

Introduction to Statistics

統計学入門

Week 11 | December 8, 2022

Take the Week 11 survey

Week 11 : 2変数の関連性




Yoh Kawano さんが新しい資料を投稿しました: Week 11 Lecture



投稿日: 昨日 (最終編集: 11:40)




Week 11 Class Survey
Google フォーム






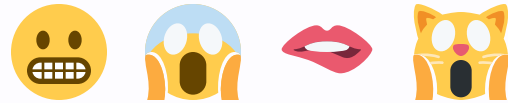
W11.pdf
PDF



クラスのコメントを追加...



Week 10 小テスト



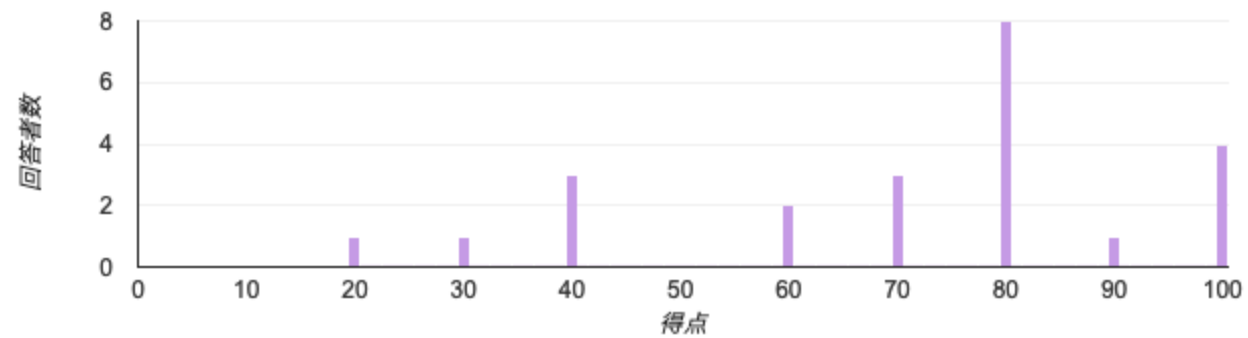
分析情報

平均
70.87/100 ポイント

中央値
80/100 ポイント

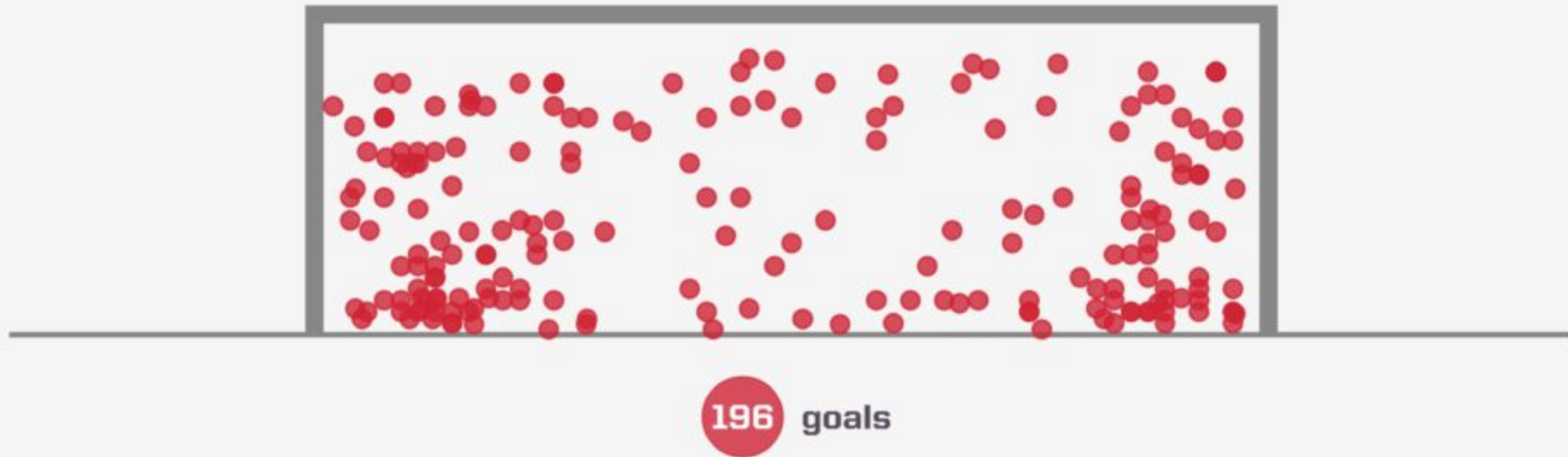
範囲
20~100 ポイント

合計点の分布



World Cup Shootout Penalties Scored

World Cups since 1982



World Cup Shootout Penalties Missed/Saved

World Cups since 1982



World Cup Shootout Penalty Conversion

World Cups since 1982



10/10 100%	8/8 100%	8/8 100%	4/4 100%	5/5 100%	8/8 100%
16/17 94.1%	6/13 46.2%	4/5 80%	2/2 100%	4/6 66.7%	17/23 73.9%
32/37 86.5%	17/30 56.7%	7/14 50%	7/10 70%	12/18 66.7%	29/38 76.3%

