

医学統計勉強会

東北大学病院循環器内科・東北大学臨床研究推進センター 共催

東北大学大学院医学系研究科EBM開発学寄附講座

宮田 敏

自己紹介

1988	4	一橋大学経済学部入学
1992	3	一橋大学経済学部卒業（経済学学士）
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
1998	3	一橋大学大学院経済学研究科博士後期課程退学
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

医学統計勉強会

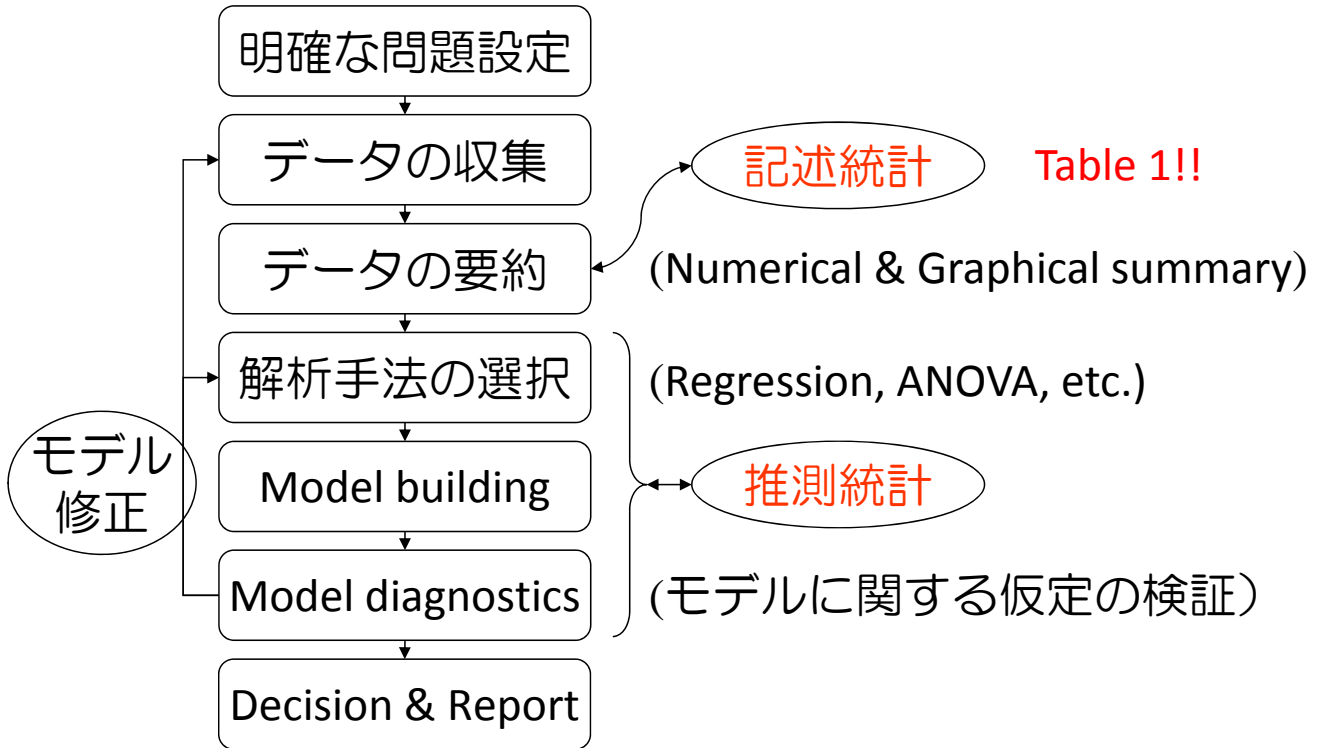
10月3日～11月27日 木曜日 19:00～20:30 臨床大講堂

第1回	基本統計量	第5回	比率と分割表
第2回	回帰分析	第6回	継時的繰り返し測定データの解析
第3回	ロジスティック回帰分析	第7回	傾向スコア
第4回	生存時間解析 生存曲線 Cox比例ハザードモデル	第8回	無作為化比較試験

**This course is not for statisticians,
not for mathematicians,
but for users of statistics!!**

- 数学的議論は最小限にとどめる.
- 医学統計で扱うデータ解析の, 基礎的概念と解析手法を扱う. (回帰分析, ロジスティック回帰分析, 生存時間解析等を含む)
- 計算機の積極的な利用が必要.

