GE2262 Tutorial

Topic 1: Introduction to Statistics

Q1

For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous.
a) Number of cell phones in a household
b) Length of the longest phone call made in a month
c) Whether the household has a land line
d) Whether there is a high-speed Internet connection in the household

> a) Numerical, discrete
> b) Numerical, continuous
> c) Categorical
> d) Categorical

Graphic evaluation should be discussed from the following aspects:

- **Layout**:
  - Y-axis should start from 0
  - Y-axis scale should be appropriate
  - X-axis and Y-axis scale should be uniform (bins should be of the same width)
  - X-label and Y-label should be included
- **Content**:
  - The title should include the variable name and measurement unit
- **Clarity**:
  - Easy to read

Q3

- Layout
  - X-axis is not uniform. Year 1990-2000-2001-2002-2003-2004 creates a gap of different width.

Q4

- Layout
  - X-axis is not uniform. Year 1925-1939-1949-1967-1971-1975 creates a gap of different width.
  - X-label (Year) is missing
  - Y-label (U.S. Dollars) should be replaced by "GBP/USD"
- Content
  - The title "Decline of British Pound" is unclear
  - It should be "Exchange Rate of British Pound to US Dollar during 1925-1975"

Q5

- Layout
  - X-axis is not uniform
- Clarity
  - The background is unnecessary and statistically misleading (gender ratio is not 1:1)

Q9

Given the sample data: 36.15, 31.00, 35.05, 40.25, 33.75, 43.00

- Mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 36.53$
- Median: $\frac{35.05+36.15}{2} = 35.60$
- Range: $x_{max} - x_{min} = 43.00 - 31.00 = 12.00$
- Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = 19.269$
- Sample standard deviation: $s = \sqrt{s^2} = 4.390$
- Skew: since mean > median, the distribution is right-skewed (positively skewed)

Q10

Given the sample data: 7, 8, 4, 5, 16, 20, 20, 24, 19, 30, 23, 30, 25, 19, 29, 29, 30, 30, 50, 56

- Mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 23.7$
- Median: $\frac{23+24}{2} = 23.5$
- Q1: $16 \cdot (1 - 0.25) + 19 \cdot 0.25 = 16.75$
- Q3: $30 \cdot (1 - 0.75) + 30 \cdot 0.75 = 30$
- Range: $x_{max} - x_{min} = 56 - 4 = 52$
- IQR: $Q_3 - Q_1 = 30 - 16.75 = 13.25$
- Sample variance: $s^2 = 176.116$

- Sample standard deviation: $s = 13.271$
- Skew: since mean > median, the distribution is right-skewed (positively skewed)

Topic 2: Basic Probability

Q1

- The probability that a new car needs warranty repair is 0.04
- The probability that a car is manufactured in US is 0.60
- The probability that a new US car needs warranty repair is 0.025

1. Construct a contingency table for the probabilities that a new car needs warranty repair

	US	Foreign	Total
Needs repair	0.025	0.015	0.04
No repair	0.575	0.385	1.00
Total	0.60	0.40	1.00

2. What is the probability that a new car selected at random

a) needs a warranty repair? (0.04)
b) needs a warranty repair and is manufactured in the US? (0.025)
c) needs a warranty repair, given that it was manufactured in the US? (0.025/0.60=0.0417)
d) was manufactured in the US, given that it needs warranty repair? (0.025/0.04=0.625)

Q2

A survey of 500 consumers was conducted to find out if they enjoyed shopping at a clothing store. The contingency table is given below:

	Male	Female	Total
Enjoyed	126	234	360
Not enjoyed	104	36	140
Total	230	270	500

1. What is the probability that a randomly selected consumer

a) Did not enjoy shopping, given that the respondent is a female?
b) Is a male, given that the respondent enjoyed shopping?

2. Are enjoying shopping and the gender of the consumer independent?
3. a)

$$P(\text{not enjoyed}|\text{female}) = \frac{P(\text{not enjoyed} \cap \text{female})}{P(\text{female})} = \frac{36/500}{270/500} = \frac{36}{270} = 0.1333$$

1. b)

$$P(\text{male}|\text{enjoyed}) = \frac{P(\text{male} \cap \text{enjoyed})}{P(\text{enjoyed})} = \frac{126/500}{360/500} = \frac{126}{360} = 0.3500$$

2.

$$P(\text{not enjoyed}) = \frac{140}{500} = 0.2800$$

Since $P(\text{not enjoyed}) \neq P(\text{not enjoyed}|\text{female})$, the two events are not independent.

Topic 3: Discrete and Continuous Probability Distributions

Q1

Weight distribution concerning age:

Age \ Weight (lb)	100	110	120	130	140
19	0.02	0.09	0.09	0.01	0.02
20	0.06	0.15	α	0.05	0.03
21	0.02	0.06	0.11	0.04	0.05

a) Find the value of α
b) Construct the probability distribution of the weights of these students.
c) A student is selected at random. What do you expect his/her weight to be?
d) What is the standard deviation of the distribution in (b)?
e) Are “age” and “weight” independent? Why or why not?

Solution:

a) $\alpha = 0.2$

b)

Weight (lb)	100	110	120	130	140
Probability	0.1	0.3	0.4	0.1	0.1

c) $\mu = \sum_{i=1}^{n} x_i \cdot P(x_i) = 100 \cdot 0.1 + 110 \cdot 0.3 + 120 \cdot 0.4 + 130 \cdot 0.1 + 140 \cdot 0.1 = 118$

d) $\sigma = \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(x_i)} = \sqrt{(100-118)^2 \cdot 0.1 + \cdots + (140-118)^2 \cdot 0.1} = sqrt116 = 10.77$

e) not independent. Take age=19 and weight=100 as an example,

$P(\text{age} = 19) = 0.23, P(\text{weight} = 100) = 0.10, P(\text{age} = 19 \cap \text{weight} = 100) = 0.02 \neq P(\text{age} = 19) \cdot P(\text{weight} = 100) = 0.023$

Q2

An airline wants to overbook flights in order to reduce the numbers of vacant seats. For a certain flight, it is known that the probabilities of 0, 1, 2 and 3 vacant seats are 0.70, 0.15, 0.10 and 0.05 respectively.

a) Find the mean and standard deviation for the number of vacant seats.
b) What is the expected total number of vacant seats on 100 such flights?

Solution:

a) $\mu = 0.5, \sigma = \sqrt{0.75} = 0.866$

b) $E(X) = 100 \cdot \mu = 50$

Q3

X	$P_A(X)$	$P_B(X)$
0	0.50	0.05
1	0.20	0.10
2	0.15	0.15
3	0.10	0.20
4	0.05	0.50

a) Compute the expected value for each distribution.
b) Compute the standard deviation for each distribution.
c) Compare the results of distributions A and B.

Solution:

$\mu_A = 1, \sigma_A = \sqrt{1.5} = 1.2247$

$\mu_B = 3, \sigma_B = \sqrt{1.5} = 1.2247$

Distribution A and B have same spread but located at different positions.

Distribution A is right-skewed (mean = 1, median = 0.5), while distribution B is left-skewed (mean = 3, median = 3.5).

(right-skewed: long tail on the right, mean > median; left-skewed: long tail on the left, mean < median)

Q4

X	Y	p
-50	-100	0.1
20	50	0.3
100	130	0.4
150	200	0.2

a) For each stock, compute the expected return and the standard deviation of return.
b) Do you think that you will invest in stock X or stock Y? Explain.

Solution:

$\bar{X} = 71, \sigma_X = \sqrt{3829} = 61.8789$

$\bar{Y} = 97, \sigma_Y = \sqrt{7101} = 84.2674$

Stock Y has higher expected return and higher risk. Risk-adverse investors may choose stock X, while risk-tolerant investors may choose stock Y.

Q5

Past records indicate that the probability of customers exceeding their credit limit is 0.05. Suppose that, on a given day, 20 customers place orders. Assume that the number of customers that the AIS detects as having exceeded their credit limit is distributed as a binomial random variable.

a) What are the mean and standard deviation of the number of customers exceeding their credit limits?
b) What is the probability that 0 customers will exceed their limits?
c) What is the probability that 1 customer will exceed his or her limit?
d) What is the probability that 2 or more customers will exceed their limits?

