# Categorical Data Analysis

(These notes are available as a pdf file on the following website: http://www.biostat.umn.edu/%7Ejohn-c/ph7460.f2006.html Go to the link called: 'kisumu.2011'. )

## Introduction

There are many kinds of data. There are *quantitative*, or measured variables like blood pressure or temperature. There are *counts*, like the number of students in a classroom who have had chickenpox. There are *times to events*, like the number of days after birth that a baby has an infectious disease. Data need not be numeric; data can be the names of drugs that a patient is taking.

*Categorical* data in general are counts. Counts of items or events are expressed as nonnegative integers. That is, counts are numbers like 0, 1, 2, 3, 4, 5, …, 100, 101, … Note that zero is a possible count: for example, if you do not own a cat, the count of the number of cats that you own is zero.

## A. One Sample

### A.1. The Simplest Possible Data

The simplest possible data describe the state of one entity. The entity might be a person. The state could be expressed as 0 or 1. Suppose the state in question is marital status. Let X represent the state. So if X = 0, the person is not married. If X = 1, the person is married.

A variable like X, which can take on only two values – 0 or 1 – is called a *Bernoulli random variable.* The term 'random' means that if you have no information about the person, you cannot predict with certainty whether X = 0 or X = 1 for that person. Again, if X represents marital status, and you meet a person for the first time, you will not know whether he or she is married. This means that X is a *random variable* representing the outcome of *one observation*.

A good example of a Bernoulli random variable is the outcome of flipping a coin. Let X = 0 if the outcome is tails, X = 1 if the outcome is heads. You cannot say with certainty in advance whether X will equal 0 or 1. If the coin is "fair", that is, not some kind of trick coin that tends to give more heads than tails, you would expect that X = 0 about half the time and X = 1 about half the time. This means that the ***expectation*** of X is

$$E(X) = 0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$$

That is, the expectation is the *average* outcome.

However, it is not usually true that the expectation of a Bernoulli random variable is 0.5. Suppose X represents the event that a person has blood type O+. That is, X = 1 if the person has blood type O+ and X = 0 otherwise. The prevalence of blood type O+ in South Africa is 39%, that is, the probability of having blood type O+ among people in South Africa is 0.39. The expectation of X in this case is

$$E(X) = 0.61 \cdot 0 + 0.39 \cdot 1 = 0.39$$

More generally, if $X \sim Ber(p),$ then $E(X) = p.$

Here is the more general rule for computing expectation. Suppose a random variable X can take on values $x_1, x_2, x_3, ..., x_n.$ Suppose that the *i-th* value $x_i$ is taken on with probability $p_i$. Then the expected value is the *weighted average,*

$$E(X) = p_1 \cdot x_1 + p_2 \cdot x_2 + ... + p_n \cdot x_n.$$

**Exercise 1:** Suppose X is a random variable which represents how many tires on a car are older than 5 years old. Suppose that:

$$prob(X = 0) = .2,$$
$$prob(X = 1) = .15,$$
$$prob(X = 2) = .25,$$
$$prob(X = 3) = .15,$$
$$prob(X = 4) = .25.$$

(Note that these probabilities must add up to 1.0).

Compute the expectation of X.

You might think Bernoulli random variables cannot tell you much: they just involve one observation on one person. The person either has the characteristic that *X* represents, or doesn't. Usually statistics involves many observations. What can you tell from one observation?

Suppose someone says to you, "There are **no people** whose eyes have different colors." You can view this as a *hypothesis*. The hypothesis could also be stated as

$$prob(X = 1) = 0,$$

where *X* is the Bernoulli random variable which is 1 if a person's eyes are of different colors and 0 if a person's eyes are of the same color. To say that a probability is zero is basically the same thing as saying: this event is **impossible.**

Then one day you happen to meet someone whose eyes are of different colors. That is, you observe $X = 1.$ You have observed a rare event. More than that: this one observation is enough to **prove** that the hypothesis above is false!

This also illustrates something about statistics. People state hypotheses. Sometimes they are stated as positive assertions. For example, "All people have eyes that are the same color." How many observations would you need to make to prove this hypothesis? Answer: You might have to examine everyone on Earth. Even this would not be enough, because by the time you got done examining everyone, an enormous number of new people would have been born that were not there before. So you would have to start over and examine all of them also. There are over 6 billion people on Earth. If you examine one person every second, it would take you over 190 years to examine everyone who is alive right now. You probably will not live that long. A great many of those 6 billion + people will have died before you get to examine them. Even if you did, billions of new people would be born before you finished. (Plus, you would need to take time to have lunch and a nap now and then.) Conclusion: there is no practical limit on how many people you would need to observe to prove the hypothesis. But only one observation would be sufficient to ***disprove*** it.

### A.2. Statistics and p-values

In general statistics do not prove or disprove hypotheses. They provide evidence. The strength of the evidence is stated in terms of probabilities. You collect data and you compile statistics and from this you compute probabilities. The probabilities are related to some underlying hypothesis. A very low probability (like, for example, 0.001) indicates that you have observed a rare event. This could be due to just luck, or it could be due to the fact that your underlying hypothesis is not true. This is how statistics get used in clinical and medical research. The probabilities that are computed are often called *p-values.*

### A.3. Bernoulli and Binomial Random Variables

Bernoulli random variables have only two possible values, 0 and 1. One number is sufficient to describe a Bernoulli random variable X. That is the probability that $X = 1$. If this probability is *p* then the probability that $X = 0$ is $1 - p$. We say that X has a Bernoulli distribution with parameter *p.* This is also written as: $X \sim Ber(p).$

Different random variables can be added together to produce new random variables. Suppose $X_1, X_2, X_{3,}..., X_n$ are all Bernoulli random variables with parameter *p.* Let $X_{sum} = X_1 + X_2 + ... + X_n.$ This is a new random variable. The smallest value it can take on is *0* (which happens only if all of the $X_i's$ are *0),* and the largest value it can take on is (multiple choice: $0,1,np,$ or $n$ ???) The expected value for $X_{sum}$ turns out to be $np.$ Now, if all the $X_i's$ are **independent,** then $X_{sum}$ has a special distribution called the **binomial** distribution. Here is what it means for the $X_i's$ to be independent: the value that is taken on by any one of the $X_i's$ is not influenced by the values that are taken on by any of the other $X_i's$. It is easiest to see what 'independent' means by thinking of examples where random variables are **not** independent. Suppose $X_1$ is the random variable that indicates that Mary Ulauwe, aged 9, has had chicken pox, and $X_2$ indicates that her brother, Samuel Ulauwe, aged 7, has had chicken pox. Then $X_1$ and $X_2$ are not independent – chicken pox is extremely contagious – if one child has it, there is a 95% chance that his or her siblings will catch it also.

An example of two random variables that are almost certainly independent: let $Y_1$ represent the event that Mary Ulauwe flips a coin and it comes up heads, and $Y_2$ similarly represents the event that her brother Samuel flips a coin and it comes up heads. The outcomes of the two coin-flips almost certainly will not influence each other.

The fact that $X_{sum}$ has a binomial distribution is represented as follows:

$$X_{sum} \sim Binom(n, p).$$

It is possible for $X_{sum}$ to take on any value between *0* and *n.* The probability that $X_{sum}$ takes on the value *j* is given by the formula:

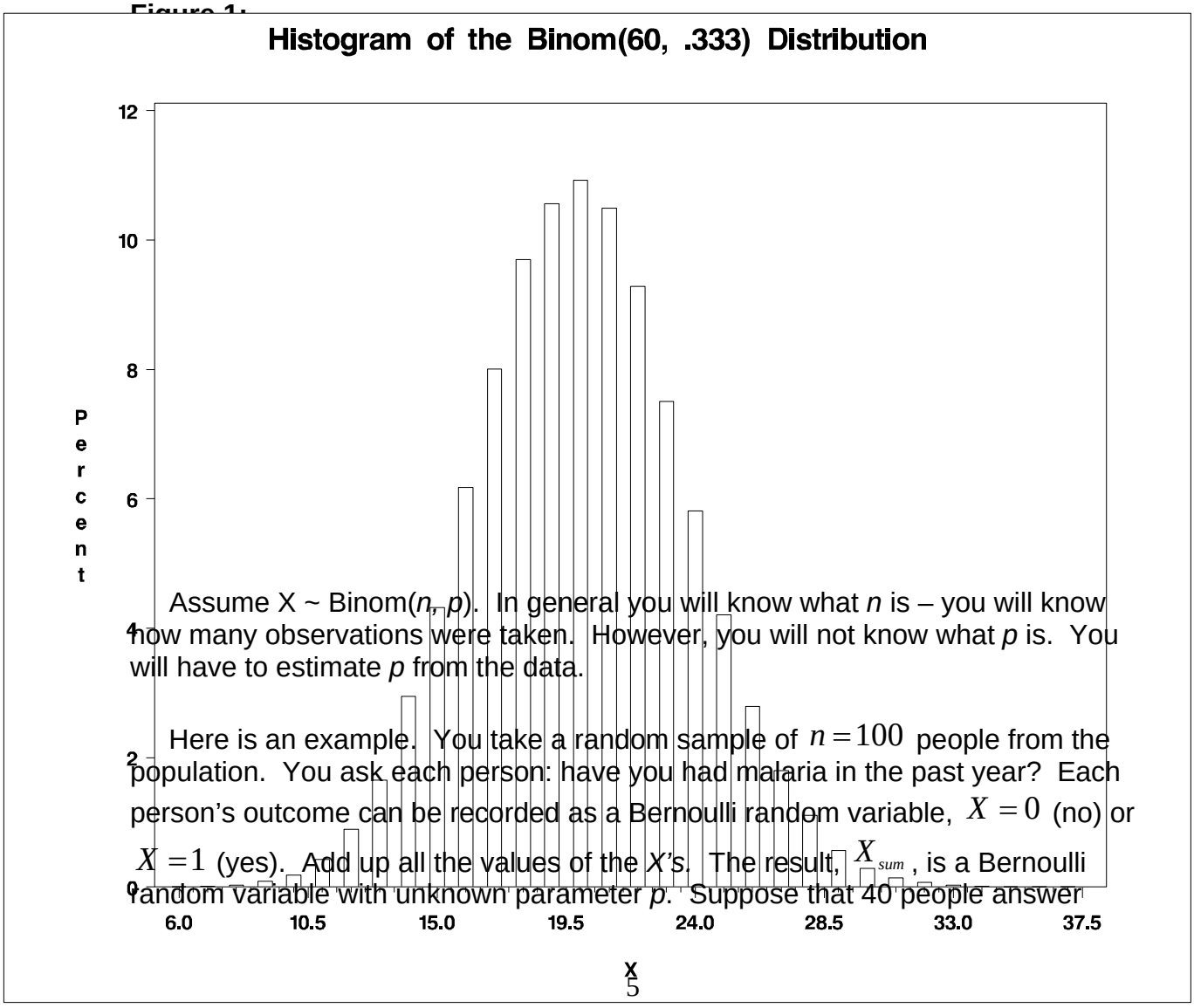
[1] $$prob(X_{sum} = j) = \binom{n}{j} \cdot p^j \cdot (1-p)^{n-j},$$

where $\binom{n}{j} = \frac{n!}{j! \cdot (n-j)!}$, and $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1.$ (pronounced as "n factorial"). Similar definitions apply for *j!* and *(n – j)!.*

**Exercise 2:** Suppose $X_1, X_2, X_3, X_4, X_5$ are independent Bernoulli random variables, all with parameter $p = .30$. What is the probability that $X_{sum}$ equals 3? What is the probability that $X_{sum} \leq 2$? What is the probability that $X_{sum} \geq 4$?

