統計教程

— モデルによる予測 —

Idiot

2022 年 12 月 9 日

第 1 章

統計モデル

この章ではついにデータが登場する。データは母集団から無作為抽出によって得られた数値であるとする。データを大文字の $X_1, X_2, \cdots, X_n$ とし、モデルからサンプリングした確率変数を小文字の $x_1, x_2, \cdots, x_n$ とする。統計モデルはデータの出現頻度や統計量などの出現区間などを予測する。まず、その予測可能なことについて列挙する。モデルとデータが異なる場合つまり、データの出現頻度をデータが予測できない場合に生じることについて説明する。

1.1　正規分布を含んだ統計モデル

次の 3 つを仮定したモデルを正規モデルと呼ぶ。

(1) 独立同分布
(2) その分布は、正規分布
(3) 正規分布の母数 (平均と分散) はそれぞれ μ, σ^2。

この正規モデルを $M(\mu, \sigma^2)$ と書く。σ^2 をある特定の値にしたときのモデルを $M(\mu)$ とし、μ を特定の値にしたモデルを $M(\sigma^2)$ とする。モデルに対してある特定の値を当てはめることができるのは、推測したい母集団について、すでに標本を得て、標本分布がわかっている場合である。この母集団を細分化した母集団について、推測した母数を使ったモデルで推測しても良いかを検討するときに、元の母集団に関する知識を使う。
母集団から無作為抽出した標本 (データの入った集合) を元にモデルを構築する。具体的には、$\mu_{ML} = \bar{X}, \sigma^2_{ML} = \frac{1}{n}\sum(X_i - \bar{X})$ とし、統計モデル $M(\mu_{ML}, \sigma^2_{ML})$ を最尤モデルと呼ぶ。

> **■最尤モデルが最も良い予測をするかはわからない**
>
> 赤池も、最尤推定量がデータを推測する上で良い推定量になっているかの根拠にならないことを指摘している**??**。

> R.A.Fisher の研究により、観測データ x が実際に $p(x|a)$ の形の確率分布に従って発生するとき、最尤法が優れた特性を示すことが示された。しかし、応用の場面では、データを生み出す確率的な構造が完全に分かっていることは無いから、Fisher の議論は、最尤法の実上用の根拠を与えない。

以下では、$M(\mu)$ による予測について説明する。

1.1.1 データが出現しやすい区間

ある決められた確率でデータが出現するとモデルが予測する区間を予測区間という。割合として、よく使われる 95% を設定したものを 95% 予測区間という。正規分布を含んだモデル $M(\mu)$ において、予測区間は比較的簡単に求めることができる。具体的には、正規分布の規格化を行い、標準正規分布に従うように変換を行い、$\frac{x-\mu}{\sigma}$ であるので、予測区間は、

$$\mu - z_{0.05}\sigma < x < \mu + z_{0.05}\sigma$$

である。この範囲に 95% のデータが生じることをモデルが予測する。実際にそのようになるかは不明であり、予測であることを意識した方が良い。
68% の確率でデータを含むと予測する区間が求められる。

$$\mu - \sigma < x < \mu + \sigma$$

1.1.2 母集団の標本が指数分布的に分布していた場合

母集団の分布形と統計モデルに含まれている確率分布関数が著しく異なる場合を考える。母集団分布として、指数分布を仮定しする。これは、自然から指数分布的なデータが得られたときのことを想定している。これを予測するモデルを正規分布の仮定されたモデルとする。
最尤モデルは、$M(\mu_{ML}, \sigma^2_{ML})$ である。ここから、

$$\mu_{ML} - \sigma_{ML} < x < \mu_{ML} + \sigma_{ML}$$

が 68% 予測区間になる。言い換えれば、標準偏差の間に、サンプルの平均が入る確率が 68% であることをモデルが予測している。このことを数値シミュレーションにより確かめる。指数分布からランダムサンプリングを行い、無作為抽出によりサンプルサイズ 10^6 の標本を得たとする。サンプルが上記の区間に入っている割合を計算する。

```
1 N = 10**6
2 sample = expon.rvs(scale=10,size=N)
3 #sample = norm.rvs(loc=0,scale=1,size=N)
4 lambd= np.average(sample)
```

```
5  print(np.average(sample),np.std(sample),np.var(sample))
6
7  mu = np.average(sample)
8  s = np.std(sample)
9
10 a,b = mu-s,mu+s
11 len(sample[np.where( (sample >a) & (sample<b) )])/N
```

この結果、期待していた値 68% よりも著しく大きな確率 86% 程度を得る。これは、モデルでは、正規分布を仮定していたが、実際には指数分布的なデータだったために生じる予測の間違いである。以上でわかるのは、統計モデルと実際の母集団が乖離している場合には、標準誤差に間違いが多くなるということである。

エラーバー (SD) から読み取れること

正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本を作り、プロットした (図 1.1.2)。それぞれの分布の右側のエラーバーは、68% 予測区間。標本が正規分布であるときには、68% 予測区間の中におよそ 68% のデータが含まれている。一方で、標本が指数分布であるときは、モデルの予測と乖離する。このことは、図から読み解くことが難しい。

エラーバー (SD) だけが描かれた図を見るとデータに対する印象が変わる (図 1.1.2)。図 1.1.2 には、正規分布と指数分布から得られた標本から、最尤推定を行ったモデル $M(\mu_{ML}, \sigma^2_{ML})$ における 68% 予測区間を描画している。データが正規分布的であるならば、データが区間に入っている割合が予測と一致する。一方で、データが指数分布的であるならば、モデルの予測 (エラーバー) から得られることと、実際のデータは乖離する。エラーバーからは、中央からデータが対称に分布しており、その中に、68% のデータが入っていることを我々は読み取ろうとする。データのばらつきを表すために SD を描いた場合、それが正規分布的な標本でない限り、データのばらつきの意味が伝わりにくくなる。実際の研究活動において、モデルを考えてエラーバーに SD を書いていると断言しがたい。例えば、SD が描かれているのに、検定においては正規分布を仮定しない統計モデルを使っていることが多々ある。これは、データの描画においては、正規分布を仮定しているにもかかわらず、仮説検定においては正規分布をもとにした推測をやめていることを意味する。この場合、著者が何を考えて SD を描いたのかを判断することが難しくなる。

測定の精度の良さを示す指標として SD を書くことがある。この場合、ただ単に SD が小さければ良い計測であることを示す意図がある。

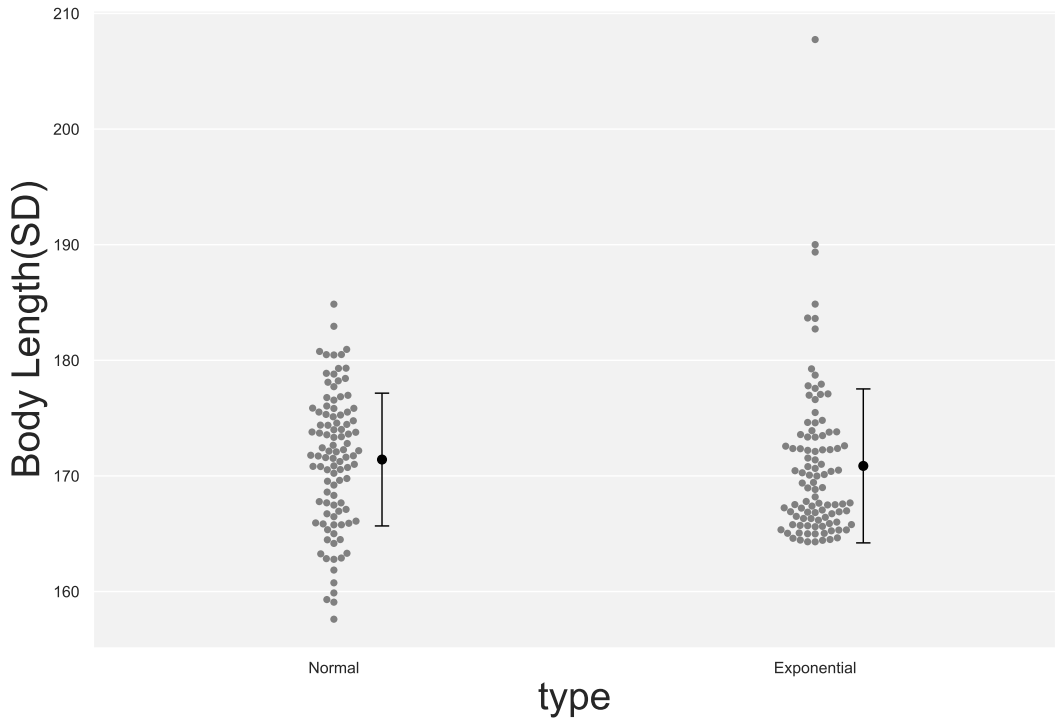

図 1.1　正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本をプロットした。それぞれの右側にあるエラーバーは、正規分布モデルが予測した 68% 予測区間。

1.1.3　平均値が出現する区間

統計モデル $M(\mu)$ では、標本の平均値 $\bar{x}$ が、以下の区間で 95% の確率で見つかる。

$$\mu - z_{0.025}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{0.025}\frac{\sigma}{\sqrt{n}}$$

この統計モデルからサンプリングした標本の標本平均 $\bar{x}$ が 95% の確率で見つかる範囲のことを 95% 信頼区間という。これも、モデルの予測である。
一般的な定義として、要約統計量 ($\bar{x}$ など) が 95% の確率で見つかることを予測する最小の区間を信頼区間という。
信頼区間の中に標本平均が含まれていることは、標本がモデルにり推測可能であることの証拠の一つになる。ただし、予測可能かの判断には、複合的に指標を見る必要がある。

