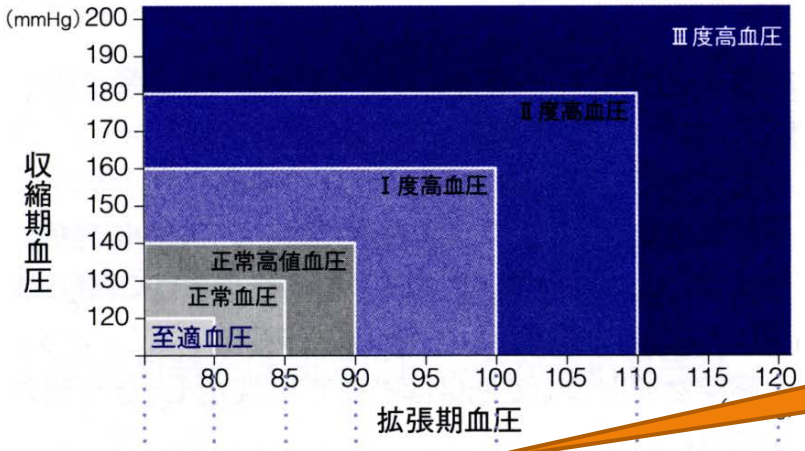
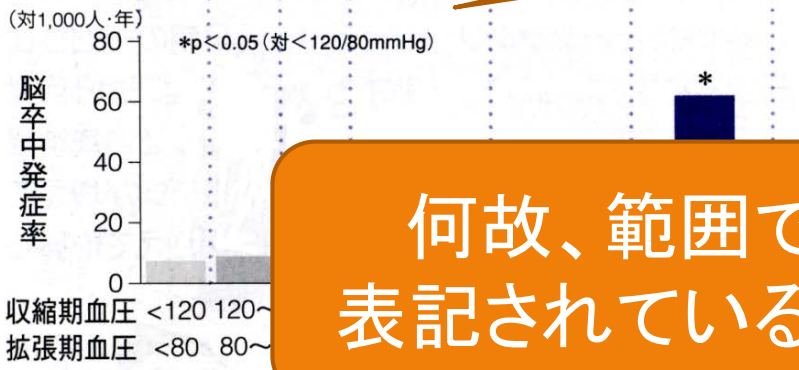
統計学

第3回　確率と確率分布

これから学ぶこと

高血圧の診断と分類





血圧と脳卒中発症率の関係【久山町研究】

【対象】1961年から久山町の循環器健診を受診した住民から設定した第一集団、60歳以上の男女580名。
【方法】追跡32年、性・年齢調整

〔日本高血圧学会. 高血圧治療ガイドライン2009より改変〕

血圧が高くなるにつれて脳卒中発症リスクは高まります。
至適血圧を目指しましょう。

	検査項目		
血中脂質	総コレステロール	140～199	mg/dl
	HDL－C	40～119	mg/dl
	中性脂肪	空 30～149 随 30～149	mg/dl
腎機能	尿素窒素	7～24	mg/dl
	クレアチニン	0.0～0.7	mg/dl
	尿酸	0.0～7.0	mg/dl
血液一般検査		.6	g/dl
		.9	%
	赤血球	360～489	×10⁴/μl
	白血球	32～85	×10²/μl
	MCV	83.1～97.7	fl
	MCH	27.0～33.6	Pg
	MCHC	31.2～36.0	%
	血小板	13.0～34.9	×10⁴/μl
	血清鉄	40～199	μg/dl
	全血比重	1.052以上	
糖尿病	血糖	空 0～99 随 0～99	mg/dl
	HbA1c	0.0～5.1	%
膵臓	血アミラーゼ	21～120	IU/l
血清等検査	CRP	0.0～0.4	mg/dl
	RA	0～19	U/ml
血沈	1h	0～15	mm
	2h	0～15	mm

この意味は？

何故、範囲で
表記されている？

目次

- 確率
- 確率変数と確率分布
- 確率の平均と標準偏差
- パーセント点とp値
- 正規分布と2項分布

教科書

- p62～p86

何故統計で確率が必要か？

- 確率と統計の違い
  - 統計：調査・観察して得られるデータを取り扱う（実測値）
  - 確率：『確率論』という理論から得られる値、又はその計算方法（理論値）
- 『理論値』と『実測値』は、普通あまりかけ離れてはいない
  - 『理論値』と『実測値』がかけ離れている→調査・観察の前提がおかしい？（検定）
  - おかしいかどうかを調べるためには『理論値』が必要

確率

確率とは何か

『確からしさ』を表す指標

- 先験的確率（数学的確率）
  - ある事象Aが「同様に確からしい」と考えられるn個の可能な場合のうちm個の場合に起こるとしたとき、Aの起こる確率はm/nである
- 経験的確率（統計的確率）
  - 多数回（n回）の試行がなされたときに、事象Aの起こる（m回）割合がほぼp（一定値）であり、試行回数nを大きくすると、m/nがpに一層近づくとき、Aの起こる確率はpとなる

確率の例

数Aの教科書参照

- １つのコインを投げる
  - 表が出る確率
    - 数学的確率：「表」と「裏」の２通りで「表」が出た＝確率１／２
    - 統計的確率：コインを何回も投げていくと、１／２の割合で表が出てるよなぁ……＝確率１／２
- １つのさいころを投げる
  - 3の目が出る確率
  - 偶数の目が出る確率
  - 素数の目が出る確率

主観的確率と客観的確率

- 客観的確率
  - 理論的に導かれる確率
  - 繰り返し実験により、「頻度」を用いて導くことができる確率
- 主観的確率
  - 理論的に導けない確率
  - 信念、確信といった「人間の主観」が入り込んでいる確率
- 「確率」は、自分が持っている情報や信念により変化するもの（ベイズ統計学）

確率＝「確信の度合い」

大数の法則

- 試行回数をどんどん増やしていくと、ある事柄Aが怒る相対度数（統計的確率）は、事柄Aが起こる数学的確率に限りなく近づいていく

確率変数と確率分布

確率変数

確率付きの変数

試行の結果によって値が定まる変数

さいころを一回投げるという試行で、さいころの目を X とおく。

（この X が「確率変数」）

X は１、２、３、４、５、６のどれかの値になる。

X が１になる確率は、$P(X=1)$　と表す。

また、X が２から４の値をとる確率は、$P(2 \leq X \leq 4)$　と表す。

確率変数の使用例

さいころを一回投げるという試行で、さいころの目を X とおく。

$$P(X=1)=\frac{1}{6}$$

1の目が出る確率

$$P(2\leq X\leq 4)=P(X=2)+P(X=3)+P(X=4)$$
$$=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{1}{2}$$

