Erasmus 2002

Suppose that during a follow–up of 10 subjects for 5 years, 4 of them died (or failed). Without the possibility of generalization this statement has no scientific interest.

To generalize the statement we assume that the data were generated by some random mechanism. Random sampling from a population is one possibility. For the 10 subjects in the follow–up study we assume a random mechanism in which the outcome for each subject is either failure or survival, but we do not specify the probability of failure.

The unknown probability of failure is denoted by the Greek letter π.

Simulation

For any given value of π we can use a computer to simulate data generated by the random mechanism. For example, if $\pi = 0.1$, a computer or calculator is used to generate random numbers between 0 and 1; 50 such numbers are shown below.

```
0.14 0.64 0.56 0.60 0.68 0.11 0.62 0.06 0.56 0.87
0.26 0.04 0.42 0.90 0.52 0.84 0.21 0.56 0.26 0.95
0.28 0.12 0.41 0.72 0.87 0.46 0.42 0.89 0.06 0.68
0.72 0.70 0.23 0.10 0.59 0.35 0.81 0.10 0.58 0.11
0.97 0.22 0.26 0.17 0.76 0.37 0.38 0.97 0.58 0.80
```

Each number corresponds to an imaginary, or simulated, subject. The subject fails if his random number is ≤ 0.1 and survives if it is > 0.1. In this way the probability of failure will be 0.1.

How many failures are there for the first row of 10 simulated subjects?

Different values of π

1=Failure, 0=Survivor

Simulation 1	0 0 0 0 0 0 0 1 0 0	$\pi = 0.1$
Simulation 2	1 1 0 0 0 0 1 0 1 0	$\pi = 0.3$
Simulation 3	0 0 1 1 0 1 0 1 0 1	$\pi = 0.5$
Simulation 4	1 1 1 0 0 1 1 0 1 1	$\pi = 0.7$
Simulation 5	0 1 1 1 1 1 1 0 1 1	$\pi = 0.9$

Choosing the value of a parameter most likely to have given rise to the data is called *estimation*.

The range of parameter values which could have given rise to the data is called the *supported range* or *confidence interval*. Wide intervals mean the parameter has been estimated with low precision.

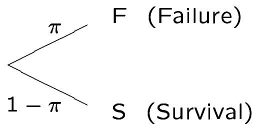
Random mechanisms for generating data are called statistical models. They depend on one or more unknown *paramaters* such as π.

In epidemiology there are three statistical models which are widely used:

- The binary model
- The survival time model
- The Gaussian metric model

This model can arise in many different situations: a five year follow–up study of failure or survival; a prevalence study in which subjects either have or do not have some characteristic; a case–control study in which cases of a disease are either exposed or unexposed.

We shall use a follow–up study for a fixed period of 5 years to illustrate the model.



The parameter π is called the *risk* of failure.

For example, if $\pi = 0.2$ then the probability of failure during the 5 years of follow-up is 0.2.

The odds of failure

The *odds* of failure vs survival are

$$\pi : (1-\pi) \quad \text{or} \quad \frac{\pi}{1-\pi} : 1$$

For example, when the risk of failure is 0.2, the odds of failure are

$$0.2 : 0.8 = 0.25 : 1$$

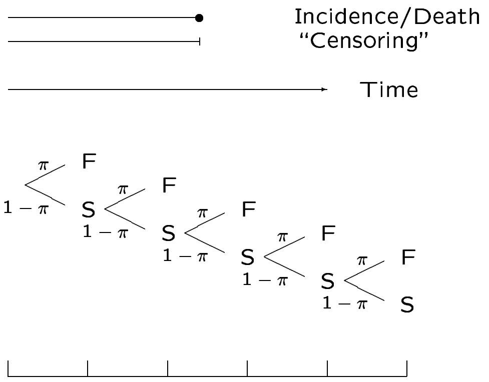
An odds of 0.25 : 1 is interpreted as 25 failures for every 100 survivors. Often the :1 is left out.

Exercise: what does it mean if the odds of failure are 1:1?

The odds parameter is denoted by Ω where

$$\Omega = \frac{\pi}{1-\pi} \quad \text{and} \quad \pi = \frac{\Omega}{1+\Omega}$$

The survival time model



One way of thinking about time-to-event data is as a series of binary trials each of which asks whether the subject has failed. If not the subject proceeds to the next trial, until finally he fails or the follow-up is censored.

The trials correspond to time bands. The probability for a subject who fails during the third band is $(1-\pi) \times (1-\pi) \times \pi$, which is the same as from 3 subjects each followed for one band. The important unit in a follow-up study is the subject×band.

Consider bands of time of length 1 month, so π is the risk of failure during the next month.

Suppose we know that $\pi = 0.005$. To simulate a survival time, generate random numbers until one which is less than 0.005 occurs.

Simulation 1: all the random numbers were > 0.005 until the 204th, which was 0.00099. Survival time is 204 months, or 17.00 years.

Simulation 2: all the random numbers were > 0.005 until the 83rd, which was 0.00227. Survival time is 83 months, or 6.92 years.

Generally a follow-up study will be halted (or analysed) after (say) 10 years, so all subjects who survive more than 10 years are *censored*.

Simulation 1: censored at 10 years.

Simulation 2: failed at 6.92 years.

The probability π now refers to the probability of failure during the next band of time, given that the subject has not yet failed.

Let the length of a band be h where h is measured in years but very very small (eg .0027 years which is one day).

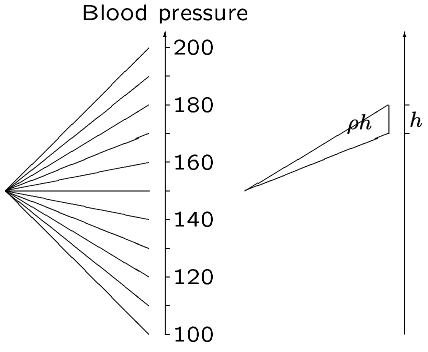
The probability π will now be very small, so instead of using π as the parameter we use the *rate* which is the probability per unit time.

$$\lambda = \pi/h \quad , \quad \pi = \lambda h$$

The rate is the probability per unit time for a *small* time band.

Models for a metric response

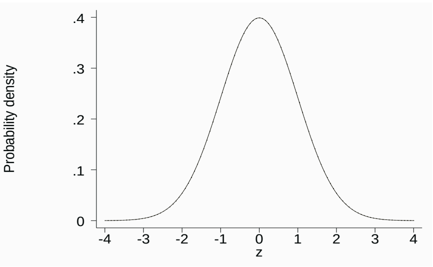
The probability is shared between many outcomes



If ρ is the probability density of values in a band of width h units, the probability of observing a value in the band is $\pi = \rho h$.

For roughly symmetric distributions the Gaussian model for the probability density is used.

The Gaussian distribution



Probability density at z is

$$0.3989 \exp\left[-\frac{1}{2}(z)^2\right]$$

If z has a standard Gaussian distribution then

$$x = \mu + \sigma z.$$

has a general Gaussian distribution. The parameter μ is called the *mean*, and the parameter σ is called *standard deviation*.

To simulate observations from a standard Gaussian model, there is a special command in Stata. Here are 20 observations (z) from a standard Gaussian distribution.:

-1.09 0.37 0.15 0.27 0.48 -1.23 0.30 -1.55 0.14 1.13

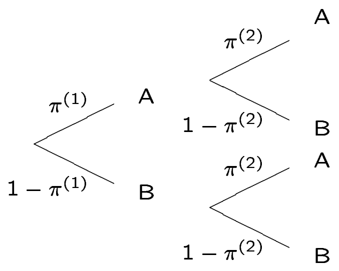
-0.66 -1.70 -0.19 1.27 0.05 1.00 -0.80 0.16 -0.63 1.62

To generate random numbers from a Gaussian distribution with mean 100 and standard deviation 15 (say) use

$$x = 100 + 15z$$

Eg, the first would be $100 + 15 \times -1.09 = 83.6$.

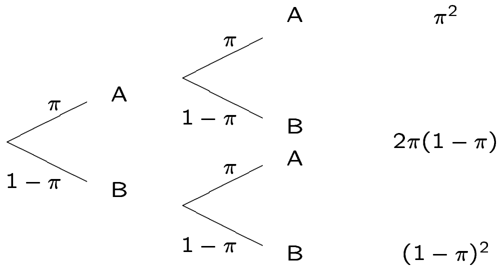
Suppose each subject is observed on two separate occasions, and that the outcome on the second occasion is completely independent of the outcome on the first occasion. The binary model now extends to



$\pi^{(1)}$ is the probability of A on the first occasion, $\pi^{(2)}$ is the probability of A on the second.

Constant probability

When the probability of A is constant at π the model becomes:



The total number of events of type A for a subject is 2, 1, or 0. If $\pi = 0.1$, then

$$\Pr(2) = 0.01,\ \Pr(1) = 0.18,\ \Pr(0) = 0.81$$

The expected number of events of type A per subject is

$$0.01 \times 2 + 0.18 \times 1 + 0.81 \times 0 = 0.20$$

The binomial distribution

For n trials per subject with constant probability π of an event of type A, the expected number of events is $n\pi$.

The actual number of events varies according to a *Binomial* distribution with

$$\begin{aligned} \text{Mean} &= n\pi \\ \text{Variance} &= n\pi(1-\pi) \end{aligned}$$

For example, when $n = 10, \pi = 0.2$, the distribution of the number of events is

```
     Number |    Percent
------------+-----------
          0 |     10.90
          1 |     27.51
          2 |     30.15
          3 |     19.76
          4 |      8.74      mean = 2.00
          5 |      2.39
          6 |      0.45       var = 1.60
          7 |      0.07
          8 |      0.02       std = 1.26
          9 |      0.01
------------+-----------
      Total |    100.00
```

The time-to-event is now time-to-next-event.

The rate parameter is still the probability of an event in the next small band of time, expressed per unit time, but it is no longer conditional on not yet having experienced an event. It is sometimes called the *intensity* parameter.

To simulate the total number of events in a fixed time, keep simulating the next event until the time is up.

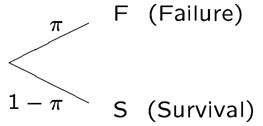
For rate λ and time y the total number of events for a subject will vary according to a *Poisson* distribution with

$$\begin{aligned} \text{mean} &= \lambda y \\ \text{Variance} &= \lambda y. \end{aligned}$$

Erasmus, 2002

The binary probability model

Models are used to calculate the proabability of the data for given values of the parameters. To illustrate this we shall use a binary model with outcome F if the subject fails during a fixed follow-up period, and S if the subject survives the period.



The probability of failure is denoted by π, and is called the *risk* parameter.

Definition

Likelihood is the probability of the observed data given the probability model which gave rise to these data.

It is used to compare candidate values for the parameters of the model; the greater the probability of the observed data, the more *likely* the parameter value.

The data: F F S F S S S F S S

The model: Prob(F) $= \pi$, Prob(S) $= 1 - \pi$

Contribution to the likelihood is

$\times\pi$ for a subject who fails
$\times(1-\pi)$ for a subject who survives

Total likelihood for π is $\pi^4(1-\pi)^6$, called the binomial or Bernoulli likelihood.

The likelihood for $\pi = 0.1$ is
$(0.1)^4 \times (0.9)^6 = 0.0000531$

The likelihood for $\pi = 0.4$ is
$(0.4)^4 \times (0.6)^6 = 0.0011944$

Thus $\pi = 0.4$ is more likely to have given rise to the data than $\pi = 0.1$.

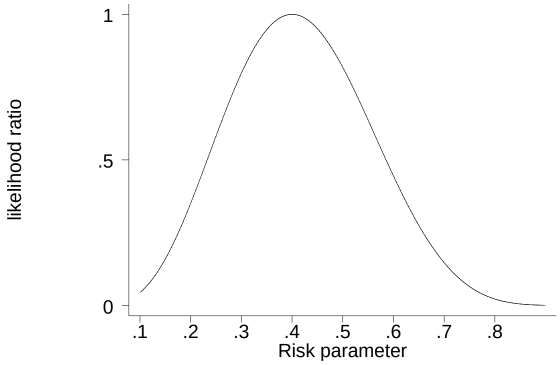
The likelihood for $\pi = 0.1$ is not, on its own, useful. It is only useful as a way to compare $\pi = 0.1$ with other values of π.

In this example $\pi = 0.4$ is the value most likely to have given rise to the data. The likelihood for $\pi = 0.4$ is larger than for any other value of π. All other values of π are compared with the most likely value using the likelihood ratio

$$LR(\pi) = \frac{L(\pi)}{L(M)}$$

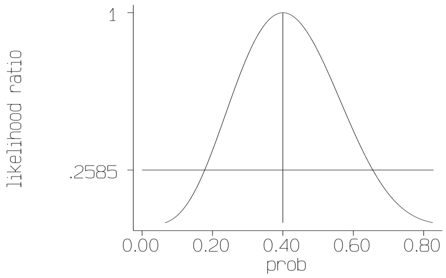
where M is the most likely value.

π	$L(\pi)$	$L(\pi)/L(M)$
0.1	5.31×10^{-5}	0.0447
0.4	119.44×10^{-5}	1.0000
0.5	97.66×10^{-5}	0.8178
0.9	0.06×10^{-5}	0.0005



Critical value and supported range

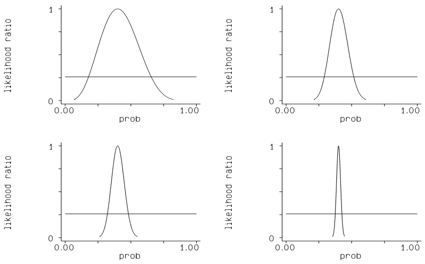
Choose a cut-point (*critical value*), such as 0.2585, and decide that values of π for which LR > 0.2585 are "supported" by the data, while values of π for which LR < 0.2585 are not.



The supported range for π consists of all values supported by the data. Here it is from 0.17 to 0.65. The width of the supported range reflects the precision with which the parameter is estimated.

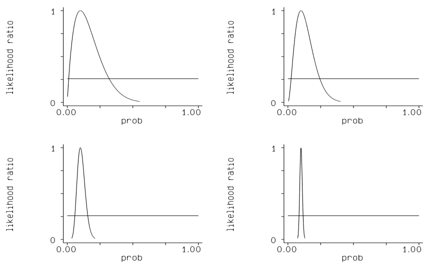
Increasing the size of a study (1)

4 : 6 20 : 30



40 : 60 400 : 600

1 : 9 2 : 18



10 : 90 100 : 900

The contribution of a subject to the log likelihood is

log(π) for a failure
log($1 - \pi$) for a survival

This can be written as

$$d \log(\pi) + (1 - d) \log(1 - \pi)$$

where $d = 1$ for a failure and $d = 0$ for a survival.

The total log lik for N subjects of whom D are failures is

$$\sum [d \log(\pi) + (1 - d) \log(1 - \pi)]$$

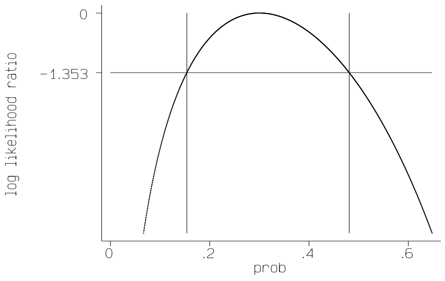
which equals

$$D \log(\pi) + (N - D) \log(1 - \pi)$$

For the 10 subjects with 4 failures the total log lik is

$$4 \log(\pi) + 6 \log(1 - \pi)$$

Log likelihood ratios



$$\text{LR}(\pi) = \frac{\text{L}(\pi)}{\text{L}(M)}$$

$$\text{LLR}(\pi) = \text{LL}(\pi) - \text{LL}(M)$$

Values of π for which LLR > -1.353 are supported by the data ($\log(0.2585) = -1.353$).

