Introduction to Multivariate Statistics
Lecture 3: Comparison of Multivariate Means

Lexin Li

University of California, Berkeley

Outline

- Motivating example: Anesthetizing effect of CO_2 ad halothane
  - $n = 19$ dogs, **each** of which was treated with 4 treatments
  - the response variable is milliseconds between heartbeats
  - 4 different treatments: 2 CO_2 pressures $\times$ halothane

Treatment	CO_2 pressure	Halothane
1	high	present
2	low	present
3	high	absent
4	low	absent

- What will change to answer the same question?
  - Suppose we select different 19 dogs for each of 4 treatments
  - Suppose we select different 3 dogs for each 4 treatments
  - Suppose the response is a binary indicator if anesthetizing is effective or not
  - Suppose there are many levels of CO_2

Outline

- What is it about:
  - Compare **quantitative** measures of **subjects** between groups that are defined by **factor(s)** with two or more **levels**
- Topics to cover:
  - Same subjects – within-subject comparison (Section 6.2)
    - multiple variables – paired comparison
    - multiple measurements – repeated measures design
  - Different subjects – between-subject comparison
    - one factor: two populations (Section 6.3) – two sample T^2 test
    - one factor: more than two populations (Section 6.4) – one-way MANOVA
    - two factors – two-way MANOVA (Section 6.7)
  - Multiple testing
  - Computing in R
- What to pay special attention:
  - One-to-one correspondence to the **univariate** comparison
  - Assumptions! (because they lead to different choices of test)

Review: hypothesis testing

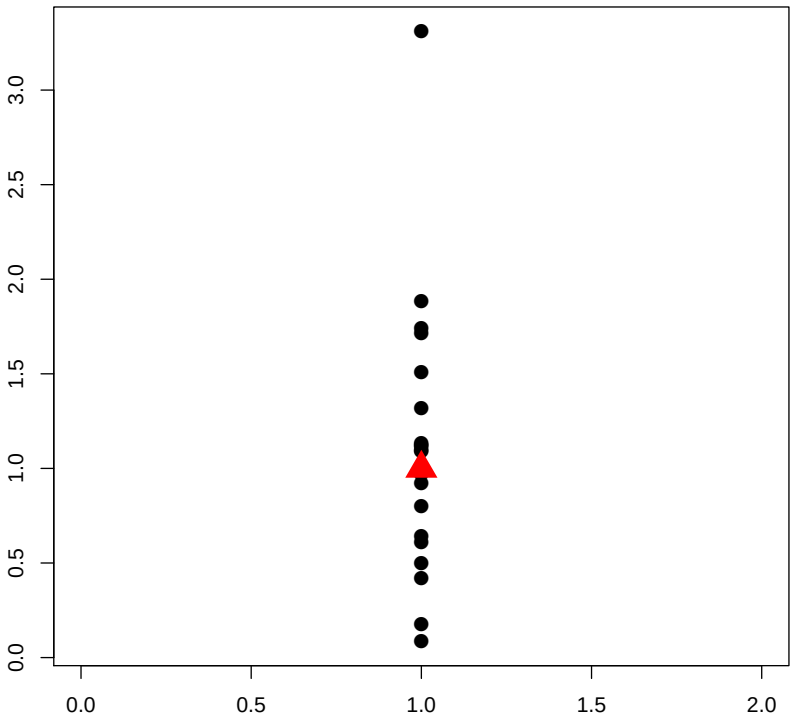
- To decide whether data (**sample**) contain enough information to cast doubt on the null hypothesis (a statement about **population**)
- Key elements:
  - A pair of **pre-specified** (null and alternative) hypotheses
  - A **test statistic** computed from the observed sample data
  - The **distribution** of the test statistic when the null hypothesis is true ← where statistical theory kicks in
  - **p-value**: the probability of obtaining a test statistic (from the population), assuming that the null hypothesis is true, at least as extreme as the one that was actually observed (from the data) – a small p-value indicates that, it would be highly unlikely to observed such data when the null hypothesis is true
  - Action: reject or do not reject the null hypothesis
- A few more things:
  - Type-I error: the chance of rejecting null when null is true
  - Type-II error: the chance of not rejecting null when null is not true
  - Sample size calculation

Paired comparison

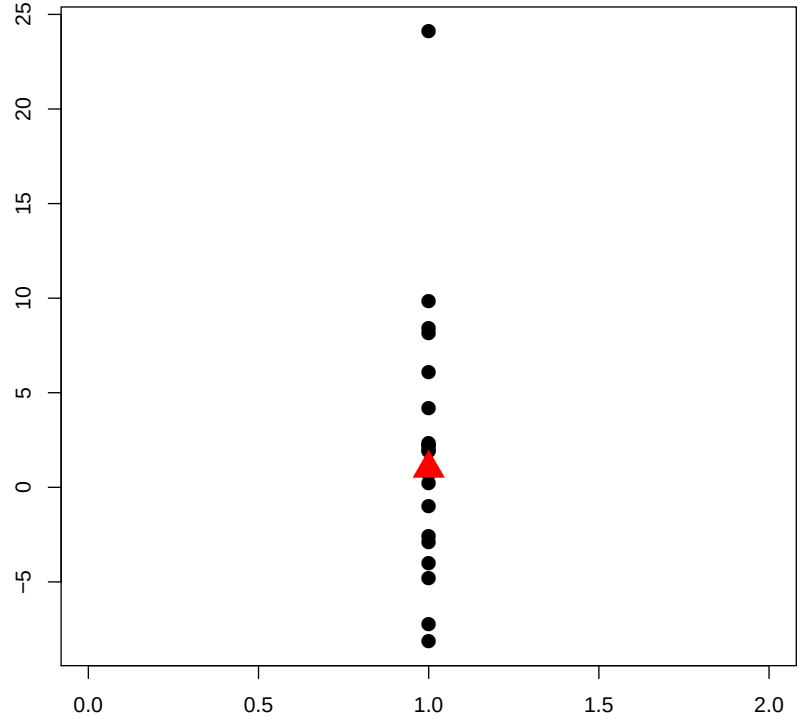
- Motivating example: Wastewater monitoring
  - Background: Municipal wastewater treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data of these self-monitoring programs led to a study in which samples of effluent were sent to two laboratories, one commercial lab and the other state lab, for testing.
  - Question: if the chemical analyses of the two labs agree with each other
  - Sample: $n = 11$ samples of wastewater, each of which is divided, with half to commercial lab, and the other half to state lab
  - Chemicals (variables) to measure: $p = 2$, biochemical oxygen demand (BOD) and suspended solids (SS)

- Remarks:
  - What if there is only **one** chemical (variable) measured?
  - What if **different** wastewater samples were sent to the two labs?

