CGS600A: Computational Tools in Cognitive Science

End semester exam

Max marks:165
Time:3 hours 20 Nov. 2018

1. *Answer all 7 questions. The paper has 5 pages and 3 pages of tables.*
2. *Questions 1 and 2 are short answer questions. Question 1 on reg. expressions and programming and question 2 on probability, statistics and hypothesis testing.*
3. *Please start the answer to a question on a fresh page and answer all parts of a question together.*
4. *You can only refer to your own handwritten class notes and nothing else.*
5. *You can use a calculator or use your mobile as a calculator after putting it in AIRPLANE mode.*
6. *Where needed assume that code fragments are typed into the Python interpreter. Note that when a program is executed it can give an error, an exception or no output.*
7. *Any code is shown using typewriter font like* `this`.

1. This question has 7 parts that require **short answers**. *Give a brief justification for each answer.*

   (a) Write a regular expression that can recognize 24HRS clock time. For example: 1pm is 13.00, 8.25am is 08.25, 1.45am is 01.45, midnight is 00.00.

   Solution:
   `([0-1][0-9]|2[0-3])\.[0-5][0-9]`

   (b)
   ```
   class A:
       val = 1

   class B(A):
       pass

   class C(A):
       pass

   print(A.val, B.val, C.val)
   B.val = 2
   print(A.val, B.val, C.val)
   A.val = 3
   print(A.val, B.val, C.val)
   ```
   What will be printed when the above code is executed?

   Solution:
   `1 1 1`

```
1 2 1
3 2 3
```

The dictionary associated with B has an entry for `val` due to the assignment `B.val=2`. C still inherits from A.

(c) What will be printed when the code fragment below is executed?

```
class Test:
    def __init__(self, v):
        self.x=v

testObj=Test(10)
testObj.__dict__['temp']=50
print(testObj.temp+len(testObj.__dict__))
```

Solution:

52

After adding 'temp' : 50 to the `testObj` __dict__ its contents are: `{'x': 10, 'temp': 50}` so length of the dictionary is 2 and value of temp in the `testObj` namespace (i.e. __dict__) is 50 so output is 52.

(d)
```
def test(l):
    """l is a list of 4 numbers"""
    a, b, c, d=l
    def testfn():
        nonlocal a,b,c,d,l
        flag=True
        while flag:
            flag=False
            if a<b:
                b,a = a,b
                flag=True
            if b<c:
                c,b = b,c
                flag=True
            if c<d:
                d,c = c,d
                flag=True
        l=[a,b,c,d]
    testfn()
    print(l)
test([4,3,8,2])
```

What will be the output of the program above? What is it doing?

Solution:

Output is `[8,4,3,2]`. It sorts a size 4 list in descending order.

(e) What is the output of the following program?

```
l=[True, 20, 30]
l.insert(2,5)
print(l, 'Sum is = ',sum(l))
```

Solution:
Note that True is 1 and False is 0. So, output is:
`[True, 20, 5, 30] Sum is =  56`

(f) What is the output when the following code fragment is executed?

```
l=[3,4]
for i in l:
    l.append(i+1)
print(l)
```

Solution:
There is no output since the loop does not terminate. The list `l` is lengthend in each iteration by appending a new element.

(g) What is the output of the program fragment below.

```
import re
pat = re.compile('\d+')
print(pat.findall("Endsem is at 16.00 on 20th Nov. 2018"))
```

Solution:
$['16','00','20','2018']$
The pattern `\d+` matches `[0-9]` one or more times.

[4×7=28]

2. This question has 8 parts that require **short answers**. *Give a brief justification for your answer where relevant.*

(a) Suppose two indistinguishable coins are tossed simultaneously. What is the sample space? [Note: normally we assume one coin is tossed first and then the second so a toss can be distinguished from another in some way.]

Solution:
$\{\{H,H\},\{T,T\},\{H,T\}\}$
Note that since the coins are indistinguishable and tossed simultaneously it is not possible to distinguish between HT and TH. So, the outcomes are now sets.

(b) What is a random variable? Define and also explain using examples.

