Hypothesis Testing

Hypothesis testing :-

Statistical technique to test some hypothesis about the parent population from which the sample is actually drawn.

population:

A complete collection of all elements to be studied.

Sample :

A sub collection of elements drawn from a population.

Estimation :-

It is to use the statistics obtained from the sample as estimate of the unknown parameters of the population from which the sample is drawn.

" A hypothesis is statistics is simply a quantitative statement about population"

procedure for Hypothesis Testing:-

① **Set up the Hypothesis :**

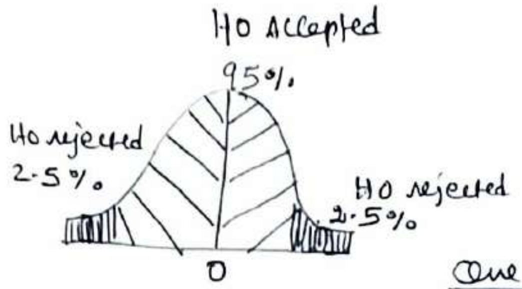
Ho (Null Hypothesis)

Ha / H1 (Alternative Hypothesis).

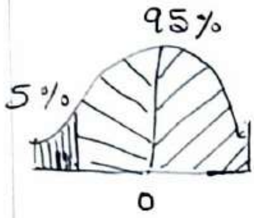
② Setup a Suitable Significant level.

Test of validity of Ho against Ha at certain level of Significant i.e 5%, 1% etc.

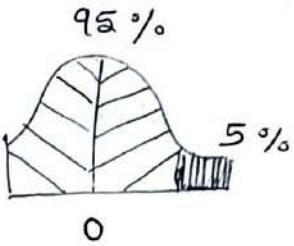
Two tailed



One tailed



One tailed



Level of Significance	10%	5%	1%	0.1%
One tailed	1.65%	1.96	2.58	3.29
Two tailed	1.28	1.64	2.33	3.10

Critical Z value (Tabulated)

Z value :-

It is a standardized score that describes how many S.D (σ) an element is from the mean.

③ Setting a test Criterion:

Selection of appropriate probability distribution for the test. i.e., t. test, χ^2 test, F test etc.

④ Doing Computation:

All calculation part is carried out.

⑤ Making Decision:

we have to draw Statistical conclusions and take decisions.

Decision	population	Condition
	Ho is true	Ho is faile
Accept Ho/ Reject Ha	Correct decision	Type II Error
Reject Ho / Accept Ha	Type I Error	Correct decision

Hypothesis testing

Introduction to test of Significance:

மிகைத்தன்மை சோதனைகள் பற்றிய முன்னுரை:

* உண்மை செயல்களில் கணக்கெடுப்புகளின் பண்புகளை ஆராய்ந்து முக்கிய முடிவெடுத்தல் இதற்கான செயல்களுக்கு எண்கள் மற்றும் மாதிரியெடுப்பு கொள்கைகள் தேவைப்படுகின்றன.

* மாதிரியெடுப்பின் பொதுப்படையான கொள்கைகள் மற்றும் அவற்றின் செயல் முறைகள். கணக்கெடுப்பிலிருந்து பொருத்தமான தன்மைகளுடன் ஒத்திருக்கின்றனவா என்று ஆராய வேண்டியுள்ளது.

* கிடைக்கும் பெரிய முடிவுகளை பல்வேறு கோட்பாடுகளை -யும் ஒப்பீடு செய்தல் மற்றும் உய்த்துணர்தல் இதற்கான தேவைகளாக வேண்டியதும் அவசியமாகிறது.

* எனவே கண்டுபிடிக்கப் பட்ட புள்ளியியல் அளவைகள் பயனுடையதாக கிடைக்கின்றன. நம்பகமான நிலையில் முடிவை ஏற்றுக் கொள்ளக் கூடிய நிலையில் உள்ளனவா என்று ஆராய வேண்டிய கட்டாயம் இருக்கிறது.

* இதற்காக பயன்படுத்தப்பட்ட மிகைத்தன்மை சோதனைகள் தேவைப்படுகின்றன.

* புள்ளியியல் ரீதியாக உய்த்துணர்தல் செய்வதற்கு கோட்பாட்டு சோதனைகள் மிகவும் அவசியம்.

* மாதிரிக்கூறின் அடிப்படையில் ஒரு முறை ஆராய்ந்து உய்த்துணர்வு கொள்ளும் சில முறை கோள்கள் (or) எடுகோள்கள் (hypotheses) தேவை கொள்கிறோம்.

Hypothesis testing:

* Most of the biological and medical investigations involve experimental studies (or) field studies with an aim to compare

  (i) a sample statistic with the population parameter (or)

  ii) statistics of samples from 2 or more populations with specific and distinct characteristics.

* Comparison of statistics from more than two popn. is also common. In all the comparisons the aim is to assess whether there is significant differences (i) between a sample statistic and the corresponding population parameter ($\bar{x}$ and μ)

  (ii) or between two parameters (μ_1 & μ_2)

  iii) or between more than two parameters ($\mu_1, \mu_2, \mu_3 \cdots$)

  using the statistics of samples ($\bar{x}_1, \bar{x}_2, \bar{x}_3 \cdots$) from the corresponding populations.

* When we talk about the "difference" between two sample statistics, and therefore, between the two popn. parameters, we mean a statistically significant difference.

* Differences may be due to 'error' it may be occur due to faulty instrument used to measure a variable.

* No two random samples from a population will be 'identical.' Some differences are bound to occur.

* Generally, an investigator has a hypothesis. The hypothesis may be that Ho(i) Sample mean is lesser than the population mean (or) $H: \bar{X} < M_1$
ii) The mean of 'treated' group is greater than the mean of the 'Control' group, (or) → $H: M_1 > M_2$
iii) The means of more than two groups are not the same. $H: M_1 \neq M_2 \neq M_3 \neq M_4$

* Whenever be the hypothesis an investigator puts forward, its Statistical Significance is obtained by subjecting the null form of the hypothesis to an appropriate test of Significance.

Null- Hypothesis: — (Ho).

* The Null form of the hypothesis (H) is called the Null-hypothesis (Ho), which means that there is no Significant difference between the mean of the two groups (or) mean of the popn. It is represented as follows.

$$H_0 : \bar{X} - M = 0$$ in other words

$$H_0 : \bar{X} = M \quad (H_0 : M_1 - M_2 = 0) \text{ (or)}$$

$$H_0 : M_1 = M_2 \quad (H_0 : M_1 = M_2 = M_3 = M_4)$$

* verbally the Ho States that there is no Significant difference between (i) Sample mean and popn. mean. (ii) or between mean of two popn. (iii) or between means of more than two populations.

Alternative Hypothesis:

* It is a Statement about the popn. parameter (or) parameters, which gives an alternative to the null hypothesis, within the range of pertinent value of the parameter i.e. if Ho is accepted, what hypothesis is to be rejected and vice versa.

* An alternative hypothesis is denoted by HA. The idea of this was originated by Neyman. for instance,

if $H_0 : \mu = 0$, the alternative are,

$H_A : \mu \neq 0$,

$H_A \mu > 0$

$H_A \mu < 0$

if $H_0 : \mu_1 = \mu_2$, the alternative are,

$H_A : \mu_1 \neq \mu_2$,

$H_A : \mu_1 > \mu_2$

$H_A \mu_1 < \mu_2$

Level of Significance:

* In testing the Null- hypothesis, we said the probability of occurance of the observed difference between Sample mean (or) more is calculated.

* This calculated probability is the probability of the occurrence of Ho, as well. If the probability for the occurrence of Ho is very low, then we reject it. Otherwise, we fail to reject the Ho.

* Level of Significance is an arbitrarily selected point in the probability scale, below which the probability is considered low, and equal to or above which the probability is considered high.

* Though any point in the probability scale can be selected as the L S, conventionally 0.05 (5%) or 0.01 (1%) is used in biostatistics.

1) எடுகோள் சோதனை (Hypothesis testing)

எந்த கணக்கெடுப்பிலிருந்து மாதிரியெடுக்கின்றோமோ அந்த கணக்கெடுப்பின் பண்புகளை சார்ந்து எடுகோள் அமைக்கப்படுகிறது.

எடுகோளின் இரண்டு வகைகள்

i) இன்மை எடுகோள்

ii) மாற்று எடுகோள்

i) இன்மை எடுகோள் (Null hypothesis)

ஆய்வுக்கு எடுத்து கொண்ட தொகுப்புகளின் புள்ளியியல் அளவு அல்லது (or) உயிர்த்தொகுதியாகும் வேறு கணக்கெடுப்பிற்கும் மாதிரியாகும் ஒரே மாதிரி உள்ளது எனவும். புள்ளியியல் வேறுபாடுகள் தற்செயலாகவோ மாதிரியெடுப்பு பிழைகளினாலோ ஏற்பட்டுள்ளன என்று எண்ணுவதே இன்மை எடுகோளாகும்

உதாரணம் இத்திறம் : $H_0 = \mu = \mu_0$ ஆகும்.

