Inference for Multivariate Means

Statistics 407, ISU

Inference for the Population Mean

This section focuses on the question:

Whether a pre-determined mean is a plausible value for the normal population mean, given the sample that we have collected.

In hypothesis testing language, this means we want to test

$$H_o : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad vs \quad H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ is the true population mean and $\boldsymbol{\mu}_0$ is the hypothesized mean.

Univariate Testing of a Mean

Recall, in the univariate situation, when we want to test

$$H_o : \mu = \mu_0 \quad vs \quad H_A : \mu \neq \mu_0$$

we calculate the *t*-statistic, as follows:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Then we would compare $|t|$ with values from a *t*-distribution with *(n-1)* degrees of freedom, and reject H_o only if $|t|$ is larger than the critical values.

Confidence Interval for a Mean

We could also compute a 100(1- α)% confidence interval for the population mean, μ, using

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

Multivariate Testing of a Mean

The multivariate analog of the *t*-statistic is:

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{S}_{n-1}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$$

It is called *Hotellings* T^2.

T^2 follows an $\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}$ distribution. (Remember? In univariate case when $t \sim t_{n-1}$ then $t^2 \sim \mathcal{F}_{1,n-1}$.)

Multivariate Testing of a Mean

The multivariate analog of the *t*-statistic is:

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}_{n-1}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

It is called *Hotellings* T^2.

T^2 follows an $\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}$ distribution. (Remember? In univariate case when $t \sim t_{n-1}$ then $t^2 \sim \mathcal{F}_{1,n-1}$.)

Multivariate Testing of a Mean

The multivariate analog of the *t*-statistic is:

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{S}_{n-1}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$$

It is called *Hotellings* T^2.

T^2 follows an $\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}$ distribution. (Remember? In univariate case when $t \sim t_{n-1}$ then $t^2 \sim \mathcal{F}_{1,n-1}$.)

Multivariate Testing of a Mean

The multivariate analog of the *t*-statistic is:

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}_{n-1}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

It is called *Hotellings* T^2.

T^2 follows an $\frac{(n-1)p}{n-p} \mathcal{F}_{p,n-p,\alpha}$ distribution. (Remember? In univariate case when $t \sim t_{n-1}$ then $t^2 \sim \mathcal{F}_{1,n-1}$.)

Multivariate Testing of a Mean

The multivariate analog of the *t*-statistic is:

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{S}_{n-1}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$$

It is called *Hotellings* T^2.

T^2 follows an $\frac{(n-1)p}{n-p} \mathcal{F}_{p,n-p,\alpha}$ distribution. (Remember? In univariate case when $t \sim t_{n-1}$ then $t^2 \sim \mathcal{F}_{1,n-1}$.)

Incorporates the dimension of the data!

Multivariate Testing of a Mean

Null hypothesis: $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$

Alternative hypothesis: $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

Multivariate Testing of a Mean

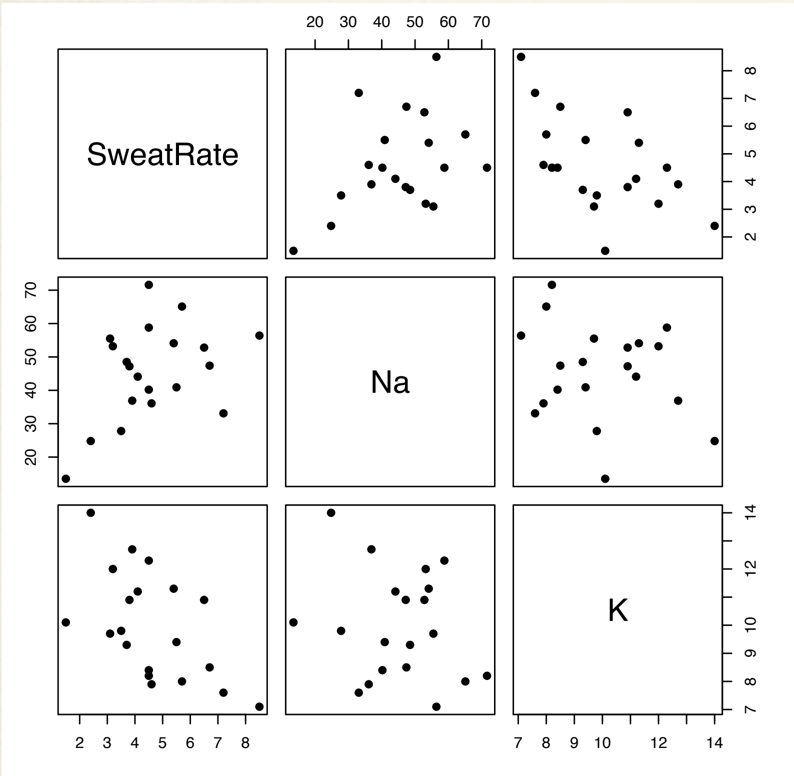
Null hypothesis: $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$

Alternative hypothesis: $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

Why aren't there one-sided alternatives?

Example: Sweat data

Perspiration of 20 healthy females for 3 components: Sweat Rate, Sodium Content and Potassium Content.



$$\bar{\mathbf{x}} = \begin{bmatrix} 4.61 \\ 45.6 \\ 9.97 \end{bmatrix}$$

$$\mathbf{S}_{n-1} = \begin{bmatrix} 2.75 & 9.37 & -1.72 \\ 9.37 & 190.8 & -5.35 \\ -1.72 & -5.35 & 3.45 \end{bmatrix}$$

$$\mathbf{S}_{n-1}^{-1} = \begin{bmatrix} 0.606 & -0.0222 & 0.268 \\ -0.0222 & 0.00630 & -0.00132 \\ 0.268 & -0.00132 & 0.422 \end{bmatrix}$$

We want to test

$$H_o : \boldsymbol{\mu} = (4 \quad 50 \quad 10)' \quad vs \quad H_A : \boldsymbol{\mu} \neq (4 \quad 50 \quad 10)'$$

Calculate:

$$\begin{aligned} T^2 &= 20 \begin{bmatrix} 4.61 - 4 \\ 45.6 - 50 \\ 9.97 - 10 \end{bmatrix}' \begin{bmatrix} 0.606 & -0.0222 & 0.268 \\ -0.0222 & 0.00630 & -0.00132 \\ 0.268 & -0.00132 & 0.422 \end{bmatrix} \\ &\qquad \begin{bmatrix} 4.61 - 4 \\ 45.6 - 50 \\ 9.97 - 10 \end{bmatrix} \\ &= 9.74 \end{aligned}$$

We want to test

$$H_o : \boldsymbol{\mu} = (4 \quad 50 \quad 10)' \quad vs \quad H_A : \boldsymbol{\mu} \neq (4 \quad 50 \quad 10)'$$

Calculate:

$$\begin{aligned} T^2 &= 20 \begin{bmatrix} 4.61 - 4 \\ 45.6 - 50 \\ 9.97 - 10 \end{bmatrix}' \begin{bmatrix} 0.606 & -0.0222 & 0.268 \\ -0.0222 & 0.00630 & -0.00132 \\ 0.268 & -0.00132 & 0.422 \end{bmatrix} \begin{bmatrix} 4.61 - 4 \\ 45.6 - 50 \\ 9.97 - 10 \end{bmatrix} \\ &= 9.74 \end{aligned}$$

We want to test

$$H_o : \boldsymbol{\mu} = \underbrace{(4 \quad 50 \quad 10)'}_{\boldsymbol{\mu}_0} \quad vs \quad H_A : \boldsymbol{\mu} \neq (4 \quad 50 \quad 10)'$$

Calculate:

$$\begin{aligned} T^2 &= 20 \begin{bmatrix} 4.61 - 4 \\ 45.6 - 50 \\ 9.97 - 10 \end{bmatrix}' \begin{bmatrix} 0.606 & -0.0222 & 0.268 \\ -0.0222 & 0.00630 & -0.00132 \\ 0.268 & -0.00132 & 0.422 \end{bmatrix} \\ &\qquad \begin{bmatrix} 4.61 - 4 \\ 45.6 - 50 \\ 9.97 - 10 \end{bmatrix} \\ &= 9.74 \end{aligned}$$

Using $\alpha = 0.10$, $\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}(0.10) = \frac{19\times 3}{17}\times 2.44 = 8.18$
then we would reject H_0 at the 10% level of significance.

Using $\alpha = 0.05$, $\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}(0.05) = \frac{19\times 3}{17}\times 3.2 = 10.7$
then we would NOT reject H_0 at the 5% level of significance.

Confidence Regions and Simultaneous Comparisons of Component Means

A $100(1-\alpha)\%$ confidence region for the mean of a *p*-dimensional normal distribution is the ellipsoid determined by all $\boldsymbol{\mu}$ such that

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}_{n-1}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq \frac{p(n-1)}{n-p}\mathcal{F}_{p,n-p}(\alpha)$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are sample observations from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population.

This equation defines an ellipse in *p*-space.

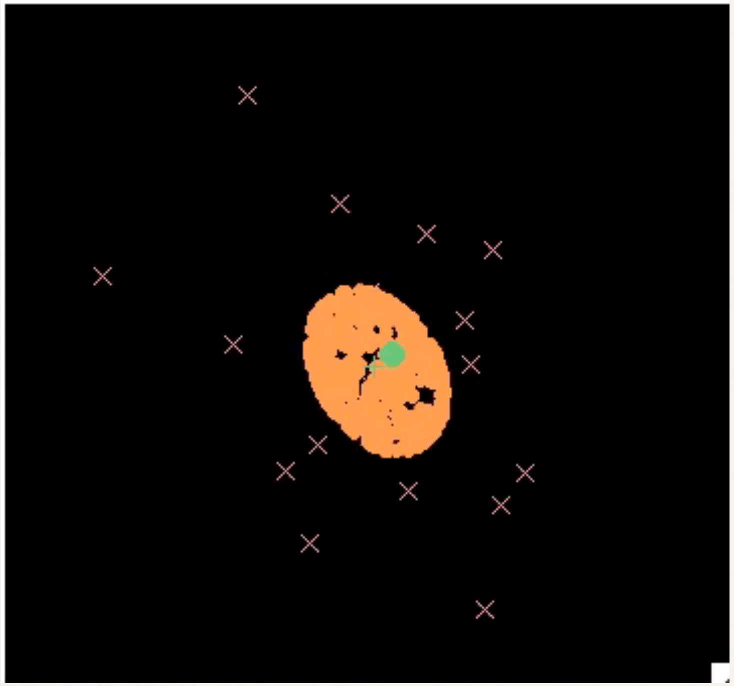
Lengths of the major axes.

The shape of the confidence ellipse is dictated by the variance-covariance matrix, $\mathbf{S}$. The orientation is given by the eigenvectors, and the lengths of the major axes of the confidence ellipse are given by

$$\sqrt{\lambda_i} \times \sqrt{\frac{p(n-1)}{n(n-p)}\mathcal{F}_{p,n-p}(\alpha)}$$

where λ_i is the *i*'th eigenvalue.

