【統計 statistics】

統計を使える人は多い．統計を説明できる人は少ない．統計を理解している人はほとんどいない．

☆統計とは何か

The time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex world-wide states that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write.

Wells, H.G. (1903). *Mankind in the making*

All knowledge is, in final analysis, history.
All sciences are, in the abstract, mathematics.
All judgements are, in their rationale, statistics.

Rao, C.R. (1989). *Statistics and truth*

・不確実性を演繹可能にする

決定論から無秩序へ

無秩序には規則性がある？

・記述統計（descriptive statistics）と推測統計（inferential statistics）

～　標本（sample）から母集団（population）へ　～

☆統計の仕組み

・分布（distribution）の利用（不確実性にはカタチがある）

二項分布　ポアソン分布　超幾何分布　幾何分布　負の二項分布　離散一様分布　多項分布
正規分布　対数正規分布　指数分布　アーラン分布　ワイブル分布　ガンマ分布　三角分布
連続一様分布　ベータ分布　多変量正規分布　標準正規分布　t 分布　カイ二乗分布　F 分布

・推定量（estimator）から母数（parameter）を推測する（点推定・区間推定）

推定量に求められる性質（推測には標本変動が含まれる）
不偏性 unbiasedness
一致性 consistency
有効性 efficiency

☆本講義の射程（高校数学レベルでのイメージ作り）

確率分布の基礎
二項分布
正規分布

チェビシェフの不等式
中心極限定理
大数の（弱）法則

区間推定
仮説検定

☆確率分布を扱うための基礎知識

・確率変数と確率分布

確率変数：偶然性に支配された事象を反映した値
確率分布：確率変数の値とその時の確率との対応関係
　　変数が連続なら関数で表し，その関数を確率密度関数という．

・同時確率と同時分布

二つ以上の確率変数が同時にある値をとる確率を同時確率と言い，その分布を同時分布という．

・期待値（expected value）

確率変数とその確率を乗じたものの総和（平均）
加法性がある　「和の平均＝平均の和」

・分散（variance）と標準偏差（standard deviation）

・確率変数の線形変換

・独立と従属

・独立な確率変数の扱い

独立な確率変数 X，Y について

・期待値

・分散

・離散型確率分布（discrete probability distribution）

・連続型確率分布（continuous probability distribution）

☆二項分布（binomial distribution）

ベルヌーイ試行（伸るか反るかの独立反復試行）の分布　離散型確率分布の代表例

※ベルヌーイ試行（Bernoulli trial）

・定義

事象 A が起こる確率をp（起きない確率を $1-p$）とし，
n回のベルヌーイ試行を行ったとき，
Aが起きた回数 X は確率変数となる．
この X について確率分布は

と表される．

この確率分布を「二項分布」と言い，$B(n,\ p)$と表す．

・二項分布の具体例

大相撲に八百長はあるか

・二項分布 $B(n,\ p)$の期待値と分散

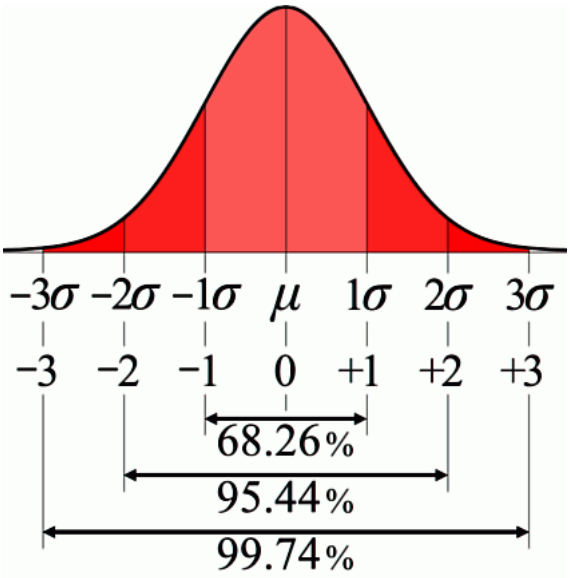
☆正規分布（normal distribution）/ ガウス分布（Gaussian distribution）

言わずと知れた分布の王様　連続型確率分布の代表例

・概要

「誤差」の分布は正規分布に従うとされており、非常に重要な連続型確率分布
正規分布に従うとされている分布の例は多い

・カタチの特徴



・正規分布の具体例

ケトレーが暴いたフランス軍徴兵試験における兵役逃れの不正

・正規分布の正確な表現

平均値がμ，分散がσ^2である正規分布を $N(\mu,\ \sigma^2)$ と表し，その確率密度関数は

と表される.

