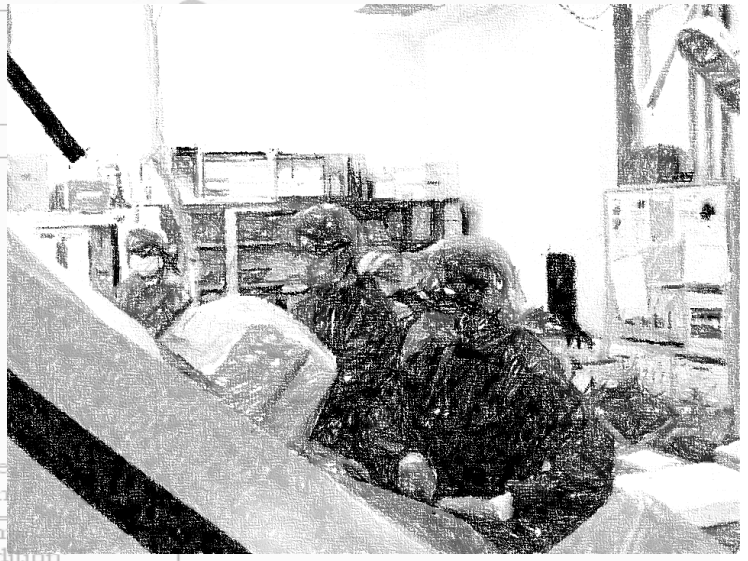


生物統計の考え方とRの紹介

沖縄県立南部医療センター・こども医療センター 諸見里　拓宏

北部地区医師会病院　中力 美和

これからの「質の高い医療体制」＝「自己進化できるシステムの成熟度が高い」

最先端医療

・医療倫理
・安全
・弱者保護
・予防医療/政策
・科学進歩への貢献
・費用対効果

医療者

分析者

論文・データを介した
世界の医療者とのコミュニケーション

データを [正しく分析] 且つ [正しく解釈]する

生物統計学　＋　疫学

分析技術

Methods for analysis (Bayesian)

研究のデザイン
分析結果の解釈

Research Design
Interpretation of results

Definition by John Orav, HSPH

リサーチクエッション

- 仮説
- 患者層(population)の設定

患者から情報を集める

- 自分の患者層から情報を収集
- サンプルサイズの決定

統計学的分析

- 統計学的検定 (信頼区間の設定)
- サンプルから全体の集団の予想

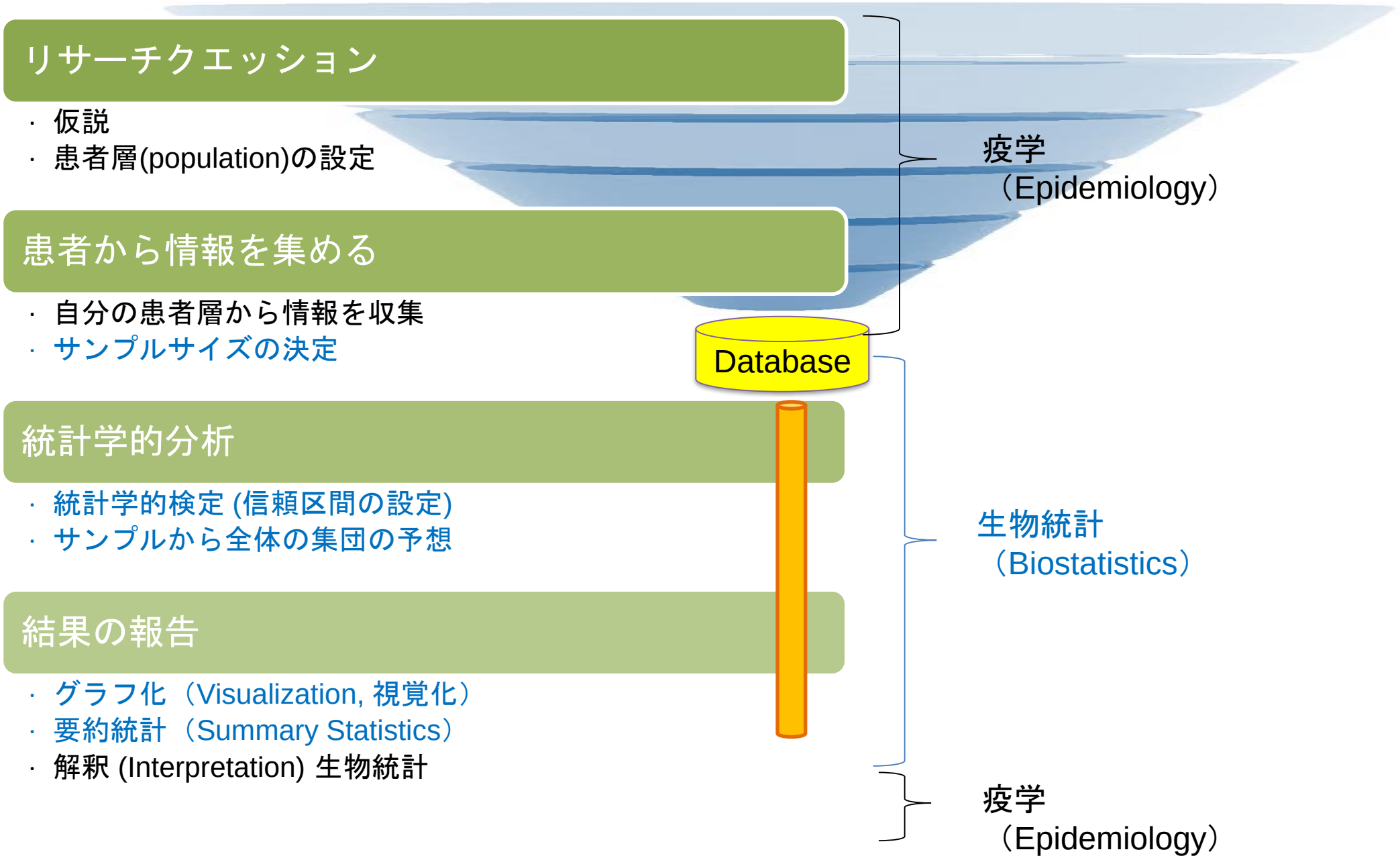
結果の報告

- グラフ化（Visualization, 視覚化）
- 要約統計（Summary Statistics）
- 解釈 (Interpretation) 生物統計

疫学
（Epidemiology）

生物統計
（Biostatistics）

疫学
（Epidemiology）


リサーチクエッション
・仮説
・患者層(population)の設定
患者から情報を集める
・自分の患者層から情報を収集
・サンプルサイズの決定
統計学的分析
・統計学的検定 (信頼区間の設定)
・サンプルから全体の集団の予想
結果の報告
・グラフ化（Visualization, 視覚化）
・要約統計（Summary Statistics）
・解釈 (Interpretation) 生物統計
Database
疫学
（Epidemiology）
生物統計
（Biostatistics）
疫学
（Epidemiology）

解釈 → 教育 → 実際の患者 → 情報集積 → 経験知 → 解釈

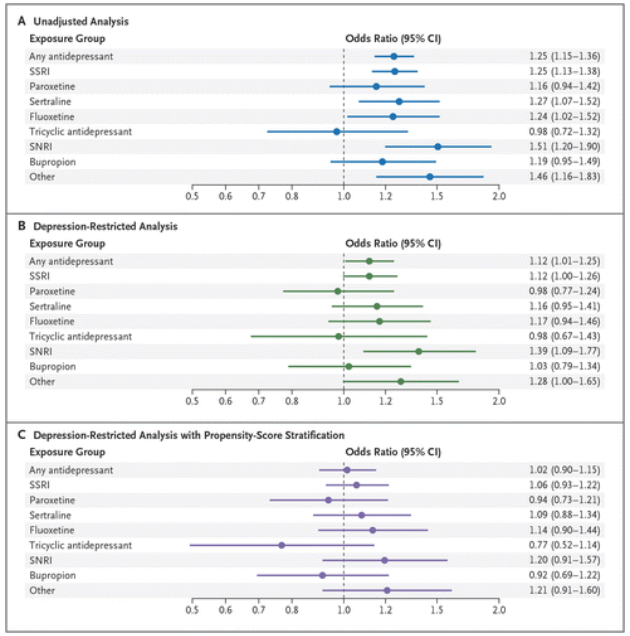
昔

生命予後

QOL

社会背景
（社会的リソース）

病態生理



平均
全体像
論文

仮想患者
への
最適解

・基礎医学
・最先端技術
・再生医療

実際の患者

現在

情報集積

データ

・国家レベルデータベース
・ビッグデータ
・NDB, DPC, JMDC

分析

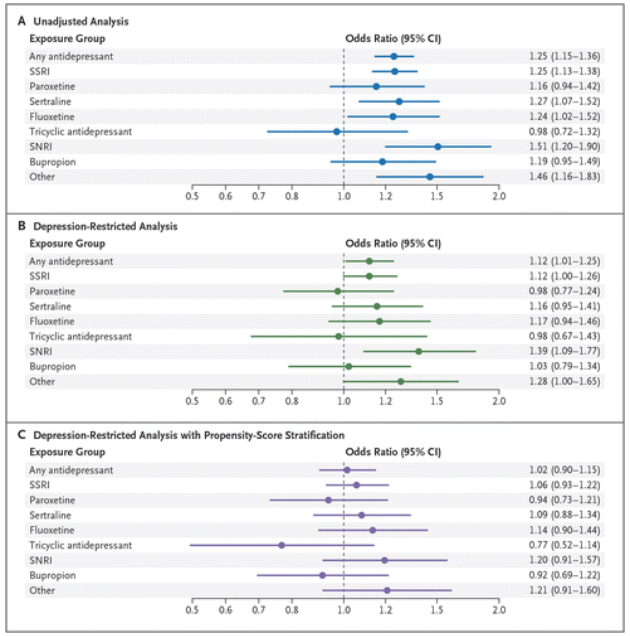
アウトカムs

QALY

保険医療解析

DPC

病態生理



CMA, treeage, BUGS

Writing skill
Visualization

平均
全体像
論文

仮想患者
への
最適解

R, Baysian model

R
STATA
SAS
Julia
etc

析

Python, R

実際の患者

電子カルテ

QALY

保険医療解析

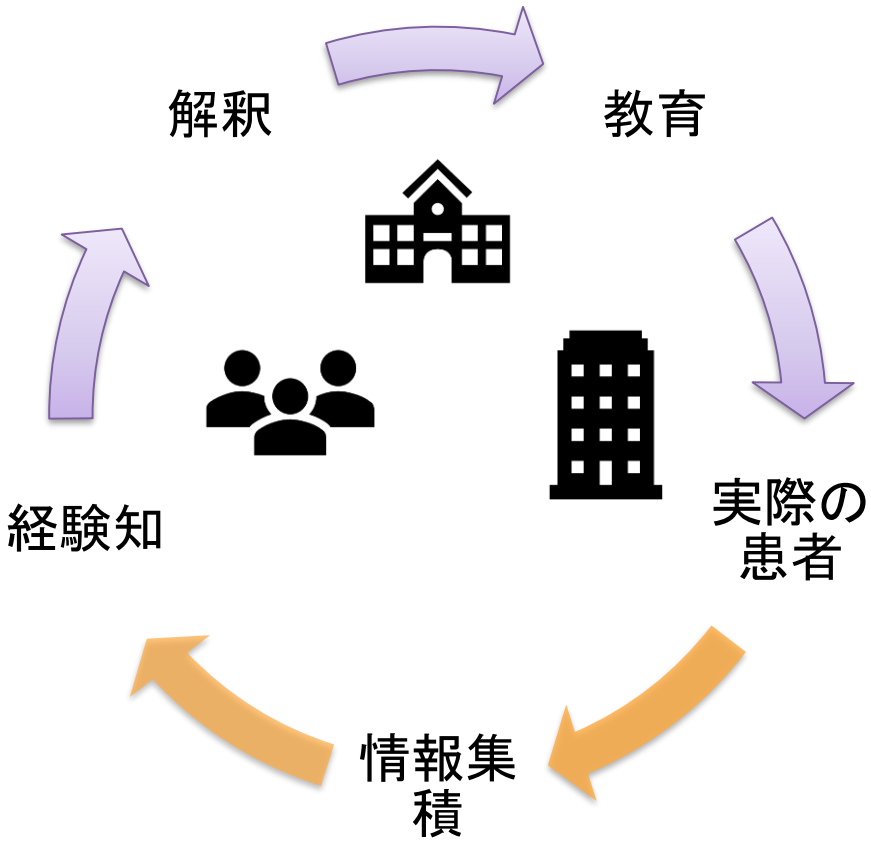
DPC

病態生理

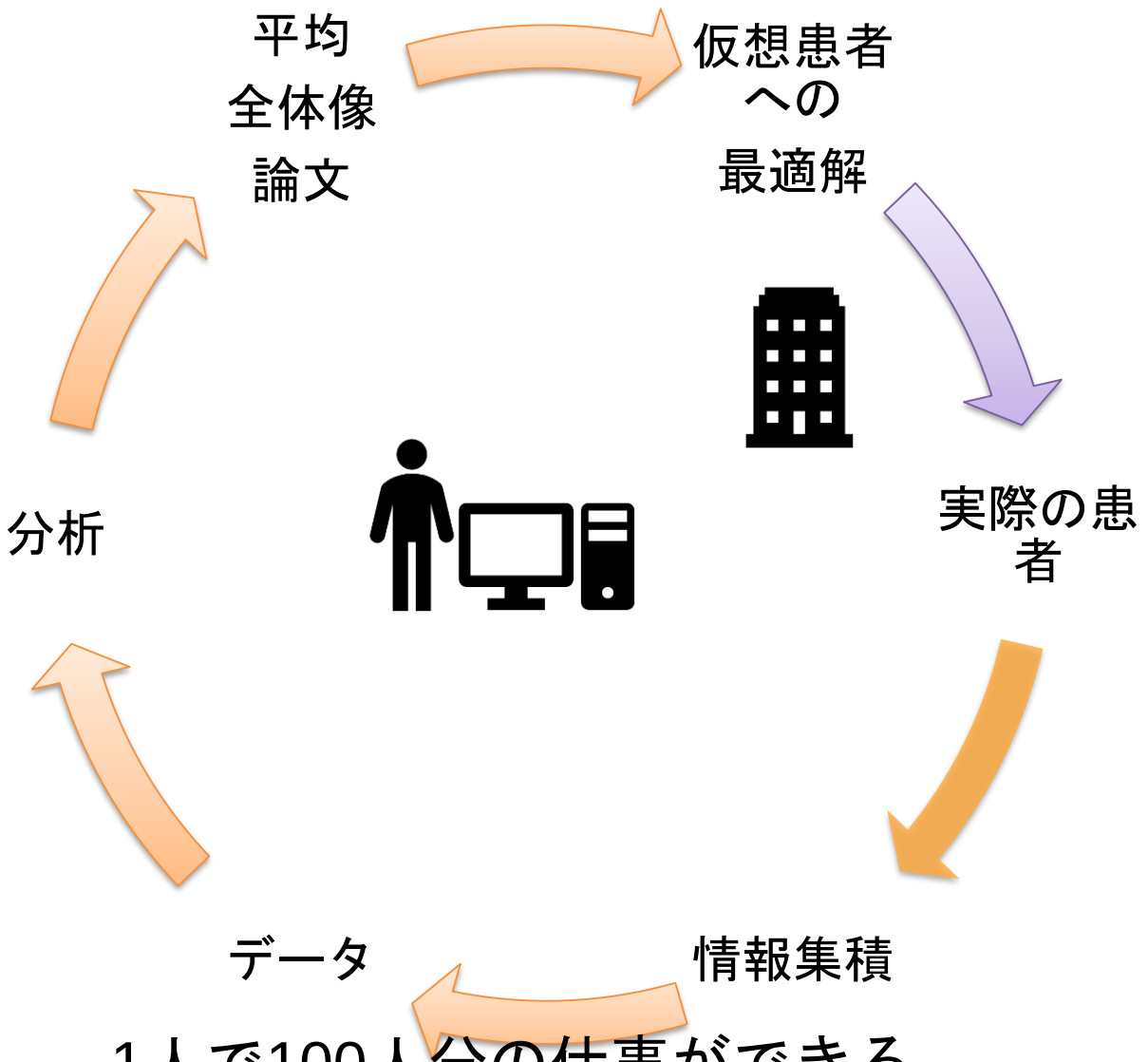
情報集積

NDB, DP
JMDC
介護データ
特定健診データ
＊SQL

データ



1人で3～5人分の仕事ができる
＝優秀な人



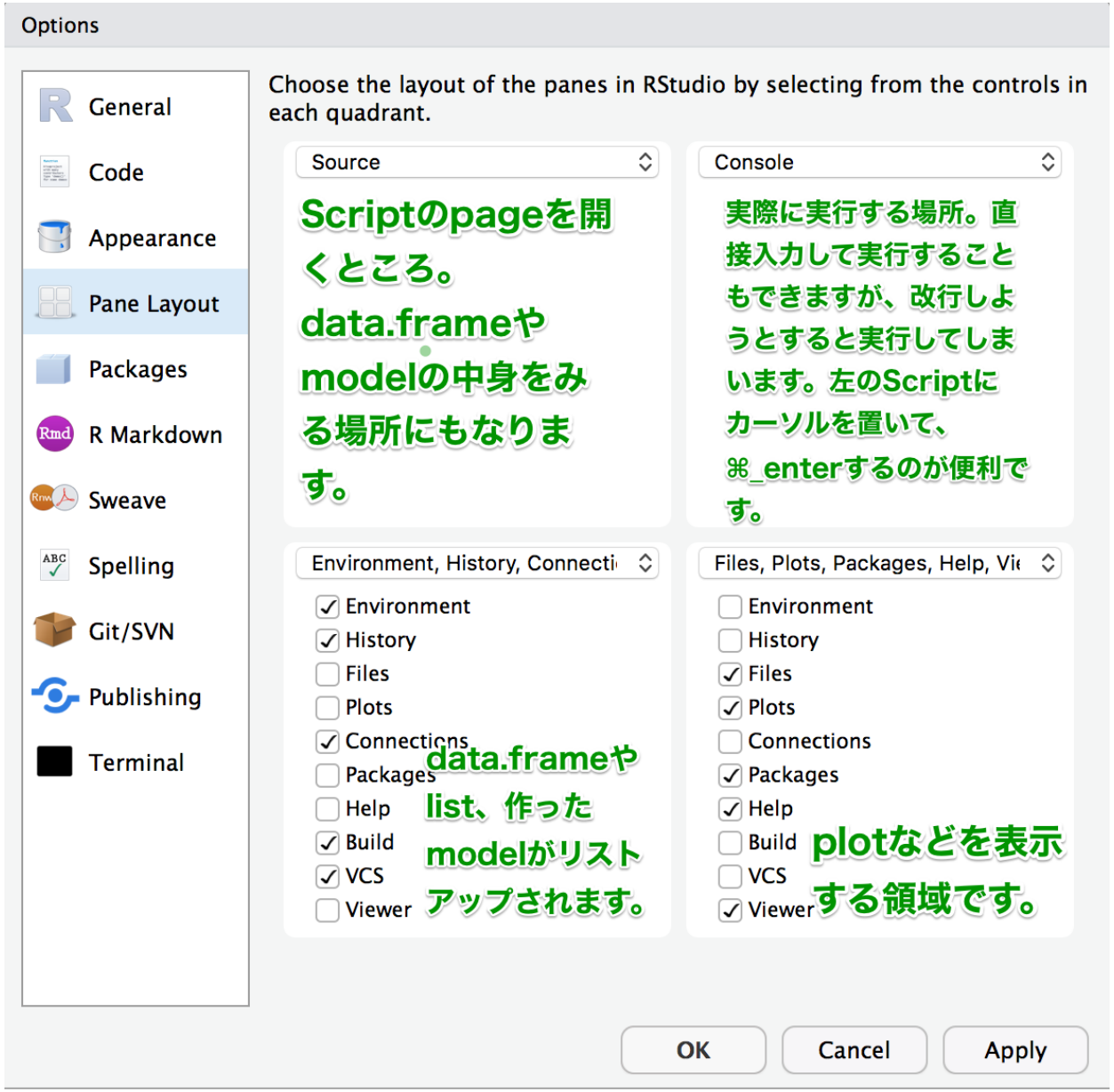
1人で100人分の仕事ができる
＝優秀な人

Part1;
RStudioの使い方

- **ペインの仕組み**

- ワーキングディレクトリ
　デフォルトワーキングディレクトリの利点

- データの保存
　.RData, .rds

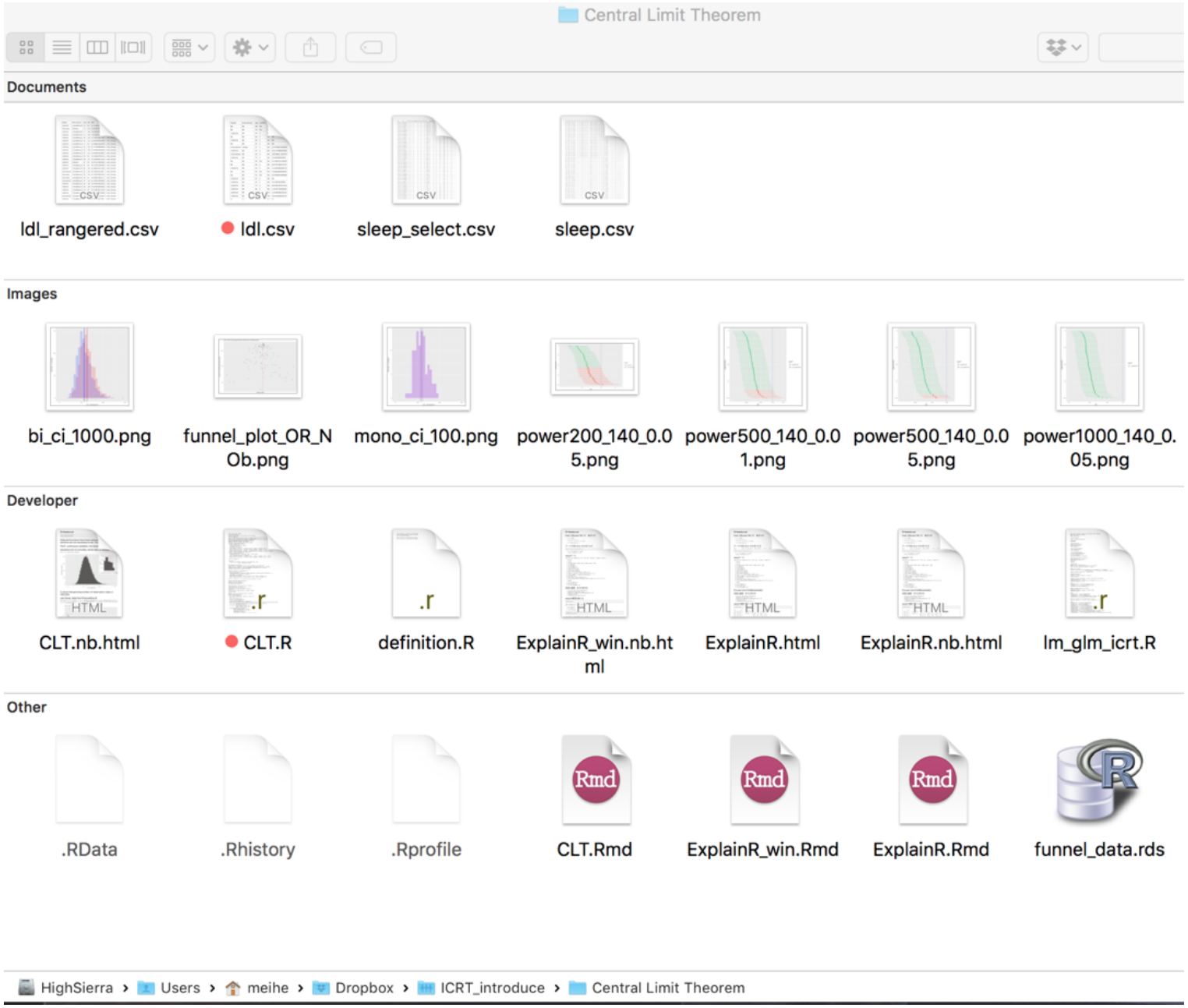


Part1;
RStudioの使い方

- ペインの仕組み

- ワークスペース
**　デフォルト ワークスペースの利点**

- データの保存
　.RData, .rds



データの保存　.RData, .rds

- save.image()

