Comparisons of Two Means
Edps/Soc 584 and Psych 594
Applied Multivariate Statistics

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Outline

- Paired Comparisons: p variables, 2 matched pairs (i.e., dependent samples):

$$H_o : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta} = \mathbf{0}$$

- Repeated measures designs: 1 variable measured as multiple times:

$$H_o : \boldsymbol{L\mu} = \mathbf{0}$$

- Two independent samples: Four Cases of

$$H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

- Missing data — later in the semester

Reading: Johnson & Wichern pages 273–296

Paired Comparisons (dependent samples)

Paired observations arise in a number of different ways:

- Every subject (case) responds twice (e.g., pre/post test)
- Cases may be matched (on relevant variables) and then randomly assigned to one of two treatments.
- Naturally occurring pairs: husbands/wifes, siblings, etc.

The plan: Review univariate and then generalize to the multivariate situation.

For $j = 1, \ldots, n$ (number of pairs), let

$X_{j1} =$ measurement (response) of the j^{th} case given treatment 1.

$X_{j2} =$ measurement (response) of the j^{th} case given treatment 2.

We want to examine the differences

$$D_j = X_{j1} - X_{j2}$$

Univariate Case

$$D_j = X_{j1} - X_{j2}$$

If $D_j \sim \mathcal{N}(\delta, \sigma^2_D)$, then the statistic

$$t = \frac{\bar{D} - \delta}{s_D/\sqrt{n}} \sim \text{ Student's } t \text{ distribution}$$

where

$$\bar{D} = (1/n)\sum_{j=1}^n D_j = (1/n)\sum_{j=1}^n (X_{j1} - X_{j2})$$

$$s^2_D = (1/(n-1))\sum_{j=1}^n (D_j - \bar{D})^2$$

- Test

$$H_o : \delta = 0 \qquad \text{versus} \qquad H_A : \delta \neq 0$$

  (or $H_o : \delta = \delta_o$ versus $H_A : \delta \neq \delta_o$).

- A $100(1-\alpha)\%$ confidence interval (estimate) of δ

$$\bar{D} \pm t_{n-1}(\alpha/2)\frac{s_D}{\sqrt{n}}$$

Advantage

The advantage of looking at differences using paired comparisons. . .

It eliminates effects of case-to-case variation, because the variance (standard deviation) of differences is reduced to the extent that the scores/measurements are positively correlated

$$\sigma_D^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 - 2\sigma_{X_1,X_2}$$

This result comes from what we know about linear combinations:

$$D = \boldsymbol{a}'\boldsymbol{X} = (1,-1)\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = X_1 - X_2$$

so

$$\mu_D = \boldsymbol{a}'\boldsymbol{\mu} \qquad \text{var}(D) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$$

where $\boldsymbol{\mu}_{2\times1}$ is the mean vector for $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_{2\times2}$ covariance matrix for $\boldsymbol{X}$.

Multivariate Situation

Record p variables for each treatment (condition) for each member of each pair.

For case j, we have

$$\begin{array}{ll}
X_{1j1} = \text{variable 1, treatment 1} & X_{2j1} = \text{variable 1, treatment 2} \\
X_{1j2} = \text{variable 2, treatment 1} & X_{2j2} = \text{variable 2, treatment 2} \\
\vdots & \vdots \\
X_{1jp} = \text{variable p, treatment 1} & X_{2jp} = \text{variable p, treatment 2}
\end{array}$$

where $j = 1, \ldots, n$ ($n =$ the number of pairs that we have).

We Study the differences

$$\begin{array}{c}
D_{j1} = X_{1j1} - X_{2j1} \\
D_{j2} = X_{1j2} - X_{2j2} \\
\vdots \\
D_{jp} = X_{1jp} - X_{2jp}
\end{array}
\quad \longrightarrow \quad
\boldsymbol{D}_j = \begin{pmatrix} D_{j1} \\ D_{j2} \\ \vdots \\ D_{jp} \end{pmatrix}$$

Needed for Statistical Inference

Assume the $\boldsymbol{D}_j \sim \mathcal{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_D)$ and $i.i.d.$ for $j = 1, \ldots, J$ where

$$\boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{pmatrix} = E(\boldsymbol{D}_j)$$

If the differences $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n$ are a random sample from a $\mathcal{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_D)$ population, then

$$T^2 = n(\bar{D} - \boldsymbol{\delta})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{D}} - \boldsymbol{\delta}) \sim \frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}$$

Modification for Large Samples: If n and (n-p) are large, then T^2 is approximately distributed as a χ^2_p random variable regardless of the distribution of $\boldsymbol{D}_j$ (i.e., $\boldsymbol{D}_j$ may not be multivariate normal, but $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}_D^{-1}$ exist).

Statistical Inference

Suppose that we have observations $\boldsymbol{d}_j' = (d_{j1}, d_{j2}, \ldots, d_{jp}$ for $j = 1, \ldots, n)$.

Descriptive statistics:

$$\bar{\boldsymbol{d}}_{p\times 1} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{d}_j \qquad \text{and} \qquad \boldsymbol{S}_{d,(p\times p)} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{d}_j - \bar{\mathbf{d}})(\boldsymbol{d}_j - \bar{\mathbf{d}})'$$

Hypothesis Test:

$$H_o : \boldsymbol{\delta} = \mathbf{0} \qquad \text{versus} \qquad H_A : \boldsymbol{\delta} \neq \mathbf{0}$$

. . . assuming $D_j \sim \mathcal{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_D)$ and $i.i.d.$

Reject H_o if

$$T^2 = n\bar{\boldsymbol{d}}'\boldsymbol{S}^{-1}\bar{\boldsymbol{d}} \geq \frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}(\alpha)$$

If you Reject $H_o : \delta = 0$

- Confidence Region:

$$n(\bar{D}-\boldsymbol{\delta})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{D}}-\boldsymbol{\delta}) \le \frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}(\alpha)$$

- Simultaneous T^2 Intervals for individual differences of components means

$$\delta_i : \qquad \bar{d}_i \pm \sqrt{\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}(\alpha)}\sqrt{s^2_{d_i}/n}$$

  where $\bar{d}_i$ is mean difference of the i^{th} variable and $s^2_{d_i}$ is the i^{th} diagonal element of S_d.

- Bonferroni $100(1-\alpha)\%$ confidence intervals

$$\delta_i : \qquad \bar{d}_i \pm t_{n-1}(\alpha/2m)\sqrt{s^2_{d_i}/n}$$

  where $m =$ the number of confidence intervals (comparisons).

Large Samples

- For Large $(n-p)$ (i.e., D_j need not be multivariate normal)

$$\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}(\alpha) \approx \chi^2_p(\alpha)$$

Example: The data

Data from Table 5.9, page 153-154 of Rencher (2007):

"Each of 15 students wrote an informal and a formal essay (Kramer, 1972, p100). The variables were recorded were the number of words and number of verbs"

$$\begin{aligned} y1 &= \text{words in informal essay} \\ y2 &= \text{verbs in informal essay} \\ y3 &= \text{words in formal essay} \\ y4 &= \text{verbs in formal essay} \end{aligned}$$

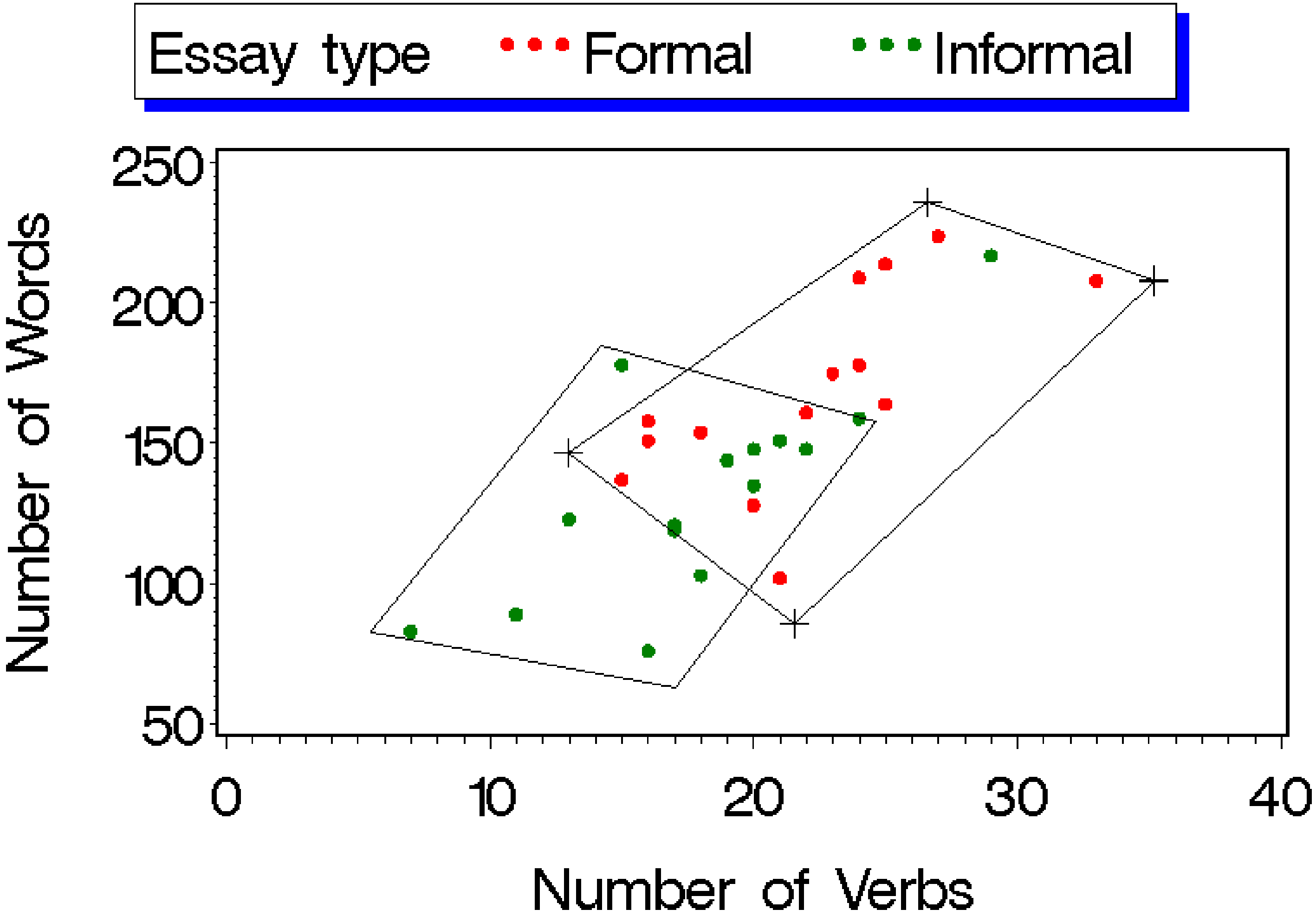
These are count data. CLT kick-in? $n = 15$ smallish

