Contents

- Data of form
  y_i= 0 or 1 for Success/Failure
- Observational Studies
  - Prospective – Relative Risks
  - Retrospective studies – Odds Ratios
- Methods
  - McNemars's Test
    - Matched Pairs
  - Logistic Modelling
    - to incorporate covariates

- **Observational Studies**
  - Epidemiological studies where characteristic is observed and relationship with other factors is inferred as distinct from a controlled study where cases and controls are determined by experimenter (e.g. by randomization)
    - e.g. heart disease and smoking
  - Prospective and retrospective

- **Prospective Studies**
  - Subjects followed forward through time
  - e.g. known to have been exposed
    - e.g. smoker
  - outcome is observed
    - risk of outcome calculated for exposed group
    - compared with risk of outcome for control group known not to be exposed

	Outcome			
	Positive	Negative	Total	Risk
Exposed	a	b	a+b	a/(a+b)
Non-exposed	c	d	c+d	c/(c+d)

Relative Risk =RR= [a/(a+b)]/[c/(c+d)]=a(c+d)/c(a+b)

$$S.E.\{\log_e(RR)\}=\sqrt{\frac{1}{a}-\frac{1}{a+b}+\frac{1}{c}-\frac{1}{c+d}}$$

	Apgar < 7			
	Yes	No	Total	Risk
Symmetric	2	14	16	12.5%
Asymmetric	33	58	91	36.2%

RR=0.345, $\log_e(RR)$= -1.065+/-0.676
so a 95%CI for the RR is (0.11, 1.05)

This includes 1.0, so evidence of effect of asymmetric growth pattern on Apgar score is not significant at 5% level

- Retrospective studies
  - Condition of interest
  - Cases identified & followed back through time to see if they were exposed to risk factor
    - Odds of being exposed calculated for cases
    - Controls not exhibiting condition of interest found
      - matched closely in other respects to cases
    - Compared with odds of exposure for controls
      - not sensible to calculate 'risk' of being a case for exposed group since can just get more or fewer controls

	Case	Control
Exposed	a	b
Non-exposed	c	d
odds	a/c	b/d

Odds ratio = [a/c]/ b/d] =ad/bc

S.E.{$\log_e$(OR)}=$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

	Enamel Erosion		
swimming	Yes	No	Total
≥ 6 hours	32	118	150
< 6 hours	17	127	144
odds	1.88	0.93	

OR= 2.03 , s.e.($\log_e$(OR)=0.323 so 95% CI for the Odds Ratio is (1.07, 3.84) and so good evidence that the odds of enamel erosion are increased for those swimming more than six hours per week

- **Matched pairs**
  - χ^2-test < –>2-sample t-test
  - McNemar's Test < – > paired t-test
- Rheumatoid arthritis study, two treatments A & B.
- Response caused? 1=yes, 0=no

		response		
		yes	no	
treatment	A	11	37	48
	B	20	28	48

- **χ^2-test????**
- **☠ INVALID ☠**
  - only 48 subjects in the table not 96
    - All subjects received both treatments

- Instead consider like/unlike responses

		B: yes	B: no	
A	yes	8	3	11
	no	12	25	37
		20	28	48

- Subjects who respond in a like way to the treatments give no information on ***differences*** between treatments.
  - 8 responded to both treatments
  - 25 responded to neither treatment
  - These give no evidence of differences

- 3 responded to A but not to B
- 12 responded to B but not to A

		B: yes	B: no	
A	yes	8	**3**	11
	no	**12**	25	37
		20	28	48

- Suggests more respond to B than to A
  - i.e. may be evidence of difference between the treatments
- McNemar's Test:
  - If no difference then number responding to A but not B ~ Binomial(M,0.5) where M=number responding differently
    - Observed 3 out of 15

- McNemar's Test
  - Can calculate (e.g. Neave 1.3)

    P[3 or fewer|B(15,0.5)]=0.035
  - i.e. H_0:no difference between treatments is rejected with p<0.05 (actually p=0.035)

- McNemar's Test
  - For large samples then use

$$\frac{(n_{10}-n_{01})^2}{n_{10}+n_{01}} \sim \chi^2_1$$

    - (Normal approx to binomial)$^2 \sim \chi^2_1$

- **Logistic Modelling**
  - Data are $y_i = 1$ or 0 (S or F)
  - Can't express y_i as a linear model
  - Idea is to model ***probability of success*** as a function of explanatory variables
  - However, must have probabilities in range 0 to 1

