A tutorial on interval estimation for a proportion, with particular reference to e-discovery

William Webber

August 2nd, 2012 (v0.1)*

1 Introduction

This article is a primer or tutorial on sampling from a binomial (two-class) population; estimating the positive proportion in that population; and setting a confidence interval on said proportion. The tutorial is particularly intended for those working in e-discovery, in which the population is a collection of documents, the two classes are relevant and irrelevant documents, and the evaluator is attempting to estimate the proportion of documents in the collection (or some section of it) that are relevant to a production request. The tutorial is specialized to e-discovery only in two regards. The first specialization is that the examples we consider contain the very low positive proportions that are frequently encountered in e-discovery; such extreme proportions mean that some common approximation methods, such as the Wald interval (Section 7), can be inaccurate. The second specialization is in the final section (Section 9), where we briefly decode contemporary e-discovery practice, particularly the ubiquitous (but often misunderstood) "$95\% \pm 2\%$".

The tutorial is aimed at a non-mathematical audience that wants a deeper understanding of what is going on in point and interval estimation. It avoids mathematical formulae, and works instead with verbal descriptions and figures. Only the simplest form of sampling, namely simple random sampling, is considered; this is, in any case, the predominant form used in e-discovery, at least as encountered by non-technical practitioners. We focus on the sampling distribution, and how this relates to (and doesn't relate to) the confidence interval.

2 Model

Assume that every document in the collection is either wholly relevant or wholly irrelevant to a topic, and that we have a reviewer who is able to make the assessment of relevance without error, without changing their conception of relevance, and without the relevance of one document influencing the relevance of another. (These are

*Comments, corrections, and suggestions for improvement welcomed; please send them to william@williamwebber.com

unrealistic assumptions, but they are necessary for the sampling model we're going to develop to be strictly valid.) We'll also ignore the distinction between documents and document families (for instance, attachments and the emails they are attached to), and assume that the unit of assessment and the unit of production is the same.

Let the number of documents in the collection be N, and the proportion of these documents that are relevant be π; this latter is the value that we want to estimate. We draw n documents at random from the collection, in such a way that any set of n of the N documents in the collection is equally likely to be sampled, thus forming a *simple random sample*. The n documents sampled are assessed for relevance, and r of them are found to be relevant; thus, the proportion p of the sample that is relevant is r/n. Sampling in this way is often pictured as drawing n balls from a bag of black and white balls, in which π of the balls are white, where white balls represent relevant documents, and black balls irrelevant ones. This is known as *sampling without replacement*.

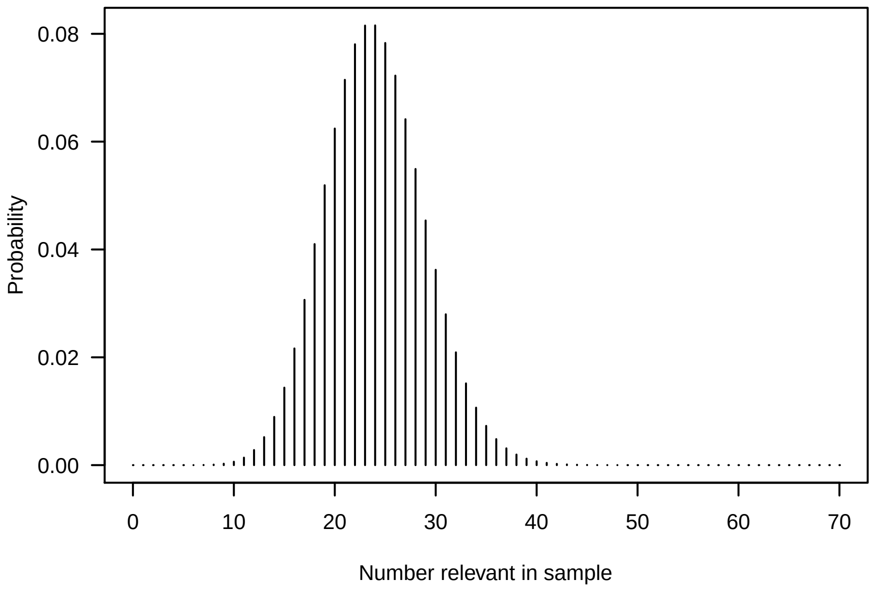
An alternative form of sampling is to choose one document at a time, until we have made n selections. Each document has the same probability of $1/N$ of being chosen at each draw, and the one document can be selected multiple times. In terms of the picture of the bag, we return each ball to the bag after it has been drawn. This form of sampling is called *sampling with replacement*. At each draw, the chance of drawing a white ball is π, which allows an even simpler picture to be applied: that of making n flips of a biased coin with probability π of turning up heads, where heads represents relevant.

Sampling without replacement gives marginally more accurate estimates, as well as being more natural in most circumstances (we wouldn't pick the same document to be assessed twice). Analysis based on sampling with replacement is easier, however, and gives a close approximation to sampling without replacement, provided the number of documents N is much larger than the sample size n ($N \gg n$). Since $N \gg n$ generally holds (for instance, we might be sampling 2400 documents from a collection with 1 million), the simpler approximation of with-replacement sampling is often used in analysis, even when the actual sampling has been without-replacement. We will perform with-replacement analysis in this tutorial.

