SAS 4:

ANOVA with Strip-Plot Example

University of Guelph

Table of Contents

SAS Availability

Faculty, staff and students at the University of Guelph may access SAS three different ways:

1. **Library computers**

   On the library computers, SAS is installed on all machines.

2. **Acquire a copy for your own computer**
   If you are faculty, staff or a student at the University of Guelph, you may obtain the site-licensed standalone copy of SAS at a cost. However, it may only be used while you are employed or a registered student at the University of Guelph. To obtain a copy, go to the CCS Software Distribution Site (**www.uoguelph.ca/ccs/download**).

3. **Central statistical computing server**
   SAS is available in batch mode on the UNIX servers (stats.uoguelph.ca) or through X-Windows.

Goals of the workshop

This workshop builds on the skills and knowledge developed in "Getting your data into SAS". Participants are expected to have basic SAS skills and statistical knowledge. This workshop will help you work through the analysis of a Strip-Plot and a Repeated Measures experimental design using both the GLM and MIXED procedures available in SAS. Specific goals:

1. To review how to build a model for a Strip-plot and a Repeated Measures experimental design
2. To learn how to build the same model in **Proc GLM** and **Proc MIXED**
3. To discover the differences between the two procedures
4. To develop a familiarity of when each procedure should be used and the correct model

Overview of Procedures in SAS used for Analysis of Variance

ANOVA

Performs analysis of variance for balanced designs. The ANOVA procedure is generally more efficient than Proc GLM for these types of designs. SAS cautions users that use this Procedure: "**Caution:** If you use PROC ANOVA for analysis of unbalanced data, you must assume responsibility for the validity of the results." (SAS 2007)

GLM

The GLM procedure is used to analyze data in the context of a **G**eneral **L**inear **M**odel (**GLM**). The SAS documentation states: "PROC GLM handles models relating one or several continuous dependent variables to one or several independent variables. The independent variables may be either *classification* variables, which divide the observations into discrete groups, or *continuous* variables." (SAS 2007)

MIXED

The MIXED procedure was the next generation of Procedures dealing with ANOVA. MIXED fits mixed models by incorporating covariance structures in the model fitting process. Some options available in MIXED are very similar to GLM but offer different functionalities.

NESTED

The NESTED procedure performs ANOVA and estimates variance components for nested random models. This procedure is generally more efficient than Proc GLM for nested models.

NPAR1WAY

Performs nonparametric one-way analysis of rank scores. This can also be done using the RANK statement in GLM.

REG

Performs simple linear regression. The REG procedure allows several MODEL statements and gives additional regression diagnostics, especially for detection of collinearity. Proc REG also creates plots of model summary statistics and regression diagnostics.

RSREG

This procedure performs quadratic response-surface regression, canonical and ridge analysis. The RSREG procedure is generally recommended for data from a response surface experiment.

TTEST

The TTEST procedure compares the means of two groups of observations. It can also test for equality of variance for the two groups. The TTEST procedure is usually more efficient than Proc GLM for this type of data.

VARCOMP

Estimates variance components for a general linear model.

For more advanced models there are a number of Procedures available today. Take a look at the SAS online documentation available at: http://support.sas.com/onlinedoc/913/docMainpage.jsp for more information

Proc GLM Syntax (source: SAS online documentation)

PROC GLM < options > ;
       CLASS variables < / option > ;
       MODEL dependents=independents < / options > ;
       ABSORB variables ;
       BY variables ;
       FREQ variable ;
       ID variables ;
       WEIGHT variable ;
       CONTRAST 'label' effect values < … effect values > < / options > ;
       ESTIMATE 'label' effect values < … effect values > < / options > ;
       LSMEANS effects < / options > ;
       MANOVA < test-options >< / detail-options > ;
       MEANS effects < / options > ;
       OUTPUT < **OUT=**SAS-data-set >
         keyword=names < … keyword=names > < / option > ;
       RANDOM effects < / options > ;
       REPEATED factor-specification < / options > ;
       TEST < **H=**effects > **E=**effect < / options > ;

Although there are numerous statements and options available in PROC GLM, many applications use only a few of them. Often you can find the features you need by looking at an example or by quickly scanning through this section.

To use PROC GLM, the PROC GLM and MODEL statements are required. You can specify only one MODEL statement (in contrast to the REG procedure, for example, which allows several MODEL statements in the same PROC REG run). If your model contains classification effects, the classification variables must be listed in a CLASS statement, and the CLASS statement must appear before the MODEL statement. In addition, if you use a CONTRAST statement in combination with a MANOVA, RANDOM, REPEATED, or TEST statement, the CONTRAST statement must be entered first in order for the contrast to be included in the MANOVA, RANDOM, REPEATED, or TEST analysis.

The following table summarizes the positional requirements for the statements in the GLM procedure.

Positional Requirements for PROC GLM Statements

Statement	Must Appear Before the	Must Appear After the	Statement	Must Appear Before the	Must Appear After
ABSORB	first RUN statement		TEST	MANOVA or	MODEL statement
BY	first RUN statement			REPEATED statement	
			WEIGHT	first RUN statement	
CLASS	MODEL statement				
CONTRAST	MANOVA, REPEATED,	MODEL statement			
	or RANDOM statement				
ESTIMATE		MODEL statement			
FREQ	first RUN statement				
ID	first RUN statement				
LSMEANS		MODEL statement			
MANOVA		CONTRAST or			
		MODEL statement			
MEANS		MODEL statement			
MODEL	CONTRAST, ESTIMATE,	CLASS statement			
	LSMEANS, or MEANS				
	statement				
OUTPUT		MODEL statement			
RANDOM		CONTRAST or			
		MODEL statement			
REPEATED		CONTRAST, MODEL, or TEST statement			

The following table summarizes the function of each statement (other than the PROC statement) in the GLM procedure:

Statements in the GLM Procedure

Statement	Description
ABSORB	absorbs classification effects in a model
BY	specifies variables to define subgroups for the analysis
CLASS	declares classification variables
CONTRAST	constructs and tests linear functions of the parameters
ESTIMATE	estimates linear functions of the parameters
FREQ	specifies a frequency variable
ID	identifies observations on output
LSMEANS	computes least-squares (marginal) means
MANOVA	performs a multivariate analysis of variance
MEANS	computes and optionally compares arithmetic means
MODEL	defines the model to be fit
OUTPUT	requests an output data set containing diagnostics for each observation
RANDOM	declares certain effects to be random and computes expected mean squares
REPEATED	performs multivariate and univariate repeated measures analysis of variance
TEST	constructs tests using the sums of squares for effects and the error term you specify
WEIGHT	specifies a variable for weighting observations

Proc MIXED Syntax (source: SAS online documentation)

PROC MIXED < options > ;
      BY variables ;
      CLASS variables ;
      ID variables ;
      MODEL dependent = < fixed-effects > < / options > ;
      RANDOM random-effects < / options > ;
      REPEATED < repeated-effect > < / options > ;
      PARMS (value-list) … < / options > ;
      PRIOR < distribution > < / options > ;
      CONTRAST'label'<fixed-effect values…>
                          < | random-effect values … > , … < / options > ;
      ESTIMATE 'label' < fixed-effect values … >
                          < | random-effect values … >< / options > ;
      LSMEANS fixed-effects < / options > ;
      WEIGHT variable ;

Items within angle brackets (< >) are optional. The CONTRAST, ESTIMATE, LSMEANS, and RANDOM statements can appear multiple times; all other statements can appear only once.

The PROC MIXED and MODEL statements are required, and the MODEL statement must appear after the CLASS statement if a CLASS statement is included. The CONTRAST, ESTIMATE, LSMEANS, RANDOM, and REPEATED statements must follow the MODEL statement. The CONTRAST and ESTIMATE statements must also follow any RANDOM statements.

The following table summarizes the statements of Proc MIXED

Statement	Description	Important Options
PROC MIXED	invokes the procedure	DATA= specifies input data set, METHOD= specifies estimation method
BY	performs multiple PROC MIXED analyses in one invocation	none
CLASS	declares qualitative variables that create indicator variables in design matrices	none
ID	lists additional variables to be included in predicted values tables	none
MODEL	specifies dependent variable and fixed effects, setting up X	S requests solution for fixed-effects parameters, DDFM= specifies denominator degrees of freedom method, OUTP= outputs predicted values to a data set, INFLUENCE computes influence diagnostics
RANDOM	specifies random effects, setting up Z and G	SUBJECT= creates block-diagonality, TYPE= specifies covariance structure, S requests solution for random-effects parameters, G displays estimated G
REPEATED	sets up R	SUBJECT= creates block-diagonality, TYPE= specifies covariance structure, R displays estimated blocks of R, GROUP= enables between-subject heterogeneity, LOCAL adds a diagonal matrix to R
PARMS	specifies a grid of initial values for the covariance parameters	HOLD= and NOITER hold the covariance parameters or their ratios constant, PDATA= reads the initial values from a SAS data set
PRIOR	performs a sampling-based Bayesian analysis for variance component models	NSAMPLE= specifies the sample size, SEED= specifies the starting seed
CONTRAST	constructs custom hypothesis tests	E displays the L matrix coefficients
ESTIMATE	constructs custom scalar estimates	CL produces confidence limits

Statement	Description	Important Options
LSMEANS	computes least squares means for classification fixed effects	DIFF computes differences of the least squares means, ADJUST= performs multiple comparisons adjustments, AT changes covariates, OM changes weighting, CL produces confidence limits, SLICE= tests simple effects
WEIGHT	specifies a variable by which to weight R	none

Review: Building a model for ANOVA

What is an Analysis of Variance or ANOVA? An analysis that partitions the variation we see in the dependent variable, or the data we have measured, into variation between and within groups or classes of observations.

Let us look at a sample dataset – a 2-by-2 Factorial design.

We have 2 factors – gender and treatment – each with 2 levels.

		Gender	
		Male	Female
Treatment	A	12, 14	20,18
	B	11, 9	17

The model used for a 2x2 factorial design:

$Y_{ijk} = \mu + Gender_i + Treatment_j + Gender*Treatment_{ij} + error_k$

Where:

Y_{ijk} = individual observation (for example: 12, 14, 20, etc…)

μ = overall mean
$Gender_i$ = effect of gender
$Treatment_j$ = effect of treatment
$Gender*Treatment_{ij}$ = the effect of the interaction between gender and treatment
$error_k$ = random error

ANOVA Table

Source	Degrees of Freedom	Sum of squares	Mean square	F
TOTAL				
Gender				
Treatment				
Gender x Treatment				
ERROR				

SAS code:

```
Data exp;
  input gender treatment Y;
  datalines;
1 1 12
1 1 14
1 2 11
1 2  9
2 1 20
2 1 18
2 2 17
;
Run;

Proc glm data=exp;
  class gender treatment;
  model Y = gender treatment gender*treatment;
Run;
```

Class statement – this is where you list the variables that identify which groups the observations fall into. Another way of looking at this – these are your independent variables or the factors in your model. In this example, that would be our **gender** and **treatment**.

Model statement – notice that this statement is a direct "translation" of the statistical model noted above. The model statement will include only the information that is listed in your dataset.

SAS Output:

```
                  The GLM Procedure

              Class Level Information

        Class          Levels     Values

        gender              2      1 2

        treatment           2      1 2
```

The first page of the output describes the variables listed in your **Class** statement.

There are 2 levels for gender and 2 levels for treatment.

