2017年12月16日（土）
日本知能情報ファジィ学会東海支部勉強会資料

ノンパラメトリック検定

古橋武

本書 では Microsoft Excel 2013 を 使用 し て い ます．Microsoft, Excel 2013 は，米国 Microsoft Corporation の 各国 における 登録商標 または 商標 です．

本稿は，日本知能情報ファジィ学会東海支部勉強会（2017 年 12 月 16 日開催）の配布資料に加筆したものです．資料中のエクセルファイルは本稿掲載の Web ページからダウンロードできます．

ノンパラメトリック検定とは？

母集団分布に正規分布などの特定の分布を仮定しないで統計的検定を行う方法.

本稿ではノンパラメトリック検定について解説します.

対応表

ノンパラメトリック検定	パラメトリック検定
スピアマンの相関係数	ピアソンの相関係数
マン・ホイットニーのU検定 ウィルコクソンの順位和検定	ステューデントのt検定
ウィルコクソンの符号付き順位和検定	対応のあるステューデントのt検定
クルスカル・ワリスの方法	一元配置分散分析
スチール・ドゥワスの方法	テューキーの方法
スチールの方法	ダネットの方法
シャーリーウィリアムズの方法	ウィリアムズの方法

3

ノンパラメトリック検定とパラメトリック検定の対応表です．表の左がノンパラメトリック検定の手法であり，右が対応するパラメトリック検定の手法です．例えば，通常よく使われる相関係数は，パラメトリック検定のピアソンの相関係数です．ノンパラメトリック検定における相関係数はスピアマンの相関係数と呼ばれます．

1. 中心極限定理(すべてに共通する重要概念)

母集団の分布がどのような分布であっても，標本平均 $\bar{x}$ はサンプルサイズ n を増やしたとき近似的に正規分布に従う.

サンプル $x_1, x_2, \cdot\cdot, x_n$

母平均: μ

母分散: σ^2

標本平均: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ➡ 正規分布に従う.

規格化: $\bar{z} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ ➡ $\bar{z}$: 標準正規分布に従う.

まず，中心極限定理です．これは以降のすべての検定において，サンプルサイズ n が大きい場合に共通して適用可能な重要概念です．中心極限定理とは，

「母集団の分布がどのような分布であっても，標本平均 $\bar{x}$ はサンプルサイズ n を増やしたとき近似的に正規分布に従う.」

というものです．いま n 個のサンプル $x_1, x_2, \cdot\cdot, x_n$ が与えられたとします．母平均をμ，

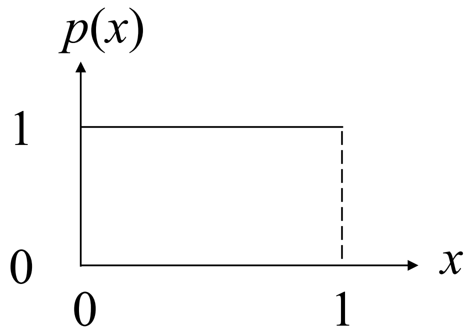
母分散をσ^2 とすると，サンプルサイズ n が大きい場合，中心極限定理により標本平均 $\bar{x}$ は平均 μ，分散 σ^2/n の正規分布に従います.

また，規格化された平均値 $\bar{z}$ は平均 0，分散 1 の標準正規分布に従います.

例えば一様乱数

確率分布 $p(x)$

$$p(x) = 1 \quad (0 < x < 1)$$



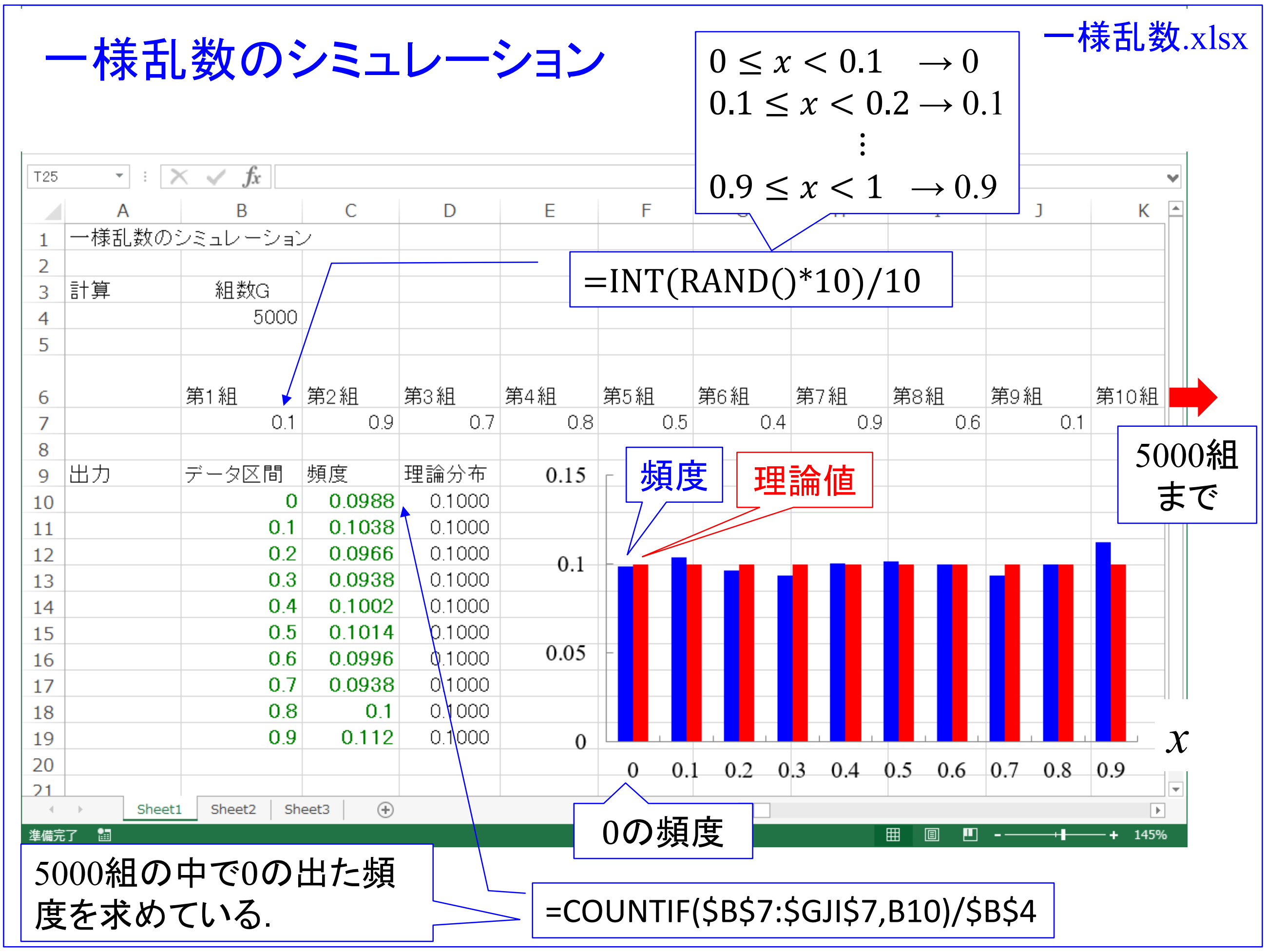
母平均 μ

$$\mu = \int_0^1 x\, p(x) dx = \int_0^1 x\, dx = \left[\frac{1}{2} x^2\right]_0^1 = 0.5$$

母分散 σ^2

$$\sigma^2 = \int_0^1 (x-\mu)^2\, p(x) dx = \int_0^1 (x-\mu)^2 = \frac{2}{3} 0.5^3$$

例えば，(0, 1)の範囲の一様乱数を考えます．この乱数は 0 より大きく 1 より小さな実数の出現確率が 1 の乱数です．この乱数の母平均をμ，母分散をσ^2 として，それぞれを定義に従って求めると，0.5, (2/3)×0.5^3 となります．



エクセルにより一様乱数のシミュレーションを実施してみましょう.「一様乱数.xlsx」のエクセルファイルを開いてください.

これは(0, 1)の範囲の一様乱数 5000 個を生成して,(0, 1) の範囲を 10 等分した各区間内の乱数の出現頻度を求めて,青いバーとしてグラフに表示するエクセルのシートです.キーボード上の F9 ボタンを押すと,シート内のすべての関数が再計算されます.そのたびに,各区間内の頻度(青いバー)は変わりますが,何度も F9 ボタンを押すと,青いバーは理論値の赤いバーの長さの周りを変化している様子が見て取れます.

INT()関数, RAND()関数, COUNTIF()関数については以降のスライドで解説します.

このエクセルファイル内でやっていることは,7 行目において,一様乱数を 5000 個生成しています.ただし,

$$
\begin{aligned}
&0 \le x < 0.1 \quad \to 0 \\
&0.1 \le x < 0.2 \to 0.1 \\
&\quad \vdots \\
&0.9 \le x < 1 \quad \to 0.9
\end{aligned}
$$

と,一様乱数 x の小数点以下 2 桁目を切り捨てています.

C 列 10 行のセル（セル C10 と略記）では，7 行目の 5000 個の乱数の中で 0 が出た回数を 5000 で割って，0 の出た頻度を求めています.

C 列 11 行のセルでは, 7 行目の 5000 個の乱数の中で 0.1 の出た頻度を求めています.

・

・

C 列 19 行のセルでは，7 行目の 5000 個の乱数の中で 0.9 の出た頻度を求めています.

D 列 10 行～D 列 19 行のセルでは 0 ～ 0.9 の出現頻度の理論値が記されています.

棒グラフの青と赤のバーは，それぞれ出現頻度とその理論値を表しています.

=INT(RAND()*10)/10

RAND() : $0 < x < 1$ の乱数 x を生成する関数.

RAND()*10 : $0 < x < 10$の一様乱数を生成.

INT(x): x の小数点以下を切り捨てて，整数部分を返す関数.

INT(x) = (RAND()*10) : $0 < x < 1 \rightarrow 0$,

$1 \leq x < 2 \rightarrow 1$,

・・・

INT(RAND()*10)/10 :

0，0.1, 0.2, ・・・, 0.9 を一様にランダムに生成します.

7 行目の各セルには

=INT(RAND()*10)/10

と 0 ～ 0.9 の 10 通りの乱数を生成する関数が書かれています.

RAND() は $0 < x < 1$ の疑似一様乱数 x を生成する関数です.

RAND()*10 は $0 < x < 10$ の一様乱数を生成する関数です.

INT(x) は x の小数点以下を切り捨てて，整数部分を返す関数です．したがって，INT(x) =INT (RAND()*10) は，$0 < x < 10$ の一様乱数を，$0 < x < 1$ の範囲は 0 に変換し，$1 \leq x < 2$ の範囲は 1 に変換します.

したがって，INT(RAND()*10)/10 は．0, 0.1, 0.2,･･･, 0.9 を一様にランダムに生成します.

=COUNTIF($B$7:$GJI$7,B10)/$B$4

B列7行目からGJI列7行目までの範囲内において，B10内の数字と同じ数字がいくつあるかを数えて，B4内の数字で割っている．

0 の出現頻度を求めている．

=COUNTIF($B$7:$GJI$7,B11)/$B$4

0.1 の出現頻度を求めている．

セル内の数式をコピーするとセル番号は自動変更される．例えば，C10の数式をC11にコピーすると，数式内のB10は自動的にB11へと変更される．$ 記号があると，その右側の記号，数字はコピーによって変更されない．

=COUNTIF($B$7:$GJI$7,B10)/$B$4

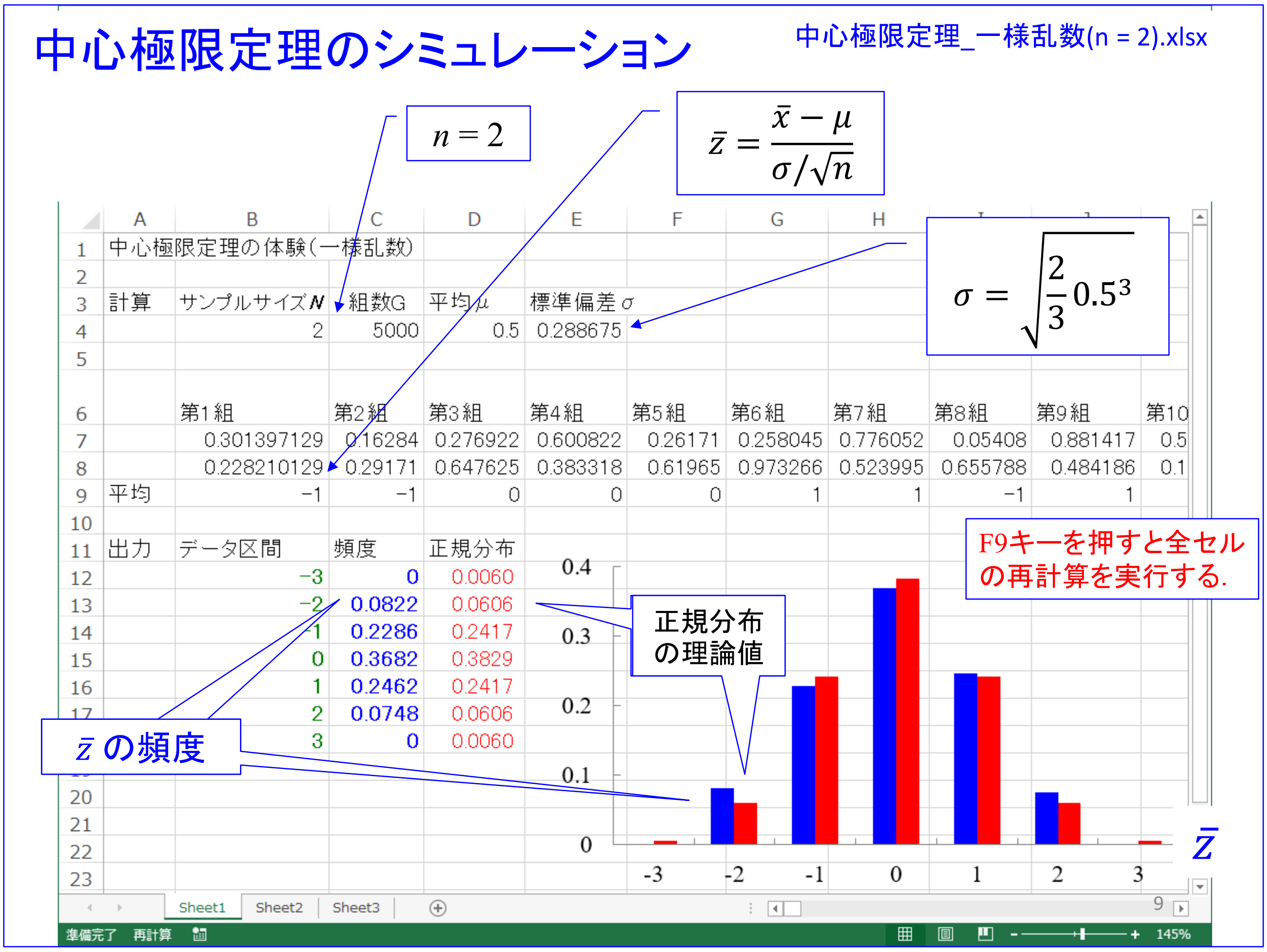
は，セル B10 内の数値と同じ数値がセル B7 からセル GJI7 までの範囲内にいくつあるかを数えて，その値をセル B4 内の数値で割っています．セル B7 内の数値が 0 なので，この数式は 0 の出現頻度を求めています．同様に

=COUNTIF($B$7:$GJI$7,B11)/$B$4

は，0.1 の出現頻度を求めています．

$記号は数式をコピーする際に意味があります．例えば，セル B10 内の数式をセル B11 にコピーすると，上記の =COUNTIF($B$7:$GJI$7,B11)/$B$4 が得られます．$記号が付いていないセル記号／番号（この例ではセル記号は B，セル番号は 11 です.）はコピーにより自動更新されます．すなわち，セル C10 内の数式がセル C11 にコピーされたので，数式内のセル番号 10 から 11 に 1 つ更新されています．（もし，セル C10 から D10 にコピーされれば，B10 → C10 とセル記号が更新されます.）

記号もしくは番号の左隣に$記号がある場合は，コピーによる更新はされません．



エクセルにより中心極限定理のシミュレーションを実施してみましょう．「中心極限定理_一様乱数(n=2).xlsx」 のエクセルファイルを開いてください．

B 列 7, 8 行のセルにて(0, 1)の間の一様乱数を生成しています．セル B9 では

=ROUND((AVERAGE(B7:B8)-$D$4)/$E$4*SQRT($B$4),0)

により，

$$\bar{z} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

を求めて，小数点以下 1 桁目を四捨五入しています．すなわち，セル B7,B 8 の乱数値の平均値を求めて，それを平均 0，分散 1 となるように規格化し，小数点以下 1 桁目を四捨五入しています．

以降，右方向の第 2 組～第 5000 組においてそれぞれサンプルサイズ $n = 2$ の場合の規格化された平均値を求めています．

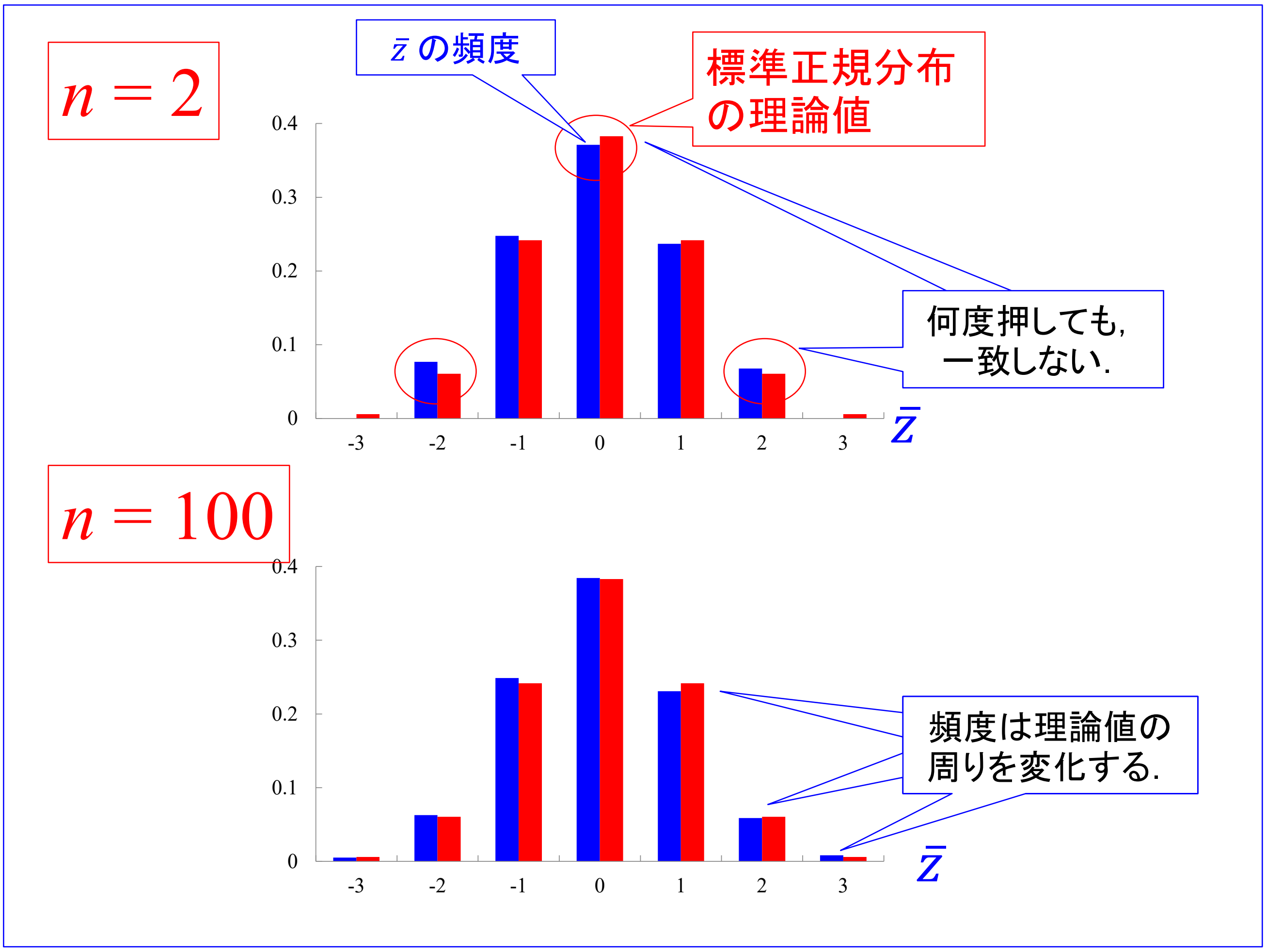
セル C12 では，第 9 行目の -3 の個数（$-3.5 \leq \bar{z} < -2.5$ にある$\bar{z}$の個数）を数えて，5000 で割ることで，$\bar{z}$ が $-3.5 \leq \bar{z} < -2.5$ にある頻度を求めています．同様に，セル C13～C18 では，$\bar{z}$ が $-2.5 \leq \bar{z} < -1.5$，$-1.5 \leq \bar{z} < -0.5$，・・・にある頻度を求めています．

セル D12～D18 では，母集団が標準正規分布に従うとした場合の，対応する区間における生起確率を求めています．

棒グラフは，セル C12～C18 の頻度を青色で，セル D12～D18 の理論確率を赤色で示しています．

F9 キーを押すと全セル内の関数を再計算させることができます．F9 キーを一回押すことで，一様乱数から 2 個のサンプルを抽出する試行を 5000 組繰り返して，得られた値の出現頻度を棒グラフに描くシミュレーションを 1 回実行することができます．F9 キーを何度も押して分かることは，$-2.5 \le \bar{z} < -1.5$，$1.5 \le \bar{z} < 2.5$ の頻度は理論値を下回ることがほとんど無いこと，$-0.5 \le \bar{z} < 0.5$ の頻度は理論値を上回ることがほとんど無いことです．

サンプルサイズが小さい場合，$\bar{z}$ は標準正規分布に従っていません．

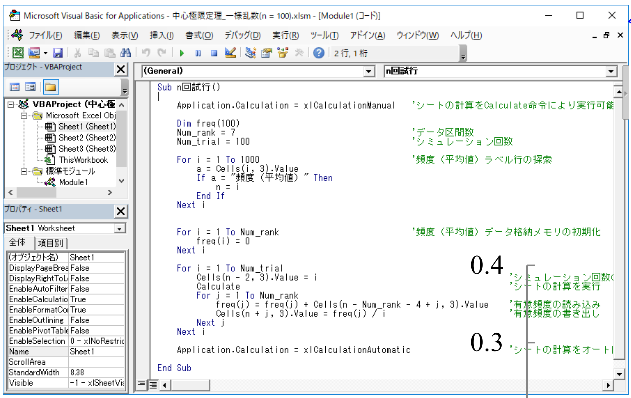


図は上側がサンプルサイズ $n = 2$ の場合の $\bar{z}$ の頻度（青色）と標準正規分布の理論値（赤色）です.