Compare mean from one population



$$\mu = 1$$
$$\sigma = 1^2$$
$$\text{p-value} = 1.415 \times 10^{-6}$$



$$\mu = 1$$
$$\sigma = 10^2$$
$$\text{p-value} = 0.211$$

Paired comparison

- **Univariate** paired comparison:
  - Setup: $Y_{11}, \ldots, Y_{1n} \sim D(\mu_1, \sigma_1)$; $Y_{21}, \ldots, Y_{2n} \sim D(\mu_2, \sigma_2)$ – same number of samples in each group, of course!
  - $H_0 : \mu_1 - \mu_2 = 0$
  - Assumption: D a normal distribution, or n is large
  - Test statistic:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{S_D/n}},$$

    where S_D is the sample variance of $D_i = Y_{1i} - Y_{2i}, i = 1, \ldots, n$
  - Distribution under H_0: $t \sim t_{n-1}$

- Remarks:
  - It is essentially a **one-sample** t test for the sample difference $D_i = Y_{1i} - Y_{2i}, i = 1, \ldots, n$

Theory behind ...

- **Univariate** normal distribution:
  - One random variable: $Y \sim N(\mu, \sigma)$ — σ is variance, not sd
  - n **i.i.d.** random variables: $Y_1, \ldots, Y_n \sim N(\mu, \sigma)$
  - Sample mean:
$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \frac{1}{n}\sigma)$$
  - Sample variance:
$$S = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
  - Distribution:
$$\frac{\bar{Y} - \mu}{\sqrt{S/n}} \sim t_{n-1}, \text{ or equivalently, } \frac{(\bar{Y} - \mu)^2}{S/n} \sim F_{1,n-1}$$
  - Central Limit Theorem: $Y_1, \ldots, Y_n \sim D(\mu, \sigma)$
$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \frac{1}{n}\sigma) \text{ approximately}$$

Paired comparison

- **Multivariate** paired comparison:
  - Setup: $\boldsymbol{Y}_{11}, \ldots, \boldsymbol{Y}_{1n} \sim D_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$; $\boldsymbol{Y}_{21}, \ldots, \boldsymbol{Y}_{2n} \sim D_p(\mu_2, \boldsymbol{\Sigma}_2)$ – same number of samples in each group, of course!
  - $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$
  - Assumption: D_p a normal distribution, or n is large
  - Test statistic:

$$T^2 = n(\bar{\boldsymbol{Y}}_1 - \bar{\boldsymbol{Y}}_2)^\top \boldsymbol{S}_D^{-1}(\bar{\boldsymbol{Y}}_1 - \bar{\boldsymbol{Y}}_2),$$

    where S_D is the sample **covariance matrix** of $\boldsymbol{D}_i = \boldsymbol{Y}_{1i} - \boldsymbol{Y}_{2i}, i = 1, \ldots, n$
  - Distribution under H_0: $T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}$

- Remarks:
  - One-to-one correspondence.

Theory behind ...

- **Multivariate** normal distribution:
  - One random vector: $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ – a $p \times 1$ vector
  - n **i.i.d.** random vectors: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Sample mean: a $p \times 1$ vector

$$\bar{\boldsymbol{Y}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Y}_i \sim N(\mu, \frac{1}{n}\boldsymbol{\Sigma})$$

  - Sample variance: a $p \times p$ matrix

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^{\mathsf{T}}$$

  - Distribution:

$$n(\bar{\boldsymbol{Y}} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{S}^{-1} (\bar{\boldsymbol{Y}} - \boldsymbol{\mu}) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

  - Central Limit Theorem: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \sim D_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\bar{\boldsymbol{Y}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Y}_i \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}) \text{ approximately}$$

Paired comparison

- Multivariate paired comparison in **numbers**:
  - Let $\boldsymbol{D}_i = \boldsymbol{Y}_{1i} - \boldsymbol{Y}_{2i}, i = 1, \ldots, n$. The test statistic can be written as

$$\begin{aligned} T^2 &= n(\bar{\boldsymbol{Y}}_1 - \bar{\boldsymbol{Y}}_2)^{\mathsf{T}} \boldsymbol{S}_D^{-1} (\bar{\boldsymbol{Y}}_1 - \bar{\boldsymbol{Y}}_2) \\ &= n\bar{\boldsymbol{D}}^{\mathsf{T}} \boldsymbol{S}_D^{-1} \bar{\boldsymbol{D}} \end{aligned}$$

  - Wastewater example:

$$\bar{\boldsymbol{D}} = \begin{pmatrix} -9.36 \\ 13.27 \end{pmatrix}, \quad \boldsymbol{S}_D = \begin{pmatrix} 199.26 & 88.31 \\ 88.31 & 418.61 \end{pmatrix}, \quad n = 11$$

$$T^2 = 11(-9.36, 13.27) \begin{pmatrix} 0.0055 & -0.0012 \\ -0.0012 & 0.0026 \end{pmatrix} \begin{pmatrix} -9.36 \\ 13.27 \end{pmatrix} = 13.6$$

p-value $= 0.021$

Paired comparison

- R code:

```
# load the data
data = read.table(file="./JW6/T06-01.dat", header=FALSE, quote="")
Y1 = data[,1:2]
Y2 = data[,3:4]
D = Y1 - Y2

# compute the sample size
n = nrow(data); n
[1] 11

# compute the mean
D.bar = apply(D,2,mean); D.bar
       V1        V2
-9.363636 13.272727

# compute the covariance
S = cov(D); S
          V1        V2
V1 199.25455  88.30909
V2  88.30909 418.61818
```

Paired comparison

- R code:

```
# compute the inverse of the covariance
S.inv = solve(cov(D)); S.inv
              V1           V2
V1  0.005536320 -0.001167908
V2 -0.001167908  0.002635186

# compute the test statistic
T2 = n * t(D.bar) %*% S.inv %*% D.bar; T2
[1,] 13.63931

# compute the p-value
1 - pf(T2*9/20,2,9)
[1,] 0.02082779
```

Repeated measures design

- Motivating example: Anesthetizing effect of CO_2 and halothane
  - Background: Improved anesthetIcs are often developed by first studying their effects on animals
  - Sample: $n = 19$ dogs, **each** of which was treated with 4 treatments
  - Response: milliseconds between heartbeats
  - 4 different treatments: 2 CO_2 pressures $\times$ halothane

Treatment	CO_2 pressure	Halothane
1	high	present
2	low	present
3	high	absent
4	low	absent

  - Hypothesis:
    - null effect: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
    - main effect of halothane: $(\mu_3 + \mu_4) - (\mu_1 + \mu_2)$
    - main effect of CO_2: $(\mu_1 + \mu_3) - (\mu_2 + \mu_4)$
    - interaction of CO_2 and halothane: $(\mu_1 + \mu_4) - (\mu_2 + \mu_3)$