$H_0 = \mu = \bar{x}$

2) மாற்று எடுகோள் (Alternative hypothesis)

இன்மை எடுகோளுக்கு எதிர்மறையான எடுகோளாகும் மாற்று எடுகோள் (H_1) எடுத்துக்காட்டாக கல்லூரி மாணவர்களின் சராசரி உயரம் 1.55 மீ என்ற இன்மை எடுகோளை நாம் சோதனைக்குட்படுத்த வேண்டுமானால்

$$H_0 = \mu = 1.55 \text{ மீ}$$

இதற்கான மாற்று எடுகோள் பின்வருமாறு இருக்கும்

i) $H_1 \quad \mu \neq 1.55$ மீ.

ii) $H_1 \quad \mu > 1.55$ மீ

iii) $H_1 \quad \mu < 1.55$ மீ

சோதனையின் முடிவாக இன்மை எடுகோள் நிராகரிக்கப்படும் சந்தர்ப்பத்தில் மாற்று எடுகோள் ஏற்றுக் கொள்ளப்படுகின்றது

பொருண்மைத்தன்மை மட்டம் (Significant level)

* உயிரியலில் நடத்தப்படும் புள்ளியியல் அடிப்படையிலான எடுகோள்களை உய்த்தறிதலை 100 சதவீதம் ஏற்றுக் கொள்ள இயலாது. மாதிரியெடுப்பு மற்றும் மதிப்பீடுகளில் ஏற்படக்கூடிய பிழைகளே இதற்கு காரணம். எனவே 5% (அல்லது) 1% பொருண்மைத்தன்மை மட்டத்தில் ஏற்படுத்தப்பட்ட உய்த்தறிதல் செய்யப்படுகிறது.

* 5 %. முக்கியத்துவ நிலையில் இருந்து பூஜ்ய கருதுகோளை (Ho) நாம் ஏற்கும் வாய்ப்புகள் என்பது 100க்கு 5 ஆகும். அதனால் நாம் பூஜ்ய கருதுகோளை (Ho) நிராகரிக்க வாய்ப்புண்டு என்று கூறுவர். அதாவது 95% உறுதியின் நம்பிக்கை என்று முடிவினை (உறுதியாக) ஏற்றுக் கொள்ளப்படுகிறது என்று கூறலாம்.

* உண்மையில் பயன்படுத்தப்பட்ட அளவுகளுக்கு 95%க்கு குறைவாக நம்பகத்தன்மை எல்லை (confidence limit) அல்லது 5% முக்கியத்துவம் மட்டும் இல்லை வேண்டும் என்பது ஆகும்.

* உரிய சோதனை முறையும் தேர்ந்தெடுத்தல் அமையும் கருதுகோளை சோதனைக்குட்படுத்தும் போது

* இச்சோதனைகள் நிகழ்தகவு பரவல் (probability distribution) அடிப்படையில் அமைந்திருக்கின்றன. எ.கா 't' சோதனை, கைவர்க்க சோதனை, Z சோதனை ; F-சோதனை

* பொருத்தமான் பெறப்பட்ட மாதிரியின் விவரங்களை சில அடிப்படை கணக்கீடுகளுக்கு பயன்படுத்த வேண்டும். சராசரி, மாறுபாடு, திட்டவிலக்கம் போன்ற புள்ளியியல் அளவைகள், மற்றும் அவற்றின் திட்டப்பிழை அடிப்படை கணக்கிடு செய்தல் மூலமே, கருதுகோள் சோதனைக்கு தேவையான புள்ளிவிவரத்தை அவை உகந்த நிலையில் கிடைக்கும்.

* தீர்மானம் செய்தல் : உரிய சோதனை முறைகள் பயன்படுத்தி கிடைக்கும் பெறப்பட்ட கணக்கீட்டு மதிப்புகள் பூஜ்ய கருதுகோளை ஏற்றுக்கொள்ள வழிவகுக்கின்றனவா அல்லது நிராகரிக்க செய்கிறதா என்பதனை தீர்மானம் செய்வதாகும்.

* ஒவ்வொரு சோதனை வகையிலும் உரிய தீர்வுகட்டமைப்பு மதிப்புகள் (critical values) பல்வேறு வரையறுப்பு பாகைகளுக்கும் (degree of freedom) முக்கியத்துவ நிலைகளுக்கும் தனியாக புள்ளியியல் அட்டவணைகளில் (Fisher and yate's tables) பார்க்க வேண்டும்

* நாம் மேற்கொண்ட சோதனையின் முடிவு அளவைகள் உரிய d.f.-ல் 95% நம்பகத்தை எல்லையால் பட்டியலிலுள்ள மதிப்புகளை விட குறைவாக இருந்தால், சூனிய எடுகோளை ஏற்றுக் கொள்ளப்படுகிறது மாறாக அளவைகள் பட்டியலிலுள்ள மதிப்புகளை விட அதிகமாக இருக்கின் நிலையால் சூனிய எடுகோள் நிராகரிக்கப்படுகிறது.

Chi - Square Test

കൈവർഗ്ഗ ശോധന

Definition:

"Chi-square test is the test of significance of overall deviation square in the observed and Expected frequencies devided by Expected frequencies."

Characteristics of χ^2 test:

- The test is based on events or frequencies and not based on mean or S.D. etc;
- The test can be used between the entire set of observed and expected frequencies.
- To draw inferences, this test is applied, especially testing the hypothesis.
- It is a general test and is highly useful in research

Application of chi-square test:

- It is used to test the goodness of fit.
- The test enables to find out whether the difference between the expected and observed value is significant or not.
- If the difference is little then the fit is good, otherwise the fit is poor

Formula :

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

where : O = observed frequencies
E = Expected frequencies
Σ = Sum oft

Steps :

1. A hypothesis is established i.e. Null hypothesis.
2. Calculate the difference between observed value and expected value (O-E).
3. Square the deviations calculated $(O-E)^2$
4. Divide the $(O-E)^2$ by its expected frequency $\frac{(O-E)^2}{E}$.
5. Add all the values obtained in Step 4

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E}\right]$$

6. Find the Chi-Square from χ^2 table at certain level of significance, usually 5% or 1% level.

Inference :

- If the calculated value of χ^2 is greater than the table value of χ^2 at certain degree of level of significance, we reject the hypothesis.
- If the calculated value of χ^2 is zero, the observed values and expected values completely coincide.

- If the calculated value of χ^2 is less than table value at certain degree of level of significance, it is said to be non-significant.

- This implies that the difference between the observed and expected frequencies may be due to fluctuations in sampling.

<u>கைவர்க்க சோதனை:</u>

புள்ளியியல்முறையில், ஒரு மாதிரியானது, அதன் அடிப்படையிலான மக்கள்தொகையை சரியாக பிரதிபலிக்கிறதா என்று கண்டறியும் சோதனை ஆகும்.

- χ^2 சோதனை மருத்துவ ஆய்வுகளிலும், மாறிகளுக்-கான மதிப்பீட்டு சோதனை செய்வதற்கும் பெரிதும் பயன்படுகிறது.

- χ^2 ன் மதிப்பு 'O' வாக இருந்தால் கண்டறிந்த அலைவெண்களுக்கும் (O f) எதிர்பார்க்கப்பட்ட அலைவெண்களுக்கும் (E f) வேறுபாடு இருக்கவில்லை என்பதும். இரண்டும் சரியாக ஒத்துப்போகின்றது என்பதும் எடுத்துக் கொள்ளலாம்.

- χ^2 - ன் மதிப்பு அதிகமானால் (O f) கண்டறிந்த அலைவெண் மதிப்புக்கும் அடிப்படையில் மதிப்புகளுக்கிடையே பெரும் மதிப்புடைய வேறுபாடும் இருக்கிறது எனக் கொள்ளலாம்.

சூத்திரம் : $\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$

O - கண்டறிந்த அலைவெண்

E - எதிர்பார்க்கப்படும் அலைவெண்.

- கணக்கிடப்பட்ட x^2 அளவு, 5% மிகை மட்டத்தில் அட்டவணையின் மதிப்பை விட அதிகமாக இருந்தால் அது மிகைத்தன்மையுடையது எனவும் இரண்டு மாதிரிகளுக்கிடையேயான வேறுபாடுகள் இருக்கின்றன அல்லது பொருந்தவில்லை எனவும் எண்ணலாம்.

- கணக்கிடப்பட்ட x^2 மதிப்பு அட்டவணை மதிப்பை விட குறைவாக இருந்தால் கணிக்கப்பட்ட முறைகளுக்கும் எதிர்பார்க்கப்பட்ட முறைகளுக்குமிடையே வேறுபாடுகள் இல்லை எனக் கூறலாம். அதாவது இரு பரவல்களும் ஒன்றோடு ஒன்று ஒத்துப்போகிறது என்ற நிலை ஏற்படுகிறது.

