

医学統計勉強会

Rの導入と使い方

帝京大学臨床研究センター（TARC）・帝京大学大学院公衆衛生学研究科 共催

帝京大学大学院公衆衛生学研究科

宮田 敏

Numerical summary: Location

データの**位置** (location) に関する要約。

x_1, x_2, \dots, x_n : 観察された標本。 n : 標本数。

平均 (Mean) : $\bar{x} = \frac{x_1 + \dots + x_n}{n} = n^{-1} \sum_{i=1}^n x_i$

中央値 (Median) : データを、最小の $x_{(1)}$ から最大の $x_{(n)}$ まで並べ直したものを $x_{(1)}, \dots, x_{(n)}$ とする。

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & : n \text{ が奇数} \\ (x_{(n/2)} + x_{(n/2+1)}) / 2 & : n \text{ が偶数} \end{cases}$$

Locationに関する, その他の要約

- **Percentile (パーセント点)**: $k\%$ percentile はデータの中の点で, 標本の $k\%$ より大きく, $(100 - k)\%$ より小さい点.
- **Quartile (四分位点)**: The first quartile (第一四分位点) = 25% percentile. The third quartile (第三四分位点) = 75% percentile.
- **Trimmed mean (刈り込み平均)**: $k\%$ trimmed mean は, データから上下 $k\%$ を取り除いた後の平均.
- **Five numbers summary**:
(min., 1st quartile, median, 3rd quartile, max.)

Rによる要約統計量の 計算 (平均, etc.)

```
> x <- rnorm(100) ## 100 random numbers 数値例 ##
>
> mean(x) ## mean ##
[1] -0.3011125
>
> median(x) ## median ##
[1] -0.2836064
>
> quantile(x) ## quantile ##
      0%      25%      50%      75%     100%
-2.6852655 -0.8990262 -0.2836064  0.4137969
 1.9382667
>
> quantile(x, 0.1) ## 10% percentile ##
      10%
-1.480168
>
> mean(x, trim=0.1) ## 10% trimmed mean ##
[1] -0.2875766
> mean(x, trim=0.5) ## 50% trimmed mean ##
[1] -0.2836064
>
> summary(x) ## Five numbers summary ##
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.6850 -0.8990 -0.2836 -0.3011  0.4138  1.9380
```

Numerical summary: Variance

データの広がり (分散, **variance**) に関する要約。

x_1, x_2, \dots, x_n : 観察された標本。 n : 標本数。

分散 (variance) : 個々の標本と標本平均との二乗距離の平均。

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

標準偏差 (Standard Deviation) $s = \sqrt{s^2}$

四分位点間距離 (Inter Quartile Range, IQR) :

$$f_s = (3^{\text{rd}} \text{ quartile} - 1^{\text{st}} \text{ quartile}) = (\text{第三四分位点} - \text{第一四分位点})$$

2025/9/24

帝京大学 医学統計勉強会 第2回Rの導入と使い方

5

**Rによる要約統計量の
計算 (分散, etc.)**

```
> var(x)## variance ##  
[1] 0.9881656  
>  
> sd(x)## standard deviation ##  
[1] 0.9940652  
> sqrt(var(x))  
[1] 0.9940652  
>  
> IQR(x) ## inter quantile distance ##  
[1] 1.312823
```

“Continuous variables were expressed as **mean \pm SD** or **median (interquartile range)**, as appropriate.”

- **Mean \pm SD** (Standard deviation): 平均(Mean)を中心にMean \pm SDの範囲に、**データ全体の60~70%**が分布している。
- **Mean \pm SE** (Standard error): **標準誤差** = **標本平均の標準偏差** s/\sqrt{n}

