

1. Consider the following data set.

ID	AGE	SEX	GPA	EXAM
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
6	28	f	2.8	400
7	25	m	1.4	150
8	18	f	2.5	520
9	19	m	2.6	510
10	17	m	3.9	670
11	19	m	2.8	513
12	18	f	3.7	672

- (a) Write a SAS program to compute

$$X := \text{GPA} + 3 \times \text{EXAM}/500$$

and store X together with the other variables in a SAS data set called **GRADES1**.

- (b) When the program in (a) is functioning correctly, add a second data step to form a data set **GRADES2** with the same variables, but with the observations corresponding to the women deleted:

```
data GRADES2;
  set GRADES1;
  if SEX = 'M';
run;
```

- (c) Add SAS instructions that compute the mean of X for the men only and print the corresponding observations in order of increasing X .
- (d) Standardize X to a variable with mean 0 and variance 1, and store the result as X in a data set **GRADES3**:

```
proc standard data=GRADES2 mean=0 std=1 out=GRADES3;
  var X;
run;
```

- (e) Verify the standardization:

```
proc means data=GRADES3;
  var X;
run;
```

Confirm that the standardized variable has mean 0 and variance 1.

2. Consider the population $S = \{1, 2, 3, 4\}$. Sampling of size 3 with replacement has sample space

$$S \times S \times S = \{(1, 1, 1), (1, 1, 2), (1, 1, 3), (1, 1, 4), (2, 1, 1), \dots, (4, 4, 4)\}.$$

Suppose the political preferences are $1, 3 \in \text{PIP}$, $2 \in \text{PNP}$, $4 \in \text{PPD}$, so that the population proportion voting PIP is 0.5.

Write a SAS program that calculates the proportion in each sample voting PIP, then calculates the mean of those sample proportions. Decide whether the sample proportion is an unbiased estimator of the population proportion. Write a short report explaining your findings. Your data set should resemble

OBS	A	B	C	PIP	PROPORTION
1	1	1	1	3	1.00000
2	1	1	2	2	0.66667
3	1	1	3	3	1.00000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
64	4	4	4	0	0.00000

3. Use SAS to generate a sample of size 100 from each of the following distributions. Use your ID as the seed in the random number generator. In each case draw a histogram with horizontal bars.

(a) $B(N, p)$, $N = 20$, $p = 0.1$

(b) $B(N, p)$, $N = 50$, $p = 0.1$

(c) $B(N, p)$, $N = 10$, $p = 0.5$

(d) $B(N, p)$, $N = 50$, $p = 0.5$

4. Test the numbers generated in the previous question for normality using

```
proc univariate data=... normal plot;
    var ...;
run;
```

Discuss both the normal plot and the p -value in the normality test.

5. Use Monte Carlo integration to estimate the definite integral of $f: [0, 1] \rightarrow \mathbb{R}$ given by

$$f(t) = -4t(1 - t).$$

Hint: a randomly generated point (X, Y) in the unit square $[0, 1] \times [0, 1]$ lies below the graph of a nonnegative function with probability equal to the area under that graph. Use 10,000 points, but do *not* print them.

6. Repeat the previous question for

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in [-2, 2].$$

How does your answer compare with the values obtained from the table at the rear of your textbook?

7. Write a SAS program that generates 250 numbers from a standard normal population and counts the percentage falling in each of the following intervals:

- (a) $-3 < X < 3$
- (b) $-2 < X < 2$
- (c) $-1 < X < 1$
- (d) $-\infty < X < 0$

Write a second data step that calculates the theoretically exact percentages using

$$\text{PROBNORM}(a) = \Pr(X < a),$$

where X has a standard normal distribution.

8. The CMS file `STATES DATA A` contains the area (in 1000 square miles) and the population (in millions) of the 50 U.S. states, arranged alphabetically. Write a SAS program that reads this file, selects all states with populations between 2 and 4 million, and for these states calculates the coefficient of correlation between area and population. Include a graph of area versus population for the selected states.

The first line of your program should contain the CMS instruction

```
CMS FILEDEF ESTADOS DISK STATES DATA A (RECFM F LRECL 80 BLKSIZE 80);
```

and the usual `input` statement should be preceded by

```
infile ESTADOS;
```

9. Write a SAS program to find, by simulation, the probability that in a set of 23 people at least two have the same birthday.

Remarks.

- A birthday may be represented by a random integer in $\{1, 2, \dots, 365\}$ (ignore leap years).
- Use a random vector X to store a set of 23 birthdays.
- Generate $N = 100$ such vectors and tally the frequency f of the event “at least two persons have the same birthday.”
- The experimental probability of an event is the relative frequency f/N .

10. The SAS manual claims that `RANUNI` generates numbers uniformly distributed on $[0, 1)$. Test this claim by generating 1000 such numbers and counting the frequencies with which they fall in each subinterval

$$\frac{k}{10} \leq X < \frac{k+1}{10}, \quad k = 0, 1, \dots, 9.$$

Run a χ^2 test on the results.