Example: Sweat data



90% confidence region

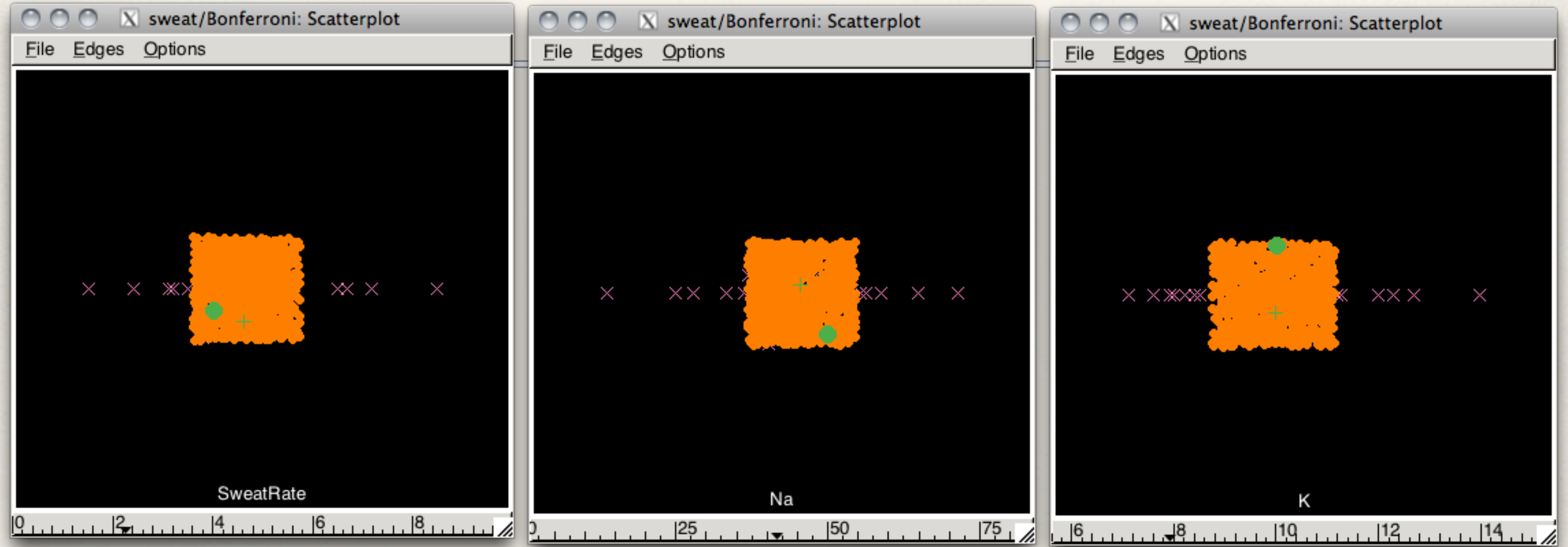
x data points
+ sample mean, $\bar{\mathbf{x}}$
● hypothesized mean, $\boldsymbol{\mu}_0$
● confidence region (3D)

Notice that $\boldsymbol{\mu}_0$ is outside the confidence region, which corresponds to the "reject" result for the hypothesis test.

Relationship between Interval and Hypothesis Test

Hypothesis testing is equivalent to examining the hypothesized value in relation to the simultaneous confidence ellipse. If the hypothesized value lies outside the ellipse, then the null hypothesis is rejected.

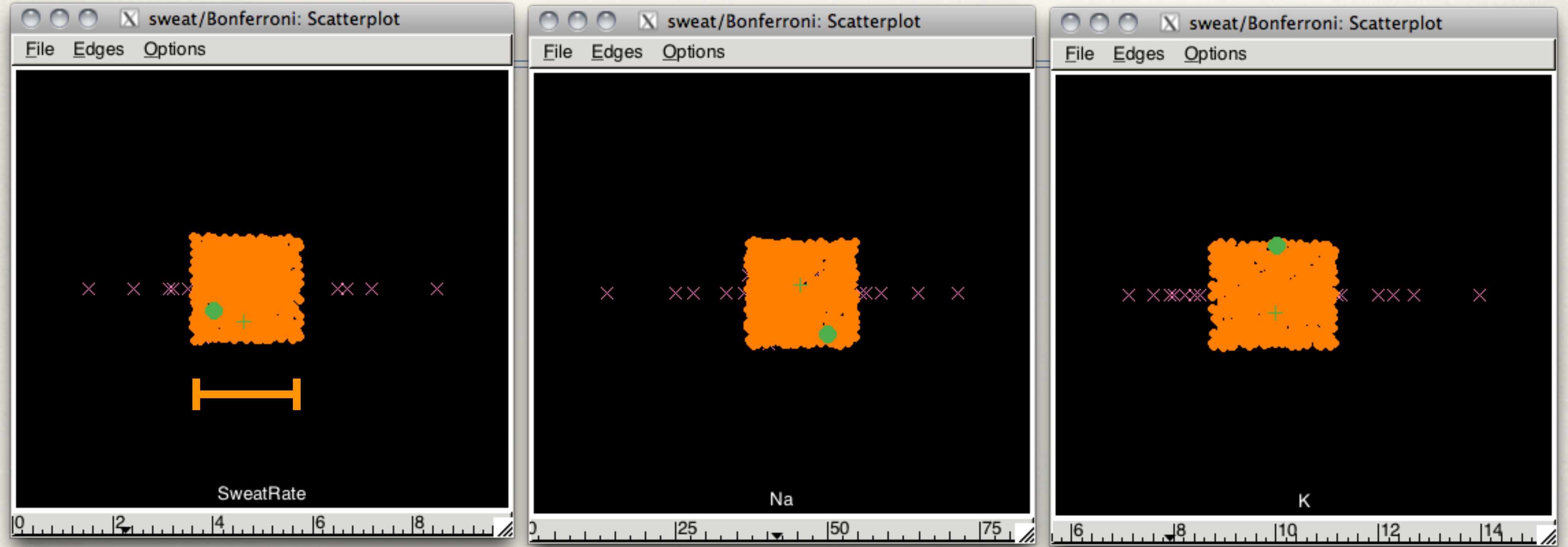
Simultaneous Confidence Intervals



Scheffe's:

$$\bar{X}_i \pm \sqrt{\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}}\frac{s_i}{\sqrt{n}}$$

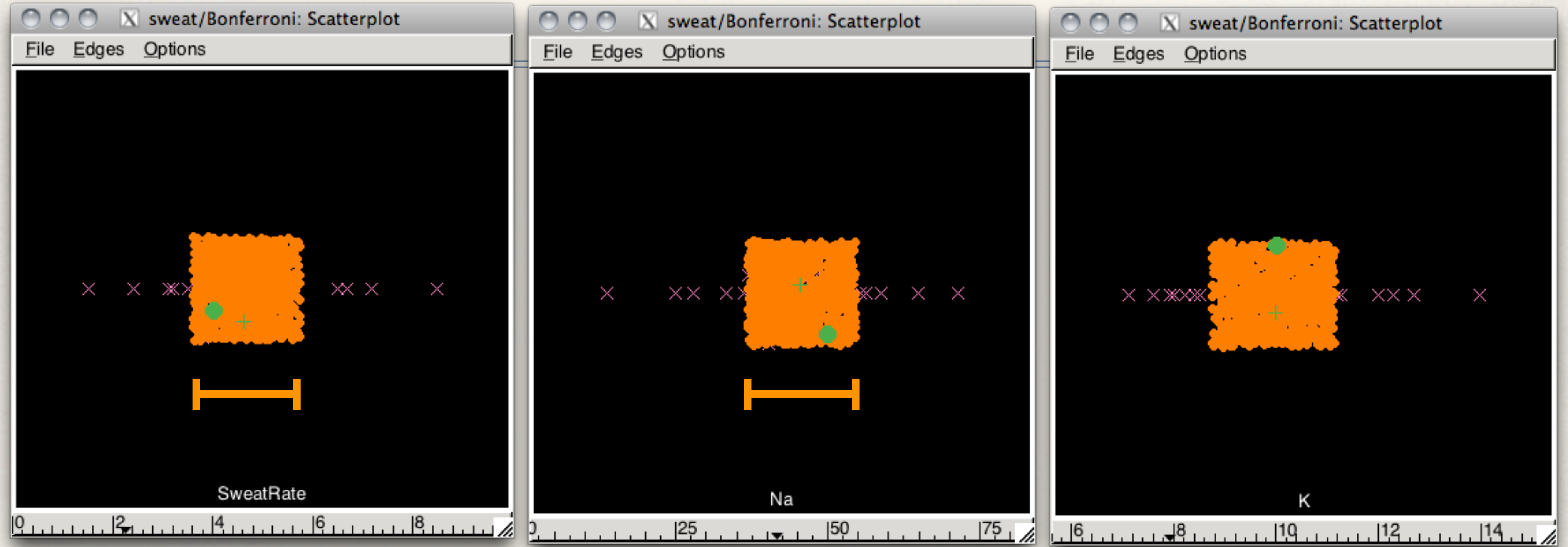
Simultaneous Confidence Intervals



Scheffe's:

$$\bar{X}_i \pm \sqrt{\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}}\frac{s_i}{\sqrt{n}}$$

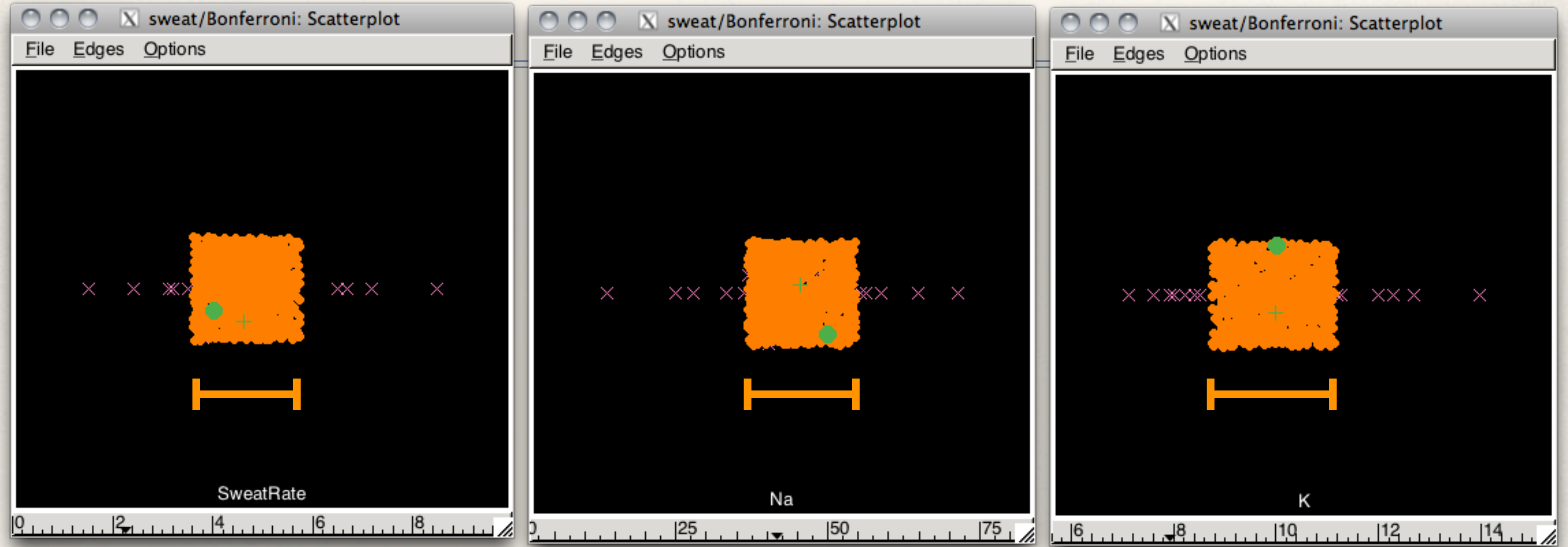
Simultaneous Confidence Intervals



Scheffe's:

$$\bar{X}_i \pm \sqrt{\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}}\frac{s_i}{\sqrt{n}}$$

