コンピュータ実習（統計処理）

生物情報工学研究室

2021.04.09（金）オンライン

内容

本実習の目的: 実験データの統計処理の基本を身につけること

- **主な統計処理の解説**
- **Excelによる統計処理の実習**

数時間の実習でできることは限られており、また、実習資料に掲載できる内容も限られていますので、各自で必要に応じて学んで下さい。
後の学生実験で使用する多群検定は補足資料として解説しています。

統計ソフトとしてExcelがベストというわけではありません。
RやPythonなどのプログラミング言語では、以下のような利点があります。
- 処理手順をプログラムで指定するため、再現性・再利用性に優れる
- データセルのようなExcel特有の操作に慣れる必要がない
- 高度な解析・グラフ作成が容易に指定できる
- 数多くのパッケージが利用できる

しかし、ここでは学生実験の基本的な処理に利用できるExcelを扱います。

内容

1. **データの分布の把握**
2. 測定による誤差とその扱い
3. 標本の統計
4. 相関と回帰
5. 有意差検定

データの分布の特徴を把握するには

例: 10匹のマウスの体重を計測した結果（単位g）

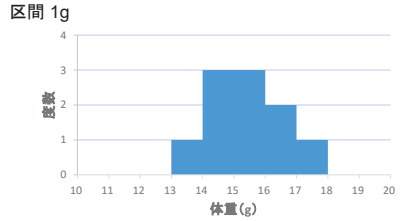
マウス	体重（g）
マウス1	15.5
マウス2	14.2
マウス3	16.3
マウス4	13.9
マウス5	17.4
マウス6	15.1
マウス7	14.7
マウス8	14.4
マウス9	16.6
マウス10	15.8

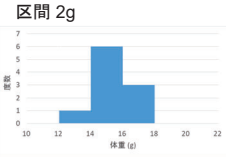
これらのデータの分布の特徴を把握したい
どのように表すことができるか？

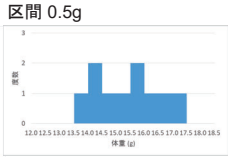
ヒストグラム

- **ヒストグラム**: データの値を階級（区間）に分けて、それぞれの階級の度数を示すグラフ

マウス	体重（g）
マウス1	15.5
マウス2	14.2
マウス3	16.3
マウス4	13.9
マウス5	17.4
マウス6	15.1
マウス7	14.7
マウス8	14.4
マウス9	16.6
マウス10	15.8







データの分布がよくわかる適切な区間を設定することがポイント

要約統計量

要約統計量（基本統計量）：データの分布の特徴を代表する値で示す

- **平均**
- **分散・標準偏差**
- **最頻値（モード）**
  - 離散分布の場合は確率関数が、連続分布の場合は確率密度関数が最大となる確率変数の値
- **中央値（メジアン）**
  - 有限個のデータを小さい順に並べたとき中央に位置する値、連続分布の場合は確率密度関数の面積を均等に分割する確率変数の値
- **四分位数**
  - 有限個のデータを小さい順に並べたとき、1/4, 2/4, 3/4番目に相当する値（4等分した位置の値）をそれぞれ第1四分位数、第2四分位数（中央値）、第3四分位数という

平均

- **平均**
  - データがどのあたりを中心に分布しているかを示す
  - 個々のデータの値を全て足して、その個数で割る

マウス	体重（g）
マウス1	15.5
マウス2	14.2
マウス3	16.3
マウス4	13.9
マウス5	17.4
マウス6	15.1
マウス7	14.7
マウス8	14.5
マウス9	16.6
マウス10	15.8

データの個数 n
データ x_i $(i = 1, 2, \ldots, n)$

平均

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$ → 15.4 g

分散と標準偏差

- データが平均からどれくらい離れているか（どれくらいばらついているか）を示す
- **分散**
  - 個々のデータと平均との差（偏差）の2乗の平均
- **標準偏差**
  - 分散の平方根

マウス	体重（g）
マウス1	15.5
マウス2	14.2
マウス3	16.3
マウス4	13.9
マウス5	17.4
マウス6	15.1
マウス7	14.7
マウス8	14.4
マウス9	16.6
マウス10	15.8

データの個数 n
データ x_i $(i = 1, 2, \ldots, n)$
平均 $\bar{x}$

分散

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$$ → 1.15 g^2

標準偏差

σ → 1.07 g

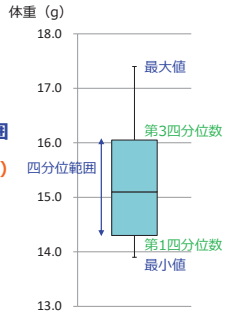
四分位数

例えば、データを
大小順に並べる

		体重（g）
上半分	マウス5	17.4
	マウス9	16.6
	マウス3	16.3
	マウス10	15.8
	マウス1	15.5
	マウス6	15.1
下半分	マウス7	14.7
	マウス8	14.4
	マウス2	14.2
	マウス11	14.0
	マウス4	13.9

第3四分位数
四分位範囲
第2四分位数（中央値）
第1四分位数

箱ひげ図



四分位範囲（interquartile range, **IQR**）

内容

測定における誤差

- 系統誤差（系統的不確かさ）
  - 同じ方法を用いて測定するとき、「真の値」に対して系統的にずれて測定されるような誤差
  - 繰り返しによらず、一定の傾向を保つ
  - 機器誤差: 測定器の特性、使用状態など
  - 理論誤差: 理論・モデルの不正確さ
  - 個人誤差: 測定者の特性、体調など
- 偶然誤差（確率的不確かさ）
  - 測定の際、偶然に生じる誤差
    - 系統誤差に帰すことができないような誤差
  - 「真の値」の周りにばらつく傾向
  - 目盛りの読み方、変動する値の計測などによる

除去が難しい

誤差について

- 誤差とは
  - 誤差 = 測定値 − 真値
  - 相対誤差 = 測定値／真値 − 1 または
    = 1 − 真値／測定値
- 実際には、
  - 誤差 = 測定値と真値と推定される値との差

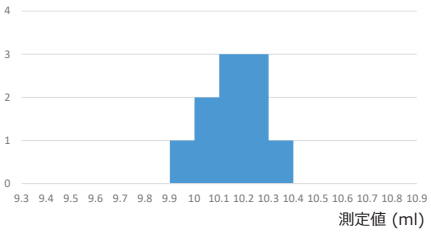
実際の真値はわからない

繰り返し測定の統計

- 測定には誤差を伴う → 繰り返し測定を行い、再現性を確認することが重要

例: ある適定実験の結果（10回の繰り返し実験）

測定回数	体積（ml）
1回目	10.26
2回目	10.02
3回目	10.31
4回目	10.24
5回目	9.95
6回目	10.18
7回目	10.07
8回目	10.23
9回目	10.19
10回目	10.11

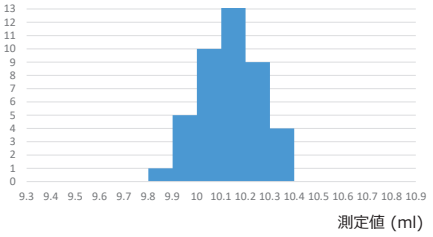


平均　10.16 ml
標準偏差　0.109 ml

繰り返し測定の統計

50回の繰り返し実験の例

10.26	9.96	10.06	10.21	10.33
10.02	10.03	10.23	10.27	10.16
10.31	10.00	10.14	10.16	10.19
10.24	10.15	9.99	10.12	10.17
9.95	10.02	10.16	10.19	10.09
10.18	10.12	10.21	10.08	10.11
10.07	10.18	10.05	9.88	10.07
10.23	10.21	10.31	10.36	10.18
10.19	10.14	10.02	10.13	10.19
10.11	10.18	10.17	10.23	9.97



平均　10.14 ml
標準偏差　0.104 ml

繰り返し測定の統計

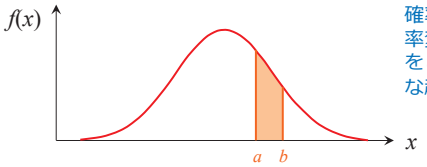
誤差（偶然誤差）を統計的に扱う

- 測定値x_iが、ある確率分布をもつと仮定
- 測定値を連続的確率変数xとし、測定値がある値をとる確率が確率密度関数$f(x)$で表されるとすると、測定値xが$a \le x \le b$となる確率は、

$$P(a \le x \le b) = \int_a^b f(x)dx$$

ただし、$f(x) \ge 0$

$$P(-\infty \le x \le +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$



確率密度関数: 連続的な確率変数xに対して、ある値をとるという事象の相対的な起こりやすさを示す

繰り返し測定の統計

誤差（偶然誤差）を統計的に扱う

測定値は「真の値」（真値）の周りにばらつく傾向

$f(x)$は、真の値に近い部分で大きな値を取り、そこから離れるにしたがい小さくなる傾向にある

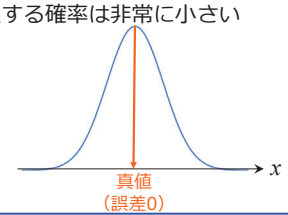
↓

繰り返し回数を増やすと一般に

- 絶対値の等しい正負の誤差は同じ確率で起こる
- 絶対値の小さい誤差は、大きい誤差よりも多く起こる
- 絶対値の非常に大きい誤差の発生する確率は非常に小さい

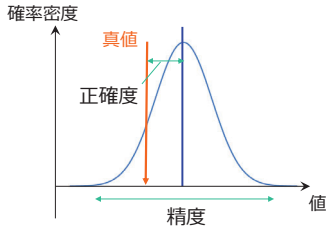
↓

測定値の確率密度は、真値（真値と推定される値）を平均とする正規分布（ガウス分布）に近づく



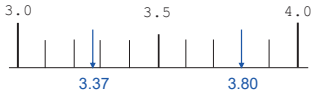
正確度と精度

- **正確度**（正確さ, accuracy）：測定値が真の値からどれだけずれているか
  - 実際には、測定値の平均と真値（真値と推定される値）との差
- **精度**（精密さ, precision）：測定値がどれくらいばらついているか
  - 測定値の標準偏差がよく用いられる



有効数字について

- 測定機器の最小の目盛り間隔の1/10（不確実な位1つ）までを有効として読み取る
  - デジタル表示の場合や特殊な目盛りでは適用されない
- 確実な位すべてと不確実な位1つを有効数字とする



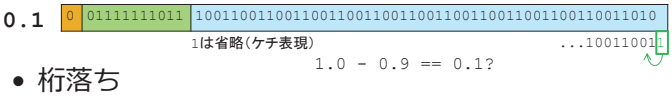
- 複数回の測定がなされた場合、結果の精度は最も精度の低い測定値で決まる
  - 有効桁数は、最も精度の低い測定値と同じ有効桁数
- 有効桁数の調整は、すべての計算が終わったところで行う
  - 計算途中の値を四捨五入したりしない

補足: 計算による誤差

浮動小数点数は有効桁数（仮数部の桁数）が決まっている

$(-1)^s \times (1+m) \times 2^e$

- 丸め誤差
  - 末尾の桁とその下の桁によって一番近い数に置き換える（丸める）ことで生じる



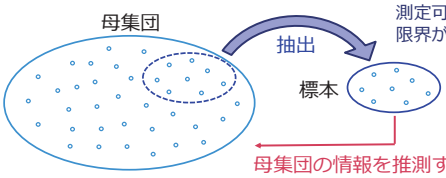
- 桁落ち
  - 値のほぼ等しい数の差の計算で発生
  - 結果がゼロに近くなり、有効桁が失われる
- 情報落ち
  - 指数部の大きく異なる数の和の計算で発生
  - 小さい方の数の仮数部で加えられない部分が生じる

内容

母集団と標本

- 母集団と標本
  - **母集団**: 統計をとるデータのすべて
  - **標本（サンプル）**：母集団から抽出された一部の集団
- 標本を抽出して、母集団の情報を推測する
  - 母集団の全数調査が難しい場合

例えば、
- （無限の）滴定量の測定値を母集団とし、そのうちの10個の滴定量の測定値からなるサンプル
- 血清アルブミン濃度を測定するために全国民から選ばれた健康な成人1000人のサンプル



標本の統計量

標本の抽出を独立かつ等確率でn回行う

標本データの平均（標本平均）$\bar{x}$ → 標本平均は、同一の母平均をもつ母集団からの標本であるが、標本によって異なり、母平均と一致するとは限らない

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

標本データの個数（標本サイズ）n
標本データ $x_i\ (i = 1, 2, \ldots, n)$

→ 標本のサイズnが大きいほど母平均との差が小さくなる傾向

標本データの分散（標本分散）$\tilde{\sigma}^2$ → 標本分散の期待値は、一般に母分散より小さくなる

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

標本データの標準偏差（標本標準偏差）$\tilde{\sigma}$

不偏分散 s^2 母分散σの推定値 ← その差を補正

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

→ 標本のサイズnが大きいほど母分散との差が小さくなる傾向

標本の統計量の計算

Excelの関数で直接計算可能

標本データの平均（標本平均）$\bar{x}$ → AVERAGE(数値1, 数値2, …)

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

標本データの分散（標本分散）$\tilde{\sigma}^2$ → VAR.P(数値1, 数値2, …)

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

$\tilde{\sigma}$ → STDEV.P(数値1, 数値2, …)

不偏分散 s^2 → VAR.S(数値1, 数値2, …)

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

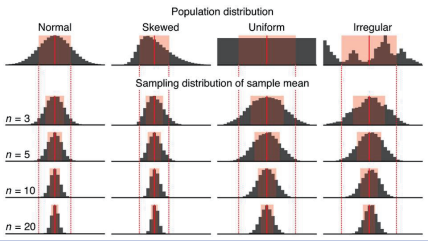
s → STDEV.S(数値1, 数値2, …)

標本		母集団	
標本平均	$\bar{x}$	母平均	μ
標本分散	$\tilde{\sigma}^2$	母分散	σ^2
不偏分散	s^2		

実習では、これらの関数を使わず、数式を入力して計算します

標本分布

- **標本分布**: 標本から計算される統計量の確率分布
- 母平均μ、母分散σ^2の母集団に対して、
  - 標本平均の平均（標本平均の期待値）= 母平均（$= \mu$）
  - 標本平均の分散 = 母分散$/n$（$= \sigma^2/n$）
  - 標本平均の標準偏差 = 母標準偏差$/\sqrt{n}$（$= \sigma/\sqrt{n}$）
    - **標本誤差**（standard error, **SE**）という
- **中心極限定理**: 標本平均の確率分布は、nが大きくなると、母集団の分布に関係なく、正規分布に近づく



標本データ数n が大きくなると「標本平均」のばらつきは小さくなる

n を大きくするほど、正しい母平均の推定が期待できる → 大数の法則

Points of significance:Importance of being uncertain. Nature Methods, 2013; 10(9):809-10.

