

# 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

9月26日～11月14日 木曜日 19:00～20:00(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, 生存時間解析等を含む)
- 計算機の積極的な利用が必要.

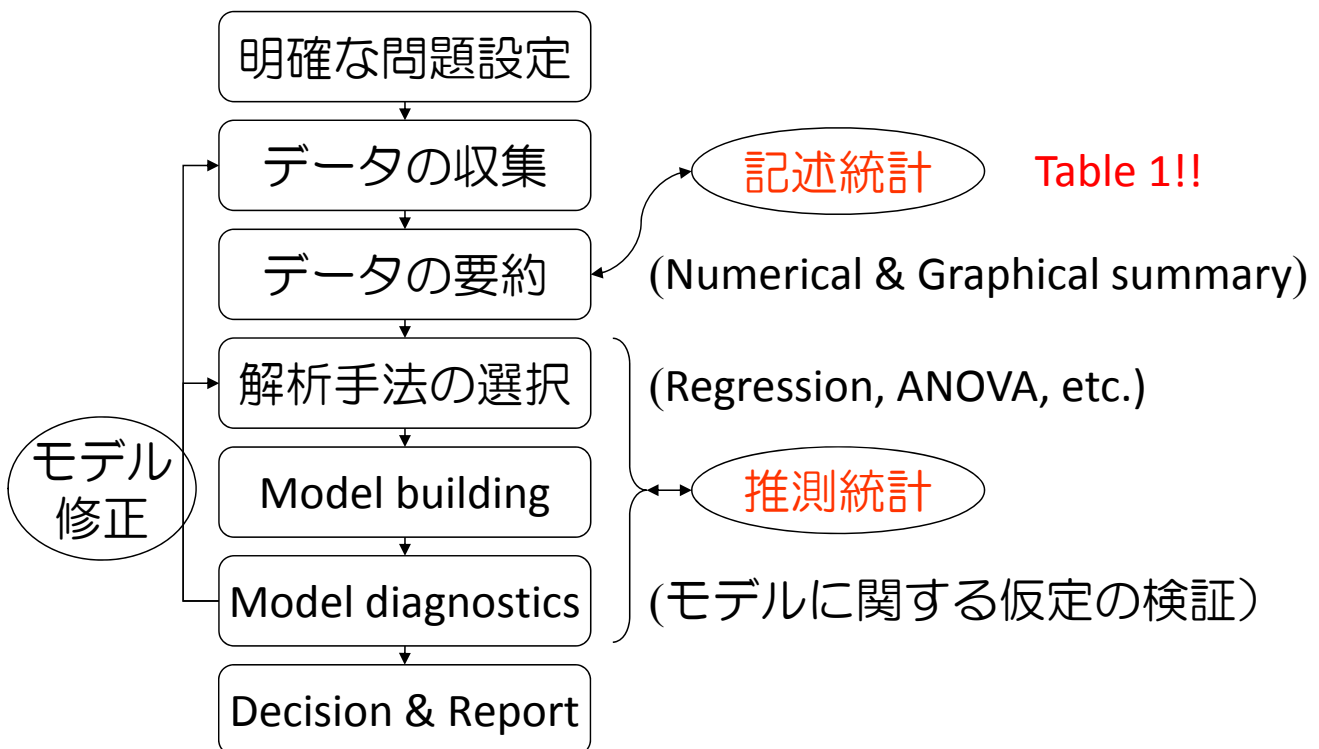
# Goal of Statistics

- 「統計学」とは、不確実性や多様性を伴った事象に対して、合理的な推論を行うための学問。
- 不確実な事象について、すべての情報を得ることは不可能。部分的な情報を集め、それを基に全体を推論することが必要。

Dr. Samuel Johnson said,

*“You don't have to eat the whole ox to know that the meat is tough.”*


## 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, 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**: データの**代表値**による要約.
- **Graphical summary**: 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 (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., 1<sup>st</sup> quartile, median, 3<sup>rd</sup> 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。

