Colored Candies and Base SAS®: A Sweet Way to Produce Chi-Square Control Charts Part 1: Tutorial

SherriJoyce King, King Information Company, Rockville, MD
Melvin T. Alexander, Westinghouse, Baltimore, MD

ABSTRACT

Colored candies are easy-to-understand tools that can illustrate Statistical Quality Control principles. Candy companies want each bag to have about the same piece counts of each color as the other bags. The objective of this tutorial is to show how Base SAS ® software can help to improve understanding of a manufacturing process where homogeneity of proportions is important, using candy color piece counts as input to SAS programs. Base SAS software will be used to produce a chi-square (χ^2) control chart that checks the stability of piece counts for candy colors during the bag-filling operation.

χ^2 is a statistical measure that compares actual values to specified values. χ^2 has three principal uses; we concentrate here on its ability to measure homogeneity of proportions. This tutorial (Part 1 of the topic) teaches the use of the SAS DATA step and the PLOT procedure (among others) for this purpose. Part 2, in the Statistics section, explores more statistical aspects of the topic. A simulation is included in each Part, where attendees count (and subsequently consume) candies to collect data on color proportions in bags they are given.

AUDIENCE

The tutorial assumes some knowledge of process control and a basic knowledge of SAS software. Concepts on which the tutorial depends will be reviewed.

OVERVIEW

The first section of the tutorial introduces basic concepts that the attendees need in order to follow the session:

- Stability of the bag-filling process; color piece counts as the indicator of process control.
- Control charts and control chart anatomy.
- χ^2.
- Analysis method.
- The data and what we need to do to prepare it and produce the χ^2 control chart.

The second section of the tutorial is a step-by-step presentation of the program used to prepare the data and generate the χ^2 control chart. The DATA step and the various procedures used (SORT, SUMMARY, TRANSPOSE, FREQ, PLOT) are explained. In overview, the program:

- Prepares and manipulates the data to put it in the format needed by the rest of the program.
- Calculates χ^2 for each bag of candies.
- Generates the χ^2 control chart.

At the end of the tutorial, attendees are active participants in a simulation of the sampling inspection portion of the bag-filling operation. The simulation has these steps:

- Data is collected and entered on the counts of colors in candy bags the attendees are given.
- The program is then run on the data collected.
- The output is reviewed, interpreted, and explained, showing how it can be used to adjust variations in color proportion.
- Finally, the data is consumed.

Only Base SAS is required for Part 1 of this presentation. For additional information on more advanced statistical techniques (SAS/QC) applicable to homogeneity of proportion, the attendees are invited to attend Part 2.

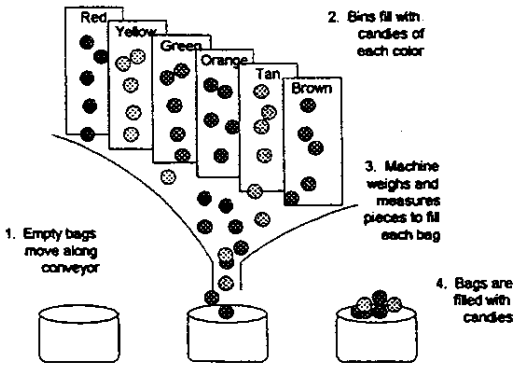
THE BASICS

Stability of the bag-filling process; color piece counts as the indicator of process control

The bag-filling process must be stable and the resulting proportions consistent. Candy companies go to a great deal of trouble to find out which proportions would be pleasing to the consumer. It doesn't make sense to be sloppy about the bag-filling process.

Figure 1-1

Bag Filling Operation



When we talk about process control, we usually talk about stabilizing a process so that the number and causes of defects (variations) are minimized. Color is analogous to a defect, in that it is an observable attribute or characteristic of the product and can easily be measured.

Control charts and control chart anatomy

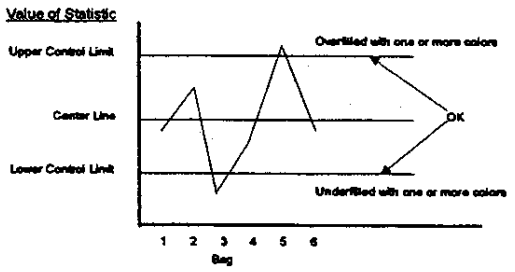
A control chart is a device that shows how stable, uniform, or consistent a specific aspect of a process is. A process is defined as a series of actions or operations performed for an identified purpose.

