1.3 欠損値の除去

アンケート調査などを実施すると、問題の中には回答されていないものが出てきます。これを欠損値と言い、統計処理では取り除いて処理する必要があります。ここではこの欠損値の取り除き方について見てみます。

以下の例について見て下さい。これにはあるテストを行った場合のデータで、受けた生徒の番号、学校（番号表示)、受験科目の点数が出ています。このデータに対していくつかの処理を考えてみましょう。

例

番号	学校	国語	数学
1	1	76	82
2	2		63
3	1	62	58
4		73	74
5	2	81	
6	2	73	65
7	1		46

まず国語の平均を考えてみます。この処理で利用されるのは何番の人でしょうか。通常考えられるのは 1, 3, 4, 5, 6 です。この処理では他の科目の点数が何であれ、国語を受験している人の点数が問題です。このように、他の項目を見ずにある項目だけで取り除く人を決める方法をデータ単位の欠損値の除去と言います。我々は後の問題用にこれを①番の方法としましょう。

次に学校別に国語の平均を求める場合は、何番の人を利用するでしょうか。これには国語の点数の他に分類に使う学校の番号も必要です。そこで利用される人は 1, 3, 5, 6 になります。国語についてはデータ単位の除去ですが、分類変数が必要ですので、このような除去の方法を（分類変数を除いて）データ単位の除去と呼びましょう。我々はこれを②番の方法とします。

最後に国語と数学についての散布図を描く場合はどうでしょうか。この場合は国語と数学両方のデータが必要ですので、利用される人は 1, 3, 4, 6 となります。このように選んだ変数すべてのデータがそろっていない場合に除去する方法を（選択変数について）レコード単位の除去と言います。ここにレコードとはデータベースで使う言葉で、横 1 列のデータ、即ちここでは個人のデータのことです。我々はこの方法を③番の方法とします。場合によっては②番の方法と③番の方法が重なる場合もありますが、これについては問題の中で話します。

さて教科書の中には、各変数ごとに欠損値を取り除く方法をデータ単位の除去、1

つのレコードに１つでも欠損値があれば取り除く方法をレコード単位の除去とあっさりと書いてある場合がありますが、ここではもう少し現実的に分類しています。

今の話をまとめると以下の表のようになります。

表 1.3.1 集計内容と欠損値の除去方法

集計内容	利用データ	欠損値の除去方法
国語の平均	1,3,4,5,6	①データ単位の除去
学校ごとの国語の平均	1,3,5,6	②（分類変数を除いて）データ単位の除去
国語の数学の散布図	1,3,4,6	③（選択変数について）レコード単位の除去

次の問題は集計の復習ですが、欠損値のある場合の一般的な処理の方法も選択するようになっています。College Analysis では変数選択の Window の中で下図のようにこの処理の方法が選べるようになっていますが、通常は自動にしておけば上のような処理をしてくれます。

図 1.3.1 欠損値の除去

問題

欠損値を含む Samples¥テキスト 9b.txt を用いて、以下の問いに答え、よく使われる欠損値の除去方法について、上の①、②、③のどれに一番近いか右の［　］に答えよ。

