Analysis of Dependent Contingency Table

CF Jeff Lin, MD., PhD.

December 17, 2005

Dependent Contingency Table

Dependent Contingency Table
Matched-Pair Data: McNemar's Test

Example: Treatments for Arthritis with Crossover Design

1. An investigator conduct a study to compare pain relief effects of two different treatments for arthritis.
2. The two treatment groups should be as comparable as possible on other prognostic factors, i.e.; age and clinical conditions.

Example: Treatments for Arthritis with Crossover Design

3. To accomplish this goal, a **matched study** is set up such that a random member of each matched pair get treatment A in the first week (period) then that member get treatment B in the second week (period), whereas the other member gets treatment B in the first week (period), then that member get treatment a in the second week (period).
4. The outcome is whether pain exists or not in the end of each week.
5. The data are displaced in a 2×2 table as shown in Table 1.

Example: Treatments for Arthritis with Crossover Design

Table 1: Treatments for Arthritis with Crossover Design: Independent 2×2 Table

	Outcome		
Treatment	No Pain	Pain	Total
A	146	54	200
B	134	66	200
Total	280	120	400

Example: Treatments for Arthritis with Crossover Design

1. There is a small difference in pain relief between two treatment groups.
2. Pain relief in treatment A group is $146/200 = 0.73$,
3. pain relief in treatment B group is $134/200 = 0.67$.
4. The Yates-corrected chi-square statistic, X^2, is 1.71 with χ^2_1, which is not significant.

Program

```
> # Independent Table
> OA.ind.tab<-matrix(c(146,54,134,66),nrow=2,byrow=T)
> OA.ind.tab
     [,1] [,2]
[1,]  146   54
[2,]  134   66
```


Program

```
> # Independent Table
> chisq.test(OA.ind.tab,correct=F)

        Pearson's Chi-squared test

data:  OA.ind.tab
X-squared = 1.7143, df = 1, p-value = 0.1904
```


Example: Treatments for Arthritis with Crossover Design

1. However, the use of this test is valid only if the two sample are independent.
2. From the manner in which the samples were selected it is obvious that they are not independent, because members of each matched are similar in age and clinical condition.
3. Thus, chi-square test or Fisher's exact test cannot be used in this situation.
4. How the can the two treatments be compared using a hypothesis test?

Example: Treatments for Arthritis with Crossover Design

1. A different type of 2×2 table arise in the context of **matched pair data**.
2. Observations are collected in pairs where the members of each pair are identical or nearly identical for a particular variable that interferes with assessing a specific relationship, as in Table 2.
3. Frequently this particular variable is called a "confounder". The matched pair is the unit of analysis and pairs are classified according to whether nor not the members of that pair had pain relief at each end of week.

Example: Treatments for Arthritis with Crossover Design

Table 2: Treatments for Arthritis with Crossover Design: Dependent 2×2 Table

	Treatment B		
Treatment A	No Pain	Pain	Total
No Pain	130	16	146
Pain	4	50	54
Total	134	66	200

Example: Treatments for Arthritis with Crossover Design

1. The above table has 200 units rather than the 400 units.
2. Furthermore, there are 130 pairs in which both subjects had no pain relief.
3. Sixteen pairs in which the treatment A had pain relief and the treatment B had pain.
4. Four pairs in which the treatment A had pain and the treatment B had pain relief.

Example: Treatments for Arthritis with Crossover Design

1. The dependence of the two samples can be illustrated by noting that

$$\begin{aligned} & p[\text{Treatment B with pain relief} \mid \text{Treatment A with pain relief}] \\ = \; & 130/146 = 0.89, \\ & p[\text{Treatment B with pain relief} \mid \text{Treatment A with pain}] \\ = \; & 4/50 = 0.08. \end{aligned}$$

2. If the samples were independent, then these two probabilities should be about the same, thus we conclude that the samples are highly dependent and that the chi-square test can not be used.

Dependent Contingency Table: Matched-Pair Data

Dependent 2×2 Table with Matched-Pair Data

Consider a dependent 2×2 table of matched-pair data as in Table 3.

Table 3: McNemar's Test: Dependent 2×2 Table

Observed a Table	Pair member B		
Pair member A	Outcome 1	Outcome 2	Total
Outcome 1	a	b	$a+b=n_{1.}$ (row 1 margin)
Outcome 2	c	d	$c+d=n_{2.}$ (row 2 margin)
Total	$a+c=n_{.1}$	$b+d=n_{.2}$	$a+b+c+d=n_{..}=n$
	column 1 margin	column 2 margin	(grand total)

Dependent 2×2 Table with Matched-Pair Data

1. A **concordant pair**, (i.e., $a+d$), is a matched in which the outcomes is the same for each member of the pair.
2. A **discordant pair**, (i.e., $b+c$), is a matched pair in which the outcomes are different for the members of the pair.

Dependent 2×2 Table with Matched-Pair Data

1. In Table 2, for 180 concordant pairs $(130+50)$, the outcomes of two treatments are the same, whereas for 20 discordant pairs $(16+4)$, the outcomes of the two treatments are different.
2. The concordant pairs provide no information about differences between treatments and will not be used in the assessment.
3. Instead, we will focus on the discordant pairs, which can be divided into two types: b and c. If each outcome of pairs are equal, then about an equal number of b and c.

McNemar's Test: Testing Hypothesis

1. Let p be the probability that a discordant pair is an element of b of the discordant pairs.
2. Thus we wish to test the hypothesis

$$H_0 : p = \frac{1}{2} \text{ versus } H_A : p \neq \frac{1}{2} \tag{1}$$

McNemar's Test: Point Estimation

Under null hypothesis H_0, the $\mathcal{E}[b]$ and $\mathbf{Var}[b]$ are

$$\mathcal{E}[b] = \frac{b+c}{2}. \tag{2}$$

$$\mathbf{Var}[b] = \frac{b+c}{4}. \tag{3}$$

McNemar's Test: Test Statistics

$$H_0 : \quad p = \frac{1}{2} \text{ versus } H_A : p \neq \frac{1}{2} \tag{4}$$

$$\mathcal{E}(b) = \frac{b+c}{2} = \frac{\#\text{ discordant pairs}}{2} \tag{5}$$

$$\mathbf{Var}(b) = \frac{b+c}{4} \tag{6}$$

$$X^2_{\text{mc}} = \frac{\left(\mid b - \frac{b+c}{2} \mid \right)^2}{\frac{b+c}{4}} \tag{7}$$

$$X^2_{\text{mc}} = \frac{(\mid b-c \mid)^2}{b+c} \text{ asym } \sim \chi^2_1. \tag{8}$$

McNemar's Test: Test Statistics

1. That is, X^2 is asymptotically approximated chi-squared distribtuted with with 1 degree of freedom.
2. The p-value is calculates as

$$p - \text{value } = P(\chi^2_1 \geq X^2). \tag{9}$$

3. For a two-sided test with significant level α, we reject H_0 if $X^2 > \chi^2_{1,1-\alpha}$.
4. We will assume that the normal approximation to the binomial distribution holds, but this assumption will valid if $npq = (b+c)/4 \geq 5$ or $(b+c) \geq 20$.

