降水確率と視聴率

1. 雨が降る確率とは？
2. テレビの視聴率とは？
3. 二項分布と正規分布
4. カイ自乗分布とティー分布

今日のテーマ

どちらが正確か？

【例1】 過去のデータによると
或る気象状況の日が100日あり
翌日雨が降ったのは 80日 だった

降水確率は
80÷100＝ 0.8

⇔

【例2】 過去のデータによると
或る気象状況の日が10,000日あり
翌日雨が降ったのは 8,000日 だった

降水確率は
8,000÷10,000＝0.8

降水確率はどちらも 80%

【例2】の方が正確かも？

定量的に評価しよう

「降水確率80%」とは？

明日の東京地方の正午から午後6時までの
降水確率は80パーセントです。

東京地方で、
明日の正午から午後6時までの間に
合計の降水量が**1ミリメートル以上**になる
確率が80パーセント

今日と同じような気象条件の日が100日あって、
翌日そうなる日数は、およそ80日

「降水確率80%」とは？

今日と同じような気象条件の日が100日あって、
翌日そうなる日数は、およそ80日

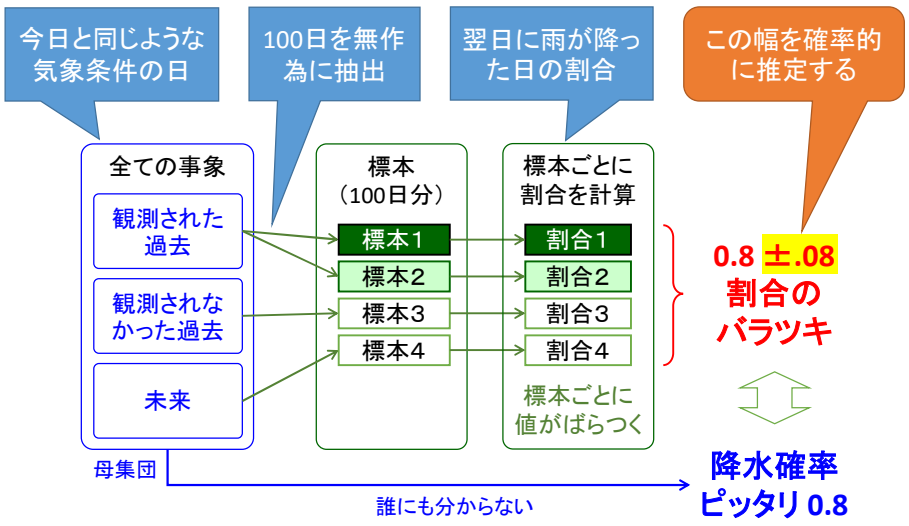
72日から88日の間にほぼ収まる（丁度80日ではない）

そうなる状況が100日のうち、
80日±8日の範囲内に
95%の確率で含まれる

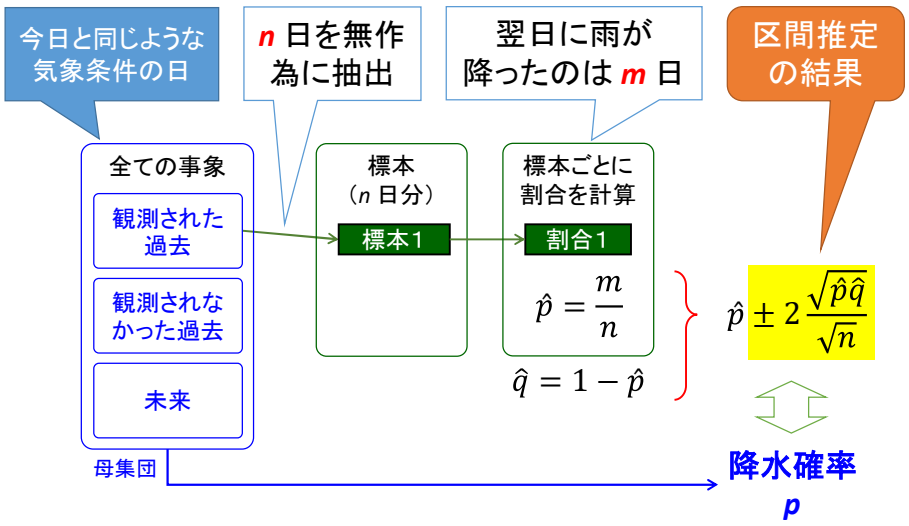
これが
← 区間推定

小林道正「データラメにひそむ確率法則」岩波書店

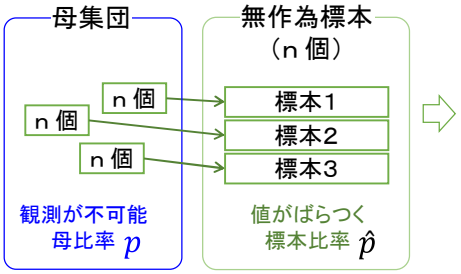
降水確率の区間推定とは？



降水確率の区間推定の手順



