How can we solve the problem of confounding?

"Treatment " at statistical analysis

- ✓ Stratification by a confounder
- ✓ Multivariable / multiple analysis

Mantel-Haenszel odds ratio

- Stratification by confounding factor
  - After stratification by confounding factor, common OR, OR_{MH}, among all strata should be calculated.
  - Assumption: there is a common OR among all strata → there is no significant difference in ORs among all strata by homogeneity test.

An example of Mantel-Haenszel estimation 1

Calculate the common OR among all strata

smoking	Case	Control	
+	a_i	b_i	M_{1i}
-	c_i	d_i	M_{0i}
Total	N_{1i}	N_{0i}	T_i

$OR_c = \Sigma W_i OR_i \,/\, \Sigma w_i$

i :"i" th stratum、W_i :weight of "i" th stratum

Practice 1
Mantel-Haenszel odds ratio(1)

1. Open the "tsunagi_v1" data by excel
   - Please refer Appendix1 for the explanation of each variable.

2. Import this data set by your statistical software (STATA, R, and SPSS …)

Mantel-Haenszel odds ratio(2)

3. Suppose, you want to examine the cancer risk by habitual alcohol drinking.

- Please **create a contingency table** of cancer and alcohol drinking.

  STATA command: **tab alc cancer, row**

- Please **calculate an odds ratio**.

  STATA command: **cc cancer alc**
  or **cs cancer alc, or**

  Same OR but 95%CI is slightly different

Case-control study

```
. cc cancer alc
                                                         Proportion
                 |   Exposed   Unexposed  |      Total      Exposed
-----------------+------------------------+------------------------
           Cases |        79          77  |        156       0.5064
        Controls |       316         738  |       1054       0.2998
-----------------+------------------------+------------------------
           Total |       395         815  |       1210       0.3264
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         2.396104       |    1.678901    3.416628 (exact)
 Attr. frac. ex. |         .5826558       |    .4043721    .7073138 (exact)
 Attr. frac. pop |         .2950629       |
                 +-------------------------------------------------
                               chi2(1) =    26.38  Pr>chi2 = 0.0000
```

Cohort study

```
. cs cancer alc, or
                 | alc                    |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        79          77  |        156
        Noncases |       316         738  |       1054
-----------------+------------------------+------------
           Total |       395         815  |       1210
                 |                        |
            Risk |        .2    .0944785  |   .1289256
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .1055215       |    .0612577    .1497852
      Risk ratio |         2.116883       |    1.584051    2.828946
 Attr. frac. ex. |         .5276074       |    .3687071    .6465115
 Attr. frac. pop |         .2671858       |
      Odds ratio |         2.396104       |    1.706524    3.364381 (Cornfield)
                 +-------------------------------------------------
                               chi2(1) =    26.38  Pr>chi2 = 0.0000
```

Mantel-Haenszel odds ratio(3)

4. Since we know that cancer risk increases with age, you may want to confirm the association between alcohol drinking and cancer risk by age group (<60, 60-69, ≥ 70).

- Please *create contingency tables* of cancer

  STATA : **by age_gp, sort: tab alc cancer, row**

- Please *calculate odds ratios for each age group*.

An example of Mantel-Haenszel estimation 1

	age	alcohol	Case	Control	OR
1	<60	+	13	129	1.54
		-	14	214	1 (ref)
2	60-69	+	32	105	3.95
		-	19	246	1 (ref)
3	≥70	+	34	82	2.62
		-	44	278	1 (ref)
Total		+	79	316	2.40
		-	77	738	1

Mantel-Haenszel odds ratio(4)

5. Is there significant difference in the odds ratio among age groups?

6. Mantel-Haenszel test: homogeneity test

STATA : **cc cancer alc, by(age_gp)**

STATA commands

OR for each age group

```
. cc cancer alc, by(age_gp)

          age_gp |       OR       5% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               1 |   1.540421      .6436834   3.653387     4.881081 (exact)
               2 |   3.945865      2.055893   7.698361     4.962687 (exact)
               3 |   2.619734      1.514075   4.493878     8.237443 (exact)
-----------------+-------------------------------------------------
           Crude |   2.396104      1.678901   3.416628              (exact)
    M-H combined |   2.692348      1.901164   3.812791
-------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     3.45  Pr>chi2 = 0.1782

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     32.88
                                                Pr>chi2 =    0.0000
```

OR_{MH}

Homogeneity test

→ no significant (you can calculate common OR!)