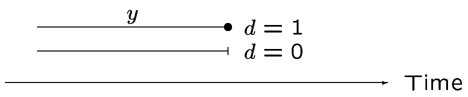
Values of π for which LLR < -1.353 are not supported by the data.

The log likelihood for a rate

A subject experiences a constant rate λ, and is followed for y years after which he either fails ($d = 1$ or his follow-up is censored (d=0).

What is the contribution to the log likelihood?

Divide the follow-up time y into N small bands of duration h so $y = Nh$, where N is very large.



The first N trials end in S; each contributes $\log(1 - \pi)$ to the log likelihood, so the total contribution is

$$N \log(1 - \pi) \approx -N\pi = -N\lambda h = -\lambda y$$

The final band contributes $\log(1-\pi)$ when the follow-up is censored and $\log(\pi)$ when the subject fails. Because $\log(1-\pi)$ is very small when N is very large it can be ignored, so the final band only makes a contribution when it ends in failure. The contribution is

$$\log(\pi) = \log(\lambda) + \log(h) = \log(\lambda) + \text{constant}$$

Ignoring the constant, the total log lik is

$-\lambda y$ when the follow-up is censored,

and

$-\lambda y + \log(\lambda)$ when the subject fails.

This can be written in the form

$$-\lambda y + d \log(\lambda)$$

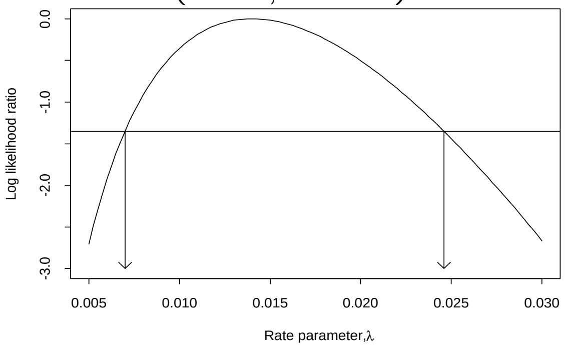
where $d = 1$ for a failure and $d = 0$ when the follow-up is censored, and is known as the Poisson or exponential log likelihood.

Total log–likelihood is

$$\sum [d \log(\lambda) - \lambda y] = D \log(\lambda) - \lambda Y$$

where D is the total number of cases and Y is the total observation time (*person–years*)

$(D = 7, Y = 500)$

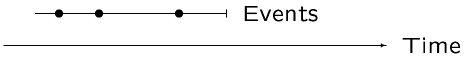


The most likely value of λ is $D/Y = 0.014$

The supported range is from 7.0/1000 to 24.6/1000.

Recurrent events

- For one subject:



- Log likelihood contribution for λ is still

$$d \log \lambda - \lambda y$$

  - d is the number of events observed
  - y is observation time

- For a homogeneous population we have

$$D \log \lambda - \lambda Y$$

where $D = \sum d, Y = \sum y$

- But note that the assumption of homogeneity (λ constant among subjects) is much more serious in this case

Information and Quadratic Approximations

Erasmus, 2002

The response

In a Gaussian model, the log likelihood for μ from a single observation x is

$$-\frac{1}{2}\left(\frac{\mu - x}{\sigma}\right)^2$$

where μ is the mean and σ the standard deviation of the Gaussian model.

For n observations the total log likelihood is

$$-\frac{1}{2}\left(\frac{\mu - \bar{x}}{S}\right)^2$$

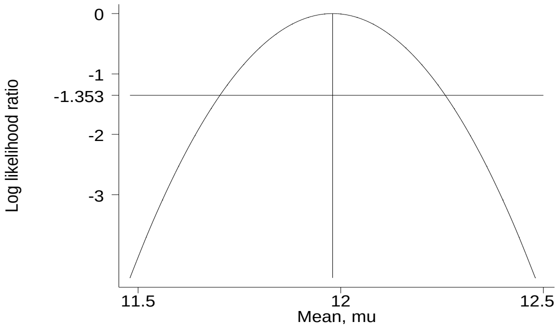
where $\bar{x}$ is mean response, and $S = \sigma/\sqrt{n}$ is the *standard error*. The most likely value of μ is $M = \bar{x}$.

The supported range for μ is found from

$$\begin{aligned} -\tfrac{1}{2}\left(\tfrac{\mu-\bar{x}}{S}\right)^2 &= -1.353 \\ \left(\tfrac{\mu-\bar{x}}{S}\right)^2 &= 2.706 \\ \mu = \bar{x} - 1.645S \quad \text{to} &\quad \mu = \bar{x} + 1.645S \end{aligned}$$

8.8	9.3	9.4	9.7	10.2	10.2	10.3	10.4	10.4	10.5
10.6	10.6	10.7	10.8	10.8	10.9	10.9	10.9	11.0	11.0
11.1	11.1	11.2	11.2	11.3	11.4	11.4	11.4	11.5	11.6
11.6	11.7	11.7	11.8	11.8	11.9	11.9	12.0	12.0	12.1
12.1	12.1	12.2	12.3	12.5	12.5	12.7	12.9	12.9	12.9
12.9	13.0	13.1	13.1	13.2	13.3	13.3	13.4	13.4	13.5
13.5	13.6	13.7	13.7	14.1	14.6	14.6	14.7	14.9	15.1

$$\bar{x} = 11.98, \quad \sigma = 1.416, \quad S = 1.416/\sqrt{70} = 0.169$$



$$\text{LLR} = -\frac{1}{2}\left(\frac{\mu - 11.98}{0.169}\right)^2$$

The supported range is from
$11.98 - 1.645 \times 0.169 = 11.70$ to
$11.98 + 1.645 \times 0.169 = 12.26$.

Score and information

For the Gaussian model the log likelihood is

$$-\frac{1}{2}\left(\frac{\mu - M}{S}\right)^2$$

where M is the mean response and $S = \sigma/\sqrt{n}$ is the standard error.

The score function is the gradient of the log likelihood:

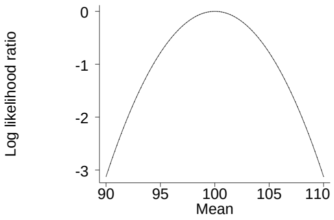
$$-\frac{(\mu - M)}{(S)^2}$$

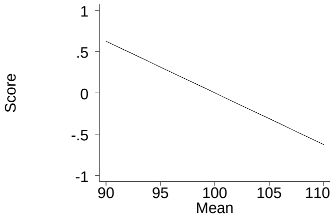
The information function is minus the gradient of the score function:

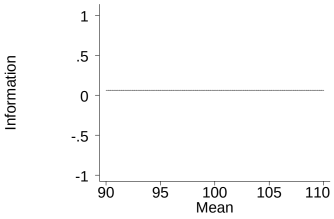
$$\frac{1}{(S)^2}$$

which does not vary with μ.

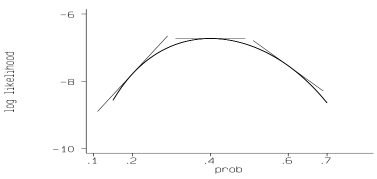
Gaussian model with $\mu = 100, S = 4$



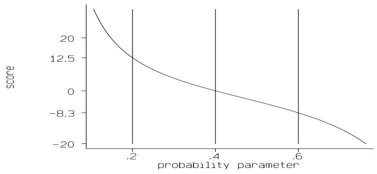




A binary response $(D = 4, N = 10)$



$$D \log(\pi) + (N - D) \log(1 - \pi)$$

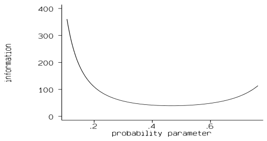


$$\frac{D}{\pi} - \frac{N-D}{1-\pi}$$

- Gradient of score curve describes *curvature* of log likelihood curve

Information

- Minus the gradient of the score function is the information function:



$$\frac{D}{(\pi)^2} + \frac{N-D}{(1-\pi)^2}$$

- Here this does not vary by very much around the most likely value (0.4) of π.

- This is true in general and justifies approximating log likelihoods by *quadratic functions* — which have constant information.

Quadratic approximations

- A quadratic approximation to the log likelihood function for a parameter β is
$$-\frac{1}{2}\left(\frac{\beta - M}{S}\right)^2$$

- M is the maximum likelihood estimate and S is its *standard error*

- Score function: $-\frac{(\beta-M)}{(S)^2}$

- Information: $\frac{1}{(S)^2}$

- Choose M, S to fit the curve around the maximum likelihood estimate:
$$S = \sqrt{1/(\text{Information at M})}$$

Approximate supported range

- For an approximate supported range corresponding to a log likelihood ratio of -1.353,
$$-\frac{1}{2}\left(\frac{\beta - M}{S}\right)^2 > -1.353$$

- The limits of the range solve the equation
$$\left(\frac{\beta - M}{S}\right)^2 = 2.706$$

- This corresponds to $\beta = M \pm 1.645S$

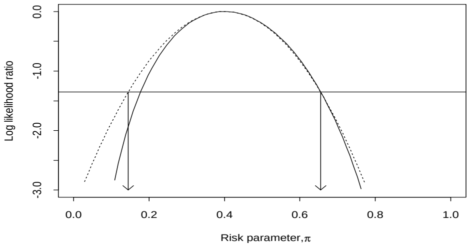
- The interval is symmetric around M

Binomial log likelihood

- $\beta = \pi$ and $M = D/N$

- Information at $\pi = M$ is
$$\frac{D}{(M)^2} + \frac{N-D}{(1-M)^2} = \frac{N}{M(1-M)}$$
so that $S = \sqrt{M(1-M)/N}$

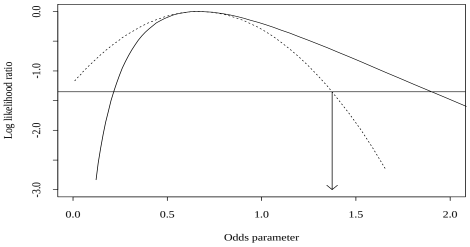
- For $D = 4, N = 10$:



Transformations of parameters

- For the *odds* parametrization
$\beta = \Omega = \pi/(1-\pi)$

- $M = D/(N-D), S = (M+1)\sqrt{M/N}$



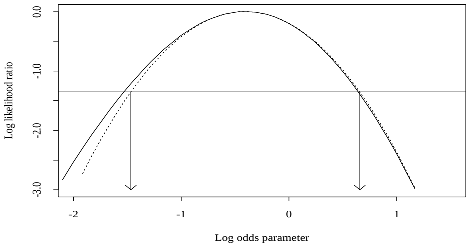
- The left hand limit is < 0 — impossible!

- The approximation is bad because the log likelihood curve is very asymmetric with respect to Ω

The logit transformation

- The asymmetry arises because the value of Ω *must* be > 0

- If we take the *log* of Ω,
$$M = \log\frac{D}{N-D}, \quad S = \sqrt{\frac{1}{D} + \frac{1}{N-D}}$$

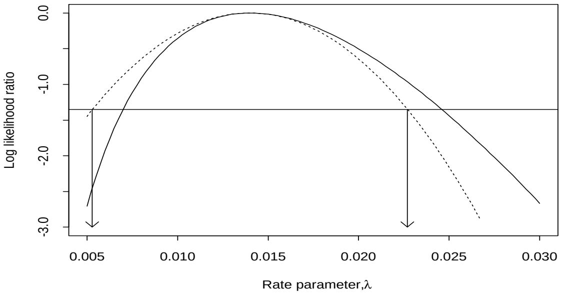


- The log odds parameter is related to the risk parameter, π by the *logit* transformation:
$$\log \text{Odds} = \log\frac{\pi}{1-\pi}$$

Poisson log likelihood

- For the rate parameter, λ,
$$M = \frac{D}{Y}, \quad S = \frac{\sqrt{D}}{Y}$$

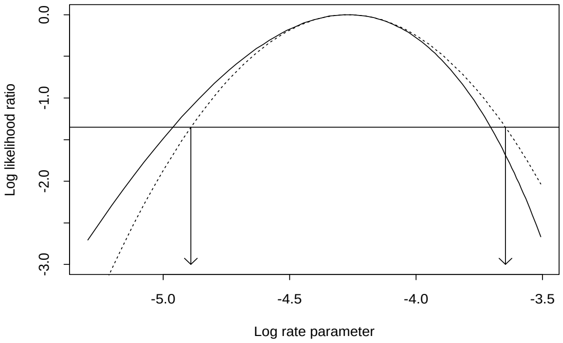
- For example, $D = 7, Y = 500$



- Since λ *must* be > 0, this is often not a good approximation

- For $\log \lambda$,

$$M = \log \frac{D}{Y}, \quad S = \sqrt{\frac{1}{D}}$$



- The confidence interval for $\log(\lambda)$ is

$$\log(7/500) \pm 1.645S.$$

- The confidence interval for λ is

$$7/500 \overset{\times}{\div} \exp(1.645S)$$

Erasmus, 2002

Questions

Most questions in statistical analysis take the form of asking whether the value which one variable takes for a given subject depends on the value taken by another variable.

Does the birth weight of a baby depends on whether it is a boy?

The variable which is of primary interest is called the *response* variable; the variable on which the response variable may depend is called the *explanatory* variable.

Here the response variable is birth weight and the explanatory variable is whether the baby is a boy.

Response variables

In biostatistics four types of response are particularly common:

1. Binary (two values coded 0/1)
2. Metric (a measurement with units)
3. Failure (does the subject fail at end of follow-up)
4. Count (aggregated failure data)

Frequency tables don't answer the question unless you say what is the response.

```
No response specified:

 low birth |       hypertens
    weight |         0          1 |     Total
-----------+----------------------+----------
         0 |       388         52 |       440
         1 |        40         20 |        60
-----------+----------------------+----------
     Total |       428         72 |       500
```

```
lowbw is the response:

 low birth |       hypertens
    weight |         0          1 |     Total
-----------+----------------------+----------
         0 |       388         52 |       440
           |     90.65      72.22 |     88.00
-----------+----------------------+----------
         1 |        40         20 |        60
           |      9.35      27.78 |     12.00
-----------+----------------------+----------
     Total |       428         72 |       500
           |    100.00     100.00 |    100.00

hyp is the response:

 low birth |       hypertens
    weight |         0          1 |     Total
-----------+----------------------+----------
         0 |       388         52 |       440
           |     88.18      11.82 |    100.00
-----------+----------------------+----------
         1 |        40         20 |        60
           |     66.67      33.33 |    100.00
-----------+----------------------+----------
     Total |       428         72 |       500
           |     85.60      14.40 |    100.00
```

Summaries

The type of response determines how it will be summarized.

