Homework 4

PSTAT 115, Fall 2021

Due on November 28, 2021 at 11:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

Problem 1. Frequentist Coverage of The Bayesian Posterior Interval.

In quiz 1 we explored the importance and difficulty of well-calibrated prior distributions by examining the calibration of subjective intervals. Suppose that $y_1, .., y_n$ is an IID sample from a $Normal(\mu, 1)$. We wish to estimate μ.

1a. For Bayesian inference, we will assume the prior distribution $\mu \sim Normal(0, \frac{1}{\kappa_0})$ for all parts below. Remember, from lecture that we can interpret κ_0 as the pseudo-number of prior observations with sample mean $\mu_0 = 0$. State the posterior distribution of μ given $y_1, .., y_n$. Report the lower and upper bounds of the 95% quantile-based posterior credible interval for μ, using the fact that for a normal distribution with standard eviation σ, approximately 95% of the mass is between $\pm 1.96\sigma$.

Our Posterior is: $e^{-\frac{1}{\sigma^2 **}\mu **^2}$

where $\sigma^2 * * = \frac{\frac{1}{k_0}}{1+n\frac{1}{k_0}}$

and $\mu * * = \frac{\frac{1}{k_0}\sum y_i}{1+n\frac{1}{k_0}}$

So, our Posterior credible interval is: $(\mu ** - 1.96 * \sqrt{\sigma^2 * *}, \mu ** + 1.96 * \sqrt{\sigma^2 * *})$

1b. Plot the length of the posterior credible interval as a function of κ_0, for $\kappa_0 = 1, 2, ..., 25$ assuming $n = 10$. Report how this prior parameter effects the length of the posterior interval and why this makes intuitive sense.

```
set.seed(120)
n <- 1:10
k_0 <- 1:25
y_i <- 1:10
CI_interval <- 1:25
CI_length <- 1:25
upper <- 1:25
lower <- 1:25
for (i in k_0){
    mu <- rnorm(1,mean=0,sd=1/i)
    y_i <- rnorm(10,mean=mu,sd=1)
    mu_star <- (1/i*sum(y_i))/(1+10*(1/i))
    sigma_star <- (1/i)/(1+10*(1/i))
    upper[i] <- mu_star+1.96*sigma_star
```

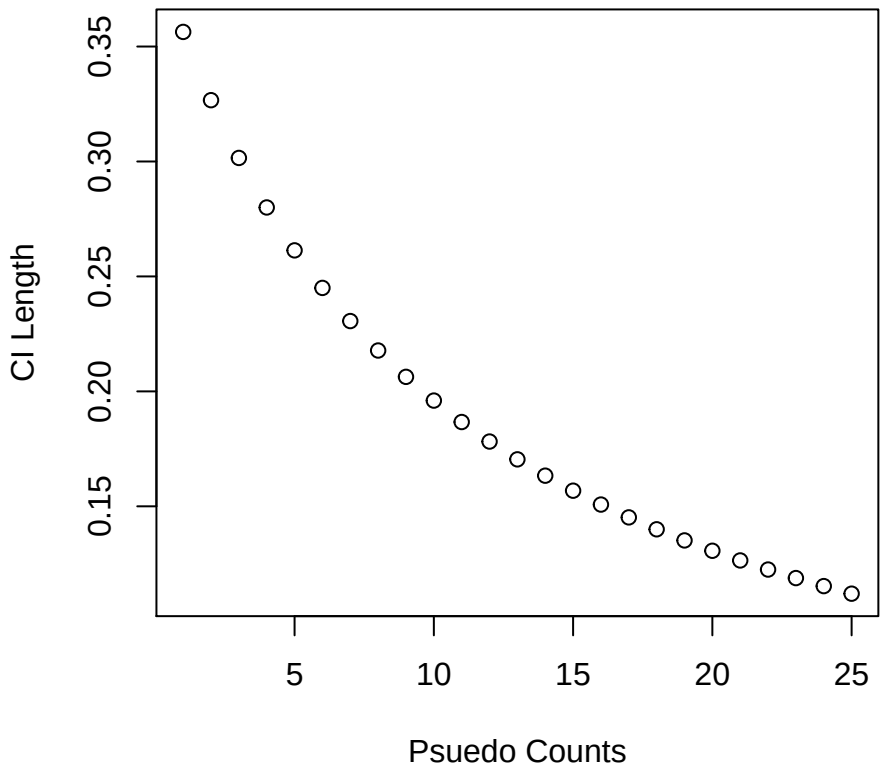
```
    lower[i] <- mu_star-1.96*sigma_star
    CI_length[i] <- upper[i] - lower[i]
}
  CI_length
```

```
##  [1] 0.3563636 0.3266667 0.3015385 0.2800000 0.2613333 0.2450000 0.2305882
##  [8] 0.2177778 0.2063158 0.1960000 0.1866667 0.1781818 0.1704348 0.1633333
## [15] 0.1568000 0.1507692 0.1451852 0.1400000 0.1351724 0.1306667 0.1264516
## [22] 0.1225000 0.1187879 0.1152941 0.1120000
```

```
  for (i in 1:25){
      print(c(upper[i],lower[i]))
  }
```

```
## [1]  0.30806267 -0.04830096
## [1] 0.39657456 0.06990789
## [1]  0.1675637 -0.1339747
## [1] 0.29715017 0.01715017
## [1] 0.262610551 0.001277218
## [1]  0.21659727 -0.02840273
## [1] -0.5463360 -0.7769242
## [1]  0.18774658 -0.03003119
## [1] 0.214229680 0.007913891
## [1]  0.11038071 -0.08561929
## [1]  0.01142652 -0.17524015
## [1]  0.06179188 -0.11638994
## [1]  0.08887150 -0.08156328
## [1]  0.12420206 -0.03913127
## [1]  0.146940409 -0.009859591
## [1] 0.1680432 0.0172740
## [1] 0.16644337 0.02125818
## [1] -0.006314136 -0.146314136
## [1] 0.16867817 0.03350575
## [1]  0.01094867 -0.11971799
## [1] -0.08836855 -0.21482016
## [1]  0.004160818 -0.118339182
## [1] 0.2968289 0.1780410
## [1] 0.18230873 0.06701461
## [1]  0.102503563 -0.009496437
```

```
plot(k_0,CI_length,xlab="Psuedo Counts",ylab="CI Length")
```

