CGS602A: Basic Statistics, Data Analysis and Inference

Mid-semester exam

Max marks: 80
Time:2 hours 18 Oct. 2020

1. *Answer all 4 questions.*
2. *You can use calculators.*
3. *Use the SND (standard normal distribution) table where needed.*
4. *Do not collaborate.*

1. (a) We want to understand the distribution of height in the Indian population. What will be the nature of this distribution? Justify your answer by clearly stating any assumptions you make. In your assumptions stick as closely as possible to approximately what is generally known about the Indian population.

Solution:
The Indian population is a union of many sub-populations. For e.g. people from certain regions of the country on average are taller than other regions. Assume, there are m sub-populations. Within each population we expect average height of males and females to be different so we end up with $2m$ sub-populations. Within a single sub-population it is reasonable to assume that natural attribute like height has a normal distribution. The size of each sub-population in the whole population also matters - let that proportion or weight be α_i for the i^{th} sub-population. Then the population distribution is a mixture of Gaussians with mean μ_i, std. dev. σ_i and mixture weight α_i with $\sum_{i=1}^{2m} \alpha_i = 1$.
The above is assuming only adult population. If children are included then more sub-populations must be considered for different age bands. However, the principle is the same.

(b) Let p be the probability that a head appears in a coin toss. Let the random variable X = number of tosses required to get a head.

i. For any $x = 1, 2, 3 \ldots$ what is $P(X = x)$?

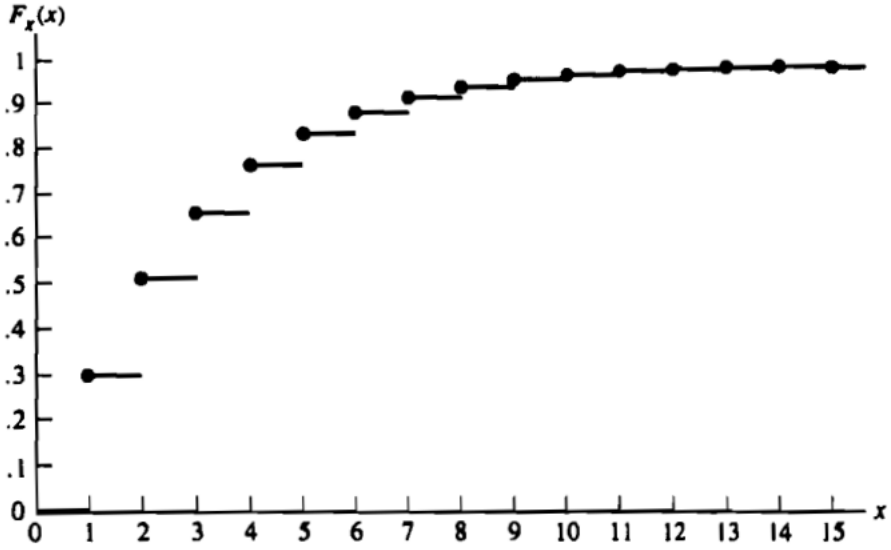
Solution:
If a H appears after x tosses then the first $(x-1)$ tosses were tails. So, $P(X = x) = (1-p)^{(x-1)}p$

ii. What is the cdf, $F_X(x) = P(X \leq x)$? Try and get a nice closed form expression for $F_X(x)$.

Solution:
$P(X \leq x) = \sum_{i=1}^{x} P(X = x) = \sum_{i=1}^{x}(1-p)^{(i-1)}p$. This is a sum of a geometric series $\sum_{i=1}^{x} u^{(i-1)}$ where $u = (1-p)$ that is $\frac{(1-u)^x}{1-u}$ and $u \neq 1$. This gives: $P(X \leq x) = 1-(1-p)^x$, $x = 1, 2, 3 \ldots$.

