OpenIntro

第 7 章：数値データの推測

OpenIntro Statistics 第 4 版（日本語版）

原著スライド：Mine Çetinkaya-Rundel（OpenIntro）

13 日の金曜日

1990 年から 1992 年にかけて、英国の研究者たちが 13 日の金曜日とその前週の金曜日（6 日）の交通量、事故件数、入院件数に関するデータを収集した。以下は交通量に関するデータの抜粋である。場所 1 と場所 2 の交通量は独立であると仮定できる。

	種別	日付	6 日	13 日	差	場所
1	交通量	1990 年 7 月	139246	138548	698	場所 1
2	交通量	1990 年 7 月	134012	132908	1104	場所 2
3	交通量	1991 年 9 月	137055	136018	1037	場所 1
4	交通量	1991 年 9 月	133732	131843	1889	場所 2
5	交通量	1991 年 12 月	123552	121641	1911	場所 1
6	交通量	1991 年 12 月	121139	118723	2416	場所 2
7	交通量	1992 年 3 月	128293	125532	2761	場所 1
8	交通量	1992 年 3 月	124631	120249	4382	場所 2
9	交通量	1992 年 11 月	124609	122770	1839	場所 1
10	交通量	1992 年 11 月	117584	117263	321	場所 2

Scanlon, T.J., Luben, R.N., Scanlon, F.L., Singleton, N. (1993), "Is Friday the 13th Bad For Your Health?," BMJ, 307, 1584-1586.

13 日の金曜日

- 13 日の金曜日と 6 日の金曜日で人々の行動が異なるかを調べたい。
- 一つのアプローチは、この 2 つの日の交通量を比較することである。
- H_0：6 日と 13 日の金曜日の平均交通量は等しい。
  H_A：6 日と 13 日の金曜日の平均交通量は異なる。

データセットの各ケースは、同じ場所・同じ月・同じ年に記録された交通量を表している：6 日の金曜日のカウントと 13 日の金曜日のカウントである。この 2 つのカウントは独立か？

仮説

6 日と 13 日の金曜日の平均交通量の差を検定するための仮説はどれか？

(a) $H_0 : \mu_{6th} = \mu_{13th}$
$H_A : \mu_{6th} \neq \mu_{13th}$

(b) $H_0 : p_{6th} = p_{13th}$
$H_A : p_{6th} \neq p_{13th}$

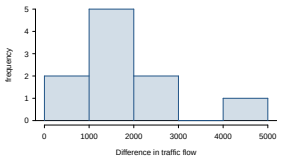
(c) $H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$

(d) $H_0 : \bar{x}_{diff} = 0$
$H_A : \bar{x}_{diff} = 0$

条件の確認

- 独立性：ケース（行）は独立であると仮定する。
- 標本サイズ／歪み：

- 標本分布は極端に歪んでいるようには見えないが、標本サイズが非常に小さいため判断が難しい。母集団分布が歪んでいるかどうかを考える必要がある——おそらく平均より交通量が少ない日と多い日は同程度に起こりうるため、歪んではいないだろう。
- σ は未知であり、n が小さすぎるため s が σ の信頼できる推定値とは言えない。



では、標本サイズが小さい場合はどうすればよいか？

復習：大標本の役割

観測値が独立であり、母集団分布が極端に歪んでいない限り、大標本によって次のことが保証される：

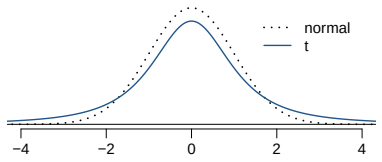
- 平均の標本分布がほぼ正規分布になる
- 標準誤差の推定値 $\frac{s}{\sqrt{n}}$ が信頼できる

正規性の条件

- 中心極限定理（CLT）によれば、母集団分布がほぼ正規であれば、いかなる標本サイズでも標本分布はほぼ正規になる。
- これは有用な特別なケースだが、小標本では正規性の確認が本質的に難しい。
- 小標本における正規性の条件を確認する際は注意が必要である。データを調べるだけでなく、データの出所についても考えることが重要である。
  - 例えば、この分布は対称であると期待できるか、また外れ値はまれであると確信できるか、という点を問うべきである。

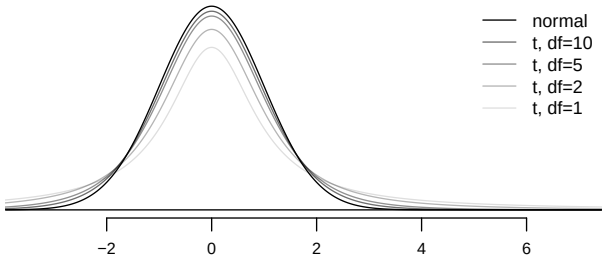
t 分布

- 母集団の標準偏差が未知の場合（ほぼ常にそうであるが）、標準誤差の推定の不確実性は t 分布という新しい分布を使用することで対処される。
- この分布もベル形状をしているが、裾が正規分布より厚い。
- したがって、正規分布と比べて、平均から 2SD を超える位置に観測値が落ちる可能性が高い。
- 裾が厚いことは、標準誤差の推定が信頼性に欠ける場合（n が小さい場合）の問題を解決するのに役立つ。