下側はサンプルサイズ $n = 100$ の場合です.「中心極限定理_一様乱数(n=100).xlsm」のエクセルファイルを開いてください．F9 キーを押すと，頻度は理論値を中心にして，上下に変化します．サンプルサイズが大きいと，規格化された平均値 $\bar{z}$ が標準正規分布に従うことをシミュレーションにより体験できます.

マクロを書いて，5000組のシミュレーションを 100回実行し，積算頻度分布を求める.

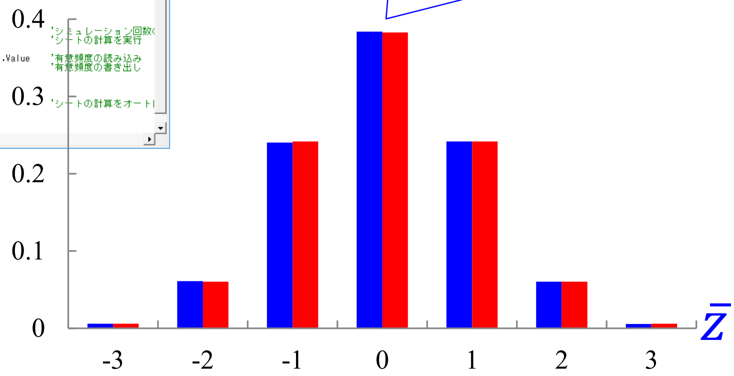
開発→マクロ→編集



頻度データの開始行を探して，そこから下行のデータを読み出して，頻度（平均値）の下行に書き出すマクロ.

5000組×100回の頻度は理論値にほぼ一致する.

Num_trial = 100の実行結果



「中心極限定理_一様乱数(n=100).xlsm」 の拡張子 xlsm はマクロ付きのエクセルファイルです．ウィンドウ上部にあるメニューから，開発 を選んで，マウスの左ボタンをクリック（左クリックと呼びます.）してください．以降，マクロ → 編集と左クリックをしていくと，マクロの編集画面を開くことができます．これは，「n 回試行」という名前のマクロです.

Num_trial = 100 により，$\bar{z}$ を 5000 組について求めて，$-3.5 \le \bar{z} < -2.5$，$-2.5 \le \bar{z} < -1.5$，・・・の頻度を求めることを 100 回繰り返して，100 回分の積算頻度を求めるマクロです．マクロはメニューバーの ▻ ボタンを押すことで実行できます．計算には時間がかかるので，セル C118 に実行中のシミュレーション回数が表示されます．また，セル C121 以下に積算頻度が表示されます．この頻度は，そのシミュレーション回数までの積算頻度が表示されます.

5000 組×100 回のシミュレーション結果では，各区間の頻度が標準正規分布の理論確率にほぼ一致します.

2. 二項検定:ノンパラメトリック検定第一歩

二つのクラス(例えばコインの裏表)からなる母集団

1 である確率 p

0 である確率 $1-p$

帰無仮説H_0:$p_1 = p, p_0 = 1-p$

対立仮説H_1:$p_1 \neq p, p_0 \neq 1-p$

n 個のサンプルにおいて, 1 が m 個である確率 $p(m)$は

$$p(m) = {}_nC_m p^m (1-p)^{n-m} \qquad (2.1)$$

となる.

さて，二項検定から見ていきましょう．これは母集団が二項分布に従う場合の検定です．冒頭にノンパラメトリック検定は母集団分布に特定の分布を仮定しない検定法であると述べましたが，二項検定は二項分布を仮定しています．以降のノンパラメトリック検定のウォーミングアップの位置づけです．

事象 X の値 x のとる値が 0，1 の 2 種類である場合を考えます．例えば，コインの裏表のように事象が 2 つのクラスからなる確率試行を考えましょう．裏が出た場合 $x=0$，表が出た場合 $x=1$ と値を対応づけ，$x=1$ となる確率を p，$x=0$ となる確率を $1-p$ とします．

事象 X が上記の二項分布に従っていないことの検定は次の通りです．

$x=1$ となる確率を p_1，$x=0$ となる確率を p_0 とします．

帰無仮説 H_0 を $p_1=p$，$p_0=1-p$ とし，対立仮説 H_1 を $p_1 \neq p$，$p_0 \neq 1-p$ とします．検定は帰無仮説が成立しているとして理論を構築します．

帰無仮説 H_0 が成立しているとすると，N 個のサンプルにおいて，1 が m 個である確率 $p(m)$は

$$p(m) = {}_NC_m p^m (1-p)^{n-m} \qquad (2.1)$$

と求められます．

二項検定の例：サンプルサイズが小さい場合($n=10$)

1	二項検定		
2		サンプルサイズn	p
3		10	0.5
4			
5		m	$p(m)$
6		0	0.000977
7		1	0.009766
8		2	0.043945
9		3	0.117188
10		4	0.205078
11		5	0.246094
12		6	0.205078
13		7	0.117188
14		8	0.043945
15		9	0.009766
16		10	0.000977

二項検定(n=10).xlsx

帰無仮説H_0：$p_0=0.5, p_1=0.5$
対立仮説H_1：$p_0\neq 0.5, p_1\neq 0.5$

母集団から10個のサンプルを抜き出したときに，1 が 1 個しか無かったとする．
帰無仮説が正しければ，1 が 1 個以下，もしくは 9 個以上である確率は

$$p_1(0)+p_1(1)+p_1(9)+p_1(10)$$
$$=0.0215$$

結論を間違う確率（危険率，有意水準）を 5% とすると，2.15 %（両側検定）は有意水準を下回る．

帰無仮説が間違っていたとして，これを棄却し，対立仮説を採用する．

検定はサンプルサイズが小さい場合と大きい場合で異なったアプローチがとられます．まず，サンプルサイズが小さい場合($n\leq 25$)です．「二項検定(n=10).xlsx」を開いてください．

サンプルサイズ　$n=10$ の場合の二項分布の確率 $p(m)$ を求めているエクセルファイルです．

帰無仮説 H_0： $p_1=0.5$，$p_0=0.5$
対立仮説 H_1： $p_1\neq 0.5$，$p_0\neq 0.5$

としています．

両側検定の考え方を，具体例で示します．

母集団から 10 個のサンプルを抜き出したときに，1 が 1 個しか無かったとします．帰無仮説が正しいとすると，1 が 1 個以下，もしくは 9 個以上である確率は

$$p_1(0)+p_1(1)+p_1(9)+p_1(10)=0.0215$$

です．すなわち，帰無仮説が正しかったとすると，目の前の 1 が 1 個しか出なかったことは，滅多に起きない（生起確率 2.15%）ことが起きたことになります．これは帰無仮説が間違っていたとして，これを棄却し，代わりに対立仮説を採用します．もちろん，本当は帰無仮説が正しかった場合，1 が 1 個以下，もしくは 9 個以上出る確率は 2.15%あります．この場合は，間違って対立仮説を採用してしまうことになります．その間違う危険率が 2.15%です．

結論を間違う確率（危険率，有意水準）が 5% 以下であれば，対立仮説を採用するとするのが，統計的検定の考え方です．

危険率は 1% が採用されることもあります．5%とするか 1%とするかには理論的根拠はありません．経験的な受容レベルです．

二項検定の例：サンプルサイズが大きい場合$(n > 25)$

二つのクラス（例えばコインの裏表）からなる母集団

帰無仮説H_0：$p_1 = p, p_0 = 1 - p$
対立仮説H_1：$p_1 \neq p, p_0 \neq 1 - p$

n 個のサンプルの値の平均値を $\bar{x}$ とすると

平均値：$\mu = p$

標準偏差：$\sigma = \sqrt{p(1-p)/n}$

規格化された平均値：$\bar{z} = \frac{\bar{x}-\mu}{\sigma} = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}}$

次に，サンプルサイズが大きい場合$(n > 25)$です．「二項検定(n=25).xlsx」を開いてください．

サンプルサイズ n が大きくなると中心極限定理により，n 個のサンプルの値の平均値を $\bar{x}$ とすると，規格化された平均値 $\bar{z}$ は標準正規分布に従います．二項分布の平均値 $\bar{x}$ の

平均値：$\mu = p$

標準偏差：$\sigma = \sqrt{p(1-p)/n}$

です．規格化された平均値は

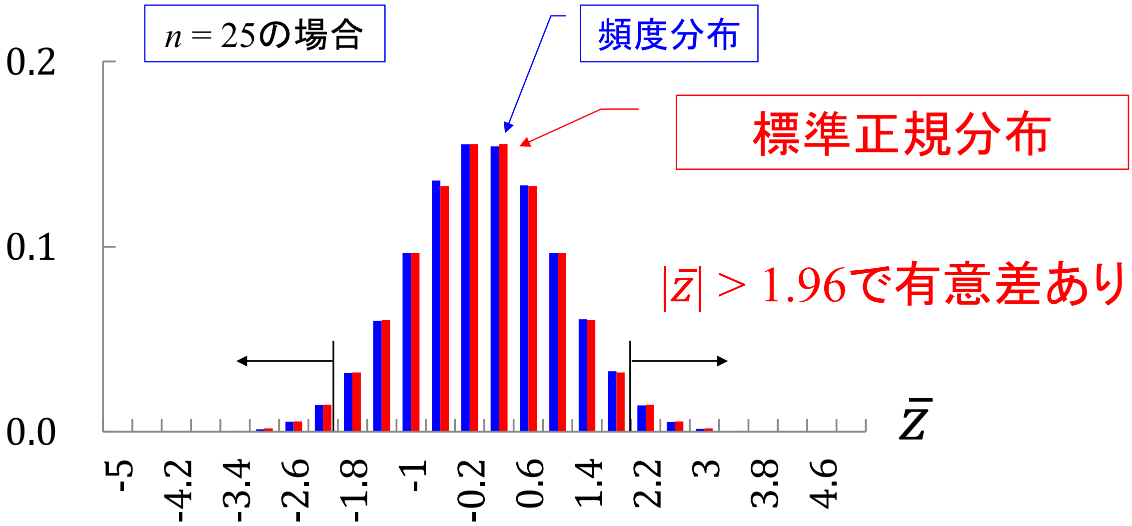
$$\bar{z} = \frac{\bar{x}-\mu}{\sigma} = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}}$$

です．

二項検定の例：サンプルサイズが大きい場合($n > 25$)

$$\bar{z} = \frac{\bar{x} - p}{\sqrt{p(1-p)/n}}$$

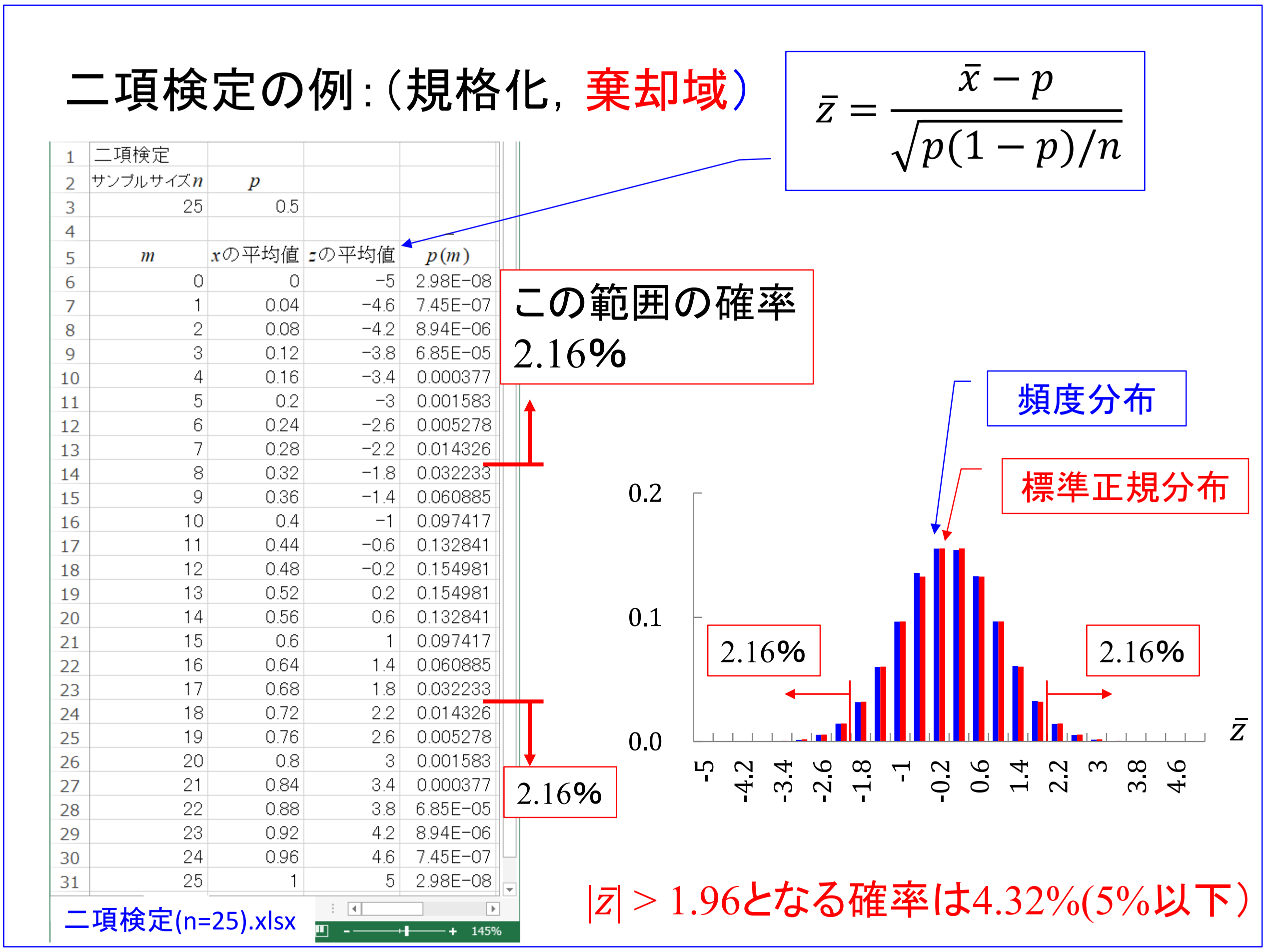
➡ 中心極限定理により n が大きいと $\bar{z}$ は標準正規分布に従う.



サンプルサイズ n が大きいと，中心極限定理により $\bar{z}$ は標準正規分布に従います.

サンプルサイズ $n = 25$, $p = 0.5$ の場合の $\bar{z}$ の頻度分布と標準正規分布の棒グラフを示します．頻度分布と標準正規分布はよく一致しています.

標準正規分布において，$|\bar{z}| > 1.96$の確率が 0.05 です.

