The data for all parts of this question are based on studies of a treatment that is intended to maintain mangroves along the Florida coast while sea level is rising. Details of the treatment are not important. The study design is important. Twenty large plots, each 10m x 10m (i.e., each is 100 m^2), were randomly located across the range of water depths where mangroves currently grow. The actual water depth in each plot was measured after the plots were established and are assumed constant during the study. Ten plots were randomly assigned to the experimental treatment; the other 10 were left alone (control plots). The response variable, Y, is the annual change in mangrove height, i.e. mangrove growth, two years after applying the treatment.

Part I

One approach to analyzing these data is to assign each plot to one of 3 depth categories: Shallow, Medium, or Deep water. The number of plots in each combination of depth category and experimental treatment is given in Table 1. As you see in Table 1, all the Deep water plots were randomly assigned to the control treatment; none received the experimental treatment.

Depth	Treatment	
Category	Control	Experimental
Shallow	2	2
Medium	5	8
Deep	3	0

Table 1: Number of plots in each combination of Depth Category and Treatment

The questions in part I consider only the Shallow and Medium depth plots. After removing the Deep water plots, there are data from 17 plots.

One model for these data is

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2), \end{aligned} \tag{1}$$

where $i = 1, 2$ indexes depth categories, $j = 1, 2$ indexes treatments, $k = 1, 2, \ldots n_{ij}$, indexes plots within each combination of depth category and treatment, and n_{ij} is the number of plots for each combination of depth category and treatment.

Table 2 gives descriptive statistics for the four combinations of depth category (Shallow or Medium) and treatment (Control or Experimental).

Table 3 gives the Error SS for model (1) and three submodels.

1. Estimate σ^2 in model (1). Show your work.
2. Calculate the F statistic for the Type III (partial) test of the null hypothesis of no effect of the treatment. Show your work.

			Growth:		
Depth Category	Treatment	n_{ij}	average	sd	se
Shallow	control	2	3.42	0.039	0.028
Shallow	experimental	2	3.49	0.13	0.094
Medium	control	5	3.15	0.11	0.048
Medium	experimental	8	3.26	0.10	0.035

Table 2: Summary statistics for each combination of two depth categories (Shallow, Medium) and treatment (Control, Experimental). Note: sd and se are based on the cell-specific variability.

Model terms	Model equation	Error SS
DepthCat, Treatment	$\mathrm{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j$	0.1351
DepthCat	$\mathrm{E}(Y_{ijk}) = \mu + \alpha_i$	0.1783
Treatment	$\mathrm{E}(Y_{ijk}) = \mu + \beta_j$	0.3186
Intercept only	$\mathrm{E}(Y_{ijk}) = \mu$	0.3462

Table 3: Error Sum-of-squares for four models fit to data from Shallow and Medium depths.

A second model, (2), adds the depth category by treatment interaction to model (1):

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\sim N(0, \sigma^2), \end{aligned} \tag{2}$$

where the notation is the same as for model (1). Table 4 is the Type III ANOVA table reported by SAS proc glm **for the data from Shallow and Medium depth plots only**.

Source	df	SS	F	p
DepthCat	1	0.184	17.91	0.0010
Treatment	1	0.026	2.54	0.13
DepthCat*Treatment	1	0.001	0.11	0.75
Error	13	0.134		

Table 4: SAS Type III ANOVA table for model (2) fit to data from the Shallow and Medium depth plots.

3. Carefully interpret the result of the interaction test (DepthCat*Treatment) in Table 4. In other words, write an appropriate one sentence conclusion for that test.
Note: "The interaction is not significant, $p > 0.05$" is not a careful interpretation.

4. Given model (2) and any information from Tables 1-4, you want to calculate the least squares means (lsmeans) estimate of the mean growth in the experimental plots. We will name this value $\hat{\mu}_{expt}$. Compute $\hat{\mu}_{expt}$. Show your work.

5. Calculate the standard error of $\hat{\mu}_{expt}$ using results from fitting model (2). Show your work.

Part II

We now consider the analysis of the data including the Deep plots, i.e., all 20 plots. The number of plots in each combination of DepthCat and Treatment is given in Table 1. For now, we will continue to use model (2), which includes the interaction. Table 5 has parts of the Type III ANOVA table reported by SAS for the 20 plots:

Source	df
DepthCat	2
Treatment	1
DepthCat*Treatment	1
Error	15

Table 5: SAS Type III ANOVA table when fitting model (2) to all 20 plots.

6. The interaction effect is reported as having 1 degree of freedom (df). You expected 2, the product of the df for DepthCat and Treatment. Explain the discrepancy.

7. Figure 1 (on next page) shows four diagnostic plots using results from model (2).

- List the assumptions made by model (2).
- Then, when possible, use the relevant plot(s) to assess each assumption.
- Make sure to briefly explain your assessments. For example, identify the plot you are looking at and the pattern you see.

Note: If a plot is unnecessary or not relevant, do not discuss it.

Still using all 20 plots, we shift back to using model (1), which **does not** include the interaction, for all 20 plots. Define μ_{ij} as the cell mean for depth category $i = 1, 2, 3$ and treatment $j = 1, 2$. You are interested in a linear contrast, τ, defined as

$$\tau = \frac{\mu_{11} + \mu_{21} + \mu_{31}}{3} - \frac{\mu_{12} + \mu_{22} + \mu_{32}}{3}.$$

8. Describe what τ represents in terms of variables used this study.

9. Under model (1), is τ estimable? Show your work that demonstrates that it is or is not estimable.

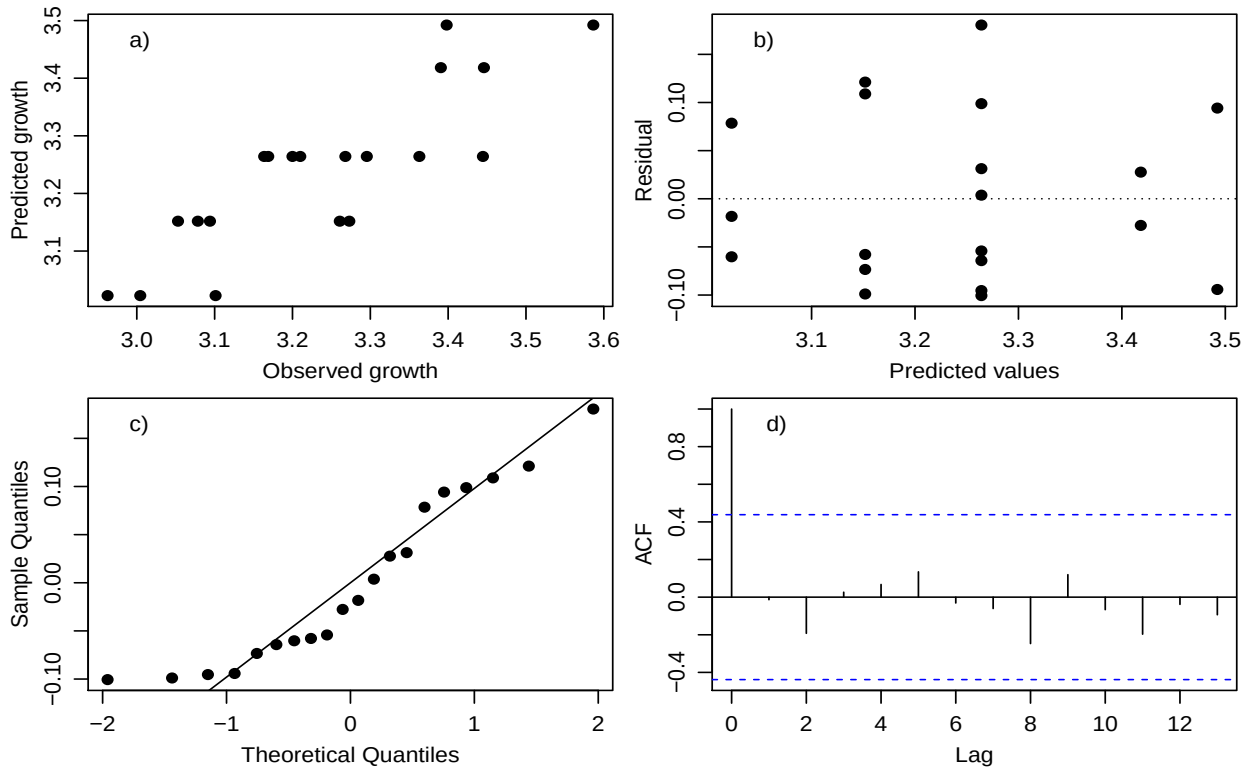

Figure 1: Diagnostic plots for model (2): a) Observed vs predicted values, b) Predicted values vs residuals, c) Normal quantile-quantile plot of residuals, and d) Autocorrelation function (ACF) of the residuals.

Part III

A second approach to analyzing the data from all 20 plots uses the measured water depth for each plot. Figure 2 shows this view of the data.

We consider model (3):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + \beta_2 T_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma_d^2), \end{aligned} \tag{3}$$

where Y_i is the mangrove growth on plot i, D_i is the water depth for plot i, and T_i is an indicator variable representing the treatment assigned to plot i: $T_i = 0$ for control plots and $T_i = 1$ for experimental plots. Table 6 gives some summary statistics from fitting model (3).

10. Carefully describe what β_2 quantifies.

11. For these data, the estimated standard deviation of the errors when fitting model (3), $\hat{\sigma}_d = 0.082$, is smaller than the estimated standard deviation when fitting model (1) to all 20 plots, $\hat{\sigma} = 0.095$. Describe a pattern in the data that would be consistent with the other outcome: $\hat{\sigma}_d > \hat{\sigma}$. You may use words or a graph, for example of growth vs depth, to describe that pattern.

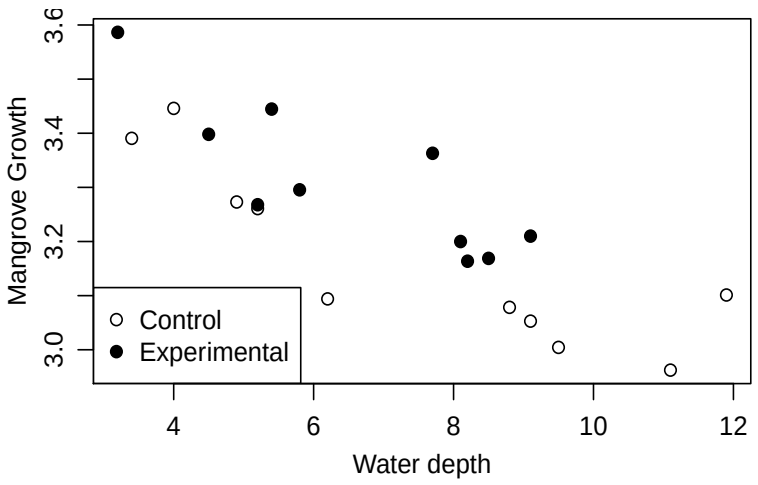

Figure 2: Growth vs water depth for all 20 plots. Symbols indicate the treatment.

Parameter	Estimate	se	p-value
β_0	3.54	0.061	< 0.0001
β_1	-0.050	0.0075	< 0.0001
β_2	0.101	0.037	0.014

Table 6: Parameter estimates and related information from fitting model (3).

12. You want to evaluate whether the relationship between growth and depth is consistent with the linear relationship specified in model (3). Outline two different approaches you can use to assess this.

