基礎統計（火１）廣松 Shike-pri

1. 統計の取り方

１－１．全数調査

母集団全体を調査。例・・・ 国勢調査、卒業文集のクラスで○○な人アンケート調査

１－２．サンプル調査

母集団の一部のみを調査。ただし、無作為に（意図的なものを排除し）サンプルを選ぶ。トリビアの種で２０００人調べればいいとかいってるあれです。

※ しかし、意図せずして偏りが生じてしまうことがある。

例・・・インターネットによる調査（ネット環境が無い人の意見が自動的に排除されてしまう）

この結果、「統計でウソをつく」ことが有り得るんだなぁー。むーん。

２．データの種類

２－１．時系列データ

ある対象についての異なった時点におけるデータ(暦年、年次、四半期（３ヶ月）など)

２－２．横断面（cross-section）データ

ある属性に関して、いくつかの異なる対象についてのデータ

例・・・各国の人口（人口という属性に対するいくつかの異なる国についてのデータ）

２－３．一次統計と二次統計

データそのものと、データが加工されたもの（物価指数とか）

３・(累積)相対度数分布

３－１・度数とは何ぞや？

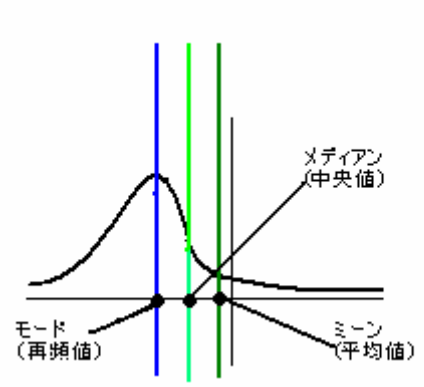
度数（frequency）とは、ある範囲に入る観測値の個数。

f_j などとあらわす。此のとき

$\frac{f_j}{n}$（n:観測地の個数）のことを相対度数という。

$\sum_{j=1}^{k} \frac{f_j}{n} = 1$　　という関係があるんでやんす。

３－２．相対度数分布（ヒストグラム）



左図が相対度数分布である。

３－３．用語の説明

モード　再頻度。分布の峰に対応する値。

メディアン　中央値（または中位数）

ミーン　平均値

例・・・集団（１，１，１，１，２，３，５，６，７）があるとき、

モードは１（４つある）

メディアンは２

ミーンは３　となりまする。

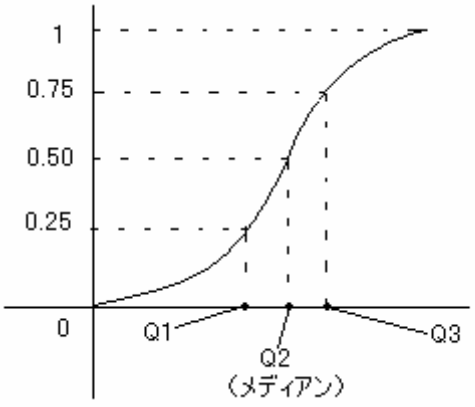
ちなみに、上の相対度数分布は「右に裾を引く分布」です。

このとき左からモード・メディアン・ミーンとなります。

左右対称のとき（正規分布のとき）は三つの値は一致します。

左に裾を引くときは左からミーン・メディアン・モードとなります。順序大事ナリ。

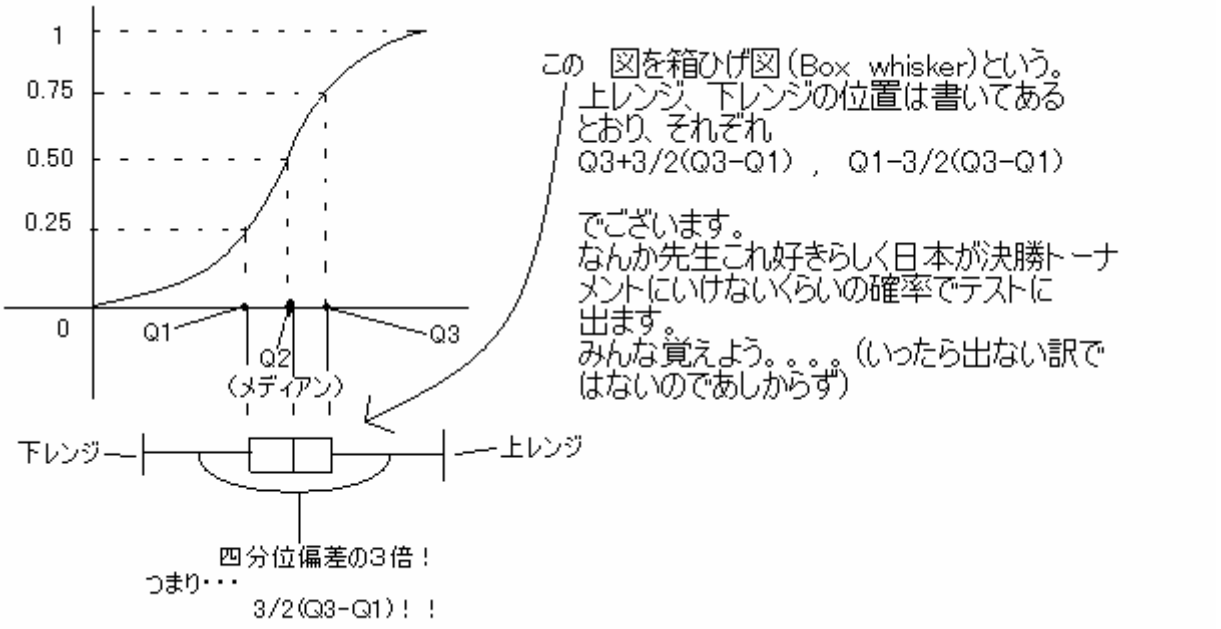
３－４．累積相対度数分布



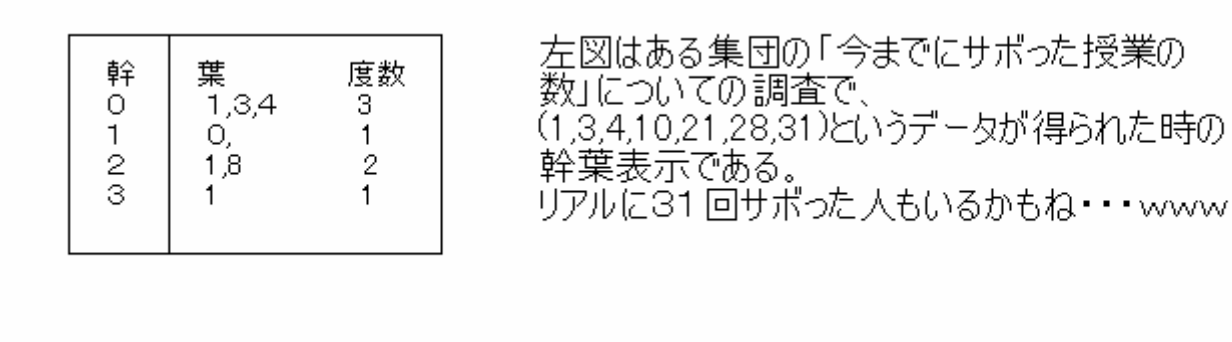
左図が累積相対度数分布である。

縦軸は相対度数の和を表す。わかって！
データの中で全体を４等分する点の値を４分位数と呼びます。
小さい順に、Ｑ１、Ｑ２、Ｑ３でＱ２はメディアンに等しいんだニャー。
また、レンジを改良したものとして四分位偏差というものがあり、
Ｑ＝１/２(Ｑ３―Ｑ１)　として定義されます。
これは散らばりの範囲をあらわすものというくらいの解釈で・・・