Mantel-Haenszel odds ratio(5)

You can also calculate OR_{MH} by yourself.

$$OR_{MH} = \Sigma(a_i * d_i / T_i) / \Sigma(b_i * c_i / T_i)$$

$$OR_{MH} = \frac{(13*214/370) + (32*246/402) + (34*278/438)}{(129*14/370) + (105*19/402) + (82*44/438)}$$

$$= 2.69$$

Practice 2

1. Using the "tsunagi_v1" data set, please examine the association between habitual alcohol drinking and cancer risk by sex stratification.

```
. cc cancer alc, by(male)

                 male |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
                    0 |   .9455128      .3977241    2.01076      7.399209 (exact)
                    1 |   1.992308      1.179265    3.441301     11.52993 (exact)
-----------------+-------------------------------------------------
                Crude |   2.396104      1.678901    3.416628              (exact)
         M-H combined |   1.583126      1.061639    2.360773
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     2.68  Pr>chi2 = 0.1014

                  Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      4.86
                                                Pr>chi2 =    0.0276
```

Q1. Is this OR_{MH} statistically significant?

Q2. Is it OK to report OR_{MH} when the homogeneity test is statistically significant?

How can we solve the problem of confounding?

"Treatment " at statistical analysis

- ✓Stratification by a confounder
- ✓Multivariable / multiple analysis

Logistic Regression Analysis

Practice 3
Multivariable analysis

1. Let's see the association between habitual alcohol drinking and cancer risk by logistic regression model.

STATA :	**logistic cancer alc**
or	**logit cancer alc, or**

2. Please examine this association adjusting for the effects of age and sex.

STATA : **logistic cancer alc male age**

```
. logistic cancer alc male age

Logistic regression                               Number of obs   =       1210
                                                  LR chi2(3)      =      67.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -431.32695                       Pseudo R2       =     0.0725

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.877452   .3864541     3.06   0.002     1.254174    2.810476
        male |   2.099375   .4306218     3.62   0.000     1.404405    3.138251
         age |   1.041058   .0093757     4.47   0.000     1.022844    1.059597
------------------------------------------------------------------------------
```

STATA : **logistic cancer alc male age_gp**

```
. logistic cancer alc male age_gp

Logistic regression                               Number of obs   =       1210
                                                  LR chi2(3)      =      66.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -431.86621                       Pseudo R2       =     0.0714

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.825007   .3736455     2.94   0.003     1.221779    2.726067
        male |   2.198312   .4481714     3.86   0.000     1.474193    3.278117
      age_gp |   1.669985   .1951521     4.39   0.000     1.328136    2.099824
------------------------------------------------------------------------------
```

STATA : **xi: logistic cancer alc male i.age_gp**

Categorical variable (>2 categories)

```
. xi: logistic cancer alc male i.age_gp
i.age_gp          _Iage_gp_1-3        (naturally coded; _Iage_gp_1 omitted)

Logistic regression                               Number of obs   =       1210
                                                  LR chi2(4)      =      66.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -431.81689                       Pseudo R2       =     0.0715

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         alc |   1.825179   .3735059     2.94   0.003     1.222123    2.725812
        male |   2.192979   .4473008     3.85   0.000     1.470332    3.270798
  _Iage_gp_2 |   1.792761   .4573496     2.29   0.022     1.087359    2.955776
  _Iage_gp_3 |    2.84385   .6937492     4.28   0.000     1.763025    4.587276
------------------------------------------------------------------------------
```

If there is no linear trend of the cancer risk by age,
it would be better to use categorical variable for age.

