しぜん言語処理のための

NLP 検定入門

@D-Lec

こんな人が対象者

教員から「検定しろ」って言われたけど
よくわかんない…

Rとかoctaveでできるけど,
中で何やっているのかわからない！

学部時代のトラウマがよみがえる…

＼怖くないよ／

NLP 検定

今日話す内容

- 二項検定
- 母比率の検定
- ２群の母比率の差の検定（Prop Test）
- マクネマー検定（McNemar Test）
- 符号検定（Sign Test）
- ウィルコクソンの符号検定（Wilcoxon Signed rank Test）
- Approximate Randamization Test for F-measure

まずは，よくある例題で
検定の気持ちを思い出してみよう！

検太くんが定子ちゃんの前で１００回コイントスを行い，表61裏39という結果になりました。

このコインは歪んでいる！

偶然じゃない？

偶然でこんなに偏るもんか！

歪んでなくたって，これぐらい偏るって

そんな確率はほとんどない！

「ほとんど」って具体的には？

今から計算してやる！　そのかわり，５%より小さかったら，素直に歪んでいることを認めろよ！

参考：プログラミングのための確率統計 平岡和幸，堀玄 共著 オーム社 (2009).

これが検定の気持ちです

簡単にまとめると

①　コインは歪んでいないと仮定。

②　１００回投げて表が６１の確率を計算。

③　計算した確率が５％よりも低ければ，
「コインは歪んでない」という仮定が
間違っていると認める。

①コインは歪んでいないと仮定

いわゆる
帰無仮説

- i 回目のコイントスの結果を

$$X_i = \begin{cases} 1(\text{表}) \\ 0(\text{裏}) \end{cases}$$

- コインは歪んでいないので,

$$P(X_i = 1) = 0.5, \quad P(X_i = 0) = 0.5$$

②100回投げて表が６１の確率を計算

いわゆる
p値

- $X_i = \begin{cases} 1(表) \\ 0(裏) \end{cases}$ なので，100回コインを投げて表の出た数は，

いわゆる
検定統計量

$$S = \sum_{i=1}^{100} X_i$$

- 各トスは独立なので，このSは二項分布に従う.

$$S \sim Bn(100, 0.5)$$

二項分布

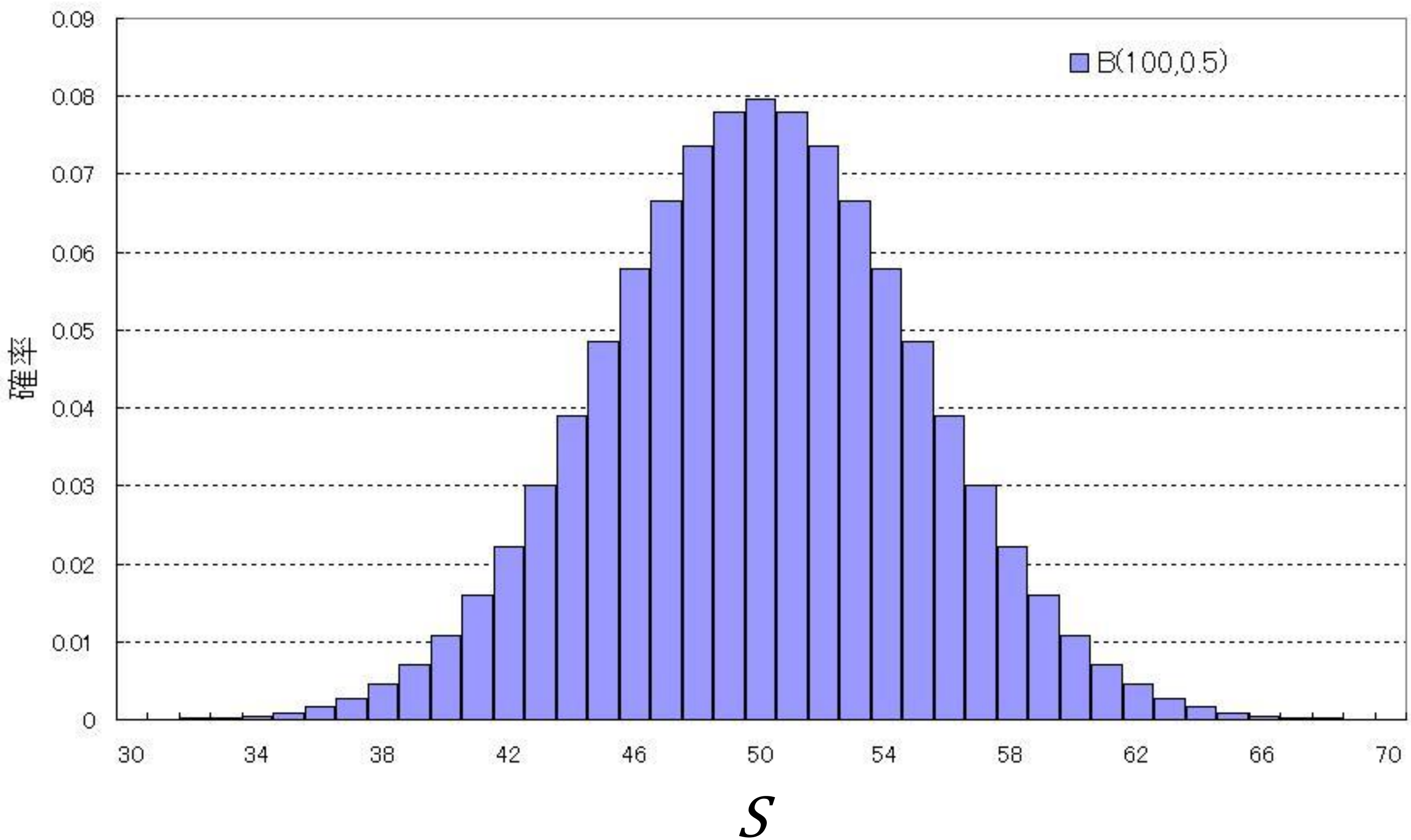
$$P(x;n,p) = {}_nC_x\, p^x (1-p)^{n-x}$$

