1 Probability models

1.1 Observation, experiments and models

Stochastic Models[1]

Normal vs *Bernoulli* and *Poisson*: We need to distinguish between *individual* observations, governed by Bernoulli and Poisson (or if quantitative rather than all-or-none or a count, Normal) and *statistics* formed by aggregation of individual observations. If a large enough number of individual observations are used to form a statistic, its (sampling) distribution can be described by a Gaussian (Normal) probability model. So, ultimately, this probability model is just as relevant.

1.1.1 Epidemiologic [subject-matter] models [JH]

We need to also make a distinction between the quantity(quantities) that is(are) of substantive interest or concern, the data from which this(these) is(are) estimated, the *statistical* models used to get to the the quantity(quantities) and the relationships of interest.

For example, of medical, public health or personal interest/concern might be the

- level of use of cell phones while driving
- average and range [across people] of reductions in cholesterol with regular use of a cholesterol-lowering medication
- amount of time taken by health care personnel to decipher the handwriting of other health care personnel
- (average) number of times people have to phone to reach a 'live' person
- reduction in one's risk of dying of a specific cancer if one is regularly screened for it.
- appropriate-size tracheostomy tube for an obese patient, based on easily easily obtained anthropometric measurements
- length of central venous catheter that can be safely inserted into a child as a function of the child's height etc.
- rate of automobile accidents as a function of drivers' blood levels of alcohol and other drugs, numbers of persons in the car, cell-phone and other activities, weather, road conditions, etc.
- Psychological Stress, Negative Life Events, Perceived Stress, Negative Affect Smoking, Alcohol Consumption and Susceptibility to the Common Cold
- The force of mortality s a function of age, sex and calendar time.
- Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction
- Are seat belt restraints as effective in school age children as in adults?
- Levels of folic acid to add to flour, so that most people have sufficiently high blood levels.
- Early diet in children born preterm and their IQ at age eight.
- Prevalence of Down's syndrome in relation to parity and maternal age.

Of broader interest/concern might be

- the wind chill factor as a function of temperature and wind speed
- how many fewer Florida votes Al Gore got in 2000 because of a badly laid-out ballot
- a formula for deriving one's "ideal" weight from one's height
- yearly costs under different cell-phone plans
- yearly maintenance costs for different makes and models of cars
- car or life insurance premiums as a function of ...
- cost per foot^2 of commercial or business rental space as a function of ...
- Rapid Changes in Flowering Time in British Plants
- How much money the City of New York should revover from Brink's for the losses the City incurred by the criminal activities of two Brink's employees (they collected the money form the parking meters, but kept some of it!).

[1] 'Stochastic' `http://www.allwords.com/word-stochastic.html` French: stochastique(fr) German: stochastisch(de) Spanish: estocstico(es) Etymology: From Ancient Greek (polytonic,), from (polytonic,) "aim at a target, guess", from (polytonic,) "an aim, a guess". Parzen, in his text on Stochastic Processes .. page 7 says: <<The word is of Greek origin; see Hagstroem (1940) for a study of the history of the word. In seventeenth century English, the word "stochastic" had the meaning "to conjecture, to aim at a mark." It is not clear how it acquired the meaning it has today of "pertaining to chance." Many writers use the expression "chance process" or "random process" as synonyms for "stochastic process."<<

1.1.2 From behaviour of statistical 'atoms' to statistical 'molecules'

1 condition' or 1 circumstance' or 'setting' [also known as "1-sample problems"]

The smallest statistical element or unit (?atom): its quantity of interest might have a Y distribution that under sampling, could be represented by a discrete random variable with '2-point' support (Bernoulli), 3-point support, $k-$point support, etc. or interval support (Normal, gamma, beta, log-normal, ...)

The *aggregate* or summary of the values associated with these elements is often a sum or a count: with e.g., a Binomial, Negative Binomial, gamma distribution. Or teh summary might be more complex – it could be some re-arrangement of the data on the individuals (e.g., the way the tumbler longevity data were summarized). This brings in the notion of "sufficient statistics".

More complex: t, F, ...

