10 章 質的データの検定

10.1 １標本の比率の検定

例

工場である期間内におきた事故の件数を曜日毎に調べたところ、以下の表が得られた。事故は曜日によるばらつきがある（一様でない）といえるか？有意水準 5%で判定せよ。

曜日	月	火	水	木	金	計
事故件数	23	14	16	11	16	80

理論　適合度検定

n 回の観測の中で、事象 1 は n_1 回、事象 2 は n_2 回、・・・、事象 k は n_k 回起こるとする。出現比率は想定比率 $p_1, p_2, \cdots, p_k$ に比べて差があるといえるか。出現の想定値を $m_1, m_2, \cdots, m_k$ （$m_i = np_i$）として、$\alpha \times 100\%$ の有意水準で判定せよ。

帰無仮説 H_0：事象 i の出現比率は p_i　　（想定比率と比べて差がない）

対立仮説 H_1：H_0 でない　　（想定比率と比べて差がある）

$$\chi^2 = \frac{(n_1 - m_1)^2}{m_1} + \frac{(n_2 - m_2)^2}{m_2} + \cdots + \frac{(n_k - m_k)^2}{m_k} \sim \chi^2_{k-1} \text{分布}$$

$p = chidist(\chi^2, k-1)$ ［$\chi^2_{k-1}(p) = \chi^2$］として、

$p < \alpha$ のとき、H_0 は棄却し、H_1 を採択する。

$p = chidist(\chi^2, d)$　両側検定

解答

帰無仮説 H_0：毎曜日一様（確率$1/5$）に起こっている。

対立仮説 H_1：一様とはいえない。

一様と考えると、$m_1 = m_2 = \cdots = m_5 = 80/5 = 16$

$$\chi^2 = \frac{(23-16)^2}{16} + \frac{(14-16)^2}{16} + \frac{(16-16)^2}{16} + \frac{(11-16)^2}{16} + \frac{(16-16)^2}{16}$$

$$= \frac{78}{16} = 4.875$$

$p = chidist(4.875, 4) = 0.300365$

$p > 0.05$ より一様でない（想定比率と差がある）といえない。

問題

ある大学（学生数 1200 名）の学生 50 人を任意抽出し、大学改革のアンケートを行ったところ、賛成 35 反対 15 であった。学生の過半数が賛成している（賛成の比率が 1/2 と異なる）といえるか、有意水準 5%で判定せよ。

解答

帰無仮説 H_0：賛成と反対は確率 1/2 である。

対立仮説 H_1：H_0 でない。

$$\chi^2 = \frac{(35-25)^2}{25} + \frac{(15-25)^2}{25} = \frac{200}{25} = 8$$

$p = chidist(8,1) = 0.004678 < 0.05$ より、賛成は過半数であるといえる。

（正確には、賛成と反対は比率 1/2 でないといえる。）

問題

上の例題で、月曜日は特に事故が起こっているといえるか。有意水準 5%で判定せよ。

解答

曜日	月曜	その他
事故件数	23	57
理論比率	1/5	4/5
理論値	16	64

帰無仮説 H_0：事故は月曜に 1/5 の確率で起きている。

対立仮説 H_1：H_0 でない。

$$\chi^2 = \frac{(23-16)^2}{16} + \frac{(57-64)^2}{64} = 3.828$$

$p = chidist(3.828,1) = 0.0504 > 0.05$ より、月曜日に多いとはいえない。

しかし、結果がぎりぎりなので考察の余地は残る。

10.2 対応のない２標本の比率の検定

1. ２×２表の検定

例

男女差が購入意欲に影響を与えるかどうか調べるために、男女によって購入意欲のありなしを分けたところ以下の結果を得た。男女差はあるといえるか。有意水準 5%で判定せよ。

	購入意欲あり	購入意欲なし	計
男	18	10	28
女	12	14	26
計	30	24	54

理論　χ^2 検定

ある事象の出現、非出現を要因の有無により分けると以下のようになった。

出現、非出現の間に要因の有無による差があるか。有意水準$\alpha \times 100\%$ で判定する。

	出現	非出現	計
要因有り	a	b	$a+b$
要因無し	c	d	$c+d$
計	$a+c$	$b+d$	$a+b+c+d=n$

H_0：２群間に差がない。

H_1：２群間に差がある。

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_1^2 \text{ 分布}$$

$p = chidist(\chi^2, 1)$［$\chi_1^2(p) = \chi^2$］として、

$p < \alpha$ ならば、H_0 を棄却し、H_1 を採択する。

注）$p = chidist(\chi^2, d)$

解答

$$\chi^2 = \frac{54 \times (18 \times 14 - 10 \times 12)^2}{28 \times 26 \times 30 \times 24} = \frac{54 \times 132^2}{524160} = 1.795055$$

$p = chidist(1.795055, 1) = 0.180312$

$p > 0.05$ より、要因による差があるとはいえない。

問題

ある案についてのアンケートで以下の結果を得た。男女間の回答（賛成の比率）に差があるといえるか。有意水準 5%で判定せよ。

	賛成	反対
男	128	86
女	107	95

解答

$\chi^2 = 1.979603$

$p = chidist(1.979603,1) = 0.159432$

$p > 0.05$ より、男女間に差があるといえない。

2. m×n表の検定

例

ある地域の女性について、ある商品の所有の有無を職業別に分類すると、以下の結果が得られた。職業間で商品所有の割合に差が認められるか。有意水準5%で判定せよ。

	所有有り	所有無し	計
主婦	90	199	289
事務	32	47	79
販売・生産	53	71	124
計	175	317	492

理論　χ^2検定

ある事象（s 種）の出現状況を要因（r 種）により分けると以下のようになる。出現頻度に要因による差が認められるか。有意水準$\alpha\times100\%$で判定する。

	事象1	事象2	…	事象s	計
要因1	x_{11}	x_{12}	…	x_{1s}	$x_{1.}$
要因2	x_{21}	x_{22}	…	x_{2s}	$x_{2.}$
⋮	⋮	⋮		⋮	⋮
要因r	x_{r1}	x_{r2}	…	x_{rs}	$x_{r.}$
計	$x_{.1}$	$x_{.2}$	…	$x_{.s}$	n

H_0：出現頻度に要因による差はない（独立である）

H_1：出現頻度に要因による差がある（独立でない）

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(x_{ij}-x_{i.}x_{.j}/n\right)^2}{x_{i.}x_{.j}/n}\sim\chi^2_{(r-1)(s-1)}\text{分布}$$　2×2表の統計量の一般形

$p=chidist(\chi^2,(r-1)(s-1))$［$\chi^2_{(r-1)(s-1)}(p)=\chi^2$］とし、

$p<\alpha$ ならば、H_0 を棄却し、H_1 を採択する。

解答

$$\chi^2=\frac{\left(90-289\times175/492\right)^2}{289\times175/492}+\frac{\left(199-289\times317/492\right)^2}{289\times317/492}$$
$$+\frac{\left(32-79\times175/492\right)^2}{79\times175/492}+\frac{\left(47-79\times317/492\right)^2}{79\times317/492}$$
$$+\frac{\left(53-124\times175/492\right)^2}{124\times175/492}+\frac{\left(71-124\times317/492\right)^2}{124\times317/492}$$
$$=6.095771$$

$p=chidist(6.095771,2)=0.047459<0.05$ より、職業間に差があるといえる。

10.3 対応のある２標本の比率の検定

例

あるキャンペーン実施の前後で、各支店の印象について客からアンケートをとり、支店毎に好印象かどうかで分類したところ、以下の結果を得た。キャンペーンは効果があった（前後で差がある）と言えるか。有意水準 5%で判定せよ。

前＼後	好印象	悪印象
好印象	40	11
悪印象	24	10

理論（McNemar 検定）

データと対照データとマッチさせて、調査結果で分類したところ以下の表を得た。データと対照データに差があると考えられるか。有意水準$\alpha\times100\%$で判定する。