2から4の目が出る確率

確率変数の種類

- 離散的確率変数
  - データが飛び飛びの値を取る
    - コイン
    - さいころ
    - 人数
- 連続的確率変数
  - データが、ある範囲のどんな値でもとりうる
    - 身長
    - 各種検査値

確率分布（分布）

離散型

確率変数 X のとりうる値が $x_1, x_2, \ldots, x_n$ であるとき、

$P(X = x_k)$ を p_k と書くことにすると、x_k と p_k の関係は

X	x_1	x_2	$\ldots$	x_n	計
P	p_1	p_2	$\ldots$	p_n	1

この対応関係を、X の**確率分布**または単に**分布**といい、

確率変数 X はこの分布に**従う**という。

分布：ある地点やある範囲における確率を表すもの

（統計学における）分布とは何か

- 分かれてあちこちにあること、また、分けてあちこちに置くこと
- その事象が空間的・時間的なある範囲内に存在すること、また、その存在する状態
- 確率分布のこと

統計学でいう「分布」：ある事象が起こる確率がどのようになっているのかを記したもの

データのバラツキには規則性がありそう

何故「分布」を考えることが必要か？

- データの「バラツキ」に「規則」を入れる為
  - データのバラツキには規則性がありそう
  - 規則を入れることにより、推定・検定が可能になる
    - 統計モデル
- 「規則」の入れ方→確率分布
  - 「バラツキ」・「誤差」に「確率」を入れる
  - ただ入れるのではなく、データの特性に沿った「規則」を入れる必要がある（正規分布、2項分布、Poisson分布等）

ランダムな現象に「規則性」？

- 例えば、さいころの目について考える
- 一回投げれば、さいころは1～6の目がランダムに出る（予測不能）
- でも、さいころを何万回も振っていると、例えば、1の目は6回に一回の割合（1/6の確率）で出ていることが分かる（規則性がある）
- その「規則性」を使えば、一回一回は予測できないけど、『目の出やすさ』は予測できる

統計学（統計科学）とは

- いくつかの対象について、観測、調査、実験などを行って得た結果を、数字として表したものを「データ」という
- （不確実性を含む）データから必要な情報を引き出すことを「統計データ解析」といい、そのための方法を「統計データ解析法」という
- 「統計学」とは、統計的方法の体系化ならびに統計的概念の本質に関する研究を課題にしている学問分野である

「分布」の概念を入れる

確率分布の例１

- 2枚のコインを投げて、表の出る枚数をXとする

X	0	1	2	計
確率	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

（裏、裏）

（表、裏）
（裏、表）

（表、表）

確率分布の例２

- さいころを1回投げる試行

出た目 X	1	2	3	4	5	6
確率	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

さいころ投げにおける理論値

これと、実際さいころを投げた時の結果を比較すると、さいころがイカサマかどうかが分かる

確率分布を関数的に表してみる

さいころを1回投げるという試行

$$f(x) = \mathrm{P}(X = x) = \begin{cases} \frac{1}{6} & (x = 1, 2, 3, 4, 5, 6) \\ 0 & (\text{それ以外}) \end{cases}$$

確率関数

確率変数を使うと、確率が「実数xについての関数」にみえる

分布関数とは何か？

- データの分布：ランダムだけど、確率で考えるとある程度の規則性がある
- その「規則性」に着目して、確率が計算できるよう「関数化」したものが分布関数

離散型と連続型

- 離散型：データは整数値しか取らない（飛び飛びの値を取る）
- 連続型：目盛さえ小さくしていけば、いくらでも数値の細かい（小数点以下の桁数が多い）データが取れる

⇒ 「分布」の形が若干違う

連続型の例

- アナログ型のストップウォッチで盤面を見ずに止めた時の秒数を測定
- 「秒数」は、小数点以下いくらでも測定可能

$$P(X = 1) = 0$$

$$P(0 \leq X < 1) = \frac{1}{60}$$

$$P(0 \leq X < 15) = \frac{1}{4}$$

$$P(0 \leq X < 0.1) = \frac{1}{600}$$

連続型の例（身長データ）

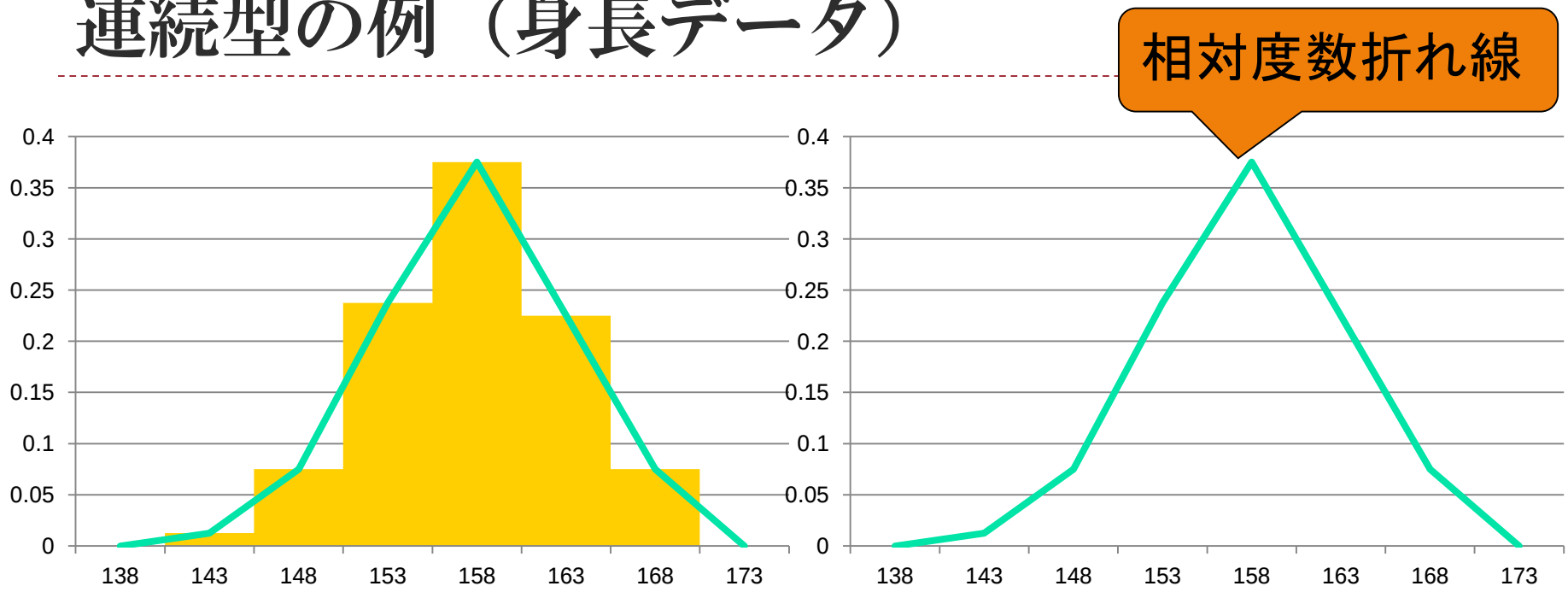
- 身長：連続データ
- 離散データのような表……

確率とみなす

階級	階級値	度数	相対度数	累積度数
141～145	143	1	0.0125	1
146～150	148	6	0.075	7
151～155	153	19	0.2375	26
156～160	158	30	0.375	56
161～165	163	18	0.225	74
166～170	168	6	0.075	80

