Lecture 1: Probability Theory for Multivariate Variables

Joint Distribution, Marginal Distribution, and Independence

Le Wang

U of Oklahoma

Questions

When making management decisions, we are often interseted in the relationships between two random variables, for example,

1. Whether changes in interest rates have any impacts on stock prices
2. Wages and Years of Schooling
3. Income and the Number of Children
4. The level of air pollution and child mortality
5. etc.

Again, these variables are **random variables** because before their values are realized, we do not know what will happen.

Motivation

To understand the relationship, we need to know what to characterize and at least whether they are related in any way (or dependent) before any serious modeling.

This lecture introduces the concepts necessary for us to understand and test the existence of any relationship.

How to Characterize the Relationship?

The uncertainty about the values of two variables is governed by a **joint** distribution.

As with Econ 5023, once we have the joint distribution, we can compute probabilities of events involving both variables and understand the relatinoship between the variables.

"Itinerary" (I go beyond financial markets!)

1. **Joint Distribution (discrete variables):**
   - 1.1 Probability Mass Function and Its Properties
   - 1.2 Illustrative Example: Play Craps
   - 1.3 R implementation and Visualization
   - 1.4 Further Use of Joint Distribution: Probability of a Set
   - 1.5 Further Use of Joint Distribution: CDF
   - 1.6 Further Use of Joint Distribution: Marginal Distribution and Implementation in R
2. **Joint Distribution and (In)dependence**
   - 2.1 Definition
   - 2.2 Statistical Tests and Implementation in R
   - 2.3 **Examples (I):** Political Affiliation and Opinion on Tax Reform
   - 2.4 **Examples (II):** Student Smoking Habit and Exercise Level
   - 2.5 **Examples (III):** Super Bowl and White Jerseys

Joint Distribution (Discrete Case)

Suppose that X and Y are two discrete random variables and that X can take on values $\{x_{1,}x_{2,}\ldots,x_n\}$ and Y $\{y_1, y_2, \ldots, y_m\}$. Then, we have $n \times m$ combinations of possible events. The ordered pair (X, Y) take values in the product $\{(x_1, y_1), (x_1, y_2\}, \ldots, (x_n, y_1), \ldots, (x_n, y_m)\}$

Definition

(Joint) Probability Mass Function (p.m.f) of X and Y is defined as follows

$$\Pr[X = x_i, Y = y_j]$$

Intuition: You can think of a ordered pair (combination) as an unit, the joint distribution is nothing but the probability of a **combination**.

We can visualize the joint probability by using the following table:

$X \backslash Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\dots$	$p(x_1, y_j)$	$\dots$	$p(x_1, y_m)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\dots$	$p(x_2, y_j)$	$\dots$	$p(x_2, y_m)$
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$	$\dots$	$p(x_i, y_j)$	$\dots$	$p(x_i, y_m)$
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$	$\dots$	$p(x_n, y_j)$	$\dots$	$p(x_n, y_m)$

Table: Joint Probability Table

A joint probability mass function must satisfy the following two properties:

1. $0 \leq p(x_i, y_j) \leq 1$
2. The total probability is 1. In other words,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$$

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) &= \sum_{i=1}^{n} [\sum_{j=1}^{m} p(x_i, y_j)] \\
&= \sum_{i=1}^{n} [p(x_i, y_1) + p(x_i, y_2) + p(x_i, y_3) + \cdots + p(x_i, y_m)] \\
&= [p(x_1, y_1) + p(x_1, y_2) + p(x_1, y_3) + \cdots + p(x_1, y_m)] \\
&\quad + [p(x_2, y_1) + p(x_2, y_2) + p(x_2, y_3) + \cdots + p(x_2, y_m)] \\
&\quad + \ldots \\
&\quad + [p(x_n, y_1) + p(x_n, y_2) + p(x_n, y_3) + \cdots + p(x_n, y_m)]
\end{aligned}$$

Example 1 (Play Craps)

Let X be the value on the first dice and Y the value on the second dice. Both X and Y take on values from 1 to 6. What is the joint distribution for these two variables?