Solution: Let X be the number of customers exceeding their credit limits. $X \sim B(20, 0.05)$

a) $E(X) = np = 20 \cdot 0.05 = 1, \sigma = \sqrt{np(1-p)} = \sqrt{20 \cdot 0.05 \cdot 0.95} = 0.9747$

b) $P(X = 0) = \binom{20}{0} \cdot 0.05^0 \cdot 0.95^{20} = 0.3585$

c) $P(X = 1) = \binom{20}{1} \cdot 0.05^1 \cdot 0.95^{19} = 0.3774$

d) $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.3585 - 0.3774 = 0.2641$

Q6

Consider a sample of 20 customers who visit an e-commerce Web site, and assume that the probability that a customer will leave the site before completing the transaction is 0.88. What is the probability that all 20 of the customers will leave the site without completing a transaction?

Solution: Let X be the number of customers who leave the site without completing a transaction. $X \sim B(20, 0.88)$

$P(X = 20) = \binom{20}{20} \cdot 0.88^{20} \cdot 0.12^0 = 0.0776$

Q7

A task force of CityU sampled 200 students after the mid-term test to ask them whether they went shopping the weekend before the mid-term test or spent the weekend studying, and whether they did well or poorly on the mid-term test. The following result was obtained.

	Did well	Did poorly	Total
Studied	90	10	100
Went shopping	30	70	100
Total	120	80	200

a) What is the probability that a randomly selected student did well on the mid-term test or went shopping the weekend before the mid-term test?
b) A random sample of 10 students is selected. What is the probability that 2 of them did well on mid-term test and studied for mid-term test the weekend before the mid-term test? What distribution are you using? Why can you use such distribution?

Solution:

a)
$P(\text{did well} \cup \text{went shopping}) = P(\text{did well}) + P(\text{went shopping}) - P(\text{did well} \cap \text{went shopping}) = \frac{120}{200} + \frac{100}{200} - \frac{30}{200} = 0.95$
Or $P(\text{did well} \cup \text{went shopping}) = 1 - P(\text{did poorly} \cap \text{studied}) = 1 - \frac{10}{200} = 0.95$

b) Let X be the number of students who did well on mid-term test and studied for mid-term test the weekend before the mid-term test. $X \sim B(10, 0.45)$

$P(X = 2) = \binom{10}{2} \cdot 0.45^2 \cdot 0.55^8 = 0.0763$

Q8

Suppose that 1,000 patrons of a restaurant were asked whether they preferred beer or wine. 70% said that they preferred beer. 60% of patrons were male. 80% of the males preferred beer.

a) What is the probability a randomly selected patron prefers wine?
b) What is the probability a randomly selected patron is female and prefers wine?
c) Suppose a randomly selected patron prefers wine, what is the probability that the patron is a male?
d) Suppose 5 patrons were selected, what is the probability that at least four of them prefer beer?

Solution:

	Beer	Wine	Total
Male	480	120	600
Female	220	180	400
Total	700	300	1000

a) $P(\text{wine}) = 1 - P(\text{beer}) = 0.3$

b) $P(\text{female} \cap \text{wine}) = 1 - [P(\text{male}) + P(\text{beer}) - P(\text{male} \cap \text{bear})] = 1 - (0.6 + 0.7 - 0.48) = 0.18$

c) $P(\text{male}|\text{wine}) = \frac{P(\text{male} \cap \text{wine})}{P(\text{wine})} = \frac{0.12}{0.3} = 0.4$

d) Let X be the number of patrons who prefer beer. $X \sim B(5, 0.7)$

$P(X \geq 4) = P(X = 4) + P(X = 5) = \binom{5}{4} \cdot 0.7^4 \cdot 0.3 + \binom{5}{5} \cdot 0.7^5 = 0.52822$

Q9

MTR Corporation has to conduct surveys regularly to evaluate its service quality. According to previous studies, 87% of the passengers refuse to take part in such surveys.

If 15 passengers are selected randomly, what is the probability that at least 2 of them will respond to the survey?

Solution:

Let X be the number of passengers who respond to the survey. $X \sim B(15, 0.13)$

$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) = 1 - \binom{15}{0} \cdot 0.87^{15} - \binom{15}{1} \cdot 0.13^1 \cdot 0.87^{14} = 0.5987$

Q10

According to Dental Association, 60% of all dentists use nitrous oxide ("laughing gas") in their practice. Let x be the number of dentists who use laughing gas in practice in a random sample of five dentists. The probability distribution of x is as follows:

x	0	1	2	3	4	5
P(x)	0.0102	0.0768	0.2304	0.3456	0.2592	0.0778

a) Find the probability that less than 2 dentists use laughing gas in a sample of five.
b) Find E(x). Interpret the result.
c) Find the standard deviation of x.
d) Based on the results of (b) and (c), show that the distribution of x is binomial with n = 5 and π = 0.6.

Solution:

a) $P(x < 2) = P(x = 0) + P(x = 1) = 0.0102 + 0.0768 = 0.0870$

b) $E(x) = \sum_{i=1}^{n} x_i \cdot P(x_i) = 0 \cdot 0.0102 + 1 \cdot 0.0768 + \cdots + 5 \cdot 0.0778 = 3$

Interpretation: on average, among 5 dentists, 3 of them use laughing gas in practice.

c) $\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(x_i) = 1.1998$

$\sigma = \sqrt{\sigma^2} = 1.0954$

d) Since $E(x) = n \cdot \pi = 5 \cdot 0.6 = 3$ and $\sigma^2 = n \cdot \pi \cdot (1 - \pi) = 5 \cdot 0.6 \cdot 0.4 = 1.2$, $x \sim B(5, 0.6)$

Q11

Given a normal distribution with $\mu = 100, \sigma = 10$, what is the probability that

a) $X > 85$?
b) $X < 80$?
c) $X < 80$ or $X > 110$?
d) $P(100 - c < X < 100 + c) = 0.80$?

Solution:

a) $P(X > 85) = P(Z > \frac{85-100}{10}) = P(Z > -1.5) = 1 - P(Z < -1.5) = 1 - 0.0668 = 0.9332$

b) $P(X < 80) = P(Z < \frac{80-100}{10}) = P(Z < -2) = 0.0228$

c)
$P(X < 80 \cup X > 110) = P(Z < \frac{80-100}{10}) + P(Z > \frac{110-100}{10}) = P(Z < -2) + P(Z > 1) = P(Z < -2) + 1 - P(Z < 1) = 0.0228 + 1 - 0.8413$

d) Let c be the desired value and $z_0 = \frac{c-100}{10}$.

$P(100 - c < X < 100 + c) = 0.80 \Rightarrow P(-z_0 < Z < z_0) = 0.80 \Rightarrow P(Z < z_0) - P(Z < -z_0) = 0.80 \Rightarrow 2P(Z < z_0) - 1 = 0.80 \Rightarrow P(Z < z_0)$

$[100 - c, 100 + c] = [100 - 1.2816 \cdot 10, 100 + 1.2816 \cdot 10] = [87.184, 112.816]$

Q12

The breaking strength of plastic bags used for packaging produce is normally distributed, with a mean of 5 pounds per square inch and a standard deviation of 1.5 pounds per square inch. What proportion of the bags have a breaking strength of

a) Less than 3.11 pounds per square inch?
b) At least 3.8 pounds per square inch?
c) Between 5 and 5.5 pounds per square inch?
d) 95% of the breaking strength will be contained between what two values symmetrically distributed around the mean?

Solution: Let X be the breaking strength of plastic bags. $X \sim N(5, 1.5^2)$

a) $P(X < 3.11) = P(Z < \frac{3.11-5}{1.5}) = P(Z < -1.26) = 0.1038$

b) $P(X \geq 3.8) = 1 - P(X < 3.8) = 1 - P(Z < \frac{3.8-5}{1.5}) = 1 - P(Z < -0.8) = 1 - 0.2119 = 0.7881$

c) $P(5 < X < 5.5) = P(\frac{5-5}{1.5} < Z < \frac{5.5-5}{1.5}) = P(0 < Z < 0.33) = P(Z < 0.33) - P(Z < 0) = 0.6293 - 0.5 = 0.1293$

d) Let c be the desired value and $z_0 = \frac{c-5}{1.5}$.

