Лабораторная работа № 2

«Интервальное оценивание статистических данных»

Выполнил: Николай Окуньков, 18ПИ-2.

Цель работы: Изучение интервального оценивания статистических данных, в том числе с использованием встроенных в базовую версию пакета R функций.

Теоретическая часть:

<u>Доверительный интервал</u> – один из типов интервальных оценок, используемых в статистике, которые рассчитываются для заданного уровня значимости. Они позволяют сделать утверждение, что истинное значение неизвестного статистического параметра генеральной совокупности находится в полученном диапазоне значений с вероятностью, которая задана выбранным уровнем статистической значимости.

<u>Доверительный интервал для оценки математического ожидания при известной дисперсии:</u>

$$P\left\{\bar{x}_n - t_{1-\frac{\alpha}{2},n-1}\frac{\hat{s}(\vec{x}_n)}{\sqrt{n}} \le m \le \bar{x}_n + t_{1-\frac{\alpha}{2},n-1}\frac{\hat{s}(\vec{x}_n)}{\sqrt{n}}\right\} = 1 - \alpha$$

где $t_{1-\alpha/2}$ – квантиль уровня ***1 – α/2*** стандартного нормального распределения ***N (0,1)***.

<u>Доверительный интервал для оценки математического ожидания при неизвестной дисперсии:</u>

$$P\left\{\frac{(n-1)\,\hat{s}^2(\vec{x}_n)}{\mathrm{X}^2_{1-\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)\,\hat{s}^2(\vec{x}_n)}{\mathrm{X}^2_{\frac{\alpha}{2},n-1}}\right\} = 1 - \alpha$$

где $t_{1-\alpha/2}$ – квантиль уровня ***1 – α/2*** распределения Стьюдента с ***n – 1*** степенями свободы.

<u>Доверительный интервал для оценки дисперсии при неизвестной математическом ожидании:</u>

$$P\left\{\bar{x}_n - t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le m \le \bar{x}_n + t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

где $X^2_{1-\frac{\alpha}{2},n-1}$ и $X^2_{\frac{\alpha}{2},n-1}$ соответственно квантили уровня ***1 – α/2*** и ***α/2*** распределения X^2 с ***n – 1*** степенями свободы.

Оценка необходимого объема выборки по заданной точности и доверительной вероятности рассчитывается по формуле:

$$n = \frac{t^2\sigma^2}{\Delta^2},$$

где ***t*** — коэффициент доверия при заданном уровне вероятности; ***σ^2*** — дисперсия; **Δ** — предельная ошибка выборки.

Ход работы:

<u>Задание 1:</u>

Построить функции для вычисления доверительного интервала по заданной выборке и величине доверительного интервала для 1) математического ожидания при известной и неизвестной дисперсии; 2) дисперсии при неизвестном математическом ожидании.

Для решения задачи были использованы функции **fun1, qnorm, sqrt, fun2, qt, fun3, qchisq**:

fun1 – функция находит доверительный интервал при известных математическом ожидании и дисперсии;

(Параметры: **array** – выборка; **var** – дисперсия; **g** – доверительная вероятность.)

qnorm – функция нужна для вычисления квантилей нормального распределения;

sqrt – функция находит квадратный корень аргумента;

fun2 – функция вычисляет доверительный интервал для математического ожидания при неизвестной генеральной дисперсии;

qt – функция нужна для вычисления квантилей распределения Стьюдента;

fun3 – функция вычисляет доверительный интервал для дисперсии при неизвестном математическом среднем;

qchisq – функция нужна для вычисления квантилей распределения хи-квадрат.

```
fun1 <- function(array, sd, g){
# доверительный интервал для математического ожидания при известной дисперсии
  array <- na.omit(array)
  N <- length(array)
  delta <- qnorm(1-(1-g)/2) * sd /sqrt(N)
  interval <- c(mean(sample1) - delta, mean(sample1) + delta)
  return(interval)
}

fun2 <- function(array, g){
# доверительный интервал для математического ожидания при неизвестной дисперсии
  array <- na.omit(array)
  N <- length(array)
  mean.c <- sqrt((N/(N-1))*var(array))
  delta <- qt(1-(1-g)/2, N-1) * mean.c / sqrt(N)
  interval <- c(mean(sample1) - delta, mean(sample1) + delta)
  return(interval)
}

fun3 <- function(array, g){
# вычисляет доверительный интервал для дисперсии при неизвестном математическом ожидании
  array <- na.omit(array)
  N <- length(array)
  mean.c <- sqrt((N/(N-1))*var(array))
  a1 <- (1 - g)/2
  a2 <- (1 + g)/2
  interval <- c(sqrt(N-1)* mean.c / sqrt(qchisq(a2, N-1)), sqrt(N-1)* mean.c / sqrt(qchisq(a1, N-1)))
}
```