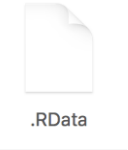
Rを終了するときに作動させる。
作ったfunctionやmodel, list, data.frameなどが全てそのまま保存できる。次の起動時にdefault workspaceから読み込まれる。

- save.image("sleep20181016?.Rdata")

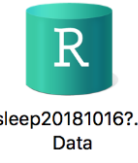
名前をつけて保存するとその時点のversionを残せる。

- saveRDS(object名, file = "わかり易い名前.rds")

data.frameの保存法。CSVより軽く、データの構造をそのまま保存できる。



save.image()でできる
不可視ファイル



名前をつけると
可視ファイルになる

saveRDS(sleep, file =
"sleep.rds")
でできる可視ファイル

データの読み込み-式の成り立ち

```
sleep　　=　　read.csv(　"sleep_select.csv"　)
#名前　　定義する　　コマンド　　コマンドの対象
```

- カーソルのある場所のコマンドを実行;
Win: ctrl + enter, Mac: ⌘ + enter

Sleep data

- 不眠と毎日の生活との関係を調べるために、1年近くに渡り、日常生活について記録をつけた。
- 1. date　日付
- 2. wk　平日を0.weekday、週末を1.weekend、日曜を2.sundayとした変数
- 3. rice　ご飯の量 0~2
- 4. fish　魚の摂取 binary

5. meat　肉の摂取量　0~3

6. vegetables　野菜の量　0~3

7. alcohol　アルコールの摂取量

8. sleep.pill　睡眠薬の使用 binary

9. bed　ベッドに入った時刻を8時を0として数値にしたもの

10. ex　運動：なしまたは少し、夕方から夜、朝 "a.little" "even" "morn"

Sleep data

11. shape.bad　頭痛肩こり・喘息発作などの体調不良 binary

12. think　考え事 binary

13. obstacle　物音、人の動きなど睡眠の妨げ binary

14. hormone　月経周期を月経期、月経後卵胞期、排卵期、黄体期に分けたもの "E1" "E2" "O" "P"

15. sleep.min　睡眠時間（分）

16. insomnia　不眠　binary

17. noct_awake　夜間覚醒　binary

変数の種類：Rでの呼び方

- カテゴリカル　　factor, ordered = FALSE, referenceあり
- 順序変数(ordinal)　　factor, ordered = TRUE
- 量的変数　　numeric
  - 離散変数　　integer 整数
  - 連続変数　　double　少数
- 文字列(string)　　character
- 二項変数(binary)　　logical (TRUE, FALSE) (1, 0)

データ種類

- **数字の性質から：**
  - **カテゴリーデータ（数字は割り振られているが 所属カテゴリーの意味のみ）**
  - **順序変数**（カテゴリー変数ではあるが, 数字の順序に意味がある）
  - **量的変数（数字の順序と数字同志の間隔に意味がある）**

- **量的変数の性質：**
  - **離散変数**（二つの数字の間に数えられる数字がある）
  - **連続変数（二つの数字の間にある数字の数は数えられないほど無限）**

- **データの頻度の点から：**
  - **正規分布 vs 非正規分布**

- **生存時間(打ち切り)データ**

BIO 206/208, HSPH
Lectures Dr. John Orav

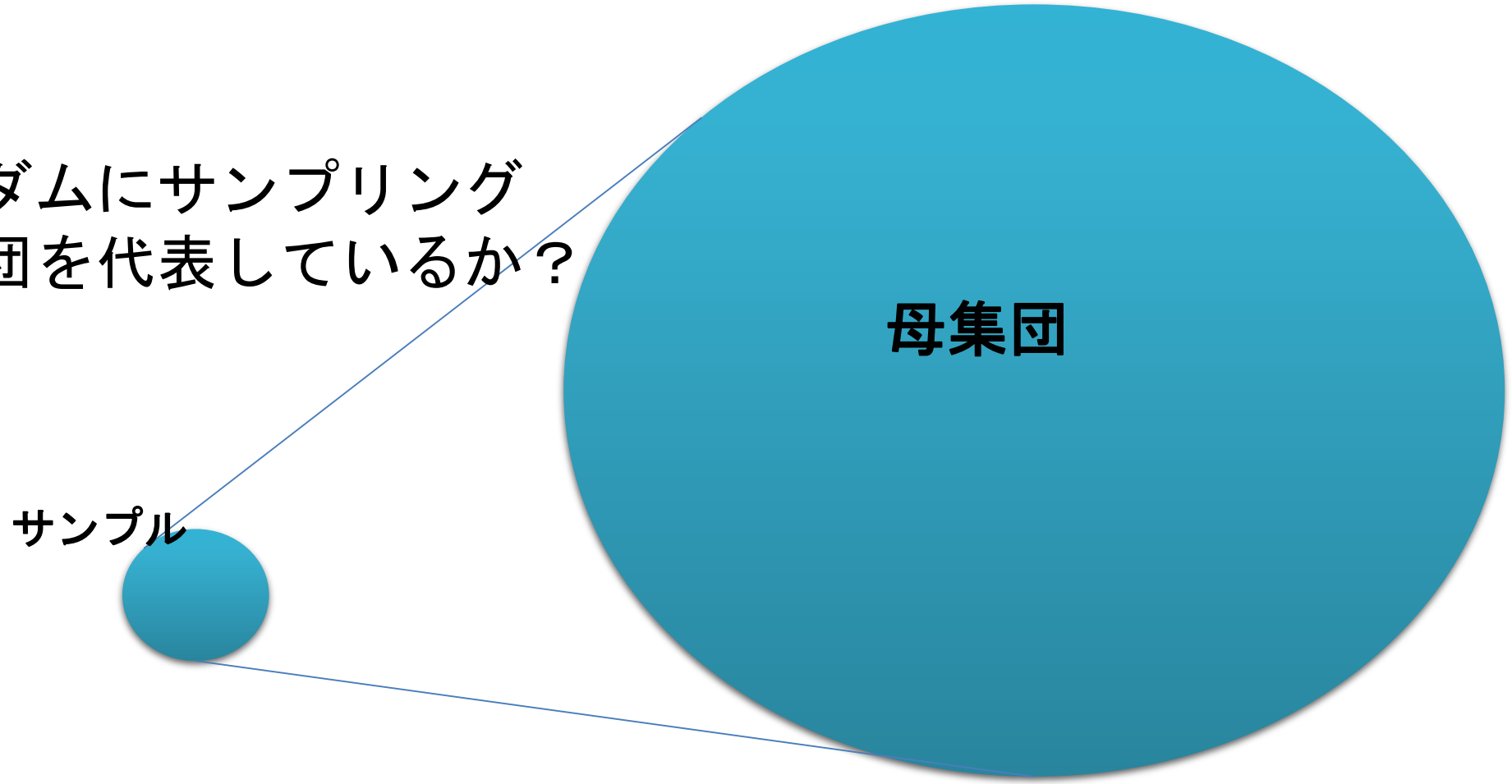
データの種類分けの例

注）政治, 宗教, 人種, 性別などに関する議論は
きわめて気を付けてください

データ	割り振られる値	変数/の種類
性別	男 , 女	二項変数
身長（cm）	- cm	量的変数 (連続変数)
体重（kg）	- kg	量的変数(連続変数)
年齢（歳）	- 歳	量的変数(離散変数)
血圧（mmHg）	-mmHg	量的変数(連続, 正規)
喫煙の有無	Yes, no, ex-smoker	カテゴリー変数
病院のタイプ	総合病院, 療養型, リハビリ病院, 公立病院, 市立病院	カテゴリー変数
保険制度に対する賛成度	反対, やや反対, 中立, やや賛成, 賛成	順序変数
日常活動度	低下, やや低下, 問題なし	順序変数
人種	Asian, African American, White, non-white	カテゴリー変数
生存期間(年)	- year	量的変数(連続, 打ち切り)

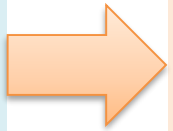
母集団とサンプルのデータ

- ランダムにサンプリング
- 母集団を代表しているか？



データの位置関係

説明変数
(曝露変数, 独立変数)
Explanatory
Exposure
Independent
(variable)



結果変数
(アウトカム, 従属変数)
Outcome
Dependent
(variable)

例：　たばこ
お酒

例：　肺がん
肝硬変

BIO 206/208, HSPH
Lectures Dr. John Orav

Part 1: sleepの概要を調べる

```
sleep = read.csv("sleep_select.csv")
summary(sleep)
```

```
         date              wk           rice          fish       
 2012-10-11:  1   0:weekday:190   Min.   :0.00   Min.   :0.000  
 2012-10-12:  1   1:weekend: 92   1st Qu.:1.00   1st Qu.:0.000  
 2012-10-13:  1   2:sunday : 46   Median :1.00   Median :0.000  
 2012-10-14:  1                  Mean   :1.18   Mean   :0.378  
 2012-10-15:  1                  3rd Qu.:2.00   3rd Qu.:1.000  
 2012-10-16:  1                  Max.   :2.00   Max.   :1.000  
 (Other)   :322                                                 
      meat          vegetables       alcohol          sleep.pill      
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :1.000   Median :2.000   Median :1.0000   Median :0.00000  
 Mean   :1.204   Mean   :1.921   Mean   :0.6341   Mean   :0.05793  
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:0.00000  
 Max.   :3.000   Max.   :3.000   Max.   :2.0000   Max.   :1.00000  
```

データクリーニングの例

- 新しい変数を作る。
- クラスを変える。
- リファレンスを変える。

```
sleep$hormone = relevel(as.factor(sleep$hormone), ref="E2")
summary(sleep$hormone)
```

```
E2  E1   O   P
 93  59  55 121
```

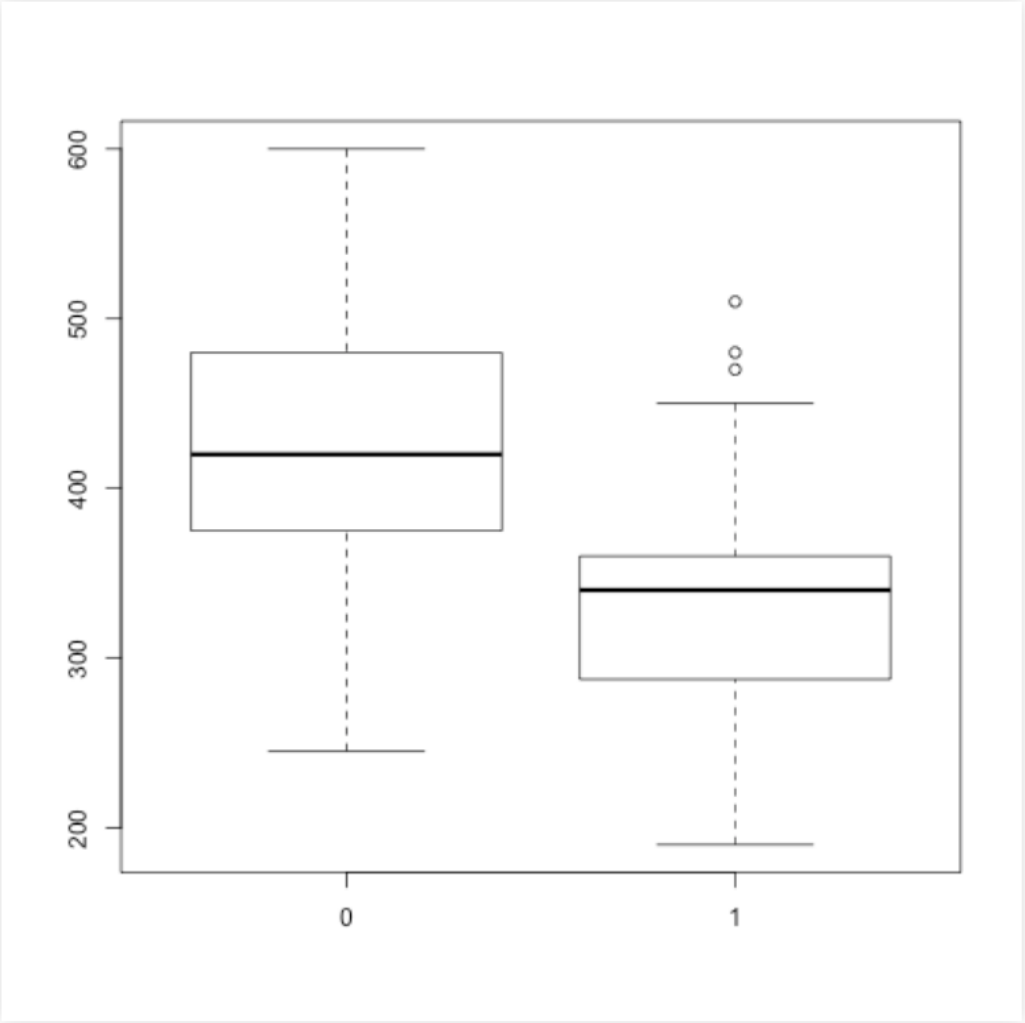
パッケージを呼び出して使う

```
#install.packages("lubridate")
library(lubridate)
sleep$date = ymd(sleep$date)
summary(sleep$date)
```

```
        Min.      1st Qu.       Median         Mean      3rd Qu.
"2012-10-11" "2012-12-31" "2013-03-26" "2013-04-01" "2013-07-05"
        Max.
"2013-09-25"
```

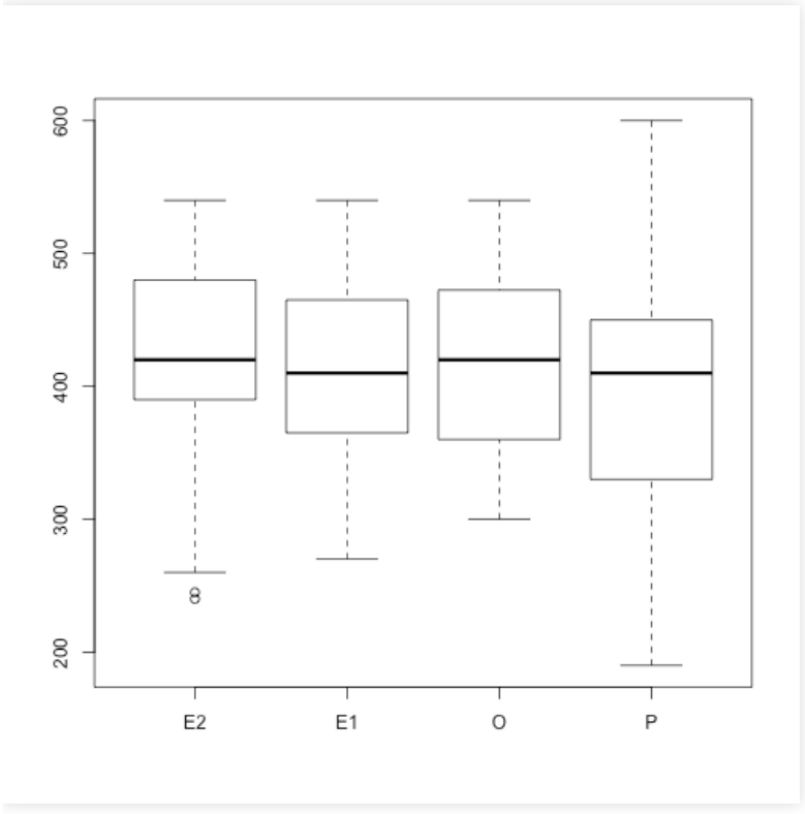
データを視覚化

```
boxplot(sleep.min ~ think, data =
sleep)
```



なぜModelに名前をつけるか

```
bp = boxplot(sleep.min ~ hormone,
data = sleep)
```




```
bp$conf
```

```
          [,1]     [,2]     [,3]
[,4]
[1,] 405.2545 389.4302 396.0322
392.7636
[2,] 434.7455 430.5698 443.9678
427.2364
```

tableで2つの変数の関係を見てみる

```
table(sleep$rice,
sleep$vegetables)
```

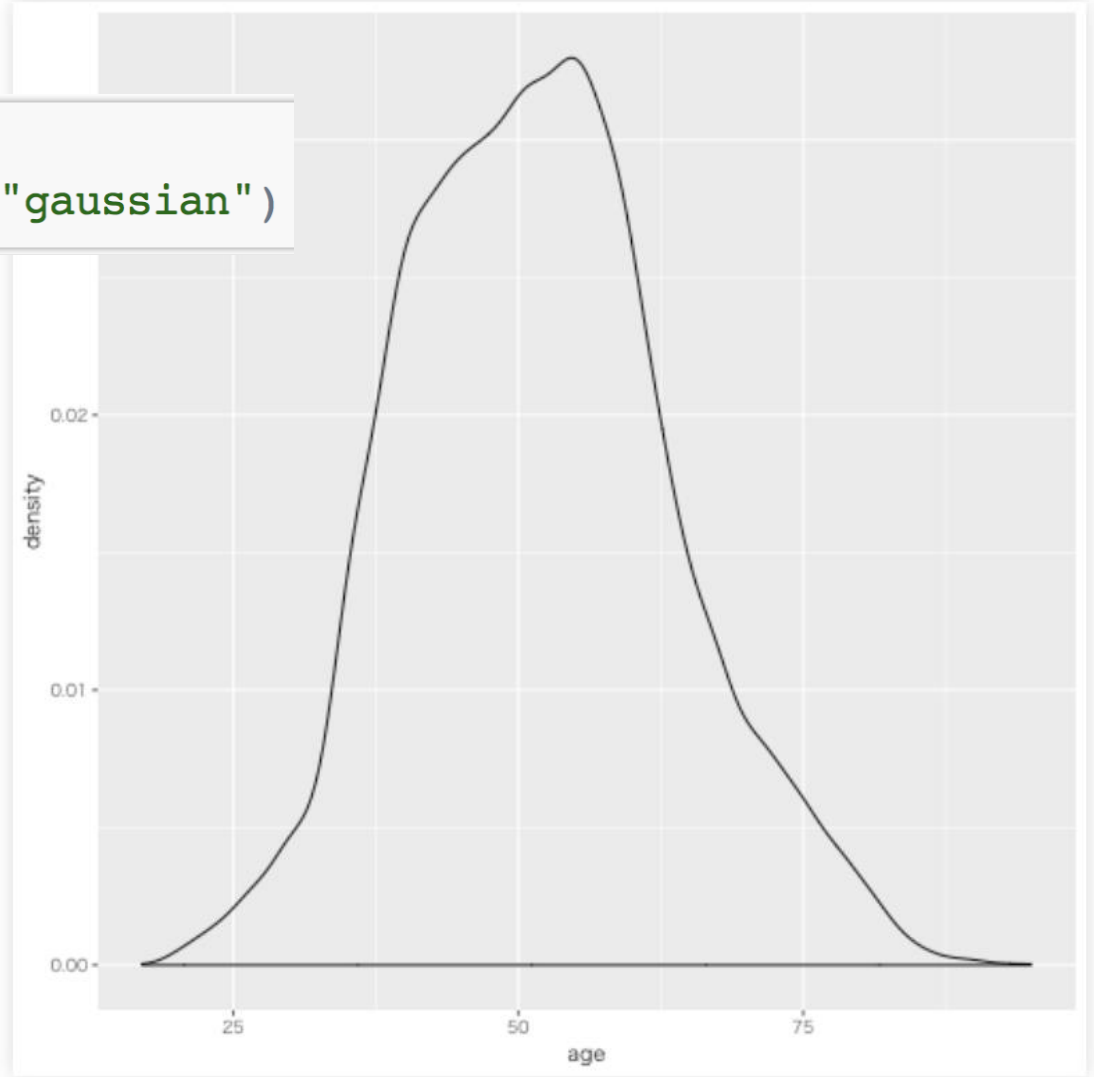
```
chisq.test(table(sleep$rice,
sleep$vegetables))
```

```
    0  1  2  3
0   3 17  9  7
1   9 83 47 58
2   0 13 36 46
```

```
        Pearson's Chi-squared test

data:  table(sleep$rice,
sleep$vegetables)
X-squared = 36.977, df = 6, p-
value = 1.779e-06
```

別のデータでさらにPlotを試す：単変量~~数値~~

```
library(ggplot2)
ggplot(ldl, aes(age)) + geom_density(kernel = "gaussian")
```

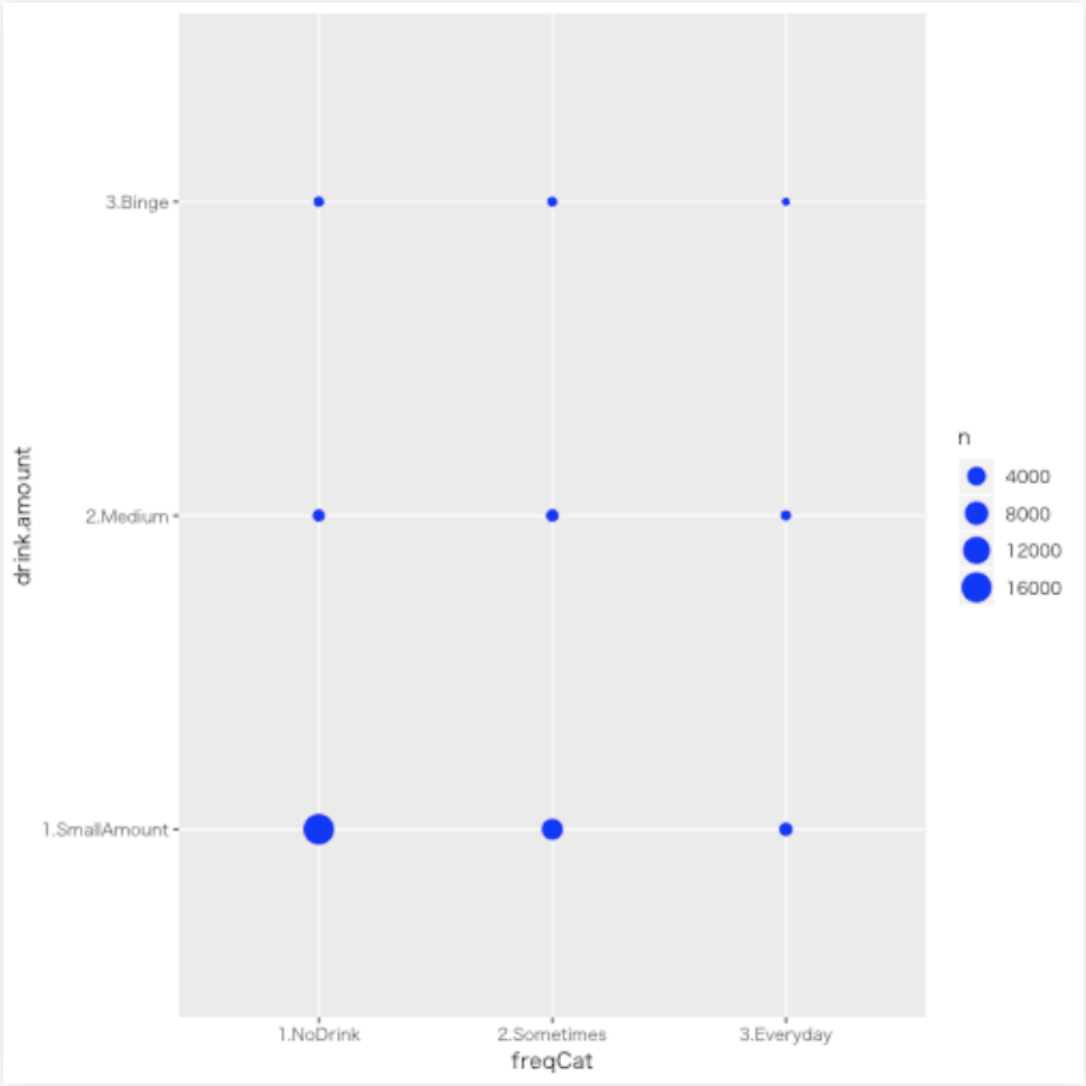


2つの変数

カテゴリカル vs カテゴリカル

factor(nominal) vs factor(nominal)

