

# örnekleme\_alıştırmaları

Hakan Mehmetcik

2024-11-14

## Alıştırma 1:

Bu alıştırmada, örnekleme verilerinin bir popülasyon hakkında gerçeği keşfetmek için nasıl kullanılabileceğini inceleyeceğiz. Popülasyon verilerimiz, Texas Üniversitesi Austin'deki istatistik öğrencilerinden toplanmıştır.

```
library(here)
```

here() starts at /Users/kobain/Desktop/IST2083

```
# veriyi R'a okutalım  
survey <- read.csv(here("data", "StudentSurvey.csv"))
```

## Soru 1:

Bu veri setinde kaç öğrenci bulunmaktadır?

```
str(survey)
```

```
'data.frame':  379 obs. of  17 variables:  
 $ ID          : int  1 2 3 4 5 6 7 8 9 10 ...  
 $ gender      : chr  "Male" "Female" "Female" "Male" ...  
 $ age         : int  17 19 18 19 19 19 18 18 19 19 ...  
 $ classification: chr  "Freshman" "Sophomore" "Freshman" "Sophomore" ...  
 $ name_letters : int  4 8 6 4 7 5 5 8 8 5 ...  
 $ happy       : int  80 76 50 75 89 90 57 60 75 100 ...  
 $ concerts    : int  2 15 3 0 1 1 0 0 1 4 ...  
 $ hair_color  : chr  "black" "black" "brown" "black" ...  
 $ own_shoes   : int  108 42 6 10 13 12 12 6 15 25 ...
```

```

$ greek      : chr  "yes" "no" "no" "yes" ...
$ live_campus : chr  "yes" "no" "yes" "no" ...
$ roomates   : int   1 1 1 1 2 3 1 1 1 3 ...
$ austin     : num   10 10 10 8 9 10 7 8 8 8 ...
$ birth_month : int   11 12 11 5 5 12 8 9 6 3 ...
$ commute    : chr   "walk" "drive by myself" "walk" "bus" ...
$ car        : chr   "no" "no" "no" "no" ...
$ sport      : chr   "yes" "no" "no" "no" ...

```

## Soru 2:

Veri setindeki ilk 10 öğrenciden kaç 5 harften uzun bir isme sahiptir?

```

head_survey <- head(survey, 10)

which(head_survey$name_letters>5)

```

```
[1] 2 3 5 8 9
```

## Soru 3:

Veri setinde %40'tan daha az mutlu olan ilk öğrencinin isminin uzunluğu nedir?

```

happy <- survey[survey$happy<40,]
head(happy)

```

	ID	gender	age	classification	name_letters	happy	concerts	hair_color
11	11	Male	18	Junior	7	2	0	black
28	28	Female	19	Sophomore	4	35	0	black
61	61	Female	19	Freshman	5	0	0	black
75	75	Male	18	Freshman	5	20	0	black
169	169	Female	18	Freshman	6	25	0	brown
232	232	Female	25	Junior	8	30	0	black

	own_shoes	greek	live_campus	roomates	austin	birth_month	commute	car	sport
11	5	no	no	1	4	5	bus	no	yes
28	4	no	no	1	5	7	walk	no	yes
61	6	no	yes	1	9	9	walk	no	no
75	5	no	no	3	6	12	bus	no	no
169	7	no	yes	1	7	12	walk	no	no
232	6	no	yes	0	7	11	walk	no	no