母平均の推定の信頼区間

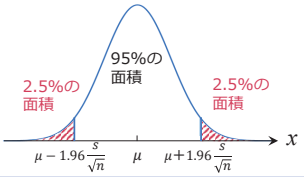
- 標本平均の分布は正規分布に従うと仮定
- さらに、母分散を不偏分散で近似
  - 平均: 母平均μ
  - 分散: 母分散$/n$ （σ^2/n）～ 不偏分散$/n$（s^2/n）→ 近似
  - 標準偏差（標準誤差, SE）：$\sigma/\sqrt{n}$～ $s/\sqrt{n}$
- 標本平均$\bar{x}$が母平均（真値）μ からどのくらいずれる可能性があるか
  → 例えば、標本平均の95%は

$$\mu - 1.96\frac{s}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{s}{\sqrt{n}}$$

標本をとってきて、その平均から95%の信頼区間をとったとき、母平均がそこに含まれるケースが95%ある

- 実際には、標本平均$\bar{x}$に対して真値μが存在すると推定できる範囲（**信頼区間**）を知りたい

$$\bar{x} - 1.96\frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{s}{\sqrt{n}}$$



一般に

$$\bar{x} - z\frac{s}{\sqrt{n}} < \mu < \bar{x} + z\frac{s}{\sqrt{n}}$$

信頼区間68.3%	$z = 1$
信頼区間95%	$z = 1.96$
信頼区間99%	$z = 2.58$
信頼区間99.7%	$z = 2.97$

信頼区間とグラフのエラーバー

例: ある滴定実験の結果（10回の繰り返し実験×5）

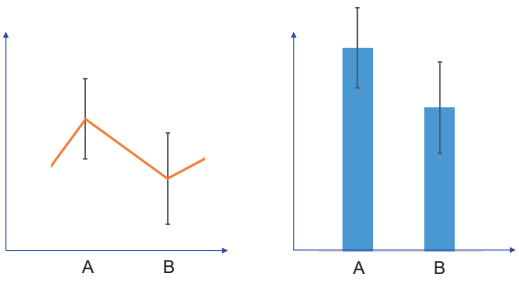
単位: ml

1	2	3	4	5
10.26	9.96	10.06	10.21	10.33
10.02	10.03	10.23	10.27	10.16
10.31	10.00	10.14	10.16	10.19
10.24	10.15	9.99	10.12	10.17
9.95	10.02	10.16	10.19	10.09
10.18	10.12	10.21	10.08	10.11
10.07	10.18	10.05	9.88	10.07
10.23	10.21	10.31	10.36	10.18
10.19	10.14	10.02	10.13	10.19
10.11	10.18	10.17	10.23	9.97

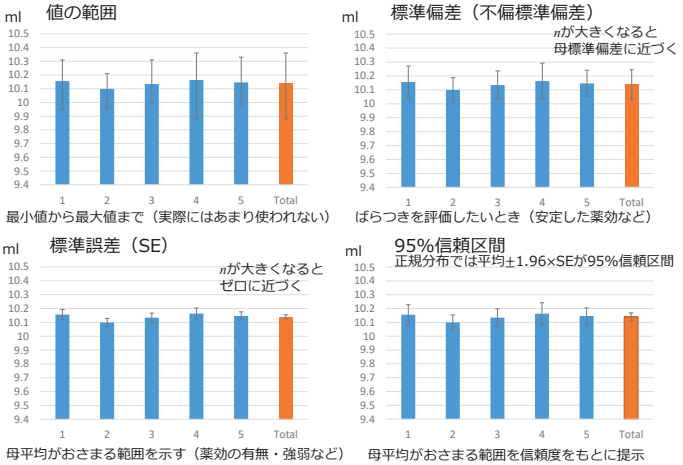
	1	2	3	4	5	全体
平均	10.156	10.099	10.134	10.163	10.146	10.140
不偏標準偏差	0.1151	0.0884	0.1023	0.1281	0.0948	0.1047
標準誤差	0.1092	0.0838	0.0971	0.1215	0.0899	0.1037

グラフのエラーバー

- データの変動性を表す
  - もとのグラフで平均値を表し、エラーバーで測定のばらつき、不確かさなどを表す
- 具体的にどのような値を表しているか、注意が必要



グラフのエラーバー



補足: 有限母集団の補正

- 母集団が有限の場合の補正
- 母平均μ、母分散σ^2の大きさNの母集団
  - 標本平均の平均（標本平均の期待値）$= \mu$
  - 標本平均の分散 $= \boxed{\frac{N-n}{N-1}} \cdot \frac{\sigma^2}{n}$

有限補正項

- 標本の数nに比べて母集団の大きさNが大きくない場合、有限補正項がつく
- Nが十分大きい場合は、有限補正項は1に近づき、無限母集団として扱われる

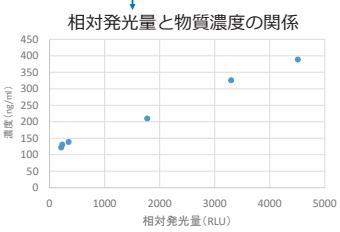
内容

2群のデータの関係を調べる

例: 微生物の生菌数を表すATP濃度を示す相対発光量（RLU）とその微生物が生産するある物質の濃度（ng/ml）との関係

相対発光量（RLU）	物質濃度（ng/ml）
2.12×10^2	122
2.33×10^2	154
3.46×10^2	181
1.97×10^3	210
3.30×10^3	283
4.51×10^3	389

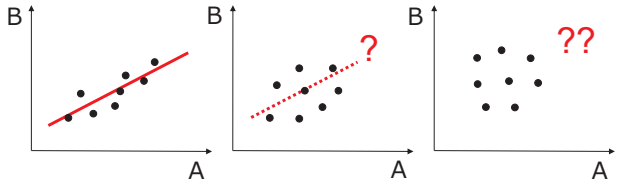
散布図を書く



散布図（scatter plot）：2種類の項目を縦軸と横軸にとり、各データがとる値を点でプロットしたもの

RLU: relative light unit

2群のデータの関係を調べる



2群のデータ間の関係を示すには？

相関係数

相関係数: 2群のデータ間の線形関係の程度を示す指標
ピアソンの積率相関係数を指すことが多い

A群: $x_1, x_2, \ldots, x_n$
B群: $y_1, y_2, \ldots, y_n$ } n 組のデータ

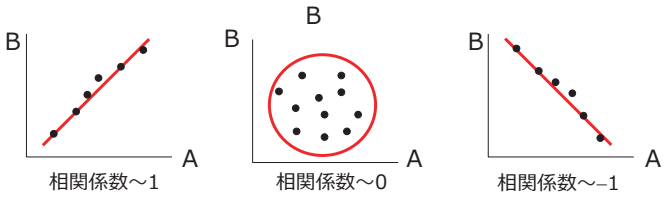
- 平均　$\bar{X} = \sum_{i=1}^{n} x_i/n,\ \bar{Y} = \sum_{i=1}^{n} y_i/n$
- 変動　$D_x = \sum_{i=1}^{n}(x_i - \bar{X})^2,\ D_y = \sum_{i=1}^{n}(y_i - \bar{Y})^2$
- 共変動　$D_{xy} = \sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$
- 相関係数　$r_{xy} = D_{xy}/(\sqrt{D_x}\sqrt{D_y})$