Part IV:

Seven additional variables were recorded at each of the 20 plots. These variables are labeled A, B, C, E, F, G, and H.

We consider model (4):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1\, D_i + \beta_2\, T_i + \beta_3\, A_i + \beta_4\, B_i + \beta_5\, C_i + \beta_6\, E_i + \beta_7\, F_i + \beta_8\, G_i + \beta_9\, H_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma_f^2), \end{aligned} \tag{4}$$

where Y_i, D_i and T_i are the same variables used in model (3) and A_i, B_i, C_i, E_i, F_i, G_i, and H_i are the values of the seven additional variables for plot i.

13. The estimated standard deviation of the errors when fitting model (4), $\hat{\sigma}_f = 0.095$, is larger than the $\hat{\sigma}_d = 0.082$ estimate from model (3) that only includes D_i and T_i. Explain how adding variables to a model can increase $\hat{\sigma}$.

14. Estimates, standard errors (se), and variance inflation factor (VIF) values for the variables in model (4) are:

Variable	Intercept	D	T	A	B	C	E	F	G	H
Estimate	3.48	-0.48	0.73	0.08	0.02	-0.26	-0.18	0.37	-0.18	0.51
se	1.98	0.12	0.99	0.49	0.32	0.68	0.29	0.72	0.25	0.51
VIF		1.9	5.4	1.3	3.5	5.1	2.7	3.2	1.5	5.0

Do you have any concerns about multicollinearity? Briefly explain your answer.

15. Most of the variables, specifically D, A, B, C, E, F, G, and H, are observed characteristics of a study plot. Some of these variables are potential confounding variables; including them in a model could reduce unwanted variability between plots. In this situation, a common analysis strategy is to use model selection to choose which observational variables to include in a model. Model selection is done only on the potential confounding variables; the randomly assigned variable, T, is omitted from the model selection and D is kept in the model. Table 7 has some model selection statistics for 10 of the possible models. The models in Table 7 are sorted by increasing number of variables.

Which variables should be included in the model? Briefly explain your choice.

Model	Nvar	Rsq	AdjRsq	Cp	AIC	BIC
D F	2	0.761	0.733	0.38	-21.67	-19.67
D F H	3	0.786	0.745	0.93	-21.80	-18.81
D B F	3	0.768	0.725	1.96	-20.26	-17.27
D A F	3	0.764	0.720	2.24	-19.87	-16.88
D C F	3	0.762	0.717	2.35	-19.71	-16.72
D F G	3	0.762	0.717	2.36	-19.70	-16.72
D F G H	4	0.796	0.741	2.31	-20.77	-16.79
D C F H	4	0.795	0.740	2.35	-20.71	-16.72
D A C F G H	6	0.813	0.728	5.23	-18.61	-12.64
D A B C E F G H	8	0.818	0.685	9.00	-15.03	-7.06

Table 7: Model selection information for 10 models fit to the data from all 20 plots. All models include variable T and an intercept. The Model column shows the names of the X variables included in that model. Nvar is the number of variables included in the model, not counting the intercept or T. Rsq and AdjRsq are the R^2 and adjusted R^2 statistics. Cp, AIC and BIC are the Mallows Cp statistic, the Akaike Information Criterion, and Schwarz's Bayesian Information Criterion.

16. After doing the model selection analysis, you are told that most of the plot characteristics (i.e., D, A, B, C, E, G, and H) were measured before the treatment was applied. Variable F was measured when the response was measured (two years after the start of the treatment). Should variable F be omitted from set of variables considered in the model selection analysis? Briefly explain why or why not.

Part I

1. Estimate σ^2 in model (1).

The full model has 3 parameters, so the error df = 17 - 3 = 14.

$$\hat{\sigma}^2 = \frac{0.1351}{14} = 0.00965$$

2. Calculate the F statistic for the type III (partial) test of the null hypothesis of no effect of the treatment.

This is a comparison of DepthCat+Treatment model to the DepthCat model. These models have 17 - 3 = 14 and 17 - 2 = 15 df, respectively.

$$F = \frac{(0.1783 - 0.1351)/(15 - 14)}{0.1351/14} = \frac{0.0432}{0.00965} = 4.4767.$$

3. Carefully interpret the result of the interaction test (GroupCat*Treatment) in Table 5.

No evidence (p = 0.75) of an interaction between two levels of the depth category (shallow and medium) and treatment. OR:
No evidence (p = 0.75) that the difference between treatments is different in the shallow and medium depth categories.

4. Compute $\hat{\mu}_{expt}$.

$$\frac{3.49 + 3.26}{2} = 3.375$$

Note: The least squares mean for a main effect in a model that fits a separate mean for each cell of the design is the equally-weighted average of the relevant cell means.

5. Calculate the standard error of $\hat{\mu}_{expt}$ using results from fitting model (2).

Model (2) assumes equal variances, so need to use the pooled estimate of $\hat{\sigma}$, not the cell-specific estimates. rMSE for model (2) = $\sqrt{0.134/13}$ = 0.1015. The two cell means are based on 2 and 8 observations so the standard errors of the two cell means are $0.1015\sqrt{1/2} = 0.0718$ and $0.0359\sqrt{1/8} = 0.0359$. The coefficients of the linear combination are $1/2$ and $1/2$, so

$$\text{se } \hat{\mu}_{expt} = \sqrt{(0.5^2)(0.0718^2) + (0.5^2)(0.0359^2)} = \sqrt{0.0016} = 0.0401$$

6. Explain why the interaction df = 1, not 2.

There are no observations for the Deep, experimental combination of factors, so the design has one missing cell.
Or, a cell means model has 5 - 1 = 4 df because there are only five observed combinations of factors. Only 1 df remains after fitting main effects of DepthCat (2 df) and Treatment (1 df).

Unequal numbers without mention of the missing cell was not accepted. If the data had unequal numbers but no missing cells, the interaction would have the expected 2 df.

7. List and evaluate assumptions

Assumption	Evaluation
Independent errors	can't tell from these plots, need to look at the design
Constant variance	residuals vs. predicted values plot (b) indicates approx. equal variances The vertical spread of residuals shows no apparent increase or decrease.
Normal distribution	normal quantile plot (c) indicates approximately normal errors The sample quantiles are close to the expected line, except for very smallest residual.

Notes:
Plot a is unnecessary. Plot b shows the same information and is more easily interpreted.
Plot d is irrelevant except when the order of observations is meaningful (e.g., when ordered by time).

This is a small data set so you don't expect textbook-perfect diagnostic plots. Other interpretations are accepted so long as they are appropriately justified.

8. Describe what τ represents in terms of variables used this study.

The population mean difference between control ($j = 1$) and experimental ($j = 2$) plots, averaged over the three depth categories.

9. Under model (1), is τ estimable?

Yes, τ is estimable. In terms of the parameters of model (1), $\mu_{ij} = \mu + \alpha_i + \beta_j$

$$\begin{aligned}
\tau &= \frac{\mu_{11} + \mu_{21} + \mu_{31}}{3} - \frac{\mu_{12} + \mu_{22} + \mu_{32}}{3} \\
&= \frac{\mu_{11} - \mu_{12}}{3} + \frac{\mu_{21} - \mu_{22}}{3} + \frac{\mu_{21} - \mu_{22}}{3} \\
&= \frac{(\beta_1 - \beta_2) + (\beta_1 - \beta_2) + (\beta_1 - \beta_2)}{3} \\
&= \beta_1 - \beta_2
\end{aligned}$$

A linear combination of parameters, $\boldsymbol{C}\,\boldsymbol{\beta}$ is estimable if it can be written as linear combination of expected values of observations, $\boldsymbol{A}\,\mathrm{E}\boldsymbol{Y}$. Under the additive model, $\mathrm{E}\,(Y_{i1k} - Y_{i2k}) = \beta_1 - \beta_2$ for any i and k. So τ is estimable.

10. Describe what β_2 quantifies.

β_2 describes the population average difference between experimental and control (as experimental - control) plots that have the same depth. OR
Geometrically, β_2 is the difference in intercept between parallel treatment and control regression lines for growth versus depth.

11. Describe a data pattern where it is likely that $\hat{\sigma}_d > \hat{\sigma}$.

Various answers are possible. All will have a relationship that is not linear in depth, so a step function, with three means, one for each depth, better describes the relationship.

12. Outline two different approaches you can use to assess linearity of the growth versus depth relationship.

There are various approaches, including:

- Look for a curved pattern in the residual vs predicted value plot.
- Look for a curved pattern in the partial residual plot for depth.
- Add a quadratic term to the model.
- Add a spline or other non-parametric relationship to the model.
- Group depth into categories with similar values and use an ANOVA lack of fit test.

13. Why might adding variables to a model increase $\hat{\sigma}$?

Consider two models; model 2 has more variables. Each has a error SS and error df. $\hat{\sigma}$ will increase when $SS_2/df2 > SS_1/df1$, i.e., $SS_2/SS1 > df_2/df1$.

In words, $\hat{\sigma}$ increases when the change in error SS is not large enough to "compensate" for the loss of error df.

14. Do you have any concerns about multicollinearity?

Need to look at the VIF values. Your answer will depend on your threshold for a "too large VIF". The most common choice is 10, although 5 is sometimes used.

Threshold	Answer
10	No concerns
5	Yes, for T and H

15. Which variables should be included in the model?

Your answer will depend on your choice of model selection criterion.

Criterion	Selected model
Cp or BIC	D F
Adj R^2 or AIC	D F H

The model with D and F has the lowest Cp and BIC statistics. The model with D, F and H has the lowest AIC and highest adjusted R^2 statistics.

The model with all 8 variables has the highest R^2, but R^2 is not an appropriate model selection statistic.

Note: Models with AIC (or BIC) values within 2 units of the best model are often considered reasonable alternative models. Using AIC, all models except for the last two are reasonable alternatives. Using BIC, both the D F and D F H models are reasonable alternatives.

16. Should you omit variable F from the analysis?

Yes, you should omit variable F unless there is no chance that the treatment could influence the values of F. If it does and a model includes both treatment and F, the estimated β for treatment no longer estimates the effect of the treatment. Some of the effect of the treatment is attributed to the F variable.

Another explanation is that in a model with both treatment and F, the β for treatment estimates the effect of the treatment when F is held constant. This is difficult to interpret when the treatment changes F.

Part I

In a study to examine the effect of 4 drugs on 3 experimentally induced diseases in dogs, each drug-disease combination was given to six randomly selected dogs. The response measurement Y of interest is the change in systolic blood pressure (mm Hg) due to treatment.

Some dogs were unable to complete the experiment. For this analysis, you can assume that the dogs with missing observations are "missing at random," meaning that the reasons for dropping out are unrelated to the responses that would have been observed for those dogs. The raw data appear in Table 2 on **Page 6**.

Consider the following model for blood pressure change Y_{ijk} of the k-th dog given the j-th disease receiving the i-th drug as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{1}$$

where μ, α_i, β_j, and γ_{ij} denote fixed effect parameters and all errors ε_{ijk} are iid following a Normal distribution with mean 0 and variance σ^2, that is, $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

Use the output of an analysis done in R on **Pages 6-7** to answer some questions below.

1. Referencing the output on **Page 6**, what parameter restrictions did R impose on Model (1) to solve the normal equations?

2. Related to the parameter restrictions in **Question 1**, complete the following table to express cell means $\mu_{ij} \equiv \mathrm{E}(Y_{ijk})$ for each drug i and disease j combination in terms of the restricted parameters.