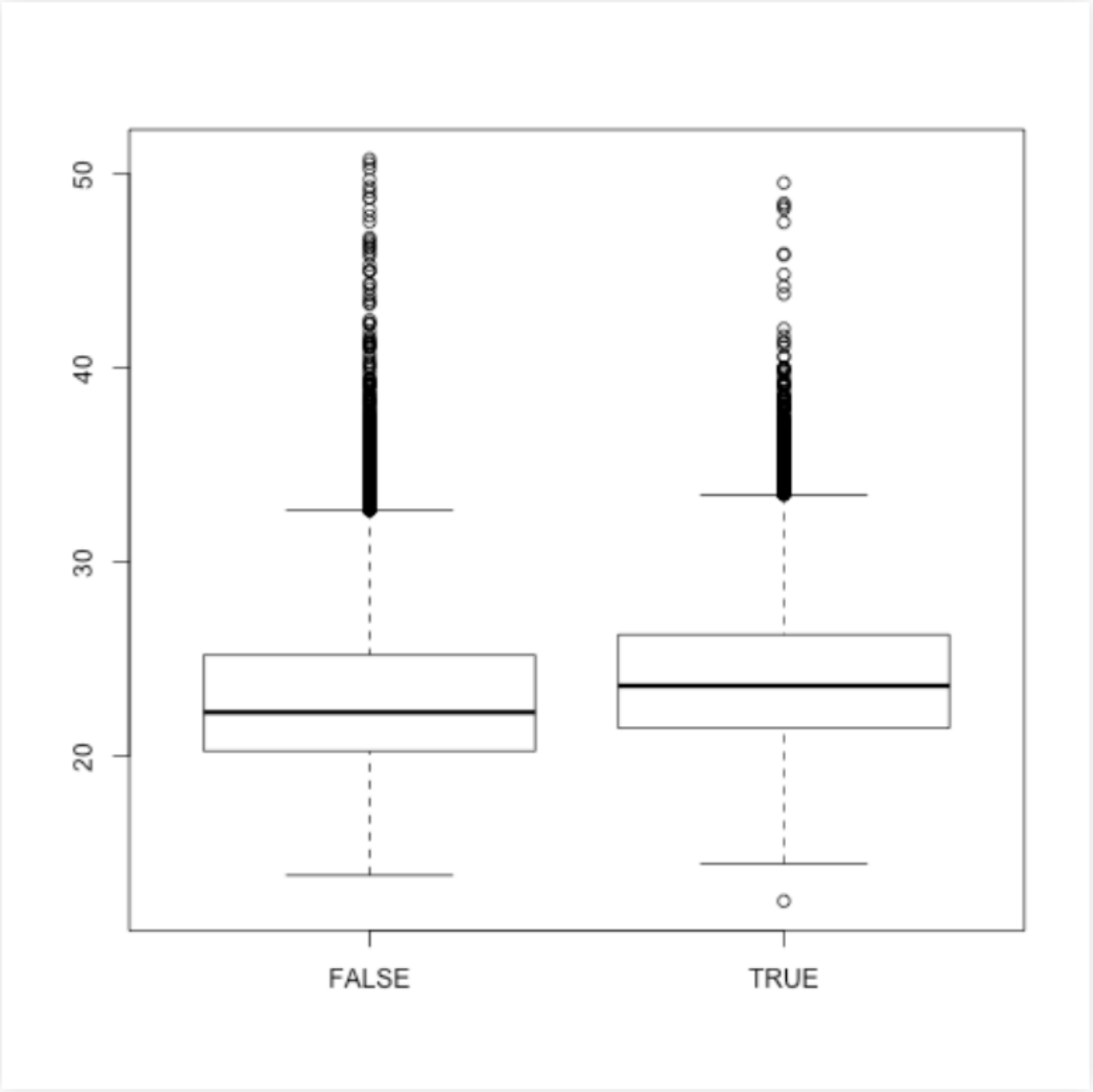
```
ggplot(ldl, aes(freqCat, drink.amount)) + geom_count(color = "blue")
```



カテゴリカル　vs　数値

factor(nominal) vs numeric

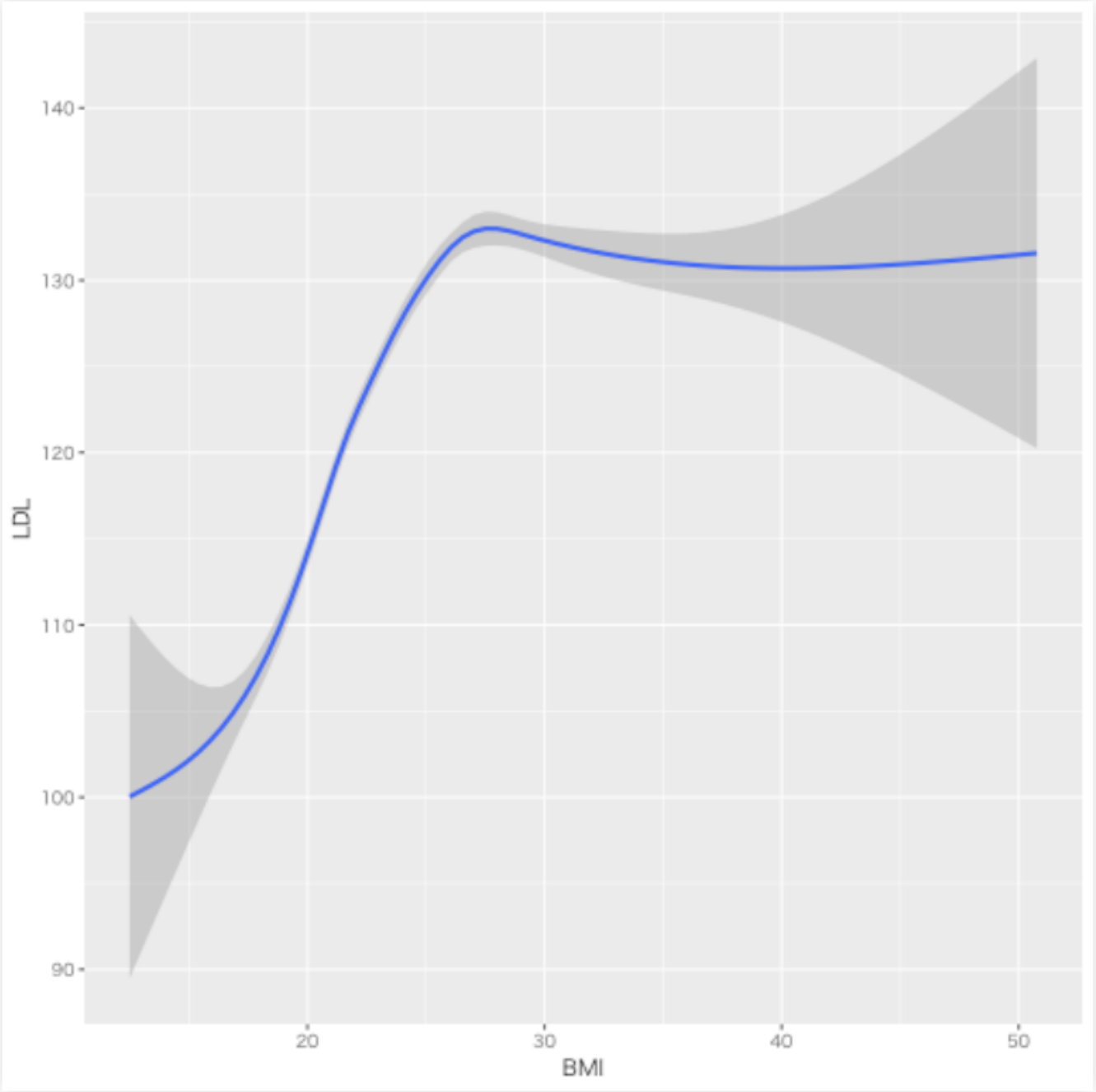
```
bp = boxplot(BMI ~ age>=50,
data = ldl)
```



数値　vs　数値

numeric vs numeric

```
ggplot(ldl, aes(BMI, LDL)) +
geom_smooth()
```

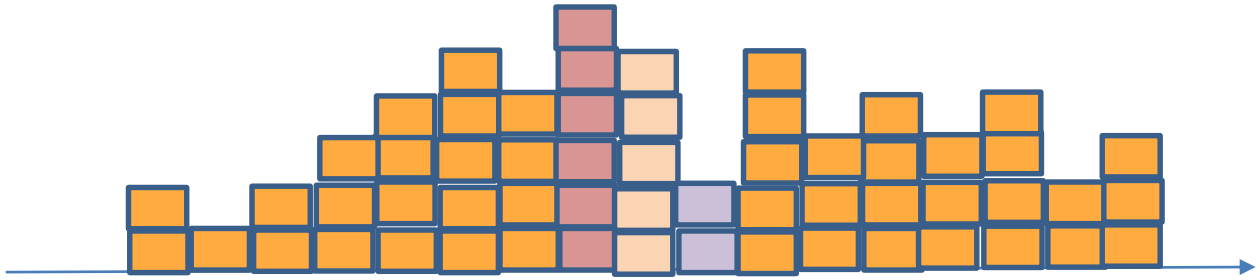


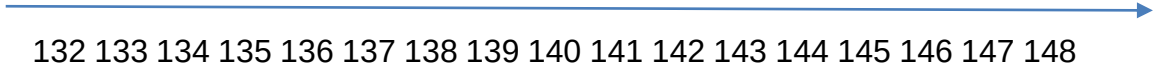
自分のデータを
二つの指標で要約する(言い換える)

連続変数の場合

代表値（平均, 中央値, 最頻値）

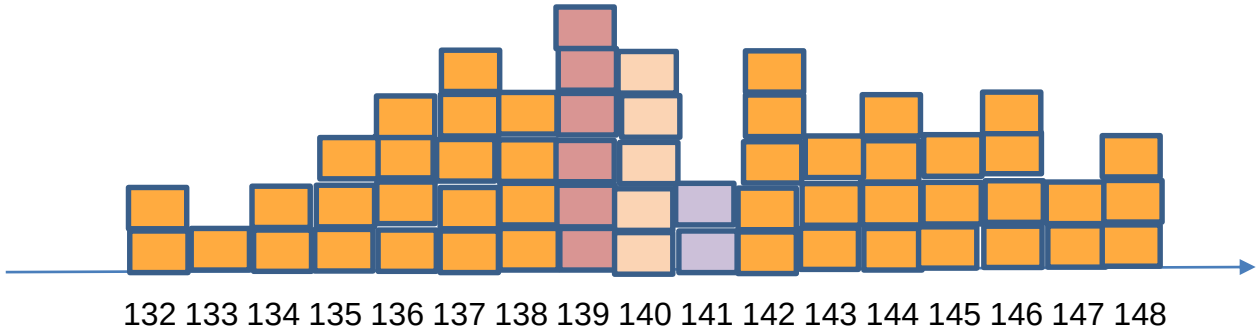
ばらつき（分散, 標準偏差）





自分のデータを
二つの指標で要約する(言い換える)

連続変数の場合



代表値（平均, 中央値, 最頻値）

ばらつき（分散, 標準偏差）

分散

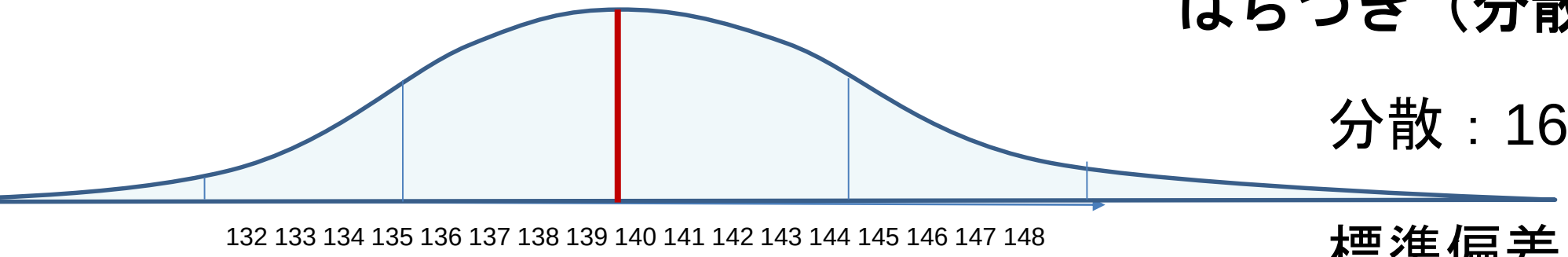
$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{(n-1)} \quad =16$$

標準偏差

$$SD = \sqrt{S^2} \quad = 4$$

自分のデータを
二つの指標で要約する(言い換える)

連続変数の場合



代表値（平均）=139.5

ばらつき（分散, 標準偏差）

分散：16

標準偏差：4

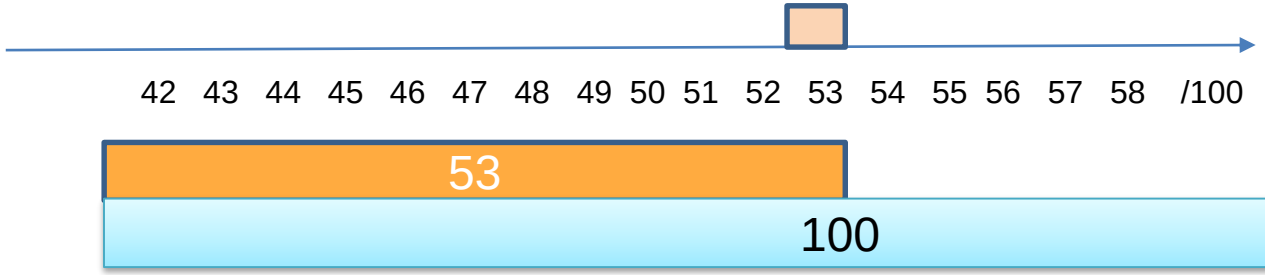
自分のデータを
二つの指標で要約する(言い換える)

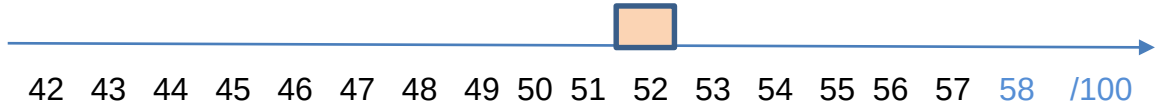
比率の場合

x = 53　n = 100

代表値（比率）$\hat{p}$ =　x/n = 0.53

真のxが起こる確率の予測値





自分のデータを
二つの指標で要約する(母集団を予測する)

比率の場合

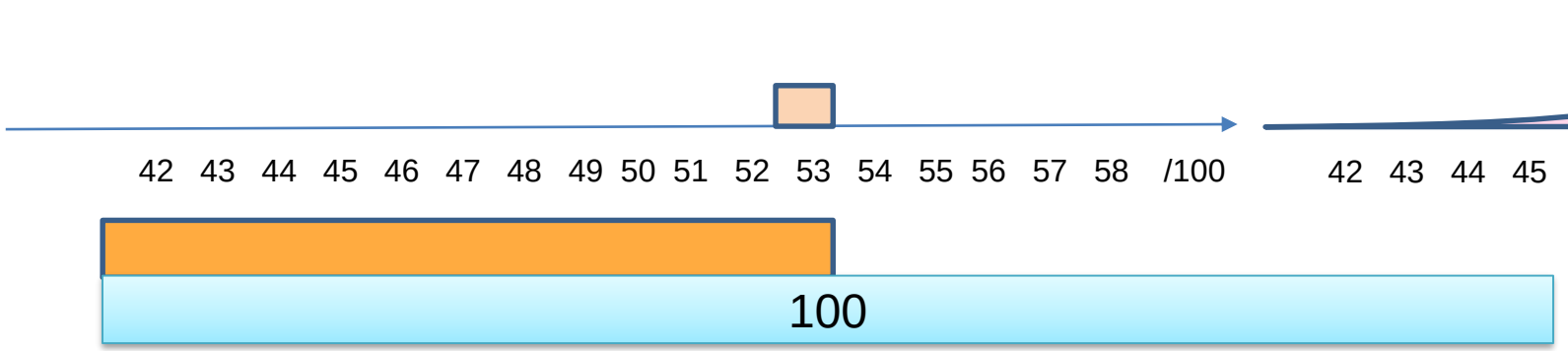
代表値（比率）$\hat{p}$ = x/n = 0.53

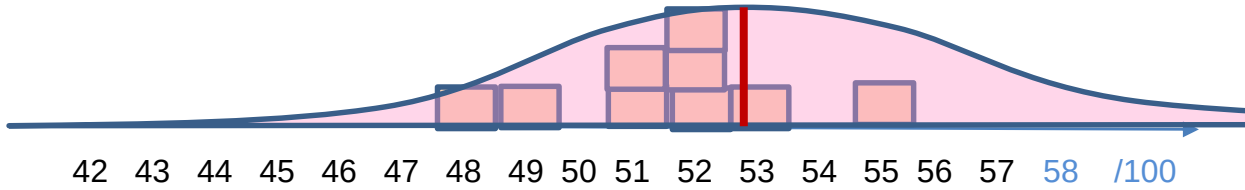
真のxが起こる確率の予測値

分散？ = ((p) x (1-p))/n = (0.53) x (0.47) / 100

標準誤差は 分散の平方根 =√ (0.53) x (0.47) / 100

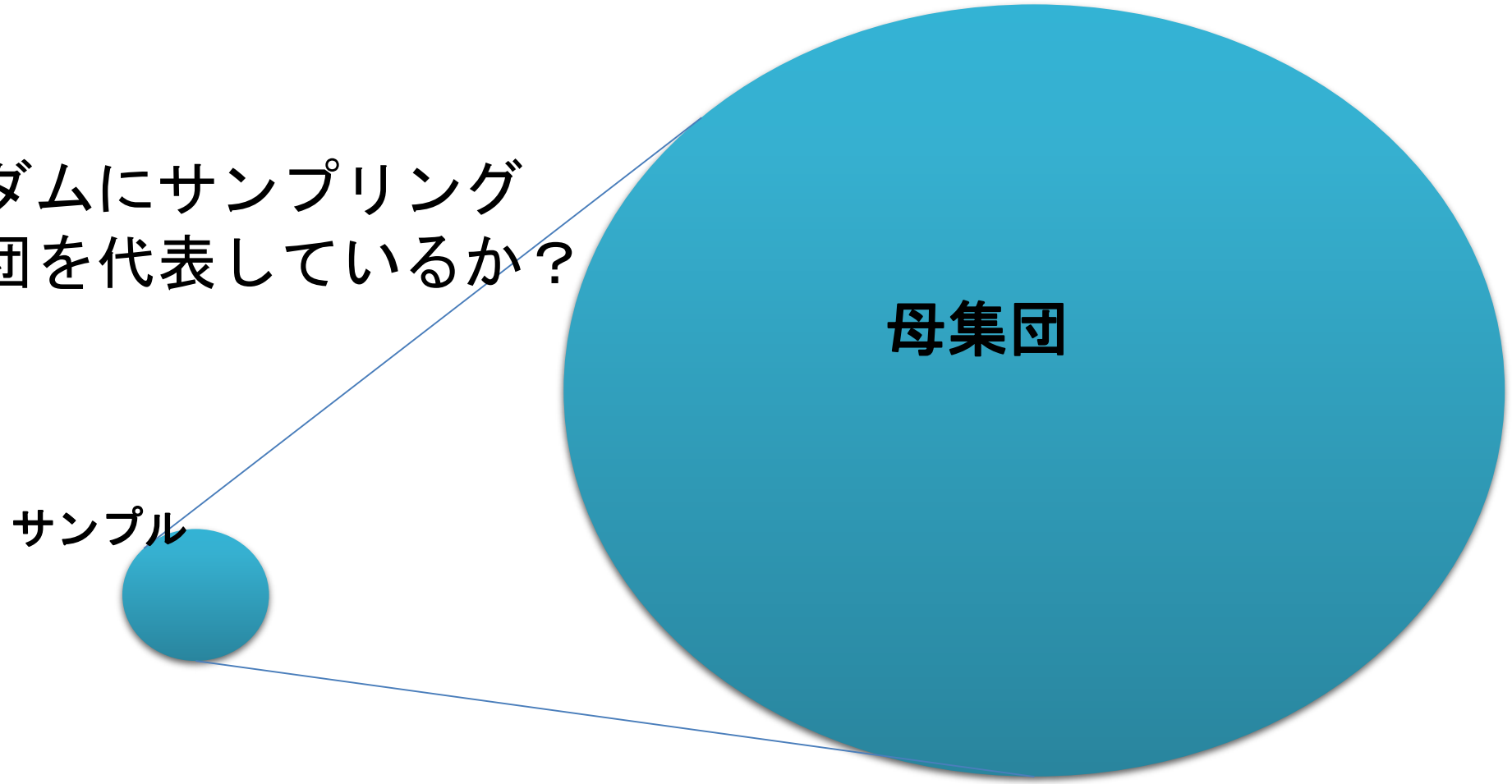
標準誤差 = 0.0499, 95%信頼区間 0.43~ 0.63





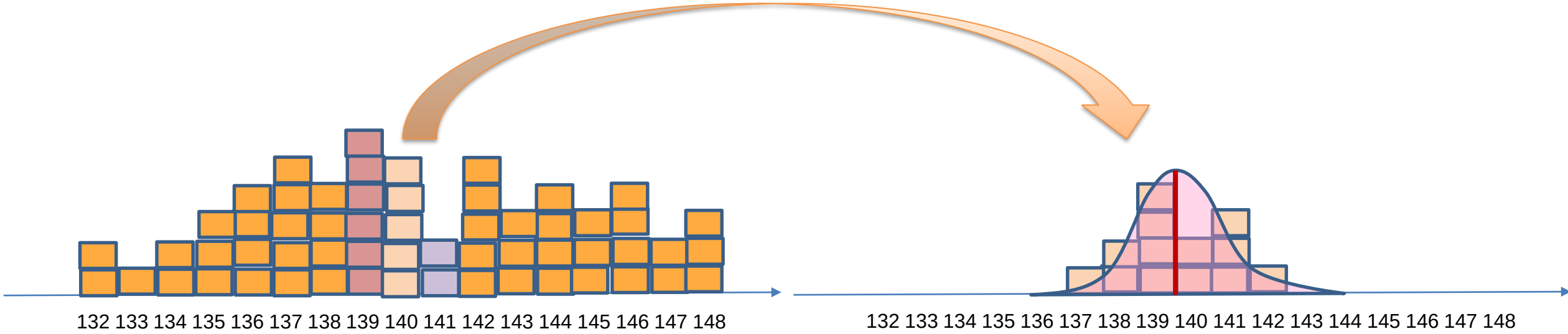
母集団とサンプルのデータ

- ランダムにサンプリング
- 母集団を代表しているか？



自分のデータを
二つの指標で要約する(母集団を予測する
)

