Paper 3561-2019

Thirteen Statistics Every Biostatistician Should Know

AnnMaria De Mars, The Julia Group, Santa Monica, CA

ABSTRACT

No one knows all of SAS® or all of statistics. There will always be some technique that you don't know. However, there are a few techniques that anyone in biostatistics should know. Make your life easier by learning to calculate those with SAS. In this session you will learn how to compute and interpret 12 of these techniques, including several statistics that are frequently confused. The following statistics are covered: prevalence, incidence, sensitivity, specificity, attributable fraction, population attributable fraction, risk difference, relative risk, odds ratio, Fisher's exact test, number needed to treat, and McNemar's test. The 13th, extra bonus tool, is SAS statistical graphics. With these 13 tools in their tool chest, even non-statisticians or statisticians who are not specialists will be able to answer many common questions in biostatistics. You're in luck because each of these can be computed with a few statements in SAS.

INTRODUCTION

No one can know all of SAS or all of statistics, no, not even that annoying guy down the hall. Especially not that annoying guy down the hall. However, certain statistics are so commonly used that every biostatistician should know them. All of these can be computed easily with SAS.

Biostatistics, broadly defined, is the application of statistics to topics in biology. However, most people when discussing biostatistics are really focused on biomedical topics, and not, say, the average lifespan of a particular species of ant-decapitating fly.
There are five ways (at least) statistics can be applied to the study of disease:

1. How common is it? This is a question of prevalence (how likely you are to have it) and incidence (how likely you are to get it). If you think those two are the same, you should take a course in epidemiology, or just finish reading this paper.
2. What causes it? What are the factors that increase (or decrease) your risk of contracting a disease?
3. What pattern(s) does it follow? What is the prognosis? Are you likely to die of it quickly, eventually or never? To determine if a treatment is effective for cancer of the eyelashes, we need to first have an idea of what the probability of disability or death is when one is left untreated and over how long of a period of time, that is, what is the "natural progression" of a disease
4. How effective are attempts to prevent or treat a disease?
5. Developing policies to minimize disease.

This paper covers basic statistics to address the first four questions. Policy should be developed based on the application of answers to those questions. While there is obviously a lot to be learned in the field of biostatistics, and a wide range of SAS procedures that can be applied, there are a basket of techniques that ought to be in everyone's hand, computable with SAS. These are prevalence, incidence, sensitivity, specificity, attributable fraction, population attributable fraction, risk difference, relative risk, odds ratio, Fisher's exact test, number needed to treat, and McNemar's test. That's only 12. What's the extra? Graphs. I admit, I cheated by including a whole category, but you'll find that ROC curves, survival curves, maps and odds ratio plots are invaluable in explaining results to a non-technical audience.

PREVALENCE AND INCIDENCE – TWO DIFFERENT ANSWERS TO THE SAME (MORE OR LESS) QUESTION

Policy makers have very good reasons for wanting to know how common a condition or disease is. It allows them to plan and budget for treatment facilities, supplies of medication, rehabilitation personnel. There are two broad answers to the question, "How common is condition X?" and, interestingly, both of these use the exact same SAS procedures.

HOW TO COMPUTE PREVALENCE USING SAS

Prevalence rate is the proportion of persons with a condition divided by the number in the population. It's often given as per thousand, or per 100,000, depending on how common the condition is. In brief, prevalence is how likely a person is to have condition X.

Assuming that your data are already cleaned and you have a variable with a binary coding for "has disease", "doesn't have disease", (pretty big assumptions) you can simply do a PROC FREQ.

```
PROC FREQ DATA = yourdatasetname ;
     TABLES variable ;
```

In the example below, from the California Health Interview Survey, approximately 11% of the respondents had been told they had Diabetes, giving a prevalence of 110 per 1,000.

Diabetes- ever				
AB22	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	4701	10.95	4701	10.95
2	38234	89.05	42935	100.00

Table 1. Output from PROC FREQ for Prevalence of Diabetes

HOW TO COMPUTE INCIDENCE USING SAS

INCIDENCE RATE is the rate at which new cases are occurring. Incidence is computed by dividing the number of new cases that occur in a specified period by the number of people in the population at risk.

The population at risk in the next example was defined as all infants born in 2014. If you are interested in birth statistics, the National Center for Health Statistics is highly recommended as a source. This example is from a public use data set of all 40,002 births in U.S. territories in the year 2014. (Winning trivial pursuit fact: The United States has 16 territories.)

Both incidence and prevalence estimates assume an accurate definition of cases, which requires understanding the data and the diagnosis. This data set required a very slight amount of coding because the Down syndrome variable at birth is coded as C for Confirmed, N for No and P for Pending (Center for Disease Control, 2014). “Pending” means that the medical personnel suspect Down syndrome but they are waiting for the results of a chromosomal analysis. “Confirmed” means the analysis has confirmed a diagnosis of Down syndrome. Based on the presumption that most experienced medical personnel recognize Down syndrome, these two categories were combined, using IF/THEN and ELSE statements.

```
LIBNAME  out “C:\Users\me\mydir\” ;
DATA  incidence ;
      SET out.birth2014 ;
      IF ca_down in (“C”,”P”) THEN down = “Y” ;
            ELSE down = “N” ;
```

Again, a PROC FREQ is used. The difference between incidence and prevalence is not in the computation but in the selection of the population and definition of the numerator. Because Down syndrome is present at birth and never acquired afterward, the new cases are going to be those children born in the year 2014 who have a diagnosis of Down syndrome and the denominator will be all births during the year.

```
PROC FREQ DATA =  out.birth2014 ;
            TABLES down;
```

Results are shown in Table 2 below.

Down Syndrome				
Down	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	39963	99.90	39963	99.90
Y	39	0.10	40002	100.00

Table 2. Output from PROC FREQ for Incidence of Down syndrome

The incidence rate is .10 or 1 per 1,000. Since this estimate falls perfectly in line with the World Health Organization (2016) estimate of between 1 in 1,000 to 1,100 live births, it appears that the case definition was appropriate.

SENSITIVITY AND SPECIFICITY – TWO ANSWERS TO A SECOND QUESTION: DO YOU HAVE A DISEASE?

