BIST P8130: Biostatistics Methods I

Recitation 02 – Probability Distributions in SAS

Zilan Chai

Fall 2017

This recitation's big ideas:

- Use SAS function to obtain:
  1. Probability from a Binomial distribution
  2. Probability from a Poisson distribution
  3. Probability from a Normal distribution
  4. Compute the cumulative distribution functions
  5. Compute the probability density (mass) functions
- Use LIBNAME statement to create a permanent SAS datasets.
- Use Output Delivery System (ODS) to generate publication-quality output.

1. The PROBBNML Function

Syntax	Arguments		
PROBBNML(p, n ,x)	p	is a numeric probability of success parameter.	Range: $0 \le p \le 1$
	n	is an integer number of independent Bernoulli trials parameter.	Range: $n > 0$
	x	is an integer number of successes random variable.	Range: $0 \le x \le n$

In SAS, the function PROBBNML(p, n ,x) can be used to find the cumulative probability (CDF) of m or less successes in n trials when the probability of success is p. Assuming n=4 and p=0.25, the cumulative probability for $x \le 1$, also expressed as $Prob(X \le 1) = Prob(X=0) + Prob(X=1)$, can be found by using the SAS commands below.

```
DATA binomial_ex1;
PROB = PROBBNML(0.25,4,1);
RUN;

PROC PRINT DATA = binomial_ex1;
RUN;
```

Obs	PROB
1	0.73828

The probability for X=1, also expressed as $Prob(X=1) = Prob(X \leq 1) - Prob(X \leq 0)$, can be found using the SAS commands below.

```
DATA binomial_ex2;
 PROB = PROBBNML(0.25,4,1)-PROBBNML(0.25,4,0);
RUN;

PROC PRINT DATA = binomial_ex2;
RUN;
```

Obs	PROB
1	0.42188

Practice: Suppose our random variable X is binomial with n=20 and p=0.65

1). What is the probability of 13 or fewer successes?

2). What is the probability of 6 or fewer successes?

3). What is the probability of EXACTLY 13 successes?

4). What is the probability of more than 13 successes?

5). What is the probability of 13 or more successes?

6). What is the probability of 8, 9, or 10 successes?

2. The POISSON Function

Syntax	Arguments		
POISSON(λ ,x)	λ	is a numeric mean parameter.	Range: $\lambda \geq 0$
	x	is an integer random variable.	Range: $x \geq 0$

To calculate Poisson probabilities, we can use the SAS command POISSON(λ ,x) to find the cumulative probability (CDF) of x or less events. For example, suppose X~Poisson(λ), we can find the probability of X=1,2,3,4 using the following SAS commands:

```
DATA POISSON_EX
;
  *Prob(X=0);
  PROB = POISSON(1,0);
  *Prob(X=1);
  PROB1 = POISSON(1,1)-POISSON(1,0);
  *Prob(X=2);
  PROB2 = POISSON(1,2)-POISSON(1,1);
  *Prob(X=3);
  PROB3 = POISSON(1,3)-POISSON(1,2);
PROC PRINT;
RUN;
```

Obs	PROB	PROB1	PROB2	PROB3
1	0.36788	0.36788	0.18394	0.061313

Practice: Suppose the random variable X is distributed Poission with $\lambda = 12.33$.

1). What is the probability of 15 or fewer occurrences?
2). What is the probability of EXACTLY 6 successes?
3). What is the probability of more than 13 successes?
4). What is the probability of 13 or more successes?
5). What is the probability of 8, 9, or 10 successes?

3. The PROBNORM Function

Syntax	Arguments		
PROBNORM(*x*)	*x*	is a numeric random variable.	Range: the real numbers

The SAS function PROBNORM(*x*) returns the cumulative probability associated with the standard normal variable. For example if we would like to find the area under the standard normal distribution's probability density function less than or equal to 1.0, we could use the command PROBNORM(1.0).

```
DATA NORMAL_EX1;
  PROB = PROBNORM(1);
PROC PRINT;
RUN;
```

Obs	PROB
1	0.84134

Since PROBNORM can be used to find the CDF for a N(0,1) variable, we can also use PROBNORM to find the probability that a *N*(80,10) variable is greater than 95 or Prob(*X*>95). We shall find the probability that the variable is less than or equal to 95 and we can then subtract that probability from 1.

```
DATA NORMCDF;
   Z=(95-80)/10;
   P_LT95=PROBNORM(Z);
   P_GT95=1-P_LT95;
PROC PRINT;
RUN;
```

Obs	Z	P_LT95	P_GT95
1	1.5	0.93319	0.066807

Practice: Suppose our random variable Z is standard normal,

1). What is the probability Z is greater than 1.24?

