Information in the Modern World

Murray Aitkin

March 17, 2008

Chapter 1

Sample surveys - should we believe what we read?

1.1 Women and Love

A major survey of sexual relationships of 4,500 US women was reported in the book *Women and Love - A Cultural Revolution in Progress*, by Shere Hite (1987).

Some of the statistics reported were staggering. For example:

- 95% of women reported emotional or psychological harassment from their husband or lover;
- 84% of women were not emotionally satisfied by their relationships;
- 70% of women married five years or more were having sex outside the marriage;
- 39% of women married 25 years or more had been struck or beaten by a husband or lover;
- 39% of women never married had been struck or beaten by a husband or lover.

What are we to make of these statistics?

That depends on how they were obtained. We will look at the sampling method used in this study, and in other studies. The sampling method has a critical effect on what we think about the study, and we will discuss the fundamental concept of “randomness” and its use to achieve a representative sample. This requires an understanding of simple probability concepts.

1.2 Would you have children?

The book *Statistics: Concepts and Controversies*, by David S. Moore, gives a simpler survey example. *Newsday*, a Long Island (NY) newspaper, runs an advice column by Ann Landers. Readers were asked to answer this question:

- If you had your life to live over again, would you have children?

Readers were asked to send a postcard, with a Yes or No answer. *Newsday* received nearly 10,000 responses, of which almost 70% said **No**. What should we conclude?

Newsday decided to commission a professional nationwide random sample of parents to ask the same question. The sample polled 1373 parents, and found that 91% said **Yes**. What should we conclude?

How can we be sure that a sample is "representative" of the population we aim to investigate?

"Representative" sampling requires that *every member of the population has a known chance of being included in the sample*, and that *this chance does not depend on the response being measured.* A sampling method that satisfies these requirements is called *unbiased*; one that does not is called *biased.*

How is unbiasedness achieved? In the most formal way, by constructing a list of all the members of the population, and then drawing the required sample size using a *random* mechanism to guarantee the known (frequently equal) chance of inclusion.

What do we conclude about the two sampling methods used? The second nationwide survey closely approximates this requirement. Survey organisations maintain population lists of families and residences based on the Census or on voter registration, and have standard random mechanisms for selecting the sample from the list.

The *Newsday* poll fails this requirement because

- we do not know what population is being sampled (readers of *Newsday*?)
- we do not know the chance of inclusion of each population member in the (voluntary) sample.

This is a characteristic feature of *voluntary response* in general.

On emotional issues, it is common to find that the chance of inclusion in the sample depends on the strength of feeling on the issue, and may be different for those with different views on the issue.

1.2.1 Bias in the *Newsday* sample

Suppose that the population size is 2,000,000. This assumption is not necessary – we would get the same result with any other population size assumption – but it makes the calculation simple. Suppose there are really 90% **Yes** parents and 10% **No** parents in the population – 1.8 million and 200,000 respectively.

Suppose our sample of exactly 10,000 contained 3,000 **Yes** and 7,000 **No** parents. Then the chance of being included in the sample was 3,000/1.8 million = 1/600 for the **Yes** parents, but 7,000/200,000 = 1/30 approximately for the **No** parents.

So a **No** parent had 20 times the chance of a **Yes** parent of being included in the sample – we have *oversampled* the **No** responders by a huge factor.

The sample is grossly biased towards **No** responders.

1.2.2 Bias in the *Women and Love* sample

What sampling method was used for *Women and Love*? Hite sent out questionnaires to

- “church groups in 34 states,
- “women’s voting and political groups in nine states,
- “women’s rights organisations in 39 states,
- “professional women’s groups in 22 states,
- “counselling and walk-in centres for women or families in 43 states,

“and a wide range of other organisations, such as senior citizen’s homes and disabled people’s organisations, in various states.

“In addition, individual women wrote for copies of the questionnaire...” (p.777) “All in all, 100,000 questionnaires were distributed, and 4,500 returned...” (p.777)

This is a response rate of 4.5%. What should we conclude? What population is being sampled?

The implication of the book is that the *target* population is the US female population. But the sampling method is very likely to oversample groups with higher proportions of women with relational difficulties. Since the questionnaires were sent to groups, we have no idea who actually answered them (they were anonymous).

The sample is clearly biased because we cannot give the chance of any woman in the US population being included in the sample.

If the sample is biased, what conclusions can we draw? Since the sampled population is undefined, we can only regard the sample AS the population. Hite has information from 4,500 women, and the percentages reported are the percentages *in her sample, which IS the population.*

These results have no knowable connection with percentages of the US female population, and cannot be used to refer to this population. To make such a connection, we would need a probability sample, in which every woman would be included with a known probability.

1.2.3 Difficulties with incomplete data

Hite's questionnaire introduced a further difficulty. The questionnaire directed respondents:

"*It is not necessary to answer every question!* There are seven headings; feel free to skip around and answer only those sections or questions you choose..." (p. 787) So the sample size may be different for different questions – even the percentages responding on different questions may not be comparable within the study:

- 95% reported emotional or psychological harassment from their husband or lover; but
- 84% of women were not emotionally satisfied by their relationships;

So it appears that between 11% and 16% of women were harassed by their husbands or lovers but were nevertheless emotionally satisfied!

These numbers surely are a consequence of non-response to one or other question. We gain a misleading impression even within the survey by comparing percentages based on different subsets of respondents.

Chapter 2

Data bases and sampling

2.1 Data bases

Data bases are large collections of data on populations, or important subpopulations, which can be used to inform policy on social, economic and political questions. Many administrative data bases (data collected as part of government or other official requirements) are now being used for this purpose, and very extensive data bases are being assembled by companies to assess the characteristics of people buying their products.

Since the data bases are often very extensive, *samples* are taken from the data bases, and conclusions about properties of the population in the data base are to be drawn from these samples. *Statistical theory* is the theory of how to relate sample properties to the population properties of interest. Statistical theory is based fundamentally on *probability theory*, and in this short course we will develop a small part of probability theory, and of statistical theory, to address the relation of sample to population in the simplest case.

For this purpose we will use the StatLab data base, from the book *StatLab: An Empirical Introduction to Statistics*, by Hodges, Krech and Crutchfield (1975). This book was one of the first to use a database to teach probability and statistical theory, and we will use its sampling method and (public access) database, though our approach to probability and statistical theory is rather different.

2.2 The StatLab database

The following description of the data base comes from pp. 318-319 of the StatLab book.

The StatLab database [called Census in the book] covers 1296 member families of the Kaiser Foundation Health Plan (a prepaid medical care programme) living in the San Francisco Bay area during the years 1961-72. These families were participating members of the Child Health and Development Study conceived and directed by Professor Jacob Yerushalmy, in the School of Public Health at the University of California, Berkeley.

On her first visit to the Oakland hospital of the Health Plan after pregnancy was diagnosed, each woman was interviewed intensively on a wide range of medical and socioeconomic matters relating both to herself and to her husband. In addition, various physical and physiological measures were made. When her child was born, further data about her and her newborn baby were recorded. Approximately 10 years later the child and mother were called in for follow-up testing, interviewing and measurement. In some instances, the husband was also interviewed and measured.

The 1296 families of the Census are divided into two equal subpopulations: 648 families consisting of a mother, father and female child; and 648 families of a mother, father and male child. The children were all born in the Kaiser Foundation Hospital, Oakland California, between 1 April 1961 and 15 April 1963. The Census does not cover any other children who may have existed in these families.

From the available data, 32 variables were selected for the Census. The 36 pages of the Census list each of these 32 variables for each of the 1296 families. The first 18 pages cover the families with girls; the second 18 pages cover the families with boys. Within each of these two sets of pages the families are listed in order of mother's age, with the youngest mothers first and the oldest last.