Both sensitivity and specificity address the same question – how accurate is a test for disease – but from opposite perspectives. Sensitivity is defined as the proportion of those who have the disease that are correctly identified as positive. Specificity is the proportion of those who do not have the disease who are correctly identified as negative. Specificity and sensitivity can be computed simultaneously, as shown in the example below using a hypothetical Disease Test. The results are in and the following table has been obtained:

	Disease	No Disease
Test Positive	240	40
Test Negative	60	160

Table 3 Results from Hypothetical Screening Test

COMPUTING SENSITIVITY AND SPECIFICITY USING SAS

Step 1 (optional): Reading the data into SAS. If you already have the data in a SAS data set, this step is unnecessary. The example below demonstrates several SAS statements in reading data into a SAS dataset when only aggregate results are available.

The ATTRIB statement sets the length of the result variable to be 10, rather than accepting the SAS default of 8 characters.

The INPUT statement uses list input, with a $ signifying character variables.

DATALINES;
a statement on a line by itself, precedes the data. (Trivial pursuit fact #2: CARDS; will also work, dating back to the days when this statement was followed by cards with the data punched on them.) A semi-colon on a line by itself denotes the end of the data.

```
      DATA diseasetest;
            ATTRIB result LENGTH= $10;
            INPUT result $ disease $ weight;
            DATALINES;
            positive present 240
            positive absent 40
            negative present 60
            negative absent 160
      ;
```

Step 2: PROC FREQ

```
      PROC FREQ DATA= diseasetest ORDER=FREQ ;
            TABLES result* disease;
            WEIGHT weight;
```

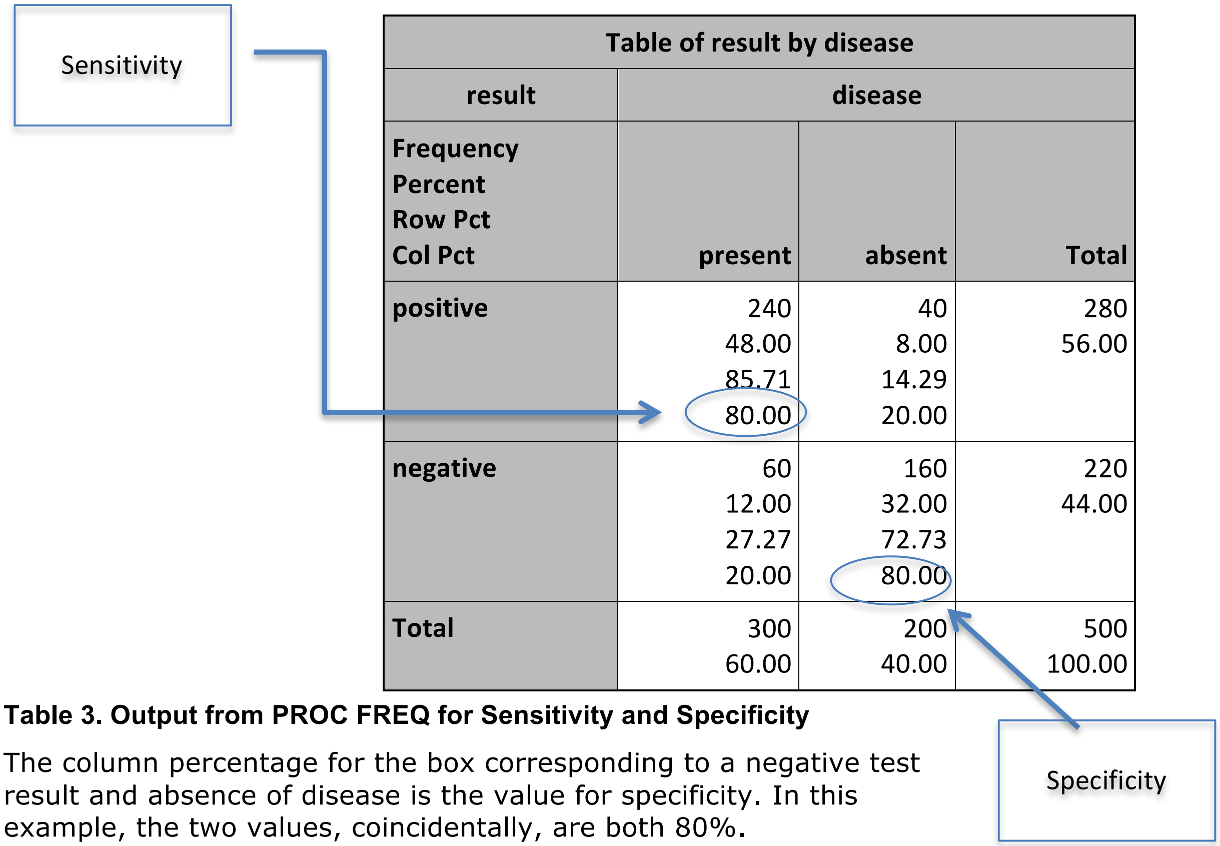
Yes, another PROC FREQ. The ORDER = FREQ option is not required but it makes the data more readable, in my opinion, because with these data the first cell will now be those who had a positive result and did, in fact, have the disease and this is the format in which sensitivity and specificity data are typically presented. The total for column 1 is the numerator for the formula for sensitivity, which is:

Sensitivity = (Number tested positive)/ (Total with disease).

TABLES variable1*variable2 will produce a cross-tabulation with variable1 as the row variable and variable2 as the column variable.

Weight weightvariable will weight each record by the value of the weight variable. The variable was named ‘weight’ in the example above but any valid SAS name is acceptable. Leaving off this statement will result in a table that only has 4 subjects, 1 subject for each combination of result and disease, corresponding to the data lines above.

Results of the PROC FREQ are shown below. The bottom value in each box is the column percent. Because the first category happens to be the "tested positive" and the first column is "disease present", the column percent for the first box in the cross-tabulation – positive test result, disease is present – is the sensitivity, 80%. This is the proportion of those who have the disease (the disease present column) who had a positive test result.



