GETTING STARTED WITH SAS (PART 2)

Christine Mauro, PhD

April 19th, 2018

Outline

- Application
- Brief Review
- Descriptive Statistics
- Visualizing Data
- Basic Hypothesis Testing

APPLICATION

Risk Factors Associated with Low Birthweight

- Lowbwt_ALL.csv

- The data on 189 births were collected at Baystate Medical Center, Springfield, Mass. during 1986. The dataset contains an indicator of low infant birth weight as a response and several risk factors associated with low birth weight. The actual birth weight is also included in the dataset.

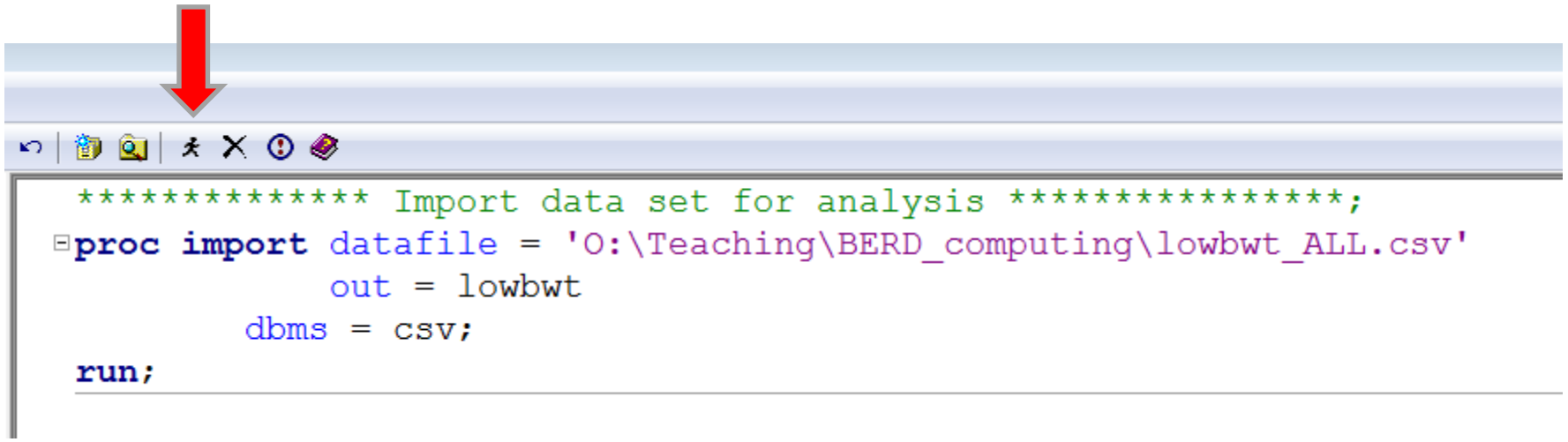
- The dataset consists of the following 10 variables:
  - low: indicator of birth weight less than 2.5kg
  - age: mother's age in years
  - lwt: mother's weight in pounds at last menstrual period
  - race: mothers race ("white", "black", "other")
  - smoke: smoking status during pregnancy
  - ht: history of hypertension
  - ui: presence of uterine irritability
  - ftv: number of physician visits during the first trimester
  - ptl: number of previous premature labours
  - bwt: birth weight in grams

BRIEF REVIEW

Day 1 Topics

- Running SAS Code
- Importing Data
- Examining Data Attributes
- Manipulating Data

Running Code



```
*************** Import data set for analysis ****************;
proc import datafile = 'O:\Teaching\BERD_computing\lowbwt_ALL.csv'
            out = lowbwt
        dbms = csv;
run;
```

PROC CONTENTS

Data Set Name	WORK.LOWBWT	Observations	189
Member Type	DATA	Variables	10
Engine	V9	Indexes	0
Created	04/18/2018 12:05:56	Observation Length	80
Last Modified	04/18/2018 12:05:56	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	817
Obs in First Data Page	189
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\cmm2212\AppData\Local\Temp\SAS Temporary Files_TD6128_SPH-F4RGZ12-BIO_\lowbwt.sas7bdat
Release Created	9.0401M0
Host Created	X64_7PRO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
2	age	Num	8	BEST12.	BEST32.
10	bwt	Num	8	BEST12.	BEST32.
8	ftv	Num	8	BEST12.	BEST32.
6	ht	Num	8	BEST12.	BEST32.
1	low	Num	8	BEST12.	BEST32.
3	lwt	Num	8	BEST12.	BEST32.
9	ptl	Num	8	BEST12.	BEST32.
4	race	Char	5	$5.	$5.
5	smoke	Num	8	BEST12.	BEST32.

```
proc contents data = lowbwt;
  run;
```

PROC PRINT

```
proc print data = lowbwt;
run;
```

The SAS System

Obs	low	age	lwt	race	smoke	ht	ui	ftv	ptl	bwt
1	0	19	182	black	0	0	1	0	0	2523
2	0	33	155	other	0	0	0	1	0	2551
3	0	20	105	white	1	0	0	1	0	2557
4	0	21	108	white	1	0	1	1	0	2594
5	0	18	107	white	1	0	1	0	0	2600
6	0	21	124	other	0	0	0	0	0	2622
7	0	22	118	white	0	0	0	1	0	2637
8	0	17	103	other	0	0	0	1	0	2637
9	0	29	123	white	1	0	0	1	0	2663
10	0	26	113	white	1	0	0	0	0	2665
11	0	19	95	other	0	0	0	0	0	2722
12	0	19	150	other	0	0	0	1	0	2733
13	0	22	95	other	0	1	0	0	0	2750
14	0	30	107	other	0	0	1	1	1	2750

