Tests for Multivariate Means

Max Turgeon

STAT 4690–Applied Multivariate Analysis

Review of univariate tests i

- Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be independently distributed, and let $\bar{X}$ and s^2 be the sample mean and variance, respectively.
- **When σ^2 is known**
  - $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, or equivalently $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1)$.
  - $100(1-\alpha)$% confidence interval: $(\bar{X} - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{X} + z_{\alpha/2}(\sigma/\sqrt{n}))$.
- **When σ^2 is unknown**
  - $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$, or equivalently $\left(\frac{\bar{X}-\mu}{s/\sqrt{n}}\right)^2 \sim F(1, n-1)$.
  - $100(1-\alpha)$% confidence interval: $(\bar{X} - t_{\alpha/2,n-1}(s/\sqrt{n}), \bar{X} + t_{\alpha/2,n-1}(s/\sqrt{n}))$.

Review of univariate tests ii

- In particular, if we want to test $H_0 : \mu = \mu_0$ when σ^2 is unknown, then we reject the null hypothesis if

$$\left|\frac{\bar{X} - \mu_0}{s/\sqrt{n}}\right| > t_{\alpha/2,n-1}, \text{ or } \left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}}\right)^2 > F_\alpha(1, n-1).$$

The multivariate tests for a single mean vector have direct analogues.

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 the 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

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

Example i

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

dataset <- gapminder %>%
  filter(year == 2012,
         !is.na(infant_mortality)) %>%
  select(infant_mortality,
         life_expectancy,
         fertility) %>%
  as.matrix()
```

Example ii

```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
```

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

Example iii

```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)

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

- It turns out 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)

```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)

drop(test_statistic) > critical_val
```

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

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

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

Confidence region for μ ii

- 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

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

Visualizing confidence regions when $p > 2$ 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:

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

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

Example

```r
sample_cov <- diag(cov(dataset))

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

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

Visualizing confidence regions when $p > 2$ (cont'd) i

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

Example (cont'd) i

```
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

```
# First create a circle of radius c
theta_vect <- seq(0, 2*pi, length.out = 100)
circle <- sqrt(critical_val) * cbind(cos(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

```r
# Eigendecomposition
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