The behavior of binomial random variables is central to much of categorical data analysis. Below is a histogram for X ~ Binom(60, .333):

**Figure 1:**




Assume X ~ Binom(*n*, *p*). In general you will know what *n* is – you will know how many observations were taken. However, you will not know what *p* is. You will have to estimate *p* from the data.

Here is an example. You take a random sample of $n = 100$ people from the population. You ask each person: have you had malaria in the past year? Each person's outcome can be recorded as a Bernoulli random variable, $X = 0$ (no) or $X = 1$ (yes). Add up all the values of the *X's*. The result, $X_{sum}$, is a Bernoulli random variable with unknown parameter *p*. Suppose that 40 people answer

yes, they have had malaria in the past year. That is, $X_{sum} = 40.$ In this case the best estimate of the unknown parameter *p* is $\hat{p} = X_{sum} / n = 40/100 = 0.40$.

Note that I have used $\hat{p}$ here to denote a ***sample estimate*** of the value of the unknown parameter $p$. It is important to distinguish between **parameters,** the true values of which generally are not known, and **estimates**, which are based on a sample of data.

In general it is difficult to compute binomial probabilities. This is because of all the factorial expressions in equation [1]. However many computing packages (for example, Stata) will do the arithmetic for you.

Suppose you flip a coin 100 times. Let $X_i$ be the Bernoulli random variable which indicates that the *i-th* flip comes up heads; that is, $X_i = 1$ if heads comes up and $X_i = 0$ if tails comes up. Let $X_{sum} = \sum X_i$ be the binomial random variable which is the sum of the 100 Bernoulli random variables. If the coin is fair, we expect that $E(X_{sum}) = 50.$ How likely is it that when you flip the coin 100 times, the observed value of $X_{sum} = 40$ or less?

The answer is obtained from equation [1]; it may be written as

[2] $$prob(X_{sum} \le k) = \sum_{j=0}^{k} \binom{n}{j} \cdot p^j \cdot (1-p)^{n-j}$$

In this case we assume:

$$n = 100$$
$$k = 40$$
$$p = .5$$

**Stata** has a function which computes the probability in equation [2]; you can use it by typing

display binomial(100, 40, .5).

Stata displays the answer as .0288.

| Exercise 3: |
|---|

3.1 Compute the probability that, if you flip a fair coin 100 times, then you will obtain 41 or more heads.

3.2 Compute manually: $prob(X \leq 2)$, where *X* is the number of heads in 5 flips of a fair coin.

3.3 Use Stata or a similar statistical package to compute $prob(X \geq 70)$, where *X* is the number of heads in 100 flips of a fair coin.

Here is how binomial probabilities get used in *statistical inference.* You assume a *null hypothesis* $H_0$. For example, $H_0$: coin is fair. Another way to state this is: $H_0$: $prob(X = 1) = 0.5$, where *X* is the variable which is 1 if the coin flip comes up heads, 0 if tails. Now suppose you flip the coin 100 times, and it comes up heads 40 times and tails 60 times. What is the probability this would happen, if the coin were fair? More precisely, if the coin is fair, what is the probability that in 100 flips, it would come up heads in 40 or fewer flips? The answer as given above is $binomial(100,40,.5) = 0.0288$. This is a fairly low probability. Should you reject the null hypothesis, and conclude that the coin you were flipping is **unfair?**

The answer to this depends on your threshold for rejection of a null hypothesis. This is the *significance level* of your statistical test. The significance level is frequently taken to be 0.05. There is nothing magic about 0.05, but it is widely used as a criterion for making the decision to reject the null hypothesis. The *p-value*, $p = 0.0288$, for the coin-flip experiment just described, is the probability that you would observe 40 or fewer heads in 100 flips of a fair coin. Since $0.0288 < 0.05$, if 0.05 is your chosen significance level, you would reject the hypothesis that this coin is fair.

This is in fact fairly weak evidence of unfairness of a coin. If you repeated the experiment of flipping the coin 100 times, you would observe 40 or fewer heads in about 2.88% of those experiments. This is not all that unlikely. It could be that in this case, the significance level of 0.05 is too large.

But now suppose you flipped the coin 1000 times. What is the probability that it would come up heads in 40% or fewer of the flips? This means that you want to compute $binomial(1000,400,.5)$. If in Stata you type in

**display binomial(1000, 400, .5)**,

Stata will print out: **1.364e-10**, which is the same thing as 0.0000000001364.

This is a REALLY small number. If something like this occurred in flipping a coin 1000 times, you would have observed a very rare event – PROVIDED it was a fair coin. You would be strongly tempted to conclude it wasn't fair, that it was weighted somehow so it is more likely to come up tails than heads.

Binomial probabilities apply to more than just fair coins. Suppose you are examining records of earthquakes of magnitude 6.0 or greater. You find records for the 1000 most recent earthquakes of magnitude 6.0 or greater. You find that 60 of these earthquakes occurred during the month of April. There are 30 days in April and, on average, 365.25 days in a year. If earthquakes are equally likely to occur on any day of the year, you would expect that the probability that an earthquake would occur in April would be

$$30/365.25 = 0.08214$$

So in 1000 earthquakes, you would expect that about 82 would occur in the month of April. How likely is it that you would observe 60 or fewer of the 1000 earthquakes to be in the month of April?

You can obtain the answer from Stata by typing:

```
display binomial(1000, 60, .08214).
```

Stata says: 0.00482563. So this is a rather small probability, less than 0.005. You would likely reject the null hypothesis. In this case the null hypothesis is

$H_0$: prob(earthquake occurs in April) = 0.08214.

That is, you would conclude that earthquakes are *less* likely to occur in April than in other months.

Exercise 4: Suppose that of the 1000 most recent earthquakes, you observed that 110 of them occurred in September. Test the hypothesis that earthquakes are no more likely to occur in September than in other months.

Exercise 5: The expected ratio of boy babies to girls is 105:100. In recent years in China, the observed ratio of boy babies to girls was 116:100. Is this ratio significantly higher than expected?

[State a null hypothesis as the probability that a baby is a boy. Then assume that among 1000 babies born in China, a certain number were boys. Then test your null hypothesis using the binomial distribution.]

## A.4 Is 49% Different from 50%?

49% is quite close to 50%. But is it significantly different?

You have a random sample of 100 people. Your null hypothesis is that the true proportion of females in the population is 0.50. However, in your sample, 49 of 100 are female and 51 are male. Stata says that

Binomial(100, 49, .5) = 0.4602.

This p-value is clearly not close to statistically significant.

But now suppose your sample is n = 1000, and again 49% (490) are female. For this, Stata says:

Binomial(1000,490,.5) = 0.2739.

That's closer to "significant", but clearly still larger than the usual threshold of 0.05.

So now you increase your sample to n = 10,000. 49% of 10,000 is 4900. Stata gives:

Binomial(10000, 4900, .5) = 0.0232.

Conclusion: whether a proportion differs significantly from the null hypothesis depends on how large the sample is. Even 49.9% will be significantly different from 50% if the sample size is big enough.

This illustrates the difference between *statistically significant* and *clinically significant.* If you were testing a new drug against an old drug, and the new drug had a 49% failure rate, whereas the old drug had a 50% failure rate, would you prefer the new drug? Not such a simple question. You would need to compare the frequency of adverse effects between the drugs. The cost would likely be different. One drug might taste worse than the other one, which sounds like a minor problem, but it could mean that some people would refuse to take the worse-tasting drug. It is true that if you were comparing a 49% failure rate to a 50% failure rate, and *the two drugs were completely equivalent otherwise,* you would prefer the new drug. But it is almost never true that the two drugs would be equivalent otherwise.

## A.5 Too Good to be True?

Suppose you are flipping a coin that you believe to be fair. You flip it 10,000 times. The **expected** number of heads is 5,000. What does this mean? Does it mean that you are more likely to come up with 5,000 heads than any other number? Or that there is a high probability that you come up with 5,000 heads? If you observe exactly 5,000 heads in 10,000 flips, is someone going to accuse you of faking the data, getting exactly 5,000 is just … too good to be true?

You can compute the probability of getting exactly 5,000 heads using the Stata Binomial function. Recall that binomial(n, k, p) is the probability of getting k OR FEWER events. So the probability of obtaining exactly 5,000 heads given that the null hypothesis is $H_0$: p = 0.50 is true is:

binomial(10000, 5000, .5) – binomial(10000,4999, .5).

Stata says:

binomial(10000, 5000, .5) = .50398932 and
binomial(10000, 4999, .5) = .49601068.

Therefore prob(exactly 5000 heads in 10000 flips, p = .5) = 0.00797864.

