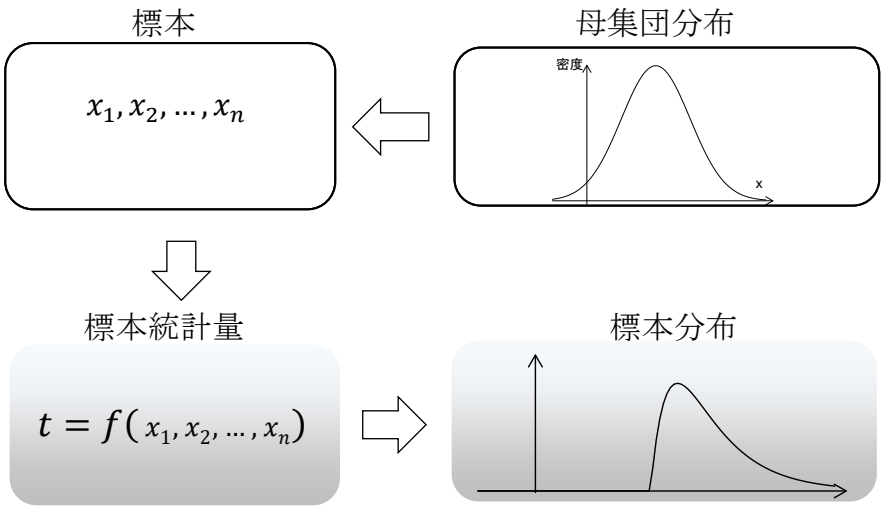
第1講　推測統計学の四つの重要概念（名義尺度編）

学びのポイント

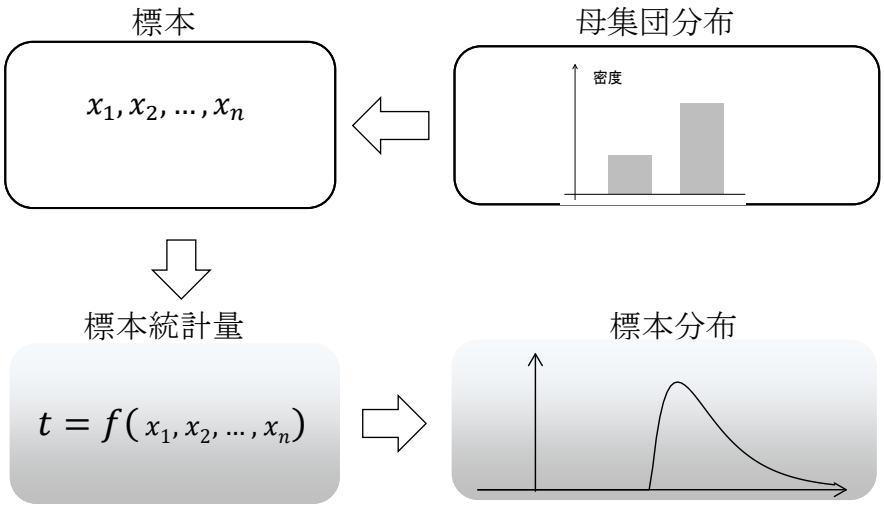
- □ 比率尺度データだけでなく、名義尺度データでも**母集団**、**標本**、**統計量**、**標本分布**という四つの概念が想定されることが分かる。
- □ 名義尺度の従属変数が従う分布には、**ベルヌーイ分布**、**二項分布**、**カテゴリカル分布**、**多項分布**が存在することが理解でき、それらの相互の関係が説明できる。
- □ 二項分布のサイズ n の値を大きくした極限に、**正規分布**、**ポワソン分布**が出現することが理解でき、どのようなときにどちらになるのか説明ができる。
- □ 標準正規分布に従う確率変数の和が従う分布としてχ^2分布という分布が提案されていることが分かる。
- □ 名義尺度データの主な関心は頻度を扱うことであることが分かり、頻度には、**粗頻度**と**相対頻度**という二つの区別があることが理解できる。
- □ 相対頻度（表）に基づく統計量として**情報量**があり、情報量の表に対してエントロピーが定義されることが分かる。
- □ **エントロピー**は、確率分布のデコボコさを測る指標であることが説明できる。
- □ 二つ以上のカテゴリーが存在する相対頻度表に対して、**結合エントロピー**や**条件付エントロピー**が定義されることが分かる。
- □ 確率分布の距離を測る統計量に、**KL 情報量**、**JS 情報量**、**相互情報量**、**χ^2値**などが利用されることが分かり、その違いを説明できる。
- □ 統計量が従う標本分布が理論的に明白ではないときに、**ブートストラップ法**を用いて標本分布を推定するアプローチがあることが理解できる。

見取り図

【前期】


標本
母集団分布
$x_1, x_2, \ldots, x_n$
密度
x
標本統計量
標本分布
$t = f(x_1, x_2, \ldots, x_n)$

【後期】


標本
母集団分布
$x_1, x_2, \ldots, x_n$
密度
標本統計量
標本分布
$t = f(x_1, x_2, \ldots, x_n)$

■ 目標

前期の「言語統計学 A」では、t 検定、重回帰分析などの具体的な統計モデルを扱う前に、これらのモデルで利用される基本的な統計量の話をしていました。

この「言語統計学 B」でも最終的な目標は第 2 講以降で扱う統計モデルを熟知して、使いこなすための知識を身に着けてもらうことですが、まずは前期同様この第 1 講では、名義尺度変数に対して計算される基本的な統計量を学びます。前期の比率尺度データについては、次の四つの概念が明確に区別されていました。

