

## HW19 due Tues May 25

Topics: Sampling distributions, effect of sample size, Central Limit Theorem

1. Reconsider Example 19-2, in which we used R to simulate fast-food service times and investigate the sampling distributions of the sample mean.

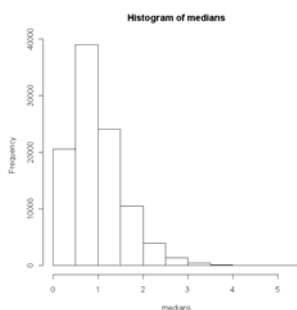
a) Determine the median of an exponential distribution with mean 1.25 minutes. In other words, determine the value  $m$  such that  $\Pr(X \leq m) = \Pr(X \geq m) = .5$ , where  $X$  has an exponential distribution with mean 1.25.

The exponential parameter is  $\lambda = 1/1.25 = 0.8$ . The cdf is  $F(x) = 1 - e^{-0.8x}$ . Setting  $F(m) = 0.5$  gives  $1 - e^{-0.8m} = .5$ , so  $e^{-0.8m} = .5$ , so  $-0.8m = \ln(.5)$ , so  $m = -\ln(.5)/.8 \approx .866$ .

b) Write R code that simulates  $N$  samples of size  $n$  from an exponential distribution with mean 1.25 minutes. This time keep track of the sample *median* rather than the sample mean or sample maximum. [Note: The sample median is defined to be the middle value if there are an odd number of observations, and it is defined to be the average of the two middle values if there are an even number of observations. The R command to calculate the sample median from the values in a vector is simply: `median(vectorname)`.] Your code should also produce a histogram of the  $N$  sample medians and calculate the mean and SD of those sample medians. Submit your code.

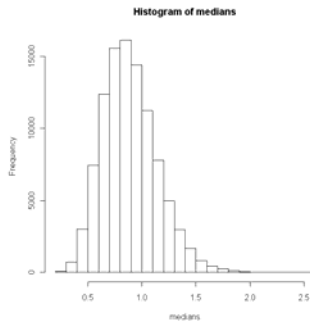
```
times = rep(NA, times = n)
medians = rep(NA, times = N)
for (i in 1:N) {
  times = rexp(n,lambda)
  medians[i] = median(times)
}
hist(medians); mean(medians); sd(medians)
```

c) Run your R code for 100,000 samples of size  $n = 5$ . Submit a histogram of the resulting sample medians, along with their mean and SD.



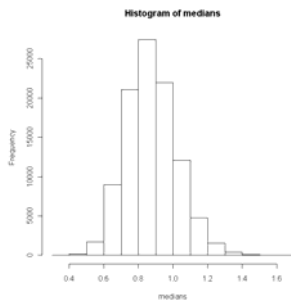
Mean of sample medians  $\approx 0.978$ , SD of sample medians  $\approx 0.577$

d) Repeat c) for 100,000 samples of size  $n = 25$ , and then for 100,000 samples of size  $n = 75$ .



$n = 25$ :  
0.252

Mean of sample medians  $\approx 0.890$ , SD of sample medians  $\approx$



$n = 75$ :  
0.145

Mean of sample medians  $\approx 0.874$ , SD of sample medians  $\approx$

e) Does the sampling distribution of the sample median appear to be approximately normal? For all sample sizes or only for large sample sizes? Explain briefly.

The sampling distribution is not very normal for  $n = 5$ , is more normal but still quite skewed for  $n = 25$ , and is getting close to normal by the  $n = 75$ .

f) Does the sampling distribution of the sample median appear to be centered around the value of the median of the (exponential) distribution? Explain briefly.

Not for smaller sample sizes, but the sample median is almost centered around the actual median with a sample size of  $n = 75$ .

g) Does the sampling distribution of the sample median have less and less variability as the sample size increases? Explain briefly.

Yes. The SD of the sample medians decreases substantially as the sample size increases.

2. Let the random variable  $H$  denote the study time (in hours per week) of a randomly selected Cal Poly student, and suppose that  $H$  has a uniform distribution on the interval  $(25, 35)$ .

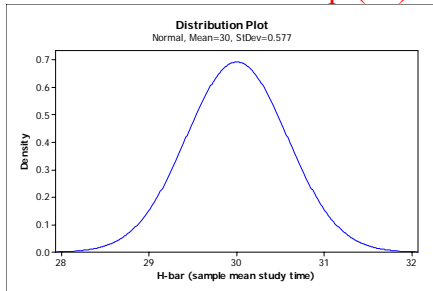
a) Determine the probability that a randomly selected student studies for between 29 and 31 hours per week.

$$\Pr(29 < H < 31) = 2/10 = 1/5 = .2.$$

Now consider a random sample of 25 students, and consider the sample mean study time.

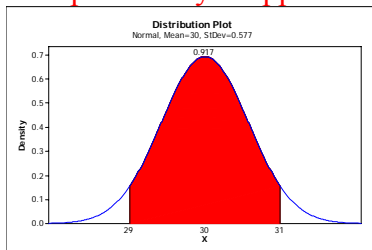
b) Specify, by name and parameter values, and sketch the (approximate) sampling distribution of this sample mean.

First note that  $E(H) = (25+35)/2 = 30$ ,  $\text{Var}(H) = (35-25)^2/12 \approx 8.333$ , and  $\text{SD}(H) = \sqrt{8.333} \approx 2.887$ . The CLT establishes that the sampling distribution of  $\bar{H}$  is approximately normal with mean 30 and  $\text{SD } 2.887/\sqrt{25} \approx 0.577$ . This is sketched below:



c) Determine the probability that the sample mean study time is between 29 and 31 hours per week.

This probability is approximately .917, as shown in the graph below:



d) What happens to this probability (that the sample mean study time is between 29 and 31 hours per week) as the sample size increases? Explain your answer.

This probability increases as the sample size increases. Because the interval from 29 to 31 encompasses the mean of 30, a larger sample makes it more likely to have such a sample mean.

3. Let the random variable  $X_i$  denote the number of typographical errors on page  $i$  of a 500-page textbook. Suppose that the  $X_i$ 's follow independent Poisson distributions with parameter  $\mu = .175$ . Determine the (approximate) probability that the book contains at least 100 typographical errors. Explain/justify your calculation.

First note that  $E(X_i) = .175$  and  $\text{Var}(X_i) = .175$ , so  $\text{SD}(X_i) = \sqrt{.175} \approx .418$ . So, the CLT establishes that with  $n = 500$ , the sample mean  $\bar{X}$  has an approximately normal distribution with mean .175 and  $\text{SD } .418/\sqrt{500} \approx .019$ . Now, finding at least 100 total errors on 500 pages is

equivalent to finding at least  $100/500 = 0.200$  errors per page on average, so we want  $\Pr(\bar{X} \geq 0.200)$ , which is .0941, as shown in the graph:

