やさしい統計学講義

夏木康志

目次

もくじ

1日目

講義開始前のアンケート

統計学の講義を始める前に簡単な受講者アンケートをしたいと思います。

１　お名前

　せっかく講義を受講してくださる受講者のお顔とお名前は全て覚えたいです。

（ オンライン講義では必ずしも受講者は顔を出す必要はないです ）

２　所属、専攻、ゼミ

　だいたいどのようなことをご専門にされているか知り、レベルにあった講義を心がけます。

３　卒業論文、ゼミのテーマ

　あなたがもし４年生であれば、卒業論文のテーマを教えてください。（ もしその論文に統計学の知識が使えるのであれば、そのような内容を優先して講義します。）

４　高校時代にやった数学の範囲

　数学 IAIIB まで、数学 III も勉強した。場合の数、確率、確率分布も履修した。（ もししてなかったら、そのレベルに合わせて講義します。）

５　大学に入ってから履修した数学の範囲

微分・積分（ 解析 ）、線形代数、確率論、統計学のどれかを履修した。（ していない人が

6　高校時代の得意科目

あとで相関係数の説明をする時に使います。 \newline

7　大学時代の得意科目

同様です。 \newline

以上です。

トーク

以前、ある国立大学で統計学の講義をした時、はじめて統計学の講義をしたこともあり、初日は特になかなか上手くゆきませんでした。

まず大学の統計学の文系向けの講義で最低限理解していただきたいのは、平均、分散、標準偏差です。これをじっくり手計算で教えて、さらに二日目の課題としては正規分布の性質についてイメージを持ってもらい、三日目では t 分布や F 分布、カイ二乗分布について知ってもらう。四日目で共分散、相関係数、回帰直線について説明して、理解してもらう。これだけで十分だと思います。

3 日目の内容と 4 日目の内容の順番を入れ替えてもよいと思います。

なので、統計学の講義をする際は、まとまった知識を受講者に与えるという方針ではなく、しぼった伝えたい内容をじっくり手計算で伝授する、という方針の方がよいと感じました。

今回も昨今の感染症事情で遠隔講義が想定され、なかなか受講者の顔と名前を覚えて、インタラクティブに講義ができるのかは、わかりませんが、できる限り、インタラクティブに講義したいです。

Ch.1. はじめに：統計学ってどんな学問

はじめに：統計学ってどんな学問？

これから四日間、統計学の講義の集中講義を担当いたします。

まず最初に統計学ってどんな学問ということについて説明し、その後、詳しい講義計画 (15 回分＋中間テスト＋最終テスト ）についてご説明いたします。この電子書籍は補助教材という位置づけで、実際の講師の説明に沿って理解されることをおすすめします。実際に、とある国立大学の統計学の講義としてこの補助教材を使ったことがありますが、この補助教材をそのまま読み上げるような講義をしたら、受講者のやる気をそいでしまいます。インタラクティブに講義は進める必要があります。

統計学とは

統計学とは科学の文法とも呼ばれ、データ・サイエンスや EBPM（ エビデンスに基づく政策決定 ）が重視される中で、無視できない学問です。また Google のエコノミストであるヴァリアン先生によるともっとも魅力的な職業は統計学を駆使したデータサイエンティストとされています。

人間社会には不確実性の中での意思決定がつきものです。将来何が起こるかわからない状態で、最適な意思決定を行うためのデータに基づく推論の手段、それが統計学です。

統計学は、確率・統計という科目として教えられるように、確率論と切っても切れない関係にあります。本来、統計学は数学とも深い関係にあり、統計学の基礎は、数理統計学とも呼ばれます。

しかし、本講義では、数学があまり得意ではない文系の受講者にも配慮して、極力数式を使わず、エッセンスを説明したいと思います。

それでは四日間の講義の予定を説明したいと思います。

やさしい統計学講義 シラバス案

1 日目

Ch1. はじめに：統計学ってどんな学問？

講義開始前の簡単な受講者アンケート

Ch2. 平均と分散、標準偏差

Ch3. 組み合わせとパスカルの三角形

Ch4. 二項分布

2 日目

Ch5. 正規分布の性質

Ch6. 正規分布の応用

Ch7. 大数の法則

Ch8. 中心極限定理

中間テスト

３日目

Ch9. 検定の概念

Ch10.t 分布、F 分布、カイ二乗分布の使い方

Ch11. 検定の応用

Ch12. 推定の概念

4 日目

Ch13. 多変数の統計

Ch14. 相関係数

Ch15. 回帰分析

Ch16. ふりかえりのテスト

講義の流れとしては、講師の方から簡単な説明を行い、受講者の皆様には適宜応用の回で、グループディカッションや演習を交えて、進めていくという方針です。

テストに関しては、講義の内容が理解できれば、単位は必ず来るようにしたいです。

毎回の講義の内容について、簡単なキーワードを交えつつ、ざっと説明したいです。

1 日目

Ch1. はじめに：統計学ってどんな学問？

今回ですね。全体の講義の計画と、統計学という学問のあらましについて説明いたします。今までの社会において、読み書きそろばん、つまり読むこと (Read)、次に書くこと (wRiting)、その次に計算・算数の技術 (aRithmetic) という順番で 3 つの R が重視されてきました。さらに現代では加えて、統計リテラシー (statistical Reasoning) という４つめの R が重視されています。これは統計的思考法を身に付けることで、不確実な中での意思決定を効率的に行うことが求められているからです。

Ch2. 平均と分散、標準偏差

ここでは平均、分散、標準偏差という統計学を学ぶ上での基礎的な概念について学びます。μ、σ^2、σについて理解します。μはギリシア文字で m に相当する文字で、統計学では mean つまり平均を意味します。σはギリシア文字で s に相当する文字で、統計学では standard deviation つまり標準偏差を意味します。ローマ字のアルファベットでは足りないために、ギリシア文字を使用するのですね。

Ch3. 組み合わせとパスカルの三角形

高校時代に順列 (!) や組み合わせ、場合の数について学ばれた経験がある方が多いと思います。ここではそれらの概念をざっと振り返り、パスカルの三角形という簡単かつ基礎的な概念を使って、今後の学習の基本をお伝えいたします。(場合の数、確率の復習の回です。)

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

. . . . . .

Ch4. 二項分布

ここでは分布の概念の基礎となる二項分布について簡単に説明します。サイコロを振った時の目の分布や、コイントスの表と裏の組み合わせの分布について説明します。

ここまでで 1 日目の講義は終了です。平均と分散、標準偏差についてよく復習しておいてください。余力のある方は教科書・参考書等で正規分布の性質について予習しておいてください。

2 日目

Ch5. 正規分布の性質

ここでは正規分布の性質について、極力数式を使わずに説明いたします。正規分布は統計学における分布の女王と呼ばれる基本的な概念であり、左右対称、分布の山が一つしかない、分布の左右の裾が広がっているという 3 つの基本的な性質があります。

Ch6. 正規分布の応用

正規分布がわかると得られる情報。正規分布と平均 (μ) と標準偏差 (σ) の関係。1 σ =68%,2 σ =95%,3 σ =99.7% といった話。

ここでは正規分布の性質をふりかえり、受講者の皆様にも演習の形式で、正規分布の性質について話し合います。

Ch7. 大数の法則

標本数 n が十分大きくなると、標本平均 μ が母平均と等しいと見ないして良いという大数の法則を説明します。大数の法則は、コイントスを繰り返すと、ほぼ表が出る確率が 1/2 に近くなるという内容を一般化した法則です。

Ch8. 中心極限定理

標本数 n が大きくなると、x1+x2+x3+...+xn が正規分布に従うという中心極限定理に

ついて説明します。

二項分布で n が大きくなると正規分布に近づくという話をヒントにします。

二日目はここまでです。受講者の皆様は、正規分布の性質について復習しておいてください。余力のある方は、検定の概念について予習しておいてください。

３日目

Ch9. 検定の概念

ここでは仮説検定について説明します。第 1 種の過誤と第 2 種の過誤について説明します。第 1 種の過誤の例は「能力の有る者に博士号を与えない」、第 2 種の過誤の例は「資質のない者を博士課程に進ませてしまう」があります。

Ch10.t 分布、F 分布、カイ二乗分布の使い方

t 分布、F 分布、カイ二乗分布を用いた検定の仕方についてご説明します。

F 分布は分散の比。カイ二乗分布は母分散についての検定です。

Ch11. 検定の応用

ここではグループディスカッション、演習を交えつつ、検定の応用例について議論します。

Ch12. 推定の概念

ここでは区間推定 (estimation) の概念について説明します。テレビの視聴率は何 %、選挙の際の当選確率は何 % といった時の推定の話題を説明します。

ここまでで三日目の講義は終わりです。検定の概念についてよく復習してください。余力のある方は、相関係数と回帰分析について、予習をしておいてください。

4 日目

Ch13. 多変数の統計

ここではクロス集計表（ 分割表 ）と多変数の統計について説明します。

Ch14. 相関係数

相関係数と共分散（ 2 変数間の分散 ）について説明します。

Ch15. 回帰分析

回帰分析について簡単に説明します。

Ch16. ふりかえりのテスト

基本的に正規分布の性質について理解できれば、単位が来るようにしたいです。ご健闘をお祈りします。

参考文献について

なかなか皆様のレベルにあった教科書を選ぶのが難しいと思っております。そんなわけで、このような電子書籍の補助教材を用意しました。

入門向けのテキストとして鳥居泰彦 (1994)『はじめての統計学』を推薦します。

一般的な大学生向けの講義のテキストとして東京大学教養学部統計学教室編 (1991)『統計学入門』を一度講義に使ったこともあるのですが、若干内容は難し目で、辞書として使った方がよろしいかも。参考書としては、石井俊全 (2012)『まずはこの一冊から意味がわかる統計学』を推薦します。

東大出版から出ている『統計学入門』は若干内容が難しいですが、定評ある教科書で、持っていて損はないです。同じく、久保川・国友 (2016)『統計学』もありますが、この本は数学の理解を前提としています。

本来統計学は微分・積分に基づく確率論の応用という側面がありますが、初学者にはできれば数学的知識を仮定せずに、統計学の基礎概念や使い方を中心に説明したいです。そのような意味では、石井著の参考書をおすすめします。

理系的な内容の本格的な統計学を学びたい方は『確率・統計』の本を経由して『数理統計学』に進まれることをお勧めします。回帰分析の先を学びたい方は、『計量経済学』を

学ばれるとよいでしょう。

また統計学を学ぶと、機械学習やネットワーク解析を学習される際のヒントとなるでしょう。統計的なパッケージとしては、R、SPSS、STATA、E-Views などがありますが、特にフリーソフトの R の利用をおすすめいたします。

それでは、統計学の世界にようこそ！！

宿題：オンライン開講を前提に、受講者の皆様にお願いしたいことがあります。フリーソフト R をお使いの端末にインストールしてください。R のインストール方法は、インターネットで検索すればすぐに出てくるはずです。

統計ソフト R をインストールすると、各種分布のグラフが簡単に描けて便利です。皆様のレポートや卒業研究にも役に立ちます。深層学習のアルゴリズムを試すことも可能です。できるだけお使いの端末に R をインストールしてください。

Ch.2. 平均と分散、標準偏差

目標：平均、分散、標準偏差について理解する

ここでは平均、分散、標準偏差という統計学を学ぶ上での基礎的な概念、情報について学びます。μ、σ 2、σについて理解します。μ（ ミュー ）はギリシア文字で m に相当する文字で、統計学では mean つまり平均を意味します。

σ（ シグマ ）はギリシア文字で s に相当する文字で、統計学では standard deviation つまり標準偏差を意味します。

平均は幾何学的には重心を意味します。分散は偏差（ 平均からの各変数の差 ）の二乗の平均です。標準偏差は分散のルートです。

分散という情報はデータのちらばり具合を意味します。統計学というデータ（ 情報 ）を扱う学問では、データを二乗した距離が重要となってきます。

データ間の距離を測るためには、絶対値（ データ同士の差をとって、マイナスの時は符号をプラスに読み換える ）だけではなく、データの差の二乗が重要になってきます。（ ある意味、面積ですね。）データの差の二乗に基づく距離概念が、データのちらばりを表す分散です。分散は二乗しているため、単位も二乗となって都合が悪いので、分散のルートをとったものを標準偏差といいます。

データの正規化についてもお話ししましょう。

データの各要素から平均を引いて、標準偏差で割ったもの (Z) を計算することをデータの標準化といいます。標準化したデータをあらためて、Z あたり 10 を掛けたものに 50 を加えたものを偏差値と呼びます。

平均値の他に中央値、最頻値も重要な概念です。中央値はソートした数値の中央の値、最頻値は度数分布の山に相当する値です。

また平均値は、算術平均、幾何平均、加重平均などさまざまな概念がありますが、まずは算術平均を基本として利用しましょう。

STUDY!

