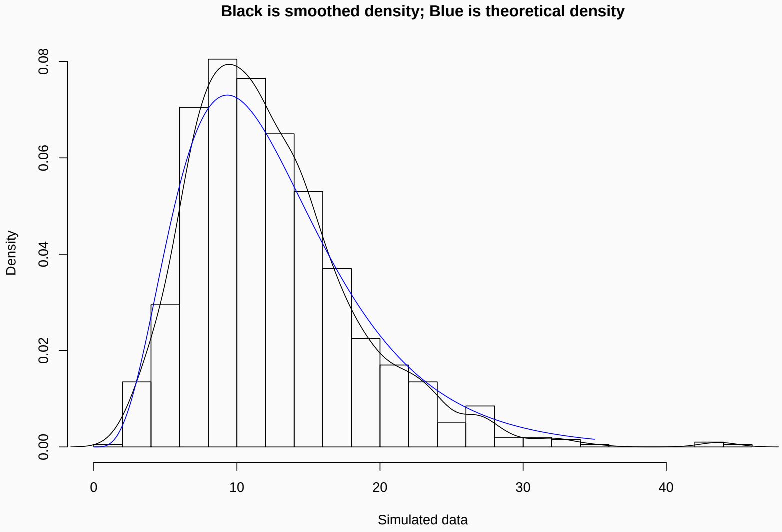
```r
# Now the t distribution
nu <- 3

null_dist_t <- replicate(B, {
  Y_t <- rmvt(n, sigma = Sigma, df = nu, delta = mu)
  mu_hat <- colMeans(Y_t)
  # Test mu = mu_0
  test_statistic <- n * t(mu_hat - mu_0) %*%
    solve(cov(Y_t)) %*% (mu_hat - mu_0)
})
```

Simulation vi

```r
# Type I error
mean(null_dist_t > critical_val)
```

```
## [1] 0.032
```

Simulation vii



Simulation viii

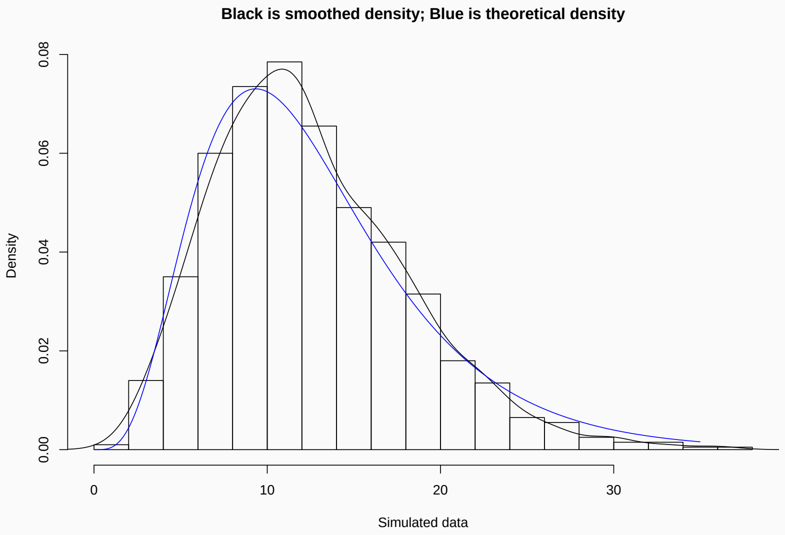
```r
# Now a contaminated normal
sigma <- 3; epsilon <- 0.25
null_dist_cont <- replicate(B, {
  Z <- rmvnorm(n, sigma = diag(p))
  Y <- sample(c(sigma, 1), size = n, replace = TRUE,
              prob = c(epsilon, 1 - epsilon))*Z
  mu_hat <- colMeans(Y)
  # Test mu = mu_0
  test_statistic <- n * t(mu_hat - mu_0) %*%
    solve(cov(Y)) %*% (mu_hat - mu_0)
})
```

