統計に関して

・時間を計って問題を解くようにする！！！！（過去問を！！）

条件付き確率に関して

P(B|A) = P(A∧B) / P(A)

P(A∧B) = P(A) + P(B) – P(A∨B)

四分位範囲→Q３　－　Q1

ヒストグラムの各度数の値は階級値（階級「範囲」の中央値）に置き換えて計算を行う

ベルヌーイ試行の分布に関しては大きく３種類ある

・二項分布　・ポアソン分布　・幾何分布

・二項分布に従う確率変数Xの期待値と分散 nは試行回数

期待値　np 分散　np(1 – p)

二項分布はｎが十分に大きいときは正規分布N(np,np(1-p))に近似する

二項分布→ベルヌーイ試行をｎ回行ってｋ回成功する確率

・ポアソン分布　平均λ回起こるのがｋ回おこる確率→離散型の分布になる

P(X=k) = (e\*\*(-λ) \* λ\*\*k ) / k!

期待値　λ　分散　λ

・幾何分布　→成功確率がｐの試行においてk回目で初めて成功する確率

P(X = k) =(1 – p)\*\*(k-1) \* p

期待値　1/p 分散　（1 – p） / p\*\*2

説明変数が大きいからといって必ずしも相関関係がおおきいとは限らない

中央値＜平均値 →　右に袖が長い

母比率ｐの信頼区間に関して

ｎが十分に大きい場合 標本比率をp^とするとこれは平均ｐ標準偏差√（p\*(1-p)/n）の正規分布にほぼ従う

・ローレンツ曲線について→不均衡さを求める

ジニ係数はローレンツ曲線と４５度線で囲まれた面積のことを指している。

数値が大きいほど不均衡である！！

系統抽出→等間隔に抽出する方法（単純無作為抽出に比べて手間が省ける）

多段抽出→最初に小さい集団を抽出しそこから無作為に抽出

層化抽出→いくつかのグループに分けて（性別とか）それぞれの層から無作為に抽出

クラスター抽出→いくつかのグループに分けて無作為の集団をすべて

標本誤差→標本のとり方による偶然に生じる誤差

非標本誤差→誤回答や未回答などに起因する誤差

１００点満点と２００点満点の標準偏差を比較する際は２００点満点の方の標準偏差を半分にすれば１００点満点のものと比較することができる。

決定係数は当てはまりの尺度のこと

回帰式のF値は回帰係数が０であるかどうかの検定のこと

その優位性を調べるのにｐ値を使用している

それぞれの回帰係数が０であるかどうかを調べるにはｐ値を見て判断する！

共分散は→cov[x,2y] = 2 cov[x,y]

・トレンドをとらえるには移動平均を使用する

・伸び率　=　（今期の値　–　前期の値）/ 前期の値

・二つの母平均の差の検定（母分散が分かっていないとき！！）

自由度はn1 + n2 - 2

＜重要＞一元配置分散分析のF値=T値の2乗

これはまじめに重要→一元配置分散分析の解説動画を見るようにする

一元配置分散分析に関しては３つ以上のクラスに差があるのかを検証している

二つの場合はｔ検定とかで二つの平均の差を求めたりして算出することができる

一元配置分散分析は差があるかないかだけを見るのでどのクラス間で差があるのかまではわからない

F = 群間変動(例えばクラスAの値をすべてクラスAの平均として考えて総平均で引いたやつの2乗和のことこれを自由度（クラスの数―１）で割る！！)/郡内変動（例えばそれぞれのクラスごとに偏差平方和を出して自由度（すべての観測された数―クラスの数）でわる）

・回帰分析にｔ値の求め方

T値　=　回帰係数　/ 標準誤差

（自由度はデータ数から切片の数を引く！！）

・AとBの相関係数が０であるとき2つは独立であるという

・無作為標本の標本平均は母平均の不偏推定値である！

・母比率の差の検定に関して

どのようにして求めるのか？？

T値（ｔ検定）とZ値（標準化）の使い分け

T値は母分散が分からず標本分散が分かっているときに使用する

Z値は母分散が分かっているときに使用する

・複利の計算の仕方前年からの変化率がｒである場合

前年の値＊（1+r/100）^n の式で表すことができる

・共分散に関して！！！相関関係を数値化したもの→共分散が表す数値は相関関係を表している

Cov[a\*X,Y] = a\*cov[x,y]

Cov[Z+X,Y] = Cov[X,Y] + Cov[Z,Y]

・mCn = m! / n!(m-n)!