2 or conditions' or 1 circumstances' or 'settings', indexed by possible values of 'X' variable(s). Think of the 'X' variable(s) as 'covariate patterns' or 'profiles.'

unknown conditions or circumstances Sometimes we don't have any measurable (or measured) 'X' variable(s) to explain the differences in Y from say family to family or person to person.There instead of the usual multiple regression approach, we use the concept of a hierarchical or random-effects or latent class or mixture model.

1.2 Binary data

It is worth recalling from the first semester, the concepts of states and events (transitions from one state to another).

Cohort studies with fixed follow-up time

Recall: *cohort* is another name for a closed population, with membership (entry) defined by some event, such as birth, losing one's virginity, obtaining one's first driver's permit, attaining age 21, graduating from university, entering the 'ever-married' state, undergoing a certain medical intervention, enrolling in a follow-up study, etc. Then the *event of interest* is the *exit* (transition) from a/the state that prevailed at entry. So *death* is the transition from the *living* state to the *dead* state, receiving a *diagnosis* of cancer changes one's state from 'no history of cancer since entry' to 'have a history of cancer', being convicted of a traffic offense changes one's state from 'clean record' to 'have a history of traffic offenses.' We can also envision more complex situations, with a transition from 'never had a stroke,' to 'have had 1 stroke,' to 'have had 2 strokes,' ... or 'haven't yet had a cold this winter,' to 'have had 1 cold,' to 'have had 2 colds,' etc.

Censoring: to be distinguished from *truncation*. Truncation implies some observations are missed by the data-gathering process, i.e., that the observed distribution is a systematic distortion of the true distribution. Note that we can have censoring of any quantitative variable, not just one that measures the duration until the event of interest. For example, the limits on say a thermometer or a weight scale or a chemical assay may mean that it cannot record/detect values below or above these limits. Also, the example in C&H implicitly refers to *right* censoring: one can have *left* censoring, as with lower limits of detection in a chemical assay, or *interval* censoring, as – in repeated cross-sectional examinations – with the date of sero-conversion to HIV.

Incidence studies: the word *new* means a change of state since entry.

"*Failure*": It is a pity that C&H didn't go one step more and use the even more generic term "*event*". That way, they would not have to think of graduating with a PhD (i.e., *getting out of – exiting from – here*) as "*failure*" and still being here" as "*survival*." This simpler and more general terminology would mean that we would not have to struggle to find a suitable label of the 'y' axis of the $1 - F(t)$, usually called $S(t)$, function. One could simply say *"proportion still in initial state,"* and substitute the term for the initial state, i.e., proportion still in PhD program, proportion event-free, etc.

N or n? D or d? JH would have preferred lower case, at least for the denominator. In *sampling* textbooks, N usually denotes the *population* size, and n the *sample* size. In the style manual used in *social sciences*, n is the sample size in each stratum, whereas N is the overall sample size. Thus, for example, a study might report on a sample of $N = 76$ subjects, composed of $n = 40$ females and $n = 36$ males.

Cohort studies with variable follow-up time: If every subject entered a study at least 5 years ago, then, in principle, one should be able to determine D and $N - D$, and the 5-year survival proportion. However, *losses to follow-up* before 5 years, and before the event of interest, lead to observations that are typically regarded as censored at the time of the loss. Another phenomenon that leads to censored observations is *staggered entry*, as in the JUPITER trial. Unfortunately, some losses to follow-up may be examples of *informative*' censoring.

Cross-sectional prevalence data

Recall again that prevalence refers to a *state*. Examples would include the

proportion (of a certain age group, say) who wear glasses for reading, or have undetected high blood pressure, or have high-speed internet at home, or have a family history of a certain disease, or a certain 'gene' or blood-type.

From a purely *statistical* perspective, the analysis of *prevalence* proportions of the form D/N and *incidence* proportions of the form D/N takes the same form: the underlying statistical 'atoms' are N Bernoulli random variables.

