Standardization

何故標準化が必要か？

2 つ以上の populations の間で疾患の rates を比較する際、confounder が存在するかもしれません。何故なら年齢、性、人種、教育レベルなど疾患の発生に関与しうる他の因子が population によって異なった分布を示すかもしれないからです。例えば地区 A では大腸癌の発症頻度が地区 B より多いとします。しかしながら地区 A では壮年から老人が多いのに比較して、地区 B では若年が多いとしますと、fair には比較できません。ある工場で物質 X による癌 Y 発生の影響を検討しましたが、一般人口の癌 Y 発生より低い傾向にありました。しかしその工場では若い人ばかりが働いているかもしれません。そこで標準化(standardization)という計算が必要となってくるのです。標準化によって２つの population を比較することができるようになります。

標準化の方法

ある population を標準に選び、問題となる因子（年齢分布など）の比率を算出し(weight)、比較しようとする population も同じ比率であるとして算出します。

標準化には直説法と間接法があります。直接法の場合、まず１つの population を標準(standard)として選定し、例えば比較する population の年齢分布が標準と同じ分布を示すように計算しなおします。直接法ではしばしばいくつかの populations と比較します。通常標準として一般人口を用います。一方間接法では２つの population 間で比較することが多く、片方を標準(A)として、もう１つの population (B)を標準と同じ年齢分布であったと仮定して計算しなおします。例えば暴露された population の構成比に対して暴露されなかった population の rate ratio を掛け合わせて Expected number を算出し、Observed number を Expected number で割る事によって計算されます。間接法において、更にもう１つの population (C)と比較しようとすると、effect modifier が存在した場合、(B)と（C）は比較できないので注意が必要です。

標準化の実際

理論を言われてもピンとこないと思います。実例を検討しながら考えてみましょう。

ある精錬所で砒素に被爆した工場員の呼吸器癌による死亡率を調査しました。

年齢		1950-9
40-49	d	5
	n	14.949
	i	0.334
50-59	d	24
	n	10.223
	i	2.348
60-69	d	24
	n	4.896
	i	4.902
70-79	d	12
	n	1.851
	i	6.483
合計	d	65
	n	31.920
	i	2.036

n: x 1000, i: / 10^3 person years

下表はアメリカの人口分布を示す表です。

年齢	1950 年	呼吸器癌の割合
40-44	67712 x 10^6	0.21896 x 10^{-3}
45-49	60190 x 10^6	
50-54	54893 x 10^6	0.80277 x 10^{-3}
55-59	48011 x 10^6	
60-64	40210 x 10^6	1.55946 x 10^{-3}
64-69	33199 x 10^6	
合計	304215 x 10^6	

直接法および間接法を用いて標準化し、工場員が同地区の人々と比較して呼吸器癌にどれくらいなりやすいか計算してください。

まずは 1950 年の人口構成を比率で表してみましょう。
40 代：67712 + 60190 / 304215 = 0.42
50 代：54893 + 48011 / 304215 = 0.34
60 代：40210 + 33199 / 304215 = 0.24

この比率(weight)を用いて 1950-9 年、工場員の呼吸器癌の頻度(standardized rate: SR)を計算します。今までの例と異なり、今回は person-years で測定しているため、相当する年齢層一般人口に疾患発生率を掛け合わせるわけにはいきません。そういったわけで一般人口においても各年齢層の比率(weight)を算出しておくのです。

$SR_{worker} = \Sigma W_{0i} I_{1I} = 0.42 \times 0.334 \times 10^{-3} + 0.34 \times 2.348 \times 10^{-3} + 0.24 \times 4.902 \times 10^{-3}$
$= 2.115/1000$ PY

この比率(weight)を用いて 1950-9 年、工場と同じ地区における一般人口の呼吸器癌の頻度を計算します。

$SR_{gen} = \Sigma W_{0i} I_{1I} = 0.42 \times 0.21896 \times 10^{-3} + 0.34 \times 0.80277 \times 10^{-3} + 0.42 \times 1.55946 \times 10^{-3}$
$= 0.7392/1000$ PY

