Mixture Models for Capture-Recapture Data

Dankmar Böhning

Invited Lecture at

Mixture Models between Theory and Applications"

Rome, September 13, 2002

How many cases **n** in a population?

- Registry identifies n_{obs} cases
- p_0 probability of being **not** identified by the registry
- **Then:**

$$n = n\,p_0 + (1 - p_0)\,n$$

$$= \text{unobserved} + \text{observed cases} = n\,p_0 + n_{obs}$$

$$n_{HTE} = n_{obs} / (1 - p_0)$$

(Horwitz-Thompson)

An Example

- A registry could identify 250 cases from a study population
- Assume that the inclusion probability

  $(1-p_0) = 0.25$ known

- Then $n_{HTE} = n_{obs} / (1 - p_0)$

  $= 250 / 0.25 = 1000$

How HAT-approach can be used for Screening

- „In a large population survey of 15000 persons were screened given **chest radiographs**, and the physicians noted possible **pulmonary artery enlargement (PAE)** in 230 of these patients. The enlargement was confirmed in a second reading in 203 of these 230 persons. A sample of 175 of the 14770 chest radiographs in which no enlargement of the pulmonary artery was noted yielded 12 radiographs that were actually postive for pulmonary enlargement.“ Levy & Lemeshow 91

What do we have ?

	PAE	No PAE	
Test +	203	27	230
Test -	? (12)	?(163)	14 770 (175)
	? **1216**	?	15000

PPV = P(D+|T+) = 203/230 = 0.8826

NPV = P(D-|T-) = 163/175 = 0.9314

using Bayes theorem

Sensitivity = 0.1670

$$\mathbf{n_{HTE} = n_{obs} / (1 - p_0) = 203 / 0.1670 = 1216}$$

Horwitz-Thompson-Approach seems easy, but ...

inclusion probability often unknown
and, consequently,

approaches differ in the way they estimate the inclusion probability, or in other words, how they

model p_0 

Information typically available in a disease (cancer) registry

- A case is identified by at least one source, typically several sources such as pathology, hospitals, physicians, death certificate, ...
- Potentially further covariates are available such as age at diagnosis, gender, ..

... in more detail

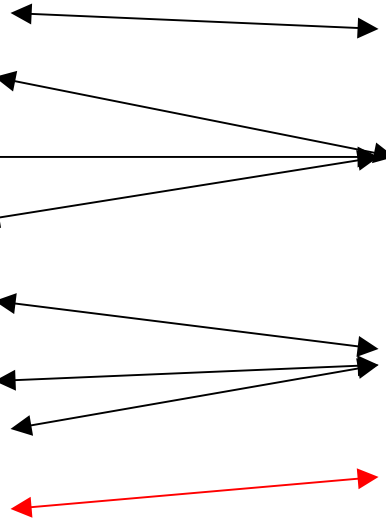
ID	Source A	Source B	Source C	Counting Sources
001	1	0	0	1
002	0	1	1	2
003	0	0	0	0
004	1	0	1	2
005	1	1	1	3
...	...	...	...	...

Two major streams of development ...

illustrated with 3 sources

- modelling a multiway contingeny table
- modelling the counting sources distribution

A	B	C	Freq
1	1	1	n_{111}
1	1	0	n_{110}
0	1	1	n_{011}
1	0	1	n_{101}
1	0	0	n_{100}
0	1	0	n_{010}
0	0	1	n_{001}
0	0	0	?



Counting Sources	Frequency
3	$n_3 = n_{111}$
2	$n_2 = n_{101} + n_{110} + n_{100}$
1	$n_1 = n_{010} + n_{011} + n_{001}$
0	$n_0 = n_{000} = ?$

An Example for first approach with two sources

- Inclusion probabilities
- Associated data

$$\begin{matrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{matrix} \qquad \begin{matrix} n_{11} & n_{10} \\ n_{01} & n_{00} \end{matrix}$$

a) estimate p_{11} as n_{11}/n

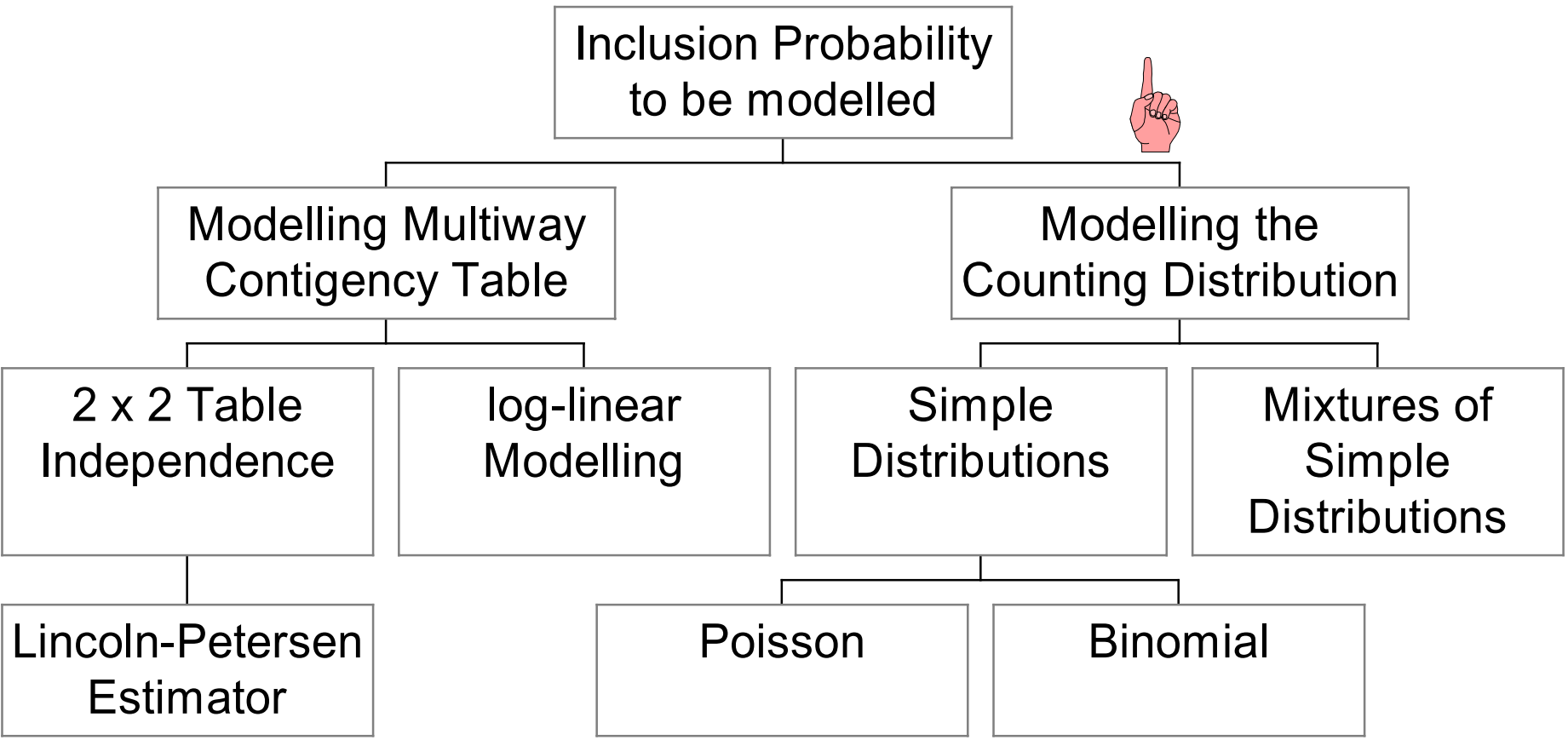
b) on the other hand, using *independence* $p_{11} = p_{1\cdot}\, p_{\cdot 1}$

one can estimate p_{11} as $n_{1\cdot}/n \times n_{\cdot 1}/n$

equating a) and b), leads to $\mathbf{n_{LP} = n_{1\cdot}\ n_{\cdot 1} / n_{11}}$

