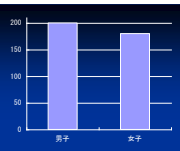
基本編：推測統計学
pp27～46

- 母集団と標本
- 無作為抽出
- 確率変数
- 期待値と分散
- 2項分布
- 正規分布
- 標準化
- 95%信頼区間
- 2項確率の信頼区間

母集団と標本

- 母集団
  - ある調査において，調査対象全体からなる集団
- 無作為抽出
  - 母集団の中から要素を選び出す
  - すべての要素が同じ確率で選ばれるようにする

例　ある高校に3年生男子200名，女子180名が在籍している。
　この男女380名を母集団とし，その中から1名を無作為に抽出したとき，その生徒が女子である確率は？



高校3年生
女子180名
男子200名

無作為抽出した生徒が女子の確率？
→ 全員が同じ確率で抽出

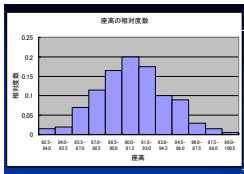
従って，確率は　180/380=9/19

変数Xを定義：
　男子→X=0，女子→X=1

X=1となる確率　P(X=1)= 9/19
X=0となる確率　P(X=0)=10/19
⇒Xは確率変数

Y=1−X と定義すると，Yも確率変数
P(Y=0) = P(X=1)=9/19

問1 上の記号でP(X=0)は何を意味するか。
また，この確率を求めなさい。



区間	度数	相対度数	累積相対度数
82.5-84.0	3	0.015	0.015
84.0-85.5	4	0.02	0.035
85.5-87.0	14	0.07	0.105
87.0-88.5	23	0.115	0.22
88.5-90.0	33	0.165	0.385
90.0-91.5	40	0.2	0.585
91.0-93.0	35	0.175	0.76
93.0-94.5	20	0.1	0.86
94.5-96.0	18	0.09	0.95
96.0-97.5	6	0.03	0.98
97.5-99.0	3	0.015	0.995
99.0-100.5	1	0.005	1
計	200	1	

無作為抽出した生徒の座高をX
座高<90の相対度数は0.385
⇒ P(X<90)=0.385

⇒ 左スソ5本の柱の面積

母集団と標本

- 調査対象全員を母集団
- 抽出された要素を標本
- 無作為抽出したとき無作為標本

無作為標本でない例
座高の高い生徒を選ぶと P(X<90)≒0
座高の低い生徒選ぶと P(X<90)≒1
相対度数分布から求めた P(X<90)=0.385**とならない**

標本が特定の区間に含まれる確率
無作為　⇒　母集団から推測可
無作為でない　⇒　推測困難

標本に含まれる要素の個数：標本サイズ
標本は母集団の部分集合
サイズは部分集合の要素数

復元と非復元
母集団から1人抽出し，もう1人抽出
- 復元抽出：最初に抽出した人を戻す
  - 同じ人を抽出する可能性有
- 非復元抽出：最初に抽出した人を除く
  - 抽出対象者が減っていく

サイコロを振って出た目を順に$X_1, X_2, \cdots, X_n$とする。
これは　S={1,2,…,6}　からの
復元抽出によるサイズnの無作為標本

クラスから1人無作為に抽出

1. 全員の名札を入れた箱から抜く
2. 大小2つの乱数サイコロを用意
   - 大コロは10の位，小コロは１の位
   - 生徒に番号をふり，サイコロをふる
   - 出た番号の生徒を抽出
   - 該当者がいないときは再度ふる
3. 乱数表を使う
   - 最初の位置を乱数サイで決定
   - 56が出たら5行6列が始点
   - 2桁の乱数なら2つずつとる
   - Excelには乱数生成機能あり

今後，標本といったら無作為標本

期待値

賞金	150,000	10,000	500	0	計
本数	2	10	200	788	1000

このくじを毎日買い続けて損をしない金額は？
長い眼で見れば全部買うのと同じ
　全部買ったときの１本当りの賞金？
　１本当り賞金−賞金総額／総本数

$$\frac{150000\times2+10000\times10+500\times200+0\times788}{1000}=500$$

これが期待値

書き直すと

$$150000\times\frac{2}{1000}+10000\times\frac{10}{1000}+500\times\frac{200}{1000}+0\times\frac{788}{1000}$$