この工場での standardized rate ratio: SRR は

$SRR = SR_{worker} / SR_{gen} = 2.115 / 0.7392 = 2.86$

となり、工場員は同地区の人々と比較して呼吸器癌に 3 倍近くの頻度でなりやすいと言えます。

それでは間接法で標準化してみましょう。ポイントは精錬所で働いた人々が肺癌で死亡する率が一般と変わらないと期待します。そうすると精錬所の人数に単純に一般人口での肺癌死亡率を掛けて得られる人数は期待値となり、精錬所で肺癌が多発していなければ、精錬所で働いたあと肺癌で死亡した数は期待値より多くないはずです。

SMR (standard mortality ratio) = Observed /Expected =
(5 + 24 + 24) / (14.949 x 0.21896 + 10.223 x 0.80277 + 4.896 x 1.55946) = 2.77

直説法に近い値となりました。

95% CI = SMR ± $1.96\sqrt{\text{var}}$ (SMR)= SMR ± $1.96\sqrt{(O/E^2)}$
= 2.77 ± $1.96\sqrt{53}/19.12^2$ = (2.02, 3.52)

この 95%CI は 1.0 を含んでいないため、統計学的に有意差があると結論付けることができます。

次にそれぞれの年齢における rate ratio を計算してみてください。

40 代：0.334/0.219 = 1.53
50 代：2.348/0.80277 = 2.93
60 代：4.896/1.560 = 3.14

どうやら年齢は effect modifier のようです。何故なら年齢が上がるにつれ rate ratio は増加しているからです。病気の潜伏期を考えると当然かもしれません。

SRR(=2.86)と SMR(=2.77)の値は微妙に違います。何故でしょうか？

年齢分布がアメリカ一般人口と工場員とで微妙に違うからです。

	一般		工場
40 代:	0.42	vs.	0.50
50 代:	0.34	vs.	0.34
60 代:	0.24	vs.	0.16

呼吸器癌の頻度は 40 代＜50 代＜60 代の傾向があるのに対して、一般人口の方に比して工場は比較的若い世代が多くなっています。ということは一般人口を標準にした直接法（SRR）の方が工場人口を標準にした間接法（SMR）より大きい値をとることになります。SRR と SMR がどれくらい違うかは、effect modification の程度と weight の差によって決まります。この研究の場合これらの差は左程大きくなかったため、SRR と SMR の差も小さかったと考えられます。

この表はアスベストに暴露された作業員について調査した retrospective cohort study です。1948 年から 1963 年の間に 58 人が癌で死亡しています。もし US 全体の人口と同じ年齢、性、人種の構成であったとしたら下表より 42.9 人が癌になっているはずであり、アスベストに暴露された作業員の方が 1.35 倍癌になりやすいことが示されました。

	PY	アメリカ白人男性の癌死亡率（/10,000）	期待される癌死（E）
1948-1952			
15 – 24 歳	1250	9.9	0.1
25 - 34	3423	17.7	0.6
35 - 44	3275	44.5	1.5
45 - 54	2028	150.8	3.1
55 - 64	1144	409.4	4.7
1953-1957			
15 - 24	544	11.2	0.1
25 - 34	3702	17.5	0.6
35 - 44	4382	44.2	1.9
45 - 54	2968	157.7	4.7
55 - 64	1552	432.0	6.7
1958-1963			
15 - 24	4	10.3	0.0
25 - 34	2206	18.8	0.4
35 - 44	4737	46.3	2.2
45 - 54	4114	164.1	6.8
55 - 64	2098	450.9	9.5
合計			42.9

Enterline PE. Mortality among asbestos products workers in the United States. Ann NY Acad Sci 132:156,1965.