t 分布（続き）

- 標準正規（z）分布と同様に、常に 0 を中心とする。
- パラメータは 1 つ：自由度（df）。



df が増加すると t 分布の形はどうなるか？

13 日の金曜日に戻ろう

	種別	日付	6 日	13 日	差	場所
1	交通量	1990 年 7 月	139246	138548	698	場所 1
2	交通量	1990 年 7 月	134012	132908	1104	場所 2
3	交通量	1991 年 9 月	137055	136018	1037	場所 1
4	交通量	1991 年 9 月	133732	131843	1889	場所 2
5	交通量	1991 年 12 月	123552	121641	1911	場所 1
6	交通量	1991 年 12 月	121139	118723	2416	場所 2
7	交通量	1992 年 3 月	128293	125532	2761	場所 1
8	交通量	1992 年 3 月	124631	120249	4382	場所 2
9	交通量	1992 年 11 月	124609	122770	1839	場所 1
10	交通量	1992 年 11 月	117584	117263	321	場所 2

$\bar{x}_{diff} = 1836$

$s_{diff} = 1176$

$n = 10$

検定統計量の計算

小標本平均の推測における検定統計量

小標本（$n < 30$）の平均に関する推測の検定統計量は、$df = n - 1$ の T 統計量である。

$$T_{df} = \frac{\text{点推定量} - \text{帰無値}}{SE}$$

この文脈では…

$$\begin{aligned} \text{点推定量} &= \bar{x}_{diff} = 1836 \\ SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \\ T &= \frac{1836 - 0}{372} = 4.94 \\ df &= 10 - 1 = 9 \end{aligned}$$

注：帰無仮説で $\mu_{diff} = 0$ と設定したため、帰無値は *0* である。

p 値の計算

- p 値は、再び t 分布の裾の面積として計算される。
- R を使う場合：

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)

[1] 0.0008022394
```

- ウェブアプリを使う場合：
  https://gallery.shinyapps.io/dist_calc/
- これらが利用できない場合は、t 表を使うことができる。

検定の結論

この仮説検定の結論は何か？

p 値が非常に小さいため、データは 6 日と 13 日の金曜日の交通量に差があることを示す強い証拠を提供していると結論付ける。

差はどれくらいか？

- 6 日と 13 日の金曜日の交通量に差があると結論付けた。
- しかし、その差が具体的にどれくらいかを知ることがより興味深い。
- 信頼区間を使ってこの差を推定することができる。

小標本平均の信頼区間

- 信頼区間は常に次の形をとる：

$$\text{点推定量} \pm ME$$

- ME は常に、臨界値と標準誤差の積として計算される。
- 小標本平均は z 分布ではなく t 分布に従うため、臨界値は $t^\star$（$z^\star$ ではなく）となる：

$$\text{点推定量} \pm t^\star \times SE$$

臨界値 $t^\star$ の計算

R を使う場合：

```
> qt(p = 0.975, df = 9)

