Excel統計（2025年2月版）

Excel Statistics (February 2025 Edition)

はじめに

Introduction

はじめに　（研修の背景）

EXCEL統計の講座を始めるにあたり、統計学の役割とその重要性、活用事例を確認しましょう。

📌 統計の役割とプロセス

問題・課題の整理⇒偏りのないデータ収集⇒分析・可視化（見える化）

インサイト（示唆）の抽出⇒報告・意思決定 ⇒行動

行動の結果を検証 ⇒戦略を見直し、繰り返す

📌 統計学が重要な理由

- 『統計学が最強の学問である』（西内 啓著-ビジネス書大賞2014「大賞」受賞）では「統計は、科学・ビジネス・医療など幅広い分野で意思決定の根拠となる」と述べられています。

📌 統計の活用事例

- フェイクデータ検知：基本統計量や相関係数を用いた異常データの検出
- 生成AIのエンジニアリング：データ前処理（クリーニング・特徴量抽出）への活用

この講座では、Excelを使いながら統計の基本を学び、実務での活用スキルを身につけていきます。

右の赤枠内がフェイクを表しています。

https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/html/nd123520.html

※リアルタイムで動画の信頼性が表示される。赤枠がディープフェイク部分を示している。

はじめに（研修の目標と目的）

📌 研修の目標

この講座では、統計の基本概念を理解し、Excelを用いて実践できるようになることを目指します。

1次元データの分析と可視化（平均・中央値・標準偏差、度数分布表などを算出し、データを見える化）

2次元データの分析と可視化（相関係数、分割表・クロス集計表とカイ2乗検定と可視化）

📌 研修の進め方

1次元データの基本分析＆ハンズオン（Excelを使用）

2次元データの基本分析＆ハンズオン（Excelを使用）

例題1問・演習問題2問・課題1問 に取り組み、理解を深める

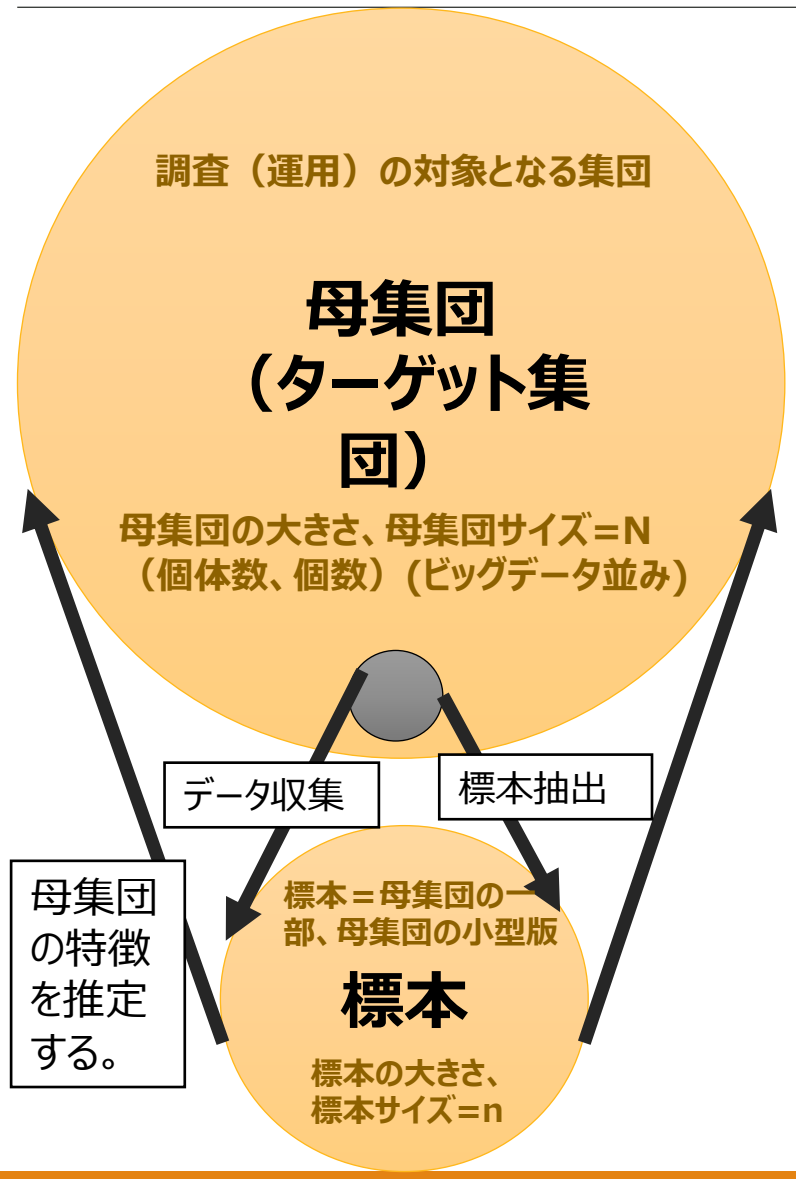
📌 研修の目的

学んだ内容を現場で活用し、成果を出すこと を目指します。

1次元データの基本分析

Basic Analysis of One-Dimensional Data

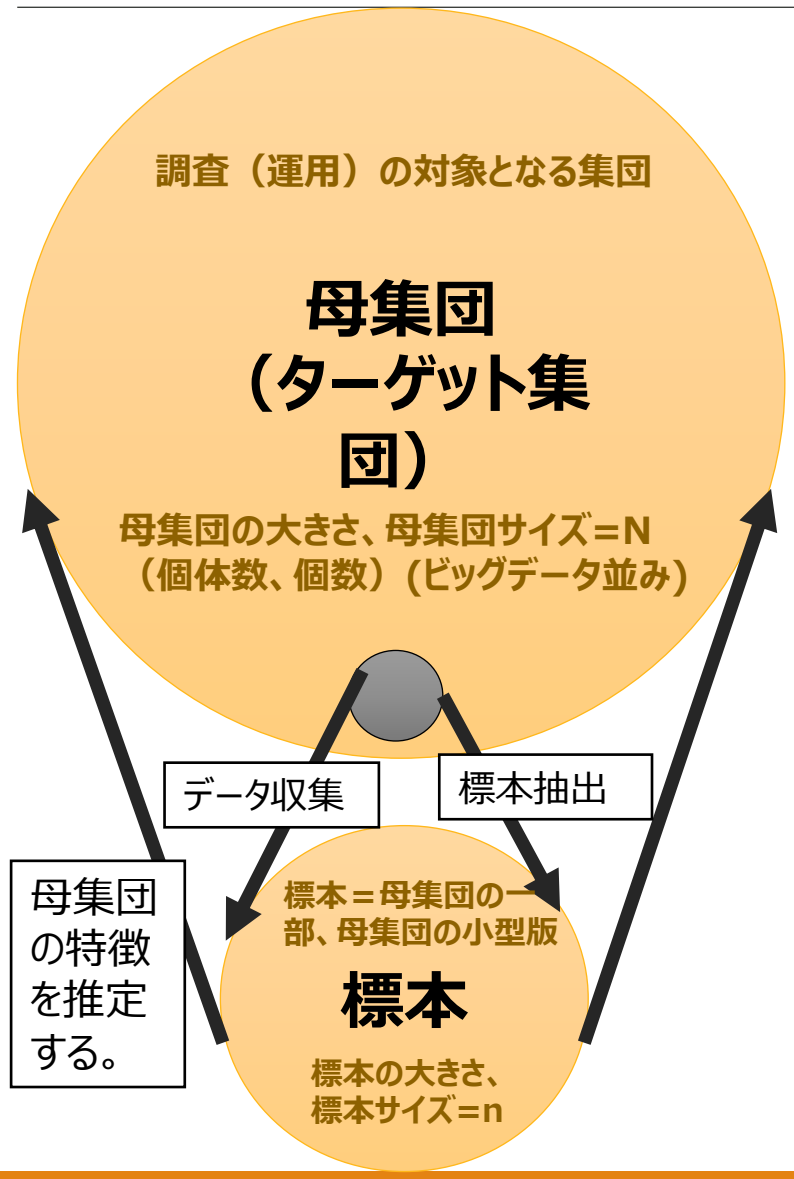
母集団



母集団：

- 調査（運用）の対象となる集団で、ターゲット集団とも言います。
- 母集団の中の個体数＝標本サイズ＝N
- 調査の例1：日本でのインターネット利用状況を調べたいです。
- 例1の対象となる母集団：日本のインターネットの利用者と非利用者（日本の住民全体）。
- 調査の例２：企業Aが顧客に新商品を提供したいです。
- 例2の対象となる母集団：企業Aの既存顧客と既存以外の新規顧客。

標本



標本：

- 標本は母集団の一部、母集団の小型版です。
- 母集団から偏りなく標本を抽出して、標本の特徴を調査して、その母集団の特徴を推定します。

調査の種類：

- 全数調査（センサス）：母集団を全て調べること。例：国民全員を調べる「国勢調査」
- 標本調査（サーベイ）：母集団の一部を調べること。例：工場での抜き取り検査、街頭アンケートなど。

データの種類

母集団から得た標本のデータは、数値であったり、文字であったりします。
基本統計量はデータの種類によって、違いますので、データの種類を理解しましょう。

