GE2262 Business Statistics

1. Introduction to Statistics

Terminology

- **Statistics**: transform data into useful information for decision making
- **Descriptive statistics**: collect, summarize (table and graph), and describe (central tendency and dispersion) data
- **Inferential statistics**: use sample data to make inferences about a population

Population and sample:

- **Population**: the entire group of interest
- **Sample**: a subset of the population
- **Parameter**: a numerical measure that describes a population
- **Statistic**: a numerical measure that describes a sample

Process of a statistical study:

1. Find **population**
2. Draw a **sample** from the population
3. Collect data from the sample, and conduct **sample statistics**
4. Use sample statistics to infer **population parameters**

Application of statistics:

- Audit sampling (determine the sample size)
- Economic indicators (e.g. GDP, unemployment rate)
- Risk and portfolio management (e.g. stock price, return, risk)
- Big data

Types of Data

- **Variable**: any characteristic, number, or quantity that can be measured or counted.
- **Data**: the values (measurements or observations) for variables. Data is the specific value of a variable
  - **Categorial data**: data that can be grouped by specific categories
  - **Numerical data**: data that can be measured by numbers
    - **Discrete data**: data that can only take on certain values (e.g. number of students in a class)
    - **Continuous data**: data that can take on any value within a range (e.g. height, weight)
    - Discrete data are usually counted, while continuous data are usually measured
  - Ordinal data: data that can be ordered. They share the properties of both categorical and numerical data (e.g. satisfaction level from 1 to 5)

Coding categorical data with numbers -> numerical data.

Describing different types of data:

- Categories: summary table, pie chart, bar chart
- Numerical: frequency distribution, histogram

Note: x-axis of a bar chart is categories, while x-axis of a histogram is values (or numerical ranges)

Frequency: the number of times a value occurs in a data set (i.e. count, appearance). Frequency is NOT a percentage.

Relative frequency = frequency / sample size

Graphs

Bar Chart

There are two types of bar charts:

- Frequency bar chart: x-axis (categories), y-axis (frequency)
- Relative frequency bar chart: x-axis (categories), y-axis (relative frequency)

The two bar charts are identical in shape, but different in vertical scale.

Remarks:

- Sort x-axis labels in alphabetical order, or in ascending/descending order of frequency
- Gap width and bar width should be consistent
- The frequency value does not need to be shown

Pie Chart

Pie chart shows **relative frequency** of each category.

Remarks:

- Sort categories in alphabetical order, or in ascending/descending order of relative frequency
- Categories of very small relative frequency can be combined into "others"
- The relative freqeuncy value should be shown

Frequency Distribution Table

Work on numerical data.

FDT requires data to be sorted in ascending order. So categorical data CANNOT be used.

1. Find the range of the data (e.g. 822.9), and decide the number of classes (e.g. 10)
2. Divide the range by the number of classes, and choose an appropriate class interval (e.g. 82.29 -> 100)

Features of a frequency distribution table:

- **exact value of each observation is lost**
- Width of each class interval is the same

Remarks:

- All class boundaries should include the left endpoint, but exclude the right endpoint (i.e. left-closed, right-open)

Excel functions:

```
=COUNTIF(A:A, "<=" & B2) - COUNTIF(A:A, "<" & C2)
```

where `A:A` is the data range, `B2` is the lower boundary of the class, and `C2` is the upper boundary of the class.

Variations of frequency distribution table:

- **Percentage distribution table**: replace frequency with relative frequency
- **Cumulative frequency distribution table**: add a column for cumulative relative frequency

Histogram

Based on a frequency distribution table.

Remarks:

- **NO gap between bars**

Universal Remarks

Prevent misleading graphs:

- Max y-axis value should be slightly larger than the largest frequency
- Min y-axis value should be 0. A break symbol can be used
- No unnecessary adornments
- **Label the axes and give the graph a title**

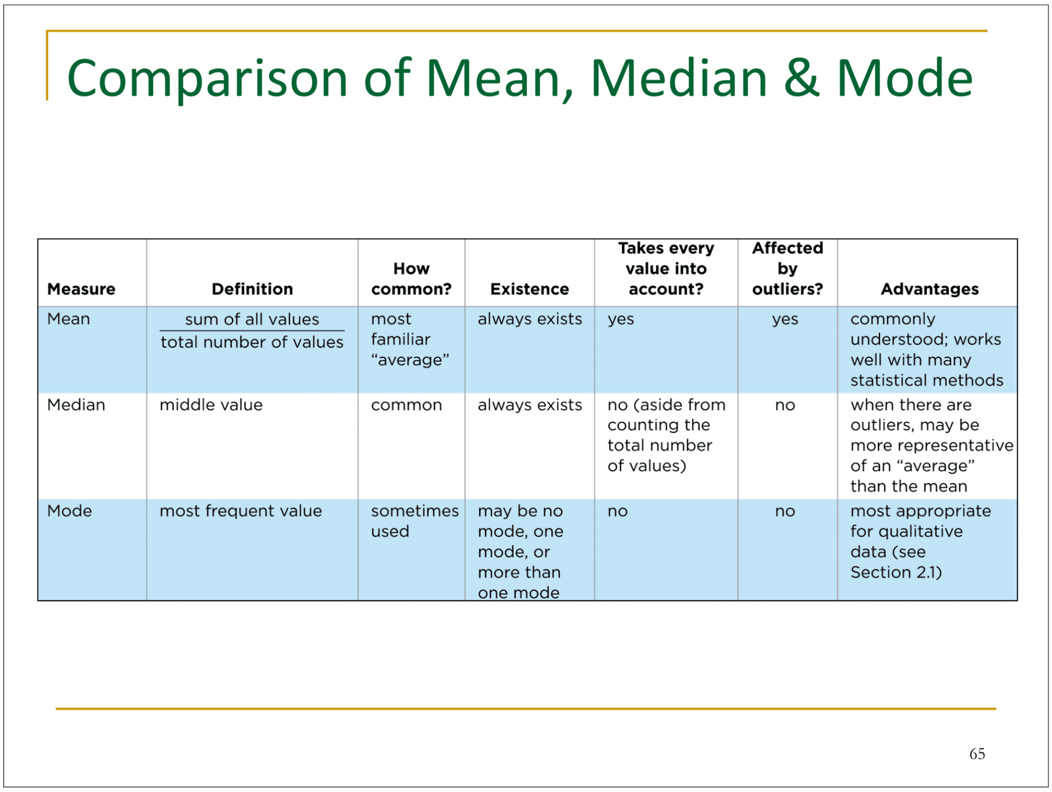
Descriptive Statistics

- **Central tendency** tells that all the values tend to cluster around a typical or central value
- **Variation** tells how spread out the data is

Measures of central tendency:

- **Mean**
  - **Sample mean**: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - **Population mean**: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
- **Median**: middle value in a ordered data set
- **Mode**: most frequent value in a data set

Comparing mean, median, and mode:



Comparison of Mean, Median & Mode

Measure	Definition	How common?	Existence	Takes every value into account?	Affected by outliers?	Advantages
Mean	$\frac{\text{sum of all values}}{\text{total number of values}}$	most familiar "average"	always exists	yes	yes	commonly understood; works well with many statistical methods
Median	middle value	common	always exists	no (aside from counting the total number of values)	no	when there are outliers, may be more representative of an "average" than the mean
Mode	most frequent value	sometimes used	may be no mode, one mode, or more than one mode	no	no	most appropriate for qualitative data (see Section 2.1)

65

Measures of variation:

- **Range**: difference between the largest and smallest values
  - $R = x_{max} - x_{min}$
- **Interquartile range (IQR)**: difference between the third (Q3) and first (Q1) quartiles
  - $IQR = Q_3 - Q_1$
- **Variance**: average of the squared differences from the mean
  - Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - Population variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- **Standard deviation**: square root of the variance

Understanding sample variance vs. population variance:

- Population variance is used when the entire population is used to calculate the variance
- Sample variance is used when a sample is used to estimate the population variance