1.2　指数分布を含んだ統計モデル

(1) 独立同分布

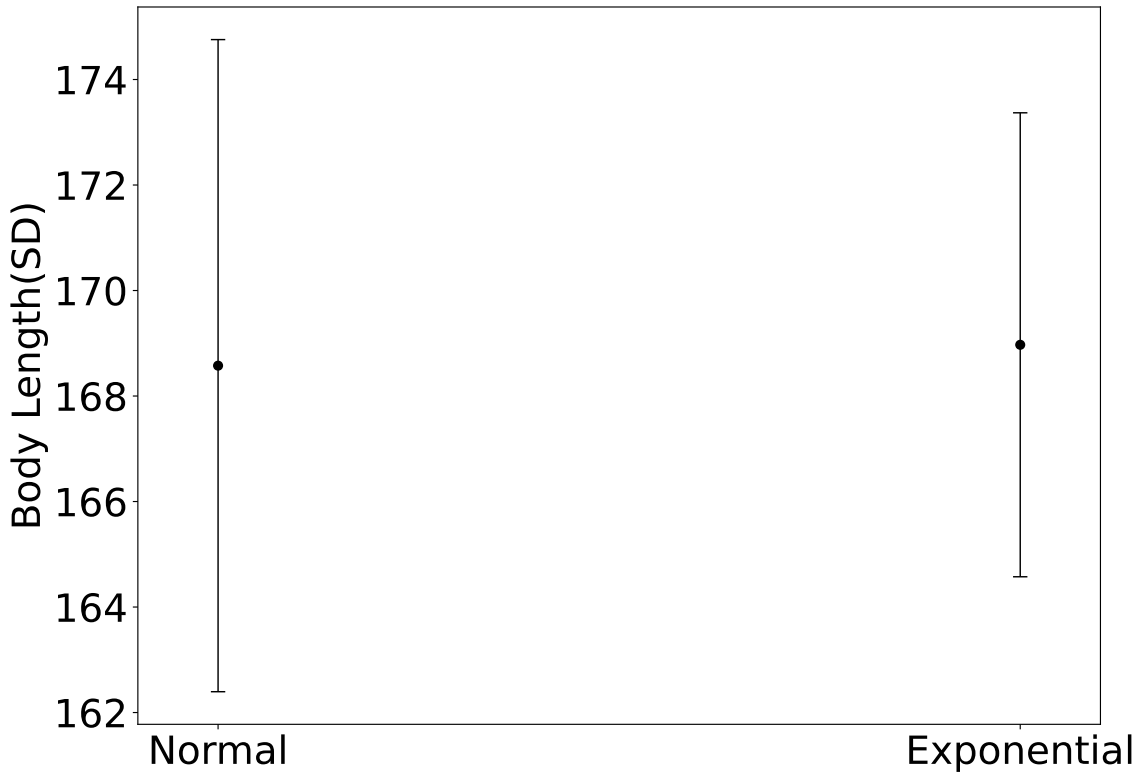

図 1.2 正規分布と指数分布それぞれからサンプルサイズ $N = 100$ の標本を得た。その標本から推測される 68% 予測区間を描画した。

(2) その分布は、指数分布 $(\lambda \exp(-\lambda x))$
(3) 指数分布の母数は λ

このモデルを $M_E(\lambda)$ とする。このモデルの 95% 予測区間は、

$$\frac{1}{\lambda}\log\frac{1}{1-\alpha/2}, \frac{1}{\lambda}\log\frac{\alpha}{2}$$

である。95% 信頼区間は式**??**である。

1.2.1 信頼区間を近似する

95% 信頼区間 (式**??**) を近似的に求める方法がある。中心極限定理を使う。このモデルでは、サンプルの平均および分散は、$E[x] = \frac{1}{\lambda}, Var[x] = \frac{1}{\lambda^2}$ である。このとき、中心極限

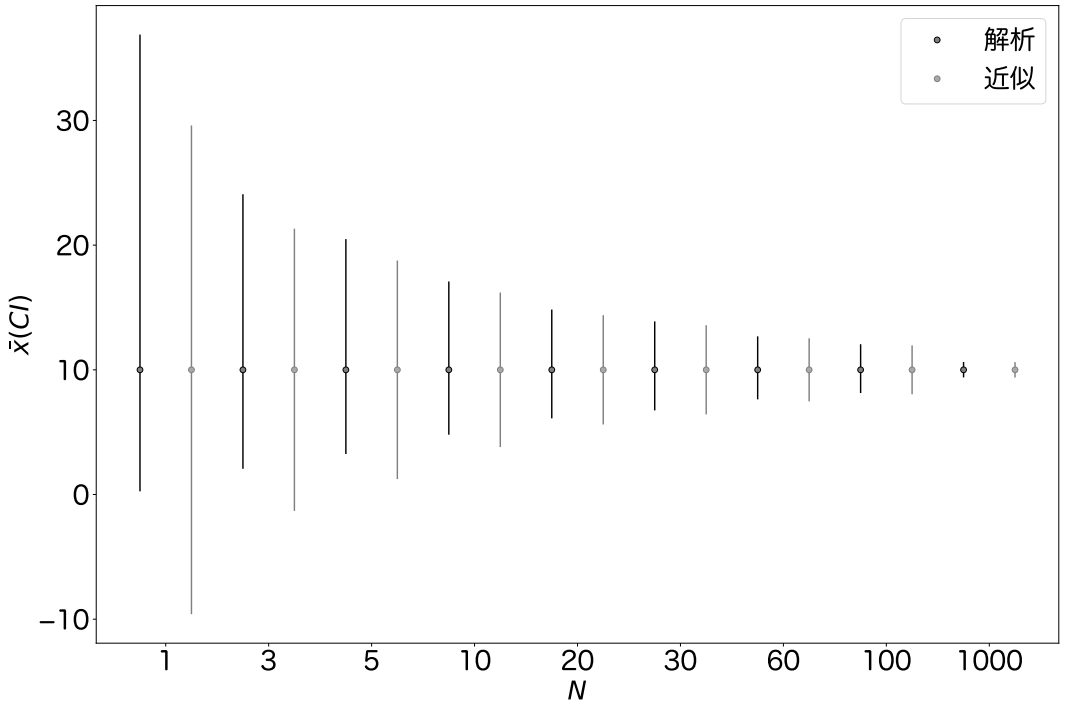

図 1.3　解析的な信頼区間と近似によって求められた信頼区間。$1/\lambda = 10$ とする。横軸にサンプルサイズ。縦軸に、平均値。エラーバーは信頼区間 (CI)。

定理により、$\bar{x} \sim N(E[x], Var[x]/n)$ である。よって、95% 信頼区間は、

$$\frac{1}{\lambda} - z_{0.05}\frac{1}{\sqrt{n}\lambda} < \bar{x} < \frac{1}{\lambda} + z_{0.05}\frac{1}{\sqrt{n}\lambda}$$

である。

解析的に求めた信頼区間と中心極限定理による近似的な信頼区間を比較する (図 1.2.1)。$\lambda = 10$ としたので、平均は全て 10 である。N が小さいと、解析と近似での信頼区間に差が生じている。近似的な信頼区間は、0 よりも小さな値も出現することを予測している。平均 10 の指数モデルでは、平均が 0 以下になることはない。このように、モデルの想定しない区間も信頼区間に含めている。N が大きくなると、解析と近似での信頼区間の違いは少なくなる。これは、中心極限定理の帰結から当然である。

1.3　モデルの選択

データの分布とモデルの分布が一致しているとき、そのモデルの予測も当たりやすい。例えば、データがある値の周りに対象に分布していないにもかかわらず、正規モデルを使う

と、その予想は当たりにくい。モデルの予測が当たると思わせるには、標本分布に関する知識が必要であり、その知識があれば、モデルの予測が当たりやすいと他の人に説得しやすくなる。このことから、まず、標本分布に関する、適当な分布の形（母数を含めた）を探索する必要がある。モデルを選択できるほどのサンプルサイズを集めて、母集団を予測しやすいモデルを決めることで、予測が当たりやすくなり、モデルの予測を信頼しやすくなる。

実際の研究では、実験のコスト増加のことを考慮すると、事前実験により標本分布を調べることは、ほとんど不可能である。そこで、何も事前研究がないなら正規モデルにより推測を行うことを検討するのも良い。モデルがデータを捉えていない場合でも、中心極限定理により信頼区間はそれなりに当たることが多い。ただし、予測区間が現実をよく捉えているかはわからない。

1.3.1 累積分布によるデータとモデルの比較

標本の累積分布のプロット方法について説明する。標本 $X_1, X_2, \cdots, X_n$ を小さいものの順に並び替えたものを、$X_{r(1)}, X_{r(2)}, \cdots, X_{r(n)}$ とする。ここで、$r(j)$ は、j 番目のデータのインデックスを返す関数である。そして、

$$(X_{r(j)}, j/n) \quad (j = 1, 2, \cdots, n)$$

をプロットする。言い換えれば、累積分布は、標本を小さい順に並べたものと、順位をサンプルサイズで割ったもののペアをプロットしたものである。

具体的なコードは次のようになる

```
1 def cummlative_norm(data):
2     sorted_data = np.sort(data) # 順番の並び替え X_{r(j)}
3     x = np.arange(len(data))/len(data) # データ数分のj/n
4     mu_ml,sigma_ml = np.mean(data),np.std(data)
5     predict_cdf = norm(mu_ml,sigma_ml).cdf(sorted_data)
6     return sorted_data,x,predict_cdf
```

累積分布の傾き

累積分布は、データの密集度が高い範囲において、傾きが大きくなり、密集度の小さい範囲では、傾きが小さくなる (図 1.3.1)。

データとモデルの比較

図 1.3.2 図 1.3.2 右側に累積分布を描いておいた。データは、(a) 正規分布、(b) 指数分布、(c) ガンマ分布からそれぞれサンプルサイズ 100 の標本である。それぞれに、正規分布の最尤モデルを重ね書きしておいたので、最尤モデルとの乖離具合が把握できる。(a) では、

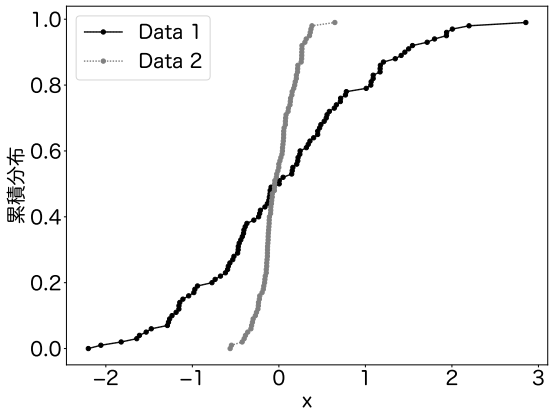

図 1.4　データの累積分布。Data 1 は、正規分布 $N(0,1)$、Data 2 は正規分布 $N(0,0.2)$ からサンプリングした。サンプルサイズは 100

最尤正規モデルとデータが一致している。(b,c) では、最尤正規モデルの曲線上に、データの累積分布の点が乗っていないので、モデルとデータが乖離していることが示唆される。このことから、このようなデータが得られたなら、モデルを再構築したほうが良い。また、サンプルサイズを 30 にした図 1.3.2 では、データが正規分布であっても、正規モデルによって推測することが良いのかはぱっと見では判断しにくく、正規モデルを確信を持って利用しにくくなる。

