Recap of Nuisance Factors and Blocking

- In the context of designed experiments we categorize factors as either:
  - *design* factors
  - *allowed-to-vary* factors not controlled, known or unknown
  - *nuisance* factors controlled
- But remember: in practice, context dictates whether a factor should be considered a design factor, a nuisance factor, or if it should be allowed to vary
  - **Example:** Browser Type (context dictates the type of factors)

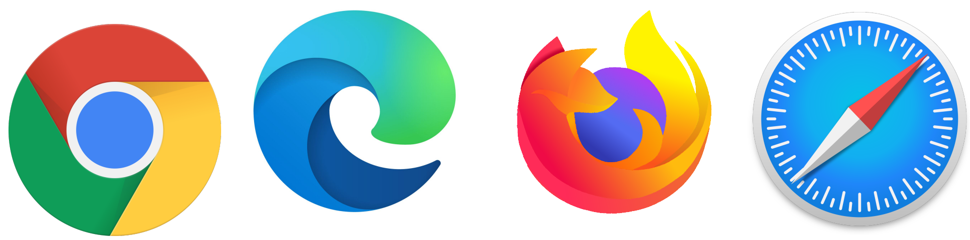

Figure 1: Four levels of the *browser* factor.

1. Usability testing involves studying the ease with which an individual uses a product or service for some intended purpose. Suppose investigators are performing a usability test to determine with which browser 70-80 year old users find it easiest to look-up the phone number of the nearest pharmacy. In this example, experimental units (70-80 year olds) are randomly assigned to one of four browser conditions, and the time it takes them to complete the task is measured.

2. Suppose that Netflix is experimenting with server-side modifications to improve (reduce) the latency of Netflix.com. Suppose that the current infrastructure serves as a control condition and the modified infrastructure is hypothesized to reduce median page load time. It is possible that a user's browser may also effect page load time, but this effect is not of interest to the investigators. To control for the potential impact of one's browser, Netflix intiallity experiments with only Firefox users.

3. Suppose that Amazon.ca is experimenting with the width of their search bar. They hypothesize that a wider search bar will minimize the amount of mouse movement required to navigate to it, thereby minimizing the average time-to-query. The experimenters do not care which browser a customer uses and so this factor is not controlled and hence is *allowed-to-vary* during their experiment.

- It's also important to understand the subtle distinction between nuisance factors and design factors in the context of a single experiment
  - Both factors are controlled during the experiment
  - With a design factor we wish to quantify its influence on the response variable
  - With a nuisance factor we do not care to quantify its effect, we wish only to *eliminate* it

- We eliminate the effect of one or more nuisance factors with **blocking**
  - To eliminate the effect of a nuisance factor, it cannot be allowed to vary on its own
  - Blocking fixes the nuisance factor at one or more levels (**blocks**)
  - By holding a nuisance factor fixed, it cannot vary and hence cannot influence the response
  - Blocking in this way represents a *randomization restriction*

Randomized Complete Block Designs

- The randomized complete block design (RCBD) is a simple experimental design that may be applied when we wish to investigate
  - a single design factor
  - while controlling for a single nuisance factor

- In a RCBD, each of the experimental conditions is carried out within every one of the blocks

- The *observed* data in such an experiment is denoted by y_{ijk}

- We assume that there are n_{jk} units in (condition, block) = (j, k) and thus an overall total of $N = \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk}$ units

- Response data of this form may be tabulated as shown below:

		Block				
		1	2	$\cdots$	b	
Condition	1	$\{y_{i11}\}_{i=1}^{n_{11}}$	$\{y_{i12}\}_{i=1}^{n_{12}}$	$\cdots$	$\{y_{i1b}\}_{i=1}^{n_{1b}}$	$\overline{y}_{\cdot 1 \cdot}$
	2	$\{y_{i21}\}_{i=1}^{n_{21}}$	$\{y_{i22}\}_{i=1}^{n_{22}}$	$\cdots$	$\{y_{2b}\}_{i=1}^{n_{2b}}$	$\overline{y}_{\cdot 2 \cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	m	$\{y_{im1}\}_{i=1}^{n_{m1}}$	$\{y_{im2}\}_{i=1}^{n_{m2}}$	$\cdots$	$\{y_{mb}\}_{i=1}^{n_{mb}}$	$\overline{y}_{\cdot m \cdot}$
		$\overline{y}_{\cdot\cdot 1}$	$\overline{y}_{\cdot\cdot 2}$	$\cdots$	$\overline{y}_{\cdot\cdot b}$	$\overline{y}_{\cdots}$