（1）標本 Sample

得られたデータのことを標本（サンプル）と呼ぶ。多くの場合全てを報告するのではなく何らかの指標で標本全体を代表させる。

（2）母集団 Population

研究者が想定する「考えられる対象をすべて含む集団」。

（3）（標本）統計量 Sample Statistic

標本から計算される指標を（標本）統計量と呼ぶ。

（4）標本分布 Sampling distribution

母集団から標本がサンプリングされる際の確率的な揺らぎのせいで、統計量が見せる確率的挙動を表したもの。

これらの四つの概念は、後期で扱う名義尺度データについてもそのまま当てはまります。しかし、後期の主眼は、あるカテゴリーが何回登場したのかという「頻度」（あるいは「頻度表」）です。連続値（実数）を取る比率尺度データとは異なり、標本が離散値（自然数）しかとらないという制約を持つので、前期とは異なる扱いが必要です。

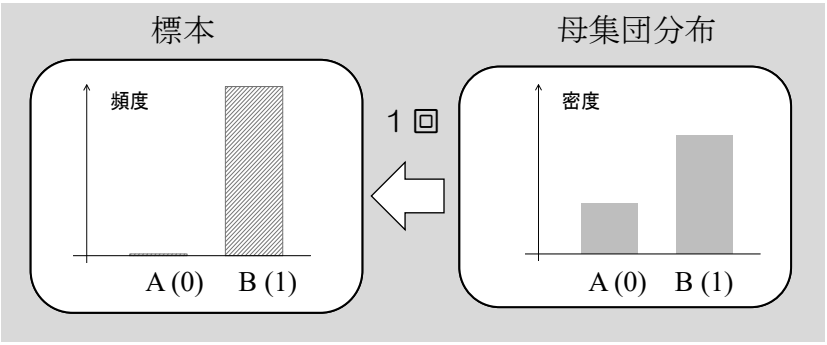
そこで、第 1 講の前半では、この頻度をもたらす確率分布たちを新しく学んでいきます。これらの分布は相互に関連しており、それぱかりか、前期に習った正規分布さえ今回習う分布から派生したものだという点が明らかになり、確率分布一般に対する理解も深まることでしょう。

第 1 講の後半では、頻度（表）にまつわる統計量とそれらが従う標本分布の想定について学んでいきます。名前にはなじみがないものが多いかもしれません。しかし、一つ一つの概念はつながっていて、丁寧に追いかけていくと、それらは非常に納得のいく理由で提唱されてきた概念であることが分かるでしょう。

