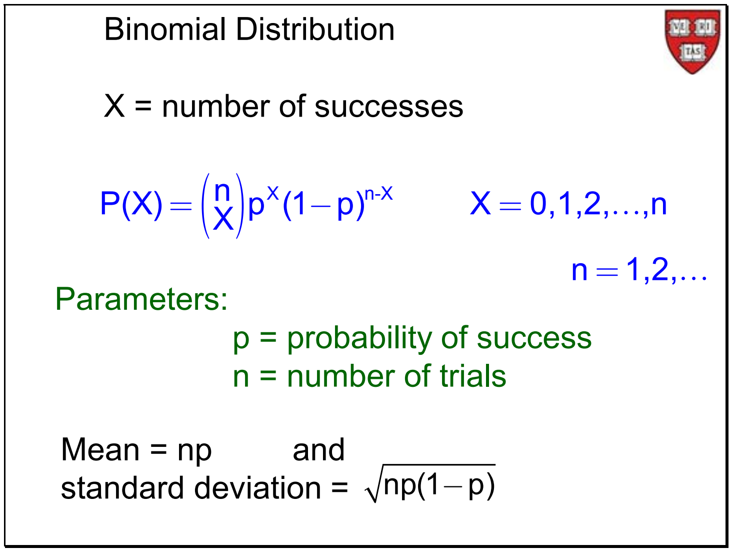
Marcello Pagano

[JOTTER 7 CONTINGENCY TABLES]

Inference about the binomial parameter p. Sample size calculations. Odds ratios. Berkson's fallacy. Yule Effect.

We now have under our belts how to test and estimate the mean of a single population, of two populations, and more than two populations, when we have normal data. In other words, when dealing with continuous measures. Now let us switch attention to the situation when we have count data. Let us start with the binomial situation. Now what we have already seen about the binomial model is that it is useful for modeling dichotomous data: yes, no; male, female; alive, dead, et cetera.



Binomial Distribution

X = number of successes

$P(X) = \binom{n}{X} p^X (1-p)^{n-X}$ $X = 0,1,2,\ldots,n$

$n = 1,2,\ldots$

Parameters:

p = probability of success

n = number of trials

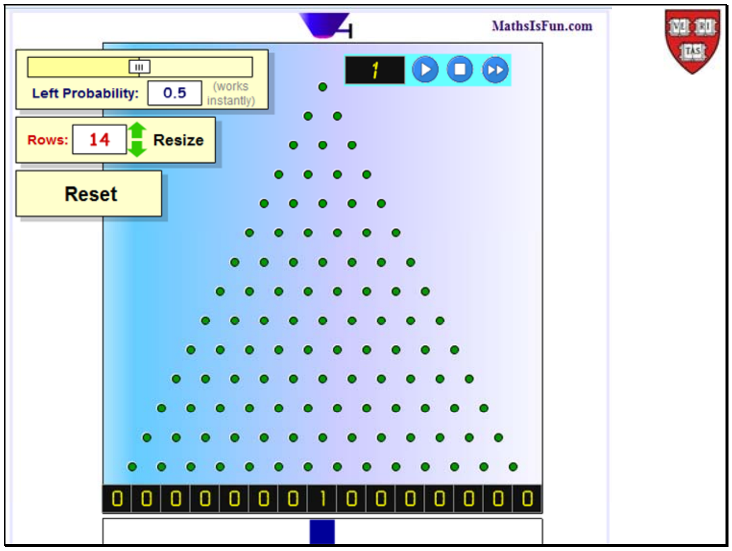
Mean = np and

standard deviation = $\sqrt{np(1-p)}$

Now our challenge is how to estimate p in the binomial model. We usually assume that n is known, and now we are going to estimate p, the probability of a success at every trial, when we have n independent trials.

For example, we might think of a village where p is the prevalence of vaccinated children, and we take a random sample of 14 children. We can model the situation of sampling from such a village by using a binomial model. So, the mean number of children we find vaccinated in such villages, as we visit village to village, is np, with a standard deviation of $\sqrt{np(1-p)}$, and we put n=14 in these formulae.

In the usual spirit of hypothesis testing, we approach inference about p by saying, if this is the truth, what do I expect to see? We can go to the Quincunx to see how the number of vaccinated children we actually do find in a village, varies from sample to sample.



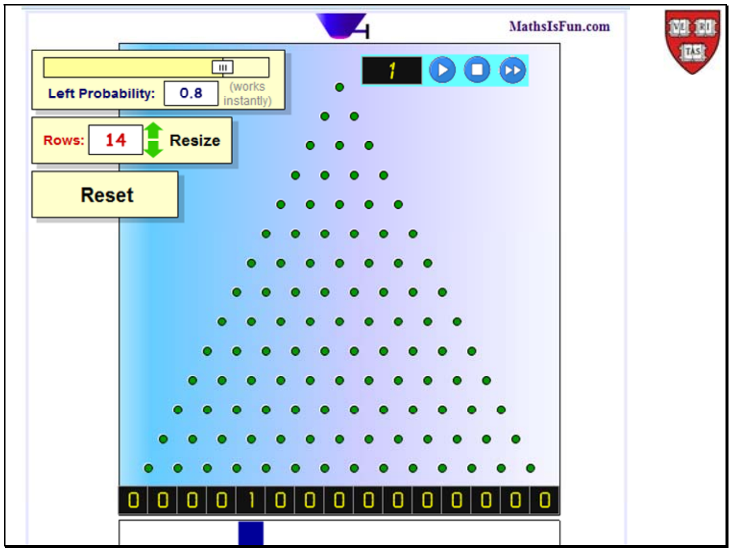
Left Probability: 0.5

Rows: 14

1

0 0 0 0 0 0 0 1 0 0 0 0 0 0 0

So here is the quincunx. I let one ball fall down with 14 rows—to simulate our asking 14 children in the village. I set the left probability at 0.5—to simulate a village with 50% vaccination coverage. We expect the observation to come down the center since at each peg it is equally likely to go left as it is to go right. It turns out, that it fell into the center bin! If the ball falls into the leftmost bin, that is finding zero children vaccinated. In the next bin to the right, is like finding one child vaccinated, and so on, all the way to the rightmost bin which represents our finding all 14 kids vaccinated.



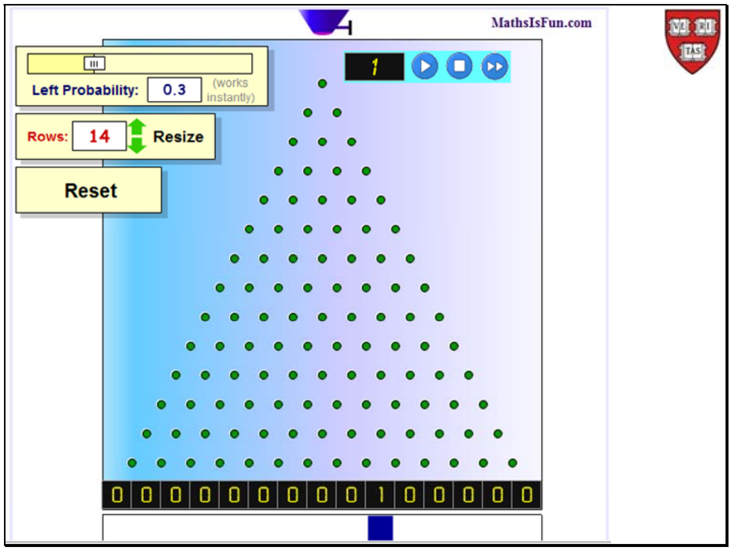
Left Probability: 0.8

Rows: 14

1

0 0 0 0 1 0 0 0 0 0 0 0 0 0 0

If I now change the “Left Probability” to 0.8 and rerun the simulation, the ball should fall left of center in our view of the Quincunx; as, indeed, it does.



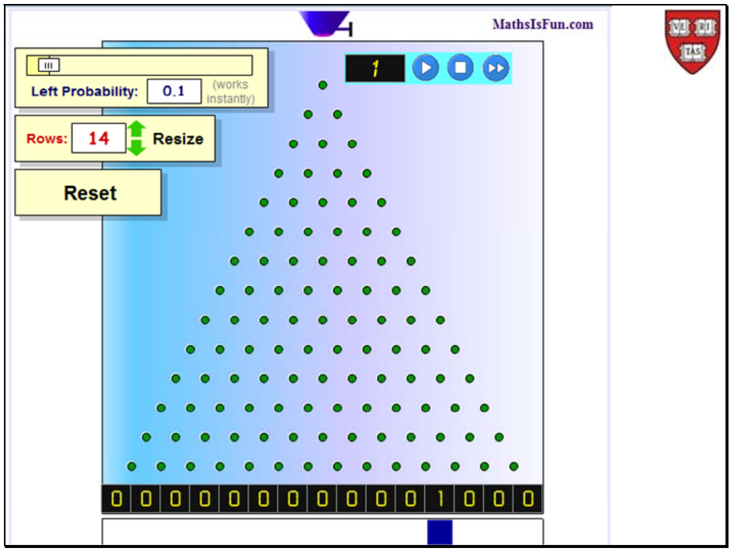
Left Probability: 0.3

Rows: 14

1

0 0 0 0 0 0 0 0 0 1 0 0 0 0 0

Next, I changed the left probability to 0.3. This should result in the ball ending up right of center, as indeed it does. (The left probability is the probability of being not vaccinated. Sorry about that, but I did not design the Quincunx!)



Left Probability: 0.1

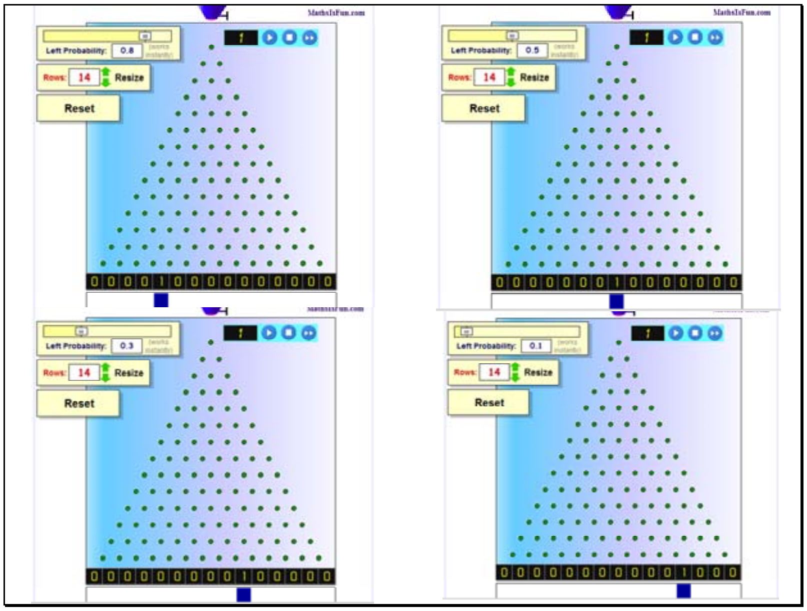
Rows: 14

1

0 0 0 0 0 0 0 0 0 0 0 1 0 0 0

Lastly, when we set the left probability to 0.1, then the ball should end up even more towards the right end, as it does.

So here is the challenge: Suppose we know where the ball ended up—how many kids did we find vaccinated in the village—determine from that information what the value of the “Left Probability” is—or what the village (not the sample, we know that) vaccination coverage, or prevalence, is.



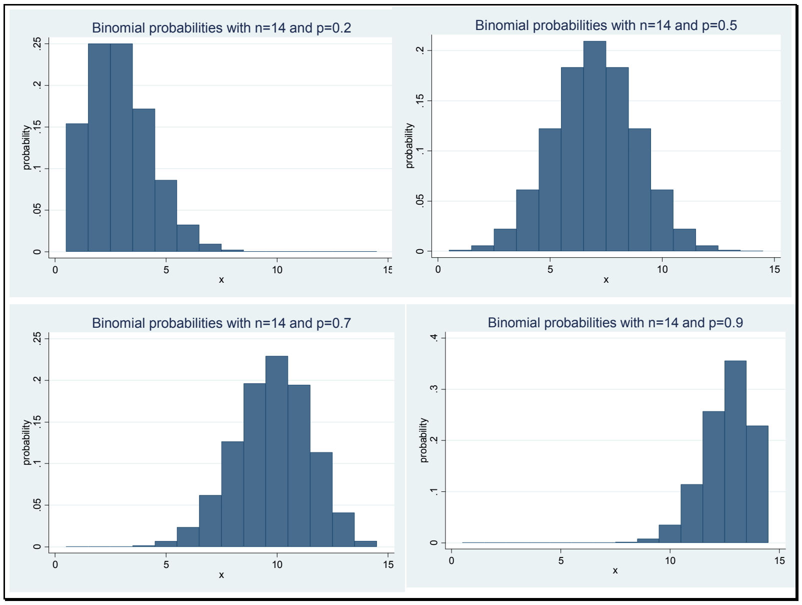
Left Probability: 0.8
Rows: 14

Left Probability: 0.5
Rows: 14

Left Probability: 0.3
Rows: 14

Left Probability: 0.1
Rows: 14

So, for example, here are the four snapshots of the Quincunx that we just looked at. It is not going to land in exactly the same spot every time, as we have seen often enough.



Binomial probabilities with n=14 and p=0.2

Binomial probabilities with n=14 and p=0.5

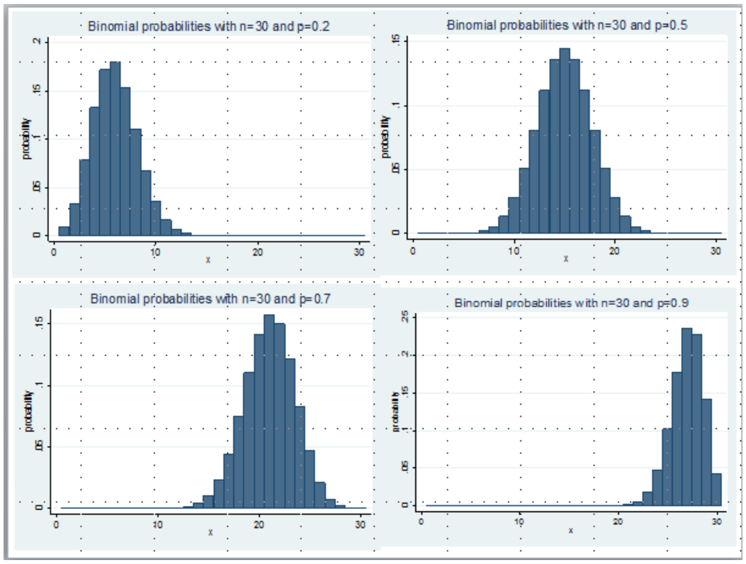
Binomial probabilities with n=14 and p=0.7

Binomial probabilities with n=14 and p=0.9

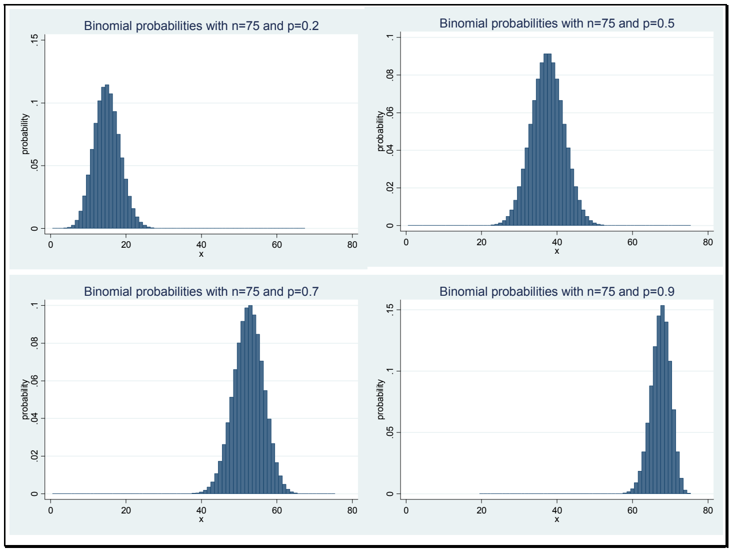
If we repeatedly run the Quincunx, we get pictures like these to reflect the distribution of the balls in the bins. These are actual theoretical binomial distributions generated by Stata, but we know that if we run the Quincunx ad infinitum, this is what we would see.

These distributions peak at points that travel from left to right and their location monotonically depends on the size of p, the probability of a success at a single trial; the peak travels from left (when p is small) to right (when p is large). Note that the p’s in these pictures correspond to (1- “Left Probability”) in the Quincunx pictures.

These peaks are important as they direct us to where we expect most of the balls to land. They mostly land around the peak because that is where most of the probability mass is. This is a fancy way of saying that when we run the Quincunx, the height of the bin tells us how popular that bin is, and the most popular bin is the one that corresponds to the peak, np.

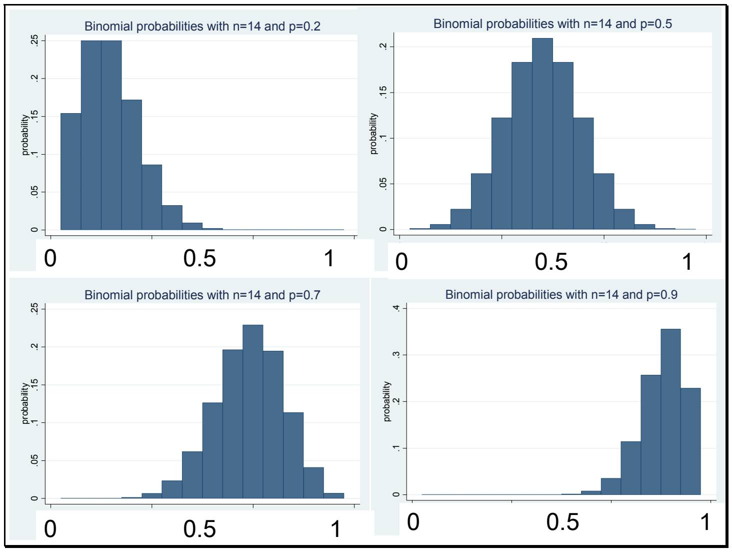


If we increase the number n—the number of children we check in the village—from 14 to 30, this is what we see: a tighter agglomeration of the probability mass around the peak, or mean. And so, just like with the flab rats, it is much easier to distinguish between any two of these four distributions than it was with n=14.

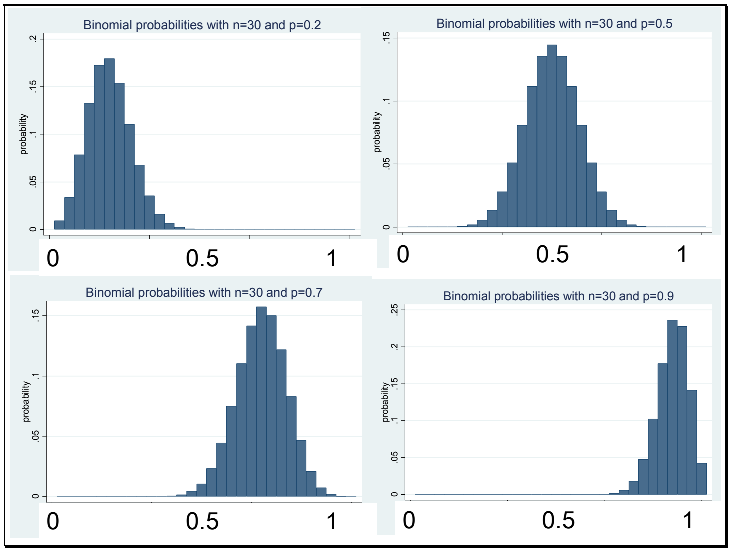


When we make n even larger, to 75, say, then it is even easier to distinguish these four cases since they are even more separated—little overlap.

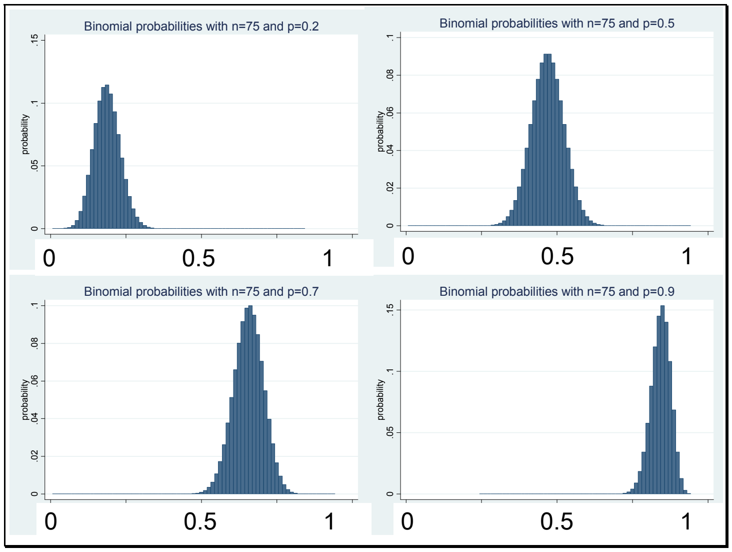
So it becomes easier to distinguish between these p's as my sample size is bigger. You notice though that the scales on the plots change as n changes, so these comparisons are not quite fair. So let us rescale the plots to reflect not our counting the number of successes, but rather looking at the proportion of successes—just divide the total number of successes by n, the sample size. In that way our horizontal axis should now go from zero to one.



