応用数理 E 第 8 回目

永幡幸生
新潟大学工学部

2024 前期

データの処理

現在はほとんどの場合コンピュータを使ってデータの処理をおこなう。分野によってはその場で、グラフを書いた方が良く分かるものもあり、定性的な性質が良く分かっているために、特定の比率でグラフを書くと良いものもあり、有名なところで片対数グラフ、両対数グラフがあり、方眼紙としてその比率で線が引かれたものもある。
重要なことは、どのように考えるべきか？どのような統計量が重要か？その量は他の人も分かってくれる量か？に答えることである。
他の人に分かってもらうというのは非常に難しい問題で、その人がその問題を知っているか知らないかにより大きく変わってしまい、学会、研究室内、会社内、などの特定の場所で話すことが前提になっている場合と、全くそのことを知らない一般の人向けに話すときではどのような統計量が重要かも変わってきてしまう。

データの処理

問題

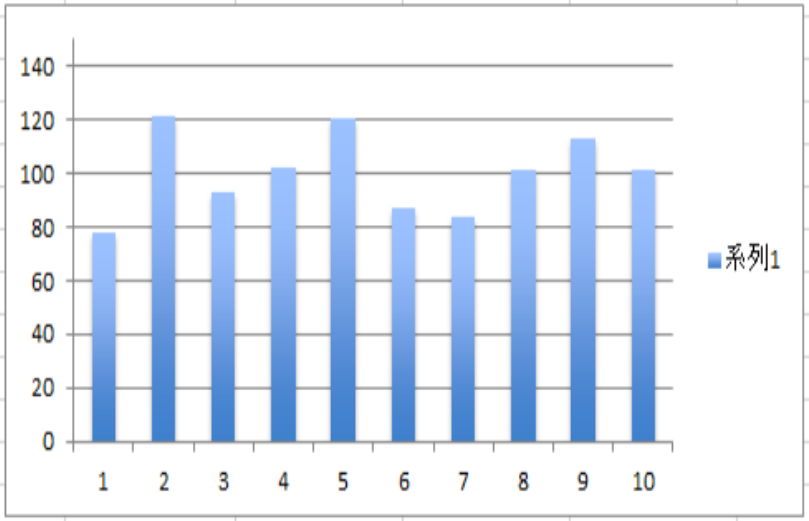
片対数グラフ、両対数グラフに関して調べよ。特にどのようなことが分かっている時にこのようなグラフを使うか、例を挙げると良い。

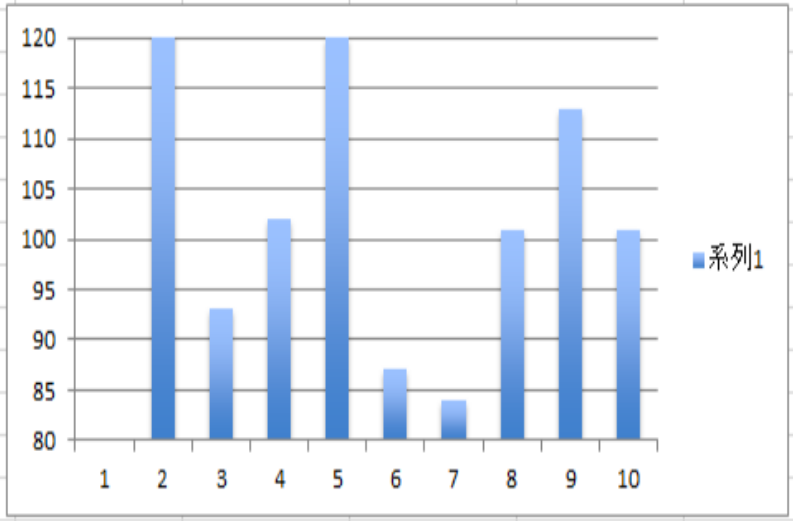
データの処理

同じデータを使っても、処理の仕方、グラフの書き方でかなり印象が変わる。
例えば次のページの２つのグラフを見ると、左のグラフは、「多少ばらつきはあるが、似たような値を取っている」と感じるし、右のグラフは、「かなり値に差がある」と感じる。
しかしよく見ると、値域が違うだけで、実は右側は、左側のグラフの一部分を切り取って、縦横の比を変えただけのものである。
このグラフは意図的に２つ作ったが、コンピュータは自動でグラフを書いてくれるため、意図と違うものが出てしまうことがあるので、注意する必要があるとともに、グラフなどを見るときは意図的に、特徴的なグラフを作って、実際とは異なる結果を主張していないか注意する必要もある。