分散を用いて以下のように表すこともできる

- 分散　$\sigma_x^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2/n,\ \sigma_y^2 = \sum_{i=1}^{n}(y_i - \bar{Y})^2/n$
- 共分散　$\sigma_{xy} = \sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})/n$
- 相関係数　$r_{xy} = \sigma_{xy}/(\sqrt{\sigma_x}\sqrt{\sigma_y})$

相関係数

正相関: 相関係数 > 0　　無相関: 相関係数 = 0　　負相関: 相関係数 < 0



平均からのずれ方が似ている → 両者は関係がある

各群のデータを、一定値シフトや定数倍しても、ピアソンの相関係数は不変

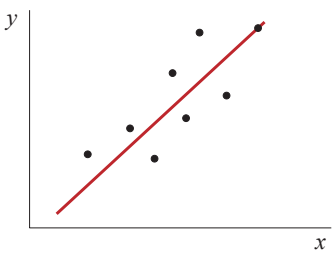
補足: 順位相関係数

- **順位相関係数**: 2群のデータの順位の相関を表す
  - ピアソンの積率相関係数で相関の検定を行うとき、データが正規分布に従うことを仮定
  - 仮定できないとき、**ノンパラメトリックな**（特定のデータの分布を仮定しない）順位相関係数を用いる
- **スピアマンの順位相関係数**: 各データの値を順位に変換してピアソンの積率相関係数を求めたもの
- **ケンドールの順位相関係数**: n 個のデータから2個を取り出したとき、それらの順位がどれだけ一致するかを計算したもの
  - $\tau = (K - L)/{}_nC_2$
    （K: 順位が一致した数、 L: 順位が不一致の数）

回帰分析

- **回帰分析**: 2群のデータ（変数）にどのような関係があるかを推定する統計学的手法

関係を表す数式を**回帰式**という



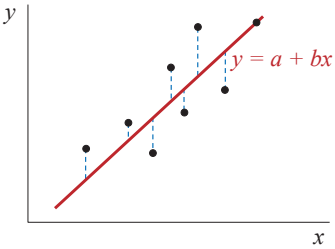
- 目的変数 y
  - 分析によって説明される側の変数
  - 結果となる変数
- 説明変数 x
  - 分析の説明に用いる変数
  - 要因となる変数

- **線形回帰**: 2群のデータに、リニアな関係があると仮定し、それらの間に最もあてはまりの良い直線を計算する

関係の強さを見るのが相関解析、
回帰分析では、その関係の内容まで求める（推定する）

線形回帰

- 変数x, yの間に回帰式 $y = a + bx$ で表される関係があるとする



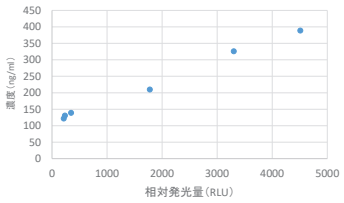
$$a = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i y_i)(\sum_{i=1}^{n} x_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} = \bar{Y} - b\bar{X}$$

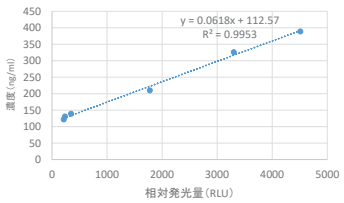
$$b = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i) - n(\sum_{i=1}^{n} x_i y_i)}{(\sum_{i=1}^{n} x_i)^2 - n(\sum_{i=1}^{n} x_i^2)} = \frac{D_{xy}}{D_x} = r_{xy} \cdot \frac{\sqrt{D_y}}{\sqrt{D_x}}$$

$\left(= \text{相関係数} \times \frac{y\text{の標準偏差}}{x\text{の標準偏差}}\right)$

データ点と直線の差の2乗和を最小にすることで求まる（最小二乗法）

Excelによる回帰直線



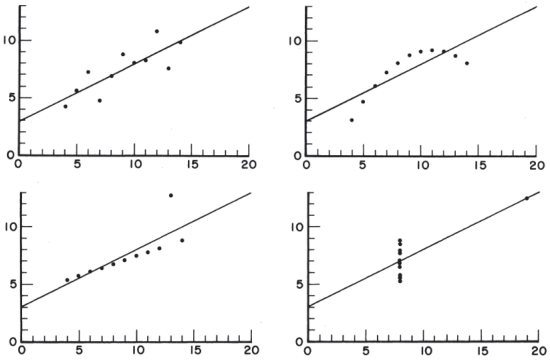


Excelの「近似曲線の追加」で回帰直線を求める（最小二乗法による）

R^2は相関係数の2乗

回帰分析を行う際の注意

アンスコムの例（Anscombe's quartet）



x, yの平均、分散、相関係数、回帰直線がほぼ同じになる

F. J. Anscombe. Graphs in Statistical Analysis. The American Statistician, 1973, 27(1):17-21.

内容

1. データの分布の把握
2. 測定による誤差とその扱い
3. 標本の統計
4. 相関と回帰
5. **有意差検定**

2群のデータの有意差

例: 野生型のマウスおよびある遺伝子のノックアウトマウスの体重増加を、食餌条件を変えた２つのグループで測定

A. 野生型　6.2, 4.7, 5.4, 6.0, 7.3, 9.7, 9.5, 8.0, 9.1, 8.2
B. ノックアウト　7.8, 4.2, 4.9, 3.1, 5.0, 5.3, 6.6, 3.7, 4.5, 3.4

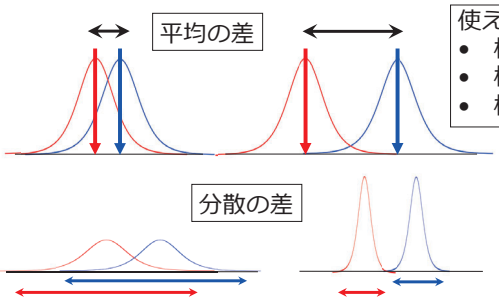
例: ある食品に含まれる錫（スズ）の定量で、試料を塩酸とともに還流煮沸する時間を変えたときの錫の質量（mg kg^{-1}）

A. 煮沸時間30分　55, 57, 59, 56, 56, 59
B. 煮沸時間75分　57, 55, 58, 59, 59, 59

- A群、B群の平均の差が十分に大きければ、A群、B群が抽出された元の母集団の平均に「差がある」としてよい？
- 「十分に大きい」をどのように判定するか？

解析結果は後ほど…

2群のデータの有意差



使えそうな統計データ
- 標本データの平均
- 標本データの分散
- 標本データ数

A群： 27, 33
B群： 20, 18
データ数の差
A群： 27, 33, 32, 29, 30, 38, 31, …
B群： 20, 18, 28, 22, 19, 23, 24, …

標本の平均差大、分散小、データ数大ほど「有意差あり」の可能性が大きくなりそう

t検定

- **t検定**: 母集団の平均（母平均）に差があるかどうかを標本データの数と平均と分散から判断する方法
  - ここでは、**2群のデータ（2標本）のt検定**をいう
- 対象となるデータが正規分布にしたがっている必要
- **t統計量**（正確には「2標本t統計量」、**t値**）を計算

M, N: A, B群のデータ数
$\bar{X}, \bar{Y}$: A, B群の平均
t: t統計量

t検定の検定統計量

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\tilde{V}}{M} + \frac{\tilde{V}}{N}}}$$