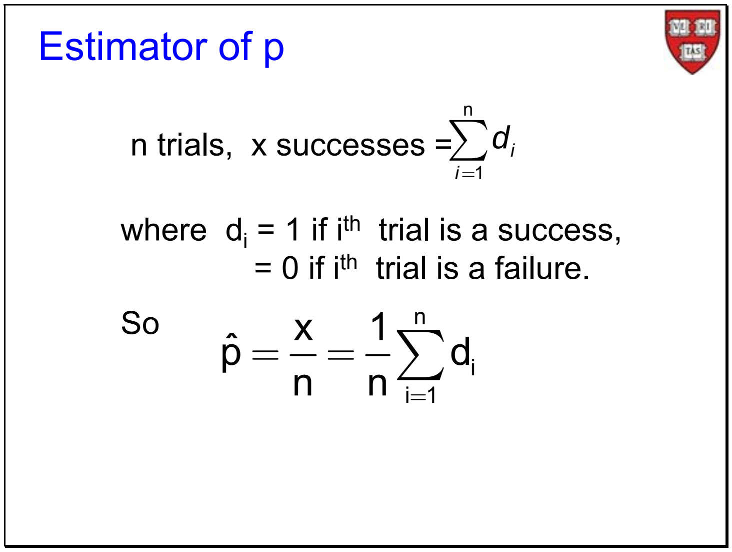
Here are the plots of the distribution of the proportion of successes when n=14.



Here are the same plots when n=30.



And here are the same plots when n=75.



Estimator of p

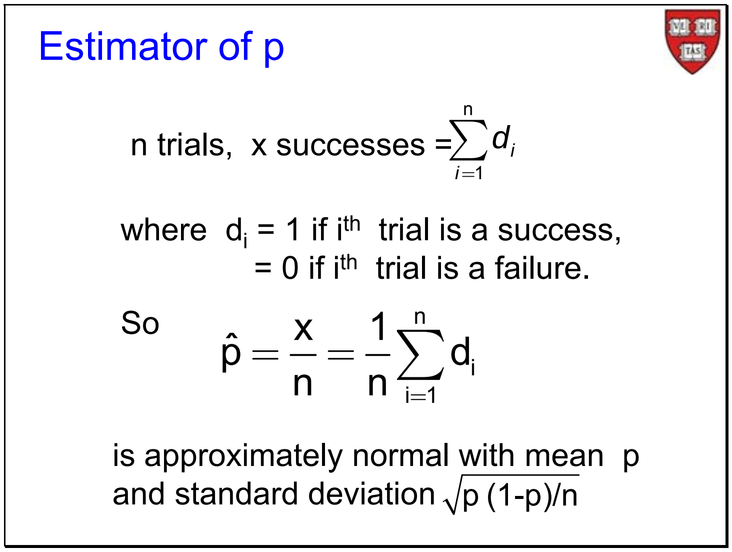
n trials, x successes $=\sum_{i=1}^{n} d_i$

where d_i = 1 if i^{th} trial is a success,
= 0 if i^{th} trial is a failure.

So $$\hat{p} = \frac{x}{n} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

So what we have drawn is the sampling distribution of the proportion of successes, $\hat{p}$, ass opposed to our earlier plots of the total number of successes. Since this is the number of successes (which peaked at np) divided by n, the number of trials, what that means is that the sampling distribution of $\hat{p}$ will peak at p, the quantity we are trying to estimate.

We should have expected this because if we define each trial to be a d that is equal to zero or one depending on whether we had a failure or a success, then $\hat{p}$ is simply the sample mean of these d's, and the central limit theorem tells us that the mean of the sampling distribution of the $\hat{p}$ is p. So on average, across the villages, we get the population value, p. So we have an *unbiased* estimator of p.



Estimator of p

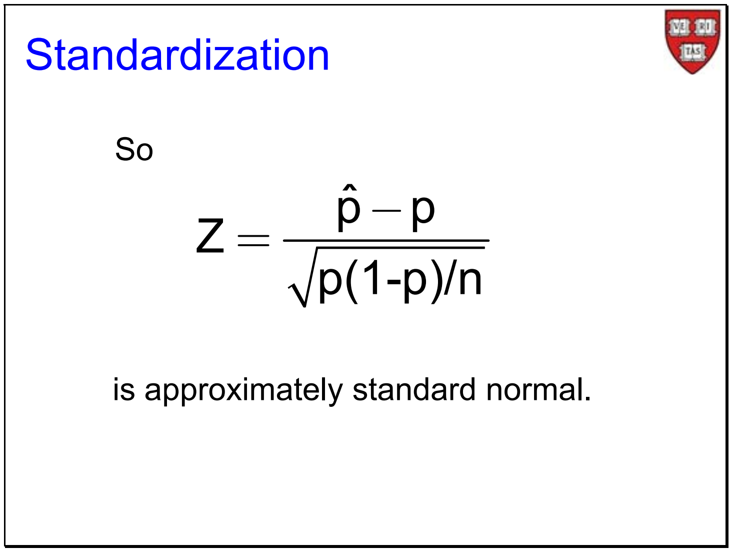
n trials, x successes $=\sum_{i=1}^{n} d_i$

where d_i = 1 if i^{th} trial is a success,
= 0 if i^{th} trial is a failure.

So
$$\hat{p} = \frac{x}{n} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

is approximately normal with mean p
and standard deviation $\sqrt{p\,(1\text{-}p)/n}$

Viewing our estimator as the sample mean, albeit of zero-one variables, allows us to appeal to the central limit theorem to say that for large sample sizes, n, this sample mean is approximately normally distributed with mean p and standard deviation (standard error) $\sqrt{p(1-p)/n}$.



Standardization

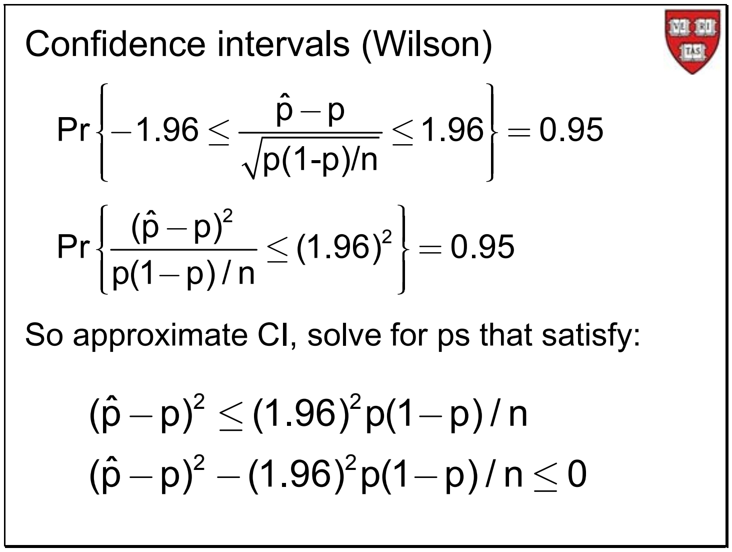
So
$$Z = \frac{\hat{p} - p}{\sqrt{p(1\text{-}p)/n}}$$

is approximately standard normal.

In order to create confidence intervals or perform hypothesis testing with respect to p, we can collect all this information and create our standardized Z, as above.

What makes this standardization a little more complex than what we have become accustomed to, in our prior inference, is that the parameter of interest, p, appears both in the numerator of Z and in its denominator, whereas before we had μ, the parameter of interest, in the numerator and a separate σ in the denominator.

One solution to too many appearances of, and this is the so-called Wald solution, is to argue that we are interested in the p inasmuch as it is the mean, so estimate the standard deviation (as we did before when we replaced the unknown σ by the sample standard deviation, s) by replacing p where it appears in the standard deviation (denominator) by $\hat{p}$. This is equivalent to using the sample standard deviation to estimate the population standard deviation, just as we did before in going from the Z to the t with continuous data. This solution is not to be recommended as it can lead to problems when p is close to either boundary (zero or one).



Confidence intervals (Wilson)

$$\Pr\left\{-1.96 \le \frac{\hat{p}-p}{\sqrt{p(1\text{-}p)/n}} \le 1.96\right\} = 0.95$$

$$\Pr\left\{\frac{(\hat{p}-p)^2}{p(1-p)/n} \le (1.96)^2\right\} = 0.95$$

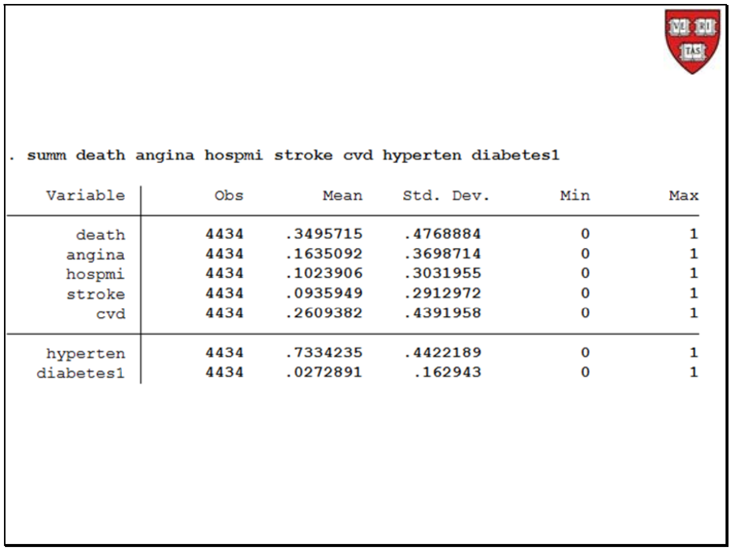
So approximate CI, solve for ps that satisfy:

$$(\hat{p}-p)^2 \le (1.96)^2 p(1-p)/n$$

$$(\hat{p}-p)^2 - (1.96)^2 p(1-p)/n \le 0$$

Another approach that also utilizes the DeMoivre result is the Wilson method. This starts by simply squaring the standardized Z. This has squaring has the advantage of eliminating both the pesky square-root in the denominator and the plus, minus interval that Z needs to fall into for the confidence interval to be satisfied. We recognize that the resulting single inequality defines a parabola in p that can subsequently be solved.

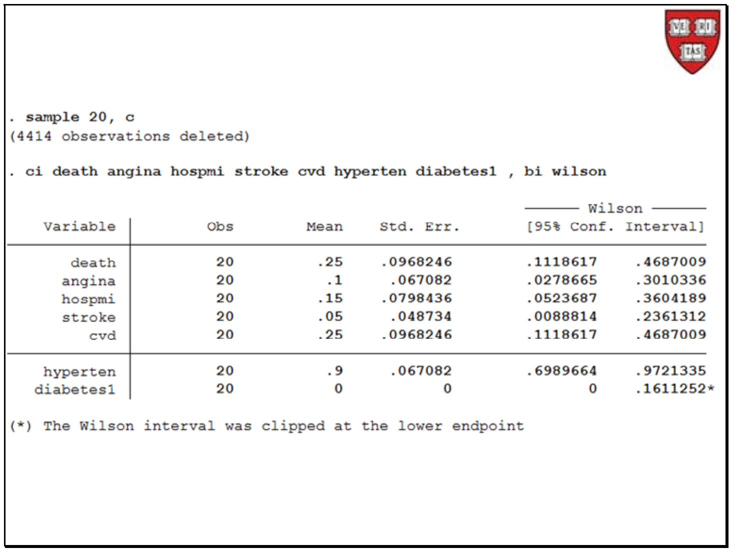
Let us apply this thinking to our Framingham Heart Study data.



```
. summ death angina hospmi stroke cvd hyperten diabetes1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      4434    .3495715    .4768884          0          1
      angina |      4434    .1635092    .3698714          0          1
      hospmi |      4434    .1023906    .3031955          0          1
      stroke |      4434    .0935949    .2912972          0          1
         cvd |      4434    .2609382    .4391958          0          1
-------------+--------------------------------------------------------
    hyperten |      4434    .7334235    .4422189          0          1
   diabetes1 |      4434    .0272891     .162943          0          1
```

Here are seven dichotomous variables with their respective proportions, p’s, in the population of 4,434 people in our data set. These are the parameters we would like to infer when we take a sample from this population. So, for example, we have that deaths are approximately 35% of the population. Angina is about 16%, and so on down to hypertension at 73%, and our old friend diabetes at the first visit is roughly 2.7%.

Now take a sample of size 20 from this population.

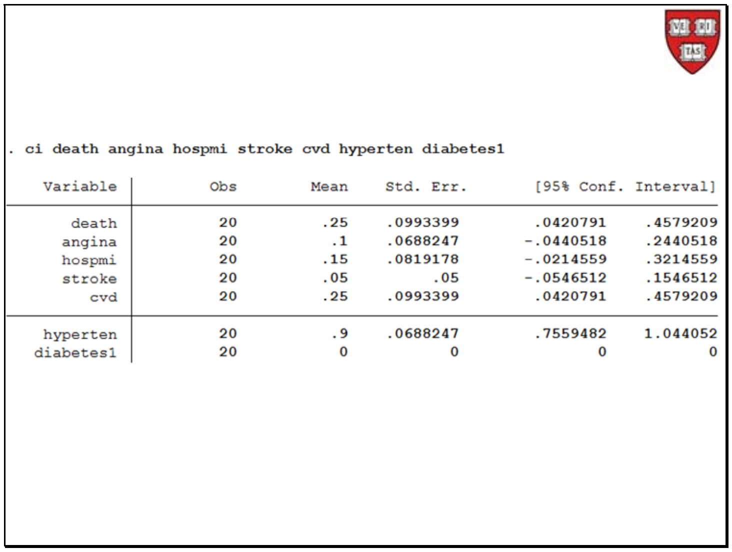


```
. sample 20, c
(4414 observations deleted)

. ci death angina hospmi stroke cvd hyperten diabetes1 , bi wilson

                                                         ------ Wilson ------
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
       death |         20         .25    .0968246        .1118617    .4687009
      angina |         20          .1     .067082        .0278665    .3010336
      hospmi |         20         .15    .0798436        .0523687    .3604189
      stroke |         20         .05     .048734        .0088814    .2361312
         cvd |         20         .25    .0968246        .1118617    .4687009
-------------+---------------------------------------------------------------
    hyperten |         20          .9     .067082        .6989664    .9721335
   diabetes1 |         20           0           0               0    .1611252*

(*) The Wilson interval was clipped at the lower endpoint
```

Now that we have the sample, the *ci* command in Stata will give us the summaries, above, and the confidence intervals, the Wilson ones in our case because we use the Wilson option. (Do not forget the *bi* option to tell Stata we have binomial data.)

Let us look at this output from Stata. We can look at each line, one by one, starting with the first one, death. First we see that $\hat{p}$ is 0.25 with a standard error of approximately 0.1. The associated (Wilson) 95% confidence interval for p is (0.11, 0.47). Remember the true value of p (which we know in our exercise because we are acting as if we know the population) is 0.35. So the confidence interval did indeed cover the population value in this case.

All the others lend themselves to similar interpretations until we get to diabetes1. We have to be careful when the mean is very small, close to zero, or very large, close to one. In those cases we sometimes run into problems as we do here. Here we have that the proportion of diabetics is small, 0.027, and as a result none appeared in our sample of 20. So $\hat{p}$ is zero and the Wilson estimator has problems. Thus the clipping message, and in this case the coverage probability may not be 95%.

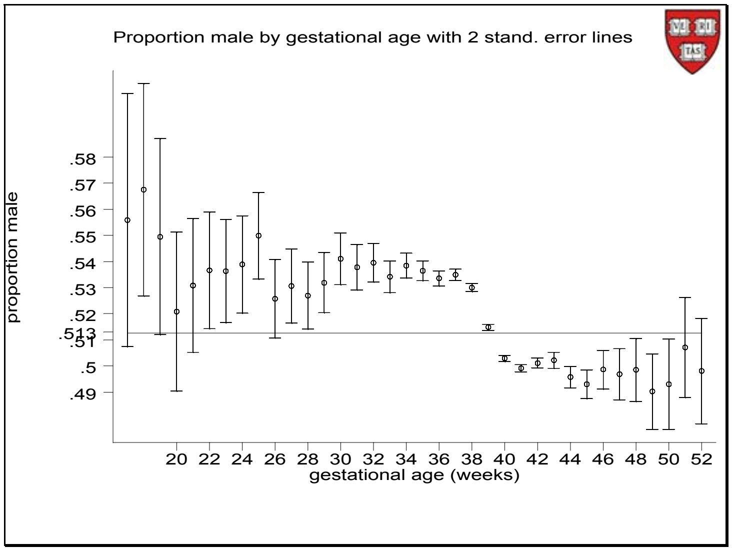


```
. ci death angina hospmi stroke cvd hyperten diabetes1

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
       death |         20         .25     .0993399        .0420791    .4579209
      angina |         20          .1     .0688247       -.0440518    .2440518
      hospmi |         20         .15     .0819178       -.0214559    .3214559
      stroke |         20         .05          .05       -.0546512    .1546512
         cvd |         20         .25     .0993399        .0420791    .4579209
-------------+---------------------------------------------------------------
    hyperten |         20          .9     .0688247        .7559482    1.044052
   diabetes1 |         20           0            0               0           0
```

In the *ci* command, had we not asked for the Wilson confidence interval what we would have gotten by default is the misleadingly called “exact” answer. First, we need look in the help file to see what this confidence interval really is since it is not properly labeled—we are just told that it is a/the(?) “95% Conf. Interval”.

Presumably it is called exact because it uses the exact, or model binomial sampling distribution of $\hat{p}$, as opposed to the DeMoivre normal approximation. In reality all these confidence intervals are approximate. Where this “exact” method does its approximating is in the confidence level—it does not give you a 95% confidence interval. What it does produce is an interval that has *at least* 95% confidence attached to it, but the true value may be much larger. It gives the exact answer to the wrong question.

Ordinarily, there is not much difference between these confidence intervals, especially for large sample sizes. If you need to trust one, go with the Wilson unless your n is tiny, in which case the confidence interval is not very informative, anyway.



Here is an example with huge n where I used the Wald approximation. Here we report on roughly 4 million singleton births in a year in the US.

What the circles refer to are the proportion males as a function of gestational age. The solid line at 0.513 is the overall average, ignoring gestational age. The plus/minus bars are capped off at plus or minus two standard errors. This is typically the graph one gets from year to year. It does not vary much.

The first thing we notice about this pattern is that when the interval is wide (small), then we have a small (large) n, since each $\hat{p}$ is roughly the same size. So we see that most of us are born roughly in the 36- to 42-week interval.

Secondly, these numbers are based on birth certificates, which are largely considered administrative documents, so I do not know how much trust to have at either end of the gestation scale. So let us concentrate mostly in the 22-week to 44-week window.

What we are seeing in this window, and even beyond, is a pattern that favors males in the early gestational periods, but that switches, rather smoothly, to favoring females at the later gestational ages. Why this pattern, I do not know.

Hypothesis Testing

$H_0 : p = 0.082$

$$z = \frac{\hat{p} - p}{\sqrt{p(1\text{-}p)/n}}$$

$$= \frac{0.115 - 0.082}{\sqrt{0.082(1 - 0.082)/52}}$$

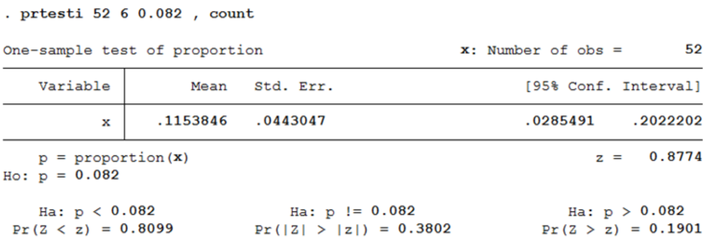
$$= 0.87$$

$$\text{p-value} = 0.384$$

Let us now investigate tests of hypotheses about the parameter p in the binomial. Once again, there are two approaches one can take. Let us first look at the central limit theorem approach, namely the normal approximation. Here the fact that p appears both in the numerator and in the denominator of the standardized Z is not an issue.

Consider a sample of 52 individuals, where 6 survived to five years post diagnosis of cancer, and we want to test the hypothesis that the survival proportion in the population is 0.082. We can set up our standardized Z and calculate it to be 0.87. Check with Stata to find out that the two-sided p-value is 0.384. So we do not reject the null-hypothesis.