確率変数xが二項分布に従うとき，

$$x \sim Bn(n,p)$$

と書く．

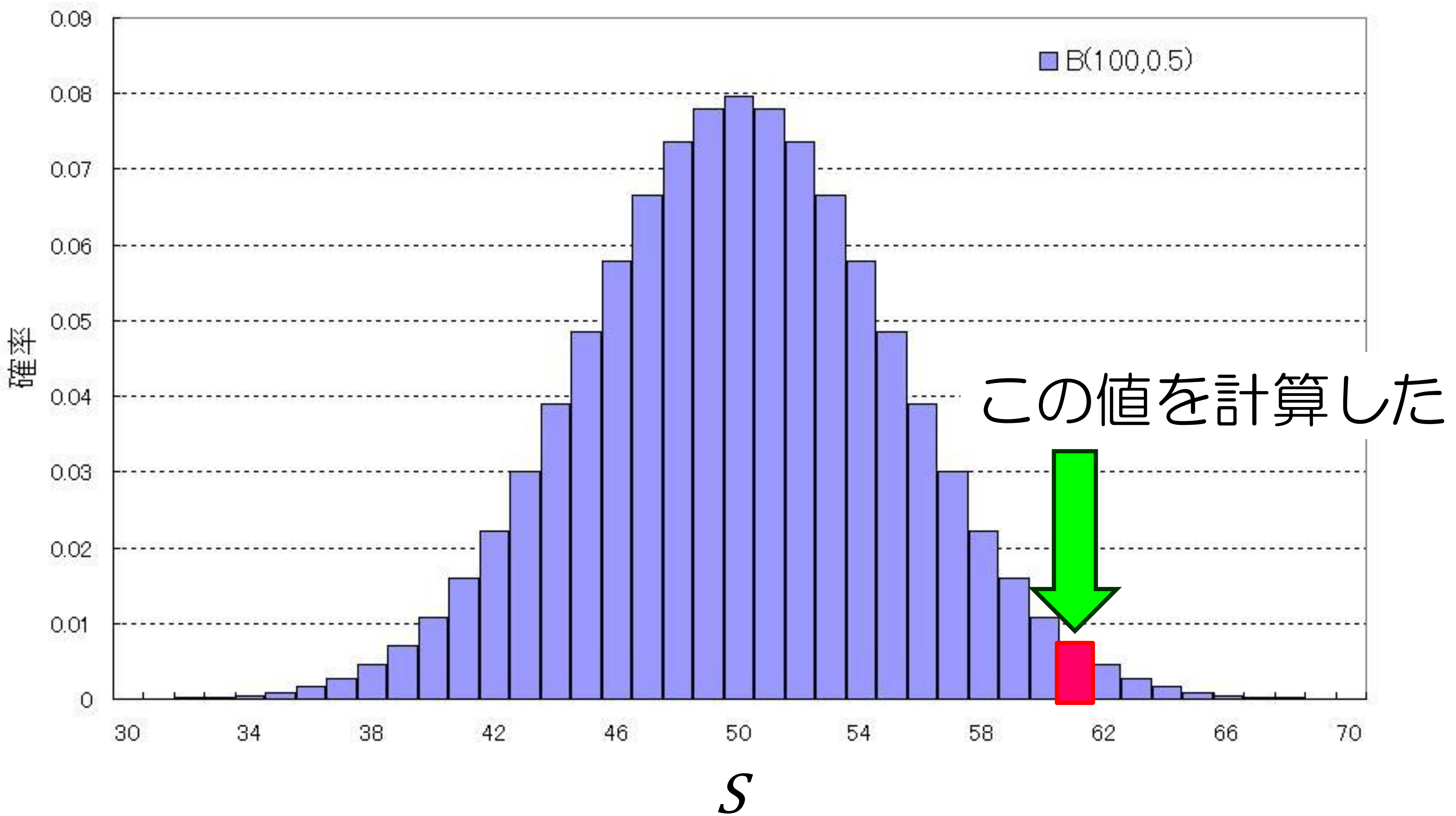
二項分布（n=100, p=0.5）



二項分布より，表61の確率を計算

$$P(S = 61) = {}_{100}C_{61} \times 0.5^{61} \times 0.5^{39}$$

二項分布（n=100, p=0.5）



