How come experimental design is central? The curious case of the oenophilist mishap

Dr. Bob L. Sturm, Lecturer
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK

December 15, 2015

Abstract

At the Centre for Digital Music, we are organising a seminar/tutorial series about statistics aimed at graduate students (and ourselves). We will address persistent and fundamental questions we all have when we go about collecting and/or analysing data, e.g., which statistical test should I use and why? How many subjects do I need? Can I get away with 20? What can I validly conclude? There is a massive literature about statistics (it is one of the greatest developments of the 20th century), but it will remain opaque as long as its fundamentals are absent in the training of researchers. This present tutorial (delivered Dec. 15, 2015) focuses on why experimental design is central to these common questions. In short, just plugging data into statistical packages and comparing p-values before considering the data collection is as senseless as shelling a clam after eating it.

1 Set-up

Consider the following scenario. As the local data science experts, we are contacted by a local chapter of oenophiles eager to have their data analysed in order to create an official ranking of local wines. They send us the table of results below with the description: "Four professional judges tasted four wines and scored each on a scale 1-5 (poor to excellent). Which wine is the best, and which is the worst, according to these judges? kthxbai!"

wine	scores
1	3 4 3 2
2	5 4 5 5
3	2 1 3 1
4	2 3 2 4

Table 1: Table of results for wine tasting.

2 Preliminaries

We start by computing some descriptive statistics of each wine from Table 1: mean $\hat{\mu}_w$, (unbiased) standard deviation $\hat{\sigma}_w$, and standard error of the mean (SEM) $\hat{\sigma}_w/\sqrt{4}$. These are shown in Table 2. Wine 2 could have the highest mean score, wine 3 could have the lowest, and wines 1 and 4 are somewhere in the middle. Are these significant? Assuming each mean is distributed Normal, but with unknown variance, we compute its 95% confidence interval (CI, $\hat{\mu}_w \pm 2.365\hat{\sigma}_w/\sqrt{4}$).[1] Figure 1 shows these intervals for each wine. We see from this that we might conclude at a significance level 0.05 that the mean score of wine 2 is significantly different from all the others. Performing ANOVA on all the scores produces a p-value

[1] This involves using a t-distribution having 3 degrees of freedom. From a table, such a distribution has 95% of its probability density in the interval $[-2.365, 2.365]$.

wine (w)	scores	$\hat{\mu}_w$	$\hat{\sigma}_w$	SEM	CI (95)
1	3 4 3 2	3.00	0.82	0.41	[2.03, 3.97]
2	5 4 5 5	4.75	0.50	0.25	[4.16, 5.34]
3	2 1 3 1	1.75	0.96	0.48	[0.62, 2.88]
4	2 3 2 4	2.75	0.96	0.48	[1.62, 3.88]

Table 2: Table of results for wine tasting. Now with basic statistics.

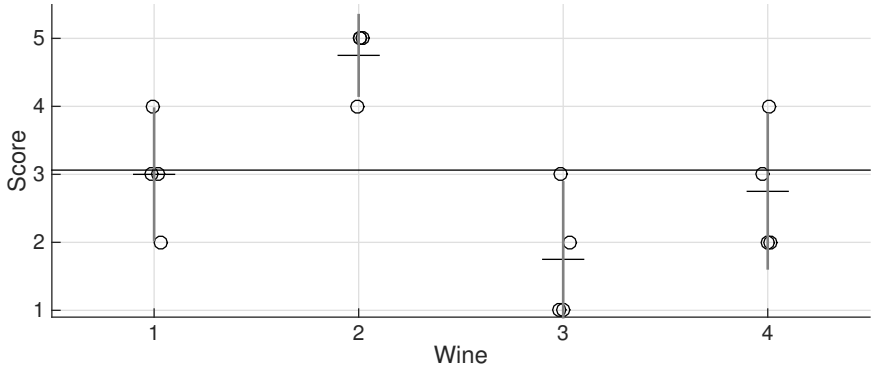


Figure 1: Scores (circles) (with random x-offset for visibility), data mean (horizontal line), wine means (short line segments), and 95% confidence intervals.

of 0.0021, which motivates rejecting the null hypothesis that there are no differences between the scores. Table 3 shows the results of pairwise comparisons on the scores of each pair of wines using ANOVA.[2] This also confirms wine 2 appears to be significantly different from the other three. It is a clear winner! We cannot conclude, however, that the mean scores of the other wines are significantly different from each other. Our estimated mean scores of wines 1 and 4 are higher than that of wine 3, but this could be due to chance. There are no clear losers, in other words.

wine a	wine b	p-value
1	2	0.04864
1	3	0.19802
1	4	0.97283
2	3	0.00126
2	4	0.02313
3	4	0.36261

Table 3: Pairwise comparisons of scores in Table 1 using ANOVA.

A serious question to consider is how our analysis is impacted by the fact that the measurements are restricted to be integers in $[1, 5]$. No doubt, we could apply a variety of different tests, depending on the different ways we define "best" and "worst;" but doing so is jumping the gun because we haven't considered the most important question first: *What can we validly conclude from the results in Table 1*?

I show below how, in fact, nothing can be concluded about these wines, no matter the scores in Table 1, because (SPOILER ALERT) the experimental design that led to its creation is "messed up." In more general terms, I show how the meaningful analysis of this data actually turns on the way it was collected in the first place, i.e., the experimental design. This demonstrates how "which statistical test to use and why" can be sensibly answered only after considering the experimental design.

[2]The MATLAB code I used is: `[p,tab,stats] = anova1(dataset(:,:)'); [c,m,h,nms] = ... multcompare(stats);`

3 Analysis

3.1 Principal aim of the experiment

Our analysis of Table 1, and indeed the process that led to its creation, hinges upon the principal aim of the experiment run by our local chapter of oenophiles, and the way in which they actually ran it. Do they want to rank the "wine quality" of these wines according to only these four particular judges? Or do they want to rank the "wine quality" of these wines according to The Wine Judging Population (TWJP), inferred from these four particular judges? These are two different experiments; but for now, let us envision that our local chapter of oenophiles wants to determine the TWJP consensus for these particular wines. This is an important clarification. If one had the resources (time, money, and wine), they could have all of TWJP score each wine, and thereby find with no uncertainty the "best" to "worst" wines with respect to TWJP.[3] Such a collection of observations produces the most high-resolution picture possible.[4] We do not have the resources however, and TWJP may not be so clearly defined or accessible.[5] Hence, we will just consider that the principal aim of this experiment is: *to determine the "best" and "worst" wines within cost according to these particular judges.*

3.2 Defining "best" and "worst"

Whatever "wine quality" is, one might not be able to directly measure and compare it; but our local chapter of oenophiles has measured "TWJP wine quality" scores. Since that is all we have to go on, we define the best wine as that which has a significantly higher mean score than all the others. The worst wine then is that which has a significantly lower mean score than all the others. We leave it to the experts to argue about whether the results have to do with the wine only (matter), or its qualia (perception), or a combination.