🕮 ノート１　母集団：代表的な確率分布

（１）ベルヌーイ分布 Bernoulli Distribution

これは 1 か 0 という二値の値を取る変数が従う確率分布。
確率πで 1、確率 $1-\pi$ で 0 を取るように設計されている。



① 密度関数　　$p(y|\pi) = \pi^y(1-\pi)^{1-y}$

② 期待値　　$E[y] = \pi$

③ 分散　　$Var[y] = \pi(1-\pi)$

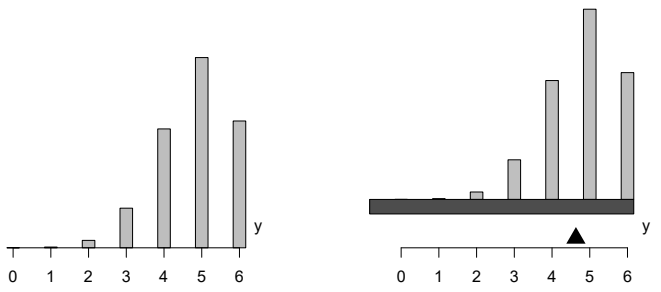
ベルヌーイ試行の具体例

（例１）□ NP の空欄が不定冠詞か定冠詞か。

（例２）「ろっかく」の変換が「六角」か「六画」か。

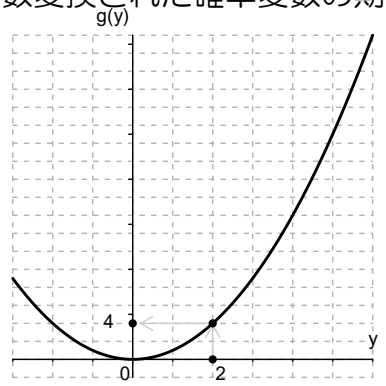
（例３）アンケート調査の解答が「はい」か「いいえ」か。

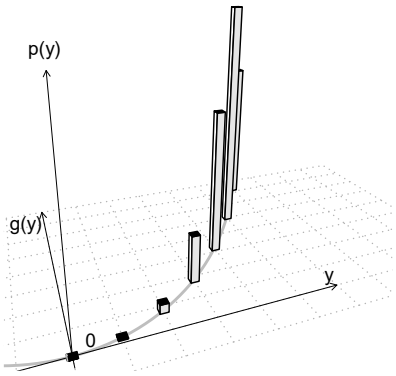
確率分布の期待値

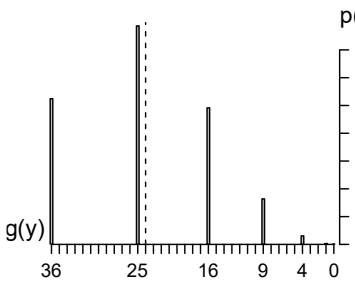


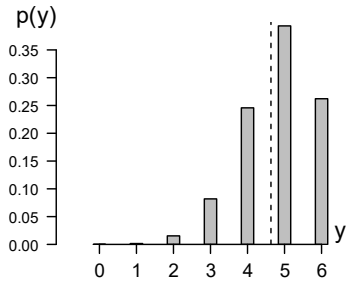
確率変数の変数変換と分散

① 変数変換された確率変数の期待値

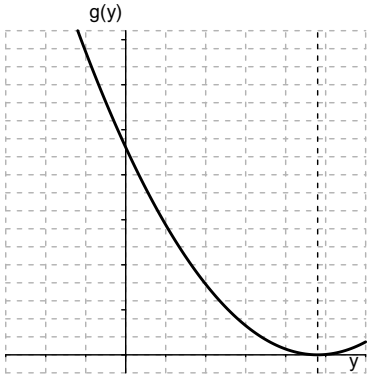


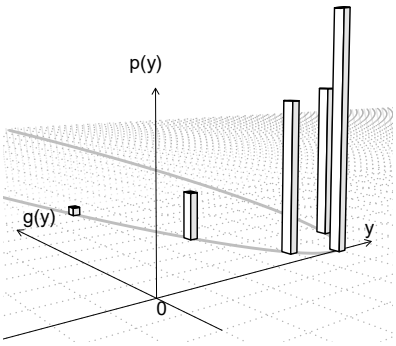


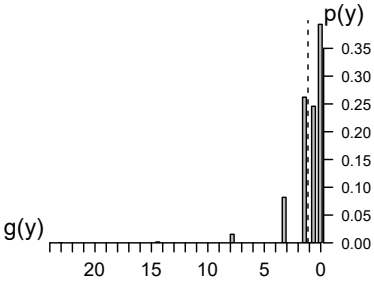


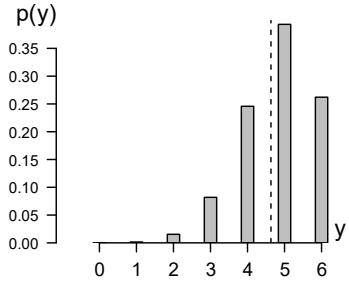


② 確率変数の分散



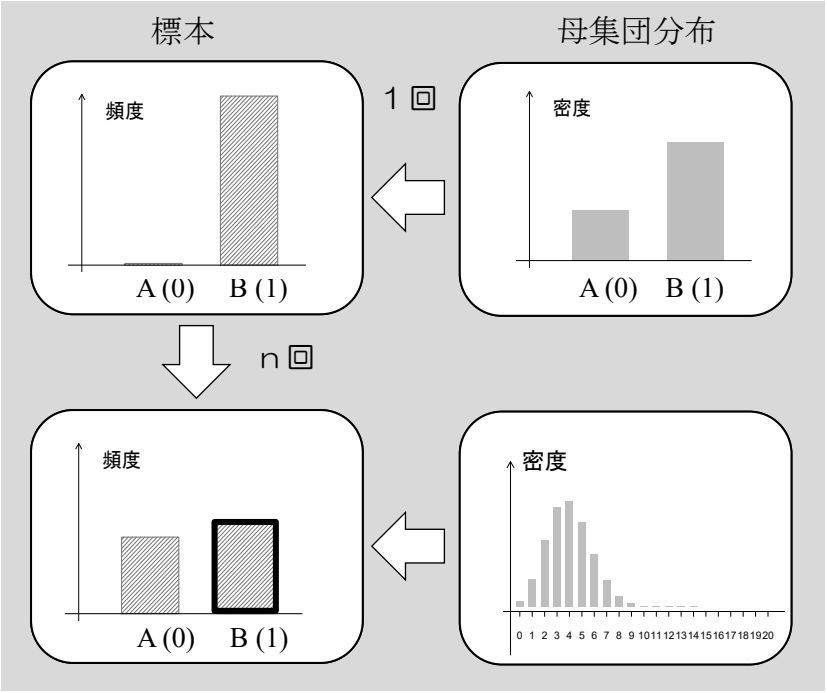






（2）二項分布 Binomial Distribution

これは、確率 π で 1 を取るベルヌーイ試行を n 回行ったときに 1 が出る回数が従う確率分布。



① 密度関数　　$p(y|n,\pi) = {}_nC_y\pi^y(1-\pi)^{n-y}$

② 期待値　　$E[y] = n\pi$

③ 分散　　$Var[y] = n\pi(1-\pi)$

二項分布に従う確率変数の具体例

（例 1）□ NP を 10 回観測したとき、そのうち何回が、空欄に不定冠詞を用いていたか。

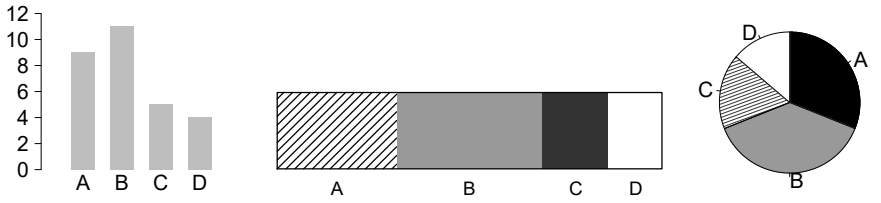
（例 2）floor を 20 人のアメリカ人に発音してもらい、そのうち何人が最後の[r]を発音するか。