In 1924, Walter Shewhart introduced the first formal use of statistical control charts for process control. Part 2 of this paper includes graphical control charts.

Figure 1-2 is a control chart:

Figure 1-2

Control Chart



The Center Line (CL) is a specified average, the expected value of the statistic under study.

The Upper Control Limit (UCL) is a line drawn at a specified distance above the Center Line, representing the largest acceptable average measurement (based on sample size) that the process should produce when it is operating in a stable manner.

The Lower Control Limit (LCL) is a line drawn at a specified distance below the Center Line, representing the smallest acceptable average measurement (based on sample size) that the process should produce when it is operating in a stable manner.

The larger the sample size, the narrower the control limits can be, providing for a greater amount of precision.

By convention, the horizontal axis shows the temporal order of production (date, shift, or time). The vertical axis shows the values being measured or counted.

Interpreting the process control chart involves examining it for points outside the control limits. When the measurement goes outside control limits, it is probable that something unexpected is happening in production. Analysis will show whether the process is stable, whether it is operating within specification.

Chi-square (χ^2)

χ^2 is a test statistic or decision indicator serving three basic functions: goodness of fit; independence of two or more (crosstabulated) variables; homogeneity of proportions of counts to the whole. The last aspect is our focus.

χ^2 gives you a single measure to help you decide whether the difference between specified and observed is statistically larger or smaller than expected. More specifically, it indicates the degree to which the difference between expected (specified) and obtained counts is larger than can be explained by sampling variation alone.

χ^2 is a discrepancy statistic. The calculation is based on the discrepancy between expected and observed; the larger the discrepancy, the larger the χ^2. Its size depends on the size of the difference between expected and actual and on the number of differences involved.

An important piece of information you need to use χ^2 is degrees of freedom (called *df*). This is the amount of essential information you need without redundancy. Degrees of freedom is the number of categories under analysis minus one.

Analysis Method

Pearson is the classical analysis method. It suffers in validity when it is applied to data with small or zero call counts. It tends to produce inconsistent measures because it depends on bin sizes. To use Pearson's, all expected call frequencies should be more than 0 and fewer than 20% of them should be less than 5. Part 2 of this paper covers the more statistical aspects of this issue. There will be more about Pearson analysis in Part 2 of this paper. Also, Part 2 will discuss likelihood ratio and Neyman's statistics as ways to calculate χ^2.

The data and what we need to do to prepare it and produce the χ^2 control chart

Table 1-1 shows the data as it appears when we receive it. This is the kind of data collected during the bag-filling operation. Each observation represents one bag and contains a variable for each color. The value for each of those color variables is the piece count for that color for that bag.

Table 1-1

ORIGINAL data set
with color piece counts, by bag

OBS	BAGNUM	RED	YELLOW	GREEN	ORANGE	TAN	BROWN	BAGTOTAL
1	1	6	12	1	1	6	0	26
2	2	5	13	4	2	2	1	27
3	3	2	6	8	5	1	2	24
4	4	4	8	5	3	4	2	26
5	5	5	10	7	1	3	2	28
6	6	7	6	6	3	2	1	25
7	7	4	9	8	4	1	1	27
8	8	3	8	6	3	0	5	25
9	9	8	8	5	1	3	1	26
10	10	9	9	3	3	3	0	27
11	11	4	8	7	3	4	1	27
12	12	5	9	4	3	2	3	26
13	13	6	5	11	2	1	2	27
14	14	6	7	7	3	2	2	27
15	15	5	6	8	3	3	1	26
16	16	7	7	6	2	2	3	27
17	17	8	3	4	3	4	4	26
18	18	3	15	3	2	2	1	26
19	19	5	12	5	1	2	2	27
20	20	11	6	2	2	2	4	27

The program in this tutorial needs to do three things:

- Manipulate the data so that it is prepared to produce the control chart.

  When we collect the data, it comes to us one observation per bag, one variable per color, showing the piece count for that color. In order for us to calculate the bag χ^2, the data needs to be organized differently.

- Calculate the χ^2 for each bag.