<u>Задание 2:</u>

Получить две выборки объема n 1 = 100 и n 2 = 200 из нормального распределения N (m, σ 2). Для обеих выборок найти доверительные интервалы с заданной доверительной вероятностью γ = 0. 90 с помощью полученных функций.

```
> N1   <- 100  # кол-во элементов в первой выборке
> N2   <- 200  # кол-во элементов во второй выборке
> mean <- 3    # мат. ожидание
> sd   <- 1.2  # среднеквадратичное отклонение
> g    <- 0.9  # доверительная вероятность
> # создаем две выбоки размером в 100 и 200 элементов
> set.seed(13)
> sample1 <- rnorm(N1, mean = mean, sd = sd); sample1
  [1] 3.6651923 2.6636737 5.1301960 3.2247841 4.3710314 3.4986314 4.4754079 3.2840156 2.5615407 4.3261731 1.6876872 3.5542451
 [13] 1.3668186 0.7727674 2.4721735 2.7672637 4.6757178 3.1207959 2.8626734 3.8426703 3.3150512 5.2033960 3.4288829 1.7455078
 [25] 3.7442210 3.1792254 1.2488198 0.5675474 1.7316507 2.1262275 2.9901472 4.0173569 2.5398102 2.3681862 2.6721288 2.2731101
 [37] 2.6005592 2.7101549 1.9646695 1.9836351 3.1204084 4.9080402 3.6797939 4.9373754 2.4376198 2.1286783 1.7719932 0.6746214
 [49] 3.3325767 4.6900244 3.3277550 3.9066301 2.5811779 2.3445711 3.2812344 2.6426061 1.9914286 3.9918124 4.7804295 3.8396108
 [61] 1.4861110 3.3579261 2.8226315 1.9332932 4.2156790 1.8953699 2.3113266 4.3804386 4.3725895 2.7126687 1.6958374 2.9262636
 [73] 2.3799632 0.7107916 3.1285878 1.5871498 5.0945123 2.5215618 3.5309273 3.5403354 2.9087254 3.3570159 1.5667743 0.6037494
 [85] 4.6662157 2.9010197 3.4710174 1.7006763 4.9225445 4.2048828 3.4558748 2.3213936 1.5434663 1.3628381 1.3006405 2.6933064
 [97] 1.5294889 3.2566011 3.0806683 4.0279621
> sample2 <- rnorm(N2, mean = mean, sd = sd); sample2
  [1]  2.55699278  3.73379692  1.04549122  3.17082921  2.68236839  2.39811502  4.46808645  3.51846262  2.99182746  5.00762881
 [11]  1.96017995  4.96337507  1.77467693  2.02588263  2.93908457  2.86489031  4.02318035  3.89681604  3.46736192  2.77677651
 [21]  4.59057704  3.54502513  2.89998174  2.91430996  1.38540445  2.75894795  2.71639384  2.93058660  1.76167219  1.96014101
 [31]  4.01713315  3.14810627  3.48182628  4.10679084  5.14391216  2.78484496  2.52901993  3.86647530  4.15022911  1.44543950
 [41]  3.40723406  1.62268478  1.69089854  0.60764543  3.51630979  5.04861504  1.10129805  2.57159577  0.96161574  1.61763559
 [51] -0.41371661  5.05749311  1.52541981  2.62311245  2.91610521  1.99487828  3.29623267 -0.09015387  3.00516232  5.81515510
 [61]  3.07672417  4.33107454  2.49111078  3.78530579  3.71241168  1.81798835  3.34050824  3.03536783  3.95963203  1.43607288
 [71]  6.32445259  2.58519809  4.20779439  6.37752496  3.99323579  6.45759768  3.42023736  1.91285021  3.97873227  2.40838888
 [81]  2.36286075  3.76186621  0.51173400  2.76716947  3.23094518  1.45761830  3.25161742  2.64038805  1.74735141  3.98287785
 [91]  4.20186595  3.52177586  3.55808004  1.68161773  3.73311107  3.01954728  3.88358088  4.49545332  2.12084619  2.91442707
[101]  4.23790235  3.98285820  2.46103500  4.32710586  4.51123393  2.27256585  1.30994642  5.03492612  1.41814213  1.90778685
[111]  0.82298301  3.22786815  1.95895321  3.55505195  0.98308788  4.53746733  1.69062722  3.30276421  2.83091520  1.06378651
[121]  2.26624673  2.82380422  2.56391772  3.85008701  2.47129057  3.66591056  2.39847205  3.18770565  4.87577136  3.61411595
[131]  2.99892497  1.17637193  2.75168982  3.85207423  0.80140566  2.72805802  2.39085329  1.52188196  1.93389163  2.38769293
[141]  3.24990017  3.21143266  0.14198216  1.67442021  2.26870874  3.60092301  2.48037403  2.59219916  0.53364767  2.78410382
[151]  3.73654111  3.49513743  2.44932760  2.19218580  3.23166516  4.65951487  3.07796081  5.76269244  0.49958666  1.61142093
[161]  3.74856677  1.17256910 -0.13326719  6.03188858  3.45346864  3.03413595  2.15276574  3.76175958  4.40259367  2.63018080
[171]  2.15546297  1.47600783  3.11469238  2.46411729  0.94874204  1.26933214  4.52780211  3.91902580  1.88352527  4.73764789
[181]  2.38501573  5.80326892  3.78567357  2.83272548  2.39522115  3.06874075  1.27214438  4.50581478  3.50398546  3.12715450
[191]  3.10791663  0.14267528  6.48682628  3.30814234  3.18753291  1.45810325  2.46168291  1.09247422  1.49278104  2.87178767
> # находим доверительные интервалы для математического ожидания при известной дисперсии
> interval1 <- fun1(sample1, sd = sd, g = g); interval1
[1] 2.728427 3.123192
> interval2 <- fun1(sample2, sd = sd, g = g); interval2
[1] 2.786239 3.065380
> # находим доверительные интервалы для математического ожидания при неизвестной дисперсии
> interval1 <- fun2(sample1, g = g); interval1
[1] 2.735403 3.116216
> interval2 <- fun2(sample2, g = g); interval2
[1] 2.770380 3.081239
> # находим доверительные интервалы для математического ожидания при неизвестной дисперсии
> interval1 <- fun3(sample1, g = g); interval1
[1] 1.027872 1.299908
> interval2 <- fun3(sample2, g = g); interval2
[1] 1.22949 1.45042
```

