**4. Farmers use apps to access information throughout the month. Using the data, check whether app usage is same or different across the four weeks of a month. Anand claims that app usage picked up after January 2016; so, test this hypothesis using data from January-2016 – May 2018.**

Solution:

1. **We test the hypothesis whether disease access is same across four weeks of a month is same or not**

H0 : Disease access is same across months. i.e week1= week2 = week 3 = week4

H1 : Disease access is not same across months.

Assumptions: Data is assumed to be normal.

Using Anova for this, we get

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
| Week1 | 31 | 17110 | 551.94 | 147469 |
| Week2 | 30 | 15666 | 522.2 | 124410 |
| Week3 | 30 | 14401 | 480.03 | 111122 |
| Week4 | 32 | 24470 | 764.69 | 510324 |

Analysis of variance gives,

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 2E+06 | 3 | 505059 | 2.2199 | 0.0894 | 2.6808 |
| Within Groups | 3E+07 | 119 | 227517 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 3E+07 | 122 |  |  |  |  |

Based on p-value, we retain H0. i.e we conclude that – disease access is same across months.

1. **To test whether app usage picked up after Jan 2016 or not.**

**Approach :** We divide the usage data into two groups, one before Jan 2016 and one after Jan 2016. We test whether usage is increased or not/

H0 U2<=U1

H1 U2> U1 right tailed test

|  |  |  |
| --- | --- | --- |
|  | Before 2016 | After 2016 |
| Mean | 287.68 | 657.7 |
| sigma | 254.4374 | 500.61 |
| n | 25 | 98 |

Calculating the below values, we retain H0.i.e we conclude that, app usage has not increased after Jan 2016.

df 121

sp^2 213739.7178

t-stat -3.57205272

t-cric 1.657544319

1. **Anand claims that number of users have increased over a period of two years. He wants to understand if app usage (number of times his app is accessed in a month by various users) has increased with the increased number of users. Prove this claim statistically. Also suggest a suitable statistical test to prove that the correlation between users and usage is non-zero.**

Part A: Whether number of users have increased over a period of two years or not.

Apporach : we divide the data into two parts, one – first two years (Jun 15 – May 17) and the other, rest of the data. We test whether there is a significant increase in number of users or not in the second set.

H0 U2<=U1

H1 U2> U1 right tailed test

|  |  |  |
| --- | --- | --- |
|  | First Two Years | Later Period |
| Mean | 56.354839 | 248.9667 |
| sigma | 38.885463 | 92.5361 |
| n | 93 | 30 |

df 121

sp^2 3201.952537

t-stat -16.21156917

t-crit 1.657544319

We retain H0. We conclude that, **number of users has not increased over two years.**

**Part B: To test whether app usage increased with number of users and to find the correlation co-efficient.**

Approach : We divide the number of users and usage into different bins and calculate the corresponding frequencies. We test whether number of users and usage are independent or not.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| User Bin\Usage Bin | Greater than 3000 | Less than 1000 | Less than 2000 |  |
| Greater than 400 |  |  | 2 | 2 |
| Less than 100 |  | 85 | 2 | 87 |
| Less than 200 |  | 16 | 1 | 17 |
| Less than 300 | 1 | 1 | 8 | 10 |
| Less than 400 |  |  | 7 | 7 |
|  | 1 | 102 | 20 | 123 |

H0 : Number of users and Usage are independent.

H1 : Number of users and usage are dependent.

We use X^2 test of independence.

**Expected frequencies** under H0 are as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| User Bin\Usage Bin | Greater than 3000 | Less than 1000 | Less than 2000 |
| Greater than 400 | 0.016260163 | 1.65853659 | 0.32520325 |
| Less than 100 | 0.707317073 | 72.1463415 | 14.1463415 |
| Less than 200 | 0.138211382 | 14.097561 | 2.76422764 |
| Less than 300 | 0.081300813 | 8.29268293 | 1.62601626 |
| Less than 400 | 0.056910569 | 5.80487805 | 1.13821138 |