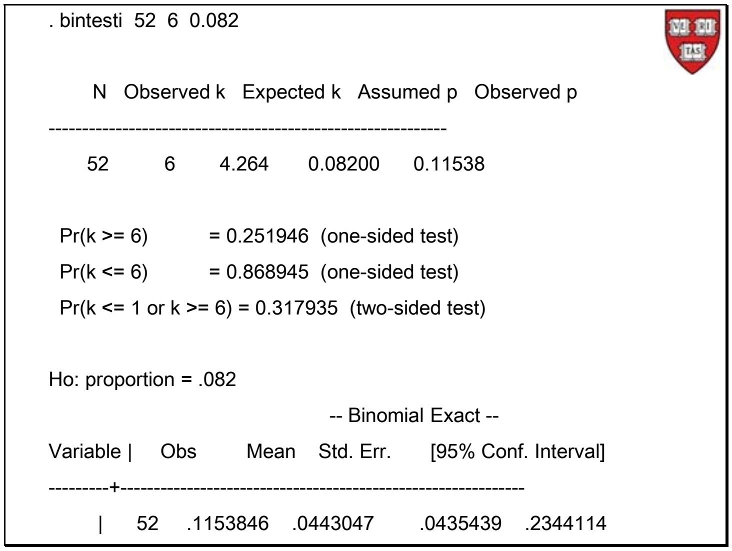
Stata Output



```
. prtesti 52 6 0.082 , count

One-sample test of proportion                   x: Number of obs =       52
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1153846   .0443047                      .0285491    .2022202
------------------------------------------------------------------------------
    p = proportion(x)                                             z =   0.8774
Ho: p = 0.082

    Ha: p < 0.082               Ha: p != 0.082                Ha: p > 0.082
 Pr(Z < z) = 0.8099         Pr(|Z| > |z|) = 0.3802          Pr(Z > z) = 0.1901
```

Rather than do any of these calculations ourselves, we can get Stata to do them for us if we use the command *prtesti,* as above.

The output from Stata looks very much like the output from the t-test. So that is using the normal approximation, which for large samples such as this is perfectly fine.

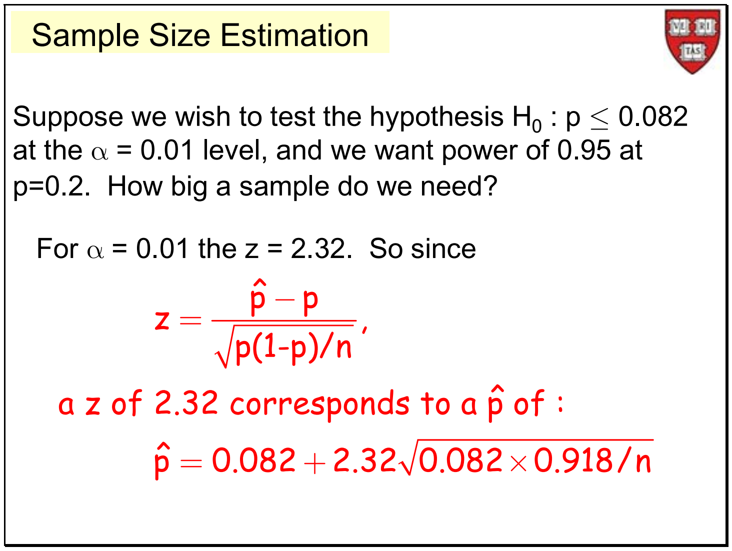


```
. bintesti  52  6  0.082

      N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
     52        6       4.264      0.08200     0.11538

 Pr(k >= 6)          = 0.251946  (one-sided test)
 Pr(k <= 6)          = 0.868945  (one-sided test)
 Pr(k <= 1 or k >= 6) = 0.317935  (two-sided test)

Ho: proportion = .082
                                        -- Binomial Exact --
Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+-------------------------------------------------------------
       |     52     .1153846    .0443047       .0435439    .2344114
```

The other approach is to argue that we have a binomial model, so why not use the calculations appropriate for a binomial and not rely on the large sample approximations afforded us by DeMoivre's central limit theorem. You may invoke this approach by using the Stata command, *bintest*. The results of a call to the "immediate" version of it, *bintesti*, is displayed above. We see that the results here look very much what the large sample approximation showed us.

This approach is fine if you are doing a one-sided test. It is slightly controversial if you are doing a two-sided test since the sampling distribution of the test statistics is not symmetric and is discrete, so it is unclear how what mass gets lumped into the rejection region.

So there are the options for you to do hypothesis testing.



Sample Size Estimation

Suppose we wish to test the hypothesis $H_0 : p \leq 0.082$ at the $\alpha = 0.01$ level, and we want power of 0.95 at p=0.2. How big a sample do we need?

For $\alpha = 0.01$ the z = 2.32. So since

$$z = \frac{\hat{p} - p}{\sqrt{p(1\text{-}p)/n}},$$

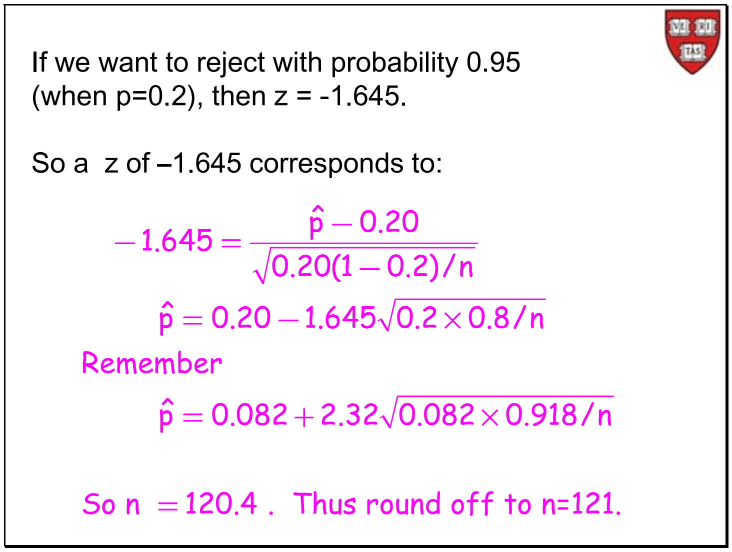
a z of 2.32 corresponds to a $\hat{p}$ of :

$$\hat{p} = 0.082 + 2.32\sqrt{0.082 \times 0.918/n}$$

Just as with the continuous variables when we calculate the sample size required to achieve a particular power, so too with the discrete variables, we can perform similar calculations. So, for example, if we are testing a hypothesis that p is less than or equal to 0.082—once again, let us just investigate the one-sided test—and we want to test this at the 0.01 level. And suppose we want a power of 95% or 0.95, when p equals 0.2. So if p=0.2 we want to be 95% sure we reject the null hypothesis that p=0.082.

Just like the continuous variable, we need these four quantities: what the null hypothesis is; at what level we wish to test that hypothesis; at what point of the alternative to calculate the power; and, what power do we wish to have. So since we chose the 0.01 level, that means we would look at a Z of 2.32. So let us look at our Z, and we want to have it equal to 2.32. This thus provides us a relationship between $\hat{p}$ and n.

Now this sort of calculation is a little bit complex, and before you get too concerned, rest assured that we appeal to Stata to do the hard work for us. This is here for those who, like me, enjoy these details!



If we want to reject with probability 0.95 (when p=0.2), then z = -1.645.

So a z of –1.645 corresponds to:

$$-1.645 = \frac{\hat{p} - 0.20}{\sqrt{0.20(1 - 0.2)/n}}$$

$$\hat{p} = 0.20 - 1.645\sqrt{0.2 \times 0.8/n}$$

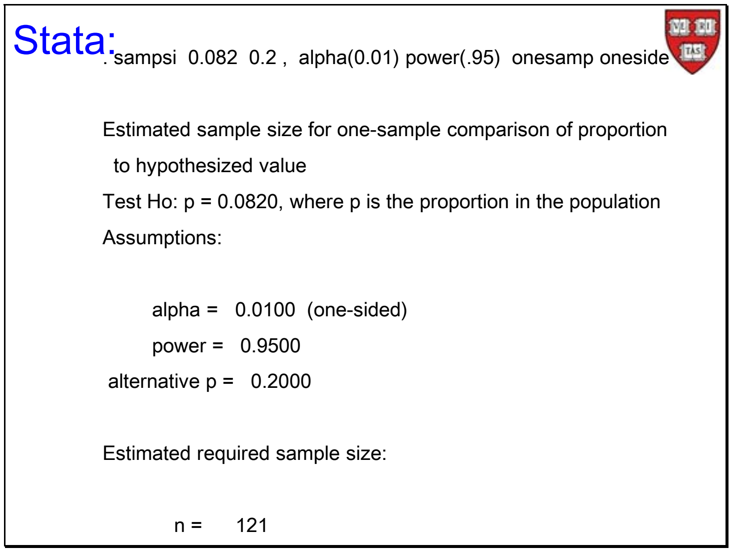
Remember

$$\hat{p} = 0.082 + 2.32\sqrt{0.082 \times 0.918/n}$$

So $n = 120.4$. Thus round off to n=121.

Now turn attention to the power. When p is equal to 0.2, we want to have a power of 0.95. That means that Z must equal -1.645 when p=0.2. That provides us a second relationship between $\hat{p}$ and n.

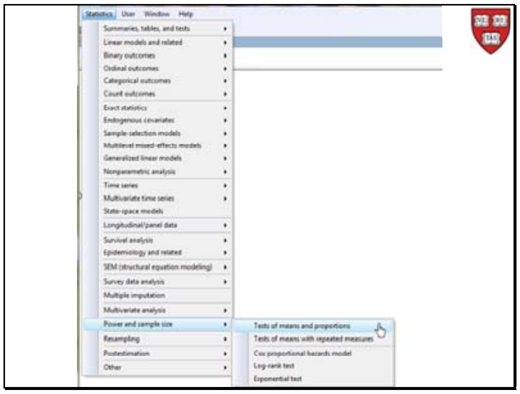
Solving these two equations relating $\hat{p}$ and n we get that n=120.4. Rounding off gives us that an n of 121 will satisfy the alpha and power requirements we require.

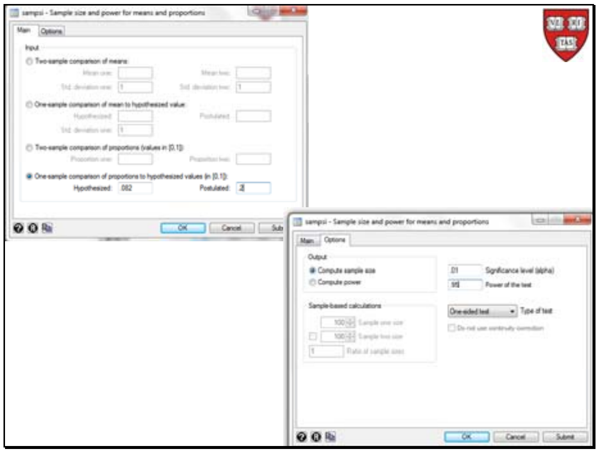


Stata:

```
. sampsi  0.082  0.2 ,  alpha(0.01) power(.95)  onesamp oneside

Estimated sample size for one-sample comparison of proportion
  to hypothesized value
Test Ho: p = 0.0820, where p is the proportion in the population
Assumptions:

         alpha =   0.0100  (one-sided)
         power =   0.9500
 alternative p =   0.2000

Estimated required sample size:

             n =      121
```

The Stata command that will do all these calculations for us is *sampsi*. You can type it in, or you can use the pull down menus:




sampsi - Sample size and power for means and proportions
Main
Options
Input
○ Two-sample comparison of means:
Mean one:
Mean two:
○ One-sample comparison of mean to hypothesized value
○ Two-sample comparison of proportions (values in [0,1]):
Proportion one:
Proportion two:
OK
Cancel
Submit
Output
○ Compute power
Sample-based calculations
Sample one size
Sample two size
Ratio of sample sizes
Do not use continuity correction

Std. deviation one: 1
Std. deviation two: 1
● One-sample comparison of proportions to hypothesized values (in [0,1]):
Hypothesized: .082
Postulated: .2

● Compute sample size
.01 Significance level (alpha)
.95 Power of the test
One-sided test Type of test

First click on Statistics, then on Power and sample size, and then on Tests of means and proportions. Go down to the One-sample comparison of proportions, and enter the hypothesized 0.082. The “postulated” slot is where you place the 0.2, namely the value of p where you wish to calculate the power.

Then you click on the tab labeled Options. This next menu is where you fill in your alpha level, the power you wish, and whether you want to use a one-sided or two-sided test. Then clicking Submit will get Stata to calculate the sample size for you. Alternatively you can clisk on

“Calculate the power” in the top left-hand corner and give Stata a sample size and it will calculate the power for you.

Two-sample situation

Now let us look at the two-sample p. What do I mean by that? What this means is we have two populations, with population I having prevalence p_1, and population II having prevalence p_2.

We take a sample from each of these two populations and on the basis of these samples, make inference about the relative sizes of p_1 and p_2. The primary hypothesis of interest is that these two are equal, but others can be entertained.

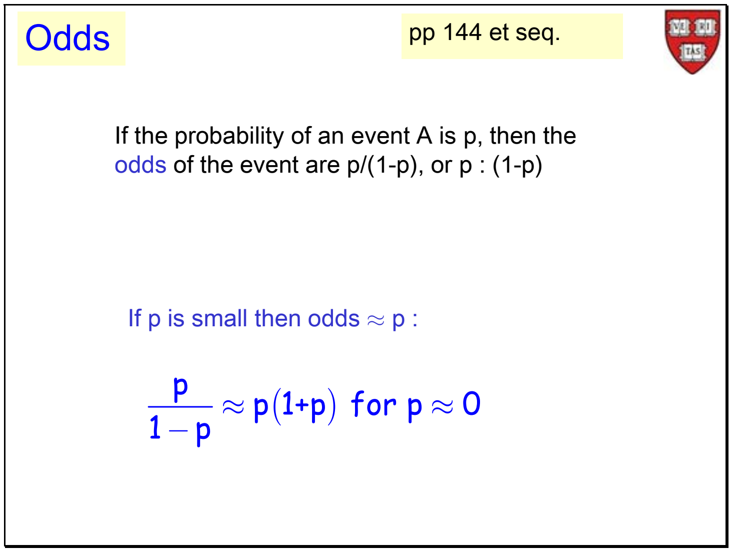
There are two ways to proceed: One is to look at their differences—so look at $p_1 – p_2$, and ask is this difference equal to zero, which is exactly what we did with the two-mean situation in the continuous case. I am going to leave this approach for you to explore. The Stata programs are all there.

Another approach this is to look at their ratio; namely p_1/p_2, or the *relative risk*. In this approach you need be careful when p_2 is equal to 0. The “null” value would be that this ratio is one.

Of course, once we think of ratios we think of our friend the odds. Because of the relationship between probabilities and odds, we know that saying the relative risk is one is equivalent to saying that the relative odds, or the odds ratio, is one.

The two approaches are equivalent as far as testing the null hypothesis that the ratio is one, but if one approach has better statistical properties, then by all means follow that approach. And that is why we look at the odds ratio in this situation.

Review of Odds



Odds

pp 144 et seq.

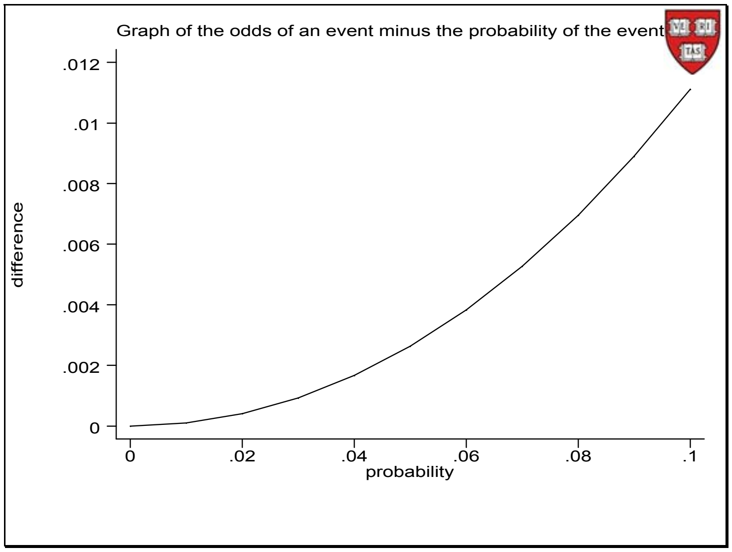
If the probability of an event A is p, then the odds of the event are p/(1-p), or p : (1-p)

If p is small then odds ≈ p :

$$\frac{p}{1-p} \approx p(1+p) \text{ for } p \approx 0$$

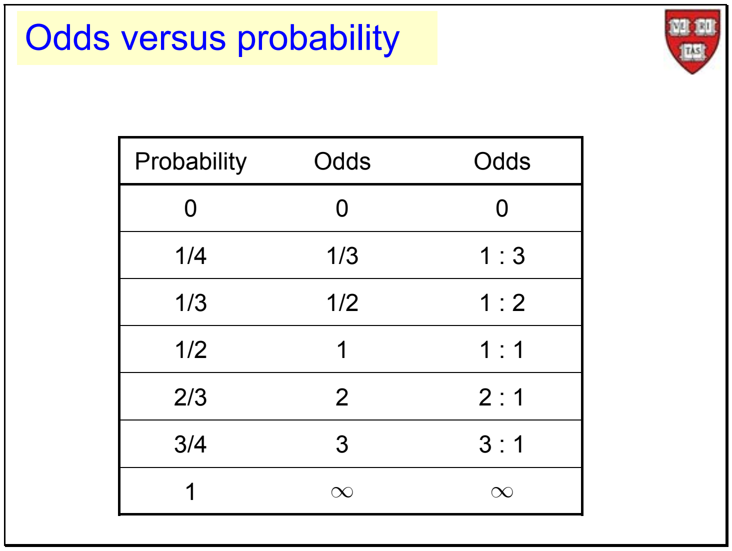
Let us take a quick refresher on relative odds. Remember, if the probability of an event A is p, then the odds of the event are p over 1 – p.

Sometimes you see this stated as “p to 1 – p”, or the ratio of p to 1– p. You will also in the literature that if p is tiny, then the odds are approximately the same as the probability, p. And that is because if you are going to expand one over 1– p for small p, then that is approximately the same as 1 + p. And so the odds are approximately equal to p, because you leave out the p^2 term.



Graph of the odds of an event minus the probability of the event

Plotted above is the difference between the odds of an event and its probability, when the probability is less than 0.1. You can see that the difference is sizable on the righty, but very small on the left.

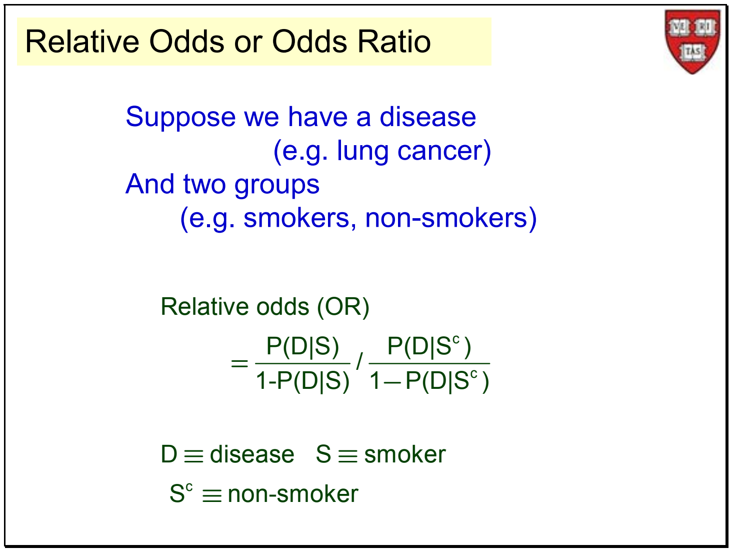


Odds versus probability

Probability	Odds	Odds
0	0	0
1/4	1/3	1 : 3
1/3	1/2	1 : 2
1/2	1	1 : 1
2/3	2	2 : 1
3/4	3	3 : 1
1	∞	∞

And just to remind you, if the probability is 0 the odds are 0. If the probability is 1, the odds are infinite. With probability smaller than ½, the corresponding odds are less than 1. With probability ½ the odds are one, and with probability greater than ½, the odds are bigger than 1. So, when the event is less likely to happen than not happen, the odds are less than one. If the event is as likely to happen as not, the odds are one. If the event is more likely to happen as not, the odds are greater than one.

Sometimes you see odds as stated in this way, other times they are stated as a ratio. For example, odds of 1/3 are sometimes stated as 1 to 3. Odds of 1/2 are 1 to 2. Evens, 1 to 1. When you get above 1, then it's 2 to 1, 3 to 1, instead of 2 it's 2 to 1, 3 to 1, cetera. So this just depends on the culture you are in.