3.3 Modelling the measurements

Consider each score in Table 1 an outcome mapped to $\{1, 2, 3, 4, 5\}$ by a random variable Y_{wn}, where n denotes the score number, and w the wine. We model this random variable as a sum of the "true" (deterministic) score of wine w, denoted τ_w, perturbed by some "noise":

$$Y_{wn} = \tau_w + Z_{wn}. \tag{1}$$

Z_{wn} is random, and captures contributions unrelated to the wine, e.g., the variability of judging and scoring, the experience of a particular judge, and particulars of the experiment. We wish to estimate the parameters $\{\tau_w : w \in \mathcal{W}\}$, and compare them in statistically valid ways. More specifically, we wish to estimate how *different* these parameters are from each other. Toward this end, we decompose this model as

$$Y_{wn} = \bar{\tau} + (\tau_w - \bar{\tau}) + Z_{wn} \tag{2}$$

where $\bar{\tau}$ is the (deterministic) mean score of all the "true" scores of the wines in $\mathcal{W}$, and $\tau_w - \bar{\tau}$ is the deviation of the "true" score for wine w from the mean of the "true" scores in $\mathcal{W}$. Seen in terms of regression, we wish to fit each measurement with the following linear model

$$Y_{wn} = \beta_0 + \beta_1\delta_{w-1} + \beta_2\delta_{w-2} + \ldots + \beta_{|\mathcal{W}|}\delta_{w-|\mathcal{W}|} + Z_{wn} \tag{3}$$

where $\beta_0 := \bar{\tau}$, and $\beta_w := \tau_w - \bar{\tau}$, and $\delta_k = 1$ if $k = 0$ and zero otherwise. In this case, we are regressing on the wine w with δ_k turning on and off each contribution.[6]

[3] We are not interested in inferring whether their proclamations hold for other populations, e.g., school children.

[4] Think of determining the mean age of the living population of a country. We could determine it with no uncertainty at time t if we had the age of every living person at time t.

[5] What is the *population* of a country? Everyone that is within the borders at a specific time? Only citizens? What about those who have just died or been born?

[6] The use of δ_k here is also called "dummy coding."

3.4 Estimation of parameters

Given N measurements of each wine in $\mathcal{W}$, e.g., the collection of scores in Table 1, we wish to estimate τ_w for each wine and $\bar{\tau}$ for all wines. Equivalently, we wish to estimate $\{\beta_i : i \in \{0, 1, \ldots, |\mathcal{W}|\}\}$. By the method of least squares, we find $\hat{\beta}_0$ from minimising the sum of deviation squares:

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{w \in \mathcal{W}} \sum_{n=1}^{N} (y_{wn} - \hat{\beta}_0)^2 = 0 \Longrightarrow \hat{\beta}_0 = \frac{1}{|\mathcal{W}|} \frac{1}{N} \sum_{w \in \mathcal{W}} \sum_{n=1}^{N} y_{wn}. \tag{4}$$

Differentiating the sum of residual squares with respect to $\hat{\beta}_w$, we find the remaining parameters:[7]

$$\frac{\partial}{\partial \hat{\beta}_w} \sum_{w \in \mathcal{W}} \sum_{n=1}^{N} \left(y_{wn} - \hat{\beta}_0 - \sum_{w' \in \mathcal{W}} \hat{\beta}_{w'} \delta_{w-w'} \right)^2 = 0 \Longrightarrow \hat{\beta}_w = \frac{1}{N} \sum_{n=1}^{N} (y_{wn} - \hat{\beta}_0). \tag{6}$$

Then our least squares approximation to Y_{wn} is $\hat{Y}_w := \hat{\beta}_0 + \hat{\beta}_w$. The expectations of these estimators are:

$$\mathrm{E}[\hat{\beta}_0] = \beta_0 + \frac{1}{|\mathcal{W}|} \frac{1}{N} \sum_{w \in \mathcal{W}} \sum_{n=1}^{N} E[Z_{wn}] \tag{7}$$

$$\mathrm{E}[\hat{\beta}_w] = \beta_w + \frac{1}{|\mathcal{W}|} \frac{1}{N} \left(\sum_{w' \in \mathcal{W} \setminus \{w\}} \sum_{n=1}^{N} E[Z_{w'n}] - (|\mathcal{W}| - 1) \sum_{n'=1}^{N} E[Z_{wn'}] \right) \tag{8}$$

and their variances are easily seen to be[8]

$$\mathrm{Var}[\hat{\beta}_0] = \frac{1}{|\mathcal{W}|^2} \frac{1}{N^2} \mathrm{Var} \left[\sum_{w \in \mathcal{W}} \sum_{n=1}^{N} Z_{wn} \right] \tag{9}$$

$$\mathrm{Var}[\hat{\beta}_w] = \frac{1}{|\mathcal{W}|^2} \frac{1}{N^2} \mathrm{Var} \left[\sum_{w' \in \mathcal{W} \setminus \{w\}} \sum_{n=1}^{N} Z_{w'n} - (|\mathcal{W}| - 1) \sum_{n'=1}^{N} Z_{wn'} \right]. \tag{10}$$

We see that increasing N decreases the variance of each of these estimators. If for all wines and scores, Z_{wn} has zero mean then these estimators converge to the true parameter values (they are "unbiased"). If for all wines and scores Z_{wn} is independently and identically distributed (iid) with variance σ^2, then the above become[9]

$$\mathrm{Var}[\hat{\beta}_0] = \sigma^2 / (|\mathcal{W}| N) \tag{11}$$

$$\mathrm{Var}[\hat{\beta}_w] = (|\mathcal{W}| - 1)\sigma^2 / (|\mathcal{W}| N) = (|\mathcal{W}| - 1)\mathrm{Var}[\hat{\beta}_0]. \tag{12}$$

This clearly shows how our uncertainty in each of our parameter estimates depends on both the number of wines and the number of scores for each wine.

Figure 2 shows a simulation of the above.[10] We randomly draw four independent scores from each of four Gaussian distributions with means $\{\beta_0 + \beta_w : w \in \mathcal{W}\}$, and variance 0.1. Thus, $\{Z_{wn}\}$ are iid, zero mean with variance $\sigma^2 = 0.1$. We then estimate the parameters, and repeat 1,000,000 times. Finally, we using histogramming to approximate distributions showing the behaviour of our estimates. Figure 3 shows

[7] We could also use the method of Lagrangian multipliers to minimise

$$\sum_{w \in \mathcal{W}} \sum_{n=1}^{N} \left(y_{wn} - \hat{\beta}_0 - \sum_{w' \in \mathcal{W}} \hat{\beta}_{w'} \delta_{w-w'} \right)^2 \text{ subject to } \sum_{w \in \mathcal{W}} \sum_{n=1}^{N} \hat{\beta}_w = 0. \tag{5}$$

[8] By the fact that each estimator only varies by Z.

