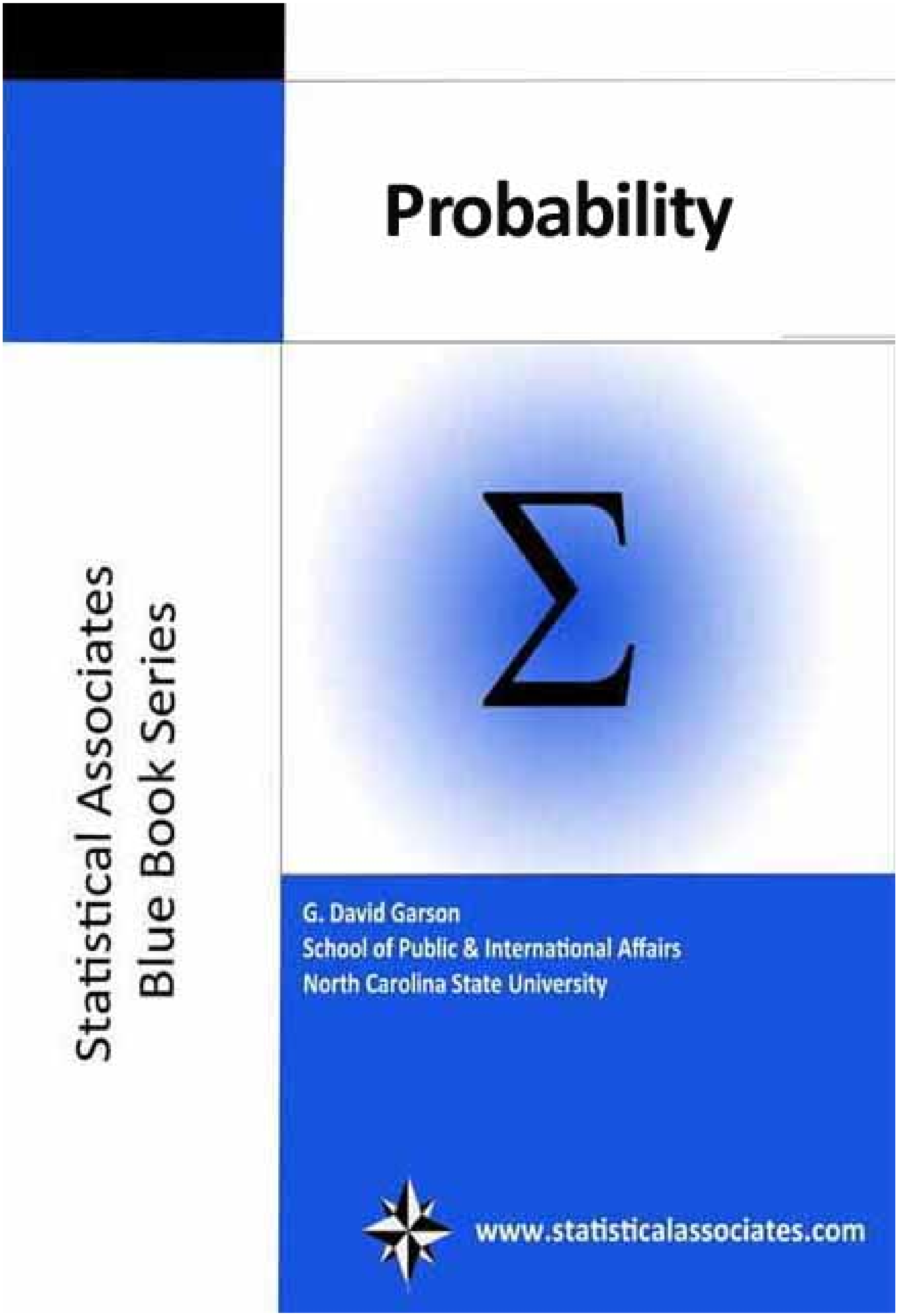

Probability
Statistical Associates
Blue Book Series
G. David Garson
School of Public & International Affairs
North Carolina State University
www.statisticalassociates.com

Contact:

G. David Garson, President
Statistical Publishing Associates
274 Glenn Drive
Asheboro, NC 27205 USA

Email: sa.publishers@gmail.com
Web: www.statisticalassociates.com

Table of Contents

PROBABILITY

Overview

The mathematics of probability underlies a variety of other topics in social science research, including sampling in survey research, significance testing, and maximum likelihood estimation methods.