11. Use a χ^2 analysis to test the null hypothesis that numbers generated by iterating the logistic map

$$f(x) = 4x(1 - x)$$

are distributed uniformly on $[0, 1]$.

Remarks.

- Generate 1000 numbers starting from $X_0 = 0.25$.
 - Obtain X_n from X_{n-1} by $X_n = f(X_{n-1})$ for $1 \leq n \leq 1000$.
 - Count the relative frequencies in the subintervals $k/10 \leq X < (k+1)/10$, $k = 0, 1, \dots, 9$.
12. Write a SAS program to generate 100 samples of size 15 from a normal distribution with mean 10 and standard deviation 2. The program should calculate a 95% confidence interval for each sample and count how many of those intervals actually contain the population mean 10.
13. The instruction

```
X = ranexp(1);
```

yields a random number from an exponential distribution with parameter 1. Write a SAS program that generates 300 values from this distribution and constructs a horizontal bar chart.

The following program generates samples from a triangular distribution with vertices $(0, 0)$, $(1, 0)$, and $(h, 2)$ where $h = 0.1$. The `output out=` statement stores the sample means in a data set `stats` as the variable `mu`.

```
data clt;
  do sample = 1 to 5;
    do i = 1 to 5;
      X = rantri(801828, 0.1);
      output;
    end;
  end;
run;

proc means noprint;
  var X;
  by sample;
  output out=stats mean=mu;
run;

proc chart data=stats;
  hbar mu;
run;
```

Modify this program to generate 100 samples of size 2 from the triangular distribution and produce a horizontal bar chart of the distribution of sample means. Repeat with sample sizes 5, 10, 20, and 30. For the samples of size 30, add instructions to test the distribution of the mean for normality. Write a report explaining your observations.

14. The CMS file `corn data A` contains

year, preseason precipitation, May temperature, June rainfall, June temperature, July rainfall, July temperature, August rainfall, August temperature, corn yield (bushels/acre).

Run a simple linear regression to predict yield using

$$W = \frac{\text{July rain}}{\text{July temperature}}$$

as the independent variable. What is your prediction for mean yield when $W = 0.04$? What is the 95% confidence interval?

Your report should include:

- (a) scatter plots
- (b) correlation and R^2
- (c) regression coefficients
- (d) tests for significance
- (e) residual analysis

15. Write a SAS program that generates 1000 values of a standard normal random variable X and calculates the coefficient of linear correlation between X and $Y = X^2$. Is there evidence of linear correlation at the 0.05 level of significance? Are X and Y correlated? Generate the random numbers with `RANNOR(seed)`, using your student ID as the seed.

16. Consider again the population $\{1, 2, 3, 4\}$ and samples of size 3 drawn with replacement. Write a SAS program that inputs each sample (you may use `cards` or a loop) and calculates the variance of each sample. Use `proc means` to calculate the mean of the sample variances. Write a separate data step to calculate the population variance. Decide whether the sample variance S^2 is an unbiased estimator of the population variance. Your data set should resemble

OBS	A	B	C	XBAR	VAR
1	1	1	1	1.00000	0.00000
2	1	1	2	1.33333	0.33333
3	1	1	3	1.66667	1.33333
4	1	1	4	2.00000	3.00000
⋮	⋮	⋮	⋮	⋮	⋮
63	4	4	3	3.66667	0.33333
64	4	4	4	4.00000	0.00000

17. Write a SAS program to reproduce the table of the normal distribution at the back of your text. Use an array X_0 – X_9 to store the columns of the table. The SAS function `PROBNORM` can be used to generate the probabilities.

18. Write a SAS program to solve $Ax^2 + Bx + C = 0$ for values of A , B , and C input with `cards`. Handle the cases

- $A = 0, B = 0$ (no equation);
- $A = 0, B, C \neq 0$ (linear equation);
- $A \neq 0, B^2 \geq 4AC$ (two solutions, possibly equal);
- $A \neq 0, B^2 < 4AC$ (no real solution).

Your output should resemble

Obs	A	B	C	SOL.1	SOL.2
1	0	1	1	-1.0000	.
2	0	0	2	.	.
3	1	2	1	-1.0000	.
4	1	-1	-2	2.0000	-1.0000
5	6	-1	-2	-0.5000	0.6667
6	1	-1	-6	3.0000	-2.0000
7	1	-6	9	3.0000	3.0000

19. The CMS file `shrimp data A` contains data on annual shrimp yield in the port of San Juan:

year, preseason catch, preseason predator population, June salinity, June temperature, July salinity, July temperature, August salinity, August temperature, shrimp yield (tons).

Run a simple linear regression to predict yield using

$$W = \frac{\text{July salinity}}{\text{July temperature}}$$

as the independent variable. What is your prediction for mean shrimp yield when $W = 0.04$? What is the 95% confidence interval?

Your report should discuss:

- scatter plots
- correlation and R^2
- regression coefficients
- tests for significance
- residual analysis