慶應女子高　高大連携科目
「理科演習」
2014年10月30日(木)

データから情報を引き出そう

慶應義塾大学 理工学部 数理科学科
南　美穂子
mminami@math.keio.ac.jp

今日の話

ガラゴはどのくらい寝るだろう？

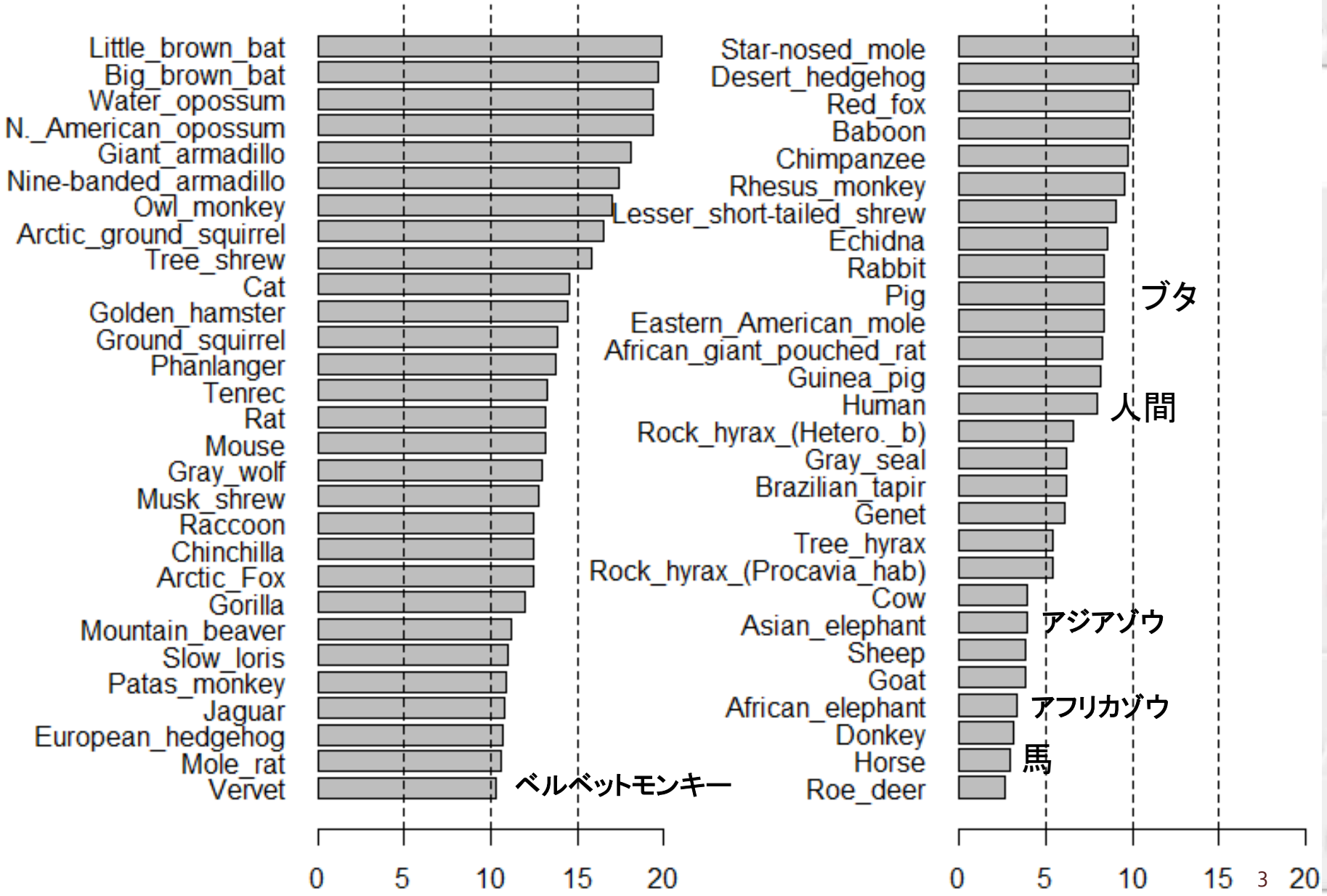
哺乳動物の睡眠時間は体重や妊娠期間などからどのくらい良く予測できるだろうか？

ガラゴ
画像

- データを眺めてみよう
- 睡眠時間と体重や妊娠期間との関係はどのようだろうか
- 睡眠時間を予測するモデルを作ろう
- ガラゴの睡眠時間を予測しよう

哺乳動物の睡眠時間

慶應女子高　高大連携科目
「理科演習」
2014年10月30日(木)



睡眠時間は体重や最高寿命から予測できるか？

Sleep in Mammals: Ecological and Constitutional Correlates"
by Allison, T. and Cicchetti, D. (1976), Science, November 12, vol. 194, pp. 732 -734.