DESCRIPTIVE STATISTICS

PROC MEANS

- Primarily used for reporting sums and means of numeric (continuous) variables

Syntax:

```
proc means data=data_name;
  ... <options>;
run;
```

- Without options, it will calculate the summary statistics for all numeric variables
  - Default statistics: N (number of non-missing obs), Mean, Std Dev, Min and Max

PROC MEANS: Options

- MAXDEC = *n* – number of decimal places to be displayed
- MISSING – treats missing values as valid summary groups
- NMISS – number of missing values
- MEDIAN, Q1 (25th percentile), Q3 (75th percentile)
- QRANGE
- MODE
- RANGE
- SUM

PROC MEANS: VAR statement

- You can control which variables to include in the report using VAR statement
  - Variables that have the same name (with the exception of a number at the end can be listed together)

Syntax:

```
proc means data=data_name;
  var var1 var2 var3 …;
run;
```

PROC MEANS: Example 1

```
proc means data=lowbwt;
  var bwt;
run;
```

The MEANS Procedure

Analysis Variable : bwt				
N	Mean	Std Dev	Minimum	Maximum
189	2944.66	729.0224169	709.0000000	4990.00

```
proc means data=lowbwt n mean std median q1 q3 ndec=2;
  var bwt;
run;
```

The MEANS Procedure

Analysis Variable : bwt					
N	Mean	Std Dev	Median	Lower Quartile	Upper Quartile
189	2944.66	729.02	2977.00	2414.00	3475.00

PROC MEANS: BY Statement

- Produces descriptive statistics for each level of another (categorical) variable
- Make sure to sort your data set by the same variable(s) listed in the BY statement

Syntax:

```
proc means data=data_name;
  var var1 var2 var3 …;
  by cat_var;
run;
```

- CLASS statement – similar to BY; no need to sort first, and generates more compact output (one table)

```
proc means data=data_name;
  var var1 var2 var3 …;
  class cat_var;
run;
```

PROC MEANS: Example 2

```
proc means data=lowbwt n mean std median q1 q3 ndec=2;
  class race;
  var bwt;
run;
```

The MEANS Procedure

Analysis Variable : bwt							
race	N Obs	N	Mean	Std Dev	Median	Lower Quartile	Upper Quartile
black	26	26	2719.69	638.68	2849.00	2367.00	3062.00
other	67	67	2804.01	721.30	2835.00	2301.00	3274.00
white	96	96	3103.74	727.72	3076.00	2575.50	3651.00

PROC FREQ

- Can be used to count frequencies of both character and numeric variables in one-way, two-way, and three-way tables
- Typically used for categorical variables
- Can create output data sets containing counts and percentages
- Computes various stats such as chi-squared test/Fisher's Exact test, odds ratio, and relative risk (more details in future lectures)

Syntax:

```
proc freq data=data_name;
                ... <options>;
run;
```

PROC FREQ

- For n-way contingency table, separate each variable name with * in the TABLES statement.

Syntax:

```
proc freq data=data_name;
  tables var1*var2;
  tables var3*var1*var2;
run;
```

- Variable 1st listed (var1) forms the rows of the table, 2nd forms the columns
- 3rd variable listed (var3) creates multiple tables – stratification by levels of var3

PROC FREQ: Example 1

```
proc freq data=lowbwt;
  tables race smoke;
run;
```

The SAS System

The FREQ Procedure

race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
black	26	13.76	26	13.76
other	67	35.45	93	49.21
white	96	50.79	189	100.00

smoke	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	115	60.85	115	60.85
1	74	39.15	189	100.00

PROC FREQ: Example 2

```
proc freq data=lowbwt;
  tables race*smoke;
run;
```

The SAS System

The FREQ Procedure

Table of race by smoke

Frequency Percent Row Pct Col Pct	smoke		
race	0	1	Total
black	16 8.47 61.54 13.91	10 5.29 38.46 13.51	26 13.76
other	55 29.10 82.09 47.83	12 6.35 17.91 16.22	67 35.45
white	44 23.28 45.83 38.26	52 27.51 54.17 70.27	96 50.79
Total	115 60.85	74 39.15	189 100.00

PROC FREQ: Example 3

```
proc freq data=lowbwt;
  tables ht*race*smoke;
run;
```

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table 1 of race by smoke Controlling for ht=0		
	smoke		
race	0	1	Total
black	14 7.91 60.87 12.96	9 5.08 39.13 13.04	23 12.99
other	51 28.81 80.95 47.22	12 6.78 19.05 17.39	63 35.59
white	43 24.29 47.25 39.81	48 27.12 52.75 69.57	91 51.41
Total	108 61.02	69 38.98	177 100.00

Frequency Percent Row Pct Col Pct	Table 2 of race by smoke Controlling for ht=1		
	smoke		
race	0	1	Total
black	2 16.67 66.67 28.57	1 8.33 33.33 20.00	3 25.00
other	4 33.33 100.00 57.14	0 0.00 0.00 0.00	4 33.33
white	1 8.33 20.00 14.29	4 33.33 80.00 80.00	5 41.67
Total	7 58.33	5 41.67	12 100.00

