Statistical inference Project

Mario Segal

August 9, 2014

This is my Class Project for the Coursera Class of Statistical Inference, August 2014

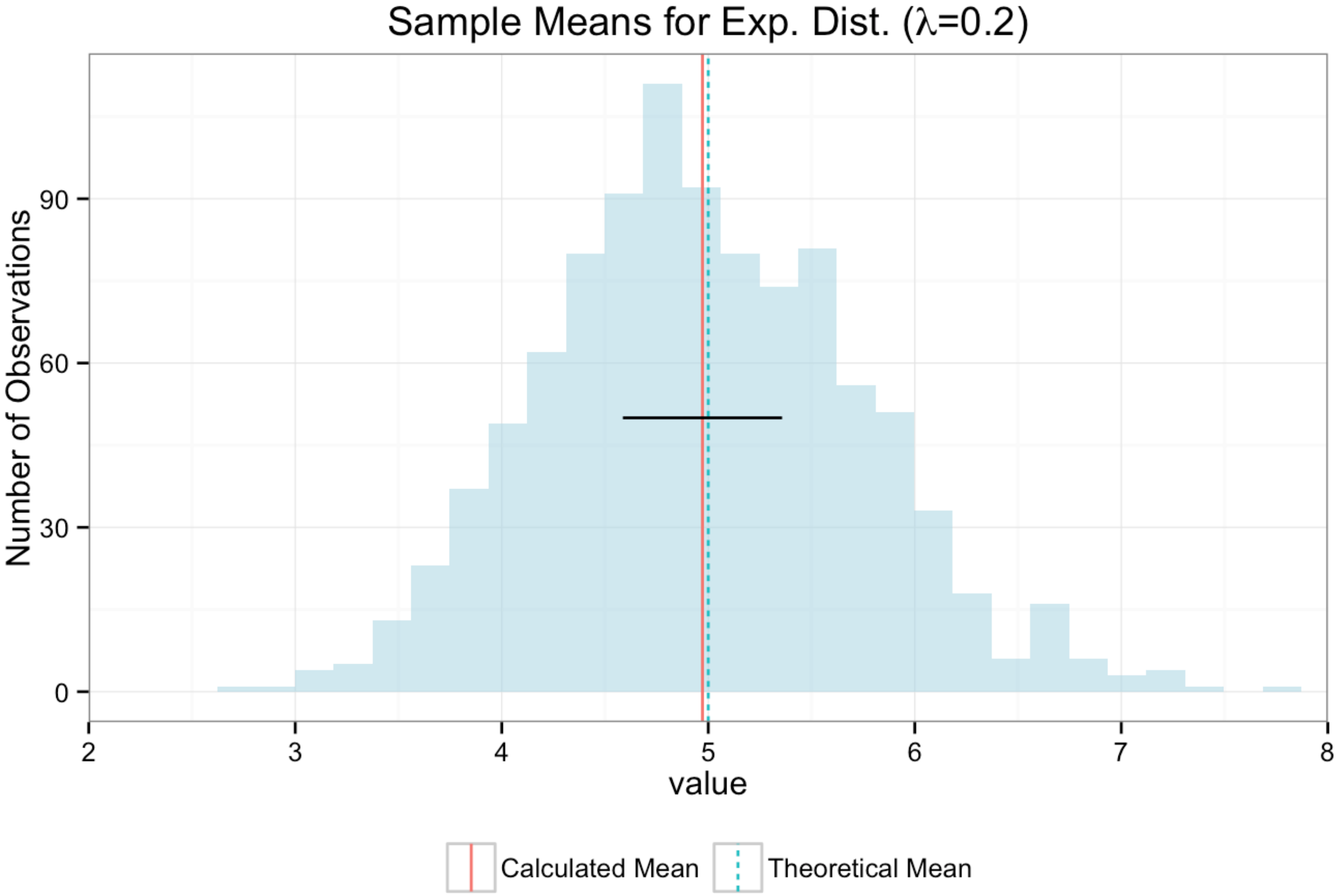
Question A: Explore Exponential Distribution with Lambda = 0.2