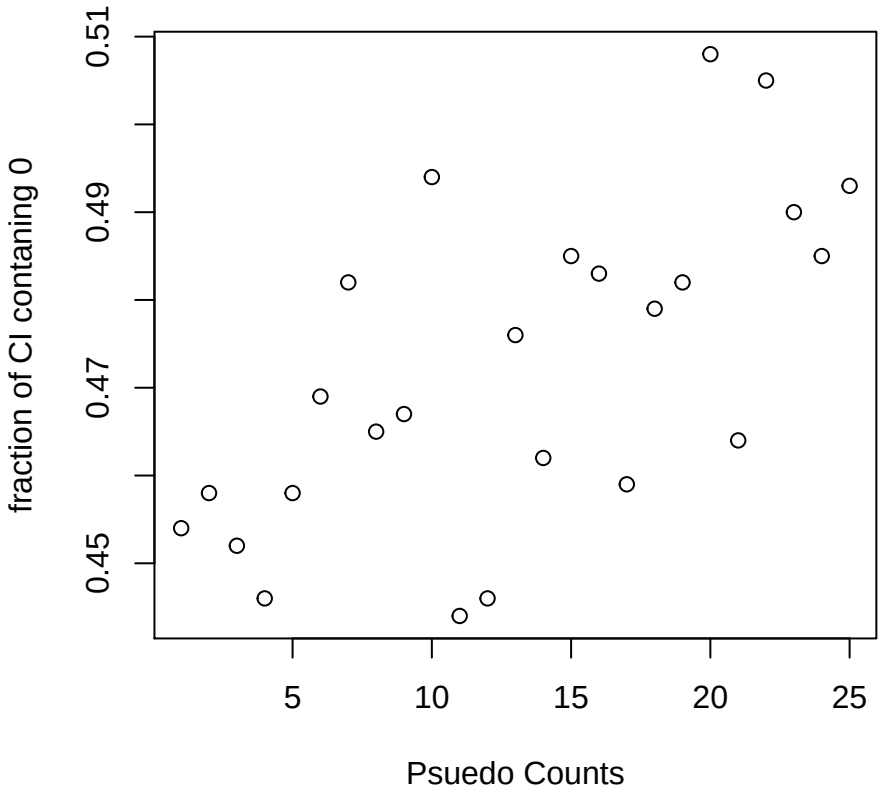


The length of the posterior credible interval shrinks as k_0 grows. This makes sense as it directly relates to the variance. As the variance gets smaller and smaller it will result in the data having less variance pushing the mean towards 0.

1c. Now we will evaluate the *frequentist coverage* of the posterior credible interval on simulated data. Generate 1000 data sets where the true value of $\mu = 0$ and $n = 10$. For each dataset, compute the posterior 95% interval endpoints (from the previous part) and see if it the interval covers the true value of $\mu = 0$. Compute the frequentist coverage as the fraction of these 1000 posterior 95% credible intervals that contain $\mu = 0$. Do this for each value of $\kappa_0 = 1, 2, ..., 25$. Plot the coverage as a function of κ_0.

