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

Bob L. Sturm

Centre for Digital Music
School of Electronic Engineering and Computer Science, QMUL
Mile End Road, London E1 4NS

December 15, 2015

Setup

A request from our local chapter of oenophiles

```
Hello.  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

Some descriptive statistics

Let's begin by computing some descriptive statistics ($N = 4$).

- Mean score ($\hat{\mu}_w$)

$$\hat{\mu}_w = \frac{1}{N} \sum_{n=1}^{N} y_{wn} \tag{1}$$

wine (w)	scores	$\hat{\mu}_w$
1	3 4 3 2	3.00
2	5 4 5 5	4.75
3	2 1 3 1	1.75
4	2 3 2 4	2.75

Some descriptive statistics

Let's begin by computing some descriptive statistics ($N = 4$).

- (unbiased) standard deviation of score ($\hat{\sigma}_w$)

$$\hat{\sigma}_w = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(y_{wn} - \hat{\mu}_w)^2} \tag{2}$$

wine (w)	scores	$\hat{\mu}_w$	$\hat{\sigma}_w$
1	3 4 3 2	3.00	0.82
2	5 4 5 5	4.75	0.50
3	2 1 3 1	1.75	0.96
4	2 3 2 4	2.75	0.96

Some descriptive statistics

Let's begin by computing some descriptive statistics ($N = 4$).

- Standard error of the mean (SEM)

$$\text{SEM}_w = \frac{\hat{\sigma}_w}{\sqrt{N}} \tag{3}$$

wine (w)	scores	$\hat{\mu}_w$	$\hat{\sigma}_w$	SEM_w
1	3 4 3 2	3.00	0.82	0.41
2	5 4 5 5	4.75	0.50	0.25
3	2 1 3 1	1.75	0.96	0.48
4	2 3 2 4	2.75	0.96	0.48

Some descriptive statistics

Let's begin by computing some descriptive statistics ($N = 4$).

- 95% confidence interval (CI)

$$[\hat{\mu}_w - 2.365 \cdot \mathrm{SEM}_w,\ \hat{\mu}_w + 2.365 \cdot \mathrm{SEM}_w] \tag{4}$$

using t-distribution having 3 degrees of freedom.

wine (w)	scores	$\hat{\mu}_w$	$\hat{\sigma}_w$	SEM_w	CI_w (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]

Some descriptive statistics

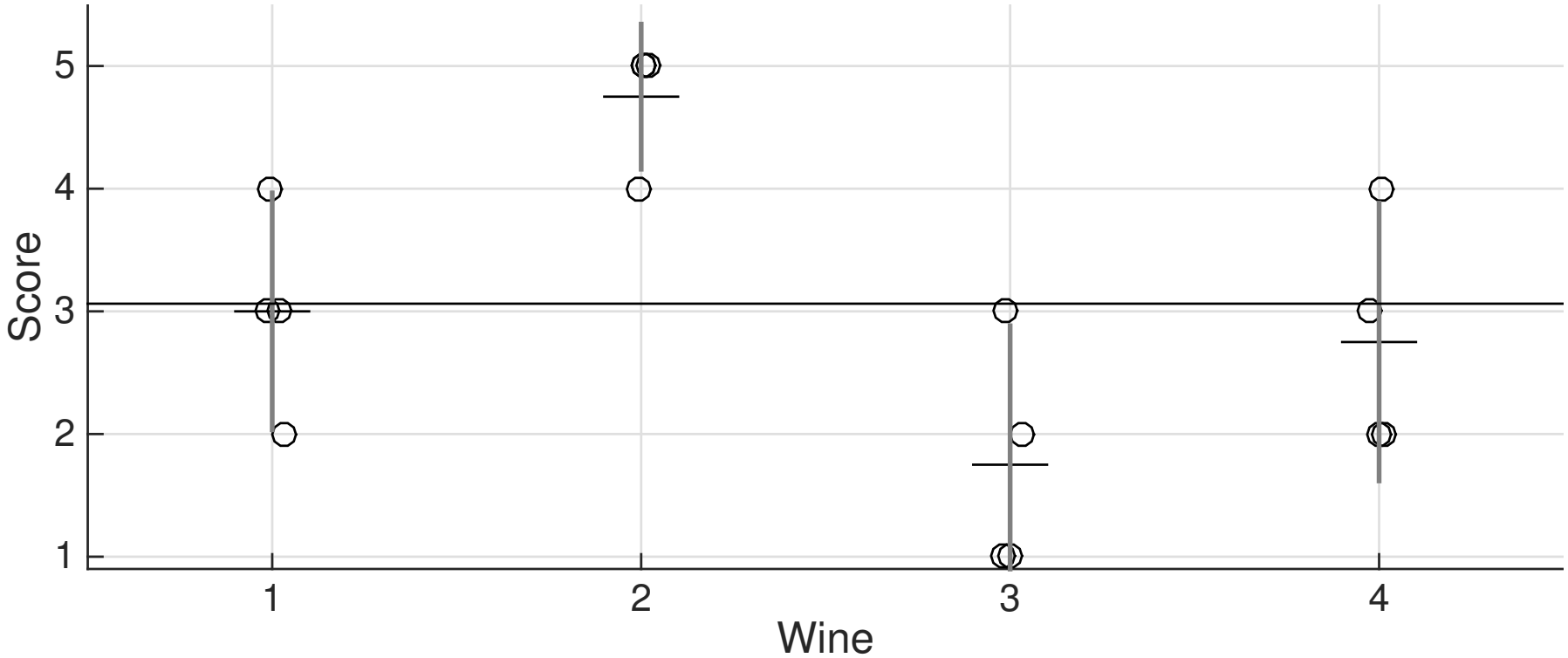


Figure: Wine scores (circles) (with random x-offset for visibility), data mean (horizontal line), wine means (short line segments), and 95% confidence intervals.

Some descriptive statistics

Let's run the data in something called "ANOVA":

```
dataset = [3 4 3 2; 5 4 5 5; 2 1 3 1; 2 3 2 4];
[p,tab,stats] = anova1(dataset')
[c,m,h,nms] = multcompare(stats)
```

This produces a p-value of 0.0021, motivating rejecting the null hypothesis that there are no differences if our α is larger, e.g., $\alpha = 0.05$.

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: Pairwise comparisons of scores using something called "ANOVA".

What have we done?

- We've input the data, and computed some basic statistics.
- We've looked at the confidence intervals of the mean scores.
- We've run something called "ANOVA," which suggests there is at least one pair of mean scores that are significantly different ($p < \alpha = 0.05$).
- Pairwise comparisons using ANOVA suggests wine 2 has a significantly higher mean score than the others (also seen in the CI).

wine (w)	scores	$\hat{\mu}_w$	$\hat{\sigma}_w$	SEM_w	CI_w (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]

What have we *not* done?