でもさ，表61を「歪んでいる」っていうことは，
表62も63も…100も「歪んでいる」って
認めることになるよね？

定子

も，もちろん！

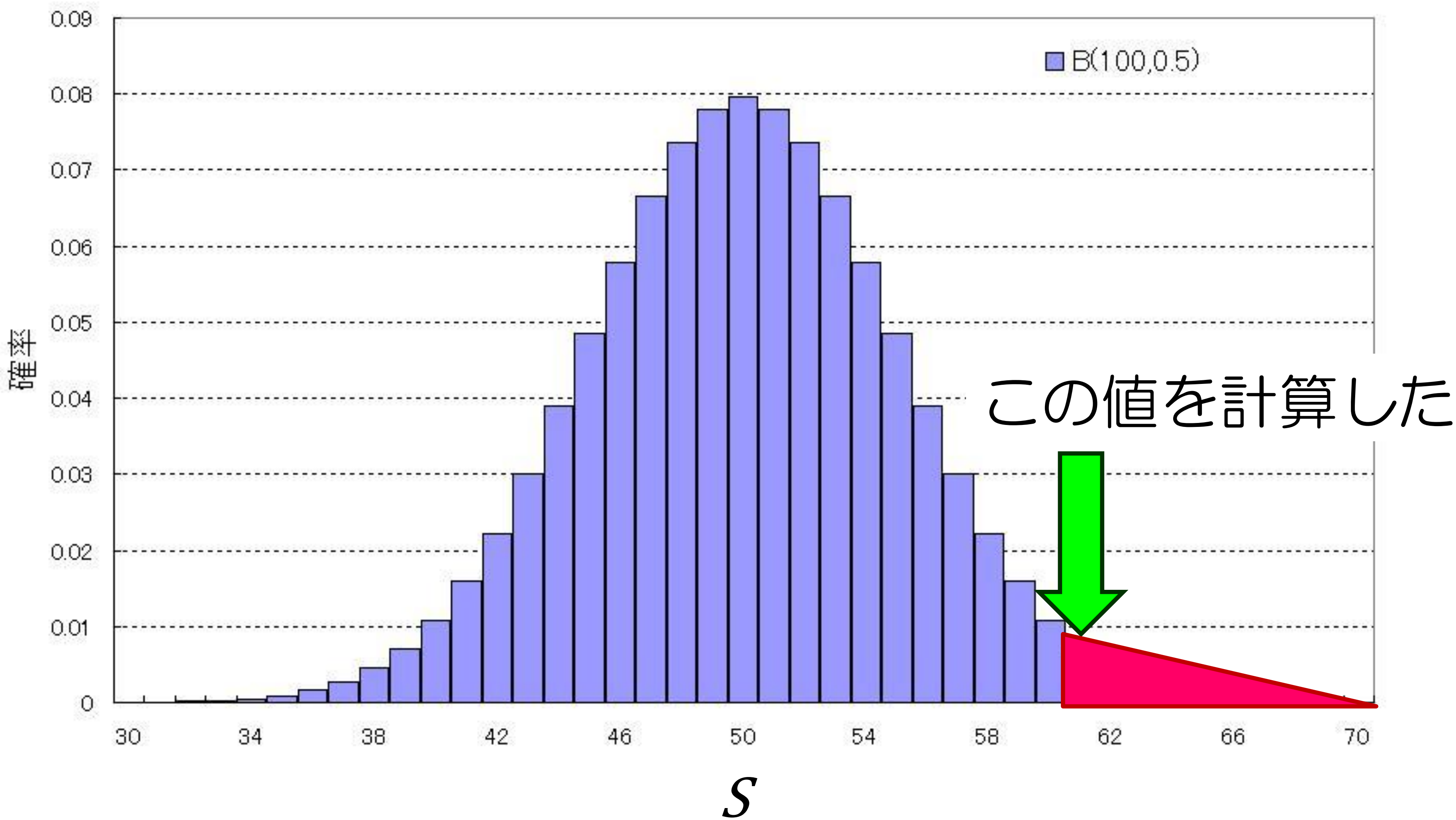
検太

これで文句ないだろ！

$$P(S \geq 61) = \sum_{i=61}^{100} {}_{100}C_i \times 0.5^i \times 0.5^{100-i}$$

片側検定