SMR を再現し 95% CI を計算してください。

SMR = O/E = 58/42.9 = 1.35

ここでは US 全体の癌の死亡率を使用しましたが、その中にはアスベストに暴露された人も含まれているので、理論的には過小評価することになります。もしアスベストに暴露されていないアメリカ白人男性の癌死亡率が判れば、より正確な値を得ることができるはずです。

95%CI を計算してみます。

95% CI = SMR ± $1.96\sqrt{\text{var}}$ (SMR)= SMR ± $1.96\sqrt{(O/E^2)}$
= 1.35 ± $1.96\sqrt{(58/42.9^2)}$ = (1.00, 1.70)
何とか１を超え有意差ありと判断できます。

クロリンを扱う工場で結膜炎が多いのではないかという危惧がありました。そこで A，B の 2 工場について若年、壮年と 2 つに分けで調査し、一般人口の頻度と比較してみたのが下の表です。

若年

	一般人口	工場 A	工場 B
結膜炎	50	50	5
Person-years	100,000	10,000	1,000
Incidence rate	0.0005	0.005	0.005

壮年

	一般人口	工場 A	工場 B
結膜炎	400	4	40
Person-years	200,000	1000	10,000
Incidence rate	0.002	0.004	0.004

２つの工場における rate ratio, SRR, SMR を計算してみてください。

それでは２つの工場における rate ratio を計算してみましょう。

工場 A において、結膜炎発症率は一般人口より何倍多いですか？
若年　RR = 0.005 / 0.0005 = 10 倍
壮年　RR = 0.004 / 0.002 = 2 倍

工場 B にについてはどうですか？
若年　RR = 0.005 / 0.0005 = 10 倍
壮年　RR = 0.004 / 0.002 = 2 倍

工場 A について直接 standardized rate ratio (SRR) を計算してみましょう。
まず Open cohort study なので、一般人口の年齢層の比率を出します。若年 10 万人と 20 万人、合計 30 万人で計算していますので、weight は若年で 1/3, 壮年で 2/3 となります。この weight を軸として incidence rate を計算し合計します。そうすれば工場 A、B の全体と一般人口の全体を fair に比較できます。

$SR_A = \Sigma W_{0i} I_{1i}$
= (100,000/300,000)(0.005 cases/py) + (200,000/300,000)(0.004 cases/py)
= 0.0043 cases /py

$SR_{gen} = \Sigma W_{0i} I_{0I}$
= = (100,000/300,000)(0.0005 cases/py) + (200,000/300,000)(0.002 cases/py)
= 0.0015 cases /py

$SRR_A = SR_A / SR_{gen} = 0.043 / 0.0015 = 2.87$

工場 B についてはどうでしょうか？

$SR_B = \Sigma W_{0i} I_{1i}$
= (100,000/300,000)(0.005 cases/py) + (200,000/300,000)(0.004 cases/py)
= 0.0043 cases /py

$SR_{gen} = \Sigma W_{0i} I_{0i}$
= = (100,000/300,000)(0.0005 cases/py) + (200,000/300,000)(0.002 cases/py)
= 0.0015 cases /py

$SRR_B = SR_B / SR_{gen} = 0.043 / 0.0015 = 2.87$

SRR_A とSRR_B は同じです。これは期待された結果ですか？最初の表で若年、壮年の工場A,Bにおける結膜炎の発症頻度は同じなのですから、weight がどうであれ同じはずです。

それでは間接的に測定してみましょう。間接的標準化はしばしば standardized morbidity rate ratio とも呼ばれるため SMR と訳されます。SMR は単純に観察された値を期待される値(工場で発生する結膜炎の割合が一般人口のそれと同じであると仮定した際の値)で割ることによって算出されます。

$SMR_A = O / E$
= (50 cases + 4 cases) / (10,000 py x 0.0005 cases/py + 1,000py x 0.002 cases/py)
= 54 / 7 = 7.71

$SMR_B = O / E$
= (5 cases + 40 cases) / (1,000 py x 0.0005 cases/py + 10,000py x 0.002 cases/py)
= 45 / (0.5 + 20) = 2.20