Solution:
Formally, a random variable (RV) is a mapping that associates an element of a set of outcomes of an experiment (the sample space $\mathcal{S}$) with elements of another set, say $\mathcal{E}$, which is

usually assumed to be $\mathbb{R}$. That is: random variable X is $X : \mathcal{S} \to \mathcal{E}$. A random variable always has an associated probability distribution - a pmf, cdf or pdf.

Examples: Choose a random student from IITK and measure his/her height and count number of sibilings. The height is a continuous RV while number of sibilings is a discrete RV. Toss 2 dice and map outcome to the sum of the values that show up on each. Here, RV X, is the mapping $X : \{1..6\} \times \{1..6\} \to \{2..12\}$.

(c) Mean, median and mode are one number summary parameters of a distribution. Suppose a philantrophist wants to give money for scholarships to one of the 23 IITs and has the three summary parameters available for the parental income distribution of students. Which should he choose and why (assuming scholarships are given to students who are needy)?

Solution:

The mean is very sensitive to extreme values (e.g. Ambani's children are in some IIT). So, it may give wrong information on whether one IIT is more needy than another. The mode also does not give a good idea of whether one IIT is more needy than another.

The median is the best choice since a lower median indicates that 50% are less than that value. Actually, quartile information can be still better. But given only these three statistics the median is the best of the three.

(d) Here are 3 statements about P-value:

a) P-value is a probability.

b) P-value is calculated assuming the null hypothesis (H_0) is true.

c) To calculate the P-value we must first decide which values of the test statistic are not as extreme as the one obtained from the sample.

One of the above statements is false. Indicate which one and correct it.

Solution:

Statement c) is false. It should be: 'To calculate the P-value we must first decide which values of the test statistic are <u>at least</u> as extreme as the one obtained from the sample.'

(e) What is the central limit theorem and where did we use it?

Solution:

Let $\mathcal{S} = \{X_1, .., X_m\}$ be a sample of size m drawn from a population with mean μ and standard deviation σ. Let $\bar{X}$ be the sample mean. Then the normalized variable $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{m}}}$ approaches a standard normal distribution $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{m}})$ as $m \to \infty$.

CLT was used to show that the distribution of the sample mean is a normal distribution with the sample mean converging to the population mean when the sample size is large enough. This is true even when the original distribution is not normal. In practice sample sizes of approx. 20 suffice even when the original distribution is not normal. This allows us to calculate CIs and is the basis of hypothesis testing.

(f) Fill in the following table with one of the following for each cell in the 2 table:
i) `Correct` ii) `Type I error` iii) `Type II` error.

H_0 in sample	H_0 in population: **False**	**True**
Accept		
Reject		

Solution:

H_0 in sample	H_0 in population: **False**	**True**
Accept	Type II error	Correct
Reject	Correct	Type I error

(g) When are large sized samples needed? Is it always good to have a large sample size? Justify your answer.

Solution:

When a small effect is known to have practical significance or when a study has low power large sample sizes will be needed to determine if the treatment or intervention is statistically significant.

However, large sample sizes can make very small effects statistically significant but often they may not be practically significant.

(h) ANOVA and/or regression can be used to analyse data from experiments that have multiple levels for one or more factors. Which will you use when explain using examples.

Solution:

The factor being investigated can have categorial levels or quantitative levels. ANOVA can be used in both cases but regression is used only when the levels of a factor are quantitative. ANOVA tests whether the different levels of a factor give a statistically significant difference in the effect being studied. For example, oil brands on cholesterol levels in blood.

Regression tries to infer a functional relationship between the predicted or dependent variable and the factor variable(s). This can allow the prediction of values for the dependent variable for currently unknown values of the factor variable values. For example, we can try and find a functional relationship between the amount of oil used and cholesterol levels in blood.

ANOVA will be used when we only want to know whether or not an effect exists. While regression is used if we suspect a functional relationship often in trying to infer causal relations.

