CGS602A: Basic Statistics, Data Analysis and Inference

End-semester exam

Max marks: 130
Time: 150 minutes 16 Dec. 2020

1. *Answer all 6 questions. Note that question 1 is a must.*
2. *You can use online tables or statistical calculators where needed.*
3. *Do not collaborate or look up answers on the internet. However, you can use your own notes and/or slides on the course website.*
4. *Upload your solutions as a .zip archive to mooKIT. Pl. avoid .jpg/.png files. Only pdf files.*

1. Please give an undertaking in your own words saying that you have not collaborated with anyone and have not used the internet (except as permitted in the instructions above) in answering the question paper.

2. In each case give a short answer and justification. Please keep your answers and justifications **precise and short**.

   (a) If the pdf of a distribution differs from a univariate normal distribution by having a skew on the left then what will be the kurtosis (not excess kurtosis) of the distribution?

   Solution:
   Kurtosis is defined as the fourth moment $(\frac{X-\mu}{\sigma})^4$. So, greater skew on either side (compared to the normal) will mean more outliers and so will increase the value beyond what it is for a normal distribution (that is 3.0). So, we expect the kurtosis to be greater than 3.0. How much greater will depend on the amount of skew.

   (b) What is the major reason to use randomization when we assign subjects to treatments?

   Solution:
   Theoretically, there are two reasons to randomize: a) Protects against confounds, bias b) Can be used as a basis for inference. While b) is almost never used in practice due to computational complexity a) is the main reason for randomization. In experiments it is not possible to account for and control all independent variables. Often we don't even know all the variables that affect the dependent variable. In such a situation randomization helps protect against confounds making the distribution of values of such uncontrolled variables similar across groups. Randomization also helps mitigate bias.

   (c) What is the key difference between an observational study and an experiment?

   Solution:
   In an experiment the experimenter, ideally, controls all the independent variables or at least the variables of interest. In an observational study the independent variables are not

controlled by the experimenter. For example, behavioural studies of animals in the wild are by necessity observational. Many longitudinal studies of human cohorts over long periods of time are also observational by necessity.

(d) We know the slogan "*correlation does not imply causation*". But what is very probable if two RVs are highly correlated and why?

Solution:
It is highly probable that there is a latent variable that highly correlates with both RVs thus giving rise to the original correlation. So, if in a city the number of schools is highly correlated with the number of criminals it clearly does not imply that schools produce criminals. Rather both are likely to be highly correlated with the population of the city. If a city has more people then it is likely to have more children and also more criminals. Of course, since variables are highly correlated one can predict one from the other.

(e) Normally, the significance level $\alpha = 0.05$. Under what conditions will you have a stricter α - that is $\alpha < 0.05$? Give a concrete example and explain.

Solution:
α gives the probability of a type I error or probability of incorrectly identifying a positive. So, if the risk associated with a false positive is very large we would significantly reduce this probability. For example, consider a clinical condition that has a very high mortality rate but is curable if diagnosed early but with substantial chances of side effects that while not fatal, reduce the quality of life to a great degree. A false positive in such a situation has serious implications. So, the probability of a type I error for the test will have to be much tighter than 0.05 perhaps 0.001 or less.

(f) Suppose a pilot experiment has shown that the effect size is small. What will you do when you plan the main study and why?

Solution:
If the effect size is known to be small or has a high probability of being small and *the effect is scientifically/practically important* then the main experiment seeking to establish the effect must have a large enough sample size so that the experiment has enough power.

However, it is important to note that even a very small effect can become statistically significant if the sample size is large enough. So, it is important that the effect size in question for which the experiment is being done be scientifically/clinically important.

(g) What are the consequences if your study is under powered and what can you do to ensure that your study has adequate power?

Solution:
Under powered studies imply high β that is higher probability for type II errors. This impies that it may not detect practially important effects even when they exist especially when the effect size is small (complement of the answer to question f)). Also, the sampling distribution

of underpowered studies has higher variance which means only large effects can be reliably detected. In practice the commonly used way to increase power is to increase sample size.

(h) How are regression and ANoVA related and when will you use what? Differentiate using examples.

Solution:

Let us consider simple linear regression and one way ANoVA. ANoVA is used whenever the independent variable is a categorial variable and the treatment groups are characterized by different values of the categorial variable - often there are just two or very few groups in such experiments. Regression is used when the independent variable is an interval or ratio variable and we can infer a quantitative linear predictive relation between the independent and dependent variables.

For examples, consider three different non-drug, behaviour change treatments for depression where depression is measured numerically on some depression scale. This is a study that will be typically analysed by ANoVA.

On the other hand if we want to study the effect different dosage levels of some anti-depressant drug (in μ-grams/kg) have on depression scores then we will use simple regression.

