二項分布に関連した検定

二項分布の例

問題： 日本人3人（甲・乙・丙）を無作為抽出した。
この3人の中のO型の人数について，
それぞれの値を取る確率を求める.

- 日本の血液型の分布は次のようにいわれている
  - O型 ： 30%
  - A型 ： 35%
  - B型 ： 25%
  - AB型 ： 10%

O型が1人もいない確率

- 記号
  - Pr(甲○)：甲がO型である確率 = 0.3
  - Pr(甲×)：甲がO型でない確率 = 1-0.3 = 0.7
- O型が1人もいない確率
  - Pr(甲×) × Pr(丙×) × Pr(丙×)

    = 0.7 × 0.7 × 0.7 = 0.343

日本の血液型の分布

- O型：30%
- A型：35%
- B型：25%
- AB型：10%

O型が1人いる確率

- O型が1人いる確率
  - Pr(甲○) × Pr(丙×) × Pr(丙×)
    = 0.3 × 0.7 × 0.7 = 0.147
  - Pr(甲×) × Pr(丙○) × Pr(丙×)
    = 0.7 × 0.3 × 0.7 = 0.147
  - Pr(甲×) × Pr(丙×) × Pr(丙○)
    = 0.7 × 0.7 × 0.3 = 0.147
  - これらの和をとることにより，求める確率は
    0.147 + 0.147 + 0.147 = 0.441

日本の血液型の分布

- O型 ： 30%
- A型 ： 35%
- B型 ： 25%
- AB型 ： 10%

O型が2人いる確率

- **O型が2人いる確率**
  - Pr(甲○) × Pr(丙○) × Pr(丙×)
    = 0.3 × 0.3 × 0.7 = 0.063
  - Pr(甲○) × Pr(丙×) × Pr(丙○)
    = 0.3 × 0.7 × 0.3 = 0.063
  - Pr(甲×) × Pr(丙○) × Pr(丙○)
    = 0.7 × 0.3 × 0.3 = 0.063
  - これらの和をとることにより，求める確率は
    0.063 + 0.063 + 0.063 = 0.189

日本の血液型の分布

- O型 ： 30%
- A型 ： 35%
- B型 ： 25%
- AB型 ： 10%

３人ともO型である確率

- ３人ともO型である確率
  - Pr(甲O) × Pr(丙O) × Pr(丙O)
    = 0.3 × 0.3 × 0.3 = 0.027
- 起こりうるすべての場合の確率の和
  - 0.343 + 0.441 + 0.027 = 1

日本の血液型の分布

- O型 ： 30%
- A型 ： 35%
- B型 ： 25%
- AB型 ： 10%

二項係数

- 二項係数
  - p：事象 A が起こる確率
  - q：事象 A が起こらない確率（$q = 1 - p$）
  - n 回の独立な試行で事象 A が x 回起こる確率は二項分布：${}_nC_x p^x q^{n-x}$ にしたがう
  - ${}_nC_x$：二項係数

$$ {}_nC_x = \frac{n(n-1)\cdots(n-x+1)}{x(x-1)\cdots 2\cdot 1}, \quad {}_nC_0 = 0 $$

  - 二項係数の性質：${}_nC_x = {}_nC_{n-x}$

例：O型の人数についての確率

- <u>二項分布を用いて，O型の人数の確率を求める</u>

<u>二項分布の確率</u>

O型の人数 x	0	1	2	$\cdots$	r	$\cdots$	n
確率	p^0q^n	${}_nC_1p^1q^{n-1}$	${}_nC_2p^2q^{n-2}$	$\cdots$	${}_nC_rp^rq^{n-r}$	$\cdots$	p^nq^0

- O型が１人もいない確率 ： ${}_3C_0(0.3)^0(0.7)^3 = 0.343$
- O型が１人いる確率 ： ${}_3C_1(0.3)^1(0.7)^{3-1} = \frac{3}{1}(0.3)(0.49) = 0.441$
- O型が2人いる確率 ： ${}_3C_2(0.3)^2(0.7)^{3-2} = \frac{3\times 2}{2\times 1}(0.09)(0.7) = 0.189$
- ３人ともO型である確率 ： ${}_3C_3(0.3)^3(0.7)^{3-3} = 0.027$