データ解析のフローチャート



2014/10/2

東北大学 医学統計勉強会

5

データの準備

元データの取り扱い

- データの形は長方形
 - 第一行目に変数名。全角文字は避ける方が無難。
 - グラフ、解析結果などを張り付けない。別ファイルで保存。
 - データの形は、長方形になるはず。

systemID	hospitalID	sex	age	height	bodyweight	
4	1185645		1	64	173	75.4
11	3329388		1	69	164	72
12	4022624		1	78	155.2	47.2
14	4402536		1	83	159.1	60
22	4862866		2	73	147.6	40.5

2014/10/2

東北大学 医学統計勉強会

6

データの準備

元データの取り扱い（続き）

ii. 元データは絶対に改変しない。

- 解析の過程で、変数を変換したり、新しい変数を定義することがある。
- 新しく作ったデータを、元データに上書きしない。
- データを改変したら、新しいファイル名で保存。
- 元データを改変すると、元データが何であるか分からなくなる。元データが分からなくなれば、意図せざるデータのねつ造まであと一歩。

データの準備

元データの取り扱い（続き）

iii. 患者さんの個人情報には記載しない。

- 残念ながら、いまだに氏名、カルテ番号など、患者さん個人を特定できる情報が付いたままのデータを見かける。
- 個人情報は、データ解析の立場からは無意味。
- 個人情報が漏れいすれば、研究は中止、研究者の辞表が何枚か必要。被害者には、お詫びの仕様がなない。
- データを受け取ったら、個人情報はすぐに匿名化もしくは削除。

データの準備

元データの取り扱い（続き）

iv. 解析記録の保存。

- 患者さんを診察すれば、医師がカルテに記録するのは**当然**。実験をすれば、実験ノートに記録するのは**常識**。統計解析の記録を残すのも、それと同じ。
- 元データと解析の記録を見れば、第三者が解析を再現できる程度の記録が必要。
 - **解析の再現性**
 - **備忘録** 「三日後の自分は遠い親戚。一週間後の自分は赤の他人」
 - 出来れば、**プログラム**を書いて解析する。

