SA2 Tutorials I & II

Manjusha Kancharla

3/26/2021

Agenda

1. Comparison of Two Populations
2. Analysis of Variance
3. Chi-square test
4. Simple Linear regression

Comparison of Two Populations

1. Inpedendent samples
   - Tests for means
   - Tests for proportions
2. Paired observations

Practice Problem 1

A nationwide retailer wants to test whether new product shelf facings are effective in increasing sales volume. New shelf facings for the soft drink Country Time are tested at a random sample of 15 stores throughout the country. Data on total sales of Country Time for each store, for the week before and the week after the new facings are installed, are given below:

- Store : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
- Before: 57 61 12 38 12 69 5 39 88 9 92 26 14 70 22
- After : 60 54 20 35 21 70 1 65 79 10 90 32 19 77 29

Using the 0.05 level of significance, do you believe that the new shelf facings increase sales of Country Time?

Practice Problem-1: Solution

What test is appropriate here?

Practice Problem-1: Solution

- Type of test: paired t-test (samples are dependent)

```r
Before <- c(57, 61, 12, 38, 12, 69, 5, 39, 88, 9, 92,
            26, 14, 70, 22)
After <- c(60, 54, 20, 35, 21, 70, 1, 65, 79, 10, 90,
           32, 19, 77, 29)
Diff <- Before-After
t.test(Diff)
```

```

    One Sample t-test

data:  Diff
t = -1.4691, df = 14, p-value = 0.1639
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -7.871876  1.471876
sample estimates:
mean of x
```

Practice Problem-1: Alternative Solution

```r
t.test(Before,After, paired = TRUE)
```

```

	Paired t-test

data:  Before and After
t = -1.4691, df = 14, p-value = 0.1639
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -7.871876  1.471876
sample estimates:
mean of the differences
                   -3.2
```

Practice Problem 2

A manufacturer of modems uses microcomputer chips from two different sources. As part of quality-control testing, the manufacturer obtains data on the rate of defective chips per thousand for each lot of chips. Study the results given below:

	Source I	Source II
Mean	13.43	15.21
Variance	32.92	38.12
Observations	10	10

What is the probability of type I error the manufacturer will commit, if he concludes that the rate of defective chips from both these lots is unequal?

Practice Problem 2: Solution

What test is appropriate here?

Practice Problem 2: Solution

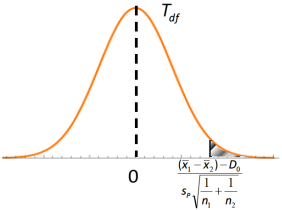
If we believe $\sigma_1 = \sigma_2$

- Calculate the "pooled" sample standard deviation

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$$

and degrees of freedom

$$df = n_1 + n_2 - 2$$



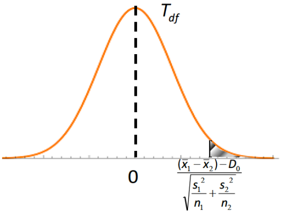
If we believe $\sigma_1 \neq \sigma_2$

- Calculate the standard error

$$se(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and degrees of freedom

$$df = \left\lfloor \frac{\left(s_1^{\,2}/n_1 + s_2^{\,2}/n_2\right)^2}{\left(s_1^{\,2}/n_1\right)^2/(n_1 - 1) + \left(s_2^{\,2}/n_2\right)^2/(n_2 - 1)} \right\rfloor$$



Practice Problem 2: Solution

- Let μ_1 denote the average number of defective chips in a lot of thousand, from source 1
- Let μ_2 denote the average number of defective chips in a lot of thousand, from source 2

Our Null and Alternative hypothesis are as follows:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$$

$$or,$$

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

Practice Problem 2: Solution

- Pooled Standard Deviation, S_p

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10 - 1)\, 32.92 + (10 - 1)\, 38.12}{10 + 10 - 2}}$$

$$= 5.96$$

The degrees of freedom are $df = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$

Practice Problem 2: Solution

Our test statistic is,

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{13.43 - 15.21}{5.96\sqrt{\frac{1}{10} + \frac{1}{10}}} = -0.67$$

The p-value for this test is,

```r
2*pt(q = -0.67, df = 18)
```

```
[1] 0.5113624
```

Practice Problem 2: Solution

We know that the Probability of type 1 error is equal to the p-value of our test statistic!

There is a 51% chance that the manufacturer will be committing to type I error by concluding that the defective chips from these 2 lots are unequal.

Practice Problem-3

In the theory of finance, a market for any asset or commodity is said to be efficient if items of identical quality and other attributes (such as risk, in the case of stocks) are sold at the same price. **A Geneva-based oil industry analyst wants to test the hypothesis that the spot market for crude oil is efficient**. The analyst chooses the Rotterdam oil market, and he selects Arabian Light as the type of oil to be studied.(Differences in location may cause price differences because of transportation costs, and differences in the type of oil—hence, in the quality of oil—also affect the price. Therefore, both the type and the location must be fixed.) A random sample of eight observations from each of four sources of the spot price of a barrel of oil during February 2007 is collected. Data, in U.S. dollars per barrel, are as follows:

Practice Problem-3: Data

U.K.	Mexico	U.A.E.	Oman
62.10	56.30	55.60	53.11
63.20	59.45	54.22	52.90
55.80	60.02	53.18	53.75
56.90	60.00	56.12	54.10
61.20	58.75	60.01	59.03
60.18	59.13	53.20	52.35
60.90	53.30	54.00	52.80
61.12	60.17	55.19	54.95

Conduct an appropriate test to help the Geneva-based oil industry analyst wants to test the hypothesis that the spot market for crude oil is efficient.

Practice Problem-3: Solution

- Need to input data in a particular format to use easily in R

```
   Location Price
1        UK 62.10
2        UK 63.20
3        UK 55.80
4        UK 56.90
5        UK 61.20
6        UK 60.18
7        UK 60.90
8        UK 61.12
9    Mexico 56.30
10   Mexico 59.45
```

Practice Problem-5: Solution

```r
fit <- aov(Price ~ factor(Location), data=q3)
summary(fit)
```

```
                 Df Sum Sq Mean Sq F value   Pr(>F)    
factor(Location)  3  188.8   62.94   11.55 4.19e-05 ***
Residuals        28  152.6    5.45                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Practice Problem-3: Solution

We perform the “Tukey’s test for pairwise comparison”:

```
TukeyHSD(fit)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Price ~ factor(Location), data = q3)

$`factor(Location)`
                diff       lwr         upr     p adj
Oman-Mexico -4.26625 -7.452753 -1.07974693 0.0054664
UAE-Mexico  -3.21250 -6.399003 -0.02599693 0.0476033
UK-Mexico    1.78500 -1.401503  4.97150307 0.4340886
UAE-Oman     1.05375 -2.132753  4.24025307 0.8033837
UK-Oman      6.05125  2.864747  9.23775307 0.0000942
UK-UAE       4.99750  1.810997  8.18400307 0.0010676
```

Practice Problem-3: Solution

Therefore we conclude at 5% level of significance, that:

- U.A.E has significantly lower average crude oil price than U.K.
- OMAN has significantly lower average crude oil price than U.K.
- U.A.E has significantly lower average crude oil price than MEXICO.
- OMAN has significantly lower average crude oil price than MEXICO.

Practice Problem - 4

A study was conducted to determine whether a relationship existed between certain shareholder characteristics and the level of risk associated with the shareholders' investment portfolios. As part of the analysis, portfolio risk (measured by the portfolio beta) was divided into three categories: low-risk, medium-risk, and high-risk; and the portfolios were cross-tabulated according to the three risk levels and seven family income levels. The results of the analysis, conducted using a random sample of 318 investors, are shown in the following contingency table. Test for the existence of a relationship between income and investment risk taking.

Income Level ($)	Low	Medium	High	Total
0-60000	15	14	9	38
61000-100000	16	13	10	39
101000-150000	22	30	11	63
151000-200000	11	20	20	51
201000-250000	18	10	14	42
251000-300000	12	10	10	32
301000 +	11	11	31	53
Total	105	108	105	318

Practice Problem - 4

- The Chi-square test of independence determines whether there is a statistically significant relationship between categorical variables.
- H_0: the variables are independent, there is no relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable
- H_1: the variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable

Practice Problem - 4 Solution

- First create a table of 'Expected Counts"
- A cell (i, j) is replaced by **(sum of row i * sum of column j) / Total sum**

Income Level	Low	Medium	High	Total
0-60000	12.5472	12.9057	12.5472	38
61000-100000	12.8774	13.2453	12.8774	39
101000-150000	20.8019	21.3962	20.8019	63
151000-200000	16.8396	17.3208	16.8396	51
201000-250000	13.8679	14.2642	13.8679	42
251000-300000	10.5660	10.8679	10.5660	32
301000 $+$	17.5000	18.0000	17.5000	53
Total	105	108	105	318

Practice Problem - 4 Solution

Compute the test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}$$

Practice Problem - 4 Solution

- First calculate $\frac{(Obs-Exp)^2}{Exp}$ for each cell:

Income Level	Low	Medium	High
0-60000	0.47950064	0.09279488	1.00280891
61000-100000	0.75721197	0.00454228	0.64292626
101000-150000	0.06900697	3.45971848	4.61866684
151000-200000	2.02505682	0.41443663	0.59312404
201000-250000	1.23118983	1.27473295	0.00125786
251000-300000	0.19460916	0.06931342	0.03032345
301000 +	2.41428571	2.72222222	10.4142857

- Sum up all these values to get $\chi^2 = \sum_{\text{all cells}} \frac{(Obs-Exp)^2}{Exp} = 32.51$

Practice Problem - 4 Solution

- We have $\chi^2 = \sum_{\text{all cells}} \frac{(Obs-Exp)^2}{Exp} = 32.51$
- Under the null hypothesis this follows a χ^2-distribution with $(r-1)(c-1) = 6 * 2 = 12$ degrees of freedom. Where $r = 7$ is the number of rows and $c = 3$ is the number of columns.
- Next we find the corresponding p-value, $P(\chi^2_{12} > 32.51)$:

```
pchisq(q = 32.51, df = 12, lower.tail = FALSE)
```

```
[1] 0.001153828
```

Practice Problem - 4 In R

- We will use the well-known iris dataset but slightly enhanced.

```
data <- iris
head(data)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

- Since there is only one categorical variable (Species) and the Chi-square test requires two categorical variables, we add the variable size which corresponds to small if the length of the petal is smaller than the median of all flowers, big otherwise

```
data$Size <- ifelse(data$Sepal.Length <
                      median(data$Sepal.Length), "small", "big")
```

Practice Problem - 4 In R

```r
with(data, table(Species, Size))
```

```
            Size
Species      big small
  setosa       1    49
  versicolor  29    21
  virginica   47     3
```

```r
with(data,chisq.test(table(Species, Size)))
```

```

    Pearson's Chi-squared test

data:  table(Species, Size)
X-squared = 86.035, df = 2, p-value < 2.2e-16
```

Practice Problem - 5

The operations manager of a company that manufactures tires wants to determine whether there are any differences in the quality of work among the three daily shifts. She randomly selects 496 tires and carefully inspects them. Each tire is either classified as perfect, satisfactory, or defective, and the shift that produced it is also recorded. The two categorical variables of interest are shift and condition of the tire produced. The data is summarized below. Does the data provide sufficient evidence at the 5% significance level to infer that there are differences in quality among the three shifts?

	Perfect	Satisfactory	Defective	Total
Shift 1	106	124	1	231
Shift 2	67	85	1	153
Shift 3	37	72	3	112
Total	210	281	5	496

Practice Problem - 5

Even if we did have a significant result for this question, we still can not rely ob the results, because there are 3 (33.3% of) cells are with expected counts < 5.0 (Check on your own!).