定性データ・カテゴリーデータ 分類や種類を区別するためのデータ	＝	質的データ・定性データ	名義尺度	分類の順序に意味が無いもの 単なるラベル、文字データ String	計算不可、カウントは可能	例） 性別、血液型、電話番号
			順序尺度	分類の順序に意味があるもの Integer, String	順序(大小)の比較、数値と近似的に見なして、計算可能にしています。	例） 順位、学年、満足度
定量データ・数値データ 数値として意味のあるデータ	＝	量的データ	連続データ	測定値（長さ、重さ、速さ、時間など） Float, double, real 型	和差積商(+,-,x,/)	例） 身長、体重、経過時間
			離散データ	カウントデータ（個数、件数、人数など） Integer型	和差積商(+,-,x,/)	例） 不良品数、故障数

1次元定量データの基本分析：基本統計量

統計量とは、平均値のように、データを入力した関数の戻り値で、データの要約・集計・計算した尺度です。基本統計量は、基本的な統計量で、用途別の定義があります。

定量データの基本統計量として、以下の指標がよく使われます。

📌 Summary Statistics（要約統計量）

平均、中央値、標準偏差など、データを簡潔に要約する数値指標

📌 Descriptive Statistics（記述統計）

分布、ヒストグラム、五数要約、箱ひげ図など、データの全体的な特徴を記述・要約する統計手法

📌 使い分け
基本的な統計量の算出 → "Summary Statistics“

データの全体的な記述・分析 → "Descriptive Statistics"
基本統計量を活用することで、データの特徴を把握することができます。

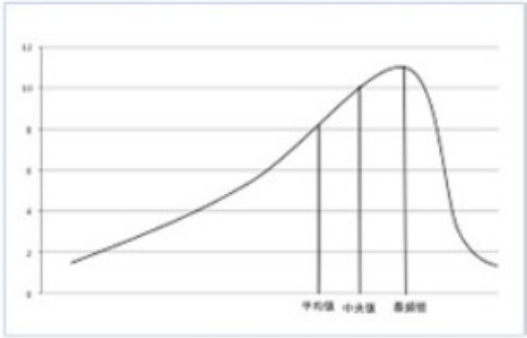
📌基本統計量の種類（定量データの分析）：
①位置の尺度/代表値（データの中心的な傾向）
②散らばり尺度
③形状（データの分布特性）/歪度と尖度

1次元定量データの基本統計量、位置の尺度/代表値

①位置の尺度または代表値（データの中心的な傾向）	説明	例
平均値（MeanまたはAverage）	全てのデータの値を足してデータの数（n）で割った値	データ：1,2,3,4,5 平均値=(1+2+3+4+5)/5=3
中央値（Median）	データを小さい順に並べたときにちょうど真ん中に来る値。真ん中のデータが1つにならず2つになるときは、その平均値	データ：1,2,4,3,5 並べ替えたデータ：1,2,3,4,5 中央値=3
最頻値（Mode）	最も頻繁に出現する値	データ：1,1,2,3,4,5,5,5 最頻値=5
最小（Minimum）	データの中で、一番小さい値	データ：1,2,3,4,5 最小値=1
最大（Maximum）	データの中で、一番大きい値	データ:1,2,3,4,5 最大値=5
順序統計量	**順序統計量：データを大きさの順に並べた際に得られる統計量として、データを4等分する指標、極端に小さい値や大きい値など。** 例：最小値、25%点(第一四分位数、Q1)、50%点・中央値（メジアン）、75%点(第三四分位数、Q2)、最大値などがあります。パーセンタイル5%点：データの下位5%にあたる値、パーセンタイル95%点：データの上位95%にあたる値など。	データ:1,2,3,4,5,6,7 中央値=4 中央値より小さいデータ：1,2,3 第一四分位数（ここでの中央値）=2 中央値より大きいデータ:5,6,7 第三四分位数（ここでの中央値）=6

