

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

【統計一口メモ 第 1 話】

＜それでも分散の F 検定やりますか？＞

名古屋市立大学大学院医学研究科 非常勤講師 薬学博士 松本一彦

統計の教科書には「Student の t 検定をやる前に分散の検定をやって、有意差があったら Welch 検定をやる」と書かれています。GLP など決定樹を使っている施設では、機械的に F 検定をやって、有意差が出たら Welch 検定に進む形式になっているところもあります。「Welch 検定って何？」それは次回のテーマとして、今回は「分散の検定って何？」をやしましょう。

§ 1. 必要な差の検定は分散ですか？平均値ですか？

2 群の差を考えると、その“差”とは何を指しているかによって方法が異なります。必ずしも分散の検定をやってから t 検定をやる必要はありません。それは次の 2 つの状況によります。

1) 平均値の差だけで、バラツキ(分散)は問題にしていない場合、すなわち等分散を仮定している場合は、「等分散の検定」をやらずに、“t 検定”を実施する。

2) 平均値の差も、バラツキ(分散)の差も、両方とも問題にしている場合は、「等分散の検定」で差が認められれば、たとえ、t 検定で差が認められなくても、差があると結論する。 丹後俊郎「医学への統計学」朝倉書店

§ 2. F 検定って？

例題：対照群と薬剤投与群で IL-X 値を測定した。両群の分散に差があるといえるだろうか？

個体番号	対照群	投与群
1	153	153
2	153	146
3	152	131
4	156	152
5	162	130
6	151	146
7	151	156
8	148	142
9	210	147
10		153

	対照群	投与群
n	9	10
平均	159.6	145.6
平方和	2986.2	730.4
自由度	8	9
平均平方	373.28	81.16

分散比の有意性検定

分子の自由度	9	
分母の自由度	8	
F 比	0.217	=81.16/373.28
上側 p 値	0.9825	= FDIST 関数 FDIST(0.217, 9, 8)
下側 p 値	0.0175*	=1-0.9825
両側 p 値	0.0351*	=2 x 0.175

F 検定は両群の平均平方の比を F 比として求め、FDIST 関数を使って上側 p 値を算出する。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

下側p値は1－上側 p 値で求め、その 2 倍を両側p値として算出する。

結果:p=0.0351 で両群の分散に差があるといえる。

※等分散性が棄却されたら Welch 検定、棄却されなかったら t 検定を用いる場合は有意水準を 5% ではなく 25% 程度にした方がよい¹⁾ (吉村功編著 毒性・薬効データの統計解析 サイエントリスト社)。

帰無仮説 $H_0: \sigma_1 = \sigma_2$ が成立するとき、それぞれから求められた平均平方の比 V_2/V_1 は 1 前後の値を取るであろう。その分布を F 分布という。F 比を計算するとき、どちらの平均平方を分母に取るかは自由である。現在は F 分布の確率が簡単に計算できるので、 $F = V_2/V_1$ と固定するのが良いであろう。両側検定の p 値は、上側 p 値と下側 p 値の内の小さい方の 2 倍とする。
(芳賀敏郎:医薬品開発のための統計解析 第 1 部 サイエントリスト社)

§ 3. F 検定より Levene (レビン) 検定がお奨め

Levene 検定は OECD の統計ガイドライン²⁾にとりあげられている他、芳賀敏郎著「医薬品開発のための統計解析 2」に紹介されて広く知られることとなった手法です。

<Levene 検定の特徴>

偏差とは“観測値－平均値”，すなわち、バラツキの大きさを示す。

偏差は正負の値を取るので、絶対値に直す。

$$|\text{偏差}| = |\text{観測値} - \text{平均値}| = |X_{ij} - \bar{X}_i|$$

絶対値をつけた偏差を観測値と見なして、「平均値の差の t 検定」をするのが Levene の検定。

<F 検定との違い>

F 検定は“平均平方”すなわち“偏差の 2 乗の和”を使っているのに対して、Levene の検定は“偏差の絶対値の和”を使っている。もしデータに外れ値が含まれているとき、偏差の 2 乗を用いる F 検定はその影響を強く受ける。Levene の検定は外れ値の影響を受け難いという利点がある。

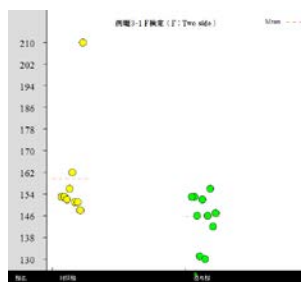
$$\text{Levene} = |X_{ij} - \bar{X}_i|$$

$$F = \frac{\sum (X_{ij} - \bar{X}_i)^2}{\dots}$$

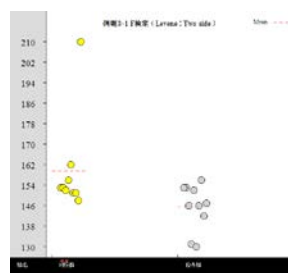
Levene の検定は、3群以上にも対応しており、F 検定よりも使い道は広い。

§ 4. 分散の検定を Pharmaco Basic で解析すると

F 検定



Levene 検定

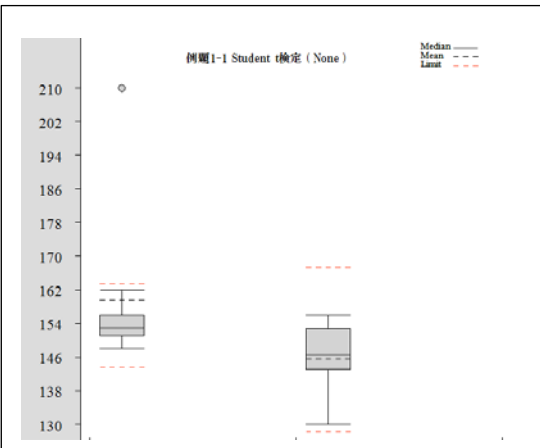


本メモは、オリジナルは「安全性評価研究会（谷学）」HP（<https://tanigaku.jp/>）に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

平方和	2986.222
平均平方	373.278
N数	9
群名	投与群
平均	145.6
平方和	730.400
平均平方	81.156
N数	10
P値	0.0351*

群名	対照群
平均	159.6
標準偏差	19.3
N	9
群名	投与群
平均	145.6
標準偏差	9.0
N	10
P値	0.3324

結果:F 検定では $p=0.035$ で有意差となるが、Levene 検定では $P=0.332$ で有意差とならない。もし、F 検定を 5%ではなく 25%とすれば、F 検定でも有意とはならないが、それでも Levene 検定より値は小さい。それは、データに外れ値があることによる。外れ値の確認は箱ひげ図で行う。



Pharmacology ホームページ:<https://Pharmacology.club/>

- 1) 永田靖、実験計画法をめぐり諸問題、品質、18:196-204,1988
- 2) 松本一彦、松田眞一、「化学物質試験のための OECD 試験ガイドライン 魚類初期生活段階毒性試験 無影響量(NOEC)決定に向けた統計ガイダンス TG210, ANNEX5」、臨床評価 26:(1)、71-82, 2018

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

【統計1口メモ 第2話】

<それでも Welch (ウェルチ) 検定やりますか?>

第1話では「それでも分散分析やりますか?」を取り上げました。多くの教科書には、2群間の平均値の差の検定で、不等分散のときに Welch 検定をするようにと書かれています。GLP で決定樹に従って解析している施設では、F 検定で有意となった場合に、機械的に Welch 検定にいくようになっているケースが多いようです。

§ 1. Welch 検定とは?

- t検定が2群の分散が等しい、すなわち等分散を条件にしているのに対して、分散が異なる不等分散のときに用いる検定です。分散が異なるため、調整した自由度(等価自由度)を用いて t検定をします。

等価自由度(ν^*) は下記の式で求めますが“少数自由度”で整数にはなりません。(※少数自由度はエクセルでは求まらない。したがって、エクセル分析ツールには Welch 検定は入っていない)。

$$\nu^* = \frac{(V_1/n_1 + V_2/n_2)^2 \times (\nu_1 \times \nu_2)}{\nu_2(V_1/n_1)^2 + \nu_1(V_2/n_2)^2}$$

帰無仮説 $H_0: \mu_x = \mu_y$			
対立仮説 $H_1: \mu_x \neq \mu_y$			
	nが等しい場合		
	癌患者群	非癌患者群	
1	153	153	
2	153	146	
3	152	138	
4	156	152	
5	158	140	
6	160	128	
7	162	116	
8	151	146	
9	151	156	
10	150	142	
n	10	10	
平均	154.6	141.7	12.9
平方和	156.4	1340.1	1496.5
自由度	9	9	18
平均平方	17.38	148.9	83.14
平均平方/n	1.74	14.89	11.07
平均値の差の標準誤差			4.08
α	0.05		
t	3.1635	自由度	
p値(両側)	0.00902	11	
p値(両側)	0.00817	12	
p値(両側)	0.00895	11.07	等価自由度
	t検定	Welch	
se	4.078	4.078	
t	3.164	3.164	← t 値は同じ
自由度	18	11.07	
p値(両側)	0.0054	0.0090	← p 値は異なる

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

- t検定と Welch 検定では平均値の差の標準誤差(s.e)の求め方が異なります。

t 検定

$$V[d] = \sigma^2/n_1 + \sigma^2/n_2 = (1/n_1 + 1/n_2)\sigma^2$$

$$\sigma^2 \sim V = \frac{S_1 + S_2}{v_1 + v_2} = \frac{S_1 + S_2}{n_1 + n_2 - 2}$$

$$s.e.[d] = \sqrt{V[d]} = \sqrt{(1/n_1 + 1/n_2)V}$$

$$t = (\bar{x}_2 - \bar{x}_1) / s.e.[d]$$

Welch 検定

$$V[d] = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

$$\sigma_1^2 \sim V_1 = \frac{S_1}{v_1} = \frac{S_1}{n_1 - 1}$$

$$\sigma_2^2 \sim V_2 = \frac{S_2}{v_2} = \frac{S_2}{n_2 - 1}$$

$$s.e.[d] = \sqrt{V[d]} = \sqrt{V_1/n_1 + V_2/n_2}$$

$$t = (\bar{x}_2 - \bar{x}_1) / s.e.[d]$$

- nが等しい場合、等分散を仮定したときと同じt値を示しますが、自由度が小さくなるため、有意となりにくくなります。ここではt検定の p=0.0054 で Welch 検定ではp=0.0090 と大きくなりました。すなわち、n数がほぼ同じ場合は、t検定で有意となるデータが Welch 検定では有意とならないケースがあるということです。

