Research Methods for Social Sciences:

Comparison between groups

Alex Teytelboym & Earth Sivakul

@t8el

June 11, 2013

Outline

1. Association between two variables
   - Comparisons of two groups
   - Testing association between two variables
   - Differences in means and confidence intervals
   - ANOVA
   - χ^2 test

Association between two variables

- Association between two variables: if the distribution of the response variable changes in some way as the value of the explanatory variable changes, we have evidence of association between the variables – even if the direction of the association is not proven.
- When there is lack of association between the variables, they are said to be independent of each other.
- Different types of variables – numerical or categorical – will require distinct methods.
- Testing whether two variables are associated or not. If one variable is categorical: testing this implies testing whether the mean values of the variable of interest for the different groups are equal (independent) or different (dependent).

Comparisons between two groups

- Two groups comparison constitute a binary or dichtomous variable: a variable having only two categories.
- In comparisons of two groups, two variables are involved (hence *bivariate* statistical methods):
  - X: explanatory (or *independent*) var $\rightarrow$ defines division by groups
  - Y: response (or *dependent*) var $\rightarrow$ on which comparisons are made.

Comparisons between two groups

Example

A comparison of Democrats and Republicans on the proportion who favor the new national health insurance policy

- A bivariate analysis
  - Response variable: opinion about national health insurance;
  - Explanatory variable: political party affiliation

Comparisons between groups

Table: Tests for comparison of between groups

		Number of groups	
		$=2$	>2
Response variable (Y)	Numerical	Difference between means: • large sample: Z test • small sample: t-test (assume $\sigma_1 = \sigma_2$)	ANOVA
	Categorical	Difference between proportions: Z-test	χ^2 test

Hypothesis testing

- Assumptions
  - about the distribution of the underlying population
  - about the variances/standard deviation in the population
  - about the sampling distribution
  - about the form of relation between variables
- Hypothesis
  - H_0 : null hypothesis - equality
  - H_A: alternative hypothesis - inequality
- Statistic
  - is the instrument to be accessed by the hypothesis. In parametric tests, the statistic follows a particular distribution (normal, t, χ^2, F, binomial, etc.). Once the data is collected, we compute the value of the statistic for the sample.

Hypothesis testing

- *P*-value
  - probability, *if the null hypothesis were true*, of obtaining a value as the one observed. The smaller the *P*-value the stronger the evidence **against** the null hypothesis.
- Conclusion-interpretation
  - Comparing with the preset level considered to be sufficient, one can either reject the null hypothesis or cannot reject it (i.e. either the evidence is consistent with the null or not).

Differences in means

Example

Do women tend to spend more time on housework than men?

- The following table reports the descriptive statistics based on data from the National Survey of Families and households in the US.

Table: Gender difference in house work hours

		House work hours (weekly)	
Gender	Sample size	Mean	Standard Deviation
Male	4252	18.1	12.9
Female	6764	32.6	18.2

Differences in means

Example

Do women tend to spend more time on housework than men?

Response: Numerical

Explanatory: Categorical= 2

- Assumptions
  - samples are random, independent values
  - both groups are large (>30) or population standard deviations are known ⇒sampling distribution of their difference is of normal distribution (by the central limit theorem)
- Alternative assumptions
  - errors are distributed normally in the population in each group and the variance of the two groups in the same ⇒ t-distribution

Differences in means

Example

Do women tend to spend more time on housework than men?

- Hypothesis
  - $H_0: \mu_2 = \mu_1$ and $H_A: \mu_2 \neq \mu_1$
  - Or we could write them as: $H_0: \mu_2 - \mu_1 = \mu = 0$ and $H_A: \mu_2 - \mu_1 = \mu \neq 0$

Differences in means

Example

Do women tend to spend more time on housework than men?

- Statistic (large sample):

$$z = \frac{(\overline{Y}_2 - \overline{Y}_1) - (\mu_2 - \mu_1)}{SE(\overline{Y}_2 - \overline{Y}_1)} = \frac{\overline{w} - \mu_w}{SE(\overline{w})} \sim N(0,1)$$

where

$$SE(\overline{Y}_2 - \overline{Y}_1) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

under the null hypothesis $\mu_2 - \mu_1 = 0$ in our example.