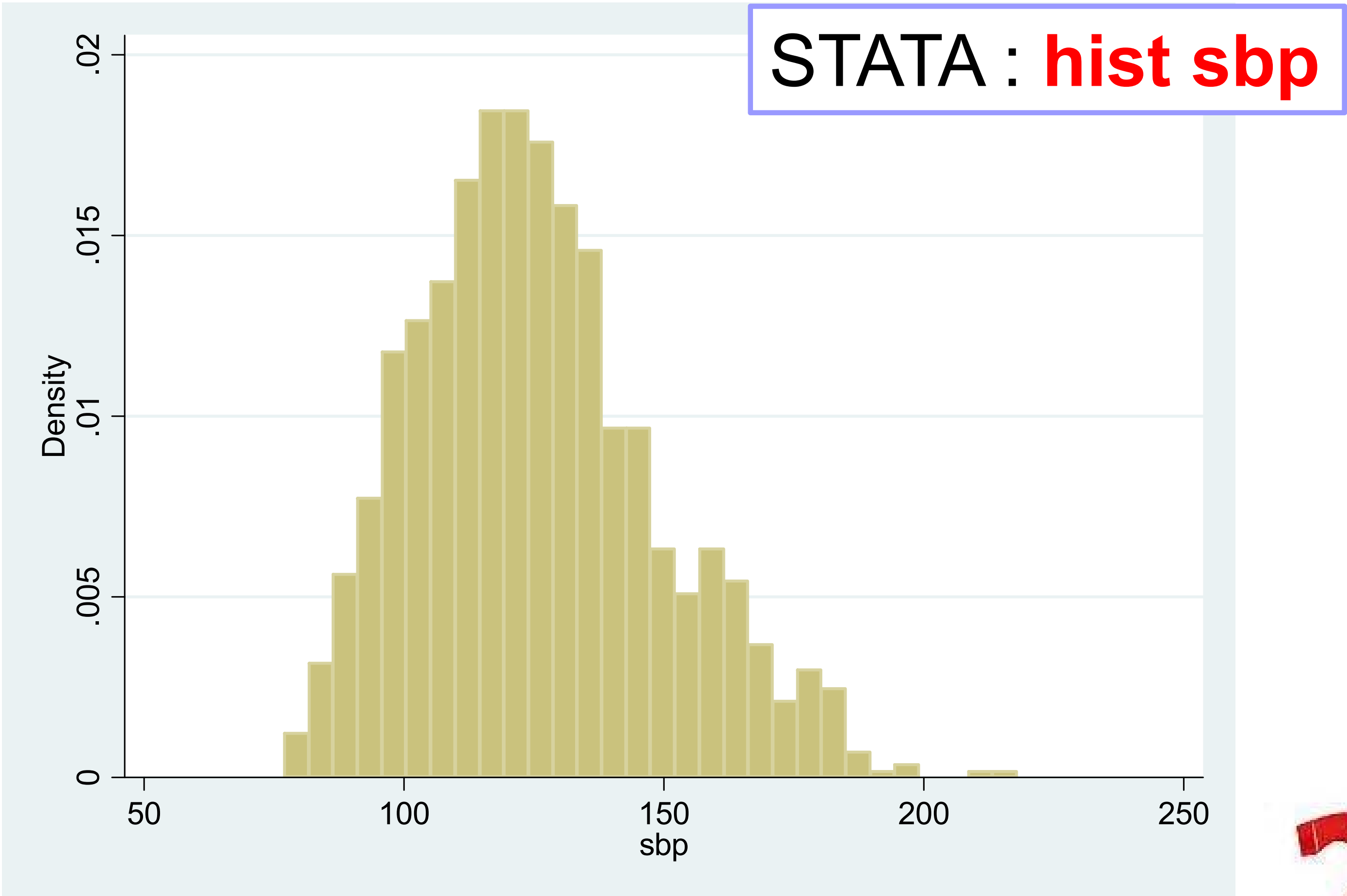
REGRESSION ANALYSIS

Practice 4
Regression analysis (1)

- Suppose, you want to know predictors of **systolic blood pressure** in the subjects of "tsunagi_v1" data.