Sensitivity
Specificity

Table of result by disease			
result	disease		
Frequency Percent Row Pct Col Pct	present	absent	Total
positive	240 48.00 85.71 80.00	40 8.00 14.29 20.00	280 56.00
negative	60 12.00 27.27 20.00	160 32.00 72.73 80.00	220 44.00
Total	300 60.00	200 40.00	500 100.00

Table 3. Output from PROC FREQ for Sensitivity and Specificity

The column percentage for the box corresponding to a negative test result and absence of disease is the value for specificity. In this example, the two values, coincidentally, are both 80%.

Three points are worthy of emphasis here:

1. While the location of specificity and sensitivity in the table may vary based on how the data and PROC FREQ are coded, the values for sensitivity and specificity will always be diagonal to one another.
2. This exact table produces four additional values of interest in evaluating screening and diagnostic tests; positive predictive value, negative predictive value, false positive probability and false negative probability. Further details on each of these, along with how to compute the confidence intervals for each can be found in Usage Note 24170 (SAS Institute, 2015).
3. The same exact procedure produces six different statistics used in evaluating the usefulness of a test. Yes, that is pretty much the same as point number 2, but it bears repeating.

RELATIVE RISK, RISK DIFFERENCE AND ODDS RATIO – THREE ANSWERS TO A THIRD QUESTION: DOES THIS CAUSE A DISEASE?

Using the sashelp.heart data set, let's look at the difference in risk of cancer between smokers and non-smokers. If you'd like to try this at home, the code below creates a data set with variables for age group, smoking status and cancer death.

```
DATA attributable ;
 SET sashelp.heart ;
 IF ageatstart < 36 THEN agegroup = "28-35" ;
       ELSE IF ageatstart < 46 THEN agegroup = "36-45" ;
       ELSE IF ageatstart < 56 THEN agegroup = "46-55" ;
```

```
      ELSE IF ageatstart > 55 THEN agegroup = "56-62" ;
IF MISSING(smoking_status) = 0 AND (smoking_status) NE "Non-smoker"
                 THEN smoker = "Yes" ;
      ELSE IF MISSING(smoking_status) = 0 THEN smoker = "_No" ;
IF deathcause = "Cancer" THEN cancer = "Yes" ;
            ELSE cancer = "_No" ;
IF cancer = "" OR smoker = "" or agegroup = "" THEN DELETE ;
```

The risk difference is the difference in risk between the exposed and non-exposed group and is an option in PROC FREQ. The relative risk is the ratio of the risk of the exposed group and non-exposed group and can also be requested from (you guessed it) PROC FREQ. The code is simply:

```
PROC FREQ DATA = attributable ;
      TABLES smoker*cancer / RELRISK RISKDIFF ;
```

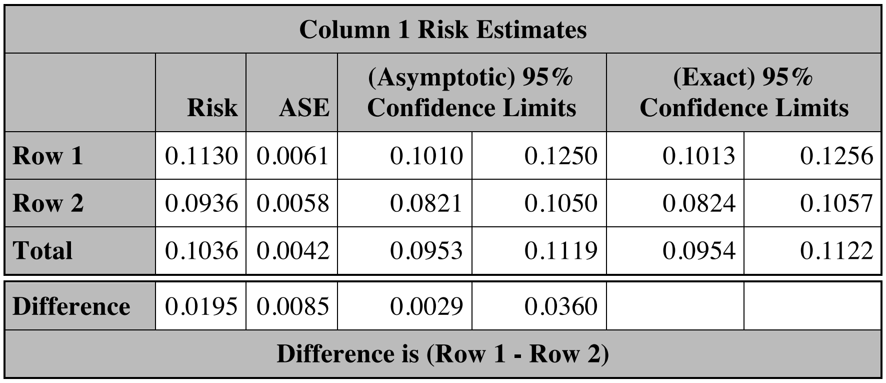
and produces the following output.

Table of smoker by cancer			
smoker	cancer		
Frequency Percent Row Pct Col Pct	Yes	_No	Total
Yes	302 5.84 11.30 56.34	2370 45.81 88.70 51.11	2672 51.65
_No	234 4.52 9.36 43.66	2267 43.82 90.64 48.89	2501 48.35
Total	536 10.36	4637 89.64	5173 100.00

Table 4. Output from PROC FREQ showing cross-tabulation

Notice that smoker and cancer were coded "_No" with an underscore to insure that the first column is prevalence in the exposed group (Yes for smoking and Yes for cancer). Note the highlighted numbers because these will look familiar in the next table.

As you can see from Table 5 below, the risk of the disease for row 1, the exposed group, of .1130 is simply the row percentage of the exposed group (in this case, smokers) with the disease. In the second row, the risk of the non-exposed is the percentage of non-smokers with the disease. The risk difference is subtracting the second row from the first. So, yes, I use the RISKDIFF option to avoid the need for subtraction (don't judge me!). The RISKDIFF option also gives 95% confidence intervals for the risk for each group and for the risk difference.



Column 1 Risk Estimates

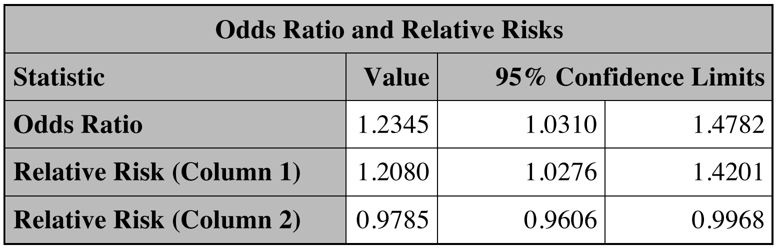
	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.1130	0.0061	0.1010	0.1250	0.1013	0.1256
Row 2	0.0936	0.0058	0.0821	0.1050	0.0824	0.1057
Total	0.1036	0.0042	0.0953	0.1119	0.0954	0.1122
Difference	0.0195	0.0085	0.0029	0.0360		

Difference is (Row 1 - Row 2)

Table 5. Output from PROC FREQ showing risks and risk difference

As you can also see above, that interval is somewhat wide, but does not include zero.

The third table produced by this analysis, not shown, is simply risk estimates for Column 2, in this case not having the disease and is the inverse of the risk difference of column 1. The fourth table, below, gives odds ratio and relative risks.