	Disease 1	Disease 2	Disease 3
Drug 1	$\mu_{11} = \mu$		
Drug 2			
Drug 3			
Drug 4			

3. Referencing the output on **Pages 6-7**, provide the following two pieces of information for each of the quantities (a)-(d) below: (i) an estimate and, when the latter is non-zero, (ii) an interpretation with respect to the mean blood pressure change $\mu_{ij} \equiv \mathrm{E}(Y_{ijk})$.

   (a) μ

   (b) α_1

   (c) γ_{23}

   (d) $\alpha_2 - \alpha_3 + \frac{1}{3}\big(\gamma_{21} + \gamma_{22} + \gamma_{23} - \gamma_{31} - \gamma_{32} - \gamma_{33}\big)$

4. The parameter restrictions used by R represent one of many possibilities for model fitting; for example, SAS would restrict the parameters in Model (1) differently to obtain a solution to the normal equations. For which of the parameters listed in (a)-(d) of **Question 3** would estimated values remain the same regardless of the parameter restrictions used? Explain.

5. Recalling the data structure (see raw data in **Table 2** on **Page 6**), state any useful inference that can be made from only the first ANOVA table **on Page 6**. Explain.

6. Letting $\bar{Y}_{ij\cdot}$ denote the sample average of responses for drug i and disease j, based on corresponding sample size denoted as n_{ij}, show that

$$F = \frac{\sum_{j=1}^{3} (n_{1j}^{-1} + n_{3j}^{-1})^{-1} (\bar{Y}_{1j\cdot} - \bar{Y}_{3j\cdot})^2}{3\widehat{\sigma}^2}$$

   has a non-central F-distribution, where $\widehat{\sigma}^2$ denotes the sum of squared residuals divided by its degrees of freedom. Report the degrees of freedom for this F-distribution as well as the non-centrality parameter, expressing the latter in terms of cell means $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$.

7. In terms of the cell means μ_{ij}, state the null hypothesis that can be tested based on the statistic from **Question 6** and interpret this null hypothesis in terms of effects on mean blood pressure change.

8. Evaluate the test statistic given in **Question 6**, report a p-value, and state your conclusion in terms of the strength of evidence supported by the p-value.

9. In communicating results, explain why it might be philosophically better to express results of a hypothesis test in terms of the strength of evidence as opposed to a conclusion in terms of rejecting or failing to reject the null hypothesis.

10. Suppose, as a statistical consultant, that you conduct a generic hypothesis test for the purpose of answering a client's research question. For simplicity, suppose that the hypothesis test is indeed appropriate to answer the underlying research question and that any necessary statistical assumptions are fulfilled. Furthermore, suppose that the p-value turns out to be 2.2e-16. What else might you consider to help the client in drawing meaningful conclusions? Briefly explain.

11. Consider the null hypothesis $\mathrm{H}_0 : \boldsymbol{C}\boldsymbol{\mu} = \mathbf{0}$, where

    $\boldsymbol{\mu} = (\mu_{11}, \mu_{21}, \mu_{31}, \mu_{41}, \mu_{12}, \mu_{22}, \mu_{32}, \mu_{42}, \mu_{13}, \mu_{23}, \mu_{33}, \mu_{43})^\top$ denotes the vector of cell means, that is, $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, and $\boldsymbol{C}$ is given as

$$\boldsymbol{C} = \begin{pmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \end{pmatrix}.$$

    Referring to the output on **Pages 6-7** (no computations are necessary), give the value of a test statistic for this hypothesis along with degrees of freedom. Briefly explain your answer.

12. Within the context of mean blood pressure change in dogs, what can you conclude from the result of the test in **Question 11**? Be sure to comment on the strength of evidence.

13. Specify an appropriate matrix $\boldsymbol{C}$ for expressing the null hypothesis of "no main drug effects" in the form $\mathrm{H}_0 : \boldsymbol{C}\boldsymbol{\beta} = \mathbf{0}$, where

    $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \gamma_{11}, \gamma_{21}, \gamma_{31}, \gamma_{41}, \gamma_{12}, \gamma_{22}, \gamma_{32}, \gamma_{42}, \gamma_{13}, \gamma_{23}, \gamma_{33}, \gamma_{43})^\top$.

14. From the output on **Pages 6-7**, how would you briefly summarize the effects of drug type and disease type on mean blood pressure change in dogs (without going into effects at specific level combinations)?

Part II

Suppose that the experiment in **Part I** was actually done at six different veterinary clinics, with one dog (before possible drop out) randomly assigned to each of the 12 drug by disease combinations at each clinic, as indicated in the following **Table 1**.

		Clinic					
Drug	Disease	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
$i=1$	$j=1$	42	44	36	13	19	22
	$j=2$	33	-	33	-	21	26
	$j=3$	31	25	-	-3	24	25
$i=2$	$j=1$	42	-	28	13	23	24
	$j=2$	-	34	36	-	31	33
	$j=3$	26	28	32	3	3	16
$i=3$	$j=1$	-	-	29	1	-	19
	$j=2$	-	11	9	-6	1	7
	$j=3$	21	9	-	1	3	-
$i=4$	$j=1$	24	-	22	-2	9	15
	$j=2$	27	16	15	-5	12	12
	$j=3$	22	25	12	-	5	7

Table 1: Observed change in systolic blood pressure (reported as integer mm Hg) for each dog by drug and disease combination across 6 clinics.

Consider a model that includes an additive block/clinic effect τ_k to account for the differences between clinics as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \tau_k + \varepsilon_{ijk}, \tag{2}$$

where all ε_{ijk} are independent and identically distributed following a Normal distribution with mean 0 and variance σ^2, that is, $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$. Now Y_{ijk} denotes the change in systolic blood pressure (mmHg) for the dog given the j-th disease and receiving the i-th drug at the k-th clinic.

15. Using **Table 1**, explain why block (or clinic) effects cannot be estimated free of the effects for diseases and drugs.

16. By considering the following ANOVA table from SAS based on Type I Sum of Squares,

Source	DF	Type I SS	Mean Square	F-Value	Pr > F
clinic	5	4634.894601	926.978920	41.99	<.0001
disease	2	663.918326	331.959163	15.04	<.0001
drug	3	2445.204851	815.068284	36.92	<.0001
drug*disease	6	517.280527	86.213421	3.91	<.0001
Error	41	905.115489	22.075988		
Corrected Total	57	9166.413793			

along with the first ANOVA table from R on **Page 6** (where clinic is not included), explain if and how variation among clinics can impact our understanding of the effects of diseases and drugs.

17. Given the data structure and the statement of Model (2), explain whether it is possible to provide an unbiased estimate of each interaction effect of the form

$$(\gamma_{ij} - \gamma_{i'j}) - (\gamma_{ij'} - \gamma_{i'j'}),$$

(i.e., free of blocks) for drugs $i \neq i' \in \{1, 2, 3, 4\}$ and diseases $j \neq j' \in \{1, 2, 3\}$.

18. Referring to the ANOVA table in **Question 16**, and in particular the interaction term, how would you proceed with inference to understand the effects of diseases and drugs (i.e., what would you consider as meaningful next steps)? Keep your response brief.

Part III

For this part we will assume that the block effects of Model (2) in **Part II** are considered random effects. This yields the following model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \tau_k + \varepsilon_{ijk}, \tag{3}$$

where all parameters and random variables are defined as in **Part II** except that terms $\tau_k \sim \mathcal{N}(0, \sigma_\tau^2)$ now denote iid random block effects instead of fixed block effects. Relevant output from analyzing the data under Model (3) starts on **Page 8**.

19. Obtain the REML estimates of the variance components σ_τ^2 and σ_ε^2 and comment on how these compare.

20. The REML method involves a likelihood function for observations known as error contrasts. How many error contrasts are involved in the REML likelihood function for the fit of Model (3)?

21. Explain why it might be useful to treat clinics as a random effect instead of a fixed effect. Also, explain if there are any disadvantages to Model (3) compared to Model (2).

Raw Data for Part I

	Disease 1	Disease 2	Disease 3
Drug 1	42, 44, 36, 13, 19, 22 $\bar{y}$=29.3333	33, 26, 33, 21 $\bar{y}$=28.25	31, −3, 25, 25, 24 $\bar{y}$=20.4
Drug 2	28, 23, 24, 42, 13 $\bar{y}$=26	34, 33, 31, 36 $\bar{y}$=33.5	3, 26, 28, 32, 3, 16 $\bar{y}$=18
Drug 3	1, 29, 19 $\bar{y}$=16.3333	11, 9, 7, 1, −6 $\bar{y}$=4.4	21, 1, 9, 3 $\bar{y}$=8.5
Drug 4	24, 9, 22, −2, 15 $\bar{y}$=13.6	27, 12, 12, −5, 16, 15 $\bar{y}$=12.83333	22, 7, 25, 5, 12 $\bar{y}$=14.2

Table 2: Observed change in systolic blood pressure (reported as integer mm Hg) for up to 6 dogs and the sample mean per drug and disease combination.