Sample Statistics:

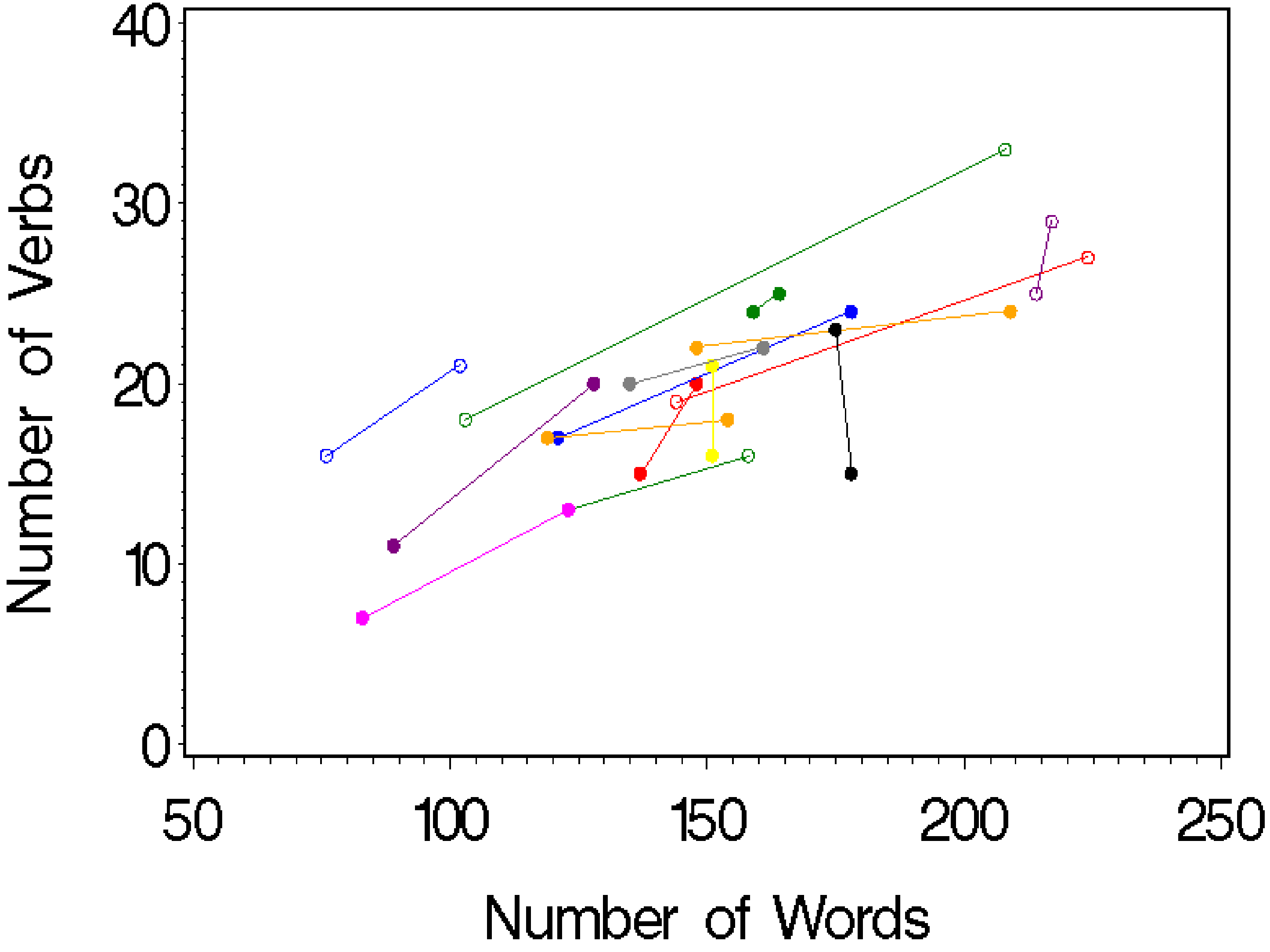
Difference: $d =$words [verbs] informal $-$ words [verbs] formal.

$$\bar{\boldsymbol{d}} = \begin{pmatrix} 32.80 \\ 3.53 \end{pmatrix} \begin{matrix} \leftarrow \text{words} \\ \leftarrow \text{verbs} \end{matrix} \qquad \boldsymbol{S} = \begin{pmatrix} 1096.03 & 139.90 \\ 139.90 & 31.55 \end{pmatrix}$$

Plot of the Data



Plot of the Data: Cases Connected



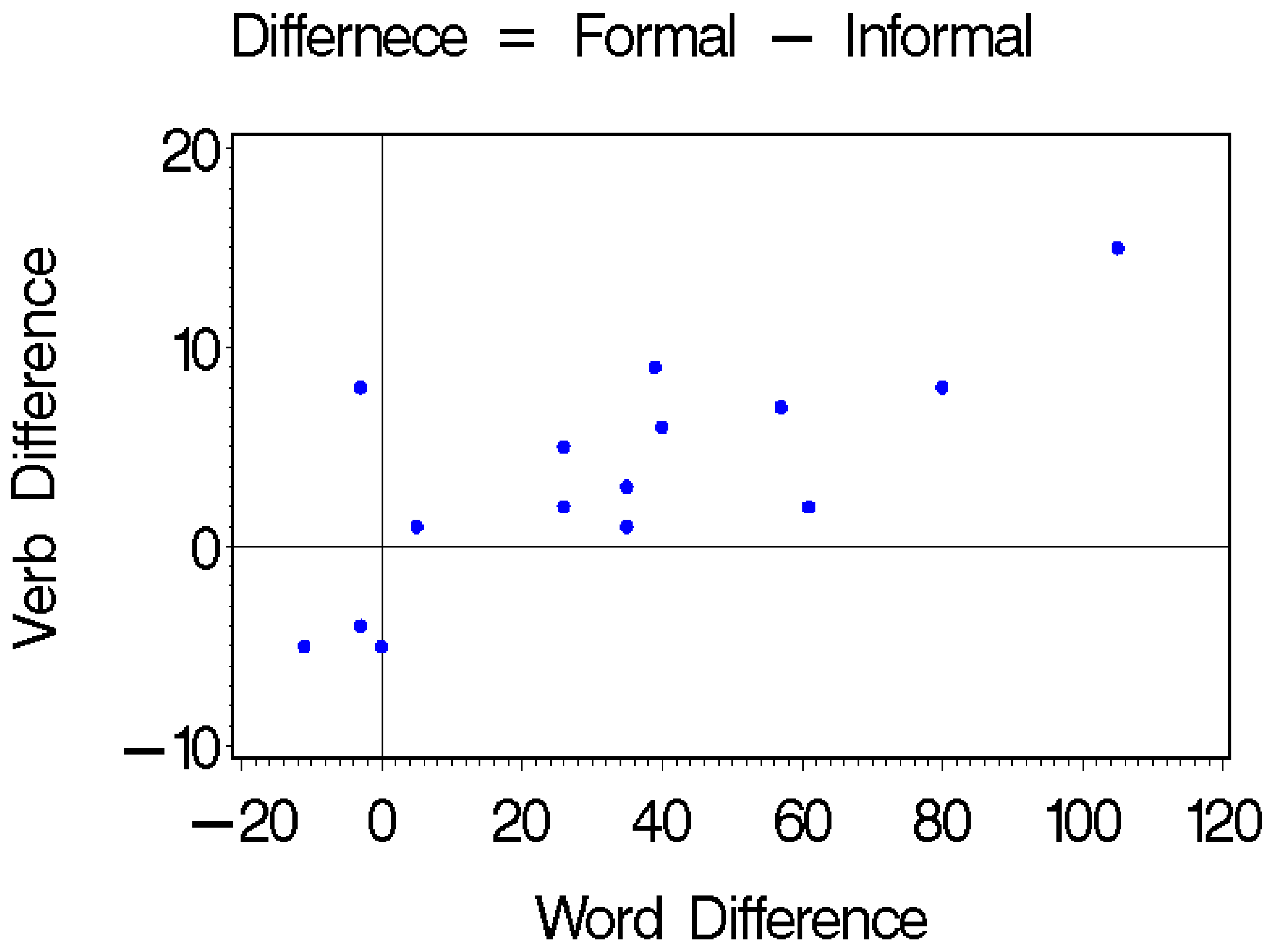
- Outline

Paired Comparisons

- Paired Comparisons (dependent samples)
- Univariate Case
- Advantage
- Multivariate Situation
- Needed for Statistical Inference
- Statistical Inference
- If you Reject $H_o : \delta = 0$
- Large Samples
- Example: The data
- Plot of the Data
- Plot of the Data: Cases Connected
- Plot of the Differences
- Example: Test
- 95% Confidence Region for δ
- 95% Confidence Region for the Mean
- SAS for the Last Figure
- Confidence Region, T^2 & Bonferroni Intervals
- Another way to calculate T^2
- Full Data Method for paired comparisons
- Full Data Method for paired comparisons
- Computations for Full Data

Comparisons of Two Means
Repeated Measures

Plot of the Differences



Example: Test

$$H_o : \boldsymbol{\delta} = \mathbf{0} \qquad \text{versus} \qquad H_A : \boldsymbol{\delta} \neq \mathbf{0}$$

(i.e., the number of words and verbs in informal and formal essays are the same).

$$\begin{aligned} T^2 &= 15 * (32.80, 3.53) \begin{pmatrix} 1096.03 & 139.90 \\ 139.90 & 31.55 \end{pmatrix}^{-1} \begin{pmatrix} 32.80 \\ 3.53 \end{pmatrix} \\ &= 15 * (32.80, 3.53) \begin{pmatrix} 0.0360156 \\ -0.047706 \end{pmatrix} \\ &= 15.191234 \end{aligned}$$

$$(14(2)/13)\mathcal{F}_{2,13}(.05) = 8.20$$

Alternatively, $(13)/((14)2)T^2 = 7.053$, which is distributed as $\mathcal{F}_{2,13}$, and has a p-value of $= .008$

Conclusion: Reject H_o. The data support the conclusion that the number of words and verbs in informal essays are not equal to the number in formal ones.

