

SDP 入試について

桃山学院大学スーパーデータサイエンスプログラム(SDP)では、社会、経営・経済、スポーツ、言語、法律、音楽、医薬など、様々な分野に興味をもち、それらの各分野で起こる諸問題の解決に、数学や情報の知識を基本としたデータサイエンスの手法を身につけ、即戦力となるデータサイエンティストを目指す受験生を広く募集しています(詳細は、<https://www.andrew.ac.jp/faculty/sdp/>を参照してください)。

総合型選抜の一環である本入試では、

- ・ 与えられた問題をしっかりと解決することができる人
- ・ 数学や情報の知識を用いて、データサイエンスを学ぶことに興味がある人

を募集するために、大学入試としては例のない

数学の予告問題制度

を導入します。

公開される予告問題は、

必修問題1題、応用問題5題

です。試験当日は、

必修問題を必ず解答

してもらい、更に

応用問題の5題の中から1題を解答

してもらいます。その他に、

教科書の章末問題程度の問題を解答

してもらいます。

必修問題、および応用問題に関しては、

出題の意図や考え方を用意

していますので、

頑張って試験当日までに問題を解決

して、本番に臨んでください。

みなさんと4月からデータサイエンスを学べることを楽しみにしています。

必修問題

xy 平面上に3点 $P_1(-1,3), P_2(0,1), P_3(1,5)$ を取り, $P_i(i=1,2,3)$ の x 座標を $x_i(i=1,2,3)$, y 座標を $y_i(i=1,2,3)$ とそれぞれ書くことにする. さらに実数 a, b と x_i に対して $ax_i + b$ の計算を行い, その結果を $\hat{y}_i(i=1,2,3)$ とする. このとき, 次の各問いに答えよ.

- (1) 次式で定義される L を a, b を用いて表せ.

$$L = \sum_{i=1}^3 (\hat{y}_i - y_i)^2$$

- (2) b をある値で固定すると, L は a の関数 $L(a)$ と考えることができる. このとき, $L(a)$ を最小にする条件1を a, b を用いた式で表せ. 同じように, a をある値で固定すると, L は b の関数 $L(b)$ と考えることができる. このとき, $L(b)$ を最小にする条件2を a, b を用いた式で表せ.

- (3) 上記で求めた条件1と条件2を同時に満たす a, b の値を求め, さらに y_i の平均値と b との関係を答えよ.

<必修問題の考え方>

大学で学ぶデータサイエンスの中でも非常に強力な分析手法の一つが

回帰分析

という分析手法である。

この分析手法は、多くの企業でも実際に利用されている。簡単な具体的例としては、

- ・ 気温による、ビールやかき氷の売り上げ予想
- ・ 駅からの距離による、家賃相場の予想
- ・ 走行距離からみる、中古車の金額予想

などが挙げられる。これらの例は、回帰分析の中でも単回帰分析といって、誤差の混入を前提としながら、一つの変数で、もう一つの変数を説明するものである。ここで、

説明される変数(ビールの売上や家賃相場など)は、目的変数

とよばれ、一方で

説明する変数(気温や駅からの距離など)は、説明変数

とよばれている。

実は、この必修問題で得られる直線は、

回帰直線

とよばれるものであり、この直線を求めることによって、上述したさまざまな問題の予想が可能になるのである。

この問題のポイントは、2つある。一つ目(前半)のポイントは、

座標平面で表された状態を把握し、 L の式を導く

ことができるかである。

題意の状態は、 xy 平面上に、与えられた3点が配置され、次に直線 $l_{a,b}: y = ax + b$ も書き込まれ、その後、各点から直線に向かって、 y 軸と平行な直線を引いた状態である。平行な直線それぞれと直線 $l_{a,b}$ との交点がそれぞれ $(x_i, \tilde{y}_i)$ ($i = 1, 2, 3$)となり、 L は、 P_i と $(x_i, \tilde{y}_i)$ ($i = 1, 2, 3$)の距離の2乗の合計である。ここから、 L の形を求めればよい。