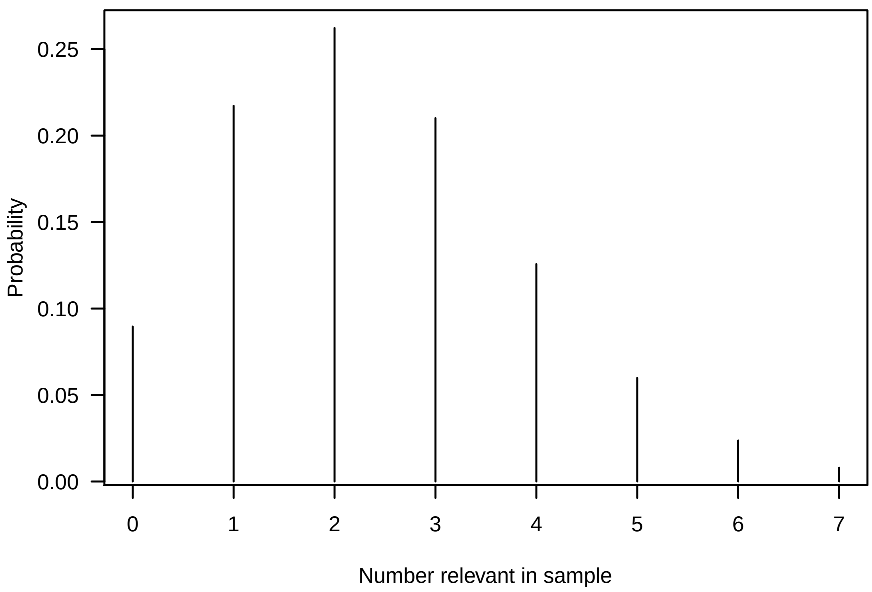
3 Sampling distribution

The number r of relevant documents in the sample will vary randomly from one sample to another, and with it the sample proportion p. The probability that a sample will contain a given number r of relevant documents, under simple random sampling with replacement, is given by a distribution known as the binomial distribution. We call r a sample statistic (which simply means some value calculated from the sample; $p = r/n$ is an alternative sample statistic), and say that the binomial distribution is the *sampling distribution of* this statistic (approximately so, if actual sampling is without replacement). The primary reason for performing sampling in a deliberately random way (rather than by judgment or by some default ordering, such as the first n documents in the collection) is so that results can be analyzed, and their sampling errors modelled, using such random distributions. Chance is more predictable than choice.

The binomial distribution for a true proportion $\pi = 1\%$ and a sample size of $n =$



(a) Sample size 2400



(b) Sample size 240

Figure 1: Binomial sampling distribution for a with-replacement sample of 2400 and of 100 documents from a collection with 1% of documents relevant. The x axis is scaled so that the proportion $p = r/n$ of relevant documents in the sample is the same.

2400 is shown in Figure 1(a); that for a sample size of $n = 240$ in Figure 1(b). The former figure shows that if we sample $n = 2400$ documents from a collection in which $\pi = 1\%$ are relevant, the probability that the sample will have 24 relevant documents is 0.0815 (around one in twelve); that it will have 18 is 0.0410 (around one in twenty-four); that it will have 12 is 0.0028 (around one in 360); and so forth. The probability does not drop to 0 for any value r of relevant documents in the range $\{0, 1, \cdots, n\}$ (though we have truncated the figure to the right). The probability of sampling 2400 relevant and no irrelevant documents from a collection in which only 1% of documents are relevant is infinitesimally small, but (at least under with-replacement sampling) it is not 0.[1]

Contemplating the binomial sampling distribution is all very comforting, but the practical reality is that for any given sample, though the statistic r we observe will come from some sampling distribution, we don't know which distribution has generated ours, because we don't know what the true proportion π is. Instead, we must use the observed statistic r as evidence to estimate what the value of π might be.

4 Point estimate

The first estimate we consider is the point estimate; a single value that we might roughly call the "best estimate" for π given the evidence, generally written $\hat{\pi}$. The commonest way of making this estimate is to ask: considering all the possible values of π, which makes the sample outcome p most likely? This estimate is known as a maximum likelihood estimate (MLE). When sampling from a proportion, the answer is straightforward (though the working to prove this answer is slightly less so[2]): the true proportion π for which the sample proportion $p = r/n$ is most likely to occur, is p itself. That is, p is the MLE of π; we write $\hat{\pi}_{mle} = p$.

Consider the scenario in which a sample of 2400 documents is drawn, and 24 are found to be relevant. We've seen above (Figure 1(a)) that the probability of sampling 24 relevant documents out of a sample of 2400 from a population with $\pi = 1\%$ relevant is is 0.0815. Let's hypothesize that π were slightly higher, say 0.011; then the probability of $r = 24$ would be lower, at 0.0728. Similarly, if π were slightly lower, say 0.009, then the probability of $r = 24$ also falls, to 0.0716. In fact, for any alternative $\pi \neq 0.01$, we'll find that the probability of sampling $r = 24$ is lower than it is for $\pi = 0.01$, as is shown in Figure 2. Therefore, $\hat{\pi}_{mle} = p = 0.01$.

A point estimate alone, however, is insufficient. Every random sample has a sampling distribution (the distribution for the previous paragraph's scenario is shown in Figure 1(a)); therefore, every random sample has the possibility of sampling error. Moreover, sampling error will vary for different sampling setups. In particular, the

[1] For those who like to contemplate such things, it is $1/10^{4800}$. The divisor here has 4800 zeroes. In comparison, the Milky Way galaxy is estimated to have in the order of 10^{69} atoms. If each of these atoms were converted into a galaxy the size of the Milky Way galaxy, then the total number of atoms in all of these galaxies would be 10^{4761}, still a duodecillion (a thousand trillion trillion trillion) times smaller than our divisor. It is an interesting question whether we could develop a sampling method so truly random as to given an event a faithful $1/10^{4800}$ probability. If such a method were proposed, don't volunteer to test it.