Repeated measures design

Repeated measures design

- Repeated measures design:
  - q different treatments are compared with respect to a **single** variable
  - each subject receives **each treatment** once over successive periods of time, assuming no left-over effect
  - The i-th subject's response for q treatments, and the mean response:

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iq} \end{pmatrix}, \quad i = 1, ..., n, \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}$$

  - All hypotheses can be formulated as $H_0 : \boldsymbol{C}\boldsymbol{\mu} = 0$ for a $\tilde{q} \times q$ matrix $\boldsymbol{C}$

Repeated measures design

- Repeated measures design:
  - q different treatments are compared with respect to a **single** variable
  - each subject receives **each treatment** once over successive periods of time, assuming no left-over effect
  - The i-th subject's response for q treatments, and the mean response:

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iq} \end{pmatrix}, \quad i = 1, ..., n, \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}$$

  - All hypotheses can be formulated as $H_0 : \boldsymbol{C}\boldsymbol{\mu} = 0$ for a $\tilde{q} \times q$ matrix $\boldsymbol{C}$

$$\begin{pmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \mu_4 - \mu_3 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}_{3\times 4} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \boldsymbol{C}\boldsymbol{\mu}$$

$$(\mu_1 + \mu_3) - (\mu_2 + \mu_4) = (1, -1, 1, -1)_{1\times 4} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \boldsymbol{C}\boldsymbol{\mu}$$

Repeated measures design

- Linear combinations: $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\boldsymbol{a}^\mathsf{T}\boldsymbol{Y} \sim N_p(\boldsymbol{a}^\mathsf{T}\boldsymbol{\mu}, \boldsymbol{a}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{a})$, where $\boldsymbol{a} \in \mathbb{R}^p$
  - $\boldsymbol{A}\boldsymbol{Y} \sim N_p(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T})$, where $\boldsymbol{A} \in \mathbb{R}^{\tilde{p}\times p}$
- For our problem: $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iq})^\mathsf{T} \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\boldsymbol{C}\bar{\boldsymbol{Y}} \sim N_{\tilde{q}}(\boldsymbol{C}\boldsymbol{\mu}, n^{-1}\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\mathsf{T})$
  - $H_0 : \boldsymbol{C}\boldsymbol{\mu} = 0$
  - Assumption: Normal distribution of $\boldsymbol{Y}_i$ or n is large
  - Test statistic:

$$T^2 = n(\boldsymbol{C}\bar{\boldsymbol{Y}})^\mathsf{T}(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\mathsf{T})^{-1}\boldsymbol{C}\bar{\boldsymbol{Y}}$$

  - Distribution under H_0: $T^2 \sim \frac{(n-1)\tilde{q}}{n-\tilde{q}} F_{\tilde{q},n-\tilde{q}}$

Repeated measures design

- Repeated measures design in **numbers**:
  - Anesthetics example:

$$\bar{\boldsymbol{Y}} = \begin{pmatrix} 368.21 \\ 404.63 \\ 479.26 \\ 502.89 \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} 2819.29 & & & \\ 3568.42 & 7963.14 & & \\ 2943.49 & 5303.98 & 6851.32 & \\ 2295.35 & 4065.44 & 4499.63 & 4878.99 \end{pmatrix}, \quad n = 19,$$

$$\boldsymbol{C}\bar{\boldsymbol{Y}} = \begin{pmatrix} 36.42 \\ 74.63 \\ 23.63 \end{pmatrix}, \quad \boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^{\mathsf{T}} = \begin{pmatrix} 3645.59 & & \\ -2034.23 & 4206.50 & \\ -590.40 & -1113.15 & 2732.05 \end{pmatrix},$$

$$\begin{aligned} T^2 &= 19(36.42, 74.63, 23.63) \begin{pmatrix} 0.00047 & & \\ 0.00028 & 0.00044 & \\ 0.00022 & 0.00024 & 0.00051 \end{pmatrix} \begin{pmatrix} 36.42 \\ 74.63 \\ 23.63 \end{pmatrix} \\ &= 116.02 \end{aligned}$$

p-value ≈ 0.

Repeated measures design

- R code:

```
# load the data
data = read.table(file="./JW6/T06-02.dat", header=FALSE, quote="")

# compute the sample size
n = nrow(data); n
[1] 19

# compute the mean
Y.bar = apply(data,2,mean); Y.bar
      V1       V2       V3       V4
368.2105 404.6316 479.2632 502.8947

# compute the variance
S = cov(data); S
         V1       V2       V3       V4
V1 2819.287 3568.415 2943.497 2295.357
V2 3568.415 7963.135 5303.991 4065.459
V3 2943.497 5303.991 6851.316 4499.640
V4 2295.357 4065.459 4499.640 4878.988
```

Repeated measures design

- R code:

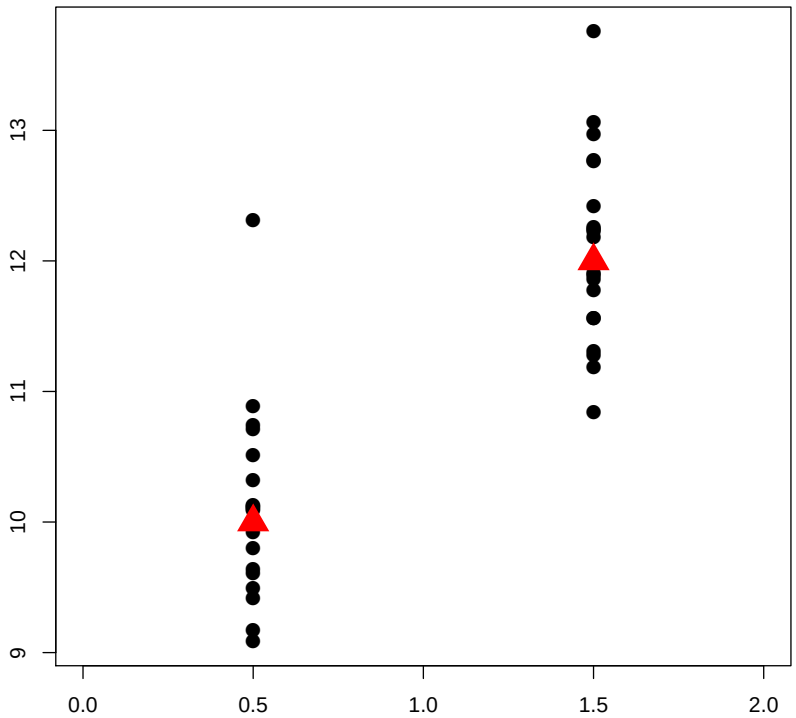
```
# specify C
C = cbind(c(-1,0,0),c(1,-1,0),c(0,1,-1),c(0,0,1)); C
     [,1] [,2] [,3] [,4]
[1,]   -1    1    0    0
[2,]    0   -1    1    0
[3,]    0    0   -1    1

# compute the test statistic
CY = C %*% Y.bar; CY
CS = C %*% S %*% t(C); CS
T2 = n * t(CY) %*% solve(CS) %*% CY; T2
1 - pf(16*T2/(18*3),3,16)
         [,1]
[1,] 36.42105
[2,] 74.63158
[3,] 23.63158
           [,1]      [,2]       [,3]
[1,]  3645.5906 -2034.225  -590.3918
[2,] -2034.2251  4206.468 -1113.1433
[3,]  -590.3918 -1113.143  2731.0234
[1,] 116.0163
[1,] 3.317767e-07
```

