# 統計解析，その前に！

# （pythonじゃないよ）

## 実験前から統計解析が必要です

担当：原　武史（岐阜大学）

**購入回数: 2回。**
最後にこの商品を購入したのは2023/2/6です。
注文の詳細を表示する

この画像を表示

# 標準 医用画像の視覚評価法 (放射線技術学スキルUPシリーズ) 単行本 – 2020/12/19

日本放射線技術学会 (監修), 白石 順二 (著)

4.7 ★★★★☆ 5個の評価　　すべての形式と版を表示

**単行本**
**￥5,280**
獲得ポイント: 159pt prime

￥8,976 より 1 中古品
￥5,280 より 10 新品
￥10,550 より 2 コレクター商品

医用画像に関する標準的な視覚評価法をまとめた、今までにない1冊
本書は、これまでに医用画像の分野で視覚評価に用いられてきた多くの手法について解説した上で、現在では視覚評価法の主流と考えられるROC(Receiver Operating Characteristic)観察者実験、FROC(Free-response ROC)観察者実験、そして、一対比較法(Paired comparison method)の理論と実験方法を、目的別、モダリティ別、手技別に解説しています。また、これら3つの視覚評価法の観察者実験を簡便かつ正確
続きを読む

不正確な商品情報を報告します。

| 本の長さ | 言語 | 出版社 |
|---|---|---|
| 215ページ | 日本語 | オーム社 |

**新品:** ￥5,280
ポイント: 159pt (3%) 詳細はこちら
prime お届け日時指定便
無料お届け日時指定便 **7月31日 月曜日, 8:00 - 12:00の間**にお届け
詳細を見る
原 武史 - 501-1112 にお届け
**残り2点（入荷予定あり）** 在庫状況について
数量: 1
カートに入れる
今すぐ買う
支払い方法 お客様情報を保護しています
出荷元 Amazon.co.jp
販売元 Amazon.co.jp
ギフトの設定
リストに追加

# Rのインストール

- C-RANミラーサイトへ：https://ftp.yz.yamagata-u.ac.jp/pub/cran/
- Download R for Linux/macOS/Windows
- **パッケージのダウンロード先を固定する**
  - chooseCRANmirro()を実行して**Yonezawaを指定**

# シナリオ

サンプル数：２（２群）

- ２つの撮影条件でどちらの条件が診断能が高いか調べたい

- 観察者実験で明らかにしたいのでROCによる実験を行いたい

サンプルサイズ：n人（１群あたり）

- 画像を何枚用意して，観察者を何人にすればよいのかを決めたい

サンプルサイズ：n（１群あたり）

# 理解するポイント

- 平均と標準偏差：母集団と標本
- 平均と分散の推定：少ないサンプルから適切に推定したい
- 信頼区間：推定の確からしさを表現したい
- 検定の考え方：差がある　とはどういう意味か？
- サンプルサイズの決め方：読影者は何人必要か？画像は何枚必要か？
- 最近の論文で必要な事項：効果量（Effect size）とp値，検出力（検定力）

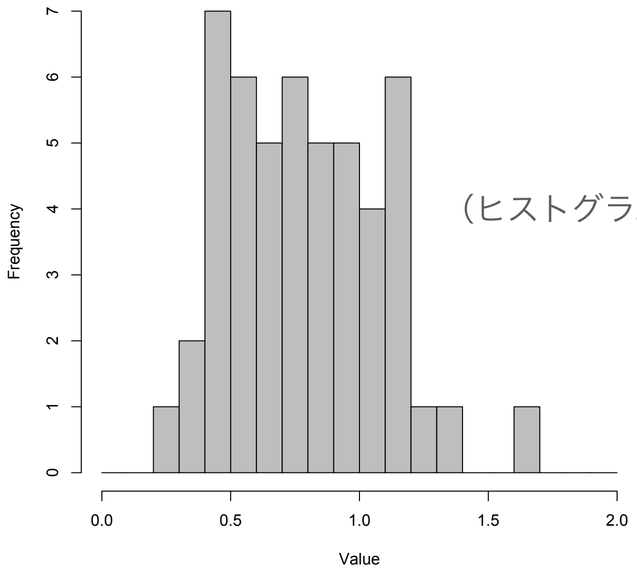
（生成した人工データ）

0.8885032 0.4908549 0.9874154 1.1454795 0.7640898 0.5963877 0.6661438 0.5879957 0.5658386 1.3026337
1.0939950 0.6117829 0.7886401 0.9142340 0.4856959 0.4089052 0.2075667 0.6893180 0.4822474 0.7464465
1.0616396 0.4526940 1.6028040 1.1852402 0.6670818 0.5160811 0.9551921 1.1191462 1.0615259 0.3574088
0.7770890 0.9546304 1.0235862 1.1025312 1.1253193 0.5781758 0.4640470 1.2571216 0.3532005 0.7350231
0.8008691 0.7409715 0.8998571 1.1304206 0.4691322 0.6239656 0.9139906 0.5979506 0.8585177 0.8121539

**基本統計量はもう大丈夫ですね？**

| | |
|---|---|
| 平均 | 0.79 |
| 標準偏差 | 0.29 |
| 最小値 | 0.2075667 |
| 第一四分位数 | 0.5806 |
| 中央値 | 0.7706 |
| 第三四分位数 | 0.7924 |
| 最大値 | 1.6028040 |

（基本統計量）




（ヒストグラム）

**数値を示しても人は理解できない**

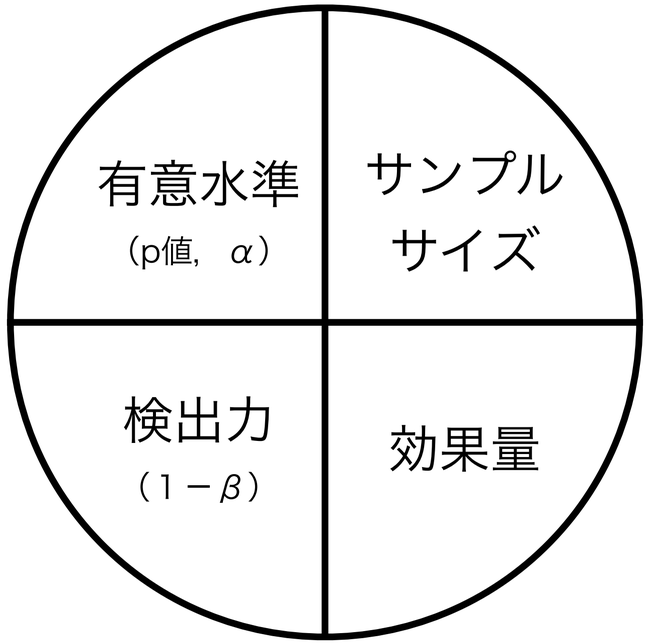
**解釈しやすい代表的な数値で**
**群の特徴を表す**

図２　人工的なデータから算出した基本統計量とヒストグラム

標準　医用画像の視覚評価法，p.106

**「何も特別なことは起こっていない」**
（帰**無**仮説：差があるとはいえ**な**い）
という事象の閾値

**観測するデータの数**
画像数，観察者数
大きくすると差が出やすくなるが
コストがかかる
また，本来なかった差があたかも
差があったようにでることになる




**本来ある差が正しく検出される確率**
サンプルサイズ，効果量，有意水準で決定
（このことをPower Analysisという）

pが小さいと検出力も低下

検出力には0.8を使う場合が多い
||
**有意差があるときに**
**80%の確率で正しく有意差ありといえる**

**２つの群の間の差や値の関連の強さ**

効果小：0.2
効果中：0.5
効果大：0.8

$$\text{Cohen's d} = \frac{|m_1 - m_2|}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_1^2}{2}}$$

図１　検定において考えるべき４つの要素

標準　医用画像の視覚評価法，p.105

表１　ある観察者実験から得られたROC曲線下面積

| 観察者群１ | ROC曲線下面積 |
|---|---|
| 観察者１ | 0.736 |
| 観察者２ | 0.676 |
| 観察者３ | 0.521 |
| 観察者４ | 0.754 |
| 観察者５ | 0.748 |
| 観察者６ | 0.946 |
| 観察者７ | 0.765 |
| 観察者８ | 0.551 |
| 観察者９ | 0.715 |

データ入力
基本統計量算出
標本標準偏差算出
95%信頼区間算出
算出された信頼区間

```
> auc1<-c(0.736, 0.676,0.521,0.754,0.748,0.946,0.765,0.551,0.715)
> summary(auc1)
   Min 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5210  0.6760  0.7360  0.7124  0.7540  0.9460
> sd(auc1)
[1] 0.1249531
> t.test(auc1)

    One Sample t-test

data:  auc1
t = 17.105, df = 8, p-value = 1.387e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.6163970 0.8084919
sample estimates:
mean of x
0.7124444
```

図３　Rによる基本統計量の算出と母平均の95%信頼区間の推定

９人は母集団（読影者の世界全体）から抽出された標本
したがって，この９人のAUCの平均は標本平均
**標本から母平均を推定できないか？**
**（ここが重要！）**
**||**
**母平均の信頼区間**
（ただし母分散が未知の場合）

t分布の表：n=9, s=0.025

$$\overline{x} - 2.262\sqrt{\frac{s^2}{n}} \leqq \mu \leqq \overline{x} + 2.262\sqrt{\frac{s^2}{n}}$$

$$0.6163970 \leqq \mu \leqq 0.8084919$$

不偏分散：$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

標準　医用画像の視覚評価法，p.113

| n | p | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 |
|---|---|---|---|---|---|---|
| | 1 | 3.07768354 | 6.31375151 | 12.7062047 | 31.820516 | 63.6567412 |
| | 2 | 1.88561808 | 2.91998558 | 4.30265273 | 6.96455673 | 9.9248432 |
| | 3 | 1.63774435 | 2.35336343 | 3.18244631 | 4.54070286 | 5.84090931 |
| | 4 | 1.53320627 | 2.13184679 | 2.77644511 | 3.74694739 | 4.60409487 |
| | 5 | 1.47588405 | 2.01504837 | 2.57058184 | 3.36493 | 4.03214298 |
| | 6 | 1.43975575 | 1.94318028 | 2.44691185 | 3.1426684 | 3.70742802 |
| | 7 | 1.41492393 | 1.89457861 | 2.36462425 | 2.99795157 | 3.4994833 |
| | 8 | 1.39681531 | 1.85954804 | 2.30600414 | 2.89645945 | 3.35538733 |
| | 9 | 1.38302874 | 1.8331[illegible]93 | 2.26215716 | 2.82143793 | 3.24983554 |
| | 10 | 1.37218364 | 1.81246112 | 2.22813885 | 2.76376946 | 3.16927267 |
| | 15 | 1.34060561 | 1.75305036 | 2.13144955 | 2.6024803 | 2.94671288 |
| | 20 | 1.32534071 | 1.72471824 | 2.08596345 | 2.527977 | 2.84533971 |
| | 30 | 1.31041503 | 1.69726089 | 2.04227246 | 2.45726154 | 2.74999565 |
| | 50 | 1.29871369 | 1.67590503 | 2.00855911 | 2.40327192 | 2.67779327 |
| | 70 | 1.2937629 | 1.66691448 | 1.99443711 | 2.38080748 | 2.64790462 |
| | 100 | 1.29007476 | 1.66023433 | 1.98397152 | 2.36421737 | 2.62589052 |
| | 200 | 1.28579879 | 1.6525081 | 1.97189622 | 2.34513708 | 2.60063444 |
| | 1000 | 1.28239872 | 1.64637882 | 1.96233908 | 2.33008267 | 2.5807547 |
| | 10000 | 1.28163623 | 1.64500602 | 1.96020124 | 2.32672084 | 2.57632105 |

t分布の表：n=9, s=0.025

データ入力
```
> auc1<-c(0.736, 0.676,0.521,0.754,0.748,0.946,0.765,0.551,0.715)
```
基本統計量算出
```
> summary(auc1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.5210  0.6760  0.7360  0.7124  0.7540  0.9460 
```
標本標準偏差算出
```
> sd(auc1)
[1] 0.1249531
```
95%信頼区間算出
```
> t.test(auc1)

   One Sample t-test

data:  auc1
t = 17.105, df = 8, p-value = 1.387e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```
算出された信頼区間
```
 0.6163970 0.8084919
sample estimates:
mean of x 
0.7124444 
```

$$0.6163970 \leqq \mu \leqq 0.8084919$$

図３　Rによる基本統計量の算出と母平均の95%信頼区間の推定

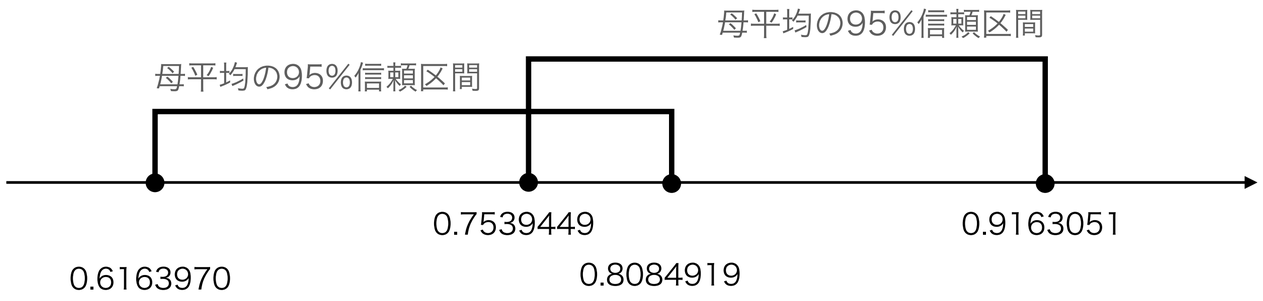
```
> auc2<-c(0.801, 0.845, 0.726, 0.997, 0.829, 0.908, 0.879, 0.696)
> summary(auc2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.6960  0.7823  0.8370  0.8351  0.8862  0.9970 
> t.test(auc2)