We've not considered the most important question:

What can we validly conclude from this table?

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

I will show why we cannot validly conclude anything with respect to wine quality.

Principal aim of the experiment

Our analysis hinges upon the principal aim of the experiment of our local chapter of oenophiles, and the way they ran it.

- Do they want to rank the "wine quality" of these wines according to only these four particular judges?
- Do they want to rank the "wine quality" of these wines according to The Wine Judging Population (TWJP), inferred from these four particular judges?

We will consider the aim of the experiment as to determine the "best" and "worst" wines according to these particular judges.

What do they mean "best" and "worst"?

Whatever "wine quality" is, one might not be able to directly measure and compare it; our local chapter of oenophiles measured "wine quality" scores.

Definitions

- The "best" wine in some set of wines is that having a significantly higher mean score than all the others.
- The "worst" wine in some set of wines is that having 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. (A wine can be stellar if paired well, or horrible if not.)

Measurement modelling

Consider each score an outcome mapped to $\{1, 2, 3, 4, 5\}$ by a random variable Y_{wn}, where n denotes the tasting, and $w \in \mathcal{W}$ the wine:

$$Y_{wn} : \{\text{“1”}, \text{“2”}, \text{“3”}, \text{“4”}, \text{“5”}\} \to \{1, 2, 3, 4, 5\}. \tag{5}$$

We model Y_{wn} as a sum of the “true” (deterministic) score of wine w (τ_w), perturbed by some “noise”:

$$Y_{wn} = \tau_w + Z_{wn}. \tag{6}$$

Z_{wn} is random, and captures contributions unrelated to wine, e.g., randomness of scoring, judge experience, particulars of the experiment.

Given a collection of measurements $\{y_{nw}\}$, we wish to estimate the parameters $\{\tau_w : w \in \mathcal{W}\}$, and compare them in statistically valid ways.

Measurement modelling

Consider each score an outcome mapped to $\{1, 2, 3, 4, 5\}$ by a random variable Y_{wn}, where n denotes the tasting, and $w \in \mathcal{W}$ the wine:

$$Y_{wn} : \{\text{“1”}, \text{“2”}, \text{“3”}, \text{“4”}, \text{“5”}\} \to \{1, 2, 3, 4, 5\}.$$

We model Y_{wn} as a sum of the “true” (deterministic) score of wine w (τ_w) perturbed by some “noise”:

$$Y_{wn} = \bar{\tau} + (\tau_w - \bar{\tau}) + Z_{wn} \tag{7}$$

$\bar{\tau}$ 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 $\mathcal{W}$.

Given a collection of measurements $\{y_{nw}\}$, we wish to estimate the parameters $\{\tau_w - \bar{\tau} : w \in \mathcal{W}\}$, and compare them in statistically valid ways.

Measurement modelling (version *regression*)

Consider each score an outcome mapped to $\{1, 2, 3, 4, 5\}$ by a random variable Y_{wn}, where n denotes the tasting, and $w \in \mathcal{W}$ the wine:

$$Y_{wn} : \{\text{“1”}, \text{“2”}, \text{“3”}, \text{“4”}, \text{“5”}\} \to \{1, 2, 3, 4, 5\}.$$

We model Y_{wn} as the linear regression:

$$Y_{wn} = \beta_0 + \beta_1 \delta_{w-1} + \beta_2 \delta_{w-2} + \ldots + \beta_{|\mathcal{W}|} \delta_{w-|\mathcal{W}|} + Z_{wn} \tag{8}$$

where $\beta_0 := \bar{\tau}$, and $\beta_w := \tau_w - \bar{\tau}$, and $\delta_k = 1$ if $k = 0$ and zero otherwise.

Given a collection of measurements $\{y_{nw}\}$, we wish to estimate the parameters $\{\beta_w : w \in \mathcal{W}\}$, and compare them in statistically valid ways.

Measurement modelling

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

$$Y_{wn} = \tau_w + Z_{wn}$$
$$Y_{wn} = \bar{\tau} + (\tau_w - \bar{\tau}) + Z_{wn}$$
$$Y_{wn} = \beta_0 + \beta_1 \delta_{w-1} + \beta_2 \delta_{w-2} + \ldots + \beta_{|\mathcal{W}|} \delta_{w-|\mathcal{W}|} + Z_{wn}$$

We would like to test the null hypothesis:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_{|\mathcal{W}|}$$
$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{|\mathcal{W}|}. \quad (9)$$

Parameter estimation

Given N measurements of each wine in $\mathcal{W}$, we estimate the parameters $\{\beta_0, \beta_w : w \in \mathcal{W}\}$ by using the method of least squares:

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{w \in \mathcal{W}} \sum_{n=1}^{N} (y_{wn} - \hat{\beta}_0)^2 = 0 \Longrightarrow \hat{\beta}_0 = \frac{1}{|\mathcal{W}|} \frac{1}{N} \sum_{w \in \mathcal{W}} \sum_{n=1}^{N} y_{wn}. \tag{10}$$

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

$$\Longrightarrow \hat{\beta}_w = \frac{1}{N} \sum_{n=1}^{N} (y_{wn} - \hat{\beta}_0). \tag{11}$$

How do these estimators behave?

The expectations of these estimators are:

$$\mathrm{E}[\hat{\beta}_0] = \beta_0 + \frac{1}{|\mathcal{W}|}\frac{1}{N}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N} E[Z_{wn}] \tag{12}$$

$$\mathrm{E}[\hat{\beta}_w] = \beta_w + \frac{1}{|\mathcal{W}|}\frac{1}{N}\left(\sum_{w'\in\mathcal{W}\setminus\{w\}}\sum_{n=1}^{N} E[Z_{w'n}] - (|\mathcal{W}|-1)\sum_{n'=1}^{N} E[Z_{wn'}]\right) \tag{13}$$

If Z_{wn} is distributed zero mean, then these are *unbiased*.

How do these estimators behave?

The variances of these estimators are:

$$\mathrm{Var}[\hat{\beta}_0] = \frac{1}{|\mathcal{W}|^2}\frac{1}{N^2}\mathrm{Var}\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N} Z_{wn}\right] \tag{14}$$

$$\mathrm{Var}[\hat{\beta}_w] = \frac{1}{|\mathcal{W}|^2}\frac{1}{N^2}\mathrm{Var}\left[\sum_{w'\in\mathcal{W}\setminus\{w\}}\sum_{n=1}^{N} Z_{w'n} - (|\mathcal{W}|-1)\sum_{n'=1}^{N} Z_{wn'}\right]. \tag{15}$$

If Z_{wn} is independently and identically distributed (iid) with variance σ^2, then the above become

$$\mathrm{Var}[\hat{\beta}_0] = \sigma^2/(|\mathcal{W}|N) \tag{16}$$

$$\mathrm{Var}[\hat{\beta}_w] = (|\mathcal{W}|-1)\sigma^2/(|\mathcal{W}|N) = (|\mathcal{W}|-1)\mathrm{Var}[\hat{\beta}_0]. \tag{17}$$

What are we doing again?