哺乳動物の睡眠時間は他の変数から予測できるだろうか?
哺乳動物各種に対して以下のデータが観測されている

sleep	1日の睡眠時間
body	体重　(kg)
brain	脳の重さ（g)
life	最高寿命（年）maximum life span
gestation	妊娠期間(日)
predation	捕食指標（1～5:　1が最も捕食されにくい)
exposure	睡眠中の危険暴露度(1～5:　1が暴露最小)

統計ソフトRを使ってみよう

- データのダウンロード

  にあるデータを　　にダウンロードして下さい

- 今日の授業では統計ソフトウェア Ｒを使います。
  画面左下のスタートボタンを左クリックして、Ｒを
  クリックしてください

- R Console と書いてある画面で
  以下のように入力してデータを読み込んで下さい

  ```
  > source("C:¥¥MihokoAll¥¥data20141030.txt")
  ```

データフレーム：mammals

	sleep	body	brain	life	gestation	predation	exposure
African_elephant	3.3	6654.000	5712.0	38.6	645	3	5
African_giant_pouched_rat	8.3	1.000	6.6	4.5	42	3	1
Arctic_Fox	12.5	3.385	44.5	14.0	60	1	1
Asian_elephant	3.9	2547.000	4603.0	69.0	624	3	5
Baboon	9.8	10.550	179.5	27.0	180	4	4
Big_brown_bat	19.7	0.023	0.3	19.0	35	1	1
Brazilian_tapir	6.2	160.000	169.0	30.4	392	4	5
Cat	14.5	3.300	25.6	28.0	63	1	2
Chimpanzee	9.7	52.160	440.0	50.0	230	1	1
Chinchilla	12.5	0.425	6.4	7.0	112	5	4
Cow	3.9	465.000	423.0	30.0	281	5	5
Donkey	3.1	187.100	419.0	40.0	365	5	5
Eastern_American_mole	8.4	0.075	1.2	3.5	42	1	1
Echidna	8.6	3.000	25.0	50.0	28	2	2
European_hedgehog	10.7	0.785	3.5	6.0	42	2	2

データの確認

- まず、データフレームmammalsの最初の20行を見よう
  >mammals[1:20,]

- 行4, 列6のデータを見る
  >mammals[4,6]
- 行4を見る
  >mammals[4,]
- 列6を見る
  >mammals[,6]
- あるいは列6の変数名を入力して列6のデータを見る
  >predation
- predationの4番目のデータを見たいときには
  >predation[4]