３－５．箱ヒゲ図、幹葉表示
☆箱ヒゲ図



☆幹葉表示



幹	葉	度数
0	1,3,4	3
1	0,	1
2	1,8	2
3	1	1

左図はある集団の「今までにサボった授業の
数」についての調査で、
(1,3,4,10,21,28,31)というデータが得られた時の
幹葉表示である。
リアルに31回サボった人もいるかもね・・・www

今６月１８日、日曜日の午前２時です。早くこんな作業やめたいです。
でも逃げちゃだめだ、逃げちゃだめだ・・・・

４・様々な（？）尺度

４－１　位置の尺度

モード、メディアン、ミーンなどです。意味はもう説明しました。
平均値は　$\bar{x}$とかあらわしちゃいます。これを踏まえて

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

という関係があります。

４－２　ばらつきの尺度

分散っていう概念があります。定義はしたのとおりでチュウ。

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

左上のが分散です。

次に、不偏分散は↓です。不偏分散は自由度を考慮したものです。

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

また

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

としたときのＳを標準偏差といいます。

４－３　その他もろもろ

他にも、ゆがみの尺度である歪度、とがり具合の尺度である尖度、積率などがあります。
こいつは分散では 2 乗であるところの乗数を３とか４とかＫとかに変えていろいろしたりしたものです。
その意味については教科書を読んだりしてください。。。

５・2 次元のデータ

今までは 1 次元のデータの話しをしてきました。1 次元のデータってのはまぁ、1 変数関数みたいな感じです。こっからは 2 次元、2 変数関数見たいな感じだと思えば良いと思うよー。

具体的にはこんな感じで。

２次元データというのは

（ｘ1，ｙ1），（ｘ2，ｙ2），（ｘ3，ｙ3），・・・，（ｘn，ｙn）

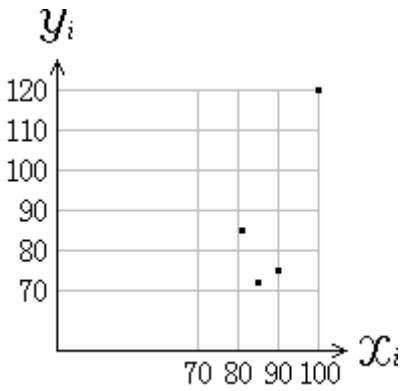
のように、１つの観測対象につき２つの観測値を持つデータです。

具体的には、下の表のようなデータです。

学生の名前	数学の点数	英語の点数
高橋	90 点	75 点
北村	81 点	85 点
野瀬	85 点	72 点
斉藤	100 点	120 点

この表を例とすると、ｉを自然数として、ｘｉは数学の点数、ｙｉは英語の点数ということになります。

そして、ｘi、ｙｉをそれぞれ横軸、縦軸に取った、下のような散布図を作ることができます。

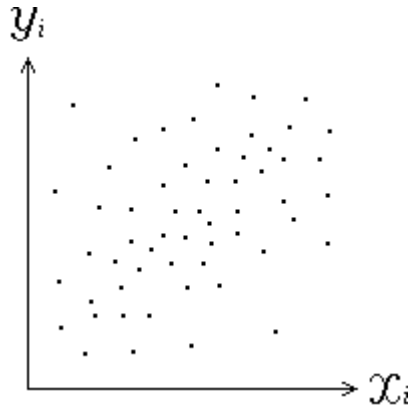


今は観測対象が４人しかいないので図中に点が４つしかありませんが、観測対象が多くなるとｘｉとｙｉの間の相関関係が見られることがあります。

相関関係とは下のようなものです。

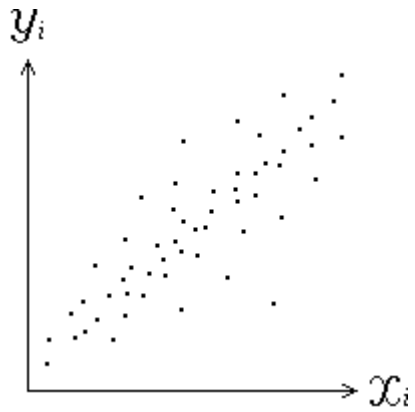
① 弱い正の相関関係

ｘｉが大きければｙｉも大きい傾向にあり、全体的にバラつきが大きい。



② 強い正の相関関係

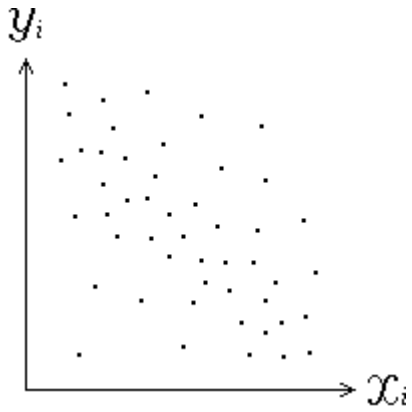
ｘｉが大きければｙｉも大きい傾向にあり、全体的にバラつきが小さい。（ほぼ一直線上に点が並ぶ。）



③ 負の相関関係

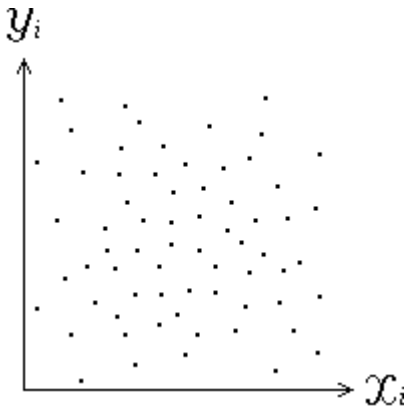
ｘｉが大きければｙｉは小さい傾向にある。

負の相関関係に関しては、あまり強弱の区別はしないみたいです。



④　相関関係無し

ｘｉの大小とｙｉの大小は無関係。



そこでこのような相関関係を調べる作業をしていきます。

１次元データのときには、データのバラつきを表す量として分散S^2を考えました。
２次元データでは、１次元の分散に相当する量として共分散Ｃxy を考えます。
共分散Ｃxy の定義式は分散S^2の定義式と似ていて、