[4×8=32]

3. (a) The following program is supposed to print in a single line all numbers between m and n (including both end points, where $m < n$) that are divisible by 3 but not divisible by 7. Some parts of the code have been replaced by `?k?`, where $k = 1, 2, 3, 4$. Supply the code at locations `?k?` so that the program works correctly.

```
def test1(m,n):
    l=[]
    for i in ?1?:
        if ?2?:
```

```
            l.append(?3?)
    print(',',?4?)
```

For example: `test1(2, 15)` will give the output:

```
3,6,9,12,15
```

Solution:

The full function defintion is given below:

```
def test1(m,n):
    l=[]
    for i in range(m,n+1):
        if i%3==0 and i%7!=0:
            l.append(str(i))
    print(',',join(l))
```

(b)

```
def addToList(val, l=[]):
    l.append(val)
    return l

l1=addToList(10)
l2=addToList(20, [])
l3=addToList(30)

print(l1)
print(l2)
print(l3)
```

When the above code fragment was executed the output was:

```
[10, 30]
[20]
[10, 30]
```

The expected output was:

```
[10]
[20]
[30]
```

Explain what actually is happening and change the definition of `addToList` in a way so that it behaves as expected without changing the basic fact that the definition intends to use a default argument.

Solution:

The default argument `l` is bound to `[]` at compile time - so only once. So, when `addToList(10)` executes `l1` becomes a reference to `l` to which `10` has been appended. When `addToList(30)` executes it appends `30` to `l` which already has `10` and `l3` becomes another reference to `l`. Now both `l1, l3` are references to `l` which has the value `[10,30]`.

In the case of `addToList(20,[])` since an explicit second argument is passed it works as expected.

So, default arguments can lead to surprising outcomes since they are bound just once to the default value at compile time. So, need to be careful when assigning default values that are mutable (like list).

The change is to give a default value `None` and initialize `l=[]` in the body so that it behaves as expected. Various other ways will either defeat the spirit of the default argument or give erroneous results. Note that in Python mutable structures (like list) when mutated by functions (like append, extend etc.) do not return any value (i.e. return None).

```
def addToList(val, l=None):
    if l==None:
        l=[]
    l.append(val)
    return l
```

[(2,3,2,3),(5,5)=20]

4. (a) In a building complex 60% of the houses get their internet service from a certain cable company, 80% get their television service from the same company and 50% get both services from the company.

If a random house is selected from the building complex find a) the probability that it gets at least one of the services from the company b) the probability that it gets exactly one service from the company.

Hint: You may find a Venn diagram useful.

Solution:

Let $A \equiv$ event that the random house gets internet from the company and $B \equiv$ event that the random house gets TV service from the company.

We have $P(A) = 0.6$, $P(B) = 0.8$ and $P(A \cap B) = 0.50$.

a) The event we want is $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.8 - 0.5 = 0.9$.

b) The event gets only internet service from the company is $A \cap B^c$, $B^c \equiv$ complement of event B. We have $A \cup B = B \cup (A \cap B^c)$. So, $P(A \cup B) = P(B \cup (A \cap B^c)) = P(B) + P(A \cap B^c)$. The second equality follows because B and $A \cap B^c$ are mutually exclusive. So, $P(A \cap B^c) = P(A \cup B) - P(B) = 0.9 - 0.8 = 0.1$. Similarly, $A^c \cap B$ is the event 'gets only TV service from the company' giving $P(A^c \cap B) = P(A \cup B) - P(A) = 0.9 - 0.6 = 0.3$. The probability of exactly one service is: $P(A^c \cap B) + P(A \cap B^c) = 0.3 + 0.1 = 0.4$.

(b) A shopkeeper sells 3 brands (A, B, C) of mobile phones. His sales figures show Brand A - 50%, Brand B - 30% and Brand C - 20%. Each brand gives a warrany of 1 year for the phone. He also finds that 25% of brand A require repairs during the warranty period. The corresponding figures for brands B and C are 20% and 10% respectively.

i. Find the probability that a randomly selected buyer has bought a brand A phone that will need repair during the warranty period.

Solution:

Let E_a, E_b, E_c be the events that the randomly selected buyer has bought a phone of brand A, brand B and brand C respectively. Their probabilities will be $P(E_a) =$

0.5, $P(E_b) = 0.3$, $P(E_c) = 0.2$. Let R be the event 'phone needs repair'. So, $P(R|E_a) = 0.25$, $P(R|E_b) = 0.20$, $P(R|E_c) = 0.10$.
The probability of the required event is: $P(E_a \cap R) = P(R|E_a)P(E_a) = 0.5 \times 0.25 = 0.125$.

ii. Find the probability that a random buyer's phone will require repair during the warranty period.

