

医学統計勉強会

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

医学統計勉強会

9月11日～12月25日（隔週水曜日）

18:30～20:00 大学棟5階 FRUプレゼンテーションルーム

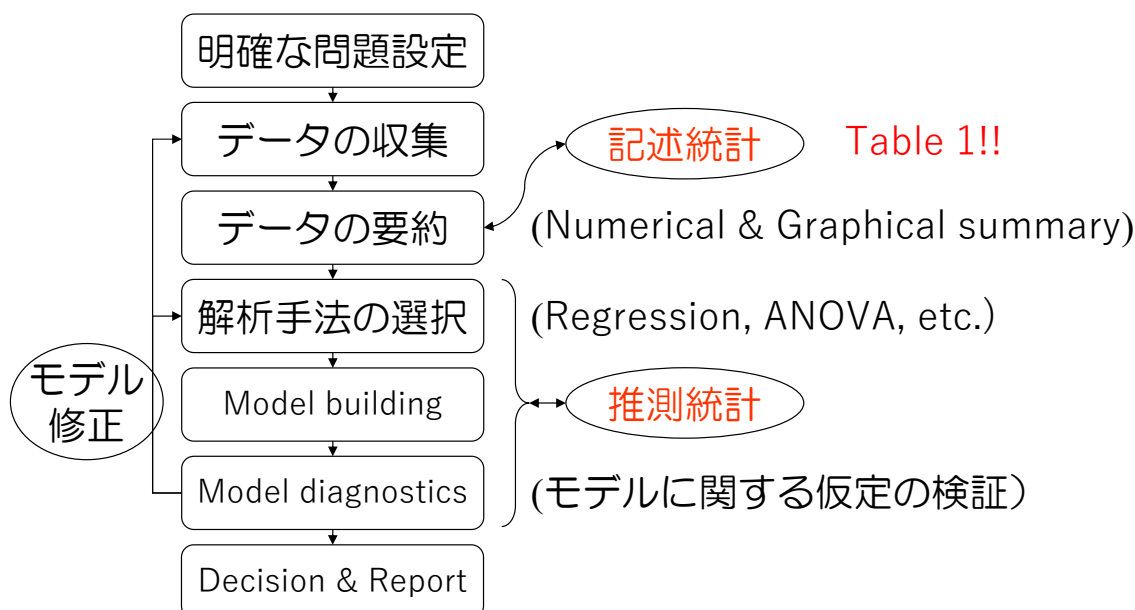
- 第1回 09/11 基本統計量 Table1を究めよう
- 第2回 09/25 推定 信頼区間 仮説検定
- 第3回 10/09 連続変数の比較
- 第4回 10/23 回帰分析
- 第5回 11/13 比率と分割表
- 第6回 11/27 ロジスティック回帰分析
- 第7回 12/11 生存時間解析
- 第8回 12/25 継時的繰り返し測定データの解析

自己紹介：宮田 敏 (smiyata@med.teikyo-u.ac.jp)

1992	4	一橋大学大学院経済学研究科修士課程入学		
1994	3	一橋大学大学院経済学研究科修士課程修了（経済学修士）		
1994	4	一橋大学大学院経済学研究科博士後期課程入学		
1995	4	一橋大学大学院経済学研究科博士後期課程休学		
1995	6	オハイオ州立大学大学院統計学部入学		
2001	8	オハイオ州立大学大学院統計学部卒業（Ph.D. 取得）		
2001	9	文部科学省統計数理研究所 講師着任		
2002	3	文部科学省統計数理研究所 講師退職		
2002	4	財団法人癌研究会ゲノムセンター情報解析部門研究員着任		
2012	3	公益財団法人がん研究会ゲノムセンター研究員 退職		
2012	4	東北大学大学院医学系研究科循環器EBM開発学 着任		
2020	3	東北大学大学院医学系研究科循環器EBM開発学 退職		
2020	4	帝京大学大学院公衆衛生学研究科 着任		

日本循環器学会誌 Circulation Journal, Statistical Consulting Editor, 2012-2016

データ解析のフローチャート



記述統計と推測統計

- **記述統計** (Descriptive statistics) : データを要約し, 分布の形状や重要な性質を記述する.
- **推測統計** (Inferential statistics) : 標本から得られた情報を一般化し, 母集団に関する推測を行う.

記述統計は, 分布の位置 (**location**), 分布の広がり (分散, **variance**), およびその**形状の要約**を得る.

- **Numerical summary** (数量的要約) : データの**代表値**による要約.
- **Graphical summary** (視覚的要約) : **図**による要約.