データの処理

値	0.000174609-0	0.10013349770	0.20009238640	0.30005127511	0.40001016381	0.49996905252	0.59992794122	0.69988682993	0.79984571863	0.89980460734
度数	78	121	93	102	120	87	84	101	113	101





データの処理

一般的には、図、グラフだけでなく、その特性量、特徴量を数値として出す方が良いが、一方で数値だけでは分かりにくい。このため両方出した方が良い。

データの処理

新潟の８月の気温（最高気温、最低気温）を例としてグラフを作る。

- 度数分布表とヒストグラムを作る

この分割を階級、階級に入った数を度数と呼ぶ。
通常階級は等分割にする。
いくつに分けるかは必ずしも決まっていない。
目安として 階級 $\cong \sqrt{\text{総データ数}}$ がある。

データの処理

- この例での特性量として

最大、最小、

標本平均： $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

標本分散： $v^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

標本中央値（メディアン）：データを小さいほうから並べなおして $x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[n]}$ として

$$\tilde{x} = \begin{cases} x_{[(n+1)/2]} & n \text{ が奇数の場合} \\ \dfrac{1}{2}\left(x_{[n/2]} + x_{[n/2]+1}\right) & n \text{ が偶数の場合} \end{cases}$$

データの処理

多くの場合 $\bar{x} \cong \tilde{x}$ であり多くの場合 $\bar{x}$ だけを求めるが、特定の分布だと、$\bar{x} \neq \tilde{x}$ となり $\tilde{x}$ の方が良い特性量になっていることが知られている。

問題

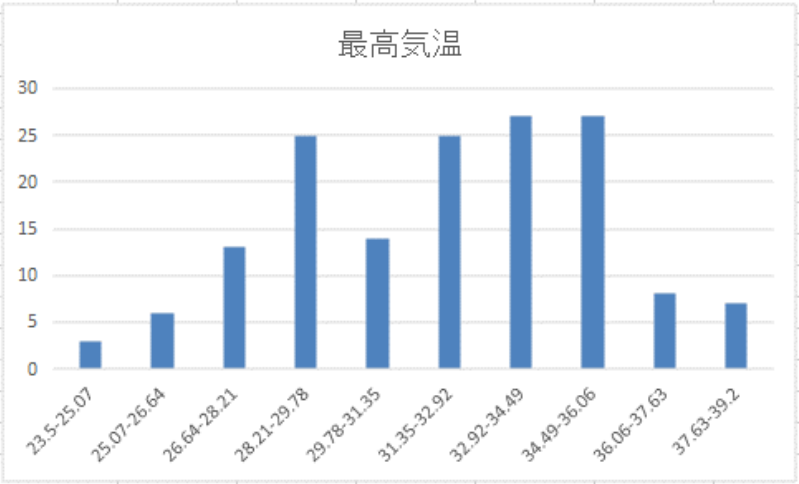
$\bar{x} \neq \tilde{x}$ となり $\tilde{x}$ の方が良い特性量になっている例を調べ、実際にそうであるか感想を述べよ。

データの処理

5年間の最高気温から度数分布表、ヒストグラムを作成

値	23.5-25.07	25.07-26.6	26.64-28.2	28.21-29.7	29.78-31.3	31.35-32.9	32.92-34.4	34.49-36.0	36.06-37.6	37.63-39.2	計
度数	3	6	13	25	14	25	27	27	8	7	155