[9] The variance of a sum of iid random variables is the sum of their variances.

[10] See the appendix for MATLAB code to produce figures like these.

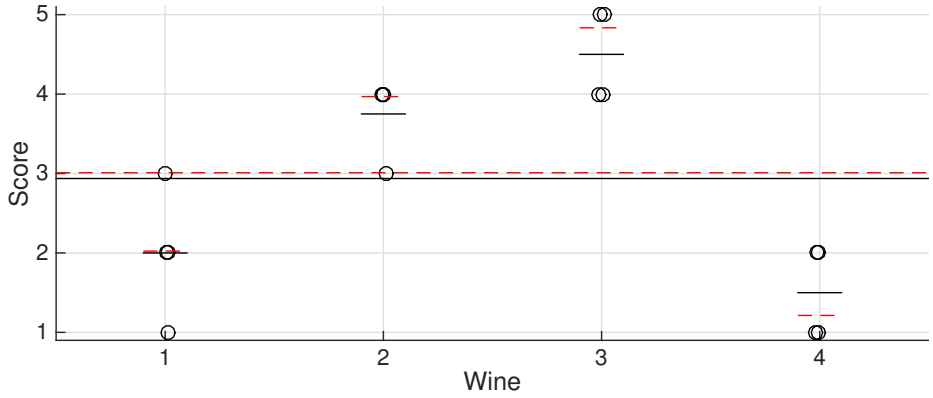
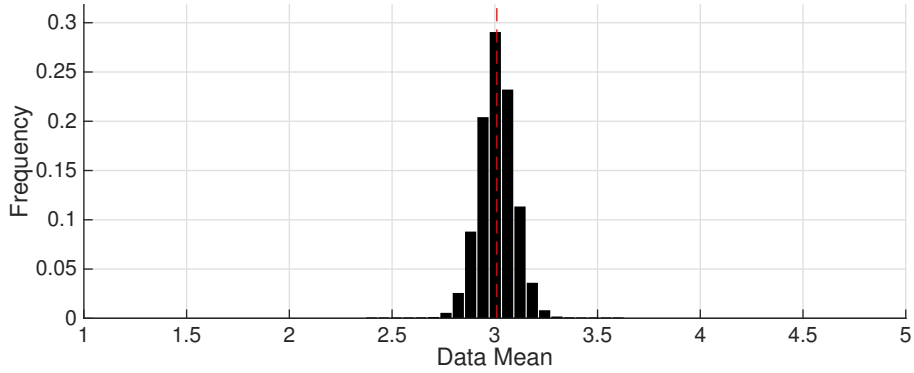
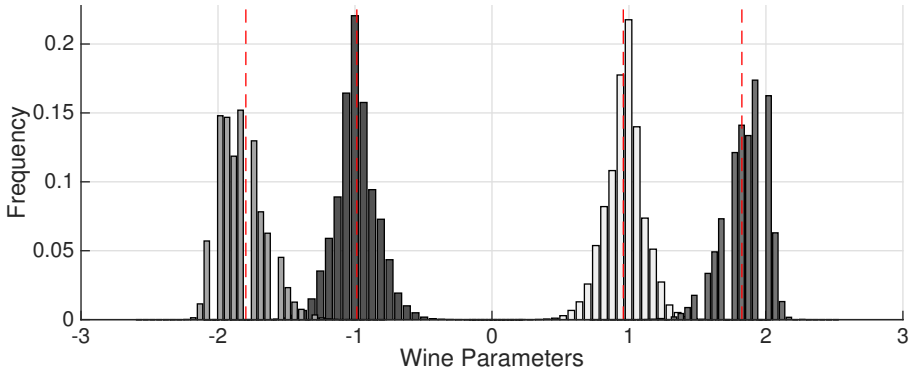


Figure 2: Simulation of 4 scores (circles, with random x-offset for visibility) of 4 wines, with Z_{wn} iid zero-mean Gaussian with $\sigma^2 = 0.1$. Data mean ($\hat{\beta}_0$) is black horizontal line, wine mean scores ($\hat{\tau}_w$) are short line segments. True mean (β_0) is red dashed line, and true wine scores (τ_w) are red dashed line segments.



(a) Mean $\hat{\beta}_0$



(b) Deviations $\hat{\beta}_w$

Figure 3: Parameter distributions for a simulation (1,000,000 trials) of 4 scores of 4 wines, with Z_{wn} iid zero-mean Gaussian with $\sigma^2 = 0.1$. True parameter values are shown as red dashed lines. These are the same parameters used in Fig. 2.

these for each parameter. While we see the variances of the deviation parameters are larger than that of the mean parameter, two observations are contrary to our predictions above. First, there does appear to be bias in the estimates even though the noise is zero mean. The bias we observe in a parameter in Fig. 3(b) is toward the nearest integer of the true parameter of the wine. Second, the variances we observe are larger than what we predict. The cause of these two differences is that our model of the measurements is not quite accurate. When we do not restrict the scores to being integers, but any number, then these

differences are greatly diminished.[11] A more accurate model of our measurements instead accounts for all scores being in fact integers:

$$Y_{wn} = \max(5, \min(1, \lfloor \tau_w + Z_{wn} \rceil)). \tag{13}$$

However, this does not lend itself easily to the analysis above. Nonetheless, our simulations show our estimates are reasonably well-behaved, but that we may need to account for this discrepancy when we draw inferences.

3.5 Null hypothesis significance testing

We want to test hypotheses about the set of true scores, $\{\tau_1, \tau_2, \ldots, \tau_{|\mathcal{W}|}\}$, or equivalently $\{\beta_1, \ldots, \beta_{|\mathcal{W}|}\}$. In particular, we wish to test whether there are no differences between the scores of the wines. More formally, we wish to test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{|\mathcal{W}|}. \tag{14}$$

Figure 3 shows the beginnings of making such an inference for the data in Fig. 2. That these distributions overlap so little shows that we are extremely likely to correctly reject H_0 using only four scores for each wine as long as there is such a large difference between at least one pair of $\{\beta_1, \ldots, \beta_{|\mathcal{W}|}\}$. We would like to do this more formally, however.