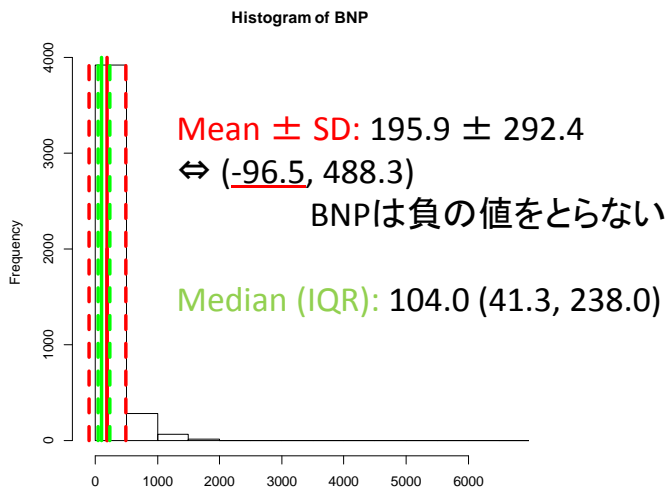
**Mean  $\pm$  SE** (Standard error): Standard error (Standard Error of Mean, SEM) = **標準誤差** = **標本平均の標準偏差** =  $s/\sqrt{n}$ .

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。

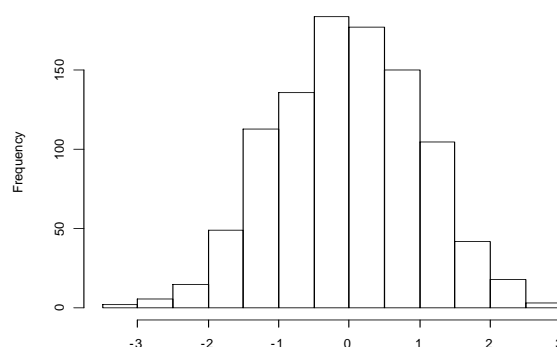


Mean  $\pm$  SDは、不合理な値(データの範囲を逸脱)をとることがある。

分布が歪んでいるときは、Median (IQR) が第一選択。

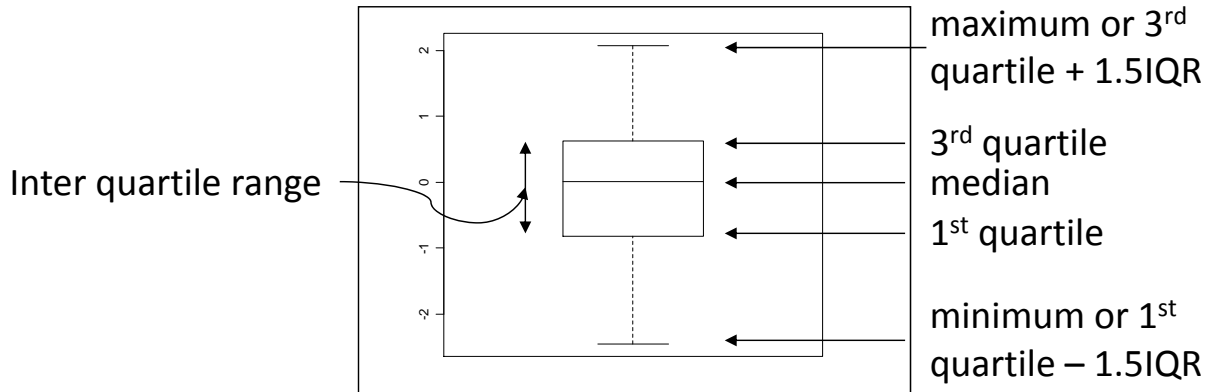
## Graphical summary: Histogram

- 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1<sup>st</sup> quartile、上限が3<sup>rd</sup> 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.