[1] 2.262157
```

小標本平均の信頼区間の構築

6 日と 13 日の金曜日の交通量の差に対する 95%信頼区間の正しい計算はどれか？

$$\bar{x}_{diff} = 1836 \qquad s_{diff} = 1176 \qquad n = 10 \qquad SE = 372$$

(a) $1836 \pm 1.96 \times 372$

(b) $1836 \pm 2.26 \times 372$ → *(995, 2677)*

(c) $1836 \pm -2.26 \times 372$

(d) $1836 \pm 2.26 \times 1176$

信頼区間の解釈

先ほど計算した信頼区間の最も適切な解釈はどれか？

$$\mu_{diff:6th-13th} = (995, 2677)$$

95%の確率で…

(a) 6 日と 13 日の金曜日の平均交通量の差は 995 台から 2,677 台の間にある。

(b) 6 日の金曜日の方が 13 日の金曜日より平均して 995 台から 2,677 台少ない車が走っている。

(c) 6 日の金曜日の方が 13 日の金曜日より平均して 995 台少ないから 2,677 台多い車が走っている。

(d) 13 日の金曜日の方が 6 日の金曜日より平均して 995 台から 2,677 台少ない車が走っている。

まとめ

仮説検定の結論は信頼区間の結果と一致しているか？

この研究の結果は、人々が 13 日の金曜日を不吉な日だと信じていることを示唆しているか？

まとめ：t 分布を用いた推測

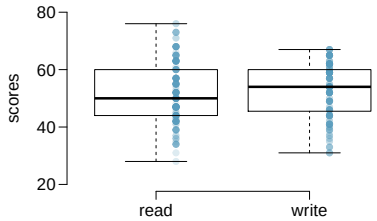
- σ が未知の場合は、$SE = \frac{s}{\sqrt{n}}$ として t 分布を使用する。
- 条件：
  - 観測値の独立性（ランダム標本により確認、非復元抽出の場合は $n <$ 母集団の 10%）
  - 極端な歪みがない
- 仮説検定：

$$T_{df} = \frac{\text{点推定量} - \text{帰無値}}{SE}, \text{ ただし } df = n - 1$$

- 信頼区間：点推定量 $\pm\ t^{\star}_{df} \times SE$

注：ここで使った例は対データの平均（従属グループ間の差）に関するものであった。観測値の差を取り、その差だけ（*1* 標本）を分析に使用したため、*1* 標本の場合と同じ手順となる。

「高校とその先」調査から 200 人の観測値が無作為に抽出された。同じ生徒が読解テストと作文テストを受け、そのスコアを以下に示す。一見して、読解と作文のテストスコアの平均に差があるように見えるか？



同じ生徒が読解テストと作文テストを受け、そのスコアを以下に示す。各生徒の読解スコアと作文スコアは互いに独立か？

	id	読解	作文
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

(a) はい

(b) いいえ

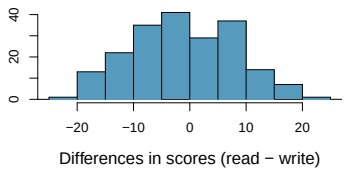
対データの分析

- 2 組の観測値がこのような特別な対応関係（独立でない）を持つとき、それらは対データと呼ばれる。
- 対データを分析するには、各ペアの観測値の結果の差を見ることが有用である。

$$\text{差} = \text{読解} - \text{作文}$$

- 常に一貫した順序で引き算することが重要である。

	id	読解	作文	差
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
⋮	⋮	⋮	⋮	⋮
200	137	63	65	-2



母数と点推定量

- 関心のある母数：すべての高校生の読解スコアと作文スコアの平均差。

$$\mu_{diff}$$

- 点推定量：標本として抽出された高校生の読解スコアと作文スコアの平均差。

$$\bar{x}_{diff}$$

仮説の設定

読解と作文の試験のスコアに実際に差がないとすれば、平均差はどれくらいと期待されるか？

読解と作文の平均スコアに差があるかどうかを検定するための仮説は何か？

H_0: 読解と作文の平均スコアに差はない。

$$\mu_{diff} = 0$$

H_A: 読解と作文の平均スコアに差がある。

$$\mu_{diff} \neq 0$$

新しいことは何もない

- 分析はこれまでと何ら変わりない。
- *1 つの*標本のデータ（差）を持っている。
- 平均差が 0 と異なるかどうかを検定する。

仮定と条件の確認

正しいのはどれか？

(a) 生徒は無作為に抽出され、全高校生の 10%未満であるため、標本中のある生徒の読解スコアと作文スコアの差は別の生徒の差と独立であると仮定できる。

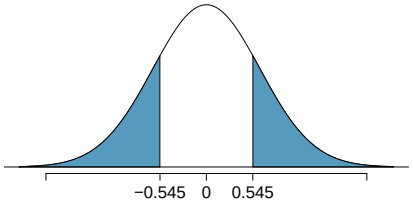
(b) 差の分布が二峰性であるため、仮説検定を続けることができない。

(c) 差をランダムにするには復元抽出で標本を取るべきであった。

(d) 生徒は無作為に抽出され、全生徒の 10%未満であるため、平均差の標本分布はほぼ正規分布になると仮定できる。

検定統計量と p 値の計算

2 つのテストの観測された平均スコア差は -0.545 点、差の標準偏差は 8.887 点であった。$\alpha = 0.05$ として、このデータは 2 つの試験の平均スコアに差があることの説得力ある証拠を提供しているか？



$$\begin{aligned} T &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} \\ &= \frac{-0.545}{0.628} = -0.87 \\ df &= 200 - 1 = 199 \\ p\text{値} &= 0.1927 \times 2 = 0.3854 \end{aligned}$$