1. A binary response is usually summarized using the proportion of 1's

2. A metric response is usually summarized using its mean or median

3. A failure response is usually summarized using the rate of failures.

Metric response

```
. tabmenu1, clear
---> select bweight as response
---> select metric as type
---> select hyp as rows
     ---> select Means as summary
     ---> click on OK

Response variable is: bweight which is metric
Row variable is: hyp
Number of records used:  500

Summary using means
----------------------
hypertens |    bweight
----------+-----------
        0 |    3198.90
        1 |    2768.21
----------------------
```

```
. tabmenu1, clear
---> select lowbw as response
---> select binary as type
---> select hyp as rows
---> click on Tables
     ---> select Proportions as summary
     ---> click on OK

Response variable is: lowbw which is binary
Row variable is: hyp
Number of records used:  500

Summary using proportions per 100
----------------------
hypertens |      lowbw
----------+-----------
        0 |       9.35
        1 |      27.78
----------------------
```

Too many values for gestation time to make a table, so we group with

```
. egen gest4=cut(getswks), at(20,35,37,39,45)
. table gest4

    gest4 |      Freq.
----------+-----------
       20 |         31
       35 |         32
       37 |        167
       39 |        260
```

Does birth weight depend on gest4

```
. tabmenu1, clear
---> select bweight as response
---> select metric as type
---> select gest4 as rows
     ---> select Means as summary
     ---> click on OK
Response variable is: bweight which is metric
Row variable is: gest4
Number of records used:    490

Summary using means

--------------------------------
    gest4 |              bweight
----------+---------------------
       20 |              1733.74
       35 |              2590.31
       37 |              3093.77
       39 |              3401.26
--------------------------------
```

A second explanatory variable

```
. tabmenu1, clear
---> select bweight as response
---> select type as metric
---> select hyp as rows
---> select sex as columns
---> click on Tables
     ---> select Mean
     ---> click on OK

Response variable is: bweight which is metric
Row variable is: hyp
Column variable is: sex
Number of records used:    500

Summary using means

          |             sex of baby
hypertens |          1                2
----------+----------------------------
        0 |    3310.75          3079.50
        1 |    2814.40          2699.72
---------------------------------------
```

id	Subject identity number
chd	Failure: 1=chd, 0 otherwise
y	Time in study (years)
doe	Date of entry
dox	Date of exit
dob	Date of birth
job	Occupation
month	month of survey
energy	Total energy (kcals per day)
height	Height (cm)
weight	Weight (kg)
hieng	High energy

Some questions

1. Does failure depend on energy?
2. Does failure depend on energy in two groups, high and low?
3. Does failure depend on energy in three groups, high and medium and low?

```
. tabmenu1, clear
---> select chd as response
---> select failure as type
---> select y as follow-up time
---> select hieng as rows
---> click on Tables
     ---> select Rates per 1000
     ---> click on OK

Response variable is: chd which is failure
Follow-up time variable is: y
Row variable is: hieng
Number of records used:    337

Summary using rates per 1000

 Energy in |
two groups |                  chd
----------+--------------------
       low |                13.60
      high |                 7.07
```

p–values and confidence

Erasumus, 2002

Likelihood and probability

- Likelihood is the probability of the data for some value of the parameter
- It is not a probability *with respect to the parameter itself* — it measures “support”
- Although some argue that this theory is sufficient for scientific purposes, it is generally felt necessary to relate support to probability in some way
- Two schools:
  - Bayesian: aims to calculate probability that parameter takes certain values
  - Frequentist: estimates probabilities of repeating a result in simulated repetitions of the study

Bayesian statistics

- In order to make statements of the form

  > the probability that π lies between ... and ... is ...

  we must interpret probabilities as *subjective* "degrees of belief"

- Further, we assume that our beliefs about the possible values of π before we saw the data can be expressed by the *prior* probability distribution $f(\pi)$

- If the data gives a likelihood function, $L(\pi)$, our "posterior" distribution for π is given by Bayes theorem:

$$\text{Posterior probability} \propto L(\pi) \times f(\pi)$$

Frequentist statistics

- This rejects subjective theories of probability — probability is defined as the relative frequency of events in a long series of repetitions

- We can think of this as *calibrating* our methods. Imagine simulating our study many times, and make two sorts of calculations:
  - Significance: calibrates the support for a particular value of the parameter in terms of a *p–value*
  - Confidence: calibrates a way of calculating a support interval by measuring the proportion of repetitions in which the true value falls within the interval

Significance

- We observe 4 failures in 10 trials. Do these data support the proposition that $\pi = 0.3$?

- The log likelihood at $\pi = 0.3$ is

$$\text{LL}(0.3) = 4\log(0.3) + 6\log(0.7)$$

  while the maximized log likelihood is

$$\text{LL}(0.4) = 4\log(0.4) + 6\log(0.6)$$

- Support for $\pi = 0.3$ is measured by their difference, the log likelihood ratio

$$\text{LLR}(0.3) = LL(0.3) - LL(0.4) = -0.2258$$

- We can calibrate this in terms of a p–value by asking ourselves what is the probability of getting a value less than or equal to this if π *really was* 0.3

Repeated studies

- We simulate the study with $\pi = 0.3$. Each repetition will give a different number of failures, and therefore a different log likelihood curve

- For example, the first 10 simulations might give

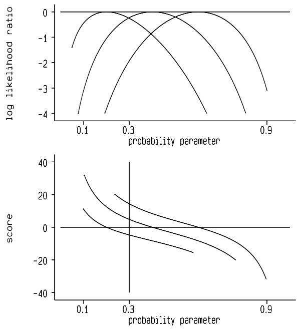
Repetition	Failures	Survivors	LLR(0.3)
1	2	8	−0.2573
2	4	6	−0.2258
3	2	8	−0.2573
4	5	5	−0.8718
5	2	8	−0.2573
6	3	7	0.0000
7	4	6	−0.2258
8	2	8	−0.2573
9	3	7	0.0000
10	4	6	−0.2258

- The p–value is the proportion of studies in which the LLR is less than (or equal to) our observed value (−0.2258)

- Here we'd estimate the p–value to be 8/10 = 0.8
- Small p–values indicate poor support for the hypothetical value under test, and 0.8 is not small!
- Of course, we really need many more repetitions to estimate the p–value reliably
- Sometimes we can use theory to calculate the p–value. This is one such case — the probabilities are given by the *binomial distribution*
- Otherwise we use computer simulation — known as *Monte Carlo testing*

The score in repeated studies

- Three repetitions, with $\pi = 0.3$:



- In repeated simulations, the value of the score at the true value of the parameter, $U(\pi)$, has expected value (mean) zero

- Its *variance* is given by the information at this point, $I(\pi)$
- or, if this varies from simulation to simulation, by its expected value (mean)
- For example, 10,000 simulated repetitions of our study ($N = 10, \pi = 0.3$) gave

D	$N - D$	LLR(0.3)	$U(0.3)$	$I(0.3)$	Freq
0	10	−3.567	−14.286	20.408	273
1	9	−1.163	−9.524	29.478	1221
2	8	−0.257	−4.762	38.549	2350
3	7	0	0	47.619	2687
4	6	−0.226	4.762	56.689	1993
5	5	−0.872	9.524	65.760	1033
6	4	−1.920	14.286	74.830	328
7	3	−3.389	19.048	83.900	94
8	2	−5.431	23.810	92.971	20
9	1	−5.341	28.571	102.041	1
10	0	−7.942	33.333	111.111	0
		Mean	−0.041	47.541	
		Variance	47.187		

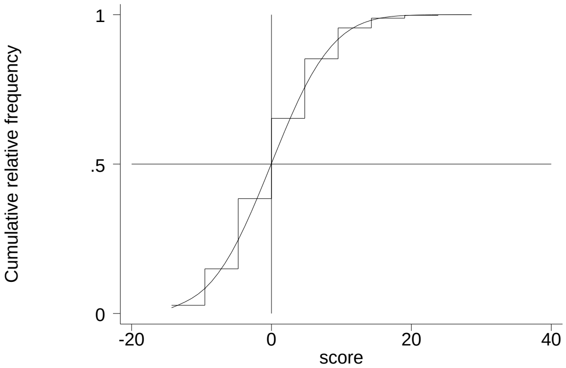
Frequency theory in large studies

- *In large studies*, the distribution of $U(\pi)$ is Gaussian (Normal), so that $U(\pi)/\sqrt{I(\pi)}$ is approximately a *standard normal deviate*
- It follows that $U^2(\pi)/I(\pi)$ is approximately distributed as chi-squared on 1 degree of freedom.
- When the log likelihood is approximately quadratic,

$$-2\text{LLR}(\pi) \approx U^2(\pi)/I(\pi)$$

  so that $-2\text{LLR}(\pi)$ is also approximately chi–squared (1 df).

The simulations when $\pi = 0.3$ were used to generate 10 000 values of $U(0.3)$ and LLR(0.3). These can be used to verify that these large sample results hold true approximately, even when N is only 10.



The graph shows that the distribution of the score in 10 000 simulations is approximately Gaussian.

Confidence of support intervals

- In what proportion of repetitions would the support interval based on a LLR of -1.353 contain the true parameter value?
- For example, 10 simulated repetitions of our study with $N = 10$ and $\pi = 0.3$ gave the ranges

D	$N - D$	Low	High
2	8	.051	.448
4	6	.178	.655
2	8	.051	.448
5	5	.257	.743
2	8	.051	.448
2	8	.051	.448
3	7	.109	.558
4	6	.178	.655
2	8	.051	.448
3	7	.109	.558
4	6	.178	.655

In this case the confidence is 100%, but it would not be in a longer series!

Approximate confidence

- In general we cannot work out exact confidence levels, but we can use our earlier results to work out approximate confidence levels in large samples
- Note that π will fall within our supported range if

  $$\text{LLR}(\pi) > -1.353$$

  or, equivalently,

  $$-2\text{LLR}(\pi) < 2.706$$
- We have seen that, in repetitions of the study, $-2\text{LLR}(\pi)$ is approximately distributed as chi–squared (1 df)
- The probability that a chi–squared (1 df) variate is < 2.706 is 0.9
- Thus, support intervals based on the -1.353 cutoff have aproximately 90% confidence

Effects and Exposures

Erasmus, 2002

The main explanatory variable is now called the *exposure*.

An effect parameter is any contrast between the parameters which measure the level of response in different groups of subjects.

Subjects are classified as exposed (1) or unexposed (0),

For a binary response, parameter π, possible measures of effect are

1. $\pi_1 - \pi_0$
2. π_1/π_0
3. $\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \Omega_1/\Omega_0$

For a failure response, parameter λ, possible measures of effect are

$\lambda_1 - \lambda_0$

λ_1/λ_0.

For a metric response, parameter μ, possible measures of effect are

$\mu_1 - \mu_0$

μ_1/μ_0.

The scale should be chosen so that the exposure has roughly the same effect regardless of level, so that we can combine information about effects from different studies.

Ratios are more likely to be constant than the differences, but interpretation also matters.

The births data

Effect of hypertension on birth weight

```
hypertens |    bweight
----------+-----------
        0 |    3198.90
        1 |    2768.21
----------------------

Effects measured as differences in means

Effect(s) of hyp

Level      Effect    95% CI

2 vs 1  -430.6959   [ -585.444 , -275.948 ]

Test for no effects

F( 1 ,   498)   =  29.757
P-value         =   0.000
```

The births data continued

The effect of hypertension on low birth weight

```
Summary using odds
hypertens |      lowbw
----------+-----------
        0 |       0.10
        1 |       0.38

Effects measured as odds ratios

Level        Effect    95% CI
2 vs 1       3.7308   [ 2.027 , 6.865 ]

Test for no effects

chi2( 1)    =  17.906
P-value     =   0.000
```

The effect of high energy on CHD

```
Summary using rates per 1000

    hieng |        chd
----------+-----------
        0 |      13.60
        1 |       7.07
----------------------
Effects measured as rate ratios

Effect(s) of hieng

Level         Effect    95% CI

2 vs 1        0.5204    [ 0.288 , 0.941 ]

Test for no effects

chi2(  1)    =    4.675
P-value      =    0.031
```

The diet data refer to a group of 337 men followed after recording a full weighed diet for two weeks. The main outcome of interest was CHD.

```
Estimated rates (per 1000) for two groups

    hieng     _D         _Y       _Rate
        0     28     2.0594      13.596
        1     18     2.5442       7.075

Rate ratio = 7.075/13.596 = 0.5203.
```

To find the SE we need a likelihood.

The log likelihood for λ_0 is

$$D_0 \log(\lambda_0) - \lambda_0 Y_0 = 28 \log(\lambda_0) - 2.0594\lambda_0$$

The log likelihood for λ_1 is

$$D_1 \log(\lambda_1) - \lambda_1 Y_1 = 18 \log(\lambda_1) - 2.5442\lambda_1$$

The log likelihood for λ_0 and λ_1 is the sum of these two log likelihoods, and the log likelihood ratio is obtained by subtracting the maximum which occurs at $\lambda_0 = 7.075$ and $\lambda_1 = 13.596$.

The rate ratio parameter

Let $\theta = \lambda_1/\lambda_0$. This is the parameter of interest; λ_0 is a nuisance parameter. To estimate the ML value of θ with a supported range we need to find the log likelihood ratio for different values of θ.

For any given value of λ_0, say 10/1000, and any given value of θ, say 0.4, we can find the corresponding value of λ_1. In this case

$$\lambda_1 = 0.4 \times 10/1000 = 4/1000.$$

Substituting $\lambda_0 = 10/1000$ and $\lambda_1 = 4/1000$ in the formula for the log likelihood ratio gives a value of -3.6

Doing this for many different values of $\lambda_)$ and θ gives the table below.

Table of LLR for different values of λ_0 & θ.

	θ					
λ_0	0.2	0.3	0.4	0.5	0.6	0.7
5.0	−30.1	−24.0	−20.1	−17.4	−15.4	−13.9
6.0	−24.3	−18.5	−14.8	−12.3	−10.6	−9.3
7.0	−19.7	−14.2	−10.8	−8.6	−7.1	−6.1
8.0	−16.2	−10.9	−7.7	−5.8	−4.5	−3.8
9.0	−13.3	−8.3	−5.4	−3.7	−2.7	−2.2
10.0	−11.0	−6.3	−3.6	−2.2	−1.4	−1.2
11.0	−9.2	−4.7	−2.3	−1.1	−0.6	−0.7
12.0	−7.8	−3.5	−1.4	−0.4	−0.2	$\boxed{-0.5}$
13.0	−6.7	−2.7	−0.8	−0.1	$\boxed{-0.1}$	−0.6
14.0	−5.8	−2.1	−0.5	$\boxed{-0.0}$	−0.3	−1.1
15.0	−5.2	−1.7	$\boxed{-0.4}$	−0.2	−0.7	−1.7
16.0	−4.8	−1.6	−0.5	−0.5	−1.3	−2.6
17.0	−4.6	$\boxed{-1.6}$	−0.8	−1.1	−2.1	−3.7
18.0	$\boxed{-4.5}$	−1.8	−1.2	−1.8	−3.1	−4.9
19.0	−4.6	−2.1	−1.8	−2.6	−4.2	−6.2
20.0	−4.8	−2.6	−2.5	−3.6	−5.4	−7.7

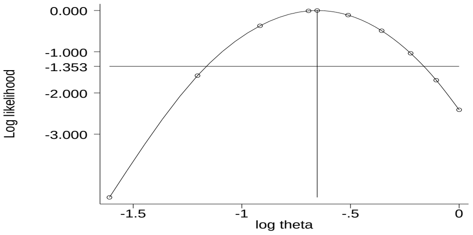
Choose a value for θ, say $\theta = 0.2$. The LLR for $\theta = 0.2$ varies from -30.1 to -4.8 according to the value of λ_0. The maximum value in the this column is -4.5, and this is the value we shall use. It is called the *profile log likelihood ratio* for $\theta = 0.2$. Similarly the profile LLR for $\theta = 0.3$ is -1.6, and so on for other values of θ.

The result of doing this for different values of θ is:

```
   theta    profile loglik
      .2        -4.533
      .3        -1.579
      .4        -0.371
      .5        -0.009
   .5203         0.000
      .6        -0.112
      .7        -0.491
      .8        -1.038
      .9        -1.691
       1        -2.409
```

Note that the maximum value of the profile LLR is at $\theta = 0.5203$.