データ＼対照データ	結果１	結果２
結果１	a	b
結果２	c	d

２つのデータによる差がないとすると

帰無仮説 H_0：２つのデータに差がない

対立仮説 H_1：２つのデータに差がある

$$\chi^2=\frac{(b-c)^2}{b+c}\sim\chi_1^2\text{分布}$$

$p=chidist(\chi^2,1)$［$\chi_1^2(p)=\chi^2$］として、

$p<\alpha$ ならば、H_0 を棄却し、H_1 を採択する。

解答

$$\chi^2=\frac{(24-11)^2}{24+11}=4.828571$$

$$p=chidist(4.828571,1)=0.027992$$

$p<0.05$ より、キャンペーン前後で差があるといえる。

10.4 比率の検定のためのデータ数の決定

例

「はい」、「いいえ」で回答するアンケート調査で、「はい」が 60%と予想されるとき、有意水準 5%で過半数である（「はい」が 50%でない）と判定するために必要なデータ数はいくらか。

理論

２つの事象の出現理論比率がそれぞれ p , $1-p$ であるとき、有意水準 $\alpha\times100\%$ で予想比率 $\hat{p}$ が理論比率と異なると判定するために必要なデータ数を求める。

$$\chi^2=\frac{(n_1-np)^2}{np}+\frac{[n_2-n(1-p)]^2}{n(1-p)}=\frac{n^2(\hat{p}-p)^2}{np}+\frac{n^2[(1-\hat{p})-(1-p)]^2}{n(1-p)}$$

$$=\frac{n(\hat{p}-p)^2}{p(1-p)}\sim\chi_1^2\text{ 分布}$$

の関係を用いて、データ数は次のように与えられる。

$$n>\frac{chiinv(0.05,1)\cdot p(1-p)}{(\hat{p}-p)^2}$$

注） $\chi_k^2(\alpha)=chiinv(\alpha,k)$

解答

$$n>\frac{chiinv(0.05,1)\times0.5\times0.5}{(0.6-0.5)^2}=96.03638$$

97 以上必要である。

問題

以下の場合、理論比率 0.5 と比較して有意差を出すために必要なデータ数はいくらか？

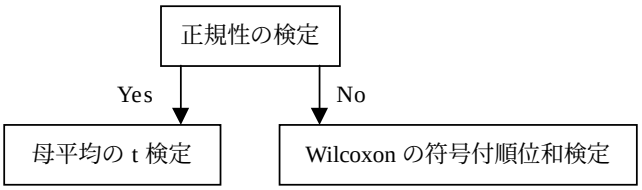
1) $\alpha=0.05$ で 0.7 で有意
2) $\alpha=0.05$ で 0.55 で有意
3) $\alpha=0.01$ で 0.6 で有意

解答

1) 25 以上　　2) 385 以上　　3) 166 以上

11 章　１標本の量的データの検定

11.1 検定手順



11.2 正規性の検定

視覚的方法

データ数が多い場合　　ヒストグラムによるグラフ化

データ数が少ない場合　　正規確率紙（MS-Excel でも可能）

数値的方法

Kolmogorov-Smirnov 検定

Shapiro-Wilk’s W-statistic

例

以下のデータの正規性を調べよ。

2.5, 2.1, 3.4, 2.8, 4.6, 3.2, 3.8, 4.8, 4.0

方法

MS-Excel を用いた視覚的方法

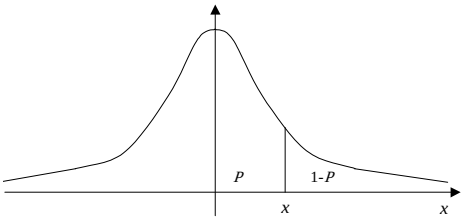
１．データを入力する。（データ数 n）

２．データを小さい順に並べ替える。（[データ－並べ替え]）

３．データに 1 から番号を振る。

４．累積比率を求める。　$p_i = \dfrac{i}{n+1}$　i は番号

５．$x = normsinv(p)$ 関数を用いて x 値を求める。



$$p = normsdist(x),\ \ x = normsinv(p)$$

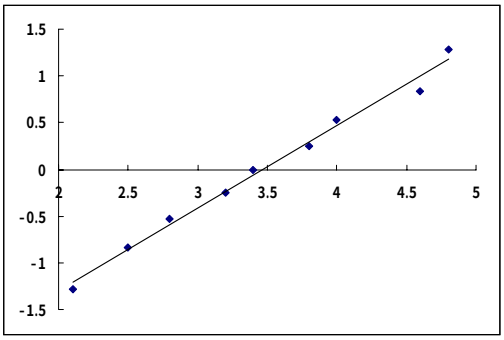
６．データと x 値を用いて散布図を描く。

７．グラフに近似曲線を加える。（[グラフ－近似曲線の追加]）

８．直線に近く並んでいるようなら正規分布

解答

番号	データ	累積比率	*x*値
1	2.1	0.1	-1.28155
2	2.5	0.2	-0.84162
3	2.8	0.3	-0.5244
4	3.2	0.4	-0.25335
5	3.4	0.5	0
6	3.8	0.6	0.253347
7	4	0.7	0.524401
8	4.6	0.8	0.841621
9	4.8	0.9	1.281551



この例題の場合、データが直線状に並んでいると認められるので、正規分布とみなせる。 (Shapiro-Wilk's W-statistic　p<0.9147)

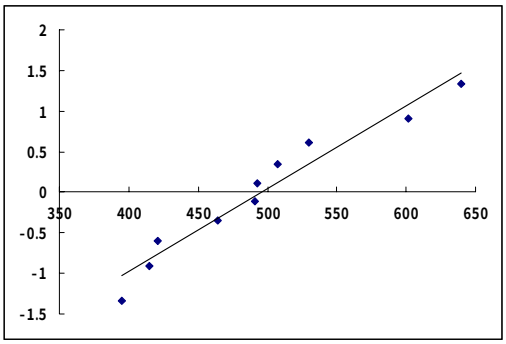
問題

以下のデータの正規性を調べよ。

507，491，421，493，415，640，464，602，530，395

解答

番号	データ	累積比率	*x*値
1	395	0.090909	-1.33518
2	415	0.181818	-0.90846
3	421	0.272727	-0.60458
4	464	0.363636	-0.34876
5	491	0.454545	-0.11418
6	493	0.545455	0.114185
7	507	0.636364	0.348756
8	530	0.727273	0.604584
9	602	0.818182	0.908458
10	640	0.909091	1.335179



この場合、ほぼ正規分布とみなせる。(Shapiro-Wilk's W-statistic　p<0.5515)

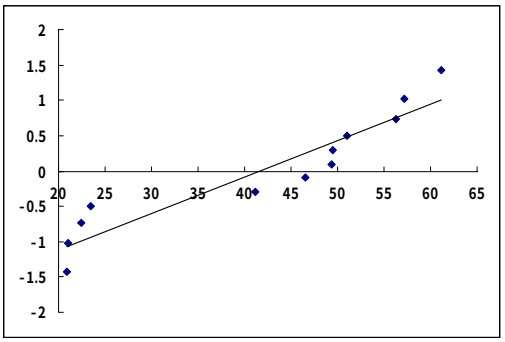
問題

以下のデータの正規性を調べよ。

20.9，61.1，57.2，51.0，46.6，41.2，21.0，56.3，49.5，49.3，22.4，23.5

解答

番号	データ	累積比率	x値
1	20.9	0.076923	-1.42608
2	21.0	0.153846	-1.02008
3	22.4	0.230769	-0.73632
4	23.5	0.307692	-0.5024
5	41.2	0.384615	-0.29338
6	46.6	0.461538	-0.09656
7	49.3	0.538462	0.096559
8	49.5	0.615385	0.293381
9	51	0.692308	0.502403
10	56.3	0.769231	0.736316
11	57.2	0.846154	1.020076
12	61.1	0.923077	1.426079



直線状に並んでいるといえないので、正規分布とはいえない。

(Shapiro-Wilk’s W-statistic　$p<0.0392$)

11.3 想定値と標本の検定（正規性あり）

例

ある会社 9 社についてある商品の１人当り売上高のデータを集めたら、正規分布し、平均 241（万円)、不偏分散から求めた標準偏差 14（万円）であった。この地域の会社の１人当り売上高は 226（万円）に比べて差があるといえるか？有意水準 5%で判定せよ。