蛇足：差を出したくないようなケースで Welch 検定を使うと喜ばしい結果になるので敢えて使う人もいるかもしれませんが、本来、統計手法は差を出したい人向けの手法で、帰無仮説 H_0 を $\mu_1 = \mu_2$ 、対立仮説を $\mu_1 \neq \mu_2$ と置いて、帰無仮説を否定して、対立仮説を採用します。

- 不等分散の場合には、しばしば外れ値があり、正規分布とはみなせない場合が多く、ノンパラメトリック検定の対象となります。本試験の場合は Wilcoxon 順位和検定でp値が 0.0080 で、Welch 検定より小さな値となりました。それだと Welch 検定より Wilcoxon 順位和検定を使うのでは？

§ 2. Welch 検定はエクセル分析ツールではできません。

<エクセル分析ツール>

t-検定：分散が等しくないと仮定した2標本による検定

	変数 1	変数 2
平均	70.8125	85.22222
分散	372.6958	945.2635
観測数	16	36
仮説平均との差異	14.4	
自由度	44	
t	-4.09274	
P(T<=t) 片側	8.96E-05	
t 境界値 片側	1.68023	
P(T<=t) 両側	0.000179	
t 境界値 両側	2.015368	

*t値やp値がデタラメ

エクセル分析ツールでは小数点自由度が求められないため、Welch検定という名称は使えない。

エクセルでは少数自由度を求めることができません。したがって、Welch 検定という名前で解析することはできません。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

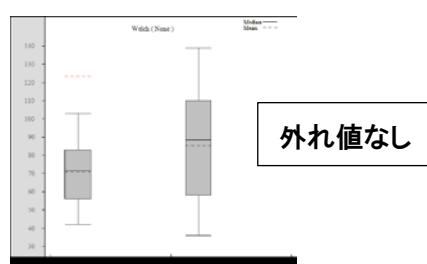
§ 3. Welch 検定の出番はないのだろうか？

外れ値があるときには正規分布が疑われるため、ノンパラの Wilcoxon 順位和検定を用いることが多くなります。n数が等しいときは、t検定と同じ結果になるので、わざわざ Welch 検定を用いる必要はありません。そうすると、Welch 検定の出る幕はないのでは？ いえ、下記のようなケースでは出番があります。もし、外れ値がなくて、それでも分散に差があつてn数が倍ほど違うような例では、t検定や Wilcoxon 検定より検出力が高い場合があります。次の例を見てください。

<Pharmaco Basic での解析>

16例	36例	
癌患者	非癌患者	
103	108	74
73	118	116
69	90	36
103	90	53
89	49	93
58	58	78
81	59	80
78	61	125
48	97	110
59	39	58
50	99	139
71	52	132
72	43	57
42	87	65
47	104	138
90	65	119
	123	111
	42	100

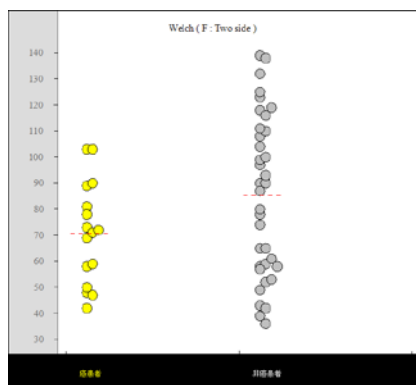
箱ひげ図



Student's t検定

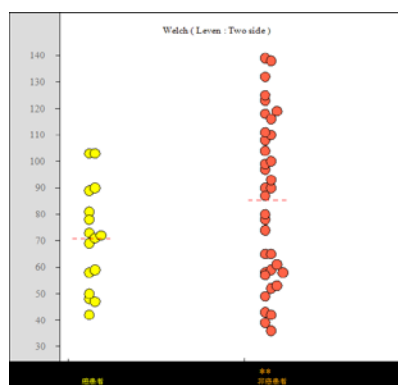


F 検定



F 検定では有意差なし(p=0.057)

Levene 検定

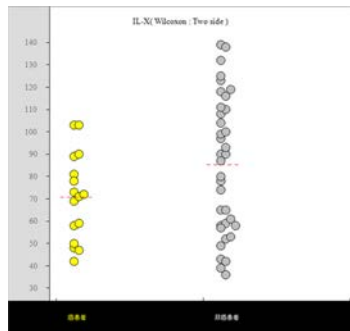


Levene 検定で有意差あり(p=0.008)

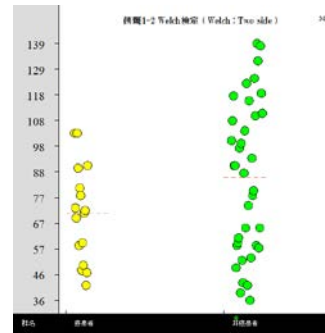
本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

外れ値がない場合は、偏差の 2 乗を使う F 検定より、偏差の絶対値を使う Levene 検定の方が検出力は高いと言われています。

Wilcoxon 順位和検定



Welch 検定



結果:2 群間の差は Welch 検定でのみ有意差がみとめられました($p=0.047$)。分散の検定では、F 検定では有意差はみとめられず、Levene 検定でみとめられたのは、外れ値がなかったことが影響していると思われます。それは、Wilcoxon 順位和検定で有意差がみとめられなかったことと一致しています($p=0.115$)。すなわち、Welch 検定の出番は「n数が倍ほど異なり、外れ値がなく、分散が異なるデータ」と言えます。

Pharmaco ホームページ:<https://pharmaco.club/>

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

【統計一口メモ 第3話】

＜t 検定の繰り返しはダメ、多重比較をせよと言われた！＞

A 氏が学会誌に投稿したときのレプリーコメントは、「t 検定の繰り返しをしている。多重比較検定でやり直すように」でした。そのデータを示します。

新抗癌剤Ascleの試験のサイトカインIL-X値（t検定）				
群	対照群	10mg群	30mg群	100mg群
例数	10	10	10	10
平均値	928.3	911.7	891.9	891.2
標準偏差	27.9	21.2	41.9 **	42.9**

t 検定は2群間比較なので、対照群と 10mg 群、対照群と 30mg 群、対照群と 100mg 群の 3 回を繰り返しています。このようなt検定の繰り返しは、有意水準を5%でやっているつもりが 14%にもなっているのです。本当は有意でないのに 有意と出てしまいます。そこで、レプリーは「多重性の調整をして 14%を 5%にしろ」というコメントをつけることになります。なぜ、5%が 14%になるのかはいつかお話することにして。今回は、レプリーの言う多重比較って何？t検定の代わりに何をすればいいのだろうか？を解説します。

多重比較には大きく分けて3つのタイプがあります。1つ目は「Dunnett (ダネット) 型」で、必ず対照群と比較します。群構成は「薬剤名」や「疾患名」のような名義尺度になります。群を入れ替えても構いません 2 つ目は「Williams (ウィリアムズ) 型」で Dunnett 検定とは異なり、群は用量になっています。なお、データは単調増加・減少を示していることが条件になっています。したがって、群の入れ替えはできません。 3 つ目は「Tukey (チューキー) 型」で全群間において総当たりするような検定法です。Tukey 検定の他に Bonferroni 検定と Holm 検定があります。今回は Dunnett 検定と Williams 検定をとりあげます。

Dunnett 型

動物No.	A剤	B剤	C剤	E剤
1	7	8	11	13
2	9	9	12	12
3	8	10	12	12
4	6	8	10	11
5	9	9	11	14
6	8	9	13	12
7	11	10	9	11
8	10	12	10	10
9	8			

Williams 型

用量 (mg/kg)	0 mg/kg	15 mg/kg	30 mg/kg	60 mg/kg	90 mg/kg
1	415	387	357	361	291
2	380	378	379	351	301
3	391	359	401	378	321
4	413	391	412	332	351
5	372	362	392	318	311
6	359	351	356	344	281
7	401	348	366	315	291

Tukey 型

対照群	Ascle 1mg	Ascle 3mg	Cisplatin 1mg	Cisplatin
12	17	18	40	28
15	16	39	21	20
14	27	20	29	39
12	9	19	78	34
15	15	37	15	48
9	18	17	32	55
17	29		18	41
	15			

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

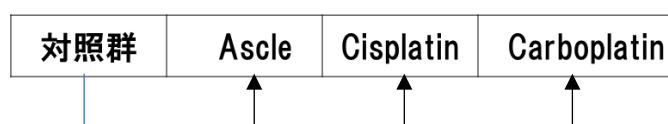
A 氏は、レフリーのコメントに応じて、市販ソフトで Dunnett 検定を実施したところ、t検定での有意差がみられなくなりました。

	平均値の差	t値		有意確率	
		t 検定	Dunnett	t 検定	Dunnett
対照群—10mg	16.6	1.495	1.067	0.152	>0.05
対照群—30mg	36.5	2.289	2.344	0.034*	>0.05
対照群—100mg	37.1	2.293	2.389	0.034*	>0.05
棄却限界値 (5%)		2.101	2.452		