Key Concepts and Terms

Counting

Permutations

Permutations are how many ways a given number of things can be sequenced. The standard notation is that ${}_nP_n$ is the permutations (sequences) of all n things in a set of n things; ${}_nP_r$ is the permutations on n things taken r at a time. The relevant formulas are below:

$${}_nP_n = n!$$

$${}_nP_r = n!/(n - r)!$$

where ! is the factorial operation (n! = 1*2*3*...*n). For example, the number of permutations of 4 things taken 3 at a time is 4!/(4-3)! = 24/1 = 24. Thus for items, A, B, C, and D, the 24 permutations are:

ABC, ACB, ADB, ABD, ADC, ACD
BAC, BCA, BDA, BAD, BCD, BDC
CAB, CBA, CBD, CDB, CAD, CDA
DAC, DCA, DAB, DBA, DBC, DCB

We could use permutation, for instance, in a study of the effects of ballot order of candidates, when 4 candidates from a possible 10 are to be listed -- we would find there are 5,040 permutations of 10 candidates taken 4 at a time.

Combinations

Combinations refer to how many subsets can be derived from a given set, regardless of permutation sequence. For instance, ABC, ACB, BCA, BAC, CAB, and CBA are 6 permutations but only 1 combination. The relevant counting formulas are below:

$$_{n}C_{n} = 1$$

$$_{n}C_{r} = n!/(n - r)!r!$$

By cancellation, n things combined 2 at a time = $_{n}C_{2}$ = n!/(n - 2)!2! = (n*(n-1))/2

For example, the number of combinations of 4 things taken 3 at a time is 4!/[(4-3)!3!] = 24/(1*6) = 4. The 4 combinations of A, B, C, and D are:

A, B, and C in any order
A, B, and D in any order
A, C, and D in any order
B, C, and D in any order

A poker example

What are the chances of getting a royal flush, which is 10 through Ace all of the same suit?

- There are four chances, one for each suit: this is $_{4}C_{1} = 4$
- The number of combinations of 5 cards: $_{52}C_{5}$ = 2,598,960
- The probability of getting a royal flush : 4/2.598,960 = 0.0000015291
- The probability of not getting a royal flush: 1 - 0.0000015291 = 0.9999984709
- Odds of getting a royal flush: 0.0000015291/0.9999984709= 0.0000015291. This is approximately equal to 1/650,000 or an odds of 1 in 650,000.

Permutations and Combinations of Sets Not All Different

The foregoing formulas assumed that we were dealing with a set of n things, each of which was different from the other. But what if we have a set of, say, individuals, some of whom are Republicans and some Democrats, and we are interested in permutations by party, not individual? The relevant formulas are below:

$${}_{n}C_{n1,\ n2,\ \ldots,\ nk} = 1$$

$${}_{n}P_{n1,\ n2,\ \ldots,\ nk} = n!/n1!n2!\ldots nk!$$

If we are going go seat all n individuals at a speaker's table and don't care about sequence, there is still only 1 combination of n people taken n at a time. However, if we do care about sequence (permutation), we need to know how many seating arrangements by party affiliation there are of, say, 4 people, 2 of them Republicans and 2 Democrats: ${}_{4}P_{2,\ 2} = 4!/(2!2!) = 24/(2*2) = 6$. The six seating sequences are:

RRDD
RDRD
RDDR
DDRR
DRDR
DRRD

Permutations with recurrences

Permutations with recurrences are sometimes needed, as when figuring out how many phone numbers exist in a 7-digit code which, of course, allows any given numeral to recur. The relevant formula is:

$${}_{n}P_{r} = n^{r}$$

Thus, for 10 numerals taken 7 at a time there are 10^{7} = 10 million permutations.

Number of combinations of any size

Sometimes we want to know the number of combinations of any size, as when we might want to know for 6 people, how many committees we could form (that is, of size 6 or size 5 or size 4, etc., down to committess on 1. The relevant formula is:

$${}_{n}C_{0,\ 1,\ldots,n} = 2^{n} - 1$$

The "-1" in the formula above subtracts out the null combination of 0 members. Thus, excluding the null committe, for six people there could be $2^6 - 1 = 63$ possible combinations for committees of any size from 1 to 6.

Software

SPSS syntax for generating permutations and combinations is linked from http://www.spsstools.net/SampleSyntax.htm#Combinations .

A SAS macro for generating permutations and combinations is located at http://support.sas.com/techsup/technote/ts498.html .

For Stata, there is a user-written .ado file for combinations: type “ssc install tuples” to install the file, then type “. help tuples” to view instructions. There is also a third-party ado file for permutations and combinations: type “ssc install percom”. Instructions are located at

http://ideas.repec.org/c/boc/bocode/s457500.html. Percom implements the commands permin and combin, generating permutations and combinations respectively from a variable list.