- பெரும் பரவல்களை இயல்நிலை சோதனைக்குட்படுத்துவது அனைத்து கணிப்புகளும் அதிகமாகவும், ஒன்றோடு ஒன்று தொடர்புடையதாகவும் இருக்க வேண்டும் - மாதிரிகளை மாற்றும்போது அவற்றின் இயல்நிலையில் பயன்படுத்தப்பட வேண்டும்.

x^2 பயன்கள்:

(i) மாதிரிகளின் சார்பற்ற தன்மையை அறியும் சோதனையில் பயன்படுகிறது. இரண்டு பண்புகளுக்கிடையே தொடர்புள்ளதா இல்லையா என்பதை இச்சோதனை செய்யலாம்.

(ii) இயல்நிலைப்பரவல் (Normal distribution), ஈருறுப்பு பரவல் (Binomial distribution) மற்றும் போய்சான் பரவல் போன்ற அறியப்பட்ட பரவல்களின் சோதனைகளைக் கொண்ட பரவல்கள் மற்றும் அறுவடைகள் ஒத்துப்போகின்றனவா என்பதை அறிய x^2 பயன்படுகிறது. இதனை பொருத்தச் செம்மை சோதனை (Goodness of fit) என்றும் கூறலாம்.

(iii) ஒரு பண்பின் ஒருமைத்தன்மையை சோதனையிடும் (test for homogenity) χ^2 சோதனை பயன்படுத்தப்படுகிறது.

<u>கணக்கிடும் முறை:</u>

1. இன்மை எடுகோளை அமைக்கவும்.
2. எதிர்பார்க்கப்பட்ட மதிப்புகளைக் கண்டுபிடிக்கவும்.
3. கண்டுபிடித்த மதிப்புகளிலிருந்து எதிர்பார்க்கப்பட்ட மதிப்புகளை கழிக்கவும் $(O-E)$
4. மேற்கண்ட வேறுபாடு மதிப்புகளின் வர்க்கம் காணவும் $(O-E)^2$
5. ஒவ்வொரு $(O-E)^2$ மதிப்பையும், அதன் எதிர்பார்க்கப்பட்ட மதிப்பால் வகுக்கவும் $\frac{(O-E)^2}{E}$
6. $\frac{(O-E)^2}{E}$ யின் கூட்டுத்தொகையை காணவும் $\sum\frac{(O-E)^2}{E}$
7. χ^2 அட்டவணை, உரிய df-ல் 5% அல்லது 1% நிலை அட்டவணையில் χ^2 மதிப்பை காண். அதனை கணக்கிடப்பட்ட $\sum\frac{(O-E)^2}{E}$ மதிப்புடன் ஒப்பிடவும்.

<u>எ.கா.</u> ஒரு மகப்பேறு மருத்துவமனையில் பிறந்த 200 குழந்தைகளில் 120 ஆண் குழந்தைகளாகவும் 80 பெண் குழந்தைகளாகவும் இருந்தன. இவ்விவரங்களைப் பயன்படுத்தி ஆண் / பெண் பிறப்பு பாகுபாடு இல்லை என்று நிரூபிக்கவும்.

இன்மை எடுகோள் ஆண் : பெண் = 1 : 1

மொத்த குழந்தைகள் = 200

எதிர்பார்க்கப்பட்ட ஆண் குழந்தை = $\frac{200 \times 1}{2} = 100$

பாலினம்	கண்டறிந்த அலைவெண் (O)	எதிர்பார்க்கப்பட்ட அலைவெண் (E)
ஆண்	120	100
பெண்	80	100

பாலினம்	கண்டறிந்த அலைவெண் (O)	எதிர்பார்க்கப்பட்ட அலைவெண் (E)	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
ஆண்	120	100	120-100 = 20	400	= 4
பெண்	80	100	80-100 = -20	400	= 4
					8

கணக்கிட்ட x^2 மதிப்பு = 800

$\left.\begin{array}{l}5\%\ \text{மட்டத்தில்} \\ \text{அட்டவணை } x^2\end{array}\right\} = 3.84$

1% ,, ,, = 6.635

df = நிகழ்வுகளின் வகைகளின் எண்ணிக்கை = 2-1 = 1

உய்த்துணர்வு:

1% மட்டத்திலேயே கண்டறிந்த x^2 மதிப்பு அதிகமாக உள்ளது (6.64 < 8) எனவே எதிர்வினை நிராகரிக்கப்படுகிறது. அதாவது ஆண் பெண் பாலின விகிதம் 1:1 என்ற கொள்கையின் நிலைமையானது, ஆண்பெருக்கம் மாற்றங்களுக்கு அதிக காரணப்படுகிறது.

Illustration 1 : A coin is tossed 100 times of which head comes 60 times and tail 40 times would you accept the hypothesis that the coin is normal having no bias for either head or tail.

Solution:

Step : 1 : Null hypothesis - i.e. the coin is normal having no bias for either head or tail.

2 : level of significance 5 %.

3. Determining expected frequencies (E)

possibilities	Observed (O) frequencies	Expected frequencies (E)
Head	60	50
Tail	40	50

4. Fixing the degree of freedom $df = n-1$

n = number of events or possibilities

i.e. head and tail $n = 2-1 = 1$

5. calculation: $x^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$

possibi-lities	Observed frequency (O)	Expected frequency (E)	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Head	60	50	$60-50=10$	$=100$	$\frac{100}{50}=2.0$
Tail	40	50	$40-50=10$	$=100$	$\frac{100}{50}=2.0$

Calculated x^2 value $= 4.00$ $\underline{=4}$

Table value at 5 %. level for one degree of freedom is $\boxed{3.84}$

Inference : The calculated x^2 value is greater than the table value. Therefore hypothesis is rejected. In other words the coin is defective with bias for head.

Illustration 2: A cross involving different genes gave rise to F2 generation of tall and dwarf in the ratio of 110:90. Test by means of chi-square whether this value is deviated from the mendel's monohybrid ratio 3:1

Solution

Steps: 1: Null hypothesis: There is no difference between 110:90 and Mendel's monohybrid ratio 3:1

2. Level of significance 5%.

3. Determining expected frequencies (E)

Mendel's Monohybrid ratio Tall : Dwarf = 3:1

Observed total number = 110+90 = 200

Expected = Tall and dwarf 3:1

= 150 : 50 = 200

4. Fixing the degrees of freedom

df = n−1 , 2−1 = 1

Calculation: $x^2 = \Sigma \left[\frac{(O-E)^2}{E}\right]$

Variables	O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Tall	110	150	-40	1600	10.6
Dwarf	90	50	40	1600	32.0

Calculated x^2 value = $\boxed{42.6}$ $\Sigma = 42.6$

For 1 df at 5% level of significance the table value = $\boxed{3.84}$

Inference:

The calculated x^2 value 42.6 is greater than the table value 3.84. Therefore the hypothesis is rejected. In otherwards the value 110:90 is derived from Mendel's monohybrid ratio 3:1

Illustration 3: when two heterozygous pea plants are crossed, 1600 plants are produced in the F2 generation. out of which 940 are yellow round. 260 are yellow wrinkled, 340 are green round and 60 are green wrinkled. By means of x^2 test wheather these values are devided from mendel's dihybrid ratio 9:3:3:1.

Solution:

Steps 1: Null hypothesis: There is no difference between Observed values and Mendel's dihybrid ratio 9:3:3:1.

2: level of Significance 5%.

3: Determining expected frequencies (E):
Mendel's dihybrid ratio 9:3:3:1

Yellow Round = 9 Total 1600 $\therefore E = \frac{9}{16} \times 1600 = 900$

yellow wrinkled = 3 " $\therefore E = \frac{3}{16} \times 1600 = 300$

Green Round = 3 " $\therefore E = \frac{3}{16} \times 1600 = 300$

Green wrinkled = $\frac{1}{16}$ Total 1600 $\therefore E = \frac{1}{16} \times 1600 = 100$

$\overline{1600}$

4: Fixing the df = n-1 = 4-1 = 3

Calculation: $x^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$

variables	O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
yellow Round	940	900	40	1600	1.77
yellow wrinkled	260	300	-40	1600	5.33
Green Round	340	300	40	1600	5.33
Green wrinkled	60	100	-40	1600	16.00
					$\Sigma = 28.43$

calculated x^2 value = $\boxed{28.43}$

For 3 df, at 5% level of significance, Table x^2 value $\boxed{7.81}$

Inference: The calculated x^2 value is greater than the table x^2 value. Therefore the hypothesis is rejected. In otherwords There is no real independent assortment or the observed value are deviated from Mendel's dihybrid ratio 9:3:3:1.

Illustration 4: A certain drug was administered to 500 people out of a total of 800 included in the sample to test its efficacy against typhoid. The result are given below: Find out the effectiveness of drug against the disease.

