

Epidemiologic Analysis: A Case-Oriented Approach

R code from Tomás Aragón, modified 2005-01-27

Chapter 1

Approaches to assess trend in proportions:

1. Start with total chi square statistic (p 4)
2. Compare to linear fit (p. 5)
3. Partition chi square statistic into linear and non-linear components (p. 8)
4. Compare two means using t-like statistic (p. 9)
5. Use correlation coefficient (p. 9)
6. Logistic model approach (p. 10)

R code running

```
12 > ##read multiple.1.data
13 > multdat <- read.table("http://www.medepi.net/selvin/multiple.1.data", header =
14 FALSE, sep = "")
15 > multdat <- data.frame(year = 1983:1991, multdat)
16 > names(multdat) <- c("year", "one+", "one")
17 > attr(multdat, "dict") <- c("year = year of birth",
18 + "one+ = one or more congenital defects (1)",
19 + "one = one congenital defect (0)")
20 > multdat
21   year one+ one
22 1 1983  369 460
23 2 1984  434 434
24 3 1985  506 487
25 4 1986  521 518
26 5 1987  526 488
27 6 1988  605 481
28 7 1989  649 477
29 8 1990  733 395
30 9 1991  688 348
31 > attributes(multdat)
32 $names
33 [1] "year" "one+" "one"
34
35 $row.names
36 [1] "1" "2" "3" "4" "5" "6" "7" "8" "9"
37
38 $class
39 [1] "data.frame"
40
41 $dict
42 [1] "year = year of birth"
43 [2] "one+ = one or more congenital defects (1)"
```

```

1 [3] "one = one congenital defect (0)"
2
3 >
4 > ##TRADITIONAL CHI-SQUARE TEST
5 > ##compare observed vs. expected counts
6 > oi <- t(as.matrix(multdat[,2:3]))
7 > colnames(oi) <- 1983:1991
8 > oi
9      1983 1984 1985 1986 1987 1988 1989 1990 1991
10 one+  369 434 506 521 526 605 649 733 688
11 one   460 434 487 518 488 481 477 395 348
12 > rtot <- apply(oi, 1, sum)
13 > ctot <- apply(oi, 2, sum)
14 > ei <- outer(rtot, ctot, "*")/sum(oi)
15 > ei
16      1983      1984      1985      1986      1987      1988      1989      1990
17 one+ 457.3636 478.8801 547.8433 573.2217 559.4291 599.1519 621.2201 622.3235
18 one  371.6364 389.1199 445.1567 465.7783 454.5709 486.8481 504.7799 505.6765
19      1991
20 one+ 571.5666
21 one  464.4334
22 >

```

$$\chi^2_T = \sum \left[\frac{(o_i - e_i)^2}{e_i} \right], d.f. = c - 1$$

```

24 > chi2.T <- sum((oi - ei)^2/ei)
25 > chi2.T
26 [1] 169.3757
27 >
28
29 > pchisq(q = chi2.T, df = 8, lower.tail = FALSE)
30 [1] 1.743111e-32
31 >
32 > ## one definition of residuals
33 >

```

$$r_i = \frac{(o_i - e_i)^2}{e_i}$$

```

35 > residi <- (oi - ei)/sqrt(ei)
36 > residi
37      1983      1984      1985      1986      1987      1988      1989
38 one+ -4.131833 -2.050883 -1.787713 -2.181172 -1.413358  0.2389174  1.114572
39 one  4.583682  2.275163  1.983213  2.419701  1.567920 -0.2650449 -1.236459
40      1990      1991
41 one+  4.436570  4.870170
42 one -4.921744 -5.402761
43 >
44 > sum(residi^2) ##chi2.T
45 [1] 169.3757
46 >
47 >
48 > #####ASSESSING LINEARITY AND FIT, p. 5

