# **2020年度 統計学 期末レポート （締切: 8月18日（火）23:59）**

担当: 社会情報科学部 山本 岳洋（[t.yamamoto@sis.u-hyogo.ac.jp](mailto:t.yamamoto@sis.u-hyogo.ac.jp)）

講義ページ: <https://tyamamot.github.io/statistics-2020/>

* 締切: 8月18日（火） 23:59
* ユニバーサルパスポートのクラスプロファイルの課題に「統計学 期末レポート」という  
  課題があるので，そこからファイルを提出
* 発展課題に取り組むのは任意です．取り組んだ人には加点します
  + 発展課題でExcelや（Googleスプレッドシート）を作成した人は  
    そのファイルについても提出してください
* GoogleドキュメントやMarkdown用のファイルはユニパとSlackに掲載しています
* 受講環境等の理由でレポートを作成できない人は相談にのりますので，  
  山本までメールやslackで事前に連絡してください．

本レポートを解答するにあたり，必要であれば下記の確率を用いよ．

確率変数が標準正規分布 に従うとき，以下が知られている.

問1.（連続型の確率変数・統計検定2級の過去問より）

連続型の確率変数 の確率密度関数を

とする．このとき，以下の問いa), b) に答えよ．

* 出典: 日本統計学会，「統計検定2級」 2016年11月問10より一部改題して出題

1. の期待値 を求めよ．答えだけでなく途中の考え方についても示せ．
2. の分散を求めよ．答えだけでなく途中の考え方についても示せ．

問2. （正規分布）  
全国の高校生に対して数学のテストを実施したところ，平均 70 点，標準偏差 10 点であった．全国の高校生のこのテストの点数が正規分布 に従うと仮定したとき，以下の問い a), b) に答えよ．なお，この問題はExcelや標準正規分布表は用いず，答えだけでなく途中の考え方についても示すこと．

1. このテストで 60 点 以上 70 点 以下 をとった学生の割合は何パーセントいるか答えよ．
2. このテストで上位5%以内に入るには何点以上取る必要があるか答えよ．答えは整数で良い．
3. このテストで95点をとった学生は，上位 0.5% 層に含まれるといえるか答えよ．

問3.（標本不偏分散）

母集団から無作為復元抽出された標本の分散について考える．ExcelのVAR()関数をはじめ，統計を扱ったソフトウェアやプログラミング言語では，デフォルトでは標本分散ではなく標本不偏分散が計算されることが多い．講義資料や補足資料などを参考に，標本分散ではなく標本不偏分散が用いられる理由についてあなたなりの言葉でまとめよ．必要があれば数式等用いてよい．

問4.（母平均の信頼区間・統計検定2級の過去問より）

NHKによる2015年国民生活時間調査によると，無作為に抽出された小学生333人の平日の睡眠時間の平均は8時間35分，標準偏差は1時間2分であった．このとき，以下の問いa)-c)に答えよ

* 出典: 日本統計学会，「統計検定2級」 2017年6月 問12より一部改題して出題

1. 小学生333人の平日の平均睡眠時間の標準誤差を答えよ．単位は分とし，平方根はそのままでよい．すなわち，回答は 　のように答えてよい．また，回答は答えだけでよい．
2. 全国の小学生の平均睡眠時間（母平均）を とする．の90%信頼区間を求めよ．単位は分とし，小数点は四捨五入し整数で回答してよい．また，回答は答えだけでよい．
3. 「の95%信頼区間の幅は，90%信頼区間よりも狭い」 この主張が正しいか誤っているか答えよ．誤っている場合は，理由についても示せ．

問5.（母比率の信頼区間）

あなたは，ある機械で製造される製品が不良品である割合（不良率）を調査した報告書を読んでいる．その報告書を確認したところ，「標本調査をした結果，標本における不良品率は20% であり，この機械で製造される製品の不良率の95%信頼区間は12.16 % 以上 27.84 % 以下である」との記載があった．この報告書を作成する際に行った標本調査のサンプルサイズを求めよ．答えだけでなく途中の考え方についても示すこと．

