Lecture 21

1. **Revision questions of L20**

   - What are the three main types of normality test?

     Answer:

   Type 1. Pearson Chi-square goodness-of-fit test.

   Type 2. Kolmogorov-Smirnov (KS) test, Cramer-von Mises (CVM) test, Anderson-Darling (AD) test.

   Type 3. Shapiro-Wilk (SW) test and the simulation based Shapiro-Francia (SF) test.

   - What are the ideas for each type of normality tests?

     Answer:

   Type 1. compare observed and expected counts over certain equal width intervals.

   Type 2. compare empirical and tested distribution functions approximated by the sample points. KS test takes for the max of such difference whereas CVM and AD tests calculate the weighted square difference. The weights for CVM test is uniform while the weights for AD test is inversely proportional to $F(y)(1-F(y))$.

   Type 3. consider R^2 of $Y_{(i)}$ and $\Phi^{-1}(\frac{0.5+i}{n})$ which are the points in the QQ plot. The SF test simulates the distribution of R^2.

2. This application demonstrates some important points in consulting: efficient communication with clients, understanding the objectives and limitations, cleaning techniques for big data, data visualisation through graphs, common features of statistical analyses for big data (outliers, missing values, high significance), etc.

Lecture 22

1. **Revision questions of L21**

   - Why are the terms in models fitted to a big data mostly significant?

     Answer: The models we consider assume normal distribution. For large data set, the assumption of normality for all observations are very strong particularly, we often see outliers. Such an assumption adds much information/power to our test leading to mostly significant results but such information may not be valid. We may consider other nonparametric test that is more scale free for the measurement. One idea is the rank and the other is to regenerate/simulate the distribution based on permutation instead of using normal.

   - There are often missing values in a large data. How to deal with them?

     Answer: Depending on the percentage of missing, it can be totally discarded (if the missing percentage is low) or imputed (if the percentage is high). There are different ways of imputation, by means, neighbour values and regression estimates, etc. Multiple imputation provides multiple imputed values to estimate and adjust the variability of imputed values.

2. **Basic concepts:**

   Treatment is the set of different experimental conditions to be tested.

   Experimental unit (EU) is the smallest unit different treatments can be applied to.

   Observational unit (OU) is the smallest unit response can be measured.

   Analytical unit (AU) is the basic unit in the statistical analysis. AU can be OU or some aggregation of OU.

   Often EU=OU=AU. EU can contains some OUs but it is a poor design if OU contains some EUs.

3. *Blocking* is the arranging of EUs in groups (blocks) that are similar to one another. Typically, a blocking factor is a source of variability that is not of primary interest to the experimenter.

4. *Confounding factor, confounder or lurking variable* is a variable that influences both the dependent variable and independent variable, causing a spurious association and bias the result. There are three general types:

   - *Selection bias:* the assignment of subjects to treatment/control is based on investigator's judgment or some selection criteria other than randomisation leading to *nonrandomised controlled experiment*.

     Use randomisation in selecting EUs to treatments and control.
   - *Observer/receiver bias:* the subjects or investigators are aware of the identity of the treatment/control groups and become biased in providing responses or evaluations as they may deliberately or subconsciously report more/less favourable results.

     Use double blind experiment with placebo.
     Placebo is a pretend treatment.
     Placebo effect is an effect which occurs from subjects respond to the idea of treatment.
     Double blind experiment refers to the case when the subjects (single blind) and investigators (both for double blind) are not aware of the identity of the treatment/control groups.

   - *Consent bias:* subjects choose whether or not to take part in an experiment.

     It is hard to deal with as it is an ethical issue in human or animal trials to withhold treatment for those in the control group or enforce treatment for those in the treatment group.

5. An example of selection bias given in the lecture:

   A portacaval shunt redirects blood flow in cases of cirrhosis of the liver. The surgery is long and dangerous, but a study in 1966 detected an increase in life expectancy compared to those without operation and claimed worthy of the surgery risk.

   However, the design of experiment was biased toward surgery, as healthier patients tended to have surgery. Hence the increase in life expectancy may due to healthier patients.

These concepts form the foundation of experimental design and statistical analyses. You may be asked to apply these concepts but not simply copy the definitions in an exam.

Lecture 23

1. **Revision questions of L22**

   - What do treatment, experimental unit, observational unit and analytical units mean?

     Answer: Refer to the previous summary on definitions.

   - Why should we consider block?

     Answer: To make EU alike so as to reduce RSS and make treatment comparison more efficient.

   - What kind of biases (confounding factors) can occur in an experiment and how to resolve them?

     Answer:
     Selection bias: use randomisation.
     Observer/receiver bias: use double blind experiment with placebo.
     Consent bias: hard to deal with as we can't withhold or enforce treatments.

2. Under the assumption of common/same variance across treatments in ANOVA models, *balance design* with equal sample size for each treatment is optimal.

3. *Completely randomised design* refers to the design in which every design or allocation of EUs to treatments is equally likely. For b blocks with t treatments each (complete design), there are $(t!)^b$ number of allocations or designs in total and they are equally likely to be adopted.

4. *Systematic design* systematically (instead of randomly) permutes treatment labels in each block. While it can avoid row effect that can occur in some adverse designs under randomisation, it may cause interaction between treatments due to the systematic pattern. The probability of selection is no longer equally likely.

5. The effect of adverse designs in complete randomisation can be averaged out by replication.

---

last adjustments: April 30, 2021 by JC