医学統計勉強会

帝京大学臨床研究センター（TARC）・帝京大学大学大学院公衆衛生学研究科　共催

帝京大学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１３日～１２月２０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

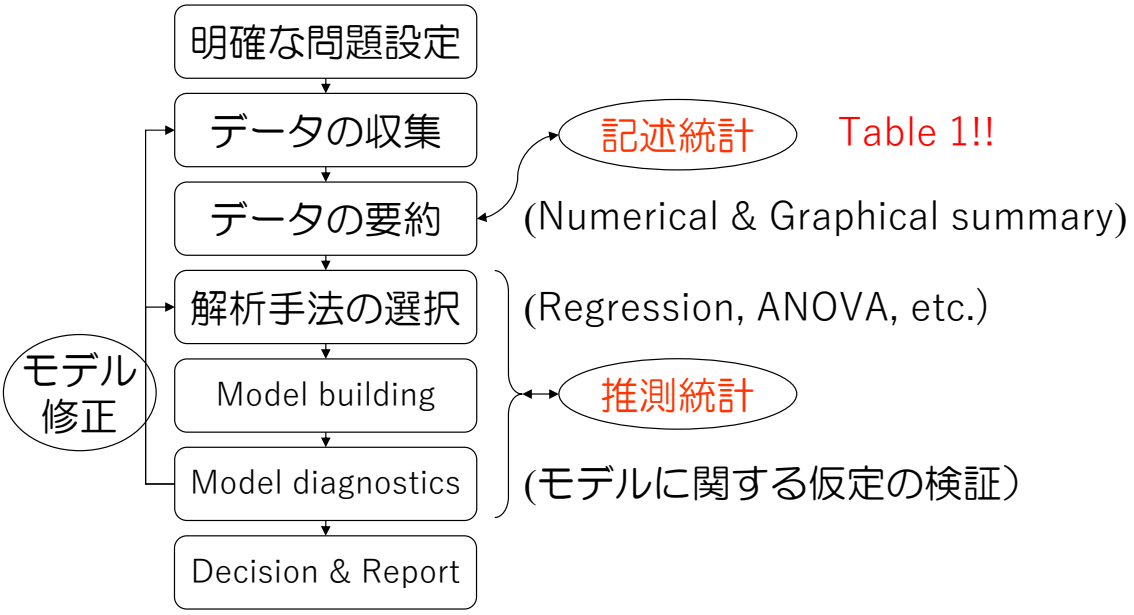
第１回　09/13　基本統計量 Table1を究めよう
第２回　09/27　推定 信頼区間 仮説検定
第３回　10/11　連続変数の比較
第４回　10/25　回帰分析
第５回　11/08　比率と分割表
第６回　11/22　ロジスティック回帰分析
第７回　12/06　生存時間解析
第８回　12/20　継時的繰り返し測定データの解析

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（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（例：平均の推定）でも，データの持つ性質により複数の解析方法が存在する場合がある．適切な解析方法を選択するために，データの特徴を把握することが重要.
- 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。

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

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

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$x_{(1)}, \ldots, x_{(n)}$　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。

$x_1, x_2, \ldots, x_n$: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 = (3rd quartile – 1st quartile)

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 ± SD, mean ± SE or median (interquartile range), as appropriate.”

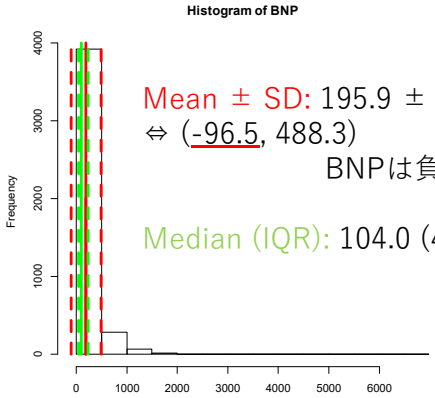
Mean ± SD (Standard deviation): 平均(Mean)を中心にMean ± SDの範囲に、データ全体の60～70%が分布している。

Mean ± SE (Standard error): Standard error (Standard Error of Mean, SEM) = 標準誤差 = 標本平均の標準偏差 = $s/\sqrt{n}$.