← 平均の差
← 合併分散

$\tilde{V}$: 合併分散（pooled variance）

$$\tilde{V} = \frac{(M-1)s_x^2 + (N-1)s_y^2}{M+N-2}$$

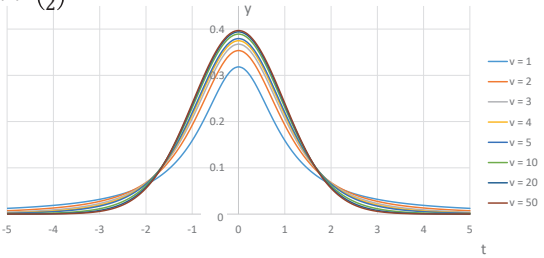
← s_x^2, s_y^2: A, B群の不偏分散
← 検定で用いる自由度

t分布

- t分布（スチューデントのt分布）：t統計量の確率密度分布

$$y = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

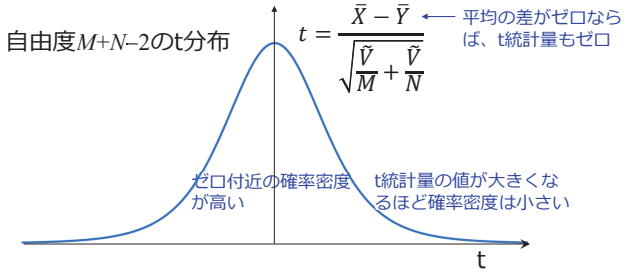
Γ: ガンマ関数, ν: 自由度



自由度νが大きいほど、正規分布に近づく

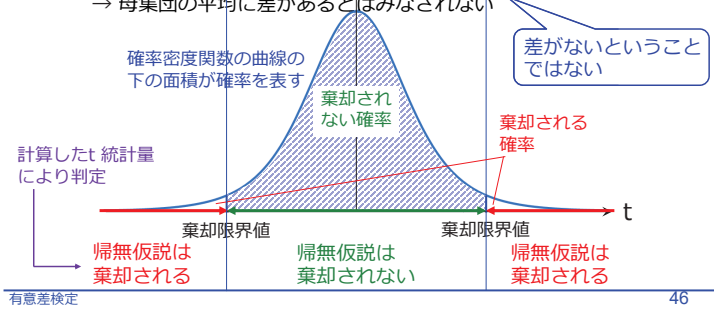
t検定

- t統計量$t = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\tilde{V}}{M}+\frac{\tilde{V}}{N}}}$は「自由度$\nu$=$M$+$N$–2のt分布」にしたがう



t検定

- 母集団の平均に差がないとする ＝ **帰無仮説**
  - 帰無仮説が成り立たないことを示す仮説を**対立仮説**という
- 母集団のデータから t 統計量を計算
- 帰無仮説を棄却するかどうかを t 統計量の値で判定
  - t統計量が棄却限界値を超えたとき、帰無仮説は棄却される
    → 母集団の平均に差があるとみなされる
  - t統計量が棄却限界値を超えないとき、帰無仮説は棄却されない
    → 母集団の平均に差があるとはみなされない

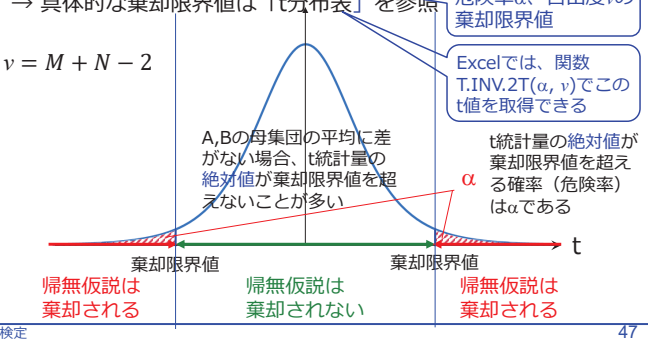


t検定

- **有意水準（危険率）**：棄却するかどうかを判断する基準
- 検定がα水準で有意（危険率α）
  → 棄却限界値を超える確率がα
- 自由度νによりt分布の曲線が決まり、さらに危険率αと合わせて、棄却限界値が決まる
  → 具体的な棄却限界値は「t分布表」を参照

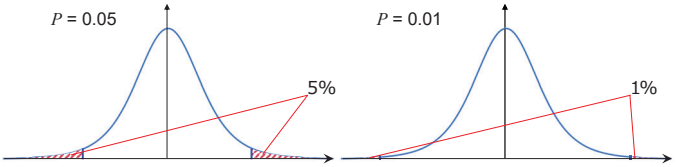
$P = \alpha$の水準で行われた有意差検定
危険率α、自由度νの棄却限界値
Excelでは、関数 T.INV.2T(α, ν)でこのt値を取得できる

$\nu = M + N - 2$



t検定（危険率）

- 危険率5%（有意水準5%, P = 0.05）とは、帰無仮説を棄却（母集団に平均差あり）とした結論が誤り（実際は差がない, 第1種の過誤）である確率が5%
- 5% → 1% とすると、標本にもっと大きな違いがないと「平均差あり」とはいえなくなる。第1種の過誤は減るが、平均差があるのに「平均差があるとはいえない」と結論する誤り（第2種の過誤）は増える
- 危険率5%はよく用いられるが、状況に応じてどの間違いをどのくらい許容するかで危険率を決める



判定	帰無仮説が正しいとき	対立仮説が正しいとき
帰無仮説を棄却	第1種の過誤	正しい判定
帰無仮説を棄却しない	正しい判定	第2種の過誤

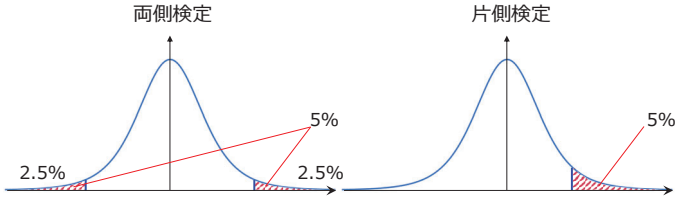
p値

- p値（p-value）：有意確率
  - 帰無仮説の下で測定データから計算された統計量よりも極端な（帰無仮説に反する）統計量が観測される確率
  - 測定データが、帰無仮説通りにならない危険率
  - 単に、測定データから計算された統計量よりも極端な（滅多に起きないことを示す）統計量が観測される確率を指すこともある
  - データの重要性を表すものではない

pはprobabilityの頭文字

t検定（両側検定と片側検定）

- 両側検定: 帰無仮説が成り立たない状態を対立仮説とする
- 片側検定: 帰無仮説の否定の「片側のみ」を対立仮説とする
  - A群とB群の大小が明らか（t統計量の片方の符号側しか判定しない）場合など
  - ただし、帰無仮説が棄却されやすくなる傾向にあり、実際の適用には注意が必要



t検定の流れ（まとめ）

1. 帰無仮説をたてる
   - 母集団の平均に差がないとする
2. 対立仮説をたてる
   - 帰無仮説が成り立たない（母集団の平均に差がある）
3. 有意水準（危険率） αを設定する
4. データ数から自由度νを計算する
5. データからt統計量（t値）を計算する
6. 危険率α、自由度νより棄却限界値が決まる
   - 危険率α、自由度νのt分布曲線による
7. 計算したt値が棄却域に入っていれば帰無仮説を棄却、入っていなければ棄却できないと判断

t検定のExcel関数

Excelの関数で直接計算可能

T.TEST(範囲1, 範囲2, 尾部, 検定の種類)