- Use a logisitic function:
  - e.g. patients on treatment or placebo
    - Code as $x_i = 1$ or 0
  - Measure Success or Failure
    - Code as $y_i = 1$ or 0
  - Model:

$$P[Y_i = 1] = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = 1 - P[Y_i = 0]$$

- Can estimate β_0 and β_1 from data:
- **R**
  - function `glm()` with `family=binomial`, `weights=freqs`
  - followed by `anova()`
- SPSS:
  - Analyze>Regression>Binary Logistic Regression
- Minitab:
  - Stat>Regression>Binary Logistic Regression
- S-PLUS:
  - Statistics>Regression>Logistic

- Interpretation
  - on placebo $P[Y_i = 1] = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
  - on treatment $P[Y_i = 1] = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$
    - β_1 is 'effect' of treatment

- More specifically:

$$\ln\left\{\frac{P[Y_i = 1]}{P[Y_i = 0]}\right\} = \beta_0 + \beta_1 x_i$$

  - $x_i = 0$ for placebo, 1 for treatment
  - i.e ***log-odds of success = linear fn. of covariate***

$$\ln\left\{\frac{P[Y_i = 1]}{P[Y_i = 0]}\right\} = \beta_0 + \beta_1 x_i$$

- If $\beta_1 > 0$ then $P[Y_i=1|x_i=1] > P[Y_i=1|x_i=0]$
  - i.e. if $\beta_1 > 0$ then treatment improves chance of success
  - i.e. if $\beta_1 < 0$ then treatment lessens chance of success

- Can extend to include lots of covariates
  - $\beta_j > 0 \Rightarrow$ P(success) ↗ as x_j ↗ & P(success) ↘ as x_j ↘
  - $\beta_j < 0 \Rightarrow$ P(success) ↘ as x_j ↗ & P(success)↗ as x_j ↘.

- Can test whether covariates affect probability of success:
- Package calculations provide standard errors of estimates

- Partial z-test:

$$H_0: \beta_j = 0$$

- test compares

$$\frac{\hat{\beta}_j}{\sqrt{\text{var}(\hat{\beta}_j)}}$$

with N(0,1) points

- Example: Y_i=1 if heart disease:

	Numerical variable	logistic coef	
factor	x_j	β_j	z-value
1:treatment	0=placebo,1=treatment	–0.32	–2.9
2:age	ln(age)	3.0	6.3
3:smoking	0=non-smok, 1=smoker	0.83	6.8
4:father's hist	0=alive, 1=dead	0.64	3.6
5:systolic BP	Systolic BP in mm Hg	0.011	3.7
6:cholesterol	Cholesterol in mg/dl	0.0095	5.6
	constant term β_0=–19.60		

- $\Phi^{-1}(.005)=z_{.005}= -2.58$ $z_{.025}= -1.96$
  - (1% level) (5% level)
  - ♦ Treatment: significant, $p < 0.01$; $\beta_1 < 0$;
- Probability of IHD is lower on treatment than on placebo

- Prognostic factors:
  - ♦ all five significant ($p < 0.01$);
  - ♦ all have positive m.l.e.'s,
  - ♦ ∴ probability of IHD increases with
    - age
    - smoking
    - 'poorer heredity'
    - high blood pressure
    - high cholesterol.

- **Relative Risks**

$$\frac{P[Y=1|x_1=1]}{P[Y=0|x_1=1]} \Big/ \frac{P[Y=1|x_1=0]}{P[Y=0|x_1=0]}$$

is *relative risk of 'success' on treatment*

$$=\exp\{\beta_1\}$$

- strictly this is an ***odds ratio*** but conventionally is interpreted as an ***approximate relative risk***

- estimated relative risk is $e^{-0.32} = 0.73 < 1$
- i.e. odds of getting IHD are 27% lower on treatment
  - (after allowing for the other prognostic factors)
- Can calculate approx confidence intervals for β_1 as
  estimate +/- 2 x s.e.

- $-0.32 \pm 2 \times 0.11 = (-0.10, -0.54)$
- $\exp\{\beta_1\}$ has 95% C. I. $(e^{-0.54}, e^{-0.1}) = (0.58, 0.90)$

- 95% confidence limits for the reduction due to treatment in odds of getting IHD are 10% and 42%.

