Dirty Data? Clean it using SAS

An Introduction to Data Cleaning Principles

CYP-C Research Champion Webinar
August 11, 2017
Giancarlo Di Giuseppe, MPH
Pediatric Oncology Group of Ontario

Outline

- SAS overview and procedures – revisited
- Fundamental principles to build a clean dataset
- Inclusion / exclusion criteria
- Visualizing data distributions
- Outliers
- Invalid or inconsistent character variables
- Dealing with missing data
- Creating data checkpoints

SAS Overview - Revisited

- For our purposes only two major things you can do in SAS
  - DATA step - Manipulate the data in some way
    - Reading in Data
    - Creating and Redefining Variables
    - Sub-Setting Data
    - Working with Dates
    - Working with Formats
  - PROCedure step
    - Analyze the data
    - Produce frequency tables
    - Estimate a regression model

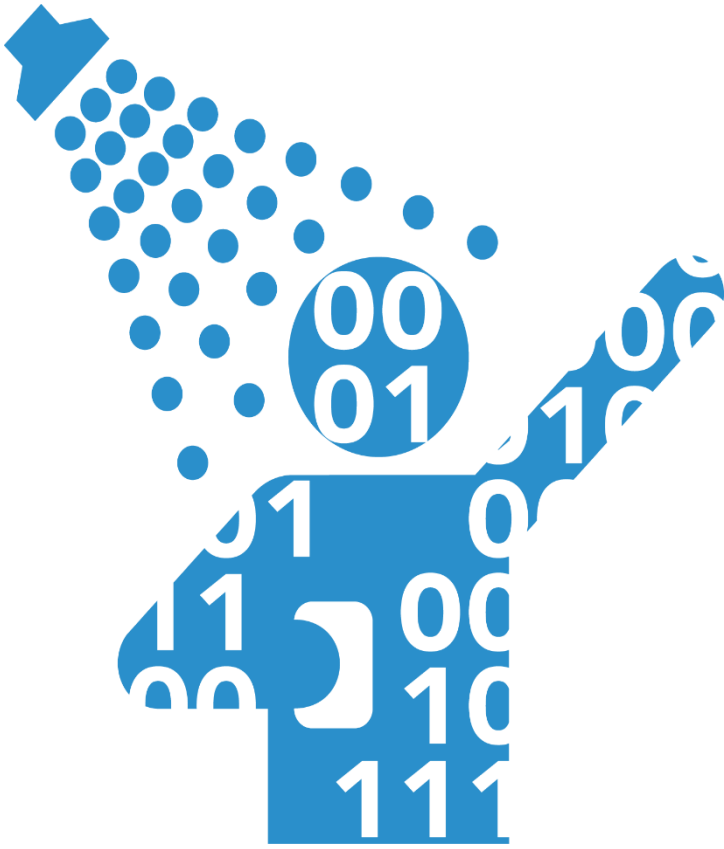
SAS Procedures – Revisited

- SAS Procedures
  - PROC FREQ
  - PROC PRINT
  - PROC MEANS
  - PROC UNIVARIATE
  - PROC SORT
  - PROC CONTENTS

PRINCIPLES FOR CLEANLINESS

Understanding Your Dirty Data Source

- No database is initially ever "*clean*"
- Databases are not constructed with our own specific research questions in mind
- Researchers must be familiar with the purpose, how variables are captured and defined, and the structure of the database



http://3rdsectorlabs.com/wp-content/uploads/2014/06/TSL-data-shower.png

Having an Analysis Plan

- Having clean data requires a sound analysis plan
  - Envision what the analysis dataset will look like with all variables and formats before performing data cleaning

- Determine what your study population denominator is **before** you begin cleaning
  - Is it patient population? Is it number of total diagnoses (therefore, multiple dx's per patient is possible)? Or is it person-time? Etc.
  - Based on the research question!

Data Manipulation and Data Cleaning: A Simultaneous Process

- Data manipulation and data cleaning are not mutually exclusive, rather they go hand-in-hand!

