今日だけスライドで

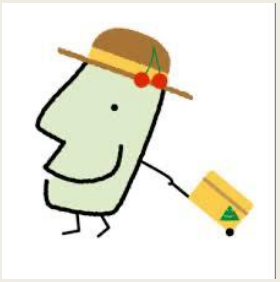
確率統計II

第1回
1変量データ

久保田 匠

自己紹介

名前　　：久保田 匠（くぼた しょう）
研究室　：自然科学棟 521 研究室
担当科目：確率統計II, プログラミング, 線形数学演習I
専門　　：スペクトルグラフ理論（線形代数＋グラフ理論）
　　量子ウォーク（確率論＋量子力学）
出身　　：山形県寒河江市（さくらんぼが有名）

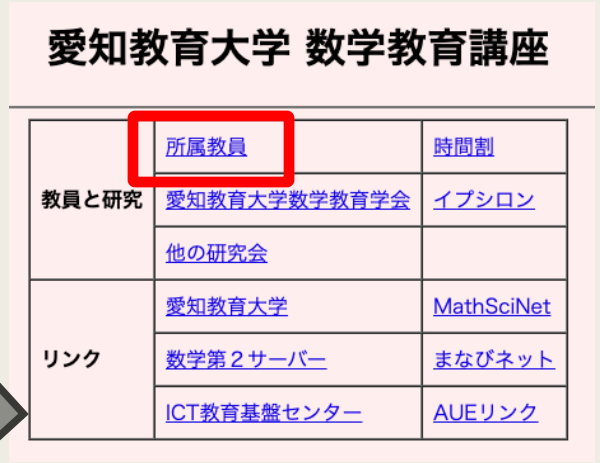


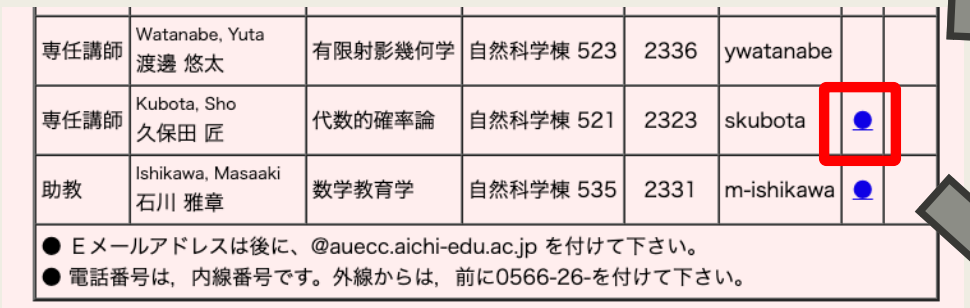
↑
←山形県公式HPより

https://machico.mu/special/detail/1827

授業資料と演習問題について

- 久保田の授業ホームページに資料がアップロードされている
- まずは「愛教大 数学」と検索してみよう。





専任講師	Watanabe, Yuta 渡邊 悠太	有限射影幾何学	自然科学棟 523	2336	ywatanabe		
専任講師	Kubota, Sho 久保田 匠	代数的確率論	自然科学棟 521	2323	skubota	●	
助教	Ishikawa, Masaaki 石川 雅章	数学教育学	自然科学棟 535	2331	m-ishikawa	●	

● Ｅメールアドレスは後に、@auecc.aichi-edu.ac.jp を付けて下さい。
● 電話番号は，内線番号です。外線からは，前に0566-26-を付けて下さい。

