

第1章の答え

1. ①連続変数、②離散変数、③連続変数、④連続変数となります。また、①横断面、②横断面、③時系列、④パネルデータとなります¹。
2. 1)サンプルサイズは1000となります。2)この大学は母集団を代表していない可能性があります。全国の大学生から無作為抽出する必要があります。
3. 銀座にいる人は、比較的高所得者が多いと考えられます。また、年収の高い人の方がアンケートに答えてくれやすい、低い人は答えにくい傾向もあるかもしれません。
4. 合格する自信のある人が事前予約していると考えられます。部屋探しのコスト（部屋を探す手間）を支払ってまで事前予約する人は、自分が合格するという自信を持っているのです。
5. 監督官庁の聞き取り調査は、その企業が不正行為を隠そうとするバイアスがかかってしまうものです。たしかに回収率は高くなりますが、本当のことを答えていない可能性があります。
6. 太郎君は見栄をはって、正直に答えなかった可能性があります。
7. 対面調査の利点は、明らかな嘘を見抜くことができること、回答者の記入漏れなどもなくなることです²。欠点は、面接員の特性が相手に何らかの影響を与えること（問6の太郎君の行動が該当します）、面接員の手抜きが生じることです³。また、面接員を前にして、道徳的また倫理的問題のある質問に正直に答えにくいことが考えられます。これ以外にも、時間面と費用面のコストもばかにできません。
8. インターネット調査の利点は、コストが低いこと、大規模調査ができること、対象者(年齢、性別など)を絞った調査ができること、回答結果が電子

¹所得は1円単位であり厳密には連続変数ではありませんが、そのとりうる値は無数にあり連続変数として扱われるのが普通です。

² 記入漏れは重要な問題です。記入漏れをする人は、何らかの特徴（ケアレスミスをしやすい、秘密主義者であるなど）がある人かもしれません。それらの人を記入漏れということでデータから除いてしまうと、セクションバイアス（特定のグループが選ばれてしまうことで生じるバイアス）が発生してしまいます。

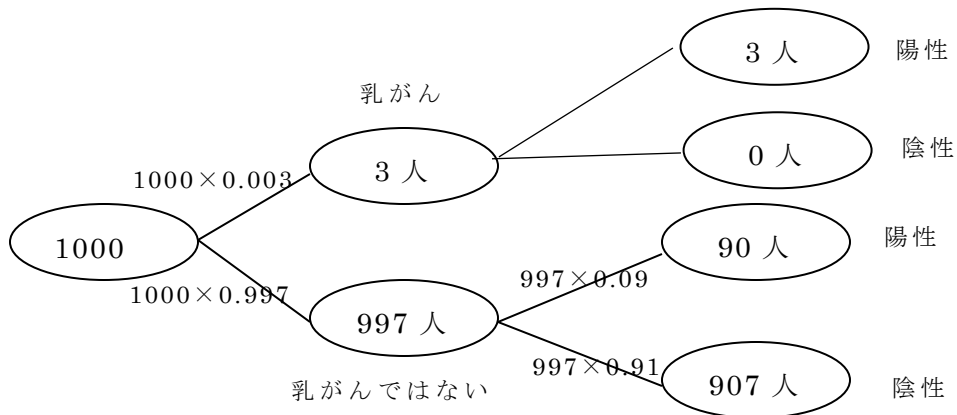
³ たとえば、面接員が調査票に適当に記入して面接したことになってしまうかもしれません。

化されることがあります。欠点は、調査に協力してくれる人は必ずしも母集団を代表していない可能性があることです。また、同一人物による複数回答の恐れもあるでしょう。

9. 正直に答えることで人種差別主義者だと思われることを恐れて、白人有権者が黒人候補者を支持していると嘘をついた可能性があります。世論調査では、全員が正直に答えてくれるわけではありません。他人に知られたくない内容（モラル、犯罪に関わること）であれば、嘘をつく可能性が高くなります。

10. 1)1000人中200人なので、20%が喫煙したことがあるといえます。2)1000人中500人は表が出たから「はい」と答えています。意味がある回答は500人だけで、意味がある「はい」は $700-500=200$ 人だけです。よって、喫煙したことがある割合は $200/500$ で40%と推定されます。3) 1)の調査では、回答者は嘘をついている可能性があります。2)の調査では、回答をランダム化しており、回答者の嘘をつく動機は軽減されており、より信頼のできる調査といえます⁴。

11. 1000人のうち、乳ガンにかかっているのは3人（0.3%）で、全員が陽性と診断されます（下図参照）。乳ガンではない人は残り997人で、このうち陽性と診断されるのは 997×0.09 ですから約90人です。つまり、1000人中陽性と診断される人は、 $3+90=93$ 人です。以上の結果から、陽性であり本当に乳ガンである確率は3.2%（ $=3/93$ ）となります。



⁴ 回答をランダム化しても嘘をつく人はいるでしょう。しかし、回答をランダム化することで、嘘をつく動機は軽減されており、正直に話す人が増えているはずですが。この調査は完璧なものではありませんが、より信頼できる調査といえます。