Data Analysis and Statistical Methods Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 25 (MWF) Tests based on proportions (one sample)

Suhasini Subba Rao

The Binomial Distribution

- The focus of this lecture is on categorical data where the outcomes are binary (just two outcomes).

- We will assume that we observe n binary variables all of which are independent of each of other and each has the sample probability of for a success.

- For example if the proportion of public who like chocolate is 70%. A sample of 100 people were taken and each asked if they liked chocolate, then the answer is yes or no and the probability of a yes is 0.7.

- The number of successes in such a sample follows a binomial distribution with $\text{Bin}(n, p)$. Refer to Lecture 7 for a summary.

- We will use this in both testing and constructing confidence intervals.

Example 1:

- Suppose a midterm is a multiple choice. There are five different answers for each question, where only one can be correct.

  If one guesses each answer, there is a 20% chance of getting the correct answer. If there are 15 questions, the number of answers one can correctly guess follows a Binomial(15,0.2). A plot of the distribution is given below.
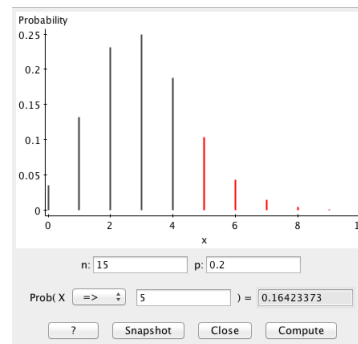
- Mike scores 5 out of 15. Does his grade suggest that he had put some effort into answering the questions correctly?

- We can articulate this as a hypothesis test. Let p denote the probability of Mike scoring a question correctly (we assume all answers are independent of each other). If $p > 0.2$, than some knowledge has been put into answering the question.

- We test $H_0 : p \leq 0.2$ against $H_A : p > 0.2$. In the exam, Mike scores 5 out of 15. How likely is it for Mike to get 5 or more correct by just randomly guessing?

- From the previous page we see this is 16.4% (this is the p-value), as this is quite large (larger than the 5% significance level), we cannot reject the null. There is nothing in his examination grade to suggest he knew the material (though we can not know for sure just from his grade).

- We used the binomial distribution in the Distribution App.

- But we can simply make JMP do the hard work.

- For JMP requires the raw data, on a spreadsheet not just 5 out of 15. See the next page for details.

The test in JMP

- On the left we see whether the student got each answer right (R) or wrong (W). Ensure this is set as a categorical variable.

- In JMP go to Analyze > Distribution. Place the variable of interest into Y, columns.

- You will see a bar chart. Click on the red triangle next to the variable (in this case grade) and choose Test Probabilities.

- Place the probability under the null in the box. In this case we place 0.2 next to R (which denotes right).

One and Two sided options in JMP

We take a closer look at the testing window.

Observe there are three options. We describe each below.

The **second option** says greater than hypothesized. This means $H_0 : p \leq 0.2$ vs $H_A : p > 0.2$.

It uses binomal distribution and exactly the method we described previously. Of course, this gives p-value derived above. Recall we cannot reject the null.

The third option says less than hypothesized. This means $H_0 : p \geq 0.2$ vs $H_A : p < 0.2$. This one uses the exact binomal distribution. It gives p-value using the binomial but for this test, it is the probability of scoring 5 or less.

▼ Test Probabilities

Level	Estim Prob	Hypoth Prob	
R	0.33333	0.20000	
W	0.66667	0.80000	

	Level	Hypoth	
Binomial Test	**Tested**	**Prob (p1)**	**p-Value**
Ha: Prob(p > p1)	R	0.20000	0.1642

▼ Test Probabilities

Level	Estim Prob	Hypoth Prob	
R	0.33333	0.20000	
W	0.66667	0.80000	

	Level	Hypoth	
Binomial Test	**Tested**	**Prob (p1)**	**p-Value**
Ha: Prob(p < p1)	R	0.20000	0.9389

The first option is a two sided test. This means $H_0 : p = 0.2$ vs $H_A : p \neq 0.2$.
Next to it is says chi-squared test. This test despite the name, simply uses the normal distribution (chi-square is simply the square of the normal). This is an approximation of the distribution, which we justify below.