連続変数の場合



自分のデータを
二つの指標で要約する(母集団を予測する)

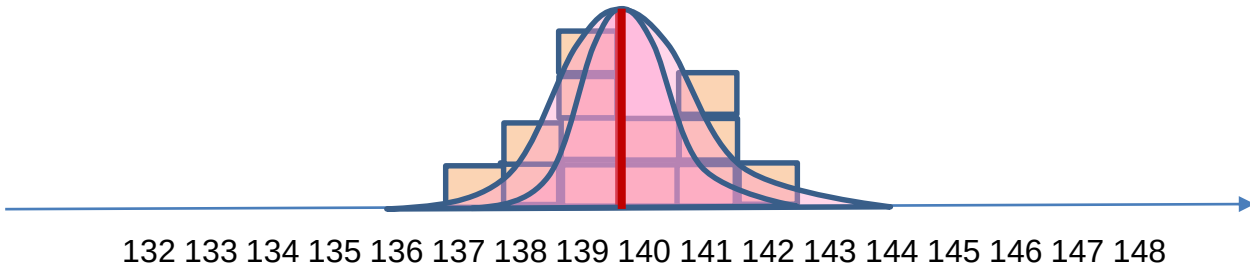
連続変数の場合

分散

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{(n-1)}$$ =16

標準偏差

$$SD = \sqrt{S^2}$$ = 4

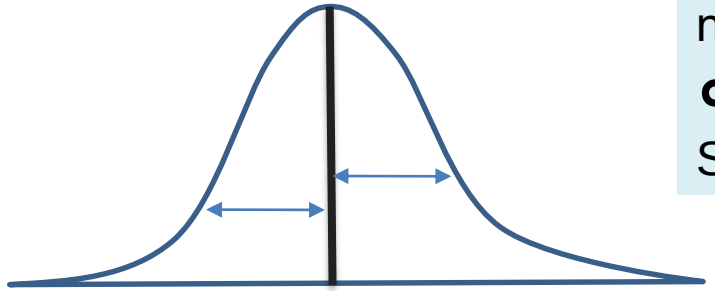


標準誤差

$$SE = sd(\overline{x}) = \sqrt{\frac{Var(x)}{n}} = \sqrt{\frac{S^2}{n}}$$ = 0.5 (n=64), = 0.2 (n=100)

３つのばらつき指標
分散, 標準偏差(SD), 標準誤差(SE)

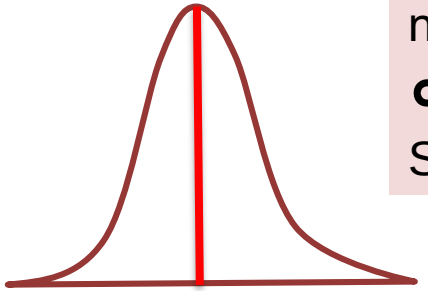
サンプルのばらつきの二乗の平均＝分散



n ↑
∝
S^2 , SD fluctuate

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-1)} \qquad SD = \sqrt{S^2}$$

サンプルの平均のばらつきの二乗の平均
＝分散/n



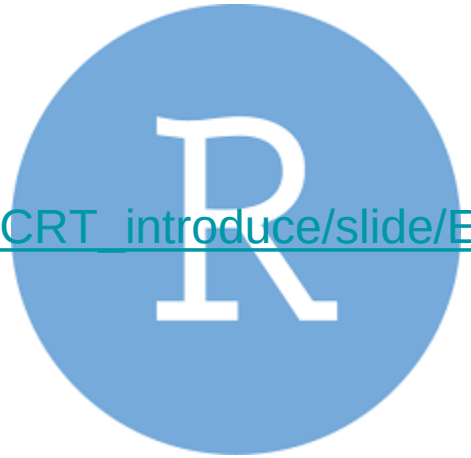
n ↑
∝
SE smaller

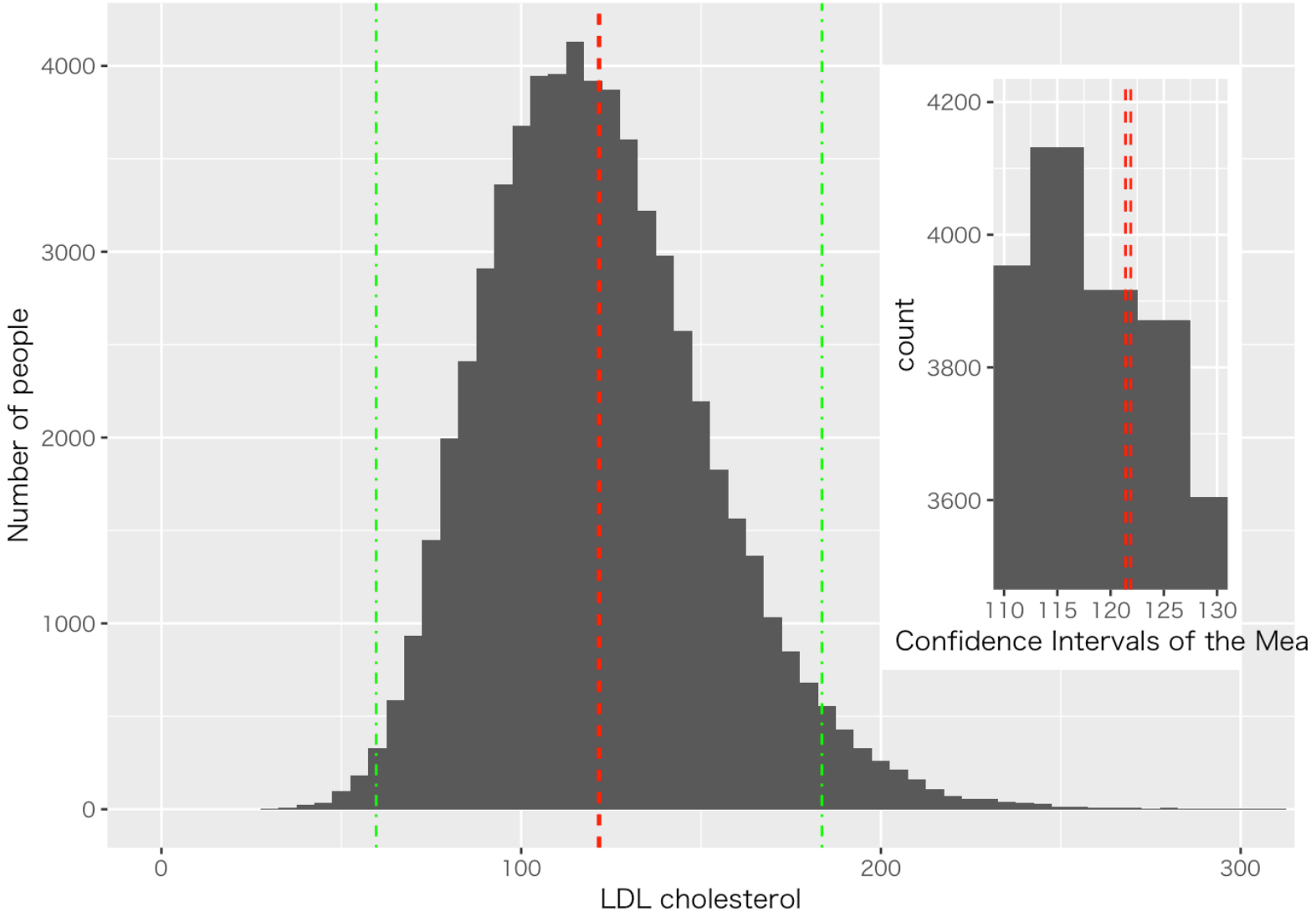
$$SE = sd(\overline{x}) = \sqrt{\frac{Var(x)}{n}} = \sqrt{\frac{S^2}{n}}$$

- サンプルのばらつきの二乗の平均　＝**分散**
- 分散の平方根＝**標準偏差**＝Standard Deviation
- 個人間のばらつきの指標（Sample SD）
- **サンプルデータ**の表現法

- 平均のばらつきの二乗の平均　＝**分散/n**
- 分散/nの平方根＝**標準誤差**＝Standard Error
- 全体の平均がどこに収まるか（Confidence Interval）
- **データ全体の予測範囲**の表現法　　信頼区間

/Users/meihe/Dropbox/ICRT_introduce/slide/ExplainR_CLT.html





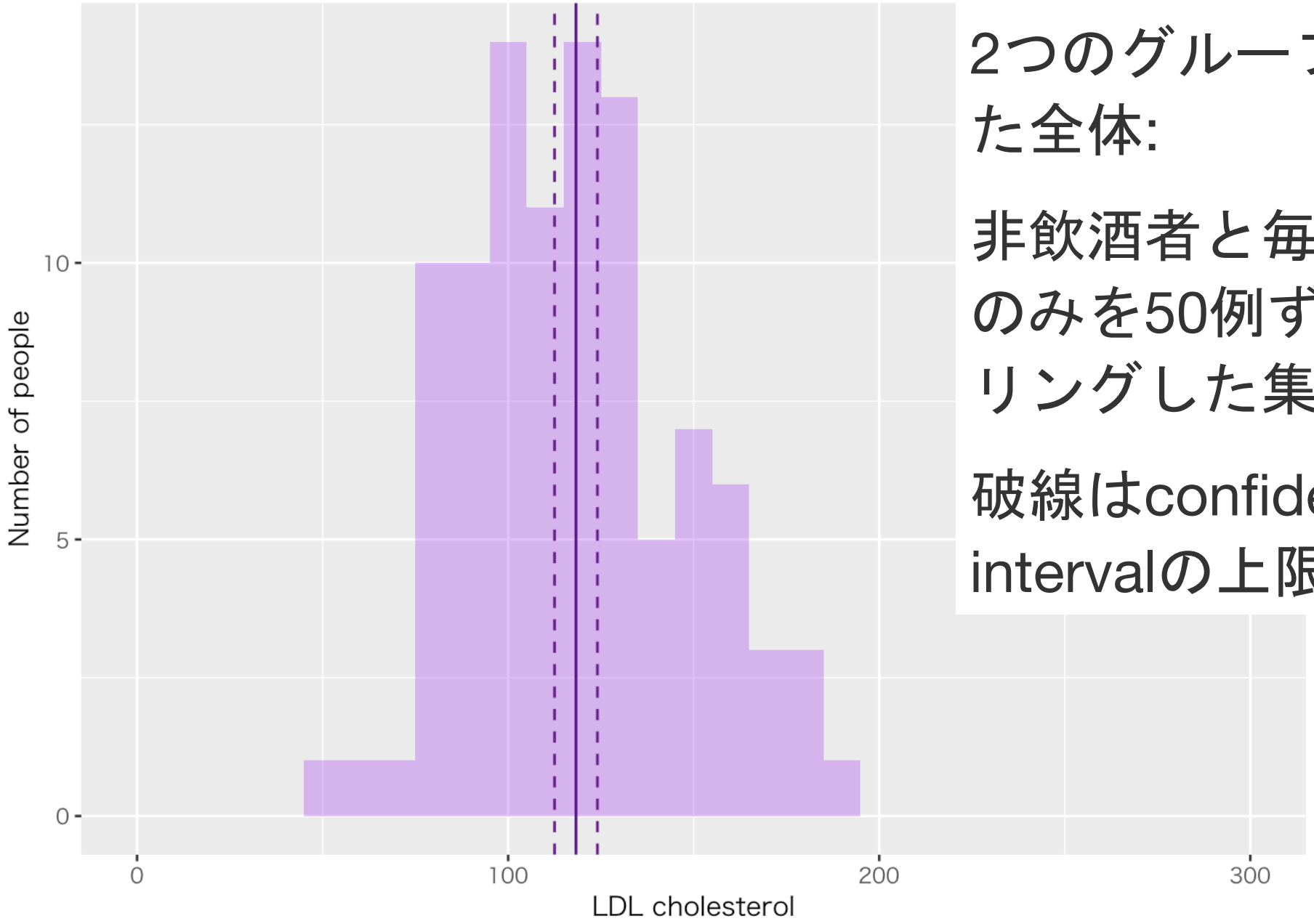
標準偏差と標準誤差

このデータを「母集団」ということにして

→非飲酒者と毎日飲酒者を同数サンプリング

→サンプルのヒストグラム そのサンプルから得られる「母集団」の平均の信頼区間を見ます。

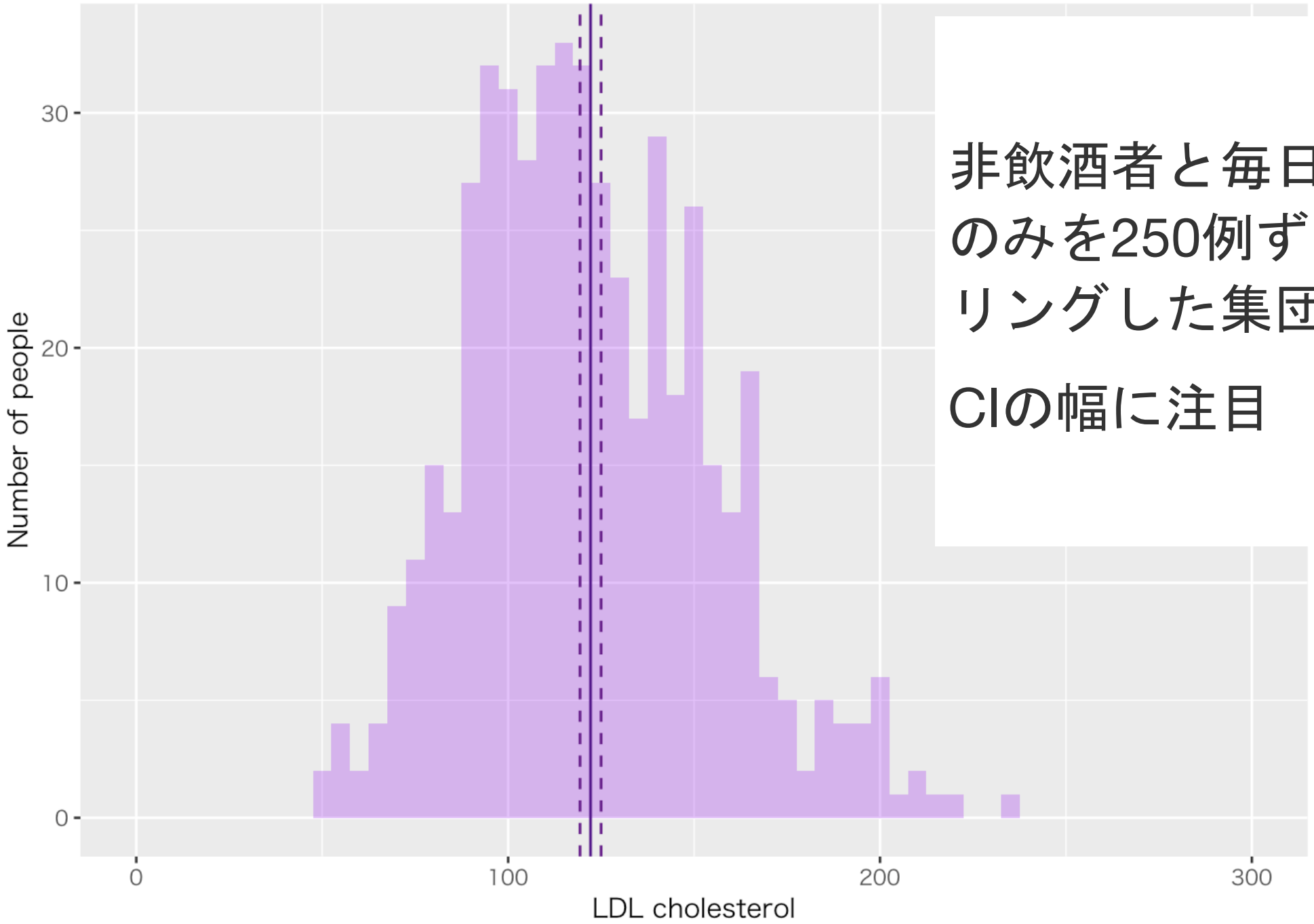
サンプル数と標準誤差との関係を見てみましょう。



2つのグループをまとめた全体:

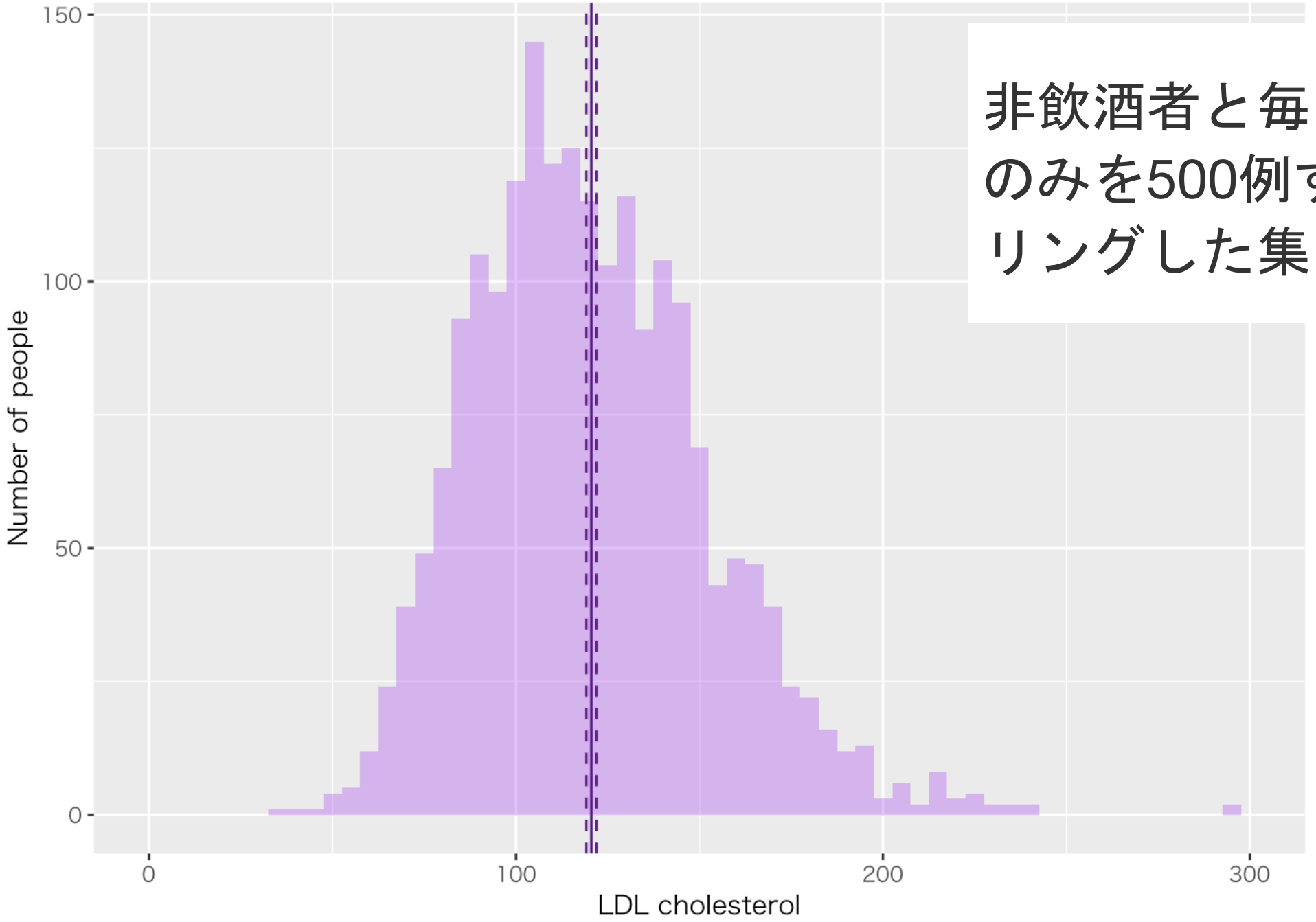
非飲酒者と毎日飲酒者のみを50例ずつサンプリングした集団

破線はconfidence intervalの上限と下限

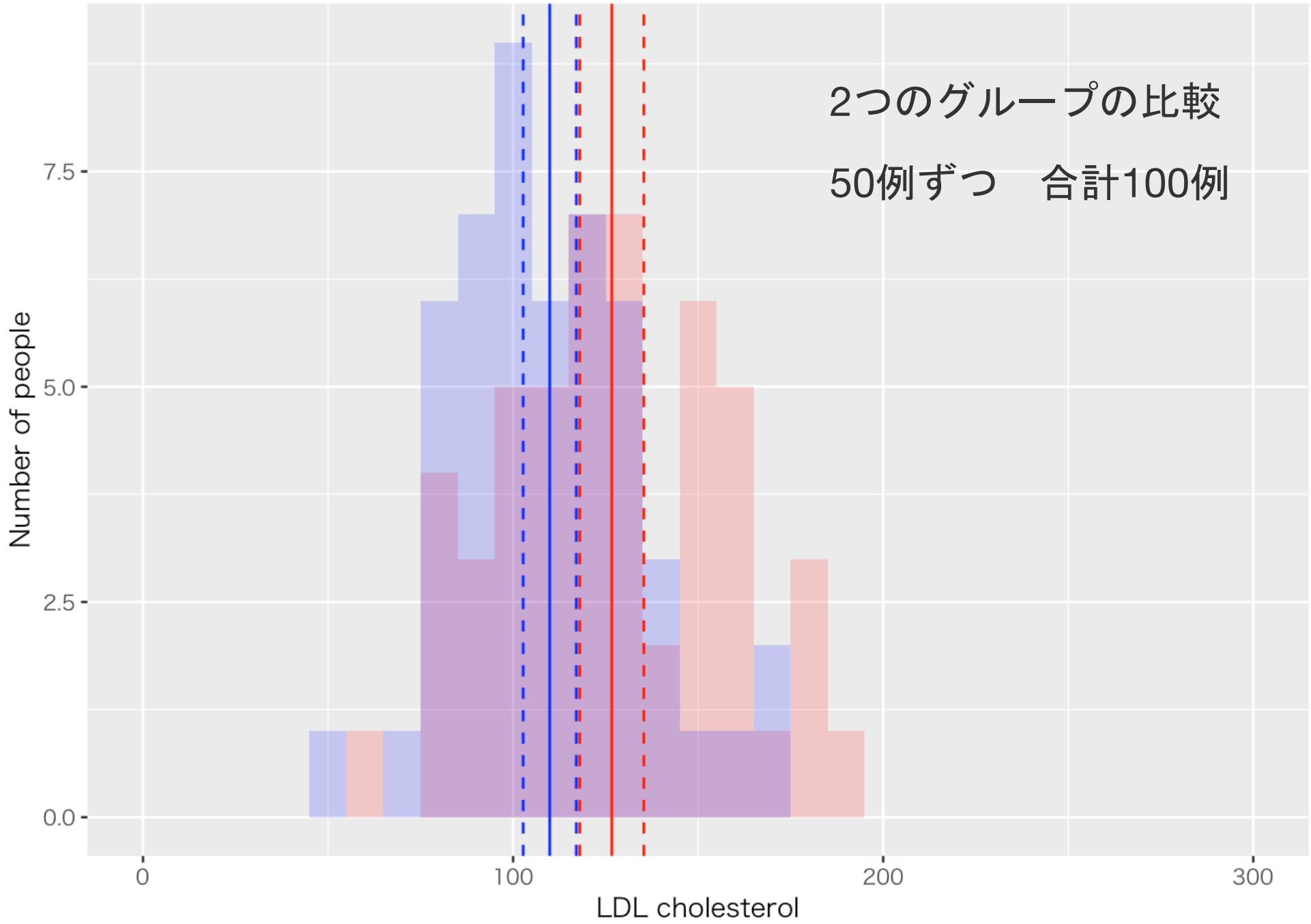


非飲酒者と毎日飲酒者のみを250例ずつサンプリングした集団

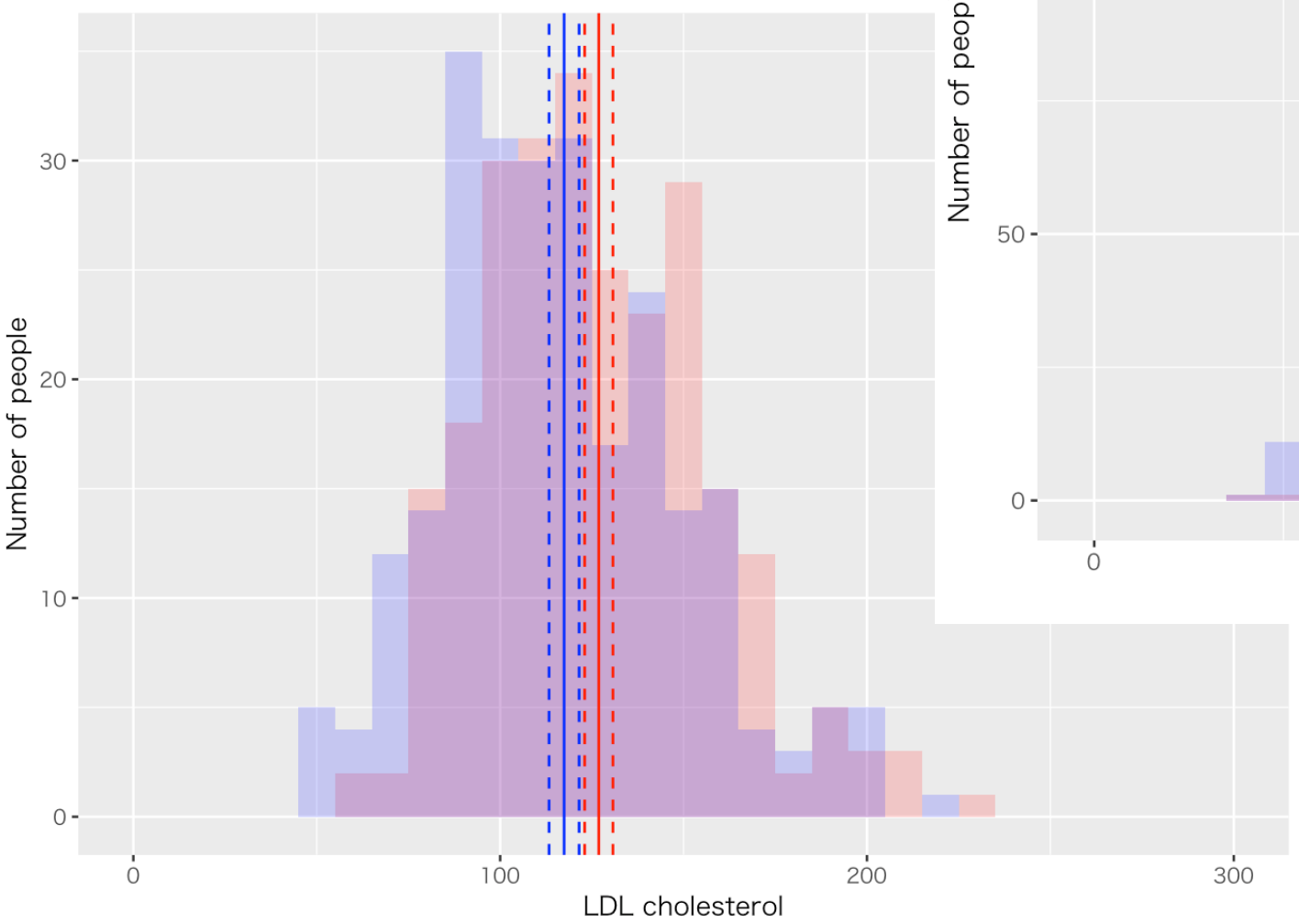
CIの幅に注目

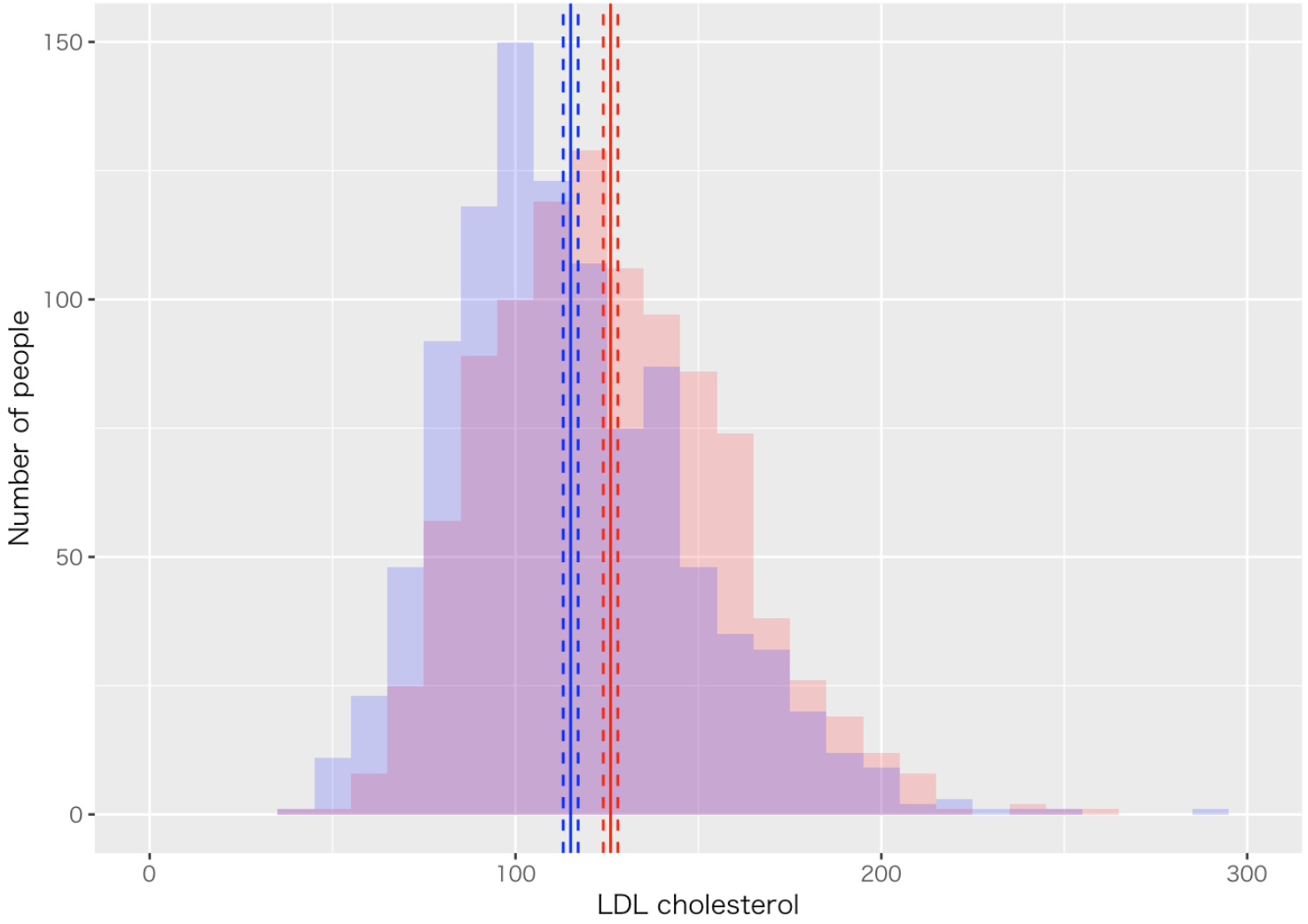


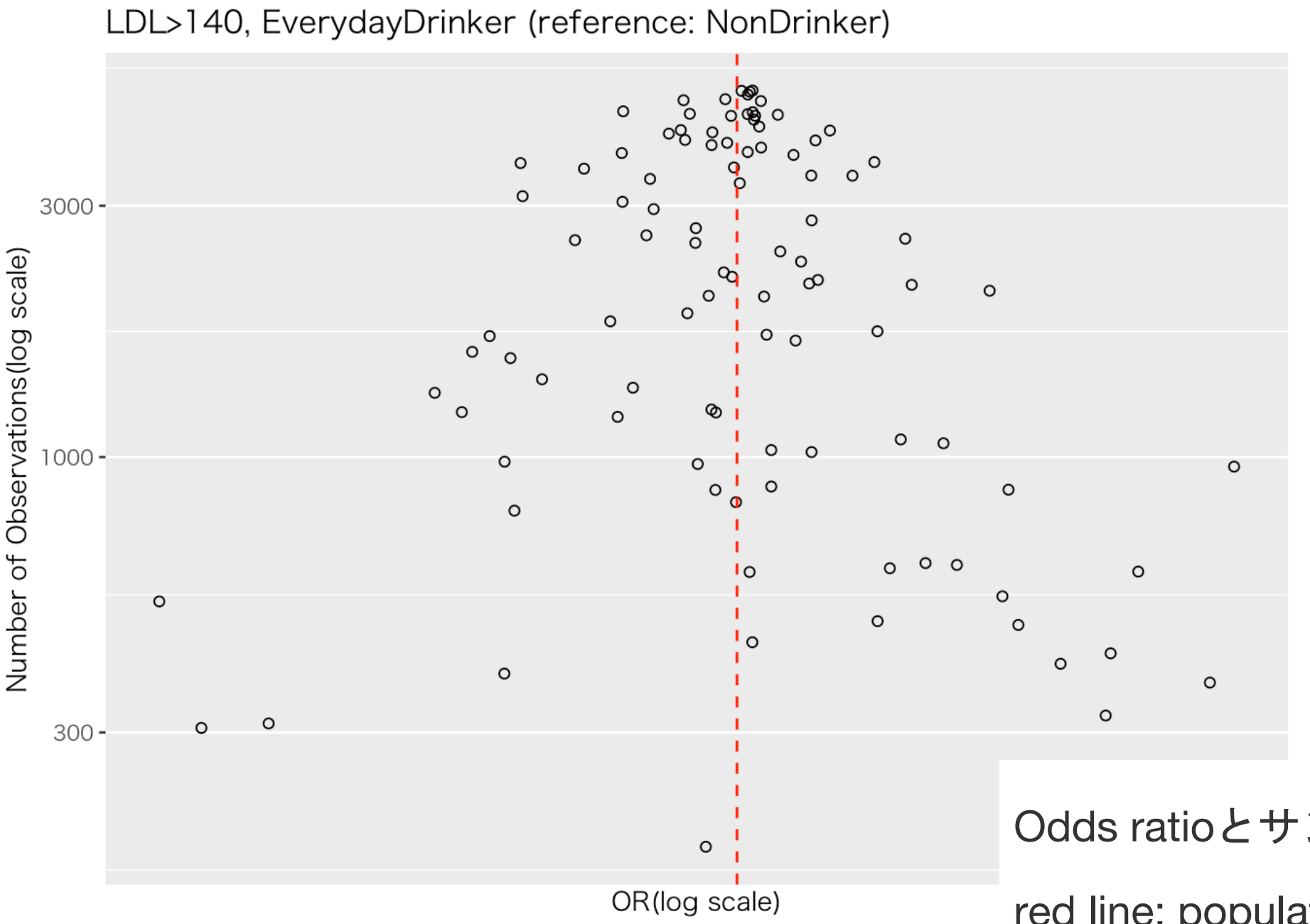
非飲酒者と毎日飲酒者のみを500例ずつサンプリングした集団


2つのグループの比較
50例ずつ　合計100例
Number of people
7.5
5.0
2.5
0.0
0
100
200
300
LDL cholesterol

↓250例ずつ　→500例ずつ

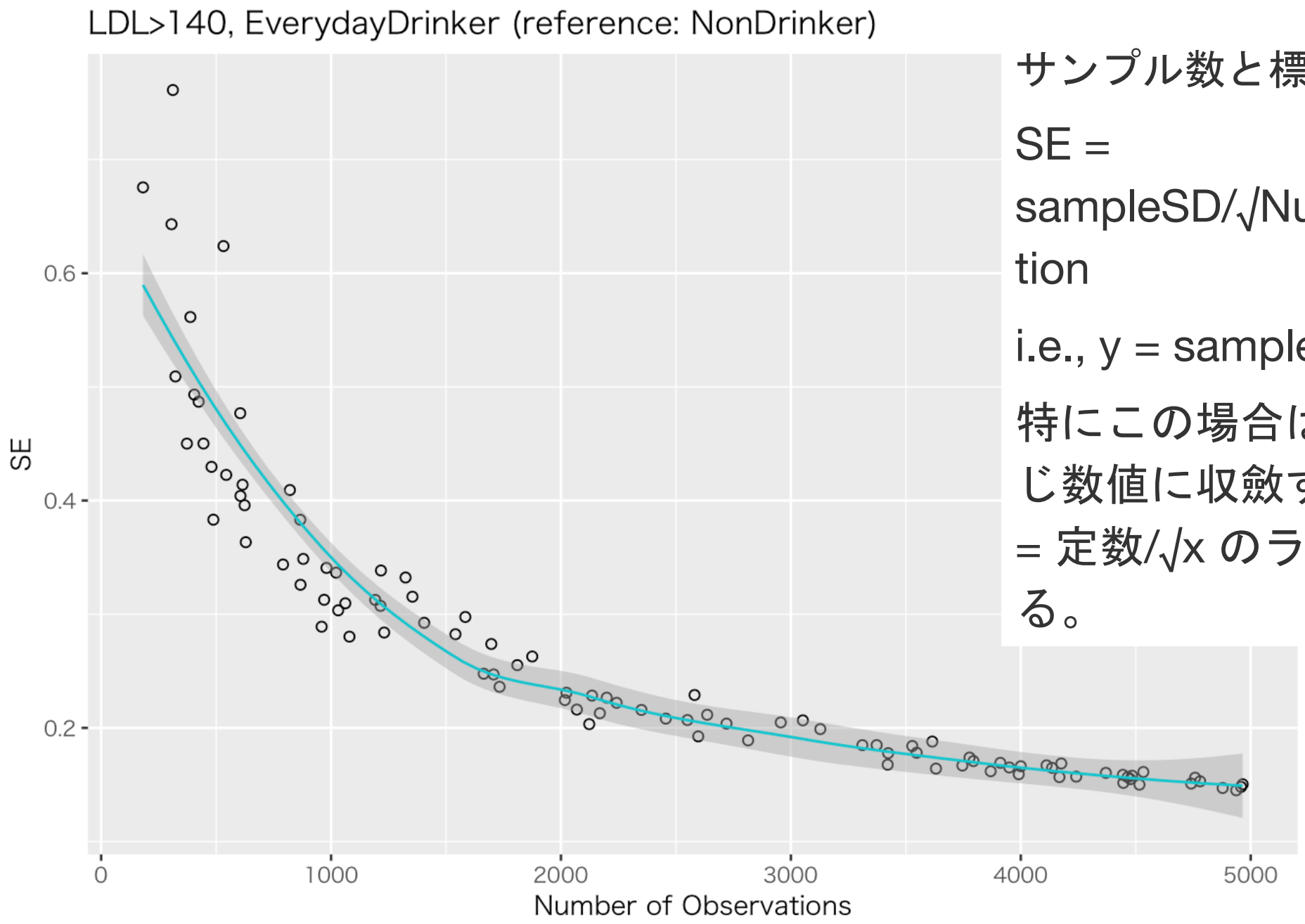






Odds ratioとサンプル数との関係

red line: population OR



サンプル数と標準誤差との関係

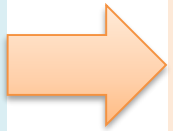
SE = sampleSD/√NumberOfObservation

i.e., y = sampleSD/√x

特にこの場合はsampleSDが同じ数値に収斂するはずなので y = 定数/√x のラインに近似される。

データの位置関係

説明変数
(曝露変数, 独立変数)
Explanatory
Exposure
Independent
(variable)



結果変数
(アウトカム, 従属変数)
Outcome
Dependent
(variable)

例：　たばこ
お酒

例：　肺がん
肝硬変

BIO 206/208, HSPH
Lectures Dr. John Orav

データの利用（関係性の説明・証明）

説明変数（独立変数）

性別（男 vs 女）
身長（cm）
体重（kg）
年齢（year-old）
喫煙の有無（yes or no）
日常活動度（低, 中等度, 高度）
人種(Asian, African American, White, non-white)

アウトカム変数（従属変数）

血圧（mmHg）

生存期間

死亡

QOL

コスト

交絡変数（Confounders）

データの利用（関係性の説明・証明）

説明変数（独立変数）

- 運動
- コーヒー
- アルコール摂取量
- 野菜の摂取量
- 肉の摂取量
- 考え事
- 喫煙の有無（yes or no）
- 睡眠薬の使用
- ベッドに入った時間

アウトカム変数（従属変数）

不眠

夜間覚醒

睡眠時間

交絡変数（Confounders）

データの利用（関係性の説明・証明）

説明変数（独立変数）

野菜摂取の既往(有 vs 無)

アウトカム変数（従属変数）

不眠

交絡変数（Confounders）

- 運動
- コーヒー
- アルコール摂取量
- 肉の摂取量
- 考え事
- 喫煙の有無（yes or no）
- 睡眠薬の使用
- ベッドに入った時間

重症度

	年齢	運動	アルコール	コーヒー	考え事	肉	睡眠薬	登録日	野菜	喫煙	睡眠時間
Case1	35	0	2	0	0	1	0	18/8/8	1	1	3
Case2	31	1	3	1	1	2	1	17/11/7	0	1	2
Case3	28	0	1	0	0	1	1	18/5/18	1	0	7
Case4	76	1	2	1	0	1	1	18/7/8	0	1	4
Case5	45	1	2	1	1	3	0	18/9/10	0	1	10
Case6	18	1	1	0	1	2	0	17/12/5	1	0	5
Case7	26	0	1	0	0	1	0	17/10/8	0	0	4
Case8	42	0	1	0	0	1	0	18/2/6	0	1	2
Case9	65	0	2	0	0	1	0	18/3/4	0	1	3
Case10	90	0	1	1	0	2	0	17/9/7	1	1	6
Case11	33	1	1	1	1	2	1	18/1/14	1	0	5
Case12	51	0	3	0	1	3	1	17/12/30	0	0	8
Case13	34	1	2	0	0	1	1	18/1/5	0	0	3
Case14	51	1	2	0	0	2	0	17/3/10	0	0	2
Case15	46	0	1	0	0	1	0	18/4/2	1	0	1
Case16	17	0	3	1	0	1	1	18/6/3	0	1	4

	年齢	運動	アルコール	コーヒー	考え事	肉	睡眠薬	登録日	野菜	喫煙	睡眠時間
Case1	35	0	2	0	0	1	0	18/8/8	1	1	3
Case2	31	1	3	1	1	2	1	17/11/7	0	1	2
Case3	28	0	1	0	0	1	1	18/5/18	1	0	7
Case4	76	1	2	1	0	1	1	18/7/8	0	1	4
Case5	45	1	2	1	1	3	0	18/9/10	0	1	10
Case6	18	1	1	0	1	2	0	17/12/5	1	0	5
Case7	26	0	1	0	0	1	0	17/10/8	0	0	4
Case8	42	0	1	0	0	1	0	18/2/6	0	1	2
Case9	65	0	2	0	0	1	0	18/3/4	0	1	3
Case10	90	0	1	1	0	2	0	17/9/7	1	1	6
Case11	33	1	1	1	1	2	1	18/1/14	1	0	5
Case12	51	0	3	0	1	3	1	17/12/30	0	0	8
Case13	34	1	2	0	0	1	1	18/1/5	0	0	3
Case14	51	1	2	0	0	2	0	17/3/10	0	0	2
Case15	46	0	1	0	0	1	0	18/4/2	1	0	1
Case16	17	0	3	1	0	1	1	18/6/3	0	1	4

検定の種類

アウトカム/予測因子	一群内比較 (Paired-data)	二群間比較	三群(以上)間比較	連続変数
離散変数	McNemar's test	Fisher/Chi-square	Fisher/Chi-square	t-test/ANOVA Wilcoxon/K-W (Logistic)
順序変数	Wilcoxon Sign Rank test	Chi-square Trend Test	Chi-square (Trend Test if ordinal)	Spearman Correlation
連続変数 （正規分布）	Paired t-test	t-test	ANOVA	Pearson Corr Spearman Corr
連続変数 (非正規分布)	Wilcoxon Sign Rank	Wilcoxon Rank Sum	Kruskal-Wallis	Spearman Corr
打ち切りデータ (生存時間)	----	Log Rank	Log Rank	Cox Regression

BIO 206/208, HSPH Lectures Dr. John Orav

検定の種類

アウトカム/予測因子	一群内比較 (Paired-data)	二群間比較	三群(以上)間比較	連続変数
離散変数	McNemar's test	Fisher/Chi-square	Fisher/Chi-square	t-test/ANOVA Wilcoxon/K-W (Logistic)
順序変数	Wilcoxon Sign Rank test	Chi-square Trend Test	Chi-square (Trend Test if ordinal)	Spearman Correlation
連続変数 （正規分布）	Paired t-test	t-test	ANOVA	Pearson Corr Spearman Corr
連続変数 (非正規分布)	Wilcoxon Sign Rank	Wilcoxon Rank Sum	Kruskal-Wallis	Spearman Corr
打ち切りデータ (生存時間)	----	Log Rank	Log Rank	Cox Regression

BIO 206/208, HSPH Lectures Dr. John Orav

解析の例: Two sample t test

- 2群間の連続変数の平均は等しい(H0) vs 等しくない(HA)
- 非飲酒者と毎日飲酒者のLDLコレステロール値の平均は等しい(H0) vs 等しくない(HA)

