2005 Biostatistics Mid-Term Exam

CF Jeff Lin, MD., PhD.

December 27, 2005

Normal Range and Reference Level

An investigator has n healthy male subjects enrolled in a study and wants to set up the "normal range" of BMI. BMI is computed as the ratio of weight in kilograms to the square of the height in meters. Suppose BMI $= D_i = \frac{X_i}{Y_i^2}$. Please derive the normal range for BMI.

Normal Range and Reference Level

Table 1: Body mass index (BMI) data of n healthy male subjects.

Patient Number	Weight (kg)	Height (m)
ID	WTKG	HTM
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
n	x_n	y_n

Normal Range or Reference Interval

1. What's is normal range or reference interval?

Normal Range or Reference Interval

1. What's is normal range or reference interval?

2. Why do we need normal range in clinical medicine?

Normal Range or Reference Interval

1. What's is normal range or reference interval?

2. Why do we need normal range in clinical medicine?

3. What's the difference between point estimation (prediction) for population mean value and individual value?

Normal Range or Reference Interval

1. What's is normal range or reference interval?

2. Why do we need normal range in clinical medicine?

3. What's the difference between point estimation (prediction) for population mean value and individual value?

4. What's the difference between interval estimation (prediction) for population mean value and individual value?

Normal Range or Reference Interval

1. What's is normal range or reference interval?

a) Who is "normal" anyway?

b) In the Taiwan population almost everyone has hard fatty deposits in their coronary arteries, which result in death for many of them. Very few Africans have this; they die from other causes.

c) So it is "normal" in the Taiwan to have an abnormality.

d) We usually say that normal people are the apparently healthy members of the local population.

e) We can draw a sample of these and make the measurement on them.

Normal Range or Reference Interval

2. Why do we need normal range in clinical medicine?

We use normal range or reference interval to compare characteristics of a marker of disease progression between affected populations.

Normal Range or Reference Interval

3. What's the difference between point estimation (prediction) for mean value and individual value?

We often use sample mean to estimate or predict population mean value or individual value or new observation value.

Normal Range or Reference Interval

4. What's the difference between interval estimation (prediction) for population mean value and individual value?

We use confidence interval as interval estimation (prediction) for population mean value. We use prediction interval as interval estimation (prediction) for individual value or new observation value. The true difference is the estimation of variances between two situations.

Normal Range or Reference Interval

1. We usually say that normal people are the apparently healthy members of the local population.

2. We can draw a sample of these and make the measurement on them.

3. If we use the range of the observations, the difference between the two most extreme values, we can be fairly confident that if we carry on sampling we will eventually find observations outside it, and the range will get bigger and bigger.

Normal Range or Reference Interval

1. To avoid this we use a range between two quantiles, usually the 2.5 centile and the 97.5 centile, which is called the normal range, 95% reference range, or 95% reference interval.

2. This leaves 5% of normals outside the "normal range", which is the set of values within which 95% of measurements from apparently healthy individuals will lie.

Normal Range or Reference Interval

1. A difficulty comes from confusion between "normal" as used in medicine and "Normal distribution" as used in statistics.

2. This has led some people to develop approaches which say that all data which do not fit under a Normal curve are abnormal!

3. Such methods are simply absurd, there is no reason to suppose that all variables follow a Normal distribution.

4. The term "reference interval", which is becoming widely used, has the advantage of avoiding this confusion.

Normal Range or Reference Interval

1. However, the most commonly used method of calculation rests on the assumption that the variable follows a Normal distribution.

2. We have already seen that in general most observations fall within two standard deviations of the mean, and that for a Normal distribution 95% are within these limits with 2.5% below and 2.5% above. If we estimate the mean, $\hat{\mu}$, and standard deviation, $s = \hat{\sigma}$, of data from a "Normal population" we can estimate the reference interval approximately as $\hat{\mu} - 2s$ to $\hat{\mu} + 2s$.

Normal Range or Reference Interval

1. Suppose the individual observation is $D_i \sim N(\mu_D, \sigma_D^2)$ and we have total n subjects in reference sample to estimate normal range. Since

$$E(D_i) = \mu_D, \quad (1)$$

$$E(\bar{D}) = \mu_D. \quad (2)$$

So we use observed sample mean $\bar{d}$ to estimate D_i.

Normal Range or Reference Interval

2. Now, suppose the individual observation is $D_i = \mu_D + \varepsilon_i$, we use $\hat{D}_i = \bar{D} + \varepsilon_i$ to estimate D_i, then

$$\begin{aligned} \mathbf{Var}(\bar{D} + \varepsilon_i) &= \mathbf{Var}(\bar{D}) + \mathbf{Var}(\varepsilon_i) && (3) \\ &= \mathbf{Var}(\bar{D}) + \mathbf{Var}(\varepsilon_i). && (4) \end{aligned}$$

Normal Range or Reference Interval

3. Now, we use $\frac{s_D^2}{n}$ to estimate $\mathbf{Var}(\bar{d})$ and use $s_D = \hat{\sigma}_D$, then

$$\begin{aligned}
\widehat{\mathbf{Var}}(\bar{D} + \varepsilon_i) &= \widehat{\mathbf{Var}}(\bar{D}) + \widehat{\mathbf{Var}}(\varepsilon_i) && (5)\\
&= \frac{s_D^2}{n} + s_D && (6)\\
&= s_D^2\left(1 + \frac{1}{n}\right). && (7)
\end{aligned}$$

Normal Range or Reference Interval

4. So the one possible method to estimate normal range is

$$\left(\bar{d}_D - 1.96 \times \sqrt{\widehat{\mathbf{Var}}(\bar{d} + \varepsilon_i)},\ \bar{d}_D + 1.96 \times \sqrt{\widehat{\mathbf{Var}}(\bar{d} + \varepsilon_i)}\right) \tag{8}$$

$$= \left(\bar{d}_D - 1.96 \times s_D\left(1 + \frac{1}{\sqrt{n}}\right),\ \bar{d}_D + 1.96 \times s_D\left(1 + \frac{1}{\sqrt{n}}\right)\right). \tag{9}$$

When $n \longrightarrow \infty$, the reference interval becomes as

$$\left(\bar{d}_D - 1.96 \times s_D,\ \bar{d}_D + 1.96 \times s_D\right). \tag{10}$$

Two-Sample t Test and Paired t Test

Two-Sample t Test and Paired t Test

A new compound $ABC\text{-}123$, is being developed for long-term treatment of patients with chronic asthma. Total $2 \times n$ asthmatic patients were enrolled in a double-blind study and randomized to receive daily oral doses of $ABC\text{-}123$ or a placebo for 6 weeks. The primary measurement of interests is the resting $FEV1$ (forced expiratory volume during the first second of expiration), which is measured at the end of the 6-week treatment period. Data (in liters) are shown in the Table 2. Does administration of $ABC\text{-}123$ appear to have any effect on $FEV1$?

Two-Sample t Test and Paired t Test

Suppose that X is FEV1 for the $ABC\text{-}123$, and X_i, $i = 1, \ldots, n$, is normally distributed as $N(\mu_X, \sigma_X^2)$. And suppose that Y_i is FEV1 for the placebo group and Y_i, $i = 1, \ldots, n$ is normally distributed as $N(\mu_Y, \sigma_Y^2)$. X_i and Y_i are assumed independent for all i and $i = 1, \ldots, n$.

1. Calculate the sample standard error of $(\bar{x} - \bar{y})$, in terms of $x_i, y_i, \mu_X, \mu_Y, \sigma_X^2$ and σ_Y^2.

2. Derive the test statistic, and call it as T^U.

Two-Sample t Test

Table 2: The results of a trial of a new chemical compound for chronic asthma: Part I

ABC-123 Group		Placebo Group	
Subject ID	6 Week fev6	Subject ID	6 Week fev6
1	x_1	1	y_1
2	x_2	2	y_2
⋮	⋮	⋮	⋮
i	x_i	i	y_i
⋮	⋮	⋮	⋮
n	x_n	n	y_n

Two Samples Inference

1. Comparison of means from two sub-population can be useful for many purpose.

2. A difference in means indicates that something caused the difference.

3. Identification of differences is a first step toward understanding the cause of the difference.

Two-Sample t Test

1. The possibility of a difference between two population mean values is usually investigated by taking a sample for each population.

2. In a **two-sample** hypothesis-testing problem, the underlying of two different populations, neither or whose values is assumed known, are compared.

3. The two sub-populations consist of different members.

4. Two separate and **independent samples** are drawn for such comparisons; one from each sub-population.

Two-Sample t Test

The two different sub-populations could be sampled as strata from the overall population as shown in Table 3.

Table 3: Notation for Two populations and two Samples

Quantity	(sub)Population 1	(sub)Population 2
Sample size	n_1	n_2
population mean	μ_1	μ_2
Population standard deviation	σ_1	σ_2
j^{th} possible observation	X_{1j}	X_{2j}

Two-Sample t Test: Sample Statistic

Two-Sample t Test: Sample Statistic

$$\text{sample mean} \quad \overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij},\ i = 1,2$$

$$\text{sample variance} \quad S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2,\ i = 1,2$$

1. Are the population mean values (represented by μ_1 and μ_2 different $(\mu_1 \neq \mu_2)$ or are they the same $(\mu_1 = \mu_2)$?

2. The comparison of two population means is change from an inference about two population parameter μ_1 and μ_2 into an inference about the difference between the two means, $\delta = \mu_1 - \mu_2$.

Two-Sample t Test

The two different sub-populations are shown in Table 4.

Table 4: Notation for Two populations and two Samples

Quantity	(sub)Population 1	(sub)Population 2
Sample size	n_1	n_2
Population mean	μ_1	μ_2
Population variance	σ_1^2	σ_2^2
Population standard deviation	σ_1	σ_2
j^{th} possible observation	X_{1j}	X_{2j}
Sample mean	$\overline{X}_1$	$\overline{X}_2$
Sample variance	S_1^2	S_2^2
Sample stand deviation	S_1	S_2
Standard error	$S_{\overline{X}_1} = \frac{S_1^2}{n_1}$	$S_{\overline{X}_2} = \frac{S_2^2}{n_2}$

Two-Sample t Test: Testing Hypothesis

Two-Sample t Test: Testing Hypothesis

We consider three different tests in Table 5 in terms of $\delta = \mu_1 - \mu_2$.