Returning to the sum of the square deviations of our data, we can decompose it as a sum of the squares of our regression errors and the square deviations of our model from the data mean:[12]

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(y_{wn} - \hat{\beta}_0)^2 = \sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(y_{wn} - \hat{y}_w)^2 + \sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn} - \hat{\beta}_0)^2 \tag{15}$$

Notice that all of these terms are proportional to variances. The term on the left is proportional to the variance of our data from the "grand mean." The first one on the right is proportional to the variance of our data from the sample mean of each wine (called, "within-group variance"). And the last one is proportional to the variance of all our predicted data to the grand mean (called, "between-group variance"). The expectations of the terms on the right are:[13]

$$\begin{aligned} E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn} - \hat{\beta}_0 - \hat{\beta}_w)^2\right] &= \frac{(N-1)}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\mathrm{Var}(Z_{wn}) + E[Z_{wn}]^2) \\ &\quad - \sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\sum_{\substack{m=1\\ m\neq n}}^{N}(\mathrm{Cov}(Z_{wn}, Z_{wm}) + E[Z_{wn}]E[Z_{wm}]) \end{aligned} \tag{16}$$

$$\begin{aligned} E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn} - \hat{\beta}_0)^2\right] &= \sum_{w\in\mathcal{W}} N\beta_w^2 + 2\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\beta_w E[Z_{wn}] \\ &\quad + \frac{(|\mathcal{W}|-1)}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n,m=1}^{N}(\mathrm{Cov}(Z_{wm}, Z_{wn}) + E[Z_{wn}]E[Z_{wm}]) \\ &\quad - \frac{1}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}\setminus\{w\}}\sum_{n,m=1}^{N}(\mathrm{Cov}(Z_{wn}, Z_{vm}) + E[Z_{wn}]E[Z_{vm}]). \end{aligned} \tag{17}$$

[11] To see this, adapt the relevant MATLAB code given in the appendix by replacing the line `y(mm,:)= min(5,max(1,tau(mm)*ones(1,N)+ z(mm,:)));` with `y(mm,:)= tau(mm)*ones(1,N)+ z(mm,:).`

[12] See the appendix for derivation.

[13] See the appendix for derivation.

If for all wines and scores, Z_{wn} is iid with zero mean and variance σ^2, then the above become

$$E\left[\frac{1}{|\mathcal{W}|-1}\sum_{w\in\mathcal{W}} N\hat{\beta}_w^2\right] = \sigma^2 + \sum_{w\in\mathcal{W}} N\beta_w^2/(|\mathcal{W}|-1) \tag{18}$$

$$E\left[\frac{1}{(N-1)|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2\right] = \sigma^2. \tag{19}$$

If in addition H_0 is in effect, then $\beta_w = 0$, and we expect these two terms to be equal. Hence, we wish to compute our estimates of these quantities and see if

$$\frac{1}{|\mathcal{W}|-1}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2 \approx \frac{1}{(N-1)|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2. \tag{20}$$

More formally, with H_0 in effect and Z_{wn} iid with zero mean and variance σ^2, then[14]

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2/\sigma^2 \sim \chi^2_{(N-1)|\mathcal{W}|} \tag{21}$$

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2/\sigma^2 = \sum_{w\in\mathcal{W}} N\hat{\beta}_w^2/\sigma^2 \sim \chi^2_{|\mathcal{W}|-1} \tag{22}$$

and so

$$F := \frac{\sum_{w\in\mathcal{W}} N\hat{\beta}_w^2/(|\mathcal{W}|-1)}{\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2/((N-1)|\mathcal{W}|)} \sim F_{|\mathcal{W}|-1,(N-1)|\mathcal{W}|}. \tag{23}$$

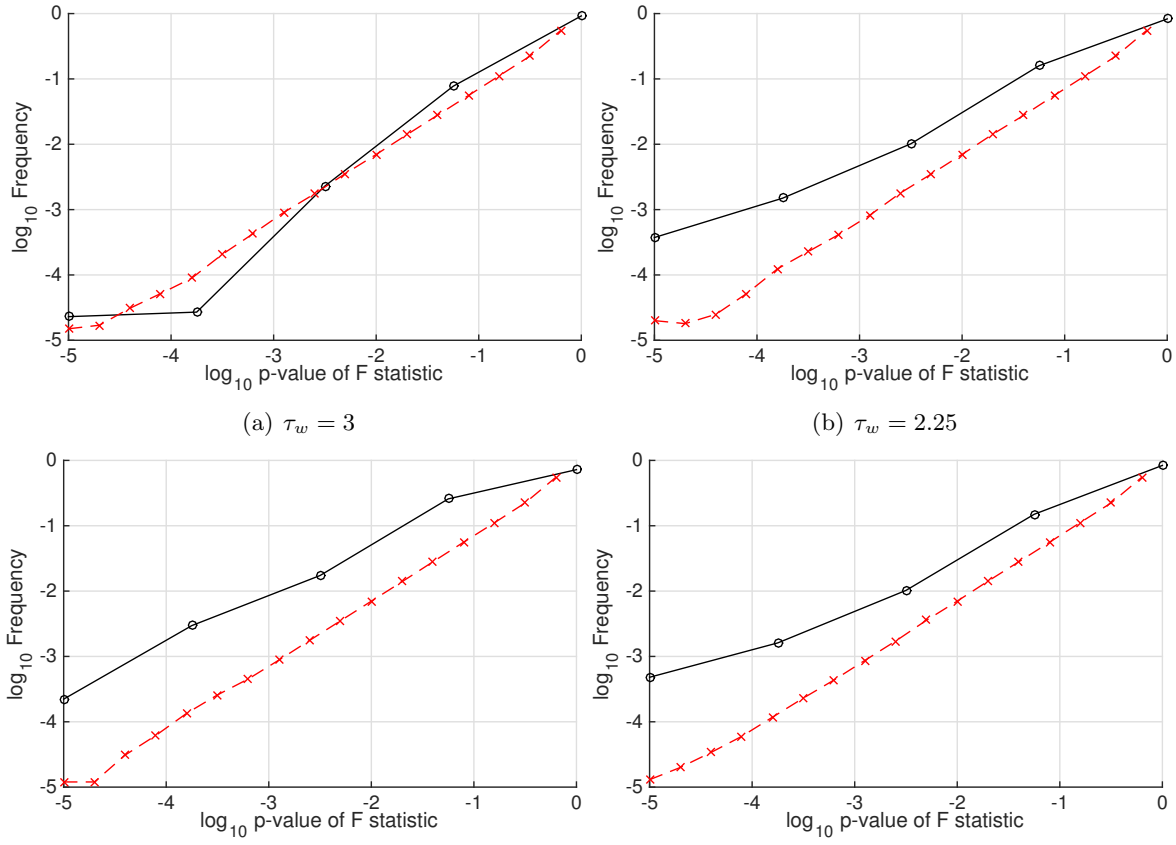
Hence, we compute the statistic f, and see whether the probability of achieving its value as or more extreme in an F-distribution with $|\mathcal{W}|-1$ and $(N-1)|\mathcal{W}|$ degrees of freedom exceeds our significance level α.