- Both can (and should) be performed within a single DATA step

http://i.telegraph.co.uk/multimedia/archive/03219/handshake1_3219777k.jpg

- Ensures efficient and easy to follow SAS programming

SUB-SETTING YOUR DATA

Receiving Your Data Cut

- Typically data is requested with slightly more information than needed
  - Allows for wiggle room if hypothesis change slightly
- No data cut is ever perfect
  - Data still needs to be cleaned
- Initial data cuts are never ready to be analyzed, they must first be cleaned

Cleaning Using Inclusion & Exclusion Criteria

```
PROC SORT DATA = T7 OUT=T7_SORT; BY CYPCID DX_DATE; RUN;

DATA T8; SET T7_SORT; BY CYPCID;
     /* INTERESTED IN PRIMARY DIAGNOSIS ONLY */
     IF FIRST.CYPCID;
     IF ORDINAL_PRIMARY IN (1);

     /* AGE INCLUSION CRITERIA - 0 TO 14 */
     IF 0 <= DX_AGE < 15;

     IF 0 <= DX_AGE < 1 THEN DX_AGE_GR=1;
          ELSE IF DX_AGE < 7 THEN DX_AGE_GR=2;
          ELSE IF DX_AGE < 11 THEN DX_AGE_GR=3;
     ELSE DX_AGE_GR=4;
     LABEL DX_AGE_GR = "AGE AT FIRST DIAGNOSIS - GROUPED";
     FORMAT DX_AGE_GR DX_AGE_GR.;

     /* SELECTS THOSE WITH A DIAGNOSIS BETWEEN 2002 & 2012 */
     IF 2002 <= YEAR(DX_DATE) <= 2012;
     DX1_YEAR = YEAR(DX_DATE);

     /* LEUKEMIA CASES */
     IF ICCC_MAIN = 1010 OR ICDO_M_CODE IN (9826, 9835, 9836,
     9837);

RUN; *N=2,492;
```

First cancers

Children aged 0 to 14

Note: Data cleaning and data manipulation done simultaneously!

Diagnosed between 2002-2012

Only concerned with leukemia cases

Keep logs of sample size in your DATA steps!!

DATA DISTRIBUTION

Recall From Last Session

- PROC FREQ produces frequency outputs
  - Can be used for numeric *or* character variables
  - Useful for counts and proportions
- PROC MEANS and UNIVARIATE produce outputs describing the data distribution for **numeric** variables
  - Checkpoint for data distributions and normality
- PROC FREQ and PROC MEANS/UNIVARIATE are used in the first step of data cleaning to <u>understand the data</u>

Duplicate Entries

```
PROC FREQ DATA=T8 ORDER=FREQ;
    TABLE CYPCID /MISSING;
RUN;
```

The FREQ Procedure

CYPCID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10015	1	0.04	1	0.04
10020	1	0.04	2	0.08
10022	1	0.04	3	0.12
10027	1	0.04	4	0.16
10044	1	0.04	5	0.20
10050	1	0.04	6	0.24
10052	1	0.04	7	0.28

Distribution of Continuous Data

```
ODS GRAPHICS ON;
PROC UNIVARIATE DATA = T8 NORMAL;
    ID CYPCID;
    VAR WBC_COUNT;
    HISTOGRAM WBC_COUNT / NORMAL;
RUN;
```

The UNIVARIATE Procedure
Variable: wbc_count

Moments

N	2008	Sum Weights	2008
Mean	1427.86822	Sum Observations	2867159.39
Std Deviation	24523.1909	Variance	601386894
Skewness	23.0529387	Kurtosis	582.218619
Uncorrected SS	1.21108E12	Corrected SS	1.20698E12
Coeff Variation	1717.46878	Std Error Mean	547.261788

Basic Statistical Measures

Location		Variability	
Mean	1427.868	Std Deviation	24523
Median	11.700	Variance	601386894
Mode	2.500	Range	720000
		Interquartile Range	44.01500

POGO
PEDIATRIC ONCOLOGY GROUP OF ONTARIO

Distribution of Continuous Data II

Tests for Normality

Test		--Statistic---		-----p Value------
Kolmogorov-Smirnov	D	0.498938	Pr > D	<0.0100
Cramer-von Mises	W-Sq	163.6925	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	758.7259	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	720000.000
99%	940.000
95%	323.800
90%	173.500
75% Q3	48.515
50% Median	11.700
25% Q1	4.500
10%	2.400
5%	1.650
1%	0.700
0% Min	0.000