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)(y_i - y)$$

分散は$(1/n)\,\Sigma\,(x_i-x)^2$でしたから、２乗の代わりに(y_i-y)が付いてると考えればいいのです。
ただ、それゆえ共分散は負の値を取ることもあります。

共分散が正のときは正の相関関係があり、負のときは負の相関関係があることも覚えておきましょう。

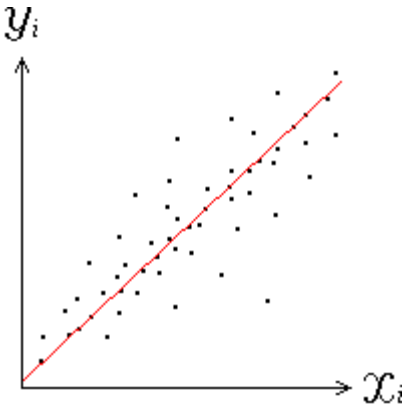
共分散で相関の正負を知ることはできますが、相関の強弱を知ることはできません。ですが、相関係数という値を求めれば、相関の強弱を知ることができます。
相関係数 r_{xy} の定義式は以下のとおりです。（Ｓｘ、Ｓｙはｘ、ｙの標準偏差のこと）

$$r_{xy} = \frac{C_{xy}}{S_x S_y}$$

Ｃxy/ＳｘＳｙの値は必ずー１以上１以下になりますので、Ｃxy/ＳｘＳｙの絶対値の大小で相関の強弱を知ることができます。

６・**回帰分析**（だんだん難しい）

回帰直線について説明します。もしも散布図を下の図のような一本の直線で表すとしたらどのような直線が適当かを分析します。



まず、その直線をy＝a＋bxと置きます。
bが直線の傾き、aが定数項であることに注意してください。（中学、高校の教科書と逆です。）
散布図を直線で表そうとしても、散布図上の全ての点がy_i＝a＋bx_iを満たす（直線y＝a＋bx上に乗る）ことは普通は起こらないので、その誤差（残差といいます）を e_i とします。つまり、y_i＝a＋bx_i＋e_iと置くわけです。（残差は d_i と表すこともある）

そして、残差が最小、すなわち、$e_1+e_2+\cdots+e_n$が最小になるようなa、bの値を求めると、（求め方は微分とか使っちゃいます。興味ある人は教科書で。でも難しい。）

$$\hat{b}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \qquad \hat{a}=\bar{y}-\hat{b}\bar{x}$$

（$\hat{b},\hat{a}$はbハット、aハットといいます。残差を最小にするようなa,bのことです。）

回帰直線には次のような性質があります。

① 回帰直線は（$\bar{x},\bar{y}$）を通る。

② $\hat{y}_i=\hat{a}+\hat{b}x_i$→理論値という。　$y_i-\hat{y}_i=e_i$→残差（y_iは観測値）

③ 関係式色々。

$$\sum y_i=\sum \hat{y}_i \qquad \bar{y}=\bar{\hat{y}} \qquad \Sigma(y_i-\bar{y})^2=\Sigma e_i^{\,2}+\Sigma(\hat{y}_i-\bar{\hat{y}})^2$$

証明は略します。。。

７・確率

7・確率

確率いきます。基本的なことはみんな高校でならったとおもうので省きます。高校でやってない可能性もある条件付確率からいきますね。

条件付き確率

袋に白、黒３つづつの玉をいれます。白の３つにはそれぞれ１、１、２の数字が書いてあり、黒の３つにはそれぞれ１，２，２の数字が書いてあります。この袋から１個だけ取り、数字が１か２かをあてる賭けをカイジが行ったとします。

まぁカイジは狡猾なので、引く瞬間に玉が白だということを確認しました。このときカイジはどっちに掛けるのが適当でしょうか？色がわからない状態では、１，２である確率はそれぞれ１／２でしたが、玉が白だとわかると、１の確率２／３、２の確率１／３となり、１にかけたほうがいいことがわかります。ざわ・・・・ざわ・・・・

このように、他の事象Ｂが起こったとわかっている場合に事象Ａが起こる確率を、

「Ｂを条件とするＡの条件付き確率」といい、**Ｐ（Ａ｜Ｂ）**と表します。

また、次のような関係があります。　　**Ｐ（Ａ｜Ｂ）＝Ｐ（Ａ∩Ｂ）／Ｐ（Ｂ）**

上の例でやってみると、Ｐ（1|白）＝Ｐ（１∩白）／Ｐ（白）＝$\frac{1/3}{1/2}$＝2/3 となります。

ジャスト２／３、条件付き確率に狂いなし。

また、上の関係式から導かれる次の関係式を乗法定理といいます。

P(A∩B)　＝　P(B)・P(A|B)　＝　P(A)・P(B|A)

また、事象Ａの起こる確率が他の事象Ｂに影響されないとき、すなわち

P(A|B)=P(A)のとき、この事象は「独立」であるといいます。

次にベイズの定理を説明します。

Ａを得られた結果、$H_1 H_2 \cdots\cdots\cdots H_k$を原因といたします。

このとき、P(Hi)を事前確率（ある原因が起こる確率)、Ｐ（Hi|Ａ）を事後確率（Ａという結果が起こったときに、その原因がHi である確率）とすると次のような関係式があります。

$$P(H_i \mid A) = \frac{P(H_i) \cdot P(A \cdot H_i)}{\sum P(H_j) \cdot P(A \mid H_j)}$$　　です。一見難しいですが上にあげた定義に

従ってみていくとそうでもありません。証明は教科書８４Ｐに書いてあり、比較的わかりやすいものです。気になる人はご覧ください。ここでは省略します。急にボールが来たもので・・・すいません。

8．確率変数

確率変数（random variable 以下 r.v と略すこともあるので注意。確率変数はＸのように

大文字を使って表す。）とは、それが取る各値に対しそれぞれ確率が与えられている変数のことを言います。たとえばサイコロの目Ｘは確率変数といえます。１から６までそれぞれ１／６という確率が与えられていますね。コインの表、裏じたいは確率変数ではないですが、表に０裏に１という数字を割り当てればそれは立派な確率変数になります。それぞれ１／２という確率が与えられています。

確率変数には二つの種類があります。一つは**離散型確率変数**です。これはサイコロの目のようにとびとびの値しかとらないものです。おもに整数っていってもいいかもです。高校でやったのはこのタイプだけでしたね。簡単に言えば、サイコロは３．２とか２．５３３とかの値はとらないでしょ？ってことでござるよケンイチくん。

離散型でそれぞれの確率$P(X=x_k)=f(x_k)$をXの確率分布といいます。

後で確率分布の例をあげていきますが、二項分布、ポアソン分布、超幾何分布などはこの離散型の確率分布の例です。

もう一つは**連続型確率変数**です。これは実数直線上の連続的な値をとるものです。

身長なんかは 172.3cm といわれるかもですが、実際は 172.3893720483….cm と無限に続くわけで。こういう場合連続型といえます。身長 50m とかの人はいないじゃないかと思うかもしれませんが（無表情な 3 分しか働けないおじ様はのぞいて)、そこの確率を０にすればいいわけです。

