- Stat-403/ 650/ 890 (BPK)
Final Exam

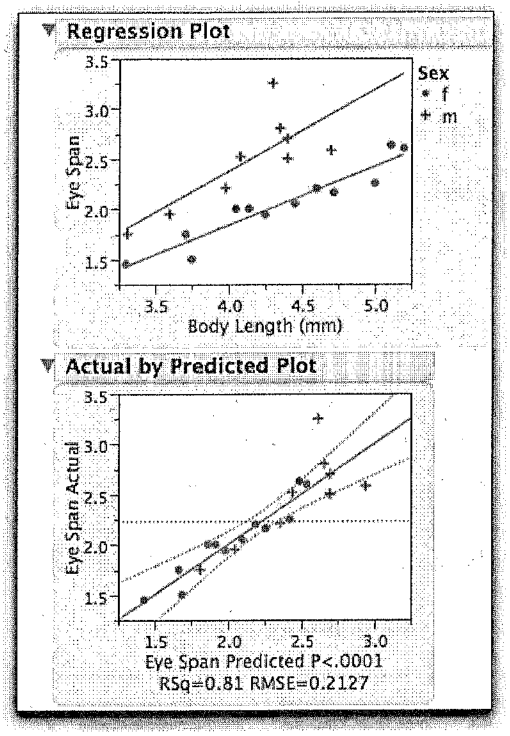
Answer all questions in the exam booklets. All parts have equal worth. Be succinct - if you are writing a thesis, you are doing far too much!

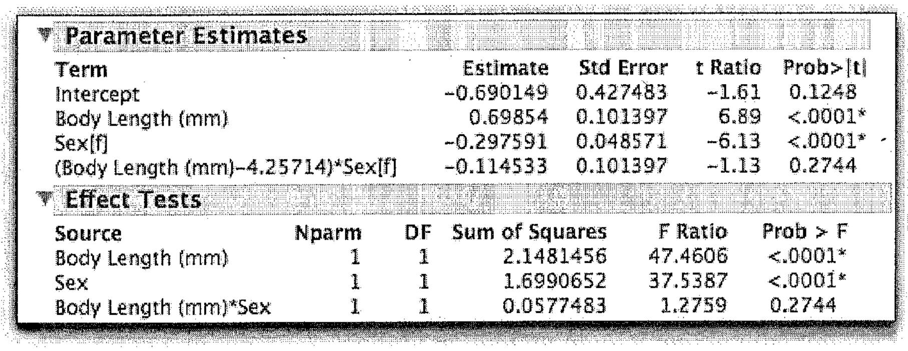
Part I: Eye span of flies caught in amber.

Kotrba M. (2004).[1] investigated the relationship between the eye span (mm) of stalk-eyed flies discovered in baltic amber and covariates such as body length (mm), body width (mm), wing length (mm) and sex.

Use the following output to answer the questions below.

[1] Kotrba M. (2004). Baltic amber fossils reveal early evolution of sexual dimorphism in stalk-eyed flies (*Diptera: Diopsidae*). Organisms Diversity and Evolution 4/4, 265-275. Data downloaded from `http://www.stat.uni-muenchen.de/service/datenarchiv/fliegen/fliegen_e.html` on 2006-07-09.



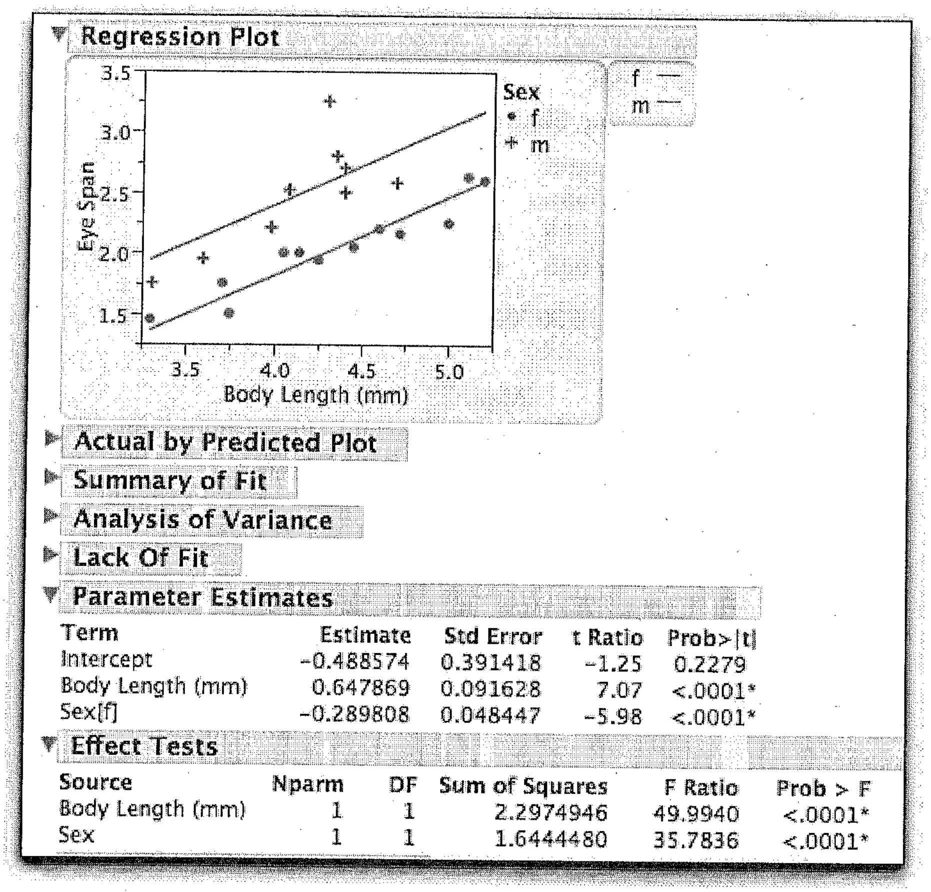


▼ Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.690149	0.427483	-1.61	0.1248
Body Length (mm)	0.69854	0.101397	6.89	<.0001*
Sex[f]	-0.297591	0.048571	-6.13	<.0001*
(Body Length (mm)-4.25714)*Sex[f]	-0.114533	0.101397	-1.13	0.2744

▼ Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Body Length (mm)	1	1	2.1481456	47.4606	<.0001*
Sex	1	1	1.6990652	37.5387	<.0001*
Body Length (mm)*Sex	1	1	0.0577483	1.2759	0.2744