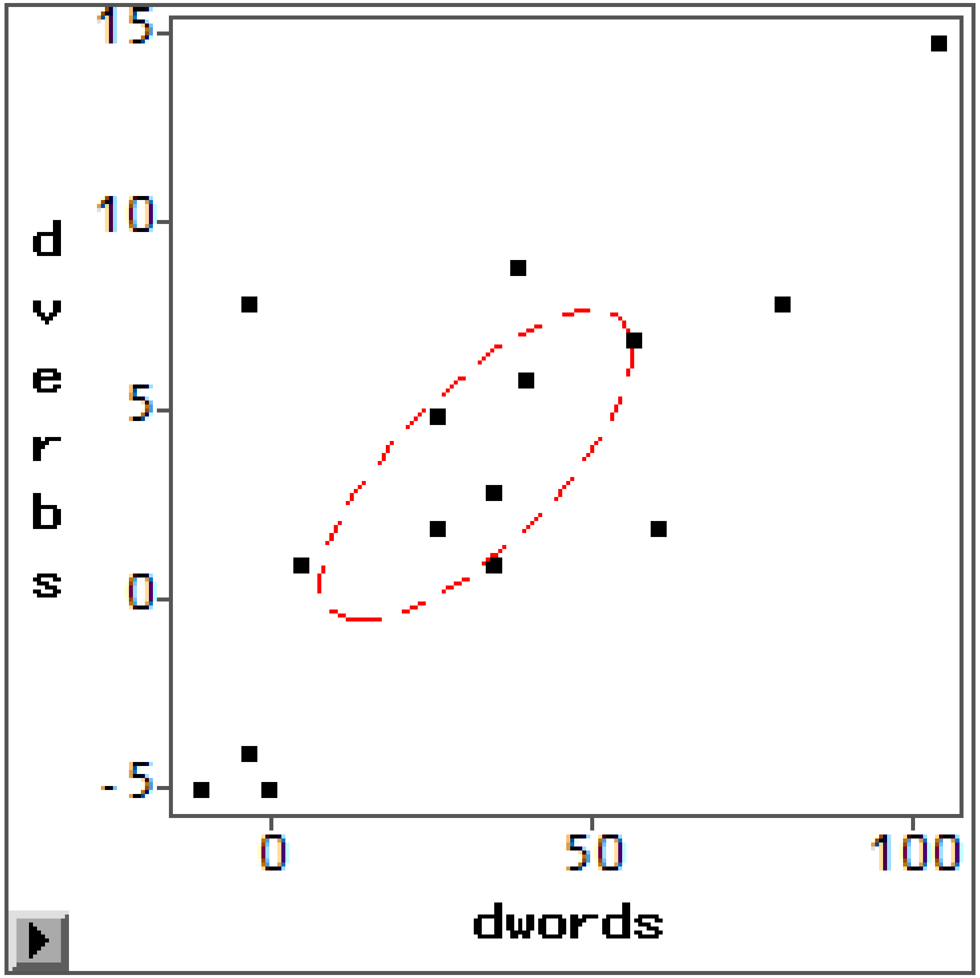
- Outline

Paired Comparisons
- Paired Comparisons (dependent samples)
- Univariate Case
- Advantage
- Multivariate Situation
- Needed for Statistical Inference
- Statistical Inference
- If you Reject $H_o : \delta = 0$
- Large Samples
- Example: The data
- Plot of the Data
- Plot of the Data: Cases Connected
- Plot of the Differences
- Example: Test
- 95% Confidence Region for δ
- 95% Confidence Region for the Mean
- SAS for the Last Figure
- Confidence Region, T^2 & Bonferroni Intervals
- Another way to calculate T^2
- Full Data Method for paired comparisons
- Full Data Method for paired comparisons
- Computations for Full Data

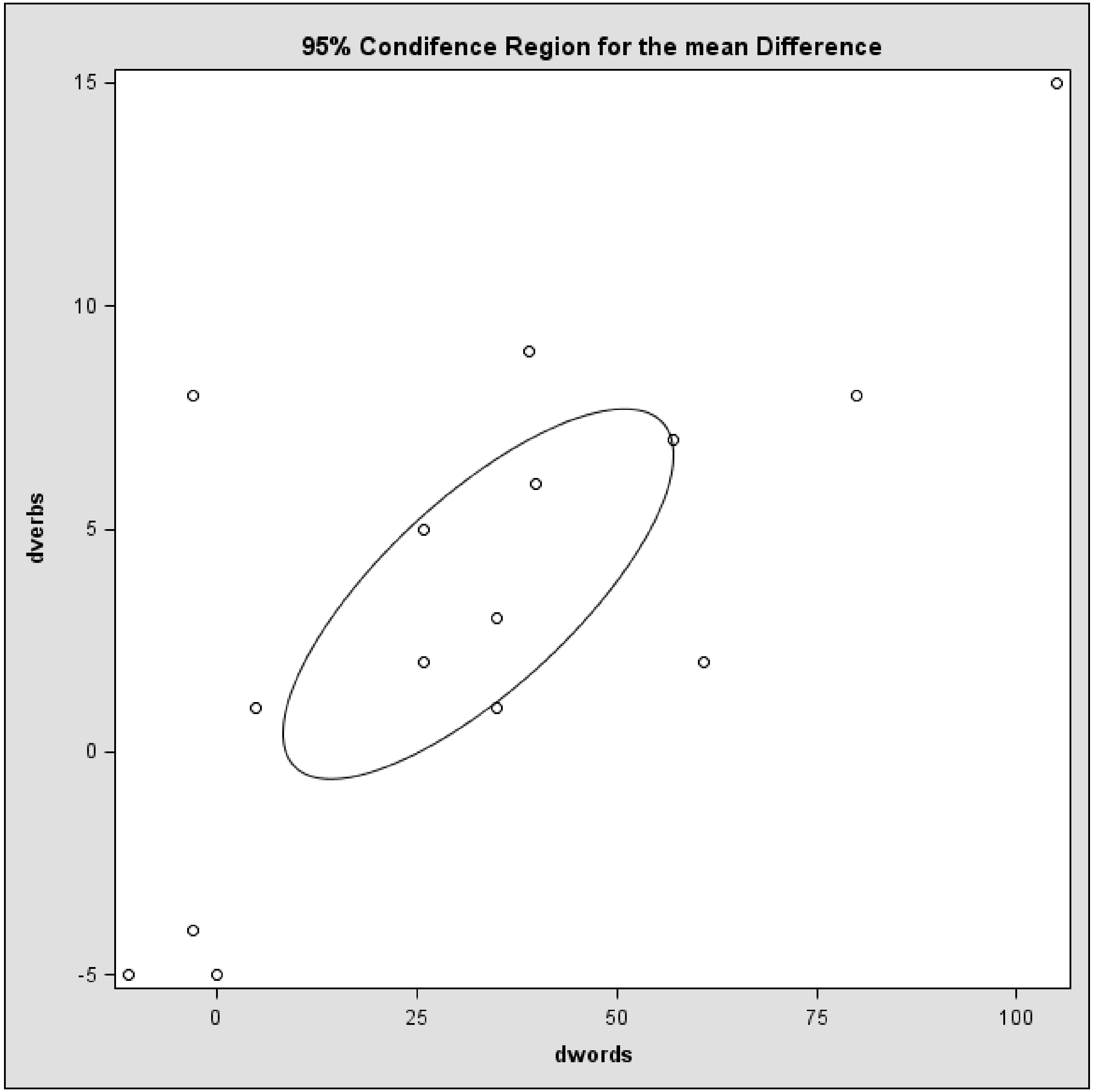
Comparisons of Two Means
Repeated Measures

95% Confidence Region for δ

From SAS>Solutions>Interactive Data Analysis
Analyze > Multivariate (scatter plot, curves)



95% Confidence Region for the Mean



SAS for the Last Figure

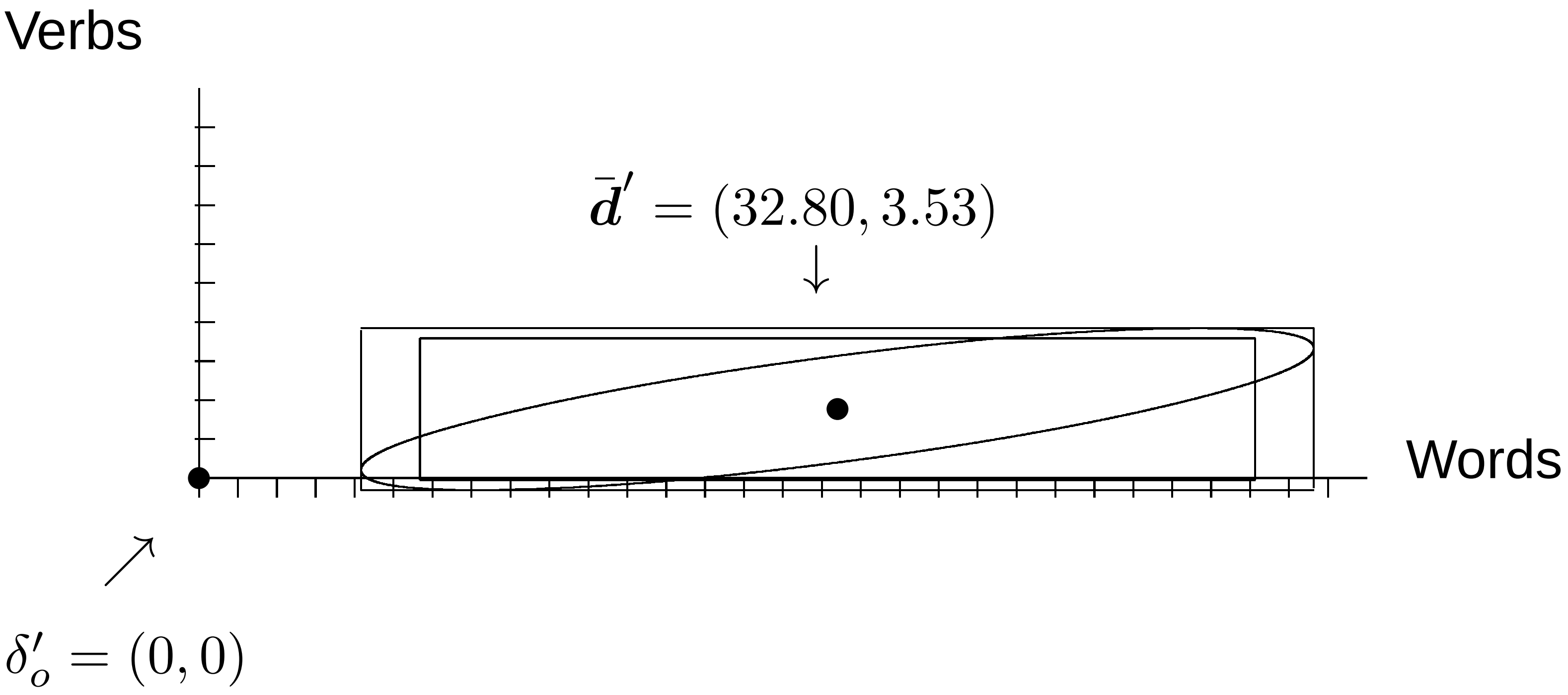
```
proc sgscatter data=essay;
    compare y= dverbs x= dwords
       / ellipse=(type=mean) ;
    title ’95% Confidence Region for the mean Difference’;
run;
```