理論　t 検定

正規分布するデータについて、標本の母平均 μ_1 と母集団の平均 μ とを比較し、差があるかどうか有意水準 $\alpha\times100\%$ で判定する。

データ数：n　　標本平均：$\bar{x}$　　不偏分散：u^2

帰無仮説 H_0：$\mu_1=\mu$　　差がない

対立仮説 H_1：$\mu_1\neq\mu$（両側検定）　　差がある

$$t=\frac{\sqrt{n}(\bar{x}-\mu)}{u}\sim t_{n-1}\text{ 分布}$$

$p=tdist(|t|,n-1,2)$ ［$t_{n-1}(p/2)=|t|$］として、

$p<\alpha$ のとき、H_0 を棄却し、H_1 を採択する。

注）$p=tdist(t,d,2)$　　両側 p

$p/2=tdist(t,d,1)$　　片側 $p/2$

$x=tinv(p,d)$　　両側 p

解答

$$t=\frac{\sqrt{9}(241-226)}{14}=3.214286$$

$$p=tdist(3.214286,8,2)=0.012345$$

$p<0.05$ より、１人当り売上高に差があるといえる。

問題

ある会社 11 社についてある商品の１人当りの売上高のデータを集めたら、以下のように与えられた（単位万円)。これらの会社の売上高は 226（万円）と比べて平均に差があるといえるか。正規分布を仮定し、有意水準 5%で判定せよ。

206, 235, 155, 172, 180, 199, 151, 172, 291, 182, 260

解答

$n = 11$, $\bar{x} = 200.2727$, $u = 44.56476$

$t = -1.91469$, $p = tdist(1.91469, 10, 2) = 0.084547$

$p > 0.05$ より、１人当りの売上高に差があるといえない。

問題

以下のデータの正規性が認められているとき、平均は母平均 5.5 と比べて差があるといえるか。有意水準 5%で判定せよ。

8.4, 4.6, 5.2, 6.3, 7.2, 5.8, 6.0, 5.4, 4.9, 6.9

解答

11.4 想定値と標本の検定（正規性なし）

例

ある会社のある商品の１人当り売上高（万円）は以下の通りである。これらの会社の１人当り売上高は、226（万円）に比べて差があるといえるか。有意水準 5%で判定せよ。

206, 235, 155, 172, 180, 199, 151, 172, 291, 182, 260

理論　Wilcoxon の符号付き順位和検定

データ	母集団の中央値	データ－母集団の中央値
$(x_1, x_2, \cdots, x_n)$	μ	$(z_1, z_2, \cdots, z_n)$

標本の中央値 m' と母集団の中央値 m を比較し、差があるかどうか $\alpha \times 100\%$ の有意水準で判定する。

帰無仮説 H_0：$m' = m$

対立仮説 H_1：$m' \neq m$（両側検定）

$|z_i|$ の小さい順に 0 を除いて順位 r_i を付け、z_i の正負で 2 群に分ける。（同数値の場合は、順位平均をとる (5+6)/2=5.5）

$z_i < 0$	$z_i > 0$
$(r_{i_1}, r_{i_2}, \cdots, r_{i_r})$	$(r_{j_1}, r_{j_2}, \cdots, r_{j_s})$

それぞれの群の順位和をとり、このうち小さい方を選ぶ。

R_r	R_s	$R = \min(R_r, R_s)$

$n = r + s$ ： $z_i = 0$ を除いたデータ数とする。

データ数が少ないとき

数表（ $p = \alpha / 2$ ）を参照し、$R \le R_n(\alpha / 2)$ のとき、H_0 を棄却して H_1 を採択する。

データ数が多いとき

$$z = \frac{|R - n(n+1)/4| - 1/2}{\sqrt{n(n+1)(2n+1)/24}} \quad \sim \quad N(0,1) \text{ 分布（正の部分）}$$

$p = 2 \cdot (1 - normsdist(z))$ ［$Z(p/2) = z$］として、

$p < \alpha$ のとき、H_0 を棄却して H_1 を採択する。

解答

データ	差	|差|	順位	訂正順位
206	-20	20	2	2
235	9	9	1	1
155	-71	71	10	10
172	-54	54	7	7.5
180	-46	46	6	6

199	-27	27	3	3
151	-75	75	11	11
172	-54	54	7	7.5
291	65	65	9	9
182	-44	44	5	5
260	34	34	4	4

$R = \min(14, 52) = 14$

$R = 14 > R_{11}(0.025) = 10$　より、中央値に差があるとはいえない。

11.5 平均の検定のためのデータ数の決定

例

母集団の身長平均が 170cm、標準偏差が 5cm であるとき、標本平均 169cm で母平均と異なることを有意水準 5%で判定するためには、いくらのデータ数が必要か。

理論

母平均が μ、母分散 σ^2 の場合、有意水準 $\alpha\times100\%$ で、標本平均 $\bar{x}$ が母平均と等しくないことを判定するために必要なデータ数はいくらか。

$Z=\dfrac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\sim N(0,1)$　を用いて、

$$n>\frac{(normsinv(1-\alpha/2))^2\cdot\sigma^2}{(\bar{x}-\mu)^2}$$

$Z(\alpha/2)=normsinv(1-\alpha/2)$ ：標準正規分布上側確率 $\alpha/2$ の x 値

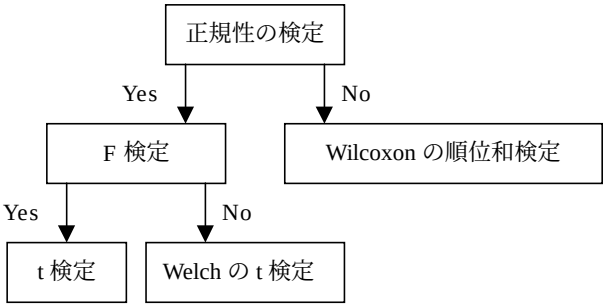
両側検定

解答

$n>\dfrac{normsinv(0.975)^2\times5^2}{1^2}=96.03619$　より、標本は 97 以上必要である。

12 章 対応のない２標本の量的データの検定

12.1 検定手順



12.2 対応のない２標本の分散の検定（正規性あり）

例

A 機を導入した会社 18 社（1 群）と B 機を導入した会社 9 社（2 群）について、機械 10 台当り 1 年間の故障発生件数を調べ、不偏分散を求めたら以下の結果を得た。分布は正規分布であると仮定して、分散に差があるといえるか有意水準 5%で判定せよ。

1 群　　10.68

2 群　　2.17

理論　F 検定

正規分布する母集団から抽出した標本１と標本２について、それぞれの母分散 σ_1^2, σ_2^2 を比較し、差があるかどうか有意水準$\alpha \times 100\%$で検定を行う。

帰無仮説 H_0：$\sigma_1^2 = \sigma_2^2$

対立仮説 H_1：$\sigma_1^2 > \sigma_2^2$（但し、$u_1^2 > u_2^2$ とする）

データ数 n_1, n_2，不偏分散 u_1^2, u_2^2 として、

H_0 のもとで

$$F = \frac{u_1^2}{u_2^2} \sim F_{n_1-1,\, n_2-1} \text{分布}$$

$p = fdist(F, n_1 - 1, n_2 - 1)$［$F_{n_1-1,\, n_2-1}(p) = F$］として、

$p < \alpha$ ならば、分散に差があるといえる。　注）$p = fdist(x, d_1, d_2)$　片側検定

解答

$$F = \frac{10.68}{2.17} = 4.922$$

$p = fdist(4.922, 17, 8) = 0.0138 < 0.05$　より、分散に差があるといえる。

問題

以下の標本データの分散には差があるといえるか。有意水準 5%で判定せよ。

標本１　152, 154, 142, 149, 148, 135, 143, 146, 136, 150, 150, 150, 138

標本２　147, 145, 138, 145, 149, 153, 152, 132, 169, 158, 133, 147, 149, 165, 159, 159