For the results shown in Fig. 2, the F-statistic is 20.36 with 3 degrees of freedom in the numerator and 12 degrees of freedom in the denominator. The probability of seeing a statistic at least that large given H_0 and Z_{wn} iid with zero mean and variance σ^2, is $p < 10^{-4}$. We are thus compelled to reject H_0. For the results in Table 1, the F-statistic is 9.06 and $p < 0.0021$. For a level of statistical significance of 0.05, we are thus also compelled to reject H_0 under limitations imposed by our measurement model and assumptions on Z_{wn}.

It is interesting to see whether there are major discrepancies in the result of hypothesis testing when using the measurement model that does not take into consideration that the responses are integers. Figure 4 compares the p-values observed for a resulting F-statistic from many simulations of 4 scores of 4 wines with H_0 in effect, and with Z_{wn} iid zero mean Gaussian with $\sigma^2 = 0.1$. We compare the results when we restrict measurements to be integers in $[1, 5]$, to those without such a restriction, for several true wine parameters. When the true parameter is an integer, the two appear to be quite in agreement. In other words, our α does reflect the probability of making a type 1 error.[15] When the true parameter is not an integer, we see that all p-values occur more frequently than we expect them to, and so our α underestimates the probability of making a type 1 error. In the worst case (τ_w between to integer scores), it seems our α should be about an order of magnitude smaller than what we specify, e.g., to ensure the probability of a type 1 error is only 0.05, then we should set $\alpha = 0.005$.

[14] This result is not easy to see, but becomes clearer when expressed in terms of projections in the measurement space [1].

[15] Rejecting H_0 when it is actually true.



(a) $\tau_w = 3$ (b) $\tau_w = 2.25$

(c) $\tau_w = 2.5$ (d) $\tau_w = 4.75$

Figure 4: Frequency of observing particular p-values of F statistic with H_0 in effect in simulations (1,000,000 trials) of 4 scores of 4 wines, and with Z_{wn} iid zero mean Gaussian with $\sigma^2 = 0.1$. Solid line is from simulations in which we restrict measurements to be integers in $[1, 5]$. Dash line is from simulations without restrictions. The results for other integer values of τ_w are highly similar.

3.6 A fatal problem reveals itself

The inflation of making a type 1 error is a problem, but it is not the fatal problem for our analysis. Upon consultation with the local chapter of oenophiles, it becomes apparent that the experiment was implemented in the following way: the person who poured the wine before judging actually poured the wine such that each judge scored the same wine four times. Hence, each row of results in Table 1 is from the same judge. In this case, Z_{wn} are no longer iid, and will be highly correlated for each w. The biggest problem however is the equivalence of the judge and wine factors. This in fact leaves us no way to separate them. With only one judge scoring one wine, we effectively have $N = 1$; and so if we plough ahead assuming Z_{wn} is iid with zero mean, then the within-group variance is distributed as the sum of *no* squared normal rvs – which does not make sense. In other words, the degrees of freedom of the within-group variance is zero. The experiment has made "false replications." We cannot say anything with regards to the wines alone because we cannot separate their scores from the judges. The implementation of the experiment does not permit a valid conclusion about the wines themselves, no matter what the scores in Table 1.

In conclusion, the validity of all statistical tests above critically relies on the distribution of the noise in the measurements, and the correct calculation of the degrees of freedom. If we cannot ensure the noise is distributed in a way that facilitates analysis,[16] then none of the tests above are statistically valid. The formal design of experiments (DOE) provides exactly that: *a formal methodology for designing and implementing an experiment such that the noise in the measurements is distributed in a way acceptable for reliably and validly testing hypotheses within one's cost constraints.*

[16]It is sufficient, but not necessary, that $\text{Cov}(Z_{wn}, Z_{vm}) = 0, n \neq m, w \neq v$ for analysis.

4 Just a brief peek at fundamental components of experimental design

Designing and implementing an experiment that is valid for a specific hypothesis entails performing several essential tasks [1]: identifying treatments, identifying plots, recognising structures in the treatments and plots, mapping plots to treatments, and specifying the measurement and its modelling. Below are definitions of the most important components. Table 4 shows examples of these for four different experiments. A future tutorial will look closely at these fundamental components, and the roles they play.

Definition 1. *Treatments* The set of things and their description applied to experimental units, $\mathcal{T} := \{i : i \in \{1, \ldots, t\}\}$.

Definition 2. *Experimental unit* The smallest unit to which a treatment is applied.

Definition 3. *Observational unit (plot)* The smallest unit on which a measurement is made.

Definition 4. *Response* The measurement made of an observational unit.

Definition 5. *Plots* The set of things mapped to treatments, $\Omega := \{\omega : \omega \in \{1, \ldots, N\}\}$.

Definition 6. *Experimental design (treatment factor)* A map $T : \Omega \to \mathcal{T}$.

Definition 7. *Plot structure* Meaningful ways (expert elicitation) of dividing up the plots. Possibilities are unstructured, blocks, etc.

Definition 8. *Treatment structure* Meaningful ways (expert elicitation) of dividing up the treatments. Possibilities are unstructured, treatment and control, etc.

Definition 9. *Plan* The translation of the experimental design into the actual plots.

Definition 10. *Response model* A mathematical relationship between the measurement (response) and the effect of a treatment. Possibilities include: simple textbook model, fixed or random effects, etc.

Treatments ($\mathcal{T}$)	*Experimental unit*	*Observational unit* ($\omega \in \Omega$)	*Treatment structure*	*Plot structure*	*Response*	*Response model*
Compost & water amount	tomato plant in a pot	tomato plant	unstructured	unstructured	tomato yield (grams)	simple textbook
New animal feed	pen	calf	new and old feeds	unstructured	weight (kilograms)	simple textbook
Local or remote learning	students in DOE 101 classroom-year	student	local, remote	majors (math, other)	test score (percentage)	fixed effects
Wines	judge	judge-tasting	none	judges	score $\{1, \ldots, 5\}$	simple textbook

Table 4: Examples of the various components for four different experiments.

Acknowledgments

Dr. Hugo Maruri-Aguilar, Dr. Ben Parker, and Dr. Heiko Grossmann.

References

[1] R. A. Bailey. *Design of comparative experiments.* Cambridge University Press, 2008.

A Appendix of derivations

A.1 Decomposition of the sum of squares (15)

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0)^2=\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0+\hat{Y}_w-\hat{Y}_w)^2=\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}((Y_{wn}-\hat{Y}_w)+(\hat{Y}_w-\hat{\beta}_0))^2$$
$$=\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{Y}_w)^2+\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{Y}_w-\hat{\beta}_0)^2+2\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{Y}_w)(\hat{Y}_w-\hat{\beta}_0). \quad (24)$$

Focusing on the last term, notice

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{Y}_w)(\hat{Y}_w-\hat{\beta}_0)=\sum_{w\in\mathcal{W}}\hat{\beta}_w\sum_{n=1}^{N}(Y_{wn}-\hat{Y}_w)=0 \quad (25)$$

because these estimators have the following properties:

$$\sum_{w\in\mathcal{W}}\hat{\beta}_w=0 \quad (26)$$

$$\sum_{n=1}^{N}\left(Y_{wn}-\hat{Y}_w\right)=0. \quad (27)$$

These properties result from the least squares, i.e., minimising the sum squared deviations from the mean.

A.2 Expectations of sum of squares

We first derive the expectation of the within-group variance, and then the between-group variance. Define $\mathbf{Y}_w$ to be a vector those N measurements with wine w, and $\mathbf{Z}_w$ a vector of the noise in those measurements. Let $\mathbf{1}$ be a vector of ones appropriately sized.

$$E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2\right]=E\left[\sum_{w\in\mathcal{W}}\left\|\mathbf{Y}_w-\mathbf{1}\hat{\beta}_0-\mathbf{1}(\hat{\tau}_w-\hat{\beta}_0)\right\|^2\right]=\sum_{w\in\mathcal{W}}E\left[\|\mathbf{Y}_w-\mathbf{1}\hat{\tau}_w\|^2\right] \quad (28)$$
$$=\sum_{w\in\mathcal{W}}E\left[\left\|\mathbf{Y}_w-\mathbf{1}\frac{1}{N}\mathbf{1}^T\mathbf{Y}_w\right\|^2\right]=\sum_{w\in\mathcal{W}}E\left[\left\|\left(\mathbf{I}-\mathbf{1}\frac{1}{N}\mathbf{1}^T\right)\mathbf{Y}_w\right\|^2\right] \quad (29)$$
$$=\sum_{w\in\mathcal{W}}E\left[\left\|\left(\mathbf{I}-\mathbf{1}\frac{1}{N}\mathbf{1}^T\right)(\mathbf{1}\tau_w+\mathbf{Z}_w)\right\|^2\right] \quad (30)$$
$$=\sum_{w\in\mathcal{W}}E\left[\left\|\mathbf{1}\tau_w-\mathbf{1}\tau_w+\left(\mathbf{I}-\mathbf{1}\frac{1}{N}\mathbf{1}^T\right)\mathbf{Z}_w\right\|^2\right] \quad (31)$$
$$=\sum_{w\in\mathcal{W}}E\left[\left\|\mathbf{Z}_w-\frac{1}{N}\mathbf{1}^T\mathbf{Z}_w\mathbf{1}\right\|^2\right]. \quad (32)$$

Define $\mathbf{u}_w$ to be a vector of length $N|\mathcal{W}|$ with ones in the rows of $\mathbf{Y}$ corresponding to measurements related to wine w, and zeros in all others. We expand the norm, and move to all noise measurements $\mathbf{Z}$.

$$E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2\right]=E[\|\mathbf{Z}\|^2]+\frac{1}{N}\sum_{w\in\mathcal{W}}\mathbf{u}_w^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_w-\frac{2}{N}\sum_{w\in\mathcal{W}}\mathbf{u}_w^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_w \quad (33)$$
$$=E[\|\mathbf{Z}\|^2]-\frac{1}{N}\sum_{w\in\mathcal{W}}\mathbf{u}_w^TE\left[\mathbf{Z}\mathbf{Z}^T\right]\mathbf{u}_w. \quad (34)$$

Converting these to sum notation

$$E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2\right]=\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}E[Z_{wn}^2]-\frac{1}{N}\sum_{w\in\mathcal{W}}\sum_{n,m=1}^{N}E[Z_{wn}Z_{wm}] \tag{35}$$

$$=\frac{1}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}NE[Z_{wn}Z_{wn}]-\frac{1}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\sum_{m=1}^{N}E[Z_{wn}Z_{wm}] \tag{36}$$

$$=\frac{1}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\left(NE[Z_{wn}Z_{wn}]-\sum_{m=1}^{N}E[Z_{wn}Z_{wm}]\right) \tag{37}$$

$$=\frac{1}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\left((N-1)E[Z_{wn}Z_{wn}]-\sum_{\substack{m=1\\m\neq n}}^{N}E[Z_{wn}Z_{wm}]\right) \tag{38}$$

$$=\frac{(N-1)}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}E[Z_{wn}Z_{wn}]-\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\sum_{\substack{m=1\\m\neq n}}^{N}E[Z_{wn}Z_{wm}] \tag{39}$$

$$=\frac{(N-1)}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\mathrm{Var}(Z_{wn})+E[Z_{wn}]^2)$$

$$-\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\sum_{\substack{m=1\\m\neq n}}^{N}(\mathrm{Cov}(Z_{wn},Z_{wm})+E[Z_{wn}]E[Z_{wm}]). \tag{40}$$

For the between-group variance, we employ the same definitions above:

$$E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2\right]=E\left[\sum_{w\in\mathcal{W}}N\hat{\beta}_w^2\right]=E\left[\sum_{w\in\mathcal{W}}N\left(\frac{1}{N}\mathbf{u}_w^T\mathbf{Y}-\bar{Y}\right)^2\right] \tag{41}$$

$$=E\left[\sum_{w\in\mathcal{W}}\left(\frac{1}{N}(\mathbf{u}_w^T\mathbf{Y})^2+N\bar{Y}^2-2\bar{Y}\mathbf{u}_w^T\mathbf{Y}\right)\right] \tag{42}$$

$$=\sum_{w\in\mathcal{W}}\left[\frac{1}{N}\mathbf{u}_w^TE[\mathbf{Y}\mathbf{Y}^T]\mathbf{u}_w+\frac{1}{N|\mathcal{W}|^2}\mathbf{u}_0^TE[\mathbf{Y}\mathbf{Y}^T]\mathbf{u}_0-\frac{2}{N|\mathcal{W}|}\mathbf{u}_0^TE[\mathbf{Y}\mathbf{Y}^T]\mathbf{u}_w\right]. \tag{43}$$