p 値 >0.05 であるため、帰無仮説を棄却できない。データは読解と作文の平均スコアに差があることの説得力ある証拠を提供していない。

p 値の解釈

p 値の正しい解釈はどれか？

(a) 読解と作文の平均スコアが等しい確率。
(b) 読解と作文の平均スコアが異なる確率。
(c) 真の平均差が 0 であると仮定したとき、200 人の学生の無作為標本において読解と作文スコアの平均差が（いずれかの方向に）少なくとも 0.545 となる確率。
(d) 帰無仮説が真であるときに誤って帰無仮説を棄却する確率。

仮説検定 ↔ 信頼区間

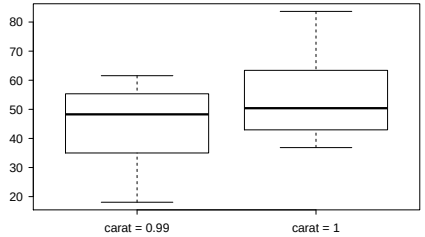
読解と作文スコアの平均差に対する 95%信頼区間を構築するとした場合、この区間は 0 を含むと期待されるか？

(a) はい
(b) いいえ
(c) 与えられた情報からは判断できない

ダイヤモンド

- ダイヤモンドの重さはカラット（carat）で測る。
- 1 カラット＝ 100 ポイント、0.99 カラット＝ 99 ポイント、など。
- 0.99 カラットのダイヤモンドと 1 カラットのダイヤモンドのサイズの違いは肉眼では識別できないが、1 カラットのダイヤモンドの価格は 0.99 カラットのダイヤモンドより高い傾向があるだろうか？
- 0.99 カラットと 1 カラットのダイヤモンドの平均価格に差があるかどうかを検定する。
- 同等の単位で比較できるよう、0.99 カラットのダイヤモンドの価格を 99 で、1 カラットのダイヤモンドの価格を 100 で割り、平均ポイント価格を比較する。

データ



	0.99 カラット	*1* カラット
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

注：このデータは *ggplot2* R パッケージのダイヤモンドデータセットからのランダムサンプルである。

母数と点推定量

- 関心のある母数：すべての 0.99 カラットと 1 カラットのダイヤモンドのポイント価格の平均差。

$$\mu_{pt99} - \mu_{pt100}$$

- 点推定量：標本として抽出された 0.99 カラットと 1 カラットのダイヤモンドのポイント価格の平均差。

$$\bar{x}_{pt99} - \bar{x}_{pt100}$$

仮説

1 カラットのダイヤモンドの平均ポイント価格（$_{pt100}$）が 0.99 カラットの平均ポイント価格（$_{pt99}$）より高いかどうかを検定するための正しい仮説の組合せはどれか？

(a) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} \neq \mu_{pt100}$

(b) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} > \mu_{pt100}$

(c) $H_0 : \mu_{pt99} = \mu_{pt100}$
$H_A : \mu_{pt99} < \mu_{pt100}$

(d) $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$
$H_A : \bar{x}_{pt99} < \bar{x}_{pt100}$

条件の確認

理論的手法でこの仮説検定を行うために満たす必要がない条件はどれか？

(a) 標本中の 0.99 カラットのダイヤモンド同士は独立、1 カラットのダイヤモンド同士も独立でなければならない。

(b) 標本中の 0.99 カラットと 1 カラットのダイヤモンドのポイント価格は独立でなければならない。

(c) 0.99 カラットと 1 カラットのダイヤモンドのポイント価格の分布は極端に歪んでいてはならない。

(d) 両グループの標本サイズはそれぞれ少なくとも 30 でなければならない。

検定統計量

2 つの小標本平均の差に関する推測の検定統計量

σ_1 と σ_2 が未知の場合、2 つの平均の差に関する推測の検定統計量は T 統計量である。

$$T_{df} = \frac{\text{点推定量} - \text{帰無値}}{SE}$$

ただし

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{および} \qquad df = \min(n_1 - 1, n_2 - 1)$$

注：df の計算は実際にははるかに複雑である。手計算の場合は上の式を使って真の df を推定する。

検定統計量（続き）

	0.99 カラット	*1* カラット
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

この文脈では…

$$\begin{aligned} T &= \frac{\text{点推定量} - \text{帰無値}}{SE} \\ &= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}} \\ &= \frac{-8.93}{3.56} \\ &= -2.508 \end{aligned}$$

検定統計量（続き）

この仮説検定の正しい df はどれか？

(a) 22
(b) 23
(c) 30
(d) 29
(e) 52

p 値

この仮説検定の正しい p 値はどれか？

$$T = -2.508 \qquad df = 22$$

(a) 0.005 から 0.01 の間
(b) 0.01 から 0.025 の間
(c) 0.02 から 0.05 の間
(d) 0.01 から 0.02 の間