Compare means from two populations

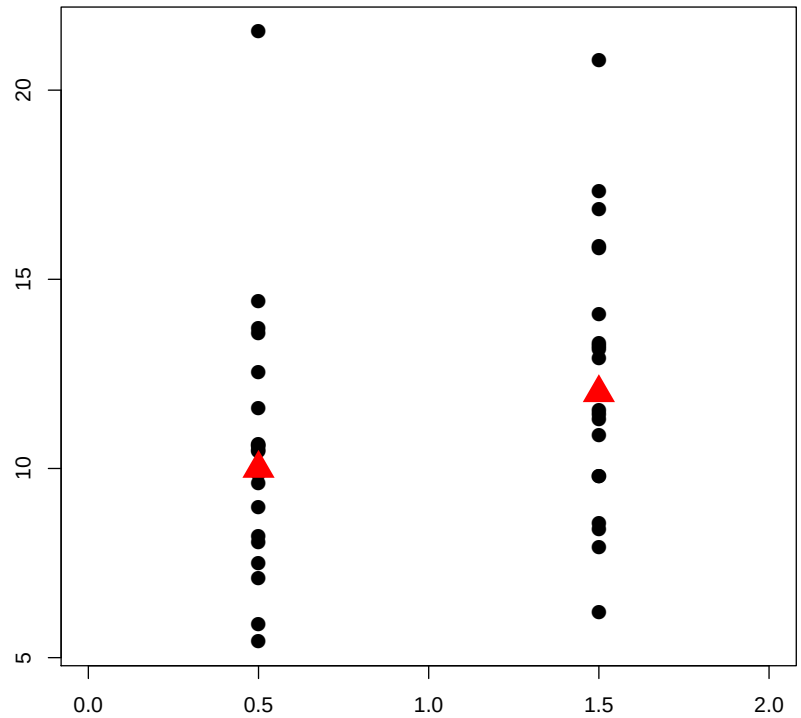
- Motivating example: Electricity consumption with and without air conditioning
  - Question: difference of electricity consumption with and without air conditioning
  - Sample: $n_1 = 45$ homes with air conditioning *vs* $n_2 = 55$ homes without
  - Measurements of electrical usage: total on-peak consumption (Y_1), and total off-peak consumption (Y_2)

- Remarks:
  - What if there is only **one** usage measure instead of two?

Compare means from two populations



$\mu_1 = 10, \mu_2 = 12$
$\sigma_1 = \sigma_2 = 1$
p-value ≈ 0



$\mu_1 = 10, \mu_2 = 12$
$\sigma_1 = \sigma_2 = 5^2$
p-value $= 0.101$

Compare means from two populations

- **Univariate** two-sample comparison:
  - Setup: $Y_{11}, \ldots, Y_{1n_1} \sim D(\mu_1, \sigma_1)$; $Y_{21}, \ldots, Y_{2n_2} \sim D(\mu_2, \sigma_2)$
    – number of samples in the two groups may differ
  - $H_0 : \mu_1 - \mu_2 = 0$
  - Assumption: two populations are **independent**; $\sigma_1 = \sigma_2$; D a normal distribution or n_1 and n_2 are large
  - Test statistic:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{S_p(\frac{1}{n_1} + \frac{1}{n_2})}},$$

    where $S_p = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$ is the pooled estimator of the variance $\sigma_1 = \sigma_2$
  - Distribution under H_0: $t \sim t_{n_1+n_2-2}$

- Remarks:
  - Compare with the paired t test
  - What if $\sigma_1 \neq \sigma_2$: the denominator becomes $\sqrt{\frac{S_1}{n_1} + \frac{S_2}{n_2}}$

Compare means from two populations

- **Multivariate** two-sample comparison:
  - Setup: $\boldsymbol{Y}_{11}, \ldots, \boldsymbol{Y}_{1n_1} \sim D_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$; $\boldsymbol{Y}_{21}, \ldots, \boldsymbol{Y}_{2n_2} \sim D_p(\mu_2, \boldsymbol{\Sigma}_2)$
    – number of samples in the two groups may differ
  - $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$
  - Assumption: two populations are **independent**; $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$; D_p a normal distribution, or n_1 and n_2 are large
  - Test statistic:

$$T^2 = (\bar{\boldsymbol{Y}}_1 - \bar{\boldsymbol{Y}}_2)^\mathsf{T} \left\{ \boldsymbol{S}_p \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{-1} (\bar{\boldsymbol{Y}}_1 - \bar{\boldsymbol{Y}}_2)$$

    where $\boldsymbol{S}_p = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$ is the pooled sample covariance matrix
  - Distribution under H_0: $T^2 \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p, n_1+n_2-p-1}$
- Remarks:
  - What if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$:
    - when $n_1 - p$ and $n_2 - p$ are both large, $T^2 \sim \chi^2_p$ approximately
    - when n_1 and n_2 are large, $\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p, n_1+n_2-p-1} \approx \chi^2_p$

    So the test assuming equal covariance is still valid approximately.

Compare means from two populations

- Multivariate two-sample comparison **in numbers**:
  - Electricity consumption example:

$$\bar{\boldsymbol{Y}}_1 = \begin{pmatrix} 204.4 \\ 556.6 \end{pmatrix}, \quad \boldsymbol{S}_1 = \begin{pmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{pmatrix}, \quad n_1 = 45$$

$$\bar{\boldsymbol{Y}}_2 = \begin{pmatrix} 130.0 \\ 355.0 \end{pmatrix}, \quad \boldsymbol{S}_2 = \begin{pmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{pmatrix}, \quad n_1 = 55$$

$$T^2 = (74.4, 201.6) \left\{ \begin{pmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{pmatrix} \left(\frac{1}{45} + \frac{1}{55} \right) \right\}^{-1} \begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} = 16.067$$

p-value $= 0.00063$

Compare means from two populations

- Why do we need multivariate test in the first place?
  - How many variables (p) are there?
  - For p as small as 2, can I simply perform **2 univariate** t tests instead of **1 multivariate** T^2 test?