問6. （仮説検定）

平均点を 60 点として作成した試験がある．100 人 の学生を無作為抽出して回答してもらった結果，平均 60.8 点，標準偏差 8 点 であった．このテストの 平均点は 60 点ではない といえるだろうか．帰無仮説を母平均 点 として 有意水準 5 % で両側検定することを考える．これに関して，以下の問い a) – d) に答えよ． なお，母集団の点数は正規分布に従うと考えてよく，母標準偏差を標本標準偏差で代用してよい．

1. 帰無仮説と対立仮説を示せ．
2. この問題における第1種の過誤とはどのような誤りか説明せよ．
3. 同様に，この問題における第2種の過誤とはどのような誤りか説明せよ．
4. 帰無仮説が正しいと仮定したとき，100 人 の学生を無作為復元抽出して得られる標本の標本平均が従う分布とその期待値と分散を示せ．
5. 実際に仮説検定を行え．結論だけでなく途中の考え方についても示すこと．

参考文献:   
（レポートを書くにあたり，参考にした書籍やウェブページがあれば，最後に記載すること）

以降は発展課題です．余裕がある人だけ，取り組めるだけ取り組んでください．

発展課題その1. （大数の法則）

大数の法則を言葉で説明すると，サンプルサイズ を大きくすれば，標本平均  が 母平均 とほぼ一致する確率は限りなく1に近づく，という意味である．そこで，二項分布に従う母集団からサンプルサイズの標本抽出を行い標本平均を求める問題を題材に大数の法則をExcelによるシミュレーションで確認してみよう．これは，当選確率50%のくじを復元抽出で回引いたとき，が大きくなるほど引いたくじの当選確率が50%に近づくことを確認することでもある．

具体的には，以下の手順a.-c.に従い，Excelによるシミュレーションを行い結果を報告せよ．

1. この母集団からサンプルサイズ の標本を無作為復元抽出することを考えよう．Excelでは， =RANDBETWEEN(0,1) と入力すると，0か1が等しい確率で出力される．これは，1なら当選，0ならはずれとみなせる．つまり，=RANDBETWEEN(0,1) の結果は二項分布 に従う母集団から要素を1つ無作為復元抽出したと見なせる．  
   いま，サンプルサイズ3の標本 を抽出し，その標本平均 を求めることを考える． これは，=RANDBETWEEN(0,1) を3つのセルに入力し，その平均（つまり足して3で割る）を求めればよい． たとえば，0.3333 という結果となれば，それは3回くじを引いたときの当選回数は1回であった，つまり，標本中の当選確率（標本比率）は1/3 であった，ということを意味している．  
   （なお，Excelで =RANDBETWEEN(0,1) のような，入力する度に異なる値が出力されるような関数を入力すると，Excelで操作をする度に値が再計算されます．値がコロコロ変わって気持ち悪い，という人は出力結果をコピーして，別の場所に，「編集→形式を選択してペースト→数値」と選びペーストすると，値だけがコピーされ，結果が固定されます．また，逆に=RANDBETWEEN(0,1)を打っても同じ値しか出力されない，という人はF9を押してみてください．）
2. さて，今説明した手順a. を の5通りで試し，それぞれについて標本平均を求め，標本平均がどのように変わっていくかを確認せよ．が大きいとき，標本平均は母平均 に近い値となっているはずである．もちろん，が小さいときも標本平均が母平均と一致する可能性はあるので， のとき標本平均が0.5となっていても間違いではない．なお，余裕があれば，もっと細かい粒度の （たとえば，10ずつ増やしていくなど）で標本平均がどのように推移するかを調べても良い．このレポートにはそれぞれのにおける標本平均を記し，作成したExcelをユニバーサルパスポート等で提出すること．

発展課題その2. （中心極限定理）

発展課題その1で大数の法則をシミュレーションで確認した．いま，中心極限定理についても同様のことをシミュレーションで確認してみよう．具体的には，発展課題その1の手順b.にて，サンプルサイズ の標本平均を求めることを回行い，サンプルサイズ における標本平均のヒストグラムをこの個のデータから描くこと考える．これをサンプルサイズ を3, 10, 500の3通りで行い，が大きいほどヒストグラムの形状が正規分布に近づく様子を確認せよ．（といっても二項分布自体が対称性を持っているため，が小さくとも正規分布のような形状になる可能性は高いです）．なお，は適宜各自で設定した数を用いてよい（たとえば，程度としてみる）．また，二項分布の成功確率を0.5ではなく，異なる値にしてみてもよい．