#### Soru 4:

Hangi değişken, her öğrencinin ismindeki harf sayısını ifade eder? Veri setinde bu değişkenin adı nedir ve bu değişken kategorik mi yoksa nicel mi?

```
str(survey)
```

```
'data.frame':  379 obs. of  17 variables:
 $ ID          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ gender      : chr  "Male" "Female" "Female" "Male" ...
 $ age        : int  17 19 18 19 19 19 18 18 19 19 ...
 $ classification: chr  "Freshman" "Sophomore" "Freshman" "Sophomore" ...
 $ name_letters : int  4 8 6 4 7 5 5 8 8 5 ...
 $ happy      : int  80 76 50 75 89 90 57 60 75 100 ...
 $ concerts   : int  2 15 3 0 1 1 0 0 1 4 ...
 $ hair_color  : chr  "black" "black" "brown" "black" ...
 $ own_shoes   : int  108 42 6 10 13 12 12 6 15 25 ...
 $ greek      : chr  "yes" "no" "no" "yes" ...
 $ live_campus : chr  "yes" "no" "yes" "no" ...
 $ roomates    : int  1 1 1 1 2 3 1 1 1 3 ...
 $ austin     : num  10 10 10 8 9 10 7 8 8 8 ...
 $ birth_month : int  11 12 11 5 5 12 8 9 6 3 ...
 $ commute    : chr  "walk" "drive by myself" "walk" "bus" ...
 $ car        : chr  "no" "no" "no" "no" ...
 $ sport      : chr  "yes" "no" "no" "no" ...
```

#### Soru 5:

Bir örnekleme dağılımı neyi ifade eder?

**Cevap:** Bir örnekleme dağılımı, örnek istatistiklerinin dağılımıdır. Örneğin, örneklem ortalamalarının dağılımı gibi.

#### Soru 6:

Merkezi Limit Teoremi, örnekleme dağılımları hakkında ne söyler?

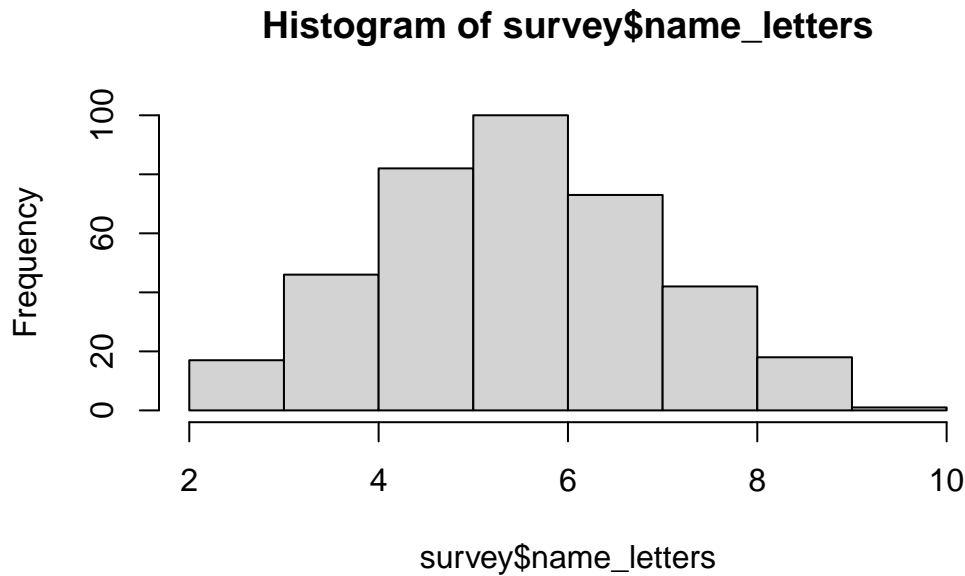
**Cevap:**

1. Örnek büyüklüğü arttıkça dağılım daha çok normal bir görünüme sahip olur.
2. Örneklem ortalamalarının değişkenliği, örnek büyüklüğü arttıkça azalır.
3. Örnekleme dağılımının merkezi, popülasyon ortalamasına eşittir.

