Empirical Probabilities – an Introduction to Statistics
Math 111 Notes

Often in the news you will read such statements as 64% of the people believe such and such or 55% of the voting age population support some candidate. These are not exact figures, but they are estimates obtained by polling some segment of the population. They are examples of empirical probabilities (see page 61 of Tan). If someone asks 200 voters which candidate they prefer and 110 prefer Senator Small, then you figure the probability anyone will vote for Senator Small is 110/200 or 55%. But because of random variations this 55% might not be correct.

Example: Suppose 60% of the voting age population of a state prefer Senator Small in the next election. You ask 200 randomly chosen voting age people in the state which candidate they prefer. What is the probability that 55% or fewer of the people you ask support Senator Small?

Let the random variable X be the number of people you ask who support Senator Small. Let the random variable Y be the proportion of the people you ask who support Senator Small. Then $Y = X/200$. Now X is a binomial random variable which we may approximate by the normal random variable

$$X \approx 200(.6) + \sqrt{200(.6)(.4)}Z = 120 + 6.928Z$$

But then Y is also approximately a normal random variable

$$Y = X/200 \approx \frac{120 + 6.928Z}{200} = .6 + .03464Z$$

So

$$\begin{aligned} P(Y < .55) &\approx P(.6 + .03464Z < .55) \\ &= P(.03464Z < .55 - .6) \\ &= P(Z < \frac{.55 - .6}{.03464}) \\ &= P(Z < -1.44) = .0749 \end{aligned}$$

So there is about a 7.5% chance your poll will claim 55% or fewer voters support the Senator.

In general then, suppose you do n trials of a binomial experiment with probability of success p. Let the random variable Y be the proportion of successes. Then

$$Y \approx \frac{np + \sqrt{npq}Z}{n} = p + \sqrt{\frac{pq}{n}}Z$$

So Y is approximately a normal random variable with mean p and standard deviation $\sqrt{pq/n}$.

The problem with the above election example is that our calculations involved the actual probability of success $p = .6$. In real life, you would not know this. If you did, there would be no need to poll the voters. So if you want to get some idea of how accurate your empirical probability is we must do more.

For example, suppose you have a coin and wish to estimate the probability it will come up heads when tossed. You can test this by flipping it a large number of times and seeing how often it comes up heads. But your results are subject to random variation. How many times must you flip the coin to be reasonably sure your empirical probability of getting heads is close to the actual probability?

Suppose you toss the coin n times and it comes up heads h of those n times. Then your empirical probability of getting heads is h/n. This is most likely different from the actual probability p of getting heads. Let the random variable Y be your empirical probability, $Y = h/n$. Then the probability that Y is within .01 of the actual probability p is

$$P(p - .01 < Y < p + .01)$$

We saw above that Y is very nearly a normal random variable with mean p and standard deviation $\sqrt{pq/n}$. So we may approximate Y by the normal random variable

$$Y \approx p + \sqrt{pq/n}Z$$

Consequently we have

$$\begin{aligned} P(p - .01 < Y < p + .01) &\approx P(p - .01 < p + \sqrt{pq/n}Z < p + .01) \\ &= P(-.01 < \sqrt{pq/n}Z < .01) \\ &= P(\frac{-.01}{\sqrt{pq/n}} < Z < \frac{.01}{\sqrt{pq/n}}) \end{aligned}$$

If we want to be 95% certain that our empirical probability Y is within .01 of p then we want

$$P(p - .01 < Y < p + .01) \geq .95$$

which means that

$$P(\frac{-.01}{\sqrt{pq/n}} < Z < \frac{.01}{\sqrt{pq/n}}) \geq .95$$

Just as in Example 2c on page 180 of Tan, we get

$$2P(Z < \frac{.01}{\sqrt{pq/n}}) - 1 \geq .95$$

and so

$$P(Z < \frac{.01}{\sqrt{pq/n}}) \geq (.95 + 1)/2 = .975$$

Looking up the table of the normal distribution we see that this is true if

$$\frac{.01}{\sqrt{pq/n}} \geq 1.96$$

Multiplying by $\sqrt{pq/n}$ we get

$$.01 \geq 1.96\sqrt{pq/n}$$

Squaring both sides we get

$$(.01)^2 \geq (1.96)^2 pq/n$$

We can now solve for n,

$$n \geq (1.96)^2 pq/(.01)^2$$

We do not know what pq is, but it can be shown that $pq \leq .25$ no matter what p and $q = 1 - p$ are†. Consequently we may take

