rule of three

Argyris Stringaris

03/04/2023

Rule of 3

Suppose you encounter a zero numerator in a study, e.g. in a trial of a drug that reports its side effects. How do you estimate the probability of a problem occuring? It is impossible in the absence of a probability, but what you can estimate is the upper CI for this, that is the maximum theoretical value, the upper bound, that would be obtained in the long run, i.e. after repeated samplings within the 95% of the sample. This will be the p below. Importantly, you can use the **Rule of Three** which I explain below.

Generally, in order to derive the risk, you can use the Bernoulli function for binary events, i.e. what you woudl use if you were wanting to find out, say, the number of heads when tossing a coin many times.

$$\binom{n}{k} p^k (1-p)^{n-k} (1)$$

where

$$\binom{n}{k} = \frac{p!}{k!(n-k)!} (2)$$

because $k = 0$, and $0! = 1$, (1) simplifies to

$$(1-p)^n$$

What needs to be satisfied is that this relationship be equal to the maximum risk, e.g. 0.05 (if using a 95% confidence interval), hence:

$$(1-p)^n = max\_risk$$

which, can be transformed to

$$(1-p) = max\_risk^{1/n}$$

and

$$p = 1 - max\_risk^{1/n}$$

NOTE: The **Rule of Three** can be intuited as follows.

If you take this relationship from above:

$$(1-p) = max\_risk^{1/n}$$

You could solve it using the natural logarithm on each side.

You can find out, by using a hand calculator that:

a) for any small p (e.g. 0.01), $ln(1-p)$ reduces to approximately $-p$. You can get this more clealry if you use a Talyor series approximation, which I am too lazy to write out.

b) that $ln(\text{max\_risk}) \approx -3$, again, verify this using your calculator.

Transforming the above, we get:

$$-p = -ln(0.05)^{1/n}$$

which after dealing with signs, and using log rules becomes:

$$p = \frac{3}{n}$$

This is how the rule of three arises.

Now, let's show this with some examples below

```r
n <-  c(10,20, 30, 40, 50,  80, 100, 120, 150, 200, 250,500,750, 1000) # these are the sample sizes

max_risk <- 1-0.95 # this is the 95% confidence interval you want to have, you could of course use a di


p = 1- max_risk^(1/n) # this is the one to get

p # the vector of probabilities
```

```
##  [1] 0.258865551 0.139108341 0.095033853 0.072157525 0.058155079 0.036754198
##  [7] 0.029513050 0.024655401 0.019773438 0.014867039 0.011911420 0.005973552
## [13] 0.003986343 0.002991250
```

As you can see, the probabilities, i.e. the upper bounds of the probabilities of a problem being prsent, decrease as the sample size in which a zero numerator was found, increases.

You can verify the values in this vector in the following way: for this rate of events, i.e. zero at each of the sample sizes and for each p-value contained in the vector, the output should be 0.05 when plugged into the binomial formula. Here is the test.

```r
pbinom(0, n,p) # where n is the sample sizes above, and p the vector of probabilities above.
```

```
##  [1] 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05
```

```r
# for which the following holds
round(pbinom(0, n,p) ,3) == rep(max_risk, length(n))
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE
```

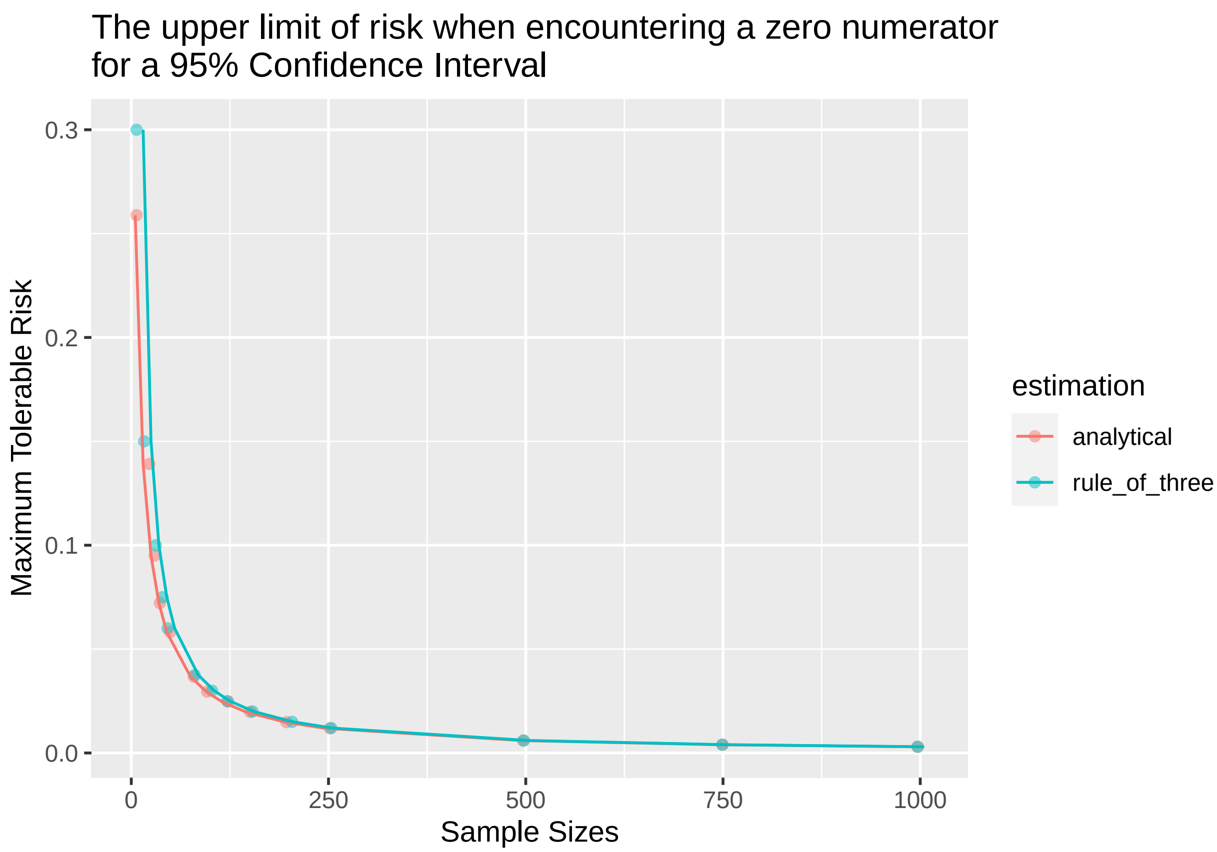
Now you can verify the above using the rule of three

