Holsenbeck_S_Midterm

Stephen Synchronicity

2017-11-02

Midterm

1

Using R, write a program to calculate all the prime numbers less than 100. A prime number is a positive integer greater than 1 that is divisible (without remainder) only by 1 and itself. Create the program by testing each number from 2 to 100 against all integers less than it using %%. Your function should return a vector of all the primes < 100. (15 pt)

A) **1**

```
prime <- function(x) {
    OP <- c(2L)
    d <- seq(2, x)
    for (i in seq(2, x)) {
        div <- seq(2, d[i - 1])
        vf <- d[i]%%div
        if (0 %in% vf == F) {
            OP[i] <- d[i]
        }
        OP <- OP[!is.na(OP)]
    }
    return(OP)
}
(p100 <- prime(100L))
```

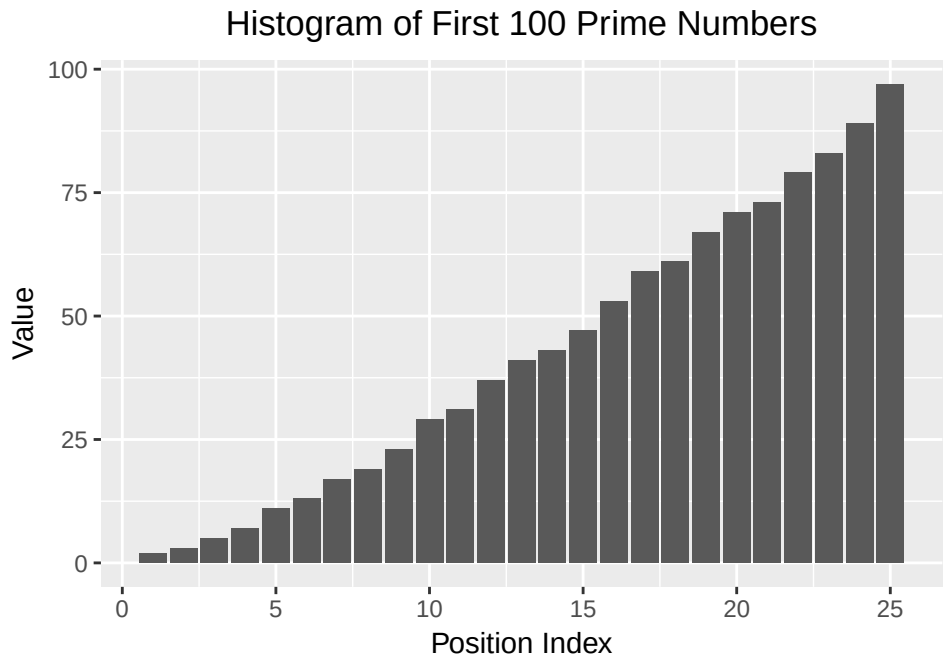
```
##  [1]  2  3  5  7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71 73 79 83
## [24] 89 97
```

2

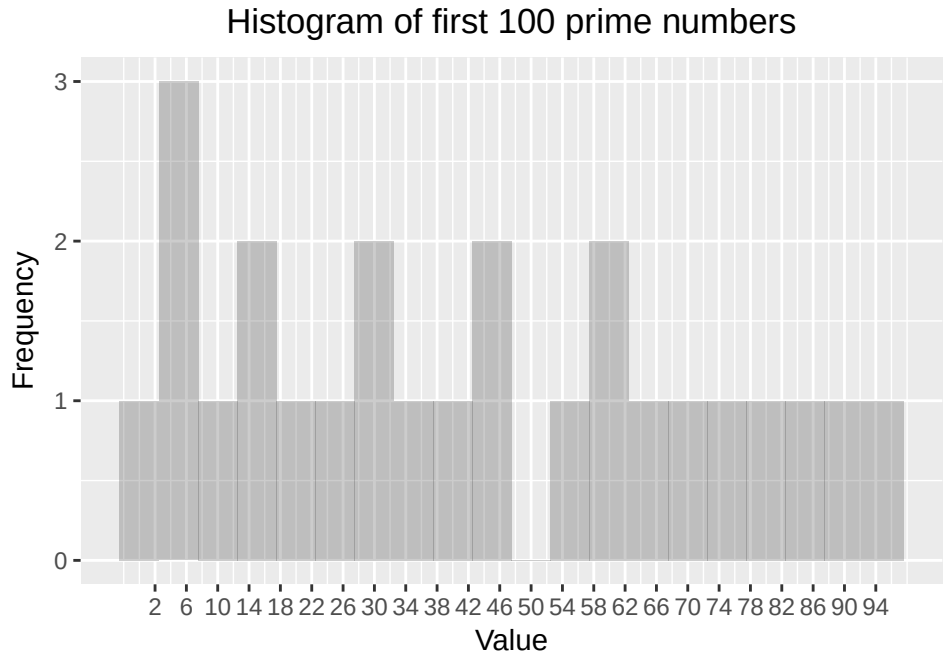
Using R, create a histogram of the result from 1 using ggplot. Be sure to nicely label your axes and title the graph. (5pt)