The profile log likelihood ratio for θ is plotted against $\log(\theta)$ below.



For the diet data the the most likely value of θ is $\theta = 7.1/13.6 = 0.520$, and the standard error of $\log(\theta)$ in the quadratic approximation is

$$\sqrt{\frac{1}{28}+\frac{1}{18}} = 0.3021$$

The supported range for $\log(\theta)$ is

$$\log(0.520) \pm 1.645 \times 0.3021$$

which is from -1.151 to -0.157. The range for θ is from 0.32 to 0.85.

Metric exposures

Metric exposures can be

- Grouped and treated as categorical
- Left as metric
- Grouped and treated as metric

It is usual to start by grouping, but for small amounts of data there is some advantage to leaving exposure as metric.

Exposure grouped and treated as categorical

In the `diet` data `energy` is a metric exposure. We shall start by grouping the values of `energy` into 3 groups.

```
. egen eng3=cut(energy),at(1500,2500,3000,4500)
. tabmenu1

     eng3 |                  chd
----------+---------------------
     1500 |                16.90
     2500 |                10.91
     3000 |                 4.88

. effmenu1

Level         Effect 95% Confidence Interval

2 vs 1        0.6452 [ 0.339 , 1.229 ]
3 vs 1        0.2886 [ 0.124 , 0.674 ]

Test for no effects
chi2(  2)     =   8.241
P-value      =   0.016
```

To estimate the effect per unit of exposure requires the assumption that the efffect per unit increase in exposure is the same throughout the range of exposure values.

Eg the effect on birth weight of an incease in gestation of 1 week is the same for 25 to 26 weeks, 26 to 27 weeks, . . . , 42 to 43 weeks.

```
. effmenu1
---> select energy as exposure
---> check metric exposure
     ---> select per 100 units

Effect per 100 unit(s) of energy

Effect    95% Confidence Interval

0.8913      [ 0.830 , 0.957 ]

Test for no effects

chi2(  1)      =   9.970
P-value        =   0.002
```

This is half way between catergorical and metric. We group `energy` into `eng3` but declare `eng3` as metric.

```
. effmenu1
---> select eng3 as exposure
---> declare eng3 as metric
     ---> select per 100 units

Effect per 100 unit(s) of eng3

Effect    95% Confidence Interval

0.9326     [ 0.889 , 0.978 ]

Test for no effects

chi2(  1)      =   8.153
P-value        =   0.004
```

The test is now called the test for trend.

Testing Hypotheses

Erasmus, 2002

The log likelihood ratio test

Hypotheses are concerned with special values of the parameter. For example, if the parameter is a rate ratio, the value 1 is special. These special values are usually called null values.

Suppose 20 subjects are asked to choose between two treatments, A and B, and that 15 prefer A while 5 prefer B. When π is the probabiliy of preferring A the null value of π is 0.5 (no preference).

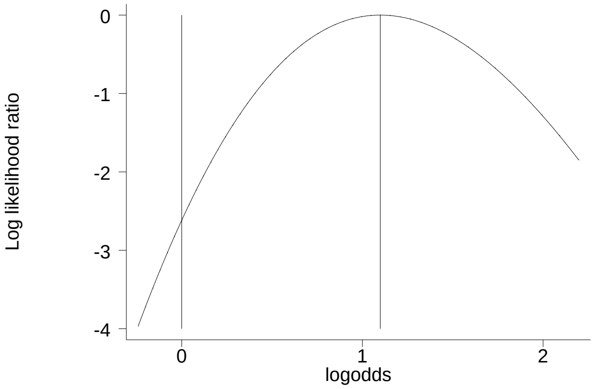
$$\text{LLR}(\pi) = 15\log(\pi) + 5\log(1-\pi)$$

$\text{LLR}(0.5) = -2.616, \quad -2\times\text{LLR}(0.5) = 5.232.$

The probability of exceeding 5.232 in a chi-squared distribution is 0.022 so $p(0.1) \approx 0.022$ and there is strong evidence against the null value $\pi = 0.5$.

This is called a (log) likelihood ratio test.

The log likelihood ratio is plotted against the logodds.



On the logodds scale the null value is $\log(.5/.5) = 0$ and the most likely value is $M = \log(15/5) = 1.099$.

The quadratic approximation with standard error S can be used in place of the LLR, so that

$$-2\text{LLR}(\pi) \approx \left(\frac{\pi - M}{S}\right)^2$$

is referred to the chi–squared distribution to get the p–value. It is usually best to do this on a scale without constraints. In this case we shall use the logodds scale and write $\beta = \log(\Omega)$.

The null value of β is 0, and

$$\begin{aligned} M &= \log(15/5) = 1.099 \\ S &= \sqrt{1/15 + 1/5} = 0.5164 \end{aligned}$$

Thus

$$-2\text{LLR}(\pi) \approx \left(\frac{0 - M}{S}\right)^2 = \left(\frac{0 - 1.099}{0.5164}\right)^2 = 4.529$$

The probability of exceeding 4.529 in a chi–squared distribution is 0., so $p(0.5) \approx 0.033$. This is called a Wald test.

Score test

The score test is also based on an approximation to the LLR, but this time it is a quadratic which has the same gradient and curvature at the null value, not at the most likely value. In general the log likelihood for π is

$$D\log(\pi) + (N - D)\log(1 - \pi)$$

Let π_{H} be the null value of π and let β_{H} be the corresponding null value of the log odds.

Using the log odds scale, the gradient at β_{H} is

$$U = D - N\pi_{\text{H}}$$

and the curvature is

$$V = N\pi_{\text{H}}(1 - \pi_{\text{H}})$$

Continued

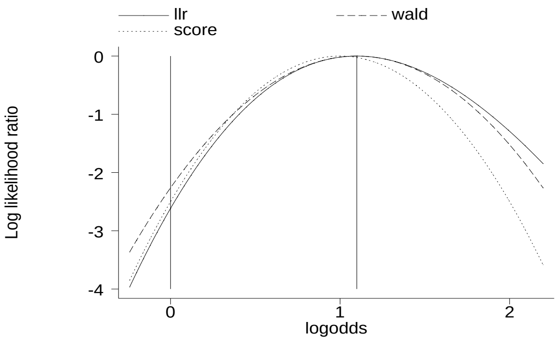
The quadratic approximation used in the score test is

$$-V(\beta - \beta_{\text{H}} - U/V)^2/2$$

The approximate LLR from this curve is $-0.5U^2/V$, so -2LLR is U^2/V.

In this example

$$\begin{aligned} U &= 15 - 20 \times 0.5 = 5 \\ V &= 20 \times 0.5 \times 0.5 = 5 \\ (U)^2/V &= 25/5 = 5 \\ p &= \text{chiprob}(1, 5) = 0.025 \end{aligned}$$



Note that the LLR and the Wald approximation both peak at log(15/5) = 1.099 but the score approximation peaks at

$$U/V = 5/5 = 1.$$

The chi-squared distribution on 1 degree of freedom is the distribution of the square of a standard Gaussian distribution.

The probability of exceeding 5 in a chi-squared distribution on 1 df is the same as the probability of being below $-\sqrt{5}$ or above $+\sqrt{5}$ in a standard Gaussian distribution. Using the chi-squared is more convenient than using the Gaussian.

The sum of two independent standard Gaussian distributions has a chi-squared distribution on 2 degrees of freedom.

When testing a hypothesis that two parameters both take their null values the LLR, or an approximation, is referred to chi-squared on 2df.

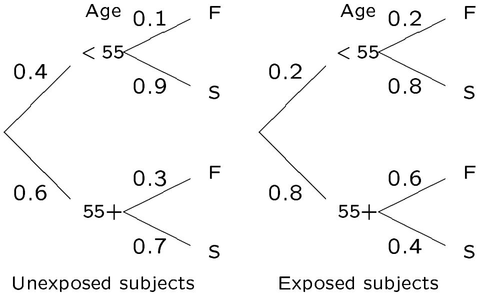
The idea extends to many parameters.

Confounding, standardization, and effect modification

Erasmus, 2002

Observational studies

- Epidemiology relies on *observational studies* of *experiments of nature*
- Often these are poor experiments – there is no control for extraneous influences which may influence exposure.
- Extraneous influences may *confound* the effect of exposure, in which case ignoring the extraneous influences will produce a biased estimate of exposure.
- A confounder is a variable whose influence we would have controlled if we had been able to design the natural experiment.



- The effect of exposure (as measured by the risk ratio) is 0.2/0.1 = 2 for the younger age group and 0.6/0.3 = 2 for the older age group.

- The overall, or *marginal*, probability of failure, for unexposed subjects is

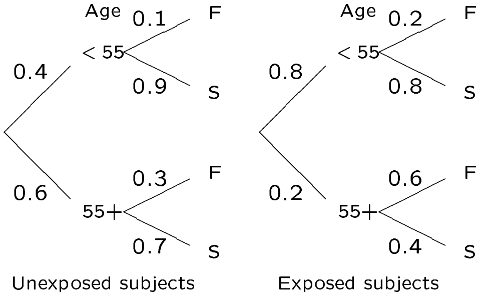
$$(0.4 \times 0.1) + (0.6 \times 0.3) = 0.22$$

- Similarly, marginal risk for exposed subjects is

$$(0.2 \times 0.2) + (0.8 \times 0.6) = 0.52$$

- The ratio of marginal risks is 0.52/0.22 = 2.36

- The exposed group was older, we over-estimates the effect of exposure.if we ignore age in the analysis



- The risk ratio is 0.2/0.1 = 2 for the younger age group and 0.6/0.3 = 2 for the older age group.

- The marginal risk for unexposed subjects is

$$(0.4 \times 0.1) + (0.6 \times 0.3) = 0.22$$

- Similarly, the marginal risk for exposed subjects is

$$(0.8 \times 0.2) + (0.2 \times 0.6) = 0.28$$

- The ratio of marginal risks is 0.28/0.22 = 1.27

- The exposed group was younger, so we underestimate the effect of exposure.if we ignore age in the analysis

Confounding

- These examples illustrate "confounding" in epidemiologucal studies.

- For a variable, V to confound the effect of exposure:

  - V must be associated with outcome (risk); *e.g.* older people have higher risk

  - V must be associated with exposure; *e.g.* older persons are more/less likely to be exposed

- However, this is not a sufficient definition. It must also make sense to envisage a "quasi-experiment" in which exposure is varied and V controlled in some way —- V should not be an *intermediate variable* on the causal path from exposure to response

Some examples to think about

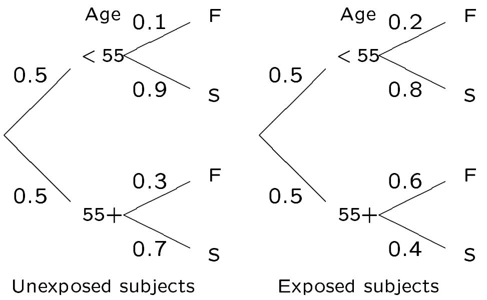
- Is birthweight a confounder of the relationship between aspects of ante-natal care and perinatal mortality?

- Does total calory intake confound the relationship between dietary fat intake and incidence of coronary heart disease or breast cancer?

variables

- In **real experiments** there are two different ways in which we may control for extraneous influences :
  1. Hold them *constant*,
  2. *Randomize* subjects to experimental groups so that the *distributions* of extraneous variables are the same.
- For the statistical control for confounding in the analysis of **observational studies**, there are likewise two approaches:
  1. Stratification, which simulates the first experimental method — holding extraneous variables constant.
  2. (Direct) standardization, which simulates the second experimental method — equalising the *distribution* of extraneous variables,

1. Estimate *age-specific* risks (or rates) in each group,
2. Calculate what marginal risks (rates) would be if the age distribution were fixed to that of some agreed *standard population*.



Marginal probability of failure is now

$(0.5 \times 0.1) + (0.5 \times 0.3) = 0.2$ for unexposed

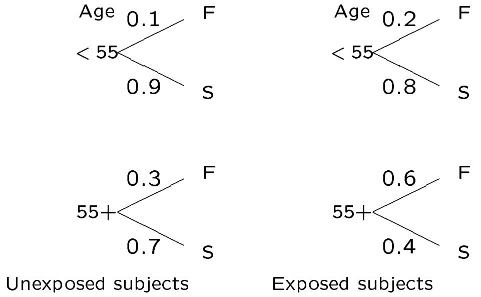
$(0.5 \times 0.2) + (0.5 \times 0.6) = 0.4$ for unexposed

The ratio of marginal risks is $0.4/0.2 = 2$.

Choice of weights

- Directly standardized risks or rates are simply weighted averages of the age-specific risks or rates
- Sometimes the overall age structure of the whole study is used
- To facilitate comparisons with other studies, sometimes a *standard* age distribution is used (*e.g.* that of the European or World population)
- Note, however, that this can be very prone to error if the standard population structure is very different from that in the our study — a small number of subjects may get very high weight

Controlling by stratification



- Hold age constant and compare the exposure groups *within age strata*.
- This leads to two studies, each of which allows estimation of the exposure effect. Here the risk ratio within both age groups is 2
- In real studies the sample size within strata will be small and the stratum–specific effects will not be reliably estimated

The model of constant effect over strata

- If the true effect of exposure varies across strata there is said to be "effect modification" — the effect of exposure within strata can then not be represented by a single number
- But, if the estimates differ only randomly, we can consider a model in which the true effect is constant. This allows us to combine the information from different strata to yield a single estimate of exposure effect
- We shall call this the estimate of effect "controlled for" the stratifying variable(s)
- Statistical tests for the presence of "effect modification" are available

trolled for job

```
Effect of hieng

Level
of job     Effect   95% CI

1          0.4103   [ 0.124 , 1.362 ]
2          0.6551   [ 0.227 , 1.888 ]
3          0.5177   [ 0.212 , 1.267 ]
```

These are sufficiently close to be combined into a single estimate.

```
Effect of hieng controlled for job

Level        Effect    95% CI

2 vs 1       0.5248   [ 0.290 , 0.949 ]
```

Parameters

Separate effect for each stratum

job	hieng=0	hieng=1
driv	λ_d	$\lambda_d\theta_d$
cond	λ_c	$\lambda_c\theta_c$
bank	λ_b	$\lambda_b\theta_b$

Common effect for each stratum

job	hieng=0	hieng=1
driv	λ_d	$\lambda_d\theta$
cond	λ_c	$\lambda_c\theta$
bank	λ_b	$\lambda_b\theta$

How do we test whether V is a confounder?

This is not a sensible question!

Use of a reference rates the SMR

- A comparison of an exposed group with an unexposed group from the same study is an *internal* comparison
- Sometimes mortality in a group of interest is compared with that in some standard reference population. This is an *external* comparison
- The SMR is a rate ratio comparison of morality in a study group with mortality in a reference population — controlling for age by stratification
- This method is called "indirect" standardization. As we have seen, this is a different approach to that of "direct" standardization

linear models

Erasmus, 2002

Commands vs menus

The commands `tabmenu1` and `effmenu1` use the classification of the response variable to make decisions about summaries and effects.

These commands are very easy to use, but there will always be a need, for some analyses, to access the underlying regression commands on which `effmenu1` depends.