1.3 The binary probability model

JH presumes they use this heading as a shorthand for 'the probability model for binary responses' (or 'binary outcomes' or binary random variables)

... to "*predict* the outcome" : JH takes this word *predict* in its broader meaning. If we are giving a patient the probability that he will have a certain *future* event *say within the next 5 years*, we can talk about predicting the outcome: we are speaking of *pro*gnosis; but what if we are giving a woman the probability that the suspicious finding on a mammogram does in fact represent an existing breast cancer, we are speaking of the *present*, of whether a phenomenon already *exists*, and we use a prevalence proportion as an estimate of the *dia*gnostic probability. Note that prevalence and incidence refer to aggregates.

The risk parameter

Risk typically refers to the *future*, and can be used when speaking to or about one person; we don't have a comparable specialized term for *the probability that a state exists* when speaking to or about one person, and would therefore just use the generic term probability.

The odds parameter

The sex-ratio is often expressed as an odds, i.e., as a ratio of males to females. If the proportion of males is 0.51, then the male:female ratio is 51:49 or (51/49):1, i.e., approximately 1.04:1. This example is a good reason why C&H should have used a more generic pair of terms than failure and survival (or success and failure).

In betting on horse races (at least where JH comes from), odds of 3:1 are the odds *against* the horse winning; i.e., the probability of winning is 1/4. When a horse is a heavy favourite so that the probability of winning was 75%, the "bookies" would give the odds as "3:1 *on*."

Rare events

One of the tricks to make events *rare* will be to slice the time period into small slices or windows.

Death, the first of the two only sure events (taxes is the other) is also rare - in the short term!

Also, it would be more correct to speak of a *rare events*, since disease is often used to describe a process, rather than a transition. And since most transitions are rapid, the probability of a transition (an event) occurring within a given short sub-interval will usually be small.

If the state of interest being addressed with cross-sectional data is uncommon (or rare), then yes, the prevalence odds and the prevalence proportion will be very close to each other.

Supplementary Exercise 1.1. If one rounds probabilities or risks or prevalences (π's), or their corresponding odds, $\Omega = \pi/(1-\pi)$, to 1 decimal place, at what value of π will the rounded values of π and Ω be differerent? Also, why use lowercase π for proportion, and uppercase Ω for odds?

1.4 Parameter Estimation

Should you be surprised if the estimate were π were other than D/N? Consult Google or Wikipedia on "the rule of succession," and on Laplace's estimate of the probability that the sun will rise tomorrow, given that it has unfailingly risen ($D = 0$) for the past 6000 years, i.e., $N \approx 365 \times 6000$.

Supplementary Exercise 1.2. One has 2 independent observations from the model

$$E[y|x] = \beta \times x.$$

The y's might represent the total numbers of typographical errors on x randomly sample pages of a large document, and the data might be $y = 2$ errors in total in a sample of $x = 1$ page, and $y = 8$ errors in total in a separate sample of $x = 2$ pages. The β in the model represents the mean number of errors per page of the document. Or the y's might represent the total weight of x randomly sample pages of a document, and the data might be $y = 2$ units of weight in total for a sample of $x = 1$ page, and $y = 8$ units for a separate sample of $x = 2$ pages. The β in the model represents the mean weight per page of the document. We gave this 'estimation of β' problem to several statisticians and epidemiologists, and to several grade 6 students, and they gave us a variety of estimates, such as $\hat{\beta} = 3.6$/page, 3.33/page, and 3.45!

How can this be?

1.5 Is the model true?

I wonder if they were aware of the quote, attributed to the statistician George Box that goes something like this

> "all models are wrong; but some are more useful than others"

`http://en.wikiquote.org/wiki/George_E._P._Box`

2 Conditional probability models

2.1 Conditional probability

JH is suprised at how few textbooks used trees to explain conditional probabilities. Probability trees make it easy to see the direction in which one is preceeding, or looking, where simply algebraic symbols can not, and make it easier to distinguish 'forward' from 'reverse' probabilities.