Use the information provided on this page to double-check that SAS read your data correctly.

```
                    Number of Observations Read          7
                    Number of Observations Used          7
                              The GLM Procedure

Dependent Variable: Y

                                   Sum of
       Source              DF      Squares      Mean Square    F Value    Pr > F

       Model                3    91.71428571     30.57142857      15.29    0.0253

       Error                3     6.00000000      2.00000000

       Corrected Total      6    97.71428571


            R-Square      Coeff Var      Root MSE       Y Mean

            0.938596      9.801480      1.414214      14.42857


       Source              DF     Type I SS      Mean Square    F Value    Pr > F

       gender               1    80.04761905     80.04761905      40.02    0.0080
       treatment            1    11.26666667     11.26666667       5.63    0.0982
       gender*treatment     1     0.40000000      0.40000000       0.20    0.6850


       Source              DF    Type III SS     Mean Square    F Value    Pr > F

       gender               1    67.60000000     67.60000000      33.80    0.0101
       treatment            1    10.00000000     10.00000000       5.00    0.1114
       gender*treatment     1     0.40000000      0.40000000       0.20    0.6850
```

1.

This is the significance of the Model.

If the p-value is <0.05 – then we say that the model is able to explain a significant amount of the variation in the dependent variable – **Y** in this example

2.

This describes the model.

R-square of 0.9386 – which says we can only explain ~ 94% of the variation in **Y** with this model.

3.

Type I vs. Type III Sum of Squares – which one do we use?

Type I should only be used for a balanced design, whereas Type III can be used for either a balanced or unbalanced design. Type III will take into account the number of observations in each treatment group. Notice the different results here – with an unbalanced design.

4.

These p-values tell you whether there are differences within each factor listed – or whether the factor is significantly contributing to explaining the variation among **Y**

If the p-value is < 0.05 then there are differences among the factor levels – however a PostHoc or means comparisons needs to be conducted to examine where the differences lie.

With a factorial design – look at the p-values from the bottom up – remember if there's a significant interaction – you need to look at the simple effects and NOT the main effects.

Conclusion:

How do we write the results of this analysis? How do we answer the research question? Do we present a table for our results? If so – what do we present? If not, why and what do we report?

Strip -plot Experimental Design

Example (source:SAS Course – Advanced General Linear Models with an emphasis on Mixed Models)

A scientist wants to compare the effects of two types of insulin and of diets containing one of three substances (aspartame, sugar and saccharin).

Eighteen cages, each containing four rats, are used for the experiment. The cages are arranged on three tables in two rows of three cages stacked on top of one another. For each table, the scientist randomly assigns one of the three diets to a stack (column) of two cages. He also administers, at random, each insulin type to one of the two rows of cages. The glucose level if measured five hours after the injection.
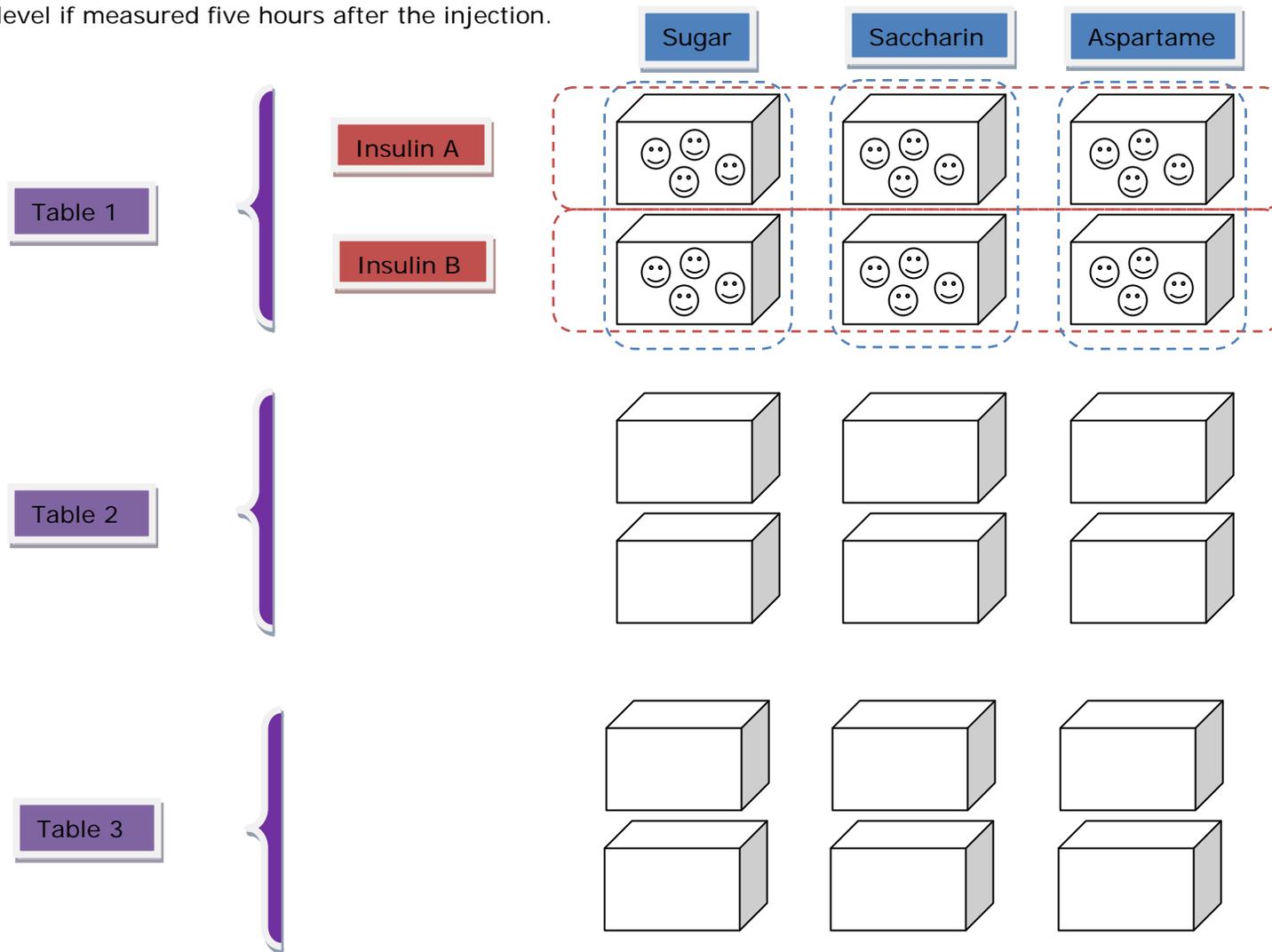
Experimental Units

What is an experimental unit?