the Lincoln-Petersen estimate of number of cases

Elaborate Developments



Reasons for *not* following the first approach

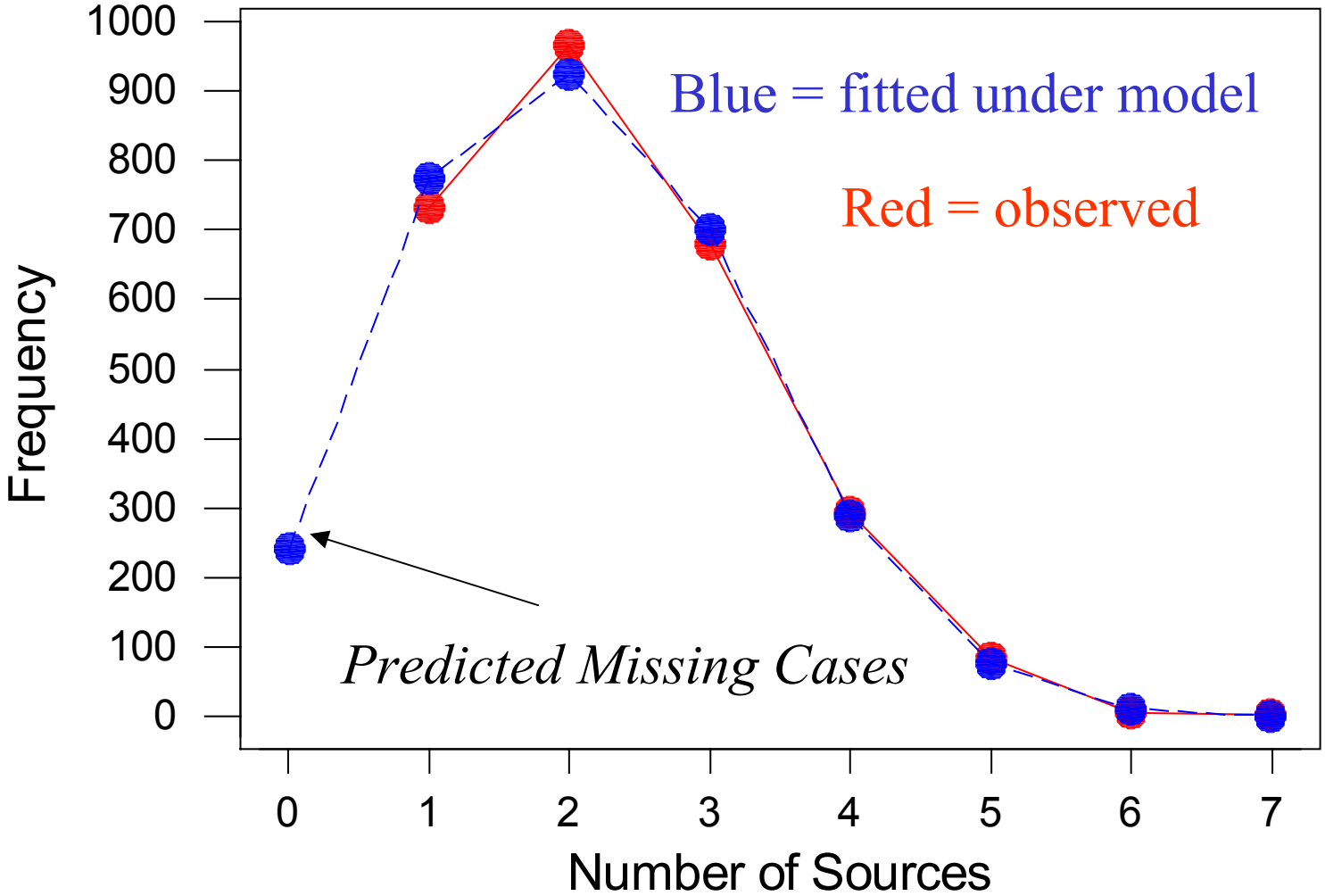
- (log-linear) modelling becomes quite complex; with 4 sources more than 100, with 5 sources more than 7000 possible models
- potential danger of one dominating source
- wortking with the counting sources distribution will only require the decision about one model

The Counting Distribution

ID	Source A	Source B	Source C	...	Counting Sources
001	1	0	0	...	1
002	0	1	1	...	2
003	0	1	0	...	1
004	1	0	1	...	2
005	1	1	1	...	3
...	...	...	...		...

Distribution of Observed and Predicted Counts of Sources

for fictional registry data



Simple Distributional Count Models

Binomial

$$f(y,\theta) = \binom{m}{y} \theta^{y} (1-\theta)^{m-y} \quad , y=0,1, \ldots,m$$

(m = Number of Sources, θ = Listing Probability)

Predicted Probability of a Zero:

$$p_0 = f(y=0,\theta) = \binom{m}{0} \theta^{0} (1-\theta)^{m-0} = (1-\theta)^{m}$$

Number of cases

Suppose, θ where *known*, then

the *estimated number of cases*

$$n_{HTE} = n_{obs} / (1 - p_0)$$

where $p_0 = (1-\theta)^m$.

Estimation of Listing Probability

in summary:

$n = n_{obs} / (1 - p_0)$, where $p_0 = (1-\theta)^m$,

also, if n is *given*

$\theta = (n_0\ 0 + n_1\ 1 + n_2\ 2 + \ldots + n_m\ m)/(n\ m)$

Estimation of Listing Probability

Consequently,

Step 0. Choose, some initial value for $\theta = \theta^{(1)}$ (for ex. $\theta = 1/2$)

Step 1. Compute $n^{(1)} = n_{obs} / (1 - p_0^{(1)})$, where $p_0^{(1)} = (1-\theta^{(1)})^m$

Step 2. Compute $\theta^{(2)} = (n_1\ 1 + \ldots + n_m\, m)/(m\ n^{(1)})$

Step 3. Continue iteration until convergence.

Estimation of Listing Probability

Step 1. Compute $n^{(2)} = n_{obs} / (1 - p_0^{(2)})$,
where $p_0^{(2)} = (1-\theta^{(2)})^m$

Step 2. Compute
$\theta^{(3)} = (n_1\ 1 + \ldots + n_m\ m)/(n^{(2)}\ m)$

..... and so on ...

Version of EM algorithm (DLR 1977)

- for finding maximum likelihood estimator of θ
- imputing the number of missing data (as by-product)
- (strong) convergence is assured (in this case)

A Demonstration

Iteration j	$\theta^{(j)}$	$\mathbf{n}^{(j)}$
1	0.50	2790
2	0.3260	2955
3	0.3078	2996
...	...	...
20	0.3021	3011
21	0.3021	3011

Simple Distributional Count Models

Poisson

$f(y,\theta) = e^{-\theta}\,\theta^{y} \,/\, y\,!\,,\ y=0,1,\ldots$

(suitable for m = Number of Sources large)

Predicted Probability of a Zero:

$p_0 = f(y=0,\theta) = e^{-\theta}\,\theta^{y} \,/\, y\,! \;= e^{-\theta}$

Estimation of Listing Parameter and Prediction of Number of Cases

Similar to the Binomial,

only difference is the way the missing cases are predicted:

Poisson: $p_0 = f(y=0,\theta) = e^{-\theta}$

Binomial: $p_0 = f(y=0,\theta) = (1-\theta)^m$

More flexible and robust approach through mixtures

- Simple counting sources distributions such as Binomial and Poisson require assumptions such as homogeneity of listing probabilities that are seldom met in reality
- allowing the listing probability to vary in unobserved sub-populations will be more realistic

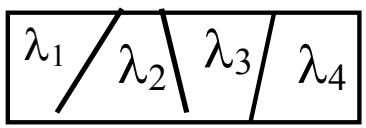
The mixture approach in a nutshell

homogeneity

one-parametric density $f(y,\theta)$

(typically $f(y,\theta)$ will be a simple density like Binomial or Poisson)

heterogeneity



density in subpop. j: $f(y,\theta_j)$

The mixture approach in a nutshell

latent variable Z describing population membership

joint density $f(x,z)$ with
$f(x,z) = f(x \mid z)f(z) = f(x,\theta_z)q_z$

marginal or mixture density:

$f(x,Q) = f(x,\theta_1)q_1 + \ldots + f(x,\theta_k)q_k$

$Q = \begin{pmatrix} \theta_1 & \ldots \theta_k \\ q_1 & \ldots q_k \end{pmatrix}$ is *mixing distribution*

Estimation of parameters works in principle as before, though technically more elaborated

in summary:

if $Q = \begin{pmatrix} \theta_1 \ldots \theta_k \\ q_1 \ldots q_k \end{pmatrix}$ given, then estimate