標本平均	32.06065
標本分散(不偏分散)	11.86903
標本中央値	32.4

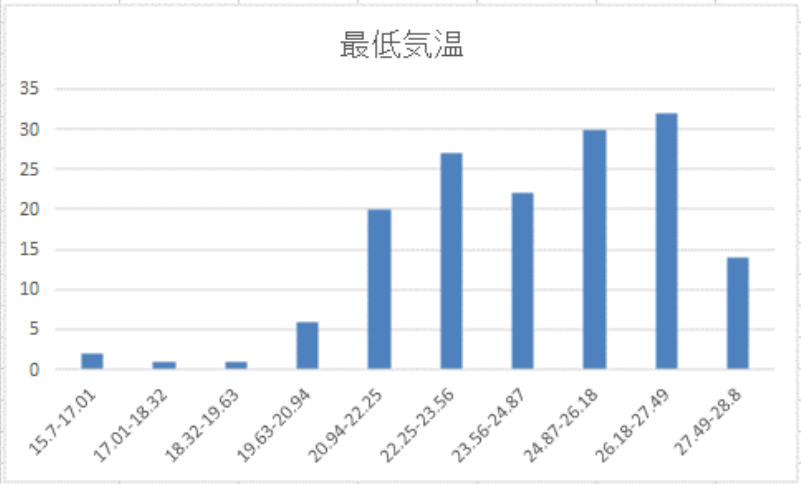


データの処理

5年間の最低気温から度数分布表、ヒストグラムを作成

値	15.7–17.01	17.01–18.3	18.32–19.6	19.63–20.9	20.94–22.2	22.25–23.5	23.56–24.8	24.87–26.1	26.18–27.4	27.49–28.8	計
度数	2	1	1	6	20	27	22	30	32	14	155

標本平均	24.43419
標本分散(不偏分散)	6.060706
標本中央値	24.7



データの処理

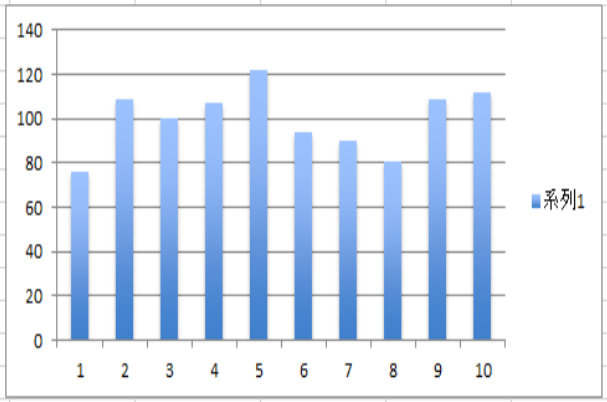
典型的な分布である、一様分布、正規分布、χ^2 分布のヒストグラムを挙げる。
(コンピュータ上で乱数を発生させ、ヒストグラムを作る。)

データの処理

一様分布

データ数	1000
最大値	0.99812844
最小値	0.000174609
分割数(固定)	10
階級	0.099795383

値	0.000174609-0	0.09996999206	0.19976537512	0.29956075818	0.39935614124	0.49915152430	0.59894690736	0.69874229042	0.79853767348	0.89833305654	計
度数	76	109	100	107	122	94	90	81	109	112	

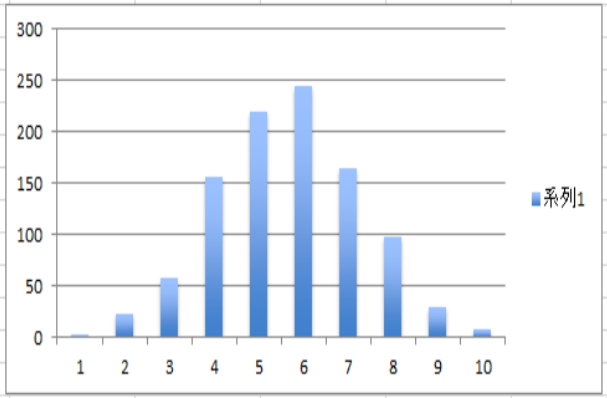


データの処理

正規分布

データ数	1000
最大値	2.993153269
最小値	-3.312787341
分割数(固定)	10
階級	0.630594061

値	-3.3127873409	-2.6821932799	-2.0515992189	-1.4210051579	-0.7904110969	-0.1598170359	0.47077702500	1.10137108600	1.73196514700	2.36255920800	計
度数	2	22	57	156	219	245	165	97	30	7	

