1. Create the following SAS dataset on 5 college students:

**DATA** COLLEGE;

INPUT ID AGE GENDER $ GPA CSCORE;

DATALINES;

1 18 M 3.7 650

2 18 F 2.0 490

3 19 F 3.3 580

4 23 M 2.8 530

5 21 M 3.5 640

;

1. Add statements necessary to compute the mean grade point average and mean college entrance exam score.
2. We want to compute an index for each subject, as follows:

INDEX = GPA + 3\*CSCORE/500;

Code:

**DATA** COLLEGE;

INPUT ID AGE GENDER $ GPA CSCORE;

INDEX=GPA + **3**\*CSCORE/**500**;

DATALINES;

1 18 M 3.7 650

2 18 F 2.0 490

3 19 F 3.3 580

4 23 M 2.8 530

5 21 M 3.5 640

;

**run**;

Title 'GPA & CSCORE Mean';

**proc** **means** data=COLLEGE mean;

var GPA CSCORE;

output out = my\_college

mean=M\_GPA CSCORE\_GPA;

**run**;

Title;

Title 'COLLEGE';

**proc** **print** data=COLLEGE;

**run**;

Title;

Log file:

4 DATA COLLEGE;

5 INPUT ID AGE GENDER $ GPA CSCORE;

6 INDEX=GPA + 3\*CSCORE/500;

7 DATALINES;

NOTE: The data set WORK.COLLEGE has 5 observations and 6 variables.

NOTE: DATA statement used (Total process time):

real time 0.02 seconds

cpu time 0.03 seconds

13 ;

14 run;

15

16 Title 'GPA & CSCORE Mean';

17

18 proc means data=COLLEGE mean;

19 var GPA CSCORE;

20 output out = my\_college

21 mean=M\_GPA CSCORE\_GPA;

22 run;

NOTE: Writing HTML Body file: sashtml.htm

NOTE: There were 5 observations read from the data set WORK.COLLEGE.

NOTE: The data set WORK.MY\_COLLEGE has 1 observations and 4 variables.

NOTE: PROCEDURE MEANS used (Total process time):

real time 0.80 seconds

cpu time 0.43 seconds

23

24 Title;

25

26 Title 'COLLEGE';

27 proc print data=COLLEGE;

28 run;

NOTE: There were 5 observations read from the data set WORK.COLLEGE.

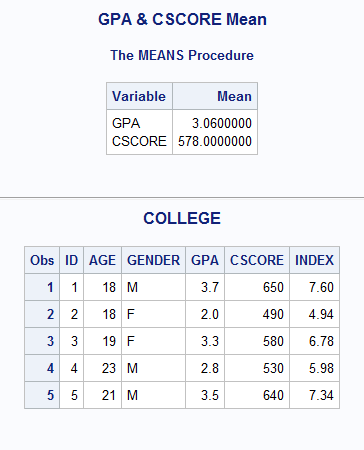
NOTE: PROCEDURE PRINT used (Total process time):

real time 0.09 seconds

cpu time 0.04 seconds

29 Title;

Output:



2 Add the necessary statements to compute the number of males and females in the previous problem.

Code:

Title 'Frequency Table for Gender';

**proc** **freq** data=COLLEGE;

table Gender;

**run**;

Title;

Log file:

328 Title 'Frequency Table for Gender';

329 proc freq data=COLLEGE;

330 table Gender;

331 run;

NOTE: There were 5 observations read from the data set WORK.COLLEGE.

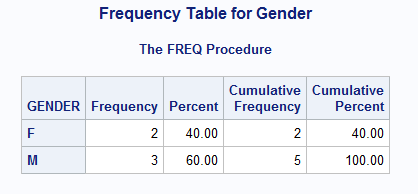
NOTE: PROCEDURE FREQ used (Total process time):

real time 0.14 seconds

cpu time 0.03 seconds

332 Title;

Output:



3 Use the data below and create a new variable (AGE\_GROUP) that has a value of 1 for ages between 0 and 35 and 2 for ages greater than 35.

