

Lab 1 Solutions

1. Analysis of cohort studies

File: Whitehall study (wha1110)

Description: This study reports a 20-year follow-up of civil servants. Subjects were recruited into the study over a 2-year period. Exposures of interest included whether and how much each person smoked, and the grade of work classified as high (administrative, professional or executive) and low (clerical or other).

1.1. Descriptive analysis

1. desc

obs:	399	Bereavement data with dates			
vars:	9	24 Oct 2009 13:31			
size:	10,773 (99.9% of memory free)				

storage	display	value			
variable name	type	format	label	variable label	

id	long	%12.0g		id of subject	
dob	int	%dDlCY		Date of birth	
doe	int	%dDlCY		Date of entry into study	
all	float	%8.0g		0 = alive, 1=died	
group	byte	%8.0g		Group	
smok	byte	%8.0g		0=never 1=former 2=0-20 3=>=20	
work	float	%9.0g	worklab		
IHD	byte	%8.0g		0=other_cause 1=death from IHD	
y	float	%9.0g			

Sorted by:					

The dataset contains **399 observations** and **9 variables**. Key variables include:

all: Indicates if the subject is alive (0) or died (1)

smok: Smoking status (0: never, 1: former, 2: 0–20 cigarettes, 3: ≥20 cigarettes)

work: Working group (low/high-grade)

IHD: Indicates death from ischemic heart disease (0: other cause, 1: death from IHD)

2. tab all

0 = alive, 1=died	Freq.	Percent	Cum.
0	121	30.33	30.33
1	278	69.67	100.00
Total	399	100.00	

There were 278 deaths from all causes.

3. tab IHD all, col

0=other_cause, 1=death from IHD	0 = alive, 1=died	Total
0	121	128
1	0	150
Total	121	278
	100.00	100.00

There were 150 deaths from coronary heart disease (IHD)

4. tab smok

0=never, 1=former, 2=0-20, 3=>=20	Freq.	Percent	Cum.
0	278	69.67	69.67
1	51	12.78	82.46
2	30	7.52	89.97
3	40	10.03	100.00
Total	399	100.00	

- Never smokers (0): 278 individuals
- Former smokers (1): 51 individuals
- Light smokers (2: 0-20 cigarettes): 30 individuals
- Heavy smokers (3: ≥20 cigarettes): 40 individuals

5. tab work

work	Freq.	Percent	Cum.
low	163	40.85	40.85
high	236	59.15	100.00
Total	399	100.00	

- Low grade working group: 163 individuals
- High-grade working group: 236 individuals

1.2 Risk Analysis

1. tab smok all

0=never			
1=former			
2=0-20	0 = alive, 1=died		
3=>=20	0	1	Total
0	101	177	278
1	12	39	51
2	4	26	30
3	4	36	40
Total	121	278	399

2. tab smok all, row

0=never			
1=former			
2=0-20	0 = alive, 1=died		
3=>=20	0	1	Total
0	101	177	278
	36.33	63.67	100.00
1	12	39	51
	23.53	76.47	100.00
2	4	26	30
	13.33	86.67	100.00
3	4	36	40
	10.00	90.00	100.00
Total	121	278	399
	30.33	69.67	100.00

3. Interpretation:

For never smokers (0), the risk of death is 63.67%

For former smokers (1), the risk of death is 76.47%

For light smokers (2: 0–20 cigarettes), the risk of death is 86.67%

For heavy smokers (3: ≥ 20 cigarettes), the risk of death is 90.00%

The risk of death increases with increased smoking intensity (i.e., number of cigarettes smoked).

4. cs all work

	work		
	Exposed	Unexposed	Total
Cases	181	97	278
Noncases	55	66	121
Total	236	163	399
Risk	.7669492	.595092	.6967419
	Point estimate		[95% Conf. Interval]
Risk difference	.1718571		.0791852 .2645291
Risk ratio	1.288791		1.115 1.48967
Attr. frac. ex.	.2240789		.1031388 .3287104
Attr. frac. pop	.1458931		
chi2(1) = 13.48 Pr>chi2 = 0.0002			

Interpretation:

$\text{Risk_exposed} = \text{Exposed Cases} / \text{Total Exposed}$

$\text{Risk_unexposed} = \text{Unexposed Cases} / \text{Total Unexposed}$

$\text{Risk difference} = \text{Risk_exposed} - \text{Risk_unexposed}$

$\text{Risk ratio (Relative Risk)} = \text{Risk_exposed} / \text{Risk_unexposed}$

Attributable (prevented) fraction exposed – an estimate of the proportion of exposed cases attributable (prevented) to the exposure = $(\text{Risk_exposed} - \text{Risk_unexposed}) / \text{Risk_exposed}$

Attributable (prevented) fraction population – an estimate of the proportion of all cases attributable (prevented) to the exposure = $(\text{Risktot} - \text{Risk_unexposed}) / \text{Risktot}$, where Risktot: the total risk in the population.