（例 3）「私は友達に贈り物をあげた」を 15 人のネイティブに翻訳してもらった時、そのうち何人が SVOO で表現したか。

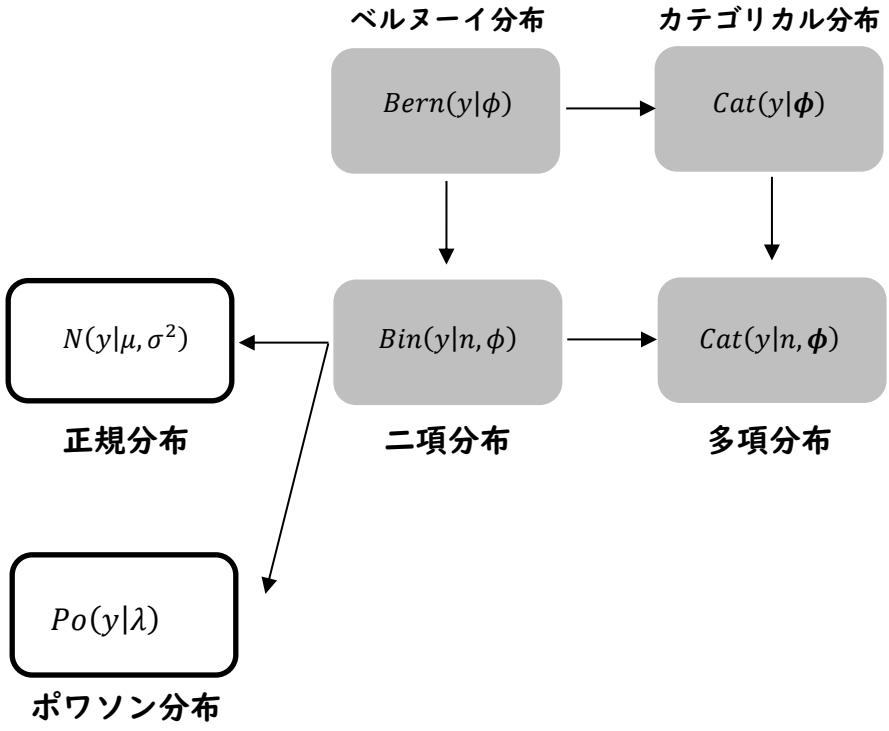
（例 4）「走りません」「走らないです」を合計を 500 例集めた。そのうち何文が「ます」を使った表現か。

視覚化

① 棒グラフ　② 帯グラフ　③ 円グラフ

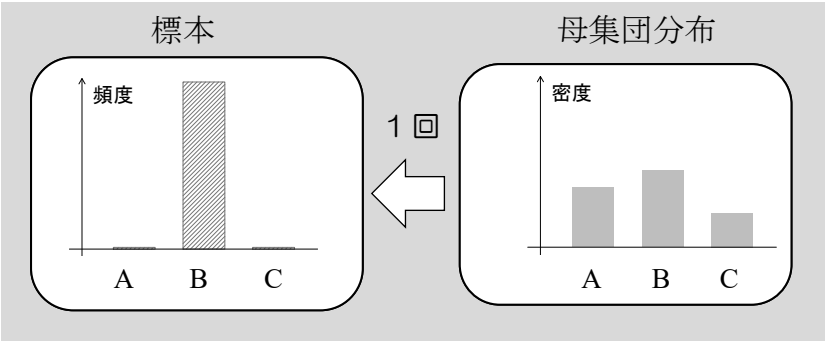


確率分布の関係（１）



（3）カテゴリカル分布 Categorical Distribution

これは、値がvになる確率がπ_vであるV個の離散値から一つ取り出す試行を1回行った時それぞれの値が出る回数が従う分布。



① 密度関数　$p(\boldsymbol{y}|\boldsymbol{\pi}) = \pi_1^{y_1}\pi_2^{y_2}\cdots\pi_V^{y_V} = \prod_{v=1}^{V}\pi_v^{y_v}$

