情報処理のレポートについて

2008 年 1 月 8 日

横 山　暁

レポートの採点での総評や，間違いが多かった部分について記述したものを Web にアップしておきます．（http://www.ae.keio.ac.jp/~satoru_y/shouei/Report07_result.pdf）
各自でチェックして復習をしてください．

＜採点・評価について＞

レポートは 25 点満点で行いました．各問題で解答をすべき内容が書いてありかつ正解であれば 21 点，その他に途中の計算方法や考察が書いてあるごとに最大 4 点加点しています．成績は，前期の出席・レポートと後期の出席・レポートを総合して，最終的な成績評価を行っています．

＜レポートの総評＞

今回のレポートは，同じデータを用いているため，Excel の計算は全員同じになります．そのため，たとえ Excel の計算は友達と協力したとしても，Word で提出してもらうことによって各自でレポートを作成し，内容の理解（復習）をして欲しいと思っていました．

しかし，残念ながら昼間部，夜間部を合わせて，同じレポートが 4 パターンほどありました．しかもその 4 パターンのすべてが，Excel の計算を間違っているだけでなく，レポートの指示に従っておらず，特に帰無仮説や対立仮説を書いてありませんでした．さらに誤字もそのまま（「対立仮説を**指示**する」になっている）で，酷いレポートでは，名前を変えただけのレポートもありました．

内容が難しかったのかもしれませんが，時間も十分にあったはずですし，基本的にすべて後期の授業で学習したことであり，講義資料もすべてウェブページに掲載してあるため，きちんとしたレポートを作成することができたはずです．現にきちんとしたレポートを提出した人も多くいました．きちんとしたレポートを提出した人は，内容もきちんと理解できているようで，レポートの出来は非常に良いように感じました．

先生によっては，同じレポートは一切採点を行わないこともあるだけでなく，その授業を不合格することすらあり得ます．今回は採点行いましたが，今後十分に注意してください．

１．記述統計

（１） 平均・分散等の計算（1 点）

基本的によくできていました．ただし，答えを入れてもらおうとしていた C 列の隣の D 列にある「ヒストグラムの階級」の列を使って計算した人がいたため，平均から最小値までと範囲の値が間違っている人がいました．

平均	24.9
分散	7.116092
標準偏差	2.6676
最大値	30.0
最小値	20.5
中央値	24.5
第 1 四分位	23
第 3 四分位	27.375
最頻値	24.5
範囲	9.5

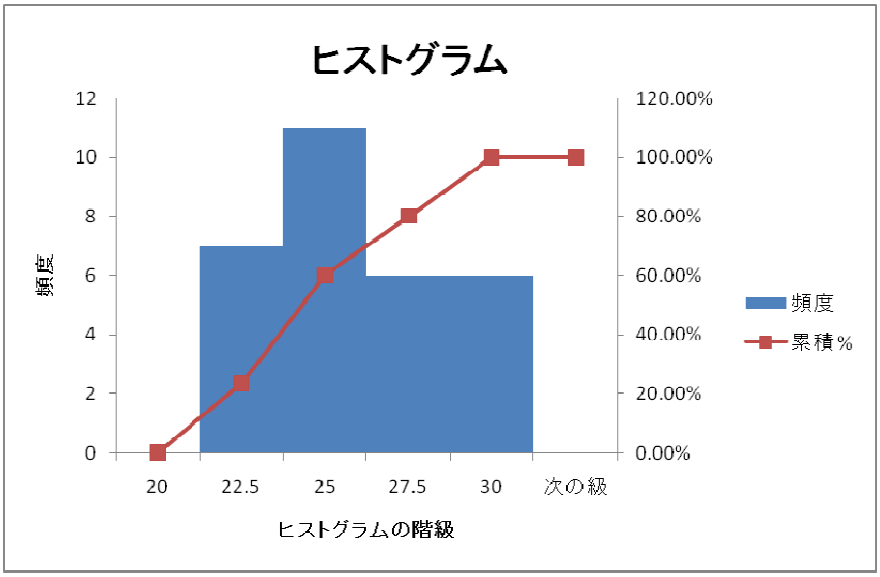
（２） 度数分布表・ヒストグラムの作成（各 1 点）

データと D 列にある「ヒストグラムの階級」の値を使って「分析ツール」の「ヒストグラム」で作成できます．その際に「累積度数分布の表示」と「グラフ作成」のチェックボックスにチェックをします．

