1.2.1 記述統計量と表計算ソフトウェア(2)

教科書：P.84～89

★表計算ソフトウェアを使った記述統計量の処理

今回の学習
- 表計算ソフトウェアを使って記述統計量を求めたり，グラフで表現したりする方法を身につけよう
- 「外れ値」と「異常値」や「疑似相関」といったデータの中に隠れた関係の考え方を理解しよう

(1) 関数とは？　〔⇒実習シート：処理〕

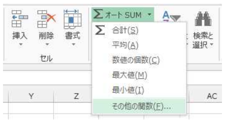
Excel では，必要な数値などを指定するだけでカンタンに計算結果を求められる「①　　　　」という機能であり，次のように表現される。

=関数名（引数1，引数2，引数3，……，引数n）

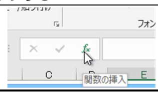
※（②　　　）：数値やセル番地，数値を入力した範囲，関数などのことである
- 複数の引数がある場合は，「，」で区切って並べる
- セル番地の範囲を指定する場合は，矩形領域（長方形上の領域）で，左上端と右下端のセル番地を「：」で区切って表す

● 関数の挿入方法（一般的なもの）

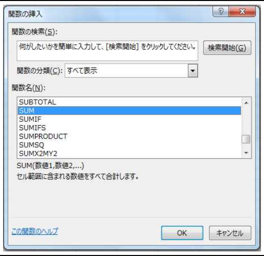
①「ホーム」タブの「編集」グループ「オートSUM」→「その他の関数」



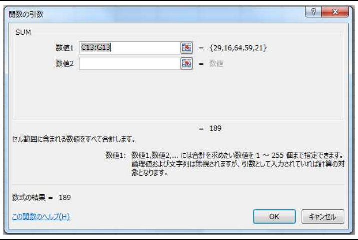
①「数式バー」にある f_x のボタンをクリックする



②出てきたウィンドウから，関数を選択
※関数の分類：「すべて表示」



③出てきたウィンドウで，引数などを設定



なお，①②については，「数式」タブの「関数ライブラリ」グループからある程度選択できる。

● データの分析で使う関数

1変数のデータ

	関数名	引数の設定
平均値	AVERAGE	
中央値	MEDIAN	
最大値	MAX	
最小値	MIN	
四分位数	QUARTILE.INC	「戻り値」の設定は，以下のようになる 「1」…第1四分位数，「2」…第2四分位数 「3」…第3四分位数 (注)　Excel では，数学の教科書とは異なる定義で計算している［最小値を0%，最大値を100%としたとき，25%，50%，75%にあたる値］
分散	VAR.P	
標準偏差	STDEV.P	

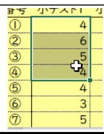
2変数のデータ

	関数名	引数の設定
共分散	COVARIANCE.P	「配列1」には「1種類目のデータ」， 「配列2」には「2種類目のデータ」を指定する
相関係数	CORREL	

(2) グラフをかく　〔⇒実習シート：処理〕

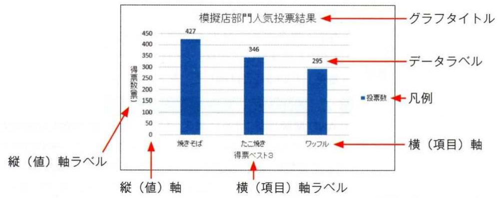
● グラフのかき方（一般的なもの）

①グラフをかきたいデータの部分を選択

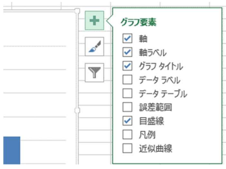


②「挿入」タブの「グラフ」グループからかきたいグラフを選択

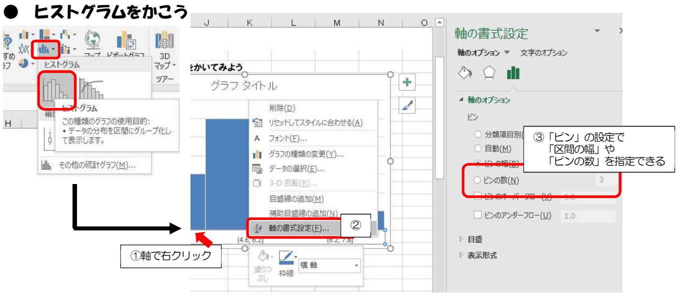
● グラフの主な要素



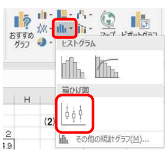
Excel では，グラフでどのような要素を表示させるかを，グラフの右上にある ＋ ボタンで選択することができる。