範囲1	検定の対象となる一方のデータ	
範囲2	検定の対象となるもう一方のデータ	
尾部	片側確率か両側確率か	
	1	片側確率
	2	両側確率
検定の種類	どのような検定を行うか	
	1	対応のあるデータのt検定
	2	2つの母集団の分散が等しい場合のt検定
	3	2つの母集団の分散が等しくない場合のt検定（ウェルチの検定）
戻り値	t分布にしたがう確率	

実習では、この関数を使わず、過程を追って計算します

t検定の例（解答）

例: 野生型のマウスおよびある遺伝子のノックアウトマウスの体重増加を、食餌条件を変えた２つのグループで測定

A. 野生型　6.2, 4.7, 5.4, 6.0, 7.3, 9.7, 9.5, 8.0, 9.1, 8.2
B. ノックアウト　7.8, 4.2, 4.9, 3.1, 5.0, 5.3, 6.6, 3.7, 4.5, 3.4

$M = N = 10 \quad \bar{x} = 7.41, \bar{y} = 4.85 \quad s_x^2 = 3.143, s_y^2 = 2.114 \quad \tilde{V} = 2.629 \quad t = 3.531$

対応する確率 $0.00250 < 0.05$

$\tilde{V} = \frac{(M-1)s_x^2 + (N-1)s_y^2}{M+N-2} \quad t = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\tilde{V}}{M}+\frac{\tilde{V}}{N}}}$

自由度 18, $P = 0.05$の棄却限界値 2.101
5%水準の有意で棄却 → 差があると判断

例: ある食品に含まれる錫（スズ）の定量で、試料を塩酸とともに還流煮沸する時間を変えたときの錫の質量（mg kg^{-1}）

A. 煮沸時間30分　55, 57, 59, 56, 56, 59
B. 煮沸時間75分　57, 55, 58, 59, 59, 59

$M = N = 6 \quad \bar{x} = 57, \bar{y} = 57.83 \quad s_x^2 = 2.8, s_y^2 = 2.567 \quad \tilde{V} = 2.683 \quad t = -0.881$

対応する確率 $0.399 > 0.05$

$\tilde{V} = \frac{(M-1)s_x^2 + (N-1)s_y^2}{M+N-2} \quad t = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\tilde{V}}{M}+\frac{\tilde{V}}{N}}}$

自由度 10, $P = 0.05$の棄却限界値 2.228
5%水準の有意で棄却されない
→ 差があると判断できない

補足: t検定について

- 2群データのt検定では、データが正規分布であることと等分散であることが必要
- 等分散とみなしてよいかどうかは、F検定を用いて判定することができる
  - 帰無仮説を「2群間の分散に差がない（等分散である）」として、棄却されなければ、等分散ではないとはいえない
  - 2群のデータの不偏分散s_x^2, s_y^2は、両者の比は自由度$M-1$, $N-1$の F分布に従う
- 等分散でない（分散が等しくない）とき、Welchのt検定が適用される
  - t統計量は

$$t = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{s_x^2}{M}+\frac{s_y^2}{N}}}$$

s_x^2, s_y^2: A, B群の不偏分散

  - 自由度νは補正された値

$$\nu \approx \frac{\left(\frac{s_x^2}{M}+\frac{s_y^2}{N}\right)^2}{\frac{s_x^4}{M^2(M-1)}+\frac{s_y^4}{N^2(N-1)}}$$

- 検定を複数回行うと、第1種の過誤の確率が大きくなるので注意

補足: t検定について

- 一般に、t検定は、帰無仮説が正しいと仮定した場合に、統計量がt分布にしたがうことを利用する検定法を総称したものをいう
- 1つの母集団（1群データ）の平均が、実際にある値であると判断してよいかという検定
  - 例えば、n個のデータからなる標本の平均が母集団の平均と等しいと考えてよいか
    - 母分散はわからなくても、母集団が正規分布にしたがっている場合、母平均をμ, 不偏分散をs^2、標本平均を$\bar{x}$とすると

$$t = \frac{(\bar{x}-\mu)}{\mathrm{SE}} = \frac{(\bar{x}-\mu)}{s/\sqrt{n}}$$

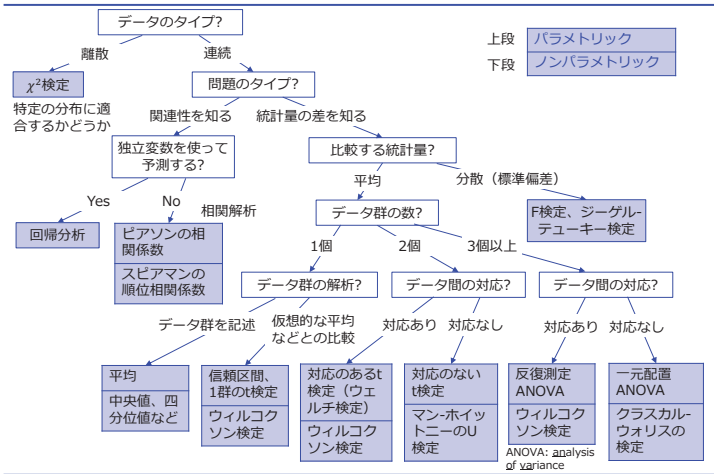
SE: 標準誤差（$= \sigma/\sqrt{n}$）
母分散を不偏分散で近似

    - 自由度$n-1$のt分布に従う
- 対応のある2群データに差があるかどうかの検定
  - n個の対応するデータの差が0であることを帰無仮説とする
    - 標本データの差の平均を$\bar{x}$、差の不偏分散をsとして、上の式（1群データの場合）を適用する

パラメトリック検定とノンパラメトリック検定

- **パラメトリックな手法**: 母集団の特性についてある仮説（母集団が正規分布であることおよび等分散性）を設ける手法
  - 母集団が正規分布にしたがう場合、その平均値、分散を使って検定を行う
- **ノンパラメトリックな手法**: 母集団の分布について一切の仮定を設けない手法
  - 標本サイズが小さい場合には，それから求められた統計量の分布型は不正確なことが多く、パラメトリックな手法を適用することは不適切になりやすい
  - ノンパラメトリックな手法は、名義尺度、順序尺度、間隔尺度、比例尺度に適用される

検定のフローチャート



補足: ノンパラメトリックな検定

- **ウィルコクソンの符号順位検定**: 2群の対応のあるデータ間における代表値（中央値）の差を検定する方法
  - 対応のあるt検定で必要とされる正規分布の仮定が満たされない場合に用いられる
- **マン・ホイットニーのU検定**: 2群の対応のないデータ間における分布の差を検定する方法
  - 正規分布の仮定が満たされない場合に用いられる
  - t検定が使える正規分布のデータでも、その95%程度の検出力が得られる（Mood, 1954）
  - **ウィルコクソンの順位和検定**も同等の検定を行う

補足: 多重比較検定

- 多重比較検定: データ群が3つ以上のときに適用
  - 対照群（比較の基準となる群）と複数の処理群（何らかの処理を行った群）
- すべての2群の組み合わせ一つ一つで2群比較のための検定（例えばt検定）を行ってはいけない
  - 危険率5%で各ペアでt検定を行ったとき、誤って有意差ありとする確率は5%、${}_3C_2$=3ペアの検定で少なくとも1つで誤りが発生する確率は$1 - 0.95^3 \approx 0.143$
  - 群の数が多くなると計算時間がかかる
- 2群の組み合わせの検定すべてにおいて有意差があるという結果のみに意味があるときは、2群比較のための検定を繰り返し行う
  - 既存薬A，Bのそれぞれの効果と配合薬Cの効果を比較し、Cが既存薬A，Bの両方よりも効果があることを示すには、「CがAよりも優れている、かつCがBよりも優れている」ということを示せばよい