Interpretation:

The risk of death in the high-grade working group (exposed) is 0.7669 or 76.69%.

The risk of death in the low-grade working group (unexposed) is 0.5951 or 59.51%.

The risk ratio (relative risk) is 1.289 (95% CI: 1.115–1.490).

This means that individuals in the high-grade working group have a 28.9% higher risk of death compared to those in the low-grade working group. The 95% confidence interval suggests that the true population risk ratio is likely between 1.115 and 1.490.

The difference is statistically significant ($p = 0.0002$).

The attributable fraction among the exposed (high-grade workers) is 0.2241, meaning that 22.41% of deaths in the high-grade group could potentially be attributed to their work grade (assuming a causal relationship).

These results suggest that there is a significant association between work grade and mortality risk, with high-grade workers having a higher risk of death. However, this finding is counterintuitive, as one might expect higher-grade jobs to be associated with better health outcomes. This unexpected result warrants further investigation into potential confounding factors such as age, job stress, lifestyle differences, or other variables not accounted for in this analysis.

5. cs all work, exact

	work		
	Exposed	Unexposed	Total
Cases	181	97	278
Noncases	55	66	121
Total	236	163	399
Risk	.7669492	.595092	.6967419
	Point estimate		[95% Conf. Interval]
Risk difference	.1718571		.0791852 .2645291
Risk ratio	1.288791		1.115 1.48967
Attr. frac. ex.	.2240789		.1031388 .3287104
Attr. frac. pop	.1458931		
1-sided Fisher's exact P = 0.0002			
2-sided Fisher's exact P = 0.0004			

The main difference between this command and the previous one is the type of statistical test used:

In the first case, we obtain a chi-square test. In the second instance, we obtain Fisher's exact tests.

For large samples like this one, the results of the chi-square test and Fisher's exact test are often very similar.

6. csi 181 97 55 66

	Exposed	Unexposed	Total
Cases	181	97	278
Noncases	55	66	121
Total	236	163	399
Risk	.7669492	.595092	.6967419
	Point estimate		[95% Conf. Interval]
Risk difference	.1718571		.0791852 .2645291
Risk ratio	1.288791		1.115 1.48967
Attr. frac. ex.	.2240789		.1031388 .3287104
Attr. frac. pop	.1458931		
chi2(1) = 13.48 Pr>chi2 = 0.0002			

1.3 Incidence Rate Analysis

1. ir all work y

	work	Exposed	Unexposed	Total
0 = alive, 1=die		181	97	278
y		18647.61	12812.45	31460.06
Incidence Rate		.0097063	.0075708	.0088366
		Point estimate		[95% Conf. Interval]
Inc. rate diff.		.0021356		.0000693 .0042018
Inc. rate ratio		1.282083		.9964124 1.658125 (exact)
Attr. frac. ex.		.220019		-.0036005 .3969092 (exact)
Attr. frac. pop		.1432498		
(midp) Pr(k>=181) =				0.0233 (exact)
(midp) 2*Pr(k>=181) =				0.0467 (exact)

The incidence rate analysis shows that the high-grade (exposed) group has a higher mortality rate (0.0097063 per person-year) compared to the low-grade (unexposed) group (0.0075708 per person-year). This suggests that high-grade workers may be at a higher risk of death.

2. The mortality rate in the total sample is estimated to be 8.8 per 1000 person-years (0.0088366 per person-year).

3. The mortality rates between the two working groups do not differ statistically significantly, as the 95% confidence interval for the incidence rate ratio includes 1 (IRR = 1.282, 95% CI: 0.996–1.658). However, there is a trend towards higher

mortality in the high-grade group that may need further investigation. This unexpected finding likely reflects uncontrolled confounding (e.g. age, duration of follow-up, job stress).

4. The incidence rate difference represents the number of deaths among the exposed that could be prevented if the exposure was completely eliminated. In this case, it suggests that if the exposure (high-grade work) was eliminated, there would be 2.14 fewer deaths per 1000 person-years (0.0021356 per person-year).