Odds Ratio and Relative Risks

Statistic	Value	95% Confidence Limits	
Odds Ratio	1.2345	1.0310	1.4782
Relative Risk (Column 1)	1.2080	1.0276	1.4201
Relative Risk (Column 2)	0.9785	0.9606	0.9968

Table 6. Output from PROC FREQ showing relative risk and odds ratio

The relative risk is the risk in the exposed divided by the risk in the non-exposed group. That is, the .1130 from Row 1 above divided by the .0936 from Row 2. Table 6 gives that value of 1.208 as well as 95% confidence limits for the relative risk.

When the RELRISK option is specified for PROC FREQ, as can be seen above, the odds ratio is quite close to the relative risk, with a value of 1.2345.

PROC FREQ is not the only SAS procedure that will compute odds ratios. For example:

```
PROC LOGISTIC  data= attributable ;
        MODEL cancer = smoker ;
```

will produce the identical odds ratio.

ODDS RATIO VS RELATIVE RISK

Think of a cross-tabulation of risk factor and disease like the table below

	DISEASE	
RISK FACTOR	YES	NO
YES	A	B
NO	C	D

Table 7. Output from PROC FREQ showing relative risk and odds ratio

Risk of exposed population = A/ (A + B)

Risk of non-exposed population = C/ (C + D)

The relative risk is, then

$$\frac{A / (A + B)}{C/(C+D)}$$

The odds ratio, in contrast, is, as the name implies, the ratio of the odds for the exposed group and the odds of the non-exposed group. The formula for odds ratio is

$$\frac{A/B}{C/D}$$

Why are these usually so close? Because most diseases, whether a person is exposed (A) or not exposed (C) have a pretty low risk relative to the total population. Let's say the risk of a disease is 1% in the exposed population and .5% in the unexposed. In a sample of 200 people in each group, exposed and unexposed, whether dividing 2/200 by 1/200 or 2/198 by 1/199, the result is going to be pretty similar.

NOTE: The above analysis is for unmatched pairs. SAS also can compute matched pairs odds ratios, see SAS Institute (2005).

ATTRIBUTABLE FRACTION AND POPULATION ATTRIBUTABLE FRACTION – TWO ANSWERS TO A FOURTH QUESTION: WHAT IS THE IMPACT ON PUBLIC HEALTH?

While relative risk and odds ratio are useful statistics for assessing the strength of a relationship, an important consideration in determining causality, two other statistics are equally or more important for public health issues (Gordis, 2014). The attributable fraction statistic answers the question, "What fraction of the disease cases in the exposed population is attributable to risk factor X?" The formula for computation is:

$$\frac{\text{Prevalence of exposed group} - \text{Prevalence of unexposed}}{\text{Prevalence of exposed group}}$$

If the prevalence of the exposed group is equal to the prevalence of the unexposed then the attributable fraction is zero. In other words, none of the disease cases can be attributable to the risk factor. Conversely, if the prevalence in the unexposed group is zero, then the attributable fraction is 1.0, i.e., 100% of the prevalence in the exposed group is due to the risk factor. Of course, the obtained fraction almost always falls between these two extremes.

The population attributable fraction statistic answers the question, "What fraction of the disease cases in the total population is attributable to risk factor X?" The computation formula is

$$\frac{\text{Prevalence overall} - \text{Prevalence of unexposed}}{\text{Prevalence overall}}$$

The attributable fraction can be high while the population attributable fraction is low, and the population attributable fraction should always be lower than the attributable fraction. Why is this? Because not everyone will be exposed to the risk factor.

SAS can be useful in computing both statistics on two fronts. First, directly, via the use of the STDRATE procedure, which computes both attributable fraction and the population attributable fraction. Second, indirectly, by creating a data set that can be used as input to the procedures, in the likely case that you don't have a data set lying around with cases in the exposed population, total count of the exposed population and cases in the non-exposed population.

COMPUTING ATTRIBUTABLE FRACTION AND POPULATION ATTRIBUTABLE FRACTION USING SAS: PROC STDRATE

The code for computing the example attributable fraction and population attributable fraction is shown below. First, I created a data set, and if those numbers look familiar, it is because they are from Table 4 above. The events (cancer) in the exposed (smoking) group = 302 and the total count in that group = 2,672. The events in the non-exposed group = 234 and the total count for the non-exposed group = 2,501.

```
      DATA std3 ;
            INPUT event_e count_e event_ne count_ne ;
            DATALINES ;
            302 2672 234 2501
            ;
```

Now that we have the data, let’s code the analysis. Surprisingly, it is not another PROC FREQ, but PROC STDRATE which computes standardized rates and risks of various types. The code below is an example of one of the simplest analyses.

```
      PROC STDRATE DATA=std3
                   REFDATA=std3
                   METHOD=INDIRECT(AF)
                   STAT=risk;
         POPULATION EVENT=event_e  TOTAL=count_e;
         REFERENCE  EVENT=event_ne TOTAL=count_ne;
```

This example uses the same data set for the exposed data and the reference data, but it should be noted that separate data sets can be specified. The method used for standardization is indirect. If you're interested in the different types of standardization, I highly recommend the 2013 SAS Global Forum paper by Yuan (2013).

The POPULATION and REFERENCE statements are required to compute attributable fractions. The POPULATION statement requires two parameters EVENT = followed by a variable with the number of cases in the exposed group and TOTAL = followed by the total number of subjects in the exposed population. The REFERENCE statement also requires two parameters, EVENT = followed by a variable with the number of cases in the non-exposed group and TOTAL = followed by the total number of subjects in the non-exposed population.

Relevant results from the PROC STDRATE are shown in Tables 8 and 9 below.

Indirectly Standardized Risk Estimates									
Study Population						Standardized Risk			
Observed Events	Number of Observations	Crude Risk	Reference Crude Risk	Expected Events	SMR	Estimate	Standard Error	95% Normal Confidence Limits	
302	2672	0.1130	0.0936	249.999	1.2080	0.1130	0.00613	0.1010	0.1250

Table 8: Partial Output from PROC STDRATE, Reference and Crude Risk

Note, once again, the risk of the study population is .1130 and of the reference population, .0936. The attributable fraction, as shown in Table 9 is .172 – in other words, 17.2% of the risk of cancer of smokers is attributable to smoking. If you recall from our first cross-tabulation, slightly over half of the population were smokers. Thus, as expected, the population attributable risk is slightly over half the attributable risk for the exposed population – 9.7% of the risk in the population is attributable to smoking.