連続型の確率分布の定義は以下のものです。

$P(a \le X \le b)=\int_a^b f(x)dx$　　です。このとき関数 f(x)を確率変数Ｘの**確率密度関数**といいます。ちなみに$\int_{-\infty}^{\infty} f(x)dx=1$です。なんとなくイメージわきますよね。

イメージしにくいひとは図が教科書に書いてあるので見てみてください。91ｐです。

連続型では例えば X=170 である確率は０となります。定義が積分だからです。あくまで範囲で決まります。このへんが離散型と大きく違うところなんだよ明智くん。

累積分布関数

ってのがあります。「ある値以下の確率」を求めるときに使ったりします。このように確率変数Ｘに対して、ｘを実数とするときｘ以下の確率 F(x)=P(Ｘ≦ｘ)をＸの累積分布関数といいます。離散型だと$F(x)=\sum_{u \le x} f(u)$　連続型だと$F(x)=\int_{-\infty}^{x} f(u)du$ です。

期待値・分散

期待値と分散の話をします。期待値は高校でもやったのでつかみやすいと思います。サイコロの目の期待値は３・５です。この３・５はいわば確率変数１・２・・・・６の重心ともいえますね。

離散型の期待値は$E(X)=\sum_x xf(x)$、連続型の期待値は$E(X)=\int_{-\infty}^{\infty} xf(x)dx$です。

期待値は今後簡単のためにμ（ミュー）と表すことが多いです。期待値には演算法則があります。$[a]E(c)=c$　$[b]E(X+c)=E(X)+c$

$[c]E(cX)=cE(X)$　$[d]E(X+Y)=E(X)+E(Y)$　（期待値の加法定理）

てなかんじです。

分散の話にいきます。（これは前あげた分散とちょっと違うらしいです。注意してケロ。

分散は期待値からの距離を基準としています。どれくらいばらついてるか、ですね。

分散は$V(X)=E\{(X-\mu)^2\}$と表せます。

しかし実際はこうやって計算するわけでなく、便利な公式があります。

$V(X)=E(X^2)-(E(X))^2$　これはめちゃくちゃ重要だって書いてあるので覚えてくれヨォ・・・・分散はσ^2と表すことが多いようです。$\sqrt{\sigma^2}$**のことを標準偏差といいます。**

チェビシェフの不等式

おれの友達にチェビシェフっていうやつがいるんですが、そいつがこの間久しぶりに電話してきてこんなん見つけたぞ！って言ってきました。深夜2時に迷惑なやつです。

$P(|X-\mu|\geq k\sigma)\leq\frac{1}{k^2}$だそうです。これを変形すると

$P(|X-\mu|\leq k\sigma)\geq 1-\frac{1}{k^2}$となるそうです。

教科書105Ｐに例があるぞ！と言い残してやつは電話を切りました・・・

期待値と分散しかわかってないのにある範囲の確率がある程度わかるし、どんな確率変数にも成立する絶対的なものらしいです。テストに出るかどうかはわかりませんが覚えててくれると奴も浮かばれるとおもいます・・・まだ生きてますが。

いよいよ次から確率分布です。ヤマともいえるところにきました・・・ここまで読んで無理だと思ってる人、まだテスト前なんだから、精一杯やって、それから単位をとりなさい。まぁおれも気持ちがくじけそうです。基礎統計は楽だってオリパンフに書いてあったのに・・・これは孔明の罠だったか・・・おなかすいたんでご飯食べます。

シーチキンと豆腐ウマス wwww うは wwww おｋwwww 夢がひろがりんぐ wwww

９・確率分布

確率分布、行きます！　ということで確率分布の話です。

離散的確率分布

① 二項分布

おれがダーツを投げてブル（真ん中）に入る確率を$\frac{1}{3}$, 入らない確率を$\frac{2}{3}$とすると、

10 回中 x 回ブルに入る確率は　$f(x)={}_{10}C_x(\frac{1}{3})^x(\frac{2}{3})^{10-x}$　となりますね。(データは今取りましたw30 本中 10 本入りました。まぁまぁです。)

こういう試行を一般化すると

$P(X=x)=f(x)={}_nC_xp^x(1-p)^{n-x}$　となります。この確率分布を**二項分布**といいます。(確率変数Xはある事象が起こる回数です)

このとき、**確率変数Xは二項分布に従う**といい、**X～Bi(n,p)**と書きます。

このとき期待値、分散はそれぞれ　**E(X)=np　V(X)=np(1-p)**　となります。

② ポアソン分布

二項分布で n とか x とかが大きくなると計算が大変になります。n=10000000000 とかやってられません。手が、手がぁぁぁぁ・・・・ってなっちゃいます。

しかし大量の試行でなかなかおきない事象を起こそうとすることは往々にしてありますね。(例・ポーカーのロイヤル・ストレート・フラッシュ、トキワの森の Lv5 のピカチュウ探し、パワプロのサクセスで最初からオールCの選手を出す、俺に彼女ができる等)

このように二項分布において n→∞、p→0 ， np→λ となる極限では、次の近似が成立。

$P(X=x)={}_nC_xp^x(1-p)^{n-x}\longrightarrow\frac{e^{-\lambda}\cdot\ \lambda^x}{x!}$　です。この確率分布を**ポアソン分布**といいます。このとき、**Xはポアソン分布に従う**といい、**X～Po(λ)**と書きます。

このとき期待値、分散はそれぞれ E(X)=np=λ　V(X)=np=λ　となります。二項分布とそっくりですね。

例を一つ。ポーカー１ゲームでロイヤル・ストレート・フラッシュが成立する確率を仮に p=0.0002 とし、10000 回ポーカーをやってロイヤル・ストレート・フラッシュが 3 回出る確率を求めると np=λ=2, x=3 ですので

$f(3)=\frac{e^{-2}\cdot(2)^3}{3!}=0.180447$ となります。ポーカー知らない人はごめんなさい。

③ 幾何分布（待ち時間分布）

ダーツやってるとなかなかブル（真ん中）に入らないときがあります。何が悪いのか試行錯誤すると余計入らなくなるものです・・・入るまで投げ続けるぞ！と意地になって投げ続けて、30 投目にやっと入った！ってこともあります。最近はほとんどないですが、ね。

このように試行の回数はあらかじめ決めないでおき、次々と続け、最初に成功するまでの試行回数を表す確率変数をＸとしたとき、X=x すなわち x 回目でようやく成功する確率