② 期待値　$E[y_v] = \pi_v$

③ 分散　$Var[y_v] = \pi_v(1-\pi_v)$

ベクトル表記

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_V \end{pmatrix}$$

カテゴリカル分布に従う確率変数の具体例

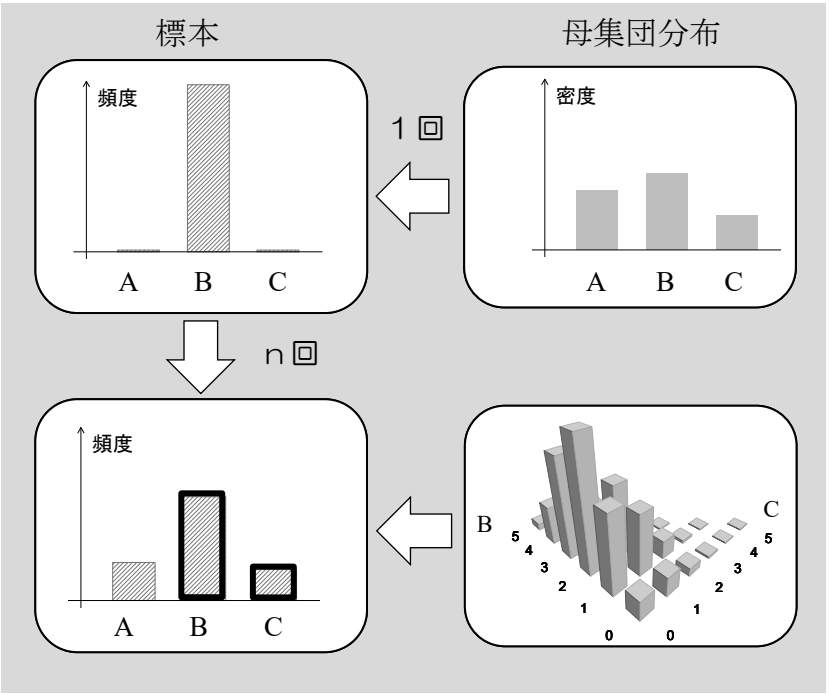
（例1）drive 代名詞 crazy 構文をコーパスで1例採取したとき、その代名詞が me であるか。

（例2）尊敬語を含む韓国語の例文を一つ翻訳するとき、「お…になる」、「…なさる」、「お…なさる」という三つのタイプの中で「…なさる」構文が使われるかどうか。

（例3）日本語の学習者が rapidity という単語を訳すとき、「速さ」「速度」「スピード」の中で「速さ」が選択されるか。

（4）多項分布 Multinomial Distribution

これは、値がvになる確率がπ_vであるV個の離散値から一つ取り出す試行をn回行った時それぞれの値が出る回数が従う分布。



① 密度関数　$p(y|n,\boldsymbol{\pi})$

$$= \frac{n!}{y_1!\,y_2!\cdots y_V!}\pi_1^{y_1}\pi_2^{y}\cdots\pi_V^{y_V}$$

$$= \frac{n!}{y_1!\,y_2!\cdots y_V!}\prod_{v=1}^{V}\pi_v^{y_v}$$

② 期待値　$E[y_v] = n\pi_v$

③ 分散　$Var[y_v] = n\pi_v(1-\pi_v)$

多項分布に従う確率変数の具体例

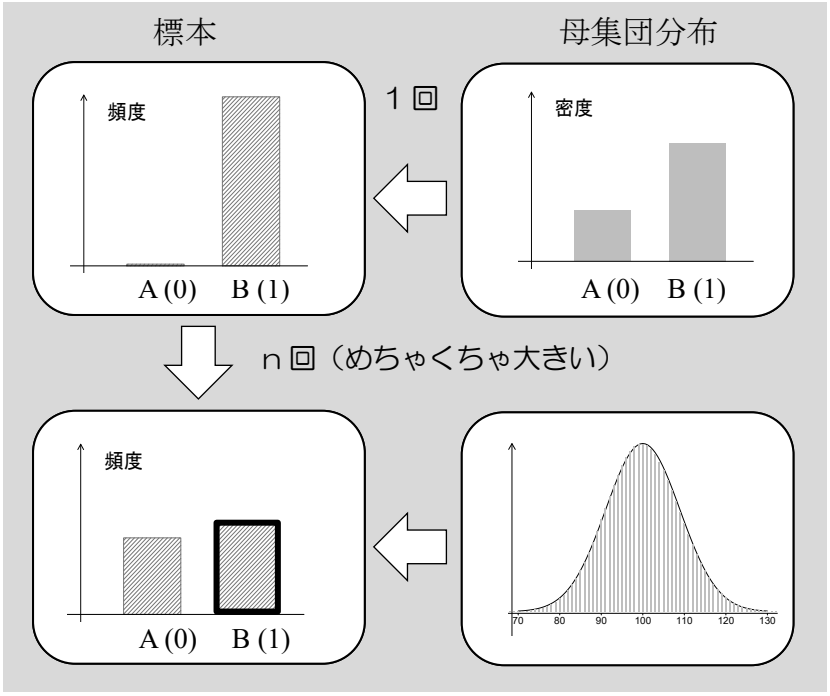
（例１）drive 代名詞 crazy 構文をコーパスで 100 例採取したとき、そのうち何例が me という代名詞を取っているか。