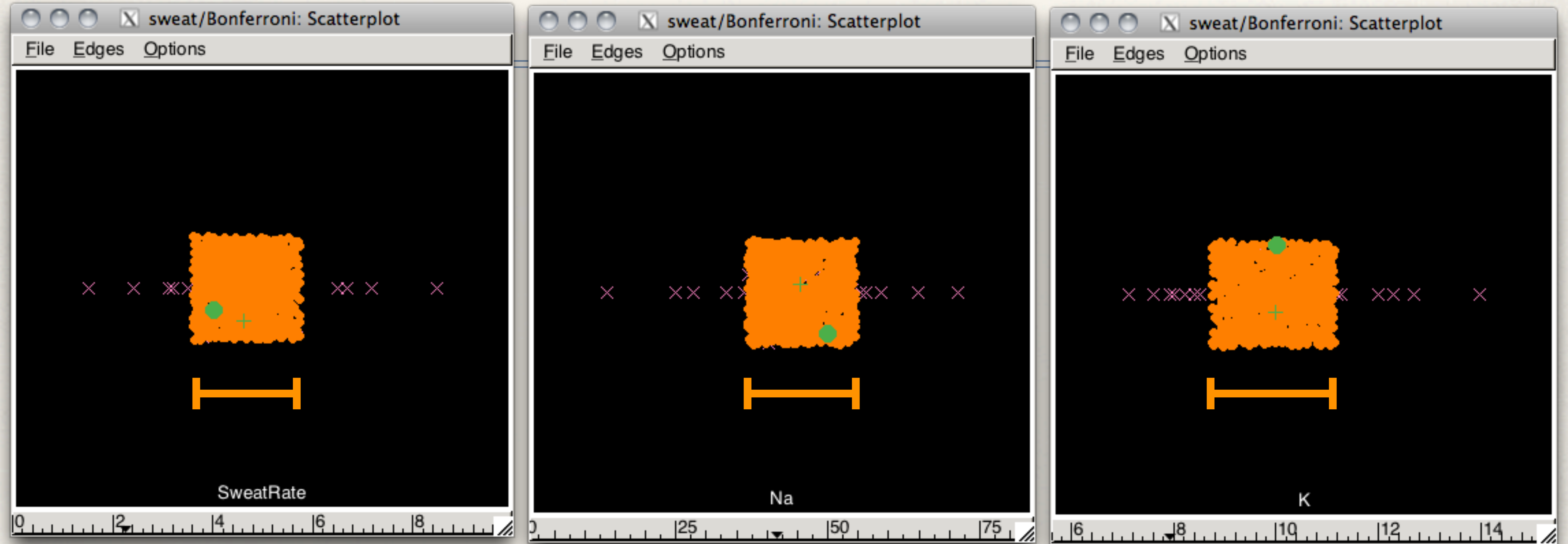
Simultaneous Confidence Intervals



Scheffe's:

$$\bar{X}_i \pm \sqrt{\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}}\frac{s_i}{\sqrt{n}}$$

Simultaneous Confidence Intervals

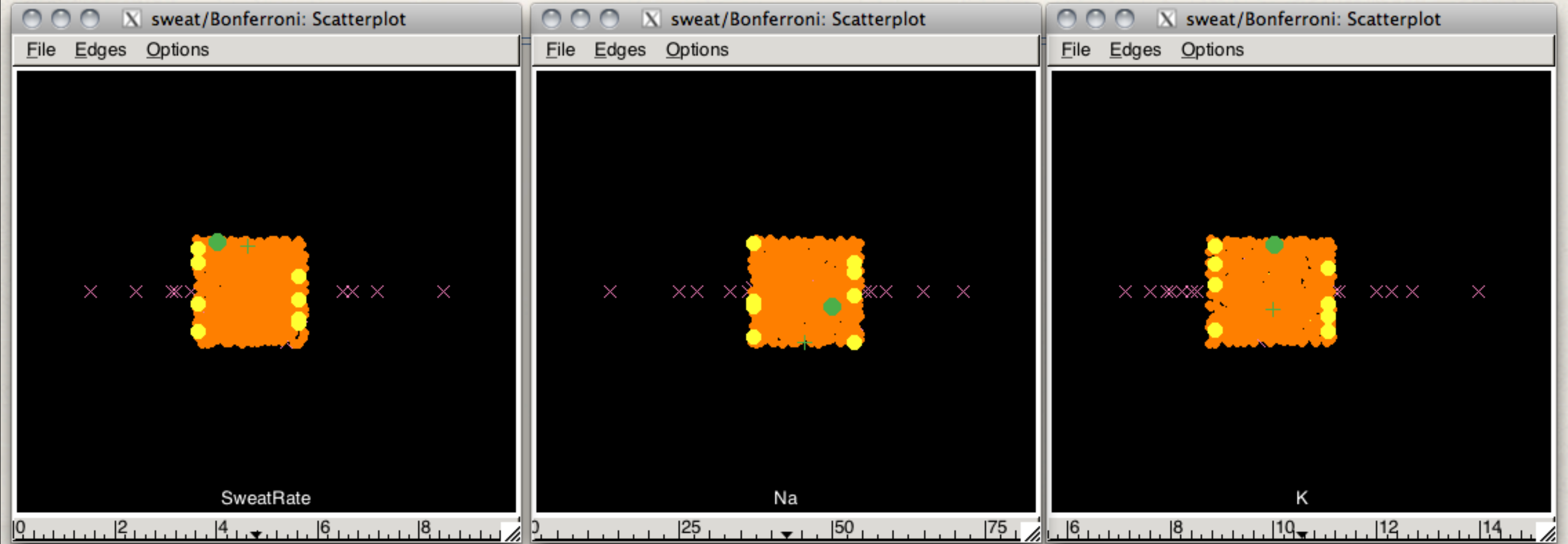


Scheffe's:

$$\bar{X}_i \pm \sqrt{\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p,\alpha}}\frac{s_i}{\sqrt{n}}$$

Hypothesized mean looks ok!