**DATA** TAXPROB;

INPUT SS SALARY AGE RACE $;

FORMAT SS SSN11.;

DATALINES;

123874414 28000 35 W

646239182 29500 37 B

012437652 35100 40 W

018451357 26500 31 W

;

Compute the number of whites (W) and blacks (B) and the number in each age group. Use the appropriate option to omit cumulative statistics from the output.

Code:

**DATA** TAXPROB;

INPUT SS SALARY AGE RACE $;

FORMAT SS SSN11.;

DATALINES;

123874414 28000 35 W

646239182 29500 37 B

012437652 35100 40 W

018451357 26500 31 W

;

**run**;

**data** tax;

set TAXPROB;

if Age<=**35** then Age\_group=**1**;

else Age\_group=**2**;

**run**;

Title 'Frequency Table of Taxprob';

**proc** **freq** data=tax;

table Race\*Age\_group / norow nocol nopercent nocum;

**run**;

Title;

Log:

370

371 Title 'Frequency Table of Taxprob';

372

373 proc freq data=tax;

374 table Race\*Age\_group / norow nocol nopercent nocum;

375 run;

NOTE: There were 4 observations read from the data set WORK.TAX.

NOTE: PROCEDURE FREQ used (Total process time):

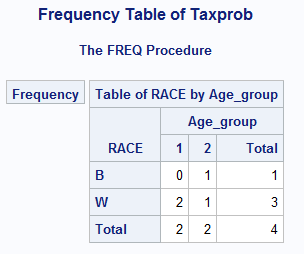
real time 0.20 seconds

cpu time 0.04 seconds

376

377 Title;

Output:



4 Use this data and PROC UNIVARIATE to produce histograms, normal probability plots, and boxplots and test the distributions for normality. Do this for variables like REACT, LIVER\_WT, and SPLEEN, first for all subjects and then separately for each of the two DOSES.

**DATA** LIVER;

INPUT SUBJ DOSE REACT LIVER\_WT SPLEEN;

DATALINES;

1 1 5.4 10.2 8.9

2 1 5.9 9.8 7.3

3 1 4.8 12.2 9.1

4 1 6.9 11.8 8.8

5 1 15.8 10.9 9.0

6 2 4.9 13.8 6.6

7 2 5.0 12.0 7.9

8 2 6.7 10.5 8.0

9 2 18.2 11.9 6.9

10 2 5.5 9.9 9.1

;

Code:

**DATA** LIVER;

INPUT SUBJ DOSE REACT LIVER\_WT SPLEEN;

DATALINES;

1 1 5.4 10.2 8.9

2 1 5.9 9.8 7.3

3 1 4.8 12.2 9.1

4 1 6.9 11.8 8.8

5 1 15.8 10.9 9.0

6 2 4.9 13.8 6.6

7 2 5.0 12.0 7.9

8 2 6.7 10.5 8.0

9 2 18.2 11.9 6.9

10 2 5.5 9.9 9.1

;

**run**;

**proc** **sort** data=LIVER

out=my\_liver;

by SUBJ;

**run**;

ods pdf file='c:\temp\sastest1.pdf' startpage=no notoc dpi=**300**;

\*options orientation=landscape;

\*ods graphics / reset=all height=4in width=7in;

ods trace on;

\*ods layout start height=7.25in width=10.5in;

\*ods region row=1 column=1;

Title 'Plots of Liver';

**proc** **univariate** data=LIVER normal plot;

var REACT LIVER\_WT SPLEEN;

Histogram REACT LIVER\_WT SPLEEN/ normal;

ods select Histogram plots;

**run**;

Title;

\*ods layout end;

Title 'Plots of liver according to Dose';

**proc** **univariate** data=LIVER normal plot;

Class Dose;

var REACT LIVER\_WT SPLEEN;

Histogram REACT LIVER\_WT SPLEEN/ normal;

ods select Histogram plots;

**run**;

Title;

ods trace off;

ods pdf close;

Log:

471 DATA LIVER;

472 INPUT SUBJ DOSE REACT LIVER\_WT SPLEEN;

473 DATALINES;

NOTE: The data set WORK.LIVER has 10 observations and 5 variables.