$$n \geq (1.96)^2(.25)/(.01)^2 = 9604$$

So if you flip the coin 9604 or more times you can be 95% certain that your empirical probability of getting heads is within .01 of the actual probability.

† To see this, let $r = p - .5$. Then $p = .5 + r$ and $q = .5 - r$. So $pq = (.5 + r)(.5 - r) = .25 - r^2$. Since $r^2 \geq 0$ we know that $pq \leq .25$.

Example: Look at Table 2.2 on page 61 of Tan. In 10,000 flips there were 5,034 heads which gives an empirical probability $Y = .5034$. So after these 10,000 coin flips you are 95% sure that the probability p of coming up heads is somewhere between .4934 and .5134. Table 2.2 also shows data for more coin flips. Flipping the coin more times increases either the certainty or the accuracy of the empirical probability, or a little bit of both if you wish.

Here is the procedure you can use to solve problems like this. Suppose you have a binomial experiment, but you do not know the probability p of success. How many trials should you perform so that the empirical probability is within d of p with certainty r? Do the following steps:

1. Using the table of normal distribution, find z so that

$$P(Z < z) = (r+1)/2$$

2. Take the number trials to be at least

$$\frac{z^2}{4d^2}$$

You may round the number of trials up a little bit to be on the safe side and so your answer does not seem to imply more precision than it actually gives. So while $\frac{z^2}{4d^2} = 9604$ in the coin flip example, a good answer would be that you should do 10,000 coin flips.

Example: You wish to determine, with 95% certainty, the probability p that a coin will come up heads. You will accept an error for p of up to .005. How many times should you toss the coin? In this example, $r = .95$ and $d = .005$. So we first find z so $P(Z < z) = (.95 + 1)/2 = .975$. We find $z = 1.96$. So we should do at least $\frac{1.96^2}{4(.005)^2} = 38416$ coin tosses. You could round this up to 40,000. If your results were as in Table 2.2, you would get an empirical probability of .5008 and you would then be 95% certain that the true probability is between .4958 and .5058.

Exercises

1. None of the 120 students in a psychology class know the answer to a multiple choice question on their final exam. They all guess by randomly choosing one of the four answers. What is the probability that more than 30% of the students guess the correct answer?

2. Suppose 10% of the raisin bran boxes coming off the assembly line have fewer than 10 raisins in them. If you sample 150 boxes, what is the probability that less than 12% have fewer than 10 raisins.

3. How many times should you toss a six sided die to estimate the probability it comes up six? You want to be 95% sure that your answer is accurate to within .001.

4. Senator Small wants to determine his popularity with his constituents. He asks you to find out, with an error of .04, what proportion of his constituents approve of the job he is doing. How many constituents must you poll to be 80% sure of your answer?

5. The treasury department wants to find out the proportion of quarters in circulation in your city dated before 1990. You estimate this by getting rolls of quarters from the local bank and determining what proportion of them are dated before 1990. How many quarters should you examine to be 99% sure your answer is accurate to within .01?

6. A biologist catches 100 trout from a lake, marks their fins and then throws them back alive. Now she wants to determine the probability p that a trout in the lake has a marked fin. She does this by catching n trout, noting whether their fins are marked. If m have marked fines, she figures p is about m/n. How many fish n should she catch to be 90% certain that the empirical probability m/n is within .03 of p? (Note, this is a way she can estimate the population of trout in the lake since the number of trout in the lake is $100/p$.)