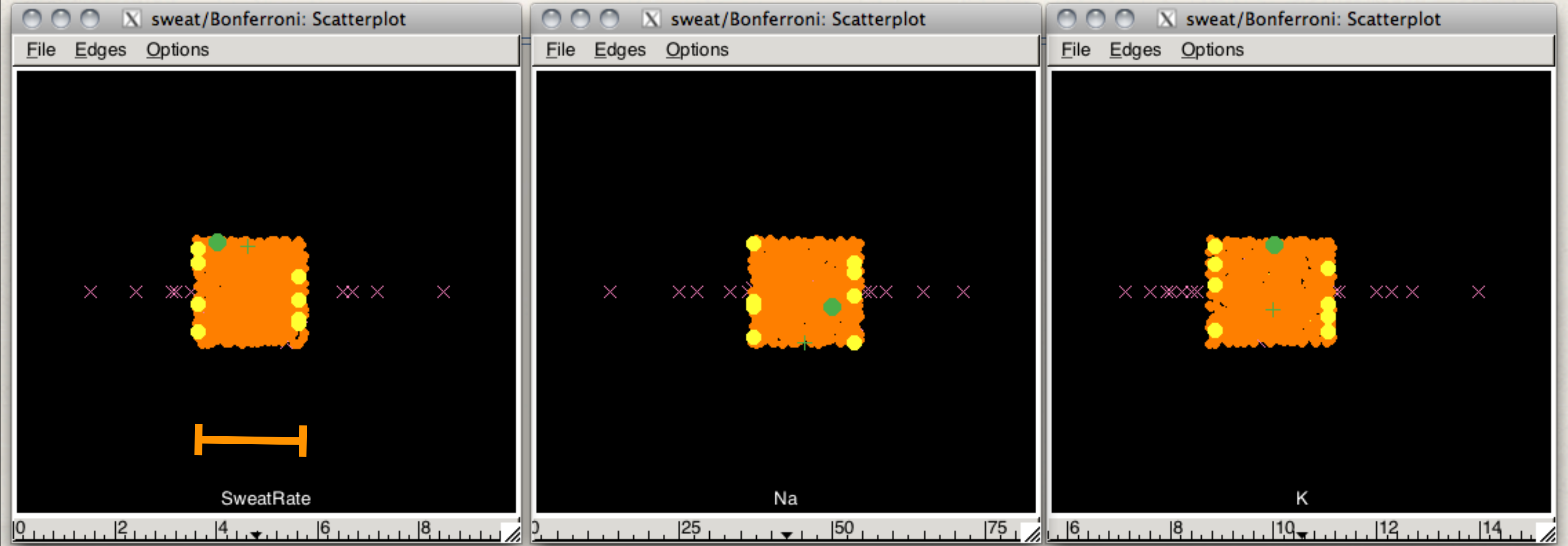
Simultaneous Confidence Intervals



Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

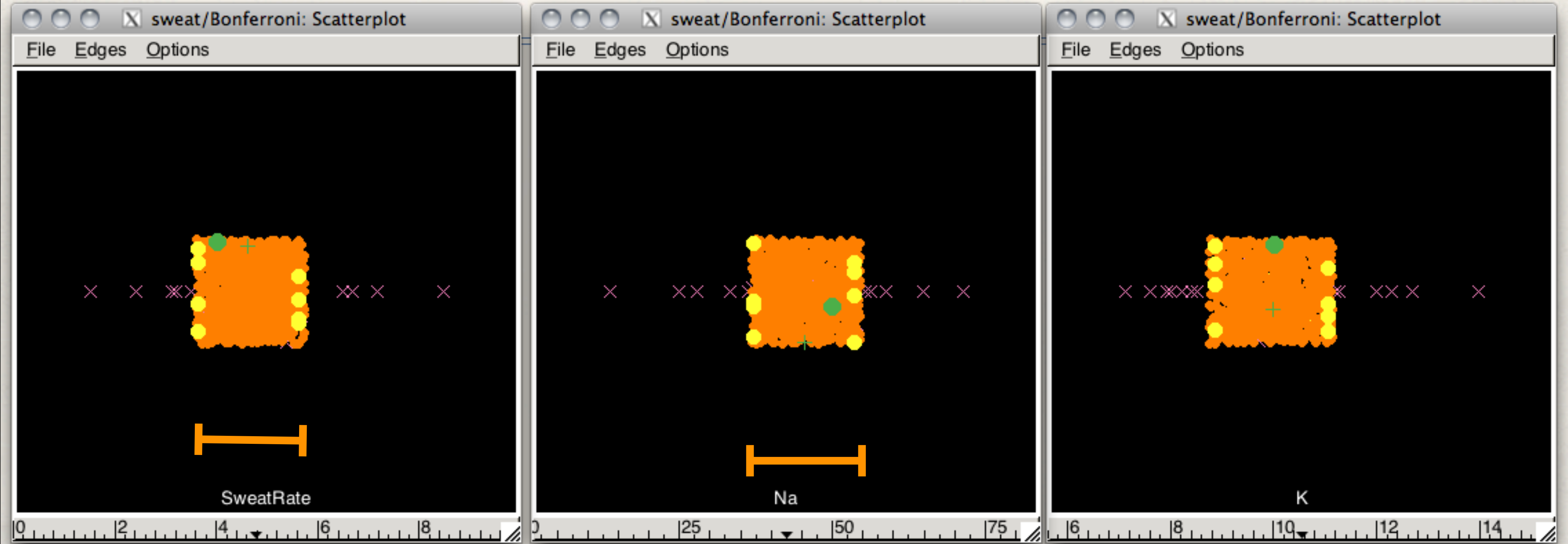
Simultaneous Confidence Intervals



Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

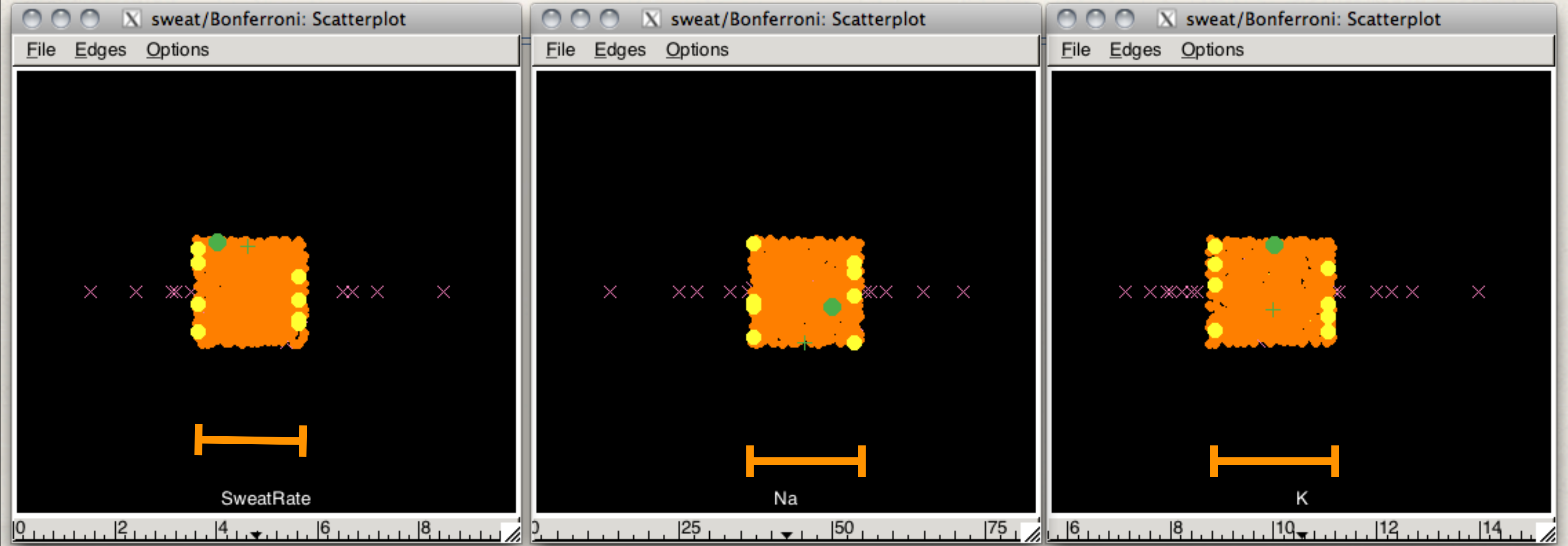
Simultaneous Confidence Intervals



Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

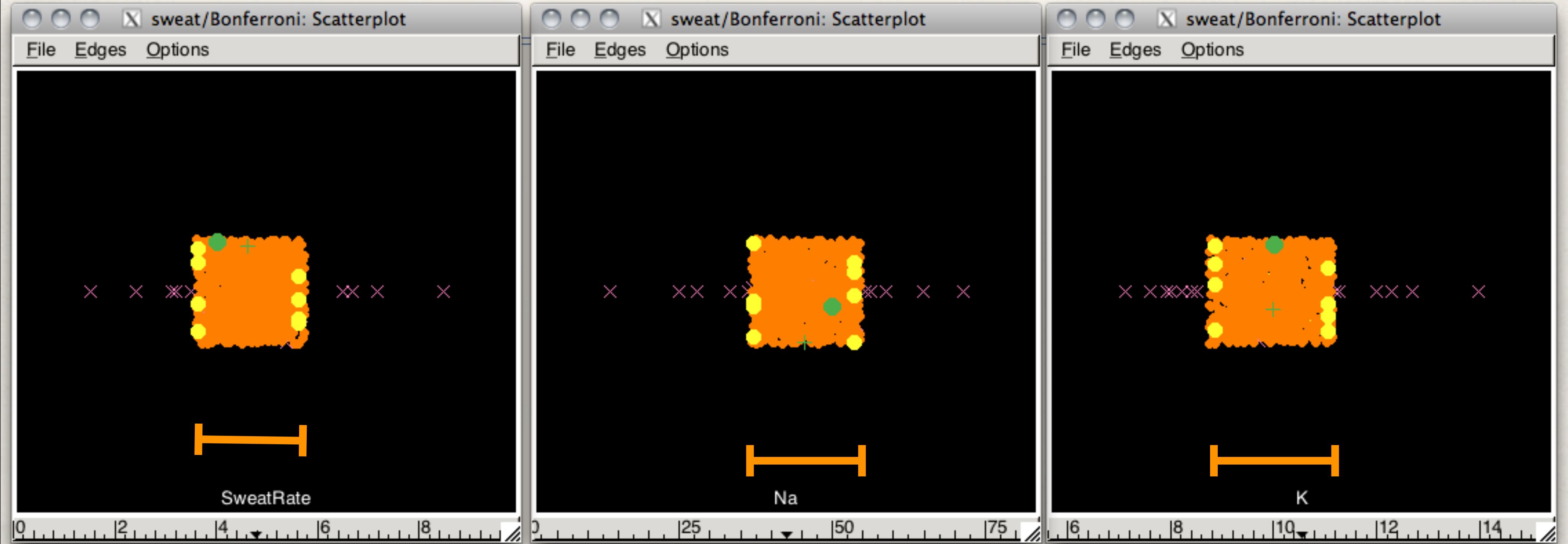
Simultaneous Confidence Intervals



Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

Simultaneous Confidence Intervals

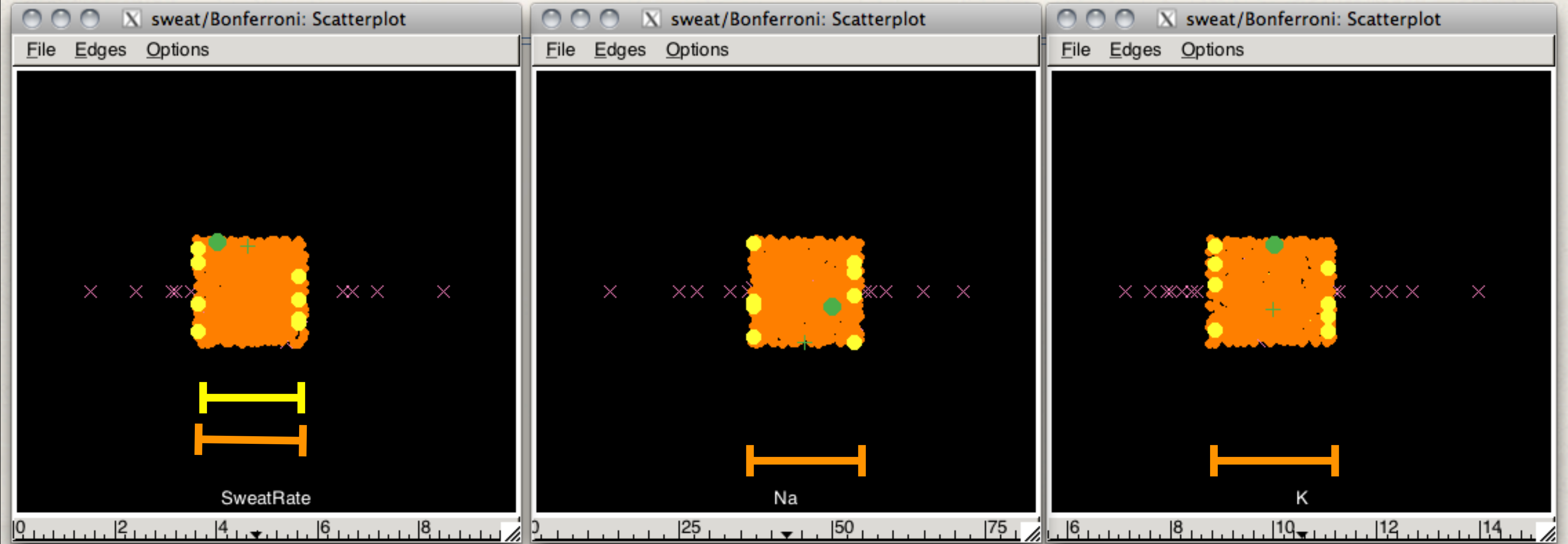


Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

Bonferroni intervals are smaller than Scheffe.

Simultaneous Confidence Intervals

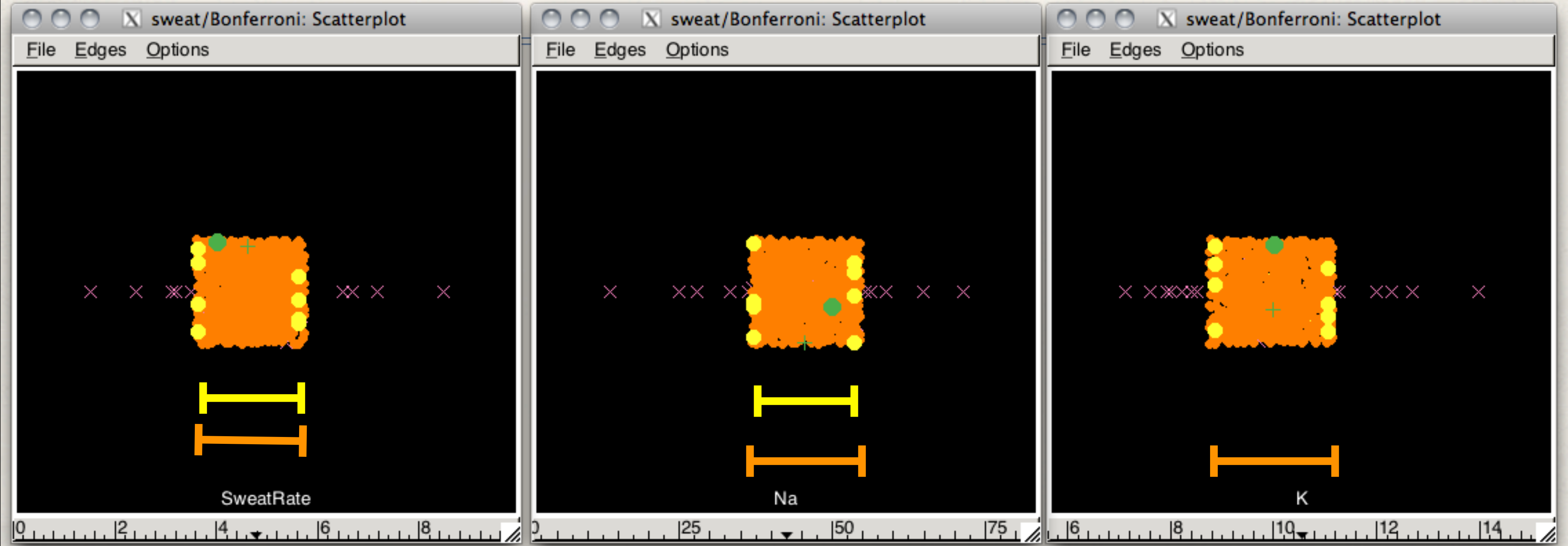


Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

Bonferroni intervals are smaller than Scheffe.