なぜ、Dunnett 検定では有意差がみられなくなったのだろうか？それには、Dunnett 検定って何をやっているのかを知る必要があります。

Dunnett 検定とは 1 つの対照群と 2 つ以上の処理群があつて、対照群と処理群の対比較を全群で同時に検定する方法です。同時にという意味は、2群のt検定と違って複数の処理群の分散を一緒にすることです。対比較とは、対照群とそれぞれの処理群の平均値の差を意味します。

Dunnett 検定は、本来、次の例のような場合に用います。



対照群を固定して各群との平均値の差を求めます。この場合の群には用量が無く群間を入れ替えても同じ結果になります。

多重比較と 2 群比較の違いは分母にくる分散の求め方の違いです。

t 検定では、分母の分散は2群だけですが、

$$t = \frac{x_1 - x_i}{\sqrt{V(1/n_1 + 1/n_i)}}$$

多重比較では分散が全群の合計 V_E となります。

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{V_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

全ての群の分散が分子にきます。

$$V_E = \frac{\sum_{i=1}^a (n_i - 1) V_i}{\phi_E} = \frac{9 \times 777.6 + 9 \times 449.1 + 9 \times 1757.9 + 9 \times 1839.1}{36}$$

従って、分子はt検定と同じ値ですが、分母が大きくなって、t値は小さくなります。検定は t 表ではなく Dunnett 表を使います(実際にはソフトが計算してくれます)。今回のデータでは全群で有意差がみられなくなりました。

A 氏はこのまま投稿を諦めなくてはならないのでしょうか？ もう少し、他の多重比較についても検討してみましょう。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

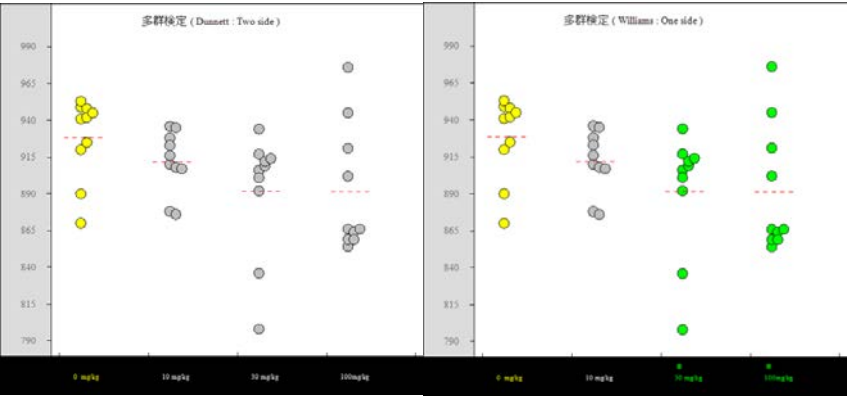
Williams 検定は Dunnett 検定とは異なり群構成は用量になっています。また、データが単調増加・減少を示していることが条件になっています。さらに、高用量群から検定をしていく下降手順となっていて、有意差がなくなった時点で、その下の用量の検定は行いません。例えば、

対照群	Ascle10mg	Ascle30mg	Ascle100mg
-----	-----------	-----------	------------

このような群構成で 100mg で有意差があり、30mg では有意差がなくなった場合は、その下の 10mg は検定せずに、保留(すなわち有意差なし)とします。もし、30mg の方が 100mg より高値(低値)の場合は、両群の平均値を使います(調整平均)。なお、単調増加・減少ということから片側検定となります。検定には Dunnett 表ではなく Williams 表を使います(実際にはソフトで計算します)。A 氏のデータを Pharmaco Basic の Williams 検定で実施してみましょう。

Dunnett 検定

Williams 検定



検定法	Dunnett (両側)			検定法	Williams (片側)		
群名	0 mg/kg			群名	0 mg/kg		
平均	928.3			平均	928.3		
標準偏差	27.889			標準偏差	27.889		
N	10			N	10		
群名	10 mg/kg			群名	10 mg/kg		
平均	911.7			平均	911.7		
標準偏差	21.066			標準偏差	21.066		
N	10			N	10		
P値	0.5226			P値	0.1993		
群名	30 mg/kg			群名	30 mg/kg		
平均	891.9			平均	891.9		
標準偏差	41.914			標準偏差	41.914		
N	10			N	10		
P値	0.0639			P値	0.0144 *		
群名	100mg/kg			群名	100mg/kg		
平均	891.2			平均	891.2		
標準偏差	42.892			標準偏差	42.892		
N	10			N	10		
P値	0.0576			P値	0.013 *		

Williams 検定ではt検定と同じ結果が得られました。A 氏のデータは用量があり、単調減少していることから検定は Williams 検定を用いるべきでしょう。A 氏は無事論文審査を通りました。

なお、JMP、SPSS、Prism ソフトは Williams 検定を搭載していません。Williams 検定を搭載して

本メモは、オリジナルは「安全性評価研究会（谷学）」HP（<https://tanigaku.jp/>）に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

いるソフトは少ない状況です。

Tukey、Bonferroni、Holm 検定は、後日とりあげることにします。

【統計1口メモ 第4話】

<生殖発生毒性試験で出現率が 0(ゼロ)多発の場合の統計手法は？>

「生殖発生毒性試験において、ラットの波状肋骨出現率に関して、平成 9 年のガイドラインでは対照群と高用量群で有意差検定を実施し、有意となった場合には、多重比較検定をすることが記載されています(平成 9 年 4 月 14 日薬審 316 号)。それに関して、2002 年に医薬安全性研究会で「ドラマ：生殖発生毒性試験の統計」という題で寸劇が実施されました(医薬安全性研究会第 90 回定例会資料)。ここでは、そのドラマの結論に沿った統計手法について解説します。なおデータは一部改変しています。

ラット波状肋骨出現率

	N=10			出現率%	
0 mg/kg	10.5	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0
15 mg/kg	12.5	12.5	11.1	0.0	0.0
	0.0	0.0	0.0	0.0	0.0
30 mg/kg	25.0	20.0	12.5	12.5	0.0
	0.0	0.0	0.0	0.0	0.0
60 mg/kg	87.5	22.2	20.0	14.3	12.3
	0.0	0.0	0.0	0.0	0.0

前回の「統計一口メモ 第 3 話」では、パラメトリック(正規分布)の場合、群構成が用量の場合は Dunnett 検定ではなく、Williams 検定を使うことを勧めました。でも、このようにゼロが多いデータは正規分布していないのでノンパラメトリック検定を使うことになります。Dunnett のノンパラメトリックには Steel(スチール)検定があり、Williams 検定には Shirley Williams(シャリー・ウィリアムズ)検定があります。先のドラマの結論は Shirley Williams 検定を使うべきだとなっていました。そこで、今回は Shirley Williams 検定について解説します。

§ 1. Shirley Williams 検定とは

【永田・吉田本】¹⁾から引用します。

用量	1	2	3	4	5	6	7	8	9	10	11	12
0 mg/kg	13	23	8	17	25	34	18	26	10	28	18	21
15mg/kg	26	22	30	38	15	24	18	11	21	30	31	23
30mg/kg	22	10	29	37	22	13	29	28	21	16	21	26
60mg/kg	26	34	30	45	17	19	27	18	36	24	25	31

①第 1 群(0mg/kg)から第 4 群(60mg/kg)を併せて小さいものから順位をつけます。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

用量					順位 r_{ik}									R_{i4} (合計)
0 mg/kg	5.5	23.5	1	9.5	27.5	43.5	12.5	30.5	2.5	34.5	12.5	17.5		220.5
10mg/kg	30.5	21	39	47	7	25.5	12.5	4	17.5	39	41.5	23.5		308.0
20mg/kg	21	2.5	36.5	46	21	5.5	36.5	34.5	17.5	8	17.5	30.5		277.0
40mg/kg	30.5	43.5	39	48	9.5	15	33	12.5	45	25.5	27.5	41.5		370.5

②第 1 群を除いて平均順位を計算します。

$$y_{24} = (R_{24} + R_{34} + R_{44}) / (n_2 + n_3 + n_4) = (308 + 277 + 370.5) / (12 + 12 + 12) = 26.542$$

$$y_{34} = (R_{34} + R_{44}) / (n_3 + n_4) = (277 + 370.5) / (12 + 12) = 26.979$$

$$y_{44} = R_{44} / n_4 = 370.5 / 12 = 30.875$$

③最大値 M_4 を求めます。

$$M_4 = \max \{y_{24}, y_{34}, y_{44}\} = 30.875$$

④統計量 U_{14} を求めます (1 から 4 群を使用しているから)。

$$U_{14} = R_{14} / n_1 = 220.5 / 12 = 18.375 \quad R_{14} \text{ は 1 群 (0mg/kg) の順位合計です。}$$

⑤分散を求めます。

$$N_4 = n_1 + n_2 + n_3 + n_4 = 12 + 12 + 12 + 12 = 48$$

$$V_4 = 1 / (N_4 - 1) \times \{ \sum \sum r_{ik}^2 - [N_4(N_4 + 1)^2 / 4] \} =$$

$$= 1 / (48 - 1) \times \{ 5.5^2 + 23.5^2 + \dots + 41.5^2 - [48 \times (48 + 1)^2 / 4] \} = 195.489$$

⑥検定統計量 t_4 を計算します。

$$t_4 = (M_4 - U_{14}) / \sqrt{V_4(1/n_4 + 1/n_1)} =$$

$$(30.875 - 18.375) / \sqrt{195.489(1/12 + 1/12)} = 2.190$$

⑦Williams 表: 上側 5% 点から

$$t_4 = 2.190 > \omega(4, \infty, 0.05) = 1.739 \text{ となるから, 「} \mu_4 \text{ は } \mu_1 \text{ より大きい} \text{」と判定します。}$$