- **Example**

	treatment	
	BP	CP
response	77	90
no response	61	45

- $Y_{11}=77$, $E(Y_{11})=138\times167/273=84.42$
  $var(Y_{11}) = 138\times133\times167\times106/(273^2\times272)$
  $=16.268$
- So $T_{MH}=3.38<3.84=\chi^2_{1;0.95}$ and $p=0.066$

Implementation in R

- Same example:
  - frequencies indicated by *Weights*

```
> bpcp
  frequency response treatment
1        77        1         1
2        61        0         1
3        90        1         0
4        45        0         0
> attach(bpcp)
> bpcp.glm<-glm(response~treatment,
    family=binomial, weights=frequency)
```


Implementation in R

```
> bpcp.glm<-glm(response~treatment,
        family=binomial, weights=frequency)
 bpcp.glm

Call:  glm(formula = response ~ treatment, family
  = binomial, weights = frequency)

Coefficients:
(Intercept)    treatment
     0.6931      -0.4602

Degrees of Freedom: 3 Total (i.e. Null); 2
  Residual
Null Deviance:      364.7
Residual Deviance:  361.3        AIC: 365.3
```


Implementation in R

```
> bpcp.glm<-glm(response~treatment,
            family=binomial, weights=frequency)

> summary(bpcp.glm)
Coefficients:
              Estimate Std. Error   z value Pr(>|z|)
(Intercept) -4.994e-16  1.000e-01 -4.99e-15  1.00000
treatment    4.055e-01  1.429e-01     2.838  0.00454
  **
---
```


Implementation in R

```
> anova(bpcp.glm)
Analysis of Deviance Table
Model: binomial, link: logit
Response: response

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
NULL                         3     364.71
treatment  1     3.40         2     361.31
```

approximately equal to the χ^2 value

- Example of combining trials in **R**
  - Data from § 8.2 on Simpson's Paradox

	centre 1			centre 2		
	S	F		S	F	
trt	30	70	30%S	210	90	70%
plac	120	180	40%S	80	20	80%
	150	250		290	110	

```
> frequency<-c(30,120,70,180,210,80,90,20)
> response<-c(rep(c(1,1,0,0),2))
> treatment<-c(rep(c(1,0),4))
> centre<-c(rep(0,4),rep(1,4))
> simpson<-cbind(frequency,response,treatment,centre)
> simpson
```


- Example of combining trials in **R**
  - Data from § 8.2 on Simpson's Paradox

	centre 1			centre 2		
	S	F		S	F	
trt	30	70	30%S	210	90	70%
plac	120	180	40%S	80	20	80%
	150	250		290	110	

```
> simpson
     frequency response treatment centre
[1,]        30        1         1      0
[2,]       120        1         0      0
[3,]        70        0         1      0
[4,]       180        0         0      0
[5,]       210        1         1      1
[6,]        80        1         0      1
[7,]        90        0         1      1
[8,]        20        0         0      1
```


- Example of combining trials in **R**

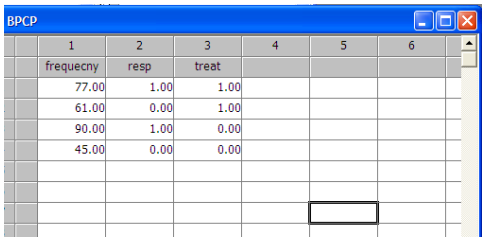
```
> simpson.glm<-glm(response~treatment+centre,
+           family=binomial, weights=frequency)
> summary(simpson.glm)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3958     0.1117  -3.543 0.000396 ***
treatment    -0.4849     0.1853  -2.616 0.008891 **
centre        1.7391     0.1859   9.357  < 2e-16 ***
```

coefficient < 0 so treatment reduces chance of success

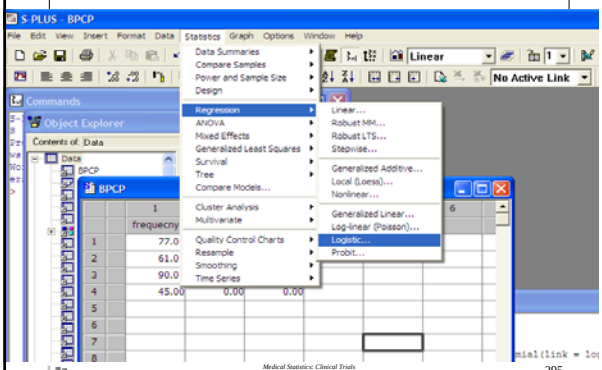
p-value small so significant evidence of effect