The Census consists of printouts numbered in consecutive dice numbers (i.e. the Census pages are numbered 11, 12, 13, 14, 15, 16, 21, 22, 23,...,65, 66). Similarly, the 36 families on each page are designated in consecutive dice numbers from 11 to 66. The identification number (ID no.) for any given family consists of two pairs of dice numbers, the first pair indicating the page and the second pair indicating the family on the page. To select a family purely at random from the population of 1296, it is necessary to throw a pair of dice twice. [The book was sold with a pair of dice, one red, one green.] If, for example, the first throw gives a red 2 and a green 6, this selects page 26. If the second throw gives a red 5 and a green 4, this selects family 54 on that page. Thus the ID number for this family is 26-54.

The 32 variables for each family are grouped by *child, mother, father*, and *family*. Part of the data were collected at the time of birth (1961-63) and the other data at the time of test (1971-72). The description and codes for each of the variables are given on the handout.

2.3 Your StatLab samples

We will be generating samples of different sizes from the data base to examine a number of questions about the population. In this short course we will restrict ourselves to three questions:

- Do mothers who were smoking at the diagnosis of pregnancy have babies with lower birthweight than mothers who were not smoking at diagnosis? [Low birth weight increases risk for babies.]

- Do mothers who were smoking at pregnancy diagnosis tend to have husbands who were also smokers?
- Do babies with low birth weight tend to have lower intelligence, when measured as children at age 10?

There are many other social/economic/medical questions which could be addressed using this data base - that is one of the values of such data bases.

The answers to these questions will be given at the end of the course, from the full population. But you will be trying to answer these questions from the random samples you have drawn. The sampling process is to generate a random sample of 40 families from the database, using the dice-throwing process described above. The family ID number, and the variables which we are going to use, are to be filled in on the data sheet supplied. This is set up to provide two sets of 10 families on the front and back of the data sheet. We will be using this structure to provide four samples of size 10, two samples of size 20, and one sample of size 40. The multiple samples of sizes 10 and 20 will show the extent of the variation between samples for each student, and will also show the variation among students in the results they obtain: this is a critically important issue in statistical theory.

Chapter 3

Probability

3.1 Representative sampling

Many students are surprised to find that, in their four samples of 10 families, they did not get exactly five boys and five girls. Since we know that the population contains equal numbers of boy- and girl-families, it seems reasonable that a "representative" sample should also have equal numbers of boys and girls.

However, our "representative" guarantee of "equal chnce" for each family to be included in the sample does not guarantee an *exact match* in proportions of boys and girls between the sample and the population.

In my own four samples of 10 families, I found 4,3,5 and 6 boys. In the pooled samples of 20 families, there were 7 and 11 boys, with 18 boys in the complete sample of 40 families. This kind of *random variation* is constantly encountered in dealing with random samples from populations. To describe this variation, we need to develop *probability models*.

We consider the sampling process step by step. We need some notation. We denote by B the event of drawing a boy family in the throw of the two dice. (It is actually only the first die that matters, since the girl families are 11-36 and the boys 41-66.) Since all pages in the data base are equally likely to be selected, the *probability* p that a boy family is selected is $18/36 = 1/2$.

Formally, if there are N *equally likely* possible outcomes, and R of these correspond to the event A of interest, then the probability p of the event A, written $\Pr[A]$ is $p = R/N$. It follows immediately that:

- if $\Pr[A] = 0$, then A *cannot occur*, or A *is impossible*;
- if $\Pr[A] = 1$, then A *is certain to occur.*

Less formally,

- if $\Pr[A]$ is *small*, say $\Pr[A] < 0.05$, then we say that A is *very unlikely* to occur.
- If $\Pr[A]$ is *large*, say $\Pr[A] > 0.95$, then A is *very likely* to occur.

However, there is a circularity in this “equally likely” definition of probability, because we have to know what “equally likely” means to use it. The *assumption* of “equally likely” outcoms when throwing a die is a *model assumption* that the die is perfectly symmetrical and balanced, so that no face is favoured to show more than any other face. We make this assumption in the absence of any evidence to the contrary, but if we carry out very large-scale studies of dice, by throwing them a very large number of times, we may be able to identify small departures from equal frequency of the six faces.

So the probability of the event B_1, that we draw a boy family at the first throw of the dice, is

$$p = \Pr[B_1] = 1/2,$$

and correspondingly

$$\Pr[G_1] = 1 - p = 1/2.$$

Now we throw the dice again to draw a second family. The family we chose at the first draw remains in the population and could be drawn again, though that would be very unlikely – its probability, by the same argument, would be 1/1296. Sampling the population in this way is called *sampling with replacement.* (What would happen if we *did* draw the same family again? We would set it aside and draw another one. Practical surveys are always drawn *without replacment*, but the two methods have very similar properties if the population is large compared to the sample.)

Since the population has not changed, and we assume that the outcome of the second throw is independent of that at the first throw, the probability of a boy family at the second throw is again $p = 1/2 = \Pr[B_2]$, and $\Pr[G_2] = 1-p = 1/2$. What possible samples of 2 families could we have? We denote them by $B_1B_2, \quad B_1G_2, \quad G_1B_2, \quad G_1G_2$.

What are the probabilities of these possible outcomes? Because of the *independence assumption*, we can multiply together the probabilities of the separate events:

$\Pr[B_1B_2]$	$p.p = p^2$	two boys
$\Pr[B_1G_2]$	$p.(1-p)$	one boy, one girl
$\Pr[G_1B_2]$	$(1-p).p$	one boy, one girl
$\Pr[G_1G_2]$	$(1-p).(1-p) = (1-p)^2$	two girls

In terms of the number r of boy families in the sample, we have:

Number of boys r	0	1	2
Probability	$(1-p)^2$	$2p(1-p)$	p^2
At $p = 1/2$	1/4	1/2	1/4

This array of the number of boy families and their probabilities is called a *probability distribution*, of the number R of boy families *in a sample of* $n = 2$ *families* from the STATLAB population.

Event			Probabilities	r Boys	Probability of r boys
$B_1B_2B_3$			p^3	3	p^3
$B_1B_2G_3$	$B_1G_2B_3$	$G_1B_2B_3$	$p^2(1-p)$	2	$3p^2(1-p)$
$B_1G_2G_3$	$G_1B_2G_3$	$G_1G_2B_3$	$p(1-p)^2$	1	$3p(1-p)^2$
$G_1G_2G_3$			$(1-p)^3$	0	$(1-p)^3$

We can extend this by direct enumeration to any sample size, though this quickly becomes tedious. For $n = 3$, we have possible samples:

At $p = 1/2$, the probabilities of 3, 2, 1, 0 boys are $1/8, 3/8, 3/8, 1/8$.

These probability distributions are particular cases of a general family, the *binomial distribution* (binomial = two names). For a general n and p, this gives the probability of r "success" events in n "trials" with "succes"s probability p, as

$$\Pr[r\,\text{successes} \mid n, p] = \binom{n}{r} p^r (1-p)^{n-r},$$

where $\binom{n}{r}$ is the *binomial coefficient* representing the number of *arrangements* of the r "successes" and $n - r$ "failure"s.

For the sample sizes $n = 10, 20$ and 40, and $p = 1/2$, the binomial distributions are:

Table 3.1: binomial distribution, $n = 10, p = 1/2$

r	0	1	2	3	4	5	6	7	8	9	10
$\Pr[r]$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

Table 3.2: binomial distribution, $n = 20, p = 1/2$

r	3	4	5	6	7	8	9	10
$\Pr[r]$	.001	.005	.015	.037	.074	.120	.160	.176
r	11	12	13	14	15	16	17	
$\Pr[r]$	.160	.120	.074	.037	.015	.005	.001	

3.2 Properties of these distributions

- The probability of an exactly equal split of the sample between boy and girl families decreases with the sample size n, from 0.246 – almost 1/4 – at $n = 10$ to 0.125 – 1/8 at $n = 40$. Most samples will *not* have equal numbers of boy and girl families.

- The probability of having between 40% and 60% of the sample as boy families increases with n.

A sample of 160 (well beyond our sample sizes) is *almost certain* to have between 40% and 60% boy families in the sample.