- We have several scores $\{y_{wn}\}$.
- We define an rv: $Y_{wn} : \{\text{“1”}, \text{“2”}, \text{“3”}, \text{“4”}, \text{“5”}\} \to \{1, 2, 3, 4, 5\}$.
- We model Y_{wn} by a linear regression:

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

- We estimate $\{\beta_0, \beta_w : w \in \mathcal{W}\}$ by the principle of least squares.
- Because we want to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{|\mathcal{W}|}$.

Let’s run some simulations!

Parameter estimation

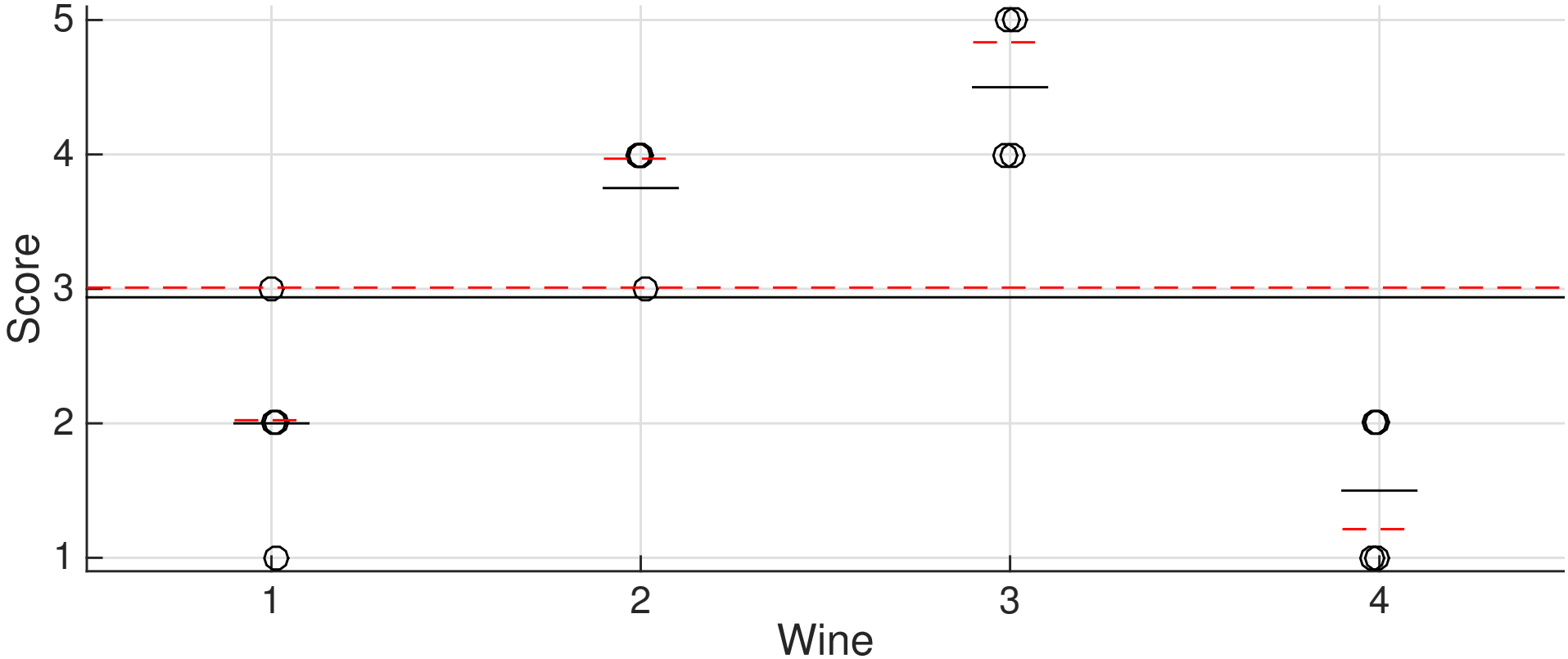


Figure: Simulation of 4 scores (circles, with random x-offset for visibility) of 4 wines, with Z_{wn} iid zero-mean Gaussian with $\sigma^2 = 0.1$. Data mean ($\hat{\beta}_0$) is black horizontal line, wine mean scores ($\hat{\tau}_w$) are short line segments. True mean (β_0) is red dashed line, and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

Parameter estimation

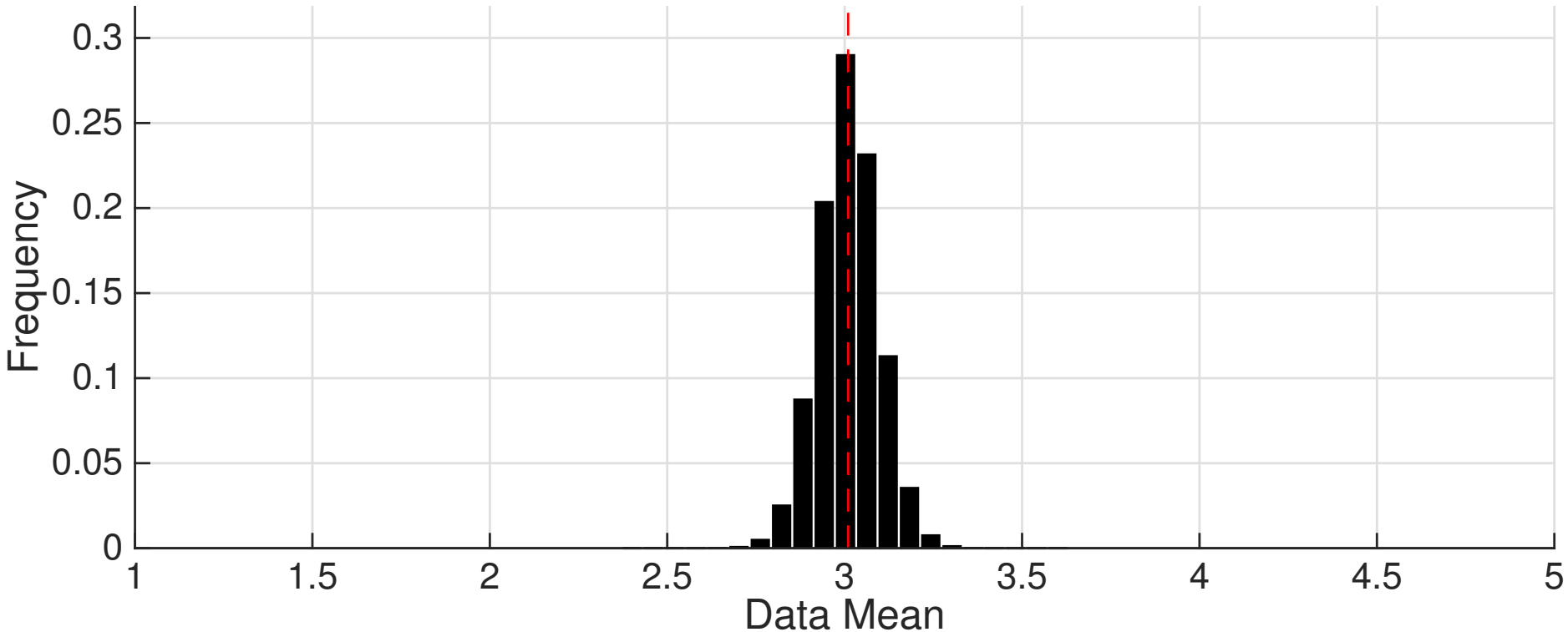


Figure: Distribution of $\hat{\beta}_0$ for a simulation (1,000,000 trials) of 4 scores of 4 wines, with Z_{wn} iid zero-mean Gaussian with $\sigma^2 = 0.1$. True parameter value is shown as red dashed line.