$P(5-c<X<5+c)=0.95 \Rightarrow P(-z_0<Z<z_0)=0.95 \Rightarrow P(Z<z_0)-P(Z<-z_0)=0.95 \Rightarrow 2P(Z<z_0)-1=0.95 \Rightarrow P(Z<z_0)=0$

$[5-c,5+c]=[5-1.96\cdot 1.5,5+1.96\cdot 1.5]=[2.06,7.94]$

Q13

A statistical analysis of 1,000 long-distance telephone calls made from the headquarters of the Bricks and Clicks Computer Corporation indicates that the length of these calls is normally distributed with μ = 220 seconds and σ = 30 seconds.

a) What is the probability that a call lasted less than 175 seconds?
b) What is the probability that a call lasted between 175 and 265 seconds?
c) What is the probability that a calls lasted between 115 and 175 seconds?
d) What is the length of a call if only 1% of all calls are shorter?

Solution: Let X be the length of the call. $X \sim N(220, 30^2)$

a) $P(X<175)=P(Z<\frac{175-220}{30})=P(Z<-1.5)=0.0668$

b)
$P(175<X<265)=P(\frac{175-220}{30}<Z<\frac{265-220}{30})=P(-1.5<Z<1.5)=P(Z<1.5)-P(Z<-1.5)=0.9332-0.0668=0.8664$

c)
$P(115<X<175)=P(\frac{115-220}{30}<Z<\frac{175-220}{30})=P(-3.5<Z<-1.5)=P(Z<-1.5)-P(Z<-3.5)=0.0668-0.00023=0.06657$

d) Let x_0 be the desired value and $z_0=\frac{x_0-220}{30}$.

$P(X<x_0)=0.01 \Rightarrow P(Z<z_0)=0.01 \Rightarrow z_0=-2.33$

$x_0=220+30\cdot(-2.33)=150.1$

Q14

The exam marks of a large class of students follow a normal distribution with mean μ and standard deviation σ. 1% of the students got 90 or above. 10% of the students got 40 or below. The passing mark is 50.

a) Find the values of μ and σ.
b) Find the chance that a randomly selected student passes the exam.

Solution: Let X be the exam mark of a randomly selected student. $X \sim N(\mu, \sigma^2)$

$P(X\geq 90)=P(Z\geq\frac{90-\mu}{\sigma})=0.01 \Rightarrow \frac{90-\mu}{\sigma}=2.33$

$P(X\leq 40)=P(Z\leq\frac{40-\mu}{\sigma})=0.10 \Rightarrow \frac{40-\mu}{\sigma}=-1.28$

Solving the two equations, we get $\mu=57.73, \sigma=13.85$

b) $P(X\geq 50)=P(Z\geq\frac{50-57.73}{13.85})=P(Z\geq -0.56)=1-P(Z<-0.56)=1-0.2877=0.7123$

Q15

The fill amount of bottles of soft drink has been found to be normally distributed with a mean amount of 2.0 liters and a standard deviation of 0.05 liter. Bottles that contain less than 95% of the listed net content (1.90 liters in this case) can make the manufacturer subject to penalty by the Consumer Council, whereas bottles that have a net content above 2.12 liters may cause excess spillage upon opening.

a) What proportion of the bottles is subject to penalty by the Consumer Council?
b) What proportion of the bottles is risking to excess spillage upon opening?
c) In an effort to reduce the possible penalty due to insufficient net content in the bottles, the manufacturer has set out the following quality control requirement: 99% of bottles should comply with the Consumer Council's standard. To achieve this, the bottler decides to set the filling machine to a new mean amount. Determine the mean amount to be set for the bottle filling machine such that the above requirement can be met.

Solution: Let X be the fill amount of a randomly selected bottle. $X \sim N(2.0, 0.05^2)$

a) $P(X<1.90)=P(Z<\frac{1.90-2.0}{0.05})=P(Z<-2)=0.0228$

b) $P(X>2.12)=P(Z>\frac{2.12-2.0}{0.05})=P(Z>2.4)=0.0082$

c) Let μ_0 be the desired value and $z_0=\frac{1.90-\mu_0}{0.05}$.

$P(X<1.90)=0.01 \Rightarrow P(Z<z_0)=0.01 \Rightarrow z_0=-2.33$

$\mu_0=1.90-0.05\cdot(-2.33)=2.0165$

Q16

At the CityU Computer Service Centre, the loading time for e-Portal page on Internet Explorer is normally distributed with mean 3 seconds.

a) Without doing the calculations, for a randomly selected student, which of the following intervals of loading time (in second) is the most likely to be: 2.9-3.1, 3.1-3.3, 3.3-3.5, 3.5-3.7? Which interval of loading time is the least likely to be? Explain.
b) What is the chance that the loading time is exactly 2 seconds?

Solution:

a) 2.9-3.1 is the most likely to be, 3.5-3.7 is the least likely to be. Since the distribution is symmetric, the interval most close to the mean is the most likely to be, and the interval most far from the mean is the least likely to be.

b) $P(X=2)=0$

Q17

The volume of a randomly selected bottle of a new type of mineral water is known to have a normal distribution with a mean of 995ml and a standard deviation of 5ml. What is the volume that should be stamped on the bottle so that only 3% of bottles are underweight?

Solution: Let X be the volume of a randomly selected bottle. $X \sim N(995, 5^2)$

Let x_0 be the desired value and $z_0 = \frac{x_0 - 995}{5}$.

$P(X < x_0) = 0.03 \Rightarrow P(Z < z_0) = 0.03 \Rightarrow z_0 = -1.88$

$x_0 = 995 + 5 \cdot (-1.88) = 985.6$

Extra Questions

You are playing a best-of-5 series with your friend. The probability of winning a game is 0.6. What is the probability of winning the series?

a) 3 games: 3W out of 3, $P(3W) = 0.6^3 = 0.216$
b) 4 games: 2W1L out of 3 + win the last game, $P(2W1L) \cdot P(W) = \binom{3}{2} \cdot 0.6^2 \cdot 0.4^1 \cdot 0.6 = 0.2592$
c) 5 games: 2W2L out of 4 + win the last game, $P(2W2L) \cdot P(W) = \binom{4}{2} \cdot 0.6^2 \cdot 0.4^2 \cdot 0.6 = 0.20736$

Total probability: $0.216 + 0.2592 + 0.20736 = 0.68256$

Topic 4: Sampling Distributions

Important note: **to prove the sampling distribution of the sample mean is normal**, either of the following conditions should be satisfied:

- **Since the population is normally distributed**, $\bar{X}$ is also normally distributed.
- **Since the sample size is large**, $\bar{X}$ is normally distributed according to CLT.

If the sample size is NOT enough (< 30) and the distribution of the population is NOT given, you should assume the population is normally distributed, in this manner:

The sample is from a unknown population and sample size N < 30, so the CLT does not apply. We assume the population is normally distributed. Therefore, the sampling distribution of the sample mean is also normally distributed.

Q1

Given a normal distribution with $\mu = 100, \sigma = 12$, given $n = 36$, find the probability that the sample mean $\bar{X}$ is:

a) Less than 95?
b) Between 95 and 97.5?
c) Above 102.2?
d) There is a 65% chance that X is above what value?

Solution:

Since the population is normally distributed $X \sim (100, 12^2)$, the sampling distribution of the sample mean is also normally distributed $\bar{X} \sim (100, \frac{12^2}{36}) = (100, 2^2)$.

a) $P(\bar{X} < 95) = P(Z < \frac{95-100}{2}) = P(Z < -2.5) = 0.0062$

b)
$P(95 < \bar{X} < 97.5) = P(\frac{95-100}{2} < Z < \frac{97.5-100}{2}) = P(-2.5 < Z < -1.25) = P(Z < -1.25) - P(Z < -2.5) = 0.1056 - 0.0062 = 0.0994$

c) $P(\bar{X} > 102.2) = P(Z > \frac{102.2-100}{2}) = P(Z > 1.1) = 1 - P(Z < 1.1) = 1 - 0.8643 = 0.1357$

d) Let x_0 be the desired value and $z_0 = \frac{x_0 - 100}{2}$.

