ST 502 HW 3 Extra Problem 1

Bruce Campbell

January 29, 2017

```
rm(list = ls())
knitr::opts_chunk$set(dev = 'png')
knitr::opts_chunk$set(fontsize=13)
knitr::opts_chunk$set(dpi=600)
knitr::opts_chunk$set(cache=TRUE)
knitr::opts_chunk$set(tidy=TRUE)
knitr::opts_chunk$set(prompt=FALSE)
knitr::opts_chunk$set(echo=FALSE)
knitr::opts_chunk$set(fig.height=6)
knitr::opts_chunk$set(fig.width=6)
knitr::opts_chunk$set(warning=FALSE)
knitr::opts_chunk$set(message=FALSE)
```

Extra Problem 1: Using the CHIS data set on the website

- (a) Find the true proportion of Asian students in the population and the true variance.
- (b) Use R to generate 5000 samples each of size n=8 from the 'Asian' indicator variable in the CHIS dataset.
- (c) Create a large sample (use normal approximation to binomial, aka CLT, with fnite sample correction) confdence interval for p for each sample. Be sure to use the estimated standard error (see page 214) rather than the truth.
- (d) Report the fraction of intervals that contained p and the average length.
- (e) Repeat the above 2 steps for n=50, n=100, n=1000. Report the fraction of intervals that contain p for each interval as well as the average width of the intervals.
- (f) What do you notice about the observed coverage probability as n grows? Give an explanation.

```
##        X1             Height          Weight           BMI
##  Min.   :   1.0   Min.   :46.00   Min.   : 50.0   Min.   : 8.94
##  1st Qu.: 700.5   1st Qu.:61.00   1st Qu.:110.0   1st Qu.:19.14
##  Median :1400.0   Median :64.00   Median :125.0   Median :21.45
##  Mean   :1400.0   Mean   :64.44   Mean   :131.5   Mean   :22.28
##  3rd Qu.:2099.5   3rd Qu.:68.00   3rd Qu.:150.0   3rd Qu.:24.29
##  Max.   :2799.0   Max.   :77.00   Max.   :240.0   Max.   :51.80
##      Asian
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.09539
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

Table 1: True proportion and variance

p	Var
0.09539	0.08629

Generate CI's

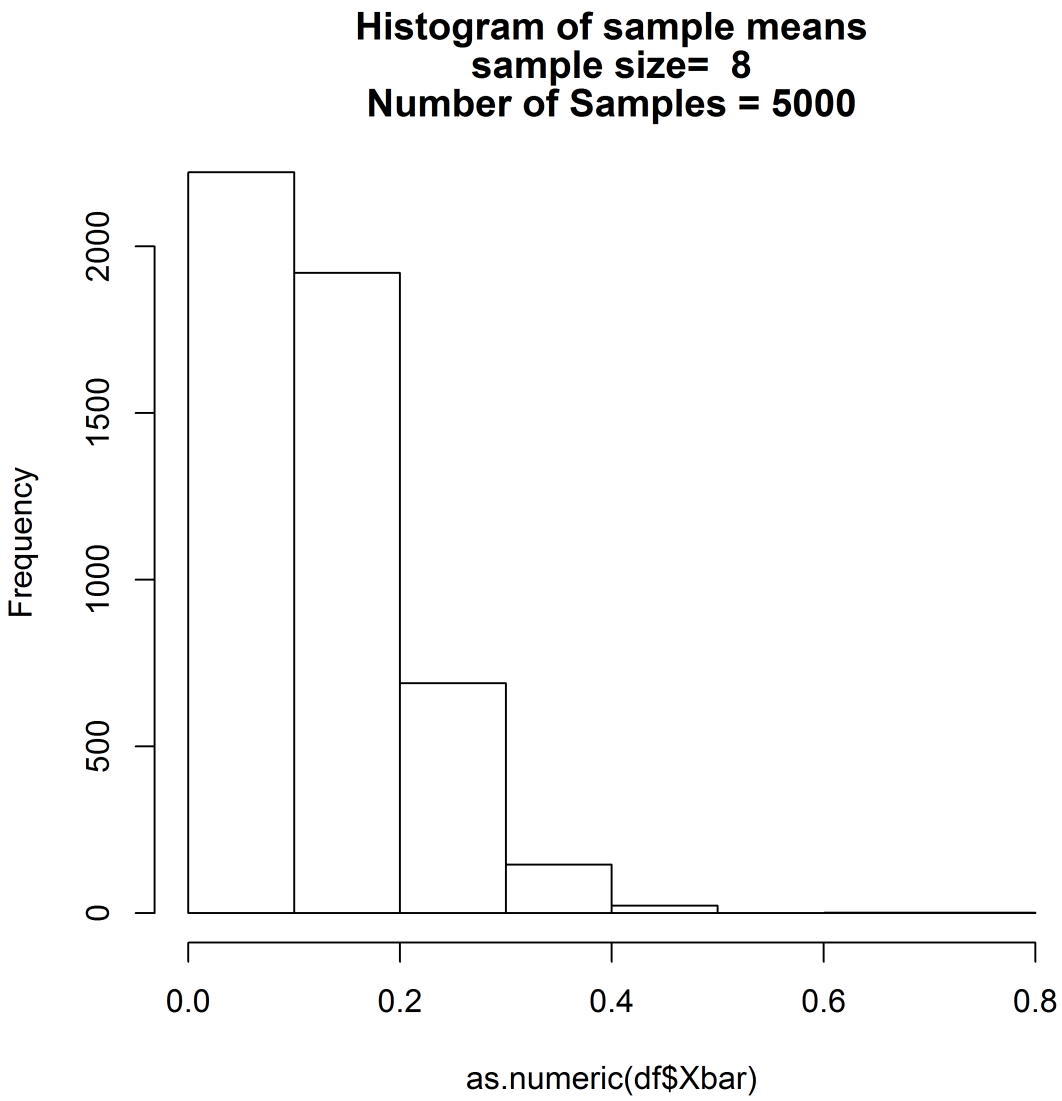