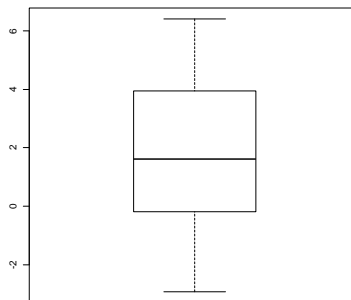
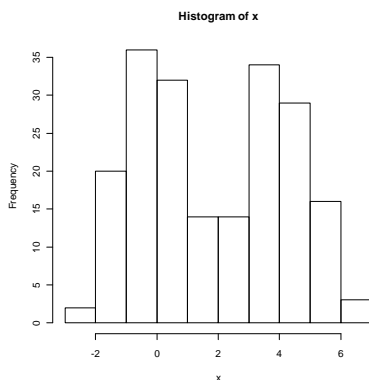


2013/926

東北大学 医学統計勉強会

15

## 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。

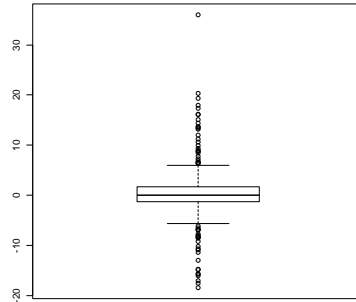
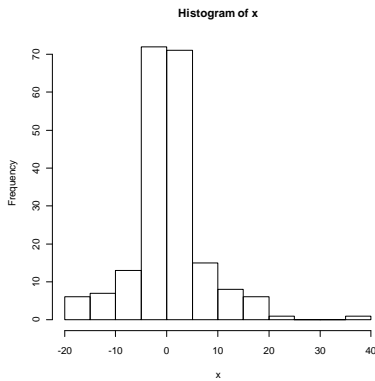
2013/926

東北大学 医学統計勉強会

16



# 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。

2013/926

東北大学 医学統計勉強会

17

## カテゴリデータの要約

度数分布：水準 カテゴリの種類  
度数 各水準のデータ数  
割合（相対度数） 度数/サンプル数 (%)

分割表 (Contingency table)：2種類のカテゴリデータの水準の組み合わせごとに、度数を求めた表。

(例)

心筋梗塞発作後6カ月の転帰

	6カ月以内死亡	6カ月以上生存
牛乳抗体陽性	29	80
牛乳抗体陰性	10	94

Davis, et al. (1974) Lancet, vol. 303, no. 7865, 1012-1014

2013/926

東北大学 医学統計勉強会

18

# 分割表の検定（独立性の検定）

	6カ月以内死亡	6カ月以上生存
牛乳抗体陽性	29	80
牛乳抗体陰性	10	94

$$\hat{p}_1 = 29/(29 + 80) = 0.266$$

$$\hat{p}_2 = 10/(10 + 94) = 0.0962$$

帰無仮説  $H_0: p_1 = p_2$  母比率が一定.

対立仮説  $H_1: p_1 \neq p_2$  母比率が異なる.

もし、二つの変数に**関連がなければ**、グループによらず（=陽性でも陰性でも）、母比率（=6カ月以内に死亡する確率）は**一定のはず**  $\Rightarrow$  **帰無仮説  $H_0$**   
関連があれば、母比率が**異なる**  $\Rightarrow$  **対立仮説  $H_1$**

2×2分割表に限り、**片側仮説**が可能。  $H_1: p_1 > (<) p_2$

## 分割表の検定（独立性の検定） 続き

**$\chi^2$ 検定 (chi-squared test)**：サンプル数が多いとき、検定統計量の分布が $\chi^2$ 分布で近似されることを利用した検定。分割表の度数は、最低**5**は必要とされる。必ず**Yatesの連続補正 (Yates's continuity correction)**を行う。