$P(\bar{X} > x_0) = 0.65 \Rightarrow P(Z > z_0) = 0.65 \Rightarrow 1 - P(Z < z_0) = 0.65 \Rightarrow P(Z < z_0) = 0.35 \Rightarrow z_0 = -0.3853$

$x_0 = 100 + 2 \cdot (-0.3853) = 99.2294$

Q2

The diameter of a brand of Ping-Pong balls is normally distributed with a mean of 1.30 inches and a standard deviation of 0.05 inches. If a random sample of 25 Ping-Pong balls is selected,

a) What is the sampling distribution of the mean?
b) What is the probability that the sample mean is less than 1.28 inches?
c) What is the probability that the sample mean is between 1.31 and 1.33 inches?
d) The probability is 60% that the sample mean will be between what two values, symmetrically distributed around the mean?

Solution:

a) Since the population is normally distributed, the sampling distribution of the sample mean is also normally distributed.

Given $X \sim N(1.30, 0.05^2)$, and the sample size $n = 25$, the sampling distribution of the sample mean is
$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(1.30, \frac{0.05^2}{25}) = N(1.30, 0.01^2)$.

b) $P(\bar{X} < 1.28) = P(Z < \frac{1.28-1.30}{0.01}) = P(Z < -2) = 0.0228$

c)
$P(1.31 < \bar{X} < 1.33) = P(\frac{1.31-1.30}{0.01} < Z < \frac{1.33-1.30}{0.01}) = P(1 < Z < 3) = P(Z < 3) - P(Z < 1) = 0.99865 - 0.8413 = 0.15735$

d) $P(\mu - k\sigma < \bar{X} < \mu + k\sigma) = 0.60$

$\Rightarrow P(Z < k) - P(Z < -k) = 0.60$

$\Rightarrow 2P(Z < k) - 1 = 0.60$

$\Rightarrow P(Z < k) = 0.80$

$\Rightarrow k = 0.8416$

$\Rightarrow (\mu - k\sigma, \mu + k\sigma) = (1.30 - 0.8416 \cdot 0.01, 1.30 + 0.8416 \cdot 0.01) = (1.2916, 1.3084)$

Q3

Time spent using e-mail per session is normally distributed with $\mu = 8, \sigma = 2$, given $n = 16$,

a) What is the probability that the sample mean is between 7.8 and 8.2 minutes?
b) What is the probability that the sample mean is between 7.5 and 8 minutes?
c) If you select a random sample of 100 sessions, what is the probability that the sample means is between 7.8 and 8.2 minutes?
d) Explain the difference in the results of (a) and (c).

Solution:

Since the population is normally distributed $X \sim (8, 2^2)$, the sampling distribution of the sample mean is also normally distributed $\bar{X} \sim (8, \frac{2^2}{16}) = (8, 0.5^2)$.

a)
$P(7.8 < \bar{X} < 8.2) = P(\frac{7.8-8}{0.5} < Z < \frac{8.2-8}{0.5}) = P(-0.4 < Z < 0.4) = P(Z < 0.4) - P(Z < -0.4) = 0.6554 - 0.3446 = 0.3108$

b) $P(7.5 < \bar{X} < 8) = P(\frac{7.5-8}{0.5} < Z < \frac{8-8}{0.5}) = P(-1 < Z < 0) = P(Z < 0) - P(Z < -1) = 0.5 - 0.1587 = 0.3413$

c) $\bar{X} \sim (8, \frac{2^2}{100}) = (8, 0.2^2)$,
$P(7.8 < \bar{X} < 8.2) = P(\frac{7.8-8}{0.2} < Z < \frac{8.2-8}{0.2}) = P(-1 < Z < 1) = P(Z < 1) - P(Z < -1) = 0.8413 - 0.1587 = 0.6826$

d) With the sample size n increases, the distribution of the sample mean becomes more concentrated around the population mean μ, which means a decrease in the standard deviation of the sampling distribution of the sample mean $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Therefore, the probability that the sample mean is between 7.8 and 8.2 minutes, an interval centered around the population mean, increases.

Q4

Weight distribution concerning age:

Age \ Weight (lb)	100	110	120	130	140
19	0.02	0.09	0.09	0.01	0.02
20	0.06	0.15	α	0.05	0.03
21	0.02	0.06	0.11	0.04	0.05

A sample of 36 first-year students is taken. Find the approximate chance that their total weight is at most 4350 lb.

Solution: $\alpha = 0.20$

Since $n > 30$, the sample mean $\bar{X}$ is normally distributed according to CLT.

Since $\mu = 118, \sigma^2 = 116$, $\bar{X} \sim (118, \frac{116}{36}) = (118, \frac{\sqrt{116}}{6}^2)$.

$P(\bar{X} \leq \frac{4350}{36}) = P(\bar{X} \leq 120.83) = P(Z \leq \frac{120.83-118}{\frac{\sqrt{116}}{6}}) = P(Z \leq 1.58) = 0.9429$

Q5

At the CityU Computer Service Centre, the loading time for e-Portal page on Internet Explorer is normally distributed with mean 3 seconds.
A random sample of 5 computers is drawn. What is the chance that their total loading time is at least 15 seconds?

Solution:

Since X is normally distributed $X \sim (3, \sigma^2)$, the sample mean $\bar{X}$ is also normally distributed $\bar{X} \sim (3, \frac{\sigma^2}{5})$.

$P(\bar{X} \geq \frac{15}{3}) = P(\bar{X} \geq 3) = P(\bar{X} > \mu) = 0.5.$

Q6

A population of size $N = 6$ has the following values: 50, 60, 70, 80, 90, 100. A sample of size $n = 2$ is to be drawn from the population (with replacement). Find the mean and standard deviation of the sampling distribution of the sample mean.

Solution:

First, calculate the mean and standard deviation of the population:

$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{50+60+70+80+90+100}{6} = 75$

$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} = \sqrt{\frac{(50-75)^2+(60-75)^2+(70-75)^2+(80-75)^2+(90-75)^2+(100-75)^2}{6}} = \sqrt{\frac{875}{3}} = 17.08$

Then, calculate the mean and standard deviation of the sampling distribution of the sample mean:

$\mu_{\bar{X}} = \mu = 75$

$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{17.08}{\sqrt{2}} = 12.08$

Thinking question: does $\bar{X}$ follow a normal distribution?

(Answer: since the population is not normally distributed, and the sample size is small, the sampling distribution of the sample mean does not follow a normal distribution.)

Q7

To investigate the length of time working for an employer, researchers at the CityU sampled 344 business students and asked them a question: Over the course of your lifetime, what is the maximum number of years you expect to work for any one employer? The resulting sample had sample mean $\bar{X}$ =19.1 years and sample standard deviation s=6 years. Assume the sample of students was randomly selected from the 5800 undergraduate students in CityU.

a) What are reasonable estimators of population mean and population standard deviation?
b) What is the sampling distribution of X ? Why?
c) If the population mean was 18.5 years, what is P($\bar{X}$ ≥19.1 years)?
d) If the population mean was 19.5, what is P(X =19.1 years)?
e) If P($\bar{X}$ ≥19.1 years) = 0.5, what is the population mean?
f) If P($\bar{X}$ ≥19.1 years) = 0.2, without calculation, can you tell that the population mean is greater or less than 19.1 years? Explain.