授業資料と演習問題について

久保田匠の授業用ホームページ

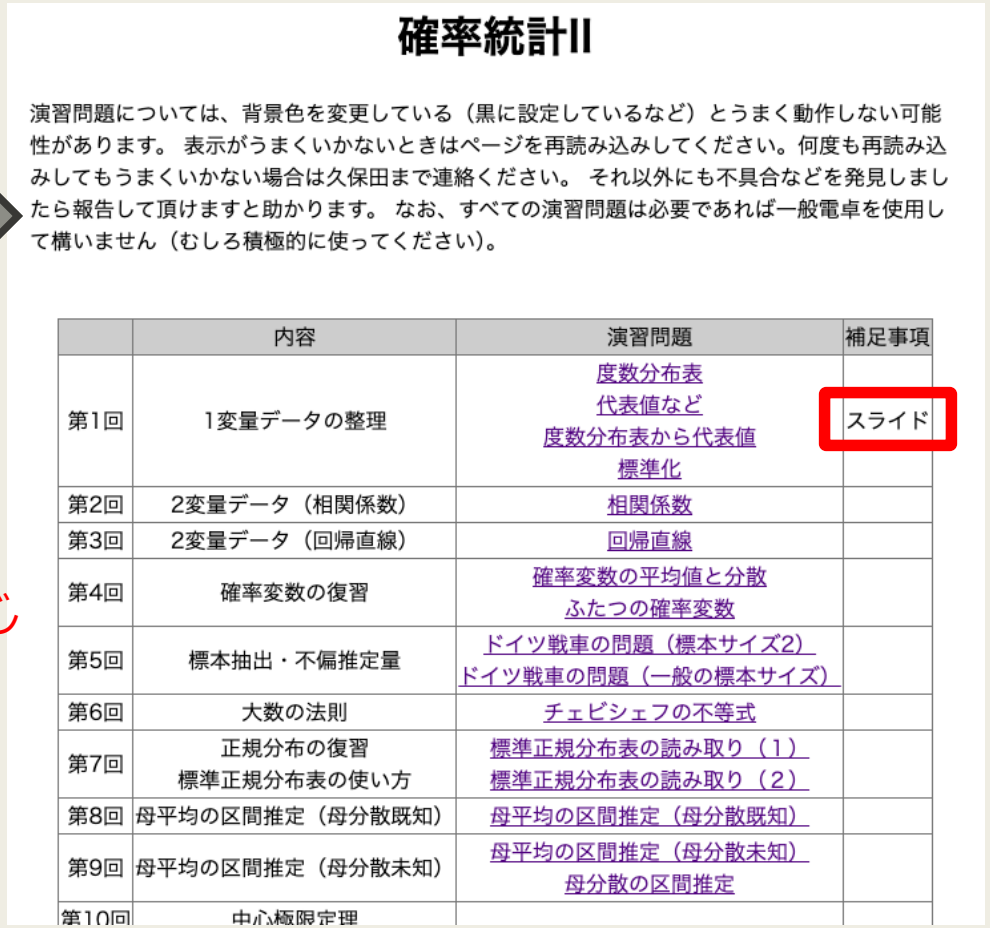
2025年度前期担当科目

	月曜	火曜	水曜	木曜	金曜
1限					
2限	確率統計II			確率統計II	
3限					確率統計II
4限	4年ゼミ				（オフィスアワー）
5限					

2025年度後期担当科目

	月曜	火曜	水曜	木曜	金曜
1限					
2限					
3限	科学リテラシー				プログラミング
4限	（オフィスアワー）	3年ゼミ？			プログラミング
5限					

月曜2限も木曜2限も金曜3限も全部同じ



確率統計II

演習問題については、背景色を変更している（黒に設定しているなど）とうまく動作しない可能性があります。 表示がうまくいかないときはページを再読み込みしてください。何度も再読み込みしてもうまくいかない場合は久保田まで連絡ください。 それ以外にも不具合などを発見しましたら報告して頂けますと助かります。 なお、すべての演習問題は必要であれば一般電卓を使用して構いません（むしろ積極的に使ってください）。

	内容	演習問題	補足事項
第1回	1変量データの整理	度数分布表 代表値など 度数分布表から代表値 標準化	スライド
第2回	2変量データ（相関係数）	相関係数	
第3回	2変量データ（回帰直線）	回帰直線	
第4回	確率変数の復習	確率変数の平均値と分散 ふたつの確率変数	
第5回	標本抽出・不偏推定量	ドイツ戦車の問題（標本サイズ2） ドイツ戦車の問題（一般の標本サイズ）	
第6回	大数の法則	チェビシェフの不等式	
第7回	正規分布の復習 標準正規分布表の使い方	標準正規分布表の読み取り（1） 標準正規分布表の読み取り（2）	
第8回	母平均の区間推定（母分散既知）	母平均の区間推定（母分散既知）	
第9回	母平均の区間推定（母分散未知）	母平均の区間推定（母分散未知） 母分散の区間推定	
第10回	中心極限定理		

- 毎回の授業の演習問題はこのページにあるのでブックマーク（お気に入り）登録しておくとよい。

授業の進め方と教科書

- 講義パート（60～70分）
  - 集中して聞く。ノートを取る。**私語厳禁。**
- 演習パート（残りの時間）
  - 授業用ホームページにある演習問題を解く。
  - この時間は休憩時間ではない。

