ANSWERS

Relationships Pen and Paper Questions

(1)

A suitable test is Chi-squared. The formula for calculating χ^2 is:

$$\chi^2 = \sum_{i=0}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where k is the number of cells in the contingency table, O_i is the ith observed value (the number in the ith cell) and E_i is the ith expected value. The expected value for a cell is calculated as:

$$E_i = \frac{n_r(i)n_c(i)}{n}$$

where n is the number of observations (the sum of the numbers in all the cells of the table, $n_r(i)$ is the number of observations in the row to which the ith cell belongs (the sum of numbers in that row) and $n_c(i)$ is the number of observations in the cell to which the ith cell belongs (the sum of numbers in that column).

The following table shows the observed values, 0_i, for all the cells in the table:

	Conservative	**Liberal Democrat**	**Labour**
Male	313	124	391
Female	344	158	388

The sum of observations in the `Male` row is:

$$n_r^{Male} = 313 + 124 + 391 = 828$$

The sum of observations in the `Female` row is:

$$n_r^{Female} = 344 + 158 + 388 = 890$$

The sum of observations in the `Conservative` column is:

$$n_c^{Conservative} = 313 + 344 = 657$$

The sum of observations in the `Liberal Democrat` column is:

$$n_c^{Liberaldemocrat} = 124 + 158 = 282$$

The sum of observations in the `Labour` column is:

$$n_c^{Labour} = 391 + 388 = 779$$

The overall number of observations is:

$$n = 313 + 124 + 391 + 344 + 158 + 388 = 1718$$

The following table shows the expected values, E_i, for all the cells in the table:

	Conservative	**Liberal Democrat**	**Labour**
Male	$\frac{828 \times 657}{1718} = 316.64$	$\frac{828 \times 282}{1718} = 135.9$	$\frac{828 \times 779}{1718} = 375.44$
Female	$\frac{890 \times 657}{1718} = 340.36$	$\frac{890 \times 282}{1718} = 146.09$	$\frac{890 \times 779}{1718} = 403.56$

The following table shows the values of $\frac{(O_i - E_i)^2}{E_i}$ for each cell in the table:

	Conservative	**Liberal Democrat**	**Labour**
Male	0.04	1.04	0.64
Female	0.04	0.97	0.60

We add up the values from the table above to get the value of χ^2:

$$\chi^2 = \sum_{i=0}^{k} \frac{(O_i - E_i)^2}{E_i} = 0.04 + 1.04 + 0.64 + 0.04 + 0.97 + 0.60 = 3.33$$

We calculate the degrees of freedom for the table (n_r is the number of rows, n_c the number of columns):

$df = (n_r - 1)(n_c - 1) = (2 - 1)(3 - 1) = 2$

Now we look up the row for $df = 2$ in the χ^2 percentage points table. We see that the calculated value does not fall in the upper 10% of the χ^2 distribution, as it is smaller than the value shown in the table for 10% (circled in red). This means that at even at a significance level of 10%, we can conclude that the voting patterns of males and females do not differ.

ν	$p(\%)$								
	Lower tail			*Upper tail*					
	0.5	2.5	5	90	95	97.5	99	99.5	99.9
1	$0.0^{4}3927$	$0.0^{3}9821$	$0.0^{2}3932$	2.706	3.841	5.024	6.635	7.879	10.83
2	0.01003	0.05064	0.1026	4.605	5.991	7.378	9.210	10.60	13.82
3	0.07172	0.2158	0.3518	6.251	7.815	9.348	11.34	12.84	16.27
4	0.2070	0.4844	0.7107	7.779	9.488	11.14	13.28	14.86	18.47

(2)

The appropriate test in the given scenario is ANOVA.

We need to use the F statistic, with formula:

$$F = \frac{MSB}{MSW} = \frac{(\sum_{g\in\{1,2,3\}} n_g(\bar{x_g} - \bar{\bar{x}})^2)/(k-1)}{(\sum_{g\in\{1,2,3\}} \sum_{i,x_i\in g} (x_i - \bar{x_g})^2)/(n-k)}$$

Overall mean and group means:

$\bar{\bar{x}} = (21 + 21 + 21 + 21 + 21 + 24 + 22 + 22 + 25)/9 = 22$

$\bar{x_A} = (21 + 21 + 21)/3 = 21 \qquad \bar{x_B} = (21 + 21 + 24)/3 = 22 \qquad \bar{x_C} = (22 + 22 + 25)/3 = 23$

$n = 9 \quad k = 3$

Calculate F-statistic:

$F = \frac{[3\times(21-22)^2+3\times(22-22)^2+3\times(23-22)^2]/(3-1)}{[(21-21)^2+(21-21)^2+(21-21)^2+(21-22)^2+(21-22)^2+(24-22)^2+(22-23)^2+(22-23)^2+(25-23)^2]/(9-3)} = 1.5$

Degrees of freedom:

$df_B = k - 1 = 2 \quad df_W = n - k = 6$

Lookup in F-statistic table for degrees of freedom above and significance level .05:

Critical value: 5.14

Comparison of statistic against critical value:

$1.5 < 5.14 \quad ? \quad TRUE$

Interpretation of comparison:

We fail to reject the null hypothesis that there is no difference between the groups.