*var_test()は2つのグループが等分散か不等分散かを判定するために、今回作った式
*sample100は各グループ50例のdata.frame

```
t.test(LDL ~ freqCat, data = sample100, var.equal = var_test(sample100))
```

```
Two Sample t-test

data:  LDL by freqCat
t = 2.9976, df = 98, p-value = 0.003448
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.668031 27.871969
sample estimates:
 mean in group 1.NoDrink mean in group 3.Everyday
                  126.67                  109.90
```

各グループ500例では

Two Sample t-test

data: LDL by freqCat
t = 7.7359, df = 1998, p-value = 1.616e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.433077 14.160923
sample estimates:
 mean in group 1.NoDrink mean in group 3.Everyday
                 126.905   115.608

*confidence intervalの幅とp-valueに注目

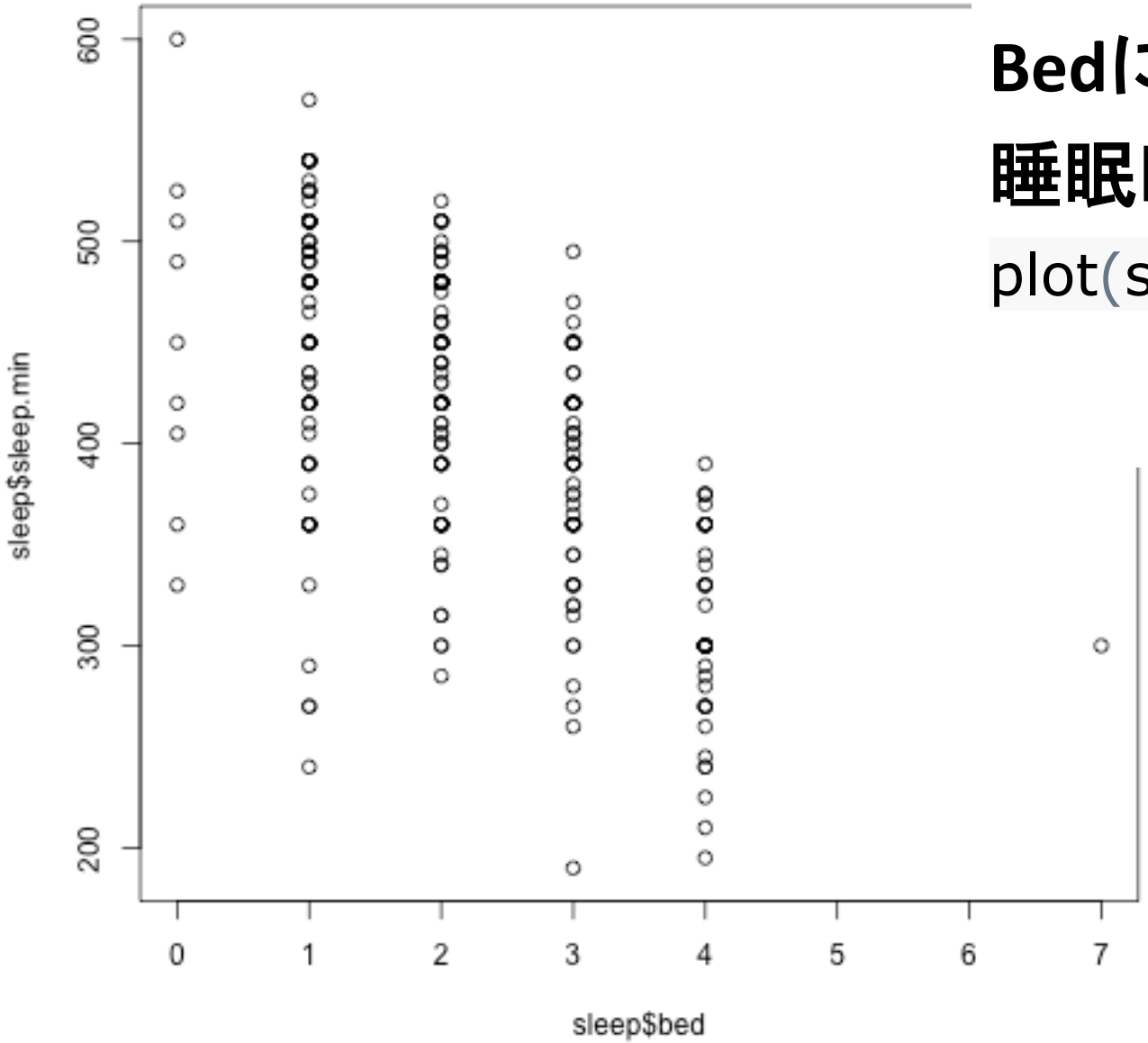
lm　線形回帰分析; 連続変数を予測するモデル

```
names(sleep)
```

```
 [1] "date"       "wk"         "rice"       "fish"       "meat"
 [6] "vegetables" "alcohol"    "sleep.pill" "bed"        "ex"
[11] "shape.bad"  "think"      "obstacle"   "hormone"    "sleep.min"
[16] "insomnia"   "noct_awake"
```

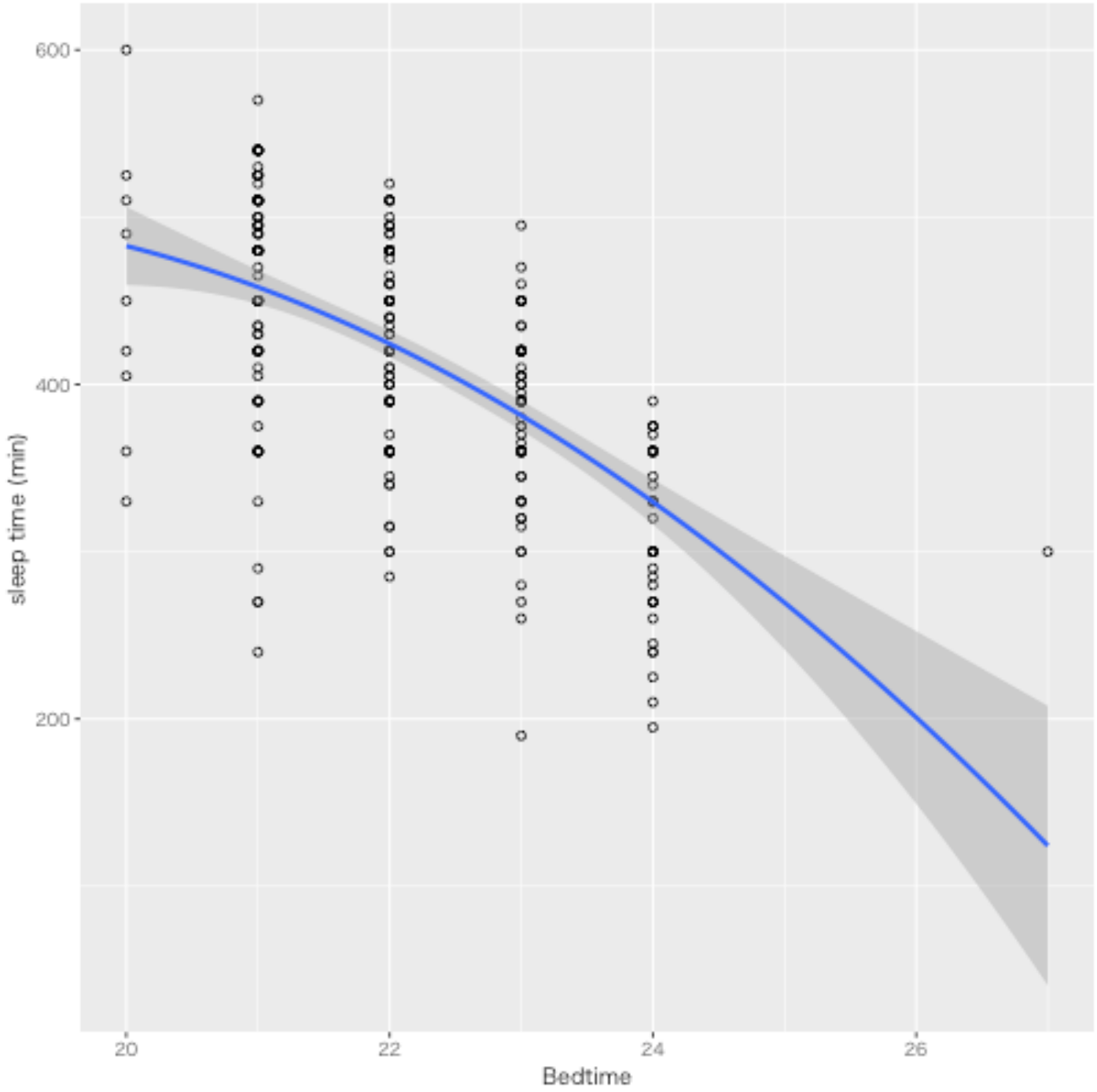
```
sleep_time_model =
  lm(sleep.min ~ . , data = dplyr::select(sleep, 2:15))

summary(sleep_time_model)
```

```
Call:
lm(formula = sleep.min ~ ., data = dplyr::select(sleep, 2:15))

Residuals:
    Min      1Q  Median      3Q     Max
-181.63  -30.77    5.28   40.31  121.71

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 500.32088   13.90234  35.988  < 2e-16 ***
wk1:weekend  -2.36511    7.60405  -0.311 0.755984
wk2:sunday   11.94443    9.78731   1.220 0.223241
rice          0.06614    5.81931   0.011 0.990940
fish        -23.91596    6.97824  -3.427 0.000692 ***
meat          1.40560    3.61801   0.389 0.697912
vegetables    7.02480    3.63499   1.933 0.054203 .
alcohol       8.18906    4.57140   1.791 0.074210 .
sleep.pill  -12.55061   13.68362  -0.917 0.359751
bed         -38.69299    2.97955 -12.986  < 2e-16 ***
exeven        5.56928    7.75173   0.718 0.473017
exmorn       -4.43720    8.55687  -0.519 0.604442
shape.bad   -16.40863   12.07859  -1.358 0.175296
think       -53.39715    9.74961  -5.477 8.96e-08 ***
obstacle     -9.93203    9.25491  -1.073 0.284033
hormoneE1     3.99091    9.47582   0.421 0.673924
hormoneO     -7.67073    9.87875  -0.776 0.438053
hormoneP    -20.03417    7.94377  -2.522 0.012170 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.39 on 310 degrees of freedom
Multiple R-squared:  0.5022,    Adjusted R-squared:  0.4749
F-statistic:  18.4 on 17 and 310 DF,  p-value: < 2.2e-16
```



Bedに入った時刻と睡眠時間との関係をPlot

```
plot(sleep$bed, sleep$sleep.min)
```




```
plot = ggplot(sleep, aes(x =
hour, y = sleep.min)) +
  geom_point(shape = 1) +
  ylab("sleep time (min)") +
  xlab("Bedtime") +
  geom_smooth(span =10)
```

線形回帰分析でLDLコレステロール値を予測する

```
Call:
lm(formula = LDL ~ ., data = ldl)

Residuals:
    Min      1Q  Median      3Q     Max
-98.205 -20.337  -3.009  17.452 191.699

Coefficients:
                                                Estimate Std. Error t value
Pr(>|t|)
(Intercept)                                     58.89094    1.35604  43.429
< 2e-16 ***
freqCat2.Sometimes                  -1.49222    0.44029  -3.389 0.000702 ***
freqCat3.Everyday                   -3.80867    0.79312  -4.802 1.58e-06 ***
drink.amount2.Medium                -5.87077    0.72928  -8.050 8.63e-16 ***
drink.amount3.Binge    -1.00691     1.12121  -0.898 0.369165
age                                             0.64048    0.01685  38.013  <
2e-16 ***
BMI                                             1.39524    0.04593  30.380  <
2e-16 ***
smoke2.Ex_Smoker                    -1.29059    0.66935  -1.928 0.053849 .
smoke3:Current_Smoker                -2.64463    0.65001  -4.069 4.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.71 on 26148 degrees of freedom
Multiple R-squared:  0.1167,    Adjusted R-squared:  0.1164
```

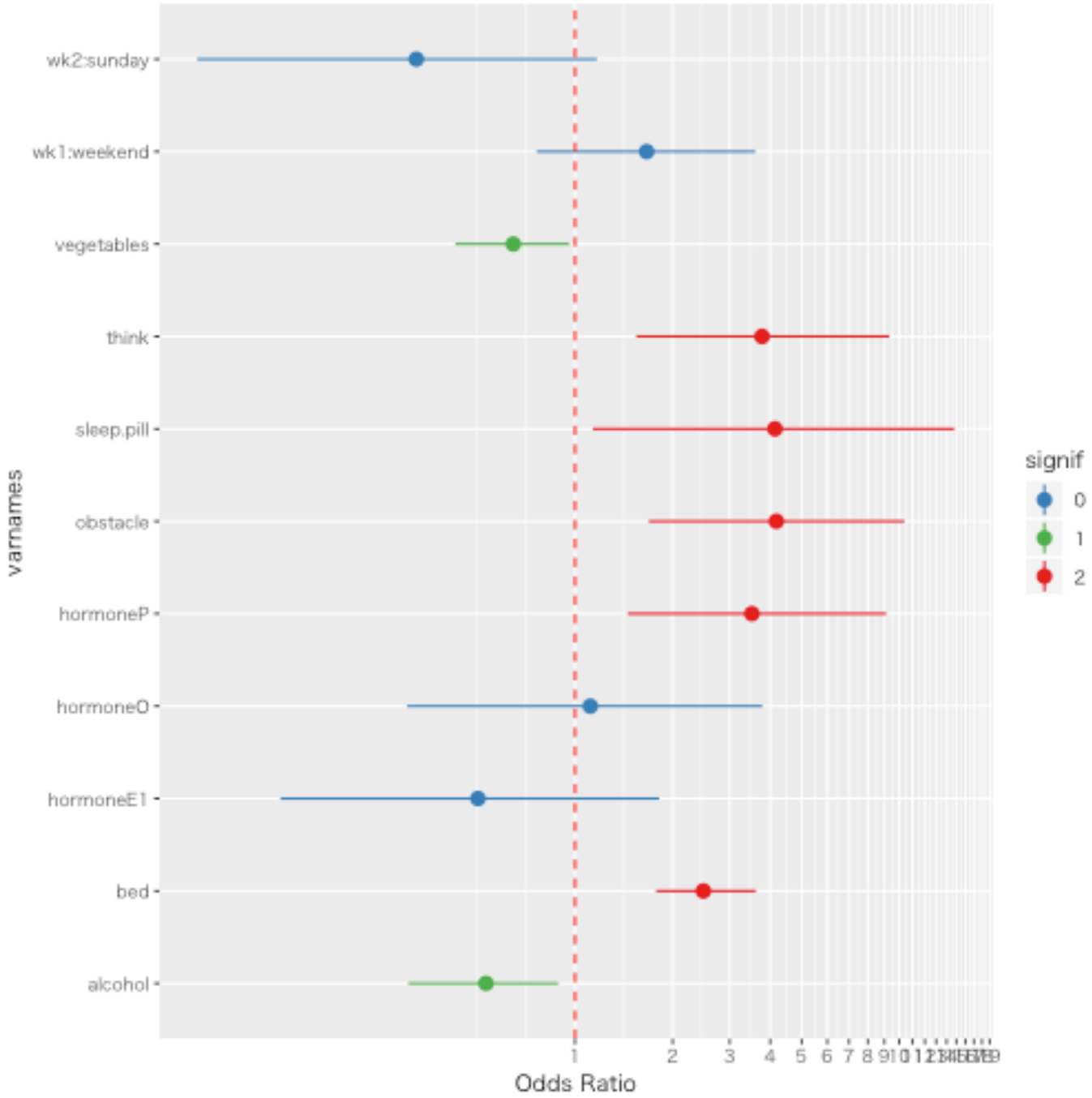
ロジスティック回帰分析で Binary outcomeを予測する

不眠(insomnia)をoutcomeにして、ロジスティック回帰分析を行う。

```
insom <- glm(insomnia ~ vegetables + meat + fish + rice + alcohol + ex + bed + think + obstacle + shape.bad + sleep.pill + wk + hormone, data = sleep, family = binomial)
```

require(MASS)

step <- stepAIC(insom, direction="both")...

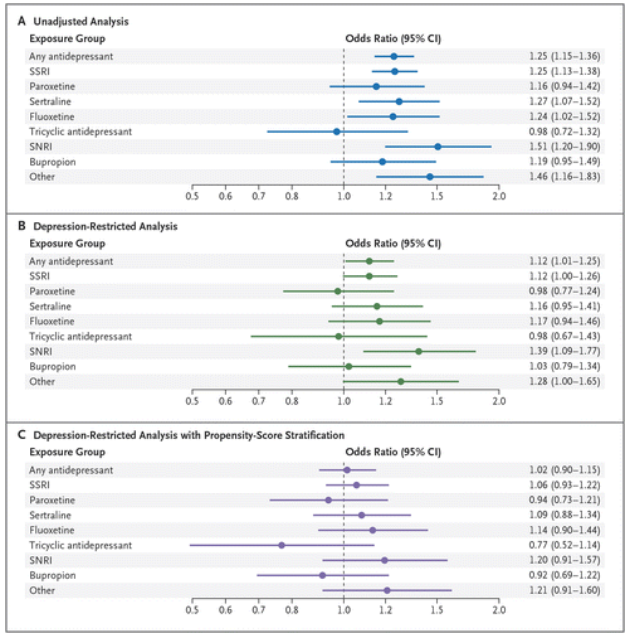


データを保存

次も同じところから始めるにはRをquitするときに.RDataを保存する。

Versionを残したい場合は、名前をつけて可視ファイルを作る。

```
save.image("sleep20181016.RData")
```



CMA, treeage, BUGS

Writing skill
Visualization

平均
全体像
論文

仮想患者
への
最適解

R, Baysian model

R
STATA
SAS
Julia
etc

析

Python, R

実際の患者

電子カルテ

QALY

保険医療解析

DPC

病態生理

情報集積

NDB, DP
JMDC
介護データ
特定健診データ
＊SQL

データ

生物統計の知識と
プログラミング言語を使った
解析システムとの対話

- 中力美和
- mcstat@oacis.org

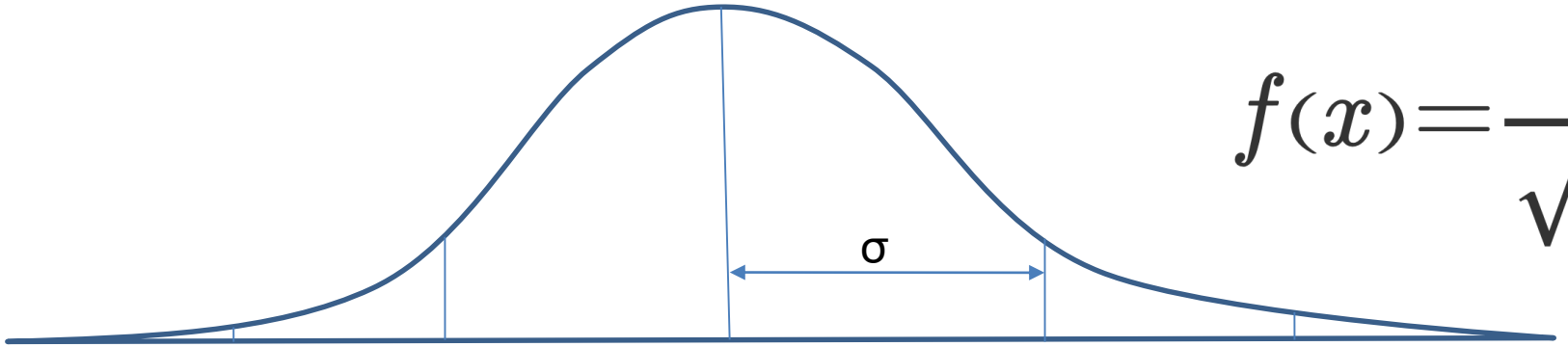
- 諸見里拓宏
- info@oacis.org

勉強会を始めていきます
興味のある方はご連絡ください

ご清聴ありがとうございました

以降は質問返答用のスライドです

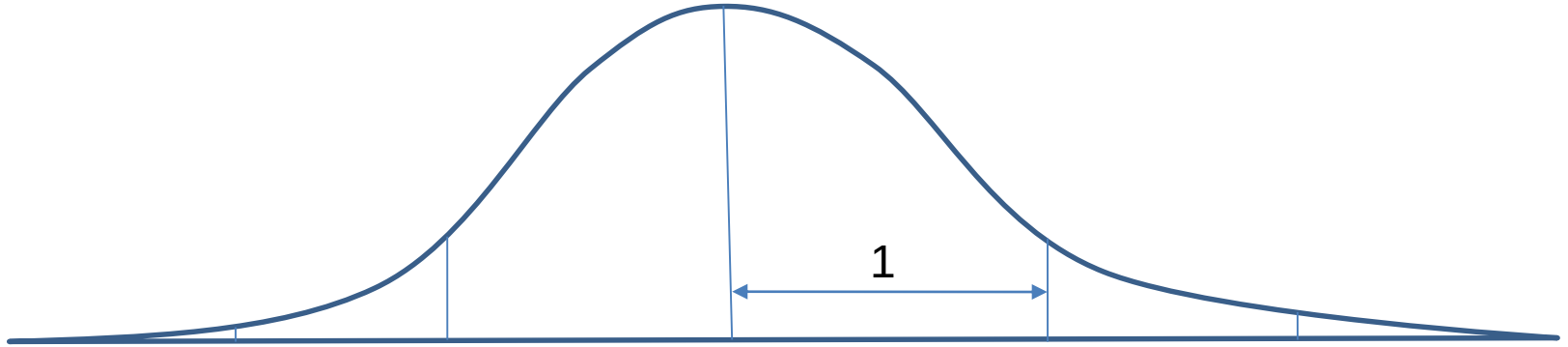
正規分布と標準正規分布



$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\times e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

μ

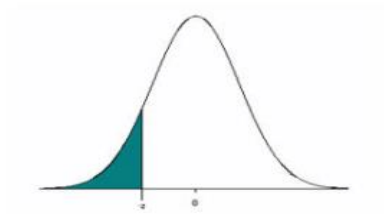
N(μ, σ)



0

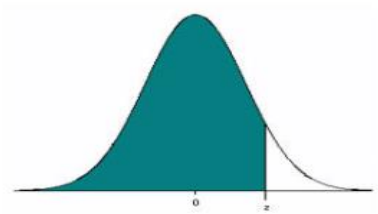
N(0 , 1)

Table of Standard Normal Probabilities for Negative Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table of Standard Normal Probabilities for Positive Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note that the probabilities given in this table represent the area to the LEFT of the z-score.
The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score

統計検定法のロジック　（統計値とは）

- サンプルデータから出た代表値(治療効果など)が1.4(1.0未満では効果なし)
- 2回目のサンプルでは0.98，３回目 1.2，４回目は1.5=毎回結論が違うかも
- 最初のサンプルで出した結果が，安定していて全体の結果に近いか？