・導出のアイデア「正規分布は誤差の分布」

この式の導出過程の解説は高校数学を超えるが，こんなにもややこしい式がどうやって導かれるのか，納得のためにアイデアだけ示しておく.（ここでは最尤値が平均値と一致するという立場から導く流儀を採用する）

誤差の三公理（ガウス）
・小さな誤差は大きな誤差より起こりやすい
・絶対値の等しい正負の誤差は等確率で起こる
・非常に大きな誤差が発生する確率は非常に小さい（ある程度以上大きな誤差は実質起こらない）

・正規分布の骨格を規格化

※ガウス積分（範囲外なので結果だけ用いる）

☆正規分布の活用

・標準正規分布

$N(0,\ 1)$ である正規分布を特に標準正規分布と言う.

この分布に関して，確率密度関数の積分値は標準正規分布表として細かく用意されており，積分計算を実行せずとも分布のおよその確率はわかる.

・二項分布の正規近似（ド・モアブルーラプラスの定理）

中心極限定理（CLT）の Special Edition

確率変数Xが二項分布 B(n, p)に従うとき，nを十分大きくとれば

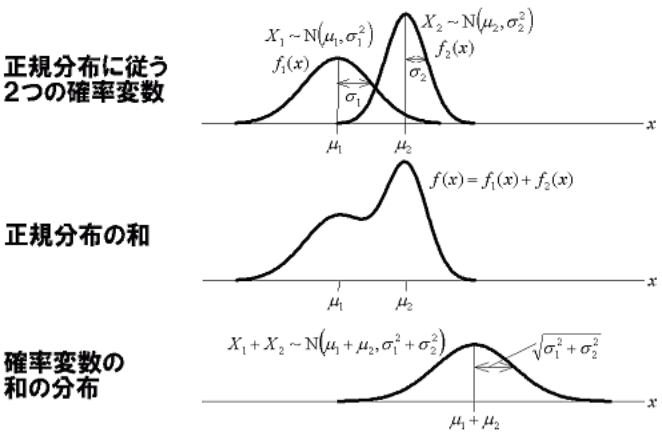
確率変数

は近似的に標準正規分布に従う.

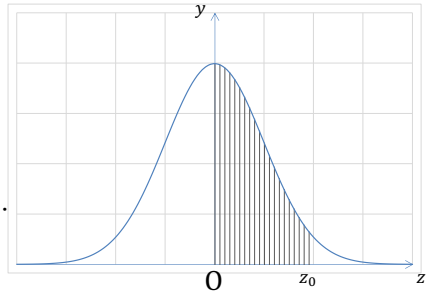
※記法「分布に従う」

・再生性（reproductive property）

二項分布と正規分布には再生性がある.



・標準正規分布表はトモダチ



次の表は，標準正規分布の分布曲線$\left(f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)$における