▶ Actual by Predicted Plot

▶ Summary of Fit

▶ Analysis of Variance

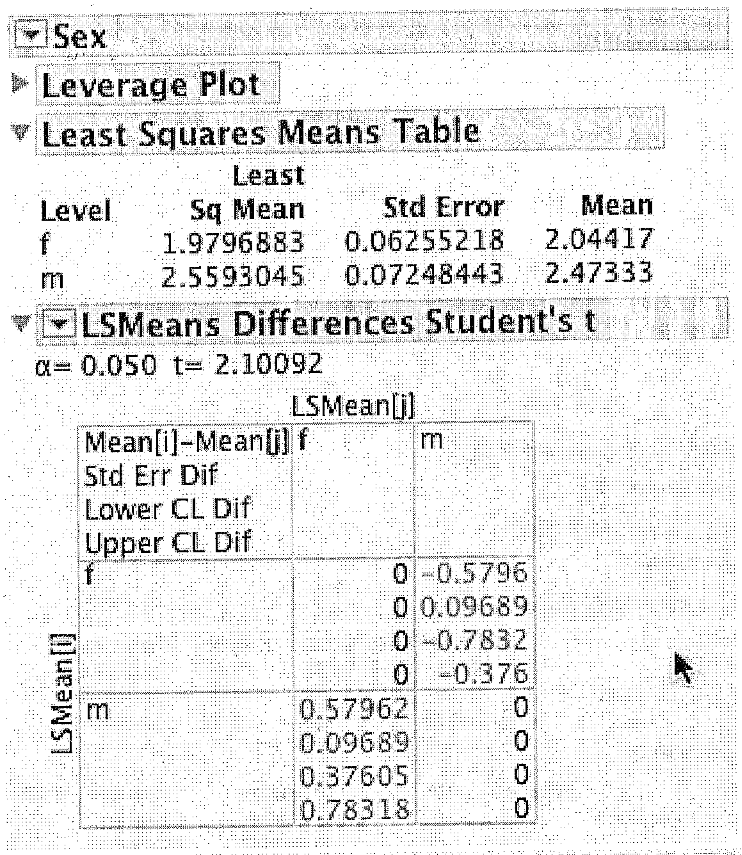
▶ Lack Of Fit

▼ Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.488574	0.391418	-1.25	0.2279
Body Length (mm)	0.647869	0.091628	7.07	<.0001*
Sex[f]	-0.289808	0.048447	-5.98	<.0001*

▼ Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Body Length (mm)	1	1	2.2974946	49.9940	<.0001*
Sex	1	1	1.6444480	35.7836	<.0001*



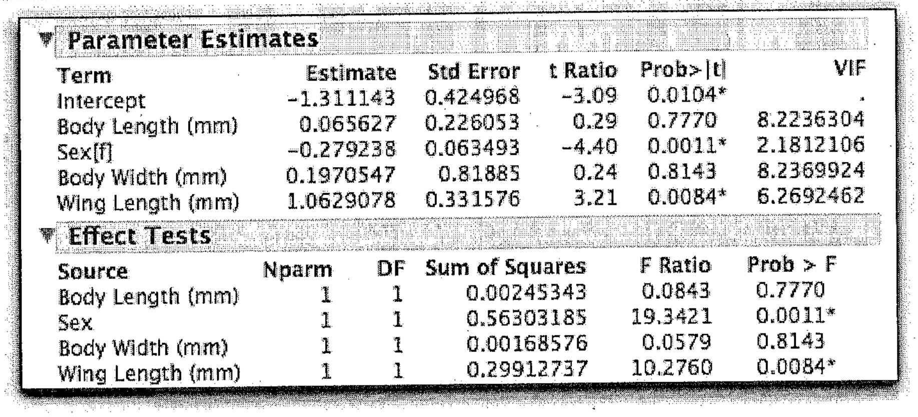
Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
f	1.9796883	0.06255218	2.04417
m	2.5593045	0.07248443	2.47333

LSMeans Differences Student's t

α= 0.050 t= 2.10092

		LSMean[j]	
	Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	f	m
LSMean[i]	f	0 0 0 0	-0.5796 0.09689 -0.7832 -0.376
	m	0.57962 0.09689 0.37605 0.78318	0 0 0 0

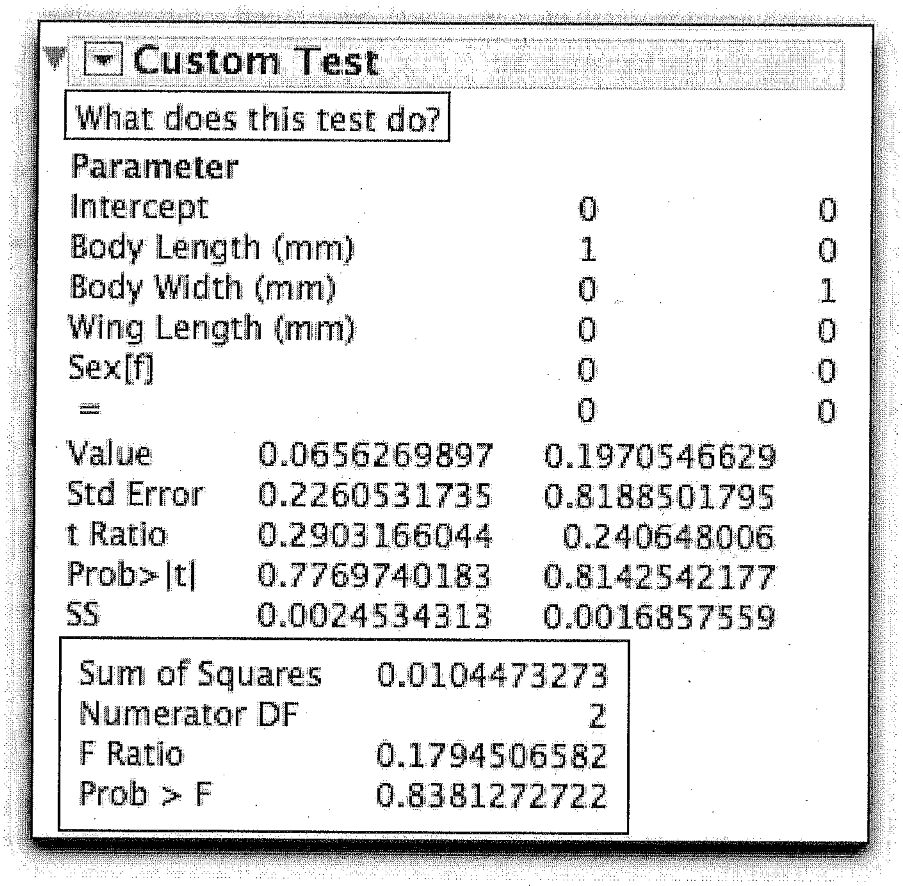


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-1.311143	0.424968	-3.09	0.0104*	.
Body Length (mm)	0.065627	0.226053	0.29	0.7770	8.2236304
Sex[f]	-0.279238	0.063493	-4.40	0.0011*	2.1812106
Body Width (mm)	0.1970547	0.81885	0.24	0.8143	8.2369924
Wing Length (mm)	1.0629078	0.331576	3.21	0.0084*	6.2692462