	Typhoid	No Typhoid	Total
Administering the drug	200	300	500
without administering the drug	280	20	300
Total	480	320	800

Solution:

Steps: 1: Null hypothesis i.e the drug is not effective in preventing typhoid

2: Level of significance 5 %

3: preparing 2x2 Contingency table (O)

	Typhoid	No typhoid	Total
Drug	200	300	500
No Drug	280	20	300
Total	480	320	800

4: preparing table for expected frequencies (E)

	Typhoid	No Typhoid	Total
Drug	$\frac{480 \times 500}{800} = 300$	$\frac{320 \times 500}{800} = 200$	500
No Drug	$\frac{480 \times 300}{800} = 180$	$\frac{320 \times 300}{800} = 120$	300
	480	320	800

5. Fixing the degrees of freedom $df = (r-1)(c-1)$

where r = row, c = column $= (2-1)(2-1) = 1$

6. Calculation : $\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
200	300	−100	10000	33.33
280	180	100	10 000	55.55
300	200	100	10 000	50.00
20	120	−100	10 000	83.33
				$\Sigma = 222.21$

Calculated χ^2 value : 222.21

For 1 d.f, at 5 % level of Significance
the table χ^2 value = 3.84

Inference :

The calculated χ^2 value (222.21) is greater than the table χ^2 value (3.84). Therefore the null hypothesis is rejected. In other words the drug is effective in preventing typhoid.

ஸ்டூடன்டின் t- சோதனை (Student t-test)

* 1905-ல் வில்லியம் கோஸட் என்பவர் சிறிய உருவளவுக்கான சோதனை முறையை கண்டுபிடித்தார்.
* ஸ்டூடன்ட் என்ற புனைப்பெயரில் இது வெளியிட்டதால் இதற்கு ஸ்டூடன்டின் t- சோ
* அவர்கள் சிறியவையாகவும், இயல்பாகவே பரவியிருக்கும் இனக்கூட்டத்தின் (Normally distributed population) திட்டவிலக்கம் தெரியாதபோதும் இங்கு பயன்படும் t- சோதனை மேற்கொள்ளப்படுகிறது.

t- சோதனையின் நோக்கங்கள்:

1. ஒரு சீரற்ற (Random) மாதிரியின் சராசரியின் முக்கியத்துவத்தை சோதனை செய்வதற்கு
2. இரண்டு மாதிரிகளின் சராசரிகளுக்கிடையேயான வேறுபாட்டின் முக்கியத்துவத்தை சோதனை செய்ய.

t- க்கான சூத்திரங்கள்:

1. இணைக்கப்பட்ட ஒரு மாதிரியும் (paired t-test)

$$t = \frac{\bar{x} - \mu}{S} \times \sqrt{n}$$

$\bar{x} - \mu$ = வேறுபாடு ∴ மதிப்பு இல்லை

S = S.D மாதிரியின் திட்டவிலக்கம்

μ = இனக்கூட்டத்தின் சராசரி (நிஜம்/ தெரிந்தோ)

$\bar{x}$ = மாதிரியின் சராசரி

n = மாதிரியின் அளவு (எண்ணிக்கை)

$$S = \sqrt{\frac{\Sigma x - \bar{x}^2}{n-1}}$$

2) இரண்டு மாதிரிகளின் சோதனைகள் (unpaired t-test) (Indepent Samples)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}}$$

$$S = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

3) தொடர்புக் கெழுவுக்கான t-சோதனை (t-test for Correlation coefficient)

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

4) இரண்டு மாதிரிகளின் சோதனை (unpaired t-test) (Dependent Samples)

$$t = \frac{\bar{d}\sqrt{n}}{S}$$

$$S = \sqrt{\frac{\Sigma (d - \bar{d})^2}{n-1}}$$

$\bar{d}$ = Mean of the differences

S = S.D of the differences

n = Size of the sample.

Student 't' test

(Basic and fundamental properties)

1. Theoretical work on t- distribution was done by W.S. Gosset; he has published his findings under the pen name "Student". That's why it is called Student's t - test.

2. Student's t - test is used when sample size is 30 or less (≤ 30) and the population Standard deviation (σ) is unknown.

3. t - distribution has been derived mathematically under the assumption of a normally distributed population.

$$f(t) = C\left(1 + \frac{t^2}{\nu \text{ (or) } df}\right)^{-\frac{\nu+1}{2}}$$

where, C = Constant required to make the area under the curve equal to unity.

ν = Degree of freedom.

Properties of t- distribution:

1. The variable t- distribution ranges from $-\infty$ to $+\infty$ ($-\infty < t < +\infty$).

2. t - distribution will be symmetrical like normal distribution, if power of t is even in probability density function (pdf).

(3) for large values of ν (i.e. increased sample size n). the t-distribution tends to a Standard normal distribution. This implies that for differ ν value and the shape of t-distribution differs.

Sample size (n)	df (ν)	t value ($t_{0.025}$)
5	4	2.776
10	9	2.262
30	29	2.045
∞	–	1.96

Z value <u>1.96</u>

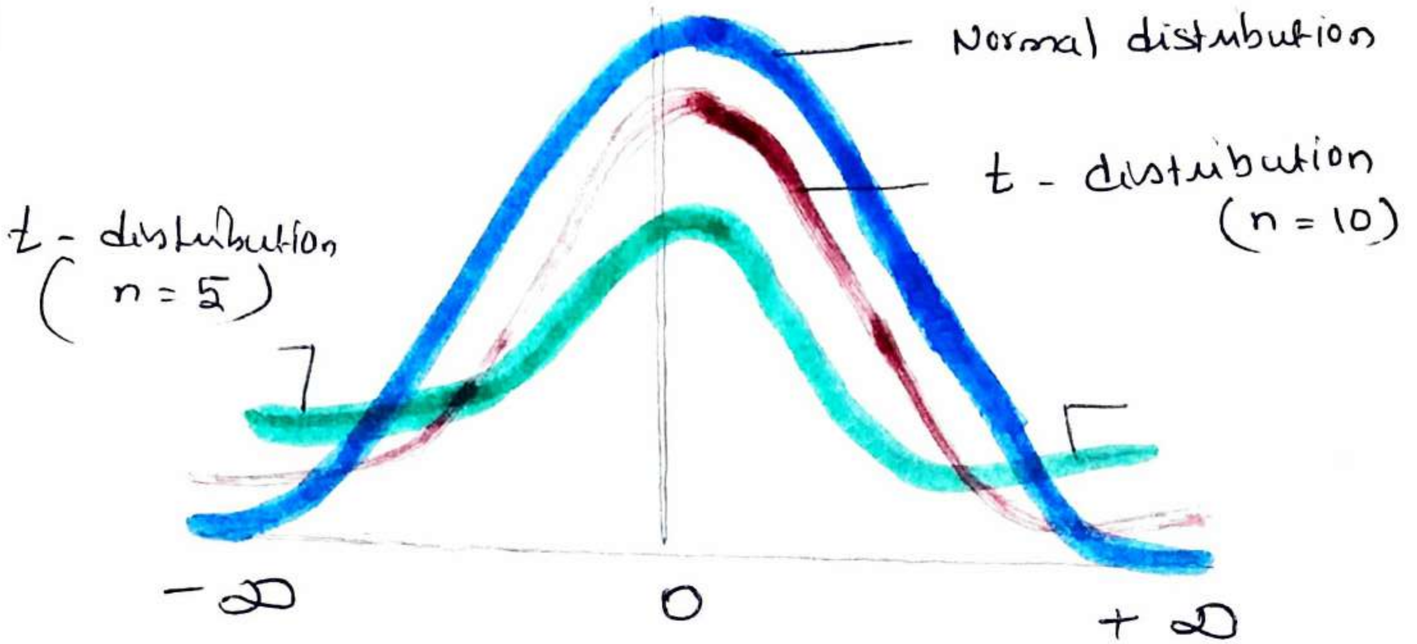
(4) The 't' distribution is less peaked than normal distribution at the centre and higher peaked in the tails.

(5) The value of y (peak height) attains highest at $t = 0$

t. Distribution table:

– Gives t value for different level of Significance and different degree of freedom.

– Calculated t-value will be compared with ~~t~~abulated t-value.


