

A Look at Probabilities and Odds and Risks

Nalini Ravishanker

Department of Statistics

University of Connecticut, Storrs, CT 06269

nalini@stat.uconn.edu

[www.stat.uconn.edu/ nalini](http://www.stat.uconn.edu/~nalini)

Conditional, Joint, Marginal Probabilities

For a specific biomedical research:

Target Population is the entire set of subjects at whom the research is aimed.

Example: Screening for Cancer in a community.

Target Population: All persons in that community who are at risk for the disease.

Let Test: Test Outcome - Negative (N), or Positive (P)

Let Status: True Disease Status - No Disease (ND), or
Disease (D)

The following table is a familiar one!

	Test Result		
Status	Positive (P)	Negative (N)	Total
Disease (D)	0.006	0.009	0.015
No Disease (ND)	0.015	0.970	0.985
Total	0.021	0.979	1.0

A general form is

		Test Result (Factor 1)		
Status (Factor 2)	Positive	Negative	Total	
Yes (D)	π_{11}	π_{12}	π_{1+}	
No (ND)	π_{21}	π_{22}	π_{2+}	
Total	π_{+1}	π_{+2}	1.0	

Marginal Probabilities:

Test:

$$P(\text{Test} = P) = 0.021 : P(\text{Positive Test Result})$$

$$P(\text{Test} = N) = 0.979 : P(\text{Negative Test Result})$$

$$P(\text{Test} = P) + P(\text{Test} = N) = 1.0$$

Addition Rule for Mutually Exclusive Events.

Status:

$$P(\text{Status} = D) = 0.015 : P(\text{Disease})$$

$$P(\text{Status} = ND) = 0.985 : P(\text{No Disease})$$

$$P(\text{Status} = D) + P(\text{Status} = ND) = 1.0$$

Joint Probabilities:

$$P(\text{Test} = \text{P}, \text{Status} = \text{D}) = 0.006$$

$$P(\text{Test} = \text{P}, \text{Status} = \text{ND}) = 0.015: \text{False Positive}$$

$$P(\text{Test} = \text{N}, \text{Status} = \text{D}) = 0.009: \text{False Negative}$$

$$P(\text{Test} = \text{N}, \text{Status} = \text{ND}) = 0.970$$

Conditional Probabilities:

$$P(\text{Test} = P \mid \text{Status} = D) = 0.006/0.015 = 0.4$$

$$P(\text{Test} = P \mid \text{Status} = ND) = 0.015/0.985 = 0.0152$$

Now, $P(\text{Test} = P) = 0.021$.

Dependence between events Test and Status.

$$P(\text{Test} = N \mid \text{Status} = D) = 0.009/0.015 = 0.6$$

$$P(\text{Test} = N \mid \text{Status} = ND) = 0.970/0.985 = 0.9848$$

$$P(\text{Status} = \text{D} \mid \text{Test} = \text{P}) = 0.006/0.021 = 0.286$$

$$P(\text{Status} = \text{D} \mid \text{Test} = \text{N}) = 0.009/0.979 = 0.0092$$

$$P(\text{Status} = \text{ND} \mid \text{Test} = \text{P}) = 0.015/0.021 = 0.714$$

$$P(\text{Status} = \text{ND} \mid \text{Test} = \text{N}) = 0.970/0.979 = 0.991$$

Real Example. A cytological test was done to screen women for cervical cancer. The results are shown in the 2×2 table below (Le, C.T.. Introductory Biostatistics, Wiley, 2003).