- The row, column and overall means are calculated as follows:

$$\overline{y}_{\cdot j \cdot} = \frac{1}{n_{j+}} \sum_{k=1}^{b} n_{jk} \overline{y}_{\cdot jk} \qquad \overline{y}_{\cdot \cdot k} = \frac{1}{n_{+k}} \sum_{j=1}^{m} n_{jk} \overline{y}_{\cdot jk} \qquad \overline{y}_{\cdot \cdot \cdot} = \frac{1}{N} \sum_{k=1}^{b} \sum_{j=1}^{m} \sum_{i=1}^{n_{jk}} y_{ijk} = \frac{1}{N} \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk} \overline{y}_{\cdot jk}$$

where $\overline{y}_{\cdot jk} =$ $\frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk}$ avg. response given (cond., block)

and $n_{j+} = \sum_{k=1}^{b} n_{jk}$ and $n_{+k} = \sum_{j=1}^{m} n_{jk}$ are respectively the overall number of units in condition j and the overall number units in condition k.

- Simple summaries such as these provide a crude assessment of whether the condition-to-condition and block-to-block variation is large.

If $\bar{y}_{\cdot j}$'s are different this suggests the design factors are influential

If $\bar{y}_{\cdot\cdot k}$'s are . . — 〃 . . . nuisance . . . – . —

- The primary analysis goal in a RCBD is to determine whether the expected response differs significantly from one condition to another
  - and if so, to identify the optimal condition
  - while controlling for the potential effect of the nuisance factor

- We've previously done this with gatekeeper tests of the form

$$H_0\colon \theta_1 = \theta_2 = \cdots = \theta_m \text{ vs. } H_A\colon \theta_j \neq \theta_k \text{ for some } k \neq j$$

- We do the same thing here, while accounting for the nuisance factor, with *appropriately defined* linear or logistic regression models which contain
  - an intercept,
  - $m-1$ indicator variables for the design factor's levels, and (x_{ij}'s)
  - $b-1$ indicator variables for the nuisance factor's levels (z_{ik}'s)

- We write the linear predictor as

$$\alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ik}$$

  - The β's jointly quantify the effect of the design factor
  - the γ's jointly quantify the effect of the nuisance factor

- Two relevant hypotheses are:

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ vs. } H_A\colon \beta_j \neq 0 \text{ for some } j$$

$$H_0\colon \gamma_1 = \gamma_2 = \cdots = \gamma_{b-1} = 0 \text{ vs. } H_A\colon \gamma_k \neq 0 \text{ for some } k$$

- These hypotheses are tested by comparing a *full* model and *reduced* models
  - We try to determine whether the full model fits the data significantly better than the reduced one

RCBD to Compare Means

- Here we're interested in testing the following hypothesis (while accounting for the influence of the nuisance factor):

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_m \text{ vs. } H_A\colon\ \mu_j \neq \mu_k \text{ for some } k \neq j$$

- We do this by testing

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0 \text{ for some } j$$

with an ANOVA in the context of the following linear regression model

$$Y_i = \alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ik} + \varepsilon_i$$

- The relevant ANOVA table is shown below:

Source	SS	df	MS	Test Stat.
Condition	SS_C	$m-1$	$MS_C = \frac{SS_C}{m-1}$	$\frac{MS_C}{MS_E}$
Block	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
Error	SS_E	$N-m-b+1$	$MS_E = \frac{SS_E}{N-m-b+1}$	
Total	SS_T	$N-1$		

- The sums of squares given in the table are:
  -
$$SS_T = \sum_{k=1}^{b} \sum_{j=1}^{m} \sum_{i=1}^{n_{jk}} \left(y_{ijk} - \overline{y}_{...}\right)^2$$
  -
$$SS_C = \sum_{k=1}^{b} \sum_{j=1}^{m} \sum_{i=1}^{n_{jk}} \left(\overline{y}_{.j.} - \overline{y}_{...}\right)^2$$

-
$$SS_B = \sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}} \left(\overline{y}_{\cdot\cdot k} - \overline{y}_{\cdot\cdot\cdot}\right)^2$$

-
$$SS_E = \sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}} \left(y_{ijk} - \overline{y}_{\cdot j \cdot} - \overline{y}_{\cdot\cdot k} + \overline{y}_{\cdot\cdot\cdot}\right)^2$$