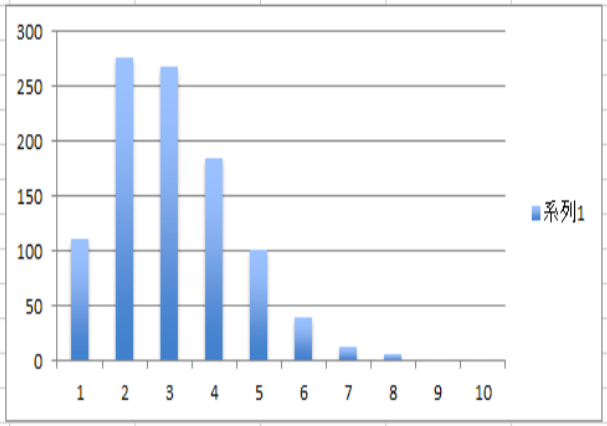


データの処理

χ^2 分布

データ数	1000
最大値	33.22902473
最小値	1.601173069
分割数(固定)	10
階級	3.162785166

値	1.601173069	4.763958236	7.926743402	11.08952857	14.25231374	17.41509891	20.57788407	23.74066923	26.90345440	30.06623957	計
度数	111	276	268	184	101	40	12	6	1	1	



データの処理

- 散布図

8 月の気温で最高気温と最低気温を同時に見た場合、何らかの相関関係があることを予想することは当然である。このようなデータに対して、横軸に最高気温、縦軸に最低気温をとってグラフを書いたものを散布図と呼ぶ。

- この例での特性量として

標本平均：$\bar{x}, \bar{y}$

標本相関係数：r

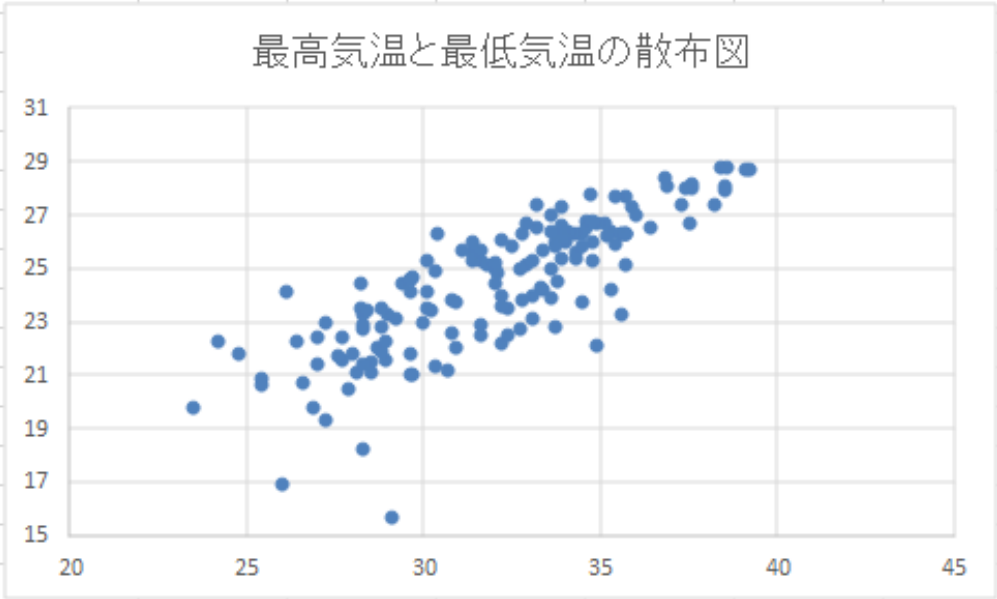
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