How many different experimental units are there in this design?

What are they?

What is the implication of different experimental units on the analysis?

Insulin treatment

Experimental unit:

Error term:

Diet treatment

Experimental unit:

Error term:

Insulin x Diet interaction

Experimental unit:

Error term:

16

Statistical Model (Linear Model)

$$Y_{ijkl} = \mu + b_i + a_j + (ba)_{ij} + \beta_k + (b\beta)_{ik} + (a\beta)_{jk} + \varepsilon_{ijk} + \gamma_{ijkl}$$

Where:

Y_{ijkl} = glucose observation on the lth rat in a cage on the ith table receiving j^{th} level of insulin and k^{th} level of diet

μ = overall mean

b_i = effect of the ith table, a blocking variable

a_j = effect of the jth level of insulin

$(ba)_{ij}$ = interaction between \the ith table and the jth level of insulin

β_k = effect of the kth level of diet

$(b\beta)_{ik}$ = interaction between the ith table and the kth level of diet

$(a\beta)_{jk}$ = interaction between the jth level of insulin and the kth level of diet

ε_{ijk} = experimental error. The error corresponding to variation between cages or cage-to-cage (cell)

γ_{ijkl} = within-cage error. The error associated with the variation observed between rats in a cage.

Fixed vs. Random effects

INSERT: Quick definitions for a fixed and a random effect

Label each factor in our model as either Fixed or Random

Factor	Fixed or Random
μ	
b_i	
a_j	
$(ba)_{ij}$	
β_k	
$(b\beta)_{ik}$	
$(a\beta)_{jk}$	
ε_{ijk}	
γ_{ijkl}	

Review: Reading a SAS Data Set

There may be situations where you may need to read from a permanent SAS data set to conduct your data analysis. This will require the use of the SET and LIBNAME statement. The SET statement refers to the filename of the permanent SAS data set and LIBNAME refers to the location of the SAS data set. In this example, the DATA step creates data set EXP2 by reading data from data set PERM.BTEMPHRT.

```
libname PERM 'C:\Documents and Settings\Desktop\IntroSAS';
DATA EXP2;
      set PERM.btemphrt;
RUN;
```

Review: Printing a SAS Data Set

PROC PRINT

This procedure either prints to screen all the observations or a subset of a specified SAS data set. The general syntax is as follows:

```
PROC PRINT data=dataset;
RUN;
```

Limiting observations when printing:
The following example specifies within PROC PRINT to display the first 50 observations in the data set. The obs keyword specifies the last observation to display.

```
PROC PRINT data=PERM.employee (obs=50);
RUN;
```

To print a subset of data to screen, specify the first observation by using firstobs keyword and the last observation with the obs keyword. That is if you wish to output to screen observations 10 to 43, the code would be as follows:

```
PROC PRINT data=PERM.employee (firstobs = 20 obs = 50);
RUN;
```

SAS code:

Look at the data graphically first, to get a sense of how your data looks and what you may find when you run the statistical analysis. We'll use a PROC MEANS and PROC GPLOT:

```
Proc means data=insulin nway noprint;
  class table diet insulin;
  var glucose;
  output out=meanins mean=meanins;
Run;
```

Proc means – we want to create a new dataset with the mean glucose value for each table-diet-insulin combination. The **nway** option in the Proc means statement will give us mean values for each combination of class variables. In this example we will get a mean for table1-sugar-A, table2-sugar-A, table3-sugar-A, table1-saccharin-A, etc…

noprint - prevents SAS from creating an output table.

Class statement – this is where you list the variables that identify which groups the observations fall into. Another way of looking at this… list the independent variables in the Class statement.

Var statement – this is where you list the variables that you are testing – in other words, the variables you would like to calculate a mean for.

Output statement – tells SAS to create a new dataset with the results of the **Proc means** procedure. The **out=** specifies the name of the new SAS dataset and the **mean=** creates a new variable that containst the mean. In this example the new variable will be called **meanins.**

```
Proc gplot data=meanins;
  plot meanins*diet=insulin / haxis=axis1 vaxis=axis2;
  by table;
  symbol v='A' c=black h=2;
  symbol v='B' c=black h=2;
  axis1 minor=none offset=(2,2);
  axis2 minor=none;
Run;
Quit;
```

Proc gplot - is one of the many graphing procedures of SAS/GRAPH. The **Proc gplot** will create a scatter plot with options to allow the users to join the dots.

Plot – identifies the graph you want to create. In this example we are looking to create a plot of the mean insulin value (Y-value) by diet (X-value) for each insulin treatment.

Haxis and **vaxis** – options to customize the horizontal (**haxis**) and vertical (**vaxis**) axes.

Symbol - are options to customize the symbols used on the graph.

Note: The Proc gplot will open the Graph window in the SAS program and will result in three plots – one for each table.

```
Proc glm data=insulin;
  class table diet insulin;
  model glucose = table|diet|insulin;
  random table table*diet table*insulin table*diet*insulin;
Run;
Quit;
```

Class statement – this is where you list the variables that identify which groups the observations fall into. Another way of looking at this – these are your independent variables or the factors in your model. In this example, that would be our **table**, **diet** and **insulin**.

Model statement – notice that this statement is a direct "translation" of the statistical model noted above. The model statement will include only the information that is listed in your dataset.