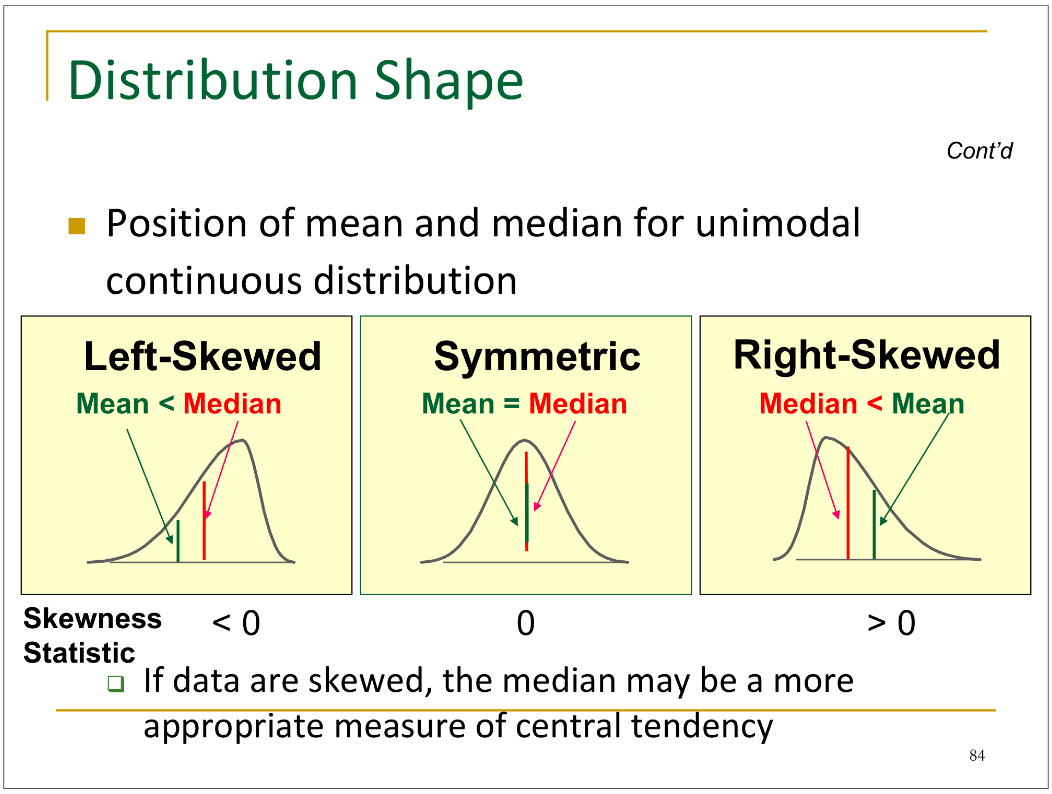
Calculating quartiles:

- Calculate position: $r_1 = \frac{1}{4}(n+1)$, $r_2 = \frac{2}{4}(n+1)$, $r_3 = \frac{3}{4}(n+1)$
- If r is an integer, $Q_r = x[r]$
- Otherwise, let $d = r - \lfloor r \rfloor$, $Q_r = x[\lfloor r \rfloor] + d(x[\lfloor r \rfloor + 1] - x[\lfloor r \rfloor])$

IQR:

- also called **midspread** because it covers the middle 50% of the data
- unaffected by outliers
- usually, values fall out of range $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ are considered outliers

Skewness: measure of the asymmetry of a distribution



Distribution Shape

Cont'd

- Position of mean and median for unimodal continuous distribution

Left-Skewed
Mean < Median

Symmetric
Mean = Median

Right-Skewed
Median < Mean

Skewness Statistic < 0 0 > 0

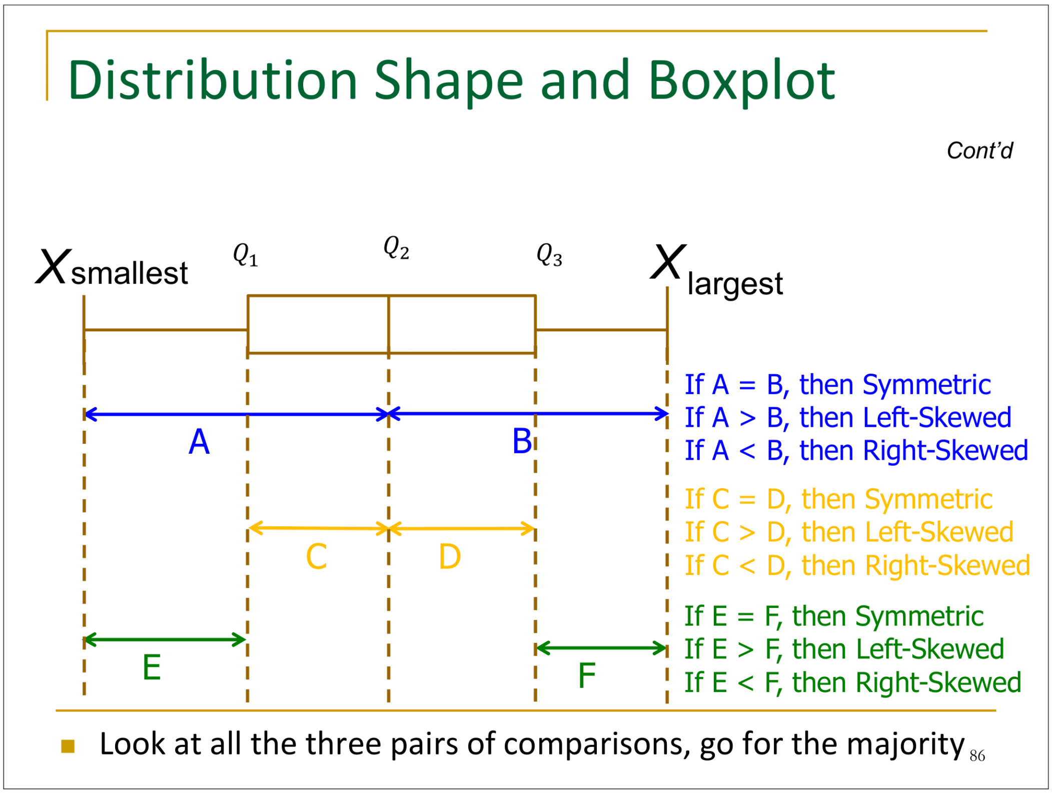
- If data are skewed, the median may be a more appropriate measure of central tendency

84

	Left-skewed	Symmetric	Right-skewed
Skewness	negative	0	positive
Mean	< median	= median	> median

Boxplot: a diagram that shows the five-number summary of a distribution

- $\min, Q_1, Q_2, Q_3, \max$
- Each range carries 25% of the data
- The length of the box is IQR
- The length of each range can be used to detect skewness