E(X+a)=E(X)+a

V(X+a)=V(X)

E(bx)=bE(X)

V(bx)=bˆ2V(X)

ただし a,b は定数。

データ

$$X = x_1, x_2, \ldots, x_n$$

平均

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

偏差

$$X - \mu = x_1 - \mu, x_2 - \mu, \ldots, x_n - \mu$$

分散

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}$$

標準偏差

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

標準化

$$Z = \frac{X - \mu}{\sigma} = \frac{x_1 - \mu}{\sigma}, \frac{x_2 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}$$

偏差値

$$50 + \mathrm{Z} \times 10$$

基礎統計学数式保存.png

Ch.3. 組み合わせとパスカルの三角形

目標：高校の場合の数を復習し、パスカルの三角形を理解する。

高校時代に順列 (!) や組み合わせ、場合の数や確率論について学ばれた経験がある方が多いと思います。ここではそれらの概念をざっと振り返り、パスカルの三角形という簡単かつ基礎的な概念を使って、今後の学習の基本をお伝えいたします。

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

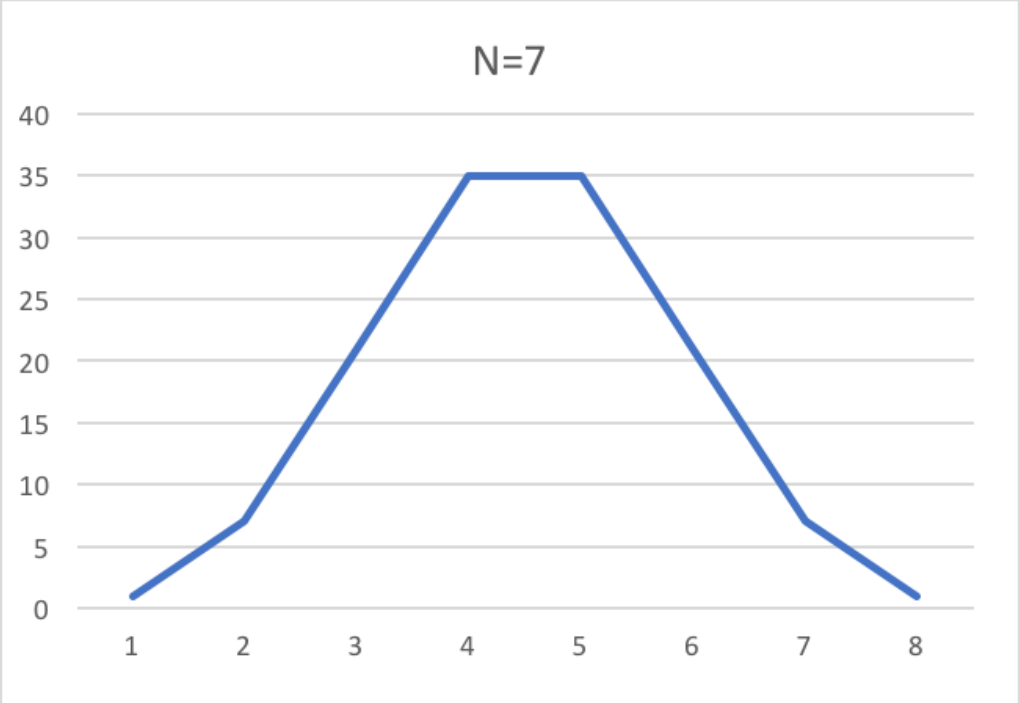
1 6 15 20 15 6 1

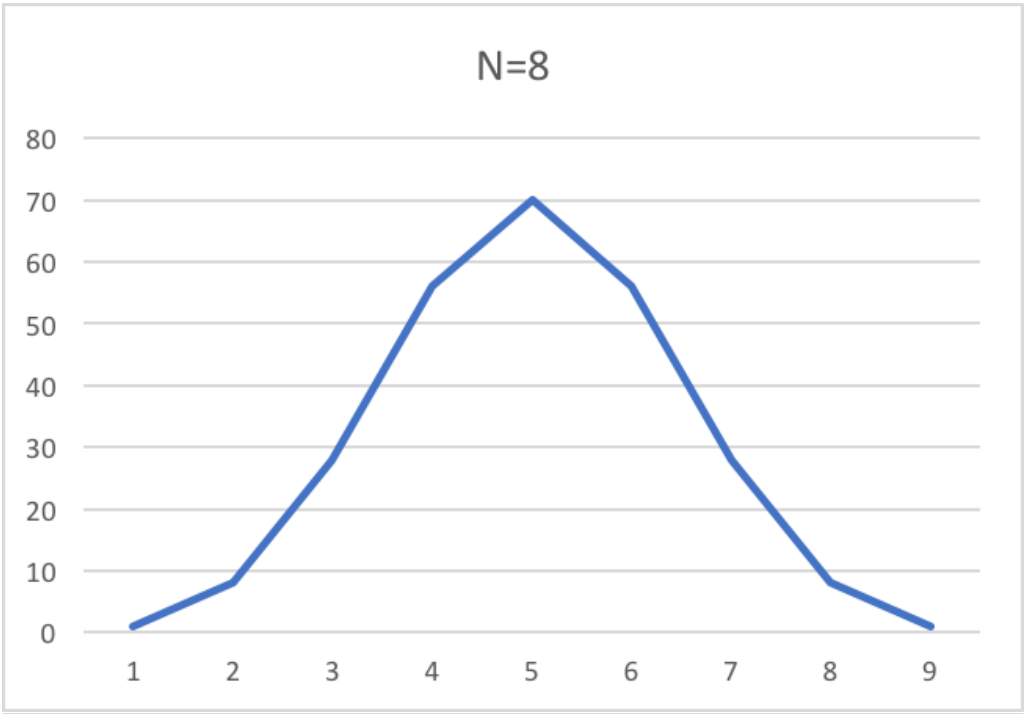
1 7 21 35 35 21 7 1

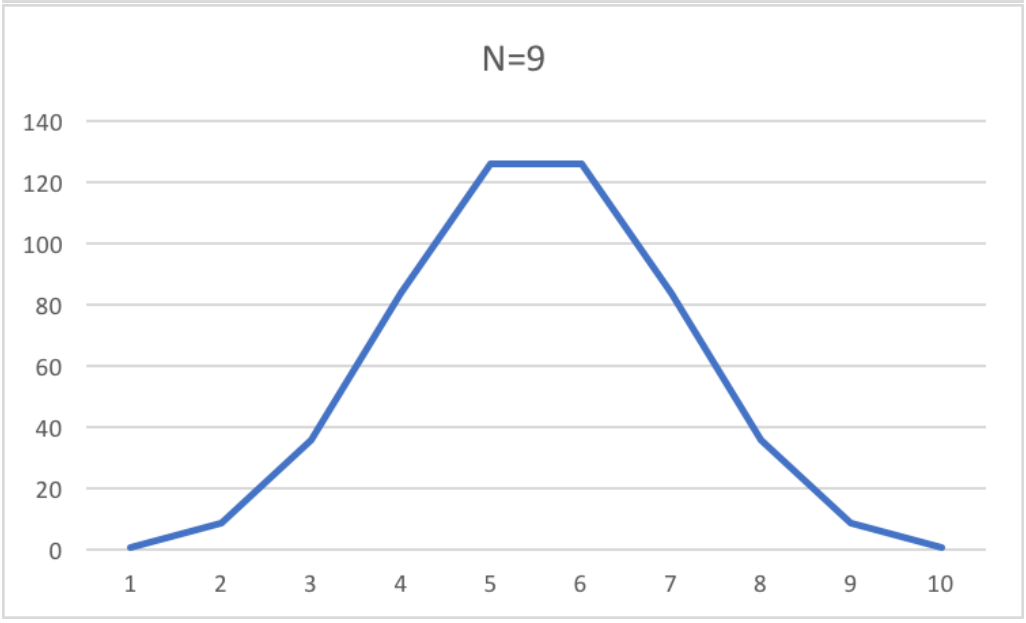
……..

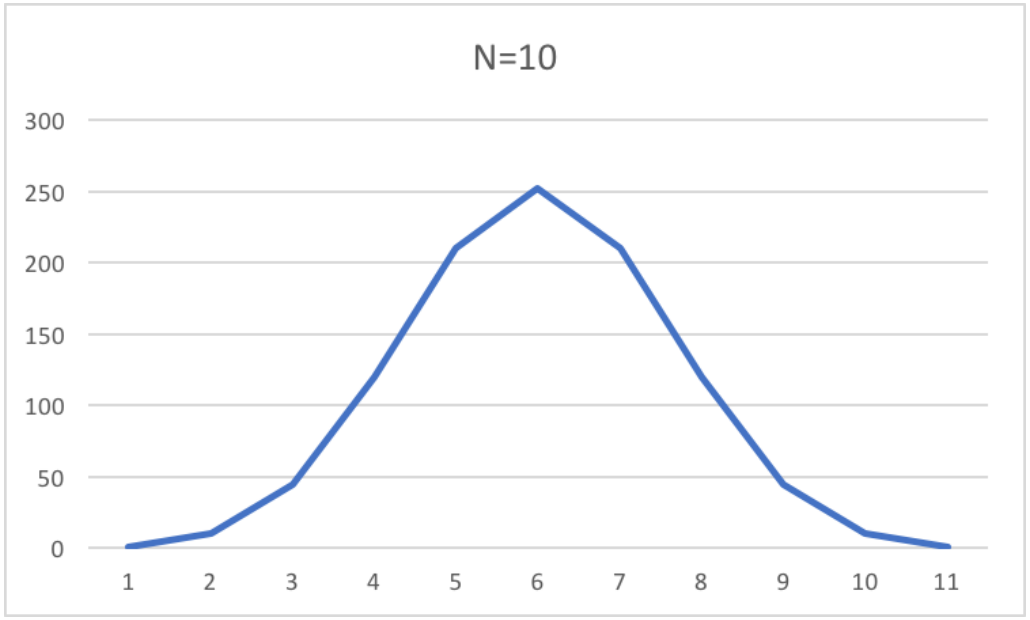
このパスカルの三角形は正規分布を近似的に導出する際にも重要なヒントとなりますので、皆様も必ずノートに書き写した上で、たとえば N=6 や N=10 の場合をエクセル等の表計算ソフトでグラフに描いて見てくださいね。パスカルの三角形は二項展開の各係数を並べたもので、N が十分大きくなると正規分布に近似できることが知られています。