（例２）「お…になる」、「…なさる」、「お…なさる」という三つのタイプの尊敬語を使った構文を 532 例集めたとき、「…なさる」構文を使うのは何例か。

（5）正規分布 Normal Distribution

これは、「稀ではない現象」を「大量に観測した」際に二項分布の極限として登場する確率分布。

（＝ランダムな誤差が積みあがると出現する分布）



① 密度関数　$p(y|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$

② 期待値　$E[y]=\mu$

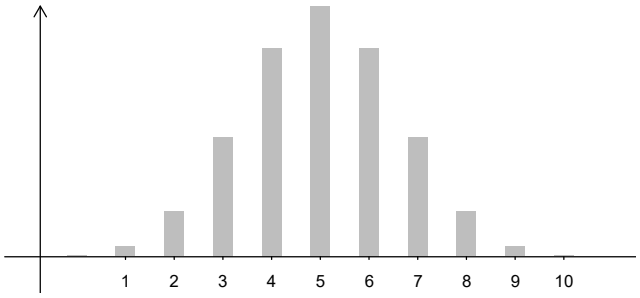
③ 分散　$Var[y]=\sigma^2$

二項分布が正規分布に近づいていくということ

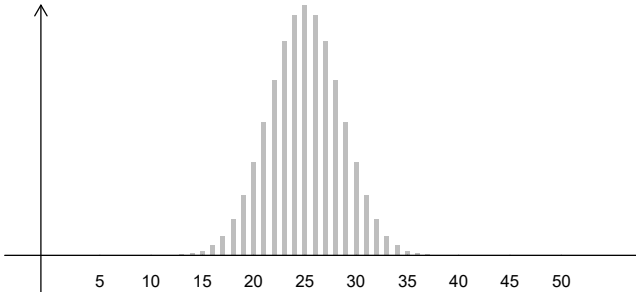
二項分布$B(y|n,\phi)$に従う確率変数Xはϕがそこそこ大きく、かつ、n が大きいとき、近似的に$N\left(n\phi, n\phi(1-\phi)\right)$に従う。