   One Sample t-test

data:  auc2
t = 24.326, df = 7, p-value = 5.052e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.7539449 0.9163051
sample estimates:
mean of x 
 0.835125 
```

$$0.7539449 \leqq \mu \leqq 0.9163051$$

図４　観察条件２の基本統計量の算出と母平均の95%信頼区間の推定




母平均の９５％信頼区間に重なりがある

それぞれの群の**母平均**に違いがあるとはいえない

||

**２つの群の平均は異なるとはいえない**

**・同じとは言っていない点に注意！**

**・有意差がある場合もある**

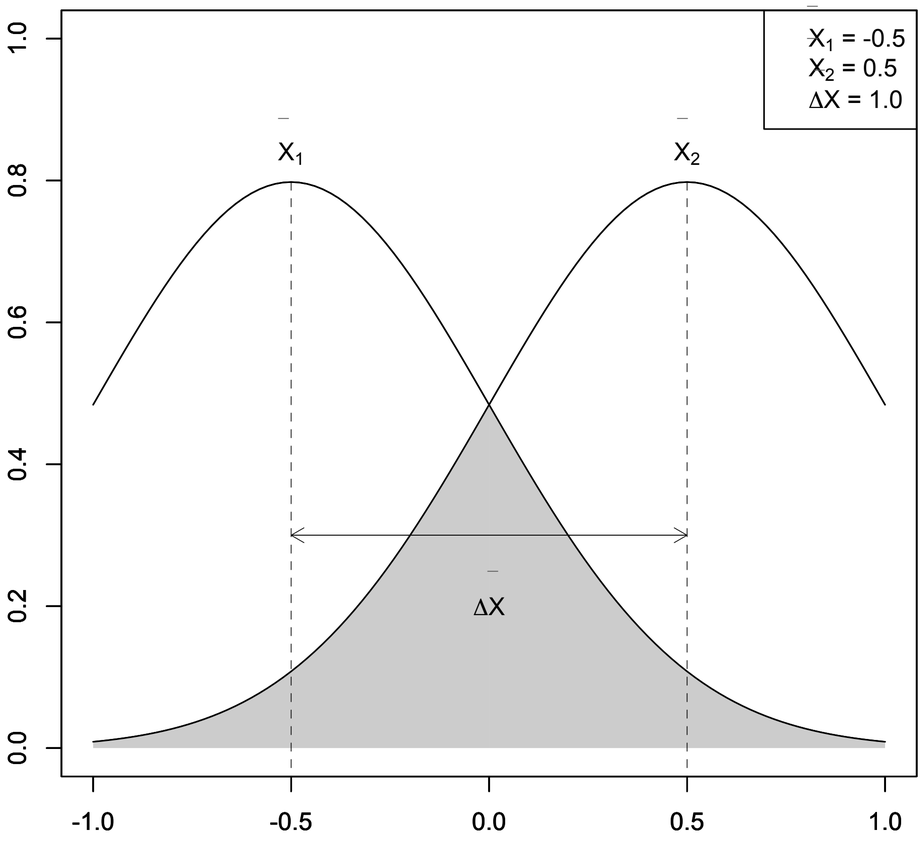
標準　医用画像の視覚評価法，p.113-4

どんなときに差が出やすいか？

**効果量**を考えましょう

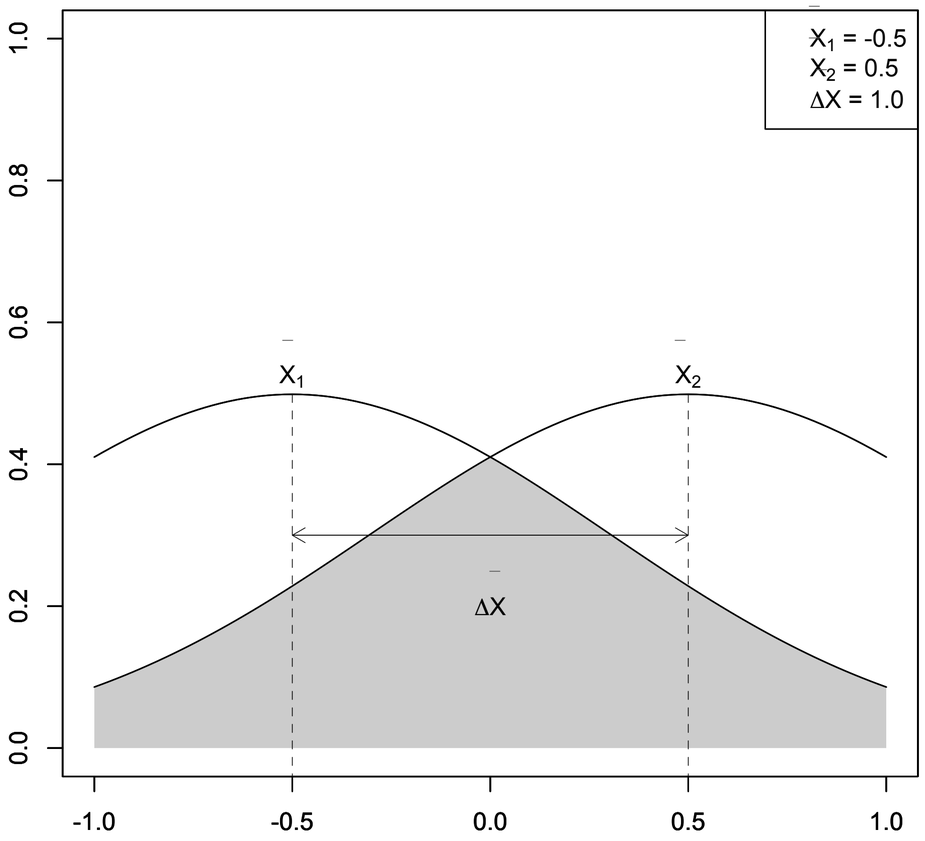
# どちらが差が大きい？

効果量： $\dfrac{|0.5-(-0.5)|}{0.5} = \dfrac{1}{0.5} = 2.0$




(a) 標準偏差 = 0.5

効果量： $\dfrac{|0.5-(-0.5)|}{0.8} = \dfrac{1}{0.5} = 1.25$

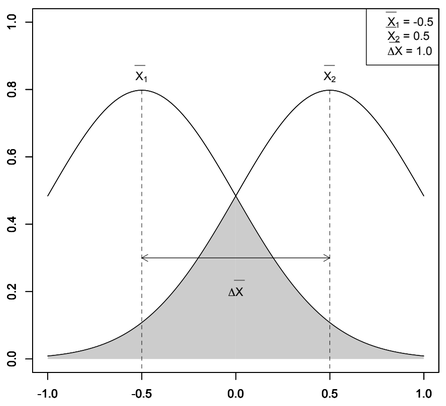



(b) 標準偏差 = 0.8

図５　差は同じで標準偏差が違う場合の重なり

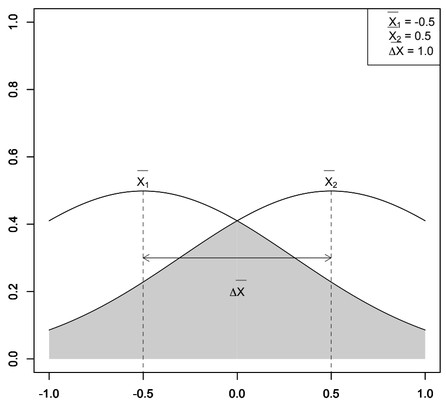
標準　医用画像の視覚評価法，p.115

効果量： $\frac{|0.5-(-0.5)|}{0.5} = \frac{1}{0.5} = 2.0$




(a) 標準偏差 = 0.5

効果量： $\frac{|0.5-(-0.5)|}{0.8} = \frac{1}{0.5} = 1.25$




(b) 標準偏差 = 0.8

図５　差は同じで標準偏差が違う場合の重なり

差が同じでも標準偏差（散らばり具合）が小さい
（すなわち値が集まっている）と
効果量が大きくなる
**＝差が出やすくなる**

差が小さい場合には標準偏差が大きいと
効果量が小さくなる
**＝差が出にくくなる**

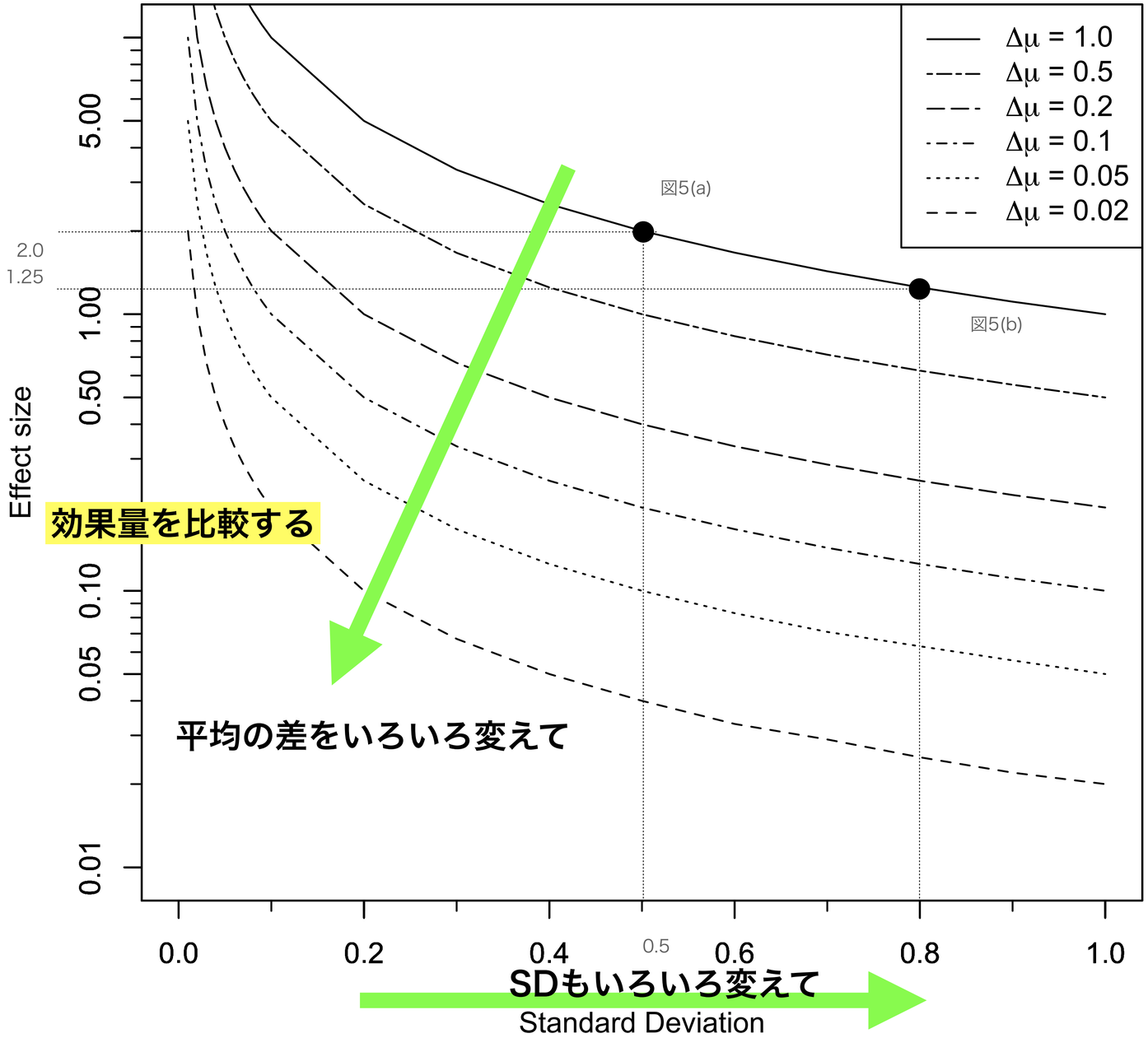



図６　分布の差と分布の標準偏差から定まる効果量

標準　医用画像の視覚評価法，p.115

# 検出力と効果量

# **サンプルサイズの設計**

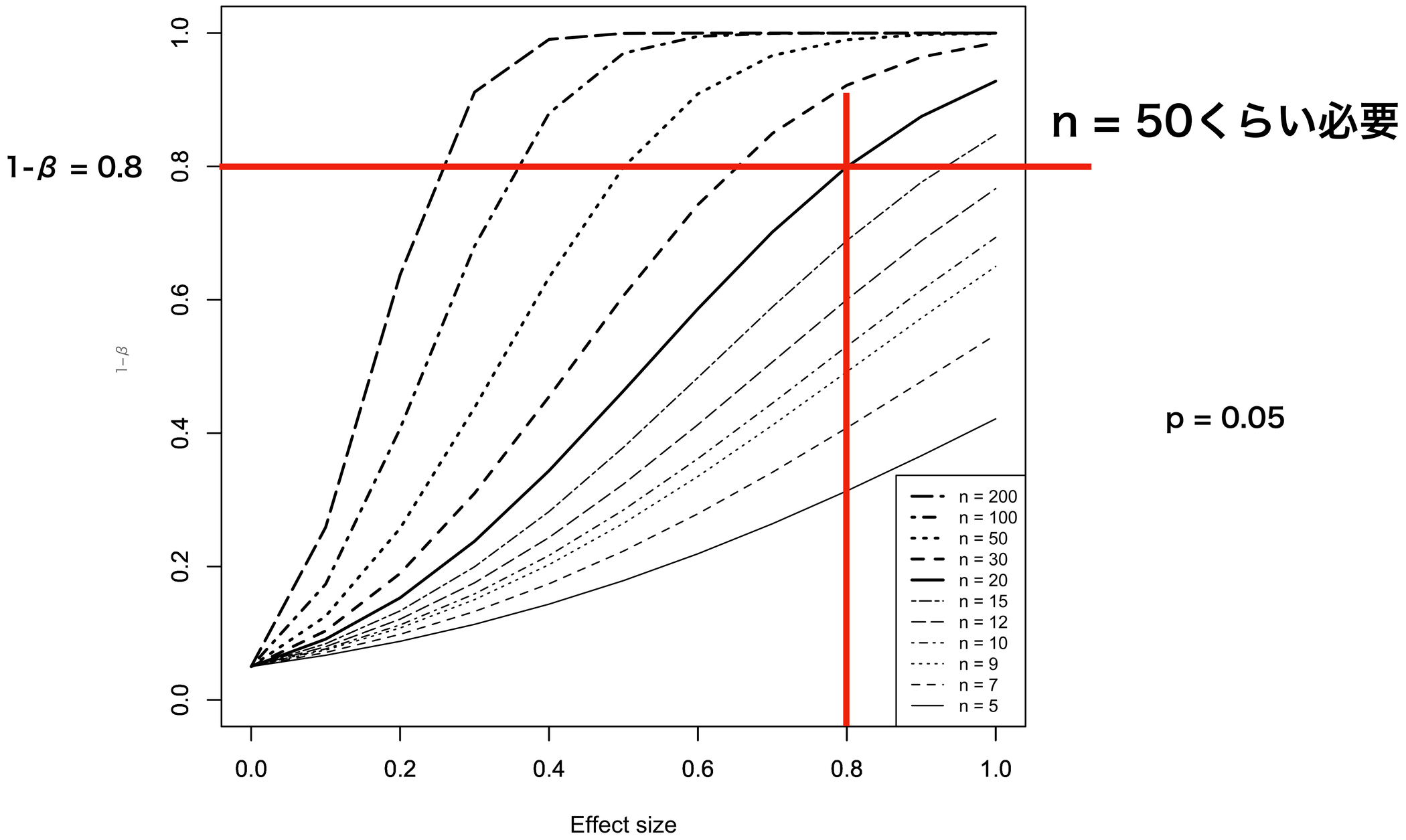


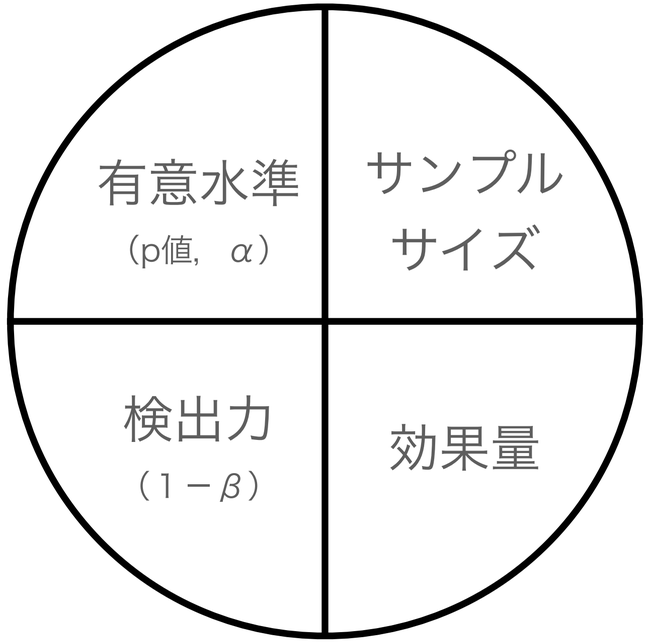

図７　サンプルサイズごとの効果量と検出力の関係

標準　医用画像の視覚評価法，p.117

# 改めて

**「何も特別なことは起こっていない」**
（帰**無**仮説：差があるとはいえ**な**い）
という事象の閾値

**観測するデータの数**
画像数，観察者数
大きくすると差が出やすくなるが
コストがかかる
また，本来なかった差があたかも
差があったようにでることになる




**本来ある差が正しく検出される確率**
サンプルサイズ，効果量，有意水準で決定
（このことをPower Analysisという）

pが小さいと検出力も低下

検出力には0.8を使う場合が多い
||
**有意差があるときに**
**80%の確率で正しく有意差ありといえる**

**２つの群の間の差や値の関連の強さ**

効果小：0.2
効果中：0.5
効果大：0.8

$$\text{Cohen's D} = \frac{|m_1 - m_2|}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_1^2}{2}}$$

図１　検定において考えるべき４つの要素

標準　医用画像の視覚評価法，p.105

# 事前分析：A priori　サンプルサイズの設計

**「何も特別なことは起こっていない」**
（帰**無**仮説：差があるとはいえ**な**い）
という事象の閾値

**p = 0.05**

**本来ある差が正しく検出される確率**
サンプルサイズ，効果量，有意水準で決定
（このことをPower Analysisという）

pが小さいと検出力も低下

検出力には0.8を使う場合が多い
||
**有意差があるときに**
**80%の確率で正しく有意差ありといえる**

**1-β = 0.8**

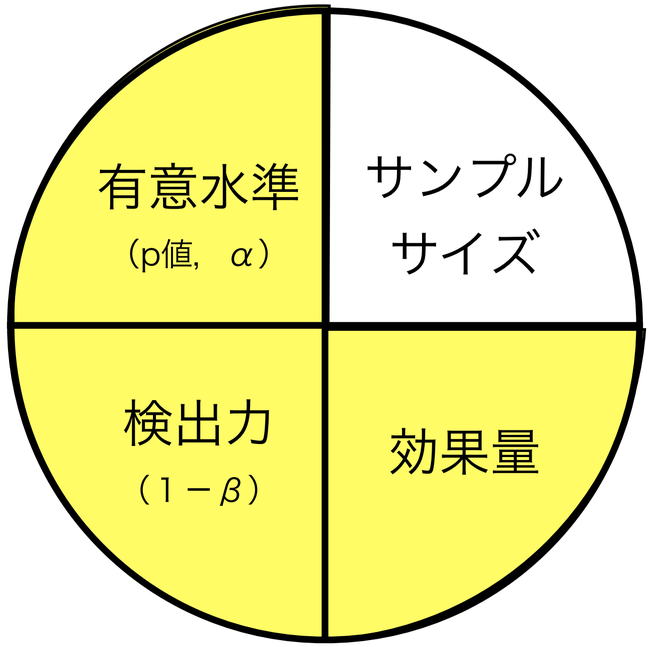



図１　検定において考えるべき４つの要素

**観測するデータの数**
画像数，観察者数
大きくすると差が出やすくなるが
コストがかかる
また，本来なかった差があたかも
差があったようにでることになる

**n = ?**

**２つの群の間の差や値の関連の強さ**

効果小：0.2
効果中：0.5
効果大：0.8

$$\text{Cohen's D} = \frac{|m_1 - m_2|}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_1^2}{2}}$$

**d = 0.5**

標準　医用画像の視覚評価法，p.105

**（問題１）**
AとB２つの群の値の平均を比較．A群の平均はB群よりも**高い**か？
効果量を0.5，1-βを0.8，有意水準0.05としてサンプルサイズを設計しなさい．