NOTE: DATA statement used (Total process time):

real time 0.20 seconds

cpu time 0.04 seconds

484 ;

485 run;

486

487 proc sort data=LIVER

488 out=my\_liver;

489 by SUBJ;

490 run;

NOTE: There were 10 observations read from the data set WORK.LIVER.

NOTE: The data set WORK.MY\_LIVER has 10 observations and 5 variables.

NOTE: PROCEDURE SORT used (Total process time):

real time 0.06 seconds

cpu time 0.07 seconds

491

492

493 ods pdf file='c:\temp\sastest1.pdf' startpage=no notoc dpi=300;

NOTE: Writing ODS PDF output to DISK destination "c:\temp\sastest1.pdf", printer "PDF".

494 \*options orientation=landscape;

495 \*ods graphics / reset=all height=4in width=7in;

496 ods trace on;

497 \*ods layout start height=7.25in width=10.5in;

498 \*ods region row=1 column=1;

499 Title 'Plots of Liver';

500 proc univariate data=LIVER normal plot;

501 var REACT LIVER\_WT SPLEEN;

502 Histogram REACT LIVER\_WT SPLEEN/ normal;

503 ods select Histogram plots;

504 run;

Output Added:

-------------

Name: Plots

Label: Plots for REACT

Template: base.univariate.Graphics.Plots

Path: Univariate.REACT.Plots

-------------

Output Added:

-------------

Name: Histogram

Label: Panel 1

Template: base.univariate.Graphics.Histogram

Path: Univariate.REACT.Histogram.Histogram

-------------

Output Added:

-------------

Name: Plots

Label: Plots for LIVER\_WT

Template: base.univariate.Graphics.Plots

Path: Univariate.LIVER\_WT.Plots

-------------

Output Added:

-------------

Name: Histogram

Label: Panel 1

Template: base.univariate.Graphics.Histogram

Path: Univariate.LIVER\_WT.Histogram.Histogram

-------------

Output Added:

-------------

Name: Plots

Label: Plots for SPLEEN

Template: base.univariate.Graphics.Plots

Path: Univariate.SPLEEN.Plots

-------------

Output Added:

-------------

Name: Histogram

Label: Panel 1

Template: base.univariate.Graphics.Histogram

Path: Univariate.SPLEEN.Histogram.Histogram

-------------

NOTE: PROCEDURE UNIVARIATE used (Total process time):

real time 5.97 seconds

cpu time 2.26 seconds

505 Title;

506 \*ods layout end;

507

508

509 Title 'Plots of liver according to Dose';

510 proc univariate data=LIVER normal plot;

511 Class Dose;

512 var REACT LIVER\_WT SPLEEN;

513 Histogram REACT LIVER\_WT SPLEEN/ normal;

514 ods select Histogram plots;

515 run;

Output Added:

-------------

Name: Plots

Label: Plots for REACT

Template: base.univariate.Graphics.Plots

Path: Univariate.REACT.'1'n.Plots

-------------

Output Added:

-------------

Name: Plots

Label: Plots for REACT

Template: base.univariate.Graphics.Plots

Path: Univariate.REACT.'2'n.Plots

-------------

Output Added:

-------------

Name: Histogram

Label: Panel 1

Template: base.univariate.Graphics.CompHistogram

Path: Univariate.REACT.Histogram.Histogram

-------------

Output Added:

-------------

Name: Plots

Label: Plots for LIVER\_WT

Template: base.univariate.Graphics.Plots

Path: Univariate.LIVER\_WT.'1'n.Plots

-------------

Output Added:

-------------

Name: Plots

Label: Plots for LIVER\_WT

Template: base.univariate.Graphics.Plots

Path: Univariate.LIVER\_WT.'2'n.Plots

-------------

Output Added:

-------------

Name: Histogram

Label: Panel 1

Template: base.univariate.Graphics.CompHistogram

Path: Univariate.LIVER\_WT.Histogram.Histogram

-------------

Output Added:

-------------

Name: Plots

Label: Plots for SPLEEN

Template: base.univariate.Graphics.Plots

Path: Univariate.SPLEEN.'1'n.Plots

-------------

Output Added:

-------------

Name: Plots

Label: Plots for SPLEEN

Template: base.univariate.Graphics.Plots

Path: Univariate.SPLEEN.'2'n.Plots

-------------

Output Added:

-------------

Name: Histogram

Label: Panel 1

Template: base.univariate.Graphics.CompHistogram

Path: Univariate.SPLEEN.Histogram.Histogram

-------------

NOTE: PROCEDURE UNIVARIATE used (Total process time):

real time 7.61 seconds

cpu time 2.45 seconds

516

517 Title;

518

519 ods trace off;

520 ods pdf close;

NOTE: ODS PDF printed 15 pages to c:\temp\sastest1.pdf.

Output:



|  |  |
| --- | --- |
| Variable | Normality |
| React | Skewness is not near zero also the mean and median are not approximately equal. So it does not have a normal distribution. |
| Liver\_wt | Skewness is near zero also the mean and median are approximately equal. So it has a normal distribution. |
| Spleen | Skewness is near zero also the mean and median are approximately equal. So it has a normal distribution. |

|  |  |
| --- | --- |
| Variable | Normality |
| React | According to Dose 1 Skewness is not near zero also the mean and median are not approximately equal. So it does not have a normal distribution.  According to Dose 2 Skewness is not near zero also the mean and median are not approximately equal. So it does not have a normal distribution. |
| Liver\_wt | According to Dose 1 Skewness is near zero also the mean and median are approximately equal. So it has a normal distribution.  According to Dose 2 Skewness is near zero also the mean and median are approximately equal. So it has a normal distribution. |
| Spleen | According to Dose 1 Skewness is not near zero also the mean and median are not approximately equal. So it does not have a normal distribution.  According to Dose 2 Skewness is near zero also the mean and median are approximately equal. So it has a normal distribution |

5 Refer to DOGS1 data. The investigator may want to ensure that the dogs allocated to each treatment group were of similar compositions with respect to gender and hair coat. Use PROC FREQ to conduct Fisher’s exact test to see if the concentration of the drug received was statistically independent of the gender of the dog. Likewise, see if the length of the coat and the drug treatment were statistically independent with Fisher’s exact test. Write your interpretation of the results of these tests.

Code:

libname college 'C:\Users\Samil\Desktop\Sem 1\Stats for programming\SAS 9.4';

/\*proc print data=college.dogs;

run;\*/

Title 'Fishers Exact Test for Drug & Gender';

**proc** **freq** data=college.dogs;

tables conc\*sex / exact;

**run**;

Title;

Title 'Fishers Exact Test for Drug & Haircoat';

**proc** **freq** data=college.dogs;

tables conc\*haircoat / exact;

**run**;

Title;

Log:

378 libname college 'C:\Users\Samil\Desktop\Sem 1\Stats for programming\SAS 9.4';

NOTE: Libref COLLEGE was successfully assigned as follows:

Engine: V9

Physical Name: C:\Users\Samil\Desktop\Sem 1\Stats for programming\SAS 9.4

379 /\*proc print data=college.dogs;

380 run;\*/

381

382 Title 'Fishers Exact Test for Drug & Gender';

383 proc freq data=college.dogs;

384 tables conc\*sex / exact;

385 run;

NOTE: There were 25 observations read from the data set COLLEGE.DOGS.

NOTE: PROCEDURE FREQ used (Total process time):

real time 0.21 seconds

cpu time 0.07 seconds

386 Title;

387

388 Title 'Fishers Exact Test for Drug & Haircoat';

389 proc freq data=college.dogs;

390 tables conc\*haircoat / exact;

391 run;

NOTE: There were 25 observations read from the data set COLLEGE.DOGS.

NOTE: PROCEDURE FREQ used (Total process time):

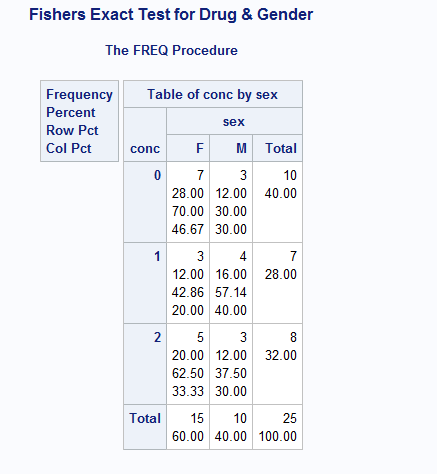
real time 0.15 seconds

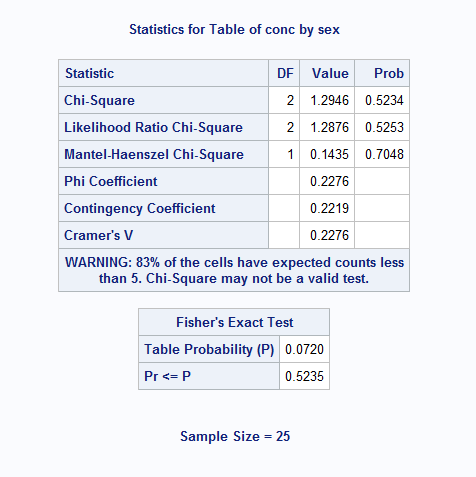
cpu time 0.09 seconds

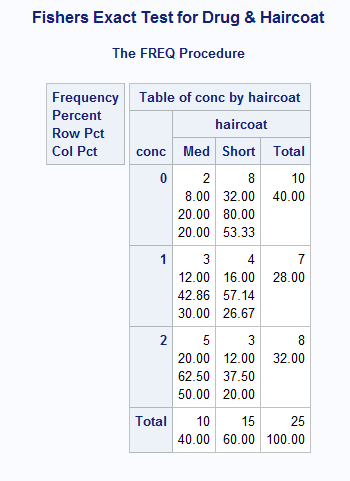
392

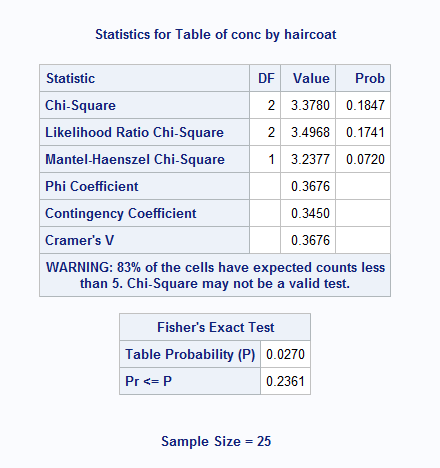
393 Title;

Output:









According to Fisher’s exact test for Drug & Gender p-value is 0.5235. Assuming a=0.05 and taking CI 95% we can say that drug concentration is dependent on gender of the dog.

According to Fisher’s exact test for Drug & Coat p-value is 0.2361. Assuming a=0.05 and taking CI 95% we can say that drug concentration is dependent on haircoat of the dog.

6 Using SAS data set Blood, generate one-way frequencies for the variables Gender, BloodType, and AgeGroup. Use the appropriate options to omit the cumulative statistics and percentages.

Code:

Title 'Frequency Table of Gender Blood\_type Age\_Group';

**proc** **freq** data=college.blood;

tables Gender Blood\_Type Age\_Group /nocum nopercent;

**run**;

Title;

Log:

394 Title 'Frequency Table of Gender Blood\_type Age\_Group';

395

396 proc freq data=college.blood;

397 tables Gender Blood\_Type Age\_Group /nocum nopercent;

398 run;

NOTE: There were 1000 observations read from the data set COLLEGE.BLOOD.