Response	`effmenu1`	Regression
binary	odds ratios	`logistic`
metric	diffs in means	`regress`
failure/count	rate ratios	`poisson`

The births data

Effect of hypertension on birth weight

```
----------------------
      hyp |    bweight
----------+-----------
        0 |    3198.90
        1 |    2768.21

. regress bweight hyp

 bweight |      Coef.  [95% Conf. Interval]
---------+--------------------------------
     hyp | -430.6959  -585.821   -275.5707
   _cons |  3198.904  3140.038     3257.77
```

The regression model used by `regress` is

$$\mu = \alpha + \beta X$$

where X is coded 0 for normal, 1 for hypertensive, so

$$\begin{array}{lcll} \mu & = & \alpha & \text{when } X = 0 \\ \mu & = & \alpha + \beta & \text{when } X = 1 \end{array}$$

The ML values of the parameters are $\alpha = 3198.9$ and $\beta = -430.7$.

The effect of hypertension on low birth weight.

```
Summary using odds

      hyp |       lowbw
----------+-----------
        0 |        0.10
        1 |        0.38

. logistic lowbw hyp

lowbw | Odds Ratio  [95% Conf. Interval]
------+--------------------------------
  hyp |   3.730769  2.027475    6.865011
```

The model used by `logistic` is

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

X is coded 0 for normal, 1 for hypertensive, so

$$\begin{aligned} \ln\left(\tfrac{\pi}{1-\pi}\right) &= \alpha && \text{when } X = 0 \\ \ln\left(\tfrac{\pi}{1-\pi}\right) &= \alpha + \beta && \text{when } X = 1 \end{aligned}$$

The parameter β is the difference between two log odds, i.e. a log odds ratio. The corresponding odds ratio is exp(β).

The ML values of exp(β) is 3.73. The ML value of α is not reported.

```
low birth | hypertens
weight    |    0     1
----------+-----------
        0 |  388    52
        1 |   40    20
----------+-----------
          |  428    72
```

$$\begin{aligned} \ln\left(\tfrac{\pi}{1-\pi}\right) &= \alpha + \beta X \\ \ln(40/388) &= \alpha \\ \ln(20/52) &= \alpha + \beta \\ \ln(20/52) - \ln(40/388) &= \beta \quad = \ln\left(\tfrac{20/52}{40/388}\right) \\ 1.3166 &= \beta \\ 3.73 &= \exp(\beta) \end{aligned}$$

Regression models

Ordinary regression

$$\mu = \alpha + \beta X$$

Logistic regression

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X$$

Poisson regression

$$\ln(\lambda) = \alpha + \beta X$$

In each case β is the effect per unit of X, measured as a change in the mean (ordinary regression); an odds ratio (logistic regression); a rate ratio (Poisson regression).

By default, in Stata regression commands, variables are metric.

Categorical exposures with more than two levels

The variable `hyp` has two levels
The variable `gest4`, created below, has 4 levels.

```
. egen gest4=cut(gestwks), at(20,35,37,39,45)
```

Using `tabmenu1` and `effmenu1` with odds

```
    gest4 |       lowbw
----------+-----------
       20 |        4.17
       35 |        0.68
       37 |        0.12
       39 |        0.01

Level    Effect    95% CI

2 vs  1  0.1642   [ 0.053 , 0.512 ]
3 vs  1  0.0290   [ 0.010 , 0.080 ]
4 vs  1  0.0028   [ 0.001 , 0.012 ]
```

To include `gest4` in regression commands we need to use indicator variables for the 4 levels.

gest4	$X1$	$X2$	$X3$	$X4$
20	1	0	0	0
35	0	1	0	0
37	0	0	1	0
39	0	0	0	1

```
. tabulate gest4, generate(X)
. logistic lowbw X2 X3 X4
```

```
----------------------------------------
lowbw | Odds Ratio   [95% Conf. Interval]
-------+--------------------------------
   X2 |   .1642105    .052703    .5116427
   X3 |   .0289933   .0104924    .0801162
   X4 |   .0028016   .0006602    .0118891
----------------------------------------
```

The variable that indicates the base line is omitted from the regression model.

```
. xi: logistic lowbw i.gest4
```

```
---------------------------------------------
     lowbw | Odds Ratio   [95% Conf. Interval]
-----------+---------------------------------
_Igest4_35 |   .1642105    .052703    .5116427
_Igest4_37 |   .0289933   .0104924    .0801162
_Igest4_39 |   .0028016   .0006602    .0118891
---------------------------------------------
```

`xi` stands for e(x)pand (i)ndicators

The baseline is, by default, the first level, but this can be changed to (say) the third level (37–) with

```
. char gest4[omit] 37
```

To re-set the default, use

```
. char gest4[omit]
```

The diet data

Effect of `job` on failure (using `tabmenu1` and `effmenu1`)

```
Summary using rates per 1000

      job |        chd
----------+-----------
        1 |       9.78
        2 |      13.42
        3 |       8.57

Level    Effect    95% CI

2 vs  1  1.3720   [ 0.635 , 2.966 ]
3 vs  1  0.8766   [ 0.429 , 1.793 ]

Test for no effects

chi2(  2)    =   1.677
P-value      =   0.432
```

Using Poisson regression

```
. xi: poisson chd i.job, e(y) irr
------------------------------------------
    chd |        IRR   [95% Conf. Interval]
--------+---------------------------------
_Ijob_2 |   1.371998   .6345893    2.966295
_Ijob_3 |   .8765802   .4285257    1.793108
      y | (exposure)
------------------------------------------
```

The option `e(y)` is how the follow-up time variable, `y`, is declared.

Hypothesis: both effects of job are 1 (zero on a log scale)

Likelihood ratio test

```
. xi: poisson chd i.job, e(y)
. lrtest, saving(0)
. xi: poisson chd, e(y)
. lrtest

chi2(2)     =       1.61
Prob > chi2 =     0.4481
```

Wald test

```
. xi: poisson chd i.job, e(y)
. testparm _Ijob_2 _Ijob_3

 ( 1)  [chd]_Ijob_2 = 0.0
 ( 2)  [chd]_Ijob_3 = 0.0

    chi2(  2) =    1.68
  Prob > chi2 =    0.4323
```

The effect of energy on failure.

```
. poisson chd energy , e(y) irr

-----------------------------------------
   chd |        IRR  [95% Conf. Interval]
-------+---------------------------------
energy |     .99885  .9981367    .9995637

. gen energy100=energy/100
. poisson chd energy100, e(y) irr

--------------------------------------------
      chd |        IRR  [95% Conf. Interval]
----------+---------------------------------
energy100 |   .8913034  .8298593    .9572968
```

Controlling for confounders using linear models

Erasmus, 2002

Controlling the effect of hieng for job

To find the effect of hieng controlled for job, using effmenu, we declare hieng as the exposure and job as the control variable:

```
Level         Effect 95% Confidence Interval

2 vs 1        0.5248 [ 0.290 , 0.949 ]
```

To find the effects of job controlled for hieng we declare job as the exposure and hieng as the control variable:

```
Level         Effect 95% Confidence Interval

2 vs 1        1.3584 [ 0.628 , 2.937 ]
3 vs 1        0.8843 [ 0.432 , 1.809 ]
```

```
. xi: poisson chd i.hieng i.job, e(y) irr

-------------------------------------
         chd |        IRR   Std. Err.
-------------+-----------------------
   _Ihieng_1 |   .5247666   .1585834
     _Ijob_2 |   1.358442   .5344426
     _Ijob_3 |   .8843023   .3229207
```

The Stata Poisson regression command makes no distinction between the exposure variable and the control variable.

The first number reported is the effect of hieng controlled for job, and the next two are the effects of job controlled for hieng.

Models and parameters

Effmenu1

job	hieng=0	hieng=1
driv	λ_d	$\lambda_d\theta$
cond	λ_c	$\lambda_c\theta$
bank	λ_b	$\lambda_b\theta$

Regression

job	hieng=0	hieng=1
driv	λ	$\lambda\theta$
cond	$\lambda\beta_c$	$\lambda\theta\beta_c$
bank	$\lambda\beta_b$	$\lambda\theta\beta_b$

Effect modification

hieng as exposure, job as modifier

job	Effect
driver	0.41
conductor	0.66
bank	0.52

To compare these effects use ratios:

job	Effect	
driver	0.41	0.41/0.41 = 1
conductor	0.66	0.66/0.41 = 1.60
bank	0.52	0.52/0.41 = 1.26

The numbers 1.60 measures how much the effect of hieng differs between conductors and drivers, while 1.26 measures how much the effect of hieng differs between bank workers and drivers.

They are called the interactions between hieng and job.

Interactions are symmetric

job as exposure, hieng as modifier

Level 2 vs level 1 of job

hieng	Effect	Ratio
low	1.14	1.14/1.14 = 1
high	1.82	1.82/1.14 = 1.60

Level 3 vs level 1 of job

hieng	Effect	Ratio
low	0.81	0.81/0.81 = 1
high	1.03	1.03/0.81 = 1.26

The interactions between hieng and job are the same as those between job and hieng.

The interpretation of interactions is NOT symmetric!

```
. xi: poisson chd i.hieng*i.job, e(y) irr
-------------------------------------
         chd |        IRR   Std. Err.
-------------+-----------------------
   _Ihieng_1 |   .4102648   .2512349
     _Ijob_2 |   1.136857   .5684285
     _Ijob_3 |    .813427   .3712769
_IhieXjob_~2 |   1.596755   1.303745
_IhieXjob_~3 |   1.261973   .9638479
```

0.41 is the effect of `hieng` when `job` is at its first level.

1.14 and 0.81 are the effects of `job` when `hieng` is at its first level.

1.60 and 1.26 are the interactions between `hieng` and `job`.

job	hieng=0	hieng=1
driv	λ	$\lambda\theta$
cond	$\lambda\beta_c$	$\lambda\theta\beta_c\gamma_{c1}$
bank	$\lambda\beta_b$	$\lambda\theta\beta_b\gamma_{b1}$

γ_{c1} and γ_{b1} are the interaction parameters. They measure deviations from the hypothesis of common effect of `hieng` in all `job` categories.

How to make Stata behave like `effmenu1`

The trick is to define 0/1 variables that match the `effmenu1` parametrization:

job	hieng=0	hieng=1
driv	λ_d	$\lambda_d\theta_d$
cond	λ_c	$\lambda_c\theta_c$
bank	λ_b	$\lambda_b\theta_b$

```
. gen ld = ( job == 1 )
. gen lc = ( job == 2 )
. gen lb = ( job == 3 )
. gen thd = ( job == 1 ) * ( hieng == 1 )
. gen thc = ( job == 2 ) * ( hieng == 1 )
. gen thb = ( job == 3 ) * ( hieng == 1 )
```

Stata output

Omitting `ld` makes the drivers the baseline.

```
. poisson chd ld lb thc thd thb, e(y) irr

-----------------------------------------------------
         chd |        IRR     [95% Conf. Interval]
-------------+---------------------------------------
          lc |   1.136857     .4266828    3.029051
          lb |    .813427     .3325064    1.989927
         thd |   .4102648     .1235412    1.362438
         thc |   .6550924     .2273009    1.888008
         thb |   .5177431      .211639    1.266581
           y | (exposure)
-----------------------------------------------------
```

The `th*`-parameters are exactly those that are reported in `effmenu1`.

Erasmus, 2002

Longitudinal studies involve repeated measurements of outcome on the same subject. As an example we consider a study of depression in which each subject visited a clinic annually for 7 years. At each visit the subject was classified as depressed or not, by dichotomizing a score.

Data for such a study can be recorded in 3 ways

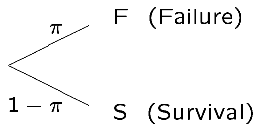
```
      id  wave    d
10048243     1    0
10048243     2    0
10048243     3    0
10048243     4    0      Long
10048243     5    1
10048243     6    1
10048243     7    1

10048243  0 0 0 0 1 1 1  Wide

10048243     3           Aggregated
```

Data management is important in longitudinal studies.

The binary model

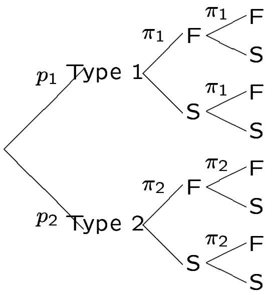


- Until now, we have assumed all subjects with the same pattern of explanatory variables share the same risk, but what if risk varies from subject to subject due to unknown or unobserved factors?
- If we cannot tell one type of subject from another, we only observe failure or survival. The probability of failure is the risk averaged over subjects – called the *subject-marginal* risk.
- In simple analyses such as logistic regression, the risks modelled are subject-marginal — we don't *really* believe that all subjects are the same! But we get the same answer whether we assume subjects have the same risks or different risks.

Manifestations of heterogeneity

- With only one observation per subject the extent of the heterogeneity in risk cannot be measured.
- With more than one observation per subject, heterogeneity is manifest in two ways:
    - Observations are not independent of one another
    - The total number of failures observed per subject is more variable in a heterogeneous population than in a homogeneous population
- In this case, we are not only able to take account of heterogeneity — we *must* do so for the analysis to be valid

- As a simple illustration, suppose there are two types of subject in the population, with probabilities p_1, p_2. When the probability of failure does not change between trials:



Subject type	Proba–bility	Probability of outcome		
		(S,S)	(S,F) or (F,S)	(F,F)
1	p_1	$(1-\pi_1)^2$	$\pi_1(1-\pi_1)$	$(\pi_1)^2$
2	p_2	$(1-\pi_2)^2$	$\pi_2(1-\pi_2)$	$(\pi_2)^2$
Marginal		$p_1(1-\pi_1)^2+$ $p_2(1-\pi_2)^2$	$p_1\pi_1(1-\pi_1)+$ $p_2\pi_2(1-\pi_2)$	$p_1(\pi_1)^2+$ $p_2(\pi_2)^2$

- Marginal probabilities of outcomes depend not only on the subject-marginal risk, $\overline{\pi}$, but also on the *variance* of π's in population, $\widetilde{\pi}$ say

	Trial 2		
Trial 1	F	S	Marginal
F	$(\overline{\pi})^2+\widetilde{\pi}$	$\overline{\pi}(1-\overline{\pi})-\widetilde{\pi}$	$\overline{\pi}$
S	$\overline{\pi}(1-\overline{\pi})-\widetilde{\pi}$	$(1-\overline{\pi})^2+\widetilde{\pi}$	$1-\overline{\pi}$
Marginal	$\overline{\pi}$	$1-\overline{\pi}$	1

- Scoring F as 1 and S as 0, the *covariance* between the two responses is $\widetilde{\pi}$ and the *correlation coefficient* is $\widetilde{\pi}$ / $[\overline{\pi}(1-\overline{\pi})]$

- The distribution of the number of failures per subject, d, is

$$\Pr\begin{pmatrix} d=2 \\ d=1 \\ d=0 \end{pmatrix} = \begin{pmatrix} (\overline{\pi})^2 & + & \widetilde{\pi} \\ 2\overline{\pi}(1-\overline{\pi}) & - & 2\widetilde{\pi} \\ (1-\overline{\pi})^2 & + & \widetilde{\pi} \end{pmatrix}$$

- The *variance* of d is

$$\underset{\left(\substack{\text{Binomial}\\ \text{variance}}\right)}{2\overline{\pi}(1-\overline{\pi})} \quad + \quad \underset{\left(\substack{\text{Extra-binomial}\\ \text{variance}}\right)}{2\widetilde{\pi}}$$

- For n trials, the variance of d is

$$n\overline{\pi}(1-\overline{\pi})+n(n-1)\widetilde{\pi}$$

Note that when $n=1$ only the binomial variance remains.

Modelling strategies: subject–specific

- Consider an exposed and an unexposed group of subjects, and denote presence or absence of exposure by $x=1$ or $x=0$.