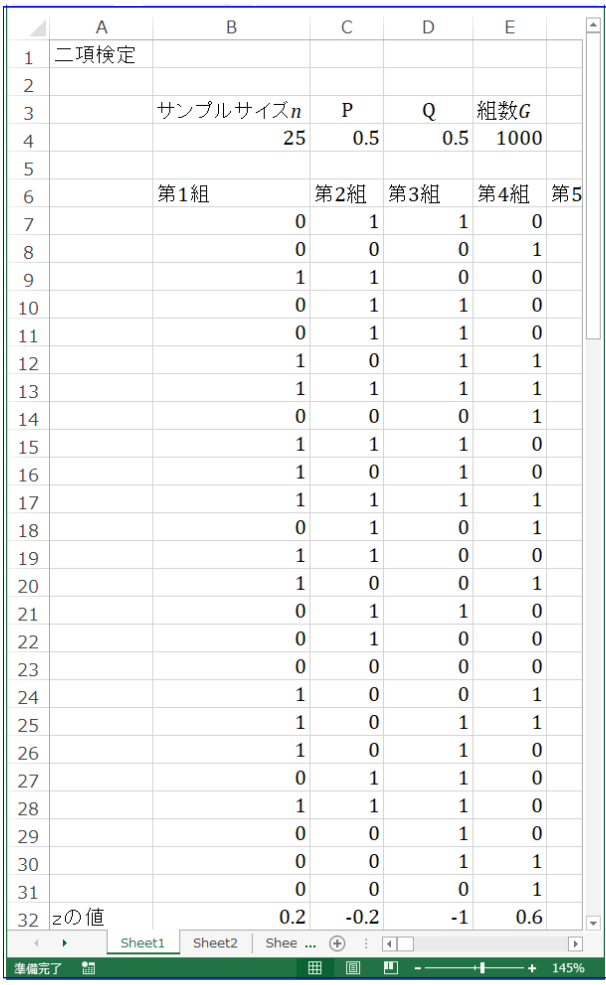


1	二項検定			
2	サンプルサイズn	p		
3	25	0.5		
4				
5	m	xの平均値	zの平均値	$p(m)$
6	0	0	-5	2.98E-08
7	1	0.04	-4.6	7.45E-07
8	2	0.08	-4.2	8.94E-06
9	3	0.12	-3.8	6.85E-05
10	4	0.16	-3.4	0.000377
11	5	0.2	-3	0.001583
12	6	0.24	-2.6	0.005278
13	7	0.28	-2.2	0.014326
14	8	0.32	-1.8	0.032233
15	9	0.36	-1.4	0.060885
16	10	0.4	-1	0.097417
17	11	0.44	-0.6	0.132841
18	12	0.48	-0.2	0.154981
19	13	0.52	0.2	0.154981
20	14	0.56	0.6	0.132841
21	15	0.6	1	0.097417
22	16	0.64	1.4	0.060885
23	17	0.68	1.8	0.032233
24	18	0.72	2.2	0.014326
25	19	0.76	2.6	0.005278
26	20	0.8	3	0.001583
27	21	0.84	3.4	0.000377
28	22	0.88	3.8	6.85E-05
29	23	0.92	4.2	8.94E-06
30	24	0.96	4.6	7.45E-07
31	25	1	5	2.98E-08

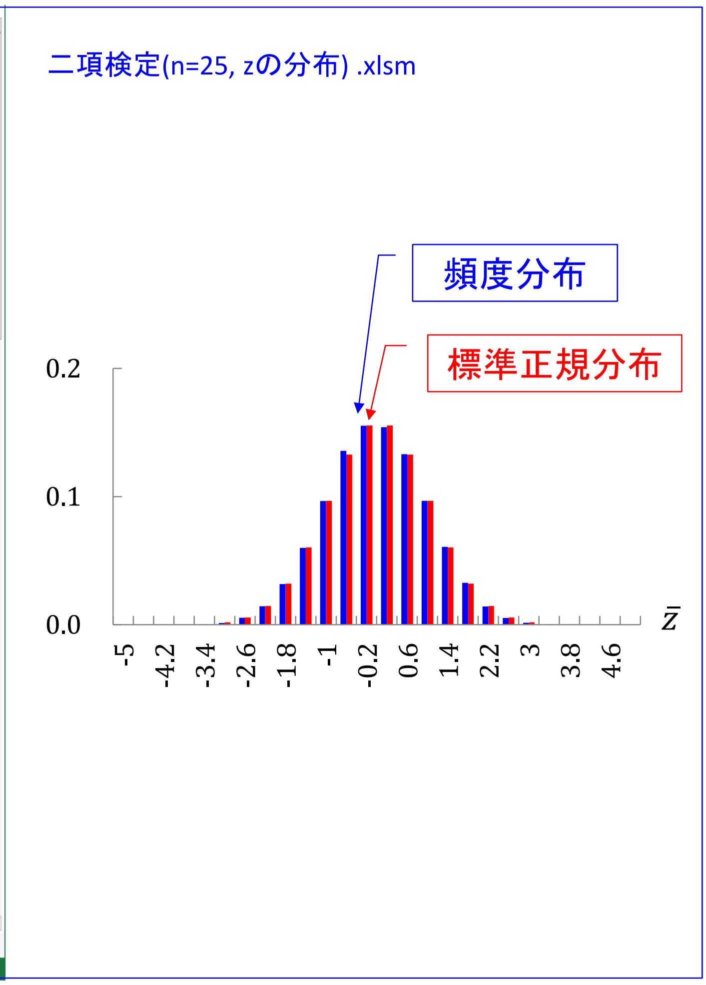
「二項検定(n=25).xlsx」を開いてください．サンプルサイズ $n = 25$，$p = 0.5$ の場合における，1 の数 m と$\bar{x} = \frac{m}{n}$，$\bar{z} = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}}$ の値を求め，25 個のサンプル中で 1 の数が m である確率 $p(m)$ を求めています．

セル D6 以下のセルに，(2.1)式に従って確率$p(m)$を求めています．$n = 25$の場合，$|\bar{z}| > 1.96$ となる確率は，4.32%です．

すなわち，1 が 7 個以下，もしくは，18 個以上である確率が 4.32%です．$n = 25$の場合，1 が 7 個以下，もしくは，18 個以上が帰無仮説を棄却できる棄却域です．規格化されたこの値では，$|\bar{z}| > 1.96$ が棄却域です．なお，4.32%は 5%より小さいのですが，仮に 1 が 8 個以下もしくは 17 個以上を棄却域とすると，そのような結果となる確率は 10.8%であり，有意水準 5%を大きく上回ってしまいます．



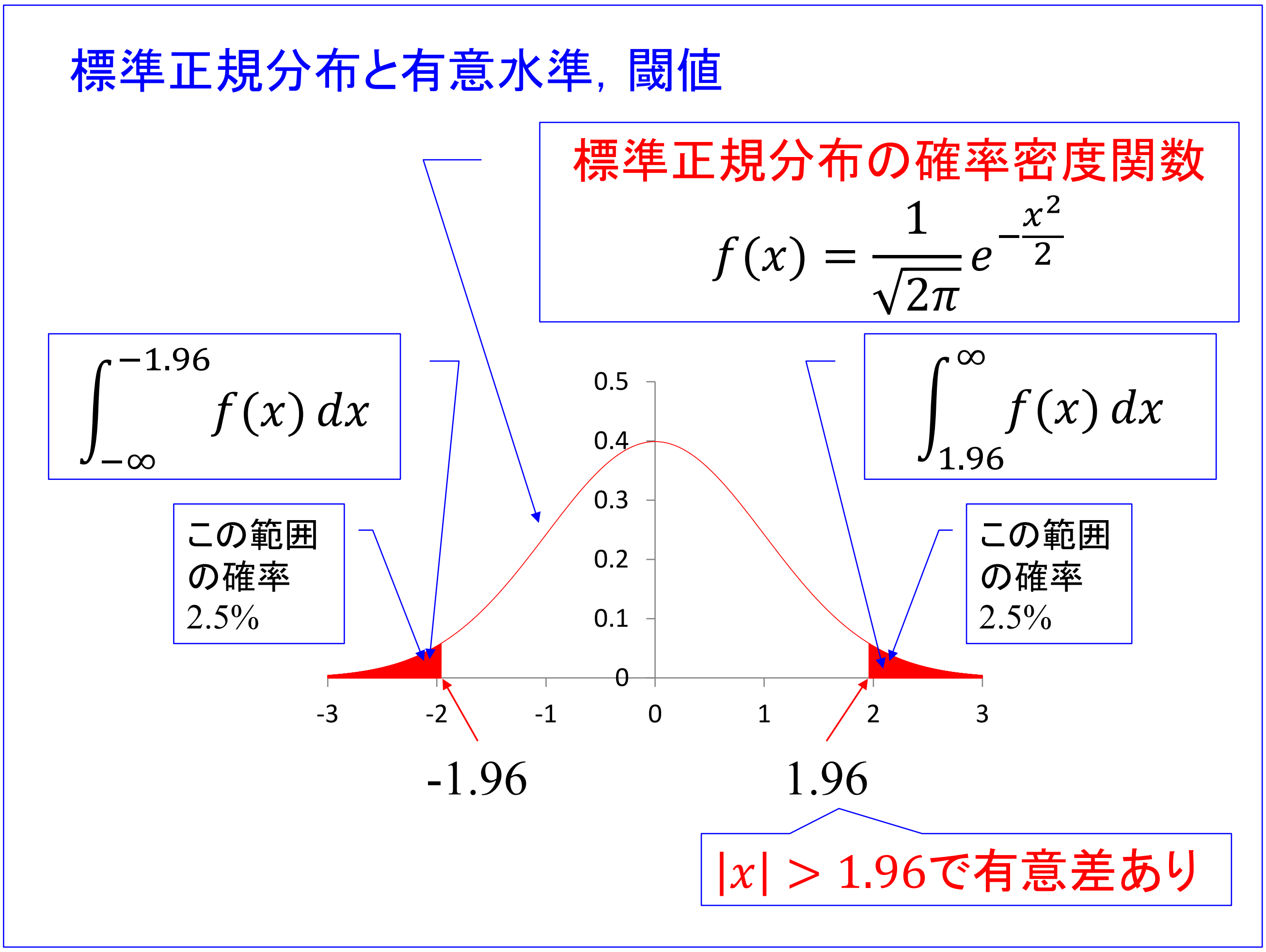
	A	B	C	D	E	
1	二項検定					
2						
3		サンプルサイズn	P	Q	組数G	
4		25	0.5	0.5	1000	
5						
6		第1組	第2組	第3組	第4組	第5
7		0	1	1	0	
8		0	0	0	1	
9		1	1	0	0	
10		0	1	1	0	
11		0	1	1	0	
12		1	0	1	1	
13		1	1	1	1	
14		0	0	0	1	
15		1	1	1	0	
16		1	0	1	0	
17		1	1	1	1	
18		0	1	0	1	
19		1	1	0	0	
20		1	0	0	1	
21		0	1	1	0	
22		0	1	0	0	
23		0	0	0	0	
24		1	0	0	1	
25		1	0	1	1	
26		1	0	1	0	
27		0	1	1	0	
28		1	1	1	0	
29		0	0	1	0	
30		0	0	1	1	
31		0	0	0	1	
32	zの値	0.2	-0.2	-1	0.6	



「二項検定(n=25, z の分布) .xlsm」を開いてください．サンプルサイズ $n = 25$, $p = 0.5$, $q(= 1 - p) = 0.5$の場合の 1000 組×100 回 のシミュレーションを行うエクセルファイルです．

32 行目の $\bar{z} = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}}$ の値は小数点以下 2 桁目で四捨五入しています．

このシミュレーションでは，1，0 の数字の出る確率が 2 項分布に従う場合，25 個のサンプルに含まれる 1 の数から得られる $\bar{z} = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}}$ の値は，組ごとにばらつきますが，1000 組×100 回 での出現頻度は標準正規分布に従うことを体験できます．



これは標準正規分布の確率密度関数 $f(x)$

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

のグラフです．$|x|>1.96$ の領域で $f(x)$と x 軸の間を赤く塗ってあります．この赤い部分の面積

$$\int_{-\infty}^{-1.96} f(x)\,dx+\int_{1.96}^{\infty} f(x)\,dx$$

は 0.05 です．この値は，事象 X が標準正規分布に従うときに，x が$|x|>1.96$ の範囲にある確率です．すなわち，事象 X が標準正規分布に従うとき，有意水準が 5%における x の閾値は 1.96 です．

3. χ^2分布：ノンパラメトリック検定の第二歩目

二つのクラス（例えばコインの裏表）からなる母集団

1 である確率　P
0 である確率　$1-P$

n 個のサンプルの値の平均値を $\bar{x}$ とすると

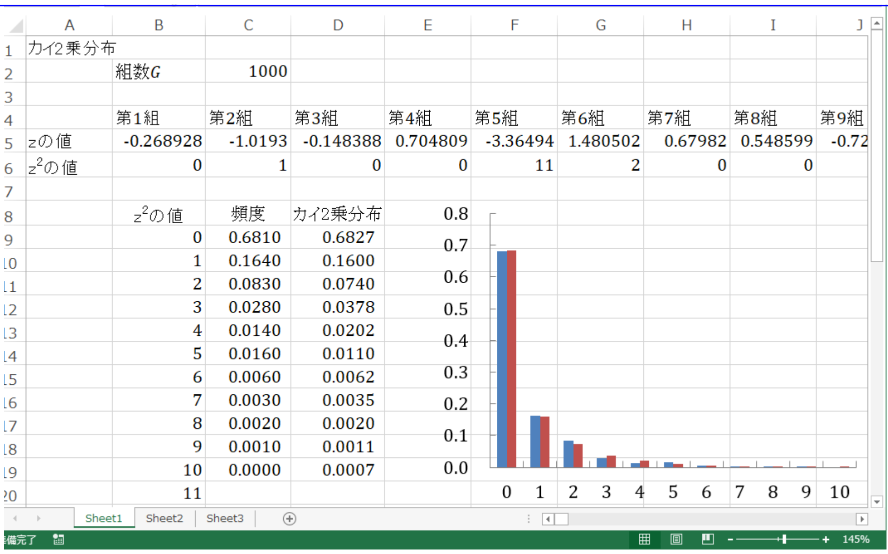
$$\bar{z} = \frac{\bar{x} - p}{\sqrt{p(1-p)/n}}$$

➡ 中心極限定理により n が大きいと $\bar{z}$ は標準正規分布に従う.

$$\bar{z}^2 = \left\{\frac{\bar{x} - p}{\sqrt{p(1-p)/n}}\right\}^2$$

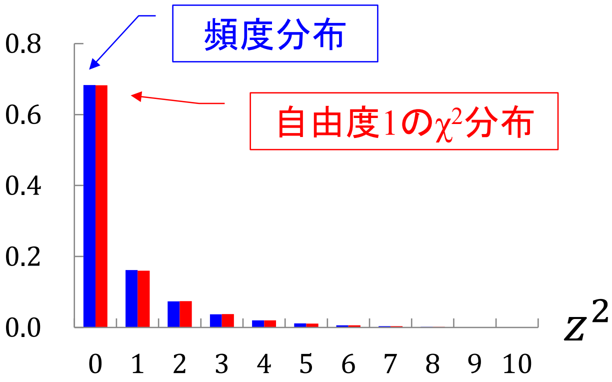
➡ $\bar{z}^2$ は自由度1の χ^2 分布に従う.

$\bar{z}$ が標準正規分布に従う場合，$\bar{z}^2$ は自由度 1 の χ^2 分布に従います.



	A	B	C	D	E	F	G	H	I	J
1	カイ2乗分布									
2		組数G	1000							
3										
4		第1組	第2組	第3組	第4組	第5組	第6組	第7組	第8組	第9組
5	zの値	-0.268928	-1.0193	-0.148388	0.704809	-3.36494	1.480502	0.67982	0.548599	-0.72
6	z^2の値	0	1	0	0	11	2	0	0	
7										
8		z^2の値	頻度	カイ2乗分布						
9		0	0.6810	0.6827						
10		1	0.1640	0.1600						
11		2	0.0830	0.0740						
12		3	0.0280	0.0378						
13		4	0.0140	0.0202						
14		5	0.0160	0.0110						
15		6	0.0060	0.0062						
16		7	0.0030	0.0035						
17		8	0.0020	0.0020						
18		9	0.0010	0.0011						
19		10	0.0000	0.0007						
20		11								

カイ2乗分布.xlsm



「カイ 2 乗分布.xlsm」のファイルを開いてください．z が標準正規分布に従う場合，z^2 は自由度 1 の χ^2 分布に従うことを体験できるシミュレーションです．

5 行目で標準正規分布に従う乱数 z を生成して，6 行目でその 2 乗値 z^2 を求めて，小数点以下 1 桁目を切り捨てることを 1000 組について実施しています．セル C9 は 6 行目における 0 の個数を数えて 1000 で割ることで，6 行目における 0 の出現頻度を求めています．以下のセルでは，それぞれ 1, 2, 3, ・・・の出現頻度を求めています．セル D9 以下では自由度 1 の χ^2 分布の区間 $0 \leq z^2 < 1$，$1 \leq z^2 < 2$，$2 \leq z^2 < 3$，・・・における理論確率を求めています．

5 行目の各セル内の

=NORMSINV(RAND())

は標準正規乱数を生成する関数です．

セル D9 内の

=CHIDIST(B9,1)-CHIDIST(B10,1)

は $f_{\chi 1}(x)$ を自由度 1 のカイ 2 乗分布の確率密度関数とすると

CHIDIST(B9,1) - CHIDIST(B10,1) = $\int_0^{\infty} f_{\chi 1}(x)dx - \int_1^{\infty} f_{\chi 1}(x)dx = \int_0^1 f_{\chi 1}(x)dx$

です.

変数 z^2 の変形

$$
\begin{aligned}
z^2 &= \left\{\frac{\bar{x}-p}{\sqrt{p(1-p)/n}}\right\}^2 = \left\{\frac{n\bar{x}-np}{\sqrt{np(1-p)}}\right\}^2 \\
&= (n\bar{x}-np)^2\frac{1}{np(1-p)} \\
&= (n\bar{x}-np)^2\frac{1-p+p}{np(1-p)} \\
&= \frac{(n\bar{x}-np)^2}{np}+\frac{(n\bar{x}-np)^2}{n(1-p)} \\
&= \frac{(n\bar{x}-np)^2}{np}+\frac{(n-n\bar{x}-n(1-p))^2}{n(1-p)} \\
&= \frac{(X_1-np)^2}{np}+\frac{(X_0-n(1-p))^2}{n(1-p)}
\end{aligned}
$$

$X_1 = n\bar{x}$：1の数
$X_0 = n - n\bar{x}$：0の数

ここで，z^2 の式を変形します．

χ^2分布の拡張

p：1の出る確率
q：0の出る確率　　$p+q=1$

np : 1 の出る回数の期待値
nq : 0 の出る回数の期待値

1の数	0の数
x_1	x_0

$$z^2 = \frac{(x_1-np)^2}{np} + \frac{(x_0-nq)^2}{nq} \quad \sim \chi_1^2 : \text{自由度1の}\chi^2\text{分布}$$

拡張（文献[1]～[3] に証明あり．）　(3.1)

1の数	2の数	・・	kの数
x_1	x_2	・・	x_k

p_i : i の出る確率

$p_1+p_2+\cdots+p_k=1$

$$z^2 = \frac{(x_1-np_1)^2}{np_1} + \frac{(x_2-np_2)^2}{np_2} + \cdots + \frac{(x_k-np_k)^2}{np_k} \quad (3.2)$$

$\boxed{\sim \chi_{k-1}^2 : \text{自由度}k\text{-1の}\chi^2\text{分布}}$

2 項分布の z^2 の式を変形すると，$z^2=\frac{(X_1-Np)^2}{Np}+\frac{(X_0-Nq)^2}{Nq}$ となります．p は 1 の出る確率，q は 0 の出る確率であり，np は 1 の出る回数の期待値，nq は 0 の出る回数の期待値です．

証明は省略して，2 項分布を k 項分布に拡張します．$i\ (i=1,2,\cdot\cdot\cdot,k)$の出る確率を$p_i$とします．このとき

$$z^2 = \frac{(X_1-Np_1)^2}{Np_1} + \frac{(X_2-Np_2)^2}{Np_2} + \cdots + \frac{(X_k-Np_k)^2}{Np_k}$$

は，自由度 $k-1$ の χ^2 分布に従います．1～k の出る確率 $p_1,p_2,\cdots,p_k$に対する制約は

$$p_1+p_2+\cdots+p_k=1$$

です．

4. 適合度検定：母集団の確率分布の検定

k 個のクラスからなる母集団

帰無仮説H_0：各クラスの度数分布は期待度数
$e_1, e_2, \cdots, e_k$と一致する.

対立仮説H_1：各クラスの度数分布は期待度数
$e_1, e_2, \cdots, e_k$と一致しない.

各クラスのサンプル数 $x_1, x_2, \cdots, x_k$であったとき,

$$z^2 = \frac{(x_1 - e_1)^2}{e_1} + \frac{(x_2 - e_2)^2}{e_2} + \cdots + \frac{(x_k - e_k)^2}{e_k}$$

$\sim \chi^2_{k-1}$：自由度k-1のχ^2分布

k 個のクラスからなる母集団から得られた散布から，母集団の確率分布を検定する方法に適合度検定があります.

各クラスの母集団の確率分布が$p_1, p_2, \cdots, p_k$ であると仮定します．サンプルサイズを n とすると，各クラスの期待*度数*$e_1, e_2, \cdots, e_k$ は $np_1, np_2, \cdots, np_k$ です.

帰無仮説 H_0，対立仮説 H_1 を以下とします.