データを眺めてみよう
どのような特徴があるだろうか

変数　sleep, body, brain, life, gestation の

- ヒストグラムを描いてみよう

  >hist(sleep)
- 平均、中央値、最小値、最大値、四分位数を見よう

  >summary(sleep)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.60    6.95   10.05   10.34   13.20   19.90
```

- >which.max(sleep)　　# sleep の最大値を取る番号
- >mammals[which.max(sleep),]　# sleepの最大値を取る行のデータ
- >mammals[which.min(sleep),]　# sleepの最小値を取る行のデータ

睡眠時間とどのような関係があるか

- 睡眠時間(sleep)と他の変数との散布図を描こう.
  体重 (body) , 脳の重さ (brain) , 最高寿命 (life), 妊娠期間 (gestation), 捕食指標 (predation),
  睡眠中危険暴露度 (exposure)

- 散布図を描く

  >plot(life, sleep)

- データ点の番号を表示する

  >identify(life, sleep)

- データ点の動物名を表示する

  >identify(life, sleep, animalname)

変数間の相関を測る指標

相関係数

- ２つの変数間の線形の相関を測る尺度

$w_i \ (i=1,\cdots,n)$（各動物の睡眠時間）

$z_i \ (i=1,\cdots,n)$（各動物の妊娠期間）

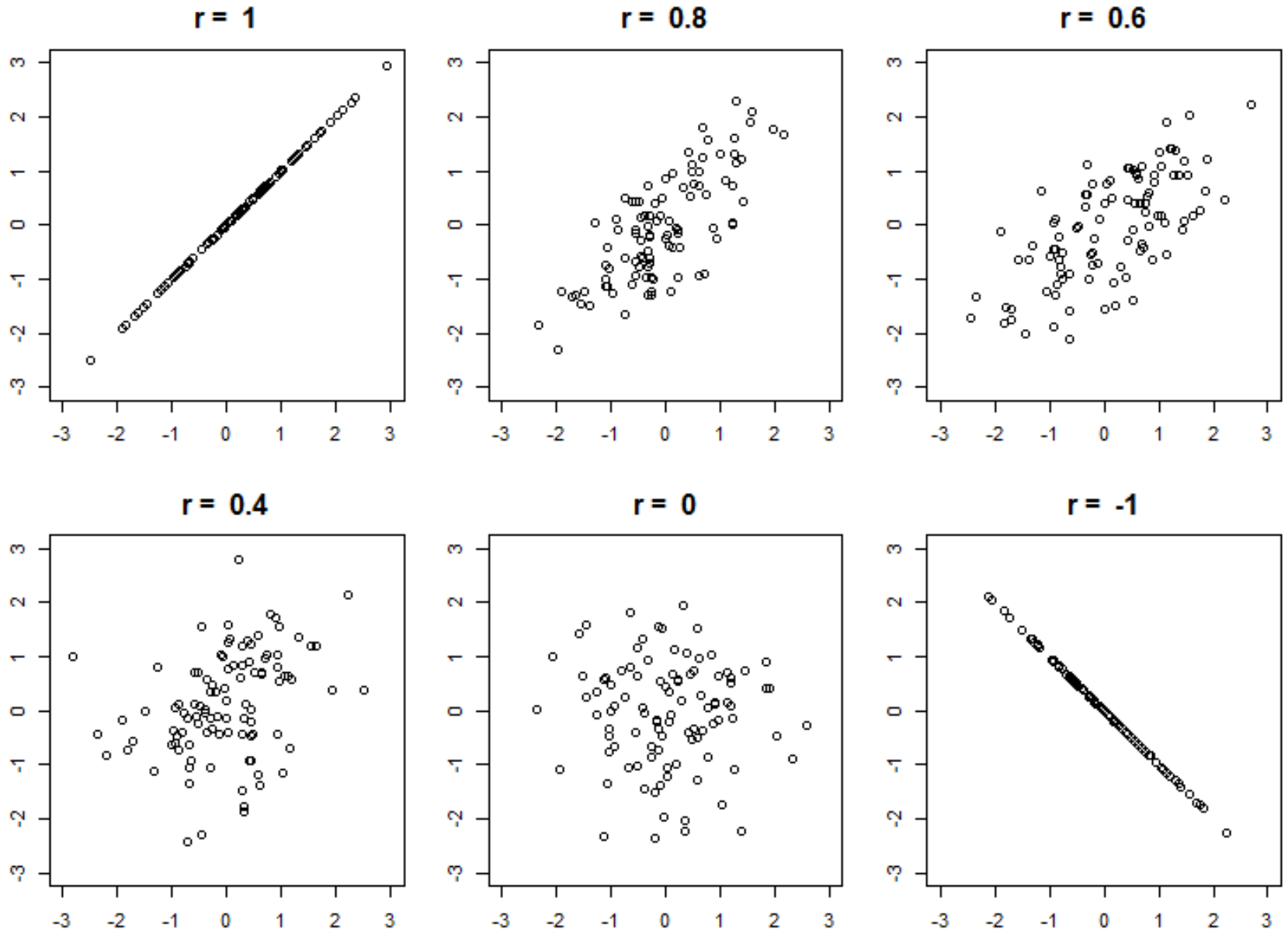
標本平均　$\overline{w}=\frac{1}{n}\sum_{i=1}^{n} w_i$　　$\overline{z}=\frac{1}{n}\sum_{i=1}^{n} z_i$

相関係数