Table 5: Two-sample Hypothesis Testing

H_0 : Null Hypothesis	H_A : Different Means	H_A : Difference between Means
$H_0 : \mu_1 = \mu_2$	$H_A : \mu_1 > \mu_2$	$H_A : \mu_1 - \mu_2 = \delta > 0$
$H_0 : \mu_1 = \mu_2$	$H_A : \mu_1 < \mu_2$	$H_A : \mu_1 - \mu_2 = \delta < 0$
$H_0 : \mu_1 = \mu_2$	$H_A : \mu_1 \neq \mu_2$	$H_A : \mu_1 - \mu_2 = \delta \neq 0$

Two-Sample t Test: Test Statistic

Two-Sample t Test: Test Statistic

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sigma^2_{\overline{X}_1 - \overline{X}_2}) \quad \text{or equivalently} \quad N(\delta, \sigma^2_D) \tag{11}$$

$$T^U = \mathsf{T} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\overline{X}_1 - \overline{X}_2}} = \frac{(\overline{D} - \delta)}{\hat{\sigma}_D}, \tag{12}$$

When the null hypothesis $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = \delta_0 = 0$ is true, then

$$T^U = \mathsf{T} \quad \sim \quad t_{n_1+n_2-2}. \tag{13}$$

T follows Student's t distribution with mean zero and $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom

Two-Sample t Test: Assumptions

1. What are the basic assumptions of two-sample t test?

Two-Sample t Test: Assumptions

1. Population 1 to be independent of population 2 and the two samples to consist of independently sampled observations.

2. The observations to be sampled from normally distributed parent populations.

3. The variance to be the same for both populations sampled ($\sigma_1^2 = \sigma_2^2$), occasionally called **homoscedasticity**.

Two-Sample t Test: Test Statistic Variance

If two populations are independent, then the variance of the mean difference $(\overline{X}_1 - \overline{X}_2)$ is

$$\mathbf{Var}(\overline{X}_1 - \overline{X}_2) = \mathbf{Var}(D) = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} = \sigma^2_D.$$

Two-Sample t Test: Test Statistic Variance

In general, σ_D^2 is unknown, so we need to estimate it. A natural estimate of the variance of the difference between two estimated mean values is

$$\hat{\sigma}_D^2 = S_{\text{un}}^2 = S_{\overline{X}_1 - \overline{X}_2}^2 = S_D^2 = S_{\overline{X}_1}^2 + S_{\overline{X}_2}^2 = \frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}. \quad (14)$$

Two-Sample t Test: Test Statistic Variance

1. Take advantage of the equal variance assumption.

2. A weighted average of the two estimated variance produces a single estimate of the common variance.

3. A pooled estimator of the common variance ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ — assumption 3)

$$\hat{\sigma}_D^2 = S^2_{\text{pooled}} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}. \tag{15}$$

Two-Sample t Test: Test Statistic Variance

1. Under the null hypothesis and assumptions, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we let the pooled variance estimator be $\hat{\sigma}^2 = S_{\text{pooled}}$, then

$$\hat{\sigma}_D^2 = S_D^2 = \frac{S_{\text{pooled}}^2}{n_1} + \frac{S_{\text{pooled}}^2}{n_2} = S_{\text{pooled}}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \tag{16}$$

2. S_{pooled}^2 estimates the variability associated with the observed difference between two sample mean values.

3. When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then $\mathcal{E}(S_{\text{pooled}}^2) = \sigma^2$.

Two-Sample t Test

1. Assume two population variances are equal as $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

2. The point estimator of difference between the two sample mean values $\mu_1 - \mu_2$ is $(\overline{X}_1 - \overline{X}_2)$.

Two-Sample t Test

3. Use the pooled estimator of the variance to test $H_0 : \mu_1 - \mu_2 = \delta_0$, the test statistic is

$$\mathsf{T} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_D} = \frac{(\overline{X}_1 - \overline{X}_2) - \delta_0}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{17}$$

where $S^2_{\text{pooled}} = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$.

4. The T-statistic has a Student's t distribution with degrees of freedom $n_1 + n_2 - 2$ when the null hypothesis $\mu_1 - \mu_2 = \delta_0 = 0$ is true.

Two-Sample t Test

To test $H_0 : \mu_1 - \mu_2 = 0$, , suppose we assume σ_1^2 and σ_2^2 are known, the test statistic

$$T^U = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\mathbf{Var}(\overline{X}_1 - \overline{X}_2)}} \tag{18}$$

$$\mathbf{Var}(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} \quad \text{(independent samples)} \tag{19}$$

When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, σ^2 is known, and $n \to \infty$,

$$T^U = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}} = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{2\sigma^2}{n}}} \tag{20}$$

Paired t Test

Paired t Test

A new compound $ABC\text{-}123$, is being developed for long-term treatment of patients with chronic asthma. Total n asthmatic patients were enrolled in a study and received daily oral doses of $ABC\text{-}123$ for 6 weeks. The primary measurement of interests is the resting $FEV1$ (forced expiratory volume during the first second of expiration), which is measured before and at the end of the 6-week treatment period. Data (in liters) are shown in the Table 6. Does administration of $ABC\text{-}123$ appear to have any effect on $FEV1$?

Paired t Test

Suppose that X is FEV1 before the treatment of $ABC\text{-}123$, and X_i, $i = 1, \ldots, n$, is normally distributed as $N(\mu_X, \sigma_X^2)$. And suppose that Y_i is FEV1 6 weeks after the treatment of $ABC\text{-}123$ and Y_i, $i = 1, \ldots, n$ is normally distributed as $N(\mu_Y, \sigma_Y^2)$. The covariance of X and Y is σ_{xy}.

1. Calculate the sample standard error of $(\bar{x} - \bar{y})$, in terms of $x_i, y_i, \mu_X, \mu_Y, \sigma_X^2$, σ_Y^2 and σ_{xy}.

2. Derive the test statistic, and call it as T^P.

Paired t Test

Table 6: The results of a trial of a new chemical compound for chronic asthma: Part II

	Before Treatment	After 6-Week Treatment
Subject	Placebo	*ABC-123*
ID	fev	fev
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
n	x_n	y_n

Paired t Test: Design

1. Two non-independent populations arise, for example, from a "before-after" experiment (research).

2. A series of experimental units ("baseline") then treated in a special way and measured again some time later.

3. The same individual is measured both at the beginning and the end of the experiment.

4. Since the same individual is measured twice, the two measurements are not independent.

Paired t Test: Design

The major different in this design is that each subject is used as his own control to compare before and after treatment.

Paired Sample Design

1. Two samples are said to be **paired** when each data point of the first sample is **matched** and is related to a unique data point of the second sample.

2. Two samples are said to be independent when the data point in one sample are unrelated to the data points in the second sample.

Paired t Test: Design

3. **paired** or **matched** study design is probably more definitive, because most other factors that influence response variable before treatment will also be present after treatment, and may influence the comparison.

4. Paired study design may be useful to control or eliminate those influential factors.

Paired t Test: Data Structure

Table 7: Data Structure of Two Nonindependent Samples for Paired t Test

Paired (matched) design Data Entry Style			
Subject	Before Treatment	After treatment	Difference between Pairs
i	X_{i1}	X_{i2}	$D_i = X_{i1} - X_{i2}$
1	X_{11}	X_{12}	$D_1 = X_{11} - X_{12}$
2	X_{21}	X_{22}	$D_2 = X_{21} - X_{22}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
i	X_{i1}	X_{i2}	$D_i = X_{i1} - X_{i2}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$n = n_1 = n_2$	X_{n1}	X_{n2}	$D_n = X_{n1} - X_{n2}$

Paired t Test

1. If $X_{i1}, \ldots, X_{n1}$ and $X_{i2}, \ldots, X_{n2}$ are normally distributed samples, then the differences $D_i = X_{i1} - X_{i2}$ are also normally distributed.

2. For large samples ($n > 30$ or so) from non-normal distributions, the mean difference $\overline{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$ typically has an approximate normal distribution (central limit theorem).

3. What's the variance?

Paired t Test

1. We assume that a random sample of size n.

2. The response variable before treatment, X_{i1} for each i^{th} subject is normal distributed with $N(\mu_i, \sigma^2)$

3. The response variable after treatment, X_{i2} is normal distributed with $N(\mu_i + \delta, \sigma^2)$.

4. If $\delta = 0$, then there is no difference of response variables between before and after treatment.

Paired t Test: Testing Hypothesis

$$H_0 : \delta = 0 \quad \text{or equivalently,} \quad H_0 : \mu_{\text{before}} = \mu_{\text{after}}.$$

Possible alternative hypotheses are:

$$H_A : \delta < 0 \quad \text{or equivalently,} \quad H_A : \mu_{\text{before}} < \mu_{\text{after}}$$

$$H_A : \delta > 0 \quad \text{or equivalently,} \quad H_A : \mu_{\text{before}} > \mu_{\text{after}}$$

$$H_A : \delta \neq 0 \quad \text{or equivalently,} \quad H_A : \mu_{\text{before}} \neq \mu_{\text{after}}.$$

Paired t Test: Test Statistic

1. The hypothesis-testing problem can thus be considered a **one-sample t test** when the variance is known.

2. The sample statistic $\overline{D}$ and sample variance of $\overline{D}$ are

$$\text{sample mean of difference:} \quad \overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_{i2} - X_{i1})$$

$$\text{sample variance:} \quad S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{(n-1)}$$

$$\text{sample standard error:} \quad S_{\overline{D}} = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D})^2/(n-1)}{n}}.$$

Paired t Test: Test Statistic

3. When the n sampled differences (D_i-values) are normally distributed, a one-sample t test ($\overline{X}$ versus μ_0) directly applies. So the sample test statistic comparing $\overline{X}$ to $\delta = 0$ becomes

$$\text{sample test statistic:} \quad \mathsf{T} = \frac{\overline{D}}{S_{\overline{D}}} \sim t_{n-1}. \tag{21}$$

Paired t Test: Test Statistic

$$\begin{aligned}
\mathbf{Var}(\overline{D}) &= \mathbf{Var}\Big(\frac{1}{n}\Big(\sum_i (X_{i1} - X_{i2})\Big)\Big) && (22)\\
&= \frac{1}{n^2}\mathbf{Var}\Big(\sum_i (X_{i1} - X_{i2})\Big) && (23)\\
&= \frac{1}{n^2}\Big(\sum_i \mathbf{Var}(X_{i1} - X_{i2})\Big) && (24)\\
&= \frac{1}{n^2}\Big(\sum_i (\mathbf{Var}(X_{i1}) + \mathbf{Var}(X_{i2}) - 2\mathbf{Cov}(X_{i1}, X_{i2})\Big) && (25)\\
&= \frac{1}{n}\Big(\mathbf{Var}(X_{i1}) + \mathbf{Var}(X_{i2}) - 2\mathbf{Cov}(X_{i1}, X_{i2})\Big) && (26)
\end{aligned}$$

