[II-C] 解答例

問 1

出題意図： 判別正答率 Kについて理解するための導入問題

解答例

$$K(\text{病状}) = \frac{58+145}{150+149} = 0.679$$

$$K(\text{治療法}) = \frac{28+129}{150+149} = 0.525$$

$$K(\text{放射線療法の有無}) = \frac{61+97}{150+149} = 0.528$$

問 2

出題意図： 事象の独立性について理解するための導入問題

解答例：

Q(病期) = 122.334

Q(病状) = 58.882

Q(治療法) = 1.525

Q(放射線療法) = 1.058

補足：$\chi 2$ 乗値の計算である。

問 3

出題意図： 判別正答率 Kと非独立指数 Qの関係についての考察能力をみる。

解答例： 判別正答率 Kが高いプロフィールは，非独立指数 Qが高い傾向にある。例えば，病期の判別正答率は 0.819 でもっとも大きく，非独立指数は 122.334 もっとも大きい。逆に治療法や放射線療法では判別正答率が，0.525 や 0.528 で，これはプロフィールを全く用いない場合の判別正答率≈0.5 とあまり変わらない（つまり，治療法や放射線療法を単一のプロフィールをして用いた場合，生存するか死亡するか判別できていない)。規則を逆にすると判別正答率は，1 から元の判別正答率を引いたものになるので，どのようなプロフィールでも判別正答率が 0.5 を超す規則が作れる。従って，0.5 からほとんど違わない判別正答率を与える規則は価値がない。こういった判別正答率をもつプロフィールは非独立指数が小さい値になっている。つまり，生存率と変数 X が非独立であれば，変数 X の値（もしくは状態）から，生存率が予測でき，判別正答率も高くなる。逆に生存率と変数 X が独立であれば，変数 X の情報は死亡・生存の予測には役立たず，一般的には判別正答率も低くなる。

病状の判別正答率は 0.679 と比較的大きいが，この数値だけでは，どれだけ有用なのかは解釈しにくい。しかし病状の非独立指数は十分に大きく，非独立で有用であると判別できる。

備考：問 5 で，各変数の判別正答率と非独立指数 Q での順番は同じにならない場合があることも記述されているので，それについて述べられることも期待する。つまり，最も判別正答率が高いものは，非独立指数 Q が高いとは限らない。問 5 でもあるように，判別正答率は母集団全体での確率にも依存する。例えば，患者全体の 90%が生存しているなら，大抵の条件で「生存している」が，80%以上の判別正答率を与えるだろう。

問 4

出題意図：　問 1－2 で計算された数値の意味を理解できているか試す。

解答例：　判別正答率と非独立指数 Q の数値をもとに考えると，がん患者の病期（初期と進行性）が患者の 10 年後の生存率に最も関係していることがわかった。病気が初期であれば 10 年後に生存している可能性が高く，逆に進行性のがん患者は 10 年後には死亡している可能性が高い。また，病期ほどではないが，病状も 10 年後の生存率を予測するのに有用だと考えられ，局部的なら生存，広範囲なら死亡といった規則が成り立つと言える。

逆に，治療法や放射線療法の有無は 10 年後の生存率とほぼ無関係だとわかった。つまり，治療法や放射線療法の有無による，10 年後の死亡・生存の判別は不可能である。

備考：　1976 年にとられたこのデータでは，治療法や放射線療法の有無は 10 年後の生存率とほぼ無関係である。最近の治療法や放射線療法の効果についての問いではない。

問 5

出題意図：　判別正答率 K と非独立指数 Q どちらが，予測に有効であるかを比較した答案を期待している。

解答例：　非独立指数 Q をもとに作成された図 2 がより有益だと考えられる。

非独立指数 Q を用いると，初期がん患者においては，治療法と生存率が独立していることがわかる。表 9 を見てみると，初期がん患者全体の死亡対生存の比（31 対 127）は，大規模治療においての死亡対生存の比（27 対 105），および小規模治療においての死亡対生存の比（4 対 22）とほぼ同等である。つまり，治療法が大規模であろうが小規模であろうが，死亡対生存の比の変化はないので，治療法による分類は，無作為におこなわれた分類となんら変わりがない。これは，治療法によって死亡・生存の確率が変化しないことを示している。

逆に病状を用いると，局部的がん患者においての死亡対生存の比と，広範囲がん患者

においての死亡対生存の比では，大きく違うことがわかる。具体的には，初期患者全体では死亡対生存は 31 : 127 であるが，初期患者でも広範囲のがんを患っている場合は，死亡対生存は 29 : 77 となり，死亡する可能性は高まると結論付けられる。