Subject domain interpretation:

There is no statistical evidence that the students grouped by age.

(3)

A parametric statistic of relatedness of two numeric variables is Pearson's correlation coefficient, for which we use the the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

The following table is an extract from a spreadsheet used to speed up the calculations:

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
	65	60	9	-5	81	25	-45
	45	60	-11	-5	121	25	55
	40	55	-16	-10	256	100	160
	55	70	-1	5	1	25	-5
	60	80	4	15	16	225	60
	50	40	-6	-25	36	625	150
	80	85	24	20	576	400	480
	30	50	-26	-15	676	225	390
	70	70	14	5	196	25	70
	65	80	9	15	81	225	135
mean: $\bar{x}$, $\bar{y}$	56	65					
variance: $\Sigma(x-\bar{x})^2/(n-1)$, $\Sigma(y-\bar{y})^2/(n-1)$					226.67	211.11	
standard deviation: $\sqrt{\text{variance}}$					15.06	14.53	
sum							1450.00

The column headers indicate what they represent, with the x and y columns containing raw data and the other columns calculated using x and y as indicated. The mean, variance and standard deviation rows show the values of those measures for x and y. The sum row contains the sum of the values in the column above.

Now can use the values from the spreadsheet to calculate Pearson's correlation coefficient:

$$r = \frac{1450}{(10-1) \times 15.06 \times 14.53} = 0.74$$

Alternatively, the simpler formula for r could be used.

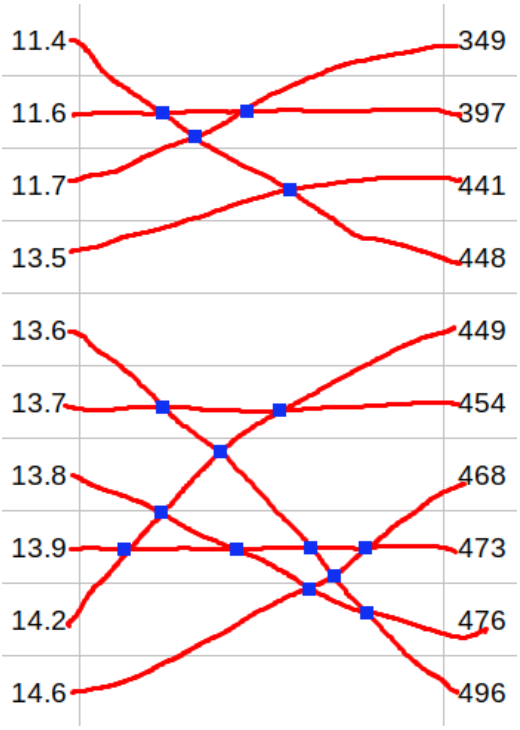
Looking up the table of critical values for r, in particular the values for $n = 10$ (the number of values in the sample), we see that the calculated correlation coefficient is greater than the critical value for 5% (.632) but not greater than the critical value for 1% (.765). This means that the correlation is significant at the 5% level but not significant at the 1% level.

n	5%	1%	n	5%	1%	n	5%	1%	n	5%	1%
4	.950	.990	7	.754	.874	10	.632	.765	13	.553	.684
5	.878	.959	8	.707	.834	11	.602	.735	14	.532	.661
6	.811	.917	9	.666	.798	12	.576	.708	15	.514	.641

(4)

Because the data are not normally distributed, the test needs to be non-parametric, such as Kendall's Tau. We pick the significance level of 0.01.

To calculate the Kendall τ, or Kendall Rank Correlation Coefficient, we order the data of each variable by rank in a column, positioning the columns side by side. Then we connect each pair of values that belong to the same data instance with an unbroken line (red in the picture) and count the number of intersection points resulting from the lines crossing each other (blue squares in the picture).



In our example the number of points of intersection is 15 and this equates to the number of discordant instance pairs, i.e. $n_d = 15$.

The number of instance pairs in the set is $n_p = \frac{n(n-1)}{2} = \frac{10 \times 9}{2} = 45$. As there are no duplicate values for either of the variables, there are no 'ties' and we can conclude that any pair that is not discordant is concordant and can calculate the number of concordant pairs as $n_c = n_p - n_d = 45 - 15 = 30$.

The Kendall τ in its simplest version (used if there are no ties) is calculated as:

$$\tau_A = \frac{n_c - n_d}{n_p} = \frac{30 - 15}{45} = \frac{1}{3}$$

To find out if this correlation has any significance and assuming that a negative correlation is not possible, we look up the table of (one-tailed) critical values for τ, for the case of 10 instances. In the table we find that at the required level of 1% the calculated correlation is not significant, as it is lower than the critical value of .600. (It also happens to be lower than the critical value for the significance level of 5%, which is .467.)

n	5%	1%	n	5%	1%	n	5%	1%	n	5%	1%
4	1.000	*	7	.619	.810	10	.467	.600	13	.359	.513
5	.800	1.000	8	.571	.714	11	.418	.564	14	.363	.473
6	.733	.867	9	.500	.667	12	.394	.545	15	.333	.467