**Fisherの直接法 (Fisher's exact test)**：サンプル数によらず、正確な**p**値を計算できる検定。

**Fisherの直接法が第一選択**。分割表が大きすぎてFisher検定が実行できない場合に限り、Yatesの連続補正をした $\chi^2$ 検定を選択する。

# 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

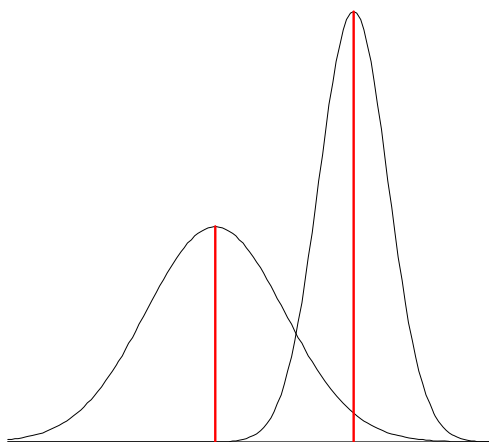
帰無仮説  $H_0: \mu_1 = \mu_2$

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。

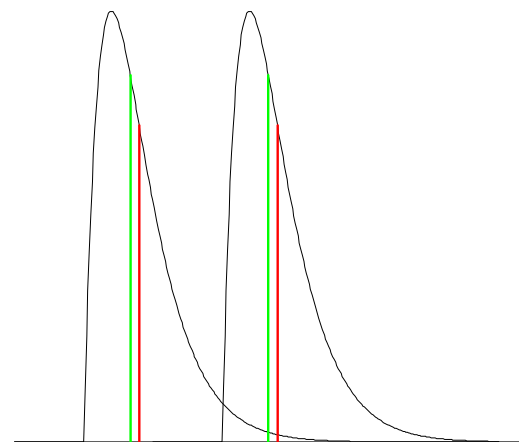
**Mann-Whitney test, Wilcoxon's rank sum test**：二群のデータは任意の**同じ形の分布**に従う。当然二群の分散は**等分散**になる。

## 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  $H_0: \mu_1 = \dots = \mu_k$ ,  $k$ : グループの数

対立仮説  $H_1$ : 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  $\mu_i$  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。

# 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。

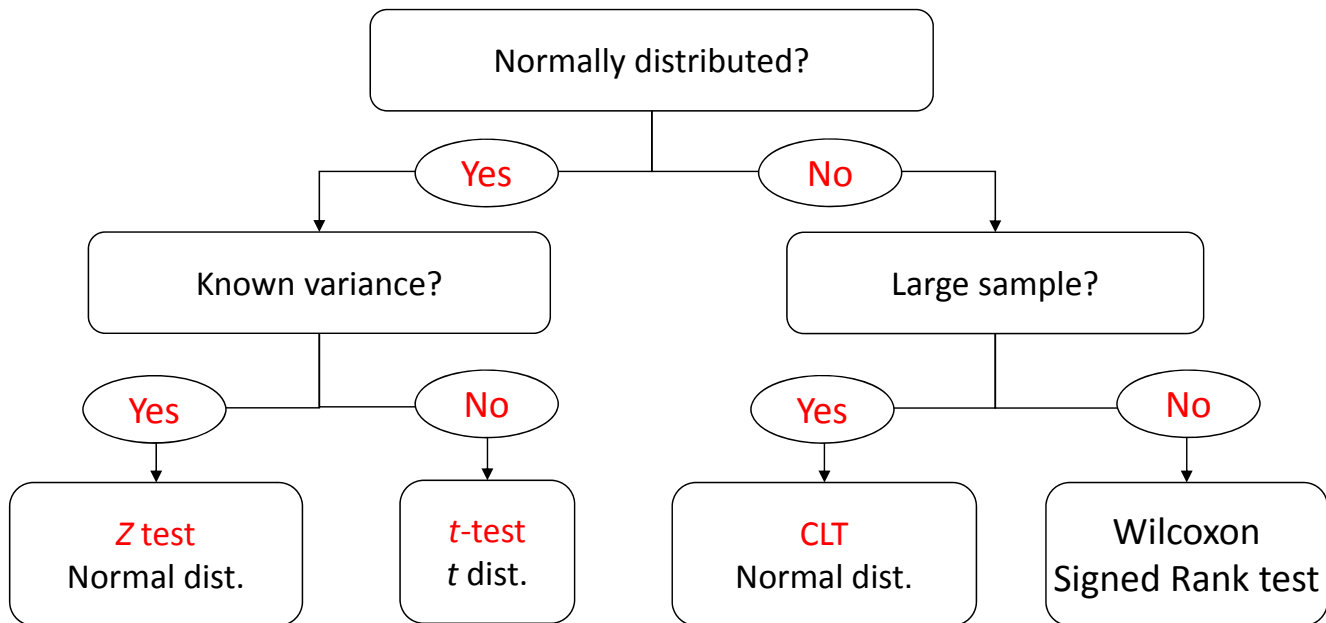
$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