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Body Length (mm)	1	1	0.00245343	0.0843	0.7770
Sex	1	1	0.56303185	19.3421	0.0011*
Body Width (mm)	1	1	0.00168576	0.0579	0.8143
Wing Length (mm)	1	1	0.29912737	10.2760	0.0084*



```
Custom Test
What does this test do?
Parameter
Intercept                                   0          0
Body Length (mm)                            1          0
Body Width (mm)                             0          1
Wing Length (mm)                            0          0
Sex[f]                                      0          0
=                                           0          0
Value        0.0656269897    0.1970546629
Std Error    0.2260531735    0.8188501795
t Ratio      0.2903166044     0.240648006
Prob>|t|     0.7769740183    0.8142542177
SS           0.0024534313    0.0016857559

 Sum of Squares     0.0104473273
 Numerator DF                  2
 F Ratio            0.1794506582
 Prob > F           0.8381272722
```

1. The researcher started by fitting a regression of eye span (Y) vs wing length. According to the output (not shown) this was based on 19 flies. The researcher also did a regression of eye span (Y) against body width. According to the output (not shown) this was based on 16 flies. Why is there a difference between the number of flies used in the two analyses.

   Solution: The pattern of missing values may be different between the two X variables. For example, the body width must have three additional missing values compared to the wing length variable.

The researcher then fit a model predicting eye span (Y) as a function of body length and sex:

2. Write out the statistical model that was fit.

   Solution: The model is:

$$EyeSpan = Length \quad Sex \quad Sex * Length$$

with normally distributed errors with mean 0 and constant variance σ^2.

This can also be written as:

$$EyeSpan_i = \beta_0 + \beta_1 LENGTH_i + \beta_2 SEX_i + \beta_3 (LENGTH_i \times Sex_i) + \epsilon_i$$

where $LENGTH_i$ is the body length of the i^{th} fly, SEX_i is an INDICATOR variable representing the sex of the fly, and ϵ_i is the residual error that has a normal distribution with mean of 0 and variance σ^2 that is constant over the entire regression surface.

You usually don't include the actual values of the estimates in a model statement.

3. *Sex* is a nominally scaled variable. How do statistical packages deal with categorical variables when used as predictor (X) variables? Give an example of how sex might be coded?

   Solution: If an categorical variable has k classes, then $k-1$ indicator variables are created. Different coding can be used. Sex could be coded at 0=female, 1=male; or 1=female, −1=male; or any two different numerical values can be used.

4. Test the hypothesis of parallel slopes. State the null and alternate hypotheses, the p-value and your conclusion.

   Solution: The hypothesis is that the slope pf the relationship between eye width and length is the same for male and females, i.e. the lines are parallel.

   This can also be written in terms of a hypothesis that $\beta_3 = 0$, i.e. the the slopes of the relationship between eye span and body length is the same for both sexes. The p-value is .27. As this is larger than $\alpha = .05$, there is no evidence that the slopes are not the same for the two sexes.

   Don't use the estimates from the *Parameter Estimates* tables as these depend on the internal parameterization used. Rather, use the F-test from the *Effect Tests* portion of the output.

The researcher then fit a model using parallel slopes for the two sexes:

5. What is the estimate of the common slope, its *se* and an approximate 95% confidence interval.

   Solution: The common slope is estimated to be .65 (SE .09) with approximate 95% confidence interval found as $.65 \pm 2(.09)$ or $(.47 \rightarrow .83)$.

6. Estimate the sex effect along with its *se*. Interpret this estimate.

   Solution: CAUTION: The value reported in the parameter estimates table is only 1/2 of the actual difference between the two lines as can be seen in the graph. So if you use that value, you need to multiply the estimate and se by 2 giving the estimate of the difference is $2(.29) = .58$ with *se* of $2(.048) = .096$. This is a DANGEROUS way to proceed as you need to know what parameterization is being used to deal with categorical variables.

   Or you can use the output from the LSMEANS section of the output which is preferred as this output does NOT depend on the parameterization used by the package. The estimated difference is $0, 58(SE.096)$ mm.

   We estimate that the MEAN eye span is 0.58 (SE .096) mm larger in male than females, AFTER ADJUSTING FOR BODY LENGTH (i.e. keeping body length fixed).

7. What hypothesis is being tested by the first line in the *Effects Test* labeled *Body Length* with an F-ratio of 49.99? What do you conclude?

   Solution: The hypothesis being tested is if the common slope for the effect of body length on the mean eye span is zero. i.e. $H : \beta_{body\ length} = 0$.

   As the p-value is much smaller than $\alpha = .05$, there is very strong evidence that the common slope is not zero, i.e. there is a relationship between Eye Span ad Body length.

The researcher continued with the analysis and fit a multiple-regression model trying to predict eye span as a function of wing length, body width, body length, and sex.

8. The research was surprised that there appeared to be no relationship between eye span and body length in this multiple-regression, but there was a very strong relationship between the two variables when a simple (univariate) regression was performed. Why?

   Solution: This could be caused by collinearity among the X variables. If there is a strong relationship between body-length and another X variable already in the model, such as body width, then the MARGINAL contribution of each variable may be small as the other variable will already have explained much of the variation in the response (Y) variable.

9. What does the column labeled *VIF* in the multiple-regression estimates tell you? Are you concerned?

   Solution: This is the Variance Inflation Factor. This indicates how much of an increase in the variance of the estimate (the se^2) is estimated to have occurred because of collinearity among the predictor variables. VIF values around 10 or higher are usually cause for concern. Here the VIF of 8 may be problematic.

Finally, the research conducted a *Custom Test* (see last output above):

10. What hypothesis is being tested by this *Custom Test*?

    Solution: This tests if the population slope coefficients associated with body length and with body width can be BOTH simultaneously dropped from the model. These tests are constructed by fitting two models. The first model contains both variables (along with the other X variables); the second model drops both variables. If both variables are useless, then SSE from both models should be similar. If both variables are needed, then the SSE between the two models should be large. The F statistic is based on the difference in SSE between the two models.

    This is NOT a test of collinearity! Both variables could be colinear (and so the simple p-values for each variable may indicate that one may be dropped if the other variable is in the model), but the variables might be important and at least one must be included (even if they are colinear). This custom tests examines if BOTH variables can be dropped simultaneously.