```
install.packages("pwr")
library(pwr)
pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="greater")
```

```
> library(pwr)
> pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="greater")

     Two-sample t test power calculation

              n = 50.1508
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = greater

NOTE: n is number in *each* group
```

**n = 51**

標準　医用画像の視覚評価法，p.131拡張

**（問題２）**
AとB２つの群の値の平均を比較．A群の平均はB群よりも**低い**か？
効果量を0.5，1-βを0.8，有意水準0.05としてサンプルサイズを設計しなさい．

```
install.packages("pwr")
library(pwr)
pwr.t.test(d=-0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="less")
```

```
> pwr.t.test(d=-0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="less")

     Two-sample t test power calculation

              n = 50.1508
              d = -0.5
      sig.level = 0.05
          power = 0.8
    alternative = less

NOTE: n is number in *each* group
```

**n = 51**

標準　医用画像の視覚評価法，p.131拡張

**（問題３）**
ＡとＢ２つの群の値の平均を比較．Ａ群の平均はB群と**異なる**か？
効果量を0.5，1-βを0.8，有意水準0.05としてサンプルサイズを設計しなさい．

```
install.packages("pwr")
library(pwr)
pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")
```

```
> pwr.t.test(d=0.5, sig.level=0.05, power=0.80, type="two.sample", alternative="two.sided")