Differences in means

Example

Do women tend to spend more time on housework than men?

- Statistic (small sample):

$$t = \frac{(\overline{Y}_2 - \overline{Y}_1) - (\mu_2 - \mu_1)}{SE(\overline{Y}_2 - \overline{Y}_1)} = \frac{\overline{w} - \mu_w}{SE(\overline{w})} \sim t_{df:\, n_1+n_2-2}$$

where

$$SE(\overline{Y}_2 - \overline{Y}_1) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

under the null hypothesis $\mu_2 - \mu_1 = 0$ in our example.

Differences in means

Example

Do women tend to spend more time on housework than men?

- *P*-value: two-sided (multiply *P*-value by 2) or one-sided depending on the alternative hypothesis
  - Two-sided: $\mu_1 \neq \mu_2$ (women do not do the same number of hours of house work as men)
  - One-sided: $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ (women do fewer/more hours of house work than men)
  - We usually use the two-sided alternative hypothesis

Differences in means

Example

Do women tend to spend more time on housework than men?

- Conclusion:
  - If the P-value is low, then there is strong evidence against the null of equality (reject the null hypothesis) $\Rightarrow$ the evidence supports the idea that means are different in the population and that there is some relationship between Y and the X variable defining the groups.
  - If P-value is large, then the evidence is consistent with the null of equality of means for both groups; rather, it is likely that the means of the two groups are similar. (do not reject the null hypothesis)

Differences in means

Example

Do women tend to spend more time on housework than men?

- Standard error of the difference

$$SE_{\overline{Y}_2-\overline{Y}_1} == \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{12.9^2}{4252}+\frac{18.2^2}{6764}} = 0.30$$

- Test statistic

$$z = \frac{\overline{Y}_2 - \overline{Y}_1}{SE_{\overline{Y}_2-\overline{Y}_1}} = \frac{32.6 - 18.1}{0.30} = 48.8$$

- P-value is essentially 0. We can conclude that the population means differ. The sample means show that the difference takes the direction of a higher mean for women.

Confidence intervals

Example

Do women tend to spend more time on housework than men?

- Confidence interval (large sample)

$$\mu_2 - \mu_1 = (\overline{Y}_2 - \overline{Y}_1) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- If the confidence interval does not contain 0, then mean of group 2 is different of the mean of group 1.
- If the confidence interval contains 0, then we do not have enough evidence to conclude which mean is larger.
- Sample size requirement: each sample is at least 30, or known standard deviations.

Confidence intervals

Example

Do women tend to spend more time on housework than men?

- Confidence interval (small sample)

$$\mu_2 - \mu_1 = \overline{Y}_2 - \overline{Y}_1 \pm t_{\alpha/2;df} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- If the confidence interval does not contain 0, then mean of group 2 is different of the mean of group 1.
- If the confidence interval contains 0, then we do not have enough evidence to conclude which mean is larger.
- No sample size requirement.

Confidence intervals

Example

Do women tend to spend more time on housework than men?

- A 99% confidence interval

$$(\overline{Y}_2 - \overline{Y}_1) \pm 2.58 \times SE_{\overline{Y}_2 - \overline{Y}_1} = 14.5 \pm 2.58 \times 0.30 = 14.5 \pm 0.8$$

- Or (13.7, 15.3)
- We infer that the true difference falls between 13.7 and 15.3 hours of house work.
- As the confidence interval contains only positive values of the difference, we are 99% confident that the hours that women spent on housework is higher than that by men.

Confidence intervals

Notes

- Violation of assumptions: if population standard deviation are different (Levene's test) and the sample size are less than 30. Alternatives: first, seek a suitable change of scale with makes the standard deviations similar; second, use other methods to test, such as non-parametric tests 1. Wilcoxon rank-sum test (based on the sum of ranks of the values in each of the two groups) 2. Mann-Whitney U-test (identical results as Wilcoxon, but more complicated) or bootstrap for confidence intervals/Monte Carlo simulations.
- Paired sample: two samples are related to each other and one numerical variable - e.g. blood pressure of patients in two circumstances (before and after treatment) $\Rightarrow$ similar procedure, change s.