11. What do you conclude and why?

    Solution: As the overall p-value is large (.82), there is no evidence that both variables cannot be dropped from the model, i.e. no evidence that both variables must be retained in the model.

Part II Miscellaneous Short Snappers

12. Recall the example of the regression of $log(TEQ)$ against year where TEQ measures the total equivalent dose of dioxin that is slowly degrading over time. The final equation was $\widehat{log(TEQ)} = 2.4 - .08year$. Hence the slope was interpreted as the change on the log-scale of TEQ for each additional year. Here "log" refers to natural logarithms.

    Show how to translate this into an expression of the change on the anti-log scale for each additional year of the study and interpret the back-transformed estimate.

    Solution: The estimated slope means that

    $$log(TEQ)_{t+1} - log(TEQ)_t = -.08$$

    Take anti-logs and
    $$\frac{TEQ_{t+1}}{TEQ_t} = e^{-.08} = .923$$

    So, the concentration each year is 92% of the previous years, or roughly an 8% loss/year.

    This can also be found directly using anti-logs as

    $$Retention = \exp(-.08) = 0.92$$

    Many students took exp(.08) forgetting that the slope is actually $-.08$, i.e. they forgot the negative sign.

13. Explain the difference between the confidence interval for the slope, the confidence interval for the mean response at a new X, and a prediction interval at at new X in the context of fitting a regression line to predicting blood pressure as a function of age. You might find diagrams helpful in your explanation.

    Solution: The confidence interval for the slope shows the range of possible slopes that would be obtained if data were to be recollected with similar X (age) values. It says nothing about the individual points. So this gives a plausible range for the change in mean blood pressure for every increase in age by 1 year.

    The confidence interval for the mean response is the range of plausible values for the mean blood pressure off all future observations with the same X (age) value.

    The prediction interval is the range of plausible values for the individual blood pressure for a single future person at that particular X (age) value.

14. Consider an experiment where a (random) sample of males and females were EACH tested using ALL four different scents (in random order) on their time to complete a maze. Write out a statistical model for this

experiment. What is the "name" of this design and what feature of the above experiment is important to note?

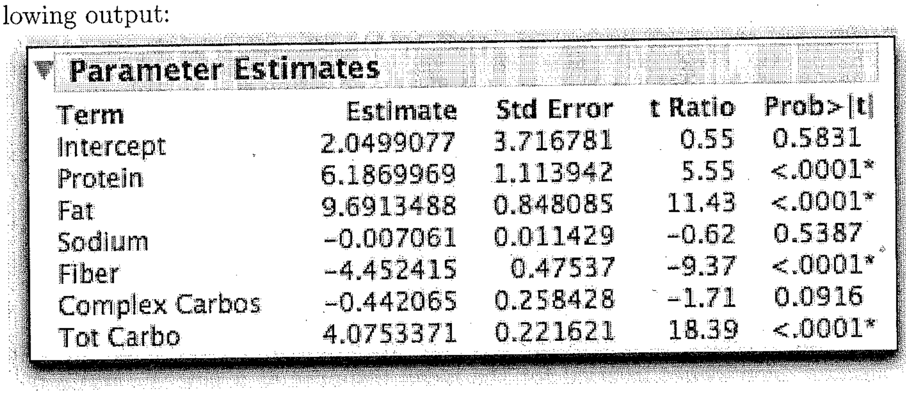
Solution: This is a split-plot design. There are two factor (*sex* and *scent*). The *sex* factor operates at the person level, while the *scent* factor operates within a person. There are two sizes of experimental units which makes it a split-plot design.

The model is

$$Time = Sex \quad Scent \quad Sex * Scent \quad Subject(R)$$

where the *Subject* variable has a unique label for each subject in the experiment. The last term could also be written as *Subject(Sex)* to indicate that every subject is different within each sex.

15. An investigator looked at the relationship between the number of calories in a serving of cereal and various attributes of the cereal and got the following output:



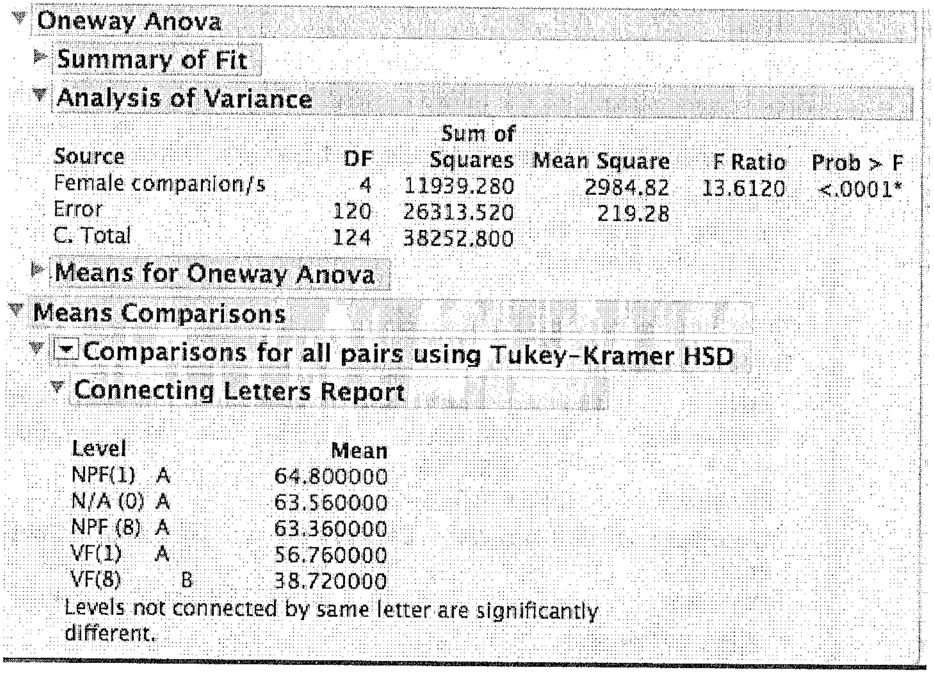
▼ Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.0499077	3.716781	0.55	0.5831
Protein	6.1869969	1.113942	5.55	<.0001*
Fat	9.6913488	0.848085	11.43	<.0001*
Sodium	-0.007061	0.011429	-0.62	0.5387
Fiber	-4.452415	0.47537	-9.37	<.0001*
Complex Carbos	-0.442065	0.258428	-1.71	0.0916
Tot Carbo	4.0753371	0.221621	18.39	<.0001*

Write a 2-3 sentences summarizing how the model was fit, the hypothesis being tested by the line labelled "Protein" and the interpretation of the coefficient.