比率の区間推定の解釈1



信頼区間（解釈1）

無作為標本を何回も繰り返すと
そのうちの 95 % は、
標本比率が

$$p \pm 2\sqrt{\hat{p}\hat{q}}/\sqrt{n}$$

の範囲に入る

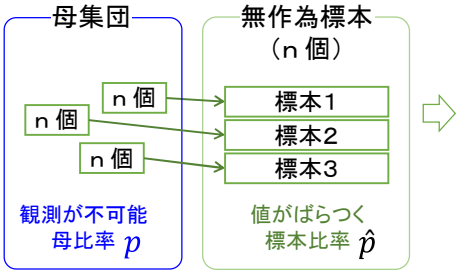
n 日を無作為に抽出

翌日に雨が降ったのは ***m*** 日

$$\hat{p} = \frac{m}{n}$$

$$\hat{q} = 1 - \hat{p}$$

比率の区間推定の解釈2



信頼区間（解釈2）

無作為標本を何回も繰り返すと
そのうちの 95 % は、
区間：

$$\hat{p} \pm 2\sqrt{\hat{p}\hat{q}}/\sqrt{n}$$

が 母比率 p を含んでいる

n 日を無作為に抽出

翌日に雨が降ったのは ***m*** 日

$$\hat{p} = \frac{m}{n}$$

$$\hat{q} = 1 - \hat{p}$$

「降水確率80%」を検証

過去のデータによると
今日と同様な気象条件が100日あり
翌日雨が降ったのは 80日 だった

n =100
m=80

翌日雨が降る日数は、100日のうち
80日 ±8日 であると
95%の確率で言える

$\hat{p} = m/n$ =0.8
$\hat{q} = 1 - \hat{p}$ =0.2

80 ±8
↑
区間推定

100日なので
×100

$$\hat{p} \pm 2\sqrt{\hat{p}\hat{q}}/\sqrt{n}$$
$$=0.8 \pm 2\sqrt{0.8 \times 0.2/100}$$
$$=0.8 \pm 0.08$$

