Binomial model: introduction

Let us suppose that we have the following Binomial model for our data, where the parameter p is the proportion of successes in n independent Bernoulli trials

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

Moreover, an estimator is a function to estimate the unknown parameter, that is p in our case, from a sample of independent observations $x_1, ..., x_n$. The particular estimated value, i.e. the estimate, is denoted by $\hat{\theta}$ and is clearly a function of the data. We have $\hat{\theta} = f(x_1, ...x_n)$.

We would like to estimate p by $\hat{p}$, the relative frequency of successes. To do so, we count the successes and divide it by the number of observations to obtain the Maximum Likelihood Estimator (MLE) for p. We thus have that $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} k_i$.

Confidence interval estimation

This study has for main reference the technical report 'Construction of Confidence Intervals', Christoph Dalitz (2017)

Suppose that we do not simply want a point estimate for p, but we wish to construct $(1-\alpha)$-Confidence Intervals (C.I.) for p. We list here some possibilities:

(i) The exact Clopper-Pearson C.I., which can be computed numerically.

(ii) The Wilson approximate C.I, based on a Normal approximation.

(iii) The Likelihood Ratio Support Interval.

(iv) The Bayesian Highest Posterior Density (HPD) Interval.

Exact Clopper-Pearson CI

A Confidence Interval (CI) is a region $[\theta_l, \theta_u]$ where the parameter falls with a high probability. θ_l denotes the lower bound and θ_u denotes the upper bound. When the fixed probability α is distributed evenly among deviations, the formal definition of the frequentist CI is given by:

$$P_{\theta=\theta_l}(\hat{\theta} \geq \theta_0) = \alpha/2 \quad \text{and}$$
$$P_{\theta=\theta_u}(\hat{\theta} \leq \theta_0) = \alpha/2$$

where $\hat{\theta}$ is the observed value of the estimator and θ_0 is one fixed value of the said estimator. Then, based on the formula on slide 1, the Exact Clopper-Pearson CI. can be computed numerically in R as:

$$1 - pbinom((k-1)/n, n, pl) = \alpha/2 \quad \text{and}$$
$$pbinom((k-1)/n, n, pu) = \alpha/2$$

Wislon approximate CI and Likelihood Ratio Support Interval

The Wilson approximate CI, which is based on a Normal approximation, and given by:

$$\frac{1}{1+z^2/n}\left[\hat{p}+\frac{z^2}{2n}\pm\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}\right]$$

where n is the sample size and p, the probability of success. These are the two parameters of a Binomial model. Then we introduce another method to construct a CI, namely the Likelihood Ratio Support Interval, given by:

$$\frac{L(p)}{L(\hat{p})}=\frac{p^k(1-p)^{n-k}}{\hat{p}^k(1-\hat{p})^{n-k}}\leq\frac{1}{K}$$

where $\hat{p}=k/n$ and is the ML estimator for p. Also, we will set $K=8$, for convenience.

Bayesian Highest Posterior Density (HPD) Interval

To obtain a Bayesian Highest Posterior Density (HPD) Interval, and assuming a noninformative prior for p we can use the following code in R:

```
library(HDInterval)
CI <- hdi(qbeta, 1-alpha, shape1=(k+1), shape2=(n-k+1))
```

Since the posterior $p(p \mid k)$ is given by:

$$
\begin{aligned}
p(p \mid k) &= \frac{\binom{n}{x} p^k (1-p)^{n-k}}{\int_0^1 \binom{n}{x} q^k (1-q)^{n-k} dq} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\ \Gamma(b)} p^{a-1} (1-p)^{b-1} \\
&= dbeba(p, a, b)
\end{aligned}
$$

where $a = k + 1$ and $b = n - k + 1$.

Simulations: empirical coverage rate of the 95% CI

We will compute the empirical coverage probability for the relative frequency $\hat{p}$. To do so, we will do proceed as follows:

(i) Simulate samples of size n and compute an estimate $\hat{p}$ of p.

(ii) Compute a confidence interval for $\hat{p}$.

(iii) Count the number of times the true value p falls into the given interval.

For this setting, we use $n = 100$ and make p vary from 0 to 1 by steps of 0.01. For each $\hat{p}$, we use $m = 10,000$ simulations.

R code to compute the approximate Wilson CI

```r
# function to compute Wilson confidence intervals
approximate.ci.binom <- function(n, k, alpha) {

  if (k == 0) {
    pl <- 0.0
    sd <- sqrt(   ( ((qnorm(1-alpha/2))^2) / (4*n^2)   )   )
    pu <- ( 1 / (1 + (qnorm(1-alpha/2)  / n) ) )  *
              (   ( qnorm(1-alpha/2) / (2*n)) + qnorm(1-alpha/2) * sd )
  }
  else if (k == n) {
    sd <- sqrt(   ( ((qnorm(1-alpha/2))^2) / (4*n^2)   )   )
    pl <- ( 1 / (1 + (qnorm(1-alpha/2)  / n) ) )  *
            ( (k/n)     + ( qnorm(1-alpha/2) / (2*n)) - qnorm(1-alpha/2) * sd )
    pu <- 1.0
  }
  else {
    sd <- sqrt( (( (k/n)* (1 - (k/n))   )/n) +  ( ((qnorm(1-alpha/2))^2) / (4*n^2)
)  )
    pl <- max(0, ( 1 / (1 + (qnorm(1-alpha/2)  / n) ) )  *  ( (k/n)     +
            ( qnorm(1-alpha/2) / (2*n)) - qnorm(1-alpha/2) * sd ) )
    pu <- min(1, ( 1 / (1 + (qnorm(1-alpha/2)  / n) ) )  *  ( (k/n)     +
            ( qnorm(1-alpha/2) / (2*n)) + qnorm(1-alpha/2) * sd ) )
  }
  return (data.frame(pl=pl, pu=pu))
}
```

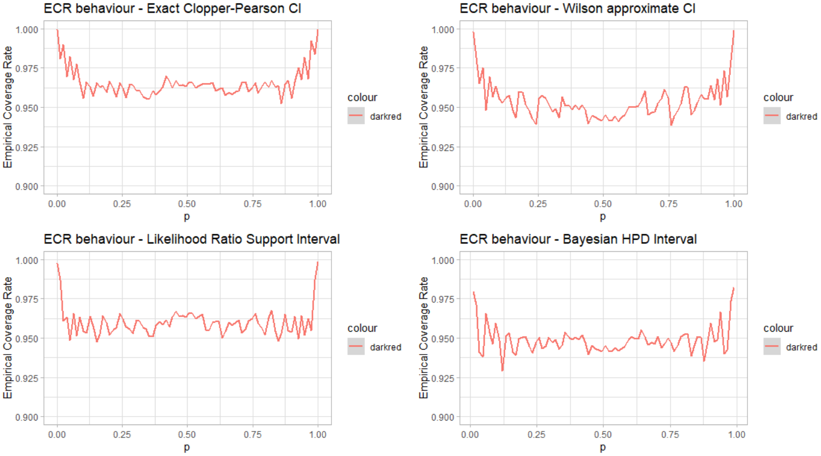
R code to obtain the empirical coverage rate: example with Wilson CI

```
n = 100 # sample size
k = 0:n # number of successes
alpha = 0.05
p = matrix(rep(k/n,10000), ncol=length(k/n), byrow = TRUE)
p_hat = matrix(rep(0,10000*(n+1)), ncol=(n+1)) # Number of replications: 1,000
ci_lowerW = matrix(rep(0,10000*(n+1)), ncol=(n+1))
ci_upperW = matrix(rep(0,10000*(n+1)), ncol=(n+1))
crW = numeric((n+1))

set.seed(1986)
for(i in 1:10000){
  for(j in 0:n) {
    p_hat[i,j] = mean(rbinom(n, 1, j/n)) # mle
  }
}
  p_hat = cbind(p_hat[,(n+1)], p_hat[,1:n]) # reorder columns

  for(i in 1:10000){
    for(j in 1:(n+1)) {

  # compute Wilson approximate CI (based on normal distribution)
   ci_lowerW[i,j] = approximate.ci.binom(n = n, k = n*p_hat[i,j], alpha=alpha)$pl
   ci_upperW[i,j] = approximate.ci.binom(n = n, k = n*p_hat[i,j], alpha=alpha)$pu

    if (p[i,j] >= ci_lowerW[i,j] & p[i,j] <= ci_upperW[i,j])  crW[j]=crW[j]+1
    }
}
```