Table 3.3: binomial distribution, $n = 40, p = 1/2$

r	12	13	14	15	16	17	18	19	20
Pr[r]	.005	.011	.021	.037	.057	.081	.103	.119	.125
r	21	22	23	24	25	26	27	28	
Pr[r]	.119	.103	.081	.057	.037	.021	.011	.005	

n	r values	Probability
10	4– 6	0.656
20	8–12	0.736
40	16–24	0.845
80	32–48	0.943
160	64–96	0.991

- Large samples provide more precise information than small samples : as the sample size increases, we obtain more precise information about the population.

Chapter 4

Statistical inference

In sampling boy families, we *know* the proportion of boy families in the population, and we can therefore give the probability of any particular number of boy families in the sample.

But the point of sampling is to obtain information about populations and their properties that we *don't* know. For example:

- What proportion of mothers in the STATLAB population were smoking when their pregnancy was diagnosed?

Check your samples of 10, 20 and 40. What are we able to say about the population proportion from these sample results?

The theory of *probability* deals with the properties of (unobserved) *samples which can be* drawn from *known populations.*

The theory of *statistical inference* deals with the properties of *unknown populations* from the (observed) samples *which have been* drawn from them.

4.1 Simple approaches

In my samples of 10, I found 4, 3, 5 and 6 boys. It seems obvious that we should *estimate* the proportion of boy families in the population by the proportion in the sample, the sample proportion of boy families.

In my samples of 10, this proportion is 0.4, 0.3, 0.5 and 0.6. In the combined samples of 20, it is 0.35 and 0.55. In the full sample of 40, it is 0.45. We know the true proportion is 0.5. Our sample estimates vary around the true value – this is an inherent feature of random sampling.

For the mothers in my sample of 40, 13 were smoking at the pregnancy diagnosis, a sample proportion of 0.325. Of the fathers, 22 were smoking, a sample proportion of 0.55. What can we say about the population proportions of mothers and fathers who were smoking?

4.2 Binomial distribution tables

The tables of the binomial distribution for $n = 10$ and $n = 20$ which I have given you allow us to get some feeling for what our population and samples could be like. (The table for $n = 40$ is too extensive for one page.)

The tables are read *downwards* to find the probabilities of each number of "successes", given the success probability p. So for $p = 0.5$, reading down the column gives the probabilities of $0, 1, 2, ..., 9, 10$ "successes", as $.001, .010, ..., .010, .001$ as we previously described. In the population the proportion of boy families is 0.5, so the probabilities of $3, 4, 5$ or 6 boy families in the sample of 10 is $.117, .205, .246$ and .205. Drawing a sample of $n = 10$ with $r = 1$ or 9 boy families is very unlikely (probability 0.010) *relative to* the probability of 3, 4, 5, 6 or 7. No-one in the last class drew such a sample but it *can* happen, though rarely, in a large class.

In a sample of $n = 20$ from $p = 0.5$, 5 or fewer boy families, or 15 or more, have correspondingly low probabilities. Again no-one in the class had such a sample -- most people had between 7 and 13 boy families.

So in random sampling from a population, the observed *sample proportion* of "successes" *varies* (across different random samples) around the true population proportion, with a variability which *decreases* with increasing sample size. An "estimate" of the population proportion is the sample proportion, but its variability has to be quantified – expressed through probability.

Now we consider the proportion of the mothers smoking at pregnancy diagnosis. My samples of $n = 10$ had 4, 3, 3 and 4 smoking mothers. How do we use the sample value to draw conclusions about the population proportion?

We use the binomial tables again, but now p is unknown - we know only the number of successes r. For $r = 4$, we read *across* the table – we can find the probability of 4 successes for each tabled value of p.

The first entry on this line is .001 at $p = .05$. If the true proportion of smoking mothers was $.05 = 1/20$, it would be *very* unlikely to find 4 smoking mothers in a sample of 10, a sample proportion of 40%. At the other end of the range, if $p = 0.85$ the probability of 4 smoking mothers is again .001 – very unlikely.

The *maximum* probability of 4 smoking mothers occurs at $p = 0.4$ – it is 0.251. Not surprisingly, *what we observe in the sample has the highest probability when the population proportion is the same as the sample proportion*!

However other values of p also give high probability – near the maximum – to $r = 4$. At $p = 0.35$ or 0.45, the probability of $r = 4$ is 0.238, so these values of p are also very *plausible.*

4.3 Confidence intervals

How do we decide what values of p could have led to our observed sample proportion, and which could not? In statistical theory we use the *ratio*, denoted by RL, of the probability at p to its *maximum possible value*; small values of this ratio make implausible the corresponding value of p.

How small a value of the ratio RL corresponds to an implausible value of p? This is a partly arbitrary decision; again in statistical theory it is *conventional* to consider the value of p to be implausible if the ratio RL is less than 0.15 – we would be unlikely to obtain the sample if this value of p were the population value.

At $r = 4$, the maximum probability is 0.251 at $p = 0.4$. For what values of p is the ratio $RL = \Pr[r = 4 \mid p]/\Pr[r = 4 \mid 0.4]$ less than 0.15? If we divide the row entries for $r = 4$ by the maximum probability 0.251, we have

p	.05	.10	.15	.20	.25	.30	.35	.40	
RL	.004	.044	.160	.351	.582	.798	.948	1.0	
p	.45	.50	.55	.60	.65	.70	.75	.80	.85
RL	.950	.818	.636	.444	.275	.147	.065	.022	.005

Values of p less than 0.15 or more than 0.7 (approximately) have relative probabilities of 4 successes of *less than* 0.15. A finer tabulation in steps of 0.001 in p shows that values of p of 0.147 and 0.698 give relative probability 0.15. So values of p in the range (0.147, 0.698) have probabilities of 4 successes which are *more than* 0.15 of the maximum probability. A simple formula provides an easy way of obtaining this range approximately without needing the tables (which become very large for larger n) or a computer. The range can be calculated using only the observed proportion $\hat{p} = r/n$ and n: the end-points of the range are

$$\frac{\hat{p} + \lambda^2/(2n)}{1 + \lambda^2/n} \pm \lambda \frac{\sqrt{\hat{p}(1 - \hat{p})/n + \lambda^2/(4n^2)}}{1 + \lambda^2/n}.$$

where $\lambda = 2$. With $\hat{p} = 0.4$ and $n = 10$, this gives the interval approximately as $(0.165, 0.692)$, which agrees reasonably well with the exact range $(0.147, 0.698)$. In the conventional language of statistical theory this interval of values of p is called a *95% confidence interval*. (The term *confidence* comes from the observation that with many students generating this interval from their independent samples, only about 95% of the intervals will cover the true value of p – the other 5% will not.)

The samples with $r = 3$ give the interval $(0.106, 0.609)$. These intervals for $n = 10$ are very wide – from a sample of 10 we learn very little about p. What about $n = 20$? In my two samples of 20 I had 7 smoking mothers in each. We read the binomial table for $n = 20$ and $r = 7$. The maximum probability is .185 at $p = 0.35$, corresponding to the sample proportion $\hat{p} = 0.35$. Dividing the row values by 0.185 give this table:

The value 0.15 of the relative probability occurs between $p = .15$ and .2, and between .55 and .6. A finer tabulation gives the values 0.169 and 0.567. The formula gives $(0.179, 0.571)$. This interval is much shorter than that for $n = 10$.

For my full sample of $n = 40$ with $r = 14$ ($p = 0.35$ again), the formula gives $(0.219, 0.508)$ and the exact interval is $(0.216, 0.503)$, which is again shorter.

p	.10	.15	.20	.25	.30	.35	.40
RL	.011	.087	.296	.610	.891	1.0	.900
p	.45	.50	.55	.60	.65	.70	
RL	.662	.401	.199	.079	.024	.006	

Note that the *length* of the interval is nearly proportional to $1/\sqrt{n}$ – a sample four times as large ($n = 40$) has an interval approximately half the length of that for $n = 10$ (.287 vs 0.551).

