MA50259: Statistical Design of Investigations

Dr. Sandipan Roy

Lecture 1

Assessment Information

- Coursework. No Exam!
- Coursework counts for 100% of the mark for the unit
- Carry out statistical analysis (in R) and write a report.
- Coursework sheet 1
  - Handed out: March 04, 2024
  - Deadline: March 18, 2024
- Coursework sheet 2
  - Handed out: April 15, 2024
  - Deadline: April 29, 2024

Course broad outline

- 75% (at least) on Design and analysis of Experiments
- 25% (at most) on Design and analysis of Observational Studies
- Lectures every Tuesday and Thursday at 12:15 on each day
- Labs will be every Friday at 12:15
- Use of R and RStudio (make sure you have a recent version of both), knowledge of R markdown will be useful although we will cover it as we go along
- R packages: Initially you will have to install the *tidyverse* package

Statistical design of investigations

Statistics is the science of *collecting, analysing, and drawing conclusions from data*. The value of statistical methods is twofold- (a) to make inference about the population and (b) make predictions on certain feature of the population.

Ways of collecting data:

- **Sampling surveys:** Used when the purpose of data collection is to estimate some property of a finite population without conducting a census (e.g. household water consumption in a city, proportion of people watching a tv program in a region etc.)
- **Observational studies and Experiments:** Used to determine the relationship between two or more measured quantities in a population (e.g. relation between smoking and incidence of lung cancer)

Observational studies vs Experiments

- **Observational studies:** data is observed in its natural environment
- **Experiments:** data is observed when the environment is controlled
- **Observational studies:** Hard to find cause and effect. Correlations may be found between variables because they are both affected by changes in a third variable that was not observed or recorded. Predictions must assume the same interrelationships among variables that existed in the past will exist in the future.
- **Experiments:** Some variables are purposely changed while others are held constant. In that way the effect that is caused by the change in the purposely varied variable can be directly observed, and predictions can be made about the result of future changes to the purposely varied variable.

Design of Experiments

Purposes of experimentation:

- determining the cause for variation in measured responses
- finding conditions that give rise to the maximum or minimum response
- comparing the response between different settings of controllable variables
- obtaining a mathematical model to predict future response values
- Planned experiments are used in many fields: engineering design, quality improvement, industrial research and manufacturing, physical and biological sciences, social sciences, psychology, business management and marketing research.

Design of Experiments: Statistical principles

The amount of variation in some experiments often confuse the results plus experiments can be time consuming and costly to carry out. This motivates to find experimental techniques that can:

- eliminate as much of the natural variation as possible
- prevent unremoved variation from confusing or biasing the effects being tested
- try to detect cause and effect with the minimal amount of experimental effort
- Have a look at list of definitions in unit's Moodle page

Example: Experiment to test plywood strength

- Plywood is a material manufactured from thin layers or "plies" of wood that are glued together
- An experiment was carried out to compare the shear strengths of birch plywood produced with 6 different glues
- Ten batches (of five test pieces of birch plywood) were tested for each type of glue
- Shear strength (in pounds per square inch) is the load that an test piece is able to withstand in a direction parallel to the face of the piece
- Collected data on the mean shear strength for each batch
- **Question:** Is there any difference in shear strength for different glues?
- Need to check if the assignment was random

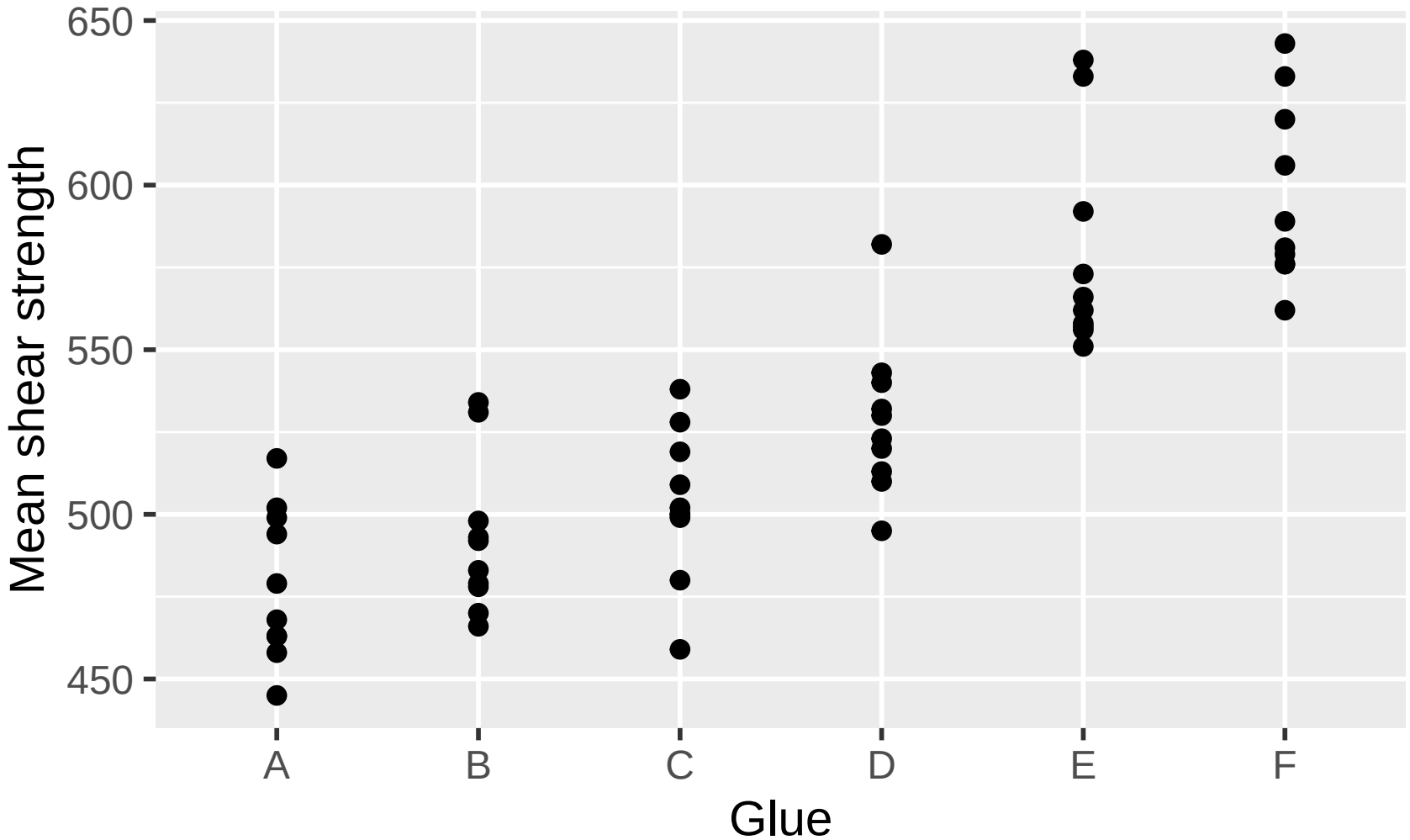
Example: Experiment to test plywood strength

```
library(tidyverse)
plywood<-read_delim(
  "http://people.bath.ac.uk/kai21/MA50259/Data/plywood.txt",delim=" ")
glimpse(plywood)
Rows: 60
Columns: 3
$ strength <dbl> 502, 458, 445, 479, 468, 463, 517, 494, 499, 463, 470, 483, 4~
$ glue     <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "B", "B", "~
$ units    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
```