・箱ひげ図

第一四分位数→下から数えて２５％に位置する数値

第二四分位数→５０％の位置にいる値のこと

第三四分位数→上から数えて２５％の位置にいる数値のこと

・分散分析にかんして

3群以上からなるデータ（例えば1組、2組、3組の算数のテスト等）や1つのデータに2つの要素を含むデータ（薬A、B、Cをそれぞれ10mg、20mg投与した場合の効果等）の母平均の差を検定する「分散分析」について説明します。分散分析は群ごとのデータのばらつきを元に、F分布を用いて検定を行います。分散分析を行うにあたっては、次に示すような「分散分析表」を作成します。帰無仮説は「各群の母平均は等しい」となります。

分散分析のポイントは「**データ全体の平均値から因子の各水準の平均値がどのくらいずれているか**」を見ること

・一元配置分散分析

1つの因子からなるデータを分析する方法で、因子に含まれる水準間の平均値の差を見ることができます。例えば、ある学校の1組、2組、3組の算数のテストのデータがある場合、一元配置分散分析を用いて、1組、2組、3組の算数のテストの平均点に差があるかどうかを検定できます。

F =　群間変動/郡内変動

1. 総変動を求める
2. 群間変動（群それぞれの変動を求める→Σ（各郡の平均-総平均）＊＊２→これを自由度（群の数-1）で割る
3. 郡内変動を求める→群の平均と各要素（各データ）の差の合計を自由度（各データの数-群の数）で割る

・二元配置分散分析

2つの因子からなるデータを分析する方法で、各因子における水準間の平均値の差を見ることができます。また、2つの要因が組み合わさることで現れる相乗効果の有無の確認もできます。例えば、薬A、B、Cをそれぞれ10mg、20mg投与した場合の効果についてのデータがある場合、二元配置分散分析を用いて薬の種類によって得られる平均値に差があるか、あるいは薬の投与量によって得られる平均値に差があるかどうかを検定できます。

・コレログラム

横軸にラグ（時間差）、縦軸に自己相関係数をとったグラフです。

コレログラムを見ると、データが周期性をもつかどうかを調べることができます

・両側検定と片側検定

薬Aに含まれるある成分Bについての分析を行います。Bの含有量を調べるため、生産された薬Aの中からランダムに25粒を抜き取り、成分Bの量を測定しました。その結果平均がmg、不偏分散がでした。この問題では帰無仮説を「薬A中の成分Bの含有量は100mgである」としたときに、3通りの対立仮説が考えられます。

１薬A中の成分Bの含有量は100mgではない

２薬A中の成分Bの含有量は100mgより多い

３薬A中の成分Bの含有量は100mgより少ない

は成分Bの含有量が100mgかどうかを調べるための検定です。

は成分Bの含有量が100mgより多いかどうかを調べるための検定です。この場合、成分Bの含有量が100mgより少ないかどうかについては考慮しません。

は成分Bの含有量が100mgより少ないかどうかを調べるための検定です。この場合、成分Bの含有量が100mgより多いかどうかについては考慮しません。1のような検定方法を「両側検定」、2と3のような検定方法を「片側検定」といいます。有意水準を5%とした場合、両側検定と片側検定の有意水準を図示すると以下のようになります。

容量は633mlよりも少ないかどうか」のような方向性のある仮説を検証するための片側検定では、平均値が633mlより大きくなってしまった時点で検定を終了し「帰無仮説を棄却できない＝633mlより少ないとは言えない」と結論付けます。

同様に対立仮説を「容量は633mlよりも大きい」と設定した片側検定では、標本の平均が633mlを下回った時点で検定を終了します。

・偏回帰係数の有意性の検定に関して　→ｔ検定を行う

偏回帰係数とは重回帰分析によって出てきた係数のこと

偏回帰係数の有意性の検定とは、定数項も含めた各偏回帰係数が0であるかについての検定結果です。

偏回帰係数\widehat{\beta}_iを[標準誤差](https://bellcurve.jp/statistics/glossary/1218.html)で割った値について、自由度(n-k-1)のt分布を用いて検定を行います。すなわち、次の式から算出される統計量t（t値）が自由度(n-k-1)のt分布に従うことを用います。nはサンプルサイズを、kは説明変数の数を表します。

