Bivariate Data: Graphical Display

The scatterplot is the basic tool for graphically displaying bivariate quantitative data.

Example:

Some investors think that the performance of the stock market in January is a good predictor of its performance for the entire year. To see if this is true, consider the following data on Standard & Poor's 500 stock index (found in SASDATA.SANDP).

Year	Percent January Gain	Percent 12 Month Gain
1985	7.4	26.3
1986	0.2	14.6
1987	13.2	2.0
1988	4.0	12.4
1989	7.1	27.3
1990	-6.9	-6.6
1991	4.2	26.3
1992	-2.0	4.5
1993	0.7	7.1
1994	3.3	-1.5

The plot your instructor is about to show you is a scatterplot of the percent gain in the S&P index over the year (vertical axis) versus the percent gain in January (horizontal axis).

How to analyze a scatterplot

The scatterplot of the S&P data can illustrate the general analysis of scatterplots. You should look for:

- Association. This is a pattern in the scatterplot.
- Type of Association. If there is association, is it:
  - Linear.
  - Nonlinear.
- Direction of Association.

For the S&P data, there is association. This shows up as a general positive relation (Larger % gain in January is generally associated with larger % yearly gain.) It is hard to tell if the association is linear, since the spread of the data is increasing with larger January % gain. This is due primarily to the 1987 datum in the lower right corner of plot, and to some extent the 1994 datum. Eliminate those two points, and the association is strong linear and positive, as the second plot shows.

There is some justification for considering the 1987 datum atypical. That was the year of the October stock market crash. The 1994 datum is a mystery to me.

TYU 6

Correlation
Pearson Correlation

Suppose n measurements, $(X_i,\ Y_i),\ i = 1, \ldots, n$ are taken on the variables X and Y. Then the Pearson correlation between X and Y computed from these data is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} X_i' Y_i',$$

where

$$X_i' = \frac{X_i - \overline{X}}{S_X} \text{ and } Y_i' = \frac{Y_i - \overline{Y}}{S_Y}$$

are the standardized data.

The following illustrate what Pearson correlation measures.

Good Things to Know About Pearson Correlation

- Pearson correlation is always between -1 and 1. Values near 1 signify strong positive linear association. Values near -1 signify strong negative linear association. Values near 0 signify weak linear association.
- Correlation between X and Y is the same as the correlation between Y and X.

- Correlation can never by itself adequately summarize a set of bivariate data. Only when used in conjunction with appropriate summary measures for X and Y (such as $\overline{X}$, $\overline{Y}$, S_X, and S_Y if X and Y are normally distributed) **and a scatterplot** can an adequate summary be obtained.
- The meaningfulness of a correlation can only be judged with respect to the sample size.

A Confidence Interval for the Population Correlation, ρ

If n is the sample size,

$$t = (r - \rho)\sqrt{\frac{n-2}{(1-r^2)(1-\rho^2)}}$$

has approximately a t_{n-2} distribution. We can use this fact to obtain a confidence interval for ρ.

Example:

Back to the S&P data, the SAS macro CORR gives a 95% confidence interval for ρ as $(-0.2775,\ 0.8345)$. As this interval contains 0, it indicates no significant linear association between JANGAIN and YEARGAIN. If we remove the 1987 and 1994 data, a different story emerges. Then the Pearson correlation is $r = 0.9360$, and a 95% confidence interval for ρ is (0.6780, 0.9880). Since this interval consists entirely of positive numbers, we conclude that ρ is positive and we estimate its value to be between 0.6780 and 0.9880.

Under $H_0 : \rho = \rho_0$, the test statistic

$$t = (r - \rho_0)\sqrt{\frac{n-2}{(1-r^2)(1-\rho_0^2)}}$$

has a t_{n-2} distribution. We can use this to conduct hypothesis tests. If t^* is the observed value of t,
For $H_{a^+} : \rho > \rho_0$, the p-value is $p^+ = P(t \geq t^*)$; For $H_{a_-} : \rho < \rho_0$, the p-value is $p^+ = P(t \leq t^*)$; For $H_{a\pm} : \rho \neq \rho_0$, the p-value is $p\pm = 2\min(p_-, p^+)$.