- Outline

Paired Comparisons
- Paired Comparisons (dependent samples)
- Univariate Case
- Advantage
- Multivariate Situation
- Needed for Statistical Inference
- Statistical Inference
- If you Reject $H_o : \delta = 0$
- Large Samples
- Example: The data
- Plot of the Data
- Plot of the Data: Cases Connected
- Plot of the Differences
- Example: Test
- 95% Confidence Region for δ
- 95% Confidence Region for the Mean
- SAS for the Last Figure
- Confidence Region, T^2 & Bonferroni Intervals
- Another way to calculate T^2
- Full Data Method for paired comparisons
- Full Data Method for paired comparisons
- Computations for Full Data

Repeated Measures

Confidence Region, T^2 & Bonferroni Intervals



Another way to calculate T^2 for paired comparisons.

So far we've "divided the sample"; that is, $\boldsymbol{D} = \boldsymbol{X}_1 - \boldsymbol{X}_2$.

Now we'll consider a "Full Sample" method that considers every case as a pair and each with p measures on each member of the pair.

Pair or "Case" Number

Conditon	1	2	$\cdots$	j	$\cdots$	n
(a)	p variables	p variables	$\cdots$	p variables	$\cdots$	p variables
(b)	p variables	p variables	$\cdots$	p variables	$\cdots$	p variables

So we have $2p$ variables measured for each case (pair). In an experimental situation, the conditions are assumed to have been randomly assigned to members of the pairs.

Full Data Method for paired comparisons

Full Data Matrix:

$$\boldsymbol{X}_{n\times 2p} = \left(\begin{array}{cccc|cccc} X_{111} & X_{112} & \cdots & X_{11p} & X_{121} & X_{122} & \cdots & X_{12p} \\ X_{211} & X_{212} & \cdots & X_{21p} & X_{221} & X_{222} & \cdots & X_{22p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n11} & X_{n12} & \cdots & X_{n1p} & X_{n21} & X_{n22} & \cdots & X_{n2p} \end{array}\right)$$

$$= (\underbrace{\boldsymbol{X}_1}_{n\times p} \,|\, \underbrace{\boldsymbol{X}_2}_{n\times p})$$

Full Sample Mean Vector:

$$\boldsymbol{X}' = (\bar{X}_{11}, \bar{X}_{12}, \ldots, \bar{X}_{1p} | \bar{X}_{21}, \ldots, \bar{X}_{2p}) = (\bar{\boldsymbol{X}}'_1 | \bar{\boldsymbol{X}}'_2)$$

- Outline

Paired Comparisons
- Paired Comparisons (dependent samples)
- Univariate Case
- Advantage
- Multivariate Situation
- Needed for Statistical Inference
- Statistical Inference
- If you Reject $H_o : \delta = 0$
- Large Samples
- Example: The data
- Plot of the Data
- Plot of the Data: Cases Connected
- Plot of the Differences
- Example: Test
- 95% Confidence Region for δ
- 95% Confidence Region for the Mean
- SAS for the Last Figure
- Confidence Region, T^2 & Bonferroni Intervals
- Another way to calculate T^2
- Full Data Method for paired comparisons
- Full Data Method for paired comparisons
- Computations for Full Data

Repeated Measures

Full Data Method for paired comparisons

Full Data Sample Covariance Matrix:

$$\boldsymbol{S}_{2p\times 2p} = \left(\begin{array}{c|c} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \hline \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{array}\right)$$

where

$\boldsymbol{S}_{11}$ is the $(p \times p)$ covariance matrix for $\boldsymbol{X}_1$

$\boldsymbol{S}_{22}$ is the $(p \times p)$ covariance matrix for $\boldsymbol{X}_2$

$\boldsymbol{S}_{12} = \boldsymbol{S}'_{21}$ is the $(p \times p)$ covariance matrix between $\boldsymbol{X}_1$ & $\boldsymbol{X}_2$.

Define a Contrast Matrix:

$$\begin{aligned} C_{p\times 2p} &= \left(\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & -1 \end{array}\right) \\ &= (I_{p\times p} | - I_{p\times p}) \end{aligned}$$

What condition do you need to have a “contrast matrix”?

Computations for Full Data

Let

$$\boldsymbol{x}_{j,(2p\times 1)} = j^{th} \text{ row of } \boldsymbol{X}_{(n\times 2p)} \text{ written as a column vector.}$$

$$\boldsymbol{d}_j = \boldsymbol{C}\boldsymbol{x}_j$$

$$\bar{d} = \boldsymbol{C}\bar{x} = \boldsymbol{C}((1/n)\sum_{j=1}^{n} \boldsymbol{x}_j)$$

Putting all of this together yields

$$\begin{aligned} T^2 &= n(\boldsymbol{C}\bar{x})'(\boldsymbol{CSC}')^{-1}(\boldsymbol{C}\bar{\boldsymbol{x}}) \\ &= n\bar{x}'\boldsymbol{C}'(\boldsymbol{CSC}')^{-1}\boldsymbol{C}\bar{\boldsymbol{x}} \end{aligned}$$

With this method, we don't have to split the data set and compute the differences.

We'll see more uses of contrast matrices. . . relatively soon.

SAS/IML code for essay example.

Repeated Measures

for comparing conditions (treatments, etc).

- This is another generalization of univariate paired t–test.
- Situation: q conditions are compared with respect to one response variable.

  Each case receives each treatment once over successive periods of time. The order of the treatments should be randomized (& counterbalanced if possible).
- Example from Cochran & Cox (1957) (I got this from Timm 1980): There are four calculator designs and each person does specified computations. Their speed is recorded for each of the four calculators. The order of the calculator use was randomly assigned.
- This is Repeated measures because each case (person) gets each treatment (calculator). . . we have repeated observations or measurements on each case.

Repeated Measures

- Let the j^{th} observation equal

$$\boldsymbol{x}_j = \begin{pmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jq} \end{pmatrix} \qquad j = 1, \ldots, n$$

where $x_{ji} =$ response or measurement of the i^{th} treatment on the j^{th} case.

- Question (hypothesis): Is there a treatment effect?

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_q$$

versus

$$H_A : \text{ Not } H_o$$

This is the same hypothesis test in univariate, repeated measures ANOVA.

Repeated Measures as a Multivariate Test

- To test this as a multivariate mean vector, we need to use contrasts of the components of $\boldsymbol{\mu}$,

$$\boldsymbol{\mu} = E(\boldsymbol{x}_j) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}$$

- Assume $\boldsymbol{X}_j \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Set up a contrast

$$\underbrace{\begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_2 \\ \vdots \\ \mu_1 - \mu_q \end{pmatrix}}_{(q-1)\times 1} = \underbrace{\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}}_{(q-1)\times q} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}}_{q\times 1} = \boldsymbol{C}_1\boldsymbol{\mu}$$

- So $H_o : \boldsymbol{C}_1\boldsymbol{\mu} = \mathbf{0}$. (no treatment effect).

Contrast Matrices

- Any contrast matrix of size $(q-1)\times q$ will do.

- For example,

$$C_2\boldsymbol{\mu} = \underbrace{\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}}_{(q-1)\times q} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}}_{q\times 1} = \begin{pmatrix} \mu_1-\mu_2 \\ \mu_2-\mu_3 \\ \vdots \\ \mu_{q-1}-\mu_q \end{pmatrix}$$

- To be a contrast matrix,
  - The rows are linearly independent.
  - Each row is a contrast vector.

Hypothesis and Test for Repeated Measures

The hypothesis of no effects due to treatment in a repeated measures design

$$H_o : \mu_1 = \mu_2 = \cdots \mu_q$$

is the same as performing Hotelling's T^2 of

$$H_o : \boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{0}$$

where $\boldsymbol{C}$ is a $(q-1) \times q$ contrast matrix

Given data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ and a contrast matrix $\boldsymbol{C}$, the T^2 test statistic equals

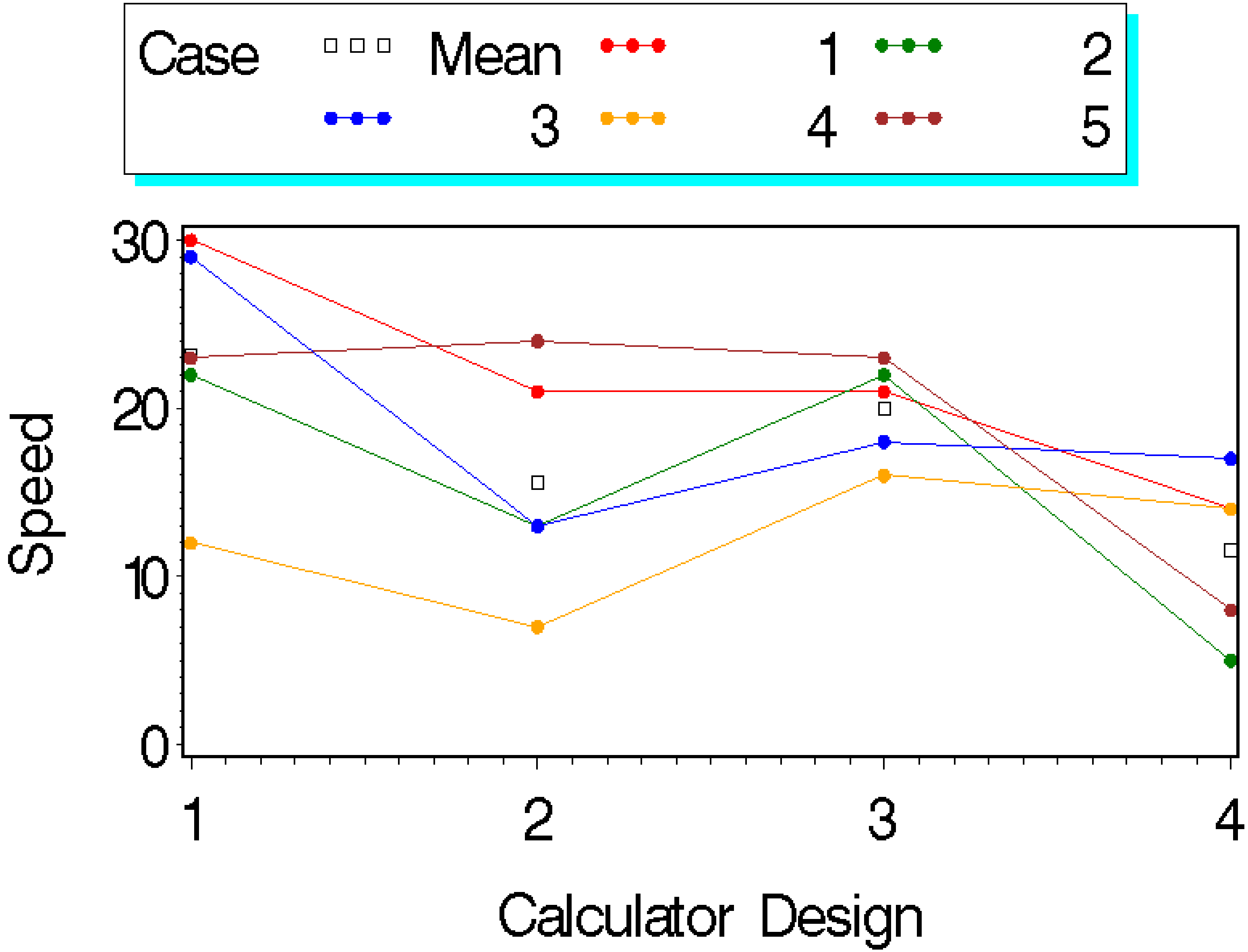
$$T^2 = n\boldsymbol{C}'\bar{\boldsymbol{x}}(\boldsymbol{CSC}')^{-1}\boldsymbol{C}\bar{\boldsymbol{x}}$$

Reject H_o if

$$T^2 > \frac{(n-1)(q-1)}{n-q+1}\mathcal{F}_{(q-1),(n-q+1)}(\alpha)$$

Now for our example. . . Plot data and then SAS/IML

(Scatter) Plot of the Calculator Data



Input 1 from SAS/IML

```
proc iml;
* A Module that computes Hotellings Tˆ2 for one sample tests;

start Tsq(X,muo,Ts,pvalue);
 n=nrow(X);
 one=j(n,1);
 Xbar = X‘*one/n;
 XbarM = one*Xbar‘;
 S=(X - XbarM)‘*(X - XbarM)/(n-1);
 Ts=n*(xbar-muo)‘*inv(S)*(xbar-muo);
 p=ncol(X);
 dfden=n-1;
 F=((n-1)*p/(n-p))*Ts;
 pvalue = 1 - cdf(’F’,F,p,dfden);
finish Tsq;
```

Input continued

```
X={ 30 21 21 14,
    22 13 22  5,
    29 13 18 17,
    12  7 16 14,
    23 24 23  8};

C1={1 -1 0 0,
    0 1 -1 0,
    0 0 1 -1};

muo={0, 0, 0};

X1 = X*C1‘;
run stats(X1,n1,Xbar1,W1,S1);
run Tsq(X1,muo,Tsq1,pvalue1);
```

Output 1 from SAS/IML

```
Data matrix (5 subjects x 4 variables) =
                   X

       30        21        21        14
       22        13        22         5
       29        13        18        17
       12         7        16        14
       23        24        23         8
                  C1


Using C1:         1        -1         0         0
                  0         1        -1         0
                  0         0         1        -1
```

Output 1 continued

```
X*C1' =          9         0         7
                 9        -9        17
                16        -5         1
                 5        -9         2
                -1         1        15
                            XBAR1

mean of C1*X1 =          7.6
                        -4.4
                         8.4
                                     TSQ1                       PVALUE1

T^2 for C1*mu=0  ---->   29.736051  with p-value =  0.0001029
```

Using Contrast Matrix 2

```
C2={1 0 0 -1,
    0 1 0 -1,
    0 0 1 -1};
X2 = X*C2‘;
run stats(X2,n2,Xbar2,W2,S2);
run Tsq(X2,muo,Tsq2,pvalue2);
```

(Partial) Output from this:

```
                     XBAR2
mean of C2*X2 =       11.6
                         4
                       8.4
                              TSQ2                    PVALUE2

T^2 for C2*mu=0  ---->   29.736051  with p-value =  0.0001029
```

With different contrast matrices, we get different $C\bar{x}$ vectors, but T^2, p–value, and conclusions are exactly the same.

T^2 and Repeated Measures

As before...

- $100(1-\alpha)\%$ Confidence region which consists of all $\boldsymbol{C\mu}$'s such that

$$n(\boldsymbol{C\bar{x}}-\boldsymbol{C\mu})'(\boldsymbol{CSC'})^{-1}(\boldsymbol{C\bar{x}}-\boldsymbol{C\mu}) \le \frac{(n-1)(q-1)}{(n-q+1)}\mathcal{F}_{(q-1),(n-q+1)}(\alpha)$$

- And Simultaneous T^2 intervals for a single contrast $\boldsymbol{c}_i'\boldsymbol{\bar{x}}$ where $\boldsymbol{c}_i'$ is the i^{th} row of matrix $\boldsymbol{C}$,

$$\boldsymbol{c}_i'\boldsymbol{\bar{x}} \pm \underbrace{\sqrt{\frac{(n-1)(q-1)}{(n-q+1)}\mathcal{F}_{(q-1),(n-q+1)}(\alpha)}}\sqrt{\frac{\boldsymbol{c}_i'\boldsymbol{S}\boldsymbol{c}_i}{n}}$$

- For Bonferroni (or one-at-time) confidence intervals, replace statistic above the brace by appropriate value from the t_{n-1} distribution.

- For large n, can use χ^2_{q-1}.

Repeated Measures ANOVA vs multivariate T^2

- The multivariate T^2 is appropriate for situations where we cannot assume that the covariance matrix for $\boldsymbol{X}$ has a particular structure.
- With repeated measures ANOVA you must assume that $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ has a special structure, in particular spherical,

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \begin{pmatrix} \sigma^2 & \tau & \cdots & \tau \\ \tau & \sigma^2 & \cdots & \tau \\ \vdots & \vdots & \ddots & \vdots \\ \tau & \tau & \cdots & \sigma^2 \end{pmatrix}$$

Unlikely but this works too: $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$.

- If the assumptions on the structure of $\boldsymbol{\Sigma}$ are met, then repeated measures ANOVA is more powerful than multivariate T^2 because the repeated measures ANOVA takes the structure of $\boldsymbol{\Sigma}$ into account.
- If assumptions on $\boldsymbol{\Sigma}$ not met, T^2 is still valid but not repeated measures ANOVA.

Two Independent Samples

Situation: Two samples, each having p measurements where we have a random sample of size n_1 from population 1 and a random sample of size n_2 from population 2.

$$\overbrace{\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}}^{\text{Sample from population 1}} \qquad \overbrace{\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}}^{\text{Sample from population 2}}$$

Sample Means

$$\bar{\boldsymbol{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \boldsymbol{x}_{1j} \qquad \bar{\boldsymbol{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_1} \boldsymbol{x}_{2j}$$

Sample Covariance matrices

$$\boldsymbol{S}_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} (\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)' \quad \boldsymbol{S}_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2} (\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)'$$

Hypothesis: $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

Assumptions

1. The sample $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ is a random sample of size n_1 from a p–variate population with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$.
2. The sample $\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_1}$ is a random sample of size n_2 from a p–variate population with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$.
3. The samples are (statistically) independent of each other. These assumptions are required when we want to test

$$H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \qquad \text{or equivalently} \qquad \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$$

$$H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \qquad \text{or equivalently} \qquad \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq 0$$

If n_1 and/or n_2 are small, then we must make two additional assumptions:

4. Both populations are multivariate normal.
5. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ This is a very strong assumption (stronger than univariate case).

Case 1: Known Σ_1 and Σ_2

To develop the test for independent populations, we'll start with supposing that we know $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ (i.e., we don't have to estimate them) and assume first 4 assumptions made on previous slide.

The test statistic would be

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \left(\frac{1}{n_1}\mathbf{\Sigma}_1 + \frac{1}{n_2}\mathbf{\Sigma}_2 \right)^{-1} (\bar{\boldsymbol{x}}_1 = \bar{\boldsymbol{x}}_2) \sim \chi^2_p$$

because

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \sim \mathcal{N}_p \left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \frac{1}{n_1}\mathbf{\Sigma}_1 + \frac{1}{n_2}\mathbf{\Sigma}_2 \right)$$

Why is $(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ multivariate normal?

When H_o is true, then $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ and the test statistic should be "small".

Case 2: Σ_1 and Σ_2 Unknown

$\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ must be estimated.