二つ目(後半)のポイントは、

関数の最大値・最小値を求める

ことである。この問題は、2 つの変数が動く場合の最大値・最小値を考える問題であり、そのままでは難しい問題となるが、問題文にあるように

2 つの変数の関数の 1 つの変数を止めて、1 変数の関数として考える

ことができれば、数学 I で学ぶ、

2 次関数の最大値・最小値を求める問題

にたちまち変わるのである。

もともと、一般論としては、「最大値・最小値を求める問題」は微分を使う必要がでてくる。その詳細を学ぶ講義も SDP では用意している。

さて、数学的には、座標平面上の状況を数式化した上で目的とする値 L を最小となる値を求めているわけであるが、この L の意味をもう一度考え直してみる。

x という説明変数で、目的変数 y を説明することを考える。誤差の混入を前提とするとすべての目的変数を厳密に表現する必要がないので、近似する関数を考えれば良いことになる。そして、近似する関数は数学的な取り扱いが簡単なものを選び、その上で誤差が最小となるように設定すればよいことになる。

この問題については、近似する関数として 1 次関数を選び、予測値と実測値の差の 2 乗の合計で誤差の大きさを測り、2 乗の合計が最小となる 1 次関数がもっとも状況を近似する関数だとして選択していることとなる。

応用問題1

3人の仲良しが地元の夏祭りに出かけた。夏祭り会場の入口付近に「いろは商店」、奥の方に「にほへ商店」があり、2つの店の片方で「くじ A」、もう一方で「くじ B」が販売されている。2つのくじは以下の通りである。

・くじ A:くじ 10 本のうち6本が当たり

・くじ B:くじ 10 本のうち 2 本が当たり

ただし、2つの店を外から見ると、どちらのくじを販売しているかを知ることができない。

2 人が順番にくじを 1 本引き、以下の結果となった。

・1 人目:「にほへ商店」でくじを引きはずれ

・2 人目:「いろは商店」でくじを引き当たり

このとき、3 人目は「いろは商店」と「にほへ商店」のどちらの店のくじを引くべきだろうか？ただし、この2人以外にくじを引いた客はいないものとする。また、2つの店のどちらが「くじ A」を売るかは同様に確からしいとし、さらに、引いたくじは元に戻さないものとする。

応用問題2

ある授業では講義後に、復習のための勉強会を図書館とコモンズで開催している。昨年度の経験から、勉強会で利用する場所について、以下のことがわかっている。

調査結果：

前回図書館を使った学生は 0.9 の確率で次回も図書館を使い、0.1 の確率で次回はコモンズに場所を変える。

前回コモンズを使った学生は 0.7 の確率で次回もコモンズを使い、0.3 の確率で次回は図書館に場所を変える。

さて、同様の授業を今年度も行う予定であるが、授業の担当講師は上記の結果を用いて、事前に任意の n 回目の講義後の勉強会の人数を予測したいと考えている。そこで、1回目の授業を始める前にアンケート調査を行った。その結果、1回目前において図書館を利用した学生は 30 人、コモンズを利用した学生は 70 人であった。

n 回目の講義後に図書館を利用した学生数を x_n 、コモンズを利用した学生を y_n として、以下の各問いにすべて答えよ。ただし、すべての講義において欠席者はおらず、全員が図書館かコモンズかのどちらかで必ず勉強会に出席するものとする。

- (1) 1 回目の講義後のそれぞれの勉強会の人数 x_1, y_1 を具体的に求めよ。
- (2) 2 回目の講義後のそれぞれの勉強会の人数 x_2, y_2 を具体的に求めよ。ただし、人数は小数第 1 位を四捨五入して整数で解答せよ。
- (3) x_n と y_n を x_{n-1} と y_{n-1} を用いて表せ。
- (4) n 回目の講義後のそれぞれの勉強会の人数 x_n, y_n を求めよ。

応用問題3

あるジュースを生産している企業がある. そのジュースを買ってくれる問屋との契約内容は生産したものはすべて購入してくれるが, その単価は販売量 x (kL)によって変化する契約であるとする.

x (kL)を生産するときの1(kL)あたりの費用 f が $f(x) = x^2 - 8x + 57 + \frac{2}{x}$ (万円/kL)であり, x (kL)を販売するときの1(kL)あたりの単価 P が $P = 45 - 0.5x$ (万円)であるとき, 以下の問いにすべて答えよ.

- (1) 生産量が x (kL)のときの売上高を x の式で表せ.
- (2) 売上高が最大となるときの生産量 x (kL)を求めよ.
- (3) 生産量が x (kL)のときの総費用を x の式で表せ.
- (4) 生産量が x (kL)のときの利益を x の式で表せ.
- (5) 利益が最大となる生産量 x (kL)を求めよ.

応用問題4

次の各問いに答えよ. 必要な場合は, $\log_{10} 2$ の近似値として 0.3010, $\log_{10} 3$ の近似値として 0.4771 を利用して良い.