度数分布表

連続型の例（身長データ）



- ヒストグラムを相対度数で描いて相対度数折れ線も描く
- 相対度数折れ線は、階級の幅を縮めていくと曲線になる
- 相対度数曲線が、身長の確率変数Xの確率分布を表しているとみなすことができる

相対度数分布曲線を使えば

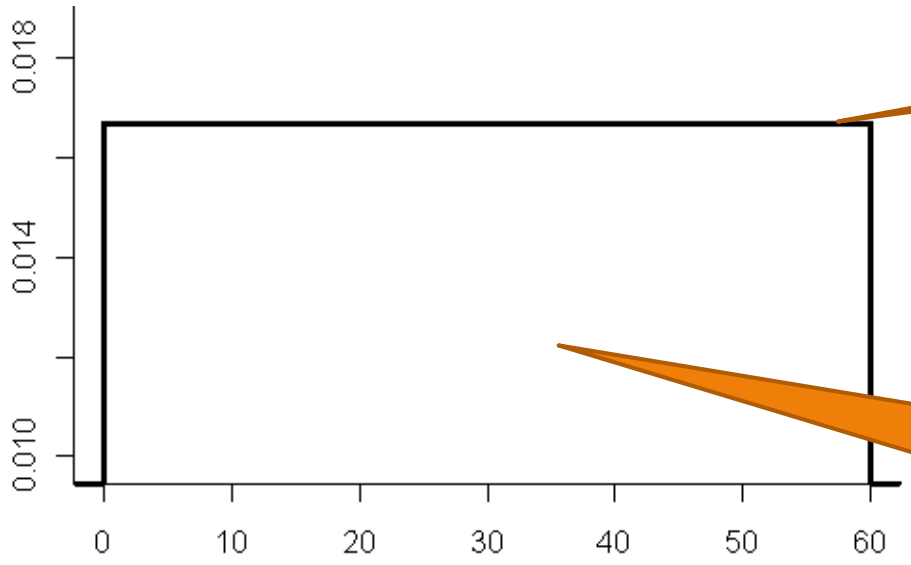
- 確率変数Xがaからbまでの値を取る確率を計算することができる
  - 積分で
- X=aである確率は計算できない（0になる）

確率密度関数

- 連続型確率分布では、離散型のときのように確率分布の表を描くことができない
- 但し、グラフは描ける
- グラフが描ける→式で表せる

確率密度関数

ストップウォッチの例での確率密度関数



確率密度関数

この部分の面積は必ず1になるようにする

$$f(x) = \begin{cases} 0 & (x < 0) \\ 1/60 & (0 \leq x < 60) \\ 0 & (x \geq 60) \end{cases}$$

累積分布関数

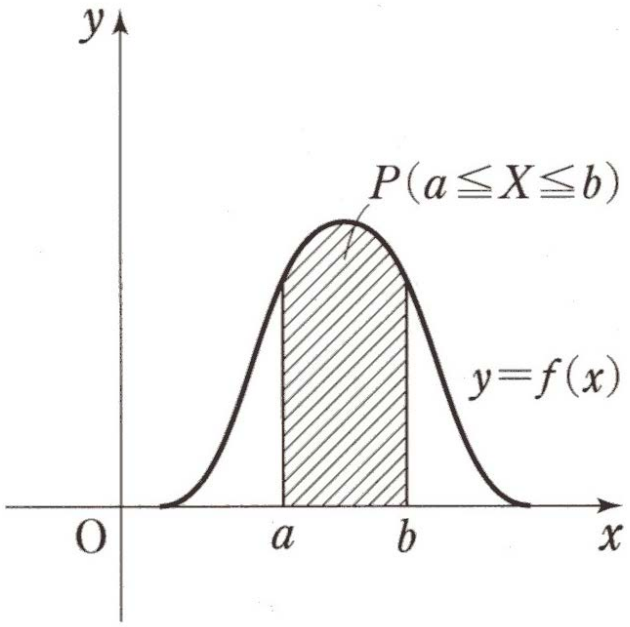
- 確率変数がx以下になる確率を表す関数
- ストップウォッチの例では

$$F(x) = \begin{cases} 0 & (x < 0) \\ x/60 & (0 \leq x < 60) \\ 1 & (x \geq 60) \end{cases}$$

度数分布表の「累積相対度数」の部分

確率密度関数の性質（連続型）

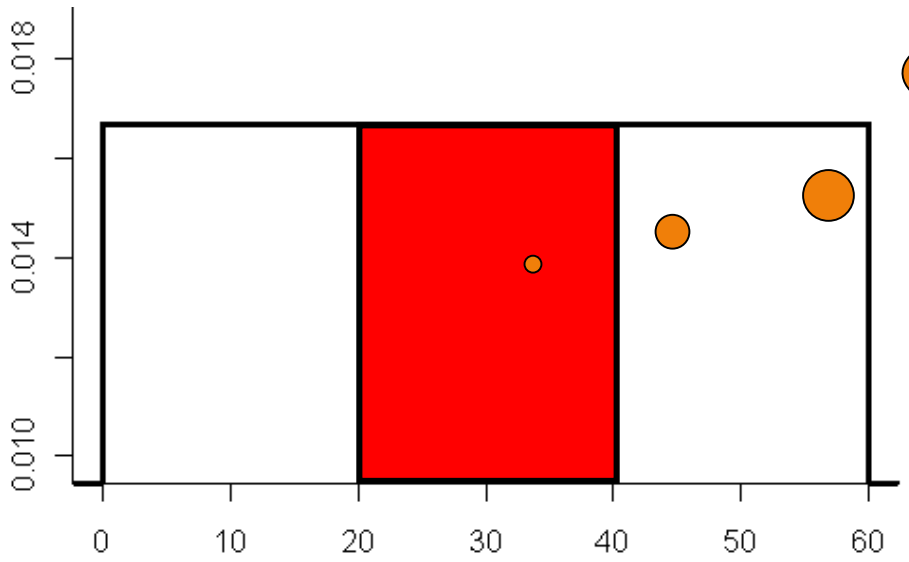
- 確率密度関数と横軸とで囲まれた部分の面積は、割合および確率と同一視できる



面積＝確率＝割合

積分と確率密度関数さえ分かれば確率が求められる

実際の例



この部分の面積が
確率になる

$$f(x) = \begin{cases} 0 & (x < 0) \\ 1/60 & (0 \leq x < 60) \\ 0 & (x \geq 60) \end{cases}$$

$$P(20 \leq x \leq 40) = \int_{20}^{40} \frac{1}{60} dx = \frac{1}{3}$$

確率の平均と標準偏差

確率の平均（期待値）と分散

確率変数 X が下の確率分布に従うものとする

X	x_1	x_2	$\cdots$	x_n	計
P	p_1	p_2	$\cdots$	p_n	1

期待値：$\mathrm{E}(X) = x_1p_1 + x_2p_2 + \cdots + x_np_n$

分散：$\mathrm{Var}(X) = (x_1 - \mathrm{E}(X))^2p_1 + \cdots + (x_n - \mathrm{E}(X))^2p_n$

離散型確率分布の場合
連続型確率分布では積分を使って定義される

標準偏差は分散の正の平方根

単純に言えば

- 度数分布表のところで、平均や分散を計算した計算方法とほぼ同じ

統計の平均・分散との違い

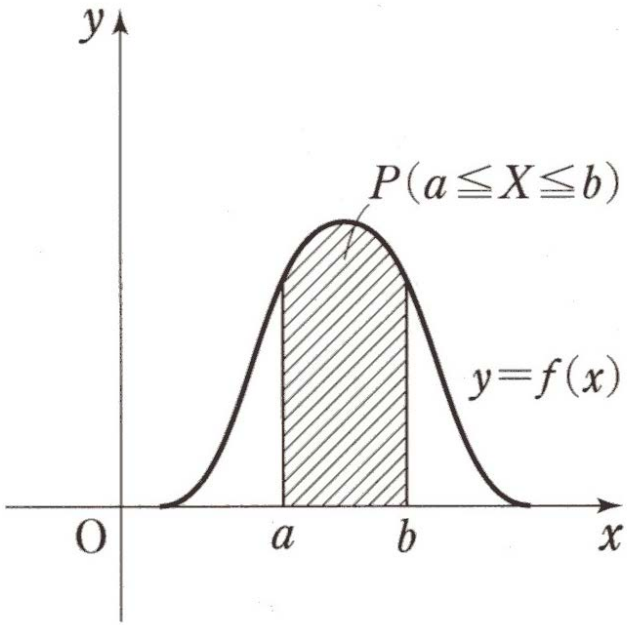
- 平均で比較
  - データの平均:「データ」と「そのデータが出てきた回数」を掛けたものの総和
  - 確率分布の平均:「確率変数の取り得る値」と「そのときの確率」を掛けたものの総和
- 分散・標準偏差も同じように考えればよい

期待値:データの平均や標準偏差などを調べたときに、その値になるであろうと期待される値

パーセント点とp値

確率密度関数の性質（連続型）

- 確率密度関数と横軸とで囲まれた部分の面積は、割合および確率と同一視できる



面積＝確率＝割合

積分と確率密度関数さえ分かれば確率が求められる

パーセント点

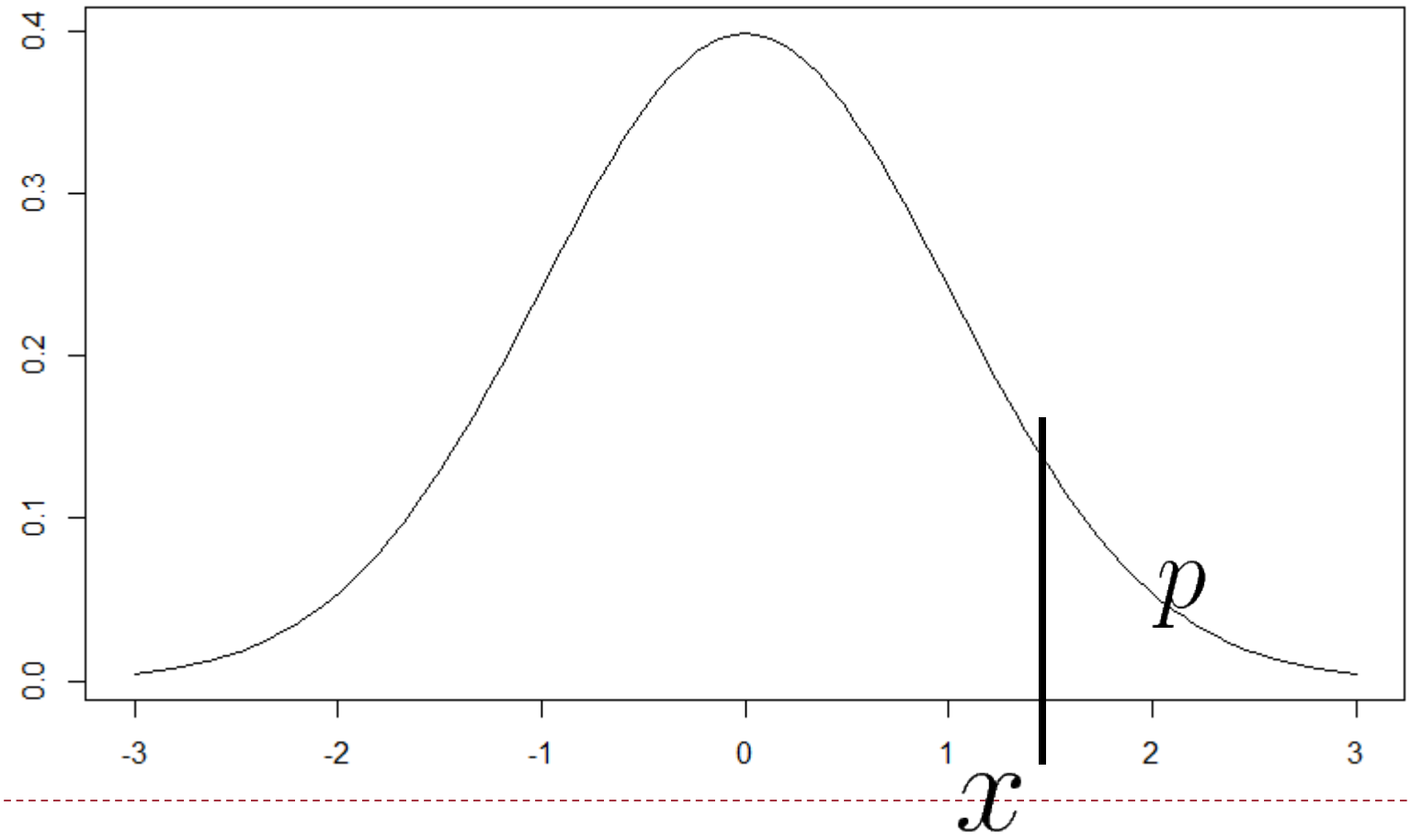
- データを小さい順に並べた時、その値よりも小さな値の割合が指定された割合になる値
  - 10%点：その値より小さいデータが全体の10%になる値
  - 50%点＝中央値
  - 0%点＝最小値、100%点=最大値
  - 上側4分位値＝75%点、下側4分位値=25%点

上側、下側、両側パーセント点

- 上側100pパーセント点：確率分布で、ある点xより右側の確率がpである、点xのこと
- 下側100pパーセント点：確率分布で、ある点xより左側の確率がpである、点xのこと
- 両側100pパーセント点：確率分布で、ある点xより右側の確率がp/2である、点xのこと

確率pが与えられている時に
点xを求める

図で表すと



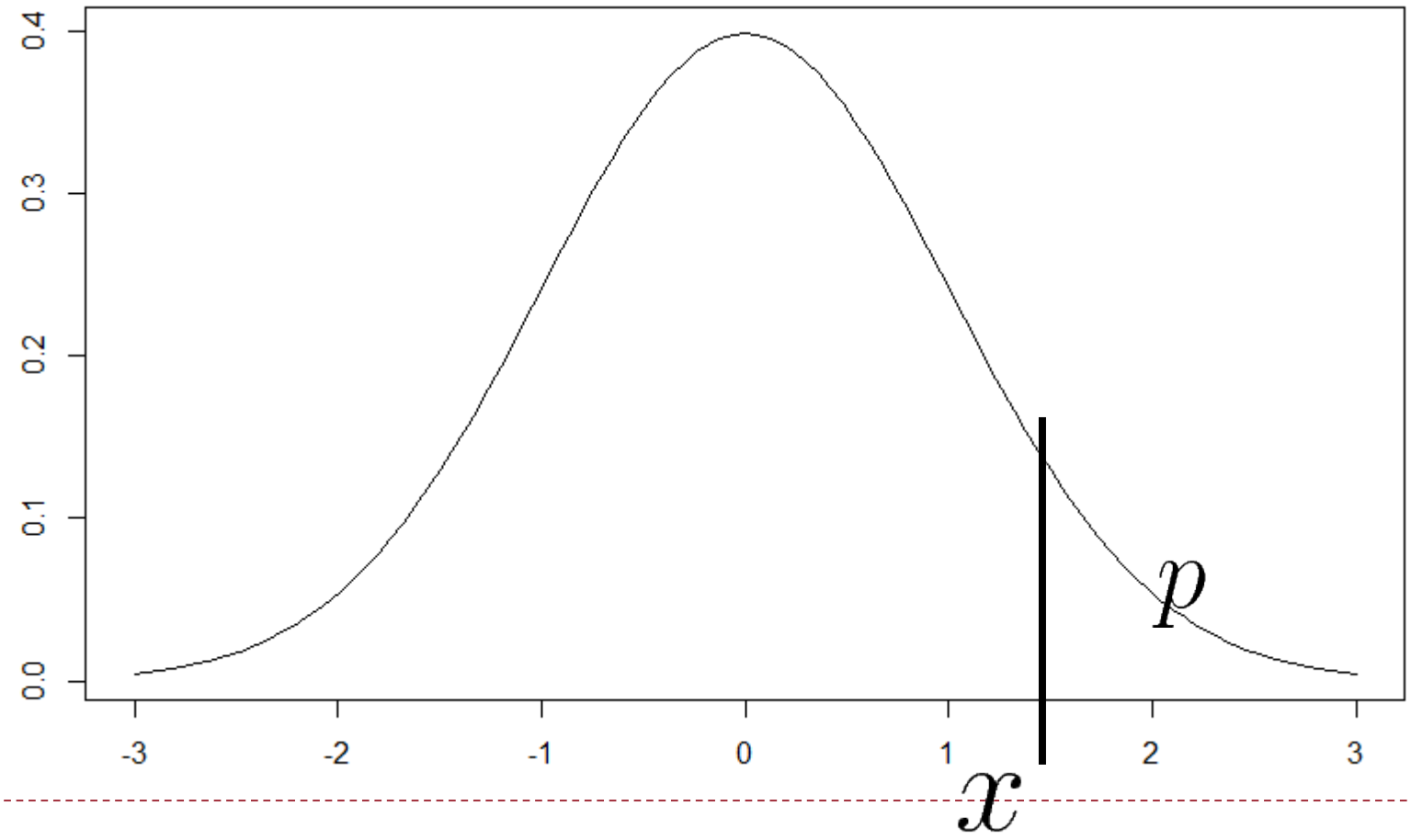
上側、下側、両側p値

- 上側p値：確率分布で、点xの右側を占める確率p
- 下側p値：確率分布で、点xの左側を占める確率p
- 両側p値：確率分布で、点xの右側を占める確率p/2

点xが与えられている時に
確率pを求める

実際は、EXCELで計算（次回以降）

図で表すと

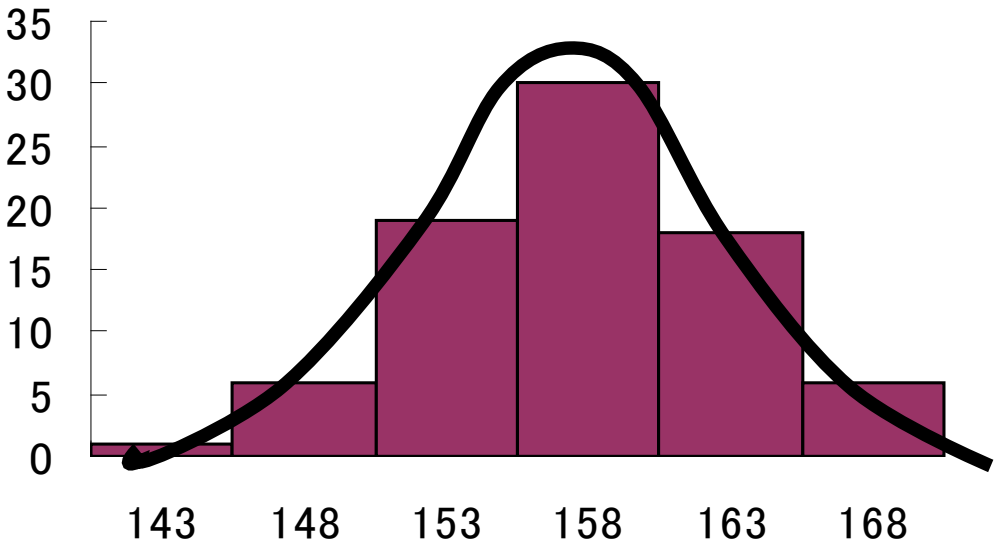


正規分布と２項分布

正規分布

- データの分布で最も代表的なもの
- 自然や社会で観測されるデータセットで非常に頻繁に現れる分布

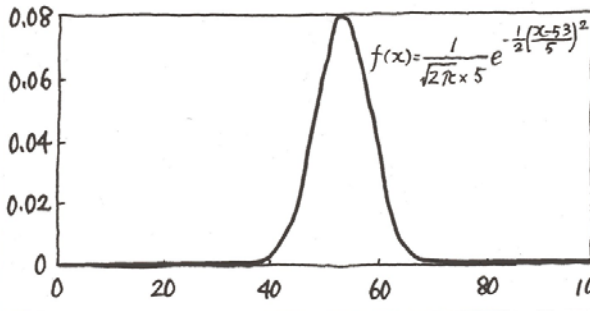