     Two-sample t test power calculation

              n = 63.76561
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

**両側検定になるとサンプル数は余分に必要になる**

**n = 64**

観察者実験では基本両側検定で行う

標準　医用画像の視覚評価法，p.131拡張

**（問題４）**

ＡとＢ２つの群の値の平均を比較．Ａ群の平均はB群の平均と**異なるか**？

予備実験では

　A群の平均と標準偏差はそれぞれ2350.2と258

　B群の平均と標準偏差はそれぞれ1872.4と420

であった．1-$\beta$を0.8，有意水準0.05としてサンプルサイズを設計しなさい．

```
> mA = 2350;   sA = 258
> mB = 1872.4; sB = 420
> es = abs(mA - mB)/sqrt((sA*sA+sB*sB)/2)
> print(es)
[1] 1.370277
> pwr.t.test(d=es, sig.level=0.05, power=0.8, type="two.sample", alternative="two.sided")

     Two-sample t test power calculation

              n = 9.424618
              d = 1.370277
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

**n = 10**

標準　医用画像の視覚評価法，p.131拡張

# 事後分析：Post hoc　検出力分析

**「何も特別なことは起こっていない」**
（帰**無**仮説：差があるとはいえ**な**い）
という事象の閾値

**p = 0.05**

**本来ある差が正しく検出される確率**
サンプルサイズ，効果量，有意水準で決定
（このことをPower Analysisという）

pが小さいと検出力も低下

検出力には0.8を使う場合が多い
||
**有意差があるときに**
**80%の確率で正しく有意差ありといえる**

**1-β = ?**

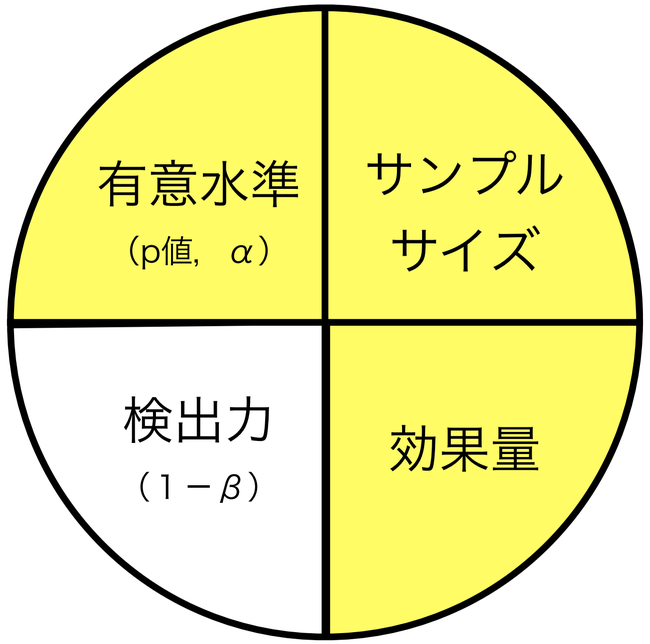



図１　検定において考えるべき４つの要素

**観測するデータの数**
画像数，観察者数
大きくすると差が出やすくなるが
コストがかかる
また，本来なかった差があたかも
差があったようにでることになる

**n = 51**

**２つの群の間の差や値の関連の強さ**

効果小：0.2
効果中：0.5
効果大：0.8

$$\text{Cohen's D} = \frac{|m_1 - m_2|}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_1^2}{2}}$$

**d = 0.5**

標準　医用画像の視覚評価法，p.105

**（問題５）**
ＡとＢ２つの群の値の平均に**違いがあるか**比較した.
効果量は0.5, サンプルサイズは50であった. p=0.05で検定した.
この場合の検出力はいくらか？

```
install.packages("pwr")
library(pwr)
pwr.t2n.test(n1=50, n2=50, d=0.5, sig.level=0.05, power=NULL, alternative="two.sided")
```

```
> pwr.t2n.test(n1=50, n2=50, d=0.5, sig.level=0.05, power=NULL, alternative="two.sided")

     t test power calculation

             n1 = 50
             n2 = 50
              d = 0.5
      sig.level = 0.05
          power = 0.6968934
    alternative = two.sided
```

標準　医用画像の視覚評価法, p.126拡張

**（問題６）**
A群の平均がB群の平均よりも**大きいか**比較した.
効果量は0.5, サンプルサイズは50であった. p=0.05で検定した.
この場合の検出力はいくらか？

```
install.packages("pwr")
library(pwr)
pwr.t2n.test(n1=50, n2=50, d=0.5, sig.level=0.05, power=NULL, alternative="greater")
```

```
> pwr.t2n.test(n1=50, n2=50, d=0.5, sig.level=0.05, power=NULL, alternative="greater")

     t test power calculation

             n1 = 50
             n2 = 50
              d = 0.5
      sig.level = 0.05
          power = 0.7989362
    alternative = greater
```

標準　医用画像の視覚評価法, p.126拡張

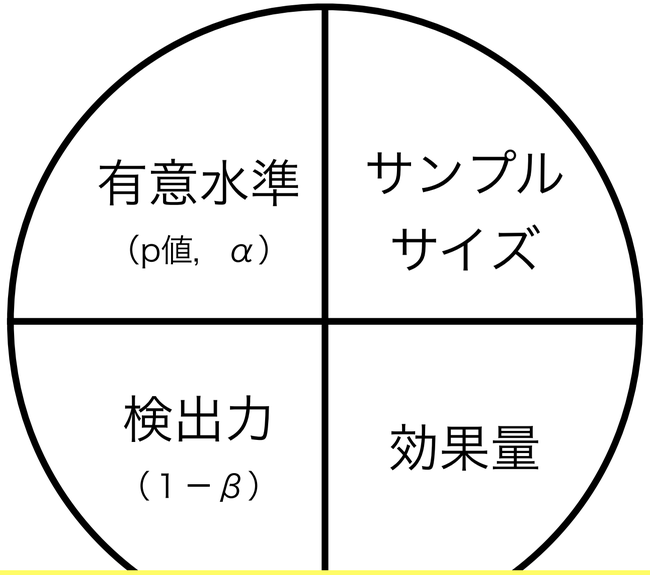
# ここまでのまとめ

- 事前分析：A prior はサンプルサイズ（観察者実験の場合には，観察者の数）を決める方法として有用である.
- 事後分析：Post hocは検出力分析で検出力（検定力）を計算できる.
- このような統計操作はいわゆる　**Power Analysis** と呼ばれる.
- 検定にはt検定を用いる方法を例として述べた.

# 改めて２

**「何も特別なことは起こっていない」**
（帰**無**仮説：差があるとはいえ**な**い）
という事象の閾値

**観測するデータの数**
画像数，観察者数
大きくすると差が出やすくなるが
コストがかかる
また，本来なかった差があたかも
差があったようにでることになる




**本来ある差が正しく検出される確率**
サンプルサイズ，効果量，有意水準で決定
（このことをPower Analysisという）

## 事前分析：A priori　サンプルサイズの設計

$$\text{Cohen's D} = \frac{|m_1 - m_2|}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_1^2}{2}}$$

検出力

## 事後分析：Post hoc　検出力分析

**有意差があるときに**
**80%の確率で正しく有意**

## どんな検定を用いるか？

標準　医用画像の視覚評価法，p.105

| データの対応 | サンプルサイズ 小さい | サンプルサイズ 大きい（n > 30） |
|---|---|---|
| なし | U検定<br>ウィルコクソン順位和検定<br>ノンパラメトリック | t検定<br>t-test<br>パラメトリック |
| あり | ウィルコクソン符号順位検定 | 対応ありt検定<br>paired t-test |

標準　医用画像の視覚評価法，p.121まとめ

データの対応

なし

表８　同一条件で観察した３つの観察者群のROC曲線下面積

| 観察者群１ | ROC曲線下面積 | 観察者群２ | ROC曲線下面積 | 観察者群３ | ROC曲線下面積 |
|---|---|---|---|---|---|
| 観察者１ | 0.736 | 観察者１０ | 0.801 | 観察者１８ | 0.851 |
| 観察者２ | 0.676 | 観察者１１ | 0.845 | 観察者１９ | 0.892 |
| 観察者３ | 0.521 | 観察者１２ | 0.726 | 観察者２０ | 0.771 |
| 観察者４ | 0.754 | 観察者１３ | 0.997 | 観察者２１ | 0.954 |
| 観察者５ | 0.748 | 観察者１４ | 0.829 | 観察者２２ | 0.865 |
| 観察者６ | 0.946 | 観察者１５ | 0.908 | | |
| 観察者７ | 0.765 | 観察者１６ | 0.879 | | |
| 観察者８ | 0.551 | 観察者１７ | 0.696 | | |
| 観察者９ | 0.715 | | | | |

**観察条件によって観察者が違う**
その結果，人数もばらばら
同じ人の変化が追跡**できない**
||
対応なし

あり

表５　ある観察者実験から得られた２つ目の条件のROC曲線下面積

| 観察者名 | ROC曲線下面積<br>観察条件１ | ROC曲線下面積<br>観察条件２ | 観察条件の差<br>条件１－条件２ |
|---|---|---|---|
| 観察者１ | 0.736 | 0.801 | -0.065 |
| 観察者２ | 0.676 | 0.845 | -0.169 |
| 観察者３ | 0.521 | 0.726 | -0.205 |
| 観察者４ | 0.754 | 0.997 | -0.243 |
| 観察者５ | 0.748 | 0.829 | -0.081 |
| 観察者６ | 0.946 | 0.908 | 0.038 |
| 観察者７ | 0.765 | 0.879 | -0.114 |
| 観察者８ | 0.551 | 0.696 | -0.145 |
| 観察者９ | 0.715 | 0.873 | -0.158 |

**観察条件が違っても観察者が同じ**
その結果，人数は同じ
同じ人の変化が追跡**できる**
**値の引き算が可能**
差の分布を考えればよい
差の母平均の信頼区間がゼロをまたぐか否か
||
対応あり

標準　医用画像の視覚評価法，p.120, 136

ノンパラメトリック

データの対応

なし

表８　同一条件で観察した３つの観察者群のROC曲線下面積

| 観察者群１ | ROC曲線下面積 |
|---|---|
| 観察者１ | 0.736 |
| 観察者２ | 0.676 |
| 観察者３ | 0.521 |
| 観察者４ | 0.754 |
| 観察者５ | 0.748 |
| 観察者６ | 0.946 |
| 観察者７ | 0.765 |
| 観察者８ | 0.551 |
| 観察者９ | 0.715 |

| 観察者群２ | ROC曲線下面積 |
|---|---|
| 観察者１０ | 0.801 |
| 観察者１１ | 0.845 |
| 観察者１２ | 0.726 |
| 観察者１３ | 0.997 |
| 観察者１４ | 0.829 |
| 観察者１５ | 0.908 |
| 観察者１６ | 0.879 |
| 観察者１７ | 0.696 |

| 観察者群３ | ROC曲線下面積 |
|---|---|
| 観察者１８ | 0.851 |
| 観察者１９ | 0.892 |
| 観察者２０ | 0.771 |
| 観察者２１ | 0.954 |
| 観察者２２ | 0.865 |

**観察条件によって観察者が違う**
その結果，人数もばらばら
同じ人の変化が追跡**できない**
||
対応なし

```
> library(exactRankTests)
> cond1<-c(0.736, 0.676, 0.521, 0.754, 0.748, 0.946, 0.765, 0.551, 0.715)
> cond2<-c(0.801, 0.845, 0.726, 0.997, 0.829, 0.908, 0.879, 0.696)
> wilcox.exact(cond1, cond2, paired=F)

    Exact Wilcoxon rank sum test

data:  cond1 and cond2
W = 16, p-value = 0.05923
alternative hypothesis: true mu is not equal to 0
```

**有意差なし**

標準　医用画像の視覚評価法，p.120, 122, 136

# パラメトリック

データの対応

なし

表８　同一条件で観察した３つの観察者群のROC曲線下面積

| 観察者群１ | ROC曲線下面積 |
|---|---|
| 観察者１ | 0.736 |
| 観察者２ | 0.676 |
| 観察者３ | 0.521 |
| 観察者４ | 0.754 |
| 観察者５ | 0.748 |
| 観察者６ | 0.946 |
| 観察者７ | 0.765 |
| 観察者８ | 0.551 |
| 観察者９ | 0.715 |

| 観察者群２ | ROC曲線下面積 |
|---|---|
| 観察者１０ | 0.801 |
| 観察者１１ | 0.845 |
| 観察者１２ | 0.726 |
| 観察者１３ | 0.997 |
| 観察者１４ | 0.829 |
| 観察者１５ | 0.908 |
| 観察者１６ | 0.879 |
| 観察者１７ | 0.696 |

| 観察者群３ | ROC曲線下面積 |
|---|---|
| 観察者１８ | 0.851 |
| 観察者１９ | 0.892 |
| 観察者２０ | 0.771 |
| 観察者２１ | 0.954 |
| 観察者２２ | 0.865 |

**観察条件によって観察者が違う**
その結果，人数もばらばら
同じ人の変化が追跡**できない**
||
対応なし

```
> cond1<-c(0.736, 0.676, 0.521, 0.754, 0.748, 0.946, 0.765, 0.551, 0.715)
> cond2<-c(0.801, 0.845, 0.726, 0.997, 0.829, 0.908, 0.879, 0.696)
> t.test(cond1, cond2, var.equal=F)