Distribution Shape and Boxplot

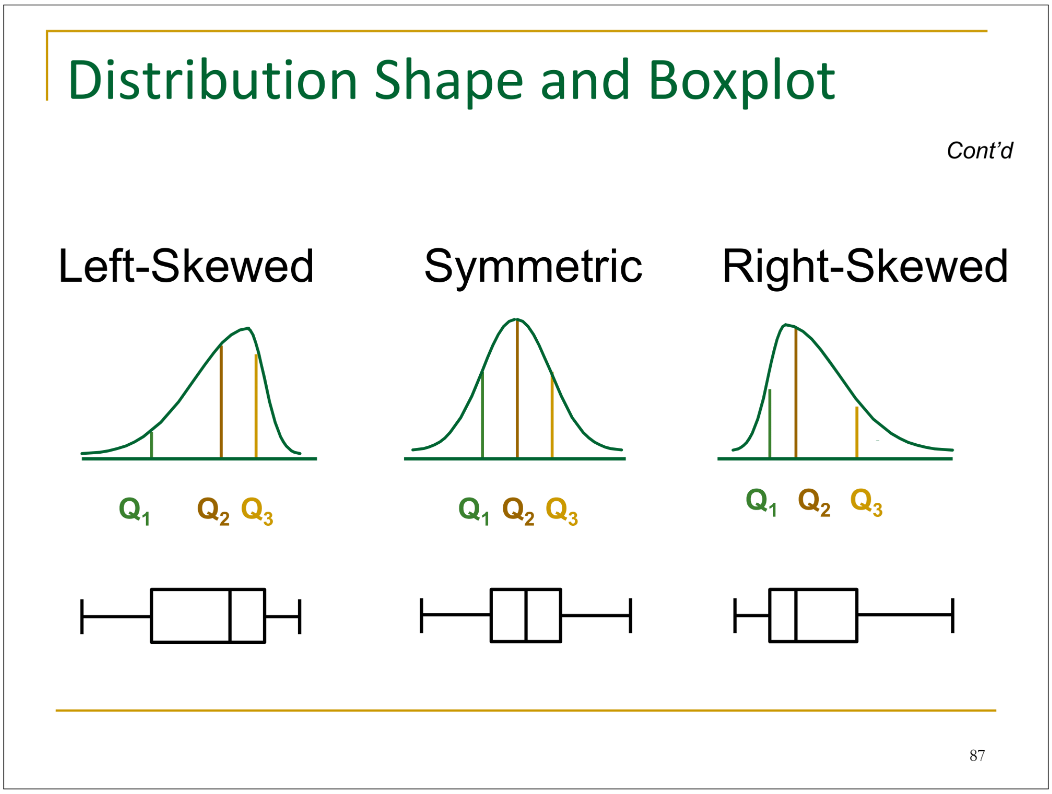
Cont'd

If A = B, then Symmetric
If A > B, then Left-Skewed
If A < B, then Right-Skewed

If C = D, then Symmetric
If C > D, then Left-Skewed
If C < D, then Right-Skewed

If E = F, then Symmetric
If E > F, then Left-Skewed
If E < F, then Right-Skewed

- Look at all the three pairs of comparisons, go for the majority



Distribution Shape and Boxplot

Cont'd

Excel functions:

```
Mean = AVERAGE()
Median = MEDIAN()
Mode = MODE() # if there are multiple modes, the smallest one is returned; if there is no mode,
#N/A is returned
Range = MAX() - MIN()
Sample Std Dev = STDEV.S()
Sample Variance = VAR.S()
Q1 = QUARTILE.EXC(data, 1)
Q3 = QUARTILE.EXC(data, 3)
```

2. Basic Probability

Terminology

- **Outcome**: a possible result of a random experiment
- **Event**: a collection of outcomes
- **Simple event**: an event that contains only one outcome, e.g. rolling a 6
- **Joint event**: an event that contains more than one outcome, e.g. rolling an even number
- **Complement** of A: $A^c = \{x \in S | x \notin A\}$
- **Sample space** S: the collection of all possible outcomes
- **Probability**:
    - numerical value representing the likelihood of an event
    - $0 \leq P(A) \leq 1$
    - **A priori** probability: probability based on prior knowledge (i.e. calculated)
    - **Empirical** probability: probability based on experimental data (i.e. observed)
    - **Subjective** probability: probability based on personal judgement
- **Impossible event**: $P(A) = 0$
- **Certain event**: $P(A) = 1$
- **Mutually exclusive events**: $A \cap B = \emptyset, P(A \cap B) = 0$
- **Collectively exhaustive events**: $A \cup B = S, P(A \cup B) = 1$

Probability Rules

- General addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - $P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive
- General multiplication rule: $P(A \cap B) = P(A) \times P(B|A)$ ($P(A) > 0$)
    - $P(A \cap B) = P(A) \times P(B)$ if A and B are independent

Conditional Probability

- $P(B|A) = \frac{P(A \cap B)}{P(A)} \Leftrightarrow P(A \cap B) = P(A) \times P(B|A)$
- If A and B are independent, $P(B|A) = P(B)$

To prove independence: prove either $P(B|A) = P(B)$ or $P(A|B) = P(A)$. Do not use the lemma in the proof.

Counting Techniques

- Permutation: ${}_nP_r = \frac{n!}{(n-r)!}$
- Combination: ${}_nC_r = \frac{n!}{r!(n-r)!}$

3. Probability Distributions

Random Variables

- **Random variable**: a variable that takes on different values as a result of random experiment
- **Discrete random variable**: a random variable that comes from counting
- **Continuous random variable**: a random variable that comes from measuring

In the context of distributions, the mean, variance, and standard deviation are **population parameters**.

Discrete Probability Distributions

Discrete probability distribution: A **mutually exclusive** and **collectively exhaustive** list of all possible values of a discrete random variable, along with their corresponding probabilities.

Properties:

- $0 \leq P(X = x_i) \leq 1$
- $\sum_{i=1}^{n} P(X = x_i) = 1$
- Expected value = mean = $\mu = E(X) = \sum_{i=1}^{n} x_i \times P(X = x_i)$
- Variance = $\sigma^2 = E(X^2) - (E(X))^2$
- Standard deviation = $\sigma = \sqrt{\sigma^2}$

Binomial Distribution

Characteristics:

- n **independent and identical** trials
- 2 mutually exclusive outcomes: success (S) and failure (F)
- Constant probability of success: $P(S) = \pi$

$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} = \frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x}$

Properties:

- Expected value = $E(X) = n \times \pi$
- Variance = $\sigma^2 = n \times \pi \times (1 - \pi)$
- Standard deviation = $\sigma = \sqrt{n \times \pi \times (1 - \pi)}$
- If $\pi = 0.5$, the distribution is symmetric
- If $\pi < 0.5$, the distribution is left-skewed
- If $\pi > 0.5$, the distribution is right-skewed
- Peak of the distribution is at $x = n \times \pi$, or the nearest integer...

Excel functions:

```
BINOM.DIST(x, n, p, cumulative)
# cumulative = TRUE for CDF = P(X <= x), FALSE for PMF = P(X = x)
```

Continuous Probability Distributions

Continuous variable: a variable that can take on **any value on a continuum**

- Probability density function (PDF): $f(x)$
- Cumulative distribution function (CDF): $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
- $P(a \leq X \leq b) = F(b) - F(a)$ = area under the $f(x)$ curve between a and b (y = relative frequency)
- $P(X = x) = 0$ for continuous random variable

Properties of PDF:

- $f(x) \geq 0$ for all x within the range
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Normal Distribution

Normal Density Function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$

Properties:

- Range: $(-\infty, +\infty)$
- Symmetric about the mean $x = \mu$
- Spread is determined by the standard deviation σ
- Mean = median = mode
- **Empirical rule**: 68-95-99.7 rule
  - 68% of the data falls within 1 standard deviation of the mean ($\mu \pm \sigma$)
  - 95% of the data falls within 2 standard deviations of the mean ($\mu \pm 2\sigma$)
  - 99.7% of the data falls within 3 standard deviations of the mean ($\mu \pm 3\sigma$)

Standard normal distribution: a normal distribution with $\mu = 0$ and $\sigma = 1$ (i.e. $Z \sim \mathcal{N}(0, 1^2)$)

Standardization: convert a normal distribution to a standard normal distribution

Translating $X \sim \mathcal{N}(\mu, \sigma^2)$ to $Z \sim \mathcal{N}(0, 1^2)$:

$$Z = \frac{X-\mu}{\sigma}$$

Excel functions:

```
NORM.DIST(x, mean, standard_dev, cumulative)
# cumulative = TRUE for CDF = P(X <= x), FALSE for PDF = f(x)
```

Reading from a normal distribution table: the table shows CDF of the standard normal distribution (i.e. $P(Z \leq z)$)

Inverse function: Given $P(X \leq x)$, find x

```
NORM.S.INV(probability) # returns z
NORM.INV(probability, mean, standard_dev) # returns x
```

4. Sampling Distributions

Sampling distribution: the distribution of a sample statistic (e.g. sample mean, sample proportion) for all possible samples of a given size from a population

i.e. Given the population, we take all possible samples of a given size, calculate their sample statistics, and plot the distribution of these sample statistics

SD of the Sample Mean $\bar{X}$

Consider a small startup with 4 staffs. Their salaries are 20, 30, 40, and 50 (unit: 1000 HKD).

The population mean is $\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{20+30+40+50}{4} = 35$

The population stddev is $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} = \sqrt{125} = 11.18$

Now, the government performs a census and randomly chooses 2 staffs to audit their salaries. This is sampling without replacement. However, for simplicity, we only consider sampling with replacement.