図の縦線部分の面積（確率）の値$\left(P(0\leq z\leq z_0)=\int_0^{z_0}f(z)dz\right)$をまとめたものである．

※標準正規分布表は他にも $P(z_0 \leq z)$や $P(z \leq z_0)$などの与えられ方がある．

与えられ方に合わせて表を読もう．

z_0	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

$Q1.\ P(0\leq z\leq 1.00)$

$Q2.\ P(0\leq z\leq 1.96)$

$Q3.\ P(-1.96\leq z\leq 1.96)$

$Q4.\ P(0\leq z\leq 2.58)$

$Q5.\ P(-2.58\leq z\leq 2.58)$

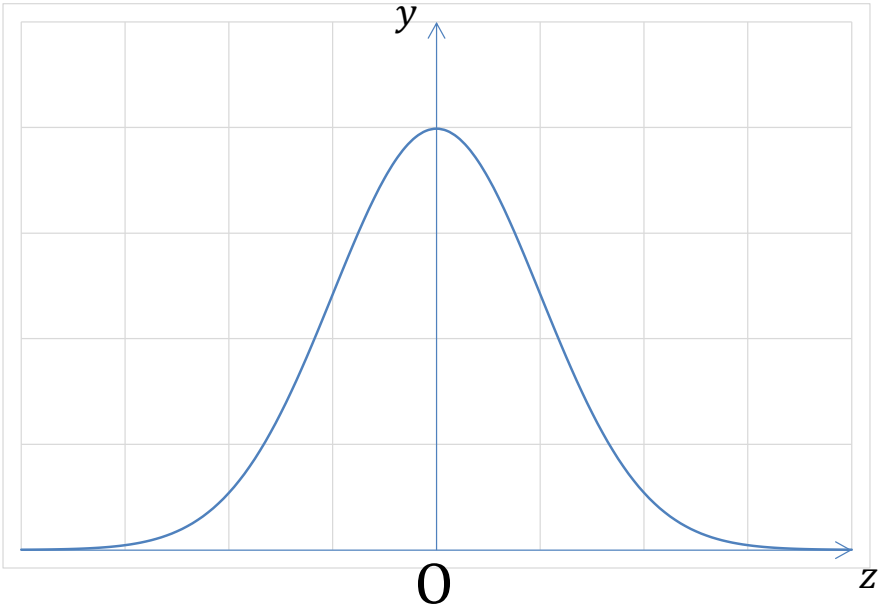
$Q6.\ P=0.95$ となる$|z|$の範囲

$Q7.\ P=0.99$ となる$|z|$の範囲

$Q8.\ P(z\leq z_0)=0.95$ となるときのz_0

$Q9.\ P(z\leq 1.96)$

$Q10.\ P(z\geq 2.58)$



☆確率と統計の架け橋（歪な現実を理論的に均す）

・チェビシェフの不等式（Chebyshev's inequality）

確率変数 X の分布が，平均 μ，分散σ^2であるとき，任意の実数$k > 0$ において

が成り立つ.

・大数の（弱）法則　（確率収束：弱い weak law of large numbers / WLLN）（概収束：強い strong … / SLLN）

・（古典的な）中心極限定理（central limit theorem / CLT）

確率変数 X の分布が，平均 μ，分散σ^2であるとき，

Xと同じ確率分布に従う独立な確率変数X_1，X_2，X_3，…，X_nについて

$Z = \frac{1}{n}\sum_{k=1}^{n} X_k$ は$N\left(\mu,\ \frac{\sigma^2}{n}\right)$に従う確率変数$Z$に法則収束する.

↓高校生向けに翻訳

☆標本調査（sample survey）

全数調査（complete survey）を回避するために

・標本調査の難しさ

無作為抽出（random sampling）… 互いに独立なデータでないと中心極限定理が使えない.
異常値（abnormal value）… 測定，記録上のミスやその他，理由があって異常の大きな値.
※外れ値（outlier）… 他と著しく外れた値．異常値を含む．なんらかの基準をもって除外する.
　外れているからと言って異常値であるとは限らない．除外すべきでない有益な外れ値もある.

・標本から母集団を推測する

不偏推定量　… 期待値が推定したい値と等しくなる量
・標本の大きさ（sample size）
　大標本（large sample）… 正規分布とみなし，標本分散をそのまま不偏推定量として用いる.
　小標本（small sample）… 標本サイズに応じた不偏推定量として（期待値を合わせた）不偏分散を用いる.
・標本平均から母集団へ

正規母集団 N（μ，σ^2）からの標本X_1，X_2，…，X_nは独立で，$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ とする.

母分散（population variance）	標本分散（sample variance）	不偏分散（unbiased variance）
$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$	$s^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$	$u^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$
母集団の分散	大標本のときは 標本分散を不偏推定量として 母分散の代わりに そのまま用いる	小標本のときは 標本分散は不偏推定量でないので 不偏分散に変換して用いる 自由度$(n-1)$の分布

$E(\bar{X})=\mu\left(\text{母平均}\right)$

$V(\bar{X})=\frac{\sigma^2}{n}\left(\frac{\text{母分散}}{\text{標本サイズ}}\right)$

・視聴率の話（標本の大きさの根拠）

・出口調査の話（当選確実の根拠）

☆推定（区間推定）estimation（interval estimation）

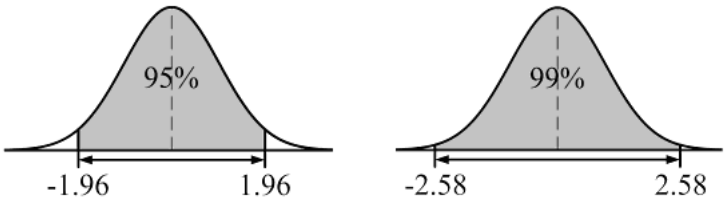
・区間推定の仕組み（信頼度と信頼区間 confidence coefficient & confidence interval）

（1）母平均（population mean）の推定（区間推定の基本）

「母平均がどの程度の確率（信頼度）でどんな範囲（信頼区間）に存在するか」を標本平均から推定する．

標本サイズが十分に大きければ，正規母集団 $\mathrm{N}\left(\mu,\ \sigma^2\right)$ からの標本平均 $\bar{X}$ は $N\left(\mu,\ \frac{\sigma^2}{n}\right)$ に従う．

$z=\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ とすると，z は $N\left(0,\ 1\right)$ に従うので，標準正規分布を利用する．