	$H_1 : \mu_1 \neq \mu_2, \dots, \mu_i \neq \mu_j, \dots, \mu_1 \neq \mu_k$	$H_1 : \mu_1 \neq \mu_2, \mu_1 \neq \mu_3, \dots, \mu_1 \neq \mu_k$	$H_1 : \mu_1 \leq \dots \leq \mu_k$
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

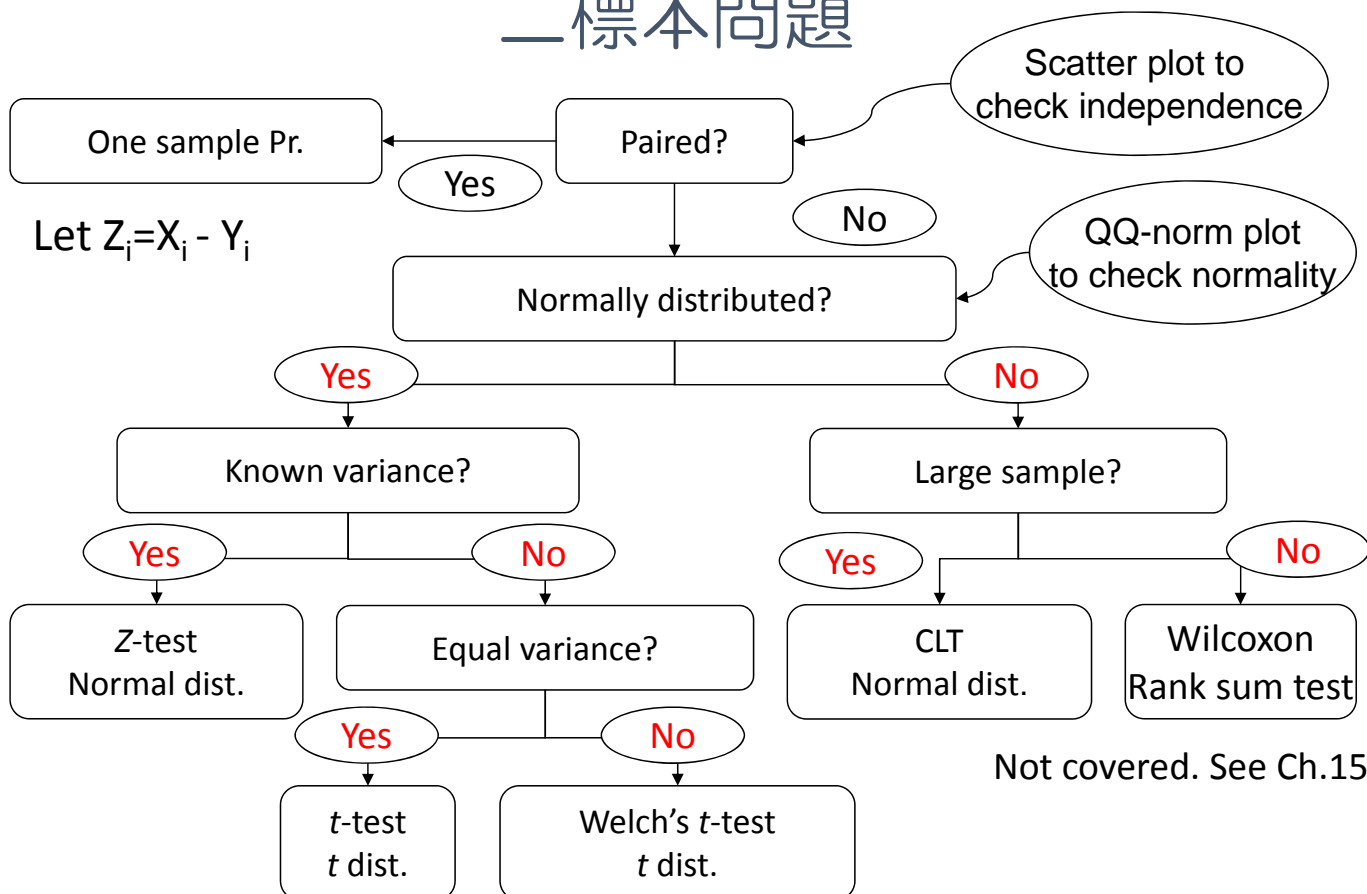