**Soru 7:**

Tipik bir Texas Üniversitesi öğrencisinin ismi kaç harften oluşur? Örneklem büyüklüğünü artırdıkça tahminimiz nasıl değişir?

```
hist(survey$name_letters)
```



```
# Calculate the population parameters  
summary(survey$name_letters)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	5.000	6.000	5.971	7.000	10.000

```
# By how many letters, on average, do names vary from the mean? (sd)  
sd(survey$name_letters)
```

```
[1] 1.49486
```

## Örnekleme Çalışması

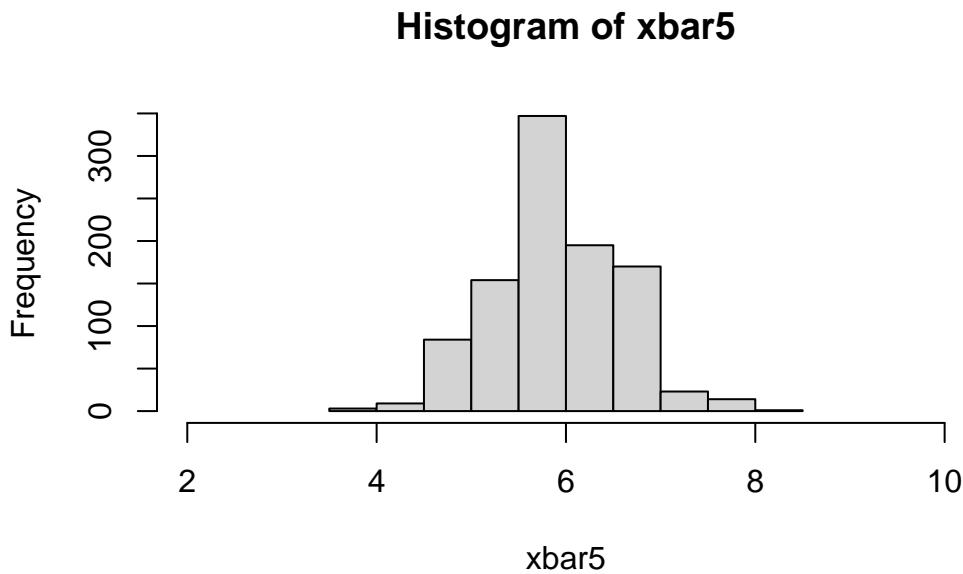
### 1. Örneklem Büyüklüğü $n=5$

1. 1000 örneklem çizelim ve her birinin ortalamasını hesaplayalım:

```
xbar5 <- rep(NA, 1000)
for (i in 1:1000) {
  x <- sample(survey$name_letters, size = 5)
  xbar5[i] <- mean(x)
}
```

2. Örnekleme dağılımının histogramını çizelim:

```
# graph the histogram of 1000 sample means
hist(xbar5, xlim = c(2,10))
```



3. Örnekleme dağılımının ortalamasını ve standart hatasını hesaplayalım:

```
# calculate mean and sd for sampling distributions
mean(xbar5)
```

```
[1] 5.954
```

```
sd(xbar5)
```

```
[1] 0.6669999
```

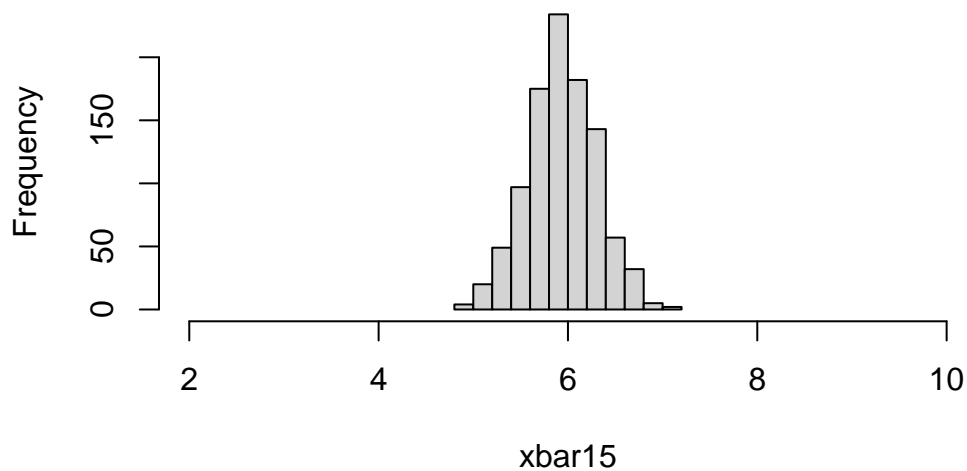
## 2. Örneklem Büyüklüğü n=15

1. Aynı işlemi n=15 için tekrarlayalım:

```
# Repeat for samples of size n=15
xbar15 <- rep(NA, 1000)
for (i in 1:1000) {
  x <- sample(survey$name_letters, size = 15)
  xbar15[i] <- mean(x)
}
```

```
# check xbar15 hist
hist(xbar15, xlim = c(2,10))
```

**Histogram of xbar15**



```
# mean and sd for xbar15
mean(xbar15)
```

```
[1] 5.9712
```

```
sd(xbar15)
```

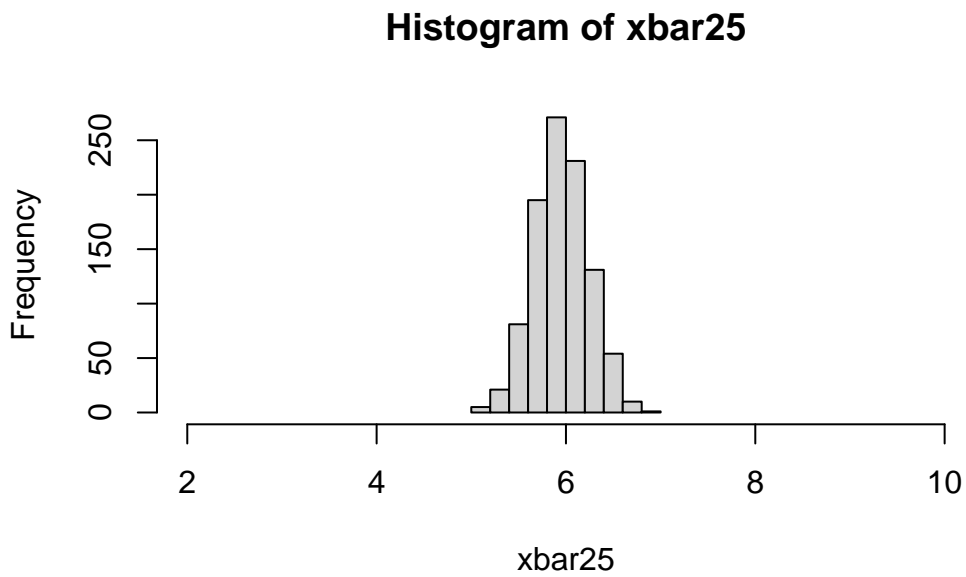
```
[1] 0.3658992
```

### 3. Örneklem Büyüklüğü n=25

1. Aynı işlemi n=25 için tekrarlayalım

```
# Repeat for samples of size n=25
xbar25 <- rep(NA, 1000)
for (i in 1:1000) {
  x <- sample(survey$name_letters, size = 25)
  xbar25[i] <- mean(x)
}
```

```
# check xbar25 hist
hist(xbar25, xlim = c(2,10))
```



```
mean(xbar25)
```

```
[1] 5.97088
```

```
sd(xbar25)
```

```
[1] 0.2890432
```

### Örnekleme Dağılımlarını Gözlemleme

1. Üç örnekleme dağılımı için ortalama yaklaşık olarak aynıydı.
2. Örneklem büyüklüğü 5'ten 25'e çıktıkça standart hata azaldı.
3. Örneklem büyüklüğü arttıkça dağılımlar daha çok normalleşti.

### Sonuç

Bu alıştırmada, üniversite öğrencilerinin isim uzunluklarının popülasyon ortalamasını bildiğimizi varsaydık:

- **Popülasyon Ortalaması:** 5.97 harf
- **Popülasyon Standart Sapması:** 1.49 harf

**Merkezi Limit Teoremini** simüle etmek için farklı boyutlarda örneklemeler çektik. Merkezi Limit Teoremi şunları söyler:

1. Örneklem büyüklüğü arttıkça, örnekleme dağılımları daha normal hale gelir.
2. Örnekleme dağılımının ortalaması, popülasyon ortalamasına eşittir.
3. Örneklem ortalamalarının değişkenliği, popülasyon standart sapmasının örneklem büyüklüğünün kareköküne bölünmesiyle tahmin edilebilir.

Simülasyon sonuçlarımız bu teoriyi doğruladı. Örneklem büyüklüğünü 5'ten 25'e çıkardıkça, örneklem ortalamalarının değişkenliği azaldı ve ortalamalar gerçek popülasyon ortalamasına daha sıkı bir şekilde yaklaştı. Başka bir deyişle, örneklem ortalamaları, popülasyon ortalamasını daha iyi tahmin etmeye başladı.



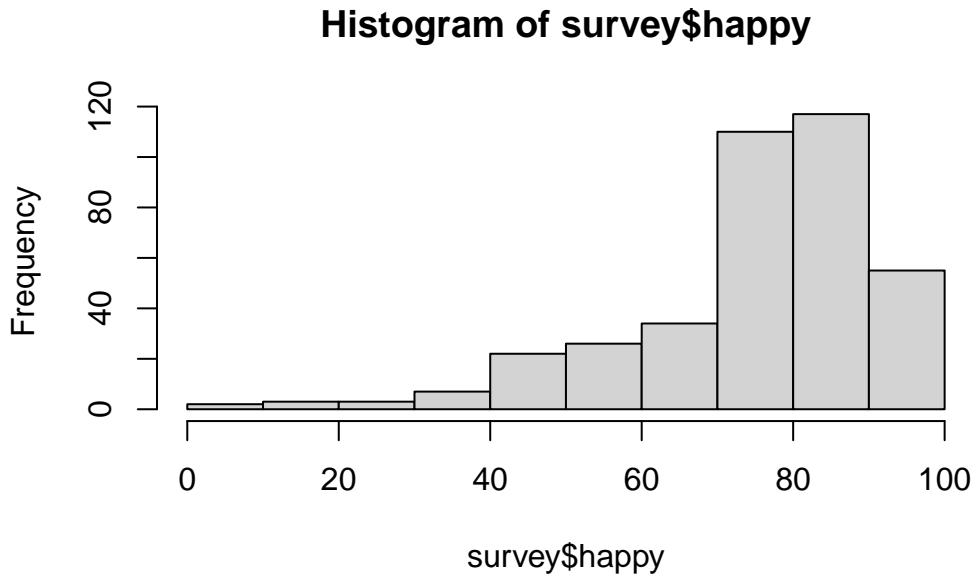
## Alıştırma 2:

Üniversite öğrencileri zamanlarının ne kadarında mutludur? Örneklem büyüklüğü arttıkça gerçek ortalama tahminimiz nasıl değişir?

### Soru 1:

Popülasyonun mutluluk skorlarının şekli nedir?

```
hist(survey$happy)
```



### Soru 2:

Üniversite öğrencileri zamanlarının ortalama olarak ne kadarında mutludur?

```
summary(survey$happy)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	70.00	80.00	78.03	90.00	100.00

### Soru 3:

Mutluluk yüzdesi skorlarının standart sapması nedir?

```
sd(survey$happy)
```

```
[1] 16.30906
```

### Farklı Örneklem Büyüklüklerinde 1.000 Örneklem Çekme

#### Örneklem büyüklüğü n=5

```
xbar5 <- rep(NA, 1000)

for (i in 1:1000) {
  x <- sample(survey$happy, size = 5)
  xbar5[i] <- mean(x)
}
```

#### Örneklem büyüklüğü n=15

```
xbar15 <- rep(NA, 1000)

for (i in 1:1000) {
  x <- sample(survey$happy, size = 15)
  xbar15[i] <- mean(x)
}
```