● ヒストグラムをかこう



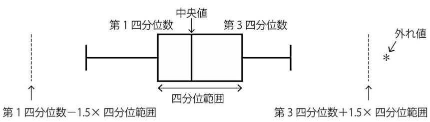
● 箱ひげ図をかこう



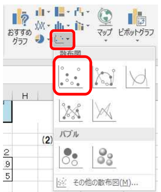
ここで，「小テスト2」で箱ひげ図をかくと，最大値は「10」であるのにひげの1つが「8」となり，「10」のところが「点」となる。これは，「10」がこのデータにとっての「③　　　　」だからである。

「③」…他の値から極端にかけ離れたデータ
（④　　　　）の（⑤　　　　）倍以上離れた値
（標準偏差の2倍以上離れた値という定義もある）

※「×」は平均値である



● 散布図をかこう



（③）は除外すべき値（＝⑥　　　　）と捉えがちだが，その背景を探ることが大切である。測定ミスや入力ミスでなければ，そこに問題発見や問題解決の手がかりがあることもあるからである。

(3)「分析1」に取り組んでみよう　〔⇒実習シート：分析1〕

イギリスの統計学者 ロナルド・フィッシャーが上げたことで有名な「アヤメ」に関するデータの分析をしてみよう。
下表は，アヤメの花のがくについて，「長さ」と「幅」を調べた結果である。
(1) この2つについて，相関関係があるか調べよ。
(2) 花が大きくなればそれにあわせてがくの「長さ」が長くなれば，「幅」も大きくなると考えられるが，相関関係としては(1)のような結果になった。なぜ，この予想に反する結果になったのか，理由を考えよ。
(3)《授業内で追加指示します》

長さ	幅
5.0	3.3
4.6	3.4
4.6	3.6

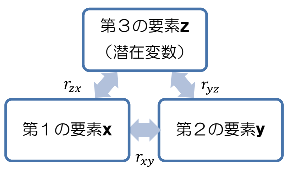
「相関関係」については，データを（⑦　　　　　）すぎたり，（⑧　　　　　）すぎたりすることで，大幅に結果が変わってくる。単に相関係数を求めるだけでなく，散布図や四分位数などを用いて，対象とするデータが適切かを慎重に評価する必要がある。

(4)「分析2」に取り組んでみよう　〔⇒実習シート：分析2〕

あなたは新しい高校の設置の準備を担当している。ある日，県民の方から次のような連絡があった。「私の住んでいる県に高校を作ってもらっては困る。高校ができると，覚せい剤を使う人の数が増えてしまうからだ。何とかならないか」
あなたは仕事の仲間に頼んで，高校の数と覚せい剤の取締送致件数を調べてもらった。この結果について，以下の2つのことについて考えてみよう。
(1) 2つのデータに相関関係があるか調べよ。
(2) 2つのデータの間に因果関係があるといえるか。言えないとすれば，何が関係していると考えられるか。

都道府県	高校数 2014年	覚せい剤取締送致件数 2012年
北海道	290	649
青森	82	52
岩手	81	44

「相関係数」を計算すると，「強い正の相関関係がある」と出てくる。しかし，この2つの要素の間に相関関係があるからといって，そこに（⑨　　　　　）があるわけではない。今回の場合，この2つの要素の間に「第3の要素」（潜在変数）がある。このような場合を「⑩　　　」という。（⑩）であるかを見るためには，潜在変数の影響を除いた相関関係である「⑪　　　」を求めればわかる。



（偏相関係数）

$$= \frac{r_{xy} - r_{zx}r_{yz}}{\sqrt{1-r_{zx}^2}\sqrt{1-r_{yz}^2}}$$

ただし，r_{xy}はxとyの相関係数，（以下略）

Excel での数式
(1) 平方根　（⑫　　　　）関数
(2) 累乗　（⑬　　　　）

1.2.3 推測統計学入門(4)