Paired t Test: Test Statistic

1. Assume $X_{i1} \sim N(\mu_1, \sigma_1^2)$ and $X_{i2} \sim N(\mu_2, \sigma_2^2)$, and $\mathbf{Cov}(X_{i1}, X_{i2}) = \sigma_{12}$, for $i = 1, \ldots, n$,

$$\begin{aligned} \mathbf{Var}(\overline{D}) &= \frac{1}{n}\Big(\mathbf{Var}(X_{i1}) + \mathbf{Var}(X_{i2}) - 2\mathbf{Cov}(X_{i1}, X_{i2})\Big) \quad (27) \\ &= \frac{1}{n}\Big(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\Big) \quad (28) \end{aligned}$$

2. When we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$\mathbf{Var}(\overline{D}) = \frac{1}{n}\Big(2\sigma^2 - 2\sigma_{12}\Big) \quad (29)$$

Paired t Test: Test Statistic

1. Plug in sample variance,

$$\widehat{\mathbf{Var}}(\overline{D}) = \frac{1}{n}\left(S_1^2 + S_2^2 - 2\hat{\sigma}_{12}\right) \tag{30}$$

2. The test statistic becomes

$$\mathsf{T} = \frac{\overline{D}}{S_{\overline{D}}} = \frac{\frac{1}{n}\sum_i(\overline{X}_{i1} - \overline{X}_{i2})}{\sqrt{\frac{1}{n}\left(S_1^2 + S_2^2 - 2\hat{\sigma}_{12}\right)}} \tag{31}$$

Paired t Test: Test Statistic

When we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and σ_{12} are known, and $n \to \infty$

$$\mathbf{Var}(\overline{D}) = \frac{1}{n}\Big(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\Big) \tag{32}$$

The sample test statistic becomes

$$\begin{aligned} T^P &= \frac{\overline{D}}{S_{\overline{D}}} = \frac{\frac{1}{n}\sum_i (X_{i1} - X_{i2})}{\sqrt{\frac{1}{n}\Big(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\Big)}} && (33) \\ &= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n}\Big(2\sigma^2 - 2\sigma_{12}\Big)}}. && (34) \end{aligned}$$

Two-Sample t Test and Paired t Test

1. When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and we assume σ^2 is known, let $n \longrightarrow \infty$.

2. The two sample t test statistic is

$$T^U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n}\left(2\sigma^2\right)}}; \tag{35}$$

3. The paired t test statistic is

$$T^P = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n}\left(2\sigma^2 - 2\sigma_{12}\right)}}. \tag{36}$$

Two-Sample t Test and Paired t Test

1. When we have positive correlation, (or covariance) between X_{i1} and X_{i2}, the paired t test statistic (from paired design) has smaller variance, therefore, the paired t test has larger power.

2. Before we conduct, we need to consider whether there exists a positive correlation.

3. However, the paired design has "regression to the mean" problem, the significant difference between before and after treatment may arise from "regression to the mean", and the significant difference may not be true.

Screening Test and Diagnostic Test

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 11 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 8: 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 (TPR)**.

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

$$\mathbf{PV}^{+} = P[D+ \mid T+] = \frac{P[D+, T+]}{P[T+]} \tag{37}$$

$$= \frac{P[T+ \mid D+]P[D+]}{P[T+ \mid D+] \times P[D+] + P[T+ \mid D-] \times P[D-]} \tag{38}$$

$$\tag{39}$$

$$= \frac{\mathbf{sensitivity} \times P[D+]}{\mathbf{sensitivity} \times P[D+] + (1 - \mathbf{specificity}) \times P[D-]} \tag{40}$$

Medical Tests: Positive Predictive Value and Negative Predictive Value

$$\mathbf{PV}^{-} = P[D- \mid T-] = \frac{P[D-, T-]}{P[T-]} \tag{41}$$

$$= \frac{P[T- \mid D-]P[D-]}{P[T- \mid D+] \times P[D+] + P[T- \mid D-] \times P[D-]} \tag{42}$$

$$= \frac{\mathbf{specificity} \times P[D-]}{(1 - \mathbf{sensitivity}) \times P[D+] + \mathbf{specificity} \times P[D-]} \tag{43}$$

Medical Tests: Sample Data as 2×2 Table

The observed data is constructed as 2×2 table as in Table 9.

Table 9: Sensitivity and specificity: 2×2 Table

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

Sensitivity and Specificity: Point Estimation

The estimated mean sensitivity and specificity are

$$\widehat{\textbf{sensitivity}} = P[T+ \mid D+] = \frac{a}{a+c} \tag{44}$$

$$\widehat{\textbf{specificity}} = P[T- \mid D-] = \frac{d}{b+d} \tag{45}$$

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

$$\widehat{\mathbf{PV}}_\star^+ = P[D+ \mid T+] = \frac{a}{a+b} \tag{46}$$

$$\widehat{\mathbf{PV}}_\star^- = P[D- \mid T-] = \frac{d}{c+d} \tag{47}$$

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

$$
\begin{aligned}
& P(\text{Test result } \neq \text{ Disease Status}) \\
= \; & \text{Disease Prevalence} \times (1 - \text{Sensitivity}) \\
& \quad + (1 - \text{Disease Prevalence}) \times (1 - \text{Specificity}) \qquad (48) \\
& P(Y \neq D\ ) \\
= \; & P(D = 1)(1 - \text{Sen}) + (1 - P(D = 1))(1 - \text{Spe}); \qquad (49)
\end{aligned}
$$

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 Estradiol ≥ 20pg/ml as an abnormal (having breast cancer),

5. The data in Table 10. 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 10: 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

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

Example: Estradiol and Breaset Cancer

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

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

$$\begin{aligned}
\text{NPV}(\text{PV}-) &= \frac{(1 - \text{Sep}) \times (1 - P(D))}{(1 - \text{Sen}) \times P(D) + (1 - \text{Sep}) \times (1 - P(D))} \\
&= \frac{(1 - 0.965)(1 - 0.02)}{(1 - 0.089)0.02 + (1 - 0.965)(1 - 0.02)} = 0.651.
\end{aligned} \tag{51}$$

Example: Estradiol and Breaset Cancer

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%).

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 11 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 11: 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

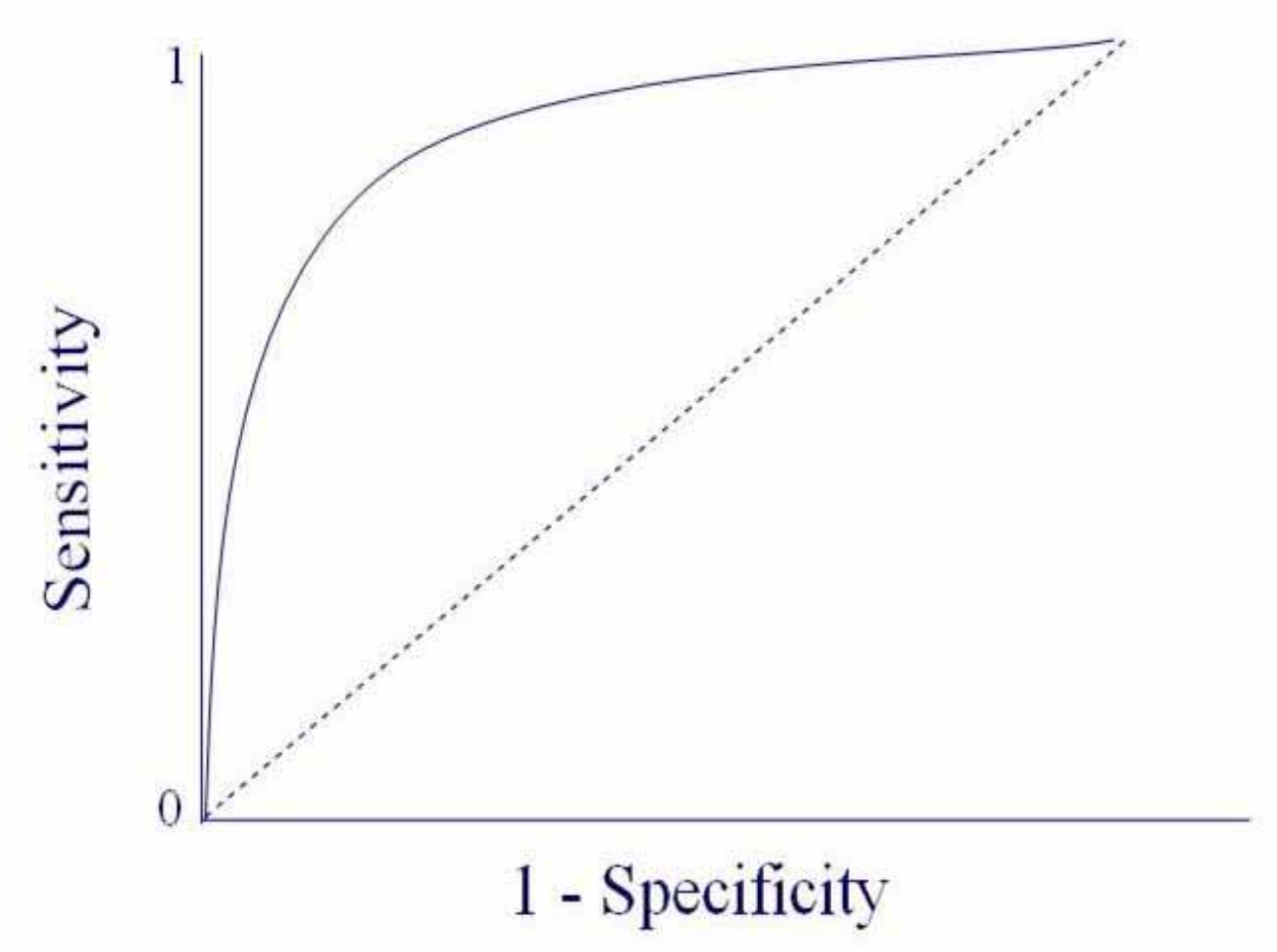

Figure 1: Receiver Operating Characteristic Curve

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 12: 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

$$\text{Sensitivity} = \frac{6}{213} = 0.028; \qquad \text{Sepecificity} = \frac{428}{432} = 0.990. \quad (52)$$

Example: Estradiol and Breaset Cancer

Cut-Off Point at Estridal ≥ 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 13: 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

$$\text{Sensitivity} = \frac{19}{213} = 0.089; \qquad \text{Sepecificity} = \frac{417}{432} = 0.965. \quad (53)$$