5. `iri 181 97 18647.61 12812.45`

	Exposed	Unexposed	Total	
Cases	181	97	278	
Person-time	18647.61	12812.45	31460.06	
Incidence Rate	.0097063	.0075708	.0088366	
	Point estimate		[95% Conf. Interval]	
Inc. rate diff.	.0021356		.0000693	.0042018
Inc. rate ratio	1.282082		.9964121	1.658125 (exact)
Attr. frac. ex.	.2200188		-.0036008	.396909 (exact)
Attr. frac. pop	.1432496			
	(midp)	Pr (k>=181) =	0.0233 (exact)	
	(midp)	2*Pr (k>=181) =	0.0467 (exact)	

Note that at the bottom of the table there are one-sided and two-sided exact significance tests. The exact probability that the number of exposed cases is 181 or more is 0.0233. The exact two-sided probability that the number of exposed cases is different from 181 is 0.0467.

6. `gen y1 = y * 100000`

`tabrate all smok, e(y1) per (100000000)`

table of cases (D), person-years (Y), and rates per 1.000e+08 person-years						
smok	_D	_Y	_rate	ci_low	ci_high	
0	177	2.2e+09	8.114	7.002	9.401	
1	39	4.0e+08	9.631	7.037	13.181	
2	26	2.4e+08	10.908	7.427	16.021	
3	36	3.2e+08	11.209	8.085	15.540	
Chisq test for unequal rates = 4.78 (3 df, p = 0.188)						

The column labelled `_D` gives the total number of events (deaths), and the column labelled `_Y` the total observation time in each group. The rate (fourth column) is multiplied by 1000, since rates are generally small. The final two columns give upper and lower confidence intervals for the rate in each group.

The chi-square test shows that the rates in the four groups are not statistically significantly different.

Notes

1. **e(y1)** (or `exposure(y1)`) specifies the variable that provides the denominator of the rate or SMR, namely the person-years or expected numbers.
2. We define $y1 = y * 100000$, which scales the person-time by 100,000 for presentation purposes. The `per(100000000)` option in the `tabrate` command then multiplies the estimated rate by 100,000,000. Hence:

```
displayed_rate = (events / y1) * 100000000
                = (events / (y * 100000)) * 100000000
                = 1000 * (events / y)
```

Thus, the displayed rates correspond approximately to 1,000 person-years of observation.

3. This rescaling is purely for readability and has no effect on relative measures (e.g., IRRs) or statistical tests.

Note — Person-time (y) vs scaled person-time (y1)

In all incidence-rate analyses, `y` denotes the original person-time (in person-years) and should be used for estimation and inference (e.g., in `ir`, `iri`, `mhrate`). For presentation purposes only, we define $y1 = y * 100000$ and call `tabrate ...`, `e(y1)` `per(100000000)`, which displays rates on a more interpretable scale — approximately per 1,000 person-years. This scaling does not alter relative measures (IRRs) or p-values; it merely rescales the numerical display. Therefore, use `e(y)` for analytical commands and `e(y1)` (with the appropriate `per(...)`) for tables and plots aimed at readability.

To examine the effect of smoking status on the risk of death we calculate the rate ratio of each smoking group vs. the previous one:

```
mhrate all smok, e(y) c(1,0)
```

```
Maximum likelihood estimate of the rate ratio
Comparing smok==1 vs smok==0

RR estimate, lower and upper 95% confidence limits, and
chi-squared test for RR=1 (1 degree of freedom)
```

RR	Lower	Upper	Chisq	p_value
1.187	0.839	1.679	0.941	0.332

```
mhrate all smok, e(y) c(2,1)
```

```
Maximum likelihood estimate of the rate ratio
Comparing smok==2 vs smok==1

RR estimate, lower and upper 95% confidence limits, and
chi-squared test for RR=1 (1 degree of freedom)
```

RR	Lower	Upper	Chisq	p_value
1.133	0.690	1.860	0.242	0.622

```
mhrate all smok, e(y) c(3,2)
```

```
Maximum likelihood estimate of the rate ratio
Comparing smok==3 vs smok==2

RR estimate, lower and upper 95% confidence limits, and
chi-squared test for RR=1 (1 degree of freedom)
```