解答

	データ数	不偏分散
標本１（２群）	13	39.08974
標本２（１群）	16	110.1333

$$F = \frac{110.1333}{39.08974} = 2.817448$$

$$p = fdist(2.817448, 15, 12) = 0.038714$$

$p < 0.05$ より、分散に差があるといえる。

問題

以下の標本データの分散には差があるといえるか。有意水準 5%で判定せよ。

標本１　112, 106, 101, 112, 102, 98, 108, 95, 101, 90, 110, 97, 95, 105, 101, 113, 114, 91

標本２　98, 88, 105, 99, 96, 93, 109, 106, 103, 87, 107, 102, 97, 91

解答

	データ数	不偏分散
標本１（１群）	18	57.91176
標本２（２群）	14	50.09341

$$F = \frac{57.912}{50.093} = 1.156$$

$$p = fdist(1.156, 17, 13) = 0.4015$$

$p > 0.05$ より、分散に差があるといえない。

12.3 対応のない２標本の検定（正規性あり・等分散性あり）

例

ある地域の同性・同年齢の児童について、ある要因の有無による２つの集団の体重を調べたところ以下のデータを得た。２つの集団の平均値に差はあるといえるか。正規性、等分散性を仮定して、有意水準 5%で判定せよ。

	データ数	平均	不偏分散
要因なし	20	40.2	25.5
要因あり	20	36.4	16.0

理論（student の）t 検定

正規分布する等分散の標本１と標本２について、母平均 μ_1, μ_2 を比較し、差があるかどうか有意水準$\alpha \times 100\%$ で検定を行う。

帰無仮説 H_0： $\mu_1 = \mu_2$

対立仮説 H_1： $\mu_1 \neq \mu_2$

データ数 n_1, n_2， 標本平均 $\bar{x}_1, \bar{x}_2$， 不偏分散 u_1^2, u_2^2 とすると、

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{(n_1 - 1)u_1^2 + (n_2 - 1)u_2^2}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2} \text{ 分布}$$

$p = tdist(|t|, n_1 + n_2 - 2, 2)$ ［$t_{n_1+n_2-2}(p/2) = |t|$］ として、

$p < \alpha$ ならば平均に差があると判定する。 両側検定

注） $p/2 = tdist(t, d, 1)$ 検定値→片側確率

$p = tdist(t, d, 2)$ 検定値→両側確率

$t = tinv(p, d)$ 両側確率→検定値

解答

$$t = \sqrt{\frac{20 \cdot 20}{40}} \frac{40.2 - 36.4}{\sqrt{\dfrac{19 \cdot 25.5 + 19 \cdot 16.0}{38}}} = 2.637999$$

$p = tdist(2.637999, 38, 2) = 0.01202$

$p < 0.05$ より、平均に差があるといえる。

問題

以下の母平均には差があるといえるか。正規性と等分散性を認めて、有意水準 5%で判定せよ。

1群　47, 45, 38, 45, 49, 53, 52, 32, 69, 58, 33, 47, 49, 65, 59

2群　47, 58, 64, 53, 64, 64, 59, 46, 42, 43, 52, 61, 57

解答

	データ数	平均	不偏分散
群1	15	49.4000	111.8286
群2	13	54.6154	64.7564

$t = -1.44996$

$p = tdist(1.44996, 26, 2) = 0.159027$

$p > 0.05$ より、平均値に差があるといえない。

問題

前節のデータの母平均には差があるといえるか。等分散性を認めて、有意水準 5%で判定せよ。

1群　112, 106, 101, 112, 102, 98, 108, 95, 101, 90, 110, 97, 95, 105, 101, 113, 114, 91

2群　98, 88, 105, 99, 96, 93, 109, 106, 103, 87, 107, 102, 97, 91

解答

	データ数	平均	不偏分散
群1	18	102.8333	57.91176
群2	14	98.64286	50.09341

$t = 1.593$

$p = tdist(1.593, 30, 2) = 0.12174$

$p > 0.05$ より、平均に差があるといえない。

12.4 対応のない２標本の検定（正規性あり・等分散性なし）

例

A 機を導入した会社 18 社（１群）と B 機を導入した会社 9 社（２群）について機械 10 台当り 1 年間の故障発生件数を調べ、平均と不偏分散を求めたところ以下の結果を得た。正規性があり、異分散であるとして、2 群間の平均に差があるかどうか有意水準 5%で判定せよ。

	平均	不偏分散
１群	10.56	10.68
２群	8.22	2.17

理論　ウェルチ(Welch)の t 検定

正規分布する分散の異なる標本１と標本２について、母平均 μ_1, μ_2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で検定を行う。

例数 n_1, n_2，標本平均 $\bar{x}_1, \bar{x}_2$，不偏分散 u_1^2, u_2^2 とすると、

帰無仮説 H_0：$\mu_1 = \mu_2$

対立仮説 H_1：$\mu_1 \neq \mu_2$　両側検定

H_0 のもとで

$$c = \frac{u_1^2/n_1}{u_1^2/n_1 + u_2^2/n_2}, \qquad d = \frac{1}{\dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}}$$ として、

自由度を $d' = \mathrm{int}(d)$ とする。

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{u_1^2/n_1 + u_2^2/n_2}} \sim t_{d'} \text{ 分布}$$

$p = tdist(|t|, d', 2)$［$t_{d'}(p/2) = |t|$］として、$p < \alpha$ ならば平均に差があるといえる。

解答

$$c = \frac{10.682/18}{10.682/18 + 2.172/9} = 0.711052, \quad d = 24.88971 \quad \rightarrow \quad d' = 24$$

一般に、自由度の小さい方が差は出にくい。

差があることを厳しく評価するなら、小数点以下を切り捨てる。

$$t = \frac{10.56 - 8.22}{\sqrt{10.68/18 + 2.17/9}} = 2.561634$$

$$p = tdist(2.561634, 24, 2) = 0.017123$$

$p < 0.05$ より、平均に差があるといえる。

問題

以下の標本データの平均には差があるとみなせるか。正規性・異分散性を認めて、有意水準 5%で判定せよ。

標本 1　358, 469, 397, 350, 329, 446, 393, 379, 443, 348,
455, 332, 311, 424, 420, 354, 353, 390, 434, 430

標本 2　335, 387, 385, 343, 394, 351, 404, 391, 330, 363,
319, 334, 348, 396, 408, 403, 415, 353, 377, 399

解答

	標本 1	標本 2
データ数	20	20
平均	390.75	371.75
不偏分散	2239.461	914.9342

$c = 0.709949 \qquad d = 32.30427 \quad \rightarrow \quad d' = 32$

$t = 1.512901$

$p = tdist(1.512901, 32, 2) = 0.140119$

$p > 0.05$ より、平均に差があるといえない。

問題

以下の標本データの平均には差があるとみなせるか。正規性・異分散性を認めて、有意水準 5%で判定せよ。

標本 1　51, 55, 46, 58, 57, 54, 48, 46, 57, 42, 44, 44, 52, 41, 52

標本 2　52, 65, 73, 59, 72, 72, 67, 50, 45, 46, 57, 68, 63

解答

	標本 1	標本 2
データ数	15	13
平均	49.8	60.6923
不偏分散	33.1714	99.3974

$c = 0.2243 \qquad d = 18.6109 \quad \rightarrow \quad d' = 18$

$t = -3.46928$

$p = tdist(3.46928, 18, 2) = 0.00274$

$p < 0.05$ より、平均に差があるといえる。

12.5 対応のない２標本の検定（正規性なし）

例

あるソフトウェアの販売において、支店の売上伸び率を２つの販売戦略グループで比較したところ、以下の結果が得られた。２群の増加は１群のそれに比べて大きいといえるか。有意水準 5%の両側検定で判定せよ。

１群：6, 5, 10

２群：12, 16, 22, 8, 17

理論　ウィルコクソン(Wilcoxon)の順位和検定

正規分布するとは限らない標本１（例数n_1）と標本２（例数n_2）について、母集団の中央値（m_1, m_2）を比較し、差があるかどうか有意水準$\alpha\times100\%$で検定を行う。