⑧次に, p を $4 - 1 = 3$ として第 1 群と第 3 群までを比較します。そのために, 第 1 群から第 3 群を併せて小さいものから順位をつけ, 先の計算を繰り返します。

$$t_3 = (M_3 - U_{13}) / \sqrt{V_3(1/n_3 + 1/n_1)} = 1.092 \text{ を得ます。}$$

$$t_3 = 1.092 < \omega(3, \infty, 0.05) = 1.716 \text{ となるので, 有意差なしとして検定作業を終了します。}$$

1 群と 2 群の検定は行いません。

以上より, 40mg/kg 群と 0mg/kg 群間に有意差があるといえます。

§ 2. ラット波状肋骨出現率を Pharmaco Basic でやってみましょう。

Pharmaco 入力形式に並べ替えます

群	A1	A2	A3	A4
用量(mg/kg)	0 mg/kg	15 mg/kg	30 mg/kg	60mg/kg
1	10.5	12.5	25.0	87.5
2	0.0	12.5	20.0	22.2
3	0.0	11.1	12.5	20.0
4	0.0	0.0	12.5	14.3
5	0.0	0.0	0.0	12.3
6	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0

本メモは、オリジナルは「安全性評価研究会（谷学）」HP（<https://tanigaku.jp/>）に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。



60mg/kgのみ緑色で5%有意を示しています。30mg/kgは p=0.053 で有意とはなりません。したがって、その下の 15mg/kgは検定しません。

§ 3. まとめ

一般的に出現率であればカテゴリカルデータ解析で2xb分割表を使いますが、多重比較検定を要求されている場合でゼロが多いデータではノンパラメトリック検定を用います。ただ、Shirly-Williams 検定が搭載されているソフトは SAS と R そして Pharmaco ぐらいだと思います。

1) 永田靖、吉田道弘「統計的多重比較法の基礎」サイエンティスト社

【統計1口メモ 第5話】

＜イヌ試験 (N=4) でノンパラやってはダメ！＞

イヌやサルのように N=3 ないし N=4 例で Wilcoxon や Steel などのノンパラメトリック検定を何の疑問もなく使っていませんか？それは、大きな過ちです。もしかしたら有意差があるのに見逃しているかもしれません、あるいはその逆も。

下記の例のように3例で対照群と試験群の比較をしたときに、明らかに試験群全例が対照群よりも低値な場合、t検定では p=0.0326 で有意となりますが、ノンパラの Wilcoxon 検定では P=0.089 で有意とはなりません。このように順位が偏っている場合に起こっている現象なので、今までは気が付かなかったのかもしれませんが、手法としては間違っています。もしかして“3例”でやるからだと思っていないですか？それでは“4 例”でやってみましょう。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

<N=3 例>

t検定				Wilcoxon検定			
	対照群	試験群			対照群	試験群	
1	153	116		1	4	1	
2	163	124		2	6	2	
3	155	143		3	5	3	
n	3	3	6	n	3	3	6
合計	461	383	78	平均	5.0	2.0	-3.0
平均	153.7	127.7	26.0	se[d]	1.528		
残差平方和	8.67	384.67	393.33	u(z)	-1.746		
se[d]	8.097		98.33	p値	0.0809		
t	3.2112				正規近似		
p値	0.0326						

偏ったデータでは全て“有意差なし”になります。

<N=4 例>

t検定				Wilcoxon検定			
	対照群	試験群			対照群	試験群	
1	153	152		1	5	4	
2	163	148		2	7	1	
3	186	150		3	8	3	
4	155	149		4	6	2	
n	3	3	6	n	4	4	6
合計	461	383	78	平均	6.5	2.5	-3.0
平均	153.7	127.7	26	se[d]	1.732		
残差平方和	8.67	384.67	393.33	u(z)	-2.309		
se[d]	8.097		98.33	p値	0.0209		
t	3.2112				正規近似		
p値	0.0326						

n=3 例とは反対に偏ったデータでは全て“有意差あり”となります。

N=3例ではどんなに差があっても有意とならず、反対に N=4例では全く差がなくても有意となることがあるということです。このようなことでは、検定手法に使うことはできませんね。

統計ソフト JMP で解析すると、「標本サイズが小さいので、この近似は精度がよくありません。統計表を使って検定してください」と書かれています。このコメントは N=6 例で消えます。

実はノンパラメトリック検定は N=4 例で実施してはダメということは、1998年の応用統計学誌で永田靖先生*が報告されています¹⁾。その論文には『ノンパラメトリック法の使用に関しては、様々な誤解があるように思う。その中で「サンプルサイズがかなり小さい場合には（正規性のチェックが困難なので）ノンパラメトリック法を用いる」という誤解がある。これが誤解であるという理由は、ノンパラメトリック法の場合は、特に多重比較を行う場合には、データがいかなる値を取っても絶対に有意にならない検定を行うはめになりやすいからである』と書かれています。

そこには、表のような結果が示されていて、N=6 ならば OK。N=5 はマージナル(どっちつかず)とな

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

っています。*永田靖先生には Pharmaco Basic の多重比較検定 (Dunnett, Williams 検定について貴重なご意見をいただいています。

＜Steel 検定を $n=4$ 以下で使ってはいけない理由¹⁾＞

全部で a 群、第1群のサンプルサイズが n でスティールの方法を用いるとき、第1群と第 j 群の比較において、第1群の n 個のデータが第 j 群のデータの最小値よりも小さいとします (つまり、もっとも有意になりやすいパターンとします)。

$$R_{ij}=1+2+3+\cdots+n=n(n+1)/2$$

$$E(R_{ij})=n(2n+1)/2$$

$$V(R_{ij})=n^2(2n+1)/12 \quad (\text{タイ=同値がないとする})$$

となり、したがって、

$$t_{ij}=(R_{ij}-E(R_{ij}))/\sqrt{V(R_{ij})}=-n/\sqrt{(2n+1)/3}$$

となります。ここで、 n の値をいくつかとって t_{ij} の値を求めると次のようになります。

$n=3$ のとき $t_{ij}=-1.964$, $n=4$ のとき $t_{ij}=-2.309$, $n=5$ のとき $t_{ij}=-2.611$, $n=6$ のとき $t_{ij}=-2.882$, $n=7$ のとき $t_{ij}=-3.130$, $n=8$ のとき $t_{ij}=-3.361$

例として 4 群でスティールの方法の棄却限界値を求めると次のようになります。

	Steel の方法の棄却限界値(群数=4 群)					
例数	3	4	5	6	7	8
$d(4,0.5,0.05)$ (両側 5%)	2.212	2.349	2.442	2.511	2.567	2.613
$ t_{ij} $	1.964	2.309	2.611	2.882	3.130	3.361

これらの結果を見比べると、例えば、4 群のとき、Steel の方法(両側検定)では、4 例で t 値が 2.309 で棄却限界値は 2.349 (両側 5%) で決して有意にならないことが分かります。

今まで、イヌやサルの大動物のみならずラットでも $n=3$ 例や 4 例で Wilcoxon や Steel 検定を使っていたのは、今回の例題のように極端な事例がなかったために気が付かなかっただけは？ でも、間違った検定手法です。 $n=3, 4$ 例ではノンパラは使わずに、パラメトリックにしましょう。

1) 永田靖: 多重比較法の実践、応用統計学 27 (2) 93-108, 1998

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

【統計1口メモ 第6話】

<95%信頼区間をどのように使っていますか？>

最近の専門誌では、検定でp値と信頼区間を併記することが要求されることが多くなりました。市販統計ソフトではそれに対応して無条件で併記されますが、研究者は信頼区間をどのような場合に有効利用しているのでしょうか？ オッズ比のようにp値が表示されない場合は、信頼区間があれば「1」が入っているかどうかを見て有意差を判定できます。でも、p値が表示されていれば、わざわざ信頼区間で確認することはしないのでは？ いったい、95%信頼区間って何の役に立つのだろーと思っていませんか？まず、信頼区間の求め方から復習してみましょう。

【母平均の推定と信頼区間】

- 母集団から標本を取り出して計算した標本平均は、母平均の推定値として使うことができる。しかし、それは母平均にぴったり一致しているわけではなく、あくまでも推定値だ。そこで、母平均を、ある幅を持って推定しようということを考える。これを「区間推定」と呼び、その区間を「信頼区間」(Confidence Interval) という。

§ 1. 例題：新規バイオマーカーの血清 IL-X 濃度 (mg/dL) を SD ラット 10 匹で測定した。
本施設における IL-X 値の母平均 μ の 95%信頼区間を知りたい。

患者No.	1	2	3	4	5	6	7	8	9	10
IL-X	147	151	140	150	149	141	131	148	162	149

「標本から推定すると、母平均はこの値からこの値までの間にはいるのではないか」という形で推定する。

$$\text{信頼区間} = \text{平均値} \pm t_{n-1}(\alpha) \times \text{標準誤差}$$

IL-X 値の標本平均は 146.8、標準偏差は 8.2 だった。

$$10 \text{ 例 (自由度} = 9) \text{ の } t_9(0.025) = 2.26$$

$$\text{標準誤差} = \text{標準偏差} / \sqrt{n} = 8.2 / \sqrt{10} = 2.58$$

$$\text{信頼区間} = 146.8 \pm 2.26 \times 2.58$$

$$= 146.8 \pm 5.83$$

$$= 141.0 \sim 152.6$$

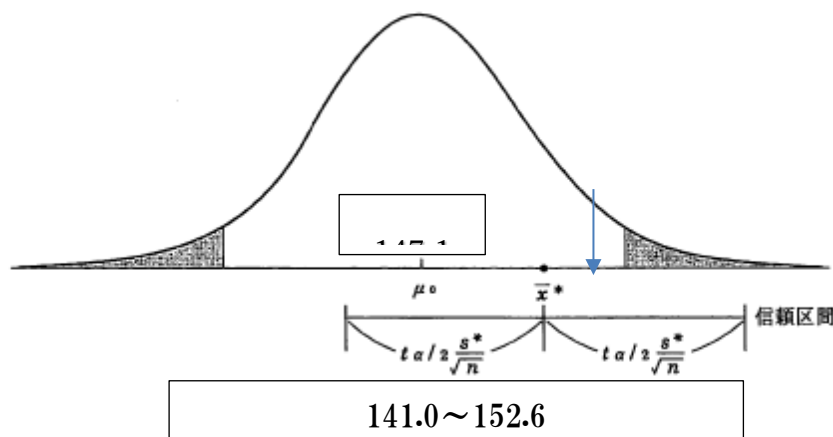
本施設で使用している SD ラットにおける IL-X 値の母平均の 95%信頼区間は 141.0～152.6 の間に含まれる。

結論. 新バイオマーカーの血清 IL-X 値は 141-153 mg/dl である。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

§ 2. 本施設内数値が学会基準値内に入っているか否かを知りたい。

- 学会基準値は平均値が 147.1 と報告されている。本施設内平均値=146.8
標準偏差=8.2、95%信頼区間=141.0~152.6 であった。学会基準値 147.1 は 95%信頼区間内にある。すなわち、学会基準値と本施設間では有意差がみられない。
- $\bar{X}^*$ が採択域にあるときは、 μ_0 は信頼区間に含まれる。



結論．本施設の数値は学会基準値内にある。

§ 3. 有意差がみられなかったときの 95%信頼区間の使い方

スネデカー・コ克蘭「統計的方法(第6版)」岩波書店より改

変。

- ある農場では、薬剤散布がトウモロコシの収穫量を増やすならば実施したいと考えている。散布費用は、1000 坪あたり 50kg の収穫量があれば採算があう。試験畑で薬剤散布と非散布の収穫量の差は 120 kg であった。散布による利益の推定値は $120 - 50 = 70\text{kg}$ となる。薬剤散布畑と非散布畑で収穫量の有意差検定をしたところ有意差がみられなかった。農場主は散布をするかしないかを選択しなければならない。
- そこで平均利益について 95%信頼区間を求めると 1000 坪あたり -25~160 kg であった。すなわち、薬剤散布で利益がでない場合は -25 kg とマイナスになるが、うまくいけば 160kg の利益がでる。その選択は農場主にまかせられるが、あなたならば散布しますか？

※ 1 エーカーを 1000 坪、1 ブッシュェルを 50 kg として略換算。

※ 95%信頼区間を p 値の代わりに使う傾向にある今日ですが、差があるかないかの p 値ではなく、実際の選択基準に具体的な数値として求めることができることに留意すべき。

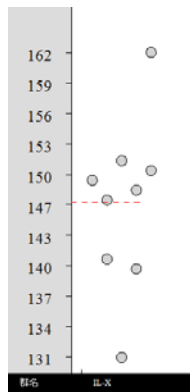
本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

<Pharmaco での解析>

Pharmaco 入力形式

IL-X
147
151
140
150
149
141
131
148
162
149

ビジュアル表示



解析結果

検定法	None
群名	IL-X
平均	146.8
標準偏差	8.162
N	10
信頼区間	140.96 ~ 152.64

§ 4. 95%信頼区間の扱いに統計家は2つに分かれている

- 広津氏は「信頼区間については、つい母平均が信頼区間に含まれる確率が 95%」といったくなるが、それは正しくない。あくまで母平均は未知数で、信頼区間あるいは信頼上下限が標本平均の関数として変化する確率変数なので、「信頼区間が母平均を含む確率が 95%」と表現しなければいけない。すなわち、実験を繰り返し、 $\alpha=0.05$ なら、100 回このような信頼区間を構成すれば 95 回は真値を含む信頼区間となっている。

広津千尋「実例で学ぶデータ科学推論の基礎」岩波書店 2018 年

- ザルツブルク氏も、「信頼区間はそのときそのときの結論ではなく、1つの時間的な仮定と見なすべきものである。長期間の試行で、95%信頼区間を毎回計算していけば、母数の真値がその信頼区間のなかにおさまる割合がその試行の 95%である。---たいていの科学者は 90%から 95%の信頼区間を計算し、まるで母数の真値がその区間に含まれていることを確信しているかのようである」。

D.サルツブルグ著「統計学を拓いた異才たち」日本経済新聞社

- Norleans 氏は平均値がその信頼区間の中にある確率が 95%であるということになる。95%の確率は、もし同じ試験を 100 回繰り返したら、100 回のうち 95 回は平均値が信頼区間の中にあるということの意味するのではないと述べている。

Mark x.Norleans「臨床試験のための統計的方法」サイエンティスト社

※以上のように Norleans 氏は広津氏やザルツブルク氏とは正反対の意見を述べている。それは、頻度論者とベイズ論者の違いで、前者は「同じ試験を 100 回繰り返したら、平均的に 95 回は真の平均値が信頼区間の中にある」と主張し、後者は「その信頼区間の中に真の平均値が存在する確率が 95%である」と主張している。実務家は時に応じて使い分けているのでは？

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

【統計一口メモ 第7話】

＜2x2分割表でイエーツ(Yate's)の修正をする?しない?＞

＜2 x 2 分割表＞

	有効	無効	計
Cisplatin	4	11	15
Ascle	7	3	10
計	11	14	25

計数値（カテゴリカルデータ）の解析で最初に出会うのが＜2 x 2 分割表＞です。「薬効の有無」や「疾患＋－」を2群で比較する場合に用いられます。

＜創薬担当者二人の主張＞

安全性研究所のA君は上の2 x 2 分割表のデータをGLP統計ソフトで解析して有意差なし（ $p=0.0841$ ）と報告しました。ところが本社の開発部B君から「なに言ってんだ $p=0.0325$ で有意差あるじゃないか！」とクレームが入りました。両者とも再計算して、結果に間違いはありません。いったいどうなってるのでしょうか？

実は、この結果はどちらも正しい結果を示しています。使っているソフトの問題です。A君の使っているGLPソフトは“イエーツの修正”が入っていて、B君の使っているJMPソフトは入っていません。イエーツの修正が入るか入らないかでこんなに違うのですが、実は統計家の間でも論議が分かれています^{1),2)}。ちなみに、医薬統計の教科書的な吉村功編集「毒性・薬効データの統計解析」³⁾はイエーツの修正をした手法を採用していますが、JMPに採用されているPearsonの方法はイエーツの修正はされていません。結果は次のようになります。

イエーツの修正あり

$$\chi^2 = 2.98 = |(4 \times 3 - 11 \times 7) - 25/2| \sqrt{2 \times 25 / (15 \times 10 \times 11 \times 14)}$$

$$\chi^2 (0.05) = 3.84 = \text{CHINV}(0.05, 1)$$

$$p \text{ 値} = 0.0841 = \text{CHDIST}(2.98, 1)$$

イエーツの修正なし

$$\chi^2 = 4.57 = |(4 \times 3 - 11 \times 7)| \sqrt{2 \times 25 / (15 \times 10 \times 11 \times 14)}$$

$$\chi^2 (0.05) = 3.84 = \text{CHINV}(0.05, 1)$$

$$p \text{ 値} = 0.0325 = \text{CHDIST}(4.57, 1)$$

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

<Pharmaco アウトプット>

検査項目	2x2
検定法	χ^2
群名	Cisplatin
N	15
Data	4 / 11
χ^2 値	0
群名	Ascle
N	10
Data	7 / 3
χ^2 値	4.573
P値	0.03249 *

<JMP アウトプット>

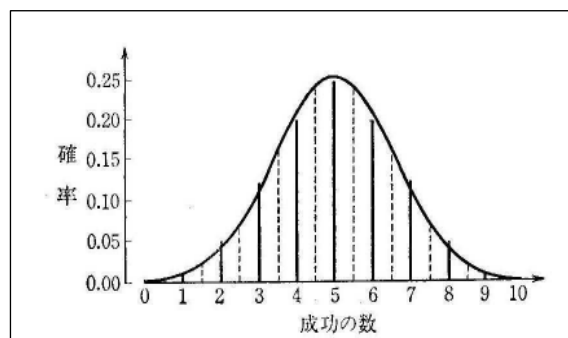
検定			
要因	自由度	(-1)*対数尤度	R2乗(U)
モデル	1	2.340874	0.1365
誤差	23	14.807371	
全体(修正済み)	24	17.148245	
N	25		
検定	カイ2乗	p値(Prob>ChiSq)	
尤度比	4.682	0.0305	
Pearson	4.573	0.0325	
Fisherの正確検定	確率	代替の仮説値	
左	0.0416	薬剤=旧法の場合のp値(効:	
右	0.9952	薬剤=新法の場合のp値(効:	
両側検定	0.0486	確率(効果=無効)は、薬剤の	

<イエーツの修正って何？>

- χ^2 検定は χ^2 の分布を連続的な分布に近似しているので誤差が生まれます（後述）。その誤差を修正するのが、「イエーツの修正」です。
- イエーツの修正をかけると、保守的な（有意差が出にくい）傾向があり，“データに5以下が含まれている場合のみに用いる”という意見もあります。
- しかし、そのような場合は一般的に Fisher の直接確率計算（正確検定）を用いるので、イエーツの修正を行ってまで χ^2 検定にこだわることはありません。
- 先に述べたように、イエーツの修正をするべきか否かは、現在、統計家の中でも意見が分かれています。
- したがって、どちらがいいとか悪いとかの問題ではありません。もし、使用するソフトを選べないならば実務家には選択の余地はないと思ってください。ソフトの開発者である統計家の考え方に従うしかありません。
- ちなみに、JMP ソフトは修正をしていません（後述）。Pharmaco は JMP の解法に従っているのも同じように修正をかけていません。

<もう少し詳しく解説すると一正規近似と連続修正>

- 下図の“縦の実線”は $n=10$ 確率 $p=0.5$ のときの r （成功の数）の2項分布を示しています。
- “曲線”は平均 $np=(10 \times 0.5)=5$ で標準偏差 $\sqrt{npq}=\sqrt{10 \times 0.5 \times (1-0.5)}=1.581$ の正規曲線です。