```
> pt(q = -2.508, df = 22)
[1] 0.0100071
```


まとめ

仮説検定の結論は何か？ ダイヤモンドを購入する際に、この結論はあなたの行動をどのように変えるか？

等価な信頼水準

$\alpha = 0.05$ の片側仮説検定と等価な信頼水準はどれか？

(a) 90%
(b) 92.5%
(c) 95%
(d) 97.5%

臨界値

0.99 カラットと 1 カラットのダイヤモンドの平均ポイント価格の差に対する信頼区間に適切な $t^\star$ はどれか？

(a) 1.32
(b) 1.72
(c) 2.07
(d) 2.82

```
> qt(p = 0.95, df = 22)

[1] 1.717144
```


信頼区間

区間を計算し、文脈の中で解釈せよ。

まとめ：2 つの小標本平均の差を用いた推測

- σ_1 または σ_2 が未知の場合、標本平均の差は $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$ の t 分布に従う。
- 条件：
  - グループ内の独立性（ランダムサンプルで確認、非復元抽出の場合は $n <$ 母集団の 10%）とグループ間の独立性
  - いずれのグループにも極端な歪みがない
- 仮説検定：

$$T_{df} = \frac{\text{点推定量} - \text{帰無値}}{SE}, \text{ ただし } df = \min(n_1 - 1, n_2 - 1)$$

- 信頼区間：

$$\text{点推定量} \pm t^\star_{df} \times SE$$

		決定	
		H_0 を棄却しない	H_0 を棄却する
真実	H_0 が真	$1-\alpha$	第一種の過誤，α
	H_A が真	第二種の過誤，β	検出力，$1-\beta$

- 第一種の過誤は H_0 を誤って棄却することであり、その確率は α（有意水準）である。
- 第二種の過誤は H_0 を棄却すべきときに棄却しないことであり、その確率は β（計算は少し複雑）である。
- 検定の検出力は H_0 を正しく棄却する確率であり、その確率は $1-\beta$ である。
- 仮説検定では α と β を低く保ちたいが、両者の間にはトレードオフが存在する。

第二種の過誤の確率

対立仮説が実際に真である場合、第二種の過誤（棄却すべきときに帰無仮説を棄却しない）を犯す可能性はどれくらいか？

- 答えは自明ではない。
- 真の母集団平均が帰無仮説の値に非常に近い場合、差を検出する（H_0 を棄却する）のが難しい。
- 真の母集団平均が帰無仮説の値から大きく離れている場合、差を検出しやすい。
- 明らかに、β は効果量（δ）に依存する。

例 - 血圧（BP）：仮説

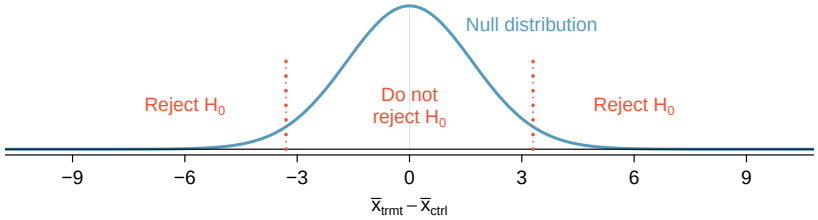
ある製薬会社が血圧を下げる新薬を開発し、臨床試験でその有効性を検証しようとしているとする。特定の標準的な血圧薬を服用している人々を募集し、被験者の半分に新薬（治療群）を投与し、残りの半分には外見が同じプラセボで現在の薬を継続服用させる（対照群）。この文脈での両側仮説検定の仮説は何か？

例 - BP：標準誤差

研究者は収縮期血圧が 140 から 180 mmHg の患者を対象に臨床試験を行うとする。以前発表された研究によれば、患者の血圧の標準偏差は約 12 mmHg で、分布はほぼ対称であるとする。グループごとに 100 人の患者がいる場合、治療群と対照群の標本平均の差の近似標準誤差はどれくらいか？

例 - BP：H_0 を棄却するために必要な最小効果量

5%の有意水準で帰無仮説を棄却するためには、治療群と対照群の観察された血圧平均の差（効果量）はどの値である必要があるか？



差は少なくとも

$$1.96 \times 1.70 = 3.332$$

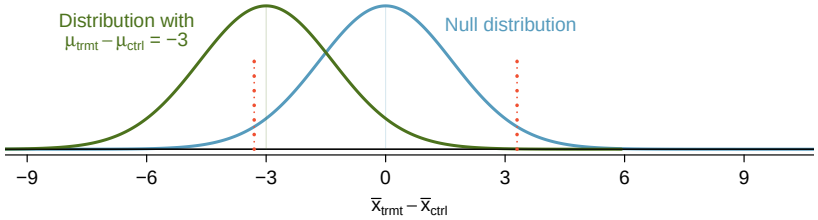
または多くとも

$$-1.96 \times 1.70 = -3.332$$

である必要がある。

例 - BP：検出力

研究者が標準薬と比べて 3 mmHg 以上の血圧への影響を検出することに関心があるとする。この効果を検出できる検定の検出力はどれくらいか？

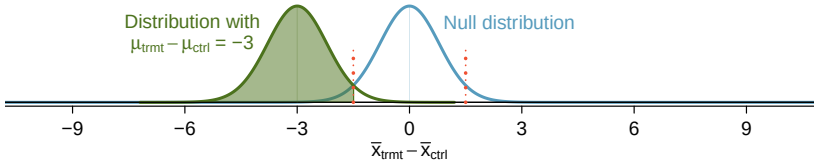


$$Z = \frac{-3.332 - (-3)}{1.70} = -0.20$$

$$P(Z < -0.20) = 0.4207$$

例 - BP：80%の検出力に必要な標本サイズ

この検定で 80%の検出力を得るためにはどれくらいの標本サイズが必要か？



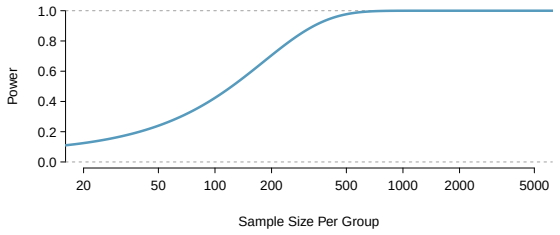
$$SE = \frac{3}{2.8} = 1.07142$$

$$1.07142 = \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

$$n = 250.88 \rightarrow n \geq 251$$