ヒストグラムは，レポートの指示にもあったように，頻度の系列において「データ系列の書式設定」より棒の間隔を 0 にします．

一部のレポートで，データをそのまま棒グラフにしているものがありましたが，間違いです．

階級	頻度	累積 %
20	0	0.00%
22.5	7	23.33%
25	11	60.00%
27.5	6	80.00%
30	6	100.00%
次の級	0	100.00%


ヒストグラム
12
10
8
6
4
2
0
頻度
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
20
22.5
25
27.5
30
次の級
ヒストグラムの階級
頻度
累積 %

２．回帰

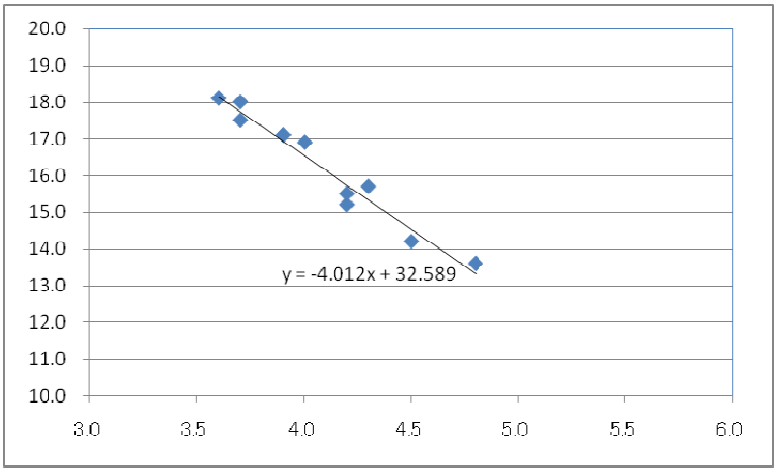
（１） 相関係数（1 点）

相関係数は CORREL という関数で計算します．-0.9786 になります．

A 列 B 列で計算すると，これらの列は「セルの書式設定」の「表示形式」で小数点以下の桁数が 1 になっているため，-1.0 と表示されてしまいますが，散布図を描いてもわかるように，データは一直線に並んでいませんので，-1 にはなりません．

（２） 散布図・回帰式（2 点）

散布図は以下の図のようになります．



回帰式は，散布図の点を右クリックして「近似曲線の追加」で「線形近似」をクリックし「グラフに数式を表示する」にチェックをすれば回帰式が求まります．

その他の求め方として，「分析ツール」の「回帰分析」からも求められます．

ただし，値は求まりますが，式の形式ではないので，きちんと式で表してください．

３．検定・信頼区間１

（Z 検定，t 検定共に仮説が書いてあって 1 点，計算ができていて 1 点，仮説が棄却されるかどうかが書いてあって 1 点，信頼区間が合っていて 1 点，合計 8 点）

（１） Z 検定・信頼区間

問題が両側検定という指示があるので，仮説は以下のようになります.

帰無仮説：データの（母集団での）平均が 20 である

対立仮説：データの（母集団での）平均が 20 でない

両側確率を求めます．授業中にはあまり触れませんでしたが，資料には書いてあります．片側（上下）確率を求め，小さい方を 2 倍します．0.0177 になりますので，これを 0.05 と比較します.

0.0177<0.05 ですので，帰無仮説を棄却し対立仮説を支持します．つまりデータの母集団の平均は 20 でないことになります.

信頼区間は，標本平均±正規分布両側 5%点×既知の標準誤差なので，

(20.05, 20.55)になります.

（２） t 検定・信頼区間

Z 検定と同じく仮説は以下のようになります.

帰無仮説：データの（母集団での）平均が 20 である

対立仮説：データの（母集団での）平均が 20 でない

t 値を求め，両側確率を TDIST で求めます．0.040 になります．これを 0.05 と比較します.

0.040<0.05 ですので，帰無仮説を棄却し対立仮説を支持します．つまりデータの母集団の平均は 20 でないことになります.

信頼区間は，標本平均±t 分布 5%点×標準誤差なので，

(20.02, 20.58)になります.

ちなみに，Z 検定の信頼区間より t 検定の信頼区間の方が広くなります.

Excel での計算は以下のようになります.