World Cup Shootout Penalty Placement

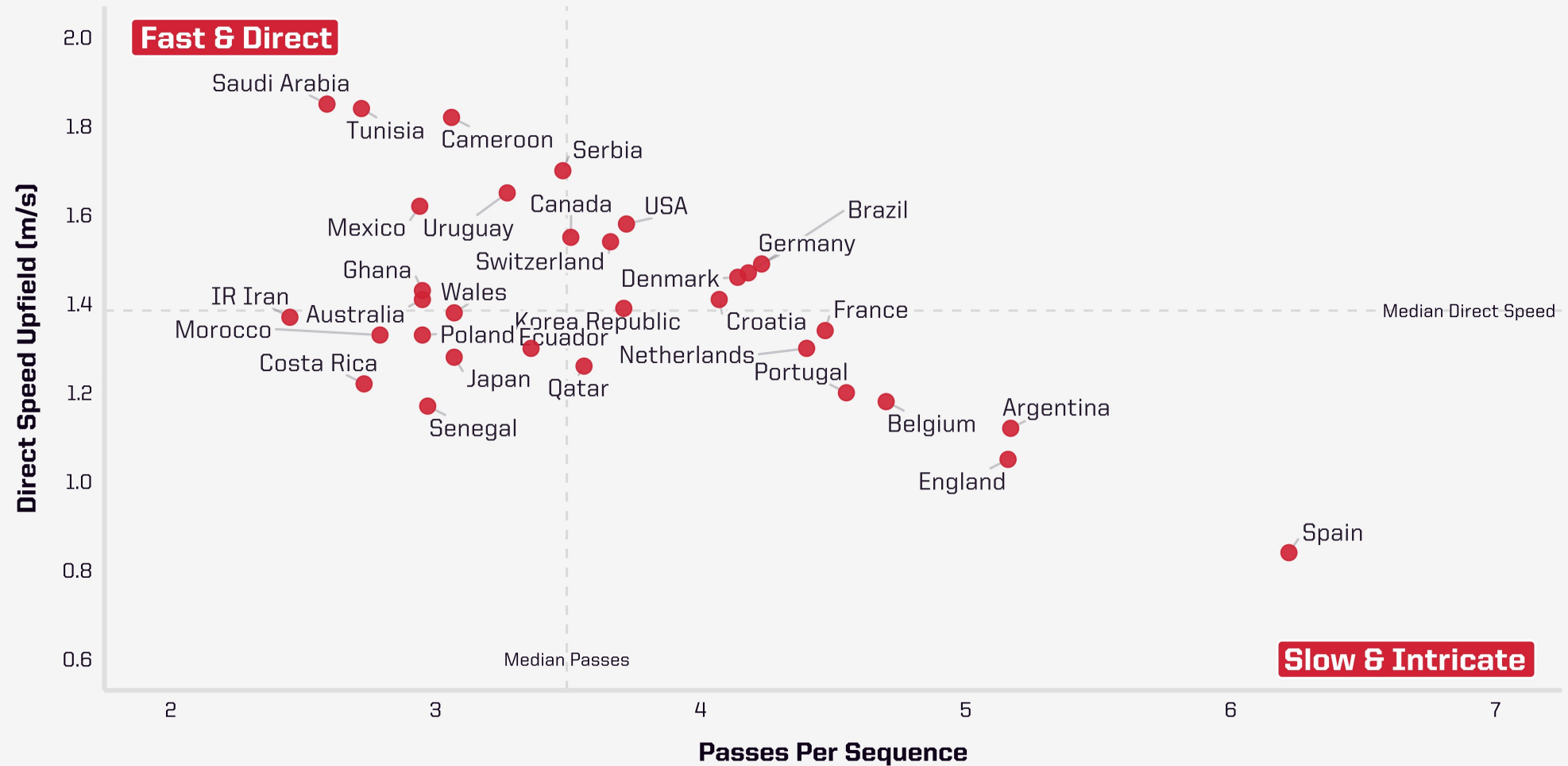
World Cups since 1982



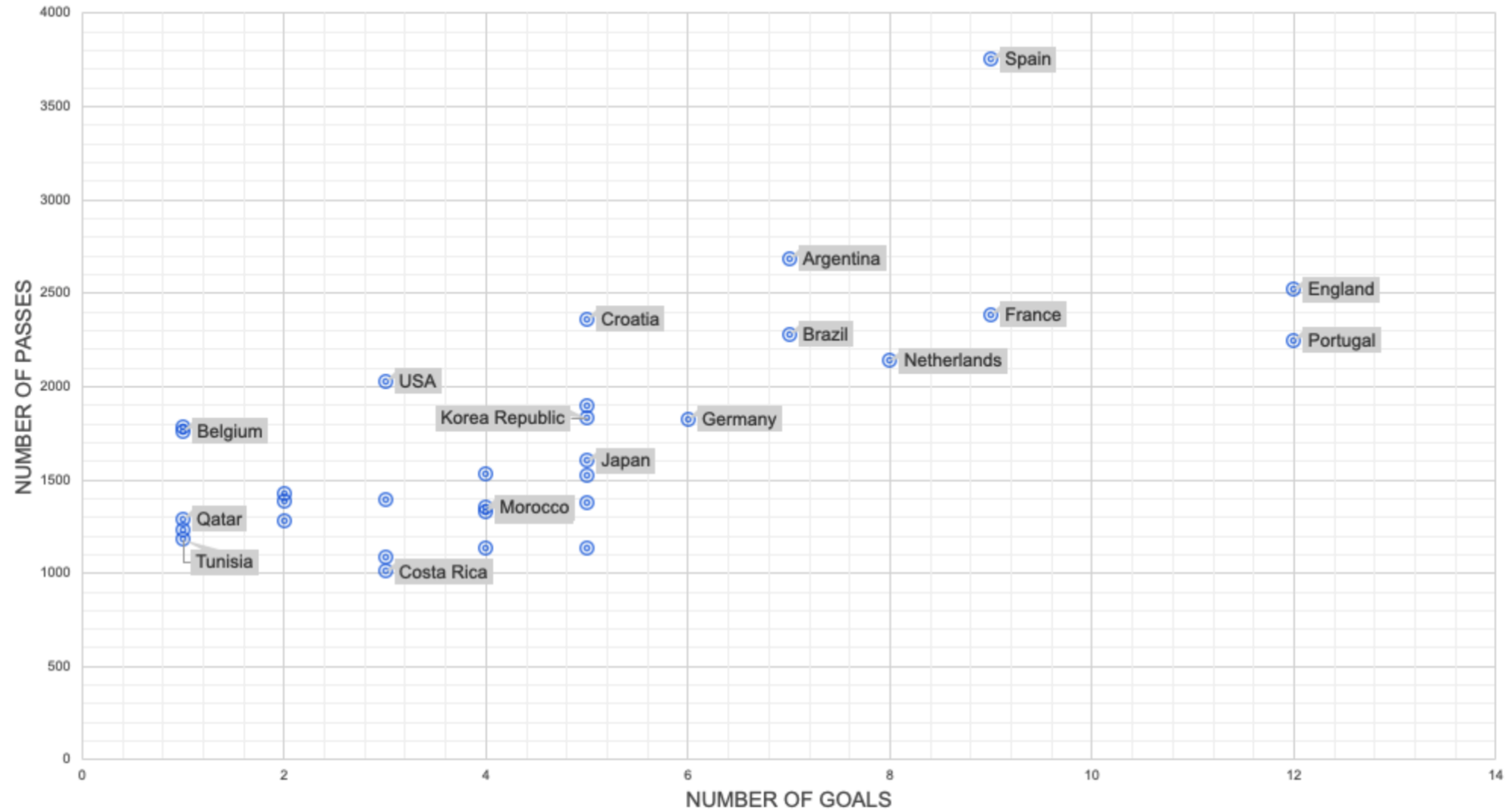
10 3.9%	8 3.1%	8 3.1%	4 1.6%	5 2%	8 3.1%
17 6.6%	13 5.1%	5 2%	2 0.8%	6 2.3%	23 9%
37 14.5%	30 11.7%	14 5.5%	10 3.9%	18 7%	38 14.8%

Team Style Comparison

World Cup 2022



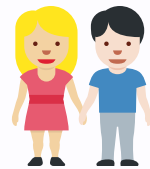
2022 Qatar World Cup



共分散と相関係数

covariance and correlation

二つの変数の関係性



分散

variance

一つの変数の平均からの散らばりを統計的に図る

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

標準偏差

standard deviation

標準偏差は分散の平方根である

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^N (x_i - \bar{x})^2}$$

共分散

covariance

二つの変数の平均からの散らばりを統計的に図る

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

相関係数

correlation

標準化した共分散の値

$$\begin{aligned} r &= \frac{s_{xy}}{s_x s_y} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