　（つまりただの偶然の産物ではないということを証明できるか？）

サンプルの中の値
What we see in the sample

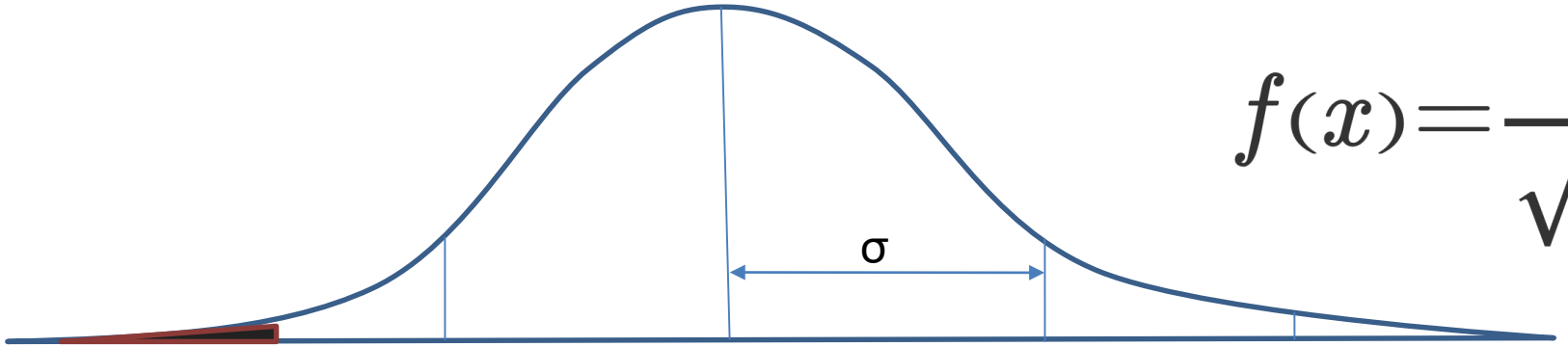
治療効果がなかった時（帰無仮説時）の値
What we expect if there is no effect

全体のデータの中に予測されるばらつき
Natural variability in the data

＞ 1.96

＝帰無仮説が棄却
＝ 偶然の産物ではないほど極端な値

正規分布と標準正規分布

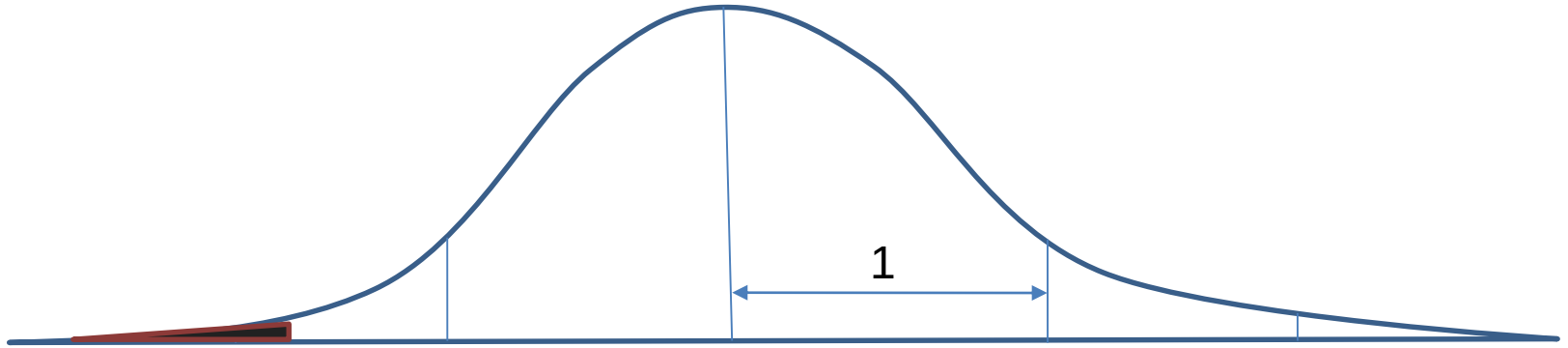


$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\times e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

x

μ

N(μ, σ)

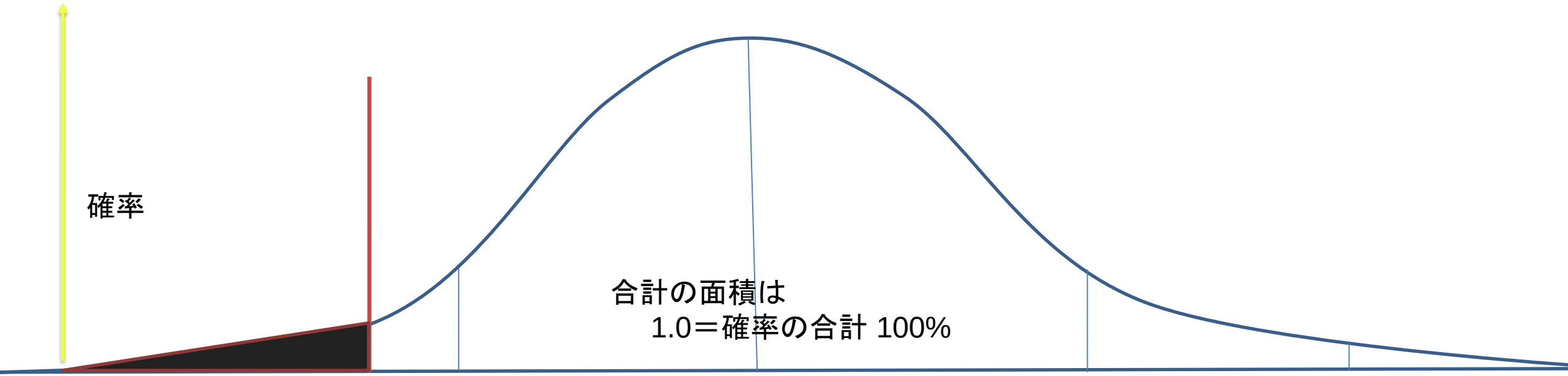


Z-score

0

N(0 , 1)

統計値から求める p 値(probability)



1.4

p値：帰無仮説下に観察された値よりも極端な値をとる確率

（関係が無いという帰無仮説下に予測される母集団下で、サンプルよりも極端な値が観察される確率）

説明変数が二項分布, アウトカム変数が連続変数であった場合の検定

1 : **帰無仮説**設定 (H_0 : μ1 = μ2)

· 関係がない時の条件を設定

2 : **データの種類**を設定

· **説明変数** : 二項変数, カテゴリー変数, 順序変数, 連続変数
· **結果変数** : 二項変数, カテゴリー変数, 順序変数, 連続変数

3 : **検定法**の選択

· 説明変数と結果変数の種類から（機序の理解）

4 : **統計値**を計算（**p-value**）

· 設定した分布に従った統計値からそれより極端な確率 (p-value) を求める

5 : 帰無仮説を**採択するかリジェクト**するか決定

6 : 結果の解釈と発表

---

1 : アトピーの既往の有無は回復までの日数と関係が無い (H_0 : 回復日数 1 = 回復日数 2)

2 : データの種類を設定

· **説明変数** : 二項変数
· **結果変数** : 連続変数, 時間のデータ（日数）

3 : 検定法の選択

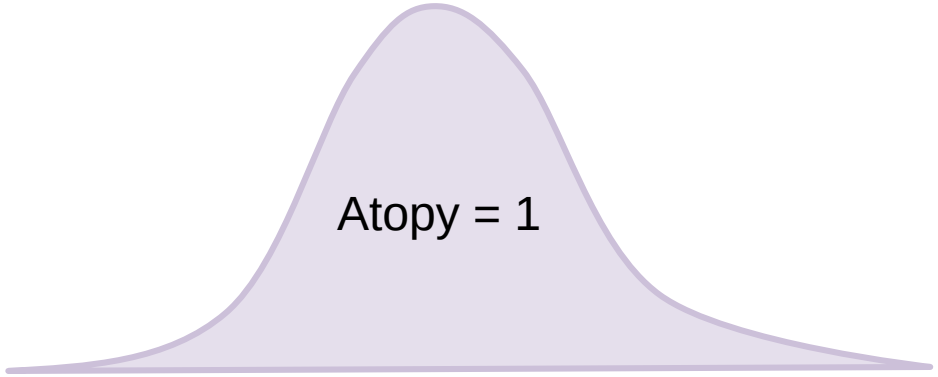
· t-test, Wilcoxon-Rank Sum, Log-Rank

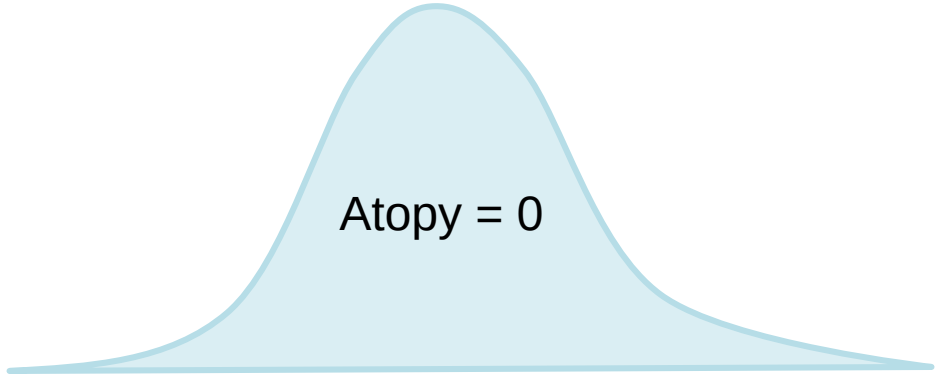
4 : 統計値を計算（p-value）

· P-value

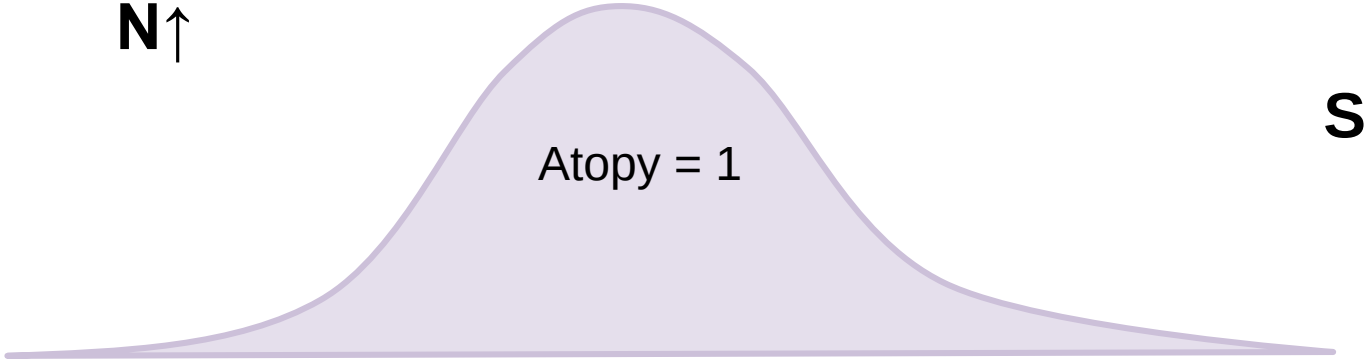
5 : 帰無仮説を**採択するかリジェクト**するか決定

6 : 結果の解釈と発表

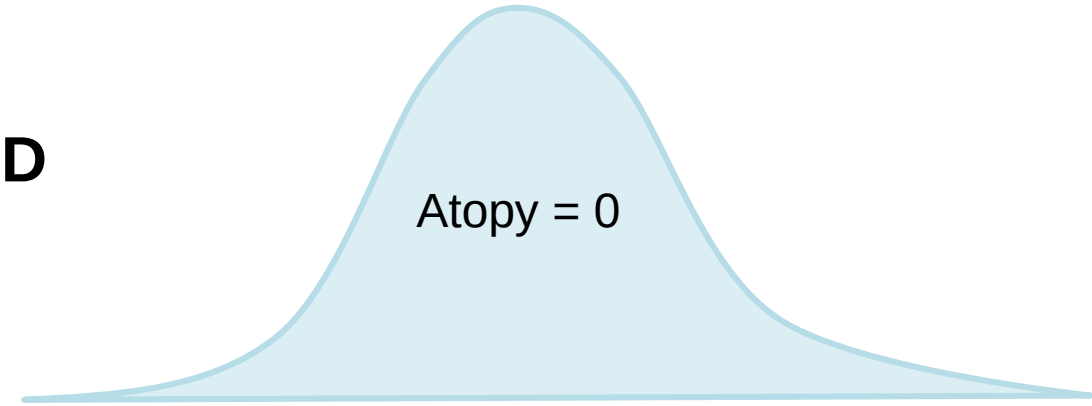




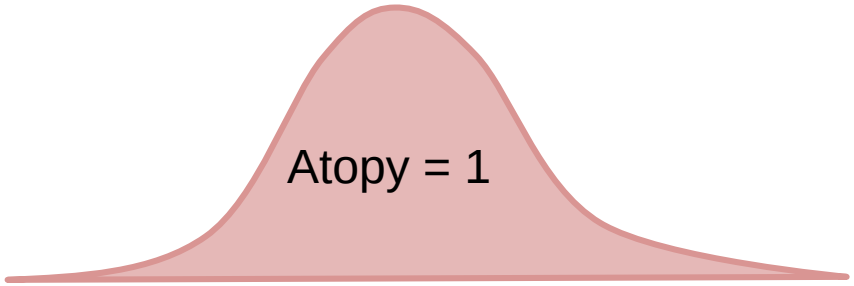
N↑



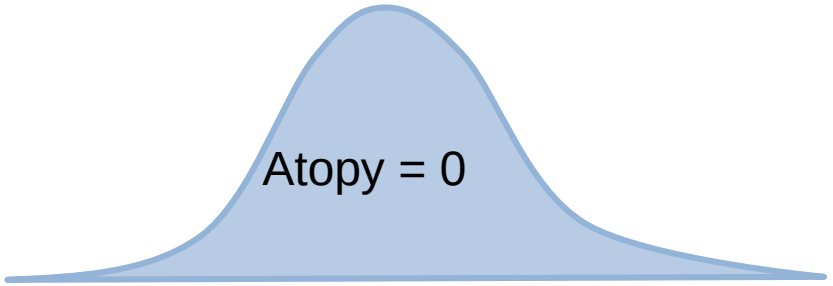
SD

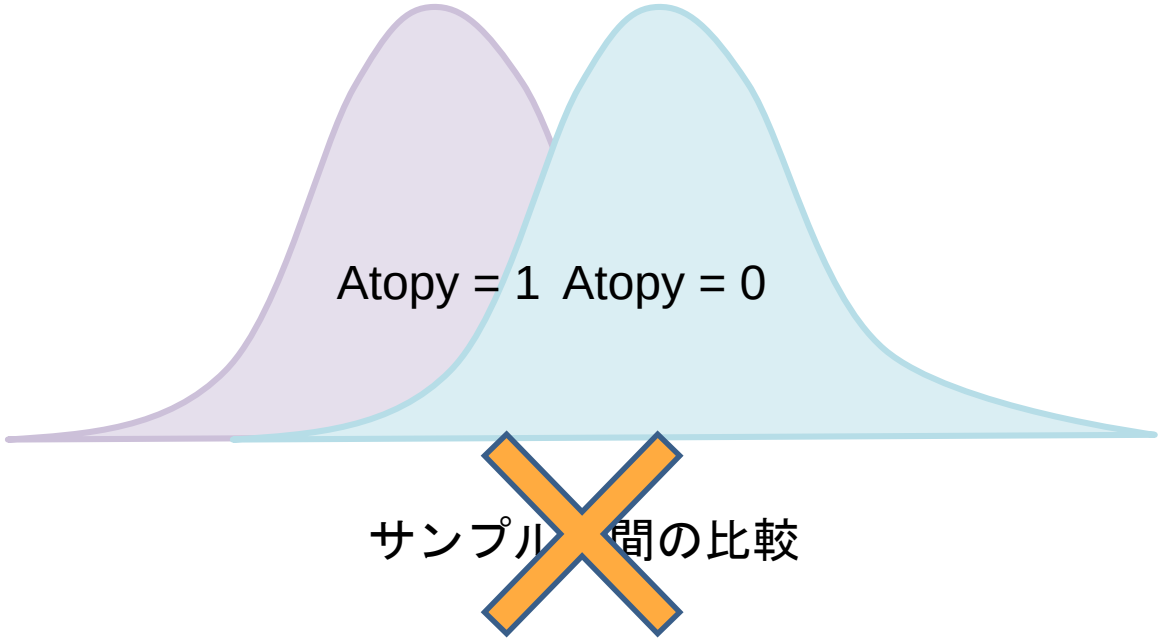


N↑

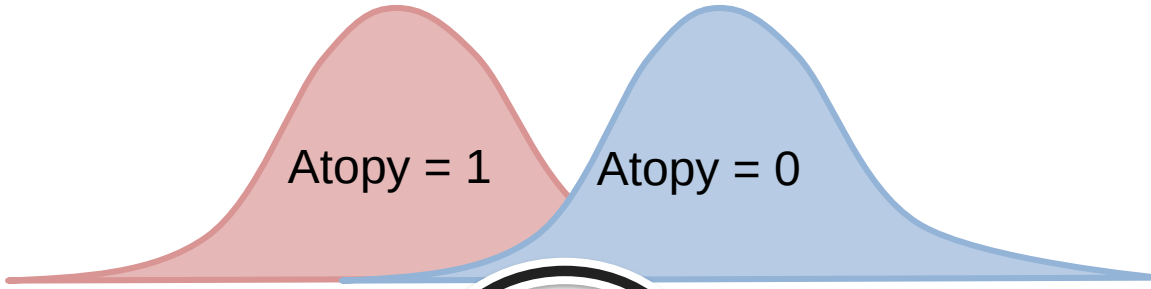


SE



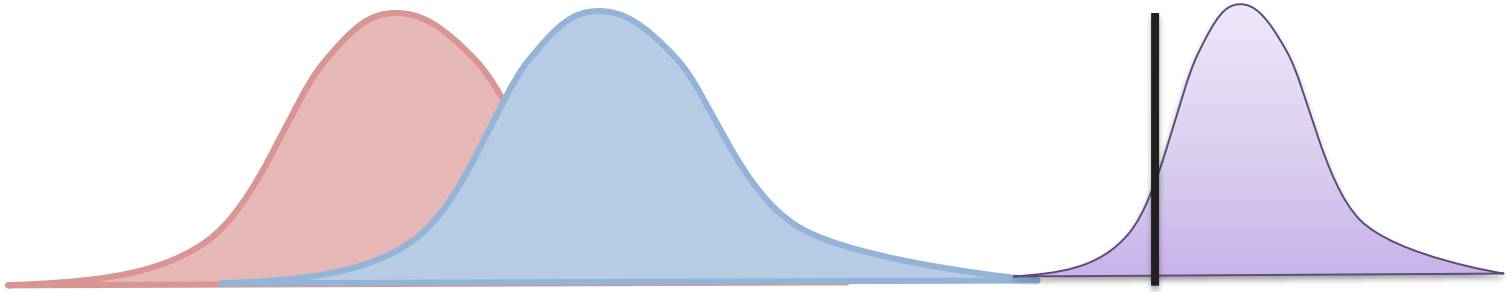


サンプル間の比較

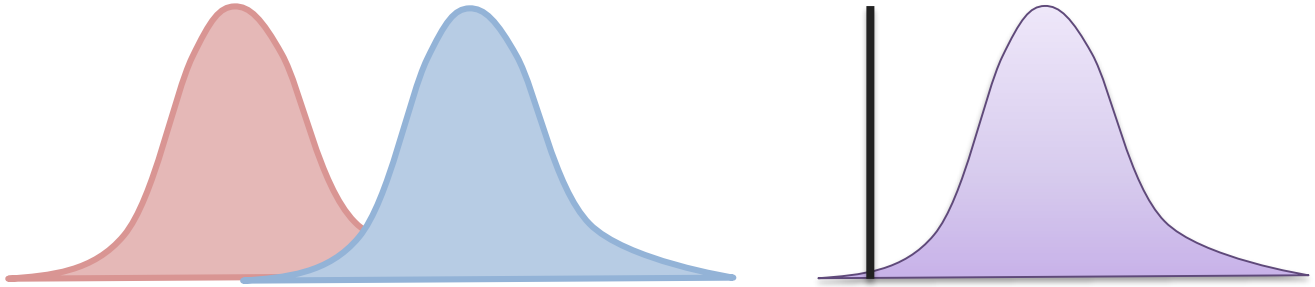


サンプルから予測される全体の集団

帰無仮説棄却できない



帰無仮説棄却



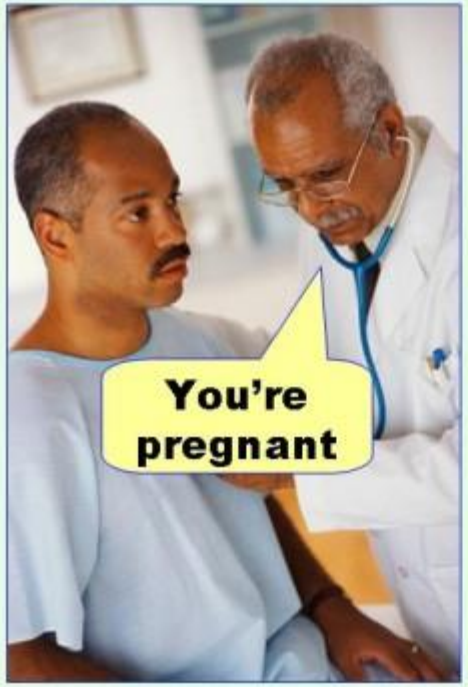
1型エラーと2型エラー

- **Type I error**: 本当は無いはずの差を見つけてしまう。
  （否定してはいけない帰無仮説を否定してしまう）
  偽陽性を生み出すエラー：αエラーと呼ばれる (5%)

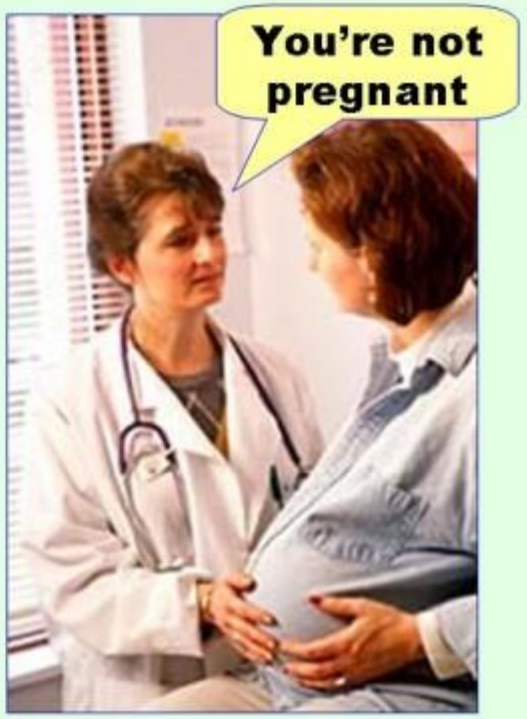
- **Type II error**: 本当は差があるのに差が無いと言ってしまう。(否定すべき帰無仮説を否定できない)
  偽陰性を生み出すエラー：βエラーと呼ばれる(20%)
  - 1-β：**Power**　：存在するはずの本当の差を発見できる能力