There are 16 possible samples:

	A	B	C	D

	A	**B**	**C**	**D**
A	20, 20	20, 30	20, 40	20, 50
B	30, 20	30, 30	30, 40	30, 50
C	40, 20	40, 30	40, 40	40, 50
D	50, 20	50, 30	50, 40	50, 50

Compute the sample mean $\bar{X}$ for each sample:

	A	**B**	**C**	**D**
A	20	25	30	35
B	25	30	35	40
C	30	35	40	45
D	35	40	45	50

List them in a frequency distribution table:

$\bar{X}$	**Frequency**	**Relative Frequency**
20	1	0.0625
25	2	0.125
30	3	0.1875
35	4	0.25
40	3	0.1875
45	2	0.125
50	1	0.0625

The population mean of the $\bar{X}$ is $\mu_{\bar{X}} = \frac{\sum_{i=1}^{N} \bar{x}_i}{N} = 35$

The population stddev of the $\bar{X}$ is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{62.5} = \sqrt{\frac{125}{2}} = 8.87$

Properties of the Sample Mean $\bar{X}$

(1) Mean of sample means

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

The formula works with or without replacement, if **the samples are unbiased** (i.e. every possible sample has the same chance of being selected)

(2) Standard deviation of sample means (**standard error of the mean**)

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

where n is the sample size.

The formula only works with replacement. (If the population is large, the formula works with or without replacement)

As n increases, $\sigma_{\bar{X}}$ decreases. If more elements are included in the sample, the sample mean becomes more stable.

Sampling from a Normal Distribution

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X}$ also follows a normal distribution:

$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

which is obtained by plug $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ into $\bar{X} \sim \mathcal{N}(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$.

Therefore, the sample mean $\bar{X}$ is normally distributed, with the same mean as the population mean, and a smaller standard deviation. Graphically, the distribution is narrower.

Example 1: It is measured that an 1-liter bottle of soft drink contains 1.025L on average, with a standard deviation of 0.02L. If 25 bottles are randomly selected, what is the probability that the sample mean is between 1.02L and 1.03L?

Solution:

Given $X \sim \mathcal{N}(1.025, 0.02^2)$ and sample size $n = 25$, we have $\bar{X} \sim \mathcal{N}(1.025, \frac{0.02^2}{25}) = \mathcal{N}(1.025, 0.004^2)$

$$P(1.02 \leq \bar{X} \leq 1.03) = P(\frac{1.02-1.025}{0.004} \leq \frac{\bar{X}-1.025}{0.004} \leq \frac{1.03-1.025}{0.004}) = P(-1.25 \leq Z \leq 1.25) = 0.89435 - 0.10565 = 0.7887$$

Therefore, it is expected that out of every 5 samples, about 4 of them will have a sample mean between 1.02L and 1.03L.

Central Limit Theorem

All non-normal distributions are naturally skewed.

When the sample size n is small, the distribution of the sample mean $\bar{X}$ is also skewed.

However, as n increases, the distribution of $\bar{X}$ becomes more and more bell-shaped, and eventually can be approximated by a normal distribution.

Central limit theorem: If the sample size n is large, the distribution of the sample mean $\bar{X}$ is approximately normally distributed, regardless of the shape of the population distribution.

For most distributions, $n \geq 30$ can be considered large.

Example 2: Weights of apples are normally distributed with a mean of 140g and a standard deviation of 20g.

(1) What is the chance that, the mean weight of a sample of 20 apples exceeds 150g?

(2) Based on (1), will the chance change if the sample size is 30?

(3) Based on (2), will the chance change if the distribution of the weights is not normal?

(4) For a sample consists of 30 apples from a non-normal distribution with the same statistics, what is the chance that the mean weight falls between 135g and 150g?

Answers:

(1) Given $X \sim \mathcal{N}(140, 20^2)$ and sample size $n = 20$, we have $\bar{X} \sim \mathcal{N}(140, \frac{20^2}{20}) = \mathcal{N}(140, \sqrt{20}^2)$

$$P(\bar{X} > 150) = 1 - P(\bar{X} \leq 150) = 1 - P(\frac{\bar{X}-140}{\sqrt{20}} \leq \frac{150-140}{\sqrt{20}}) = 1 - P(Z \leq 2.24) = 1 - 0.9875 = 0.0125$$

(2) Given $n = 30$, we have $\bar{X} \sim \mathcal{N}(140, \frac{20^2}{30}) = \mathcal{N}(140, \sqrt{\frac{40}{3}}^2)$

$$P(\bar{X} > 150) = 1 - P(\bar{X} \leq 150) = 1 - P(\frac{\bar{X}-140}{\sqrt{\frac{40}{3}}} \leq \frac{150-140}{\sqrt{\frac{40}{3}}}) = 1 - P(Z \leq 2.74) = 1 - 0.9968 = 0.0032$$

(3) The chance will not change. According to the central limit theorem, when the sample size is large, the distribution of the sample mean is approximately normally distributed, regardless of the shape of the population distribution.

(4) According to the CLT, Given $\mu = 140, \sigma = 20$, and $n = 30$, we have $\bar{X} \sim \mathcal{N}(140, \frac{20^2}{30}) = \mathcal{N}(140, \sqrt{\frac{40}{3}}^2)$

$$P(135 \leq \bar{X} \leq 150) = P(\frac{135-140}{\sqrt{\frac{40}{3}}} \leq \frac{\bar{X}-140}{\sqrt{\frac{40}{3}}} \leq \frac{150-140}{\sqrt{\frac{40}{3}}}) = P(-1.37 \leq Z \leq 2.74) = 0.9968 - 0.0853 = 0.9115$$

5. Confidence Intervals Estimation

Terminology

- **Estimation**: the process of inferring the value of a population parameter from a sample statistic
- **Point estimation**: use a (one) sample statistic to estimate a population parameter
- **Confidence interval** (CI): a range of values that is likely to contain the value of a population parameter
    - based on **one sample** from the population (point estimation)
    - **level of confidence**: the probability that the CI contains the population parameter

	Population Parameters	**Sample Statistics**
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Proportion	π	p

Confidence Interval for the Mean μ

Normal Distribution (Known σ)

Given a population that

- **is normally distributed**, AND
- **standard deviation σ is known**

If, population is not normally distributed, but $n \geq 30$, the sample mean $\bar{X}$ is approximately normally distributed, according to the central limit theorem.

Given a sample that is normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$, the sample mean $\bar{X}$ is also normally distributed $\bar{X} \sim \mathcal{N}(\mu_{\bar{X}}, \frac{\sigma}{\sqrt{n}}^2)$

Then, the $100(1-\alpha)\%$ CI for μ is:

$\bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$

where

- $Z_{\alpha/2}$ (critical value) is the z-score such that $P(z \geq Z_{\alpha/2}) = \frac{\alpha}{2}$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ is the standard error (= standard deviation of the sample mean)
- $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the sampling error (margin of error)

Colloquials:

- $P(z \leq -Z_{\alpha/2}) = P(z \geq Z_{\alpha/2}) = \frac{\alpha}{2}$
- $P(-Z_{\alpha/2} \leq z \leq Z_{\alpha/2}) = 1 - \alpha$

Understanding **level of confidence**:

If you repeat sampling many times, and construct a CI for each sample, then about $100(1-\alpha)\%$ of the CIs will contain the population mean.

Conventially speaking, we are $100(1-\alpha)\%$ confident that the population mean falls within the CI.

Example 1: It is measured that an 1-liter bottle of soft drink contains 1.025L on average, with a standard deviation of 0.02L. If 25 bottles are randomly selected, what is the 95% CI for the mean amount of soft drink in a bottle?

Solution:

Given 95% CI, we have $\alpha = 0.05$, and $Z_{\alpha/2} = Z_{0.025} = 1.96$

The 95% CI for μ is $1.025 \pm 1.96 \times \frac{0.02}{\sqrt{25}} = 1.025 \pm 0.00784 = (1.01716, 1.03284)$

Example 2: A random sample of 15 stocks traded on the Hang Seng Index showed an average 215,000 shares. The population standard deviation is 195,000 and the shares traded are very close to a normal distribution. What is the 99% CI for the mean number of shares traded?

Solution:

Given 99% CI, we have $\alpha = 0.01$, and $Z_{\alpha/2} = Z_{0.005} = 2.576$

The 99% CI for μ is $215000 \pm 2.576 \times \frac{195000}{\sqrt{15}} = 215000 \pm 129698.5 = (85301.5, 344698.5)$

Interpretation of the CI:

- If **all possible samples** of size 15 are taken, and a 99% CI is constructed for each sample, then about 99% of the CIs will cover the population mean.
- We are 99% confident that the population mean falls within the CI.
- Incorrect interpretation: There is a 99% chance that the population mean falls within the CI. (The population mean is a constant, so the outcome is either 0% or 100%. There are no probabilities involved.)