McNemar's Test: Test Statistics

1. For discrete binomial correction, we sometime use

$$X^2_{\text{mc}^\star} = \frac{\left(\mid b - \frac{b+c}{2} \mid - \frac{1}{2} \right)^2}{\frac{b+c}{4}} \tag{10}$$

$$X^2_{\text{mc}^\star} = \frac{(\mid b-c \mid -1)^2}{b+c} \text{ asym } \sim \chi^2_1. \tag{11}$$

2. That is, X^2_{mc} is asymptotically approximated chi-squared distribtuted with with 1 degree of freedom.

Difference and Odds Ratio of Proportions of Paired Samples

Difference of Proportions of Paired Samples

1. Let π_{1+} be the outcome 1 (as success) in pair member A, and let π_{+1} be the outcome 1 in pair member B.
2. For binary responses, if each outcome of matched pairs are equal, the null hypothesis is $H_0 : \pi_{1+} = \pi_{+1}$.
3. The inference for the dependent proportions is

Difference of Proportions of Paired Samples

$$d = \pi_{1+} - \pi_{+1}, \tag{12}$$

$$\mathbf{Var}(d) = \sigma_d^2 = \frac{[\pi_{1+}(1-\pi_{1+}) + \pi_{+1}(1-\pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})]}{n} \tag{13}$$

Difference of Proportions of Paired Samples

$$\hat{d} = \hat{\pi}_{1+} - \hat{\pi}_{+1}, \tag{14}$$

$$\widehat{\mathbf{Var}}(\hat{d}) = \hat{\sigma}_{\hat{d}}^2 = \frac{[\hat{\pi}_{1+}(1-\hat{\pi}_{1+}) + \hat{\pi}_{+1}(1-\hat{\pi}_{+1}) - 2(\hat{\pi}_{11}\hat{\pi}_{22} - \hat{\pi}_{12}\hat{\pi}_{21})]}{n}. \tag{15}$$

$$\approx \frac{\left[(\hat{\pi}_{12} + \hat{\pi}_{21}) - (\hat{\pi}_{12} - \hat{\pi}_{21})^2\right]}{n} \tag{16}$$

Thus, $(1-\alpha) \times 100\%$ C.I. for $\hat{d}$: $\hat{d} \pm Z_{1-\alpha/2} \times \sqrt{\hat{\sigma}^2(\hat{d})}$ (17)

Difference of Proportions of Paired Samples

1. Under H_0, an simple alternative estimated variance is

$$\hat{\sigma}^2(d) \approx \frac{\hat{\pi}_{1+} + \hat{\pi}_{+1} - 2\hat{\pi}_{11}}{n} = \frac{b+c}{n^2} \tag{18}$$

2. The score test statistic

$$\text{z} = \frac{\hat{d}}{\hat{\sigma}(d)}, \tag{19}$$

$$\text{simplified to} \quad \text{z} = \frac{b-c}{(b+c)^{1/2}} \tag{20}$$

Odds Ratio of Proportio of Paired Samples

1. The **odd ratio** comparing the odds of outcome 1 (success) at treatment B to treatment A is estimated by:

$$\widehat{\text{OR}} = \frac{c}{b} \tag{21}$$

2. Confidence intervals can be obtained as described in Breslow and Day (1981), section 5.2, or in Armitage and Berry (1987), chapter 16.

Example: Treatments for Arthritis with Crossover Design

Table 4: Treatments for Arthritis with Crossover Design: Dependent 2×2 Table

	Treatment B		
Treatment A	No Pain	Pain	Total
No Pain	130	16	146
Pain	4	50	54
Total	134	66	200

Example: Treatments for Arthritis with McNear's Test

1. In Table 4, McNemar's Test showed $X^2_{\text{mc}} = 7.2$, and p-value is 0.0073.
2. The marginal proportion of pain relief for treatment A is 0.73 and for treatment B is 0.67, the difference between two proportion $(B - A)$ is -0.06.
3. The 95% C.I.for the difference is $(-0.103, -0.0170)$, treatment A is more effective than treatment B.

Treatments for Arthritis with McNear's Test: Program

```
> # Dependent Table
> (OA.tab<-matrix(c(130,16,4,50),nrow=2,byrow=T))
     [,1] [,2]
[1,]  130   16
[2,]    4   50
```


Treatments for Arthritis with McNear's Test: Program

```
> mcnemar.test(OA.tab, correct=FALSE)

        McNemar's Chi-squared test

data:  OA.tab
McNemar's chi-squared = 7.2, df = 1, p-value = 0.00729
```


Treatments for Arthritis with McNear's Test: Program

```
> mcnemar.test(OA.tab, correct=TRUE)

        McNemar's Chi-squared test with continuity correction

data:  OA.tab
McNemar's chi-squared = 6.05, df = 1, p-value = 0.01391
```


Treatments for Arthritis with McNear's Test: Program

```
> binom.test(OA.tab[1,2], (OA.tab[1,2]+OA.tab[2,1]), p=0.5)

        Exact binomial test
data:  OA.tab[1, 2] and (OA.tab[1, 2] + OA.tab[2, 1])
number of successes = 16, number of trials = 20, p-value = 0.01182
alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:
 0.563386 0.942666
sample estimates:
probability of success
                   0.8
```


Treatments for Arthritis with McNear's Test: Program

```
> marginal prop diff
> (OA.tab.prop<-prop.table(OA.tab))
     [,1] [,2]
[1,] 0.65 0.08
[2,] 0.02 0.25
> margin.table(OA.tab.prop,2)[1]
[1] 0.67
> margin.table(OA.tab.prop,1)[1]
[1] 0.73
```