教科書：（なし）

★正しい？ それとも誤っている？ 〜検定〜

今回の学習
・ものごとについて数学的に「正しい」というための方法を使えるようになろう。

(1) 検定とは何か？

次の例題から，「検定」とはどのようなものかを考えてみる。

問題1
ある立方体のさいころを 180 回投げた結果，１の目が 59 回出た。このさいころの１の目が出る確率は，1/6 と考えてよいか？

[問題を整理する]
さいころの１の目が出る確率を p とする。
このとき，仮説「p は 1/6 である」が「正しい」(＝妥当である) と言えるのか「誤っている」(＝妥当ではない) と言えるのかを，標本から得られる情報を使って判断する問題である。
このように，標本から得られる情報を使って妥当であるか否かを判断することを（①　　　　）という。

[考え方と解答]

(Step 1) 帰無仮説と対立仮説を（数式で）記述する

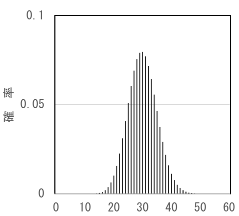
（ ① ）において，検定すべきもととなる「※成立しない」と思われる仮説を（②　　　　）[記号：(③　　　)]，（ ② ）に相対する「※成立する」と思われる仮説を（④　　　　）[記号：(⑤　　　)] という。
※ （ ② ）を等式で立てる必要がある。そのため，成立するか否かが逆転することもある。
今回は，帰無仮説 $H_0 : p = 1/6$　　　　対立仮説 $H_1 : p \neq 1/6$

(Step 2) 有意水準を確認する

（⑥　　　　　　）：めったに起こらないことが起こる確率，すなわち誤った結論を導き出す確率のこと。多くの場合は，「5%」または「1%」である
今回は，5%で調べてみる

(Step 3) 検定統計量 U の従う分布を確認する

H_0 の下での U を確認しておく
今回であれば，１の目が出た回数 U は（⑦　　　　）の（⑧数式　　　　　）に従う。このときのグラフは右のようになる



(Step 4) 棄却域を確認する

有意水準以下の確率で起こる範囲（＝⑨　　　　）を求める
今回は U が従う二項分布 $B(180, 1/6)$ から，直接確率が 5%以下になるときの U の範囲を求めることは大変である。しかし，n=180 は大きいので中心極限定理により，U は近似的に（⑩　　　　　）の（⑪数式　　　　　）に従うといえる。また，

$$Z = \frac{U-30}{\sqrt{25}} = \frac{U-30}{5} \quad \cdots(*)$$

と変換すれば，Z は標準正規分布 $N(0, 1)$に従う。このとき，（ ⑨ ）は正規分布表から，確率が 5%のときの $z < -1.96,\ 1.96 < z$ となる。

(Step 5) 検定統計量の値が棄却域にあるか確認する

検定統計量 U の値が（ ⑨ ）に入らなかった場合は，今回起こったことが「めったに起こらないこと」とはいいきれないので，(⑫　　　　　　)を棄却できない(このとき，(⑫) を受容する)。一方，U の値が棄却域に入った場合は，今回起こったことが「めったに起こらないこと」といいきることができるので，(⑫) を棄却し，(⑬　　　　　　) を採択する。今回の実験の結果の「59 回」は，(*)より

$$z = \frac{59-30}{5} = 5.8$$

である。これは（ ⑨ ）に入っているので，（ ⑫ ）は棄却され，（ ⑬ ）が採択される。すなわち，さいころの１の目が出る確率は，1/6 と考えられない。

（解答）
帰無仮説 $H_0 : p = 1/6$　　　対立仮説 $H_1 : p \neq 1/6$　とする
有意水準を 5%として，検定を行う
H_0 の下で検定統計量 U は，二項分布 $B(180, 1/6)$に従う
ここで，n=180 は大きいので中心極限定理により，U は近似的に正規分布 $N(30, 25)$に従うといえる。また，

$$Z = \frac{U-30}{\sqrt{25}} = \frac{U-30}{5} \quad \cdots(*)$$

と変換すれば，Z は標準正規分布 $N(0, 1)$に従う。
このとき，棄却域は正規分布表から $z < -1.96,\ 1.96 < z$ と決まる。
今回の実験の結果の「59 回」は，(*)は $z = \frac{59-30}{5} = 5.8$ である。
これは，棄却域に入っている (＝5%水準で有意な) ので，帰無仮説は棄却され，対立仮説が採択される。すなわち，さいころの１の目が出る確率は，1/6 と考えられない