記述統計 (Table 1) の重要性

- 記述統計はデータを要約し, データの持つ全体的な**特徴**, **傾向**を把握する.
- 同じ目的でも, データの性質により**複数の解析方法**が存在する場合がある. 適切な方法を**選択**するために, データの特徴を把握する.
- サンプルは, 母集団から偏り無く抽出されているはず. 研究で用いられるサンプルが, 想定される母集団とどの程度同じでどの程度異なるかを記述統計によって明らかにする.
- データの収集が, **公正**に行われていることを示す.
 - 比較対照の際, 対照のための条件以外の背景因子に, 極端な差がないことを示す.
 - データに異常な値がないことを確認.

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
```

2024/9/11

帝京大学 医学統計勉強会 第1回基本統計量

9

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) :

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

2024/9/11

帝京大学 医学統計勉強会 第1回基本統計量

10

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。

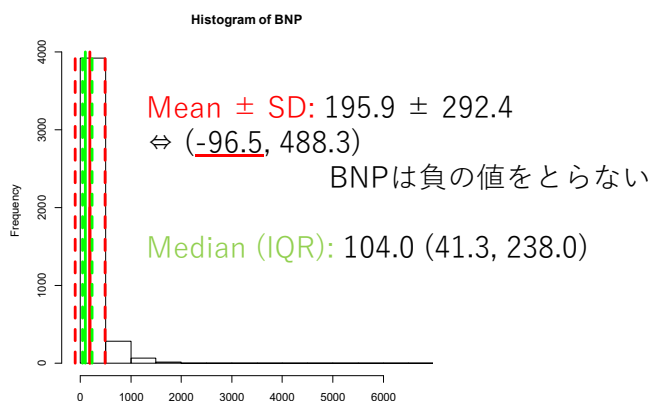
2024/9/11

帝京大学 医学統計勉強会 第1回基本統計量

11

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。

Median (interquartile range, IQR): 中央値(Median)を中心に、IQRの範囲にデータ全体の**50%**が分布している。



Mean \pm SDは、不合理な値(データの範囲を逸脱)をとることがある。

分布が**歪んでいる**ときは、Median (IQR) が第一選択。

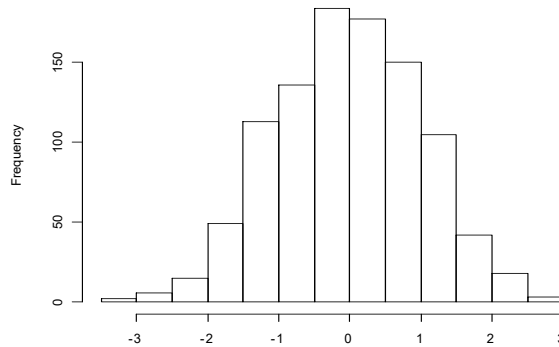
2024/9/11

帝京大学 医学統計勉強会 第1回基本統計量

12