Relative Odds or Odds Ratio

Suppose we have a disease
(e.g. lung cancer)
And two groups
(e.g. smokers, non-smokers)

Relative odds (OR)

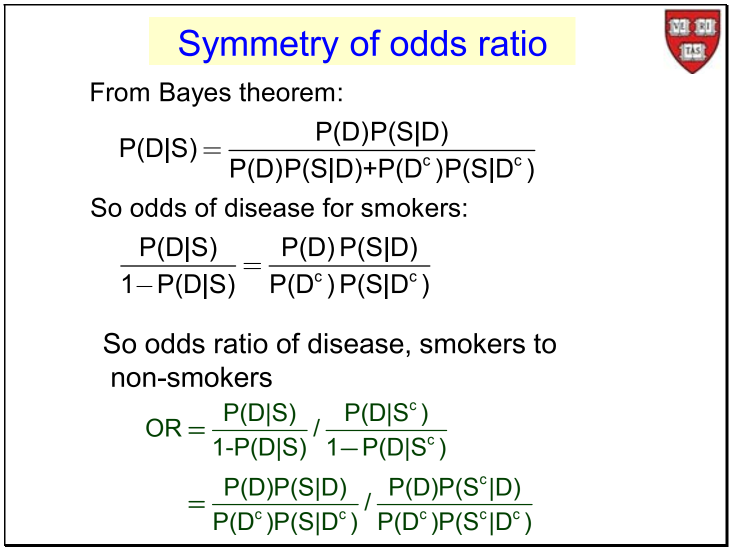
$$= \frac{P(D|S)}{1\text{-}P(D|S)} \Big/ \frac{P(D|S^c)}{1-P(D|S^c)}$$

$D \equiv$ disease $\quad S \equiv$ smoker

$S^c \equiv$ non-smoker

Now here is the real strength of the odds, and that is when you look at the odds ratio or relative odds. They are closely related to our use of Bayes' theorem: Suppose we have a disease such as lung cancer. And we have got two groups of people, smokers and nonsmokers. Now what are the relative odds of the disease, for smokers versus nonsmokers? So we first find the odds of the disease for smokers. Then do the same for non-smokers. Then take the ratio of these two odds.

So that's the relative odds of the disease for smokers relative to nonsmokers.



Symmetry of odds ratio

From Bayes theorem:

$$P(D|S) = \frac{P(D)P(S|D)}{P(D)P(S|D)+P(D^c)P(S|D^c)}$$

So odds of disease for smokers:

$$\frac{P(D|S)}{1-P(D|S)} = \frac{P(D)\,P(S|D)}{P(D^c)\,P(S|D^c)}$$

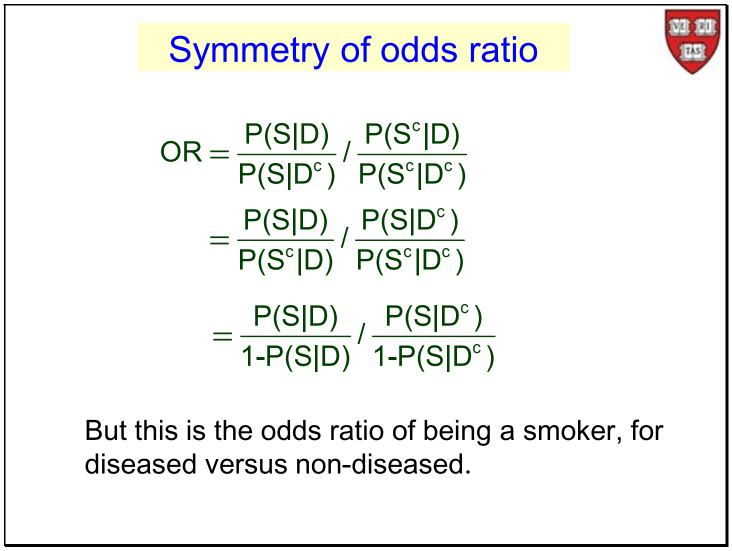
So odds ratio of disease, smokers to non-smokers

$$OR = \frac{P(D|S)}{1\text{-}P(D|S)} / \frac{P(D|S^c)}{1-P(D|S^c)}$$

$$= \frac{P(D)P(S|D)}{P(D^c)P(S|D^c)} / \frac{P(D)P(S^c|D)}{P(D^c)P(S^c|D^c)}$$

Now recall Bayes' theorem, first for smokers and then for non-smokers. We see that we can write the odds of the disease for smokers as a ratio where the numerator is the product of the probability of having the disease and the probability of being a smoker, given that one has the disease. Similarly, the denominator is the same product but for those without the disease.

After developing a similar expression for the odds of disease for the non-smokers, we can take their ratio, and simplify by canceling some common terms.



Symmetry of odds ratio

$$
\begin{aligned}
OR &= \frac{P(S|D)}{P(S|D^c)} / \frac{P(S^c|D)}{P(S^c|D^c)} \\
&= \frac{P(S|D)}{P(S^c|D)} / \frac{P(S|D^c)}{P(S^c|D^c)} \\
&= \frac{P(S|D)}{1\text{-}P(S|D)} / \frac{P(S|D^c)}{1\text{-}P(S|D^c)}
\end{aligned}
$$

But this is the odds ratio of being a smoker, for diseased versus non-diseased.

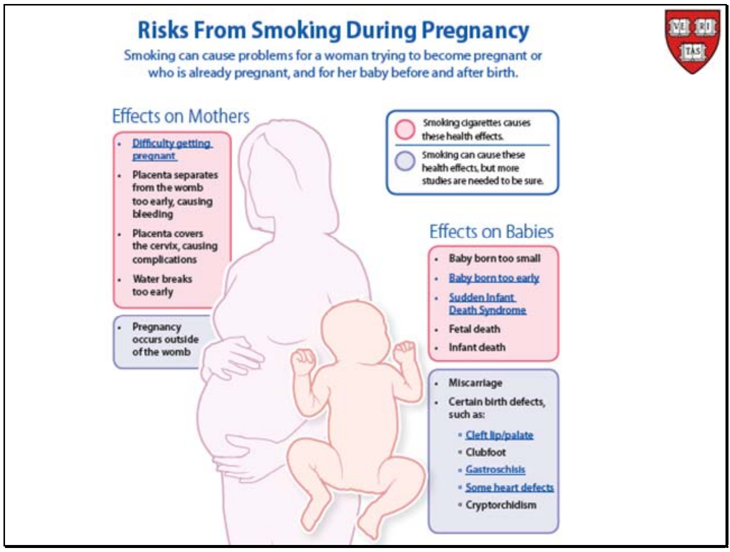
After rearranging the terms we recognize that the odds ratio can also be written as the ratio of the odds of being a smoker, given one has the disease, and the odds that one is a smoker amongst those non-diseased!

This rather surprising reversal says that the relative odds of disease amongst smokers and non-smokers, is exactly the same as the odds of smoking amongst those diseased and non-diseased. So if you are interested in the relative odds of getting the disease by smoking, then that is exactly the same as the relative odds of smoking and having the disease. So sampling amongst smokers and non-smokers and then determining their disease status in time, is the same as sampling amongst diseased and non-diseased individuals and determining their smoking status.

That means you can follow smokers over a lifetime to see what the odds are of getting lung cancer. Do the same for non-smokers. Take the ratio of those two odds to see the relative odds. Alternatively, find a group inflicted with lung cancer and see what the odds are that they are smokers. Do the same with a comparable group of individuals who do not have lung cancer. Calculate the relative odds for these two groups, and the Bayes' theorem tells us that these two relative odds are the same.

That is the theory behind case control studies.

Case Control Example



Risks From Smoking During Pregnancy

Smoking can cause problems for a woman trying to become pregnant or who is already pregnant, and for her baby before and after birth.

○ Smoking cigarettes causes these health effects.

○ Smoking can cause these health effects, but more studies are needed to be sure.

Effects on Mothers

- Difficulty getting pregnant
- Placenta separates from the womb too early, causing bleeding
- Placenta covers the cervix, causing complications
- Water breaks too early

- Pregnancy occurs outside of the womb

Effects on Babies

- Baby born too small
- Baby born too early
- Sudden Infant Death Syndrome
- Fetal death
- Infant death

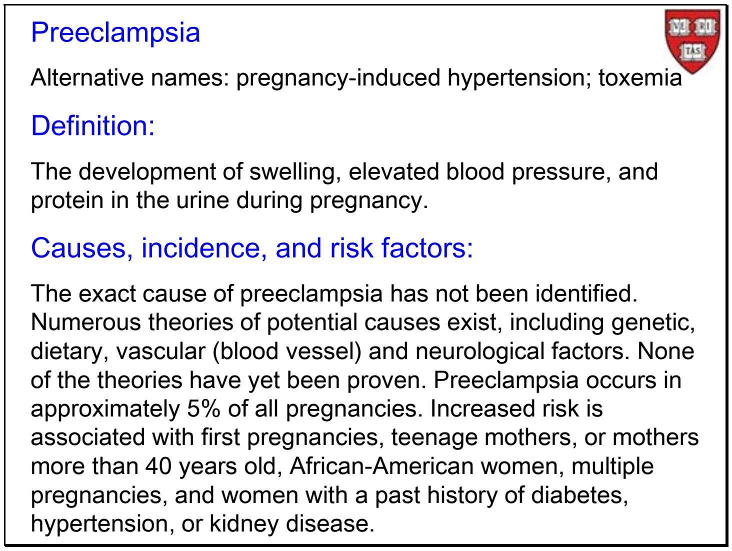
- Miscarriage
- Certain birth defects, such as:
  - Cleft lip/palate
  - Clubfoot
  - Gastroschisis
  - Some heart defects
  - Cryptorchidism

[1]

Let us look at an example of a case control study that deals with the risks of smoking during pregnancy. This poster if from the Centers for Disease Control, and is part of their campaign to alert pregnant women about the ill effects of smoking during pregnancy. It lists the effects on the mother, on the left, and the effects on babies, on the right, and none of these look very good for you.

One outcome I want you to pay attention to is fetal death. All these side effects are serious, of course, but fetal death is particularly pertinent to the study at hand:

[1] http://www.cdc.gov/reproductivehealth/TobaccoUsePregnancy/PDF/SmokingPregRisk.pdf



Preeclampsia

Alternative names: pregnancy-induced hypertension; toxemia

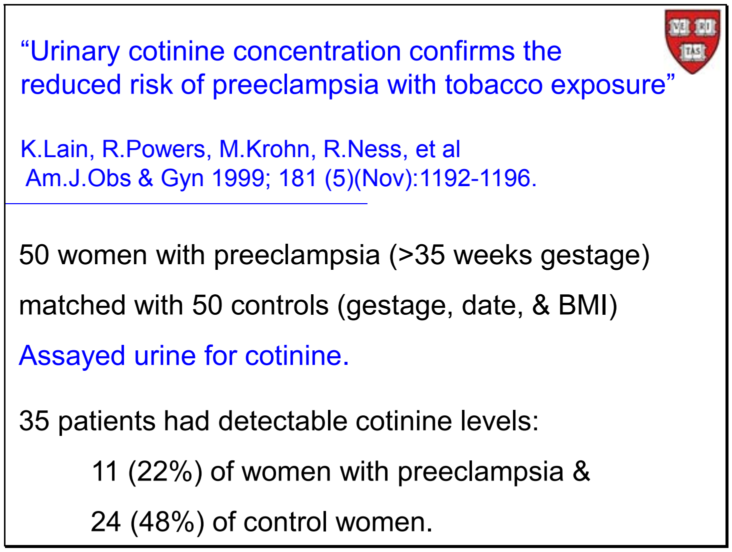
Definition:

The development of swelling, elevated blood pressure, and protein in the urine during pregnancy.

Causes, incidence, and risk factors:

The exact cause of preeclampsia has not been identified. Numerous theories of potential causes exist, including genetic, dietary, vascular (blood vessel) and neurological factors. None of the theories have yet been proven. Preeclampsia occurs in approximately 5% of all pregnancies. Increased risk is associated with first pregnancies, teenage mothers, or mothers more than 40 years old, African-American women, multiple pregnancies, and women with a past history of diabetes, hypertension, or kidney disease.

The study deals with preeclampsia, also known as toxemia, so let me define it briefly for those of you who do not know what it is. First, the condition does not occur too often. It has a prevalence of only about 5% of all pregnancies. It typically happens late in the pregnancy, and is very dangerous. Indeed, the only way of saving the mother's life is to deliver the baby immediately; typically with a ceasarian-section.



“Urinary cotinine concentration confirms the reduced risk of preeclampsia with tobacco exposure”

K.Lain, R.Powers, M.Krohn, R.Ness, et al
Am.J.Obs & Gyn 1999; 181 (5)(Nov):1192-1196.

50 women with preeclampsia (>35 weeks gestage)

matched with 50 controls (gestage, date, & BMI)

Assayed urine for cotinine.

35 patients had detectable cotinine levels:

11 (22%) of women with preeclampsia &

24 (48%) of control women.

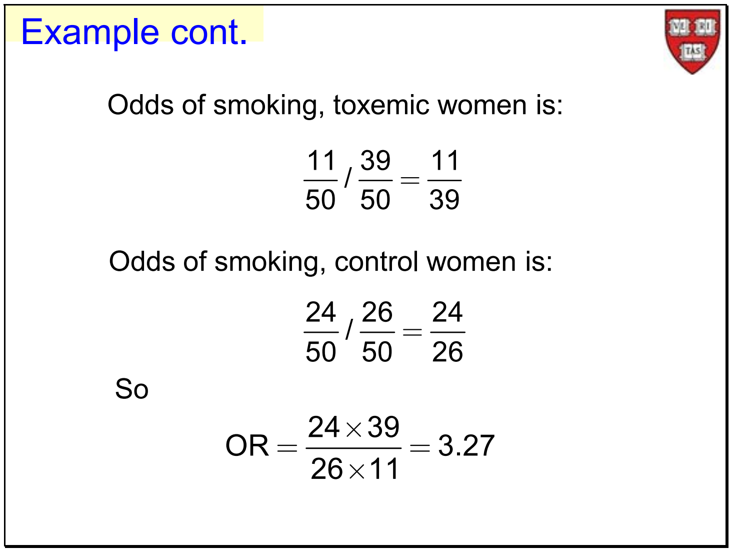
Let us look at the study called, "Urinary cotinine concentration confirms the reduced risk of preeclampsia with tobacco exposure." The reason it attracted my attention is the rather surprising claim in the title that there might be a beneficial outcome to smoking during pregnancy.

Remember the cotinine concentration. Cotinine is a metabolite for nicotine, and so mothers with high cotinine in their urine presumably can be identified as being smokers, or ingesting nicotine.

The study is a case control study, and they took 50 women with preeclampsia—here it was determined at more than 35 weeks of gestational age—and they matched those 50 women with 50 controls. And they were matched on gestational age, on date, and BMI, Body Mass Index.

They found that there were 35 patients of the 100, with cotinine in the urine, and how they split up is very interesting: 11 of the women with preeclampsia, 50, had high cotinine, whereas 24 of the control had high cotinine.

So just looking at this you say, wait a minute, there are many more smokers amongst the control group, in fact more than twice as many. So it looks like, indeed, the title is correct.



Example cont.

Odds of smoking, toxemic women is:

$$\frac{11}{50} / \frac{39}{50} = \frac{11}{39}$$

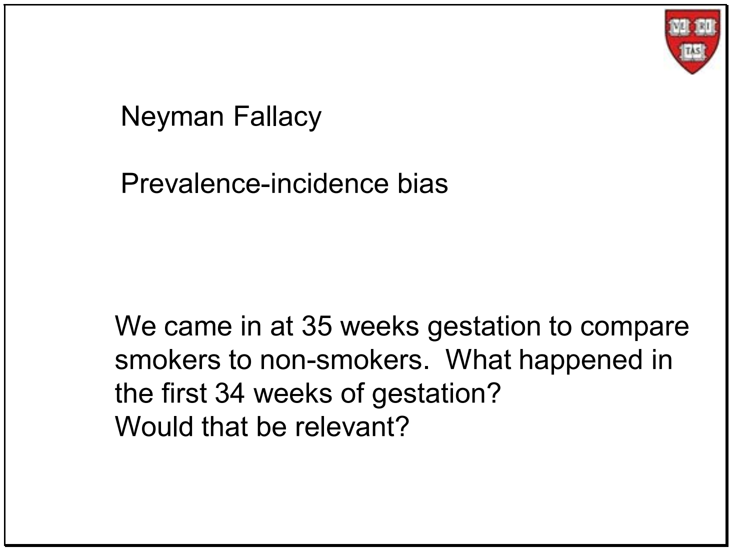
Odds of smoking, control women is:

$$\frac{24}{50} / \frac{26}{50} = \frac{24}{26}$$

So

$$OR = \frac{24 \times 39}{26 \times 11} = 3.27$$

Let us calculate the odds ratio, and it turns out to be 3.27. So we have more than a tripling of the odds for the toxemic women. What is going on?



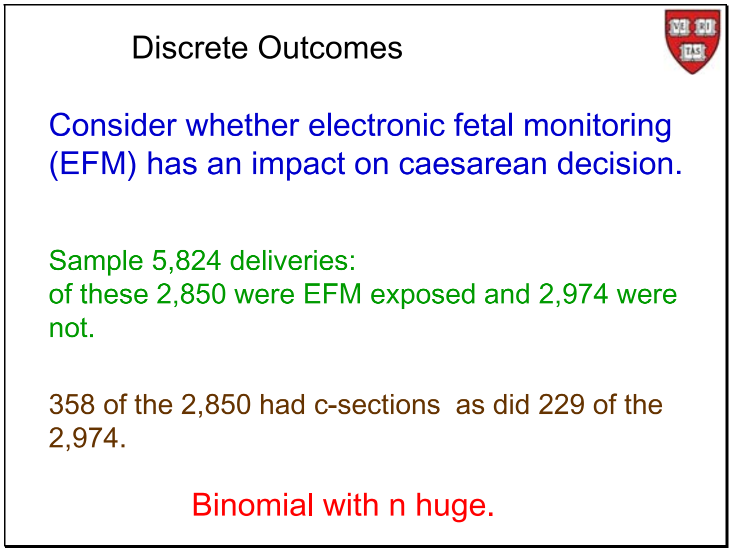
Neyman Fallacy

Prevalence-incidence bias

We came in at 35 weeks gestation to compare smokers to non-smokers. What happened in the first 34 weeks of gestation?
Would that be relevant?

We need to be careful when considering case control studies that evolve over time where the condition for entry into the study might be related to a measure associated with the study. One of the most important problems you can have is what is called the Neyman Fallacy. It is also called the Prevalence-incidence bias. It's very much like the healthy worker effect.

Entry into the study required that a woman be pregnant for 35 weeks, whence they compared smokers to non-smokers. But what about what happened in the first 34 weeks of gestation? Could that be relevant? Remember, the CDC warned about fetal deaths being attributed to smoking. So if you took 50 women who smoke at the beginning of their pregnancy, and took 50 comparable women who did not smoke, at the beginning of their pregnancy, then after 35 weeks would you still have two groups of 50? Chances are, you would not. So looking for the effects of smoking at the tail end of the pregnancy, when in fact it has had an effect throughout the pregnancy, is misleading. Just as it would have been in the Pearl smoking and longevity study to only look at the three groups he studied, after they turned 70. This is a nonsense study.



Discrete Outcomes

Consider whether electronic fetal monitoring (EFM) has an impact on caesarean decision.

Sample 5,824 deliveries:
of these 2,850 were EFM exposed and 2,974 were not.

358 of the 2,850 had c-sections as did 229 of the 2,974.

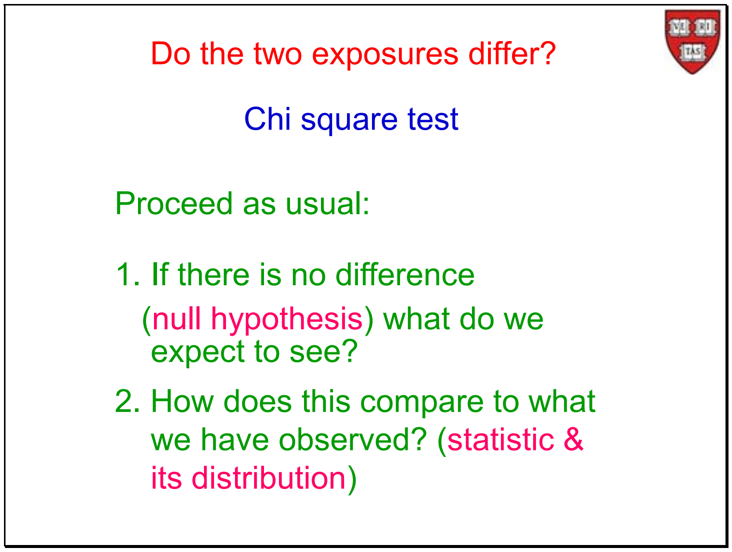
Binomial with n huge.

Returning to our primary aim, which is to look at discrete outcomes, let us start with whether two binary outcomes are related. Let me introduce this topic with an example.

Consider whether Electronic Fetal Monitoring, let's call it EFM, has an impact on the decision to have a Cesarean section, not. In this study they looked at 5,824 deliveries. Of these, in roughly half, 2,850, the women were subjected to EFM, and in other roughly half, 2,974, they were not.

Now consider how many of the women had c-sections: 358 of the 2,850 who had EFM had a c-section, and of the remaining 2,974 who were not exposed to EFM, 229 had a c-section.

So we could treat this as two binomials and use the methods already developed, but the n is exceptionally huge, so let us explore another avenue, one that is impervious to large n, and one that is more easily generalizable to the situation when we have more than two groups and also when we deal with nominal variables that take on more than two possible values.



Do the two exposures differ?

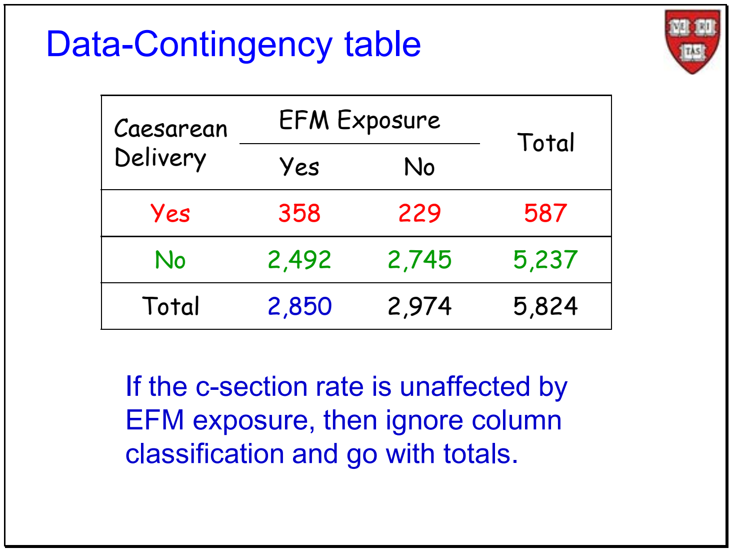
Chi square test

Proceed as usual:

1. If there is no difference (null hypothesis) what do we expect to see?
2. How does this compare to what we have observed? (statistic & its distribution)

The next test we look at is the very popular chi square test. It proceeds very much like all the tests we have encountered so far: You start with an if statement established by your null-hypothesis. Then you ask, if the null hypothesis is true—for example, that there is no difference between those who were EFM exposed and those who were not, if that is our null our null hypothesis—then what do we expect to see? What did we see? Compare the two.

We are going to have to come up with a statistic that allows us to compare the two, and establish how this statistic varies from sample to sample—in other words, determine the sampling distribution of the statistic. In other words, the same approach we have repeatedly in this course.

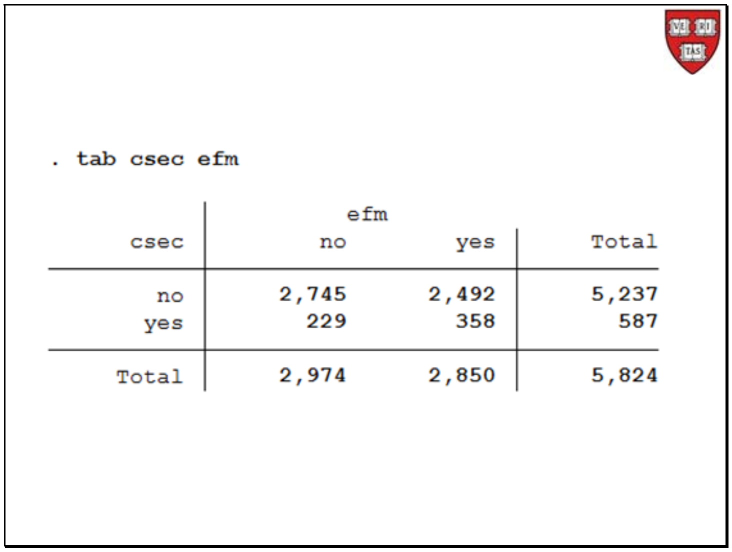


Data-Contingency table

Caesarean Delivery	EFM Exposure		Total
	Yes	No	
Yes	358	229	587
No	2,492	2,745	5,237
Total	2,850	2,974	5,824

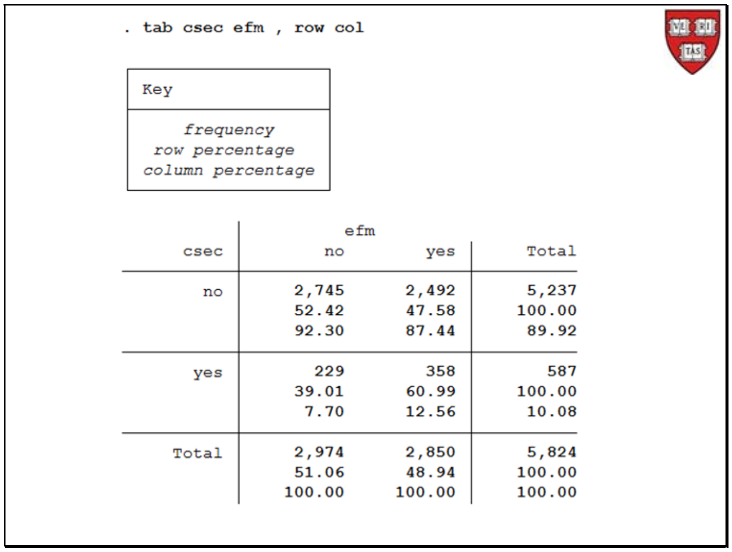
If the c-section rate is unaffected by EFM exposure, then ignore column classification and go with totals.

Let us go back to our numbers, displayed above in a two by two table. If the null-hypothesis that the c-section rate is the same for those exposed to EFM as it is for those not exposed to EFM, then the 587 with c-sections should be distributed amongst the two groups in the same ratio as the membership in the two groups; namely in the ratio 2850 : 2974. Since 587/5824 is approximately 10%, we expect 10% of the 2850 and 10% of the 2974 to have had c-sections. This is ideally what to expect, but we also expect some sampling variability around this ideal. How to account for that?



```
. tab csec efm

                      efm
      csec         no        yes      Total
-------------------------------------------
        no      2,745      2,492      5,237
       yes        229        358        587
-------------------------------------------
     Total      2,974      2,850      5,824
```

So for example when we go and we get Stata to do these calculations for us,



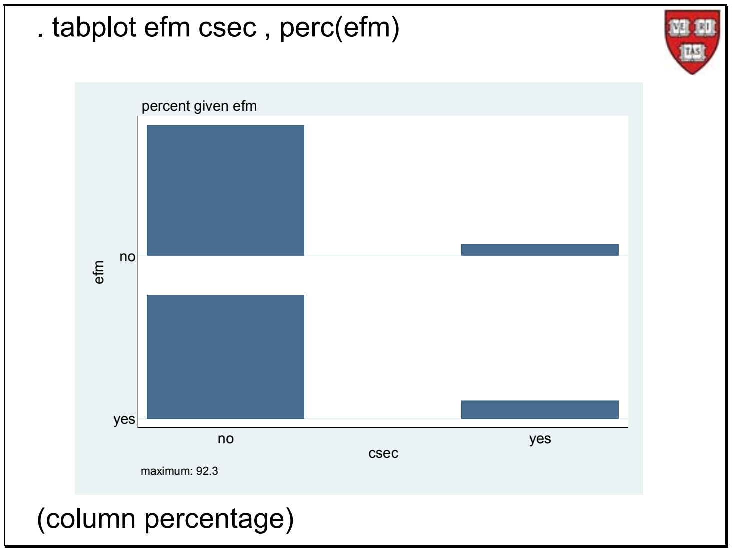
```
. tab csec efm , row col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
+-------------------+

                      efm
      csec         no        yes      Total
-------------------------------------------
        no      2,745      2,492      5,237
                52.42      47.58     100.00
                92.30      87.44      89.92
-------------------------------------------
       yes        229        358        587
                39.01      60.99     100.00
                 7.70      12.56      10.08
-------------------------------------------
     Total      2,974      2,850      5,824
                51.06      48.94     100.00
               100.00     100.00     100.00
```

we can ask Stata to show us the actual frequency, that's just like the last slide, but also show me the row percentages. So for example, 52.42% of the no c-section did not have EFM, and the other 47% did have EFM.

Inspecting this table by row, we see that of those who did have a c-section, 39% did not have EFM exposure, and 61% did. This distribution is suspiciously different from the overall tally where 51% had EFM exposure, and 48.94% did not—by design, this was supposed to be 50-50.

Alternatively we can look at this table by column. We see from the margin that about 10% had c-sections. In those not exposed to EFM we see that only 7.7% had c-sections, in contrast to those in the exposed to EFM column where 12.56% had c-sections.

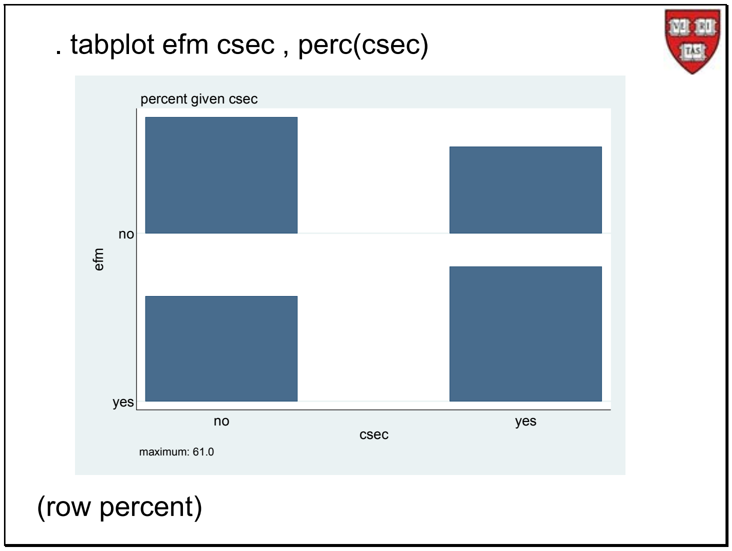
This table does not look balanced. It seems like there is a relationship between the row classification and the column classification. Is this difference attributable to chance alone?



```
. tabplot efm csec , perc(efm)
```

(column percentage)

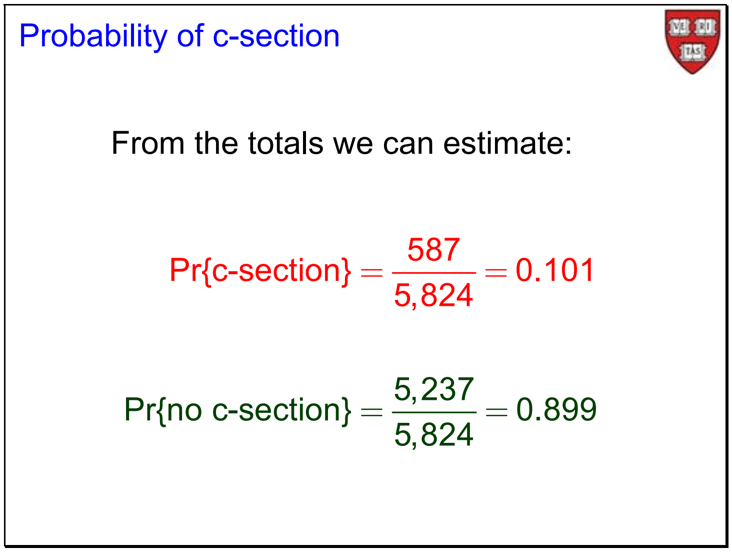
Another way of looking at this table is to plot it. You can use the *tabplot* command, to draw a bar graph for the cell counts. Here I have used the percentage EFM option; that means that we condition on EFM. So it is like asking Stata for the row percentage. So horizontally these should sum up to 100%. It looks like, the bottom right cell is a little bit thicker than the one above, in other words these two patterns do not seem to be parallel to each other.



```
. tabplot efm csec , perc(csec)
```

(row percent)

We could have taken the column percentages, instead. We now see the lack of column parallelism much more evident than before.



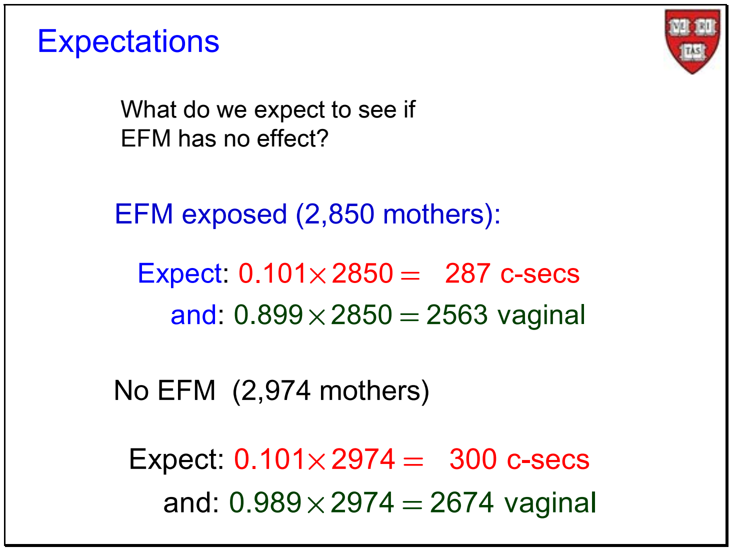
Probability of c-section

From the totals we can estimate:

$$\text{Pr\{c-section\}} = \frac{587}{5{,}824} = 0.101$$

$$\text{Pr\{no c-section\}} = \frac{5{,}237}{5{,}824} = 0.899$$

To judge whether we can attribute the evident differences between what we expected to see (parallelism) and what we saw, let us return to the numbers to quantify our expectations under the null: From the totals we can estimate, as we did before, the probability of a c-section—and that is 0.101, or roughly 10%. That means that the probability of no c-section is 0.899. So let's apply this probability of c-section to each of the two groups, those who had the EFM and those who did not.



Expectations

What do we expect to see if
EFM has no effect?

EFM exposed (2,850 mothers):

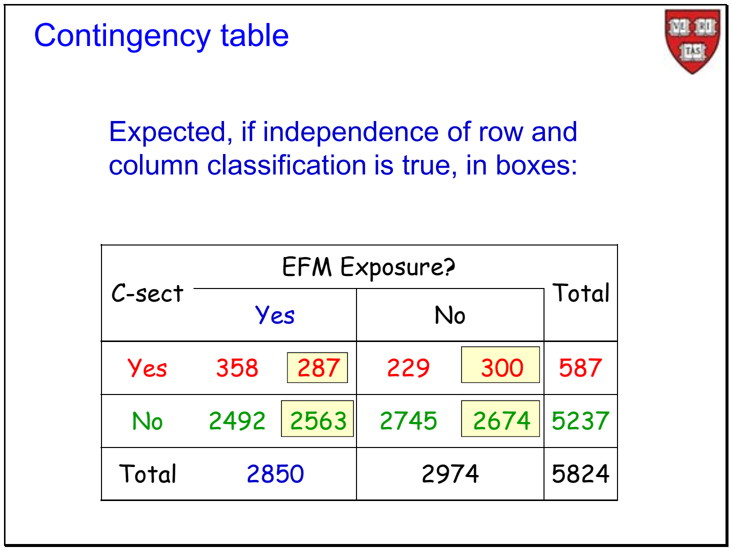
Expect: $0.101 \times 2850 = 287$ c-secs
and: $0.899 \times 2850 = 2563$ vaginal

No EFM (2,974 mothers)

Expect: $0.101 \times 2974 = 300$ c-secs
and: $0.989 \times 2974 = 2674$ vaginal

So what do we expect to see if EFM has no effect? Well then we'd expect that of the EFM exposed mothers, all 2,850 of them, roughly 10% should have a c-section. So we expect 0.101 times the 2850. So that's 287. So we expect that 287 of these mothers would have had a c-section. And that means that roughly 90% percent of them, 2850, which gives us 2563, would have had a vaginal delivery.

Let us do the same thing now for the mothers who were not EFM exposed. So of these 2974, roughly 10% percent of them, which comes out to be 300, we expect to have had c-sections, and the other 2,674 to have had vaginal deliveries.



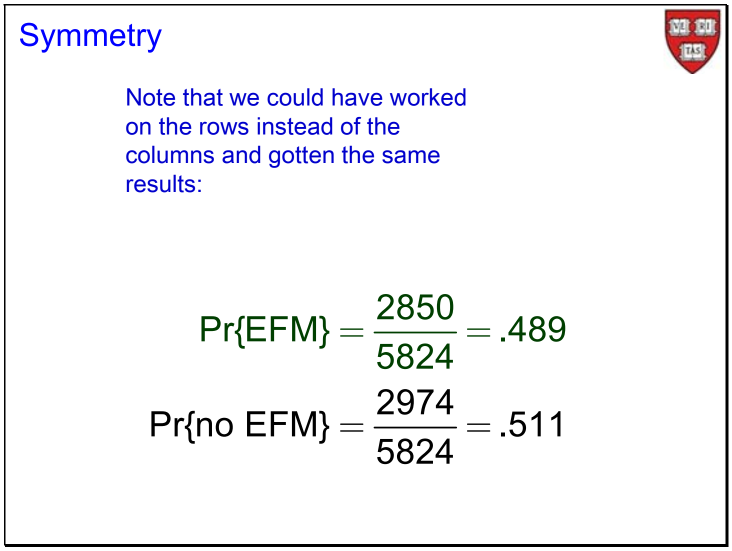
Contingency table

Expected, if independence of row and column classification is true, in boxes:

C-sect	EFM Exposure?				Total
	Yes		No		
Yes	358	287	229	300	587
No	2492	2563	2745	2674	5237
Total	2850		2974		5824

Let us gather these numbers and place them in boxes within our two by two table so we can contrast what we expect with what we actually saw.

Within each cell we see that actually observed. We see that in the yes-yes cell (top left hand corner) we expected to see 287, but we actually saw many more, namely 358. Since across the table, for each row and column, the sum of the observed and the sum of the expecteds is the same, if one expected value is too large, that means that the entry in the next row, or column, must be too small.



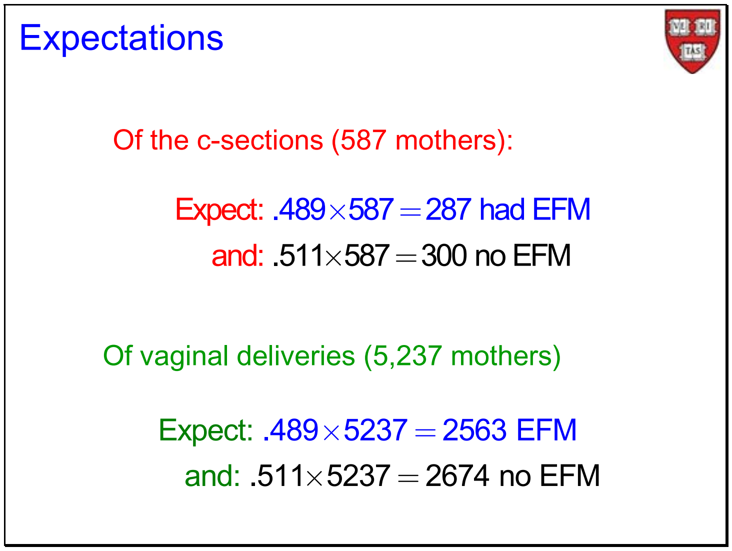
Symmetry

Note that we could have worked on the rows instead of the columns and gotten the same results:

$$\Pr\{\text{EFM}\} = \frac{2850}{5824} = .489$$

$$\Pr\{\text{no EFM}\} = \frac{2974}{5824} = .511$$

Note we could have worked along the rows instead of the columns and gotten the same results. We could have argued, for example, that 48.9% of all the women were subjected to EFM.



Expectations

Of the c-sections (587 mothers):

Expect: $.489 \times 587 = 287$ had EFM

and: $.511 \times 587 = 300$ no EFM

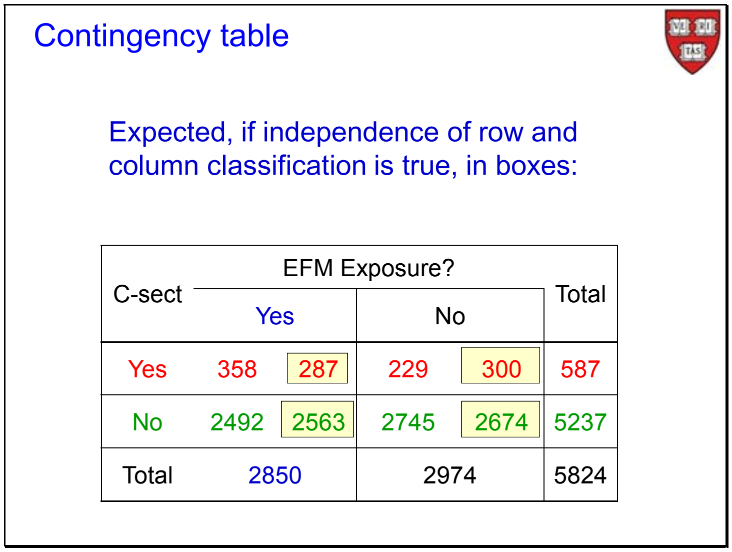
Of vaginal deliveries (5,237 mothers)

Expect: $.489 \times 5237 = 2563$ EFM

and: $.511 \times 5237 = 2674$ no EFM

So applying the probability of EFM to the mothers who had c-sections, we expect that 48.9% of these 587, or 287 had an EFM, and the remaining 300 not to have had EFM. Similarly, of the 5,237 who had vaginal deliveries, 48.9%, or 2,563, to have been exposed to EFM, and the other 2674, to have not been exposed to EFM.