Simulation ix

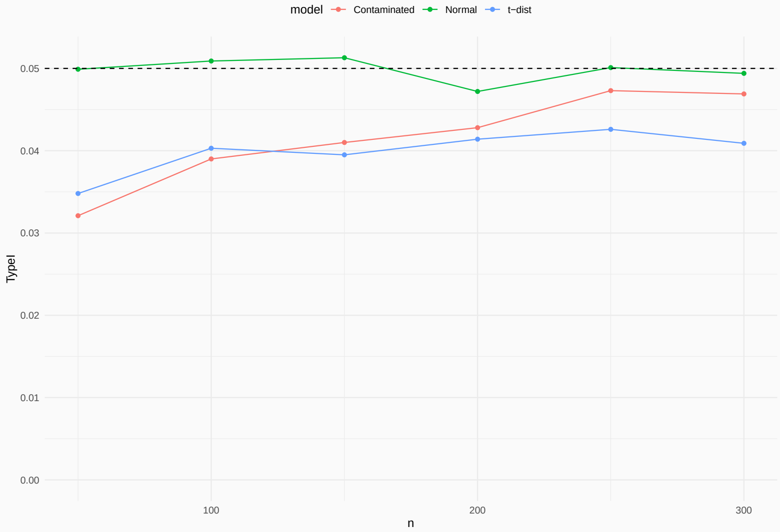
```r
# Type I error
mean(null_dist_cont > critical_val)
```

```
## [1] 0.025
```

Simulation x



Simulation xi



Robust T^2 test statistic

- One potential solution:
    - Fix the distribution, and derive an approximation of the null distribution.
- However, you could potentially get a different approximation for each distribution, and it is not clear which one to use for a given dataset.
- A different solution:
    - Replace the sample mean and sample covariance with **robust estimates** and derive an approximation under general assumptions.
- Generally valid for a large class of distributions, but it will typically at a cost of lower efficiency (i.e. lower power).

Minimum Covariance Determinant Estimator i

- This is a robust estimator of the mean and the covariance introduced by Rousseeuw (JASA, 1984).
  - *Robustness* can mean many things; in this setting, it means that the estimators are stable in the presence of outliers.
- It is defined as follows:
  - Let h be an integer between n (i.e. the sample size) and $\lfloor (n + p + 1)/2 \rfloor$ (where p is the number of variables).
  - Let $\bar{\mathbf{Y}}_{MCD}$ be the mean of the h observations for which the determinant of the sample covariance matrix is minimised.
  - Let S_{MCD} be the corresponding sample covariance scaled by a constant C.

- It can be shown that the smaller h, the more robust $(\bar{\mathbf{Y}}_{MCD}, S_{MCD})$.
- However, there is cost in efficiency. This is can be counterbalanced by *reweighting* the estimators:
    - Let $d_i^2 = (\mathbf{Y} - \bar{\mathbf{Y}}_{MCD})^T S_{MCD}^{-1} (\mathbf{Y} - \bar{\mathbf{Y}}_{MCD})$ be the Mahalanobis distances under the original MCD estimate.
    - Define a weighting function $W(d^2) = I(d^2 \leq \chi^2_{0.975}(p))$.
    - Compute the reweighted MCD estimates:

$$\bar{\mathbf{Y}}_R = \frac{\sum_{i=1}^n W(d_i^2)\mathbf{Y}_i}{\sum_{i=1}^n W(d_i^2)}$$

$$S_R = C\frac{\sum_{i=1}^n W(d_i^2)(\mathbf{Y}_i - \bar{\mathbf{Y}}_R)(\mathbf{Y}_i - \bar{\mathbf{Y}}_R)^T}{\sum_{i=1}^n W(d_i^2)}.$$

Minimum Covariance Determinant Estimator iii

- This reweighted estimator $(\bar{\mathbf{Y}}_R, S_R)$ has the same robustness properties as $(\bar{\mathbf{Y}}_{MCD}, S_{MCD})$, but with higher efficiency.
    - This makes sense, as we are generally including more data points when reweighting, but still controlling for outliers.