1.3.2　qq プロットによるデータと正規分布の比較

qq プロットについて説明する。まず上記、$(X_{r(j)}, j/n)$ について、j/n を、$F^{-1}(j/n)$ によって変換する。ここで、累積標準正規分布の逆関数を $F^{-1}(p)$ とする。つまり、

$$(F^{-1}(j/n), X_{r(j)}) \quad (j = 0, 1, \cdots, n)$$

をプロットする。

```
1    def qq_plot(data,ax):
2        sorted_data = np.array(sorted(data))
3        p = np.arange(len(data))/len(data)
4        x_ = norm(0,1).ppf(p)
5        return np.c_[x_,sorted_data]
```

qq プロットを図 1.3.2 図 1.3.2 左側に描いておいた。直線に乗っているデータは、正規モデルの推測が当たりやすいと考えられる。

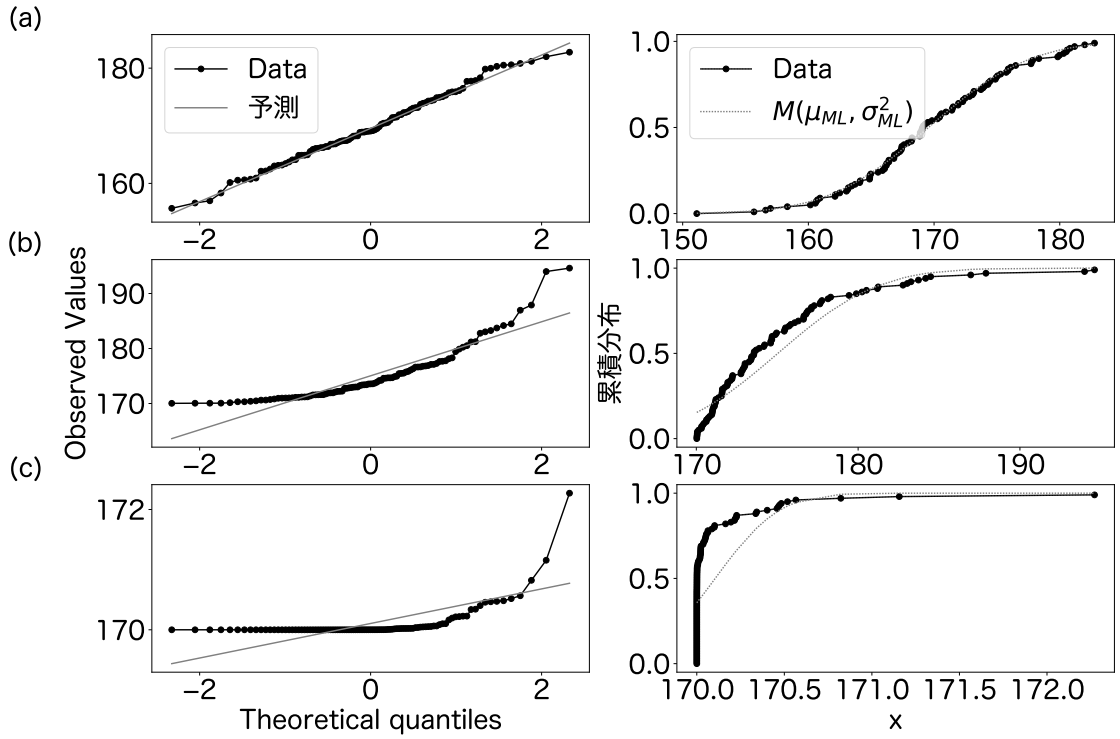

図 1.5　左には qq プロット、右は累積分布と最尤モデルの累積分布。サンプルサイズは 100(a) 正規分布 $N(170, 5.8)$(b) 指数分布 $\lambda = 5.8$(c) ガンマ分布 $s = 0.1$)

1.3.3　AIC の比較

AIC は、対数尤度に対して、データ由来のパラメータ分、ペナルティを与えたものである。AIC が低いモデルは、相対的に良くデータに当てはまるモデルであり、そのモデルからデータが生成されたことを示唆するものではない。また、AIC の差が 10 あったから良いとか悪いとかではなく、AIC が低いものが相対的に良いモデルと判断されがちになるだけである。AIC が小さいから、良い予測をするということは一般にない。
正規モデルのデータに対する AIC を計算する。母数を最尤推定により決定した最尤モデルのパラメータ数は 2 である[*1]。過去の研究データから、平均 μ を決定し分散については最尤推定量により決定したモデル $M(\sigma^2_{ML})$ のパラメータ数は 1 である。

[*1] データ由来の母数 2 つあるので

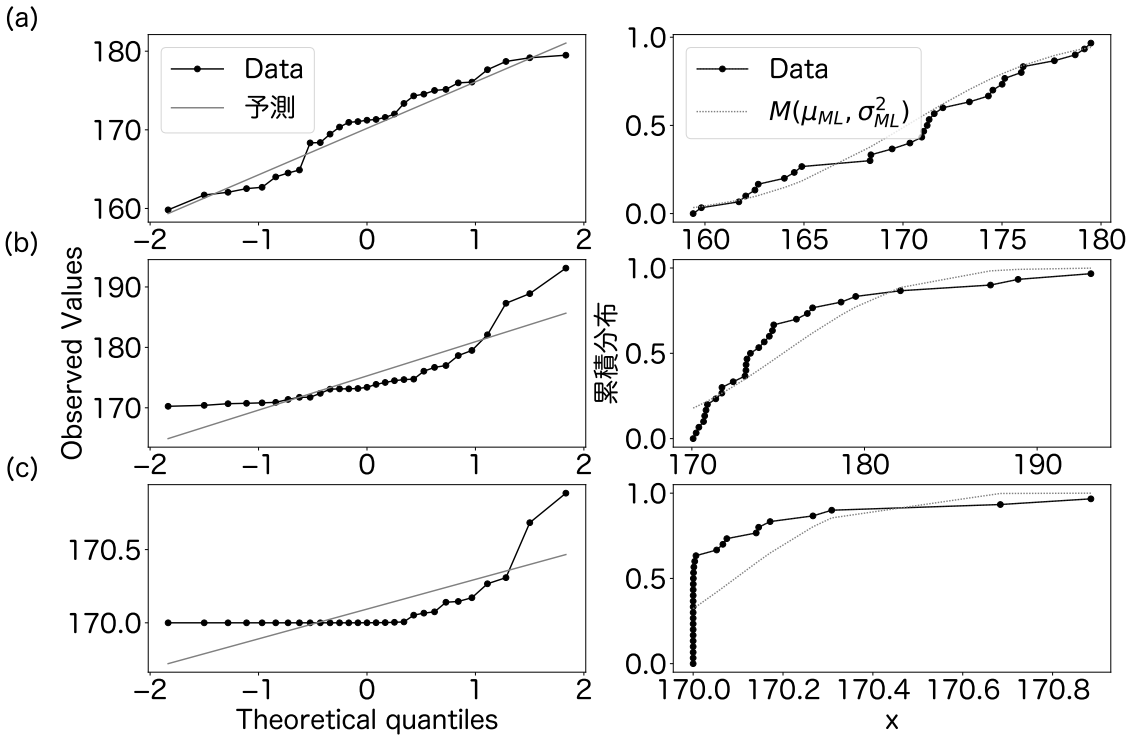

図 1.6　左には qq プロット、右は累積分布と最尤モデルの累積分布。サンプルサイズは 30(a) 正規分布 $N(170, 5.8)$(b) 指数分布 $\lambda = 5.8$(c) ガンマ分布 $s = 0.1$)

AIC が低いモデルは良い予測をするモデル？

AIC が低いから、良い予測ができるかは不明である。一般に、比較対象のモデルの中で、データへの適合度が相対的に高いモデルである。まず、AIC が低いモデルでも、データの出現を予測しにくい事例を紹介する。

1.4　モデルでデータを推測可能

1.4.1　ひとつのデータから推測する

サンプルサイズ 1 の標本が正規モデル $M(\mu)$ により予測できるのかを考える。$M(\mu)$ であれば、95% の確率で、$\mu - \sigma z_{0.025} \sim \mu - \sigma z_{0.025}$ の間でデータが見つかることを予測する。この中に入っていることが予測可能の目安の一つにはなる。

サンプルサイズが小さい場合では、標本分布の形がどの分布に適合するのかを推測しにくいので、このモデルが現実を予測していると言い切ることはできない。

第 2 章

統計量を使ったモデルとデータの比較

統計モデルからサンプリングを行った標本から、統計量を計算すると、その統計量より偏った値が出現する確率が得られる。この確率が低いとき、標本がモデルからサンプリングされたものではないと一定の割合で判断を下すことにする。このことを利用して、標本をモデルからサンプリングしたと判断するものとそうでないものに仕分けを行う。

2.1 自己標本の批判

統計モデルからサンプリングした標本の統計量が従う確率密度関数が理論的に求められる。正規モデルであれば、

$$Z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0, 1)$$

である。このことを利用すれば、Z 値よりも偏った値のモデルの中で出現する確率が計算できる。例えば、$Z = 0$ であれば、これ以上に偏った値がモデル内で出る確率は 0.5 程度なので、よくある統計量であることがわかる。$Z = 1.96$ であれば、これ以上に偏った値が得られるモデル内での確率は 0.025 程度なので、なかなかのレアさであると判断することがある。ここで注意しなければならないのは、p 値はモデル内における Z 値よりも偏った値の出現する確率であるということである。
言い換えを書いておく。

標本 $\rightarrow$ 統計量 $\rightarrow$ p 値 (p 値の小ささが標本のモデル上での得られにくさの指標になる)

p 値が小さいなら、統計量 Z 値以上に偏った値の得られる確率が低いということであるので、Z 値の元の標本もそもモデルからは得られにくいということを示唆する。つまり、モデルから標本の得られにくさの指標の一つが p 値であるとも言え、p 値が小さいほど、その標本はそのモデルから得られにくい。
モデルが生成したはずの標本であるが、閾値を決めてモデルから生成されたものではない

とするのである。モデルが自身から得られた標本を批判するのであるから、自己の標本を批判するのである。これは、モデルを元に、標本がモデルにより予測できるかどうかを考えている。

定義 2.1.1. 統計モデルにおいて、標本の統計量以上に偏った（大きいまたは小さな）値が得られる確率を p 値と呼ぶ。

2.1.1 p 値の計算練習

$Z(\bar{x}, \mu) \sim N(0,1)$ により、Z 以上の値が得られる確率も計算できる。つまり、

$$p = \Phi(Z(\bar{x}, \mu) > x)$$

を計算させる。モデルからえた標本から計算した統計量を $\bar{x} = 172.4$、サンプルサイズを $n = 10$。モデルとして、正規モデル $\mu = 168, \sigma^2 = 6.8$ とする。このとき、Z 値は $Z(\bar{x}, \mu) = 2.04$ であり、これを元に以下のスクリプトを実行すれば、$p = 0.04$ であることがわかる。

```
1 xbar = 172.4
2 mu = 168
3 sigma2 = 6.8**2
4 n=10
5 Z = np.sqrt(n)*(xbar-mu)/np.sqrt(sigma2)
6 print(Z)
7 p=1-norm.cdf(Z,0,1)
8 print(p*2)
```

2.1.2 自己標本の否定確率

あるサンプルサイズの標本をモデルから 100 標本を得たとすると、それぞれの統計量 Z_i をそれぞれ計算できる。全体のうち 95 個の標本についてはモデルから生成されたと判断し、残りの 5 個についてはモデルから生成されていないと判断することにする。これは自己標本の批判を元にすれば可能である。具体的には、正規モデルを利用すれば、その統計量 Z が $N(0,1)$ に従うことがわかっている。Z の値が偏った値になっていれば、その出現頻度は低くなるので、$P(|z| < Z) = 95/100$ となる Z を計算する。この Z は具体的に計算でき、$Z_{0.95} = 1.96$ である。ここから、$|z_i| < Z_{0.95} = 1.96$ となる z_i の個数を数えればおよそ 95 になる。また、$|P(|z_i| > 0.95)| > 1.96$ ならば、$|z_i| > Z_{0.95} = 1.96$ である。

式を展開する。

$$
\begin{aligned}
& |z_i| > Z_{0.95} \\
\rightarrow & \frac{\sqrt{n}|\bar{x} - \mu|}{\sigma} > Z_{0.95} = 1.96 \\
\rightarrow & \mu - \frac{\sigma}{\sqrt{n}} Z_{0.95} << \mu + \frac{\sigma}{\sqrt{n}} Z_{0.95}
\end{aligned} \tag{2.1}
$$

いくつか言葉を定義しておく。

定義 2.1.2. モデルからサンプリングされた標本のうち、モデルから生成されたものではないと判定する割合を α とし、有意水準と呼ぶ。言い換えれば、α 値は、統計モデルからサンプリングされた値について、これが元の統計モデルからサンプリングなのかどうかを判定する頻度に関する閾値[*1]である。式 (2.1) の範囲を信頼区間といい、これ以外の範囲を棄却域と言う。