$$r(\mathbf{w},\mathbf{z})=\frac{\sum_{i=1}^{n}(w_i-\overline{w})(z_i-\overline{z})}{\sqrt{\sum_{i=1}^{n}(w_i-\overline{w})^2}\sqrt{\sum_{i=1}^{n}(z_i-\overline{z})^2}}$$

- $-1 \le r \le 1$
- ２つの変数の値が直線上にあり、
  - 一方が大きくなると他方も大きくなるとき、$r=1$
  - 一方が大きくなると他方は小さくなるとき、$r=-1$

相関係数



睡眠時間データの相関係数

- 以下のように入力するとmammalsの変数間の相関係数が計算される

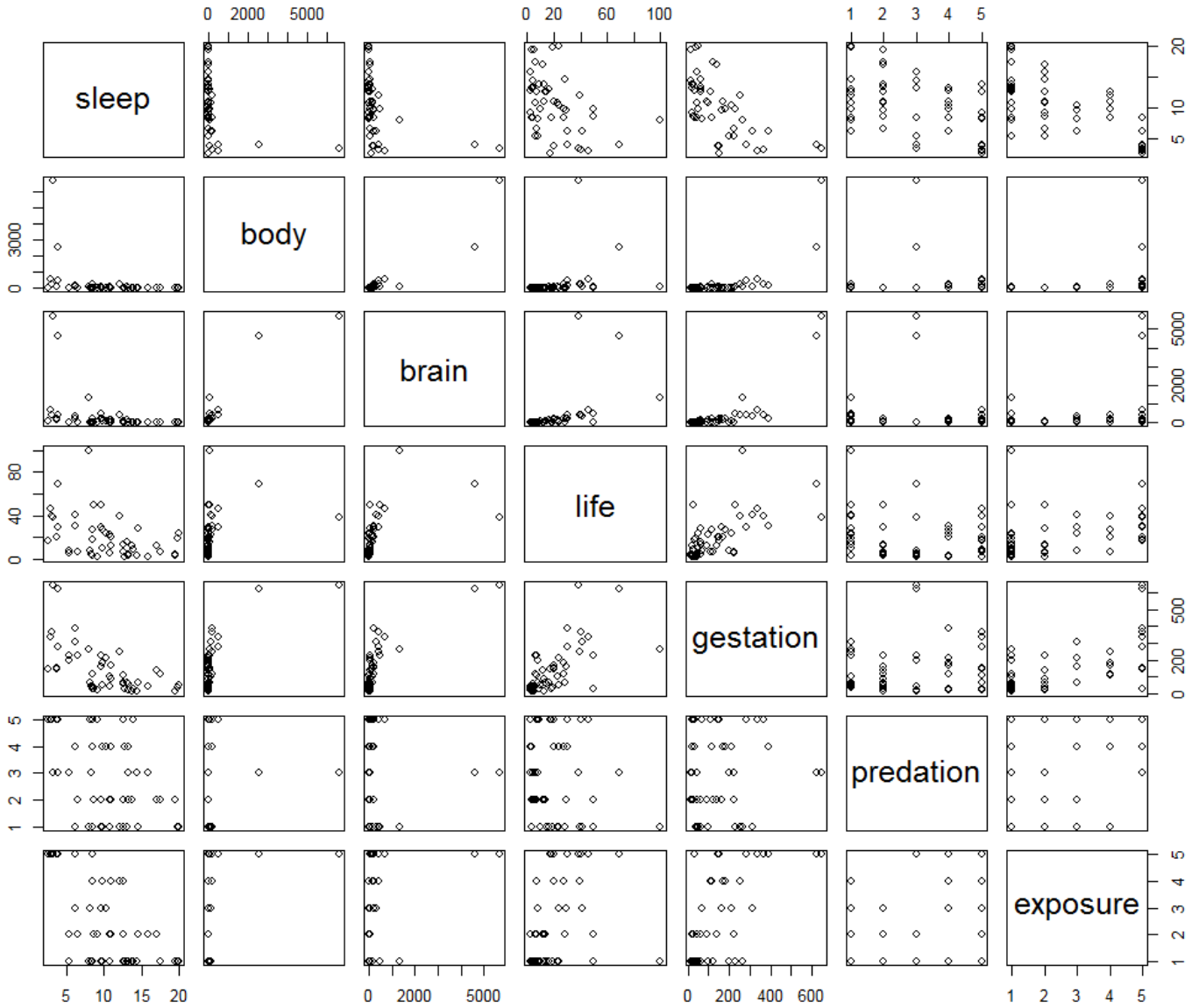
```
>cor(mammals)
```

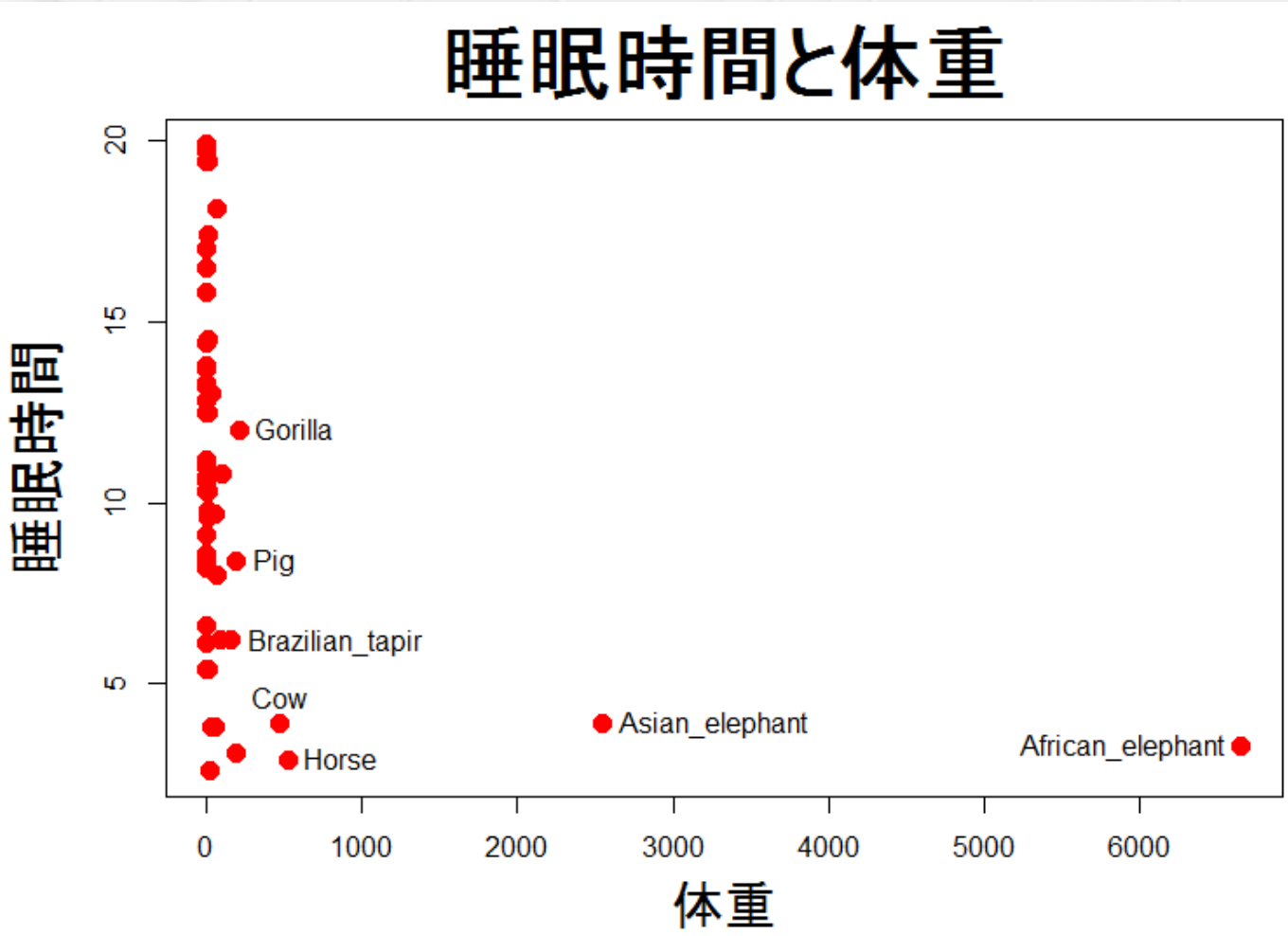
	sleep	body	brain	life	gestation	predation	exposure
sleep	1.0000	-0.31655	-0.368270	-0.3966	-0.6290	-0.461046	-0.6686
body	-0.3165	1.00000	0.933957	0.3005	0.6897	0.046309	0.3504
brain	-0.3683	0.93396	1.000000	0.5106	0.7846	0.009188	0.3717
life	-0.3966	0.30053	0.510562	1.0000	0.6377	-0.132247	0.3580
gestation	-0.6290	0.68971	0.784584	0.6377	1.0000	0.139496	0.6252
predation	-0.4610	0.04631	0.009188	-0.1322	0.1395	1.000000	0.6258
exposure	-0.6686	0.35040	0.371724	0.3580	0.6252	0.625819	1.0000

- 睡眠時間と最も線形相関が強いのはどの変数？
- 線形相関が最も強いのはどの変数間？
- 線形相関が 0 に最も近いのはどの変数間？

- 以下のように入力すると全変数の散布図が描かれる