本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

- 2 項分布は $r = 0, 1, 2, \dots, 10$ という値でしか確率をもたない離散型であるのに対し、正規分布は $-\infty$ から ∞ までのどの区間においても連続した確率をもちます。
- $r = 8$ の確率を推定するのに、正規曲線のどの部分を近似として用いればいいのでしょうか？
- 2 項分布とは正規分布をその単位級間隔ごとの組み分けしたものと考えます。すなわち、 $r = 8$ の場合は、 $7.5 \sim 8.5$ までの点線内の面積に対応します。
- そこで、8 回以上の確率を計算したいときに、その正確な確率は $r = 8, 9, 10$ に対する 2 項確率を加えることによって得られます。
- $r = 8, 9, 10$ それぞれの確率とその合計は

$$0.0495 + 0.0097 + 0.00098 = 0.0547$$

- 正規曲線下の対応する面積は $7.5 \sim \infty$ までであって、 $8 \sim \infty$ までの面積ではありません。
- したがって、

$$\text{正規偏差は} \quad z = (7.5 - 5) / 1.581 = 1.58127$$

$$\text{近似的な確率は} \quad p = 1 - \text{NORMSDIST}(1.581) = 0.0569$$

※正規偏差 z 検定と χ^2 検定は同じもの。

- これを 7.5 ではなく 8 でやると

$$z = (8 - 5) / 1.581 = 1.89753$$

$$p = 1 - \text{NORMSDIST}(1.8975) = 0.0288$$

- 以上の結果、 8 より 7.5 の方がより真実 (0.0547) に近いことから、一般原則として $(r - np)$ の絶対値から $1/2$ だけ引くことにします。
- この $1/2$ 減算を＜連続性の修正＞と呼びます。
- 大きさ n の標本で r 個の成功を得る確率を一般式であらわすと、

$$p_r = {}_n C_r p^r (1-p)^{n-r} \quad {}_n C_r = n! / (r!(n-r)!) = \text{二項係数}$$

$n = 10, p = 0.5, r = 8$ のときの確率は

$$Pr = 10! / 8!(10-8)! \times 0.5^8 \times (1-0.5)^{(10-8)} = 45 \times 0.00391 \times 0.25 = 0.04395$$

$n = 10, p = 0.5, r = 9$ のときの確率は $= 0.009765$

$n = 10, p = 0.5, r = 10$ のときの確率は $= 0.000976$

$$\text{合計} = 0.05469$$

スネデカー・コ克蘭「統計的方法」原書第 6 版 岩波書店 1977 年より抜粋

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

＜エクセルを用いた二項分布の計算法＞

大きさ n の標本で、事象 E の起こる確率を p とするとき、そのうち x 個に E が起こる確率は $\text{BINOMDIST}(r, n, p, \text{false})$ で計算できます。

$n=10, r=8, p=0.5$ のときは $\text{BINOMDIST}(8, 10, 0.5, \text{false}) = 0.04395$

＜何で JMP はイエーツの修正をかけない？＞

イエーツの修正をかけない代表に JMP の考え方を芳賀敏郎先生が「医薬品開発のための統計解析」第3部⁴⁾で詳しく解説しています。簡単に言うと「2項分布の正確な確率を用いた実質 α は、第1種の誤りの確率である名目 α に比べて低くなるため、平均をとる手法 (mid- p) が好ましい。それには連続修正を加えない正規近似がよい」ということになります。それでも、「第1種の誤りの確率を厳格に保証しなければならない場合（保守的な立場）には、2項分布の正確な確率計算をする（イエーツの修正をかける）」という言葉も心にとめておかねば。

- 1) 丹後俊郎、医学への統計学（第3版）、朝倉書店、（2013年）
- 2) 永田 靖、統計的方法のしくみ、日科技連、（1996年）
- 3) 吉村 功、毒性・薬効データの統計解析、サイエンティスト社、（1987年）
- 4) 芳賀敏郎、医薬品開発のための統計解析 第3部、サイエンティスト社、（2016年）

以上

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

【統計一口メモ 第8話】

<Fisher 正確検定は片側？それとも両側？>

A 君の質問です。「学会誌に論文を投稿したら、査読者（レフリー）から「セル内に5以下の数値があるので χ^2 検定ではなく、Fisher 正確検定で求め、

ました。実施した χ^2 検定は両側検定なので Fisher 正確検定
両側検定をしようと思っているのですが、教科書¹⁾には片側

	有効	無効	計
Cisplatin	4	11	15
Ascle	7	3	10
計	11	14	25

しか出ていませんが、JMP のアウトプットでは片側と両側が出ています。でも、両側は片側の2倍ではありません、どうしてなのでしょう？」という内容でした。

R.A.Fisher の正確検定は本来片側検定として提案されたものですが、両側検定だと思い込んでいる人も少なからずおられます。その両側検定の求め方が統計家（統計ソフト）によって異なるので

要注意です。そもそも「Fisher 正確検定って何？」から紐解いていきましょう。

<目的> 2 x 2 分割表で4つのマス目（セル）内に5以下の数値がある場合に、2群間の出現率に差があるかどうかを検出する。

<定義> 4つのセル内の数値で求められる確率は、次の式で定義される。

$$p = \frac{{}_e C_a \times {}_f C_c}{{}_n C_g} = \frac{e! f! g! h!}{n! a! b! c! d!}$$

要因 A	要因 B		合計
	B ₁	B ₂	
A ₁	a	b	e
A ₂	c	d	f
合計	g	h	n

${}_n C_g$ は n 個から g 個を選ぶ組み合わせ（Combination）の数で

$${}_n C_g = \frac{n!}{g!(n-g)!}$$

として求められ、2項係数と呼ばれます。ちなみに Excel では n! は=FACT(n)で、2項係数 ${}_n C_g$ は

=COMBIN(n, g)で求めることができます。

上の投稿論文の例で解析してみましょう。

手順1. Cisplatin 投与群 15 例から有効 4 例の取り出し方は ${}_{15}C_4 = 1365$ 通り

Ascle 投与群 10 例から有効 7 例の取り出し方は ${}_{10}C_7 = 120$ 通り

Cisplatin 投与 15 例から 4 例、Ascle 10 例から 7 例の取り出し方は、 ${}_{15}C_4 \times {}_{10}C_7 = 1365 \times 120 = 163800$ 通りある。

手順2. 全体の人数 25 例から有効症例 4 + 7 = 11 例の取り出し方は ${}_{25}C_{11} = 4457400$ 通りある。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

手順3. 2 x 2 分割表の生起確率 p は

4	11
7	3

P=0.0367

$$\frac{{}^{15}C_4 \times {}^{10}C_7}{{}^{25}C_{11}} = \frac{163800}{4457400} = 0.036748 \quad \text{である。}$$

手順4. 分割表でセル内の最小値がゼロになるまでの確率を求め、それを合計する。

3	12
8	2

P=0.0046

2	13
9	1

P=0.0002

1	14
10	0

P=0.0000

$p = 0.0367 + 0.00046 + 0.0002 + 0.0000 = 0.0416$

片側検定で p 値は 0.0416 となる。

< JMP での両側確率の求め方 >

JMP では、両側検定を反対側にゼロになるセルを作成して片側で求めた P 値に加えます。

4	11
7	3

P=0.0367

5	10
6	4

P=0.1415

途中略

9	6
2	8

P=0.0505

10	5
1	9

P=0.0067

11	4
0	10

P=0.0003

$P = 0.0070 + 0.0505 = 0.0575$ $P = 0.0067 + 0.0003 = 0.0070$

両側検定の p 値は片側検定の p 値 0.0416 より小さい累積 p 値(0.0070)を足した $0.0070 + 0.0416 = 0.0486$ となります。なお、0.0575 は 0.0416 より大きいので、より小さな 0.0070 が採用されます。

	有効	無効				
シスプラチン	4	11				
アスクル	7	3				
f11	f12	f21	f22	p	pL	pU
1	14	10	0	0.0000	0.0000	1.0000
2	13	9	1	0.0002	0.0002	1.0000
3	12	8	2	0.0046	0.0048	0.9998
4	11	7	3	0.0367	0.0416	0.9952
5	10	6	4	0.1415	0.1831	0.9584
6	9	5	5	0.2830	0.4660	0.8169
7	8	4	6	0.3032	0.7692	0.5340
8	7	3	7	0.1732	0.9424	0.2308
9	6	2	8	0.0505	0.9930	0.0576
10	5	1	9	0.0067	0.9997	0.0070
11	4	0	10	0.0003	1.0000	0.0003
					両側検定	0.0486

JMP では次のように表示されます。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

検定	カイ2乗	p値(Prob>ChiSq)
尤度比	4.682	0.0305*
Pearson	4.573	0.0325*

Fisherの正確検定	p値	対立仮説
左片側検定	0.9952	処方o=旧処方の場合の確率(効果o=有効)は、処方o=新処方の場合よりも大きい
右片側検定	0.0416*	処方o=新処方の場合の確率(効果o=有効)は、処方o=旧処方の場合よりも大きい
両側検定	0.0486*	「効果o=有効」である確率は、「処方o」の水準間で異なる

<もう一つの両側検定>

丹後俊郎氏は次のように述べています²⁾。「両側検定の場合のp値の計算方法はいくつか考えられているが、推奨する方法は“片側p値を単純に2倍”することである」。

また、竹森幸一氏は論文の中で「片側確率はいずれの解説書でも同じであるが、両側確率は次の2つの型にわかれる」と述べています³⁾。

1. 片側確率に観察パターンの起こる確率と同等あるいはそれより小さい確率を全て加えた場合

2. 片側確率を2倍したもの。

この1の方法は JMP の方法で、2. は前述の丹後氏が述べている方法、Pharmaco の方法です。

< JMP と Pharmaco の 両 側 確 率 の 求 め 方 の 違 い >

	Cisplatin	Ascle		Cisplatin	Ascle		Cisplatin	Ascle
有効	3	0	有効	3	1	有効	3	2
無効	1	6	無効	1	5	無効	1	4

