Асимптотична статистика
4 курс, статистика, Шкляр Ірина Володимирівна

Завдання 1, варіант 7

У роботі ми маємо перевірити за змінною Var1, чи можна стверджувати, що риби належать до різних популяцій на основі тесту χ^2. Отже, нульова гіпотеза –

H0 = “риби належать одній популяції” ,

і альтернативна –

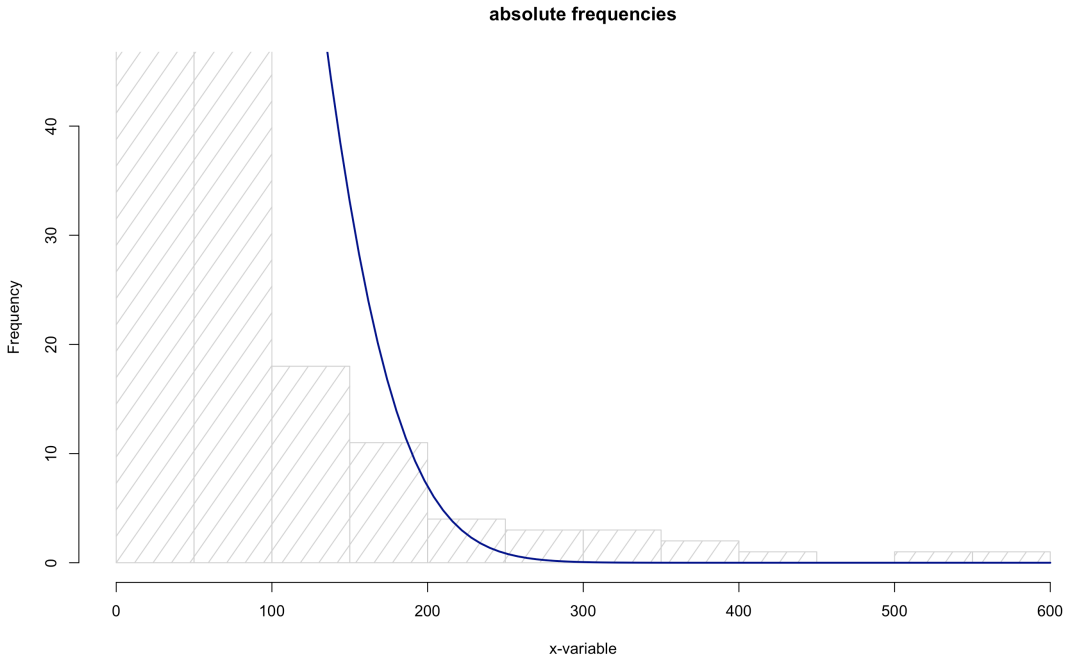
H1 = “риби належать до різних популяцій” .

Інакше кажучи, можемо переформулювати гіпотези:

H0 = “дані мають нормальний або логнормальний розподіл” ,
H1 = “дані мають в сумі нормальний або логнормальний розподіл” .

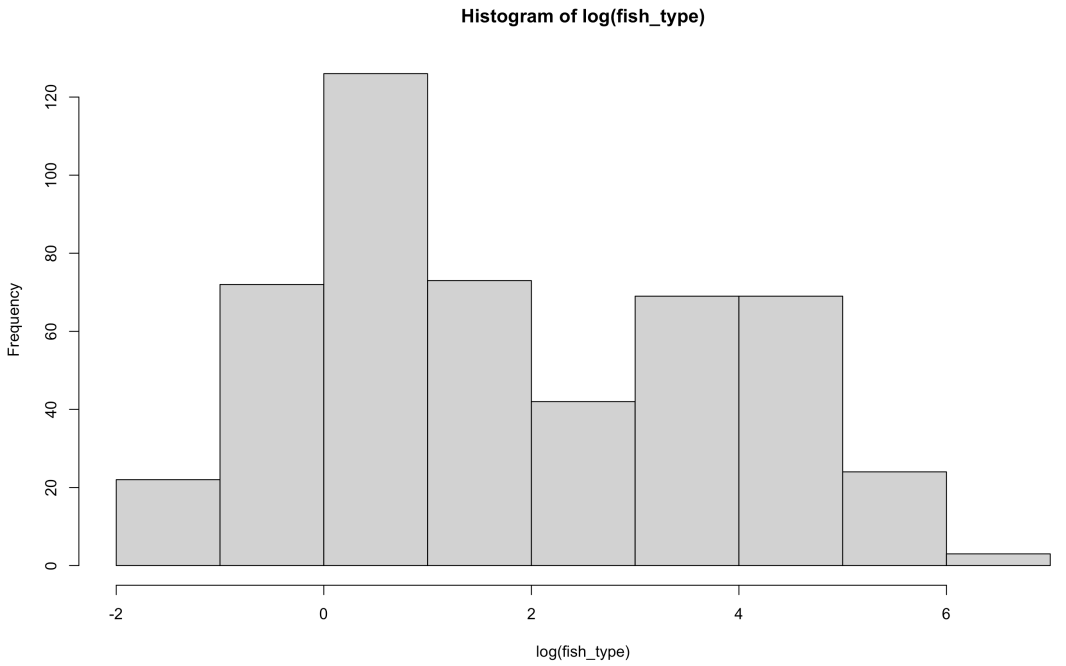
Завантажимо дані для роботи та нарисуємо гістограму даних разом із теоретичною щільністю нормального розподілу:

```
> fish <- read.table(file="~/Downloads/regrasympt/fish.txt",header=T)
>
> fish_type <- fish$Var1
>
> m <- mean(fish_type)
> std <- sqrt(var(fish_type))
>
> hi <- hist(fish_type, density=20, breaks=10, xlab="x-variable", ylim=c(0, 45),
main="absolute frequencies")
> curve(dnorm(x, mean=m, sd=std)*length(fish_type)*(hi$breaks[2]-hi$breaks[1]),
col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



З гістограми бачимо, що розподіл може бути логнормальним. Мені захотілося перевірити, можливо цей розподіл є експоненційним? Тому я прологарифмувала дані і отримала такий графік:

> hist(log(fish_type))



Отриманий графік схожий на суміш двох нормальних розподілів (в сумі вони не завжди дають нормальний розподіл). Тому, на мій погляд, маємо отримати відхилення нульової гіпотези.

Перевіримо, чи дійсно це є так, використовуючи тест хі-квадрат.
Для застосування тесту ми оцінимо математичне сподівання m та середньоквадратичне відхилення sd за групованими даними, а при оцінці sd використаємо поправку Шеппарда.

```
> r <- hist(fish_type, breaks=10, plot = FALSE)
>
> nn<-r$counts                  # емпіричні частоти
> tt<-r$breaks                  # межі комірок
> h<-tt[2]-tt[1]                # ширина комірки
> x<-tt[-length(tt)]+h/2        # середини комірок
> m<-sum(x*nn)/sum(nn)          # оцінка мат сподівання

# оцінка середньо-квадратичного відхилення з поправкою Шеппарда
> s<-sqrt(sum((x-m)^2*nn)/sum(nn)+h^2/12)
> pp<-pnorm(tt,mean=m,sd=s)                          # теоретичні імовірності
> pp[c(1,length(tt))]<-c(0,1)                        # розширюємо крайні комірки
> nth<-length(fish_type)*(pp[-1]-pp[-length(pp)])    # теоретичні частоти
> chi2emp<-sum((nn-nth)^2/nth)                       # статистика тесту хі-квадрат
> 1-pchisq(chi2emp,df=length(tt)-4)                  # досягнений рівень значущості
```

Отримали результат 0. Отже, досягнутий рівень значущості p = 0 (можливо дуже близький до нуля, такий, що R округлив це значення). А тому дійсно ми маємо відхилити основну гіпотезу (дані не є нормально чи логнормально розподіленими), отже, риби належать до різних популяцій.