として

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

データの処理

5年間の最高気温、最低気温から散布図を作成

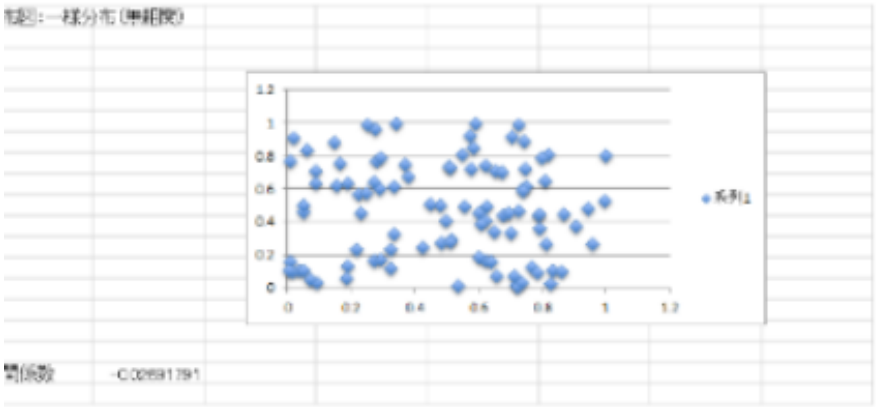


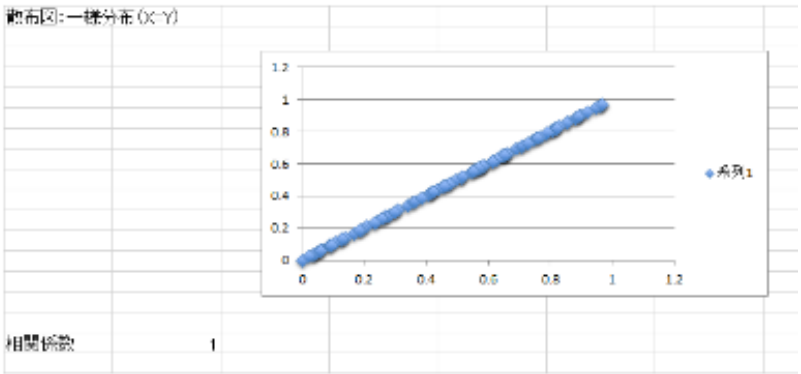
	平均	分散	共分散	相関係数
最高気温	32.06065	11.86903	6.879666	0.811144
最低気温	24.43419	6.060706		