4.4 Formal Theory

The binomial distribution gives the *probability* of r successes in n trials with success probability p at each trial:

$$\Pr[r\,\text{successes}|n,p] = \binom{n}{r} p^r(1-p)^{n-r},\ r = 0, 1, ...n.$$

Probability theory enables us to make *predictive statements* about the *relative frequency* or *proportion* of events which we will observe in repeated "experiments" of performing the trials. It gives a mathematical representation of the variability observed when we draw different random samples from the same population.

The probability above, written $\Pr[r \mid n, p]$ – the probability of r *given* n and p – is a *function* of the *argument* r – as r takes all its possible values $0, 1, 2, ..., n, \Pr[r \mid n, p]$ takes its probability values, which sum to 1 over all values of r, since *one* of them must occur.

When the sample has been drawn, we have now observed the r value – *we have the data.* The value of p is unknown, and we want to use the observed value of r to draw conclusions about it. Now we have to consider all the probability distributions for r for *different* values of p which give appreciable probability to what we have observed. For this purpose we define the *likelihood function*, which is formally identical to the probability above, but is *a function of* p*, not of* r:

$$L(p \mid n, r) = \binom{n}{r} p^r(1-p)^{n-r},\ 0 \le p \le 1.$$

This definition formalises our reading of the binomial tables across the row instead of down the column – each row of the table is the likelihood function of p for the given number of successes r.

The value of p which *maximises* the likelihood function is $\hat{p} = r/n$; this is called the *maximum likelihood estimate* of p.

The *ratio* of the likelihood to its *maximum* is called the *relative likelihood function*, written

$$RL(p) = L(p \mid n, r)/L(\hat{p} \mid n, r).$$

(We use the same notation RL as before, but it is now represented as an *explicit* function of p.) This has a maximum value of 1 at $p = \hat{p}$ – it is just a *rescaling*

– a *scale change* of the likelihood function. Our conclusions about the true population value of r are based on the function $RL(p):$ if this is "small" for some values of p – that is, much less than 1 – then these values of p are "implausible", in the sense of giving relatively low probability to what we observed. This is formalised in the idea of a *confidence interval* – an interval of p which contains the "plausible" values of p, that is, those with *high relative likelihood.*

The figures show the relative likelihoods for my samples, for $r = 3, 4, 5$ or 6 successes (boy families) out of $n = 10$, for $r = 7$ and 11 boy families out of $n = 20$, and for $r = 18$ boy families out of $n = 40$. The straight line is drawn at a relative likelihood of 0.15; this gives an interval of p with *95% confidence coefficient*; this term is used because, if many such intervals are constructed from different random samples, approximately 95% of them will cover the true population value, while 5% will not. For greater confidence we need a smaller relative likelihood than 0.15; a relative likelihood value of 0.037 will give *99% confidence* in the intervals constructed – only 1% of these intervals will not cover the true value.

With modern computing software these intervals are easily calculated on a computer. Most text books and practical users however rely on a simpler approximation to the 95% interval than the one we have given; the simpler approximation is

$$\hat{p} \pm 2\sqrt{\hat{p}(1-\hat{p})/n}.$$

This interval is symmetric about $\hat{p}$ and is easily calculated on a hand calculator. Here are the exact, the simple and the more accurate approximate 95% intervals for my sample results for the proportion of smoking mothers:

Table 4.1: 95% confidence intervals for p

r	n	Exact	Accurate approx.	Simple approx.
3	10	(0.085, 0.605)	(0.106, 0.609)	(0.010, 0.590)
4	10	(0.147, 0.698)	(0.165, 0.692)	(0.090, 0.710)
7	20	(0.169, 0.567)	(0.179, 0.571)	(0.137, 0.563)
14	40	(0.216, 0.503)	(0.219, 0.508)	(0.199, 0.501)

The simple approximation is quite inaccurate for small n and interval endpoints near 0 or 1. The "accurate" approximation is much better but is not very accurate for $n = 10$.

4.5 Derivation of the approximate confidence interval

How do we know theoretically that the approximate confidence interval works? It is based on a *mathematical approximation* to the likelihood function, or actually to the *natural logarithm* of the likelihood function. We have

$$L(p) = \binom{n}{r} p^r (1-p)^{n-r}$$

for r successes in n trials. Taking the natural log, we have the *log-likelihood function*

$$\ell(p) = \log_e L(p) = \log \binom{n}{r} + r\log(p) + (n-r)\log(1-p).$$

We *approximate* the log-likelihood by a *quadratic* function of p, of the form

$$\ell^*(p) = a \cdot (p-\hat{p})^2 + b,$$

where a and b are determined so that the two functions, and their first and second derivatives, agree at $p = \hat{p}$. At $p = \hat{p}$,

$$\begin{aligned}
\ell(\hat{p}) &= \log \binom{n}{r} + r\log(\hat{p}) + (n-r)\log(1-\hat{p}) \\
\ell'(p) &= \frac{r}{p} - \frac{n-r}{1-p} \\
\ell'(\hat{p}) &= 0 \\
\ell''(p) &= -\frac{r}{p^2} - \frac{n-r}{(1-p)^2} \\
\ell''(\hat{p}) &= -\frac{n}{\hat{p}(1-\hat{p})} \\
\ell^*(\hat{p}) &= b \\
{\ell^*}'(p) &= 2a(p-\hat{p}) \\
{\ell^*}'(\hat{p}) &= 0 \\
{\ell^*}''(p) &= 2a
\end{aligned}$$

So equating terms for $\ell(p)$ and $\ell^*(p)$,

$$\begin{aligned}
b &= \log \binom{n}{r} + r\log(\hat{p}) + (n-r)\log(1-\hat{p}) \\
a &= -\frac{n}{2\hat{p}(1-\hat{p})}
\end{aligned}$$

and the approximation is, for the log-likelihood,

$$\ell(p) \doteq \ell^*(p) = \log \binom{n}{r} + r\log(\hat{p}) + (n-r)\log(1-\hat{p}) - \frac{n(p-\hat{p})^2}{2\hat{p}(1-\hat{p})},$$

and the corresponding approximation for the likelihood is

$$L(p) \doteq L^*(p) = \binom{n}{r}\hat{p}^r(1-\hat{p})^{n-r} \cdot \exp\left[-\frac{n(p-\hat{p})^2}{2\hat{p}(1-\hat{p})}\right].$$

The *relative* likelihood is then approximated by

$$\begin{aligned} RL(p) \doteq RL^*(p) &= L^*(p)/L^*(\hat{p}) \\ &= \exp\left[-\frac{n(p-\hat{p})^2}{2\hat{p}(1-\hat{p})}\right]. \end{aligned}$$

The interval for p for 95% confidence is defined by $RL(p) = 0.15$. If we use the approximation $RL^*(p) = 0.15$, this is equivalent to

$$\log_e RL^*(p) = \log_e 0.15$$

which is

$$-\frac{n(p-\hat{p})^2}{2\hat{p}(1-\hat{p})} = -1.9,$$

which is equivalent to

$$(p-\hat{p})^2 = 3.8\frac{\hat{p}(1-\hat{p})}{n},$$

or

$$p = \hat{p} \pm 1.95\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

and we round off the value 1.95 to 2.

The more accurate approximate interval comes from replacing a in the approximating log-likelihood by $a' = -n/[2p(1-p)]$, where p is *not* replaced by $\hat{p}$.

4.6 Sample design

We often want to be able to estimate a population proportion to a given degree of precision. For example, we may want a 95% confidence interval for a proportion to be not more than a specified length. This usually requires a quite large sample, for which the simple approximation for the confidence interval is quite accurate.

Suppse we want the 95% confidence interval for p to be not more than 0.04 in length. The simple approximate interval is $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, and if this is to be of length not more than 0.04, we must have

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 0.01,$$

which means that $n > 10^4 \cdot \hat{p}(1-\hat{p})$. Since the maximum value of $\hat{p}(1-\hat{p})$ is $1/4$, the interval length requirement will be satisfied, whatever the sample outcome, if $n > 2500$. In general, if the 95% confidence interval for p is to be of length not more than δ, then this is guaranteed if the sample size exceeds $4/\delta^2$.