As an example, for the S&P data we test

$$\begin{aligned} H_0 : &\quad \rho &= &\quad 0 \\ H_{a\pm} : &\quad \rho &\neq &\quad 0 \end{aligned}$$

by computing

$$t^* = 0.4295\sqrt{\frac{8}{(1-0.4295^2)}} = 1.3452,$$

and comparing this with a t_8 distribution. The resulting values are

$$p^+ = P(t_8 > 1.3452) = 0.1077,$$

and

$$p_- = P(t_8 < 1.3452) = 0.8923,$$

so that

$$p = 2\min(0.1077, 0.8923) = 0.2154.$$

TYU 7

Simple Linear Regression

The SLR model attempts to quantify the relationship between a single predictor variable Z and a response variable Y. This reasonably flexible yet simple model has the form

$$Y = \beta_0 + \beta_1 X(Z) + \epsilon,$$

where ϵ is a random error term, and $X(Z)$ is a function of Z, such as Z, Z^2, or $\ln(Z)$.

By looking at different functions X, we are not confined to linear relationships, but can also model nonlinear ones. The function X is called the **regressor**. Often, we omit specifying the dependence of the regressor X on the predictor Z, and just write the model as

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Example: An experiment was conducted on the effect of number of days of training received (Z) on performance (Y) in a battery of simulated sales presentations. The data are found in sasdata.knn_ex.

Model Fitting

The term "model fitting" refers to using data to estimate model parameters. We will fit the simple linear regression model to a set of data (X_i, Y_i), $i = 1, \ldots, n$. As with the C+E model, two options (there are others as well) are least absolute errors, which finds values b_0 and b_1 to minimize

$$\text{SAE}(b_0, b_1) = \sum_{i=1}^{n} \mid Y_i - (b_0 + b_1 X_i) \mid,$$

or least squares, which finds values b_0 and b_1 to minimize

$$\text{SSE}(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2.$$

We'll concentrate on least squares. Using calculus, we find the least squares estimators of β_0 and β_1 to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Example: For the S&P data, we would like to fit a model that can predict YEARGAIN (the response) as a function of JANGAIN (the predictor). Since a scatterplot reveals no obvious nonlinearity, we will take the regressor to equal the predictor: that is,

$$\text{YEARGAIN} = \beta_0 + \beta_1 \text{JANGAIN} + \epsilon.$$

The relevant SAS/INSIGHT output for the regression of YEARGAIN on JANGAIN looks like this:

And the relevant SAS/INSIGHT output for the regression of YEARGAIN on JANGAIN, with the years 1987 and 1994 removed, looks like this:

TYU 8

Residuals, Predicted and Fitted Values

- The **predicted value** of Y at X is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

- For $X = X_i$, one of the values in the data set, the predicted value is called a **fitted value** and is written

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

- The **residuals**, e_i, $i = 1, \ldots, n$ are the differences between the observed and fitted values for each data value:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Tools to Assess the Quality of the Fit

- Residuals. Residuals should exhibit no patterns when plotted versus the X_i, $\hat{Y}_i$ or other variables, such as time order. Studentized residuals should be plotted on a normal quantile plot.

- Coefficient of Determination. The coefficient of determination, r^2, is a measure of (take your pick):
  - The proportion of the variation in the response "explained" by the predictor.
  - The proportion the variation in the response is reduced by knowing the predictor.

The notation r^2 comes from the fact that the coefficient of determination is the square of the Pearson correlation. Check out the quality of the two fits for the S&P data:

TYU 9

Model Interpretation

- The Fitted Slope. The fitted slope may be interpreted in a couple of ways:
  - **As the estimated change in the mean response per unit increase in the regressor.** This is another way of saying it is the derivative of the fitted response with respect to the regressor:

$$\frac{d\hat{Y}}{dx} = \frac{d}{dx}(\hat{\beta}_0 + \hat{\beta}_1 x) = \hat{\beta}_1.$$