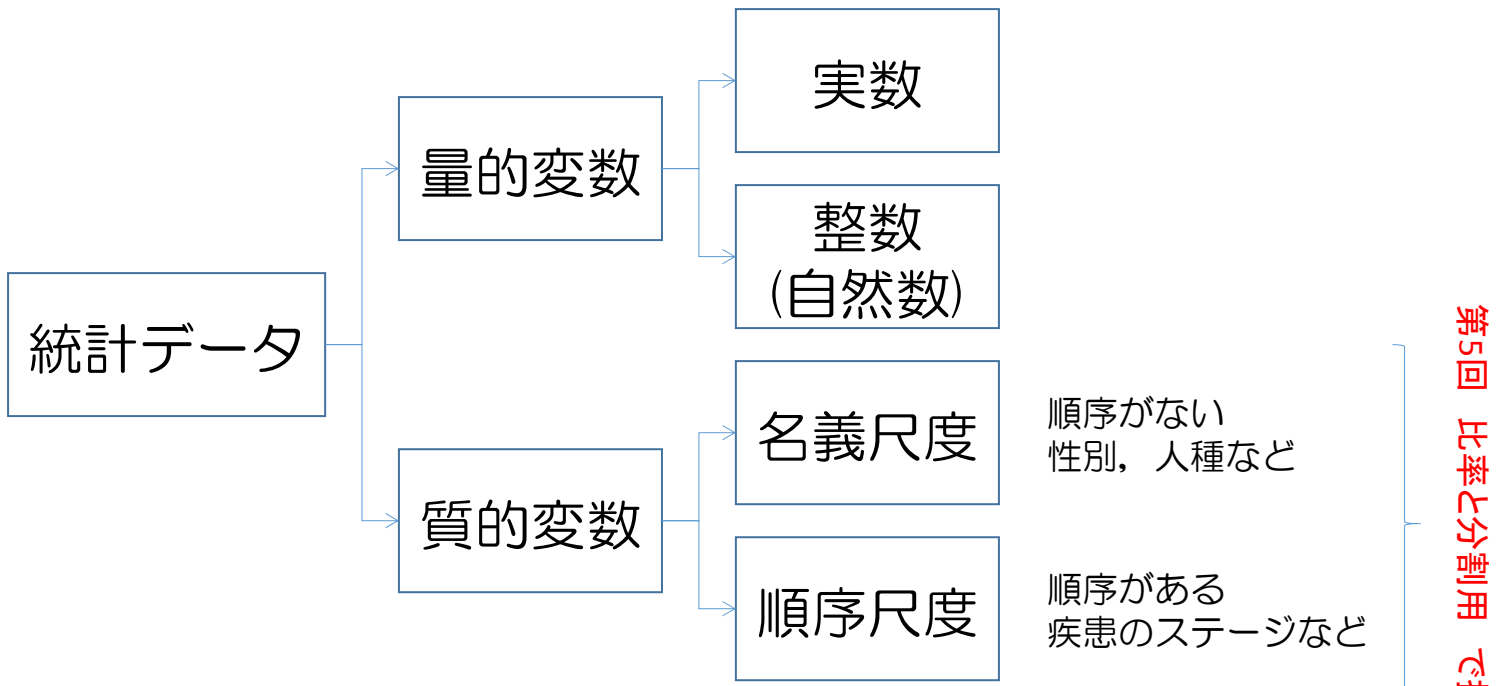
データの準備

データ入手時にすべきこと：**入力ミス**、**異常値**の発見

表計算ソフトの**フィルター機能**が便利

- **データの範囲**： 本来正の値をとるはずが、負の値をとる。小数点の間違いで、体重35kgが3.5kgになる、等。
- **全角文字と半角文字**の混在：“w”と“w”など。
- **質的変数の数字表記**： 男性→1, 女性→2など。男性→M, 女性→Fのように書き直す。
- **異常な値**の検出：“3.14”と“3,14”など。
- **欠測値の数**： 欠測値の数が想定より多い場合、データが正常に認識されていないことがある。

統計データと尺度



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, \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.)

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):

$$f_s = (3^{\text{rd}} \text{ quartile} - 1^{\text{st}} \text{ quartile})$$

“Continuous variables were expressed as **mean \pm SD**, **mean \pm SE** 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。

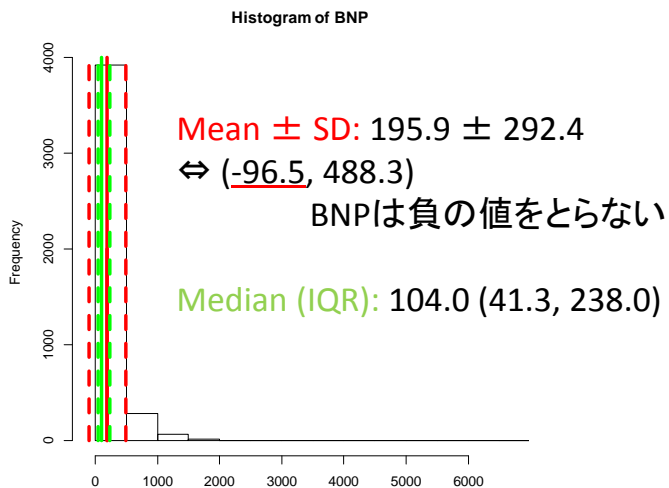
Mean \pm SE (Standard error): Standard error (Standard Error of Mean, SEM) = **標準誤差** = **標本平均の標準偏差** = s / \sqrt{n} .