  The χ^2 calculation requires the number of pieces for each color by bag (we are calling the intersection of a color and a bag a *cell*), the total pieces per bag, and the ratio each color represents of the total number of pieces for all bags. The calculation for cell χ^2 looks like this:

$$\text{cell } \chi^2 = \frac{(\text{cell piece count} - (\text{bagtotal} * \text{ratio}))^2}{(\text{bagtotal} * \text{ratio})}$$

  The bag χ^2 is simply the sum of the cell χ^2 for the bag.

- Produce the control chart.

  The control chart requires a specification for the Center Line (CL), the Upper Control Limit (UCL) and the Lower Control Limit (LCL).

THE PROGRAM

The first DATA step in the program prefixes the bag number with the string "BAG" so that it will merge properly later on. It also separates the BAGTOTAL variable into a different data set called BAGTOTAL.

```
data eachbag(drop=bagtotal)
     bagtotal(keep=bagid bagtotal);
  set original;
  length bagid $5;
  drop bagnum;
  if bagnum<10 then
    bagid="BAG0"||put(bagnum,$1.);
  else
    bagid="BAG"||put(bagnum,$2.);
run;
```

This next PROC TRANSPOSE reverses the variables and the observations in the data set. The result is the COLORS data set, where each observation represents a color and there are 20 variables in each observation showing how many pieces for that color in each bag.

```
proc transpose data=eachbag out=colors
  name=color prefix=bag;
run;
```

As in most procedures, the DATA= option names the input data set, while the OUT= option names the output data set.

For the TRANSPOSE procedure, the NAME= option gives a name to the variable in the output data set that stores the name of the variable(s) being transposed.

Table 1-2 shows that the COLOR variable in the output data set COLORS stores the names of the colors, which were variables in the input data set EACHBAG.

The PREFIX= option specifies the prefix to use in building the names of transposed variables. Table 1-2 shows variables named BAG1 through BAG20.

Without a VAR statement, seen in a PROC TRANSPOSE later in the program, all numeric variables will be transposed.

Table 1-2

PROC TRANSPOSE — COLORS data set

	C																			
	O		B	B	B	B	B	B	B	B	A	A	A	A	A	A	A	A	A	A
O	L		A	A	A	A	A	A	A	A	G	G	G	G	G	G	G	G	G	G
B	O		G	G	G	G	G	G	G	G	1	1	1	1	1	1	1	1	1	2
$	R		1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0
1	RED		6	5	2	4	5	7	4	3	8	9	4	5	6	6	5	7	8	3	5	11
2	YELLOW		12	13	6	8	10	6	9	8	8	9	8	9	5	7	6	7	3	15	12	6
3	GREEN		1	4	8	5	7	6	8	6	5	3	7	4	11	7	8	6	4	3	5	2
4	ORANGE		1	2	5	3	1	3	4	3	1	3	3	3	2	3	3	2	3	2	1	2
5	TAN		6	2	1	4	3	2	1	0	3	3	4	2	1	2	3	2	4	2	2	2
6	BROWN		0	1	2	2	2	1	1	5	1	0	1	3	2	2	1	3	4	1	2	4

The second PROC TRANSPOSE flattens the file so that there is one observation per bag/color cell. The data must be sorted first by COLOR so that PROC TRANSPOSE can use the BY COLOR statement.

```
proc sort data=colors;by color;run;

proc transpose data=colors out=cells
  name=bagid prefix=pieces;
  var bag1-bag20;
  by color;
run;
```

The VAR statement names the variables to transpose. The BY statement specifies that there will be one observation generated for each variable being transposed for each BY group.

The BAGID variable needs to be in the format "BAG01" - "BAG20" so that it will merge successfully later with the BAGTOTAL data set; this next DATA step renames the bag indicators.

```
data cells;
  set cells;
  if length(trim(bagid))=4 then
    bagid=substr(bagid,1,3) ||
    "0" || substr(bagid,4,1);
run;
```

Partial results are shown in Table 1-3.