21

Random statement – this is where you list the factors of your model that you've identified as

SAS output:

```
                        The GLM Procedure

                    Class Level Information

        Class          Levels    Values

        table               3    1 2 3

        diet                3    aspartame saccharin sugar

        insulin             2    A B


              Number of Observations Read          72
              Number of Observations Used          72


Dependent Variable: glucose

                                   Sum of
   Source               DF        Squares     Mean Square    F Value    Pr > F

   Model                17     9.22104028      0.54241413       6.78    <.0001

   Error                54     4.31842500      0.07997083

   Corrected Total      71    13.53946528


        R-Square     Coeff Var     Root MSE     glucose Mean

        0.681049     5.963786      0.282791       4.741806


   Source               DF       Type I SS     Mean Square    F Value    Pr > F

   table                 2     0.54495278      0.27247639       3.41    0.0404
   diet                  2     1.88611944      0.94305972      11.79    <.0001
```

The first page of the output describes the variables listed in your **Class** statement.

There are 3 levels for table, 3 levels for diet and 2 levels for insulin.

Use the information provided on this page to double-check that SAS read your data correctly.

1.
This is the significance of the Model.

If the p-value is <0.05 – then we say that the model is able to explain a significant amount of the variation in the dependent variable – **glucose** in this example

2.
This describes the model.

R-square of 0.6820 – which says we can only explain ~ 68% of the variation in **glucose** with this model.

3.
Type I vs. Type III Sum of Squares – which one do we use?

Type I should only be used for a balanced design, whereas Type III can be used for either a balanced or unbalanced design. Type III will take into account the number of observations in each treatment group. Notice with a balanced design the results are the same.

table*diet	4	0.99273056	0.24818264	3.10	0.0227
insulin	1	4.76890139	4.76890139	59.63	<.0001
table*insulin	2	0.33385278	0.16692639	2.09	0.1339
diet*insulin	2	0.01093611	0.00546806	0.07	0.9340
table*diet*insulin	4	0.68354722	0.17088681	2.14	0.0887

Source	DF	Type III SS	Mean Square	F Value	Pr > F
table	2	0.54495278	0.27247639	3.41	0.0404
diet	2	1.88611944	0.94305972	11.79	<.0001
table*diet	4	0.99273056	0.24818264	3.10	0.0227
insulin	1	4.76890139	4.76890139	59.63	<.0001
table*insulin	2	0.33385278	0.16692639	2.09	0.1339
diet*insulin	2	0.01093611	0.00546806	0.07	0.9340
table*diet*insulin	4	0.68354722	0.17088681	2.14	0.0887

4.
These p-values tell you whether there are differences within each factor listed – or whether the factor is significantly contributing to explaining the variation among **glucose**.

If the p-value is < 0.05 then there are differences among the factor levels – however a PostHoc or means comparisons need to be analyzed.

Source	Type III Expected Mean Square
table	Var(Error) + 4 Var(table*diet*insulin) + 12 Var(table*insulin) + 8 Var(table*diet) + 24 Var(table)
diet	Var(Error) + 4 Var(table*diet*insulin) + 8 Var(table*diet) + Q(diet,diet*insulin)
table*diet	Var(Error) + 4 Var(table*diet*insulin) + 8 Var(table*diet)
insulin	Var(Error) + 4 Var(table*diet*insulin) + 12 Var(table*insulin) + Q(insulin,diet*insulin)
table*insulin	Var(Error) + 4 Var(table*diet*insulin) + 12 Var(table*insulin)
diet*insulin	Var(Error) + 4 Var(table*diet*insulin) + Q(diet*insulin)
table*diet*insulin	Var(Error) + 4 Var(table*diet*insulin)

5.
What is the experimental unit in this experiment? See p. 12-13. What are the implications of these? Are the results presented in this output correct? We specified a **random** statement – we see the output above – what do we do with this?

When SAS ran the **Proc glm** code – there was no obvious way to inform SAS what experimental units were used for each treatment effect, therefore SAS used the lowest unit in the dataset for the error term. Remember the dataset lists the observations by rat – so SAS assumes the experimental unit for the analysis is the individual rat. Since we are conducting a strip-plot design we know that this is NOT the case. Each treatment has a separate experimental unit – see p 12-13 to review.

To capture this information we need to add **test** statements to the **Proc glm** code. Add one test statement for each treatment effect: diet, insulin, and the diet*insulin interaction.

Test statement – states the hypothesis (the treatment effect we are testing) along with the correct error term that should be used to test the hypothesis.

To test the **Diet** effect:

Null hypothesis:

$H_o : \mu_{sugar} = \mu_{saccharin} = \mu_{aspartame}$

Alternate hypothesis:

$H_a : \mu_{sugar} \neq \mu_{saccharin} \neq \mu_{aspartame}$

Our experimental unit for **diet** was the column of cages – not the 8 rats within the cages. The correct error term would be **table*diet**. To test differences between the diets – we use the variation among the diets between the tables as the error.

SAS code:

Test h=diet e=table*diet;

To test the **insulin** effect:

Null hypothesis:
$$H_o : \mu_A = \mu_B$$

Alternate hypothesis:
$$H_a : \mu_A \neq \mu_B$$

Our experimental unit for **insulin** was the row of cages – not the 12 rats within the cages. The correct error term would be **table*insulin**. To test differences between the two different insulins – we use the variation among the insulin treatments between the tables as the error.

SAS code:

Test h=insulin e=table*insulin;

To test the interaction between **diet*insulin** effect:

Null hypothesis:
$$H_o : \mu_{ij} = \mu_{ij}$$

Alternate hypothesis:
$$H_a : \mu_{ij} \neq \mu_{ij}$$

Our experimental unit for **diet*insulin** is the individual cage and not the 4 rats within the cage. The correct error term would be **table*diet*insulin**.