- The situation where the risk is constant from trial to trial within a subject, but varies according to the (unobserved) type of subject, is represented by the model

$$\log\left(\frac{\pi}{1-\pi}\right)=\alpha+\beta x+\gamma_{\text{type}}$$

  - π is a *subject–specific* risk for any trial
  - γ_{type} are effects of (unobserved) subject type
  - Even though type is not observed, such a model can be fitted. It is a *random effects* model

Modelling strategies: subject–marginal

- The marginal relationship can also be written as a logistic regression model:

$$\log\left(\frac{\overline{\pi}}{1-\overline{\pi}}\right)=\alpha+\beta x$$

  - $\overline{\pi}$ is a *subject-marginal* risk
  - In fitting such a model, we must allow for *overdispersion* and *correlation* of responses

- But there is a snag. Unless the risks are small (less than 0.10), the effects represented by β in the two models are not the same.

ratios

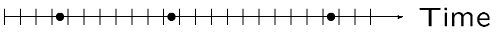
- Even when the subject-specific odds ratio for exposure is the same for all types of subject, this is not the same as the subject-marginal odds ratio.

- Eg – 3 types of subject, equally frequent in the population, and an exposure with an odds ratio of 2 for each type:

	Exposed		Unexposed		Odds
Type	Risk	Odds	Risk	Odds	ratio
1	0.889	8.0	0.8	4.00	2.0
2	0.667	2.0	0.5	1.00	2.0
3	0.333	0.5	0.2	0.25	2.0
Marginal	0.630	1.7	0.5	1.0	1.7

- The risk we observe in the population is the average of the risks. The population odds are the odds corresponding to that risk, *not* the average of the odds.

- When dealing with recurrent events in time; we divide the time scale into "clicks" of short duration, h say:



The risk of failure in any click is λh where λ is the *rate* or *intensity* parameter

- Subject-specific effects are the same as subject-merginal effects.

- In the absence of heterogeneity, the number of events per subject, d, in follow-up time y follows a Poisson distribution with mean λy and variance also equal to λy

- If subject–specific rates are heterogeneous in the population of subjects with mean and variance $\overline{\lambda}$ and $\widetilde{\lambda}$, the mean and variance of d are $\overline{\lambda}y$ and $\overline{\lambda}y + \widetilde{\lambda}(y)^2$. There is *extra-Poisson variation*

Time-to-event data

- If an event cannot recur, we can observe it at most once. Is subject heterogeneity an issue in this case?

- Example: cycle–specific risks of conception in a life table of consecutive ovulatory cycles

Cycle	Women	Conceptions	Risk
1	1538	194	0.126
2	1332	136	0.102
3	1176	119	0.101
4	1022	101	0.099
5	914	76	0.083
...			
9	503	32	0.064
10	454	33	0.073
11	402	25	0.062
12	349	20	0.057

- These are time-to-event data, "time" being measured as the (discrete) number of cycles until conception

- Conception risk seems to fall with time

Two explanations

1. *Subject–specific* risk of conception at each cycle is the same for all subjects, but decreases with time

2. Subject-specific risks of conception vary between subjects but are constant over time; the *subject-marginal* risk falls with time because the most fecund women conceive early and are removed from the population studied

- The latter explanation is the only plausible one. In this case, subject heterogeneity is manifest even though we observe no more than one event per subject

- We should fit models which assume a constant subject-specific risk over time, but allow for heterogeneity between subjects. Indeed this is the usual way to model fecundibility data

Erasmus, 2002

Let the rate parameter for subject i be λ_i, and suppose the subject experiences d_i events in time y_i.

In repetitions of the study with fixed y_i the total number of events for a subject has a Poisson distribution with mean and variance both equal to $\lambda_i y_i$.

There is rarely enough data to estimate each subject parameter. Instead assume

$$\lambda_i = f_i \overline{\lambda}$$

where $\bar{\lambda}$ is the marginal rate parameter and f_i is the *frailty* for subject i chosen at random from a distribution, mean 1, variance κ.

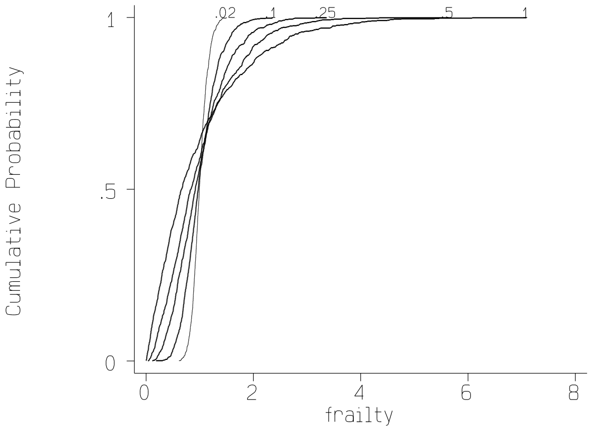
Under this model

$$\text{Mean}(d_i) = \overline{\lambda} y_i \qquad \text{Var}(d_i) = \overline{\lambda} y_i + \kappa(\overline{\lambda} y_i)^2$$

When $\kappa = 0$ the subjects are homogeneous; when $\kappa > 0$ the subjects are heterogeneous.

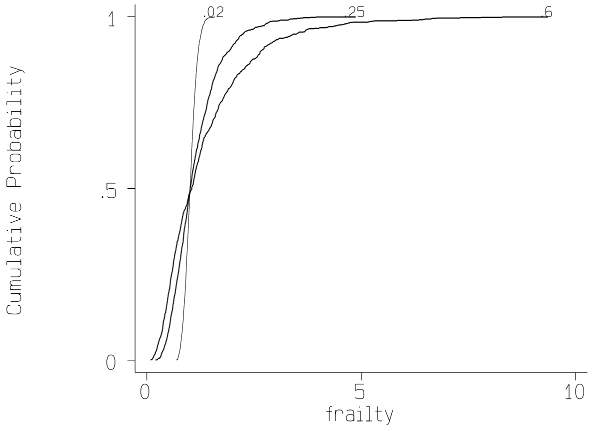
The gamma frailty distribution

Gamma frailty distributions with mean 1 and variance 1, 0.5, 0.25, 0.1, 0.02 are shown below.



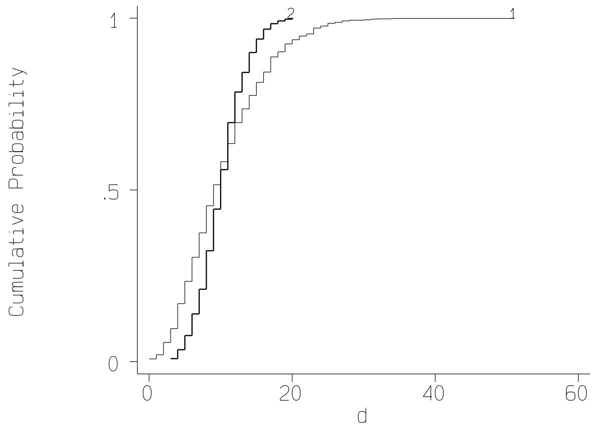
The log normal frailty distribution

The distribution of the log frailty is normal (Gaussian) with mean zero and variance κ. The variance of the frailties is then also (approximately) κ provided κ is not too large.



When the frailties are gamma the distribution of the number of events is negative binomial.

The graph shows the cdf of a Poisson distribution with mean 10 and the same distribution combined with gamma frailies with variance 0.25.



tribution

The likelihood for the negative binomial distribution is a complicated expression which depends on $\bar{\lambda}$ and κ.

However, the score function for $\log(\bar{\lambda})$ is simple:

$$\sum_i \frac{1}{1+\kappa\bar{\lambda}y_i}(d_i - \bar{\lambda}y_i)$$

The score function for κ is rather more complicated!

Comparison of poisson with negative binomial estimates

The score equation for a log rate parameter under the poisson model is

$$\sum(d_i - \lambda y_i) = 0$$

Under the negative binomial model this becomes

$$\sum \frac{1}{1+\kappa\bar{\lambda}y_i}(d_i - \bar{\lambda}y_i)$$

Each takes the form

$$\sum w_i(d_i - \lambda y_i) = 0$$

but for the poisson the weights are all equal, while for the negative binomial the weight for a subject with long follow–up (hence high events) is less than for a subject with a short follow–up.

The effect of exposure

At the subject level the effect of exposure is θ in the model

$$\lambda_{i1} = \theta\lambda_{i0}$$

At the marginal level it is θ in the model

$$\bar{\lambda}_1 = \theta\bar{\lambda}_0$$

where $\bar{\lambda}_1$ is the marginal rate for exposed subjects and $\bar{\lambda}_0$ for unexposed subjects. The two θ's are the same.

The models can also be expressed in the form

$$\begin{aligned} \log(\lambda_i) &= \alpha + \beta x + \gamma_i \\ \log(\bar{\lambda}) &= \alpha + \beta x \end{aligned}$$

where $x = 0$ for unexposed subjects and $x = 1$ for exposed subjects.

The stata command to fit a regression model using the negative binomial likelihood is **nbreg**. We shall use the epilepsy data as an example.

```
. use epilep1, clear
. nbreg d trt, e(y)

       d |      Coef.   Std. Err.
---------+-----------------------
     trt |  -.0750871   .2514438
   _cons |   1.456328   .1821562
       y | (exposure)

/lnalpha |  -.1054406    .173741
---------+----------------------
   alpha |    .899928   .1563544
--------------------------------
Likelihood ratio test of alpha=0:
chi2(1) =  1867.07   Prob > chi2 = 0.0000
```

The frailty variance (alpha here kappa in notes) is 0.9 which is very high.

- d_i is the number of failures out of i trials in subject i. In repeated studies the distribution of d_i is binomial with mean $n_i\pi_i$ and variance $n_i\pi_i(1-\pi_i)$.

- The natural model for the odds is
$$\Omega_i = f_i\overline{\Omega}$$
or
$$\log(\Omega_i) = \gamma_i + \log(\overline{\Omega})$$
where $\overline{\Omega}$ is the marginal odds and γ_i is a random subject effect with mean zero.

- There is no distribution which for the subject effects which leads to simple exact calculations, but such a model may still be fitted using numerical approximations. A common assumption is that the subject effects have a Gaussian distribution with mean 0 and variance κ.

The effect of exposure

At the subject level the effect of exposure is θ in the model
$$\Omega_{i1} = \theta\Omega_{i0}$$
At the marginal level it is θ in the model
$$\overline{\Omega}_1 = \theta\overline{\Omega}_0$$

where $\overline{\Omega}_1$ is the marginal odds for exposed subjects and $\overline{\Omega}_0$ for unexposed subjects. Unles the Ω's are small the two θ's are not the same.

The models can also be expressed in the form
$$\begin{aligned}\log(\Omega_i) &= \alpha + \beta x + \gamma_i\\ \log(\overline{\Omega}) &= \alpha + \beta x\end{aligned}$$
where $x = 0$ for unexposed subjects and $x = 1$ for exposed subjects.

Generalized Linear Models and Quasi–likelihood

Erasmus, 2002

- A metric response (z_i) for subject i is assumed to have a Gaussian distribution with mean μ_i and variance $v_i = (\sigma_i)^2$.

- For a single explanatory variable

$$\mu_i = \alpha + \beta x_i$$

- The score function for β is

$$\sum_i w_i(z_i - \mu_i)x_i$$

where $w_i = (v_i)^{-1}$.

- When the variance is constant ($v_i = v$) all the weights are equal and the score function is

$$\sum_i (z_i - \mu_i)x_i$$

- A generalization is to assume that the linear part of the model holds on a transformed scale, for example

$$\log(\mu_i) = \alpha + \beta x_i$$

- More generally we might assume

$$g(\mu_i) = \alpha + \beta x_i$$

where $g(\mu)$ is called the *link function*.

- The score function now takes the form

$$\sum_i w_i(z_i - \mu_i)x_i$$

where the weights are $w_i = 1/([v_i g'(\mu_i)]$, and $g'(\mu)$ denotes the slope of the graph of $g(\mu)$ against μ.

The exponential family

For a wide class of probability models, called the *exponential family* the score function for β takes the same form.

The family includes Gaussian, binary, binomial, Poisson, negative binomial, and these differ only in the variance of the response.

- Gaussian: z_i is the metric response for subject i
  - $\mathrm{E}(z_i) = \mu_i$
  - $v_i = (\sigma_i)^2$
- Poisson: $z_i = d_i$, the number of events for subject i
  - $\mathrm{E}(z_i) = \lambda_i y_i = \mu_i$
  - $v_i = \lambda_i y_i = \mu_i$
- Binary: $z_i = d_i$, coded 0/1 for subject i
  - $\mathrm{E}(z_i) = \pi_i = \mu_i$
  - $v_i = \pi_i(1 - \pi_i) = \mu_i(1 - \mu_i)$
- Negative binomial: $z_i = d_i$, coded 0/1 for subject i
  - $\mathrm{E}(z_i) = \lambda_i y_i = \mu_i$
  - $v_i = \mu_i + \kappa(\mu_i)^2$

The natural links for the Poisson and binary models

- For the Poisson model $v_i = \mu_i$, and the choice of $g(\mu_i) = \log(\mu_i)$ as the link function makes $w_i = 1$. The score function for β is then

$$\sum_i (z_i - \mu_i)x_i = \sum_i (d_i - \lambda_i y_i)x_i$$

- For a binary model $v_i = \mu_i(1 - \mu_i)$, and the choice of $g(\mu_i) = \log[\mu_i/(1 - \mu_i)]$ as the link function (the *logit* link), makes $w_i = 1$. The score function for β is then

$$\sum_i (z_i - \mu_i)x_i = \sum_i (d_i - \pi_i)x_i$$

- Other links are possible, and the general form of the score function is

$$\sum \text{Weight} \times (\text{Observed} - \text{Expected}) \times x$$

- GLM's are linear models in which the probability model for the response is a member of the exponentila family.

- The expected response is related to a linear function of explanatory variables by a *link function*, $g(\mu)$ say:

$$g(\mu) = \alpha + \beta x + \cdots$$

- The score function for a regression coefficient, β, takes the form

$$\sum_i \underset{\text{Weight}}{\frac{1}{v(\mu_i)g'(\mu_i)}} \qquad \underset{\text{(Observed − Expected)}}{(z_i - \mu_i)} \qquad \underset{x}{x_i}$$

- Components of the weights are:
  - $v(\mu)$ the variance of the response when its expected value is μ, and
  - $g'(\mu)$ the slope of the graph of the link function, $g(\mu)$, evaluated at μ

- For the "natural" link function for model (log for Poisson, logit for binary, etc.) the weights are 1.

- This generalization allows rationalization of software and more flexibility – for example *additive* models for rates can be fitted, using the identity link function.

Using the glm command in Stata

When using the `glm` command it is necessary to specify the family and the link. For example, for Poisson regression we use

```
. glm d x, family(poisson) link(log)
```

where d is the number of events and x refers to the variables in the model.

For logistic regression (d coded 0/1) we use

```
glm d x, family(binomial) link(logit)
```

- The binary model is a special case of the binomial with $n_i = 1$.

- GLM's can be used to estimate risk ratios instead of odds ratios using `family(binomial) link(log)`.

- Rate differences can be estimated using `family(poisson) link(identity)`

Log offsets

The GLM with family Poisson is expressed in terms of the mean. In follow-up studies $\mu = \lambda y$ where y is the follow-up time, and the parameter of interest is λ.

On a log scale

$$\log(\mu) = \log(y) + \log(\lambda)$$

so when $\log(\mu) = \alpha + \beta x$

$$\log(\lambda) = -\log(y) + \alpha + \beta x$$

The term $-\log(y)$ is called an *offset*. In Stata the `glm` command should include the option `lnoff(y)`.