比率の区間推定（まとめ）

【例1】 過去のデータによると
今日と同様な気象条件が100日あり
翌日雨が降ったのは 80日 だった

今日の様な気象条件だと
翌日雨が降る確率は
0.8 ±0.08 であると
0.95 の確率で言える

【例2】 過去のデータによると
今日と同様な気象条件が10,000日あり
翌日雨が降ったのは 8,000日 だった

今日の様な気象条件だと
翌日雨が降る確率は
0.8 ±0.008 であると
0.95 の確率で言える

データ数を100倍にすると、区間の幅は1/10倍

$$\hat{p} \pm 2\sqrt{\hat{p}\hat{q}}/\sqrt{n}$$

区間推定の公式

データ数を a 倍にすると、区間の幅は $1/\sqrt{a}$ 倍

精度を上げる（区間を狭める）には，十分な数の正確なデータが必要

降水確率と視聴率

1. 雨が降る確率とは？
2. テレビの視聴率とは？
3. 二項分布と正規分布
4. カイ自乗分布とティー分布

今日のテーマ

断言できるか？

テレビの視聴率 番組Aは10％ 番組Bは13％	番組Bの方が 視聴率が高いと 断言できるか？

例1　100世帯で調査した

別の100世帯で調査したら

逆転するかも…

番組Aの方が
高かったかも…

⇔

例2　10,000世帯で調査した

別の10,000世帯で調査しても

逆転しない？

番組Bの方が
確実に高い？

区間推定の使い方

テレビの視聴率
番組Aは　10%
番組Bは　13%

番組Bが高いと
断言できるか？

区間推定に
できること

調査した数が 100 世帯ならば、
区間推定の幅は 6 %

番組Aは　10±6% ＝ 4 ～ 16%

番組Bは　13±6% ＝ 7 ～ 19%

区間推定で
わかること

別の100世帯を調査していたら
番組Aの方が高かったかもしれない

区間推定の使い方

テレビの視聴率
番組Aは　10%
番組Bは　13%

番組Bが高いと
断言できるか？

調査した数が 100 世帯ならば、
区間推定の幅は 6 %

・調査数を 100倍
・区間幅は1/10倍

調査数が 10,000 世帯ならば、
区間推定の幅は 0.6 %

番組Aは　10±0.6% ＝　9.4～10.6%

番組Bは　13±0.6% ＝ 12.4～13.6%

番組Bの方が高いと断言できる

100人に聞いた視聴率10%

100人を無作為に選び電話調査
番組Aを視聴したのは 10人 だった

n =**100**
m=10

番組Aの視聴率は
10±6 % であると
95%の確率で言える

$\hat{p} = m/n$ =0.1
$\hat{q} = 1 - \hat{p}$ =0.9

$$\hat{p} \pm 2\sqrt{\hat{p}\hat{q}}/\sqrt{n}$$
$$= 0.1 \pm 2\sqrt{0.1 \times 0.9/100} = 0.1 \pm 0.06$$

(95%だと約2)

10,000人に聞いた視聴率10%

10,000人を無作為に選び電話調査
番組Aを視聴したのは 1,000人 だった

n =**10,000**
m=1,000

番組Aの視聴率は
10±0.6 % であると
95%の確率で言える

$\hat{p} = m/n$ =0.1
$\hat{q} = 1 - \hat{p}$ =0.9

$$\hat{p} \pm 2\sqrt{\hat{p}\hat{q}}/\sqrt{n}$$
$$= 0.1 \pm 2\sqrt{0.1 \times 0.9/10{,}000} = 0.1 \pm 0.006$$

(95%だと約2)

まとめ

テレビの視聴率 番組 A は 10 % 番組 B は **13** %	番組 B の方が 視聴率が高いと 断言できるか？

例1　100 世帯で調査した

↓

別の100世帯で調査したら
番組 A は　**16**%
番組 B は　7%

↓

番組 A の方が高い
逆転もあり得る

⇔

例2　10,000 世帯で調査した

↓

別の10,000世帯で調査したら
番組 A は　10.6%
番組 B は　**12.4**%

↓

番組 B の方が高い
ほぼ確実に！

詳細

データ数と信頼区間

経済的 ← データ数　少 →

データ数 100 → $\alpha = 0.05$, $n = 100$, 信頼度 95%

$$0.1 \pm 1.96\frac{\sqrt{0.1\times0.9}}{\sqrt{100}} = 0.1 \pm 0.0588$$

← 信頼区間

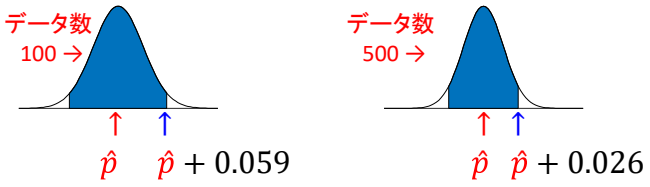
データ数 500 → $\alpha = 0.05$, $n = 500$, 信頼度 95%

