

data manipulation hands-on

Devin Pastoor

Sept 26, 2016

```
library(PKPDmisc)
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##   filter, lag
##
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
library(knitr)
library(ggplot2)

opts_chunk$set(echo = FALSE, warnings = F, message = F, fig.width = 8, fig.height = 6)
```

DATA IMPORT

Objectives:

- Import datasets and documents
- Perform basic data manipulation upon importing the data.

Task-I

Use the .csv files `demog`, `IV`, and `Oral` provided into the data object folder.

1. Read in all three csv files

ID	TIME	DV	AMT	DOSE
1	0.00	NA	100	100
1	0.25	1273.5	NA	100
1	0.50	995.38	NA	100
1	1.00	1254.7	NA	100
1	2.00	1037.6	NA	100
1	3.00	1135.4	NA	100

ID	TIME	DV	AMT	DOSE
1	0.00	NA	100	100
1	0.25	206.29	NA	100
1	0.50	514.88	NA	100
1	1.00	602	NA	100

ID	TIME	DV	AMT	DOSE
1	2.00	1051.2	NA	100
1	3.00	1021.5	NA	100

ID	SEX	WT	AGE	RACE
1	Female	56.8	28	Hispanic
2	Female	58.5	30	Black
3	Male	69.0	48	Black
4	Female	54.8	50	Asian
5	Male	78.1	35	Caucasian
6	Male	67.4	34	Hispanic

DATA MANIPULATION

The goals of this section:

- Use data manipulation tools to prepare the dataset for analysis

Task-II

1. Add a Formulation column and label IV/Oral for each dataset
2. Create one integrated dataset with both IV and Oral data.
3. Rename “DV” column as “COBS”
4. Appropriately merge the demographics dataset into the IV and Oral dataset

Individuals only found in PK:

ID TIME COBS AMT DOSE FORM — — — — —

Individuals found only in demographics:

ID	SEX	WT	AGE	RACE
52	Female	70	33	Asian
51	Male	60	28	Caucasian

Given missing PK information, we will only retain individuals that also have PK data

ID	TIME	COBS	AMT	DOSE	FORM	SEX	WT	AGE	RACE
1	0.00	NA	100	100	IV	Female	56.8	28	Hispanic
1	0.25	1273.5	NA	100	IV	Female	56.8	28	Hispanic
1	0.50	995.38	NA	100	IV	Female	56.8	28	Hispanic
1	1.00	1254.7	NA	100	IV	Female	56.8	28	Hispanic
1	2.00	1037.6	NA	100	IV	Female	56.8	28	Hispanic
1	3.00	1135.4	NA	100	IV	Female	56.8	28	Hispanic

5. Perform the following tasks:

- a. Ensure that the following columns are numeric and not text: TIME, COBS, WT, AGE, AMT and

- DOSEs
- Change the following:
 - fix COBS to be numeric
 - Filter the dataset such that you remove all rows that were non-numeric
 - Create a new column called "GENDER" where:
 - Female = 0
 - Male = 1
 - Create a new column called RACEN where:
 - Caucasian = 0
 - Asian = 1
 - Black = 2
 - Hispanic = 3

```
## Classes 'tbl_df', 'tbl' and 'data.frame': 1200 obs. of 10 variables:
## $ ID : int 1 1 1 1 1 1 1 1 1 1 ...
## $ TIME: num 0 0.25 0.5 1 2 3 4 6 8 12 ...
## $ COBS: chr NA "1273.5" "995.38" "1254.7" ...
## $ AMT : int 100 NA NA NA NA NA NA NA NA ...
## $ DOSE: int 100 100 100 100 100 100 100 100 100 100 ...
## $ FORM: chr "IV" "IV" "IV" "IV" ...
## $ SEX : chr "Female" "Female" "Female" "Female" ...
## $ WT : num 56.8 56.8 56.8 56.8 56.8 56.8 56.8 56.8 56.8 56.8 ...
## $ AGE : int 28 28 28 28 28 28 28 28 28 28 ...
## $ RACE: chr "Hispanic" "Hispanic" "Hispanic" "Hispanic" ...

## Warning in ifelse(!is.na(as.numeric(c(NA, "1273.5", "995.38", "1254.7", , :
## NAs introduced by coercion
```

ID	TIME	COBS	AMT	DOSE	FORM	SEX	WT	AGE	RACE	NNFLG
20	24	BQL	NA	100	IV	Male	80.9	31	Asian	1
20	24	BQL	NA	100	ORAL	Male	80.9	31	Asian	1

- Save the above modifications as a new csv file, dropping the columns SEX and RACE

Descriptive Statistics

Objectives

- How to make summaries of the data using descriptive statistics and other data manipulation tools (dplyr, base R functions etc)