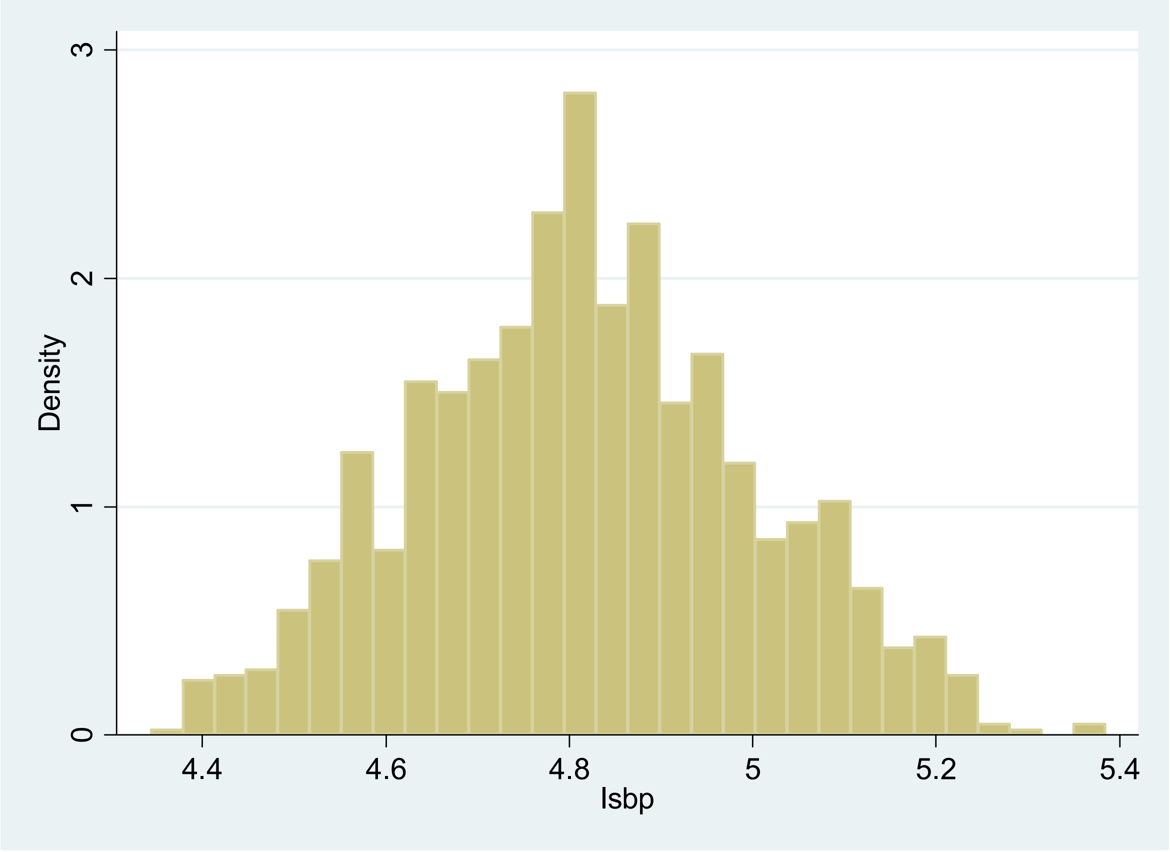
- What do you have to check first?

Distribution of systolic blood pressure



Log-transformation may work…

STATA : **gene lsbp=log(sbp)**
hist lsbp



Practice 4
Regression analysis (2)

- Age is one of the predictors of systolic blood pressure.

- Please conduct regression analysis using "age" as a explanatory variable.

STATA : **reg lsbp age**

STATA commands

```
. reg lsbp age

      Source |       SS       df       MS              Number of obs =    1210
-------------+------------------------------           F(  1,  1208) =   91.90
       Model |  2.84842138     1  2.84842138           Prob > F      =  0.0000
    Residual |  37.4400935  1208  .030993455           R-squared     =  0.0707
-------------+------------------------------           Adj R-squared =  0.0699
       Total |  40.2885149  1209  .033323834           Root MSE      =  .17605

------------------------------------------------------------------------------
        lsbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0044274   .0004618     9.59   0.000     .0035213    .0053334
       _cons |   4.531922   .0301251   150.44   0.000     4.472819    4.591026
------------------------------------------------------------------------------
```

SBP= 4.531922 + **0.0044274***age

Practice 4
Regression analysis (3)

- Please transform age variable into 10-year age group.

STATA : **gene age10=floor(age/10)**

- Let's see the association between age and systolic blood pressure using this variable (age10).
- What do you expect?

```
. reg lsbp age10

      Source |       SS       df       MS              Number of obs =    1210
-------------+------------------------------           F(  1,  1208) =   90.90
       Model |  2.8196114     1  2.8196114           Prob > F      =  0.0000
    Residual |  37.4689035  1208  .031017304           R-squared     =  0.0700
-------------+------------------------------           Adj R-squared =  0.0692
       Total |  40.2885149  1209  .033323834           Root MSE      =  .17612

------------------------------------------------------------------------------
        lsbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age10 |   .0431853   .0045294     9.53   0.000     .0342989    .0520717
       _cons |   4.557933      .0276   165.14   0.000     4.503784    4.612083
------------------------------------------------------------------------------
```

SBP= 4.557933 + **0.0431853***age(10)

cf. SBP= 4.531922 + **0.0044274***age

Practice 4
Regression analysis (4)

- Suppose, hemoglobin level may be one of the predictors of systolic blood pressure.
- Please pick-up other potential predictors (other than hemoglobin) for systolic blood pressure in this data set based on your knowledge.
- And, conduct regression analysis.