・第一種の過誤、第二種の過誤

第一種の過誤…帰無仮説が正しいのに棄却する　=有意水準

第二種の過誤…帰無仮説が偽であるのにもかかわらずそれを真として棄却しない誤りのこと。対立仮設が正しいのに採択されない→対立仮設での分布から求める

・検出力と有意水準

有意水準…本当は[帰無仮説](https://bellcurve.jp/statistics/glossary/899.html)H_{0}が正しいのに、誤ってH_{0}を棄却してしまう確率

検出力…帰無仮説H_{0}が正しくないときに、正しくH_{0}を棄却する確率　１－第二の過誤　で求めることができる。帰無仮説が正しくない場合に、正しく帰無仮説を棄却する確率のことである。

・検定統計量に関して

検定統計量とは検定するときには検定するための値に変換しなければならないこのようにして算出された値のこと<https://bellcurve.jp/statistics/course/9317.html>

統計量Zは標準正規分布に従う

統計量Tはｔ分布に従う（不偏分散を使用する）

・ｐ値とは→ｐ値＝0.05の時これが起こる確率は0.05の確率である。

ｐ値→帰無仮説があっている確率のこと。ｐ値が低いということは帰無仮説がとても低い確率であるため棄却されるよねてきなやつ

P値：帰無仮説が正しいとした仮定とき、観測した事象よりも極端なことが起こる確率のことです。「観測した事象よりも極端な事象が起こる確率」であることから、これは累積確率となっています。コインの問題では、「6.25%」がP値になります。仮に有意水準を５％に設定していたとしてｐ値が５％より小さい場合は帰無仮説が起こることはめったにないということになり帰無仮説を棄却する。p値は、「帰無仮説が正しい場合に、それよりも極端なデータが観測される確率」

・自由度に関して→ある代表値や合計値があるときに自由に値をとれる数

m\*n分割表の自由度は（m-1）\*(n-1)となる

詳しくはこちら！！　→　<https://best-biostatistics.com/contingency/degree-freedom.html>

・F分布に関して→等分散性の検定（三郡以上の比較をしたいときに分散分析を行う）

母集団X,YがありXから取り出したｍ個の標本不偏分散をUx\*\*2 、Yから取り出したｎ個の標本不偏分散をUy\*\*2とすると二つの母集団の母分散が等しいときF＝Ux\*\*2 / Uy\*\*2は自由度（m-1,n-1）のF分布に従う。

・偏相関係数に関して

偏相関係数はAとBに強い相関関係があった場合、これはCによる第３の因子によってAとBは強い相関関係がもたらされている可能性がある。このような見かけ上の相関がある場合は第三の因子の影響を除いた相関係数「偏相関係数」を用いて相関関係を評価しなければならない。

・残差変動とは

残差変動とは、実際のデータと推定された回帰式から得られた予測値との差を表す

・有意水準の考え方→有意水準よりも低い結果が出たときは仮説を棄却する。

例えば基準を５％だとして今回の確率が0.04であった場合、５％よりも低いということはこれはめったに起こることではないと解釈する→よって仮説が間違っていると考え帰無仮説を棄却する。

・適合度の検定→カイ２乗分布を使用する

調査によって得られたクロス集計表がある場合、実測度数がある特定の分布に適合（　一致）するかどうかを検定することを適合度の検定といいます。適合度の検定では、カイ二乗分布を用いて検定を行います。適合度検定ではこの「理論値」からの「実測値」のズレを算出し、検定を行います。

カイ２乗X\*\*2 =(実測値　–　理論値)\*\*2/理論値

適合度の自由度はn-1になる！

実測値は実際に測定したときの値

理論値は　理論値

・2つの変数の期待値と分散

期待値

E(x + y) = E(x) + E(y)

E(x – y) = E(x) - E(y)

分散☆ものすごく大事！！！

V(x + y) = V(x) + V(y)　+ 2cov(x,y) ☆ｘとｙが独立でない場合

V(x + y) = V(x) + V(y)　☆XとYが独立である場合

V(x - y) = V(x) + V(y)　☆XとYが独立である場合

V(x) = E(X\*\*2) – {E(X)}\*\*2

・指数分布

連続型確率分布の一つで、機械が故障してから次に故障するまでの期間や、災害が起こってから次に起こるまでの期間のように、次に何かが起こるまでの期間が従う分布。ある期間に平均して\lambda（ラムダ）回起こる現象が、次に起こるまでの期間Xが指数分布に従うとき、X=xとなる[確率密度関数](https://bellcurve.jp/statistics/glossary/811.html)f(x)は次の式で表されます。\lambdaは指数分布のパラメータであり、必ず正の値をとります。