- When the distribution of z is not from the exponential family but we know its variance as a function of its mean

$$v_i = v(\mu_i)$$

we can still use

$$\sum_i w_i(z_i - \mu_i)x_i$$

with weights $w_i = 1/[v(\mu_i)g'(\mu_i)]$ as a "quasi" score function which has all the right properties. In particular:

  - Var("Score") = "Information"
  - Estimates based on "score" equations, although not ML, have the smallest possible SEs

- Use of the method in these circumstances is called "quasi–likelihood" estimation

- Assume frailties drawn from *any distribution* with mean 1 and variance κ

- The predicted mean and variance of subject event counts, d_i, are

$$\mu_i = \overline{\lambda} y_i, \qquad v(\mu_i) = \mu_i + \kappa(\mu_i)^2$$

— as for the negative binomial model.

- To estimate the effect of exposure in the marginal model

$$\overline{\lambda}_1 = \theta\overline{\lambda}_0$$

when κ is known, we can use quasi-likelihood with the negative binomial variance function.

Variance parameters

- When $v(\mu)$ involves unknown parameters (such as κ) we need a method to estimate these

- Negative binomial regression uses a ML estimate, but this makes the strong assumption that frailties have a gamma distribution

- Another way is to choose κ to make

$$\sum \frac{(\text{observed} - \text{expected})^2}{v(\mu)} = \text{df}$$

- In Stata, the expression above is called the *Pearson chi-squared*.

Example of choosing κ

In the PNG data, the effect of vaccination in a model with

$$\text{Var}(z) = \mu + \kappa(\mu)^2$$

where $\kappa = 1$ can be found by using

```
. glm d vacc, fam(nbinomial 1) lnoff(y)

Generalized linear models         No. of obs      =       1390
Optimization     : ML             Residual df     =       1388
                                  Scale param     =          1
Deviance         =  1262.297541   (1/df) Deviance =   .9094363
Pearson          =  1027.288414   (1/df) Pearson  =   .7401213

Var function:  V(u) = u+(1)u^2
Link function: g(u) = ln(u)

Log likelihood   =  -2415.96973

------------------------------------------------
     d |      Coef.   Std. Err.      z    P>|z|
-------+----------------------------------------
  vacc |  -.0792613   .0683746   -1.16   0.246
 _cons |   .5948046   .1064859    5.59   0.000
------------------------------------------------
```

The Pearson chi-squared divided by its degrees of freedom is 0.74.

Because we want this to be 1, κ is too large.

Trying $\kappa = 0.5$:

```
. glm d vacc, fam(nbinomial 0.5) lnoff(y)
Generalized linear models         No. of obs      =       1390
Optimization     : ML             Residual df     =       1388
                                  Scale param     =          1
Deviance         =  1660.453165   (1/df) Deviance =   1.196292
Pearson          =  1485.379326   (1/df) Pearson  =   1.070158

Var function: V(u) = u+(0.5)u^2
Link function: g(u) = ln(u)

Log likelihood   = -2401.356425

-----------------------------------------------------
           d |      Coef.   Std. Err.      z    P>|z|
-------------+---------------------------------------
        vacc |  -.0792613   .0568729    -1.39   0.163
       _cons |   .5948046   .0883881     6.73   0.000
-----------------------------------------------------
```

The Pearson chi-squared divided by its df is now 1.07. Trial and error (iteration) will usually converge to a value of κ which makes the Pearson chi-squared exactly equal to its df, but there are better ways of doing this.

Erasmus, 2002

What happens if we use the wrong variance function?

- The variance function for the Poisson family is $v(\mu) = \mu$. The corresponding score function for λ is
$$\sum_i (d_i - \lambda y_i)$$
and this is wrong when there is subject heterogeneity.

- The ML value of λ is $\sum_i d_i / \sum_i y_i$ which makes the score zero.

- When rates are heterogeneous, E$(d_i) = \overline{\lambda} y_i$, so the equation
$$\sum_i \left(d_i - \overline{\lambda} y_i\right) = 0$$
still has expectation zero, and therefore leads to a consistent estimate for $\overline{\lambda}$, again $\sum_i d_i / \sum_i y_i$.

But the SE's are wrong

- Because the likelihood is incorrect for heterogenous subjects, the variance of the score is no longer equal to the information

- If the Poisson model is true
$$\text{Var}\left[\sum_i (d_i - \lambda y_i)\right] = \sum_i (\lambda y_i)$$
but, under heterogeneity,
$$\text{Var}\left[\sum_i (d_i - \overline{\lambda} y_i)\right] = \sum_i [\overline{\lambda} y_i + \tilde{\lambda}(y_i)^2]$$

- Thus, significance tests based on this likelihood give wrong p–values, and support intervals have incorrect coverage

- In addition, our estimate of $\overline{\lambda}$ may not be as good as it might be

Poisson regression

- These ideas generalize to the case where we wish to fit a regression model for the marginal rate as a function of explanatory variables

- A simple example is the model

$$\overline{\lambda}_1 = \theta \overline{\lambda}_0$$

used to estimate the effect of exposure.

- The rates $\overline{\lambda}_1$, $\overline{\lambda}_0$, are the marginal rates for exposed and unexposed subjects respectively.

- Poisson regression produces consistent estimates for the effect of exposure, even in the presence of heterogeneity, but p–values and coverage probabilities are incorrect.

The standard error S of an estimate M of a parameter β has been defined as

$$S = \sqrt{1/I(M)}$$

where $I(M)$ is the information at M.

Using frequency theory we showed that the coverage probability of the supported range

$$M \pm 1.645S$$

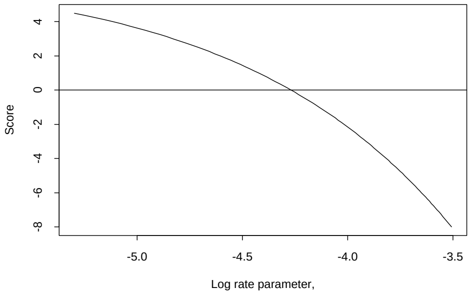
is approximately 90%.

When the model is wrong this supported range no longer has coverage probability 90%, so we need to find an interval which does have coverage probability 90%.

Recall: the score function for log(λ)

- In the simple Poisson model of homogeneous failure rate, λ, the score function for log(λ) is

$$U = D - \lambda Y$$



- ML estimate solves the *score equation*

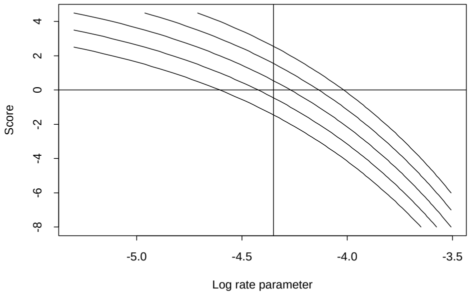
$$D - \lambda Y = 0$$

- The information at $M = \log(D/Y)$ is

$$I(M) = D$$

The variance of M

Each curve is the score curve for a repetition.



The curves are approximately linear in the region of the true value of $\beta = \log(\lambda)$, so

$$I(\beta) \approx I(M) = I$$

and

$$M - \beta \approx U(\beta)/I$$

so in repetitions of the study

$$\text{Var}(M) = \frac{\text{Var}(U)}{(I)^2} = \frac{I}{(I)^2} = \frac{1}{I} = \frac{1}{D}$$

Robust estimates of the variance of M

- The score, U, is calculated by adding independent contributions, u_i, from each subject

$$U = \sum u_i$$

- In the case of estimating a rate

$$u_i = d_i - \overline{\lambda} y_i$$

where d_i, y_i are event counts and observation times for subject i

- The estimated variance of U is

$$N \frac{\sum (u_i - \bar{u})^2}{N-1} = N \sum (u_i)^2 / (N-1)$$

The u_i are evaluated at the estimated parameter value so $\bar{u} = 0$.

- As before, the variance of M, the estimated value of $\log \overline{\lambda}$, is

$$\text{Var}(M) \approx \frac{\text{Var}(U)}{(I)^2}$$

but this no longer simplifies to $1/I$. Instead

$$\text{Var}(M) = \frac{N \sum (u_i)^2 / (N-1)}{(I)^2}$$

We shall refer to the square root of this estimated variance as S_{R}, the robust SE.

- The supported range

$$M \pm 1.645 S_{\text{R}}$$

now has approximately 90% coverage.

- Another way of writing of writing $\text{Var}(U)/(I)^2$ is

$$I^{-1} V I^{-1}$$

the "information sandwich"

An example

Epileptic seizures in 28 subjects

Subject	d_i	$u_i = d_i - \lambda y$	$(u_i)^2$
1	14	-20.32	412.96
2	14	-20.32	412.96
3	11	-23.32	543.89
4	13	-21.32	454.60
5	55	20.68	427.60
...			
8	93	58.68	3443.17
...			
18	123	88.68	7863.89
...			
25	143	108.68	11811.03
...			
28	53	18.68	348.89
Total	961	0.00	33088.11

- Fitted value for λy is $961/28 = 34.32$
- Information based SE of estimate of $\log \lambda$, using the Poisson model, is

$$S = \sqrt{1/I} = \sqrt{1/961} = 0.032$$

- If the Poisson model were correct, sums of d_i and $(u_i)^2$ would be approximately equal
- The robust SE is

$$S_{\text{R}} = \sqrt{\frac{28 \times 33088.11}{27 \times (961)^2}} = 0.192$$

Robust SE's in stata

```
. use epilep1, clear
. poisson d if trt==0, e(y)
       d |      Coef.   Std. Err.
---------+-----------------------
   _cons |   1.456328   .0322581
---------------------------------

. di 8*exp(1.456)
34.310161

. poisson d if trt==0, e(y) robust

       d |      Coef.   Std. Err.
---------+-----------------------
   _cons |   1.456328   .1927568
---------------------------------
```

random effects

Erasmus, 2002

A group of subjects who are followed over time is called a cohort in epidemiology, and a panel in the social sciences.

The members of the panel are interviewed or visit a clinic at regular intervals.

Both the outcome and explanatory variables can change from one visit to another. A new record is used for each visit.

The outome can be time-to-next-event, binary, or metric.

We shall be concerned with the first two of these.

Three examples

1. In the PNG data the outcome is the time to the next infectious episode. Explanatory variables are vaccination status and age. Age changes with time but vaccination status is constant.

2. In the epilepsy data the outcome is the number of seizures in each two week period following treatment. Explanatory variables include treatment, a base-line for the number of seizures before treatment, and period. Only period changes with time.

3. In the DPS data the outcome is binary, namely whether or not the subject is depressed at the time of the visit, assessed by dichotomizing a score. Explanatory variables include sex, age at interview, care of handicapped member of family, and whether in a full time job. All variables except sex change with time.

How to record longitudinal data

In a typical longitudinal study there will be some information which changes with time, and some which does not. It is best to record these in separate files.

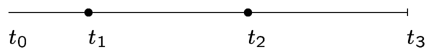
For example, date of birth and sex are constant, and would go in a file with one record per subject. Measurements made at each visit will change with time and should go in another file, with one record per visit (long coding).

The two files are merged for analysis.

Separating the two kinds of information saves repeating time constant information in all of the time varying records.

A special case is where events are recorded.

A typical event history



Subject			Event			
id	dob	etc.	id	tin	tout	diag
10	12/2/56	...	10	t_0	t_1	3
			10	t_1	t_2	17
			10	t_2	t_3	0

- The subject file contains information which is constant for a subject.

- The event file contains information about each event.

The file **pngsubj** contains the variables

`id`	`identity number for the subject`
`sex`	`1=male 2=female`
`vacc`	`vaccination 1=P 2=V`
`dob`	`date of birth`

The file **pngevent** contains the variables

`id`	`identity number for the subject`
`diag`	`diagnosis code (80,81 = ALRI)`
`timein`	`date at start of event record`
`timeout`	`date at end of event record`

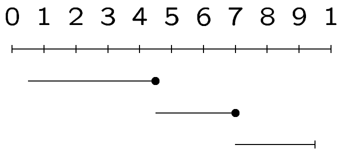
After defining the events of interest (ALRI, ALLI) the two files would be merged and aggregated to form

`id`	`identity number for the subject`
`sex`	`1=male 2=female`
`vacc`	`vaccination 1=P 2=V`
`dob`	`date of birth`
`d`	`total number of events`

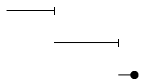
Splitting records into time-bands

To take account of the possibility that rates might vary on some time scale, first choose the time scale, eg time in study, age, etc., then split the records into time-bands on this scale.

For example, the records for a subject might look like this on an age scale.



To take account of age using (say) 2 year age bands, we must first split them into records like this:



etc., etc.

then collapse the data into the number of events and total follow–up time during each two year age band.

Shared random effects (frailties)

Longitudinal data are nearly always coded long, so that there are several records per subject, each record referring to a visit with binary data (A/B), or to an event, with time-to-event data.

If risks or rates are assumed to be constant over time, the data can be aggregated so that there is one record per subject.

When risks or rates are allowed to vary with time the data can only be aggregated at the subject x time-band level, so there are still several records per subject.

Records from the same subject share the same random effect, and are therefore correlated or clustered, and this must be taken account of when fitting models.

Epilepsy seizure data

```
      id    period         d

       1         1         5
       1         2         3
       1         3         3
       1         4         3

      18         1        37
      18         2        29
      18         3        28
      18         4        29

      35         1        22
      35         2        17
      35         3        19
      35         4        16
```

Note how the counts from records from the same subject are very similar.

```
. xi: xtpois d trt, e(y) i(id)

--------------------------------------------
         d |      Coef.  [95% Conf. Interval]
----------+---------------------------------
       trt |  -.0750871  -.5679078    .4177337
     _cons |   1.456328   1.099309    1.813348
         y | (exposure)
----------+---------------------------------
  /lnalpha |  -.1054406  -.4459667    .2350855
----------+---------------------------------
     alpha |    .899928   .6402051    1.265017
--------------------------------------------
```

The option `i(id)` tells the command that the data are clustered by subject (id).

Binary data

The same ideas apply to panel data with a binary outcome, as in the DPS data.

```
. use dps, clear
. xtlogit d care, i(id)

       d |      Coef.   Std. Err.
---------+------------------------
    care |   .5659821   .0904988
   _cons |  -1.639452   .0330635
---------+------------------------
/lnsig2u |   1.023291   .0398481
---------+------------------------
 sigma_u |   1.668034    .033234
     rho |   .7356132   .0077499
----------------------------------
```

$$\rho = (\sigma_u)^2/(1+(\sigma_u)^2)$$

There is no distribution of random subject effects that leads to a simply computed exact likelihood. However, numerical approximations can be used. Converting σ_u^2 to ρ is a computational aid, and is not helpful in interpretation.

Longitudinal data and GEE

Erasmus, 2002

Recurrent events

- We observe counts of events in the subject during different time periods, d_i^t

- At the subject level
  - $\mathsf{E}(d_i^t) = \lambda_i^t y_i^t$
  - $\mathsf{Var}(d_i^t) = \lambda_i^t y_i^t$
  - The responses are independent.

- At the marginal level the mean and variance of the count d_i^t are

  $$\mu_i^t = \overline{\lambda}^t y_i^t, \qquad v(\mu_i^t) = \mu_i^t + \kappa(\mu_i^t)^2$$

  and there is now correlation between the counts for the same subject

  $$\text{Covariance}(d_i^s, d_i^t) = \kappa \mu_i^s \mu_i^t$$

Binary data

- Consider a set of trials in the same subject, with outcomes $d^1, d^2, \ldots = 0$ or 1

- At the subject level

  — $\mathsf{E}(d_i^t) = \pi_i^t$
  — $\mathsf{Var}(d_i^t) = \pi_i^t(1 - \pi_i^t)$
  — The responses are independent.