実際に



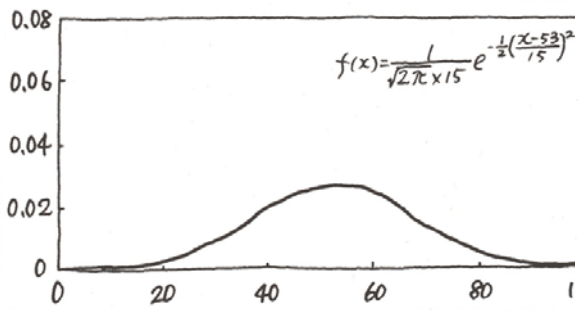
- 大抵のデータはヒストグラムで表すと正規分布のベル型カーブのような感じになる

正規分布のグラフ

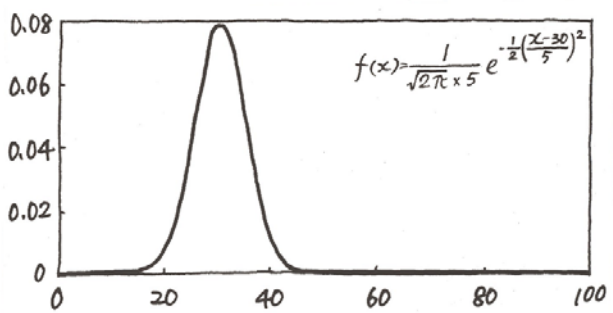
平均が53で標準偏差が5



平均が53で標準偏差が15

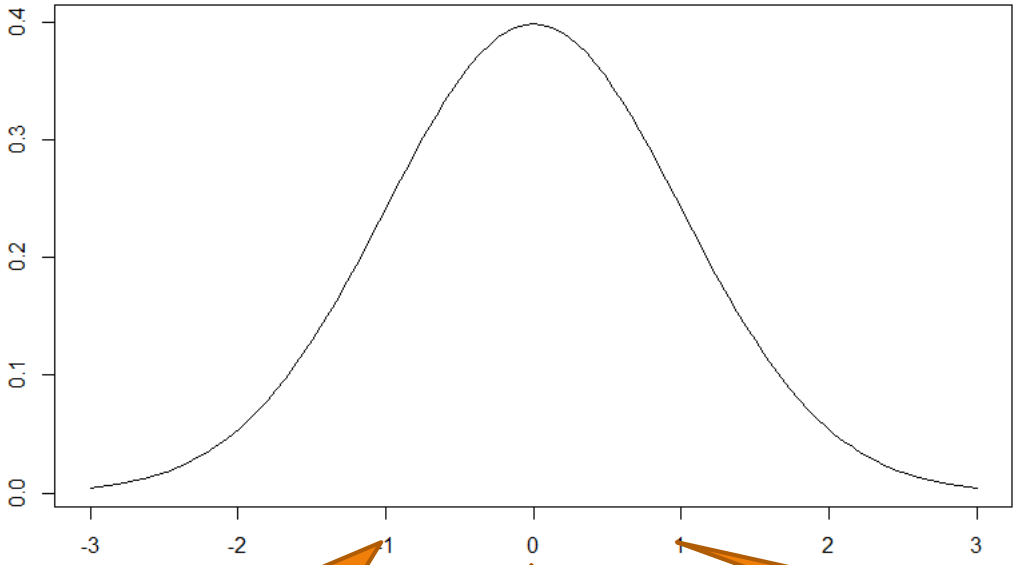


平均が30で標準偏差が5



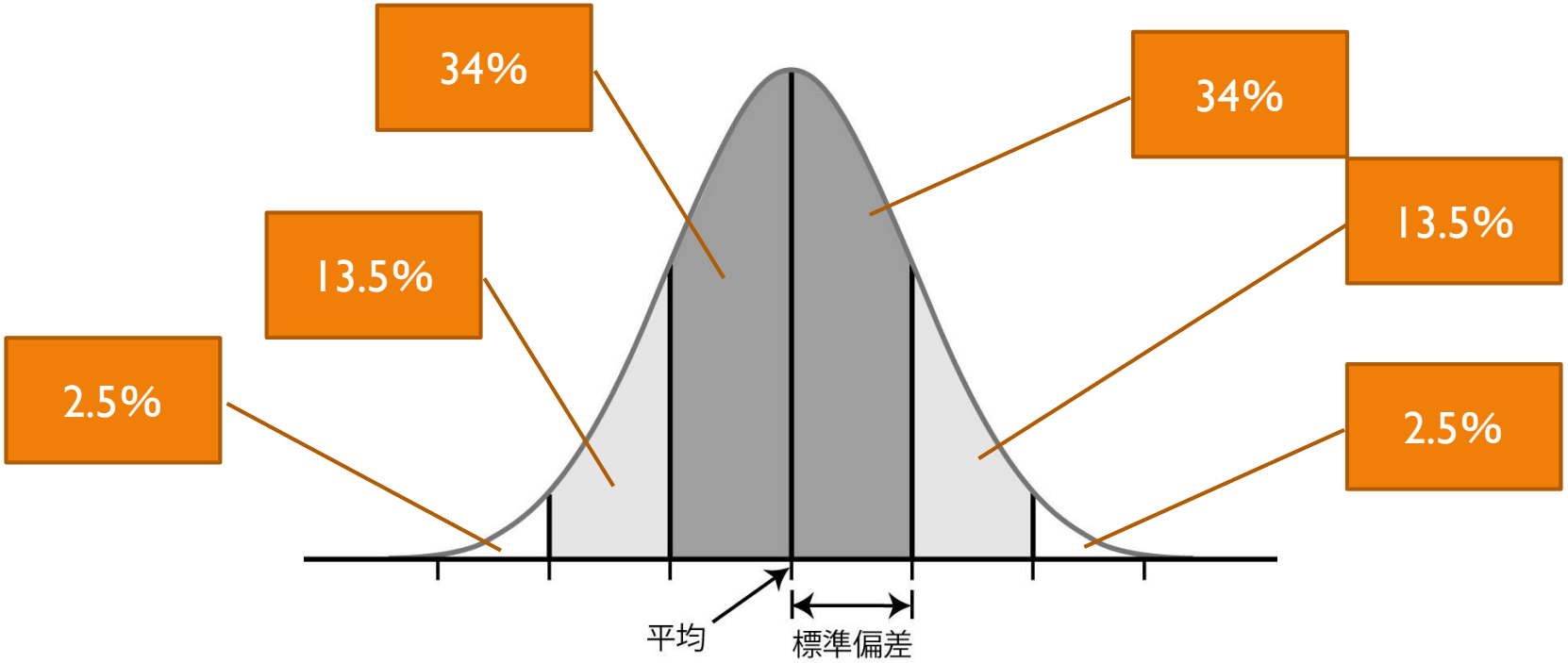
正規分布の性質１

- 左右対称、山型
- 山の頂上に対するXの値が平均
- 山の中腹に相当するXの値が平均±標準偏差



平均－標準偏差　平均　平均＋標準偏差

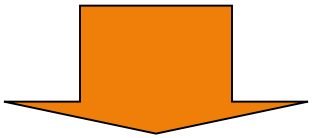
正規分布の性質２



- 全てのデータの約68%が、平均から標準偏差1つ分離れた範囲内に収まっている
- 全てのデータの約95%が、平均から標準偏差2つ分離れた範囲内に収まっている

標準化