- 教科書
  - 尾畑伸明『データサイエンスのための確率統計』

- 確率統計Iと同じなので購入済み…？
  - 「教科書の p.** を見てください」のような指示をすることがあるのでまだ持ってない人は購入してね。

成績評価

- 期末試験（100%）
- ただし…
- 試験の際は 一般電卓 を使用してもよい。
  - ただし、関数電卓やスマートフォンは使用できない。
  - 「一般電卓を使用してもよい」というのは「一般電卓がないと単位取得は厳しい」という意味。
- なるべく早く購入して電卓の操作に慣れておくこと。
- 授業のときはスマホの電卓機能を使ってもOK。

平方根（√）を計算する機能は必須

確率統計II の授業で学ぶこと

- 大雑把には統計学の基礎。
- 主に次の3つを学ぶ。

推測統計（本講義のメイン）

記述統計　　区間推定　　仮説検定

データの整理
7割は高校までの復習

- 推測統計とは、すべてのデータを調べることが難しい状況（データの数が膨大、調査コストが高い、データを取ることで商品価値が失われる、など）において、サンプル（標本）をもとに、全体の平均や分散を「推測」するための統計的手法の総称のこと。

推測統計ヨ 区間推定, 仮説検定

- 学生の平均睡眠時間を調べてみよう。
- 30人の学生にアンケートをとると平均は 6.8時間 だった。
  - 本当の平均も 6.8時間 に近いと思われるが…
  - 調査したのはたまたま選んだ30人。
- 本当の平均は「ぴったり 6.8時間である」と言い切るのはちょっと自信がない。
- そこで登場するのが 区間推定。
- 区間推定では「このくらいの幅の中に本当の平均があるはず」というゆとりをもたせた推定を行う。
  - 例えば、「95%の確信で、本当の平均は 6.5時間～7.1時間の間にある」のように。
- 何を推定するか？何の情報を既知とするか？で扱う分布は変わってくるので、 4,5パターンくらいおさえる必要がある。

推測統計ヨ区間推定, 仮説検定

「背理法」＋「確率」

- コイントスで遊んでいたところ、使っているコインはどうも表が多めに出ているような気がしたとする。

- 「このコインが公平であるか？」をテストしたい。
- 「コインが公平である」という仮説を立てる。

帰無仮説という

- 実際に100回コインを投げてみたら62回表が出た。
  - この結果は偶然かもしれないし、仮説が誤りなのかもしれない。
- 「62回表が出る」以上に極端な結果が出る確率を計算する。

$$P(\text{表が出た回数が38回以下 or 62回以上}) \fallingdotseq 0.0210$$

- 5%を下回ったので「コインが公平である」という仮説は正しくなさそうだ、と判断する。
- 本来は検定を行う前に何%の水準で行うかを決めておく。5%は一般的。

1変量データの整理

- しばらくは中学高校の復習。
- 「データ」の例は、試験の点数など。
- n人の学生が数学の試験を受けたとしたら、n個のデータ

$$x_1, x_2, \ldots, x_i, \ldots, x_n$$

が集まる。

- 気象データや通販サイトの購入履歴では、データの数は数十万を超える。
- しかし、大量の数値データをそのまま見ても傾向を掴むのは難しいため、適切な方法を使ってデータを整理する必要がある。
- 多くの場合、まず度数分布表やヒストグラムを作成する。
- 平均値や分散だけでデータを理解しようとしないこと。

度数分布表

- ローデータ（raw data, 生のデータ）が手に入ったらまずは度数分布表を作ろう。
- 次は試験結果（100点満点）のローデータである。

86, 92, 92, 95, 78, 80, 92, 95, 68, 99, 73, 93, 91, 92, 72,
80, 93, 92, 86, 95, 99, 86, 92, 81, 69, 93, 65, 38, 81, 52,
67, 65, 65, 76, 68, 86, 97, 78, 46, 97, 76, 70, 81, 85, 51,
60, 78, 73, 73, 76, 95, 92, 92, 76, 89, 78, 86, 93, 95, 51