今度は値が大きく異なります。どのように考えますか？

Effect modification が存在するとき、異なる標準を使用することはできません。この例題では、共通のrate を用いましたがSMR_Aに対してはAの年齢person-years分布をSMR_Bに対してはBの年齢person-years分布を用いました。若年と壮年に層化(stratification)していますが、壮年で明らかに結膜炎の頻度が高くなっています(effect modification)。工場Aでは若年が 10,000 と壮年の 10 倍であるのに対して、工場Bでは壮年が 10,000 と若年の 10 倍であり、逆の関係になっています。このような場合、SMRを用いずSRRを用います。

以下 STATA を用いた標準化(standardization)の方法を紹介します。

コロラド州とルイジアナ州の乳幼児死亡率について調査してみました。まずデータを入力します。

```
. edit
- preserve

. list

        state        race       births      deaths
  1.  Colorado       black         3166          52
  2.  Colorado       white        48805         469
  3.  Colorado       other         1837           6
  4. Louisiana       black        29670         525
  5. Louisiana       white        42749         344
  6. Louisiana       other         1548           3
  7.       USA       black       641567       11461
  8.       USA       white      2992488       25810
  9.       USA       other       175339        1137
```

次に下記暗号をコンピュータに指示します。この暗号では、dstdize が直説法により標準化、deaths が分子、births が分母、race が標準化する因子を示しています。

```
. dstdize   deaths births race , by(state) base(USA)

-------------------------------------------------------------
-> state= Colorado
                              -----Unadjusted-----   Std.
                                    Pop.   Stratum   Pop.
  Stratum         Pop.       Cases   Dist. Rate[s] Dst[P]   s*P
-------------------------------------------------------------
    black         3166          52   0.059 0.0164   0.168 0.0028
    other         1837           6   0.034 0.0033   0.046 0.0002
    white        48805         469   0.907 0.0096   0.786 0.0075
-------------------------------------------------------------
Totals:          53808         527      Adjusted Cases:     563.1
                                            Crude Rate:    0.0098
                                         Adjusted Rate:    0.0105
                          95% Conf. Interval: [0.0094, 0.0115]
```

コロラド州における黒人の数は 3166 人であり、全体に占める割合は 5.9%、黒人の出生時死亡率は 1.64%です。一方 USA 全体における黒人の割合は 16.8%であり、この 2 者を掛け合わせると 0.28%となります。この 0.28%に黒人人口 3166 人を掛け、同様な計算を白人、その他の人種に対しても行ない合計すると Adjusted cases 563.1 人が算出されます。この数値は、もしコロラド州の人種構成が US 全体のものと同一であればコロラド州で生まれた 53808 人の新生児のうち 563.1 人が 1 歳までに死亡していたであろうことが予想されます。

```
-------------------------------------------------------------
-> state= Louisiana
                              -----Unadjusted-----   Std.
                                    Pop.   Stratum   Pop.
```

```
  Stratum          Pop.       Cases    Dist. Rate[s] Dst[P]    s*P
-------------------------------------------------------------
     black        29670         525    0.401 0.0177    0.168 0.0030
     other         1548           3    0.021 0.0019    0.046 0.0001
     white        42749         344    0.578 0.0080    0.786 0.0063
-------------------------------------------------------------
Totals:           73967         872      Adjusted Cases:      694.6
                                             Crude Rate:     0.0118
                                          Adjusted Rate:     0.0094
                              95% Conf. Interval: [0.0087, 0.0101]
```

ルイジアナ州に関しても同様の計算を行ないます。

```
-> state= USA
                               -----Unadjusted-----    Std.
                                      Pop.   Stratum    Pop.
  Stratum          Pop.       Cases    Dist. Rate[s] Dst[P]    s*P
-------------------------------------------------------------
     black       641567       11461    0.168 0.0179    0.168 0.0030
     other       175339        1137    0.046 0.0065    0.046 0.0003
     white      2992488       25810    0.786 0.0086    0.786 0.0068
-------------------------------------------------------------
Totals:         3809394       38408      Adjusted Cases:   38408.0
                                             Crude Rate:     0.0101
                                          Adjusted Rate:     0.0101
                              95% Conf. Interval: [0.0100, 0.0102]

Summary of Study Populations:
     state                N        Crude       Adj_Rate          Confidence Interval
 ------------------------------------------------------------------------------
 Colorado             53808     0.009794       0.010465      [    0.009449,      0.011482]
Louisiana             73967     0.011789       0.009391      [    0.008672,      0.010109]
        USA         3809394     0.010082       0.010082      [    0.009982,      0.010183]
```