Selected R Output for Part I

```
> dogs <- read.table(file="~/Library/.../dogclinic.dat",
+                    col.names=c("Drug","Disease","Y", "Clinic"))
> dogs$Drug <- as.factor(dogs$Drug)
> dogs$Disease <- as.factor(dogs$Disease)
> dogs$Clinic <- as.factor(dogs$Clinic)
> lm.outP1 <- lm(Y~Drug*Disease,data=dogs)

> anova(lm.outP1)
Analysis of Variance Table

Response: Y
             Df Sum Sq Mean Sq F value    Pr(>F)
Drug          3 2992.8  997.61  9.0513 8.047e-05 ***
Disease       2  365.7  182.86  1.6591    0.2015
Drug:Disease  6  737.9  122.98  1.1158    0.3680
Residuals    46 5070.0  110.22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(lm.outP1)
```

```
lm(formula = Y ~ Drug * Disease, data = dogs)

Residuals:
    Min      1Q  Median      3Q     Max
-23.400  -6.775   0.950   6.650  16.000

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     29.3333     4.2860   6.844 1.56e-08 ***
Drug2           -3.3333     6.3571  -0.524   0.6026
Drug3          -13.0000     7.4235  -1.751   0.0866 .
Drug4          -15.7333     6.3571  -2.475   0.0171 *
Disease2        -1.0833     6.7767  -0.160   0.8737
Disease3        -8.9333     6.3571  -1.405   0.1667
Drug2:Disease2   8.5833     9.7735   0.878   0.3844
Drug3:Disease2 -10.8500    10.2326  -1.060   0.2945
Drug4:Disease2   0.3167     9.2918   0.034   0.9730
Drug2:Disease3   0.9333     8.9903   0.104   0.9178
Drug3:Disease3   1.1000    10.2326   0.107   0.9149
Drug4:Disease3   9.5333     9.1924   1.037   0.3051
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.5 on 46 degrees of freedom
Multiple R-squared:  0.4469,Adjusted R-squared:  0.3146
F-statistic: 3.379 on 11 and 46 DF,  p-value: 0.001754
-------------------------------------------------------------
Analysis of Variance Table
Type III Sum of Squares

Response: Y
             Df Sum Sq    Mean Sq   F value   Pr(>F)
Drug          3 2851.058  950.3527  8.622558  0.0001194 ***
Disease       2  371.711  185.8557  1.686270  0.1964555
Drug:Disease  6  737.888  122.9814  1.115811  0.3680099
Residuals    46 5069.983  110.2170
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Selected SAS Output for Part III

The Mixed Procedure

Model Information	
Data Set	WORK.SET2
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information		
Class	Levels	Values
clinic	6	1 2 3 4 5 6
drug	4	1 2 3 4
disease	3	1 2 3

Dimensions	
Covariance Parameters	2
Columns in X	26
Columns in Z	6
Subjects	1
Max Obs per Subject	58

Number of Observations	
Number of Observations Read	58
Number of Observations Used	58
Number of Observations Not Used	0

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	271.17027937	
1	1	271.17027937	0.00000000

Convergence criteria met but final Hessian is not positive definite.

Covariance Parameter Estimates	
Cov Parm	Estimate
clinic	3638.45
Residual	22.0760

Fit Statistics	
-2 Res Log Likelihood	271.2
AIC (Smaller is Better)	275.2
AICC (Smaller is Better)	275.5
BIC (Smaller is Better)	274.8

Solution for Fixed Effects

Effect	clinic	drug	disease	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept				9.7292	60.3704	0	0.16	.
clinic	1			12.1192	85.3323	0	0.14	.
clinic	2			7.6296	85.3346	0	0.09	.
clinic	3			6.6233	85.3297	0	0.08	.
clinic	4			-14.8946	85.3322	0	-0.17	.
clinic	5			-4.0182	85.3290	0	-0.05	.
clinic	6			0	.	.	.	.
disease			1	3.9048	3.0091	41	1.30	0.2016
disease			2	1.8609	2.8602	41	0.65	0.5189
disease			3	0	.	.	.	.
drug		1		10.5036	3.0048	41	3.50	0.0012
drug		2		7.0276	2.8602	41	2.46	0.0183
drug		3		-1.4382	3.2097	41	-0.45	0.6564
drug		4		0	.	.	.	.
drug*disease		1	1	3.9525	4.1497	41	0.95	0.3464
drug*disease		1	2	2.4752	4.3249	41	0.57	0.5702
drug*disease		1	3	0	.	.	.	.
drug*disease		2	1	5.3724	4.1245	41	1.30	0.2000
drug*disease		2	2	12.3237	4.2146	41	2.92	0.0056
drug*disease		2	3	0	.	.	.	.
drug*disease		3	1	6.8946	4.7502	41	1.45	0.1543
drug*disease		3	2	-4.8199	4.3046	41	-1.12	0.2694
drug*disease		3	3	0	.	.	.	.
drug*disease		4	1	0	.	.	.	.
drug*disease		4	2	0	.	.	.	.
drug*disease		4	3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
clinic	5	0	0.03	.
disease	2	41	12.56	<.0001
drug	3	41	36.32	<.0001
drug*disease	6	41	3.91	0.0036

Part I

1. Essentially, any effects which do not have estimates listed in the R output are set to zero, i.e., $\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0, \quad i = 1, 2, 3, 4; j = 1, 2, 3.$

2. The answers are

	Disease 1	Disease 2	Disease 3
Drug 1	$\mu_{11} = \mu$	$\mu_{12} = \mu + \beta_2$	$\mu_{13} = \mu + \beta_3$
Drug 2	$\mu_{21} = \mu + \alpha_2$	$\mu_{22} = \mu + \alpha_2 + \beta_2 + \gamma_{22}$	$\mu_{23} = \mu + \alpha_2 + \beta_3 + \gamma_{23}$
Drug 3	$\mu_{31} = \mu + \alpha_3$	$\mu_{32} = \mu + \alpha_3 + \beta_2 + \gamma_{32}$	$\mu_{33} = \mu + \alpha_3 + \beta_3 + \gamma_{33}$
Drug 4	$\mu_{41} = \mu + \alpha_4$	$\mu_{42} = \mu + \alpha_4 + \beta_2 + \gamma_{42}$	$\mu_{43} = \mu + \alpha_4 + \beta_3 + \gamma_{43}$

3. (a) $\widehat{\mu} = 29.333$; this represents estimated the mean change with Disease 1 and Drug 1 (i.e., μ_{11})

   (b) $\widehat{\alpha}_1 = 0$ by restriction

   (c) $\widehat{\gamma}_{23} = 0.933$; this represents an estimated interaction contrast. Namely, this contrast is difference of differences $(\mu_{23} - \mu_{21}) - (\mu_{13} - \mu_{11}) = (\mu_{23} - \mu_{13}) - (\mu_{21} - \mu_{11})$ between the mean change in Disease 3 and Disease 1 over Drug 2 vs. the mean change in Disease 3 and Disease 1 over Drug 1 (or Drug 2 vs Drug 1 over Disease 3 against Drug 2 vs Drug 1 over Disease 1).

   (d) $\widehat{\alpha}_2 - \widehat{\alpha}_3 + \frac{1}{3}\big(\widehat{\gamma}_{21} + \widehat{\gamma}_{22} + \widehat{\gamma}_{23} - \widehat{\gamma}_{31} - \widehat{\gamma}_{32} - \widehat{\gamma}_{33}\big) = 9.667 + \frac{1}{3}\big(0 + 8.583 + 0.933 - 0 + 10.85 - 1.1\big) = 16.089$; this is estimated mean change comparing Drug 2 to Drug 3 overall (i.e., $\frac{1}{3}(\mu_{21} + \mu_{22} + \mu_{23}) - \frac{1}{3}(\mu_{31} + \mu_{32} + \mu_{33})$).

4. The only estimable parameter under Model (1) (without restrictions) is (d); this parameter is obtainable from the row space of the design matrix. Its OLS estimator would be $\frac{1}{3}(\bar{Y}_{21\cdot} + \bar{Y}_{22\cdot} + \bar{Y}_{23\cdot}) - \frac{1}{3}(\bar{Y}_{31\cdot} + \bar{Y}_{32\cdot} + \bar{Y}_{33\cdot})$.

5. Due to the missing observations the data are not balanced and the Type I and Type III sums of squares do not agree with the exception of the interaction term between drug and disease. Type I sums of squares are then not appropriate for interpreting the main effects of drug or disease. However, the interaction term can be cleanly interpreted and the p-value associated with the interaction term is 0.3680, suggesting little to no evidence that the effect of drug differs depending on disease.

6. The averages $\bar{Y}_{1j\cdot} \sim \mathcal{N}(\mu_{1j}, \sigma^2/n_{1j})$ and $\bar{Y}_{3j\cdot} \sim \mathcal{N}(\mu_{3j}, \sigma^2/n_{3j})$ are normal and mutually independent for $j = 1, 2, 3$ and independent of $\widehat{\sigma}^2$. The 3×1 vector

$$W \equiv \left(\frac{\bar{Y}_{11\cdot} - \bar{Y}_{31\cdot}}{\sigma(n_{11}^{-1} + n_{31}^{-1})^{1/2}}, \frac{\bar{Y}_{12\cdot} - \bar{Y}_{32\cdot}}{\sigma(n_{12}^{-1} + n_{32}^{-1})^{1/2}}, \frac{\bar{Y}_{13\cdot} - \bar{Y}_{33\cdot}}{\sigma(n_{13}^{-1} + n_{33}^{-1})^{1/2}} \right)^\top$$

is multivariate normal with mean

$$\delta \equiv \left(\frac{\mu_{11} - \mu_{31}}{\sigma(n_{11}^{-1} + n_{31}^{-1})^{1/2}}, \frac{\mu_{12} - \mu_{32}}{\sigma(n_{12}^{-1} + n_{32}^{-1})^{1/2}}, \frac{\mu_{13} - \mu_{33}}{\sigma(n_{13}^{-1} + n_{33}^{-1})^{1/2}} \right)^\top$$

and a 3×3 identity covariance matrix. Hence, $W^\top W$ is chi-square with 3 degrees of freedom and non-centrality parameter $\delta^\top \delta$. We also know $SSR/\sigma^2 = df \times \widehat{\sigma}^2/\sigma^2 \sim \chi^2_{df}$, where df denote degrees of freedom for error (here $df = 46$). The ratio of two independent chi-square variables, divided by their corresponding degrees of freedom, has a F-distribution

$$\frac{W^\top W}{3} \Big/ \frac{\widehat{\sigma}^2}{\sigma^2} = F,$$

which has degrees of freedom $(3, 46)$ and non-centrality parameter $\delta^\top \delta = \sigma^{-2} \sum_{j=1}^{3} (\mu_{1j} - \mu_{3j})^2/(n_{1j}^{-1} + n_{3j}^{-1})$.

7. The non-centrality parameter $\delta^\top \delta = 0$ if and only if $\mu_{1j} = \mu_{3j}$ for $j = 1, 2, 3$. The null hypothesis states that there is no difference between the average blood pressure change for Drug 1 compared to Drug 3 over any of the diseases.

8.

$$\begin{aligned} F &= \frac{\sum_{j=1}^{3} (n_{1j}^{-1} + n_{3j}^{-1})^{-1} (\bar{Y}_{1j\cdot} - \bar{Y}_{3j\cdot})^2}{3\widehat{\sigma}^2} \\ &= \Big[(6^{-1} + 3^{-1})^{-1} (29.33 - 16.33)^2 + (4^{-1} + 5^{-1})^{-1} (28.25 - 4.4)^2 \\ &\qquad + (5^{-1} + 4^{-1})^{-1} (20.4 - 8.5)^2 \Big] / (3 \cdot 110.22) \\ &= 5.797, \end{aligned}$$

where the p-value is 0.0019. There is strong evidence that mean blood pressure change is not the same for Drugs 1 and 3 with respect to all diseases.

9. Two concerns come to mind. (1) By placing emphasis on the decision about H_0, we present the outcome of the test as a dichotomous one, even while the evidence

exists on a continuous scale. Depending on the chosen level of significance, e.g. 0.05, we might statistically distinguish between two p-values such as 0.049 and 0.051, even though these practically stand for the same strength of evidence; and the same time, we would not distinguish, for example, between a p-value < 0.0001 and 0.02, which do not represent the same levels of evidence. (2) Although evidence in favor of H_a is often an important interest of statistical inference, we de-emphasize the uncertainty that is still associated with a binary decision about H_0. While stating a decision based on a level of significance has a nuance understood by statisticians, this is not always understood by practitioners and the uncertainty can easily be forgotten.

10. Obtain a confidence interval for the parameter of interest and/or estimate the effect size if it does not coincide with the parameter of interest. A goal should be to ensure that the statistical evidence is indeed also practically significant.

11. From the output, the test statistic for this is given in the Type III ANOVA table (main effect of disease) as $F = 1.686$ with degrees of freedom 2,46. Note that the null hypothesis corresponds to a test of

$$\frac{1}{4}\sum_{i=1}^{4}\mu_{i1} = \frac{1}{4}\sum_{i=1}^{4}\mu_{i2}, \qquad \frac{1}{4}\sum_{i=1}^{4}\mu_{i1} = \frac{1}{4}\sum_{i=1}^{4}\mu_{i3},$$

or a test about equality among overall mean effects $\sum_{i=1}^{4}\mu_{ij}/4$ of diseases $j = 1, 2, 3$ (averaging across drugs).

12. The p-value associated with the test above is 0.1964. There is not strong evidence to suggest differences among the overall effects of disease type on mean blood pressure change.

13. The null hypothesis corresponds to whether

$$\theta_i \equiv \frac{1}{3}\sum_{j=1}^{3}\mu_{ij} = \mu + \alpha_i + \frac{1}{3}\sum_{j=1}^{3}(\beta_j + \gamma_{ij})$$

is the same for $i = 1, 2, 3, 4$. This is equivalent to testing that three pairwise differences $\theta_1 - \theta_2$, $\theta_1 - \theta_3$ and $\theta_1 - \theta_4$ are zero, or

$$0 = \theta_1 - \theta_i = \alpha_1 - \alpha_i + \frac{1}{3}\sum_{j=1}^{3}\gamma_{1j} - \frac{1}{3}\sum_{j=1}^{3}\gamma_{ij}, \quad i = 2, 3, 4.$$

One may write

$$\boldsymbol{C} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & -\frac{1}{3} & \frac{1}{3} & 0 & 0 & -\frac{1}{3} & \frac{1}{3} & 0 & 0 & -\frac{1}{3} \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{1}{3} \end{pmatrix}.$$