- Example: Lizards data
  - Background: A zoologist collected lizards in the southwestern US and compared the two types
  - Question: any difference between the two types of lizards
  - Sample: $n_1 = 20$ Cnemidophorus lizards *vs* $n_2 = 40$ Sceloporus lizards
  - Measurements: mass in grams (Y_1), and snout-vent length in millimeters (Y_2)

- Results: pre-determined significance level $\alpha = 0.05$

Compare means from two populations

- Why do we need multivariate test in the first place?
  - How many variables (p) are there?
  - For p as small as 2, can I simply perform **2 univariate** t tests instead of **1 multivariate** T^2 test?

- Example: Lizards data
  - Background: A zoologist collected lizards in the southwestern US and compared the two types
  - Question: any difference between the two types of lizards
  - Sample: $n_1 = 20$ Cnemidophorus lizards *vs* $n_2 = 40$ Sceloporus lizards
  - Measurements: mass in grams (Y_1), and snout-vent length in millimeters (Y_2)

- Results: pre-determined significance level $\alpha = 0.05$
  - The p-values for the two univariate t tests are, **0.46 and 0.08**, respectively

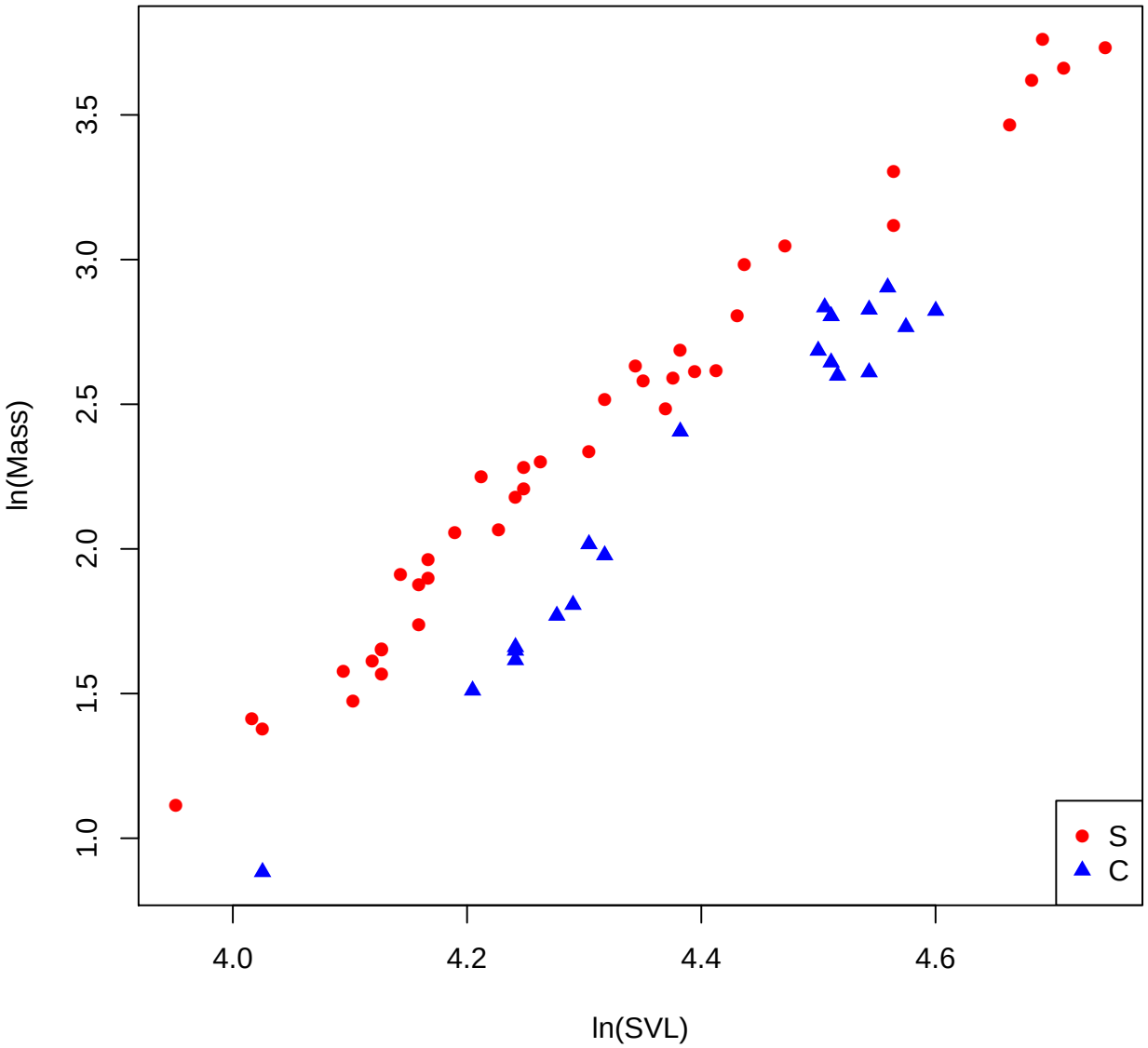
Compare means from two populations

- Why do we need multivariate test in the first place?
  - How many variables (p) are there?
  - For p as small as 2, can I simply perform **2 univariate** t tests instead of **1 multivariate** T^2 test?

- Example: Lizards data
  - Background: A zoologist collected lizards in the southwestern US and compared the two types
  - Question: any difference between the two types of lizards
  - Sample: $n_1 = 20$ Cnemidophorus lizards *vs* $n_2 = 40$ Sceloporus lizards
  - Measurements: mass in grams (Y_1), and snout-vent length in millimeters (Y_2)

- Results: pre-determined significance level $\alpha = 0.05$
  - The p-values for the two univariate t tests are, **0.46 and 0.08**, respectively
  - The p-value for the multivariate T^2 test is **0.0001**