Treatments for Arthritis with McNear's Test: Program

```
> prop.diff<-(margin.table(OA.tab.prop,2)[1]
    -margin.table(OA.tab.prop,1)[1])
> prop.diff
[1] -0.06
>
> # EQ 16
> (prop.diff+c(-1,1)*qnorm(0.975)*
     sqrt((sum(off.diag)-diff(off.diag)^2)/sum(OA.tab)))
[1] -0.10303003 -0.01696997
```


Treatments for Arthritis with McNear's Test: Program

```
> (off.diag<-diag(OA.tab.prop[1:2,2:1]))
[1] 0.08 0.02
> # C.I.
> sum(off.diag)
[1] 0.1
> diff(off.diag)
[1] -0.06
> sum(OA.tab)
[1] 200
> # EQ 16
> (prop.diff+c(-1,1)*qnorm(0.975)*
     sqrt((sum(off.diag)-diff(off.diag)^2)/sum(OA.tab)))
[1] -0.10303003 -0.01696997
```


Treatments for Arthritis with McNear's Test: SAS

```
title "McNemar Test: Matched-Paired Data";
data mcne2 ;
input A B count ;
cards;
1 1 130
1 0 16
0 1 4
0 0 50
run;
```


Treatments for Arthritis with McNear's Test: SAS

```
proc freq data=mcne2 order=data page;
    tables A*B / agree;
    weight count;
    exact agree;
run;
```


Treatments for Arthritis with McNear's Test: SAS

A	B		
Frequency			
Percent			
Row Pct			
Col Pct	1	0	Total
1	130	16	146
	65.00	8.00	73.00
	89.04	10.96	
	97.01	24.24	
0	4	50	54
	2.00	25.00	27.00
	7.41	92.59	
	2.99	75.76	
Total	134	66	200
	67.00	33.00	100.00

Treatments for Arthritis with McNear's Test: SAS

McNemar's Test

Statistic (S)	7.2000
DF	1
Asymptotic Pr > S	0.0073
Exact Pr >=S	0.0118

Reliability Studies with Kappa Statistic

Two Surveys with Same Diet Questionnaire

1. A diet questionnaire was administered by mail to 537 female American nurses on two separate occasions several months apart, Rosner (2000).
2. The questions asked included quantities eaten of more than 100 separate food items. The data obtained from the two surveys for the amount of beef consumption are presented in Table 5.
3. How can the reproducibility of response for the beef consumption data be quantified?

Two Surveys with Same Diet Questionnaire

Table 5: 2×2 table of paired nutrition data

	The Second Survey		
The First Survey	$1 <=$ serving/week	> 1 serving/week	Total
$1 <=$ serving/week	136	92	228
> 1 serving /week	69	240	309
Total	205	332	537

Two Surveys with Same Diet Questionnaire

A chi-square test for association between the survey 1 and survey 2 responses could be performed. However, this test would not give a quantitative measure of reproducibility. There are $(136 + 240)/537 = 70.7\%$ concordant responses.

Reliability Studies with Kappa Statistic

How can the reproducibility of response be quantified in a dependent 2×2 table?

Table 6: Kappa: reliability of 2×2 table

	The Second Survey		
The First Survey	$1 <=$ serving/week	> 1 serving/week	Total
$1 <=$ serving/week	$n_{11} = a$	$n_{12} = b$	$n_{1.}$
> 1 serving /week	$n_{21} = c$	$n_{22} = d$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Reliability Studies with Kappa Statistic

1. We would like to compare the observed concordance rate $\mathbf{\Pi}_O$ and the expected concordance $\mathbf{\Pi}_E$ whether the responses of the subjects were statistically independent.
2. The motivation behind this definition is that the questionnaire would be virtually worthless if the frequency of consumption reported at one survey had not relationship to the frequency of consumption reported at a second survey.

Reliability Studies with Kappa Statistic

1. The perfect **agreement** corresponds to $\mathbf{\Pi}_O = 1$. Thus $\mathbf{\Pi}_O - \mathbf{\Pi}_E$ is the excess of the observers agreement.
2. We could use $\mathbf{\Pi}_O - \mathbf{\Pi}_E$ as the measure of reproducibility.
3. However, it is preferable to use a measure that equals to 1 in the case of perfect agreement and 0 if the responses on the two surveys are completely independent.
4. The maximum possible value for $\mathbf{\Pi}_O - \mathbf{\Pi}_E$ is $1 - \mathbf{\Pi}_E$, which is achieved with $\mathbf{\Pi}_O = 1$.

Reliability Studies with Kappa Statistic

5. Therefore, the $\boldsymbol{\kappa}$ statistic, which is defined as $(\mathbf{\Pi}_O - \mathbf{\Pi}_E)/(1 - \mathbf{\Pi}_E)$, is used as the measure of reproducibility.

Reliability Studies with Kappa Statistic

1. Supposes there k response categories and the probability of response in the i^{th} is π_{i+} for the first survey and π_{+i} for the second survey.
2. These probabilities can be estimated from the row and column margins of the contingency table.
3. The expected concordance rate $\mathbf{\Pi}_E$ if the survey responses are independent is given

$$\boldsymbol{\kappa} = \frac{\mathbf{\Pi}_O - \mathbf{\Pi}_E}{1 - \mathbf{\Pi}_E}, \tag{22}$$

$$\mathbf{\Pi}_O = \sum_{i=1}^{k} \frac{n_{ii}}{n}, \tag{23}$$

Reliability Studies with Kappa Statistic

Under H_0:

$$\pi_{i+} = \frac{n_{i.}}{n}, ; \tag{24}$$

$$\pi_{+i} = \frac{n_{.i}}{n}, ; \tag{25}$$

$$\pi_{ii} = \pi_{i+}\pi_{+i}, ; \tag{26}$$

$$\mathbf{\Pi}_E = \sum_{i=1}^{k} \pi_{ii} = \sum_{i=1}^{k} \pi_{i+}\pi_{+i}, . \tag{27}$$

Reliability Studies with Kappa Statistic

The Kappa statistic is

$$\boldsymbol{\kappa} = \frac{\mathbf{\Pi}_O - \mathbf{\Pi}_E}{1 - \mathbf{\Pi}_E} \tag{28}$$

$$\sigma^2[\boldsymbol{\kappa}] = \frac{1}{n(1-\mathbf{\Pi}_E)^2} \times \left\{ \mathbf{\Pi}_E + \mathbf{\Pi}_E^2 - \sum [\pi_{ii}(\pi_{i+} + \pi_{+i})] \right\} \tag{29}$$

$$= \frac{\mathbf{\Pi}_O - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}} \tag{30}$$

1. $\boldsymbol{\kappa}$ equals 0 when the agreement equals that expected by chance, and it equals 1 when there is perfect agreement.
2. The stronger the agreement, the higher the values occur, for a given pair of marginal distribution. Negative values occur when agreement is weaker than expected by chance.