$n = n_{obs} / (1 - p_0)$, where $p_0 = f(y=0,Q)$,

also, if n is *given*, then estimate

Q by the NPMLE

Special Mixtures

Mixtures of Binomials

$$f(y, \theta_j, q_j) = \sum_{j=1}^{k} q_j \binom{m}{y} \theta_j^y (1-\theta_j)^{m-y}, \quad y=0,1,\ldots,m$$

(m = Number of Sources,
θ_j = Listing Probability in sub-population j,
q_j = weight of sub-population)

Predicted Probability of a Zero:

$$p_0 = f(y=0, \theta_j, q_j) = \sum_{j=1}^{k} q_j (1-\theta_j)^m$$

Special Mixtures

Mixtures of Poissons

$$f(y, \theta_j, q_j) = \Sigma_{j=1}^{k} \; q_j \exp(-\theta_j) \, \theta_j^{y} \, / y \, ! \; , \; y=0,1, \ldots$$

(θ_j = Listing parameter in sub-population j,
q_j = weight of sub-population)

Predicted Probability of a Zero:

$$p_0 = f(y=0, \theta_j, q_j) = \Sigma_{j=1}^{k} \; q_j \exp(-\theta_j)$$

Results of Analysis for Cancer Registry of Saarland

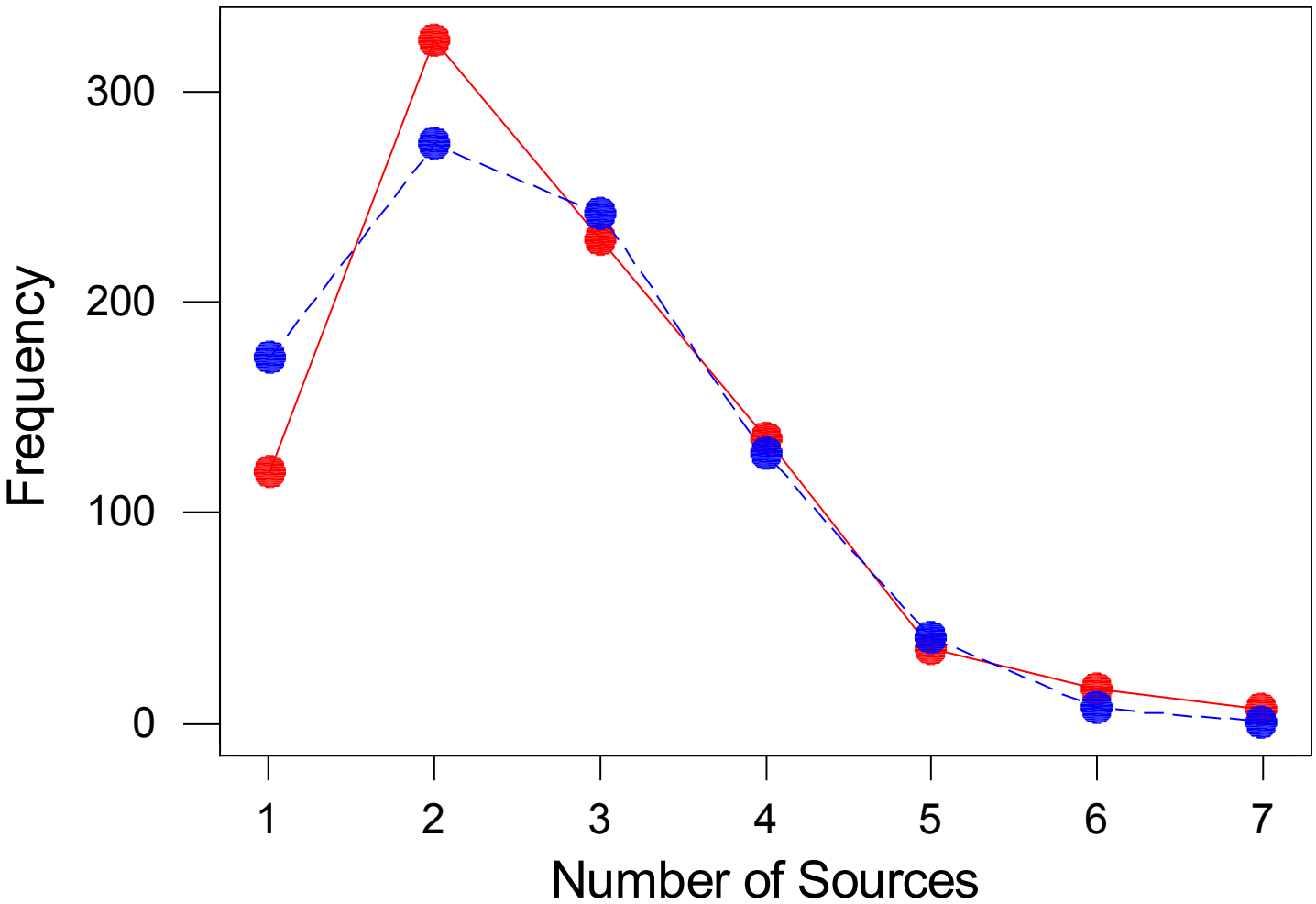
- Joint project with Robert Koch Institute, Berlin, Dachorganisation Krebs (Dr. Dieter Schön)
- Six main sources, and subsidiar sources which occur from 40 hospital categories and 31 departmental categories
- counting sources variable seldom > 10
- years considered: 1994 - 1998
- here 3 sites: lung cancer, female breast cancer and prostata cancer

Lung Cancer

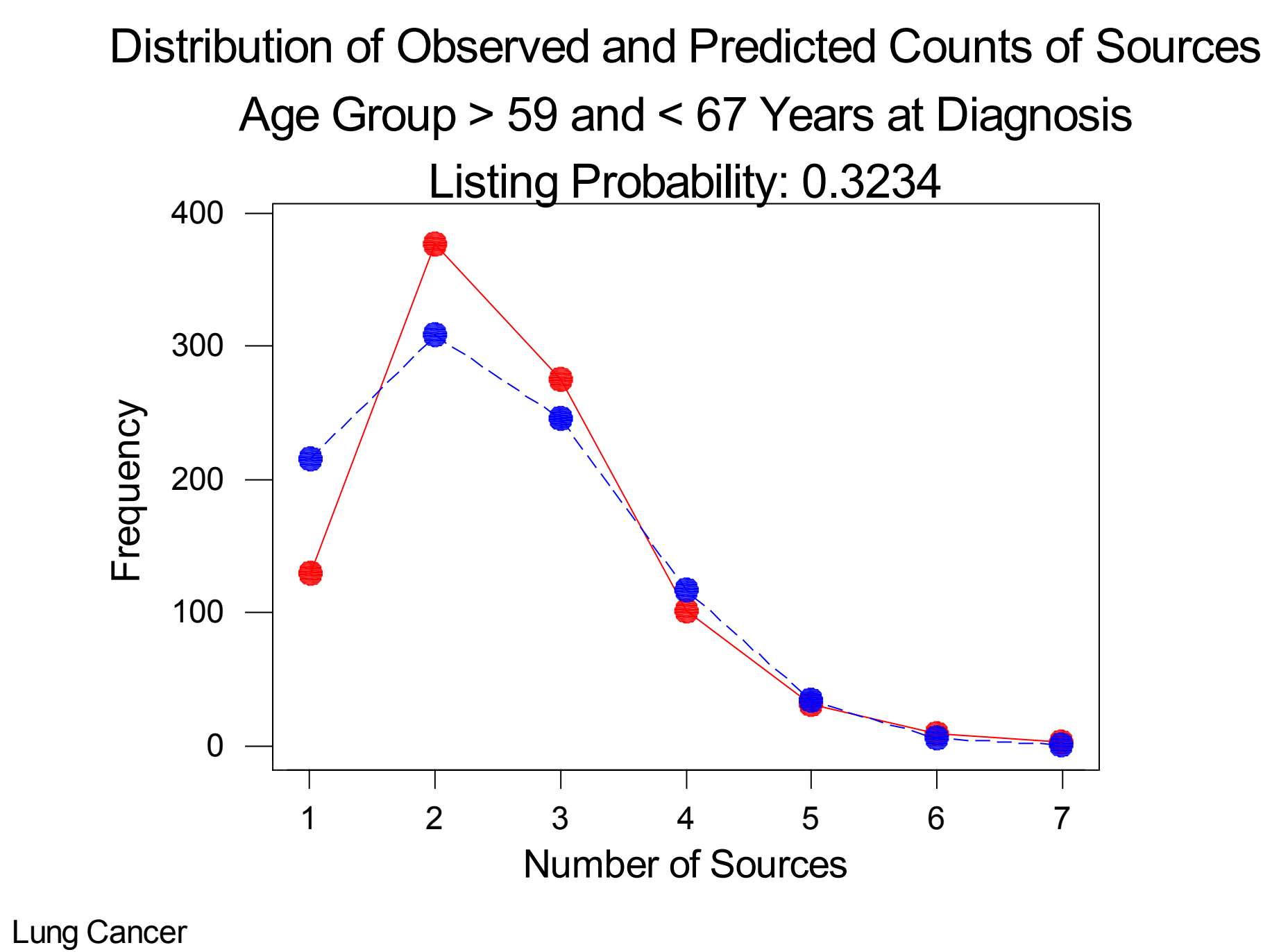
(using mixtures of binomials)