期待値とは(賞金×確率)の和

平均値（期待値）の定義

確率分布表

X　$x_1, x_2, \ldots\ldots, x_m$

P　$p_1, p_2, \ldots\ldots, p_m$

平均値　$E(X)= x_1p_1+ x_2p_2+\ldots\ldots+ x_mp_m$

例：乱数サイをふって出る目をX

X　0,　1,　…,　9

P　0.1, 0.1,　…,　0.1

$E(X)=0\cdot0.1+1\cdot0.1+\cdots+9\cdot0.1$
$=4.5$

Xの10倍を賞金Yとすると

$E(Y)=10\cdot0\cdot0.1+\cdots$
$+10\cdot9\cdot0.1$
$=45=10\,E(X)$

分散・標準偏差

- $V=(X-平均値)^2$とすると
  $E(V)=(0-4.5)^2\cdot0.1+(1-4.5)^2\cdot0.1+$
  $\cdots+(9-4.5)^2\cdot0.1$
  $=8.25$　これが分散：$V(X)$と書く
- $SD=\sqrt{V(X)}=\sqrt{8.25}$　これが標準偏差

2項確率・・・確率のシーラカンス

P(表)=p, P(裏)=1−p　のコイン:

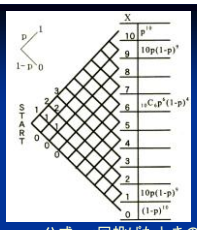
1回目	2回目	結果	表の数	確率
表	表	表 表	2	pp
	裏	表 裏	1	p(1−p)
裏	表	裏 表	1	(1−p)p
	裏	裏 裏	0	(1−p)(1−p)

P(表)=p, P(裏)=1-p　のコイン

P=0.5のとき	
P(表が2)=p^2	0.25
P(表が1)=2p(1-p)	0.5
P(表が0)=$(1-p)^2$	0.25

試行を増やしてみると

1回	p	1-p			
表の数	1	0			
2回	p^2	2p(1-p)	$(1-P)^2$		
表の数	2	1	0		
3回	p^3	$3p^2(1-p)$	$3p(1-p)^2$	$(1-P)^3$	
表の数	3	2	1	0	
4回	p^4	$4p^3(1-p)$	$6p^2(1-p)^2$	$4p(1-p)^3$	$(1-p)^4$
表の数	4	3	2	1	0



表が出る確率がpのコインを10回投げる

二項確率 B(10,p) と呼ばれる

公式：n回投げたときの表の回数をT
$P(T=k)={}_nC_k\,p^k(1-p)^{n-k}$,　k=0, 1, 2,...,n
T～B(n,p)

二項分布

- 公式：
  n回投げたときの表の回数をT
  $P(T=k)={}_nC_k\,p^k(1-p)^{n-k}$,　k=0, 1, 2,...,n
- T～B(n,p)
- Excel関数：
  BINOMDIST(*k,n,p,FALSE*)

組み合わせの数

- 相異なるn個のものからr個を取って得られる集合の数
- ${}_nC_r$ または $\binom{n}{r}$ とかく
- 公式
  $${}_nC_r=\frac{n(n-1)(n-2)\cdots(n-r+1)}{r!}$$
  $$=\frac{{}_nP_r}{r!}=\frac{n!}{(n-r)!r!}$$
- Excel関数：COMBIN(*n,r*)

二項分布の期待値と分散

P(表)=p, P(裏)=1-pのコイン

定義　X：表ならX=1，裏ならX=0

$E(X)=1p+0(1-p)=p$

$V(X)=p(1-p)^2+(1-p)(0-p)^2$
$=p(1-p)$

公式

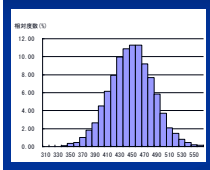
期待値：　$E(X)= x_1p_1+ x_2p_2+\cdots+ x_mp_m$

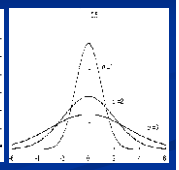
分　散：　$V(X)=p_1(x_1-E(X))^2+\cdots+p_n(x_n-E(X))^2$

コインをn回投げる

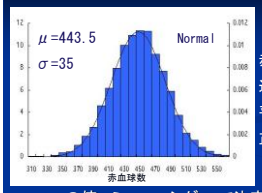
- 定義　X_k : k回目が
  表なら$X_k=1$　裏なら$X_k=0$
  $E(X_k)=p$, $V(X_k)=p(1-p)$
- 定義　$T=X_1+\cdots+X_n$
  表の数～B(n,p)
  $E(T)=E(X_1+\cdots+X_n)$
  $=E(X_1)+\cdots+E(X_n)$
  $=np$
  $V(T)=V(X_1+\cdots+X_n)=np(1-p)$