順列の復習ですが、!(階乗) についてまず復習した後、組み合わせについても復習します。

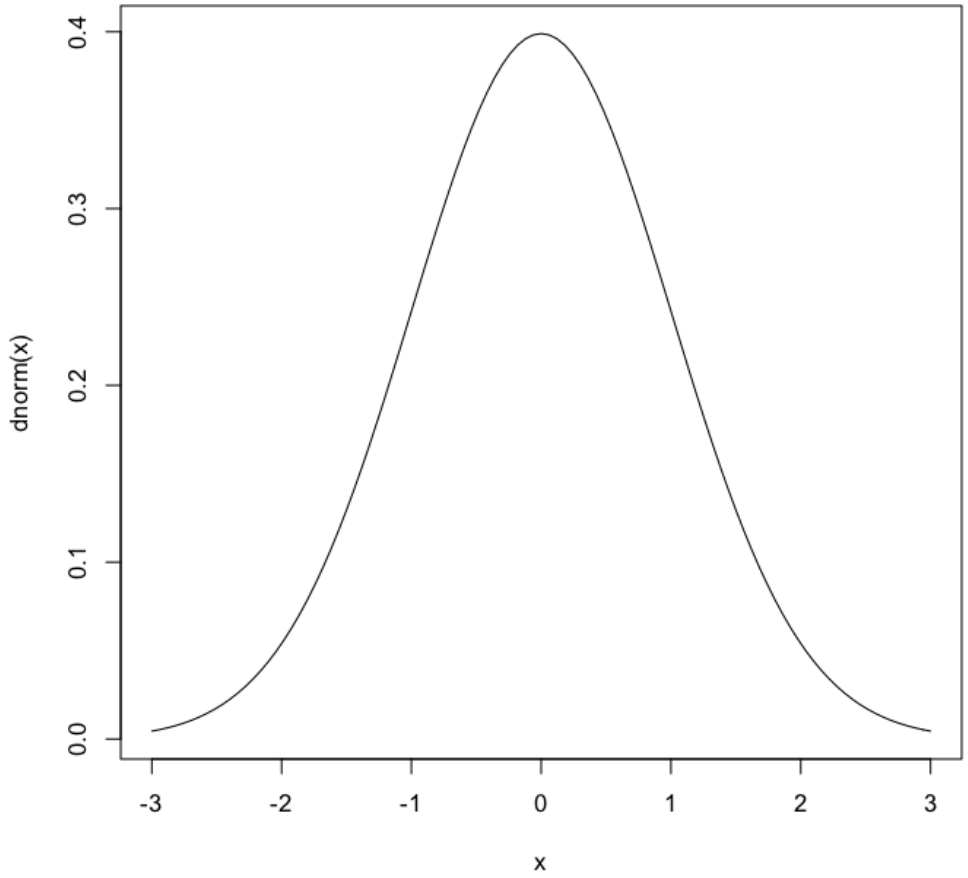



N=8
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9


N=9
140
120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10



.....N $\rightarrow\infty$



確率論の考え方には、先験的アプローチ、経験的アプローチ、主観的アプローチ、公理的アプローチなどがあるとされます。数理的な分析では公理的アプローチが採用されます。公理的アプローチでは、確率は 0 から 1 の間をとる、全ての確率の和は１、お互いに排反な確率の和は個別の確率の和に等しいといったルール、公理から始まり、演繹的に議論を組み立てていくアプローチです。

Ch.4. 二項分布

目標：まず平均と分散、標準偏差を復習した後、二項分布について理解する。

ここでは分布の概念の基礎となる二項分布について簡単に説明します。サイコロを振った時の目の分布や、コイントスの表と裏の組み合わせの分布について説明します。

二項分布の重要な性質。

μ=np

σ^2=npq

ポワソン分布とは、めったにおこならないことを記述した分布で、自由度 m に対して、

μ=m

σ^2=m

ポワソン分布は平均と分散は常に等しく、自由度と一致する分布です。

2日目

Ch.5. 正規分布の性質

ここでは正規分布の性質について、極力数式を使わずに説明いたします。正規分布は統計学における分布の女王と呼ばれる基本的な概念であり、左右対称、分布の山が一つしかない（ 単峰性 ）、分布の左右の裾が広がっているという 3 つの基本的な性質があります。正規分布の平均は μ、標準偏差は σ であらわされます。
標準正規分布の平均は 0, 標準偏差は 1 です。

なぜ正規分布が重要な概念なのでしょうか。

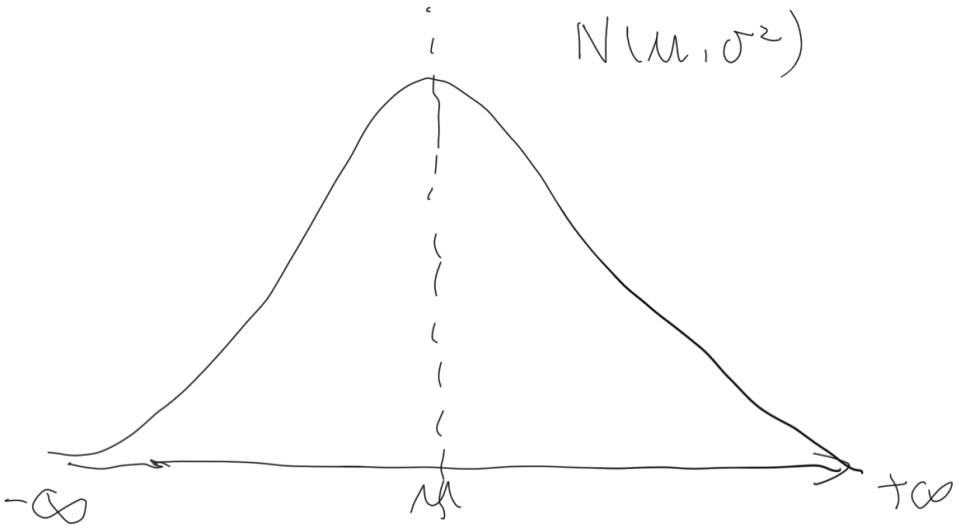
正規分布の再生産性とはなんでしょうか。

正規分布の再生産性とは、お互いに独立な正規分布の間で正規分布 I に正規分布 II を加えたものもの、やはり正規分布するという性質です。

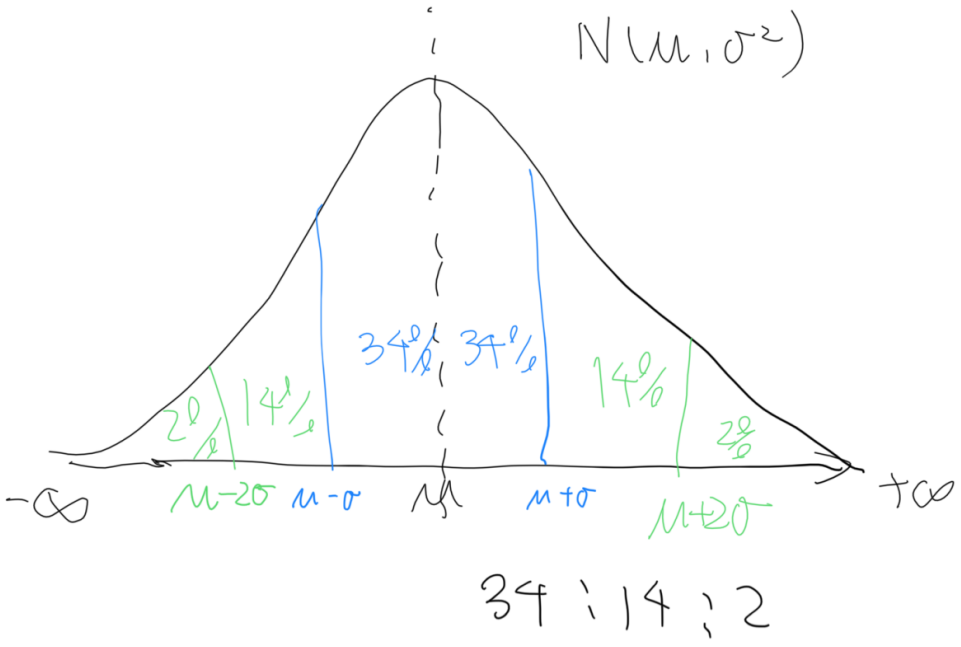
平均 μ は中央にある。平均値と中央値、最頻値が一致するという性質も重要です。

B(n,p) n →∞ → N(np,npq)

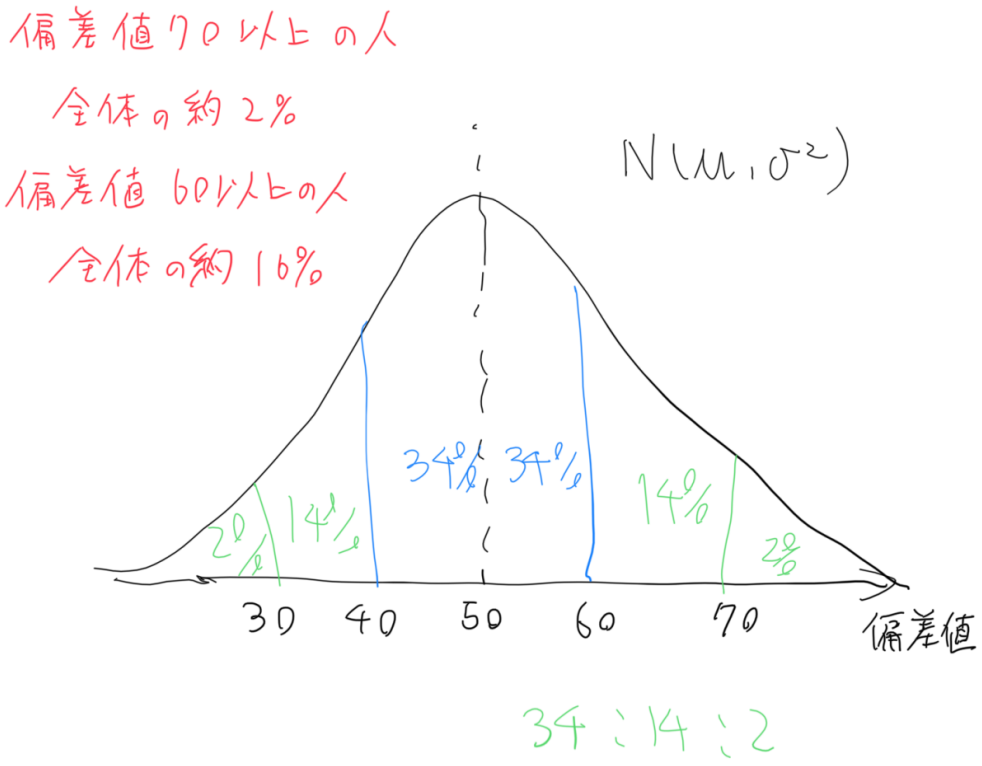
この式は、二項分布で n が十分大きくなると、正規分布に近似できるという考え方をあらわしています。



正規分布 1.png



正規分布の説明 2.png



正規分布の説明 1.png

Ch.6. 正規分布の性質の応用

正規分布がわかると得られる情報。正規分布と平均 (μ) と標準偏差 (σ) の関係。1 σ =68%,2 σ =95%,3 σ =99.7% といった話。

ここでは正規分布の性質をふりかえり、受講者の皆様にも演習の形式で、正規分布の性質について話し合います。

たとえばクラスで 2 つ以上のチームを作った上で、正規分布の性質について議論して代表者にプレゼンテーション (5 分間) をしてもらいます。

富士山と筑波山ではどちらが正規分布の形状に近いと言えるか？ 根拠をあげて説明して見てください。

Ch.7. 大数の法則

標本数 n が十分大きくなると、標本平均 μ が母平均と等しいと見ないして良いという大数（ たいすう ）の法則を説明します。大数の法則は、コイントスを繰り返すと、ほぼ表が出る確率が 1/2 に近くなるという内容を一般化した法則です。

n(試行回数) を増やせば増やすほど、その事象の起こる割合は一定の値 p に近づく。

Ch.8. 中心極限定理

標本数 n が大きくなると、x1+x2+x3+...+xn が正規分布に従うという中心極限定理について説明します。X1,X2,.....,XN が互いに独立で、平均μ、分散σ 2 をもつ同一の分布に従うとする。N を大きくした時、X1+X2+X3+...+XN の分布は正規分布に近く。

二項分布で n が大きくなると正規分布に近づくという話をヒントにします。

大数の法則と中心極限定理の包含関係について説明できるようにしてください。

(大数の法則を一般化したものが、中心極限定理です。)

いかなる分布でもその標本平均 X バーは、標本サイズ N が大きくなるにつれて、平均μ、標準偏差

σ / √ n の正規分布に近づく。

中間テスト

0　（ ）カッコの中身を埋めてください。(復習問題)