$$0.1 \pm 1.96\frac{\sqrt{0.1\times0.9}}{\sqrt{500}} = 0.1 \pm 0.0263$$

← 信頼区間

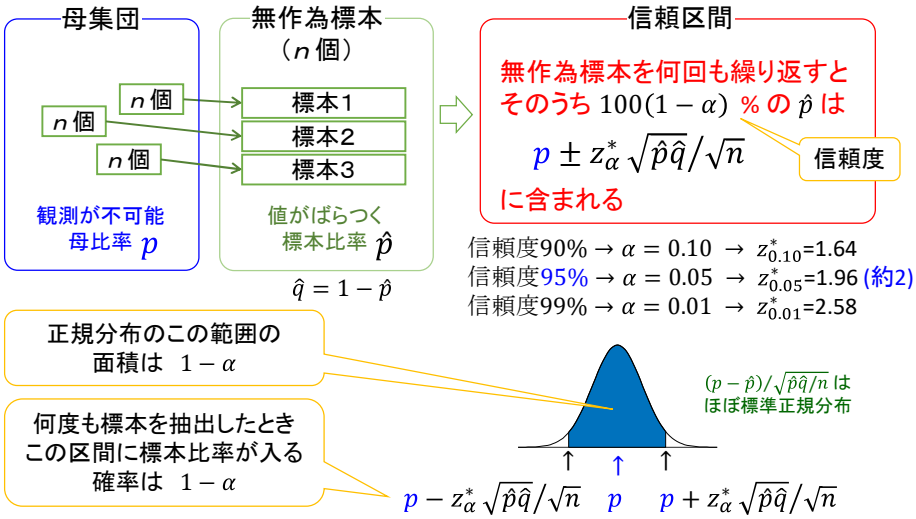
← 狭い　信頼区間　広い

正確



詳細

比率の区間推定



降水確率と視聴率

1. 雨が降る確率とは？
2. テレビの視聴率とは？
3. 二項分布と正規分布
4. カイ自乗分布とティー分布

ベルヌーイ試行と二項分布

視聴率
番組を見た →1
番組を見ない→0

コイン投げ
表が出た →1
裏が出た →0

くじ
当たった →1
はずれた →0

試行
成功 →1
失敗 →0

試行
成功 → 1
失敗 → 0

試行を n 回行った
x 回、成功
$n-x$ 回、失敗

ベルヌーイ試行

割合を計算
成功の割合 $x/n = p$
失敗の割合 $1-p=q$

確率変数 X の確率関数
$f(x|p) = \binom{n}{x} p^x q^{n-x} \equiv B(n,p)$

二項分布
（離散的な分布）

平均 $E(X) = np$
分散 $V(X) = npq$

二項分布と正規分布

確率変数 X の確率関数
$B(n,p) = \binom{n}{x} p^x q^{n-x}$

二項分布
（離散的な分布）

確率変数 X の密度関数
$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

正規分布
（連続的な分布）

X が二項分布に従う: $X \sim B(n,p) \cong N(np,npq)$ ← データ数 n が大きい時

$X-np$ が正規分布に従う: $X - np \sim N(0,npq)$

標準正規分布に従う: $\frac{X-np}{\sqrt{npq}} \sim N(0,1)$ ← $\frac{X-\text{平均}}{\text{標準偏差}}$

成功の確率 は 正規分布 する

$$\frac{X-np}{\sqrt{npq}} \sim N(0,1)$$

左辺の分母と分子を n で割る

$$\frac{Z-p}{\sqrt{pq/n}} \sim N(0,1)$$

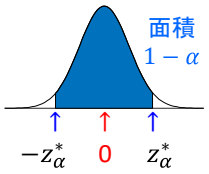
$$\left|\frac{Z-p}{\sqrt{pq/n}}\right| \leq z_\alpha^*$$