Normal distubution
t - distubution
(n = 10)
t - distubution
(n = 5)
- ∞
0
+ ∞

Student t - test

(1) To test the Significance of the mean of a random sample (Sample Size is 30 (or) less than 30)

$$t = \frac{(\bar{x}-\mu)}{S} \times \sqrt{n}, \quad S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1 \ (df)}}$$

$\bar{x}$ = Mean of the Sample
μ = Mean of the population
n = Sample Size
S = Standard deviation of the Sample

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{n}}$$

→ for population

Confidence interval estimate (for α level of Significance)

One tailed test $\bar{x} \pm t\alpha \times \frac{S}{\sqrt{n}}$ $\quad \alpha = 0.05$

Two tailed test $\bar{x} \pm t\alpha/_2 \times \frac{S}{\sqrt{n}}$ $\quad \alpha = \frac{0.05}{2} = 0.025$

(2) To test the difference between Means of the Two Samples (Indepent Samples) (unpaired test)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \times \sqrt{\frac{n_1 n_2}{n_1+n_2}},$$

$$S = \sqrt{\frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-x_2)^2}{n_1+n_2-2 \ (df)}}$$

x_1 = Mean of the Sample (1)
x_2 = Mean of Sample (2)

n_1 = Sample Size of Sample (1)
n_2 = Sample Size of Sample (2)

S = Standard deviation (Combined)

(3) To test the difference between Means of Two Samples (Depened Samples or Matched pair observations).

$$t = \frac{\bar{d}\sqrt{n}}{S}$$

$$S = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1 \ (df)}}$$

$\bar{d}$ = Mean of the differences
S = SD of the difference
n = Size of the Sample.

(4) Testing the Significance of an observed Correlation Coefficient.

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2} \ (df)$$

r = Correlation Coefficient

n = Size of the Sample.

Testing the Significance of the Mean
(For Random Sample)

Q ① The manufacturer of a certain make of LED bulb claims that his bulbs have a mean life of 20 months.
A random sample of 7 such bulb gave the following values. Life of bulbs in months:

19, 21, 25, 16, 17, 14, 21, can you regard the producer's claim to be valid at 1 %. level of significance?

Solution:

Given data, popn. mean (μ) = 20 months
Life of Bulbs (in months) : 19, 21, 25, 16, 17, 14, 21

level of significance: 1 %.

$H_0 : \mu = \bar{x}$
$H_a : \mu \neq \bar{x}$

$$t = \frac{\bar{x} - \mu}{S} \times \sqrt{n} \; ; \quad S = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

Calculation of $\bar{x}$ and S

x	$(x-\bar{x})$	$(x-\bar{x})^2$
19	0	0
21	+2	4
25	+6	36
16	−3	9
17	−2	4
14	−5	25
21	−2	4
133		82

$\bar{x} = \frac{\Sigma x}{n} = \frac{133}{7} = 19$

$\Sigma (x-\bar{x})^2 = 82$

$S = \sqrt{\frac{82}{7-1}} = \sqrt{\frac{82}{6}}$

$S = \sqrt{13.67} = 3.7$

$t = \frac{19-20}{3.7} \times \sqrt{7}$

$$t = \frac{1 \times 2.65}{3.7}$$

$t = \boxed{0.716}$ calculated 't' value.

$df(\nu) = n-1 = 7-1 = 6$ $t_{0.001}$ (tabulated value)

$t_{0.001} = \boxed{3.707}$

Accept the Hypothesis i.e.

→ Ho is passed and accepted
→ There is no difference between the sample mean and population mean life of bulb.
→ claim of the producer is correct.

---

eg: (2) A random Sample of Size 15 has 50 as mean, the Sum of the Squares of the deviation taken from mean is 130,
→ can this Sample be regarded as taken from the population having 53 as mean?

→ obtain 95% and 99% Confidence limits of the mean for the population.

Given data,

$\bar{x} = 50, \ \mu = 53, \ n = 15, \ \underline{\Sigma(x-\bar{x})^2} = 130$

$H_0 : \mu = \bar{x}, \quad H_a : \mu \neq \bar{x},$

$$t = \frac{\bar{x} - \mu}{S} \times \sqrt{n}, \quad S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$S = \sqrt{\frac{130}{15-1}} = \sqrt{\frac{130}{14}}$$

$$S = \sqrt{9.29} = \underline{\underline{3.05}}$$

$$t = \frac{\overset{(\bar{x} - \mu)}{50 - 53}}{3.05} \times \sqrt{15} = \frac{-3 \times 3.87}{3.05}$$

$$= \frac{11.61}{3.05} = \boxed{3.81} \text{ (calculated 't' value)}$$

ν (or) $df = n - 1 = 15 - 1 = 14$

95% ⇒ Limit for 5% (or) ...

99% ⇒ 1% level of significance

Condifidance Limit

$t_{0.05}^{(95\%)} = \boxed{2.145}$

$t_{0.01}^{(99\%)} = \boxed{2.977}$ (Two tail test)

(df = 14 ; Level of signficance = 1% / 5% (two tailed)

<u>Sample mean for 95% Confidence limit</u>:

$$= \bar{x} \pm \frac{S}{\sqrt{n}} \times t_{0.05}$$

$$= 50 \pm \frac{3.05}{\sqrt{15}} \times 2.145 = 50 \pm \frac{3.05}{3.87} \times 2.145$$

$$= 50 \pm \frac{6.54}{3.87} = 50 \pm 1.69$$

Limit = $\underline{48.31}$ to $\underline{51.69}$

Sample mean for 99 % confidence limit.

$$\bar{x} \pm \frac{s}{\sqrt{n}} \times t_{0.001}$$

$$= 50 \pm \frac{3.05}{\sqrt{15}} \times 2.977 = 50 \pm \frac{3.05}{3.87} \times 2.977$$

$$= 50 \pm \frac{9.08}{3.87} = 50 \pm 2.35$$

Limit. 47.65 to 52.35

TV ↓ = CV ↑ = Significant difference

Ho will be failed (rejected)

i.e, There is a difference between sample mean & popn. mean

For Two Independent Samples (unpaired test)

Q ① Two type of drugs were used on 6 and 5 patients for reducing their weight. Drug A was imported and Drug B was indegenous. The increase in the weight (in kg) after using the drug for 90 days was given below:

Drug A : 8, 10, 12, 9, 14, 13 - kg

Drug B : 7, 9, 14, 12, 8. - kg

is there a significant difference in a efficacy of the drug?

Solution:

H_0 : Drug A = Drug B

H_a : Drug A $\neq$ Drug B

Given data:

$n_1 = 6$

$n_2 = 5$,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}}$$

$$S = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

Calculation of $\bar{x}_1$, $\bar{x}_2$ and S :

x_1	$(x_1 - \bar{x}_1)$	$(x_1 - \bar{x}_1)^2$	x_2	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
8	$8-11 = -3$	9	7	$7-10 = -3$	9
10	$10-11 = -1$	1	9	$9-10 = -1$	1
12	$12-11 = +1$	1	14	$14-10 = +4$	16
9	$9-11 = -2$	4	12	$12-10 = +2$	4
14	$14-11 = +3$	9	8	$8-10 = -2$	4
13	$13-11 = +2$	4			
			50		
66		28			34

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{66}{6} = 11$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{50}{5} = 10$$

$$S = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

$$S = \sqrt{\frac{28 + 34}{6+5-2}} = \sqrt{\frac{66}{9}} = \sqrt{6.88} = \underline{\underline{2.62}}$$

$$t = \frac{11-10}{2.62} \times \sqrt{\frac{6\times5}{6+5}}$$

$$t = \frac{1}{2.62} \times \sqrt{\frac{30}{11}} = \frac{1}{2.62} \times \sqrt{2.72}$$

$$t = \frac{1}{2.62} \times 1.65 = \frac{1.65}{2.62} = \underline{\underline{0.63}}$$

$t = \boxed{0.63}$ (Calculated Value)

df (or) $\nu = n_1 + n_2 - 2 = 6 + 5 - 2 = 9$

$t_{0.05} = \boxed{2.262}$ (Tabulated Value)

Ho is passed and accepted, so there is no significant difference between the efficacy of Drug A and Drug B

Q(2) Two Laboratories A and B carry out independent estimates of Sugar Content in a Chocolate made by a firm named X. A sample is taken from each batch and devided into two parts; and send to the two Laboratories. The sugar Content Obtained by the laboratories are regarded below (in mg/g)

Lab A : 8, 9, 9, 6, 4, 6

Lab B : 7, 8, 6, 4, 5, 6

Is there a significant difference between the mean sugar content obtained by two laboratories A & B.

solution

Ho : Lab A = Lab B

Ha : Lab A ≠ Lab B

Given data, (Sugar Content in chocolate in mg/g)

Lab A : 8, 9, 9, 6, 4, 6 $= n_1 = 6$

Lab B : 7, 8, 6, 4, 5, 6 $= n_2 = 6$

Calculation of $\bar{x}_1$, $\bar{x}_2$ and S.

x_1	$(x_1 - \bar{x}_1)$	$(x_2 - \bar{x}_2)^2$	x_2	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
8	8-7 = 1	1	7	7-6 = 1	1
9	9-7 = 2	4	8	8-6 = 2	4
9	9-7 = 2	4	6	6-6 = 0	0
6	6-7 = -1	1	4	4-6 = -2	4
4	4-7 = -3	9	5	5-6 = -1	1
6	6-7 = -1	1	6	6-6 = 0	0
42		20	36		10

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{42}{6} = \boxed{7}$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{36}{6} = \boxed{6}$$

$$S = \frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$$S = \frac{20 + 10}{6 + 6 - 2} = \frac{30}{10} = \sqrt{3} = \underline{1.732}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S}$$

$$t = \frac{7 - 6}{1.732} \times \sqrt{\frac{6 \times 6}{6 + 6}} = \frac{1}{1.732} \times \frac{36}{12}$$

$$t = \frac{1}{1.732} \times \sqrt{3} = \cancel{\frac{1}{1.732}} \times \cancel{1.732} = \underline{1}$$

$t = \boxed{1}$ (calculated value)

$\nu\,(df) = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$

$t_{0.005} = \boxed{2.228}$ (tabulated 't' value)

Ho is correct and accepted. So, there is no significant difference between the sugar content obtained by Lab A and Lab B.