両群のデータの小さい順に順位を付ける。ただし、同じ値にはそれらが異なると考えた場合の順位の平均値を付ける。

母集団の中央値	m_1	m_2
標本	$(x_1, x_2, \cdots, x_{n_1})$	$(y_1, y_2, \cdots, y_{n_2})$
順位	$(r_1, r_2, \cdots, r_{n_1})$	$(s_1, s_2, \cdots, s_{n_2})$

ここに　$n_1 \le n_2$ とする。

帰無仮説 H_0：$m_1 = m_2$

対立仮説 H_1：$m_1 \neq m_2$（通常は両側検定）

$n_2 \le 20$　の場合

$W = \sum_{i=1}^{n_1} r_i$　→　確率 $p = \alpha/2$（両側検定）として数表を参照

データ数(n_1, n_2)の組で（$U_{1-p}; U_p$）の値を求め、

$W \le U_{1-p}$ または $W \ge U_p$ であれば、両群の中央値に差があると判定する。

$n_2 > 20$　の場合

$$z = \frac{|W - n_1(n_1+n_2+1)/2| - 1/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \sim N(0,1)$$ 分布（正の部分）

$p = 2\cdot(1 - normsdist(z))$　［$Z(p/2) = z$］として、

$p < \alpha$ であれば、両群の中央値に差があると判定する。

解答

１群	２群	１群順位	２群順位
6	12	2	5
5	16	1	6

10	22 8 17	4	8 3 7
順位和		7	29

データ数　　3, 5

データ数の少ない１群の順位和を求める。　　$W = 7$

数表　　$n_1 = 3, n_2 = 5$ の場合、両側検定 5%で、 6; 21

6<7<21 であるので、群１と群２の中央値は異なるといえない。

問題

ラットの体重増加(g)を、条件を変えた２つのグループで測定したところ、以下の結果が得られた。２群の体重増加に差は認められるか、有意水準 5%で判定せよ。

１群：7.2, 8.3, 5.4, 6.0, 7.3, 11.7, 10.5, 8.0, 9.1

２群：10.1, 13.2, 7.4, 9.1, 16.2, 14.5, 6.3, 11.2, 12.4, 7.4, 12.5, 9.1, 17.0

解答

群	データ	群	データ	順位	訂正順位
1	7.2	1	5.4	1	1
1	8.3	1	6	2	2
1	5.4	2	6.3	3	3
1	6	1	7.2	4	4
1	7.3	1	7.3	5	5
1	11.7	2	7.4	6	6.5
1	10.5	2	7.4	6	6.5
1	8	1	8	8	8
1	9.1	1	8.3	9	9
2	10.1	1	9.1	10	11
2	13.2	2	9.1	10	11
2	7.4	2	9.1	10	11
2	9.1	2	10.1	13	13
2	16.2	1	10.5	14	14
2	14.5	2	11.2	15	15
2	6.3	1	11.7	16	16
2	11.2	2	12.4	17	17
2	12.4	2	12.5	18	18
2	7.4	2	13.2	19	19
2	12.5	2	14.5	20	20
2	9.1	2	16.2	21	21
2	17	2	17	22	22

１）群に番号を付け、群別にデータを入力する。

２）データの大きさ順に並べ替える。［データ－並べ替え］

３）データに順位を付ける。　　注）rank(数値,範囲,順序) 関数を利用する。

順序：0 または省略で降順, 0 以外で昇順

４）同順位のものに訂正を加える。

例　6, 6 → 6.5, 6.5　　10, 10, 10 → 11, 11, 11　[(10+11+12)/3]

５）群別に順位合計をとる。　　注）sumif(範囲,検索条件,合計範囲) 関数を利用する。

例　sumif(C2:C23, “=1”, F2:F23)

または、群別に並べ直し、各群の順位合計をとる。

	データ数	順位合計
1 群	9	70
2 群	13	183

６）Wilcoxon の順位和検定数値表により、検定する。

$n_A = 9, n_B = 13$　表より $\alpha = 0.025$ のとき　→　73;134 両側検定

データ数の少ない 1 群の順位合計は 70 であるから、上記の範囲に入らない。

よって、有意水準 5%で差があるといえる。

問題

正規分布しない 2 群のデータで順位和を求めたところ、以下の結果を得た。それらの中央値に差があるかどうか、有意水準 5%で判定せよ。

	データ数	順位合計
1 群	30	1265
2 群	40	1220

解答

$$Z = \frac{|W - n_1(n_1 + n_2 + 1)/2| - 1/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} = \frac{1265 - 30 \times 71/2 - 1/2}{\sqrt{30 \times 40 \times 71/12}} = 2.367629$$

$p = 2 \cdot (1 - normsdist(2.367629)) = 0.017902 < 0.05$　より、

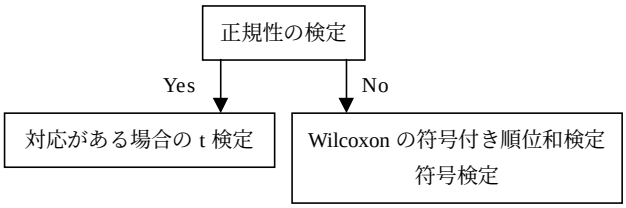
中央値に差があるといえる。

注）標準正規分布　$normsdist(x)$　検定値→累積確率値

$normsinv(p)$　累積確率値→検定値

13 章 対応のある２標本の量的データの検定

13.1 検定手順



13.2 対応のある２標本の検定（正規性あり）

例

ある商品の陳列位置を変える前と後とで売上高（千円）を規模の等しい８つの支店で比較したところ、以下の結果を得た。標本間の差が正規分布するとして有意水準 5%で差があるかどうか判定せよ。

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396

理論

正規分布する対応のある（正確には差が正規分布する）標本１と標本２の母平均 μ_1, μ_2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で判定する。

帰無仮説 H_0： $\mu_1 = \mu_2$

対立仮説 H_1： $\mu_1 \neq \mu_2$

各データの差（z_i =標本１－標本２）について、データ数 n，平均 $\bar{z}$，不偏分散 u_z^2

$$t = \frac{\sqrt{n}\,\bar{z}}{u_z} \sim t_{n-1} \text{ 分布}$$

$p = tdist(|t|, n-1, 2)$ ［$t_{n-1}(p/2) = |t|$］ として、

$p < \alpha$ のとき、H_0 を棄却し、H_1 を採択する。

解答

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396
差	-11	29	-111	-74	-10	12	-90	-54

注）ここでは横方向だが、Excel でデータは縦方向に入力する。

$n = 8$， $\bar{z} = -38.625$， $u_z = 50.82726$

$t = 2.149398$

$p = tdist(2.149398, 7, 2) = 0.068675$

$p > 0.05$ より、差があるとはいえない。

13.3 対応のある２標本の検定（正規性なし）

例

ある商品の陳列位置を変える前と後とで売上高（千円）を規模の等しい８つの支店で比較したところ、以下の結果を得た。各標本が正規分布しないものとして有意水準 5% で売上高に差があるかどうか判定せよ。

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396

理論　Wilcoxon の符号付き順位和検定

任意の分布に従う対応のある標本１と標本２の母集団の中央値 m_1, m_2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で判定する。

帰無仮説 H_0：$m_1 = m_2$

対立仮説 H_1：$m_1 \neq m_2$

対応する各標本の差（z_i =標本１－標本２）について、$|z_i|$ の小さい順に 0 を除いて順位 r_i を付け、z_i の正負で 2 群に分ける。（同数値の場合は、順位平均をとる (5+6)/2=5.5）

$$z_i < 0 \qquad z_i > 0$$

$$(r_{i_1}, r_{i_2}, \cdots, r_{i_r}) \qquad (r_{j_1}, r_{j_2}, \cdots, r_{j_s})$$

それぞれの群の順位和をとり、このうち小さい方を選ぶ。

$$R_r \qquad R_s \qquad R = \min(R_r, R_s)$$

$n = r + s$ ： $z_i = 0$ を除いたデータ数とする。

データ数が少ないとき

数表（$p = \alpha / 2$）を参照し、$R \le R_n(\alpha / 2)$ のとき、H_0 を棄却して H_1 を採択する。

データ数が多いとき

$$z = \frac{|R - n(n+1)/4| - 1/2}{\sqrt{n(n+1)(2n+1)/24}} \quad \sim \quad N(0,1) \text{ 分布（正の部分）}$$