Attributable Fraction Estimates			
Parameter	Estimate	95% Confidence Limits	
Attributable Risk	0.17219	0.07381	0.25167
Population Attributable Risk	0.09702	0.03446	0.15552

Table 9: Partial Output from PROC STDRATE, Attributable Risk and Population Attributable Risk

From a public health perspective, e.g., in examining potential benefits of an intervention, the distinction between attributable fraction and population attributable fraction is important. If a risk factor is uncommon, say, working in a coal mine, even if the attributable fraction is large, reducing the risk substantially may have little impact on the population risk.

STRATIFIED RISK ESTIMATES: CREATING A DATA SET FOR USE BY PROC STDRATE

If it seems as if that attributable risk is a little low, it might be because another analysis (not shown) found a substantial relationship between age group and smoking status in this sample. The younger subjects were much more likely to be smokers, but they were also less likely to die, because (Captain Obvious alert) older people are more likely to die. The solution, then, is to re-compute the attributable fraction and population attributable fraction stratified by age.

CREATING THE DATA SET OF FREQUENCIES STRATIFIED BY AGE

PROC STDRATE is great if you happen to have a dataset with the number of cases in the exposed group, total in the exposed group, number of cases in the non-exposed group and total in the non-exposed group, conveniently sorted by strata. Seriously, though, what is the probability you just have that lying around? You could do a PROC FREQ and then type in the data in a data step. Another alternative is to follow the steps below.

There are actually more statements than strictly necessary, but I like things to be neat and tidy.

```
PROC FREQ DATA=attributable ;
      TABLES agegroup*smoker*cancer /OUT=freqcount;
```

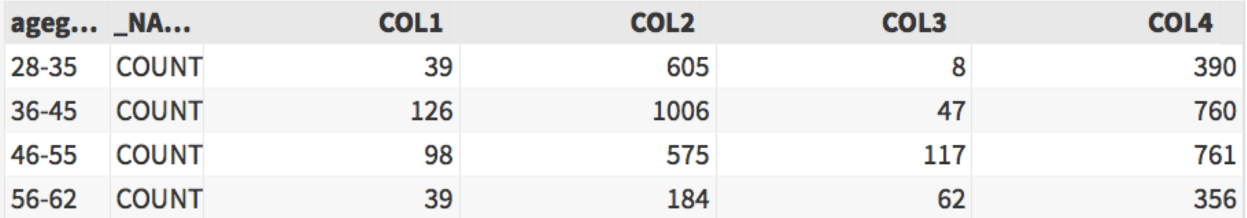
This creates an output dataset with counts and percentages of agegroup (strata) by smoking status (risk) by cancer death (event) and outputs it to a dataset named ‘freqcount’. The records will be in ascending order by value, the default for PROC FREQ.

```
DATA freqcount ;
      SET freqcount ;
      DROP PERCENT;
```

The step above simply drops percent as a variable in the data set, since we don’t need it.

```
PROC TRANSPOSE DATA=freqcount   OUT=transf  ;
      BY agegroup ;
```

The TRANSPOSE step above transposes the data set by age group so that each age group has four variables and outputs the results to a dataset named ‘transf’. The dataset created is shown below.



ageg...	_NA...	COL1	COL2	COL3	COL4
28-35	COUNT	39	605	8	390
36-45	COUNT	126	1006	47	760
46-55	COUNT	98	575	117	761
56-62	COUNT	39	184	62	356

Output Data 1: Data set created by TRANSPOSE procedure

The default name is col1 – col4. Since “Yes” comes before “_No” in SAS alphabetical order, the first two columns are for Smoker (Yes) when Cancer = “Yes” and when Cancer = “_No”.

```
DATA std4 ;
      SET transf ;
      count_e = col1+ col2 ;
      count_ne = col3 + col4 ;
      RENAME
            col1 = event_e
            col3 = event_ne ;
```

The final step above creates the count variables needed for PROC STDRATE. Since col1 and col2 are the number of smokers who did die of cancer (cancer = “Yes”) and the number of smokers who did not die of cancer (cancer = “_No”) the sum of these is the count of smokers, i.e., the count of exposed. Similarly, the sum of col3 and col4 is the count of non-exposed. The procedure has two additional statements from the PROC STDRATE above, one required and one optional.

```
PROC STDRATE DATA=std4

             REFDATA=std4

             METHOD=indirect(af)

             STAT=RISK

             PLOTS(STRATUM=HORIZONTAL);

   POPULATION EVENT=event_e  TOTAL=count_e;

   REFERENCE  EVENT=event_ne TOTAL=count_ne;

   STRATA agegroup / STATS;
```

STRATA statement stratifies by the designated variable, in this case, age group.

Indirectly Standardized Risk Estimates									
Study Population			Reference Crude Risk	Expected Events	SMR	Standardized Risk			
Observed Events	Number of Observations	Crude Risk				Estimate	Standard Error	95% Normal Confidence Limits	
302	2672	0.1130	0.0936	201.632	1.4978	0.1401	0.00755	0.1253	0.1549

Table 10: Partial Output from PROC STDRATE, Reference and Crude Risk with STRATA statement

A close comparison of Table 10 and Table 8 finds some similarities and some differences. The observed events, number of observations and crude risk in the study population are all the same, as is the reference crude risk for the reference population. However, the expected events and Standardized Mortality Rate (SMR) have changed. Given the stratification by age, if smokers and non-smokers were distributed equally across age groups, 202 deaths would have been expected, not the 302 observed in the study population.