Differences in means

Example

Exploring the relation between education and type of place of residence

- Ghana 2003, DHS. Women between 15 – 65 years old.
- Education: quantitative, continuous variable (in years)
- Place of residence: qualitative, categorical variable, 1:urban; 0: rural

Differences in means

Example

Exploring the relation between education and type of place of residence

Table: Residence and education in Ghana

		Education in single years	
Residence	Sample size	Mean	Standard Deviation
Urban	2364	7.52	4.468
Rural	3306	3.74	4.246

Differences in means

Example

Exploring the relation between education and type of place of residence

- Null hypothesis: mean levels of education urban and rural Ghanian women are equal. Alternative hypothesis: they are not.
- $z = 32.06$
- P-value(32.06)$= 0.000 \Rightarrow$very strong evidence against the null hypothesis

Differences in means

Example

Exploring the relation between education and type of place of residence

- Confidence Interval: $(3.55, 4.01)$
- As the confidence interval contains only positive values of the difference in education, we are 95% confident that the level of education of women in urban areas is higher than that of the rural areas. In other words, education level is **not** independent of the place of residence.

ANOVA

- We use ANOVA to compare several groups
- Response (Y) variable is numerical and explanatory variable (X): categorical ≥ 2
- ANOVA – 'analysis of variance' -
  - significance test, using F-distribution for detecting evidence of differences among population means.
- Logic: Are the differences between the groups large enough to reject the null hypothesis and justify the conclusion that the populations represented by the groups are different?

ANOVA

Example

Political ideology by party identification

- Response variable: political ideology, measured on 1-7 scale, from extreme liberal to extreme conservative
- Explanatory variable: party identification, 3 groups: Democrat, Independent, Republican

ANOVA

Example

Political ideology by party identification

Table: Party affiliation and ideology in the US

Group	Political ideology							Sample		
	1	2	3	4	5	6	7	Size	Mean	Std D
Dem	11	50	60	139	35	39	6	340	3.82	1.32
Ind	8	33	47	142	37	40	6	313	3.99	1.27
Rep	2	19	30	99	65	61	14	290	4.53	1.28

ANOVA

- Assumptions
  - the variable in the population is normally distributed for each of the groups $Y_i \sim N(\mu_i, \sigma_i)$
  - the standard deviations in each group are the same $\sigma_1 = \sigma_2 = \sigma$
- Hypothesis
  - $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$ (where g is the number of groups)
  - H_1: at least one of the means differs from others

ANOVA

- Statistic (under the null):

$$F = \frac{BSS/(g-1)}{WSS/(N-g)} \sim F_{g-1,N-g}$$

where

  - $BSS = \sum_{i=1}^{g} n_i(\overline{Y}_i - \overline{Y})^2$: variability between sample means (how group means vary compared to overall mean)
  - $WSS = \sum_{i=1}^{g}(n_i - 1)s_i^2 = (n_1 - 1)s_1^2 + \ldots + (n_g - 1)s_g^2$: variability within each sample (how each observation varies compared to its group mean)

- $TSS = BSS + WSS$: total variability about the means $\Rightarrow$ WSS is the part not explained by groups mean differences ("error sum of squares").

ANOVA

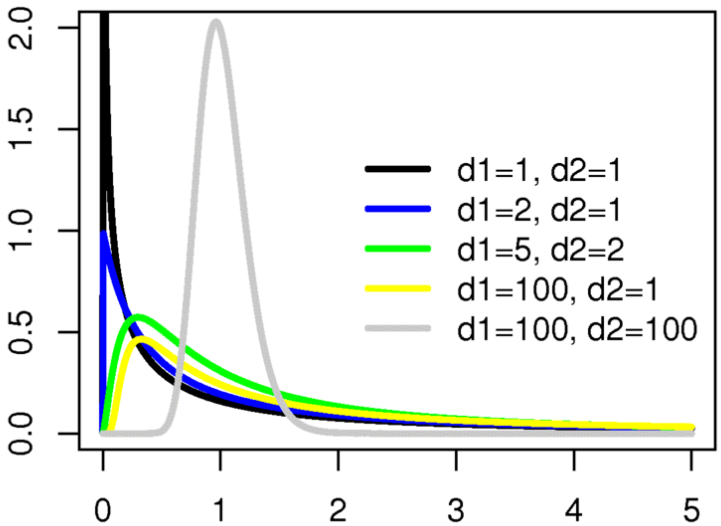
- P-value
  - gives the right hand tail of the F distribution, with degrees of freedom of the numerator and of the the denominator. Large F values provide strong evidence against H_0.
- Conclusion
  - if the P-value is low, there is strong evidence against the null $\Rightarrow$ evidence of some relationship between variables $Y$and X.
  - if the P-value is larger than 0.05 or so, there is no strong evidence against the null hypothesis of equal means; that is, we cannot reject the equality of the means hypothesis.