Task III

- Summarize counts stratified by gender and race

SEX	RACE	n
Female	Asian	4
Female	Black	6
Female	Caucasian	13
Female	Hispanic	5
Male	Asian	4
Male	Black	6

SEX	RACE	n
Male	Caucasian	4
Male	Hispanic	8

2. Add a column with the count the number of males/females in the dataset

SEX	CNTGEN
Female	28
Male	22

ID	TIME	COBS	AMT	DOSE	FORM	SEX	WT	AGE	RACE	NNFLG	CNTGEN
1	0	NA	100	100	IV	Female	56.8	28	Hispanic	0	28
2	0	NA	100	100	IV	Female	58.5	30	Black	0	28
3	0	NA	100	100	IV	Male	69.0	48	Black	0	22
4	0	NA	100	100	IV	Female	54.8	50	Asian	0	28
5	0	NA	100	100	IV	Male	78.1	35	Caucasian	0	22
6	0	NA	100	100	IV	Male	67.4	34	Hispanic	0	22

4. summarize the min, mean, and max values for WT, AGE:

- a. non-stratified (WT only)
b. by Gender

minWT	meanWT	maxWT
52.3	64.06795	80.9

SEX	AGE_min	WT_min	AGE_mean	WT_mean	AGE_max	WT_max
Female	20	52.3	36.96429	59.45714	51	69.0
Male	28	64.3	40.49049	69.95856	59	80.9

5. Calculate the Median, 5th, and 95th percentile concentration at each time point for each formulation and dose level. hint: you can use `?quantile` to calculate various quantiles

FORM	DOSE	TIME	q05	q50	q95
IV	100	0.25	822.818000	1715.6000	2750.6600
IV	100	0.50	999.604000	1536.9000	2815.1400
IV	100	1.00	1071.120000	1422.6000	2629.3800
IV	100	2.00	750.034000	1215.9000	1979.6400
IV	100	3.00	789.454000	1059.2000	1802.8400
IV	100	4.00	560.956000	909.4700	1277.5800
IV	100	6.00	356.112000	695.9600	1014.3740
IV	100	8.00	174.490000	528.9600	810.1720
IV	100	12.00	79.720000	280.4500	462.8960
IV	100	16.00	19.021600	153.9800	302.8760
IV	100	24.00	3.807910	55.0940	169.2525
IV	250	0.25	2184.120000	4292.5000	9036.4400
IV	250	0.50	2680.400000	4720.8000	6643.0600

FORM	DOSE	TIME	q05	q50	q95
IV	250	1.00	2657.800000	3892.5000	7872.3200
IV	250	2.00	2096.660000	3062.7000	5637.6800
IV	250	3.00	1644.920000	3096.2000	5029.8600
IV	250	4.00	1300.120000	2928.6000	4014.5600
IV	250	6.00	894.444000	1837.5000	3379.0200
IV	250	8.00	593.554000	1532.7000	2700.6400
IV	250	12.00	254.336000	961.4400	1871.9800
IV	250	16.00	127.992000	617.0800	1133.0000
IV	250	24.00	16.746000	138.3200	644.1220
ORAL	100	0.25	116.238000	243.9600	729.1820
ORAL	100	0.50	219.358000	418.8000	1095.5180
ORAL	100	1.00	356.370000	613.1500	1288.1000
ORAL	100	2.00	459.004000	850.0400	1283.4600
ORAL	100	3.00	504.412000	733.0300	1096.0600
ORAL	100	4.00	417.484000	779.8000	1037.5400
ORAL	100	6.00	249.916000	511.2300	823.0760
ORAL	100	8.00	156.652000	434.4900	754.0580
ORAL	100	12.00	69.605600	235.0000	461.8440
ORAL	100	16.00	22.982000	131.5400	237.3840
ORAL	100	24.00	4.061125	43.5375	129.3972
ORAL	250	0.25	275.032000	709.5700	1837.4600
ORAL	250	0.50	522.846000	1152.3000	2785.5800
ORAL	250	1.00	703.254000	1824.5000	4055.5800
ORAL	250	2.00	1194.720000	2307.1000	3532.7200
ORAL	250	3.00	1362.160000	2366.9000	4142.9800
ORAL	250	4.00	1269.500000	2248.5000	3805.1200
ORAL	250	6.00	807.978000	1771.1000	2807.2200
ORAL	250	8.00	695.654000	1296.5000	1922.8600
ORAL	250	12.00	268.334000	855.5100	1432.2400
ORAL	250	16.00	113.282000	466.3600	970.7700
ORAL	250	24.00	11.168400	144.8100	558.9920

BONUS: visualize the concentration time data with the quantiles underlaid