▼ Test Probabilities

Level	Estim Prob	Hypoth Prob		
R	0.33333	0.20000		
W	0.66667	0.80000		

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	1.4618	1	0.2266
Pearson	1.6667	1	0.1967

- The two-sided test assumes that the binomial distribution is close to normal. So the accuracy of this p-value is based on how close the following binomial distribution is to normal:
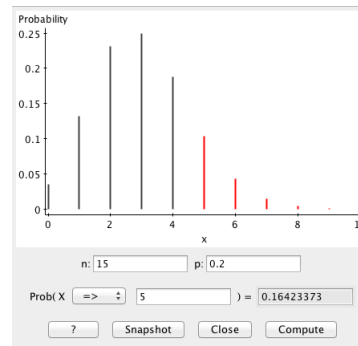
- For this example, it is clear that it not the case (and this is the main reason one gets two different p-values for two different statistical tests). This means we have to take these p-values with a pinch of salt. However, for larger sample sizes it will be closer to the truth. We show this below.

- In this class the focus will be on the value given for "Pearson".

Scores in an Exam: Example 2

- The exam was multiple choice, again with each question having a choice of 5 possible answers.

- A student scores 33 out of 100. We test $H_0 : p \leq 0.2$ against $H_A : p > 0.2$. Using the data we need to calculate the probability of scoring 33 or more when he is randomly guessing.
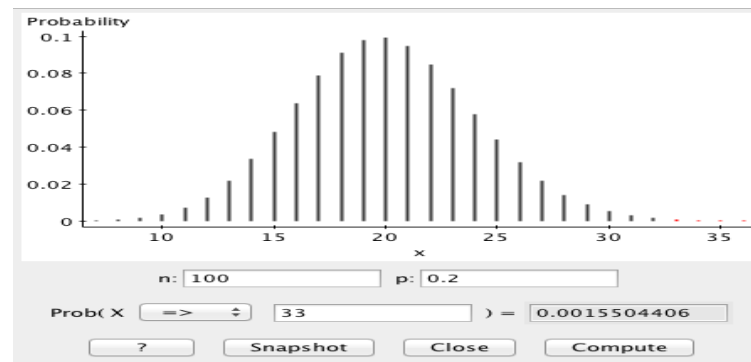
- This gives gives the p-value $= 0.15\%$. Since $0.15\% < 5\%$, this strongly suggests suggests Rick may not have been guessing and had used some of knowledge in obtaining that score that he did.

- As we mentioned above, the one-sided case will always give the correct p-value for binary data, since it uses the correct distribution to obtain the p-value.

- But suppose our aim is to test $H_0 : p = 0.2$ vs $H_A : p \neq 0.2$. This means testing to see if the student put some effort into the exam (including the case of trying to get answers wrong!).

- However, for a two-sided test we have to use the normal distribution, which means the binomial distribution should be sufficiently close to the normal distribution for the p-value to be meaningful.

- We observe from the plot of the binomial distribution that it does look quite normal (the sample size $n = 100$ is quite large and $p = 0.2$ is not too small for the skew to dominate).

- The JMP test is

Test Probabilities		
Level	Estim Prob	Hypoth Prob
R	0.33000	0.20000
W	0.67000	0.80000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	9.2884	1	0.0023*
Pearson	10.5625	1	0.0012*

This means the p-value corresponding to the Pearson test 0.12% is sufficiently close to the true people value to reject the null at the 5% level. Thus, based on the grade there is evidence to suggest that the student was not guessing.

Polls of Same-sex marriage

A recent Gallup poll found that 64.2% (90 people) of a sample of 140 individuals were favourable to same-sex marriage.