Distribution of Continuous Data III

The UNIVARIATE Procedure
Variable: wbc_count

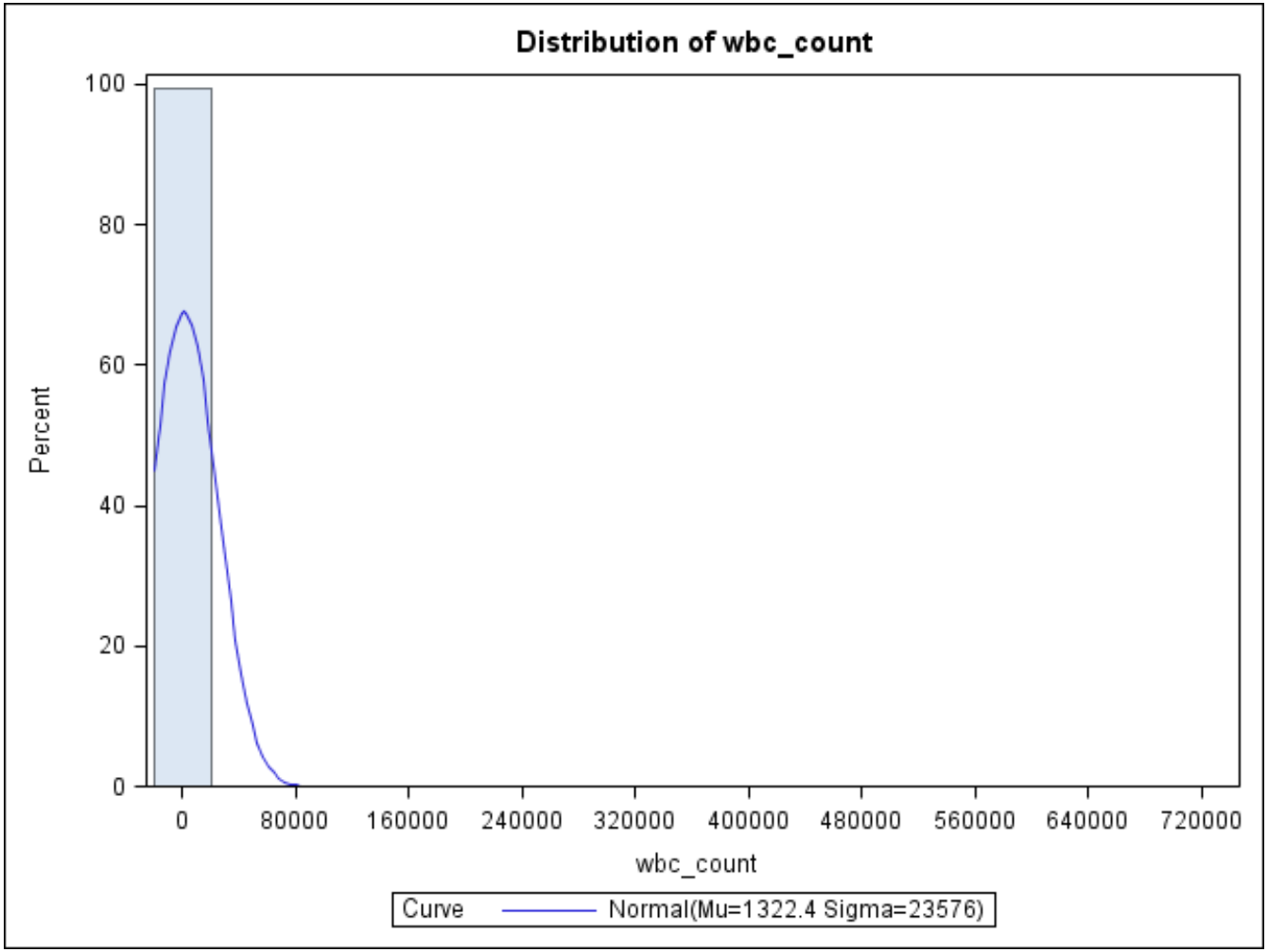
Extreme Observations

Lowest			Highest		
Value	CYPCID	Obs	Value	CYPCID	Obs
0.0	13486	375	190000	12667	279
0.0	12973	319	202000	13664	391
0.0	12941	315	470000	10881	103
0.1	13623	384	570000	11423	159
0.2	A001141	1099	720000	12086	216

Missing Values

Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	484	19.42	100.00

Normality of Continuous Data



OUTLIERS

Dealing With Outliers

- If there are many outliers, these will introduce bias in your study
- Many options to handle these skewed data:
  - Median + IQR instead of mean
  - Use a logical range of values and assign any outlier the upper bound of the range
  - Categorize your data based on the distribution or clinically meaningful ranges
- Whichever approach used should be justified!