SAS code:

Test h=diet*insulin e=table*diet*insulin;

25

Updated SAS Code:

```
Proc glm data=insulin;
  class table diet insulin;
  model glucose = table|diet|insulin;
  random table table*diet table*insulin table*diet*insulin;
  test h=diet e=table*diet;
  test h=insulin e=table*insulin;
  test h=diet*insulin e=table*diet*insulin;
Run;
Quit;
```

SAS output: (same first 4 pages as above – with the addition of this output)

```
                        The GLM Procedure

Dependent Variable: glucose

      Tests of Hypotheses Using the Type III MS for table*diet as an Error Term

   Source                 DF     Type III SS    Mean Square   F Value    Pr > F

   diet                    2     1.88611944     0.94305972       3.80    0.1189
```

This p-values tell you whether there are differences among the levels of **diet** – using the columns of diet as the experimental unit and applying the **table*diet** as the error term.

```
       Tests of Hypotheses Using the Type III MS for table*insulin as an Error Term

   Source                 DF     Type III SS    Mean Square   F Value    Pr > F

   insulin                 1     4.76890139     4.76890139      28.57    0.0333
```

This p-values tell you whether there are differences among the levels of **insulin**– using the columns of diet as the experimental unit and applying the **table*insulin** as the error term.

```
      Tests of Hypotheses Using the Type III MS for table*diet*insulin as an Error Term

   Source                 DF     Type III SS    Mean Square   F Value    Pr > F

   diet*insulin            2     0.01093611     0.00546806       0.03    0.9688
```

This p-values tell you whether there are differences among the levels of **diet*insulin** – using the columns of diet as the experimental unit and applying the **table*diet*insulin** as the error term.

Please take note of the different p-values between these test results and the overall model results.

Note that you can draw incorrect conclusions if the correct error terms are not used for your experimental design.

By using **Proc glm,** you must be vigilant of the experimental design you used, develop the correct model and develop the correct hypotheses tests for **test** statements. In any analysis the model is key to ensure proper analysis. **Proc mixed,** the next generation of ANOVA tools is another tool to use when analysing experimental data.

SAS code:

```
Proc mixed data=insulin;
  class table diet insulin;
  model glucose = diet insulin diet*insulin;
  random table table*diet table*insulin table*diet*insulin;
Run;
```

Notice the similarity in the code used for **Proc glm** and **Proc mixed**. The class statement is used in the same way for both procedures, however the model statement is different. When using **Proc mixed,** the model statement contains ONLY the fixed effects. With **Proc glm**, the entire model is included in the **model** statement. The **random** statement contains only the random effects – similar purpose and use in both procedures. With **Proc mixed** you do not need to include **test** statements.

SAS output:

The output is very different than the **Proc glm** output.

```
                      The Mixed Procedure


                      Model Information


    Data Set                    LIBRARY.INSULIN
    Dependent Variable          glucose
    Covariance Structure        Variance Components
    Estimation Method           REML
    Residual Variance Method    Profile
    Fixed Effects SE Method     Model-Based
    Degrees of Freedom Method   Containment
```

> This information provides us with an overview of how the analysis was conducted.
> The name of the dataset **LIBRARY.INSULIN,** dependent variable, and estimation methods

```
                    Class Level Information

    Class      Levels    Values

    table          3     1 2 3
    diet           3     aspartame saccharin sugar
    insulin        2     A B

                        Dimensions

        Covariance Parameters            5
        Columns in X                    12
        Columns in Z                    36
        Subjects                         1
        Max Obs Per Subject             72


                   Number of Observations

        Number of Observations Read          72
        Number of Observations Used          72
        Number of Observations Not Used       0


                     Iteration History

    Iteration    Evaluations    -2 Res Log Like     Criterion

          0             1          52.91864023
          1             3          46.97846584     0.00000001


                  Convergence criteria met.


                 Covariance Parameter Estimates

        Cov Parm                 Estimate

        table                    0.001008
        table*diet               0.009829
```

Similar to the **Proc glm** output this part of the output describes the variables listed in your **Class** statement.

Information about the matrices used in the analysis

Iteration history – remember **Proc mixed** uses an iterative process for the analysis.

Covariance Parameter Estimates – provides estimates of the random effects included in your model

28

```
              table*insulin                 0
              table*diet*insulin      0.02239
              Residual                0.07997


                      Fit Statistics

              -2 Res Log Likelihood          47.0
              AIC (smaller is better)        55.0
              AICC (smaller is better)       55.6
              BIC (smaller is better)        51.4


                  Type 3 Tests of Fixed Effects

                      Num      Den
          Effect       DF       DF     F Value     Pr > F

          diet          2        4        3.80     0.1189
          insulin       1        2       28.13     0.0338
          diet*insulin  2        4        0.03     0.9685
```

Fit Statistics – use these when comparing models

Type 3 test of Fixed Effects – this is where most of us will be interested in – please note the similar results as the second Proc glm run on p. 23

Conclusion:

How do we write the results of this analysis? How do we answer the research question? Do we present a table for our results? If so – what do we present? If not, why and what do we report?