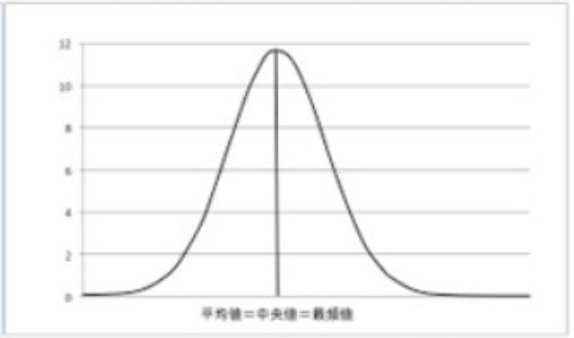
1次元定量データの基本統計量：代表値を求める意義

図　平均値、中央値、最頻値の違いA



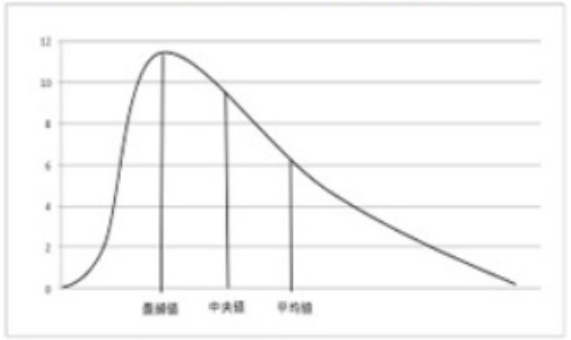
平均値＜中央値＜最頻度

図　平均値、中央値、最頻値の違いB



平均＝中央値＝最頻度

図　平均値、中央値、最頻値の違いC



平均値＞中央値＞最頻度

平均、中央値や最頻値を比較して、データの代表値の意義を理解します。

https://www.stat.go.jp/naruhodo/5_tokucho/chushin.html

1次元定量データの基本統計量：散らばり尺度

②散らばり尺度	説明	例
分散（Variance）	データの値が平均からどれくらい離れているか、全データの偏差（平均との差）の2乗を平均した指標	データ：1,2,3 平均=2 偏差：-1, 0, 1 分散 （不偏分散）= ((-1)^2 + 0^2 + 1^2)/(3-1)=2/2=1
標準偏差（Standard Deviation）	分散の平方根、データのばらつきを示す	データ:1,2,3 平均=2 分散=1 標準偏差 =√(分散)=1
範囲（Range）	データの最大値からデータの最小値を引いた値	データ:1,2,3 範囲 = 3-1 =2
四分位範囲（Interquartile Range, IQR）	第1四分位数（Q1）と第3四分位数（Q3）の差	データ:1,2,3,4,5,6,7 中央値=4 中央値より小さいデータ：1,2,3 第一四分位数（ここでの中央値）=2 中央値より大きいデータ:5,6,7 第三四分位数（ここでの中央値）=6 四分位範囲=6-2=4

1次元定量データの基本統計量：散布度を求める意義

- 以下の3つのデータセットは同じ平均値5を持ちます。
- グラフの縦軸の5の値と、平均の直線を確認してください。
- 平均から各データまでの散らばりの度合いは違います。

分散・標準偏差を用いて、データのばらつきを確認します。

図　散らばりA

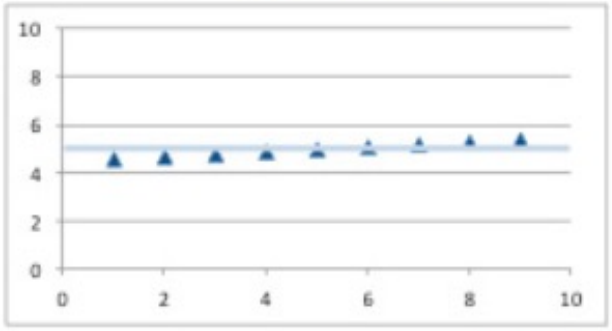


図　散らばりB

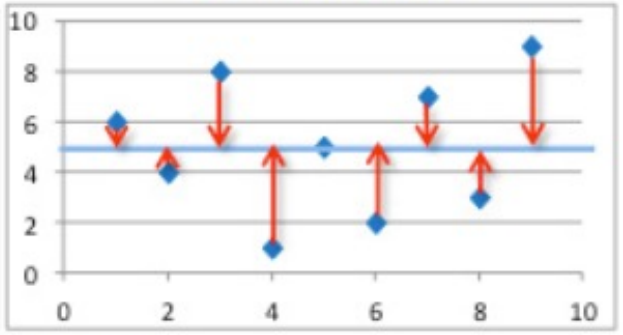
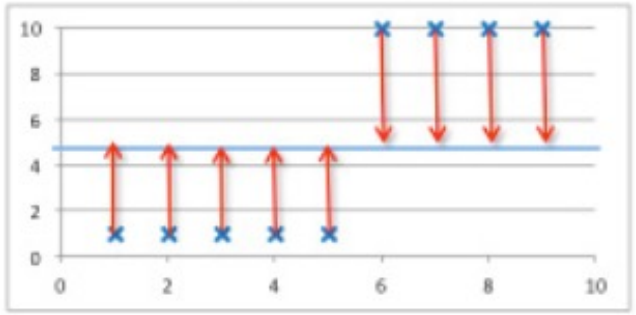
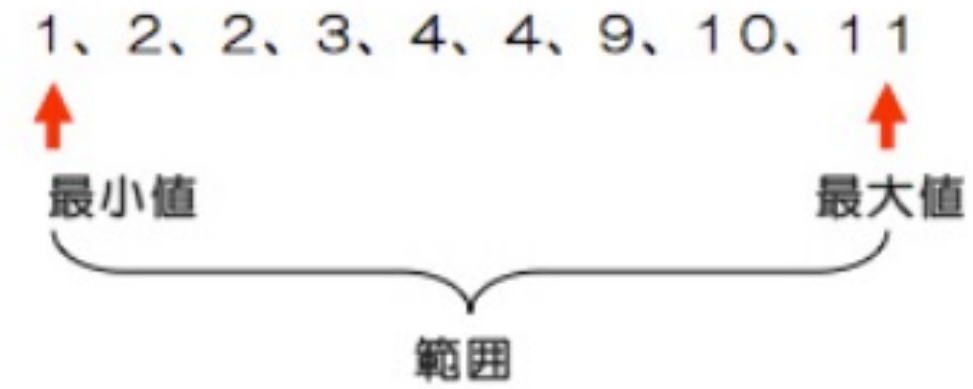