```
> pairs(mammals)
```



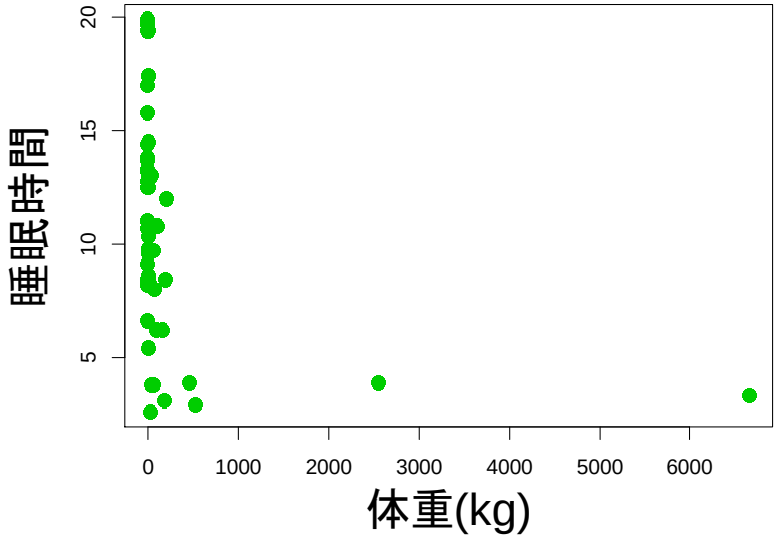


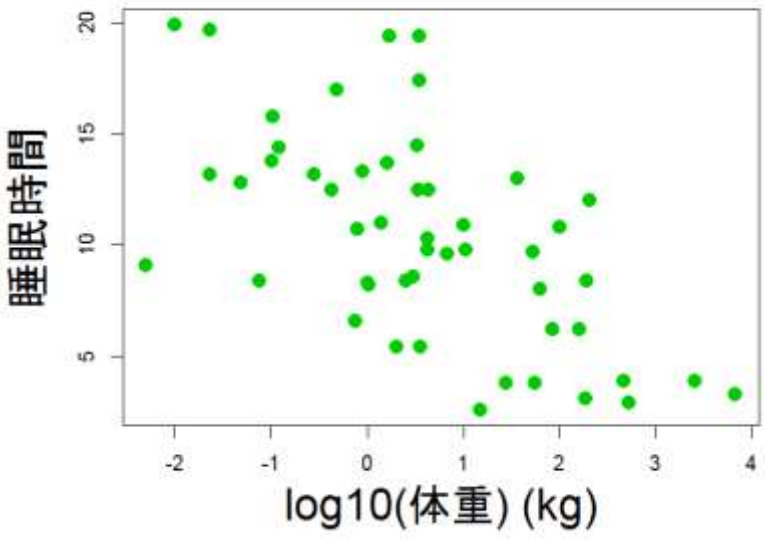
相関係数は小さいけれど、体重が重いほど睡眠時間が短いという傾向はありそう

変数の変換

睡眠時間と体重の 相関係数　= -0.317
睡眠時間と脳の重さの相関係数
= -0.368

適当な単調関数を選んでより
強い線形相関を持つように変換
できる





体重、脳の重さを対数関数で変換すると
睡眠時間と線形な相関を持つようになる

睡眠時間と log10(体重)の相関係数
= -0.611
睡眠時間と log10(脳の重さ)の相関係数
= -0.618

睡眠時間を予測するモデル

被説明変数　y　　（睡眠時間）

説明変数　　$x_1, x_2, \cdots, x_p$　　（最高寿命、妊娠期間など）

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + e$$

例えば

$$\text{睡眠時間} = b_0 + b_1 \times \text{最高寿命} + b_2 \times \text{妊娠期間} + b_p \times \log(\text{体重}) + e$$

◆ 回帰係数　$b_0, b_1, \cdots, b_p$　をどのように推定するか？

◆ どの変数をモデルに含めるか？

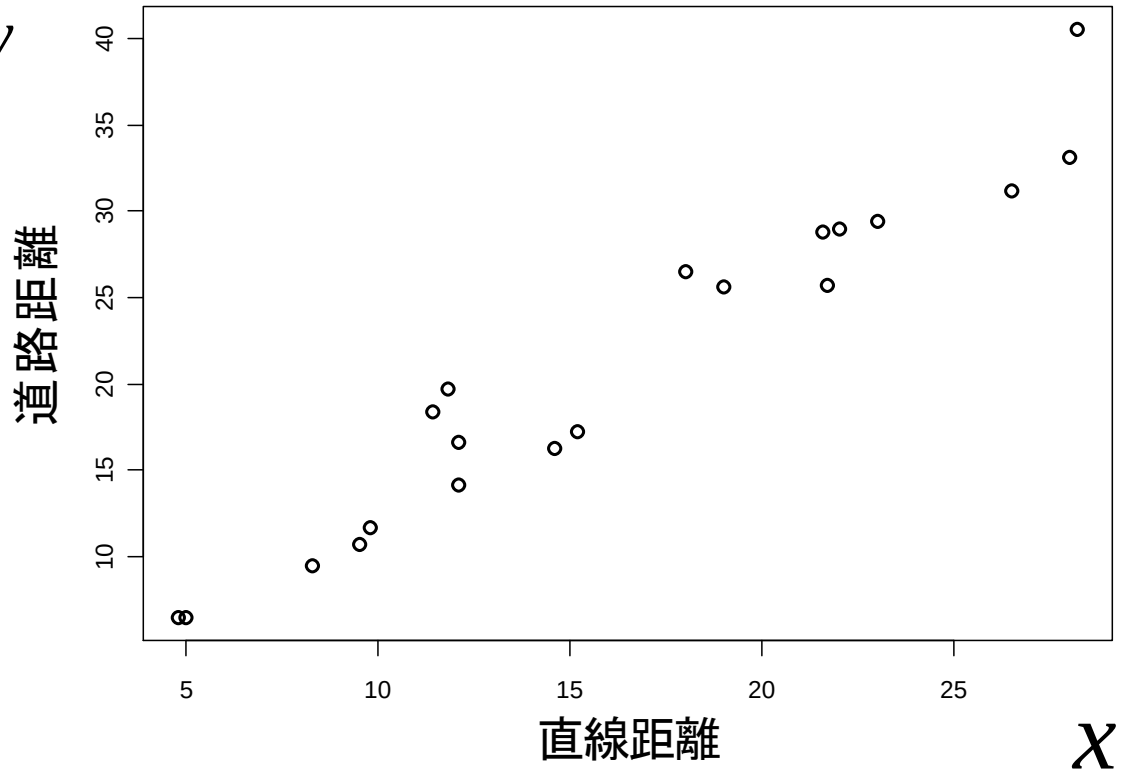
道路距離と直線距離

- 道路距離 y を直線距離 x の関数で表すと？

- 道路距離と直線距離の関係は

$$y_i = b\,x_i + e_i$$
$$(i = 1,2,\cdots,n)$$

b：定数



1単位は25000分1の地図の1cm
つまり距離は1単位=250m

Gilchrist (1984)