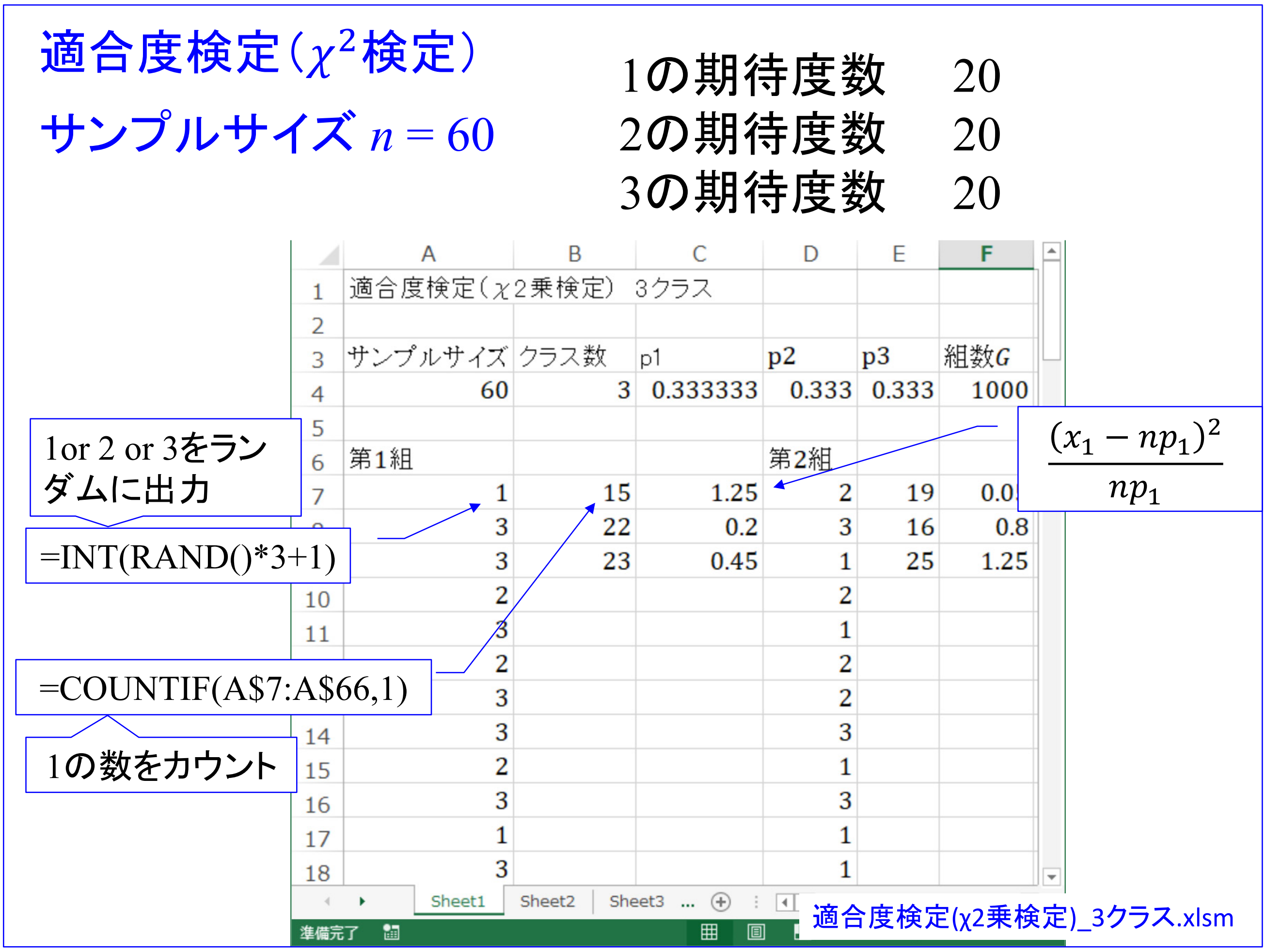
帰無仮説 H_0：*各クラスの度数分布は期待度数*$e_1, e_2, \cdots, e_k$ と一致する.

対立仮説 H_1：*各クラスの度数分布は期待度数*$e_1, e_2, \cdots, e_k$ と一致しない.

このとき，各クラスのサンプル数が $x_1, x_2, \cdots, x_k$であったとすると,

$$z^2 = \frac{(x_1 - e_1)^2}{e_1} + \frac{(x_2 - e_2)^2}{e_2} + \cdots + \frac{(x_k - e_k)^2}{e_k}$$

は，自由度 k - 1 の χ^2分布: χ^2_{k-1}に従います.



	A	B	C	D	E	F
1	適合度検定(χ2乗検定) 3クラス					
2						
3	サンプルサイズ	クラス数	p1	p2	p3	組数G
4	60	3	0.333333	0.333	0.333	1000
5						
6	第1組			第2組		
7	1	15	1.25	2	19	0.0
8	3	22	0.2	3	16	0.8
9	3	23	0.45	1	25	1.25
10	2			2		
11	3			1		
12	2			2		
13	3			2		
14	3			3		
15	2			1		
16	3			3		
17	1			1		
18	3			1		

「適合度検定(χ2 乗検定)_3 クラス.xlsm」のファイルを開いてください．3 クラスでサンプルサイズ *n=60*，各クラスの期待度数 $e_1=20, e_2=20$, $e_3=20$ としたときの

$$z^2=\frac{(x_1-e_1)^2}{e_1}+\frac{(x_2-e_2)^2}{e_2}+\frac{(x_3-e_3)^2}{e_3}$$

の分布を体験するシミュレーションです．

=INT(RAND()*3+1)

により，1 or 2 or 3 を等確率でランダムに生成します．

=COUNTIF(A$7:A$66,1)

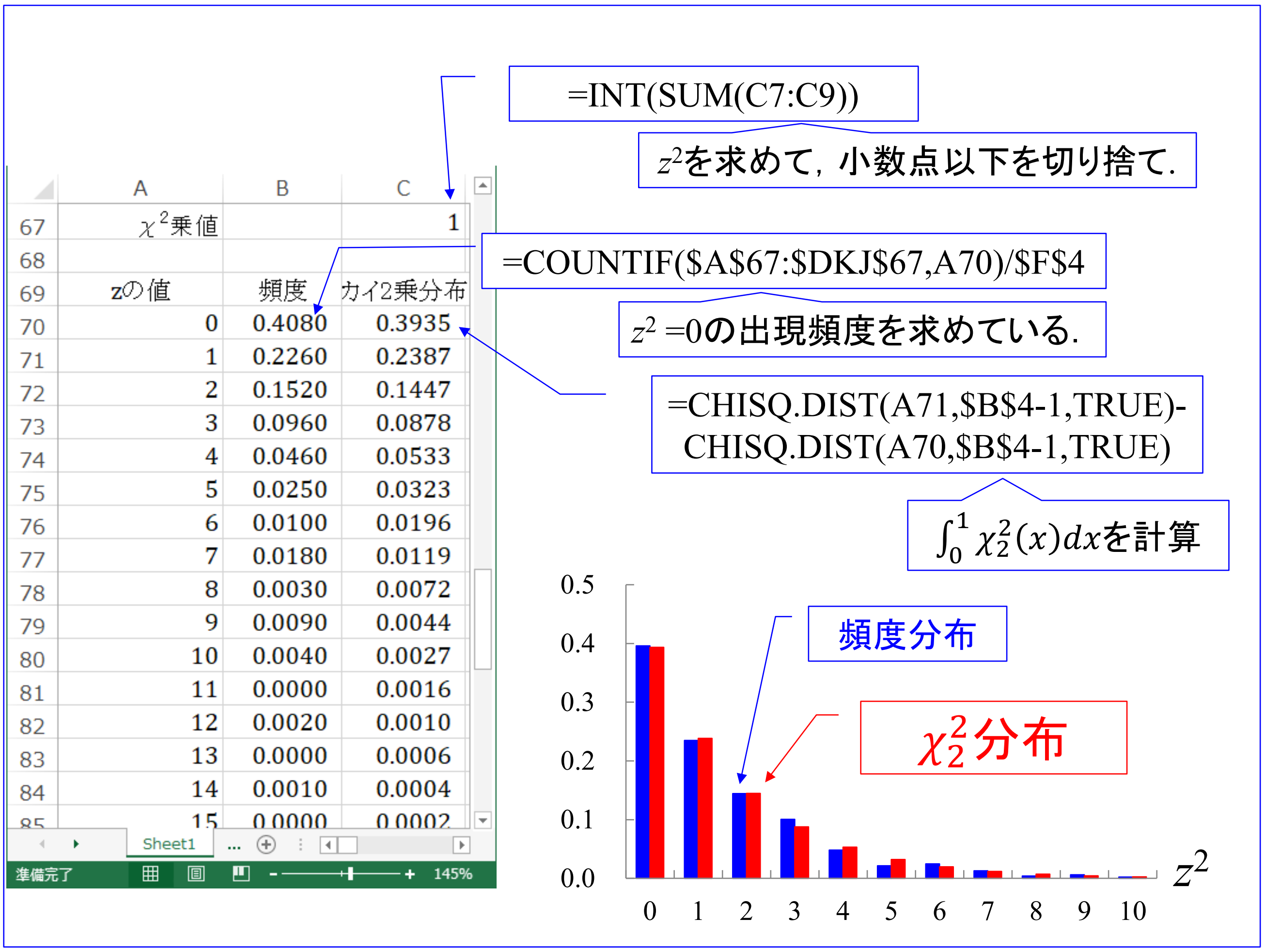
は，A 列の 60 個のサンプルから 1 の数を数えます．

セル C7 では

$$z_1^2 = \frac{(x_1 - np_1)^2}{np_1}$$

により，z_1^2 を計算します．

67 行目では z^2 の値を求めて，小数点以下 1 桁目を切り捨てています．



	A	B	C
67	χ²乗値		1
68			
69	zの値	頻度	カイ2乗分布
70	0	0.4080	0.3935
71	1	0.2260	0.2387
72	2	0.1520	0.1447
73	3	0.0960	0.0878
74	4	0.0460	0.0533
75	5	0.0250	0.0323
76	6	0.0100	0.0196
77	7	0.0180	0.0119
78	8	0.0030	0.0072
79	9	0.0090	0.0044
80	10	0.0040	0.0027
81	11	0.0000	0.0016
82	12	0.0020	0.0010
83	13	0.0000	0.0006
84	14	0.0010	0.0004
85	15	0.0000	0.0002

セル B70 では，

=COUNTIF($A$67:$DKJ$67,A70)/$F$4

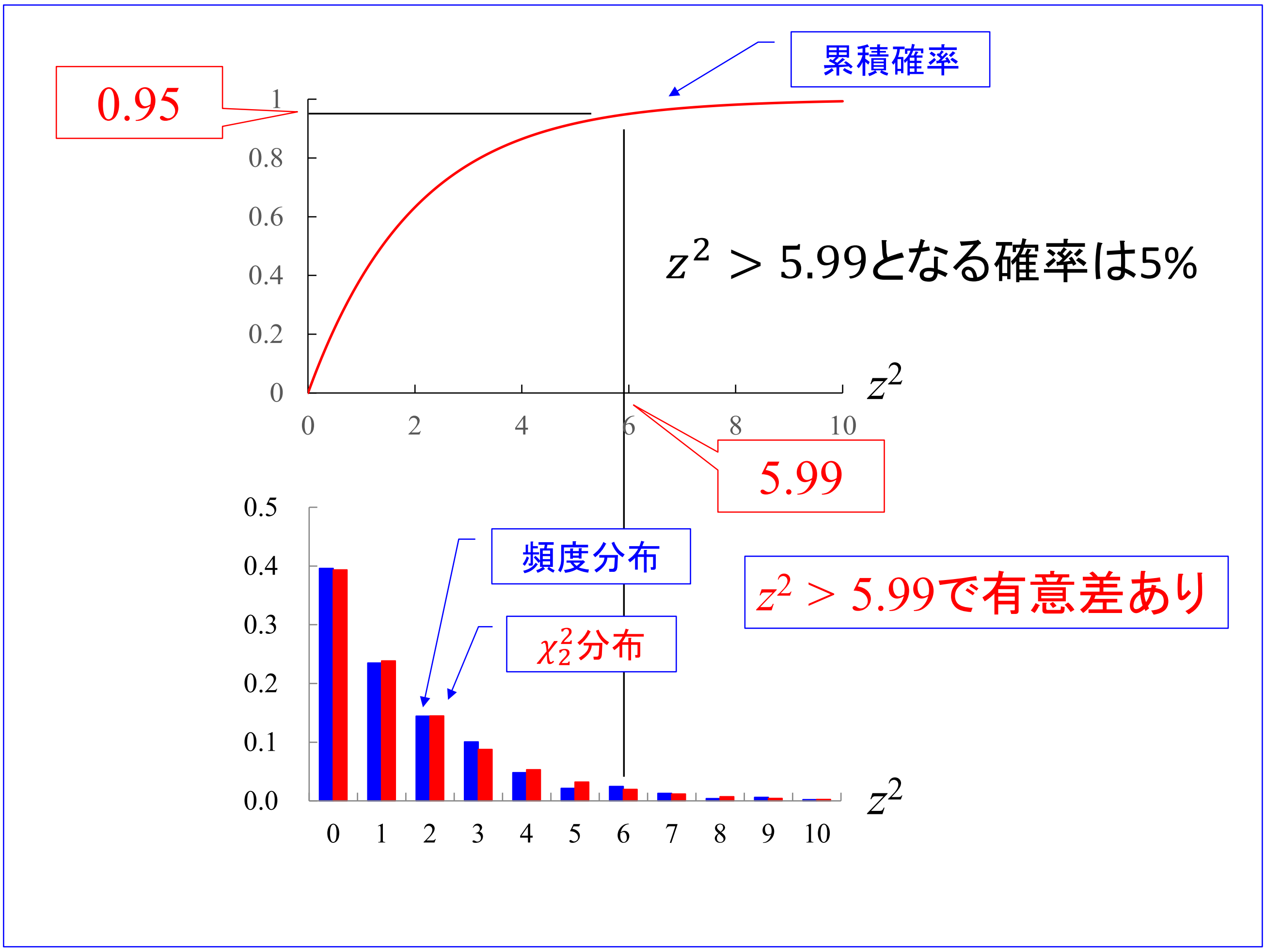
により，67 行目にある 1000 組の z^2 の値の中から，0 の出現頻度を求めています．やっていることは，A 列 70 行目で指定されている数値: 0 の数を数えて，1000 で割っています．

セル C70 では

$$0 < z^2 \leq 1$$

の範囲における χ_2^2 分布の確率を求めています．

図の棒グラフは，1000 組×100 回のシミュレーションにより得られた z^2 の頻度分布と χ_2^2 の確率分布です．



$k = 3$ のとき

$$0 \leq z^2 \leq 5.99$$

の範囲の確率は 95%です．$z^2 = 0$からの　χ_2^2 分布の確率を積算することで，$z^2 = 5.99$における累積確率が 95%となります．従って，

$z^2 > 5.99$となる確率は 5%です．

$z^2 > 5.99$において帰無仮説を棄却し，対立仮説を採用します．

5. ウィルコクソンの順位和検定：小標本の場合

ステューデントのt検定のノンパラメトリック版

帰無仮説H_0：$F_1(x) = F_2(x)$
対立仮説H_1：$F_1(x) \neq F_2(x)$

母集団１からのサンプル：$x_{11}, x_{12}, \cdots, x_{1n_1}$
母集団2からのサンプル：$x_{21}, x_{22}, \cdots, x_{2n_2}$

例：$n_1 = 2,\ n_2 = 3$とする．母集団1，2からのサンプルをそれぞれA，Bと記し，小さい順に並べる．BBABAであったとする．
先頭から1, 2, 3, ・・・と順位をつける．サンプルサイズの小さい方の順位の総和 t_1 を求める．
$t_1 = 3 + 5 = 8$

大きい順に並べる：ABABB
先頭から1, 2, 3, ・・・と順位をつける．サンプルサイズの小さい方の順位の総和 t_2 を求める．
$t_2 = 1 + 3 = 4$

t_1，t_2 の内小さい方の値を t 値として検定に用いる．

ステューデントのｔ検定のノンパラメトリック版がウィルコクソンの順位和検定です．
ステューデントのｔ検定は，互いに独立な事象 $X_{11}, X_{12}, \cdots, X_{1N_1} \sim N(\mu_1, \sigma_1^2)$, $X_{21}, X_{22}, \cdots, X_{2N_2} \sim N(\mu_2, \sigma_2^2)$,であり，かつ，$\sigma_1^2 = \sigma_2^2 = \sigma^2$であるときに，

帰無仮説 H_0 ：$\mu_1 = \mu_2$
対立仮説 H_1 ：$\mu_1 \neq \mu_2$

とする検定です．ここで，$X \sim N(\mu, \sigma^2)$は，事象 X が，平均 μ，分散 σ^2 の正規分布に従うことを意味します．
ステューデントの t 検定は事象 X_{1i}，X_{2i} の母平均 μ_1，μ_2 が等しいかどうかの検定です．各事象が正規分布に従い，母分散が等しいと見なせるとき，ステューデントの t 検定を適用できます．もし，各事象は正規分布に従うが，母分散が等しいと見なせないときには，ウェルチの t 検定を適用できます．
各事象の従う分布が正規分布でない場合には上記の検定を適用できません．ウィルコクソンの順位和検定は，各事象の従う分布が正規分布でない場合に適用できる検定法です．

帰無仮説 H_0 ：$F_1(x) = F_2(x)$

対立仮説 H_1 ：$F_1(x) \neq F_2(x)$

$F_1(x)$，$F_2(x)$はそれぞれ事象 X_{1i}，X_{2i} の分布関数です．分布関数 $F(x)$ と確率密度関数 $f(x)$ の関係は

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

です．

母集団１からのサンプル：$x_{11}, x_{12}, \cdots, x_{1n_1}$，母集団２からのサンプル：$x_{21}, x_{22}, \cdots, x_{2n_2}$ とすると，

帰無仮説 H_0 ：サンプル x_{1i}，x_{2i} が同じ分布から抽出された
対立仮説 H_1 ：サンプル x_{1i}，x_{2i} は異なる分布から抽出された

と書き換えることができます．

まず，サンプルサイズが小さい（小標本の）場合です．検定統計量 t をスライド中の例のように決定します．すなわち，母集団１，２からのサンプル全てを値の小さい順に一列に並べます．そして，先頭から 1, 2, 3, …と順位をつけます．母集団１，２からのサンプルのうち，サンプルサイズの小さい方について，順位の総和 t を求めます．

ウィルコクソンの検定：小標本の場合（つづき）

$n_1 = 2, n_2 = 3$の場合の，全組み合わせ．

BBBAA　BBABA　BABBA　ABBBA　BBAAB　BABAB

ABBAB　BAABB　ABABB　AABBB

この内，$t \leq 4$ となるのはAABBB, ABABBのとき．従って，$t \leq 4$ となる確率は 2/10×2 = 4/10.（2倍しているのは両側検定のため）

$\alpha = 0.05$　両側検定

n_1 \ n_2	3	4	5	6	7	8	9	10	11	12	13	14	15
3			0	1	1	2	2	3	3	4	4	5	5
4		0	1	2	3	4	4	5	6	7	8	9	10
5			2	3	5	6	7	8	9	11	12	13	14
6				5	6	8	10	11	13	14	16	17	19
7					8	10	12	14	16	18	20	22	24
8						13	15	17	19	22	24	26	29
9							17	20	23	26	28	31	34
10								23	26	29	33	36	39
11									30	33	37	40	44
12										37	41	45	49
13											45	50	54
14												55	59
15													64

例えば，$n_1 = 2, n_2 = 3$ の場合を考えます．組み合わせの総数が少ないので，全組み合わせを書き出すと

BBBAA　BBABA　BABBA　ABBBA　BBAAB　BABAB　ABBAB　BAABB　ABABB　AABBB

です．$t \leq 4$ となるのは AABBB, ABABB のときです．従って，$t \leq 4$ となる確率は 2/10 × 2 = 4/10 = 0.4 です．確率を 2 倍しているのは両側検定のためです．

小標本の場合は数表が作られています．この表は有意水準 $\alpha = 0.05$ の両側検定のためのものです．小さい方のサンプルサイズを n_1 とし，大きい方のそれを n_2 としています．例えば，$n_1 = 3, n_2 = 5$ のとき，表中の値は 0 です．$n_1 = 3$ のとき，順位和 t の最小値 t_{min} は 6 です．表の数値はこの最小値を引いた値です．$n_1 = 3, n_2 = 5$ のとき，表の値が 0 (すなわち，$t = 6$)である確率は 0.05 以下であることを意味します．$n_1 = 4$ のとき $t_{min} = 10$ です．例えば，$n_1 = 4, n_2 = 5$ のとき，表の値は 1 ($t = 11$)です．この例のサンプルサイズでは $t \leqq 11$ となる確率が 0.05 以下です．

ウィルコクソンの順位和検定：大標本の場合

帰無仮説H_0：$F_1(x) = F_2(x)$
対立仮説H_1：$F_1(x) \neq F_2(x)$

母集団1からのサンプル：$x_{11}, x_{12}, \cdots, x_{1n_1}$
母集団2からのサンプル：$x_{21}, x_{22}, \cdots, x_{2n_2}$

1. 小さい順に並べて順位をつける.
2. サンプルサイズの小さい方の順位の総和t_1を求める.
3. 大きい順に並べて順位をつける.
4. サンプルサイズの小さい方の順位の総和t_2を求める.
5. t_1，t_2 の内小さい方の値を t 値として検定に用いる.