- (c) Create a large sample (use normal approximation to binomial, aka CLT, with fnite sample correction) confdence interval for p for each sample. Be sure to use the estimated standard error (see page 214) rather than the truth.
- (d) Report the fraction of intervals that contained p and the average length.Use R to generate 5000 samples each of size n=8 from the 'Asian' indicator variable in the CHIS dataset.

The $1-\alpha$ CI for $\hat{p}$ is given by $\bar{X} \pm z(\frac{\alpha}{2})s^2_{\hat{p}}$. Where $s^2_{\hat{p}}$ is the estimage SE given by

$S^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}(1-\frac{n}{N})$

Here we choose an α of 0.05 which yields a 95% CI for p

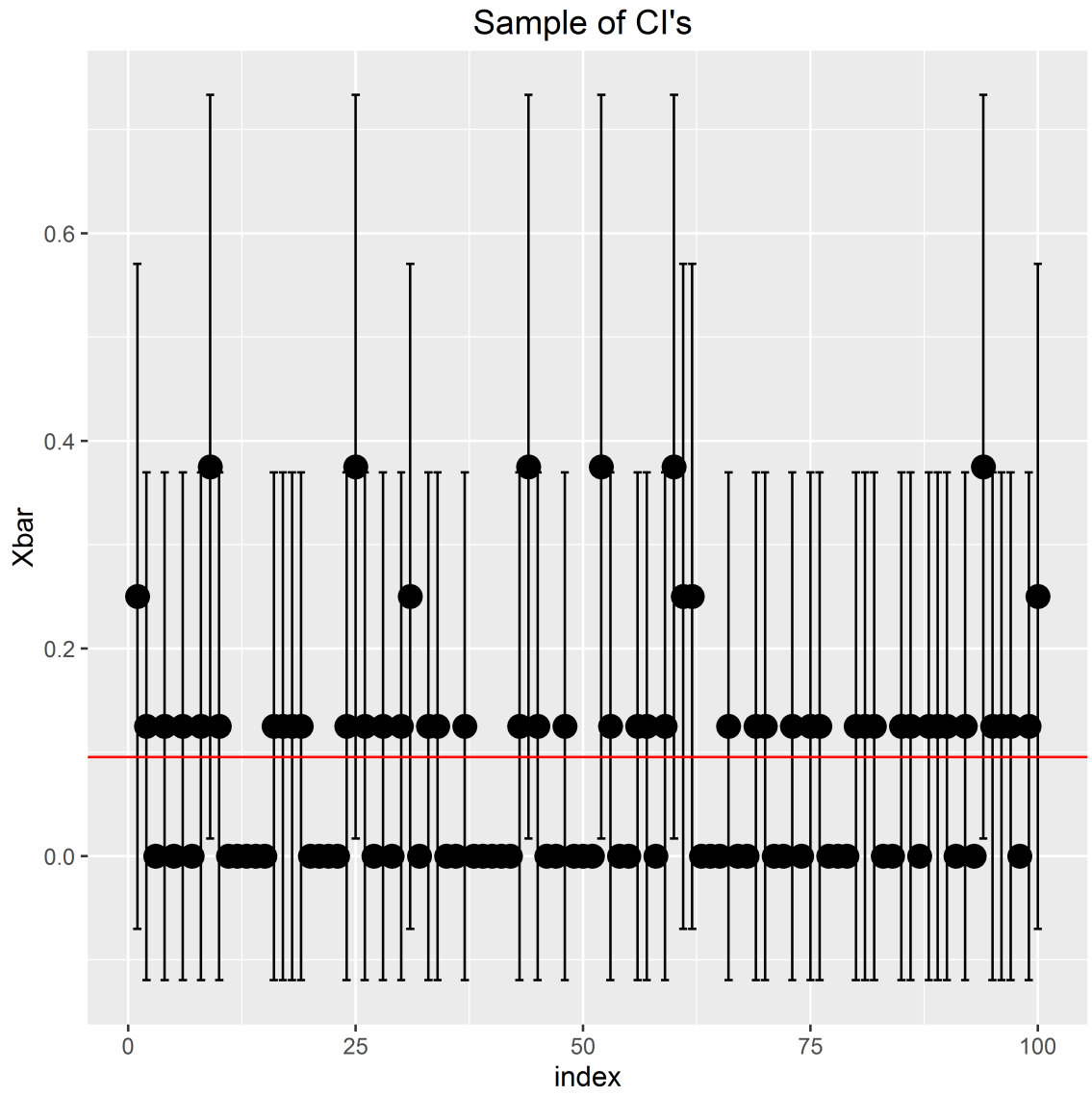
```
## [1] 8
## [1] 5000
```



proportion

0.5508

Table: Proportion of CI's containing the true mean

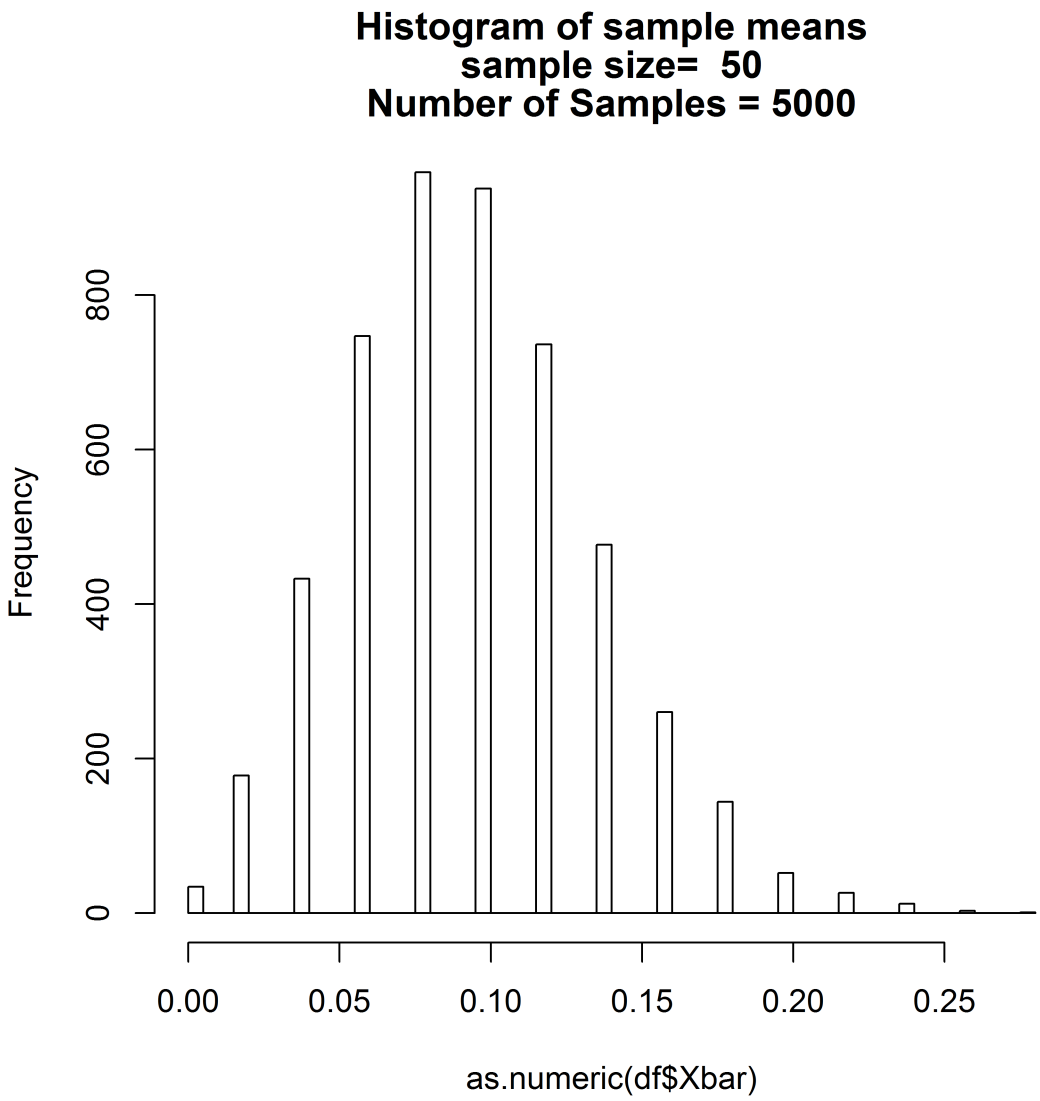


Repeat the above 2 steps for n=50, n=100, n=1000. Report the fraction of intervals

that contain p for each interval as well as the average width of the intervals.