本書では、歴史的な習慣にしたがって、$\alpha = 0.05$ を利用して、計算を行う。

2.1.3 μ の変化に応じた信頼区間

信頼区間は、サンプルサイズ n、有意水準 α およびモデルの母数 μ, σ^2 により決まる。ここでは、μ の変化に応じて、信頼区間が変化する様子を確かめる。
図 2.1 には、モデル毎の平均値と信頼区間を描いた。μ の大きさににによらず信頼区間の幅は同じである。各 μ に対して、信頼区間の内側で $\bar{x}$ が 95% の確率で見つかることを統計モデル $M(\mu)$ が推測する。この外側にある $\bar{x}$ になる標本については統計モデルにより推測できないのではないかと疑いがかけられる。

2.2 正規モデルにおける中心間の距離 (効果量)

ここでは、正規モデル $M_N(\mu; \sigma^2)$ について考える。分散が等しい二つの正規モデル $M_a = M(\mu_a), M_b = M(\mu_b)$ とする。標準偏差を 1 にすると、それぞれのモデルは、$M_{a'} = M(\frac{\mu_a}{\sigma}; 1^2), M_{b'} = M(\frac{\mu_b}{\sigma}; 1^2)$ である。$M_{a'}$ の中心から $M_{b'}$ の中心への距離は、$D = \frac{|\mu_a - \mu_b|}{\sigma}$ となる。D を効果量と呼ぶ。式を変形すれば、$D\sigma = |\mu_a - \mu_b|$ であり、中心からの距離が σ 何個分かを D が示す。$D = 1$ であれば、σ 分 M_a と M_b の中心は離れている。$D = 2$ であれば、σ2 つ分 M_a と M_b の中心は離れている。$D = 0.5$ であれば、σ の半分の距離で M_a と M_b の中心は離れている。

[*1] 閾値（読み：いきち）＝限界値

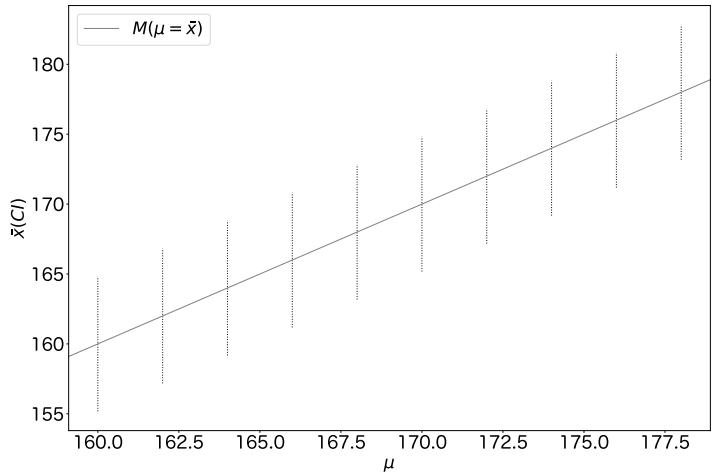

図 2.1　横軸にモデルの母数 μ、縦軸に、モデルが予測する平均値 $\bar{x}$、エラーバーに 95% 信頼区間を描いた。$N = 10, \sigma^2 = 6.8^2$

2.3　統計量をもとにしたモデル間類似度 (検出力)

母数の異なる二つの統計モデル M_a, M_b について考察する。M_a の信頼区間内の統計量が M_b の統計量において出現する確率を検出力という。言い換えれば一方で出現する統計量が他方のモデルにおいて出現する確率を計算する。これは、M_a から M_b へのモデル間の類似度と言える。

2.3.1　検出力の定義

M_a の棄却できない統計量の範囲 (信頼区間 A) に M_b の統計量が出現する確率を β とする。β を検出力という[*2]。β は、二つの異なるモデルを比較するための指標で、一方のモデルで棄却できない母数がもう一方のモデルで出現する確率である。M_a に対する M_a の検出力 β は、$1-\alpha$ であり、M_a を棄却する閾値を低く設定すると、β は大きな値になる。二つの統計モデルの母数がよく一致するならば、β は $1-\alpha$ に近い値を取り、一致していないならば、β は 0 に近い値を取る。具体的に、α, β を式で書くと、

$$P_a(\mu \in R_a) = \alpha$$
$$P_b(\mu \in A_a) = \beta$$

ここで、R_a, A_a はそれぞれ統計モデル M_a の棄却域、信頼区間、P_a, P_b は、それぞれ統計モデル M_a, M_b における統計量に関する確率密度関数。

[*2] 検出力を検定力または統計力と呼ぶこともある。
https://id.fnshr.info/2014/12/17/stats-done-wrong-03/

2.3.2 正規分布モデルの検出力

具体的に、$P_a(\mu \in R_a), P_b(\mu \in A_a)$ を計算してみる。σ^2 がすでに与えられた正規モデルを $M(\mu)$ とし、$M_a = M(\mu_a), M_b = M(\mu_b)$ とする。M_a または、M_b からサンプリングされた確率変数 $x_1, x_2, \cdots, x_n$ の平均値は、それぞれ $\bar{x}_a \sim N(\mu_a, \sigma/n)$ または $\bar{x}_b \sim N(\mu_b, \sigma/n)$ である。M_a の信頼区間 A_a は、$|\bar{x}_a| < \mu_a + \sigma/\sqrt{n} z_{2.5\%}$ である。このとき、P_a を $N(\mu_a, \sigma)$ の確率密度関数とすると、

$$P_a(\mu \in A_a) = \alpha$$

であるのは定義から明らか。また、P_b を $N(\mu_b, \sigma)$ の確率密度関数とすると、

$$P_b(\mu \in A_a) = \beta$$

である。μ_a と、μ_b が一致していれば、$P_b(\mu \in A_a) = 1 - \alpha$ である。μ_b が μ_a から離れていくと、$P_b(\mu \in A_a) = 0$ に近づいていく。

検出力と α の領域を図示した (図 2.2)。M_a の 95% 信頼区間は、$|\mu| < \mu_a + z_{0.025} \frac{\sigma}{\sqrt{N}}$ である。信頼区間は、図 2.2(a) において灰色で塗った x 軸の範囲である。α は図 2.2(c) の灰色で塗りつぶした領域の面積である。検出力 $1 - \beta$ は、M_b における M_a の信頼区間の外側の領域の面積なので、図 2.2(b) の濃い灰色の範囲である。

α を 0 に近づけていくと、信頼区間は徐々に大きくなり、β は大きくなる。α を 1 に近づけていくと、信頼区間は徐々に狭くなり、β は小さくなる。

α、M_a の母数 μ_a、M_b の母数 μ_b を固定したまま、サンプルサイズを変化させ, β の変化を表す (図 2.3)。$\bar{x}$ の確率密度関数 ($N(\mu, \sigma^2/n)$) の分散がサンプルサイズによって変化することは明らかである。このことから、サンプルサイズが大きくなると、信頼区間は徐々に狭くなり、$1 - \beta$ は大きくなる。サンプルサイズが小さいときは、$1 - \beta$ も小さくなる。

μ_a を固定し、μ_b を変化させたときの検出力 $1 - \beta$ を図 2.3.2 に示した。サンプルサイズが大きければ、$1 - \beta$ も大きくなることがわかる。

β を定義したことにより、β の数値を決定し、M_a, M_b の違いが β になるために必要なサンプルのサイズが推測できる。ここでは、μ_a, μ_b が固定されている状況を考える。検出力 $1 - \beta$ は 1 に近いほど、M_a, M_b が違うと主張できる。あらかじめ決めたおいた基準の $1 - \beta$ を閾値を設定し、それ以上の $1 - \beta$ となるサンプルサイズを推測する。サンプルサイズが小さければ、M_a と M_b の違いは曖昧であり、サンプルサイズが大きくなると、はっきりとモデルの違いがわかる。

2.3.3 β の計算

正規モデル M_a, M_b を使って、β を計算してみる。M_a の信頼区間は、

$$-z_{0.025} \leq \frac{\sqrt{n}(\bar{x} - \mu_a)}{\sigma} \leq z_{0.025}$$

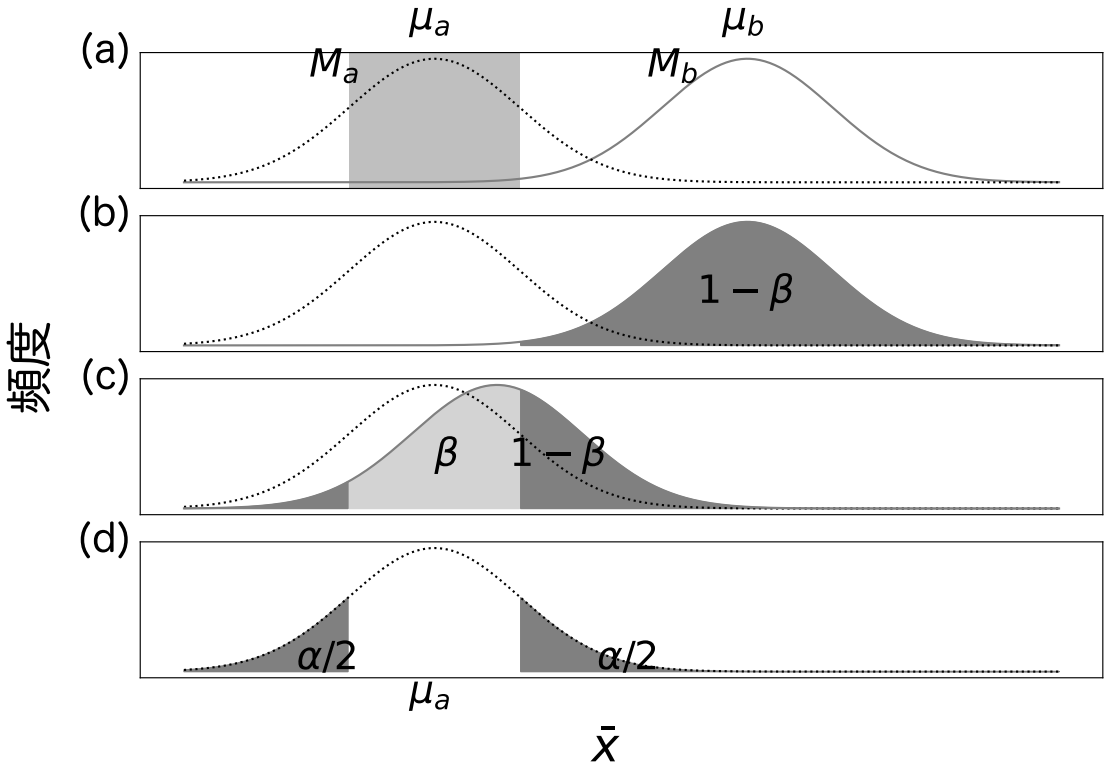

図 2.2　統計モデル M_a, M_b から計算された統計量 $\bar{x}$ の確率分布 P_a, P_b。(a) 灰色の範囲は M_a の信頼区間。(b) 灰色の領域は、$1-\beta$ の領域を示している。β の領域が小さいので、描画できなかった (c)μ_b が μ_a に近いときの β と $1-\beta$ の領域。(d) 灰色の範囲の面積が α を示している。