Solution: A multiple regression was used to fit a model using least squares to predict the number of calories in a serving of cereal as a function of the amount of protein, fat, sodium, fibre, complex carbohydrates, and total carbohydrates. There was strong evidence against the hypothesis that the marginal effect of protein was zero (i.e. strong evidence that the calories/serving increased with the amount of protein, keeping all other variables fixed). The estimated marginal effect of protein was 6.2 (SE 1.1). calories per serving per increase of one gram of protein, keeping all other variables fixed.

16. A researcher obtained the following output when studying the relationship between the survival time of fruit flies and the number and type of sexual partners (none, 1 virginal, 8 virginal, 1 experienced, 8 experienced). Each individual male fly was measured in individual flasks with the suitable female type and number.



```
▼ Oneway Anova
  ▶ Summary of Fit
  ▼ Analysis of Variance
                                       Sum of
    Source                   DF       Squares   Mean Square   F Ratio   Prob > F
    Female companion/s        4     11939.280       2984.82   13.6120    <.0001*
    Error                   120     26313.520        219.28
    C. Total                124     38252.800
  ▶ Means for Oneway Anova
▼ Means Comparisons
  ▼ [▾] Comparisons for all pairs using Tukey-Kramer HSD
    ▼ Connecting Letters Report

      Level                Mean
      NPF(1)   A      64.800000
      N/A (0)  A      63.560000
      NPF (8)  A      63.360000
      VF(1)    A      56.760000
      VF(8)       B   38.720000
     Levels not connected by same letter are significantly
     different.
```

Write a 2-3 sentences summarizing the results suitable for a scientific journal. The codes for the levels are

- NPF(1) - one non-virginal female
- NPF(8) - eight non-virginal females
- VF(1) - one virginal female
- VF(8) - eight virginal females
- N/A (0) - no female of either status

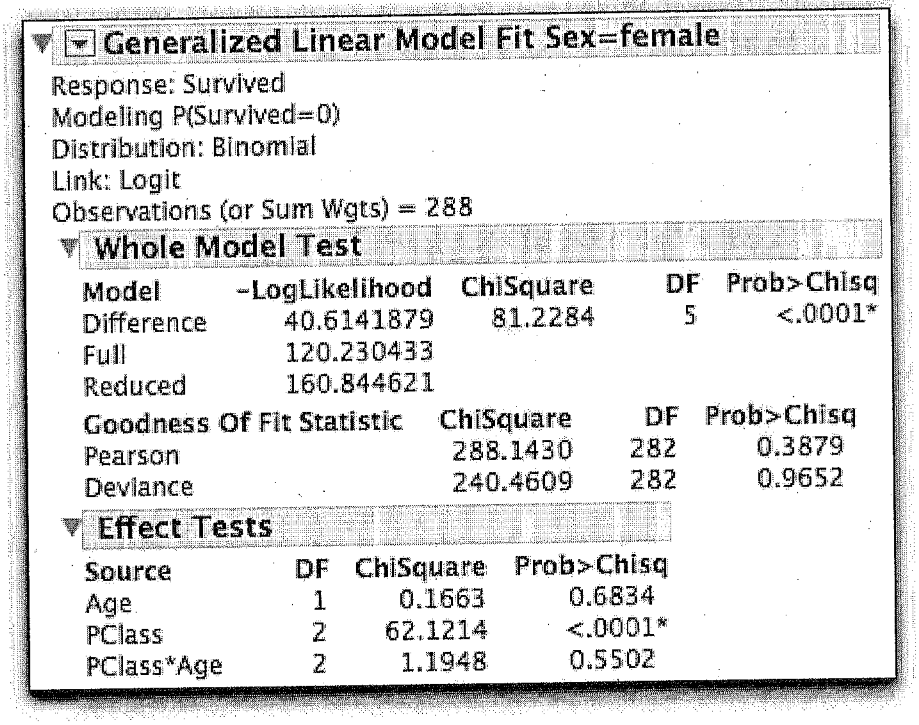
Solution: A single factor CRD ANOVA showed strong evidence that the mean lifetime of male flies differed among the five treatment groups (F=13.6, p<.0001). A Tukey multiple-comparison procedure indicated that the mean lifetime for males with 8 virginal females was less than the other four groups, but we were unable to distinguish among the mean male lifetimes of the other four treatment groups.

DO ONE OF PART III or PART IV

Part III: Surviving the Titanic

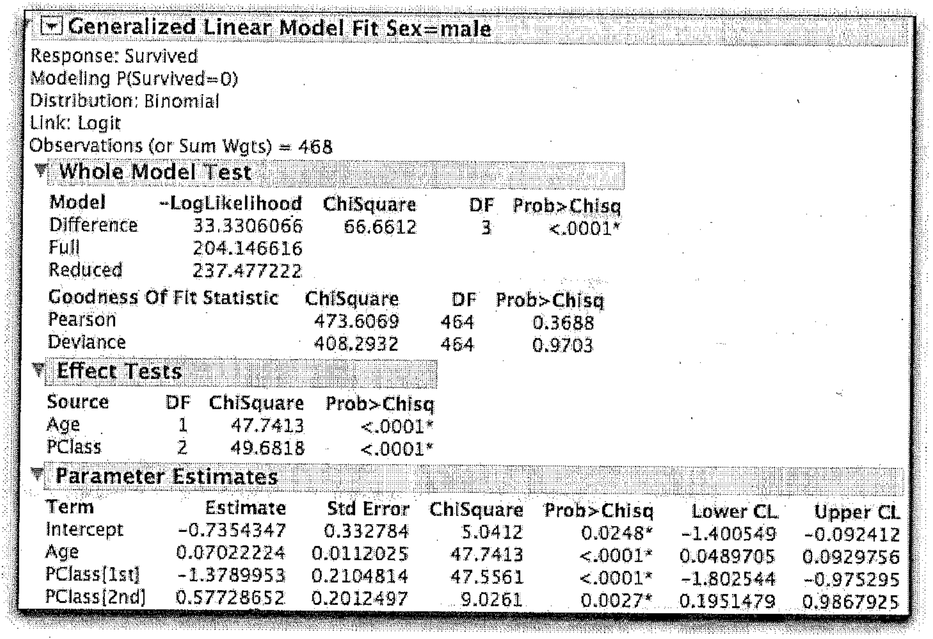
Recall the assignment where you investigated survival rates of passengers aboard the Titanic.