PROC FREQ: Options

- Use any of these in the TABLES statement:

```
tables var1 / <options>; * one-way table;
or
tables var1*var2 / <options>; * two-way table;
```

- NOFREQ – suppress cell frequencies
- NOPERCENT - suppress cell percentages
- NOCUM - suppress cumulative counts and percentages
- NOROW - suppress row percentages (for n-way tables)
- NOCOL - suppress column percentages (for n-way tables)

DATA VISUALIZATION

Visualizing Data

- Histograms
- Boxplots
- Scatterplots

PROC SGPLOT

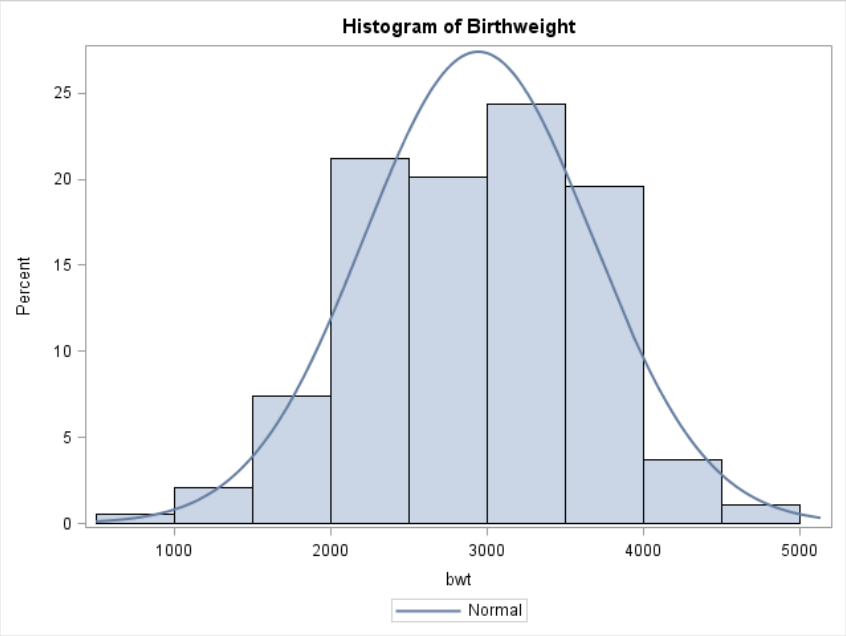
- Creates one or more plots and overlays them on a single set of axes
- Graphics include: scatterplots, line plots, histograms, boxplots, regression plots, etc.

Syntax:

```
proc sgplot data=data_name;
statement variable_name/ <options>;
run;
```

- Statement specifies the type of graph to construct

PROC SGPLOT: Histogram 1

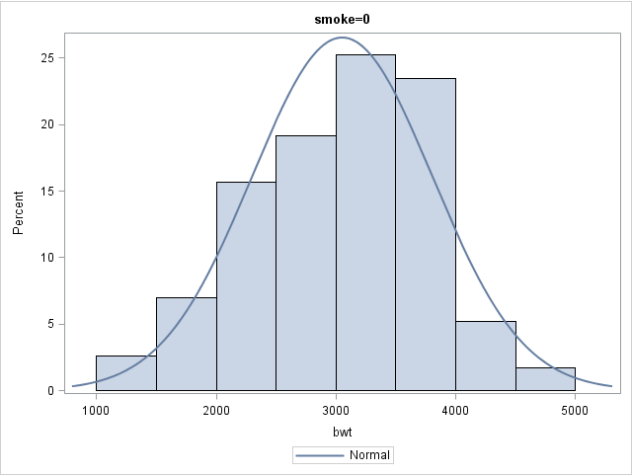


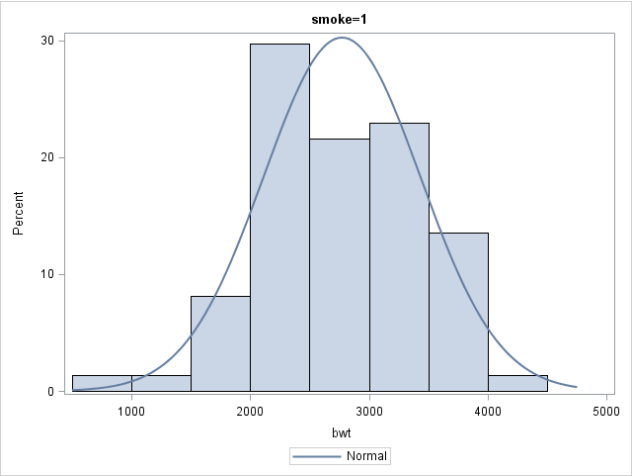
SAS code:

```
proc sgplot data=lowbwt;
  histogram bwt;
  density bwt;
  title "Histogram of Birthweight";
run;
title;
```

* Use to view continuous variables

PROC SGPLOT: Histogram 2





SAS code:

```
proc sort data=lowbwt;
   by smoke;
run;

proc sgplot data=lowbwt;
   by smoke;
   histogram bwt;
   density bwt;
run;
```