より、

$$A_a = \{\mu; \mu_a - \frac{\sigma}{\sqrt{n}} z_{0.025} \leq \mu \leq \mu_a + \frac{\sigma}{\sqrt{n}} z_{0.025}\}$$

である。ここで、$a = \mu_a - \frac{\sigma}{\sqrt{n}} z_{0.025}, b = \mu_a + \frac{\sigma}{\sqrt{n}} z_{0.025}$ とおく。棄却域は A_a 以外の μ である。M_b の標本平均 $\bar{x}_b$ は、$N(\mu, \frac{\sigma^2}{n})$ に従うので、A_a の区間で、$N(\mu_b, \frac{\sigma^2}{n})$ の面積を計算すれば良い。ここで、$\frac{\sqrt{n}(\bar{x}_b - \mu_b)}{\sigma} \sim N(0,1)$ である。このことを利用すると、a, b は、$N(\mu_b, \frac{\sigma^2}{n})$ の確率変数だとすると、

$$\begin{aligned} A &= \frac{\sqrt{n}(a - \mu_b)}{\sigma} \\ &= \frac{\sqrt{n}(\mu_a - \frac{\sigma}{\sqrt{n} z_{\alpha/2}})}{\sigma} \\ &= -z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b) \end{aligned}$$

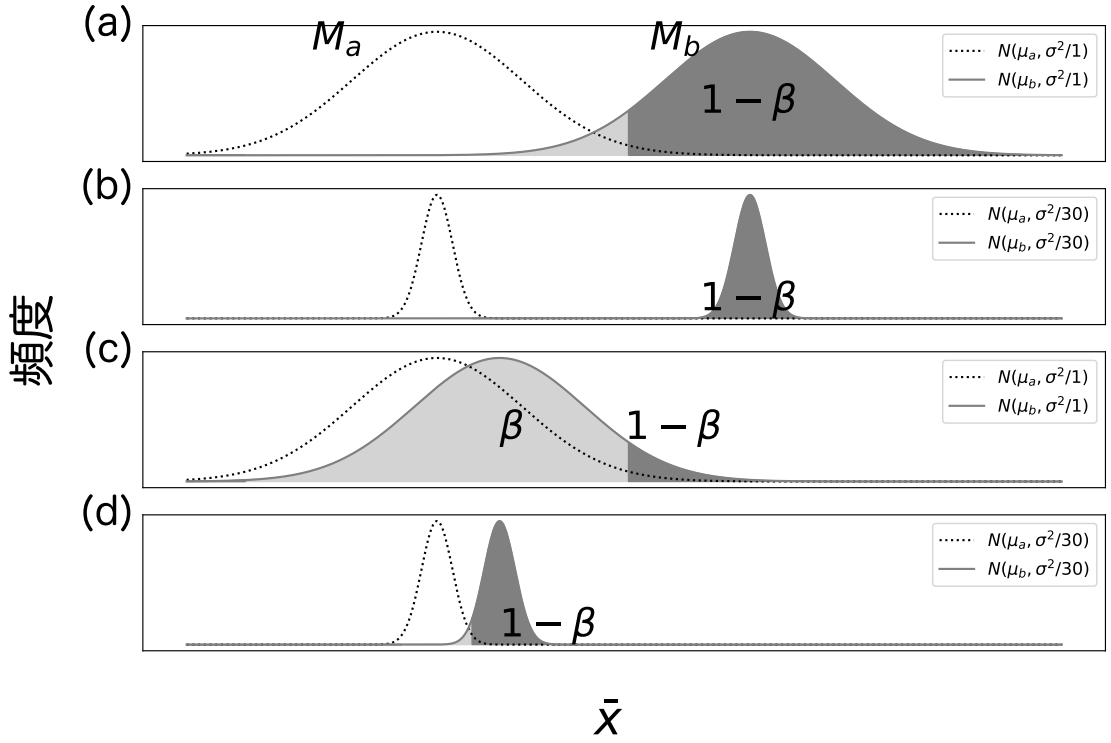

図 2.3 統計モデル M_a, M_b から計算された統計量 $\bar{x}$ の確率分布 P_a, P_b。(a)μ_a, μ_b のサンプルサイズ 1 の平均値がしたがう確率密度関数 $N(\mu_a, \sigma^2/1), N(\mu_a, \sigma^2/1)$。(b)(a) と同じ μ_a, μ_b に対して、サンプルサイズを 30 にした場合の確率密度関数。(c)μ_a, μ_b が (a) よりも近いときの $\bar{x}$ の確率密度関数。(d)(c) と同じ μ_a, μ_b に対してサンプルサイズを 30 にした場合の $\bar{x}$ の確率密度関数。

同様に、

$$\begin{aligned} B &= \frac{\sqrt{n}(b-\mu_b)}{\sigma} \\ &= \frac{\sqrt{n}(\mu_a - \frac{\sigma}{\sqrt{n}z_{\alpha/2}})}{\sigma} \\ &= z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b) \end{aligned}$$

である。以上より、確率密度関数 $N(0,1)$ において、$-z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b) \leq x \leq z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_a - \mu_b)$ の間で積分すれば良い。
$d = \frac{\mu_a - \mu_b}{\sigma}$ とおく。$d = 0.6, n = 9$ とする。このときの β を計算してみる。$N(0,1)$ において、$-z_{\alpha/2} - 0.6\sqrt{n} \leq x \leq z_{\alpha/2} + 0.6\sqrt{n}$ の区間で積分する。

```
1 A,B = norm.interval(0.95,0.,1)
2 N = 9
```

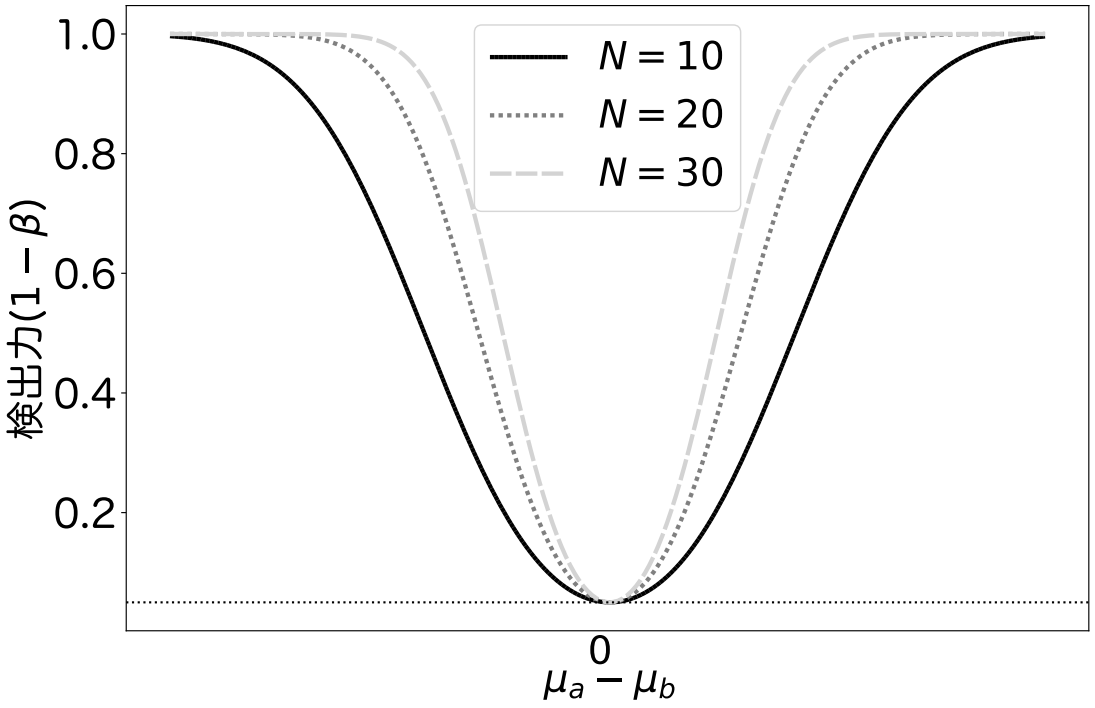

図 2.4　μ_a を変数にしたときの検出力 (検出力関数)。

```
3 d = 0.6
4 a,b = A+d*np.sqrt(N),B+d*np.sqrt(N)
5 print(a,b)
6 norm.cdf(b,0,1)-norm.cdf(a,0,1)
```

答えは、0.564

サンプルサイズ

d と検出力を指定したときに、M_a, M_b の類似度を検出力以上にするためのサンプルサイズが計算できる。$\beta = 0.1, \doteq 0.8$ とし、この β を満たすように N を計算した。

```
1 A,B = norm.interval(0.95,0.,1)
2 beta = 0.1
3 d = 0.8
4 for N in range(10,200,2):
5     a,b = A+d*np.sqrt(N),B+d*np.sqrt(N)
6     beta_ = norm.cdf(b,0,1)-norm.cdf(a,0,1)
```

```
7       if beta_ < beta:
8           break
9   print(N)
```

計算を実行すると、18 であることがわかる。

2.3.4 最尤モデルでの β の計算

データを元にしたモデルとモデルの類似度

統計モデル A を $M(\mu = 170)$ とし、統計モデル B を $M(\bar{X})$ とする。ここで、$\bar{X}$ は、無作為抽出によって得られた標本の平均であり、標本の大きさを 100 とする。モデル A,B の間の検出力が計算可能である。$d = \frac{170-\bar{X}}{6.8}$、$n = 100$ であるので、$\bar{X} = 168$ を得たとすると、

```
1   A,B = norm.interval(0.95,0,1)
2   N = 100
3   d = (170-168)/(6.8)
4   a,b = A+d*np.sqrt(N),B+d*np.sqrt(N)
5   print(a,b)
6   norm.cdf(b,0,1)-norm.cdf(a,0,1)
```

その検出力は、0.163

2.4 過誤

これまでの議論をまとめる。モデル M_a からサンプリングを行った標本について、モデル M_a に関する標本であるかを判定する。モデルから生成された標本であるが、偏った統計量出会った場合は、モデルから生成されていないと判断する。この頻度を α とした。このように、モデル M_a から生成されたのに、統計量の出現頻度から、このモデルから生成したものではないと言う誤った判断を行う事になる。このことを判断の間違いであると言うことから第 1' の過誤と呼ぶ[*3]。

今度は、モデル M_b からサンプリングを行った標本が、別のモデル M_a からサンプリングされたかを判定する。統計量が信頼区間に入っているかどうかを確認し、入っていなければ、モデル M_a からサンプリングされていないと判定できる。問題が生じるのは、統計量が信頼区間に入っている場合である。これは、実際には、M_a からサンプリングされていないのにもかかわらず間違って、サンプリングされたと判断する事になる。この判断の間違いを第 2' の過誤と呼ぶ。

[*3] Neyman-Pearson とは異なる過誤を定義したので、1' および 2' とした。仮説検定において、Neyman-Pearson と Fisher を混ぜ合わせて過誤を定義することが現在の主流である。こちらの定義では、さまざまな誤解が生じている [1]

表 2.1　モデル M_a による自己標本批判

	M_a の信頼区間に 標本の統計量が入っていない	M_a の信頼区間に 標本の統計量が入っている
モデル M_a の標本	モデル M_a の標本ではないと判定 (第 1' の過誤)	モデル M_a の標本と判定
モデル M_b の標本	モデル M_a の標本ではないと判定	モデル M_a の標本であると判断 (第 2' の過誤)

■過誤はデータとモデルを比較したときに生じる判断ミス

データとモデルを比べたときに、誤ってモデルが間違いと判定することを第一の過誤と教科書において紹介していることがある。誤ってモデルが間違いと判断するとはどのようなことなのかの定義がないので、この定義の意図がわからない。
本書では、モデルからサンプリングした標本とモデルを比較したときに生じる間違いとして過誤を定義した。標本は無作為抽出によって得られたものではない。

■正解と回答の違い

あるデータ群に対してそのデータの特徴を元に、Yes または No とアノテーションをつける。データからその Yes または No を予測する手順を開発する。その手順によって得た回答と、正解（真の値）の一致と不一致は以下の通りになる (表 2.2)。回答と一致したら、True、一致しないなら False。Yes と予測したら Positive、No と予測したら Negative とする。回答が Yes な問題に、Yes と答えることは (手順が正しい予測を行なった)、True Positive といい、No と答える (手順が間違えた予測を行なった) ことは False Negative という。回答が No な問題に、Yes と答えることを、False Positive、また、No と答えることを True negative という。
モデル M_a の標本に Yes を対応ずけ、モデル M_b の標本に No を対応付ける。標本を元に、Yes または No を判定する手順をモデル M_a を元にした統計検定を利用する。この問題において回答が FP となったものを第 1' の過誤であり、FN となったものが第 2' の過誤である。

表 2.2　正解と回答の違い

	負例 (真の値)	正例 (真の値)
正例 (予測値)	偽陽性 (FP) 予測が外れた	真陽性 (TP) 予測が当たった
負例 (予測値)	真陰性 (TN) 予測が当たった	偽陰性 (FN) 予測が外れた

2.5 自己否定の過推定

統計モデルの中で、統計モデルを統計量により検査するときに、モデル自身を絶対にダメなモデルと判断しすぎてしまうことを自己否定の過推定と言う。この過誤は 2 つの要因に分解でき、[*4]、不適切な統計量を使用することで、棄却域と統計量の違いにより生じる α_1、そして、検定を繰り返して生じる α_2 である $(0 < \alpha_2 \leq 1)$。$\alpha_2 = \alpha$ となっていれば、有意水準 α の検定ができる。α_1 は、統計モデルと、その統計量の関数になっており、言い換えれば、統計量が統計モデルの中で設計通りの振る舞いをしているかを測る指標である。正規モデルを使い、統計量 T を使った場合、$\alpha_1 \approx 0$ であるが、指数モデルを使い、統計量 T を使った場合、α_1 が指定した α よりも多くなる。これを見ていく。α_2 は、$\alpha \times 2$ 以上になる場合、軽視されることはないが、α_1 が同程度の隔たりになる場合においては無視され、α_1 は α_2 よりも軽視されがちであることも説明する。

2.5.1 どんな統計モデルでも T 統計量で調べよう (α_1)

統計モデルの分布の仮定が正規分布以外の場合においても、T 統計量を使ってモデル自身を検証できるのかを調べる。次の統計モデル $M_E(\lambda)$ を構築する。

1. $X_1, X_2, \cdots, X_n$ は i.i.d
2. 指数分布
3. λ

母数 $\lambda = 1$ とした統計モデルを $M_E(1)$ とする。$M(1)$ からランダムサンプリングした確率変数 $x_1, x_2, \cdots, x_n$ から次の統計量を計算する。

$$T = \frac{\bar{X} - 1}{\sqrt{\frac{\sigma^2}{n}}}$$

ここで、$T \sim t(n-1)$ とする。T 値が $t(n-1)$ の棄却域に入っている頻度を数値計算により計算する。具体的に、平均 1 の指数分布または、平均 1、標準偏差 1 の正規分布からサンプルを得て標本を作る。その標本を 100000 回取得する。このとき、T 値を計算し、T 値いじょの値が得られる確率 p を計算する。その p が $p < 0.05$ となる割合を計算する。以上をサンプルサイズを変化させてシミュレーションを行なった。平均 1、標準偏差 1 の正規分布の場合、T 値は $t(n)$ 分布に従ので、$p < 0.05$ となる頻度も、5% 程度になることが期待される。一方で、平均 1 の指数分布の場合、T は $t(n-1)$ 分布に従うとはいえない。このことから、$p < 0.05$ となる頻度は計算してみなければわからない。
シミュレーションの結果、正規分布から標本を得た場合、$p < 0.05$ になる割合は、サンプ

[*4] α_2 は α_1 に関係するので実際には、分解できない。気持ちとしては、α_2 は、α_1 を変数に持つ関数である $\alpha = \alpha_2(\alpha_1)$。

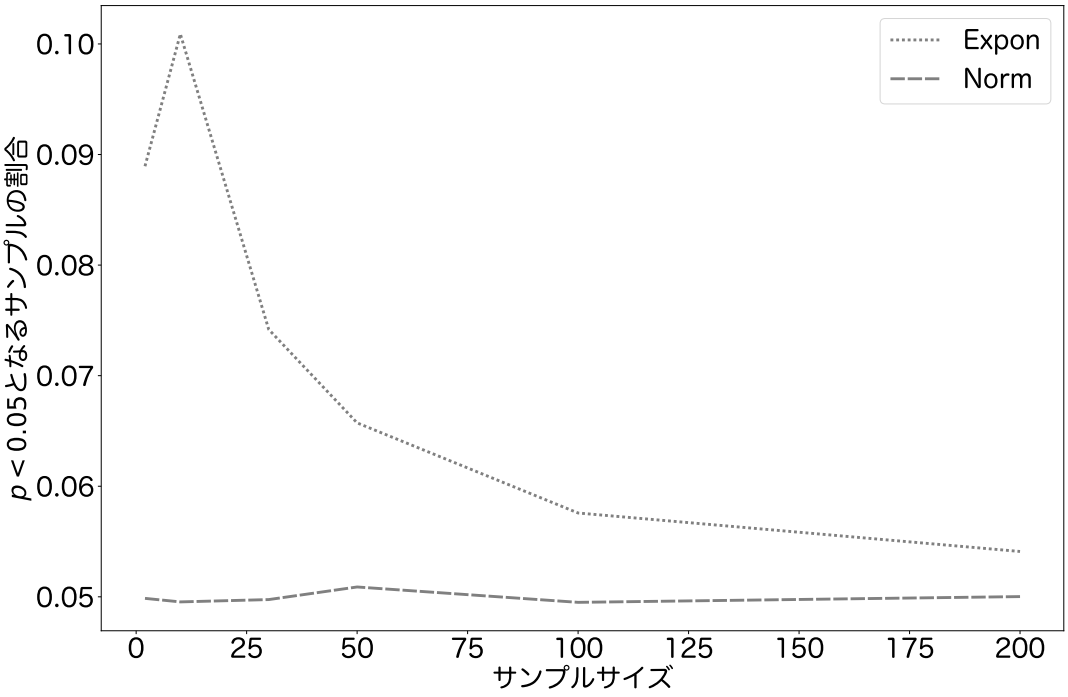

図 2.5　正規分布または指数ぶんぷから得た標本の T 値から計算した p 値で、$p < 0.05$ 以下になる割合

ルサイズに依存せず、5% 程度であり、期待通りである。一方で、指数分布から標本を得た場合、$p < 0.05$ になる割合はサンプルサイズに応じて変化しており、また、どのサンプルサイズでも $p < 0.05$ となる割合は 5% より多い。
このことから、指数モデルの α_1 は、$\alpha_1 > 0.05$ であることがわかり、統計量を正しく選ばなかったことで、自己否定の過誤が期待した 0.05 よりも大きくなっていることがわかる。

■サンプルサイズが xx 以上あるから t 検定

サンプルサイズがある値以上あるので、中心極限定理により、t 検定が利用できるというものもある[a]。このロジックが読み込めなかったので、その謎を明らかにすべく我々はアマゾンの奥地へ向かった。
データが指数分布的であるときに、t 検定を使うときに生じる問題は上でみた通りであり、$p < 0.05$ となる標本の割合が多くなっているので、間違った推測をする可能性が高くなる。他の分布関数でもおそらく同じような現象が現れる。このことから、我々は「t 検定が利用可能である」は正確ではなく、「t 検定を使うことができるが、間違った推測である確率が高くなる」ということだと推察した。
業界によっては、サンプルサイズが xx 以上であれば、過誤を無視して良いというふ

うに言われることもある。実際には、設計したモデルと

[a] http://id.ndl.go.jp/bib/024660739

2.6　検定を繰り返し使おう (α_2)

ここまでは、一つの標本に対して、統計モデル $M(\mu)$ により推測できるかを考えていた。ここでは、$\sigma^2 = 10^2$ とした正規モデル $M(\mu)$ によって複数の標本について推測できるかを仮説検定を指標にし考える。標本が 3 個あるとする。このとき、それぞれの標本の統計量 T が信頼区間に入っている確率は、$(1-\alpha)$ である。全ての標本の統計量 T が信頼区間に入っている確率は、その積 $(1-\alpha)\times(1-\alpha)\times(1-\alpha) = (1-\alpha)^3$ であり、この確率で統計モデルは棄却されない。一方で、棄却される確率は、$1-(1-\alpha)^3$ である。表 2.3

表 2.3　標本数に応じた α_2

標本数	$\alpha = 0.05$	$\alpha = 0.01$
1	0.05	0.01
2	0.0975	0.0199
3	0.142	0.0297
4	0.185	0.0394

は、標本数に応じた α_2 である。標本数が大きくなるについれて、α_2 が大きくなることがわかる。

α_1 がレベル α の検定になっていない場合、α_2 はさらに有意水準 α から隔たりの多い数値になる。

2.7　類似度の過誤

統計モデルの間の類似度を検出力といった。統計モデルに対して、不適切な統計量を与えたとき、検出力を歪める。これを類似度の過誤といい、その確率を β' で表す。直接またはシミュレーションを行い β を計算することがおそらくできそうだが、面倒なので行わない。

第 3 章

尤度比を使ったモデルとデータの比較

モデル M において得られるデータ元に、母数を最尤推定する。新たに作られた最尤モデル上での尤度と元のモデル M での尤度の比がある分布に従うことがわかっている。このことを利用して、もともモデル M でデータを予測してもいいのかを考察する。

3.1 尤度比検定

母数の個数が k 個のモデル $M(\theta)$ とする (θ は k 次元ベクトル)。モデル $M(\theta)$ からサンプリングしたサンプルサイズ n の標本 $x = (x_1, x_2, \cdots, x_n)$ を得たとする。この標本 X から θ のうち r 個の母数に関する最尤推定量を $\bar{\theta}$ 得たとする。$\bar{\theta}$ のうち $k - r$ 個はモデル由来の母数であり、r 個は標本から推定した母数である。このことから、$\bar{\theta}$ は自由度 r の母数のベクトルと言う。
もとのモデル $M(\theta)$ における標本 X に対する尤度は、$L(\theta, x)$ とする。また、最尤モデル $M(\bar{\theta})$ での尤度は、$L(\bar{\theta}, x)$ とする。このとき、これら尤度の比がカイ二乗分布分布に従うことがわかっている[*1]。つまり、

$$-2 \log \lambda(X) \sim \chi^2_{k-r}$$

ただし、

$$\lambda(X) = \frac{L(\theta, x)}{L(\bar{\theta}, x)}$$

である。

[*1] ただしいくらかの条件がある

3.2　正規モデルにおける尤度比検定

σ_0^2 を設定した正規モデル $M(\mu_0;\sigma_0^2)$ について考察する。この正規モデルからサンプリングを行なった標本 X とする。標本から得た最尤正規モデルを $M(\bar{x};\sigma_0^2)$ とする。それぞれのモデル内での標本 X の尤度を $L(\mu_0,X), L(\bar{x},X)$ とする。具体的な数式は、

$$L(\mu_0,X) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\sum(x_i-\mu_0)}{2\sigma^2}) \tag{3.1}$$

$$L(\bar{X},X) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\sum(x_i-\bar{X})}{2\sigma^2}) \tag{3.2}$$

$$\tag{3.3}$$

これらから $\lambda(X)$ を計算すると、

$$-2\log\lambda(X) = -2(-\frac{n}{2\sigma_0^2}(\bar{x}-\mu_0)^2) \tag{3.4}$$

$$= \frac{n}{\sigma_0^2}(\bar{x}-\mu_0)^2 \sim \chi_1^2 \tag{3.5}$$

$$\tag{3.6}$$

である。

数値実験

モデルと同じ確率密度関数からサンプリングを行い、尤度比検定を行なってみる。

数値実験を行なってみる。具体的に、正規分布 $N(170,5.8^2)$ からサンプリングした標本 1000 個を集める。標本から平均値を求め、これを最尤推定量とする (xbar)。この最尤モデル $M(\mu;\sigma^2=5.8^2)$ における標本の尤度を計算する (loglike2)。同様に、モデル $M(170;\sigma^2=5.8^2)$ における標本の尤度を計算する (loglike)。以上から尤度比を計算し、それが χ_1^2 分布と一致することを確かめる。以下がコードである。

```
1  norm_ = norm(170,5.8)
2  data_ = norm.rvs(170,5.8,size=(1000,10))
3  xbar = np.average(data_,axis=1)
4  loglike_ = np.prod(norm_.pdf(data_),axis=1)
5  #loglike2_ = np.prod(norm(xbar,5.8).pdf(data_),axis=0)
6  #print(np.prod(norm(xbar,5.8).pdf(data_),axis=1),xbar)
7
8  loglike2_ = []
9  for item in data_:
10     #print(item.shape)
11     a = norm(np.average(item),5.8).pdf(item)
```

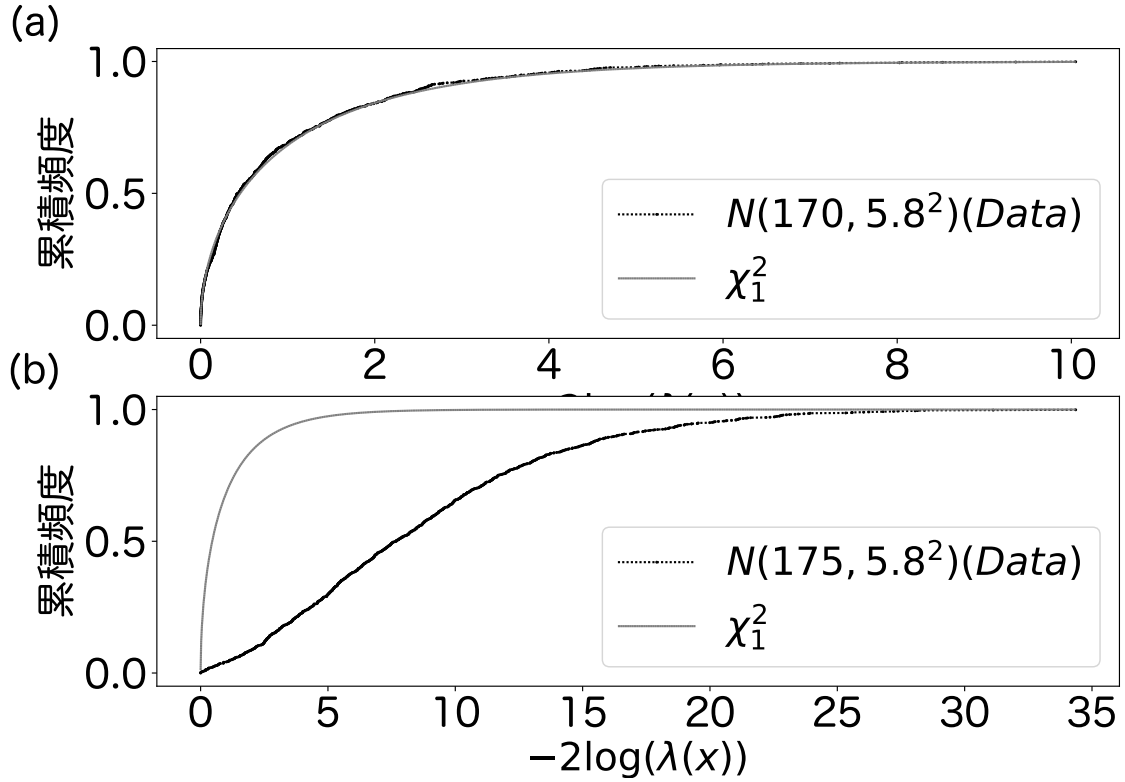

図 3.1　対数尤度比の累積頻度。モデルは正規モデル $M(170; \sigma^2 = 5.8^2)$。(a) 標本を $N(170, 5.8^2)$ からサンプリングした結果。(b) 標本を $N(175, 5.8^2)$ からサンプリングした標本。

```
12      loglike2_.append(np.prod(a))
13
14  y = -2*np.log(loglike_/loglike2_)
15  x = sorted(y)
16  y_ = np.arange(len(y))/len(x)
17  plt.plot(x,y_)
18  plt.plot(x,chi2.cdf(x,df=1))
19  plt.show()
```

$N(170, 5.8^2)$ と $N(175, 5.8^2)$ と言う 2 種類の密度関数からサンプリングを行いそれぞれ結果を図 3.1(a) および (b) に示す。図 3.1(a) は、モデルとデータの分布が一致していることから、累積分布が χ^2_1 の累積分布にかなり近いことがわかる。図 3.1(b) は、モデルとデータが一致していない状況での結果を示している。尤度比の多くが右に移動しており、標本の多くが χ^2_1 において珍しいと判定されやすくなっている。

3.2.1　データとモデルの乖離を検証する

モデル上において、その標本を元にした最尤モデルにおける尤度比が χ_1^2 に従うことを示した。このことを元に、データをモデルによって予測可能かを調べる。手順は、

1. 標本を x とする。
2. モデル M における最尤推定量を計算する。
3. モデル M および最尤モデル M_{MLE} における標本 x に対する尤度を計算する
4. 尤度比および $-2\log\lambda(x)$ を計算し、χ_1^2 において珍しい値なのかを検証する。

実際に、正規モデルにおいてこの手順をなぞってみる。$M(\mu;\sigma^2)$ における最尤モデルは、$M(\bar{x};\sigma^2)$ である。それぞれのモデルにおける尤度を計算し、$-2\log\lambda x$ を計算すればよい。

3.3　複雑なモデルでの尤度比検定

次のモデル $M(\beta_1,\beta_2)$ を考える。

1. x_i は定数。
2. y_i は以下に示す分布 $p(y_i;\lambda_i)$ に従う。
3. $\lambda_i = \exp(\beta_1+\beta_2 x_i)$
4. $y_i \sim p(y_i;\lambda_i) = \frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}$

無作為抽出した標本 x における 2 つの最尤モデルを考える。最初のモデルは、$\beta_2=0$ とした上で、β_2 に関する最尤推定を行なったモデル $M_1 = M(\hat{\beta_1},\beta_2=0)$ である。このモデルでは、x_i に応じて、λ_i が変化しないので、λ が常に一定のモデルになる。言い換えれば、y が母数 $\lambda=\exp(\beta_1)$ のポアソン分布となるモデルである。次のモデルは、β_1,β_2 の両方について最尤推定を行なったモデル $M_2 = M(\hat{\hat{\beta_1}},\hat{\hat{\beta_2}})$ である。このモデルにおいて、(x_i,y_i) はペアになっており、x_i に応じて y_i が揺らぎを持って決まる。
ここで、M_1 における尤度比が χ_1^2 に従うことを確かめる。手順は以下の通りである。

1. M_1 においてサンプリングを行い、(x_i,y_i) からなる標本 X を得る。x_i は、既存の標本 x のものを使う。(x_i,y_i) に関してバラバラになった標本が得られる。
2. M_1 における標本 X の尤度 L_1 を計算する。
3. M_2 における標本 X の尤度 L_2 を計算する。
4. $-2\log\frac{L_1}{L_2}$ を計算する。以上を繰り返す。

以上を行うと、χ_1^2 に従うことがわかる。図 3.2a,b に結果を載せている。コードを書いておく。

```
1 df = pd.read_csv("https://raw.githubusercontent.com/tushuhei
```

```
    /statisticalDataModeling/master/data3a.csv")

def get_dd(d):
    d['y_rnd'] = np.random.poisson(np.mean(d.y),len(d.y))
    model1 = smf.glm(formula='y_rnd~1',data=d,
    family=sm.families.Poisson())
    model2 = smf.glm(formula = 'y_rnd~x',data=d,family=sm.
        families.Poisson())
    #print(fit1.summary())
    fit1 = model1.fit()
    fit2 = model2.fit()
    return fit1.deviance - fit2.deviance

l = []
for i in range(1000):
    l.append(get_dd(df))


x = sorted(l)
y = np.arange(len(l))/len(l)
plt.plot(x,y)
plt.plot(x, chi2.cdf(x,df = 1))
plt.show()
```

データと、最尤モデル M_1 との比較は同様に、

1. 標本 x の尤度 L_1 を M_1 上で計算する。
2. 標本 x の尤度 L_2 を M_2 上で計算する。
3. $-2\log\frac{L_1}{L_2}$ を計算する。

最尤モデル M_1 においてデータ x が予測できないなら、$-2\log\frac{L_1}{L_2}$ が大きな値を取る。
最尤モデル M_1 からサンプリングされた標本の尤度と、最尤モデル M_2 での尤度を比較すると、χ_1^2 に従う。なぜならば、ここにおける最尤モデル M_2 のパラメータ β_2 はほとんど 0 と変わりなく、小さな値をとるので、M_1 と違いが少ない。標本が M_1 からサンプリングされていないなら、モデル 2 での最尤推定の結果、β_2 も 0 から離れてしまい、尤度比も大きくなるはずである。M_1 で標本 x を予測しない方がよくないことを示す証拠の一つになる。ただし、M_2 が良い予測モデルであるのかは不明である。
M_1 からサンプリングした標本で、M_1 および M_2 を推定したモデルの尤度比は χ_1^2 に従う。実際の標本を元に、M_1, M_2 を推定し、そのモデルの尤度比は、M_1 による予測がで

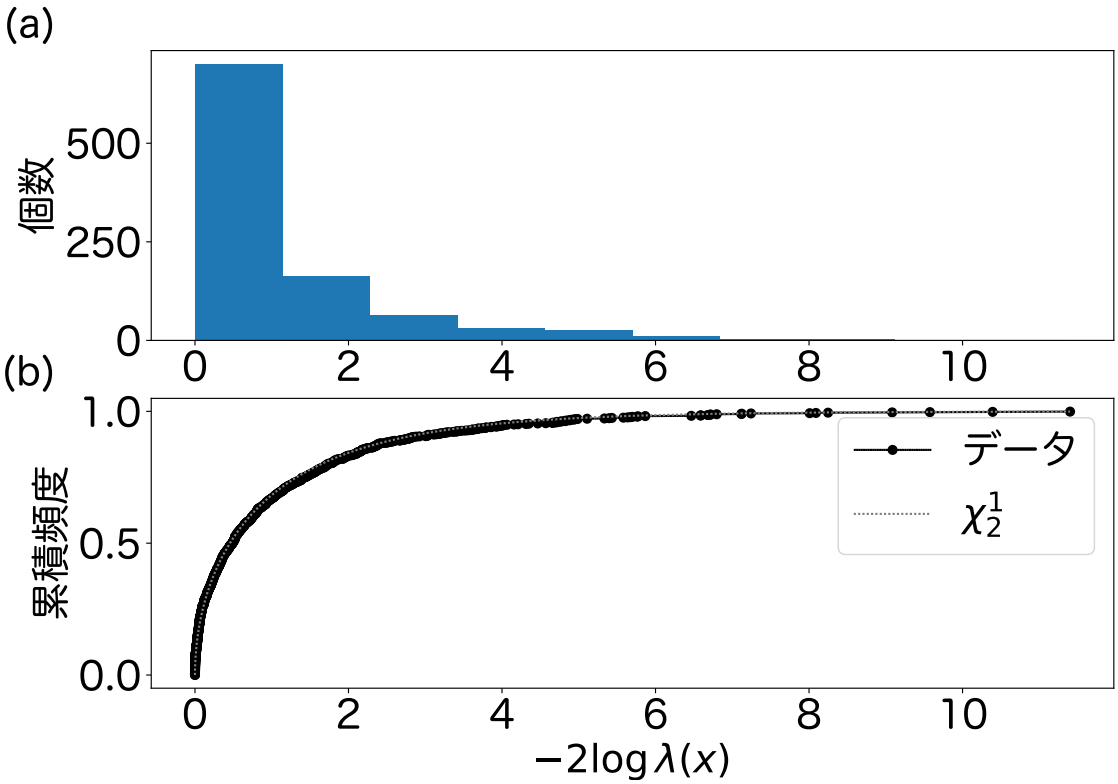

図 3.2　M_1 における対数尤度比の累積頻度。(a) ヒストグラム (b) 累積分布

きるならば、χ^2_1 程度だと考えられる。実際にこの例では、尤度比が大きくなり、χ^2_1 においては珍しい値になったので、M_1 により予測しない方が良さそうと言う根拠の一つになりうる。M_2 の $\hat{\beta}_2$ のパラメータがどうなっているかなども気にした方が良さそう。

付録 A

仮説検定の 3 つの枠組み

現在利用されている仮説検定の枠組みを紹介する。それぞれの枠組みは前提が異なっており、それぞれの問題で良い推測を与える。適切な枠組みを選ばなければ間違った解釈をしてしまう事になる。我々の行なっている科学では、F 型の枠組みを使う。以降の章では、F 型の問題について考える。

■有意性検定・仮説検定

Fisher は、帰無仮説を設定し、帰無仮説とデータを比較検討する方法を構築した。これを、有意性検定という。これに対して、Neyman-Pearson らは、帰無仮説に加えて、対立仮説を設定し、データを元に帰無仮説を棄却するかの判断を有意水準により行う意思決定の枠組みを構築した。これを、仮説検定と呼ぶことがある [1]。現代の科学の多くは、Fisher と Neyman-Pearson の両方を組み合わせ、帰無仮説・対立仮説を設定し、帰無仮説とデータの乖離を p 値によって調べ、棄却するかを検討する。ここでの p 値は、Neyman-Pearson の解釈から、20 回に 1 回程度のことを有意と呼ぶことにしている。この流派をハイブリッド仮説検定法 [2] と呼ぶことにする。

A.1　F 型

A.1.1　解決できる問題

A.1.2　データとモデルの比較

ここで、いくつかのことを定義しておく。

定義 A.1.1. 統計モデルと標本を比較して、モデルが母集団のことを予測できないとさまざまな指標をもとに判断するとき、統計モデルを却下すると宣言する。

ここで、母集団から無作為抽出した標本 (モデルから生成された標本ではない) を正規モデルにより、予測できるかを考える。上記の議論と同様に、標本から、統計モデルにあっ

た統計量を計算し、統計量よりも偏った値が出現する確率 (p 値) を計算する。p 値が小さければ、モデルにより予測できないと考え、値が 1 に近いほど、もしかしたらモデルで予測できるのかもしれないと考える[*1]。標本を元に、モデルにより予測ができないかを考えている。

以上のことは、托卵行動に例えることができる。モズは、カッコウに対して卵を託す托卵を行い、カッコウは、モズの卵とは気が付かず、そのまま育てる。ここで言い換えたいのは、カッコウは統計モデルであり、卵は標本そして、モズは科学者である。統計モデルは、モデルからのサンプリングされた標本を巣穴に置いている。卵の情報を要約した統計量が、モデル由来であることをモデルはその統計量の出現頻度を推測できる。出現頻度が p 値である。モデルの巣に自然から無作為抽出した標本を科学者が置く。その標本の統計量の出現頻度をモデルは推測できる。得られた推測から、標本がモデルの卵であることを判定するのは科学者である。この手順だけでは科学者はモデル鳥と標本卵を比較しているだけであり、標本卵を構成しているデータそのものとモデル鳥を比較していないということに注意しなければならない。

■偶然の差が生じたかを確かめたい

「偶然の差が生じたかを確かめたい」や「こんなことが起こる確率は 5% くらい」という言葉を統計学の教科書で見たことがある。これらは、本書での説明とは異なる前提をもとに議論を進めており、本書と解釈の互換性はない。

科学では、実験で得られたデータは、同様の実験を行った場合、同様のものが得られるということが前提になっている。このことを現象に再現性があると言う。再現性のないデータを現状の統計学で扱うことや、現実の現象が得られる確率を議論することは困難である。

本書の前提を元にすれば、「こんなこと（これ以上に偏った統計量値）が（モデル内で）起こる確率は 5% くらい」ということを省略して「こんなことが起こる確率は 5% くらい」と言うことはできる。また、現実において起こりやすいのかどうかについては議論できない。

A.1.3　p 値を使った判断に関する注意

p 値を元に統計モデルとデータの不一致を考えるとき、p 値はモデルとデータの乖離を示す指標の一つであると言うことを意識しなければならない。このことを忘れてしまい、次の間違った判断を行うことがある。

1. p 値が 0 に近いならば、統計モデルによりデータを予測できないと判断する
2. p 値が 1 に近いならば、統計モデルによりデータを予測できると判断する

[*1] p 値だけで判断してはいけない

図 A.1　統計量を使ったモデルとデータの比較に関する概念図

それぞれのデータがどのようなものなかのかを確認してみる。

p 値が 0 に近い → 統計モデルによりデータを予測できないと判断
p 値が 1 に近い → 統計モデルによりデータを予測できると判断

A.2　NP 型

NP 型では、すでに前提になっていることから著しく外れたことが起きたことを検出するための方法である。言い換えるなら、まず、何度も検証を行っていてモデルが事象をよく予測することがわかっている状況を構築する。そして、その事象から、いくつかの無作為抽出し、計測を行う。最後に、計測の平均がモデルの予測を著しく外れているならば、前提に狂いが生じたのではないかと疑う。

A.2.1　解決できる問題

NP 型で扱う問題をいくつか挙げる。

ある調味料の製造ラインでは, 砂糖の含有量 (g) は, 原料の不均一や製造ラインの狂いなどから変動するが, 標準偏差は常に一定で $\sigma = 3$ の正規分布に従っているとしてよい. 各製品の砂糖の含有量が $\mu = 60$ になるように調整してラインを稼働させて, しばらくしてから, 25 個の製品を抜取検査したところ, 砂糖の含有量の平均値は $\bar{x} = 61.63$ であった. その時点で製造ラインは $\mu = 60$ を保持していると言えるだろうか.

正規モデル $M(60, 3^2)$ によって予測ができるという前提条件を満たしている。この前提を元に、無作為抽出した 25 個がそのモデルにより予測できるのかを調べる。標本全体とモデルの累積分布などを比較する方法もあるが、ここでは、検定によって調べてみる。このモデルでは、

$$Z = \sqrt{n}\frac{\mu - \bar{x}}{\sigma} \sim N(0, 1)$$

である。変数を入れれば、$Z = 2.72$ となる。$vavrPhi(Z) < \Phi(1/20 = 0.05)$ であり、モデル内で、20 回に 1 回よりも少ない頻度で観測されないようなことが現実で起きている。または、$\Phi(Z) < \Phi(2\sigma = 2)$ であり、$2\sigma = 2$(標準正規分布の中で) よりも珍しいことが起きているので、モデルでの予測ができないことが起きている。偶発的に生じた可能性も捨てられないが、製造過程に不具合が生じているのではないかと推測される。

A.2.2　解釈

A.3　ハイブリッド型

A.3.1　解決できる問題

A.4　モデルの設定

帰無仮説 $\mu = \mu_0$ をを含む統計モデル $M(\mu_0)$ を帰無モデル (M_{H_0})、対立仮説 $\mu \neq \mu_0$ を含む統計モデル $M(\mu \neq \mu_0)$ を対立モデル (M_{H_1}) と呼ぶ。一般に、統計モデルの否定したい母数 μ_0 を帰無仮説と言い、その母数ではないという $\mu \neq \mu_0$ を対立かせつと言います。つまり、次のように帰無仮説を含む統計モデル M_0 を構築します。
具体的には、データがある特定の母数 μ をもつ統計モデルの信頼区間に含まれるか否かによって、統計モデルが棄却されるかを調べます。

- i.i.d
- 数学関数
- 統計モデルの母数を μ とし、$\mu = \mu_0$

一番最後の仮説が帰無仮説と言います。対立仮説を含む統計モデル M_1 は、M_0 と同様の仮説 (1),(2) から構成されますが、仮説 3 は統計モデル M_0 と M_1 で異なります。

- i.i.d

- 数学関数
- 統計モデルの母数を μ とし、$\mu \neq \mu_0$

一番最後の仮説が対立仮説です。M_1 の最後の仮説は、M_0 の最後の仮説の否定系になります。
二つの統計モデルを作って、M_0 で計算される信頼区間に、データから得られる統計量が入らないなら、M_0 は棄却されます。逆に、統計量が信頼区間に入るなら、何も起こりません。このように、否定したい仮説を設定し、少なくとも帰無仮説を含む統計モデルはだめだったと判断します。

参考文献

[1] 祐作大久保, 健大會場. p 値とは何だったのか : Fisher の有意性検定と neyman-pearson の仮説検定を超えるために. 生物科学 = Biological science, Vol. 70, No. 4, pp. 238–251, 04 2019.

[2] 土居淳子. 帰納的推論ツールとしての統計的仮説検定ー有意性検定論争と統計改革 ー. 京都光華女子大学人間関係学会 年報人間関係学, No. 13, 2010.