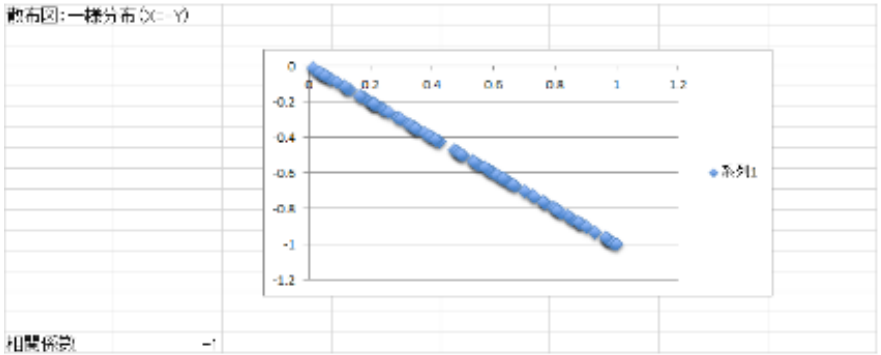
データの処理

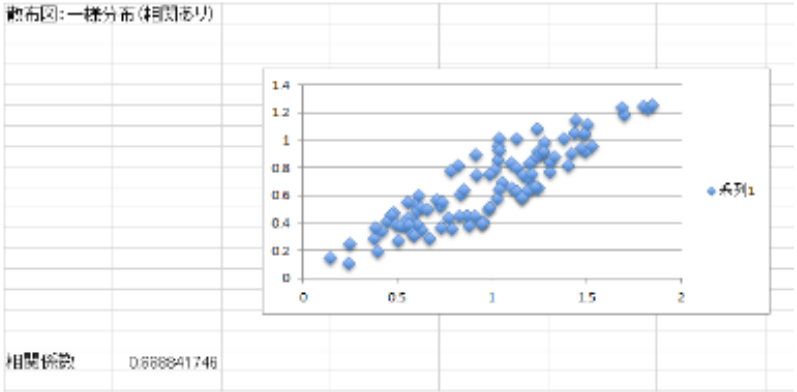
一様分布と、正規分布に関して、適当に相関を変えた散布図を挙げる。

データの処理

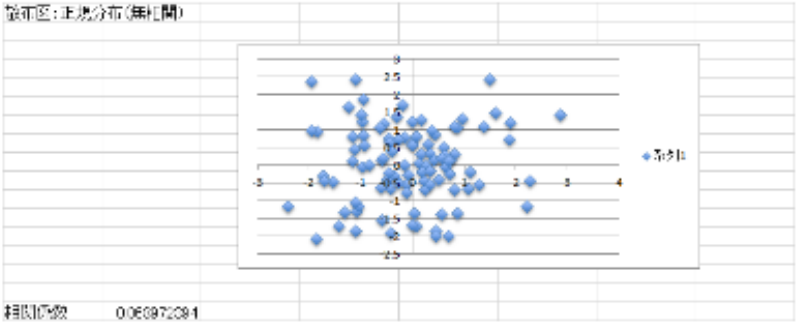

散布図：一様分布（無相関）
系列1
相関係数

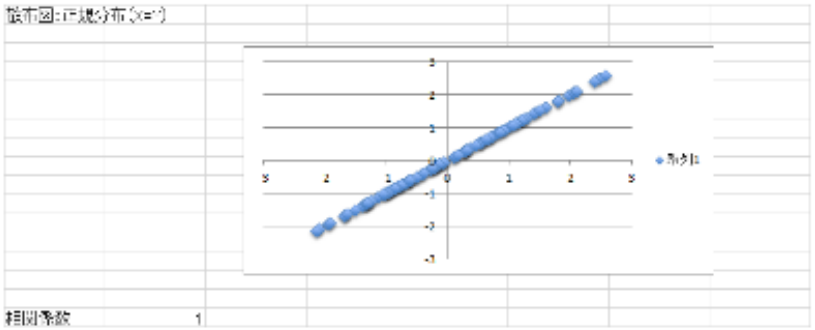

散布図：一様分布（X=Y）
系列1
相関係数
1

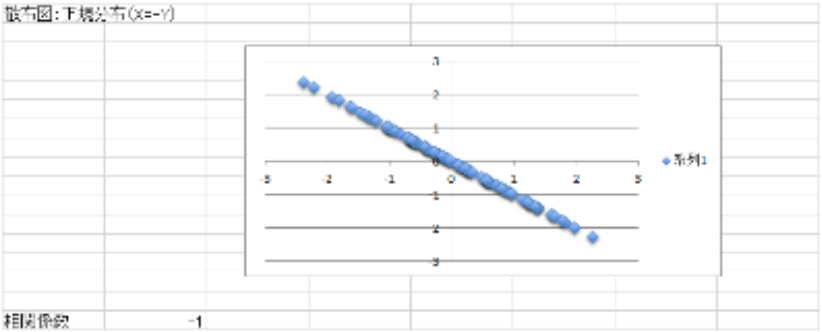

散布図：一様分布（X=−Y）
系列1
相関係数

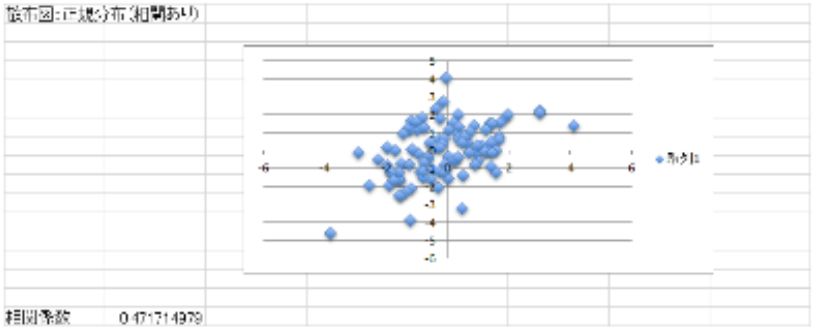

散布図：一様分布（相関あり）
系列1
相関係数
0.888841746

データの処理


散布図：正規分布(無相関)
系列1
相関係数 0.063972081


散布図：正規分布(x=y)
系列1
相関係数 1


散布図：正規分布(x=-y)
系列1
相関係数 -1


散布図：正規分布(相関あり)
系列1
相関係数 0.471714979