図　散らばりC



https://www.stat.go.jp/naruhodo/5_tokucho/chirabari.html

1次元定量データの基本統計量：最小値、最大値と範囲を求める意義



https://www.stat.go.jp/naruhodo/5_tokucho/chirabari.html

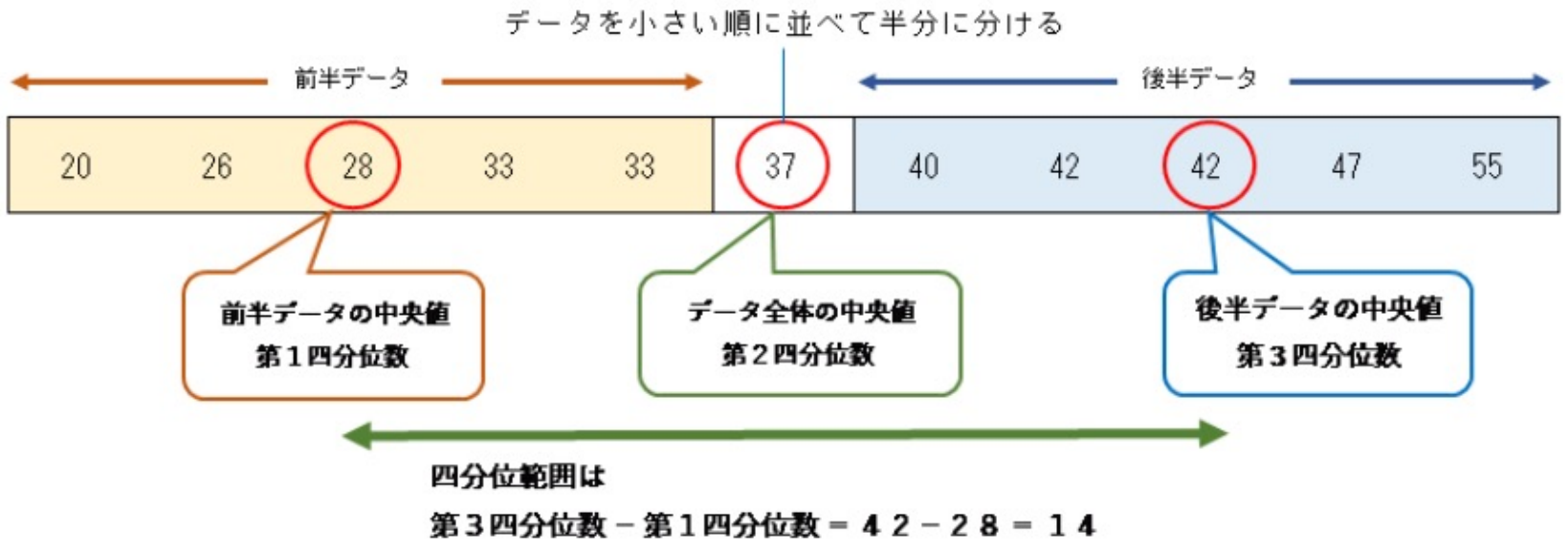
気温データ、化学製品では、平均値よりも最小値、最大値、範囲を求めます。

例：1日の最低気温、1日の最高気温

その意味を確認します。

多くの場合、物理的データや化学製品の仕様として、最大値や最小値がある基準値を超えてはならないという条件があるからです。

1次元定量データの基本統計量：順序統計量を求める意義

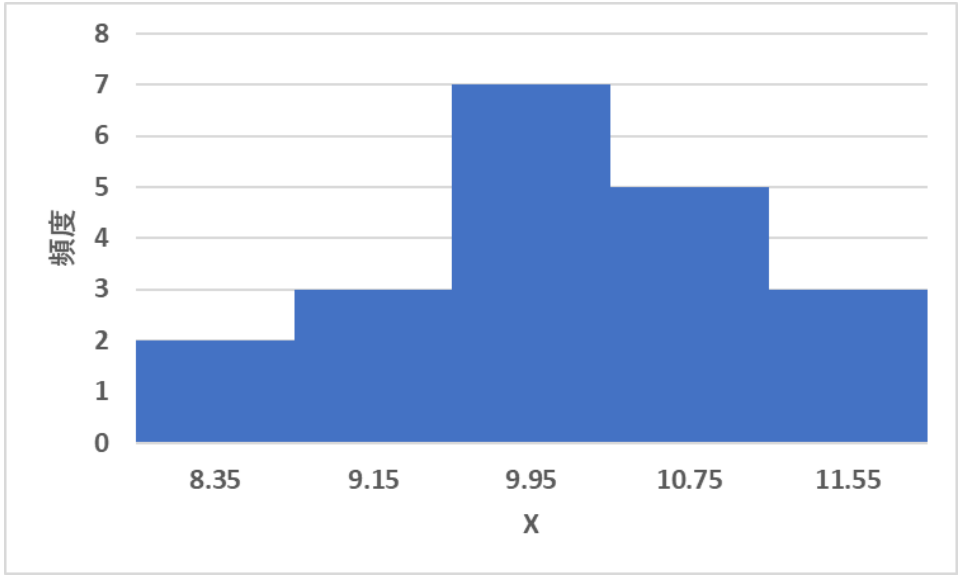


データを4等分する指標として、第1四分位数（Q1）、第2四分位数（Q2, 中央値）、第3四分位数（Q3）を用います。求まった、それぞれの位置で、データの特徴を確認する。

https://www.stat.go.jp/naruhodo/5_tokucho/chirabari.html

1次元定量データの可視化：ヒストグラム

基本統計量とヒストグラムをセットで見る



データ区間下限	データ区間上限	中間値	頻度
7.95	8.75	8.35	2
8.75	9.55	9.15	3
9.55	10.35	9.95	7
10.35	11.15	10.75	5
11.15	11.95	11.55	3