補足: 多重比較検定

- パラメトリック検定
  - テューキー-クレーマー（Tukey-Kramer）法, テューキーの範囲検定: 母平均について群間ですべてのペアの比較を行う
  - ダネット（Dunnett）法: 1つの対照群と複数の処理群において、母平均について対照群と処理群のペアの比較のみを行う（多対一）
  - ボンフェローニ（Bonferroni）法: 危険率α、実施する検定の数をNとするとき、各検定の有意水準をα/Nに調整する
    - 対応のあるデータ群にも適用できる
- ノンパラメトリック検定
  - スティール・ドゥワス（Steel-Dwass）法: テューキー-クレーマ法をノンパラメトリックで行う
  - スティール（Steel）法: ダネット法をノンパラメトリックで行う

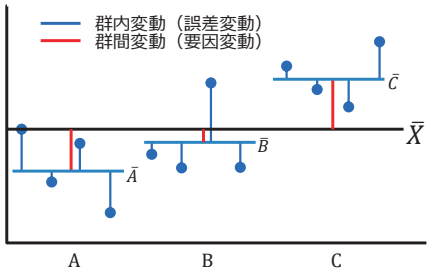
補足: ANOVA

- 分散分析（analysis of variance, **ANOVA**）: 母集団の平均（母平均）に差があるかどうかをデータの分散から判断する手法
  - 帰無仮説は「各群の母平均は等しい」＝「標本が同じ平均値をもつ母集団から取られた」
- 一元配置分散分析
  - 標本群が1つの因子によって分類される（一元配置）として、その因子に含まれる群間の平均値の差を調べる
- 二元配置分散分析
  - 標本群が1つの因子によって分類される（一元配置）として、その因子に含まれる群間の平均値の差、2つの因子の口語作用の有無を調べる

補足: ANOVA

- 群内変動: データとそのデータが属する群の平均との差
- 群間変動: 各群の平均の総平均との差
- $F = \dfrac{\text{群間変動の平方和/群の自由度}}{\text{群内変動の平方和/データの自由度}}$

→ F分布に従う → F検定を用いる



群内変動に対して群間変動が大きくなると、群間になんらかの変動が起きたと判断する（群間比較はできない）

補足: ANOVAの適用例

例: 蛍光試薬溶液を4つの条件で保存したときの溶液の蛍光強度を示したものである。条件による差が偶然誤差によっては説明できないほど大きいか判定したい

条件		反復測定値	平均
A	調製直後	102, 100, 101	101
B	暗所に１時間保存	101, 101, 104	102
C	柔らかい光の所に１時間保存	97, 95, 99	97
D	明るい光の所に１時間保存	90, 92, 94	92
			総平均 98

帰無仮説は、条件による蛍光強度の差は偶然誤差によるものである
群内変動の平方和
Aの変動$^2 = (102-101)^2+(100-101)^2+(101-101)^2 = 2$
同様に、Bの変動$^2 = 6$, Cの変動$^2 = 8$, Dの変動$^2 = 8$
よって、群内変動の平方和 $= 2+6+8+8 = 24$
データの自由度は、データ数 − 群数(平均値の個数) $= 12-4 = 8$
群内変動の平方和 /データの自由度 $= 24/8 = 3$
群間変動の平方和 $= \{(101-98)^2+(102-98)^2+(97-98)^2+(92-98)^2\} \times 3 = 186$
群間変動の平方和 /群の自由度 $= 186/3 = 62$
$F = 62/3 = 20.7$
F分布の$P = 0.05$のときの棄却統計量は4.066、よって帰無仮説は棄却された

補足: テューキー-クレーマー法

- ANOVAで有意差を検定しなくても、F統計量を用いない多重比較（テューキー-クレーマー法、ダネット法、ボンフェローニ法など）では前もってANOVAを実行しなくてもよい
- テューキー-クレーマー法: すべての可能な平均の対を比較し、母集団の平均（母平均）に差があるかどうかを判断する
  - t検定の拡張とみなすことができる
- テューキー-クレーマー法の手順
  1. 帰無仮説群を立てる: すべての２群(i,j)の組み合わせについて母平均は等しい $(H_{(i,j)})$
  2. 群の標本平均$\bar{x_i}$と不偏分散s_i^2を求める
  3. 誤差自由度 $\nu = \sum_{i=1}^{m} n_i - m$および誤差分散 $V_E = \frac{\sum_{i=1}^{m}(n_i-1)s_i^2}{\nu}$を求める
  4. すべての(i,j)の組み合わせについて、t統計量 $t_{ij} = \frac{\bar{x_i}-\bar{x_j}}{\sqrt{V_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$を求める
  5. 誤差自由度 νと危険率αから, t統計量の限界値$q(m,\nu;\alpha) / \sqrt{2}$を求める
     この$q(m,\nu;\alpha)$は「スチューデント化された範囲の上側100α%点」と呼ばれ、m,ν,αに対応する点を巻末の付表2 − 1または2 − 2から読みとることで求められる.
  6. t統計量の絶対値が5.で求めた限界値より大きければ帰無仮説$H_{(i,j)}$を棄却し、「第i群と第j 群が抽出されたもとの母集団の平均には有意差がある」と結論づける

補足: カイ二乗検定

- **カイ二乗検定, χ^2検定**: 帰無仮説が正しければ検定統計量が漸近的にカイ二乗分布に従うような統計的検定法
- 適合度、独立性（関連があるかないか）を検定
- 帰無仮説: 適合しない、独立である（関連がない）
- 離散したデータに適用される
- χ^2値の計算
  - χ^2値 $= \sum \frac{(O-E)^2}{E}$　　O: 観測度数, E: 期待度数
  - 期待度数: 帰無仮説が成り立つとき、期待される度数
  - 期待度数と観測度数の差をもとにχ^2値を計算

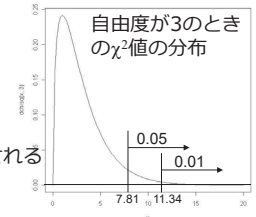
補足: カイ二乗検定の適用例

例: 4つの実験グループの1年間の実験器具の破損数が下記の通りであった。
24, 17, 11, 9
グループによって破損しやすさに差があると言えるか？

帰無仮説: グループによって破損しやすさに差はない
自由度はデータ数 − 1 = 3

観測度数 O	期待度数 E	$O-E$	$(O-E)^2/E$
24	15.25	8.75	5.020
17	15.25	1.75	0.201
11	15.25	−4.25	1.184
9	15.25	−6.25	2.561
合計61		0.00	$\chi^2 = 8.966$

$$f(x)=\frac{x^{\frac{\phi}{2}-1}e^{-\frac{x}{2}}}{\Gamma\left(\frac{\phi}{2}\right)2^{\frac{\phi}{2}}},\quad 0<x<\infty$$
ϕは自由度

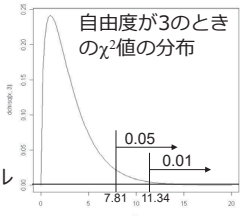


帰無仮説（差がない）は5%の有意水準で棄却される
1%の有意水準では棄却されない

補足: カイ二乗検定の適用例

例: メンデルはエンドウ豆の交配実験で、以下のものを観測した。
しわのない黄色　315個
しわの寄った黄色　101個
しわのない緑色　108個
しわの寄った緑色　32個
合計　556個
メンデルの（分離の）法則から、これらは、9:3:3:1の割合で現れるはず
上の観測はメンデルの法則に合っていると言えるか？