So that is a pretty small number. Not likely to occur by chance. But you can similarly find that

prob(exactly 4999 heads in 10000 flips, p = .5) = .49601068 - .48803363

= .00797705

And prob(exactly 5001 heads in 10000 flips, p = .5) = .51196637 - .30398932

= .00797705,

both of which are just slightly smaller than the probability of getting exactly 5000 heads.

You conclude that (1) the probability of getting exactly 5000 heads is not very high, and (2) the probability of getting exactly 5000 heads is still higher that the probability of getting any other number, assuming the coin is fair (p = .5).

The issue of 'too good to be true' has arisen regarding the work of Gregor Mendel, a pioneering geneticist, who carried out experiments on varieties of peas. Mendel tabulated the numbers of peas with certain inherited characteristics. He developed some basic theory to compute the expected numbers, assuming random mixtures of genes from the parent pea plants. He then reported his **observed** numbers. They were in extremely close agreement

with the expected numbers. The agreement was so close that the famous statistician, R. A. Fisher, suspected that Mendel was changing his data to agree better with the expectations:

**Table 1:**

**Mendel's Data on Inherited Characteristics of Peas**

| Category | Predicted Number and Percent | Reported (by Mendel) Number and Percent |
|---|---|---|
| 1 | 5493 (75%) | 5474 (74.74%) |
| 2 | 1831 (25%) | 1850 (25.25%) |

What do you think? Do Mendel's data look too good to be true?

## A.6 Confidence Intervals

Take a random sample of 1000 people in Kenya. Ask each one: have you had malaria? A total of 663 of them say yes. So in your sample, the proportion of people who have had malaria is 0.663. Of course you don't know the true proportion in the whole population. It is not practical for you to ask everyone, since there are over 37 million people in Kenya. You would like to say: the proportion of people who have had malaria is 0.663. But you are not certain of this answer. Maybe your sample was not typical. You need to qualify your estimate by a statement about the uncertainty of the estimate.

This is done by specifying a **confidence interval.** It will turn out that a 95% confidence interval for the true proportion who have had malaria is

(0.632, 0.692)

You can obtain this from Stata by entering:

```
cii 1000 663
```

What does this mean? What is a 95% confidence interval?

Here first is what it **doesn't** mean. It **doesn't** mean that there is a 95% chance that the true value of the proportion is between 0.632 and 0.692.

That is what you would like for it to mean, but it is not correct. Here is the right interpretation. Suppose you take repeated samples of size 1000. You get results like the following

**Table 2: Results of Repeated Sampling**

| Number In Sample | Number who have had malaria | Proportion who have had malaria | 95% Confidence Interval |
|---|---|---|---|
| 1000 | 663 | 0.663 | (0.632, 0.692) |
| 1000 | 650 | 0.650 | (0.619, 0.680) |
| 1000 | 677 | 0.667 | (0.637, 0.696) |
| 1000 | 620 | 0.620 | (0.589, 0.650) |
| 1000 | 701 | 0.701 | (0.671, 0.729) |
| … | … | … | … |

Note that the 95% confidence interval changes with each sample. The true value is a fixed but unknown number. It either lies inside of a given confidence interval or it doesn't. If it is inside of a given confidence interval, the probability that it is inside is 1.00. If it is not, that probability is 0.00. The point being, the probability that the true value is inside of a specified 95% confidence interval is not 0.95. It is either 0 or 1. Say, for example, the true value is 0.625. In the 5 samples above, 0.633 lies within the 95% confidence interval in samples 2 and 4.

So what does "95% Confidence Interval" actually mean?

Here is how it works. You take a sample. There is an algorithm for computing the lower and upper bounds in the 95% confidence interval. Stata and other statistical packages use that algorithm. Those lower and upper bounds are **variables;** they depend on what your sample says. To say that (lower_bound, upper_bound) is a 95% confidence interval for the true value means that if you repeat the whole experiment of sampling 1000 people over and over again, and compute the confidence interval for each sample, then 95% of the time, the true value will lie between lower_bound, upper_bound.

Exercise 6: Assume you take a random sample of 400 college students. You find that 168 of them are male and 232 are female. Use Stata to compute a 95% confidence interval for the true proportion of female students.

You can compute **approximate** 95% confidence intervals for proportions in the following way, without having to use Stata. First, take a random sample of size *n*. Assume that *m* people in your sample have the characteristic you are interested in (for example, *m* of the *n* people sampled say they have had malaria. The sample proportion is

$$\hat{p} = m/n .$$

Then an approximate 95% confidence interval for the true proportion is:

[3] (95% CI):
$$lower\_bound = \hat{p} - 1.96 \cdot \sqrt{\hat{p} \cdot (1-\hat{p})/n},$$
$$upper\_bound = \hat{p} + 1.96 \cdot \sqrt{\hat{p} \cdot (1-\hat{p})/n}.$$

Exercise 7: Repeat Exercise 6, but using the formulas just given for the lower and upper bounds for the 95% confidence interval. How different are your answers from those given by Stata?

It is possible to compute confidence intervals for other levels of confidence. For example, you might want to compute 99% confidence intervals for a certain proportion. This can be done in Stata by entering:

cii 1000 663, level(99)

You should think about this a bit. Is the 99% confidence interval going to be bigger, or smaller, than the 95% confidence interval?

The answer that Stata comes back with is (0.632, 0.701). So the 99% confidence interval is bigger.

You can compute this by hand also as given above in [3], but one thing has to change: you need to replace 1.96 by 2.57:

[4] (99% CI):
$$lower\_bound = \hat{p} - 2.57 \cdot \sqrt{\hat{p} \cdot (1-\hat{p})/n}$$
$$upper\_bound = \hat{p} + 2.57 \cdot \sqrt{\hat{p} \cdot (1-\hat{p})/n}.$$

Again you can see that the 99% confidence interval has to be bigger. Note that in these formulas, the sample estimate $\hat{p}$ is exactly in the middle of the confidence interval.

Exercise 8: Use [4] to compute the 99% CI for the true proportion, given that 120 out of a random sample of 150 people say they have had malaria.

### A.7 Binomial Variables in Data Files

In Stata, as in most statistical packages, you can analyze data in files. The following data file, "hsb2.dta" related to data on people in high school and beyond. It can be downloaded from the internet by typing in the following command in Stata:

use http://www.ats.ucla.edu/stat/stata/hsb2

This file has 200 observations, and a number of variables for each observation. One of the variables is **female**. This variable takes on only two values: 1 (for female students) or 0 (for male students). If you type in

summarize female

in Stata, the following table is displayed:

| variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| female | 200 | .545 | .4992205 | 0 | 1 |

("Std. Dev.", standard deviation, will be explained below).

Note that a mean of .545 implies that 54.5% of the people represented on the file were female, that is, the sample proportion is $\hat{p} = 109/200 = 0.545$.

If you want the 95% confidence interval for the true proportion of females, you can enter the following in Stata:

ci female

This gives the 95% confidence interval as (0.475, 0.615).

Now, to define standard deviation: you first have to define **variance.** Given a random variable *X*, the variance is defined in terms of the expectation. Here is a reminder of the definition of expectation:

If a random variable *X* takes on values $x_1, x_2, x_3, ..., x_n$, where the probability of taking on the value $x_i$ is $p_i$. Then the expected value is the *weighted average,*

$$E(X) = p_1 \cdot x_1 + p_2 \cdot x_2 + ... + p_n \cdot x_n.$$

Given this, the definition of *Var(X)* is:

$$Var(X) = E(X^2) - (E(X))^2 .$$

This may look like double-talk: you could read it as, the expectation of X squared minus the expectation of X squared. But there is an important difference between the two expectations on the right side. In the first one, $E(X^2)$, the variable *X* is first squared, then its expectation is computed. In the second term, $(E(X))^2$, the expectation of *X* is computed, and then that value is squared.

Here is how it works for the simplest of random variables, the Bernoulli random variable *X*. Let $p = prob(X = 1).$ Since X can take on only the two values 0 and 1, and since 0-squared is 0 and 1-squared is 1, $X^2$ is always equal to $X$. Therefore $E(X^2) = E(X).$ But $E(X) = p$, so $E(X^2) = p$ also. On the other hand, $(E(X))^2 = p^2$. This means that

$$Var(X) = E(X^2) - (E(X))^2 = p - p^2 = p \cdot (1 - p).$$

The **standard deviation** of a random variable X is by definition equal to the square root of the variance:

$$SDev(X) = \sqrt{Var(X)} .$$

Therefore, for a Bernoulli random variable *X* with probability *p* of being equal to 1, the standard deviation is:

$$SDev(X) = \sqrt{p \cdot (1 - p)} .$$

Now, a binomial random variable is the sum of several independent Bernoulli random variables:

$$X_{sum} = X_1 + X_2 + ... + X_n.$$

There is a fact about independent random variables: namely, the variance of a sum of random variables equals the sum of the variances. (This is **not true** for random variables which are not independent.) This means that

$$Var(X_{sum}) = Var(X_1) + Var(X_2) + ... + Var(X_n).$$

But if all the $X_i$ are Bernoulli random variables with parameter *p* and are independent, then

$$Var(X_{sum}) = p \cdot (1-p) + p \cdot (1-p) + ... + p \cdot (1-p) = n \cdot p \cdot (1-p).$$

Therefore,

$$SDev(X_{sum}) = \sqrt{n \cdot p \cdot (1-p)}.$$

Example: Suppose *X* is the count of females in a random sample of size 200, where the probability in the general population of being female is 0.5. Then

$$SDev(X) = \sqrt{200 \cdot 0.5 \cdot 0.5)} = \sqrt{50} = 7.071.$$

But what are standard deviation and variance, really?

The variance and standard deviation of *X* are measures of how "dispersed" the values of *X* are, or how "spread out" they are. A random variable *X* which is very concentrated around its expected value will have a small variance and small standard deviation. A random variable which is highly dispersed takes on a wide range of values; it will have a large variance and a large standard deviation. Examples: levels of sodium in human blood have a very small standard deviation (about 5 milli-equivalents per liter where the average is about 140 meq/L), whereas low-density lipoprotein (LDL) levels in serum have a standard deviation of about 25 mg/dL, where the average is about 100 mg/dL. The reason for this is, the human body cannot tolerate levels of sodium that deviate much from the average, whereas the body can tolerate large deviations away from the average in serum LDL levels.