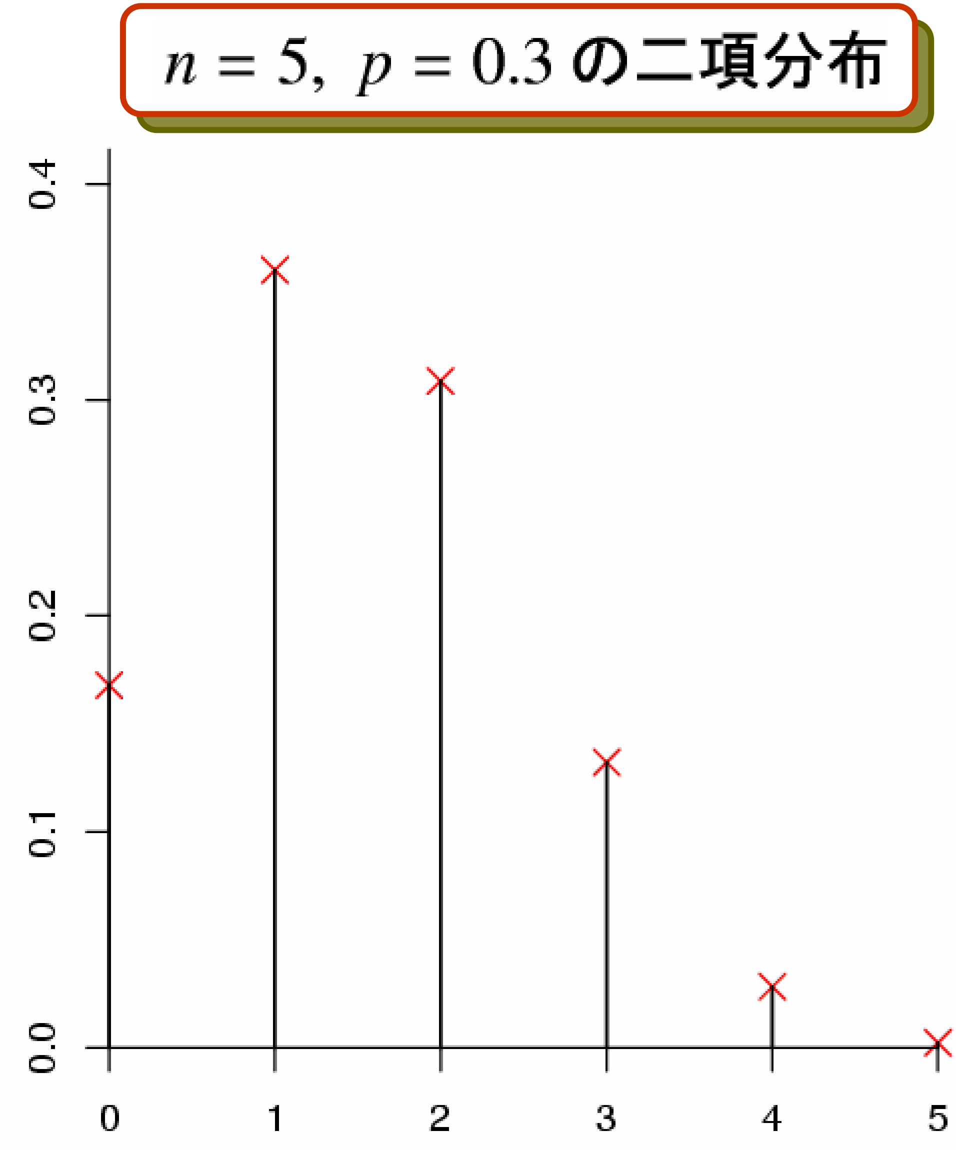
比率の推定と検定

比率の推定と二項分布の正規近似

- 比率が未知の場合： x/n で推定
  - 100人抽出した中に，O型が32人 ⇒ $p = 0.32$ と推定
- O型の人の人数： 二項分布
  - 標本数が大きく，母集団比率があまり小さくないとき