二群以上を比較するときは、平均を比較しているので **Mean \pm 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 \pm SD** も可能。

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。

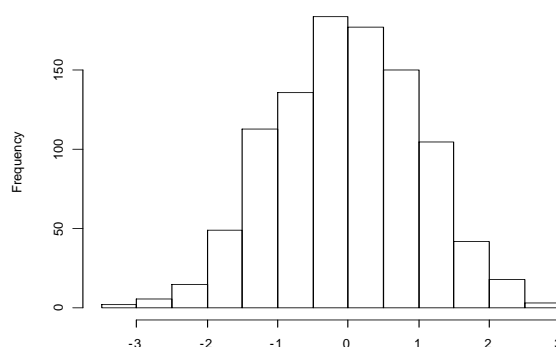


Mean \pm SDは、不合理な値(データの範囲を逸脱)をとることがある。

分布が歪んでいるときは、Median (IQR) が第一選択。

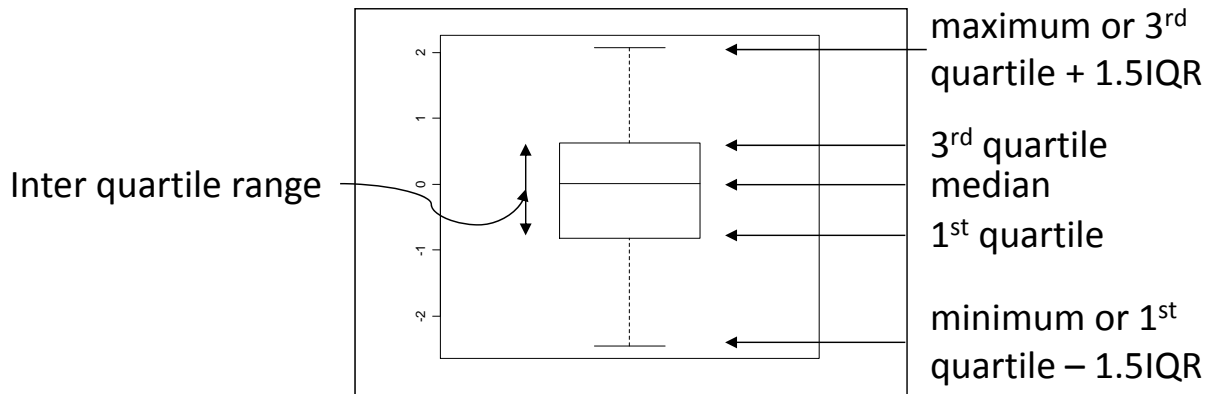
Graphical summary: Histogram

- 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): 頻度もしくはは相対頻度を表した棒グラフ.



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.

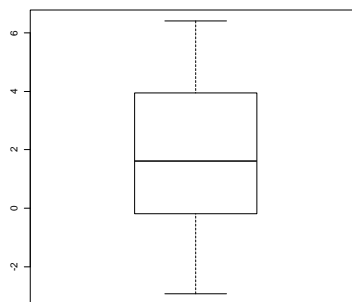
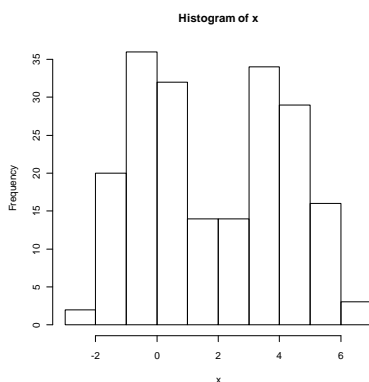


2014/10/2

東北大学 医学統計勉強会

19

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。

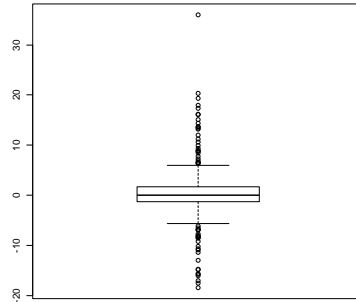
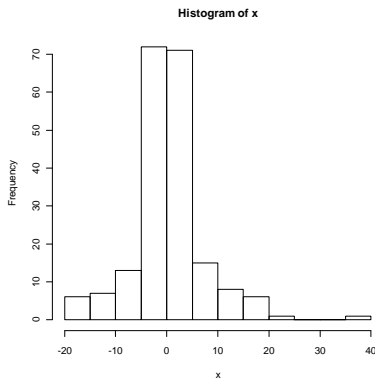
ヒストグラムは分布の特徴の、**全体的な傾向**をとらえるのに適している。