1. Repeatedly sample 40 exponential random numbers and plot their averages

```
require(plyr)
require(ggplot2)
require(scales)
set.seed(12345)
n = 40
lambda = 0.2
samples = 1000
data <- data.frame(Measure = "Sample Means", value = sapply(1:samples, function(x) mean(rexp(n,

    lambda))), stringsAsFactors = F)
# data <- ddply(data,.(),mutate,calc=mean(value),theory=1/lambda)
means <- data.frame(Type = c("Calculated Mean", "Theoretical Mean"), mean = c(mean(data$value),

    1/lambda))
sd_calc <- sd(data$value)
title <- expression(paste("Sample Means for Exp. Dist. (", lambda, "=", "0.2",
    ")", sep = ""))
ggplot(data, aes(x = value)) + geom_histogram(alpha = 0.6, fill = "lightblue") +
    theme_bw() + theme(legend.position = "bottom") + ggtitle(title) + coord_cartesian(xlim = c(
2,
    8)) + geom_vline(data = means, aes(xintercept = mean, color = Type, linetype = Type),
    show_guide = TRUE) + guides(color = guide_legend(title = NULL), linetype = guide_legend(tit
le = NULL)) +
    scale_y_continuous("Number of Observations", labels = comma) + geom_line(data = NULL,
    aes(x = c(mean(data$value) - sd_calc/2, mean(data$value) + sd_calc/2), y = c(50,
        50)), color = "black")
```