For this more realistic case, we must also assume

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$$

Since $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, we will estimate $\boldsymbol{\Sigma}$ by pooling the data from the two samples:

$$\begin{aligned} \boldsymbol{S}_{pool} &= \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2} \\ &= \frac{\sum_{j=1}^{n_1}(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)' + \sum_{j=1}^{n_2}(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)'}{n_1+n_2-2} \end{aligned}$$

$\boldsymbol{S}_{pool}$ is an estimator of $\boldsymbol{\Sigma}$ with $df = n_1 + n_2 - 2$.

Distribution of Linear Combination

Consider the linear combination of two random vectors $\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2$

$$E(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2} &= \text{cov}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \\
&= \text{cov}(\bar{\boldsymbol{x}}_1) + \text{cov}(\bar{\boldsymbol{x}}_x) \qquad \leftarrow \text{ independent samples} \\
&= \frac{1}{n_1}\boldsymbol{\Sigma} + \frac{1}{n_2}\boldsymbol{\Sigma} \\
&= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}
\end{aligned}$$

which is estimated by $(\frac{1}{n_1} + \frac{1}{n_2})\boldsymbol{S}_{pool}$.

When $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}$ is a random sample of size n_1 from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}$ is a random sample of size n_2 from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ then the test statistic for $H_o : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_o$

$$T^2 = ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}_o)' \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{S}_{pool}\right)^{-1} ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}_o)$$

Distribution of Test Statistic

- Outline

Paired Comparisons

Repeated Measures

Two Independent Samples
- Two Independent Samples
- Assumptions
- Case 1: Known Σ_1 and Σ_2
- Case 2: Σ_1 and Σ_2 Unknown
- Distribution of Linear Combination
- Distribution of Test Statistic
- Example: Two Independent Samples T^2
- Example continued
- $100(1-\alpha)\%$ Confidence Region for $\mu_1 - \mu_2$
- Example:Confidence Region
- Computing the Axes of the Ellipse
- Figure of 95% Confidence Region
- Simultaneous T^2 Intervals
- Simultaneous T^2 continued
- Example: Simultaneous T^2 intervals
- Bonferroni and One-at-a-Time Intervals
- Example: Simultaneous T^2 and Bonferroni

The test statistic

$$T^2 = ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}_o)' \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{pool}\right)^{-1} ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}_o)$$

has a sampling distribution that is

$$\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} \mathcal{F}_{p,(n_1+n_2-p1-1)}$$

or we could just refer

$$\frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p} T^2 \qquad \text{to} \qquad \mathcal{F}_{p,(n_1+n_2-p1-1)}$$

Note:

$$\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{pool}\right)^{-1} = \left(\left(\frac{n_1+n_2}{n_1 n_2} S_{pool}\right)\right)^{-1} = \frac{n_1 n_2}{n_1+n_2} (S_{pool})^{-1}$$

So sometimes you'll see

$$T^2 = \frac{n_1 n_2}{n_1+n_2} ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}_o)' \boldsymbol{S}_{pool}^{-1} ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}_o)$$

Example: Two Independent Samples T^2

From Johnson & Wichern: Wisconsin homeowners without airconditioning ($n_1 = 45$) and those with airconditioning ($n_2 = 55$).

$X_1 =$ total on-peak consumption of electricity July 1977 (in kilowatts)
$X_2 =$ total off-peak consumption of electricity July 1977(in kilowatts)

$$\bar{\boldsymbol{x}}_1 = (204.4, 556.6)' \qquad \bar{\boldsymbol{x}}_2 = (130.0, 355.0)'$$

and $$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = (74.4, 201.6)$$

$$\boldsymbol{S}_1 = \begin{pmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{pmatrix} \qquad \boldsymbol{S}_2 = \begin{pmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{pmatrix}$$

$$\boldsymbol{S}_{pool} = \frac{44\boldsymbol{S}_1 + 54\boldsymbol{S}_2}{98} = \begin{pmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{pmatrix}$$

Example continued

The estimated covariance matrix of $(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ is

$$
\begin{aligned}
\boldsymbol{S}_{\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2} &= (\frac{1}{n_1} + \frac{1}{n_2})\boldsymbol{S}_{pool} \\
&= (\frac{1}{45} + \frac{1}{55})\begin{pmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{pmatrix} \\
&= \begin{pmatrix} 442.98 & 868.91 \\ 868.91 & 2572.12 \end{pmatrix}
\end{aligned}
$$

To test $H_o : \boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{0}$, compute test statistic

$$
\begin{aligned}
(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}_{\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) &= (74, 201.6)\begin{pmatrix} 442.98 & 868.91 \\ 868.91 & 2572.12 \end{pmatrix}^{-1}\begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} \\
&= 16.06
\end{aligned}
$$

For $\alpha = .05$: $(98(2)/97)\mathcal{F}_{2,97}(.05) = 2.02(3.1) = 6.24$.

Conclusion. . .

$100(1-\alpha)\%$ Confidence Region for $\mu_1 - \mu_2$

Is the set of all $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$'s such that

$$\frac{n_1 n_2}{n_1 + n_2}((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta})' \boldsymbol{S}_{pool}^{-1}((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}) \le c^2$$

where

$$c^2 = \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} \mathcal{F}_{p,(n_1+n_2-p-1)}(\alpha)$$

To study the ellipsoid, we can focus on the eigenvalues and eigenvectors of $\boldsymbol{S}_{pool}$.

The axes of the ellipsoid are

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \pm \sqrt{\lambda_i} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2} \boldsymbol{e}_i \qquad i = 1, \ldots, p$$

where λ_i and e_i are the eigenvalues and eigenvectors of $\boldsymbol{S}_{pool}$.

Example:Confidence Region

The 95% Confidence Region (Ellipse):

The set of all possible $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ that satisfy the following equation:

$$((74.4-\delta_1),(201.6-\delta_2))\begin{pmatrix} 442.98 & 868.91 \\ 868.91 & 2572.12 \end{pmatrix}^{-1}\begin{pmatrix} (74.4-\delta_1) \\ (201.6-\delta_2) \end{pmatrix} \leq c^2$$

where $c^2 = (98(2)/97)\mathcal{F}_{2,97}(.05) = 2.02(3.1) = 6.26$.

Eigenvalues and Eigenvectors of $\boldsymbol{S}_{pool}$ are

$$\lambda_1 = 71323.426, \quad \boldsymbol{e}_1 = \begin{pmatrix} 0.3356 \\ 0.9420 \end{pmatrix}$$

and

$$\lambda_2 = 3301.572, \quad \boldsymbol{e}_2 = \begin{pmatrix} 0.9420 \\ -0.3356 \end{pmatrix}$$

Computing the Axes of the Ellipse

Major axis

$$\begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} \pm \sqrt{\lambda_1}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)c^2}\boldsymbol{e}_1$$

$$\pm \sqrt{71323.426}\sqrt{\left(\frac{1}{45}+\frac{1}{55}\right)6.2441}\begin{pmatrix} 0.3356 \\ 0.9420 \end{pmatrix}$$

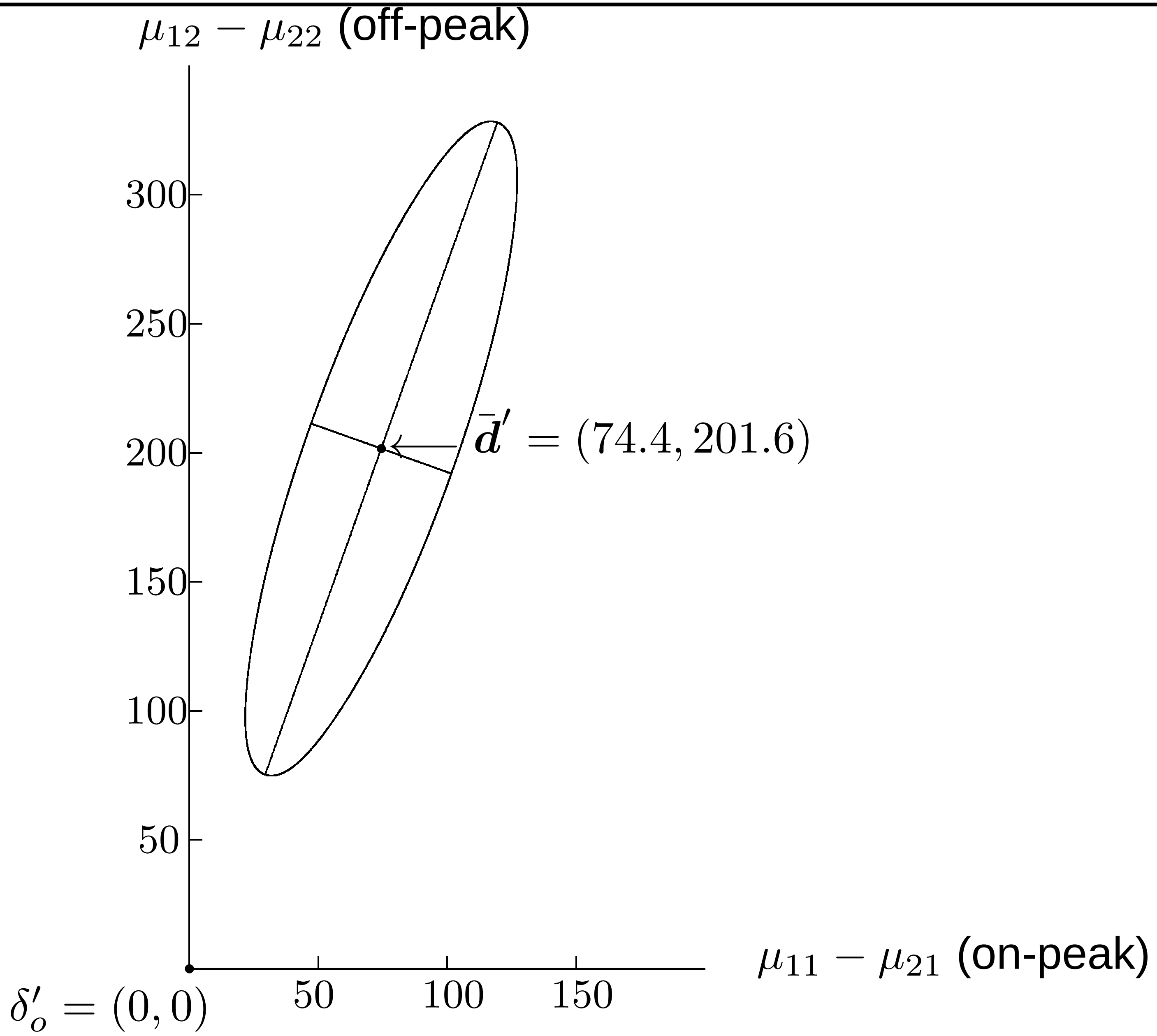
$$\rightarrow \begin{pmatrix} 29.38 \\ 75.24 \end{pmatrix}, \begin{pmatrix} 119.42 \\ 327.96 \end{pmatrix}$$

Minor axis

$$\begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} \pm \sqrt{3301.572}\sqrt{\left(\frac{1}{45}+\frac{1}{55}\right)6.2441}\begin{pmatrix} 0.9420 \\ -0.3356 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 47.21 \\ 211.29 \end{pmatrix}, \begin{pmatrix} 101.59 \\ 191.91 \end{pmatrix}$$

Figure of 95% Confidence Region



Simultaneous T^2 Intervals

Let

$$c^2 = \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} \mathcal{F}_{p,(n_1+n_2-p-1)}(\alpha)$$

With “confidence $100(1-\alpha)\%$”

$$\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \pm c\sqrt{\boldsymbol{a}'\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{S}_{pool}\boldsymbol{a}}$$

will cover $\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for all possible $\boldsymbol{a}$.