Compare means from two populations

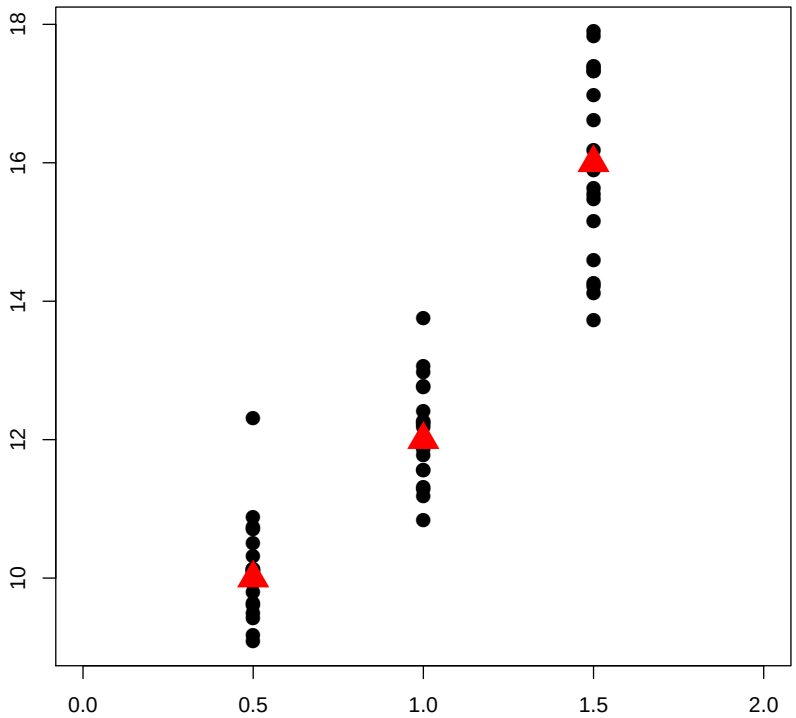


Compare means from many populations

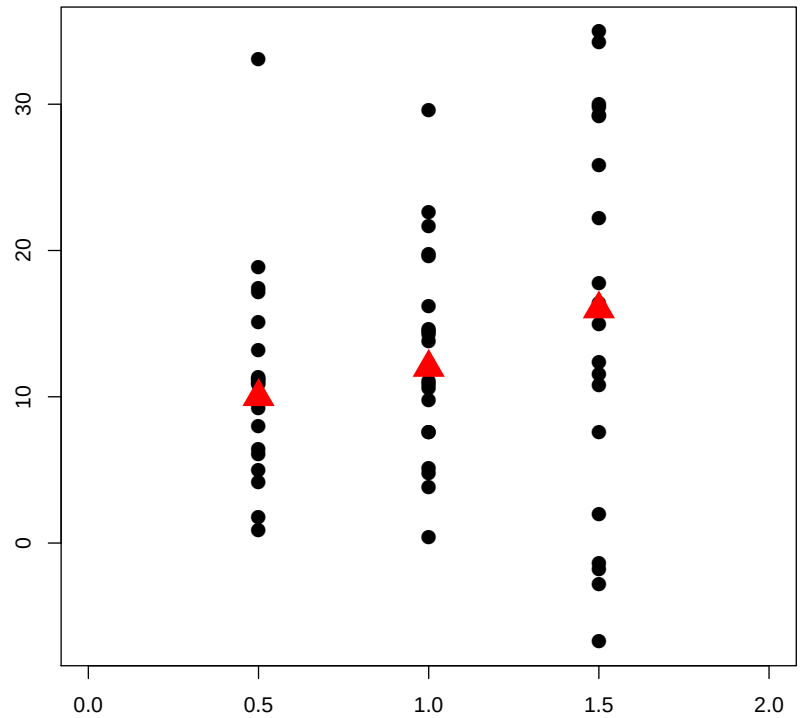
- Motivating example: Wisconsin nursing home
  - Background: The State of Wisconsin reimburses nursing homes and develops a set of formulas for rates for each facility, based on factors such as level of care, mean wage rate, etc.
  - Question: to investigate the effects of ownership on costs; to compare 3 groups of ownership, private, nonprofit, government
  - Samples: $n_1 = 271$, $n_2 = 138$, $n_3 = 107$
  - Measurements: 4 cost variables, Y_1 = cost of nursing labor, Y_2 = cost of dietary labor, Y_3 = cost of plant operation and maintenance labor, Y_4 = cost of housekeeping and laundry labor

- Remarks:
  - What if there is only **one** cost variable instead of four?
  - What if one is interested in the effect of both **ownership** and **certification** (skilled nursing facility, intermediate care facility, or a combination of the two)

Compare means from many populations



$\mu_1 = 10, \mu_2 = 12, \mu_3 = 16$
$\sigma_1 = \sigma_2 = 1$
p-value ≈ 0



$\mu_1 = 10, \mu_2 = 12, \mu_3 = 16$
$\sigma_1 = \sigma_2 = 10^2$
p-value $= 0.125$

Compare means from many populations

- **Univariate one-way** analysis of variance (ANOVA):
  - Setup:

$$\begin{aligned} Y_{11}, \ldots Y_{1n_1} &\sim N(\mu_1, \sigma_1^2) \\ Y_{21}, \ldots Y_{2n_2} &\sim N(\mu_2, \sigma_2^2) \\ &\ldots \\ Y_{g1}, \ldots Y_{gn_g} &\sim N(\mu_g, \sigma_g^2) \end{aligned}$$

  - Hypothesis $H_0 : \mu_1 = \ldots = \mu_g$
  - Assumption: **independent**; $\sigma_1 = \ldots = \sigma_g$; normal; $g \geq 2$
  - Intuition: let's look at some graphs!

Compare means from many populations

- **One-way ANOVA table:**

Source of variation	Sum of squares	Degrees of freedom
Treatment	$\text{SSTr} = \sum_{\ell=1}^{g} n_\ell(\bar{x}_\ell - \bar{x})^2$	$g-1$
Residual (error)	$\text{SSE} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \bar{x}_\ell)^2$	$\sum_{\ell=1}^{g} n_\ell - g$
Total	$\text{SST} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \bar{x})^2$	$\sum_{\ell=1}^{g} n_\ell - 1$

- $H_0 : \mu_1 = \ldots = \mu_g$
- Test statistic:

$$F = \frac{\text{SSTr}/(g-1)}{\text{SSE}/(\sum_{\ell=1}^{g} n_\ell - g)} \sim F_{g-1, \sum_{\ell=1}^{g} n_\ell - g}$$

- Reject H_0 if $F > F_{g-1, \sum_{\ell=1}^{g} n_\ell - g}(\alpha)$
- For the Wisconsin nursing home example, if there is only one cost variable, then the reference distribution is $F_{?,?}$
  — The answer is (**wake-up call**):

Compare means from many populations

- **One-way ANOVA table:**

Source of variation	Sum of squares	Degrees of freedom
Treatment	$\text{SSTr} = \sum_{\ell=1}^{g} n_\ell(\bar{x}_\ell - \bar{x})^2$	$g-1$
Residual (error)	$\text{SSE} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \bar{x}_\ell)^2$	$\sum_{\ell=1}^{g} n_\ell - g$
Total	$\text{SST} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \bar{x})^2$	$\sum_{\ell=1}^{g} n_\ell - 1$

- $H_0 : \mu_1 = \ldots = \mu_g$
- Test statistic:

$$F = \frac{\text{SSTr}/(g-1)}{\text{SSE}/(\sum_{\ell=1}^{g} n_\ell - g)} \sim F_{g-1,\sum_{\ell=1}^{g} n_\ell - g}$$