Example i

```
# Recall our dataset
dataset <- filter(gapminder, year == 2012,
                  !is.na(infant_mortality))

dataset <- dataset[,c("infant_mortality",
                      "life_expectancy")]
dataset <- as.matrix(dataset)
```

```
# The sample estimators
colMeans(dataset)
```

Example ii

```
## infant_mortality  life_expectancy
##         25.82416         71.30843
```

```
cov(dataset)
```

```
##                  infant_mortality life_expectancy
## infant_mortality         557.0787      -168.81173
## life_expectancy         -168.8117        67.36145
```

Example iii

```r
# The MCD estimators
library(rrcov)

mcd <- CovMcd(dataset)
getCenter(mcd)
```

```
## infant_mortality  life_expectancy
##         11.42203         75.90424
```

```r
getCov(mcd)
```

Example iv

```
##                    infant_mortality life_expectancy
## infant_mortality         132.91885       -60.71957
## life_expectancy          -60.71957        45.54039
```

Robust T^2 test statistic i

- To get a robust T^2 statistic, we can simply replace the sample estimates by the (reweighted) MCD estimates:

$$T^2_{MCD} = n(\mathbf{Y} - \bar{\mathbf{Y}}_R)^T S_R^{-1}(\mathbf{Y} - \bar{\mathbf{Y}}_R).$$

- Unfortunately, the finite-sample properties of $(\bar{\mathbf{Y}}_R, S_R)$ are unknown. BUT:
    - There exists a constant κ such that $\bar{\mathbf{Y}}_R \approx N_p\left(\mu, \frac{\kappa}{n}\Sigma\right)$.
    - There exist constants c, m such that $mc^{-1}S_R \approx W_p(m, \Sigma)$ and $E(S_R) = c\Sigma$.
    - $\bar{\mathbf{Y}}_R$ and S_R are independent.

Robust T^2 test statistic ii

- Putting all of these together, we can deduce that

$$T^2_{MCD} \approx \kappa c^{-1} \frac{mp}{m-p+1} F(p, m-p+1).$$

- The constants κ, m, c can be estimated (Hardin & Rocke, 2005).
- Alternatively, the null distribution of T^2_{MCD} can be estimated using resampling techniques (Willems *et al*, 2002).

Robust T^2 test statistic iii

Algorithm (Willems *et al*, 2002)

1. Rewrite the approximation with only two parameters: $T^2_{MCD} \approx dF(p, q)$.
2. Compute the theoretical mean and variance of $dF(p, q)$ as a function of d, q, p.
3. For several values of n, p, generate multivariate normal variates $N_p(0, I_p)$ and compute T^2_{MCD}.
4. Compute the sample mean and variance of T^2_{MCD}, and use the method of moments to estimate d, q.

Example (cont'd) i

```r
n <- nrow(dataset); p <- ncol(dataset)

# Classical T2
mu_0 <- c(25, 70)
test_statistic <- n * t(mu_hat - mu_0) %*%
  solve(cov(dataset)) %*% (mu_hat - mu_0)

critical_val <- (n - 1)*p*qf(0.95, df1 = p,
                             df2 = n - p)/(n-p)
```

Example (cont'd) ii

```r
c(drop(test_statistic), critical_val)
```

```
## [1] 26.883440  6.129242
```

```r
drop(test_statistic) > critical_val
```

```
## [1] TRUE
```

```r
# Robust T2
t2_robust <- T2.test(dataset, mu = mu_0, method = "mcd")
t2_robust
```

Example (cont'd) iii

```
##
##  One-sample Hotelling test (Reweighted MCD Location)
##
## data:  dataset
## T2 = 42.678, F = 18.000, df1 = 2, df2 = 178, p-value =
## alternative hypothesis: true mean vector is not equal
##
## sample estimates:
##              infant_mortality life_expectancy
## MCD x-vector         16.97192        73.82329
```

Example (cont'd) iv

```r
t2_robust$p.value
```

```
## [1] 7.597764e-08
```

Summary

- We looked at Hotelling's T^2 statistic for tests of one or two multivariate means.
- We described the link between T^2 and the LRT test statistic.
- We discussed confidence regions, simultaneous confidence intervals, and Bonferroni correction.
- We looked at a robust version of T^2 and how to estimate its null distribution.