$p = 2 \cdot (1 - normsdist(z))$ ［$Z(p/2) = z$］として、

$p < \alpha$ のとき、H_0 を棄却して H_1 を採択する。

解答

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396
差	-11	29	-111	-74	-10	12	-90	-54
|差|	11	29	111	74	10	12	90	54
訂正順位	2	4	8	6	1	3	7	5

注）ここでは横方向だが、Excel でデータは縦方向に入力する。

$R = \min(7, 29) = 7$

数表より、$R = 7 > R_8(0.025) = 3$　より、差があるとはいえない。

14 章 相関係数の検定

14.1 Pearson の相関係数

例

２つの商品 A, B の地域別使用率 (%) のデータは以下の通りである。それぞれの商品の使用率に相関が認められるか。正規性を仮定して、有意水準 5%で判定せよ。

A(%)	33	24	30	50	42	15	15	56	13	45	44	21	18	31	27	40
B(%)	20	34	50	20	58	23	12	34	26	56	42	5	25	51	19	27

理論

２変数が２変量正規分布に従うとき、母相関係数 ρ が 0 かどうか（0 と差があるか）有意水準 $\alpha \times 100\%$ で判定する。

帰無仮説 H_0： $\rho = 0$

対立仮説 H_1： $\rho \neq 0$　（両側検定）

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \text{ 分布}$$

$p = tdist(|t|, n-2, 2)$ ［$t_{n-2}(p/2) = |t|$］として、

$p < \alpha$ ならば、H_0 を棄却し H_1 を採択する。

注意　MS-Excel　相関係数：correl(範囲 1, 範囲 2)

解答

$r = 0.453786$, $n = 16$

$t = 1.905387$

$p = tdist(1.905387, 14, 2) = 0.077476$

$p > 0.05$ より、相関がある（相関係数が 0 と異なる）といえない。

問題

以下の 2 変数間の相関係数を求め、正規性を仮定して、相関係数が 0 と異なるかどうか有意水準 5%で判定せよ。

変数1	65	86	78	83	85	89	83	80	85	93	75	85	79	80
変数2	162	210	224	179	217	230	223	204	224	197	186	189	172	185

解答

$r = 0.557714$　　$t = 2.327588$

$p = tdist(2.327588, 12, 2) = 0.038237$

$p < 0.05$ より、相関があるといえる。

問題

以下の 2 変数間の相関係数を求め、正規性を仮定して、相関係数が 0 と異なるかどうか有意水準 5%で判定せよ。

変数 1	35	26	43	36	36	58	26	27	46	28	38	47	15	20	23
変数 2	41	50	65	28	40	67	33	23	56	45	43	49	20	18	41

解答

$r = 0.786115,\ \ n = 15$

$t = 4.585775$

$p = tdist(4.585775, 13, 2) = 0.000511$

$p < 0.05$ より、相関がある（相関係数が 0 と異なる）といえる。

14.2 Spearman の順位相関係数

例

前節の問題で、それぞれの商品の使用率に相関が認められるか。正規性を仮定せずに、有意水準 5%で判定せよ。

理論

一般の分布に従う 2 変数について、順位相関係数 r_s を求め、母相関係数 ρ が 0 かどうか（0 と差があるか）を判定する。　注）=correl(範囲 1, 範囲 2)

帰無仮説 H_0： $\rho = 0$

対立仮説 H_1： $\rho \neq 0$　（両側検定）

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{n-2} \text{ 分布}$$

$p = tdist(|t|, n-2, 2)$ ［$t_{n-2}(p/2) = |t|$］として、

$p < \alpha$ ならば、H_0 を棄却し H_1 を採択する。

解答

各変量ごとに小さい順に順位を付ける。ただし、同順位の場合は異なる順位とした場合の平均とする。

A (%)	B (%)	順位A	順位B	訂正A	訂正B
33	20	10	4	10	4.5
24	34	6	10	6	10.5
30	50	8	13	8	13
50	20	15	4	15	4.5
42	58	12	16	12	16
15	23	2	6	2.5	6
15	12	2	2	2.5	2
56	34	16	10	16	10.5
13	26	1	8	1	8
45	56	14	15	14	15
44	42	13	12	13	12
21	5	5	1	5	1
18	25	4	7	4	7
31	51	9	14	9	14
27	19	7	3	7	3
40	27	11	9	11	9

それらの順位について、相関係数を求める。

$r_s = 0.461312$　　$t = 1.945443$

$p = tdist(1.945443, 14, 2) = 0.072084 > 0.05$　より、相関があるとはいえない。

15 章 区間推定

15.1 母比率の区間推定

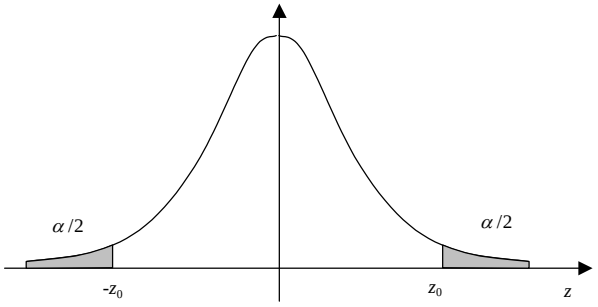
例

ある制度についてのアンケート調査をランダムに抽出された 100 人に対して行ったところ、賛成 65 人、反対 35 人であった。母集団の賛成の比率を、信頼区間 95%（有意水準 5% に相当）で推定せよ。また、調査数 1000 人ではどうか。

理論

データ数 n、標本比率 $\hat{p}$ の標本から、母比率 p を信頼区間 $(1-\alpha)\times 100\%$ で推定する。

$z = \dfrac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$ 分布　を利用し、$z \cong \dfrac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$　の近似を考える。



$$-z_0 \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le z_0 \quad \text{より、} \qquad z_0 = normsinv(1-\alpha/2)$$

$$\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0 \le p \le \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0$$

解答

$$\hat{p} = 0.65, \quad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.047697, \quad z_0 = norminv(0.975) = 1.959961$$

$$\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0 = 0.556516, \quad \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0 = 0.743484$$

$$0.556516 \le p \le 0.743484$$

1000 人では、以下のように精度が上がる。

$$0.620438 \le p \le 0.679562$$

問題

ある 500 人に対する調査で支持 205 人、不支持 295 人という結果を得た。母集団における支持の比率を信頼区間 95%で推定せよ。信頼区間 99%ではどうか。

解答

$$\hat{p} = 0.41 \qquad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0 = 0.04311 \qquad z_0 = 1.959961$$

$$0.36689 \le p \le 0.45311$$

問題

ある選挙において有効投票数の 3 割で当選することが分っており、信頼区間 99%の範囲が 3 割を超えると当選確実が打てるものとする。今ある候補が 3156 票の開票で 1083 票の得票を得た。この候補には当選確実が打てるか。

解答

$$\hat{p} = 0.343156 \qquad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.008451 \qquad z_0 = 2.575835$$

$$0.321388 \le p \le 0.364924$$

これより、当選確実は打てる。

15.2 正規母集団の母平均の区間推定

例

ある標本データから所得について集計したところ以下のデータを得た。母集団は正規分布するとして母平均を信頼区間 95%（有意水準 5% に相当）で推定せよ。

データ数 30，平均 620，標準偏差 90

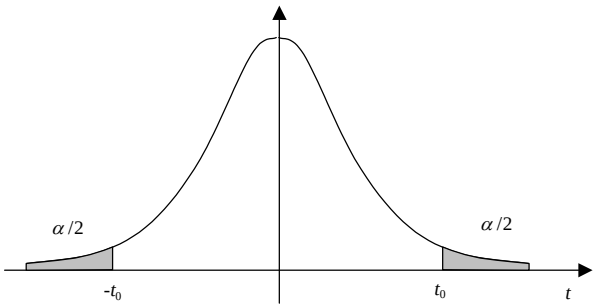
また、データ数を 100 にすると結果はどう変わるか？

理論

正規分布する母集団から得られた標本より、母平均 μ を信頼区間 $(1-\alpha)\times 100\%$ で推定する。

データ数 n，標本平均 $\bar{x}$，不偏分散 u^2

$t=\dfrac{\sqrt{n}(\bar{x}-\mu)}{u}\sim t_{n-1}$ 分布　　の性質を用いる。



$-t_0\le\dfrac{\sqrt{n}(\bar{x}-\mu)}{u}\le t_0$　より、

$\bar{x}-\dfrac{u}{\sqrt{n}}t_0\le\mu\le\bar{x}+\dfrac{u}{\sqrt{n}}t_0$

注）MS-Excel　　$t_0=tinv(\alpha,d)$　を用いる。

解答

$n=30$，　$\bar{x}=620$，　$u=90$

$t_0=tinv(0.05,29)=2.045231$

$\dfrac{u}{\sqrt{n}}t_0=33.60657$，　$\bar{x}+\dfrac{u}{\sqrt{n}}t_0=653.6066$，　$\bar{x}-\dfrac{u}{\sqrt{n}}t_0=586.3934$