Commonly used α and corresponding z-values:

α	z_α	$z_{\alpha/2}$
0.01	2.326	2.576
0.05	1.645	1.960
0.10	1.282	1.645

Factors Affecting the Width of the CI (t-Distribution)

Observing the formula $\bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$, we can see that the width of the CI is affected by:

- Population standard deviation:
  - $\sigma \uparrow \rightarrow \frac{\sigma}{\sqrt{n}} \uparrow \rightarrow$ CI width $\uparrow$
- Sample size:
  - $n \uparrow \rightarrow \frac{\sigma}{\sqrt{n}} \downarrow \rightarrow$ CI width $\downarrow$
- Level of confidence:
  - $1 - \alpha \uparrow \rightarrow |Z_{\alpha/2}| \uparrow \rightarrow$ CI width $\uparrow$

Student's t-Distribution

$T \sim t(v)$, where $v > 0$ is the degrees of freedom.

Degree of freedom: the number of independent pieces of information used to estimate a parameter.

Degree of freedom = Number of observations - Number of intermediate parameters estimated from the observations

When calculating sample variance s^2, we have to first estimate the sample mean $\bar{x}$, which is a parameter. Therefore, the degree of freedom is $n - 1$.

Properties of the t-distribution $t(v)$:

- $\mu_T = 0$ for all $v > 1$
- Standard deviation $\sigma_T = \sqrt{\frac{v}{v-2}}$ for $v > 2$ (∞ for $1 < v \leq 2$, undefined for $v \leq 1$)

Properties of PDF $f(t)$:

- Infinite range: $(-\infty, +\infty)$
- Bell-shaped, symmetric, median = mode = 0
- As v increases, the t-distribution approaches the standard normal distribution $\mathcal{N}(0, 1)$

Reading from a t-distribution table: 2 parameters are required:

- the degrees of freedom v
- the area to the right of the critical value ($P(t \geq t_{\alpha/2}) = \frac{\alpha}{2}$)

And the value given by tuple $(v, \frac{\alpha}{2})$ is the critical value $t_{\alpha/2}$.

- Given fixed $\frac{\alpha}{2}$, as v increases, $t_{\alpha/2}$ decreases (reason: as v increases, the sample statistics are more and more reliable, the sample mean becomes more and more stable, and the CI becomes narrower)
- Given fixed v, as $\frac{\alpha}{2}$ decreases, $t_{\alpha/2}$ decreases (reason: as the level of confidence increases, the CI becomes wider)

Excel functions:

```
T.DIST(x, degrees_freedom, cumulative)
# cumulative = TRUE for CDF = P(X <= x), FALSE for PDF = f(x)

T.INV(probability, degrees_freedom)
```

`T.DIST(x, v, TRUE)` calculates $f(t)$ for given t and v.

`T.INV(probability, v)` calculates the critical value $t_{\alpha/2}$ for given v and $\frac{\alpha}{2}$.

Normal Distribution (Unknown σ)

If the population standard deviation σ is unknown, we use the sample standard deviation s to estimate σ.

Recall:

- Population variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

The transformed variable $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ follows a t-distribution with $n - 1$ degrees of freedom.

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

Then, the $100(1-\alpha)\%$ CI for μ is:

$$\bar{X} \pm t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}}$$

where

- $t_{\alpha/2,n-1}$ is the critical value
- $s_{\bar{X}} = \frac{s}{\sqrt{n}}$ is the standard error
- $E = t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$ is the sampling error

Example 3: The monthly salary of brokers is found to be normally distributed. A random sample of 25 brokers has a mean monthly salary of 80K, with a standard deviation of 16K. What is the 95% CI for the mean monthly salary of all brokers?

Read the question carefully: the mean and standard deviation are sample statistics, not population parameters. Population parameters are always explicitly given (e.g. the standard deviation of the population is known to be 16K). Otherwise, it can be assumed that all the given values are sample statistics.

Solution:

Given 95% CI, we have $\alpha = 0.05$, and $t_{\alpha/2,n-1} = t_{0.025,24} = 2.0639$

The 95% CI for μ is $80000 \pm 2.0639 \times \frac{16000}{\sqrt{25}} = 80000 \pm 6604.48 = (73395.52, 86604.48)$

Example 4: A random sample of 200 customers showed that the average amount spent at a pet supply store is 213.4, with a standard deviation of 92.2. What is the 99% CI for the mean amount spent at the pet supply store?

Solution:

Given 99% CI, we have $\alpha = 0.01$, and $t_{\alpha/2,n-1} = t_{0.005,199} = 2.8387$

The 99% CI for μ is $213.4 \pm 2.8387 \times \frac{92.2}{\sqrt{200}} = 213.4 \pm 18.5 = (194.9, 231.9)$

Factors Affecting the Width of the CI

Observing the formula $\bar{X} \pm t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}}$, we can see that the width of the CI is affected by:

- Sample standard deviation:
  - $s \uparrow \rightarrow \frac{s}{\sqrt{n}} \uparrow \rightarrow$ CI width $\uparrow$
- Sample size:
  - $n \uparrow \rightarrow \frac{s}{\sqrt{n}} \downarrow \rightarrow$ CI width $\downarrow$
  - $n \uparrow \rightarrow t_{\alpha/2,n-1} \downarrow \rightarrow$ CI width $\downarrow$
- Level of confidence:
  - $1 - \alpha \uparrow \rightarrow |t_{\alpha/2,n-1}| \uparrow \rightarrow$ CI width $\uparrow$

Sample Size Determination

Since sampling takes time and effort, it is important to balance the sample size and the accuracy of the estimation.

Observing the formula $\bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$, we can see that the width of the CI is affected by the sample size n. While the mean μ is unknown, the sampling error $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ can be controlled by adjusting the sample size n.

If σ is known, and the sampling error cannot exceed a certain value E, then the sample size n can be determined by:

$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2$

Example 5: Based on 679 randomly selected graduates, the average salary is 20,501 with a standard deviation of 3,875.

(1) Calculate the 95% CI for the mean salary of all graduates.

(2) If the sampling error cannot exceed 500, what is the minimum sample size required?

Solution:

(1) is a case of normal distribution with unknown σ.

Given 95% CI, we have $\alpha = 0.05$, and $t_{\alpha/2,n-1} = t_{0.025,678} = 1.9635$

The 95% CI for μ is $20501 \pm 1.9635 \times \frac{3875}{\sqrt{679}} = 20501 \pm 291.99 = (20209.01, 20792.99)$

(2) Given $E = 500, \alpha = 0.05, \sigma = 3875$, we have $n = \left(\frac{1.9635 \times 3875}{500}\right)^2 = 231.6$

So a sample size of 232 is required. (In this question, σ is unknown, so we assume $\sigma = s$)

However, this requires knowledge of the population standard deviation σ. If previous data is available, then σ can be estimated from the sample standard deviation s. If not, then σ can be estimated by range/4.

Example 6: The government wants to estimate the average number of hours per week that university students spend on part-time jobs. The range of the number of hours is 0 to 40. What is the minimum sample size required to estimate the average number of hours per week with a 95% CI, if the sampling error cannot exceed 2 hours?

Solution:

Estimate σ by range/4: $\sigma = \frac{40-0}{4} = 10$

Given $E = 2, \alpha = 0.05, \sigma = 10$, we have $n = \left(\frac{1.96 \times 10}{2}\right)^2 = 96.04$

So a sample size of 97 is required.

6. Hypothesis Testing

Step 1: Define the Hypotheses

- **Hypothesis testing**: a statistical method that uses sample data to evaluate a hypothesis about a population parameter
- **Null hypothesis** (H_0): a statement about the value of a population parameter that is assumed to be true
    - Always contains a population parameter (e.g. μ) but not a sample statistic
    - Always contains an equality condition ($=, \leq, \geq$)
    - Always **assumed to be true** at start
    - The final decision would be either **reject** or **not to reject** (but NOT accept)
- **Alternative hypothesis** (H_1 or H_a): a statement that contradicts the null hypothesis
    - The **opposite** of the null hypothesis
    - Always contains a population parameter and an inequality condition
    - **Mutually exclusive and collectively exhaustive** with the null hypothesis
- Types of hypothesis testing
    - **Two-tailed test**: $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
    - **Lower-tailed test**: $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$
    - **Upper-tailed test**: $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$

Example: Coka-Cola claims that the average amount of soft drink in a 1L bottle is 1.025L. The government wants to test this claim.