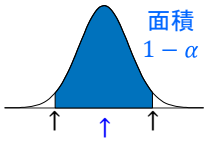
$$p - z_\alpha^*\sqrt{pq/n} \leq Z \leq p + z_\alpha^*\sqrt{pq/n}$$

成功の確率
$Z = X/n$

確率 1−α で
発生する区間

正規分布





$p - z_\alpha^*\sqrt{pq/n}$ p $p + z_\alpha^*\sqrt{pq/n}$

MATLAB program

```
close all; clear all;
figure('Position',[10 10 200 100]);

sgm=10;
myu=50;

x=30:0.5:70;
y=exp(-((x-
myu).^2)/2/sgm^2)./sqrt(2*pi)/sgm;
area(x,y); hold on;

x=0:0.5:100;
y=exp(-((x-
myu).^2)/2/sgm^2)./sqrt(2*pi)/sgm;
plot(x,y,'k'); grid on;

axis([0,100,0,0.05]);
```

降水確率と視聴率

1. 雨が降る確率とは？
2. テレビの視聴率とは？
3. 二項分布と正規分布
4. カイ自乗分布とティー分布

参考

正規分布から誘導される分布

$Z_1, Z_2, \cdots, Z_n \sim N(0,1)$

Z_i は標準正規分布に従う独立な確率変数

$Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ のとき

$Y \sim \chi_n^2$

Y は自由度 n のカイ自乗分布に従う

$X \sim N(0,1),\ Y \sim \chi_n^2$ のとき

$$T = \frac{X}{\sqrt{Y/n}} \sim t_n$$

T は自由度 n のティー分布に従う

参考

様々な分布と密度関数

正規分布　$X \sim N(\mu, \sigma)$ ⇨ $f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

自由度 n の
カイ自乗分布　$X \sim \chi_n^2$ ⇨ $f(x|n) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} exp\left(-\frac{x}{2}\right)$

$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt$（ガンマ関数）

自由度 n の
ティー分布　$X \sim t_n = \frac{N(0,1)}{\sqrt{\chi_n^2/n}}$ ⇨ $f(x|n) = \frac{1}{\sqrt{n}\,B\left(\frac{n}{2},\frac{1}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$

$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$（ベータ関数）

降水確率と視聴率

1. 雨が降る確率とは？
2. テレビの視聴率とは？
3. 二項分布と正規分布
4. カイ自乗分布とティー分布
5. 2つの比率の差の区間推定

2つの比率の差の区間推定

母集団が2つあり、母比率の差 $p_1 - p_2$ を区間推定したい。
母比率は未知だが、標本比率 $\bar{X}$, $\bar{Y}$ ならば観測できる。
これらを使えば、母比率の差を区間推定できる。

【例】 ある候補者の支持率は、
男と女で、差があるか？

$p_1 - p_2$ の、$100(1-\alpha)$ % の信頼区間は、近似的に以下となる。

$$\bar{X} - \bar{Y} \pm z_{\alpha}^{*} \sqrt{\frac{1}{m}\bar{X}(1-\bar{X}) + \frac{1}{n}\bar{Y}(1-\bar{Y})}$$

出典：裳華房_数理統計学_稲垣宣生

支持率は 男と女で 差があるか？

【例】 ある候補者を支持するか否かを調査した。

- 男性 **100** 人中の **55** 人が支持、
- 女性 **60** 人中の **48** 人が支持すると答えた。

男女の支持率の差について、95%の信頼区間は？

【答え】
差の値には幅があるが
確実に負なので
確実に差がある

出典：裳華房_数理統計学_稲垣宣生

【解説】

$$\bar{X} = \frac{55}{100} = 0.55$$

$$\bar{Y} = \frac{48}{60} = 0.80$$

$$\bar{X} - \bar{Y} = -0.25 \pm \Delta$$

$$\alpha = 0.05$$

$$\Delta = z_{\alpha}^{*} \sqrt{\frac{1}{m}\bar{X}(1-\bar{X}) + \frac{1}{n}\bar{Y}(1-\bar{Y})}$$