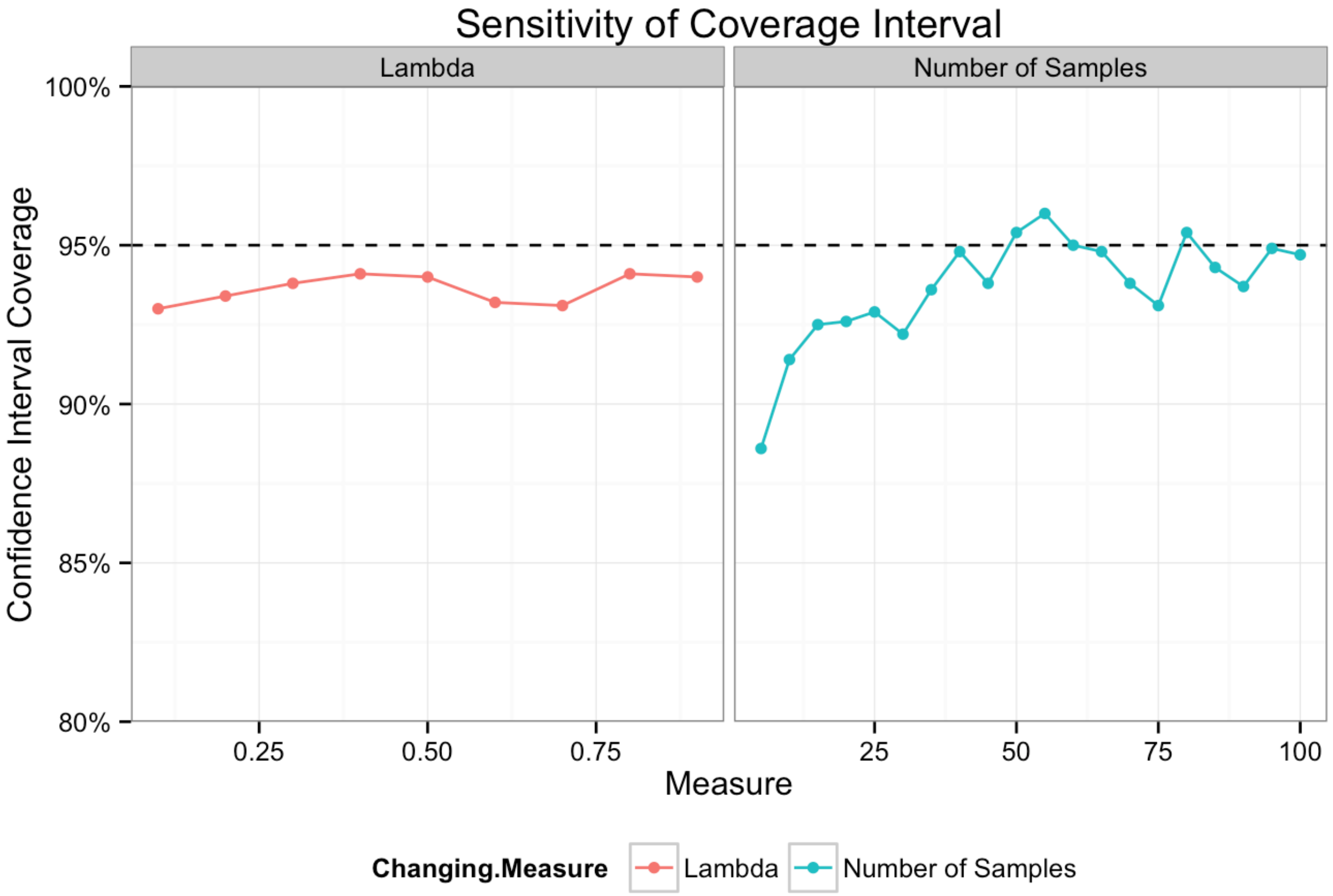


The figure above shows the distribution of means from 1,000 repeated samples of 40 exponential random numbers with λ =0.2. The red line is the calculated average of the 1,000 sample means, while the blue line denotes the theoretical mean of an exponential distribution which is $\frac{1}{\lambda}$ = 5. The Caluclatd Mean is 4.972 with a 95% confidence interval of 4.9241 to 5.0198. The black horizontal line shows the calculated standard deviation of the sample means centered around the calculated sample mean which is equal to 0.77. According to the central limit theorem the standard deviation of the sample means (the standard error) is equal to the standard deviation of the actual distribution divided by $\sqrt{n}$, where n is the sample size (40 in this case), given that with some basic algebra we calculate the estimated population standard deviation to be 4.88 which is close to the theoretical value of $\frac{1}{\lambda} = 5$ As expected the distribution of sample means appears to be normal and centered around the distribution mean

```
n = 40
coverage1 <- function(lambda) {
    sapply(lambda, function(lambda) {
        lhats <- mean(rexp(n, lambda))
        ll <- lhats - qnorm(0.975) * (lhats/sqrt(n))
        ul <- lhats + qnorm(0.975) * (lhats/sqrt(n))
        (ll < (1/lambda) & (1/lambda) < ul)
    })
}

coverage2 <- function(n1) {
    sapply(lambda, function(lambda) {
        lhats <- mean(rexp(n1, lambda))
        ll <- lhats - qnorm(0.975) * (lhats/sqrt(n1))
        ul <- lhats + qnorm(0.975) * (lhats/sqrt(n1))
        (ll < (1/lambda) & (1/lambda) < ul)
    })
}

set.seed(5443)
dat <- data.frame(`Changing Measure` = "Lambda", Measure = seq(0.1, 0.9, 0.1),
    coverage = sapply(seq(0.1, 0.9, 0.1), function(y) mean(sapply(1:samples,
        function(z) coverage1(y)))))
lmbda = 0.2
dat1 <- data.frame(`Changing Measure` = "Number of Samples", Measure = seq(5,
    100, 5), coverage = sapply(seq(5, 100, 5), function(y) mean(sapply(1:samples,
    function(z) coverage2(y)))))
data <- rbind(dat, dat1)
ggplot(data, aes(x = Measure, y = coverage, color = Changing.Measure)) + geom_hline(yintercept
= 0.95,
    col = "black", linetype = "dashed") + geom_line() + geom_point() + coord_cartesian(ylim = c
(0.8,
    1)) + scale_y_continuous("Confidence Interval Coverage", labels = percent) +
    facet_wrap(~Changing.Measure, scales = "free_x") + theme_bw() + theme(legend.position = "bo
ttom") +
    ggtitle("Sensitivity of Coverage Interval")
```



The figure above shows the sensitivity of the Confidence Interval from the sampled exponential distribution. The left panel simulates different values or lambda for 1,000 simulations of 40 exponential samples that come close but do not quite reach the 95% threshold. The right panel simulates 1,000 simulations of different exponential samples for $lambda = 0.2$, showing that predictability improves with sample size as expected however the 95% threshold is not always reached as sample size increases