Parameter estimation

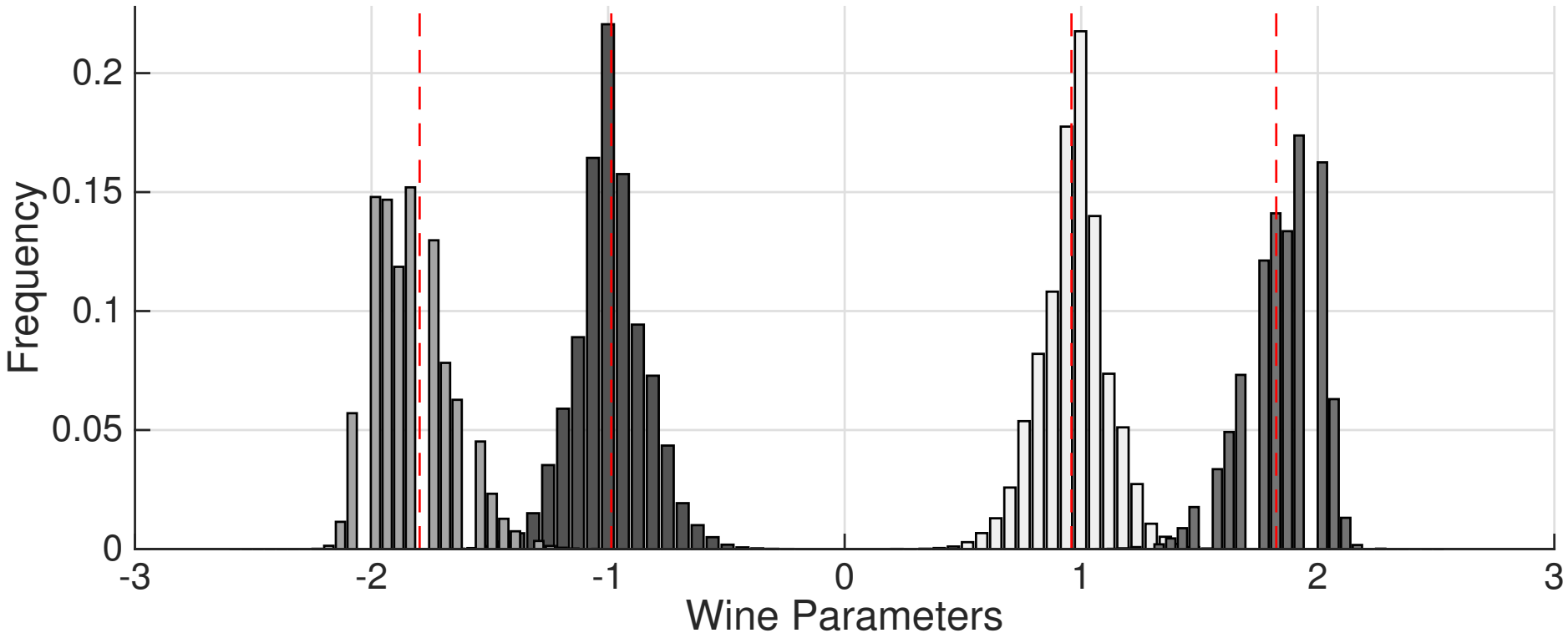


Figure: Distributions of $\{\beta_w : w \in \mathcal{W}\}$ for a simulation (1,000,000 trials) of 4 scores of 4 wines, with Z_{wn} iid zero-mean Gaussian with $\sigma^2 = 0.1$. True parameter values are shown as red dashed lines.

What have we learned?

- We have several scores $\{y_{wn}\}$.
- We define an rv: $Y_{wn} : \{\text{“1”}, \text{“2”}, \text{“3”}, \text{“4”}, \text{“5”}\} \to \{1, 2, 3, 4, 5\}$.
- We model Y_{wn} by a linear regression:

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

- Our least-squares estimates $\{\hat{\beta}_0, \hat{\beta}_w : w \in \mathcal{W}\}$ are reasonably distributed, except there does exist bias when Z_{wn} is zero mean.