Crude data ではルイジアナ州の方がコロラド州より乳児死亡率が高く見えましたが、標準化により得た adjusted data ではむしろコロラド州の方が高めで、ルイジアナ州の方が低い逆の結果となりました。もともと黒人の方が乳幼児死亡率が高く、ルイジアナ州では黒人人口が多いため、乳幼児死亡率が高くなっていたのでした。ルイジアナ州の小児科レベルが悪いわけではなさそうです。

1962 年のスエーデンとパナマの死亡率を比較しようと思います。

	スエーデン		パナマ	
年齢	人口	死亡	人口	死亡
0―29 歳	3,145,000	3,523	741,000	3,904
30―59 歳	3,057,000	10,928	275,000	1,421
60 歳以上	1,294,000	59,104	59,000	2,456

直感的にもパナマでは若いうちの死亡が多いことに気が付きます。
まずはデータを入力します。

```
. input cases exposed level time

        cases      exposed        level         time
  1. 3904 1 1 741000
  2. 3523 0 1 3145000
  3. 1421 1 2 275000
  4. 10928 0 2 3057000
  5. 2456 1 3 59000
  6. 59104 0 3 1294000
  7. end
```

まずは直接法で。

```
. ir cases exposed time, by(level) es

             level |        IRR        [95% Conf. Interval]          Weight
-------------------+-------------------------------------------------------
                 1 |   4.703267         4.49271    4.923845         3145000   (exact)
                 2 |   1.445493        1.366817    1.527792         3057000   (exact)
                 3 |    .911368         .874962    .9489276         1294000   (exact)
-------------------+-------------------------------------------------------
             Crude |   .7376398        .7205147    .7550885                   (exact)
 E. Standardized |     1.17234        1.139856    1.205749

. ir cases exposed time, by(level) es ird

             level |        IRD        [95% Conf. Interval]          Weight
-------------------+-------------------------------------------------------
                 1 |   .0041484         .003979    .0043177         3145000
                 2 |   .0015925        .0013156    .0018694         3057000
                 3 |  -.0040483       -.0057353   -.0023613         1294000
-------------------+-------------------------------------------------------
             Crude |  -.0025744       -.0027502   -.0023987
 E. Standardized |    .0016911        .0013999    .0019823
```

.Crude ではパナマの方が死亡率が低いことになってしまいました。何故なら死亡年齢を考慮に入れていないからです。実際単純計算するとスエーデンの死亡率は 9.8×10^{-3} でありパナマは 7.23×10^{-3} となり、パナマの死亡率の方が低いことになります。しかし年齢によって標準化することにより若干パナマの死亡率の方が高いことになりました。

次に間接法で。

. ir cases exposed time, by(level) is

level	IRR	[95% Conf. Interval]		Weight	
1	4.703267	4.49271	4.923845	741000	(exact)
2	1.445493	1.366817	1.527792	275000	(exact)
3	.911368	.874962	.9489276	59000	(exact)
Crude	.7376398	.7205147	.7550885		(exact)
I. Standardized	1.726055	1.685313	1.767783		

. ir cases exposed time, by(level) is ird

level	IRD	[95% Conf. Interval]		Weight
1	.0041484	.003979	.0043177	741000
2	.0015925	.0013156	.0018694	275000
3	-.0040483	-.0057353	-.0023613	59000
Crude	-.0025744	-.0027502	-.0023987	
I. Standardized	.0030447	.0029521	.0031373	

.間接法も同様です。