SPSS

Analyse→Compare Means→one-way ANOVA.

ANOVA

- F sampling distribution
  - Takes only non-negative values and is skewed to the right.
  - Determined by two degrees of freedom terms: d1=$g-1$ and d2=$N-g$



ANOVA

Example

Political ideology by party identification

- $H_0: \quad \mu_{dem} = \mu_{ind} = \mu_{rep}$
- $F = \frac{BSS/(g-1)}{WSS/(N-g)} = \frac{85.2/2}{1570.84/940} = 25.6$
- The between groups estimate is over 25 times the within groups estimate. $P<0.001$. This provides strong evidence against H_0

ANOVA

Example

Political ideology by party identification

Table: Political ideology by party identification - ANOVA

Source	Sum of squares	Deg of free-dom	Mean square	F	*P*-Value
Between	85.38	2	42.69	25.6	0.001
Within	1570.84	940	1.67		
Total	1656.22	942			

ANOVA

Limitations of the $F-$test:

- If H_0 is rejected, test does not specify which means are different or how different they are.
- ANOVA only valid for analysis of effect of one variable (eg. party membership) on another (eg. political attitude). Multivariate analyses require different techniques.
- Stringent assumptions (normal population distribution with identical std devs.)
- Dependent variable to be numerical: level, interval, ratio, etc.

ANOVA

Example

Educational levels and ethnic groups

Table: Education and ethnicity in Ghana

		Education in single years	
Ethnic group	Sample size	Mean	Standard Deviation
Akan	2476	7.14	4.08
Ga/Adangbe	434	6.72	4.75
Ewe	697	6.24	4.48
Others	2063	2.52	4.15

ANOVA

Example

Educational levels and ethnic groups

Table: Education and ethnicity in Ghana - ANOVA

Source	Sum of squares	Deg of free-dom	Mean square	F	*P*-value
Between	25759	3	8586	483.3	0.000
Within	100659	5666	17.8		
Total	126418	5669			

- Conclusion: Strong evidence against the null of equality of average education levels for different ethnic groups.

χ^2 test

χ^2 test

- Topic for self-study
- Reading: Agresti and Finlay (2003), pp. 253-268
- Requirements: understand the problem setting that this method is appropriate, the logic of this test and its limitations.
- The following slides are for the benefit of those, who are interested in the χ^2-test

χ^2 test

- We use χ^2 test to understand the relationship between categorical variables
- RESPONSE (Y): Categorical (≥ 2) EXPLANATORY(X): Categorical (≥ 2)
- Examples:
  - Religious affiliation: Catholic, Protestant, Hindu, Buddhism, other
  - Ethnic origin: White, Black, Asian, Chinese, other
  - Association between the two variables
- Note: comparing proportions for two groups are special cases of one considered here.

χ^2 test

Example

Bed-nets and residence in Ghana

Table: Bednets and residence in Ghana

Bednet while sleeping	Rural	Urban	Total for Ghana
No (%)	70.7	88.0	77.9
Yes (%)	29.3	12.0	22.1
Total	100	100	100
Sample size	3315	2368	5683

χ^2 test

- Assumptions
  - at least 80% of the expected frequences are greater than or equal to 5 (in each cell)
  - if many cells have less than 5 observations, we can try combining categories, in a meaningful way. Otherwise, more observations is required.
- Hypothesis
  - H_0 : there is no association between the categories of one factor and the categories of the other factors (i.e. variables are not related)
  - H_A: the factors are related (i.e. variables are dependent)

χ^2 test

- Statistic (under the null):