Example: Different Estradiol Cut-Off Points

Table 14: 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

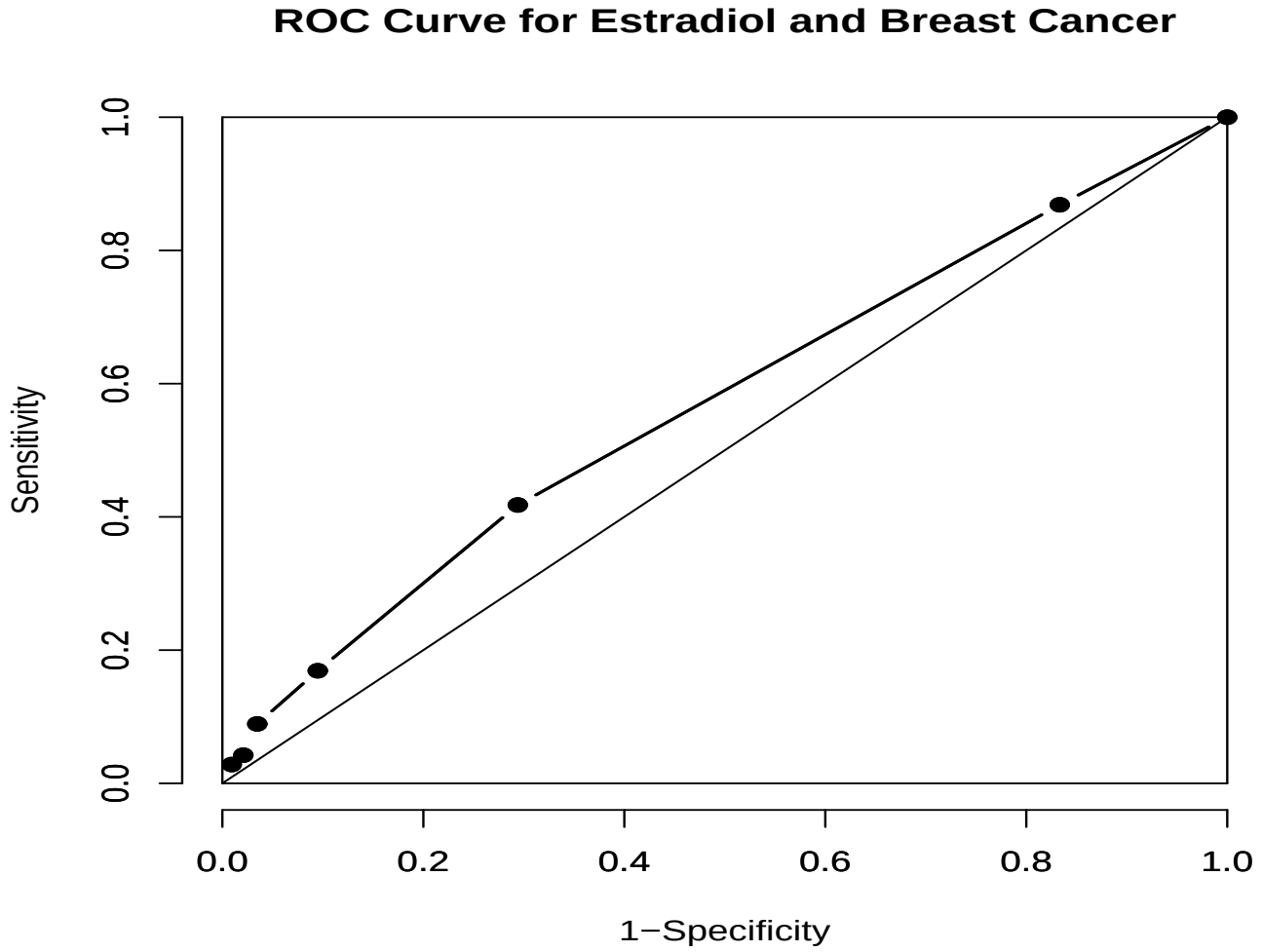

Figure 2: ROC Curve for Estradiol and Breaset Cancer

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 15: 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

$$\text{Sensitivity} = \frac{19}{213} = 0.089; \qquad \text{Sepecificity} = \frac{417}{432} = 0.965. \quad (54)$$

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

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

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

$$\begin{aligned}
\text{NPV(PV}-\text{)} &= \frac{(1 - \text{Sep}) \times (1 - P(D))}{(1 - \text{Sen}) \times P(D) + (1 - \text{Sep}) \times (1 - P(D))} \\
&= \frac{(1 - 0.965)(1 - 0.02)}{(1 - 0.089)0.02 + (1 - 0.965)(1 - 0.02)} = 0.651.
\end{aligned} \tag{55}$$

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 16: 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