$H_0 : \mu \geq 1.025$ vs $H_1 : \mu < 1.025$

Step 2: Collect the Data and Identify the Rejection Region

Given a population $P(N^?, \mu^?, \sigma^?)$ (population parameters are fixed but unknown), and a sample $S(n, \bar{x}, s)$ (sample statistics are observed), we want to test the null hypothesis $H_0 : \mu = \mu_0$.

Rejection region: the range of values of the sample statistic that leads to the rejection of the null hypothesis

- i.e. if the sample statistic falls within the rejection region, the null hypothesis is rejected
- **size** or **level of significance** (α): the probability of rejecting the null hypothesis when it is true
    - the false-positive rate (reporting rejection but actually not)
    - typically 0.01, 0.05, or 0.10
    - provides the **critical value** of the hypothesis test

Different types of hypothesis testing have different rejection:

- If X is known to be normally distributed, OR the sample size $n \geq 30$
    - If population stddev σ is known, use the standard normal distribution, i.e. Z-test
    - If population stddev σ is unknown, use the t-distribution, i.e. t-test
- If X is not normally distributed, and the sample size $n < 30$, assume the population is normally distributed and use the t-distribution

The testing types are similar to the confidence interval types.

Step 3: Compute the Test Statistic

Denote test statistic u as Z or t based on the type of hypothesis testing. Recall:

- Z test statistic:
    - Given that population is normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$, and population stddev σ is known
    - Sample mean is also normally distributed $\bar{X} \sim \mathcal{N}(\mu, (\frac{\sigma}{\sqrt{n}})^2)$
    - $Z = \frac{\bar{X} - \mu(\bar{X})}{\sigma(\bar{X})} = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$
    - where σ^2 is the population variance $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- t test statistic:
    - Given that population is normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$, but population stddev σ is unknown
    - Sample mean is also normally distributed $\bar{X} \sim \mathcal{N}(\mu, (\frac{s}{\sqrt{n}})^2)$

    - $t = \frac{\bar{X}-\mu(\bar{X})}{s(\bar{X})} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
    - where s^2 is the sample variance $s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$

Step 4: Make Statistical Decision

There are 2 approaches:

(1) **Critical value approach**:

Denote critical value $c_{\alpha/2}$ as $Z_{\alpha/2}$ or $t_{\alpha/2,n-1}$ based on the type of hypothesis testing. Recall:

- $Z_{\alpha/2}$: $P(Z \geq Z_{\alpha/2}) = \frac{\alpha}{2}$
- $t_{\alpha/2,n-1}$: $P(T \geq t_{\alpha/2,n-1}) = \frac{\alpha}{2}$ under $T \sim t(n-1)$

For two-tailed test:

- If H_0 is true, we are $100(1-\alpha)\%$ confident that the sample mean falls within the CI
- i.e. $P(-c_{\alpha/2} \leq Z \leq c_{\alpha/2}) = 1-\alpha$
- If $Z \leq -c_{\alpha/2}$ or $Z \geq c_{\alpha/2}$, reject H_0

For lower-tailed test: (lower tail is the rejection region)

- $P(Z \leq -c_\alpha) = \alpha$
- If $Z \leq -c_\alpha$, reject H_0

For upper-tailed test: (upper tail is the rejection region)

- $P(Z \geq c_\alpha) = \alpha$
- If $Z \geq c_\alpha$, reject H_0

(2) **P-value approach**:

Convert the test statistic to a p-value, which is the probability of observing a sample statistic as extreme as the one computed from the sample, assuming that the null hypothesis is true. (i.e. p = the probability that H_0 is true, given the sample data. A more extreme sample statistic leads to a smaller p-value.)

Then, compare the p-value with the level of significance α: if $p \leq \alpha$, reject H_0.

- For two-tailed test: If $P(Z \leq -|u|) + P(Z \geq |u|) \leq \alpha$, reject H_0
- For lower-tailed test: If $P(Z \leq u) \leq \alpha$, reject H_0
- For upper-tailed test: If $P(Z \geq u) \leq \alpha$, reject H_0

Types of Errors

Decision \ Truth	H_0 **true**	H_0 **false**
Do not reject H_0	Level of Confidence $(1-\alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Power of the test $(1-\beta)$

- Type I error: reject a true H_0
    - $\alpha = P\left(\text{reject } H_0 | H_0 \text{ true}\right)$
    - α is the level of significance (the smaller, the better)
    - $(1-\alpha)$ is the level of confidence
- Type II error: fail to reject a false H_0
    - $\beta = P\left(\text{do not reject } H_0 | H_0 \text{ false}\right)$
    - $(1-\beta)$ is the power of the test
- α is specified before the test (e.g. 0.01, 0.05, 0.10)
- β depends on the true value of the population parameter which is unknown
- Ways to reduce the probability of Type II error:
    - Increase α (preferred if the cost of Type II error exceeds the cost of Type I error)
    - Increase the sample size n (preferred if the resources permit)

Example

(1) A random sample of $n = 25$ boxes of cereals gave a mean $\bar{X} = 364.5$ g.

The company has specified the population distribution is normal and the standard deviation $\sigma = 15$ g.

Test at $\alpha = 0.05$ level of significance whether the average weight is close to $\mu_0 = 368$ g.

Solution:

$H_0 : \mu = 368$ vs $H_1 : \mu \neq 368$

Since $X \sim \mathcal{N}(\mu, \sigma^2)$, the sample mean $\bar{X}$ is also normally distributed $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) = \mathcal{N}(\mu, \frac{15^2}{25}) = \mathcal{N}(\mu, 3^2)$

(a) Critical value approach:

Since $\alpha = 0.05$ and the test is double-tailed, we have $Z_{\alpha/2} = Z_{0.025} = 1.96$. H_0 will be rejected if $Z \leq -1.96$ or $Z \geq 1.96$.

$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{364.5-368}{15/\sqrt{25}} = -1.167$

Since Z is not in the rejection region, we do not reject H_0.

(b) P-value approach:

Given that $Z = -1.167$, the double-tailed p-value $P(Z \leq -1.167) + P(Z \geq 1.167) = 0.2432 > \alpha = 0.05$. Therefore, we do not reject H_0.

(2) Customers are concerned about the amount of cereal being less than 368g. Test at $\alpha = 0.05$ level of significance whether the average weight is less than $\mu_0 = 368$ g.

$H_0 : \mu \geq 368$ vs $H_1 : \mu < 368$

(a) Critical value approach:

Since $\alpha = 0.05$ and the test is lower-tailed, we have $Z_\alpha = Z_{0.05} = 1.645$. H_0 will be rejected if $Z \leq -1.645$.

$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{364.5-368}{15/\sqrt{25}} = -1.167$

Since Z is not in the rejection region, we do not reject H_0.

(b) P-value approach:

Given that $Z = -1.167$, the lower-tailed p-value $P(Z \leq -1.167) = 0.1216 > \alpha = 0.05$. Therefore, we do not reject H_0.

(3) Each soft drink bottle should contain $\mu = 1$ L of drink as specified on the label. A random sample of $n = 40$ bottles showed a mean of $\bar{X} = 1.03$ L and $\mu = 0.08$ L. The company suspects that the machine is not working properly given the large deviation. Check at $\alpha = 0.05$ level of significance.

$H_0 : \mu = 1$ vs $H_1 : \mu \neq 1$

Since $X \sim \mathcal{N}(\mu, \sigma^2)$, the sample mean $\bar{X}$ is also normally distributed $\bar{X} \sim \mathcal{N}(\mu, \frac{s^2}{n}) = \mathcal{N}(\mu, \frac{0.08^2}{40}) = \mathcal{N}(\mu, 1.6 \times 10^{-4})$

(a) Critical value approach:

Since $\alpha = 0.05$ and the test is double-tailed, we have $t_{\alpha/2,n-1} = t_{0.025,39} = 2.0227$. H_0 will be rejected if $t \leq -2.0227$ or $t \geq 2.0227$.