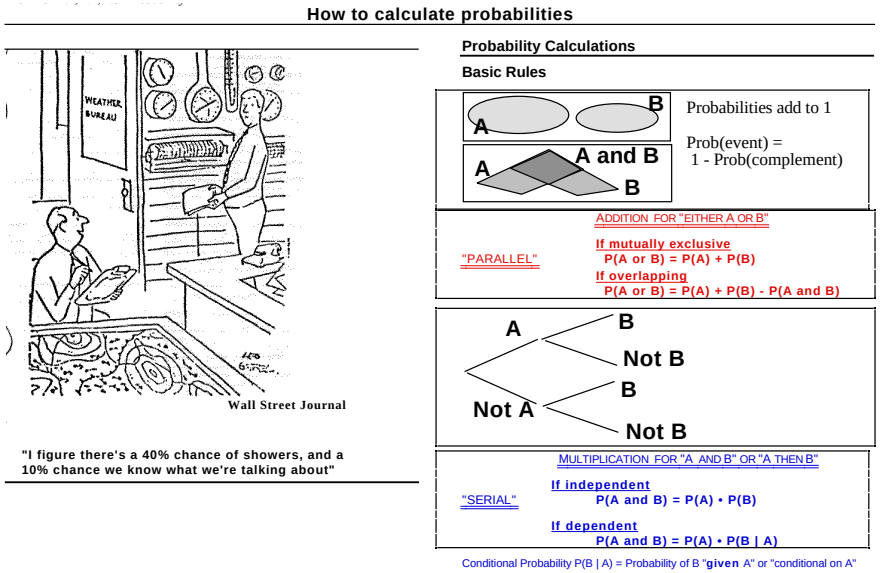

Figure 1: From JH's notes for EPIB607, introductory biostatistics for epidemiology

Trees show that the probability of a particular sequence is always a fraction of a fraction of a fraction .. , and that if we start with the full probability of 1 at the single entry point on the extreme left, then we need at the right hand side to account for all of this (i.e., the 'total') probability.

Statistical dependence and independence

JH likes to say that with independence, one doesn't have to look over one's shoulder to the previous event to know which probability to multiple by.. The illustrated example on the gender composition of 2 independent births, and of a sample of 2 persons sampled (without replacement) from a pool of 5 males and 5 females, show this distinction: in the first example, when one comes to the second component in the probability product, $Pr(y_2 = male)$ is the same

2.3.2 Twins: Excerpt from an article by Bradley Efron

Modern science and the Bayesian-Frequentist controversy

> Heres a real-life example I used to illustrate Bayesian virtues to the physicists. A physicist friend of mine and her husband found out, thanks to the miracle of sonograms, that they were going to have twin boys. One day at breakfast in the student union she suddenly asked me what was the probability that the twins would be identical rather than fraternal. This seemed like a tough question, especially at breakfast. Stalling for time, I asked if the doctor had given her any more information. "Yes", she said, "he told me that the proportion of identical twins was one third". This is the population proportion of course, and my friend wanted to know the probability that her twins would be identical.
>
> Bayes would have lived in vain if I didn't answer my friend using Bayes' rule. According to the doctor the prior odds ratio of identical to nonidentical is one-third to two-thirds, or one half. Because identical twins are always the same sex but fraternal twins are random, the likelihood ratio for seeing "both boys" in the sonogram is a factor of two in favor of Identical. Bayes' rule says to multiply the prior odds by the likelihood ratio to get the current odds: in this case 1/2 times 2 equals 1; in other words, equal odds on identical or nonidentical given the sonogram results. So I told my friend that her odds were 50-50 (wishing the answer had come out something else, like 63-37, to make me seem more clever.) Incidentally, the twins are a couple of years old now, and "couldnt be more non-identical" according to their mom.

Supplementary Exercise 2.1. Depict Efron's calculations using a probability tree.

Supplementary Exercise 2.2 Use a probability tree to determine the best strategy in the Monty Hall problem

```
( http://en.wikipedia.org/wiki/Monty_Hall_problem )
```

Supplementary Exercise 2.3 A man has exactly two children: you meet the *older* one and see that it's a boy. A woman has exactly two children; you meet *one* of them [don't know if its the younger/older] and see is a boy. What is the probability of the man's younger child being a boy, and what is the probability of the woman's "other" child being a boy?

3 Likelihood

"*We need a way of choosing a value of the parameter(s) of the model*" (1st paragraph): It is clear from the later text that they do not mean to give the impression that one is only interest in a single value or point-estimate. For any method to be worthwhile, it needs to be able to provides some measure of uncertainty, i.e. an interval or range of parameter values.