A) **2**

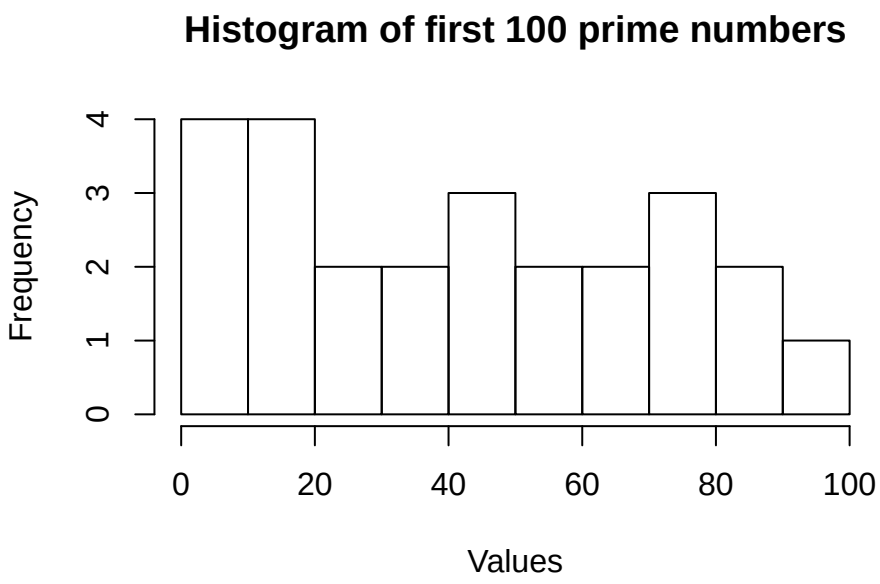
```
library(ggplot2)
index <- seq(1, 25)
df.p100 <- as.data.frame(cbind(index, p100))
ggplot(data = df.p100, mapping = aes(x = index, y = p100)) + geom_histogram(stat = "identity") +
    xlab("Position Index") + ylab("Value") + ggtitle("Histogram of First 100 Prime Numbers") +
    theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```




```
ggplot(data = df.p100, mapping = aes(x = p100)) + geom_histogram(binwidth = 5, alpha = 0.3) +
    ggtitle("Histogram of first 100 prime numbers") + xlab("Value") + ylab("Frequency") +
    scale_x_continuous(breaks = round(seq(min(p100), max(p100), by = 4), 1)) + theme(plot.title = eleme
    plot.subtitle = element_text(hjust = 0.5))
```




```
hist(p100, breaks = 10, xlab = "Values", main = "Histogram of first 100 prime numbers")
```



3

You flip a coin five times. a. What's the chance of getting three or more heads in a row? (5 pt) b. What's the chance of getting three or more heads in a row conditional on knowing the first flip was a heads? (5 pt)

a) $P(A = H|B = H|C = H) = 0.5^3 = 0.125$

b) $P(B = H, C = H)$ given $P(A = H) = 1 = 1 * .5^2 = 0.25$

3

```
0.5^3
```

```
## [1] 0.125
```

```
0.5^2
```

```
## [1] 0.25
```

4

NASA has declared that the Earth is likely to be hit by an asteroid this year based on an astronomical observation it has made. These things are hard to judge for certain, but it is known that the test NASA used is pretty good – it has a sensitivity of 99% and a false positive rate of only 1%. It is further known that the general probability of an asteroid hitting earth in any given year is 1 in 100,000. What is the probability we will actually be hit by an asteroid this year given NASA's test? (10 pt)

A)

$$P(A|+) = \frac{P(+|A)P(A)}{P(+|A)P(A) + P(+|\neg A)P(\neg A)}$$

4

```
s <- 0.99
fp <- 0.01
```

```
pA <- 1/1e+05
false.pos <- function(pA, s, fp) {
    s * pA/(s * pA + fp * (1 - pA))
}
false.pos(pA, s, fp)
```

```
## [1] 0.0009890307
```

A) Funny NASA would have data on the false positive rate and sensitivity for such a test. We must be talking about small asteroids if they have that data. The chance that an asteroid hits earth given a positive test is 0.000989

5

The average number of snow days in Boston in a winter month is 1. Assuming these events follow a poisson distribution, calculate (using R) the probability of getting 5 or more snow days in a month. (5 pt)

A) **5**