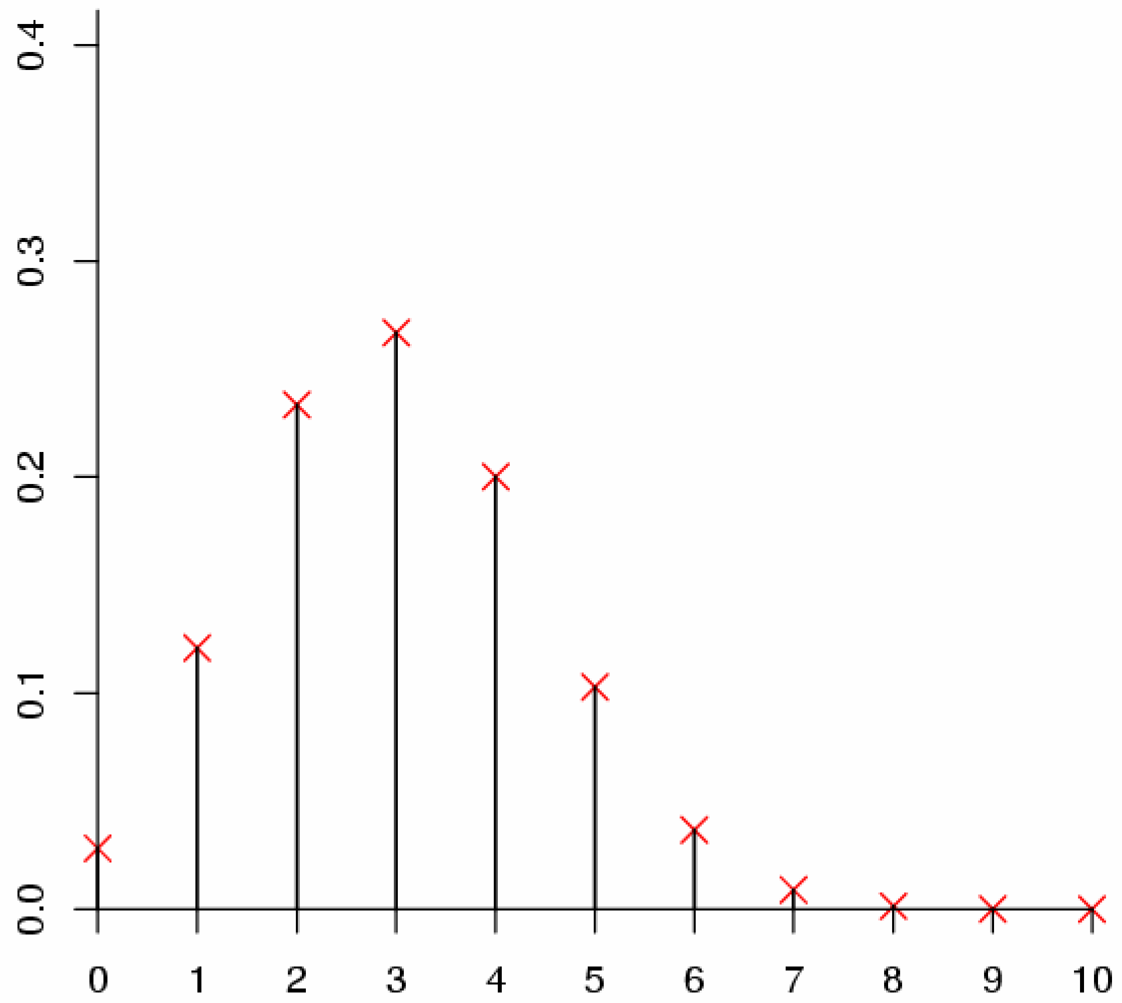
    ⇒ 平均 np, 分散 npq の正規分布で近似できる

標本数５，比率０．３の二項分布



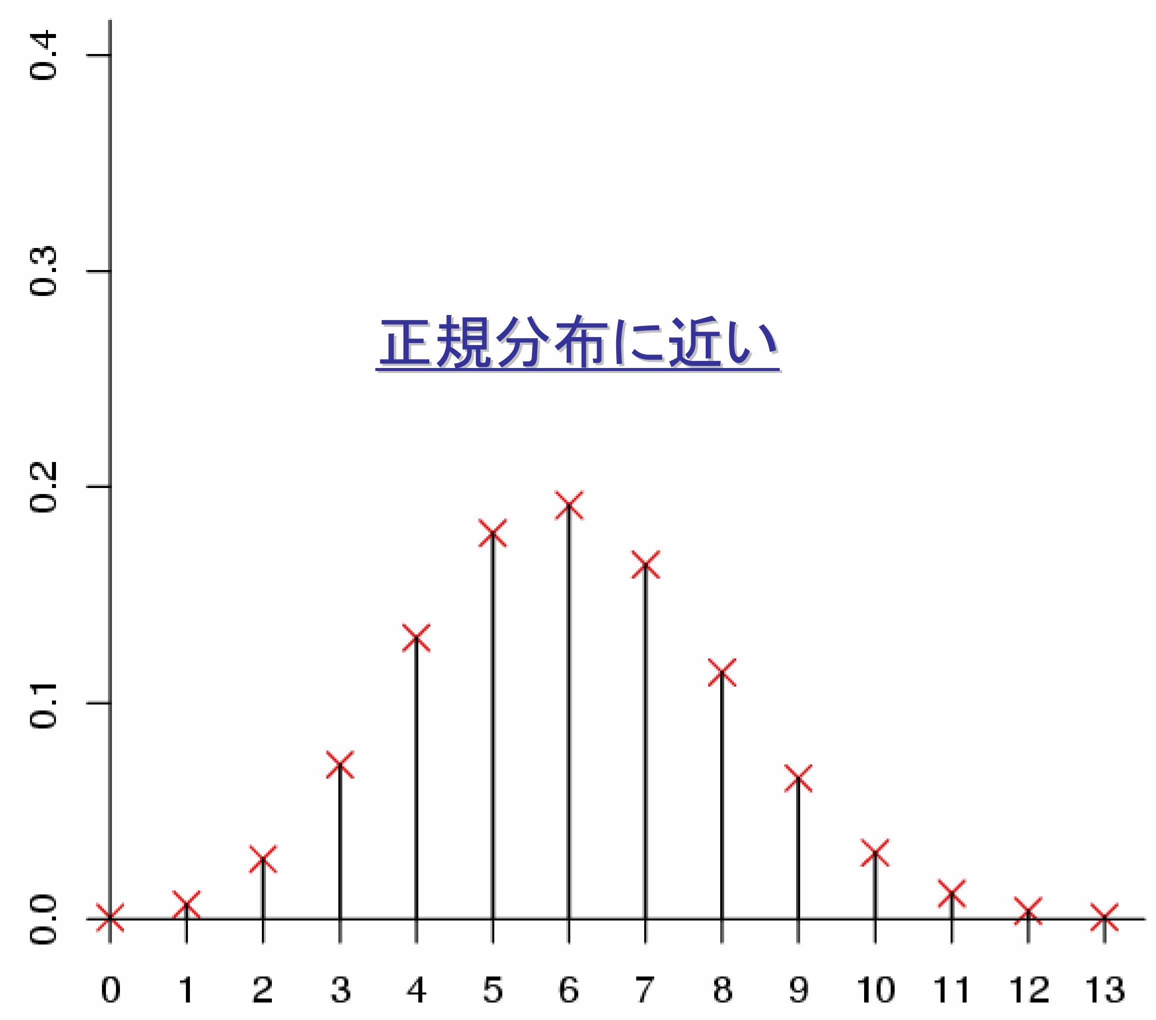
標本数 10，比率 0．3 の二項分布