は $P(X=x)=f(x)=p(1-p)^{x-1}$ と表されます。大丈夫ですよね（＾＿＾）

この確率分布を**幾何分布**と呼びます。幾何分布は離散的な**待ち時間分布**です。

このとき、**Ｘは幾何分布に従う**といい、**Ｘ～Ge(p)**と書きます。

幾何分布の期待値、分散は次の通りです。

$$E(X)=\frac{1}{p} \qquad V(X)=\frac{1-p}{p^2}$$ 期待値のほうはイメージわきますね。

主な離散的確率分布は以上です。他にも負の二項分布、超幾何分布などがありますが、過去のシケプリなどをみたところ言及していないものが多かったです。きになるひとは教科書を見ておいてくださいね。出たらその時は・・・再就職だな。

だいぶ疲れてきましたががんばります。だって、おれこのシケプリ作り戦争が終わったら結婚するんだ・・・

連続的確率分布

① 正規分布

正規分布はもっとも代表的な連続型の確率分布です。こいつがわからないと仕様がありません。気合入れてやっていきましょう。

正規分布の密度関数は $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ であらわされます。こんなんの覚えられるわけないので覚えなくていいです。教科書持込可なので。

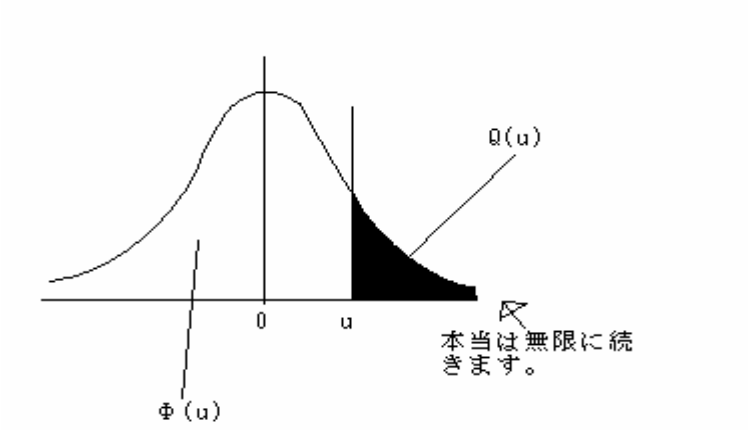
このとき**確率変数Ｘは正規分布に従う**といい、**Ｘ～N(μ,σ²)**と書きます。

正規分布の期待値は μ、分散は σ^2 です。正規分布は期待値と分散を決めると一意的に決まります。

連続型の確率変数Ｘが正規分布Ｎ(μ,σ²)に従うとき、ａＸ＋ｂは正規分布Ｎ(ａμ＋ｂ , a ²σ²)に従います。ここで、ａ＝１/σ、ｂ＝－μ/σ として確率変数Ｚ＝(Ｘ－μ)/σ を定めると、Ｚは

正規分布N(０,１)に従います。こうしてできた期待値０、分散１の正規分布を**標準正規分布**と言います。
教科書 280p には標準正規分布表が載っていますね。こいつの見方を説明します。
ここに載っているのはすべて「上側確率」です。累積分布関数だと下側なんですが、ね。
この上側確率を Q(u)と表してます。累世分布関数はΦ(u)と表します。
まぁわかるとは思いますが　　Q(u)=1-Φ(u)です。だから上側がわかれば下側もわかります。

$Q(u) = 1 - \Phi(u) = \int_u^{\infty} \phi(u)du$ です。



関係しきとして大事なものがあります。**Φ(-z)=1-Φ(z)**　は大事。表にはプラスの値しか書いてありませんが、これで u がマイナスでもできますね。
例として、$P(-1 \le u \le 2.23)$ を求めてみましょう。
表をみてみると、Q(1)=0.15866 ですね。よって Q(-1)=1-Q(1)=1-0.15866=0.84134 です。
また Q(2.23)=0.12874 です。
$P(-1 \le u \le 2.23)$ =Q(－１)－Q(2.23)=0.84134-0.12874=0.7126　となります。おｋ？
さて、ここでの u っていうのはあくまで標準化した時の確率変数Ｚ＝(Ｘ－μ)/σ の値であることに注意してください。例えばある 100000 人の団体が受けたテストの得点が正規分布に従い、平均（期待値）が 50 点、分散が10^2だったとし、得点 T として 50≦T≦51　となる人数を求めたいとする。
すると
P(50≦T≦51)＝P(0≦T－50≦1) =P(0≦T－50／10≦0.1)
=Q(0)－Q(0.1)=0.50000－0.46017
=0.03983
となります。ということは 100000 人受ければ大体 3983 人くらいこの範囲に入ってるだろう、ってことになります。
確率変数Ｔを標準化して、標準正規分布に従うようにする。これが上の変形です。

間違っても、教科書の表に 50 がない！51 がない！って言わないでください・・・おれはあわややるところでしたが、ねｗ

② 指数分布

故障率が一定のシステムの偶発的な故障までの待ち時間、つまり寿命だとか耐用年数とかあるいは災害までの年数なんかもこいつに従うらしいです。

つまり、こいつも待ち時間分布の性質を持ってます。

確率密度関数は

$f(x)=\lambda e^{-\lambda}$ (x≧０) , 0 (x<0) となります。

累積分布関数は

$F(x)=P(X\le x)=1-e^{-\lambda}$ (x≧0) , 0 (x<0) です。

期待値、分散については

$E(X)=\frac{1}{\lambda}$ $V(X)=\frac{1}{\lambda^2}$ です。

このとき**確率変数Ｘは指数分布に従う**といい、**Ｘ～Ex(λ)**と書きます。

主なものは以上です。難しいですね・・・がんばっていきましょう。

と、思いましたが、いったんＵＰします。

7 章からが微妙なんでね・・・もうやめたいです。でもあきらめたらそこで不可ですよね安西先生・・・・

10・多次元の確率分布

第 7 章にあたります・・・なんですが、第 7 章のどこが削られるのかがいまいちはっきりしません。７．２と７．４が削られるという説が有力ではあるんですが。

ノートを見てみると 7 章に当たる分は半ページだけでした・・・なのでここは飛ばしちゃいます。過去問を見たところここからの積極的な出題はないみたいです。後でここの知識が必要な場合はその都度載せます。ただ、7 章、多次元の確率分布はお読みください。

11・標本分布

第 9 章にあたり、9・4 は省かれています。

これからは**統計的推測**の話をします。よくクイズ番組で「東大生の正解率！」とかありますね。あれは別に東大生全員に聞いたわけではなくて、数十人の東大生を選んでその結果をあたかも東大生全体の結果としてやっているわけです。この場合東大生が**母集団**, 選ばれた数十人の東大生は**標本**といい、その選び方を**標本抽出**といいます

しかし標本の選び方によってはばらつきが当然生まれてしまいます。これに対応するために出てくるのが**標本分布**です。

標本によって母集団を推定するのが目的なんですが、その手がかりとなるのが

母数（パラメータ）です。パラメータとしては

$\mu = E(X) = \sum_x xf(x)$　（離散的）　$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$　（連続的）