RR	Lower	Upper	Chisq	p_value
1.028	0.620	1.702	0.011	0.916

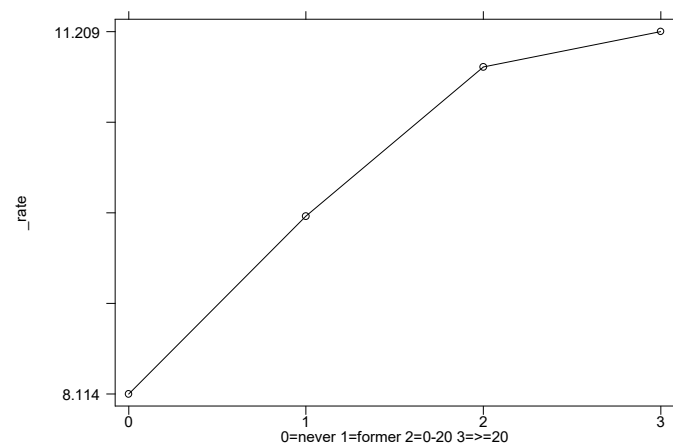
In this case we estimate the rate ratio by comparing each level of exposure to its previous one. Results show that successive groups do not differ statistically significantly with respect to the rate of deaths.

Therefore, analysis of smoking status using `tabrate` and `mhrate` commands shows a trend of increasing mortality rates with increased smoking intensity, although the differences between successive groups are not statistically significant.

The mortality rates per 100,000,000 person-years are:

- Never smokers: 8,114
- Former smokers: 9,631
- Light smokers (0-20 cigarettes): 10,908
- Heavy smokers (≥ 20 cigarettes): 11,209

7. `tabrate all smok, e(y1) per(100000000) graph border
xlab(0(1)3)`



8. `mhrate all smok, e(y)`

Score test for trend of rates with smok

RR estimate, lower and upper 95% confidence limits, and
chi-squared test for trend (1 degree of freedom)

The RR estimate is an approximate estimate of the
rate ratio for one unit increase in smok

RR	Lower	Upper	Chisq	p_value
1.137	1.011	1.278	4.619	0.032

When using the `mhrate` command with all smoking categories, the rate ratio represents the change in mortality rate for each increase in smoking category (i.e., from never to former, former to light smoker, light to heavy smoker). In this case we treat exposure as a continuous variable, and we estimate the rate ratio for one unit increase in smok without any grouping. In this case, the rate ratio of 1.137 (95% CI:

1.011–1.278) suggests that for each increase in smoking category, the mortality rate increases by about 13.7%.

```
9.    gen smok2 = .
      replace smok2 = 0 if smok == 0 | smok == 1
      replace smok2 = 1 if smok == 2 | smok == 3
```

Using the `ir` command with `smok2` would give a single rate ratio comparing smokers to non-smokers. The `tabrate/mhrrate` commands with `smok2` would provide similar information but in a different format. Both analyses would likely show a higher mortality rate for smokers compared to non-smokers, but with less granularity than the four-category analysis. This binary categorization might show a statistically significant difference that wasn't apparent in the more detailed analysis, due to increased statistical power.

2. Analysis of case-control studies

```
1.    tab case ed
```

Group	0=none	1=elementary	2=medium	3=high	Total
0	161	27	17	21	226
1	117	24	13	19	173
Total	278	51	30	40	399

The odds of HIV according to educational level can be calculated from the data provided in the "tab case ed" output:

No education (0): $117/161 = 0.727$

Elementary (1): $24/27 = 0.889$

Medium (2): $13/17 = 0.765$

High (3): $19/21 = 0.905$

There doesn't appear to be a clear trend in the odds as the education level increases. However, to decide this, we need to perform the test of trend.

2. tabodds case ed

ed	cases	controls	odds	[95% Conf. Interval]	
0	117	161	0.72671	0.57273	0.92208
1	24	27	0.88889	0.51292	1.54044
2	13	17	0.76471	0.37143	1.57438
3	19	21	0.90476	0.48643	1.68285
Test of homogeneity (equal odds):					
			chi2(3)	=	0.75
			Pr>chi2	=	0.8626
Score test for trend of odds:					
			chi2(1)	=	0.48
			Pr>chi2	=	0.4890

The output of "tabodds case ed" command shows:

- Odds ratios for each education level compared to the reference level (0 = no education)
- 95% confidence intervals for these odds ratios
- Test of homogeneity: $\chi^2(3) = 0.75$, $p = 0.8626$
- Score test for trend of odds: $\chi^2(1) = 0.48$, $p = 0.4890$

The high p-values for both tests suggest that there's no significant difference in odds of HIV infection across education levels, and no significant trend.

