# Big Data
# Solutions: sampling for frequent itemset mining

**Exercise 1**

(a) We have $\hat{\pi}_Z > 0.2$ if $m > 20$. With $\pi_Z = 0.1$ and $n = 100$ the probability is:

```
> pbinom(20,100,.1, lower.tail=F)
[1] 0.0008075739
```

(b) Since $\pi_Z = 0.1$, we have that if $\hat{\pi}_Z > 0.2$, then $\hat{\pi}_Z - \pi_Z > 0.1$. This is the information we have to use to exploit the Hoeffding inequality. We get:

$$\mathbb{P}(\hat{\pi}_Z - \pi_Z > 0.1) \leq e^{-2\times 0.1^2 \times 100} = e^{-2} \approx 0.1353$$

Notice that since we bound the difference between $\hat{\pi}_Z$ and $\pi_Z$ in only one direction, we take only half the value of the two-sided bound.

The bound provided by the Hoeffding inequality is about 169 times as big as the actual binomial probability.

(c) To compute the exact binomial probability, we need to know the true support $\pi_Z$. If we knew the true support, then there would be no reason to mine for frequent itemsets in the first place. The Hoeffding bound, on the other hand, does not depend on $\pi_Z$.

**Exercise 2**

We start with:

$$\delta \geq 2e^{-2\epsilon^2 n}.$$

Division by 2 gives:

$$\frac{\delta}{2} \geq e^{-2\epsilon^2 n}.$$

Taking the natural log of both sides we obtain:

$$\ln \frac{\delta}{2} \geq -2\epsilon^2 n.$$

Notice that the inequality sign remained in the same direction because the logarithmic function is monotonically increasing. Finally, to isolate $n$, we divide both sides by $-2\epsilon^2$ and obtain:

$$\frac{-\ln \frac{\delta}{2}}{2\epsilon^2} \leq n$$

Note that since we divided by a negative number, the inequality sign has flipped. Using the properties of logarithms, we can rewrite $-\ln \frac{\delta}{2}$ as $\ln \frac{2}{\delta}$, to finally obtain:

$$n \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}.$$

**Exercise 3**

(a) Since with $k$ items, there are $2^k$ itemsets, the union bound combined with the Hoeffding bound gives:

$$\mathbb{P}\left( \bigcup_{Z \subseteq \{1,\ldots,k\}} |\pi_Z - \hat{\pi}_Z| > \epsilon \right) \leq 2^k 2e^{-2\epsilon^2 n} = 2^{k+1} e^{-2\epsilon^2 n}$$

This gives (see exercise 2):

$$n \geq \frac{1}{2\epsilon^2} \ln \frac{2^{k+1}}{\delta}.$$

(b) $n \geq 199,779$.

(c) The assumption is not realistic at all. For example, one itemset could be a subset of the other. The event that we mis-estimate the support of an itemset by more than $\epsilon$, will be positively correlated to the event that we mis-estimate the support of one of its subsets by more than $\epsilon$.

(d) The Hoeffding inequality states that

$$\mathbb{P}(|\pi_Z - \hat{\pi}_Z| > \epsilon) \leq 2e^{-2\epsilon^2 n}$$

It follows that

$$\mathbb{P}(|\pi_Z - \hat{\pi}_Z| \leq \epsilon) \geq 1 - 2e^{-2\epsilon^2 n}$$

Also, we have

$$\mathbb{P}\left( \bigcup_{Z \subseteq \{1,\ldots,k\}} |\pi_Z - \hat{\pi}_Z| > \epsilon \right) = 1 - \mathbb{P}\left( \bigcap_{Z \subseteq \{1,\ldots,k\}} |\pi_Z - \hat{\pi}_Z| \leq \epsilon \right)$$