NOTE: PROCEDURE FREQ used (Total process time):

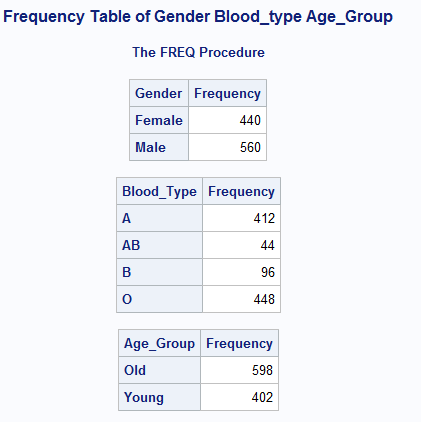
real time 0.11 seconds

cpu time 0.03 seconds

399

400 Title;

Output:



7 Using data set Blood, produce frequencies for the variable Chol (cholesterol). Use a format to group the frequencies into three groups: low to 200 (normal), 201 and higher (high) and missing. Run PROC FREQ twice, once using the MISSING option and once without. Compare the percentages in both listings.

Code:

**proc** **format**;

value chol low-**200**="LOW"

**201**-High="HIGH"

**.**="Missing";

**run**;

Title 'Frequency Table of Cholesterol without missing';

**proc** **freq** data=college.blood;

table Cholesterol;

format Cholesterol chol.;

**run**;

Title;

Title 'Frequency Table of Cholesterol with missing';

**proc** **freq** data=college.blood;

table Cholesterol / missing;

format Cholesterol chol.;

**run**;

Title;

Log:

435 proc format;

436 value chol low-200="LOW"

437 201-High="HIGH"

438 .="Missing";

NOTE: Format CHOL is already on the library WORK.FORMATS.

NOTE: Format CHOL has been output.

439 run;

NOTE: PROCEDURE FORMAT used (Total process time):

real time 0.00 seconds

cpu time 0.00 seconds

440

441 Title 'Frequency Table of Cholesterol without missing';

442 proc freq data=college.blood;

443 table Cholesterol;

444 format Cholesterol chol.;

445 run;

NOTE: There were 1000 observations read from the data set COLLEGE.BLOOD.

NOTE: PROCEDURE FREQ used (Total process time):

real time 0.22 seconds

cpu time 0.07 seconds

446 Title;

447

448 Title 'Frequency Table of Cholesterol with missing';

449 proc freq data=college.blood;

450 table Cholesterol / missing;

451 format Cholesterol chol.;

452 run;

NOTE: There were 1000 observations read from the data set COLLEGE.BLOOD.

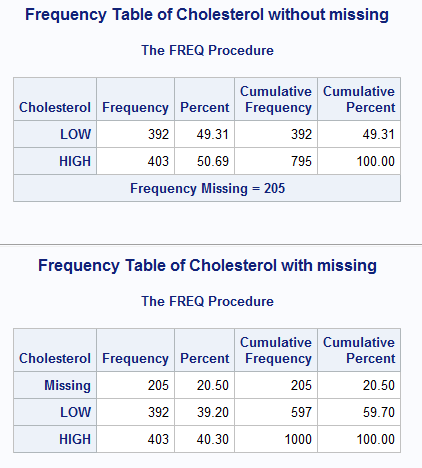
NOTE: PROCEDURE FREQ used (Total process time):

real time 0.14 seconds

cpu time 0.04 seconds

453 Title;

Output:



When we don’t use missing command in the proc freq than the missing values are not taken into consideration and its frequency, percent , cumulative frequency and cumulative percent are not calculated.

When we use missing command in the proc freq than the missing values are taken into consideration and its frequency, percent , cumulative frequency and cumulative percent are calculated.

8 What’s wrong with this program?

[You have to specify an error (if there is one) in each line (line numbers are given for your convenience). If you don't see any error, you can write "No Error". You are also welcome to write how you could improve the code, if you feel so.]

Your answer will be in the format,

Line 1 : <Your reason/observation>

Line 2:...

...

Line 11: ..

