Ch-6 Multivariate Statistics

Two sample T-test for Difference in means

$$t_{data} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

$df = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$

df will be minimum of

Q Page - 189

ie $n_1 - 1 = 2529 - 1$

$n_2 - 1 = 804 - 1 = 803$

S-1 State hypotheses

$H_0 = \mu_1 = \mu_2$ (no-differences)

$H_1 : \mu_1 \neq \mu_2$ (means are different)

This is two-tail test.

S-4 conclusion

$P = 0.508 > 0.05$

do not reject the null hypothesis.

S-2 compute test statistics

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} = \frac{1.5714 - 1.5361}{\sqrt{\frac{(1.3126)^2}{2529} + \frac{(1.3251)^2}{804}}}$$

$= 0.6595$

S-3 P-value (use z-table because sample is so equal to t-test equal to z-test)

P-value $= 2 \cdot P(t > t_{data}) = P(t > 0.6595)$

$2 \times 0.257 = 0.514$

$= 2 \times (0.254) = 0.508$ 0.7421

$1 - 0.7421 = 0.257$

from z-table row 0.6 & column 0.06 = 0.6 + 0.06 = 0.66

value is 0.7454, from right tail $= 1 - 0.7454 = 0.2546$)

Two Sample Z-Test for Difference in Proportions

$$Z_{data} = \frac{P_1 - P_2}{\sqrt{P_{pooled} \cdot (1 - P_{pooled})\left(\left(\frac{1}{n_1}\right) + (1/n_2)\right)}}$$

where $P_{pooled} = \dfrac{x_1 + x_2}{n_1 + n_2}$

Ex

$x_1 = 707$ of $n_1 = 2529$ customers in the training set belonging to the voicemail plan while $x_2 = 215$ of $n_2 = 804$ customers in the tests set.

So $P_1 = \dfrac{x_1}{n_1} = \dfrac{707}{2529} = 0.2796$

$P_2 = \dfrac{x_2}{n_2} = \dfrac{215}{804} = 0.2674$

$P_{pooled} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{707 + 215}{2529 + 804} = 0.2766$

S-1 Hypothesis

$H_0 : \pi_1 = \pi_2$ Vs $H_1 : \pi_1 \neq \pi_2$

S-2

$$Z_{data} = \frac{P_1 - P_2}{\sqrt{P_{pooled} \cdot (1 - P_{pooled})\left(\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)\right)}}$$

$$= \frac{0.2796 - 0.2674}{\sqrt{0.2766 \cdot (0.7234)\left((1/2529) + \left(\frac{1}{804}\right)\right)}}$$

$$= 0.6736$$

S-3 p-value is

$$\text{p-value} = 2 \cdot P(Z > 0.6736)$$

$$= 0.5006$$

from Z-table
0.6 (row) & 0.07 (column)
0.7485

$1 - 0.7485 = 0.2515$

$= 0.251 \times 2 = 0.502$

S-4 conclusion

$$\text{p-value} = 0.5006 > 0.05$$

so do not reject H_0.

TEST FOR HOMOGENEITY OF PROPORTIONS (chi-square distribution)

observed frequency (Marital status)

Dataset	Married	Single	other	Total
Training set	410	340	250	1000
Test set	95	85	70	250
Total	505	425	320	1250

S-1 Hypothesis set

H_0 : $P_{married,\ training} = P_{married,\ test}$

$P_{single,\ training} = P_{single,\ test}$

$P_{other,\ training} = P_{other,\ test}$

H_a : At least one of the claims in H_0 is wrong.

S-2 Expected frequency

$$\text{married, training} = \frac{(1000)(\ldots)}{1250} = 4\ldots$$

$$\boxed{\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}}$$

Expected frequencies

Dataset	Married	Single	Other	Total
Training set	404	340	256	1000
Test set	101	85	64	250
Total	505	425	320	1250

S-3 calculate test statistic χ^2_{data}

cell	Observed frq.	Expected frq.	$\frac{(Obs - Exp)^2}{Exp}$
Married, training	410	404	$\frac{(410-404)^2}{404} = 0.09$
Married, test	95	101	0.36
Single, training	340	340	0
Single, test	85	85	0
Other, training	250	256	0.14
Other, test	70	64	0.56

$$\text{sum} = \chi^2_{data} = 1.15$$

p-value $= P(\chi^2 > \chi^2_{data}) = P(\chi^2 > 1.15) = 0.5627$

① test of independent : $df = (r-1)(c-1) = 2$

ie $df = (\text{no. of rows} - 1)(\text{no. of column} - 1)$

$= (2-1)(3-1) = 1 \times 2 = 2$

~~where p-value $= 2 \chi 0.5627$~~

② goodness of fit

$df = k-1$ (k = no. of categories)

$= 6-1$

S-5

Conclusion

p-value $< \alpha$ then reject Ho

but then $0.5627 \not< 0.05$ then observed frequencies represent proportions that are significantly different for the training & test data.


Lenovo
Pause
Insert
Backspace
Home
End
NmLk
/
*
-
P
Delete
PgUp
PgDn
7
Home
8
9
+
Enter
4
5
6
Shift
1
End
2
3
PgDn
Enter
0
Ins