であらわされる**母平均μ**や同じようにして**母分散**σ^2などがあげられます。
母集団分布がわからなくても母平均や母分散がわかれば多くのことがわかります。
一方、大きさnの標本X_1X_2・・・・X_nからとった

標本平均$\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$　や

標本分散$s^2 = \dfrac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$

など、標本を要約し、母集団の母数のいろいろな推測に使われるものを**統計量**といいます。
これら統計量の優れているのは次の関係式で母集団につながっているからだお（ ＾ω＾）

$E(\overline{X}) = \mu$　　$E(s^2) = \sigma^2$　この性質かなり大事です。おｋ？

ここでちょっと。標本分散の分子が n-1 なのはなんでか？むかーしみた分散はたしか n だったはずです。これはまぁよくわかりませんが自由度が関係しているらしいです。
上に記した標本分散のことを**不偏分散**といいます。この不偏分散の n-1 を自由度といいます。「自由に動ける変数の数」という意味らしいです。
統計量は一般的に標本の関数$t(X_1X_2$・・・・$X_n)$で表せます。この統計量の確率分布をその統計量の**標本分布**といいます。この統計量の値の出方から母集団分布が求められます。
べ、べつにあんたのために母集団分布を求めるんじゃないんだからねっ・・・ξ////)ξ

ところで、標本和$X_1 + X_2 + \cdots + X_n$や標本平均$\overline{X}$の具体的な標本分布は母集団分布に依存します。

二項母集団

母集団分布が母数pのベルヌーイ分布ならば二項分布Bi(1,p)、$X_1 + X_2 + \cdots + X_n$の分布は二項分布Bi(n,p)に従う。

ポアソン母集団

母集団分布が母数λのポアソン分布Po(λ)ならば、$X_1 + X_2 + \cdots + X_n$はポアソン分布Po(nλ)に従います。

正規母集団

母集団分布が母数μ、σ^2の正規分布N(μ, σ^2)ならば、$X_1 + X_2 + \cdots + X_n$は

正規分布N(nμ, nσ^2)に従い、$\overline{X}$は正規分布N(μ、$\dfrac{\sigma^2}{n}$)に従います。←これ大事ス wwww

眠いです・・今午前5時25分です。もう倒れそうです・・・でも大丈夫。おれが倒れてもきっとジェバンニが一人でやってくれますから・・・

12　正規分布からの標本

第10章に当たります。10.5と10.6は省かれてます。

標本が正規母集団からとられているという過程は統計学の理論の中心です。このとき正規母集団からの標本に基づく統計量の標本分布を計算することが必要となるが、この計算を行うのが、**正規標本論**です。

分散が既知のときの標本平均の標本分布

上で$\overline{X}$は正規分布N(μ、$\frac{\sigma^2}{n}$)に従うことを言いました。正規分布に従う以上標準化することができます。この場合標準化変数$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$は標準正規分布N(0,1)に従います。

つまり、σ^2さえわかっていれば$\overline{X}$の標本分布は結局標準正規分布N(0,1)を見ることに帰着します。

χ^2分布（カイ2乗分布）

定義いきます。

Z_1Z_2・・・Z_kを独立な、かつ標準正規分布N(0,1)に従う確率変数とします。

今、　　$\chi^2=Z_1{}^2+Z_2{}^2+\cdots+Z_k{}^2$　とすると、

確率変数χ^2が従う確率分布を**自由度kのχ^2分布といいます。**いいわね？鉄郎・・・うん、メーテル。

χ^2分布を用いると、正規母集団からの標本に基づく標本分散s^2の標本分布は次のようにまとめられます。

標本分散$s^2=\frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n-1}$　とするとき、

$\chi^2\equiv(n-1)\frac{s^2}{\sigma^2}$　は自由度（n-1）のχ^2分布χ^2(n-1)に従うといえる・・・じっちゃんの名にk（以下略）

χ^2分布の上側確率のパーセント点の表が282、283ページにあります。

分散が未知のときの標本平均の標本分布

世の中うまくいかないものです。上で分散がわかっている時の話をしましたが、分散がわかっていることはあまり現実的じゃないです。

標本平均の標本分布の標準化$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$　は分散（σ^2）がわからないと計算できません。

そこで、少なくともわかる標本（不偏）分散s^2を代わりに使っちゃおう、っていうのが人情ってもんです。現実的ですしね。

ここで定義されたのがスチューデントの**ｔ統計量**　$t = \dfrac{\overline{X} - \mu}{\sqrt{s^2/n}}$　です。

これは既にもう標準正規分布には従いませんね。
いろいろ変形していくと

$t = \dfrac{\dfrac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2}/(n-1)}}$　という豪勢な式になります。これを死神の目でよーくみると

と・・・寿命がみえｒ（以下略）気を取り直してよく見ると・・・

分子の$\dfrac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$は標準正規分布N(0,1)に、

分母にある　$(n-1)\dfrac{s^2}{\sigma^2}$は自由度（n-1）の$\chi^2$分布$\chi^2$(n-1)に従います。ｔはこれらの比になっております。これらの組み合わせでｔの密度関数が決まるので、正規標本論ではこれらを新たに**ｔ分布**と呼んでいます。定義はしたのとおりです。

ｔ分布　二つの確率変数ＹとＺが次の条件を満たすものとする。

① Ｚは標準正規分布N(0,1)に従う。

② Ｙは自由度ｋのχ^2分布χ^2(n-1)に従う。

③ ＺとＹは独立である。

今、確率変数ｔを

$t = \dfrac{Z}{\sqrt{Y/k}}$　と定義すると、ｔが従う確率分布を**自由度ｋのｔ分布**といいます。

自由度ｋのt-分布をt(k)と表します。

これによると、さっき出てきた

$t = \dfrac{\overline{X} - \mu}{\sqrt{s^2/n}}$ は自由度n-1のｔ分布t(n-1)に従います。

また、$\overline{X}$の標準偏差$\dfrac{s}{\sqrt{n}}$を標本平均の**標準誤差**といいます。

10 章終わりました。いよいよあと二つ。がんばって生きたい・・んですが、もう限界近いです。ほんとに倒れそう・・・でも大丈夫・・・私が倒れてもかわり（まっすん）がいるもの・・・

13 **推定**

第 11 章の内容です。ここからの内容は「統計理論の中心であり、頂点である」そうなので、心していくんだよのび太くん・・・

ここは 11．5．2 と 11．5．3 を省くと聞きましたが・・・間違ってたらごめんなさい。

さて行きましょう。標本 X_1X_2・・・・X_n から求めた統計量を一般に**推定量**といいます。それらから未知である母集団の母数（パラメータ）を**推定**するんです・・・どうです、なんだかワクワクするでしょう？？？

大きく分けて推定の仕方には点推定と区間推定とがあります。

こ、これから頑張るのよっ・・・ってあんたのために言ってるんじゃないからね！ξ////)ξ

点推定

母集団の未知の母数をある一つの値で推定する方法を点推定といいます。標本平均 $\overline{X}$ で母平均 μ の推定をする、みたいなかんじです。そうなるとどんな統計量を推定量とするかが大事ですね。そこでそいつをどうやって求めるのかについて話します。最尤法です。