These are some news headlines, are they accurate?

- The New York Times: Survey suggests that the majority of the general public a favourable to same-sex marriage.

- The Wall Street Journal: Poll suggest that **over** 60% of the population are favourable to same-sex marriage.

Are these reports accurate? To answer this question we write them as a hypothesis tests.

12

The New York Time reporting

- Majority means a proportion is over 50%. This means testing $H_0 : p \leq 0.5$ against $H_A : p > 0.5$.

- The Binomial is below:
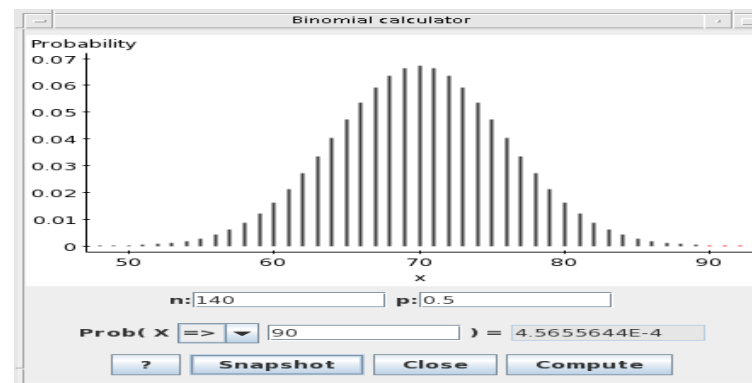
From the plot we see that the probability is 0.045%. This probability tells is that the chance of obtaining a sample where 90 or more people

out of 140 say they are favourable to same sex marriage when public opinion is equally divided is 0.045%, as this is very small (and far below the 5% significance level) we can reject the null.

- Interpretation The data suggests that public opinion is in favour of same-sex marriage, and the New York Times has accurately reported the findings.

The Wall Street Journal reporting

- Based on the data is there evidence that **over** 60% of the population support same-sex marriage, this corresponds to testing $H_0 : p \leq 0.6$ against $H_A : p > 0.6$.

- We calculate the chance that in a sample of 140 people, 90 or more say they are in favour given that the proportion of the public that suppose same-sex marriage is only 60%. We use the binomial distribution
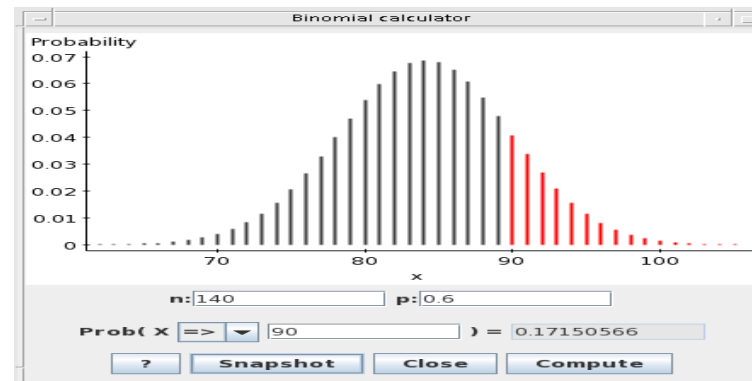
From the plot we see that the probability is 17.1%. This probability tells is that the chance that 90 or more people out of 140 say they are in favour of same-sex marriage when public is opinion is divided 60:40 is 17.1%.

- As 17.1% is relatively large (over the 5%), the data could easily come from a population where public opinion is 60%. There is no evidence to reject the null.

- Interpretation Therefore even though in the sample 64.2% were in favour of same-sex marriage, this does not prove that the proportion in the public who support same sex marriage is over 60%.

Why bother with the normal approximation?

- As we illustrated above, obtaining the exact p-values by using the true, Binomial distribution, is very useful.

- However, we require the normal approximation for

  - Two-sided tests
  - To construct a confidence interval for the true proportion.