$X \backslash Y$	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Table: Joint Probability Table (Play Craps): $p(x_i, y_j) = \frac{1}{36}$

A joint probability mass function must satisfy the following two properties:

1. $0 \leq p(x_i, y_j) \leq 1$
2. The total probability is 1. In other words,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$$

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) &= \sum_{i=1}^{n} [\sum_{j=1}^{m} p(x_i, y_j)] \\
&= \sum_{i=1}^{n} [p(x_i, y_1) + p(x_i, y_2) + p(x_i, y_3) + \cdots + p(x_i, y_m)] \\
&= [p(x_1, y_1) + p(x_1, y_2) + p(x_1, y_3) + \cdots + p(x_1, y_m)] \\
&\quad + [p(x_2, y_1) + p(x_2, y_2) + p(x_2, y_3) + \cdots + p(x_2, y_m)] \\
&\quad + \ldots \\
&\quad + [p(x_n, y_1) + p(x_n, y_2) + p(x_n, y_3) + \cdots + p(x_n, y_m)]
\end{aligned}$$

Example 1 (Play Craps)

Two-step summation (fix one first and then compute along that dimension): Let X be the value on the first dice and Y the value on the second dice. Both X and Y take on values from 1 to 6.

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)$$

$X \backslash Y$	1	2	3	4	5	6	Total Probability
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$= \frac{1}{6}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$= \frac{1}{6}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$= \frac{1}{6}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$= \frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$= \frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$= \frac{1}{6}$
							$= 1$

Table: Joint Probability Table (Play Craps): $p(x_i, y_j) = \frac{1}{36}$

Implementation in R

```r
# Example: Play Craps
# Step 1: Simulate Data first
set.seed(123456)
x<-sample(1:6,100000, replace = T)
y<-sample(1:6,100000, replace = T)

# Step 2: Calculate counts for each cell
mytable<-table(x,y)

# Alternatively, combine these two into a data frame
data<-as.data.frame(cbind(x,y))
mytable<-xtabs(~ x+y, data=data)

# Step 3: Cell proportions
prop.table(mytable)
```

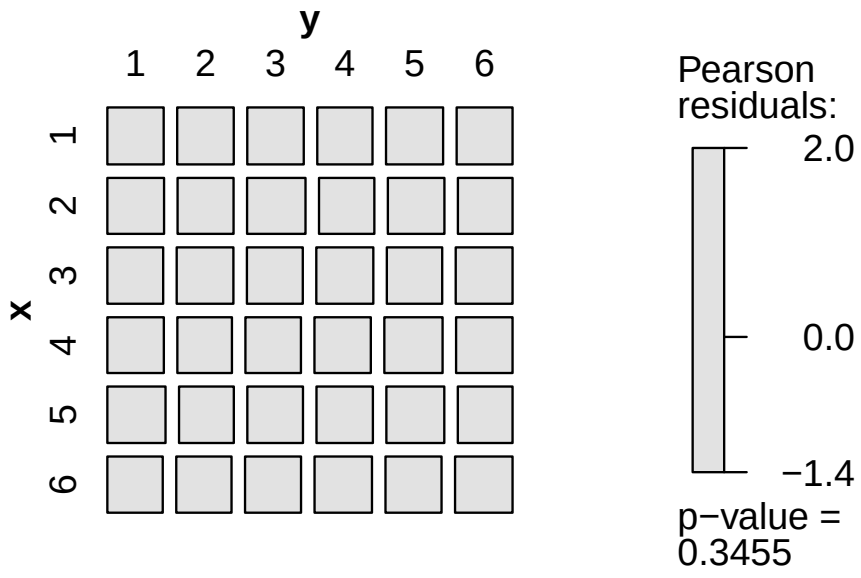
```
##    y
## x          1       2       3       4       5       6
##   1 0.02814 0.02825 0.02822 0.02760 0.02852 0.02793
##   2 0.02783 0.02748 0.02867 0.02715 0.02775 0.02681
##   3 0.02760 0.02804 0.02762 0.02819 0.02737 0.02810
##   4 0.02762 0.02669 0.02715 0.02757 0.02848 0.02761
##   5 0.02881 0.02713 0.02756 0.02793 0.02869 0.02711
##   6 0.02730 0.02724 0.02759 0.02830 0.02762 0.02833
```

More on xtabs

xtabs() Create a contingency table (optionally a sparse matrix) from cross-classifying factors, usually contained in a data frame, using a formula interface with the cross-classifying variables (separated by +) on the right hand side (or an object which can be coerced to a formula). Interactions are not allowed. On the left hand side, one may optionally give a vector or a matrix of counts; in the latter case, the columns are interpreted as corresponding to the levels of a variable. This is useful if the data have already been tabulated, see the examples below.