Let Test be an event with two outcomes:

Negative (N), Positive (P)

Let True Disease Status be an event with two outcomes:

Disease (D), No Disease (ND)

	Test Result		
Status	Positive (P)	Negative (N)	Total
Disease (D)	154	225	379
No Disease (ND)	362	23,362	23,724
Total	516	23,587	24,103

A general form is

Test Result (Factor 1)				
Status (Factor 2)	Positive (P)	Negative (N)	Total	
Yes (D)	n_{11}	n_{12}	n_{1+}	
No (ND)	n_{21}	n_{22}	n_{2+}	
Total	n_{+1}	n_{+2}	n	

Marginal Proportions:

Test Results:

$$\text{Proportion of Test Positive} = \frac{516}{24,103} = 0.021408$$

$$\text{Proportion of Test Negative} = \frac{23,587}{24,103} = 0.978592$$

Disease Status:

$$\text{Proportion of Disease} = \frac{379}{24,103} = 0.015724$$

$$\text{Proportion of No Disease} = \frac{23,724}{24,103} = 0.984276$$

Joint Proportions:

$$\text{Proportion of Test=P and Status=D} = \frac{154}{24,103} = 0.006389$$

$$\text{Proportion of Test=P and Status=ND} = \frac{362}{24,103} = 0.015019$$

= Proportion of False Positives

$$\text{Proportion of Test=N and Status=D} = \frac{225}{24,103} = 0.009335$$

= Proportion of False Negatives

$$\text{Proportion of Test=N and Status=ND} = \frac{23,362}{24,103} = 0.969257$$

Think of these as estimated probabilities (true unknown proportions) in a 2×2 table

Table of Joint and Marginal Proportions

	Test Result		
Status	Positive (P)	Negative (N)	Total
Disease (D)	0.006389	0.009335	0.015724
No Disease (ND)	0.015019	0.969257	0.984276
Total	0.021408	0.978592	1.0

A general form is

		Test Result (Factor 1)		
Status (Factor 2)	Positive	Negative	Total	
Yes (D)	p_{11}	p_{12}	p_{1+}	
No (ND)	p_{21}	p_{22}	p_{2+}	
Total	p_{+1}	p_{+2}	1.0	

Conditional Proportions:

Proportion of Test = P given Status = D

= Proportion of Diseased who will test Positive

= $154/379 = 0.406$: **Sensitivity**

Out of 379 persons with cancer, 40.6% test Positive!

In the entire population, only 2.1% test Positive.

Implies a strong "statistical" association between Test and Status.

Sensitivity is the Number of Diseased Persons who Screen Positive divided by the Number of Diseased Persons.

Specificity is the Proportion of No Disease Persons who will Screen Negative

$$\textbf{Specificity} = P(\text{Test}=\text{N} \mid \text{Status}=\text{ND}) = 23362/23724 \\ = 0.985$$

Out of 23724 persons without cancer, 98.5% test Negative.

In the entire population, 97.9% test Negative.

Specificity is the Number of Non Diseased Persons who Screen Negative divided by the Number of Non Diseased Persons.

The Screening Test in this example is:

- Highly Specific, and
- Not very Sensitive.

Implications:

- Test a woman with No Disease. Result would almost surely be Negative.

- Test a woman with Disease. There is a 59.4% chance

Test will be Negative!

Given that Test=P, what proportion of persons have

Status=D? **Positive Predictivity**

$$P(\text{Status} = D \mid \text{Test} = P) = 154/516 = 0.298$$

Given that Test=N, what proportion of persons have

Status=ND? **Negative Predictivity**

$$P(\text{Status} = ND \mid \text{Test} = N) = 23362/23587 = 0.9904$$

Positive and Negative Predictivity depend on

- the efficiency of the screening test,
- the disease prevalence of the target population.

Scenario A:

	Test Result		
Status	Positive (P)	Negative (N)	Total
Disease (D)	45,000	5,000	50,000
No Disease (ND)	5,000	45,000	50,000
Total	50,000	50,000	100,000

$P(\text{Disease}) = \text{Prevalence} = .50$

$\text{Positive Predictivity} = P(D|P) = 0.90$

Scenario B:

	Test Result		
Status	Positive (P)	Negative (N)	Total
Disease (D)	9,000	1,000	10,000
No Disease (ND)	9,000	81,000	90,000
Total	18,000	82,000	100,000

$P(\text{Disease}) = \text{Prevalence} = .10$

Positive Predictivity = 0.50

Implications:

- Suppose screening test is highly sensitive and specific.
- Target Population has low disease prevalence (rare disease)
- Then, Positive Predictivity will be Low!