二項分布（n=100, p=0.5）



あ。あとさ，表61が歪んでいるっていうなら，
裏61も歪んでるんだよね？

定子

そ，そうだね…

検太

これならどうだ！

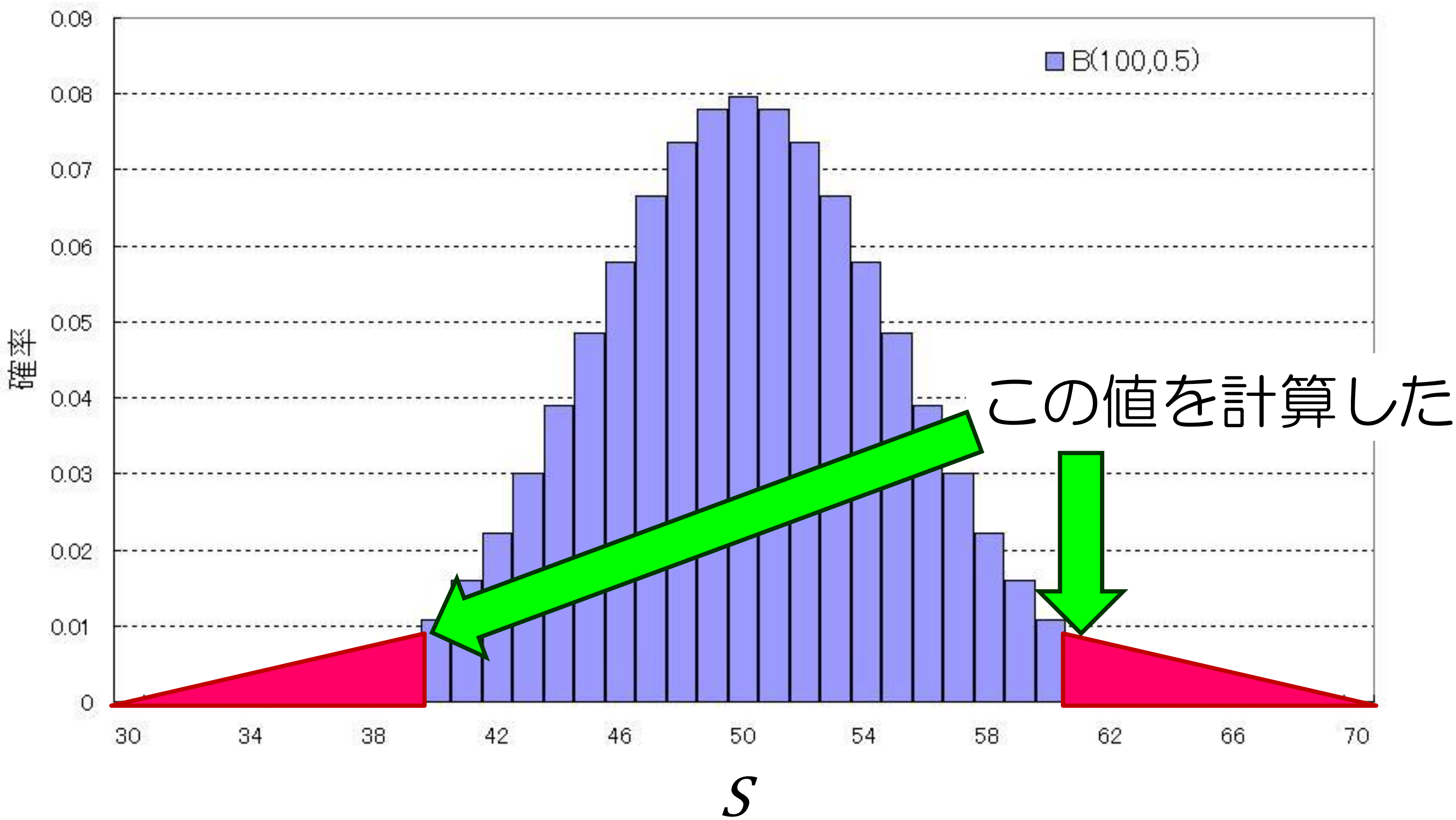
$$P(S \geq 61) = \sum_{i=61}^{100} {}_{100}C_i \times 0.5^i \times 0.5^{100-i}$$