区間の個数と幅を決めてヒストグラムを作成する。

=1+LOG(20,2)
5.3

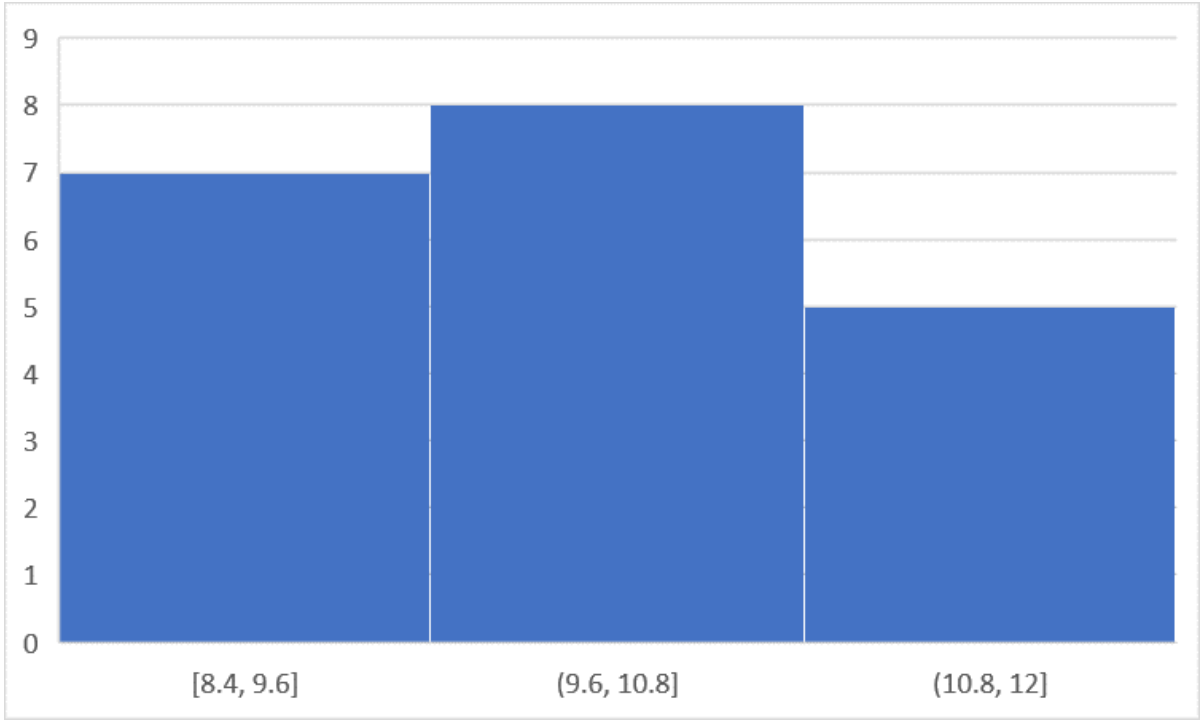
データ
8.4
8.5
9.2
9.5
9.5
9.6
9.6
10
10
10
10.1
10.2
10.6
10.7
10.8
10.9
11.1
11.2
11.4
11.7

データを可視化してみる。

- データは定量値なので、ヒストグラムを作成する。
- ヒストグラム作成の厳格なルールな存在しない。
- スタージェスの式で、ヒストグラムの区間の個数・ビン数は $k = 1 + log2(n)$
- ここでは、$\mathbf{k=1+log2}(20)=5.3$⇒区間の個数は５とする。
- **データの最小値8.4と最大値11.7を含む区間8~12を使うとする。**
- 範囲12－8＝4を5区切りにすると、(12-8)/5=0.8である。
- データは小数点以下1桁あり、単位は0.1である（データは0.1の倍数）。
- 8－データ**単位0.1の半分である0.05**を引いて、7.95からデータ区間を作る。
- EXCELのメニューのデータ⇒データ分析⇒ヒストグラムをクリック⇒データは20件の標本、データ区間はデータ区間上限、出力先を指定し、グラフにチェック⇒OK
- 表とグラフを調整する。