備考：　治療法で分類すると判別正答率が高くなる理由

治療法を用いた分類が高い判別正答率を得られたのは，大規模治療を受けた患者の数が小規模治療を受けた患者より相対的に多いからである。厳密に証明することは難しいが，二分した際に片方が圧倒的多数となると，判別正答率が高くなりやすいことが知られている。もし大規模治療と小規模治療を受けた患者の比が，局部的と広範囲患者の比と同等のものであれば，治療法による判別正答率は病状による判別正答率を下回ると考えられる。

問 6

出題意図：　分析の結果を予測に応用する方法を組み立てる能力をみる。

解答例：　表 1 を全て図に置き換えると，枝の数が 15（本来は 16 であるが，ここでは，進行性，小規模，放射線使用無し，局部的の患者が 0 なので 15 になる。）になる時である。この時に誤分類が最も少なくなる，しかし，表 1 の様に，極端にデータ数が少ないグループある場合（例：進行性，小規模，放射線使用無し，局部的の患者），ごく少人数の変化によって結果が全く逆になり得る。例えば，表 1 では進行性，小規模，放射線使用有り，局部的の患者は一人（生存）だけだったが，もう 1 人現れ死亡した場合，生存率は 100%から 50%へ急激に下がる。このように，関係図を複雑にしていくと，少量のデータの変化によって分析結果および予測が全く変わることがあるので，（データの量に対して）必要以上に複雑な関係図の作成は避けるべきであると考えられる。つまり，関係図を複雑にすると（枝の数を増やす）と一般的には現状データにおいての誤分類は減る。しかし「必要以上」の関係を模索しようとすると，偶然による本来は関係のない変数の間に，偽の関係を見いだす危険性がある。

次のような点について論じている答案も高く評価する。

— 関係図を複雑にすると結果の解釈が難しくなり生存率と患者のプロフィール間の構造の理解を促し難くなる。

— 逆にデータ数が大きい場合は，相対的に複雑な図でも安定しているので，より複雑な関係図は正確な判別が可能。つまり，適度な「複雑」度合いはデータ数と相対的である。

— また，同じようにデータが母集団の代表的なものであれば，複雑な図でも安定しているので，より複雑な関係図は正確な判別が可能。つまり，適度な「複雑」度合いはデータ精度と相対的である。

— 非独立指数が大きい場合や，枝の先端で判別が極端な数値をとる場合は，複雑にしても正確な判別が可能な場合がある。

問 7

出題意図：ここで紹介した分析法の応用力をみる。

解答例 A： まず，データを生存と死亡の二つに分割し，それぞれの年齢分布図を作成する。二つの分布図が交わらない場合は両端の中間点を分割点すれば，誤分類なく分割できる。二つの分布が交わる場合は，誤分類法免れない。分割点を下げると，年齢が低いグループでの正答率は高くなるが，年齢の高いグループでの正答率が下がる。両者での正答率を同じ程度にする分割点が，全体の誤分類を最も小さくする合理的な分割点である。

図に示された例の場合，年取った方は若いかたに比べて死亡する確率が高いが，分割点より年齢が下なのに死亡した数と分割点より年齢が上なのに生存しているといった，誤分類を最小にする年齢の数値を特定する。これは，二つの分布図が交わる点が応用できる。例えば，下の図 7.1 で説明すると，ここには二つの分布がある，右側の分布を死亡した患者で，左側の分布を生存している患者としよう。その場合，下のように，二つの分布が交わる点（Z）で 2 分割するのが誤分類を最小化できる。