We thus have a solution to the problem of inference about a single population proportion, including the sample size required for a specified precision. We now want to extend this more generally.

Chapter 5

Two-population problems

5.1 Relationships

In the first assignment the last question asked whether fathers' and mothers' smoking was related. How do we investigate *relationship* questions?

For two binary variables, this is relatively simple. We can *stratify*, or divide, the population into two sub-populations: families with a smoking father, and those with a non-smoking father. In my sample of 40 families, there were 22 with a smoking father, and 18 with a non-smoking father. In the 22 smoking-father families, there were 10 in which the mother also smoked. In the 18 non-smoking father families, there were 3 in which the mother smoked.

Let p_{FS} be the (sub-)population proportion of smoking mothers in the sub-population of smoking fathers. We call p_{FS} a *conditional probability* – it is the probability of a smoking mother, *given that* the father smokes.

We write similarly p_{FNS} for the proportion of smoking mothers in the sub-population of *non*-smoking fathers – p_{FNS} is the conditional probability of a smoking mother *given that* the father is a non-smoker.

If the father smokes, my *sample* proportion of smoking mothers is $10/22 = 0.45$ which is the sample estimate p_{FS}. If the father does not smoke, my sample proportion of smoking mothers is $3/18 = 0.17$. This is much smaller than 0.45 – it appears that if the father does not smoke, a smaller proportion of mothers smoke than if the father does smoke. We say that mothers and fathers smoking appears to be *positively associated*, or *positively correlated* – if one smokes, the other is more likely to smoke.

However, as with all samples from populations, these *estimates* of the sub-population proportions vary across different random samples, and so we need a method to decide whether the difference in conditional proportions we see in the sample reflects a real difference in the populations.

5.2 Confidence intervals for a difference in proportions

We could construct confidence intervals for the sub-populations proportions from our samples. Using the accurate approximation method, the 95% confidence interval for p_{FS} is $(0.266, 0.657)$, and the simple approximation gives $(0.243, 0.667)$, while the approximations for p_{FNS} are $(0.057, 0.397)$ and $(-0.009, 0.343)$. The last approximation gives an impossible negative endpoint, though we know that any proportion p must lie in the range (0,1). The simple approximation breaks down from the small sample size and the low proportion p_{FNS}.

Since these intervals overlap in the interval (0.266 – 0.397), it appears that the two population proportions could be equal. However we really want a confidence interval for the *difference* $p_{FS} - p_{FNS}$. Obtaining this interval requires more theory than we will develop, but the approximate result is very simple, and is an extension of the simple approximate method for a single p.

An approximate 95% confidence interval for $p_{FS} - p_{FNS}$ is

$$\hat{p}_{FS} - \hat{p}_{FNS} \pm 2\sqrt{\frac{\hat{p}_{FS}(1-\hat{p}_{FS})}{n_{FS}} + \frac{\hat{p}_{FNS}(1-\hat{p}_{FNS})}{n_{FNS}}}.$$

The first term is the *difference* between the observed proportions, and the second is the square root of the *sum* of the terms in the square roots of the separate confidence intervals.

For my sample this is

$$0.455 - 0.167 \pm 2\sqrt{\frac{0.455 * 0.545}{22} + \frac{0.167 * 0.833}{18}} = 0.288 \pm 0.276 = (0.012, 0.564).$$

The interval extends from just above zero to 0.564 – it is possible, though at the extreme of acceptability, that there is no difference between p_{FS} and p_{FNS}, but it is also possible that they could differ by a large amount, as much as 0.56 at the other extreme of acceptability.

Our small samples in each sub-population do not give much precision but my sample suggests pretty strongly that there *is* a positive association in the population. Yours will vary from this result ... The population reality will be revealed!

5.3 Population breakdowns

The true population structure is as follows:

The difference in proportion of mothers smoking, between smoking fathers and non-smoking fathers, is $0.483 - 0.232 = 0.251$. This is a substantial difference – the father's smoking behaviour affects the mother's (or is it the other way around?). We can break this table down by the sex of the child:

Table 5.1: Population smoking

		Father	smoking	
		Yes	No	Total
Mother	Yes	345	135	480
smoking	No	370	446	816
Total		715	581	1296
	p	0.483	0.232	

Table 5.2: Population smoking by sex of child

		Girl	child		Boy	child	
		Father	smoking		Father	smoking	
		Yes	No	Total	Yes	No	
Mother	Yes	190	67	257	155	68	223
smoking	No	178	213	391	192	233	425
Total		368	280	648	347	301	648
	p	0.516	0.239		0.447	0.226	

The proportions of smoking mothers in the “father smoking” category are different – 0.52 in the girl baby and 0.45 in the boy baby sub-populations. Is this important? This is a genetic and environmental issue – is it possible that the relation could be the other way around – that the smoking behaviour of father and mother could affect the sex of the child? That is, that the smoking behaviour of the parents could change the probability of the child being a boy? Such questions are common in *epidemiology* – the study of diseases in human populations – where the same kind of data collected. We find *cases* with the disease or condition, and look for *causal factors* which might *cause* or *explain* the disease or condition, by comparing the frequency of a causal factor amongst the *cases* with the frequency of the factor amongst *controls* – people who are normal, or healthy, without the disease or condition.

Observational studies of this kind are called *case-control* studies. Since there is no randomisation of people to “case” or “contro” condition, differences between these groups may be due to many other possible factors.

Studies of new drugs and their effectiveness in treating patients are carried out in a very controlled way, in designed studies called *Randomised Clinical Trials*.

Chapter 6

The Randomised Clinical Trial(RCT)

6.1 Definition

An important application of statistical inference using the binomial distribution is to the comparison of new medical or surgical treatments for disease or illness in a *randomised clinical trial*. Such trials have certain characteristic features:

- A new treatment which has found to be effective in small studies on selected patients is to be evaluated in a large study, compared with the *current best treatment*.

- Patients taking part in the study are assigned to receive either the new treatment or the current best treatment. Assignment to one treatment or the other by *randomisation*: in the simplest form (not used in practice) by tossing a coin for each patient – heads means the new treatment, tails means the current best treatment.

- The treatment received by the patient *appears* to be the same, so the patient is not aware of which treatment her or she is receiving. This requires *informed consent* by the patient in advance: he or she must be told what the treatments are, and that he or she will be randomised to one or the other treatment.

- With new drug treatments, the pills or capsules are made to look the same for each treatment, though for the current best treatment group the pill may have no active component – it may be an inert *placebo*.

6.2 Example – RCT of depepsen for the treatment of duodenal ulcers (1969)

This study was carried out at the Royal North Shore hospital in Sydney by Professor D.W. Piper and co-workers. The drug depepsen (a trade name for sodium amylosulphate) had been found effective in the treatment of gastric (stomach) ulcers, and it was believed that because of its known physiological action in the treatment of this condition, and the similarity of the two conditions, it should also be effective for duodenal ulcers. The criterion for "success" of the treatment was taken as the complete healing of the ulcer within a period of 8 weeks after the beginning of treatment. The existence of an ulcer, and its healing, were positively identified by fibre-optic duodenoscopy, in which a flexible tube is swallowed by the patient, and the lining of the duodenum examined visually through the optical tube.

The standard (current best treatment) for duodenal ulcers was then sedation and bed rest. About 50% of patients recovered, that is healed completely in 8 weeks, with this treatment alone, without any additional drug treatment. However this proportion could not be assumed to apply to the hospital subjects in the study – it had to be assessed from patients randomly assigned to the *control group* receiving the current best treatment.

The study encountered a difficulty – the drug depepsen was very expensive to produce, and only 20 doses were available for the trial, so only 20 patients could be treated with depepsen. The dose given in the *depepsen treatment* group was 5ml containing 500mg of depepsen, to be taken six times daily, one and three hours after each main meal. Patients randomised to the *placebo control* group received a "dose" of 5ml of flavoured liquid at the same frequency.