- 度数分布表は次のスライドで。

度数分布表

86, 92, 92, 95, 78, 80, 92, 95, 68, 99, 73, 93, 91, 92, 72,
80, 93, 92, 86, 95, 99, 86, 92, 81, 69, 93, 65, 38, 81, 52,
67, 65, 65, 76, 68, 86, 97, 78, 46, 97, 76, 70, 81, 85, 51,
60, 78, 73, 73, 76, 95, 92, 92, 76, 89, 78, 86, 93, 95, 51

$$\frac{30+40}{2}$$

階級	階級値	試験の成績	
		度数	相対度数
以上 未満			
30 ～ 40	35	1	0.02
40 ～ 50	45	1	0.02
50 ～ 60	55	3	0.05
60 ～ 70	65	8	0.13
70 ～ 80	75	13	0.22
80 ～ 90	85	12	0.20
90 ～ 100	95	22	0.37
合計	-	60	1

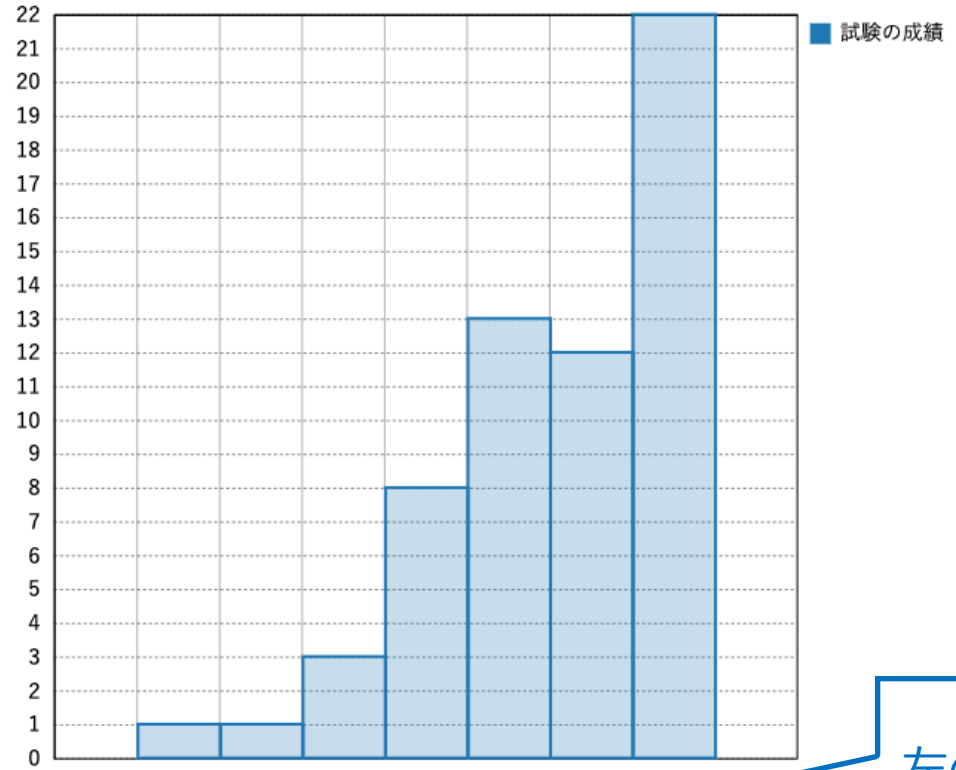
70以上80**未満**の区間に
13個のデータがある

$$= \frac{13}{60}$$

ヒストグラム

- 度数分布表を作ったらグラフにまとめる。
- 以下のようなグラフを ヒストグラム という。

階級	階級値	試験の成績	
		度数	相対度数
以上 未満			
30 ～ 40	35	1	0.02
40 ～ 50	45	1	0.02
50 ～ 60	55	3	0.05
60 ～ 70	65	8	0.13
70 ～ 80	75	13	0.22
80 ～ 90	85	12	0.20
90 ～ 100	95	22	0.37
合計	-	60	1



「棒」の範囲は通常
左の数以上 右の数未満

- 実践的には、エクセルやその他適当なツールにデータを入力すると度数分布表もヒストグラムもコンピュータが作ってくれる。

普通の意味の平均値

- データ $x_1, \ldots, x_n$ に対して、

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

  を 平均値 という。

- 一方で、状況によってはローデータが手に入らず、手元にあるのが度数分布表だけという場合もある。
- 度数分布表（だけの情報）からローデータを復元することはできない。
- このような状況においては、度数分布表の情報だけをもとにして得られる適切な「平均値」を算出することになる。