```
ppois(5, 1, F)
```

```
## [1] 0.0005941848
```

6

You want to know how many hours of sleep the average college student gets. You start out with a preliminary survey of 10 people, and get the following data (in hours): 7,6,5,8,6,6,4,5,8,7. You hypothesize that despite what the doctors say, the average college student does not get 7 hours of sleep a night. What does your survey say? State your null hypothesis, research hypothesis (two tailed), and calculate your threshold value, test statistic, and p value. Do you reject the null or not? (10 pt)

A)

$$H_0 : \mu = 7$$
$$H_a : \mu \neq 7$$
$$t_{stat} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$
$$t_{stat} = \frac{6.2 - 7}{\frac{1.316561}{\sqrt{10}}}$$
$$t_{stat} = -1.92153785$$
$$p_{value} = 2 * \int_{-\infty}^{x} f(-|t_{stat}|)dt$$
$$p_{value} = 0.08684229$$

6

```
sh <- c(7, 6, 5, 8, 6, 6, 4, 5, 8, 7)

tCrit(0.95, sh, 2, 7)
```

```
## # A tibble: 6 x 2
##       Variable      Values
##          <chr>       <dbl>
```

```
## 1    Low tCrit -2.26215716
## 2   High tCrit  2.26215716
## 3 Conf_Int.Low  5.25818887
## 4  Conf_Int.Hi  7.14181113
## 5  T Statistic -1.92153785
## 6       pValue  0.08684229
```

At the 95% confidence level, we fail to reject the null hypothesis, and conclude that the average amount of sleep per night among college students is not significantly different from ~7 hrs.

7

Despite the disappointing results in 6, you are confident in your hypothesis. Assuming your sample standard deviation and mean do not change and you want to survey as few people as possible, how many additional people would you have to survey to reject the null at the 0.05 level? (5 pt)

A) Algebraically, ~14, or 4 additional people would need to be surveyed to reject the null. Experimentally, the addtl function adds survey participants with the rNorm function using the mean and sd of the supplied data. It outputs the number of participants it added when the p_{value} becomes significant. When this trial is replicated 100 times, the number of people needed to be surveyed is ~18 or 8 additional.

7 - Algebraic Approach

$$t_{stat} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < -2.26215716 \mid t_{stat} > 2.26215716$$

$$-2.26215716 = \frac{6.2 - 7}{\frac{1.316561}{\sqrt{n}}}$$

$$-2.26215716 \frac{1.316561}{\sqrt{n}} = -0.8$$

$$1.316561 = 0.3536447\sqrt{n}$$

$$n = 13.8595$$

7 - Experimental Approach

```
(1.31656/-0.3536447)^2
addtl <- function(cl, iVector, mu = 0) {
    input <- c(mean(iVector), mu, sd(iVector), length(iVector))
    par <- tCrit(cl, input, 2, mu)
    while (par$Values[6] > (1 - cl)) {
        iVector <- append(iVector, abs(rnorm(1, mean(iVector), sd(iVector))), after = (length(iVector))
        input <- c(mean(iVector), mu, sd(iVector), length(iVector))
        par <- tCrit(cl, input, 2, mu)
        if (par$Values[6] < (1 - cl)) {
            return(length(iVector))

            break
        }
    }
    return(length(iVector))
}
addtl(0.95, sh, 7)
round(mean(replicate(100, addtl(0.95, sh, 7), simplify = T)))
```

8

You survey the same 10 people during finals period, and get the following hours: 5,4,5,7,5,4,5,4,6,5. Do college students get significantly less sleep than usual during finals? (10 pt)

A) According to the t-test at the 95% confidence interval, yes, we can reject the null in favor of the hypothesis that college students get significantly less sleep during finals. The 95% confidence interval indicates that the difference in sleep duration could be anywhere from .116 to 2.284 hours in the negative direction. **8**

```
fsh <- c(5, 4, 5, 7, 5, 4, 5, 4, 6, 5)
t.test(sh, fsh)
```

```
##
##  Welch Two Sample t-test
##
## data:  sh and fsh
## t = 2.3434, df = 16.309, p-value = 0.03209
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1161141 2.2838859
## sample estimates:
## mean of x mean of y
##       6.2       5.0
```

9

You are a very bad gardener, and hypothesize that feeding houseplants vodka might help them relax and grow better. You perform an experiment to test your hypothesis, giving 15 houseplants water spiked with vodka, and 15 houseplants water alone. These are your results: This looks pretty bad for the treatment, but being as good at statistics as you are bad at gardening, you test it using the chi-square test. What are your results? (15 pt)

9 Data