[2] http://en.wikipedia.org/wiki/Maximum_likelihood#Discrete_distribution.2C_continuous_parameter_space

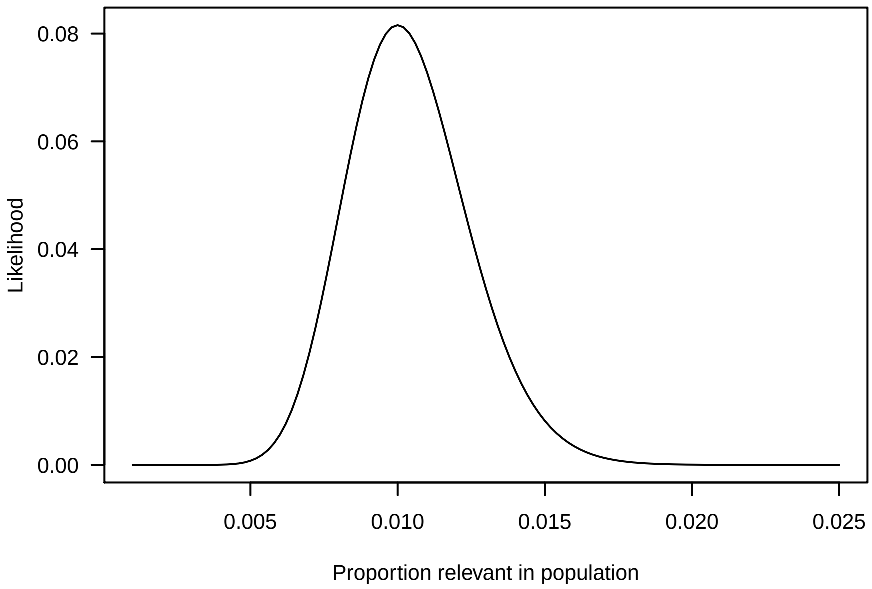


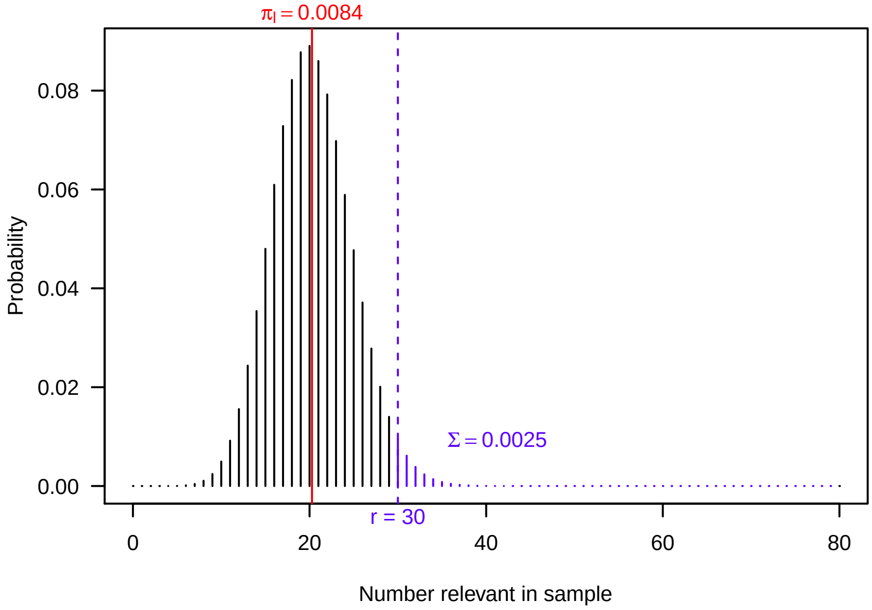
Figure 2: Likelihood of drawing 24 relevant documents in a sample of 2400 from a collection with a given proportion relevant.

smaller the sample, the greater the likely error. The inverse relationship between sample size and likely error can be seen by comparing, with the 2400 sample, the greater spread (in terms of sample proportion, $p = r/n$) of the 240-sample distribution in Figure 1(b). A sample proportion of 1 in 60, almost twice the true proportion, has a probability of one in eight in the 240 sample, but less than one in a thousand for the 2400 sample. When we quote an estimated result, we need to express the uncertainty inherent in our random sampling and estimation setup.

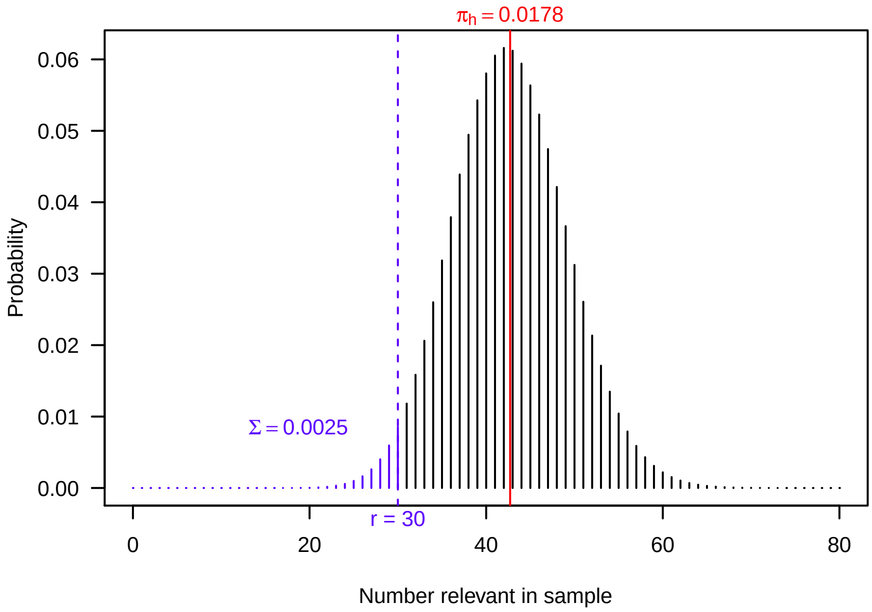
5 Confidence intervals

A common way of expressing the uncertainty of random sample estimation is through a second form of estimate known as a confidence interval. This interval provides a range of values, and states as a percentage our degree of confidence that the true value of π is within that range. We might say, for instance, that π is between 0.006 and 0.016 with 95% confidence.

A confidence interval is derived by reasoning about sampling distributions, but in a different way from the reasoning that leads to the MLE point estimate. For the point estimate, we ask what proportion $\widehat{\pi}$ makes the observed sample statistic r most likely. For confidence intervals, we instead look for bounding proportions π_l and π_h that each give the observed sample statistic a particular degree of unlikeliness.