- Reject H_0 if $F > F_{g-1,\sum_{\ell=1}^{g} n_\ell - g}(\alpha)$
- For the Wisconsin nursing home example, if there is only one cost variable, then the reference distribution is $F_{?,?}$
  — The answer is (**wake-up call**): $F_{2,513}$

Compare means from many populations

- One-way MANOVA Table:

Source	Sum of squares	Df
Treatment	$\boldsymbol{B} = \sum_{\ell=1}^{g} n_\ell (\bar{\boldsymbol{x}}_\ell - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_\ell - \bar{\boldsymbol{x}})^\mathsf{T}$	$g-1$
Residual	$\boldsymbol{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell)(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell)^\mathsf{T}$	$\sum_{\ell=1}^{g} n_\ell - g$
Total	$\boldsymbol{B} + \boldsymbol{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}})^\mathsf{T}$	$\sum_{\ell=1}^{g} n_\ell - 1$

- $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$
- Test statistic:

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B} + \boldsymbol{W}|} = \frac{1}{|\boldsymbol{W}^{-1}\boldsymbol{B} + \boldsymbol{I}_p|} \sim \text{Table 6.3}$$

- Other forms of test statistics = all functions of $\boldsymbol{W}^{-1}\boldsymbol{B}$

Compare means from many populations

- **Two-way ANOVA Table:**

Source of variation	Sum of squares	Degrees of freedom
Factor 1	$\mathrm{SS}_{\mathrm{fac1}} = \sum_{\ell=1}^{g} bn(\bar{x}_{\ell\cdot} - \bar{x})^2$	$g-1$
Factor 2	$\mathrm{SS}_{\mathrm{fac2}} = \sum_{k=1}^{b} gn(\bar{x}_{\cdot k} - \bar{x})^2$	$b-1$
Interaction	$\mathrm{SS}_{\mathrm{int}} = \sum_{\ell=1}^{g}\sum_{k=1}^{b} n(x_{\ell k} - \bar{x}_{\ell\cdot} - \bar{x}_{\cdot k} + \bar{x})^2$	$(g-1)(b-1)$
Residual	$\mathrm{SS}_{\mathrm{res}} = \sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{\ell kr} - \bar{x}_{\ell k})^2$	$gb(n-1)$
Total (Corrected)	$\mathrm{SS}_{\mathrm{c.tot}} = \sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{\ell kr} - \bar{x})^2$	$gbn-1$

  - main effect of factor 1; of factor 2; interaction between the two
  - test statistics:

$$\frac{\mathrm{SS}_{\mathrm{fac1}}/(g-1)}{\mathrm{SS}_{\mathrm{res}}/\{gb(n-1)\}}, \quad \frac{\mathrm{SS}_{\mathrm{fac2}}/(b-1)}{\mathrm{SS}_{\mathrm{res}}/\{gb(n-1)\}}, \quad \frac{\mathrm{SS}_{\mathrm{int}}/\{(g-1)(b-1)\}}{\mathrm{SS}_{\mathrm{res}}/\{gb(n-1)\}}$$

- **Two-way MANOVA Table:**
  - replace $(\bar{x}_{\ell\cdot} - \bar{x})^2$ with $(\bar{\boldsymbol{x}}_{\ell\cdot} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{\ell\cdot} - \bar{\boldsymbol{x}})^{\mathsf{T}}$, ...
  - test statistics:

$$\Lambda^*_{\mathrm{fac1}} = \frac{|\mathrm{SSP}_{\mathrm{res}}|}{|\mathrm{SSP}_{\mathrm{fac1}} + \mathrm{SSP}_{\mathrm{res}}|}, \quad \Lambda^*_{\mathrm{fac2}} = \frac{|\mathrm{SSP}_{\mathrm{res}}|}{|\mathrm{SSP}_{\mathrm{fac2}} + \mathrm{SSP}_{\mathrm{res}}|},$$

$$\Lambda^*_{\mathrm{int}} = \frac{|\mathrm{SSP}_{\mathrm{res}}|}{|\mathrm{SSP}_{\mathrm{int}} + \mathrm{SSP}_{\mathrm{res}}|}$$

MANOVA

- Example: Plastic film quality
  - 3 quality measurements: Y_1 = tear resistance, Y_2 = gloss, Y_3 = opacity
  - 2 factors, each with 2-levels: rate of extrusion (low/high), amount of an additive (low/high)
  - 5 replications with each combination of the factor levels

- R code:

```
# load the data
data<-read.table(file="./JW6/EX-06-13.dat", header=FALSE, quote="")
y<-as.matrix(data[,3:5])

rate<-as.factor(data[,1]); rate
[1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
Levels: 0 1

additive<-as.factor(data[,2]); additive
[1] 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1
Levels: 0 1
```

MANOVA

- R code:

```
# generate factors
rate <- factor(gl(2,10), labels=c("Low", "High")); rate
[1] Low  Low  Low  Low  Low  Low  Low  Low  Low  Low
High High High High High High High High High High
Levels: Low High

additive <- factor(gl(2,5,len=20), labels=c("Low", "High")); additive
[1] Low  Low  Low  Low  Low  High High High High High
Low  Low  Low  Low  Low  High High High High High
Levels: Low High

# one-way anova
summary(aov(y[,1] ~ rate))
            Df Sum Sq Mean Sq F value  Pr(>F)
rate         1  1.740  1.7405   12.41 0.00243 **
Residuals   18  2.525  0.1403

summary(aov(y[,3] ~ additive))
            Df Sum Sq Mean Sq F value Pr(>F)
additive     1    4.9   4.901   1.273  0.274
Residuals   18   69.3   3.850
```

MANOVA

- R code:

```
# one-way manova
summary(manova(y ~ rate))
          Df  Pillai approx F num Df den Df   Pr(>F)
rate       1 0.58638    7.561      3     16 0.002273 **
Residuals 18


# two-way manova, additive model
summary(manova(y ~ rate + additive), test="Wilks")
          Df   Wilks approx F num Df den Df  Pr(>F)
rate       1 0.38684   7.9253      3     15 0.00212 **
additive   1 0.55384   4.0279      3     15 0.02753 *
Residuals 17


# two-way manova, interaction model
summary(manova(y ~ rate * additive), test="Wilks")
              Df   Wilks approx F num Df den Df   Pr(>F)
rate           1 0.38186   7.5543      3     14 0.003034 **
additive       1 0.52303   4.2556      3     14 0.024745 *
rate:additive  1 0.77711   1.3385      3     14 0.301782
Residuals     16
```