3. Null hypotheses for homogeneity and trend tests:

For both tests, the null hypothesis is $H_0: OR_i = 1$ for all i (no association between education and HIV).

The alternative hypotheses differ:

Homogeneity test: H_1 : At least one $OR_i \neq 1$

Trend test: H_1 : The true ORs increase (or decrease) with increasing level of exposure

What changes between these tests is the alternative hypothesis:

- The test of homogeneity looks for any difference in odds ratios across the education levels. It's sensitive to any pattern of association, not just linear trends.
- The score test for trend specifically looks for a monotonic increase or decrease in the odds ratios as the education level increases. It's designed to detect dose-response relationships.

Both tests are examining whether there's an association between education level and HIV risk, but they're looking for different patterns in that association. The

homogeneity test is more general, while the trend test is looking for a specific type of relationship.

4. mlogit case ed

Score test for trend of odds with ed				
Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
1.072459	0.48	0.4890	0.879655	1.307522
Note: The Odds Ratio estimate is an approximation to the odds ratio for a one unit increase in ed.				

This treats education (exposure) as a continuous variable and estimates the odds ratio for a one-unit increase in education level. The result suggests a slight increase in odds with each level of education, but it's not statistically significant.

5. recode skin 9 = . cc case skin

	Exposed	Unexposed	Total	Proportion exposed
Cases	76	97	173	0.4393
Controls	73	152	225	0.3244
Total	149	249	398	0.3744
	Point estimate		[95% conf. interval]	
Odds ratio	1.631408		1.060123	2.509562 (exact)
Attr. frac. ex.	.3870325		.0567136	.6015241 (exact)
Attr. frac. pop	.1700259			
chi2(1) = 5.51 Pr>chi2 = 0.0189				

We need to recode the missing value (9) to ".", to ensure that Stata treats these values as truly missing rather than as a separate category in our analysis.

This suggests that the odds of HIV are on average 63% higher for women with skin incisions or tattoos than for women without incisions and tattoos. This difference is statistically significant at the 5% level, since the 95% CI does not include the 1 and the p-value of chi-square test is 0.0189.

6. cc case skin, exact

	Exposed	Unexposed	Total	Proportion Exposed	
Cases	76	97	173	0.4393	
Controls	73	152	225	0.3244	
Total	149	249	398	0.3744	
	Point estimate		[95% Conf. Interval]		
Odds ratio	1.631408		1.060123	2.509562	(exact)
Attr. frac. ex.	.3870325		.0567136	.6015241	(exact)
Attr. frac. pop	.1700259				
1-sided Fisher's exact P = 0.0125					
2-sided Fisher's exact P = 0.0216					

The main difference between this command and the previous one is the type of statistical test used:

In the first case, we obtain a chi-square test. In the second instance, we obtain Fisher's exact tests.

7. cc case skin, level(90) exact

	Exposed	Unexposed	Total	Proportion Exposed	
Cases	76	97	173	0.4393	
Controls	73	152	225	0.3244	
Total	149	249	398	0.3744	
	Point estimate		[90% Conf. Interval]		
Odds ratio	1.631408		1.131906	2.348826	(exact)
Attr. frac. ex.	.3870325		.1165347	.5742554	(exact)
Attr. frac. pop	.1700259				
1-sided Fisher's exact P = 0.0125					
2-sided Fisher's exact P = 0.0216					

As expected, the 90% CI is narrower than the 95% CI.

8. `cci 76 97 73 152, exact`

	Exposed	Unexposed	Total	Proportion Exposed	
Cases	76	97	173	0.4393	
Controls	73	152	225	0.3244	
Total	149	249	398	0.3744	
	Point estimate		[95% Conf. Interval]		
Odds ratio	1.631408		1.060123	2.509562	(exact)
Attr. frac. ex.	.3870325		.0567136	.6015241	(exact)
Attr. frac. pop	.1700259				
1-sided Fisher's exact P = 0.0125					
2-sided Fisher's exact P = 0.0216					

This produces the same results as the "cc" command used earlier.

9. `gen ed2 = (ed > 0)`

Then, using `cc case ed2, exact` and `tabodds case ed2` followed by `mhodds case ed2`, we would get odds ratios comparing any education to no education.

In conclusion, this analysis suggests that while skin incisions or tattoos are associated with increased odds of HIV infection, educational level does not show a clear or significant association with HIV status in this population.