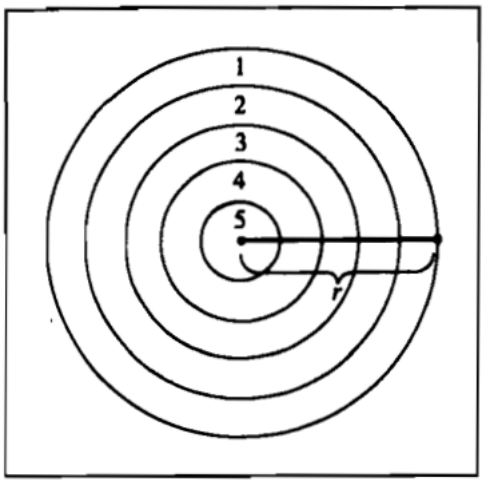
iii. Sketch $F_X(x)$ for $-\infty < x < \infty$.

Solution:
With $p = 0.3$ the figure below shows the cdf. For non-negative values between whole numbers the cdf is unchanged till the next whole number. It is 0 for $-\infty < x < 1$.



[8,(3,(2,3),4)=20]

2. Consider the figure below which shows a dart board on a wall with area A.



Assume when a dart is thrown it will always hit the wall but may or may not hit the board. The numbers indicate the points earned if the throw lands in the corresponding ring. Also, assume that the circles have the radii 1, 2, 3, 4, 5 in some unit (i.e. $r = 5$) and the probability of hitting a ring is proportional to its area. Answer the questions below:

(a) What is the function P(scoring i points) for $i = 0..5$, where 0 points are given if the throw misses the dart board completely and hits the wall?

Solution:
$P(\text{score 0 points}) = 1 - \frac{pi \times 5^2}{A}$. Similarly, $P(\text{scoring 5 points}) = \frac{\pi \times 1^2}{A}$, $P(\text{scoring 4 points}) = \frac{\pi(2^2-1^2)}{A}$, ..., $P(\text{scoring 1 point}) = \frac{\pi(5^2-4^2)}{A}$. The general expression is:

$$P(\text{scoring i points}) = \begin{cases} 1 - \frac{\pi \times 5^2}{A} & i = 0 \\ \pi \times \frac{(6-i)^2-(5-i)^2}{A} & i = 1..5 \end{cases}$$

(b) Calculate the conditional probability $P(\text{scoring i points} \,|\, \text{dart board is hit})$.

Solution:
$P(\text{scoring i points} \,|\, \text{hit dart board}) = \frac{P(\text{scoring i points} \cap \text{hit dart board})}{P(\text{hit dart board})}$. For $i = 1..5$ the event *hit dart board* is a sub-event of *scoring i points*, so $P(\text{scoring i points} \cap \text{hit dart board}) = P(\text{scoring i points}) = \pi \times \frac{(6-i)^2-(5-i)^2}{A}$ and $P(\text{scoring i points} \,|\, \text{hit dart board}) = \frac{A}{\pi 5^2} \times \pi \times \frac{(6-i)^2-(5-i)^2}{A} = \frac{(6-i)^2-(5-i)^2}{5^2}$.

(c) Suppose you are an expert player and never completely miss the dart board. Derive an expression for $P(\text{scoring i points})$, $i = 1..5$.

Solution:
$P(\text{scoring 5 points}) = \frac{\pi \times 1^2}{\pi 5^2}$, $P(\text{scoring 4 points}) = \frac{\pi(2^2-1^2)}{\pi 5^2}$, ..., $P(\text{scoring 1 point}) = \frac{\pi(5^2-4^2)}{\pi 5^2}$. The general expression is: $P(\text{scoring i points}) = \frac{(6-i)^2-(5-i)^2}{5^2}$.

(d) What is the relation between the probability functions in (b) and (c)? Justify briefly.

Solution:
They are identical.
Since the dart board is always hit $P(\text{dart board hit}) = 1$ and the conditional probability $P(\text{scoring i points} \,|\, \text{dart board hit}) = P(\text{scoring i points})$ for $i = 1..5$.

(e) Show that the function $P(\text{scoring i points})$ in (c) is indeed a probability function by showing that it satisfies all the Kolmogorov axioms.