まとめ

- 目標の検出力のレベルに必要な標本サイズを計算する。
- 様々な標本サイズに対して検出力を計算し、目標の検出力（通常 80%または 90%）が得られる標本サイズを選ぶ。



望ましい検出力を達成する方法

検出力を高める（第二種の過誤率を下げる）方法はいくつかある：

1. 標本サイズを増やす。
2. 標本の標準偏差を減らす。これは実質的に標本サイズを増やすのと同じ効果をもたらす（標準誤差が小さくなる）。s が小さいほど、帰無値と観察された点推定量を区別しやすくなる。確保するのは難しいが、慎重な測定プロセスと母集団をより均質に限定することで助けになる場合がある。
3. α を増やすことで H_0 を棄却しやすくする（ただし、これには第一種の過誤率が増加するという副作用がある）。
4. より大きな効果量を考慮する。母集団の真の平均が対立仮説にあるが帰無値に近い場合、差を検出するのが難しくなる。

- テネシー州のウルフ川は、殺虫剤産業がクロルダン（農薬）、アルドリン、ディエルドリン（いずれも殺虫剤）などの廃棄物を捨てていたかつての廃棄場跡地のそばを流れている。
- これらの高毒性有機化合物は、様々な種類のがんや先天性異常を引き起こす可能性がある。
- これらの物質が川に存在するかどうかを検定する標準的な方法は、深さの 10 分の 6 の深さで採取することである。
- しかし、これらの化合物は水よりも密度が高く、分子が堆積粒子に付着する傾向があるため、中深度よりも底部付近でより高い濃度で見られる可能性が高い。

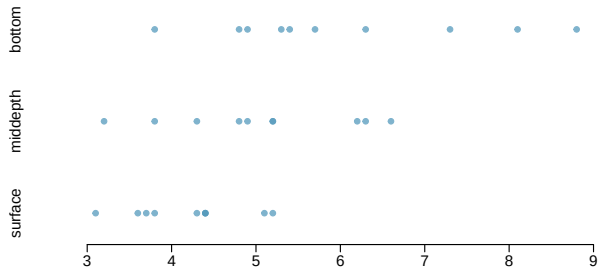
データ

3 つの深さレベルでのアルドリン濃度（ナノグラム/リットル）。

	アルドリン	深さ
1	3.80	底部
2	4.80	底部
…		
10	8.80	底部
11	3.20	中深度
12	3.80	中深度
…		
20	6.60	中深度
21	3.10	表面
22	3.60	表面
…		
30	5.20	表面

探索的分析

3 つの深さレベルでのアルドリン濃度（ナノグラム/リットル）。



	n	平均	標準偏差
底部	10	6.04	1.58
中深度	10	5.05	1.10
表面	10	4.20	0.66
全体	30	5.10	1.37

研究の問い

3 つの深さレベル間でアルドリン濃度の平均に差があるか？

- 2 グループの平均を比較するには Z または T 統計量を使う。
- 3 グループ以上の平均を比較するには、*ANOVA*（分散分析）という新しい検定と F という新しい統計量を使う。

ANOVA（分散分析）

ANOVA は、目的変数の平均がカテゴリ変数の異なるレベルで異なるかどうかを評価するために使われる。

H_0 : すべてのカテゴリで目的変数の平均は同じである。

$$\mu_1 = \mu_2 = \cdots = \mu_k,$$

ここで μ_i はカテゴリ i の観測値の目的変数の平均を表す。

H_A : 少なくとも 1 つの平均が他と異なる。

条件の確認

1. 観測値はグループ内およびグループ間で独立でなければならない。
   - データが母集団の 10%未満の単純無作為標本であれば、この条件は満たされる。
   - データが独立かどうかを注意深く考える（例えば、対になっていないか）。
   - 常に重要だが、確認が難しい場合もある。
2. 各グループ内の観測値はほぼ正規分布に従う必要がある。
   - 標本サイズが小さい場合に特に重要である。

正規性はどのように確認するか？

3. グループ間の変動性はほぼ等しい必要がある。
   - グループ間で標本サイズが異なる場合に特に重要である。

この条件はどのように確認できるか？

z/t 検定 vs. ANOVA - 目的

z/t 検定	*ANOVA*
*2つ*のグループの平均を比較し、観察された差がサンプリング変動によって合理的に説明できないほど離れているかどうかを見る。	*2つ以上*のグループの平均を比較し、観察された差がすべてサンプリング変動によって合理的に説明できないほど離れているかどうかを見る。
$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

z/t 検定 vs. ANOVA - 方法

z/t 検定	ANOVA
検定統計量（比）を計算する。	検定統計量（比）を計算する。
$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$	$F = \frac{\text{群間変動}}{\text{群内変動}}$

- 大きな検定統計量は小さな p 値をもたらす。
- p 値が十分小さければ H_0 は棄却され、母集団平均は等しくないと結論付ける。

z/t 検定 vs. ANOVA

- グループが 2 つだけの場合、検定統計量の分母にプールされた分散を使えば t 検定と ANOVA は等価となる。
- グループが 3 つ以上の場合、ANOVA は標本平均を全体の総平均と比較する。

仮説

3 つの深さレベル間でアルドリン濃度の平均に差があるかを検定するための正しい仮説はどれか？