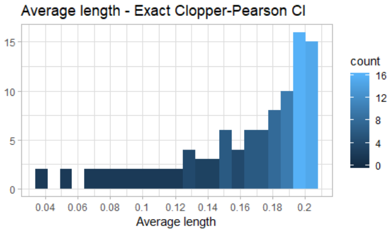
R code to generate plots and compute the average length of the CI

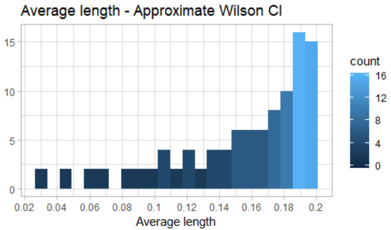
```
# coverage rate
crW/rep(10000,(n+1))
# [1] 1.0000 0.9960 0.9960 0.9432 0.9794 0.9484 0.9665 0.9697 0.9499 0.9645 ...
# [99] 0.9964 0.9973 1.0000
# plot with ggplot2
library("ggplot2")
x = k/n
y = crW/rep(10000,(n+1))
qplot(x,y, geom='smooth', span =0.15, color="darkred") +
  ggtitle("Empirical Coverage Rate ...") +
  xlab("p") + ylab("Empirical Coverage Rate") +
  ylim(0.9, 1) +
  theme_light()

# average length of the confidence intervals
CI_lengthW=matrix(rep(0, 2*(n+1)), ncol=2)
for(a in 1:(n+1)) {
  CI_lengthW[a, ] = apply(cbind(ci_lowerW[,a], ci_upperW[,a]),2, mean)
}
av_lengthW <- CI_lengthW[,2] - CI_lengthW[,1]
mean(av_lengthW)  # [1] 0.1534113

# Plot of average length
qplot(av_lengthW, geom="histogram", fill=..count..,
          bins = length(seq(0.02,0.24, by=0.01)+2)) +
  ggtitle("Histogram of average length interval - Approximate Wilson CI (0.1534)") +
  scale_x_continuous(name = "Average length", breaks = seq(0.02,0.22, by=0.01),
                     labels = seq(0.02,0.22, by=0.01)) +
  theme_light()
```

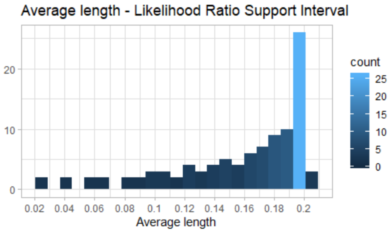
Behaviour of the empirical coverage rate for the different CI

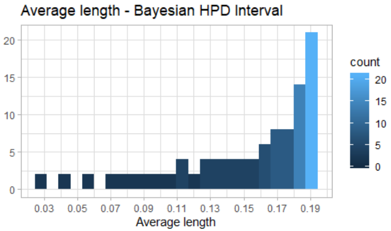


Average length of the CI









References

C. Dalitz, Construction of Confidence Intervals (2017), Technical Report No. 2017-01, pp. 15-28, Hochschule Niederrhein, Fachbereich Elektrotechnik Informatik (2017) Construction of Confidence Intervals Christoph Dalitz

The R Project for Statistical Computing: https://www.r-project.org/