



Agriculture and
Agri-Food Canada

Agriculture et
Agroalimentaire Canada



Getting What You Want from PROC MEANS and PROC UNIVARIATE

Marjorie Smith, Cereal Research Centre

Canada

PROC MEANS

- provides data summarization tools to compute descriptive statistics for variables
 - across all observations
 - within groups of observations

PROC UNIVARIATE

- Used to explore the data distributions of variables
 - summarize, visualize, analyze, and model the statistical distributions of numeric variables

MOISTURE CONTENT OF WHEAT GRAIN HARVESTED AT DIFFERENT MATURITIES AND STORED UNDER VARIOUS CONDITIONS

DAY	RH	HT	REP	MC %
1	65	E	1	12.72
1	65	E	2	12.71
1	65	E	3	13.07
1	75	E	1	13.92
1	75	E	2	14.72
1	75	E	3	14.36
1	85	E	1	15
1	85	E	2	15.35
1	85	E	3	15.54
2	65	E	1	13.23
2	65	E	2	
2	65	E	3	13.3
2	75	E	1	15.01

....more data lines....

- DAY = days in storage
- RH = % relative humidity
- HT = harvest time (E = early, L = late, N = normal)
- MC % = % moisture content

Data set input and format

```
proc format;
```

```
  value $fht 'E'='early' 'L'='late' 'N'='normal';
```

```
data a;
```

```
  infile 'example_data.txt' dlm='09'x dsd firstobs=4 missover;
```

```
  input day rh ht $ rep mc;
```

```
  label rh='% RH' ht='harvest time' mc='% moisture';
```

```
  format ht fht.;
```

```
title 'MOISTURE CONTENT OF WHEAT GRAIN HARVESTED AT  
DIFFERENT MATURITIES';
```

```
title2 'AND STORED UNDER VARIOUS CONDITIONS';
```

```
proc means data=a;
```

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum
day		198	3.8181818	2.1296249	0
rh	% RH	198	74.5454545	8.2656070	65.0000000
rep		198	2.0000000	0.8185663	1.0000000
mc	% moisture	198	15.2317677	1.9559897	10.3600000

Variable	Label	Maximum
day		7.0000000
rh	% RH	85.0000000
rep		3.0000000
mc	% moisture	22.7300000

```
proc means data=a fw=8;  
var mc;
```

The MEANS Procedure

Analysis Variable : mc %% moisture

N	Mean	Std Dev	Minimum	Maximum
198	15.2318	1.9560	10.3600	22.7300

```
proc means data=a n mean var fw=8 maxdec=3;
  by ht rh day;
  var mc;
```

```
harvest time=early RH=65 day=1
```

```
The MEANS Procedure
```

```
Analysis Variable : mc moisture
```

N	Mean	Variance
3	12.833	0.042

```
harvest time=early RH=65 day=2
```

```
Analysis Variable : mc moisture
```

N	Mean	Variance
2	13.265	0.002

```
harvest time=early RH=65 day=3
```

```
Analysis Variable : mc moisture
```

N	Mean	Variance
3	12.323	2.891

- Each BY group listed in a separate table
- Difficult to compare statistics for each group if there are many groups

```
proc means data=a n mean var fw=8 maxdec=3;
  class ht rh day;
  var mc;
```

Analysis Variable : mc moisture						
harvest time	RH	day	N Obs	N	Mean	Variance
early	65	1	3	3	12.833	0.042
		2	3	2	13.265	0.002
		3	3	3	12.323	2.891
		4	3	3	13.363	0.005
		5	3	3	13.630	0.022
		6	3	3	13.630	0.008
		7	3	3	13.313	0.009
	75	1	3	3	14.333	0.161
		2	3	3	15.500	0.292
		3	3	3	15.867	0.933
		4	3	3	15.030	0.292
		5	3	3	15.697	0.002
		6	3	3	15.843	0.002
		7	3	3	15.533	0.020
	85	1	3	3	15.297	0.075