帰無仮説: メンデルの法則に適合する
自由度はデータ数 − 1 = 3
期待度数は、556にメンデルの法則の比率を掛けて、
しわのない黄色　313個
しわの寄った黄色　104個
しわのない緑色　104個
しわの寄った緑色　35個
合計　556個
となるので $\chi^2 = \frac{(315-313)^2}{313}+\frac{(101-104)^2}{104}+\frac{(108-104)^2}{104}+\frac{(32-35)^2}{35}=0.51$



帰無仮説は5%の有意水準で棄却されず、メンデルの法則に適合しないとはいえない

補足: フィッシャーの正確確率検定

- **フィッシャーの正確確率検定, フィッシャーの直接確率検定**: 2つのカテゴリーに分類されたデータの分析に利用される
  - 標本サイズが小さい場合にも適用できるノンパラメトリックな検定
  - カイ二乗検定などのように分布にあてはめるのではなく、直接有意確率を計算する

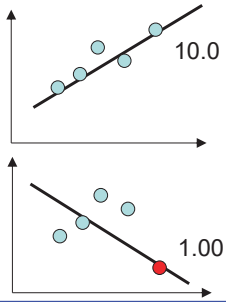
A \ B	B_1	B_2
A_1	x_{11}	x_{12}
A_2	x_{21}	x_{22}

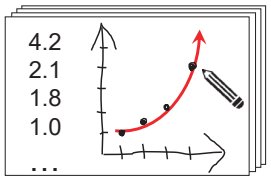
仮説の設定のしかた
- 帰無仮説: AとBは独立である
- 対立仮説: AとBは関連がある

有意確率 $P=\frac{(x_{11}+x_{12})!(x_{21}+x_{22})!(x_{11}+x_{21})!(x_{12}+x_{22})!}{(x_{11}+x_{12}+x_{21}+x_{22})!x_{11}!x_{12}!x_{21}!x_{22}!}$

データ処理の注意

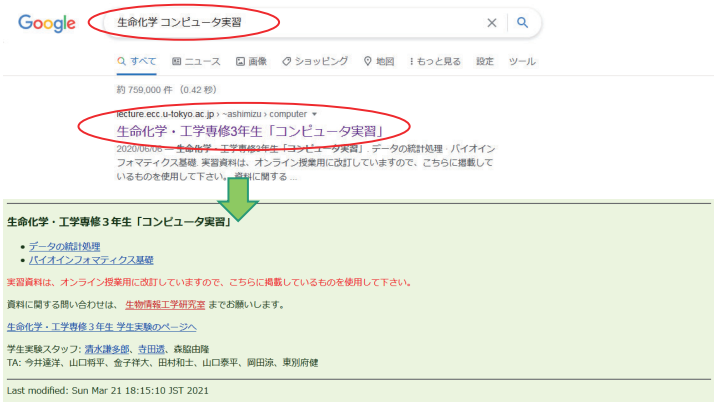
- 適切な統計手法を選べるようにする
- 操作を追うだけでなく、内容も理解する
- 得られたデータが妥当か、注意を払う





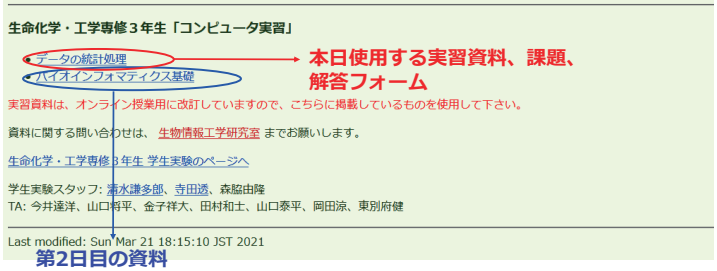
実習の資料について

- 「生命化学 コンピュータ実習」で検索
  - https://lecture.ecc.u-tokyo.ac.jp/~ashimizu/statistics/

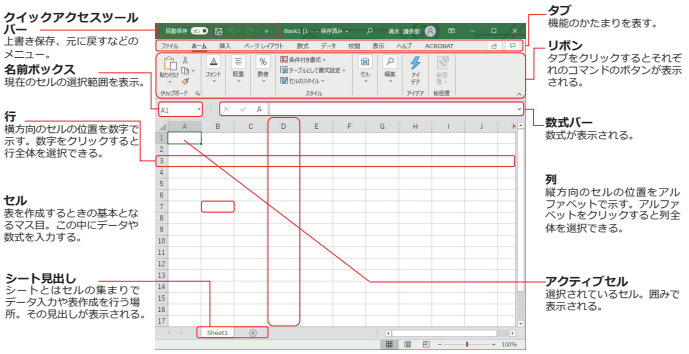


実習の資料について

- 「コンピュータ実習」のトップページ
  - https://lecture.ecc.u-tokyo.ac.jp/~ashimizu/computer/

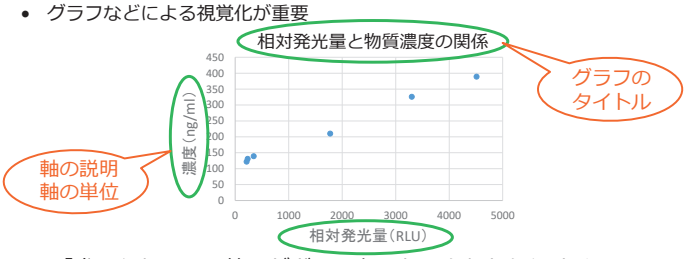


第2日目の資料
実習資料・課題はWebにアップし、配布しません
（第2日目の当日、リンクを設定します）



レポートに関する注意

- 平均・分散などの値に単位を忘れずに
- 有効数字を意識する
  - 各実験ごとの指示にしたがう（学生実験では通常3桁程度？）
  - Excelを用いて計算するときは、途中で丸めるなどの操作は行わず、最後に結果を提示するときに有効数字を意識すればよい
  - 「200」のようなゼロで終わる整数は、有効数字がわかるように指数表記
- グラフなどによる視覚化が重要



- 「求められている答えがどこにあるか」をわかりやすく
- 他人にわかりやすいレポートを
  - 数ヵ月後の自分は「他人」と同じ

提出のしかた

- 以下のURLから、解答用フォームをダウンロード
  - https://lecture.ecc.u-tokyo.ac.jp/~ashimizu/statistics/form.xlsx
- 解答用フォームに課題1～3の解答を入力してGoogle Formsで提出し、チェックを受ける
  - https://docs.google.com/forms/d/1cE9zR22EhJ8wiZk58-_ko4wJ3aBxLhL34SfMpZMVXF4/edit?usp=sharing

質問がありましたら、教員、TA宛てにチャットして下さい
提出した解答はチェックし、zoomのチャットで修正点があれば指摘します
修正を指摘されたら再提出して下さい
全問正解の通知を受け取るまで退室しないで下さい

- 提出する前に確認すること

- 課題の解答が所定の位置に記入されているか
- グラフの軸ラベルと単位を記入しているか
- 数値に（無次元数以外は）単位がついているか
- 有効数字が正しいか

- 実習資料をよく読んでください
- とくに「★重要ポイント」を必ず理解すること
- 疑問点や質問などはスタッフ・TAまで！

では、はじめましょう！