問題2
海外旅行中，とある路地裏で１枚のコインで賭けをする賭博師を見かけた。「表が出たらお客さんが１ドル，裏が出たらこっちが１ドル。さあ，この賭けをやらないかい」。コインに不正があるようで，実際に調べたところ８回中７回裏が出た。「そのコインの表裏の出る確率は等しくないのでは？」と声をかけると，賭博師は事も無げに「そんなことはない。偶然だよ」と応えた。
そのコインの表の出る確率 p と裏の出る確率 q は等しいといえるか？　有意水準を５%として検定せよ。また，実験回数が８回と少ないが，中心極限定理を用いてよい。

（解答）
帰無仮説 H_0：(1　　　　)　　対立仮説 H_1：(2　　　　)　とする
有意水準を 5%として，検定を行う
H_0 の下で検定統計量 U は，(3　　　　　　　　　)に従う
ここで，中心極限定理により，U は近似的に (4　　　　　　　　) に従うといえる。
また，　　　　5
　　　　　　　　　　　　　　$\cdots(*)$
と変換すれば，Z は標準正規分布 $N(0, 1)$に従う。
このとき，棄却域は正規分布表から，(6　　　　　　　　　) と決まる。
今回の実験の結果の「表１回」は，(*)より (7　　　　　　　　　)。
これは，棄却域に入って (8　　　　　) (＝5%水準で有意な) ので，帰無仮説は (9　　　　) され，対立仮説が (10　　　　) される。すなわち，コインの表裏の出る確率は等しくないと考えられ (11　　　)

問題3
エンドウ豆の交配で，雑種第二代において，黄色と緑色の豆のできる割合はメンデルの法則に従えば３：１である。ある実験で黄色が 428 個，緑色が 132 個得られたという。この結果はメンデルの法則に反するといえるか。有意水準を５%として検定せよ。ただし，$\sqrt{105} = 10.25$ を用いてもよい。

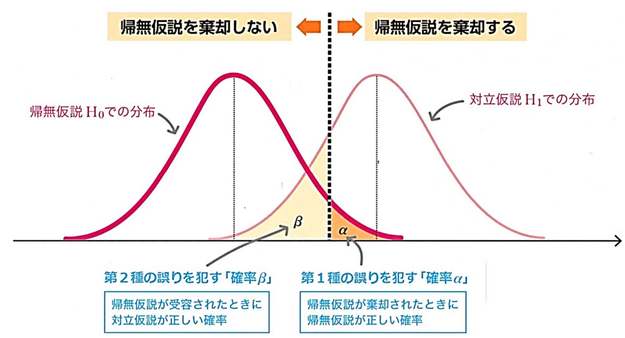
（解答）
帰無仮説 H_0：(1　　　　　　　　　　)，
対立仮説 H_1：(2　　　　　　　　　　)　とする
560 個中の黄色の豆の個数を U，有意水準を 5%として，検定を行う
H_0 の下で検定統計量 U は，(3　　　　　　　　　)に従う
ここで，中心極限定理により，U は近似的に (4　　　　　　　　　) に従うといえる。
また，　　　　5
　　　　　　　　　　　　　　$\cdots(*)$
と変換すれば，Z は正規分布 $N(0, 1)$ に従う。
このとき，棄却域は正規分布表から，(6　　　　　　　　　) と決まる。
メンデルの実験の結果の「黄色が 428 個」は，(*)より (7　　　　　　　　　)。
これは，棄却域に入って (8　　　　　) ので，帰無仮説は (9　　　　) され，対立仮説が (10　　　　) される。すなわち，メンデルの法則に反すると考えられ (11　　　)