    Welch Two Sample t-test

data:  cond1 and cond2
t = -2.2729, df = 14.771, p-value = 0.03842
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.237883675 -0.007477436
sample estimates:
mean of x mean of y
0.7124444 0.8351250
```

**有意差あり**

**信頼区間**

標準　医用画像の視覚評価法，p.120, 122, 136

データの対応

なし

```
> library(exactRankTests)
> cond1<-c(0.736, 0.676, 0.521, 0.754, 0.748, 0.946, 0.765, 0.551, 0.715)
> cond2<-c(0.801, 0.845, 0.726, 0.997, 0.829, 0.908, 0.879, 0.696)
> wilcox.exact(cond1, cond2, paired=F)

    Exact Wilcoxon rank sum test

data:  cond1 and cond2
W = 16, p-value = 0.05923
alternative hypothesis: true mu is not equal to 0
```

ノンパラメトリック

有意差なし

**検定を何度も繰り返すと**

**検定も失敗する**

ということに似ている

```
748, 0.946,  765, 0.551, 0.715)
829, 0.908,  879, 0.696)
```

こんなんいいの？

いいんです！

有意差あり

```
data:  cond1 and cond2
t = -2.2729, df = 14.771, p-value = 0.03842
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.237883675 -0.007477436
sample estimates:
mean of x mean of y
0.7124444 0.8351250
```

パラメトリック

標準　医用画像の視覚評価法，p.120, 122, 136

データの対応

あり

```
> library(exactRankTests)
> cond1<-c(0.736, 0.676, 0.521, 0.754, 0.748, 0.946, 0.765, 0.551, 0.715)
> cond3<-c(0.801, 0.845, 0.726, 0.997, 0.829, 0.908, 0.879, 0.696, 0.873)
> wilcox.exact(cond1, cond3, paired=T)

    Exact Wilcoxon signed rank test

data:  cond1 and cond3
V = 1, p-value = 0.007812
alternative hypothesis: true mu is not equal to 0
```

有意差あり

ノンパラメトリック

表５　ある観察者実験から得られた２つ目の条件のROC曲線下面積

| 観察者名 | ROC曲線下面積 観察条件１ | ROC曲線下面積 観察条件２ | 観察条件の差 条件１－条件２ |
|---|---|---|---|
| 観察者１ | 0.736 | 0.801 | -0.065 |
| 観察者２ | 0.676 | 0.845 | -0.169 |
| 観察者３ | 0.521 | 0.726 | -0.205 |
| 観察者４ | 0.754 | 0.997 | -0.243 |
| 観察者５ | 0.748 | 0.829 | -0.081 |
| 観察者６ | 0.946 | 0.908 | 0.038 |
| 観察者７ | 0.765 | 0.879 | -0.114 |
| 観察者８ | 0.551 | 0.696 | -0.145 |
| 観察者９ | 0.715 | 0.873 | -0.158 |

**観察条件が違っても観察者が同じ**

その結果，人数は同じ

同じ人の変化が追跡**できる**

**値の引き算が可能**

差の分布を考えればよい

差の母平均の信頼区間がゼロをまたぐか否か

||

対応あり

標準　医用画像の視覚評価法，p.120, 136

データの対応

あり

```
> cond1<-c(0.736, 0.676, 0.521, 0.754, 0.748, 0.946, 0.765, 0.551, 0.715)
> cond3<-c(0.801, 0.845, 0.726, 0.997, 0.829, 0.908, 0.879, 0.696, 0.873)
> t.test(cond1, cond3, paired=TRUE)

    Paired t-test

data:  cond1 and cond3
t = -4.5522, df = 8, p-value = 0.001869
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.19116747 -0.06261031
sample estimates:
mean of the differences
              -0.1268889
```

有意差あり

パラメトリック

信頼区間

表５　ある観察者実験から得られた２つ目の条件のROC曲線下面積

| 観察者名 | ROC曲線下面積 観察条件１ | ROC曲線下面積 観察条件２ | 観察条件の差 条件１－条件２ |
|---|---|---|---|
| 観察者１ | 0.736 | 0.801 | -0.065 |
| 観察者２ | 0.676 | 0.845 | -0.169 |
| 観察者３ | 0.521 | 0.726 | -0.205 |
| 観察者４ | 0.754 | 0.997 | -0.243 |
| 観察者５ | 0.748 | 0.829 | -0.081 |
| 観察者６ | 0.946 | 0.908 | 0.038 |
| 観察者７ | 0.765 | 0.879 | -0.114 |
| 観察者８ | 0.551 | 0.696 | -0.145 |
| 観察者９ | 0.715 | 0.873 | -0.158 |

**観察条件が違っても観察者が同じ**

その結果，人数は同じ

同じ人の変化が追跡**できる**

**値の引き算が可能**

差の分布を考えればよい

差の母平均の信頼区間がゼロをまたぐか否か

||

対応あり

標準　医用画像の視覚評価法，p.120, 136

# 事前分析：A priori　サンプルサイズの設計

## シナリオ

対応あり　か　対応なしか？

サンプル数：２（２群）

t検定　か　ウィルコクソンか？

- ２つの撮影条件でどちらの条件が診断能が高いか調べたい

対応あり　か　対応なしか？

AUCの比較？　FOMの比較？

- 観察者実験で明らかにしたいのでROCによる実験を行いたい

サンプルサイズ：n人（１群あたり）

- 画像を何枚用意して，観察者を何人にすればよいのかを決めたい

サンプルサイズ：n（１群あたり）

期待されるAUCの差，効果量を予備実験で調べる

実験の後には

# 事後分析：Post hoc　検出力分析

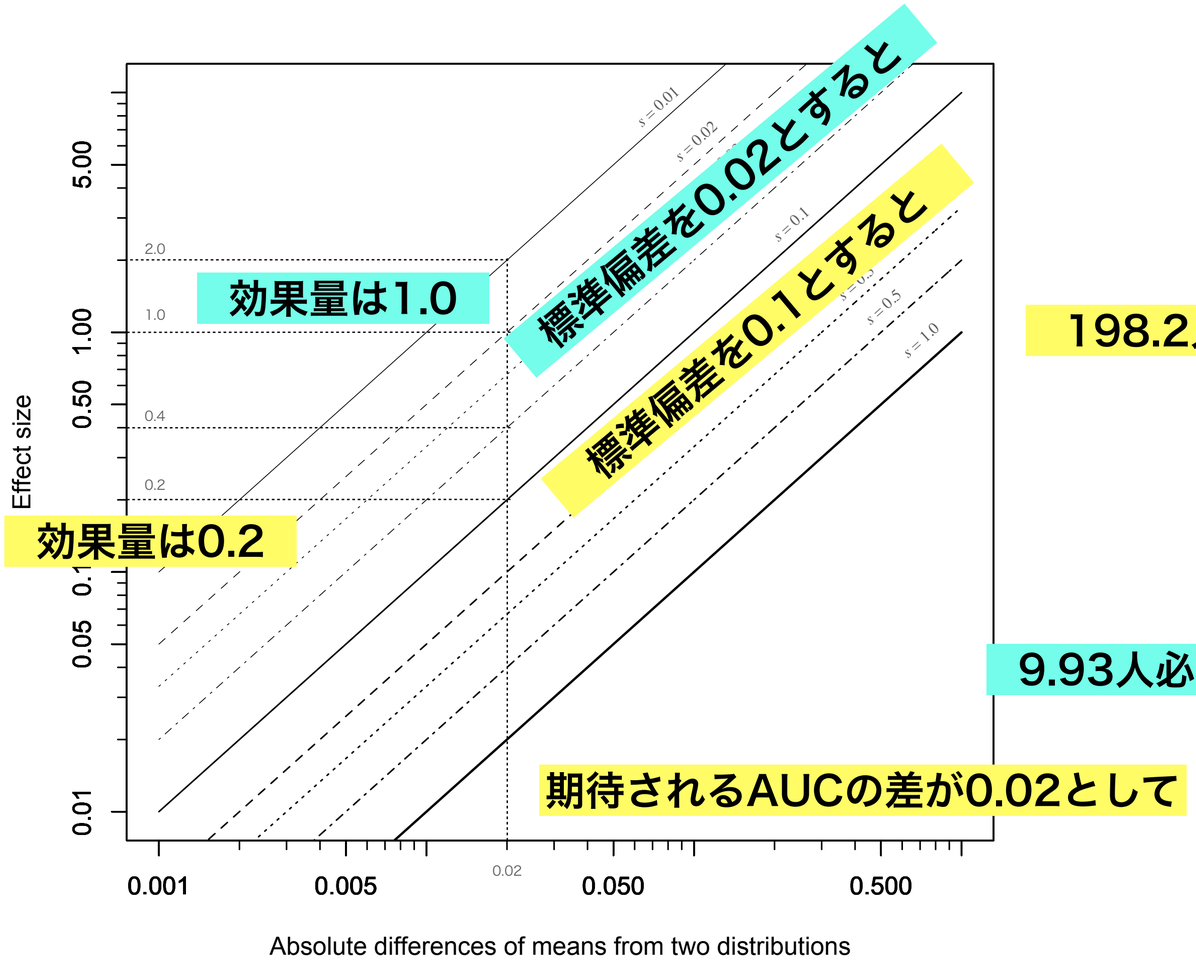
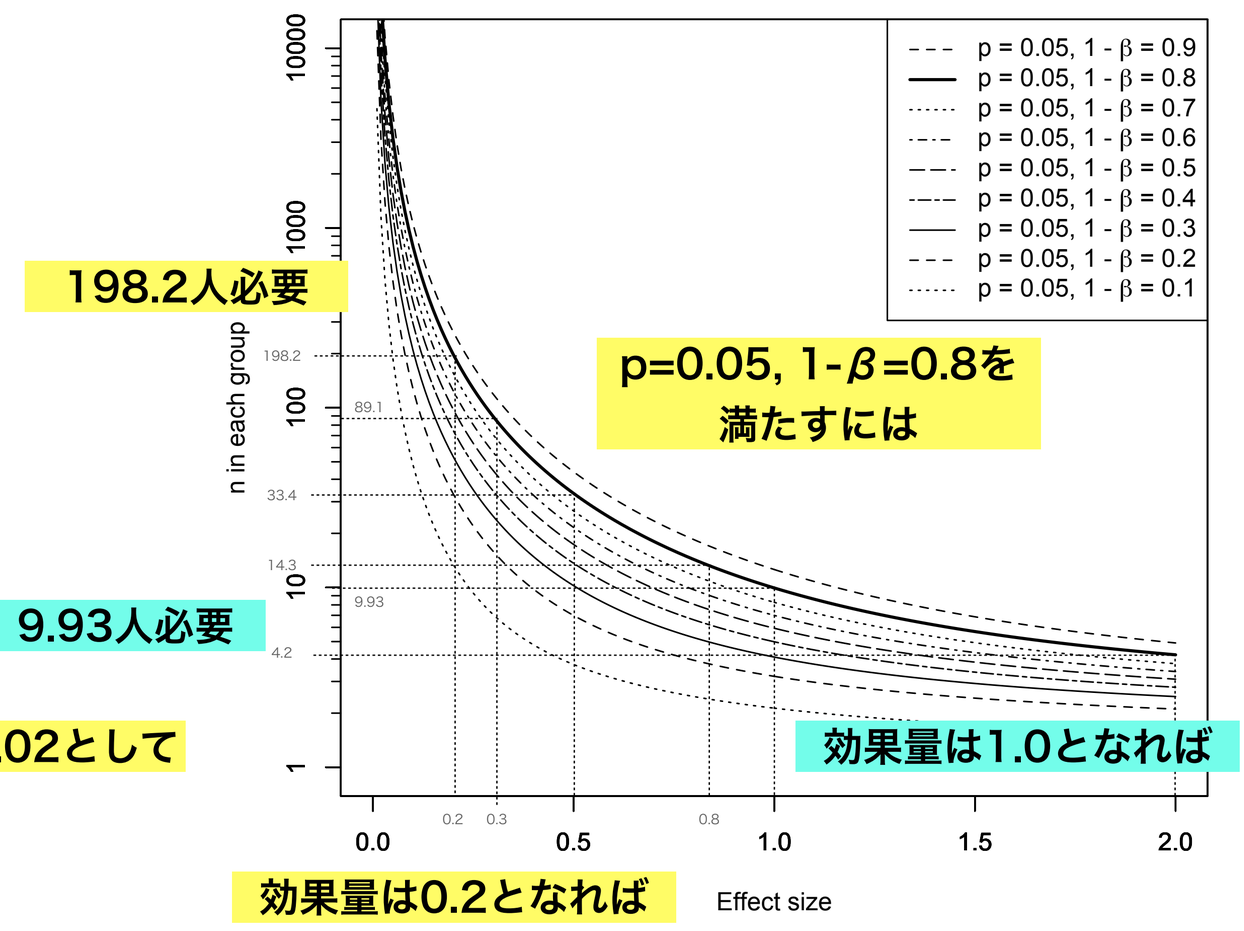



図２２　ROC解析で用いられる値の範囲で推定される効果量変化




(b)　対応ありの母平均の検定

図２０　有意水準と検出力と効果量から設計されるサンプルサイズ

# 事前分析：A priori　サンプルサイズの設計

標準　医用画像の視覚評価法，p.128, 130