## Take Home Message

1. 統計学とは
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

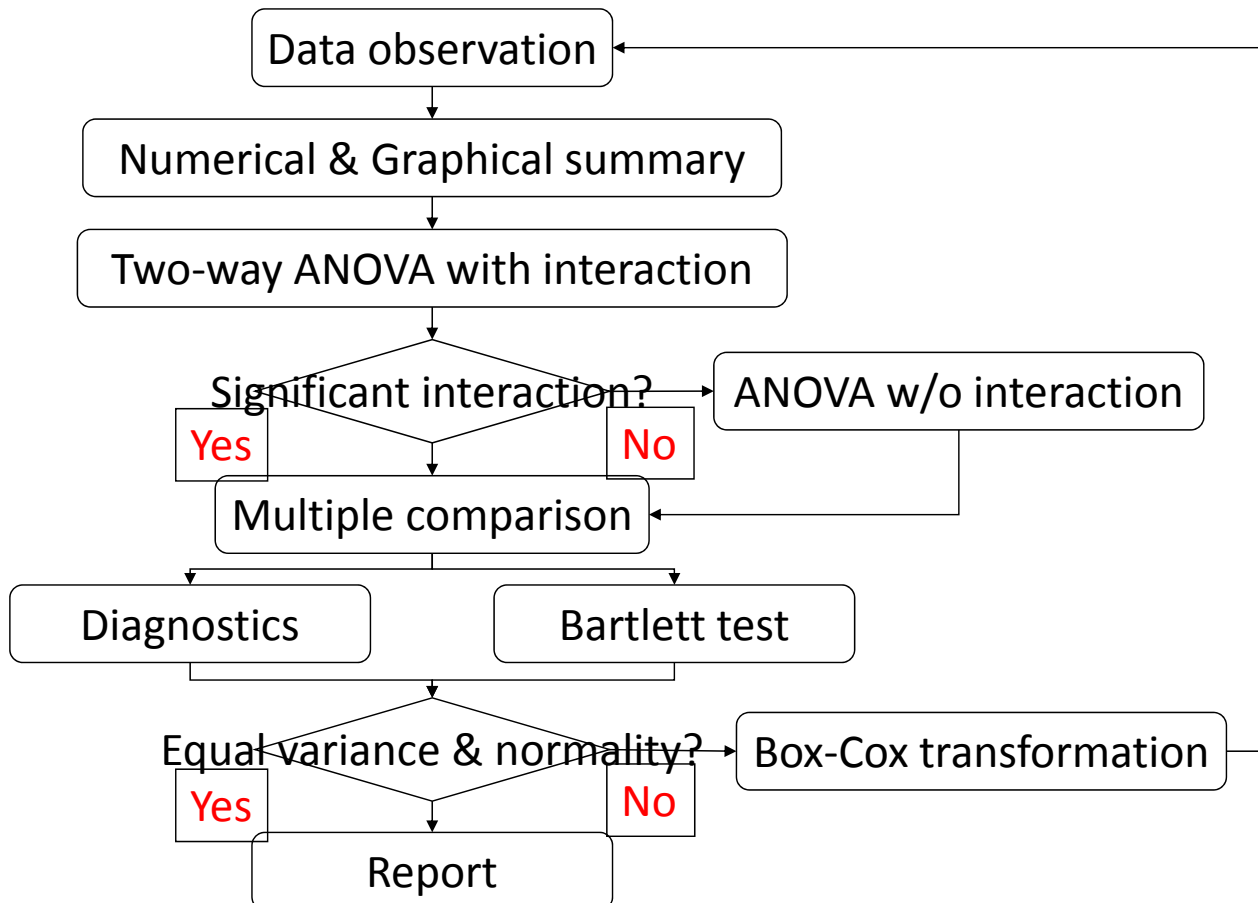
# 一標本問題



# 二標本問題



# 分散分析



2013/926

東北大学 医学統計勉強会

29

## 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

2013/926

東北大学 医学統計勉強会

30