Reliability Studies with Kappa Statistic

1. To test the two-sided (one-sided) hypothesis

$$H_0 : \quad \boldsymbol{\kappa} = 0$$
$$\text{versus} H_A : \quad \boldsymbol{\kappa} \neq 0, \quad (H_A : \boldsymbol{\kappa} > 0), \tag{31}$$

2. Use the test statistic

$$z = \frac{\boldsymbol{\kappa}}{s.e.(\boldsymbol{\kappa})} \tag{32}$$

$$(1-\alpha) \times 100\% \quad \text{C.I. for } \boldsymbol{\kappa}\boldsymbol{\kappa} \pm Z_{1-\alpha/2} \; s.e.(\boldsymbol{\kappa}), \tag{33}$$

where $s.e.(\boldsymbol{\kappa}) = \sqrt{\sigma^2[\boldsymbol{\kappa}]} = \sigma[\boldsymbol{\kappa}]$

Two Surveys with Same Diet Questionnaire

1. $\boldsymbol{\kappa} = 0.378$, $s.e.(\boldsymbol{\kappa}) = 0.043$, $z = 8.8$, p-value=$2(1 - \Phi(8.8)) < 0.001$, and 95% confidence interval is $(0.298, 0.457)$.
2. It is rarely plausible that agreement is no better than expected by chance.
3. Thus rather than testing $H_0 : \boldsymbol{\kappa} = 0$, it is more important to estimate strength of agreement, by constructing a confidence interval for $\boldsymbol{\kappa}$.

Reliability Studies with Kappa Statistic: Guildlines

1. $\boldsymbol{\kappa} > 0.75$ denotes excellent reproducibility.
2. $0.4 < \boldsymbol{\kappa} \leq 0.75$ denotes good reproducibility.
3. $\boldsymbol{\kappa} \leq 0.4$ denotes marginal reproducibility.

In general, reproducibility is not good for many items in dietary survey, indicating the need for multiple dietary assessments to reduce variability.

Reliability Studies with Kappa Statistic: Notes

1. A weight form of the kappa statistic allows you to assign weights, or scores, to the various categories so that you can incorporate such consideration into the construction of the test statistic.
2. **Gamma**, **Kendalll's tau-b**, **Kendalll's tau-c**, **Somer's D** statistics are all based on concordant and discordant pairs, that is, they use the relative ordering on the levels of the variables to determine whether association is negative, positive, or present at all. They differ mainly on their strategies for adjusting for ties and sample size.

Reliability Studies with Kappa Statistic Program

```
> # load package vcd
> library(vcd)
> help(Kappa)
> Kappa(survey.tab)
               value        ASE
Unweighted 0.3781906 0.04100635
Weighted   0.3781906 0.05504449

> confint(Kappa(survey.tab))
Kappa              lwr       upr
  Unweighted 0.2978196 0.4585616
  Weighted   0.2703054 0.4860758
```


Reliability Studies with Kappa Statistic Program

```
> # MACRO function from
> # http://www.itc.nl/~rossiter/teach/R/R_ac.pdf
> survey.tab<-matrix(c(136,92,69,240),nrow=2,byrow=T)
> survey.kappa<-kappa(survey.tab)
> summary.kappa(survey.kappa)
```


Reliability Studies with Kappa Statistic Program

```
kappa.stat <- function(o, w=FALSE)
{
    n <- sum(o)
    e <- outer(apply(o, 1, sum), apply(o, 2, sum))/n
    if (is.matrix(w) == FALSE) {
        qo <- 1-(po <- sum(diag(o))/n)
        qe <- 1-(pe <- sum(diag(e))/n)
        kappa <- 1-qo/qe
        sk <- sqrt(po*qo/(n*qe^2))
        sk0 <- sqrt(pe/(n*qe))
        stopifnot(kappa >= 0)
        z <- kappa/sk0
        c("kappa"=kappa, "sigma-kappa"=sk, "sigma-kappa-0"=sk0,
            "95% lcl"=kappa-qnorm(0.975)*sk,
            "95% ucl"=kappa+qnorm(0.975)*sk,
```

```
        "Z value"=z, "P value"=pnorm(z, lower=FALSE)*2)
    }
    else {
        qow <- sum(w*o)/n
        qow2 <- sum(w*w*o)/n
        qew <- sum(w*e)/n
        qew2 <- sum(w*w*e)/n
        kw <- 1-qow/qew
        skw <- sqrt((qow2-qow^2)/n/qew^2)
        skw0 <- sqrt((qew2-qew^2)/n/qew^2)
        stopifnot(kw >= 0)
        zw <- kw/skw0
        c("kappa-w"=kw, "sigma-kappa-w"=skw, "sigma-kappa-w0"=skw0,
        "95% lcl"=kw-qnorm(0.975)*skw, "95% ucl"=kw+qnorm(0.975)*skw,
        "Z value"=zw, "P value"=pnorm(zw, lower=FALSE)*2)
    }
}
```


Reliability Studies with Kappa Statistic Program

```
> survey.tab<-matrix(c(136,92,69,240),nrow=2,byrow=T)
> kappa.stat(survey.tab)

        kappa   sigma-kappa sigma-kappa-0
 3.781906e-01  4.100635e-02  4.472105e-02

      95% lcl       95% ucl       Z value       P value
 2.978196e-01  4.585616e-01  8.456657e+00  2.751526e-17
```


Medical Tests: Diagnostic Tests and Screening Tests

Screening Test and Diagnostic Test

Breast cancer is considered largely a hormonal disease. An important hormone in breast-cancer resection is estradiol. The data in Table 10 on serum estradiol were obtained from 213 breast-cancer cases and 432 age-matched controls. All women were age 50-59 years.