	A	B	C
1	No	データ	B 列の計算式
2	1	20.3	
3	2	20.5	

4		3	19.8	
5		4	20.2	
6		5	20.4	
7		6	20.8	
8		7	20.7	
9		8	20.7	
10		9	19.7	
11		10	19.9	
12	平均		20.3	'=AVERAGE(B2:B11)
13	標準偏差		0.394405	'=STDEV(B2:B11)
14	分散		0.155556	'=VAR(B2:B11)
15	標本の大きさ		10	'=COUNT(B2:B11)
16	従来の母平均		20	
17	既知の母標準偏差		0.4	
18	信頼度		0.95	
19	Z 検定			
20	既知の標準誤差		0.126491	'=B17/SQRT(B15)
21	z		2.371708	'=(B12-B16)/B20
22	下側確率		0.991147	'=NORMSDIST(B21)
23	上側確率		0.008853	'=1-B22
24	両側確率		0.017706	'=2*MIN(B22:B23)
25	区間推定			
26	両側 5%点		1.959964	'=NORMSINV(1-(1-B18)/2)
27	下側信頼限界		20.05208	'=B12-B26*B20
28	上側信頼限界		20.54792	'=B12+B26*B20
29	t 検定			
30	標準誤差		0.124722	'=B13/SQRT(B15)
31	自由度		9	'=B15-1
32	t		2.405351	'=(B12-B16)/B30
33	両側確率		0.039549	'=TDIST(B32,B31,2)
34	区間推定			
35	両側 5%点		2.262157	'=TINV(1-B18,B31)
36	下側信頼限界		20.01786	'=B12-B35*B30
37	上側信頼限界		20.58214	'=B12+B35*B30

４．検定・信頼区間２

（分散の検定，平均の差の t 検定共に仮説が書いてあって 1 点，計算ができていて 1 点，仮説が棄却されるかどうかが書いてあって 1 点，平均の佐野 t 検定の信頼区間が合っていて 1 点，合計 7 点）

（１） 分散の検定

分散の検定は分散が等しいかを検定しますので，仮説は以下のようになります．

帰無仮説：2 つのデータの分散が等しい

対立仮説：2 つのデータの分散が等しくない

分散比を計算し，FDIST から片側確率を求め，片側確率の小さい方を 2 倍することで両側が求まります．0.928 になります．明らかに 0.05 より大きいので，帰無仮説は棄却できません．つまりほぼ間違いなく 2 つのデータの分散は等しくないことになります．

（２） 平均の差の t 検定

分散に差がないことが分かったので，Welch の検定ではなく，平均の差の t 検定を行います．仮説は以下のようになります．帰無仮説：まず DEVSQ で偏差平方和を求め共通の分散の推定値を求めます．さらに平均の差から平均の差の標準偏差を求めます．この 2 つの値から t 値を求め，TDIST で両側確率を求めます．0.024 になります．0.024<0.05 より帰無仮説を棄却します．つまり平均に差があることになります．

信頼区間は平均の差±t 分布 5%点×標準誤差なので，

(-7.39, -0.58)になります．

Excel での計算は以下のようになります．

	A	B	C	D
1	No.	データ 1	データ 2	B 列の計算式
2	1	78	81	
3	2	80	84	
4	3	79	82	
5	4	83	88	
6	5	82	86	
7	6	85	83	
8	7	78	78	

9	8	74	84	
10	9	76	89	
11	10	84		
12	平均	79.9	83.88889	'=AVERAGE(B2:B11)
13	標準偏差	3.573047	3.443996	'=STDEV(C2:C11)
14	分散	12.76667	11.86111	'=VAR(C2:C11)
15	標本の大きさ	10	9	'=COUNT(C2:C11)
16	自由度	9	8	'=C15-1
17	等分散性の検定			
18	分散比	1.076347		'=B14/C14
19	下側確率	0.464063		'=FDIST(B18,B16,C16)
20	上側確率	0.535937		'=1-B19
21	両側確率	0.928126		'=2*MIN(B19:B20)
22	平均の差の t 検定 or Welch の方法			
23	偏差平方和	114.9	94.88889	'=DEVSQ(C2:C11)
24	共通の分散の推定値	12.34052		'=(B23+C23)/(B16+C16)
25	平均の差	-3.98889		'=B12-C12
26	平均の差の標準誤差	1.61407		'=SQRT((1/B15+1/C15)*B24)
27	t 値	-2.47132		'=B25/B26
28	両側確率	0.024333		'=TDIST(ABS(B27),B16+C16,2)
29	t 分布 5%点	2.109816		'=TINV(0.05,B16+C16)
30	95%下側信頼限界	-7.39428		'=B25-B29*B26
31	95%上側信頼限界	-0.5835		'=B25+B29*B26