Here is some output for the questions that follow: **Output A follows**



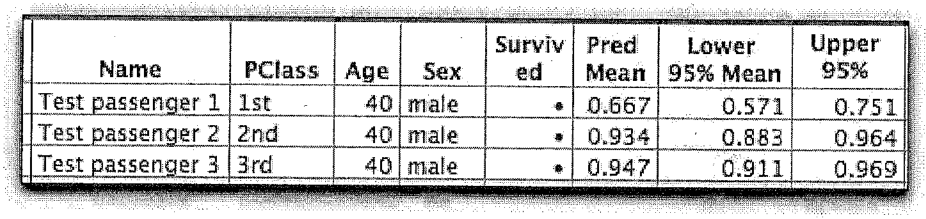
```
Generalized Linear Model Fit Sex=female
Response: Survived
Modeling P(Survived=0)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 288
 Whole Model Test
  Model       -LogLikelihood  ChiSquare   DF  Prob>Chisq
  Difference      40.6141879    81.2284    5     <.0001*
  Full           120.230433
  Reduced        160.844621
  Goodness Of Fit Statistic  ChiSquare    DF  Prob>Chisq
  Pearson                     288.1430   282      0.3879
  Deviance                    240.4609   282      0.9652
 Effect Tests
  Source      DF  ChiSquare  Prob>Chisq
  Age          1     0.1663      0.6834
  PClass       2    62.1214     <.0001*
  PClass*Age   2     1.1948      0.5502
```

Output B follows



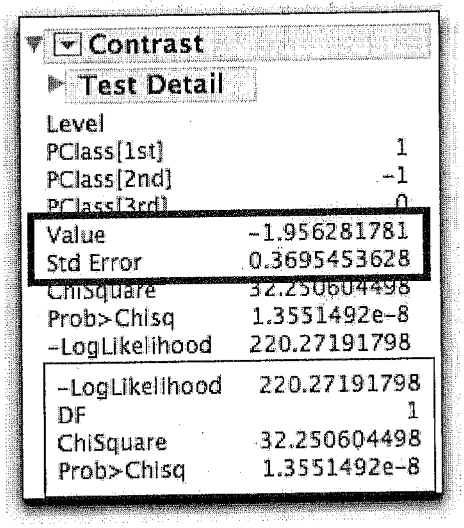
```
Generalized Linear Model Fit Sex=male
Response: Survived
Modeling P(Survived=0)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 468
Whole Model Test
  Model        -LogLikelihood  ChiSquare   DF  Prob>ChiSq
  Difference       33.3306066    66.6612    3    <.0001*
  Full            204.146616
  Reduced         237.477222
  Goodness Of Fit Statistic   ChiSquare    DF  Prob>ChiSq
  Pearson                      473.6069   464      0.3688
  Deviance                     408.2932   464      0.9703
Effect Tests
  Source   DF  ChiSquare  Prob>ChiSq
  Age       1    47.7413     <.0001*
  PClass    2    49.6818     <.0001*
Parameter Estimates
  Term            Estimate   Std Error  ChiSquare  Prob>ChiSq    Lower CL     Upper CL
  Intercept     -0.7354347    0.332784     5.0412     0.0248*   -1.400549   -0.092412
  Age           0.07022224   0.0112025    47.7413     <.0001*   0.0489705   0.0929756
  PClass[1st]   -1.3789953   0.2104814    47.5561     <.0001*   -1.802544   -0.975295
  PClass[2nd]   0.57728652   0.2012497     9.0261     0.0027*   0.1951479   0.9867925
```

Output C follows

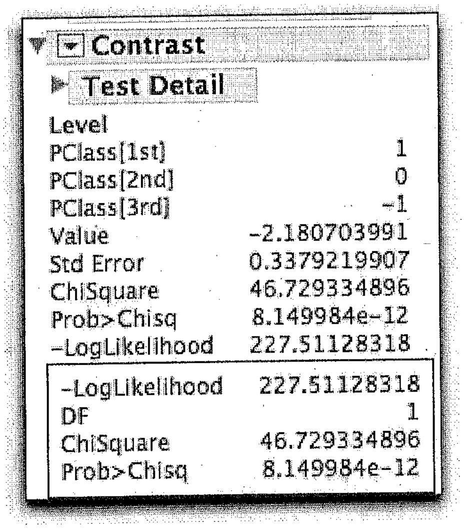


Name	PClass	Age	Sex	Survived	Pred Mean	Lower 95% Mean	Upper 95%
Test passenger 1	1st	40	male	•	0.667	0.571	0.751
Test passenger 2	2nd	40	male	•	0.934	0.883	0.964
Test passenger 3	3rd	40	male	•	0.947	0.911	0.969

Output D follows

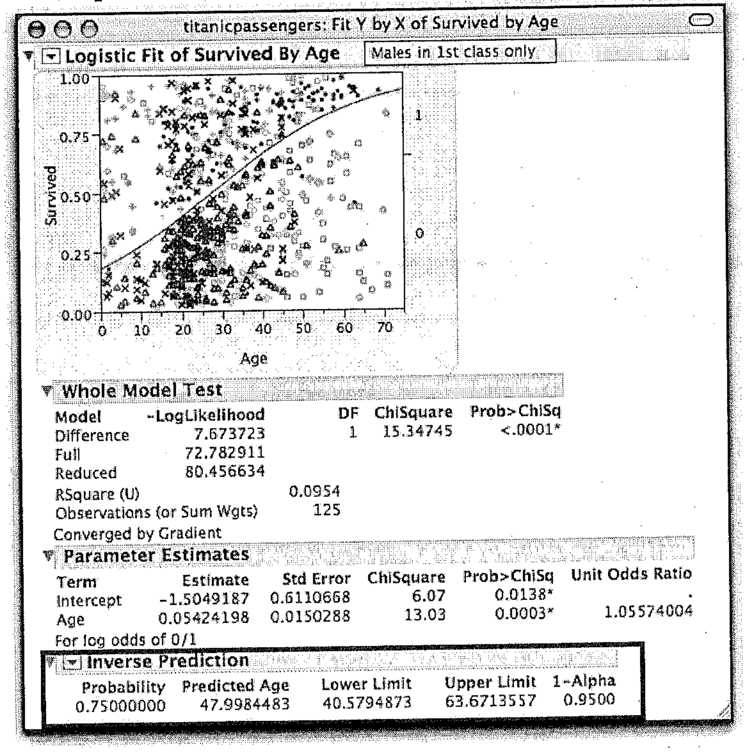


```
Contrast
  Test Detail
Level
PClass[1st]               1
PClass[2nd]              -1
PClass[3rd]               0
Value          -1.956281781
Std Error      0.3695453628
ChiSquare      32.250604498
Prob>Chisq     1.3551492e-8
-LogLikelihood 220.27191798

-LogLikelihood  220.27191798
DF                         1
ChiSquare       32.250604498
Prob>Chisq      1.3551492e-8
```



```
Contrast
  Test Detail
Level
PClass[1st]               1
PClass[2nd]               0
PClass[3rd]              -1
Value          -2.180703991
Std Error      0.3379219907
ChiSquare      46.729334896
Prob>Chisq     8.149984e-12
-LogLikelihood 227.51128318

-LogLikelihood  227.51128318
DF                         1
ChiSquare       46.729334896
Prob>Chisq      8.149984e-12
```