**A.8: Binomial Proportions:**

We need to define a new kind of random variable which is related to the binomial random variables. This is the *binomial proportion*. This may lead to some confusion. If $X_{sum}$ is a binomial random variable representing, for example, the number of people in a sample of 1000 who have had malaria, then $X_{sum}$ is nothing more than the **count** of people in the sample who have had malaria. The ***proportion*** of people who have had malaria is the count divided by the total; that is,

$$binomial\_proportion = \hat{p} = X_{sum} / n.$$

A binomial random variable will range between *0* and *n*. A binomial *proportion* will range between 0 and 1.

We will want to compute the variance and standard deviation of the binomial proportion $\hat{p}$. To do this, you need to know the following little fact about variances of random variables:

$$Var(f \cdot X) = f^2 \cdot Var(X)$$

(If you are inclined to do the math, this fact is easy to prove from the definition of variance given above.)

Here the factor $f$ is just a constant.

Referring to [5], the expression for a binomial proportion,

$$\hat{p} = X_{sum} / n.$$

So in this expression, let $f = 1/n.$ This means that

$$\begin{aligned} Var(\hat{p}) &= f^2 \cdot Var(X_{sum}) = (1/n^2) \cdot Var(X_{sum}) = (1/n^2) \cdot n \cdot p \cdot (1-p) \\ &= p \cdot (1-p)/n. \end{aligned}$$

And since standard deviation is the square root of variance,

$$SDev(\hat{p}) = \sqrt{p \cdot (1-p)/n}.$$

(This should remind you of part of the formula for lower and upper confidence limits!)

Exercise 9. Use the formulas above to compute the standard deviation for the sample proportion of females in the hsb2 file.

Exercise 10. Polling organizations like to make predictions of regarding elections. Assume there are two candidates, Mr. A and Mr. B. The polling organization typically assembles a random sample of potential voters of size about n = 1000. They ask each person in the sample if they are going to vote for Mr. A or Mr. B. They count up the answers. Suppose for example 470 people

prefer Mr. A, and 530 people prefer Mr. B. If the proportion favoring one candidate is significantly different from 0.50, then the polling organization will say "Mr. A (or Mr. B) holds a significant lead in the race". Or if the proportion is not significantly different from 0.50, they will say "Mr. A and Mr. B are statistically tied."

So: with the split being 470 (A) and 530 (B), what will the polling organization say?

What is a 95% confidence interval for the true proportion who would vote for Mr. A?

If you were to state a null hypothesis for this situation, what would it be?

# B. Two or More Samples

### B.1 2 x 2 Tables

What we have done in Section A is about estimates of sample proportions for a single group, and comparisons of those estimates with hypothesized 'true' proportions. We can estimate proportions, do tests to see if they are significantly different from hypothesized proportions, and compute confidence intervals for the true proportions.

In this section we will consider situations where you want to estimate and compare proportions for two or more groups.

Assume you carry out a clinical trial for prevention of herpes, using a new vaccine. You find a sample of 200 people who have not had herpes. You randomize 100 of them to have the active vaccine (group A) and 100 to have an inactive placebo (group P). You follow each person for a 3-year period. At the end you count up the number of people in each group who had a diagnosis of herpes.

You can represent the data in a 2 x 2 table, as follows:

**Table 3:** A 2 x 2 Table: Herpes Clinical Trial

| | E+ | E- | Row margins |
|---|---|---|---|
| D+ | *a= 28* | *b= 40* | *n = 68* |
| D- | *c= 72* | *d= 60* | *m = 132* |
| Col margs. | *r=100* | *s= 100* | *N =200* |

Here E+ and E- represent drug treatment groups; E+ = active vaccine, and E- = placebo vaccine. D+ and D- represent disease outcomes: D+ indicates that the patients were positive for herpes, while D- indicates they did not have herpes. The main entries in the table a, b, c, and d are counts of people; for example, 60 people are in the (D+, E-) cell indicating positive for herpes and placebo vaccine.

The numbers n and m are called **row margins;** r and s are called **column margins.**

The object of a clinical trial like this one is to determine which drug is better. The better drug is the one which has a higher probability of a good outcome. The table makes it possible to estimate the proportion of good outcomes in the two groups. Note that in this case, a good outcome is D-. The proportion of people who have a good outcome in the E+ group is $\hat{p}_{E+} = 72/100 = 0.72$, while the proportion of people having a good outcome in the E- group is $\hat{p}_{E-} = 60/100 = 0.60$. This looks like a fairly strong difference, but is it significant?

The 'true' proportions, $p_{E+}$ and $p_{E-}$, are not known. $\hat{p}_{E+}$ and $\hat{p}_{E-}$ are **estimates** of these unknown true values. Unlike what was done in the previous section, we do not compare $\hat{p}_{E+}$ and $\hat{p}_{E-}$ to some fixed constants. We need to compare them to each other. Obviously they are different. The question is, do they differ from what you might expect to get by chance? Can we assign a probability to this?

You can state a null hypothesis. It would be written as $H_0$: $p_{E+} = p_{E-}$. (Note that it would be a logical mistake to state the null hypothesis as $H_0$: $\hat{p}_{E+} = \hat{p}_{E-}$. The

true proportions are not known; the observed proportions $\hat{p}_{E+}$ and $\hat{p}_{E-}$ are estimates of the true proportions, and we already know that they are not equal.)

There are several ways to test the null hypothesis. Some you can do 'by hand' and some require a computer. Here are some possibilities:

1. Use a web app called GraphPad:
   http://www.graphpad.com/quickcalcs/contingency1.cfm

This program will ask you to enter the 4 numbers in a 2 x 2 table. If you enter the numbers shown above (28, 40, 72, 60) the program will then ask which test you want to use. The choices are Fisher's Exact Test, Chi-square with Yates' correction, or Chi-square without Yates' correction. The 'recommended' test for a variety of reasons is Fisher's Exact test. The program also asks if you want to do a two-tailed test or a one-tailed tests. In each case the program gives you back a p-value. In most cases you will choose the two-tailed test. The p-values returned by the program are the following:

Fisher's Exact test: p = 0.1003

Chi-square with Yates' Correction: p = 0.1006

Chi-square without Yates' Correction: p = 0.0733

None of these would be considered significant by the usual 0.05 criterion.
The p-value for the chi-square without Yates' correction is smaller than that for the Fisher Exact test or the chi-square with Yates' correction. This test tends to be less 'conservative' than the other two tests – that is, it is more likely to reject the null hypothesis even if it is true. This is part of the reason for preferring the Fisher Exact test – it is not necessarily a good thing to reject the null hypothesis when it is in fact true.

2. Use a statistical package like Stata. This is not quite as easy. Stata has a function to compute either the Fisher or Chi-square p-values, but the function operates on a data file rather than having you just enter the four numbers. For the data above, you would need a data file with 200 observations. The file would be structured as follows:

| Observation | E | D |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |

| | | |
|---|---|---|
| … | … | … |
| 99 | 1 | 1 |
| 100 | 1 | 0 |
| 101 | 0 | 1 |
| 102 | 0 | 0 |
| … | … | … |
| 199 | 0 | 0 |
| 200 | 0 | 1 |

Here all of the first 100 observations are in Group E+, and all of the 2nd 100 are in group E-. Among the first 100, 28 have outcome D+ and 72 have outcome D-. In the second 200, all are in Group E- and 40 are D+, and 60 are D-. This needs to be the active data set (*.dta) in Stata. The command for Fisher's Exact test is

```
tabulate e d, exact
```

where e and d are the column headings in the dataset, and e is coded as 1 for E+, 0 for E-, and d is coded similarly. In this case Stata returns
(1) the two-sided Fisher Exact test p-value, 0.100, and (2) the one-sided p-value (0.050). [Note: It is not always true that the one-sided p-value is half of the two-sided p-value.]

3. To compute an approximate p-value by hand: Go back to the 2 x 2 table above, and compute

$$\chi^2 = \frac{N \cdot (ad - bc)^2}{n \cdot m \cdot r \cdot s}$$

This is the uncorrected chi-square statistic. To find the associated p-value, you need chi-square tables. Here is an excerpt from a typical table:

**Table 4:** The Chi-square Distribution

| df | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | --- | --- | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 |
| … | | | | | | | | | |

The numbers across the top correspond to p-values. The first column labeled 'df' stands for 'degrees of freedom'. The entries in the table are values of the chi-square statistic. The degrees of freedom in this case is 1, so we need only look at the first row in the table. The table does not give exact p-values. What it says,

for example, is that if the chi-square statistic $\chi^2$ as computed above is between 3.841 and 5.024, then the p-value is between 0.05 and 0.025.

In this case,

$$\chi^2 = 200 \cdot (28 \cdot 60 - 40 \cdot 72)^2 / (88 \cdot 112 \cdot 100 \cdot 100) = 3.209,$$

which, in the first row of the table above, is between 2.706 and 3.841. This means that the p-value is between 0.05 and 0.10. Note that this is in agreement with the uncorrected chi-square p-value (0.0733) obtained from GraphPad. As a rule, the uncorrected chi-square p-value is smaller than the Fisher Exact test p-value.

4. The **Yates-corrected** chi-square statistic is defined as

$$\chi_c^2 = \frac{N \cdot (|ad - bc| - N/2)^2}{n \cdot m \cdot r \cdot s}.$$

For the data in the table above, one obtains $\chi_c^2 = 2.696.$ From the top row of the chi-square table above, you can see that the p-value must be just slightly bigger than 0.10. The Yates-corrected chi-square agrees more closely with the Fisher Exact test than does the uncorrected chi-square.

The theoretical underpinnings of these tests are beyond the scope of these notes. Briefly, the Fisher Exact test is based on a distribution of numbers in 2 x 2 tables, called the hypergeometric distribution. The chi-square tests are based on the fact that the binomial distribution is fairly well approximated by the **normal** distribution (which will not be discussed here). Because the chi-square tests are based on approximations, they should not be used when the sample sizes are too small. How small? The sample size in each cell of the 2 x 2 table should be at least 5.