- Consider the example of a two-party election. In an opinion poll 44.11% out of 1000 people that are randomly sampled said they were would vote for party A.

- We want to use 44.11% to construct an interval where we reliably believe the true proportion in the public will lie. This is a confidence interval.

The normal approximation to the binomial distribution

- Let S_n denote the number of "successes" out of n and $\widehat{p} = S_n/n$ (the proportion of "successes" in the sample). The CLT is

$$\widehat{p} \sim \mathcal{N}\left(p, \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{standard error}}\right),$$

  where p denotes the true proportion.

- You saw the 'normality' happen in the above slides.

- Like all estimators as n grows, the standard error becomes smaller and the estimator improves.

- General practical recommendations for CLT to roughly hold The number of "successes" and "failures" in a sample is greater than 5.

  This assumption essentially ensures that the skew in the corresponding normal distribution not too large.

Testing using the normal approximation

- We test $H_0 : p = p_0$ against the alternative $H_A : p \neq p_0$ (we can also do one-sided tests using the normal approximation i.e. $H_0 : p \leq p_0$ against $H_A : p > p_0$), here p_0 is some given value such as $0.2, 0.5$ or 0.6 (in the previous example).

- Under the null the standard error is $\sqrt{p_0(1 - p_0)/n}$. For example, in the case we test $H_0 : p = 0.2$ against $H_A : p \neq 0.2$ we use the standard error $\sqrt{0.2 \times 0.8/n}$.

- To do the test we make the z-transform

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}},$$

and look up the z-tables according to the correct hypothesis.

Example 1 (testing)

Elections are due to take place next week. Ernie and Bert are the two candidates. So far the polling has been neck and neck. Is any of the candidates going to get the majority.

In a recent poll of 1000 monsters, 44.1% of the monsters said they would be vote for Ernie.

Using this data, is there any evidence to suggest that one candidate will get the majority?

Solution 1

- Based on the above we need to test $H_0 : p = 0.5$ vs $H_A : p \neq 0.5$.

- Using the formula the standard error $= \sqrt{0.5 \times 0.5/1000} = 0.016$.

The z-score is

$$z = \frac{0.441 - 0.5}{0.016} = -3.7.$$

Using this approximation we obtain the p-value
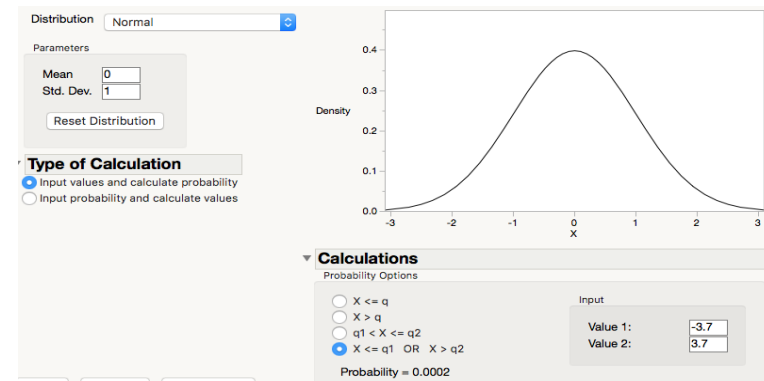
- The p-value is the sum of area to the left of -3.7 and right of 3.7. This is 0.02%. This is substantially smaller than 5%, thus there is strong evidence to suggest that one of the candidates will win the majority.