Ｅ(ｘ)　=　１/λ　　V(X)　=　1 / λ\*\*2

・標準化

Z =　（Ｘ – μ）　/ σ

・期待値の求め方　E(x)

離散型確率変数の場合　Σ　x ＊p

連続型確率変数の場合　∫　ｘ　＊　f(x) dx

分散の求め方V(x)

V(x) = E(X\*\*2) – {E(X)}\*\*2

・確率質量関数・確率密度関数にかんして

確率質量関数は離散型確率変数Xがある値xをとる確率を関数f(x)とした場合、f(x)は「確率質量関数」と呼ばれます。

確率密度関数は連続型確率変数Xがある値xをとる確率密度を関数f(x)とすると、f(x)を「確率密度関数」と呼びます。確率とは異なり、f(x)\geq 1になる場合もあります。確率密度関数のみ確率密度であるために１を超えることがある。確率密度関数を用いて確率を求めるときは一点における確率は求めることができないため必ず範囲で求めることをしなければならない。

・ｔ分布(標準正規分布に似ている)に関して（母平均を推定したいとき）

母分散が分からないときの母平均の推定など

T =　(標本平均　－　母平均)　/ √（不偏分散/標本数）　→この時は自由度n-1に従う

確率変数Xが自由度ｍのｔ分布に従っているとき、Xの期待値E(X)と分散V(X)は次のようになる

E(X)　＝０　V(X)　＝　m　/ (m – 2)

自由度＝　データの数　-群の数

★★二つの標本をｔ検定で行う場合。。。★★

2標本t検定を行う場合、前提条件の1つとして「2つの母集団の分散が等しいこと」が必要です

「等分散性の検定」で2標本の母分散が等しいかどうか検定（28-3章参照）

等分散ではないとは言えない場合（帰無仮説「2標本の母分散は等しい」が棄却されない場合）はt検定

等分散ではないと言える場合（帰無仮説「2標本の母分散は等しい」が棄却される場合）はWelchのt検定（※後述します）

ただし、統計学では検定を繰り返し行うと「多重性の問題」が生じるため、最近では2標本のt検定を行う場合には等分散性の検定は行わず、等分散かどうかを考慮する必要のない「Welchのt検定」を行ったほうが良いという考え方も一般的になりつつあります。

☆☆カイ２乗検定はカテゴリカルデータを対象とした検定手法、ｔ検定は連続データを対象とした検定手法

期待度数とは？？

もし関係がなかったらきっとこうなるであろうという回数を求める

カイ２乗検定は独立性の検定（関係があるかないかの検定）

・カイ2乗分布（母分散を推定したいとき）独立性の検定、２つの変数間で関連していない

カイ２乗値Σ（観測データ　–　期待度数）\*\*2 / 期待度数　＜分子だけ二乗であることに注意！＞

カイ2乗分布は標本から母分散を推定するときに使用する（母平均が分かっているとき）

母平均　μ　母分散　σ\*\*2

標本を標準化する→Z=(x1 – μ)/（√(σ\*\*2/n))

カイ２乗値　X\*\*2 =((n-1）\* s\*\*2) / σ\*\*2 (s\*\*2は不偏分散)（σ\*\*2は母分散）

V =　Z1\*\*2　+　Z2\*\*2 + …Zn\*\*2 が自由度ｎのカイ2乗分布する

確率変数Xが自由度ｋのカイ２乗分布に従っているとき、

E(x) = k V(x) = 2k

このVから９５％の信頼区間を求める

（母平均が分かっていないとき）→標本平均から推定する→この時カイ2乗分布の標本平均を使用するときは自由度がn-1になる

Z=(x1 – μ)/σ　この標準化の時のμが標本平均に変わることに注意

・区間推定に関して（母集団の従う分布が正規分布であるとき）（母平均の推定）

1. 母分散が分かっているとき

母分散のσ\*\*2を使用し標準正規分布から導く

1. 母分散が分かっていないとき

不偏分散s\*\*2の値を使い、ｔ分布を用いて信頼区間を算出する。

・損失関数に関して

損失関数とは正解値と予測値の誤差を表す関数のこと。この関数の誤差を最小化するような値を求める方法を考える。この時、損失関数を微分するとある瞬間に損失関数がどの程度傾いているのか知ることができる。この傾きの大きさを徐々に小さくし損失関数の最小値を求める方法を勾配降下法という。