## B.2: 2 x 2 Tables: Some Definitions and Terminology

**Table 5: Generic 2 x 2 Table**

| | *E*+ | *E*- | Row margins |
|---|---|---|---|
| *D*+ | *a* | *b* | *n* =*a*+*b* |
| *D*- | *c* | *d* | *m* = *c*+*d* |
| Col margs. | *r*=*a*+*c* | *s*=*b*+*d* | *N* |

**Exposure:** 'Exposure' in a clinical trial refers to exposure to the drug or agent being tested. If *E*+ denotes active drug and *E*- denotes placebo, then we say the *E*+ group is exposed and the *E*- group is unexposed.

**Risk:** Risk is the probability of having the disease outcome of interest, that its, D+. In general you expect that risk is different for the exposed people than for the unexposed people. The risks for an exposed person and an unexposed person respectively are:

risk of disease given exposure = $prob(D+ \mid E+).$

risk of disease given no exposure = $prob(D+ \mid E-).$

These are the 'true' risks in the population. The entries *a, b, c*, and *d* in the 2 x 2 table are **data** from a sample of the population. These data provide estimates of the risks, as follows:

estimated risk given exposure = $a/(a+c)$, and

estimated risk given no exposure = $b/(b+d).$

Note that, since risk is a probability, risk is always a number between 0 and 1.

**Odds:** Odds are closely related to risk. The term is related to betting odds, which you might encounter if you spend time at a horse-race track. Specifically, odds is defined as

$$odds = risk/(1-risk).$$

For the 2 x 2 table, we refer to odds of disease given exposure and odds of disease given no exposure. These are defined as:

odds of disease given exposure = $\dfrac{prob(D+|E+)}{1-prob(D+|E+)}$, and

odds of disease given no exposure = $\dfrac{prob(D+|E-)}{1-prob(D+|E-)}$.

Again, these are population parameters which must be estimated from the data in the 2 x 2 table:

estimated odds of disease given exposure = $\dfrac{a/(a+c)}{c/(a+c)} = a/c$, and

estimated odds of disease given no exposure = $\dfrac{b/(b+d)}{d/(b+d)} = b/d$.

Note that odds is always a non-negative number, but it is not restricted to being less than 1. Also, note that odds can be undefined (infinite) if one of the denominators is 0.

**Exercise 11:** Compute the estimates of risk and odds given exposure and of risk and odds given no exposure, for the data in the 2 x 2 table for the clinical trial involving herpes.

### B.3: Risk Ratio and Odds Ratio:

Risk ratio (also called **relative risk**) is defined as the ratio of the risk for exposed people divided by the risk for unexposed people. Specifically, it is:

$$\text{Risk ratio} = \frac{prob(D+\,|\,E+)}{prob(D+\,|\,E-)}.$$

The risk ratio can be estimated from the data in the 2 x 2 table:

$$\textbf{Estimated risk ratio} = \frac{a/(a+c)}{b/(b+d)} = \frac{a\cdot(b+d)}{b\cdot(a+c)}.$$

Odds ratio (or relative odds) has an analogous definition:

$$\text{Odds ratio} = \frac{odds(D+\,|\,E+)}{odds(D+\,|\,E-)}.$$

This can be estimated as follows:

$$\text{Estimated odds ratio} = \frac{a/c}{b/d} = \frac{ad}{bc}.$$

This is sometimes called the *cross-product ratio.*

**Exercise 12:** Compute the risk ratio and the odds ratio for the data in the herpes clinical trial.

Now, suppose that the herpes vaccine has **no effect** on a person's chances of getting herpes. This means that

$$prob(D+\,|\,E+) = prob(D+\,|\,E-).$$

Note that, from the definition of risk ratio given above, this implies that

Risk ratio = 1.00.

The same is true for the odds ratio: Odds ratio = 1.00.

In a typical clinical trial, the null hypothesis is that there is no difference between the two drug groups; the probability of D+ is the same whether you get E+ or E-. Therefore another way to state the null hypothesis is:

Risk ratio = 1.00 or

Odds ratio = 1.00.

If you carried out a clinical trial and you obtained the following data,

**Table 6:** A Hypothetical Clinical Trial

| | *E+* | *E-* | Row Margins |
|---|---|---|---|
| *D+* | *36* | *48* | *84* |
| *D-* | *54* | *72* | *126* |
| Col Margs. | *90* | *120* | *210* |

then when you compute the relative risk or the odds ratio, you will get exactly 1.00 for both. This happens essentially because the second row is a multiple of the first row: 1.5 *(first row) = second row. In this case there is no evidence whatsoever that *E*+ has any different effect from *E*- on the disease outcome.

### B.4: Observed vs. Predicted

There is another way to think of this which can be helpful. Suppose we knew the row and column margins of the table, but not what the cell numbers are inside:

**Table 7: Known Margins, Unknown Cells**

| | *E+* | *E-* | Row Margins |
|---|---|---|---|
| *D+* | *a* | *b* | *84* |
| *D-* | *c* | *d* | *126* |
| Col Margs. | *90* | *120* | *210* |

How could we compute what the numbers would be in the cells, if the null hypothesis is true?

It turns out that, for example, to figure out what the "b" cell would be, we compute

$$84 \cdot 120 / 210 = 48.$$

That is, we compute the product of the row margin and the column margin corresponding to the "b" cell, and divide by the total in the table.

This is the **predicted value** of the "b" cell, under the null hypothesis. If all the cell values are computed in this way, you will get back Table X above.

To put this another way: if you compute what all the cell contents would be from the margins as we just did for the "b" cell, you will get a table of predicted values for which the null hypothesis is true – that is, the relative risk and the odds ratio for that table will both equal exactly 1.00.

Let's go back to the herpes clinical trial. The 2 x 2 table was:

**Table 8:** Herpes Clinical Trial - Observed

| | E+ | E- | Row Margins |
|---|---|---|---|
| D+ | *28* | *40* | 68 |
| D- | *72* | *60* | *132* |
| Col margs. | *100* | *100* | *200* |

This is the **observed** data. We can compare this to the **predicted** data, where the predicted values are computed as above:

**Table 9:** Herpes Clinical Trial - Predicted

| | E+ | E- | Row Margins |
|---|---|---|---|
| D+ | *34* | *34* | *68* |
| D- | *66* | *66* | *132* |
| Col margs. | *100* | *100* | *200* |

One way to compare the observed and predicted values is to subtract the predicted values from the observed. This would give you the following table:

**Table 10:** Herpes Clinical Trial: Observed - Predicted

| | E+ | E- | Row Margins |
|---|---|---|---|
| D+ | *-6* | *+6* | *0* |
| D- | *+6* | *-6* | *0* |
| Col margs. | *0* | *0* | *0* |

From this you can see **how** the observed data deviates from the data you would expect if the null hypothesis were true: the observed data in the "a" cell, for example, is smaller than the predicted; the observed data in the "b" cell is larger than the predicted. In terms of the medical outcomes: from the "a" cell, there are fewer cases of herpes than you would predict if the vaccine had no effect. That is, the vaccine seems to be effective. In the "b" cell, there are more cases of *D+* than would be predicted under the null hypothesis: which means that in the placebo group, there is an excess of cases of herpes. Again this suggests that the vaccine is effective. However, bear in mind that the Fisher Exact test p-value

for this table was $p = 0.1003$, which means that, although the effect of the vaccine seems to be beneficial, there is not a **statistically significant** effect observed in this clinical trial.

The observed – predicted table gives another way of computing the chi-square statistic. The following is called *Pearson's chi-square*:

$$\chi^2_{Pearson} = \sum_{i=1}^{4} \{(obs_i - pred_i)^2 / pred_i\},$$

where the summation is over all 4 cells in the table. If you do this computation for the herpes clinical trial, you get:

$$\chi^2_{Pearson} = (-6)^2/34 + 6^2/34 + (-6)^2/66 + 6^2/66 = 3.209.$$

This happens to be exactly the same value that you get from the formula for the uncorrected chi-square,

$$\chi^2 = \frac{N \cdot (ad - bc)^2}{n \cdot m \cdot r \cdot s},$$

as was shown above. And as you may recall, the corresponding p-value was 0.0733.

We thus now have 4 ways to compute p-values for 2 x 2 tables: Fisher's Exact test (where you usually need a computer program to do the computations, the Yates-corrected chi-square, the uncorrected chi-square, and Pearson's chi-square. The one which is most often recommended is the Fisher Exact test, but the Yates-corrected chi-square is often used also. There are at least two other ways to compute p-values for 2 x 2 tables, one of which we will explain when we get to logistic regression.

**Exercise 13:** Compute the Pearson chi-square statistic for the following 2 x 2 table:

Table 11: Hypothetical Clinical Trial

| | *E+* | *E-* | Row Margins |
|---|---|---|---|
| *D+* | *100* | *200* | *300* |
| *D-* | *300* | *400* | *700* |
| Col margs. | *400* | *600* | *1000* |

**Back to Relative Risks and Odds Ratios**

Below are two tables where the relative risks are different:

**Table 12A:** Relative Risk = $\frac{20/100}{40/100} = 0.50.$

| | *E+* | *E-* | Row Margins |
|---|---|---|---|
| *D+* | *20* | *40* | *68* |
| *D-* | *80* | *60* | *132* |
| Col margs. | *100* | *100* | *200* |

**Table 12B:** Relative Risk = $\frac{20/820}{40/640} = 0.39$.

| | *E*+ | *E*- | Row Margins |
|---|---|---|---|
| *D*+ | *20* | *40* | 60 |
| *D*- | *800* | *600* | *1400* |
| Col margs. | *820* | 640 | *1460* |

However, the odds ratios for these two tables are exactly the same:

Odds ratio, Table A: $\frac{20 \cdot 60}{40 \cdot 80} = 0.375$

Odds ratio, Table B: $\frac{20 \cdot 600}{40 \cdot 800} = 0.375$.

Why is this? And why is the difference between the odds ratio and the relative risk large in Table A (.5 vs. .375), but not so large in Table B (.39 vs. .375) ?

**Fact 1:** If the positive outcomes (that is, *D*+ outcomes) are relatively rare, then the relative risk and the odds ratio are approximately equal.