For two Dependent Samples (paired test)

Q(1) A drug is given to 8 patients and the difference in their blood pressure were recorded to be:

Before Drug A : 112, 113, 118, 120, 119, 113, 110, 122

After Drug A : 116, 120, 117, 125, 126, 111, 111, 117,

Is it responsible to believe that the drug has no effect on change of Blood pressure?

Solution:

Ho : Blood pressure before drug A = BP After drug A

Ha : BP before drug A $\neq$ BP After drug A

Given data:

$$t = \frac{\bar{d}\sqrt{n}}{S}$$

$$S = \sqrt{\frac{\Sigma (d-\bar{d})^2}{n-1}}$$

Calculation of $\bar{d}$ and S

Before drug	After drug	d	$(d-\bar{d})$	$(d-\bar{d})^2$
112	116	116-112 = +4	4 - 2 = 2	4
113	120	120-113 = +7	= 5	25
118	117	117-118 = -1	= -3	9
120	125	125-120 = +5	= +3	9
119	126	126-119 = +7	= 5	25
113	111	111-113 = -2	= -4	16
110	111	111-110 = +1	= -1	1
122	117	117-122 = -5	= -7	49
		16		138

$$\bar{d} = \frac{\Sigma d}{n} = \frac{16}{8} = 2$$

$$S = \sqrt{\frac{138}{8-1}} = \sqrt{\frac{138}{7}} = \sqrt{19.71} = 4.44$$

$$t = \frac{\bar{d} \times \sqrt{n}}{S} = \frac{2 \times \sqrt{8}}{4.44} = \frac{2 \times 2.83}{4.44} = \frac{8.49}{4.44} = 1.912$$

$t = \boxed{1.912}$ (calculated value)

$df = n-1 = 8-1 = 7$

$t_{0.05} = \boxed{2.365}$ (Tabulated 't' value)

Ho is correct and accepted. So, drug has no Significant role in the change of the blood pressure.

Q (2) 7 Students were given intensive coaching and 5 tests were conducted in a month, the scores of test 1 and 5 are given below:

Test 1: 52, 43, 52, 27, 36, 43, 61
Test 5: 63, 41, 62, 36, 31, 53, 70

Does the scores from 1 to 5 show an improvement?

Solution:

H_0: marks before coaching = marks after coaching

H_0: marks before coaching $\neq$ marks after coaching

Given data:

$$n = 7, \quad t = \frac{\bar{d} \times \sqrt{n}}{S}, \quad S = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}$$

marks in test (1)	marks in test (5)	d	$(d-\bar{d})$	$(d-\bar{d})^2$
52	63	11	5	25
43	41	-2	-8	64
52	62	10	4	16
27	36	9	3	9
36	31	-5	-11	121
43	53	10	4	16
61	70	9	3	9
		42		260

$$\bar{d} = \frac{\Sigma d}{n} = \bar{d} = \frac{42}{7} = \underline{6}$$

$$S = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}} = \sqrt{\frac{260}{7-1}} = \sqrt{\frac{260}{6}} = \sqrt{43.33} = 6.58$$

$$t = \frac{\bar{d} \times \sqrt{n}}{S} = \frac{6 \times \sqrt{7}}{6.58} = \frac{6 \times 2.65}{6.58} = \frac{15.9}{6.58} = 2.416$$

$t = \boxed{2.416}$ (calculated t value)

$df = n-1 = 7-1 = 6$

$t_{0.05} = \boxed{2.447}$ (Table t value)

H0 is correct and accepted. There is no improvement in the marks after the intesive coaching.

For Observed Coefficient :

Q 1 A random sample of 27 pair of observations from a normal population gives a correlation coefficient of 0.55.
Is it likely that the variables in the population are uncorrelated ?

Solution :

H0 : Correlation = Not Significant
Ha : Correlation ≠ not Significant

Given data,

$n = 27$, $r = 0.55$

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

r = Correlation coefficient
n = Sample size.

$$t = \frac{0.55}{\sqrt{1-(0.55)^2}} \times \sqrt{27-2}$$

$$t = \frac{0.55}{\sqrt{1-0.3025}} \times \sqrt{25} = \frac{0.55}{\sqrt{0.6975}} \times 5 = \frac{2.75}{0.835}$$

$t = \boxed{3.293}$ (calculated value)

$df = n-2 = 27-2 = 25$

$t_{0.05} = \boxed{2.060}$ (Tabulated 't' value)

Ho failed and rejected, So variable in the population are uncorrelated.

Q(2) Is a correlation coefficient of 0.6 significant? If obtained from a random sample of 11 pair of values from a normal population.

Given data,

$n = 11$, $r = 0.6$

Ho : Correlation = Not Significant

Ha : Correlation ≠ not Significant

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

r = Correlation Coefficient

n = Sample size.

$$t = \frac{0.6}{\sqrt{1-(0.6)^2}} \times \sqrt{11-2}$$

$$t = \frac{0.6}{\sqrt{1-(0.36)}} \times \sqrt{9} = \frac{0.6}{\sqrt{0.64}} \times 3 = \frac{0.6}{.8} \times 3 = \frac{1.8}{0.8}$$

$t = \boxed{2.25}$ (calculated 't' value)

$df = n-2$; $11-2 = 9$

$t_{0.05} = \boxed{2.262}$ (table value)

H_0 is passed and accepted. So, correlation coefficient is not significant.

மாறுபாட்டுப் பரவற்படி ஆய்வு

(Analysis of Variance)

* இரண்டிற்கு மேற்பட்ட மாதிரிகளுக்கிடையேயான சராசரிகளை ஒப்பிட்டுப் பார்க்க அனோவா உதவுகிறது.

* அனோவா என்ற முறையை ஃபிஷர் (R.A. Fisher) என்பவர் பயன்படுத்தினார். எனவே இதற்கு F சோதனை என்ற பெயர் உண்டு.

* அனோவாவிற்கான சார்பில சோதனை அளவை

$$F = \frac{\text{மாதிரிகளுக்கிடையேயான மாறுபாடு}}{\text{மாதிரிகளுக்குள் உள்ள மாறுபாடு}} = \frac{\text{Variance between Samples}}{\text{Variance within Samples}}.$$

"அனோவா"வின் நோக்கம்:

i) முதல் வகை பிழையை அதிகரிக்கச் செய்யாமல், பல இனக் குழுக்களின் சராசரிகளை ஒப்பீடு செய்யப் பயன்படுகிறது (உண்மையான இன்மை எடுகோளை (H0) தவறுதலாக நிராகரிப்பது முதல்வகைப் பிழையாகும்.

ii) மாதிரிகளின் மொத்த வேறுபாடுகள் பல பகுதிகளாகப் பிரிக்கப்பட்டு ஆய்வுக்கு உபயோகப்படுத்தப்படுகிறது.

iii) சோதனை ஆராய்ச்சிகளிலும் கணக்கெடுப்பு ஆராய்ச்சிகளிலும் இது பயன்படுகிறது.

அனோவாவின் அடிப்படை எடுகோள்கள்:

i). ஒவ்வொரு மாதிரியும் இயல்நிலைப் பரவலிலிருந்து பெறப்பட்டது என்று கருதப்படுகிறது.

ii) தொகுதியின் (குழுக்கள்) மாறுபாடுகள் ஒரே மாதிரியானவை என்று கருதப்படும் அதாவது

$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ என

iii) ஒவ்வொரு மதிப்பீடும் மற்றவற்றை சாராதது என்று கொள்கிறோம்.

மாறுபாட்டுப் பகுப்பாய்வின் வகைகள்:

i) ஒரு வழி மாறுபாட்டு பகுப்பாய்வு (One way ANOVA)

ii) இருவழி மாறுபாட்டு " " (Two way - ANOVA)

ஒரு வழி அனோவா:

தரவுகளை ஒரே ஒரு காரணியின் அடிப்படையில் வகைப்படுத்தும் முறையாகும்

கணக்கீடு முறை:

(1) ஒவ்வொரு பகுதியிலும் உள்ள மதிப்புகளின் கூட்டுத்தொகை மற்றும் அதன் சராசரியை கண்டுபிடித்தல்.

(2) பகுதிகளின் சராசரியை மொத்தமாக கூட்டல் செய்து மொத்த தொகையை கண்டுபிடித்தல்.