Output E follows



```
titanicpassengers: Fit Y by X of Survived by Age
Logistic Fit of Survived By Age   Males in 1st class only

Whole Model Test
Model        -LogLikelihood      DF   ChiSquare   Prob>ChiSq
Difference         7.673723       1    15.34745     <.0001*
Full              72.782911
Reduced           80.456634
RSquare (U)                   0.0954
Observations (or Sum Wgts)       125
Converged by Gradient

Parameter Estimates
Term         Estimate    Std Error   ChiSquare   Prob>ChiSq   Unit Odds Ratio
Intercept   -1.5049187   0.6110668        6.07      0.0138*                 .
Age          0.05424198  0.0150288       13.03      0.0003*        1.05574004
For log odds of 0/1

Inverse Prediction
  Probability   Predicted Age   Lower Limit   Upper Limit   1-Alpha
  0.75000000     47.9984483     40.5794873    63.6713557    0.9500
```

Here is a simple table of the count of survival status by sex averaged over all classes of passenger:

	Survived	
Sex	No	Yes
Female	154	308
Male	709	142

17. Compute the odds and logit of survival vs. death for females.

    Solution The odds of survival for females is $\frac{308}{154} = 2:1$. The $logit_F(survival) = log(2) = .69$.

18. Compute the odds ratio of survival vs. death for females vs. males. How do odds ratios differ from raw probabilities.

    SolutionThe odds ratio is $odds(survival)_{F\ vs\ M} = \frac{\frac{308}{154}}{\frac{142}{709}} = 10.0$ The odds-ratio differ from raw probabilities in that the base rate is unknown. All that is known is that the odds of survival for females are 10× that of males, but the base rate for males is unknown.

19. Explain the difference between a prospective and retrospective study.

    Solution In prospective studies, patients are followed over time to see the final outcome. In retrospective studies, subjects with observed outcomes are selected, and their initial status is ascertained after the fact.

The above analysis is a gross picture of what happens as it averages over the various classes of passengers and ages of passengers. Subsequent analyses investigated the survival rates as a function of passenger class, age, and sex.

Output for this and subsequent analyses appeared previously. The variable *Survival* is coded as 0=death and 1=survived.

20. What is the model that was used to give the results in Output A? It is not necessary to insert estimates into the model (yet).

    Solution The model is:

    $$Y_i \text{ distributed } Binomial(p_i)$$

    $$\phi_i = logit(p_i)$$

    $$\phi_i = \beta_0 + \beta_1 AGE + \beta_2 I1_{class} + \beta_3 I2_{class} + \beta_4 AGE * I1_{class} + \beta_5 AGE * I2_{class}$$

    where $I1$ and $I2$ are two indicator variables created to represent the three passenger classes, and p_i is the probability of DEATH (see below) for the i^{th} passenger.

    The last line can also be written in the short hand syntax of:

    $$DEATH = AGE \;\; CLASS \;\; AGE * CLASS$$

    Notice that the JMP output shows that the probability of $Survived = 0$, i.e. death is being modelled.

21. What hypothesis (in words) is being tested by the last effect test in Output A that has a p-value of .5502? Why is this hypothesis of interest?

    Solution This tests if the relationship (on the logit scale) between age and DEATH is parallel for the three classes of passengers. If the relationship is parallel, then comparison among DEATH rates for the three passenger classes are the same for all ages.

A subsequent model (for a different sex) was fit whose results are shown in Outputs B, C and D.

22. Explain what is being estimated by the value of -1.95 that is highlighted in Output D? Estimate the odds-ratio along with its *se*.

    Solution This estimates the difference in log-odds of DEATH between 1st and 2nd class male passengers. The difference in log-odds of SURVIVAL will be the negative of this or 1.95.

    Because differences on the log-scale correspond to log(ratios), the odds ratio is estimated as $OR_{1st\ vs\ 2nd}(survival) = e^{1.95} = 7.02$.

    The *se* can be found in two ways. First, find the 95% confidence interval for the odds-ratio by taking the anti-log of the confidence interval for the log-odds ($e^{1.95-2(.37)} = 3.35 \rightarrow e^{1.95+2(.37)} = 14.73$). Then the *se* is approximately the width of the c.i./4 or $se \approx \frac{14.73-3.35}{4} = 2.84$.

    Or, the delta method can be used. This gives $se \approx .37(.702) = 2.59$.

23. Show how this relates to the predicted values as shown in Output C?

    Solution The proportion of survival for a male in 1st class is $1 - .667 = .333$. The probability of survival for a male in 2nd class is $1 - .934 = .066$. [Recall that JMP modeled the probability of death so gives estimates of the probability of death.] Hence the odds-ratio is $OR = \frac{\frac{.333}{.667}}{\frac{.066}{.934}} = 7.06$ which matches the estimate above except for rounding error..

Finally, an analysis was performed for only the males in 1st class. Results are found in Output E.

24. Write out the equation of the fitted model in terms of **Survival rates.**

    Solution JMP models the probability of death. Because survival is the complement of death, the odds of survival are the inverse of the odds of death. This implies that the logit of survival is the negative of the logit of death. The fitted line is therefore:

$$logit(\widehat{survival}) = 1.505 - .0543(age)$$

25. Estimate the odds-ratio of **surviving** (along with a *se*) between males who were 20 and 25 years old.

    Solution The regression equation shows the change in log-odds ratio per year of change. It is for the log-odds of death, so the log-odds of survival is simply the negative of this. For for a five year difference, the change in log-odds is $5 \times .0542 = .271$ with a *se* of $5 \times .015 = .075$. This can be inverted to give the odds-ratio $OR = e^{.271} = 1.31$.

The *se* can be found by inverting the confidence interval on the log-odds scale ($e^{.271-2(.075)} = 1.12 \rightarrow e^{.271+2(.075)} = 1.52$). The *se* is estimated as the width/4 or $\frac{1.52-1.12}{4} = .1$. Alternatively, the delta method can be used. The *se* is then $se \approx .075(1.31) = .098$.