Solution:

a) Sample mean and sample standard deviation are reasonable estimators of population mean and population standard deviation.

b) Since the sample size is large, the sampling distribution of the sample mean is normally distributed according to CLT.
$\bar{X} \sim (19.1, \frac{6^2}{344}) = (19.1, (\frac{6}{\sqrt{344}})^2)$

c) $P(\bar{X} \geq 19.1) = P(Z \geq \frac{19.1-18.5}{6/\sqrt{344}}) = P(Z \geq 1.8547) = 1 - P(Z \leq 1.8547) = 1 - 0.9678 = 0.0322$

d) $P(\bar{X} = 19.1) = 0$

e) $P(\bar{X} \geq 19.1) = 0.5 \Rightarrow P(Z \geq \frac{19.1-\mu}{6/\sqrt{344}}) = 0.5 \Rightarrow \frac{19.1-\mu}{6/\sqrt{344}} = 0 \Rightarrow \mu = 19.1$

f) Since the probability that $\bar{X}$ is greater than 19.1 years is 0.2, the population mean is less than 19.1 years. (For a normal distribution, the mean is the median, since $P(\bar{X} \geq 19.1) = 0.2 < 0.5$, $\bar{X} < 19.1$)

Topic 5: Confidence Interval Estimation

Q1

$\bar{X} = 120, \sigma = 24, n = 36$, construct a 99% confidence interval for the population mean.

Known: sample mean $\bar{X} = 120$, population standard deviation $\sigma = 24$, sample size $n = 36$, level of confidence $1 - \alpha = 0.99$.

Step 1 - Since $n > 30$, the sample mean $\bar{X}$ is normally distributed according to CLT. Since σ is known, z-distribution is used.

Step 2 - $\alpha = 0.01, z_{\alpha/2} = z_{0.005} = 2.576$

Step 3 - The 99% CI is $\mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 120 \pm 2.576 \cdot \frac{24}{\sqrt{36}} = 120 \pm 10.304 = [109.696, 130.304]$

Step 4 - We are 99% confident that the population mean is between 109.696 and 130.304.

Q2

A stationery store wants to estimate the mean retail value of greeting cards that it has in its inventory. A random sample of 100 greeting cards indicates a mean value of $2.65 and a standard deviation of $0.44.

Assuming a normal distribution, construct a 95% confidence interval estimate of the mean value of all greeting cards in the store's inventory.

Known: sample mean $\bar{X} = 2.65$, sample standard deviation $s = 0.44$, sample size $n = 100$, level of confidence $1 - \alpha = 0.95$.

Step 1 - Since X is normally distributed, the sample mean $\bar{X}$ is also normally distributed. Since σ is unknown, t-distribution is used.

Step 2 - $\alpha = 0.05, t_{\alpha/2,n-1} = t_{0.025,99} = 1.9842$

Step 3 - The 95% CI is $\mu \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = 2.65 \pm 1.9842 \cdot \frac{0.44}{\sqrt{100}} = 2.65 \pm 0.0873 = [2.5627, 2.7373]$

Step 4 - We are 95% confident that the population mean is between $2.5627 and $2.7373.

Q3

One very important measure of tire performance is the tread wear index, which indicates the tire's resistance to tread wear compared with a tire graded with a base of 100. This means that a tire with a grade of 200 should last twice as long, on average, as a tire graded with a base of 100. A consumer organization wants to estimate the actual tread wear index of a brand name of tires that claims "graded 200" on the sidewall of the tire. A random sample of n=18 indicates a sample mean tread wear index of 195.3 and a sample standard deviation of 21.4.

a) Assuming that the population of tread wear indexes is normally distributed, construct a 95% confidence interval estimate of the population mean tread wear index for tires produced by this manufacturer under this brand name.
b) Do you think that the consumer organization should accuse the manufacturer of producing tires that do not meet the performance information provided on the sidewall of the tire? Explain.
c) Explain why an observed tread wear index of 210 for a particular tire is not unusual, even though it is outside the confidence interval developed in (a).

Known: sample mean $\bar{X} = 195.3$, sample standard deviation $s = 21.4$, sample size $n = 18$, level of confidence $1 - \alpha = 0.95$.

Step 1 - Since X is normally distributed, the sample mean $\bar{X}$ is also normally distributed. Since σ is unknown, t-distribution is used.

Step 2 - $\alpha = 0.05, t_{\alpha/2,n-1} = t_{0.025,17} = 2.1098$

Step 3 - The 95% CI is $\mu \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = 195.3 \pm 2.1098 \cdot \frac{21.4}{\sqrt{18}} = 195.3 \pm 10.6419 = [184.6581, 205.9419]$

Step 4 - We are 95% confident that the population mean is between 184.6581 and 205.9419.

Question b) Since the 95% CI contains 200, the consumer organization should not accuse the manufacturer of producing tires that do not meet the performance information provided on the sidewall of the tire.

Question c) the corresponding z-score of 210 is $\frac{X-\mu}{\sigma} = \frac{210-195.3}{21.4} = 0.6869$. Since X is normally distributed, the closer the z-score is to 0, the more likely the value is to be observed. Since 0.6869 is close to 0, the observed tread wear index of 210 for a particular tire is not unusual.

Q4

A food inspector, examining 10 bottles of a certain brand of honey, obtained the following percentages of impurities:

($n = 10, \bar{X} = 21.25, s = 1.9896$)

With 95% confidence, what is the sampling error if the inspector used the sample mean to estimate the mean percentage of impurities in this brand of honey?

Known: sample mean $\bar{X} = 21.25$, sample standard deviation $s = 1.9896$, sample size $n = 10$, level of confidence $1 - \alpha = 0.95$.

Step 1 - Since $n < 30$, CLT does not apply. We have to assume the population is normally distributed. Therefore, the sampling distribution of the sample mean is also normally distributed. Since σ is unknown, t-distribution is used.

Step 2 - $\alpha = 0.05, t_{\alpha/2,n-1} = t_{0.025,9} = 2.2622$

Step 3 - Margin of error $E = t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = 2.2622 \cdot \frac{1.9896}{\sqrt{10}} = 1.4233$

Q5

A sample of 12 observations is obtained from an infinite population with normal distribution. Based on the sample data, the 95% confidence interval for the population mean is calculated to be [20, 30]. Form this 95% confidence interval, determine the mean and standard deviation of the sample.

Known: sample size $n = 12$, level of confidence $1 - \alpha = 0.95$, 95% CI $[20, 30]$.

Step 0 - Let $\bar{X}, s$ be the sample mean and standard deviation, respectively.

Step 1 - Since X is normally distributed, the sample mean $\bar{X}$ is also normally distributed. Since σ is unknown, t-distribution is used.

Step 2 - $\alpha = 0.05, t_{\alpha/2,n-1} = t_{0.025,11} = 2.2010$

Step 3 - The 95% CI is $\mu \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = \bar{X} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = \bar{X} \pm 2.2010 \cdot \frac{s}{\sqrt{12}} = [20, 30]$

Therefore, $\bar{X} - 2.2010 \cdot \frac{s}{\sqrt{12}} = 20$ and $\bar{X} + 2.2010 \cdot \frac{s}{\sqrt{12}} = 30$

Solve the equations to get $\bar{X} = 25$ and $s = 7.8694$

Q6

An advertising agency that serves a major radio station wants to estimate the mean amount of time that the station's audience spends listening to the radio on a daily basis. From past studies, the standard deviation is estimated as 45 minutes. Assume the population has the normal distribution.

a) What sample size is needed if the agency wants to be 90% confident of being correct to within ±10 minutes?
b) If 99% confidence is desired, how many listeners need to be selected?

Question a) Known: population standard deviation $\sigma = 45$, level of confidence $1 - \alpha = 0.90$, margin of error $E = 10$.

Step 1 - Since the population is normally distributed, the sampling distribution of the sample mean is also normally distributed. Since σ is known, z-distribution is used.

Step 2 - $\alpha = 0.10, z_{\alpha/2} = z_{0.05} = 1.645$

Step 3 - Margin of error $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{1.645 \cdot 45}{10}\right)^2 = 54.7970 \approx 55$

Step 4 - The sample size needed is 55.

Question b)

Step 2 - $\alpha = 0.01, z_{\alpha/2} = z_{0.005} = 2.576$

Step 3 - Margin of error $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{2.576 \cdot 45}{10}\right)^2 = 134.3745 \approx 135$

Step 4 - The sample size needed is 135.