* Use to view continuous variables by a categorical variable

PROC SGPLOT: Options for histogram

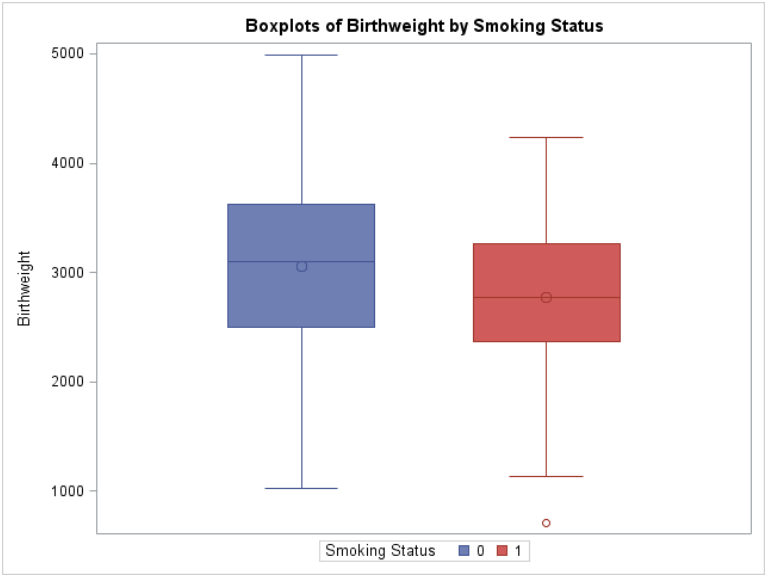
- binstart = n; n is the midpoint of first bin
- binwidth = n; n is the bin width
- nbins = n; n is the number of bins
- scale = percent, count, or proportion;
- showbins; places tick marks at midpoints of bins

PROC SGPLOT: Boxplots 1

SAS code:

```
proc sgplot data=lowbwt;
   vbox bwt / group=smoke ;
   keylegend / title="Smoking Status";
   yaxis label="Birthweight";
   title "Boxplots of Birthweight by Smoking Status";
run;
title;
```

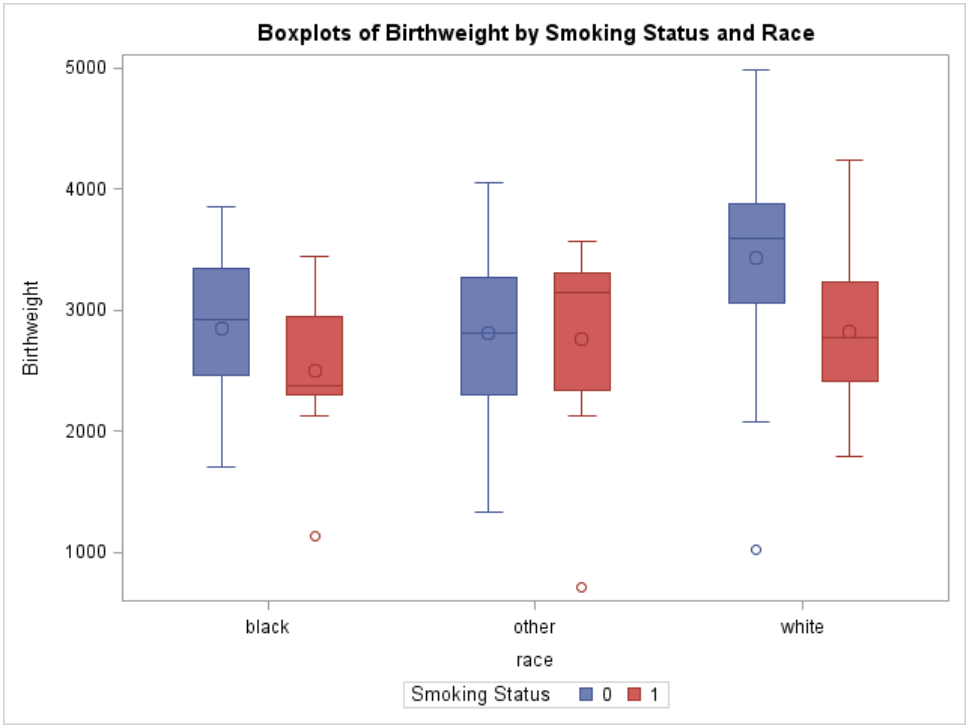
- **bwt** is the continuous variable
- **Smoke** is the categorical variable
- **vbox / hbox** – options for vertical /horizontal boxplots



* Use to view continuous variables by a categorical variable

PROC SGPLOT: Boxplots 2

```
proc sgplot data=lowbwt;
  vbox bwt / group=smoke category=race;
  keylegend / title="Smoking Status";
  yaxis label="Birthweight";
  title "Boxplots of Birthweight by Smoking Status and Race";
run;
title;
```