**1** DATA **123**;

**2** INPUT AGE STATUS PROGNOSIS DOCTOR GENDER STATUS2

**3** STATUS3;

**4** (data lines)

;

**5** PROC CHART DATA=**123** BY GENDER;

**6** VBAR STATUS

**7** VBAR PROGNOSIS;

**8** RUN;

**9** PROC PLOT DATA=**123**;

**10** DOCTOR BY PROGNOSIS;

**11** RUN;

Line 1 We cannot use a name of a data set starting with numeric. It should start with underscore or

characters.

Line 2 In the input statement it will consider all inputs to be numeric. Data type should be

mentioned like Gender should be a character data type.

Line 4 Datalines should be written together as it a key word. There is no need of brackets and no

dataset to look for values.

Line 5 We cannot use a name of a data set starting with numeric. It should start with underscore or

characters. We will also have to sort the data set first than we can group by a variable. After

writing the name of the data set put a semi colon. Write the By statement in next line.

Line 6 Semi colon is missing.

Line 7 No need to write it separately. We can combine line 6 & 7.

Line 8 No error.

Line 9 The name of the data set is wrong. It should start with underscore or

characters.

Line 10 It should be plot Doctor\*Prognosis.

Line 11 No error.

9 Using SAS data set COLLEGE, create two-way table of Scholarship (rows) by ClassRank (columns). Use a user-defined format to group class rank into two groups: 70 and lower, and 71 and higher.

Code:

libname college 'C:\Users\Samil\Desktop\Sem 1\Stats for programming\SAS 9.4';

\*options fmtsearch=(college);

/\*proc print data=college.college;

run;\*/

**proc** **format**;

value Rank low-**70**="LOW"

**71**-High="HIGH"

**.**="MISSING";

**run**;

Title 'Frequency table for Scholarship by ClassRank';

**proc** **freq** data=college.college;

tables Scholarship\*ClassRank / missing;

format ClassRank Rank.;

**run**;

Title;

Log:

459 proc format;

460 value Rank low-70="LOW"

461 71-High="HIGH"

462 .="MISSING";

NOTE: Format RANK is already on the library WORK.FORMATS.

NOTE: Format RANK has been output.

463 run;

NOTE: PROCEDURE FORMAT used (Total process time):

real time 0.00 seconds

cpu time 0.00 seconds

464

465 Title 'Frequency table for Scholarship by ClassRank';

466 proc freq data=college.college;

467 tables Scholarship\*ClassRank / missing;

468 format ClassRank Rank.;

469 run;

NOTE: There were 100 observations read from the data set COLLEGE.COLLEGE.

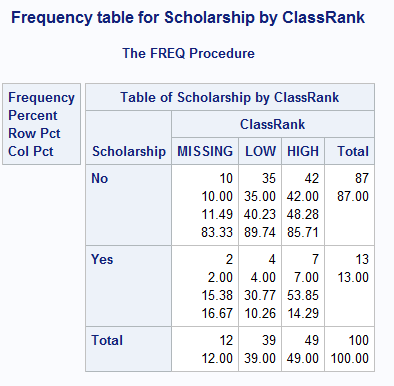
NOTE: PROCEDURE FREQ used (Total process time):

real time 0.16 seconds

cpu time 0.03 seconds

470 Title;

Output:



10 Using SAS data set blood, produce a table of frequencies for BloodType, in frequency order.

Code:

libname college 'C:\Users\Samil\Desktop\Sem 1\Stats for programming\SAS 9.4';

Title 'Frequency Table of Blood\_type';

**proc** **freq** data=college.blood order=freq;

table blood\_type;

**run**;

Title;

Log:

454 Title 'Frequency Table of Blood\_type';

455 proc freq data=college.blood order=freq;

456 table blood\_type;

457 run;

NOTE: There were 1000 observations read from the data set COLLEGE.BLOOD.

NOTE: PROCEDURE FREQ used (Total process time):

real time 0.13 seconds

cpu time 0.03 seconds

458 Title;

Output:

