04.11.2022

**Домашнее задание №3**

**Отчет**

Количество циклов химиотерапии, требующихся пациенту в дебюте заболевания Z, является случайной величиной со следующим распределением:

Таблица 1 – Распределение количества циклов химиотерапии в дебюте заболевания Z

|  |  |  |
| --- | --- | --- |
| Количество циклов (X) | 1 | 2 |
| Вероятность (p) | 0,5 | 0,5 |

При рецидиве распределение является следующим:

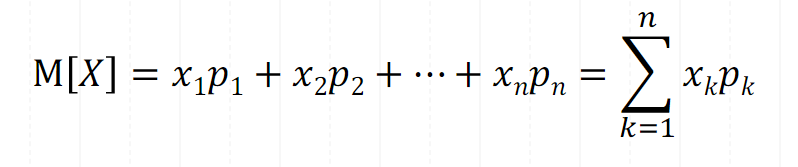
Таблица 2 – Распределение количества циклов химиотерапии при рецидиве заболевания Z

|  |  |  |
| --- | --- | --- |
| Количество циклов (X) | 1 | 2 |
| Вероятность (p) | 0,25 | 0,75 |

1. Рассчитаем математическое ожидание и дисперсию числа циклов терапии при дебюте (первичном выявлении) и при рецидиве.

Формула расчета математического ожидания:

Формула 1 – Расчёт математического ожидания M[X]



Формула расчёта дисперсии:

Формула 2 – Расчёт дисперсии D[X]

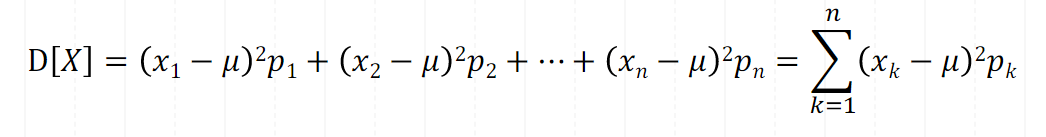


Таблица 3 – Результат расчёта M[X] и D[X] при дебюте и рецидиве

|  |  |  |  |
| --- | --- | --- | --- |
| **Метрика** | **Обозначение** | **Дебют (первичное выявление)** | **Рецидив** |
| **Математическое ожидание** | M(X) | 1,50 | 1,75 |
| **Дисперсия** | D(X) | 0,25 | 0,19 |

*Прим. Расчёт произведен в excel. См. Приложение 1, лист №1*

1. Предположим, что мы изучаем только рецидивировавших пациентов.

Построим таблицу распределения общего числа циклов терапии у рецидивировавших пациентов («дебютных» + «рецидивных»):

Таблица 4 – Распределение количества циклов химиотерапии у рецидивировавших пациентов

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Дебют + рецидив** | | |
| **Количество циклов (X)** | 2 | 3 | 4 |
| **Вероятность (p)** | 0,125 | 0,500 | 0,375 |

*Прим. Расчёт произведен в excel. См. Приложение 1, лист №2*

Рассчитаем математическое ожидание и дисперсию для полученного распределения (таблица 4):

Таблица 5 – Результат расчёта M[X] и D[X] при дебюте + рецидиве

|  |  |  |
| --- | --- | --- |
| **Метрика** | **Обозначение** | **Дебют + рецидив** |
| **Математическое ожидание** | M(X) | 3,25 |
| **Дисперсия** | D(X) | 0,44 |

*Прим. Расчёт произведен в excel. См. Приложение 1, лист №2*

1. Посмотрим график теоретической зависимости стандартной ошибки (SE) при оценке среднего числа циклов от объема выборки N. Величину N возьмём в диапазоне 10 – 160.

Для этого рассчитаем стандартное отклонение по формуле:

Формула 3 – Расчёт стандартного отклонения (X)

Таблица 6 – Результат расчёта (X) при дебюте + рецидиве

|  |  |  |
| --- | --- | --- |
| **Метрика** | **Обозначение** | **Дебют + рецидив** |
| **Стандартное отклонение** | (X) | 0,66 |

Далее, построим таблицу распределения теоретической зависимости стандартной ошибки (SE) при значениях N в диапазоне от 10 до 160. Рассчитаем SE по формуле:

*Формула 4 – Расчёт стандартной ошибки SE*

*Расчёт произведен в excel. См. Приложение 1, лист №3.*

Наконец построим график теоретической зависимости стандартной ошибки (SE) при значениях N в диапазоне от 10 до 160.

Рисунок 1 – График теоретической зависимости стандартной ошибки (SE) при значениях N в диапазоне от 10 до 160

1. Для значений N = 10, 40, 160 произведем виртуальный эксперимент.

Сформируем выборку указанного количества (10, 40, 160) с помощью функции *sample,* оценим по выборке стандартное отклонение (), оценим стандартную ошибку. Выполним эти операции с помощью кода в R:

library('readr')

library('dplyr')

cycles <- c(2,3,4) # Количество циклов химиотерапии для дебют+рецидив (из задания 2)

pr <- c(0.125, 0.500, 0.375) # Веротяности для циклов терапии дебют+рецидив (из задания 2)

df <- data.frame(cycles, pr) # Дата-фрейм

n1 <- 10 # Количество пациентов

n2 <- 40 # Количество пациентов

n3 <- 160 # Количество пациентов

group1 <- df[sample(1:nrow(df), size = n1, replace = TRUE), ]

group2 <- df[sample(1:nrow(df), size = n2, replace = TRUE), ]

group3 <- df[sample(1:nrow(df), size = n3, replace = TRUE), ]

true\_mean\_1 <- sum(cycles\*pr) # Истинное среднее (математическое ожидание)

variance\_1 <- sum((cycles-true\_mean\_1)^2\*pr) # Дисперсия

standard\_deviation\_1 <- sqrt(variance\_1)

print(standard\_deviation\_1)

se\_1 <- standard\_deviation1/sqrt(n1)

true\_mean\_2 <- sum(cycles\*pr) # Истинное среднее (математическое ожидание)

variance\_2 <- sum((cycles-true\_mean\_2)^2\*pr) # Дисперсия

standard\_deviation\_2 <- sqrt(variance\_2)

se\_2 <- standard\_deviation1/sqrt(n2)

true\_mean\_3 <- sum(cycles\*pr) # Истинное среднее (математическое ожидание)

variance\_3 <- sum((cycles-true\_mean\_3)^2\*pr) # Дисперсия

standard\_deviation\_3 <- sqrt(variance\_3)

se\_3 <- standard\_deviation1/sqrt(n3)

df\_finish <- data.frame(N = c(n1,n2,n3), SE = c(se\_1, se\_2, se\_3))

df\_finish %>% write\_delim("hw3/task3\_df.xls", delim = "\t")

Код лежит в репозитории: <https://github.com/golubnikova/BioStat_2022/tree/main/biostat/hw3>

В результате запуска кода получим следующий результат:

Таблица 7 – Значения при N = 10, 40, 160

|  |  |
| --- | --- |
| **N** |  |
| 10 | 0,21 |
| 40 | 0,10 |
| 160 | 0,05 |

Представим результат в графическом виде:

Рисунок 2 – График зависимости оценочной стандартной ошибки при значениях N=10, 40, 160

Сравним оценку и истинное значение SE, полученное в п.3. В таблице 8 приведены значения и SE для N = 10, 40, 160.

Таблица 8 – Сравнение оценочной и истинной SE

| **N** |  | **SE** |
| --- | --- | --- |
| 10 | 0,21 | 0,21 |
| 40 | 0,10 | 0,10 |
| 160 | 0,05 | 0,05 |

**Выводы:**

* Чем больше выборка, тем меньше стандартная ошибка ;
* Оценочное значение совпадает с истинным (рассчитанным в п.2).