Election

Frequencies

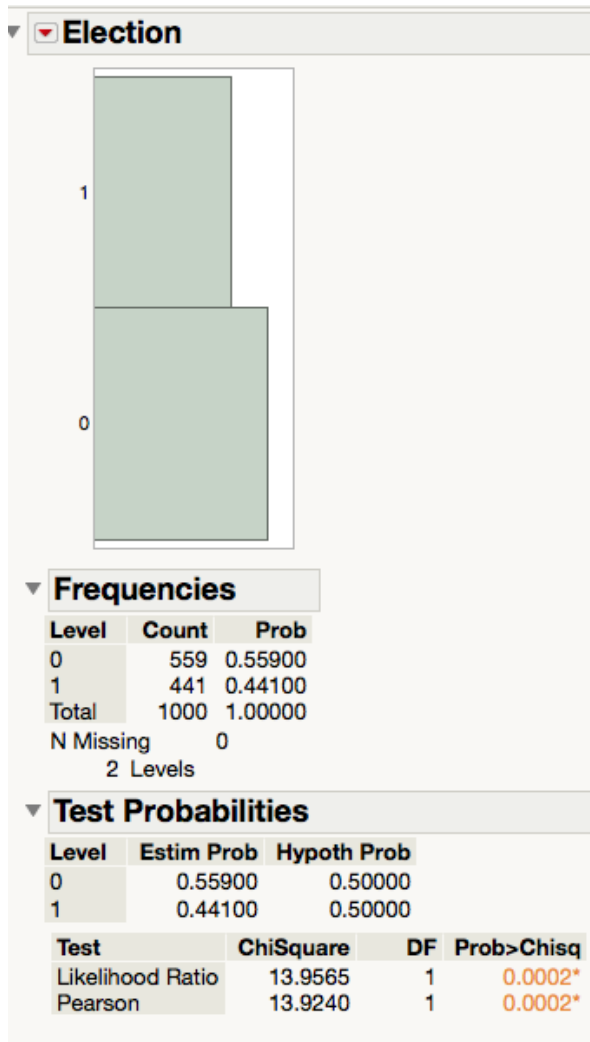| Level	Count	Prob
0	559	0.55900
1	441	0.44100
Total	1000	1.00000
N Missing	0	
2 Levels		

Test Probabilities

Level	Estim Prob	Hypoth Prob
0	0.55900	0.50000
1	0.44100	0.50000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	13.9565	1	0.0002*
Pearson	13.9240	1	0.0002*

In JMP

Just as we calculated we observe that the p-value for the Pearson test to be 0.02%. It coincides with the Likelihood ratio because the sample size $n = 1000$ (and $p = 0.5$, which means the binomial is symmetric) is sufficiently large for the normality assumption to hold.

In some of the old exams I have used $\widehat{p}$ instead of p when doing the test. Please don't use this.

Confidence intervals

- Using the normality result $\widehat{p} \sim \mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}})$ the 95% confidence interval for the true proportion is

$$\left[\widehat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}, \widehat{p} + 1.96\sqrt{\frac{p(1-p)}{n}} \right],$$

- Of course p is unknown which is why we are constructing a CI for it. Thus the standard error $\sqrt{\frac{p(1-p)}{n}}$ is unknown (n is known but not p). Instead we replace p with its estimator $\widehat{p}$ to give the estimated standard error $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$.

- This yields the confidence interval

$$\left[\widehat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \widehat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right].$$

- Observe that we **do not** use the t-distribution when dealing with proportions.

JMP: confidence intervals for one proportion
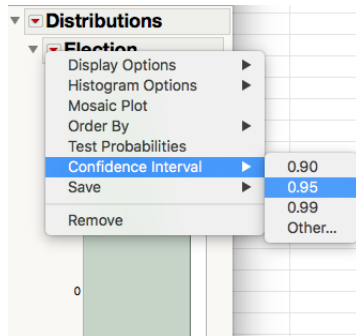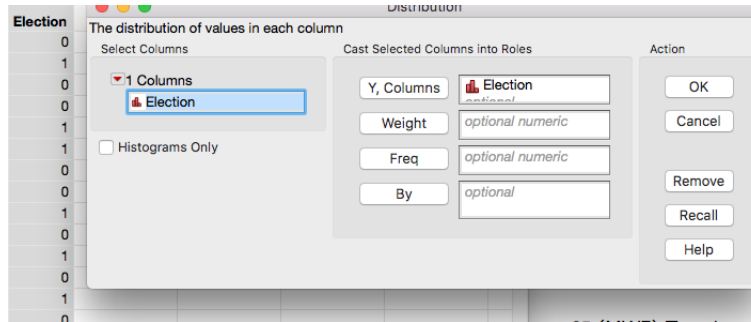
You need the raw data. Go To Analyze > Distribution. Ensure your variable is categorical (red symbol on left). You should obtain the output below:

2 Levels

Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
0	559	0.55900	0.528059	0.589489	0.950
1	441	0.44100	0.410511	0.471941	0.950
Total	1000				

Example 2 (confidence intervals)

Elections are due to take place next week. Ernie and Bert are the two candidates.

In a recent poll of 1000 monsters, 44.1% of the monsters said they would be vote for Ernie.

Construct a confidence interval for the proportion of people who will vote for Ernie.

Solution 2

- Since p is unknown we use its estimate $\widehat{p} = 0.441$ to estimate the standard standard error.

- The estimated standard error is

$$\sqrt{\frac{0.441 \times 0.559}{1000}} = 0.0157$$

  Note this is different to the standard error used for the test, since do not have a candidate value for p.

- The 95% confidence interval is

$$[0.441 - 1.96 \times 0.0157, 0.441 + 1.96 \times 0.0157] = [0.41, 0.47].$$
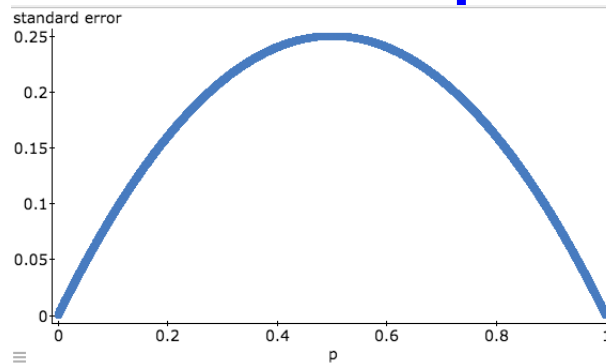
- In other words, the proportion of the monsters likely to vote for Ernie is 44.1% with a margin of error 3%.

- This is exactly what is given in the JMP output

2 Levels

▼ ▾ Confidence Intervals					
Level	Count	Prob	Lower CI	Upper CI	1-Alpha
0	559	0.55900	0.528059	0.589489	0.950
1	441	0.44100	0.410511	0.471941	0.950
Total	1000				

Standard Errors - fixed sample size different p

- The above is a plot of how the standard error of an estimator changes with the true proportion.

- When the true proportion is $p = 0\%$, the standard error is zero.

  If the entire population hates chocolate and a sample of 10 people is taken. We know for sure that noone in the sample will like chocolate. There is no sampling variability. If there is no difference between samples, the standard error is zero.

30

- When the true proportion is $p = 100\%$, the standard error is zero.

  If the entire population loves chocolate and a sample of 10 people is taken. We know for sure that everyone in the sample will like chocolate. There is no sampling variability. If there is no difference between samples, the standard error is zero.

- Usually p lies somewhere between $0 - 100$ and the standard error varies accordingly.

- The standard error is maximum when $p = 0.5$.

  This means if the entire population is split 50:50 on a issue. For any sample of say 100 people, the opinions will be varied. For example, one group of a 100 may be split 60:40 whereas another may split 40:60 (the other way).

Determining the sample size

- Let us return to the example of Ernie and Bert. The 95% confidence interval for those who will vote for Ernie is

$$[0.441 \pm 0.03] = [44.1 \pm 3]\% = [41.1, 47.1]\%$$

- The Margin of Error is $0.03 = 1.96\sqrt{\frac{0.441 \times 0.559}{1000}}(= 3\%)$.