	JMP	Pharmaco		JMP	Pharmaco		JMP	Pharmaco
片側検定	0.0333	0.0333	0.1190	0.1190	0.2619	0.2619		
両側検定	0.0333	0.0667	0.1905	0.2381	0.5238	0.5238		

片側=両側 片側x2倍=両側

JMP の両側確率の求め方は、左の例では片側=両側で、真中の例では片側確率+ α で

右の例では片側x2倍となり、データによって求め方が異なり、一定の方式をとっていません。

したがって、Pharmaco は両側=片側x2倍方式を採用しています。なお、右の例は有効と無効の合計が5で等しくなっています。そのようなケースでは、JMP も Pharmaco も結果は同じになります。

統計一口メモ第7話の2x2分割表の χ^2 検定もイエーツの補正をするソフトとしないソフトがあり、結論が異なることを述べました。また、今回も Fisher 正確検定で両側確率を求める場合はソフトによって結論が異なることがあることにふれました。使用しているソフトがどちらを選択しているのか知っておく必要がありますね。なお、Fisher 正確検定は Fisher の直接確率検定とも呼ばれています。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

<文献>

- 1) 吉村功編著「毒性・薬効データの統計解析」サイエンティスト社（2005 年）
- 2) 丹後俊郎「医学への統計学」第 3 版 朝倉書店（2013 年）
- 3) 竹森幸一「わが国の統計解説書にみられる Fisher の直接確率法の両側確率と片側確率をめぐる混乱」弘前医療福祉大学紀要 5(1), 77-81, 2014

以上

【統計一口メモ 第 9 話】

<両側検定と片側検定 使い分けていますか？>

第 8 話で Fisher 正確検定の両側検定について、統計ソフトによって手法はまちまちであることを述べました。今回はさらに、片側検定の使い方について統計家のニュアンスが異なることと検証的な臨床試験と探索的な前臨床試験では有意水準が異なることについても触れてみようと思います。

<帰無仮説と対立仮説>

2 種の薬剤のうち A 剤は標準品で B 剤は新薬。in vitro では B 剤の有用性が確認されています。

仮説には「2 つの薬剤に差がある」と 「A 剤より B 剤の新薬の方が優れている」が考えられます。

帰無仮説は、 “差がない” $H_0: \mu_1 = \mu_2$

対立仮説は、 “差がある” $H_1: \mu_1 \neq \mu_2$

“優れている” $H_1: \mu_1 < \mu_2$ or $\mu_1 > \mu_2$

対立仮説は前者を選ぶと両側検定、後者は片側検定となります。

<片側検定を行う際の帰無仮説の取り方>

対立仮説を $H_1: \mu_1 < \mu_2$ に設定するとします。このとき、帰無仮説を $H_0: \mu_0 \geq \mu_1$ としてはダメ！

帰無仮説は便宜的な“仮説”に過ぎず、成り立っているかどうかを知りたい仮説ではありません。成り立っているかどうかを知りたいのは対立仮説の方です。したがって、 $H_0: \mu_1 = \mu_2$ としなければなりません。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

＜片側検定の方で有意になったのでこちらを採用しました！＞ ※影の声「オイオイ」

永田先生は「統計量が 1.80 でした。両側検定では $1.80 < 1.960$ で“有意差なし”でした。でも、片側検定では $1.645 < 1.80$ で“有意差がみられた”ので、こちらを採用しました！」という報告に

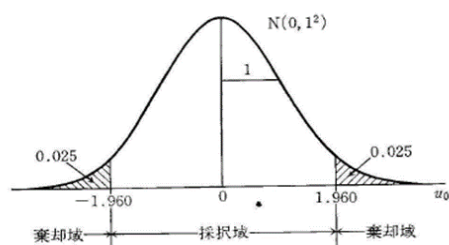
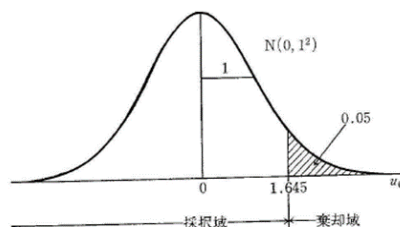


図 3.1 F

両側検定



片側検定

「これは、データを見たあとに仮説や棄却域を自分の都合のよいように設定すること、いわば、“反則行為”です。仮説や棄却域の設定は、データを取る前に行わなければならない¹⁾」と述べています。

※論文執筆者の行為は恣意的と言われても仕方ありません。片側検定をするかしないかは、永田先生の言われるようにデータを取る前、プロトコルを書く段階で決めておかなければいけません。そのことについて、多くの統計家が著書の中で、いろいろな表現で述べています。それを紹介しながら片側検定の使い方について学んでいきましょう。私の考えは※で述べていきます。

＜片側検定は稀！いつも両側を＞

Dr. Altman は「片側検定が適切なことは稀である。事前に強い期待、例えば新治療は旧治療より悪いはずがないという期待をもっているときでさえも、それが正しいことを確信することはできない。確実にあれば実験する必要がない！片側検定が本当に適切であると思うのであれば、その決定はデータを解析する前になされなければならない；結果がどうであつたかによるべきではない。いつも両側を使うことを勧める」と述べています²⁾。

※適切であるかどうかをどうやって判断するのでしょうか。解析前に対立仮説で $\mu_1 < \mu_2$ を設定し、それを検定することで判断するものと思うのですが、Dr. Altman がレフリーの論文では片側検定はボツになりますね。

＜両側検定を用いる方が問題になることが少ない＞

丹後先生は「一般に、一方の群が、他方の群に比べて、大小関係が予想できる場合に片側検定を用いることができる。しかし、大小関係が予測できたとしてもその関係が、研究者間で認められていない場合には片側検定を用いるのは望ましくなく、さらに、それは、両側検定に比べて、有意でない差を有意としてしまう可能性が高くなるので、両側検定を用いる方が問題になることは少なくすむ」³⁾。

※「研究者間で認められる」ということってどういう場合なんだろう？あたらしい事象に関

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

しては研究者は自分だけかもしれない。「問題になる」のを避けるために両側検定にするという行為は多くの統計家が取っている。

<両側検定としておくのが安全>

Dr. Armitage は「両側検定よりも片側検定を用いたいという誘惑が生じるかもしれない。

データをながめて、仮説からのずれの方向を見定めてから片側検定の使用を決定することは決して許されるべきものではない。仮説からのある方向への乖離は常に偶然にしか生じず、従ってどんなにずれが大きくてもそれは意味のないものであると見なせることが確実な場合に限って片側検定を用いることを決めてよい。しかも、その決定はデータを調べる前に行うのである。このような状況は現実には稀であり、有意性検定はほとんどいつも両側検定としておくのが安全であろう」⁴⁾

※対立仮説の $H_1: \mu_1 < \mu_2$ をよほど稀な場合にしかとってはならぬという主張。血沈のような上にあがることのないようなデータしか使えないということだろうか？ここでも両側検定に安全の地を求めている。

<あえて両側検定にしておくのが賢明>

佐藤、小西先生は「片側検定か両側検定か？というのも、どうもよく分かりにくいところです。あまり片側検定を使っている例は見かけません。どうしてでしょう。ある薬の効果を調べるためにプラシーボとの比較を行う際、プラシーボの方が実際の薬より優れた効果があると仮定することはないでしょう。したがってこの場合、理屈で言えば片側検定を使うのが適当だと考えるのが自然です。でも、片側検定による報告がされていることは稀です。その理由の1つには、片側検定では「有意差が出やすくなる」ということで、厳しい査読委員に好まれない傾向があると言えるでしょう。したがって、結論としては、両者の違いを理論的によく理解したうえで、あえて両側検定を使用しておくのが賢明かと思います」⁶⁾。

※理屈で言えば片側検定を使うのが適当と言いながら、最終的には両側検定を使うのが賢明とレフリーに付度しているのはいただけません。実務者は再試験ができない緊迫した状態にある場合に付度している余裕はありません。レフリーに結果を見る前から片側を設定したことをしっかり述べるべきです。

<片側検定は事前に決めておけばよい>

浜田先生は「両側検定では対立仮説の方向を欲張るため片側検定に比べて、p 値が大きくなり意になり難くなります。多くの場合、両側検定の p 値は片側検定のその 2 倍となります。通常何も断らずに検定といえば、両側検定を意味します。両側検定では有意にならず、片側検定では有意となる微妙なケースでは、事後的に両側検定を片側検定に変えたい

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

誘惑にかられますが、片側検定と両側検定のどちらを用いるかは、検定を行う前に決めておく必要があります」⁵⁾

※検定を行う前に「片側検定を使う」と宣言しておけば、それで OK という実務者の立場を理解した記述です。

<芳賀先生の私見>

厚労省の「承認申請のための試験では、片側検定の第1種の誤りを両側検定で慣例的に用いている値の半分に設定する方法が好ましい」という提言に対して、「これは“企業は自社に有利になるように両側・片側検定を設定する危険がある“という性悪説に基づいており、企業の統計担当者を無能力扱いしているように思われる。統計担当者が社内で権威ある立場を維持できるようになり、妥当な指針が出されるようになることが期待される」⁷⁾と述べています。

※芳賀先生は片側検定を対立仮説に基づいて使用すべきであることを主張されていました。もちろん、

後付けのような使い方ではなく、検定前に宣言しておくことは当たり前として、当局やレフリーの顔を伺うような卑屈な態度はとるべきではないと。

< χ^2 検定は片側か両側か>

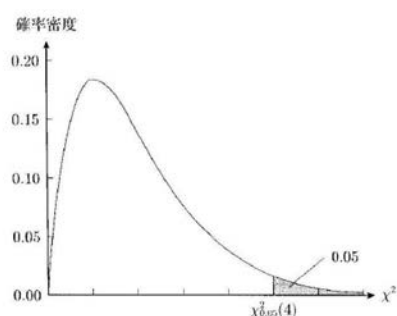


図 5 χ^2 分布の密度関数(自由度 4)