```
dt9 <- tibble::tribble(~condition, ~live, ~die, "treatment", 4L, 11L, "control",
    8L, 7L)
```

A) H_{n1} The proportion of plants that lived is independent of treatment. H_{a1} The proportion of plants that lived is dependent on treatment. H_{n2} The proportion of plants that died is independent of treatment. H_{a2} The proportion of plants that died is dependent on treatment. H_{n3} The numbers of plants living or dying is independent of treatment. H_{a3} The numbers of plants living or dying is dependent on treatment. H_{n4} The survival rate is independent of treatment. H_{a4} The survival rate is dependent on treatment. **9**

```
# Method to add Sums dt9 <- dt9 %>% rowwise() %>% mutate(Sum = sum(live,die))
# (dt9 <- rbind(dt9,c('Sums',colSums(dt9[,-1]))))
chisq.test(dt9[, 2])  #Hyp 1 Living
```

```
##
##  Chi-squared test for given probabilities
##
## data:  dt9[, 2]
## X-squared = 1.3333, df = 1, p-value = 0.2482
```

```
chisq.test(dt9[, 3])  #Hyp 2 Dying
```

```
##
##  Chi-squared test for given probabilities
##
## data:  dt9[, 3]
## X-squared = 0.88889, df = 1, p-value = 0.3458
```

```
chisq.test(dt9[, c(2:3)])  #Hyp 3 #Taken together (dependence across groups)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dt9[, c(2:3)]
## X-squared = 1.25, df = 1, p-value = 0.2636
```

```
chisq.test(c(4/11, 8/7))  #Hyp 4 #Survival Rate
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(4/11, 8/7)
## X-squared = 0.40305, df = 1, p-value = 0.5255
```

The Chi-squared test, shows that the data regardless of it's segmentation, indicates that the treatment and control groups are independent, whether a plant lives, or a plant dies is not dependent on what they are watered with. The small sample size may be part of the reason that no statistically significant correlation can be found.

10

Perhaps you got things backwards, and plants need more stimulation to thrive. So you adjust your experiment into three treatment groups: water, vodka, and coffee. These are your results: The overall mean is 50 days (as we said, you're a bad gardener). Use an F test to determine if there is any significant difference among these three groups. (15 pt)

10 data

```
dt10 <- tibble::tribble(~condition, ~mean.days.alive, ~sd, ~n, "water", 50L, 10L,
    20L, "vodka", 45L, 7L, 10L, "coffee", 55L, 4L, 10L)
dt10 <- as.data.frame(dt10)
```

A) $H_0 : \mu_w = \mu_v = \mu_c$: The average lifespan is the same across conditions.
$H_a : \neg(\mu_D = \mu_I = \mu_R)$ The average lifespan is different across conditions.

$$f_{stat} = \frac{\text{average variance between groups}}{\text{average variance within groups}}$$

$$\text{between groups} = \frac{n_1(\bar{y}_1 - \bar{y})^2 + ... + n_G(\bar{y}_G - \bar{y})^2}{df = G - 1}$$

$$\text{within groups} = \frac{(n_1 - 1)s_1^2 + ... + (n_G - 1)s_G^2}{df = N - G}$$

$$\text{where } N = \text{sum(n) in all}, G = \text{no. of Groups}$$

$$\text{compare } f_{stat} \text{ to qf}(cl, df_1, df_2)$$

$$\text{or compare } p_{value} = \text{1-pf}(f_{stat}, df_1, df_2) \text{ to } \alpha$$

10

```r
f.test <- function(df) {
    bgt <- vector("numeric")
    wgt <- vector("numeric")
    for (i in seq(1:nrow(df))) {
        ybar <- df[i, 2]
        s <- df[i, 3]
        n <- df[i, 4]
        (nums <- sapply(df, is.numeric))
        df1 <- ncol(df[, nums]) - 1
        df2 <- sum(df[, 4]) - ncol(df[, nums])
        bgt[i] <- n * (ybar - mean(df[, 2]))^2
        wgt[i] <- (n - 1) * s^2
    }
    wg <- sum(bgt)/df1
    bg <- sum(wgt)/df2
    f <- bg/wg
    fcrit <- qf(0.95, df1, df2)
    pv <- 1 - pf(f, df1, df2)
    OP <- tibble::tribble(~Stat, ~Value, "fStat", f, "fCrit", fcrit, "pValue", pv)
    return(OP)
}
f.test(dt10)
```

```
## # A tibble: 3 x 2
##     Stat     Value
##    <chr>     <dbl>
## 1  fStat 0.2686486
## 2  fCrit 3.2519238
## 3 pValue 0.7658900
```

Based on these results, we fail to reject the null hypothesis at the 95% confidence level in support of the claim that the average lifespan is centered around a mean of 50 days across conditions, and that the treatments being studied do not have a significant influence on the average days of the life span. Whatever plant is being tested appears to be very resilient.