統計学より

2012.12.27
スローサークル
2013.1.5
高校サークル・サロン

飯島 光治

はじめに　この所 統計学を読んできました。(定年5〜6年目)。学生時代 統計が全然わからず 教員になり、多少勉強しましたが。それゆえ 本だけは購入していました。それらの本の中から。

なお統計学には、数理統計学と社会統計学があり、社会統計学からは、数理統計学はデータを所与のものとして扱い、吟味(目的、方法等)していないのではの批判を感じました。

1　有意抽出、有意である、有意でない、有意水準 のことばの意味について

はじめに　統計でよく有意ということばがでてきます。本を読んでいて、有意を偶然でないとおきかえるとよいのではと思いました。

○ 有意は significant の訳で 辞書では
1 重要な　2 意味ありげな　3 a 意味のある、意義深い
b (…を)意味して、表わして　4 かなりの、著しい
(例) significant change 著しい変化
とあります。

(1) 有意抽出 とは

これについて書いてある本は少なく、あまり使われない方法からだと思います。次の説明が。

「標本のとり方には、大別して2つのしかたがある。その1つは標本が全体を代表させうるように意識的に判断を加えて一部分を有意選出するのであるが そのときの標本を有意標本という。*　無作為抽出によって得られる標本を無作為標本または任意標本という。

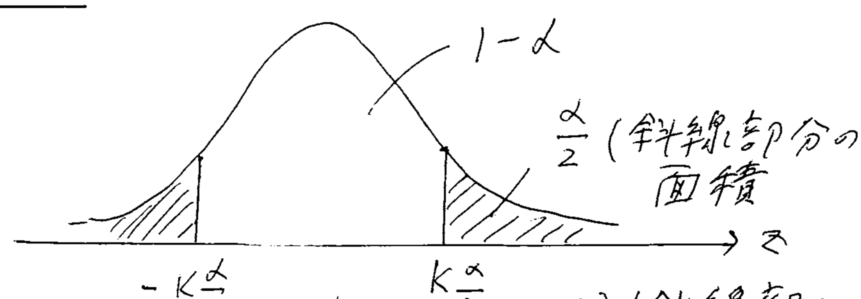
*NHKの「のど自慢 熱唱熱演名場面」(2012.12.16)は有意抽出です。

* 層別は (例 20代、30代、…の年令別)

無作為抽出による標本抽出は、確率という
客観的尺度で規定されるから無作為標本を
もとにして、母集団を推測するときにこそ、有意標本
とはちがって、客観的な妥当性がその推測に付与される。
この点が統計的推測で本質的に大切な点である。
無作為標本は確率の法則の適用できる標本
という意味で、確率標本ともいう。
統計的推測の諸方法は無作為標本についてのみ
展開されるのである。今後無作為標本を単に
標本とよぶ。」（「統計学通論」(第2版)（北川敏男、稲葉三男共著 共立出版 1960刊 1979版 P95）
（傍線 飯島）

(2) 有意とは

1) 「



有意水準αと棄却域（斜線部）

「正規分布によれば、標準正規分布の場合、棄却域との境界値は $\pm K_{\alpha/2}$ で与えられ、$\alpha=0.05$ のとき $K_{0.025}=1.96$ となっている。したがって確率変数を Z として $-1.96 \leqq Z \leqq 1.96$ の範囲にある Z であれば、有意となる。」

（「キーポイント確率・統計」（和達三樹 十河清 岩波 1993刊 P105）
傍線 飯島）

* α を有意水準という。目安として、ふつう $\alpha=0.05$ か 0.01。

<コメント> この場合有意は意味がある。
信頼区間（$1-\alpha$ を信頼度という）に入っている
となります。すると有意でないは、その反対
となり、Z が棄却域に入れば仮説は
採用されないことに。

2） なお 次の説明が。

「仮説検定の基本的な考え方は、仮説H_0が正しいとすると標本観察の結果との間に大きなくい違いがある時は、H_0を否定しようという考え方であるが、H_0が事実正しくとも偶然的な変動（標本変動）で大きなくい違いが生じうる。

しかし その確率が 5%以下とか1%以下というように小さい時には、それは偶然的な原因によるものではなくH_0が誤りであるという意味のある原因によるものと考えようというわけである。これが有意という言葉の意味である。」（「基本統計学」（宮川公男 有斐閣 1977刊 1982版 P180）

ここで 5%や1%を 有意水準という。

従って 有意でないは、有意の否定で、H_0を仮説したことは、その確率が、許容範囲に入っていて、H_0を否定できないことになります。"信頼範囲

なお上記に、標本変動ということばがでてきましたが

「変動＝①考えられたいずれかの原因に帰因させることのできるもの＋②考えられたどの原因にも帰因させることのできないもの＝①原因による変動＋

②残差変動。 統計学では ①を有意な（意味のある意味づけできる）変動 ②を有意でない（意味づけのできない）変動という。」とあります。

（「統計学でリスクと向き合う」宮川公男 東洋経済新報社 2003刊 P93）

<コメント> 同じ著者によるものですが、有意の説明が「基本統計学」と「リスクと向き合う」では違っています。私は「キーポイント確率統計」に従い、「リスクと向き合う」の説明に同意します。

もし、確率が棄却域に入った場合、有意であるとあったらそれは「基本統計学」にある意味と解釈します。

$\boxed{2}$ 二項分布の係数の計算を (1) ポアソン分布で
(2) 正規分布で 計算する。— 計算の簡便化 —

(1) ポアソン分布で近似する。

二項分布 $B(n, p)$ の平均は np になりますが
$np = \lambda$ (一定) で p が小さい時、
ポアソン分布 $\dfrac{\lambda^x}{x!} e^{-\lambda}$ ($x = 0, 1, 2, \cdots$ e は自然対数)
で近似できる。

(例) 毎回の作業において、事故のおこる確率が 0.001 であるとき 2000回の作業中に、4回事故のおこる確率は?

→ 2項分布を用いると ${}_{2000}C_4 (0.001)^4 (0.999)^{1996}$ を計算しなければならない(!)。これをポアソン分布で近似すると $\dfrac{2^4}{4!} e^{-2}$ でよい。(2は $np = 2000 \times 0.001$ より。)

$\left(e^{-2} = \dfrac{1}{e^2} \fallingdotseq \dfrac{1}{2.7182^2}\right.$、電卓で $\fallingdotseq \dfrac{1}{7.39} \fallingdotseq 0.135$

$\left. \therefore \dfrac{\cancel{2}\cdot\cancel{2}\cdot\cancel{2}\cdot 2}{\cancel{4}\cdot 3\cdot\cancel{2}\cdot 1} \times 0.135 = \dfrac{0.27}{3} = \underline{0.09} \right)$

(2) 正規分布で近似する。

二項分布は、n を大きくすると正規分布で近似できる。(ド・モアブル・ラプラスの定理。中心極限定理*の1つ。)
*分布の極限が正規分布になることを主張する定理群につけられた名称。

- n を大きくするの目安は、$p \leqq \frac{1}{2}$ のときは $np > 5$、$p > \frac{1}{2}$ のときは $nq > 5$ $(p + q = 1)$ 程度の n であれば、正規分布をあてはめてよい。
- 二項分布 $B(n, p)$ の平均値 np、分散 npq と同じ正規分布 $N(np, npq)$ に近づく。

(例1) サイコロを 60回振って 1の目が 9回から11回出る確率は？*

* このとき $p=\frac{1}{6}<\frac{1}{2}$ で $np=60\cdot\frac{1}{6}=10>5$

→ 二項分布では ${}_{60}C_{9}\left(\frac{1}{6}\right)^{9}\left(\frac{5}{6}\right)^{51}$ などの計算。

しかし 正規分布 $N(10,\ 2.9^2)$* をあてはめて

$P(8.5<X<11.5)=P(-0.52<Z<0.52)$

$=0.1985\times 2=\underline{0.3970}$ とできる。

(注1) 上記で、*$\underline{2.9^2}$ は $npq=60\times\frac{1}{6}\times\frac{5}{6}=\frac{50}{6}=\frac{25}{3}=8.333\cdots$

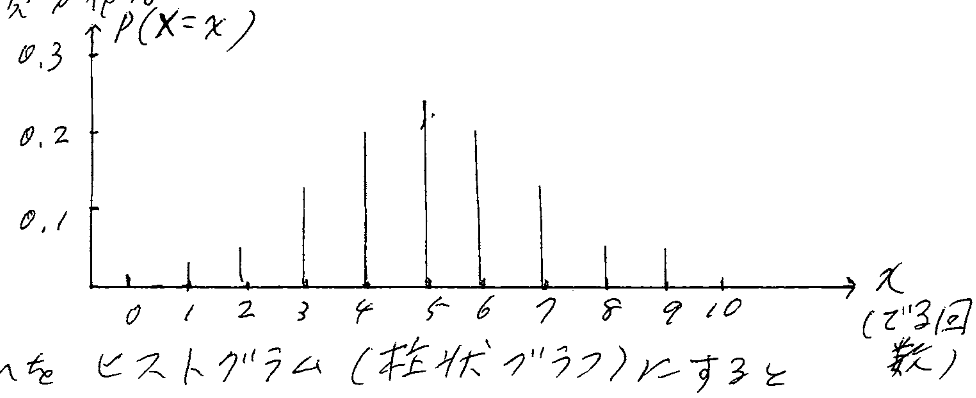
$\sigma=\sqrt{8.333}\fallingdotseq 2.886\cdots\fallingdotseq 2.9$ として

0.52は 標準化で $\frac{X-\overset{np}{\mu}}{\sigma}\nearrow\frac{11.5-10}{2.9}$ を計算して。

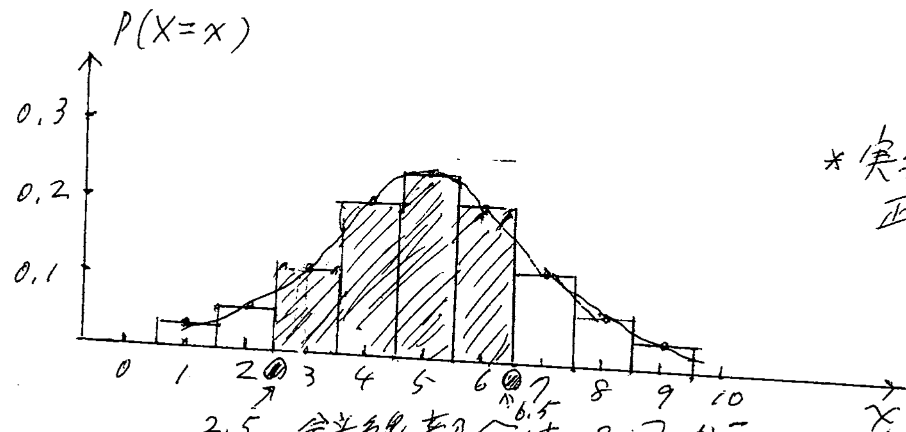
(注2) 上記で $8.5<X<11.5$ としたのは、

たとえば 硬貨1枚を 10回投げて表がでる場合

まず 二項分布は



これを ヒストグラム（柱状グラフ）にすると



*実線は正規分布のグラフ

斜線部分は 3回から6回表がでる確率（＝各面積の和）

この時 横幅＝1より 1つの長方形の面積は高さが表す。

3回～6回は 2.5～6.5 となっている。

8.5～11.5 も同じ理由によります。

（以上 例と解説は「基礎統計学」（青柳雅計 開成出版 1982刊 1992版 P49～P50 による））

3 世論調査はなぜ3000人程度でよいか。

これをきちんと理解するのは、私には簡単ではなく理解する過程が、統計学の内容に。

(準備) ある事象がおきる確率がp、おきない確率が$(1-p)$になっているとする。おきる場合を成功、おきない場合を失敗とよび、それぞれ数1、0を対応させる。

ここで重要な仮定は「毎回の確率事象が同一でかつ独立である。」(たとえばコイン投げの場合何回やっても表・裏のでる確率は同一でかつ独立) このような仮定でのくり返しの試行をベルヌーイ試行という。

ここで n回のベルヌーイ試行の結果、そのうちx回が成功であったとすると、xは0からnまでの整数値をとる。xの従う確率分布が2項分布です。

確率関数$p(x)$は $p(x) = {}_nC_x p^x (1-p)^{n-x}$ となる。

ここで支持率調査の場合

まず母集団の性質は支持率p*によって特徴づけられる。(pを母集団比率という。)

*任意に有権者を1人選んだ時その人が支持すると答える確率。

支持する確率p、支持しない確率$(1-p)$からなるので、2項母集団という。

ここで2項母集団から選び出されたn人の標本$\{x_1, x_2, \cdots, x_n\}$を考える。ここでそれぞれの$x_i$は、支持するとき1、支持しないとき0という値をとることにする。

このときのx_iの総和$\sum_{i=1}^{n} x_i$ ($=n\bar{x}$ 即ち標本平均のn倍)は0からnまでの値を確率的にとる。*

*0はn人全員が支持しない、nはn人全員が支持の時。

この総和の従う分布は、上にのべたように2項分布$B(n, p)$に他ならない。(標本平均$\bar{x}$は この標本の示す標本支持率)

＜まとめ＞

「母集団比率 p をもつ２項母集団の場合、
n 個からなる標本では、その総和 すなわち
標本平均の n 倍は、２項分布 $B(n, p)$ に従って
確率分布する。」

このとき 平均は np、分散は $np(1-p)$。

さらに 「総和 $\sum_{i=1}^{n} x_i$ は、近似的に 正規分布
$N(np, np(1-p))$ に従う。」

さらに y を標本値として $\dfrac{y-np}{\sqrt{np(1-p)}}$ で標準化して
標準正規分布を用いる。（この時 $y=\sum_{i=1}^{n} x_i$）

さらに（確率）$=Prob\left\{\left|\dfrac{\sum x_i - np}{\sqrt{np(1-p)}}\right| < K_{\frac{\alpha}{2}}\right\} = 1-\alpha$ ① がいえ
（このレポート P2の図）

*$|\hat{p}-p| = |p-\hat{p}|$

これは $\sum_{i=1}^{n} x_i / n = \hat{p}$ とおくと

①は $Prob\left\{|p-\hat{p}|^{*} < K_{\frac{\alpha}{2}}\sqrt{\dfrac{p(1-p)}{n}}\right\} = 1-\alpha$ ①′ がいえる。

次の例題は、この式を用いると

例「内閣支持率 p を精度±2％以内で
推定するには、標本サイズ n は何人以上
必要か。信頼度95％で考えよ。」

解）推定精度を 2％以内におさえるには
上記の①′より $K_{\frac{\alpha}{2}}\sqrt{\dfrac{p(1-p)}{n}}$ が2％以内におさまれ
ばよい。信頼度95％のとき $K_{\frac{\alpha}{2}}=1.96$ より

$$1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqq 0.02$$

（ここで①′の p を $\hat{p}$ で近似した。）

$$\rightarrow n \geqq \left(\frac{1.96}{0.02}\right)^2 \times \hat{p}(1-\hat{p})$$

ここで $\hat{p}(1-\hat{p})$ は $\hat{p}=\frac{1}{2}$ のとき最大*

$\hat{p} \rightarrow x$ として
$x(1-x)$
$= x - x^2$
$= -(x^2 - x)$
$= -(x-\frac{1}{2})^2 + \frac{1}{4}$
より。

ゆえに 少なくとも

$$n \geqq \left(\frac{1.96}{0.02}\right)^2 \times \frac{1}{2} \times \frac{1}{2} = 2401（人）$$

以上が世論調査で3000人程度でよいの根拠です。

（以上 レポートP6の準備から例題までの説明は、「キーポイント確率・統計」によります。）

＜補足＞ 以上の内容と関連ありませんが、「キーポイント確率・統計」を読んでいて

1) 正規分布の平均を求めるのに

「平均 $=\int_{-\infty}^{\infty} x\cdot f(x)dx$（←定義）

$=\int_{-\infty}^{\infty}\{(x-\mu)+\mu\}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$

ここで μ は x の中央値とする*）

* ふつう μ は平均値ですが、平均を求めるのに平均値 μ を用いるのはで、中央値としました。（注 ふつう こういう説明はありませんが）

を得る。ここで{ }内の第1項には奇関数の** 積分値0であることを用い、…」とあります。（P64）

** $y=f(x)$ で $f(-x)=-f(x)$ の時、$f(x)$ は奇関数。

第1項は $(x-\mu)\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ ① でこれを $f(x)$ とおくと このまゝでは奇関数になりません。

ここで $\frac{x-\mu}{\sigma}=y$ とおくと $x-\mu=\sigma y\rightarrow dx=\sigma dy$ より

$①=\sigma y\cdot\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2}}\cdot\sigma dy=\underbrace{\sigma y\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}}_{②}dy$

* dy を除く。

②は奇関数になります。すると①の正体は②ということでしょうか。 平均は μ に。

2)「正規分布 $N(\mu,\sigma^2)$ の特性関数は

$f(\xi)=\int_{-\infty}^{\infty}e^{i\xi x}f(x)dx=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}e^{i\xi x-\frac{(x-\mu)^2}{2\sigma^2}}dx$

から計算される。上式の指数関数の肩は

$i\xi x-\frac{1}{2\sigma^2}(x-\mu)^2=-\frac{1}{2\sigma^2}(x-\mu-i\xi\sigma^2)^2+i\mu\xi-\frac{1}{2}\sigma^2\xi^2$ Ⓐ

と変形できるので、」とあります。（P65）

Ⓐの右辺→左辺はいえますが、左辺→右辺はどうやってでしょうか。？