(a) Lower bound



(b) Upper bound

Figure 3: Sampling distributions for the hypothesized lower and upper bound on a 95% exact binomial confidence interval on the proportion of relevant documents in a collection, for a sample of 2400 documents, of which 30 are relevant.

Let's return to our scenario of finding $r = 30$ relevant documents in a simple without-replacement random sample of $n = 2400$ documents. Our goal is to calculate a 95% confidence interval on π from this sample result. Starting with the lower bound, we ask ourselves, for what true proportion π_l would the probability of observing 30 *or more* relevant documents in the sample be 2.5%? As it turns out, this proportion is 0.84%. The sampling distribution of $\pi = 0.84\%$ is shown in Figure 3(a). The height of the bars that are at or above $r = 30$ sum to 0.025; that is, when π is 0.84%, there is a 2.5% probability that our sample will have 30 or more relevant documents in it. This sets the lower bound of our interval.

Next, we ask, for what true proportion π would the probability of observing 30 *or fewer* relevant documents in the sample have been 2.5%? This proportion works out to $\pi_h = 1.78\%$, as shown in Figure 3(b). That in turns set the upper bound of our interval. We now have two hypothesized values, one high, one low, under each of which the probability of a result at least as extreme as our observed result is 2.5%. So $[0.84\%, 1.78\%]$ is our (100% - 2.5% - 2.5% =) 95% confidence interval on π.

To be precise, the formal definition of a confidence interval requires us to go a couple of steps further. A confidence interval on a proportion has 95% confidence if the following holds: for any true proportion π, if an approaching-infinite number of samples were drawn, and for each a sample a confidence interval were calculated using the same procedure, then at least 95% of these confidence intervals would include π. The reasoning with the sampling distributions of hypothesized bounding π values is a procedure that satisfies this formal requirement.

The method described above for calculating a confidence interval on a proportion is known as the exact binomial confidence interval, because it is based on the exact (binomial) sampling distribution of the statistic (though, in fact, the binomial distribution itself is an approximation if sampling is without replacement, as it generally is). (The method is also known as the Clopper-Pearson interval, after its discoverers.) The exact interval guarantees 95% coverage, in the formal sense described in the previous paragraph. For most π values and sample sizes, coverage will actually be above 95%, making the interval conservative. Despite (or because of) this conservatism, it is the interval one would recommend for certification purposes. Several approximate intervals, however, have also been developed, for analytical convenience or reduced conservatism. We're going to look at two of these next, the Wilson (Section 6) and the Wald (Section 7) intervals, both of which use so-called normal approximations. The Wald does so in a particular simplifying way, making it widely used in exposition and rough reckoning, but also helping spawn some of the misconceptions about confidence intervals that we will discuss in Section 8.

6 An accurate approximate interval: the Wilson

A distribution that pops up all the time in statistics is the normal distribution, colloquially known as the bell curve. The formula for the normal distribution is not particularly simple, but its properties are familiar and computationally convenient, so it is a preferred analytic tool. As it happens, the binomial distribution is approximately normal, more closely so as the sample size increases. Whereas the binomial distribution takes sample size and proportion as its parameters, the normal distribution takes mean (the center of the distribution) and variance (which gives the width of the distribution). A binomial distribution with sample size n and proportion π is approximated by a normal distribution of mean $\mu = n\pi$ and variance $\sigma^2 = n\pi(1 - \pi)$.

We can use the normal distribution to approximate the sampling distribution of the binomial in deriving a confidence interval on a proportion. As with the binomial, we find the lower-bound π_l for which a normal sampling distribution has a 2.5% probability of generating the observed sample r or higher, using π_l as the mean and $\pi_l(1-\pi_l)/n$ as the variance, and conversely for the upper-bound π_h. The approximate normal confidence interval on the proportion described above is known as the Wilson (or score) interval.

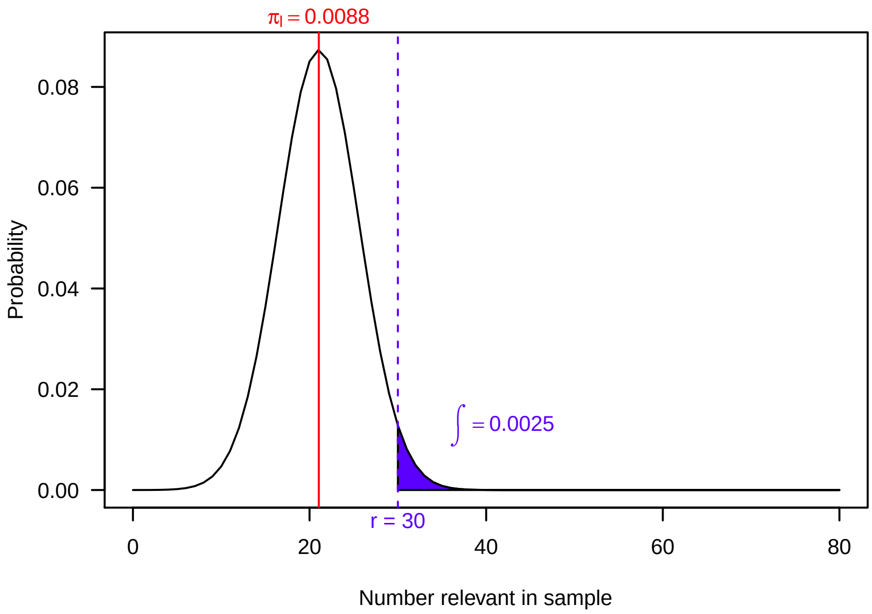
The binomial distribution is discrete, giving probabilities only to whole samples r such as $0, 1, 2$, and so forth, whereas the normal is continuous, giving probabilities (more formally, probability densities) to fractional samples.[3] Therefore, graphically, in calculating the 2.5% tails of the bounding sampling distributions, we are measuring the area under a curve, not summing the probabilities at sample points. Taking again our scenario of a sample size of 2400 with 30 relevant documents, the lower-bound normal approximate sampling distribution is shown in Figure 4(a), and the upper-bound sampling distribution in Figure 4(b). The Wilson interval for this sample outcome is $[0.88\%, 1.78\%]$.