Distribution of Observed and Predicted Counts of Sources

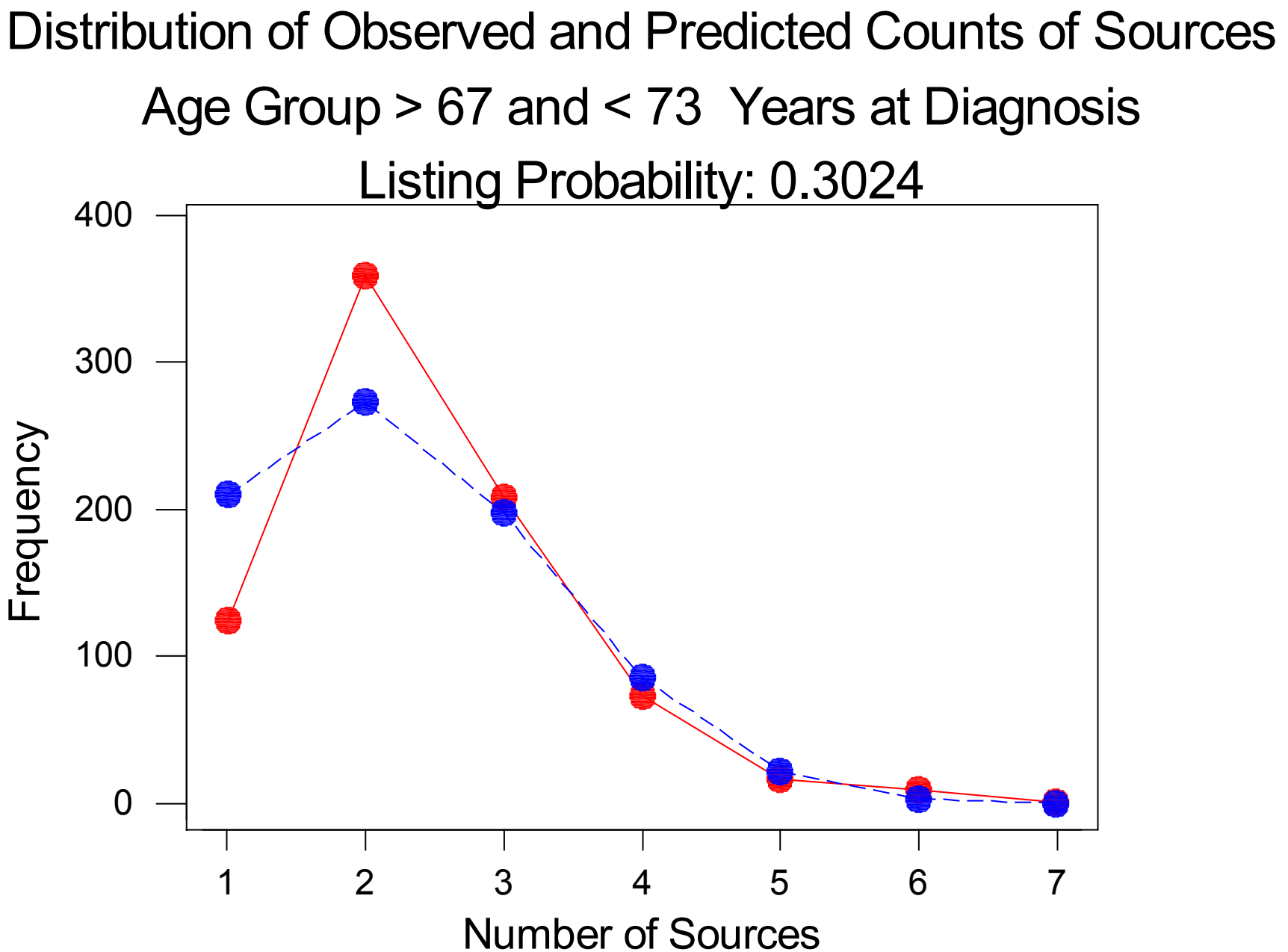
Age Group < 59 Years at Diagnosis
Listing Probability: 0.3684



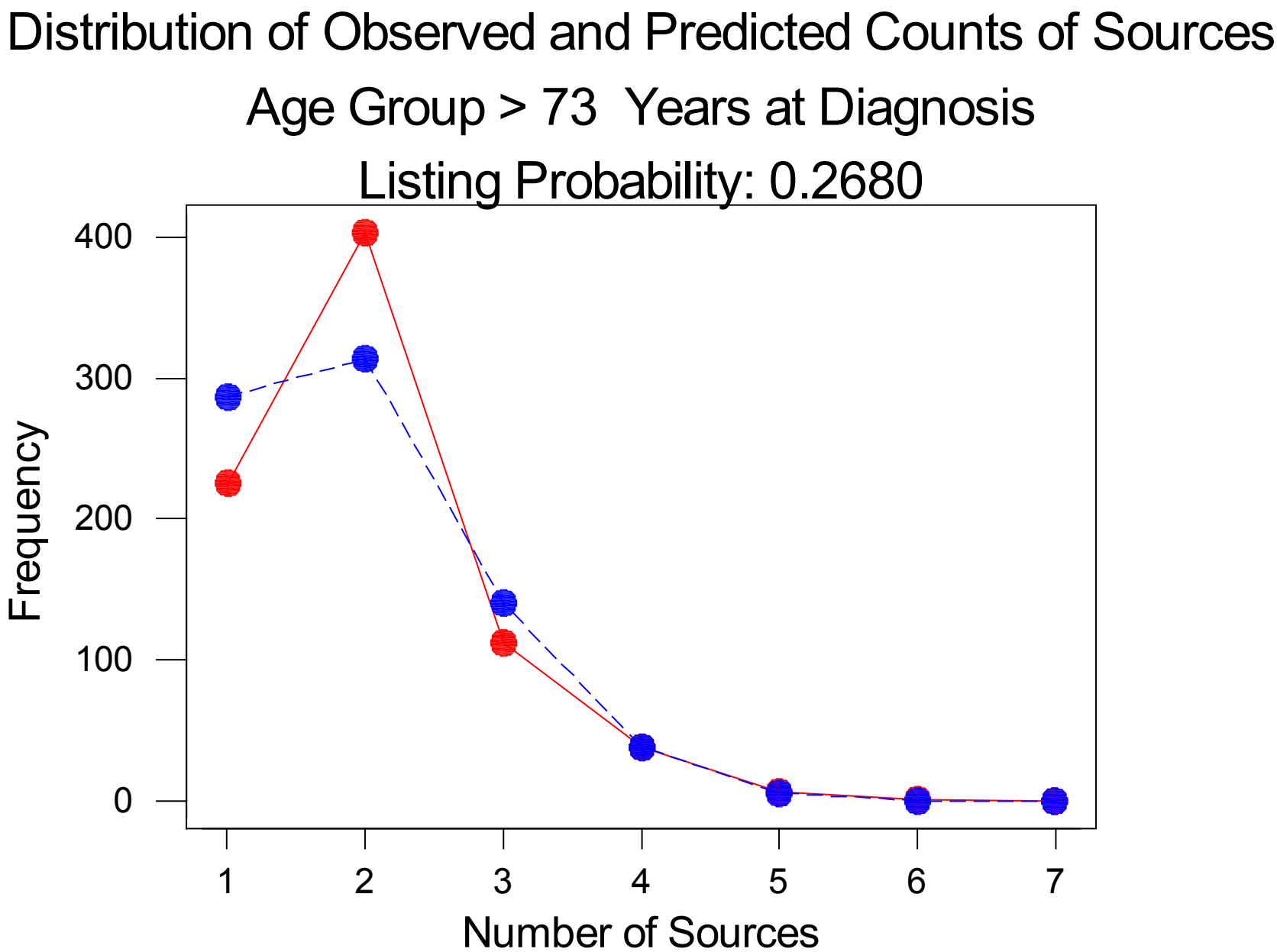
Lung Cancer


Distribution of Observed and Predicted Counts of Sources
Age Group > 59 and < 67 Years at Diagnosis
Listing Probability: 0.3234
Frequency
400
300
200
100
0
1
2
3
4
5
6
7
Number of Sources