- **In terms of the estimated change in the mean response per unit increase in the predictor.** In this formulation, if the regressor X, is a differentiable function of the predictor, Z,

$$\frac{d\hat{Y}}{dz} = \frac{d}{dz}(\hat{\beta}_0 + \hat{\beta}_1 X) = \hat{\beta}_1 \frac{dX}{dz},$$

so

$$\hat{\beta}_1 = \frac{d\hat{Y}}{dz} \Big/ \frac{dX}{dz}$$

- **In terms of the estimated change in the mean response per unit increase in the predictor.** In this formulation, if the regressor X, is a differentiable function of the predictor, Z,

$$\frac{d\hat{Y}}{dz} = \frac{d}{dz}(\hat{\beta}_0 + \hat{\beta}_1 X) = \hat{\beta}_1 \frac{dX}{dz},$$

so

$$\hat{\beta}_1 = \frac{d\hat{Y}}{dz} \Big/ \frac{dX}{dz}$$

- The Fitted Intercept. The fitted intercept is the estimate of the response when the regressor equals 0, provided this makes sense.
- The Mean Square Error. The mean square error or MSE, is an estimator of the variance of the error terms ϵ, in the simple linear regression model. Its formula is

$$\text{MSE} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

It measures the "average squared prediction error" when using the regression.

Example: Consider the S&P data with the 1987 and 1994 observations omitted. The fitted model is

$$\widehat{YEARGAIN} = 9.6462 + 2.3626 JANGAIN.$$

- The Fitted Slope. The fitted slope, 2.3626 is interpreted as the estimated change in YEARGAIN per unit increase in JANGAIN.
- The Fitted Intercept. The fitted intercept, 9.6462, is the estimated YEARGAIN if JANGAIN equals 0.
- The Mean Square Error. The MSE, 21.59, estimates the variance of the random errors.

TYU 10

Classical Inference for the SLR Model
Estimation of Slope and Intercept

Level L confidence intervals for β_0 and β_1 are

$$(\hat{\beta}_0 - \hat{\sigma}(\hat{\beta}_0)t_{n-2,\frac{1+L}{2}}, \hat{\beta}_0 + \hat{\sigma}(\hat{\beta}_0)t_{n-2,\frac{1+L}{2}}),$$

and

$$(\hat{\beta}_1 - \hat{\sigma}(\hat{\beta}_1)t_{n-2,\frac{1+L}{2}}, \hat{\beta}_1 + \hat{\sigma}(\hat{\beta}_1)t_{n-2,\frac{1+L}{2}}),$$

respectively, where

$$\hat{\sigma}(\hat{\beta}_0) = \sqrt{\text{MSE}\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]},$$

and

$$\hat{\sigma}(\hat{\beta}_1) = \sqrt{\text{MSE} \Big/ \sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Example: For the reduced S&P data (i.e. with 1987 and 1994 removed) the SAS/INSIGHT output shows the estimated intercept is $\hat{\beta}_0 = 9.65$ with $\hat{\sigma}(\hat{\beta}_0) = 1.77$. Since $t_{6,0.975} = 2.45$, a 95% confidence interval for β_0 is

$$9.65 \pm (1.77)(2.45) = (5.31, 13.98).$$

A similar computation with $\hat{\beta}_1 = 2.36$ and $\hat{\sigma}(\hat{\beta}_1) = 0.36$, gives a 95% confidence interval for β_1 as

$$2.36 \pm (0.36)(2.45) = (1.47, 3.25).$$

NOTE: Whether the interval for β_1 contains 0 is of particular interest. If it does, it means that we cannot statistically distinguish β_1 from 0. This means we have to consider plausible the model for which $\beta_1 = 0$:

$$Y = \beta_0 + \epsilon$$

This model, which is just the C+E model, implies that there is no association between Y and X.