平均はデータの（ ）をデータの個数で割ったものである。

分散はデータの（ ）の二乗の平均である

標準偏差は（　）のルートである

データの平均はギリシア文字 (　) であらわされる。

データの標準偏差はギリシア文字 (　) であらわされる。

標準化したデータ Z は各データから（　）を引いたものを、標準偏差で割ったものである。
偏差値とは上の Z に１０を掛け、( ) を足したものである。

偏差値の平均値は ( ) である。

ヒント：　μ σ　総和　偏差　分散　50

１　正規分布の性質について最低３つ以上列挙して説明してください。

　図もあればなおよし。

２　仮にあるクラス (10 人) の小テスト結果が１０点満点で

{1,2,3,5,5,6,6,7,8,9}

であったとします。 \newline

電卓を使うか、手計算で、平均、分散、標準偏差を計算してください。 \newline

2-2

5 人が受けたテスト結果が 100 点満点で

{34,69,70,82,90} \newline

であったとします。電卓を使うか、手計算で、平均、分散、標準偏差を計算してください

3　パスカルの三角形を描いてください。

パスカルの三角形と二項分布の関係について説明してください。

4　あるテストの分布が正規分布で近似できるとすると、偏差値 60 以上の人は全体の何 % でしょうか。同様に、偏差値 70 以上の人 (応用：偏差値 80 以上の人) についても求めてください。
hint: 標準正規分布表を参照しましょう。

5 予習問題です。

t 分布、F 分布、カイ二乗分布について、それぞれ性質を説明してください。

6　予習問題 (発展) です

正規分布、二項分布、ポワソン分布、t 分布、F 分布、カイ二乗分布、以外の分布について調べ、その名前と性質を説明してください。

IV 15.9%,2.3%,(0.1%)

3日目

Ch.9. 検定の概念

ここでは仮説検定について説明します。第 1 種の過誤と第 2 種の過誤について説明します。第 1 種の過誤の例は「能力の有る者に博士号を与えない」、第 2 種の過誤の例は「資質のない者を博士課程に進ませてしまう」があります。

検定力についても説明いたします。

帰無仮説 H0、対立仮説 H1 について説明します。仮説検定の特徴は背理法にあるといえます。

帰無仮説 H0 は間違っていないと証明（ 検証 ）したい仮説のことを意味します。対立仮説は H0 が間違っていないと検証したい時に基準となる対立する仮説です。帰無仮説と対立仮説はお互いに排反（ どちらかしか成り立たない ）の関係にあり、帰無仮説が間違っていないと検証したい場合に、排反な対立仮説が成り立たないことを示すことで、ある確率の元で少なくとも帰無仮説が間違っていない（ 必ず正しいとは限らない ）ことを言うという論法であり、一種の背理法です。

第一種の過誤は、正しいはずの帰無仮説を間違っていると認識することで、逆に第二種の過誤は間違っている対立仮説を正しいと認識してしまうことです。第 1 種の過誤と第二種の過誤の確率を同時に下げることは難しいです。どちらかを立てると、どちらかが立たなくなり、そのような中で正しい結論を導き出すことが重要です。第一種の過誤はあわて者の誤り、第二種の過誤はぼんやり者の誤りと通称されることもあるようです。

	真実	H0が誤り
H0を採択	T	F2
H0を棄却	F1	T

Ch.10.　さまざまな分布（ t 分布、F 分布、カイ二乗分布 ）の使い方

目標：t 分布、F 分布、カイ二乗分布について理解する。時間をかけて講師から内容を説明しますが、適宜参考書を熟読されるなどして、補足的に理解を深めてください。

t 分布、F 分布、カイ二乗分布を用いた検定の仕方についてご説明します。

F 分布は分散の比。カイ二乗分布は母分散についての検定で用います。

t 分布を用いた検定は、回帰分析の出力結果などで頻繁に用いられます。t 値の絶対値が 2 以上であれば、ほぼ有意（ その値はゼロと有意に異なるという意味 ）といえます。母平均μの推定に使います。

カイ二乗分布は、自由度を m とすると、μ＝ m, σ ^2=2m、σ＝√ (2m) となる分布です。

今まで平均、分散という統計量についてお話ししました。また標本分散と不偏分散について説明をします。

平均や標本分散の値、標本分散の比が既知な場合に全体の分布に当てはめて検定する手法について確認します。

これら３つの分布はまずカイ二乗分布から導出されます。

カイ二乗分布の特徴は、自由度 (n) によって分布の形が大きく変化することです。自由度 (n) が大きくなる (n>60) とほぼ正規分布と同じ形になります。

t 分布は左右対称型の分布で、正規分布よりも裾野が厚く、自由度が大きくなる (n>30) と正規分布にほぼ近似されます。

F 分布は m,n という２つの自由度によって形が決まる左右非対称の歪んだ分布です。分散の比の分析に利用されることが多いです。

R コマンド（ 重要！ ）

F 分布を表示する R のコマンドを載せておきます。

curve(df(x,1,1)

自由度 (1,1) の f 分布の密度関数のグラフが書けます。

curve(dt(x, 15),-4,4)

自由度 (15) の t 分布の密度関数のグラフが書けます。

curve(dchisq(x, 1))

自由度 (1) のカイ二乗分布の密度関数のグラフが書けます。

- - - - - -

参考 ３つの分布の表示法（ エクセル関数 ）

t 分布

T.DIST(t, 自由度,T/F)

Q(t)=T.DIST.RT(t, 自由度 ）

t α=T.INV(1-α, 自由度 ）

F 分布

F.DIST(x, 自由度 1, 自由度 2,T/F)

F.DIST.RT(x, 自由度 1, 自由度 2)

F α=D.INV.RT(α, 自由度 1, 自由度 2)

カイ二乗分布

CHISQ.DIST(x、自由度、T/F)

Q(x)=CHISQ.DIST.RT(x, 自由度)

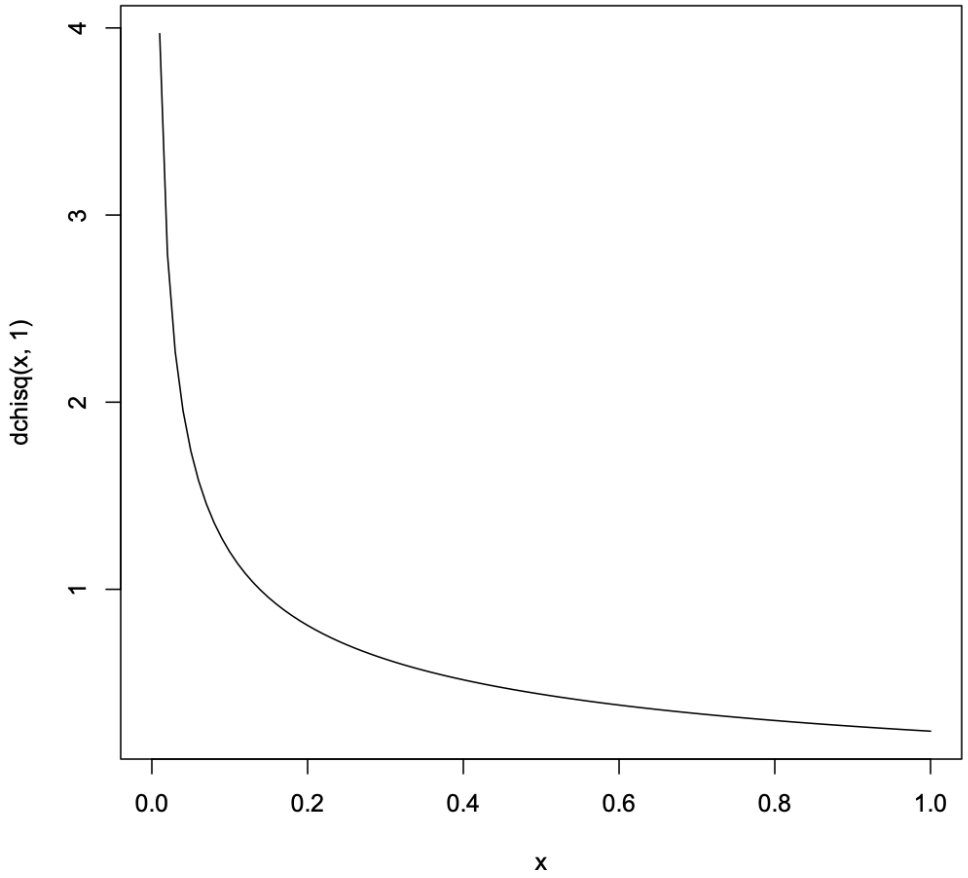
カイ二乗α=CHISQ.INV.RT(α、自由度)

参考正規分布

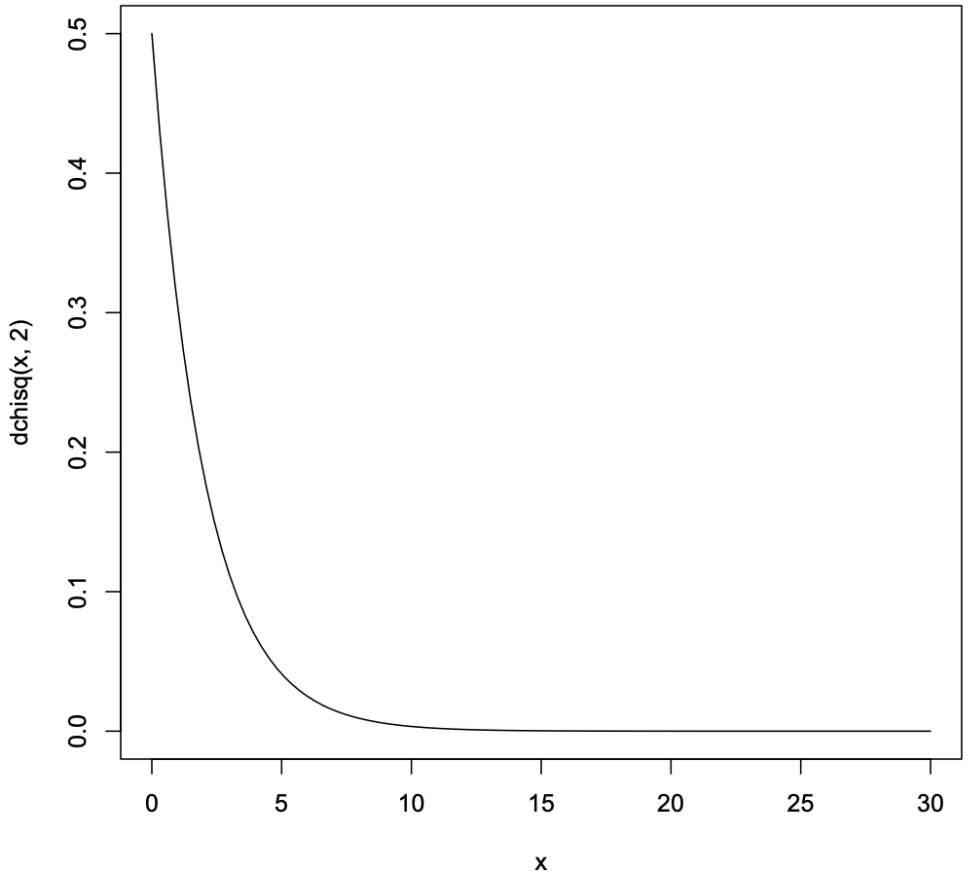
NORM.S.DIST(X,T/F)