1次元定量データの可視化：Excelでのヒストグラム

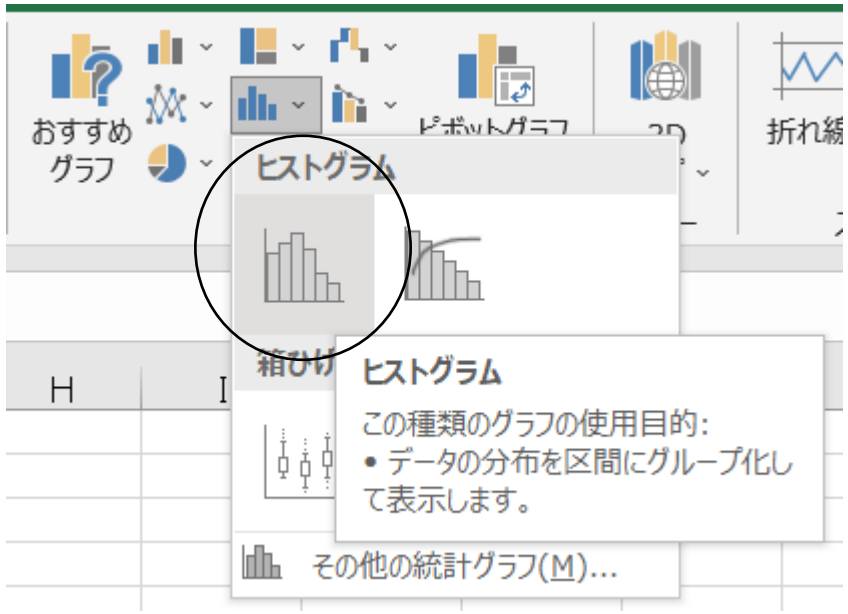
自動でヒストグラムを作成する。



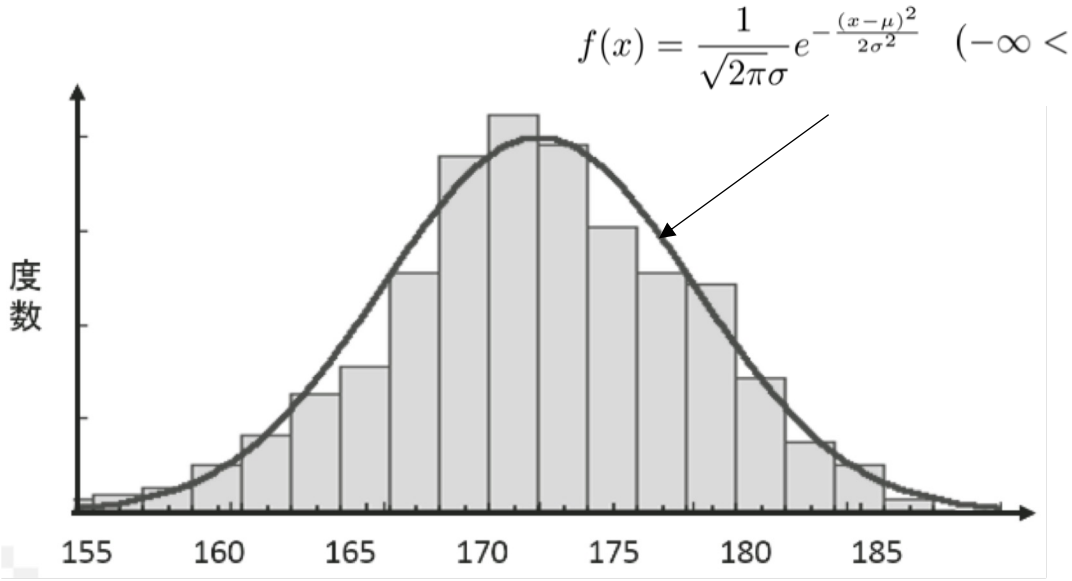
データ
8.4
8.5
9.2
9.5
9.5
9.6
9.6
10
10
10
10.1
10.2
10.6
10.7
10.8
10.9
11.1
11.2
11.4
11.7