(a) $H_0 : \mu_B = \mu_M = \mu_S$
$H_A : \mu_B \neq \mu_M \neq \mu_S$

(b) $H_0 : \mu_B \neq \mu_M \neq \mu_S$
$H_A : \mu_B = \mu_M = \mu_S$

(c) $H_0 : \mu_B = \mu_M = \mu_S$
H_A : 少なくとも 1 つの平均が他と異なる。

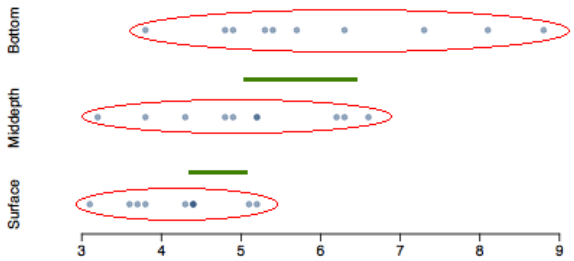
(d) $H_0 : \mu_B = \mu_M = \mu_S = 0$
H_A : 少なくとも 1 つの平均が他と異なる。

(e) $H_0 : \mu_B = \mu_M = \mu_S$
$H_A : \mu_B > \mu_M > \mu_S$

検定統計量

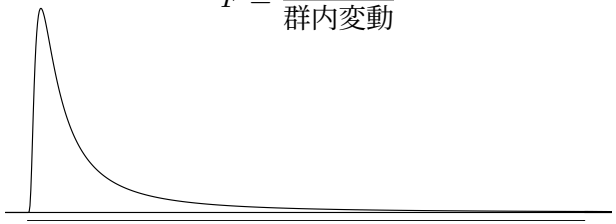
グループ内の変動は大きく見えるか？ グループ間はどうか？

$$F = \frac{\text{群間変動}}{\text{群内変動}}$$



F 分布と p 値

$$F = \frac{\text{群間変動}}{\text{群内変動}}$$



- H_0 を棄却するためには小さな p 値が必要であり、そのためには大きな F 統計量が必要である。
- 大きな F 統計量を得るためには、標本平均間の変動がグループ内の変動よりも大きい必要がある。

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

ANOVA に関連する自由度

- グループ：$df_G = k - 1$，ここで k はグループ数
- 全体：$df_T = n - 1$，ここで n は総標本サイズ
- 誤差：$df_E = df_T - df_G$
- $df_G = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 30 - 1 = 29$
- $df_E = 29 - 2 = 27$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

グループ間の平方和，SSG

グループ間の変動を測る

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

ここで n_i は各グループサイズ，$\bar{x}_i$ は各グループの平均，$\bar{x}$ は全体（総）平均。

	n	平均
底部	10	6.04
中深度	10	5.05
表面	10	4.2
全体	30	5.1

$$\begin{aligned} SSG &= \left(10 \times (6.04 - 5.1)^2\right) \\ &+ \left(10 \times (5.05 - 5.1)^2\right) \\ &+ \left(10 \times (4.2 - 5.1)^2\right) \\ &= 16.96 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

全体の平方和，SST

全体の変動を測る

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

ここで x_i はデータセットの各観測値を表す。

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \cdots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \cdots + (0.1)^2 \\ &= 1.69 + 0.09 + 0.04 + \cdots + 0.01 \\ &= 54.29 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

誤差の平方和，SSE

グループ内の変動を測る：

$$SSE = SST - SSG$$

$$SSE = 54.29 - 16.96 = 37.33$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

平均二乗誤差

平均二乗誤差は平方和を自由度で割って計算される。

$$\begin{aligned} MSG &= 16.96/2 = 8.48 \\ MSE &= 37.33/27 = 1.38 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

検定統計量，F 値

先に述べたように、F 統計量はグループ間変動とグループ内変動の比である。

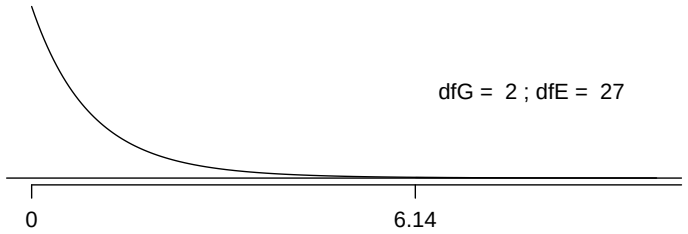
$$F = \frac{MSG}{MSE}$$

$$F = \frac{8.48}{1.38} = 6.14$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

p 値