Screening Test and Diagnostic Test

Table 7: Serum-Estradiol Data

Serum estradiol (pg/ml)	Case ($N = 213$)	Controls ($N = 432$)
01–04	28	72
05–09	96	233
10–14	53	86
15–19	17	26
20–24	10	6
25–29	3	5
30+	6	4

Screening Test and Diagnostic Test

Screening Test and Diagnostic Test

1. Evaluate the **accuracy** of the estradiol level as a diagnostic test. (What is the optimal cut-off point?)
2. The preceding sample was selected to oversample cases. In the general population, the **prevalence** of breast cancer is about 2% among women 50 to 59 years old. Evaluate the usefulness of the estradiol level as a diagnostic test. (What is the optimal cut-off point when you consider the prevalence?)

Screening Test and Diagnostic Test

1. What is the accuracy of a diagnostic test?

Screening Test and Diagnostic Test

1. What is the accuracy of a diagnostic test?
2. What are the sensitivity and specificity?

Screening Test and Diagnostic Test

1. What is the accuracy of a diagnostic test?
2. What are the sensitivity and specificity?
3. What are the predictive positive value and predictive negative value?

Screening Test and Diagnostic Test

1. What is the accuracy of a diagnostic test?
2. What are the sensitivity and specificity?
3. What are the predictive positive value and predictive negative value?
4. What is the ROC curve?

Screening Test and Diagnostic Test

1. What is the accuracy of a diagnostic test?
2. What are the sensitivity and specificity?
3. What are the predictive positive value and predictive negative value?
4. What is the ROC curve?
5. How to decide the cut-off point?

Medical Tests:
Diagnostic Tests and Screening Tests

Medical Tests:
Diagnostic Tests and Screening Tests

1. The purpose of **diagnostic testing** is to obtain objective evidence of the presence or absence of a particular condition.
2. This evidence can be obtained to detect disease at its earliest stages among asymptomatic persons in the general population, a process referred to as screening.
3. **Screening** is an application of a test or procedure to asymptomatic, apparently well individuals, in order to separate those with a relatively high probability of having a given disease from those with a relatively low probability of having the disease.

Medical Tests:
Diagnostic Tests and Screening Tests

1. Investigators often conduct a study to evaluate a simple new screening test compared to **"gold standard test"**.
2. The disease status is usually defined by **"gold standard" test**.
3. In the simplest case the test will simply be classified as having a positive (disease likely) or negative (disease unlikely) finding.

Medical Tests:
Diagnostic Tests and Screening Tests

4. Further, suppose that there is a **"gold standard"** that tells us whether or not a subject actually has the disease.
5. The definite classification might be based upon data from follow-up, invasive radiographic or surgical procedures, or autopsy results.
6. In many cases, the **"gold standard"** itself will only be relatively correct, but nevertheless the best classification available.

Medical Tests:
Diagnostic Tests and Screening Tests

7. Ideally, those with the disease should all be classified as having disease, and those without disease should be classified as non-diseased.
8. For this reason, two indices of the performance of a test consider how often such correct classification occurs.
9. However, classification of disease is not perfect, errors in measurement lead to misclassification of outcome or exposure.

Medical Tests:
True Positive Test and True Negative Test

1. A test is **true positive test** if the test is positive and the subject has the disease.
2. A test is **true negative test** if the test is negative and the subject does not have the disease.

The Simplest Medical Tests with a Dependent 2×2 Table

We can summarize a medical test results as 2×2 table as shown in Table .

captionTrue Positive Test and True Negative Test

	Disease	
Medical Test	Present (D+)	Absent (D-)
Positive (T+)	true positive	false positive
Negative (T-)	false negative	true negative

Medical Tests: Sensitivity and Specificity

1. The **sensitivity** of a screening test of a disease is the probability that the screening test of an individual is positive and test classify that individual as having the disease given that person has the disease.
2. The **specificity** of a screening test of a disease is the probability that the screening test of an individual is negative and test classify that individual as not having the disease given that person does not have the disease.

Medical Tests: Sensitivity and Specificity

$$\textbf{Sensitivity} = P[T+ \mid D+] = P[\text{Test Positive} \mid \text{Disease Present}]$$
$$\textbf{Specificity} = P[T- \mid D-] = P[\text{Test Negative} \mid \text{Disease Absent}]$$

1. Sensitivity is sometimes called **true positive rate (TFR)**.
2. Specificity is sometimes called **true negative rate (TNR)**.

Medical Tests: False Positive Test and False Negative Test

1. A **false positive test** if the test is positive and the subject does not have the disease.
2. A **false negative test** if the test is negative and the subject has the disease.

Medical Tests: False Positive Test and False Negative Test

1. **false-positive rate (FPR)** is that 1 minus sensitivity.
2. **false-negative rate (FNR)** is that 1 minus sensitivity.

Medical Tests: Positive Predictive Value and Negative Predictive Value

1. The **positive predictive value (PPV)**, $\mathbf{PV}^+$, is the predictive value of a positive test and is defined as the probability that a person has a disease given that the test is positive (also known as **predictive value positive**).
2. The **negative predictive value (NPV)**, $\mathbf{PV}^-$, is the predictive value of a negative test and is defined as the probability that a person does not have a disease given that the test is negative (also known as **predictive value negative**).

Medical Tests: Positive Predictive Value and Negative Predictive Value

The $\mathbf{PV}^+$ and $\mathbf{PV}^-$ are depend on the probability of disease occurrence **(prevalence)**, $P[D+]$, in population such that $P[D+] + P[D-] = 1$.

Medical Tests: Positive Predictive Value and Negative Predictive Value

$$\begin{aligned}
\mathbf{PV}^+ &= P[D+ \mid T+] = \frac{P[D+, T+]}{P[T+]} && (34)\\
&= \frac{P[T+ \mid D+]P[D+]}{P[T+ \mid D+] \times P[D+] + P[T+ \mid D-] \times P[D-]} && (35)\\
& && (36)\\
&= \frac{\mathbf{sensitivity} \times P[D+]}{\mathbf{sensitivity} \times P[D+] + (1 - \mathbf{specificity}) \times P[D-]} && (37)
\end{aligned}$$

Medical Tests: Positive Predictive Value and Negative Predictive Value

$$\begin{aligned}
\mathbf{PV}^- &= P[D- \mid T-] = \frac{P[D-, T-]}{P[T-]} && (38)\\
&= \frac{P[T- \mid D-]P[D-]}{P[T- \mid D+] \times P[D+] + P[T- \mid D-] \times P[D-]} && (39)\\
&= \frac{\mathbf{specificity} \times P[D-]}{(1 - \mathbf{sensitivity}) \times P[D+] + \mathbf{specificity} \times P[D-]} && (40)
\end{aligned}$$

Medical Tests: Sample Data as 2×2 Table

The observed data is constructed as 2×2 table as in Table 8.