参考文献 盛山和夫 (2015)『統計学入門』ちくま学芸文庫

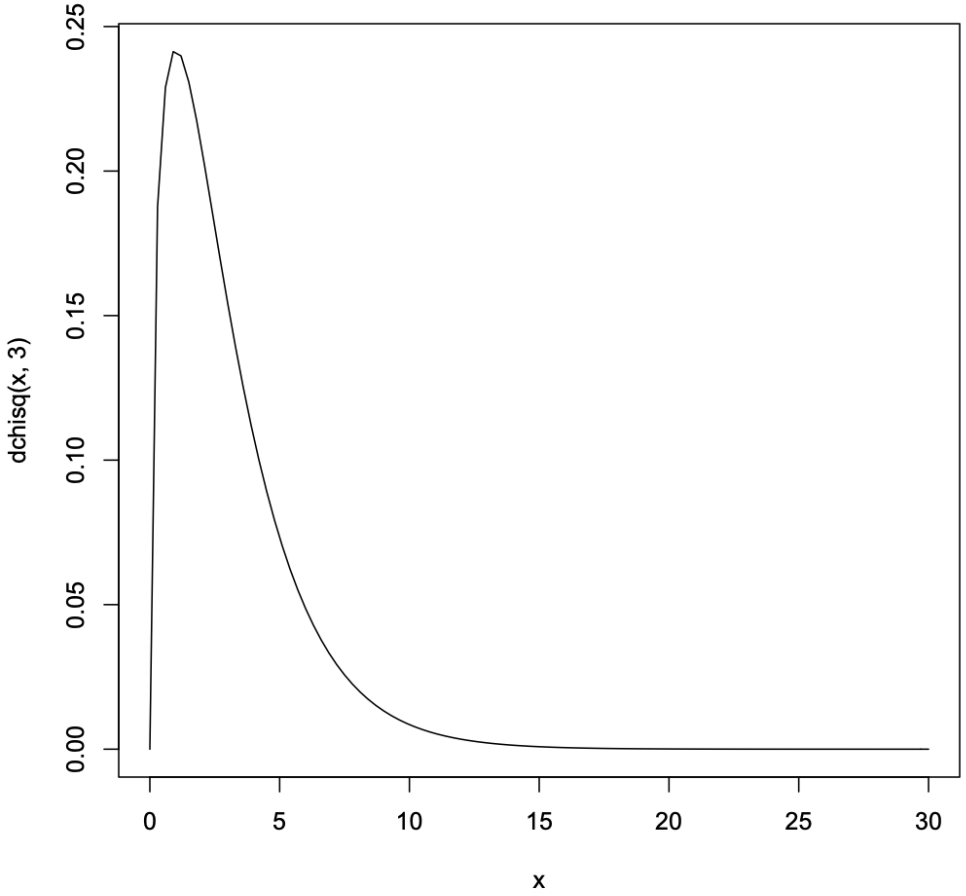
カイ二乗分布の例



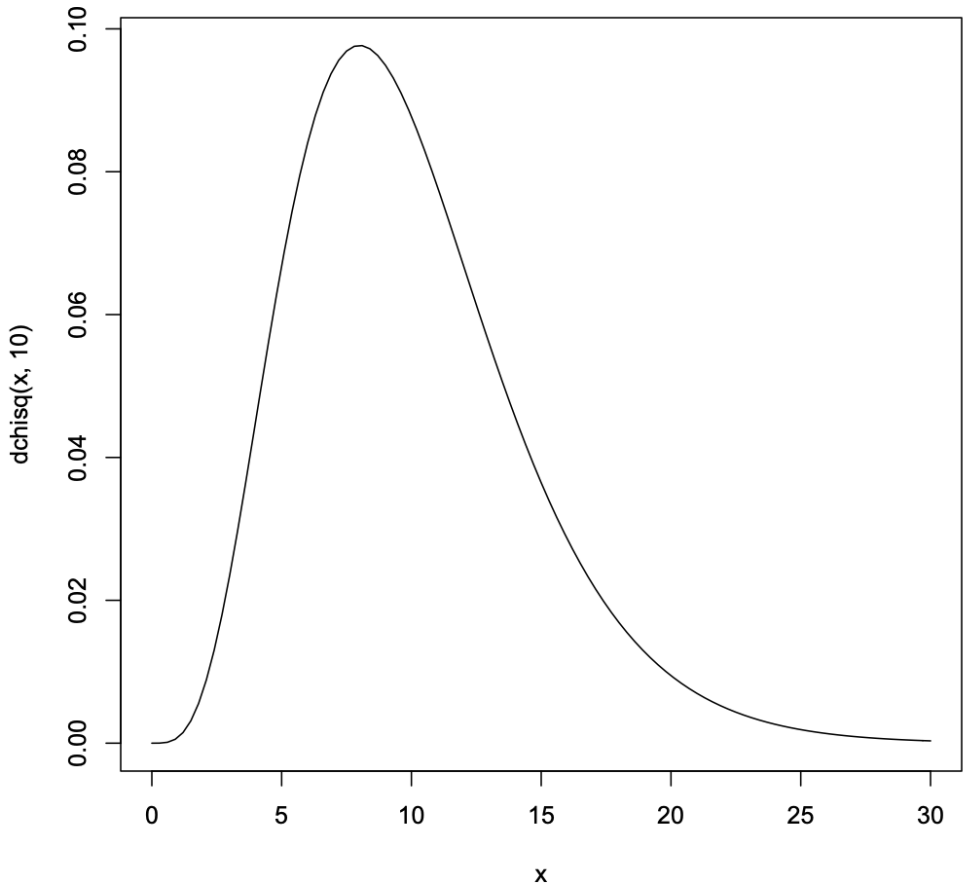
chisquare1.png



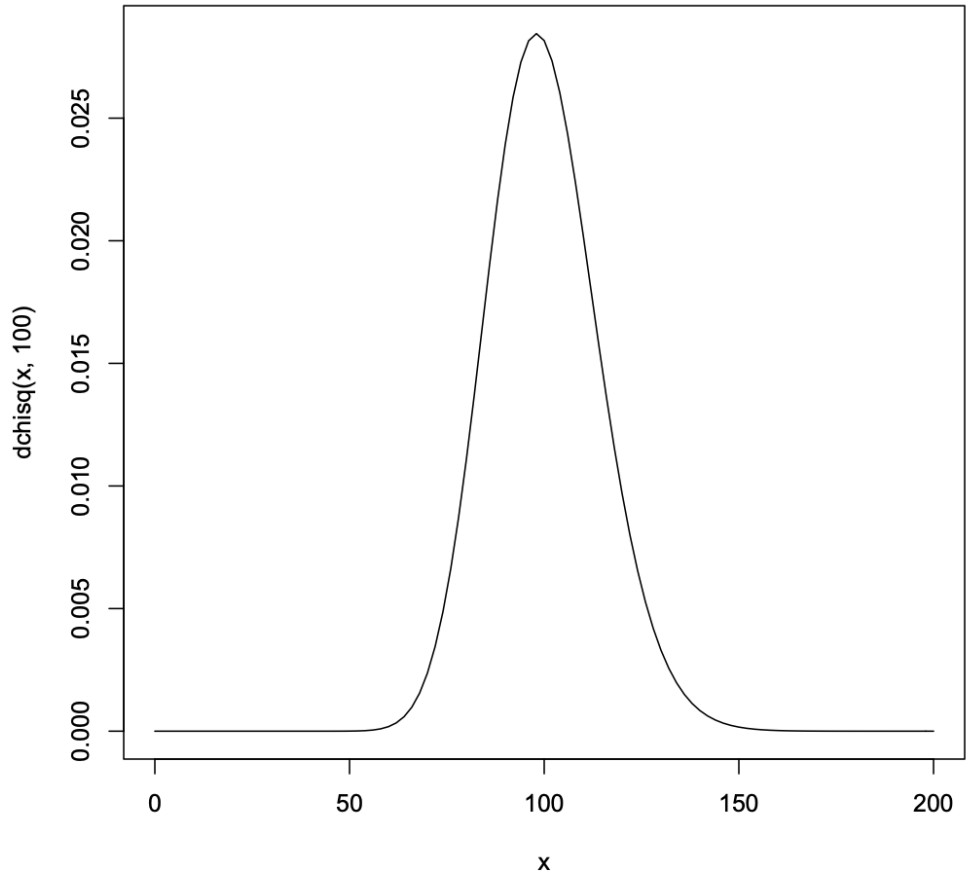
chisquare2.png



chisquare3.png

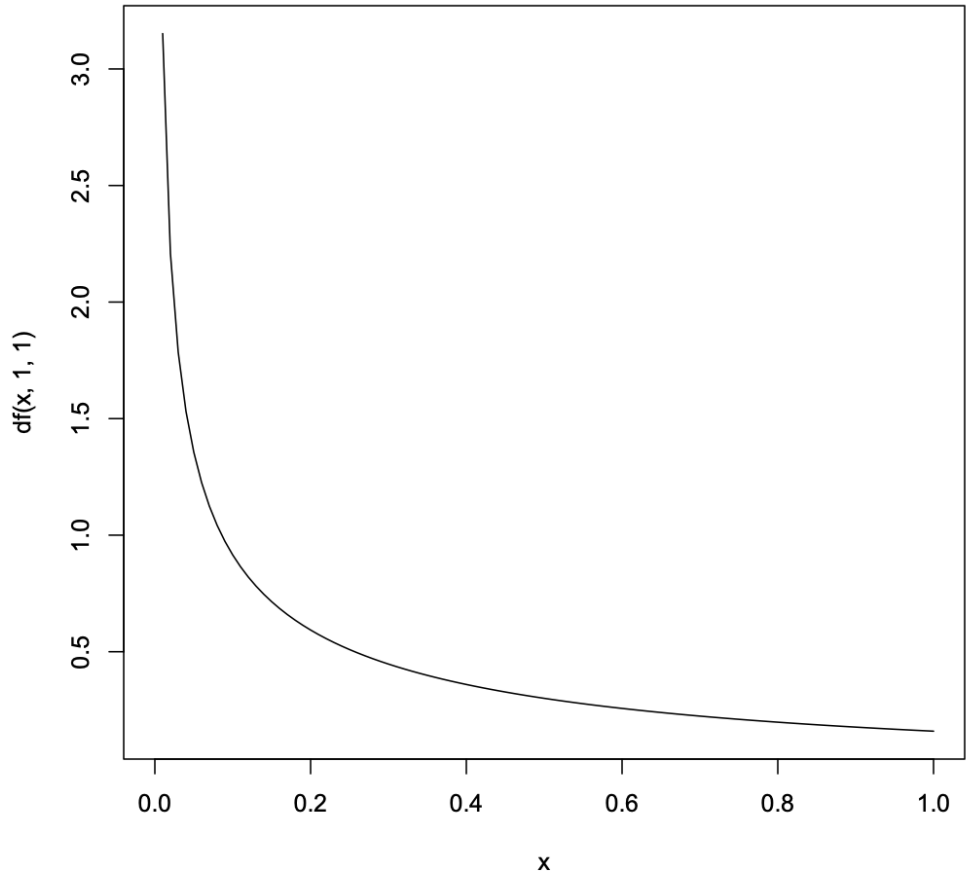


chisquare10.png

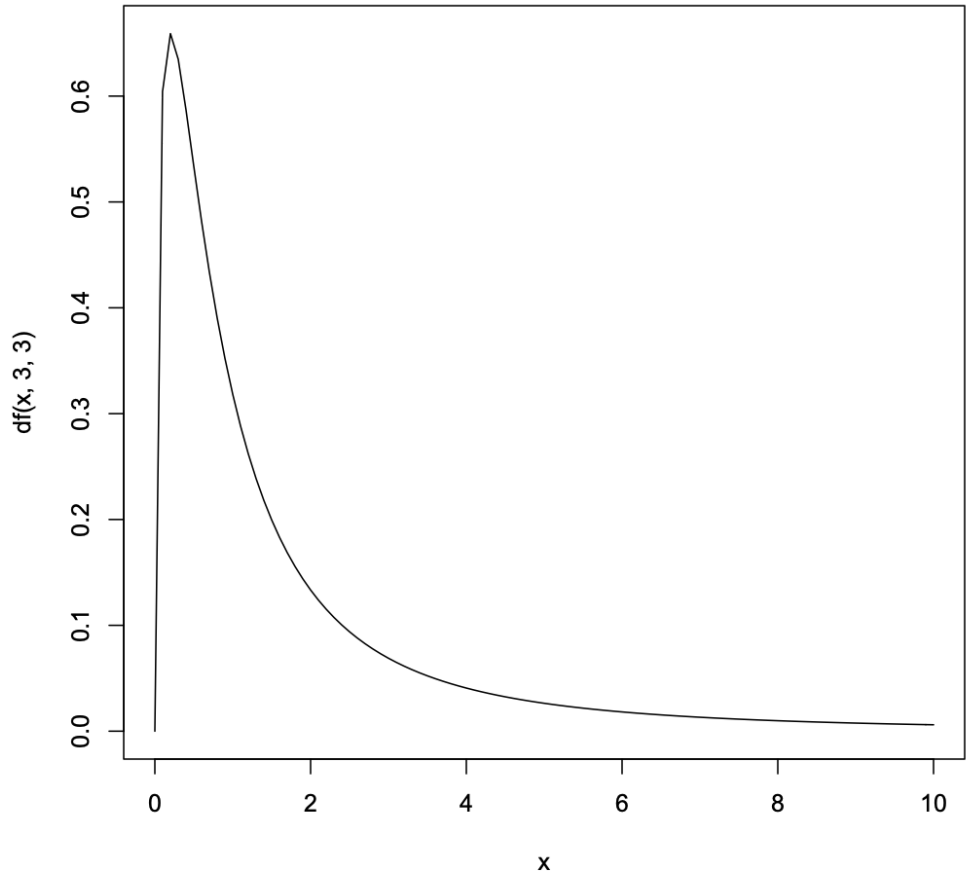


chisquare100.png

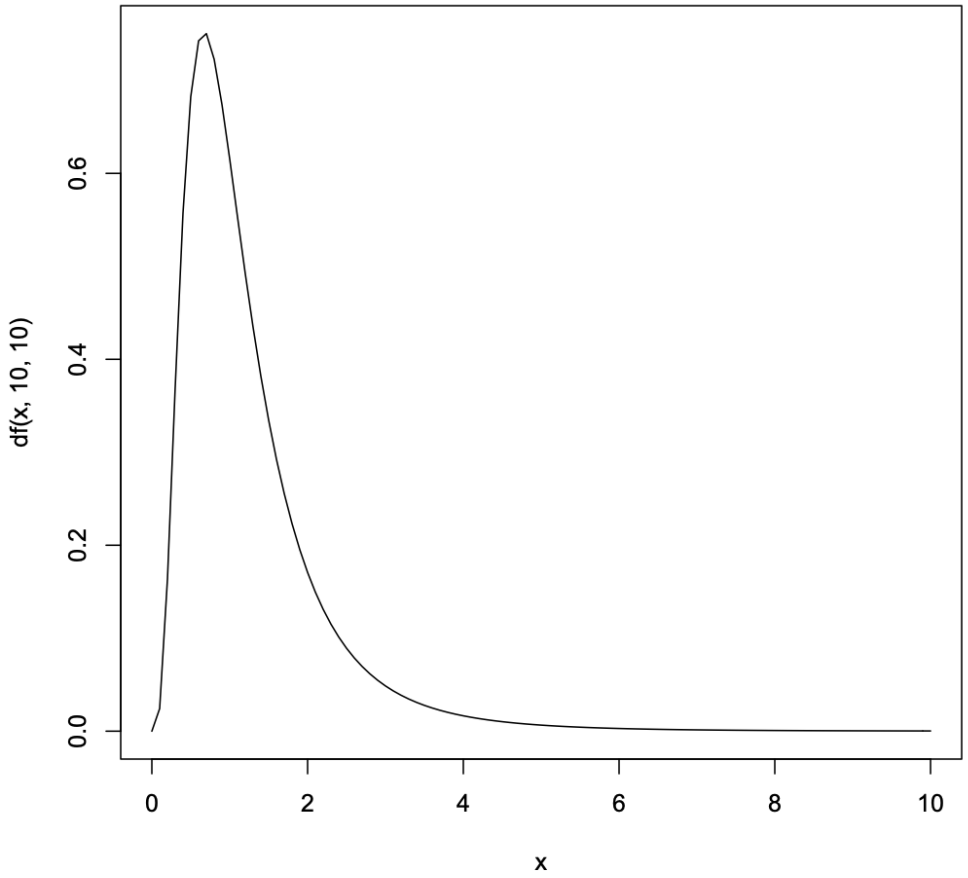
F 分布の例



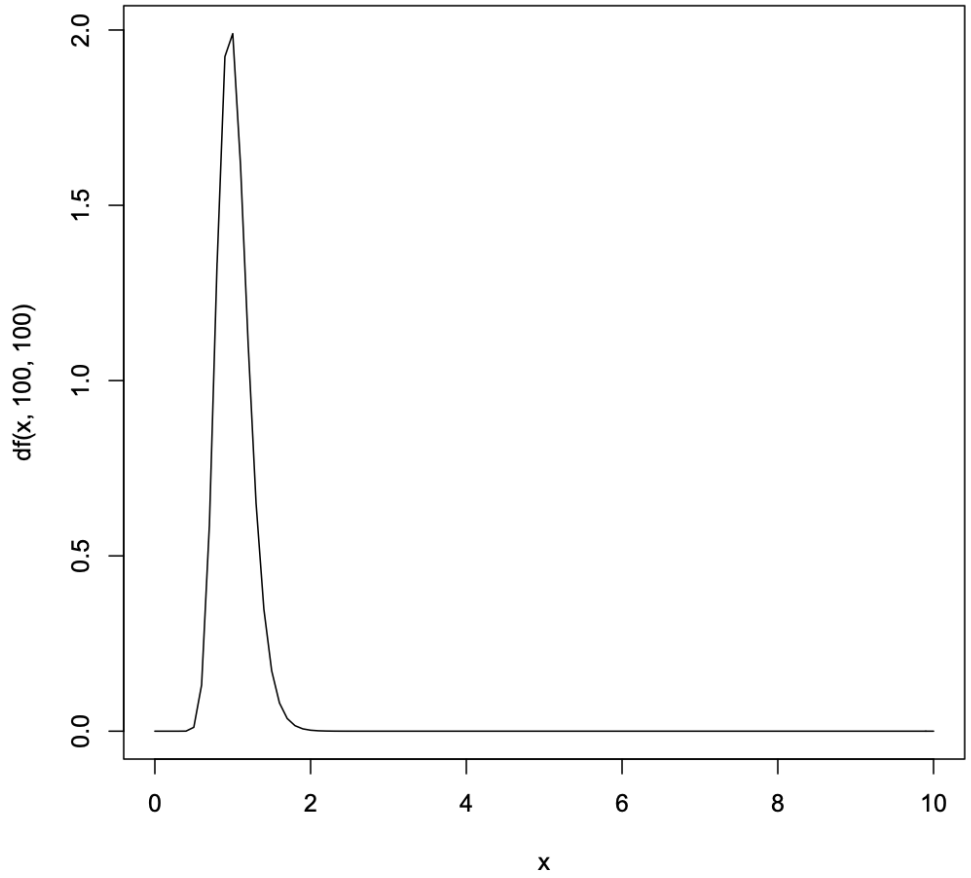
f＼__1＼__1.png



f ╲__ 3 ╲__ 3.png

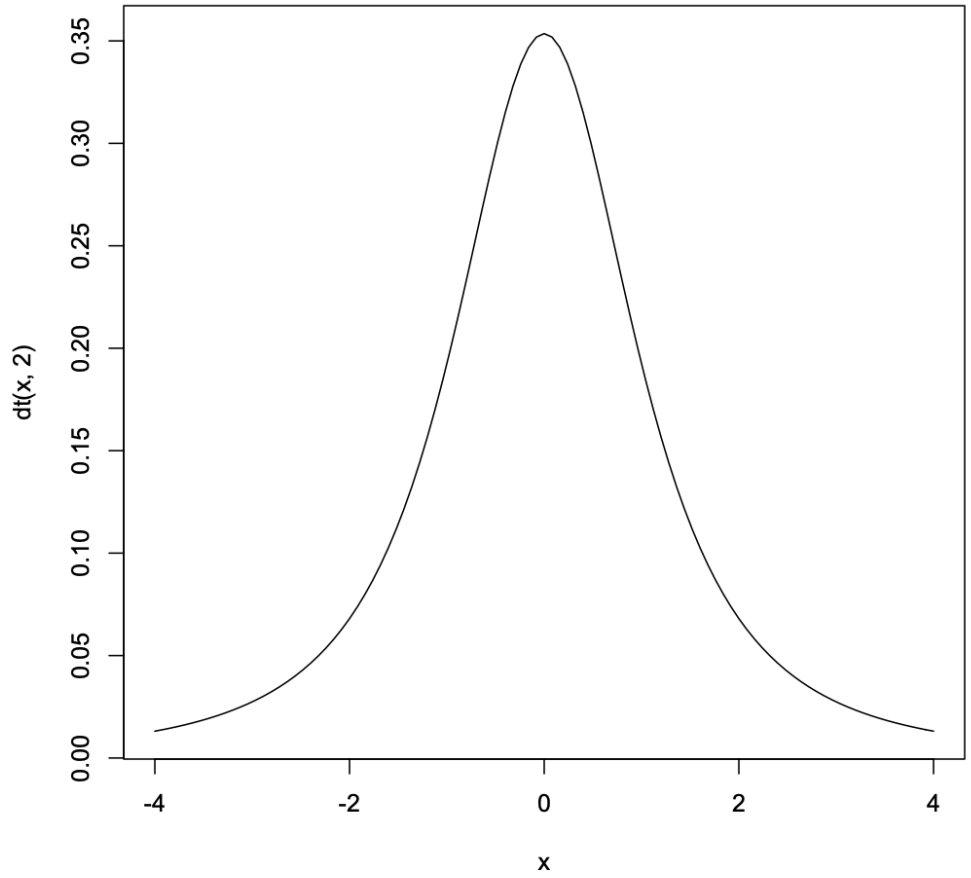


f ╲__ 10 ╲__ 10.png

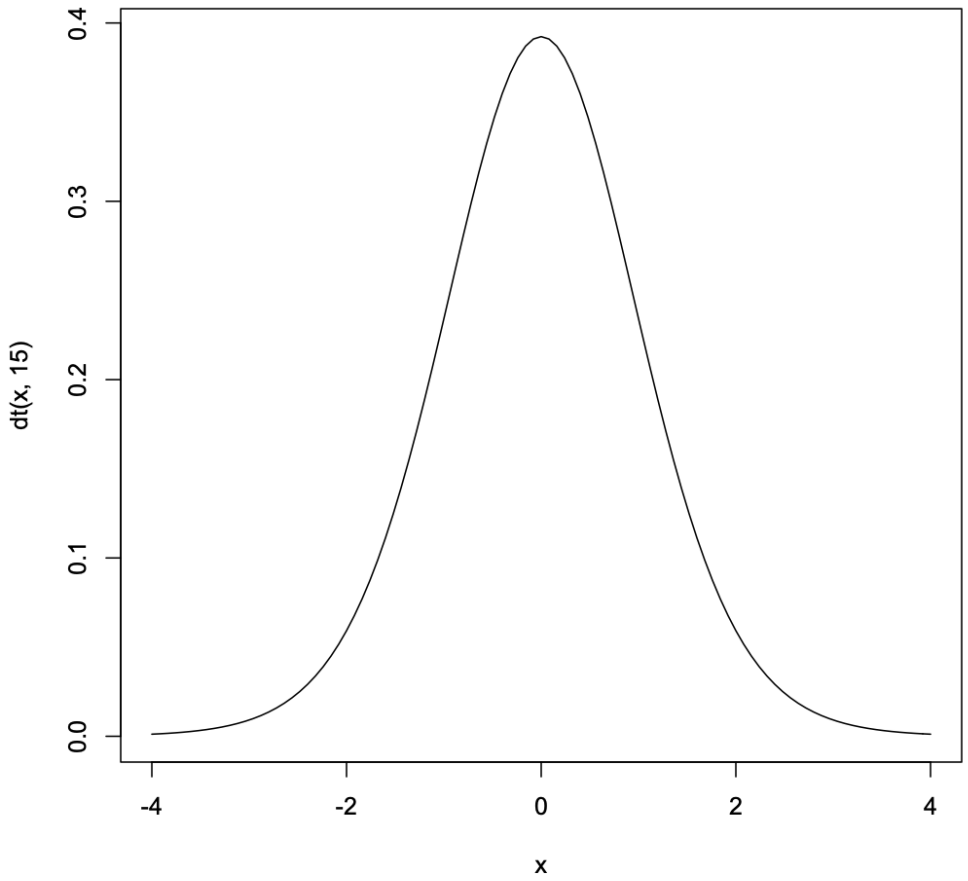


f ＼__ 100 ＼__ 100.png

t 分布の例



t2.png



t15.png

Ch.11. 検定の応用

ここではグループディスカッション、演習を交えつつ、検定の応用例について議論します。どのようなケースにどのような分布の検定を利用するのが望ましいか、議論します。

まずは基本となる平均、分散、標準偏差の概念を復習します。

その後、正規分布について復習します。

正規分布について復習した後、t 分布、F 分布、カイ二乗分布について分布をグラフで眺めるなどして、復習します。

これらの分布の使い道について整理します。

Ch.12. 推定の概念

ここでは区間推定 (estimation) の概念について説明します。テレビの視聴率は何 %、選挙の際の当選確率は何 % といった時の推定の話題を説明します。

点推定 X バーを μ の推定値として表示して、X バーには ś$Z \sigma / \sqrt{n}$ の誤差があることを表示する。

ここは例題を使って、２題くらい説明して、あとで皆さんにも演習問題を解いていただき、理解を深めてもらう予定です。

4日目

Ch.13. 多変数の統計

ここではクロス集計表と多変数の統計について説明します。クロス集計表は、元のデータ表から複数のコラムを抽出して作成する二次元の表です。クロス集計表は分割表とも呼ばれます。１変数だけではなく、多変数の統計、多変量解析の初歩についても説明いたします。