Solution:
$P(\text{scoring i points}) = \frac{(6-i)^2-(5-i)^2}{5^2} \geq 0$ since $1 \leq i \leq 5$. $P(S) = P(\text{dart board hit}) = 1$ and for any $1 \leq i,\ j \leq 5$, $i \neq j$ we have $P(i \cup j) = \frac{\text{area of ring i+area of ring j}}{25} = P(i) + P(j)$. So, all axioms are satisfied and $P(\text{scoring i points})$ is a probability function.

[5,5,5,4,6=25]

3. (a) Suppose you draw a simple random sample of size 20 from a population whose distribution is not known. Can you claim that the sampling distribution of the mean will be a normal distribution? Justify.

Solution:
No.
Since the population distribution is not known and could be very far from normal the minimum size of a simple random sample should be at least 30 to ensure that the sampling distribution is close to normal. With a sample size of 20 we cannot claim that the sampling distribution will be normal/close to normal.

(b) Now you draw 100 simple random samples of size 8 each from a normally distributed population. What can you claim about the nature of the sampling distribution of the mean in this case? Justify.

Solution:
The sampling distribution will be normal since the population is normally distributed, the sample size is 8 which is greater than the minimum sample size limit of 4 or 5 that is necessary for a single sample. In addition the number of samples is 100.

(c) There are 6 identical tiles in a box with numbers 1 to 6 written on them. You select a simple random sample of size 2 without replacement. What will be the probability that the mean of a sample that you choose is 3.0.

Solution:
We have $\binom{6}{2} = 15$ total number of samples of size 2. The following are possible (along with means):
$(1, 2, 1.5), (1, 3, 2), (1, 4, 2.5), (1, 5, 3), (1, 6, 3.5), (2, 3, 2.5),$
$(2, 4, 3), (2, 5, 3.5), (2, 6, 4), (3, 4, 3.5), (3, 5, 4), (3, 6, 4.5),$
$(4, 5, 4.5), (4, 6, 5), (5, 6, 5.5)$. Only two of them have a mean of 3 so the probability is $\frac{2}{15}$.

[4,4,7=15]

4. You are a plant biologist working to improve the size of melons. After measuring a sufficiently large number of melons of the new variety you conclude that the melon diameter D is normally distributed with a mean of $\mu = 11.5cm$ and a standard deviation of $\sigma = 1.15cm$.

(a) What is the probability that the melon diameter is between $10cm$ and $13cm$?

Solution:
We must first transform the values to Z values (that is normalized values) and then use the SND tables to answer all the questions.
$Z_{10} = \frac{10-11.5}{1.15} = -1.304$, $Z_{13} = \frac{13-11.5}{1.15} = 1.304$. From table we get $P(Z \leq 1.30) = .9032$. From symmetry, $P(Z \leq -1.304) = 1 - 0.9032 = .0968$. So, $P(-1.304 \leq Z \leq 1.304) = 0.9032 - 0.0968 = 0.8064$.

(b) To advertise this as an **improved variety** of melon it is required that at least 80% of the melon diameters be between $10cm$ and $13cm$ and at 90% be greater than equal to $10cm$. Can you claim that your new variety is an **improved variety**? Justify.

Solution:
We see that 80.64% are between $10cm$ and $13cm$ and 90.32% are less than equal to $13cm$ so you can claim the variety is an **improved variety**.

(c) If you had made measurements on 5000 melons how many melons will be i) between $10cm$ and $13cm$ in diameter, ii) greater than equal to $10cm$ in diameter iii) less than equal to $13cm$ in diameter?

Solution:
i) Number of melons between $10cm$ and $13cm$ is $5000 \times 0.8064 = 4032$.
ii) Number of melons greater than equal to $10cm$ from symmetry is $5000 \times 0.9032 = 4516$. Again, number of melons greater than equal to $13cm$ is $5000 * .9032 = 4516$. We can say this directly using symmetry of SND.

[Hint: you will need the tables given for the standard normal distribution for this question.] [7,7,(2,2,2)=20]