データ全体の散らばりの範囲に興味があれば **Mean \pm SD** が第一選択。

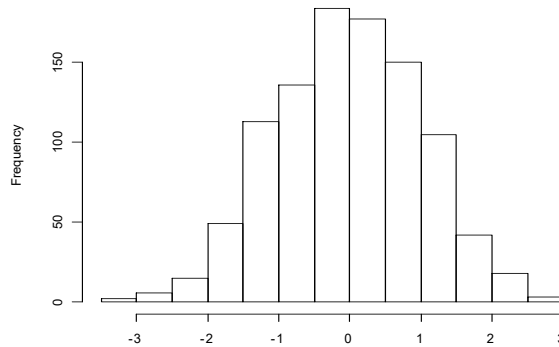
2025/9/24

帝京大学 医学統計勉強会 第2回Rの導入と使い方

6

Graphical summary: Histogram

- 階級 (Classes/Bins): Sub-interval of the sample range
- 度数 (Frequency): それぞれの階級のなかの標本数.
- 相対度数 (Relative Frequency): = 度数/標本数.
- ヒストグラム (Histogram): 頻度もしくは相対頻度を表した棒グラフ.



2025/9/24

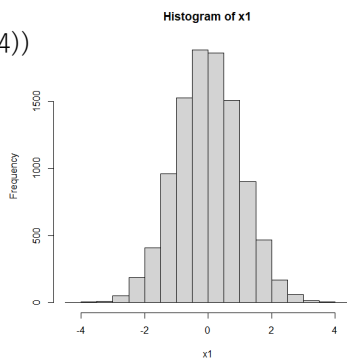
帝京大学 医学統計勉強会 第2回Rの導入と使い方

7

N = 10000

```
> hist(x1, ylim=c(0, 1884))
```

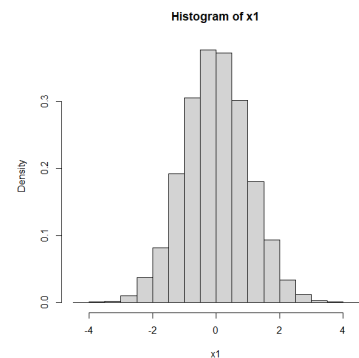
frequency



N = 10000

```
> hist(x1, freq=F)
```

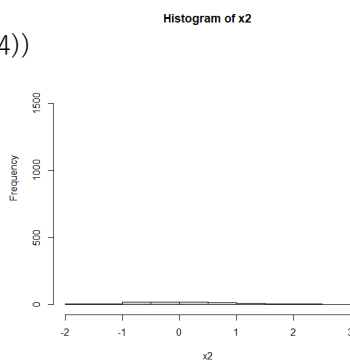
relative frequency



N = 100

```
> hist(x2, ylim=c(0, 1884))
```

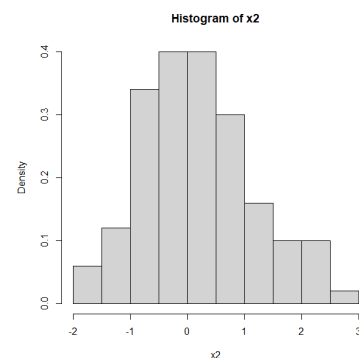
frequency



N = 100

```
> hist(x2, freq=F)
```

relative frequency



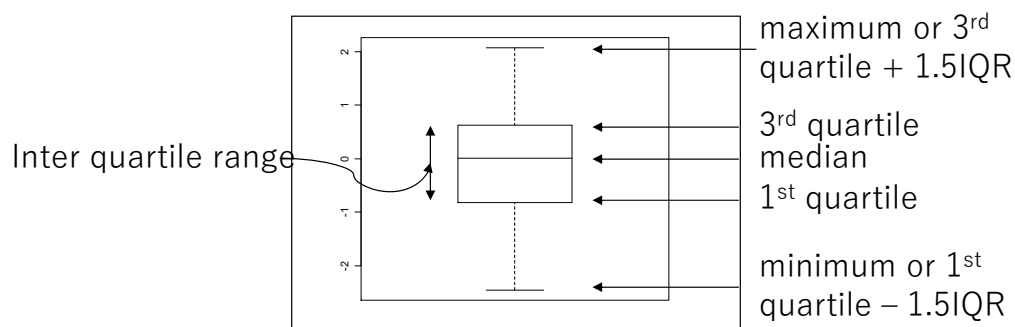
2025/9/24

市原ハナ 医学統計勉強会 第2回Rの導入と使い方

8

Graphical summary: Box-plot

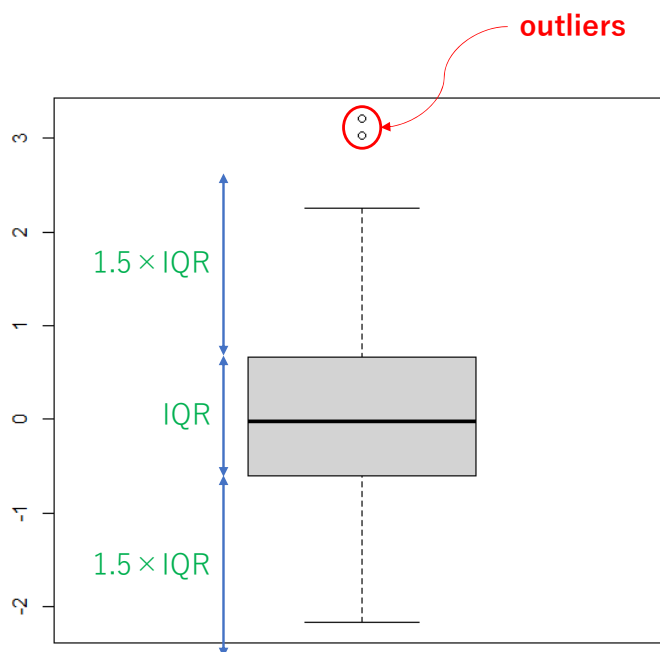