```
> power.t.test(n=NULL, delta=1, sig.level=0.05, power=0.8, type="paired")

     Paired t test power calculation

              n = 9.937864
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

効果量が1.0ならば
10人の観察者でよい

(a) 効果量1.0，有意水準5%，検出力0.8，対応ありのt検定の場合

```
> power.t.test(n=NULL, delta=0.2, sig.level=0.05, power=0.8, type="paired")

     Paired t test power calculation

              n = 198.1513
          delta = 0.2
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

効果量が0.2ならば
199人の観察者が必要

(b) 効果量0.2，有意水準5%，検出力0.8，対応ありのt検定の場合

図２３　サンプルサイズの推定

# 画像枚数はどれだけ必要か？

- **明確な指標はない（と思う）**
  - 妥当な枚数，観察者の負担になりすぎない枚数.
  - 予備実験で一人の観察者から効果量を求め十分な効果量が確認する.
  - . . . . .

# 予備実験でAUCを計算

```
> ps<-c(
+ 0.6943926, 0.7908950, 0.8437792, 0.6333895, 0.9489886,
+ 0.8356554, [illegible], 0.8932312,
+ 0.9357352, [illegible], 0.7431715,
+ 0.9104615, [illegible], 0.6756603,
+ 0.8212344, [illegible], 0.7956368,
+ 0.8595545, 0.8478187, 0.8239134, 0.9281065, 0.7693703)
>
> ns<-c(
+ 0.6714567, 0.7451352, 0.6025817, 0.6553833, 0.7820320,
+ 0.5931192, 0.6317957, 0.7334455, 0.8191236, 0.8852585,
+ 0.5633350, [illegible], 0.6381366,
+ 0.6710714, [illegible], 0.7186398,
+ 0.7734048, [illegible], 1.0134752,
+ 0.8306870, 0.7658674, 0.5302526, 0.8650005, 0.6679478)
>
> pgs<-rep(1, casenum); ngs<-rep(0, casenum)
>
> res <- prediction(c(ps, ns), c(pgs, ngs))
> roc <- performance(res, "tpr", "fpr")
> auc <- performance(res, "auc")
>
> plot(unlist(roc@x.values), unlist(roc@y.values), type="l",
+   xlab="False positive fraction",
+   ylab="True positive fraction",
+   asp = 1)
>
> print(unlist(auc@y.values))
[1] 0.8166667
```

陽性の確信度

陰性の確信度

AUC

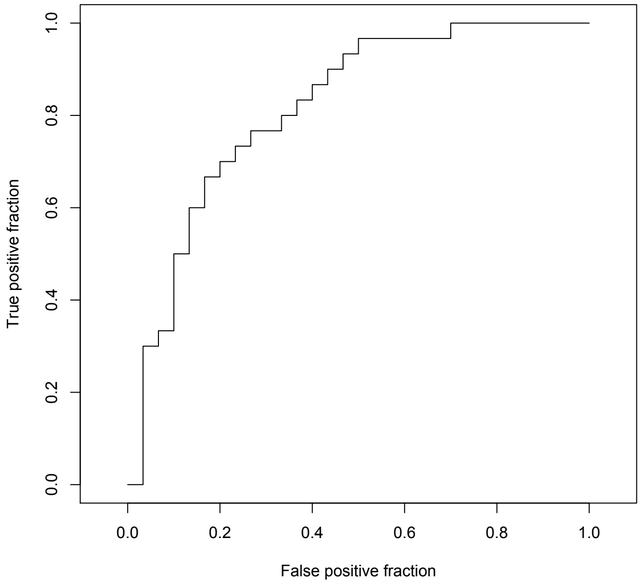



図31　確信度からROC曲線を生成しAUCを算出する例

陽性と陰性の確信度を乱数で作成

陽性：0.8

陰性：0.2

どちらもSD：0.1

```
> library(ROCR)
> library(pROC)
>
>
> set.seed(12345)
> casenum <- 20
> ps <- rnorm(casenum, 0.8, 0.1)
> ns <- rnorm(casenum, 0.2, 0.1)
>
> ps
 [1] 0.8585529 0.8709466 0.7890697 0.7546503 0.8605887 0.6182044 0.8630099 0
[10] 0.7080678 0.7883752 0.9817312 0.8370628 0.8520216 0.7249468 0.8816900 0
[19] 0.9120713 0.8298724
> ns
 [1] 0.27796219 0.34557851 0.13556716 0.04468626 0.04022905 0.38050975 0.151
 [9] 0.26121235 0.18376890 0.28118732 0.41968335 0.40491903 0.36324456 0.225
[17] 0.16759134 0.03379498 0.37677339 0.20258010
>
> pgs <-rep(1, casenum); ngs<-rep(0, casenum)
> res <- prediction(c(ps, ns), c(pgs, ngs))
> roc <- performance(res, "tpr", "fpr")
> roc <- performance(res, "auc")
> as.numeric(unlist(roc@y.values))
[1] 1
```

AUC 1.0

**AUCの大量生産ができそう**

標準　医用画像の視覚評価法，p.140，改

# 予備実験でAUCを計算

**様々な条件でAUCを量産**

```
estimateAUC <- function(pm, nm, sd, num){
    #pmは陽性の平均，nmは陰性の平均
    #sdはそれぞれの標準偏差，numはそれぞれの回答数
    tps <- rnorm(num, pm, sd); #print(tps)
    tns <- rnorm(num, nm, sd); #print(tns)
    tpgs <-rep(1, num); tngs<-rep(0, num)
    tres <- prediction(c(tps, tns), c(tpgs, tngs))
    troc <- performance(tres, "tpr", "fpr")
    troc <- performance(tres, "auc")
    return(as.numeric(troc@y.values))
}
```

**条件１のAUCの平均とSD**

```
> estimateAUC(0.8, 0.2, 0.5, 30)
[1] 0.8233333
> estimateAUC(0.8, 0.2, 0.5, 30)
[1] 0.8211111
[1] 0.7522222
> estimateAUC(0.8, 0.2, 0.5, 30)
[1] 0.77
[1] 0.7
> estimateAUC(0.8, 0.2, 0.5, 30)
[1] 0.8122222
> estimateAUC(0.8, 0.2, 0.5, 30)
[1] 0.8788889
```

**条件２のAUCの平均とSD**

```
> estimateAUC(0.6, 0.4, 0.5, 30)
[1] 0.58
> estimateAUC(0.6, 0.4, 0.5, 30)
[1] 0.6611111
> estimateAUC(0.6, 0.4, 0.5, 30)
[1] 0.7444444
> estimateAUC(0.6, 0.4, 0.5, 30)
[1] 0.5611111
> estimateAUC(0.6, 0.4, 0.5, 30)
[1] 0.7044444
> estimateAUC(0.6, 0.4, 0.5, 30)
[1] 0.5933333
> estimateAUC(0.6, 0.4, 0.5, 30)
[1] 0.6933333
```

**それぞれの群のAUCのSDは サンプルサイズに依存**

**AUCの差が決まる**

**サンプルサイズ30のときの 効果量が推定できる**

標準　医用画像の視覚評価法，p.140，改

# 予備実験でAUCを計算

## 様々な条件でAUCを量産：20回

```
repeatEstimation<-function(ma, mb, sd, num, iter){
    #ma：陽性群の平均，mb：陰性群の平均
    #sd：それぞれの標準偏差，num：画像枚数
    #iter：試行回数（観察者の人数）
    auclist <-c()
    for (cnt in 1:iter){
        tmp<-estimateAUC(ma, mb, sd, num)
        #print(tmp)
        auclist <- c(auclist, tmp)
    }
    return(auclist)
}
```

**条件１のAUCの平均とSD**

陽性の確信度の平均：0.6

陰性の確信度の平均：0.4

SD：0.3

画像数：30+30

**条件２のAUCの平均とSD**

陽性の確信度の平均：0.7

陰性の確信度の平均：0.3

SD：0.3

画像数：30+30

```
> cond1auc <- repeatEstimation(0.6, 0.4, 0.3, 30, 20)
> cond2auc <- repeatEstimation(0.7, 0.3, 0.3, 30, 20)
> mean(cond1auc);sd(cond1auc)
[1] 0.6940556
[1] 0.06404928
> mean(cond2auc);sd(cond2auc)
[1] 0.8247222
[1] 0.05358236
>
>
> cond1auc
 [1] 0.6088889 0.7322222 0.7533333 0.6322222 0.7166667 0.8155556 0.6977778 0.7422222 0.7300000
[10] 0.6211111 0.6355556 0.7033333 0.6044444 0.7022222 0.6566667 0.6677778 0.6200000 0.8277778
[19] 0.6944444 0.7188889
> cond2auc
 [1] 0.9388889 0.7466667 0.8255556 0.7822222 0.8755556 0.8022222 0.8366667 0.8211111 0.8444444
[10] 0.7922222 0.7511111 0.7866667 0.8377778 0.8500000 0.8500000 0.8144444 0.8633333 0.7133333
[19] 0.8922222 0.8700000
```

**AUCの平均と標準偏差**

| | |
|---|---|
| 条件１ | 0.694, 0.064 |
| 条件２ | 0.825, 0.054 |

標準　医用画像の視覚評価法，p.140，改

# 予備実験でAUCを計算

```
effectsize <- function(dat1, dat2){
    #２つの群の分布から
    #簡易的にCohen's dを求める
    diff <- abs(mean(dat1)-mean(dat2))
    sds <- sqrt((sd(dat1)*sd(dat1)+sd(dat2)*sd(dat2))/2.0)
    return (diff/sds)
}

eslist <- c()
for (n in 1:100){
    cond1auc <- repeatEstimation(0.6, 0.4, 0.3, n, 20)
    cond2auc <- repeatEstimation(0.7, 0.3, 0.3, n, 20)
    eslist <- c(eslist, effectsize(cond1auc, cond2auc))
}
```

**条件１のAUCの平均とSD**

陽性の確信度の平均：0.6

陰性の確信度の平均：0.4

SD：0.3

画像数：30+30

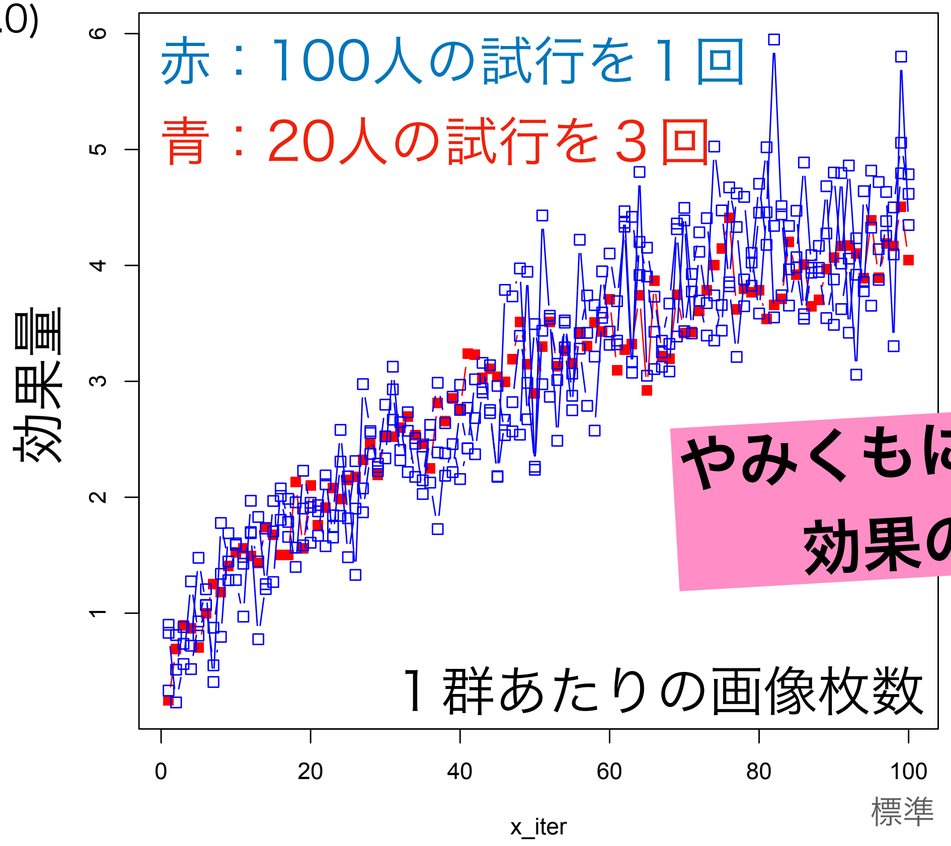
**条件２のAUCの平均とSD**

陽性の確信度の平均：0.7

陰性の確信度の平均：0.3

SD：0.3

画像数：30+30




**やみくもに枚数を増やしても効果の上昇は小さい**

標準　医用画像の視覚評価法，p.140，改

# 観察者実験が終わりました
# あとは統計処理！

p値を出す：不等号では書かない

信頼区間を書く

効果量と検出力も明らかにする

**（実験結果１）**

ＡとＢ２つの観察条件でAUCの値の平均に**違いがあるか**比較した.
対応ありのt検定で行った．効果量は0.37015, サンプルサイズは8であった．p=0.05で検定した.
有意差は出たけど，この場合の検出力はいくらか？