7 A less accurate approximate interval: the Wald

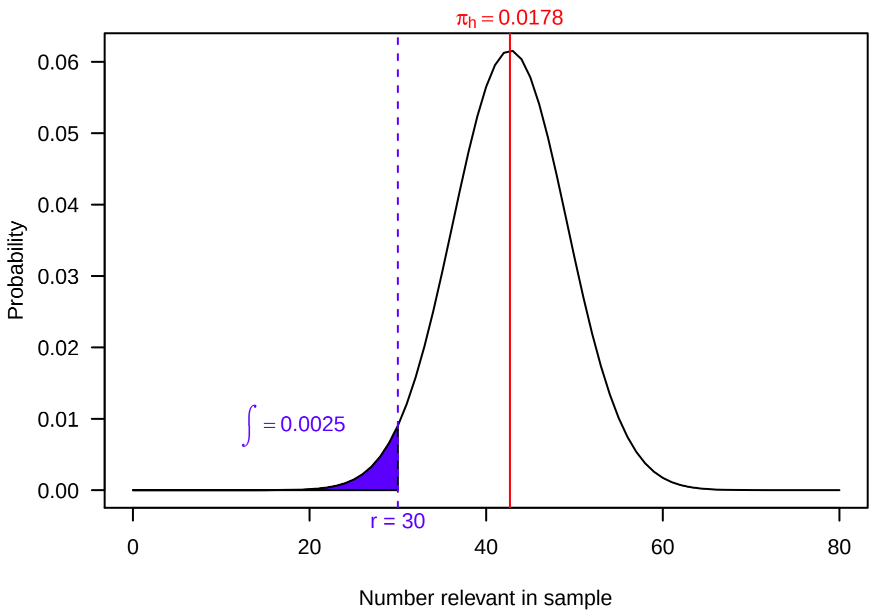
The Wilson interval is still analytically inconvenient because the bounding sampling distributions have different variances and hence different widths. We can simplify matters further if we give each bounding distribution the same variance. Since the variance of the normal approximation to the binomial is derived from the proportion π, this is equivalent to using the same π to calculate the variance of both bounding distributions, rather than the actual π values at the hypothesized bounds. A simple choice for this proportion is the actual proportion observed in the sample, p, from which we get the variance $p(1 - p)/n$. We then need to find the hypothetical low and high bounds that fit these equal-width sampling distributions. This leads to bounding distributions with the same shape, differing only in location, as we see for our example scenario in Figure 5. This confidence interval is known as the Wald interval. The Wald interval for our sampling example is $[0.81\%, 1.69\%]$.

Since the bounding sampling distributions of the Wald interval are identical and symmetric, it follows that the interval is symmetric, as wide below the MLE point estimate as above. And indeed the interval for the example scenario is symmetric in this way; the point estimate is $r/n = 30/2400 = 1.25\%$, and the interval can be expressed as $1.25\% \pm 0.44\%$. Moreover, if you were to move the two intervals so that they were centered on the sample value r, then not only would they overlap, but also the outer 2.5% tails of the melded interval would sit on the boundaries of the interval,

[3] This makes little modelling sense, but it allows us to adjust the boundary values more precisely to give average coverage of 95% (though at the price of under-coverage for certain proportions π).