Lung Cancer


Distribution of Observed and Predicted Counts of Sources
Age Group > 67 and < 73 Years at Diagnosis
Listing Probability: 0.3024
Frequency
400
300
200
100
0
1
2
3
4
5
6
7
Number of Sources

Lung Cancer

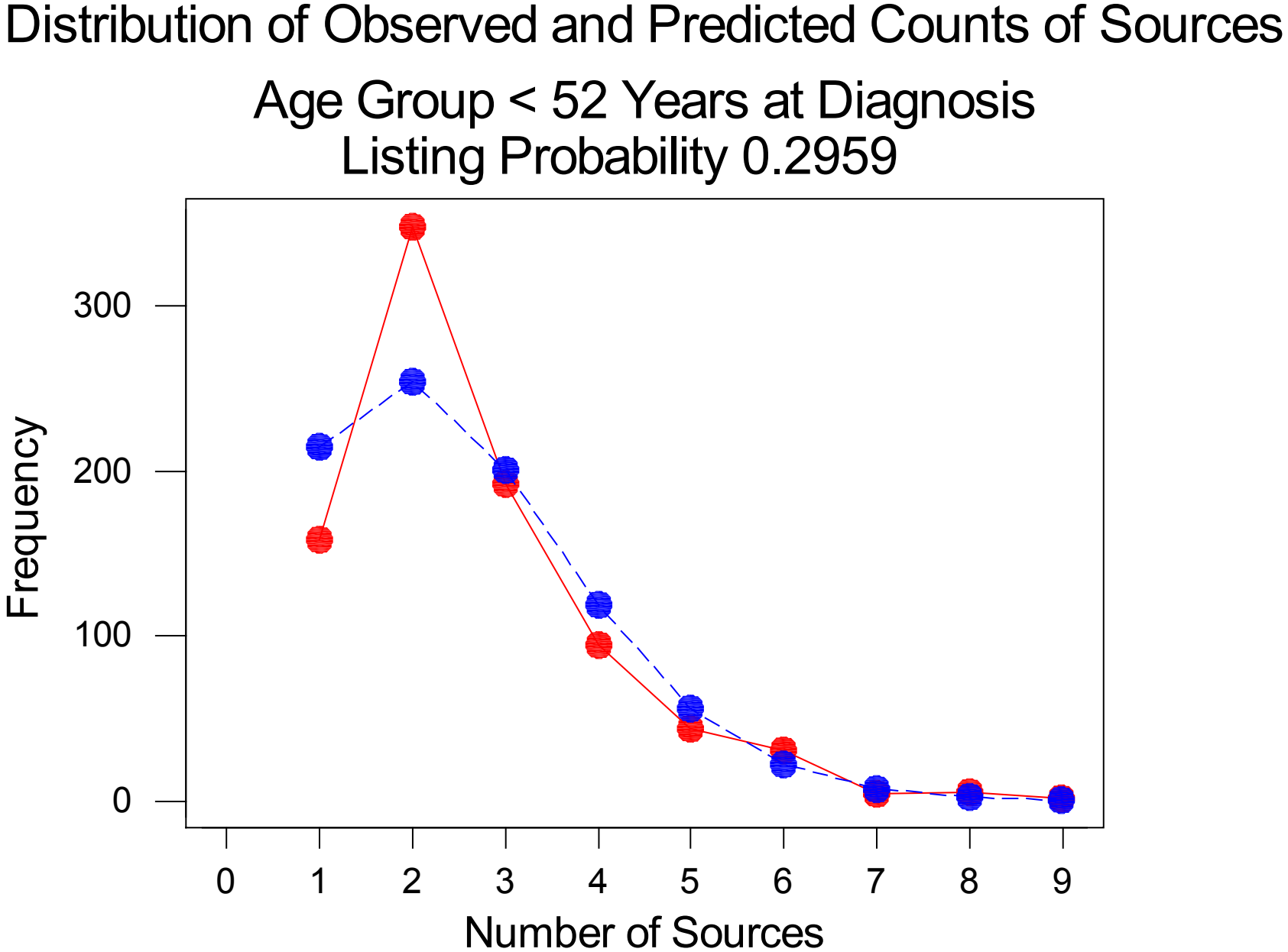

Distribution of Observed and Predicted Counts of Sources
Age Group > 73 Years at Diagnosis
Listing Probability: 0.2680
Frequency
400
300
200
100
0
1
2
3
4
5
6
7
Number of Sources

Lung Cancer

Summary Table for Lung Cancer

Age group	n_{obs}	n_o	Listing Probability	Complete-ness (%)
- 59	866	47	0.3684	94.8521
- 67	926	64	0.3234	93.5354
- 73	792	67	0.3057	92.2002
> 73	785	143	0.2680	84.5905
Total	**3369**	**321**	**0.3125**	**91.3008**

Breast Cancer
(using mixtures of Poissons)