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \sim \chi^2_{df=(r-1)(c-1)}$$

where

  - r: number of rows and c: number of columns
  - f_0 : observed frequency
  - $f_e = \frac{(rowtotal)(columntotal)}{TotalSampleSize}$: expected frequency, if variable were independent

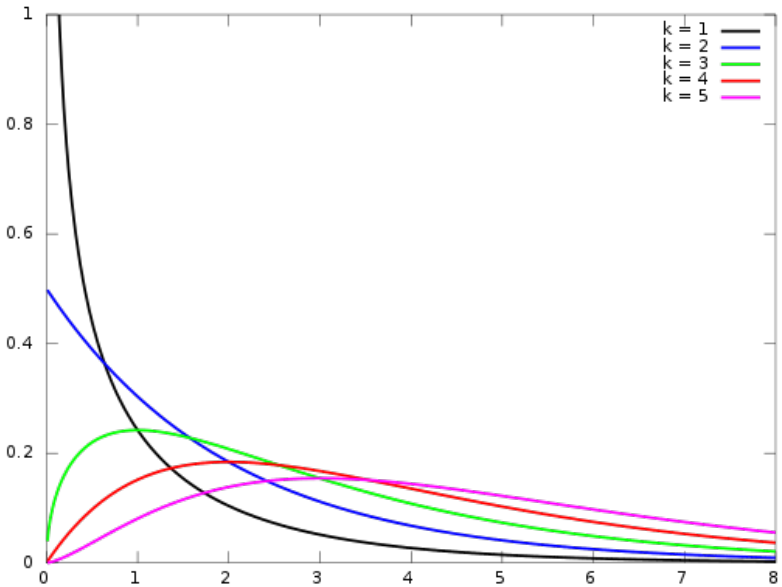
- Logic: if there is no relation between the explanatory (grouping) variable and the response variable, the marginal distribution for all groups should be similar and similar to the marginal total distribution. Then, the expected frequencies should be close to the observed frequency. Instead, if the expected frequency is very distant to the observed frequency, the χ^2-statistic will be high. That is, high values of the χ^2–statistic indicate that the variables are related (dependent).

χ^2 test

- *P*-value
  - right-hand tail probability of the χ^2 graph.
- Conclusion
  - if the *P*-value is relatively low, there is strong evidence against the null of no association between variables, or that the marginal distributions are similar for different groups.

χ^2 test

- χ^2 distribution
  - Takes only non-negative values and is skewed to the right.
  - $\mu = k$ number of degrees of freedom



χ^2 test

Example

Bed-nets and residence in Ghana

Table: Bednets and residence in Ghana

Bednet while sleeping		Rural	Urban	Total
No	Count (f_o)	2343	2084	4427
	Exp. count (f_e)	**2582.4**	**1844.6**	**4427.0**
Yes	Count (f_o)	972	284	1256
	Exp. count (f_e)	**732.6**	**523.4**	**1256.0**
Total	Count (f_o)	3315	2368	5683.0
	Exp. count (f_e)	**3315.0**	**2368.0**	**5683.0**

χ^2 test

Example

Bed-nets and residence in Ghana.
We can conclude that the frequency of bed-net use is different in urban and rural Ghana.

Table: χ^2 test

	Value	Degrees of freedom	*P*-value
χ^2 test	240.9028	1	0.000

χ^2 test

Limitations of χ^2 test

- Test statistic does not say by how much individual cells deviate from independence
- Rejection of independence by chi-square test says nothing about strength of association.
- Value of χ^2 directly proportional to n for a given degree of association. Independence likely to be rejected in large samples even if association is very weak.

Question 1

Based on a random sample of people in a large developing country, a researcher hypothesizes that people who live in districts with higher levels of industrialisation suffer from worse health conditions than people in less industrialised districts. Health conditions are measured as an index from 0 to 100 with higher scores indicating better health conditions. She divides the districts into two groups according to the share of urban population in total population in each district. The following table reports the sample statistics:

Table: Health conditions

	High industrialisation	Low industrialisation
Mean	67	75
St. Dev.	10	11
N	200	200

Question 1(a) I

(a) *Test the hypothesis at the 5% significance level. Give details for each step.* [30%]