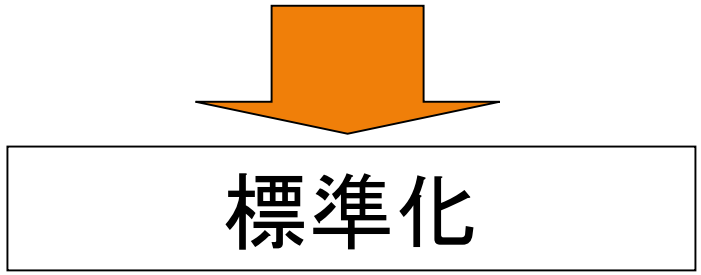
- データを、（データー平均値）÷標準偏差で加工すると、できたデータの平均と標準偏差は以下のようになる
  - 平均＝0
  - 標準偏差＝1



標準化

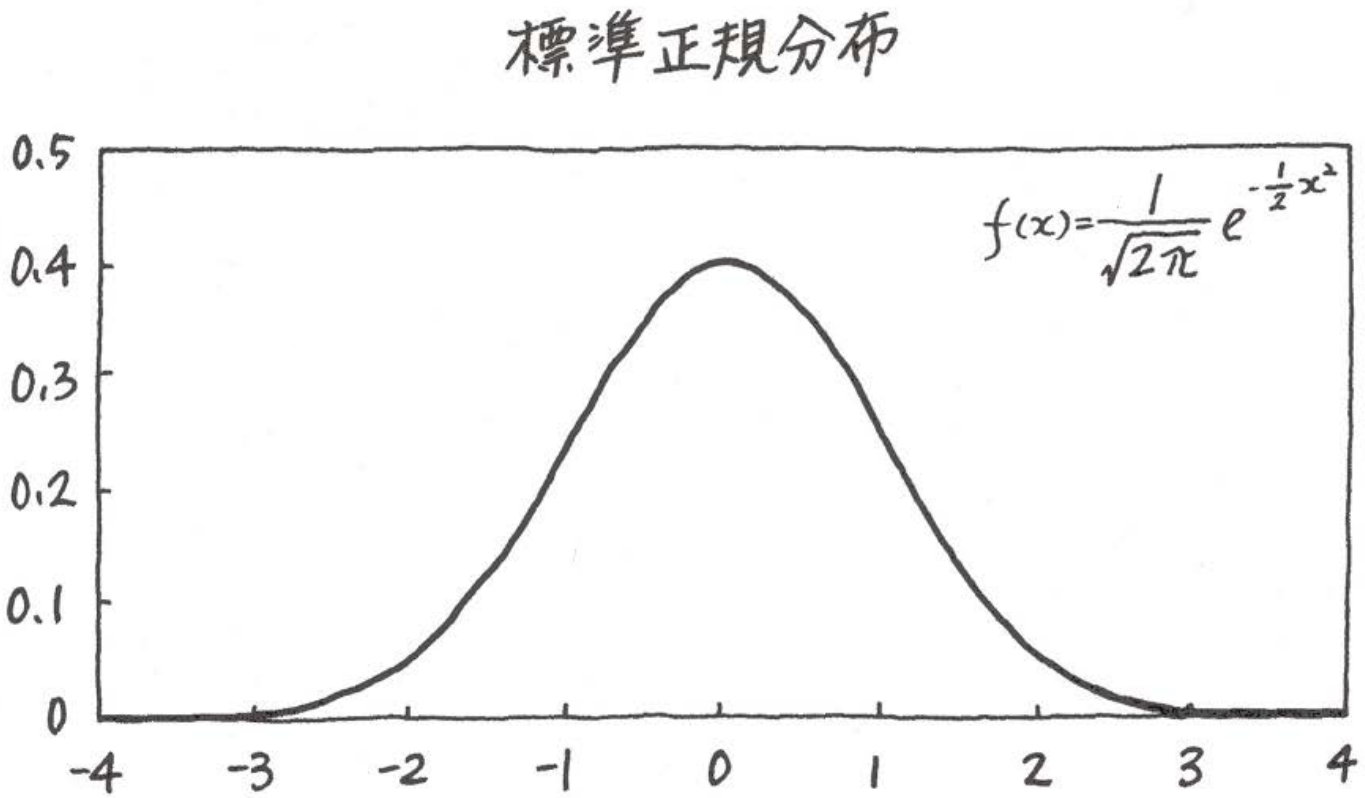
正規分布の標準化

- データを、（データー平均値）÷標準偏差で加工すると、できたデータの平均と標準偏差は以下のようになる
  - 平均＝0
  - 標準偏差＝1



- 標準化すると色々使いやすい
  - あらゆる平均と標準偏差の正規分布が標準正規分布の世界に置き換えられる

標準正規分布のグラフ



標準正規分布の性質

- 平均値＝０
- 標準偏差＝１
- 標準正規分布に従うデータが－１～＋１（平均から標準偏差１個以内の範囲）の値となる確率は0.6826（＝７０％弱）
- 標準正規分布に従うデータが－２～＋２（平均から標準偏差１個以内の範囲）の値となる確率は0.9544（＝９５％強）
  - 標準正規分布では、標準偏差２個の範囲内に殆どのデータが入ってしまう