Q7

The personal director of a large corporation wants to study absenteeism among clerical
workers at the corporation's central office during the year. A random sample of 25 clerical workers reveals the following:

$\bar{X} = 9.7, s = 4.0$

a) Set up a 95% confidence interval estimate of the average number of absences for clerical workers. Give a practical interpretation of the interval obtained.
b) What assumption must hold in order to perform the estimation in (a)?
c) If the personnel director also wants to take a survey in a branch office, what sample size is needed if the director wants to be 95% confident of being correct to within ±1.5 days and the population standard deviation is assumed to be 4.5 days?

Known: sample mean $\bar{X} = 9.7$, sample standard deviation $s = 4.0$, sample size $n = 25$, level of confidence $1 - \alpha = 0.95$.

Step 1 - Since $n < 30$, CLT does not apply. We have to assume the population is normally distributed. Therefore, the sampling distribution of the sample mean is also normally distributed. Since σ is unknown, t-distribution is used.

Question b) Two assumptions must hold:

- The population is normally distributed.
- The population standard deviation σ is equal to the sample standard deviation s.

Step 2 - $\alpha = 0.05, t_{\alpha/2,n-1} = t_{0.025,24} = 2.0639$

Step 3 - The 95% CI is $\mu \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = 9.7 \pm 2.0639 \cdot \frac{4.0}{\sqrt{25}} = 9.7 \pm 1.6511 = [8.0489, 11.3511]$

Step 4 - We are 95% confident that the average number of days of absence for clerical workers is between 8.0489 and 11.3511.

Question c) Known: population standard deviation $\sigma = 4.5$, level of confidence $1 - \alpha = 0.95$, margin of error $E = 1.5$.

Step 1 - Since no sample size is given, CLT does not apply. We have to assume the population is normally distributed. Therefore, the sampling distribution of the sample mean is also normally distributed. Since σ is known, z-distribution is used.

Step 2 - $\alpha = 0.05, z_{\alpha/2} = z_{0.025} = 1.96$

Step 3 - Margin of error $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{1.96 \cdot 4.5}{1.5}\right)^2 = 34.5744 \approx 35$

Step 4 - The sample size needed is 35.

Q8

From past experience, the numbers of vitamin supplements sold per day in a health food store well fit a normal distribution with variance 9. The number of vitamin supplements sold per day in a sample of 11 days is obtained:

($n = 1, \bar{X} = 22.8182, s = 3.7099$)

a) Construct a 95% confidence interval for population average number of vitamin supplements sold per day in the store.
b) Someone suggests constructing the confidence interval by t-distribution. Do you agree? Explain your opinion.
c) The data analyst found that there is a data-entry mistake. The last observation should be smaller than 30. How will this affect the confidence interval?
d) If the company would like to limit the sampling error within ±0.5, what is the least sample size needed for a 90% confidence interval?

Question a) Known: sample mean $\bar{X} = 22.8182$, sample standard deviation $s = 3.7099$, sample size $n = 11$, level of confidence $1 - \alpha = 0.95$.

Population variance $\sigma^2 = 9$, so population standard deviation $\sigma = 3$.

Step 1 - Since X is normally distributed, the sample mean $\bar{X}$ is also normally distributed. Since σ is known, z-distribution is used.

Step 2 - $\alpha = 0.05, z_{\alpha/2} = z_{0.025} = 1.96$

Step 3 - The 95% CI is $\mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 22.8182 \pm 1.96 \cdot \frac{3}{\sqrt{11}} = 22.8182 \pm 1.7729 = [21.0453, 24.5911]$

Step 4 - We are 95% confident that the population average number of vitamin supplements sold per day in the store is between 21.0453 and 24.5911.

Question b) Since the population variance is known, the z-distribution should be used.

Sample standard deviation is less reliable than population standard deviation due to the small sample size. Therefore, t-distribution should not be used.

Question c) If the last entry is smaller than 30, the sample mean will be smaller. However, since σ is known, the margin of error will not change.

The change of CI is that, both the lower and upper bounds will be smaller, but the length of the interval will not change.

Question d) Known: level of confidence $1 - \alpha = 0.90$, margin of error $E = 0.5, \sigma = 3$.

Step 2 - $\alpha = 0.10, z_{\alpha/2} = z_{0.05} = 1.645$

Step 3 - Margin of error $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{1.645 \cdot 3}{0.5}\right)^2 = 97.4169 \approx 98$

Step 4 - The least sample size needed is 98.

Topic 6: Hypothesis Testing

Q1

Business Week reported that at the top 50 business schools, students studied an average of 14.6 hours. Set up a hypothesis test to try to prove that the mean number of hours studied at your school is different from the 14.6 hour benchmark reported by Business Week.

- Null hypothesis: $H_0 : \mu = 14.6$
- Alternative hypothesis: $H_1 : \mu \neq 14.6$

Decision \ Truth	**H_0 true**	**H_0 false**
Don't Reject H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

- Type I error: Given that H_0 is true, you still reject H_0.
- Type II error: Given that H_0 is false, you fail to reject H_0.

The cost of Type I error and Type II error depends on the context of the problem. In this problem, there is no specific cost. Since the researchers have less control over Type II error, α should be set to as small as possible to minimize the probability of Type I error.

Q2

Suppose that at a particular branch the population mean amount of money withdrawn from ATMs per customer transaction over the weekend is $160 with a population standard deviation of 30$.

a) If a random sample of 36 customer transactions is examined and the sample mean withdrawal is $148, is there evidence to believe that the population average withdrawal is less than 160$? (Use a 0.05 level of significance.)

b) Compute the p-value and interpret its meaning.

Known: sample mean $\bar{X} = 148$, population mean $\mu = 160$, population standard deviation $\sigma = 30$, sample size $n = 36$, level of significance $\alpha = 0.05$.

Step 0 - Denote X as the amount of money withdrawn from ATMs per customer transaction over the weekend. (Optional, but if there are multiple variables, it is better to define them first.)

Step 1 - Despite X is from an unknown distribution, since $n > 30$, the sample mean $\bar{X}$ is normally distributed according to CLT. Since $\sigma = 30$ is known, z-distribution is used.

Step 2 - Define the hypotheses: $H_0 : \mu \geq 160$, $H_1 : \mu < 160$. (Note the question asks for "less than", which does not contain the equal sign. So the result should be inverted.)

Step 3 - $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{148 - 160}{30/\sqrt{36}} = -2.4$

Step 4 - Make the decision:

(a) Critical value:

The lower-tail test critical value is $z_\alpha = 1.645$.

Since $Z = -2.4 < z_\alpha = -1.645$, reject H_0.

(b) P-value:

The p-value is $p = P(Z < -2.4) = 0.0082$.

Since $p = 0.0082 < \alpha = 0.05$, reject H_0.

The meaning of p-value: the probability of observing a sample mean as extreme as the one observed, assuming H_0 is true.

In this example: The probability of observing a sample mean $\bar{X} \leq 148$, given that the average withdrawal is at least \$160, is 0.0082.

Step 5 - (H_1 is not rejected) There is evidence to believe that the population average withdrawal is less than \$160.

Q3

The mean life of a certain electric tube is required to be not less than 1200 hours. The standard deviation based on past experience is 150 hours. A lot is received from a supplier, and 25 specimens of the tubes have been tested. The mean life of the test tubes is found to be 1080 hours. Should the lot be accepted at the 5% level of significance under the assumption of normal population?

Known: sample mean $\bar{X} = 1080$, population mean $\mu = 1200$, population standard deviation $\sigma = 150$, sample size $n = 25$, level of significance $\alpha = 0.05$.

Step 0 - Let X be the life of a certain electric tube.

Step 1 - Since $n > 30$, the sample mean $\bar{X}$ is normally distributed according to CLT. Since $\sigma = 150$ is known, z-distribution is used.

Step 2 - Define the hypotheses: $H_0 : \mu \geq 1200$, $H_1 : \mu < 1200$. ("not less than" $\rightarrow H_0 : \mu \geq 1200$)

Step 3 - $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1080 - 1200}{150/\sqrt{25}} = -4.0$

Step 4 - Make the decision:

(a) Critical value:

The lower-tail test critical value is $z_\alpha = 1.645$.