By appropriate choices for $\boldsymbol{a}$, we can get component intervals:

$$\boldsymbol{a}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \boldsymbol{a}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \qquad \ldots, \qquad \boldsymbol{a}_p = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Simultaneous T^2 continued

So the component intervals are

$$\begin{array}{ccc}
(\bar{x}_{11} - \bar{x}_{21}) & \pm & c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{pool,11}} \\
(\bar{x}_{12} - \bar{x}_{22}) & \pm & c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{pool,22}} \\
\vdots & \vdots & \vdots \\
(\bar{x}_{1p} - \bar{x}_{2p}) & \pm & c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{pool,pp}}
\end{array}$$

where

$$c = \sqrt{\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} \mathcal{F}_{p,(n_1+n_2-p-1)}(\alpha)}$$

Example: Simultaneous T^2 intervals

Consider the linear combination vectors:

$$\boldsymbol{a}_1 = (1,0)' \qquad \text{So} \qquad \boldsymbol{a}_1'\boldsymbol{\delta} = \boldsymbol{a}_1'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mu_{11} - \mu_{21} = \delta_1$$

and

$$\boldsymbol{a}_2 = (0,1)' \qquad \text{So} \qquad \boldsymbol{a}_2'\boldsymbol{\delta} = \boldsymbol{a}_2'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mu_{12} - \mu_{22} = \delta_2$$

Using these we get the intervals for on-peak

$$74.4 \pm (2.502)\sqrt{442.98} \longrightarrow 21.81 \le \delta_1 \le 126.99$$

and for off-peak

$$201.6 \pm (2.502)\sqrt{2572.12} \longrightarrow 74.87 \le \delta_2 \le 328.33$$

Note: $\sqrt{c^2} = \sqrt{6.26} = 2.502$

Bonferroni and One-at-a-Time Intervals

For Bonferroni and One-at-a-Time (i.e., univariate method) intervals, you simply need to change the value of c.

Bonferroni

$$c = t_{n_1+n_2-2}(\alpha/2m)$$

where $m =$ number of intervals formed (probably p, but no more). These should be planned *a priori*.

One-at-a-Time

$$c = t_{n_1+n_2-2}(\alpha/2)$$

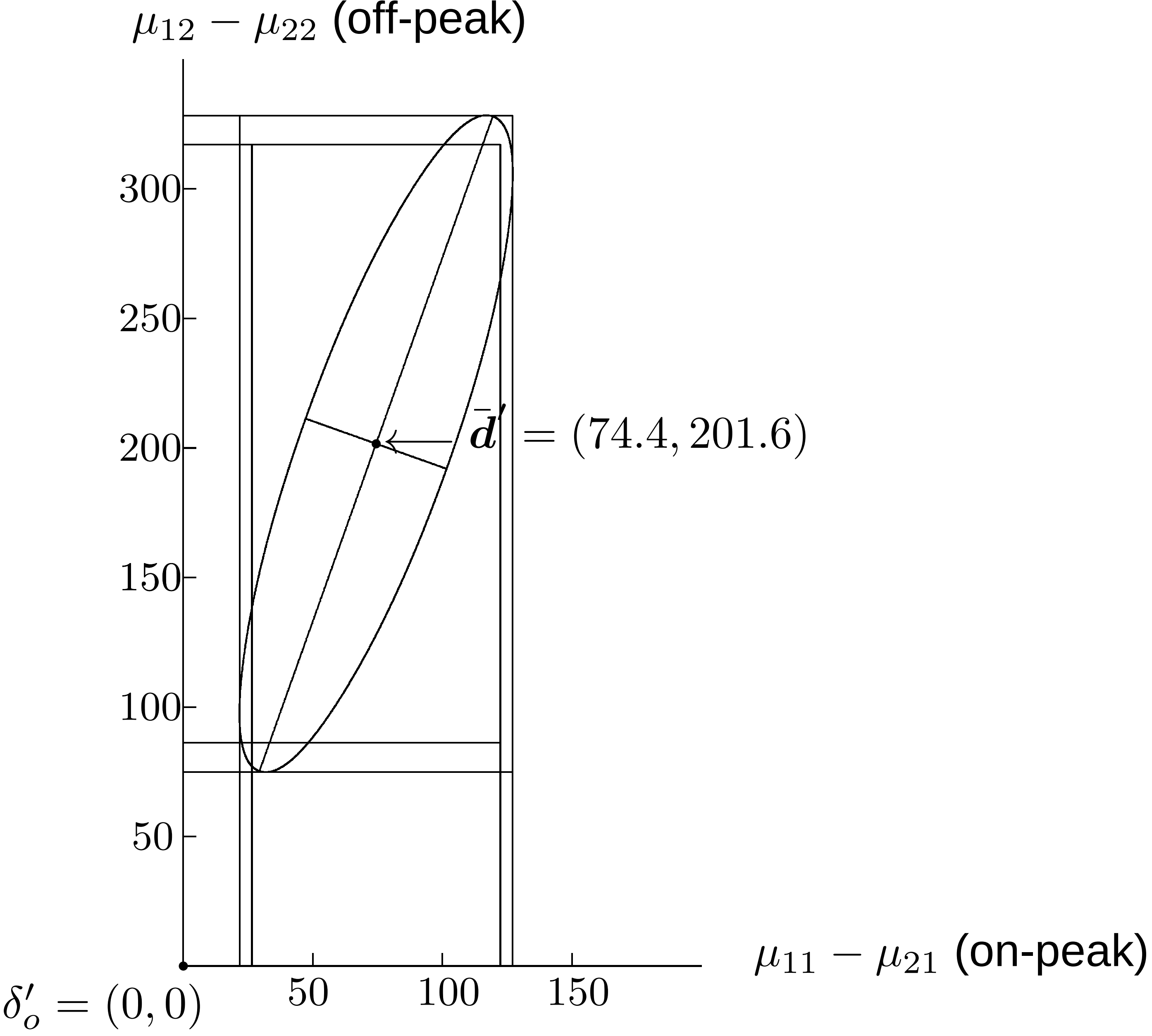
Example: Simultaneous T^2 and Bonferroni

- Outline

Paired Comparisons

Repeated Measures

Two Independent Samples
- Two Independent Samples
- Assumptions
- Case 1: Known Σ_1 and Σ_2
- Case 2: Σ_1 and Σ_2 Unknown
- Distribution of Linear Combination
- Distribution of Test Statistic
- Example: Two Independent Samples T^2
- Example continued
- $100(1-\alpha)\%$ Confidence Region for $\mu_1 - \mu_2$
- Example:Confidence Region
- Computing the Axes of the Ellipse
- Figure of 95% Confidence Region
- Simultaneous T^2 Intervals
- Simultaneous T^2 continued
- Example: Simultaneous T^2 intervals
- Bonferroni and One-at-a-Time Intervals
- Example: Simultaneous T^2 and Bonferroni



Case 3: Large $n_1 - p$ and $n_2 - p$

If $n_1 - p$ and $n_2 - p$ are large, then we do NOT need to assume:

- $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.
- $\boldsymbol{x}_{1j} \sim$ multivariate normal.
- $\boldsymbol{x}_{2j} \sim$ multivariate normal.

We do need to assume that

- Observations between populations are independent.
- $\boldsymbol{x}_{11}, \ldots \boldsymbol{x}_{1,n_1}$ are a random sample from population 1 with $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$.
- $\boldsymbol{x}_{21}, \ldots \boldsymbol{x}_{2,n_2}$ are a random sample from population 2 with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$.

If $n_1 - p$ and $n_2 - p$ are large, then an approximate sampling distribution for the test statistic T^2 is χ^2_p.

Large Sample Case

- To test Estimate the covariance matrix of the differences $\boldsymbol{\Sigma}_{\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2}\ldots$ remember case 1?

$$\begin{aligned}\boldsymbol{\Sigma}_{\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2} &= \boldsymbol{\Sigma}_{\bar{\boldsymbol{x}}_1}+\boldsymbol{\Sigma}_{\bar{\boldsymbol{x}}_2}\\ &= \frac{1}{n_1}\boldsymbol{\Sigma}_1+\frac{1}{n_2}\boldsymbol{\Sigma}_2\end{aligned}$$

which we can estimate using

$$\frac{1}{n_1}\boldsymbol{S}_1+\frac{1}{n_2}\boldsymbol{S}_2$$

- Test statistic for $H_o : \boldsymbol{\mu}_1-\boldsymbol{\mu}_2=\boldsymbol{\delta}_o$

$$T^2=((\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)-\boldsymbol{\delta}_o)'\left(\frac{1}{n_1}\boldsymbol{S}_1+\frac{1}{n_2}\boldsymbol{S}_2\right)^{-1}((\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)-\boldsymbol{\delta}_o)\sim\chi^2_p$$

Large Sample Case continued

- A $100(1-\alpha)\%$ Confidence region (ellipsoid) for $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is the set of all $\boldsymbol{\delta}$ that satisfy

$$((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta})' \left(\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right)^{-1} ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \boldsymbol{\delta}) \leq \chi^2_p(\alpha)$$

- For $100(1-\alpha)\%$ simultaneous χ^2 intervals

$$\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\boldsymbol{a}'\left(\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right)\boldsymbol{a}}$$

- Let’s try this for the air conditioner data. . .

Example using Large Sample

What if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$?

n_1 and n_2 may be large enough to use the large sample theory.

$$\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2 = \frac{1}{45}\begin{pmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{pmatrix} + \frac{1}{55}\begin{pmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{pmatrix}$$

$$= \begin{pmatrix} 464.17 & 886.08 \\ 886.08 & 2642.15 \end{pmatrix}$$

$$\left[\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right]^{-1} = \begin{pmatrix} 59.874 & -20.08 \\ -20.08 & 10.519 \end{pmatrix} \times 10^{-4}$$

Example: Large Sample Test Statistic

- Outline

Paired Comparisons

Repeated Measures

Two Independent Samples
- Two Independent Samples
- Assumptions
- Case 1: Known Σ_1 and Σ_2
- Case 2: Σ_1 and Σ_2 Unknown
- Distribution of Linear Combination
- Distribution of Test Statistic
- Example: Two Independent Samples T^2
- Example continued
- $100(1-\alpha)\%$ Confidence Region for $\mu_1 - \mu_2$
- Example:Confidence Region
- Computing the Axes of the Ellipse
- Figure of 95% Confidence Region
- Simultaneous T^2 Intervals
- Simultaneous T^2 continued
- Example: Simultaneous T^2 intervals
- Bonferroni and One-at-a-Time Intervals
- Example: Simultaneous T^2 and Bonferroni

Test $H_o : \boldsymbol{\delta} = \mathbf{0}$: Test statistic is

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \left[\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right]^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

$$= ((204.4-130.0),(556.6-355.0)) \begin{pmatrix} 59.874 & -20.08 \\ -20.08 & 10.519 \end{pmatrix} (10^{-4}) \begin{pmatrix} 204.4-130.0 \\ 556.6-355.0 \end{pmatrix}$$

$$= 15.66$$

which for $\alpha = .05$, the critical value from χ^2_p of 5.99 (the p-value $< .005$)

Compare this with $T^2 = 16.06$ using $\boldsymbol{S}_{pool}$ (where we assumed that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$).

Large Sample χ^2–Intervals

Using the same the linear combination vectors as above:

$$\boldsymbol{a}_1 = (1,0)' \quad \text{so} \quad \boldsymbol{a}_1'\boldsymbol{\delta} = \boldsymbol{a}_1'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mu_{11} - \mu_{21}$$

and

$$\boldsymbol{a}_2 = (0,1)' \quad \text{so} \quad \boldsymbol{a}_2'\boldsymbol{\delta} = \boldsymbol{a}_2'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mu_{12} - \mu_{22}$$

$$(204.4 - 130.0) \pm \sqrt{5.99}\sqrt{464.17} = (21.7, 127.1)$$

$$(556.6 - 355.0) \pm \sqrt{5.99}\sqrt{2642.15} = (75.8, 327.4)$$

which are very similar to the T^2 intervals given previously

Note: $\mathcal{X}_2^2(.05) = 5.99$

Sample Sample with $n_1 = n_2$

We obtained similar results in our large and small sample procedures; however, one possible reason stems from $n_1 \approx n_2$.

Note that when $n_1 = n_2 = n$

$$\frac{(n-1)}{n+n-2} = \frac{1}{2}$$

$$\begin{aligned}
\frac{1}{n}\boldsymbol{S}_1 + \frac{1}{n}\boldsymbol{S}_2 &= \frac{1}{n}(\boldsymbol{S}_1 + \boldsymbol{S}_2) = \underbrace{2\left(\frac{(n-1)}{n+n-2}\right)}_{=1}\frac{1}{n}(\boldsymbol{S}_1 + \boldsymbol{S}_2) \\
&= \frac{2}{n}\left(\frac{(n-1)S_1 + (n-1)\boldsymbol{S}_2}{n+n-2}\right) \\
&= \left(\frac{1}{n} + \frac{1}{n}\right) S_{pool}
\end{aligned}$$

This implies that with equal samples, the large sample procedure for computing an estimate of $\boldsymbol{\Sigma}_{\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2}$ is essentially the same as the procedure based on pooled covariance matrix.

Case 4: Small sample with $\Sigma_1 \neq \Sigma_2$

- Outline

Paired Comparisons

Repeated Measures

Two Independent Samples
- Two Independent Samples
- Assumptions
- Case 1: Known Σ_1 and Σ_2
- Case 2: Σ_1 and Σ_2 Unknown
- Distribution of Linear Combination
- Distribution of Test Statistic
- Example: Two Independent Samples T^2
- Example continued
- $100(1-\alpha)\%$ Confidence Region for $\mu_1 - \mu_2$
- Example:Confidence Region
- Computing the Axes of the Ellipse
- Figure of 95% Confidence Region
- Simultaneous T^2 Intervals
- Simultaneous T^2 continued
- Example: Simultaneous T^2 intervals
- Bonferroni and One-at-a-Time Intervals
- Example: Simultaneous T^2 and Bonferroni

We should consider whether $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is a reasonable assumption.

If $n_1 - p$ and $n_2 - p$ are small and $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, then there's no "nice" measure like T^2 whose distribution does not depend on $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

Rule-of-Thumb for when to worry about $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$:

Don't worry if ratios $\sigma_{1,ik}/\sigma_{2,ik} \leq 4$ (or $\sigma_{2,ik}/\sigma_{1,ik} \leq 4$).

Our air conditioner example:

$$\left.\begin{array}{lrl} (1,1) & 13825.3/8632.0 & = 1.60 \\ (1,2) & 23823.4/19616.7 & = 1.21 \\ (2,2) & 73107.4/55964.5 & = 1.31 \end{array}\right\} \text{all } \leq 4$$

Testing whether $\Sigma_1 = \Sigma_2$

We could use Bartlet's test, but this assumes

- Data are multivariate normal (not just that the means are multivariate normal).
- $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

So if you reject H_o (significant test statistics), it could be because

- $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$
- Data are not normal.
- Or both $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ and Data are not normal.

Additionally for a valid test you need large samples, but if you have large samples you don't need to assumed that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ (or normality of the data).

Revisiting Examining “Why”

- Our motivation for computing confidence intervals for components of mean vector was to come to conclusion about individual means.

- The simultaneous T^2 intervals hold for any $\boldsymbol{a}$.

- The $\boldsymbol{a}$ that leads to the largest population difference is proportional to

$$\boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = \boldsymbol{a}^*$$

- If null hypothesis using T^2 is rejected, then $\boldsymbol{a}^{*\prime}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ has the largest possible statistic

$$\boldsymbol{a}^{*\prime}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

  which is a multiple of T^2.

- $\boldsymbol{a}^*$ is useful for interpreting and describing why H_o was rejected.

Interpretation

- For the air conditioner data (using large sample), $\boldsymbol{a}^*$ is proportional to

$$(10^{-4})\begin{pmatrix} 59.874 & -20.080 \\ -20.080 & 10.519 \end{pmatrix}\begin{pmatrix} 74.4 \\ 201.6 \end{pmatrix} = \begin{pmatrix} .041 \\ .063 \end{pmatrix}$$

- So the difference in X_2 (off-peak consumption) contributes more ($.063 > .041$) to the rejection of $H_o : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ via T^2 test than X_1 (on-peak energy consumption).
- Note:

$$\boldsymbol{a}^{*\prime}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{pmatrix} .041(\mu_{11} - \mu_{21}) \\ .063(\mu_{12} - \mu_{22}) \end{pmatrix}$$

Summary regarding Inferences about μ

- Outline

Paired Comparisons

Repeated Measures

Two Independent Samples

Summary
- Summary regarding Inferences about μ
- Error Rates & More Reasons
- Reason 4
- A couple of final notes

Four reasons for taking a multivariate approach to hypothesis testing:

Reason 1:

If you do p univariate (t) tests, you have an inflated type I error rate (i.e., actual α larger than you want it to be).

With a multivariate test, the exact α level is under your control.

.g., If $p = 5$ and you perform p separate univariate tests all at $\alpha = .05$, then

Prob$\{$at least 1 false rejection$\}$ $=$ Prob$\{$at leat 1 Type I error$\}$ $> .05$

In the extreme case where all the variables are independent, if H_o is true

$$\begin{aligned}\text{Prob}\{\text{at least 1 false rejection}\} &= 1 - \text{Prob}\{\text{all } P \text{ retained}\}\\ &= 1-(1-\alpha)^p\end{aligned}$$

Error Rates & More Reasons

Overall error rates are somewhere between
For $p = 5 \implies .05$ and $.23$
For $p = 10 \implies .05$ and $.40$.

Reason 2:

Univariate tests ignore (completely) the correlations between the variables. Multivariate tests make direct use of the covariance matrix.

Reason 3:

Multivariate tests are more powerful (in most cases).
Sometimes all p univariate tests fail to reach significance, but multivariate test is significant because small effects combine to jointly indicate significance.

Note: For a given sample size, there is a limit to the number of variables a multivariate test can handle without losing power.

Reason 4

Many multivariate procedures and tests of mean have as a by-product the construction of a linear combination(s) of variables that reveals information about how the variables combine to lead to rejection of H_o.

A couple of final notes

- Mahalanoba's Generalized Distance

$$
\begin{aligned}
D^2 &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}_{pool}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \\
&= \left(\frac{1}{n_1} + \frac{1}{n_2} \right) T^2 \\
&= \left(\frac{n_1 + n_2}{n_1 n_2} \right) T^2
\end{aligned}
$$

This is the distance between two centroids in $\boldsymbol{S}_{pool}$ metric.

For large sample where $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, can use $(1/n_1)\boldsymbol{S}_1 + (1/n_2)\boldsymbol{S}_2$ to define the metric.

- The two-independent sample T^2 test generalizes to a g–sample test $\longrightarrow MANOVA$, which is also a generalization of univariate $ANOVA$ to multivariate situation.