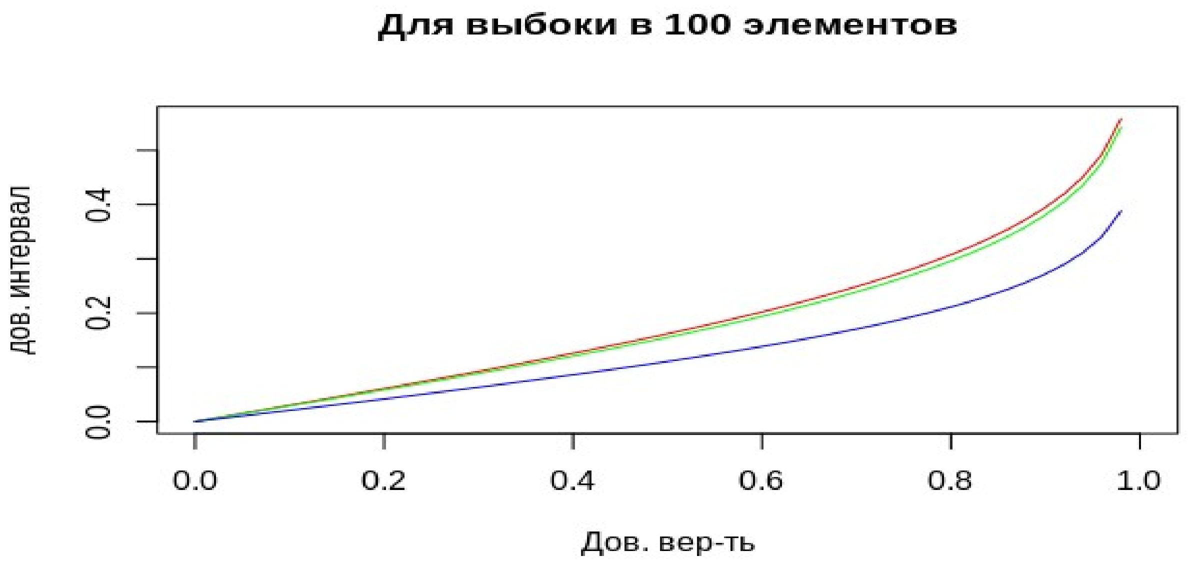
В выводе первая граница интервала левая, вторая – правая. Видно, что данные интервалы отличаются десятыми долями. Данное явление можно наблюдать из-за разницы в способе вычисления.