- A more realistic model is

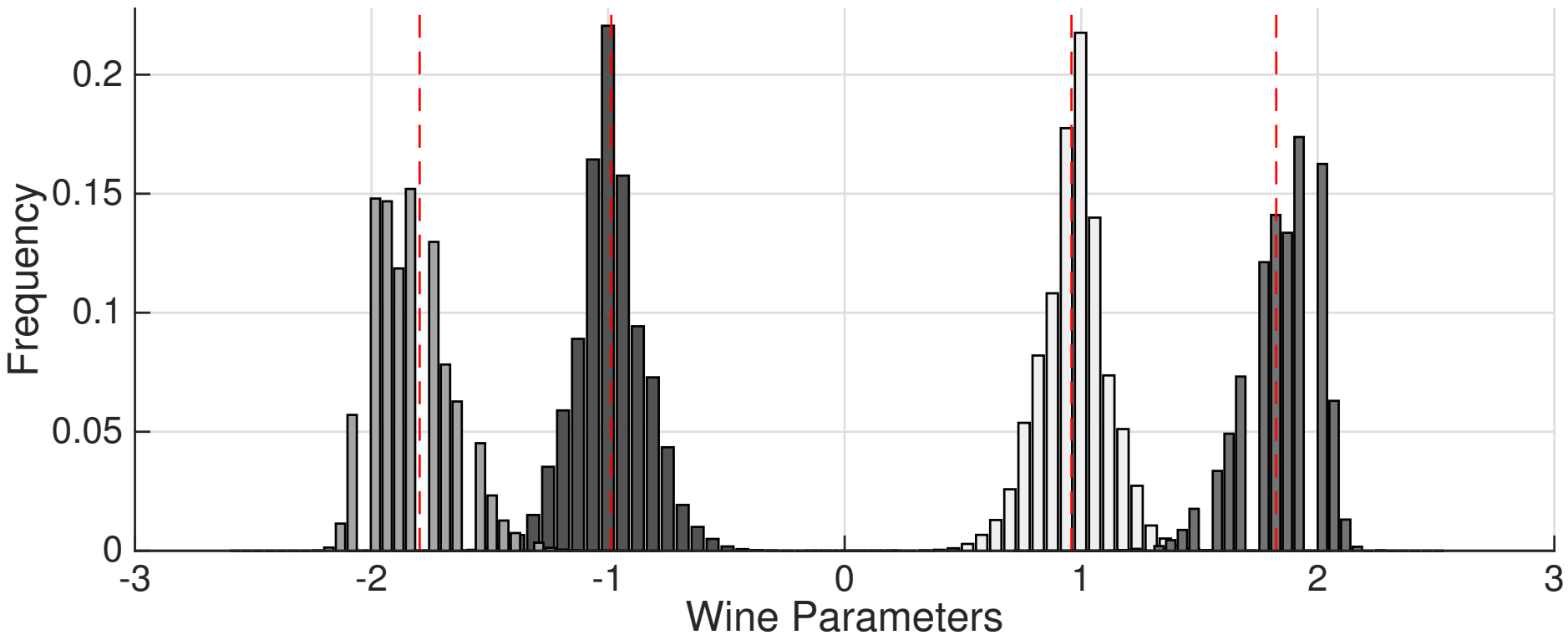
$$Y_{wn} = \max(5, \min(1, \lfloor \tau_w + Z_{wn} \rceil)). \tag{18}$$

This is not so easy to analyse, but we may need to consider the bias in our hypothesis testing.

Null hypothesis significance testing

Given N measurements of each wine in $\mathcal{W}$, and our measurement model, we wish to test the null hypothesis:

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



Decomposing the sum of squares

Returning to the sum of squares we minimised to find the estimator of β_0, we can decompose it in the following way:

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(y_{wn}-\hat{\beta}_0)^2=\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(y_{wn}-\hat{y}_w)^2+\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2. \tag{19}$$

- The left-hand term is proportional to the variance of our data from the "grand mean" (called, "total sum of squares").
- The first term on the right is proportional to the variance of our data from the sample mean of each wine (called, "within-group variance").
- The last term is proportional to the variance of all our predicted data to the grand mean (called, "between-group variance").

Expectations of the sum of squares

The expectation of the within-group variance is:

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

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

$$-\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\sum_{\substack{m=1\\m\neq n}}^{N}(\mathrm{Cov}(Z_{wn},Z_{wm})+E[Z_{wn}]E[Z_{wm}]) \quad (20)$$

Expectations of the sum of squares

The expectation of the between-group variance is:

$$E\left[\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2\right]=\sum_{w\in\mathcal{W}}N\beta_w^2+2\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}\beta_w E[Z_{wn}]$$
$$+\frac{(|\mathcal{W}|-1)}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n,m=1}^{N}(\mathrm{Cov}(Z_{wm},Z_{wn})+E[Z_{wn}]E[Z_{wm}])$$
$$-\frac{1}{N|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{v\in\mathcal{W}\setminus\{w\}}\sum_{n,m=1}^{N}(\mathrm{Cov}(Z_{wn},Z_{vm})+E[Z_{wn}]E[Z_{vm}]). \quad (21)$$

Expectations to hypothesis testing

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

$$E\left[\frac{1}{|\mathcal{W}|-1}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2\right]=\sigma^2+\sum_{w\in\mathcal{W}}N\beta_w^2/(|\mathcal{W}|-1) \quad (22)$$

$$E\left[\frac{1}{(N-1)|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2\right]=\sigma^2. \quad (23)$$

If H_0 is *also* in effect, then we expect these two terms to be equal. Hence, we wish to compute our estimates of these quantities and see if

$$\frac{1}{|\mathcal{W}|-1}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn}-\hat{\beta}_0)^2\approx\frac{1}{(N-1)|\mathcal{W}|}\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(y_{wn}-\hat{\beta}_0-\hat{\beta}_w)^2. \quad (24)$$

Back to hypothesis testing

If Z_{wn} iid zero mean and variance σ^2 and $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{|\mathcal{W}|}$, then

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn} - \hat{\beta}_0 - \hat{\beta}_w)^2/\sigma^2 \sim \chi^2_{(N-1)|\mathcal{W}|} \tag{25}$$

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn} - \hat{\beta}_0)^2/\sigma^2 = \sum_{w\in\mathcal{W}} N\hat{\beta}_w^2/\sigma^2 \sim \chi^2_{|\mathcal{W}|-1} \tag{26}$$

and so

$$F := \frac{\sum_{w\in\mathcal{W}} N\hat{\beta}_w^2/(|\mathcal{W}|-1)}{\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn} - \hat{\beta}_0 - \hat{\beta}_w)^2/((N-1)|\mathcal{W}|)} \sim F_{|\mathcal{W}|-1,(N-1)|\mathcal{W}|}. \tag{27}$$

What the hell did we just do? (Single-factor ANOVA!)