Statistical modeling

係数 b の推定

- できるだけ誤差

$$y_i - b\,x_i$$

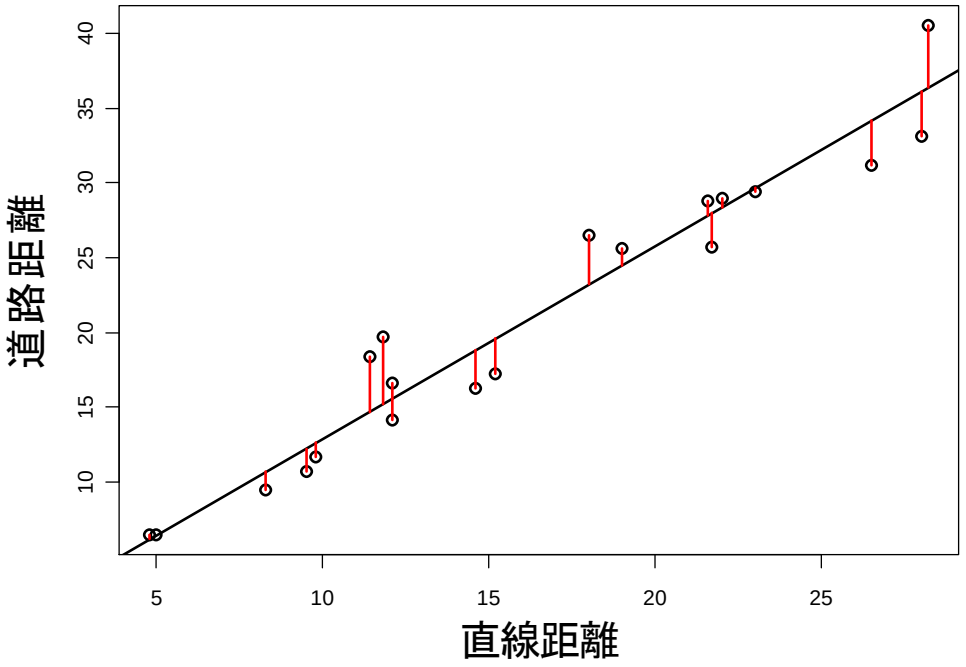
を小さくしたい。

- 個々の誤差を全体で
どのように評価するか？

誤差２乗和

$$\sum_{i=1}^{n}\left(y_i - b\,x_i\right)^2$$

が最小になるような直線を求める



最小2乗法：誤差2乗和の最小化

- 誤差２乗和をｂの関数として見る

$$S(b)=\sum_{i=1}^{n}\left(y_i-b\,x_i\right)^2$$

$$=\sum_{i=1}^{n}\left(y_i{}^2-2b\,x_iy_i+b^2x_i{}^2\right)$$

$$=b^2\sum_{i=1}^{n}x_i{}^2-2b\sum_{i=1}^{n}x_iy_i+\sum_{i=1}^{n}y_i{}^2$$

$$=Ab^2-2Bb+C$$

ここで $A=\sum_{i=1}^{n}x_i{}^2, B=\sum_{i=1}^{n}x_iy_i, C=\sum_{i=1}^{n}y_i{}^2$. また $A>0$ である.

問題： $S(b)$ を最小にする b を求めよう

線形回帰モデル

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + e$$

被説明変数 y ： 睡眠時間 sleep

説明変数 $x_1, x_2, \cdots, x_p$ の候補

- lbody, lbrain
- life, gestation
- predation, exposure

なお　lbody は log10(body), lbrain は log10(brain)

Rによる回帰モデルのあてはめ

例：sleep を life と lbody で説明するモデルをあてはめる