14. Drug type appears to be the main factor with a significant effect. Regardless of the Types of Sum of Squares, the p-value associated with Drug is very small (≈ 0.00012 using the Type III Sum of Squares and slightly smaller using the Type I.) Disease does not appear to have a substantial effect on mean blood pressure change in dogs nor does the effect of Drug appear to depend on type of Disease (i.e., little evidence of interaction in this analysis).

Part II

15. Due to missing data, block differences cannot be estimated free of the effects for diseases and drugs. For example, Disease 1 is used in Clinic 2 only once while it shows up at least three times in any of the other clinics. Likewise, Drug 3 is used three times in Clinic 4, but only used once in Clinic 1. So, we cannot separate disease and drug effects from the effects of blocks (different clinics).

16. Yes, according to the analysis, there appears to be substantial variability between clinics. Specifically, the estimated MSE in the model not accounting for clinic is about 110. In comparison, by including clinic as a fixed a effect, the overall amount of residual variability decreases to 22 – a substantial drop. A consequence of being able to attribute a large portion of the variability to clinics is that we can detect a significant drug*disease interaction in the ANOVA table including clinic; this is true despite partial confounding of clinic with the main/overall effects of disease and drug.

17. Just based on observations in clinic 6, one could estimate $(\gamma_{ij}-\gamma_{i'j})-(\gamma_{ij'}-\gamma_{i'j'})$ for $i \neq i' \in \{1,2\}$ and $j \neq j' \in \{1,2,3\}$. Just based on observations in clinic 3, one could estimate $(\gamma_{ij} - \gamma_{i'j}) - (\gamma_{ij'} - \gamma_{i'j'})$ for $i \neq i' \in \{2,4\}$ and $j \neq j' \in \{1,2,3\}$. And just based on observations in clinic 1, one could estimate $(\gamma_{ij} - \gamma_{i'j}) - (\gamma_{ij'} - \gamma_{i'j'})$ for $i \neq i' \in \{1,4\}$ and $j \neq j' \in \{1,2,3\}$. Then, any interaction of interest is indeed estimable.

18. Consider profile plots and compare cell means of interest or contrasts within the levels of a factor using appropriate tests. The significant interaction suggests that overall drug and disease effects are not of much interest, particularly given their confounding with clinic as well.

Part III

19. $\widehat{\sigma}^2_\tau = 3638.45$ and $\widehat{\sigma}^2_\varepsilon = 22.0760$. Relatively speaking, the estimates confirm our earlier findings that a considerable amount of variability stems from the clinics.

20. The number of error contrasts involved is $n - r$ where $r = rank(X)$, the model matrix. Here $r = rank(X) = 17$, hence we need 58-17=41 error contrasts.

21. The main reason for modeling clinic as a random effect is tied to the generalizability of the results related to the fixed effects. When modeling clinic as a random effect, we treat the observed clinics as a random sample of some larger population of clinics. This introduces additional variability and a variance components associated with clinic that requires estimation and thus additional uncertainty. The *cost* associated with modeling the random effects can come in the form of less precise inference for the fixed effects. However, for these data, this turns out to not be the case though. Additionally, with random effects, we use modeling assumptions, such as normality, where the appropriateness of such assumptions is difficult to check due to the small number (6) of clinics

1 Background and Data

In a typical wine competition, panels of four or five qualified judges evaluate "flights" of wines grouped by type. Although there are some differences among competitions, usually each judge rates each wine according to categories Gold, Silver, Bronze, No Award. After the votes there is typically a discussion among members of the panel in which consensus is sought and usually achieved. Note that in this phase of the competition wines are not ranked against one another, they are simply judged to be in one of the quality groups. In fact, our concern will only be whether a wine wins an award (Gold, Silver or Bronze) or does not win an award.

There are a number of key characteristics of wine that can be measured and quantified. These include (1) alcohol content, (2) acidity, and (3) tannin content. The context of this question is the analysis of measures of these three characteristics in wines, relative to values that have resulted in tasting outcomes of an award (Gold, Silver, or Bronze).

To make the problem manageable we will consider only one type of wine, Pinot Noir, produced from grapes grown in only one region of the world, the west coast of the United States. We will deal with measurements of alcohol in units of percent by volume (ABV), measurements of acidity in terms of pH (low pH is high acidity), and tannins in milligrams per liter (mg/L) catechin equivalents as determined by a protein precipitation assay. While there is variability among Pinot Noir wines in terms of our three characteristics of interest, there are also ranges for each, outside of which a wine could not be considered to be Pinot Noir. For our purposes here, we will take these ranges to be $12.5 - 16.0$ for alcohol, $3.2 - 5.0$ for pH, and $40 - 1000$ for tannins.

We have available measurements of alcohol for 70 award-winning Pinot Noir wines and paired values of pH for 32 of those wines. We have values for tannins for 49 wines, also award winning, but these values are not paired with either measurements of alcohol or pH.

2 Modeling Marginal Distributions

As a first step in our analysis of this problem, consider fitting marginal theoretical probability distributions to the data displayed as stem plots on pages 9–10 and as histograms in Figure 1.

1. At this point we are concerned only with fitting marginal distributions to the three separate characteristics of alcohol, pH, and tannins. Define notation for the random variables appropriate for this problem. Assume that, for a given characteristic, measurements across bottles of wine are independent.

We have assumed the allowable ranges of values for a wine to be considered a Pinot Noir to be $(12.5, 16.0)$ for alcohol, $(3.2, 5.0)$ for pH, and $(40, 1000)$ for tannins. One of our challenges in choosing theoretical probability distributions will be to match these ranges with the support of the chosen distributions.

To develop a model for alcohol content, we might note that the stem plot for this characteristic and the histogram in the top panel of Figure 1 appear to describe a distribution that is both unimodal and skew left.

We might recall that there is a version of an extreme value distribution that has support on the positive line but can be both unimodal and skew to the left. So one possible choice of a model for alcohol is the left-skew version of an extreme value distribution. We could truncate the model below at 12.5 and above at 16.0 and then fit the truncated model to the data. A plot showing this model fitted to the available data and overlain on the histogram of alcohol values is presented in Figure 2. This fitted distribution is truncated both below (at 12.5) and above (at 16.0) although that is not visually obvious in Figure 2.

The possible values of alcohol content are percentages, which can be easily expressed as proportions between 0 and 1, and this might make us think of a beta distribution to model alcohol, possibly truncated at values of 0.125 and 0.160.

2. A standard beta distribution for a random variable X has density, with $\alpha > 0$ and $\beta > 0$,

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \, x^{\alpha-1} \, (1 - x)^{\beta-1}; \quad 0 < x < 1. \tag{1}$$

A few facts about this distribution are that it has expected value $\alpha/(\alpha + \beta)$, if $\alpha > 1$ and $\beta > 1$ it has a unique mode at $(\alpha - 1)/(\alpha + \beta + 2)$, and it is skew left if and only if $\alpha > \beta > 1$. Given this, would it be possible to obtain either a standard or truncated beta distribution that is unimodal, skew left, and for the standard beta, places negligible probability outside of the interval $(0.125, 0.160)$?

Note: We are using Karl Pearson's definition of skewness, which applies only to unimodal distributions, which does not include distributions that are J−shaped (or L−shaped or U−shaped for that matter).

3. Because a beta distribution has bounded support, another option is to map alcohol (as a proportion) onto the interval $(0.125, 0.160)$. Suppose that X is a random variable having a standard beta distribution with parameters $\alpha > 0$ and $\beta > 0$ and density (1). With $0 < a < 1$ and $0 < b < 1$, give the transformation of X that results in a random variable Y with possible values on the interval (a, b) and derive the log likelihood – you may omit any terms that are constants for the log likelihood.

Figure 2 shows the fitted distribution from **Question 3**. At least visually, this fitted distribution gives little indication of skewness, certainly much less so than the truncated extreme value distribution also shown in this figure.

The choices of distributions for pH could be considered to be normal or extreme value distributions, with both of these being truncated on the left at a value of 3.3 and the right at 5.0. The fits of these two models to the pH data are shown in Figure 3. There appears to be little reason to prefer one model over the other from a visual standpoint.

The distribution of tannin concentrations appears skew right in form, particularly when represented as a stem plot. The natural choice for a model is a gamma distribution. A fitted gamma distribution truncated on the left at 40 and on the right at 1000 is shown in Figure 4. Similar to the upper truncation of distributions fit to pH in Figure 3, the lower truncation of this gamma model for tannins is not evident on the graph of Figure 4.

4. We have arrived at two potential models for alcohol. These models differ not only in distributional shape, but also in the scale on which support is expressed, $(0.125, 0.160)$ for the beta-based models and $(12.5, 16.0)$ for the extreme value model. We have posited two models for pH, normal and extreme value, with both truncated at the same values below and above. And we have suggested one model for tannnin concentrations, again truncated at both extremes. None of these models appear visually to be wonderful descriptions of the available data (Figures 2 to 4) but we know that we should not rely on visual evidence

alone in drawing conclusions. Suppose that we fit these distributions to the available data using maximum likelihood. We would like to both compare potential models for a given wine characteristic and also have some sense of goodness of fit for a chosen model. It would be preferable to avoid using multiple procedures based on multiple theoretical constructs to achieve assessment of potential models. Outline a single procedure that could be used to (1) compare competing models for alcohol and pH and (2) determine how well the chosen models, as well as the gamma-based model for tannins, describe the marginal distributions in the observed data.

Assume that, based on whatever procedure we develop, we are willing to accept the following models as representations of the three wine characteristics of concern.

- Alcohol: Left-skew extreme value distribution truncated to have support on $(12.5, 16.0)$. For random variables $A_i \;\; i = 1, \ldots, n_A$ defined in Question 1 and having possible values $a \in (12.5,\, 16.0)$, and parameters $-\infty < \lambda < \infty$, $\theta > 0$ the common probability density function is,