S-Plus calculation

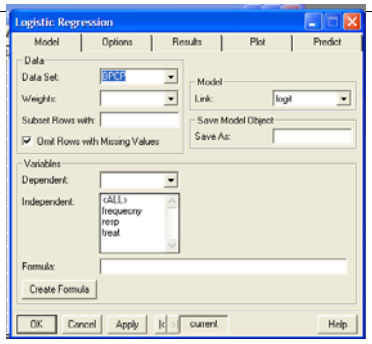
- Same example:
  - Frequencies indicated by *Weights*



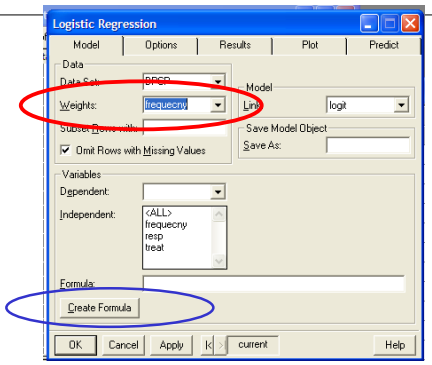
S-Plus calculation



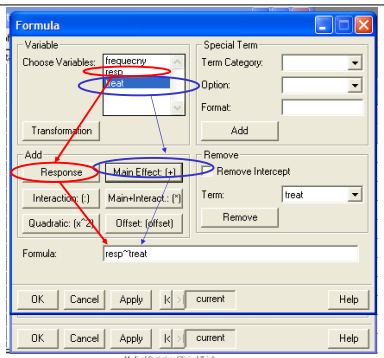
S-Plus calculation



S-Plus calculation



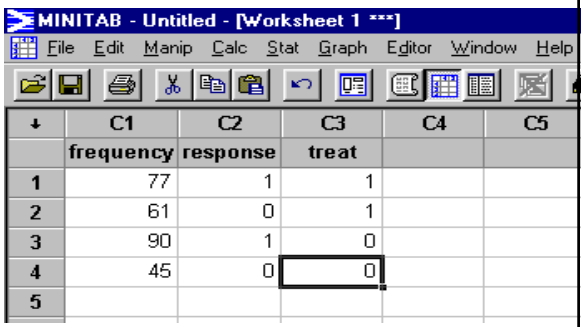
S-Plus calculation



S-Plus calculation

```
Coefficients:
                Value Std. Error  t value
(Intercept)  0.693033   0.181784  3.81239
      treat -0.460102   0.249845 -1.84155

(Dispersion Parameter for Binomial family
taken to be 1 )

    Null Deviance: 364.713 on 3 degrees of
freedom

Residual Deviance: 361.308 on 2 degrees of
freedom

Number of Fisher Scoring Iterations: 2
```

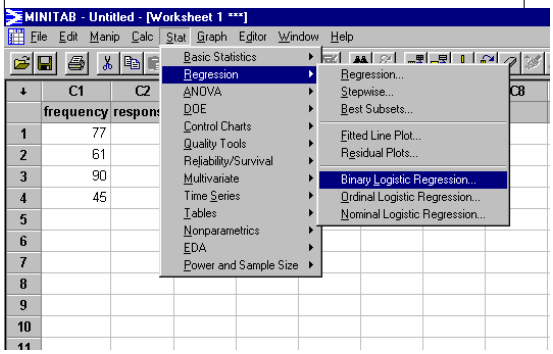

Minitab calculation



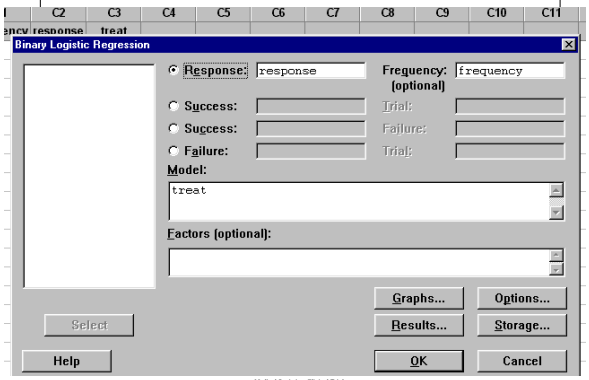
	C1	C2	C3	C4	C5
	frequency	response	treat		
1	77	1	1		
2	61	0	1		
3	90	1	0		
4	45	0	0		
5					