(3) 検定に関する注意点

検定の方法は，「帰無仮説 H_0 を棄却して，対立仮説 H_1 を採択できるか」を統計的に検証する方法である。この場合，考えられる判定の可能性は下図のように４つの場合がある。

事実 判定	H_0 が正しい H_1 が誤り	H_0 が誤り H_1 が正しい
H_0 を棄却しない ＝H_1 を採択しない	○(正しい判定) ※H_0 が支持されたわけではない	×「⑮　　　　　」 確率 β
H_0 を棄却する ＝H_1 を採択する	×「⑭　　　　　」 確率 α	○(正しい判定) 確率 $\gamma = 1 - \beta$

「 ⑭ 」は，標本調査の結果が棄却域に入ったときに起こる。棄却域の「確率 α」を有意水準と呼ぶが，帰無仮説 H_0 が正しいときでも，標本調査の結果が棄却域に入ることは「確率 α」で起こる。そこで，「 ⑭ 」を犯す確率は，棄却域を決める「有意水準 α」と考えられる。
「 ⑮ 」は，帰無仮説 H_0 が受容されたときに対立仮説 H_1 が正しい確率を表す（このときの確率を β とする)。
下図からわかるように，「α を小さくすれば，β は大きくなる」という関係にある。



「⑯　　　　　　」以上に起こってはならない誤りが「⑰　　　　　　」である。「帰無仮説 H_0 が誤っているときに，帰無仮説 H_0 が棄却される確率 γ」のことを (⑱　　　　) という。
このとき，検出力 γ ＝ (⑲数式　　　　) と表せる。

(4) 母平均の検定

(注) この節では，母集団は正規分布に従っているものとする。

母平均の検定も，これまでの検定と同じ手順で行われる。しかし，母分散がわかっているか否かによって，検定統計量やそれが従う分布が異なる。

母分散が	検定名称	検定統計量	検定統計量が従う分布
わかっている or わかっていないが 標本の大きさが 大きいとき (標本分散を σ^2 とする)	u 検定	$Z=\dfrac{\overline{U}-m}{\dfrac{\sigma}{\sqrt{n}}}$ m：母平均 σ：母標準偏差 $\overline{U}$：標本平均	標本平均 $\overline{U}$ は 正規分布 $N\left(m,\ \dfrac{\sigma^2}{n}\right)$ に従う。 また，Z の変換により， 標準正規分布 $N(0, 1^2)$ に従う。
わかっていない	t 検定	$T=\dfrac{\overline{U}-m}{\dfrac{s}{\sqrt{n}}}$ m：母平均 s：(不偏分散の) 標準偏差 $\overline{U}$：標本平均	T は 自由度 $n-1$ の t 分布 に従う。

問題4
あるスーパーで売られている肉のパックは 1kg と表示されているが，16 個を抽出して測ったところ，平均値が 998g だった。このスーパーでは故意に表示と異なる価格でパックしているといえるだろうか。次の３つの場合について，有意水準 5%で検定せよ。
(1) この店の秤が古くて，母標準偏差が 4.0g であることがわかっている場合

(解答)
母平均を m，標本平均を $\overline{U}$ とする。
帰無仮説 H_0：(1　　　　　)　対立仮説 H_1：(2　　　　　)　とする
有意水準を 5%として，(注意)両側で (3　　　　) を行う
H_0 の下で検定統計量 $\overline{U}$ は，(4　　　　　　　　) に従う
ここで，　5　　　　　　…(*)

と変換すれば，Z は標準正規分布 $N(0, 1)$ に従う。
このとき，棄却域は正規分布表から (6　　　　　　　　) と決まる。
今回の平均値「998g」は，(*)より (7　　　　　　　　)。これは，棄却域に入って (8　　　　) ので，帰無仮説は (9　　　) され，対立仮説が (10　　　) される。
すなわち，このスーパーでは故意に表示と異なる価格でパックしていると (11　　　　)