ADDITIONAL DESIGNS

A. Split-plot Experimental Design

Let's rearrange the experiment so we still have 18 cages – but now we are randomly assigning the three diets to the columns of cages (same as the Strip-plot), but we will randomly assign the insulin treatment to the two cages within each diet.
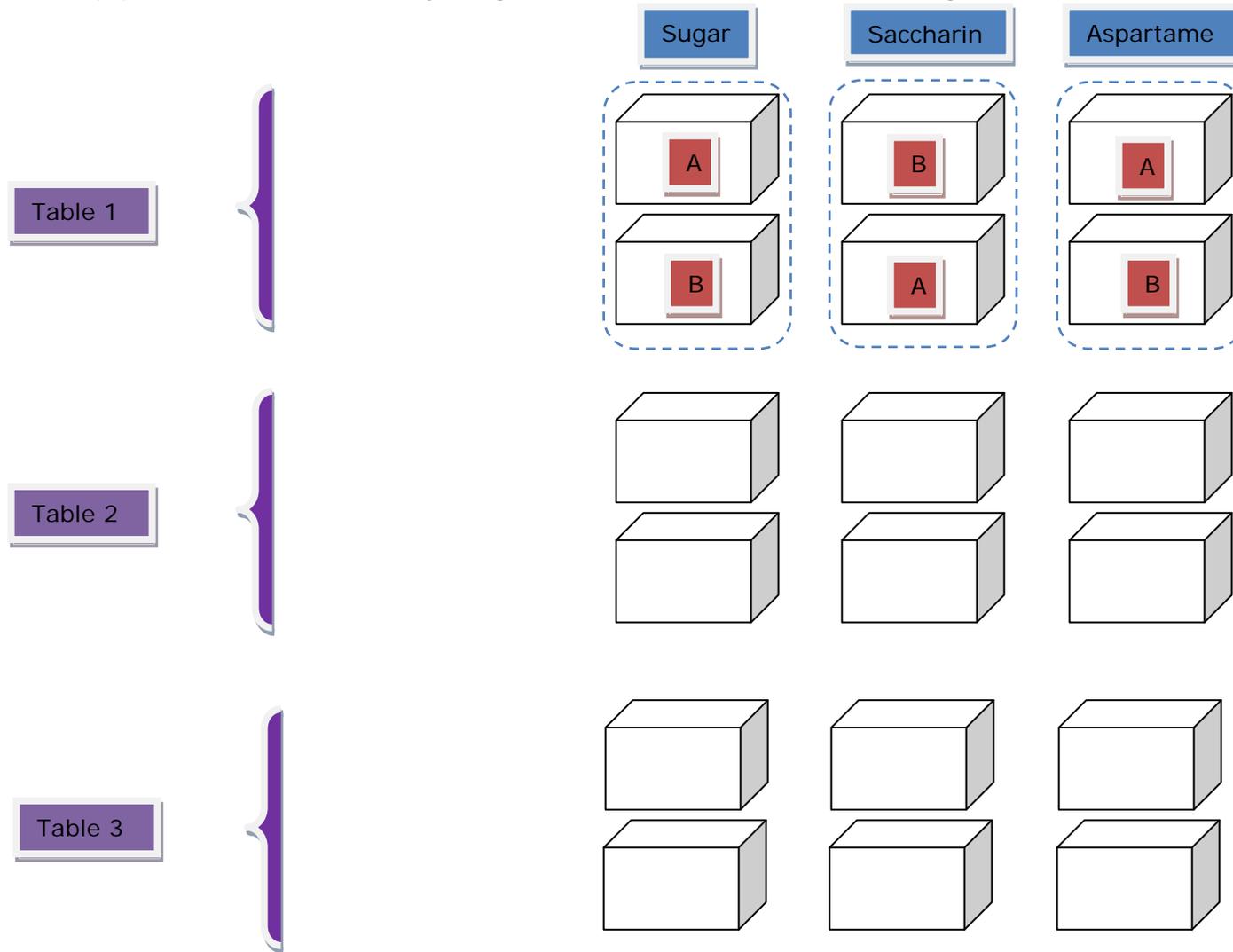
Questions

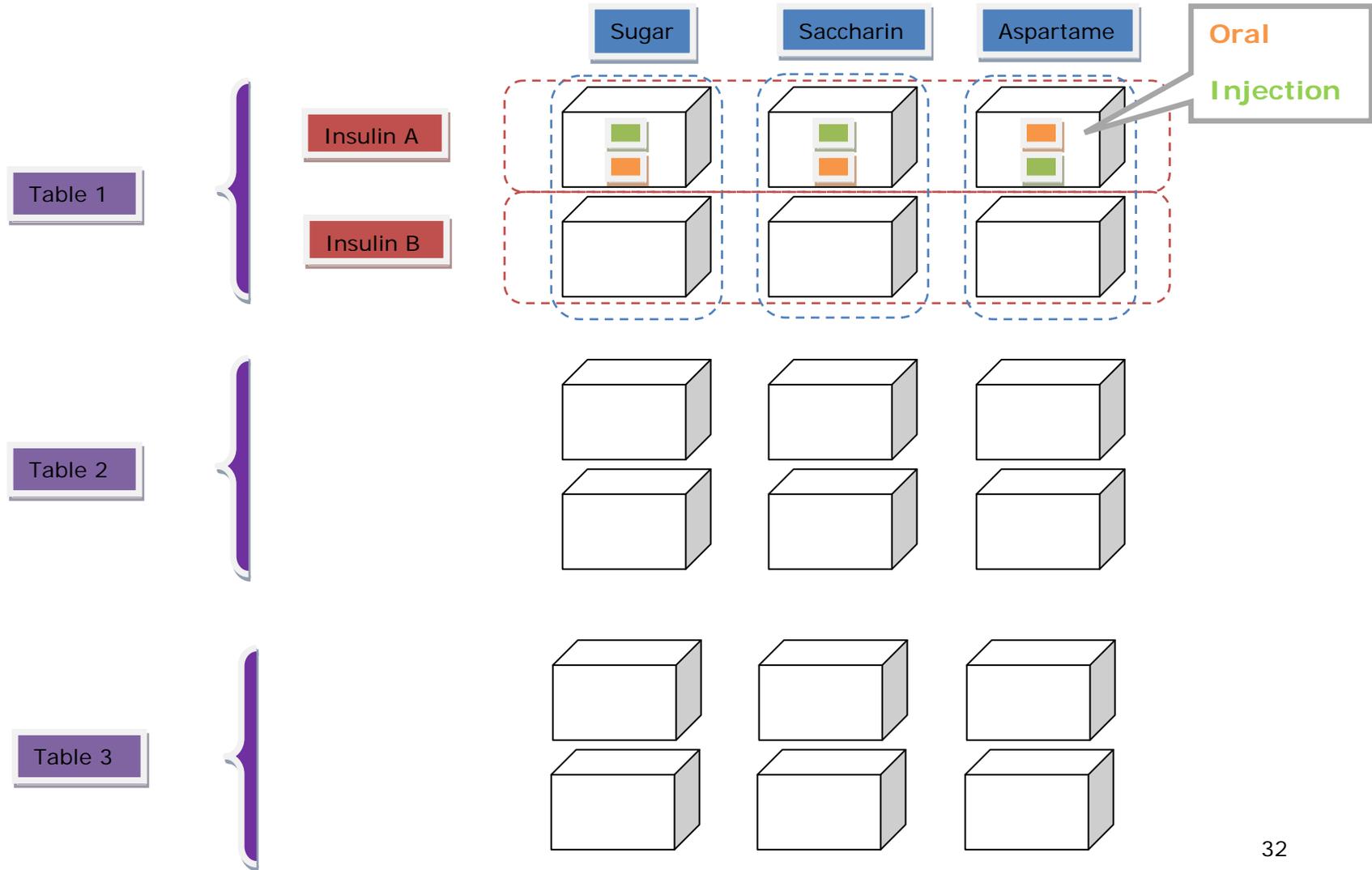How many different experimental units are there in this design? What are they?