- At the marginal level

  — $\mathsf{E}(d_i^t) = \overline{\pi}^t$
  — $\mathsf{Var}(d_i^t) = \overline{\pi}^t(1 - \overline{\pi}^t)$
  — For small κ
  Covariance$(d_i^s d_i^t) = \kappa\overline{\pi}^s(1 - \overline{\pi}^s)\overline{\pi}^t(1 - \overline{\pi}^t)$

Multivariate quasi–likelihood

- When (Observed – Expected) terms within subjects

  $$(z_i^1 - \mu_i^1), (z_i^2 - \mu_i^2), \ldots, (z_i^t - \mu_i^t), \ldots$$

  are correlated , the weights given to them must also reflect this

  - correlation means that, to some extent, the same information is being repeated!

- The mathematical details of the calculation of the weights in the quasi–score functions need not concern us. It is sufficient to know that they depend on:

  - the slopes of the link function
  - the variances of the responses (as a function of their expected values)
  - the *covariances* between responses within clusters (again as a function of expected values)

- The quasi score function with these weights is also known as a *generalized estimating equation* (GEE).

PNG data – GEE

```
model      :       xtgee d, e(y) i(id) fam(pois) link(log)
exposure   :       vacc      (categorical)
modifier   :       ageband   (categorical)

Number of records used in the fit :  2520

Effects of vacc on the ratio scale

Level 2 versus level 1

Level or value
of ageband       Effect   95% Confidence Interval

0                0.9494   [ 0.840 , 1.073 ]
2                0.8764   [ 0.764 , 1.005 ]
4                1.0101   [ 0.815 , 1.251 ]
6                0.6226   [ 0.340 , 1.141 ]

Overall test for effect modification
chi2(  3)   =    3.207
P-value     =    0.361
```

The command **`xtgee`** uses an exchangeable correlation structure by default.

Following **xtgee** the correlations can be displayed.

```
. xtcorr

        c1      c2      c3      c4
r1  1.0000
r2  0.4122  1.0000
r3  0.4122  0.4122  1.0000
r4  0.4122  0.4122  0.4122  1.0000
```

Other structures can be used, but require a time variable. With the unstructured option:

```
. xtcorr

        c1      c2      c3      c4
r1  1.0000
r2  0.4527  1.0000
r3  0.4035  0.5691  1.0000
r4  0.2741  0.3023  0.4331  1.0000
```

```
. xtgee d care, fam(bin) i(id)

        d |      Coef.   Std. Err.
---------+----------------------
     care |   .3980332    .062634
    _cons |  -1.113458   .0214789

. xtgee d care, fam(bin) i(id) rob

          |              Semi-robust
        d |      Coef.   Std. Err.
---------+-----------------------
     care |   .3980332   .0697908
    _cons |  -1.113458   .0214465
---------------------------------
```

Correlation structures

```
. xtcorr


      c1     c2     c3     c4     c5     c6     c7
r1  1.00
r2  0.29   1.00
r3  0.29   0.29   1.00
r4  0.29   0.29   0.29   1.00
r5  0.29   0.29   0.29   0.29   1.00
r6  0.29   0.29   0.29   0.29   0.29   1.00
r7  0.29   0.29   0.29   0.29   0.29   0.29   1.00

. xtgee d care, fam(bin) i(id)  corr(uns) t(wave)
. xtcorr

      c1     c2     c3     c4     c5     c6     c7
r1  1.00
r2  0.32   1.00
r3  0.27   0.34   1.00
r4  0.26   0.30   0.36   1.00
r5  0.22   0.26   0.30   0.34   1.00
r6  0.23   0.25   0.26   0.31   0.35   1.00
r7  0.22   0.23   0.24   0.28   0.28   0.32   1.00
```

“Robust” SE’s

- If the variance is mis-specified, standard errors calculated from the inverse information will be wrong

- The robust SE generalizes easily to clustered data

- The score contributions, u_i^t, are correlated within subjects so the usual estimate of the score variance:
$$\sum_{i,t}(u_i^t)^2$$
does not work

- But, the score contributions of *subjects*, $u_i = \sum_t u_i^t$, are independent and we can use
$$V = \sum_i (u_i)^2$$
in the usual way

Placebo group only: $u_i^t = d_i^t - \lambda y_i^t$

id	period	d_i^t	u_i^t	$(u_i^t)^2$	u_i	$(u_i)^2$
1	1	5	−3.58	12.82		
1	2	3	−5.58	31.14		
1	3	3	−5.58	31.14		
1	4	3	−5.58	31.14	−20.32	412.90
2	1	3	−5.58	31.14		
2	2	5	−3.58	12.82		
2	3	3	−5.58	31.14		
2	4	3	−5.58	31.14	−20.32	543.82
3	1	2	−6.58	43.30		
3	2	4	−4.58	20.98		
3	3	0	−8.58	73.62		
3	4	5	−3.58	12.82	−23.32	412.90
etc						
			0.00	11935.28	00.00	33088.11

- Fitted value for λy is 961/112=8.58
- "Naive" SE for log(λ) is

$$\sqrt{1/961} = 0.032$$

- Robust SE of log(λ) taking account of clusters is

$$\sqrt{28/27}\sqrt{33088.11/(961)^2} = 0.192$$

```
. use epilep2, clear
. xi: poisson d if trt==0, e(y) clus(id)

         |               Robust
       d |      Coef.   Std. Err.
---------+-----------------------
   _cons |   1.456328   .1927568
```

The bootstrap

Erasmus, 2002

Inference based on a probability model

Likelihood theory tells us that the distribution of M in repeated samples from the model where θ is the true value of the parameter, is approximately Gaussian with mean θ and standard deviation $S = 1/\sqrt{I}$, where I is the information.

The SD of the sampling distribution is also called the SE of M.

In 90% of samples M will lie in the interval

$$\theta - 1.645S \ , \ \theta + 1.645S$$

which means that in 90% of samples the interval

$$M - 1.645S \ , \ M + 1.645S$$

will include θ. Thus $M \pm 1.645S$ is a 90% confidence interval for θ.

What to do now?

Generate (say) 1000 bootstrap samples from the original data

$$z_1, z_2, \cdots, z_N$$

by sampling *with replacement*. For each sample work out M and obtain

- The distribution of M in repeated samples.

- The SD of this distribution (S).

- Use $M \pm 1.645S$ if the distribution is Gaussian.

- Use the 5 and 95 percentiles of the distribution if not.

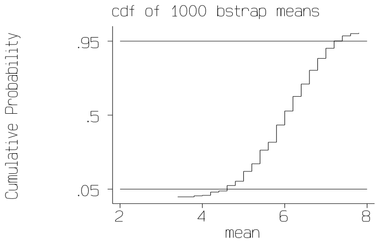
Model: z has distribution with mean μ

Original data: 3 7 5 8 7

Estimate of μ is $M = 6.00$ (the mean)

Random samples of size 5 with replacement

5 3 8 7 7 — $M = 6.0$
7 5 7 5 5 — $M = 5.8$
etc.



Bootstrap confidence limits

SD of 1000 bstrap means is 0.8395

90% confidence interval for μ is

$$6.00 \pm 1.645 \times 0.8395$$

which is from 4.6 to 7.4.

5/95 percentiles of the distribution of bootstrap means are 4.6 – 7.2

The percentiles are more reliable, but take more bootstrap samples than using the SD.

The PNG data

The log rate ratio comparing vaccinated subjects with those receiving placebo is (for all infections) -0.1003.

The SE of the log rate ratio, based on the Poisson likelihood, is $\sqrt{1/1204 + 1/1038} = 0.0423$

The SE of the log rate ratio based on the bootstrap is 0.0620.

The robust SE using the Poisson likelihood is 0.0609.

The larger value for the bootstrap and robust compared to Poisson reflects the subject heterogeneity.

Erasmus, 2002

Examples

- Many study designs in epidemiology deliberately collect *incomplete data*
- All variables are only collected in a subset; the remainder are incomplete
- Some examples:
  - Nested case–control studies
  - Case–cohort studies
  - 2–phase case–control studies
  - 2–phase prevalence studies
- We have seen that a likelihood based approach works for the first of these; the others are not so straightforward

Case–cohort studies

- Select a *sample* of the cohort and add to this *all* cases which arise during follow–up.
- This differs from the nested case–control design where controls are selected *independently* from the risk set for each case
- This has several advantages:
  - Often cohort studies have many disease "endpoints" of interest. The nested design creates a different set of controls for each endpoint while, in the case–cohort approach, the same controls can be used throughout
  - Controls can be sampled at start of study — this sometimes simplifies the logistics of studies, saving costly repeat visits

2–phase case–control studies

This is very similar to a case–cohort study, but the first phase of the study is a case–control study rather than a cohort study

- **Phase 1:** Draw cases and controls in the usual way
  - Some measurements may be difficult or expensive to collect
  - Instead we collect baseline variables, possibly plus *surrogates* for the remainder
- **Phase 2:** Collect remaining data in subsamples of cases and controls
  - It is usually advantageous to target the second phase using the surrogate measures

- For example, this is attractive for *rare exposures*:
  - **Either** use surrogate exposure measure to target definitive exposure measurements in phase 2 on subjects likely to be exposed,
  - **Or** independently subsample the 4 groups defined by

	Cases	Controls
Exposed		
Unexposed		

  for measuring confounders in phase 2

- Complete set of exposure and confounder measurements only available for the phase 2 subsample
- Sampling design of the phase 2 study must usually be taken into account in the analysis

Study

- Routine information system records hospital and a few baseline variables for all births
- Data are insufficiently detailed to provide much insight into the epidemiology of perinatal death
- Case–control design: For every perinatal death draw a control live birth. For logistic reasons it is convenient to use the next live birth on the hospital register
- Analysis as a conventional matched case–control study treats hospital as a confounder — rarely appropriate
- Better to treat the routine system as phase 1 and the case–control study as phase 2 of a 2–phase study

2-phase prevalence studies

- A design used when definitive measures of disease state are difficult or expensive to carry out at population level, e.g.
  - psychiatric diagnoses often require skilled assessment
  - some conditions may require detailed investigations such as serology, X–ray, etc. to confirm
- **Phase 1:** Collect explanatory variables, plus simple and inexpensive surrogate for the presence of disease
  - psychiatric inventories/scales
  - indicative symptoms (fever etc.)
- **Phase 2:** establish definitive diagnosis in a subsample
  - Usually advantageous to take larger sample amongst those likely to be affected

Inverse probability weighting

- A simple method of analysis is to try to reconstruct the data as it might have been had it been complete
- Analyse subjects with complete data (the "complete cases") but use *weights* to correct for the sampling design
- If π, which may depend on variables measured in phase 1, is the probability of selection for phase 2, we weight by $1/\pi$ — Eg. subjects with $\pi = 0.25$ selected for phase 2 are weighted by 4×
- Otherwise, use conventional methods — Poisson regression, logisic regression etc.
- "Naive" standard errors are wrong — use "robust" methods such as the information sandwich

Example: Case–cohort study

- Say we use a 10% control sample, plus *all cases*
  - $\pi = 1$ for cases ($d = 1$) and $\pi = 0.10$ for controls (non–cases in the control sample) ($d = 0$)
  - Weights are 1 (if $d = 1$) or 10 (if $d = 0$)
- Person–years observation in the cohort estimated by

$$Y = \begin{array}{c}\text{Person–years (Cases)}\\ + \\ 10 \times \text{Person–years (Non-case controls)}\end{array}$$

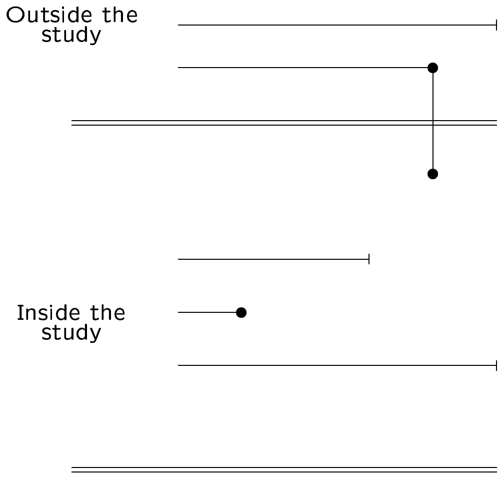
  and estimated rate is D/Y
- Generalizes to Poisson regression and Cox regression — using information–sandwich for SEs

- Inverse probability weighting works because, using correct weights, the score function has zero expectation at the true parameter values
- In Cox regression, score function for an explanatory variable, x, has the form

$$\sum_{\text{Risk sets}} \left(x_{\text{Case}} - \overline{x}_{\text{Risk set}}\right)$$

  where $\overline{x}_{\text{Risk set}}$ is the mean of x in the risk set, weighted by the fitted rate ratios
- In case–cohort study, we only have, say, 10% of non–cases in risk sets but all the cases → bias unless we use inverse probability weights
- An alternative is to include those cases not drawn in the original control sample *only in their own risk set* — again we must use a robust method for SEs

Prentice's method – pictorially



Estimating the weights?

- Consider a cohort of size 5000, from which we draw 500 controls
- Let 200 cases arise — 30 from the control sample and 170 from the rest of the cohort
- Using our knowledge of how the sample was drawn, for non–cases $\pi = 0.1$ and weight is 10
- Without this knowledge we would estimate π by 470/4800
- Counter–intuitively we do better by estimating the weights, but we should allow for this when calculating SEs — eg. by bootstrap, recalculating the weights for each bootstrap sample

	First phase		Second phase	
	Disease	Healthy	Case	Control
Male	14867	2971	249	250
Female	2944	2934	249	248

- Odds ratio for the sex effect is 14867 × 2934/2944 × 2971 = 4.987 from phase 1 and 249 × 248/249 × 250 = 0.992 from unweighted analysis of phase 2

- For male cases $\pi = 249/14867$ etc.

- With inverse probability weighting,

$$\text{Odds ratio} = \frac{\frac{14867}{249}249 \times \frac{2934}{248}248}{\frac{2944}{249}249 \times \frac{2971}{250}250} = 4.987$$

- To take account of the confounders use logistic regression with these weights — and robust estimates of SEs

A 2–phase prevalence studies

- A study of depressive illness
  - GHQ is the diagnosis from a general health questionnaire, and
  - GP is the diagnosis by the general practitioner (present or absent)
  - Definitive diagnosis (SCAN) shown as + or −

Sex	GHQ	GP	N_1	N_2	−	+	Weight
M	−	−	227	18	16	2	227/18=12.61
M	+	−	60	22	11	11	60/22=2.727
M	−	+	17	5	4	1	17/5=3.400
M	+	+	20	14	4	10	20/14=1.429
F	−	−	287	24	18	6	287/24=11.96
F	+	−	133	67	37	30	133/67=1.985
F	−	+	19	12	6	6	19/12=1.583
F	+	+	60	41	15	26	60/41=1.463

- Prevalence estimate is sum of weights for the cases (+) divided by the total sum of weights for all subjects — logistic regression generalization is straightforward

Efficiency

- Inverse probability weighting is a simple method, but may not be very efficient

- Maximum likelihood would be efficient, but would need us to specify a distribution model for all missing data

- This may be difficult, and we might misspecify such a model

- Recently a theory of *semi–parametric efficient* estimation has been developed, aimed at finding the best possible method which can be devised *without* specifying parametric models for the missing data

- Semi–parametric efficient estimates can be derived by an extension of the IPW approach