- (1) $\log_{10} 5$ が無理数であることを証明せよ.
- (2) 100 万円を投資することについて試算した. 2つのビジネスについて検討したところ, A というビジネスは 6 年で 150 万円になり, B というビジネスは 10 年で 200万円になるという. どちらのビジネスも1年ごとに投資額は一定の割合で増加しているものとする. 1 年あたりの増加率が高いのはどちらのビジネスの方か.

応用問題5

お祭り会場のある店では「50%の確率でかわいいぬいぐるみが当たる!？」とのメッセージがあり、「番号が書かれたくじを引いて、出た番号の末尾が偶数の場合は『当たり』で、好きなぬいぐるみを選ぶことができる。出た番号の末尾が奇数の場合は『はずれ』で、残念賞となる」との説明を受けた。

いま、この店でくじを引いてかわいいぬいぐるみを当てたいので、その参考として、自分が引く前にこのくじを引いた50人の結果を観察することにした。その結果、50人はそれぞれ一度ずつくじを引いたが、ぬいぐるみが当たった人は11人であった。

そこで、お店のいう、「50%の確率で当たる」というのは本当だろうか？有意水準5%で検定することによって判断せよ。ただし、正規分布による近似の方法を利用してもよい。その際には、別紙の標準正規分布の数表を利用せよ。

<応用問題のポイント>

データサイエンスに限らず、問題解決の際には、対象を数学的に記述し、表現された数学的内容を数学的処理することを通じて現状把握や未来予測が行われる。言うならば、数学的な模型を作ってその模型について調べることによって問題解決をするわけであるが、このような「数学的な模型」のことを「**数理モデル**」という。

ここで示した応用問題は、すべて数学的には高等学校数学ⅠAⅡBの範囲の数学ではあるが、色々な形で数理モデル化の実例でもある。数学の応用という視点だけでなく、問題解決の手法という観点もこめて問題に取り組んでいただきたい。

<応用問題1の考え方>

データサイエンスでは、観測結果を反映させることで、モデル内で設定している確率に修正をかけ、より現実に近い数理モデルにしていくことが多々ある。この時有用なのが

ベイズの定理の利用

である。

たとえば皆さんは、

以前は迷惑メール扱いされていなかったメールが

いつのまにか迷惑メールとして振り分けられている

ことに気づいたことはないだろうか？実は、この背景には、

ベイズの定理の考え方が利用されている

のである。

問題に戻るとこの問題のポイントは、

当たりやすいくじを売る店が、入口付近と奥の方のどちらに配置されているかの確率を

観測事実にベイズの定理を応用して再計算

することである。その確率がわかれば、後はそれぞれの場合での3人目が当たる確率を求めればよい。

いま、以下のように事象を考えよう。

事象 X_1 : 入口付近に当たりやすいくじを売る店が配置されている

事象 X_2 : 奥の方に当たりやすいくじを売る店が配置されている

事象 Y : 1人目が奥の方の店でくじを引いてはずれ、2人目が入口付近の店でくじを引いてあたる

問題文より, 以下の確率が事前の情報としてわかる.

$$P(X_1), P(X_2), P(Y|X_1), P(Y|X_2)$$

大切なので少し言い換えて繰り返すが, 我々が知りたいのは,

「事象 Y が起こったもとで, 事象 X_1 , および事象 X_2 が起こる条件付確率」

であるので, ベイズの定理を使って以下の確率を求めればよいのである.

$$P(X_1|Y), P(X_2|Y)$$

<応用問題2の考え方>

この問題は、未来の予測を可能とする数理モデルの作成に利用されている数理的な手法の一つであるマルコフ過程を基として作成されている。マルコフ過程とは、簡単にいうと、

「現在から未来への状態変化が、これまでどのような状態であったかとは全く無関係に、現在の状態のみから決定される」

ことを確率を使って表現したものである。つまり、

過去は振り返らない、これからが大事や！

を表現したものであるといえる。

例えば、天気予報、人口の推移、コイントスなど、我々の身近なさまざまな場面でマルコフ過程はみられるのである。

この問題のポイントは2つある。一つ目は

遷移する状態を漸化式で記述する

ことである。つまり、

(1) の問題の状況を式としてモデル化できるか

がポイントになる。

1回目と2回目の関係式が書けると、まったく同様に考えて n 回目と $n+1$ 回目の関係式をモデル化することができる。結局は、

$n+1$ 回目の状態は、 n 回目の状態によって決まる

となり、まさにこれはマルコフ過程であるといえる。

二つ目のポイントは、この漸化式を変形して最初の状態から n 回目の状態を計算する式を導くことである。 $x_n + y_n = 100$ を使うことにより、確率漸化式に帰着できる問題ではあるが、確率漸化式以外の解き方を使っても解くことができる。