2014/10/2

東北大学 医学統計勉強会

20

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。

2014/10/2

東北大学 医学統計勉強会

21

平均・中央値の差の検定

二標本問題：二つのグループの平均値の差の検定

帰無仮説 $H_0: \mu_1 = \mu_2$

対立仮説 $H_1: \mu_1 \neq \mu_2$

Welch's t-test（ウェルチのt検定）：二群のデータがそれぞれ正規分布に従う。不等分散を仮定する。

⇔ Mean \pm SE, Mean \pm SD に対応。

Mann-Whitney test, Wilcoxon's rank sum test：二群のデータは任意の同じ形の分布に従う。当然二群の分散は等分散になる。⇔ Median (IQR) に対応。

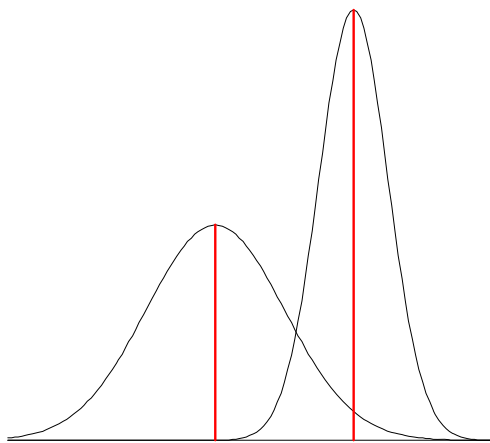
検定をしたら、必ずp値を明記する。

2014/10/2

東北大学 医学統計勉強会

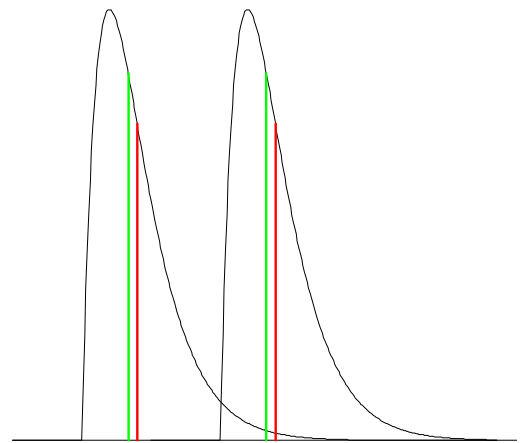
22

T検定に適した比較



- 二群とも正規分布に従う
- 分散は異なってもよい

MW検定に適した比較



- 正規分布に従わなくてもよい
- 分布の形状は同じ。分散も同じ。

データの分布が正規分布に従わず、分散も等しくない場合、取りあえず元データを**対数変換**するなどして、**等分散**に近づける。それでもだめなら、専門家にご相談ください。

三群以上の比較：

帰無仮説 $H_0: \mu_1 = \dots = \mu_k, k: \text{グループの数}$

対立仮説 $H_1: \text{少なくとも一つの母平均が他から異なる}$

分散分析 (Analysis of Variance, ANOVA)：各群のデータがそれぞれ**正規分布**に従う。**等分散**を仮定する。

Kruskal-Wallis test：各群のデータは任意の**同じ形の分布**に従う。当然各群の分散は**等分散**になる。

データの分布が正規分布に従わず、分散も等しくない場合、やはり**対数変換**などで、**等分散**に近づける。

Box-Cox変換：分散の安定化と正規性の向上を同時に達成する変換。詳細は、ご相談ください。

多重比較 (Multiple Comparison)

分散分析の帰無仮説 $H_0 : \mu_1 = \dots = \mu_k$ が棄却されたとき、どの μ_i が他から有意に異なるかが知りたい。

Tukey's HSD (Honestly Significant Difference) : すべての対比 ($\mu_i - \mu_j$) についての検定を同時に行う。可能な対比の組み合わせは、 $k(k-1)/2$ 通り。

Dunnettの方法 : グループの一つがコントロール群である時、コントロール群と他の $(k-1)$ の対照群との比較を同時に行う。

Williamsの方法 : 対立仮説 $H_1 : \mu_1 \leq \dots \leq \mu_k$ (あるいはその逆) を検定する。

2014/10/2

東北大学 医学統計勉強会

25

多重比較 (Multiple Comparison)