Errors in hypothesis testing

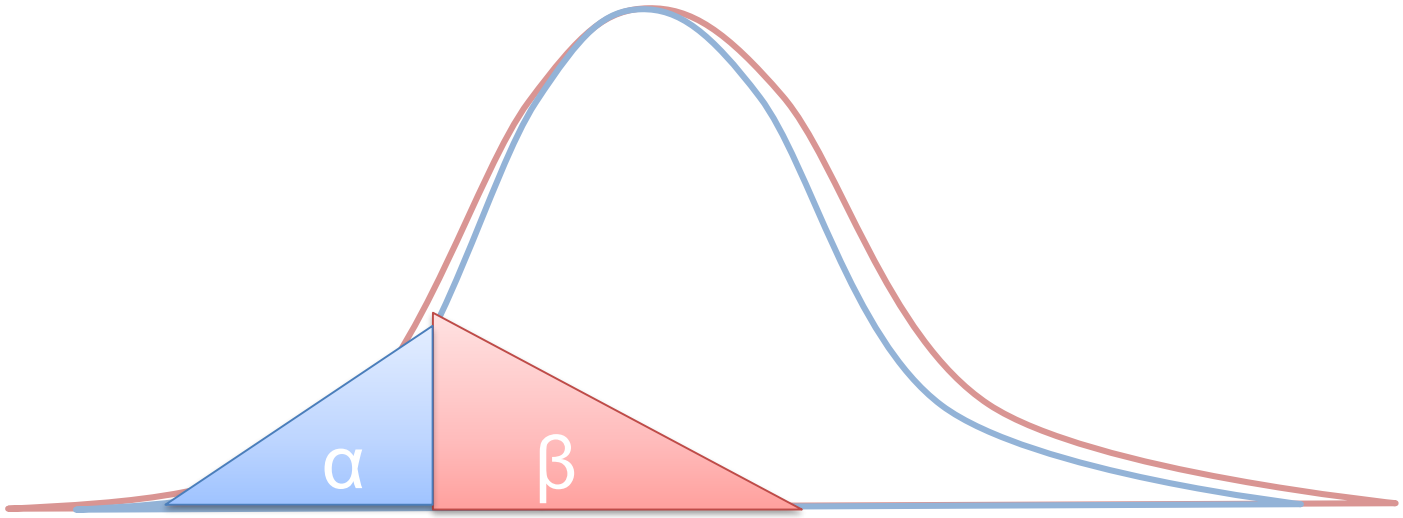
Type I error
(false positive)



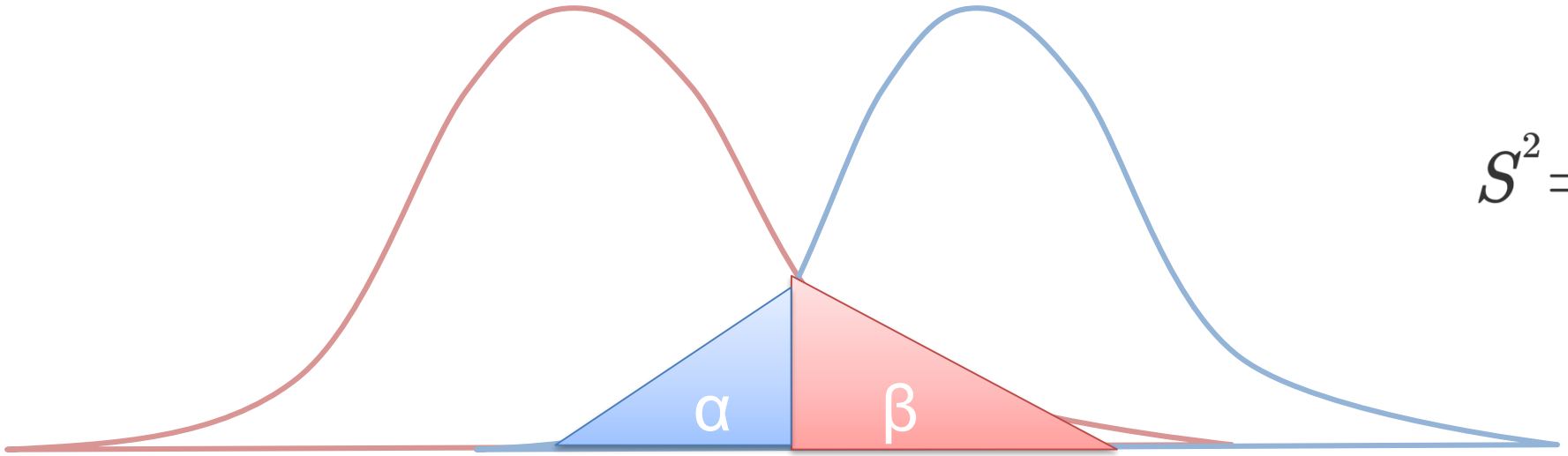
Type II error
(false negative)



サンプルサイズ
(ばらつき, αエラー, βエラー, 効果値)



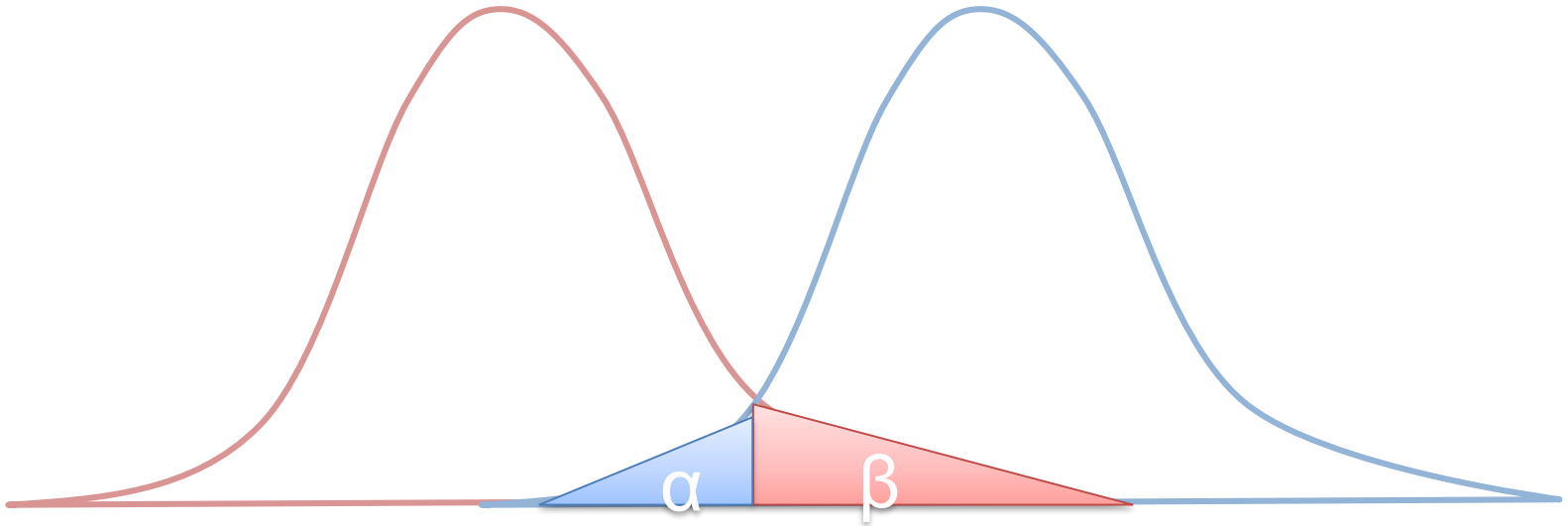
サンプルサイズ
(ばらつき, αエラー, βエラー, 効果値)



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

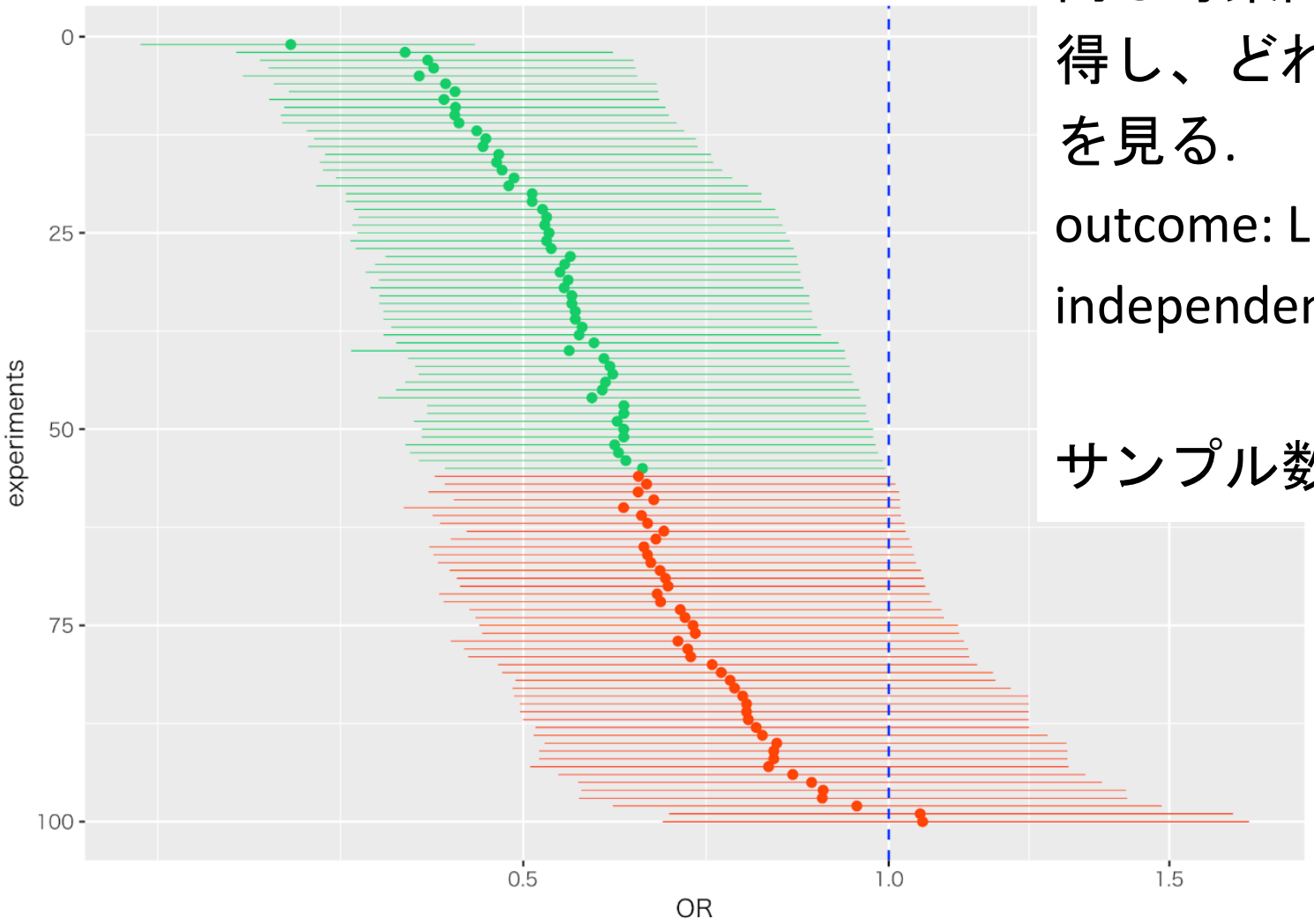
$$SD = \sqrt{S^2}$$

サンプルサイズ
(ばらつき, αエラー, βエラー, 効果値)



$$SE = sd(\bar{x}) = \sqrt{\frac{Var(x)}{n}} = \sqrt{\frac{S^2}{n}}$$

powerを理解する



同じ母集団から100回同数のサンプルを取得し、どれくらいの割合で有意差が出るかを見る.

outcome: LDL cholesterol > 140

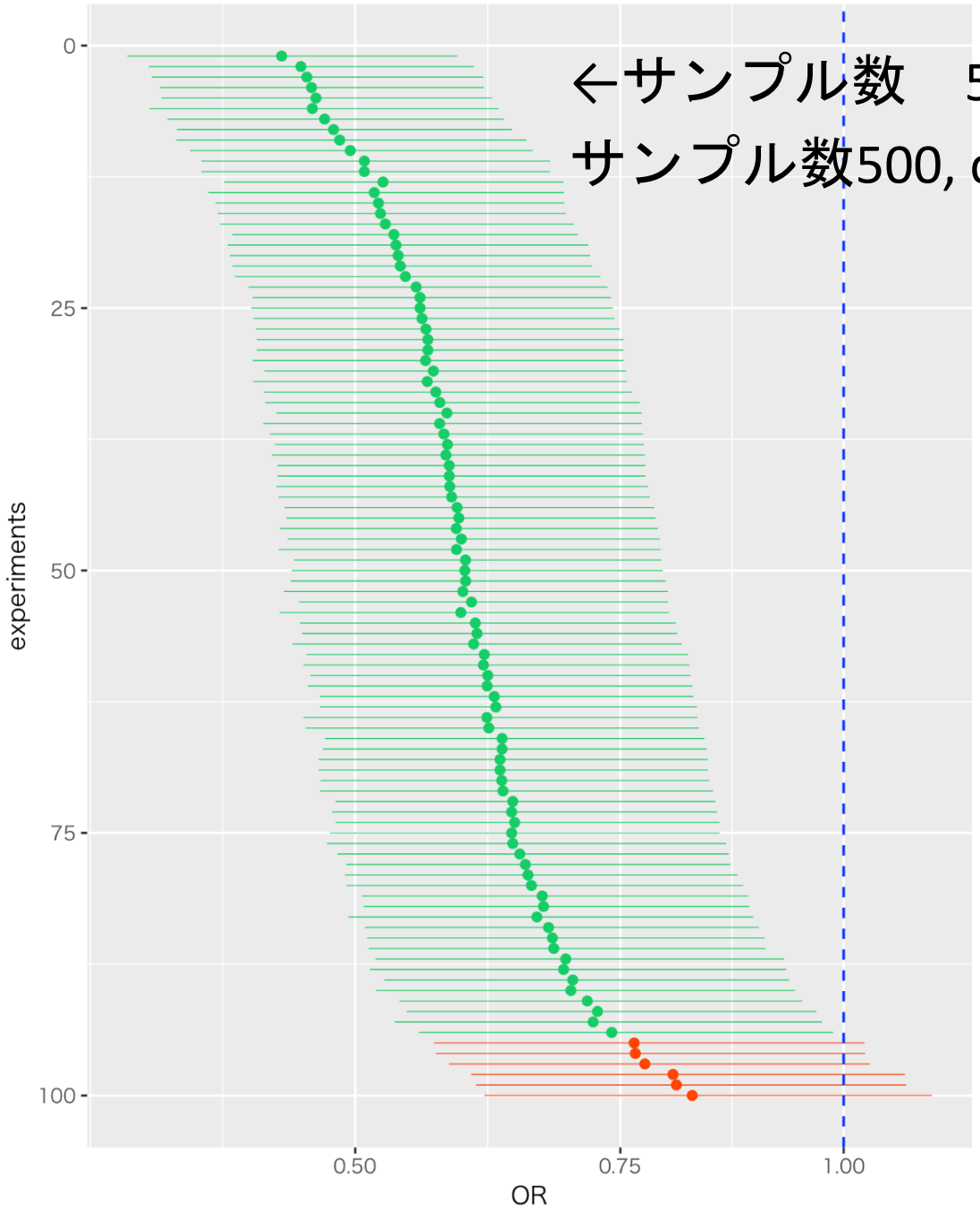
independent variable:

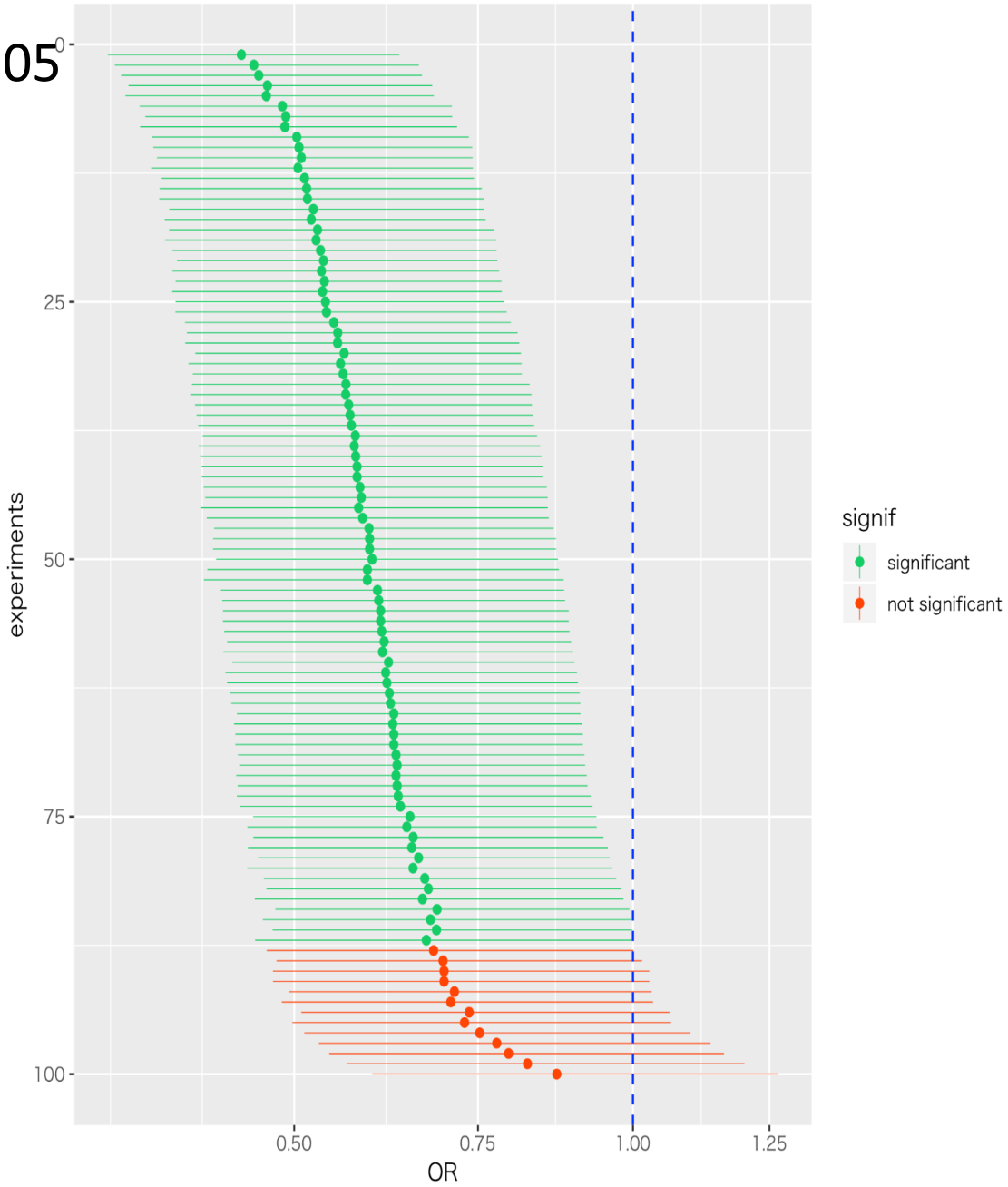
毎日飲酒 vs 非飲酒

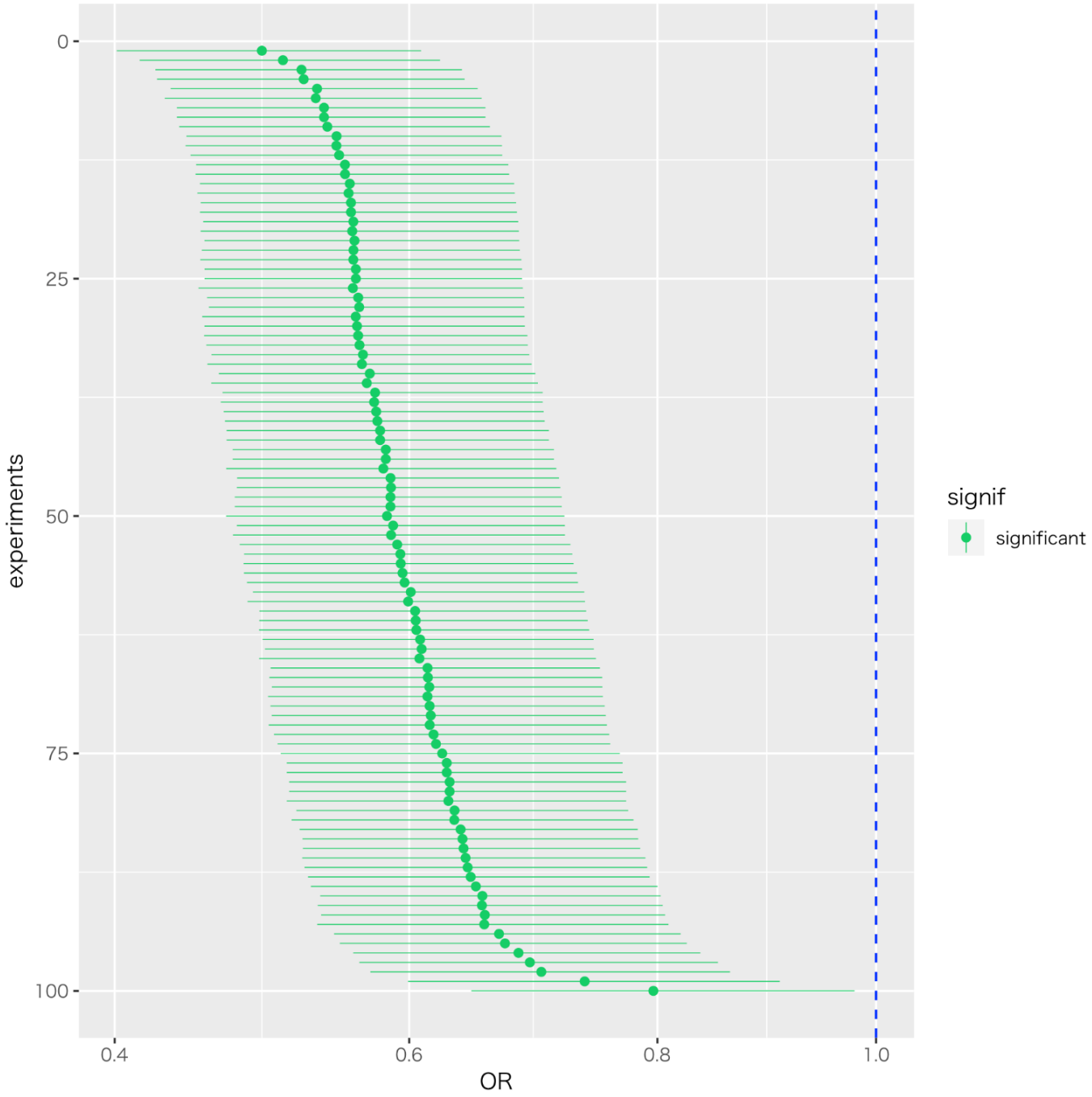
サンプル数　200, α　0.05

←サンプル数　500, α　0.05

サンプル数500, α　0.01→







サンプル数　1000, α　0.05

生物統計を”正しく使い”且つ”正しく解釈する”
ための基礎

- ①：データの分類
- ②：データが生まれてくるプロセスを知る
- ③：何を目的としてデータを集めるのか
  （データ同士の全体での関係性）
- ④：仮説検定の統計検定法の種類
- ⑤：サンプルと母集団の違い：ばらつきを知る
- ⑥：統計検定法のロジック／帰無仮説の棄却
- ⑦：αエラー(第一のエラー), βエラー(第二のエラー), p値
- ⑧：サンプルサイズ検定法