$$= 1.96 \sqrt{\frac{0.55 \times 0.45}{100} + \frac{0.8 \times 0.2}{60}}$$

$$= 0.1$$

答え：
差の信頼区間は -0.25±0.1 である
（最低で -0.35、最高で -0.15）
確実に負の値、つまり、差は確実！

2つの品種、発芽率に差がない？

【例】 2つの品種のタネを蒔いて、発芽率を調査した。

- 品種Aは、**220** 個蒔いて **160** 個が発芽した
- 品種Bは、**200** 個蒔いて **160** 個が発芽した

発芽率の差について、95%の信頼区間は？

【答え】
差の値には幅があり
正にも負にも成り得るので
差は不確実

出典：裳華房_数理統計学_稲垣宣生

【解説】

$$\bar{X} = \frac{160}{220} = 0.73$$

$$\bar{Y} = \frac{160}{200} = 0.80$$

$$\bar{X} - \bar{Y} = -0.073 \pm \Delta$$

$$\alpha = 0.05$$

$$\Delta = z_{\alpha}^{*}\sqrt{\frac{1}{m}\bar{X}(1-\bar{X}) + \frac{1}{n}\bar{Y}(1-\bar{Y})}$$

$$= 1.96\sqrt{\frac{0.73 \times 0.27}{220} + \frac{0.8 \times 0.2}{200}}$$

$$= 0.081$$

答え：
差の信頼区間は 0.073±0.081である
（最低で -0.08、最高で +0.15）
負にも正にも成り得るので
確実に差があるとは言えない

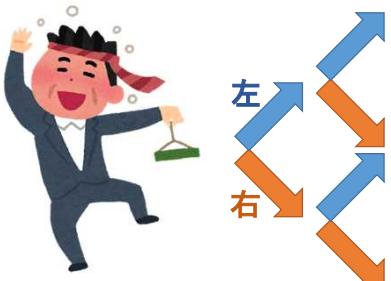
酔っぱらい 何処へ行く
ランダムウォークと正規分布

- 千鳥足にも理屈がある
- 行きつく先は正規分布
  （ランダムウォークの数理モデル）
- エレベータに乗った酔っ払い
- 行きつく先は逆ガウス分布
  （地震予測の数理モデル）

酔っ払いは何処へ行くか？

左か右かの
確率は1／2

どちらへ行くかは
全くの出鱈目（ランダム）

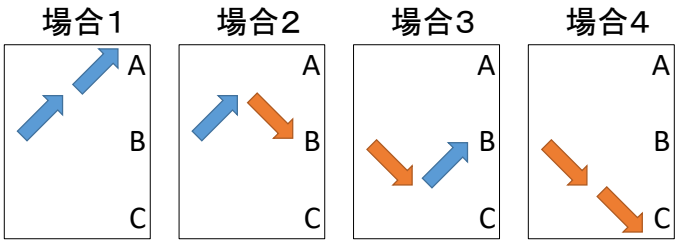


酔っ払いの
らんだむ をーく

【問1】
A,B,Cに辿り着く確率は？

	ア）	イ）	ウ）
A地点	0.33	0.25	0.1
B地点	0.33	0.5	0.8
C地点	0.33	0.25	0.1

全ての組み合わせを列挙する



回数

A地点に辿り着く→１回
B地点に辿り着く→２回
C地点に辿り着く→１回

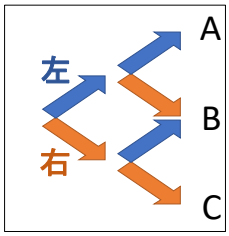
計４回

確率

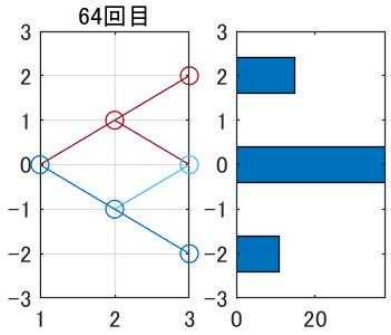
A地点に辿り着く→ **1/4**
B地点に辿り着く→ **1/2**
C地点に辿り着く→ **1/4**