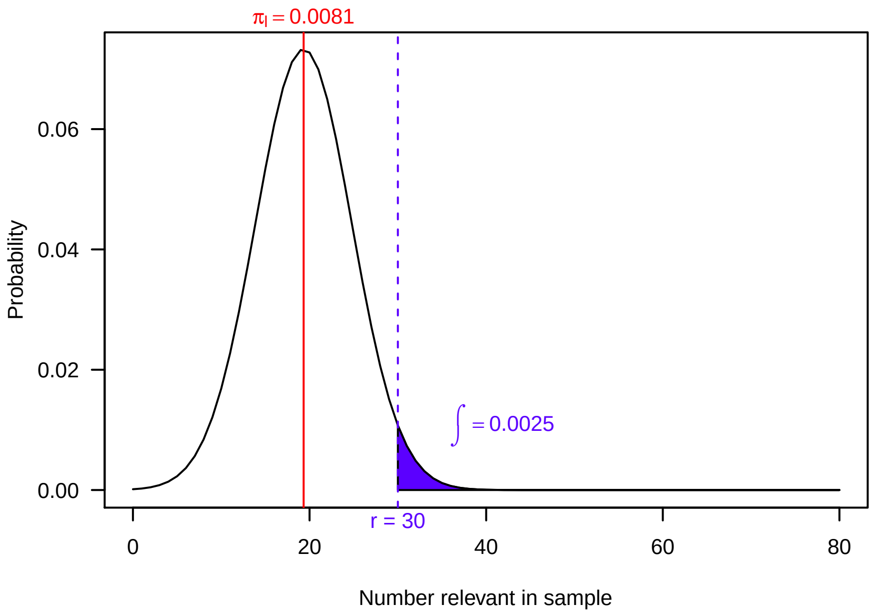


(a) Lower bound

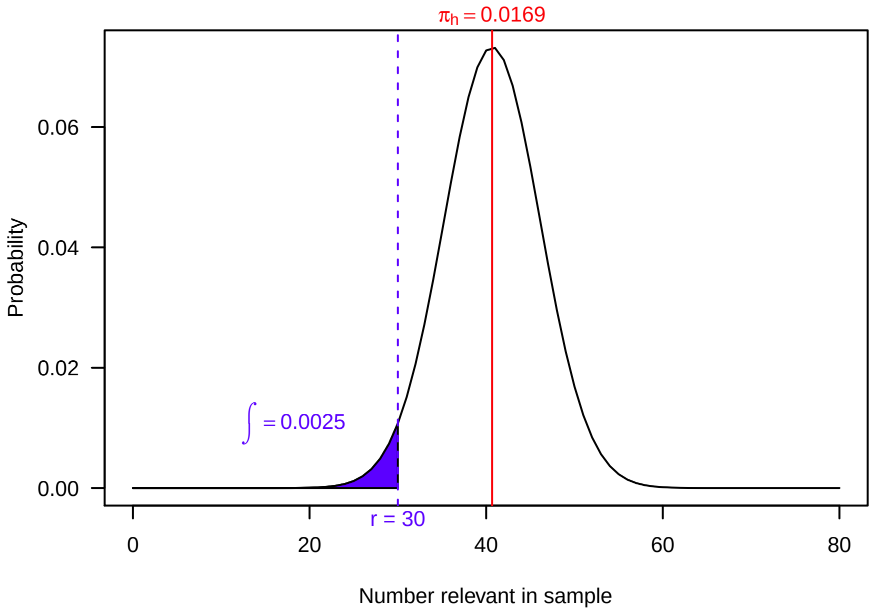


(b) Upper bound

Figure 4: Normal approximation sampling distributions for the hypothesized lower and upper bound on a 95% Wilson confidence interval on the proportion of relevant documents in a collection, for a sample of 2400 documents, of which 30 are relevant.



(a) Lower bound



(b) Upper bound

Figure 5: Normal approximation sampling distributions for the hypothesized lower and upper bound on a 95% Wald confidence interval on the proportion of relevant documents in a collection, for a sample of 2400 documents, of which 30 are relevant.

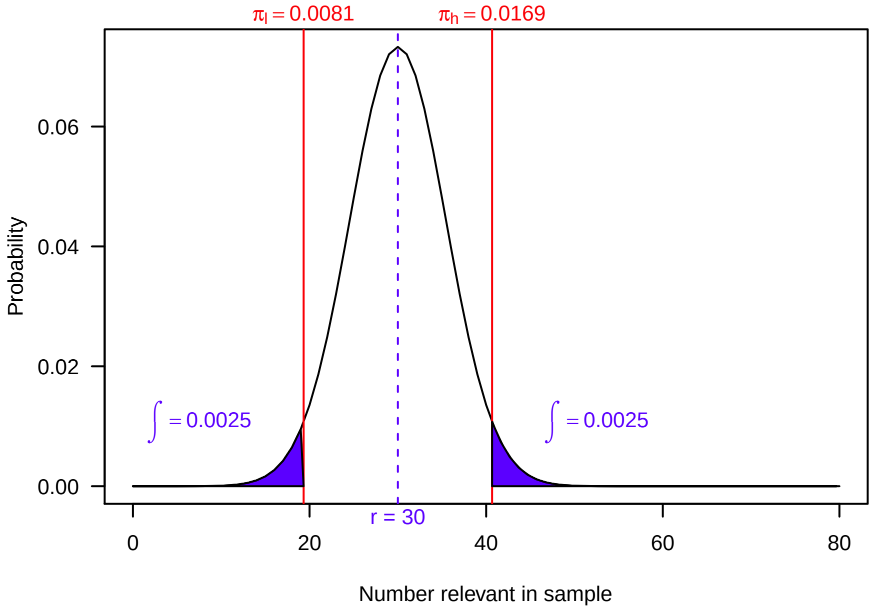


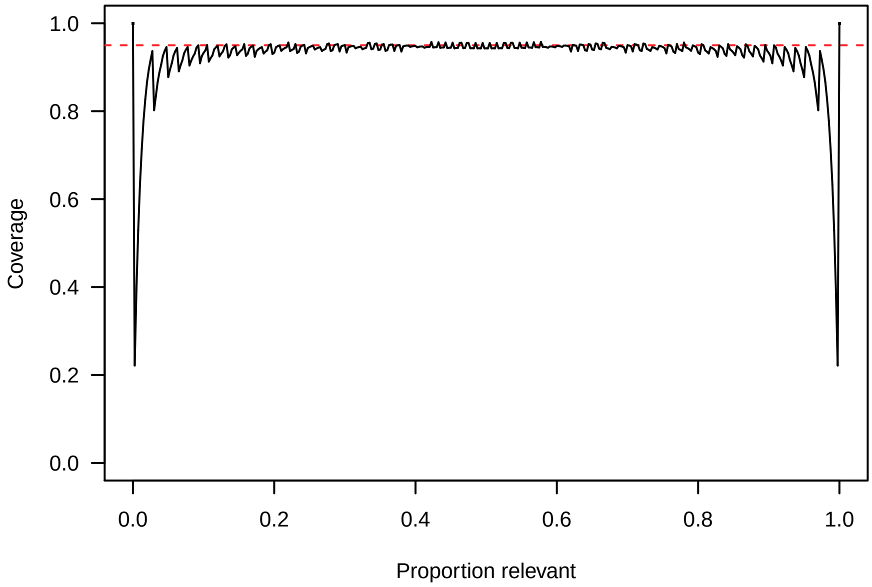
Figure 6: Wald interval interpreted as a margin of error.

just as the inner 2.5% tails of the bounding intervals did, as illustrated in Figure 6. This relationship holds for any sample, and for any confidence level.

