実習試験2
問題１解説

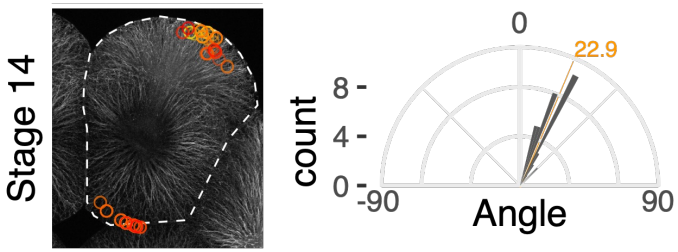
東京都立大学・作題チーム

出題意図

- 生命科学で使用される統計・データ解析の方法論に触れてほしい
  - 検定（何が違うか）、回帰解析やモデル化・モデル選択（何が要因か）
- 一般的に使用されるツールに触れてほしい
  - 画像解析、遺伝子配列などそれぞれにツール
- 今後何を勉強するべきか：「量・数」を使って何を問い、どう説明するか

使用ソフトの選択意図

- Rの使用
  - 膨大なデータのハンドリング：エクセルは100万行しか使えない。生命科学でもすぐ一杯。
  - 統計関数の多彩さ、可視化するツールが豊富であること。フォーマットの自由度が高い
  - 汎用性（Pythonなどのソフトと文法や表記法が類似）
- ImageJ/FIJI
  - 多彩なツール
  - 非常に広い汎用性（さまざまな形式の画像に対応）



将来広く活躍してほしい

第1問

```
5  df = read.csv("Data4.csv")
6  df
7
8  #問 1
9  c1 = c(df[1,1], df[8,1], df[15,1],df[22,1])
10 c1 = df[df$Temperature==24,1]
11 c1
```

- Rでデータを読み込む基本作業ができるか
- 読み込んだデータを参照できるか。
- Data frameを表示させて目視で探しても解答可能。上記のような方法でも可能。
- 細かな入力ミスが目立った。

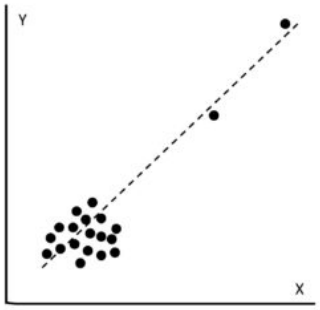
第2問

```r
#問2
df = data.frame(matrix(nrow=4, ncol = 2))
df[1,1] =11.48886
df[2,1] =10.68633
df[3,1] =10.01386
df[4,1] =13.30461
df[1,2] =15.09553
df[2,2] =15.09553
df[3,2] =15.45261
df[4,2] =17.67019
df
t.test(df[,1], df[,2])
```

```r
df2 = data.frame(matrix(nrow=4, ncol = 2))
df2[1,1] = df[1,1]
df2[2,1] = df[8,1]
```

- 生命科学で差があるかどうか見る方法の基本：t検定
- データフレームを作成し、値を入力ができるか（24℃、30℃のデータを抜き出そうとしているか部分点）
- t検定が出来るかどうか（発展的要素）

第3問




```
44 #問3
45 df2 = df[15:21,]
46 df2 = df[df$pH==8,]
47 gg = ggplot(df2, aes(Temperature, Dry_weight)) + geom_point()
48 gg
49
50 plot(x = df2$Temperature, y = df2$Dry_weight)
```

- 相関を判断する際に必要なステップ（外れ値の影響）
- データフレームからのデータの抽出（46行目は発展的な内容）。第2問と同じ方法でも可能。
- グラフの作成は「お知らせ」のとおり。応用できたかどうかは、コマンドの意味や成り立ちを理解しようとしたかどうかで差が出たのではないか。
- XY軸を逆にしても減点しない。打ち間違えも部分点。

第4問：相関係数の計算

```
52 #問4
53 df2 = df[15:21,]
54 df2
55 cor.test(df2[,3], df2[,1])
56
```

```
        Pearson's product-moment correlation

data:  df2[, 3] and df2[, 1]
t = 9.8692, df = 5, p-value = 0.0001821
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8368162 0.9964806
sample estimates:
      cor
0.9752806
```

- データを抽出できるか（統計ソフトのキモ：大量のデータを条件に応じて抽出）
- cor.test関数を使えるか。相関係数の評価ができるか
- 生物学的な考察ができるか：「言えること」の意識が生物の振る舞いに向いているか

第5問

```r
#問5
res = lm(data = df, formula = Dry_weight ~ pH)
summary(res)

```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.3014     2.2543   4.126 0.000336 ***
pH            0.5832     0.2973   1.962 0.060566 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.759 on 26 degrees of freedom
Multiple R-squared:  0.1289,    Adjusted R-squared:  0.09545
F-statistic: 3.849 on 1 and 26 DF,  p-value: 0.06057
```

- 線形回帰分析を行えるか
- Y ~ aX + bの線形モデルが立てられるか（部分点。打ち間違いにより動いてない例）
- 結果が係数の推定を行なっていることが理解できているか（*p*値の理解）

第6問