$$P(S \leq 39) = \sum_{i=0}^{39} {}_{100}C_i \times 0.5^i \times 0.5^{100-i}$$

$$P(S \geq 61) + P(S \leq 39) = 0.035200 \ldots$$

両側検定

二項分布（n=100, p=0.5）



いわゆる
有意水準

③計算した確率が５％よりも低ければ，
「コインは歪んでない」という仮定が
間違っていると認める。

いわゆる
対立仮説

確率は約3.5%だ！
5%より低いぞ！

検太

間違っていたと認める…

定子

簡単にまとめると

帰無仮説

①　コインは歪んでいないと仮定。

p値

②　１００回投げて表が６１の確率を計算。

有意水準

③　計算した確率が５％よりも低ければ，
「コインは歪んでない」という仮定が
間違っていると認める。

対立仮説

実はこれが二項検定

でも，二項検定の計算って

$\sum$ が，めんどくさいんだよな…

そんなときは…

中心極限定理

説明しよう！

- **中心極限定理**とは，$S=\sum_{i=1}^{n}X_i$ が平均 μ，分散 σ^2の分布に従うとき，nが大きければ，元の分布に関わらず

$$S=\sum_{i=1}^{n}X_i \sim N(n\mu, n\sigma^2)$$

と近似できる便利な定理である．

正規分布

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

x は平均 μ，分散 σ^2 の正規分布に従う：

$$x \sim N(\mu,\sigma^2)$$

さらに正規分布の標準化を使うと

$$\mathrm{Z}(S) = \frac{S - \mu}{\sigma} \sim N(0,1)$$

つまり…？

$$S \sim Bn(n,p)$$

- 二項分布の平均：np
- 二項分布の分散：np(1-p)より

$$\mathrm{Z}(S) = \frac{S - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

コイントスの問題に適用すると…

検定統計量：

$$Z(61) = \frac{61 - 100 \times 0.5}{\sqrt{100 \times 0.5 \times 0.5}} = 2.2$$

統計の本に載ってる巻末の正規分布表より：

$$\mathrm{P}\big(Z \geq Z(61)\big) = 0.0139$$

結果

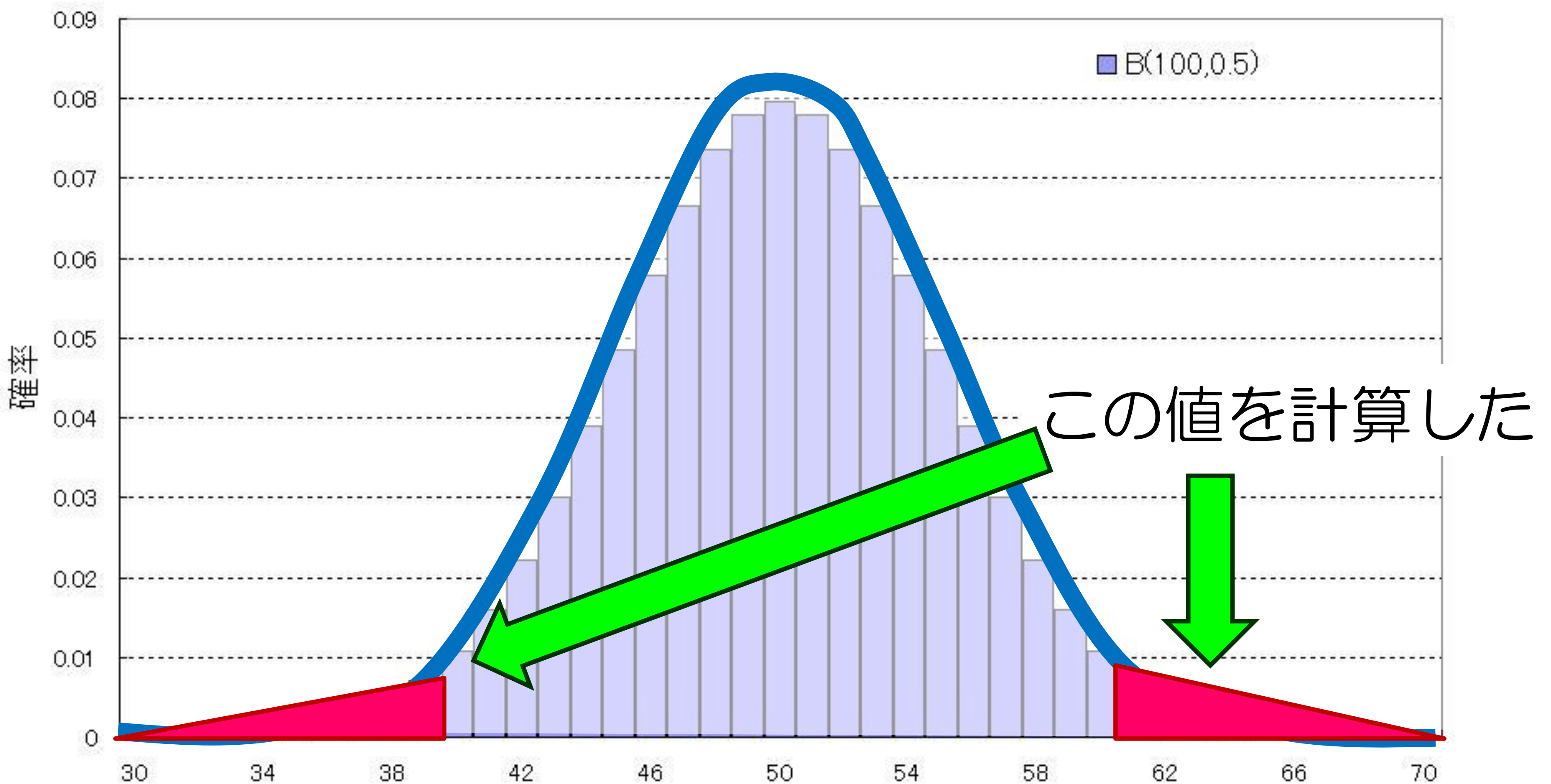
$$P\left(Z \geq Z(61)\right)+P\left(Z \leq Z(39)\right)$$