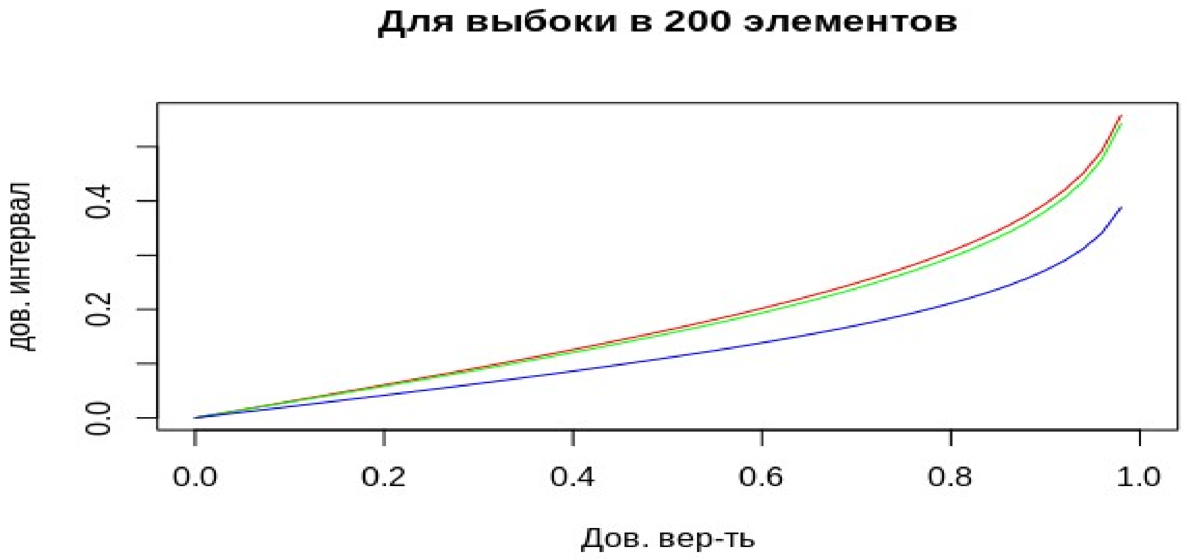
Задание 3:

Для всех трёх случаев в одной системе координат построить графики зависимости длины доверительного интервала от величины доверительной вероятности. При этом γ придать минимум 50 разных значений. Проанализировать взаимное расположение полученных графиков и объяснить его.

```r
findLen <- function(interval){
# функция для нахождения длины интервала
  return(abs(max(interval) - min(interval)))
}


# строим графики зависимости длины доверительного интервала от величины доверительной вероятности.
par(mfcol=c(2, 1))

G <- seq(from=0, to=1, by=0.02); G
len1 <- c(1:length(G))
len2 <- c(1:length(G))
len3 <- c(1:length(G))

for(i in 1: length(G)){
  len1[i] <- findLen(fun1(sample1, sd = sd, g = G[i]))
  len2[i] <- findLen(fun2(sample1, g = G[i]))
  len3[i] <- findLen(fun3(sample1, g = G[i]))
}

plot(G, len1,  col = "red", type = "l",
     xlab = "Дов. вер-ть",
     ylab = "дов. интервал",
     main = "Для выбоки в 100 элементов
     ")
lines(G, len2,  col = "green")
lines(G, len3,  col = "blue")

for(i in 1: length(G)){
  len1[i] <- findLen(fun1(sample2, mean = mean, g = G[i]))
  len2[i] <- findLen(fun2(sample2, g = G[i]))
  len3[i] <- findLen(fun3(sample2, g = G[i]))
}

plot(G, len1,  col = "red", type = "l",
     xlab = "Дов. вер-ть",
     ylab = "дов. интервал",
     main = "Для выбоки в 200 элементов
     ")
lines(G, len2,  col = "green")
lines(G, len3,  col = "blue")
```

Результат:


Для выбоки в 100 элементов
дов. интервал
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
Дов. вер-ть


Для выбоки в 200 элементов
дов. интервал
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
Дов. вер-ть

Анализ графиков:

Первый график показывает зависимость длины доверительного интервала от доверительной вероятности для выборки из 100 элементов. Второй график – для выборки из 200 элементов. Красной линией обозначены результаты функции расчёта доверительного интервала для мат. ожидания при известной дисперсии, зеленой линией – результаты функции расчёта доверительного интервала для мат. ожидания при неизвестной дисперсии, синей линией – результаты функции расчёта доверительного интервала для дисперсии при неизвестном мат. ожидании. Мы можем наблюдать, что при увеличении объёма выборки длины уменьшились, значит при увеличении объема выборки можно получить меньший по длине интервал при неизменной вероятности. Причина этому – корень из длины выборки, стоящий в знаменателе при расчётах.

Вывод

Выполнение данной работы помогло мне изучить способы нахождения доверительных интервалов для различных случаев, рассмотреть и применить возможности пакета R для их построения. Благодаря графикам я смог выявить зависимость ширины доверительного интервала от величины доверительной вероятности.