Answer

μ_h :mean of health condition in high-industrialisation region

μ_l :mean of health condition in low-industrialisation region

1. Assumptions:
    1. samples are random, independent values
    2. both groups are large (as $n_h = n_l = 200 > 30$) $\Rightarrow$sampling distribution of their difference is of normal distribution (by the central limit theorem)$\Rightarrow$ z-test
2. Hypothesis:
   $H_0 : \mu_h = \mu_l$
   $H_A : \mu_h < \mu_l$

Question 1(a) II

3. Test statistic:

$$\begin{aligned} SE_{(\bar{X}_h - \bar{X}_l)} &= \sqrt{\frac{s_h^2}{n_h} + \frac{s_l^2}{n_l}} \\ &= \sqrt{\frac{10^2}{200} + \frac{11^2}{200}} \\ &= 1.051 \end{aligned}$$

$$z = \frac{(\bar{X}_h - \bar{X}_l) - (\mu_h - \mu_l)}{SE(\bar{X}_h - \bar{X}_l)} = \frac{(67-75)-0}{1.051} = -7.61$$

4. Decision:

   $\alpha = 0.05$ (one-sided test)

   $z_{critical} = z_{0.05} = 1.64$

   $|z| = 7.61 > z_{Critical}$

   $\therefore$ Reject H_0

   1. $P-$value (this is an alternative way of arriving at the same decision). Here the z-score is so great that it's not in the table, so the $P-$value is going to be tiny i.e. much less than 0.05 so we reject the H_0.

5. Conclusion: sufficient evidence to reject the null hypothesis that the health conditions are the same in the areas of high and low industrialisation.

Question 1(b) I

(b) *It is also hypothesized that the type of industrial activity that a region specialises in may affect the health condition of its residents. According to the theory, industrial activities are classified into four major types: labour-intensive, resource-based, medium-technology and knowledge-intensive industries. How could the researcher use ANOVA to test the hypothesis? Give a verbal explanation; show all your steps and discuss the logic of this test. You are not required to carry out the actual computation.* [30%]

Question 1(b) I

Answer

ANOVA: analysis of variance using F-distribution for detecting evidence of differences between population means.

Logic: Are the differences between the groups large enough to reject H_0 and justify the conclusion that the populations represented by the groups are different? (Between group difference $>$ Within group difference)

Steps: 5 steps as usual.

1. Assumptions:
   1. Normal distribution for each group.
   2. Standard deviations in each group are the same.
2. Hypothesis:
   H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
   H_A: *at least one* of the means differs from the others.

Question 1(b) II

3. Test statistic:
   $F = \frac{Between\ sum\ of\ squares/(g-1)}{Within\ sum\ of\ squares/(N-g)} \sim F_{df_1, df_2}$
   $df_1 = g - 1$
   $df_2 = N - g$
   (TSS = BSS + WSS, $F = \frac{BSS/(g-1)}{WSS/(N-g)}$) where $g = 4$
4. Decision/P-value. Either F is greater than $F_{critical}$ (check F-distribution table). Or find the P-value and compare it against α.
5. Conclusions:
   - Good answers should include some examples.
   - Discuss the limitations of ANOVA; is it too weak of a test; good answers use an example in the context of the question.

Question 1(c) I

(c) *Can the test result in (b) tell us whether resource-based industrialisation has worse impact on health of local residents than knowledge-driven industrialisation? If not, what test should the researcher employ?* [20%]

Answer

No, ANOVA cannot tells which group is different from the others. You should give at least one proper test for this hypothesis, for example:

1. Select the two appropriate sub-samples, do a z-test for comparison between two groups
2. Or regression test using dummy (more on this in Classes 5 & 6).

Good answers should give detail on the steps, the interpretation using examples.

Question 1(d) I

(d) *If the health condition of individuals was rated in 3 categories: unhealthy, normal and very healthy, what test would you use to test the hypothesis about the relationship between industrialisation and health? Explain your answers in detail.* [20%]
Now the dependent variable changed to a categorical variable. You need to give details of at least one test. Possible tests are:

1. χ^2-test for three groups.
2. Comparison of proportions between each of the two groups (z-test for large sample). But be careful if you ape comparing all of them (standard error correction)!
3. If you know: ordered probit/logit, multinomial logit is fine to mention, but won't be required or expected in the exam.

Good answers should provide details with examples. Also discuss the limitations!