Calculating the (e-e0)2/e0

|  |  |  |  |
| --- | --- | --- | --- |
| User Bin\Usage Bin | Greater than 3000 | Less than 1000 | Less than 2000 |
| Greater than 400 | 0.016260163 | 1.65853659 | 8.62520325 |
| Less than 100 | 0.707317073 | 2.29001962 | 10.4291001 |
| Less than 200 | 0.138211382 | 0.25673053 | 1.12599235 |
| Less than 300 | 10.38130081 | 6.41327116 | 24.9860163 |
| Less than 400 | 0.056910569 | 5.80487805 | 30.1882114 |

X^2 stat 103.0779593

df 8

X2 critical 15.50731306

p-val 1.00156E-18

We reject H0, i.e we conclude that Users and Usage are dependent

Correlation coefficient between users and usage is 0.85 and the graph of usage(y-axis) and number of users(x-axis) is below.

**7. If a disease is likely to spread in particular weather condition (data given in the disease index sheet), then the access of that disease should be more in the months having suitable weather conditions. Help the analyst in coming up with a statistical test to support the claim for two districts for which the sample of weather and disease access data is provided in the data sheet. Identify the diseases for which you can support this claim. Test this claim both for temperature and relative humidity at 95% confidence.**

We take cities as proxy for different whether conditions and test whether disease access is same in different cities(whether conditions)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Belgavi | | | Dharwad | | | Hypothesis, H0 : Bel = Dhar, H1 : Bel <> Dhar | | | | | | |
| Mean | Std. Dev | n1 | Mean | Std. Dev | n2 | Sp^2 | d.f | t-stat | T-left | T-right | P-value | Conclusion |
| D1 | 21.545 | 25.035 | 24 | 14.494 | 16.684 | 22 | 460.47 | 44 | 1.11 | -2.02 | 2.02 | 0.27 | Retain H0 |
| D2 | 15.909 | 15.647 | 24 | 13.878 | 21.655 | 22 | 351.78 | 44 | 0.37 | -2.02 | 2.02 | 0.72 | Retain H0 |
| D3 | 22.091 | 22.996 | 24 | 29.976 | 43.681 | 22 | 1187.10 | 44 | -0.78 | -2.02 | 2.02 | 0.44 | Retain H0 |
| D4 | 15.331 | 13.147 | 24 | 15.798 | 20.677 | 22 | 294.40 | 44 | -0.09 | -2.02 | 2.02 | 0.93 | Retain H0 |
| D5 | 14.859 | 16.105 | 24 | 13.269 | 18.066 | 22 | 291.36 | 44 | 0.32 | -2.02 | 2.02 | 0.75 | Retain H0 |
| D6 | 47.539 | 47.259 | 24 | 43.429 | 74.933 | 22 | 3847.32 | 44 | 0.22 | -2.02 | 2.02 | 0.82 | Retain H0 |
| D7 | 27.434 | 29.441 | 24 | 21.280 | 27.689 | 22 | 819.01 | 44 | 0.73 | -2.02 | 2.02 | 0.47 | Retain H0 |
| D8 | 21.165 | 19.089 | 24 | 29.434 | 53.459 | 22 | 1554.47 | 44 | -0.71 | -2.02 | 2.02 | 0.48 | Retain H0 |
| D9 | 23.837 | 22.189 | 24 | 21.893 | 35.227 | 22 | 849.64 | 44 | 0.23 | -2.02 | 2.02 | 0.82 | Retain H0 |
| D10 | 25.750 | 30.783 | 24 | 15.219 | 27.707 | 22 | 861.71 | 44 | 1.22 | -2.02 | 2.02 | 0.23 | Retain H0 |
| D11 | 22.924 | 25.577 | 24 | 16.737 | 24.239 | 22 | 622.39 | 44 | 0.84 | -2.02 | 2.02 | 0.41 | Retain H0 |

1. **Anand is of the opinion that disease and variety access depends on the week of the month. Check the validity of Anand’s belief using an appropriate hypothesis test.**

Part A: Week of the month and disease access are independent or not

H0 : Disease access and week of the month are independent.

H1 : Disease access and week of the month are dependent.