Actual Screening Tests:

Developmental Stage: try on a pilot population.

Get efficiency of screening test via Sensitivity and Specificity.

Application Stage: apply to a target population.

Find predictivity values.

- Data on Status is not available.
- Suppose Disease prevalence is available from national surveys.
- Suppose Sensitivity, Specificity available after developmental stage.

We can use Bayes' Theorem to compute Predictivity

$$\text{Obtain } P(P) = P(P | D)P(D) + P(P | ND)P(ND)$$

$$= \text{Prevalence} \times \text{Sensitivity} + (1 - \text{Prevalence}) \times (1 - \text{Specificity})$$

$$\text{Then, } P(D | P) = P(P | D)P(D) / P(P)$$

Odds Ratio: Compares odds of Test Positive in Disease Group to the odds of Test Positive in No Disease Group.

Let $p_1 = n_{11}/n_{1+}$ and $p_2 = n_{21}/n_{2+}$.

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

$0 < OR < \infty$; It estimates a Population Odds Ratio = Num/Denom, where

Num = $P(\text{Test}=\text{P}, \text{Status}=\text{D}) P(\text{Test}=\text{N}, \text{Status}=\text{ND})$,

Denom = $P(\text{Test}=\text{P}, \text{Status}=\text{ND}) P(\text{Test}=\text{N}, \text{Status}=\text{D})$.

$OR = 1$ implies no association between Test and Status.

$OR > 1$ implies Disease group is more likely than No Disease Group to Test Positive.

$OR < 1$ implies Disease group is less likely than No Disease Group to Test Positive.

For our data, $OR = 44.17$

Log Odds ratio = $\log(OR) = 3.78805$;

$Var(\log(OR)) = [1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}]$

$100(1 - \alpha)$ C.I. for population $\log(OR)$ is

$$\log(OR) \pm z_{\alpha/2} \sqrt{Var(\log(OR))}$$

For our example, it is $(3.5587, 4.0174)$.

$100(1 - \alpha)$ C.I. for population OR is

$$\exp\{\log(OR) \pm z_{\alpha/2} \sqrt{Var(\log(OR))}\}$$

For our example, it is $(35.1175, 55.5565)$.

Relative Risk: RR :

$$\begin{aligned} RR &= p_1/p_2 \\ &= OR \times \frac{\{1 + (n_{21}/n_{22})\}}{\{1 + (n_{11}/n_{12})\}} \end{aligned}$$

For cross-sectional data, called **prevalence ratio**.

Chi-Square Test for Association

H_0 : No association between Test and Status, i.e.,

$$\pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all } i, j = 1, 2.$$

Let $m_{ij} = np_{i+}p_{+j}$,

with $p_{i+} = \frac{n_{i+}}{n}$ and $p_{+j} = \frac{n_{+j}}{n}$.

Screening Test example: $m_{11} = 8.1$; $m_{12} = 370.9$; $m_{21} = 507.9$; $m_{22} = 23216$.

Pearson's Chi-square Test Statistic:

$$Q_P = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

If $m_{ij} > 5$ for all i, j : expected cell counts,

then $Q_P \sim \chi_1^2$ under H_0 .

Screening Test example: $Q_P = 2723.3$ SIG

Randomization Chi-square Statistic

Assume n_{i+} for $i = 1, 2$ are fixed by the sampling design.

n_{+j} for $j = 1, 2$ are fixed under H_0 . $f(n_{ij})$ corresponds to

HG dist. under H_0 .

$$m_{ij} = E(n_{ij}|H_0) = \frac{n_{i+}n_{+j}}{n},$$

$$v_{ij} = Var(n_{ij}|H_0) = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n^2(n-1)}$$

For large n ,

$$Q_R = \frac{(n_{11}-m_{11})^2}{v_{11}} \sim \chi_1^2$$

Screening Test example: $Q_R = 2723.2$ SIG