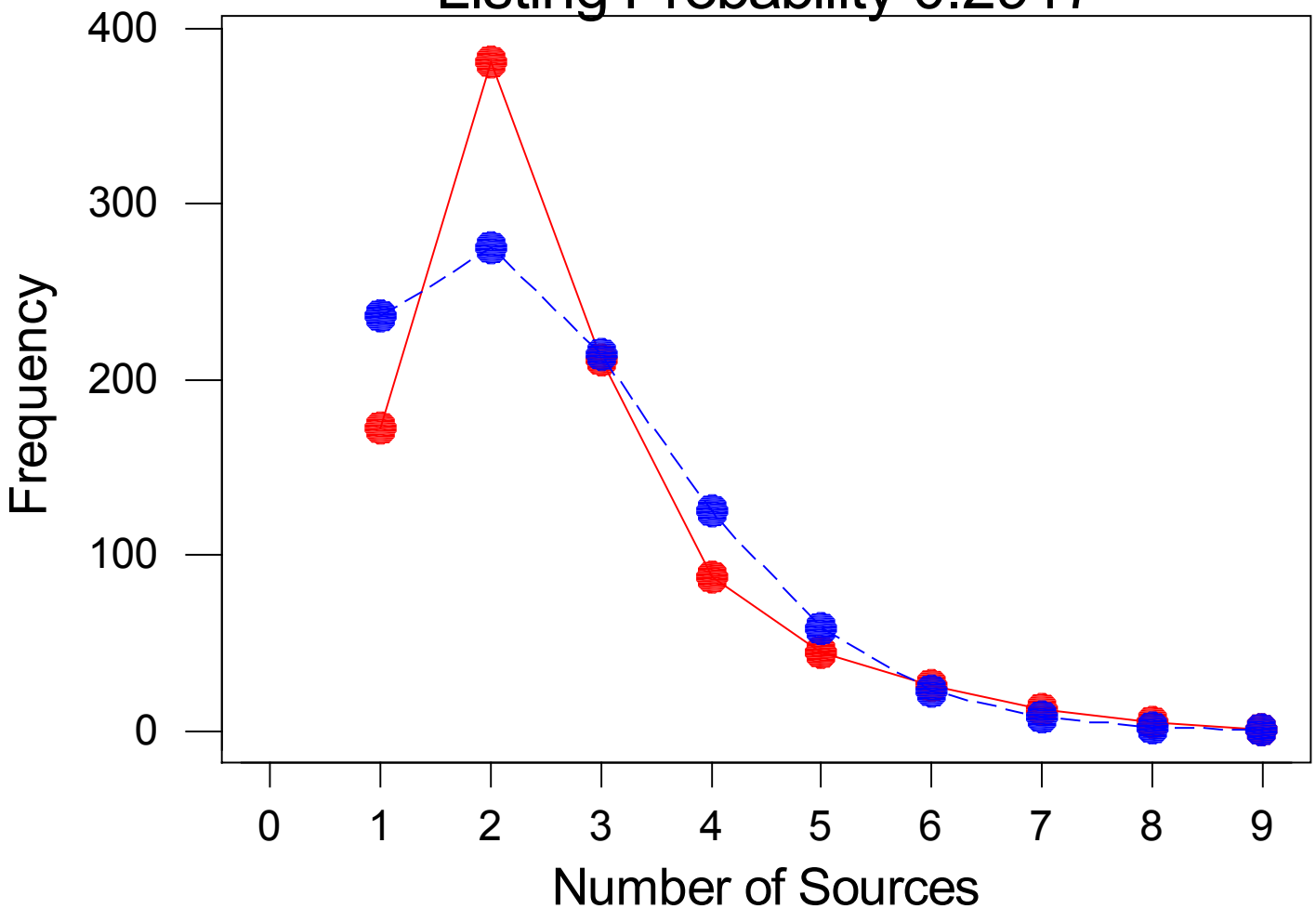


Breast Cancer

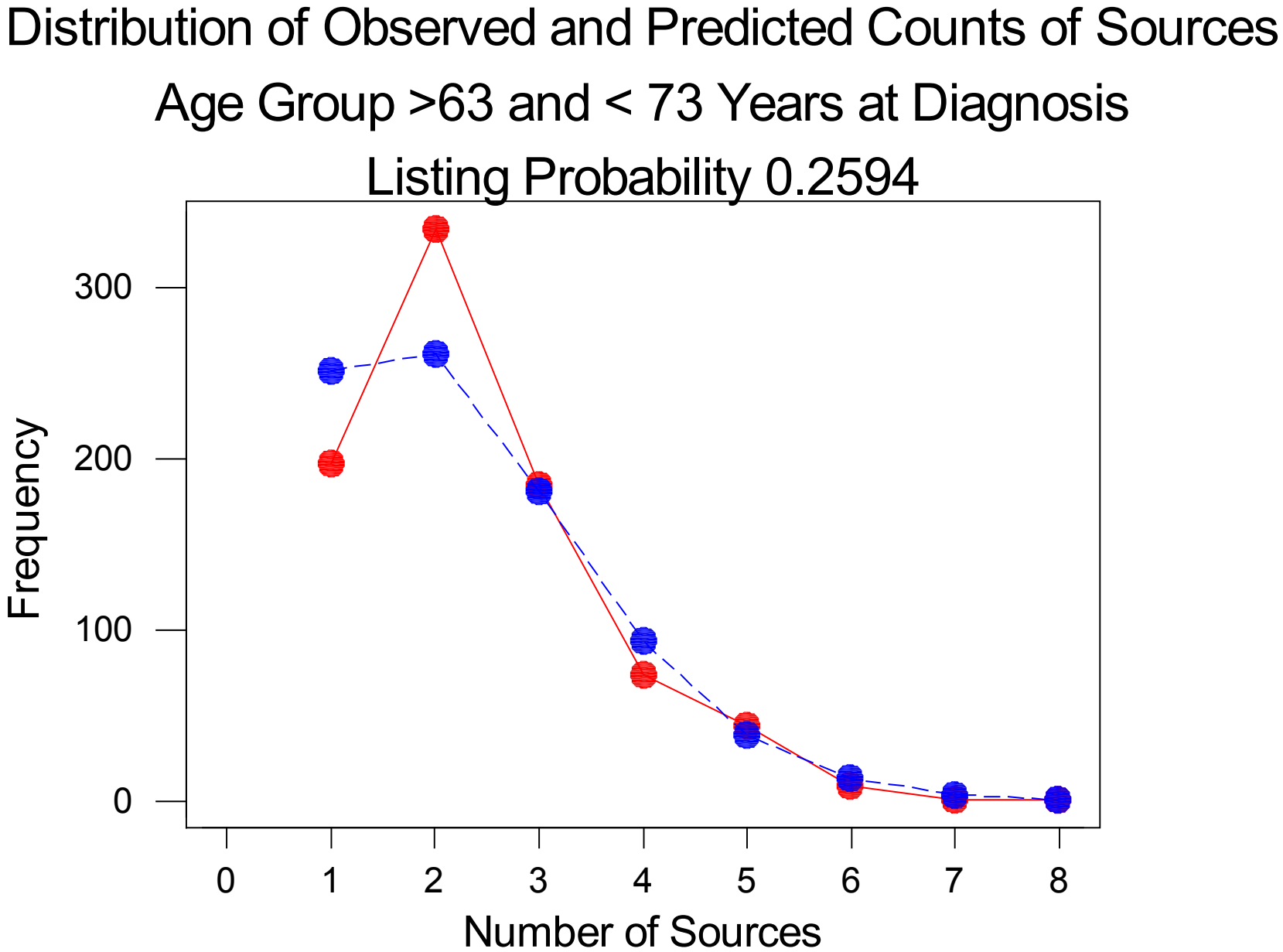
Distribution of Observed and Predicted Counts of Sources