4 日目の内容は多変数の統計分析です。共分散、相関係数と回帰分析について理解します。共分散の計算の仕方がわかれば、相関係数と回帰分析もわかるようになるので、例題を使って、共分散を計算してみます。

パネルデータと時系列データについても説明する予定です。

Ch.14. 相関係数

相関係数と共分散（ ２変数間の分散 ）について説明します。相関係数は-1 から１までの範囲をとる係数で、２つの変数の共分散を各変数の標準偏差の積で割ったものです。相関係数が例えば１に近いときに、本当に正の相関があると言えるかは、実際にグラフをプロットして確認する必要があります。
順位相関係数もあります。

相関係数は高校の時にならったコサインと同じ概念です。そんなに難しくないので、じっくり理解しましょう。

例題：五人のグループの数学のテストの点数

{90,88,70,64,55}

と物理のテストの点数

{98,70,66,50,48}

の場合の数学と物理のテストの点数の共分散と相関係数を計算してみてください。

X と Y の共分散

$$COV(X,Y) = \sigma_{XY} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

ただし

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

X と Y の相関係数

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

ただし

$$\sigma_X = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

$$\sigma_Y = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n}}$$

X と Y の共分散.png

$$b = \frac{s_{xy}}{s_{xx}} = \frac{\sum_i (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

相関係数

$$\rho_{xy} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

$$-1 \leq \rho_{xy} \leq 1$$

相関係数の説明.png

Ch.15. 回帰分析

回帰分析について説明します。

y=bx+a

y: 従属変数、被説明変数

x: 独立変数、説明変数

最小二自乗法とは誤差の二乗を最小化する手法です。

重回帰分析とは X がベクトルで与えられる場合をいいます。

決定係数 R2 とは y の分散の内、X によって説明される部分の割合を示します。

回帰分析では、データをあてはまりのよい直線に近似する手法です。「あてはまりのよい直線」は、データの平均値を必ず通ります。

回帰分析の説明

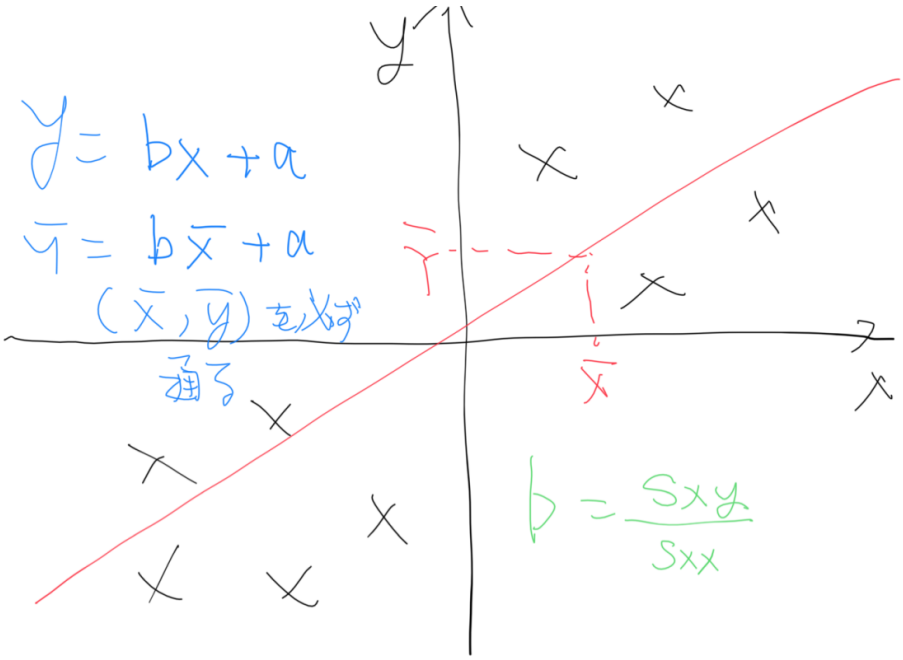
$$y = bx + a$$
$$\bar{y} = b\bar{x} + a$$

点$(\bar{x},\bar{y})$を必ず通る。ここで、$\bar{x}$はxの平均を表します。

$$b = \frac{S_{xy}}{S_{xx}}$$
$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

回帰直線の傾きは,XとYの共分散をXの分散で割ったもの

回帰分析の説明.png



回帰分析の説明.png

演習問題

今回は最終回ですね。今まで四日間お疲れ様でした。テストはやさしい統計学講義での皆様にお願いしたい到達点、理解の復習という位置付けです。

Q1. 正規分布の性質について箇条書きで説明してください。(ポイントは最低３つです。)

Q2. 相関係数の概念を使って、英語が文科系（ 国語、社会 ）の科目であるか、理数系の科目であるか説明できるでしょうか？

Q3-1.10000 人に向けてテレビの視聴率調査をしました。ある番組を見た人は 2500 人でした。この番組の視聴率を信頼係数 95% で推定してください。

Q4. 第 1 種の過誤、第 2 種の過誤について、事例をあげて説明してください。

Q5. 相関係数を用いた分析を例にとって、相関関係と因果関係の違いについて説明してください。

ご健闘をお祈りしております。

Appendix

受けよう統計検定

　本書の読者、やさしい統計学講義の受講者にオススメしたい資格があります。

統計検定です。

本書のレベルであれば、３級から。

他の本や講義も受講する方であれば、２級。

エキスパートを目指す方であれば、準１級。

プロを目指す方であれば、１級。

を受験してください。

　統計学を修めたデータサイエンティストを社会では需要しています。

　データサイエンティストは第四次産業革命の進展する現代で最も魅力的な職業の一つにカウントされます。

　皆さまもぜひ統計検定を受験して、データサイエンティストとしての素養、腕を磨いてください。
本書はあくまで入門書、スタートであり、必要最小限度の内容です。

　今後も統計学を学習・勉強され、統計検定２級以上を目指されることをオススメいたします。

Appendix. アドバンスト統計解析

やさしい統計学講義のおまけとして、上級の統計解析や社会での統計解析の使われ方についてトピック単位で説明します。

1 社会における統計解析の使われ方

統計学は記述統計学と推測統計学に大きく分かれます。国の人口や GDP を調査によってまとめる仕事は記述統計学、逆に既にある統計調査の結果から、アンケート調査の推計結果の妥当性や信頼性を判断するのは推測統計学です。

社会において統計学のスキルが重要とされるのは、基本的に推測統計学の要素であると言えるでしょう。

2 機械学習における統計分析のスキルの重要性

機械学習（ 大量のデータを使って計算機に学習させ、分析結果を掃き出させること ）において、分析結果を人間が解釈するのに必要なのが統計学の基礎知識です。

たとえば CHAID というアルゴリズムはカイ二乗統計量を決定木が分岐する際の統計量として用います。一般に決定木というアルゴリズムでは、木の分岐をジニ係数やカイ二乗統計量、AIC などの統計量で判断しますが、この結果を理解するのには統計学の知識が必須です。

また機械学習の分析結果を理解するにあたっても F 値という統計量の知識が必須です。

３ 経済物理における統計分析の重要性

経済データを物理学の理論を使って分析する経済物理・社会物理では、ベキ分布やパレート分布など、統計学の基礎知識が必須です。

４ ネットワーク解析における統計分析の重要性

Fb のつながりや、産業構造、疾病構造などを分析するネットワーク解析においても、統計分析のスキルは重要です。隣接行列というものをつくって、ネットワークグラフを記述するネットワーク解析では、ページランク、次数（ つながりの数 ）、といった概念に加えて、豊かな分析をするには統計分析の知識が必須です。

5 モデルの評価

統計モデルがどれくらいフィットしているか比較評価するためには情報量基準という概念が重要です。赤池情報量基準、ベイズ情報量基準、カルバック＝ライブラ情報量基準などの概念が有名です。これらの情報量基準は統計モデルの妥当性を比較検討する際に、有効な情報を提供します。

6 統計学とプログラミング

現代社会における統計分析に役立つプログラミング言語としては python と R が挙げられると思います。python は機械学習や AI の分析にも役立つ汎用的な言語で、R は統計分析に特化した統計言語です。やさしい統計学講義では、アドバンスト統計解析、つまり上級向けの統計解析に R を使用することをお勧めします。確かにエクセルでも相関係数や回帰分析を行うことができますが、解析結果の妥当性を検証する検定や推定にはあまりエクセルは向いているとは言えません。また数値計算の妥当性（ どこまでの有効数

字を扱っているか ）という点で、エクセルはたとえば医学分析など厳密な検証が必要な分野では使われていません。

では代わりに医学系などの分析では STATA という統計分析パッケージが使われています。しかし、STATA や SPSS,E-views などの統計解析ソフトは有料で、非常に高い単価です。確かにこれらの統計パッケージは学生の内にアカデミック版で入手して使い方を覚えるということも考えられますが、もし、研究者を目指すのであれば R、理系研究者を目指すのであれば、python もマスターしたいところです。

R でも H20 というパッケージを用いることで深層学習を試すことができます。python はもっと機械学習のアルゴリズムを試すのに向いています。

python と R という二つのプログラミング言語を挙げましたが、皆様が社会に出てから活躍されて行く上で、新しいプログラミング言語や主流となるプログラミング言語は、どんどん変化していきます。そういった中で、python や R を利用してプログラミングの考え方、統計解析プログラミングの概念を掴むという姿勢が重要であると思います。

7 やってみようプログラミング

筆者は Xojo (VB と互換性のあるクロスプラットフォームの basic 言語 ）で相関係数や回帰分析のプログラム、関数電卓等を記述した経験があります。皆様も得意な言語で簡単な解析プログラムをハンドメイドで作ってみてはいかがでしょうか？ R も個人の貢献が重なって作られた一種の統計プラットフォームだと思います。

8 ビッグデータ解析

ビッグデータ解析では大量のデータを処理することが求められます。できるだけクラウド環境を利用して、クラウド上で R,python を利用するといった姿勢が求められます。

9 やっとけばよかった統計学

筆者は数量経済分析という計量経済学の授業しか統計学を履修したことがありません。しかし、社会に出るにあたって、統計課や計量分析室といった統計学や計量経済分析のスキルが求められる部署に所属することを余儀なくされました。数学や統計学といったスキルは、社会に出てからもさまざまな場面で要求されるものです。やり直し学習（ リカレント学習 ）の必要性があります。もちろん、社会に出てから必要最小限なところだけ学習して切り抜ければよいという考え方もあります。

しかし世の中で AI や機械学習、ビッグデータ解析といった概念が重要視されるにつれて、数理的思考法や統計リテラシーを身に付けることは重要性が増しており、学生時代に数理的思考法や統計リテラシーについて数学、数理科学の講義や、統計学の講義を履修して学習しておくことは重要であるといえるでしょう。

もちろん社会に出てから自習したり、社会人大学院で学習したり、場合によっては OJT や研修で統計リテラシーを身につけることも可能でしょうが、まずは時間と余裕とエネルギーがある大学時代に統計リテラシーのコアを把握することで、応用が利く豊かな思考ができるようになりたいものです。

10 さわってみよう統計

SNA、国勢調査、産業連関表といった統計はインターネットで公開されています。総務省の統計サイトや地方自治体の統計のページから統計表をダウンロードして分析してみましょう。