TYU 11

To test the hypothesis

$$H_0 : \quad \beta_0 \quad = \quad \beta_{00}$$

versus one of the alternative hypotheses

$$\begin{array}{lccc} H_{a_-} : & \beta_0 & < & \beta_{00} \\ H_{a^+} : & \beta_0 & > & \beta_{00} \\ H_{a\pm} : & \beta_0 & \neq & \beta_{00}, \end{array}$$

where β_{00} is a known constant, we make use of the fact that under the distribution theory developed above, if H_0 is true,

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\hat{\sigma}(\hat{\beta}_0)} \sim t_{n-2}.$$

If t^* denotes the observed value of t, the p-value of the tests of H_0 versus H_{a_-}, H_{a^+} and $H_{a\pm}$ are $p_- = P(T \leq t^*)$, $p^+ = P(T \geq t^*)$ and $p\pm = P(|T| \geq |t^*|) = 2\min(p_-, p^+)$, respectively, where $T \sim t_{n-2}$.

Similarly, to test the hypothesis

$$H_0 : \quad \beta_1 \quad = \quad \beta_{10}$$

versus one of the alternative hypotheses

$$\begin{array}{lccc} H_{a_-} : & \beta_1 & < & \beta_{10} \\ H_{a^+} : & \beta_1 & > & \beta_{10} \\ H_{a\pm} : & \beta_1 & \neq & \beta_{10}, \end{array}$$

we make use of the fact that under the distribution theory developed above, if H_0 is true,

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}(\hat{\beta}_1)} \sim t_{n-2}.$$

If t^* denotes the observed value of t, the p-value of the tests of H_0 versus H_{a_-}, H_{a^+} and $H_{a\pm}$ are $p_- = P(T \leq t^*)$, $p^+ = P(T \geq t^*)$ and $p\pm = P(|T| \geq |t^*|) = 2\min(p_-, p^+)$, respectively, where $T \sim t_{n-2}$.

Most often, but not always, the test of greatest interest is whether, given the SLR model, the response depends on the regressor as specified by the model or not. The appropriate hypothesis test for this purpose is

$$H_0 : \quad \beta_1 \quad = \quad 0$$

versus

$$H_{a\pm} : \quad \beta_1 \quad \neq \quad 0.$$

The p-value for this test is automatically output on SAS regression output (and that of most other statistics packages).

Estimation of The Mean Response

The mean response at $X = x_0$ is

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

The point estimator of μ_0 is

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

A level L confidence interval for μ_0 is

$$(\hat{Y}_0 - \hat{\sigma}(\hat{Y}_0) t_{n-2,\frac{1+L}{2}}, \hat{Y}_0 + \hat{\sigma}(\hat{Y}_0) t_{n-2,\frac{1+L}{2}}),$$

where

$$\hat{\sigma}(\hat{Y}_0) = \sqrt{\text{MSE}\left[\frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]}.$$

Prediction of a Future Observation

A level L prediction interval for a future observation at $X = x_0$ is

$$(\hat{Y}_{new} - \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-2,\frac{1+L}{2}},$$

$$\hat{Y}_{new} + \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-2,\frac{1+L}{2}})),$$

where

$$\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_0,$$

and

$$\hat{\sigma}(Y_{new} - \hat{Y}_{new}) = \sqrt{\text{MSE}\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]}.$$

The macro REGPRED will compute confidence intervals for a mean response and prediction intervals for future observations for each data value and for other user-chosen X values.

Example:

The SAS macro REGPRED was run on the reduced S&P data, and estimation of the mean response and prediction of a new observation at the values JANGAIN= 3.3 and 13.2 were requested (These are the values for the two omitted cases). Here are the results

JANGAIN	YEARGAIN	Level 0.95 Prediction Interval	
3.3	-1.5	5.32	29.57
13.2	2.0	25.11	56.55

The Relation Between Correlation and Regression

If the standardized responses and regressors are

$$Y_i' = \frac{Y_i - \overline{Y}}{S_Y},$$

and

$$X_i' = \frac{X_i - \overline{X}}{S_X},$$

Then the regression equation fitted by least squares can be written as

$$\hat{Y}' = r \cdot X',$$

Where X' is any value of a regressor variable standardized as described above, and r is the Pearson correlation between X and Y.

The Regression Effect refers to the phenomenon of the standardized predicted value being closer to 0 than the standardized regressor. Equivalently, the unstandardized predicted value is fewer Y standard deviations from the response mean than the regressor value is in X standard deviations from the regressor mean.

For the S&P data $r = 0.4295$, so for a January gain X' standard deviations (S_X) from $\overline{X}$, the regression equation estimates a gain for the year of

$$\hat{Y}' = 0.4295 \cdot X'$$

standard deviations (S_Y) from $\overline{Y}$.
With 1987 and 1994 removed, the estimate is

$$\hat{Y}' = 0.9360 \cdot X',$$

which reflects the stronger relation.

The Relationship Between Two Categorical Variables

Analysis of categorical data is based on counts, proportions or percentages of data that fall into the various categories defined by the variables.

Some tools used to analyze bivariate categorical data are:

- Mosaic Plots.
- Two-Way Tables.

Example:

A survey on academic dishonesty was conducted among WPI students in 1993 and again in 1996. One question asked students to respond to the statement "Under some circumstances academic dishonesty is justified." Possible responses were "Strongly agree", "Agree", "Disagree" and "Strongly disagree". Here are the 1993 results:

TYU 12

Inference for Categorical Data with Two Categories Methods for comparing two proportions can be used (estimation from chapter 5 and hypothesis tests from chapter 6. See homework problem 6.24.).

Inference for Categorical Data with More Than Two Categories: One-Way Tables

Suppose the categorical variable has c categories, and that the population proportion in category i is p_i. To test

$$\begin{aligned} H_0: \quad p_i &= p_i^{(0)}, \ i = 1, 2, \ldots, c \\ H_a: \quad p_i &\neq p_i^{(0)} \text{ for at least one } i \end{aligned}$$

for pre-specified values $p_i^{(0)}, i = 1, 2, \ldots, c$, use the **Pearson χ^2 statistic**

$$X^2 = \sum_{i=1}^{c} \frac{(Y_i - np_i^{(0)})^2}{np_i^{(0)}},$$

where Y_i is the observed frequency in category i, and n is the total number of observations.

Note that for each category the Pearson statistic computes **(observed-expected)^2/expected** and sums over all categories. Under H_0, $X^2 \sim \chi^2_{c-1}$. Therefore, if x^{2*} is the observed value of X^2, the p-value of the test is $P(\chi^2_{c-1} \geq x^{2*})$.

Example:

Historically, the distribution of weights of "5 pound" dumbbells produced by one manufacturer have been normal with mean 5.01 and standard deviation 0.15 pound. It can be easily shown that 20% of the area under a normal curve lies within ± 0.25 standard deviations of the mean, 20% lies between 0.25 and 0.84 standard deviations of the mean, 20% lies between -0.84 and -0.25 standard deviations of the mean, 20% lies beyond 0.84 standard deviations above the mean, and another 20% lies beyond 0.84 standard deviations below the mean.

This means that the boundaries that break the $N(5.01, 0.15^2)$ density into five subregions, each with area 0.2, are 4.884, 4.9725, 5.0475 and 5.136.
A sample of 100 dumbbells from a new production lot shows that 25 lie below 4.884, 23 between 4.884 and 4.9725, 21 between 4.9725 and 5.0475, 18 between 5.0475 and 13 above 5.136. Is this good evidence that the new production lot does not follow the historical weight distribution?

Solution:

We will perform a χ^2 test. Let p_i be the proportion of dumbbells in the production lot with weights in subinterval i, where subinterval 1 is $(-\infty, 4.884]$, subinterval 2 is $(4.884, 4.9725]$, and so on. If the production lot follows the historical weight distribution, all p_i equal 0.2. This gives our hypotheses:

$$\begin{aligned} H_0: \quad p_i &= 0.2, \ i = 1, 2, \ldots, 5, \\ H_a: \quad p_i &\neq 0.2, \end{aligned}$$

for at least one i, $i = 1, 2, \ldots, 5$.

Since $np_i^{(0)} = 20$ for each i, the test statistic is

$$x^{2*} = \frac{(25-20)^2}{20} + \ldots + \frac{(13-20)^2}{20} = 4.4$$

The p-value is $P(\chi_4^2 \geq 4.4) = 0.3546$, so we cannot reject H_0.