t 値の平均値：$\mu = \frac{n_1(n_1+n_2+1)}{2}$

t 値の標準偏差：$\sigma = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$

順位和をサンプルサイズで割ったものはその群の平均順位です．サインプルサイズが大きくなれば（大標本となれば）中心極限定理により順位和 t は正規分布に従います．
n_1, n_2 が数表にないほどに大きい場合，t 値は，平均値

$$\mu = \frac{n_1(n_1+n_2+1)}{2}$$

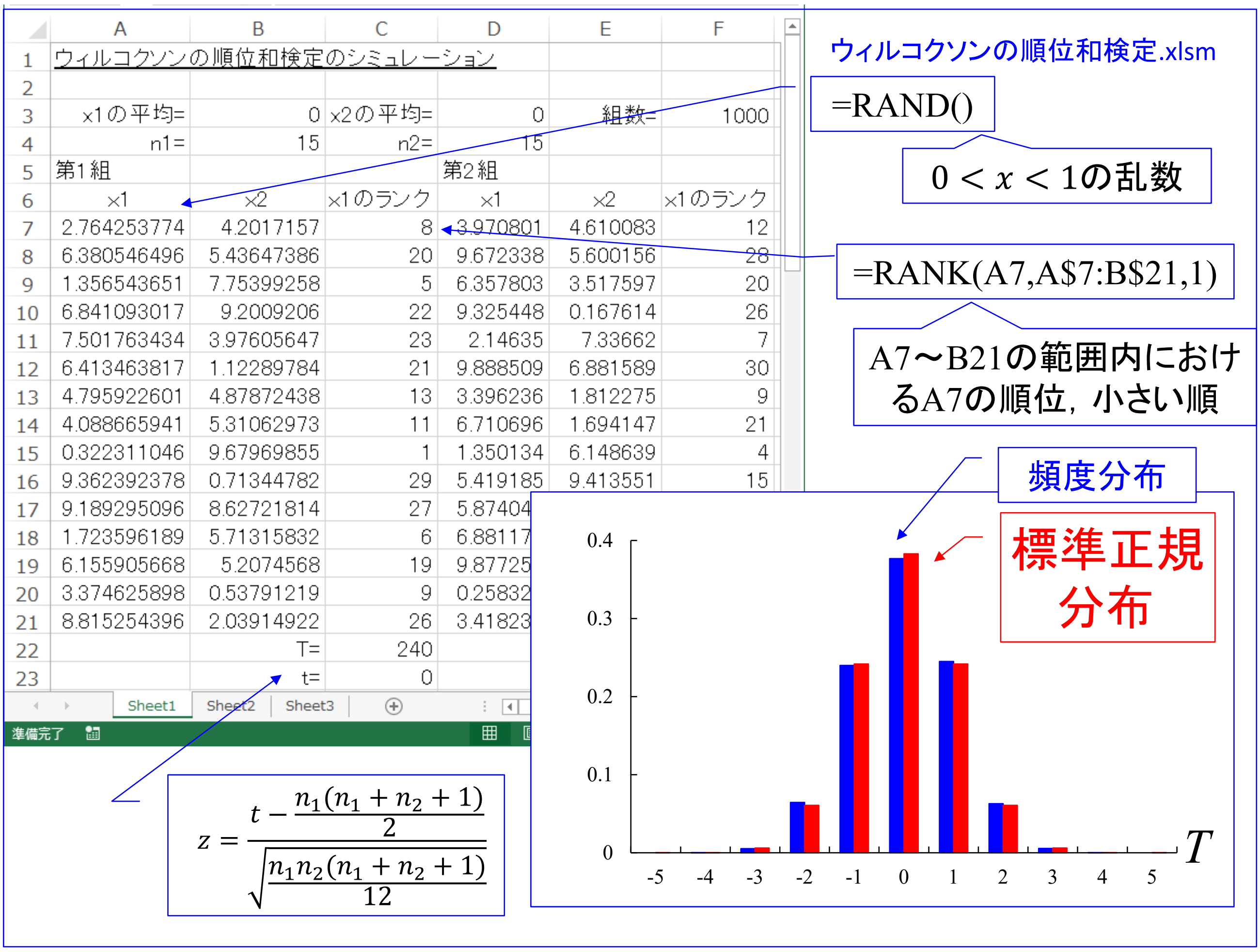
標準偏差

$$\sigma = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$$

の正規分布にほぼ従います．従って，

$$z = \frac{t - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

は平均 0, 分散 1 の標準正規分布にほぼ従います.



	A	B	C	D	E	F
1	ウィルコクソンの順位和検定のシミュレーション					
2						
3	x1の平均=	0	x2の平均=	0	組数=	1000
4	n1=	15	n2=	15		
5	第1組			第2組		
6	x1	x2	x1のランク	x1	x2	x1のランク
7	2.764253774	4.2017157	8	3.970801	4.610083	12
8	6.380546496	5.43647386	20	9.672338	5.600156	28
9	1.356543651	7.75399258	5	6.357803	3.517597	20
10	6.841093017	9.2009206	22	9.325448	0.167614	26
11	7.501763434	3.97605647	23	2.14635	7.33662	7
12	6.413463817	1.12289784	21	9.888509	6.881589	30
13	4.795922601	4.87872438	13	3.396236	1.812275	9
14	4.088665941	5.31062973	11	6.710696	1.694147	21
15	0.322311046	9.67969855	1	1.350134	6.148639	4
16	9.362392378	0.71344782	29	5.419185	9.413551	15
17	9.189295096	8.62721814	27	5.87404		
18	1.723596189	5.71315832	6	6.88117		
19	6.155905668	5.2074568	19	9.87725		
20	3.374625898	0.53791219	9	0.25832		
21	8.815254396	2.03914922	26	3.41823		
22		T=	240			
23		t=	0			

n_1=15, n_2=15 の場合について，z 値を求め，その頻度分布を求めるシミュレーションを行うエクセル画面のスナップショットです．「ウィルコクソンの順位和検定.xlsm」ファイルを開いてください．

RAND()により，$0 < x < 1$の範囲の乱数を生成します．

セル C8 の RANK(A7, A$7:B$21, 1) により，セル A7 の数値がセル A7〜B21 の範囲内の数値の中で小さい方から何番目にあるかを求めます．A$7 としておくと，セル A7 の関数がセル F7 にコピーされた際に，A$7 → F$7 と自動的に列記号は更新されますが，行番号の 7 は固定となり，更新されません．もし，$A$7 としておくと，A も固定となります．関数の最後の引数 1 は，数値の小さい方から 1, 2, 3, ・・・と順位をつけること（昇順）を指定します．この引数を 0 とすると，数値の大きい方から順位をつけること（降順）を指定します．

マクロを実行すると，23 行目の z 値の頻度分布が得られます．頻度分布は標準正規分布にほぼ等しい分布となることが分かります．

同順位（タイ, ties）がある場合

サンプル値の中に同じ値がある場合，順位付はどうすればよいでしょうか？

例：母集団1からのサンプル：$\{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\} = \{2, 3, 7, 7, 9\}$
　　母集団2からのサンプル：$\{x_{21}, x_{22}, x_{23}, x_{24}, x_{25}\} = \{1, 3, 7, 11, 15\}$
であったとします．サンプル値の昇順に並べます．ただし，同順位の組の中の順番はどちらでも良いとします．順位は以下の通りにつけます．

	x_{21},	x_{11},	x_{12},	x_{22} ,	x_{13},	x_{14},	x_{23},	x_{15},	x_{24} ,	x_{25}
値	1	2	3	3	7	7	7	9	11	15
順位	1	2	3.5	3.5	6	6	6	8	9	10

すなわち，同じ値が無かったとしたときにつけられるはずの順位の平均値がつけられます．

31

サンプル値の中に同じ値がある場合，順位付はどうすれば良いでしょうか？具体例で説明します

母集団１からのサンプル：$\{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\} = \{2,\ 3,\ 7,\ 7,\ 9\}$
母集団２からのサンプル：$\{x_{21}, x_{22}, x_{23}, x_{24}, x_{25}\} = \{1,\ 3,\ 7,\ 11,\ 15\}$

であったとします．

これらサンプルを値の昇順に並べます．ただし，同順位の場合はどちらが先であってもかまわないとします．同順位のサンプルの順位は，同じ値が無かったとしたときにつけられるはずの順位の平均値をつけます．すなわち

	x_{21},	x_{11},	x_{12},	x_{22},	x_{13},	x_{14},	x_{23},	x_{15},	x_{24},	x_{25}
値	1	2	3	3	7	7	7	9	11	15
順位	1	2	3.5	3.5	6	6	6	8	9	10

とします．

同順位がある場合，かつ大標本の場合

t 値の平均値：$\mu = \frac{n_1(n_1+n_2+1)}{2}$

t 値の標準偏差：

$$\sigma = \sqrt{\frac{n_1 n_2}{12n(n-1)}\left(n^3 - n - \sum_{i=1}^{s}\left(d_i^3 - d_i\right)\right)}$$

$n = n_1 + n_2$

s :同順位となる組数

d_i: 同順位となる組 i に属するサンプルの数

前スライドの例では，$s = 2, d_1 = 2, d_3 = 3$ です. 32

サンプルサイズが大きくなれば（大標本となれば）中心極限定理により順位和 t は正規分布に従います．t 値は，平均値

$$\mu = \frac{n_1(n_1 + n_2 + 1)}{2}$$

標準偏差

$$\sigma = \sqrt{\frac{n_1 n_2}{12n(n-1)}\left(n^3 - n - \sum_{i=1}^{s}\left(d_i^3 - d_i\right)\right)}$$

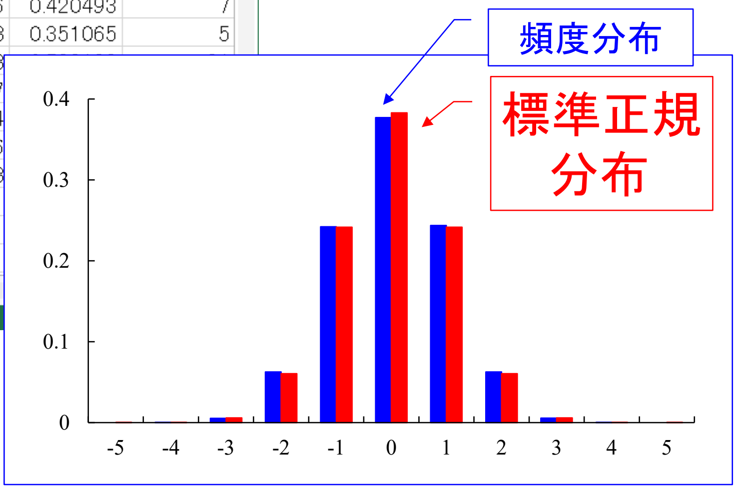
の正規分布にほぼ従います．ただし，$n = n_1 + n_2, s$:同順位となる組数，d_i: 同順位となる組 i に属するサンプルの数です．前スライドの例では，$s = 2$, $d_1 = 2$, $d_2 = 3$ です．

標準偏差は同順位があるために少し小さくなります．

	A	B	C	D	E	F
1	ウィルコクソンの順位和検定のシミュレーション（同順位がある場合）					
2						
3	n1	n2	d1	d2	d3	組数
4	15	15	3	2	2	1000
5	第1組				第2組	
6	x1	x2	x1のランク	x1	x2	x1のランク
7	0.102896869	0.98508036	7	0.070108	0.211554	1
8	0.102896869	0.25290704	7	0.211554	0.814214	10
9	0.102896869	0.53077877	7	0.110195	0.211554	4
10	0.014589727	0.33850908	2.5	0.792875	0.956959	20
11	0.920977294	0.01458973	28	0.956959	0.55453	28.5
12	0.840686321	0.51153255	23.5	0.101677	0.146261	2.5
13	0.840686321	0.50391614	23.5	0.294229	0.101677	12
14	0.479321539	0.9063158	16	0.192352	0.666516	8
15	0.253201663	0.26643813	11	0.149726	0.420493	7
16	0.252546608	0.65426755	9	0.134498	0.351065	5
17	0.056465566	0.965289	4	0.813118		
18	0.080066264	0.64949179	5	0.761767		
19	0.887443244	0.27240479	26	0.369184		
20	0.0085724	0.56441283	1	0.96216		
21	0.291361749	0.86460435	14	0.934928		
22		t	184.5			
23		z=	-2			
24		補正z=	-2			

補正z

$$= \frac{t - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2}{12n(n-1)}\left(n^3 - n - \sum_{i=1}^{s}\left(d_i^3 - d_i\right)\right)}}$$



「ウィルコクソンの順位和検定（同順位のある場合）.xlxm」ファイルを開いてください．メニューバーから「開発」→「マクロ」→「編集」と左クリックすると，VBA のマクロが表示されます．このマクロは第 1 組から第 1000 組までそれぞれ母集団 1 からのサンプル 15 個，母集団 2 からのサンプル 15 個の総計 1,000 組×30 個=30,000 個の乱数を生成します．その直後に同順位データを生成して上書きします．同順位となるサンプルの数はセル C4, D4, E4 で設定しています．これらセルの数値は，マクロの実行前に，人手で書き換えて設定します．d1 + d2 + d3 ≦ mini(n1, n2)とする必要があります．マクロはこれらのセルの値を読み込んで，まず，乱数を一つ生成し，セル A7 もしくは B7 のいずれかをその生成した値で書き換えます．いずれの選択確率も 0.5 です．次に，セル A8 もしくは B8 のいずれかを最初に生成したものと同じ値で書き換えます．以降，A9 or B9, ... と d1 で指定され行数について書き換えていきます．次は，新たな乱数を一つ生成して，d2 で指定された行数だけ同様に書き換えます．最後に，もう一つ新しく乱数を一つ生成して，d3 で指定された行数だけ同様に書き換えます．このスライドの例では A7,A8,A9 が同順位となり，次は A10, B11 が同順位となり，最後は A12, A13 が同順位となっています．

以降マクロは第 2 組～第 1000 組について，同様の書き換えを行っていきます．

24 行目のセルでは同順位を考慮した場合の z 値

$$補正z = \frac{t - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2}{12n(n-1)}(n^3 - n - \sum_{i=1}^{s}(d_i^3 - d_i))}}$$

を計算しています.

エクセルシートは 1000 組の補正 z 値の頻度分布を棒グラフに描きます．以上を 1 回の試行として，マクロは 1000 組×100 回の試行の補正 z 値の頻度分布を求めた棒グラフに描きます.

100 試行の計算には時間がかかります．セル C40 に今，何試行目であるかが表示されます．スライドの右下には得られた頻度分布の例を示します．ほぼ標準正規分布に近い分布が得られています.

なお，この例では，補正項 $\sum_{i=1}^{s}(d_i^3 - d_i)$ の値は $3^3 - 3 + 2^3 - 2 + 2^3 - 2 = 36$ です．一方，$n^3 - n = 30^3 - 30 = 26970$ です．補正項により標準偏差は 0.07%大きくなります.

6. マン・ホイットニーのU検定：小標本の場合

ステューデントのt検定のノンパラメトリック版

帰無仮説H_0：$F_1(x) = F_2(x)$
対立仮説H_1：$F_1(x) \neq F_2(x)$

母集団1からのサンプル：$x_{11}, x_{12}, \cdots, x_{1n_1}$
母集団2からのサンプル：$x_{21}, x_{22}, \cdots, x_{2n_2}$

例：$n_1 = 2$, $n_2 = 3$とする．母集団1，2からのサンプルをそれぞれA，Bと記す．
小さい順に並べる：BBABA

Aより右にあるBの数を各Aについて数え，総和を求める．

$u_1 = 1 + 0 = 1$
Aより左にあるBの数を各Aについて数え，総和を求める．
$u_2 = 2 + 3 = 5$

値の小さい方の u 値を検定に用いる．

ステューデントの t 検定のノンパラメトリック版にはマン・ホイットニーの U 検定もあります．

　ウィルコクソンの順位和検定とマン・ホイットニーの U 検定は本質的に同じ検定法です．ほぼ同時期に発表されています[1][2]．両者は検定統計量の求め方が異なります．
　検定統計量 u はスライド中の例のように決定します．

[1] Mann, H. B. and Whitney, D. R., “On a test of whether one of two random variables is stochastically larger than the other,” The Annals of Mathematical Statistics, Vol. 18, no. 1 (Mar., 1947), pp. 50-60
[2] Wilcoxon, F., “Individual Comparison by Ranking Method,” Biometrics Bulletin, Vol. 1, no. 6, (Dec., 1945), pp.80-83

マン・ホイットニーのU検定：大標本の場合

1. 小さい順に並べて順位をつける
2. サンプル1，2群それぞれの順位の総和 t_1, t_2 を求める.
（この t 値はウィルコクソンの順位和検定の t 値と同じ. ）
3. u_1, u_2 を次式により求める.

$$u_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - t_1$$

$$u_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - t_2$$

4. u 値の

平均値：$\mu = \dfrac{n_1 n_2}{2}$

標準偏差：$\sigma = \sqrt{\dfrac{n_1 n_2(n_1+n_2+1)}{12}}$.

35

大標本の場合

$$平均値：\mu = \frac{n_1 n_2}{2}$$

$$標準偏差：\sigma = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}.$$

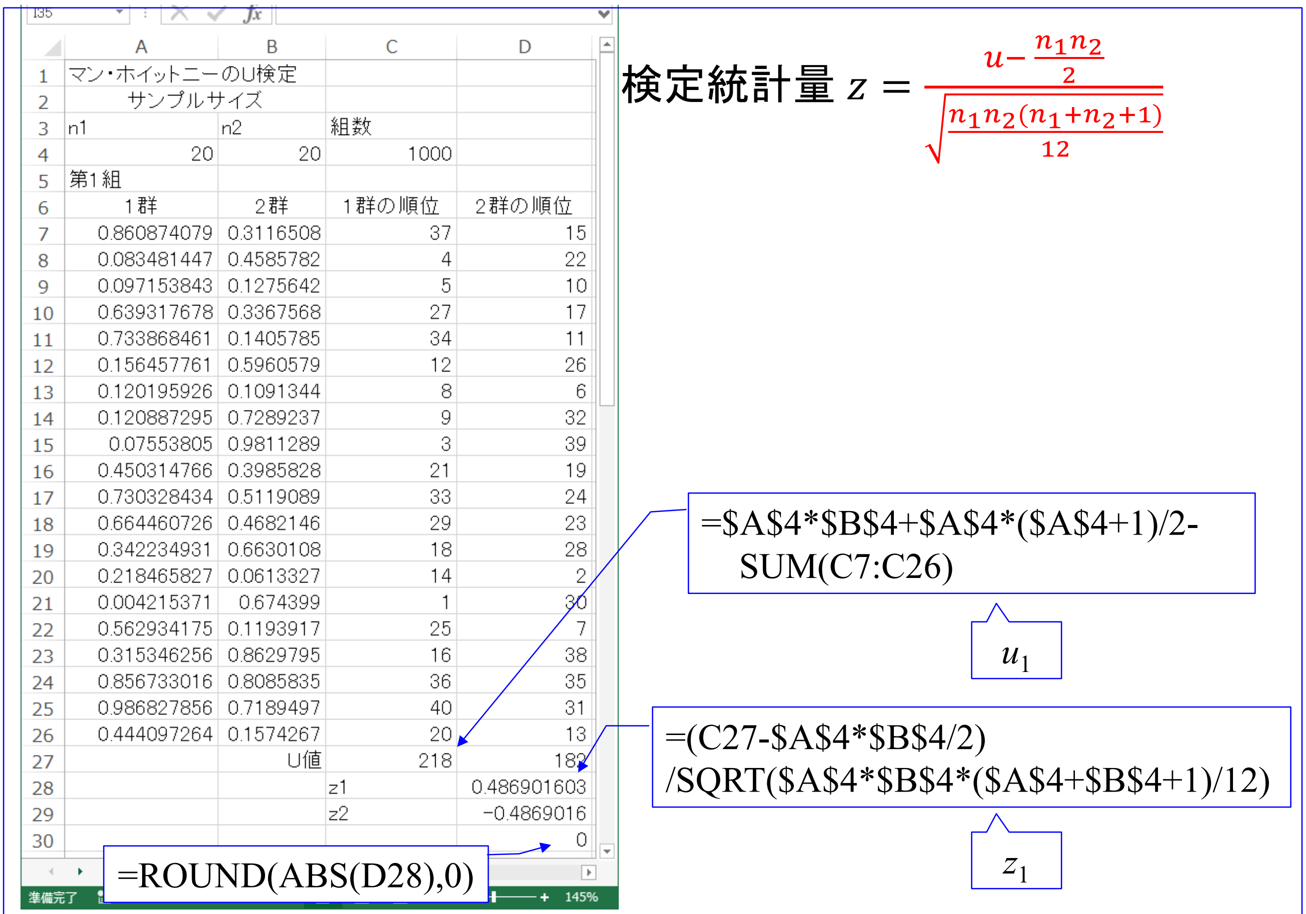
の正規分布に従います．従って，

$$z = \frac{u - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}}$$

は平均 0，分散 1 の標準正規分布にほぼ従います．

u_1，u_2は，順位和 t_1，t_2 を定数から引いた値です．よって，u 値と t 値は平均値が異なり，標準偏差は同じになります．

$$
\text{検定統計量 } z = \frac{u - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}
$$