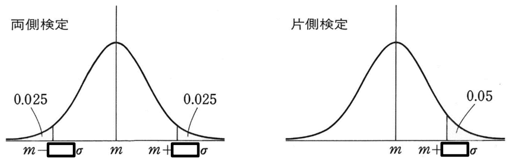
(2) 秤については不明だが，標本の不偏分散が $16.0g^2$ の場合

(解答)
母平均を m，標本平均を $\overline{U}$ とする。
帰無仮説 H_0：(1　　　　　)　対立仮説 H_1：(2　　　　　)　とする
有意水準を 5%として，(注意)両側で (3　　　　) を行う
H_0 の下で検定統計量 $\overline{U}$ は，
　4　　　　　　…(*)

と変換すれば，T は自由度 (5　　　) の (6　　　) に従う。
このとき，棄却域は t 分布表から (7　　　　　　　) と決まる。
今回の平均値「998g」は，(*)より (8　　　　　　　　)。これは棄却域に入って (9　　　　) ので，帰無仮説は (10　　　) され，対立仮説が (11　　　) される。すなわち，このスーパーは故意に表示と異なる価格でパックしていると (12　　　)

(注意) 問題4では，仮説に対して，平均値が異常に大きくても，または異常に小さくても仮説が棄却されるように，棄却域を両側にとっている。このような検定を (⑳　　　　　) という。
これに対して，平均値が「異常に大きい」または「異常に小さい」に限定した上で，棄却域を片側にとる検定を (㉑　　　　　) という。
問題として出されるときは，次のように考えて差し支えない
「……と異なるといえるか」　　　　なら　(㉒　　　　　)
「……より大きい (小さい) といえるか」　なら　(㉓　　　　　)
※「……より大きい」の場合，棄却域を平均値の (㉔　　　　) に設定することから (㉕　　　　　)，「……より小さい」の場合，棄却域を平均値の (㉖　　　　) に設定することから (㉗　　　　　) ということがある

(有意水準 5%のときの正規分布のグラフ)
□の中については，正規分布表を見て自分で埋めてみよう



(3) この店の秤が古くて，母標準偏差が 4.0g であることがわかっている場合
このスーパーでは故意に少なめにパックしているといえるか

(解答)
母平均を m，標本平均を $\overline{U}$ とする。
帰無仮説 H_0：(1　　　　　)　対立仮説 H_1：(2　　　　　)　とする
有意水準を 5%として，(3　　　　　　) で (4　　　　) を行う
H_0 の下で検定統計量 $\overline{U}$ は，(5　　　　　　　　) に従う
ここで，　6　　　　　　…(*)

と変換すれば，Z は標準正規分布 $N(0, 1)$ に従う。
このとき，棄却域は正規分布表から (7　　　　　) と決まる。
今回の平均値「998g」は，(*)より (8　　　　　　　)。これは，棄却域に入って (9　　　　) ので，帰無仮説は (10　　　) され，対立仮説が (11　　　) される。すなわち，このスーパーは故意に少なめにパックしていると (12　　　　)

問題5
ある化学製品の合成時の反応温度はこれまで 70.0 度とされてきた。最近原料の購入先を変更したので，反応温度が変わるかどうかを検証することになった。そのためにデータを 10 個とったものが，下記の値である。有意水準 5%で両側検定せよ。ただし，$\sqrt{5.38}=2.32$，$\sqrt{10}=3.16$ を利用してもよい。　69，74，74，71，69，72，73，68，74，70

(解答)
母平均を m，標本平均を $\overline{U}$，不偏分散を s^2 とする。

$\overline{U}=$ (1　　　　　　　　　　　　　　　　　　)

$s^2=$ (2　　　　　　　　　　　　　　　　　　)

帰無仮説 H_0：(3　　　　　)　対立仮説 H_1：(4　　　　　)　とする
有意水準を 5%として，両側で (5　　　　) を行う
H_0 の下で検定統計量 $\overline{U}$ は，
　6　　　　　　…(*)

と変換すれば，T は自由度 (7　　) の (8　　　) に従う。
このとき，棄却域は t 分布表から (9　　　　　　　) と決まる。
今回の平均値「71.4g」は，(*)より (10　　　　　　　　)。これは棄却域に入って (11　　　　) ので，帰無仮説は (12　　　) され，対立仮説が (13　　　) される。すなわち，最近原料の購入先を変更したので，反応温度が変わると (14　　　)