Age Group >52 and < 63 Years at Diagnosis

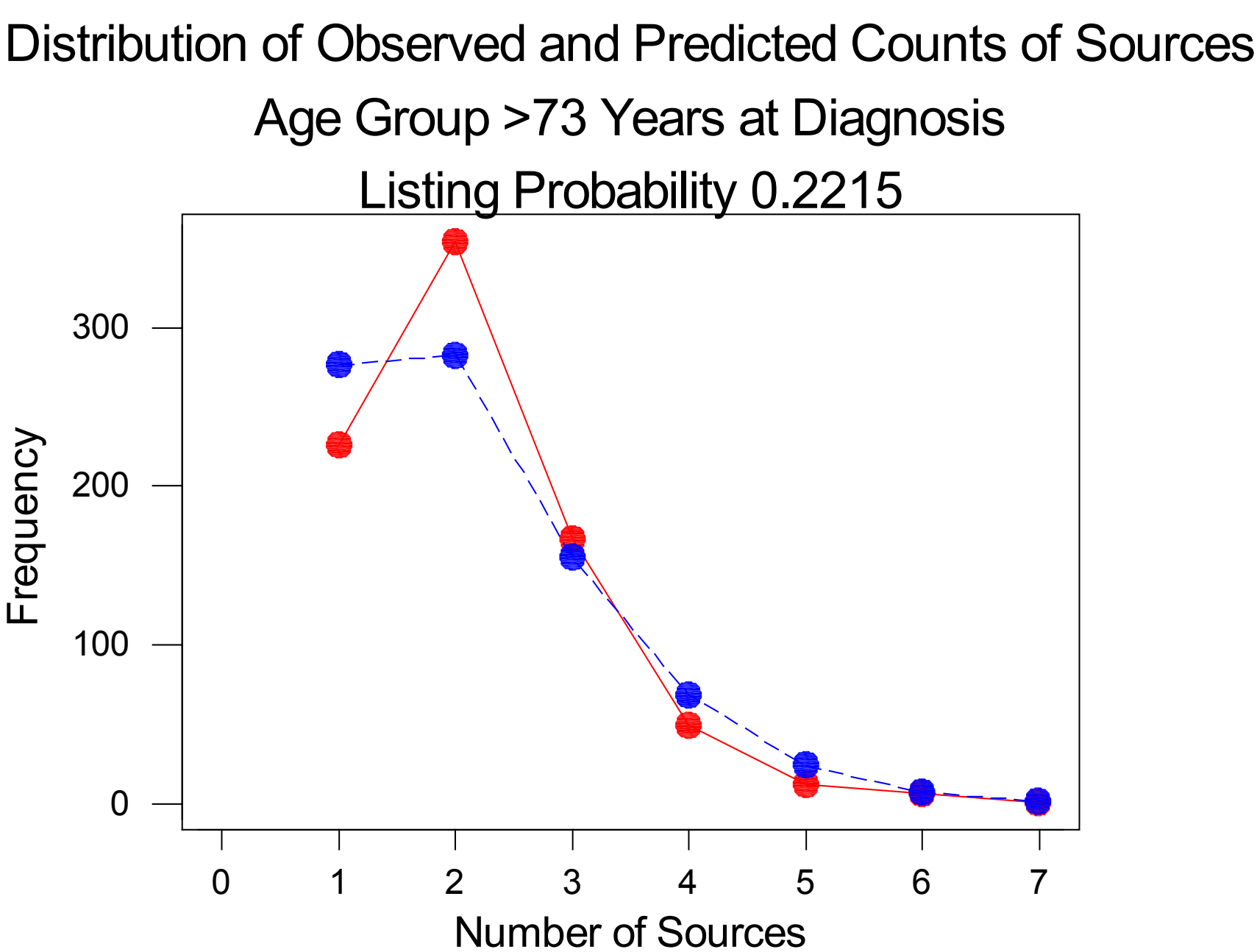
Listing Probability 0.2917



Breast Cancer



Breast Cancer

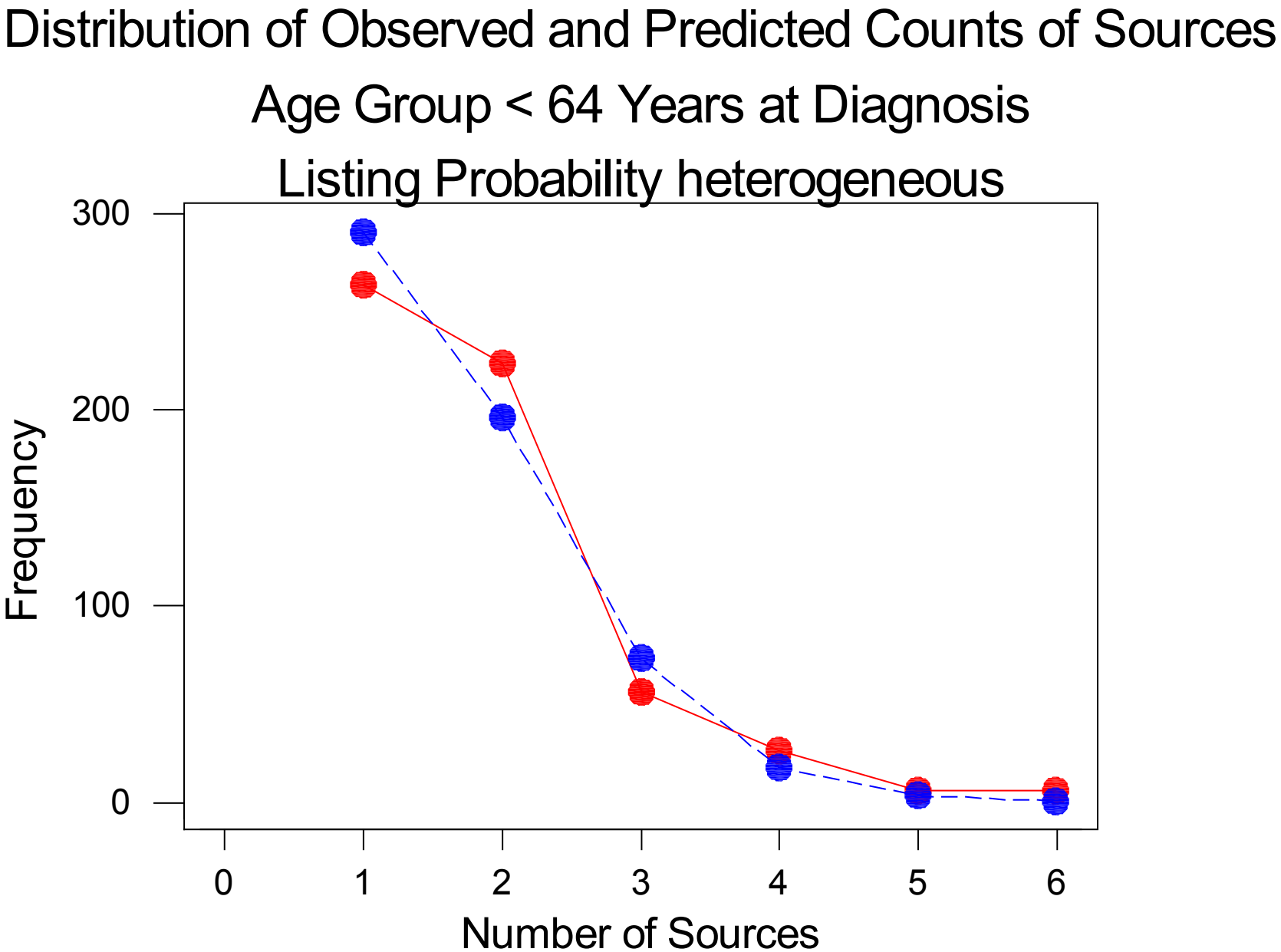


Breast Cancer

Summary Table for Breast Cancer

Age group	n_{obs}	n_o	Listing Probability	Complete-ness (%)
- 52	877	91	0.2959	90.5992
- 63	942	101	0.2917	90.3164
- 73	845	121	0.2594	87.4741
> 73	817	143	0.2215	85.1042
Total	**3481**	**456**	**0.2687**	88.4176

Prostate Cancer
(using mixtures of Binomials)