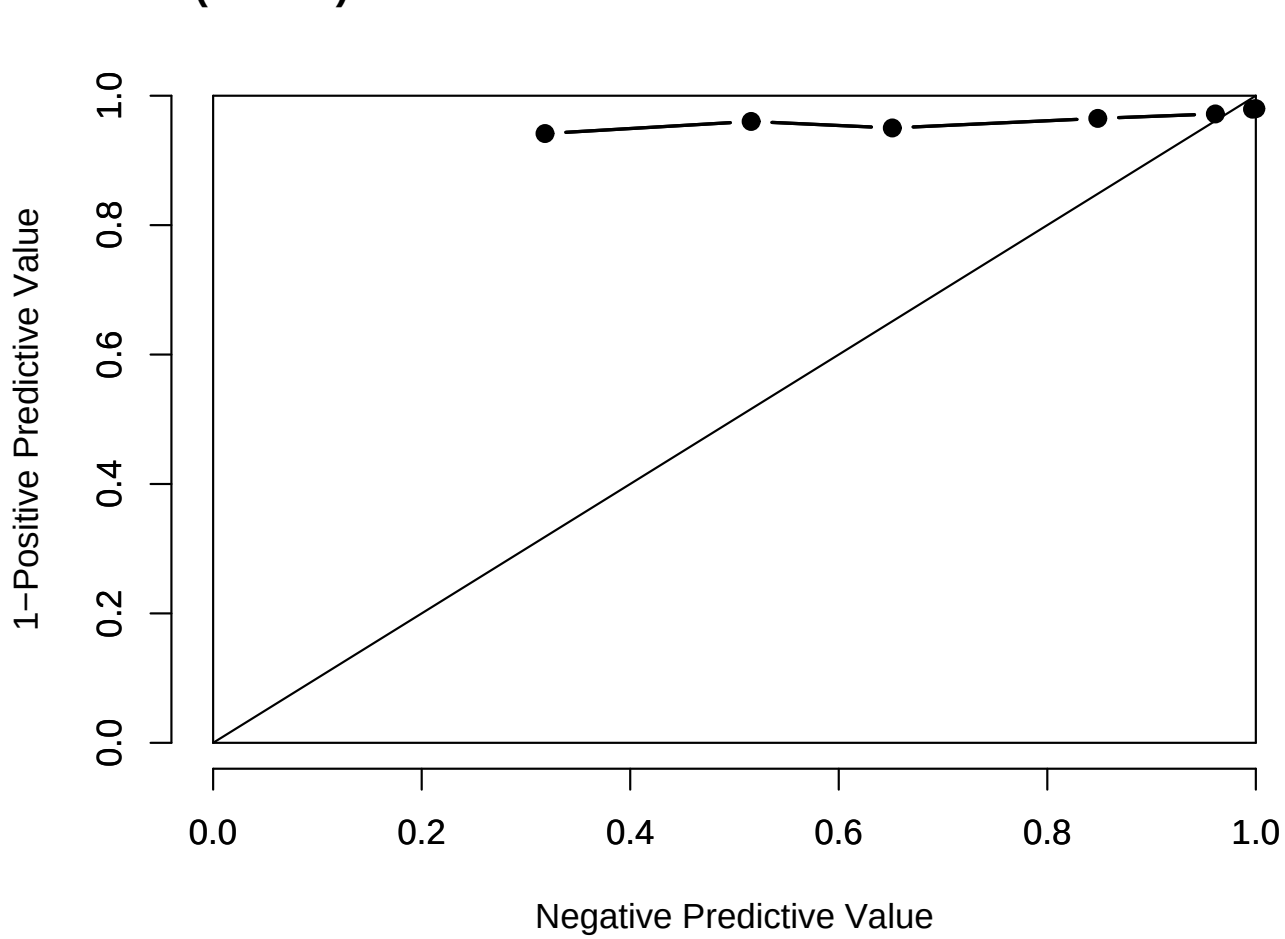

Figure 3: (1-PPV) versus NPV Curve for Estradiol and Breaset Cancer

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) #
```

Prevalence and Incidence

Prevalence and Incidence

1. **Prevalence** measures frequency of disease in a defined population at a specified point in time.

2. **Incidence** measures the frequency at which new disease is occurring in a defined population at risk over time.

Prevalence and Incidence

1. Fundamentally, **prevalence** is a static measure of disease frequency – a "snapshot" view, with time frozen.

2. Fundamentally, **incidence** is a dynamic measure of disease frequency – requires that people be observed over a period of time.

Prevalence and Incidence

1. Prevalence

   (a) (Point) Prevealence
   (b) Lifetime prevalence
   (c) Period Prevalence

2. Incidence

   (a) Cumulative incidence (incidence proportion)
   (b) Incidence density (incidence rate)

Prevalence

The "Point in time" at which prevalence is determined may refer to any of several time scales.

1. Calendar time – e.g., prevalence of HIV infections in Taiwan on January 1, 2005.

2. Age – e.g., prevalence of HIV infections among 20-year-old military recruits (regardless of whether they achieved age 20 in calendar time).

3. Time since some event – e.g., prevalence of depression among widows/widowers 6 months after the death of a spouse.

Point Prevalence

Point prevalence is defined as the proportion of a population affected by a disease at a given time point and expressed as a percentage.

$$\begin{aligned} &\text{point prevalence} \\ = &\frac{\text{number of cases of disease at a given time point}}{\text{population exposed (at risk) at a given time point}} \end{aligned} \tag{56}$$

Point Prevalence: Example

Lung cancer in a community, Jan 1, 1980:

Population	3,500,000
Cases	95,000
Prevalence	2.7%

Period Prevalence

Period Prevalence is defined as the proportion of a population affected by a disease during a time period and expressed as a percentage.

$$\begin{aligned} &\text{period prevalence} \\ = \quad &\frac{\text{number of cases of disease between a specific time period } (T_0, T_1)}{\text{population exposed during that time period } (T_0, T_1)} \end{aligned} \tag{57}$$

Period Prevalence: Example

Lung cancer in a community, Jan 1 – Dec 31, 1980:

Population	3,500,000	
Cases	96,250	(1250 new cases)
Prevalence	2.75%	

Cumulative Incidence

Cumulative Incidence is the proportion of the population will develop illness during the specified time period.

$$\begin{aligned} &\text{Cumulative Incidence (C.I.)} \\ = &\frac{\text{number of NEW cases of disease during a period}}{\text{population exposed during this period}} \end{aligned} \tag{58}$$

Cumulative Incidence: Example

Lung cancer in a community, Jan 1 – Dec 31, 1980:

Population	3,500,000	
Cases	96,250	(1250 new cases)
Cumulative incidence	0.36/1000	per year
Prevalence	2.71%	

Incidence Rate

1. Most subjectss in studies enter the study over a period of time, often over several years.

2. Others will become lost to contact during the follow-up period so that their information is not available at the end of the study.

3. The length of time of the study or follow-up will therefore not be the same for each subject.

4. This can be seen in Figure 4 below.

Incidence Rate

Example: Prevalence and Incidence Rate

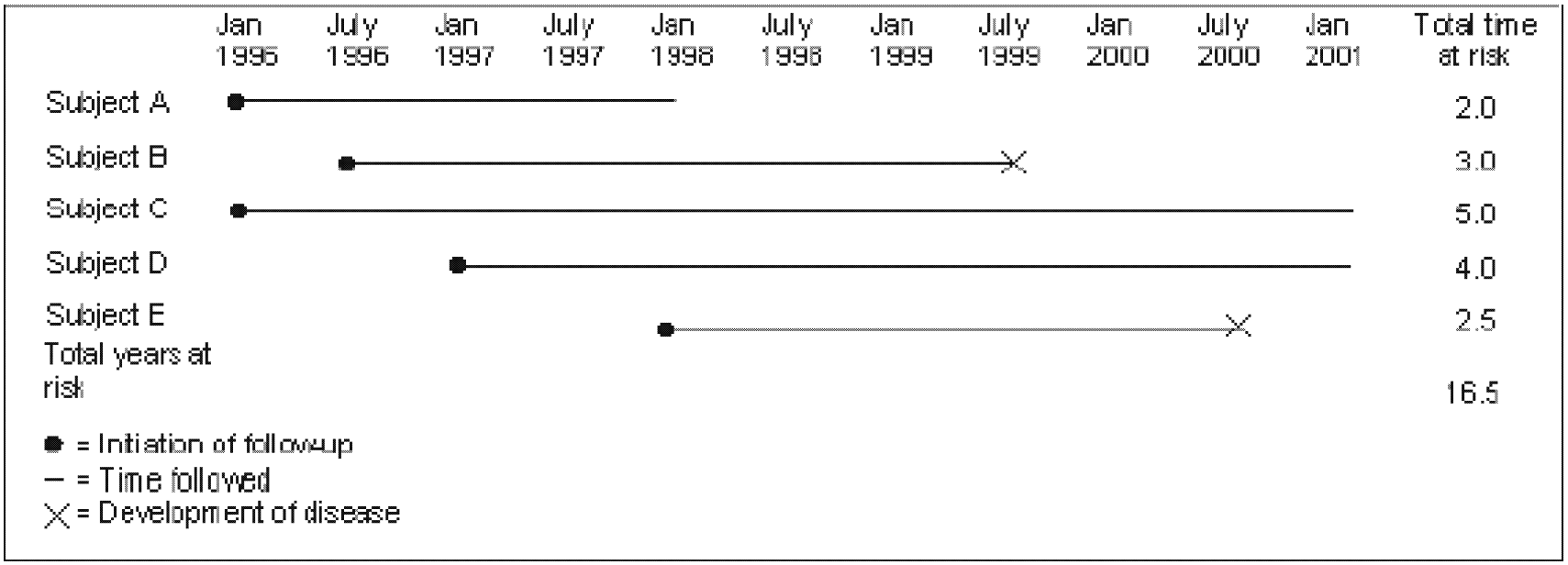

Figure 4: Incidence Rate and Follow-Up with Pearson-Time

Incidence Rate: Pearson-Time

Pearson-Years

1. Person-time is the sum of the amount of time each individual is observed while free of disease.

2. pearson-years is the sum of the amount of **years** each individual is observed while free of disease.

3. Each subject may contribute a different amount of person-years.

Incidence Rate: Pearson-Years

Person-time at risk is the denominator for incidence rates of disease

$$
\begin{aligned}
& 1000 \text{ person-years at risk} \\
= \; & 100{,}000 \text{ people for } 1/100 \text{ years} \qquad (59) \\
= \; & 10{,}000 \text{ people for } 1/10 \text{ years} \\
= \; & 1000 \text{ people for } 1 \text{ years} \\
= \; & 100 \text{ people for } 10 \text{ years} \\
= \; & 20 \text{ people for } 50 \text{ years}
\end{aligned}
$$

Incidence Rate: Pack-Years for Smoking

1×365 pack-year

$= 0.5 \times 365$ for 2 years

$= 2 \times 365$ for 0.5 years

Incidence Rate

An **incidence rate** (**incidence density**) is defined as the number of new cases of disease during a defined period of time, divided by the **total person-time** of observation.

$$\begin{aligned} &\text{Incidence Rate (I.R.)} \\ = \quad &\frac{\text{number of NEW cases of disease during a period}}{\text{total person-time of observation}} \end{aligned} \tag{60}$$

Example: Period Prevalence and Incidence Rate

Incidence and Pearson–years

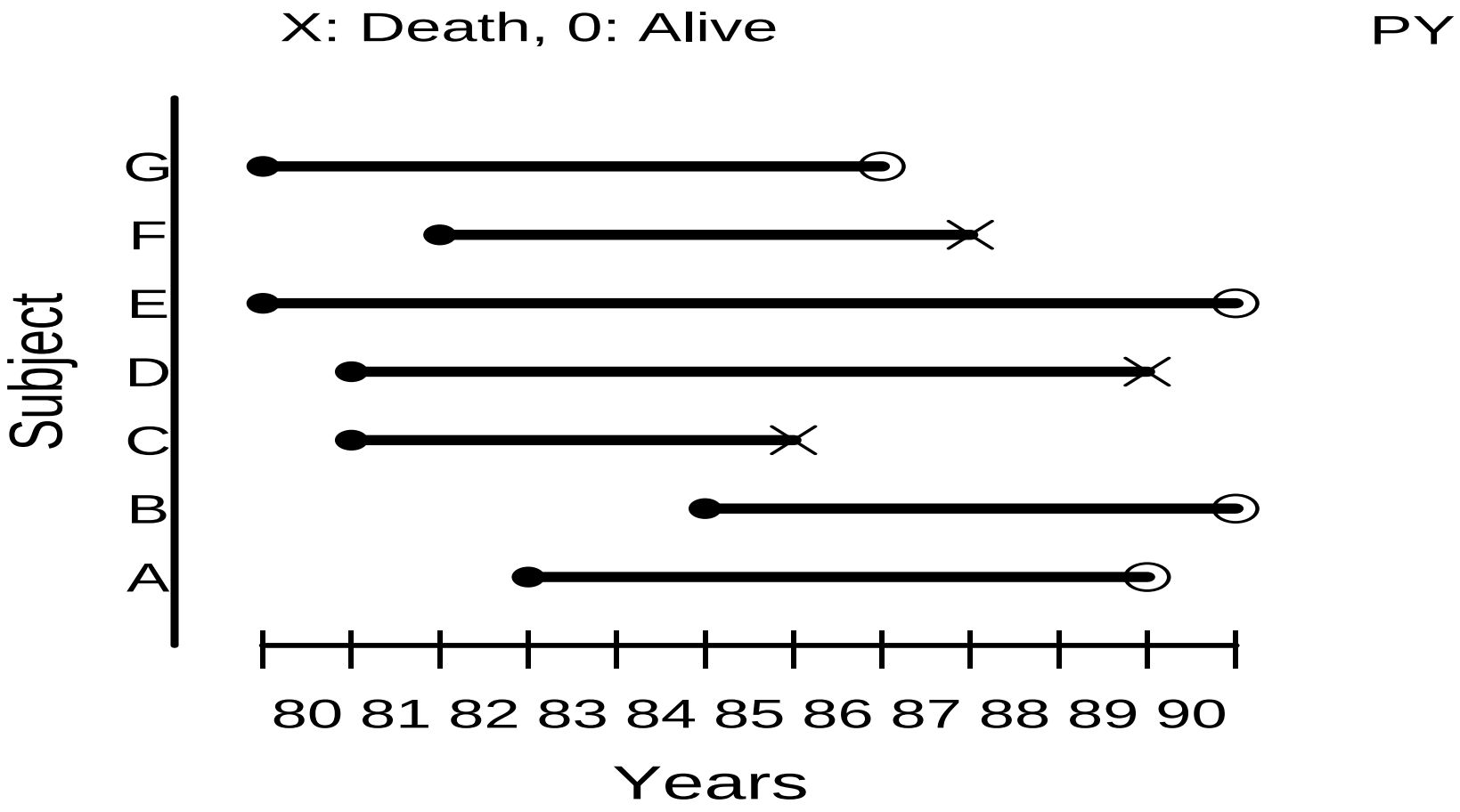

Figure 5: Incidence and Prevalence

Example: Period Prevalence and Incidence Rate

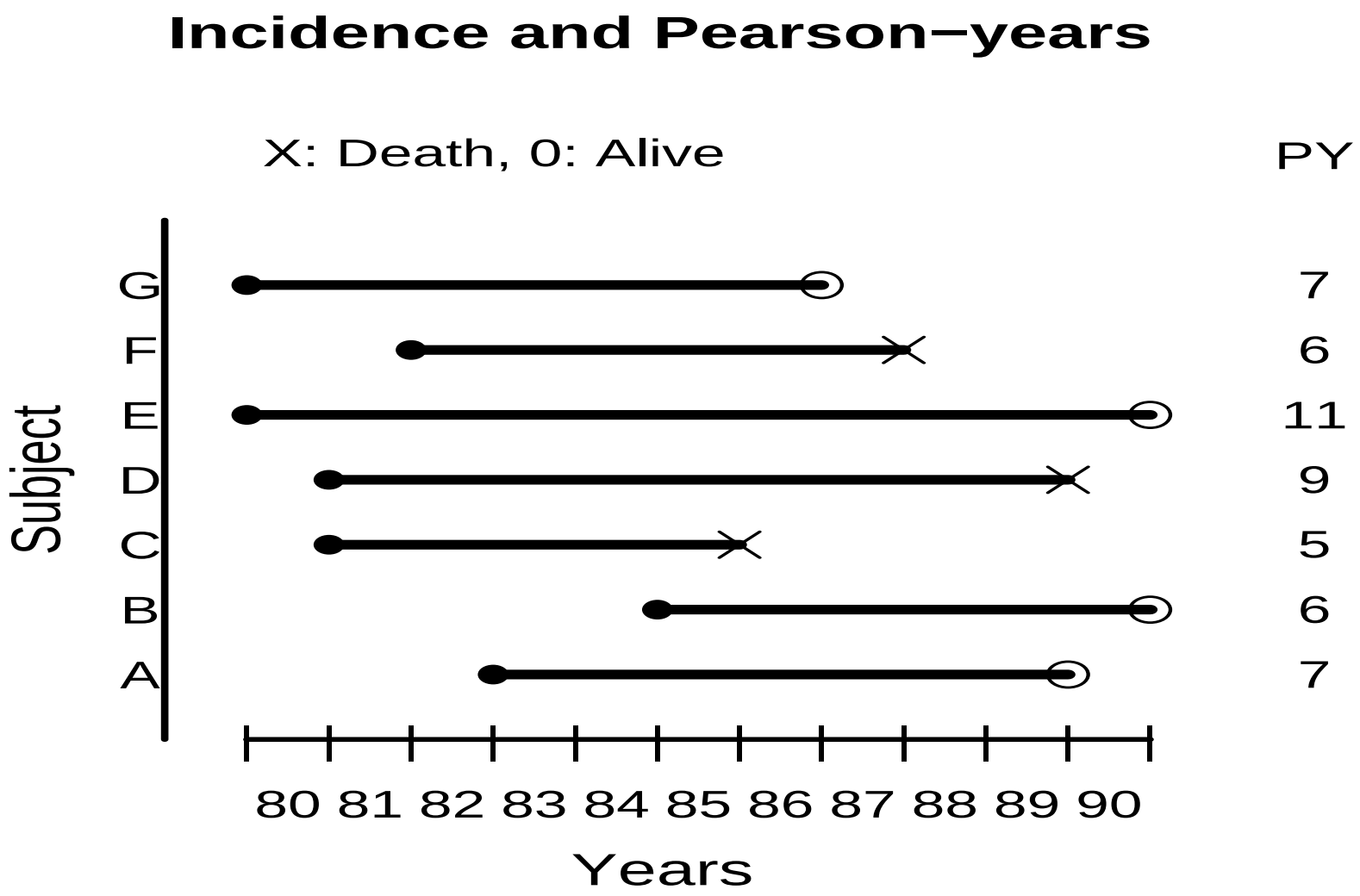

Figure 6: Incidence and Prevalence

Example: Period Prevalence and Incidence Rate

$$\text{Period Prevalence} = \frac{3}{7} = 0.428 \tag{61}$$

$$\text{Incidence Rate} = \frac{3}{\sum(7+6+\cdots+7)} = \frac{3}{51} = 0.058 \tag{62}$$

Risk

1. Risk is the probability of occurrence of an outcome in an outcome free population during a specified time period.

$$\begin{aligned} \text{If} \quad d &= \text{number of new cases} \\ \text{And} \quad N &= \text{population initially at risk,} \\ \text{Then,} \quad & \text{Risk (over a defined period)} = \frac{d}{N} \end{aligned} \tag{63}$$

2. Risk is usually applied to non-recurrent diseases or to the first episode of a disease.

Attack Rate

1. Attack Rate is defined as a cumulative incidence during an outbreak of a disease.

2. Usually expressed for the entire epidemic period, from the first to the last case.

Attack Rate

Outbreak of cholera in country X in March 1999

$$\begin{aligned} \text{Number of cases} &= 490 \\ \text{Population at risk} &= 18,600 \\ \text{Attack rate} &= 2.6\% \end{aligned}$$

Crude Rate and Specific Rate

Crude Rate

1. **Crude Death Rate** is the number of deaths in a year divided by the total population of interest.

2. The total population is estimated at the midpoint of the year.

$$\begin{array}{l}\text{Crude Death Rate} \\ = \quad \dfrac{\text{Number of deaths in a calendar year}}{\text{Population at midpoint of the year}}\end{array} \tag{64}$$

Specific Rate

1. Detailed understanding of disease experience in different population subsets

2. Homogeneous subgroups and detailed rates

3. Age-specific, sex-specific

4. Cumbersome in calculation

Age-Specific Rate

1. Age-specific rates: a rate used for a specific age group.

2. Numerator and denominator refer to the same age group.

$$\text{Age-Specific Rate} = \frac{\text{Total number of deaths from all causes in 1 yr per age group}}{\text{Number of subjects in the population at mid-year per age group}} \quad (65)$$

Crude Rate and Age-Specific Rate

Table 17: Crude Death Rate and Age-Specific Death Rate

Age Group (yrs)	Population	Deaths	Age-Specific Rate
00 – 04	97,870	383	3.9
05 – 09	221,452	75	0.3
10 – 24	284,956	440	1.5
25 – 34	265,885	529	2.0
35 – 44	207,564	538	2.6
45 – 54	193,505	1,107	5.7
55 – 64	175,579	2,164	12.3
65 – 74	152,172	3,789	24.9
≥ 75	107,114	7,834	73.1
Total (Crude)	1,706,097	16,859	9.9

Crude Rate and Age-Specific Rate
Comparing Mortality in Different Populations

1. Crude rate

2. Specific rate

3. Standardization rate

   (a) Direct (i.e., Age Adjustment)

   (b) Indirect (i.e., Standardized Mortality Ratio)

Crude Rate and Age-Specific Rate
Comparing Mortality in Different Populations

4. Cohort Analysis

5. Life-table Analysis

6. Median survival

7. Life expectancy

Comparing Mortality in Different Populations: Crude Rate and Age-Adjusted Rate