1)縦軸に変数値をとる. 2)下限が1st quartile、上限が3rd quartileとなる”Box”を描く. 3)medianの位置に線を描く. 4)Boxの上下辺からmax., min.まで線を引く. 5)上下辺から $1.5 \times \text{IQR}$ 以上離れた標本ははずれ値 (Outlier) として, 点で表す.



2025/9/24

帝京大学 医学統計勉強会 第2回Rの導入と使い方

9

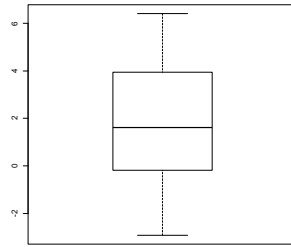
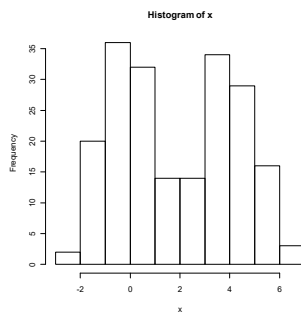


2025/9/24

帝京大学 医学統計勉強会 第2回Rの導入と使い方

10

ヒストグラムとボックスプロット：二峰型

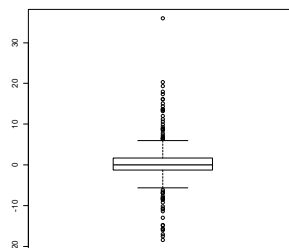
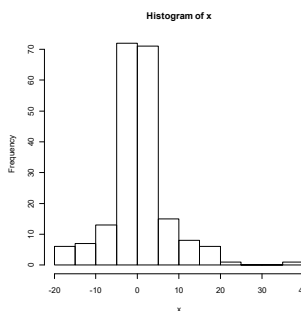


```
x1 <- rnorm(100, mean=0)
x2 <- rnorm(100, mean=4)
x <- c(x1, x2)
hist(x)
boxplot(x)
```

データの分布が「二峰型」の場合、ヒストグラムはその特徴をとらえているが、ボックスプロットではピークが二つあるという特徴がつかめない。

ヒストグラムは分布の特徴の、**全体的な傾向**をとらえるのに適している。

ヒストグラムとボックスプロット：裾が重い



```
x1 <- rnorm(100, mean=0, sd=1)
x2 <- rnorm(100, mean=0, sd=10)
x <- c(x1, x2)
hist(x)
boxplot(x)
```

データの裾が重い分布の場合、ボックスプロットのほうが「極端に大きい（小さい）**異常値**」をとらえるのに適している。

結局、ヒストグラムとボックスプロットは両方検討する必要がある。さらに、このような**分布の形状**に関する情報は、**数値的な要約では得られない**ことに留意する。