Dealing With Outliers II

```
DATA T8; SET T8;
    /* UPPER LIMIT TO OUTLIERS */
    IF WBC_COUNT >= 500 THEN WBC_COUNT_CLEAN = 500;
    ELSE WBC_COUNT_CLEAN = WBC_COUNT;

    /* CREATING CLINICAL CATEGORIES */
    IF WBC_COUNT ^= . THEN DO;
    IF WBC_COUNT < 50 THEN WBC_GROUP = 1;
        ELSE IF WBC_COUNT < 100 THEN WBC_GROUP = 2;
        ELSE IF WBC_COUNT < 200 THEN WBC_GROUP = 3;
        ELSE IF WBC_COUNT < 300 THEN WBC_GROUP = 4;
        ELSE IF WBC_COUNT < 400 THEN WBC_GROUP = 5;
    ELSE WBC_GROUP = 6; END;
RUN;
```

DO loop

Dealing With Outliers III

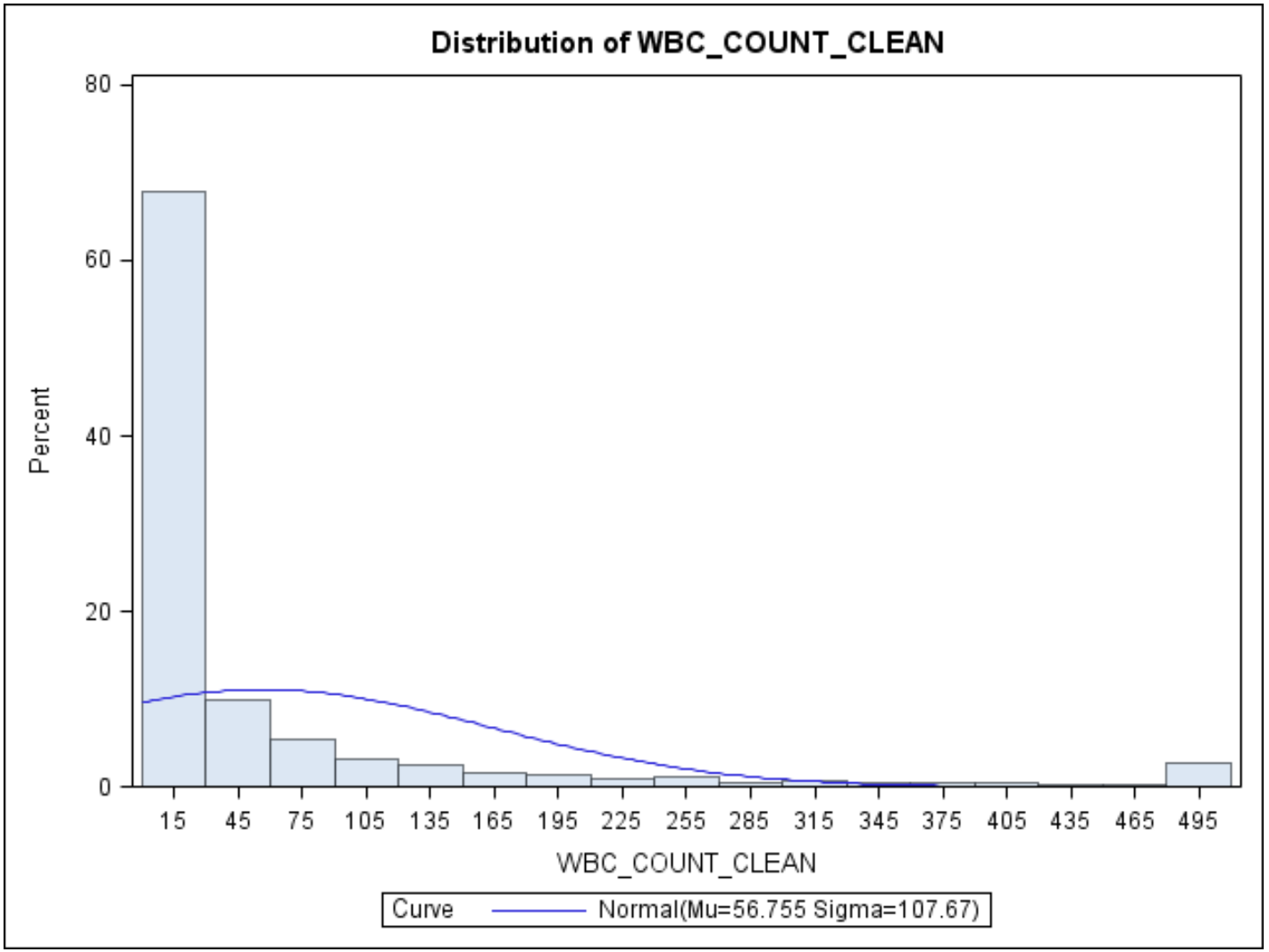
```
/* MEAN VS MEDIAN + IQR */
PROC MEANS DATA=T8 MEAN MIN MAX Q1 MEDIAN Q3;
     VAR WBC_COUNT_CLEAN;
RUN;

/* DATA CATEGORIZATION */
PROC FREQ DATA=T8;
     TABLES WBC_GROUP /MISSING;
RUN;
```