This set of expected numbers is exactly what we calculated above, and if there is no relationship between the row and column classifications, then this is what we expect to see.



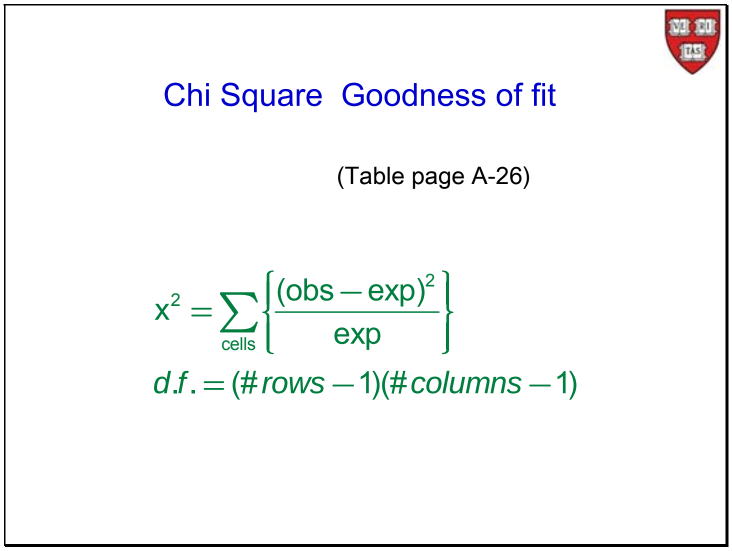
Contingency table

Expected, if independence of row and column classification is true, in boxes:

C-sect	EFM Exposure?				Total
	Yes		No		
Yes	358	287	229	300	587
No	2492	2563	2745	2674	5237
Total	2850		2974		5824

To compare what we have seen to what we expected under the null, we can attempt to do what we did before, but we have four contrasts to make here, whereas in the past (with the mean, for example) we only had one comparison to make. We could take the four differences and sum them up, but when we do that we always get zero; just like when we tried to average out all the distances from the mean. In that case we were led to the standard deviation by taking the squares of the deviations and we can do the same thing here.

A problem with this approach is that the cells are not the same size to begin with. A discrepancy of 10 observations in the first row is of much more consequence, when comparing expected to observed, in a row where the average size is 300, than a discrepancy of ten in the second row where the average cell size is about 2,600. So we can standardize before summing by dividing by the expected value in the cell.



Chi Square Goodness of fit

(Table page A-26)

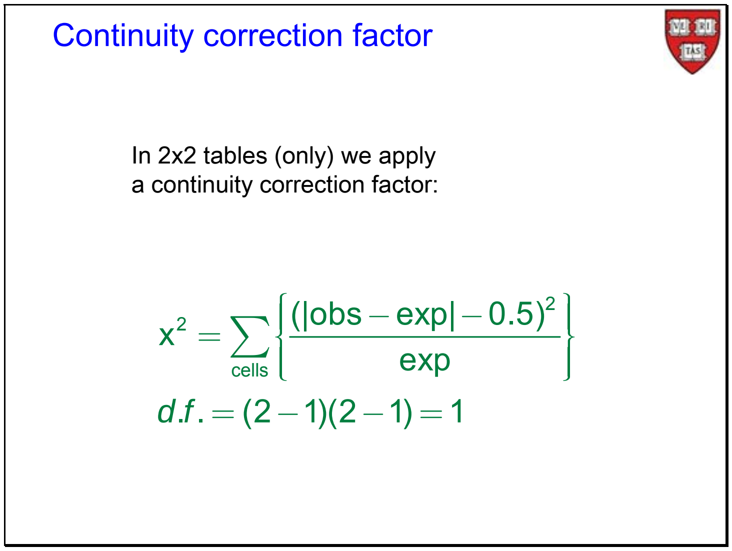
$$x^2 = \sum_{\text{cells}} \left\{ \frac{(\text{obs} - \text{exp})^2}{\text{exp}} \right\}$$

$$d.f. = (\#rows - 1)(\#columns - 1)$$

The X^2 statistic, introduced by Pearson, aggregates, over all the cells, the observed minus the expected squared, divided by the expected. Its sampling distribution is approximated by a Chi-squared distribution. The surprising thing is that this is almost independent of the number of observations we have. It is also quite robust to the sampling plan—for example, it works here even though we chose 50% of the women to have EFM, and it would have worked the same whether we had just spun a coin for each woman to determine whether she gets EFM or not.

And just as with the t, we need degrees of freedom. And degrees of freedom are calculated as the number of rows minus 1 times the number of columns minus 1. In this case it was a 2x2 table, so it's 2 minus 1 times 2 minus 1, which is equal to 1.

Intuitively, the reason why this is so is because once you fix one of the cell numbers, the other three are available by subtraction. So you have only one degree.



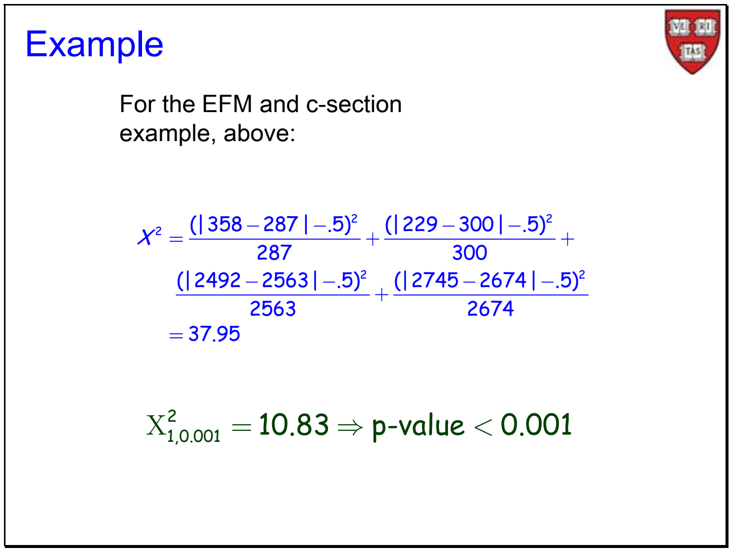
Continuity correction factor

In 2x2 tables (only) we apply
a continuity correction factor:

$$x^2 = \sum_{\text{cells}} \left\{ \frac{(|\text{obs} - \text{exp}| - 0.5)^2}{\text{exp}} \right\}$$

$$d.f. = (2-1)(2-1) = 1$$

There is an exception to the above rule. In the single instance when we have a 2x2 table, as opposed to the forthcoming bigger tables, we get a better approximation to the sampling distribution of the X^2 statistic if we make a so called *continuity correction*. Before squaring each cell, we decrease the absolute difference within the cell by 0.5.



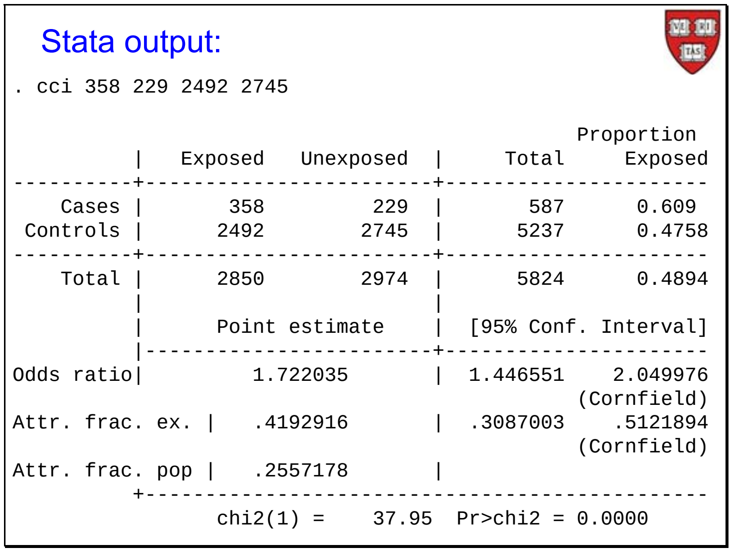
Example

For the EFM and c-section example, above:

$$X^2 = \frac{(|358-287|-.5)^2}{287} + \frac{(|229-300|-.5)^2}{300} + \frac{(|2492-2563|-.5)^2}{2563} + \frac{(|2745-2674|-.5)^2}{2674} = 37.95$$

$$X^2_{1,0.001} = 10.83 \Rightarrow \text{p-value} < 0.001$$

So here is a sample calculation one would do for the current study. The p-value is less than 0.05, so we reject the null hypothesis that there is no relation between the row and column classifications. That means that we conclude that it is unlikely that the discrepancy between the observed and expected happened just by chance and that we believe that the EFM caused more c-sections to be performed than would have been done without the EFM.

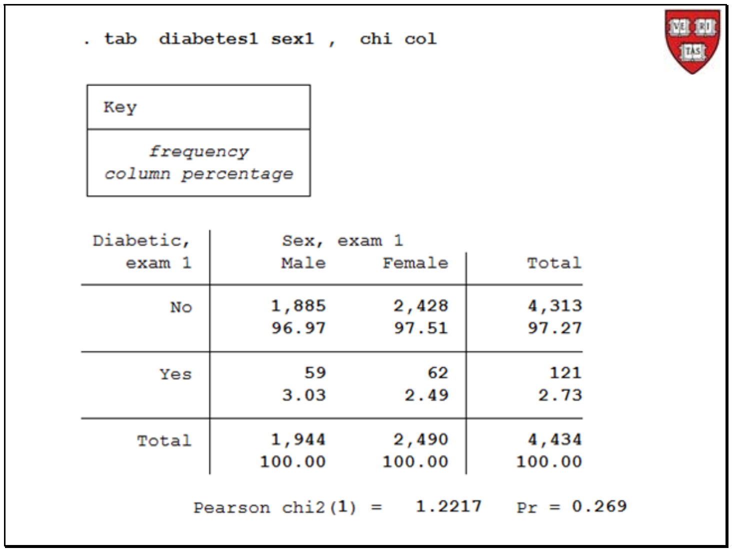


Stata output:

```
. cci 358 229 2492 2745

                                               Proportion
          |   Exposed   Unexposed  |      Total     Exposed
----------+------------------------+----------------------
     Cases |       358         229  |       587        0.609
 Controls |      2492        2745  |      5237       0.4758
----------+------------------------+----------------------
     Total |      2850        2974  |      5824       0.4894
          |                        |
          |      Point estimate    |   [95% Conf. Interval]
          |------------------------+----------------------
Odds ratio|         1.722035       |   1.446551    2.049976
                                               (Cornfield)
Attr. frac. ex. |   .4192916       |   .3087003    .5121894
                                               (Cornfield)
Attr. frac. pop |   .2557178       |
          +-----------------------------------------------
                 chi2(1) =     37.95  Pr>chi2 = 0.0000
```

If we go straight to Stata, we would use the *cc* command (case control). We see that this agrees with our hand calculation.

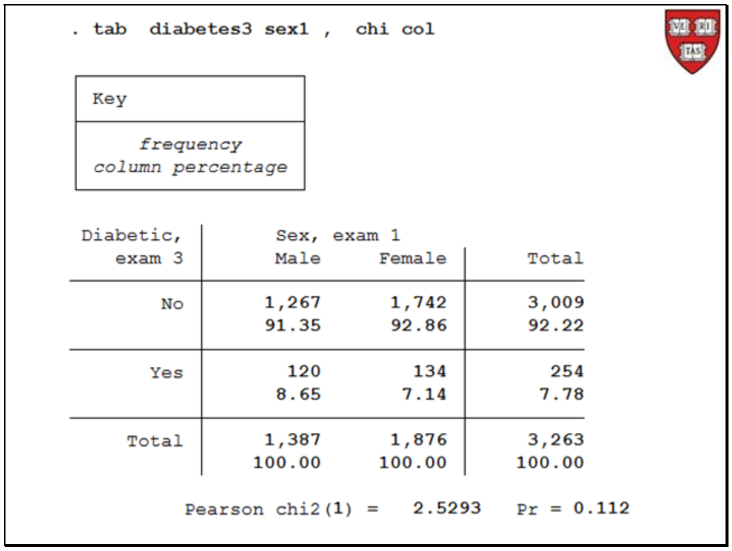
Rxc Tables



```
. tab  diabetes1 sex1 ,  chi col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

 Diabetic, |      Sex, exam 1
    exam 1 |      Male     Female |     Total
-----------+----------------------+----------
        No |     1,885      2,428 |     4,313 
           |     96.97      97.51 |     97.27 
-----------+----------------------+----------
       Yes |        59         62 |       121 
           |      3.03       2.49 |      2.73 
-----------+----------------------+----------
     Total |     1,944      2,490 |     4,434 
           |    100.00     100.00 |    100.00 

          Pearson chi2(1) =   1.2217   Pr = 0.269
```

Returning to our Framingham Heart Study we can look at the relationship between diabetes at visit one and the subjects' sex. The above analysis says that there does not appear to be any relationship and we do not reject the null hypothesis of independence of the row and column classifications.

The nice thing about the chi squared test is that we can extend it to more than two by two tables. In fact, we can extend it to, as we say in the jargon, r by c tables; that means any number of rows (r) and any number of columns (c).

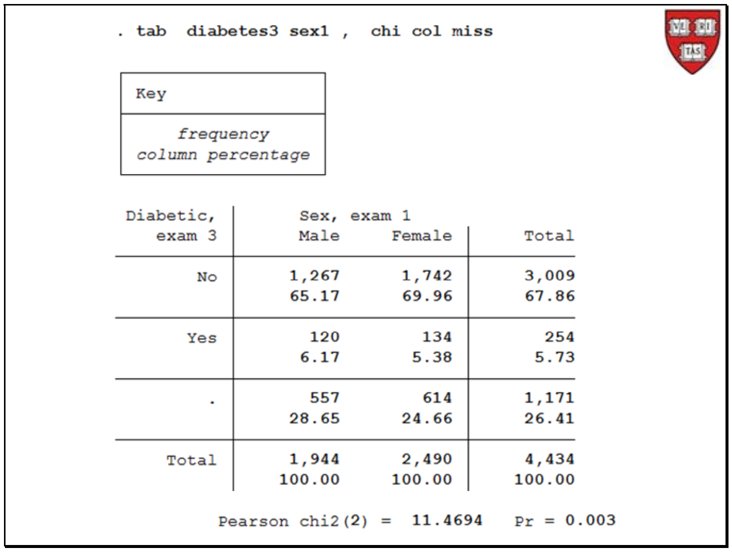


```
. tab  diabetes3 sex1 ,  chi col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

 Diabetic, |      Sex, exam 1
    exam 3 |      Male     Female |     Total
-----------+----------------------+----------
        No |     1,267      1,742 |     3,009 
           |     91.35      92.86 |     92.22 
-----------+----------------------+----------
       Yes |       120        134 |       254 
           |      8.65       7.14 |      7.78 
-----------+----------------------+----------
     Total |     1,387      1,876 |     3,263 
           |    100.00     100.00 |    100.00 

          Pearson chi2(1) =   2.5293   Pr = 0.112
```

Consider extending this very same example by looking at the diabetes status at the third exam. We get that the p-value is 0.112, so at first blush we do not reject the null hypothesis that by the time they did the exam 3 that these two classifiers, diabetes status and sex, are still independent of each other.

But recall that for a longitudinal study (one that progresses over time) we have to take care especially if there are subjects dropping out of the study. Are we still comparing the same group at visit 3 as we were at visit 1? Well at visit 3 we have 3,263 in contrast to the 4,434 we had at visit 1. So we have lost 1,171 people. We have lost a considerable number of people.

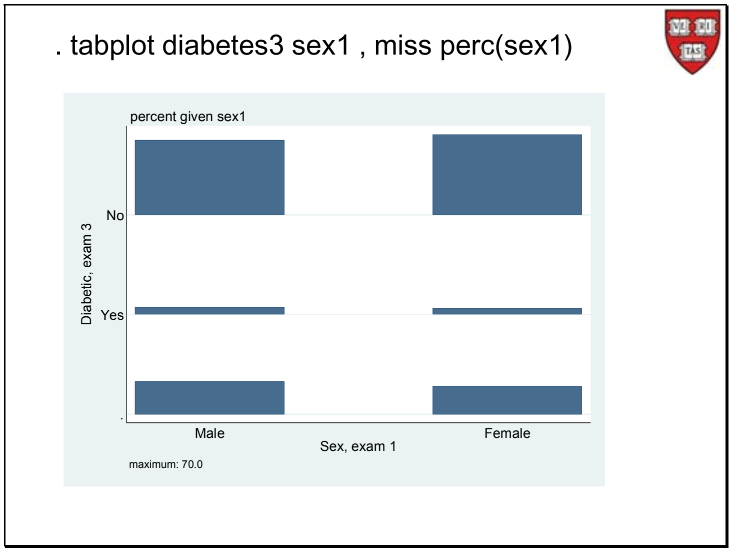
If these 1,171 were lost at random, then we might not be concerned. The problem arises if there is any relationship between the classifiers under study; diabetic status and sex. If there were a relationship, we might now get a distorted impression by visit 3.



```
. tab  diabetes3 sex1 ,  chi col miss

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

 Diabetic, |      Sex, exam 1
    exam 3 |      Male     Female |     Total
-----------+----------------------+----------
        No |     1,267      1,742 |     3,009 
           |     65.17      69.96 |     67.86 
-----------+----------------------+----------
       Yes |       120        134 |       254 
           |      6.17       5.38 |      5.73 
-----------+----------------------+----------
         . |       557        614 |     1,171 
           |     28.65      24.66 |     26.41 
-----------+----------------------+----------
     Total |     1,944      2,490 |     4,434 
           |    100.00     100.00 |    100.00 

          Pearson chi2(2) =  11.4694   Pr = 0.003
```

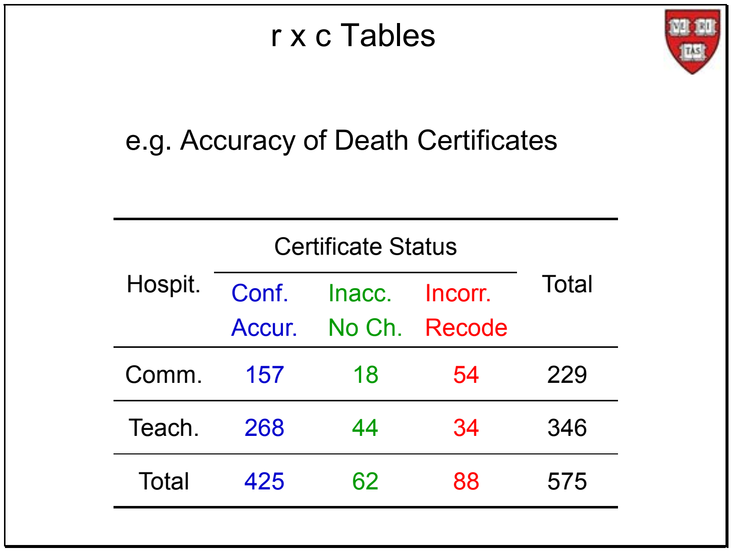
If we ask Stata to include the missings by appending the option *miss* in our tab command we get the above output.