１）意見１に関する１次元分割表を描け。　［　　］

意見１:1	意見１:2	合計

２）意見１と意見２に関する２次元分割表を描け。　［　　］

	意見２:1	意見２:2	意見２:3	合計
意見１:1				
意見１:2				
合計				

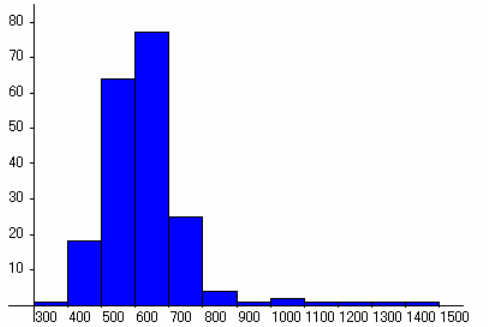
3）年収と支出に関する以下の基本統計量を求めよ。　［　　］

	最小値	最大値	平均値	中央値	標準偏差
年収					
支出					

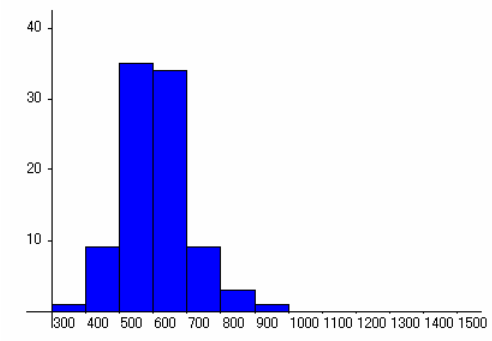
4）地域別の年収に関する基本統計量を求めよ。　［　　］

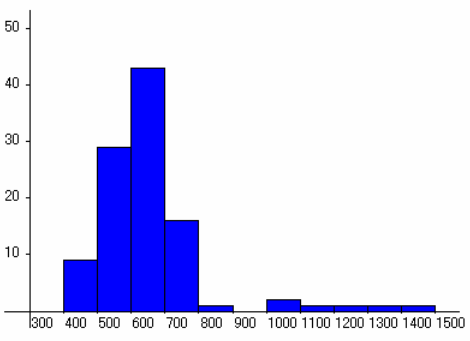
	最小値	最大値	平均値	中央値	標準偏差
地域:1					
地域:2					

5）年収に関するヒストグラムを描け。　［　　］

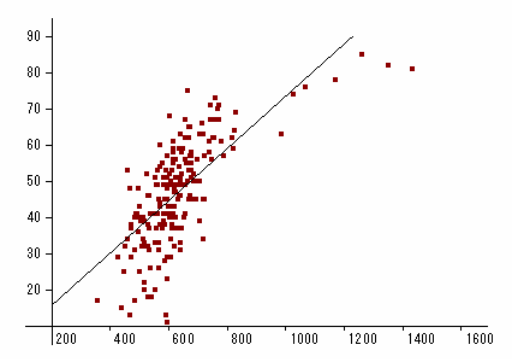


6）地域 1, 2 の年収に関するヒストグラム　［　　］





7）年収と支出に関する散布図を描け。　　［　　］



8）年収と支出に関する相関係数を求めよ。　　［　　］

相関係数＝［　　］

9）支出を年収で予測する回帰式を求めよ。　　［　　］

支出＝［　　］×年収＋［　　］

解答

1）意見１に関する１次元分割表を描け。　　［　①　］

これは１つの変数だけを見ているので、データ単位の除去。

意見１:1	意見１:2	合計
84	113	197

2）意見１と意見２に関する２次元分割表を描け。　　［②または③］

これは意見１を分類変数とみれば（分類変数を除いて）データ単位の除去、意見１と意見２を同等とみれば（選択変数について）レコード単位の除去。

	意見２:1	意見２:2	意見２:3	合計
意見１:1	31	23	29	83
意見１:2	40	31	41	112
合計	71	54	70	195

3）年収と支出に関する以下の基本統計量を求めよ。　　［　①　］

これは同時に２つの変数を見ているようで、実は単独に見ているのでデータ単位の除去。

	最小値	最大値	平均値	中央値	標準偏差
年収	356	1432	631.36	613.5	139.91
支出	11	85	46.79	47.5	14.35

4）地域別の年収に関する基本統計量を求めよ。　　［　②　］

これは地域を分類変数としているので（分類変数を除いて）データ単位の除去

	最小値	最大値	平均値	中央値	標準偏差
地域:1	356	986	607.90	603	97.39
地域:2	426	1432	652.12	619.5	166.59

5）年収に関するヒストグラムを描け。　［　①　］

１変数だけ見ているのでデータ単位の除去。　省略

6）地域 1, 2 の年収に関するヒストグラム　［　②　］

地域を分類変数としているので（分類変数を除いて）データ単位の除去。　省略

7）年収と支出に関する散布図を描け。　［　③　］

年収と支出のどちらも必要なので（選択変数について）レコード単位の除去。

省略

8）年収と支出に関する相関係数を求めよ。　［　③　］

年収と支出のどちらも必要なので（選択変数について）レコード単位の除去。

相関係数＝［ 0.7030 ］

9）支出を年収で予測する回帰式を求めよ。　［　③　］

年収と支出のどちらも必要なので（選択変数について）レコード単位の除去。

支出＝［ 0.0722 ］×年収＋［ 1.3744 ］