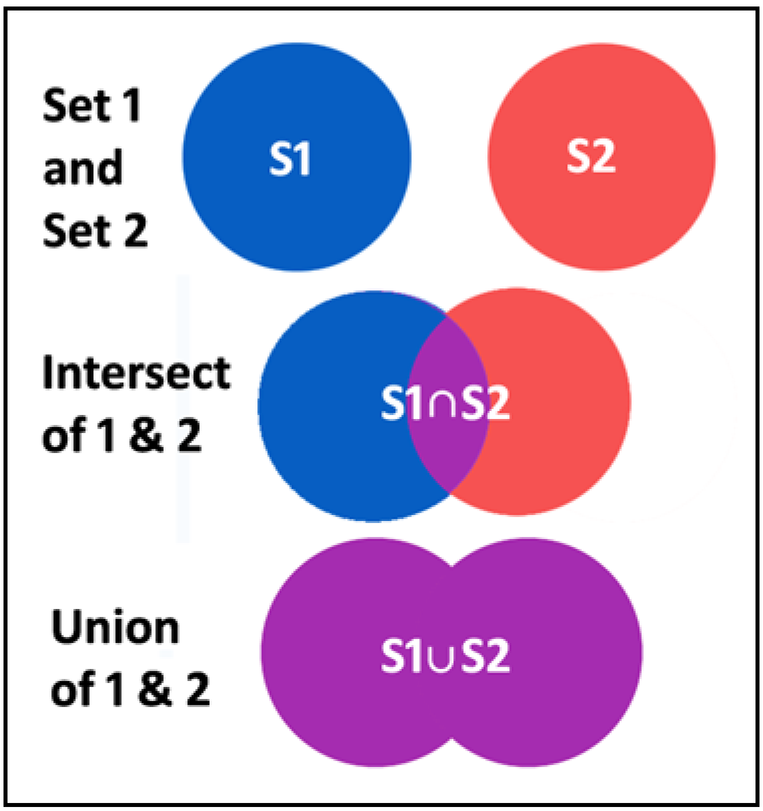
Probability

Notation for probability

Below is a common way by which researchers refer to various types of probability.

Probability Terms	
$P(A)$	The probability from 0 to 1.0 that A will occur
$P(A \cup B)$	The probability of A or B occurring
$P(A \cap B)$	The probability of A and B both occurring.
$P(A\|B)$	The probability of A occurring, given that B has occurred (conditional probability)

In Venn diagram form, this generalizes to sets 1 and 2 as depicted below.



Independent and dependent events

Two events are *independent* if any given outcome for the first event does not affect the probabilities of any outcome for the second event. Coin tosses are independent events. Two events are *dependent* if the outcome of the first <u>does</u> affect the second. For instance, the chances of being picked at random to be on a committee is a dependent event because each pick eliminates one person from the pool, changing the probability of being selected for everyone else on the succeeding pick.

Probability rules

The rules for probability operations can vary depending on whether one is dealing with independent or dependent events:

Probability Rules		
	Independent Events	Dependent Events
P(A∩B)	P(A)*P(B)	P(A)*P(B|A)
P(A∪B)	P(A)+P(B) - P(A∩B)	same
P(A|B)	P(A)	(P(A∩B)/P(B)

Examples for independent events

For independent events like coin tosses, let P(A) be the probability of getting heads on the first toss, equal to 1/2. Let P(B) be the probability of getting heads on the second toss, also 1/2. The probability of getting two heads in a row (P(A∩B)) is 1/2 times 1/2 = 1/4. That is, the joint probability of two independent events is the product of their individual probabilities. The probability of getting heads on the first toss <u>or</u> getting heads on the second toss is the sum of their individual probabilities minus their joint probability: P(A)+P(B) - P(A∩B) = 1/2 + 1/2 -.1/4 = 3/4. The conditional probability is the same as the probability. Thus, the probability of getting heads on the second toss, given that the first toss is heads, is unchanged: P(B|A) = P(B) = 1/2.

Examples for dependent events

For dependent events like picking committee members from a pool, let P(A) be the probability of picking a Democrat from a pool with two Democrats and two Republicans, so P(A) = 1/2. Let P(B) be the chance of getting a Democrat second, regardless of the first choice. If a Democrat has been selected first, there will be 1 Democrat and 2 Republicans left, giving a conditional probability of picking the Democrat second of 1/3. However if a Republican is picked first, the conditional probability will be 2/3. Overall, the chance is the mean, or P(B) = 1/2, since either condition is equally likely. Now let P(B|A) be the probability that the second person picked is a Democrat, given that the first person picked is a Democrat. Once a Democrat has been picked, there will be 1 Democrat and 2 Republicans left, so the conditional probability of picking a Democrat is 1/3. For dependent events, the joint probability is the probability of the first event times the conditional probability of the second event. Thus P(A∩B) = P(A)*P(B|A) = 1/2*1/3 = 1/6. That is, there is one chance in six of picking a Democrat first and another Democrat second for the committee, from a pool of 2 Democrats and 2 Republicans. The chance of getting a Democrat first or getting a Democrat second -- P(A∪B) -- is computed as for independent events and is also 3/4. The conditional probability of picking a Democrat second, given a Democrat has already been picked, P(B|A), equals P(A∩B)/P(B) = (1/6)/(1/2) = 1/3, as noted above.