総和は1

シミュレーションで確かめる



← $N = 2$ →



64で割ると確率

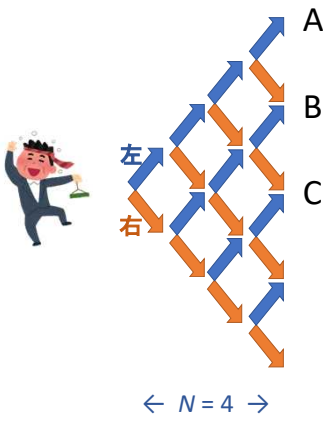
回数	A	B	C
32	0.13	0.63	0.25
64	0.17	0.59	0.23
128	0.28	0.48	0.23
256	0.26	0.50	0.24

各地点に辿り着く確率

MATLAB プログラム

```
clear variables; close all;

N=3; R=N;
M=32;

rng(0);
y=zeros(2*N+1,1); x=zeros(N,1);
figure('Position',[10 10 200 150]);

for j=1:M
    x(1)=0;
    for i=2:N
        s=2*(rand(1)>0.5)-1;
        x(i)=x(i-1)+s;
    end
    y(x(N)+R+1)=y(x(N)+R+1)+1;
    subplot(1,2,1); plot(x,'o-'); hold off; grid on;
    axis([1 N -R R]); title(strcat(num2str(j),'回目'));

    subplot(1,2,2); grid on;
    barh(-R:R,y); axis([0 Inf -R R]);
    pause;
end
```

酔っ払いは何処へ行くか？



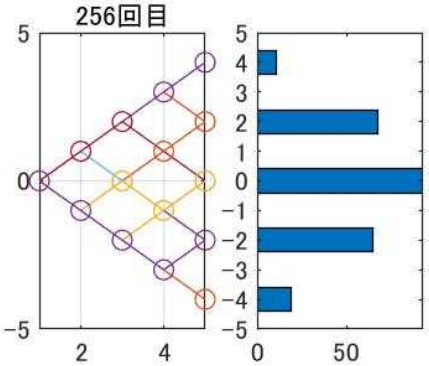
← $N = 4$ →

左の確率は p =0.5
右の確率は1-p

【問2】A,B,Cに着く確率は？

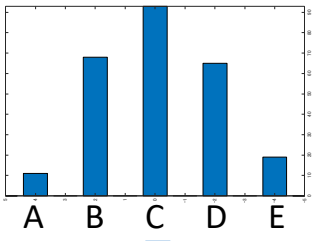
	ア）	イ）	ウ）
A地点	0.1	0.05	0.05
B地点	0.2	0.15	0.2
C地点	0.4	0.6	0.5

シミュレーションで確かめる

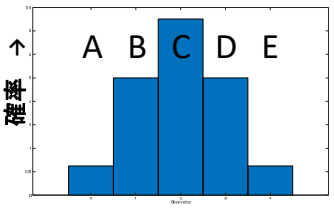


回数	A	B	C	D	E
64	0.08	0.25	0.38	0.28	0.02
128	0.06	0.26	0.37	0.28	0.03
256	0.07	0.25	0.36	0.27	0.04
512	0.08	0.22	0.38	0.25	0.07
1024	0.06	0.22	0.38	0.25	0.07