To see this, think of risk as a probability, *p*. Then by definition, $odds = \frac{p}{1-p}$. If *p* is small, then *1 – p* is close to 1. This means that

$$risk = p \approx \frac{p}{1-p} = odds.$$

To see this even more concretely, assume risk is small, for example, $risk = .01.$ Then $odds = .01/.99 = 0.0101.$ If the risk is small in both the *E*+ column and the *E*- column, the corresponding odds for both columns will be approximately equal to the risks, and the odds ratio will approximately equal the risk ratio.

**Exercise 14:** Find two 2 x 2 tables which have the same relative risk, but the odds ratios are different.

### B.6: Case-Control Sampling

There are many study designs other than clinical trials. One of the most useful in epidemiologic studies is the case-control design. The basic idea here is the following. You are interested in knowing whether a certain risk factor causes a certain disease. For example, does cigarette smoking cause breast cancer? You identify a group of women who have breast cancer. These are the cases. You also identify a group of women who do not have breast cancer. This is the control group. You may try to select the controls to have similar ages or to live in the same neighborhood as the cases. The two groups do not have to have equal sizes, and in fact in many studies, the control group is 2-4 times larger than the case group.

An important principle here: if you want to study smoking, you must not select either the cases or the controls based on their history of smoking. That is, your selection of both cases and controls must be ***independent*** of their smoking habits or history.

Once you have selected the cases and controls, you then ask them about their smoking history: for example, "Do you now smoke cigarettes? Have you ever smoked at least 100 cigarettes in your lifetime? (if yes) When did you first start smoking cigarettes? For how many years did you smoke? During the years that you smoked, how many cigarettes did you smoke per day, on average?" These are standard questions to determine **exposure** to cigarette smoke.

Another important principle: if you want to study a certain exposure factor in a case-control study, you must not match cases and controls on that factor. So if you matched cases and controls on, for example, body mass index, you would not be able to study the effects of body mass index on incidence of breast cancer. (Do you see why?)

After you have collected the smoking exposure data, you are ready to do statistical analysis. You tabulate what percent of the cases smoked cigarettes and what percent of the controls. You can summarize this in a 2 x 2 table:

| | *E+* | *E-* | Row Margins |
|---|---|---|---|
| *D+* | *50* | *30* | *80* |
| *D-* | *100* | *140* | *240* |
| Col Margs. | *150* | *170* | *320* |

The column labeled '*E+*' denotes smokers (women exposed to smoking), and '*E-*' denotes nonsmokers. The cases are designated by '*D+*' and the controls by '*D-*'. Note that the total number of controls is 3 times larger than the total number of cases.

The statistical analysis in this study is basically the same as it was for the herpes clinical trial. You compute the Fisher Exact test p-value or the chi-square statistic and the corresponding p-value. GraphPad gives the following results:

Fisher Exact test p-value: 0.0018

Yates-corrected chi-square and p-value: $\chi_c^2 = 9.638, p = 0.0019$.

Also, you note that the proportion of smokers is higher among the cases than it is among the controls:

Proportion of smokers among cases: $50/80 = 0.625$,

Proportion of smokers among controls: $100/240 = 0.417$.

Further, you can compute an odds ratio for the table above:

Odds ratio = $\dfrac{50 \cdot 140}{30 \cdot 100} = 2.33$.

This is a fairly large odds ratio; it says that the odds that a case is a smoker are 2.33 times larger than the odds that a control is a smoker.

**Important warning!**

Now, one thing you **cannot** do in a case-control study is compute an estimate of the relative risk. In this case, 'risk' refers to the risk of breast cancer. To

estimate this risk among smokers and nonsmokers, you would need to have a different study structure. You would need to start with a group of smokers and another group of nonsmokers, and then follow them for, say, 30 years (cancer has a long latency period) and then determine which ones had breast cancer during that time. These groups are called 'cohorts'. In the table above, we have a group of smokers (the people in the *E*+ column) and a group of nonsmokers (the people in the *E*- column), but **they were not selected as cohorts**. They were selected **after** it had been determined that they were cases or controls. If you tried to estimate relative risk in the usual way, you would get:

$$relative\_risk\,??? = \frac{prob(D+\,|\,E+)}{prob(D+\,|\,E-)} = \frac{50/150}{30/170} = 1.89\,.$$

But this is wrong, because, for example, *50/150* is not an estimate of $prob(D+\,|\,E+)$. To estimate the latter, you would need to start with a group of smokers and follow them for a period of time. That, however, is not how the cases and controls were selected. First we determined their case or control status, and then we determined their smoking history.

But the estimate of the odds ratio given above is valid! How can this be?

**Table 13: Population Categories and Case-Control and Cohort Sampling**

| Case-Control Status | Exposure Status E+ | E- | Case-Control Sampling |
|---|---|---|---|
| | | | Sampling Fraction $f_1$ |
| D+ | A | B | ----------> $f_1$*A  $f_1$*B |
| | | | Sampling Fraction $f_2$ |
| D- | C | D | ----------> $f_2$*C  $f_2$*D |
| Cohort Sampling | \| $g_1$ V | \| $g_2$ V | |
| | $g_1$*A | $g_2$*B | |
| | $g_1$*C | $g_2$*D | |

In the 2 x 2 table shown here, the entries A, B, C, and D represent numbers of people in the population in the various cells. For example, there are B people in the population who are D+ and E-. **Case-control** sampling is represented by the arrows pointing to the right. In case-control sampling, the sample of cases selected ($f_1$) is the same whether the people are exposed (E+) or unexposed (E-). Similarly there is a constant fraction of controls sampled. Note that $f_1$ and $f_2$ need not be the same. In **cohort sampling,** represented below the table by the vertical lines, the sampling fractions $g_1$ and $g_2$ are the same regardless of the disease status (D+ or D-).

These sampling fractions result in the 2 x 2 tables on the right and below the original population table.

For the lower table, generated by cohort sampling, the **relative risk** would be computed as:

$$RR_{sample} = \frac{g_1 A/(g_1 A + g_1 C)}{g_2 B/(g_2 B + g_2 D)} = \frac{A/(A+C)}{B/(B+D)} = RR_{population}$$

because the factors $g_1$ and $g_2$ cancel out in the numerator and denominator respectively.

For the odds ratio, the same table gives:

$$OR_{sample} = \frac{g_1 A \cdot g_2 D}{g_2 B \cdot g_1 C} = \frac{AD}{BC} = OR_{population}.$$

This means that for **cohort sampling,** the sample estimates of both the relative risk and of the odds ratio would be expected on average to agree with the population relative risk and odds ratio respectively.

It is also easy to show that the estimate of the odds ratio for case-control sampling would be expected to agree with the population odds ratio.

However, if we try to compute the relative risk for the case-control sample, we get the following:

$$RR_{sample} = \frac{f_1 A/(f_1 A + f_2 C)}{f_1 B/(f_1 B + f_2 D)},$$

and the factors $f_1$ and $f_2$ do **not** all cancel. What this means is, case-control samples cannot give valid estimates of relative risk.

**Exercise 15:** Refer to Table 13 above. Assume the following values for counts of people in the population:

A = 50 B = 10

C = 1000 D = 1000

(1) Compute the odds ratio and relative risk for this population data.

(2) Assume a case-control sample is taken with sampling fractions:

$f_1 = 0.8,$
$f_2 = 0.2$

Compute the cell sizes and the odds ratio and relative risk for this sample.

In spite of what I said above, that relative risk cannot be obtained from case-control studies, you will see relative risks in papers on case-control studies, even in good journals. An example is given below. Why is this?

As I noted above also, if the probability of the outcome of interest is small, relative risk and odds ratio are approximately equal. That's one reason. The other reason is, it is easier to understand what a relative risk is – it's just the ratio of two probabilities – but the odds ratio is harder to understand, because in general people don't know exactly what odds are. It's true, there is an exact correspondence between odds and probability – for example, if odds of an event is 1/4 then the probability of that event is 1/5; if the odds equal 1.0, then the probability of the event is 0.50. The relationship between odds and probability is:

$$odds = \frac{prob}{1 - prob} ,$$

so general if you want to solve for the probability in terms of the odds, you get

$$prob = \frac{odds}{1 + odds} .$$

However, given an odds **ratio,** you cannot compute what the corresponding risk ratio (relative risk) would be.

Main point here: You can estimate the odds ratio in clinical trials, cohort studies, and case-control studies. If the disease condition you are interested is rare (low probability), then the relative risk is well approximated by the odds ratio. However in general you cannot reliably estimate relative risk from a case-control study.

There is another reason that odds ratio is frequently used instead of relative risk. This is because if you are analyzing data using **logistic regression** (which will be explained later), you can obtain estimates of odds ratios, but not relative risks.