Table 1-3

CELLS data set from PROC TRANSPOSE

OBS	COLOR	BAGID	PIECES1
1	BROWN	BAG01	0
2	BROWN	BAG02	1
3	BROWN	BAG03	2
4	BROWN	BAG04	2
5	BROWN	BAG05	2
6	BROWN	BAG06	1
7	BROWN	BAG07	1
8	BROWN	BAG08	5
9	BROWN	BAG09	1
10	BROWN	BAG10	0
11	BROWN	BAG11	1
12	BROWN	BAG12	3
13	BROWN	BAG13	2
14	BROWN	BAG14	2
15	BROWN	BAG15	1
16	BROWN	BAG16	3
17	BROWN	BAG17	4
18	BROWN	BAG18	1
19	BROWN	BAG19	2
20	BROWN	BAG20	4
21	GREEN	BAG01	1
22	GREEN	BAG02	4
23	GREEN	BAG03	8

This PROC SUMMARY calculates the color totals. These will be used in the divisor in the ratio needed for the χ^2 formula.

```
proc summary data=cells;
  class color;
  var pieces1;
  output out=cellsum sum= ;
```

Table 1-4

CELLSUM summary data set

OBS	COLOR	_TYPE_	_FREQ_	PIECES1
1		0	120	527
2	BROWN	1	20	38
3	GREEN	1	20	110
4	ORANGE	1	20	50
5	RED	1	20	113
6	TAN	1	20	49
7	YELLOW	1	20	167

This DATA step calculates the ratio each color represents of the total number of pieces.

```
data ratio (drop=_freq_ _type_);
  set cellsum
    (rename=(pieces1=colrpiec));
  retain bigtotal;
  if _type_=0 then do;
    bigtotal=colrpiec;
    delete;
  end;
  else ratio=colrpiec/bigtotal;
run;
```

Table 1-5

RATIO data set from DATA step

OBS	COLOR	COLRPIEC	BIGTOTAL	RATIO
1	BROWN	38	527	0.07211
2	GREEN	110	527	0.20873
3	ORANGE	50	527	0.09488
4	RED	113	527	0.21442
5	TAN	49	527	0.09298
6	YELLOW	167	527	0.31689

This next PROC FREQ computes cell χ^2 but won't put it into a data set YET -- a limitation of PROC FREQ that we hope will be corrected. Instead, what we'll do is use DATA step output to generate cell χ^2.

```
proc freq data=cells;
  tables color*bagid / cellchi2;
  weight pieces1;
run;
```

Table 1-6

PROC FREQ printed output

TABLE OF COLOR BY BAGID

COLOR(NAME OF FORMER VARIABLE)

BAGID(NAME OF FORMER VARIABLE)

Frequency / Cell Chi-Square / Percent / Row Pct / Col Pct	BAG01	BAG02	BAG03	BAG04	Total
BROWN	0	1	2	2	38
	1.8748	0.4605	0.042	0.0084	
	0.00	0.19	0.38	0.38	7.21
	0.00	2.63	5.26	5.26	
	0.00	3.70	8.33	7.69	
Total	26	27	24	26	527
	4.93	5.12	4.55	4.93	100.00

What we need to do instead is MERGE data sets we have already and compute χ^2 in a DATA step. This next DATA step MERGEs BAGTOTAL and CELLS by BAGID into a data set called INCLTOTL (to include the total). The two data sets must be sorted by the BAGID variable for the MERGE to be successful.

```
proc sort data=bagtotal;by bagid;run;
proc sort data=cells;by bagid;run;

data incltotl
     (rename=(pieces1=cellpiec));
  merge bagtotal cells; by bagid;
run;
```

Table 1-7

CELLS and BAGTOTAL data sets merged by bagid

OBS	BAGTOTAL	BAGID	COLOR	CELLPIEC
1	26	BAG01	BROWN	0
2	26	BAG01	GREEN	1
3	26	BAG01	ORANGE	1
4	26	BAG01	RED	6
5	26	BAG01	TAN	6
6	26	BAG01	YELLOW	12
7	27	BAG02	BROWN	1
8	27	BAG02	GREEN	4
9	27	BAG02	ORANGE	2
10	27	BAG02	RED	5
11	27	BAG02	TAN	2
12	27	BAG02	YELLOW	13
13	24	BAG03	BROWN	2
14	24	BAG03	GREEN	8

This next DATA step performs the calculation of the cell χ^2 after merging the INCLTOTL and RATIO data sets. The two data sets must both be sorted by COLOR for the MERGE to be successful.

```
proc sort data=ratio;by color;run;

proc sort data=incltotl;by color;run;

data cellchi2;
  merge incltotl ratio;
  by color;
  cellchsq =
    (cellpiec -(bagtotal*ratio))**2
    /(bagtotal*ratio);
run;
```