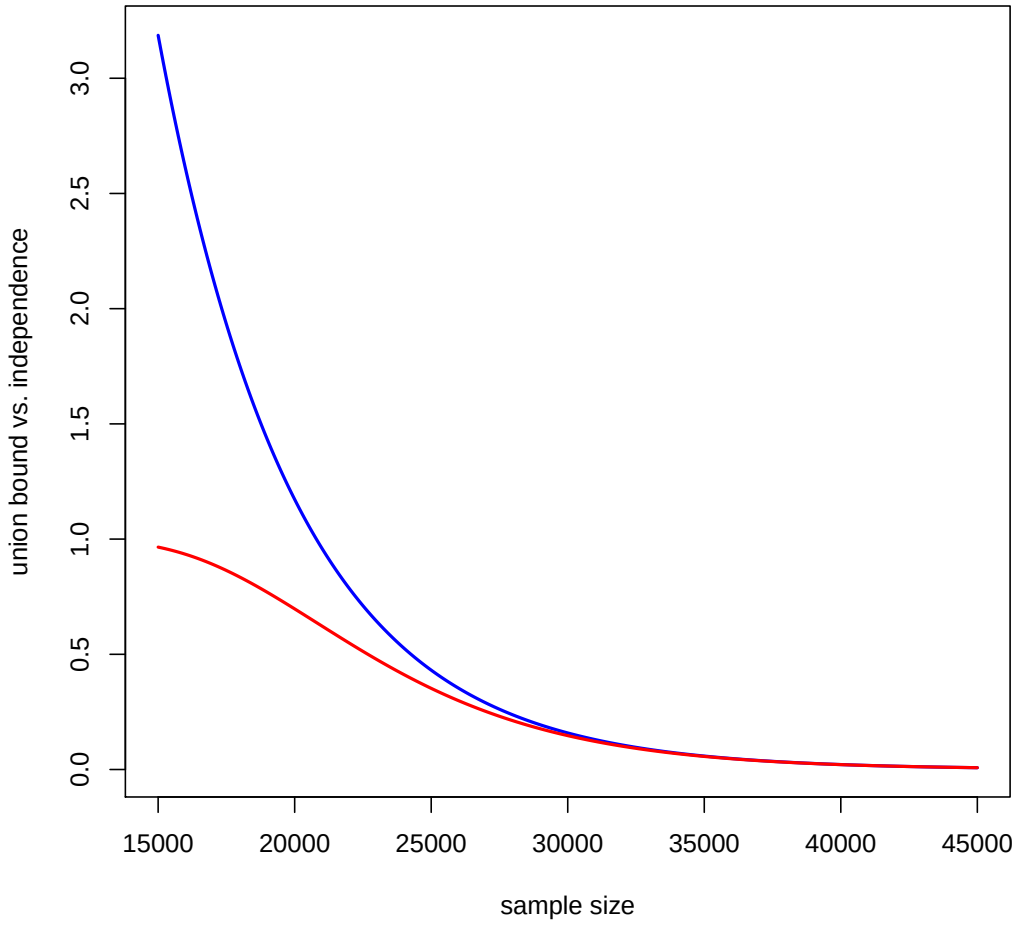
Recall that if $A$ and $B$ are independent events, then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Using the Hoeffding inequality, and the independence assumption, we thus have

$$\mathbb{P}\left( \bigcap_{Z \subseteq \{1,\ldots,k\}} |\pi_Z - \hat{\pi}_Z| \leq \epsilon \right) \geq (1 - 2e^{-2\epsilon^2 n})^{2^k}$$

Finally, we can conclude that

$$\mathbb{P}\left(\bigcup_{Z\subseteq\{1,\ldots,k\}} |\pi_Z - \hat{\pi}_Z| > \epsilon\right) \leq 1 - (1 - 2e^{-2\epsilon^2 n})^{2^k}$$

When we plot the two bounds for $\epsilon = 0.01$, $k = 5$, and $n$ in the range from $15,000$ to $45,000$, we get the following graph




The blue curve gives the union bound, and the red curve gives the bound based on the independence assumption. Although the bound that uses the independence assumption is always smaller, it appears that in the range of acceptable values for $\delta$, the difference between the bounds is negligible.

**Exercise 4: Lowering the threshold**

The formula for the lowered threshold $t'$ is:

$$t' = t - \sqrt{\frac{1}{2n} \ln \frac{1}{\mu}}$$

With $t = 0.01$, $n = 100,000$, and $\mu = 0.001$ we find that

$$t' = 0.01 - \sqrt{\frac{1}{200,000} \ln \frac{1}{0.001}} = 0.01 - 0.006 = 0.004$$

Now with the normal approximation of the binomial distribution. We want:

$$\mathbb{P}(\hat{\pi} < t') \leq \mu$$

We have

$$\begin{aligned} \mathbb{P}(\hat{\pi} < t') &= \mathbb{P}\left( \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} < \frac{t' - \pi}{\sqrt{\pi(1-\pi)/n}} \right) \\ &= \mathbb{P}\left( Z < \frac{t' - \pi}{\sqrt{\pi(1-\pi)/n}} \right), \end{aligned}$$

where $Z$ is a standard normal random variable. Define $\mathbb{P}(Z \leq z_\mu) = \mu$. Filling in the pessimistic value $\pi = 0.5$ and putting

$$\frac{t' - \pi}{\sqrt{0.25/n}} = z_\mu$$

we obtain

$$t' = \pi + z_\mu \sqrt{\frac{0.25}{n}}$$

Following through we get (fill in $\pi = t$):

$$t' = t + z_{0.001} \sqrt{\frac{0.25}{n}} = 0.01 - 3.09 \sqrt{\frac{0.25}{100,000}} = 0.01 - 0.00489 = 0.00511$$