$$= 0.0139 \times 2 = 0.0278$$

p値は約2.8%だ！
やっぱり5%より低いぞ！

検太

二項分布を正規分布に近似



S

これがなんと
母比率の検定

Sが二項分布ではなく
「正規分布に従う」
と近似したのが二項検定と違う点

母比率って何？

母比率とは

$S \sim Bn(n, p)$ の p

コイントスの問題で言うところの
「表の出る確率：0.5」

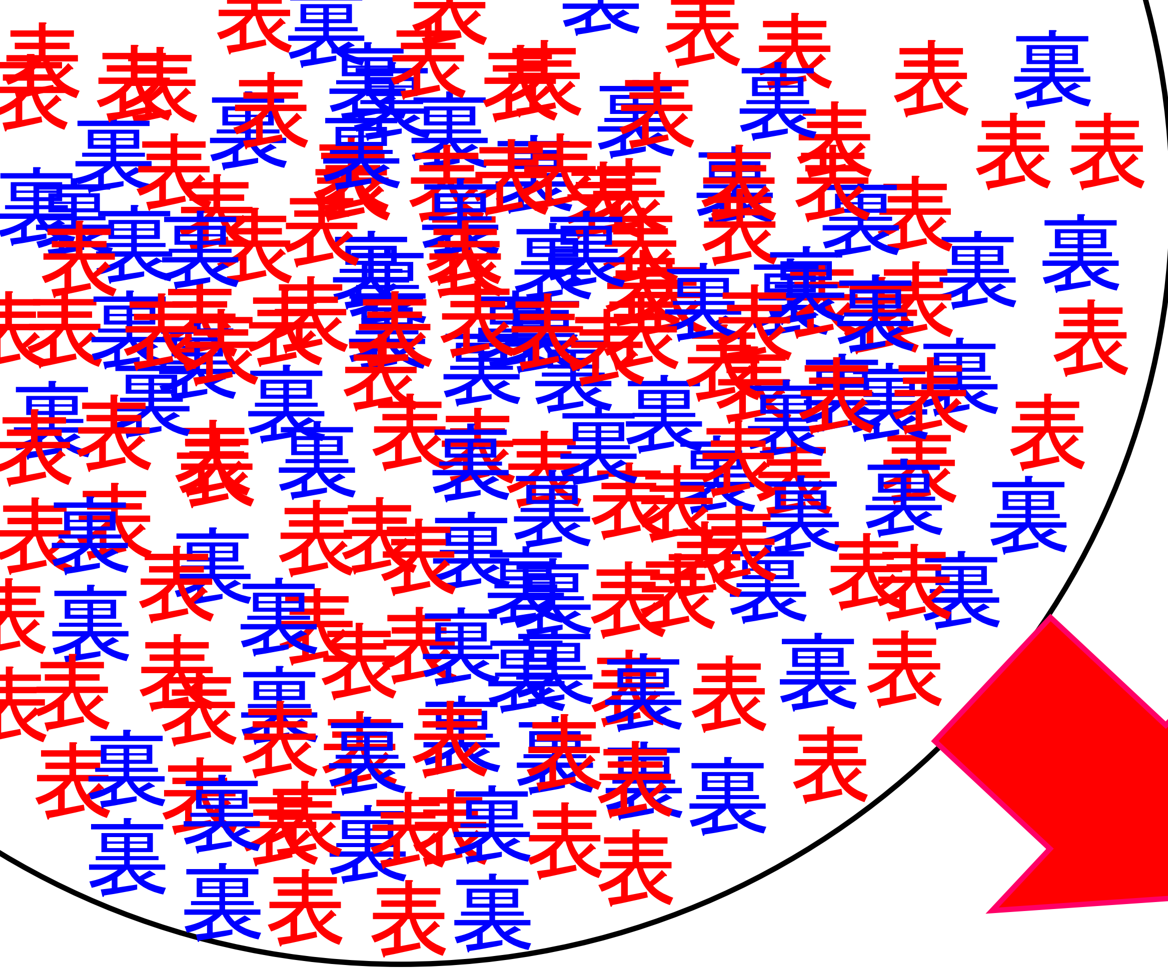
ポイント:
コイントスの問題では,
母比率pが0.5でないことを検定により示した.

もっと本質的な話をすると…

- １００回のコイントスは,
- 「表」と「裏」が

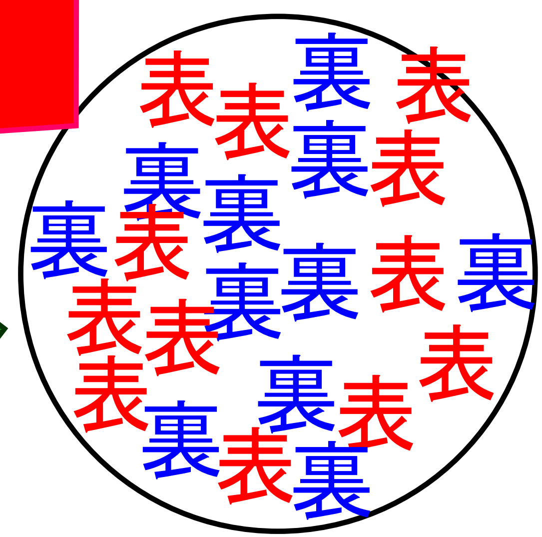
$$p : 1\text{-}p$$

の比率で含まれるめちゃくちゃ大きな集まり（母集団）からランダムに要素を１００個取り出してくることと見なすことができる.



この集まりが
母集団

母集団の比率が
母比率 p



１００個取り出したのが
標本

標本中の比率が
標本比率 $\bar{X} = S/100$

ポイント：標本比率は母比率に近い値になりやすい

母比率の検定とはすなわち…

- 母比率がpでないことを標本比率を使って認めさせる方法。

母比率の検定まとめ

① 母比率をpと仮定。

② 検定統計量（S∝標本比率）は，試行回数nが小さければ二項分布に従い，大きければ正規分布で近似する。

③ Sからp値（Sの出る確率）を計算。

④ p値が有意水準より低ければ，母比率がpであることを棄却し，母比率がpでないことを認める。

さて次は，
２群の母比率の差の検定

２群の母比率の差の検定

- 標本比率の差から，２つの標本の母比率が同じではないことを示す。

またコイントスの例で考えよう！

検太くんが別のコインを今度は２００回投げて，
表130，裏70という結果になりました。

標本比率は130/200=0.65だ！
さっきのコインは（0.61）だったから
0.04も大きい！
つまりも表が出やすい！
（母比率が違う！）

これくらいの差は
偶然の誤差じゃないの？

なら，また検定だ！
母比率が同じだとしたときに
これくらいの差が出る確率が５％
より小さかったら，
母比率が違うことを認めろよ！

最初に投げたコインをコインA,
今投げたコインをコインBとします。

まず，コインAとBの出やすさ，
つまり母比率pは同じだと仮定します。

$$p_A = p_B = p$$

これが今回の帰無仮説。

次に，検定統計量

- 母比率の検定の時と同じく，表の出る回数Sは二項分布に従います。

$$S_A \sim Bn(100, p)$$

平均: 100p
分散: 100p(1-p)

$$S_B \sim Bn(200, p)$$

平均: 200p
分散: 200p(1-p)