- We have several scores $\{y_{wn}\}$, which we model by an rv Y_{wn} and

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

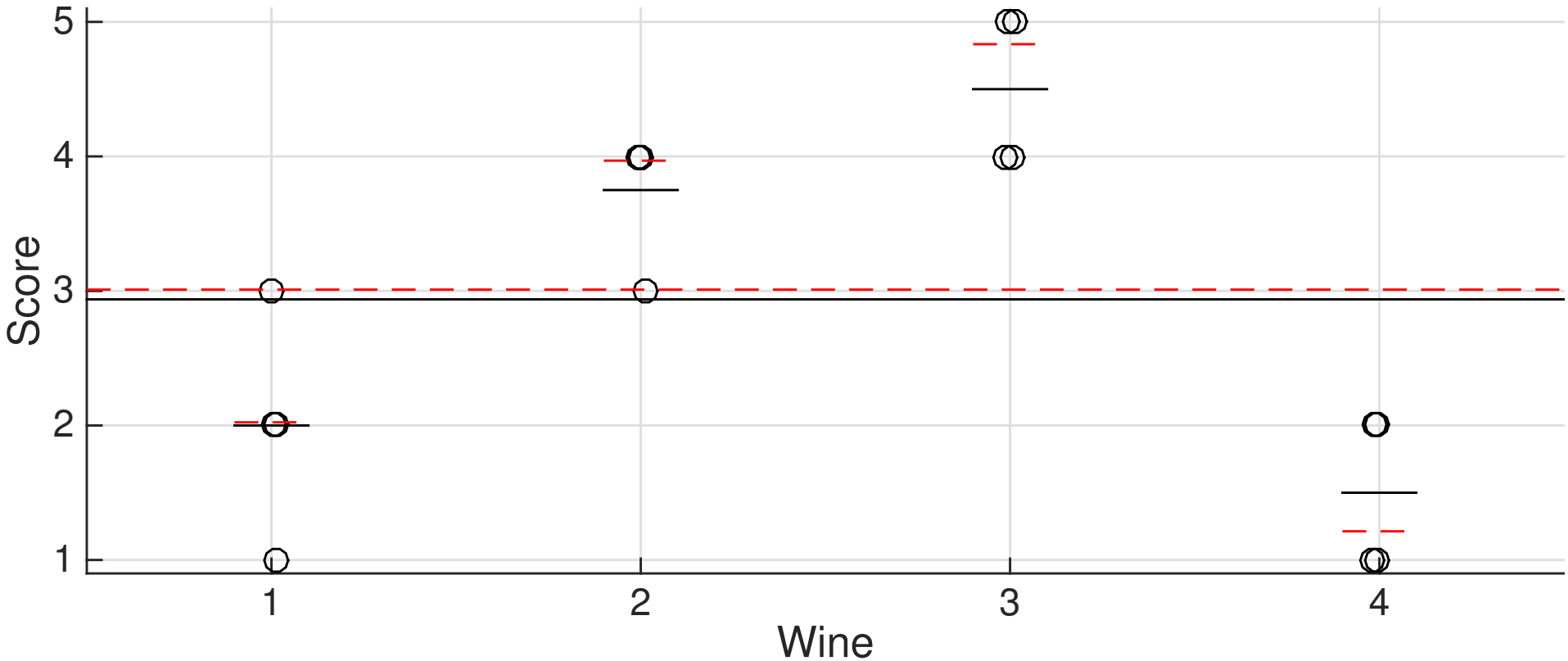
- We estimate $\{\beta_0, \beta_w : w \in \mathcal{W}\}$ by the principle of least squares.
- We assume a form of Z_{wn} (e.g., iid zero mean and variance σ^2).
- We compute the statistic

$$f = \frac{\sum_{w \in \mathcal{W}} N \hat{\beta}_w^2 / (|\mathcal{W}| - 1)}{\sum_{w \in \mathcal{W}} \sum_{n=1}^{N} (y_{wn} - \hat{\beta}_0 - \hat{\beta}_w)^2 / ((N-1)|\mathcal{W}|)}. \tag{28}$$

- We look that statistic up in the F-table $(3, 12)$ to find p in order to reject $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{|\mathcal{W}|}$ given α.

Let's run some simulations!

Null hypothesis significance testing

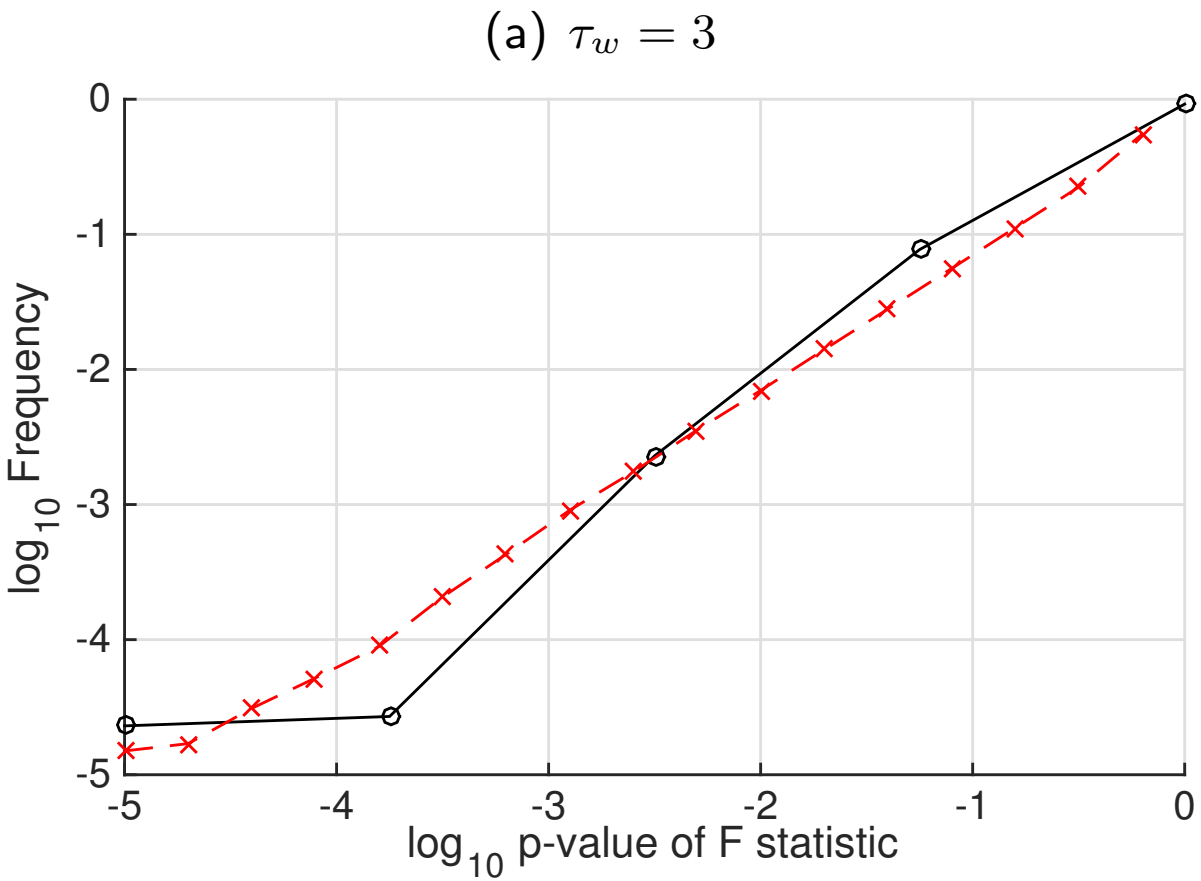


Here, $f = 20.36$. The probability of seeing a statistic at least that large given Z_{wn} iid zero mean and variance σ^2 and H_0 true is $p < 10^{-4}$.

Null hypothesis significance testing

What about our more realistic measurement model?