"In simple statistical analyses, these stages of model building and estimation may seem to be absent, the analysis just being an intuitively sensible way of summarizing the data." Part of the reason is that (as an example) a sample mean may simply seem like a natural quantity to calculate, and it does not seem to require an explicit statistical model. The mean can also be seen as the least squares estimate, in the sense that the sum of the squared deviations of the sample values from any other value than the sample mean would be larger than the sum of the squared deviations about the mean itself, i.e., the sample mean is a least squares estimate. But that purely arithmetic procedure still does not require any assumptions about the true value of the parameter value μ, or about the shape of the distribution of the possible values on both sides of μ. For the grade 6 exercise about the mean number of errors per page, it seemed to make sense to divide the total number of errors by the total number of pages; but what if the task was to estimate the mean weight of the pages? We discussed in class at least two different statistical models – that would lead to different estimates.

"In modern statistics the concept which is central to the process of parameter estimation is likelihood." Older and less sophisticated methods include the method of moments, and the method of minimum chi-square for count data. These estimators are not always efficient, and their sampling distributions are often mathematically intractable. For some types of data, the method of weighted least squares is a reasonable approach, and we will also see that iteratively-reweighed least squares is a way to obtain ML estimates without formally calculating likelihoods.

Likelihood is central not just to obtain frequentist-type estimators per se, but also to allow Bayesian analyses to combine prior beliefs about parameter values to be updated with the data at hand, and arrive at what one's post-data beliefs should be.

Likelihood provides a very flexible approach to combining data, provided one has a probability model for them. As a simple example, consider the challenge of estimating the mean μ from several independent observations for a $N(\mu, \sigma)$ process, but where each observation is recorded to a different degree of numerical 'rounding' or 'binning.' For example, imagine that because of

the differences with which the data were recorded, the $n = 4$ observations are $y_1 \in [4, 6),\ y_2 \in [3, 4),\ y_3 \in [5, \infty),\ y_4 \in [-\infty, 3.6)$. Even if we were told the true value of σ, the least squares method cannot handle this uni-parameter estimation task.

"*The main idea is simply that parameter values which make the data more probable are better supported than values which make the data less probable.*" Before going on to *their first example*, with a parameter than in principle could take any values in the unit interval, consider a *simpler example* where there are just two vlaues of π. We have sample of candies from one of two sources: American, where the expected distribution of colours is 30%:70% and the other Canadian where it is 50%:50%. In our sample of $n = 5$, the observed distribution is 2:3. Do the data provide more support for the one source than the other?

3.1 Likelihood in the binary model

Notice the level of detail at which the observed data are reported in Figure 3.1: not just the numbers of each (4 and 6) but the actual *sequence* in which they were observed. The Likelihood function uses the probability of the observed data. Even if we did not know the sequence, the probability of observing 4 and 6 would be ${}^{10}C_4 = 210$ times larger; however since we assume there is no order effect, i.e., that π is constant over trials, the actual sequence does not contain any information about π, and we would not include this multiplier in the Likelihood. In any case, we think of the likelihood as a function of π rather than of the observed numbers of each of the two types.: these data are considered fixed, and π is varied.. contrast this with the tail area in a frequentist p-values, which includes other non-observed values more extreme than that observed. Likelihood and Bayesian methods do not do this.

"$\pi = 0.5$ *is more likely than* $\pi = 0.1$" Please realize that this statement by itself could be taken to mean that we should put more money on the 0.5 than the 0.1. It does not mean this. in the candy source example, knowing where the candies were purchased, or what they tasked like, would be additional information that might in and of itself make one source more likely than the other. The point here is not to use terms that imply a prior or posterior probability distribution on π. The likelihood function is based just on the data, and in real life any extra prior information about π would be combined with the information provided by the data. It would have been better if the authors had simply said "*the data provide more support for* "$\pi = 0.5$ *than* $\pi = 0.1$." Indeed, I don't think "$\pi = 0.5$ *is more likely than* $\pi = 0.1$" is standard terminology. The terminology "0.4 is the ML estimate of π" is simpler and less ambiguous.