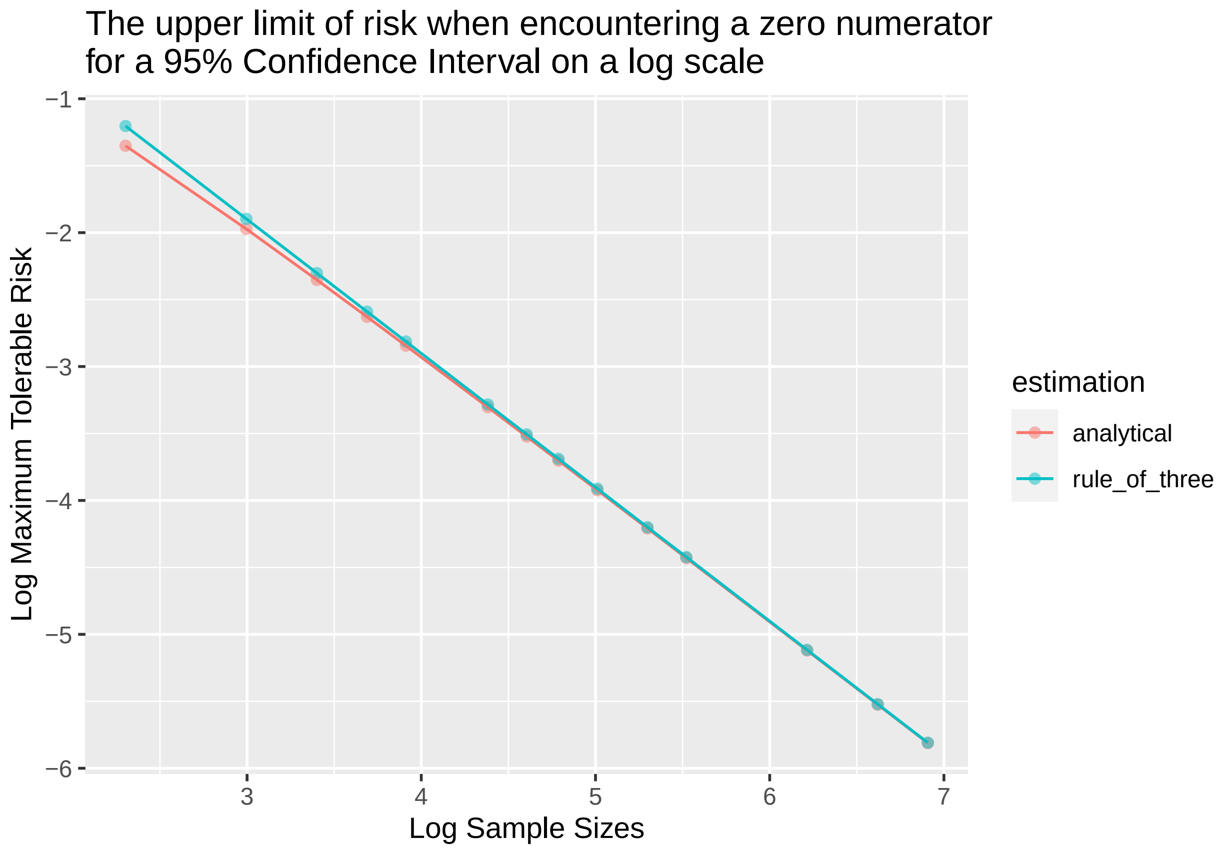
```r
rule_three  <- 3/n # this is the rule of three
rule_three
```

```
##  [1] 0.3000 0.1500 0.1000 0.0750 0.0600 0.0375 0.0300 0.0250 0.0200 0.0150
## [11] 0.0120 0.0060 0.0040 0.0030
```

Now plot all this to show differences and overlap between analytical and rule of three approximation

```
## Warning: position_dodge requires non-overlapping x intervals
```





The upper limit of risk when encountering a zero numerator
for a 95% Confidence Interval on a log scale
Log Maximum Tolerable Risk
Log Sample Sizes
estimation
analytical
rule_of_three