$586.3934\le\mu\le 653.6066$

データ数を 100 にすると、以下のように精度が高まる。

$602.142\le\mu\le 637.858$

問題

正規分布を仮定して、以下の身長データ (cm) から母平均を信頼区間 95%で推定せよ。また、信頼区間 99%ではどうか。

184, 170, 164, 176, 177, 170, 171, 159, 174, 170,
165, 170, 171, 183, 175, 169, 181, 172, 171, 164

解答

95%信頼区間

$n = 20$, $\bar{x} = 171.8$, $u = 6.379243$, $t_0 = 2.0930$, $\frac{u}{\sqrt{n}} t_0 = 2.985578$ より、

$$168.8144 \le \mu \le 174.7856$$

99%信頼区間

$t_0 = 2.860943$, $\frac{u}{\sqrt{n}} t_0 = 4.080969$ より、

$$167.719 \le \mu \le 175.881$$

問題

正規分布を仮定して、以下のデータから母平均を信頼区間 95%で推定せよ。また、信頼区間 99%ではどうか。

52, 63, 41, 70, 67, 61, 46, 42, 67, 32, 37, 37, 56, 29, 57, 52, 45, 44, 64, 51, 61, 58

解答

95%信頼区間

$n = 22$, $\bar{x} = 51.45455$, $u = 11.99495$, $t_0 = 2.079614$, $\frac{u}{\sqrt{n}} t_0 = 5.318263$ より、

$$46.13628 \le \mu \le 56.77281$$

99%信頼区間

$t_0 = 2.831366$, $\frac{u}{\sqrt{n}} t_0 = 7.240742$ より、

$$44.2138 \le \mu \le 58.69529$$

15.3 正規母集団の母分散の区間推定

例

ある標本データから所得について集計したところ以下のデータを得た。母集団は正規分布するとして母分散を信頼区間 95%で推定せよ。

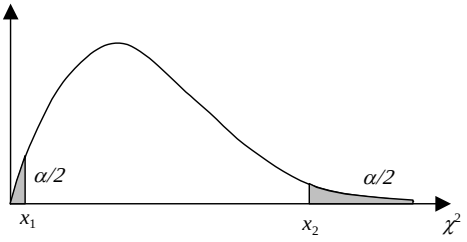
データ数 30，平均 620，不偏分散 8100

理論

正規分布する母集団から得られた標本より、母分散σ^2を信頼区間$(1-\alpha)\times 100\%$で推定する。

データ数n，不偏分散u^2

$\chi^2 = \dfrac{(n-1)u^2}{\sigma^2} \sim \chi^2_{n-1}$ 分布　　の性質を用いる。



$x_1 \le \dfrac{(n-1)u^2}{\sigma^2} \le x_2$　より、

$\dfrac{(n-1)u^2}{x_2} \le \sigma^2 \le \dfrac{(n-1)u^2}{x_1}$

注）MS-Excel　　$chiinv(p,d) = \chi^2_d(p)$　を用いる。

解答

$n = 30$，　$u^2 = 8100$，

$x_1 = chiinv(0.05, 29) = 17.70838$

$x_2 = chiinv(0.05, 29) = 42.55695$

$\dfrac{(n-1)u^2}{x_1} = 13264.91$，　$\dfrac{(n-1)u^2}{x_2} = 5519.663$　より、

$5519.663 \le \sigma^2 \le 13264.91$

問題

身長(cm)についての以下の標本データを用いて、母集団の母分散を有意水準 5%で推定せよ。

184, 170, 164, 176, 177, 170, 171, 159, 174, 170,
165, 170, 171, 183, 175, 169, 181, 172, 171, 164　　（例数 20）

解答

$n = 20$, $u^2 = 40.69474$,

$x_1 = chiinv(0.05,19) = 10.11701$

$x_2 = chiinv(0.05,19) = 30.14351$

$\frac{(n-1)u^2}{x_1} = 76.42577$, $\frac{(n-1)u^2}{x_2} = 25.65063$　より、

$$25.65063 \le \sigma^2 \le 76.42577$$

16 章 回帰分析

例

下の表のデータを用いて、身長により体重を推定する式を考える。ただし、式は 1 次式（体重 $= a \times$ 身長 $+ b$）と仮定し、その有効性を検討せよ。

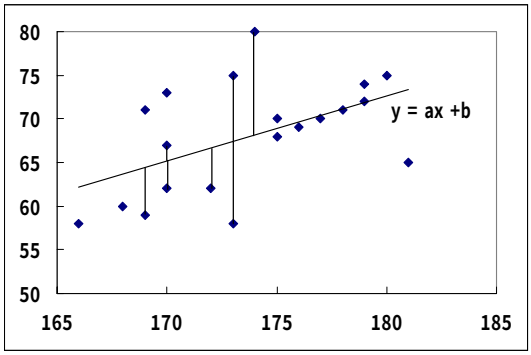
体重	71	68	67	72	69	80	75	65	74	71
身長	169	175	170	179	176	174	173	181	179	178
体重	62	75	70	70	62	58	60	58	59	73
身長	170	180	177	175	172	166	168	173	169	170

理論

回帰式の決定

2 変数の関係を、$y = ax + b$ の直線で表わすとすると、x を説明変数、y を目的変数と呼ぶ。

データ点からこの直線へ垂直におろした線の長さの 2 乗が最小となるように係数 a, b を決める。



平均 $\bar{x}, \bar{y}$，標準偏差 u_x, u_y，

相関係数 r とすると

$$a = r\frac{u_y}{u_x}, \quad b = \bar{y} - r\frac{u_y}{u_x}\bar{x}$$

回帰式の有効性の検討

重相関係数 R　目的変数の実測値と回帰式による予測値の相関係数

この場合　$R = r$

寄与率（重決定係数）R^2　目的変数の変動のうち回帰式が説明する割合

回帰式の有効性の検定　回帰式は無意味と考えられる確率で検討する。

（これは Excel 分析ツールを利用する。）

解答

$\bar{x} = 173.7$，$\bar{y} = 67.95$

$u_x = 4.402153$，$u_y = 6.378211$，$r = 0.513047$

$a = 0.743346$，$b = -61.1692$

回帰式　$y = 0.743346x - 61.1692$

重相関係数　$R = 0.513047$

寄与率　$R^2 = 0.263217$

Excel の分析ツールを用いた解答例

回帰統計	
重相関 R	0.513047
重決定 R2	0.263217
補正 R2	0.222285
標準誤差	5.624827
観測数	20

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	203.4538	203.4538	6.430541	0.020703
残差	18	569.4962	31.63868		
合計	19	772.95			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-61.1692	50.93303	-1.20097	0.245327	-168.176	45.83721
X 値 1	0.743346	0.293135	2.535851	0.020703	0.127492	1.3592