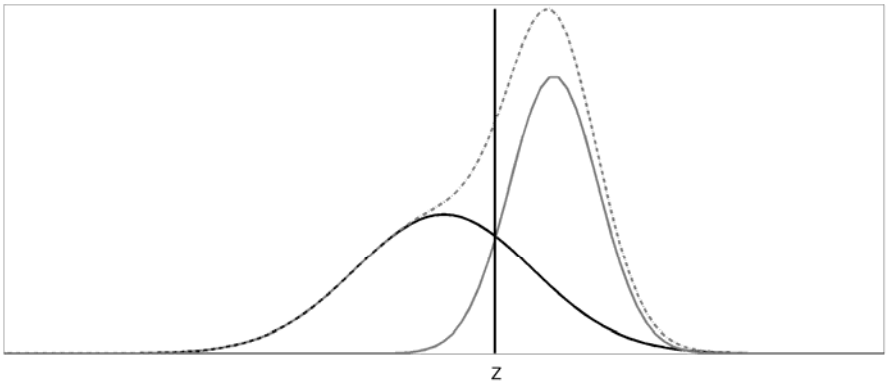


図 7.1

Z 点で分割した場合の誤分類の割合は，図 7.2 の縦じま(若いが死亡)と横じま(年取っているが生存)の面積の和で表される。

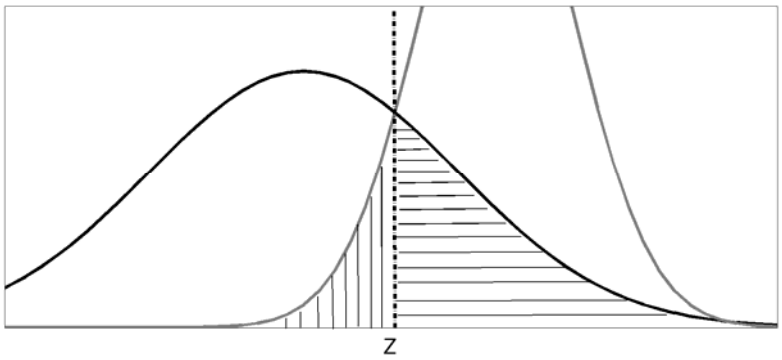


図 7.2 図 7.1 の拡大図

下の図 7.3 の様に　Z 点より低い数値（D 点）で分割した場合，縦じまの面積は減るが横じまの面積は増える。従って，誤分類は太枠で囲った面積だけ増える。

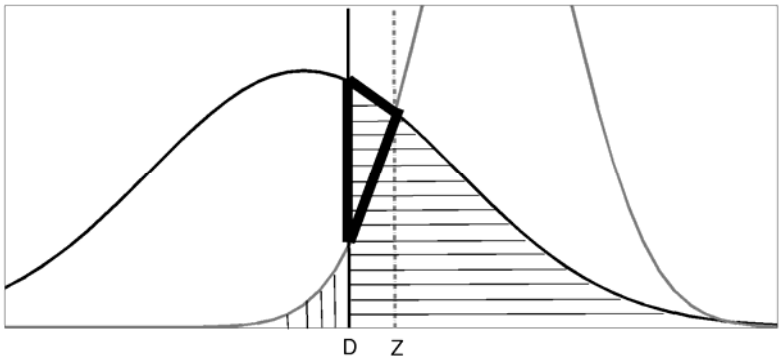
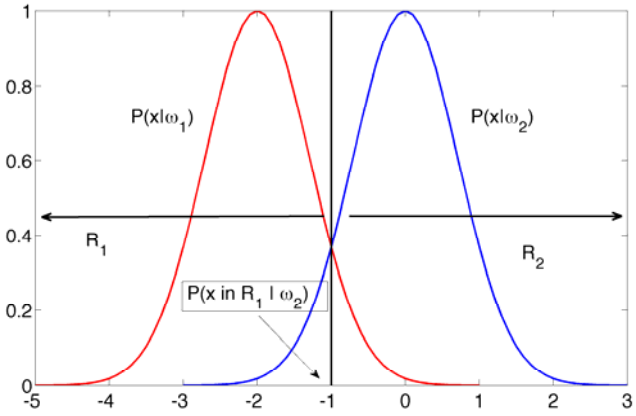


図 7.3　図 7.1 の拡大図

解答例 B：まず生存者全体を半分ずつに分ける年齢(中央値)を求める。同じ要領で，死亡者全体の中央値となる年齢を求める。この二つの中央値の平均で区切る。
備考：この方法のほうが簡単であるが，解答例 A のほうが誤分類を減らすという意味ではよい答えを与える。

先生方への補足：　解答例 A は，ベイズ統計法で説明できる。



もし $P(\omega 1 \mid x) > P(\omega 2 \mid x)$, x を ω1 と分類する。

もし $P(\omega 2 \mid x) > P(\omega 1 \mid x)$, x を ω2 と分類する。

といったルールを応用すると：

$$
\begin{aligned}
P(error) &= P(x \in R_2 \mid \omega_1) + P(x \in R_1 \mid \omega_2) \\
&= P(\omega_1)\int_{R_2} P(x \mid \omega_1)dx + P(\omega_2)\int_{R_1} P(x \mid \omega_2)dx \\
&= \int_{R_2} P(x \mid \omega_1)P(x)dx + \int_{R_1} P(x \mid \omega_2)P(x)dx
\end{aligned}
$$

が最小化される。