二群以上を比較するときは，平均を比較しているので Mean ± SE が第一選択.
一群の時は，データ全体の散らばりの範囲に興味があれば Mean ± SD も可能.

Mean ± SD (Standard deviation): 平均(Mean)を中心にMean ± 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 ± SD: 195.9 ± 292.4
⇔ (-96.5, 488.3)
BNPは負の値をとらない

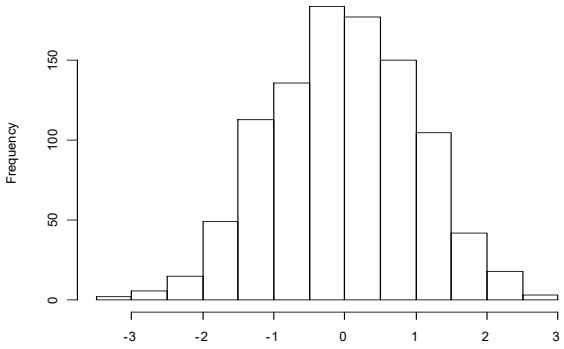
Median (IQR): 104.0 (41.3, 238.0)

Mean ± 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.

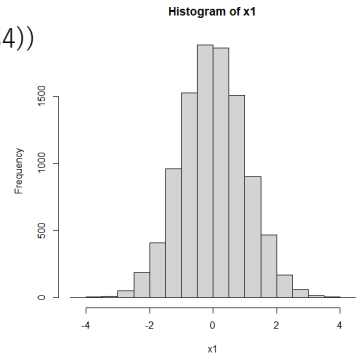
Graphical summary: Histogram

- 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.



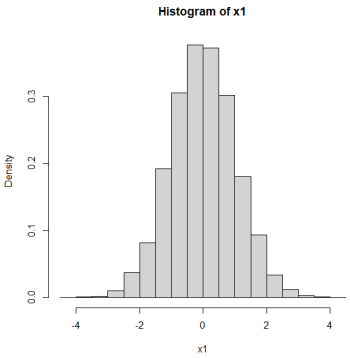
N = 10000
> hist(x1, ylim=c(0, 1884))

frequency



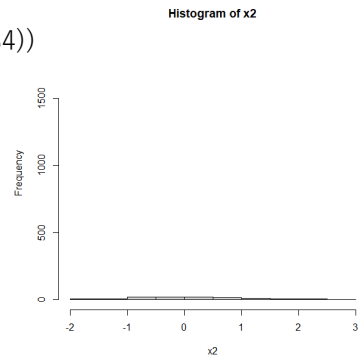
N = 10000
> hist(x1, **freq=F**)

relative frequency



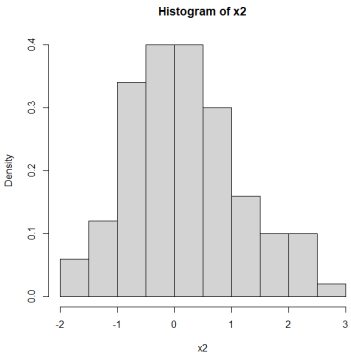
N = 100
> hist(x2, ylim=c(0, 1884))