```

```

1 >
2                                     
$$E[p_i] = p_i = a + b(x_i - 1982)$$

3 > ##pi = a + b*(xi - 1982)
4 >
5 > ##create complete data set
6 > xi <- c(rep(1983:1991, oi["one+",]), rep(1983:1991, oi["one",]))
7 > yi <- c(rep(rep(1, ncol(oi)), oi["one+",]), rep(rep(0, ncol(oi)), oi["one",]))
8 >
9                                     
$$S_{xx} = \sum_{i=1}^{919} (x_i - \bar{x})^2$$

10 > Sxx <- sum((xi - mean(xi))^2)
11 > Sxx
12 [1] 58000.32
13 >
14                                     
$$S_{yy} = \sum_{i=1}^{919} (y_i - \bar{y})^2$$

15 > Syy <- sum((yi - mean(yi))^2)
16 > Syy
17 [1] 2255.371
18 >
19                                     
$$S_{xy} = \sum_{i=1}^{919} (x_i - \bar{x})(y_i - \bar{y})$$

20 > Sxy <- sum((xi - mean(xi))*(yi - mean(yi)))
21 > Sxy
22 [1] 1483.174
23 >
24                                     
$$\hat{b} = \frac{S_{xy}}{S_{xx}}$$

25 > bhat <- Sxy/Sxx
26 > bhat
27 [1] 0.02557183
28 > ##calculate fitted proportions, p. 7
29 >
30                                     
$$\hat{a}' = \bar{y} - \hat{b}\bar{x}$$

31 > ahat.prime <- mean(yi) - bhat*mean(xi)
32 > ahat.prime
33 [1] -50.2649
34 >
35                                     
$$\hat{a} = \bar{y} - \hat{b}(x_i - 1982)$$

36 > ahat <- mean(yi) - bhat*mean(xi-1982)
37 > ahat
38 [1] 0.4184592
39 >
40 > ##calculate observed proportions
41 > phati <- sweep(oi, 2, ctot, "/")["one+",]
42 > phati
43      1983      1984      1985      1986      1987      1988      1989      1990
44 0.4451146 0.5000000 0.5095670 0.5014437 0.5187377 0.5570902 0.5763766 0.6498227

```

```

1      1991
2 0.6640927
3 > years <- 1983:1991
4 >
5
6 
$$\tilde{p}_i = \bar{a} + \hat{b}(x_i^* - 1982), \text{ where } x_i^* = \{1983 \text{ to } 1991\}$$

7 > ptildei <- ahat + bhat * (years - 1982)
8 > ptildei
9 [1] 0.4440310 0.4696028 0.4951746 0.5207465 0.5463183 0.5718901 0.5974619
10 [8] 0.6230338 0.6486056
11 > ni <- ctot
12 >
13 
$$z_i = \frac{\hat{p}_i - \tilde{p}_i}{S_{\hat{p}_i}}, \text{ where } S_{\hat{p}_i} = \sqrt{\hat{p}_i \frac{(1 - p_i)}{n_i}}$$

14 > Spi <- sqrt(phati * (1 - phati)/ni)
15 > Spi
16      1983      1984      1985      1986      1987      1988      1989
17 0.01726078 0.01697111 0.01586412 0.01551174 0.01569083 0.01507320 0.01472563
18      1990      1991
19 0.01420322 0.01467385
20 > zi <- (phati - ptildei)/Spi
21 > zi
22      1983      1984      1985      1986      1987      1988
23 0.06277877 1.79111413 0.90722543 -1.24439780 -1.75775439 -0.98186718
24      1989      1990      1991
25 -1.43188375 1.88611567 1.05541883
26 >
27 
$$\chi_{fit}^2 = \sum z_i^2$$

28 > chi2.fit <- sum(zi^2)
29 > chi2.fit
30 [1] 16.35901
31 >
32 
$$1 - Pr(X \leq x)$$

33 > pchisq(q = chi2.fit, df = 7, lower.tail = FALSE)
34 [1] 0.02203132
35 >
36 > ##Table 1.4. p. 7
37 > round(cbind(phati, ptildei, pdiffi = phati-ptildei, Spi, zi),3)
38      phati ptildei pdiffi  Spi  zi
39 1983 0.445 0.444 0.001 0.017 0.063
40 1984 0.500 0.470 0.030 0.017 1.791
41 1985 0.510 0.495 0.014 0.016 0.907
42 1986 0.501 0.521 -0.019 0.016 -1.244
43 1987 0.519 0.546 -0.028 0.016 -1.758
44 1988 0.557 0.572 -0.015 0.015 -0.982
45 1989 0.576 0.597 -0.021 0.015 -1.432
46 1990 0.650 0.623 0.027 0.014 1.886
47 1991 0.664 0.649 0.015 0.015 1.055
48 >
49 > #####PARTITIONING CHI-SQUARE, p. 8