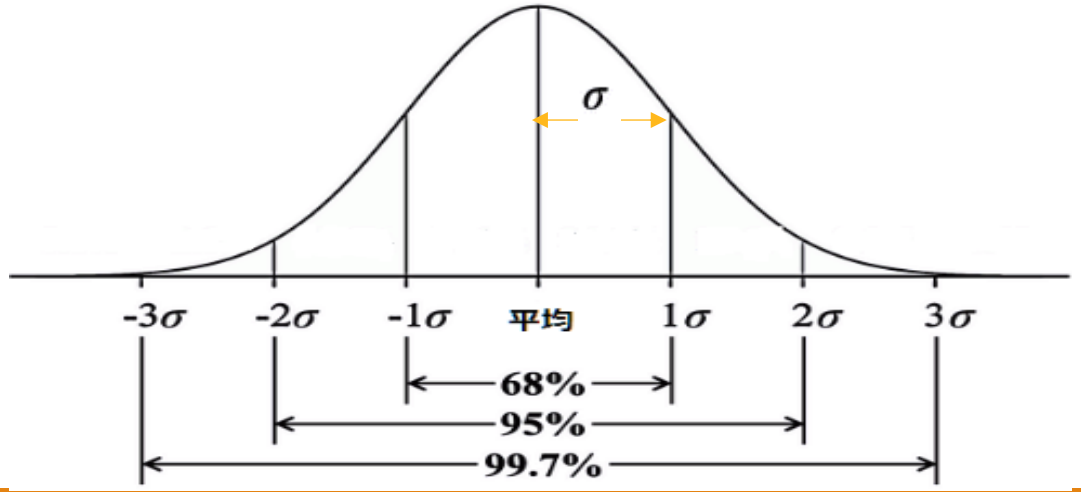
自動でヒストグラムを作成する。

20件のデータをマウスで選択⇒メニューの挿入⇒グラフ⇒ヒストグラムをクリック⇒左上のヒストグラムをクリック⇒グラフを調整する。



1次元定量データの可視化：正規分布とヒストグラム

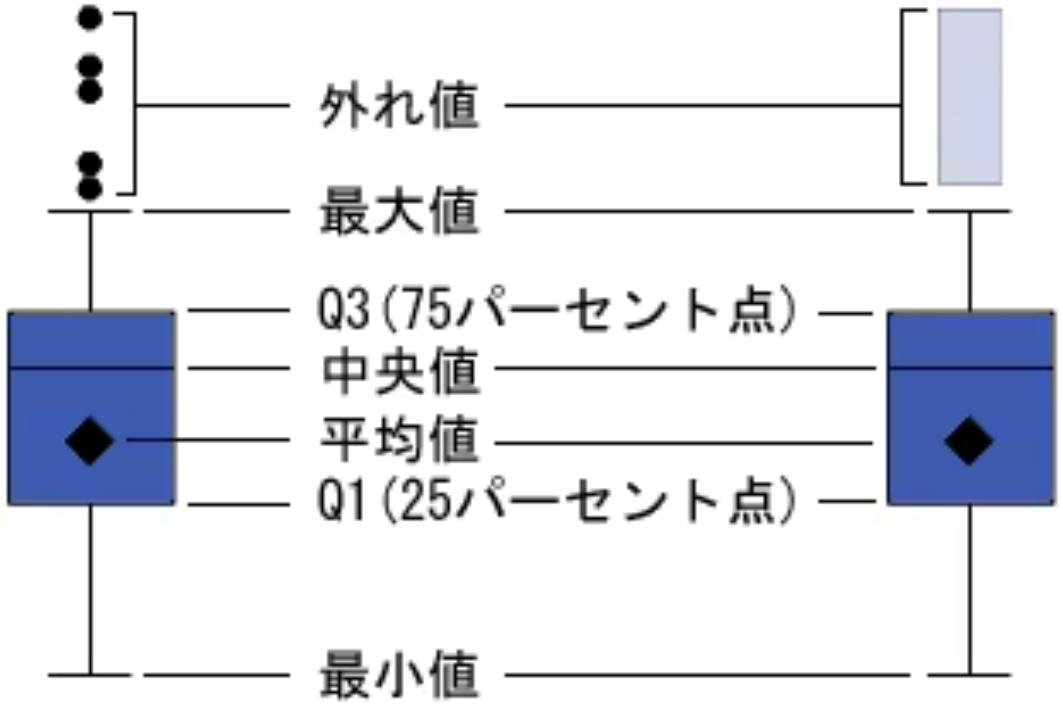




多くのデータは正規分布に従う。要するに、ヒストグラムに、近似的に山の形をした曲線が当てはまる。正規分布の密度関数f(x)が当てはまると、近似的に確率が求まったり、母集団の推定が簡単になる。

- 平均を中心に山は左右対称で、以下の特徴がある。
- 平均±１標準偏差に約68%のデータがある。
- 平均±2標準偏差に約95%のデータがある。
- 平均±3標準偏差に約99.7%のデータがある。

1次元定量データの可視化：箱ひげ図の例

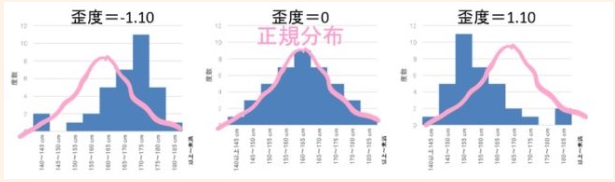
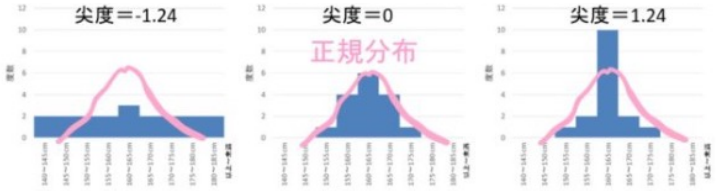


https://support.sas.com/documentation/cdl_alternate/ja/vaug/67500/HTML/default/n0kzo3n26iuhvhn1c1u3hwwa2qtt.htm

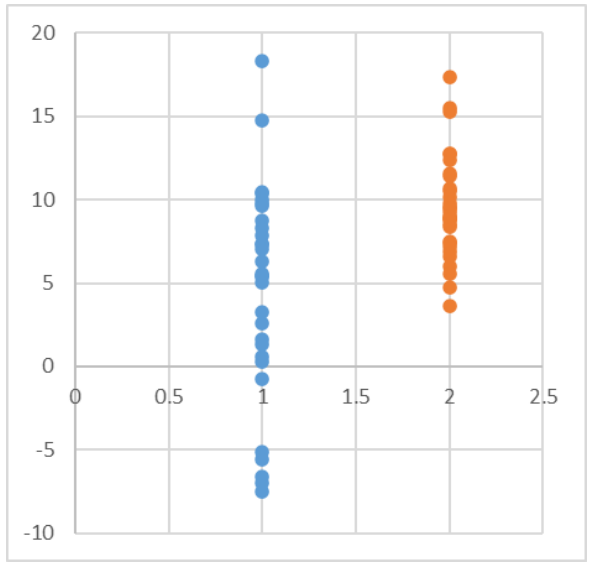
具体例

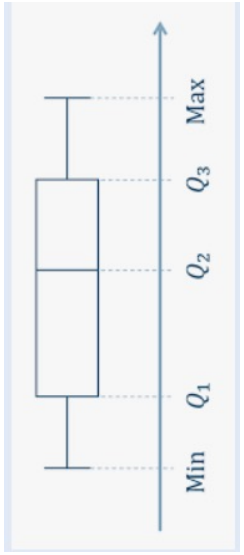
- **例えば、あるクラスの数学のテスト点数を箱ひげ図で表すと、**
  - 箱の長さが短い場合 → 得点のばらつきが小さい（みんな似た点数）。
  - 箱が長い場合 → ばらつきが大きい（高得点と低得点が混ざっている）。
  - ひげが極端に長い場合 → データの分布が広がっている。
  - 外れ値がある場合 → 極端に高得点・低得点の生徒がいる。

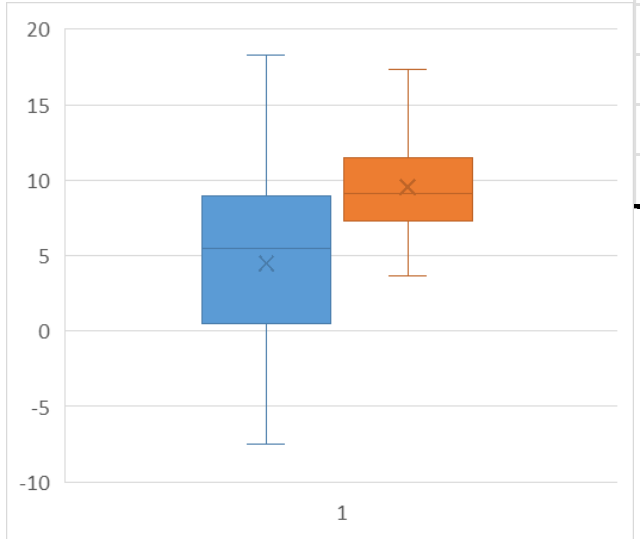
1次元定量データの基本統計量：形状の尺度

③形状（データの分布特性）	説明	例
歪度(わいど Skewness)	データの対称性を示す指標（正なら右に歪み、負なら左に歪み）。歪度 ＞ 0：「右裾が長い」。歪度 ＜ 0：左裾が長い。歪度 ＝ 0：左右対称。 $歪度= n\frac{\sum_1^n\left(\frac{xi-x}{s}\right)^3}{(n-1)(n-2)}$ 	データ：1,2,3 平均=2 偏差：-1, 0, 1 分散 （不偏分散）= ((-1)^2 + 0^2 + 1^2)/(3-1)=2/2=1 標準偏差=√(分散)=1 歪度=3/(2*1)*((-1/1)^3+(0/1)^3+(1/1)^3)=1.5*((-1)+1)=0
尖度(せんど Kurtosis)	データの分布のとがり具合を示す指標。これらの統計量を活用することで、データの特徴を簡単に把握し、分析の基盤を築くことができる。尖度の種類があり、以下の式では、尖度が大きいほど、正規分布より尖った形の分布、小さいほど、正規分布より偏平な形の分布。尖度＝0で、正規分布の形。 $歪度= \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_1^n\left(\frac{xi-x}{s}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$ 	データ：1,2,3,4,5 平均=3 偏差：-2,-1, 0, 1,2 分散 （不偏分散）= ((-2)^2+(-1)^2 + 0^2 + 1^2+2^2)/(5-1)=(4+1+0+1+4)/4=5/2 標準偏差=√(分散)=√(5/2) 尖度 =(30/24)*(16+1+1+16)/((2.5)^2)-8=-1.2