Table 1-8

CELLCHI2 data set

OBS	BAGTOTAL	BAGID	COLOR	CELLPIEC	INCLTOTL	RATIO	CELLCHSQ
1	26	BAG01	BROWN	0	38	527	0.07211	1.87476
2	27	BAG02	BROWN	1	38	527	0.07211	0.46051
3	24	BAG03	BROWN	2	38	527	0.07211	0.04195
4	26	BAG04	BROWN	2	38	527	0.07211	0.00837
5	28	BAG05	BROWN	2	38	527	0.07211	0.00018
6	25	BAG06	BROWN	1	38	527	0.07211	0.35739
7	27	BAG07	BROWN	1	38	527	0.07211	0.46051
8	25	BAG08	BROWN	5	38	527	0.07211	5.67108
9	26	BAG09	BROWN	1	38	527	0.07211	0.40816
10	27	BAG10	BROWN	0	38	527	0.07211	1.94687
11	27	BAG11	BROWN	1	38	527	0.07211	0.46051
12	26	BAG12	BROWN	3	38	527	0.07211	0.67537
13	27	BAG13	BROWN	2	38	527	0.07211	0.00145
14	27	BAG14	BROWN	2	38	527	0.07211	0.00145
15	26	BAG15	BROWN	1	38	527	0.07211	0.40816
16	27	BAG16	BROWN	3	38	527	0.07211	0.56968
17	26	BAG17	BROWN	4	38	527	0.07211	2.40918
18	26	BAG18	BROWN	1	38	527	0.07211	0.40816
19	27	BAG19	BROWN	2	38	527	0.07211	0.00145
20	27	BAG20	BROWN	4	38	527	0.07211	2.16519
21	26	BAG01	GREEN	1	110	527	0.20873	3.61121
22	27	BAG02	GREEN	4	110	527	0.20873	0.47473
23	24	BAG03	GREEN	8	110	527	0.20873	1.78525

To get the bag χ^2, we need to add up the cell χ^2 for all colors within each bag. We do this easily with the SUMMARY procedure.

```
proc summary data=cellchi2 nway;
  output out=bagchi2 sum=bagchi2;
  var cellchsq;
  class bagid;
run;
```

Table 1-9

BAGCHI2 data set -- PROC SUMMARY on cellchi2 by bagid

OBS	BAGID	_TYPE_	_FREQ_	BAGCHI2
1	BAG01	1	6	13.4164
2	BAG02	1	6	3.5781
3	BAG03	1	6	8.0253
4	BAG04	1	6	1.6451
5	BAG05	1	6	1.6331
6	BAG06	1	6	1.6539
7	BAG07	1	6	3.7449
8	BAG08	1	6	9.3192
9	BAG09	1	6	2.5161
10	BAG10	1	6	5.1536
11	BAG11	1	6	2.3388
12	BAG12	1	6	1.3675
13	BAG13	1	6	7.6250
14	BAG14	1	6	0.8012
15	BAG15	1	6	2.5515
16	BAG16	1	6	1.3563
17	BAG17	1	6	8.3219
18	BAG18	1	6	8.3912
19	BAG19	1	6	2.6229
20	BAG20	1	6	10.1909

The next DATA step calculates the upper and lower control limits and the center line using the CINV (chi-square inverse) function. This will prepare the control limits for the control chart.

- The upper control limit (UCL) is set to 97.5% (+2 sigma).
- The lower control limit (LCL)is set to 2.5% (-2 sigma).
- The center line (CL)is set to 50%.

UCL, LCL, and CL are all calculated with 5 degrees of freedom for the six colors (6 - 1 = 5).

(These variables have special names because of the way we use them in Part 2.)

```
data ctrllims;
  set bagchi2;
  drop _type_ _freq_;
  _ucle_ = cinv(.975,5);
     /* Lower control limit */
  _lcle_ = cinv(.025,5);
     /* Upper control limit */
  _mean_ = cinv(.5,5);
     /* The center line is the mean */
run;
```