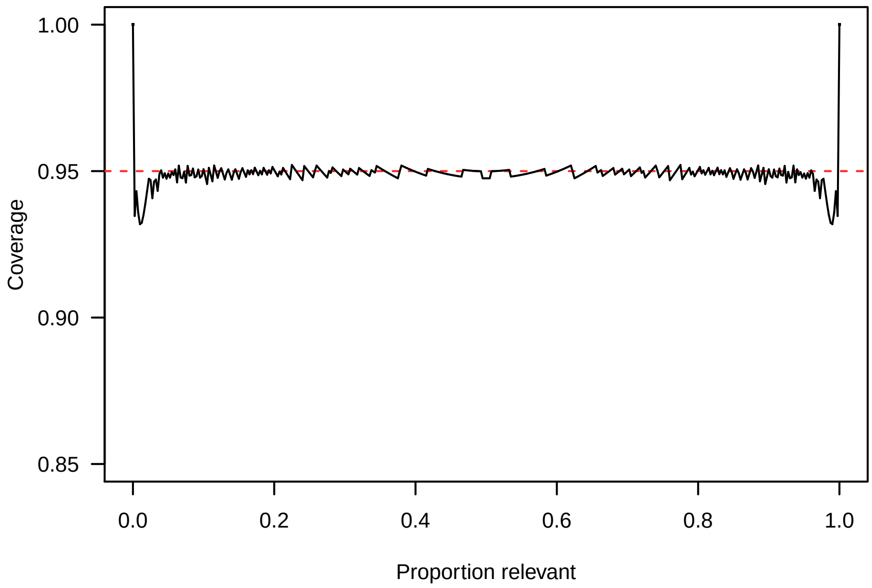
What we have in effect achieved is to replace the confidence interval derived from bounding sample distributions with one taken from the tails of a single sampling distribution, which is the same as the (approximate normal) sampling distribution of the MLE point estimate for π. This seductively encourages us to stop thinking about bounding distributions altogether, and start to think of the confidence interval as expressing some sort of distribution of error around the point estimate itself.

Once we make this simplification, all sorts of possibilities open up to us. For the mathematical statistician, methods of working with normal distributions are numerous; for instance, we can estimate the intervals of compound measures such as recall by the technique of "propagation of error". For the back-of-the-envelope statistician, the confidence interval has the simple form $p \pm 1.96 * \sqrt{p(1-p)/n}$, and we can see that (for instance) to halve the width of the confidence interval, we need to quadruple the size of our sample.

Unfortunately, the Wald interval is quite inaccurate in some circumstances, particularly when sample sizes are small and the true proportion π is close to 1 or (as it often is in e-discovery) 0. The accuracy of a confidence interval method for a given sample size can be measured by computing the probability that different true proportions π will be contained in the interval. Figure 7 displays coverage for the Wald interval, for sample sizes of 100 and 2400. Though coverage of 95% is the goal, it can fall as low as 25% for the smaller sample, when the proportion relevant in the population is low;



(a) Sample size 100



(b) Sample size 2400

Figure 7: Coverage of the Wald confidence interval on a proportion for sample size of 100 and sample size of 2400. Note the different y axes.

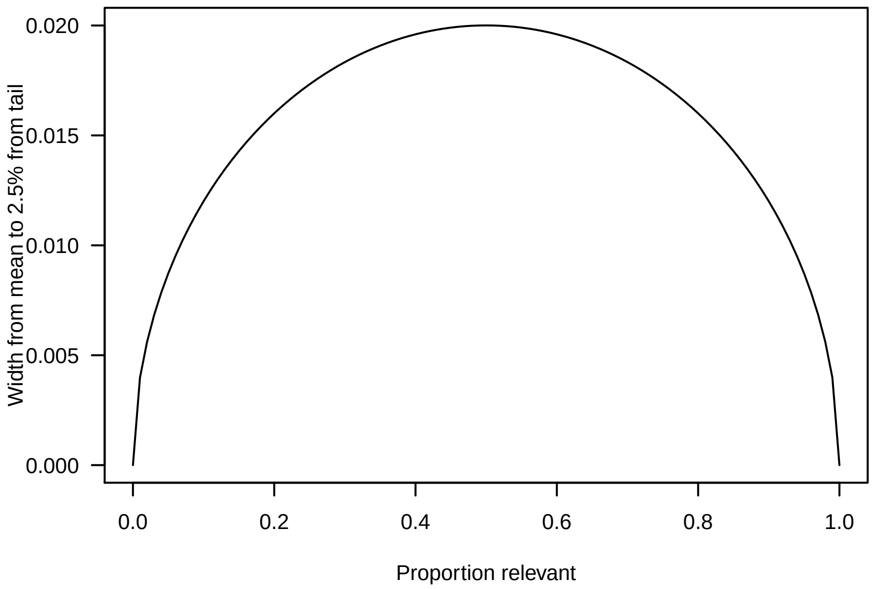


Figure 8: Width to the 2.5% tail for the normal approximation to the binomial sampling distribution for different true proportions π, on a sample size of 2400.

for the larger sample, undercoverage is limited to 93%.

The reason for the inaccuracy of the Wald interval is that, by using p for the variance of both bounding distributions, the Wald interval enforces a specious symmetry on the interval. The true interval is asymmetric; for samples with $p < 0.5$, the upper bound sampling distribution is wider than the lower; and this disparity grows as p approaches 0, as Figure 8 shows. A particular crisis for the Wald interval occurs when there are no relevant documents in the sample; here, p is 0, and an anomalous interval of $[0, 0]$ is produced, no matter how small the sample size. This interval is clearly incorrect: one can sample from a collection with relevant documents in it and have no relevant documents in the sample.[4] For these reasons, though useful as an analytic tool, the Wald interval should be avoided in practice.

We tabulate the interval estimates from the three interval methods we've considered in Table 1. The Wald interval is symmetric ($1.25\% \pm 0.44\%$), whereas the exact and Wilson intervals are asymmetric, wider on the side towards 50%. The symmetry means that, compared to the other two intervals, the Wald interval is slightly longer on the low end, and decidedly shorter on the upper end. The exact and Wilson intervals are identical (within rounding) at the upper end; the Wilson is slightly shorter at the lower end. We can't actually say which of these intervals is more "accurate" in this case,

[4] A human, seeing a $[0, 0]$ interval on a modest sample will realize something is wrong, even if they're not sure what; but these testing regimes are increasingly computerized, and a computer seeing this will carry blithely on.