Solution:
The probability of the required event is $P(R) = P(R|E_a)P(E_a) + P(R|E_b)P(E_b) + P(R|E_c)P(E_c)$. $P(R) = 0.125 + 0.06 + 0.02 = 0.205$. The second and third terms are calculated in the same way as the first term was in part i.

iii. If a buyer return's to the shop for repair of her/his phone during the warranty period what is the probability that the phone is a brand A phone; probability that it is a brand C phone.

Solution:
The two events whose probabilities are needed are: $E_a|R$ and $E_c|R$. We can use Bayes' theorem or calculate directly using: $P(E_a|R) = \frac{P(E_a \cap R)}{P(R)} = \frac{0.125}{0.205} = 0.61$. Similarly, $P(E_c|R) = \frac{P(E_c \cap R)}{P(R)} = \frac{0.02}{0.205} = 0.098 \approx 0.10$. We are using $P(R)$ and the other probabilities from part ii.

[(3,5),(3,5,4)=20]

5. (a) If X is a RV argue that $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$, where Var(.) is the variance operator and E(.) is the expectation operator.
Hint: $Var(X) = \sum_{x \in \mathcal{S}} (x - E(X))^2 P(x)$. $E(X)$ *is mean* μ *and is a constant.*

Solution:
Note that variance is actually an expectation for the random variable $(x - E(X))^2$.

$$\begin{aligned} Var(X) &= E((X - E(X)^2) \\ &= E(X^2 - 2XE(X) + E((X)^2) \\ &= E(X^2) - E(2XE(X)) + E(E(X)^2) \quad \text{by linearity of expectation} \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \quad \text{linearity and since } E(X) = \mu \text{ is a constant} \\ &= E(X^2) - E(X)^2 \end{aligned}$$

(b) In a shooting experiment the probability of hitting the target (outcome H) is p and of missing the target (outcome M) is q. Once the target is hit shooting stops. Assume all shots are independent.

i. What is the sample space?

Solution:

$\mathcal{S} = \{H, MH, MMH, MMH, \ldots\}$. It is an infinite sample space with each outcome being a sequence of misses (M) followed by a hit (H).

ii. Write the expression for the expectation for RV, X, the number of shots.

Solution:
If we stop after i shots then we have missed $(i-1)$ times and hit at the i^{th} shot. Since each shot is independent the corresponding probability $P(i)$ is $q^{i-1}p$. For expectation we have to consider all possible i so

$$E(X) = \sum_{i=1}^{\infty} i \times P(i) = \sum_{i=1}^{\infty} i \times q^{i-1}p$$

iii. Simplify the above expression using $\sum_{i=1}^{\infty} ix^{i-1} = \frac{1}{(1-x)^2}$.

Solution:
Since $q = (1-p)$ we get:

$$\begin{aligned} E(X) &= \sum_{i=1}^{\infty} i \times (1-p)^{i-1}p \\ &= p\sum_{i=1}^{\infty} i \times (1-p)^{i-1} \\ &= \frac{p}{p^2} \quad \text{using the given identity with } x = 1-p \\ &= \frac{1}{p} \end{aligned}$$

iv. Give a simple intuitive justification for the simplified expression for expectation you got above.

Solution:
If the expectation is k then k shots on average are required to hit the target so the probability of hitting the target $p = \frac{1}{k}$ which means $k = \frac{1}{p}$ - that is expectation is $\frac{1}{p}$.

[6,(3,4,3,4)=20]

6. (a) Normally, we write $\Phi(z)$ for the area under the standard normal curve for $Z \leq z$, that is $P(Z \leq z)$. Write the equations for the P-value for both the one-sided upper and one-sided lower tailed tests and two-sided test for the z-statistic. Assume the computed value for the z-statistic is z.

Solution:
P-value $= (1 - \Phi(z))$ - one-sided upper
P-value $= \Phi(z)$ - one-sided lower

P-value $= 2(1 - \Phi(|z|))$ - two tailed

(b) The target thickness for silicon wafers used in a certain integrated circuit is 245 μm. A sample of 50 wafers is obtained and the thickness of each is measured, giving a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm.

What is the P-value? If wafer thickness limits are important for the integrated circuit what α will you choose? What are H_0, H_a? Does the data above call for rejection of H_0? Justify your answers in each case and show calculation details where needed. Also, clearly state any assumptions you make.