History: there is some dispute as to who first used the principle of ML for the choice of parameter value. The name of Gauss is often mentioned. The 1912 paper by Fisher, while still a student, is a nice clean example, and shows how

The usual reference is to papers by Fisher in the early 1920's, where he worked of many of the properties of ML estimators.

One interesting feature of the 1912 paper is that Fisher never defined the likelihood as a *product* of probabilities; instead he defined the log-likelihood as a *sum* of log-probabilities. This is very much in keeping with his summation of information over observations. indeed, there is a lot in his writings about choosing the most informative configurations at which to observe the experimental or study units.

3.2 Supported range

The choice of critical value is much less standardized or conventional than say the one for a significance test, or confidence level, or a highest posterior density.

Fig 3.4 (based on 20/50) vs. Fig 3.3 (based on 4/10): the authors don't say it explicitly, but the sharpness of the likelihood function is measured formally by the second derivative at the point where it is a maximum.

3.3 The log likelihood

The (log-)likelihood is invariant to alternative monotonic transformations of the parameter, so one often chooses a parameter scale on which the function is more symmetric.

3.4 Censoring in follow-up studies

3.5 Other fitting methods

We mentioned earlier that the method of least squares does not make an explicit assumption about the distribution of the deviations from or even that the observed data are a sample from a larger universe. Another older method, that does not make explicit assumptions about the variations about the postulated means, is the method of minimum chi-square. It was used for fitting simpler models for dose response data involving count data. This

minimum chi-square criterion does not lead to simple methods of estimation, or to estimators with easily derived sampling distributions. Nevertheless, it is one of the thee methods (the others are ML – which requires a fully specified model for the variations, and LS, that does not) used in the java applet `http://www.biostat.mcgill.ca/hanley/MaxLik3D.swf`. The applet allows you to fit a linear model to the above-described 2-point data, and to monitor how the log-likelihood, the sum of squared deviations, and the chi-square goodness of fit statistics vary as a function of the entertained values of β.

The applet shows that the LS method which measures lack of fit on the same scale that the y's are measured on (cf the two red lines). The min-X^2 method – applied to y's that represent counts or frequencies, is similar, in that the "loss function" is $\sum(y-\hat{y})/\hat{y}^2$. The criterion for the ML fitting of a Poisson model is very different, in that it is measured on the probability or log-probability scale, a scale that is shown in blue, and projecting out from the $x-y$ plane.

Under some Normal models with homoscedastic variation, the LS and ML methods give the same estimates for the parameter(s) that make up the mean. If $y|x \sim Normal(\mu_x, \sigma^2)$, then Lik $= \prod(1/\sigma)\exp[-\{(y_i-\beta x_i)^2/2\sigma^2\}]$. This is maximized when the exponentiated quantity is minimized. The minimization is the same one involved in the LS estimation.

Supplementary Exercise 3.1. Grouped Normal data (from Fisher's paper[6]).Three hundred observed measurement errors (ϵ's) from a $N(0,\sigma)$ distribution are grouped (binned) in nine classes, positive and negative values being thrown together as shown in the following table:-

Bin	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	All
Frequency (f)	114	84	53	24	14	6	3	1	1	300

Estimate σ^2 ...

1. as $(1/300)\sum f \times \epsilon^2_{mid}$. Note that we estimate it using a divisor of n rather than $n-1$, since we do not have to estimate μ : the errors are deviations from *known* values, so $\mu = 0$ (structurally).

2. Using Sheppard's correction for the grouping, i.e, by subtracting $w^2/12$, where w is the width of each bin, in this case 1. Incidentally, can you figure out why Sheppard subtracts this amount? Shouldn't grouping *add* rather than subtract noise?

3. Using the method of Minimum χ^2.

4. Using the method of Maximum Likelihood.

[6] On the Mathematical Foundations of Theoretical Statistics, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 222 (1922), pp. 309-368

3.6 Other Applications: exercises

3.6.1 2 datapoints and a model

One has 2 independent observations from the (no-intercept) model

$$E[y|x] = \mu_{y|x} = \beta \times x.$$

The y's might represent the total numbers of typographical errors on x randomly sampled pages of a large document, and the data might be $y = 2$ errors in total in a sample of $x = 1$ page, and $y = 8$ errors in total in a separate sample of $x = 2$ pages. The β in the model represents the mean number of errors per page of the document. Or the y's might represent the total weight of x randomly sample pages of a document, and the data might be $y = 2$ units of weight in total for a sample of $x = 1$ page, and $y = 8$ units for a separate sample of $x = 2$ pages. The β in the model represents the mean weight per page of the document.

We gave this 'estimation of β' problem $\{ (x, y) = (1, 2)$ & $(2, 8)\}$ to several statisticians and epidemiologists, and to several grade 6 students, and they gave us a variety of estimates, such as $\hat{\beta} = 3.6$/page, 3.33/page, and 3.45!

Supplementary Exercise 3.2

How can this be? The differences have to do with (i) what model they (implicitly or explicitly) used for the variation of each $y \mid x$ around the mean $\mu_{y|x}$ and (ii) the method of fitting.

1. From 1st principles derive both the LS and (if possible the) ML estimators of β when

   (a) $y \mid x \sim ???(\mu_{y|x})$
   (b) $y \mid x \sim Poisson(\mu_{y|x})$
   (c) $y \mid x \sim N(\mu_{y|x}, \sigma)$ [assume σ is known]
   (d) $y \mid x \sim N(\mu_{y|x}, \sigma^2 = x \times \sigma_0^2)$ [assume σ_0^2 is known]

2. Where possible, match the estimators with the various numerical estimates above.

3. One of the numerical estimates came from another fitting method, namely the (now seldom-used) method of Minimum Chi-square, which seeks the

value of β that minimizes $\sum \frac{(O-E)^2}{E} = \sum \frac{(y-\beta x)^2}{\beta x}$ in this example. Verify that the one remaining estimate of unknown origin is in fact obtained using this estimator.

See the (Flash) applet on `http://www.biostat.mcgill.ca/hanley/software/`

One of the messages of this exercise is that for one to use a likelihood approach, one must have a fully-specified probability model so that one can write the probability of each observed observation.

And, with different distributions of the y's around the mean $\mu_{y|x} = E(y|x) = \beta \times x$, the probabilities (and thus the overall likelihood, and its maximum, would be different.

3.6.2 Application: Distribution of Observations in a Dilution Series.

(Again, Text from Fisher's 1922 paper). An important type of discontinuous distribution occurs in the application of the dilution method to the estimation of the number of micro-organisms in a sample of water or of soil. The method here presented was originally developed in connection with Mr. Cutler's extensive counts of soil protozoa carried out in the protozoological laboratory at Rothamsted, and although the method is of very wide application, this particular investigation affords an admirable example of the statistical principles involved.

In principle the method consists in making a series of dilutions of the soil sample, and determining the presence or absence of each type of protozoa in a cubic centimetre of the dilution, after incubation in a nutrient medium. The series in use proceeds by powers of 2, so that the frequency of protozoa in each dilution is one-half that in the last. The frequency at any stage of the process may then be represented by

$$m = \frac{n}{2^x},$$

when x indicates the number of dilutions. Under conditions of random sampling, the chance of any plate receiving 0, 1, 2, 3 protozoa of a given species is given by the Poisson series

$$e^{-m}\left(1, m, \frac{m^2}{2!}, \frac{m^3}{3!}, \dots\right),$$

and in consequence the proportion of sterile plates is

$$p = e^{-m},$$

and of fertile plates

$$q = 1 - e^{-m}.$$

In general we may consider a dilution series with dilution factor a so that

$$\log p = -\frac{n}{a^x},$$

and assume that s plates are poured from each dilution. The object of the method being to estimate the number n from a record of the sterile and fertile plates, we have

$$L = S_1(\log p) + S_2(\log q),$$

when S_1 stands for summation over the sterile plates, and S_2 for summation over those which are fertile.