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.



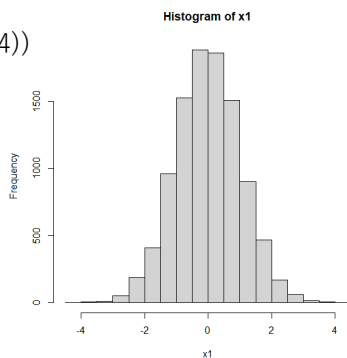
2024/9/11

帝京大学 医学統計勉強会 第1回基本統計量

13

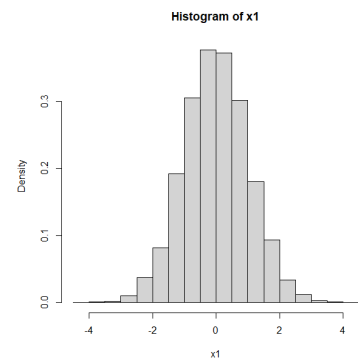
N = 10000
`> hist(x1, ylim=c(0, 1884))`

frequency



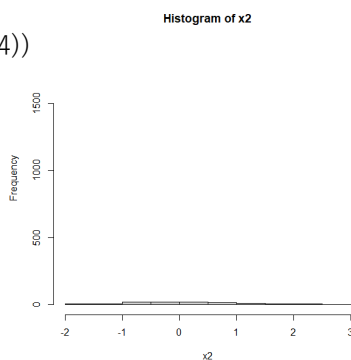
N = 10000
`> hist(x1, freq=F)`

relative frequency



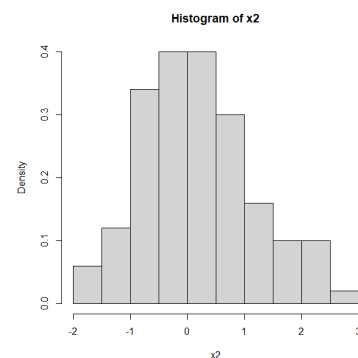
N = 100
`> hist(x2, ylim=c(0, 1884))`

frequency



N = 100
`> hist(x2, freq=F)`

relative frequency



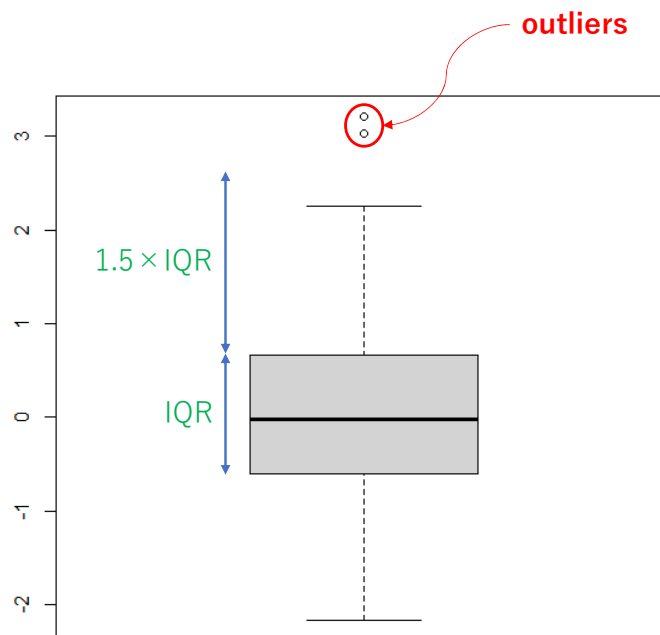
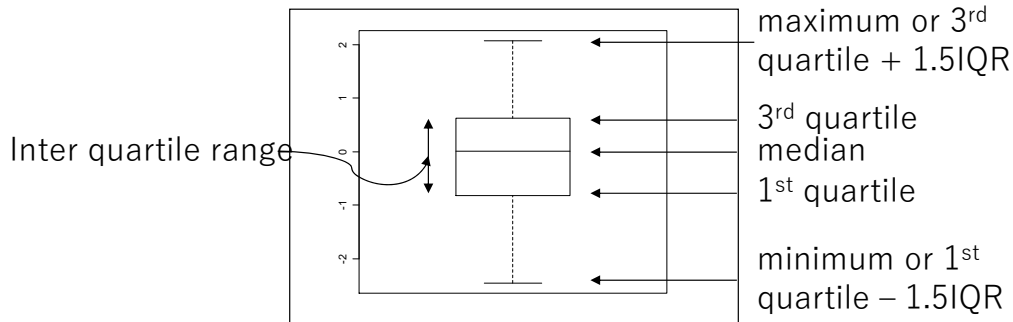
2024/9/11

甲 京 大 学 医 学 統 計 勉 強 会 第 1 回 基 本 統 計 量

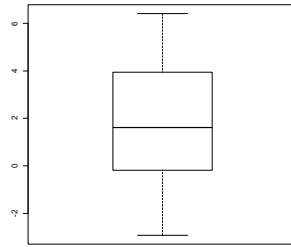
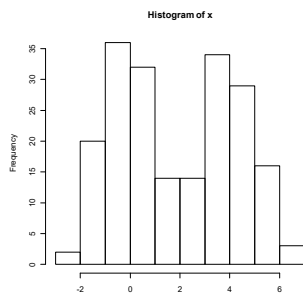
14

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×IQR以上離れた標本ははずれ値 (Outlier) として、点で表す.



ヒストグラムとボックスプロット：二峰型

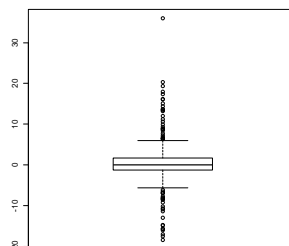
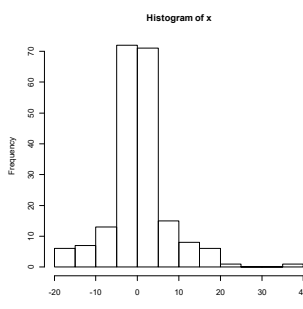


```
x1 <- rnorm(100, mean=0)
x2 <- rnorm(100, mean=4)
x <- c(x1, x2)
hist(x)
boxplot(x)
```

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。