ここで

中心極限定理

中心極限定理を使うと…

$$S_A \sim N(100p, 100p(1-p))$$

$$S_B \sim N(200p, 200p(1-p))$$

さらに正規分布の線形性より…

コインAの標本比率

$$\bar{X}_A = \frac{S_A}{100} \sim N\left(p, \frac{1}{100} p(1-p)\right)$$

コインBの標本比率

$$\bar{X}_B = \frac{S_B}{200} \sim N\left(p, \frac{1}{200} p(1-p)\right)$$

正規分布の再生性

n個の独立な確率変数X_i (i=1,...,n)がそれぞれ$N(\mu_i, \sigma_i{}^2)$に従っているとき,

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i{}^2 \sigma_i{}^2\right)$$

が成立する。ただし, a_iは定数とする。

なにがしたいかというと…

- 正規分布の再生性を使うと…

$$X_1 \sim N(\mu_1, {\sigma_1}^2)$$

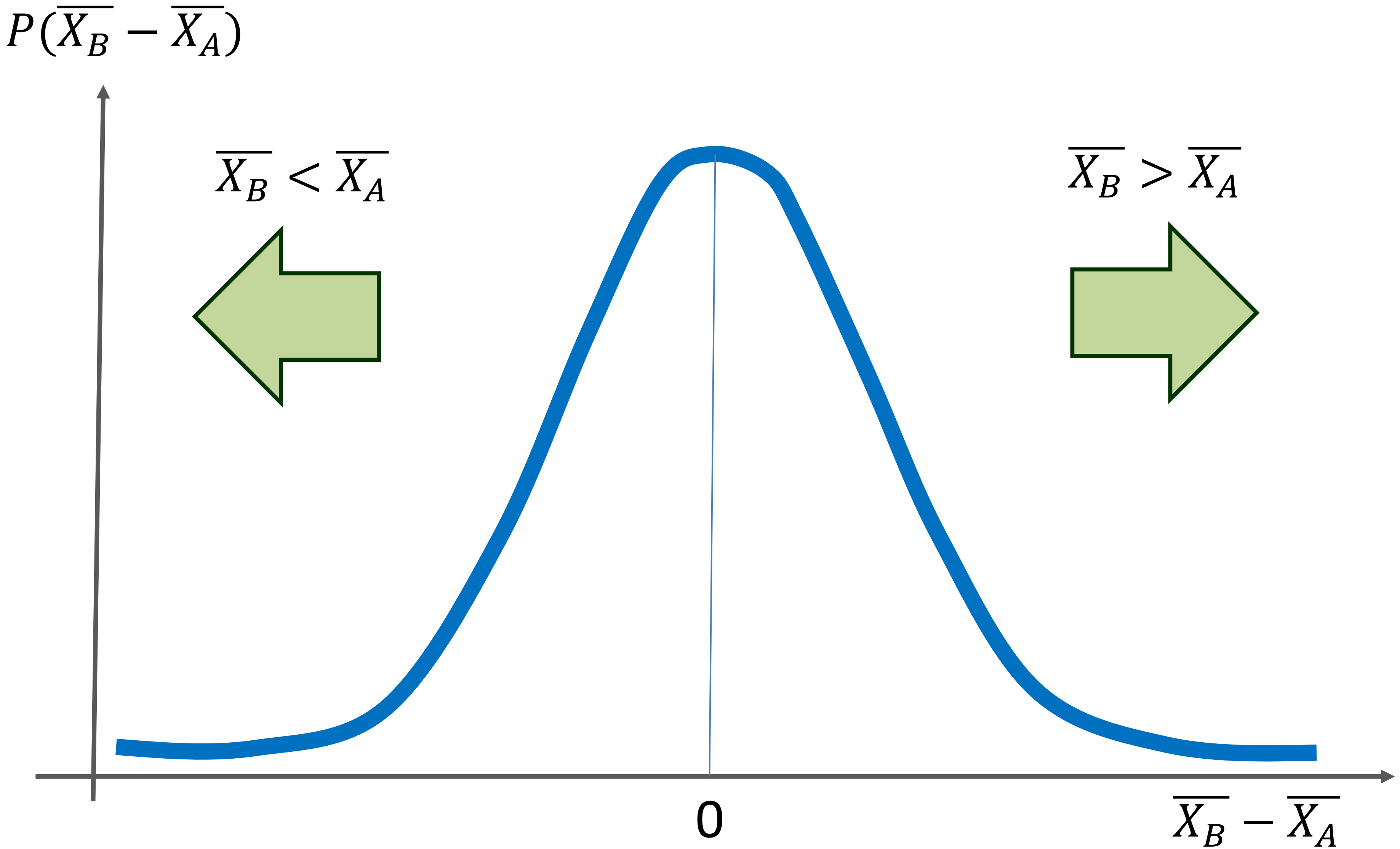
$$X_2 \sim N(\mu_2, {\sigma_2}^2)$$

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$$

つまり，

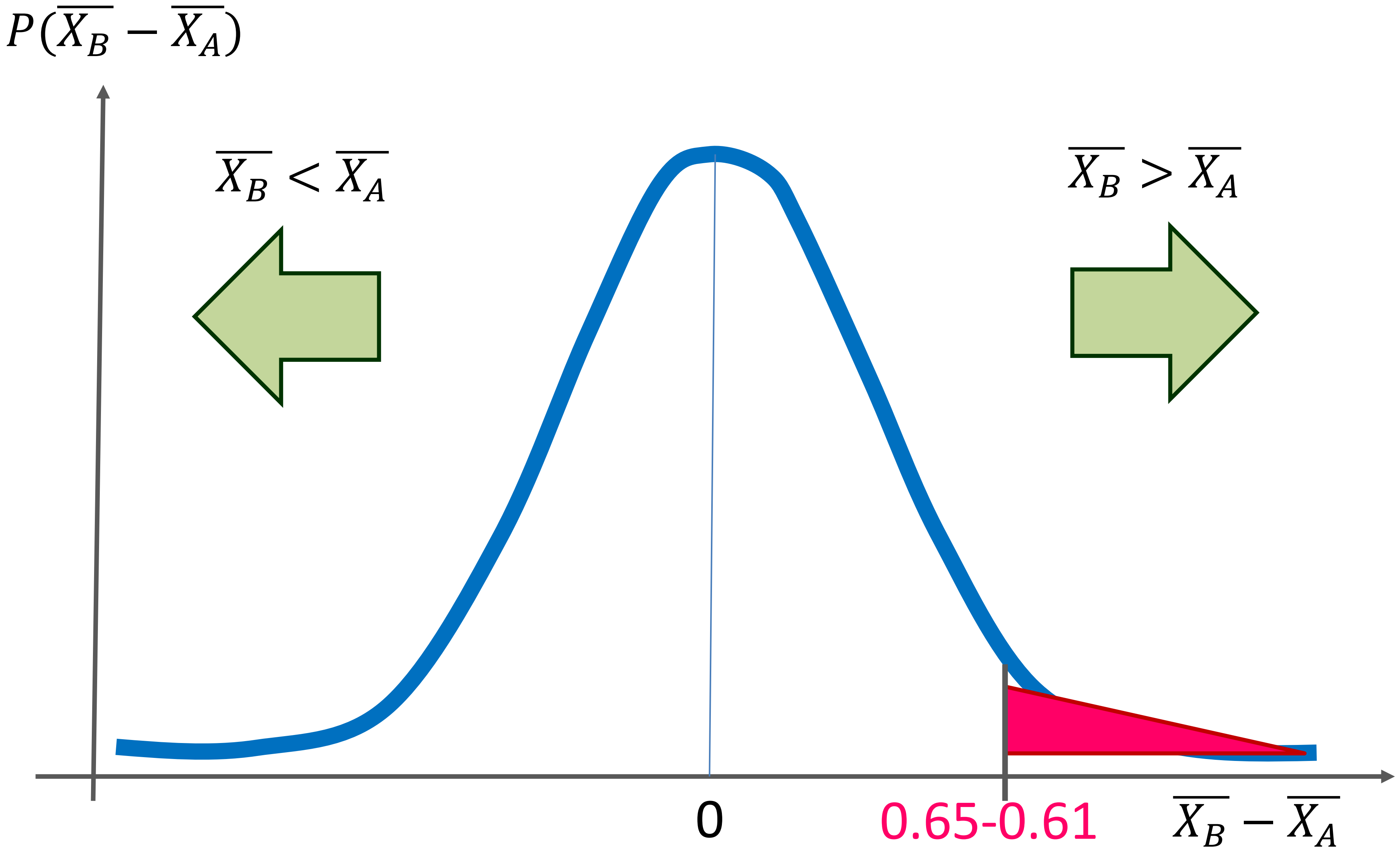
$$\overline{X_B} - \overline{X_A} \sim N\left(0, \left(\frac{1}{200} + \frac{1}{100}\right) p(1-p)\right)$$

標本比率の差の上に
