

Activity 2.14

[S2008/S2008-DC3]

Sometimes categorical information is represented numerically. In the early days of computing, it was very common to represent everything with a number. For instance the categorical variable for sex, with levels male or female, might be stored as 0 or 1. Even categorical variables like race or language, with many different levels, can be represented as a number. The codebook provides the interpretation of each number (hence the word “codebook”).

Here is a very small part of a dataset from the 1960s used to study the influence of smoking and other factors on the weights of babies at birth.[?] Gestation

gest.	wt	race	ed	wt.1	inc	smoke	number
284	120	8	5	100	1	0	0
282	113	0	5	135	4	0	0
279	128	0	2	115	2	1	1
244	138	7	2	178	98	0	0
245	132	7	1	140	2	0	0
351	140	0	5	120	99	3	2
282	144	0	2	124	2	1	1
279	141	0	1	128	2	1	1
281	110	8	5	99	2	1	2
273	114	7	2	154	1	0	0
285	115	7	2	130	1	0	0
255	92	4	7	125	1	1	5
261	115	3	2	125	4	1	5
261	144	0	2	170	7	0	0

At first glance, all of the data seems quantitative. But read the codebook:

gest. - length of gestation in days

wt - birth weight in ounces (999 unknown)

race - mother's race

0=5=white 6=mex 7=black 8=asian

9=mixed 99=unknown

ed - mother's education

0= less than 8th grade,

1 = 8th -12th grade - did not graduate,

DA
ges

2= HS graduate--no other schooling ,
3= HS+trade,
4=HS+some college
5= College graduate,
6&7 Trade school HS unclear,
9=unknown

marital 1=married, 2= legally separated, 3= divorced,
4=widowed, 5=never married

inc - family yearly income in \$2500 increments
0 = under 2500, 1=2500-4999, ...,
8= 12,500-14,999, 9=15000+,
98=unknown, 99=not asked

smoke - does mother smoke? 0=never, 1= smokes now,
2=until current pregnancy, 3=once did, not now,
9=unknown

number - number of cigarettes smoked per day
0=never, 1=1-4, 2=5-9, 3=10-14, 4=15-19,
5=20-29, 6=30-39, 7=40-60,
8=60+, 9=smoke but don't know, 98=unknown, 99=not asked

Taking into account the codebook, what kind of data is each variable? If the data have a natural order, but are not genuinely quantitative, say "ordinal." You can ignore the "unknown" or "not asked" codes when giving your answer.

Gestation	<u>categorical</u>	<u>ordinal</u>	<u>quantitative</u>	Activity 2.14-1
Race	<u>categorical</u>	<u>ordinal</u>	<u>quantitative</u>	Activity 2.14-2
Marital	<u>categorical</u>	<u>ordinal</u>	<u>quantitative</u>	Activity 2.14-3
Inc	<u>categorical</u>	<u>ordinal</u>	<u>quantitative</u>	Activity 2.14-4
Smoke	<u>categorical</u>	<u>ordinal</u>	<u>quantitative</u>	Activity 2.14-5
Number	<u>categorical</u>	<u>ordinal</u>	<u>quantitative</u>	Activity 2.14-6

The disadvantage of storing categorical information as numbers is that it's easy to get confused and mistake one level for another. Modern software makes it easy to use text strings to label the different levels of categorical variables. Still, you are likely to encounter data with categorical data stored numerically, so be alert.

A good modern practice is to code missing data in a consistent way that can be automatically recognized by software as meaning missing. Often, NA is used for this purpose. Notice that in the number variable, there is a clear order to the categories until one gets to level 9, which means "smoke but don't know." This is an unfortunate choice. It would be better to store number as a quantitative variable telling the number of cigarettes smoked

per day. Another variable could be used to indicate whether missing data was "smoke but don't know," "unknown", or "not asked."