前項の方法は、すべて正規性の仮定を必要とするパラメトリックな方法。正規性を必要としない、ノンパラメトリックな方法も存在する。

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

	$H_1 : \mu_1 \neq \mu_2, \dots, \mu_i \neq \mu_j, \dots, \mu_1 \neq \mu_k$	$H_1 : \mu_1 \neq \mu_2, \mu_1 \neq \mu_3, \dots, \mu_1 \neq \mu_k$	$H_1 : \mu_1 \leq \dots \leq \mu_k$
パラメトリック検定	Tukeyの方法	Dunnettの方法	Williamsの方法
ノンパラメトリック検定	Steel-Swassの方法	Steelの方法	Shirley-Williamsの方法

2014/10/2

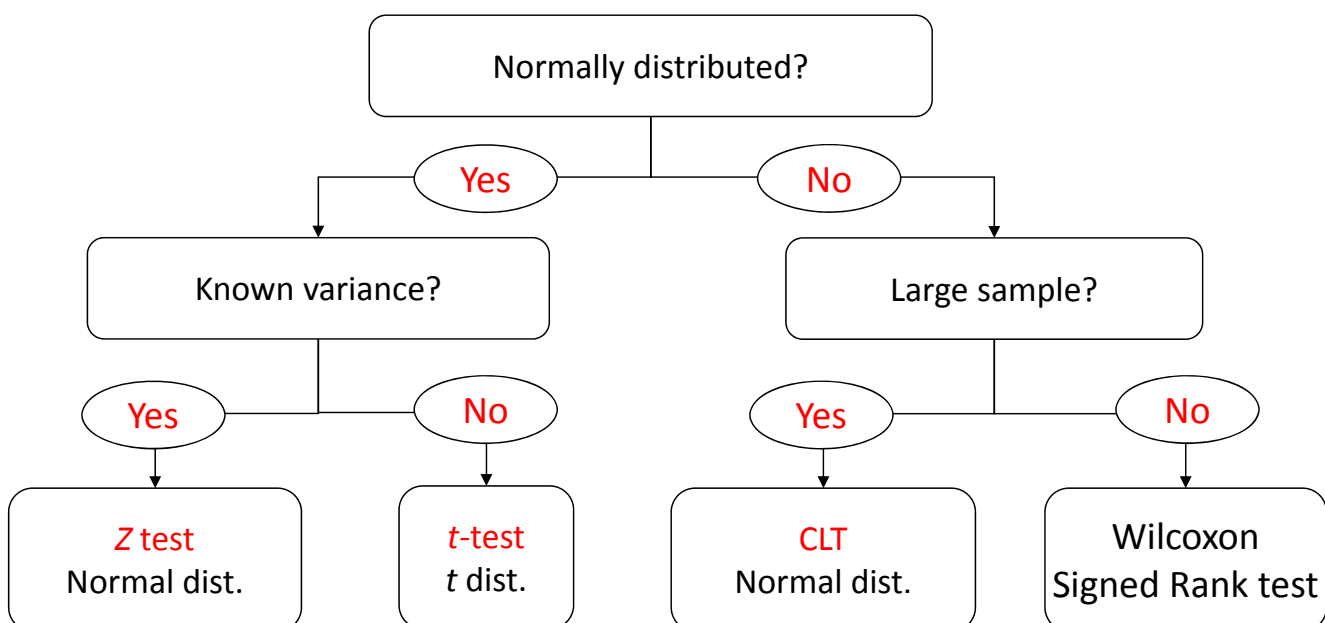
東北大学 医学統計勉強会

26

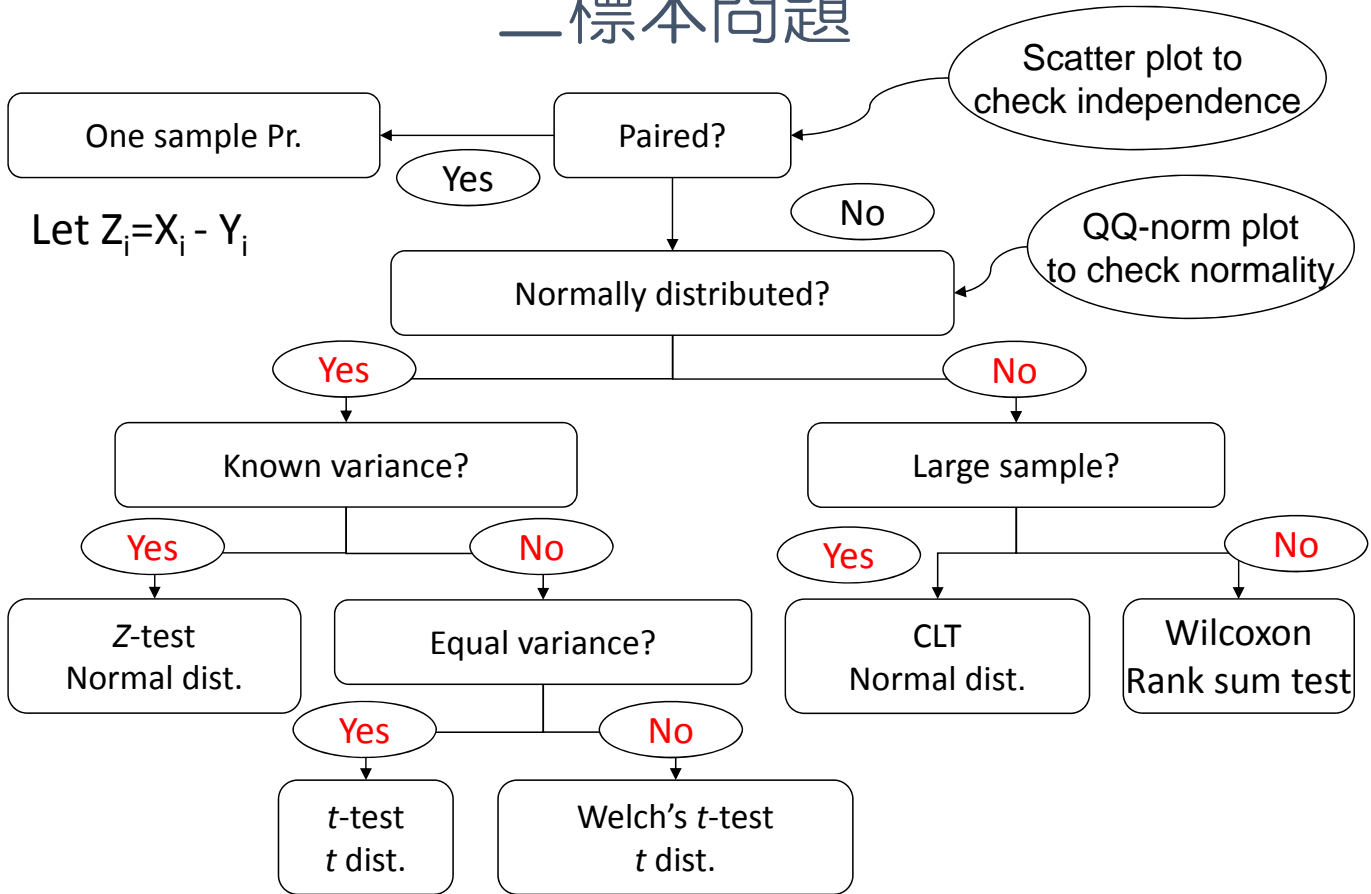
Take Home Message

1. 統計学とは
2. データの準備
2. 記述統計
 - 数値的要約：平均、中央値、分散、標準偏差、IQR
 - 視覚的要約：ヒストグラム、ボックスプロット
3. カテゴリデータの要約と比較
4. 平均・中央値の比較
 - 二標本問題：Welch's t-test, Mann-Whitney test
 - 三群以上の比較：分散分析、Kruskal-Wallis test
 - 多重比較