$n = 10,\ p = 0.3$ の二項分布

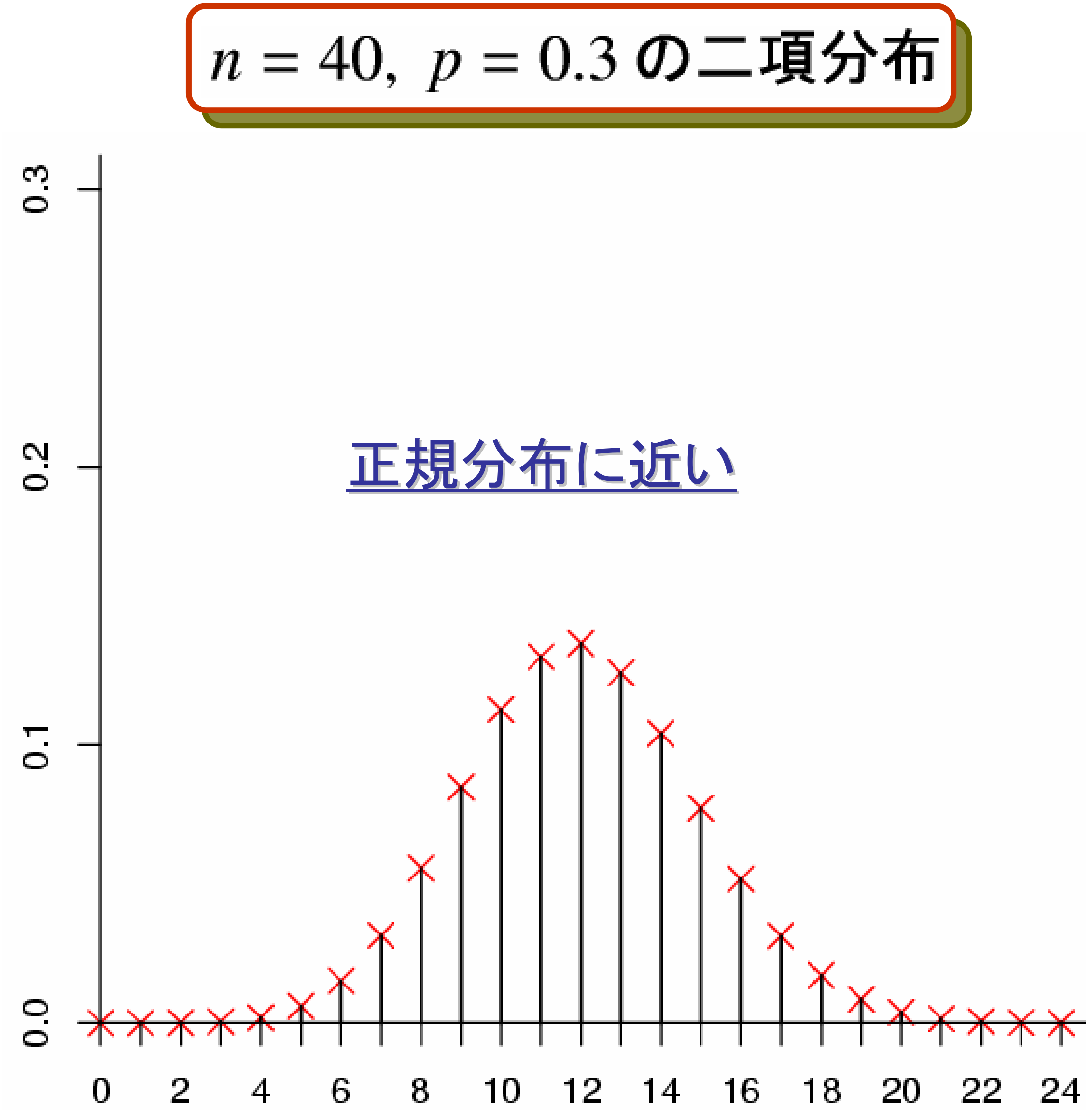


標本数２０，比率０．３の二項分布

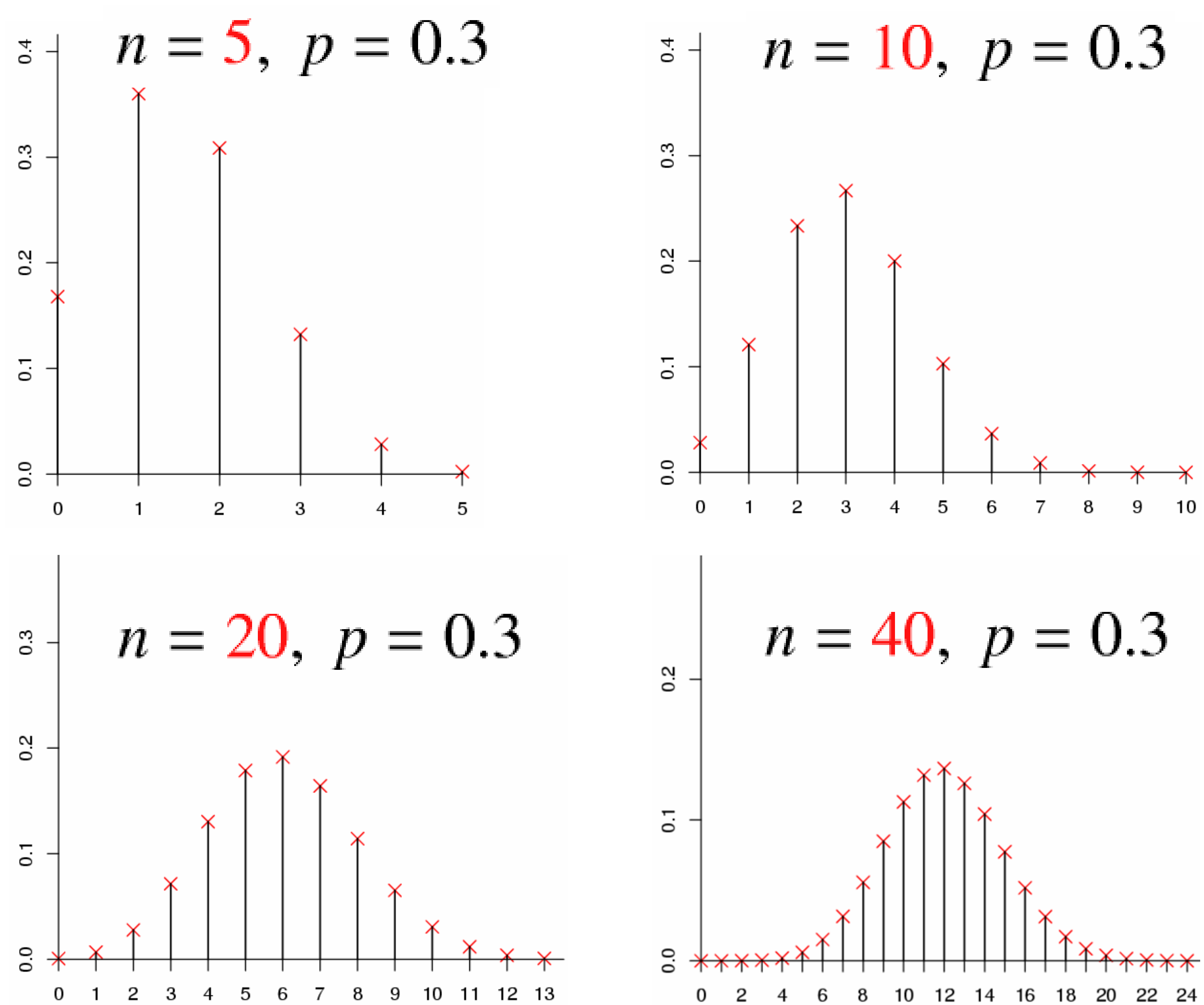
$n = 20,\ p = 0.3$ の二項分布



標本数４０，比率０．３の二項分布



二項分布



$n = 5,\ p = 0.3$
0.0 0.1 0.2 0.3 0.4
0 1 2 3 4 5
$n = 10,\ p = 0.3$
0.0 0.1 0.2 0.3 0.4
0 1 2 3 4 5 6 7 8 9 10
$n = 20,\ p = 0.3$
0.0 0.1 0.2 0.3
0 1 2 3 4 5 6 7 8 9 10 11 12 13
$n = 40,\ p = 0.3$
0.0 0.1 0.2
0 2 4 6 8 10 12 14 16 18 20 22 24

二項分布の正規近似

- 次のとき，二項分布は正規分布でよく近似できる
  - $p < 0.5$ のとき，$np > 5$
  - $p > 0.5$ のとき，$n(1-p) > 5$

- 仮説検定
  - 二項分布の，正規分布への近似が妥当であれば，仮説

$$H_0 : p = p_0$$

のもとで，次のことを用いて検定を行なうことができる。

$$z = \frac{(x/n) - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(p_0, p_0(1-p_0)/n)$$

例：比率の検定

- 種子の発芽率に関する調査
  - A試験所の種子の発芽率 ： 60%
  - B試験所の種子 ：100粒中48粒発芽 ⇒ 発芽率48%と推定
- 帰無仮説
  - B試験所の種子の発芽率(p)と A試験所の種子の発芽率(p_0)は等しい $\quad H_0 : p = p_0$
  - 有意水準5%で帰無仮説の検定を行う

統計量

- 帰無仮説が真のとき, $p_0 = 0.60$ より x/n の平均と分散は
  (x : B 試験所の種子 n 粒のうち発芽したものの数)

  平均 : $p_0 = 0.60$

  分散 : $p_0 q_0/n = p_0(1 - p_0)/n$
  $$= 0.6 \times 0.4/100 = 0.0024 = 0.049^2$$

- $p_0 = 0.6 > 0.5,\ nq_0 = 40 > 5$ より正規分布で近似できる

⇩

仮説のもとでは, x/n の分布は
平均 0.60, 分散 0.049^2 の正規分布で近似できる

検定統計量

- **検定統計量**
  - $$z = \frac{(x/n) - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.48 - 0.60}{0.049} = -2.45$$
  - 標準正規分布の両側5%点：1.96
- **検定結果**
  - 帰無仮説を有意水準5%で棄却する
  - B試験所の種子の発芽率はA試験所の種子の発芽率と異なるといえる

符号検定

例：符号検定

A社とB社のアイスクリームでどちらがおいしいかを調査

- 帰無仮説
  - A社とB社のアイスクリームのおいしさに差はない
- 基本的な考え方
  - 「A社の方がおいしい」という回答が得られる確率と，「B社の方がおいしい」という回答が得られる確率を比較し，これらの間に大きな差があるか調べる

データ

- データ
  - 若い女性20人を無作為に抽出し，A社とB社のアイスクリームを試食してもらい，どちらがおいしいか回答を得た
  - 両者が同じ条件で比較されるよう、配慮して試験を行った(後述)

被験者番号	1	2	3	4	5	6	7	8	9	10
反応	−	+	−	−	+	−	−	−	−	−
被験者番号	11	12	13	14	15	16	17	18	19	20
反応	+	−	−	−	−	−	−	+	+	−

+ ：A社の商品がおいしいと回答，− ：B社の商品がおいしいと回答

帰無仮説

- 記号の定義
  - p : A 社のアイスクリームをおいしいと答える確率
  - q : B 社のアイスクリームをおいしいと答える確率
- 帰無仮説
  - おいしいという回答を得る確率はA社もB社も等しい

$$H_0 : p = q = \frac{1}{2}$$

回答についての確率

- 帰無仮説のもとで,「A社の方がおいしい」と答える確率は二項分布にしたがう
  - n 人中 x 人が A 社の商品をおいしいと答える確率

$$ {}_nC_x p^x (1-p)^{n-x} = {}_nC_x p^x q^{n-x} $$

  - B 社についても同様に

$$ {}_nC_x q^x (1-q)^{n-x} = {}_nC_x q^x p^{n-x} $$

帰無仮説のもとでの確率

- **帰無仮説が真のときの確率**
  - 帰無仮説のもとでは

$$H_0 : p = q = \frac{1}{2}$$

より, 「おいしい」と回答を得る確率は両社で等しく,

n 人中 x 人が A 社の商品をおいしいと答える確率

$=$ n 人中 x 人が B 社の商品をおいしいと答える確率

$$= {}_nC_x \left(\frac{1}{2}\right)^n$$

基本的な考え方

基本的な考え方

n 人中 x 人「A社の方がおいしい」という回答が得られる確率と，
n 人中 x 人「B社の方がおいしい」という回答が得られる確率を比較し，
これらの間に大きな差があるか調べる