(3) மாதிரிகளின் வர்க்கங்களின் கூட்டல் (CSS between)

(4) மாதிரிகளின் வர்க்க கூட்ட சராசரி (MS between)

(5) உள்ளிருக்கும் வர்க்கங்களின் கூட்டல்

- (SS - within)

(6) உள்ளிருக்கும் வர்க்கசராசரி

(MS - within)

(7) ANOVA அட்டவணை கீழ்கண்டவாறு எழுதலாம்

வேறுபாட்டின் ஆதாரங்கள்	சுதந்திர மாறிகள் (df)	வர்க்கங்களின் கூட்டல் (S-S)	வர்க்கசராசரி (M.S)	கணக்கிடப்பட்ட F-மதிப்பு
குழுக்களுக்கிடையே				
குழுக்களுக்குள்				
மொத்தம்				

- கணக்கிடப்பட்ட F மதிப்பு பட்டியல் 'F' மதிப்பை விட அதிகமாக இருந்தால் ($P < 0.01$). கருதுகோள் நிராகரிக்கப் படுகிறது. அதாவது வித்தியாசம் உள்ளது என்று அனுமானம் கொள்ளலாம்.
- கணக்கிடப்பட்ட 'F' மதிப்பு பட்டியல் 'F' மதிப்பைவிட குறைவாக இருந்தால் ($P > 0.01$). கருதுகோள் ஏற்றுக்கொள்ளப்படுகிறது. அதாவது எந்தவொரு வேறுபாடும் இல்லை (குழுக்களுக்கிடையேயும், குழுக்களுக்குள்ளும்) என்று அனுமானம் கொள்ளலாம்.

Analysis of Variance (ANOVA)

- The term ANOVA was first proposed by R.A Fisher.
- Analysis of variance refers to the examination of differences among the samples.
- It is an extermely useful technique concerning research in Biology
- It is used to examine the significance of the difference amongst more than two sample means at the same time.
- The analysis of variance has been classified into
  - One way classification &
  - Two-way classification

Principles:

We take two estimates of population variance i.e., one based on between samples variance and the other within samples variance. Then these two estimates of population variance are compared with 'F' test as follows

$$F = \frac{\text{Variance between Samples}}{\text{Variance within Samples}}$$

The value of F is to be compared to the F-limit for a given degrees of freedom. If the calculated F value exceeds the F-table value we can say that there are significant variance between the sample means.

Steps involved in the Analysis are:

Step: 1

Find out the means of each samples $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4 ---- \bar{x}_k$

Step: 2

Find out the Combined mean of the samples

$$X = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 \ldots \bar{X}_K}{\text{No. of Samples}}$$

Step: 3

Sum of Squares between the Samples (or) SS-between

$$\therefore \text{SS-between} = n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + n_3(\bar{x}_3-\bar{x})^3 + n_4(\bar{x}_4-\bar{x})^2 \ldots n_k(\bar{X}_k-\bar{X})^2$$

n = number of items in the corresponding Samples.

Step: 4

Mean square between the Samples (or) M.S-between.

$$\therefore \text{MS-between} = \frac{\text{SS- between}}{\text{degrees of freedom between the Samples.}}$$

Step: 5

Sum of squares within the Samples (or) SS-within

$$\therefore \text{SS within} = \Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2 + \Sigma(x_3-\bar{x}_3)^2 \; \Sigma(x_4-\bar{x}_4)^2 \ldots \Sigma(x_k-\bar{x}_k)^2$$

Step: 6

Mean square within the Samples (or) M-S within.

$$\therefore \text{M.S within} = \frac{\text{SS- within}}{\text{degrees of freedom within the Samples}}$$

Step: 7 Make ANOVA Table:

Source of Variance	Sum of Square (SS)	Degree of freedom	Mean Square (MS)
Between Sample			
Within samples			
Total			

Step: 8 Find out F - value

$$F = \frac{\text{Variance between Samples}}{\text{Variance within Samples}} = \frac{\text{MS between}}{\text{MS within}}$$

If the calculated F-value is less than F-Table value, there is no significant.

Illustration : ANOVA — One way.

A certain manure was used on four plots of land A. B. C. and D. Four beds were prepared in each plot and the manure used. The output of the crop in the beds of plots. A, B, C and D is given below

A	B	C	D
6	15	9	8
8	10	3	12
10	4	7	1
8	7	1	3

using ANOVA find out whether the difference in the means of the production of crops of the plots is significant or not.

Solution :

Step 1: Find out the means of each samples.

	Sample I (x_1)	Sample II (x_2)	Sample III (x_3)	Sample IV (x_4)
	6	15	9	8
	8	10	3	12
	10	4	7	1
	8	7	1	3
Tot :	32	36	20	24
$\bar{x}$	8	9	5	6

Step : 2 Find out the Combined mean of the Samples.

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 - \bar{X}_4}{\text{No. of Samples}} = \frac{8+9+5+6}{4}$$

$$= \frac{28}{4} = 7 \qquad \bar{X} = 7$$

Step : 3 Sum of Squares between the samples
(or) SS- between

$\therefore$ SS between $= n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + n_3(\bar{x}_3-\bar{x})^2 + n_4(\bar{x}_4-\bar{x})^2$

$= 4(8-7)^2+4(9-7)^2+4(5-7)^2+4(6-7)^2$

$= 4(1)^2+4(2)^2+4(-2)^2+4(-1)^2$

$= 4(1)+4(4)+4(4)+4(1)$

$= 4+16+16+4$

$= 40$

Step : 4 - Mean Square between the samples
(or) MS - between

$\therefore$ MS-between $= \dfrac{\text{SS-between}}{\text{degree of freedom between the samples}}$

There are four samples so the degrees of freedom are $4-1 = 3$

$\therefore$ MS- between $= \dfrac{40}{3} = 13.33$

Step : 5 Sum of Squares within the Samples (or) SS-within.

$\therefore$ SS- within $= \Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2 + \Sigma(x_3-\bar{x}_3)^2 + \Sigma(x_4-\bar{x}_4)^2$

x_1	$(x_1-\bar{x}_1)$ $\bar{x}_1 = 8$	$(x_1-\bar{x}_1)^2$	x_2	$(x_2-\bar{x}_2)$ $\bar{x}_2=9$	$(x_3-\bar{x}_3)^2$
6	$6-8=-2$	4	15	$15-9=6$	36
8	$8-8=0$	0	10	$10-9=1$	1
10	$10-8=2$	4	4	$4-9=-5$	25
8	$8-8=0$	0	7	$7-9=-2$	4
		8			66

x_3	$(x_3-\bar{x}_3)$ $\bar{x}_3=5$	$(x_3-\bar{x}_3)^2$	x_4	$(x_4-\bar{x}_4)$ $\bar{x}_4-6$	$(x_4-\bar{x}_4)^2$
9	$9-5=4$	16	8	$8-6=2$	4
3	$3-5=-2$	4	12	$12-6=6$	36
7	$7-5=2$	4	1	$1-6=-5$	25
1	$1-5=-4$	16	3	$3-6=-3$	9
		40			74

$\therefore$ SS-within $= 8+66+40+74$
$= 188$

Step :6 - Mean square within the samples (or) MS- within

$$\therefore \text{MS-within} = \frac{\text{SS-within}}{\text{degree of freedom within the samples}}$$

There are 16 items within the 4 samples

$\therefore$ degrees of freedom $16-4=12$

$\therefore$ MS- within $= \frac{188}{12} = \underline{15.66}$

Step : 7 make ANOVA Table :

Source of Variance	Sum of Squares SS	Degree of freedom (dof)	Mean Square (MS)
Between Sample	40	3	13.33
within Sample	188	12	15.66
Total	228	15	

Step : 8 Find out F value

$$F = \frac{\text{Variance between Samples}}{\text{Variance within Samples}} = \frac{13.33}{15.66} = \underline{\underline{0.851}}$$

The table value of F for $V_1 = 3$ and $V_2 = 10$ } at 5% level of significance $= \underline{\underline{3.49}}$

Step : 9 Inference:

The calculated value (0.851) is lesser than the table value (3.49). Therefore the difference in the means of the production of crops of the plots is not significant.

SPSS மென் பொருள்

(ஸ்டாடிஸ்டிக்கல் பாக்கேஜ் ஃபார் சோசியல் சயன்ஸ்)

முன்னுரை :

SPSS என்பது புள்ளியல் ஆராய்ச்சிக்கான ஒரு கணினி மென்பொருள். நீ மற்றும் ஹல் என்னும் இரு விஞ்ஞானிகளால் உருவாக்கப்பட்டது.

நீ இதனை உருவாக்கும்பொழுது விஞ்ஞானத்தில் மேற்படிப்பு பட்டதாரியாக இருந்தார். ஹல் ஸ்டான்ஃபோர்டு பல்கலைக்கழகத்தில் ஆராய்ச்சி பேராசிரியராகவும், சிகாகோ பல்கலைக்கழகத்தில் [illegible] பேராசிரியராகவும் பணியாற்றினார்.