1次元定量データの基本統計量の意義







	青	赤
平均	4.490726	9.541388
標準誤差	1.17471	0.577397
中央値（メジアン）	5.479446	9.10606
最頻値（モード）	#N/A	#N/A
標準偏差	6.434153	3.162533
分散	41.39833	10.00161
尖度	−0.20566	0.382518
歪度	−0.28263	0.559091
範囲	25.8	13.72542
最小	−7.51283	3.61562
最大	18.28717	17.34104
合計	134.7218	286.2416
データの個数	30	30
信頼度(95.0%)(95.0%)	2.402552	1.180909

基本統計量は、何のために使うのでしょうか？

- **2つまたは複数の独立したグループを比較する。**
- **１つのグループの特徴を記述する。**

1次元定性データの基本分析

定性データ（カテゴリーデータ）の基本統計量として、以下の指標がよく使われます。

📌度数（Frequency）と度数分布（Frequency Distribution）

各カテゴリの出現回数をカウントします。

例：アンケートの回答（「満足」「普通」「不満」）のそれぞれの数。

📌相対度数（Relative Frequency）

各カテゴリの度数を全体の合計で割った値（割合）。

例：「満足」50人、「普通」30人、「不満」20人なら、「満足」の相対度数は50%（50/100）。

📌最頻値（Mode）

最も多く出現したカテゴリ。

例：「満足」「普通」「不満」の中で「満足」が最も多ければ、それが最頻値。

📌カテゴリの多様性を示す指標

エントロピー（Shannon Entropy）：カテゴリーの散らばり具合を測る指標。値が大きいほど多様性が高い。

ジニ係数: カテゴリの不均一性を測定。

📌クロス集計（満足度 × 性別）

インデックス	満足度	性別
1	満足	男性
2	普通	女性
3	不満	男性
...	...	...

1. 度数（Frequency）各カテゴリの出現回数：
満足: 56人普通: 33人不満: 11人

2. 相対度数（Relative Frequency）各カテゴリの割合：
満足: 56%普通: 33%不満: 11%

3. 最頻値（Mode）最も多く選ばれたカテゴリ：
「満足」

4. エントロピー（多様性の指標）Shannon エントロピー:
1.35（値が大きいほど多様性が高い）

1次元定性データの基本分析：度数分布表と可視化

データの集計	説明	例
度数（Frequency）と度数分布（Frequency Distribution）	各カテゴリの出現回数をカウントします。	例：アンケートの回答のそれぞれの数（例：「満足」50人、「普通」30人、「不満」20人）
相対度数（Relative Frequency）	各カテゴリの度数を全体の合計で割った値（割合）。	例：「満足」50人、「普通」30人、「不満」20人なら、「満足」の相対度数は50%（50/100）。
最頻値（Mode）	最も多く出現したカテゴリ。	例：（例：「満足」50人、「普通」30人、「不満」20人）の中で「満足」が最も多いので、それが最頻値。

1次元定性データの基本分析：カテゴリの多様性を示す指標

カテゴリの多様性を示す指標	説明	例
エントロピー（Shannon Entropy）	カテゴリーの散らばり具合を測る指標。 値が大きいほど多様性が高い。 エントロピーが高い → データのカテゴリ分布が均等エントロピーが低い → ある特定のカテゴリに偏りがある。 エントロピー H(X) は、次の式で計算されます。 $H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$	ケースA：「満足」「普通」「不満」がそれぞれ 1/3ずつ 含まれる分布が均等 → エントロピーは 高い H(X)=−(0.33log_2(0.33)+0.33log_2(0.33)+0.33log_2(0.33))=−(0.33×−1.585+0.33×−1.585+0.33×−1.585=1.585 ケースB：「満足」が 90% で「普通」「不満」が各5%ほとんど「満足」 → エントロピーは 低い H(X)=−(0.9log_2(0.9)+0.05log_2(0.05)+0.05log_2(0.05))=−(0.9×−0.152+0.05×−4.322+0.05×−4.322)=0.64
ジニ係数	ジニ係数は データの不平等の度合い を測る指標で、特に 所得格差 や カテゴリデータの偏り を評価する際に使われます。 値の範囲: G≈0⇒完全に均等、すべてのカテゴリが同じ割合（例: 3つのカテゴリが33%ずつ） G=0.2 ～ 0.4⇒やや均等、ある程度の偏りがある（例: 40%・35%・25%） G=0.4 ～ 0.6⇒中程度の不平等、明らかに偏っている（例: 60%・30%・10%） G≈1⇒極端な偏り、ほぼ1つのカテゴリが支配的（例: 90%・5%・5%） $G = 1 - \sum_{i=1}^{n} P(x_i)^2$	ケースA: 「満足」「普通」「不満」がそれぞれ 1/3ずつ 含まれる G=1−(0.33^2 +0.33^2 +0.33^2) =1−(0.1089+0.1089+0.1089 = 0.673 ケースB：「満足」が 90% で「普通」「不満」が各 5% G=1−(0.9^2 +0.05^2 +0.05^2) =1−(0.81+0.0025+0.0025) =0.185

出典

Appendices

出典元URLの説明

スライド　掲載URL

3　https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/html/nd123520.html

＞総務省　令和5年版　情報通信白書

11他　https://www.stat.go.jp/naruhodo/5_tokucho/chushin.html

＞統計局「なるほど統計学園」