$$f(a|\lambda, \theta) = \frac{1}{K_{EV}(\lambda, \theta)} \exp\left(\frac{a-\lambda}{\theta}\right) \exp\left[-\exp\left(\frac{a-\lambda}{\theta}\right)\right], \tag{2}$$

  where

$$K_{EV}(\lambda, \theta) = \int_{12.5}^{16.0} \exp\left(\frac{t-\lambda}{\theta}\right) \exp\left[-\exp\left(\frac{t-\lambda}{\theta}\right)\right] \, dt$$

- pH: Normal distribution truncated to have support on $(3.3,\, 5.0)$. For random variables $P_i;\;\; i = 1, \ldots, n_P$ associated with pH and having possible values $p \in (3.3,\, 5.0)$ and parameters $-\infty < \mu < \infty$ and $\sigma^2 > 0$, the common probability density function is,

$$g(p|\mu, \sigma^2) = \frac{1}{K_N(\mu, \sigma^2)} \exp\left[-\frac{1}{2\sigma^2}(p-\mu)^2\right], \tag{3}$$

  where

$$K_N(\mu, \sigma^2) = \int_{3.3}^{5.0} \exp\left[-\frac{1}{2\sigma^2}(t-\mu)^2\right] \, dt.$$

- Tannins: Gamma distribution truncated to have support on $(40, 1000)$. For random variables $T_k;\;\; k = 1, \ldots, n_T$ associated with tannins and having possible values $t \in (40, 1000)$

and parameters $\alpha > 0$, $\beta > 0$, the common probability density function is,

$$h(t|\alpha, \beta) = \frac{1}{K_G(\alpha, \beta)}\, t^{\alpha-1}\, \exp(-\beta\, t), \tag{4}$$

where,

$$K_G(\alpha, \beta) = \int_{40}^{1000} u^{\alpha-1}\, \exp(-\beta\, u)\, du.$$

We will consider the three distributions just listed as the marginal distributions that correspond to a "population" of award worthy Pinot Noir wines from the west coast of the U.S.

5. We would like to define a "target value" for each of the three characteristics based on what we know about the distributions of these factors in award winning wines. These would represent values which a newly released Pinot Noir wine should strive to meet in order to belong to the population of award winning wines that produced our data. One choice for such target values are the expected values of the individual characteristics of alcohol, pH, and tannins. There might be a number of justifications for choosing the expected value of an individual characteristic as a target value for that characteristic. Give one of them.

6. In principle, we could construct confidence intervals for each of the three expected values individually. Using pH as an example, let P be a random variable associated with the pH value of an arbitrarily chosen award winning wine. Based on our development to this point, we assume that P has a distribution with density (3). The target value for pH described in Question 5 is $E(P) = k(\mu, \sigma^2)$. Assume that we have estimated μ and σ^2 using maximum likelihood and that regularity conditions needed for asymptotic normality apply. Let I_P^{-1} denote either the expected or observed inverse information matrix for our model. The delta method would indicate an appropriate variance for computation of a Wald interval is,

$$V_P = \left(\frac{\partial k(\mu, \sigma^2)}{\partial \mu}, \frac{\partial k(\mu, \sigma^2)}{\partial \sigma^2} \right) \; I_P^{-1} \; \left(\begin{array}{c} \frac{\partial k(\mu, \sigma^2)}{\partial \mu} \\ \frac{\partial k(\mu, \sigma^2)}{\partial \sigma^2} \end{array} \right) \Bigg|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2}.$$

Write out the forms that would need to be computed to obtain values for $k(\mu, \sigma^2)$ and its derivatives. These expressions should be in terms of $g(p|\mu, \sigma^2)$, not in terms of the

particular formula for $g(p|\mu, \sigma^2)$ given in (3).

Hint: The expected value of a random variable that follows density (3) is not μ. Again, your answer should be written in terms of $f(p|\mu, \sigma^2)$.

7. Suppose we compute a set of confidence intervals for each of our three target values using confidence coefficients $\{0.90, 0.80, 0.70, 0.60, 0.50\}$. Again using pH as an example, consider two unclassified wines with pH values of p_1 and p_2. Suppose that p_1 is contained in the confidence intervals computed at levels of $0.50, 0.60$ and 0.70, but not those computed at levels of 0.80 or 0.90. Suppose that p_2 is contained in all of the intervals. Briefly explain why it would be incorrect to conclude that p_2 is closer to the target value for pH, namely $E(P)$, than is p_1.

3 Relations Among Characteristics

To this point we have dealt only with marginal distributions of our three wine characteristics of alcohol, pH, and tannins. It is well known that these characteristics are related to one another in various and sometimes complex ways. There are, however, several scales at which such associations function: global, local, and taste perception. By global association we mean connections across wine types. For example, red wines tend to have high tannins, high alcohol, and high pH, while dry white wines have low tannins, low alcohol, and low pH. Local associations, if they exist, would be realized for a single given wine type, with alcohol, pH, and tannins all contained within their respective ranges for that type. Figure 5 shows a scatterplot of pH against alcohol for the 32 wines on which we have paired alcohol and pH data. There is no association evident in this plot. While this is not definitive evidence that alcohol and pH are uncorrelated, yet alone independent, it does support an assumption of mutual independence to some degree. Given that we have no information about tannins jointly with either pH or alcohol, we will assume our three characteristics are independent at the local level. The implication for our analysis here is that we will assume random variables associated with alcohol, pH and tannins are independent within bottles of wine as well as among bottles of wine.

Despite apparent independence among alcohol and pH for at least the wines on which we have

data, it is known that these factors are physically related in the fermentation of a given batch of wine. A vintner can exercise at least some control over alcohol and pH during fermentation, during which alcohol increases and pH decreases. Tannins are largely determined by the grapes themselves, not fermentation, but knowing the level of tannins in a wine may cause a vintner to make lesser or greater efforts to modify alcohol and pH during the production process.

The relations between alcohol and pH during fermentation, and their influence on how concentration of tannins could be potentially manipulated are important because taste perceptions caused by changes in one characteristic can be, to some extent, compensated for by changes in other characteristics. In particular, it is possible to some degree to manipulate pH and alcohol to "compensate" for tannins that are judged to be too high or too low. To offset higher than desired tannins, one can attempt to increase alcohol or pH. To offset lower than desired tannins, one can attempt to decrease alcohol or pH. Because the process of fermentation increases alcohol and decreases pH, it is not possible to manipulate both of these factors in the same direction. To make matters even more complex, higher alcohol can also offset lower pH, because alcohol is sweet and sweetness masks acidity.

To simplify this somewhat, we have two sets of taste-related relations that may be important:

TR1: High tannins can be offset by higher alcohol or higher pH and low tannins can be offset by lower alcohol or lower pH.

TR2: Lower pH can be offset by higher alcohol and higher pH can be offset by lower alcohol.

We would like to construct an index or several indices of quality that can be applied to Pinor Noir wines. We might base such indices on the difference of values of alcohol, pH, and tannins from the respective target values of $E(A)$, $E(P)$, and $E(T)$ for wines in our population of award eligible wines. Let $\hat{\mu}_A$, $\hat{\mu}_P$ and $\hat{\mu}_T$ denote estimated values of $E(A)$, $E(P)$ and $E(T)$, respectively. For a wine with measured alcohol a_0, pH p_0, and tannins t_0, an index based on total deviation would be the $L1$ norm,

$$Q = |a_0 - \hat{\mu}_A| + |p_0 - \hat{\mu}_P| + |t_0 - \hat{\mu}_T|, \tag{5}$$

but this would not take into account the information in **TR1** and **TR2**. In addition, both the ranges of possible values and the units of measurement differ for alcohol, pH, and tannins.

A unit change in pH may mean something quite different than a unit change in alcohol, for example.

8. Suggest a modification to Q in (5) that would attempt to take into account the information in **TR1** and mitigate the difficulty that alcohol, pH, and tannins are measured in different units. If possible, suggest an additional modification to what you already have that would attempt to take into account the information in both **TR1** and **TR2**. In your answer, use the notation already given in (5) plus $\hat{\sigma}_A^2$, $\hat{\sigma}_P^2$, and $\hat{\sigma}_T^2$ to denote estimated variances for alcohol, pH, and tannins, respectively.
Hint: Address the issue of different units of measurement first, to create a unitless version of Q. Then modify that quantity to account for **TR1** *and* **TR2**.

4 Bayesian Considerations

Now suppose we choose to take a Bayesian approach to estimation and inference in this problem. Given assumed independence among random variables both within and among sampling units (bottles of wine), and supposing that we do not formulate a joint prior that incorporates dependencies among data model parameters from different variables, the problem reduces to three separate Bayesian procedures which will require Markov Chain Monte Carlo for approximation of posterior distributions. We may consider, however, that one iteration of the overall estimation algorithm consists of one iteration of all three portions to produce a value from the joint posterior

$$p(\lambda, \theta, \mu, \sigma^2, \alpha, \beta | \boldsymbol{a}, \boldsymbol{p}, \boldsymbol{t}) = p(\lambda, \theta | \boldsymbol{a})\, p(\mu, \sigma^2 | \boldsymbol{p})\, p(\alpha, \beta | \boldsymbol{t}).$$

9. In a typical normal model with a normal prior for the expected value and an inverse gamma prior for the variance, we obtain conditional posteriors through conditional conjugacy. In this question you are asked to determine whether conditional conjugacy holds in the case of a truncated data model. For this question, use random variables $Y_1, \ldots, Y_n$ that have normal distributions with parameters μ and σ^2, truncated to the interval (a, b). Assume that we specify a prior for μ that is $N(\mu_0, \tau_0^2)$ also truncated to the interval (a, b). Derive the conditional posterior of μ given a known σ^2.

10. Let whatever index of quality we have arrived at in **Question 8** be denoted as Q. A certain sommelier believes that for a particular wine tasting event if a new, previously unevaluated wine lies in the first quartile of the distribution of index Q among award worthy wines, it has a 50% chance of winning an award, if it lies between the first and second quartiles it has a 70% chance of winning an award, and if it lies between the second and third quartiles it has a 90% chance of winning an award. Outline, in algorithmic form, a procedure that will take the MCMC output representing draws from the joint posterior and determine the values of Q that correspond to these cut-points. In this algorithm you do not need to know the formula for Q, you can just say "compute Q".

Note on Data Sources: These data were taken from a Pinot Noir blog called the PinotFile Vol. 10, Issue 14, July 12,2015 (https://www.princeofpinot.com/article/1706/). Data for tannins in 49 wines had to be simulated, but that was done to match the marginal distribution of tannins shown in Figure 2(B) in Harbertson *et al.* (2008), Variability of tannin concentration in red wines. American Journal of Enology and Viticulture **59**, 210-214.

5 Figures

1. Alcohol