Simultaneous Confidence Intervals

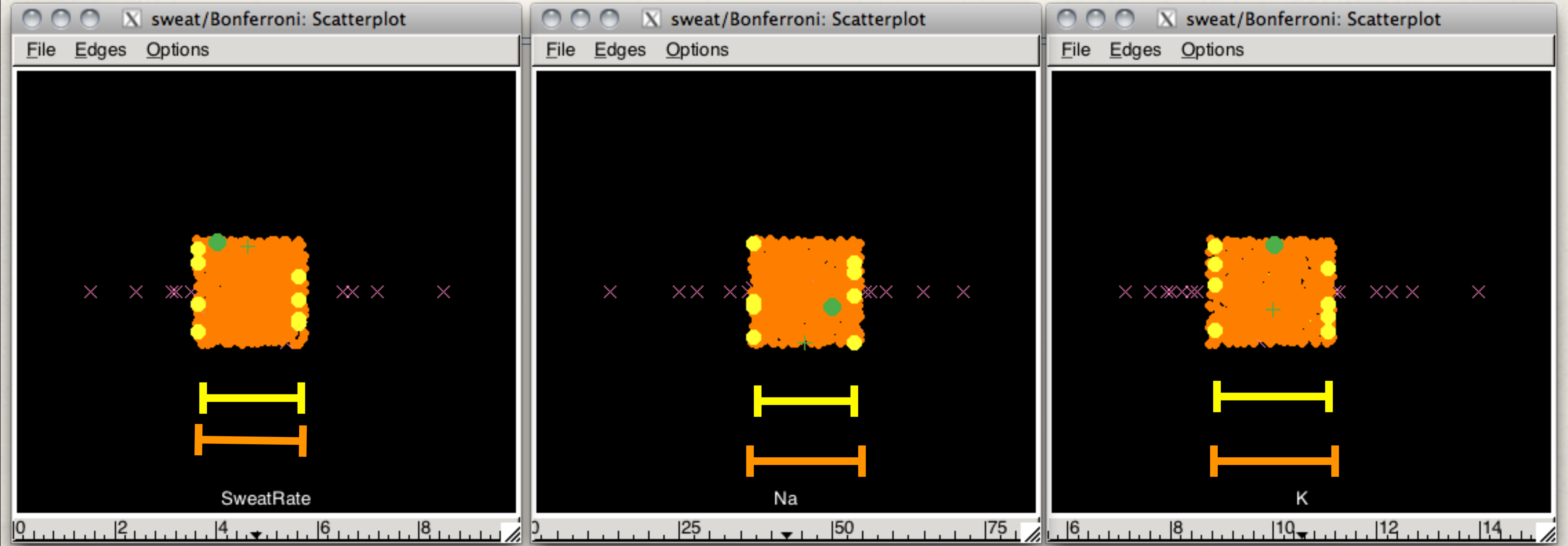


Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

Bonferroni intervals are smaller than Scheffe.

Simultaneous Confidence Intervals

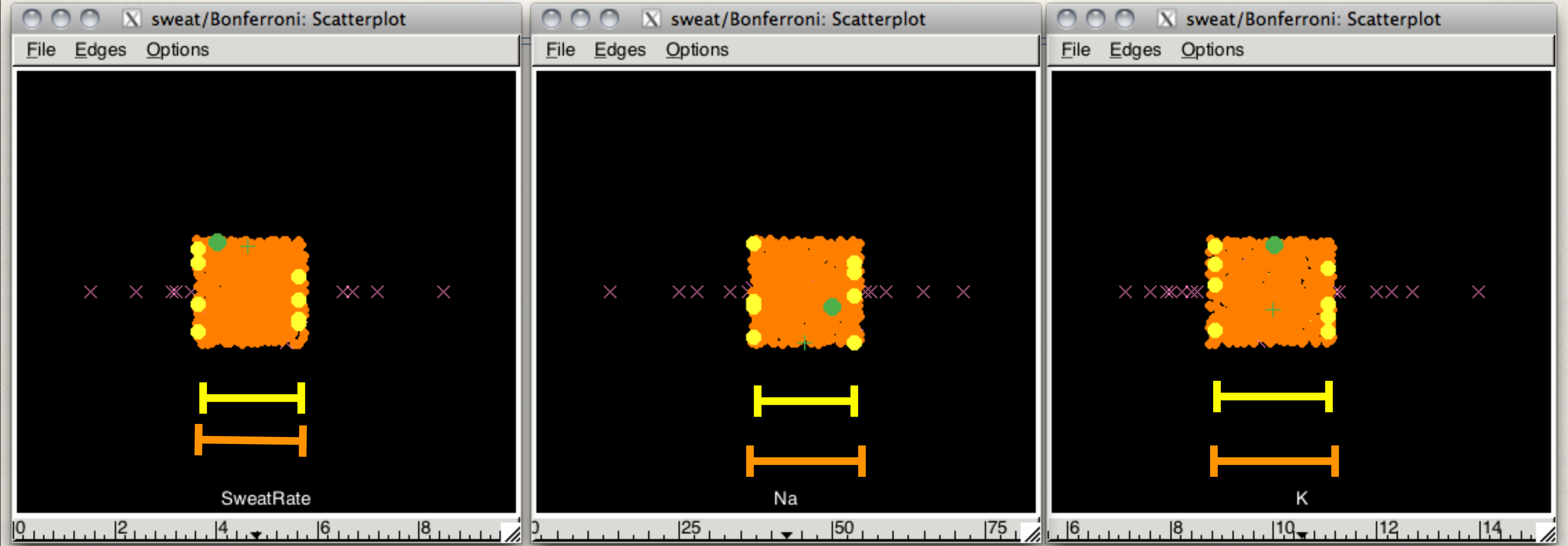


Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

Bonferroni intervals are smaller than Scheffe.

Simultaneous Confidence Intervals

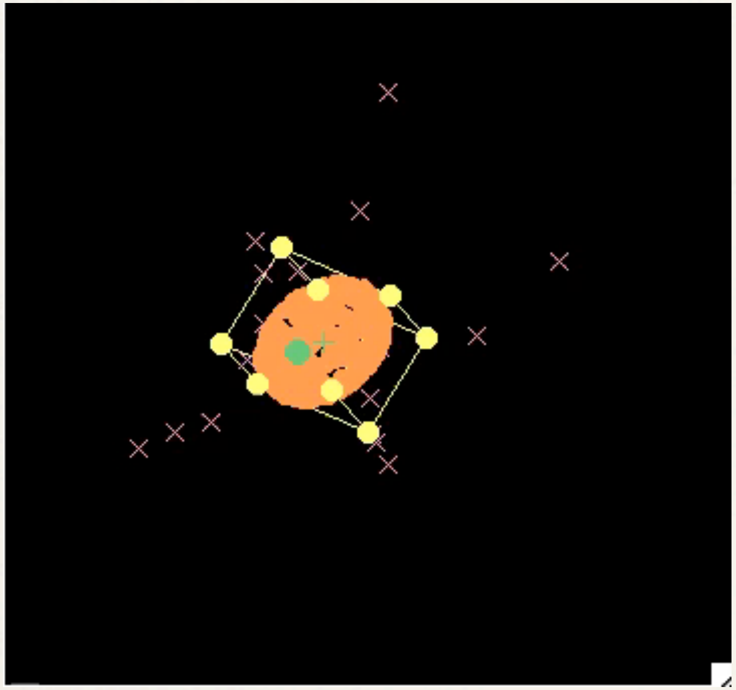


Bonferroni:

$$\bar{X}_i \pm t_{n-1,\frac{\alpha}{2p}} \frac{s_i}{\sqrt{n}}$$

Bonferroni intervals are smaller than Scheffe.

Intervals vs Region



90% confidence region

x data points

+ sample mean, $\bar{\mathbf{x}}$

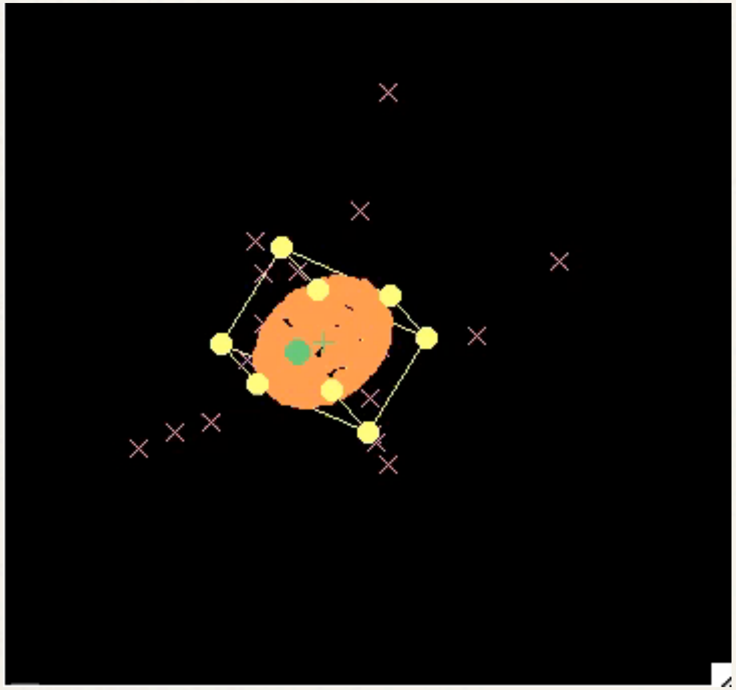
● hypothesized mean, μ_0

● confidence region (3D)

● Bonferroni confidence interval, region

Ellipse extends outside of box, in some directions. Hypothesized mean is inside box always.

Intervals vs Region



Bonferroni don't effectively use variance-covariance.

90% confidence region

x data points

+ sample mean, $\bar{\mathbf{x}}$

● hypothesized mean, μ_0

● confidence region (3D)

● Bonferroni confidence interval, region

Ellipse extends outside of box, in some directions. Hypothesized mean is inside box always.