2). What is the probability Z is less than 1.24?
3). What is the probability Z is between -0.79 and 1.16?
4). What if X is normal with mean 266 and standard deviation 16, what is the probability X is between 260 and 280?

4. The CDF Function

Syntax:	
CDF('BINOMIAL',*x*,*p*,*n*)	Returns a value from a (Binomial) cumulative probability distribution.
CDF('POISSON',*n*,*x*)	Returns a value from a (Poisson) cumulative probability distribution.
CDF('NORMAL', *x*, μ, σ)	Returns a value from a (Normal) cumulative probability distribution.

Details for other distributions (e.g. T-distribution, Chi-Square distribution) can be found at: http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000208980.htm

5. The PDF Function

Syntax:	
PDF('BERNOULLI',*x*,*p*)	Returns a value from a (Binomial) probability mass distribution.
PDF('POISSON',*n*,*x*)	Returns a value from a (Poisson) probability mass distribution.
PDF('NORMAL', *x*, μ, σ)	Returns a value from a (Normal) probability density distribution.

Example of comparing the CDF function and the PROBNORM function for a continuous r.v,:

```
data two;
input x @@;
cprob1=PROBNORM(x);
cprob2=CDF('NORMAL',x,0,1);
CARDS;
-3 -2 -1 0 1 2 3
;
run;
Proc print data=two label;
title "Normal probability distribution,
       mu=0, sigma=1";
title2 "The empirical rule";
label cprob1="P(X<=x)" cprob2="P(X<=x)";
run;
```

Obs	x	P(X<=x)	P(X<=x)
1	-3	0.00135	0.00135
2	-2	0.02275	0.02275
3	-1	0.15866	0.15866
4	0	0.50000	0.50000
5	1	0.84134	0.84134
6	2	0.97725	0.97725
7	3	0.99865	0.99865

Example of the CDF and PDF functions for a discrete r.v. :

```
DATA cdf_ex2;
INPUT x @@;
  cprob=CDF('BINOMIAL',x,.3,6);
  prob=PDF('BINOMIAL',x,.3,6);
CARDS;
1 2 3 4 5 6
;
RUN;
Proc print data=cdf_ex2 label;
title "Binomial probability distribution,
       n=6, p=.3";
label prob="P(X=x)" cprob="P(X<=x)";
run;
```

Obs	x	P(X<=x)	P(X=x)
1	1	0.42018	0.30253
2	2	0.74431	0.32413
3	3	0.92953	0.18522
4	4	0.98907	0.05953
5	5	0.99927	0.01021
6	6	1.00000	0.00073

6. The LIBNAME Statement – Permanent SAS Data Set

SAS stores data as "SAS data set" (.sas7bdat). There are 2 types of SAS data sets: temporary and permanent.

Temporary dataset exists only during the current job and is erased when exit SAS. It has an one-level name, and automatically stored in the WORK library

Permanent dataset remains when the job/session is finished. It has a two-level name, and never stored in WORK library

Syntax for generating a new SAS data library

```
LIBNAME libref 'Path\directory';
```

A *libref* is a name corresponding to the location of a SAS data library. Any *libref* (name of library) that you create exits only for the current SAS session. If the *libref* is WORK, then your data set will be temporary.

Suppose you want to create a permanent SAS data set called SBP in the O:\Teaching\Statistical.Computing folder:

```
libname rec01 'C:\Users\bioguest.SPH-6DSKSX1-BIO\Downloads';
  data rec01.sbp;
    input pat_name $ pat_id gender $ year1 year2 year3;
    cards;
       John  1002 M 90 120 125
       Alice 1003 F 140 148 116
       Mike  1004 F 121 130 117
       Barbara 1005 M 151 144 148;
  run;
```

7. Output Delivery System (ODS)

Each SAS procedure creates *output objects* that can be sent (separately) to destinations as:

- HTML (files can be used as Web pages; look pretty)
- RTF (easy to incorporate into a Word document)
- PDF (portable document format)
- SAS data sets
- CSV file

ODS is a great tool to format your PROC and DATA output, it maintains the format of the data variables that sometimes are not exported correctly (dates, times, etc.)

Example – convert SAS data set into a .csv file; output procedure results into .rtf/.pdf/.html files:

```
ODS CSV FILE='Path\myfile.csv';
  proc print data=test;
  run;
ODS CSV CLOSE;

ODS RTF FILE='pathname\filename.rtf';
  All procedures;
ODS RTF CLOSE;

ODS PDF FILE='pathname\filename.pdf';
  All procedures;
ODS PDF CLOSE;

ODS HTML FILE='pathname\filename.html';
  All procedures;
ODS HTML CLOSE;
```