```
The decimal point is 1 digit(s) to the left of the |

132 | 0
134 | 000
136 | 0
138 | 0000
140 | 0000
142 | 000000
144 | 00000000
146 | 0
148 | 0000
```

2. pH

```
The decimal point is 1 digit(s) to the left of the |

34 | 0335679
35 | 89
36 | 0112668
37 | 00114559
38 | 000569
39 | 3
40 | 0
```

3. Tannins

```
The decimal point is 2 digit(s) to the right of the |

 0 | 666
 1 | 003345679
 2 | 02355777
 3 | 3678899
 4 | 256667
 5 | 001238889
 6 | 127
 7 | 37
 8 | 05
```

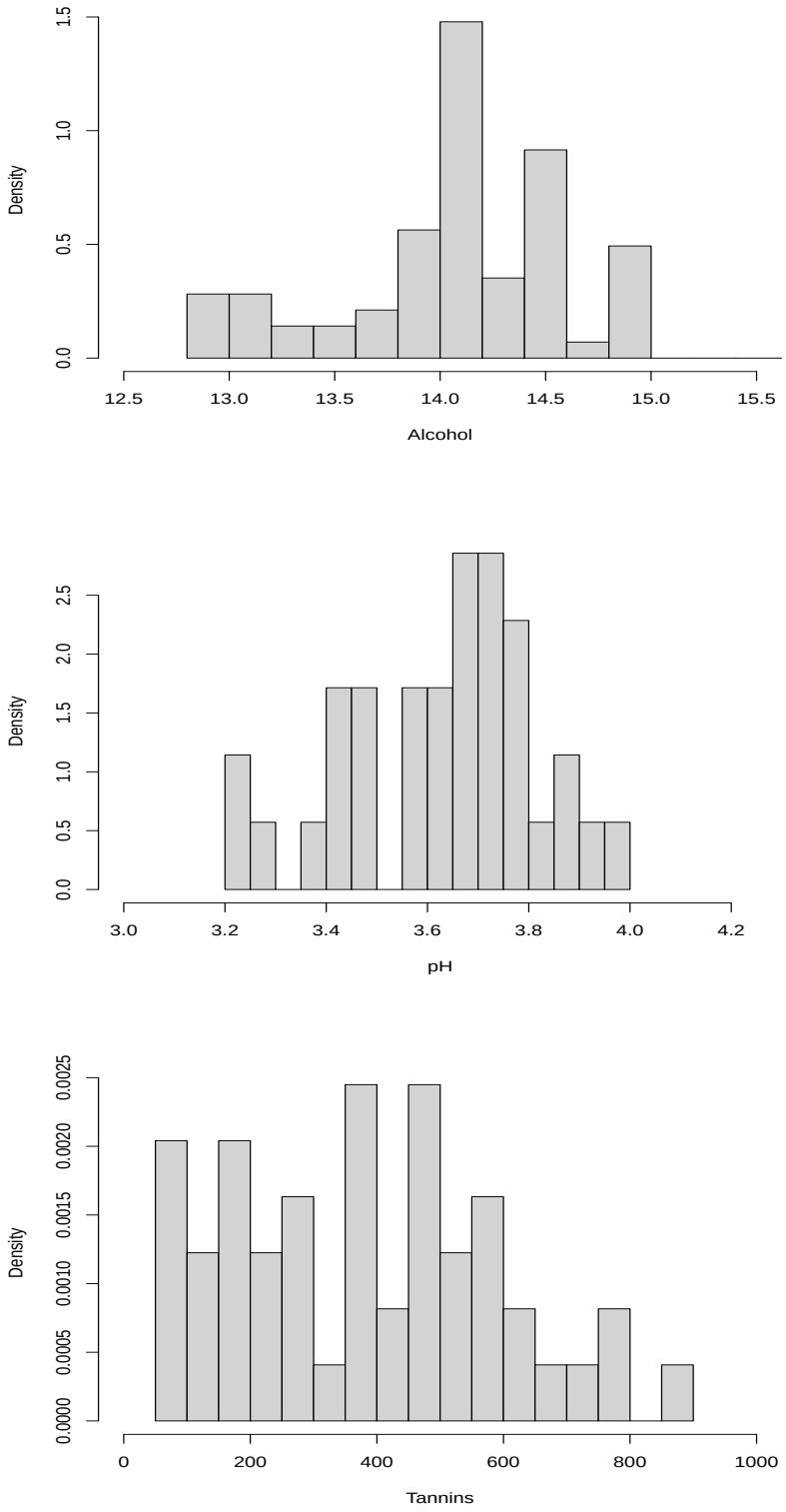

Figure 1: Histograms of alcohol (top), pH (middle), and tannins (bottom) in award-winning pinot noir wines.

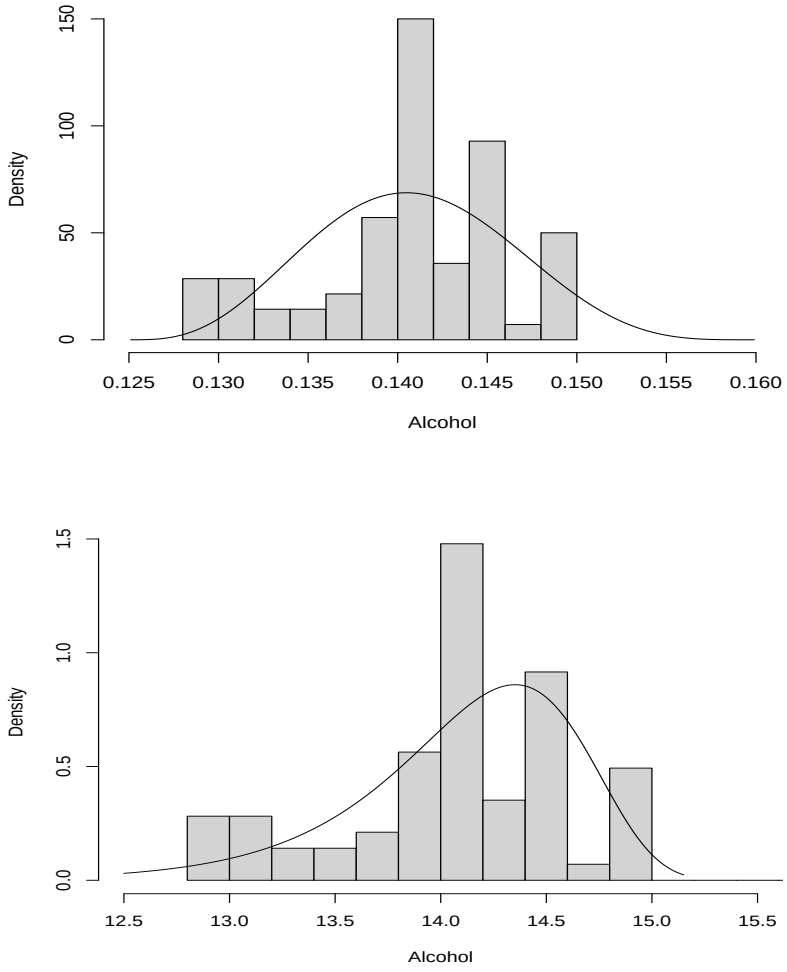

Figure 2: Histograms of alcohol with fitted models based on the beta distribution (top) and extreme value distribution (bottom).

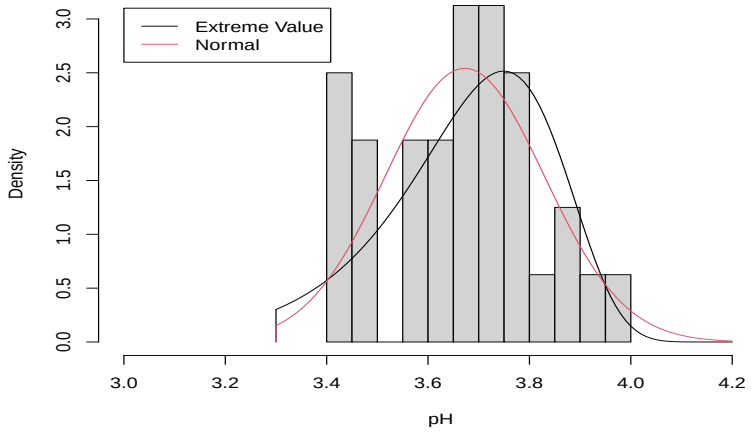

Figure 3: Histogram of pH with fitted models based on the normal and extreme value distributions.

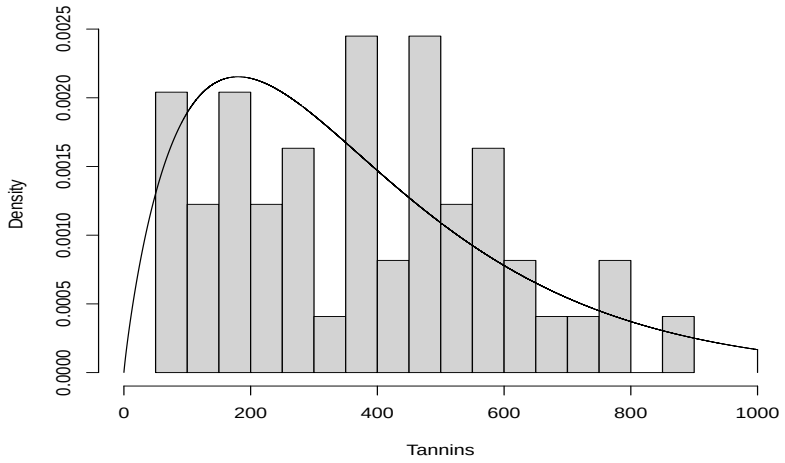

Figure 4: Histogram of tannin concentrations with fitted gamma distribution truncated at 1000.

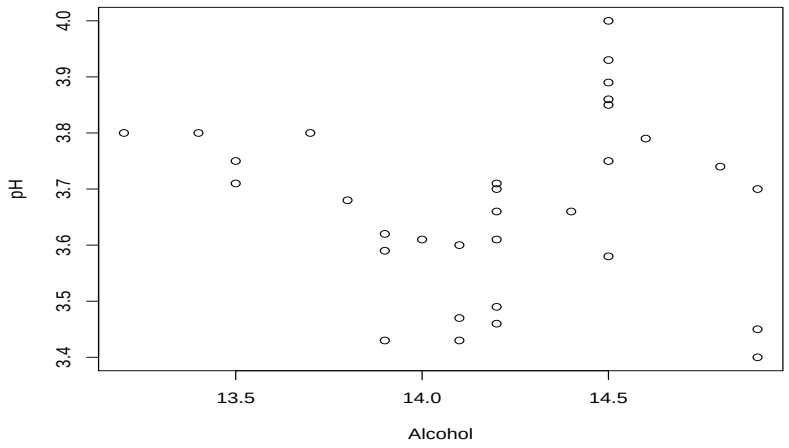

Figure 5: Scatterplot of pH versus alcohol for award winning Piont Noir wines.

These are a sketch of the answers hoped for. Other possibilities might exist for some of the questions that would be entirely adequate if they are both technically correct and logically consistent.

1.
- Define random variables A_i to be connected with the alcohol content of award worthy wines, $i = 1, \ldots, n_A$. Here, $n_A = 70$.
- Define random variables P_i to be connected with the pH of award worthy wines, $i = 1, \ldots, n_P$. Here, $n_P = 32$. Note that this assumes that the first n_P wines are those that have paired values of alcohol and pH associated with them.
- Define random variables T_k to be connected with the concentration of tannins in award worthy wines, $k = 1, \ldots, n_T$. Here, $n_T = 49$.

2. It would not be possible to formulate a unimodal beta density that is skew left and either restricted to the interval (0.125, 0.160) or one that places all but negligible probability on that interval. Consider first a standard beta density with support on the unit interval. Using the usual parameterization, if X has a beta distribution with parameters α and β and $\alpha > \beta > 1$ (so that it has a unique mode and is skew left), then

$$E(X) = \frac{\alpha}{\alpha + \beta} > 0.50,$$

and the mode is greater than $E(X)$, because it is skew left. This cannot occur if X is restricted to the interval $(0.125, 0.160)$ or places the majority of its probability there.

Because truncating any beta distribution does not change the shape of the density within the interval of truncation, it is not possible for a beta distribution with unique mode greater than 0.50 to be truncated to an interval entirely less than 0.50 and then have a mode in that interval.

3. If $X \sim Beta(\alpha, \beta)$ then let $Y = a + (b - a)X$. The density of Y is then, for $\alpha > 0$

and $\beta > 0$,

$$g(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{1}{(b-a)^{\alpha+\beta-1}} (y-a)^{\alpha-1} (b-y)^{\beta-1};\ ,\ \ a \le y \le b.$$

The log likelihood for a sample $y_1, \ldots, y_n$ is then,

$$\begin{aligned} \ell &= \sum_{i=1}^{n} \ell_i \\ \ell_i &= (\alpha-1)\log(y_i-a) + (\beta-1)\log(b-y_i) - (\alpha+\beta-1)\log(b-a) + \log\{\Gamma(\alpha+\beta)\} \\ &- \log\{\Gamma(\alpha)\} - \log\{\Gamma(\beta)\}. \end{aligned} \tag{1}$$

4. Note: this questions has multiple possible answers that would be adequate. What is presented here is one of those.

The problem involves comparison of non-nested models for two sets of data, and goodness of fit for three sets of data. To arrive at one procedure that can handle these needs we can look to methods based on generalized residuals or possibly empirical distribution functions for discrete versions of the fitted models. For our three characteristics of alcohol content, pH, and concentration of tannins:

- Alcohol.

  For independent and identically distributed random variables $A_1, \ldots, A_{n_a}$, let $g_a(a|\hat{\alpha},\hat{\beta})$ denote the (fitted) common beta-based model restricted to (0.125, 0.16) and let $f_a(a|\hat{\lambda},\hat{\theta})$ denote the (fitted) common extreme value model truncated to (12.5, 16.0). For the set of observed values $\{a_i :, i = 1, \ldots, n_a\}$, define generalized residuals,

$$\begin{aligned} u_{1,i} &= \int_0^{a_i} f_a(t|\hat{\lambda},\hat{\theta})\, dt \\ u_{2,i} &= \int_0^{a_i} g_a(t|\hat{\alpha},\hat{\beta})\, dt \end{aligned} \tag{2}$$

  The empirical distribution functions of these residuals are, for $0 < u < 1$,

$$\begin{aligned} \tilde{F}_a(u) &= \frac{1}{n_a} \sum_{i=1}^{n_a} I(u_{1,i} \le u) \\ \tilde{G}_a(u) &= \frac{1}{n_a} \sum_{i=1}^{n_a} I(u_{2,i} \le u) \end{aligned} \tag{3}$$

Here, we could compare $\tilde{F}(u)$ and $\tilde{G}(u)$ either visually as a diagnostic or first compute Kolmogorov statisitcs for tests of these distributions against a theoretical Uniform $(0, 1)$ and then compare as a test procedure. If neither appear reasonably close to uniform we could also conduct a test of equal distribution using a two-sample Kolmogorov-Smirnov procedure.

- pH.

  We again have two potential models for pH. Let $P_1, \ldots, P_{n_p}$ denote independent and identically distributed random variables associated with pH. Let $g_p(p|\hat{\mu}, \hat{\sigma}^2)$ denote the (fitted) common normal-based model truncated to (3.3, 5.0) and let $f_p(y|\hat{\psi}, \hat{\phi})$ denote the (fitted) common extreme value model truncated to (3.3, 5.0). For the set of observed values $\{p_i :, i = 1, \ldots, n_p\}$, define generalized residuals exactly as in (2) with g_p and f_p replacing f_a and g_a. Compute empirical distribution functions $\tilde{F}_p$ and $\tilde{G}_p$ for these sets of residuals as in (3) with the obvious modifications for the variable pH rather than alcohol. As for modeling alcohol content, the empirical distribution functions may be compared to each other or to theoretical Uniform distributions.

- Tannins.

  Here we have only one posited model, that being a truncated gamma distribution. For independent and identically distributed random variables $T_1, \ldots, T_{n_t}$ let the (fitted) common gamma-based density function be denoted as $f_t(t|\hat{\alpha}, \hat{\beta})$ and define generalized residuals for the set of observed values $\{t_k : i = 1, \ldots, n_t\}$ as,

  $$u_k = \int_0^{t_k} f_t(u|\hat{\alpha}, \hat{\beta})\, du$$

  and denote the empirical distribution function of these residuals as, for $0 < u < 1$,

  $$\tilde{F}_t(u) = \frac{1}{n_t} \sum_{i=k}^{n_t} I(u_k \leq u)$$