Table 18: Mortality Rate in Sweden and Panama

	Sweden			Panama		
	Mortality by Age-Group			Mortality by Age-Group		
Age	Deaths Number	Population Number	Death Rate	Deaths Number	Population Number	Death Rate
0-29	3,523	3,145,000	1.1	3,904	741,000	5.3
30-59	10,928	3,057,000	3.6	1,421	275,000	5.2
60+	59,104	1,294,000	45.7	2,956	59,000	50.1
All ages	73,555	7,496,000	9.8	8,281	1,075,000	7.7

Death Rate: per 10^3 pearson-years

Comparing Mortality in Different Populations: Crude Rate and Age-Adjusted Rate

1. The crude mortality rate in Sweden is large than that of Panama.

2. The age-secific mortality rates in Sweden are all smaller than those in Panama.

3. Why?

Comparing Mortality in Different Populations: Crude Rate and Age-Adjusted Rate

1. Comparisons with crude rate is suitable only when populations similar in all respects.

2. To account for these differences, **adjusted rates** are used in the comparison.

Comparing Mortality in Different Populations: Specific Rate

Disease rates in dissimilar populations can be compared by adjusting for known confounding factors (e.g. age).

Comapring Mortality by Standardization

Questions: What do we want? Answer: Want to compare rates.

1. Why we standardize?

2. What do we need? Standardized Population and Rate

3. How to standardize? Direct and Indirect

Comapring Mortality by Standardization

Why we standardize – Fleiss (1981, p. 240)

1. Comparing single (standardized) summaries easier than comparing tables.

2. If some age groups contain very small numbers, standardization can help.

3. For some subpopulations of interest, accurate age-specific rates may not exist.

Standardization: Age-Adjusted Rate

1. Direct adjustment:

   (a) Use data from a "standard population" to adjust rates

2. Indirect adjustment:

   (a) Adjust using "standard rates"
   (b) Determine expected rates
   (c) Compare observed and expected

Standardization: Age-Adjusted Rate

What Do We Need?

1. A **reference (standardized) population** with associated age structure.

2. Use proportion in each age class.

3. Examples: United States 1940, 1970, 2000 populations have all been used as standard population.

Standardization: Age-Adjusted Rate

What Do We Do With It?

- A standardized rate is a weighted average of age-specific values.

So, what are the weights?

Direct Standardization: Age-Adjusted Rate

1. Goal of adjustment
   - Reflect similar distributions
2. Reference population
   - Numbers used as weights to form weighted averages for both populations
   - Choice of reference population
3. Same set of numbers must be applied to both populations

Direct Standardization: Age-Adjusted Rate

1. Fictional!

2. Weighted average of age-specific rates

3. Directly comparable provided they refer to the same standard population

4. need age-specific rates for all observed populations

5. Standard population ("ideal" = world standard population)

Direct Standardization: Age-Adjusted Rate

Table 19: Direct Standardization Age-Adjusted Rate: Notation

	Observed i^{th} Population			Standardized Population		
	Mortality by Age-Group			Mortality by Age-Group		
Age	Deaths Number	Population Number	Death Rate	Deaths Number	Population Number	Death Rate
1	d_{i1}	n_{i1}	r_{i1}	d_{s1}	$n_{(s)1}$	$r_{(s)1}$
2	d_{i2}	n_{i2}	r_{i2}	$d_{(s)2}$	$n_{(s)2}$	$r_{(s)2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	d_{ij}	n_{ij}	r_{ij}	$d_{(s)j}$	$n_{(s)j}$	$r_{(s)j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
J	d_{iJ}	n_{iJ}	r_{iJ}	$d_{(s)J}$	$n_{(s)J}$	$r_{(s)J}$
Total	d_{i+}	n_{i+}	r_{i+}	$d_{(s)+}$	$n_{(s)+}$	$r_{(s)+}$

Direct Standardization (Age-Adjusted Rate): Notation

1. Suppose we the i^{th} observed population and j^{th} age-specific group, where $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

2. Let d_{ij} = number of death (cases) in age group j of population i.

3. Let n_{ij} = number at risk in age group j of population i.

4. Let

$$r_{ij} = \frac{d_{ij}}{n_{ij}} = \text{observed incidence proportion in age group } j \text{ of population } i.$$

Direct Standardization (Age-Adjusted Rate): Notation

5. Let $d_{(s)i}$, $n_{(s)j}$, and $r_{(s)j}$ be the same values for the standard population.

6. Let d_{i+}, n_{i+}, $d_{(s)+}$, and $n_{(s)+}$ be totals summed across all J age groups.

7. Let

$$r_{(s)j} = \frac{d_{(s)j}}{n_{(s)j}} \tag{66}$$

be the death rate of standardized population.

Direct Standardization (Age-Adjusted Rate): Notation

- Question: **How many cases** would we observe in the standard population if the observed age-specific rates applied?

- What do we need?

1. Obseved age-specific rates, r_{ij}

2. Number at risk in standardized population, $n_{(s)j}$

3. Total number observed in standardized population, $n_{(s)+}$

Direct Standardization (Age-Adjusted Rate)

1. The expected number of cases (death) for i^{th} observed population in each age group in the standardized population.

$$E_{i(s)j} = r_{ij} \times n_{(s)j} = \frac{d_{ij}}{n_{ij}} \times n_{(s)j}. \tag{67}$$

2. The overall expected number of cases (death) for i^{th} observed population in the standardized population

$$E_{i(s)+} = \sum_j E_{i(s)j} = \sum_j (r_{ij} \times n_{(s)j}) = \sum_j \left(\frac{d_{ij}}{n_{ij}} \times n_{(s)j} \right). \tag{68}$$

Direct Standardization (Age-Adjusted Rate)

3. The overall expected rate for i^{th} observed population in the standardized population

$$r_{i(s)E} = \frac{E_{i(s)+}}{n_{(s)+}} = \frac{\sum_j (r_{ij} \times n_{(s)j})}{n_{(s)+}} \tag{69}$$

$$= \frac{\sum_j \left(\frac{d_{ij}}{n_{ij}} \times n_{(s)j} \right)}{n_{(s)+}} \tag{70}$$

$$= \text{Weighted Average Rate with Weight } n_{(s)j}. \tag{71}$$

Direct Standardization (Age-Adjusted Rate): cumulative Mortality Figure (CMF)

1. cumulative mortality figure (CMF) is the ratio for comparing the number of cases and the expected death for i^{th} observed population in the standard population $(E_{i(s)+} = \sum_j E_{i(s)j})$ with number observed $(d_{(s)})$.

$$\text{CMF} = \frac{E_{i(s)+}}{d_{(s)+}} = \frac{\sum_j E_{i(s)j}}{\sum_j d_{(s)j}} \tag{72}$$

2. The directly standardized rate for i^{th} observed population is

$$r_{i(s)E} = \frac{E_{i(s)+}}{n_{(s)+}} = \text{CMF} \times \frac{d_{(s)+}}{n_{(s)+}} \tag{73}$$

Standardized (Age-Adjusted) Mortality Rate in Sweden

Table 20: Standardized (Age-Adjusted) Mortality Rate in Sweden

	Sweden			Standardized Population		
	Mortality by Age-Group			Mortality by Age-Group		
Age	Deaths Number	Population Number	Death Rate	Deaths Number	Population Number	Death Rate
0-29	3,523	3,145,000	1.1		56,000	
30-59	10,928	3,057,000	3.6		33,000	
60+	59,104	1,294,000	45.7		11,000	
All ages	73,555	7,496,000	9.8		100,000	

Death Rate: per 10^3 pearson-years

Standardized (Age-Adjusted) Mortality Rate in Sweden

Age-standardized (adjusted) mortality rate in Sweden

$$\begin{aligned} r_{i(s)E} &= \frac{(1.1 \times 56,000) + (3.6 \times 33,000) + (45.7 \times 11,000)}{100,000} \\ &= 6.8 \text{per 1000 person-years} \end{aligned} \tag{74}$$

Standardized (Age-Adjusted) Mortality Rate in Panama

Table 21: Standardized (Age-Adjusted) Mortality Rate in Sweden

	Panama			Standardzed Population		
	Mortality by Age-Group			Mortality by Age-Group		
Age	Deaths Number	Population Number	Death Rate	Deaths Number	Population Number	Death Rate
0-29	3,904	741,000	5.3		56,000	
30-59	1,421	275,000	5.2		33,000	
60+	2,956	59,000	50.1		11,000	
All ages	8,281	1,075,000	7.7		100,000	

Death Rate: per 10^3 pearson-years

Standardized (Age-Adjusted) Mortality Rate in Panama

Age-standardized (adjusted) mortality rate in Panama

$$\begin{aligned} r_{i(s)E} &= \frac{(5.3 \times 56,000) + (5.2 \times 33,000) + (50.1 \times 11,000)}{100,000} \\ &= 10.2 \text{per 1000 person-years} \end{aligned} \tag{75}$$

Comapring Age-Adjusted Rate: CMF Between Sweden and Panama