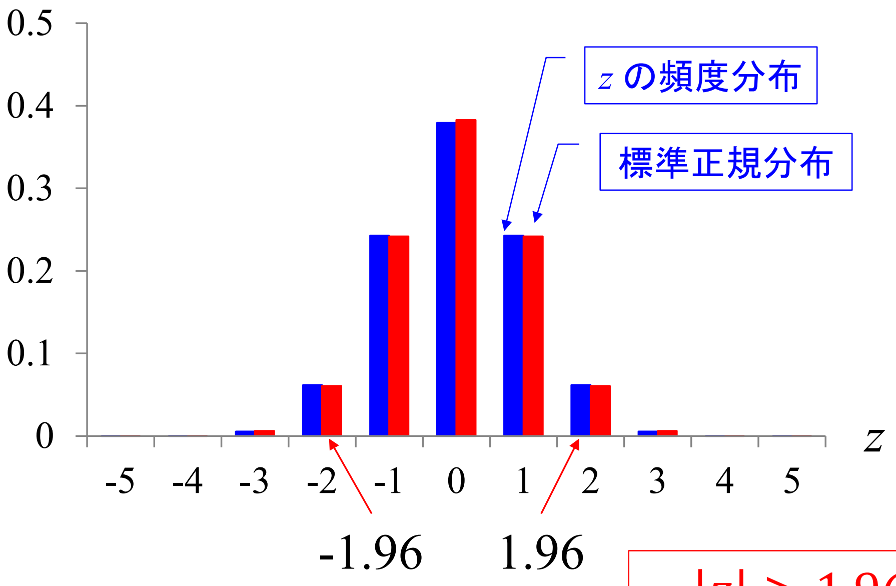
	A	B	C	D
1	マン・ホイットニーのU検定			
2	サンプルサイズ			
3	n1	n2	組数	
4	20	20	1000	
5	第1組			
6	1群	2群	1群の順位	2群の順位
7	0.860874079	0.3116508	37	15
8	0.083481447	0.4585782	4	22
9	0.097153843	0.1275642	5	10
10	0.639317678	0.3367568	27	17
11	0.733868461	0.1405785	34	11
12	0.156457761	0.5960579	12	26
13	0.120195926	0.1091344	8	6
14	0.120887295	0.7289237	9	32
15	0.07553805	0.9811289	3	39
16	0.450314766	0.3985828	21	19
17	0.730328434	0.5119089	33	24
18	0.664460726	0.4682146	29	23
19	0.342234931	0.6630108	18	28
20	0.218465827	0.0613327	14	2
21	0.004215371	0.674399	1	30
22	0.562934175	0.1193917	25	7
23	0.315346256	0.8629795	16	38
24	0.856733016	0.8085835	36	35
25	0.986827856	0.7189497	40	31
26	0.444097264	0.1574267	20	13
27		U値	218	182
28			z1	0.486901603
29			z2	-0.4869016
30				0

マン・ホイットニーのU検定.xlsm

「マン・ホイットニーの U 検定.xlsm」を開いてください．母集団 1，2 からのサンプルをそれぞれ 1 群，2 群として小さい順に順位をつけ，u 値，z 値を求めることを 1000 組について実行し，z 値の頻度分布を描くエクセルファイルです．

検定統計量 z

➡ 中心極限定理により標準正規分布となる.



|z| > 1.96で有意差あり

　図は 1000 組×100 回のシミュレーションを繰り返したときの z の頻度分布と標準正規分布です.

　検定統計量 z がほぼ標準席分布となることが分かります．有意水準 5%の両側検定では，|z| > 1.96 のときに有意差ありとして帰無仮説を棄却できます.

同順位がある場合，かつ大標本の場合

u 値の平均値 : $\mu = \dfrac{n_1 n_2}{2}$

u 値の標準偏差 :

$$\sigma = \sqrt{\frac{n_1 n_2}{12n(n-1)}\left(n^3 - n - \sum_{i=1}^{s}\left(d_i^3 - d_i\right)\right)}$$

$n = n_1 + n_2$
s :同順位となる組数
d_i: 同順位となる組 i に属するサンプルの数

38

マン・ホイットニーの U 検定において，同順位がある場合の扱いは，ウィルコクソンの順位和検定の場合と同じです．大標本の場合，u 値は，平均値

$$\mu = \frac{n_1 n_1}{2}$$

標準偏差

$$\sigma = \sqrt{\frac{n_1 n_2}{12n(n-1)}\left(n^3 - n - \sum_{i=1}^{s}\left(d_i^3 - d_i\right)\right)}$$

の正規分布にほぼ従います．ただし，$n = n_1 + n_2$，s :同順位となる組数，d_i: 同順位となる組 i に属するサンプルの数です．

標準偏差は同順位があるために少し小さくなります．

7. ウィルコクソンの符号付き順位和検定：小標本の場合（対応のある場合）

対応のあるステューデントのt検定のノンパラメトリック版

帰無仮説 $H_0 : F_1(x) = F_2(x)$
対立仮説 $H_1 : F_1(x) \neq F_2(x)$

製品no.: 1 2 n
案1のサンプル：$x_{11}, x_{12}, \cdots, x_{1n}$
案2のサンプル：$x_{21}, x_{22}, \cdots, x_{2n}$

1. 製品ごとにサンプル値の差 $(x_{1i} - x_{2i})$ を求める.
2. 差の絶対値 $|x_{1i} - x_{2i}|$ をとる.
3. 各絶対値の n 個の中での順位を求める.
4. 差がマイナスである製品について順位の総和 t を求める.

対応のあるステューデントの t 検定のノンパラメトリック版がウィルコクソンの符号付き順位和検定です.

帰無仮説，対立仮説はウィルコクソンの順位和検定と同じです.

帰無仮説 H_0 ： $F_1(x) = F_2(x)$
対立仮説 H_1 ： $F_1(x) \neq F_2(x)$

ウィルコクソンの符号付き順位和検定がウィルコクソンの順位和検定と異なるのは, 例えば

製品 no.: 1 2 ･･･ n
案１のサンプル：x_{11}， $x_{12}, \cdots$， x_{1n}
案２のサンプル：x_{21}， $x_{22}, \cdots$， x_{2n}

のように, 案１と案２に対応がある点です. この例では n 個の製品それぞれに案１，２を適用して，サンプル x_{1i}，x_{2i} を抽出しています．製品ごとにサンプル値が大きく異なるとすると，案１，２の

効果の違いに有意差が見られない結果となってしまうことがあります．サンプル値の差$(x_{1i} - x_{2i})$を求めれば，製品の違いによる影響を取り除くことができます．

ウィルコクソンの符号付き順位和検定の t 値はスライドに記した手順により求めます．

例：$n = 6$とする．　案Aの値が小さい（差がマイナスの）場合をA，案Bが小さい場合をBと表す．

案Aの値がすべて大きければ，絶対値の小さい順に並べると

BBBBBB

と表される．このとき，$t = 0$.

案Aの値が小さい場合が1つだけの場合：

	ABBBBB	BABBBB	BBABBB	BBBABB	・・・
$t=$	1	2	3	4	

案Aの値が小さい場合が2つの場合：

	ABABBB	ABBABB	ABBBAB	ABBBBA	・・・
$t=$	4	5	6	7	

案Aの値がすべて小さければ

AAAAAA

このとき $t = \frac{n(n+1)}{2} = 21$.

サンプルサイズ $n = 6$ の場合の例を示します．案 A の値が小さい（差がマイナスの）場合を A，案 B が小さい場合を B と表します．

案 A の値がすべて大きければ，絶対値の小さい順に並べると

BBBBBB

と表されます．このとき，$t = 0$ です．

スライドには A, B の組み合わせとそのときの t 値の一部を示してあります．
案 A の値がすべて小さければ

AAAAAA

であり，このとき $t = \frac{n(n+1)}{2} = 21$です．

$n = 6$のとき

$t = 0$ or 21 となる確率

全 2^6 通りの内 2 通り.

$$\frac{2}{2^6} = 0.03 < 0.05$$

(両側検定)有意差あり.

$t \leq 1$ or $t \geq 20$となる確率

全 2^6 通りの内 4 通り.

$$\frac{4}{2^6} = 0.0625 > 0.05$$

(両側検定)有意差無し.

Wilcoxonの符号付き順位和検定	
n	両側検定5%で有意となるt値
6	0
7	2
8	4
9	6
10	8
11	11
12	14
13	17
14	21
15	25
16	30
17	35
18	40
19	46
20	52
21	59
22	66
23	73
24	81
25	89

サンプルサイズ $n = 6$ のとき，$t = 0$ or 21 となる確率は，全 2^6 通りの組み合わせの内の 2 通りなので，

$$\frac{2}{2^6} = 0.03 \quad < 0.05$$

です．もし，t 値がこれらの値となれば， 両側検定で有意差ありです．

$t \leq 1$ or $t \geq 20$となる確率は，全 2^6 通りの内の 4 通りなので，

$$\frac{4}{2^6} = 0.0625 \quad > 0.05$$

です．$t = 1$ or 20 の場合は，有意差はありません．

$n > 25$のとき: 大標本の場合

t 値の平均値

$$\mu = \frac{n(n+1)}{4}$$

t 値の分散

$$\sigma^2 = \frac{n(n+1)(2n+1)}{24}$$

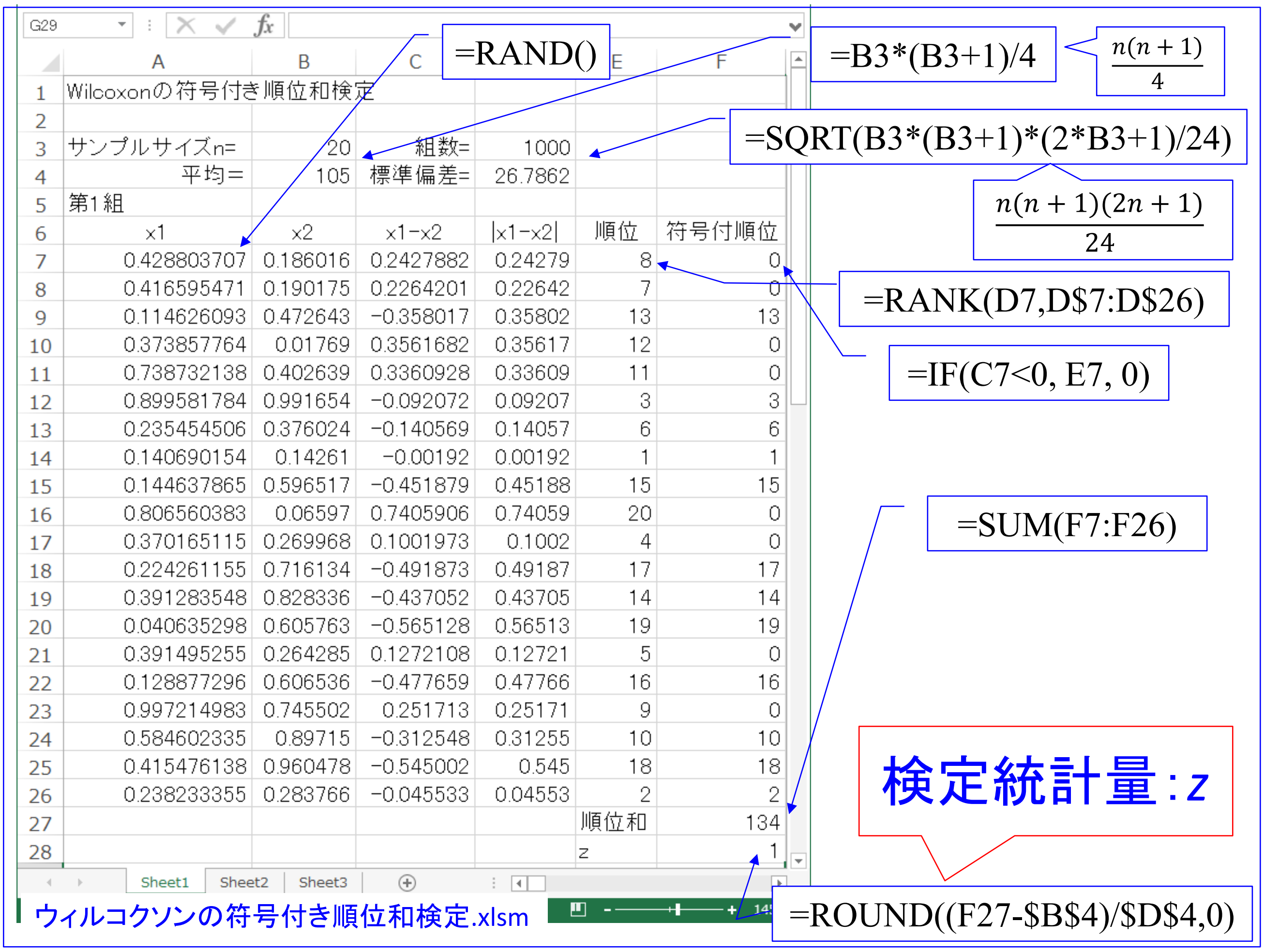
大標本の場合には，t 値の平均値は

$$\mu = \frac{n(n+1)}{4}$$

となり，分散は

$$\sigma^2 = \frac{n(n+1)(2n+1)}{24}$$

となります．



	A	B	C	D	E	F
1	Wilcoxonの符号付き順位和検定					
2						
3	サンプルサイズn=	20	組数=	1000		
4	平均=	105	標準偏差=	26.7862		
5	第1組					
6	x1	x2	x1−x2	|x1−x2|	順位	符号付順位
7	0.428803707	0.186016	0.2427882	0.24279	8	0
8	0.416595471	0.190175	0.2264201	0.22642	7	0
9	0.114626093	0.472643	−0.358017	0.35802	13	13
10	0.373857764	0.01769	0.3561682	0.35617	12	0
11	0.738732138	0.402639	0.3360928	0.33609	11	0
12	0.899581784	0.991654	−0.092072	0.09207	3	3
13	0.235454506	0.376024	−0.140569	0.14057	6	6
14	0.140690154	0.14261	−0.00192	0.00192	1	1
15	0.144637865	0.596517	−0.451879	0.45188	15	15
16	0.806560383	0.06597	0.7405906	0.74059	20	0
17	0.370165115	0.269968	0.1001973	0.1002	4	0
18	0.224261155	0.716134	−0.491873	0.49187	17	17
19	0.391283548	0.828336	−0.437052	0.43705	14	14
20	0.040635298	0.605763	−0.565128	0.56513	19	19
21	0.391495255	0.264285	0.1272108	0.12721	5	0
22	0.128877296	0.606536	−0.477659	0.47766	16	16
23	0.997214983	0.745502	0.251713	0.25171	9	0
24	0.584602335	0.89715	−0.312548	0.31255	10	10
25	0.415476138	0.960478	−0.545002	0.545	18	18
26	0.238233355	0.283766	−0.045533	0.04553	2	2
27					順位和	134
28					z	1

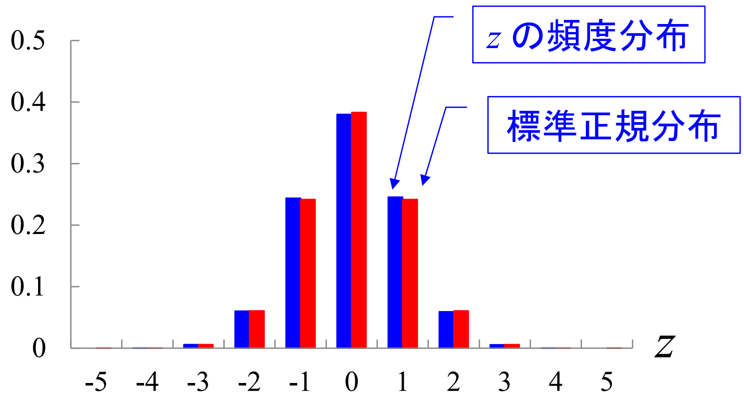
「ウィルコクソンの符号付き順位和検定.xlsm」のエクセルファイルを開いて下さい．

セル A7 から B26 では，それぞれ $0 < x < 1$ の範囲の乱数を生成しています．これら乱数は行ごとに対応があるとしています．C 列で 1 群と 2 群のサンプル値の差を行ごとに求め，D 列でその絶対値を求めています．E 列では D 列の値に対して大きい順に番号を付けています．（小さい順ではありませんが，本質的な違いはありません．）F 列では C 列で負の値となった場合だけ E 列の順位をそのままコピーしています．セル F27 で F 列の順位の総和を求めています．セル F28 では，検定統計量

$$z = \frac{t - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$$

を計算しています．

検定統計量: $z = \dfrac{t - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$



$|z| > 1.96$で有意差あり

1000 組について z 値を求め，その頻度分布を求め，さらに 1000 組×100 回のシミュレーション結果の頻度分布がスライドの図です． z の頻度分布はほぼ標準正規分布に従うことが分かります．

同値，同順位（タイ, ties）がある場合

サンプル対の中に同じ値がある場合，また，サンプル対の差の絶対値に同順位がある場合，順位付はどうすればよいでしょうか？

例：案1からのサンプル：$\{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\} = \{2, 3, 7, 7, 9\}$
　　案2からのサンプル：$\{x_{21}, x_{22}, x_{23}, x_{24}, x_{25}\} = \{1, 3, 7, 8, 15\}$

であったとします．各対の差の絶対値は

$$\{|x_{11} - x_{21}|, |x_{12} - x_{22}|, |x_{13} - x_{23}|, |x_{14} - x_{24}|, |x_{15} - x_{25}|\}$$
$$= \{1, 0, 0, 1, 6\}$$

となります．差が0の対は順位付から外します．

	$\lvert x_{11} - x_{21}\rvert$,	$\lvert x_{14} - x_{24}\rvert$,	$\lvert x_{15} - x_{25}\rvert$
差の絶対値	1	1	6
順位	1.5	1.5	3

同順位の場合，同順位が無かったとしたときにつけられるはずの順位の平均値がつけられます．

サンプル値の中に同じ値がある場合，順位付はどうすれば良いでしょうか？スライドでは具体例で説明してあります．スライドの説明を読んでください．

同順位がある場合，かつ大標本の場合

t 値の平均値と分散

$$\mu = \frac{n'(n'+1)}{4}$$

$$\sigma^2 = \frac{n'(n'+1)(2n'+1)}{24} - \sum_{i=1}^{s} \frac{d_i\left(d_i^2-1\right)}{48}$$

$n' = n - d_0$
d_0:差の無い対の個数
s :同順位となる組数
d_i: 同順位となる組 i に属するサンプル対の数

ウィルコクソンの符号付き順位和検定において，n 個の対があるとします．その中で，差の値が 0 の対が　d_0 個あったとします．これらの対は検定の対象から除去しますので，残った対数 $n' = n - d_0$ です．次に，絶対値が同じ値である対を一つの組にまとめます．

大標本の場合の t 値の平均値と分散は

$$\mu = \frac{n'(n'+1)}{4}$$

$$\sigma^2 = \frac{n'(n'+1)(2n'+1)}{24} - \sum_{i=1}^{s} \frac{d_i\left(d_i^2-1\right)}{48}$$

となります．ただし，$n' = n - d_0$，d_0:差の無い対の個数，s :同順位となる組数，d_i: 同順位となる組 i に属するサンプル対の数です．

8. クルスカル・ワリスの検定：小標本の場合

一元配置分散分析のノンパラメトリック版

k 群の互いに独立な標本

帰無仮説 $H_0: F_1(x) = F_2(x) = \cdots = F_k(x)$

対立仮説 $H_1: F_1(x) \neq F_2(x)\ or\ F_1(x) \neq F_3(x)\ or\ \cdots or\ F_{k-1}(x) \neq F_k(x)$

第1群からのサンプル：$x_{11}, x_{12}, \cdots, x_{1n_1}$

第2群からのサンプル：$x_{21}, x_{22}, \cdots, x_{2n_2}$

⋮

第k群からのサンプル：$x_{21}, x_{22}, \cdots, x_{2n_k}$

$n\left(= \sum_{i=1}^{k} n_i\right)$個の観測値に順位をつける.