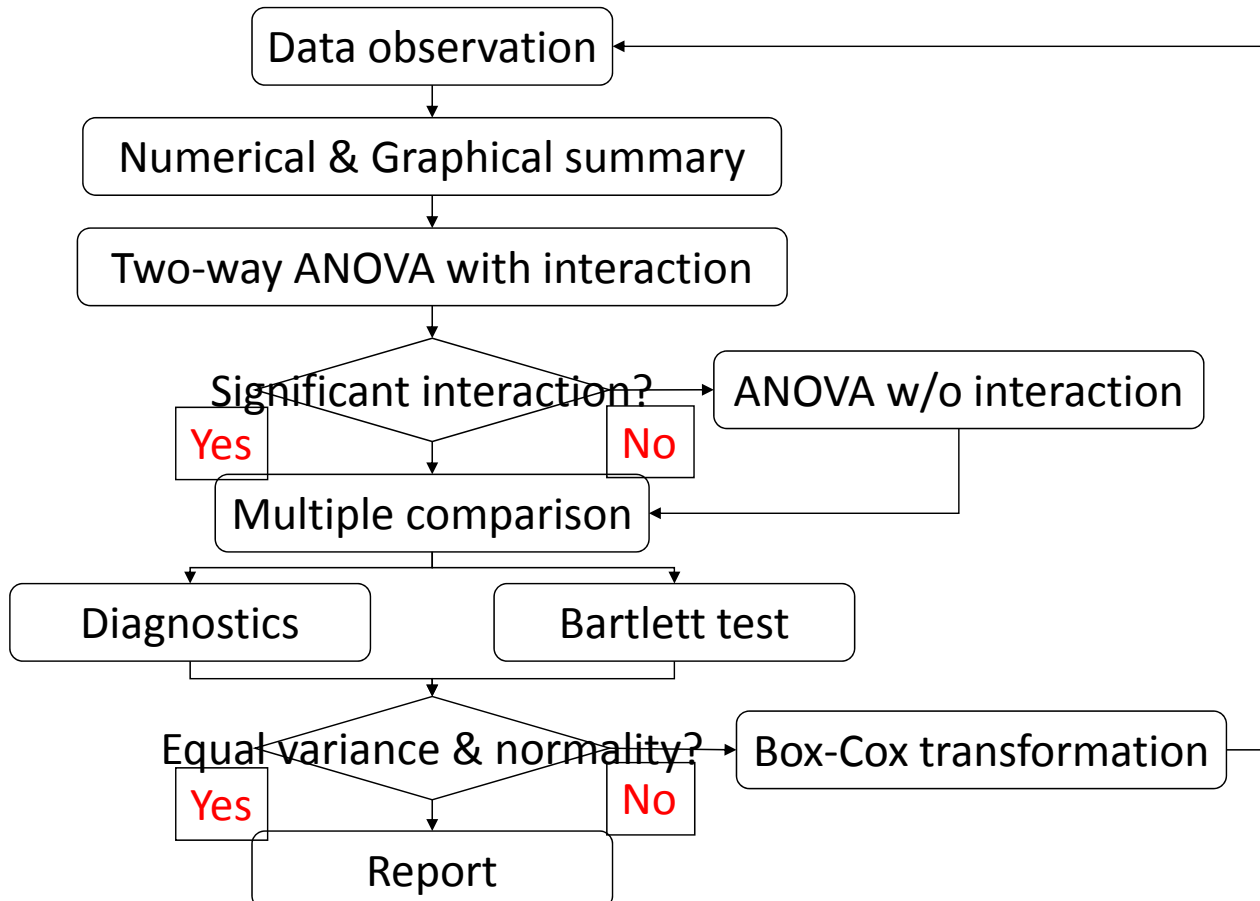
一標本問題



二標本問題



分散分析



参考文献：

- 丹後 俊郎 (著)「新版 医学への統計学」朝倉書店; 新版 (1993/09)ISBN-10: 4254125461
- 丹後 俊郎 (著)「統計学のセンスーデザインする視点・データを見る目」朝倉書店(1998/10) ISBN-10: 4254127510
- 東京大学教養学部統計学教室 (編集)「統計学入門 (基礎統計学)」東京大学出版会 (1991/7/9) ISBN-10: 4130420658