آزمون نرمال بودن داده‌ها

الف- آزمون شاپیرو-ویلک

یکی از آزمون هایی که برای فرض نرمال بودن داده‌ها وجود دارد، آزمون شاپیرو-ویلک است که به صورت زیر می‌توان انجام داد.

```
>shapiro.test(x)
```

مثال: در مجموعه داده آمادهٔ women در R که مربوط به قد و وزن ۱۵ نفر از زنان آمریکایی است، می‌خواهیم بررسی کنیم که آیا قد و وزن این افراد دارای توزیع نرمال هستند یا خیر.

```
>women
     height    weight
1      58        115
2      59        117
3      60        120
⋮      ⋮          ⋮

> Shapiro.test(women[,1])

        Shapiro-Wilk normality test

W = 0,96  ,  P-Value = 0,75

>Shapiro.test(Women[,2])
```

تمرین: در مجموعهٔ داده آمادهٔ CO2 در R که مربوط به جذب دی اکسید کربن در گیاهان علفی است، بررسی کنید آیا میزان جذب (متغیر uptake) دارای توزیع نرمال هست یا خیر؟!

تمرین ۲: در مجموعهٔ دادهٔ Info_1 که در تکلیف دوم از آن استفاده کرده اید، بررسی کنید آیا فشارخون افراد براساس جنسیت دارای توزیع نرمال هست؟!

(به تفکیک جنسیت بررسی کنید آیا توزیع فشارخون (BP) نرمال است؟)

ب- آزمون کلموگروف-اسمیرنوف

آزمون کلموگروف-اسمیرنوف نیز برای فرض نرمال بودن داده‌ها مورد استفاده قرار می‌گیرد.
این آزمون در مقایسه با آزمون شاپیرو-ویلک از توان آزمون کمتری برای بررسی نرمال بودن
داده‌ها برخوردار است. این آزمون با تابع ks.test( ) در R قابل انجام است.

```
>ks.test(x, "pnorm", mean(x), sd(x))
>ks.test(women[,1], "pnorm", mean(women[,1]), sd(women[,1]))
```

ج- آزمون نرمال بودن با استفاده از آزمون نیکویی-برازش

در این روش ابتدا باید داده‌های پیوسته را به صورت رسته‌ای بنویسیم و طبقه‌بندی کنیم و
سپس باقی مراحل را همانند آنچه که در آزمون نیکویی برازش گفته شد انجام می‌دهیم.
به عنوان مثال می‌خواهیم بررسی کنیم آیا میزان جذب (متغیر uptake) در مجموعه داده
CO2 دارای توزیع نرمال است یا خیر.

```
> x = CO2[,5]
```
→ برای راحتی کار، داده‌های این متغیر را داخل بردار x قرار دادم.

```
>range(x)
[1]  7.7   45.5
```
→ بدست آوردن دامنه تغییرات متغیر x

```
>max(x) - min(x)
[1]  37.8
```

حال باید داده‌ها را به صورت رسته‌ای بنویسیم. برای این کار فرض کنید می‌خواهیم داده‌ها را به
۶ رسته تقسیم بندی کنیم. بنابراین طول هر دسته را بدست می‌آوریم.

```
>(max(x) - min(x))/6
[1]  6.3
```

حال دسته های ما به صورت زیر می شوند.

$$
\left.
\begin{array}{ll}
7.7 \leq x < 14 \quad \equiv \quad x < 14 & \\
14 \leq x < 20.3 & \\
20.3 \leq x < 26.6 & \\
26.6 \leq x < 32.9 & \\
32.9 \leq x < 39.2 & \\
x \geq 39.2 &
\end{array}
\right\}
\begin{array}{l}
\to Z = 1 \\
\to Z = 2 \\
\to Z = 3 \\
\to Z = 4 \\
\to Z = 5 \\
\to Z = 6
\end{array}
$$

و مرحله بعد بدست آوردن فراوانی های هر دسته است که می توان این کار را به صورت زیر باهم ادغام کنیم:

```
>W=c(   )
>W[1]=Length(which(x<14))
>W[2]=Length(which(x<20.3 & x≥14))
   ⋮
>W[6]=Length(which(x≥39.2))
>W
[1]  13  17  7  18  15  14
```

بنابراین نمایش دسته‌ای داده‌ها به صورت زیر می‌شود:

دسته ها	Z=1	Z=2	Z=3	Z=4	Z=5	Z=6
فراوانی ها	13	17	7	18	15	14

آزمون مسأله:

$$
\begin{cases}
H_0: & X \sim N(\mu, \sigma^2) \\
H_1: & O.W
\end{cases}
$$

حال باید احتمال های مورد نظر با هر دسته را محاسبه کنیم. (تحت فرض H_0)

$$P_1 = P_{01} = P(Z=1) = P(X<14)$$ $$X \overset{H_0}{\sim} N(\mu, \sigma^2)$$

چون μ و σ^2 نامعلوم پس باید برآورد کنیم.

$$\hat{\mu} = \bar{X} \qquad \hat{\sigma}^2 = S_x^2$$

```
>mean(x) -> mu          ≡     >mu = mean(x)
>sd(x) -> sd                  >sd = sd(x)
```

$$\Rightarrow P_{01} = P_{H_0}(X<14) = P\left(\frac{X-mu}{sd} \leq \frac{14-mu}{sd}\right) = \Phi\left(\frac{14-mu}{sd}\right)$$

به همین ترتیب باید احتمال های هر دسته را محاسبه کنیم.

```
>P=c( )
>P[1]=pnorm(14, mean=mu, sd=sd)
>P[2]=pnorm(20.3, mu, sd) -pnorm(14, mu, sd)
 ⋮
>P[6]=pnorm(39.2, mu, sd, lower.tail=F)
```

پس به این ترتیب بردار احتمالات مربوط به هر دسته مشخص شد.

★ چون برای انجام این آزمون، دو پارامتر را محاسبه کردیم بنابراین باید این آزمون را بدون استفاده از تابع `Chisq.test( )` انجام دهیم.

```
>n=length(x)
>n
[1]  84
>e=c( )
>e=n*p   ->
```

بردار مربوط به مقادیر مورد انتظار

حال باید چک کنیم ببینیم آیا مقادیر مورد انتظار از 5 بالاتر هستند یا نه که در این مثال تمامی مقادیر e بالای 5 هستند.

حال باید مقدار آماره آزمون (X_0^2) و P-value را بدست بیاوریم

$$X_0^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\text{P-value} = P(X^2 > X_0^2) \qquad X^2 \overset{H_0}{\sim} \chi^2_{(k-1-2)}$$

6

۲ پارامتر برآورد کردیم

```
> X2 = sum(((w-e)^2)/e)

> P_value = pchisq(X2, 3, lower.tail=F)
```

تمرین ۳ : بررسی کنید آیا فشارخون افراد حاضر در فایل info_1 به تفکیک جنسیت دارای توزیع نمایی هستند یا خیر.

آزمون اول
$H_0: BP_0 \sim \exp(\lambda)$

فشارخون افراد با کد جنسیت ۰

$H_1:$ O.W

آزمون دوم
$H_0: BP_1 \sim \exp(\lambda)$

$H_1:$ O.W