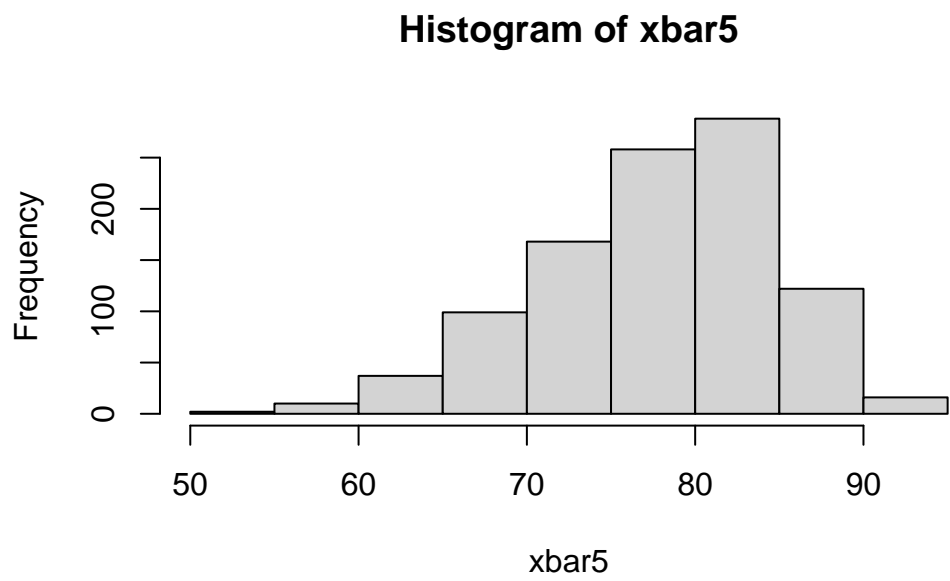
#### Örneklem büyüklüğü n=25

```
xbar25 <- rep(NA, 1000)

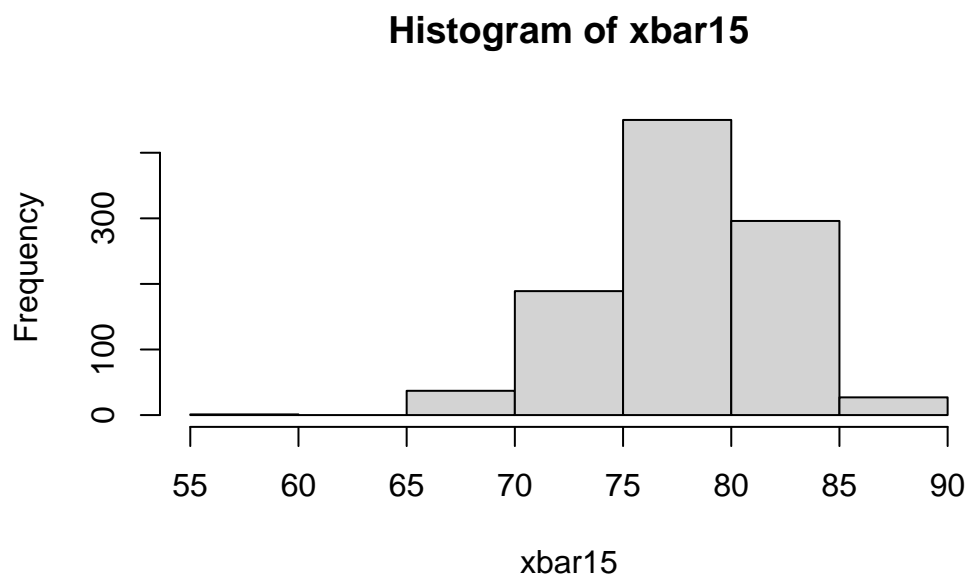
for (i in 1:1000) {
  x <- sample(survey$happy, size = 25)
  xbar25[i] <- mean(x)
}
```