「 X^2 の実現値がこの分布の上側確率 0.05 の点より上の値であれば、仮説が正しければ頻繁には起こらないことが生じているとして、帰無仮説 H_0 を有意水準 0.05 で棄却する」⁸⁾。

JMP アウトプット

検定			
要因	自由度	(-1)*対数尤度	R2乗(U)
モデル	1	2.340874	0.1365
誤差	23	14.807371	
全体(修正済み)	24	17.148245	
N	25		
検定	カイ2乗	p値(Prob>ChiSq)	
尤度比	4.682	0.0305	
Pearson	4.573	0.0325	

JMP の χ^2 検定は「Pearson」として出力される。

Pearson の方法は2項分布の正規近似を用いる両側検定であり、連続修正は加えられていない。標準正規分布を2乗したものが自由度1の χ^2 分布であり、標準正規分布で、正規分布の両側確率が χ^2 分布の上側確率になる。

※ χ^2 検定は χ^2 分布でみると上側確率 0.05 の片側検定です。しかし、通常正規近似で計算され JMP のように Pearson として両側検定で表記されます。どちらにしても、確率 0.05 には変わりはありません。でも、 χ^2 検定として表記するのであれば、片側検定ということになります。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

- 1) 永田靖「統計的方法のしくみ」日科技連 1997 年
- 2) D.G. Altman「医学研究における実用統計学」サイエンティスト社 1999 年
- 3) 丹後俊郎「医学への統計学」第 3 版 朝倉書店 2013 年
- 4) P. Armitage, G. Berry「医学研究のための統計的方法」サイエンティスト社 2001 年
- 5) 浜田知久馬「学会・論文発表のための統計学」真興交易医書出版部 1999 年
- 6) 佐藤敏彦、小西宏明「StatView4.0 日本語版」BNN1994 年
- 7) 芳賀敏郎「医薬品開発のための統計解析」第 1 部 サイエンティスト社 2016 年
- 8) 広津千尋「データ科学推論の基礎」岩波書店 2018 年

【統計 10 口メモ 第 10 話】

＜なぜ 3 群以上の比較に t 検定を使ってはダメなんだろう？＞

A さんの疑問：専門誌に对照群と各群で、2 群間の比較をして投稿したら「t 検定の繰り返しはダメ」とレフリーコメントが返ってきました。なぜ、ダメなのかがわかりません。

新抗癌剤 Ascle の試験のサイトカイン IL-X 値 (t 検定)				
群	对照群	10mg 群	30mg 群	100mg 群
例数	10	10	10	10
平均値	928.3	911.7	891.9	891.2
標準偏差	27.9	21.2	41.9 **	42.9**

§ 1. t 検定とは？

t 検定の目的は「平均値の差」が「誤差」の何倍あるかを知ること。

まず t 統計量を下記の式から計算します。

$$t = \frac{|\text{平均値の差}|}{\text{平均値の差の標準誤差}} = \frac{|x - y|}{\sqrt{(1/n_1 + 1/n_2) \times V}}$$
$$V = (S_1 + S_2) / (\nu_1 + \nu_2)$$

<0mg/kg vs 100mg/kg>

V の計算が「平方和の和」を「自由度の和」で割ることに注意です。

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

t検定	0 mg/kg	10 mg/kg
1	890	878
2	949	910
3	870	916
4	920	928
5	953	876
6	948	936
7	925	908
8	941	923
9	942	935
10	945	907
n	10	10
平均	928.3	911.7
平方和 S	7000.10	3994.10
自由度 v	9	9

エクセル関数

$S=DEVSQ(890:945) = 7000.1$

$P(\text{片側})=TDIST(ABS(1.502),18,1)=0.075$

$P(\text{両側})=TDIST(ABS(1.502),18,2)=0.150$

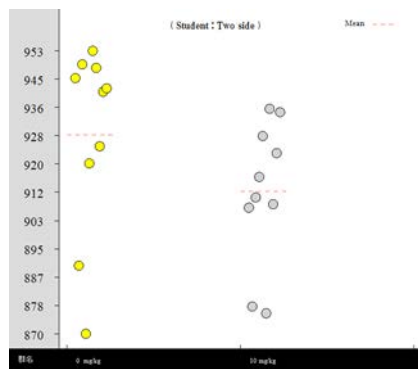
平均値の差の有意差検定

t値	1.502
p値(片側)	0.07523
p値(両側)	0.15046

上の計算式とエクセル関数の使用で P 値まで出す方法がわかりました。

簡単にいうと分子の「平均値の差」を分母の「標準誤差の和」で割るということ。大事なことは、この分母は 2 群しか足していない ことです。論文のように多群の場合は、分子は 2 群の差で、分母は全部の標準誤差が足されます。すなわち数が大きくなり t 値が小さくなります。有意差が出にくくなります。I

こ表示されます。



検定法	Student (両側)
群名	0 mg/kg
平均	928.3
標準偏差	27.889
N	10
群名	10 mg/kg
平均	911.7
標準偏差	21.066
N	10
P値	0.1505
平均値の差	-16.6
信頼区間	-39.82~6.62

§ 2. 質問への回答

<回答その1>：多群で t 検定を使って、対照群と投与群を比較すると、各検定は独立ではなくなります。つまり、対照群が共通ですから、これが変わるとそれに伴って、全部の投与群が影響を受けます。それが Dunnett 氏が注意した点です。

<回答その2>：もう少し、確率論的立場から解説します。2 群の比較において、独立な 10 回の比較をくり返すと、“少なくとも 1 回”は 5% 有意で差ありと出る確率はいくらでしょうか？

Brown 氏の例が、よく目にします。それは次のような例です。

「あるレストランの地下にはたくさんのワインが貯蔵されていますが、20 本に 1 本、すなわち 5% の割合で、酔っぱくなった不良品が混入している。客がやってきて 1 本のワインを注文するとき、ソムリエは長い目で見て 20 回に 1 回は冷汗をかいて謝ることになる。－

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

方、パーティには 10 本単位でワインをだすことになっているが、1 本でも酔っぱいのが混ざっていると、パーティはパーになる。

1 本あたり 5% で不良品というときにはパーティあたり 40% およそ 5 回に 2 回の割合で謝るしまつになる (Brown)。く佐久間昭：医学統計 Q&A、金原出版)」

酸っぱいワインが 5% もあること自体は現実的ではないのですが、そこは我慢して、5% という数字にこだわしましょう。さて、ここで、なぜパーティあたり 40% にもなるのか考えてみましょう。

まず、確率の復習からはいります。サイコロを 6 回ふって、「少なくとも 1 回は 1 がでる確率」を知りたいものとします。まともに計算するならば、1 が 1 回だけ出る確率、1 が 2 回出る確率-----中略-----1 が 6 回出る確率のすべてを計算して加え合わせなければなりません。

この場合には「少なくとも 1 回は 1 が出る」ことが起こらない確率、すなわち「1 回も 1 が出ない確率」をまず求めて、それを 1 から差し引く方がずっと楽です。

1 回ごとに 1 が出ない確率は $1 - 1/6$ だから、6 回とも 1 が出ない確率はそれを 6 回かけ合わせればいい。したがって、「少なくとも 1 回は 1 が出る確率」は、

$$1 - (1 - 1/6)^6 = 0.665 = 66.5\%$$

になります。そこで、先ほどのワインの問題ですが、10 本単位で出したときに、パーティあたり

$$1 - (1 - 0.05)^{10} = 0.4013$$

約 40% に不良品が入る可能性があるということになります。

ここで、本題に入りましょう。A, B, C の 3 群間の比較を行うときに、各 2 群間の組み合わせで t 検定をしようとすれば A vs B, A vs C, B vs C の 3 通りの t 検定を行うことになります。

各 t 検定の危険率を 5%、すなわち「実際には差がないのに、たまたま差があるように判定される確率が 5% ずつある」とすると「少なくとも 1 つの組み合わせがたまたま有意差あり」と判定される確率は 1 から 3 つの組み合わせすべてが「有意差なし」となる確率を引いたものだから、

$$1 - (1 - 0.05)^3 = 1 - 0.953 = 0.14$$

となり、この検定全体では危険率が 14% となってしまいます。5% の危険率で検定しているつもりが、14% の危険率、すなわち過剰評価していることになり、レフリーはそれを指摘するわけです。

群の数が増えれば増えるほど、全体の危険率は 5% よりはるかに大きくなるわけです。実際には差がないのに、たまたま有意差ありとされる確率が増えてしまいます。

以上の理由で、多処理を同時的に試験して、つぎつぎに 2 群を取り出しては 5% の危険率で 2 標本の t 検定を繰り返してはいけないということになります。

※僕のつぶやき：回答 1 で Dunnett さんが言ったのは、対照群、A, B, C とあったときに、

本メモは、オリジナルは「安全性評価研究会（谷学）」HP (<https://tanigaku.jp/>) に継続掲載されているものです。谷学 HP 委員会のご厚意により「臨床評価」HP に転載させていただきました。今後も数話まとめてこの HP に転載させていただく予定です。

対照群が繰り返し使うことになるからダメと言ったのは理解できけど、回答2で確率論的に解釈したのとは、少し違う気がするんだけど。対照群が違えば、その2群の検定は独立しているように見えるのもダメなんだ。t検定を並べてやることだってあるんだけど、その都度 Bonferroni 補正して α を $1/k$ (回数)で割らなければいけないと統計家はおっしゃる。

とりあえずレフリーコメントに応えられるように多重比較をしっかり勉強しておかねば。

以上