度数分布表から得られる平均値

- 右のように、度数分布表（のみ）が与えられている場合。
- 階級値と度数を見て、

  階級 I_j には、a_j のデータが f_j 個ある

  とみなして計算する。

階級	階級値	度数
I_1	a_1	f_1
$\vdots$	$\vdots$	$\vdots$
I_j	a_j	f_j
$\vdots$	$\vdots$	$\vdots$
I_k	a_k	f_k
合計		n

- 度数分布表からの平均値 は

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} a_j f_j$$

と定める。

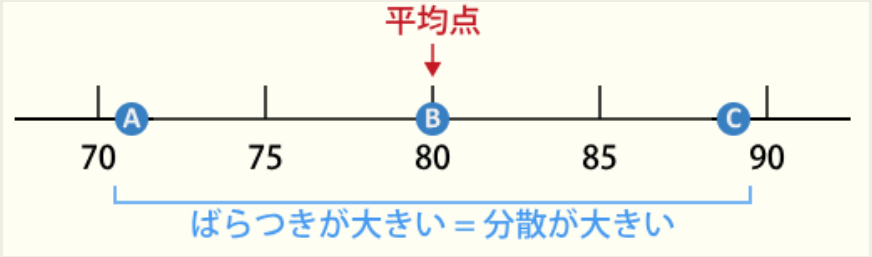
分散と標準偏差

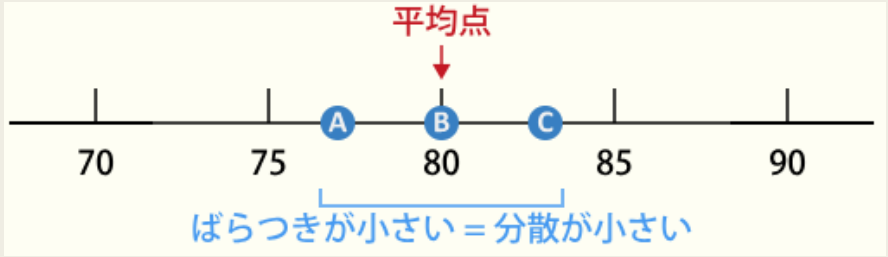
- データ $x_1, \ldots, x_n$ に対して、

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

  を 分散 という。

- どの変量についての分散かを明示したいときは s_x^2 とかく。
- $s(>0)$ を 標準偏差 という。
- 分散はデータのばらつきを表す。





https://sci-pursuit.com/math/statistics/variance.html

分散公式

- 高校で習ってる…？

データをすべて
2乗したものの平均

定理

$$s^2 = \overline{x^2} - \bar{x}^2$$

証明.

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}2\bar{x}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 \\
&= \overline{x^2} - \frac{2\bar{x}}{n}\sum_{i=1}^{n}x_i + \frac{\bar{x}^2}{n}\sum_{i=1}^{n}1 \\
&= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 \\
&= \overline{x^2} - \bar{x}^2
\end{aligned}$$

分散の定義 vs 分散公式

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

平均値が汚いと
この項が汚い

$$s^2 = \overline{x^2} - \bar{x}^2$$

- 分散を計算する場合、多くのケースでは分散公式を使う方が簡単。
- ただし、平均値が整数になる場合は定義を使った計算の方が多分簡単。

不偏分散

- データ $x_1, \ldots, x_n$ に対して、

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

  を 不偏分散 という。

- 推測統計ではむしろ不偏分散の方が重要。
- なぜ n-1 で割るのかについては現時点では説明できないが、理論的に重要な意味はある。
  - 第5回くらいの授業で説明します。
- なお、不偏分散の正の平方根 u(>0) を不偏標準偏差とは言わないので注意。
  - 不偏推定量（←第5回の授業で説明予定）にならないため。
- 不偏分散の正の平方根 u は「不偏分散の正の平方根」という。

例題1

例. サイズ 4 のデータが 14, 16, 22, 23 のとき、中央値、分散s^2、標準偏差s、不偏分散u^2 を求めよ。

- 中央値は「データを小さい順に並べたときの中央の数」だが、データのサイズが偶数のときは中央に最も近いふたつの値の平均。つまり、(16+22)/2 = **19** である。
- 分散は、分散公式を使って計算する。
  - 与えられたデータを x として、例えば次の表を作る。