Example: Experiment to test plywood strength

```
ggplot(plywood,aes(glue,strength)) + geom_point() +

xlab("Glue") + ylab("Mean shear strength")
```



Example: Experiment to test plywood strength

Design: Completely randomized design with one treatment factor

- What are the experimental units?
- How many replicates?
- What is random?
- What is the independent variable (treatment factor)
- What is the dependent (response) variable?
- What is controlled by the experimenter?
- Any lurking variables?
- Can we conclude something from the plot?

Completely randomized design (CRD)

- One treatment factor with t levels (Independent variable)
- the n experimental units are divided **randomly** into t groups and each group is then subject to only one of the levels of the treatment factor
- If $n = tr$ is a multiple of t, then each level of the factor will be applied to r experimental units, and there will be r replicates of each run with the same level of the treatment factor [balanced design]
- If n is not a multiple of t, then there will be an unequal number of replicates of each factor level
- The response is observed for all the experimental units [timing?]
- This design should be used when there is only one factor under study and the experimental units are homogeneous
- Any other known independent variables are held constant so that they will not bias any conclusions

Our own randomised experiment

```r
set.seed(5678) # to control randomness
r<-10 ;t<-6;
levels<-c("level 1","level 2","level 3","level 4", "level 5","level 6")
f <- rep(levels,each = r) %>% factor()
n<-r*t
fac <- f %>% sample(size=n) # randomisation
units <- 1:n # unit labels
crd <- tibble( units=units, treatment=fac ) # creates data.frame (tibble)
glimpse(crd)
Rows: 60
Columns: 2
$ units     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
$ treatment <fct> level 4, level 5, level 6, level 3, level 3, level 6, level ~
crd<- crd %>% arrange(treatment) # arrange by treatement level
glimpse(crd)
Rows: 60
Columns: 2
$ units     <int> 7, 12, 14, 19, 23, 31, 33, 39, 54, 55, 18, 28, 37, 38, 41, 4~
$ treatment <fct> level 1, level 1, level 1, level 1, level 1, level 1, level ~
```

Note:

- The units are not grouped by treatment levels!
- We have not observed the response yet!

Replication

- Replication dictates that r different batches of plywood are tested at each of the t different types of glue rather than a single test piece of plywood being re-manufactured t times!
- Replicates in each level of the treatment factor means that the variance of the experimental error can be estimated from the data
- If the variability among the treatment means is not larger than the experimental error variance, the treatment differences are probably due to differences of the experimental units assigned to each treatment.
- Without replication it is impossible to tell if treatment differences are real or just a random manifestation of the particular experimental units used in the study
- The number of replicates can be determined by considering which difference in the response value is of practical importance and also on historical estimates of variability, if known

Randomisation

- The random division of experimental units into groups is called randomisation, and it is the procedure by which the validity of the experiment is guaranteed against biases caused by lurking variables
- In the plywood experiment, randomisation would prevent bias from lurking variables, such as variability in the strength from batch to batch and trends in the strength measurement technique over time (attributed to test machine or human error)
- When experimental units are randomized to treatment factor levels, a statistical test of the hypothesis that the treatment effect is zero can be performed
- The completely randomized design (CRD) was chosen so that differences in lurking variables between batches would be unlikely to correspond to changes in the factor levels

Our own randomised experiment, simulating the response

In order to understand statistical analysis later, we can think of how to simulate the response in our own randomised experiment.

- Let Y_{ij} be the response for the jth experimental unit subject to the ith level of the treatment factor, $i = 1, \ldots, t$ and $j = 1, \ldots, r_i$. Y_{ij} is treated as a **random variable**
- r_i is the number of replicates in ith level of the treatment factor. Note that $r_i = r$ for all i, if the design is balanced
- Let μ_i be the **mean response** (long-run average of all possible experiments) at the ith level of the treatment factor, then

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- ϵ_{ij} is the (additive) experimental error and is also treated as a **random variable**