最尤（さいゆう）とは「もっともっともらしい」ということです。これから使うのは「最尤原理」というもので、「現実の標本は確率最大のものが実現した」という仮定です。

ちょっと教科書にある例をあげてみませふ。（こっからは教科書に書いてあることです。教科書を読みたい人はそちらをどうぞ！！）

1 になる確率が p, 0 になる確率が 1-p のベルヌーイ分布 Bi(1, p)が母集団分布の場合を考えます。もちろん何を推定するかというと p ですね。

じゃあ 5 回くらいやってみましょう。すると・・・

$X_1=1, X_2=1, X_3=1, X_4=1, X_5=0$　とでました。まぁ見た目たぶん p=0.8 だろうな、と推定できますね。最尤原理より、↑の標本はある p において一番起きやすいものが起きてるとします。このとき、この標本が得られる確率は

$L(p)=p^4(1-p)$　となりますよね。このとき L(p)は p のいろいろな値でのもっともらしさ**（尤度）**を表している関数とみなせます。そしてこの関数を**尤度関数**といいます。

最尤法とは尤度関数を最大にするものを推定値や推定量とする考え方です。おｋ？

尤度関数を最大にする値が最尤推定値、関数としてだと最尤推定量となります。

上に上げた例だと、　$\dfrac{dL(p)}{dp}=p^3(5-4p)$ なので p=0.8 が最尤推定値です。尤もですね。

一般に、大きさ n の標本で上の p のような未知の母数を θ とすると、尤度関数は

X_1X_2・・・・X_n の同時確率分布を θ の関数とみなしたものとなり、X_i の確率分布を

F(x, θ)とすると、尤度関数は $L(\theta)=\prod_{i=1}^{n} f(x_i,\theta)$ と掛け算になります。

これは母数がたくさん出てきたときでも同じです。

そんで、最尤推定量を出すときは、$\frac{\partial L(\theta)}{\partial \theta}=0$ を計算するみたいです。

掛け算で出てきているのでいったん対数をとって和にして考える方法（**対数尤度**を考える方法）もあります。この辺正直自分も微妙なんでなんじゃらほいと思った人は教科書を読んでみてくださいね！教科書にいっぱい例が出てます。ぜひチェック。

あと点推定の基準とかがあります。複数推定量が出てきたときにどれにするのか決めないといけないときとかに使うみたいです。

不偏性

$\hat{\theta}$ を推定量、θ を母数としたときに $E(\hat{\theta})=\theta$ となる性質のことです。

標本平均はつねに母平均の不偏推定量です。s^2 なんかももちろんそうですね。

他に標本の大きさ n が大きくなるに従い推定量が真の母数に近づく**一致性、**

出来るだけ分散が小さいほうがいいという**有効性**なんかがありますです。

見た感じ区間推定のほうが大事みたいですね・・こっちも気合入れましょう。

区間推定

見た感じ、出そうです。ここ。過去問にもいくつか・・・やっていきましょう。

今までの点推定では母数θをある一つの値として推定してきましたが、今度はθに対して確率の考え方を用いてちょっと幅をつけて推定します。真の母数の値θが**ある区間（L、U）**に入る確率を**１－α**（つまり、αはθが区間に入らない確率ですね。）以上になるように保証する方法です。L,U はそれぞれ**下側、上側信頼限界**といい、１－αを**信頼係数**といいます。そして区間[L,U]を 100(１－α)％信頼区間と呼びます。

１－αは通常 0.99 や 0．95 などに設定されることがおおいです。正規分布（時には t 分布）の場合でやってみます。226 ページにぐだぐだといろいろ書いてございます。これは一応読んでください。ただ、ここでは過去問や練習問題を用いてちょっくらやってみます。

2006．2/7　廣松　試験問題

第 6 問　(1)

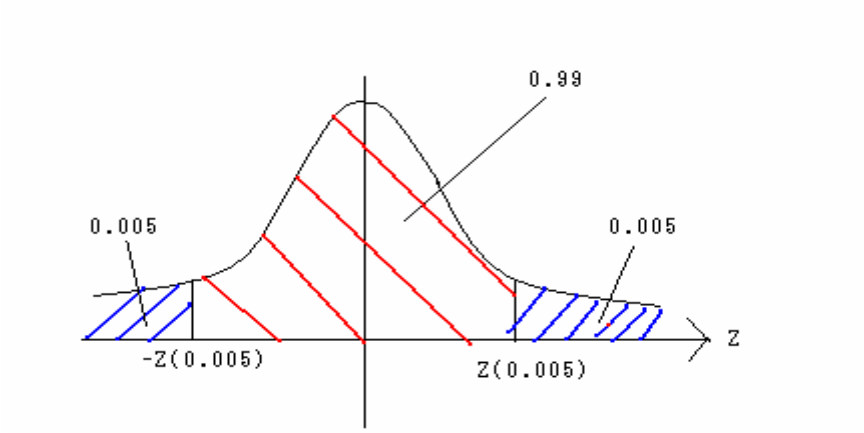
母平均=μ、母分散=25 の正規母集団について、次の問いに答えよ。

この母集団から大きさ n の標本を取り出して、母平均μを信頼水準 99％で区間推定したい。

信頼区間の幅を 4 以内にするためには n の大きさをどのように取ればよいか。

正規母集団の問題ならおそらく後ろの表を使うはず。そうなると**標準化**が必須です。

標準化すると $Z=\dfrac{\overline{X}-\mu}{5/\sqrt{n}}$　ですね。さて、下の図を見てください。



連続型の確率密度関数だと確率は面積で表されるといいました。つまり、どこからどこまでだと面積が 0・99 になるのかを考えないといけません。そこで $Z_{0.005}$ をとりました。これは「こっから先の確率が 0.005 になる点」です。その対称点 $-Z_{0.005}$ は「こっから後の確率が 0・005 になる点」です。二つあわせて 0.01 ですね。つまり、この間にあれば
確率は 0・99 なわけです。つまり、

$-Z_{0.005}<Z=\dfrac{\overline{X}-\mu}{5/\sqrt{n}}<Z_{0.005}$　となればおｋです。

$Z_{0.005}$ を表で調べてみると・・・2.58 が一番近いようですね。こいつを使います。
上の不等式に $Z_{0.005}$= 2.58 を代入して整理すると・・・

$\dfrac{-12.9}{\sqrt{n}}+\overline{X}<\mu<\dfrac{12.9}{\sqrt{n}}+\overline{X}$　となります。さて、問題は「幅が 4 以内」でしたね。