Table 8: Sensitivity and specificity: 2×2 Table

	Disease		
Medical Test	Present (D+)	Absent (D-)	Total
Positive (T+)	$O_{1,1} = a$	$O_{1,2} = b$	$a + b = n_{1\cdot}$ (row 1 margin)
Negative (T-)	$O_{2,1} = c$	$O_{2,2} = d$	$c + d = n_{2\cdot}$ (row 2 margin)
Total	$a + c = n_{\cdot 1}$ column 1 margin	$b + d = n_{\cdot 2}$ column 2 margin	$a + b + c + d = n_{\cdot\cdot} = n$ (grand total)

Sensitivity and Specificity: Point Estimation

The estimated mean sensitivity and specificity are

$$\begin{aligned}
\widehat{\mathbf{sensitivity}} &= P[T+ \mid D+] = \frac{a}{a+c} && (41)\\
\widehat{\mathbf{specificity}} &= P[T- \mid D-] = \frac{d}{b+d} && (42)
\end{aligned}$$

Point Estimation: Positive Predictive Value and Negative Predictive Value

1. The estimated mean $\mathbf{PV}^+$ and $\mathbf{PV}^-$ actually depend on the disease prevalence.
2. However, we can seen many clinical literatures calculated the $\mathbf{PV}^+$ and $\mathbf{PV}^-$ as

$$\begin{aligned}
\widehat{\mathbf{PV}}_\star^+ &= P[D+ \mid T+] = \frac{a}{a+b} && (43)\\
\widehat{\mathbf{PV}}_\star^- &= P[D- \mid T-] = \frac{d}{c+d} && (44)
\end{aligned}$$

3. The above two calculations are not exact the definition of original $\mathbf{PV}_\star^+$ and $\mathbf{PV}_\star^-$.

Point Estimation: Positive Predictive Value and Negative Predictive Value

The difficulty in that we usually have no information about the disease prevalence.

Medical Tests: Accuracy

1. Vague term
2. Missclassification probability

$$\begin{aligned} & P(\text{Test result } \neq \text{ Disease Status}) \\ = \ & \text{Disease Prevalence} \times (1 - \text{Sensitivity}) \\ & +(1 - \text{Disease Prevalence}) \times (1 - \text{Specificity}) \qquad (45) \\ & P(Y \neq D\ ) \\ = \ & P(D=1)(1-\text{Sen}) + (1-P(D=1))(1-\text{Spe}); \qquad (46) \end{aligned}$$

Where $Y = 1$ if test result is postiive, $Y = 0$ if test result is negative; and $D = 1$ for disease and $D = 0$ for non-disease.

Example: Breast Cancer and Estradiol Levels

1. Breast cancer is considered largely a hormonal disease.
2. In the population, the prevalence of breast cancer is about 2%.
3. An important hormone in breast-cancer is estradiol.
4. Investigators chose Estridal $\geq$ 20pg/ml as an abnormal (having breast cancer),
5. The data in Table 9. on serum estradiol were obtained from 213 breast-cancer cases and 432 age-matched controls, and all women were age 50-59 years.

Example: Estradiol and Breaset Cancer

Table 9: Estradiol and Breast Cancer: Case-Control Study

	Breast		
Estradiol Test	Case (D+)	Control (D-)	Total
Positive (T+) $\geq$ 20pg/ml	19	15	34
Negative (T-) $<$ 20pg/ml	194	417	611
Total	213	432	645

$$\text{Sensitivity} = \frac{19}{213} = 0.089; \qquad \text{Sepecificity} = \frac{417}{432} = 0.965. \qquad (47)$$

Example: Estradiol and Breaset Cancer

In the population, the prevalence of breast cancer is about 2%.

$$\begin{aligned} \text{PPV(PV+)} &= \frac{\text{Sen} \times P(D)}{\text{Sen} \times P(D) + (1 - \text{Sep} \times (1 - P(D))} \\ &= \frac{0.089(0.02)}{0.089(0.02) + (1-0.965)(1-0.02)} = 0.050; \\ \text{NPV(PV}-) &= \frac{(1-\text{Sep}) \times (1-P(D))}{(1-\text{Sen}) \times P(D) + (1-\text{Sep}) \times (1-P(D))} \\ &= \frac{(1-0.965)(1-0.02)}{(1-0.089)0.02 + (1-0.965)(1-0.02)} = 0.651. \end{aligned} \qquad (48)$$

Example: Estradiol and Breaset Cancer

Thus, there is a 5% probability of breast cancer among 50-59-year-old women with serum Estradiol $\geq$ 20pg/ml. This is about 2.5 times the general population rate (2%).

Screening Test and Diagnostic Test

1. Sometimes, a new screening test is not a simple screening test.
2. The new screening test may provide several categories of response rather than simply test positive or test negative.
3. In other instances, the results of the test are reported as continuous variable.
4. In either case, the designation of a cut-off point for distinguishing test positive versus test negative is arbitrary.

Medical Tests: ROC Curve
Receiver Operating Characteristic Curve

Example: Breast Cancer and Estradiol Levels

Breast cancer is considered largely a hormonal disease. An important hormone in breast-cancer resection is estradiol. The data in Table 10 on serum estradiol were obtained from 213 breast-cancer cases and 432 age-matched controls. All women were age 50-59 years.

Example: Breast Cancer and Estradiol Levels

Table 10: Serum-Estradiol Level and Breaset Cancer Data

Serum estradiol (pg/ml)	Case $(N = 213)$	Controls $(N = 432)$
01–04	28	72
05–09	96	233
10–14	53	86
15–19	17	26
20–24	10	6
25–29	3	5
30+	6	4

Example: Breast Cancer and Estradiol Levels

1. Evaluate the **accuracy** of the estradiol level as a diagnostic test. (What is the optimal cut-off point?)
2. The preceding sample was selected to oversample cases. In the general population, the **prevalence** of breast cancer is about 2% among women 50 to 59 years old. Evaluate the usefulness of the estradiol level as a diagnostic test. (What is the optimal cut-off point when you consider the prevalence?)