### Örnekleme Dağılımlarının Histogramları

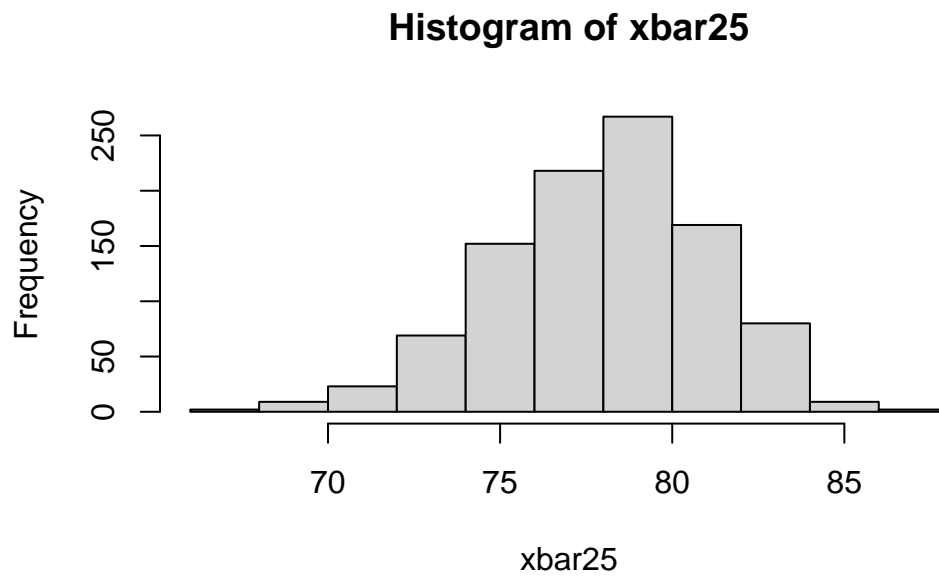
```
hist(xbar5)
```



```
hist(xbar15)
```



```
hist(xbar25)
```



Örneklem Ortalamaları

```
mean(xbar5)
```

```
[1] 78.0464
```

```
mean(xbar15)
```

```
[1] 78.00413
```

```
mean(xbar25)
```

```
[1] 77.96416
```

Örneklem Standart Sapmaları

```
sd(xbar5)
```

```
[1] 6.97502
```

```
sd(xbar15)
```

```
[1] 4.08698
```

```
sd(xbar25)
```

```
[1] 3.062461
```

### Örnekleme Dağılımları İçin Gözlemler:

1. Tüm örnekleme dağılımlarında ortalama yaklaşık olarak aynıydı.
2. Örneklem hatası (SESESE) örneklem büyüklüğü arttıkça azaldı.
3. Örneklem büyüklüğü arttıkça dağılımlar daha normal hale geldi.

**Genel Sonuç:** Eğer örneklem büyüklüğü yeterince büyükse, örnekleme dağılımı, popülasyonun şekli ne olursa olsun, normal bir dağılıma sahip olacaktır.

---

### Sonuç Çıkarımı

Yukarıdaki alıştırmada, üniversite öğrencilerinin zamanlarının ne kadarında mutlu olduklarının popülasyon ortalamasını biliyorduk:

- **Popülasyon Ortalaması:** 78.0378.0378.03
- **Popülasyon Standart Sapması:** 16.3116.3116.31
- **Mutluluk Skorlarının Şekli:** Negatif çarpık

Farklı büyüklüklerde örneklem çekerek **Merkezi Limit Teoremi’ni (CLT)** simüle ettik. Merkezi Limit Teoremi şu üç şeyi söyler:

1. Örneklem büyüklüğü arttıkça, örnekleme dağılımları daha normal hale gelir.
2. Örnekleme dağılımının ortalaması, popülasyon ortalamasına eşit olur.

3. Örneklem ortalamalarının değişkenliği veya standart hatası, popülasyon standart sapmasının örneklem büyüklüğünün kareköküne bölünmesiyle tahmin edilebilir.

**Simülasyon Sonuçları:** Simülasyon sonuçlarımız bu teoriyle uyumluydu:

- Örneklem büyüklüğünü 5'ten 25'e çıkardıkça, örneklem ortalamalarının değişkenliği azaldı.
- Örneklem ortalamaları, gerçek popülasyon ortalamasına daha sıkı bir şekilde yaklaştı.
- Dağılımın şekli örneklem büyüklüğü arttıkça daha normal hale geldi.

Başka bir deyişle, örneklem ortalamaları, popülasyon ortalamasını daha iyi tahmin etmeye başladı.