p 値は、すべてのグループの平均が等しいとした場合に、「グループ間」変動と「グループ内」変動の比が観察された値以上となる確率である。自由度 df_G と df_E の F 曲線の下で観察された F 統計量より右側の面積として計算される。



結論 - 文脈の中で

仮説検定の結論は何か？

データは、アルドリンの平均濃度が

(a) すべてのグループで異なることの説得力ある証拠を提供している。

(b) 表面では他のレベルよりも低いことの説得力ある証拠を提供している。

(c) 少なくとも 1 つのグループで異なることの説得力ある証拠を提供している。

(d) すべてのグループで同じであることの説得力ある証拠を提供している。

結論

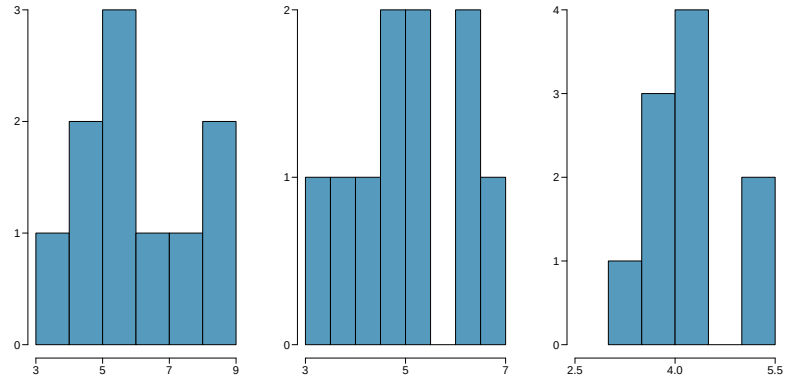
- p 値が小さい（α 未満）場合、H_0 を棄却する。データは少なくとも 1 つの平均が他と異なることの説得力ある証拠を提供している（ただし、どのグループが異なるかは分からない）。
- p 値が大きい場合、H_0 を棄却しない。データは少なくとも 1 対の平均が互いに異なることの説得力ある証拠を提供しておらず、標本平均の観察された差はサンプリング変動（偶然）によるものと考えられる。

(1) 独立性

この条件は満たされているように見えるか？

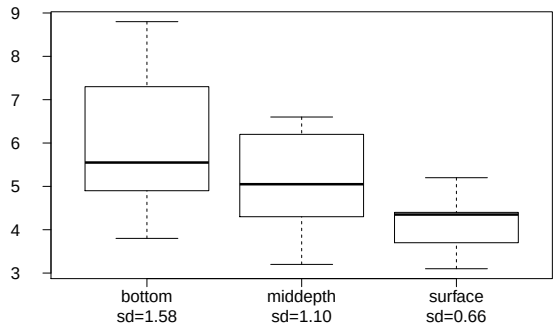
(2) ほぼ正規

この条件は満たされているように見えるか？



(3) 等分散性

この条件は満たされているように見えるか？



どの平均が異なるか？

- 先ほど少なくとも 1 対の平均が異なると結論付けた。自然に続く問いは「どの平均が？」である。
- グループの全ての可能なペアについて 2 標本 t 検定を行うことができる。

このアプローチに潜在的な問題点は見えるか？

- 多くの検定を行うと、第一種の過誤率が増加する。
- この問題は修正された有意水準を使用することで解決される。

多重比較

- 多くのペアのグループを検定するこのシナリオを多重比較（対比較）と呼ぶ。
- ボンフェローニ補正は、これらの検定にはより厳格な有意水準が適切であることを示唆する：

$$\alpha^\star = \alpha/K$$

ここで K は検討している比較の数である。

- k グループがある場合、通常すべての可能なペアが比較され、$K = \frac{k(k-1)}{2}$ となる。

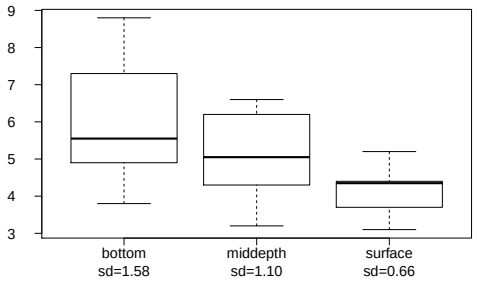
修正 α の決定

アルドリンデータセットでは深さに 3 つのレベル（底部、中深度、表面）がある。$\alpha = 0.05$ の場合、どのペアの平均に有意差があるかを判断する 2 標本 t 検定の修正有意水準はどれであるべきか？

(a) $\alpha^* = 0.05$
(b) $\alpha^* = 0.05/2 = 0.025$
(c) $\alpha^* = 0.05/3 = 0.0167$
(d) $\alpha^* = 0.05/6 = 0.0083$

どの平均が異なるか？

以下の箱ひげ図に基づいて、どの平均が有意に異なると予想されるか？



(a) 底部 & 表面
(b) 底部 & 中深度
(c) 中深度 & 表面
(d) 底部 & 中深度；中深度 & 表面
(e) 底部 & 中深度；底部 & 表面；中深度 & 表面

どの平均が異なるか？（続き）

ANOVA のグループ間で等分散性の仮定が満たされる場合、すべてのグループのデータを使って変動性を推定できる：

- グループ内の標準偏差を $\sqrt{MSE}$（これが s_{pooled}）で推定する
- t 分布には誤差の自由度 $n-k$ を使う

2 つの平均の差：ANOVA 後

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

底部と中深度での平均アルドリン濃度に差があるか？

	n	平均	標準偏差
底部	10	*6.04*	1.58
中深度	10	*5.05*	1.10
表面	10	4.2	0.66
全体	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$\begin{aligned} T_{df_E} &= \frac{(\bar{x}_{\text{底部}} - \bar{x}_{\text{中深度}})}{\sqrt{\frac{MSE}{n_{\text{底部}}} + \frac{MSE}{n_{\text{中深度}}}}} \\ T_{27} &= \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87 \\ 0.05 &< p\text{値} < 0.10 \quad \text{（両側）} \\ \alpha^\star &= 0.05/3 = 0.0167 \end{aligned}$$

H_0 を棄却しない。データは底部と中深度での平均アルドリン濃度に差がある説得力ある証拠を提供していない。

対比較

底部と表面での平均アルドリン濃度に差があるか？