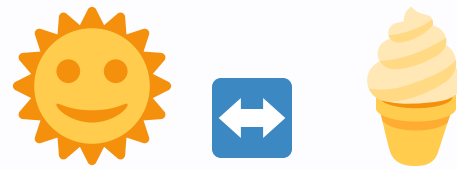
この計算だと、**相関係数**は必ず

-1 から $+1$

の間の数値になる

相関係数 r の値	相関
$-1 \leq r \leq -0.7$	強い負の相関
$-0.7 \leq r \leq -0.4$	負の相関
$-0.4 \leq r \leq -0.2$	弱い負の相関
$-0.2 \leq r \leq 0.2$	ほとんど相関がない
$0.2 \leq r \leq 0.4$	弱い正の相関
$0.4 \leq r \leq 0.7$	正の相関
$0.7 \leq r \leq 1$	強い正の相関

では実際に計算してみよう



天気とアイスクリームの関係ってどうなん？



天気とおでんの関係ってどうなん？



天気とゲームの関係ってどうなん？



10°C	50
20°C	100
0°C	50
30°C	200

この関係性を統計学な数値で表すには？

まずは共分散を計算する

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

☀️ → x 🍦 → y

☀の平均

$$\overline{x} = (10 + 20 + 0 + 30) / 4 = 15$$

🍦 の平均

$$\bar{y} = (50 + 100 + 50 + 200) / 4 = 100$$



$$x_i - \bar{x}$$

10°C

50

10-15=-5

20°C

100

20-15=5

0°C

50

0-15=-15

30°C

200

30-15=15

 $x_i - \bar{x}$ $y_i - \bar{y}$

10°C	50	-5	50-100=-50
20°C	100	5	100-100=0
0°C	50	-15	50-100=-50
30°C	200	15	200-100=100



		$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) * (y_i - \bar{y})$
--	--	-----------------	-----------------	-------------------------------------

10°C	50	-5	-50	250
20°C	100	5	0	0
0°C	50	-15	-50	750
30°C	200	15	-100	1500



	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
--	-----------------	-----------------	----------------------------------

10°C	50	-5	-50	250
------	----	----	-----	-----

20°C	100	5	0	0
------	-----	---	---	---

0°C	50	-15	-50	750
-----	----	-----	-----	-----

30°C	200	15	-100	1500
------	-----	----	------	------

-	-	-	-	2500
---	---	---	---	-------------

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$s_{xy} = \frac{2500}{4} = 625$$

625?

正の相関関係なのはわかるけど、その強度は？ 🧐

比較するものがないので、わからん！ 😓

そこで相関係数 (correlation) の出番

$$\begin{aligned} r &= \frac{s_{xy}}{s_x s_y} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

すでに s_{xy} はやったので

$$r = \frac{625}{s_x s_y}$$

☀️ → s_x 🍦 → s_y

標準偏差を計算！

 の標準偏差は

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



$$x_i - \bar{x} \quad (x_i - \bar{x})^2$$

10°C	-5	25
20°C	5	25
0°C	-15	225
30°C	15	225
-	-	500



$$s_x = \sqrt{\frac{500}{n}} = \sqrt{\frac{500}{4}} = \sqrt{125} = 11.18$$



の標準偏差は

$$s_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

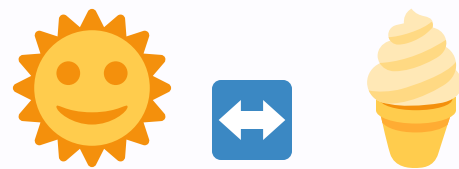


$$y_i - \bar{y} \quad (y_i - \bar{y})^2$$

50	-50	2500
100	0	0
50	-50	2500
200	-100	20000
-	-	25000



$$s_y = \sqrt{\frac{25000}{n}} = \sqrt{\frac{25000}{4}} = \sqrt{6250} = 79.06$$



相関係数（correlation）は！

$$r = \frac{625}{s_x s_y} = \frac{625}{11.18 * 79.06} = 0.71$$

相関係数 r の値	相関
$-1 \leq r \leq -0.7$	強い負の相関
$-0.7 \leq r \leq -0.4$	負の相関
$-0.4 \leq r \leq -0.2$	弱い負の相関
$-0.2 \leq r \leq 0.2$	ほとんど相関がない
$0.2 \leq r \leq 0.4$	弱い正の相関
$0.4 \leq r \leq 0.7$	正の相関
$0.7 \leq r \leq 1$	強い正の相関 🌞 ↔ 🍦

では実際にやってみよう

- choose a variable of your choice (x)
- create a correlation with happiness (y)

Example: sleep and happiness

- create a scatter plot 散布図
- calculate covariance 共分散
- calculate correlation 相関係数