・ユークリッド距離

ユークリッド距離は点Aと点Bの距離のことｎ次元の距離を求めることが可能

・平均と分散に関して

各データにｂを加えると

平均はｂ増え　,　分散は変化なし

各データをb倍すると

平均はｂ倍　分散はｂ\*\*２になる

・条件付きの確率

Pa(B) = P(a∧B)/P(A)

・不偏性

標本平均には不偏性はあるが、標本分散に関しては不偏性はない。

不偏分散　＝　標本分散＊｛n/(n-1)｝＝　（Σ偏差の2乗）/ （n – 1）

不偏分散を使用したときは自由度はn-1になるので気を付ける！！！！！

・標本誤差　=　標準偏差　/ √ｎ

標準誤差（SE：standard error）は推定量の標準偏差であり、標本から得られる推定量そのもののバラつき（＝精度）を表すものです。標準誤差は、一般的に「標本平均の標準偏差」を意味します

標本平均と母平均の誤差みたいな感じ

推定量の推定制度を表すもの

推定量の標準偏差

・推定と検定に関して

推定：母集団を特徴づける母数（パラメーター：平均など）を統計学的に推測すること。

検定：母集団から抽出された標本の統計量に関する仮説が正しいかを統計学的に判定すること。

＜日本人全員（母集団）からランダムに100人を抽出した標本を例に取ると＞

推定：抽出された100人の身長から日本人全員の平均身長を推測すること。

検定：日本人の平均身長が165cmと言われているが、抽出された100人の平均は167cmだった。このとき、100人の平均値は妥当かどうかを判定すること。

・母比率　→　母集団においてある事象が起こる確率のこと

母比率の信頼区間の求め方

標本比率±ｚ√｛標本比率×（１－標本比率）÷標本の数｝

この母比率は標準正規分布を利用して求める

標本比率を使用した標準化を行うとき(p=母比率)

Z＝　（標本比率　–　ｐ）/ √｛ｐ\*(１-p)/n｝

・母分散を標本から推定する方法はカイ２乗分布を用いて行う

Wを使用するときは自由度がn-1 になるので注意！

・母平均を標本から推定する場合はｔ分布を使用する

この時は統計量Tを使用している。

T=（標本平均 – 母平均）\* √（n – 1）/ 標本の標準偏差

このT分布はn-1の自由度になる。

・大数の法則→　母集団から標本を抽出する場合、抽出する数が大きくなればなるほど標本平均は母平均に近づく。

☆標本平均の分布

母平均　μ　母分散　σ\*\*２であるとき

標本の数を大きくするにつれて

標本平均　μ　標本分散　σ\*\*２/ ｎ

に近づく

標本分散s\*\*2を母分散を使用して

中心極限定理はサンプルサイズを大きくすれば大きくするにつれて正規分布標本平均　μ　標本分散　σ＊＊２/ 2

に近づくというもの

・変動係数　→　平均に対してデータがどの程度ばらついているのかを評価するもの

異なる種類のデータ同士を比較するときに使うデータのばらつきの大きさの比率を表したもの

変動係数が大きいほうがばらつきも大きくなる

変動係数　 ＝　標準偏差　/ 平均

・歪度と尖度

正規分布の歪度と尖度はどちらも０

歪度とは　どれだけ歪んで（ゆがんで）いるのかを表したもの

歪度＝０は左右対称　歪度＝負の値　は左に歪んだ（右にデータが集中している）

歪度　＝正の値は右に歪んだ（左にデータが集中している）

尖度とは　どれだけとがっているかを表している

尖度が正の値の時　はデータが平均付近に集中している

尖度が負の値の時　はデータが散らばっておりグラフが平たんになっている

確率変数の場合

確率を求める場合は確率密度関数とｘ軸との囲まれる面積を求める

期待値　＝　平均みたいな役割がある

分散　＝　（ｘ　－　期待値）＊確率　の合計