Expected Values

Expected values in tables

Chi-square is one of the most common significance tests. In a crosstabulation, for example, it can test if one variable is related significantly to another variable. The chi-square formula is a function of the difference between the observed cell count and what is expected by chance. What is expected by chance is a function of probabilities, illustrated in the figure below.

In this table, there are 15 males in a sample of 30 people. Thus the probability of being male is 15/30 = 50% = .5. There are 15 Republicans out of 30 people, so the probability of being Republican is also 15/30 = 50% = .5. The joint probability that a given person in the sample is a male Republican is the product of these two probabilities: .5*.5 = 25% = .25. Since there are 30 people total, and since 25% of 30 is 7.5, the expected number of male Republicans is 7.5. (As a computational shortcut, the expected number will always be the product of the marginals divided by the total: (15*15)/30 = 7.5). The full explanation of chi-square is found in the separate Statistical Associates' volume on "Significance Testing," but in general chi-square is based on the observed minus expected counts for each cell in the table. The larger these differences, the more significant the relationship in the table.

Probability and Expected Values

	A	B	C	D	E	F	G	H	I	J
2										
3	COMPUTING PEARSON CHI-SQUARE FOR A 2-BY-2 TABLE									
4										
5			Party * Sex Crosstabulation							
6				Sex						
7				Male	Female	Total		(O - E)	$(O - E)^2$	$(O - E)^2/E$
8	Party	Rep	Count	10	5	15		2.5	6.25	0.833333
9			Expected	7.5	7.5			-2.5	6.25	0.833333
10								-2.5	6.25	0.833333
11		Dem	Count	5	10	15		2.5	6.25	0.833333
12			Expected	7.5	7.5					
13								SUM =		3.333333
14	Total		Count	15	15	30				
15			Expected	15	15			This is the Pearson		
16								chi-square.		
17			Note: Expected = (15*15)/30 = 7.5							

PROBABILITY OF BEING MALE: 15/30 = .5

PROBABILITY OF BEING REPUBLICAN: 15/30 = .5

PROBABILITY OF BEING A MALE REPUBLICAN = .5*.5 = .25

EXPECTED NUMBER OF MALE REPUBLICANS = .25 * 30 = 7.5

**CHI-SQUARE USES OBSERVED MINUS EXPECTED
AS THE BASIS OF SIGNIFICANCE TESTING**

Bibliography

Bertsekas, Dimitri P. & Tsitsiklis, John N. (2008). *Introduction to probability, 2nd edition*. Belmont, MA: Athena Scientific.

Klein, Grady & Dabney, Alan (2013). *The cartoon introduction to statistics. NYC:* Hill and Wang.

Mosteller, Frederick, Robert Rourke, and George Thomas (1961). *Probability with statistical applications*. Reading, MA: Addison-Wesley. A classic text.

Roussas, George G. (2006). *Introduction to probability*. Salt Lake City, UT: Academic Press.

Statistical Associates Publishing
Blue Book Series

Association, Measures of
Canonical Correlation
Case Studies
Cluster Analysis
Content Analysis
Correlation
Correlation, Partial
Correspondence Analysis
Cox Regression
Creating Simulated Datasets
Crosstabulation
Curve Estimation & Nonlinear Regression
Delphi Method in Quantitative Research
Discriminant Function Analysis
Ethnographic Research
Evaluation Research
Factor Analysis
Focus Group Research
Game Theory
Generalized Linear Models/Generalized Estimating Equations
GLM (Multivariate), MANOVA, and MANCOVA
GLM (Univariate), ANOVA, and ANCOVA
Grounded Theory
Life Tables & Kaplan-Meier Survival Analysis
Literature Review in Research and Dissertation Writing
Logistic Regression: Binary & Multinomial
Log-linear Models,
Longitudinal Analysis
Missing Values & Data Imputation
Multidimensional Scaling
Multiple Regression

Narrative Analysis
Network Analysis
Neural Network Models
Nonlinear Regression
Ordinal Regression
Parametric Survival Analysis
Partial Correlation
Partial Least Squares Regression
Participant Observation
Path Analysis
Power Analysis
Probability
Probit and Logit Response Models
Research Design
Scales and Measures
Significance Testing
Social Science Theory in Research and Dissertation Writing
Structural Equation Modeling
Survey Research & Sampling
Testing Statistical Assumptions
Two-Stage Least Squares Regression
Validity & Reliability
Variance Components Analysis
Weighted Least Squares Regression

Statistical Associates Publishing
http://www.statisticalassociates.com
sa.publishers@gmail.com