$t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{1.03-1}{0.08/\sqrt{40}} = 2.3717$

Since t is in the rejection region, we reject H_0.

(b) P-value approach:

Since $t_{0.025,39} = 2.0227, t_{0.01,39} = 2.4258, t_{0.025,39} < 2.3717 < t_{0.01,39}$, it can be inferred that $0.01 < P(T \geq 2.3717) < 0.025$.

$P(T \leq -2.3717) + P(T \geq 2.3717) < \alpha = 0.05$. Therefore, we reject H_0.

It is not recommended to use the p-value approach for the t-test, as the p-value is not directly available from the t-table.

(4) Check at $\alpha = 0.05$ level of significance whether the average amount is more than $\mu_0 = 1$ L.

Be careful: the question asks for $\mu > 1$, but the null hypothesis **must contain an equality condition**.

The correct null hypothesis is $H_0 : \mu \leq 1$ vs $H_1 : \mu > 1$

Critical value approach:

Since $\alpha = 0.05$ and the test is upper-tailed, we have $t_{\alpha,n-1} = t_{0.05,39} = 1.6859$. H_0 will be rejected if $t \geq 1.6859$.

$t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{1.03-1}{0.08/\sqrt{40}} = 2.3717$

Since t is in the rejection region, we reject H_0.

Therefore, we do not reject that the average amount is more than 1L.

(5) The freeze point of natural milk is distributed with a mean of $\mu = -0.545^\circ$ C. $n = 14$ randomly selected bottles of milk shows a mean $\bar{X} = -0.550^\circ$ C and a standard deviation $s = 0.016^\circ$ C. Test at $\alpha = 0.05$ level of significance whether the milk contains excess water.

Step 1: **The sample is from an unknown distribution, and the sample size $n < 30$. Assume the population is normally distributed,** according to CLT, the sampling distribution of the sample mean is also normally distributed.

Step 2: Hypothesis: $H_0 : \mu = -0.545$ vs $H_1 : \mu \neq -0.545$

Step 3: Given $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\bar{X} \sim \mathcal{N}(\mu, \frac{s^2}{n}) = \mathcal{N}(\mu, \frac{0.016^2}{14}) = \mathcal{N}(\mu, (0.004276)^2)$ and t-distribution is used.

Step 4: Critical value approach

Given $\alpha = 0.05$ and the test is double-tailed, we have $t_{\alpha/2,n-1} = t_{0.025,13} = 2.1604$. H_0 will be rejected if $t \leq -2.1604$ or $t \geq 2.1604$.

$t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{-0.550-(-0.545)}{0.016/\sqrt{14}} = -1.1693$

Since t is not in the rejection region, we do not reject H_0.

Therefore, at $\alpha = 0.05$ level of significance, we do not have enough evidence to conclude that the milk contains excess water.

7. CI Estimation and Hypothesis Testing for Proportions

Proportion

The observation is a **categorial variable** which only take **two values** (e.g. yes/no, success/failure).

Given n observations, and Y observations are of one category, the **sample proportion** is $p = \frac{Y}{n}$.

Population proportion π is equal to the probability of success in a single observation.

$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$

Sampling Distribution of Sample Proportion

Consider $N = 4$ ballots received at a polling station, where 3 votes are for candidate A. The sample proportion is $pi = \frac{3}{4} = 0.75$.

A random sample of size $n = 2$ is taken with replacement.

Denote Y as the number of votes for candidate A in the sample. Therefore Y obeys a binomial distribution $Y \sim B(2, 0.75)$.

- $\mu = np = 2 \times 0.75 = 1.5$
- $\sigma = \sqrt{np(1-p)} = \sqrt{2 \times 0.75 \times 0.25} = 0.6124$

Note μ and σ are population parameters.

Now we want to study the sample proportion $p = \frac{Y}{n}$.

Consider all 16 possible samples of size 2:

p	M1	M2	M3	F
M1	1	1	1	0.5
M2	1	1	1	0.5
M3	1	1	1	0.5
F	0.5	0.5	0.5	0

The frequency distribution of p is:

p	0	0.5	1
$P(p)$	1/16	6/16	9/16

- $\mu_p = \frac{1}{16} \times 0 + \frac{6}{16} \times 0.5 + \frac{9}{16} \times 1 = 0.75$
- $\sigma_p = \sqrt{\frac{1}{16} \times (0 - 0.75)^2 + \frac{6}{16} \times (0.5 - 0.75)^2 + \frac{9}{16} \times (1 - 0.75)^2} = 0.3062$

Properties of Sampling Distribution of Sample Proportion

(1) Mean of the sample proportion $\mu_p = \pi$

(2) Standard deviation of the sample proportion $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

(3) According to CLT, we approximate the sampling distribution by a normal distribution if: $n \geq 30$ and $np \geq 5$ and $n(1-p) \geq 5$

The approximated distribution is $p \sim \mathcal{N}(\pi, \frac{\pi(1-\pi)}{n})$

Z-score is $Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$

Example:

Suppose 40\% of the depositors at a bank have multiple accounts. A random sample of 200 depositors is taken. What is the probability that the sample proportion of depositors with multiple accounts is less than 0.3?

Solution:

Given $\pi = 0.4, n = 200, p = 0.3$

- Step 0: Let p be the sample proportion of depositors with multiple accounts.
- Step 1: Since $n \geq 30$, $np = 80 \geq 5$ and $n(1-p) = 120 \geq 5$, p is normally distributed according to CLT.
- Step 2: $Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.3-0.4}{\sqrt{\frac{0.4\times0.6}{200}}} = -2.8868$
- Step 3: $P(p < 0.3) = P(Z < -2.8868) = 0.0020$

Estimation of Population Proportion

If population proportion π is unknown, the standard deviation σ_p of the sample proportion can be estimated by the sample deviation s_p.

$\sigma_p = s_p = \sqrt{\frac{p(1-p)}{n}}$

If π is unknown, use $np \geq 5$ and $n(1-p) \geq 5$ to verify the CLT condition.

Confidence Interval for Population Proportion

Given a sample proportion p, the $100(1-\alpha)\%$ CI for the population proportion π is:

$$p \pm Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$$

where

- $Z_{\alpha/2}$ is the critical value
- $\sqrt{\frac{p(1-p)}{n}}$ is the standard error σ_p
- $Z_{\alpha/2} \times \sigma_p$ is the sampling error E

If $p - Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}} < 0$, set the lower bound to 0.

If $p + Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}} > 1$, set the upper bound to 1.

Example 1: Among a random sample of 200 depositors, 95 have multiple accounts. Set up a 95% CI for the population proportion of depositors with multiple accounts.

Solution:

Given $95/200 = 0.475, n = 200, \alpha = 0.05$

- Step 0: Let p be the sample proportion of depositors with multiple accounts.
- Step 1: Since population proportion π is unknown, π is estimated by $p = 0.475$. Since $n \geq 30$, $np = 95 \geq 5$ and $n(1-p) = 105 \geq 5$, p is normally distributed according to CLT.
- Step 2: $Z_{\alpha/2} = Z_{0.025} = 1.96$
- Step 3: Calculate the CI:
  $p \pm Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}} = 0.475 \pm 1.96 \times \sqrt{\frac{0.475 \times 0.525}{200}} = 0.475 \pm 0.0692 = (0.4058, 0.5442)$

Factors Affecting the Width of the CI

Observing the formula $p \pm Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$, we can see that the width of the CI is affected by:

- Level of confidence
  - $(1-\alpha) \uparrow \rightarrow |Z_{\alpha/2}| \uparrow \rightarrow$ CI width $\uparrow$
- Sample size
  - $n \uparrow \rightarrow \frac{p(1-p)}{n} \downarrow \rightarrow$ CI width $\downarrow$
- Sample proportion
  - $p \in (0, 0.5) \uparrow \rightarrow \sqrt{\frac{p(1-p)}{n}} \uparrow \rightarrow$ CI width $\uparrow$
  - $p \in (0.5, 1) \uparrow \rightarrow \sqrt{\frac{p(1-p)}{n}} \downarrow \rightarrow$ CI width $\downarrow$

Sample Size Determination

If the sampling error $E = Z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$ should not exceed a certain value E_0, then the sample size n can be determined by:

$$E_0 = Z_{\alpha/2} \times \sqrt{\frac{\pi(1-\pi)}{n}} \Rightarrow n = \frac{Z^2 \times \pi(1-\pi)}{E_0^2}$$

> Note this formula is given in population proportion π.