What is the statistical model for this design?

For each experimental unit – what is the correct error term?

Write the appropriate Proc GLM followed by Proc MIXED SAS code:

B. Strip Split-plot Experimental Design

Now let's combine the Strip-plot and Split-plot designs to create a Strip-split-plot design. Same design as the Strip-plot, but adding a third treatment effect – administration method (oral or by injection). Two of the four rats in each cage are randomly selected to receive the insulin treatment orally or by injection.

Questions

How many different experimental units are there in this design? What are they?

What is the statistical model for this design?

For each experimental unit – what is the correct error term?

Write the appropriate Proc GLM followed by Proc MIXED SAS code:

33

APPENDIX

A. Split-plot Experimental Design answers

How many different experimental units are there in this design? What are they?

- There are 3 experimental units
- Table will remain as the blocking variable
- Diets are randomly assigned to the column of cages – making the column of cages the experimental unit for diet. The diet factor is also referred to as the *Whole Plot Factor*
- Insulin is randomly assigned to a cage within the column of diets – making a cage within the diet column the experimental unit for insulin. The insulin factor is also referred to as the *Subplot Factor*
- The experimental unit for the diet*insulin effect will be the individual cage.

What is the statistical model for this design?

$$Y_{ijkl} = \mu + b_i + \alpha_j + (b\alpha)_{ij} + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

Where:

Y_{ijkl} = glucose observation on the lth rat in a cage on the ith table receiving j^{th} level of insulin and k^{th} level of diet

μ = overall mean

b_i = effect of the ith table, a blocking variable

α_j = effect of the jth level of diet

$(b\alpha)_{ij}$ = interaction between the ith table and the jth level of diet

β_k = effect of the kth level of insulin

$(\alpha\beta)_{jk}$ = interaction between the jth level of diet and the kth level of insulin

ε_{ijk} = experimental error. The error corresponding to variation between rats in cages

For each experimental unit – what is the correct error term?

- Diet – the correct error term is the interaction between table and diet -> $(b\alpha)_{ij}$
- Insulin – the correct error term is -> experimental error
- Diet*insulin – the correct error term is -> experimental error

Write the appropriate Proc GLM followed by Proc MIXED SAS code:

PROC GLM

```
Proc glm data=insulin;
  class table diet insulin;
  model glucose = table diet table*diet insulin diet*insulin;
  random table table*diet;
  test h=diet e=table*diet;
Run;
Quit;
```

PROC MIXED

```
Proc mixed data=insulin;
  class table diet insulin;
  model glucose = diet insulin diet*insulin;
  random table table*diet;
Run;
```

B. Strip split-plot Experimental Design answers

How many different experimental units are there in this design? What are they?

- There are 4 experimental units
- Table will remain as the blocking variable
- Diets are randomly assigned to the column of cages – making the column of cages the experimental unit for diet.
- Insulin is randomly assigned to the row of cages – making the row of cages the experimental unit for insulin.
- The experimental unit for the diet*insulin effect will be the individual cage.
- Form of insulin was randomly assigned to rats within a cage – therefore the experimental unit for Form of insulin is the individual rat.

What is the statistical model for this design?

$$Y_{ijkl} = \mu + b_i + a_j + (ba)_{ij} + \beta_k + (b\beta)_{ik} + (a\beta)_{jk} + \Omega_l + (a\Omega)_{jl} + (\beta\Omega)_{kl} + (a\beta\Omega)_{jkl} + \varepsilon_{ijk} + \gamma_{ijkl}$$

Where:

Y_{ijkl} = glucose observation on the lth rat in a cage on the ith table receiving j^{th} level of insulin and k^{th} level of diet

μ = overall mean

b_i = effect of the ith table, a blocking variable

a_j = effect of the jth level of diet

$(ba)_{ij}$ = interaction between the ith table and the jth level of diet

β_k = effect of the kth level of insulin

$(b\beta)_{ik}$ = interaction between the ith table and the kth level of insulin

$(a\beta)_{jk}$ = interaction between the jth level of diet and the kth level of insulin

ε_{ijk} = experimental error. The error corresponding to variation between cages (table*diet*insulin)

γ_{ijkl} = within-cage error. The error associated with the variation observed between rats in a cage.

For each experimental unit – what is the correct error term?

- Diet – the correct error term is the interaction between table and diet
- Insulin – the correct error term is the interaction between table and insulin
- Diet*insulin – the correct error term is the interaction between table,diet, insulin -> experimental error
- Form – the correct error term is -> within-cage error
- Form – the correct error term is -> within-cage error
- Form – the correct error term is -> within-cage error
- Form – the correct error term is -> within-cage error

Write the appropriate Proc GLM followed by Proc MIXED SAS code:

PROC GLM

```
Proc glm data=insulin;
  class table diet insulin form;
  model glucose = table|diet|insulin form form*diet form*insulin form*diet*insulin;
  random table table*diet table*insulin table*diet*insulin;
  test h=diet e=table*diet;
  test h=insulin e=table*insulin;
  test h=diet*insulin e=table*diet*insulin;
Run;
Quit;
```

PROC MIXED

```
Proc mixed data=insulin;
  class table diet insulin;
  model glucose = diet insulin diet*insulin form form*diet form*insulin form*diet*insulin;
  random table table*diet table*insulin table*diet*insulin;
Run;
```