The following table is from an article in the American Journal of Epidemiology, 2006: Lonn S et al.: Mobile phone use and risk of parotid gland tumor. V164 pp 637-643:

**TABLE 2. Odds ratio* of malignant parotid gland tumors and benign pleomorp mobile phone use in Denmark and Sweden, 2000–2002‡**

| | Malignant parotid gland tumors | | | | Benig | |
|---|---|---|---|---|---|---|
| | No. of cases | No. of controls | Odds ratio | 95% confidence interval | No. of cases | c |
| Frequency of use | | | | | | |
| Never or rarely§ | 35 | 280 | 1.0 | | 35 | |
| Regular use¶ | 25 | 401 | 0.7 | 0.4, 1.3 | 77 | |
| Duration (years) of regular use | | | | | | |
| <5 | 15 | 237 | 0.7 | 0.3, 1.4 | 48 | |
| 5–9 | 8 | 125 | 0.7 | 0.3, 1.8 | 24 | |
| ≥10 | 1 | 30 | 0.3 | 0.0, 2.5 | 5 | |
| Time (years) since first regular use | | | | | | |
| <5 | 14 | 228 | 0.7 | 0.3, 1.3 | 47 | |
| 5–9 | 8 | 128 | 0.7 | 0.3, 1.7 | 23 | |
| ≥10 | 2 | 36 | 0.4 | 0.1, 2.6 | 7 | |
| Cumulative use (hours) | | | | | | |
| <30 | 7 | 110 | 0.7 | 0.3, 1.6 | 20 | |
| 30–449 | 11 | 184 | 0.7 | 0.3, 1.4 | 34 | |
| ≥450 | 5 | 90 | 0.6 | 0.2, 1.8 | 22 | |
| Cumulative no. of calls | | | | | | |
| ≤624 | 5 | 101 | 0.5 | 0.2, 1.3 | 13 | |
| 625–7,349 | 12 | 190 | 0.7 | 0.3, 1.6 | 40 | |
| ≥7,350 | 6 | 95 | 0.7 | 0.3, 2.0 | 21 | |

* Adjusted for age, gender, geographic region, and education.
† Included in only two regions (Stockholm area and Göteborg municipality).
‡ Totals for variables are not equal because of missing responses to several qu
§ Referent category.
¶ "Regular use" defined as use of a mobile phone on average once per week or

risks, and the odds ratio did not increase for use of mobile The present study

The Lonn article is based on a case-control study of 60 people who had malignant parotid [salivary] gland tumors, 112 benign pleomorphic adenomas, and 681 controls. The risk factor of interest was mobile phone use. If you focus on the main question of interest, upper left section of the table gives you a 2 x 2 table as follows:

| | *E+* | *E-* | Row Margins |
|---|---|---|---|
| *D+* | *25* | *35* | *60* |
| *D-* | *401* | *280* | *681* |
| Col Margs. | *426* | *315* | *741* |

Here *E+* denotes the exposed group, where 'exposure' is to regular use of a mobile phone use, while *E-* denotes people who never or rarely use a mobile phone.

The odds ratio estimate is

$$OR = \frac{25 * 280}{35 * 401} = 0.488,$$

which differs from the finding in the paper (OR = 0.70). The difference is due to the fact that their analysis was adjusted for age, gender, geographic region, and education (using logistic regression, to be discussed later). Note that the proportion of people in the case group who were exposed was 25/60 = 42%, while the proportion of people in the control group *D-*) who were exposed was 401/681 = 59%. There is thus no evidence at all that cases are more likely than controls to have been exposed; in fact the estimated odds ratio goes the other way.

**Fact:** that the odds of exposure given disease are the same as the odds of disease given exposure. This is another reason case-control studies and the odds ratio are of value.

Note that they also provide 95% confidence limits for the true odds ratio: $(0.4,1.3)$. This interval includes the number 1.0, which indicates that the data are consistent with the hypothesis that there is no effect of exposure on the odds of acquiring a parotid gland malignancy.
But how would you compute a 95% confidence interval for the odds ratio?

Here is the methodology:

1. Compute the estimated odds ratio as above. Result: *OR = 0.488*

2. Take the natural logarithm of the estimated odds ratio:
$y = \ln(OR) = \ln(.488) = -.718$

3. Compute the variance of *y = ln(OR).* For a 2 x 2 table with cell entries a, b, c, and d, this turns out to be:

$$\text{var}(\ln(OR)) = 1/a + 1/b + 1/c + 1/d \ ,$$

that is, it is just the sum of the reciprocals of the counts in each cell.

4. Find the standard error of *ln(OR).* This is just the square root of the estimated variance:

$$serr(\ln(OR)) = \sqrt{\text{var}(\ln(OR)} = \sqrt{1/a + 1/b + 1/c + 1/d} \ .$$

5. Compute the 95% confidence interval for the true values of ln(OR):

lower 95% bound for *ln(OR) = ln(OR) – serr(ln(OR)), and*

*upper 95% bound for ln(OR) = ln(OR) + serr(ln(OR)).*

6. Finally, to get things back on the original odds ratio scale, you do the opposite of taking the natural logarithm of the upper and lower bounds in [3]: that is, you apply the inverse function of the natural logarithm – you exponentiate the lower and upper limits by using the exponential function (note that *exp(A) = $e^A$).*

Lower 95% bound for *OR* = exp(lower 95% bound for *ln(OR)*),

Upper 95% bound for OR = exp(upper 95% bound for *ln(OR))*

**Exercise 16:** Compute the 95% confidence interval for the true odds ratio for the parotid gland study data given in the 2 x 2 table above.

The reason you first compute the confidence limits on the log scale and then exponentiate them is that confidence limit estimates usually depend on the fact that certain statistics are well-approximated by the normal distribution function. The odds ratio itself is **not** well-approximated by the normal distribution, but *ln(OR)* is. It is possible to compute an estimated standard error for the odds ratio itself, but the resulting confidence limits are not very accurate.

Note that, in this exercise, you will not get exactly the same answers as in Lonn's table because their computation was adjusted for age, gender, geographic region, and education.

The SAS statistical package, applied to the data in the 2 x 2 table for the parotid gland study, gives:

1. Fisher exact test p-value (two-sided): p = .0093
2. Uncorrected chi-square p-value : p = .0075
3. Corrected chi-square p-value : p = .0112
4. Odds ratio estimate : $\hat{OR} = 0.488$
5. 95% Conf Limits for true OR : *(0.29, 0.85)*.

Note that the both the estimated odds ratio and the confidence limits differ somewhat from what appears in the table of Lonn et al. – this is, again, because the analysis shown here is not 'adjusted' for other covariates (age, gender, geographic region, education).

# C. Logistic Regression

### C.1 The logistic function.

**Assume that you are studying the relationship between a *predictor X* and an outome variable, Y**. The outcome variable *Y* has only two states, 0 and 1 (think: 0 = dead, 1 = alive). The predictor *X* can take on several different values. For example, *Y* may be represent the event of having a in the next year. *X* may be the person's systolic blood pressure at the beginning of the year. The higher *X* is, the more likely it is that he/she will have a stroke. This is expressed by the equation:

$$prob(Y = 1 \mid X) = \frac{1}{1 + \exp(-b_0 - b_1 \cdot X)}.$$

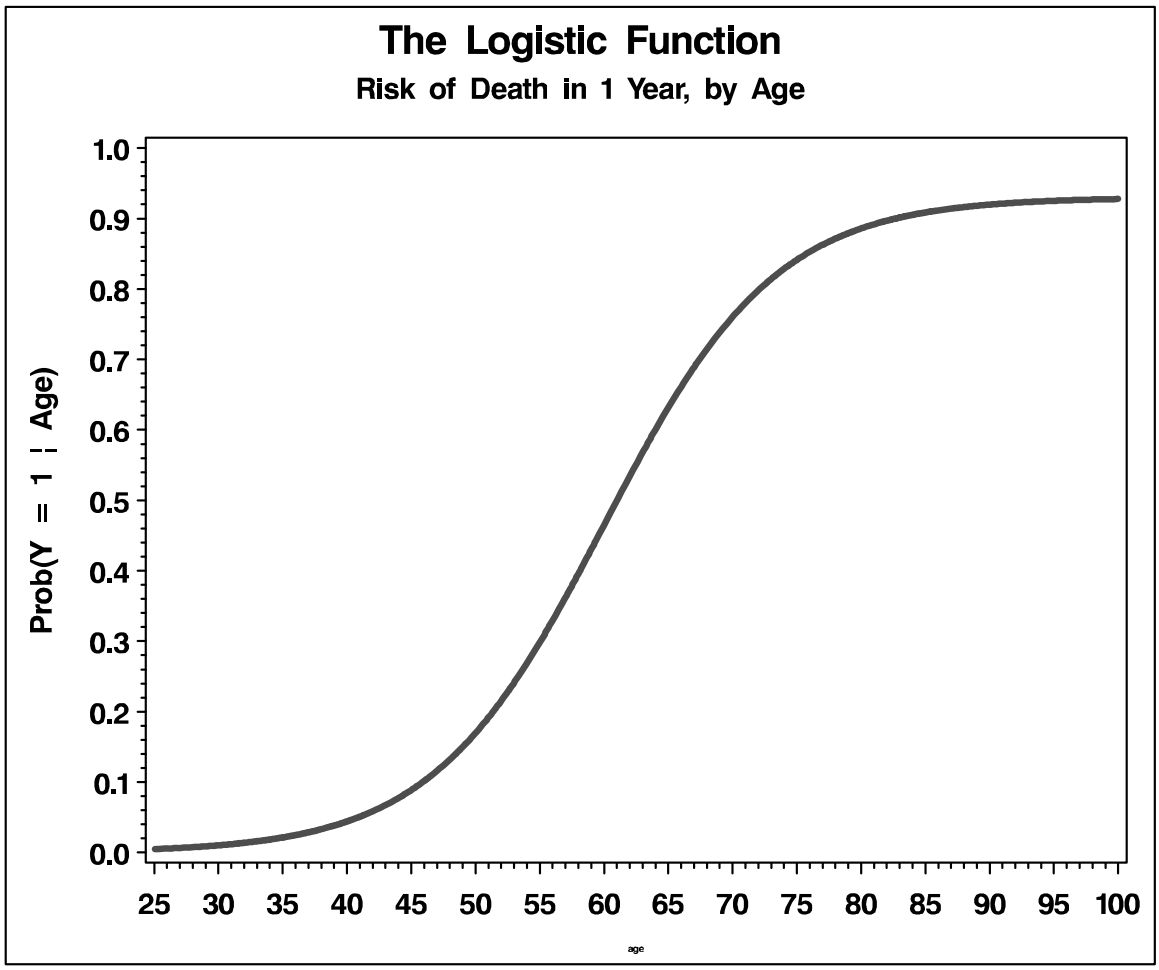
This is the **logistic function** or the **logistic risk function.** There are several things to note about it:

1. $b_0$ and $b_1$ are unknown constants.
2. The function is always > 0. This is true because 1 is nonnegative and exp(anything) is always nonnegative.
3. The function is always < 1. This is true because (a) the first term in the denominator is 1 and the second term is positive. Therefore (b) the denominator is larger than the numerator, so the whole fraction is < 1.
4. If $b_1$ is positive, then when *X* increases, the risk of an event (Y = 1) increases.

The objective with logistic regression is to estimate the unknown coefficients $b_0$ and $b_1$, and to perform tests regarding these parameters. In some special cases,

you can do this with pencil and paper. However in most cases you will need to use computer software.