Frequencies as per the given data:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 |  |
| Week1 | 653 | 480 | 639 | 490 | 428 | 1044 | 607 | 641 | 573 | 493 | 461 | 6509 |
| Week2 | 598 | 444 | 607 | 450 | 392 | 940 | 601 | 543 | 531 | 486 | 385 | 5977 |
| Week3 | 516 | 390 | 587 | 407 | 342 | 878 | 530 | 476 | 528 | 414 | 373 | 5441 |
| Week4 | 892 | 638 | 931 | 685 | 589 | 1433 | 848 | 800 | 807 | 643 | 637 | 8903 |
|  | 2659 | 1952 | 2764 | 2032 | 1751 | 4295 | 2586 | 2460 | 2439 | 2036 | 1856 | 26830 |

Expected frequencies if week and disease access are independent.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 |
| Week1 | 645 | 474 | 671 | 493 | 425 | 1042 | 627 | 597 | 592 | 494 | 450 |
| Week2 | 592 | 435 | 616 | 453 | 390 | 957 | 576 | 548 | 543 | 454 | 413 |
| Week3 | 539 | 396 | 561 | 412 | 355 | 871 | 524 | 499 | 495 | 413 | 376 |
| Week4 | 882 | 648 | 917 | 674 | 581 | 1425 | 858 | 816 | 809 | 676 | 616 |

Calculating the (e-e0)2/e0

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Row Labels | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 |
| Week1 | 0.1 | 0.09 | 1.48 | 0.02 | 0.02 | 0 | 0.66 | 3.27 | 0.59 | 0 | 0.26 |
| Week2 | 0.05 | 0.19 | 0.12 | 0.02 | 0.01 | 0.3 | 1.08 | 0.05 | 0.28 | 2.32 | 1.96 |
| Week3 | 1 | 0.09 | 1.25 | 0.06 | 0.48 | 0.06 | 0.06 | 1.05 | 2.25 | 0 | 0.03 |
| Week4 | 0.11 | 0.15 | 0.21 | 0.17 | 0.11 | 0.04 | 0.12 | 0.33 | 0.01 | 1.57 | 0.72 |

X^2 22.7393298

df 30

X2 critical 43.77297183

p-val 0.825822415

we retain H0, i.e we conclude, there is no dependency between disease access and week.

Part B: Week of the month and variety access are independent or not

H0 : Variety access and week of the month are independent.

H1 : Variety access and week of the month are dependent.

Frequencies as per the given data.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |  |
| Week1 | 1230 | 1029 | 755 | 658 | 710 | 799 | 686 | 517 | 639 | 775 | 7798 |
| Week2 | 1117 | 939 | 661 | 576 | 614 | 700 | 636 | 500 | 606 | 699 | 7048 |
| Week3 | 1059 | 909 | 643 | 574 | 624 | 633 | 558 | 453 | 577 | 694 | 6724 |
| Week4 | 1966 | 1508 | 1073 | 917 | 1129 | 1148 | 1048 | 802 | 1010 | 1171 | 11772 |
|  | 5372 | 4385 | 3132 | 2725 | 3077 | 3280 | 2928 | 2272 | 2832 | 3339 | 33342 |

Expected frequencies, if Week and Variety access are independent.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
| Week1 | 1256 | 1026 | 733 | 637 | 720 | 767 | 685 | 531 | 662 | 781 |
| Week2 | 1136 | 927 | 662 | 576 | 650 | 693 | 619 | 480 | 599 | 706 |
| Week3 | 1083 | 884 | 632 | 550 | 621 | 661 | 590 | 458 | 571 | 673 |
| Week4 | 1897 | 1548 | 1106 | 962 | 1086 | 1158 | 1034 | 802 | 1000 | 1179 |

Calculating the (e-e0)2/e0

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
| Week1 | 0.55 | 0.01 | 0.69 | 0.67 | 0.13 | 1.32 | 0 | 0.39 | 0.82 | 0.04 |
| Week2 | 0.3 | 0.16 | 0 | 0 | 2.04 | 0.06 | 0.47 | 0.81 | 0.09 | 0.07 |
| Week3 | 0.55 | 0.69 | 0.2 | 1.09 | 0.02 | 1.23 | 1.79 | 0.06 | 0.06 | 0.63 |
| Week4 | 2.53 | 1.04 | 0.97 | 2.12 | 1.67 | 0.09 | 0.2 | 0 | 0.1 | 0.05 |

X^2 23.73310631

df 27

X2 critical 40.11327207

p-val 0.64507612

Retain the null. There is a no dependency between variety access and week