Since $\mathbf{Y}\mathbf{Y}^T=(\boldsymbol{\tau}+\mathbf{Z})(\boldsymbol{\tau}+\mathbf{Z})^T=\boldsymbol{\tau}\boldsymbol{\tau}^T+\mathbf{Z}\mathbf{Z}^T+\boldsymbol{\tau}\mathbf{Z}^T+\mathbf{Z}\boldsymbol{\tau}^T$ where $\boldsymbol{\tau}:=\tau_1\mathbf{w}_1+\ldots+\tau_{|\mathcal{W}|}\mathbf{w}_{|\mathcal{W}|}$, then

$$\begin{aligned}E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2\right]=&\sum_{w\in\mathcal{W}}\left[\frac{1}{N}\mathbf{u}_w^T\boldsymbol{\tau}\boldsymbol{\tau}^T\mathbf{u}_w+\frac{1}{N|\mathcal{W}|^2}\mathbf{u}_0^T\boldsymbol{\tau}\boldsymbol{\tau}^T\mathbf{u}_0-\frac{2}{N|\mathcal{W}|}\mathbf{u}_0^T\boldsymbol{\tau}\boldsymbol{\tau}^T\mathbf{u}_w\right]\\&+\sum_{w\in\mathcal{W}}\left[\frac{1}{N}\mathbf{u}_w^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_w+\frac{1}{N|\mathcal{W}|^2}\mathbf{u}_0^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_0-\frac{2}{N|\mathcal{W}|}\mathbf{u}_0^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_w\right]\\&+\sum_{w\in\mathcal{W}}\left[\frac{1}{N}\mathbf{u}_w^TE[\mathbf{Z}]\boldsymbol{\tau}^T\mathbf{u}_w+\frac{1}{N|\mathcal{W}|^2}\mathbf{u}_0^TE[\mathbf{Z}]\boldsymbol{\tau}^T\mathbf{u}_0-\frac{2}{N|\mathcal{W}|}\mathbf{u}_0^TE[\mathbf{Z}]\boldsymbol{\tau}^T\mathbf{u}_w\right]\\&+\sum_{w\in\mathcal{W}}\left[\frac{1}{N}\mathbf{u}_w^T\boldsymbol{\tau}E[\mathbf{Z}^T]\mathbf{u}_w+\frac{1}{N|\mathcal{W}|^2}\mathbf{u}_0^T\boldsymbol{\tau}E[\mathbf{Z}^T]\mathbf{u}_0-\frac{2}{N|\mathcal{W}|}\mathbf{u}_0^T\boldsymbol{\tau}E[\mathbf{Z}^T]\mathbf{u}_w\right].\end{aligned} \tag{44}$$

Collecting terms and gradually converting to sum notation:

$$E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2\right] = \sum_{w\in\mathcal{W}} N\left[\tau_w^2+\bar{\tau}^2-2\bar{\tau}\tau_w\right] + \sum_{w\in\mathcal{W}}\left[\tau_w\mathbf{u}_w^TE[\mathbf{Z}]+\frac{1}{|\mathcal{W}|}\bar{\tau}\mathbf{u}_0^TE[\mathbf{Z}]-2\bar{\tau}\mathbf{u}_w^TE[\mathbf{Z}]\right]$$
$$+\sum_{w\in\mathcal{W}}\left[\tau_w\mathbf{u}_w^TE[\mathbf{Z}]+\frac{1}{|\mathcal{W}|}\bar{\tau}\mathbf{u}_0^TE[\mathbf{Z}]-\frac{2}{|\mathcal{W}|}\tau_w\mathbf{u}_0^TE[\mathbf{Z}]\right]$$
$$-\frac{1}{N|\mathcal{W}|}\mathbf{u}_0^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_0+\sum_{w\in\mathcal{W}}\frac{1}{N}\mathbf{u}_w^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_w \tag{45}$$

$$=\sum_{w\in\mathcal{W}}N(\tau_w-\bar{\tau})^2+2\sum_{w\in\mathcal{W}}\tau_w\mathbf{u}_w^TE[\mathbf{Z}]+2\bar{\tau}\mathbf{u}_0^TE[\mathbf{Z}]-2\bar{\tau}\mathbf{u}_0^TE[\mathbf{Z}]$$
$$-2\bar{\tau}\mathbf{u}_0^TE[\mathbf{Z}]-\frac{1}{N|\mathcal{W}|}\mathbf{u}_0^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_0+\sum_{w\in\mathcal{W}}\frac{1}{N}\mathbf{u}_w^TE[\mathbf{Z}\mathbf{Z}^T]\mathbf{u}_w \tag{46}$$

$$=\sum_{w\in\mathcal{W}}N\beta_w^2+2\sum_{w\in\mathcal{W}}(\tau_w-\bar{\tau})\mathbf{u}_w^TE[\mathbf{Z}]-\frac{1}{N|\mathcal{W}|}\sum_{w,v\in\mathcal{W}}\sum_{n,m=1}^{N}E[Z_{wn}Z_{vm}]$$
$$+\frac{1}{N}\sum_{w\in\mathcal{W}}\sum_{n,m=1}^{N}E[Z_{wn}Z_{wm}] \tag{47}$$

$$=\sum_{w\in\mathcal{W}}N\beta_w^2+2\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\beta_wE[Z_{wn}]$$
$$+\frac{1}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n,m=1}^{N}\left(|\mathcal{W}|E[Z_{wn}Z_{wm}]-\sum_{v\in\mathcal{W}}E[Z_{wn}Z_{vm}]\right) \tag{48}$$

$$=\sum_{w\in\mathcal{W}}N\beta_w^2+2\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\beta_wE[Z_{wn}]+\frac{(|\mathcal{W}|-1)}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n,m=1}^{N}E[Z_{wn}Z_{wm}]$$
$$-\frac{1}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}\setminus\{w\}}\sum_{n,m=1}^{N}E[Z_{wn}Z_{vm}] \tag{49}$$

Finally, we see that

$$E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2\right] = \sum_{w\in\mathcal{W}}N\beta_w^2+2\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\beta_wE[Z_{wn}]$$
$$+\frac{(|\mathcal{W}|-1)}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n,m=1}^{N}(\mathrm{Cov}(Z_{wm},Z_{wn})+E[Z_{wn}]E[Z_{wm}])$$
$$-\frac{1}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}\setminus\{w\}}\sum_{n,m=1}^{N}(\mathrm{Cov}(Z_{wn},Z_{vm})+E[Z_{wn}]E[Z_{vm}]). \tag{50}$$