frequency



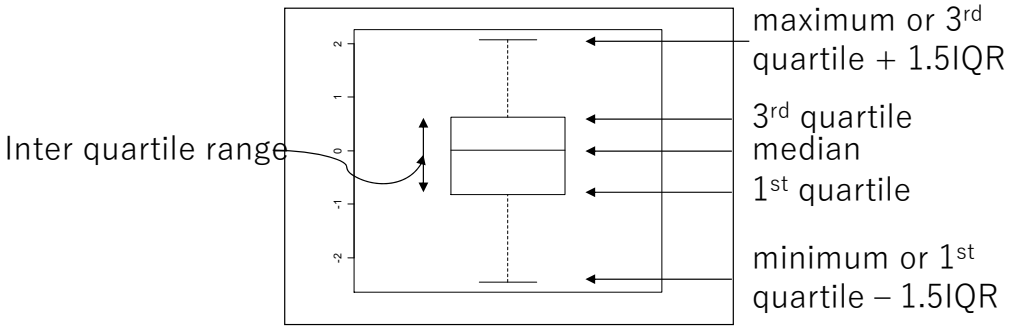
N = 100
> hist(x2, **freq=F**)

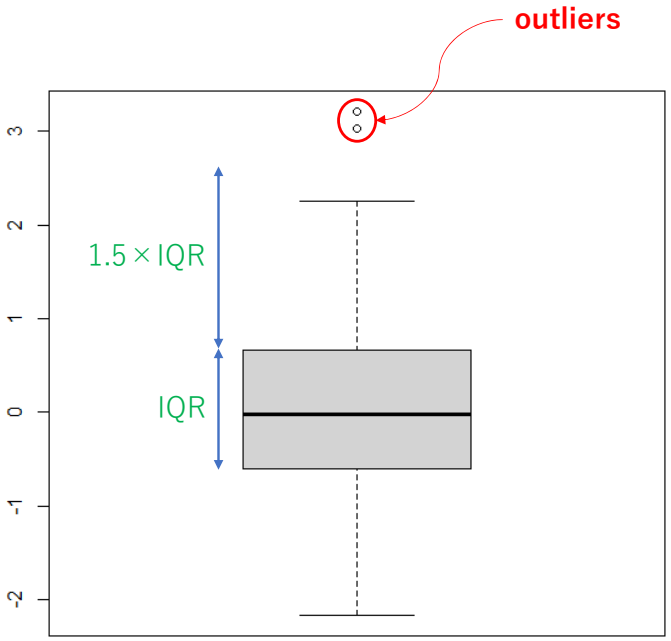
relative frequency



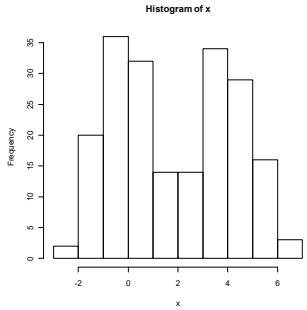
Graphical summary: Box-plot

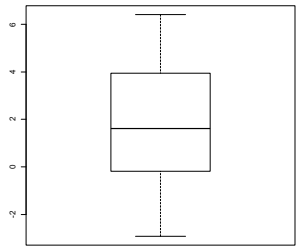
1)縦軸に変数値をとる．2)下限が1st quartile、上限が3rd quartileとなる”Box”を描く． 3)medianの位置に線を描く． 4)Boxの上下辺からmax., min.まで線を引く． 5)上下辺から1.5×IQR以上離れた標本ははずれ値 (Outlier) として，点で表す．





ヒストグラムとボックスプロット：二峰型



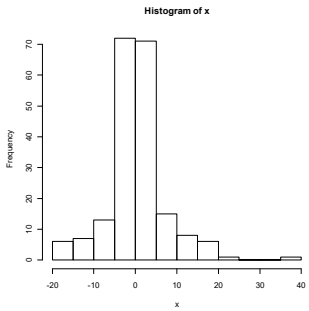


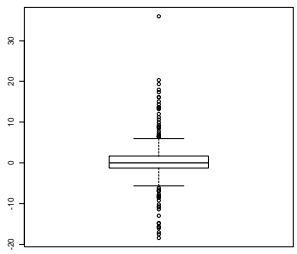
```
x1 <- rnorm(100, mean=0)
x2 <- rnorm(100, mean=4)
x <- c(x1, x2)
hist(x)
boxplot(x)
```

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．