Assume that *X* is age in years, and Y represents whether the person died during the years following entry (*Y = 1)* or the person survived during the one-year follow-up period (Y = 0). The function could be graphed as follows:




This is the usual 'sigmoidal' shape of the logistic function. It indicates that the risk of death within the next year is very low at age 25, but is very high at age 100.

You can carry out logistic regression in most statistical software packages (SAS, R, Stata, SPSS, SYSTAT, etc.). In general you will be analyzing a file which has the following structure:

| Observation | X | Y |
|---|---|---|
| 1 | 54 | 0 |
| 2 | 86 | 1 |

| | | |
|---|---|---|
| 3 | 23 | 0 |
| … | … | … |

The logistic risk function has a lot of flexibility. It can be used to model risk as a function of several variables. Assume your data file has this structure:

| Obs | Y | Age | SBP | LDL | Cigs |
|---|---|---|---|---|---|
| 1 | 1 | 67 | 138 | 180 | 35 |
| 2 | 0 | 43 | 122 | 105 | 0 |
| 3 | 0 | 34 | 150 | 189 | 20 |
| 4 | 1 | 86 | 103 | 98 | 0 |
| 5 | 0 | 72 | 122 | 144 | 0 |
| … | … | … | …. | …. | … |

Here the outcome variable, Y, indicates whether the person had a heart attack within 5 years of study entry. Age in age in years, SBP is systolic blood pressure, LDL is low-density lipoprotein, and Cigs is how many cigarettes per day the person smokes. A logistic risk function in this case could be:

$$prob(Y = 1 \mid riskfactors) = 1/(1 + \exp(-b_0 - b_1 Age - b_2 SBP - b_3 LDL - b_4 Cigs))$$

.

To obtain estimates of the coefficients using Stata, you would write:

```
. logit Y age sbp ldl cigs
```

The logistic model can also be used to carry out the odds ratio analysis discussed above for 2 x 2 tables. Refer again to the data on parotid gland malignancy:

**Table 16: Parotid Gland Malignancy and Exposure to Mobile Phones**

| | *E+* | *E-* | Row Margins |
|---|---|---|---|
| *D+* | *25* | *35* | *60* |
| *D-* | *401* | *280* | *681* |
| Col Margs. | *426* | *315* | *741* |

The idea here is, represent the exposure status with a variable *E* and the disease status with a variable *D.* The data file would have the following structure:

| Obs | D | E | |
|---|---|---|---|
| 1-280 | 0 | 0 | (this line is repeated 280 times) |
| 281-681 | 0 | 1 | (this line is repeated 401 times) |
| 682-716 | 1 | 0 | (this line is repeated 35 times) |
| 717-741 | 1 | 1 | (this line is repeated 25 times) |

The statistical model here is:

$$prob(D=1|E)=\frac{1}{1+\exp(-b_0-b_1E)}.$$

and in Stata, the logistic analysis would be:

```
. logit  d e
```

It is a good idea when analyzing categorical data in Stata to first examine the file. In this case, the 'tabulate' command is appropriate. If you enter

```
.  tabulate d e
```

In Stata, the follow table will be printed:

**Table 17: Stata output from Tabulate command:**

| d | e = 0 | e = 1 | Total |
|---|---|---|---|
| 0 | 280 | 401 | 681 |
| 1 | 35 | 25 | 60 |
| Total | 315 | 426 | 741 |

Stata output from logistic regression:

**Table 18: Stata output from logit command:**

```
. logit d e

Iteration 0:   log likelihood = -208.32186
Iteration 1:   log likelihood = -205.09666
     …               …              …

Iteration 4:   log likelihood = -205.02284

Logistic regression                             Number of obs =     741
                                                LR chi2(1)    =    6.60
                                                Prob > chi2   = 0.0102
Log likelihood  =  -205.02284                   Pseudo R2     = .0158

------------------------------------------------------------------------------
           d |      Coef   Std. Err.      z    P>|z|     [ 95% Conf Int ]
-------------+----------------------------------------------------------------
           e |  -.6956441    .273197    -2.55   0.011     -1.2311  - .1610
       _cons |  - 2.079442   .1792843  -11.60   0.000     -2.4308  -1.7280
------------------------------------------------------------------------------
```

Here is how this is interpreted. The logistic model was

$$prob(d=1|e)=\frac{1}{1+\exp(-b_0+b_1e)}$$

The estimate of the coefficient $b_1$ is in the first row of the table under 'Coef'. That is $\hat{b}_1=-.6956441$. The estimate of $b_0$ is in the row labeled '_cons', which means, the constant term (also called the intercept), that is, $\hat{b}_0=-2.079442$. In general, the term you are most interested in is $b_1$.

Now, a curious observation: if you exponentiate the term $b_1$, you get

$$\exp(-.6956441)=0.499.$$

This happens to be the same as the estimated odds ratio from the simpler analysis – not a coincidence

**Fact: Relationship between logistic coefficients and odds ratios:**

In the logistic model,

$$prob(Y=1|X_1,X_2,...,X_p)=\frac{1}{1+\exp(-b_0-b_1X_1-b_2X_2-...-b_pX_p)}.$$

the odds ratio corresponding to a 1-unit increase in the variable $X_j$ is

$$OR_j=\exp(b_j).$$

So if, for example, you are analyzing data in which the outcome variable is death, and in which one of the predictors $X_j$ is age: if the coefficient $b_j$ is estimated to be *0.300* then the odds of death for Mr. Smith, who is 1 year older than Mr. Jones, is $\exp(0.300)=1.35$ times larger than the odds of death for Mr. Jones.

Now, what if Mr. Smith is 10 years old than Mr. Jones. What is the odds ratio corresponding to that increase in age?

$$\exp(0.300\cdot 10)=20.09.$$

In general, if the variable $X_j$ is increased by $\Delta X_j$ units, then the corresponding odds ratio is

$$\exp(\hat{b}_j\cdot\Delta X_j).$$

**Exercise 17:** Assume gender is coded as 0 for men and 1 for women. Assume that the coefficient of gender in a logistic regression where the outcome variable is onset of Type 2 diabetes is -0.105. What is the odds ratio of Type 2 diabetes for men versus women?

**Exercise 18:** The Stata output shown above showed the coefficient of the exposure variable, but it did not show the corresponding odds ratio for exposed vs. nonexposed people. Nor did it show 95% confidence limits for the true odds ratio. Can you see a way to compute the 95% confidence interval from what is given in the table above?

**Exercise 19:** Reference the 'High School and Beyond' datafile, which can be accessed from Stata by the following command (if your PC is connected to the Web):

use http://www.ats.ucla.edu/stat/stata/notes/hsb2

This data file has 200 observations and a number of variables which you can see by using the **describe** command in Stata. One of the variables is '**female'.** This variable is coded as 1 if the person is female, 0 if male.

Use logistic regression to explore whether other variables in the dataset 'predict' whether the person is female. The dataset includes scores on math, science, reading and writing.

The **logit** procedure in Stata, as shown above, produces estimates of parameters and their 95% confidence intervals. The **logistic** command will produce estimates of the odds ratios associated with a 1-unit increase in the predictor. The basic syntax is the same – for example,

```
. logistic d e
```

will produce an estimate of the odds ratio for the data on parotid gland cancer and mobile phone devices show in Table 16.

**Exercise 20:** Use **logistic** Stata on the hsby2 file to find estimates of the odds ratio of being female corresponding to a 10-point increase in the math test score. Check that this agrees with the estimates you can derived from the **logit** procedure.

### C.2 Testing for Interaction Using Logistic Regression

Suppose age and gender are important predictors of coronary heart disease. The logistic model in this case would be:

$$prob(CHD \mid age, gender) = \frac{1}{1+\exp(-b_0 - b_1 * age - b_2 * gender)} \quad .$$

It may be that the effect of age is different for men that it is for women. This means that there is an **interaction** between age and gender. The interaction is often represented by adding another term to the logistic model, on the right side of the equation. That term is often just the product of age and gender:

$$prob(CHD \mid age, gender) = \frac{1}{1+\exp(-c_0 - c_1 * age - c_2 * gender - c_3 * age * gender)}$$

.

Note that the coefficients $c_0, c_1, c_2$ in this model are not the same, and will not have the same estimates, as the coefficients $b_0, b_1, b_2$ in the first model.

To compute logistic coefficients for the second model, you must first create a new variable, as follows:

```
. generate agegender = age * gender
```

Then you would use either the **logit** or **logistic** command:

```
logit chd age gender agegender.
```

Evaluating the significance of an interaction term is not so easy. The **right** way to do it is as follows:

Model 1: logit chd age gender
Model 2: logit chd age gender agegender

Then find the value of log(likelihood) for each model. This is in the Stata printout. The value for Model 2 will be larger than that for Model 1. Both of the

values will very likely be negative numbers, but the Model 2 log likelihood will be larger.

Then compute -2 * log(likelihood) for each model.

Then subtract the value of -2 * log(likelihood) for Model 2 from the value for Model 1:

$$diff = -2 \cdot \log(likehood)_{Model\,1} - (-2 \cdot \log(likelihood)_{Model\,2}).$$

This should always be a positive number.

To find a p-value for the interaction, you have to compare *diff* to the chi-square distribution with 1 degree of freedom.

**Exercise 21:** In the hsby2 dataset, with **female** as the outcome variable, test for whether there is an interaction between the **writing** score and the **math** score. What would it mean if there were such an interaction?

**References:**

There are many, many references on categorical data analysis. Two textbooks that I would recommend are:

1. Joseph L. Fleiss, *Statistical Methods for Rates and Proportions, 3rd Edition (2003),* John Wiley & Sons, New York.

2. Alan Agresti, *Categorical Data Analysis, 2nd Edition (2002),* Wiley-Interscience, NY.

Another extremely useful reference is the UCLA statistics website:

http://www.ats.ucla.edu/stat/

This has a huge collection of resources, examples, tutorials, and data files. The following page from that site is especially useful:

http://www.ats.ucla.edu/stat/stata/whatstat.htm#bitest

Another reference on both categorical data and Stata is:

3. J. Scott Long and Jeremy Freese, *Regression Models for Categorical Dependent Variables Using Stata, 2nd Edition.(2006).* Stata Press. This is a paperback, costs about $58 US.