群ごとに順位の総和 t_i を求める.

検定統計量：

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{t_i^2}{n_i} - 3(n+1)$$

一元配置分散分析のノンパラメトリック版がクルスカル・ワリスの検定です.

これまでの検定は 2 群の分布間の差の検定でした．ここでは，k 群の分布間に差があるかを検定します.

帰無仮説は「k 群の分布間に差が無い」であり，対立仮説は「k 群のいずれかの分布間に差がある」です.

検定統計量は全サンプルを値の小さい順に並べて順位を付けます．そして，群ごとに順位の総和 t_i を求めます．検定統計量は

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{t_i^2}{n_i} - 3(n+1)$$

です.

$n_1 = 3, n_2 = 2, n_3 = 2$とする．第1, 2, 3群に所属するサンプルをそれぞれA, B, Cとする．

AAABBCC $t_1 = 6, t_2 = 9, t_3 = 13,$

$$h = 0.214\left(\frac{36}{3} + \frac{81}{2} + \frac{169}{2}\right) - 24 = 5.357$$

BBCCAAA $t_1 = 18, t_2 = 3, t_3 = 7, h = 5.357$

BBAAACC $t_1 = 12, t_2 = 3, t_3 = 13, h = 5.357$

全組み合わせ $\frac{7!}{3!2!2!} = 210$通り．

h_{max}となる組み合わせ : 3群の順列 = 6通り

$$p = \frac{6}{210} = 0.029$$

例えば，$n_1 = 3, n_2 = 2, n_3 = 2$とします．第 1，２，３群に所属するサンプルをそれぞれ A, B, C とします．

スライドには 3 通りの極端な例を示してあります．これらの検定統計量 h はいずれも 5.357 です．A が３つ，B が２つ，C が２つのとき，これらの並べ方は，7!/3!2!2! = 210 通りです．７つの記号が全て異なるとすれば，並べ方は順列の 7! 通りです．この並べ方の内，重複が A で 3! 通り，B で 2! 通り，C で 2! 通りあります．この重複分を除くことは，7! を 3!2!2! で割ることでできます．

$h_{max} = 5.357$ となる組み合わせは 6 通りあります．従って，h_{max} の出現する確率は 6/210 = 0.029 です．

AAABCBC $t_1 = 7, t_2 = 10, t_3 = 12,$

$$h = 0.214\left(\frac{49}{3} + \frac{100}{2} + \frac{144}{2}\right) - 24 = 4.714$$

AAACBCB　BCBCAAA　CBCBAAA

$h = 4.714$となるのは以上の4通り.

$h \geq 4.714$となるのは10通り.

$$p = \frac{10}{210} = 0.048$$

サンプルサイズ			h	p
n_1	n_2	n_3		
2	1	1	2.7	0.5
2	2	1	3.6	0.2
2	2	2	4.571	0.067
			3.714	0.2
3	1	1	3.2	0.3
3	2	1	4.286	0.1
			3.857	0.133
3	2	2	5.357	0.029
			4.714	0.048
			4.5	0.067

この例(AAABCBC)では，h は前スライドの例より小さくなります.

$h = 4.714$ となるのは，AAABCBC, AAACBCB, BCBCAAA, CBCBAAA の 4 通りあります．$h \geq 4.714$ となるのは，前スライドの例と合わせて 10 通りです．従って，$h \geq 4.714$ となる確率は 10/210 = 0.048 です.

スライド中の表はサンプルサイズが小さい場合に，h の取り得る値の内，大きいほうの値と，その値が出現する確率を計算した例です.

例えば，$n_1 = 2, n_2 = 1,\ n_3 = 1$ の場合，A, A, B, C の並びは全部で 4!/2!1!1! = 12 通りです．これらの内 $h = 2.7$ となるのは AABC, AACB, BAAC, CAAB, BCAA, CBAA の 6 通りです． 従って，$h = 2.7$ となる確率 $p = 0.5$ です．$n_1 = 2, n_2 = 1,\ n_3 = 1$ の場合は p 値が 0.05 を下回る並びはありません．この場合は検定をするまでもなく有意差のある結果を得ることはありません.

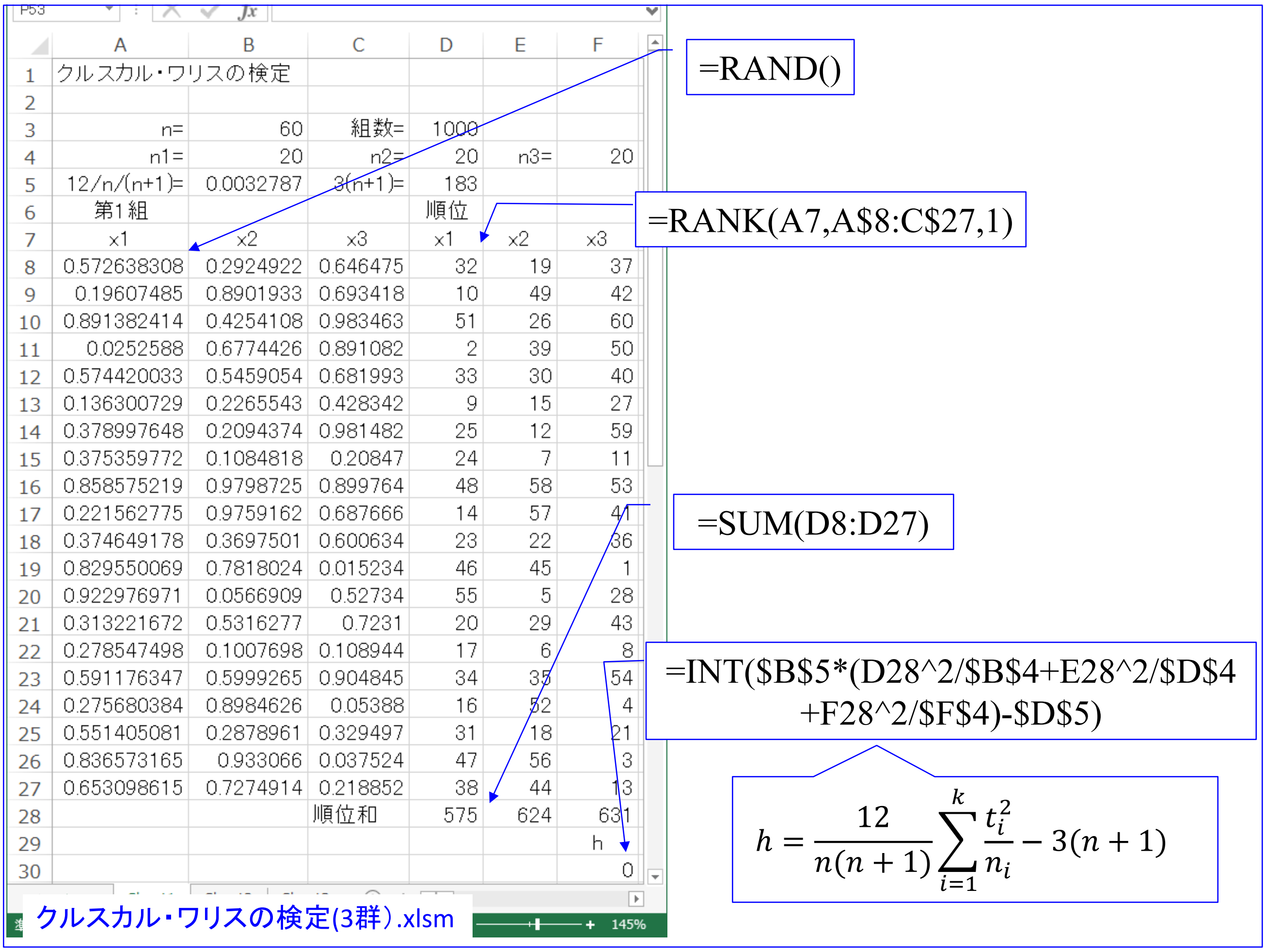
$n_1 = 3, n_2 = 2,\ n_3 = 2$ の場合において，$h \geq 4.714$ であれば，有意水準 5% で帰無仮説を棄却できます.

クルスカル・ワリスの検定：大標本の場合

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{t_i^2}{n_i} - 3(n+1)$$

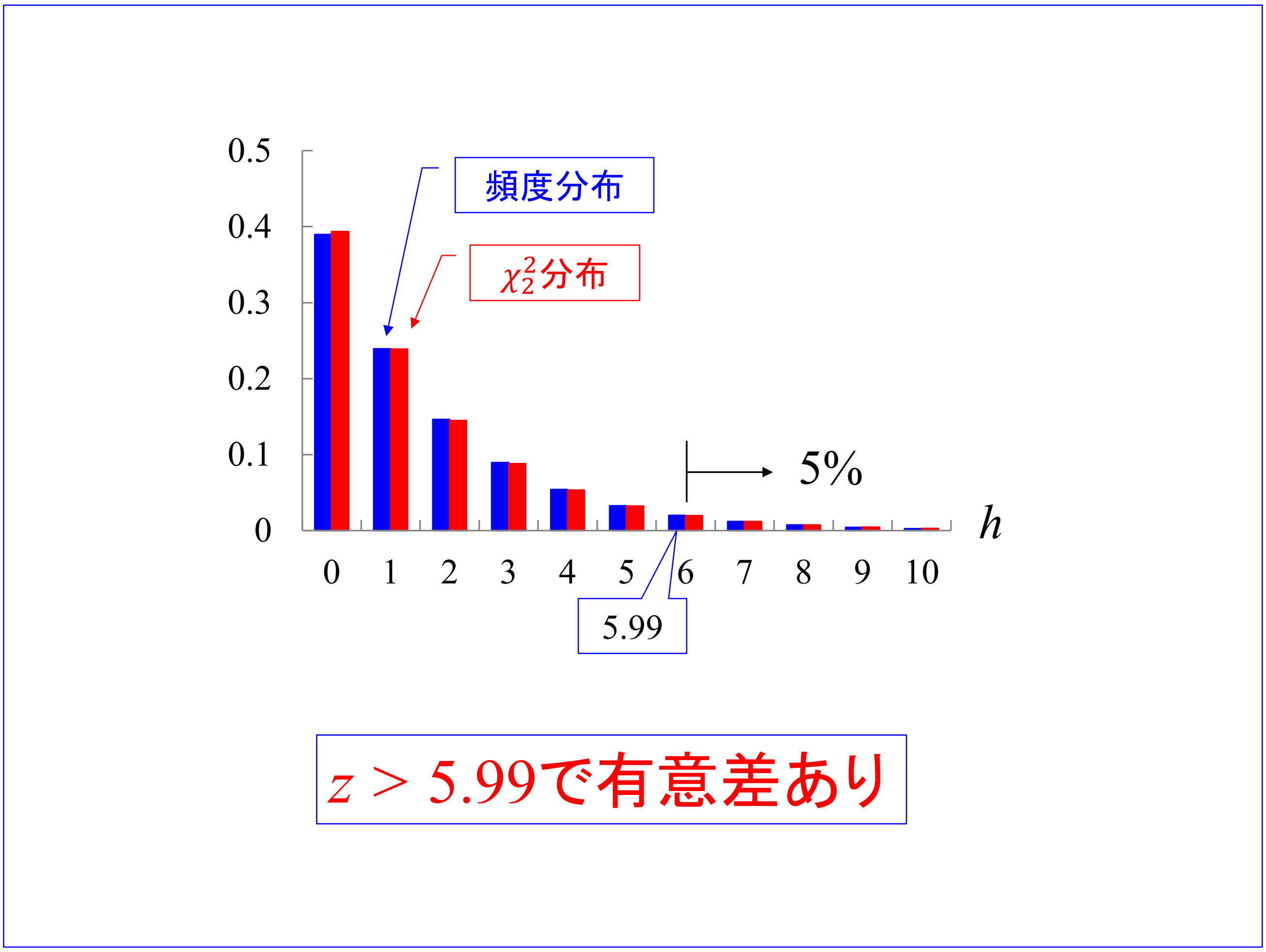
$$\xRightarrow{(n \to \infty)} \chi^2_{k-1}$$

k 群の場合，$n \to \infty$ において h は自由度 $k-1$ の χ^2 分布に従います．



	A	B	C	D	E	F
1	クルスカル・ワリスの検定					
2						
3	n=	60	組数=	1000		
4	n1=	20	n2=	20	n3=	20
5	12/n/(n+1)=	0.0032787	3(n+1)=	183		
6	第1組			順位		
7	x1	x2	x3	x1	x2	x3
8	0.572638308	0.2924922	0.646475	32	19	37
9	0.19607485	0.8901933	0.693418	10	49	42
10	0.891382414	0.4254108	0.983463	51	26	60
11	0.0252588	0.6774426	0.891082	2	39	50
12	0.574420033	0.5459054	0.681993	33	30	40
13	0.136300729	0.2265543	0.428342	9	15	27
14	0.378997648	0.2094374	0.981482	25	12	59
15	0.375359772	0.1084818	0.20847	24	7	11
16	0.858575219	0.9798725	0.899764	48	58	53
17	0.221562775	0.9759162	0.687666	14	57	41
18	0.374649178	0.3697501	0.600634	23	22	36
19	0.829550069	0.7818024	0.015234	46	45	1
20	0.922976971	0.0566909	0.52734	55	5	28
21	0.313221672	0.5316277	0.7231	20	29	43
22	0.278547498	0.1007698	0.108944	17	6	8
23	0.591176347	0.5999265	0.904845	34	35	54
24	0.275680384	0.8984626	0.05388	16	52	4
25	0.551405081	0.2878961	0.329497	31	18	21
26	0.836573165	0.933066	0.037524	47	56	3
27	0.653098615	0.7274914	0.218852	38	44	13
28			順位和	575	624	631
29						h
30						0

「クルスカル・ワリスの検定(3 群).xlsm」のエクセルファイルを開いてください．x_1, x_2, x_3 の 3 群で，各群のサンプルサイズが $n_1 = n_2 = n_3 = 20$ の場合における検定統計量 h の分布を求めるシミュレーションを実施するファイルです．セル F30 では h の値に対して INT()関数を用いることで，小数点以下を切り捨てています．ファイル内ではこの h を求める計算を 1000 組について実行し，$h = 0, 1, 2, 3,$ ・・・の出現頻度を求めています．



図中の青色のバーが得られた h の頻度分布の例です．横軸は h の値，縦軸が h の値の出現頻度です．赤色のバーはχ_2^2分布の理論値です．前スライドのエクセルファイルを開いた状態で（windows パソコンであれば）F9 キーを押すことで，1000 組の全乱数を再生成させて, h の頻度分布を求め直すことができます. 頻度分布は F9 キーを押す度に変化しますが, 赤色のバーの周辺を変化することが見て取れます． χ_2^2分布においては $h \geq 5.99$ となる確率が 0.05 です．従って，$h > 5.99$ のとき，有意差ありとして，帰無仮説を棄却することができます．

9. スティール・デュワスの検定

テューキーの方法のノンパラメトリック版

k 群の互いに独立な標本

帰無仮説H_0:$F_i(x) = F_j(x)$
対立仮説H_1:$F_i(x) \neq F_j(x)$　　$i \neq j$

第1群からのサンプル:$x_{11}, x_{12}, \cdots, x_{1n_1}$
第2群からのサンプル:$x_{21}, x_{22}, \cdots, x_{2n_2}$
⋮
第k群からのサンプル:$x_{k1}, x_{k2}, \cdots, x_{kn_k}$

第i 群と第j 群に対してウィルコクソンの順位和検定を適用する.

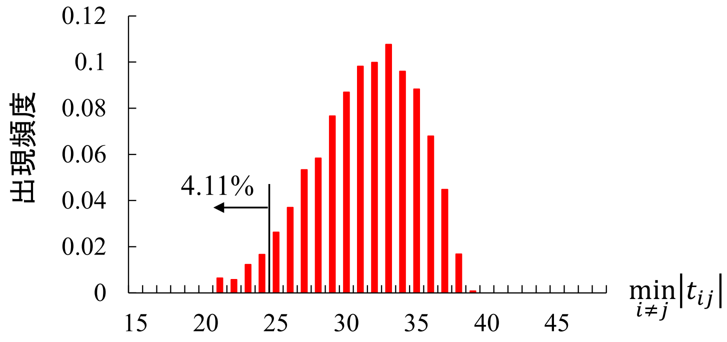
検定統計量 $t_{ij} < t_{th}$であれば帰無仮説$F_i(x) = F_j(x)$を棄却し，対立仮説$F_i(x) \neq F_j(x)$を採用する.

テューキーの方法のノンパラメトリック版がスティール・デュワスの検定です．
k 群の互いに独立な標本の任意の対（第 i 群と第 j 群，$i \neq j$）において

帰無仮説 H_0 : $F_i(x) = F_j(x)$
対立仮説 H_1 : $F_i(x) \neq F_j(x)$

とします．各群のサンプルサイズ$n_1, n_2, \cdots, n_k$は等しいとします．
第 i 群と第 j 群に対してウィルコクソンの順位和検定を適用します．多重比較を考慮して，検定統計量 t_{ij} の閾値 t_{th} を，ウィルコクソンの順位和検定より厳しく設定します．

閾値 t_{th} は，それ以下の値となる $\min_{i \neq j} |t_{ij}|$ の出現確率が有意水準（例えば5%）以下となるときの $\min_{i \neq j} |t_{ij}|$ の値である．例えば，下図のヒストグラムは，$n_1 = \cdots = n_k = 6, k = 3$ のとき，$\min_{i \neq j} |t_{ij}|$ を求めることを10万回繰返して，$\min_{i \neq j} |t_{ij}|$ の出現頻度を求めた結果である．$\min_{i \neq j} |t_{ij}| \leq 24$ の出現頻度は4.11%，$\min_{i \neq j} |t_{ij}| \leq 25$ の出現頻度は6.74%であった．



サンプルサイズ $n_1 = \cdots = n_k = ?$	サンプル群数 k							
	3	4	5	6	7	8	9	10
5	16	15						
6	24	23	22	21				
7	34	32	31	30	29	28	28	
8	45	43	42	40	40	39	38	38
9	59	56	54	53	52	51	50	49
10	74	71	68	67	66	64	63	63

t_{th} は，その値以下となる $\min_{i \neq j} |t_{ij}|$ の出現確率が有意水準（例えば 5%）以下となる閾値です．例えば，スライド内のヒストグラムは，$n_1 = \cdots = n_k = 6,\ k = 3$ のとき，$\min_{i \neq j} |t_{ij}|$ を求めることを 10 万回繰返して，$\min_{i \neq j} |t_{ij}|$ の出現頻度を求めた結果です．$\min_{i \neq j} |t_{ij}| \leq 24$ の出現頻度は 4.11%，$\min_{i \neq j} |t_{ij}| \leq 25$ の出現頻度は 6.74%でした．これより，閾値 $t_{th} = 24$ です．

スライド内の数表は，文献[3]からの抜粋です．サンプルサイズ $n_1 = \cdots = n_k =?$ とサンプル群数 k に応じて，閾値 t_{th} が示されています．上記のシミュレーション例は，表中の〇印の場合が対応します．

[3] R. G. D. Steel, “Some rank sum multiple comparisons tests,” Biometrics, Vol. 17, No. 4, 1961

9. スティール・デュワスの検定：大標本の場合

テューキーの方法のノンパラメトリック版

第i群と第j群に対してウィルコクソンの順位和検定を適用する場合.

$$\text{検定統計量 } z_{ij} = \frac{t_{ij} - \frac{n_i(n_i+n_j+1)}{2}}{\sqrt{\frac{n_i n_j (n_i+n_j+1)}{12}}}$$

第i群と第j群に対してマン・ホイットニーのU検定を適用する場合.

