ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Extending univariate methods

Applied Multivariate Statistics – Spring 2013

Overview

- Multivariate t-test (one sample, two samples)
- MANOVA
- Multivariate Linear Regression

Revision: One-sample z-Test

1. Model: $X_1, \dots, X_n \sim N(\mu, \sigma_X^2)\ iid, \sigma_X$ known
2. Hypotheses: $H_0: \mu = \mu_0,\ H_A: \mu \neq \mu_0$
3. Test statistics:

$$T = \frac{\overline{X}_n - \mu_0}{\sigma_{\overline{X}_n}} = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_X}$$

   If H_0 is true: $\overline{X}_n \sim N(\mu_0, \frac{\sigma_X^2}{n})$ and thus $T \sim N(0,1)$

4. Make observation of test statistics: t
5. Compute p-value: Probability of seeing something as extreme as t or even more extreme than t if H_0 is true:

$$P(|T| > |t|)$$

Revision: One-sample t-Test

1. Model: $X_1, \ldots, X_n \sim N(\mu, \sigma_X^2)\ iid, \sigma_X$ unknown
2. Hypotheses: $H_0: \mu = \mu_0,\ H_A: \mu \neq \mu_0$
3. Test statistics:

$$T = \frac{\overline{X}_n - \mu_0}{\hat{\sigma}_{\overline{X}_n}} = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\hat{\sigma}_X}$$

   If H_0 is true: $\overline{X}_n \sim N(\mu_0, \frac{\sigma_X^2}{n})$ and thus $T \sim t_{n-1}$

4. Make observation of test statistics: t
5. Compute p-value: Probability of seeing something as extreme as t or even more extreme than t if H_0 is true:

$$P(|T| > |t|)$$

Hotelling's one-sample T-Test: Σ known

1. Model: $X_1, \ldots, X_n \sim N(\mu, \Sigma)\ iid, \Sigma$ known; p dimensions
2. Hypotheses: $H_0: \mu = \mu_0,\ H_A: \mu \neq \mu_0$
3. Test statistics:

$$T = n\boxed{(\overline{X}_n - \mu_0)^T \Sigma^{-1} (\overline{X}_n - \mu_0)}$$

Squared Mahalanobis Distance between sample mean and μ_0

If H_0 is true: $T \sim \chi^2_p$

4. Make observation of test statistics: t
5. Compute p-value: Probability of seeing something as extreme as t or even more extreme than t if H_0 is true:

$$P(|T| > |t|)$$

Hotelling's one-sample T-Test: Σ unknown

1. Model: $X_1, \ldots, X_n \sim N(\mu, \Sigma)\ iid, \Sigma$ unknown; p dimensions
2. Hypotheses: $H_0: \mu = \mu_0,\ H_A: \mu \neq \mu_0$
3. Test statistics:

$$T = n\boxed{(\overline{X}_n - \mu_0)^T S^{-1} (\overline{X}_n - \mu_0)}$$

Estimated Sq. Mahalanobis Distance between sample mean and μ_0

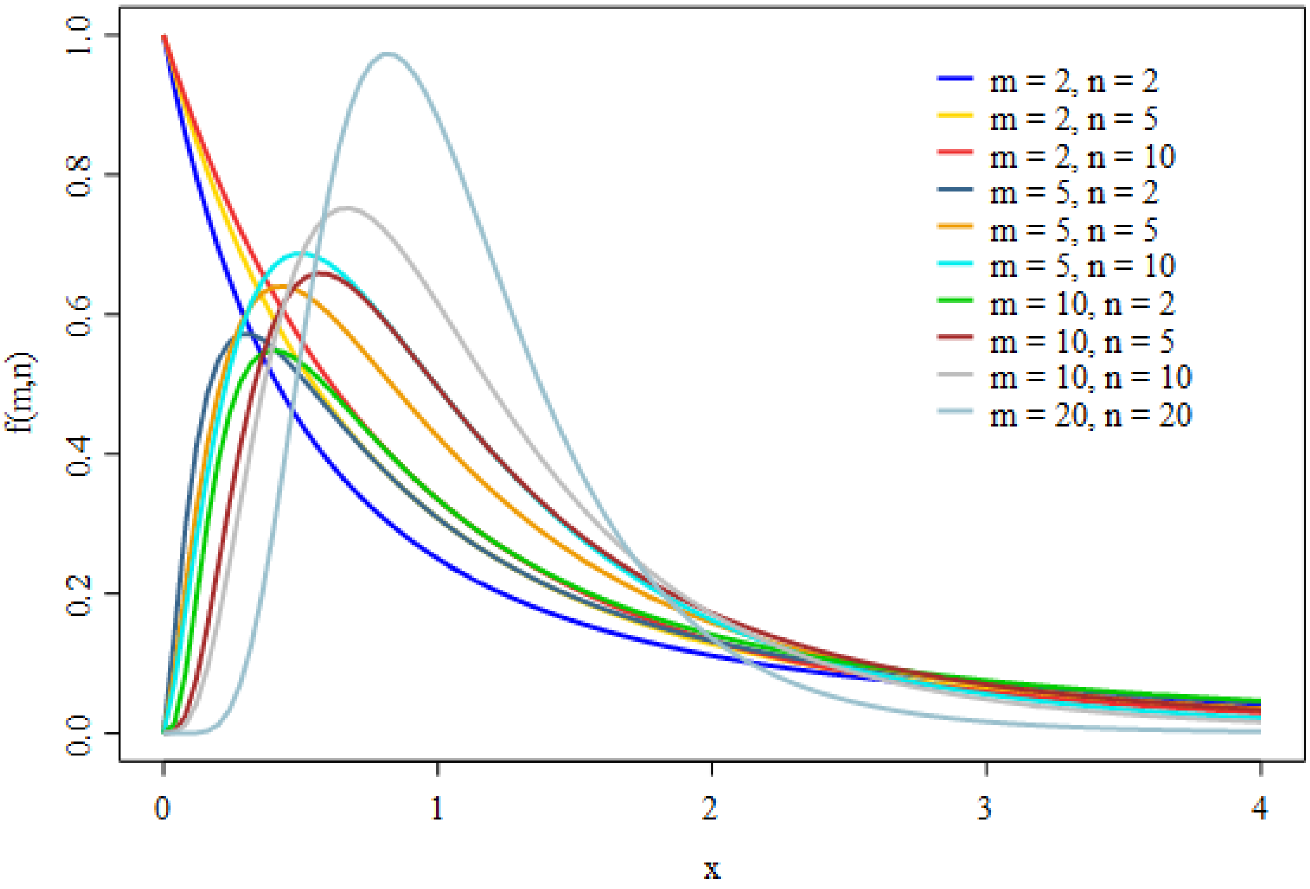
If H_0 is true: $T \sim F_{p,n-p}$

4. Make observation of test statistics: t
5. Compute p-value: Probability of seeing something as extreme as t or even more extreme than t if H_0 is true:

$$P(|T| > |t|)$$

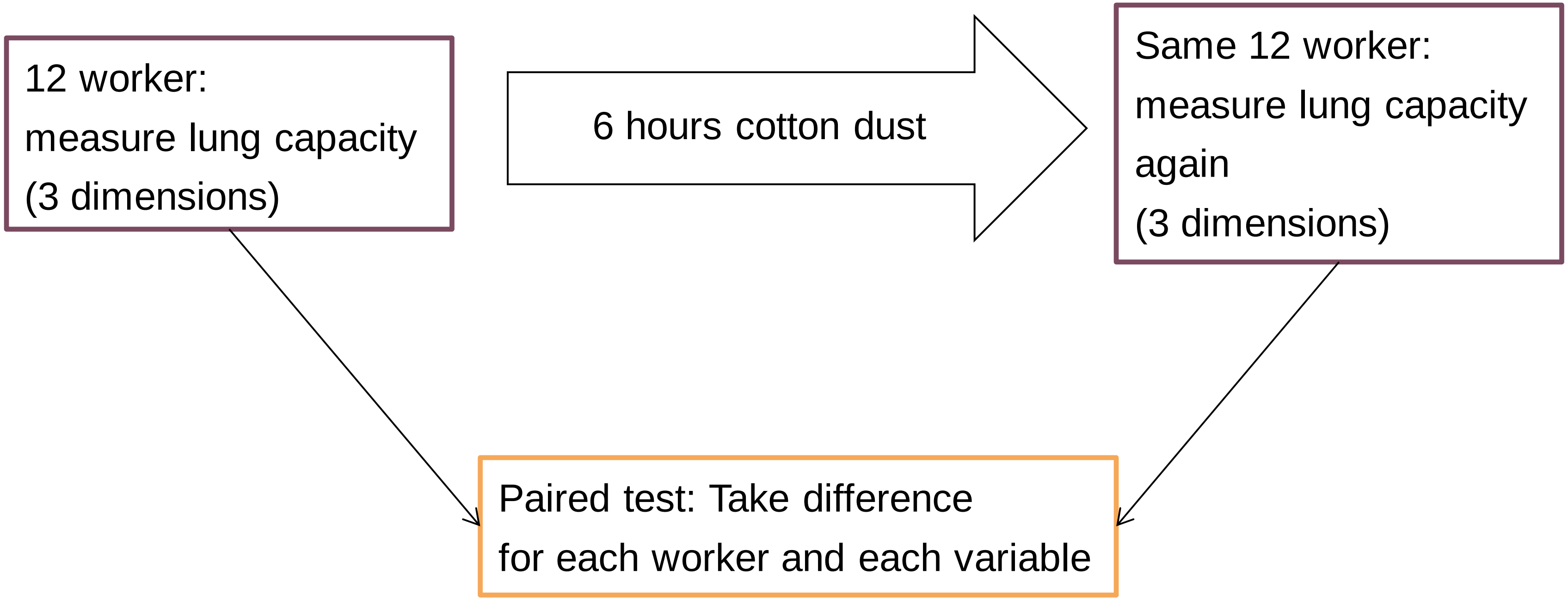
R: Function "HotellingsT2" in package "ICSNP"

F distribution



$$F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n}$$

Example: Change in Pulmonary Response after Exposure to Cotton Dust



Revision: Two-sample t-Test

1. Model: $X_1, \dots, X_n \sim N\left(\mu_X, {\sigma_X}^2\right)\ iid, \sigma_X$ unknown
   $$Y_1, \dots, Y_m \sim N\left(\mu_Y, {\sigma_X}^2\right)\ iid$$
2. Hypotheses: $H_0: \mu_X = \mu_Y,\ H_A: \mu_X \neq \mu_Y$
3. Test statistics:
   $$T = \frac{(\overline{X}_n - \overline{Y}_n) - (\mu_X - \mu_Y)}{\hat{\sigma}_{\overline{X}_n - \overline{Y}_n}}$$
   If H_0 is true: $T \sim t_{n+m-2}$
4. Make observation of test statistics: t
5. Compute p-value: Probability of seeing something as extreme as t or even more extreme than t if H_0 is true:
   $$P(|T| > |t|)$$

Can be extended to

$\sigma_X \neq \sigma_Y$

Hotelling's Two-Sample T-Test: Σ unkown, but equal in both groups

1. Model: $X_1, \ldots, X_n \sim MVN(\mu_X, \Sigma)\ iid, \Sigma$ unknown, p dims.
   $Y_1, \ldots, Y_m \sim MVN(\mu_Y, \Sigma)\ iid$
2. Hypotheses: $H_0 : \mu_X = \mu_Y,\ H_A : \mu_X \neq \mu_Y$
3. Test statistics:

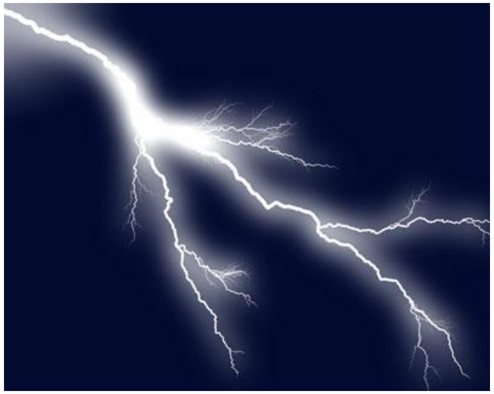
   $$T = \frac{(n+m-p-1)nm}{(n+m)(n+m-2)p}(\overline{X}_n - \overline{Y}_n)^T S^{-1}(\overline{X}_n - \overline{Y}_n)$$

   If H_0 is true: $T \sim F_{p,n+m-p-1}$
4. Make observation of test statistics: t
5. Compute p-value: Probability of seeing something as extreme as t or even more extreme than t if H_0 is true:

   $$P(|T| > |t|)$$

R: Function "HotellingsT2" in package "ICSNP"

Example: Quality control for screws



20 screws:
- winding [mm]
- length [mm]
- diameter [mm]

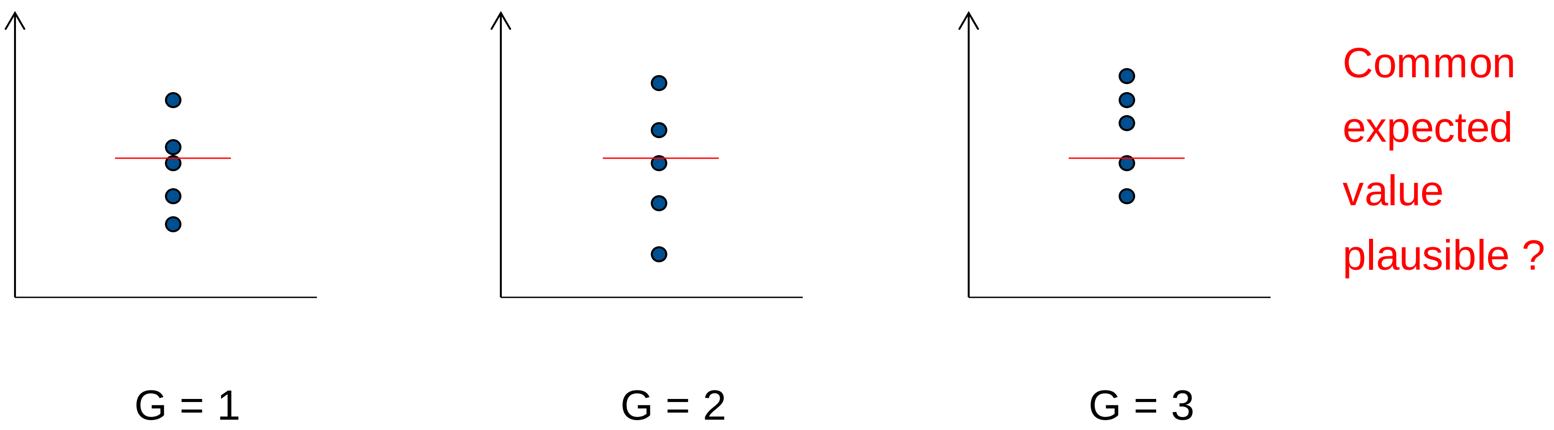
Plant struck by lightning:

Machines still adjusted correctly?

15 screws:
- winding
- length
- diameter

Revision: One-way ANOVA

- Are the expected values in three groups the same?