Since $Z = -4.0 < z_\alpha = -1.645$, reject H_0.

(b) P-value:

The p-value is $p = P(Z < -4.0) = 0.00003$.

Since $p = 0.00003 < \alpha = 0.05$, reject H_0.

Step 5 - The lot should be rejected.

Q4

A manufacturer of chocolate candies uses machines to package candies as they move along a filling line. Although the packages are labeled as 8 ounces, the company wants the packages to contain a mean of 8.17 ounces so that virtually none of the packages contain less than 8 ounces. A sample of 50 packages is selected periodically, and the packaging process is stopped if there is evidence that the mean amount packaged is different from 8.17 ounces. Suppose that in a particular sample of 50 packages, the mean amount dispensed is 8.159 ounces, with a sample standard deviation of 0.051 ounce.

a) Is there evidence that the population mean amount is different from 8.17 ounces? (Use a 0.05 level of significance.)
b) Compute the p-value and interpret its meaning.

Known: sample mean $\bar{X} = 8.159$, population mean $\mu = 8.17$, sample standard deviation $s = 0.051$, sample size $n = 50$, level of significance $\alpha = 0.05$.

Step 0 - Let X be theamount of chocolate candies in a package.

Step 1 - Since $n > 30$, the sample mean $\bar{X}$ is normally distributed according to CLT. Since σ is unknown, t-distribution is used.

Step 2 - Define the hypotheses: $H_0 : \mu = 8.17$, $H_1 : \mu \neq 8.17$. ("different from" $\rightarrow H_1 : \mu \neq 8.17$)

Step 3 - $t = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{8.159 - 8.17}{0.051/\sqrt{50}} = 1.5251$

Step 4 - Make the decision:

(a) Critical value:

The double-tail test critical value is $t_{\alpha/2, n-1} = t_{0.025, 49} = 2.0096$.

The rejection region is $t < -2.0096$ or $t > 2.0096$.

Since t is not in the rejection region, don't reject H_0.

(b) P-value:

The p-value is $p = P(T < -1.5251) + P(T > 1.5251) = 2P(T > 1.5251) = 0.1333$. (You need a p-distribution calculator to get the value. t-table does not yield the exact value.)

The meaning of p-value: the probability of observing a sample mean $\bar{X}$ more than 8.159 or less than $2\mu - \bar{X} = 8.181$, given that the average amount is 8.17, is 0.1333.

Since $p = 0.1333 > \alpha = 0.05$, don't reject H_0.

Step 5 - (H_1 is rejected) There is no evidence that the population mean amount is different from 8.17 ounces.

Q5

The Glen Valley Steel Company manufactures steel bars. If the production process is
working properly, it turns out that steel bars are normally distributed with mean length of at least 2.8 feet. Longer steel bars can be used or altered, but shorter bars must be scrapped. You select a sample of 25 bars, and the mean length is 2.73 feet and the sample standard deviation is 0.20 feet. Do you need to adjust the production equipment? ($\alpha = 0.05$)

Known: sample mean $\bar{X} = 2.73$, population mean $\mu = 2.8$, sample standard deviation $s = 0.20$, sample size $n = 25$, level of significance $\alpha = 0.05$.
Solution:

Step 0 - Let X be the length of a steel bar.

Step 1 - Since X is normally distributed, the sample mean $\bar{X}$ is also normally distributed.

Step 2 - Define the hypotheses: $H_0 : \mu \geq 2.8$, $H_1 : \mu < 2.8$. ("at least" $\rightarrow H_0 : \mu \geq 2.8$)

Step 3 - $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{2.73 - 2.8}{0.20/\sqrt{25}} = -1.75$

Step 4 - Make the decision:

(a) Critical value:

The lower-tail test critical value is $z_\alpha = 1.645$.

The rejection region is $z < -1.645$.

Since $Z < -1.645$, reject H_0.

(b) P-value:

The p-value is $p = P(Z < -1.75) = 0.0401$.

Since $p = 0.0401 < \alpha = 0.05$, reject H_0.

Step 5 - The production equipment needs to be adjusted.

Q6

A bank branch located in a commercial district of a city has developed an improved process for serving customers during the 12:00 to 1 p.m. peak lunch period. The waiting time in minutes (operationally defined as the time the customer enters the line to the time he or she is served) of all customers during this hour is recorded over a period of a week. A random sample of 15 customers is selected, and the results are as follows:

($n = 15, \bar{X} = \frac{643}{150} = 4.2867, s = 1.6380$).

At the 0.05 level of significance, is there evidence that the average waiting time at a bank branch in a commercial district of the city is less than five minutes during the lunch period?

Known: sample mean $\bar{X} = 4.2867$, population mean $\mu = 5$, sample standard deviation $s = 1.6380$, sample size $n = 15$, level of significance $\alpha = 0.05$.

Step 0 - Let X be the waiting time in minutes.

Step 1 - Since $n < 30$, CLT does not apply. We have to assume X is normally distributed. Therefore, the sample mean $\bar{X}$ is also normally distributed. Since σ is unknown, t-distribution is used.

Step 2 - Define the hypotheses: $H_0 : \mu \geq 5$, $H_1 : \mu < 5$. ("less than" $\rightarrow H_1 : \mu < 5$)

Step 3 - $t = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{4.2867 - 5}{1.6380/\sqrt{15}} = -1.6866$

Step 4 - Make the decision: (critical value)

The lower-tail test critical value is $t_{\alpha, n-1} = t_{0.05,14} = -1.7613$.

The rejection region is $t < -1.7613$.

Since $t > -1.7613$, don't reject H_0.

Step 5 - (H_1 is rejected) There is no evidence that the average waiting time at a bank branch in a commercial district of the city is less than five minutes during the lunch period.

Q7

A television documentary on over-eating claimed that Americans are about 10 pounds
overweight on average. To test this claim, 18 randomly selected individuals were examined, and their average excess weight was found to be 12.4 pounds, with a sample standard deviation of 2.7 pounds.

a) What assumption(s) is(are) required for performing the hypothesis testing in (ii) below?
b) At a significance level of 0.01, is there any reason to doubt the validity of the claimed 10-pound value?
c) Define the probability of type I error α and that of type II error β according to the context of this part.

Known: sample mean $\bar{X} = 12.4$, population mean $\mu = 10$, sample standard deviation $s = 2.7$, sample size $n = 18$, level of significance $\alpha = 0.01$.

Step 0 - Let X be the excess weight in pounds.

Step 1 - Since $n < 30$, CLT does not apply. We have to assume X is normally distributed. Therefore, the sample mean $\bar{X}$ is also normally distributed. Since σ is unknown, t-distribution is used.

Question a) Two assumptions are made:

- The population is normally distributed.
- The population standard deviation σ is estimated by the sample standard deviation s.

Step 2 - Define the hypotheses: $H_0 : \mu = 10$, $H_1 : \mu \neq 10$. ("about", implies "equal" $\rightarrow H_0 : \mu = 10$)

Step 3 - $t = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{12.4 - 10}{2.7/\sqrt{18}} = 3.7712$

Step 4 - Make the decision: (critical value)

The double-tail test critical value is $t_{\alpha/2, n-1} = t_{0.005,17} = 2.8982$.

The rejection region is $t < -2.8982$ or $t > 2.8982$.

Since $t > 2.8982$, reject H_0.

Step 5 - There is reason to doubt the validity of the claimed 10-pound value.

Question c)

- Type I error: Given that the average excess weight is 10 pounds, you still doubt the validity of the claimed 10-pound value.
- Type II error: Given that the average excess weight is not 10 pounds, you fail to doubt the validity of the claimed 10-pound value.

Q8

A management consultant has introduced new procedures to a reception office. He claims
that the receptionist should not do more than 10 minutes of paperwork in each hour. A check is made on 40 random hours of operation. The sample mean and sample standard deviation of the time spent on paperwork are found. Based on these figures, the null hypothesis that the new procedures meet specifications is rejected at a 1% level of significance.

a) After the consultant has asked the data entry clerk to show him the original data, he finds that the sample size should be 41, instead of 40. Should the null hypothesis that the new procedures meet specifications be rejected? Why or why not?

b) Peter, the manager of the reception office, asks the consultant to test the same
hypothesis with a new level of significance of 5%. Should the null hypothesis that the
new procedures meet specifications be rejected? Why of why not?