$\phi = 0.5$のとき

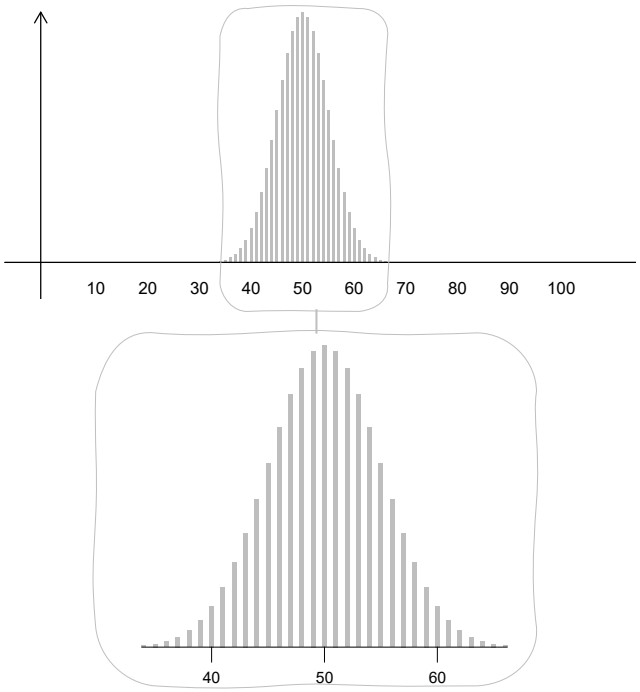
$n = 10$



$n = 50$

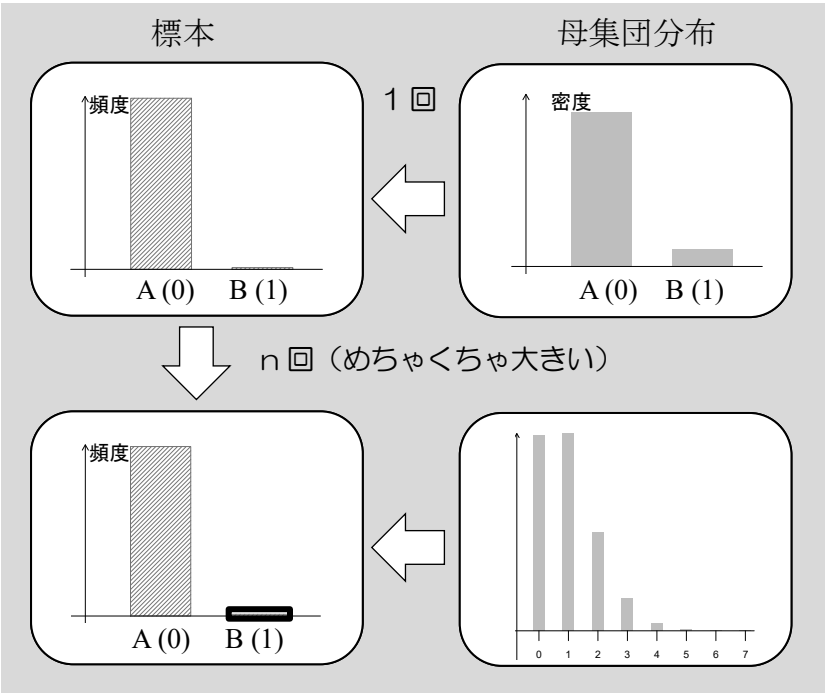


$n = 100$



（6）ポワソン分布 Poisson Distribution

これは、「稀な現象」を「大量に観測した」際に、得られる発生回数が従う二項分布の極限として登場する確率分布。



① 密度関数　　$Po(y|\lambda) = \frac{\lambda^y}{y!}\exp[-\lambda]$

② 期待値　　$E[y] = \lambda$

③ 分散　　$Var[y] = \lambda$

期待値がλなので「単位時間あたりに平均λ回起こる現象が、単位時間にy 回起きる確率現象」のモデルとして使われる。

ポワソン分布に従う確率変数の具体例

（例１）○×新聞記事の一面で「言語」という単語は何回登場するか。

（例２）ある釣り場で１時間釣りをしたときに、何匹の魚を釣ることができるか。

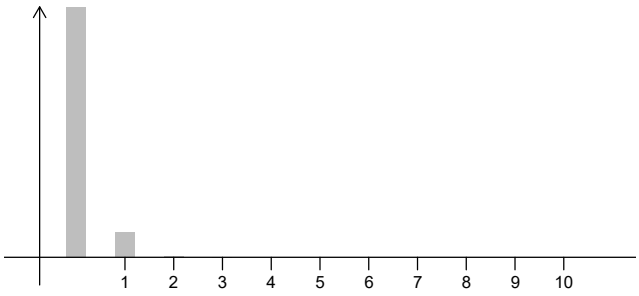
（例３）一年間で何回兵士が馬に蹴られて怪我をしてしまうか。

二項分布がポワソン分布に近づいていくということ

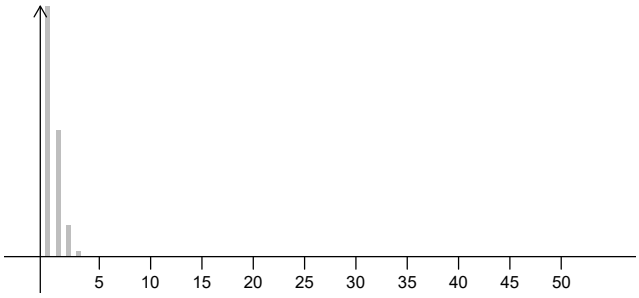
二項分布$B(y|n,\phi)$に従う確率変数 X はϕがとても小さく、かつ、n が大きいとき、近似的に$Po(n\phi)$に従う。