At the end of the eight week period, duodenoscopy was performed again to determine whether the ulcer had completely healed. 20 patients were randomly assigned to the treatment group and 18 to the control group, but three patients had to be excluded from the study because they did not comply with the *protocol* – the instructions for the treatment – they all took *all* the medication in the first week. Two of these were in the treatment group and one in the control group. Of the 35 remaining patients, 13 of the 18 receiving depepsen healed, while 10 of the 17 receiving placebo healed. Does this indicate a real superiority in healing of depepsen over placebo?

Classifying the patients by treatment and recovery, we have the 2 x 2 table:

Table 6.1: Clinical trial of depepsen

	Depepsen	Placebo	Total
Healed	13	10	23
Not healed	5	7	12
Total	18	17	35

Write p_D for the probability of recovery with depepsen, and p_P for the probability of recovery with placebo. Then $\hat{p}_D = 13/18 = 0.722$, and $\hat{p}_P =$

$10/17 = 0.588$. A higher *sample* proportion of patients recover with depepsen, but is this true in the population, or could it be that $p_D = p_P$, or $p_D < p_P$? We construct a confidence interval to answer this question.

The approximate 95% confidence interval for $p_D - p_P$ is

$$\begin{aligned} p_D - p_P &\in \hat{p}_D - \hat{p}_P \pm 2\sqrt{\frac{\hat{p}_D(1-\hat{p}_D)}{n_D} + \frac{\hat{p}_P(1-\hat{p}_P)}{n_P}} \\ &= 0.134 \pm 2\sqrt{\frac{0.722 * 0.278}{18} + \frac{0.588 * 0.412}{17}} \\ &= 0.134 \pm 0.319 \\ &= (-0.185, +0.453). \end{aligned}$$

So the difference in recovery proportions in the two populations could be as much as 0.453 in favour of depepsen, or as much as 0.185 in favour of placebo. Our trial is so small that this small difference in sample proportions is a poor indicator of the difference in the population proportions.

So this trial is *inconclusive*, as are many small trials – the sample sizes are too small to give any precision in the difference in the response proportions. How large would the trial *need* to be to find that this difference *did* indicate the superiority of depepsen?

We can work this out using the same method as for a single proportion. We have as before $\hat{p}(1-\hat{p}) \leq 1/4$ for all p, and assuming $n_D = n_P = n$, we have

$$\sqrt{\frac{\hat{p}_D(1-\hat{p}_D)}{n_D} + \frac{\hat{p}_P(1-\hat{p}_P)}{n_P}} \leq \sqrt{\frac{1}{4n} + \frac{1}{4n}} = \sqrt{\frac{1}{2n}}.$$

So if $\hat{p}_D - \hat{p}_P > 2\sqrt{\frac{1}{2n}}$, the 95% confidence interval will not include zero. This means

$$0.134 > \sqrt{\frac{2}{n}},$$

so that

$$\sqrt{\frac{n}{2}} > \frac{1}{0.134}, \sqrt{n} > \frac{\sqrt{2}}{0.134}, n > \frac{2}{0.134^2} = 111.4.$$

So in a large trial with 112 patients in each of the treatment and control groups, an observed difference of $0.722 - 0.588$ would indicate the superiority of depepsen over placebo, because the 95% confidence interval would not include zero. The *actual* sample sizes are less than 1/6 of the required size – the trial was far too small to establish this difference as real.

Soon after this trial, a different drug treatment for duodenal ulcers – cimetidine (trade name Tagamet) – was found to be effective, and trials of depepsen for the treatment of duodenal ulcers were abandoned.

In the last few years, these drug treatments which are based on reducing acidity in the stomach have been replaced by an entirely different treatment with antibiotics – it was discovered that most ulcers develop from a bacterial infection which responds almost immediately to antibiotic drug treatment – many ulcers heal completely in one week!

Chapter 7

Measuring the strength of association

7.1 Measures of association

We frequently want to express the strength of association between two binary variables by a single number – a measure of association on a common scale. Consider the association between mothers' and fathers' smoking. For the population we have:

Table 7.1: Parents' smoking

		Father	smoking	
		Yes	No	Total
Mother	Yes	345	135	480
smoking	No	370	446	816
	Total	715	581	1296

We would like the *scale* for our measure of association to range from -1 to $+1$, these extremes representing complete *negative* association and complete *positive* association, with the midpoint 0 representing *no* association.

What does complete association mean? Suppose the table looked like Table 7.2. Then every couple either smokes, or does not smoke – the mother and

Table 7.2: Parents' smoking – complete positive association

		Father	smoking	
		Yes	No	Total
Mother	Yes	715	0	715
smoking	No	0	581	581
	Total	715	581	1296

father have *identical* smoking behaviour. We want this pattern to correspond to an association of +1.

Now suppose the table looked like Table 7.3. Now if one parent smokes, the other does not – we have identically *reversed* smoking behaviour. We want this pattern to correspond to an association of -1.

Table 7.3: Parents' smoking – complete negative association

		Father	smoking	
		Yes	No	Total
Mother	Yes	0	581	581
smoking	No	715	0	715
	Total	715	581	1296

Now suppose the table looked like this:

Table 7.4: Parents' smoking independent

		Father	smoking	
		Yes	No	Total
Mother	Yes	265	215	480
smoking	No	450	366	816
	Total	715	581	1296

From our conditional probability approach we have

$$p_{FS} = 265/715 = 0.370, \, p_{FNS} = 215/581 = 0.370,$$

so *it makes no difference*, to the probability of the mother smoking, whether the father is smoking or not. We want this pattern, of *independence* of fathers' and mothers' smoking, to correspond to an *association of zero.*

7.2 The correlation coefficient

We now adopt a standard notation for the 2 x 2 contingency table:

Table 7.5: Parents' smoking – in general

		Father	smoking	
		Yes	No	Total
Mother	Yes	a	b	$a+b$
smoking	No	c	d	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d$

We denote the "cell" entries in the table by , b, c and d. If we condition on father's smoking, we have for the probability of the mother smoking,

$$\begin{aligned}
p_{FS} &= \frac{a}{a+c} \\
p_{FNS} &= \frac{b}{b+d} \\
p_{FS} - p_{FNS} &= \frac{a}{a+c} - \frac{b}{b+d} \\
&= \frac{a(b+d) - b(a+c)}{(a+c)(b+d)} \\
&= \frac{ad-bc}{(a+c)(b+d)}.
\end{aligned}$$

If we consider the table the other way around, and ask what is the effect of mother's smoking on father's smoking, we have

$$\begin{aligned}
p'_{MS} &= \frac{a}{a+b} \\
p'_{MNS} &= \frac{c}{c+d} \\
p'_{MS} - p'_{MNS} &= \frac{a}{a+b} - \frac{c}{c+d} \\
&= \frac{a(c+d) - c(a+b)}{(a+b)(c+d)} \\
&= \frac{ad-bc}{(a+b)(c+d)}.
\end{aligned}$$

The numerator is the same, but the denominator is different. We now define the *correlation coefficient* of the two binary variables, denoted by ρ, as

$$\rho = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \sqrt{(p_{FS} - p_{FNS})(p'_{MS} - p'_{MNS})}.$$

The correlation is *symmetric* in the two variables. Its value for the smoking status is

$$\rho = \frac{345 \cdot 446 - 135 \cdot 370}{\sqrt{480 \cdot 816 \cdot 715 \cdot 581}} = 0.258 = \sqrt{0.251 \cdot 0.266},$$

where 0.251 and 0.266 are the differences in conditional probabilities in each direction.

It is convenient to have a verbal scale to interpret the correlation coefficient. We say that $0 < \rho < 0.3$ is a *low* (positive) correlation, $0.3 \leq \rho < 0.6$ is a *moderate* (positive) correlation, and $0.6 \leq \rho < 1$ is a *high* correlation, with similar definitions for negative correlations.