BONUS STATISTIC: STANDARDIZED MORTALITY RATE

The Standardized Mortality Rate = Observed Number of Deaths Per Year

Expected Number of Deaths Per Year

Attributable Fraction Estimates			
Parameter	Estimate	95% Confidence Limits	
Attributable Risk	0.33235	0.25356	0.39609
Population Attributable Risk	0.18725	0.12531	0.24481

Table 11: Partial Output from PROC STDRATE, Attributable Risk and Population Attributable Risk with STRATA statement

Taking into account the stratification by age, that is, controlling for the fact that smokers in this sample were younger than non-smokers, the attributable risk estimate is now 33.2% and the population attributable risk 18.7%

The PLOTS statement produces plots of the crude estimate of the risk by strata, with the reference group risk as a single line. If you look at the graph below you can see several useful measures. First, the blue dots are the risk estimate for the exposed group at each age group and the vertical blue bars represent the 95% confidence limits for that risk. The red crosses are the risk for the reference group at each age group. The horizontal, solid blue line is the crude estimate for the study group, i.e., smokers, and the dashed, red line is the crude estimate of risk for the reference group, in this case, the non-smokers.

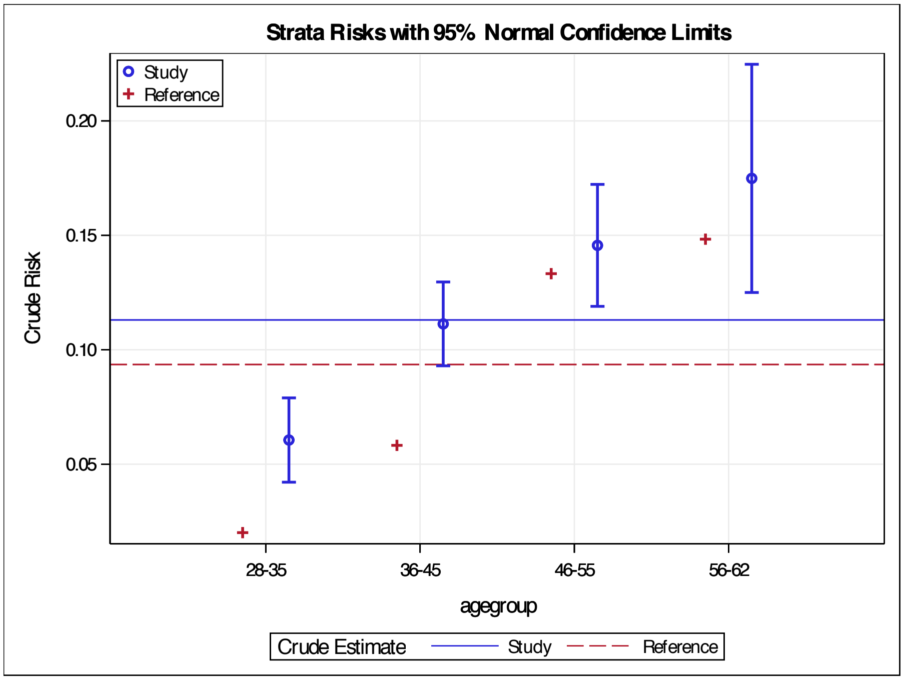


Figure 1: Graph from PROC STDRATE, PLOTS option, of Risks by Strata

Several observations can be made at a glance.

1. The crude risk for non-smokers is lower than for smokers.
2. As expected, the younger age groups are below the overall risk of mortality from cancer.
3. At every age group, the risk is lower for the non-exposed group.
4. The differences between exposed and non-exposed are significantly different for the two younger age groups only, for the other two groups, the non-smokers, although having a lower risk, do fall within the 95% confidence limits for the exposed group.

NUMBER NEEDED TO TREAT

Let's move away from risk factors to protective factors. While it is natural to feel that no expense is too much to save a life, the fact is that resources are finite and policy makers may need to evaluate relative benefits of treatments. How many people must be treated with a medication or other intervention to save a single life or prevent one adverse outcome? If there is a PROC NNT, I've never found it. Fortunately, computing the number needed to treat is simple. The formula is:

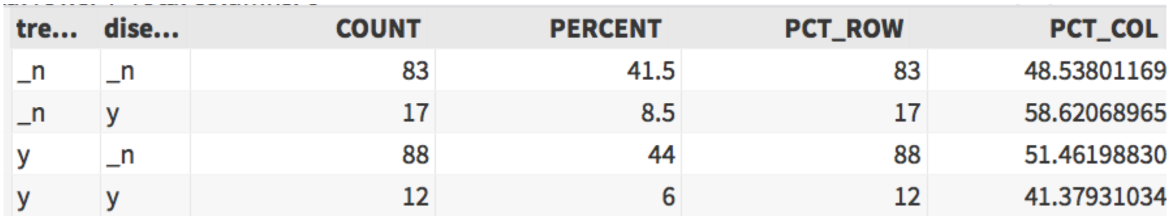
$$\frac{1}{\text{(Rate of untreated group – Rate of treated group)}}$$

One simple way to compute number needed to treat using SAS is to use a FREQ procedure to create an output data set is to compute the rate of the untreated group and the treated group in cross-tabulation and then use a calculator to plug the numbers for rate of untreated group and rate of treated group into the equation above, but what fun would that be?

The following code yields both the cross-tabulation and the number needed to treat nicely formatted. Tip: If you code the untreated as "_N" or "_No" and treated as "Y" or "Yes", the rate for untreated will always be COL2 and the rate for treated will be COL4. In the code below, the LABEL, VAR and ID statements are unnecessary, as is the SPLIT option in the PRINT statement. These all simply make the final result look nice.

The OUT= and OUTPCT options are required. OUT = creates a dataset and OUTPCT specifies that the dataset will include percentages

```
PROC FREQ DATA =nnt ;
      TABLES treated*disease /OUT =nnt2 OUTPCT ;
```



tre...	dise...	COUNT	PERCENT	PCT_ROW	PCT_COL
_n	_n	83	41.5	83	48.53801169
_n	y	17	8.5	17	58.62068965
y	_n	88	44	88	51.46198830
y	y	12	6	12	41.37931034

Output Data 2: Data set created by PROC FREQ with OUTPCT option

The following code transposes the data set, computes the number needed to treat in the DATA step and prints it nicely formatted in the PRINT step.

```
PROC TRANSPOSE DATA=nnt2 OUT=nnt3 ;
DATA ans ;
      SET nnt3 ;
      IF _name_ = "PCT_ROW" THEN
      numn = 1/(col2/(col1 + col2) - col4/(col3 + col4)) ;
LABEL numn = "Number Needed/To Treat" ;
PROC PRINT DATA= ans SPLIT="/" ;
      ID numn ;
      VAR ;
      WHERE numn NE . ;
```

Number Needed To Treat
20

Table 12: Number Needed to Treat from FREQ, TRANSPOSE, DATA and PRINT steps

FISHER'S EXACT TEST- FOR SMALL SAMPLE SIZES

These next two statistics apply to some special cases where the typical chi-square just won't do. You have collected data and have a simple design, a control group and treatment group. Your variable of interest, though, is low incidence and your SAS printout cautions, you

"WARNING: 50% of the cells have expected counts less
than 5. Chi-Square may not be a valid test."