Analysis Variable : WBC_COUNT_CLEAN

Mean	Minimum	Maximum	Lower Quartile	Median	Upper Quartile
56.7551245	0	500.0000000	4.5000000	11.7000000	48.5150000

Dealing With Outliers III



Dealing With Outliers III

```
/* MEAN VS MEDIAN + IQR */
PROC MEANS DATA=T8 MEAN MIN MAX Q1 MEDIAN Q3;
     VAR WBC_COUNT_CLEAN;
RUN;

/* DATA CATEGORIZATION */
PROC FREQ DATA=T8;
     TABLES WBC_GROUP /MISSING;
RUN;
```

The FREQ Procedure

WBC_GROUP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	484	19.42	484	19.42
1	1513	60.71	1997	80.14
2	176	7.06	2173	87.20
3	147	5.90	2320	93.10
4	63	2.53	2383	95.63
5	38	1.52	2421	97.15
6	71	2.85	2492	100.00

CLEANING CHARACTER VARIABLES

CApITOLIzATioN Matters!

```
PROC FREQ DATA=T8;
    TABLES PROTOCOL_NAME;
RUN;
```

ALL PROTOCOL A	47	0.61
ALL PROTOCOL AB	33	0.43
ALL PROTOCOL B	27	0.35
ALL PROTOCOL BFM-90	1	0.01
ALL PROTOCOL C	129	1.67
ALL PROTOCOL C & CCSG107	1	0.01
ALL PROTOCOL C (107)	1	0.01
ALL PROTOCOL C (CCG 107)	1	0.01
ALL PROTOCOL C - MODIFIED	1	0.01
ALL PROTOCOL C MODIFIED	1	0.01
ALL PROTOCOL C- (CCG-107)	1	0.01
ALL PROTOCOL C- MODIFIED	1	0.01
ALL PROTOCOL C-MODIFIED	2	0.03
ALL PROTOCOL C/CCG 107	2	0.03
ALL PROTOCOL C/CCG107	3	0.04
ALL PROTOCOL C/MODIFIED CCG107	1	0.01
ALL PROTOCOL CC 1891	1	0.01
ALL PROTOCOL_BERLIN	1	0.01
ALL Protocol C	2	0.03
ALL Standard Risk 1991	1	0.01
ALL Standard Risk POG 9605	1	0.01
ALL protocol C	1	0.01

CAPITOLIZATION Matters! Use UPCASE

```
DATA T8; SET T8;
    PROTOCOL_NAME = UPCASE(PROTOCOL_NAME);
RUN;
PROC FREQ DATA=T8; TABLES PROTOCOL_NAME; RUN;
```