```

```

1 >
2
3 > ##need to get chi2.L, then chi2.NL
4 > n <- length(xi)
5 >
6
7 > S2bhat <- Syy/(n*Sxx)
8 > S2bhat
9 [1] 4.264227e-06
10 >
11 > z2 <- chi2.L <- (bhat/sqrt(S2bhat))^2
12 > z2
13 [1] 153.3498
14 >
15 > chi2.NL <- chi2.T - chi2.L
16 > chi2.NL
17 [1] 16.02589
18 >
19 > ##proportion of chi2.T variation "explained" by straight line, p. 8
20 >
21
22 > chat <- chi2.L/chi2.T
23 > chat
24 [1] 0.9053826
25 >
26 > #####COMPARING TWO MEAN VALUES, p. 9
27 > tapply(xi, yi, mean)
28      0      1
29 1986.848 1987.505
30 >
31 > ##alternative
32 > xbar1 <- mean(xi[yi==1])
33 > xbar1
34 [1] 1987.505
35 > xbar2 <- mean(xi[yi==0])
36 > xbar2
37 [1] 1986.848
38 >
39 > m1 <- rtot["one+"]
40 > m2 <- rtot["one"]
41 >
42
43 > Sx1x2 <- sqrt((Sxx/n)*(1/m1 + 1/m2))
44 > z <- (xbar1 - xbar2)/Sx1x2
45 > z

```

$$\chi_T^2 = \chi_L^2 + \chi_{NL}^2$$

$$z = \frac{\hat{b}}{S_{\hat{b}}}, \text{ where } S_{\hat{b}} = \frac{S_{yy}}{n S_{xx}}, \text{ and } z^2 = \chi_L^2$$

$$\hat{c} = \frac{\chi_L^2}{\chi_T^2}$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{S_{\bar{x}_1 - \bar{x}_2}}, \text{ where } S_{\bar{x}_1 - \bar{x}_2}^2 = \frac{S_{xx}}{n} \left[\frac{1}{m_1} + \frac{1}{m_2} \right]$$

```

1     one+
2 12.38345
3 >
4
5 > z^2 ## = chi2.L
6     one+
7 153.3498
8 >
9 > #####USE CORRELATION COEFFICIENT
10 >
11
12 > rxy <- Sxy/sqrt(Sxx*Syy)
13 > rxy
14 [1] 0.1296785
15 > n*rxy^2 ## = chi2.L
16 [1] 153.3498
17 >
18 > #####USE LOGISTIC MODEL
19 >
20
21
22 > ##using full data
23 > xi.ctr <- xi-1982
24 > logm <- summary(glm(yi~xi.ctr, family = binomial(link = logit)))
25 > logm
26
27 Call:
28 glm(formula = yi ~ xi.ctr, family = binomial(link = logit))
29
30 Deviance Residuals:
31     Min       1Q   Median       3Q      Max
32 -1.443  -1.214   0.933   1.098   1.276
33
34 Coefficients:
35             Estimate Std. Error z value Pr(>|z|)
36 (Intercept) -0.332660   0.048488  -6.861 6.86e-12 ***
37 xi.ctr       0.104332   0.008469  12.319 < 2e-16 ***
38 ---
39 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
40
41 (Dispersion parameter for binomial family taken to be 1)
42
43     Null deviance: 12544  on 9118  degrees of freedom
44 Residual deviance: 12390  on 9117  degrees of freedom
45 AIC: 12394