```r
#問 6
res3 = lm(data = df, formula = Dry_weight ~ Temperature + pH)
summary(res3)

```

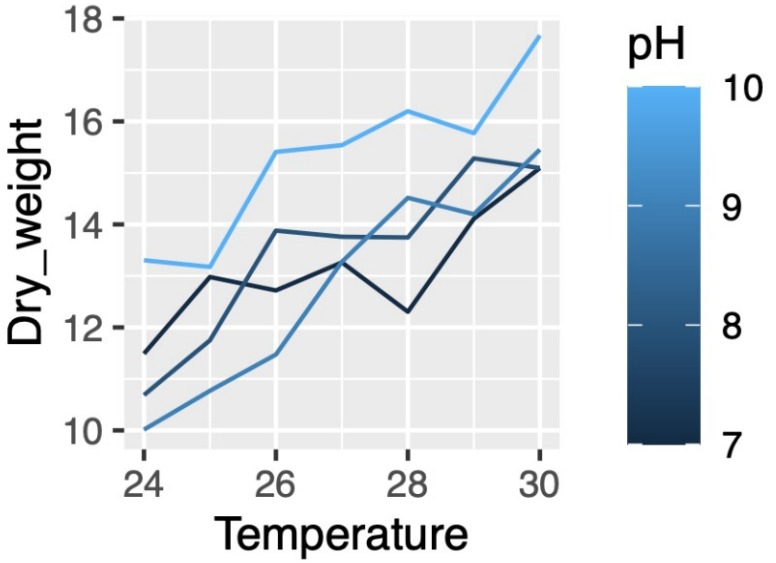
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.53345    2.90738  -3.279  0.00306 **
Temperature  0.69759    0.09622   7.250 1.35e-07 ***
pH           0.58324    0.17212   3.388  0.00233 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.018 on 25 degrees of freedom
Multiple R-squared:  0.7192,    Adjusted R-squared:  0.6968
F-statistic: 32.02 on 2 and 25 DF,  p-value: 1.271e-07
```

- 重回帰分析を行えるか
- $Y \sim aX_1 + bX_2+c$の線形モデルが立てられるか（部分点）
- 結果が係数の推定を行なっていることが理解できているか

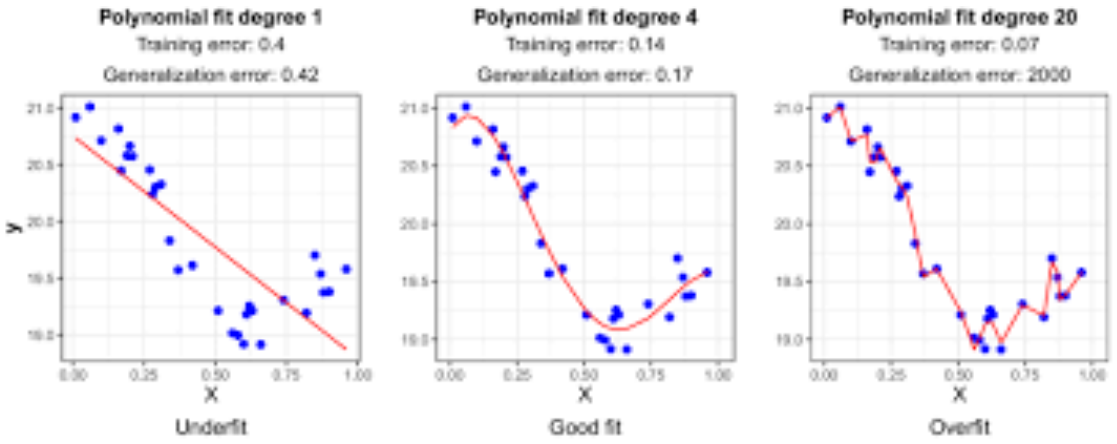
第7問

- 重回帰分析の結果と単回帰分析の結果でどこが違っているか理解できるか：「pHで説明できるか」
- 単純な精度の違いではないことが理解できるか
- 乾燥重量の増加に温度とpHが共に影響している
- データの重なりなども良い説明。
- では、とったデータ全部使ってモデルを立てればいいのか？



モデル選択：説明変数を探る

- 何のためにモデルを立てていたのか：説明変数（原因・要因）を探る
- 細胞の運命　～　遺伝子１＋遺伝子２＋遺伝子３＋・・・・
- 原因を探ったら次は実験により検証する。全部検証するのか？
- Over fittingの問題：次のデータをうまく説明できない。



出題意図とTake home message

- 統計ソフトに馴染んでほしい：文法や操作方法は互いに似ている
- 検定やモデリング・回帰分析による判断という方法論
- 新しいソフトやプログラムを短期間で消化するためには？（自分なり）
- 生物学的な判断や考察から離れない
- More is betterではありませんよ。