注）標準誤差：線形回帰式における予測値と実測値とのずれの標準偏差
ここで、分散はデータ数でなく自由度で割ったものとしている。

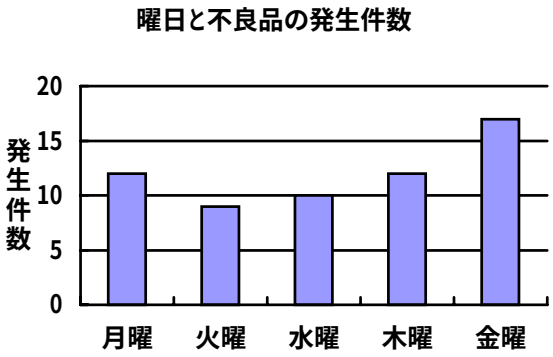
演習問題

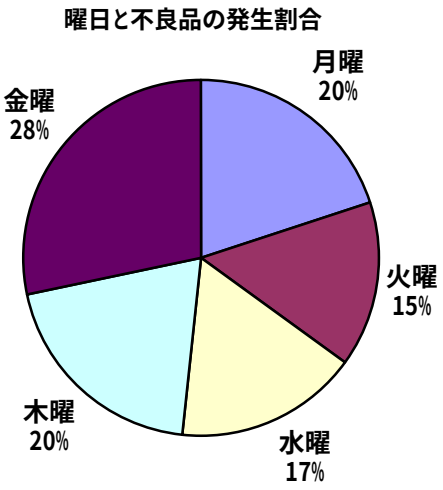
問題３－１　以下のデータで棒グラフと円グラフを描け

表　曜日と不良品の発生件数

曜日	月曜	火曜	水曜	木曜	金曜	合計
発生件数	１２	９	１０	１２	１７	６０

解答



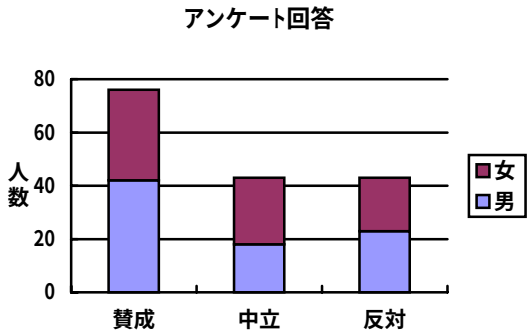


問題３－２　以下のデータで積み上げ棒グラフを描け。

表　アンケート回答

	賛成	中立	反対
男	４２	１８	２３
女	３４	２５	２０

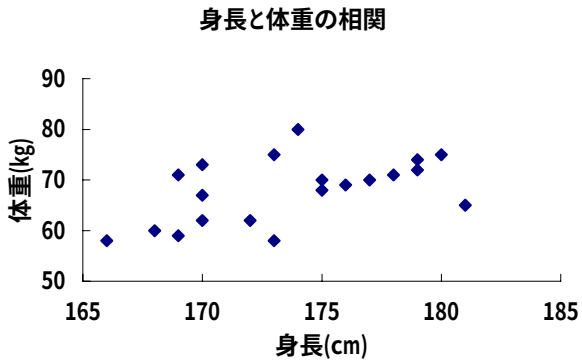
解答



問題３－３　以下のデータで相関図（散布図）を描け。

身長(cm)	体重(kg)	身長(cm)	体重(kg)
169	71	170	62
175	68	180	75
170	67	177	70
179	72	175	70
176	69	172	62
174	80	166	58
173	75	168	60
181	65	173	58
179	74	169	59
178	71	170	73

解答



相関係数 = 0.513

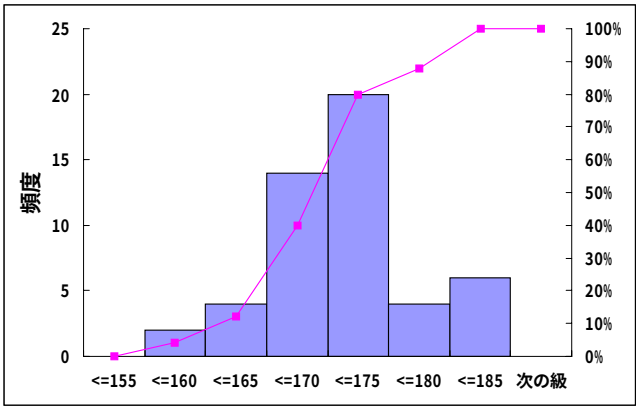
問題４－１　以下の 50 人の身長データ(cm)で度数分布表を作り、ヒストグラムを描け。

184.9, 170.6, 164.7, 165.3, 165.1, 170.5, 171.2, 159.8, 167.2, 170.2,
165.9, 170.3, 171.9, 183.7, 158.5, 169.8, 181.5, 172.2, 171.7, 164.5,
166.0, 171.1, 178.5, 173.6, 180.4, 165.1, 169.4, 172.4, 174.2, 164.6,
176.9, 180.6, 170.4, 178.7, 166.2, 172.5, 172.6, 166.2, 170.2, 170.2,
165.0, 175.3, 165.6, 174.8, 169.7, 169.3, 169.6, 174.0, 180.5, 172.2

解答

Excel 使用例

ﾃﾞｰﾀ区間	頻度	累積 %
<=155	0	.00%
<=160	2	4.00%
<=165	4	12.00%
<=170	14	40.00%
<=175	20	80.00%
<=180	4	88.00%
<=185	6	100.00%
次の級	0	100.00%



統計ソフト[Statistica] 使用例

階級	度数	相対度数(%)	累積度数	累積相対度数(%)
155 < x <= 160	2	4	2	4
160 < x <= 165	4	8	6	12
165 < x <= 170	14	28	20	40
170 < x <= 175	20	40	40	80
175 < x <= 180	4	8	44	88
180 < x <= 185	6	12	50	100
計	50	100		

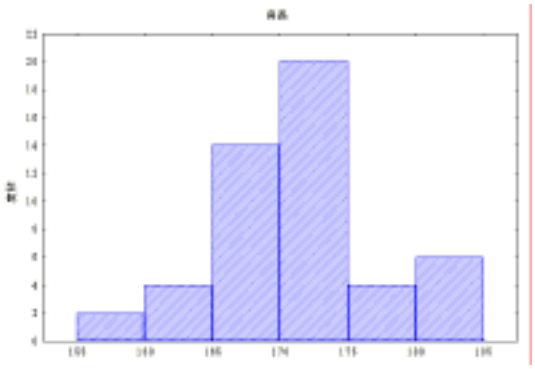


図　身長分布のヒストグラム

問題５－２　以下の表の身長のデータで、MS-Excel を用いて平均、分散、不偏分散、標準偏差（２種類）及び、基本統計量を求めよ。

解答

表計算機能を用いた場合

身長	(身長-平均)^2	身長^2
184.9	222.755625	34188.01
170.6	0.390625	29104.36
164.7	27.825625	27126.09
165.3	21.855625	27324.09
165.1	23.765625	27258.01
170.5	0.275625	29070.25
171.2	1.500625	29309.44
159.8	103.530625	25536.04
167.2	7.700625	27955.84
170.2	0.050625	28968.04
165.9	16.605625	27522.81
170.3	0.105625	29002.09
171.9	3.705625	29549.61
183.7	188.375625	33745.69
158.5	131.675625	25122.25
169.8	0.030625	28832.04
181.5	132.825625	32942.25
172.2	4.950625	29652.84
171.7	2.975625	29480.89
164.5	29.975625	27060.25
平均	分散(1)	分散(2)
169.975	46.043875	46.043875
標準偏差 (関数)	標準偏差	標準偏差
6.961841484	6.78556372	6.78556372
範囲	不偏分散	
26.4	48.46723684	
	不偏標準偏差	
	6.961841484	

統計機能を用いた場合

身長	
平均	169.975
標準誤差	1.556715
中央値 (ﾒｼﾞｱﾝ)	170.25
最頻値 (ﾓｰﾄﾞ)	#N/A
標準偏差	6.961841
分散	48.46724
尖度	0.578674
歪度	0.742487
範囲	26.4
最小	158.5
最大	184.9
合計	3399.5
標本数	20

使用する関数

=sum(範囲)	合計
=average(範囲)	平均
=stdev(範囲)	標準偏差
=max(範囲)	最大
=min(範囲)	最小