```

$$z^2 = \chi_L^2$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \text{ where } \chi_L^2 = nr_{xy}^2$$

$$\log\text{-odds}_i = \log\left(\frac{p_i}{1-p_i}\right) = A + B(x_i - 1982)$$

$$p_i = \frac{1}{1 + e^{-[A+B(x_i-1982)]}}$$

```

1
2 Number of Fisher Scoring iterations: 4
3
4 > A <- logm$coef[1]
5 > A
6 (Intercept)
7 -0.3326601
8 > B <- logm$coef[2]
9 > B
10 xi.ctr
11 0.1043317
12 >
13 > ##using table of counts and weights
14 > yrs.ctr <- years-1982
15 > logm.wt <- summary(glm(phati~yrs.ctr, weights = ctot,
16 + family = binomial(link = logit)))
17 > logm.wt
18
19 Call:
20 glm(formula = phati ~ yrs.ctr, family = binomial(link = logit),
21 weights = ctot)
22
23 Deviance Residuals:
24 Min 1Q Median 3Q Max
25 -1.812 -1.271 0.113 1.149 1.873
26
27 Coefficients:
28 Estimate Std. Error z value Pr(>|z|)
29 (Intercept) -0.332660 0.048488 -6.861 6.86e-12 ***
30 yrs.ctr 0.104332 0.008469 12.319 < 2e-16 ***
31 ---
32 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
33
34 (Dispersion parameter for binomial family taken to be 1)
35
36 Null deviance: 171.212 on 8 degrees of freedom
37 Residual deviance: 17.206 on 7 degrees of freedom
38 AIC: 87.259
39
40 Number of Fisher Scoring iterations: 3
41
42 > A <- logm.wt$coef[1]
43 > A
44 (Intercept)
45 -0.3326601
46 > B <- logm.wt$coef[2]
47 > B
48 yrs.ctr
49 0.1043317
50 >
51 >
52 > ptildeprimei <- 1/(1 + exp(-(A + B * (years - 1982))))
53 > ptildeprimei
54 [1] 0.4431646 0.4690405 0.4950839 0.5211541 0.5471095 0.5728112 0.5981258

```

```

1 [8] 0.6229277 0.6471021
2 > ni <- ctot
3 > Spi <- sqrt(phati * (1 - phati)/ni)
4 > Spi
5      1983      1984      1985      1986      1987      1988      1989
6 0.01726078 0.01697111 0.01586412 0.01551174 0.01569083 0.01507320 0.01472563
7      1990      1991
8 0.01420322 0.01467385
9 > zi <- (phati - ptildeprimei)/Spi
10 > zi
11      1983      1984      1985      1986      1987      1988      1989
12 0.1129721 1.8242481 0.9129436 -1.2706745 -1.8081782 -1.0429769 -1.4769650
13      1990      1991
14 1.8935807 1.1578779
15 > chi2.logistic.fit <- sum(zi^2)
16 > chi2.logistic.fit
17 [1] 17.25379
18 > pchisq(q = chi2.logistic.fit, df = 7, lower.tail = FALSE)
19 [1] 0.01583100
20 >
21 > ##Table 1.5, p. 7
22 > round(cbind(phati, ptildeprimei, pdiffi = phati-ptildeprimei, Spi, zi),3)
23      phati ptildeprimei pdiffi  Spi  zi
24 1983 0.445      0.443 0.002 0.017 0.113
25 1984 0.500      0.469 0.031 0.017 1.824
26 1985 0.510      0.495 0.014 0.016 0.913
27 1986 0.501      0.521 -0.020 0.016 -1.271
28 1987 0.519      0.547 -0.028 0.016 -1.808
29 1988 0.557      0.573 -0.016 0.015 -1.043
30 1989 0.576      0.598 -0.022 0.015 -1.477
31 1990 0.650      0.623 0.027 0.014 1.894
32 1991 0.664      0.647 0.017 0.015 1.158

```

33

1 **See also**

2 **##stats**

3 glm
4 pchisq

5 **##operations**

6 :
7 rep
8 t
9 matrix
10 as.matrix
11 apply
12 tapply
13 outer
14 sweep

15 **##arithmetic**

16 sum
17 round
18 +
19 -
20 *
21 /

22 **##manipulation**

23 "<-"
24 read.table
25 data.frame
26 names
27 attr
28 attributes
29 colnames
30 rownames
31 summary
32 nrow
33 ncol