This definition of the correlation coefficient applies to the *population.* We can define correspondingly the *sample* correlation coefficient, denoted by r. For my sample of 40, we have the corresponding sample correlation coefficient:

Table 7.6: Parents' smoking – my sample

		Father	smoking	
		Yes	No	Total
Mother	Yes	15	2	17
smoking	No	8	15	23
	Total	23	17	40

$$r = \frac{15 \cdot 15 - 2 \cdot 8}{\sqrt{17 \cdot 23 \cdot 23 \cdot 17}} = 0.535 = \sqrt{(15/17 - 8/23) \cdot (15/23 - 2/17)} = \sqrt{0.535 \cdot 0.535}.$$

The sample correlation in this sample is equal to the conditional probability differences which are the same. The sample correlation is far away from the true population correlation: sample correlations vary substantially in different small random samples.

Chapter 8

Continuous Variables

8.1 Definitions

So far we have considered only binary variables, taking one of two values. However most of the variables in the StatLab database, and in practical surveys and experiments, are not binary, but are either *categorical* with more than two categories, or "continuous", measured on a continuous scale, like birth weight, age, or income. In theory, any value of these "continuous" variables is *possible*, but the values we can *record* are limited by the precision of the measuring instrument which records them. So birth weight is recorded to the nearest 0.1 pound (though it was actually measured in pounds and ounces), mother's weight to the nearest pound, and family income in $100 units.

If we tabulate the frequency of each possible value as a "bar chart", we obtain a *histogram* or *frequency diagram* of the population values for the variable. Figure 1 shows a histogram of mothers weight for girls in the STATLA Bpopulation. The figure is very "spiky", as many possible values have no mothers and others (particularly at "round" values like 150 and 165) have many mothers.

8.2 Population mean and median

We will not be concerned with this amount of *detail* in the histogram, and will be content with some overall *summaries* of the population histogram. We first define some notation. Let the possible values of the variable Y in the population be represented by $Y_{(1)}, Y_{(2)}, ..., Y_{(D)}$ where these are in *increasing order*, from the smallest $Y_{(1)}$ to the largest $Y_{(D)}$. The *number* of values of $Y_{(I)}$ in the population will be written N_I, so the *sum* of the N_I over the range $I = 1, 2, ..., D$ is the *population size* N: we write $N = \sum_{I=1}^{D} N_I$. If we divide each N_I by the population size N we obtain the *proportion* of the values of $Y_{(I)}$ in the population: we define

$$p_I = N_I / N$$

to be the *population proportion* of the value $Y_{(I)}$. If we add up, or *cumulate*, the population proportions p_I, we obtain the *cumulative proportions* C_I:

$$C_I = \sum_{J \leq I} p_J = \sum_{J \leq I} N_J/N.$$

Figure 2 shows the cumulative proportions for mothers weight for girl babies. It is quite smooth, although Figure 1 is very rough.

We now define two commonly-used summary quantities. The *population* (arithmetic) *mean* of the variable Y is denoted by the symbol μ (the Greek letter "mew") with a subscript for the variable:

$$\begin{aligned} \mu_Y &= \sum_{I=1}^{D} N_I y_{(I)}/N \\ &= \sum_{I=1}^{D} p_I y_{(I)}; \end{aligned}$$

it is the simple *average* of all the values in the population, which is equal to the *weighted* average of all the *possible* values, weighted by their *proportions* in the population.

The population *median* of the variable Y, denoted by ν_Y (the Greek letter "new"), is (roughly) the value below and above which 50% of the population lie. More precisely, ν_Y is the value of Y, say $Y_{(K)}$, such that $C_{K-1} < 0.50$ but $C_K \geq 0.50$. So as Y increases, the cumulative proportion increases also, and the median is the value of Y at which the cumulative proportion *changes* from being less than 0.5 to being greater than or equal to 0.5.

For the mother's weight population, the mean mother weight is 130.9 pounds and the median is 128 pounds. These are quite close, and they are often close for *symmetrical* populations of variable values.

Figure 3 gives the population histogram for boy birth weights. It is quite symmetrical. The mean boy birth weight is 7.64 pounds and the median is 7.60 pounds.

8.3 Sample Median

Given a sample from the population with measured values of a continuous variable, we define the *sample* median analogously. We sort the observations into increasing order, and then find the "middle" observation. The definition of the sample median is slightly different depending on whether the sample size n is even or odd. If it is *odd*, then we define $r = (n+1)/2$, and the sample median is the r-th observation in increasing order. So if $n = 39$ for example, then $r = 20$, and the sample median is the 20-th observation in increasing order.

If n is *even*, define $r = n/2$, and the sample median is the *average* of the r-th and $(r+1)$-th observations in increasing order. So if $n = 40$, the median is the average of the 20-th and 21-st observations.

So the sample median defines a "cut" of the sample into "hih" and "low" values – we can convert the continuous variable into a binary variable: "high/low" on the continuous scale. This allows us to use the previous theory of analysis for binary variables applied to the high/low categorization of continuous variables.

Sorting the sample observations is tedious in large samples, for which we use computer sorting routines. For small samples this sorting process is done conveniently by constructing a *stem-and-leaf* plot.

8.4 Example 1 – boy and girl birth weights

It is common knowledge that boys are larger and heavier at birth than girls (in some overall population sense). Do we have evidence for this in our sample of 40? My sample birth weights came from 18 boys and 22 girls:

```
Boys (18)
---------
7.6  7.2  8.3  7.1  6.9  8.8  7.6  9.8  6.9  8.4
9.9  6.3  7.3  7.4  7.0  6.4  7.3  6.9

Girls (22)
----------
7.2  6.8  5.9  6.2  7.8  6.9  7.3  6.7  7.6  3.1
8.9  6.4  7.0  7.1  4.9  6.4  6.9  7.5  5.9  7.9
5.7  7.1
```

We record each value (separately for boys and girls) as the "leaves" (using the decimal place value) on the "stem" (the weight in pounds):

```
Weight          Boys            Girls
------          ----            -----
3                               1
4                               9
5                               997
6               99349           8297449
7               62163403        283601591
8               384             9
9               89
```

The visual picture is clear – the "histogrm" for girls is shifted upwards – towards smaller values – than that for boys. But could that be just a sample fluctuation (do *your* samples show this too?) consistent with the same median weight in the population for boys and girls?

To check this, we determine the *combined sample median.* We first re-order the weights in increasing order in each row, and pool them to form a combined sample:

```
Weight(pounds)  Boys      Girls      Combined
------          ----      -----      --------
3                         1          1
4                         9          9
5                         799        799
6               34999     2447899    234447899999
7               01233466  011235689  00111223334566689
8               348       9          3489
9               89                   89
```

The 20th and 21st observations in increasing order are both 7.1, so the sample median is 7.1 pounds for the combined sample.

If the population medians for the boy and girl birth weight are the same, then the same proportion of boys and girls (0.5) lie above and below their common *population* median. So if we classify the *sample* values as above and below the *combined sample* median, we should see approximately the same sample proportions above and below for boys and girls.

We define "low" birth weight to be *less than or equal to 7.1 pounds*, and "high" birth weight to be greater than 7.1 pounds. This does not mean that "low" birthweight is abnormal, or even unusual – it simply means that half the sample were in this below-sample-median category. Classifying the birthweights, we have the 2 × 2 table:

Table 8.1: Birthweights – my sample

	Boys	Girls	Total
Low	7	15	22
High	11	7	18
Total	18	22	40

Note that the numbers of "low" and "high" birth weights are not 20, because the median is an observed value and so median values are classified into the "low" category.

Does this table provide evidence of different population proportions of boy and girl birth weights below and above 7.1 pounds? The sample proportion of "high" birth weights for boys is $11/18 = 0.611$, and that for girls is $7/22 = 0.318$ which is substantially less. A 95% confidence interval for the true difference in population proportions is

$$0.611 - 0.318 \pm 2\sqrt{\frac{0.611 \cdot 0.389}{18} + \frac{0.318 \cdot 0.682}{22}}$$

which is 0.293 ± 0.304, or $(-0.011, 0.597)$.