つまり、$\dfrac{25.8}{\sqrt{n}}<4$　となればよいわけです。こいつを解くと 41.6＜n ですなわち
42≦n　となります。　おｋ？

じゃあ次はｔ分布を使う例もやってみます。
問題
ある正規母集団から大きさ n=5 の標本
9.75　7.95　12.80　8.25　9.86　を得た。母平均 μ の信頼係数 95％の信頼区間を求めよ。

母分散がわからないので方針はｔ分布ということになりますね。まず出すもん出しましょう。
$\overline{X}=9.72, s=1.92$ です。これは電卓を駆使してだしませふ。
さて、上と似たような感じです。今度はｔ分布になっただけと思ってください。

$t = \dfrac{\overline{X} - \mu}{\sqrt{s^2/n}}$ は自由度 n-1 のｔ分布 t(n-1)に従うことを考えると次の不等式がすぐわかります。

$-t_{0.025}(4) < \dfrac{9.72-\mu}{1.92/\sqrt{5}} < t_{0.025}(4)$ です。表より $t_{0.025}(4)=2.776$ ですから、後はせっせと計算して、7.33<μ<12.11　　なので、95%信頼区間は[7.33, 12.11]　となります。
こんなかんじです。母分散の信頼区間なんかもあるみたいですね・・・これはχ二乗分布を使うんでしょう。目を通しておくのも一興かと。
いよいよ次で最終しょうです。しっかり頑張りましょう。

14・**仮説検定**

ここもよく出題されています。しっかりがんばるのよ・・・鉄郎・・・うん、メーｔ（以下略）
例えば、サッカーで日本がブラジルに勝つ確率は 50%である、という**仮説**を立てたとします。じゃあやってみようということで 20 回試合をしたところ日本は 1 勝 19 負でした。
このとき、「日本がブラジルにサッカーで勝つ確率は 50%である」という仮説は支持できるでしょうか？**仮説が正しいと仮定して**確率を二項分布で計算するとおそらくとんでもなく低い確率になると思います。そうなるとほぼ「起こりえない」こととなり、この仮説は**「誤っている」**と判断せざるを得ません。このとき仮説は**棄却**されるといいます。
でも、一応確率は０でないわけです。例えば 10 パーセント以下は「起こりえない」と判断する人ならこの仮説は「誤っている」というでしょうが、0.00000000000・・・どこまで続くのやら・・・１パーセント以下じゃないと「起こりえない」と判断しない心の広い人にとってはこの仮説は「正しい」ことになります。
このように、仮説に対してあらかじめどの程度の希少確率を考えるかにより、仮説が有意か否かがわかります。この基準の確率を**有意水準**といいます。

さて、日本がブラジルに勝つ確率について何か積極的な判断をしたいなら、「50%」である、と同時に「50%ではない」という仮説を立てておき、前者が棄却されたとき後者が採択されたとします。このときもとの仮説を**帰無仮説**、それと対立する仮説を**対立仮説**といいます。お互いはそれぞれ否定の関係にあり、それぞれH_0, H_1などと表します。
ちなみに帰無仮説を棄却するかしないかについては次の 4 つの場合があります。
①H_0が真実で、H_0を棄却しない。（採択する）・・・正しい！
②H_0が誤りである（H_1正しい）のに、H_0を棄却しない。・・・誤り！
この誤りのことを**第二種の誤り**といったりします。品質管理に照らし合わせて**消費者のリスク**といったりもします。
③H_0が真実なのに、H_0を棄却する。・・・誤り！
この誤りのことを**第 1 種の誤り**といったりします。品質管理に照らし合わせて**生産者のリスク**といったりもします。
④H_0が誤りで、H_0を棄却する。・・・正しい！

母平均に関する検定
こちらも実際に問題を解いていく形で説明します。
検定には二種類あって**両側検定、片側検定**とあります。片側検定には**左片側検定、右片側対立検定**とあります。
両側検定の帰無仮説、対立仮説はH_0：$\mu=\mu_0$　H_1：$\mu\neq\mu_0$
左片側検定の帰無仮説、対立仮説はH_0：$\mu=\mu_0$　　H_1：$\mu<\mu_0$
右片側検定の帰無仮説、対立仮説はH_0：$\mu=\mu_0$　　H_1：$\mu>\mu_0$ です。
11 章と同じように、分散が既知の場合は正規分布表を、未知の場合は t 分布表を使用します。では問題いきましょー！いえぇー！！！！！！！！！！
問題
空調システムの作動状況を調べるために、設定温度を 25 度とし、7 日間にわたって室内温度を測定したところ、次のような結果を得た。
24.2　25.3　26.2　25.7　24.4　25.1　25.6
このシステムが正しく働いているか堂かを 5%の有意水準で検定する。
いきます。この場合、帰無仮説H_0：μ=25.0 を　両側対立仮説H_1：$\mu\neq$25.0 に対して検定する。
今$\overline{X}=25.21, s=0.715$　　であるから、
仮定が正しいとすると

$$t=\frac{25.21-25.0}{0.715/\sqrt{7}}=0.777$$　　である。ここで 11 章の図を思い出して欲しい。5%以下のゾーンにこの値が入っていなければ良いのである。この場合 t 分布なので

表を見ると$t_{0.025}(6)=2.447$である。　－2.447<0・777<2.447
なので上２．５％、下２．５％のどちらにも入らず間に入っているのでおｋ。
したがって、有意水準5%で帰無仮説は棄却できない。
次は片側。
問題
英語の特別講義の効果を調べるため、10 名の受講者に対して講義の前後で英語の試験を行い、二つの試験の得点差を求めたところ、
-1 3 4 5 3 0 7 4 2 -2 の 10 の標本が得られた。
この講義の効果について調べる。得点差は正規分布 N(μ, σ^2)に従うとし、
帰無仮説H_0 : μ=0 を　右方側対立仮説H_1 : μ>0 に対して有意水準 5%で検定する。
ここで$\overline{X}=2.5, s=2.8$なので

$t=\dfrac{2.5-0.0}{2.8/\sqrt{10}}$=2.82　となる。上と同じようにやってみる。

$t_{0.05}(9)=1.833$で　1.833<2.82　であるから
有意水準5%でH_0は棄却される。すなわちH_1が採択される。よって、この講義には効果があったと認められる。

こんな感じです。左側もおんなじです。
このあとχ2 乗検定とかもあるんですが、過去問をみたところ出てないので省略します。出たらごめんなさい。不安な人は教科書を読んでおいてください。
シケプリは以上となります。長くてごめんなさい。公式だけを素直にまとめることもできたんですが、80%自分の理解のためにやってたので許してください。ただよく読めば理解できるようなものにちゃんとなったと思います。
これを読んで理解したら、時間があれば過去問や練習問題にも目を通して問題が解けるようになってください。同じ問題が繰り返し出ていたりしますので・・・
理Ｉ15 組に答えつきの問題がありました。本とは書いちゃだめなんでしょうけど活用するのも手かもしれませんね。。。基礎統計で皆さんが見事単位をとれますよう
お祈りしています・・・May the unit be with you…　by　Hori-ken