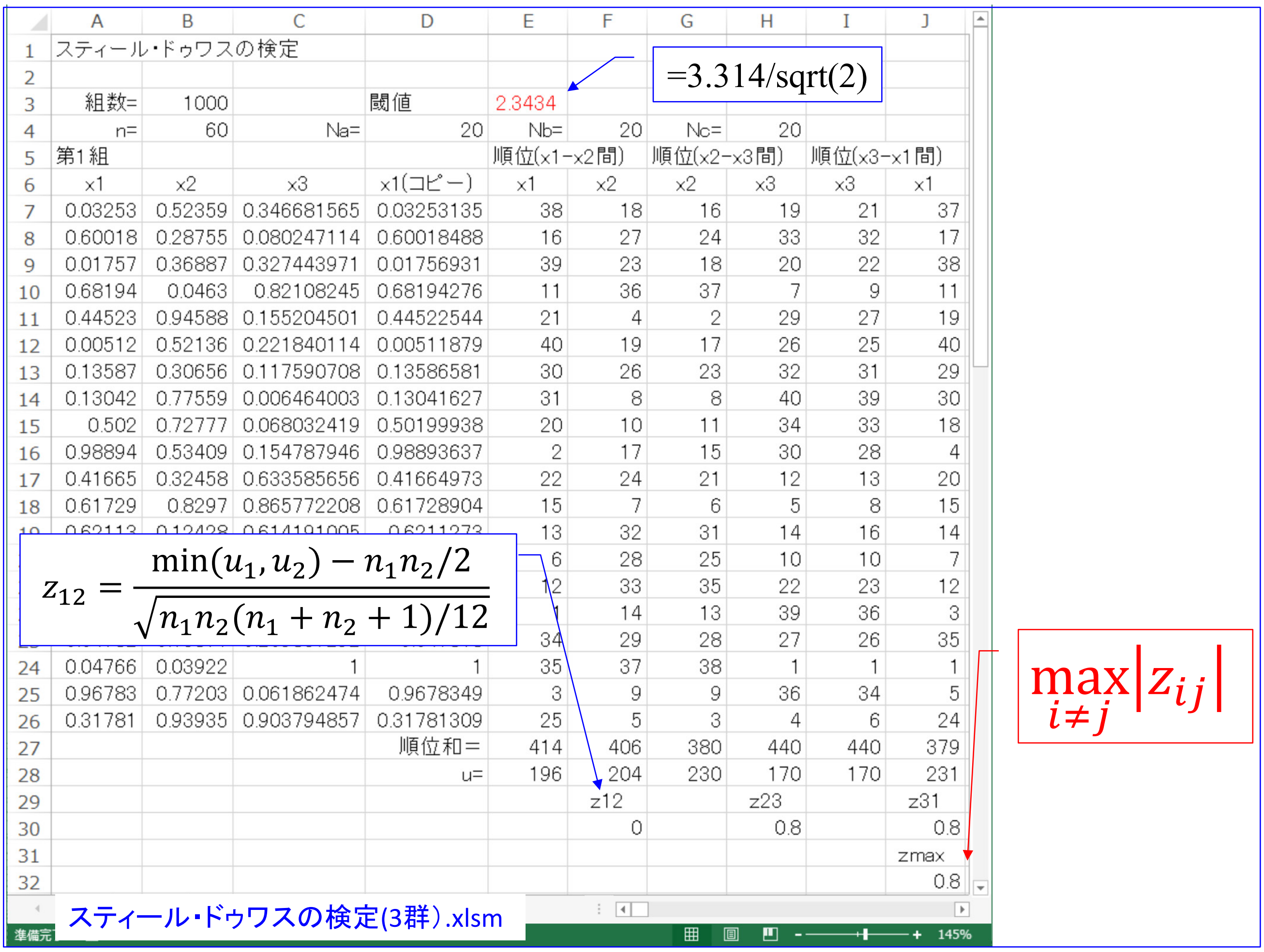
$$\text{検定統計量 } z_{ij} = \frac{u_{ij} - \frac{n_i n_j}{2}}{\sqrt{\frac{n_i n_j (n_i+n_j+1)}{12}}}$$

検定統計量 $z_{ij} > z_{th}$であれば帰無仮説$F_i(x) = F_j(x)$を棄却し，対立仮説$F_i(x) \neq F_j(x)$を採用する.

大標本の場合には検定統計量z_{ij}は標準正規分布に従います．ちなみに，マン・ホイットニーの U 検定を適用した場合にも，検定統計量z_{ij}は標準正規分布に従います.

テューキーの方法では, 2 群のペアごとにステューデントの t 検定を実施します. そして, z 値の境界値を，多重比較を考慮して決定します．例えば，2 群で，$n_1 = n_2 = 20$，有意水準 5%で両側検定の場合，z_{12} 値の境界値 $z_{th} = 2.02$です．3 群で，$n_1 = n_2 = n_3 = 20$，有意水準 5%で両側検定の場合，z_{ij} 値の境界値 z_{th}には「ステューデント化された範囲」と呼ばれる値を用いなければなりません．数表より，データ群数 3，自由度∞の場合のステューデント化された範囲の 5%点は 3.314 です．これより，$z_{th} = \frac{3.314}{\sqrt{2}} = 2.34$（ステューデント化された範囲の数表の値はそのまま用いるのではなく，$\sqrt{2}$で割る必要があります.）とします．そして，検定統計量 $z_{ij} > z_{th}$であれば帰無仮説$F_i(x) = F_j(x)$を棄却し，対立仮説$F_i(x) \neq F_j(x)$を採用します.

t 分布はサンプルサイズが大きくなると標準正規分布に近づきます．そこで，大標本の場合には，ウィルコクソンの順位和検定（マン・ホイットニーの U 検定）の検定統計量z_{ij}に対して，テューキーの方法におけるサンプルサイズが大きい場合（自由度∞）の閾値を適用できます.



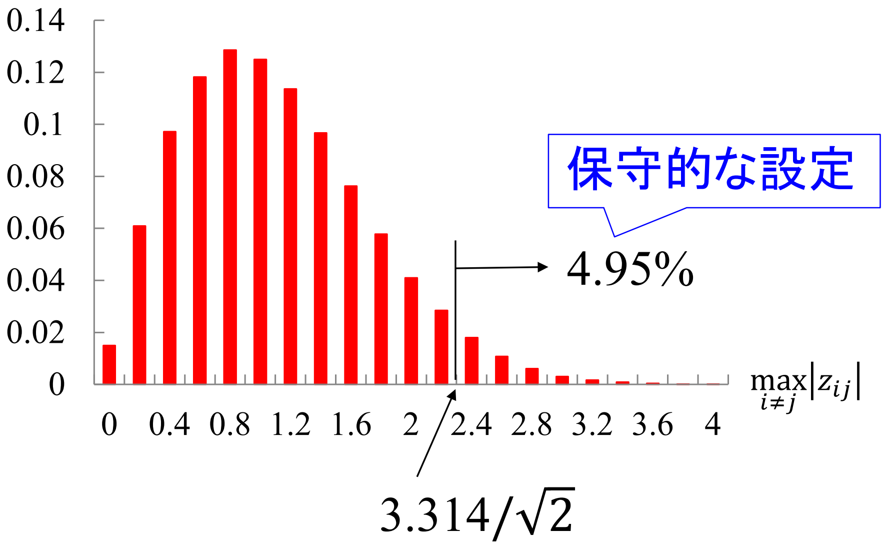
	A	B	C	D	E	F	G	H	I	J
1	スティール･ドゥワスの検定									
2										
3	組数=	1000		閾値	2.3434					
4	n=	60	Na=	20	Nb=	20	Nc=	20		
5	第1組				順位(x1−x2間)		順位(x2−x3間)		順位(x3−x1間)	
6	x1	x2	x3	x1(コピー)	x1	x2	x2	x3	x3	x1
7	0.03253	0.52359	0.346681565	0.03253135	38	18	16	19	21	37
8	0.60018	0.28755	0.080247114	0.60018488	16	27	24	33	32	17
9	0.01757	0.36887	0.327443971	0.01756931	39	23	18	20	22	38
10	0.68194	0.0463	0.82108245	0.68194276	11	36	37	7	9	11
11	0.44523	0.94588	0.155204501	0.44522544	21	4	2	29	27	19
12	0.00512	0.52136	0.221840114	0.00511879	40	19	17	26	25	40
13	0.13587	0.30656	0.117590708	0.13586581	30	26	23	32	31	29
14	0.13042	0.77559	0.006464003	0.13041627	31	8	8	40	39	30
15	0.502	0.72777	0.068032419	0.50199938	20	10	11	34	33	18
16	0.98894	0.53409	0.154787946	0.98893637	2	17	15	30	28	4
17	0.41665	0.32458	0.633585656	0.41664973	22	24	21	12	13	20
18	0.61729	0.8297	0.865772208	0.61728904	15	7	6	5	8	15
19	0.62113	0.12428	0.614191005	0.6211273	13	32	31	14	16	14
					6	28	25	10	10	7
					12	33	35	22	23	12
					1	14	13	39	36	3
					34	29	28	27	26	35
24	0.04766	0.03922	1	1	35	37	38	1	1	1
25	0.96783	0.77203	0.061862474	0.9678349	3	9	9	36	34	5
26	0.31781	0.93935	0.903794857	0.31781309	25	5	3	4	6	24
27				順位和=	414	406	380	440	440	379
28				u=	196	204	230	170	170	231
29						z12		z23		z31
30						0		0.8		0.8
31										zmax
32										0.8

「スティール・デュワスの検定(3 群).xlsm」のファイルを開いてください．マン・ホイットニーの U 検定を各対に適用し，検定統計量 z_{ij} の絶対値の最大値を求めています．ファイルは 3 群の場合で， $n_1 = n_2 = n_3 = 20$ のときの検定統計量 z_{12}（セル F30），z_{13}（セル H30），z_{31}（セル I30）を求め，その絶対値の中の最大値をセル J32 で求めています．

$\max_{i \neq j} |z_{ij}| > q_{k,\infty}/\sqrt{2}$ のとき有意差あり

$q_{k,\infty}$: k 群，自由度∞のステューデント化された範囲

$k = 3$ のとき $q_{3,\infty}/\sqrt{2} = 3.314/\sqrt{2}$



ステューデント化された範囲の数表

k	$q_{k,\infty}$
2	2.772
3	3.314
4	3.633
5	3.858
6	4.03
7	4.17
8	4.286
9	4.387
10	4.474

図は頻度グラフです．3 群で，$n_1 = n_2 = n_3 = 50$ の場合です．横軸が$\max_{i \neq j} |z_{ij}|$の値で，縦軸が頻度です．「スティール・デュワスの検定(3 群).xlsm」のファイルのマクロを実行すると，1000 試行を 100 回（=10 万試行を）実行して，図中の頻度グラフを得ています．頻度は，例えば，横軸の値が 0 と表記されている場合は，[0, 0.2)（0 以上，0.2 未満）の範囲, 0.4 であれば[0.4, 0.6)，の出現頻度です．

100 万回試行のシミュレーションを実行したところ，$\max_{i \neq j} |z_{ij}|$が閾値 $3.314/\sqrt{2}$を超えた頻度は 4.95%でした．5%よりは小さく，第 1 種の過誤が起きにくい（保守的な，有意差が出難い）設定となっています．

ではサンプルサイズが小さい場合に，閾値 $3.314/\sqrt{2}$を適用して良いでしょうか？サンプルサイズを変えてそれぞれ 100 万回試行のシミュレーションを実行して，$\max_{i \neq j} |z_{ij}|$が閾値 $3.314/\sqrt{2}$を超えた頻度を求めてみました．

$n_1 = n_2 = n_3 = 5$の場合　　4.27%

$n_1 = n_2 = n_3 = 20$の場合　　4.77%

$n_1 = n_2 = n_3 = 100$の場合　4.95%
$n_1 = n_2 = n_3 = 250$ の場合　4.96%

スティール･デュワスの検定では, サンプルサイズが十分大きいことが前提となっています. サンプルサイズが小さい場合に同じ閾値($q_{k,\infty}$)を用いると有意差が出難くなることが分かりました．したがって，サンプルサイズが小さい場合でも，閾値($q_{k,\infty}$)を適用することは間違いではありません.

10. スピアマンの順位相関係数

ピアソンの相関係数のノンパラメトリック版

2変量(X, Y)

帰無仮説H_0 : X, Y 間に相関は無い.
対立仮説H_1 : X, Y 間に相関がある.

$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

1. 各変量について，値の小さい順に順位をつける.
   $(r_{x1}, r_{y1}), (r_{x2}, r_{y2}), \cdots, (r_{xn}, r_{yn})$
2. ペア間の順位の差 $d_i = r_{xi} - r_{yi}$を求める.
3. 順位相関係数 r_s を求める.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

スピアマンの順位相関係数はピアソンの相関係数のノンパラメトリック版です.

2 変量 X, Y 間に相関があることの検定です.

1. n 対のサンプル$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$が得られたとして，変量ごとに値の小さい順に順位を付けます．すなわち，$x_1, x_2, \cdots, x_n$についてその中で順位をつけます．つけられた順位を $r_{x1}, r_{x2}, \cdots, r_{xn}$とします. 同様に, $y_1, y_2, \cdots, y_n$の中の順位を $r_{y1}, r_{y2}, \cdots, r_{yn}$とします.
2. ペア間の順位の差$d_i = r_{xi} - r_{yi}$を求めます.
3. 順位相関係数r_sを求めます.

r_sの検定：小標本の場合

$n = 3$のとき

r_x 123 123 123 123 123 123 ・・・

r_y 123 132 213 231 312 321 ・・・

$r_s = 1$ 0.5 0.5 -0.5 -0.5 -1 ・・・

全組み合わせ $3! \times 3! = 36$通り

$r_s = 1$: 6通り

0.5: 12通り

-0.5: 12通り

-1: 6通り

r_s =1となる確率は0.167 → 有意差無し

n	r_sの臨界値
4	1
5	0.9
6	0.829
7	0.714
8	0.643
9	0.6
10	0.564
12	0.506
14	0.456
16	0.425
18	0.399
20	0.377
22	359
24	0.343
26	0.329
28	0.317
30	0.306

小標本の場合は全組み合わせを調べます．サンプルサイズ $n = 3$ の場合，
例えば，$r_{x1} = 1$，$r_{x2} = 2, r_{x3} = 3$（123 と略記します．）に対して，r_{y1}, r_{y2}, r_{y3}の組み合わせは，

123 132 213 231 312 321

の 6 通りがあります．r_{x1}, r_{x2}, r_{x3}=123，r_{y1}, r_{y2}, r_{y3}=123 のとき，順位相関係数 r_s=1 です．

全組み合わせは $3! \times 3! = 36$ 通りです．この中で

$r_s = \ 1$ となる組み合わせは 6 通り

$r_s = \ 0.5$ となる組み合わせは 12 通り

$r_s = -0.5$ となる組み合わせは 12 通り

$r_s = -1$ となる組み合わせは 6 通り

です．従って，$r_s = 1$ となる確率は $6/36 = 0.167$ です．サンプルサイズ $n = 3$ の場合は，どのような組み合わせでも有意差は出ません．

スライド中の右の表には有意差ありとなる順位相関係数 r_sの臨界値が示されています．サンプルサイズ $n = 5$ の場合で，$r_s = 0.9$ 以上となった場合には，片側検定にて，有意差ありとして帰無仮説を棄却できます．

r_sの検定：大標本の場合

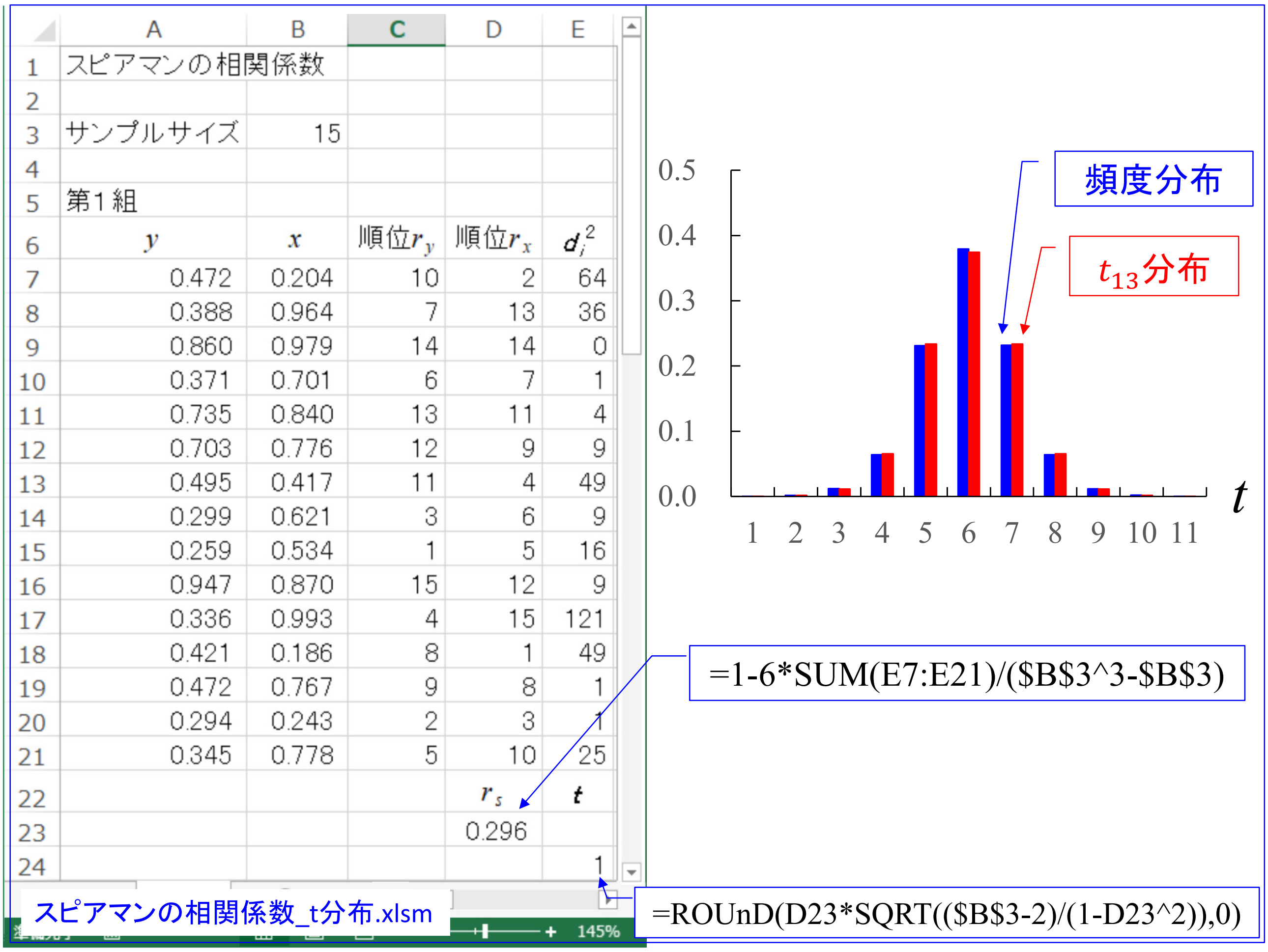
$n > 10$のとき

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$ 自由度 n-2 の t 分布に従う

スピアマンの順位相関係数は大標本の場合（サンプルサイズ $n > 10$ にて）

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

が自由度 $n-2$ の t 分布に従います.



	A	B	C	D	E
1	スピアマンの相関係数				
2					
3	サンプルサイズ	15			
4					
5	第1組				
6	y	x	順位r_y	順位r_x	d_i^2
7	0.472	0.204	10	2	64
8	0.388	0.964	7	13	36
9	0.860	0.979	14	14	0
10	0.371	0.701	6	7	1
11	0.735	0.840	13	11	4
12	0.703	0.776	12	9	9
13	0.495	0.417	11	4	49
14	0.299	0.621	3	6	9
15	0.259	0.534	1	5	16
16	0.947	0.870	15	12	9
17	0.336	0.993	4	15	121
18	0.421	0.186	8	1	49
19	0.472	0.767	9	8	1
20	0.294	0.243	2	3	1
21	0.345	0.778	5	10	25
22				r_s	t
23				0.296	
24					1

「スピアマンの相関係数_t 分布.xlsm」のファイルを開いてください．$0 < n < 1$の一様乱数を 15 組（サンプルサイズ $n = 15$）生成して，変量y, xごとに小さい方からそれぞれ順位を付けています．セル D23 にて順位相関係数 r_s を求め，セル E24 にて t 値を求めています．t 値は小数点以下 1 桁目を四捨五入しています．

1000 組×100 回のシミュレーションを行うマクロを実行することにより，頻度分布を求めた結果が図のヒストグラムです．青いバーが得られた頻度分布，赤いバーが自由度(n - 2 =) 13 の t 分布です．両者はほぼ一致しています．