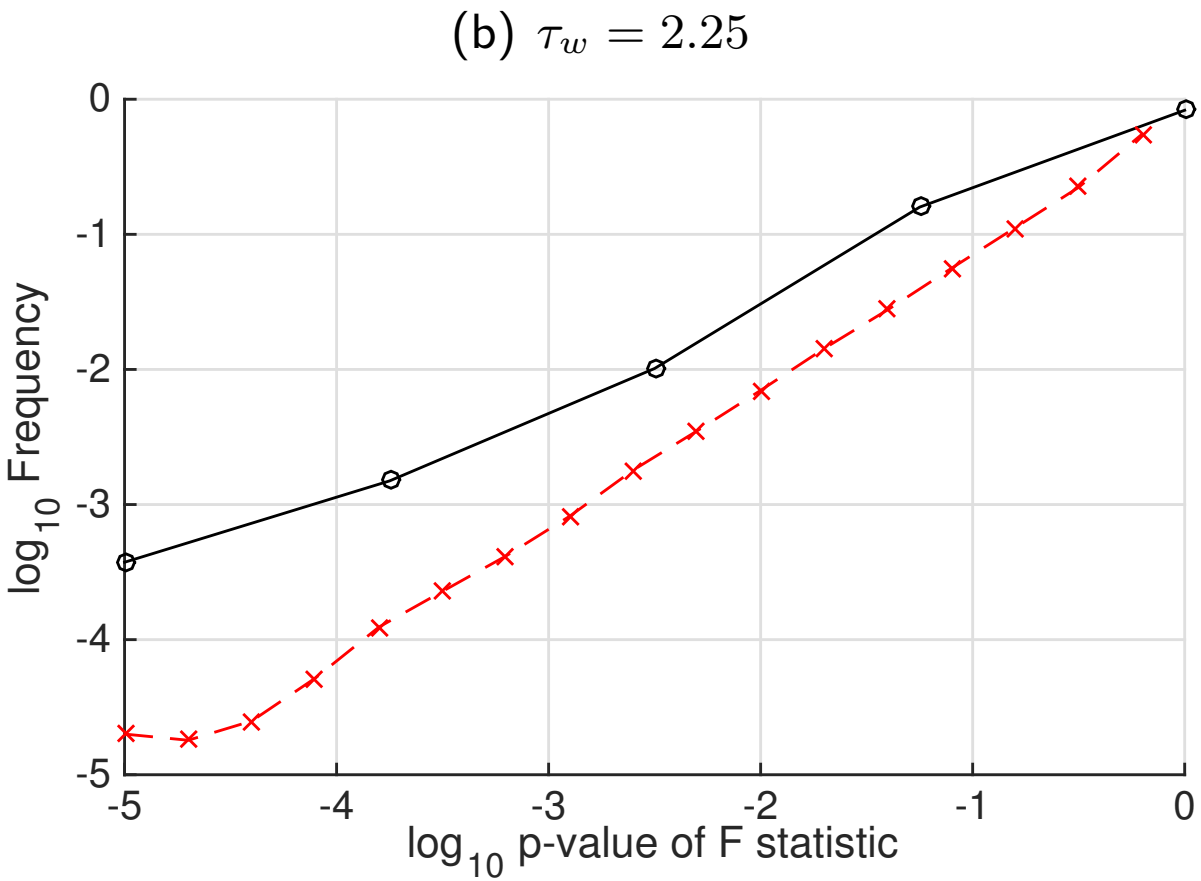
$$Y_{wn} = \max(5, \min(1, \lfloor \tau_w + Z_{wn} \rceil)).$$



Null hypothesis significance testing

What about our more realistic measurement model?

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

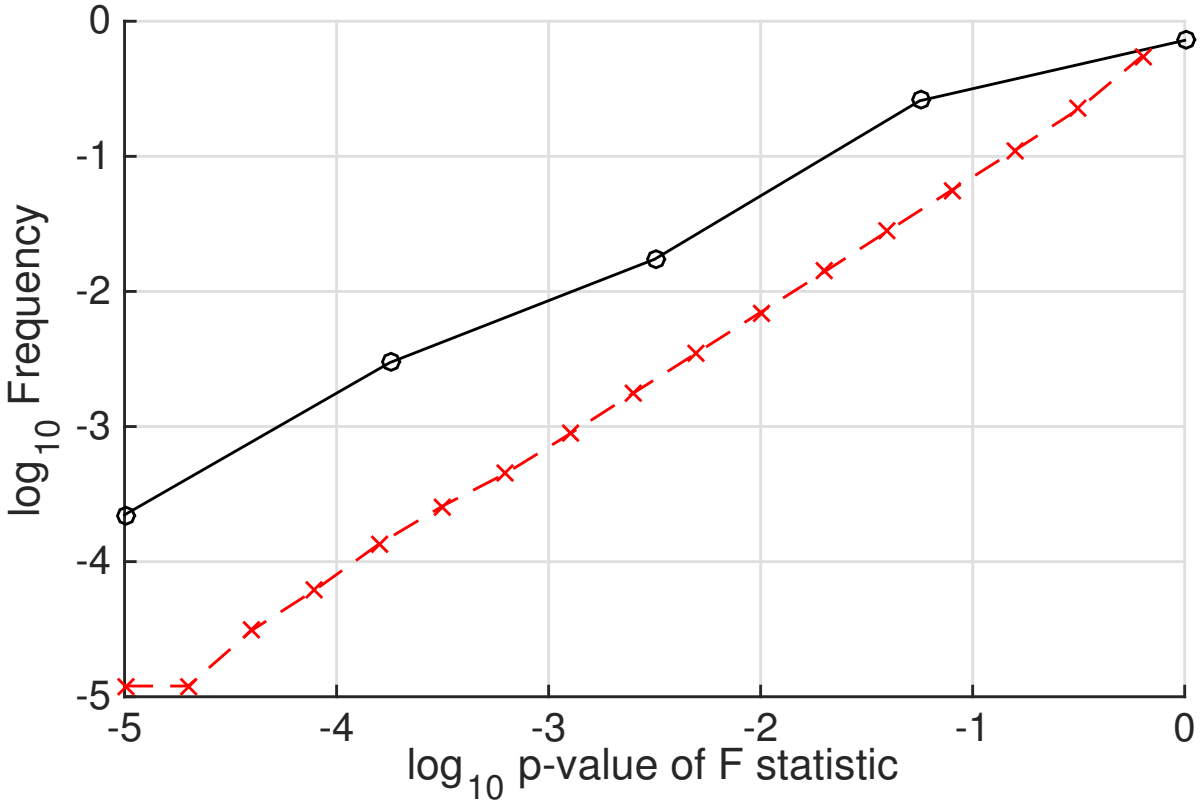


Null hypothesis significance testing

What about our more realistic measurement model?

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

(c) $\tau_w = 2.5$



Null hypothesis significance testing on our table

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

Here, $f = 9.06$. The probability of seeing a statistic at least that large given Z_{wn} iid zero mean and variance σ^2 and H_0 true is $p < 0.0021$.

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

What have we done?

- We've specified a single-factor measurement model.
- We've derived least-squares estimators of the parameters.
- We've decomposed the dataset variance (total sum of squares).
- We've specified a null hypothesis.