- In general the Margin of Error of the 95% confidence interval is

$$MoE = 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Two factors determine how large the margin of error is (i) the sample size n and (ii) the proportion p.

- In general, the margin of error for a $100(1 - \alpha)\%$ CI is

$$MoE = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

- To obtain the sample size for a given MoE we use the rearrange the above formula

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{MoE^2}.$$

For example, if we want to construct a 95% CI, then we use $z_{\alpha/2} = 1.96$ and

$$n = \frac{1.96^2 \times p(1-p)}{MoE^2}.$$

- However p is unknown. But we have shown that when the sample size is fixed that the confidence interval is largest when $p = 1/2$.

- Thus as in Lecture 13, we choose the p (in Lecture 13 we chose the σ) which maximized the mean squared error in the margin of error calculation, which is p

$$n = \frac{1.96^2 \times 0.5^2}{MoE^2}.$$

This will ensure that the true margin of error cannot be larger than the pre-specified margin of error (regardless of the true value of p).

Example 1

- What sample size should we use in order to achieve a margin of error of **no more** than 2% for a 90% confidence interval for the proportion of arthritis patients taking ibuprofen who suffer severe side effects?

- Solution Since the standard error is maximised at $p = 0.5$ we replace p with 0.5 and solve the equation

$$n = \left(\frac{1.64^2 \times 0.5 \times 0.5}{0.02^2} \right) = 1681.$$

This means that a large number of participants have to be involved in the study.

- However, for medication usually the proportion who experience a reaction tends to be far less than 50%. Often we can place an upper bound on the proportion.

Example 2

- Suppose we **know** that the proportion of the population who get a side effect cannot be greater than 15%. What sample size should we choose to ensure that the MoE is 2%, but the sample size is not as large?

- Solution Since we know that the true p is such that $0 \leq p \leq 0.15$, we should include this in the sample size calculation as this will reduce the sample size required.

- We need to use in the margin of error calculation the p which maximises the standard error, but we also know that $p \leq 0.15$.

We use the plot below.
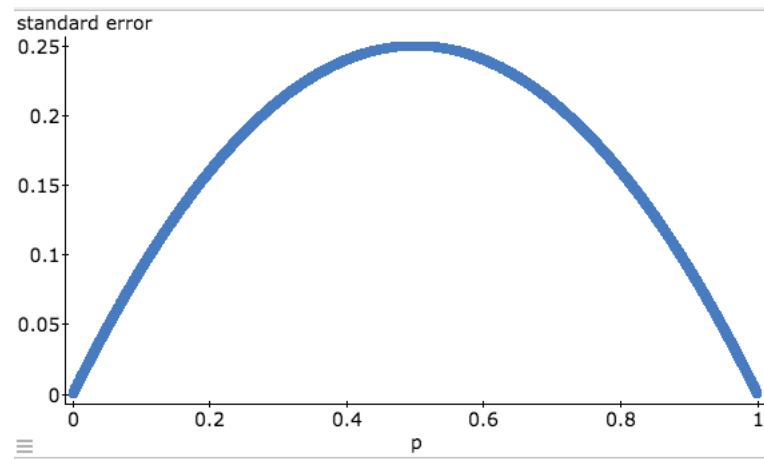
Using the plot we see that margin of error (for a given sample size) is maximised at 0.15 (if the proportion is restricted to less than 0.15). Using this information the sample size should be:

$$n = \left(\frac{1.64^2 \times 0.15 \times 0.85}{0.02^2} \right) = 858.$$

- Compared with the previous sample size of 1681, we observe that we have managed to substantially reduce the sample size when we use prior information about the proportion.

- Comment If we know that the true proportion is greater than or equal to 0.85, then using the same argument the sample size that we should use to be sure that the $MoE < 0.02$ is

$$n = \left(\frac{1.64^2 \times 0.15 \times 0.85}{0.02^2} \right) = 858.$$