Known: population mean $\mu = 10$, sample size $n = 40$, level of significance $\alpha = 0.01$.

Consider the original hypothesis test:

Step 0 - Let X be the time spent on paperwork in minutes per hour.

Step 1 - Since $n > 30$, the sample mean $\bar{X}$ is normally distributed according to CLT. Since σ is unknown, t-distribution is used.

Step 2 - Define the hypotheses: $H_0 : \mu \leq 10$, $H_1 : \mu > 10$. ("not more than" $\rightarrow H_0 : \mu \leq 10$)

Step 3 - $t = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\bar{X} - 10}{s/\sqrt{40}}$

Step 4 - Make the decision: (critical value)

The upper-tail test critical value is $t_{\alpha, n-1} = t_{0.01, 39}$

The rejection region is $t > t_{0.01, 39}$.

Since the null hypothesis is rejected, we know that $t > t_{0.01, 39}$.

Question a)

In this case, t changes to $t' = \frac{\bar{X} - 10}{s/\sqrt{41}} > \frac{\bar{X} - 10}{s/\sqrt{40}} = t$.

On the other hand, as n increases, $t_{\alpha, n-1}$ decreases. Therefore, $t_{0.01, 41-1} < t_{0.01, 40-1}$.

Therefore: $t' > t > t_{0.01, 40-1} > t_{0.01, 41-1}$.

Since $t' > t_{0.01, 41-1}$, reject H_0.

Question b)

The upper-tail test critical value is $t_{\alpha, n-1} = t_{0.05, 40-1}$.

As α increases, $t_{\alpha, n-1}$ decreases. Therefore, $t_{0.05, 40-1} < t_{0.01, 40-1}$.

(Note: $t_{\alpha, n-1}$ is a decreasing function of α because, as α increases, the area in the tail increases, and the critical value decreases.)

Therefore: $t > t_{0.01, 40-1} > t_{0.05, 40-1}$.

Since $t > t_{0.05, 40-1}$, reject H_0.

Q9

A retail chain knows that on average, sales in its stores are 20% higher in June than in May. For the current year, a random sample of six stores was selected. Their percentage June sales increases were found to be

($n = 6, \bar{X} = 19.5, s = 0.7668$).

a) Test the null hypothesis that the true mean percentage sales increase is 20, against the two-sided alternative, at the 10% significant level. What assumption is required?
b) What is the final decision if one wants to test whether the true mean percentage sales increase more than 20%? Explain without computation.

Known: sample mean $\bar{X} = 19.5$, population mean $\mu = 20$, sample standard deviation $s = 0.7668$, sample size $n = 6$, level of significance $\alpha = 0.10$.

Step 0 - Let X be the percentage sales increase.

Step 1 - Since $n < 30$, CLT does not apply. We have to assume X is normally distributed. Therefore, the sample mean $\bar{X}$ is also normally distributed. Since σ is unknown, t-distribution is used.

Question a) Two assumptions are made:

- The population is normally distributed.
- The population standard deviation σ is estimated by the sample standard deviation s.

Step 2 - Define the hypotheses: $H_0 : \mu = 20$, $H_1 : \mu \neq 20$.

Step 3 - $t = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{19.5 - 20}{0.7668/\sqrt{6}} = -1.5972$

Step 4 - Make the decision: (critical value)

The double-tail test critical value is $t_{\alpha/2, n-1} = t_{0.05, 5} = 2.5705$.

The rejection region is $t < -2.5705$ or $t > 2.5705$.

Since t is not in the rejection region, don't reject H_0.

Step 5 - The null hypothesis that the true mean percentage sales increase is 20\% is not rejected.

Question b) Define hypotheses: $H_0 : \mu \leq 20$, $H_1 : \mu > 20$. ("more than" $\rightarrow H_1 : \mu > 20$)

The rejection region is $t > t_{\alpha, n-1} = t_{0.10, 5}$, which is a positive value. Since t is negative, H_0 is not rejected.

Therefore, H_1 is rejected. There is no evidence that the true mean percentage sales increase more than 20\%.

Topic 7: Proportion

Takeaways:

- sample standard deviation (standard error): $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$
- z-statistic: $z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$
- $100(1-\alpha)\%$ CI: $p \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$
- Sample size: $E = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \Rightarrow n = \frac{z_{\alpha/2}^2 \cdot p(1-p)}{E^2}$
- Hypothesis test: $H_0 : \pi = \pi_0$, $H_1 : \pi \neq \pi_0$, $Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

Q1

A survey of 40 college students showed that 8 of them own shares of stock.

a) Find the sample proportion of college students who own shares.
b) Find the standard error of the sample proportion.

Solution:

a) $p = \frac{X}{n} = \frac{8}{40} = 0.20$

b) $\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20 \cdot 0.80}{40}} = 0.0632$

Q2

You plan to conduct a marketing experiment in which students are to taste one of two different brands of soft drink. Their task is to correctly identify the brand tasted. You select a random sample of 200 students and assume that the students have no ability to distinguish between the two brands. (Hint: If an individual has no ability to distinguish between the two soft drinks, then each brand is equally likely to be selected.)

a) What is the probability that the sample will have between 50% and 60% of the identifications correct?
b) The probability is 90% that the sample percentage is contained within what symmetrical limits of the population percentage?
c) What is the probability that the sample percentage of correct identifications is greater than 65%?
d) Which is more likely to occur – more than 60% correct identifications in the sample of 200 or more than 55% correct identifications in a sample of 1,000? Explain.

Known: sample size $n = 200$, population proportion $\pi = 0.5$.

Step 1 - Since $n > 30$ and $np > 5$ and $n(1-p) > 5$, the sample proportion p is normally distributed according to CLT.

a) Step 2 - $z_1 = \frac{0.50 - 0.50}{\sqrt{\frac{0.50 \cdot 0.50}{200}}} = 0$, $z_2 = \frac{0.60 - 0.50}{\sqrt{\frac{0.50 \cdot 0.50}{200}}} = 2.8284$

Step 3 - $P(0.50 < p < 0.60) = P(0 < z < 2.8284) = P(z < 2.8284) - P(z < 0) = 0.9977 - 0.5000 = 0.4977$

b) $P(0.5 - k\sigma_p < p < 0.5 + k\sigma_p) = 0.90 \Rightarrow P(-k < z < k) = 0.90$

$P(-k < z < k) = P(z < k) - P(z < -k) = P(z < k) - (1 - P(z < k)) = 2P(z < k) - 1 = 0.90$

$P(z < k) = 0.95 \Rightarrow k = 1.645$

$0.5 \pm 1.645 \cdot \sqrt{\frac{0.50 \cdot 0.50}{200}} = 0.5 \pm 0.0582 = [0.4418, 0.5582]$

c) Step 2 - $z = \frac{0.65 - 0.50}{\sqrt{\frac{0.50 \cdot 0.50}{200}}} = 4.2426$

Step 3 - $P(p > 0.65) = P(z > 4.2426) = 1 - P(z < 4.2426) = 1 - 0.99999 = 0.00001$

d) Step 1 - In both cases, the sample proportion p is normally distributed according to CLT.

For $n = 200, p = 0.60$: Step 2 - $z = \frac{0.60 - 0.50}{\sqrt{\frac{0.50 \cdot 0.50}{200}}} = 2.8284$

Step 3 - $P(p > 0.60) = P(z > 2.8284) = 1 - P(z < 2.8284) = 1 - 0.9977 = 0.0023$

For $n = 1,000, p = 0.55$: Step 2 - $z = \frac{0.55 - 0.50}{\sqrt{\frac{0.50 \cdot 0.50}{1,000}}} = 34.7850$

Step 3 - $P(p > 0.55) = P(z > 34.7850) = 1 - P(z < 34.7850) = 1 - 1 = 0$

Therefore, it is more likely to occur that more than 60% correct identifications in the sample of 200.