標本数がある程度大きいとき，下のような偏ったデータが
得られたときは，帰無仮説を棄却すると考える

$+$	$+$	$+$	$+$	$+$	$\cdots$	$+$
$-$	$-$	$-$	$-$	$-$	$\cdots$	$-$

全ての回答が ”+” か ”－”

$+$	$-$	$+$	$+$	$+$	$\cdots$	$+$
$-$	$-$	$+$	$-$	$-$	$\cdots$	$-$

1人だけが ”+” か ”－”

確率の計算

- 全ての回答が “+” か “−” である場合の確率

$$ {}_nC_0\left(\frac{1}{2}\right)^n + {}_nC_0\left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n + \left(\frac{1}{2}\right)^n = 2\left(\frac{1}{2}\right)^n $$

- 1人が “−” で残りが “+”，もしくは
  1人が “+” で残りが “−” である確率

$$ {}_nC_1\left(\frac{1}{2}\right)^n + {}_nC_1\left(\frac{1}{2}\right)^n = n\left(\frac{1}{2}\right)^n + n\left(\frac{1}{2}\right)^n = 2n\left(\frac{1}{2}\right)^n $$

棄却限界点の設定

- 以下同様にして確率が求まる。ゆえに,

$$2\,{}_nC_0\left(\frac{1}{2}\right)^n + 2\,{}_nC_1\left(\frac{1}{2}\right)^n + 2\,{}_nC_2\left(\frac{1}{2}\right)^n + \cdots + 2\,{}_nC_k\left(\frac{1}{2}\right)^n \leq 0.05$$

  を満たす最大のkが，棄却限界点となる
- 記号
  - $\alpha = 2({}_nC_0 + {}_nC_1 + {}_nC_2 + \cdots + {}_nC_k)\left(\frac{1}{2}\right)^n$

棄却限界点

$$\alpha = 2({}_nC_0 + {}_nC_1 + {}_nC_2 + \cdots + {}_nC_k)\left(\frac{1}{2}\right)^n$$

k	${}_{20}C_k$	確率 α
0	1	0.0000
1	20	0.0000
2	190	0.0004
3	1140	0.0026
4	4845	0.0118
5	15504	0.0414
6	38760	0.1153

1%点

5%点

検定結果

- アイスクリームの例
  - A社の方がおいしい：5人
  - B社の方がおいしい：15人
- 検定結果
  - 右の表より，有意水準5%で帰無仮説を棄却する
  - 若い女性について，「B社のアイスクリームの方がおいしい」と判定された

k	${}_{20}C_k$	確率 α
0	1	0.0000
1	20	0.0000
2	190	0.0004
3	1140	0.0026
4	4845	0.0118
5	15504	0.0414
6	38760	0.1153

試験における注意事項

- 試験では，判定に影響する要因をできるだけ取り除き，<u>同じ条件で</u>比較できるよう配慮する必要がある．
- 今回の例では・・・
  - アイスクリームを食べる<u>順番</u>を，<u>乱数表</u>を用いて決める
    - 各被験者ごとに乱数を割り付け，偶数ならばA社，奇数ならばB社のアイスクリームを先に食べる
    - 先に食べた方がおいしいと感じる偏りを無くすため
  - アイスクリームを食べる前に<u>水でうがい</u>をする
    - 先に食べたアイスクリームの味を消すため

正規近似を用いた符号検定

- 標本数が大きい場合には，次の統計量の分布が標準正規分布で近似できることを用いて検定を行うことができる

$$\frac{n - 2k - 1}{\sqrt{n}}$$

  - この値が1.96よりも大きい ⇒ 有意水準5%で棄却
  - この値が2.58よりも大きい ⇒ 有意水準1%で棄却
  - k の値は，”+” か “ー” の数が小さい方とする