Now that the missing 1,171 are included in our analysis, we see that there is a 4% differential between the male missing and the female missing. In fact, if we look at the distribution down the two columns we see a difference. The chiu-squared p-value is 0.003, so we would reject the null hypothesis that the row classification and the column classification are independent of each other.



```
. tabplot diabetes3 sex1 , miss perc(sex1)
```

If we *tabplot* this table, we see the above. We see a larger proportion males missing than females. These two columns are not parallel. Why? What is happening? I leave it for you to find out. I have no answer. But the statistic here is telling us that something is going on that might be of interest.



r x c Tables

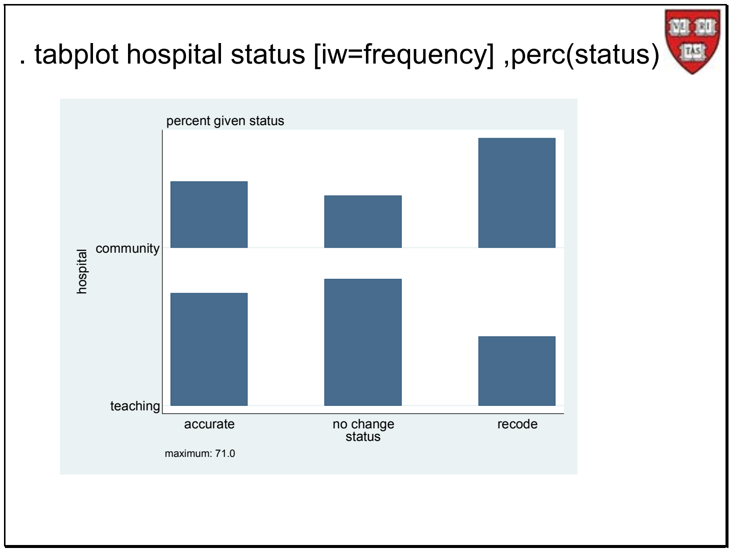
e.g. Accuracy of Death Certificates

Hospit.	Certificate Status Conf. Accur.	 Inacc. No Ch.	 Incorr. Recode	Total
Comm.	157	18	54	229
Teach.	268	44	34	346
Total	425	62	88	575

Here is another example. It approaches another, general interesting question. How accurate are death certificates? This study looked at 575 death certificates. And the row classification is whether the hospital that completed the death certificate was a community hospital or whether it was a teaching hospital. I believe this was in the state of Connecticut.

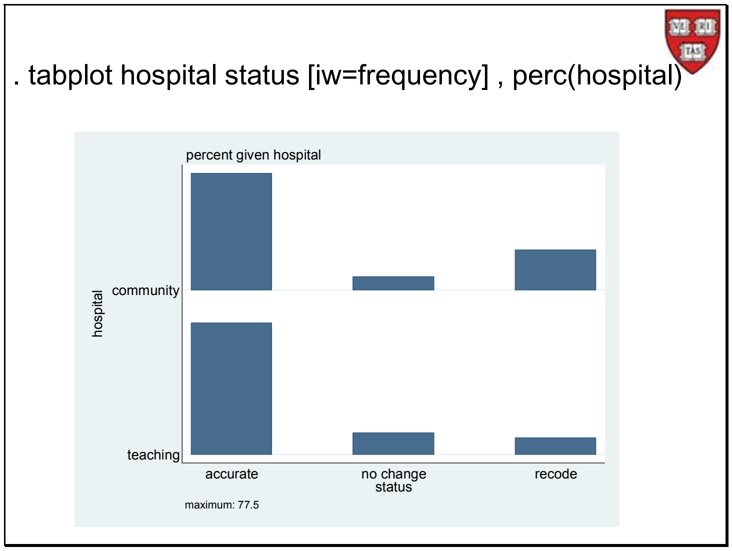
Now of these 575, 229 came from the community hospital and 346 came from the teaching hospitals. So roughly a third came from the community hospitals and 2/3 from the teaching hospitals. They compared these 575 death certificates to the medical records to check for any inaccuracies and found that of 425 of the 575 were accurately filled in. Of the remaining ones that had inaccuracies they classified them into two groups, those where the inaccuracies were not serious enough to require recoding of the death certificates (62 certificates), and those that did require a recoding of the death certificates (88 certificates).

Now the null hypothesis is that the row classification is independent of the column classification. That means that the accuracy with which these certificates were recorded was independent of the type of hospital. So since roughly a third of the certificates came from the community hospitals, that one-third ratio should maintain in each of the three columns, under the null hypothesis.



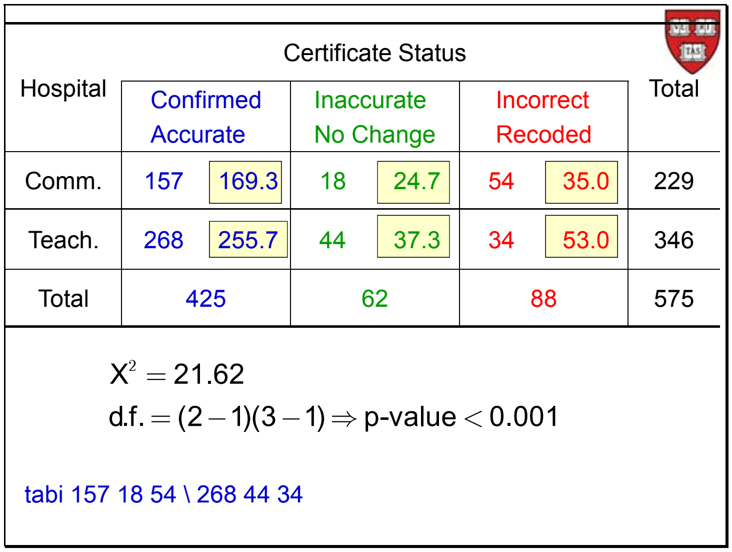
```
. tabplot hospital status [iw=frequency] ,perc(status)
```

Now if we look at the tabplot for this table with the *perc(status)* option, so that the column sums are 100%, we see some variability around this roughly one-third ratio; the largest deviation is in the “recode” column. These two rows of bar graphs do not seem to be parallel.



```
. tabplot hospital status [iw=frequency] , perc(hospital)
```

We can look at it in the other direction, namely with the perc(hospital) option so that the row sums are 100%, again we see that the two rows of bar graphs are not parallel.



Hospital	Certificate Status						Total
	Confirmed Accurate		Inaccurate No Change		Incorrect Recoded		
Comm.	157	169.3	18	24.7	54	35.0	229
Teach.	268	255.7	44	37.3	34	53.0	346
Total	425		62		88		575

$$X^2 = 21.62$$

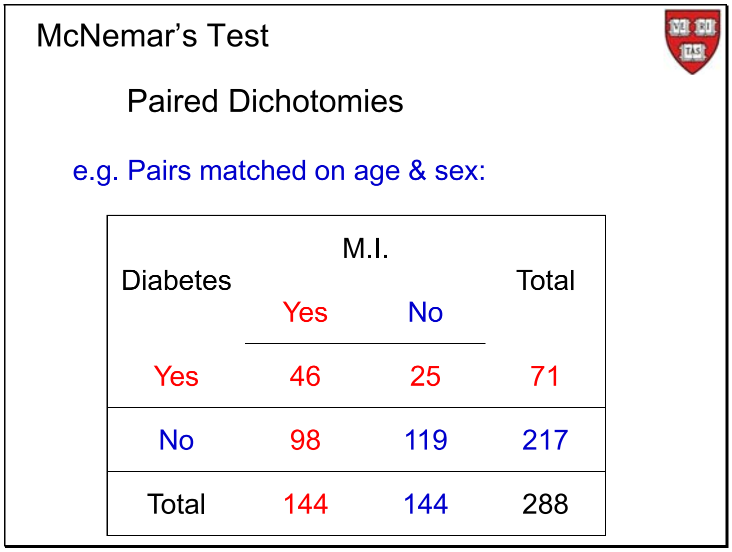
$$\text{d.f.} = (2-1)(3-1) \Rightarrow \text{p-value} < 0.001$$

```
tabi 157 18 54 \ 268 44 34
```

Our suspicions are confirmed when we submit the table to Stata to find that the p-value associated with the X^2 is less than 0.001. So we reject the null hypothesis that the row and column classifiers are independent of each other.

We thus see that the X^2 statistics extends simply to the larger tables.

McNemar Test



McNemar's Test

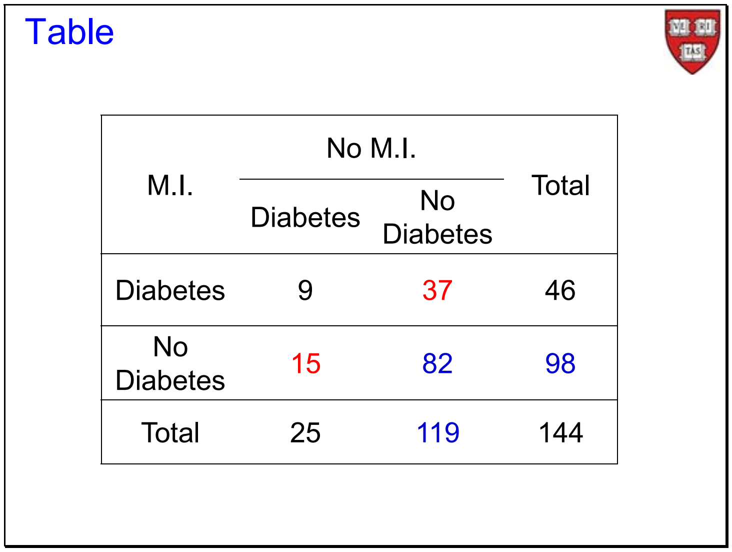
Paired Dichotomies

e.g. Pairs matched on age & sex:

Diabetes	M.I. Yes	M.I. No	Total
Yes	46	25	71
No	98	119	217
Total	144	144	288

An important two by two table is when the data actually represent matched pairs. For example, this study matched 144 pairs of individuals on their diabetes status and on whether they had suffered a myocardial infarction or not. They were matched on age and sex. So we have 288 observations, true, but they should more properly be analyzed as 144 couplets.

The test designed for such an analysis is called the McNemar Test. This is the discrete analog of what we did with the two sample, dependent t.



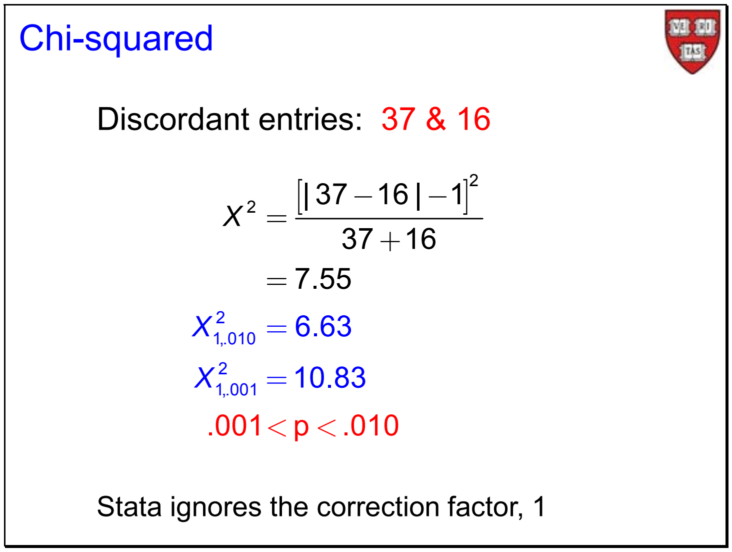
Table

M.I.	No M.I. Diabetes	No M.I. No Diabetes	Total
Diabetes	9	37	46
No Diabetes	15	82	98
Total	25	119	144

We reclassify the 144 couplets by registering the status of each pair according to the diabetes status of the MI individual versus the diabetes status of the member of the pair without MI. The argument then goes, if we are interested whether there is a relationship between MI and diabetes, then consider the diagonal elements in the above table. If both members of the couple have diabetes, or neither do, then that is not going to provide any information about the relationship between MI and diabetes. So the diagonal elements are non-informative and we can discard them.

Now consider the off diagonal cells. If those with diabetes but no MI are the same in number as those with MI but no diabetes, then there is no relationship between MI and diabetes. SO we can test for a relationship by looking at these off diagonal cells and test whether it is plausible to think that their difference can be attributed purely to chance. That is the McNemar test.

Some find it troubling that we ignore the diagonal cells, feeling somehow that there is some information there; for example you could have a zillion on this diagonal and 37 and 16 in the off diagonal cells, as we have here, and the McNemar test would give the same answer we are about to calculate.



Chi-squared

Discordant entries: 37 & 16

$$X^2 = \frac{[|37-16|-1]^2}{37+16}$$

$$= 7.55$$

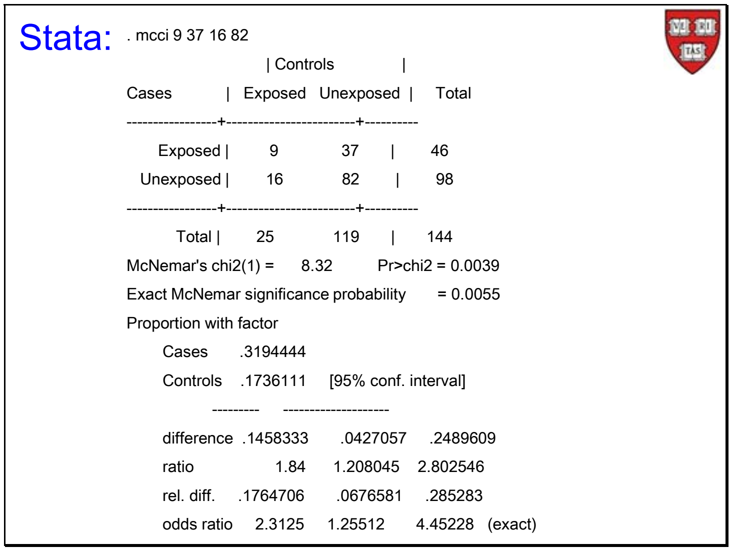
$$X^2_{1,.010} = 6.63$$

$$X^2_{1,.001} = 10.83$$

$$.001 < p < .010$$

Stata ignores the correction factor, 1

Returning to the McNemar test, we see the formula above. The 1 is a continuity correction that Stata ignores. We then compare this X^2 to a Chi-square with one degree of freedom to see that the p-value is less than 0.05, so we reject the null hypothesis, and conclude that there is a relationship between diabetes and MI.

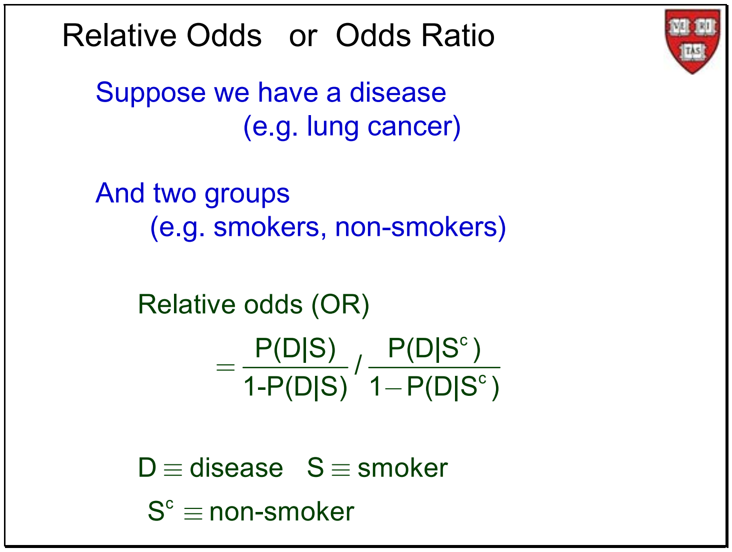


Stata:

```
. mcci 9 37 16 82
                 | Controls              |
Cases            |  Exposed  Unexposed  |   Total
-----------------+------------------------+----------
         Exposed |       9        37     |      46
       Unexposed |      16        82     |      98
-----------------+------------------------+----------
           Total |      25       119     |     144
McNemar's chi2(1) =      8.32    Pr>chi2 = 0.0039
Exact McNemar significance probability       = 0.0055
Proportion with factor
        Cases       .3194444
        Controls    .1736111     [95% conf. interval]
                   ---------     --------------------
        difference  .1458333     .0427057    .2489609
        ratio           1.84     1.208045    2.802546
        rel. diff.  .1764706     .0676581     .285283
        odds ratio    2.3125      1.25512     4.45228   (exact)
```

This is the Stata output. They get 8.32 for the statistic, instead of the 7.55 we got, but qualitatively the conclusions are the same.

Odds Ratio Review



Relative Odds or Odds Ratio

Suppose we have a disease
(e.g. lung cancer)

And two groups
(e.g. smokers, non-smokers)

Relative odds (OR)

$$= \frac{P(D|S)}{1\text{-}P(D|S)} / \frac{P(D|S^c)}{1 - P(D|S^c)}$$

$D \equiv$ disease $S \equiv$ smoker

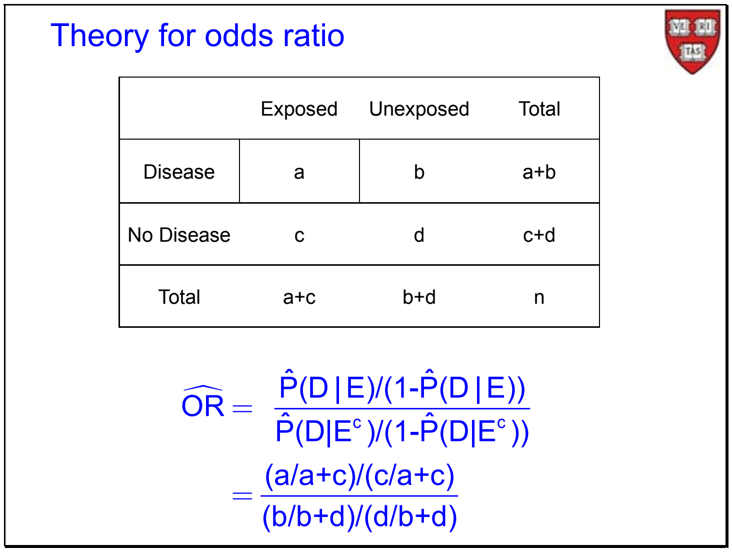
$S^c \equiv$ non-smoker

Let us return to the odds ratio for quantifying the relationship between two dichotomous classifications—the row and the column classification in our 2x2 table.

We saw the advantage of the odds ratio for case control studies, but it also quantifies these relationships in general.

Recall that 1 is the null value. It is the value for which there is no relationship between the row and column classifiers. So this pivotal value of 1 can be used as a point from which we can measure dependence; the further away the larger the dependence. The only hitch is that the differences are not symmetric around one. For example, m and 1/m are equally distant from 1 in some sense; for example odds of 3:1 are the same as 1:1/3. So the odds of 1/3:1 is just a reversal of what we consider success and what we consider failure. So if the distance from 1 is important, some people restrict themselves to restructuring the problem so as to just deal always with odds bigger than 1. We can always do this just by rephrasing the statement.

Another view is to always deal with the logarithm of odds and feeling that that domain is closer to linearity when dealing with ratios; the log of 1/3 is minus the log of 3, and that provides symmetry around 1 whose log is zero. We deal with this more fully when we look at logistic regression in the penultimate week.



Theory for odds ratio

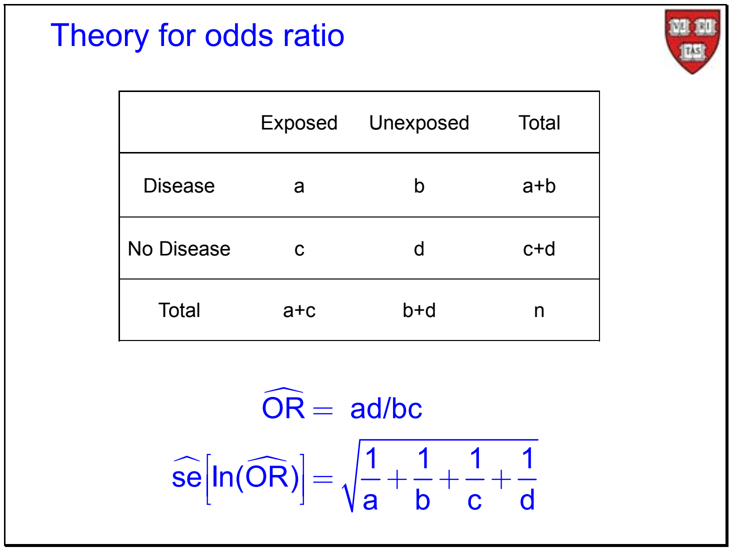
	Exposed	Unexposed	Total
Disease	a	b	a+b
No Disease	c	d	c+d
Total	a+c	b+d	n

$$\widehat{OR} = \frac{\hat{P}(D\,|\,E)/(1-\hat{P}(D\,|\,E))}{\hat{P}(D|E^c)/(1-\hat{P}(D|E^c))}$$

$$= \frac{(a/a+c)/(c/a+c)}{(b/b+d)/(d/b+d)}$$

Treating the odds ratio as a population parameter—unfortunately we do not have a Greek letter reserved for this, we just use OR, so we are being inconsistent—we are interested in estimating it on the basis of a sample from that population. The natural estimator is the one above, obtained from using the obvious estimators of the probabilities shown.

After we cancel the a+c and the b+d we are left with ac divided by bd—the product along one diagonal divided by the product along the other diagonal; a simple formula to remember.