一般の正規分布

- 一般の正規分布は、標準正規分布の全てのデータに一定数を掛けて、そのあと一定数を加えることによって得られる
  - 一般の正規分布のデータ＝σ(標準偏差)×標準正規分布表の値＋μ(平均)

標準正規分布の確率密度関数、累積分布関数の値から、一般の正規分布の値を求めることができる

二項分布

１回の試行で事象 A の起こる確率が p であるとき、この試行を n 回行う反復試行において A が r 回起こる確率は ${}_nC_r p^r q^{n-r}$。

このような反復試行において、A の起こる回数を X とすると、確率変数 X の確率関数は

$$p(x) = {}_n C_x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

と表される。このような確率分布を二項分布といい、$B(n, p)$ で表す。

但し、$0 < p < 1$、$q = 1 - p$ とする。

「反復試行の確率」をまとめたもの

二項分布の例

1 個のさいころを 3 回投げるとき、1 の目が出る回数 X は二項分布 $B\left(3, \frac{1}{6}\right)$ に従う。

1 枚の硬貨を 5 回投げるとき、表が出る回数 X は二項分布 $B\left(5, \frac{1}{2}\right)$ に従う。

不良率 1 %の製品の山から 50 個取り出したときの不良品の個数 X は二項分布 $B\left(50, \frac{1}{100}\right)$ に従う。

二項分布の確率分布例

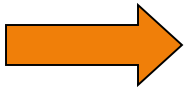
1個のさいころを3回投げるとき、1の目が出る回数 X は二項分布 $B\left(3, \frac{1}{6}\right)$ に従う。

$$p(x) = {}_3C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{3-x}$$

X	0	1	2	3	計
p	${}_3C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^3$	${}_3C_1 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^2$	${}_3C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^1$	${}_3C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^0$	1

二項分布の平均と分散

X：二項分布 $B(n,p)$ に従う

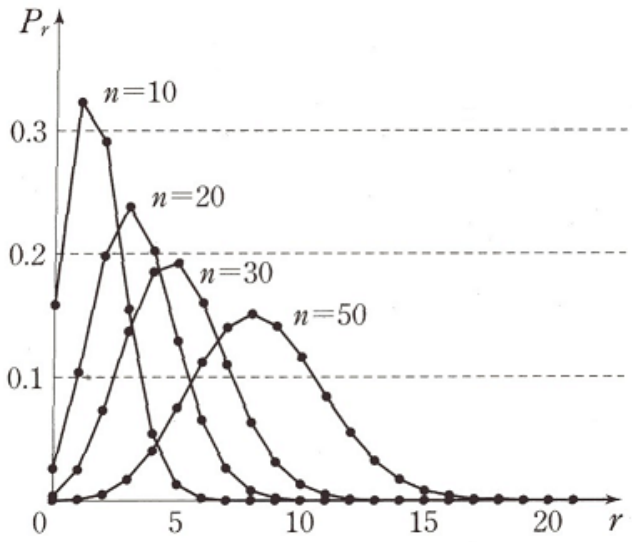


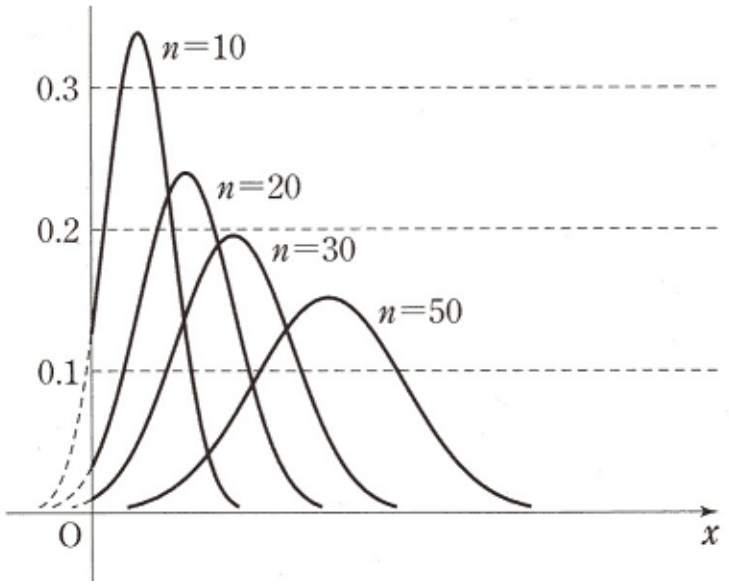
期待値：$E(X) = np$

分散：$V(X) = np(1-p)$

標準偏差：$\sigma(X) = \sqrt{np(1-p)}$

二項分布の正規近似





二項分布 $B(n,p)$ に従う確率変数 X は、n が十分大きいとき、近似的に正規分布 $N(np,npq)$ に従う。但し、$q=1-p$ とする。

Poisson分布

二項分布 $B(n,p)$ において、p が十分小さく（$p \leq 0.1$ 程度）、n が十分大きいとき、$\lambda = np$ とすると、その確率は

$$p_x = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0, 1, 2, \cdots)$$

と表すことができる。

確率変数 X が非負整数値（0 以上の整数）で、その確率関数が上の式で表されるとき、確率変数 X は母数 λ の Poisson 分布に従うという。

（$e = 2.7182\cdots$：自然対数の底、又は Napier 数）

Poisson分布の例

生起の希な現象を表す

- あるクラスの欠席者数
- 交通事故の死亡者数
- 電車に乗る一定時間間隔の乗客数
- 一定の時間内に電話交換台に入る回数

一様分布

- 連続型と離散型がある
- 離散型
  - コインを一回投げる
  - サイコロを一回投げる
- 連続型
  - アナログ型ストップウォッチの例
  - くじを引くための「一様乱数」を発生させるのに使われる