ヒストグラムを連続関数で近似する





正規分布



赤血球数は
近似的に
平均443.5, SD35の
正規分布に従う

- 2つの値(μミュー, σシグマ)で決定
- 平均μ, 標準偏差(SD)σ, 分散σ^2

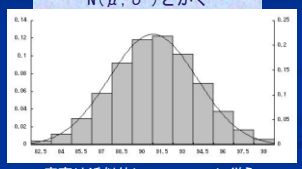
正規確率関数

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

- 様々な実測データによく適合
- $e=2.71828\cdots$自然対数の底
- μ(ミュー), σ(シグマ)の2つのパラメータで決まる
- 正規確率密度関数ともいう
- Excel関数 : **NORMDIST**(***x***,μ,σ,***FALSE***)

データの特徴を2つの数で表現

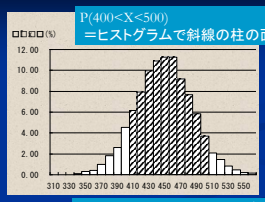
平均μ, 分散σ^2 の正規分布を
N(μ, σ^2) とかく



座高は近似的にN(91, 3.23^2)に従う

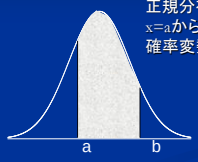
赤血球数の分布からの標本をXとすると

P(400<X<500)
＝ヒストグラムで斜線の柱の面積



≒正規分布N(443.5, 35^2)のグラフで
x=400とx=500の間の面積

抽象化 : 正規確率変数 X の定義



- P(a<X<b)
  正規分布のグラフの
  x=aからx=bまでの面積となる
  確率変数 X を考える。

Xは正規確率変数
と呼ばれる。

クイズ

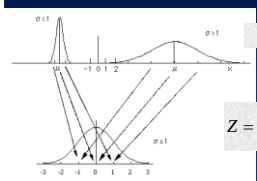
- ケトレーは, フランス兵士10万人の身長のヒストグラムに正規分布を重ねた結果, 身長が低いと偽って徴兵免除された者が2000人いると結論した。
  徴兵免除者を調査せず何故推定できたのか？

μ=0, σ=1の正規分布 N(0,1) は標準正規分布

身長 X～N(170,5^2)

- 全員に10cmの下駄を履かせて測れば,
  分布が10cm平行移動
  X～N(170,5^2) → X+10～N(180,5^2)
  - X～N(μ,σ^2) → X+c～N(μ+c,σ^2)
- 身長をmで測れば, 100分の1
  X～N(170,5^2) → X/100～N(1.7,0.05^2)
  - X～N(μ,σ^2) → cX～N(cμ,$c^2\sigma^2$)
- 標準化とは
  X～N(170,5^2) → X-170～N(0, 5^2) → $\frac{X-170}{5}\sim N(0,1)$

標準化定理



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

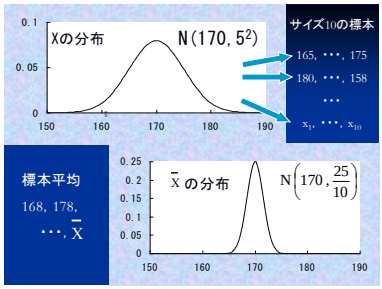
$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

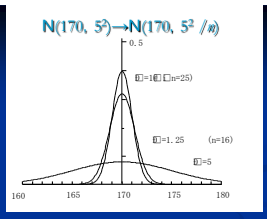
- Xが正規分布$N(\mu,\sigma^2)$に従うとき，
  $Z=(X-\mu)/\sigma$
  は標準正規分布$N(0,\ 1^2)$に従う

標本平均の定理

- 無作為標本 $X_1,\cdots,X_n \sim N(\mu,\sigma^2)$
  標本平均 $\bar{X} = \frac{X_1+\ldots+X_n}{n}$
- 新たに，$Y_1,\cdots,Y_n \sim N(\mu,\sigma^2)$
  標本平均 $\bar{Y}$
- さらに，$Z_1,\cdots,Z_n \sim N(\mu,\sigma^2)$
  標本平均 $\bar{Z}$
- ・・・，標本平均を無数に得たそれら標本平均の分布は

  正規分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$ となる





- $N(170,5^2/16)=N(170,1.25^2)$
- $N(170,5^2/25)=N(170,1)$
- nが大きいと分布がμ に集中する