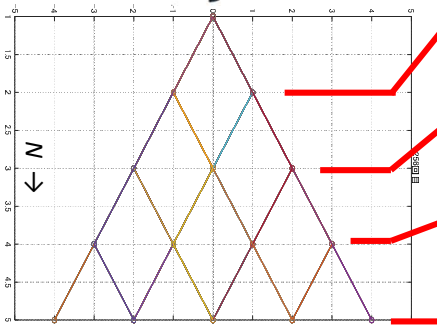
90度
回転



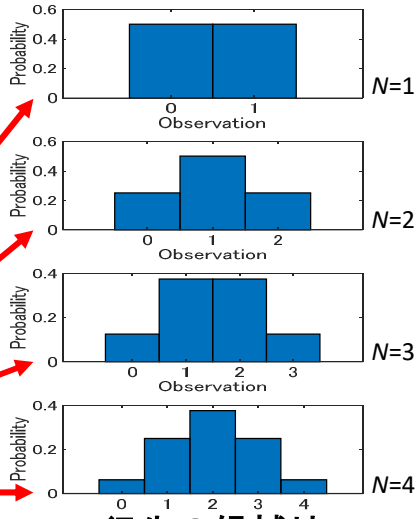
これが「**2項分布**」だ



二項分布



行先の候補地



行先の候補地

二項分布と正規分布

$$f(x) = \binom{N}{x} p^x (1-p)^{N-x}$$

$\frac{N!}{x!(N-x)!}$

平均： Np

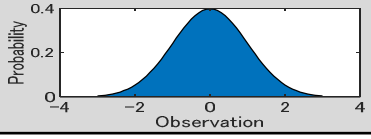
分散：$Np(1\text{-}p)$

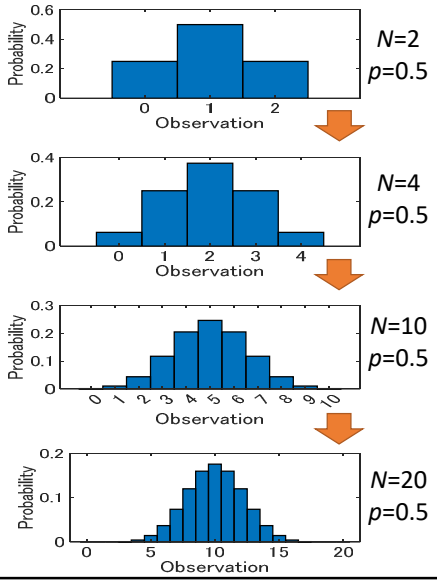
二項分布

中心極限定理 ⏸ $N \to \infty$ で同じ

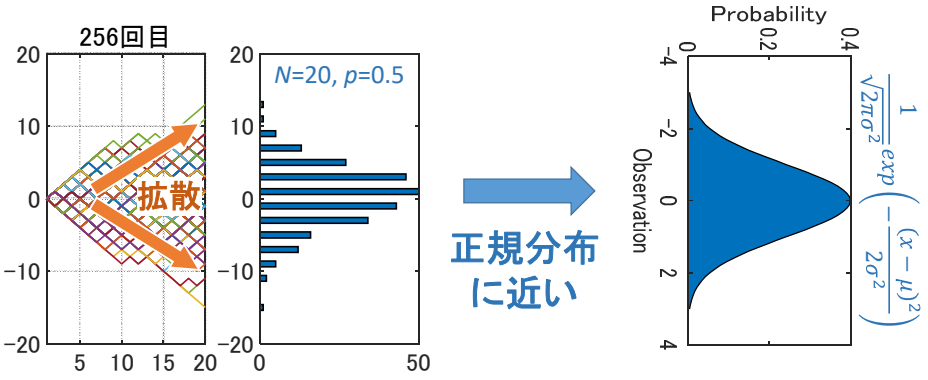
正規分布

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$





シミュレーションで確かめる



何かが拡散すると正規分布となる
花粉、煙、熱、インク、etc.

正規分布と拡散方程式

正規分布

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

とくに、時刻 t に比例して
分散 σ^2 が大きくなるとき　← 例えば $\sigma^2=t$

拡散方程式

$$\frac{\partial}{\partial t}f(x,t)=D\frac{\partial^2}{\partial x^2}f(x,t)$$

を満たす。