発展課題その3. （母比率の区間推定）  
東京と大阪の住人について，無作為抽出による標本調査により新型コロナウィルスの抗体保有率を調査したところ，東京については1,971人中2人が，大阪については2970人中5人が抗体保有者であった．このことについて，以下の問いa), b)に答えよ．

出典: 朝日新聞デジタル2020年6月16日「抗体検査の陽性率発表　東京0.1%、大阪0.17%」

* https://www.asahi.com/articles/ASN6J3H1YN6JULBJ001.html

1. 講義や教科書で扱った計算式に従い，母比率，すなわち東京の住民の抗体保有率95%信頼区間と大阪の住民の抗体保有率95%信頼区間をそれぞれ求めてみよ．単位はパーセントで表し，小数点をどこまで表すかは自由に設定してよい．
2. で求めた信頼区間に対して不思議に思った点があれば述べよ．また，講義資料や教科書に記載されている，母比率の信頼区間の計算式の導出を参照しながら，その不思議なことが起こった理由についてあなたなりに考察してみよ．

発展課題その4.（第1種の過誤，第2種の過誤）

仮説検定における第1種の過誤と第2種の過誤について，下記に示すようなExcelを用いたシミュレーションで理解を深めよ．作成したExcelをユニバーサルパスポートで提出するなどし，自分で課題を行ったことが分かるようにすれば自由に回答してよい．

a) （第1種の過誤） 母集団が正規分布 に従うとしよう．ここからサンプルサイズ30の標本をExcelで無作為復元抽出し，標本平均を計算する．この標本平均をもとに，帰無仮説: 母平均 という仮説を立て，有意水準 5% で両側検定を行う．今説明した，標本抽出から検定までを100回繰り返し，100回中帰無仮説が棄却された回数を確認せよ．つまり，このシミュレーションは帰無仮説が正しい（母平均 ）にもかかわらず，帰無仮説を棄却してしまった回数を表す．

b) (第2種の過誤) 母集団が今度は正規分布 に従うとしよう．ここからサンプルサイズ30の標本をExcelで無作為復元抽出し，標本を用意する．この標本の標本平均をもとに，帰無仮説: 母平均 という仮説を立て，有意水準 5% で両側検定を行う．この手続きを100回繰り返し，有意でない（つまり，値が有意水準より大きい）とき，帰無仮説を採択すると結論づけた場合，100回中，帰無仮説が採択された回数を確認せよ．つまり，このシミュレーションは帰無仮説が正しくない（母平均 という仮説は今回の例は正しくない）にも関わらず，帰無仮説を採択してしまった回数を表す．

* [補足] 講義中でも説明しまたが，有意水準しか考慮しない一般的な仮説検定において，「有意でない→帰無仮説を採択」 は間違いです．この発展課題の目的は，第2種の過誤が起こる確率と有意水準αが異なることを確認することが目的です（なので，有意でない→帰無仮説を採択と結論づけてはいけない）．

c) さらに余裕があれば，サンプルサイズを増やすと第1種の過誤と第2種の過誤がどのように変化するかも確認してみよ．

この発展課題においてExcel（Googleスプレッドシート）で使いそうな関数を補足しておきます．  
（AVERAGEやSQRTも使います）

* 正規分布 からの無作為抽出: =NORM.INV(RAND(), , )
* 母平均の両側検定において，値が有意水準を下回るかどうかは，結局のところ が であるかを判定すればよいですね．たとえば，セルB2にこの の値があるとき， =IF(ABS(B2)>=1.96, 1, 0) と入力すると，セルB2の値の絶対値が1.96以上のとき1，そうでないとき0が出力されます．この結果を用いると，帰無仮説を棄却するのか（この場合1），そうでないのか（この場合は0）を簡単に判定できます

以上．