(i) Most tests make certain assumptions, typically the normality assumption is very common. When is it possible to do tests without any assumptions?

Solution:

The assumptions are needed to ensure that the sampling distribution is of a particular type which then allows us to calculate critical values for the test statistic for comparison. If the assumptions do not hold then this is not possible. The alternative is to actually generate the sampling distribution by computer simulation using Monte Carlo methods and bootstrap to get random samples. The sampling distribution that is generated by the simulation is then directly used to get the critical values for comparing the test statistic. No assumptions are needed about the data in this case.

Of course, a non-parametric test is another possibility. But then we are changing the test itself and doing a different test.

(j) Give an experiment of your own invention (don't choose any discussed in class) that requires a within subjects design. What is the main advantage of a within subjects design?

Solution:

Any treatment that makes comparative before-after claims can use a within subjects design. Consider for example, claims aboutr cosmetic performance (e.g. fairness creams) or about individual performance. For example, if I claim I can tell the alcohol content of any alcoholic beverage to within 1% then the experiment to test my claim has to be a within subjects design.

One main advantage of a within subjects design is that it reduces the number of independent subjects. Another important advantage is that the effect of variables that are not controlled between the treatments are mitigated since the same subject is involved.

[10×5=50]

3. Express the following intutition formally as an inequality, justify it formally and then say where it is used in hypothesis testing.

   Intuition: Given a finite or countable set of events, the probability that at least one of the events happens is less than or equal to the sum of the probabilities of the individual events.

Solution:

This result is often called the *union bound*. Formally, if $E_i, i = 1..n$ are n events with probabilities $P(E_i)$ then $P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$.

To justify it note that $P(E_1) \leq P(E_1)$ so it holds when $n = 1$. Assume it holds for $n-1$. We know that $P(\cup_{i=1}^{n} E_i) = P(\cup_{i=1}^{n-1} E_i) + P(E_n) - P(\cup_{i=1}^{n-1} E_i \cap E_n)$. This follows from the fact that $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$. The last term in the earlier equality is positive since it is a probability. This implies: $P(\cup_{i=1}^{n} E_i) \leq P(\cup_{i=1}^{n-1} E_i) + P(E_n) \leq \sum_{i=1}^{n-1} P(E_i) + P(E_n) = \sum_{i=1}^{n} P(E_i)$.

We apply this principle in the Bonnferroni correction where the type I error probability α is divided by m when there are m tests happening together.

[3,5,2=10]

4. A football coach wants to test whether a new training program will improve the strength and flexibility of members in his team. To test it he creates an obstacle course, takes 15 randomly chosen members from his team and times each of them. Then he puts them through the new training regimen for a week and again makes them take the obstacle course and times them. The difference Δ = Time before training - Time after training is given below. Clearly shorter times on the course are better. The Δ values for the 15 members are given below:

   $\Delta = 1.1, -0.48, 0.62, 0.04, 1.75, 0.30, 0.43, 0.26, -0.97, 0.45, 0.42, -0.18, 0.18, 0.25, -0.57.$

   Help the coach to determine whether or not the new training regimen is better. You will have to determine the hypothesis, the test you will use, what α level and come up with a decision after applying the test. Give all necessary details.

Solution:

The null hypothesis H_0: $\mu_{\text{before}} = \mu_{\text{after}}$ or equivalently $\mu_\Delta = 0$. The alternate hypothesis is H_a: $\mu_{\text{after}} < \mu_{\text{before}}$ or equivalently $\mu_\Delta > 0$. Note the test is one sided since the claim is it will improve performance.

We have a standard within subjects experiment with a before-after comparison so the obvious test is a paired t-test. So, we do a one-sample t-test for $\mu_\Delta = 0$. We choose $\alpha = 0.05$.

$t = \frac{\bar{\Delta} - \mu_\Delta}{s/\sqrt{N}}$, where $\bar{\Delta} = 0.24$ is the sample mean, $N = 15$ is the sample size, $s = 0.662$ is the sample std. deviation and $\mu_\Delta = 0$. So, $t = \frac{0.24}{0.662/\sqrt{15}} = 1.404$. From the statistical calculator/tables we see that a one tailed t-critical value is 1.76. So t is less than the critical value and the null hypothesis cannot be rejected.

[15]

5. (a) What assumptions must hold when linear regression is to be used for predictive purposes to generalize to unseen values that are part of the population?

Solution:
Independent random sampling, linearity of relation Y, X, normal distribution of Y for each value of X, homoscedasticity - for each value of X the Y variable has the same variance - also called homogeneity of variance.