What do you do now? Answer: Use Fisher's exact test.
How do you get it? If your analysis involves a 2 x 2 table, simply perform a chi-square analysis and look at the third table in your output. For 2 x 2 chi-square analyses, SAS automatically produces a Fisher exact test. In the example below, the code requests a chi-square analysis for very low birthweight – less than 1 pound – by marital status of the mother.

```
PROC FREQ DATA=vlow ;
      TABLES pnd*married / CHISQ ;
```

Table of pnd by Married			
Pnd	Married(Married Mother)		
Frequency Percent Row Pct Col Pct	0	1	Total
0	1452 29.04 29.06 99.86	3545 70.90 70.94 99.97	4997 99.94
1	2 0.04 66.67 0.14	1 0.02 33.33 0.03	3 0.06
Total	1454 29.08	3546 70.92	5000 100.00

Table 13: Result of PROC FREQ with 50% of cells having expected counts less than 5

As can be seen in the table above, 50% of the cells have an expected count less than 5. Rather than interpret the chi-square, skip over that table and look at the next table SAS produced.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1452
Left-sided Pr <= F	0.2045
Right-sided Pr >= F	0.9754
Table Probability (P)	0.1799
Two-sided Pr <= P	0.2045

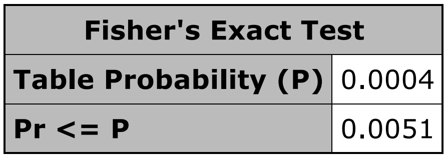
Table 14: Fisher's Exact Test produced by PROC FREQ with CHISQ option

Although twice as many extremely low birthweight babies were born to unmarried mothers, you can see from the Fisher’s result that this difference is not statistically significant from zero.

How to get a Fisher’s Exact Test with more than two categories

It’s great that SAS automatically produces a Fisher’s Exact Test for 2 x 2 tables but what if you have more than two categories? What if your birthweight variable is “very low”,”low” and “normal”? Simple. Just replace the CHISQ in your TABLES statement with FISHER. You’ll still get the same cross-tabulation , chi-square and a table that gives the exact probability of this table.

```
PROC FREQ DATA=vlow ;
TABLES bwt*married / FISHER ;
```



Fisher's Exact Test	
Table Probability (P)	0.0004
Pr <= P	0.0051

Table 14: Fisher’s Exact Test produced by PROC FREQ with FISHER option with > 2 categories

MCNEMAR (AND, FOR THAT MATTER, KAPPA) FOR MATCHED PAIRS

Chi-square is an extremely useful test for a wide variety of situations (and so generic that it is not reviewed here), but it has a few assumptions. One, addressed above, is that there is an expected count of at least five per cell. A second assumption is that the data represent independent observations. In many biomedical applications, the latter assumption is violated. The same subjects are measured pre- and post-treatment for symptoms. The same test results are read by two raters. The same statement in PROC FREQ will address both of these situations.

McNemar Test of Marginal Homogeneity

Suppose you have a drug that has a side effect of nausea for some patients. You hypothesize that eating before taking the drug will reduce these effects. For a sample of 200 patients, you record whether your patients experience nausea when taking the drug. Then, for the next administration, you require them to eat a small meal within five minutes of taking the medication. You record any indications of nausea again. Your hypothesis is that the distribution of side effects (nausea present or absent) will differ based on whether or not they ate food with the medication. That is the marginal probabilities will be different.

```
PROC FREQ DATA=med_eff ;
      TABLES drug1*drug2;
      TEST AGREE ;
```

Table of drug1 by drug2			
drug1	drug2		
Frequency Percent Row Pct Col Pct	nausea	normal	Total
nausea	15 7.50 71.43 68.18	6 3.00 28.57 3.37	21 10.50
normal	7 3.50 3.91 31.82	172 86.00 96.09 96.63	179 89.50
Total	22 11.00	178 89.00	200 100.00

Table 15: Result of PROC FREQ

Clearly, the above table is no different than the typical PROC FREQ output. However, the TEST statement with AGREE keyword produces an additional table, which gives the probability of obtaining these results if the null hypothesis is true, that the marginal probabilities are in fact equal. As can be seen below, this hypothesis is accepted.

McNemar's Test	
Statistic (S)	0.0769
DF	1
Pr > S	0.7815

Table 16: McNemar Test produced by PROC FREQ with TEST statement

In case you were wondering, the equation for the test is very simple $(B-C)^2$ / (B+C) - in the example above, 1/13.

Why Kappa results are completely different from McNemar

Because both are produced by using the same TEST AGREE statement, and both are used with matched pairs, it can be confusing when McNemar and Kappa give dramatically different results. This confusion can be cleared up, though, when one realizes that these are two different tests with different hypotheses that are computed completely differently. While McNemar tests the hypothesis that the marginal probabilities are the same, Kappa tests the hypothesis that the agreement observed is greater than the agreement expected by chance and is not nearly as simple of a computation (although not really all that difficult, either).

If, in the example above, rather than the same subjects being assessed under two conditions, the experiment involved two raters assessing the same subjects, the exact same reports of symptoms were rated by two nurse practitioners as either indicating nausea or not, the same PROC FREQ would be used with the same TEST statement. The only addition is an optional ODS GRAPHICS ON statement and a required KAPPA option on the TABLES statement. Actually, leaving off the TEST statement has no effect, the same results will be produced.

```
ODS GRAPHICS ON ;

PROC FREQ DATA=med_eff ;
      TABLES drug1*drug2 / KAPPA;
      TEST AGREE ;
```

Results from the Kappa test are shown in Table 16, with a Kappa coefficient of .66 indicating a moderate degree of agreement.