ALL PROTOCOL A	47	0.61
ALL PROTOCOL AB	33	0.43
ALL PROTOCOL B	27	0.35
ALL PROTOCOL BFM-90	1	0.01
ALL PROTOCOL C	132	1.70
ALL PROTOCOL C & CCSG107	1	0.01
ALL PROTOCOL C (107)	1	0.01
ALL PROTOCOL C (CCG 107)	1	0.01
ALL PROTOCOL C - MODIFIED	1	0.01
ALL PROTOCOL C MODIFIED	1	0.01
ALL PROTOCOL C- (CCG-107)	1	0.01
ALL PROTOCOL C- MODIFIED	1	0.01
ALL PROTOCOL C-MODIFIED	2	0.03
ALL PROTOCOL C/CCG 107	2	0.03
ALL PROTOCOL C/CCG107	3	0.04
ALL PROTOCOL C/MODIFIED CCG107	1	0.01
ALL PROTOCOL CC 1891	1	0.01
ALL PROTOCOL- BERLIN	1	0.01
ALL STANDARD RISK 1991	1	0.01
ALL STANDARD RISK POG 9605	1	0.01

FINDing, Cleaning, and Manipulating

```
DATA T9; SET T8;
    PROTOCOL_NAME = UPCASE(PROTOCOL_NAME);

    IF FIND(PROTOCOL_NAME,"ALL PROTOCOL C") THEN DO;
        PROTOCOL_NAME = "ALL PROTOCOL C";
        ALL_RISK = "HIGH RISK";
    END;
RUN;
PROC FREQ DATA=T9; TABLES PROTOCOL_NAME; RUN;
```

DO loop

ALL PROTOCOL A	47	0.61
ALL PROTOCOL AB	33	0.43
ALL PROTOCOL B	27	0.35
ALL PROTOCOL BFM-90	1	0.01
ALL PROTOCOL C	163	2.10
ALL PROTOCOL- BERLIN	1	0.01
ALL STANDARD RISK 1991	1	0.01
ALL STANDARD RISK POG 9605	1	0.01

Use Caution When Searching Text

- When performing character search functions in SAS, be wary of the phrase being used
- Can lead to errors in data cleaning
- Searched term should be unique enough to prevent unwanted matches
- If "ALL PROTOCOL B" was searched using FIND(), then the BFM-90 protocol would have been misclassified as Protocol B

ALL PROTOCOL A	47	0.61
ALL PROTOCOL AB	33	0.43
ALL PROTOCOL B	27	0.35
ALL PROTOCOL BFM-90	1	0.01
ALL PROTOCOL C	163	2.10
ALL PROTOCOL- BERLIN	1	0.01
ALL STANDARD RISK 1991	1	0.01
ALL STANDARD RISK POG 9605	1	0.01

MISSING DATA

Recall: Viewing Missing Data

```
PROC FREQ DATA =  T8;
    TABLES STAGE_CODE /MISSING;
RUN;
```

The FREQ Procedure

STAGE_CODE

STAGE_CODE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	7494	69.36	7494	69.36
1	268	2.48	7762	71.84
1 A	2	0.02	7764	71.86
1 B	1	0.01	7765	71.87
1/2	2	0.02	7767	71.89
11	1	0.01	7768	71.90
1A	64	0.59	7832	72.49

Understanding Your Missing Data

```
PROC FREQ DATA =  T8;
    WHERE DX1_GRP = 2;
    TABLES STAGE_CODE /MISSING;
RUN;
```

- Staging not done for the leukemia's which represent a high % of childhood cancers
- Staging important for lymphomas
- Know your data!

The FREQ Procedure

STAGE_CODE

STAGE_CODE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	534	32.07	534	32.07
1	44	2.64	578	34.71
1 A	2	0.12	580	34.83
1 B	1	0.06	581	34.89
11	1	0.06	582	34.95
1A	56	3.36	638	38.32

DATA CHECKPOINTS

Date Checkpoints I

```
DATA FLAGS; SET T8 (KEEP=PATIENT_ID DOB DX_DATE1 DOD);
    IF DOD < DX_DATE1 AND DOD ^=. THEN DEATH_FLAG = 1;
        ELSE DEATH_FLAG=0;
    IF DX_DATE1 < DOB THEN DX_FLAG = 1;
        ELSE DX_FLAG = 0;
RUN;
PROC FREQ DATA=FLAGS; TABLES DEATH_FLAG DX_FLAG; RUN;
```