Solution:

Assumptions: We assume a normal population. This is reasonable since most manufactured items have dimensional variations that are normally distributed around the required dimension that is the mean. Since $m = 50$ which is reasonably large we use $\sigma \approx s$ and therefore the z-statistic.

$H_0 \equiv (\mu = 245\mu\text{m})$, $H_a \equiv (\mu \neq 245\mu\text{m})$.

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{(246.18-245)}{3.60/\sqrt{50}} = 2.32$.

Test is two-tailed so: P-value$=2(1 - \Phi(2.32)) = 0.0204$ - using value from standard normal table.

Since wafer thickness is important we choose $\alpha = 0.01$ instead of 0.05. So P-value > 0.01 and H_0 cannot be rejected though at 0.05 it can be rejected.

[(2,2,4),(6,2,2,2)=20]

7. An automobile model is known to sustain no visible damage 25% of the time in 15kmph crash tests. A modified bumper design has been proposed in an effort to increase this percentage. Let p denote the proportion of all crashes with this new bumper that result in no visible damage. The null hypotheses to be tested is: $H_0 : p = 0.25$. The alternate hypothesis is: $H_a : p > 0.25$.

The test will be based on an experiment involving $m = 20$ independent crashes with prototypes of the new design. Intuitively, H_0 should be rejected if a substantial number of the crashes show no damage. Consider the following test procedure:

Test statistic:X = the number of crashes with no visible damage
Rejection region: $x \geq 8$ where x is the observed value of the test statistic. That is: $x \in \{8..20\}$

(a) What is the distribution of X when H_0 is true?

Solution:

X is binomially distributed since in an experiment the car either does not have visible damage or it has. The probability for the event of interest is $p = 0.25$. So, $X \sim B(20, 0.25)$.

(b) What is P(Type I error)?

Solution:

$$
\begin{aligned}
P(\text{Type I error}) &= P(H_0 \text{ is rejected when it is true}) \\
&= P(X \geq 8 \text{ when } X \sim B(20, 0.25)) \\
&= 1 - B(7; 20, 0.25) \\
&= 1 - 0.898 \quad \text{value from Binomial distribution table} \\
&= 0.102
\end{aligned}
$$

(c) What is α?

Solution:
α is the same as Type I error so $\alpha = 0.102$. So this is not significant at the usual 95% level.

(d) For $p = 0.30$ what is the value of β - that is $P(\text{Type II error for } p = 0.30)$?

Solution:

$$
\begin{aligned}
\beta &= P(\text{Type II error when } p = 0.30) \\
&= P(H_0 \text{ is accepted when it is false with } p = 0.30) \\
&= P(X \leq 7 \text{ when } X \sim B(20, 0.30)) \\
&= B(7; 20, 0.30) \\
&= 0.772 \quad \text{value from the Binomial table}
\end{aligned}
$$

(e) Re-calculate α for the rejection region $x \geq 9$ and then recalculate β again for $p = 0.30$.

Solution:

$$
\begin{aligned}
P(\text{Type I error}) &= P(X \geq 9 \text{ when } X \sim B(20, 0.25)) \\
&= 1 - B(8; 20, 0.25) \\
&= 1 - 0.959 \quad \text{value from Binomial distribution table} \\
&= 0.041
\end{aligned}
$$

For β we must now repeat the earlier calculation for $X \leq 8$

$$
\begin{aligned}
\beta &= P(X \leq 8 \text{ when } X \sim B(20, 0.30)) \\
&= B(8; 20, 0.30) \\
&= 0.887 \quad \text{value from the Binomial table}
\end{aligned}
$$

Now though the result is significant at 95% level the probability of type II error is almost 0.9. So, the study is very underpowered. This is to be expected since the alternate value 0.30 is very close to 0.25.

(f) For a fixed alternate value (that is $p = 0.3$) what can you say about the relationship between α and β based on the above. Give an intuitive reason for the relationship.

Solution:
For a fixed alternate value of p we see that α and β are actually complements of each other with respect to the distribution curve. So, if α goes down β goes up and vice versa. So, the only way to improve both numbers simultaneously is to increase the sample size.

[3,5,2,5,(3,3),4=25]