  A goodness of fit test can be conducted through the use of a Kolmogorov statistic comparing $\tilde{F}_t$ to a theoretical Uniform $(0, 1)$.

5. One justification for the use of expected values for a target quantity is provided by consideration of a prediction problem using squared error loss. The expected value minimizes prediction mean squared error for a random variable from a given distribution.

6. Note that the purpose of the hint was to emphasize that the usual relations between parameters and moments for particular models no longer hold when those models are truncated. What is needed is to use definitions. Using the interval for $\mu_P = E(P)$ as an example, and assuming we can interchange integration and differentiation, what we would need to compute are

$$\begin{aligned} \mu_P &= k(\mu, \sigma^2) = \int_{3.3}^{5.0} p\, g(p|\mu, \sigma^2)\, dp \\ \frac{\partial\, k(\mu, \sigma^2)}{\partial\, \mu} &= \int_{3.3}^{5.0} \frac{\partial}{\partial\, \mu}\, p\, g(p|\mu, \sigma^2)\, dp \\ \frac{\partial\, k(\mu, \sigma^2)}{\partial\, \sigma^2} &= \int_{3.3}^{5.0} \frac{\partial}{\partial \sigma^2} p\, g(p|\mu, \sigma^2)\, dp \end{aligned}$$

7. The interpretation would not be valid because it implies that the confidence regions obtained have a systematic relation with the true but unknown value of the quantity being estimated, in this case $\mu_P = E(P)$. It would not be correct to claim that a value contained in the 0.80 region is necessarily closer to the true but unknown value than a value in the 0.50 region, for example. Note that this is the same fallacy as interpreting values closer to the center of an interval as being closer to the truth than values farther from the center of the interval.

8. Following the Hint, we can construct a unitless version of the index Q through studentization. Using notation suggested in the question, let

$$\begin{aligned} \tilde{a}_0 &= \frac{a_0 - \hat{\mu}_A}{(\hat{\sigma}_A^2)^{1/2}}, \\ \tilde{p}_0 &= \frac{p_0 - \hat{\mu}_P}{(\hat{\sigma}_P^2)^{1/2}}, \\ \tilde{t}_0 &= \frac{t_0 - \hat{\mu}_T}{(\hat{\sigma}_T^2)^{1/2}}. \end{aligned}$$

A unitless version of the index Q given in the question is then,

$$\tilde{Q} = |\tilde{a}_0| + |\tilde{p}_0| + |\tilde{t}_0|.$$

The taste relations given in **TR1** indicate that large values of $\tilde{t}_0$ can be offset by large values of $\tilde{a}_0$ and/or $\tilde{p}_0$, which suggests dropping absolute values and modifying $\tilde{Q}$ as,

$$\tilde{Q}_1 = \tilde{t}_0 - \tilde{a}_0 - \tilde{p}_0.$$

But we know that during fermentation alcohol and pH change in opposite directions, which might cause us to use, instead,

$$\tilde{Q}_2 = [\tilde{t}_0 - \max\{\tilde{a}_0, \tilde{p}_0\}]I(\tilde{t}_0 > 0) + [\tilde{t}_0 - \min\{\tilde{a}_0, \tilde{p}_0\}]I(\tilde{t}_0 < 0).$$

The taste relations given in **TR2** then suggest that low pH can be offset by high alcohol and vice versa. A simple way to incorporate this information would be to then add $\tilde{a}_0 + \tilde{p}_0$ to what we already have as,

$$\tilde{Q}_3 = \tilde{Q}_2 + (\tilde{a}_0 + \tilde{p}_0).$$

9. For this question we have the common density function of $Y_1, \ldots, Y_n$ being,

$$f_Y(y|\mu, \sigma^2) = \frac{1}{K_y(\mu, \sigma^2)} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right],$$

where

$$K_Y(\mu, \sigma^2) = \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{b} exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] dy - \int_{-\infty}^{a} exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] dy\right).$$

The prior for μ is,

$$\pi(\mu|\mu_0, \tau_0^2) = \frac{1}{K_\mu} \frac{1}{(2\pi\tau_0^2)^{1/2}} \exp\left[-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right].$$

where

$$K_\mu = \frac{1}{(2\pi\tau_0^2)^{1/2}} \left(\int_{-\infty}^{b} exp\left[-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right] d\mu - \int_{-\infty}^{a} exp\left[-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right] d\mu\right).$$

Note that $K_Y(\mu, \sigma^2)$ is a function of the data model parameters μ and σ^2, while K_μ is a constant with respect to either of those parameters (μ_0 and σ_0^2 having fixed values chosen as part of prior specification).

The conditional posterior for μ given σ^2 based on a sample $Y_1, \ldots, Y_n$ is then,

$$p(\mu|\boldsymbol{y}, \sigma^2) \propto \frac{1}{K_Y(\mu, \sigma^2)} \exp\left[-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 - \frac{1}{\tau_0^2}(\mu - \mu_0)^2\right].$$

Note that $K_Y(\mu, \sigma^2)$ is now a function of μ alone since σ^2 is considered given. In the expression for $p(\mu|\boldsymbol{y}, \sigma^2)$ we may complete the square in the exponentiated term in the usual manner, giving

$$p(\mu|\boldsymbol{y}, \sigma^2) \propto \frac{1}{K_Y(\mu, \sigma^2)} \exp\left[-\frac{1}{2V}(\mu - M)^2\right]$$

where

$$M = \frac{n\tau_0^2\bar{y} + \sigma^2\mu_0}{n\tau_0^2 + \sigma^2};\ \text{and}\ V = \frac{\tau_0^2\sigma^2}{n\tau_0^2 + \sigma^2}.$$

The problem is that $K_Y(\mu, \sigma^2)$ is still a function of the argument μ and cannot be folded into the constant of proportionality. As a result, the posterior will not be normal or truncated normal. That is,

$$\int_a^b \frac{1}{K_Y(\mu, \sigma^2)} \exp\left[-\frac{1}{2V}(\mu - M)^2\right] d\mu \neq \frac{1}{K_Y(\mu, \sigma^2)} \int_a^b \exp\left[-\frac{1}{2V}(\mu - M)^2\right] d\mu.$$

10. What is desired are quartiles of the posterior predictive distribution of Q_1. To obtain these values,

(a) For each value of $\boldsymbol{\psi}_m;\ m = 1, \ldots, M$, simulate a value of a_m from $f_a(a|\lambda_m, \theta_m)$. This may require a Metropolis or rejection algorithm.

(b) Also for each value of $\boldsymbol{\psi}_m$, simulate a value of p_m from $g(p|\mu_m, \sigma_m^2)$ and a value of t_m from $h(t|\alpha_m, \beta_m)$. These are easily accomplished using the built-in R functions *rnorm* and *rgamma*, respectively. If the simulated value of a_m is below 12.5 or above 16.0 it is discarded and another draw is made with the same values of λ_m and θ_m. If the simulated value of p_m is below 3.3 or above 5.0 it is discarded and another draw is made with the same values of μ_m and σ_m^2. If

the simulated value of t_i is below 40 or above 1000 it is similarly discarded and another draw is made with the same values of α_m and β_m.

(c) If M is sufficiently large, all simulated value of a_m, p_m and t_m can be used. If M is only moderate, it might be advisable to make use of thinning to decrease correlation between successive values. For each set of values (a_m, p_m, t_m) and $\boldsymbol{\psi}_m$, compute Q_m.

(d) Calculate the quartiles of the empirical distribution of Q_m.