$$\text{Comparative Mortality Figure (CMF)} \quad = \quad \text{Age-Standarized Rate Ratio} \tag{76}$$

1. CMF for Panama versu Sweden

$$\text{CMF}_{P/W} = \frac{10.2}{6.8} = 1.5 \tag{77}$$

2. Mortality in Panama is 50% higher than in Sweden and excess is independent of age.

Indirect Standardization

1. Reverses role of study and standard populations.

2. Question: How many cases would we observe in the study population if the standard age-specific rates applied?

3. What do we need?

   (a) Standard age-specific rates, $r_{(s)j}$
   (b) Number at risk in the study observed population, n_{ij}
   (c) Total observed in study observed population, d_{i+}

Indirect Standardization

1. The number expected in each age group in the i^{th} study observed population

$$E_{ij} = r_{(s)j} \times n_{ij} = \frac{d_{(s)j}}{n_{(s)j}} \times n_{ij}. \tag{78}$$

2. The overall expected number in the i^{th} study observed population

$$E_{i+} = \sum_j (E_{ij}) = \sum_j (r_{(s)j} \times n_{ij}) = \sum_j \left(\frac{d_{(s)j}}{n_{(s)j}} \times n_{ij} \right) \tag{79}$$

Indirect Standardization: Standardized Mortality/Morbidity Ratio (SMR)

The **standardized mortality/morbidity ratio (SMR)** is the ratio of the total number observed in the i^{th} study area (d_{i+}) and the total number expected based on age-specific rates in the standard population (E_{i+}).

$$\textbf{SMR} = \frac{d_{i+}}{E_{i+}} = \frac{\text{Observed number of cases}}{\text{Expected number of cases}} \tag{80}$$

Indirect Standardization

Indirectly standardized rates is multiplying SMR by overall rate in standard population

$$\text{ISR} = \text{SMR} \times \frac{d_{(s)+}}{n_{(s)+}}. \tag{81}$$

Indirect Standardization

Indirect standardization is used when

- age-specific rates in one population is not known
- age-specific rates excessively variable because of small numbers

Indirect Standardization

Table 22: Mortality Rate in Sweden and Panama

	Sweden			Panama		
	Mortality by Age-Group			Mortality by Age-Group		
Age	Deaths Number	Population Number	Death Rate	Deaths Number	Population Number	Death Rate
0-29	3,523	3,145,000	1.1		741,000	
30-59	10,928	3,057,000	3.6		275,000	
60+	59,104	1,294,000	45.7		59,000	
All ages	73,555	7,496,000	9.8	8281	1,075,000	

Death Rate: per 10^3 pearson-years

Indirect Standardization

Calculate **expected deaths** in the Panamanian population (treat Sweden as standard population):

$$\begin{aligned} 0-29 \quad & 0.0011 \times 741,000 \quad = 814.5 \\ 30-59 \quad & 0.0036 \times 275,000 \quad = 990.0 \\ 60+ \quad & 0.0457 \times 59,000 \quad = 2696.3 \end{aligned}$$

Total expected deaths is 4501.4

Indirect Standardization

Table 23: Indrect Standardized Mortality Rate in Panama (Sweden as Standardized Population): Death Rate: per 10^3 pearson-years

	Sweden			Panama		
	Mortality by Age-Group			Mortality by Age-Group		
Age	Deaths Number	Population Number	Death Rate	Expected Deaths Number	Population Number	Death Rate
0-29	3,523	3,145,000	1.1	814.5	741,000	
30-59	10,928	3,057,000	3.6	990.0	275,000	
60+	59,104	1,294,000	45.7	2696.3	59,000	
All ages	73,555	7,496,000	9.8	4501.4	1,075,000	

Indirect Standardization

Standardized Mortality Ratio (SMR) of Paname

$$\text{SMR}_{P/S} = \frac{\text{Observed deaths}}{\text{Expected Death}} = \frac{O}{E} = \frac{8281}{501.4} = 1.84 = 184\% \quad (82)$$

Direct versus Indirect Standardization

1. CMF has observed (study) totals in the denominator and (directly) standardized expectations in the numerator.

2. By convention, SMR has observed (study) values in the numerator and (indirectly) standardized expectations in the denominator.

Direct versus Indirect Standardization

3. Direct standardization depends on the age structure of the standard population.

4. Indirect standardization depends on the age structure of the study population.

5. So, can compare directly standardized rates, if using same standard population.

Direct versus Indirect Standardization

6. Direct and indirect standardization produce identical (or at least proportional) results:

   (a) If age fractions are identical in study and standard populations.
   (b) If age-specific rates are identical in study and standard populations.
   (c) If age-specific rates in study population are proportional to those in standard population.

Direct versus Indirect Standardization

7. Direct standardization requires accurate assessment of age-specific incidence proportions for study population.

8. For rare diseases, $r_{ij} = \frac{d_{ij}}{n_{ij}}$ may be statistically unstable (add/remove single case can change r_{ij} dramatically).

9. Often, standard population is larger and $r_{(s)j} = \frac{d_{(s)j}}{n_{(s)j}}$ is more stable than r_{ij}.

Direct versus Indirect Standardization

10. Also, d_{ij} might not be as available as d_{i+} (age-specific counts not available but total count reported instead).

11. However, may have n_{ij} (e.g., from census).

12. In such cases, direct standardization not available (we don't have what we need).

Direct versus Indirect Standardization

13. Choice between direct and indirect standardization often reduces to the type of data available.

14. If age-specific incidence counts are available for the standard population but not the study population, indirect standardization is the only option available.

Direct versus Indirect Standardization

15. Must choose standard population.

16. Indirect standardization often assumes proportionality to allow comparability.

17. Direct standardization is inherently comparable across analyses using the same standard, but does not allow comparison across standards.

18. Krieger and Williams (2004) discuss the impact of changing from the US 1940 standard population to the US 2000 standard population (particularly with respect to measures of health disparity).

Standardization: Summary

1. Standardization seeks to summarize information across risk (age) groups, giving a single number (standardized count or rate).

2. Idea: remove variation due to known risk factors so remaining variation must be due to other (unknown) risk factors.

3. Inherent loss of information moving from age-specific to standardized rates.

Standardization: Summary

4. Standardization provide summaries (and as such give up some detail)

5. Direct standardization requires age-specific incidence from study population.

6. Indirect standardization requires age-specific incidence from standard population.

Standardization: Summary

7. Direct standardization comparable for same standard population.

8. Direct standardization not comparable across different standard populations.

9. Indirect standardization comparable only in special circumstances (proportionality assumption), but remember – all models are wrong, but some are useful (George Box).

Standardization: Reference

1. Fleiss, JL (2004) Statistical Methods for Rates and Proportions (3nd ed). NY: Wiley.

2. Inskip, H (1998) Standardization methods. In P. Armitage and T. Colton (eds.) Encyclopedia of Biostatistics, Chichester: Wiley. pp. 4237-4250.

3. Pickle, LW and White, AA (1995) Effect of the choice of age-adjustment on maps of death rates. Statistics in Medicine 14, 615-627.

4. Selvin, S (1991) Statistical Analysis of Epidemiologic Data. NY: Oxford University Press.

Standardization: Reference

5. Hennekens and Burin - Epidemiology in Medicine

6. Mausner and Kramer - Epidemiology: An Introductory Text

7. Bland - Introduction to Medical Statistics

Example: One

Table 24: Death numbers between two populations

Age	Pop. One		Pop. Two		Pop. Standard	
Group	Total	Death	Total	Death	Total	Death
00-20	10,000	50	20,000	90	300,000	1,400
21-40	10,000	80	10,000	70	200,000	1,500
41-60	20,000	100	30,000	50	500,000	1,500
60+	15,000	150	20,000	100	350,000	2,500
Total	55,000	380	80,000	310	1,350,000	6,900

Example: One

