1. Boy/girl ratios

(a) $z = \dfrac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \dfrac{0.4846 - 0.5}{\sqrt{0.5(1-0.5)/43{,}200}} = -6.38$, p-value is less than 0.0001.

There is strong evidence that π is not 0.5.

(b) Approximate 95% confidence interval for π is

$$p \pm z_{0.975}\sqrt{\frac{p(1-p)}{n}} = 0.4846 \pm 1.96\sqrt{\frac{0.4846(1-0.4846)}{43{,}200}} = (0.480, 0.489)$$

where $p = 20{,}937/43{,}200 = 0.4846$.

We are 95% confident that the proportion of females is between 0.480 and 0.489.

(c) $(0.4846 - 0.4812) \pm 1.96\sqrt{\dfrac{0.4846(1-0.4846)}{43{,}200} + \dfrac{0.4812(1-0.4812)}{7{,}344}} = (-.009, 0.016)$

(d) No. The tests and confidence intervals assume that events (male or female child) are independent. Alternatively, each child has the same probability of being female. If that probability varies between families, the events aren't independent.

2. Distribution of counts of boys and girls

(a) Let $X \sim Binomial(n = 6, \pi = 0.485)$. Expected # with 0 boys = expected # with 6 girls = (6!/(6! 0!) $\pi^6(1-\pi)^0$ = 93.71 if you use π=0.485.

(b) $H_0$ : proposed model $Binomial(n = 6, \pi = 0.485)$ is correct . I give you the contributions to Chi-square for most of the cells. For the 6 girl cell, $(E-O)^2/E = (93.71 - 113)^2/ 93.71 = 3.971$. I gave you the sum for the rest of the cells, so Chi-square = 3.971 + 12.894 = 16.866. This has 6 d.f., so

$$p - value = P(\chi^2_6 > 16.866) = 0.0098$$

There is strong evidence to reject the null hypothesis, the data do not follow the proposed binomial distribution.

* The Chi-square statistic has 6 d.f. because there are 7 counts and one constraint (total observed = 7200), so 7-1=6 d.f..

3. Case-control study of breast cancer and drinking

(a) This could be done by computing a z statistic or using the Chi-square test. The Chi-square statistic is 4.19 with a p-value of 0.040. There is evidence of an association between breast cancer and drinking.

(b) The odds ratio is $\dfrac{62/147}{97/153} = 0.665$

(c) The s.e. of the log odds ratio is estimated as $\sqrt{\frac{1}{97}+\frac{1}{62}+\frac{1}{153}+\frac{1}{147}} = 0.199$

(d) The 95% ci for the log odds ratio is log(0.665) +/- 1.96*0.199 = ( -0.799, -0.017), so the 95% ci for the odds ratio is ( exp(-0.799), exp(-0.017) ) = (0.45, 0.98).

* My SAS code, which gets everything except the se is:

```
data bc;
  input cancer $ drink $ n;
cards;
Yes L 97
Yes H 62
No L 153
No H 147
;

proc freq;
  table cancer*drink / chisq measures cl;
  weight n;
  title 'Stat 500: HW 6, problem 3';
  run;
```

Note that SAS reports the odds of heavy drinking in the 'No' / odds of heavy drinking in 'Yes', which are the reciprocals of the desired answers. The measures option produces the s.e.; the cl option produces the confidence limits (i.e. the confidence intervals). The weight statement indicates how many individuals had that combination of characteristics; if this were omitted you would need one line of data for each individual, a total of 459 lines.

4. Comparison of case-control and random sampling:

(a) Incidence of breast cancer = 159/459 = 0.35 = 35%

(b) This is very different because the case-control data provide no information about the incidence of breast cancer. The relative proportions of cases and controls are determined by the investigators.

(c) The incidences in each subgroup are: 97/250 = 0.388 =38.8% in light drinkers and 62/209 = 0.297 = 29.7% in heavy drinkers. The difference is 0.091 = 9.1%

(d) The incidences of heavy drinking in each subgroup are: 62/159 = 39.0% in the cases and 147/300 = 49.0% in the controls. The difference is 10.0%

(e) The s.e. of the difference in incidence of heavy drinking is $\sqrt{\frac{0.390(1-0.390)}{159}+\frac{0.490(1-0.490)}{300}} = 0.048 = 4.8\%$

(f) Even when the case-control study provides a valid estimate of some parameter, the values of the two estimates (from the case-control and from the random sample) will not be exactly the same, because of discreteness (people only come in whole units) and sampling variation. However, if the

two estimates are close, the case-control study is probably providing a valid estimate; if the two estimates are quite different, the case-control study is not. We looked at:

Estimates of:
Incidence of breast cancer: quite different
Incidence of b.c. in subgroups and their difference: quite different
Incidence of heavy drinking in subgroups and their difference: very similar
Odds ratio: very similar

So, a case-control design provides estimates of the odds ratio, of the subgroup estimates of the 'environmental variable', here drinking, and of the difference in incidence of drinking.

(g) The standard errors (e.g. of differences or of the log odds ratio) from the case-control study of 459 people are a bit larger than those from the random sample of 8499 people, but those from the case-control study of 8500 people are considerably smaller still.

Advantages:
Better precision (smaller s.e.) for the same number of people for estimable quantities, including the odds ratio.
Or in other words, need many fewer people in a case-control study to get the same precision.

Disadvantages:
Can not estimate certain quantities (e.g. incidence of the rare event).

* A case-control study of 660 people (ca 430 controls and 230 cases) will provide the same s.e. of the log odds ratio as the random sample of 8500 people.