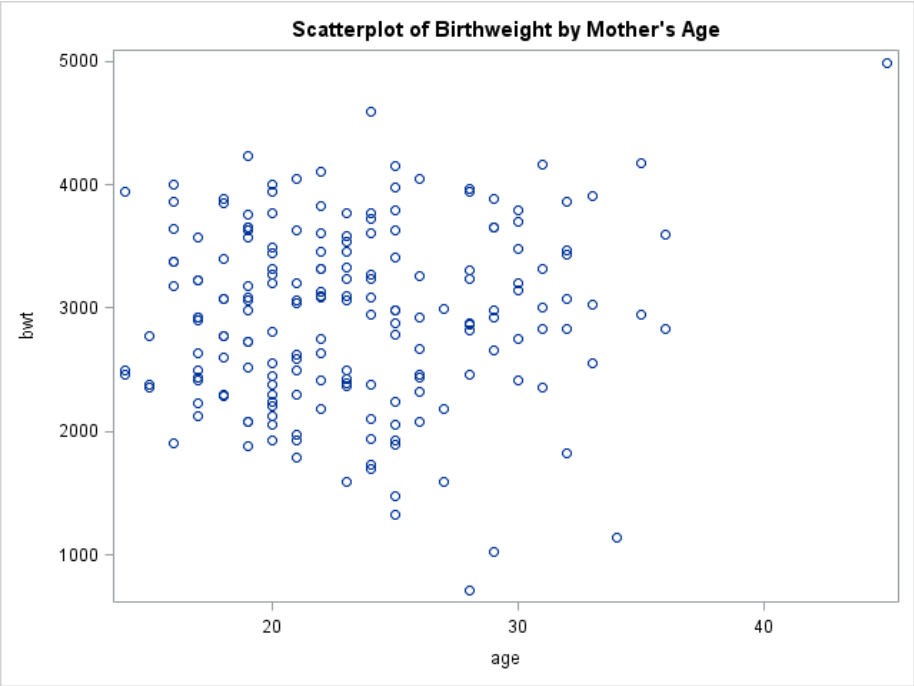
* Use to view continuous variables by **two** categorical variables

PROC SGPLOT: Scatterplot 1

SAS code:

```
proc sgplot data=lowbwt;
 scatter x=age y=bwt ;
 title "Scatterplot of Birthweight by Mother's
Age";
run;
title;
```

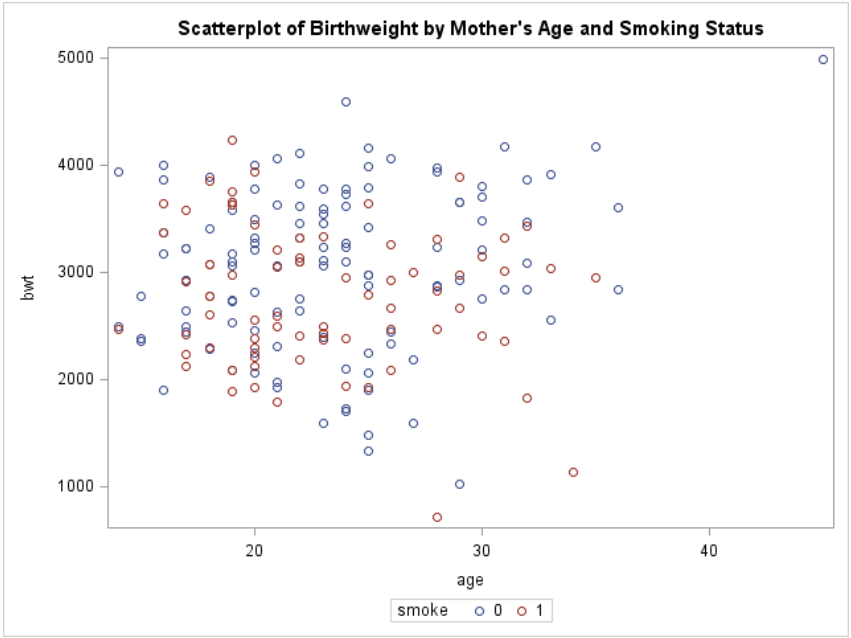
* Use to view a continuous variable by a continuous variable



PROC SGPLOT: Scatterplot 2

SAS code:

```
proc sgplot data=lowbwt;
 scatter x=age y=bwt / group=smoke;
 title "Scatterplot of Birthweight by Mother's Age and
Smoking Status";
run;
title;
```



* Use to view a continuous variable by a continuous variable at different levels of a categorical variable

PROC UNIVARIATE

- Also produces simple summary statistics (mean, standard deviation)
- Generates histograms, boxplots, normal probability plots (QQ plots)
- Conducts tests for normality

Syntax:

```
proc univariate data=data_name;
  <options>;
run;
```

PROC UNIVARIATE: Ex 1

```
proc univariate data=lowbwt;
   var bwt;
run;
```

Moments			
N	189	Sum Weights	189
Mean	2944.65608	Sum Observations	556540
Std Deviation	729.022417	Variance	531473.684
Skewness	-0.210171	Kurtosis	-0.0814157
Uncorrected SS	1738735950	Corrected SS	99917052.6
Coeff Variation	24.7574724	Std Error Mean	53.0285779

Basic Statistical Measures			
Location		Variability	
Mean	2944.656	Std Deviation	729.02242
Median	2977.000	Variance	531474
Mode	2495.000	Range	4281
		Interquartile Range	1061

Note: The mode displayed is the smallest of 4 modes with a count of 4.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	55.52961	Pr > |t|	<.0001
Sign	M	94.5	Pr >= |M|	<.0001
Signed Rank	S	8977.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	4990
99%	4593
95%	3997
90%	3884
75% Q3	3475
50% Median	2977
25% Q1	2414
10%	1970
5%	1790
1%	1021
0% Min	709