Medical Tests: ROC Curve

1. Most tests have some quantitative aspect.
2. For Example, biomarkers for Cancer, PSA, CA–125.
3. Tests that invove an element of subjective assessment are often ordinal in nature.
4. For example, radiologist's reading images as "definitely", "probably", "possibly", "definite not".

Medical Tests: ROC Curve

1. The same statistical approach can be used only if we can select a **cut off point** to distinguish "normal" from "abnormal," which is not a trivial problem.
2. The decision rule is based on whether or not the test result (or some transformation of it) exceed a **threshold** value.
3. The choice a suitable threshold will vary with circumstances.
4. The choice threshold depends on the trade-off that is acceptable between failing to detect disease and falsely identifying disease with the test.

Medical Tests: ROC Curve

The **ROC** curve is a device that simply describes the range of trade-offs that can be ahieved by the test.

Medical Tests: ROC Curve

1. Firstly, we can investigate to what extent the test results differ among people who do or do not have the diagnosis of interest.
2. The receiver operating characteristic (ROC) plot is one way to do this.
3. These plots were developed in the 1950s for evaluating radar signal detection. Only recently have they become commonly used in medicine.

Medical Tests: ROC Curve

A receiver operating characteristic plot is obtained by calculating the sensitivity and specificity of every observed data value at several defined cut-off pointsf (5-10 or more) and plotting sensitivity against $1 -$ specificity,

Medical Tests: ROC Curve

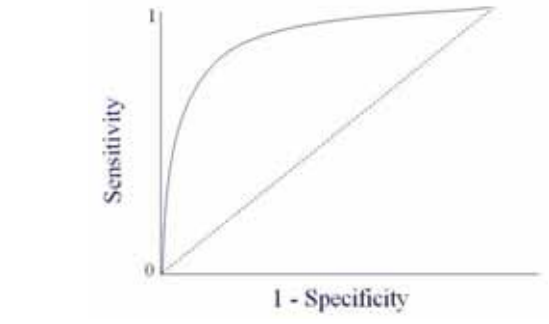

Figure 1: Receiver Operating Characteristic Curve

Medical Tests: ROC Curve

We just want to calculate sensitivity and specificity for this test, we have to choose a "cutpoint" which separates "normal" from "abnormal".

Example: Estradiol and Breaset Cancer Cut-Off Point at Estridal ≥ 30pg/ml

If we chose Estridal ≥ 30pg/ml as an abnormal (having breast cancer), we can "collapse" some rows and get the following familiar 2×2 table:

Example: Estradiol and Breaset Cancer Cut-Off Point at Estridal ≥ 30pg/ml

Table 11: Estradiol ≥ 30pg/ml as a Cut-Off Point

	Breast		
Estradiol Test	Present (D+)	Absent (D-)	Total
Positive (T+) ≥ 30pg/ml	6	4	10
Negative (T-) < 30pg/ml	207	428	635
Total	213	432	645

$$\text{Sensitivity} = \frac{6}{213} = 0.028; \qquad \text{Sepecificity} = \frac{428}{432} = 0.990. \tag{49}$$

Example: Estradiol and Breaset Cancer Cut-Off Point at Estradiol ≥ 20pg/ml

If we chose Estridal ≥ 20pg/ml as an abnormal (having breast cancer), we can "collapse" some rows and get the following familiar 2×2 table:

Example: Estradiol and Breaset Cancer Cut-Off Point at Estridal ≥ 20pg/ml

Table 12: Estradiol ≥ 20pg/ml as a Cut-Off Point

	Breast		
Estradiol Test	Present (D+)	Absent (D-)	Total
Positive (T+) ≥ 20pg/ml	19	15	34
Negative (T-) < 20pg/ml	194	417	611
Total	213	432	645

$$\text{Sensitivity} = \frac{19}{213} = 0.089; \qquad \text{Sepecificity} = \frac{417}{432} = 0.965. \tag{50}$$

Example: Different Estradiol Cut-Off Points

Table 13: Sensitivity and Specificity of Different Estradiol Cut-Off Points for Breast Cancer

Serum estradiol Cut Point	Sensitivity	Specivity
≥ 30 pg/ml	0.0281	0.990
≥ 25 pg/ml	0.0422	0.979
≥ 20 pg/ml	0.0892	0.965
≥ 15 pg/ml	0.1690	0.905
≥ 10 pg/ml	0.4178	0.706
≥ 5 pg/ml	0.8685	0.166
≥ 0 pg/ml	1.0000	0.000

Example: Estradiol and Breaset Cancer

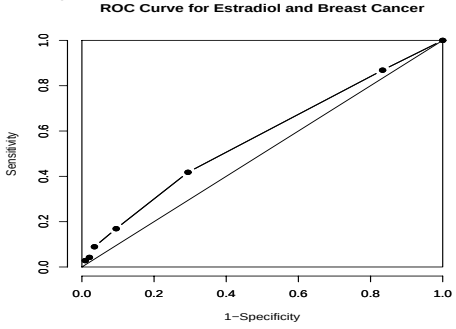

Figure 2: ROC Curve for Estradiol and Breaset Cancer

Example: Estradiol and Breaset Cancer PPV and NPV

1. When choose a different "cutpoint" which separates "normal" from "abnormal", we will have different sensitivity and specificity.
2. We will have different positive predictive value and negative predictive palue

Example: Estradiol and Breaset Cancer Cut-Off Point at Estridal ≥ 20pg/ml

In the population, the prevalence of breast cancer is about 2%.