(b) Suppose RVs X and Y are perfectly correlated then what is the relation between z_X and z_Y where z_X, z_Y are the corresponding z values of X and Y.

Solution:
Since X and Y are perfectly correlated the corresponding z values - that is mean shifted, std. dev. scaled values - will be equal. That is $\hat{z}_Y = z_X$.

(c) If X, Y have a correlation coefficient of r then we can write the estimate $\hat{z}_Y = rz_X$. What will be the regression equation for Y in terms of X and population parameters?

Solution:
We have $\hat{z}_Y = rz_X$. Assuming μ_X, μ_Y and σ_X σ_Y are the population means and standard deviations of X and Y respectivly we get:

$$\begin{aligned}
\hat{z}_Y &= rz_X \\
\frac{\hat{Y} - \mu_Y}{\sigma_Y} &= r \times (\frac{X - \mu_X}{\sigma_X}) \\
\hat{Y} &= r\frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \mu_Y \\
\hat{Y} &= r\frac{\sigma_Y}{\sigma_X}X + (\mu_Y - r\frac{\sigma_Y}{\sigma_X}\mu_X)
\end{aligned}$$

(d) Suppose for a hypothetical group of men the correlation between waist size and height is $r = +0.6$. The mean height $\mu_X = 69$ and $\sigma_X = 3$; also $\mu_Y = 32$, $\sigma_Y = 4$.

i. What is the slope of the regression line between Y and X?

Solution:
From the equation derived earlier the slope is $r\frac{\sigma_Y}{\sigma_X} = 0.6 \times \frac{4}{3} = 0.8$.

ii. What is the value of the Y-intercept?

Solution:
The intercept is given by: $\mu_Y - r\frac{\sigma_Y}{\sigma_X}\mu_X = 32 - 0.8 \times 69 = -23.2$

iii. Is the value in ii above a sensible value? Justify.

Solution:
It is not a sensible value since it says that for a height of 0 the waist size is -23.2. A

negative waist size is not possible.

[8,3,5,(3,3,3)=25]

6. In an experiment to test whether or not noise inhibits learning 15 subjects are assigned to 3 groups, 5 to each group: i) no noise ii) moderate noise iii) high noise. Subjects are given 20 minutes to memorize 10 nonsense syllables and they are told they will tested the next day. In the no noise condition subjects are in a quiet room with no ambient noise. For the moderate noise condition they do the task while listening to classical music being played and in the high noise condition the task is done while listening to rock music. The number of syllables correctly recalled the next day are shown below:

Group	**No. recalled**
i)	8, 10, 9, 10, 9
ii)	7, 8, 5, 8, 5
iii)	4, 8, 7, 5, 7

Based on the above answer the questions below:

(a) State the null and alternate hypothesis.

Solution:
If the three group means are: μ_1, μ_2, μ_3 then:
Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$.
Alternate Hypothesis H_a : *not* H_0 - that is at least one equality should not hold.

(b) Compute the relevant quantities: $SS_{\text{betweengroups}}$, $SS_{\text{withingroups}}$, $df_{\text{betweengroups}}$, $df_{\text{withingroups}}$, $s^2_{\text{betweengroups}}$, $s^2_{\text{withingroups}}$, F ratio.

Solution:
$SS_{\text{betweengroups}} = 26.53$
$SS_{\text{withingroups}} = 22.80$
$df_{\text{betweengroups}} = 2$
$df_{\text{withingroups}} = 12$
$s^2_{\text{betweengroups}=\frac{26.53}{2}} = 13.265$
$s^2_{\text{withingroups}} = \frac{22.80}{12} = 1.9$
$F = \frac{13.265}{1.9} = 6.98$ ratio

(c) Find the critical F value, compare the F ratio to the critical value and report the decision at the relevant α level(s) in the form of a standard ANoVA summary table.

Solution:
The critical values for F using the F calculator/tables are:
$F_{0.05} = 3.89$, $F_{0.01} = 6.93$. So, H_0 is rejected at both $\alpha = 0.05$ and $\alpha = 0.01$. The ANoVA summary table will look as follows:

Variation source	**SS**	**df**	**MS**	**F**
Between groups	26.53	2	13.265	6.98**
Within groups	22.80	12	1.9	
Total	49.33	14		

$\star\star$ - significant at $\alpha = 0.01$. Single $\star$ means $\alpha = 0.05$.

(d) If now you want to compare i), ii) and iii) pairwise what test and α will you use and why?

Solution:
A simple pairwise comparison can be made for each pair using a t-test and the chosen α will be $\alpha = \frac{\alpha_{\text{original}}}{m}$ where m is the number of pairs being tested. We are using the Bonferroni correction on α since multiple comparisons are being done.

[(2,3),12,7,6=30]