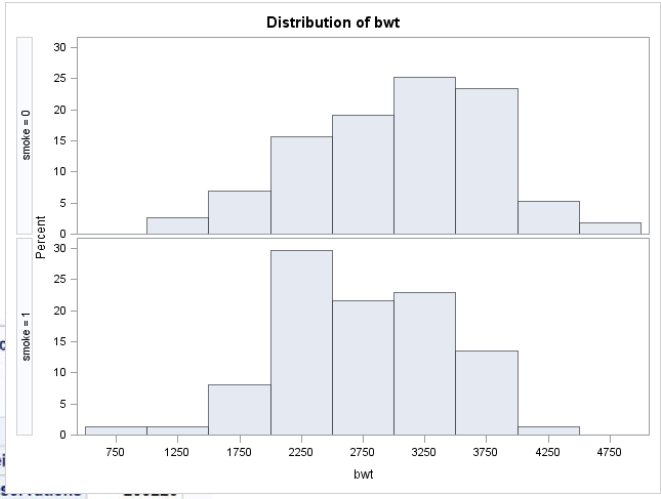
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
709	160	4167	83
1021	87	4174	84
1135	161	4238	159
1330	88	4593	85
1474	89	4990	86

PROC UNIVARIATE: Options

- **PLOTS** option in the PROC UNIVARIATE line
  - Gives a stem-and-leaf plot, a boxplot, and a QQ plot
- **HISTOGRAM** statement generates histograms and offers several options (e.g., change the width of the bars, the colors, etc.)
- **PROC UNIVARIATE** is great for checking distributional assumptions (normality) and generate summary statistics for continuous variables

PROC UNIVARIATE: Ex 2

```
proc univariate data=lowbwt;
   class smoke;
   var bwt;
   histogram bwt ;
run;
```



The UNIVARIATE Procedure
Variable: bwt
smoke = 0

Moments			
N	115	Sum Weights	115
Mean	3054.95652	Sum Observations	351320
Std Deviation	752.409013	Variance	566119.323
Skewness	-0.2911419	Kurtosis	-0.2268633
Uncorrected SS	1137804928	Corrected SS	64537602.8
Coeff Variation	24.6291234	Std Error Mean	70.1625022

Basic Statistical Measures			
Location		Variability	
Mean	3054.957	Std Deviation	752.40901
Median	3100.000	Variance	566119
Mode	2920.000	Range	3969
		Interquartile Range	1134

Note: The mode displayed is the smallest of 3 modes with a count of 3.

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	43.54116	Pr > |t|	<.0001
Sign	M	57.5	Pr >= |M|	<.0001
Signed Rank	S	3335	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	4990

The UNIVARIATE Proc
Variable: bwt
smoke = 1

Moments			
N	74	Sum Wei	
Mean	2773.24324	Sum Obs	
Std Deviation	660.075168	Variance	435699.228
Skewness	-0.2955374	Kurtosis	0.42038264
Uncorrected SS	600931022	Corrected SS	31806043.6
Coeff Variation	23.8015605	Std Error Mean	76.7321797

Basic Statistical Measures			
Location		Variability	
Mean	2773.243	Std Deviation	660.07517
Median	2775.500	Variance	435699
Mode	2466.000	Range	3529
		Interquartile Range	893.00000

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	36.14185	Pr > |t|	<.0001
Sign	M	37	Pr >= |M|	<.0001
Signed Rank	S	1387.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	4238.0
99%	4238.0

BASIC HYPOTHESIS TESTING

T-tests in SAS

- PROC TTEST
  - Can be used for one-, two-sample
  - Independent and paired samples
- Always remember to check model assumptions before inferences
  - Normality (for small samples)
  - Independent observations within the group(s) (not repeated)
- For a T-test, check the normality assumption
  - QQPLOT is a statement in PROC UNIVARIATE
  - Listing NORMAL in the PROC line will output tests for normality (*with a p-value > α, the null hypothesis that the data came from a normally distributed population cannot be rejected*)

One-sample T-test

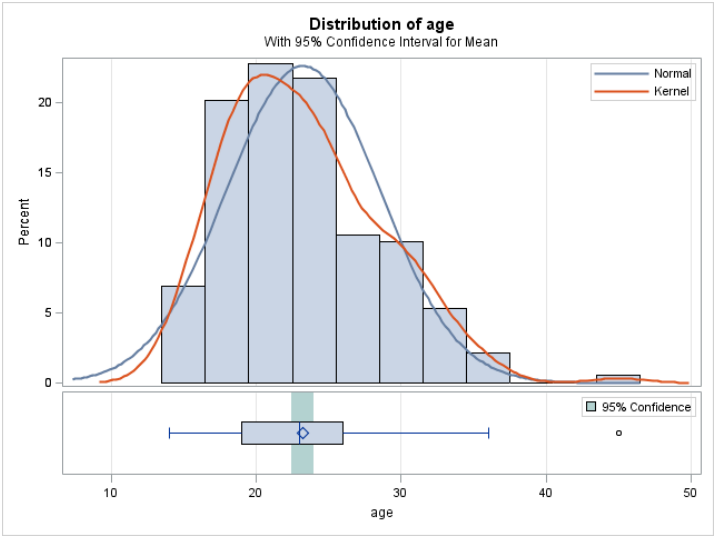
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

```
proc ttest data=data_name HO=nullvalue;
     var var_name;
run;
```

- The default value for H_0 is zero (μ_0=0)
- The default α is 0.05
  - Specify alpha=… in the PROC line
- The default hypothesis is two-sided (≠) (SIDES=2)
  - Specify SIDES=L (<) or SIDES=U (>) in the PROC line