Table 14: Estradiol ≥ 20pg/ml as a Cut-Off Point

	Breast		
Estradiol Test	Present (D+)	Absent (D-)	Total
Positive (T+) ≥ 20pg/ml	19	15	34
Negative (T-) < 20pg/ml	194	417	611
Total	213	432	645

$$\text{Sensitivity} = \frac{19}{213} = 0.089; \qquad \text{Sepecificity} = \frac{417}{432} = 0.965. \tag{51}$$

Example: Estradiol and Breaset Cancer Cut-Off Point at Estridal ≥ 20pg/ml

In the population, the prevalence of breast cancer is about 2%.

$$\begin{aligned}
\text{PPV(PV+)} &= \frac{\text{Sen} \times P(D)}{\text{Sen} \times P(D) + (1 - \text{Sep} \times (1 - P(D))} \\
&= \frac{0.089(0.02)}{0.089(0.02) + (1 - 0.965)(1 - 0.02)} = 0.050; \\
\text{NPV(PV}-\text{)} &= \frac{(1 - \text{Sep}) \times (1 - P(D))}{(1 - \text{Sen}) \times P(D) + (1 - \text{Sep}) \times (1 - P(D))} \\
&= \frac{(1 - 0.965)(1 - 0.02)}{(1 - 0.089)0.02 + (1 - 0.965)(1 - 0.02)} = 0.651.
\end{aligned} \tag{52}$$

Example: Estradiol and Breaset Cancer Cut-Off Point at Estridal ≥ 20pg/ml

Thus, there is a 5% probability of breast cancer among 50-59-year-old women with serum Estradiol ≥ 20pg/ml. This is about 2.5 times the general population rate (2%).

Example: Different Estradiol Cut-Off Points

Table 15: PPV and NPV of Different Estradiol Cut-Off Points for Breast Cancer

Serum estradiol Cut Point	PPV	NPV
≥ 30 pg/ml	0.058	0.318
≥ 25 pg/ml	0.039	0.515
≥ 20 pg/ml	0.049	0.651
≥ 15 pg/ml	0.035	0.848
≥ 10 pg/ml	0.028	0.961
≥ 5 pg/ml	0.020	0.996
≥ 0 pg/ml	0.020	1.000

Example: Estradiol and Breaset Cancer

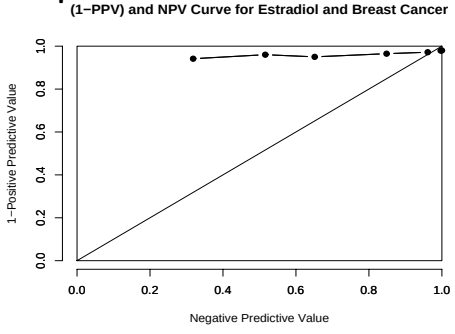

Figure 3: (1-PPV) versus NPV Curve for Estradiol and Breaset Cancer

Example: Estradiol and Breaset Cancer

Example: Estradiol and Breaset Cancer

```
Est.mat<-matrix(
c(5,28,72,
10,96,233,
15,53,86,
20,17,26,
25,10,6,
30,3,5,
60,6,4)
,nrow=7,ncol=3,byrow=T)
Est.mat<-Est.mat[rev(rank(Est.mat[,1])),]
```


Example: Estradiol and Breaset Cancer

```
Est.row.sum<-matrix(apply(Est.mat,1,sum),7,1) # row sum
Est.col.sum<-matrix(rep(matrix(apply(Est.mat,2,sum),1,3),7)
                    ,7,3,byrow=T)  # col sum
Est.col.cum<-apply(Est.mat,2,cumsum)          # col culmulative sum
Neg.mat<-Est.col.sum-Est.col.cum
sen.mat<-matrix(Est.col.cum[,2]/Est.col.sum[,2],7,1) # [1:6,]
sep.mat<-matrix(Neg.mat[,3]/Est.col.sum[,3],7,1)     # [1:6,]
sen.sep<-cbind(sen.mat,sep.mat)
```


Example: Estradiol and Breaset Cancer

```
Est.mat
Est.row.sum
Est.col.sum
Est.col.cum
Neg.mat
sen.sep
```


Example: Estradiol and Breaset Cancer

```
prevD<-0.02
PPV<-(prevD*sen.mat)/(prevD*sen.mat+(1-sep.mat)*(1-prevD))
NPV<-((1-sep.mat)*(1-prevD))/
    ((1-sen.mat)*prevD+(1-sep.mat)*(1-prevD))

PPV.NPV<-cbind(PPV,NPV)
PPV.NPV
```


Example: Estradiol and Breaset Cancer

```
plot(1-sep.mat,sen.mat,xlab="1-Specificity", type="n", bty="n",
    ylab="Sensitivity", xlim=c(0,1), ylim=c(0,1),
    main="ROC Curve for Estradiol and Breast Cancer")
points(1-sep.mat,sen.mat,pch=19,type="b", lwd=1)
```

Example: Estradiol and Breaset Cancer

```
# ROC
plot(1-sep.mat,sen.mat,xlab="1-Specificity", type="b", bty="n",
    axes=T, lty=1, lwd=1.5, pch=19,
    main="ROC Curve for Estradiol and Breast Cancer",
    ylab="Sensitivity", xlim=c(0,1), ylim=c(0,1))
points(1-sep.mat,sen.mat,pch=19,type="b", lwd=1.5, lty=1)
axis(1,outer=FALSE,tick=1,lty=0)
axis(2,outer=FALSE,tick=1,lty=0)
```


Example: Estradiol and Breaset Cancer

```
lines(c(0,1),c(0,0),lty=1) #  x=0
lines(c(1,1),c(0,1),lty=1) #  x=1
lines(c(0,0),c(0,1),lty=1) #  y=0
lines(c(0,1),c(1,1),lty=1) #  y=1
lines(c(0,1),c(0,1),lty=1) #
```


Example: Estradiol and Breaset Cancer

```
# PPV, NPV
plot(NPV,1-PPV, type="b", bty="n", cex=0.7,
    axes=T, lty=1, lwd=1.5, pch=19,
    main="(1-PPV) and NPV Curve for Estradiol and Breast Cancer",
    xlab="Negative Predictive Value",
    ylab="1-Positive Predictive Value",
    xlim=c(0,1), ylim=c(0,1))
points(NPV,(1-PPV),pch=19,type="b", lwd=1.5, lty=1)
axis(1,outer=FALSE,tick=1,lty=0)
axis(2,outer=FALSE,tick=1,lty=0)
```


Example: Estradiol and Breaset Cancer

```
lines(c(0,1),c(0,0),lty=1) #  x=0
lines(c(1,1),c(0,1),lty=1) #  x=1
lines(c(0,0),c(0,1),lty=1) #  y=0
lines(c(0,1),c(1,1),lty=1) #  y=1
lines(c(0,1),c(0,1),lty=1) #
```