```
> aa = lm(sleep~lbody+gestation)
> summary(aa)

Call:
lm(formula = sleep ~ lbody + gestation)
---
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.802953   0.773434  16.553   <2e-16 ***
lbody       -0.461771   0.255113  -1.810   0.0767 .
gestation   -0.012820   0.005636  -2.275   0.0275 *
---
Residual standard error: 3.628 on 47 degrees of freedom
Multiple R-squared:  0.435,     Adjusted R-squared:  0.411
F-statistic:  18.1 on 2 and 47 DF,  p-value: 1.488e-06
```

Rによる回帰モデルのあてはめ

例：sleep を life と lbody で説明するモデルをあてはめる

```
> aa = lm(sleep~lbody+gestation)
> summary(aa)

Call:
lm(formula = sleep ~ lbody + gestation)
---
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.802953   0.773434  16.553   <2e-16 ***
lbody       -0.461771   0.255113  -1.810   0.0767 .
gestation   -0.012820   0.005636  -2.275   0.0275 *
---
Residual standard error: 3.628 on 47 degrees of freedom
Multiple R-squared:  0.435,     Adjusted R-squared:  0.411
F-statistic:  18.1 on 2 and 47 DF,  p-value: 1.488e-06
```

係数推定値：推定されたモデル

sleep = 12.8 – 0.46 lbody – 0.013 gestation + e

誤差 e の標準偏差の推定値

決定係数 R^2

R^{2*}

モデルの評価基準

推定値： $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$

残差　： $\hat{e}_i = y_i - \hat{y}_i$

- Multiple R-squared（決定係数）R^2
  回帰モデルによって説明される分散の割合

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{e}_i^{\,2}}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- Adjusted R-squared　R^{2*}
  説明変数の数の違いを調整した決定係数

睡眠時間を説明するモデルを探そう

- R^{2*} (Adjusted R-squared)に基づいて睡眠時間をより良く予測するモデルを探そう.
- 今回は以下の手順でモデルを探す.

1. lbody, lbrain の１つだけを説明変数とするモデルのどちらかを選ぶ
2. 1. で選択した変数に life, gestation のどちらかを加えたモデルのどちらかを選ぶ
3. 2. で選択した変数に predation, exposureの1つ、あるいは、両方加えたモデルから最も良いモデルを選ぶ

モデルの選択:ステップ1

1. lbody, lbrain の１つだけを説明変数とするモデルのどちらかを選ぶ

```
> a1 = lm(sleep~lbody)
> summary(a1)
------
Multiple R-squared:  0.3728,     Adjusted R-squared:  0.3598
------
> a2 = lm(sleep~lbrain)
> summary(a2)
------
Multiple R-squared:  0.3819,     Adjusted R-squared:  0.369
------
```

⇒ lbrain を選択する

モデルの選択:ステップ2

2. 1.で選択した lbrain に life, gestation のどちらかを加えたモデルのどちらかを選ぶ

```
> b1 = lm(sleep~lbrain+life)
> summary(b1)
------
Multiple R-squared:  0.3875,     Adjusted R-squared:  0.3614
------
> b2 = lm(sleep~lbrain+gestation)
> summary(b2)
------
Multiple R-squared:  0.4384,     Adjusted R-squared:  0.4145
------
```

⇒ lbrain を選択する

モデルの選択:ステップ3

3. 2.で選択した lbrain と　gestation に predation, exposureの1つ、あるいは、両方加えたモデルから最も良いモデルを選ぶ