One-sample Example

- Is the average age of mother's in this sample different from 26?




```
proc ttest data=lowbwt H0=26;
  var age;
run;
```

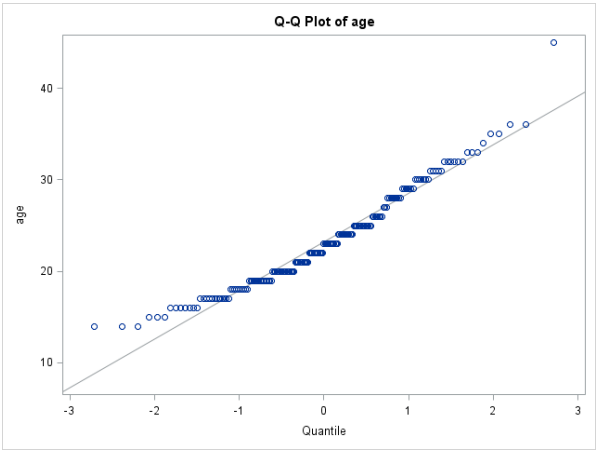
The TTEST Procedure

Variable: age

N	Mean	Std Dev	Std Err	Minimum	Maximum
189	23.2381	5.2987	0.3854	14.0000	45.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
23.2381	22.4778	23.9984	5.2987	4.8129	5.8943

DF	t Value	Pr > |t|
188	-7.17	<.0001



Outcome: Continuous

Two-sample independent T-test

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

```
proc ttest data=data_name HO=nullvalue;
  class group_variable;
  var var_name;
run;
```

- In the two-sample case, remember to test for the equality of variances!!
  - Equality of Variance test is part of the output when you use a CLASS statement

Two-sample Example

- Is the birthweight of babies born to smokers significantly different than the birthweight to babies born to non-smokers?

The TTEST Procedure

Variable: bwt

smoke	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	115	3055.0	752.4	70.1625	1021.0	4990.0
1	74	2773.2	660.1	76.7322	709.0	4238.0
Diff (1-2)		281.7	717.8	107.0		

smoke	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3055.0	2916.0	3193.9	752.4	666.1	864.6
1		2773.2	2620.3	2926.2	660.1	568.2	787.7
Diff (1-2)	Pooled	281.7	70.6927	492.7	717.8	651.8	798.7
Diff (1-2)	Satterthwaite	281.7	76.4668	487.0			

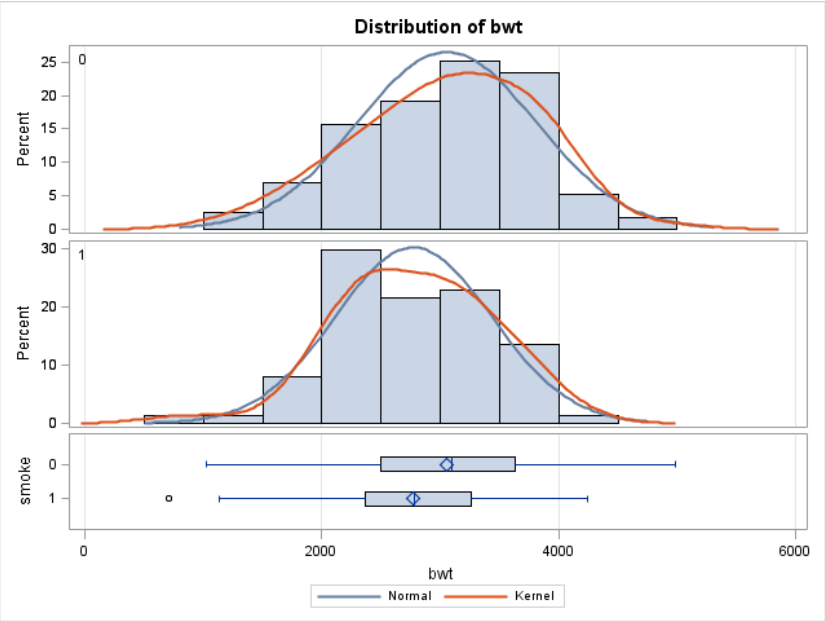
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	187	2.63	0.0092
Satterthwaite	Unequal	170	2.71	0.0074

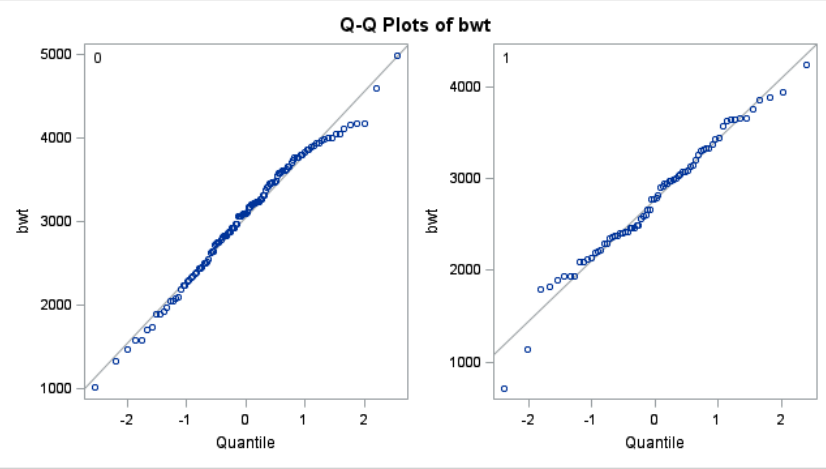
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	114	73	1.30	0.2290

Outcome: Continuous
Predictor: Binary

Two-sample Example

- Is the birthweight of babies born to smokers significantly different than the birthweight to babies born to non-smokers?