<応用問題3の考え方>

データサイエンスに限らず、数理モデルによる問題解決でよく使われる方法が、目的とする値を最大(あるいは最小)にする状況を探すことである。そして、最大(あるいは最小)となる状況を探すための基本的な手段が

微分の応用

である。

この問題のポイントは二つある。一つ目は

文章で示された状況の数理モデル化

である。

まず、「売上」「総費用」「利益」は、基本的に、

$$\text{売上} = \text{単価} \times \text{販売量}$$

$$\text{総費用} = \text{単位あたりの費用} \times \text{生産量}$$

$$\text{利益} = \text{売上} - \text{総費用}$$

としてモデル化する。もちろん、現実社会では状況に応じた値引き等もあるわけだが、最初に構築するモデルは、それらの細かなことは切り落として構築する。もちろん、必要に応じてもっと複雑なモデル化を行うことはあるが、それはこの問題の範囲ではない。

二つ目のポイントは

構築した数式の最大を与える状況を決定する

である。先述した様に、このような場合は微分の考え方を使うことによって、統一的に扱うことができる。

<応用問題4の考え方>

これまでの応用問題でもわかるように、データサイエンスでは、日常や社会の現象をモデル化し、そのモデルを数学や確率・統計、情報の知識を用いて解き、評価することが重要である。その中では、ベイズの定理(応用問題1)、マルコフ過程(応用問題2)や微分・積分(応用問題3)などが利用されているが、本問題で取り上げる

指数・対数が利用される

ことも非常に多い。指数関数・対数関数はそれぞれ数学的に利用しやすい性質を持つため、数理モデルを分析するための基本的な道具である。その「利用しやすい性質」は高等学校の授業の中で示されているが、「利用しやすさ」の実感はないかもしれない。

この問題のポイントも二つある。

$\log_{10} 2$ の値が無理数であることや $\log_{10} 2$ と $\log_{10} 5$ の関係式を活用できるのか

という問題が、一つ目のポイントである。言い換えると、対数関数の性質を把握できているかがポイントである。

二つ目のポイントは、ビジネスの成長に関する数理モデル化である。

ビジネスの成長に関しては、年に一定の割合で増加するとして n 年で a 倍となるなら、1年での増加率は $a^{1/n}$ であると言える。問題について言うと、AB 両方のビジネスの1年での増加率を求めた上で、それを比較すればよいが、その比較の際にも対数関数の性質を使うことができる。

<応用問題5の考え方>

大学での学びは、高校までの学びと違って、

「なぜならばに答える」

ものである。そのなぜは、

つねに我々の日常

にある。

数学は、科学のなぜに回答を与える学問

であり、

データサイエンスは、その数学を武器に、

より複雑な日常のなぜに回答する学問

である。この問題は、まさに出題者が実際に秋祭りで体験したものをベースに考えたものであり、その解決方法を数学やデータサイエンスの手法を使って知ることができるのである。

この問題のポイントは、

仮説検定の基本的な手法をマスターしているかどうか

である。

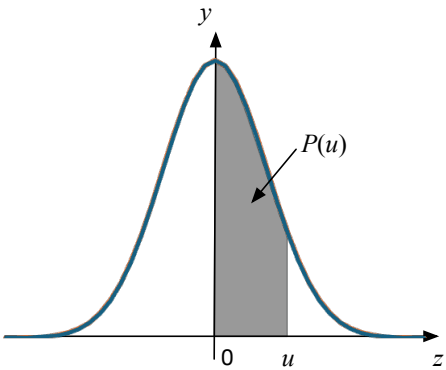
くじ引きの毎回の試行は、ベルヌーイ試行($p = 0.5$)であるので、 $n = 50$ は十分大きいと考えて、当たる回数 S は、二項分布 $B(n, p)$ に従うが、正規分布 $N(np, np(1 - p))$ で近似することができる。このとき、

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

は、標準正規分布 $N(0, 1)$ に従う。

あとは、有意水準 $\alpha = 0.05$ の条件のもと、上の実現値が、棄却域に入っているか否かを教科書の標準正規分布の数表を頼りに確認すればよい。

正規分布表



<i>u</i>	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