加工統計に分類されると思いますが、RESAS も面白い情報を提供してくれます。

AppendixII.R の使い方

はじめに

なぜ R を使うか？

SPSS,STATA,E-Views,MATLAB などの統計パッケージ、数値計算パッケージと比べると,R は無料です。GUI ベースの SPSS と比べると R は使い勝手が悪く、敷居が高いようですが、R は柔軟で新機能の実装が早いです。Python は機械学習や人工知能の分野と相性がよいですが、R の方が統計言語（ 統計解析専用のプログラミング言語 ）としてはとっつきやすいです。R はエクセルより統計解析の機能が実装されており、専門的です。この章では、R のパッケージ、MASS、igraph、svs、h2o の使い方を概観します。

MASS は多変量解析の一種である対応分析、igraph はネットワーク分析、svs は PLSA、h2o は深層学習のパッケージです。

MASS で行う対応分析はマーケティングなどで使われる多変量解析手法の一種です。igraph は SNS の可視化に使われるネットワーク分析に対応したパッケージです。svs はクラスタリングの手法、plsa（ 確率的潜在的意味解析 ）という POS データを分析するのに向いた手法が試せるパッケージです。h2o は深層学習が試せるパッケージです。

四則演算・基本演算

+-＼ */ 四則演算

^ ベキ

%/% 整数商

%% 剰余

sqrt(2) ルート 2

基本数学関数

sqrt ルート

abs 絶対値

exp 2.81828...

log ログ

log10

log2

sin

cos

tan

CSV の読み書き

data=read.csv(“abc.csv”,header=TRUE)

write.csv(data, “abc.csv”,guote=FALSE)

＼* CSV の読み書きはデータロード、データセーブの基本です。

回帰分析

result=lm(y~x,data=data1)

summary(result)

回帰分析（ データ例 ）

str(iris)

result<-lm(Petal.Length~Sepal.Length,data=iris)

summary(result)

＼ *iris はアヤメのサンプルデータです。

＼* パッケージのインストールコマンド（ 近くのサーバーからパッケージをダウンロードします ）

install.packages(“MASS”)

MASS →対応分析

library(MASS)

result=corresp(data,nf=2)

biplot(result)

MASS →対応分析 (データ例)

library(MASS)

caith

result=corresp(caith,nf=2)

biplot(result)

igraph →ネットワーク解析

library(igraph)

data1=read.csv(“abc.csv”,header=TRUE)

x=as.matrix(data1)

g=graph.adjacency(x,mode= “undirected”)

g=simplify(g)

plot(g)

＼*ネットワーク解析では、隣接行列という概念を使って、グラフを描画します。

svs → PLSA

library(svs)

h=fast __ plsa(data1,r,symmetric=FALSE,tol=1e-8)

write.csv(h $ prob0, “plsa __ test0 __ result0.csv”,quote=FALSE)

write.csv(h $ prob1, “plsa __ test0 __ result1.csv”,quote=FALSE)

write.csv(h $ prob2, “plsa __ test0 __ result2.csv”,quote=FALSE)

＼* plsa は大規模 POS データの分析に使われる手法です。

h2o →深層学習（ サンプルデータ ）

library(h2o)

localH2O <- h2o.init(ip= “localhost”,port=54321, startH2O=TRUE,nthreads=1)

irisPath=system.file(“exdata”, “iris.csv”,package= “h2o”)

iris.hex=h2o.importFile(localH2O,path=irisPath)

h2o.deeplearning=(x=1:4,y=5,data=iris.hex,activaton= “Tanh”)

h2o.shutdown(localH2O)

＼*h2o は深層学習のパッケージで、ローカルの仮想サーバを立てて、深層学習を行います。深層学習は機械学習の一手法で、大量のデータを必要としますが、比較的高い精度の学習結果が得られる手法とされています。

ここに挙げたRのコードは筆者がこれまで研究やビッグデータ解析で使ったコードの一部です。まだまだ応用の余地が多分にあります。皆様もRを使って実りある解析結果が得られることを期待して筆をおきたいです。

AppendixIII. 数学付録：パレートからケインズへ

数学付録　パレートからケインズへ

パレート(1982)『初期応用経済学講義集』およびケインズ(1921)『確率論』から、統計学に必要な数式展開を引用、まとめておく。

1　パレート(1982)『初期応用経済学講義集』第 3 章「統計データの利用法」pp.27-34.より

（標準）正規分布は二項分布（ベルヌーイ分布）から n を十分大きくした時の極限として利用できるというアイデアを利用する

$$\mu = m + n$$

$\frac{1.2.3....n}{1.2...m\,1.2...n} p^m q^n, \frac{m}{n} = p, \frac{n}{m} = q$（Pareto(1982),p.29 より）

上式を書き換えると

$$\frac{n!}{m!n!} p^m q^n$$

$\frac{1}{\sqrt{2\pi\mu pq}} e^{-\frac{h^2}{2\mu pq}}, \frac{m}{\mu} = p$（Pareto(1982),p.30 より）

$P_\alpha = \frac{2}{\sqrt{\pi}} \int_0^\beta e^{-t^2}, \beta = \frac{\alpha}{\sqrt{2\mu pq}}$（Pareto(1982),p.30 より）

$K = \sqrt{2\mu pq},\ K^2 = \frac{\theta+2}{2s_2}$（Pareto(1982),p.30 より）

K:精度、K^2:荷重（ウェート）という

最終的に

$p = \frac{K}{\sqrt{\pi}} e^{-k^2 z^2}$ (Pareto(1982),p.33 より)

パレートの 19 世紀末のローザンヌ大学での応用経済学講義のノートによると上記のような内容でベルヌーイ分布から（標準）正規分布を導出する方法が当時講義されていたことがわかる。パレートは統計学を経済学に応用したという意味で計量経済学の先駆者であった。

2 ケインズ(1921)『確率論』第５部「統計的推定」pp357-468.より

ケインズ『確率論』は蓋然性の哲学と呼ばれる主観的確率論の書物とされるが「統計的推定」の部に関してはパレートの「統計データの利用法」と一部重なる内容を議論している。

ケインズ(1921)『確率論』Ch.29.「先駆的確率論の統計的頻度予測のための利用法：ベルヌーイ、ポワソン、チェビシェフの定理」ではベルヌーイ＝ラプラスの定理について議論がある。

$\frac{m!}{(mq-h)!(mq+h)!}p^n q^{m-n}$ (Keynes(1921),p.370 より)

上式をスターリングの定理を用いて近似

$\frac{1}{\sqrt{2\pi mpq}}exp\left[-\frac{h^2}{2mpq}\right]$ (Keynes(1921),p.370 より)

基本的にはベルヌーイ分布をスターリングの定理を用いて近似して正規分布を導くというアイデア。

$\frac{1}{\sqrt{2\pi mpq}}\int_{-a}^{+a} exp\left[-\frac{z^2}{2mpq}\right]dz, \frac{z}{\sqrt{2\pi mpq}} = t$と置換して

$\frac{2}{\sqrt{\pi}}\int_0^{\frac{2}{\sqrt{2pmq}}} exp[-t^2]\,dt$ (Keynes(1921),p.371 より)

が得られる。

ローザンヌ大学のパレート(1848-1923)、ケンブリッジ大学のケインズ(1883-1946)、どちらの経済学者も講義や著作の中で正規分布の導出を紹介していたことがわかる。

パレートは数学、物理学の研究からスタートして、一般均衡理論を中心とする経済学の研究をして、最終的に一般社会学研究にウェートを移していった。一方、ケインズは数学、確率論の研究からスタートして、貨幣論、銀行論（金融論）を経過してマクロ経済学を構築した。両者の研究スタンスの違いは、単にミクロ経済学とマクロ経済学といった対象の違いだけではなく、統計学と確率論の志向性の違いとして整理できないだろうか。

演習問題の略解

Q1. 正規分布の特徴は単峰性、裾野が広い、左右対称、平均値、中央値、最頻値が一致する。ベル型の形状などが挙げられる。加えて標準正規分布では、平均ゼロ、標準偏差が 1 で、分散は 1 の二乗である。

Q.5

相関関係と因果関係の違いは、基本的には因果関係が原因と結果の関係であるのに対して、相関関係は複数の因果関係が絡み合った状態と言える。つまり因果関係があっても、必ずしも相関関係があるとは言えない。

疑わしき相関とは、たとえばクリスマスシーズンや年末になるとケーキの売り上げがあがるかもしれないが、これは、景気と相関しているわけではなく、季節による影響である。このため見せかけの相関があっても、時間に従うデータでは、季節調整を行う必要がある。(この問題は論理的思考力や例示力をチェックするためのもので、本来、100%の模範解答は存在しない。この解答例は一例である。)

Q3-1. 10000 人に向けてテレビの視聴率調査をしました。ある番組を見た人は 2500 人でした。この番組の視聴率を信頼係数 95%で推定してください。

略解）

公式

$$x - 1.96\frac{\sqrt{np(1-p)}}{n} \leq p \leq x + 1.96\frac{\sqrt{np(1-p)}}{n}$$

に数値をあてはめて解く。

（この公式はアンケート調査等が二項分布 B(n, p)で記述できることを前提に、n が十分大きいことを前提に

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

が N(0, 1^2)に従うとみなせることから導かれています。また信頼区間 95%の Z のとる範囲は-1. 96 と 1. 96 の間です。）

X=2500, n=10000

p=X/n=0. 25

を公式に代入して、

(2500-1. 96√10000*0. 25(1-0. 25)/10000 <= p <=(2500+1. 96√10000*02. 5(1-0. 25))/10000

$$\frac{2500 - 1.96\sqrt{100 * 0.25 * (1 - 0.25)}}{10000} \leq p \leq \frac{2500 + 1.96\sqrt{100 * 0.25 * (1 - 0.25)}}{10000}$$

$$0.25 - 1.96\sqrt{0.25(1 - 0.25)} \leq p \leq 0.25 + 1.96\sqrt{0.25(1 - 0.25)}$$

$$0.25 - 1.96\frac{\sqrt{3}}{4} \leq p \leq 0.25 + 1.96\frac{\sqrt{3}}{4}$$

略解数式.png

$$0.25 - 0.24 * 1.732 \leq p \leq 0.25 + 0.24 * 1.732$$

$$0.24151 \leq p \leq 0.25848$$

より

95%の信頼係数で視聴率が 24.2%以上 25.8%以下である。

ただし$\sqrt{3}$＝1.732 として計算。

Q3-2. 人口 100 万人の C 市でアンケート調査（サンプル調査）を行うとしたら、統計的に有意と言えるサンプル数は何人くらいでしょうか？ただし事前調査によりアンケートの回収率は 25%とわかっているものとします。答えだけでなく、考え方も説明してください。

略解）

信頼係数を 95%と仮定して解く。

信頼区間の幅は 1.96*√(p(1-p)/n)

$$1.96 * \sqrt{\frac{p(1-p)}{n}}$$

を前後にとったものなので

$$2 * 1.96 * \sqrt{\frac{p(1-p)}{n}}$$

信頼区間は 5%以下なので

$$2 * 1.96 * \sqrt{\frac{p(1-p)}{n}} \leq 0.05$$

略解 2.png

これに p=0.25 を代入して n について解く

$$2 * 1.96 * \sqrt{0.25 * 0.75} / 0.05 \leq \sqrt{n}$$

67.89 <=$\sqrt{n}$

両辺は正なので二乗して

4609.65 <= n

ゆえに約 4610 人にアンケートを行えば良い。

もし回収率もわからない時は、仮に 50%として計算します。

その場合は、

$$2 * 1.96 * \sqrt{\frac{p(1-p)}{n}} \leq 0.05$$

p=0.5 を代入して n について解く

2*1.96*√(0.5*0.5) / 0.05 <=√ n

98 <=$\sqrt{n}$

両辺は正なので二乗して

9604 <=n

ゆえに 9604 人にアンケートを行えば良い。

（一万人アンケートのおそらくの根拠）

略解３.png

奥付

奥付

やさしい統計学講義

https://puboo.jp/book/122941

著者：夏木康志

著者プロフィール：https://puboo.jp/users/ynatsuki/profile

感想はこちらのコメントへ

https://puboo.jp/book/122941

電子書籍プラットフォーム：パブー（ https://puboo.jp/ ）

運営会社：株式会社トゥ・ディファクト

やさしい統計学講義

著　　夏木康志

制　作　Puboo
発行所　デザインエッグ株式会社