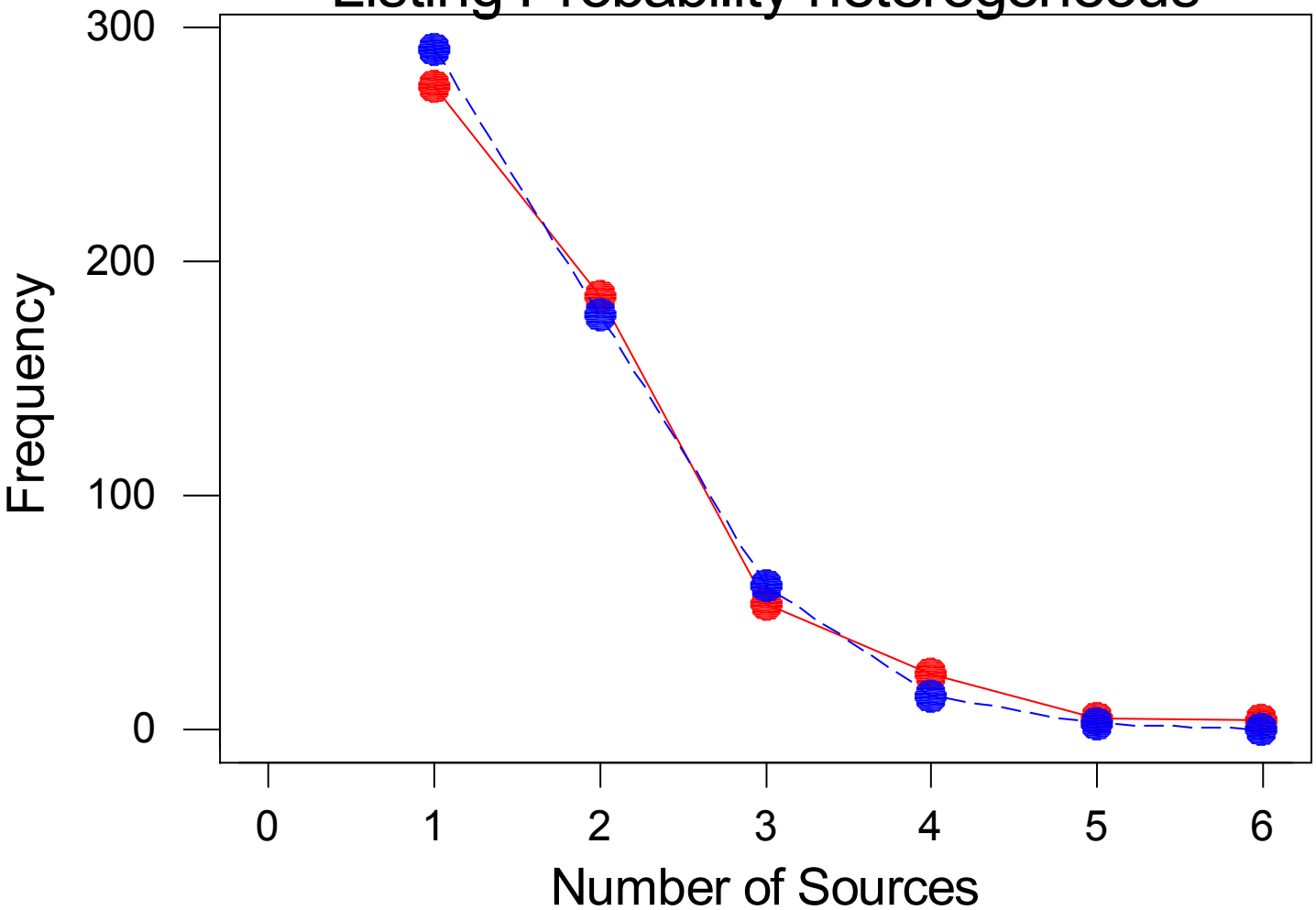


Prostate Cancer

Distribution of Observed and Predicted Counts of Sources

Age Group > 64 and < 70 Years at Diagnosis

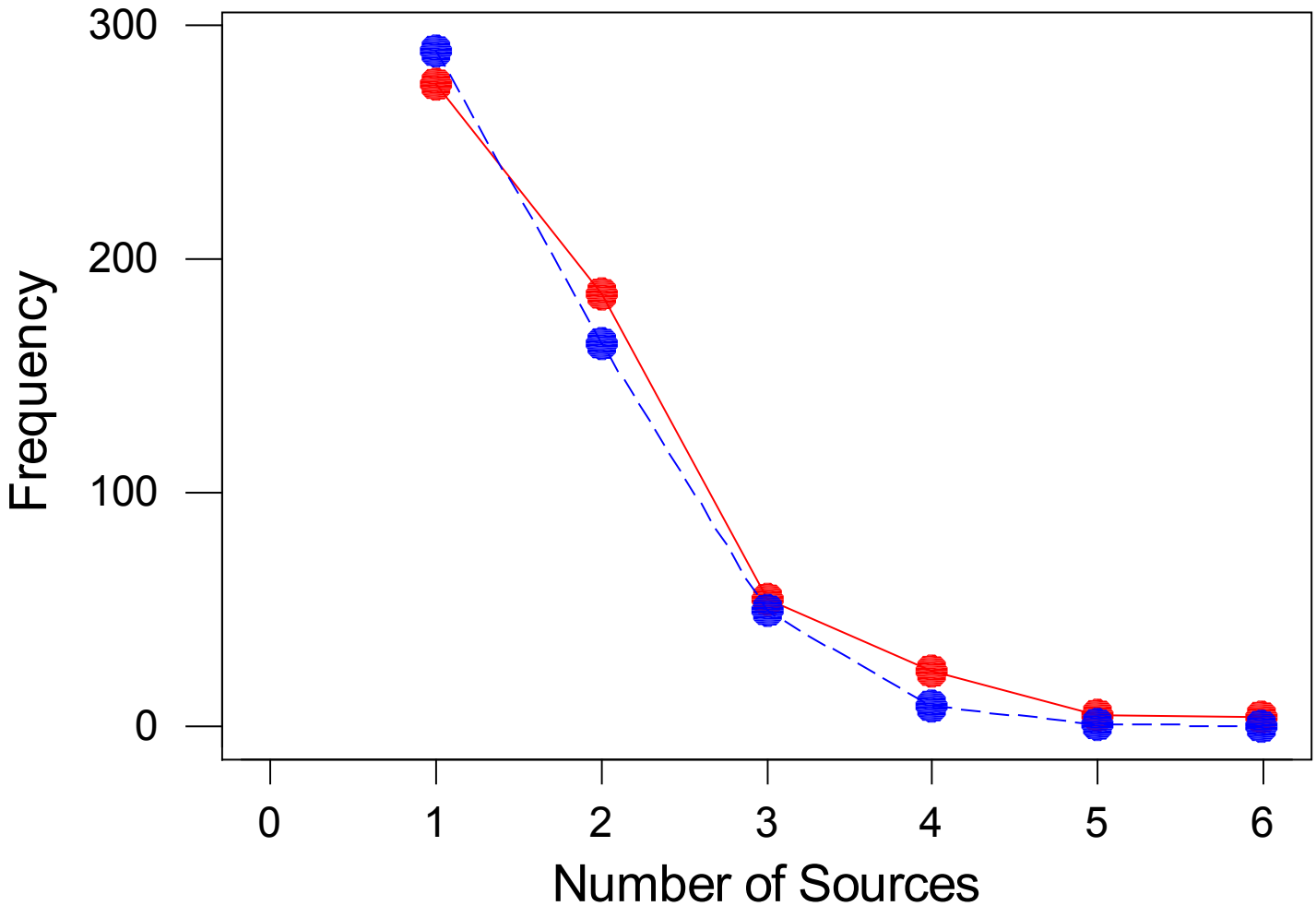
Listing Probability heterogeneous



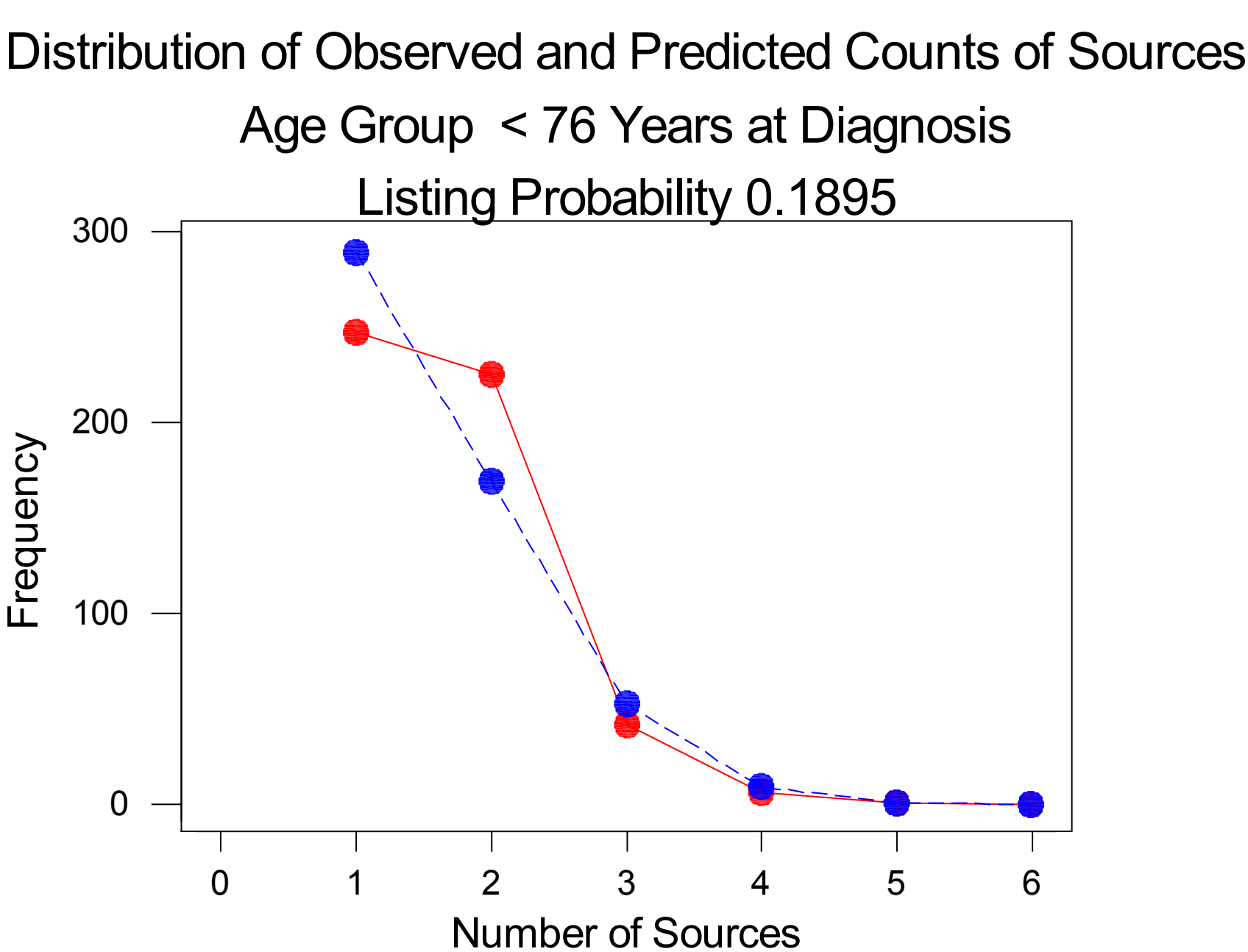
Prostate Cancer

Distribution of Observed and Predicted Counts of Sources

Age Group > 70 and < 76 Years at Diagnosis
Listing Probability 0.1848



Prostate Cancer



Prostate Cancer

Summary Table for Prostate Cancer

Age group	n_{obs}	n_o	LP 1.	LP 2.	Weight 2. Comp.	Complete-ness (%)
- 64	582	181	0.2111	0.5534	0.0184	76.2779
- 70	547	203	0.1910	0.4525	0.0368	72.9333
- 76	512	213	0.1848	-	1.0000	70.6207
> 76	521	206	0.1895	-	1.0000	71.6644
Total	**2162**	**803**	**0.2010**	**-**	**1.0000**	**72.9174**

Thank You!