Theory for odds ratio

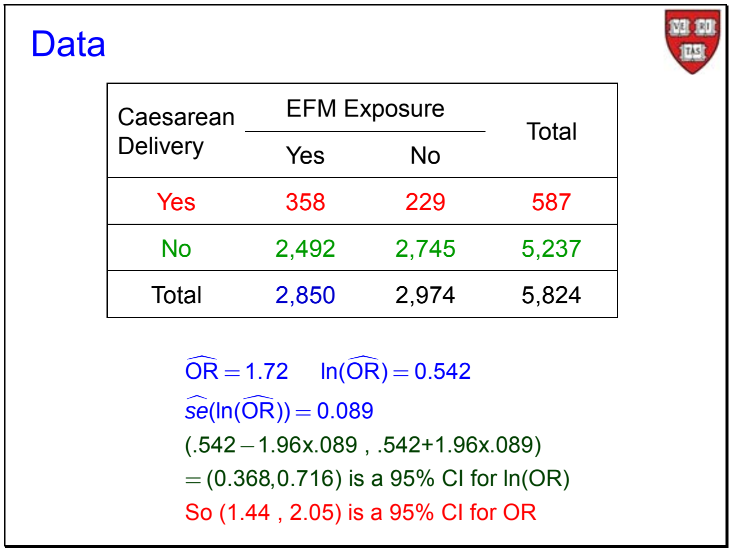
	Exposed	Unexposed	Total
Disease	a	b	a+b
No Disease	c	d	c+d
Total	a+c	b+d	n

$$\widehat{OR} = \text{ ad/bc}$$

$$\widehat{se}\left[\ln(\widehat{OR})\right] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

It turns out that the sampling distribution of the log of the estimator of the odds ratio, (ln(ad/bc)) is approximately normal and we can estimate the standard error simply by taking the sum of the reciprocals of the cell entries.[2] This is a very cute formula.

That is the theory for odds ratios.



Data

Caesarean Delivery	EFM Exposure		Total
	Yes	No	
Yes	358	229	587
No	2,492	2,745	5,237
Total	2,850	2,974	5,824

$\widehat{OR} = 1.72 \qquad \ln(\widehat{OR}) = 0.542$

$\widehat{se}(\ln(\widehat{OR})) = 0.089$

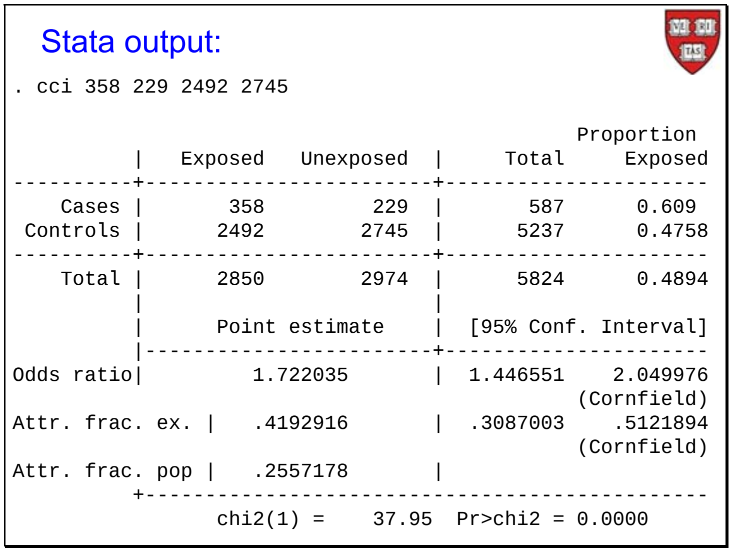
$(.542 - 1.96\text{x}.089\ ,\ .542{+}1.96\text{x}.089)$

$= (0.368, 0.716)$ is a 95% CI for ln(OR)

So (1.44 , 2.05) is a 95% CI for OR

Returning to our electronic fetal monitoring exposure and c-section study where we found that the chi-squared analysis had a small p-value. Approaching the same data using what we have learnt about odds ratios, we have the above analysis. The confidence interval for OR does not include one, so our inference here is in agreement with the chi-squared analysis. This analysis, though, is more informative because we quantify the dependence via the OR.

[2] If a cell entry is zero, just replace it with 0.5 for the sake of this formula.

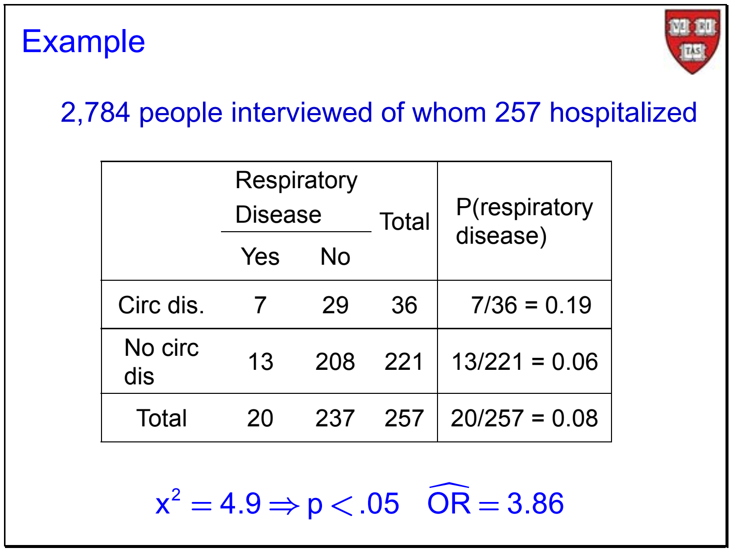


Stata output:

```
. cci 358 229 2492 2745

                                               Proportion
          |   Exposed   Unexposed  |      Total     Exposed
----------+------------------------+----------------------
    Cases |       358         229  |        587       0.609
 Controls |      2492        2745  |       5237      0.4758
----------+------------------------+----------------------
    Total |      2850        2974  |       5824      0.4894
          |                        |
          |      Point estimate    |  [95% Conf. Interval]
          |------------------------+----------------------
Odds ratio|         1.722035       |  1.446551    2.049976
                                                (Cornfield)
Attr. frac. ex. |   .4192916       |  .3087003     .5121894
                                                (Cornfield)
Attr. frac. pop |   .2557178       |
          +-----------------------------------------------
                 chi2(1) =    37.95  Pr>chi2 = 0.0000
```

The Stata output requires some studying, and I leave that to you to do.

Berkson’s Fallacy



Example

2,784 people interviewed of whom 257 hospitalized

	Respiratory Disease		Total	P(respiratory disease)
	Yes	No		
Circ dis.	7	29	36	7/36 = 0.19
No circ dis	13	208	221	13/221 = 0.06
Total	20	237	257	20/257 = 0.08

$$x^2 = 4.9 \Rightarrow p < .05 \quad \widehat{OR} = 3.86$$

We must include an example of what can go wrong with the study of discrete data; although this is certainly not peculiar to dichotomous data. This is an example of what is called Berkson's fallacy. Berkson is the one who quantified the fallacy, not commit the error. It is an example of sampling error, and it is the same error committed by Ray Pearl, about 100 years ago. He noticed from autopsy records that very rarely did people die of both cancer and tuberculosis. So he decided that the tuberculosis must be a protector against cancer. So as a treatment for the cancer patients, he proposed to inject patients with the tuberculin.

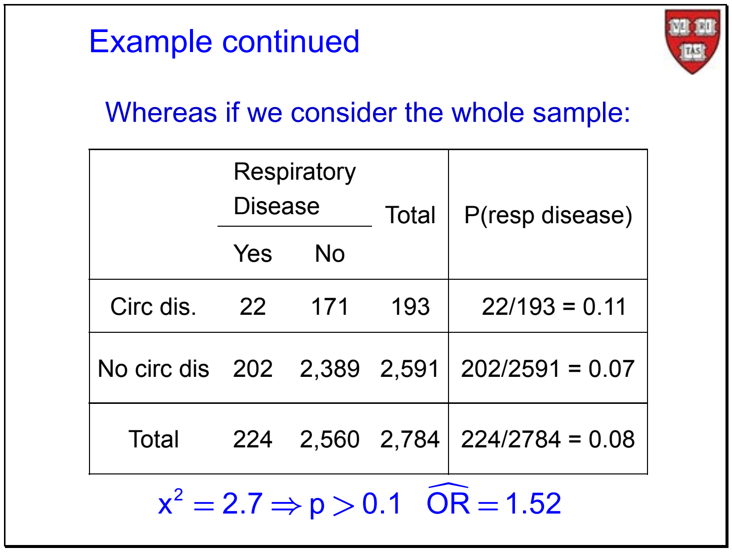
This was before the days of informed consent, and he actually did treat one patient thusly. The patient passed away, and he was about to treat a second patient when the authorities made him put a stop to his study. Apart from the obvious ethical issues involved judging by our current sensitivities, the question we need to ask is who gets autopsied? Does this pool of autopsied patients represent a random sample of dead patients? If not, is it then correct to infer what would happen to general patients from autopsied patients?

Consider the study above where 2,784 people were interviewed. Of those, 257 were hospitalized. So let us look at those 257: each was classified as either having a circulatory disease or not, and whether they had a respiratory disease or not. We see the resultant 2x2 table above.

When we subject this table to a chi-squared analysis we see that the X^2 value is 4.9, the p-value is less than 0.05, and we reject the null hypothesis that there is no relationship between the row and column classifiers. In other words, this sure looks like there is a relationship between respiratory and circulatory ailments. Indeed, the odds ratio is a sizable 3.86.

We can understand what is happening if we look at the people with circulatory disease, seven of them had respiratory disease; about 19%. On the other hand, who had no circulatory disease, the probability of respiratory disease is 6%. That explains the estimated odds ratio.

But let us return to the original 2,784 who were interviewed. Who were these 287 that we chose of those 2,784? In a sense these were the sickest patients. These were the ones who were hospitalized. They certainly were not a random subset.



Example continued

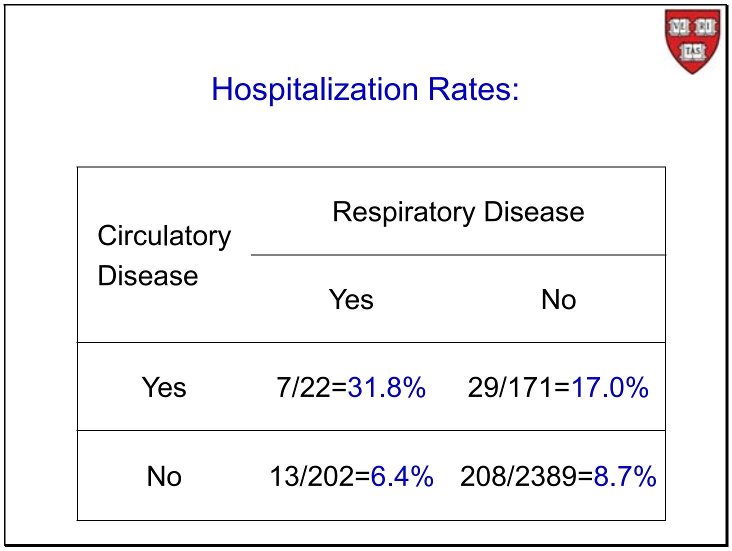
Whereas if we consider the whole sample:

	Respiratory Disease		Total	P(resp disease)
	Yes	No		
Circ dis.	22	171	193	22/193 = 0.11
No circ dis	202	2,389	2,591	202/2591 = 0.07
Total	224	2,560	2,784	224/2784 = 0.08

$$x^2 = 2.7 \Rightarrow p > 0.1 \quad \widehat{OR} = 1.52$$

Let us return to the whole sample only to find that this time the chi-squared analysis does not lead to the rejection of the null hypothesis and the odds ratio estimate is now reduced to 1,52. So on the basis of the 2,784 patients we conclude that there is no relationship no relationship between respiratory and circulatory diseases.

So what happened? What happened is we looked at the sub sample. Those who were more sick than the rest.



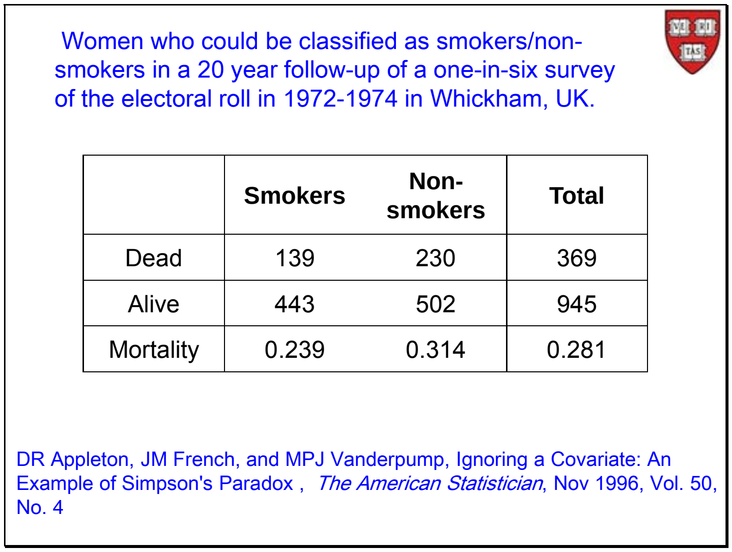
Hospitalization Rates:

Circulatory Disease	Respiratory Disease	
	Yes	No
Yes	7/22=31.8%	29/171=17.0%
No	13/202=6.4%	208/2389=8.7%

Indeed, if we look at the hospitalization rates and how they vary with respect to the classifications we were interested in studying, we see a large discrepancy with an over representation of the group that provided the impetus for the results we saw.

Thus we got a distorted view of the relationship between respiratory diseases and circulatory diseases. Thus Berkson's fallacy results from not getting a representative sample.

Yule Effect—Simpson's Paradox



Women who could be classified as smokers/non-smokers in a 20 year follow-up of a one-in-six survey of the electoral roll in 1972-1974 in Whickham, UK.

	Smokers	**Non-smokers**	**Total**
Dead	139	230	369
Alive	443	502	945
Mortality	0.239	0.314	0.281

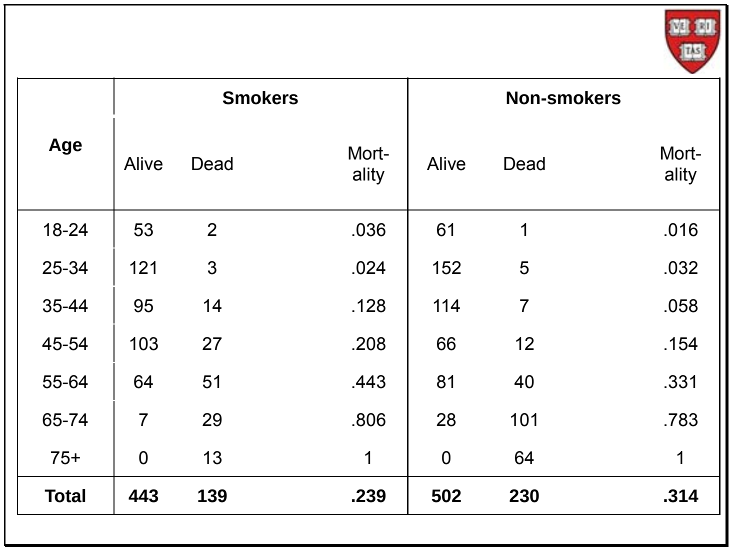
DR Appleton, JM French, and MPJ Vanderpump, Ignoring a Covariate: An Example of Simpson's Paradox , *The American Statistician*, Nov 1996, Vol. 50, No. 4

One last topic: I cannot leave contingency tables without showing you this topic. This is something that is called the Yule effect, or, Simpson's Paradox, and that is because Simpson wrote about this about 50 years after Yule, so it makes sense that we call it the Simpson Paradox.

Let me introduce it by example. The researchers chose one in six persons in the electoral roll between 1972 and 1974, in Wickham, in the UK. Let us focus on the women in the sample, because that is the group the study reported.

Each woman was classified as a smoker or a non-smoker, and then they returned 20 years later to do a follow up, and saw who had survived and who had passed away. What they discovered is shown above, where we see that the mortality rate amongst smokers is 0.239, whereas amongst the non-smokers it is a higher 0.314.

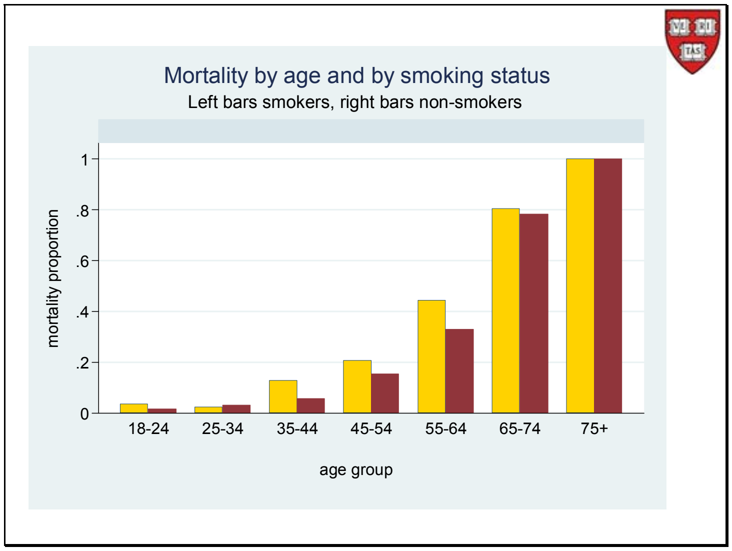
What is going on here? This is very odd. Before you run off and buy yourself some cigarettes, let us look a little more closely and see what is going on here.



Age	**Smokers**			**Non-smokers**		
	Alive	Dead	Mort-ality	Alive	Dead	Mort-ality
18-24	53	2	.036	61	1	.016
25-34	121	3	.024	152	5	.032
35-44	95	14	.128	114	7	.058
45-54	103	27	.208	66	12	.154
55-64	64	51	.443	81	40	.331
65-74	7	29	.806	28	101	.783
75+	0	13	1	0	64	1
Total	**443**	**139**	**.239**	**502**	**230**	**.314**

Let us break down the two groups into age categories because we learnt that the mean can hide a lot of sins and the composition formula warned us of what can happen when comparing two different groups.

If we look at each of these age groups, we see that, except for the 25-34 age group, the mortality rates for the smokers are higher than the non-smokers. In fact, in the 25-34 age group if there had been one more death amongst the smokers, then the mortality in that group would also fall in line.

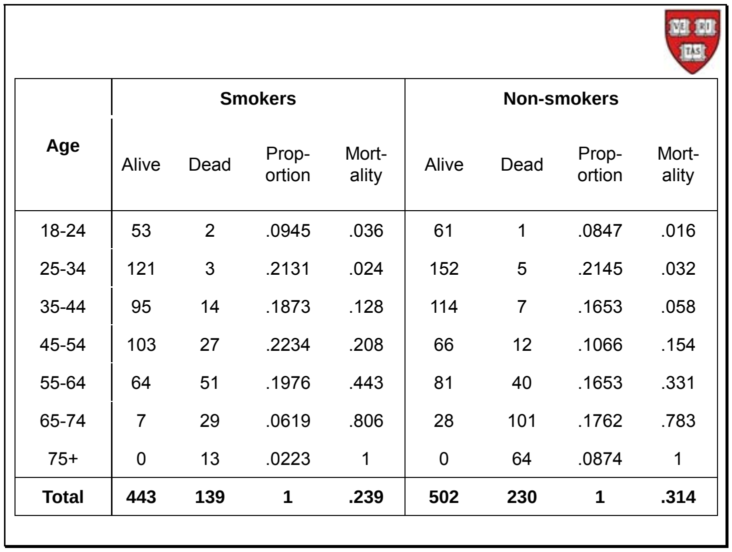


Mortality by age and by smoking status

Left bars smokers, right bars non-smokers

The point is well made in this graphic. We see that the yellow bars are bigger than the red bars, except in the second age group. But that has such a small mortality that it is unlikely that could be driving the overall average being higher for the red group than for the yellows.

Having learnt from the composition formula we look for the makeup of the two groups of women.



Age	**Smokers**				**Non-smokers**			
	Alive	Dead	Prop-ortion	Mort-ality	Alive	Dead	Prop-ortion	Mort-ality
18-24	53	2	.0945	.036	61	1	.0847	.016
25-34	121	3	.2131	.024	152	5	.2145	.032
35-44	95	14	.1873	.128	114	7	.1653	.058
45-54	103	27	.2234	.208	66	12	.1066	.154
55-64	64	51	.1976	.443	81	40	.1653	.331
65-74	7	29	.0619	.806	28	101	.1762	.783
75+	0	13	.0223	1	0	64	.0874	1
Total	**443**	**139**	**1**	**.239**	**502**	**230**	**1**	**.314**

When we look at the composition of the two groups we see that the smokers are younger than the non-smokers: the proportions in the in the under 64 age groups are all higher for the smokers, except for the 25-34 age group which is larger in the third decimal place for the non-smokers. The older age groups, the 65 and over, are where the non-smokers have higher representation than the smokers.