```
install.packages("pwr")
library(pwr)
power.t.test(n=8, delta=0.37015, sig.level=0.05, power=NULL, type="paired")
```

```
> power.t.test(n=8, delta=0.37015, sig.level=0.05, power=NULL, type="paired")

     Paired t test power calculation

              n = 8
          delta = 0.37015
             sd = 1
      sig.level = 0.05
          power = 0.1467562
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

**検出力が0.1467と低い**
なので有意差はあったけど
これはたまたま（１－β）で
あまり信用できないだろう

**１－β を改めて**
β：第２種の誤り
本当は「ある」のに「ない」という間違い
つまりぼんやり「見落とし」
１－βなので
**見落としていない確率**

**つまり検出力が0.1467とは**
**見落としていない確率が0.1467**

標準　医用画像の視覚評価法，p.126拡張

**（実験結果１）**

ＡとＢ２つの観察条件でAUCの値の平均に**違いがあるか**比較した.
対応ありのt検定で行った．効果量は0.37015, サンプルサイズは8であった．p=0.05で検定した.
有意差は出たけど，この場合の検出力はいくらか？

**これは論文になるのか？**

```
install.packages("pwr")
library(pwr)
power.t.test(n=8, delta=0.37015, sig.level=0.05, power=NULL, type="paired")
```

```
> power.t.test(n=8, delta=0.37015, sig.level=0.05, power=NULL, type="paired")

     Paired t test power calculation

              n = 8
          delta = 0.37015
             sd = 1
      sig.level = 0.05
          power = 0.1467562
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

**意図的に隠蔽？！**

**検出力が0.1467と低い**
なので有意差はあったけど
これはたまたま（１－β）で
あまり信用できないだろう

**１－β を改めて**
β：第２種の誤り
本当は「ある」のに「ない」という間違い
つまりぼんやり「見落とし」
１－βなので
**見落としていない確率**

**つまり検出力が0.1467とは**
**見落としていない確率が0.1467**

標準　医用画像の視覚評価法，p.126拡張

**（実験結果２）**

ＡとＢ２つの観察条件でAUCの値の平均に**違いがあるか**比較した.
対応ありのt検定で行った．効果量は0.1215, サンプルサイズは25であった．p=0.05で検定した.
有意差は出たけど，この場合の検出力はいくらか？

```
install.packages("pwr")
library(pwr)
power.t.test(n=25, delta=0.1215, sig.level=0.05, power=NULL, type="paired")
```

```
> power.t.test(n=25, delta=0.2215, sig.level=0.05, power=NULL, type="paired")

     Paired t test power calculation

              n = 25
          delta = 0.2215
             sd = 1
      sig.level = 0.05
          power = 0.1850255
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

**検出力が0.1850255**

**サンプルサイズの設計では199人必要だったのに**
**２５人でしか実験できなかった**
（２５人で有意差が出たのでやめた！）

**当然，検出力が不足する**

標準　医用画像の視覚評価法，p.126拡張

**（実験結果２）**

ＡとＢ２つの観察条件でAUCの値の平均に**違いがあるか**比較した.
対応ありのt検定で行った．効果量は0.1215, サンプルサイズは25であった．p=0.05で検定した.
有意差は出たけど，この場合の検出力はいくらか？

```
install.packages("pwr")
library(pwr)
power.t.test(n=25, delta=0.1215, sig.level=0.05, power=N
```

**これは論文になるのか？**

```
> power.t.test(n=25, delta=0.2215, sig.level=0.05, power=NULL, type="paired")

     Paired t test power calculation

              n = 25
          delta = 0.2215
             sd = 1
      sig.level = 0.05
          power = 0.1850255
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

**意図的に隠蔽？！**

**検出力が0.1850255**

**サンプルサイズの設計では199人必要だったのに**
**２５人でしか実験できなかった**
（２５人で有意差が出たのでやめた！）

**当然，検出力が不足する**

標準　医用画像の視覚評価法，p.126拡張

**（実験結果２）**

ＡとＢ２つの観察条件でAUCの値の平均に**違いがあるか**比較した.
対応ありのt検定で行った．効果量は0.1215, サンプルサイズは25であった．p=0.05で検定した.
有意差は出たけど，この場合の検出力はいくらか？

```
install.packages("pwr")
library(pwr)
power.t.test(n=25, delta=0.1215, sig.level=0.05, power=N
```

**これは論文になるのか？**

```
> power.t.test(n=25, delta=0.2215, sig.level=0.05, power=NULL, type="paired")

     Paired t test power calculation

              n = 25
          delta = 0.2215
             sd = 1
      sig.level = 0.05
          power = 0.1850255
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

**意図的に隠蔽？！**

**検出力が0.1850255**

**サンプルサイズの設計では199人必要だったのに**
**２５人でしか実験できなかった**
（２５人で有意差が出たのでやめた！）

**当然，検出力が不足する**

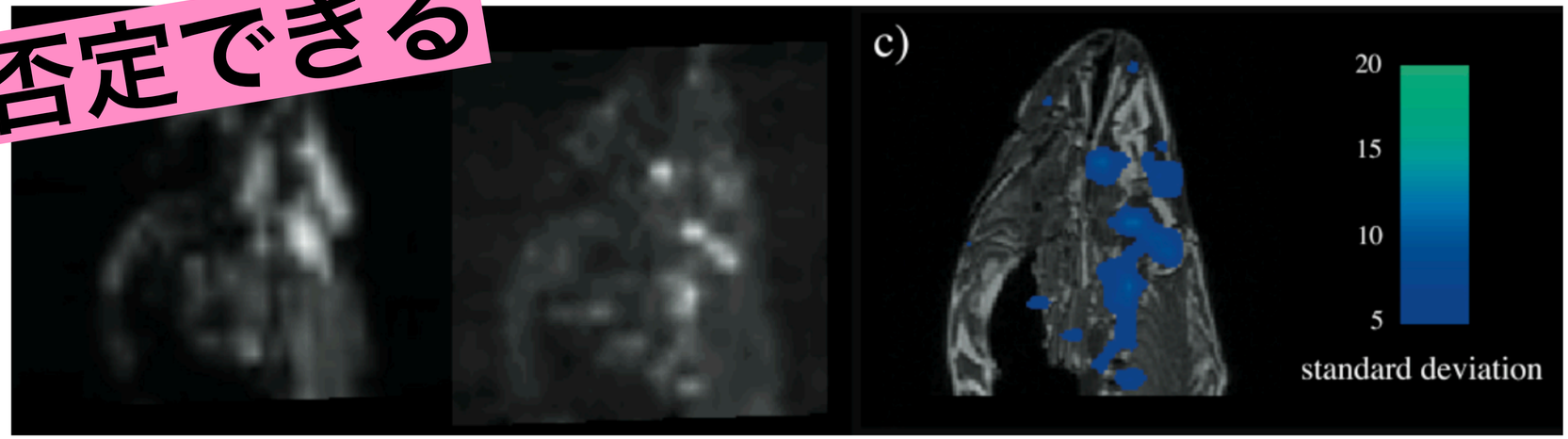
標準　医用画像の視覚評価法，p.126拡張

# 有意差が「ある」と言っても？

- 死んだサーモンのfMRIでも賦活領域に統計的有意差ありとでる

**VOXELWISE VARIABILITY**

検出力で否定できる



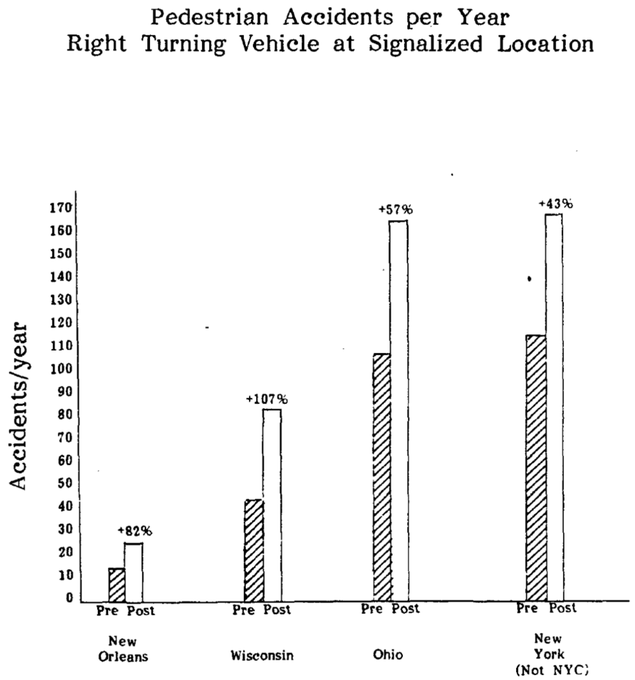

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

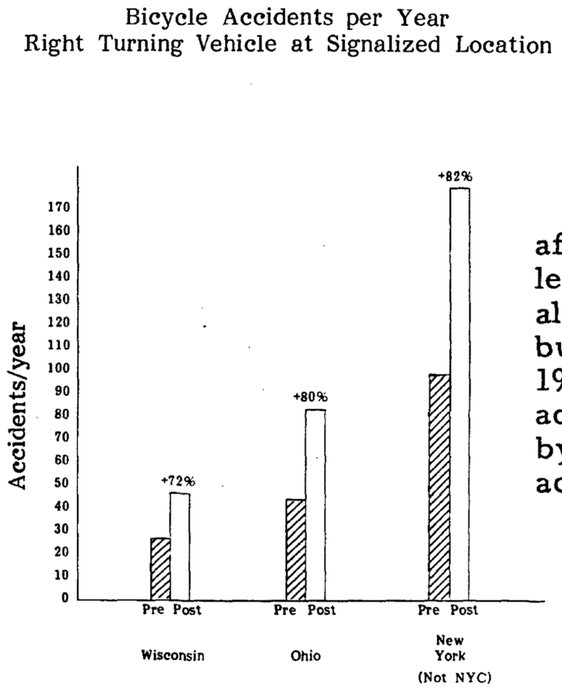
C. Bennett, A. Baird, M. Miller, G. Wolford. Neural correlates of in- terspecies perspective taking in the post-mortem atlantic salmon: An argument for proper multiple comparisons correction. Journal of Serendipitous and Unexpected Results, 1:1-5, 2010.

# 有意差が「ない」と言っても？

- 有意差が「ない」は「差がない」ではなく「差があるとはいえない」

The increases ranged from 43% to 107% for pedestrian accidents and 72% to 123% for bicyclist accidents. Over half of the accidents in which a vehicle turned right at a signalized location after the adoption of Western RTOR involved a right turn on a red signal. *




**論文から切り取って説明するのはよくないですが．．．**

In the Other Outstate areas, pedestrian accidents rose sharply in 1978 after dropping slightly in 1977, the first year after Western RTOR became legal. The average increase, 20%, was not statistically significant. Taken alone, the 1978 increase of 53% over baseline would be statistically significant but would require an interpretation that, although Western RTOR was legal in 1977, it was not practiced until 1978 in these areas of the state. Bicycle accidents in the same areas were higher than baseline in both 1977 and 1978, by an average of about 30%; the difference, consistent with the other bicycle accident series although somewhat smaller, did reach statistical significance.