```
> setwd("C://temp//Rdata")
> # age  pop1 death1  pop2 death2    pop0 death0
> exam1<-read.csv("2005SMR01.csv", header = TRUE,
sep = ",", dec=".")
> exam1
  age  pop1 death1  pop2 death2    pop0 death0
1  20 10000     50 20000     90  300000   1400
2  40 10000     80 10000     70  200000   1500
3  60 20000    100 30000     50  500000   1500
4  80 15000    150 20000    100  350000   2500
5 100 55000    380 80000    310 1350000   6900
> attach(exam1)
```

Example: One

```
> ## Direct Age-Adjusted Rate pop1
> (r1<-death1/pop1)
[1] 0.005000000 0.008000000 0.005000000 0.010000000 0.006909091
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (r1.s<-r1*pop0)
[1] 1500.000 1600.000 2500.000 3500.000 9327.273
> # Expected # of Death in pop1
> (r1.s.E<- ( sum(r1.s[1:4])/pop0[5] )*10**6 )
[1] 6740.741
```

Example: One

```
> ## Direct Age-Adjusted Rate pop2
> (r2<-death2/pop2)
[1] 0.004500000 0.007000000 0.001666667 0.005000000 0.003875000
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (r2.s<-r2*pop0)
[1] 1350.0000 1400.0000  833.3333 1750.0000 5231.2500
> # Expected # of Death in pop2
> (r2.s.E<- ( sum(r2.s[1:4])/pop0[5] )*10**6 )
[1] 3950.617
```

Example: One

```
> ############################
> # CMF pop1 vs. pop2 ######
> (CMF12<-r1.s.E/r2.s.E)
[1] 1.70625
```

Example: One

```
> # Indirect adjusted Rate pop1
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (E1<-r0*pop1)
[1]  46.66667  75.00000  60.00000 107.14286 281.11111
> (E1.tot<-sum(E1[1:4]) )
[1] 288.8095
> # SMR pop1
> (SMR1<-death1[5]/E1.tot)
[1] 1.315746
```

Example: One

```
> # Indirect adjusted Rate pop2
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (E2<-r0*pop2)
[1]  93.33333  75.00000  90.00000 142.85714 408.88889
> (E2.tot<-sum(E2[1:4]) )
[1] 401.1905
> # SMR pop2
> (SMR2<-death2[5]/E2.tot)
[1] 0.7727003
```

Example: One

```
> # SMR pop1 vs. pop2
> (SMR1/SMR2)
[1] 1.702790
```

Example: Two

Table 25: Death numbers between two populations

Age	Pop. One		Pop. Two		Pop. Standard	
Group	Total	Death	Total	Death	Total	Death
00-20	10,000	50	20,000	90	30,0000	1,400
21-40	10,000	Unknown	10,000	70	200,000	1,500
41-60	20,000	100	30,000	50	500,000	1,500
60+	15,000	Unknown	20,000	100	350,000	2,500
Total	55,000	380	80,000	310	1,350,000	6,900

Example: Two

```
> exam2<-read.csv("2005SMR02.csv", header = TRUE, sep = ",",
  dec=".",as.is=TRUE) ## Special Care: as.is=TRUE
> exam2
  age  pop1 death1  pop2 death2    pop0 death0
1  20 10000    50   20000     90  300000   1400
2  40 10000  NA     10000     70  200000   1500
3  60 20000   100   30000     50  500000   1500
4  80 15000  NA     20000    100  350000   2500
5 100 55000   380   80000    310 1350000   6900
> attach(exam2)
```

Example: Two

```
> ##############################################
> # Unknown pop1 death rate
> # Only Indirect adjusted rate can be calculated
> ##############################################
```

Example: Two

```
> # Indirect adjusted Rate pop1
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (E1<-r0*pop1)
[1]  46.66667  75.00000  60.00000 107.14286 281.11111
> (E1.tot<-sum(E1[1:4]) )
[1] 288.8095
> # SMR pop1
> (SMR1<-death1[5]/E1.tot) # NOT WORK
Error in death1[5]/E1.tot : non-numeric argument to binary operator
> # You must use as.is=TRUE in read.csv
> (SMR1<-as.numeric(death1[5])/E1.tot)
[1] 1.315746
```

Example: Two

```
> # Indirect adjusted Rate pop2
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (E2<-r0*pop2)
[1]  93.33333  75.00000  90.00000 142.85714 408.88889
> (E2.tot<-sum(E2[1:4]) )
[1] 401.1905
> # SMR pop2
> (SMR2<-as.numeric(death2[5])/E2.tot)
[1] 0.7727003
```

Example: Two

```
> # SMR pop1 vs. pop2
> (SMR1/SMR2)
[1] 1.702790
```

Example: Three

Table 26: Death numbers between two populations

Age	Pop. One		Pop. Two		Pop. Standard	
Group	Total	Death	Total	Death	Total	Death
00-20	1,000	5	20,000	90	300,000	1,400
21-40	1,000	8	10,000	70	200,000	1,500
41-60	2,000	10	30,000	50	500,000	1,500
60+	1,500	15	20,000	100	350,000	2,500
Totals	5,500	38	80,000	310	1,350,000	6,900

Example: Three

```
> exam3<-read.csv("2005SMR03.csv", header = TRUE,
  sep = ",", dec=".")
> exam3
  age pop1 death1  pop2 death2    pop0 death0
1  20 1000      5 20000     90  300000   1400
2  40 1000      8 10000     70  200000   1500
3  60 2000     10 30000     50  500000   1500
4  80 1500     15 20000    100  350000   2500
5 100 5500     38 80000    310 1350000   6900
> attach(exam3)
```

Example: Three

```
> # Small number of death, unatable r1
> # Use indirect Adjusted Rate
```

Example: Three

```
> # Indirect adjusted Rate pop1
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (E1<-r0*pop1)
[1]  4.666667  7.500000  6.000000 10.714286 28.111111
> (E1.tot<-sum(E1[1:4]) )
[1] 28.88095
> # SMR pop1
> (SMR1<-death1[5]/E1.tot)
[1] 1.315746
```

Example: Three

Example: Three

```
> # Indirect adjusted Rate pop2
> (r0<-death0/pop0)
[1] 0.004666667 0.007500000 0.003000000 0.007142857 0.005111111
> (E2<-r0*pop2)
[1]  93.33333  75.00000  90.00000 142.85714 408.88889
> (E2.tot<-sum(E2[1:4]) )
[1] 401.1905
> # SMR pop2
> (SMR2<-death2[5]/E2.tot)
[1] 0.7727003
```

Example: Three

```
> # SMR pop1 vs. pop2
> (SMR1/SMR2)
[1] 1.702790
```

Example: Three

```
> # Consider pop2 as standardinzed population
>
>
> # Indirect adjusted Rate pop1
> (r2<-death2/pop2)
[1] 0.004500000 0.007000000 0.001666667 0.005000000 0.003875000
> (E1<-r2*pop1)
[1]  4.500000  7.000000  3.333333  7.500000 21.312500
> (E1.tot<-sum(E1[1:4]) )
[1] 22.33333
> # SMR pop1
> (SMR1<-death1[5]/E1.tot)
[1] 1.701493
```

Example: Three

```
> # Consider pop2 as standardinzed population
> # Indirect adjusted Rate pop2
> (r2<-death2/pop2)
[1] 0.004500000 0.007000000 0.001666667 0.005000000 0.003875000
> (E2<-r2*pop2)
[1]  90  70  50 100 310
> (E2.tot<-sum(E2[1:4]) )
[1] 310
> (E2.tot<-sum(death2[1:4]) )
[1] 310
> # SMR pop2
> (SMR2<-death2[5]/E2.tot)
[1] 1
```

Example: Three

```
> # SMR pop1 vs. pop2
> (SMR1/SMR2)
[1] 1.701493
```

Example: Three

```
> ### Consider pop1 as standardized pop
> # Indirect adjusted Rate pop1
> (r1<-death1/pop1)
[1] 0.005000000 0.008000000 0.005000000 0.010000000 0.006909091
> (E1<-r1*pop1)
[1]  5  8 10 15 38
> (E1.tot<-sum(E1[1:4]) )
[1] 38
> # SMR pop1
> (SMR1<-death1[5]/E1.tot)
[1] 1
```

Example: Three

```
> ### Consider pop1 as standardized pop
> # Indirect adjusted Rate pop2
> (r1<-death1/pop1)
[1] 0.005000000 0.008000000 0.005000000 0.010000000 0.006909091
> (E2<-r1*pop2)
[1] 100.0000  80.0000 150.0000 200.0000 552.7273
> (E2.tot<-sum(E2[1:4]) )
[1] 530
> # SMR pop2
> (SMR2<-death2[5]/E2.tot)
[1] 0.5849057
```

Example: Three

```
> # SMR pop1 vs. pop2
> (SMR1/SMR2)
[1] 1.709677
```