N=50

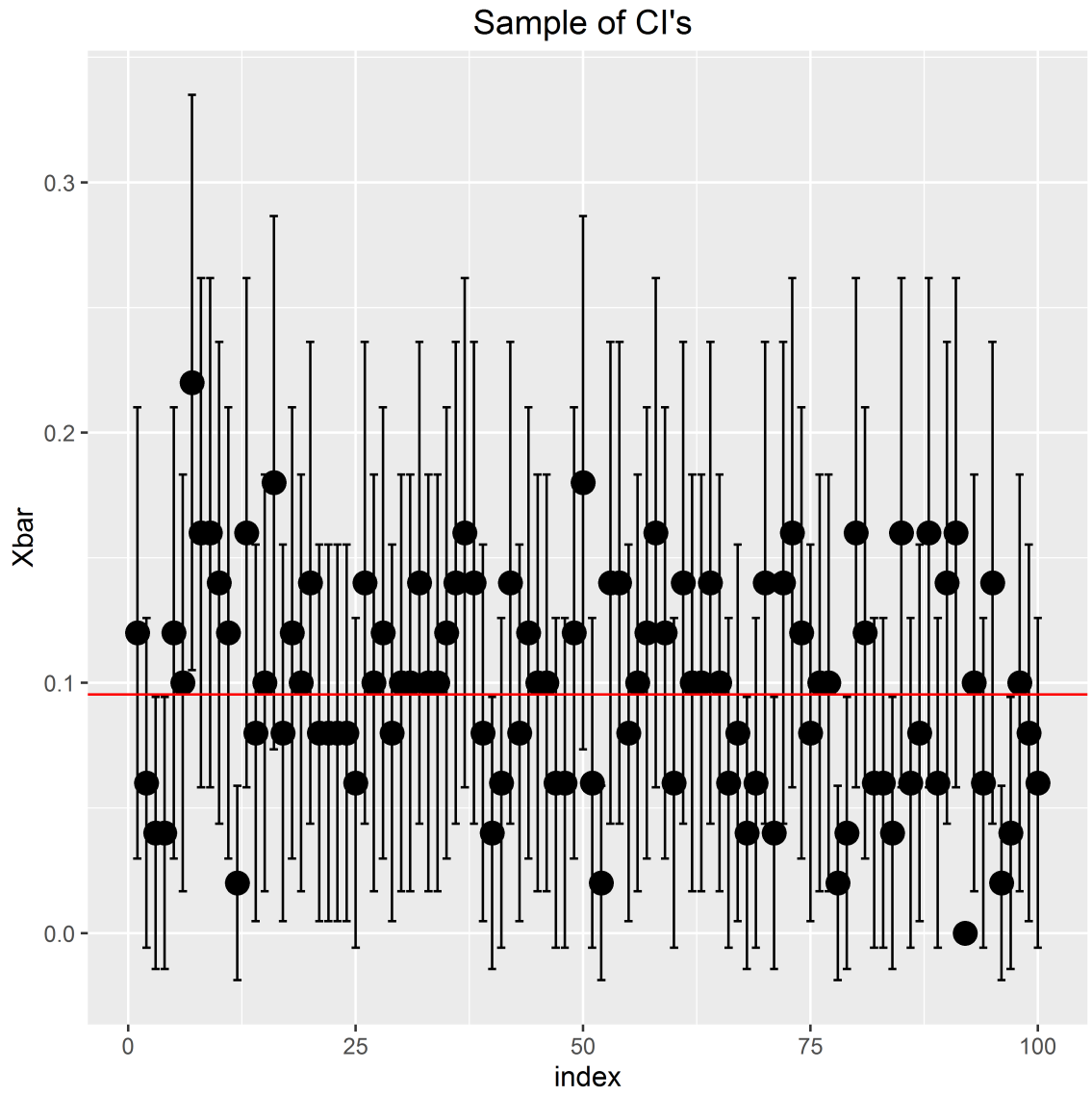
```
## [1] 50
## [1] 5000
```



proportion

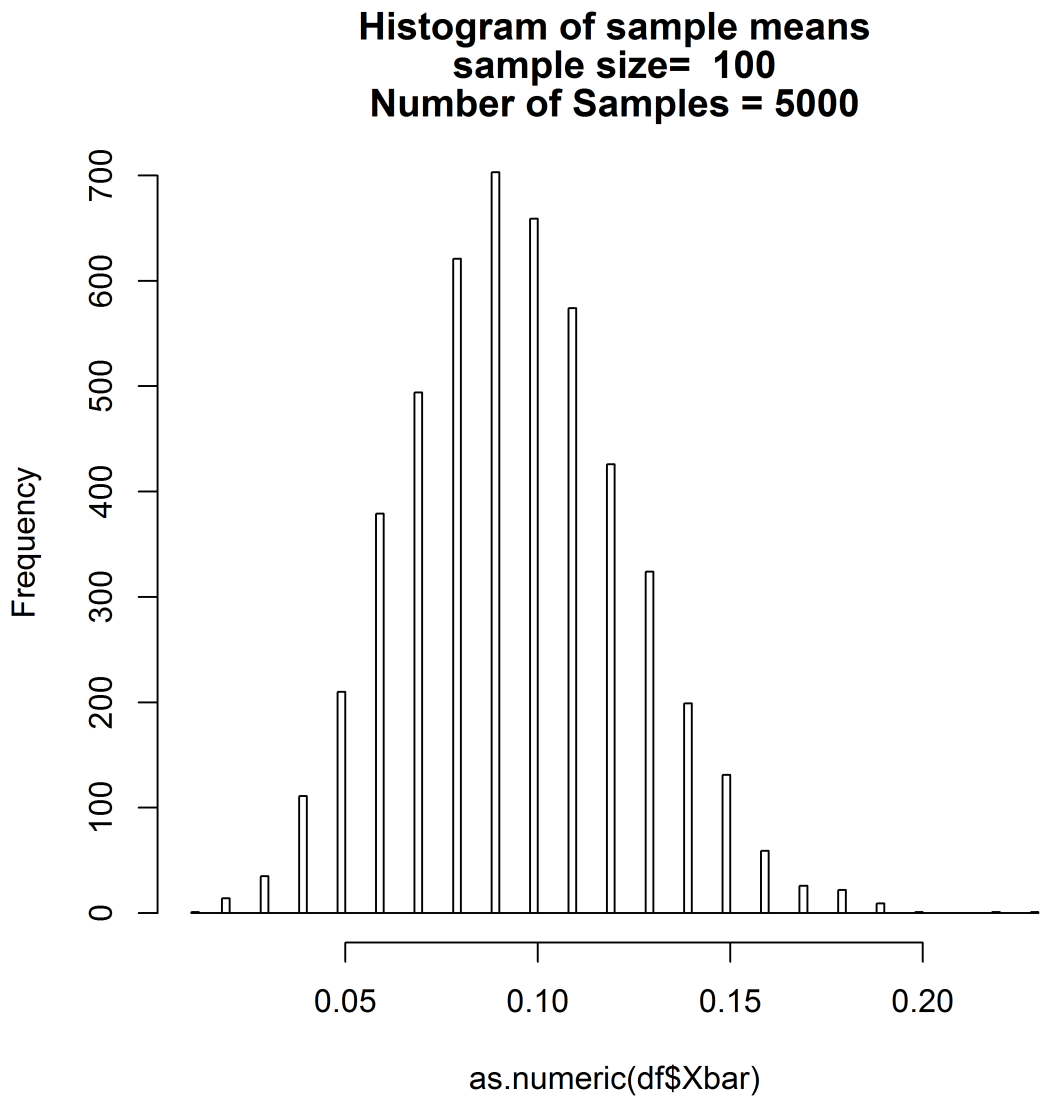
0.8626

Table: Proportion of CI's containing the true mean



N=100

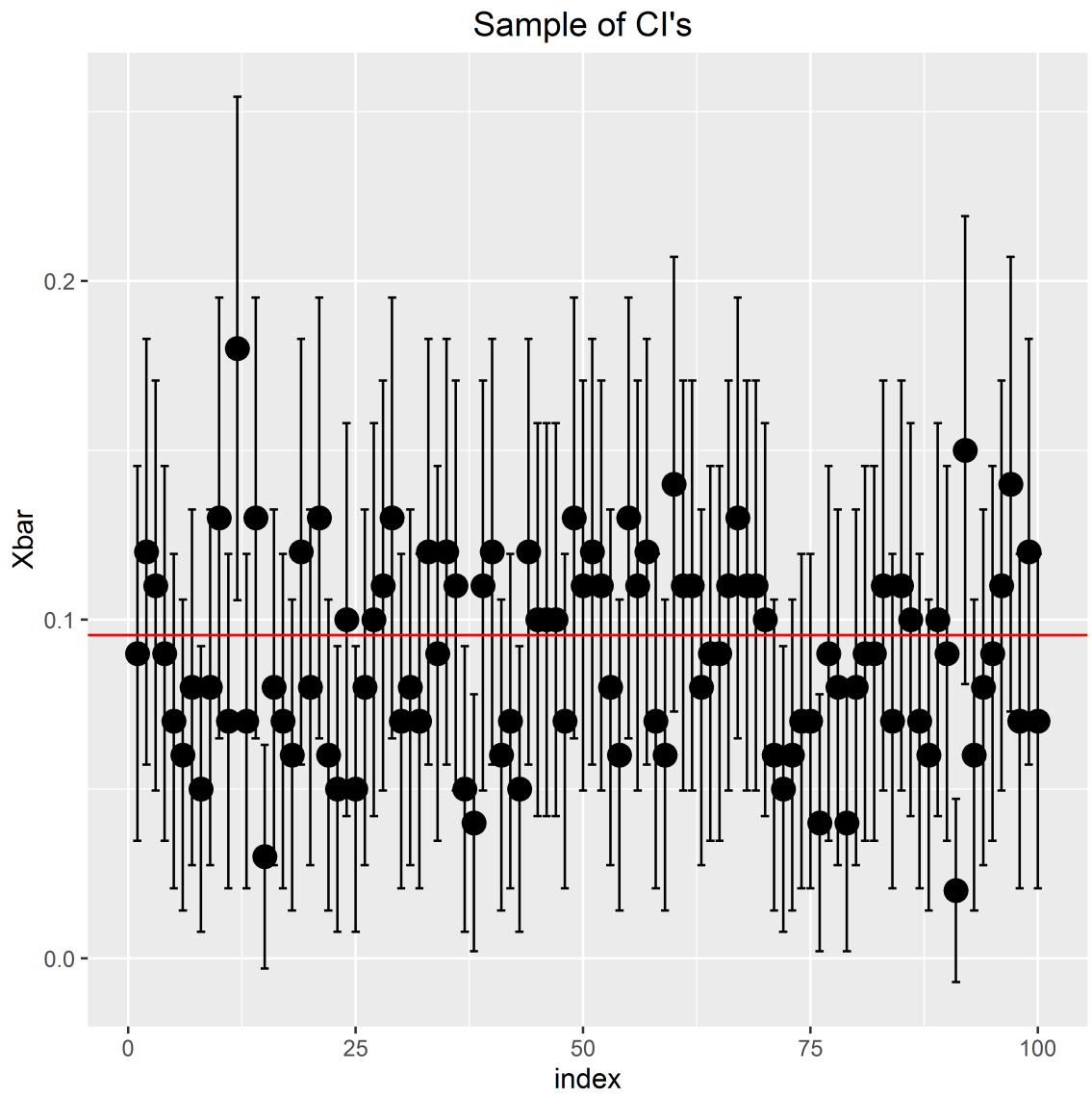
```
## [1] 100
## [1] 5000
```



proportion

0.9138

Table: Proportion of CI's containing the true mean



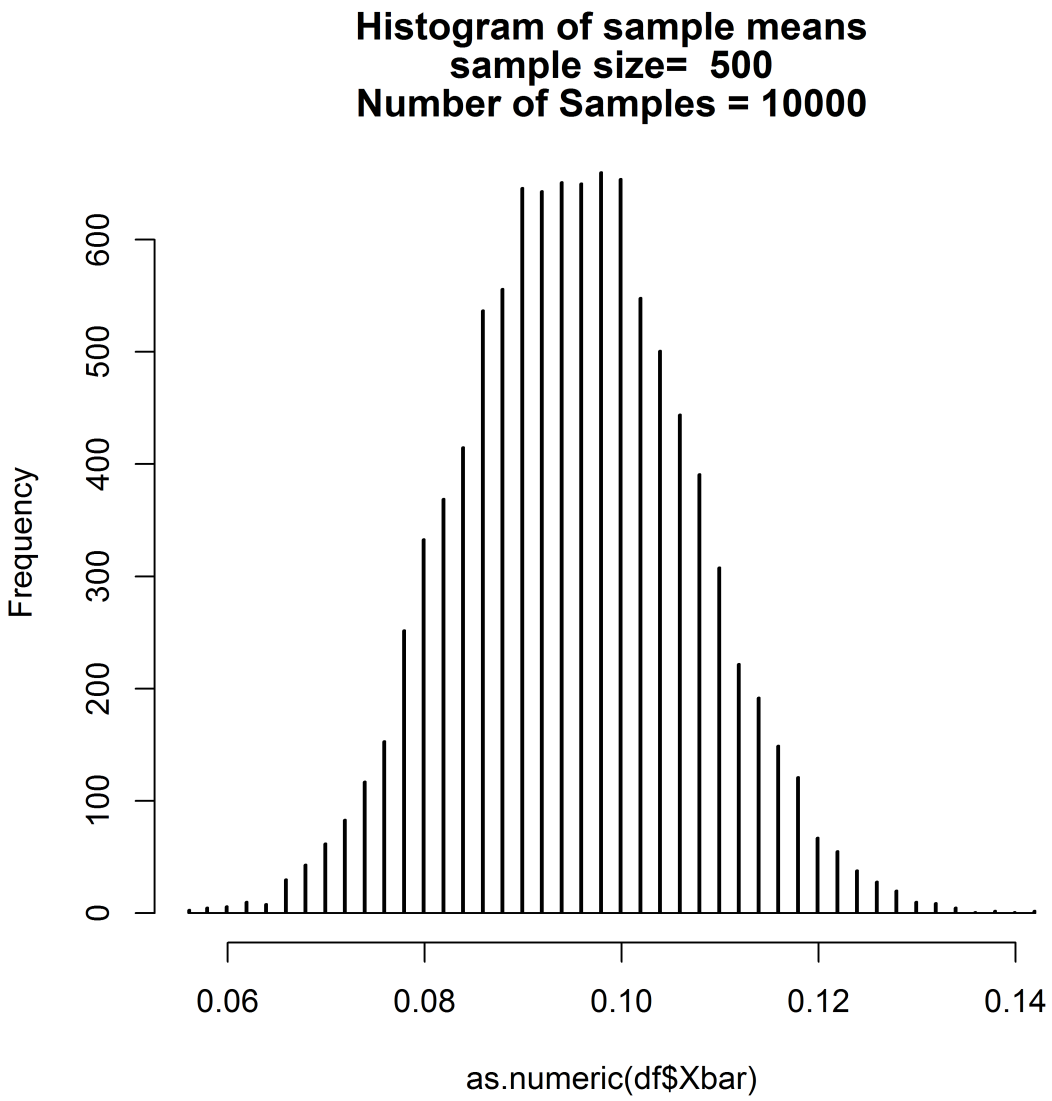
What do you notice about the observed coverage probability as n grows? Give an

explanation.