	treatment	
	BP	CP
response	77	90
no response	61	45

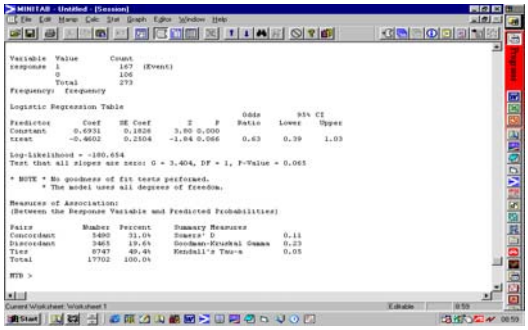
Minitab calculation



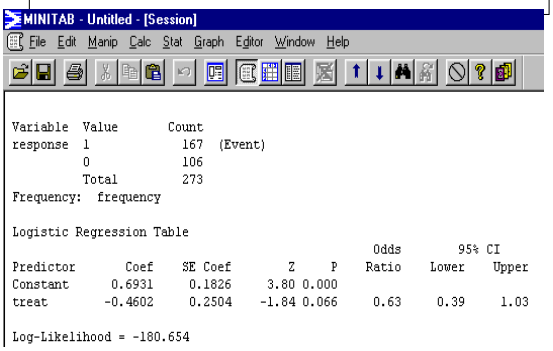
Minitab calculation



Minitab calculation



Minitab calculation



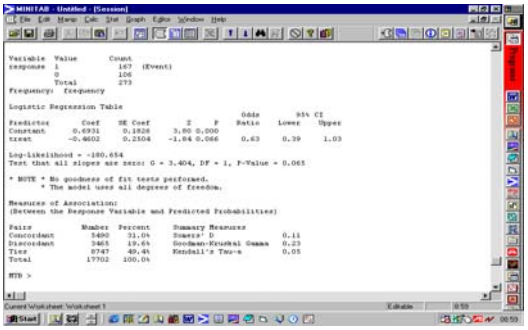
Minitab calculation

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P
Constant	0.6931	0.1826	3.80	0.000
treat	-0.4602	0.2504	-1.84	0.066

- Recall T_{MH}=3.38 with p=0.066
  - Note $(-1.84)^2 = 3.386$

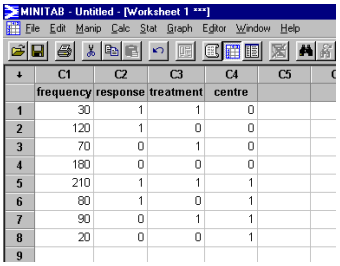
Minitab calculation



Minitab calculation

~~(S-plus left as exercise)~~

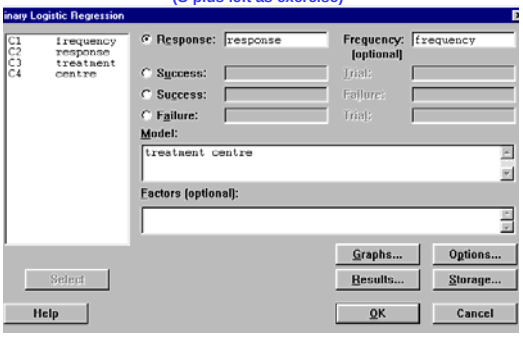
- Example of combining trials
  - Data from § 8.2 on Simpson's Paradox



	C1	C2	C3	C4	C5
	frequency	response	treatment	centre	
1	30	1	1	0	
2	120	1	0	0	
3	70	0	1	0	
4	180	0	0	0	
5	210	1	1	1	
6	80	1	0	1	
7	90	0	1	1	
8	20	0	0	1	
9					

Minitab calculation

~~(S-plus left as exercise)~~



Minitab calculation

~~(S-plus left as exercise)~~

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-0.3958	0.1117	-3.54	0.000			
treatmen	-0.4849	0.1854	-2.62	0.009	0.62	0.43	0.89
centre	1.7391	0.1859	9.36	0.000	5.69	3.95	8.19

Log-Likelihood = -496.323

- This shews that there is strong evidence that the odds of success on treatment are markedly lower than on placebo, even allowing for the difference in success rates between the centres.
  - Note also strong evidence of a higher success rate in centre 2

- **Summary and Conclusions**
  - **Observational Studies**
    - Prospective and retrospective studies
  - Prospective:–
    - Calculate risk of outcome
    - Look at ***relative risks***
  - Retrospective:–
    - Calculate odds of exposure
    - Look at ***odds ratios***

- **Summary and Conclusions**
  - McNemar's test
    - Matched pairs binary responses: uses only **unlike pairs**
  - Logistic Regression
    - models log-odds as a linear in the covariates
    - positive coefficients indicate factor increases the risk of 'success'

- Logistic models allow estimation of **relative risks** (including C.I.s)
- Logistic models can be implemented in most standard statistical packages
  - SPSS
  - S-PLUS / R
  - Minitab