Table 1-10

CTRLLIMS data set with control limits for plotting

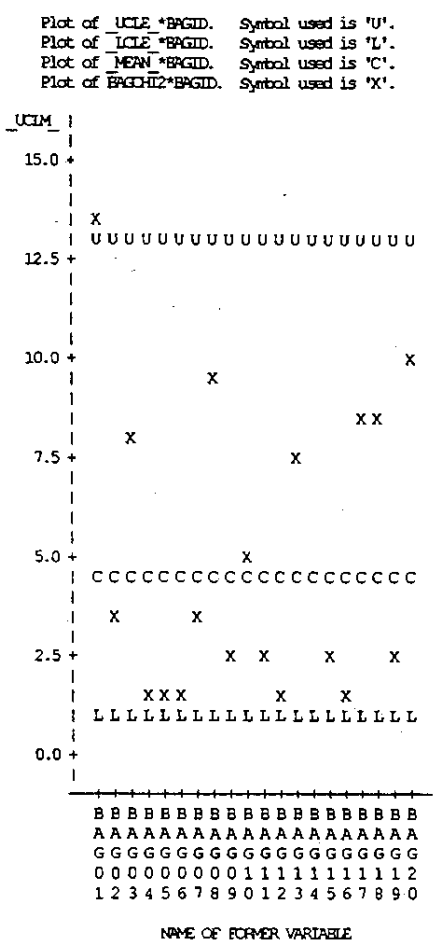
OBS	BAGID	BAGCHI2	_UCLE_	_LCLE_	_MEAN_
1	BAG01	13.4164	12.8325	0.83121	4.35146
2	BAG02	3.5781	12.8325	0.83121	4.35146
3	BAG03	8.0253	12.8325	0.83121	4.35146
4	BAG04	1.6451	12.8325	0.83121	4.35146
5	BAG05	1.6331	12.8325	0.83121	4.35146
6	BAG06	1.6539	12.8325	0.83121	4.35146
7	BAG07	3.7449	12.8325	0.83121	4.35146
8	BAG08	9.3192	12.8325	0.83121	4.35146
9	BAG09	2.5161	12.8325	0.83121	4.35146
10	BAG10	5.1536	12.8325	0.83121	4.35146
11	BAG11	2.3388	12.8325	0.83121	4.35146
12	BAG12	1.3675	12.8325	0.83121	4.35146
13	BAG13	7.6250	12.8325	0.83121	4.35146
14	BAG14	0.8012	12.8325	0.83121	4.35146
15	BAG15	2.5515	12.8325	0.83121	4.35146
16	BAG16	1.3563	12.8325	0.83121	4.35146
17	BAG17	8.3219	12.8325	0.83121	4.35146
18	BAG18	8.3912	12.8325	0.83121	4.35146
19	BAG19	2.6229	12.8325	0.83121	4.35146
20	BAG20	10.1909	12.8325	0.83121	4.35146

Now we plot the control chart. In Base SAS, this is done with the PLOT procedure. In the PLOT statement of that procedure, we request four plots, one for each control limit, one for the center line, and one for the bag chi square; then we present them all in one plot using the OVERLAY option of the PLOT statement. In Base SAS, we are operating without graphic capability, so we need to assign a meaningful text character to each plot so that we can tell them apart when they are overlayed.

```
proc plot data=ctrllims;
  plot _ucle_*bagid="U"
       _lcle_*bagid="L"
       _mean_*bagid="C"
       bagchi2*bagid="X" / overlay;
title 'Chi Square Control Chart';
quit;
run;
```

Table 1-11

Bag Chi Square Control Chart



NOTE: 1 obs hidden.

AUTHOR CONTACT

SherriJoyce King
King Information Company
16 Scotch Mist Court
Rockville, MD 20854-2929
Voice: (301) 251-6207
Fax: (301) 738-6873
Internet: 72764.164@compuserve.com

Mel Alexander
Integrated Product Development
Westinghouse Electronic Systems Group
7323 Aviation Boulevard
POBox 746, MS G16
Baltimore, MD 21203-0746
Voice: (410) 993-1478
Fax: (410) 765-1485
Internet: 71006.1534@compuserve.com

REFERENCES

Alexander, Melvin T. "Exposing Statistical Quality Control Ideas to Middle School Students."

Cullen, Francis J. "The Anatomy of a Control Chart."

Rowntree, Derek. Statistics Without Tears: A Primer for Non-Mathematicians. New York: Charles Scribner's Sons. 1981.

Schlotzhauer, David C. "Computing Goodness-of-fit Chi-square Statistics with SAS Software." *Observations*, Vol. 1, No. 1, pp. 17-24.

Williams, Frederick. Reasoning With Statistics: Simplified Examples in Communications Research. New York: Hold, Rinehart and Winston, Inc. 1968.