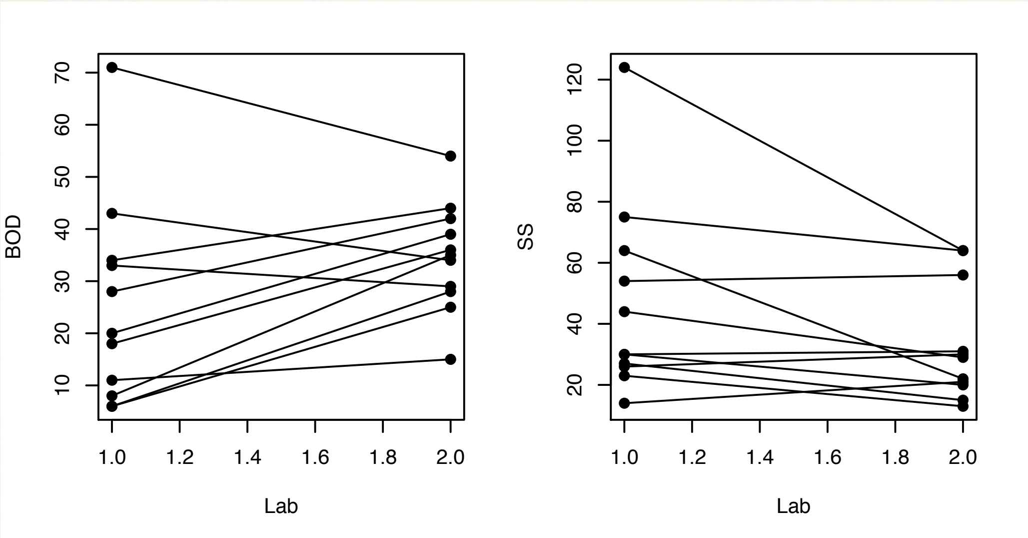
Paired Samples

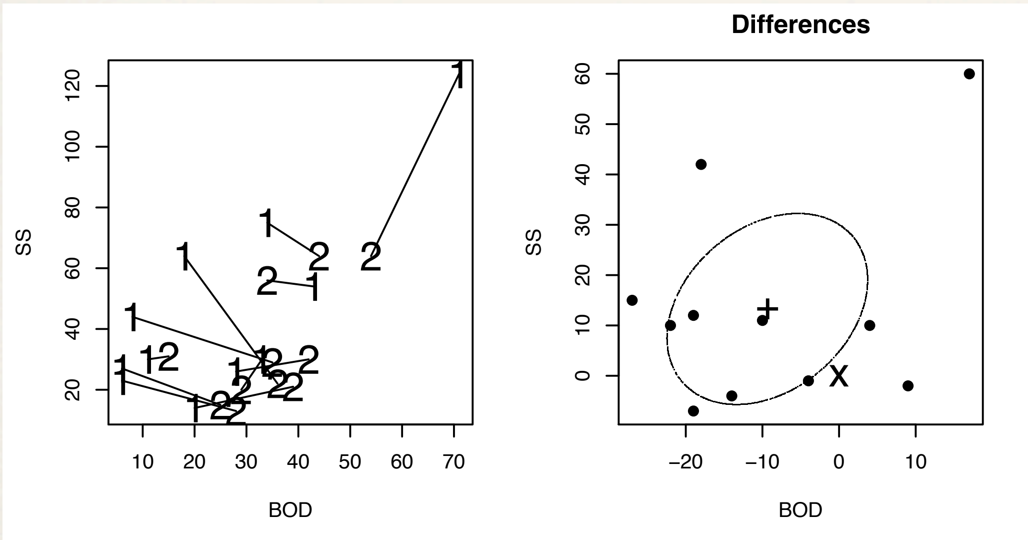
- Subtract the second value from the first and treat as a one sample problem.
- Example: "Municipal wastewater treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data from one of these self-monitoring programs led to a study in which samples of effluent were divided and sent to two laboratories were divided and sent to two laboratories for testing. One half of each sample was sent to the Wisconsin state lab and the other half sent to a private commercial lab routinely used in the monitoring program. Measurements were made on biochemical oxygen demand (BOD) and suspended solids (SS) were obtained, on $n=11$ samples."
- Are the analyses producing equivalent results?

Data

Sample	BOD	SS	BOD	SS	BOD	SS
1	6	27	25	15	-19	12
2	6	23	28	13	-22	10
3	18	64	36	22	-18	42
4	8	44	35	29	-27	15
5	11	30	15	31	-4	-1
6	34	75	44	64	-10	11
7	28	26	42	30	-14	-4
8	71	124	54	64	17	60
9	43	54	34	56	9	-2
10	33	30	29	20	4	10
11	20	14	39	21	-19	-7

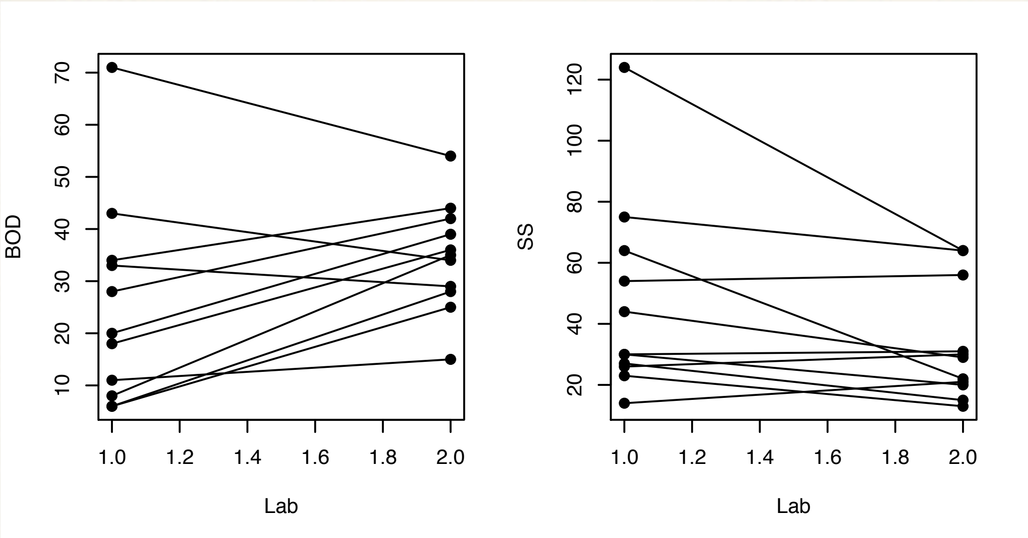
Plots

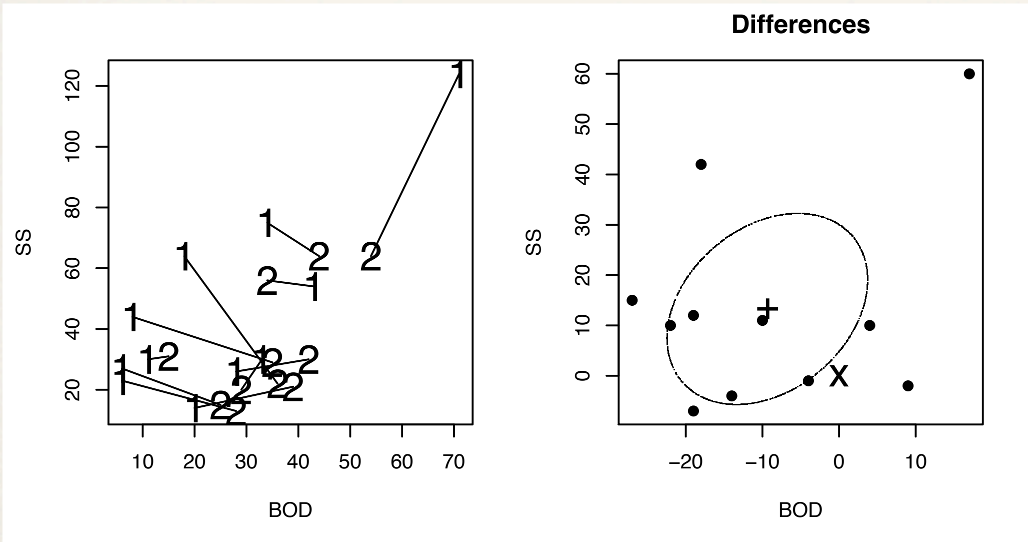




- Plot vars vs repeats
  - Are the values for BOD the same from labs 1 & 2?
  - Are the values for SS the same from labs 1 & 2?
- Plot repeats in multivariate plots.
  - Is there a systematic shift from one rep to another.
- Plot differences, along with sample mean and hypothesized mean.
  - How similar are the sample mean and hyp mean?

Plots





- Plot vars vs repeats
  - Are the values for BOD the same from labs 1 & 2?
  - Are the values for SS the same from labs 1 & 2?
- Plot repeats in multivariate plots.
  - Is there a systematic shift from one rep to another.
- Plot differences, along with sample mean and hypothesized mean.
  - How similar are the sample mean and hyp mean?

Reject null, SS higher in lab 1, BOD lower in lab 1.

Two Sample Hotellings T^2

For two samples collected independently from multivariate normal populations, $N(\boldsymbol{\mu}_1, \Sigma), N(\boldsymbol{\mu}_2, \Sigma)$, the test statistic for
$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ Vs $H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ is

$$T^2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'((\frac{1}{n_1} + \frac{1}{n_2})\mathbf{S}_{pooled})^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$

distributed as $\frac{(n_1+n_2-2)p}{n_1+n_2-1-p}\mathcal{F}_{p,n_1+n_2-p-1,\alpha}$ and

$$\mathbf{S}_{pooled} = \frac{n_1-1}{n_1+n_2-2}\mathbf{S}_1 + \frac{n_2-1}{n_1+n_2-2}\mathbf{S}_2$$

Test for Homogeneity of Variance-covariances

To test the hypothesis $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g$ against the alternative that they are not all equal, SAS PROC DISCRIM uses a LRT test

$$\Lambda = \frac{\prod_{i=1}^{g} |\mathbf{S}_i|^{(n_i-1)/2}}{|\mathbf{S}_{pooled}|^{(n-p)/2}}$$

(extends Bartlett's univariate test.) No exact distribution, but for large data χ^2 can be used to obtain *p*-values. Also Box's M test,

$(n-g)\log|\mathbf{S}| - \sum_{i=1}^{g}(n_i-1)\log|\mathbf{S}_i|$, there's R code for this. BUT, both tend to be sensitive, to very often reject the null!!

Comparing several Multivariate Means - MANOVA

- **Data:** g samples of size $n_1, \ldots n_g$; p variables.

- **Population:** g populations; different means $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g)$ same covariance (Σ), multivariate normal.

- Arises from (1) designed experimental studies, where there are multiple responses measured for one or more treatments, (2) in classification problems, where we want to classify a new observation into a class, based on two or more measured variables.

Recall Univariate ANOVA

We have random samples $X_{l1}, X_{l2}, \ldots, X_{ln_l}$ from $N(\mu_l, \sigma^2)$ *l=1, ..., g*. We are interested to know if the population means of the groups are different, that is, if $\mu_1 = \mu_2 = \ldots = \mu_g$.

The problem is parametrized into the following model formulation:

$$X_{lj} = \mu \quad + \quad \tau_l \quad + \quad e_{lj}$$

with the constraint $\sum_{l=1}^{g} n_l \tau_l = 0$, which leads to the null hypothesis notation:

$$H_o : \tau_1 = \tau_2 = \ldots = \tau_g = 0$$

The model is fit to the data by splitting each observation into components:

$$X_{lj} = \bar{X} \quad + \quad (\bar{X}_l - \bar{X}) \quad + \quad (X_{lj} - \bar{X}_l), \quad j = 1, \ldots, n_l; \; l = 1, \ldots, g$$

and computing the sums of squares for each component:

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X})^2 = \sum_{l=1}^{g} n_l (\bar{X}_l - \bar{X})^2 + \sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2$$

The model is fit to the data by splitting each observation into components:

$$X_{lj} = \bar{X} \quad + \quad (\bar{X}_l - \bar{X}) \quad + \quad (X_{lj} - \bar{X}_l), \quad j = 1, \ldots, n_l;\ l = 1, \ldots, g$$

and computing the sums of squares for each component:

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X})^2 = \sum_{l=1}^{g} n_l (\bar{X}_l - \bar{X})^2 + \sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2$$

SS CorTot

The model is fit to the data by splitting each observation into components:

$$X_{lj} = \bar{X} \quad + \quad (\bar{X}_l - \bar{X}) \quad + \quad (X_{lj} - \bar{X}_l), \quad j = 1, \ldots, n_l;\ l = 1, \ldots, g$$

and computing the sums of squares for each component:

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X})^2 = \sum_{l=1}^{g} n_l (\bar{X}_l - \bar{X})^2 + \sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2$$

SS CorTot SS Tr

The model is fit to the data by splitting each observation into components:

$$X_{lj} = \bar{X} \quad + \quad (\bar{X}_l - \bar{X}) \quad + \quad (X_{lj} - \bar{X}_l), \quad j = 1, \ldots, n_l; \; l = 1, \ldots, g$$

and computing the sums of squares for each component:

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X})^2 = \sum_{l=1}^{g} n_l (\bar{X}_l - \bar{X})^2 + \sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2$$

SS CorTot SS Tr SS Res

ANOVA Table

Source	SS	df
Treatment (Between)	SS_{Tr}	$g-1$
Error (Within)	SS_{Res}	$\sum_{l=1}^{g} n_l - g$
Total	SS_{CorTot}	$\sum_{l=1}^{g} n_l - 1$

Then we test the hypothesis using

$$F = \frac{SS_{Tr}/(g-1)}{SS_{Res}/(\sum_{l=1}^{g} n_l - g)} \sim \mathcal{F}_{g-1,\sum n_l - g}(\alpha)$$

Multivariate ANOVA

The corresponding model formulation is

$$\mathbf{x}_{lj} = \boldsymbol{\mu} \qquad + \qquad \boldsymbol{\tau}_l \qquad + \qquad \mathbf{e}_{lj}$$

And the corresponding data decomposition

$$\mathbf{x}_{lj} = \bar{\mathbf{x}} \quad + \quad (\bar{\mathbf{x}}_l - \bar{\mathbf{x}}) \quad + \quad (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l), \quad j = 1, \ldots, n_l; \; l = 1, \ldots, g$$

The corresponding sums of squares is given by:

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{X}_{lj}-\bar{\mathbf{X}})(\mathbf{X}_{lj}-\bar{\mathbf{X}})' = \sum_{l=1}^{g} n_l(\bar{\mathbf{X}}_l-\bar{\mathbf{X}})(\bar{\mathbf{X}}_l-\bar{\mathbf{X}})' +$$
$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{X}_{lj}-\bar{\mathbf{X}}_l)(\mathbf{X}_{lj}-\bar{\mathbf{X}}_l)'$$

MANOVA Table

Source	SS	df
Treatment (Between)	$\mathbf{B}$	$g-1$
Error (Within)	$\mathbf{W}$	$\sum_{l=1}^{g} n_l - g$
Total	$\mathbf{B}+\mathbf{W}$	$\sum_{l=1}^{g} n_l - 1$

The corresponding sums of squares is given by:

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{X}_{lj}-\bar{\mathbf{X}})(\mathbf{X}_{lj}-\bar{\mathbf{X}})' = \underbrace{\sum_{l=1}^{g} n_l(\bar{\mathbf{X}}_l-\bar{\mathbf{X}})(\bar{\mathbf{X}}_l-\bar{\mathbf{X}})'}_{\text{B}} +$$

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{X}_{lj}-\bar{\mathbf{X}}_l)(\mathbf{X}_{lj}-\bar{\mathbf{X}}_l)'$$

MANOVA Table

Source	SS	df
Treatment (Between)	$\mathbf{B}$	$g-1$
Error (Within)	$\mathbf{W}$	$\sum_{l=1}^{g} n_l - g$
Total	$\mathbf{B}+\mathbf{W}$	$\sum_{l=1}^{g} n_l - 1$

The corresponding sums of squares is given by:

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{X}_{lj}-\bar{\mathbf{X}})(\mathbf{X}_{lj}-\bar{\mathbf{X}})' = \underbrace{\sum_{l=1}^{g} n_l(\bar{\mathbf{X}}_l-\bar{\mathbf{X}})(\bar{\mathbf{X}}_l-\bar{\mathbf{X}})'}_{\mathbf{B}} +$$

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{X}_{lj}-\bar{\mathbf{X}}_l)(\mathbf{X}_{lj}-\bar{\mathbf{X}}_l)'}_{\mathbf{W}}$$

MANOVA Table

Source	SS	df
Treatment (Between)	$\mathbf{B}$	$g-1$
Error (Within)	$\mathbf{W}$	$\sum_{l=1}^{g} n_l - g$
Total	$\mathbf{B}+\mathbf{W}$	$\sum_{l=1}^{g} n_l - 1$

Testing the hypothesis

$$H_o : \boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_l = \mathbf{0}$$

The big complication here is that we are comparing matrices now instead of single numbers. The matrices correspond to variance ellipses in *p*-D space. So there are numerous test statistics. The most popular is Wilks' lambda:

$$\Lambda^* = \frac{|W|}{|B+W|}$$

If Λ^* is too small then reject H_o. If $\sum n_l = n$ is large,

$-(n-1-(p+g)/2)\ln\left(\frac{|W|}{|B+W|}\right)$ is distributed as $\chi^2_{p(g-1)}(\alpha)$. For smaller *n*, some combinations of *p*,*g* the test statistic has an *F* distribution.

Note that we can also express Λ^* as

$$\Lambda^* = \prod_{i=1}^{\min(p,g-1)} \frac{1}{1+\lambda_i}$$

where $\lambda_i; \quad i = 1, \ldots, \min(p, g-1)$ are the eigenvalues of $W^{-1}B$.

Also note, being based on determinants of the matrices that Wilks lambda is essentially comparing volumes of the corresponding ellipses. Another way of calculating volumes of ellipses is the products of the lengths of the leading axes (eigenvalues).

Other Test Statistics

- Roy's largest eigenvalue of $\mathbf{BW}^{-1}$
- Lawley and Hotellings: $T = trace(\mathbf{BW}^{-1})$
- Pillai's trace: $V = trace(\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1})$

Example: Dominican Republic Student Testing

82 Students, enrolled in one of Technology (23), Architecture (38), Medical Technology (21) programs.

The students were given 4 tests: Aptitude, Math, Language, General Knowledge.

Is there is a difference in the average test scores for each student type?

n=82, n_1=23, n_2=38, n_3=21, p=4, g=3

Summary Statistics

n=82, n_1=23, n_2=38, n_3=21, p=4, g=3

$$\bar{\mathbf{X}} = \begin{bmatrix} 49.1 \\ 46.8 \\ 73.1 \\ 22.3 \end{bmatrix} \quad \bar{\mathbf{X}}_1 = \begin{bmatrix} 39.0 \\ 47.4 \\ 70.6 \\ 19.2 \end{bmatrix} \quad \bar{\mathbf{X}}_2 = \begin{bmatrix} 67.2 \\ 51.2 \\ 77.3 \\ 21.5 \end{bmatrix} \quad \bar{\mathbf{X}}_3 = \begin{bmatrix} 27.4 \\ 38.1 \\ 68.1 \\ 27.2 \end{bmatrix}$$

$$\mathbf{S}_1 = \begin{bmatrix} 1067.7 & 224.1 & 97.5 & 74.0 \\ 224.1 & 192.9 & 34.6 & 37.9 \\ 97.5 & 34.6 & 68.9 & 3.2 \\ 74.0 & 37.9 & 3.2 & 50.5 \end{bmatrix} \qquad \mathbf{S}_2 = \begin{bmatrix} 488.9 & 25.8 & 37.8 & 67.5 \\ 25.8 & 235.7 & 46.0 & -28.3 \\ 37.8 & 46.0 & 57.2 & 0.57 \\ 67.5 & -28.3 & 0.57 & 68.2 \end{bmatrix}$$

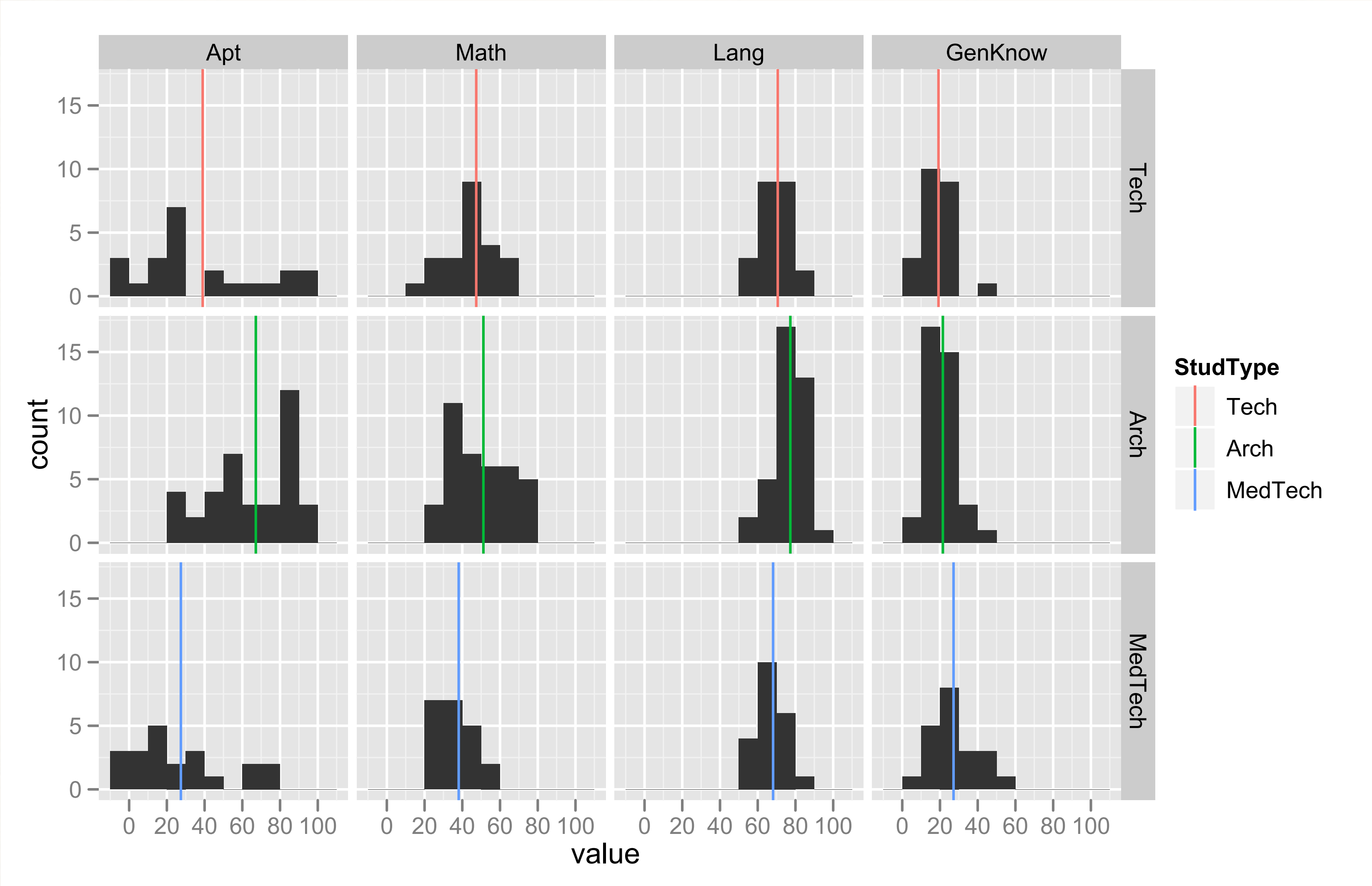
$$\mathbf{S}_3 = \begin{bmatrix} 672.9 & 112.9 & 101.2 & 68.3 \\ 112.9 & 81.2 & 7.8 & 4.4 \\ 101.2 & 7.8 & 56.3 & -9.7 \\ 68.3 & 4.4 & -9.7 & 170.4 \end{bmatrix}$$

Can we consider that the population variance-covariances might be equal?

Descriptive graphics

Can we consider that the population variances might be equal?

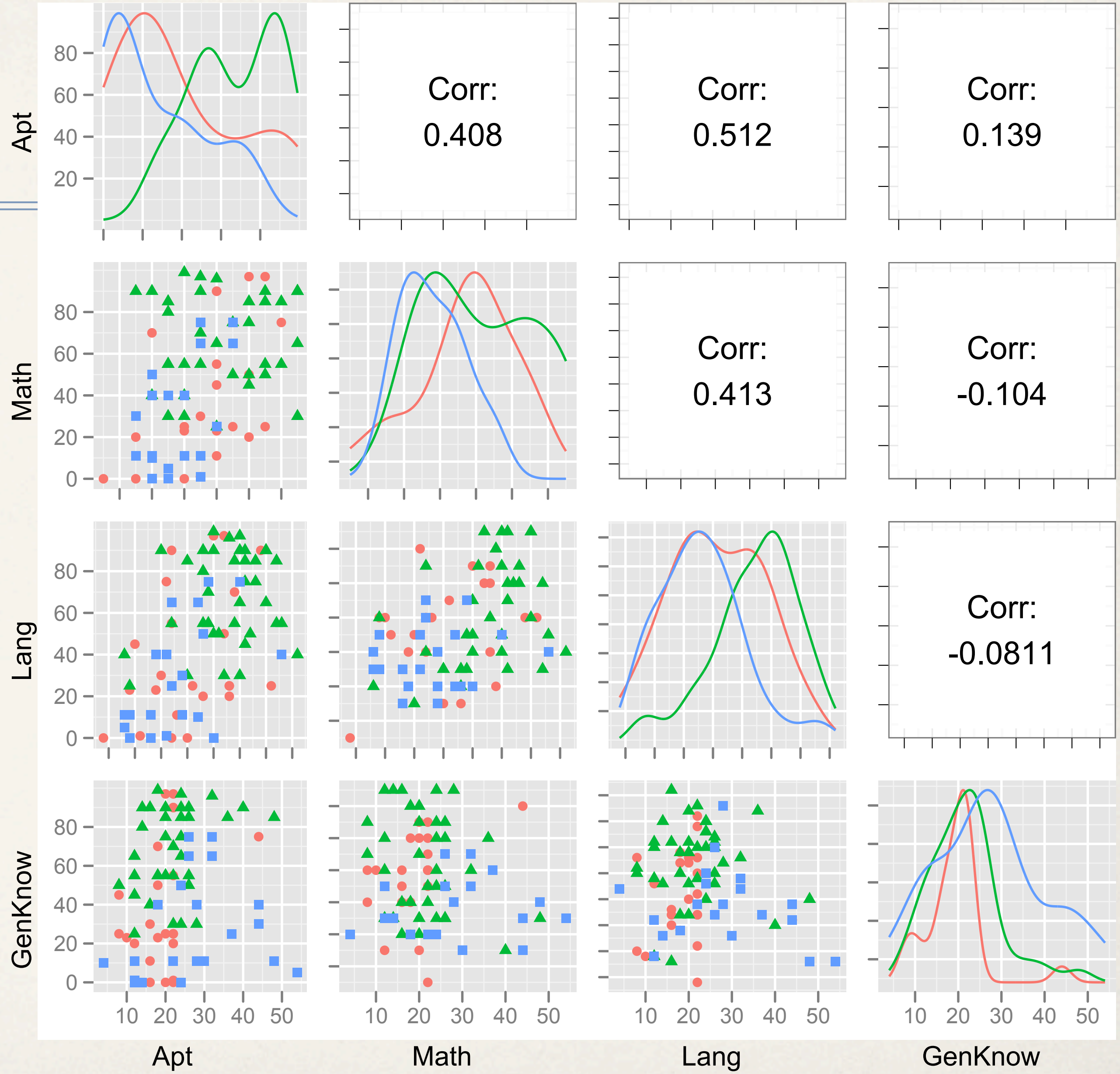
Are the population means different?



Descriptive graphics

Can we consider that the population variance-covariances might be equal?

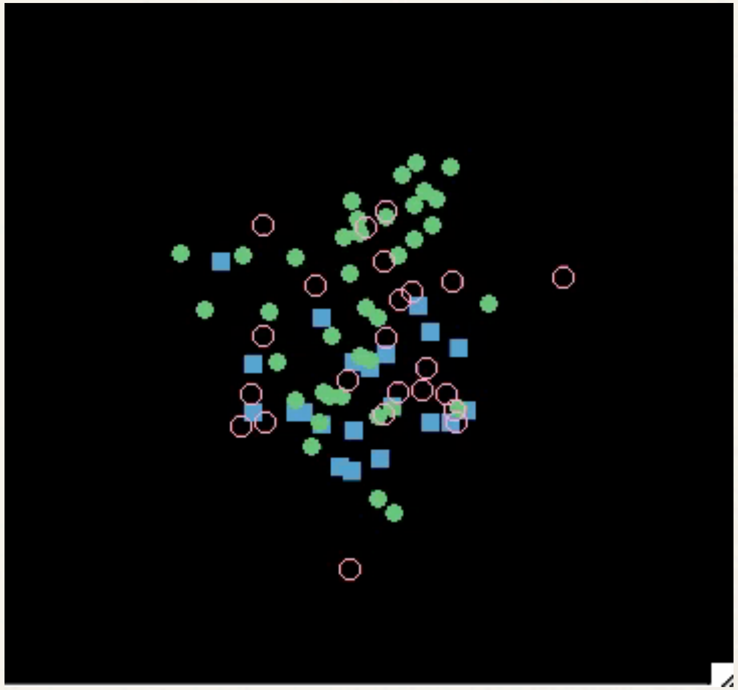
Are the means different?



Descriptive graphics

Can we consider that the population variance-covariances might be equal?

Are the means different?



Test

$$H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 \quad \text{OR}$$

$$H_o : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3 = \mathbf{0}$$

$$\mathbf{B} = \begin{bmatrix} 24600.2 & 6832.6 & 5709.5 & -2040.4 \\ 6832.6 & 2329.6 & 1570.2 & -1064.7 \\ 5709.5 & 1570.2 & 1325.7 & -455.6 \\ -2040.4 & -1064.7 & -455.6 & 743.5 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} 55036.2 & 8140.0 & 5570.0 & 5490.1 \\ 8140.0 & 14589.0 & 2619.2 & -128.0 \\ 5570.0 & 2619.2 & 4759.9 & -102.4 \\ 5490.1 & -128.0 & -102.4 & 7040.6 \end{bmatrix}$$

Test

$$H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 \quad \text{OR}$$
$$H_o : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3 = \mathbf{0}$$

$$\mathbf{B} = \begin{bmatrix} 24600.2 & 6832.6 & 5709.5 & -2040.4 \\ 6832.6 & 2329.6 & 1570.2 & -1064.7 \\ 5709.5 & 1570.2 & 1325.7 & -455.6 \\ -2040.4 & -1064.7 & -455.6 & 743.5 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} 55036.2 & 8140.0 & 5570.0 & 5490.1 \\ 8140.0 & 14589.0 & 2619.2 & -128.0 \\ 5570.0 & 2619.2 & 4759.9 & -102.4 \\ 5490.1 & -128.0 & -102.4 & 7040.6 \end{bmatrix}$$

```
> summary(manova(cbind(Apt, Math, Lang, GenKnow)~StudType,
        domrep), test="Wilks")
          Df   Wilks approx F num Df den Df    Pr(>F)
StudType   2 0.54345   6.7736      8    152 1.384e-07 ***
Residuals 79
```

Which variables?

Analysis of variance on each variable

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Apt	2	24600	12300.1	17.656	4.589e-07	***
Math	2	2329.6	1164.80	6.3074	0.002875	**
Lang	2	1325.7	662.85	11.001	6.097e-05	***
GenKnow	2	743.5	371.74	4.1712	0.01896	*

All variables have significant difference between means! In order of importance: Apt, Lang, Math, GenKnow

Which groups?

Tukey's multiple comparisons

Apt	diff	lwr	upr	p adj
Arch-Tech	28.16	11.50	44.81	0.00
MedTech-Tech	-11.57	-30.60	7.46	0.32
MedTech-Arch	-39.73	-56.87	-22.59	0.00

Math	diff	lwr	upr	p adj
Arch-Tech	3.79	-4.78	12.37	0.54
MedTech-Tech	-9.30	-19.09	0.50	0.07
MedTech-Arch	-13.09	-21.92	-4.26	0.00

Lang	diff	lwr	upr	p adj
Arch-Tech	6.68	1.78	11.58	0.00
MedTech-Tech	-2.47	-8.06	3.13	0.55
MedTech-Arch	-9.15	-14.19	-4.11	0.00

GenKnow	diff	lwr	upr	p adj
Arch-Tech	2.31	-3.65	8.27	0.63
MedTech-Tech	7.97	1.17	14.78	0.02
MedTech-Arch	5.66	-0.47	11.80	0.08

Apt: Arch different from Tech and MedTech

Math: MedTech different from Arch and Tech

Lang: Arch different from Teach and MedTech

GenKnow: MedTech different from Tech

Summary

In general, using a multivariate test, such as MANOVA, is preferable to multiple univariate tests, such as ANOVA. This controls for the correlation structure in the data. Its possible to get non-significant results in the univariate tests, but significant results in the multivariate tests. That is, the univariate tests may not detect the differences, in the presence of correlation between the responses.