Simple Kappa Coefficient	
Kappa	0.6613
ASE	0.0873
95% Lower Conf Limit	0.4901
95% Upper Conf Limit	0.8324

Table 17: Kappa coefficient produced by PROC FREQ with KAPPA option

As can be seen from the final table produced by this procedure, the null hypothesis is rejected.

Test of H0: Kappa = 0	
ASE under H0	0.0707
Z	9.3551
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Table 18: Test of Kappa =0, produced by PROC FREQ with Kappa option

ODS Statistical Graphics

Kappa plots can be produced specifically by a PLOTS=KAPPAPLOT option on a TABLES statement, or, as in this example, by the use of an ODS GRAPHICS ON statement. ODS GRAPHICS ON will not only produce the Kappa plot, shown below, for this particular procedure, but will also continue to produce graphics as well as tabular output for all of the statistical procedures in your program until you turn it off by specifying ODS GRAPHICS OFF.

Although discussing all of the statistical graphics are far beyond the scope of this paper, the reader is strongly encouraged to examine some of the vast number of options produced either by default when ODS GRAPHICS ON is specified or by specific request, as in the STDRATE procedure above.

As with the risk plots above, the Kappa plot illustrates multiple statistics simultaneously. The size of each square indicates frequency, so it can be seen at a glance that normal ratings far outnumber ratings of nausea. The darker areas indicate agreement and it is also clear that there is more agreement on normal ratings than on abnormal (nausea) ratings.

While relatively old in software terms, the paper by Rodriguez (2004) gives an overview to a wide range of statistical graphics commonly used and is highly recommended.

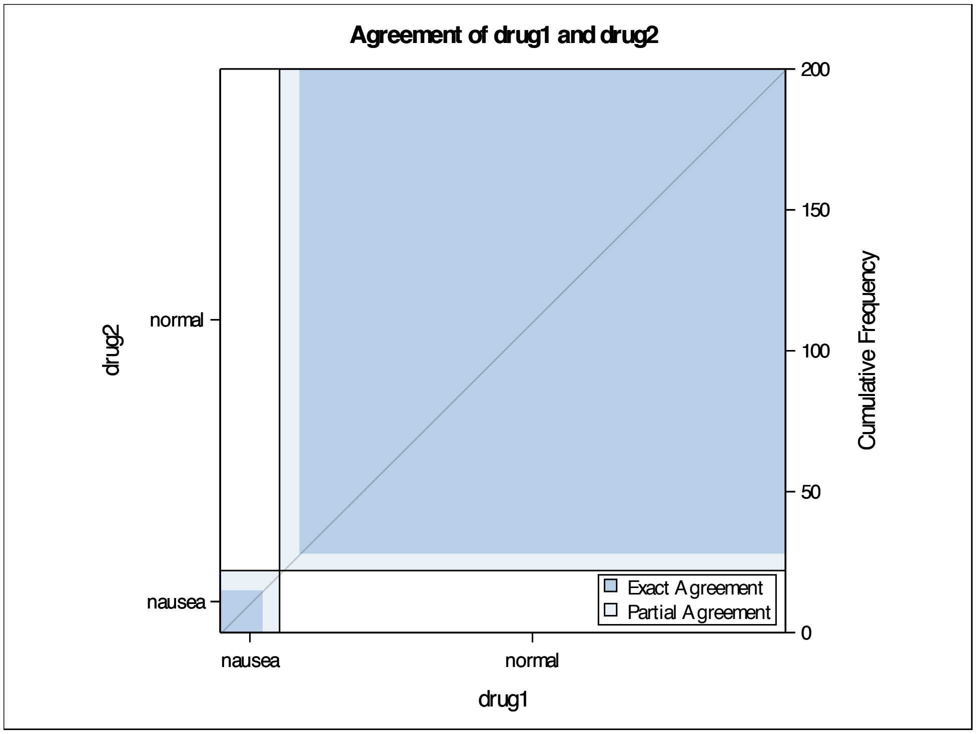


Figure 2: Graph from PROC FREQ with Kappa option, produced with ODS GRAPHICS ON

CONCLUSION

It is impossible to know "all of SAS" and anyone who claims to possess such all-encompassing knowledge is either severely misguided or a pathological liar of such proportion that a career in politics is recommended. However, by identifying broad questions of interest in research in public health and biomedical research, the biostatistician can quickly become adept in using procedures that address common issues.

To further expand his or her tool kit and become an even more invaluable member of the team, two additional recommendations are:

1. Become familiar with SAS procedures for data manipulation to create data sets for analysis. As we have seen, PROC TRANSPOSE is your friend, and
2. Take advantage of SAS statistical graphics. In presenting statistical results and conclusions to non-technical audiences, a picture is truly worth a thousand words.

REFERENCES

Center for Disease Control (2014). User Guide to the 2014 Public Use Natality Data. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/natality/UserGuide2014.pdf

Gordis, L. (2014) Epidemiology (5th ed.) Philadelphia, PA: Saunders, Elsevier.

Rodriguez, R. (2004). An Introduction to ODS for Statistical Graphics in SAS□ 9.1 http://www2.sas.com/proceedings/sugi29/204-29.pdf

SAS Institute (2005). Usage note *23127:* estimating the odds ratio for matched pairs data with binary response**. http://support.sas.com/kb/23/127.html**

SAS INSTITUTE, INC. (2015). Usage note *24170:* estimating sensitivity, specificity, positive and negative predictive values, and other statistics**.** **http://support.sas.com/kb/24/170.html**

World Health Organization (2016). Genes and human disease. http://www.who.int/genomics/public/geneticdiseases/en/index1.html

Yuan, Y. (2013). Computing direct and indirect standardized rates and risks with the STDRATE procedures. Proceedings of the SAS Global Forum. https://support.sas.com/resources/papers/proceedings13/423-2013.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

AnnMaria De Mars
The Julia Group
1502 Broadway, Ste. 303
Santa Monica, CA 90404
(310) 717-9089
annmaria@thejuliagroup.com
www.thejuliagroup.com