DO ONE OF PART III or PART IV

Part IV: Estimating breeding pairs

A survey was conducted to estimate the number of a breeding pairs of a ducks. The study area was first subdivided into two parts - high quality habitat and low quality habitat. For example, in the prairies, high quality habitat would be the "pot-hole" area where there are many small ponds ringed with good cover while poor quality habitat would be agricultural fields. High quality habitat has a very high density of suitable nesting area; low quality habitat has very few suitable nesting areas. The high quality habitat was divided into 4 hectare (10,000 square meters) quadrats and a simple random sample of quadrats was selected. Crews then visited the selected quadrats and intensively looked for breeding pairs. In the low quality habitat, crews walked along a transects of varying lengths and looked for breeding pairs in 100 m on either side of the transect. Here is the summary information below:

Habitat	Total		
Type	Area (ha)	Statistics	se(statistic)
High	4,000	Sample mean = 10 pairs/quadrat	se(sample mean) = 2 pairs/quadrat
Low	10,000	Sample ratio = 0.05 pairs/ha	se(sample ratio) = 0.01 pairs/ha

26. Estimate the total number of pairs in each habitat type along with a standard error.

```
High  Est Total =  1000 quadrats  x 10 pairs/quadrat = 10,000 pairs.
      se(Total) =  1000 x 2 = 2000 pairs.

Low   Est Total =  10000 ha        x .05 pairs/ha     =     500 pairs.
      se(Total) =  10,000 x.01 = 100 pairs.
```

Because each quadrat is 4 ha, there were 4000/4 = 1000 quadrats.

Note that both the estimates and their se must have units.

27. Estimate the total number of pairs overall along with a standard error.

Solution: Estimated total = 10,000 + 500 = 10,500. *se* (estimated total)= $\sqrt{2000^2 + 100^2}$ = 2002.5 pairs.

28. Suppose that because of bad weather, the actual number of quadrats visited in each stratum is smaller than allocated. Will this cause a problem? Explain?

Solution: As long as the missingness is MCAR this does not cause any problems. For example, suppose that only 50% of quadrats could actually be sampled. Then, if the 50% are randomly chosen from the planned quadrats, this is exactly equivalent to having a smaller (randomly chosen) sample size as the original plan.

29. Suppose you had 100 person-days of crew time available. Discuss how you would apportion your crew time between the two habitat areas. Explain the rationale for your allocation to your significant-other who has not taken a course in statistics.

    Solution: A critical assumption is that the cost to search either habitat is about the same.

    Equal allocation would not be appropriate because a precise estimate for the low density areas is not likely needed.

    Proportional allocation to stratum size would place a lot of effort in the *Low* stratum, but the variability is very low in that stratum.

    Proportional allocation to the density would place most effort in the *High* stratum.

    Neyman allocation (not covered this year) would try and take account both stratum size and stratum variability. Here area can be used as a measure of each stratum. We don't know the actual variabilities in each stratum, but do know that the stratum standard deviations are in a 50:1 ratio. We can use any values that are in this ratio as the proportionality constant will cancel out. Neyman allocation finds the best allocation to get the best precision for a fixed sample size, and requires allocation proportional to the product of stratum size and stratum standard deviation. We get:

Stratum	Size	Std dev	Size x Std Dev
High	4000	50	200,000
Low	10000	1	10,000
Total			210,000

    The effort that is allocated to the *High* density stratum is then $100 \times 200,000/210,000 = 95$ days.

    This is sensible as effort is allocated to strata that are very VARIABLE – why put effort into the low density stratum, even though it is very large, if the densities are very consistent.

    Note: the fact that the *High* density area has a high density of ducks is irrelevant – what is important is the variation in the reading, i.e. the same allocation would have taken place if the High/Low labels were reversed.

30. Why did the survey designer divide the habitat into high and low quality habitat before conducting the survey?

    Solution: The designer stratified to group the habitat into more homogeneous sets. The homogeneity is based on the variation of the readings and not necessarily the density. By stratifying, some of the variation in the readings will be accounted for by the strata leaving less "unexplained" variation (noise) in the data.

    As well, the stratification allowed two different method of surveying to be used as the conditions in the high density areas (lots of small ponds) are quite different than conditions in the low density areas (agricultural land).

    The designer may also have wanted estimates for both habitats – perhaps a policy issue is involved on the preservation of different habitats.

    Lastly, the designer may have some other administrative reasons for the split.

31. Some of the transects in the low density area had NO breeding pairs. Explain why it is necessary to include these transects with 0 counts in the survey above. What would the effect of simply ignoring the transects with 0 counts.

    Solution: If zeroes are excluded the estimated density will be biased high. In addition, the sample standard deviation will be reduced which will cause the estimated standard error to be misleading too precise, i.e. not fully reflecting the variability in the estimate.

    The apparent sample size is also smaller than the actual sample size which can also affect precision.

32. A research associate noted that it was awkward to travel to the individual 4 ha selected quadrats and suggested that the survey be modified to go to a select point and then measure 4 quadrats at that selected point. By reducing the number of randomly selected points by a factor of 4 and measuring 4 quadrats at each point, the survey effort would be the same, but travel time would be reduced – saving money. Comment on this proposal.

    Solution: This would be a form of cluster sampling. A naive analysis of a cluster sample would lead to standard errors that are under reported, i.e. the results look more certain then they actually are. The key problem with cluster designs is that the observations within a cluster are no longer independent of each other.

33. It turns out that these species of ducks are high gregarious and tend to aggregate together, i.e. the distribution of ducks within each stratum is very patchy. Would this indicate the use of fewer larger or many smaller quadrats? Hint: think of the impact of the patchiness on the variability within a stratum.

Solution: HIghly aggregated data would favour fewer larger quadrats. The reason is that the larger quadrats would be more likely to actually include an aggregation of the birds and so you would have less variation in the counts across the quadrats. The small quadrats would tend to be either 0 or have large counts leading to higher variation.

As a matter of interest, if animals are randomly scattered on the landscape independently of each other (called a Poisson process), then the quadrat size is irrelevant and 10 quadrats of 10 square meters is equally precise as 25 quadrats of 4 square meters. The total area sampled drives the precision.

34. Explain the difference between bias and precision. Using suitable diagrams, explain under what conditions you might prefer a very precise but biased estimate over an imprecise but unbiased estimate.

Solution: Bias is the difference between the average value of the estimator over repeated sampling from the same population and the true parameter value.

Precision is the variation of the estimator around its average value over repeated sampling from the same population.

If the bias were very small relative to the precision, or of known size, a biased but precise estimator may be preferred over an unbiased but imprecise estimator:

```
              X = true population value
Unbiased but not precise             Biased but Precise
              X                                X
  .  . ...  .  ... .. .               ......
```

[Technically, an overall measure of accuracy of an estimate is the Mean Square Error (MSE) = $Bias^2 + se^2$. A low bias and low *se* could give rise to a lower MSE than 0 bias and a large *se*.