The FREQ Procedure

DEATH_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10799	99.95	10799	99.95
1	5	0.05	10804	100.00

DX_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10804	100.00	10804	100.00

Date Checkpoints II

```
PROC PRINT DATA=T8 NOOBS;
    WHERE DOD < DX_DATE1 AND DOD ^=. ;
    VAR PATIENT_ID DX_DATE1 DOD;
RUN;
```

Treatment Checkpoints

```
DATA TX_FLAGS;
    MERGE T8 (IN=MASTER) CHEMO (IN=A) SURG (IN=B)
          BMT (IN=C) RAD (IN=D);

    BY CYPCID;

    IF A THEN CHEMO = 1; ELSE CHEMO = 0;
    IF B THEN SURGERY = 1; ELSE SURGERY = 0;
    IF C THEN BMT = 1; ELSE BMT = 0;
    IF D THEN RAD = 1; ELSE RAD = 0;

    NUM_TX_MODALITIES = SUM(CHEMO,SURGERY,BMT,RAD);

    IF FIRST.CYPCID;
    IF MASTER THEN OUTPUT;
RUN;
```

Treatment flags

REMEMBER: All datasets involved in a merge **must** be sorted by the common identifier (ie.CYPCID)

Treatment Checkpoints II

```
PROC FREQ DATA=TX_FLAGS;
    TABLES DX1_GRP * (CHEMO SURGERY BMT RAD);
    TABLES DX1_GRP * NUM_TX_MODALITIES;
RUN;
```

Frequency
Row Pct
Col Pct

Table of DX1_GRP by CHEMO

by SURGERY

DX1_GRP(DIAGNOSIS OF FIRST NEOPLASM GROUPED)	CHEMO 0. NO	CHEMO 1. YES	CHEMO Total	SURGERY 0. NO	SURGERY 1. YES	SURGERY Total
1. LEUK	107	3037	3144	3046	98	3144
	3.40	96.60		96.88	3.12	
	3.50	39.21		57.09	1.79	
2. LYMPHOMA AND RETIC	327	1338	1665	844	821	1665
	19.64	80.36		50.69	49.31	
	10.69	17.28		15.82	15.01	
3. CNS	1475	1045	2520	915	1605	2520
	58.53	41.47		36.31	63.69	
	48.22	13.49		17.15	29.35	
4. SNS TUMORS AND RETINO	346	542	888	156	732	888
	38.96	61.04		17.57	82.43	
	11.31	7.00		2.92	13.38	
5. KIDNEY	83	453	536	26	510	536
	15.49	84.51		4.85	95.15	
	2.71	5.85		0.49	9.33	

Treatment Checkpoints II

```
PROC FREQ DATA=TX_FLAGS;
    TABLES DX1_GRP * (CHEMO SURGERY BMT RAD);
    TABLES DX1_GRP * NUM_TX_MODALITIES;
RUN;
```

Frequency
Row Pct
Col Pct

Table of DX1_GRP by NUM_TX_MODALITIES

DX1_GRP(DIAGNOSIS OF FIRST NEOPLASM GROUPED)	NUM_TX_MODALITIES 0	1	2	3	4	Total
1. LEUK	103	1756	905	356	24	3144
	3.28	55.85	28.78	11.32	0.76	
	12.80	39.41	25.71	20.86	7.59	
2. LYMPHOMA AND RETIC	120	552	673	283	37	1665
	7.21	33.15	40.42	17.00	2.22	
	14.91	12.39	19.12	16.58	11.71	
3. CNS	426	993	590	436	75	2520
	16.90	39.40	23.41	17.30	2.98	
	52.92	22.28	16.76	25.54	23.73	
4. SNS TUMORS AND RETINO	35	370	234	102	147	888
	3.94	41.67	26.35	11.49	16.55	
	4.35	8.30	6.65	5.98	46.52	
5. KIDNEY	6	92	224	205	9	536
	1.12	17.16	41.79	38.25	1.68	
	0.75	2.06	6.36	12.01	2.85	

Topics Covered

- Key principles to build a clean dataset
- Using Inclusion / exclusion criteria
- Visualizing data distributions
- Handling data outliers
- Cleaning character variables
- Dealing with missing data
- Creating data checkpoints

THANK YOU!