- We've assumed a form of Z_{wn}.
- We've tested H_0 using single-factor ANOVA.

But we've yet to consider the most important question:

What can we validly conclude from this table?

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

Highly dependent upon Z_{wn}!

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

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

What can we say?

What can we validly conclude from this table?

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

What we can validly say hinges entirely upon Z_{wn}.

A request to our local chapter of oenophiles

```
Hello.  It looks like the wine 2 scores are significantly
larger than the rest, but can you tell us which judges
scored which wines?
```

Setup

A response from our local chapter of oenophiles

Hai! We spoke with the people we hired to set up the judging and they said they had poured the wines such that each judge scored four glasses of wine (per our instructions), but that the four glasses of each judge had the same wine. So each judge scored one wine. Please tell me that's ok.

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

Implementation

The disappointing **response**

Thank you for that key information. It is in fact not ok. The experiment has false replication. There is no way to disentangle the judge and wine factors.

wine/judge	scores
1	3 4 3 2
2	5 4 5 5
3	2 1 3 1
4	2 3 2 4

Had each glass of each judge been *appropriately* mapped to the wines, then we could have said something. Now we cannot say anything.
(BTW, this invalidates our conclusion about wine 2.)

Disappearing degrees of freedom

Let's plough ahead, assuming Z_{wn} iid zero mean and variance σ^2 and try to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{|\mathcal{W}|}$:

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn} - \hat{\beta}_0 - \hat{\beta}_w)^2/\sigma^2 \sim \chi^2_{(N-1)|\mathcal{W}|} \tag{29}$$

$$\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(\hat{y}_{wn} - \hat{\beta}_0)^2/\sigma^2 = \sum_{w\in\mathcal{W}} N\hat{\beta}_w^2/\sigma^2 \sim \chi^2_{|\mathcal{W}|-1} \tag{30}$$

$$F := \frac{\sum_{w\in\mathcal{W}} N\hat{\beta}_w^2/(|\mathcal{W}|-1)}{\sum_{w\in\mathcal{W}}\sum_{n=1}^{N}(Y_{wn} - \hat{\beta}_0 - \hat{\beta}_w)^2/((N-1)|\mathcal{W}|)} \sim F_{|\mathcal{W}|-1,(N-1)|\mathcal{W}|}. \tag{31}$$

However, now $N = 1$, which leaves us with no degrees of freedom for the within-group variance. This experiment has "false replication."

Conclusion

- The validity of all any statistical test critically relies on the distribution of the noise in the measurements.
- If we cannot ensure it is distributed in a way that facilitates analysis (it is sufficient, but not necessary, that the noise be iid for analysis), then none of the tests above are statistically valid.
- Our statistical machinery is agnostic to that — *it produces numbers no matter what.*
- The formal design of experiments provides exactly what we need:

Design of Experiments (DOE)

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.

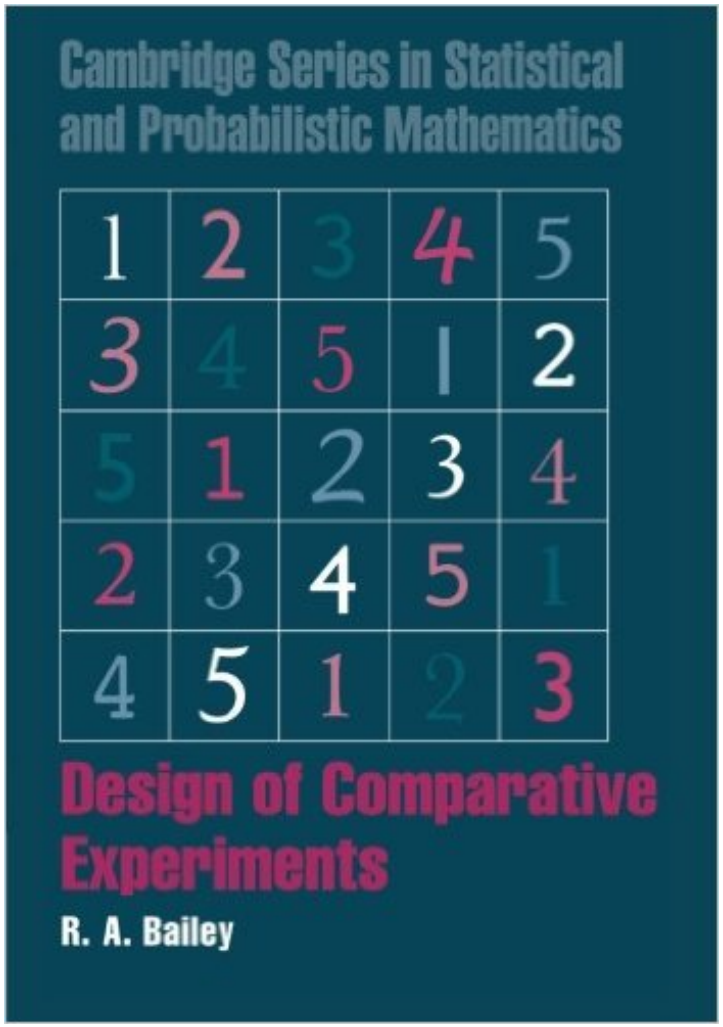
Experimental design is central to this pursuit.

Just a brief peek at fundamental components of DOE

Designing and implementing an experiment that is valid for a specific hypothesis entails performing several essential tasks:

- identifying treatments,
- identifying plots,
- recognising structures in the treatments and plots,
- mapping plots to treatments,
- specifying the measurement and its modelling.

See R. A. Bailey (2008)!



Just a brief peek at fundamental components of DOE

- *Treatments*: The set of things and their description applied to experimental units, $\mathcal{T} := \{i : i \in \{1, \ldots, t\}\}$.
- *Experimental unit*: The smallest unit to which a treatment is applied.
- *Observational unit (plot)*: The smallest unit on which a measurement is made.
- *Response*: The measurement made of an observational unit.
- *Plots*: The set of things mapped to treatments, $\Omega := \{\omega : \omega \in \{1, \ldots, N\}\}$.
- *Experimental design (treatment factor)*: A map $T : \Omega \to \mathcal{T}$.

Just a brief peek at fundamental components of DOE

- *Plot structure*: Meaningful ways (expert elicitation) of dividing up the plots. Possibilities are unstructured, blocks, etc.
- *Treatment structure*: Meaningful ways (expert elicitation) of dividing up the treatments. Possibilities are unstructured, treatment and control, etc.
- *Plan*: The translation of the experimental design into the actual plots.
- *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.

Just a brief peek at fundamental components of DOE

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	unstruct.	unstruct.	tomato yield (grams)	simple textbook
New animal feed	pen	calf	new and old feeds	unstruct.	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: Examples of the various components for four different experiments.