முதல் முதல் உருவாக்கிய போது இம்மென்பொருள் P.A.S.W என அழைக்கப்பட்டது அதாவது ப்ரடிக்டிவ் அனாலிடிக் சா:ப்ட்வேர். இதனை 2009-ல் IBM கம்பெனியானால் 1.2 பில்லியன் டாலருக்கு வாங்கப்பட்டு பின் SPSS - ஆனது IBM கம்பெனி என வெளியிடப்பட்டது.

பயன்படுத்துவோர் :

1. SPSS பெரும்பான்மையாக சோசியல் சயன்ஸ் பிரிவினரால் புள்ளியல் ஆராய்ச்சிகளுக்கு பயன்படுத்தப்படுகிறது.

2. மேலும் சந்தை ஆராய்ச்சியாளர்கள், உடல் நல ஆராய்ச்சியாளர்கள், சர்வே கம்பெனிகள், அரசாங்கம், கல்வி ஆராய்ச்சியாளர்கள், வியாபார அமைப்புகள் போன்ற பல பிரிவினராலும் பயன்படுத்தப்படுகிறது.

SPSS என்ற மென்பொருளை பயன்படுத்தி பின்வரும் பணிகளை நாம் எளிதில் செய்ய முடியும்.

1. புள்ளியல் ஆராய்வு

இதில் நான்கு வகை ஆராய்ச்சிகள் செய்யலாம்

i) விவரிக்கும் புள்ளியல்:

தொகுக்கப்பட்ட புள்ளி விவரங்களை அட்டவணையாக்குதல், எத்தனை முறை ஒவ்வொரு மதிப்பும் உள்ளது என்று எண்ணிக்கையிடுதல் போன்றவை.

ii) இருவேறுபாட்டு புள்ளியியல்:

சராசரி, முகடு, t-சோதனை, அனோவா, தொடர்புபடுத்துதல் போன்றவை.

iii) எண் மதிப்புகளிலிருந்து அடுத்த போக்கை யூகித்தல்:

தொடர்பு போக்கு (ரிக்ரஷன்)

iv) புகைப்படத்தின் மூலம் ஒத்த கூட்டங்களை கண்டறிதல்

2. வரைபடத்தை நிரூபித்தல்:

எ.கா. காரணி சோதனை

கொத்து சோதனை

3. வரைபடத்தை பதிவுசெய்தல்: documentation.

SPSS-ன் பல பயன்பாடுகள்:

1. SPSS-ன் பல பயன்பாடுகள் கணினியின் திரையில் தெரியும். சிறுபடங்கள் அமைப்புகள் மெனுவின் மூலம் பயன்படுத்த முடியும்.

2. அதிகப் பயனுள்ள மென்பொருளை கட்டளை கொடுக்கும் மொழியின் மூலம் பயன்படுத்தலாம்

3. பைதான் மொழியின் மூலம் புது பயன்பாடுகளை மென்பொருளாக்க முடியும்

4. புள்ளி விவரங்கள் சீராகவும் ஒழுங்காகவும் திரையில் தெரியும் வண்ணம் தாளில் (Spread sheet) உள்ள கட்டங்களில் பதிவு செய்யலாம். அதில் வரிசைகளும், கலங்களும் இருக்கும். அந்த கட்டங்களில் எண் (or) எழுத்து மட்டுமே பதிவு செய்யலாம்.

5. பாஸ்வேர்டுகளை சேமித்து வைக்க இயலாது.

6. புள்ளிவிவரம் சேகரிக்கும் சர்வே போன்றவற்றில் மதிப்புகள் மிக அதிகமாக இருக்கும். அப்பொழுது ஒவ்வொரு மதிப்பையும் பதிவு செய்தல் கடினமானது. அதனால் கீழ்க்கண்ட முறைகள் கையாளப்படுகிறது.

i) டேட்டா என்ட்ரி ஆபரேட்டர்

ii) கணினி மூலம் தொலைபேசி நேர்காணல் செய்தல்.

iii) புள்ளிவிவரத்தை ஸ்கேன் செய்தல்

iv) அடையாள குறியீடு உணர்தல், அடையாள மார்க் உணர்தல் மென்பொருள் (OMR) பயன்படுத்துதல்

v) ஆன்லைனில் கேள்வித்தாளை பதிவு செய்தல் கேள்விகளில் பெற்ற மதிப்புகளை SPSS-ல் எளிதாக நுழைக்க முடியும்.

vi) SPSS என்பது ASC II டெக்ஸ்ட் ஃபைல்கள், மற்ற புள்ளியியல் மென்பொருள்கள் டேட்டாபேஸ் போன்றவற்றிலிருந்து மதிப்புகளை வாசிக்கவும், எழுதவும் செய்கின்றது.

vii) SPSS-ன் வெளியீடு சாளரத்தை ஃபைலாக (.sp. ஃபைல்) அதை சேமிக்கும். இதனை MS-word, எக்ஸல், பிடிஎஃப் மற்றும் HTML. XML போன்ற எந்த ஒரு ஃபைலாகவும் மாற்ற இயலும் இல்லை JPEG, BMP போன்ற வரைபடங்களாகவும் மாற்ற இயலும்.

SPSS-ன் வரையறை (Limitations):

SPSS-ல் மதிப்புகளை மாற்ற இயலாது பயன்படுத்துவோருக்கு புள்ளியியல் வெளிப்பாட்டின் மிக குறைந்த கட்டுப்பாடே உள்ளது. அதனால் ஒரு நபர் தான் விரும்புவது போல் வெளியீடு கிடைக்கவில்லை என நினைத்தால் அவர் வேறு மற்ற மென்பொருட்களை பயன்படுத்த வேண்டும் (or) SPSS-ன் வரைமுறைக்கு உட்பட்டு செயல்பட வேண்டும்.

Introduction to statistical package-SPSS.

What is SPSS?

SPSS (Statistical Package for the Social Sciences) is among the most widely used programs for statistical analysis in social science. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and conduct complex statistical analyses.

1. It is developed by Norman H. Nie and C. Hadlai Hull of IBM Corporation in the year 1968.
2. Long produced by SPSS Inc., it was acquired by IBM in 2009. Current versions (post 2015) have the brand name: IBM SPSS Statistics
3. It is compatible with Windows, Linux, UNIX About & Mac operating systems.
4. The original SPSS manual Nie, Bent & Hull, 1970 has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis.
5. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the data file) are features of the base software.Statistics included in the base software:
   - Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics.
   - Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests, Bayesian
   - Prediction for numerical outcomes: Linear regression.
   - Prediction for identifying groups: Factor analysis, cluster analysis (two-step, K-means, hierarchical).
   - Geo spatial analysis, simulation R extension (GUI), Python.
6. The many features of SPSS Statistics are accessible via pull-down menus or can be programmed with a proprietary 4GL *command syntax language*.
7. Command syntax programming has the benefits of reproducible output, simplifying repetitive tasks, and handling complex data manipulations and analyses.
8. SPSS is among the most widely used programs for statistical analysis in social science and Market research. Government, Higher education, Consumer packaged goods, Retail, Manufacturing, Healthcare, Insurance, Finance, Banking, Telecommunications.
9. It offers complete plotting, reporting and presentation features.

10. It provides in-depth statistical capabilities.

11. It includes a full range of data, management system and editing tools.

12. It is easy to learn and use

13. Features of SPSS:
    - Importing data from other file formats
    - Importing data from an ASCII file
    - Opening existing SPSS system files
    - Creating new SPSS data files
    - Getting data into SPSS
14. One can have only one data file open at a time.
    - One can create new data files or modify existing ones.
    - This window displays the content of the data file.
    - The data editor offers a simple and efficient spreadsheet like facility for entering data and browsing the working data file.
15. Entering Data Editor._ Displays variable definition information, including defined variable and value labels, data type, etc., Displays the actual data values or defined value labels.
16. This editor provides two views of the data, DATA VIEW

17. _We can edit text, swap data in rows and columns.

18. Output can be modified in many ways with is editor, and can create multidimensional tables. Ex: Editing Data PIVOT TABLE EDITOR.

19. High-resolution charts and plots can be modified in chart windows. Text output not displayed in pivot tables can be modified with the text output editor.

Advantages:

SPSS offers a user friendliness that most packages are only now catching up to. It is popular, and though that is certainly not a reason for choosing a statistical package, many data sets are easily loaded into it and other programs can easily import SPSS files.

Disadvantage:

- It is expensive, sometimes ridiculously so, and even when you do buy your really only leasing, and its license is definitely not user friendly.
- Its menu offerings are typically the most basic of an analysis and sometimes lacking even then, and it makes doing an inappropriate analysis very easy.
- For academic use SPSS lags notably behind SAS, R and even perhaps others that are on the more mathematical rather than statistical side for modern data analysis.