$\phi = 0.01$のとき

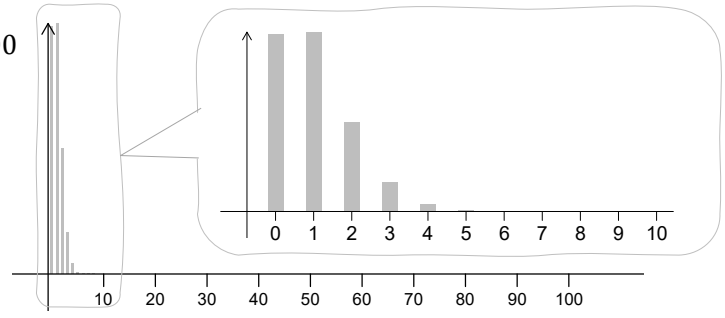
$n = 10$



$n = 50$

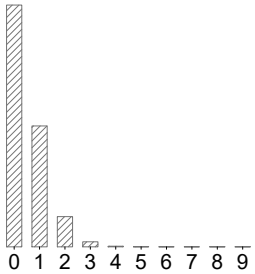


$n = 100$

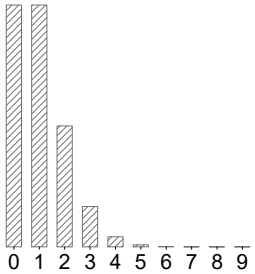


ポワソン分布の例

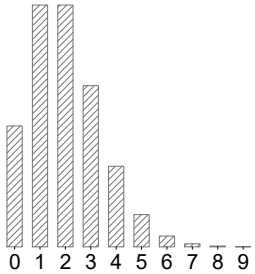
$\lambda = 0.5$



$\lambda = 1$

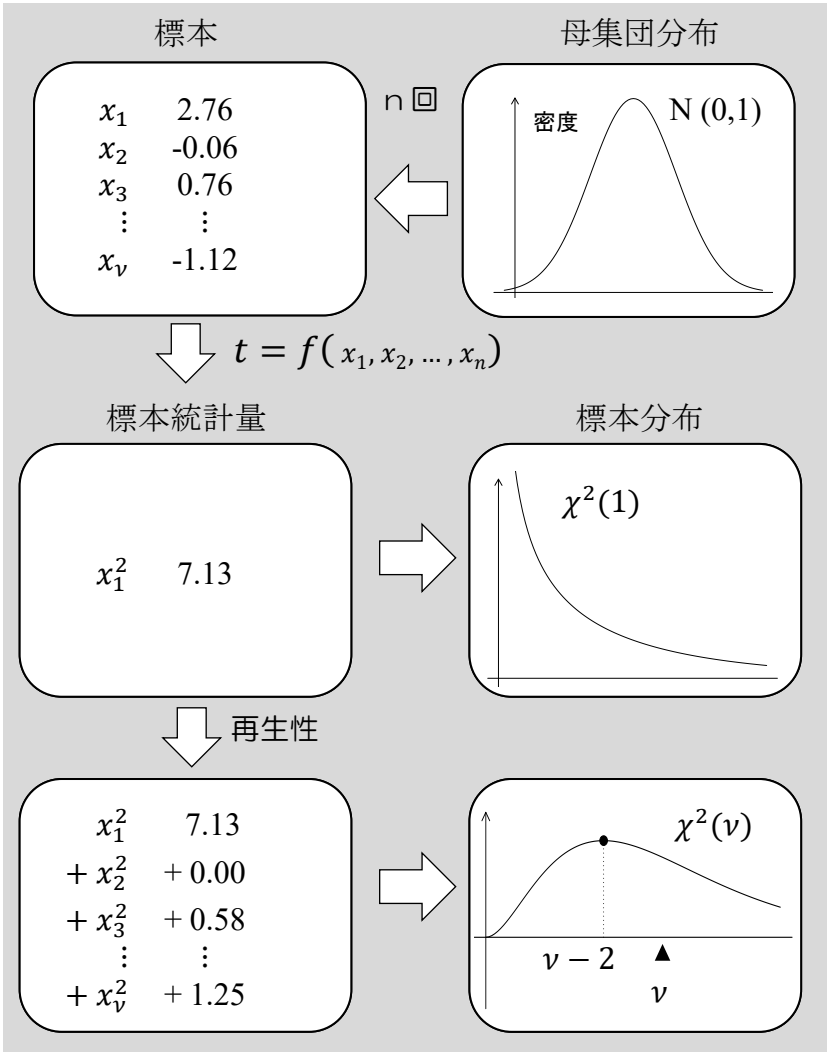


$\lambda = 2$



（7）χ^2分布 Chi-square Distribution

これは、N(0,1)に従うν個の確率変数の二乗和が従う分布。



① 密度関数　$\chi^2(y|\nu) = \dfrac{2^{-\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} y^{\frac{\nu}{2}-1} \exp\left[-\dfrac{y}{2}\right]$

② 期待値　$E[y] = \nu$

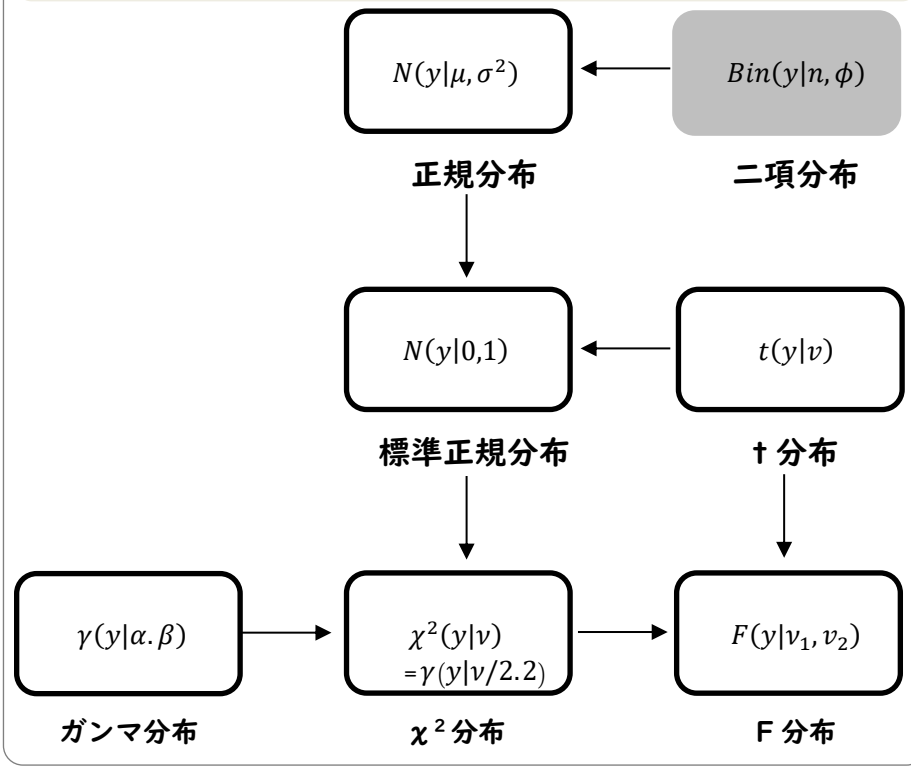
③ 分散　$Var[y] = 2\nu$

標準正規分布

これは、平均が０，分散が１の正規分布。

$$z_i \sim N(0,1)$$

確率分布の関係（２）



分布の再生性　Reproductive Property

同一分布に従う複数の独立な確率変数の和が元の分布に従うとき、その分布には再生性があると言います。

① 正規分布

$$\begin{gathered} X_1 \sim N(\mu_1, \sigma_1^2) \\ X_2 \sim N(\mu_2, \sigma_2^2) \end{gathered}$$

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

② 二項分布

$$\begin{gathered} X_1 \sim Binom(n_1, \phi) \\ X_2 \sim Binom(n_2, \phi) \end{gathered}$$

$$X_1 + X_2 \sim Binom(n_1 + n_2, \phi)$$

③ χ^2分布

$$\begin{gathered} X_1 \sim \chi^2(n_1) \\ X_2 \sim \chi^2(n_2) \end{gathered}$$

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$