CLASS statement

Statistics for groups
listed in one table

Formatting

- fw = field width (default is 12)
- Maxdec = number of decimal places shown

```

proc means data=a n mean var
  fw=8 maxdec=3;
  by ht;
  class rh day;
  var mc;

```

Each BY group produces a separate table

```
harvest time=early
```

```
The MEANS Procedure
```

```
Analysis Variable : mc moisture
```

RH	day	N Obs	N	Mean	Variance
65	1	3	3	12.833	0.042
	2	3	2	13.265	0.002
	3	3	3	12.323	2.891
	4	3	3	13.363	0.005
	5	3	3	13.630	0.022
	6	3	3	13.630	0.008
	7	3	3	13.313	0.009
75	1	3	3	14.333	0.161
	2	3	3	15.500	0.292
	3	3	3	15.867	0.933
	4	3	3	15.030	0.292
	5	3	3	15.697	0.002
	6	3	3	15.843	0.002
	7	3	3	15.533	0.020
85	1	3	3	15.297	0.075
	2	3	3	16.597	0.042
	3	3	3	17.207	0.021
	4	3	3	17.900	0.023
	5	3	3	18.120	0.050
	6	3	3	18.210	0.052
	7	3	3	18.127	0.010

Using CLASSDATA to display a subset of variable combinations

```
data classtypes;
  input ht $ rh;
  format ht fht.;
  datalines;
E 65
E 85
L 65
L 85
;
proc means data=a n mean var fw=8
  maxdec=3 classdata=classtypes
  exclusive;
class ht rh;
var mc;
```

The MEANS Procedure

Analysis Variable : mc moisture

harvest time	RH	N Obs	N	Mean	Variance
early	65	21	20	13.191	0.521
	85	21	21	17.351	1.084
late	65	21	21	13.908	0.163
	85	21	20	17.581	0.648

- EXCLUSIVE restricts the class levels to the variable combinations that are in the 'classtypes' dataset

Output data sets

- Use the OUTPUT statement to:
 - Have greater control over how the output data looks
 - Save the output statistics to a SAS data set you can manipulate
 - Use more than one OUTPUT statement to create several OUT= data sets
- If you only want the OUT= data set, use the NOPRINT option in the PROC MEANS statement

```

proc means data=a alpha=0.05 mean lclm uclm noprint;
  by ht rh;
  var mc;
  output out=out1 n=n mean=MeanMoistureContent
    lclm=LowerLimit uclm=UpperLimit;

```

```

proc print;

```

Obs	ht	rh	_TYPE_	_FREQ_	n	Mean Moisture Content	Lower Limit	Upper Limit
1	early	65	0	21	20	13.1905	12.8527	13.5283
2	early	75	0	21	21	15.4005	15.0982	15.7028
3	early	85	0	21	21	17.3510	16.8771	17.8248
4	late	65	0	21	21	13.9081	13.7244	14.0918
5	late	75	0	21	21	14.8186	14.3294	15.3078
6	late	85	0	21	20	17.5805	17.2036	17.9574
7	normal	65	0	21	21	13.2052	13.0364	13.3741
8	normal	75	0	21	21	15.5800	14.6031	16.5569
9	normal	85	0	21	21	16.9500	16.2775	17.6225

```
proc means data=a alpha=0.05 mean lclm uclm noprint;
  class ht rh;
  var mc;
  output out=out2 n=n mean=MeanMoistureContent
    lclm=LowerLimit uclm=UpperLimit;
```

```
data b;
  set out2;
  format MeanMoistureContent LowerLimit UpperLimit 4.1;
```