Two-sample paired T-test

$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 \neq \mu_2 \text{ or } \mu_1 - \mu_2 \neq 0$$

```
proc ttest data=data_name HO=nullvalue;
     paired before*after;
run;

proc ttest data=data_name HO=nullvalue;
    var diff; /* diff=after-before */
run;
```

Analysis of Variance (ANOVA)

- If we want to compare the (continuous) outcomes of 3 or more groups
- Model assumptions:
  - Independent samples
  - Responses within the groups are independent and identically distributed (i.i.d.)
  - Residuals are normally distributed
  - Equality of variances across groups

One-Way ANOVA

- **One** categorical factor with multiple levels
- Use PROC GLM
  - Can also accommodate multiple factors and interactions

- Simplest syntax (only one categorical factor):

```
proc glm data=data_name;
  class categorical_factor;
  model continuous_response=categorical_factor;
run;
```

ANOVA: Example

- Does birthweight significantly differ by race?

```
proc glm data=lowbwt;
   class race;
   model bwt=race;
run;
```

Outcome: Continuous

Predictor: Categorical
(3 or more levels)

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5070607.63	2535303.82	4.97	0.0079
Error	186	94846445.01	509927.12		
Corrected Total	188	99917052.65			

R-Square	Coeff Var	Root MSE	bwt Mean
0.050748	24.25043	714.0918	2944.656

Source	DF	Type I SS	Mean Square	F Value	Pr > F
race	2	5070607.632	2535303.816	4.97	0.0079

Source	DF	Type III SS	Mean Square	F Value	Pr > F
race	2	5070607.632	2535303.816	4.97	0.0079

Categorical Data Analysis

- Categorical outcome (Y) with 2 levels (binary) or ≥3 levels (nominal or ordinal)
- Examples:
  - Nominal: race/ethnicity
  - Ordinal: clothing sizes (S, M, L, XL)
  - Binary: Disease/ No Disease; Republican/Democrat

- Predictor variables (X) can take on any form: binary, categorical, and/or continuous

Chi-squared Test of Independence

- Use two categorical variables (row and column) to test whether they are independent or associated
- Hypotheses:

H0: variables A and B are independent

vs

H1: variables A and B are not independent

Test statistics: $$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Chi-squared Test of Independence

- Create a (r X c) table
  - r represents the number of levels for the row variable
  - c represents the number of levels for the column variable
  - Most common example is a 2 x 2 table
- Use the observed and expected counts in each cell to calculate the chi-squared statistics
- If low expected cell counts (< 5), use Fisher's Exact test instead

Chi-squared Test in SAS

Syntax:

```
proc freq data=data_name;
      tables var1*var2/expected exact chisq;
run;
```

- `Chisq`–computes Chi-squared statistics/p-value
- `Exact`–computes Fisher's Exact test/p-value
- `Expected`–produces expected values

Chi-Squared Test: Example

- Is there an association between history of uterine irritability and having a low birthweight baby?

```
proc freq data=lowbwt;
  table ui*low / chisq
expected exact;
run;
```

Outcome: Categorical
Predictor: Categorical

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct	Table of ui by low		
	low		
ui	0	1	Total
0	116 110.74 61.38 72.05 89.23	45 50.259 23.81 27.95 76.27	161 85.19
1	14 19.259 7.41 50.00 10.77	14 8.7407 7.41 50.00 23.73	28 14.81
Total	130 68.78	59 31.22	189 100.00

Fisher's Exact Test	
Cell (1,1) Frequency (F)	116
Left-sided Pr <= F	0.9934
Right-sided Pr >= F	0.0198
Table Probability (P)	0.0132
Two-sided Pr <= P	0.0269

Statistics for Table of ui by low

Statistic	DF	Value	Prob
Chi-Square	1	5.4008	0.0201
Likelihood Ratio Chi-Square	1	5.0761	0.0243
Continuity Adj. Chi-Square	1	4.4227	0.0355
Mantel-Haenszel Chi-Square	1	5.3722	0.0205
Phi Coefficient		0.1690	
Contingency Coefficient		0.1667	
Cramer's V		0.1690	

Other Options for Proc Freq

- **/agree** McNemar's test for matched pair analysis (proportion)
- **/chm** Chi-square test for 2X2Xk table (k strata) Breslow-Day Test, Cochran-Mantel Haenszel test
- **/noprint** Do not print tables in the output
- **/trend** Cochran-Armitage trend test for proportions

Need advanced methods?

- Elementary Statistics Using SAS (Schlotzhauer).
- Categorical Data Analysis Using SAS, 3rd edition (Stokes, Davis, and Koch).
- UCLA Institute for Digital Research and Education

https://stats.idre.ucla.edu/sas/

Thank you!

BERD EDU link:
http://irvinginstitute.columbia.edu/resources/biostat_educational_initiatives.html

Acknowledgements: Cody Chiuzan, PhD