The confidence interval *just* includes zero, so we do *not quite* have convincing evidence that the population proportions differ.

8.5 Example 2 – mother's smoking and child's birthweight

Do mothers who smoke at the diagnosis of pregnancy have babies with "low" birthweight? This question is badly worded – we mean that the birthweight *average* – mean or median – is lower for mothers who smoke than for mothers who don't smoke. Does our sample provide any evidence of this?

This question is complicated by the difference in average birthweights for boys and girls. We did not give these earlier, but from the stem-and-leaf plot we se that the sample median birthweight for boys is 7.3 pounds, while that for girls is 6.9 pounds. We will ignore this difference, and classify the sample by high/low birthweight as defined above, and by mother smoking or not at diagnosis. The resulting 2 × 2 table is:

Table 8.2: Birthweights and mothers smoking – my sample

		Mother's	smoking	
		Smoker	Non-smoker	Total
Birth	Low	10	12	22
weight	High	5	13	18
	Total	15	25	40

The proportion of low birthweight babies in the sample is 10/15 = 0.667 for smoking mothers, and 12/25 = 0.480 for non-smoking mothers. Smoking appears to be positively associated with low birth weight, but the correlation of 0.18 is quite low, and the 95% confidence interval for the difference in population proportions is

$$0.187 \pm 2\sqrt{\frac{0.667 \cdot 0.333}{15} + \frac{0.48 \cdot 0.52}{25}} = 0.187 \pm 0.315 = (-0.128, 0.502)$$

which includes zero.

Thus this sample does *not* provide convincing evidence of association between smoking at diagnosis and low birthweight.

8.6 Changes in an individual on repeated measurement

When we measure an individual on two occasions, we are often interested in the *change* in the individual's measurement between the two occasions, and in particular whether there has been *any* change in the population median. The weight of mothers in the StatLab population is an example.

The weights in pounds of the 40 mothers in my sample are given below, at diagnosis (D) and follow-up (F), with the change C = F - D:

```
D: 110 147 212 115 247 170 140 140 106 140
F: 113 147 206 120 185 168 135 144 112 156
C:  +3   0  -6  +5 -62  -2  -5  +4  +6 +16

D: 133 145 140 130 135 105 112 104 172 118
F: 145 140 164 139 123 126 112 118 193 114
C: +12  -5 +24  +9 -12 +21   0 +14 +21  -4

D: 160 117 145 162 121 145 148  89  97 122
F: 161 128 169 160 118 200 165  97 109 151
C:  +1 +11 +24  -2  -3 +55 +17  +8 +12 +29

D: 104 132 128 115  94 124 108 140  95 158
F: 115 141 123 120 100 134 106 157 109 207
C: +11  +9  -5  +5  +6 +10  -2 +17 +13 +49
```

The stem-and-leaf plot of the weight changes is shown below, for the observations in the above order, and then re-ordered. There are some large individual changes, both positive and negative, but do these represent any *general* population change in median weight?

```
 C              units           units
---             -----           -----
-60             2               2
-50
-40
-30
-20
-10             2               2
- 0             625542352       222345556
+ 0             305469018956    001345566899
+10             6241721073      0112234677
+20             41149           11449
+30
+40             9               9
+50             5               5
```

The plot is fairly symmetrical about zero. The median is the average of 6 and 8, so is 7. So in the sample, half the mothers gained 8 pounds or more, the other half gained 6 pounds or less. Could the median in the population of the weight changes be zero? This would mean that, in the population, as many mothers gained weight as lost weight over the 10-year follow-up period.

The number of negative changes in the sample is 11, and the number of positive changes is 27. Two mothers did not change weight – we exclude these from consideration. So of the 38 mothers who changed weight, a proportion

$27/38 = 0.711$ *increased*. Could this be consistent with a population proportion of 0.5?

Based on $\hat{p} = 0.711$ in $n = 38$, the 95% confidence interval for the population proportion is $(0.549, 0.832)$. This does *not* include 0.5, so the sample proportion who increased in weight is *not* consistent with 0.5 – a median weight *gain* definitely occurred over the 10-year follow-up period. The sample median increase of 7 pounds is definitely not consistent with zero.

8.7 Confidence interval for the median

If the population median is not zero, what can we say about it? We could check other hypothetical values of the median to see whether they are consistent with the data. For example, could the median weight change be 3 pounds? There are 13 sample values less than 3, one equal to 3, and 26 greater than 3. Of the 39 mothers with weight gains not equal to 3 pounds, the sample proportion $\hat{p} = 26/39 = 0.667$ had weight gains greater than 3 pounds. The 95% confidence interval for the population proportion p with more than 3 pounds gain is $(0.507, 0.796)$ which still excludes zero. So the value of 3 pounds is also not a possible value for the population median.

We don't want to have to repeat this calculation for every possible value of the median! Fortunately, we can find those values of the median in the confidence interval directly from the formula. Recall that $\hat{p} = r/n$, where r is the number of observations in the sample greater than the hypothetical median. If the confidence interval for p includes 0.5, for a given value of the median ν, then this value of ν is included in the confidence interval for the median. So the *extreme endpoints* of the 95% confidence interval for ν are given by the values of $\hat{p}$, or r, corresponding to the extreme values of $\hat{p}$ for which the 95% confidence interval for p just includes 0.5.

So the endpoints of the 95% confidence interval for ν are given by the solutions, in $\hat{p}$, of

$$\frac{\hat{p} + \lambda^2/(2n)}{1 + \lambda^2/n} \pm \lambda \frac{\sqrt{\hat{p}(1-\hat{p})/n + \lambda^2/(4n^2)}}{1 + \lambda^2/n} = 0.5.$$

After some algebra we find that the two solutions to this equation are

$$\hat{p} = \frac{1}{2} \pm \frac{1}{\sqrt{n}}$$

which is equivalent to

$$r = n\hat{p} = \frac{n}{2} \pm \sqrt{n}.$$

These values of r define the observations which are the endpoints of the 95% confidence interval for ν. For $n = 40$, these values are

$$r = 20 \pm \sqrt{20} = 20 \pm 4.47 = (15.53, 24.47).$$

Of course r has to be an integer; we take the integer values of r just *inside* this interval, 16 and 24. The 16-th observation in order is 4, the 24-th is 10. So the 95% confidence interval for the population median change ν in weight is (4, 10) pounds.

This method can be used for any continuous variable. For mother's weights at diagnosis and at test, we have the stem-and-leaf plots below:

```
tens   D(units)    F(units)
----   --------    --------
 80    9
 90    457         7
100    44568       0699
110    025578      2234588
120    1248        003368
130    0235        459
140    0000055578  01457
150    8           167
160    02          014589
170    02
180                5
190                3
200                067
210    2
...
240    7
   (med. 131)  (med. 137)
```

So the 95% confidence interval for the median weight at diagnosis is (121, 140) and for the median weight at follow-up is (123, 144).

8.8 Association between two continuous variables

We examine the association between *two* continuous variables by cutting *both* variables at their medians and classifying them as high/low in a 2×2 table.

Is there an association between low birthweight and intelligence, as measured by the Peabody test score? For my sample of 40, the Peabody test scores give the stem-and-leaf plot below:

```
tens  units
----  -----
50    7
60    22345788
70    1355556677899
80    0011122233445779
90    01
```

The sample median is 78.5. Classifying by “low” (≤ 78) or “high” (≥ 79) and by low or high birthweight gives for my sample:

Table 8.3: Birthweights and Peabody score – my sample

		Peabody	score	
		Low	High	Total
Birth	Low	10	10	20
weight	High	10	10	20
	Total	20	20	40

This is an unusual result! The correlation between birthweight and Peabody score is zero! There is certainly no evidence from this sample of any association in the population.