- So how do we use this table?
  - We test
  $$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$$
  using $t_C \equiv \frac{MS_C}{MS_E}$
  - The p-value is: $Pr\left(F_{(m-1,\, n-m-b+1)} > t_C\right)$
  - We test
  $$H_0\colon\ \gamma_1 = \gamma_2 = \cdots = \gamma_{b-1} = 0$$
  using $t_B \equiv \frac{MS_B}{MS_E}$
  - The p-value is: $Pr\left(F_{(b-1,\, n-m-b+1)} > t_B\right)$ $\boxed{t_B \sim F_{(b-1,\, n-m-b+1)}}$

this null dist. is approximated because of the randomization restriction

The Gap Example

The Gap has three versions of an online weekday promotion that a customer sees when they go to gapcanada.ca:

- Version 1: 50% discount on 1 item
- Version 2: 20% discount on your entire order
- Version 3: Spend \$50 and get a \$10 gift card

Interests lies in determining whether there is a difference in the average purchase total (i.e, the average dollar value of a customer's purchase) between promotion versions. However, the amount of money one spends may also be influenced by the nuisance factor, day of week. As such, a randomized complete block design was run with $m = 3$ experimental conditions (corresponding to the three promotions) and $b = 5$ blocks (corresponding to the day of the week). Here $n_{jk} = 50\ \forall j, k$, and so the design was "balanced". For each visitor to gapcanada.ca, their purchase total (in dollars) was recorded. The regression model fit to these response observations is

$$Y_i = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \gamma_4 z_{i4} + \varepsilon_i$$

Source	SS	df	MS	Test Stat.
Condition	49618.34	2	24809.17	2165.39
Block	19258.30	4	4814.58	420.22
Error	8512.67	743	11.46	
Total	77389.32	749		

where x_{i2} and x_{i3} are condition indicators for promotions 2 and 3 (promotion 1 is the baseline) and $z_{i1}, \ldots, z_{i4}$ are block indicators for Monday-Thursday (Friday is the baseline). The ANOVA table obtained from the observed data is shown in the table above.

RCBD to Compare Proportions

- Here we're interested in testing the following hypothesis (while accounting for the influence of the nuisance factor):

$$H_0\colon \pi_1 = \pi_2 = \cdots = \pi_m \text{ vs. } H_A\colon \pi_j \neq \pi_k \text{ for some } k \neq j$$

- We do this by testing

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ vs. } H_A\colon \beta_j \neq 0 \text{ for some } j$$

  with a likelihood ratio test (LRT) in the context of the following logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ik}$$

  - The likelihood ratio test compares the full model to the one with out the x's

- Similarly, we test
$$H_0\colon \gamma_1 = \gamma_2 = \cdots = \gamma_{b-1} = 0 \text{ vs. } H_A\colon \gamma_k \neq 0 \text{ for some } k$$
with a LRT that compares the full model to the reduced one without the z's

H0: Reduced model fit as well as the full model

HA: full model better

- The observed test statistic for both of these tests is

$t \sim \chi^2_{( )}$

$$\begin{aligned} t &= 2 \times \log\left(\frac{\overset{\text{(maximized)}}{\text{Likelihood}_{\text{Full Model}}}}{\underset{\text{(maximized)}}{\text{Likelihood}_{\text{Reduced Model}}}}\right) \\ &= 2 \times [\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}] \end{aligned}$$

  - The p-value is: $Pr(\chi^2(m-1) > t)$ for design factors
    $Pr(\chi^2(b-1) > t)$ for nuisance factors

The Enterprise Example

Enterprise is experimenting with $m = 3$ banner ads as a mechanism to drive traffic to their website. Since there are known regional differences in consumer preferences in the US, they wish to control for the nuisance factor "region" with $b = 4$ blocks corresponding to the four major US geographic regions: Northeast (NE), Northwest (NW), Southeast (SE), and Southwest (SW). A total of $n_{jk} = 5000\ \forall j, k$ people were randomized to each ad condition in each region.

Interest lies in determining whether or not the different ads perform similarly with respect to click-through-rate (CTR) – and we wish to determine which one maximizes CTR – but we want to control for the effect of region. We do so with the following logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3}$$

where x_{i2} and x_{i3} are condition indicators for ads 2 and 3 (ad 1 is the baseline) and z_{i1}, z_{i2}, z_{i3} are block indicators for the NW, SE, SW regions (NE is the baseline).

in this example, can't randomize w.r.t. regions (rand. restriction)
because people are stuck with one region

Optional Exercises:

- Proofs: 20
- R Analysis: 18
- Communication: 1(d)