ht	rh	_TYPE_	_FREQ_	n	Mean Moisture Content	Lower Limit	Upper Limit
.	.	0	189	187	15.3	15.1	15.6
.	65	1	63	62	13.4	13.3	13.6
.	75	1	63	63	15.3	14.9	15.6
.	85	1	63	62	17.3	17.0	17.6
early	.	2	63	62	15.3	14.9	15.8
late	.	2	63	62	15.4	15.0	15.8
normal	.	2	63	63	15.2	14.7	15.8
early	65	3	21	20	13.2	12.9	13.5
early	75	3	21	21	15.4	15.1	15.7
early	85	3	21	21	17.4	16.9	17.8
late	65	3	21	21	13.9	13.7	14.1
late	75	3	21	21	14.8	14.3	15.3
late	85	3	21	20	17.6	17.2	18.0
normal	65	3	21	21	13.2	13.0	13.4
normal	75	3	21	21	15.6	14.6	16.6
normal	85	3	21	21	17.0	16.3	17.6

TYPE automatic variable

- unique value for each combination of class variables
- indicates which combination of the class variables PROC MEANS uses to compute the statistics

The available keywords
to include in the
PROC statement

Specifies which
statistics to compute
and the order to
display them in the
output

(list from SAS 9.2 documentation –
PROC MEANS)

Descriptive statistic keywords

CLM	NMISS
CSS	RANGE
CV	SKEWNESS SKEW
KURTOSIS KURT	STDDEV STD
LCLM	STDERR
MAX	SUM
MEAN	SUMWGT
MIN	UCLM
MODE	USS
N	VAR

Quantile statistic keywords

MEDIAN P50	Q3 P75
P1	P90
P5	P95
P10	P99
Q1 P25	QRANGE

Hypothesis testing keywords

PROBT PRT	T
-----------	---

TESTING FOR RESISTANCE TO WHEAT MIDGE IN SEVERAL WHEAT LINES

line	position	instar2	instar3	dead
3001	19	0	37	3
3001	19	1	50	3
3001	11	4	14	14
3001	11	2	0	8
3002	22	0	0	1
3002	22	0	0	6
3002	18	0	0	0
3002	18	0	0	3
3040	25	1	46	2
3040	25	0	50	0
3040	16	0	50	0
3040	16	0	50	0
3024	4	2	30	11
3024	4	1	50	2

.....more data lines.....

- position = position in cage
- instar2 = second stage larvae
- instar3 = third stage larvae
- dead = dead larvae

```

proc summary mean std sum;
  class line;
  var instar2 instar3 dead;
  output out=out1(rename=( _freq_ =SampleSize))
    mean(instar3)=MeanInstar3 std(instar3)=StdInstar3
    sum=TotalInstar2 TotalInstar3 TotalDead;

```

```
proc print data=out1;
```

MIDGE INFESTATION RATE ON SPIKES OF SEVERAL WHEAT LINES

Obs	line	_TYPE_	Sample Size	Mean Instar3	Std Instar3	Total Instar2	Total Instar3	Total Dead
1		0	27	20.4444	20.8000	19	552	83
2	3001	1	3	21.3333	25.7941	7	64	25
3	3002	1	4	0.0000	0.0000	0	0	10
4	3024	1	4	31.2500	14.7281	8	125	32
5	3026	1	4	4.2500	4.3493	1	17	2
6	3034	1	2	3.0000	2.8284	0	6	2
7	3035	1	2	29.0000	5.6569	2	58	1
8	3037	1	2	43.0000	9.8995	0	86	1
9	3040	1	4	49.0000	2.0000	1	196	2
10	3053	1	2	0.0000	0.0000	0	0	8

PROC SUMMARY

- virtually the same as PROC MEANS but the default option is NOPRINT

Rename the automatic variable _FREQ_

OUT= data set includes mean and std for stage 3 larvae only (instar3)

PROC UNIVARIATE

- descriptive statistics:
 - Moments, quantiles or percentiles, frequency tables, extreme values
- histograms
- goodness-of-fit tests for a variety of distributions
- create output data sets containing summary statistics, histogram intervals, and parameters of fitted curves

- An important first step in data analysis:
 - find key features of distributions
 - identify outliers and extreme observations
 - determine the need for data transformations
 - compare distributions

(summarized from SAS 9.2 documentation – PROC UNIVARIATE)



Canada 