Multiple testing

- Why should one be concerned?
  - Setup: test a number of, say m, hypotheses **simultaneously**
  - Probability of claiming at least one significant result while all results are truly insignificant:

$$\begin{aligned} p &= \text{Prob(at least one result claimed significant)} \\ &= 1 - \text{Prob(all results are claimed insignificant)} \\ &= 1 - (1-\alpha)^m \end{aligned}$$

  - Let $\alpha = 0.05$:

Multiple testing

- Why should one be concerned?
  - Setup: test a number of, say m, hypotheses **simultaneously**
  - Probability of claiming at least one significant result while all results are truly insignificant:

$$\begin{aligned} p &= \text{Prob(at least one result claimed significant)} \\ &= 1 - \text{Prob(all results are claimed insignificant)} \\ &= 1 - (1-\alpha)^m \end{aligned}$$

  - Let $\alpha = 0.05$:
    if $m = 1$, then $p = 0.05$
    if $m = 2$, then $p = 0.10$
    if $m = 20$, then $p = 0.64$
    if $m = 100$, then $p = 0.99$

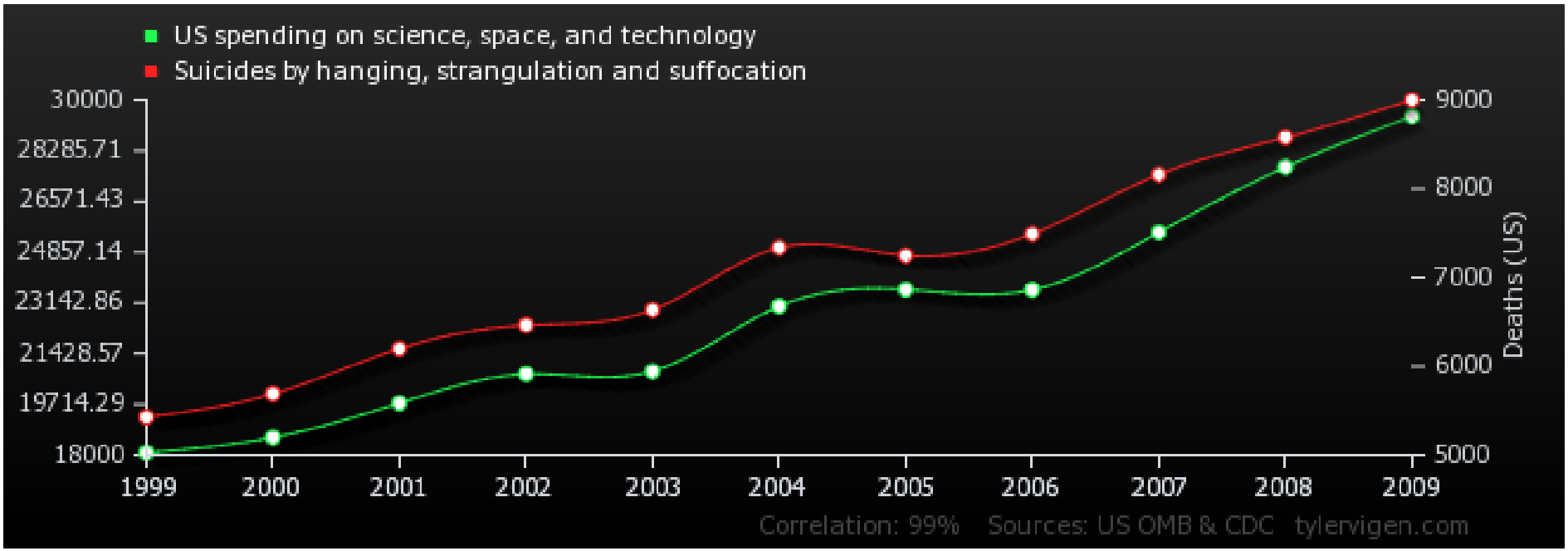
- Solutions:
  - Bonferroni correction; false discovery rate; many others ...

Association and causation

- **Association $\neq$ Causation**
  - A correlation / association between two variables does **not** necessarily imply that one causes the other, due to existence of potential **confounders**
  - Most statistical analysis infer about correlation between variables; a very small number of analyses infer about causation (causal inference)

- Examples (funny and not so funny ones):
  - Ex.1: sleeping with one's shoes on is strongly correlated with waking up with a headache — therefore, sleeping with one's shoes on causes headache
  - Ex.2: as ice cream sales increase, the rate of drowning deaths increases sharply — therefore, ice cream consumption causes drowning

Association and causation



Association and causation

- Examples (continued):
  - Ex.3: young children who sleep with the light on are much more likely to develop myopia in later life — therefore, sleeping with the light on causes myopia [published in ***Nature***, 1999, and received much coverage in the popular press; later study did not find that infants sleeping with the light on caused the development of myopia; it did find a strong link between parental myopia and the development of child myopia, also noting that myopic parents were more likely to leave a light on in their children's bedroom; in this case, the cause of both conditions is parental myopia, and the above-stated conclusion is false]
  - Ex.4: HDL ("good") cholesterol is negatively correlated with incidence of heart attack — therefore, taking medication to raise HDL will decrease the chance of having a heart attack [later research called this conclusion into question; instead, it may be that other underlying factors, like genes, diet and exercise, affect both HDL levels and the likelihood of having a heart attack]

Association and causation

- Some lessons:
  - We laugh at obvious mistakes but often forget how easy it is to make subtle errors any time an attempt is made to use statistics to prove causality [*Little Handbook of Statistics* by G.E. Dallal]
  - With that said, we should also bear in mind that,
    - Fully randomized study
    - With some not very complicated statistical techniques, plus some **additional assumptions** we wish to make, one can do **causal inference** with the **observational data** too

Discussion

- What is this chapter about: **in a bigger picture**
  - Study the **association** between **quantitative** variable(s) and **qualitative** variable(s)

- important concepts:
  - Quantitative (continuous) variable
    - What are they?
    - How many of them? $\rightarrow p \rightarrow$ univariate or multivariate test
  - Qualitative (categorical) variable $\Leftrightarrow$ Factor
    - What are they?
    - How many **factors**? $\rightarrow$ one-way or two-way (M)ANOVA
    - How many **levels** does each factor have? $\rightarrow$ two-sample test or (M)ANOVA
  - Subjects / samples
    - What constitute the samples?
    - How many samples? $\rightarrow n$

Discussion

- Things to pay attention to:
  - Data setup / conditions — What is the right choice of the test, while there are so many out there?
  - Extension from a **univariate** test to a **multivariate** test — Which parts are similar and which are different? Why do we need a multivariate version at all?
  - Multiple testing
  - **Association $\neq$ Causation**
- Topics covered:
  - **Parametric**
  - **Quantitative vs qualitative**
  - a **small** number of variables
- Topics not covered:
  - **Nonparametric**
  - **Qualitative vs qualitative**
  - a **large** number of variables