Inference for Categorical Data with More Than Two Categories: Two-Way Tables

Suppose a population is partitioned into rc categories, determined by r levels of variable 1 and c levels of variable 2. The population proportion for level i of variable 1 and level j of variable 2 is p_{ij}. These can be displayed in the following $r \times c$ table:

	Column				Marginals
row	1	2	...	c	
1	p_{11}	p_{12}	...	p_{1c}	$p_{1\cdot}$
2	p_{21}	p_{22}	...	p_{2c}	$p_{2\cdot}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	p_{r1}	p_{r2}	...	p_{rc}	$p_{r\cdot}$
Marginals	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot c}$	1

We want to test

H_0 : row and column variables are independent

H_a : row and column variables are not independent.

To do so, we select a random sample of size n from the population. Suppose the table of observed frequencies is

	Column				Totals
row	1	2	...	c	
1	Y_{11}	Y_{12}	...	Y_{1c}	$Y_{1\cdot}$
2	Y_{21}	Y_{22}	...	Y_{2c}	$Y_{2\cdot}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	Y_{r1}	Y_{r2}	...	Y_{rc}	$Y_{r\cdot}$
Totals	$Y_{\cdot 1}$	$Y_{\cdot 2}$	...	$Y_{\cdot c}$	n

Under H_0 the expected cell frequency for the ij cell is given by

$$\begin{aligned} E_{ij} &= \frac{\text{row i total} \times \text{column j total}}{\text{sample size}} \\ &= \frac{Y_{i\cdot} Y_{\cdot j}}{n} \\ &= n\hat{p}_{i\cdot}\hat{p}_{\cdot j}, \end{aligned}$$

where $\hat{p}_{i\cdot} = Y_{i\cdot}/n$ and $\hat{p}_{\cdot j} = Y_{\cdot j}/n$.
To measure the deviations of the observed frequencies from the expected frequencies under the assumption of independence, we construct the Pearson χ^2 statistic

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}}.$$

Note that for the test to be valid, we require that $E_{ij} \geq 5$.

Example: A polling firm surveyed 269 American adults concerning how leisure time is spent in the home. One question asked them to select which of five leisure activities they were most likely to partake in on a weeknight. The results are broken down by age group in the following table:

As an example of the computation of table entries, consider the entries in the (1,2) cell (age: 18-25, activity: Read), in which the observed frequency is 3. The marginal number in the 18-25 bracket is 62, while the marginal number in the Read bracket is 40, so $\hat{p}_{1\cdot} = 62/269$, while $\hat{p}_{\cdot 2} = 40/269$, so the expected number in the (1,2) cell is $269(62/269)(40/269) = 9.22$. The Pearson residual is $(3 - 9.22)/\sqrt{(9.22)} = -2.05$.
The value of the χ^2 statistic is 38.91, which is computed as the sum of the squares of the Pearson residuals. Comparing this with the χ^2_{16} distribution, we get a p-value of 0.0011.

Synopsis:

For data in a two-way table with r rows and c columns, we want to test

H_0 : row and column variables are independent

H_a : row and column variables are not independent.

Test statistic:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}},$$

where Y_{ij} is the observed count in cell ij and
$E_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = Y_{i\cdot}Y_{\cdot j}/n$ is the expected count in cell ij if H_0 is true.

***p*-value:** $P(\chi^2_{(r-1)(c-1)} \geq x^{*2})$, where x^{*2} is the observed value of the test statistic.

TYU 13

Association is NOT Cause and Effect

Two variables may be associated due to a number of reasons, such as:

1. X **could cause** Y**.**
2. Y **could cause** X**.**
3. X **and** Y **could cause each other.**
4. X **and** Y **could be caused by a third (lurking) variable Z.**
5. X **and** Y **could be related by chance.**
6. **Bad (or good) luck.**

The Issue of Stationarity

- When assessing the stationarity of a process in terms of bivariate measurements X and Y, always consider the evolution of the relationship between X and Y, as well as the individual distribution of the X and Y values, over time or order.
- Suppose we have a model relating a measurement from a process to time or order. If, as more data are taken the pattern relating the measurement to time or order remains the same, we say that the process is stationary **relative to the model**.