$P(-\alpha \leq z \leq \alpha)=0.95$ のとき

$P(-\alpha \leq z \leq \alpha)=0.99$ のとき

（2）母比率（population proportion）の推定

（3）標本の大きさ（sample size）の決定

・区間推定の練習問題

問題 1（母平均）

ある工場で作られるチョコレートについて，無作為に 100 個のサンプルを抽出して重さを調べたところ，標本平均は 50g で，標本標準偏差が 2g であった．このチョコレートの重さの母平均に対する信頼度 95%の信頼区間を求めよ．

問題 2（母比率）

ある原野に A 種，B 種の 2 種のネズミが生息している．任意に 400 匹のネズミを捕獲したところ，A 種が 80 匹いた．この原野全体において A 種のネズミは何%生息していると考えられるか．信頼度 95%で推定せよ．ただし，標準正規分布において，$P(|z| \leq z_0) = 0.95$ となるとき$z_0 = 2$としてよい．

問題 3（標本の大きさ）

ある年度における 20 歳以上の日本人女性の身長は，平均がμ(cm)，標準偏差が 6.3 の正規分布をなす．大きさ 121 の標本を無作為抽出して平均を求めると $153.1(cm)$であった．

(1)信頼度 95%でμの信頼区間を求めよ．

(2)信頼度 99%でμの信頼区間を求めよ．

(3)標本平均とμの差が 0.5 以下になる確率を 95%以上にしたい．標本の大きさをどの程度にすればよいか．

☆仮説検定（statistical hypothesis testing）

・仮説検定の仕組み（危険な背理法）

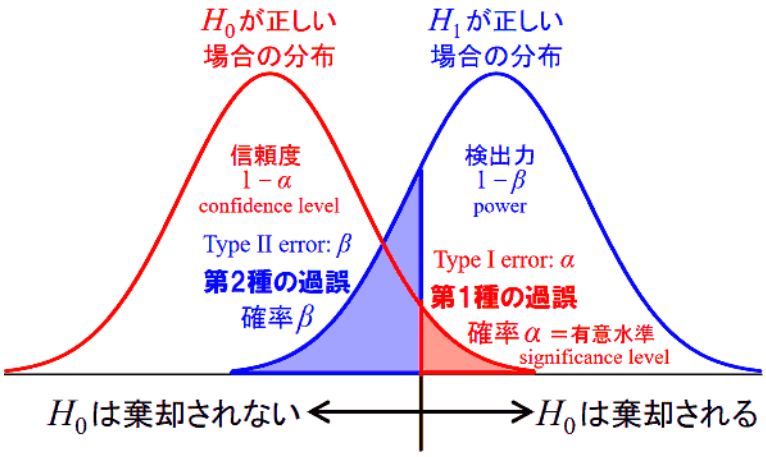
・帰無仮説と対立仮説（有意でないか有意であるか）

仮説検定では，仮説を否定することでもう一つの仮説の正しさを選択する．否定することを前提として立てる仮説のことを帰無仮説と言い，帰無仮説が否定されることを棄却されると言う．帰無仮説と対立するもう一つの仮説を対立仮説と言い，帰無仮説が棄却された場合は対立仮説が正しいとされる．帰無仮説が棄却されない場合は帰無仮説が採択されると言うが，それは帰無仮説を否定することはできないというほどの意味でしかなく，積極的に帰無仮説を認めるものではない．

・2 種類の過ち

・有意水準（危険率）

帰無仮説が正しいときに，
検定統計量（実測値）が設定した極端な範囲，
すなわち危険域（棄却域）に入る確率のこと．



・棄却域と採択域

・第 1 種の過誤

・第 2 種の過誤

・α と β はトレードオフ

・仮説検定の練習問題

問題 4（両側検定）

あるコインを 100 回投げたところ，60 回表が出た．このとき，このコインには何らかの細工がされていると言えるか．有意水準（危険率）5%で検定せよ．

問題 5（片側検定）

ある大学の入学試験は 550 点満点の試験に対し，平均点が 250 点，標準偏差が 77 点であった．このとき N 高校からの受験者は 49 名おり，その平均点は 268 点であった．N 高校からの受験者は受験者全体に比べて優秀であると言えるか．有意水準 5%で検定せよ．

問題 6（有意水準を求める）

ある工場で生産される製品が平均して 1%の不良品を含む．ある日，この工場で生産された製品から無作為に 100 個のサンプルを抽出し検品したところ，2 個の不良品が含まれていた．

(1)合格品，不良品に対して，各々確率変数 1 と 0 を与えるとき，標本平均および標本標準偏差を求めよ．

(2)ある有意水準で検定すると，この日異常があったことにはならない．その有意水準の最大値を，%で少数第 1 位まで求めよ．