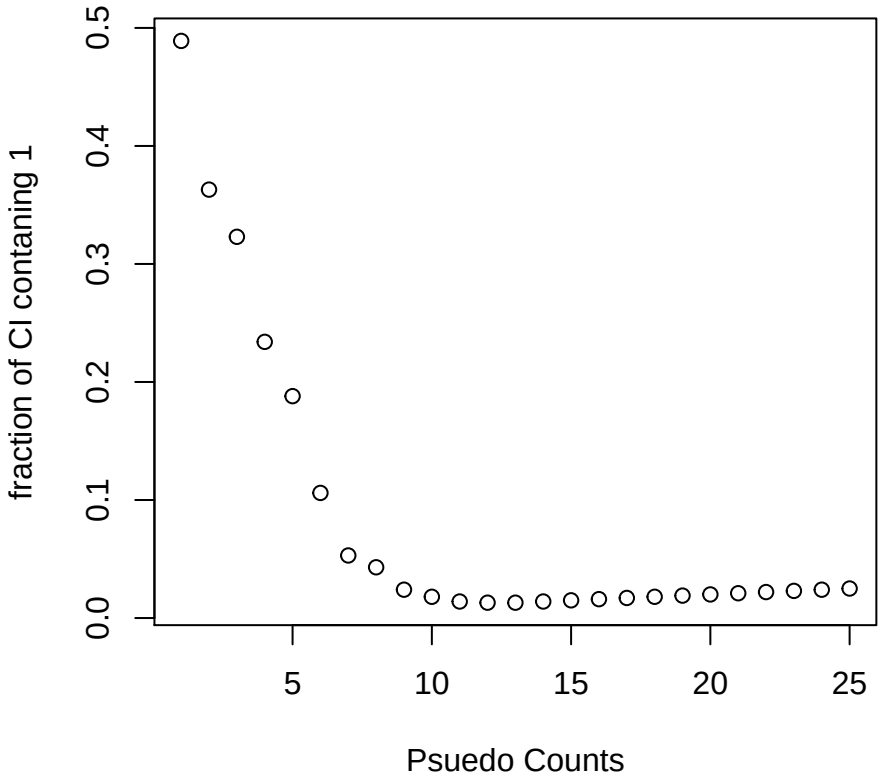
```
set.seed(150)
S <- 1:1000
mu <- 0
CI_contain_zero <- 1:25
for (i in k_0){
        for (s in S){
        y_i <- rnorm(10,mean=mu,sd=1)
        mu_star <- (1/i*sum(y_i))/(1+10*(1/i))
        sigma_star <- (1/i)/(1+10*(1/i))
        upper[i] <- mu_star+1.96*sigma_star
        lower[i] <- mu_star-1.96*sigma_star
        if(upper[i] > 0 && lower[i] < 0){
            CI_contain_zero[i] <- sum(CI_contain_zero[i])+1
        }
```

```
}
}
plot(k_0,CI_contain_zero/1000,xlab="Psuedo Counts",ylab="fraction of CI contaning 0")
```



1d. Repeat the 1c but now generate data assuming the true $\mu = 1$.

```
set.seed(100)
S <- 1:1000
mu <- 1
CI_contain_zero <- 1:25
for (i in k_0){
        for (s in S){
        y_i <- rnorm(10,mean=mu,sd=1)
        mu_star <- (1/i*sum(y_i))/(1+10*(1/i))
        sigma_star <- (1/i)/(1+10*(1/i))
        upper[i] <- mu_star+1.96*sigma_star
        lower[i] <- mu_star-1.96*sigma_star
        if(upper[i] > 1 && lower[i] < 1){
            CI_contain_zero[i] <- sum(CI_contain_zero[i])+1
        }
}
}
plot(k_0,CI_contain_zero/1000,xlab="Psuedo Counts",ylab="fraction of CI contaning 1")
```



1e. Explain the differences between the coverage plots when the true $\mu = 0$ and the true $\mu = 1$. For what values of κ_0 do you see closer to nominal coverage (i.e. 95%)? For what values does your posterior interval tend to overcover (the interval covers the true value more than 95% of the time)? Undercover (the interval covers the true value less than 95% of the time)? Why does this make sense?

As the values of k_0 grow, we can see that the values start to have nominal coverage. When the mu=0 we see the values tend to overcover where as when mu=1 the values tend to undercover. This makes sense as if the average stays the same while the variance shrinks, it will shrink towards zero, making the mu=0 appear in more intervals where as mu=1 will slowly shrink it out of the interval