The coverage probability increases as the sample size grows. This is due to the the convergence of the sample mean to the true mean and the convergence of the sampling distribution of the sample mean to the normal distribution. This convergence of the sampling distribution is what makes the confidence interval limits accurate. By the design of the randmo confidance interval we expect the coverage to converge to .95 as the sample size grows and the number of samples increases.

For illustration we calculate one more simulation with a large sample size and a high number of samples.

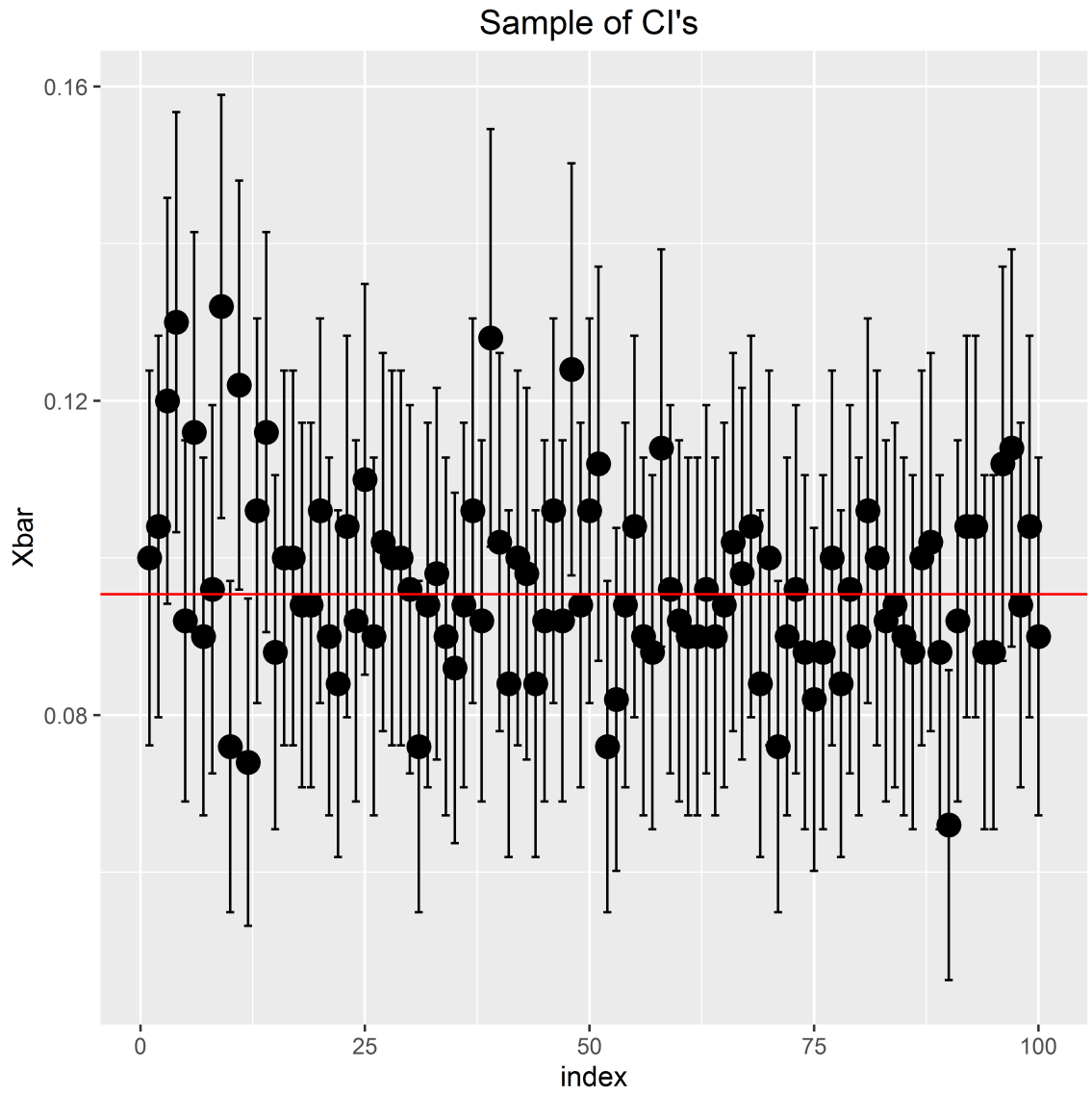
```
## [1] 500
## [1] 10000
```



proportion

0.9462

Table: Proportion of CI's containing the true mean


Sample of CI's
0.16
0.12
0.08
Xbar
0
25
50
75
100
index