Method	Interval		1.25% ± width	
	Bottom	Top	Lower	Upper
Exact	0.84	1.78	0.41	0.53
Wilson	0.88	1.78	0.37	0.53
Wald	0.81	1.69	0.44	0.44

Table 1: Intervals and interval widths for the 95% exact binomial, Wilson, and Wald confidence intervals, for a sample of 2400 documents, in which 30 were relevant.

though, since we don't know what the true proportion π is (and even if we did, an interval only needs to cover it 95% of the time, not always).

8 Confidence interval misconceptions

We've covered the difficult ground in our discussion of confidence intervals on the proportion; now it is time to use what we've learnt to clear up some misconceptions about confidence intervals.

8.1 Confidence intervals are not necessarily symmetric

It is very common to think of a confidence interval as some sort of symmetric "margin of error" on the point estimate, and express the interval in a form like "0.2 ± 0.03". However, the exact confidence interval on the proportion (and accurate approximations to it) is only symmetric in one special case, where the observed sample proportion p is 0.5. For every other value of p, the interval is asymmetric, longer on the inward than the outward side, and sometimes significantly so. The Wald interval does always give symmetric intervals, but this is the main cause of its inaccuracy.

8.2 Confidence level does not equal width

Don't make the mistake of thinking that the confidence level of an interval (say, 95% or 99%) has a simple mapping to the interval's width. A 99% interval is not simply 4% wider than a 95% interval, at least not in the space of the proportion parameter. If we used the Wald interval, then we can consider the 99% interval to be 4% wider in a sort of probability space – that is, we go out a further 2% at each end of normal distribution. But this is much more than 2% wider when expressed in proportions of the population. In fact, a 99% interval on a proportion is more than 30% wider relative to a 95% interval; for instance, if the 95% interval is $[0.4, 0.6]$, the 99% interval is $[0.37, 0.63]$. And a 99.9% interval is 30% wider than a 99% interval; and so forth. This leads on to the next point.

8.3 A 100% confidence interval is largely meaningless

Newcomers to estimation occasionally ask for a 100% confidence interval, or wonder whether we settle for 95% just out of some laziness or vagueness ("after all, it's only another 5% . . . "). A 100% interval, however, will be at least $(0, 1)$; the rounded brackets mean we can rule out 0 or 1, but only if we see at least one relevant or one irrelevant document in the sample. We can only rule out true proportions that have no probability of producing the observed sample. But even a true proportion of 99.9% has some probability of producing a sample holding no relevant documents (although again we may be off counting atoms within galaxies within atoms within galaxies). When we sample, we have to accept some degree of uncertainty; what is under our control is the degree of uncertainty we're willing to accept.

8.4 The width of the confidence interval cannot generally be known in advance of sampling

We would like to know before we design our sample how wide our confidence interval will be, for a given sample size. Unfortunately, as we have seen, the width of a confidence interval on a proportion depends on the sample proportion actually observed. All we can do is say what the maximum confidence interval width will be, which occurs when the sample proportion is 0.5.

8.5 A confidence interval is not simply the percentiles of a sampling distribution around the point estimate

A frequent misconception is that a confidence interval can simply be taken from the percentiles of the sampling distribution of the point estimate; that, in other words, the confidence is simply a sampling "margin of error" around the point estimate.[5] Rather, as we have seen, a confidence interval is formed from the inward-facing tails of the sampling distributions around the upper and lower hypothesized bounds on the interval. This is only equivalent to percentiles of a sample-centered when the sampling distribution is (or is approximated as) symmetric and identically-shaped for all parameters. That is the case for the Wald approximation; but, as we've seen, the Wald approximation is often not a close one.

9 Confidence intervals in e-discovery practice: the meaning of $95\% \pm 2\%$

Now – finally! – we're ready to look at the use of confidence intervals in e-discovery. Of course, there's an enormous amount that could be said here, so I'm going to restrict myself to just one point: clarifying the meaning of the sometimes mysterious expression "$95\% \pm 2\%$" that is frequently quoted in e-discovery practice, along with the magical sample sizes 2399 and 2401 that accompany it.

[5] People commonly misapply bootstrap and other resampling methods in this way.

First, let's unpack "$95\%\pm2\%$". It must be immediately clarified that this expression is properly made about a planned sampling task (or, at least, retrospectively about what was planned), not about an actual estimate; the reason has to do with the "$\pm2\%$", as we'll see in a moment. The "95%" indicates that what is being planned is a 95% confidence interval, typically on the proportion of relevant documents either in the whole collection or in some part of it (such as the subset of the collection not being produced). That is, the evaluation designer wants to be able to say at the end of the sampling something like "we have 95% confidence that the true proportion of relevant documents lies within the interval [x, y]".

The "$\pm2\%$" is stating the desired width of the confidence interval. The "2%" is a proportion of the entire collection, not of the point estimate; the evaluator is imagining a statement like "$44\% \pm 2\%$". As we said before, however, the exact width depends upon the observed sample proportion; here, "$\pm2\%$" is the maximum width. When we've drawn the sample, we can calculate the exact interval; it is a mistake to simply apply the $\pm2\%$ to the observed sample proportion p. The interval on a proportion is widest when the sample proportion p is 0.5. This can be observed for normal approximations in Figure 8, but the same holds true for the exact interval and closer approximations to it. So to achieve the "$\pm2\%$" guarantee, the evaluator must choose a sample size large enough to give an interval this width if the sample proportion turned out to be $p = 0.5$.

We mentioned above that the exact interval is not generally symmetric, and so that stating it in terms of a point estimate plus or minus a margin of error is incorrect. For the special case that the sample proportion p is 0.5, the exact interval is, however, symmetric, so "$\pm2\%$" is correct as a statement of maximum width, though misleading if we take it to imply that the actual interval will be symmetric.

What sample size is necessary to achieve the worst-case goal of "$\pm2\%$"? For the exact binomial confidence interval, 2399 samples are required, which is where this magical number comes from. If the approximate Wald interval is used instead, the number is 2401. We've seen that the Wald interval can be quite inaccurate; you may prefer to associate with e-discovery statisticians who talk about "2399" than those talking about "2401". You'll notice I've sat on the fence in this tutorial by working with 2400.