```
> c1 = lm(sleep~lbrain+gestation+predation)
> c2 = lm(sleep~lbrain+gestation+exposure)
> c3 = lm(sleep~lbrain+gestation+predation+exposure)
```

当てはめた結果を図に描く

当てはめ結果をffに保存したとき

◘ 睡眠時間の推定値と観測値の散布図を描く

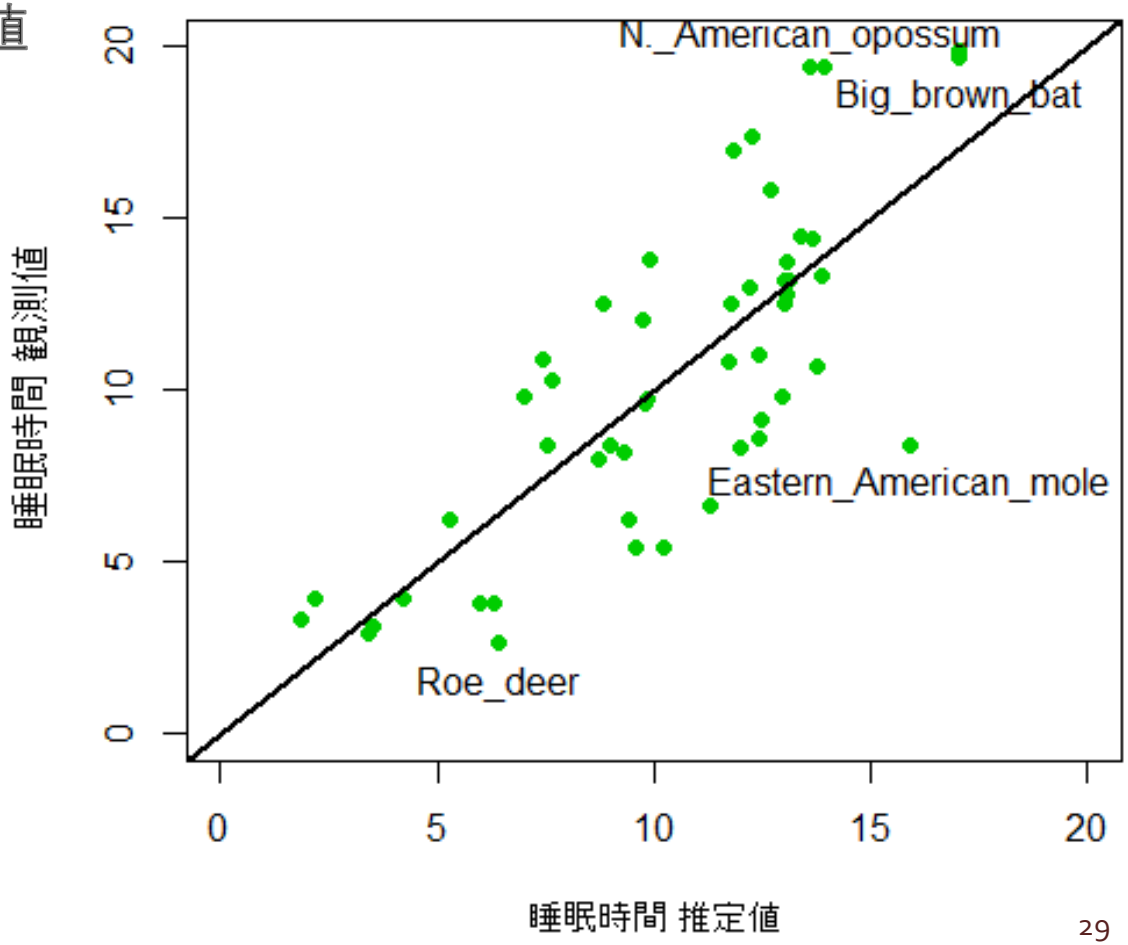
```
> plot(ff$fitted, sleep)
> plot(ff$fitted, sleep, xlab=‘睡眠時間 推定値’, ylab=‘睡眠時間 観測’,
  col=3,pch=16)
   abline(0,1, lwd=2)
```

◘ 推定値と誤差の散布図を描く

```
> plot(ff$fitted, ff$resid, xlab=‘睡眠時間 推定値’, ylab=‘睡眠時間 誤差’,
  col=4,pch=15)
> abline(h=0, lwd=2)
```

選択したモデルの当てはめ結果

- 睡眠時間 = 17.724ー1.763×log10(脳の重さ) ー0.0083 ×妊娠期間
  ー1.310×捕食度
- 誤差の標準偏差推定値
  3.068



ガラゴの睡眠時間を予測しよう

- ガラゴの情報：

体重(body)	0.2 kg	捕食指標	2
脳の重さ(brain)	5 g	睡眠中の危険暴露度	2
最高寿命	10.4年	危険度	2
妊娠期間	120日		

- ガラゴの情報はデータフレーム Galago に入っている
モデルの当てはめ結果を ff とすると

```
> predict (ff, newdata=Galago)
```

とするとこのモデルでの予測値が表示される

ガラゴの睡眠時間予測値　=

誤差の標準誤差の推定値　=　3.068

ガラゴの睡眠時間は？

- ガラゴの睡眠時間

10.7 時間

統計学は、

- データから情報を引き出すための方法の科学です.
- 学際的な学問で、医学、経済学、生物学、環境科学などあらゆる分野で応用されています.