Supplementary Exercise 3.3 Estimate n from the following dilution series data:

Dilution:	0.25	0.5	1	2	4	8	16	32	64	128
Number of plates:	5	5	5	5	5	5	5	5	5	5
Number of fertile plates:	5	5	5	5	4	3	2	2	0	0

3.6.3 Application: Pooled testing:- old and new uses

The following excerpts are from a 1976 article "Group testing with a new goal, estimation", in Biometrika, 62, 1, p. 181 by authors Sobel and Elashoff. They begin by referring to the Dorfman, whose article, in the Annals of Mathematical Statistics, 1943, first used the ideas of group testing, with a binomial model, to reduce the number of medical tests necessary to find all members of a group of size N that have the syphilis antigen. They continued...

> Another aspect of the group-testing problem arises when one is interested not in the *classification of all the individuals* but in the *estimation of the frequency* of a disease, or of some property, when group-testing methods can be used. Given a random sample of size N, say, from a binomial population, the best estimate of the prevalence rate p, in the sense of minimizing the mean square error, will be obtained by testing each unit separately. However, if N is large and the tests are costly, then a different criterion, that includes testing costs, may indicate that group-testing designs should be used. We might expect benefits from group testing to increase as p decreases.
>
> [....] Example: Rodents are collected from the harbour of a large city, and, after being killed, dissected, etc., their liver is to be carefully examined under a microscope for the presence or absence of a

specific type of bacterium. The goal of the study is to estimate the proportion p of rodents that carry this bacterium using an economical experimental design. In this application the cost of obtaining the animals is negligible compared to the cost of testing, i.e. the microscopic search. It was proposed that an economical design to estimate p should be possible by combining in a single sample a small portion of the liver from each of several test animals and then carrying out a microscopic search on a homogeneous mixture of these liver portions. The problem is to find the best number, say A, of liver portions to combine and how to estimate the prevalence rate p from such a design. In addition, if this bacterial type is present in some particular tests, then the pathologists want to know whether they should carry out another test on a subset of these same animals or go on to test a new group of A animals.

[...] Thompson (1962) estimated the proportion of insect vectors capable of transmitting asteryellows virus in a natural population of the six-spotted leafhopper, an aphid. Instead of putting one insect with a previously unexposed aster test plant, he puts several insects with one test plant, for eoonomic reasons, and waits to see if the plant develops the symptoms of this virus. If it does, then at least one of these insects carried the virus; otherwise it is assumed that none carried it. The statistical problem is to choose an optimal number A of insects to be put with one test plant.

Contemporary uses: (can also Google *Minipool testing*)

The following text is an excerpt form Canadian Blood Services : Customer Letter #2005-18, 2005-05-17, entitled "Planned Measures to Protect the Blood Supply from West Nile Virus (WNV) - 2005 Season."

Dear Colleague:

West Nile season is approaching once again and this letter is to inform you about enhanced measures Canadian Blood Services has put in place to further protect the safety of the blood supply during the 2005 season.

For the summer of 2005, Canadian Blood Services will again use single-unit testing (SUT) to enhance the sensitivity of the West Nile Virus nucleic acid test. Minipool testing (6 samples/pool) is used throughout the year.

- In the summer of 2005, a 'trigger' will be used to initiate SUT. SUT will be initiated in a health region when a presumptive positive blood donor is detected using minipool testing, OR the prevalence of recent confirmed human cases in the preceding two weeks exceeds 1/1,000 population in rural areas, or 1/2,500 in urban areas.
- SUT will cease in a health region when there have been no positive donors for two weeks or the occurrence of WNV cases in the population falls below the aforementioned population triggers.

Supplementary Exercise 3.4 Suppose that in order to estimate the prevalence (π)of a characteristic in a population, one tests N randomly sampled objects by pooling them into n_b batches of size k (so that $N = n_b \times k$) and determining, for each batch, i.e. collectively, if at least one of its members is positive. Suppose that n_{b+} batches are found to be positive. Develop estimators of π using the method of moments, and using minimum χ^2 and Maximum Likelihood criteria.

3.7 Bayesian approach to parameter estimation

Given that the Bayesian approach is a very important and conceptually different way of making inference about the parameters of a model, and even though they mentioned Bayes rule in Chapter 2, it is surprising that Clayton and Hills do not make a statement about the Bayesian approach until Chapter 10; and even then, they do not give it much space. Maybe it's because they wanted the reader to become quite comfortable with Likelihood (which provides the Bridge between the prior and posterior distributions) before doing so.