Example 2: Per each 10,000 transactions, 22 are suspected of fraud. What is the minimum sample size required to estimate the population proportion of fraud transactions with a 99% CI, if the sampling error cannot exceed 0.001?

Solution:

Given $p = \frac{22}{10000} = 0.0022, \alpha = 0.01, E_0 = 0.001$

- Step 1: Despite n is unknown, we can first assume $n \geq 30$ and $np \geq 5$ and $n(1-p) \geq 5$, such that p is normally distributed according to CLT.

- Step 2: $Z_{\alpha/2} = Z_{0.005} = 2.576$
- Step 3: $n = \frac{Z^2 \times \pi(1-\pi)}{E_0^2} = \frac{2.576^2 \times 0.0022 \times 0.9978}{0.001^2} = 14566.59$

So a sample size of 14567 is required.

Hypothesis Testing for Population Proportion

Z-distribution is always used for hypothesis testing of population proportion.

If the population proportion π is unknown, use the sample proportion p to estimate π.

Example 3: The bank has a business objective of serving 80\% of the customers within 5 minutes. A random sample of 45 customers is taken, and 39 customers are served within 5 minutes. Test the claim at $\alpha = 0.05$ level of significance.

Solution:

Given $n = 45, p = \frac{39}{45} = 0.8667, \pi = 0.8, \alpha = 0.05$

- Step 1: Since $n \geq 30$, $n\pi = 36 \geq 5$ and $n(1-\pi) = 9 \geq 5$, p is normally distributed according to CLT.
- Step 2: Hypothesis: $H_0 : \pi = 0.8$ vs $H_1 : \pi \neq 0.8$
- Step 3: $Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.8667-0.8}{\sqrt{\frac{0.8\times0.2}{45}}} = 1.1180$
- Step 4:

(a) Critical value approach:

The double-tailed critical value $Z_{\alpha/2} = Z_{0.025} = 1.96$. H_0 will be rejected if $Z \leq -1.96$ or $Z \geq 1.96$. Since Z is not in the rejection region, we do not reject H_0.

(b) P-value approach:

The double-tailed p-value $P(Z \leq -1.1180) + P(Z \geq 1.1180) = 0.2636 > \alpha = 0.05$. Therefore, we do not reject H_0.

8. Linear Regression

Covariance

Population covariance: $\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)}{N}$

Sample covariance: $s_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$

Dividing the (x, y) scatter plot into 4 quadrants:

- Quadrant I: $x > \bar{x}, y > \bar{y}$
- Quadrant II: $x < \bar{x}, y > \bar{y}$
- Quadrant III: $x < \bar{x}, y < \bar{y}$
- Quadrant IV: $x > \bar{x}, y < \bar{y}$

With positive linear association, the majority of the points are in quadrants I and III. With negative linear association, the majority of the points are in quadrants II and IV.

A non-linear association cannot be detected by covariance.

Note the unit of covariance is the product of the units of X and Y. Therefore, you cannot compare the covariance of different pairs of variables.

Coefficient of Correlation

Coefficient of correlation measures the strength and direction of the linear relationship between two variables. It is **unitless**.

Population coefficient of correlation: $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$

Sample coefficient of correlation: $r_{XY} = \frac{s_{XY}}{s_X s_Y}$

Properties:

- The sign of ρ_{XY} (r_{XY}) is the same as that of σ_{XY} (s_{XY}).
- $-1 \leq \rho_{XY}, r_{XY} \leq 1$

Correlation does not imply causation. (Casuation: changes in one variable cause changes in another variable.)

Linear Regression

Simple linear regression is a statistical method that models the relationship between two variables by fitting a linear equation to the observed data, and then makes **predictions** based on that line.

Denote the **dependent variable** as Y and the **independent variable** as X.

Assumption of linearity: $E(Y|X=x) = \beta_0 + \beta_1 x$

Discrepancy: $\epsilon = Y - (\beta_0 + \beta_1 X)$

The **population regression line** is defined as $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- $\beta_0 + \beta_1 X_i$ is the **expected value** of Y_i given X_i
- ϵ_i is the **error term** (residual) that represents the discrepancy between the observed value and the expected value

Least Squares Estimation

Residual $e_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i)$

Sum of squared errors (SSE): $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

To minimize the SSE, we differentiate the SSE with respect to β_0 and β_1 and set them to 0.

The result is (b_0, b_1 are called least squares estimators):

$b_0 = \bar{Y} - b_1 \bar{X}$

$b_1 = \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Also $b_1 = r_{XY} \frac{s_Y}{s_X}$

Therefore b_1 will have the same sign as r_{XY} (and thus s_{XY}).

Interpreting the Estimated Coefficients:

If the independent variable goes out of the sample range, the prediction may not be accurate. (The association may become approximately linear outside the sample range.)

Total sum of squares (SST): $SST = \sum_{i=1}^{n} (Y_i - \hat{Y}_i^*)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

Coefficient of determination (R-squared): $R^2 = 1 - \frac{SSE}{SST} \in [0, 1]$

In simple linear regression, $R^2 = r_{XY}^2$

Inference in Slope

Assume that residuals ϵ_i are normally distributed with mean 0 and variance σ^2.

Therefore the dependent variables Y_i are also normally distributed with mean $\beta_0 + \beta_1 X_i$ and variance σ^2.

Therefore b_1 is normally distributed:

- $E(b_1) = \beta_1$
- $\sigma_{b_1}^2 = \frac{\sigma^2}{\sigma_X^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- $\sigma_{b_1}^2$ can be estimated by $s_{b_1}^2 = \frac{s_e^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{SSE/(n-2)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/(n-2)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Mean square error (MSE): $s_e^2 = \frac{SSE}{n-2}$

Understanding $n-2$ degrees of freedom: 2 parameters, β_0 and β_1, are estimated from the sample.

Confidence interval: The $100(1-\alpha)\%$ CI for β_1 is: $b_1 \pm t_{\alpha/2, n-2} \times s_{b_1}$

t-statistic for CI: $t = \frac{b_1 - \beta_1}{s_{b_1}} \sim t(n-2)$

Hyphothesis testing: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

t-statistic for hypothesis testing: $t = \frac{b_1}{s_{b_1}}$

Example:

Given a sample of $n = 10$ observations.

X	621.0	359.7	530.0	492.1	70.5	567.0	125.5	50.6	353.3	263.6
Y	299.6	207.7	325.0	336.3	48.6	400.3	91.2	39.1	268.6	214.3

Observations:

- $\bar{X} = 343.33, \bar{Y} = 223.07$
- $\sum(X_i - \bar{X})^2 = 398,408.321$
- $\sum(Y_i - \bar{Y})^2 = 144,538.641$
- $s_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = 227,844.609$

Least squares estimation: $b_1 = \frac{s_{XY}}{\sum(X_i - \bar{X})^2} = \frac{227,844.609}{398,408.321} = 0.571887$

$b_0 = \bar{Y} - b_1\bar{X} = 223.07 - 0.571887 \times 343.33 = 26.723976$

Therefore $\hat{Y} = 26.723976 + 0.571887X$

Slope inference:

From calculator we know $r_{XY} = 0.949473$

$R^2 = r_{XY}^2 = 0.901499$

$\because R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{\sum(Y_i - \bar{Y})^2}$

$\therefore SSE = (1 - R^2) \times \sum(Y_i - \bar{Y})^2 = 14,237.20068$

$m_e^2 = \frac{SSE}{n-2} = \frac{14,237.20068}{8} = 1,779.650085$

$s_{b_1}^2 = \frac{m_e^2}{\sum(X_i - \bar{X})^2} = \frac{1,779.650085}{398,408.321} = 4.46900 \times 10^{-3}$

$s_{b_1} = 0.066835$

Application of Regression

- Time series analysis: $\hat{Y}_t = \beta_0 + \beta_1 t$
- Centa-City Index (CCI) and property price
  - CCI indicates the **current movement of property prices**
  - Multiple linear regression model is used to predict property prices