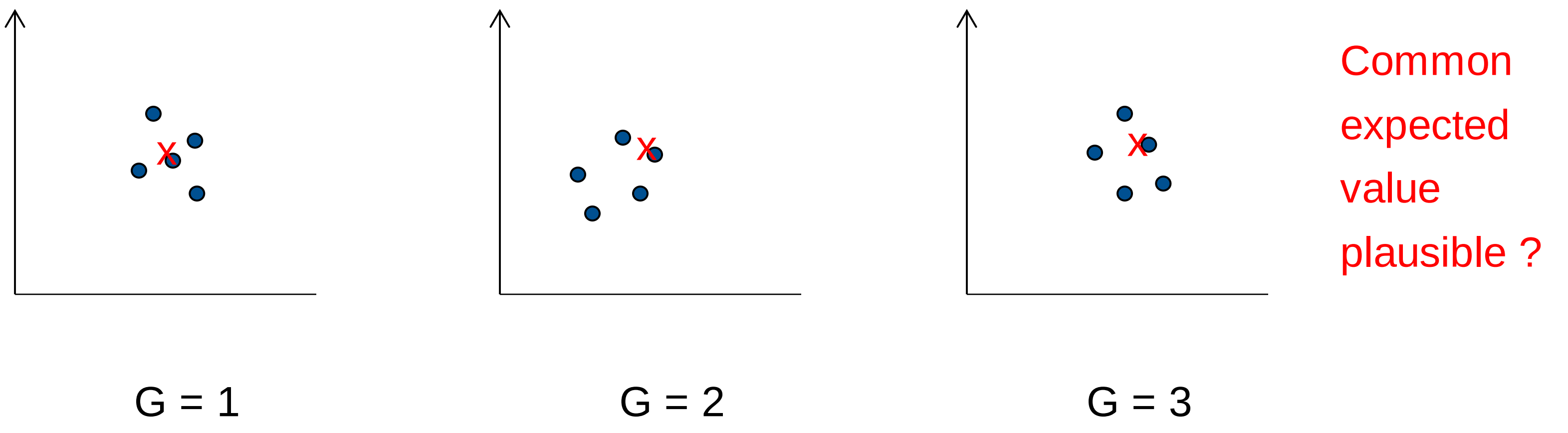


- ANOVA:
  - Compare variation within groups and between groups
  - Assume normality

➡ p-Values can be computed

MANOVA

- Are the multi-dimensional expected values in three groups the same?



- MANOVA:
  - Compare within groups and between groups covariance matrices (test statistics based on eigenvalues)
  - Assume normality

  Wilks test: p-Values can be computed
- R: Function “manova” and “summary(…, test = “Wilks”)”

Revision: Univariate (Multiple) Linear Regression

- N samples, p predictors, 1 response
- Univariate Linear Regression model:

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon = f(X) + \epsilon$$

For N samples using matrix notation:

$$Y = X\beta + E$$

where

Y: N*1 matrix, X: N*(p+1), β: (p+1)*1, E: N*1

- Criterion to optimize: $RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$

- Solution: $\hat{\beta} = (X^T X)^{-1} X^T Y$

Multivariate (Multiple) Linear Regression

- N samples, p predictors, K responses
- Univariate Linear Regression model for each response:

  $$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \epsilon_k = f_k(X) + \epsilon_k$$

  $Cov(\epsilon) = \Sigma$, errors between responses can be correlated

  For N samples using matrix notation:

  $$Y = XB + E$$

  where

  Y: N*K matrix, X: N*(p+1), B: (p+1)*K, E: N*K
- Criterion to optimize: $RSS(B;\Sigma) = \sum_{i=1}^{N} (y_i - f(x_i))^T \Sigma^{-1} (y_i - f(x_i))$
- Solution: $\hat{B} = (X^T X)^{-1} X^T Y$

- Surprising result: Estimates and even confidence intervals are the same if doing K univariate multiple regressions!

Is MANOVA and Multivariate Linear Regression useful?

- Multivariate Regression, MANOVA not well supported in statistical software (including R)
- Useful, if you want to test if a predictor has an influence on any response
- Possible in theory, but not well supported:
  - simultaneous confidence intervals for several parameters
  - Tests among parameters of different responses

- R: Function “lm” with matrix as y and “summary(…, test = “Wilks”)”

Concepts to know

- Hotelling’s T-test
- Idea of MANOVA and Multivariate Regression

R functions to know

- “HotellingsT2”
- “Manova”
- “lm” with y being a matrix