「2」「3」「×」「=」で23^2

x	14	16	22	23
x^2	196	256	484	529

$$\bar{x} = 18.75$$

$$\overline{x^2} = 366.25$$

  - 分散公式より $s^2 = \overline{x^2} - \bar{x}^2 = 366.25 - 18.75^2 = 14.6875$
  - 小数第2位まで求めるなら分散は **14.69**
- 標準偏差は $\sqrt{14.6875} = 3.832427$ より **3.83**

誤差が出るので丸める前の数を使う。

例題1

例. サイズ 4 のデータが 14, 16, 22, 23 のとき、中央値、分散s^2、標準偏差s、不偏分散u^2 を求めよ。

- 不偏分散は通常の分散から計算するのがオススメ。

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

より、

$$u^2 = \frac{n}{n-1}s^2 = \frac{4}{3} \cdot 14.6875 = 19.583333$$

なので、不偏分散は **19.58**

誤差が出るので丸める前の数を使う。

例題2

例. 次の度数分布表のデータについて中央値と最頻値を求めよ。

階級	度数
40–60	4
60–80	6
80–100	10
合計	20

- 度数分布表のみが与えられている場合は、ローデータの情報は失っているので階級値を使って計算する。
- 中央値は、データの数が偶数個なので今回は10番目と11番目のデータの（階級値の）平均をとって、(70+90)/2 = **80**
- 最頻値は度数が最も多い階級の階級値なので **90**
  - 度数が最も多い階級が複数ある場合はすべて答える。

ax+b の平均と分散

定理. 変量 x のデータ $x_1, \ldots, x_n$ と実数 a, b に対して、次が成り立つ。

$$\overline{ax+b} = a\bar{x} + b, \qquad s^2_{ax+b} = a^2 s^2_x$$

証明.

$$\begin{aligned}
\overline{ax+b} &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\
&= \frac{1}{n}\left(a\sum_{i=1}^{n} x_i + bn\right) \\
&= a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + b \\
&= a\bar{x} + b
\end{aligned}$$

$$\begin{aligned}
s^2_{ax+b} &= \frac{1}{n}\sum_{i=1}^{n}\left(ax_i + b - (\overline{ax+b})\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(ax_i + b - (a\bar{x} + b)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} a^2(x_i - \bar{x})^2 \\
&= a^2\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= a^2 s^2_x
\end{aligned}$$

b が消える理由

a^2 が出る理由

データの標準化

定理. 変量 x のデータ $x_1, \ldots, x_n$ に対して変量 z を

$$z = \frac{x - \bar{x}}{s_x}$$

x の **標準化** という。
確率統計II では重要。

で定める。このとき、z の平均は 0 であり、分散（と標準偏差）は 1 である。

証明. 前のスライドの定理を使う。

$$z = \frac{1}{s_x}(x - \bar{x}) = \frac{1}{s_x}x - \frac{\bar{x}}{s_x}$$

$$\bar{z} = \frac{1}{s_x}x - \frac{\bar{x}}{s_x} = \frac{1}{s_x}\bar{x} - \frac{\bar{x}}{s_x} = 0$$

$$s_z^2 = s^2_{\frac{1}{s_x}x - \frac{\bar{x}}{s_x}} = s^2_{\frac{1}{s_x}x} = \frac{1}{s_x^2}s_x^2 = 1$$

例題3

例題1と同じデータ

例. サイズ 4 のデータ 14, 16, 22, 23 を標準化せよ。

$$z = \frac{x - \bar{x}}{s_x}$$

- データの平均値と標準偏差は例題1 で求めたように、

  平均値 18.75　　標準偏差 3.83247

誤差が出るので
丸める前の数を使う

- よって、標準化されたデータは

  -1.24, -0.72, 0.85, 1.11

$$\frac{x - 18.75}{3.832427}$$

- なお、偏差値 は平均値が50, 標準偏差が10になるように変換された指標。事務的な事情により導入された。

演習

- 久保田の授業ホームページにアクセスして問題演習。

「度数分布表」

「代表値など」

「度数分布表から代表値」

「標準化」

このふたつから
取り組んでみよう

- 第2回以降の授業も、残った時間は問題演習の時間。
- 授業を聞いて分かったつもりになるのではなく、実際に手を動かし、自分がその日の内容を理解できているか問題演習を通して確認しよう。