正規分布が乗ったぞ！



２群の母比率の差の検定では,
この標本比率の差が
偶然起こるにしては確率が低すぎる
ことを示します.

この確率を計算(片側検定)



まずは標準化

$$Z = \frac{\overline{X_B} - \overline{X_A}}{\sqrt{\left(\frac{1}{200} + \frac{1}{100}\right) p(1-p)}} \sim N(0,1)$$

あれ？　母比率pは？

$$p_A = p_B = p$$

及び，標本比率は母比率pに近い値と考えられるため，

母比率pを2つの標本をマージした標本比率で代用する。

$$p = \frac{61 + 130}{100 + 200}$$

$$Z = \frac{\frac{130}{200} - \frac{61}{100}}{\sqrt{\left(\frac{1}{200} + \frac{1}{100}\right)\frac{61+130}{100+200}\left(1 - \frac{61+130}{100+200}\right)}}$$

$\fallingdotseq 0.67$

統計の教科書の付録の正規分布表より,

$$P(Z \geq 0.67) = 0.24825$$

確率は約25%で
5%よりずっと高い！

定子

だから，「２つのコインの母比率
は同じではない」とは言えない。

2群の母比率の差の検定まとめ

① ２つの標本の母比率は同じであると仮定。

② 検定統計量（標本比率の差）を計算。

③ 「母比率に差はない」と仮定した上での正規分布上で $\overline{X_B} - \overline{X_A}$ の確率(p値)を計算。

④ p値が有意水準以下ならば，①の仮定を棄却し，「２つの標本の母比率が同じではない」ことを認める。

＼ね。全然怖くない／

NLP 検定