B Appendix of MATLAB code for figures

B.1 Figure 2

Due to randomness, the figure produced below will likely not match Figure 2.

```
M = 4; % number of wines
N = 4; % number of observations of each wine
sigma2 = 0.1; % noise variance

% set true tau_w, random number in [1,5]
tau = 1+4*rand(M,1);
% compute true mean, beta_0
beta0 = mean(tau);
% compute true parameters beta_w
betaw = tau - beta0;

% create noise data for each observation of each wine
z = sqrt(sigma2)*randn(M,N);
% synthesize observations
y = zeros(M,N);
for mm=1:M
  y(mm,:) = max(1,min(5,round(tau(mm)*ones(1,N) + z(mm,:))));
end

% plot dataset (with random x-offset for visibility)
figure; hold on; grid on;
for mm=1:M
  plot(mm+0.01*randn(1,N),y(mm,:),'ko','MarkerSize',10);
end
axis([0.5 M+0.5 0.9 5.1]);
set(gca,'XTick',[1:M],'YTick',[1:5],'FontSize',16);
ylabel('Score'); xlabel('Wine');

% plot true data mean and our estimate of it
hh = line([0.5 M+0.5],beta0*ones(2,1));
set(hh,'Color','r','LineStyle','--');
beta0hat = mean(y(:));
hh = line([0.5 M+0.5],beta0hat*ones(2,1));
set(hh,'Color','k');

% plot true wine parameters and our estimates of them
betawhat = mean(y,2)-beta0hat*ones(M,1);
for mm=1:M
  hh = line(mm+[-0.1 0.1],(beta0hat+betawhat(mm))*ones(2,1));
  set(hh,'Color','k','LineStyle','-');
  hh = line(mm+[-0.1 0.1],(beta0+betaw(mm))*ones(2,1));
  set(hh,'Color','r','LineStyle','--');
end
```

B.2 Figure 3

Due to randomness, the figures produced below will likely not match those in Figure 3.

```
M = 4; % number of wines
N = 4; % number of observations of each wine
sigma2 = 0.1; % noise variance
numsimulations = 1000000; % number of simulations

% set true tau_w, random number in [1,5]
tau = 1+4*rand(M,1);
% compute true mean, beta_0
beta0 = mean(tau);
% compute true parameters beta_w
betaw = tau - beta0;

% create place to store estimates
beta0hat = zeros(numsimulations,1);
betawhat = zeros(numsimulations,M);
% run simulations
for ii=1:numsimulations
  z = sqrt(sigma2)*randn(M,N); % synthesise noise
  y = zeros(M,N); % synthesise measurements
  for mm=1:M
    y(mm,:) = min(5,max(1,tau(mm)*ones(1,N) + z(mm,:)));
  end
  beta0hat(ii) = mean(y(:)); % estimate beta_0
  betawhat(ii,:) = mean(y,2)-beta0hat(ii)*ones(M,1); % estimate beta_w
end

%% plot distribution of beta_0 estimates (magic numbers to make histogram look nice)
figure; hold on; grid on;
[Num,X] = hist(beta0hat,[1:0.05:5]);
prop = Num./sum(Num);
newlims = prop*X'*[0.8 1.2];
[Num,X] = hist(beta0hat,linspace(min(newlims),max(newlims),21));
prop = Num./sum(Num);
hh = bar(X,prop);
set(hh,'FaceColor',zeros(3,1));
hh = line(beta0*ones(2,1),[0 1]);
set(hh,'Color','r','LineStyle','--');
axis([1 5 0 1.1*max(prop)]);
ylabel('Frequency'); xlabel('Data Mean');

%% plot distributions of beta_w estimates (magic numbers to make histogram look nice)
figure; hold on; grid on;
for mm=1:M
  [Num,X] = hist(betawhat(:,mm), ...
    linspace(min(betawhat(:,mm))-0.2,max(betawhat(:,mm))+0.2,37));
  prop = Num./sum(Num);
  hh = bar(X,prop);
  set(hh,'FaceColor',rand*ones(3,1));
  hh = line(betaw(mm)*ones(2,1),[0 1]);
  set(hh,'Color','r','LineStyle','--');
end
axis([-3 3 0 1.5*max(prop)]);
ylabel('Frequency'); xlabel('Wine Parameters');
```

B.3 Figure 4

Due to randomness, the figures produced below will likely not match those in Figure 4.

```
M = 4; % number of wines
N = 4; % number of observations of each wine
sigma2 = 0.1; % noise variance
numsimulations = 1000000; % number of simulations

% set true tau_w
tau = 2.5*ones(M,1);
% compute true mean, beta_0
beta0 = mean(tau);
% compute true parameters beta_w
betaw = tau - beta0;

% places to store F statistics
Fstat = zeros(numsimulations,1);
Fstat_norestrictions = zeros(numsimulations,1);

% run simulations
for ii=1:numsimulations
  z = sqrt(sigma2)*randn(M,N); % compute noise
  % synthesise measurements
  y = zeros(M,N); y_norestrictions = zeros(M,N);
  for mm=1:M
    y(mm,:) = max(1,min(5,round(tau(mm)*ones(1,N) + z(mm,:))));
    y_norestrictions(mm,:) = tau(mm)*ones(1,N) + z(mm,:);
  end
  % estimate beta_0 parameter
  beta0hat = mean(y(:));
  % estimate beta_w parameters
  betawhat = mean(y,2)-beta0hat*ones(M,1);
  % compute sum of squares for residual and parameters
  sumofsquares_residual = ...
    norm((y-beta0hat)-repmat(betawhat,1,N),'fro')^2;
  sumofsquares_betaw = betawhat'*betawhat;
  % compute F statistic
  Fstat(ii) = (sumofsquares_betaw/(M-1))/(sumofsquares_residual/((N-1)*M*N));

  % estimate beta_0 parameter with no restrictions on measurements
  beta0hat = mean(y_norestrictions(:));
  % estimate beta_w parameters with no restrictions on measurements
  betawhat = mean(y_norestrictions,2)-beta0hat*ones(M,1);
  % compute sum of squares for residual and parameters with no restrictions on measurements
  sumofsquares_residual = ...
    norm((y_norestrictions-beta0hat)-repmat(betawhat,1,N),'fro')^2;
  sumofsquares_betaw = betawhat'*betawhat;
  % compute F statistic with no restrictions on measurements
  Fstat_norestrictions(ii) = (sumofsquares_betaw/(M-1))/(sumofsquares_residual/((N-1)*M*N));
end

% remove NaN
Fstat(isnan(Fstat)) = 1;
Fstat_norestrictions(isnan(Fstat_norestrictions)) = 1;

% plot results
figure; hold on; grid on;
p = cdf('F',Fstat,M-1,N*(M-1),'upper');
[Num,X] = hist(log10(p),[-5:1.25:0]);
prop = Num./sum(Num);
plot(X,log10(prop),'ko-');
p = cdf('F',Fstat_norestrictions,M-1,N*(M-1),'upper');
[Num,X] = hist(log10(p),[-5:0.3:0]);
prop = Num./sum(Num);
plot(X,log10(prop),'rx--');
axis([-5 0 -5 0]);
ylabel('log_{10} Frequency'); xlabel('log_{10} p-value of F statistic');
set(gca,'YTick',[-5:1:0],'XTick',[-5:1:0]);
```