**有意差がない　＝　差がない　（ここが間違い！）**

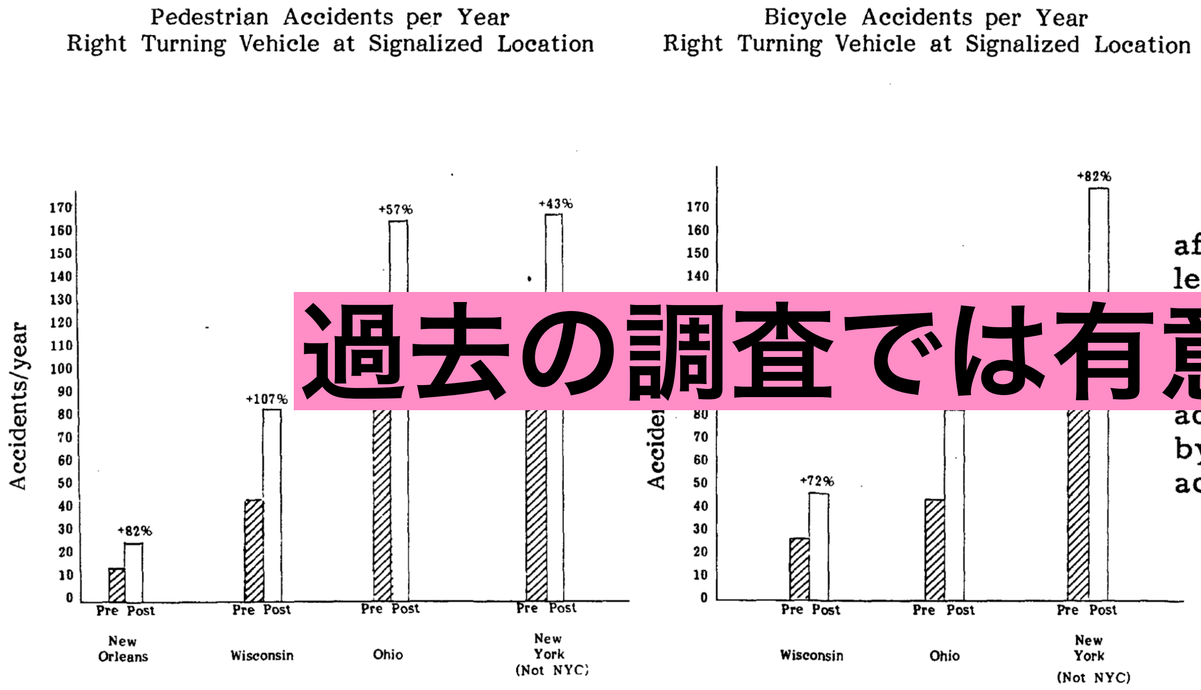
**＝　赤信号で右折（日本なら左折）しても事故は増えない**

* D. F. Preusser, W. A. Leaf, K. B. DeBartolo, R. D. Blomberg, M. M. Levy. The effect of right-turn-on-red on pedestrian and bicyclist accidents. Journal of Safety Research, 13:45—55, 1982.

# 有意差が「ない」と言っても？

- 有意差が「ない」は「差がない」ではなく「差があるとはいえない」

The increases ranged from 43% to 107% for pedestrian accidents and 72% to 123% for bicyclist accidents. Over half of the accidents in which a vehicle turned right at a signalized location after the adoption of Western RTOR involved a right turn on a red signal. *




**過去の調査では有意差なし**

**論文から切り取って説明するのはよくないですが．．．**

In the Other Outstate areas, pedestrian accidents rose sharply in 1978 after dropping slightly in 1977, the first year after Western RTOR became legal. The average increase, 20%, was not statistically significant. Taken of 53% over baseline would be statistically significant erpretation that, although Western RTOR was legal in d until 1978 in these areas of the state. Bicycle accidents in the same areas were higher than baseline in both 1977 and 1978, by an average of about 30%; the difference, consistent with the other bicycle accident series although somewhat smaller, did reach statistical significance.

**有意差がない　＝　差がない　（ここが間違い！）**

**＝　赤信号で右折（日本なら左折）しても事故は増えない**

* D. F. Preusser, W. A. Leaf, K. B. DeBartolo, R. D. Blomberg, M. M. Levy. The effect of right-turn-on-red on pedestrian and bicyclist accidents. Journal of Safety Research, 13:45—55, 1982.

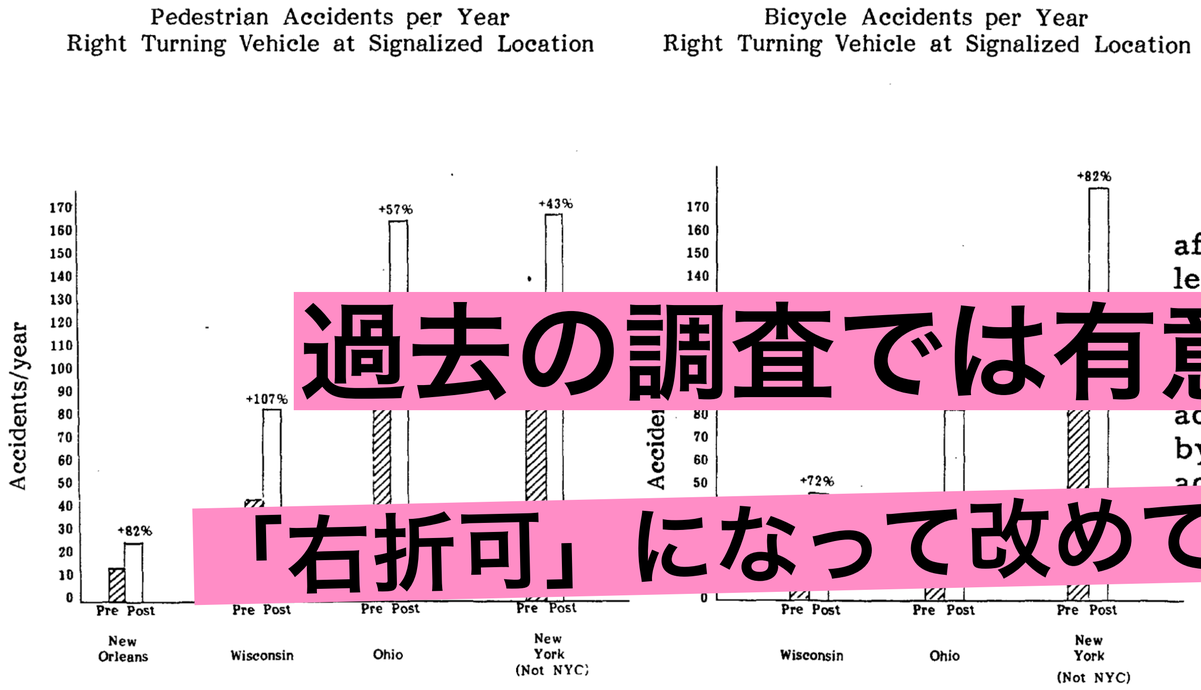
# 有意差が「ない」と言っても？

- 有意差が「ない」は「差がない」ではなく「差があるとはいえない」

The increases ranged from 43% to 107% for pedestrian accidents and 72% to 123% for bicyclist accidents. Over half of the accidents in which a vehicle turned right at a signalized location after the adoption of Western RTOR involved a right turn on a red signal. *

**同等性，非劣性の統計**

論文から切り取って説明するのはよくないですが．．．




**過去の調査では有意差なし**

In the Other Outstate areas, pedestrian accidents rose sharply in 1978 after dropping slightly in 1977, the first year after Western RTOR became legal. The average increase, 20%, was not statistically significant. Taken of 53% over baseline would be statistically significant erpretation that, although Western RTOR was legal in d until 1978 in these areas of the state. Bicycle accidents in the same areas were higher than baseline in both 1977 and 1978, by an average of about 30%; the difference, consistent with the other bicycle

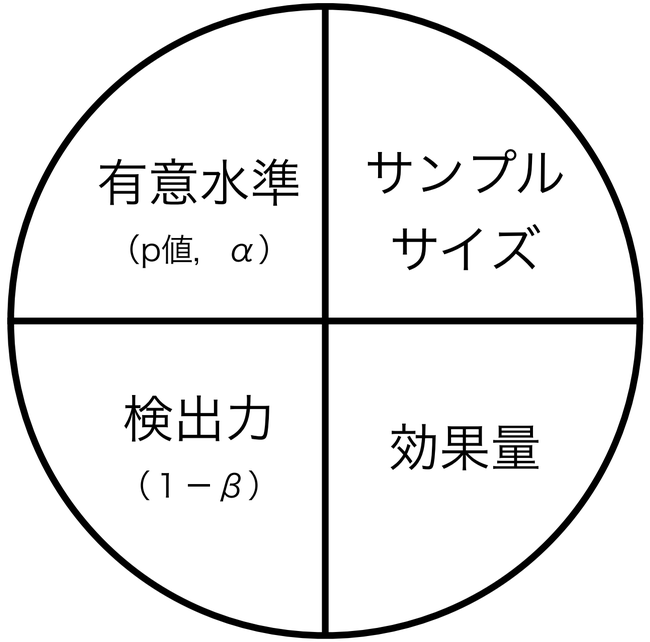
**「右折可」になって改めて調査すると有意に事故が増えた**

有意差がない　＝　差がない　（ここが間違い！）

＝　赤信号で右折（日本なら左折）しても事故は増えない

* D. F. Preusser, W. A. Leaf, K. B. DeBartolo, R. D. Blomberg, M. M. Levy. The effect of right-turn-on-red on pedestrian and bicyclist accidents. Journal of Safety Research, 13:45—55, 1982.

# まとめ




- **事前分析：サンプルサイズの設計．そのためには効果量が必要．**
  - 効果量の算出には標本の平均と標準偏差が必要になる．
  - それは事前の予備実験や論文で調べておく必要がある．
  - どんな検定を用いるのか決める必要がある．
  - 倫理委員会に提出する計画書にはそれなりの記述が必要である場合が多い．
- **事後分析：検出力を明らかにする（1－βの計算）**
  - 用いたデータの効果量が必要．Cohen’s dが用いられる．
  - 論文ではp値だけではなく**効果量**と**検出力**の両方の記載が求められる．
- **厳格な統計基準を用いるとサンプルサイズはかなり多く必要**
  - サンプルサイズ＝費用なので大変です．
  - しっかりした設計は無駄な実験や費用を減らすと考えましょう．

**Power Analysisは必須**

# p値に関する12の誤解 *

* S. N. Goodman. A dirty dozen: Twelve p-value misconceptions. Seminars in Hematology, 45:135—140, 2008.

1. $p = 0.05$ ならば，帰無仮説が真である確率は 5% しかない.
2. $p \geq 0.05$ のような有意でない結果は，グループ間に差がないことを意味する.
3. 統計的に有意な発見は客観的に重要である.
4. p値が0.05より大きい研究と小さい研究は矛盾する.
5. p値が同じ研究は帰無仮説に対して同等の証拠を提供する.
6. $p = 0.05$ は、帰無仮説のもとで 5% しか起こりえないデータを観察したことを意味する。
7. $p = 0.05$ と $p \leq 0.05$ は同じことである.
8. p 値は不等式の形で書かれるべきものである(例えば，$p = 0.015$ のときは $p \leq 0.02$ とする).
9. $p = 0.05$ は，帰無仮説を棄却したとしたら，第一種の誤りの確率が 5% しかないことを示す.
10. 有意水準 $p = 0.05$ の下で，第一種の誤りの確率は 5% になる.
11. ある方向を向いた結果やその方向の結果がありえない差異を気に留めないのであれば，片側の p 値を用いるべきである.
12. 科学に関する結論や処置の方針は p 値が有意であるかどうかに基づくべきである.

https://repun-app.fish.hokudai.ac.jp/pluginfile.php/6536/mod_resource/content/1/ダメな統計学.pdf
p.44, 訳者コラム