Visualization in R (Mosaic Plot)

```r
library(vcd)
mosaic(mytable, shade=TRUE, legend=TRUE)
```

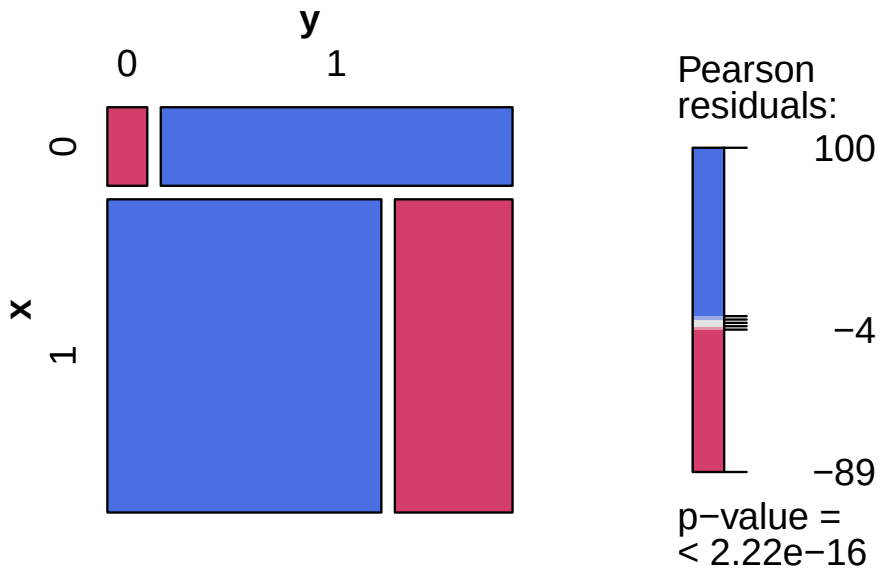


More on Mosaic Plot: Another Example

```
mytable<-xtabs(~ x+y, data=data)
mytable

##    y
## x      0     1
##   0  2043 18007
##   1 55928 24022

prop.table(mytable)

##    y
## x         0       1
##   0 0.02043 0.18007
##   1 0.55928 0.24022
```

More on Mosaic Plot: Another Example

```
mosaic(mytable, shade=TRUE, legend=TRUE)
```



We learn several things (which **combination** is more likely):

1. $X = 1, Y = 0$ is most likely, while $X = 0, Y = 0$ least likely
2. We will use this to examine the conditional probability (defined later) as well.

How can we further use the joint p.m.f?

$X \backslash Y$	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

What is probability of obtaining $X + Y = 3$?

In other words, What is $\Pr[X + Y = 3]$?

It can be obtained only when either $(X = 1, Y = 2)$ or $(X = 2, Y = 1)$.

Use 2: Cumulative Distribution Function

Definition

Joint Cumulative Distribution Function

$$F(x, y) = \Pr[X \leq x, Y \leq y]$$

In the discrete case, the CDF is a double sum

$$\begin{aligned} F(x, y) &= \Pr[X \leq x, Y \leq y] \\ &= \sum_{x_i \leq x,\, y_j \leq y} \sum p(x_i, y_j) \end{aligned}$$

Example

$X \backslash Y$	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

What is $\Pr[X \leq 3.5, Y \leq 4]$?

$= \frac{12}{36}$

Use 3: Marginal Distribution

Although X and Y are jointly distributed random variables, we sometimes are only interested in each variable itself (either X or Y). In this case, we need to find out the p.m.f. of X without Y, or the p.m.f. of Y without X. This is called marginal p.m.f..

It is nothing but $\Pr[X = x_i]$!

X	Probability
1	
2	
3	
4	
5	
6	

Table: Marginal Distribution of X in the experiment of Playing Craps

From Joint to Marginal Distribution

$X \backslash Y$	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

X	Probability
1	
2	
3	
4	
5	
6	

Marginal p.m.f

From the example above, the marginal p.m.f can be obtained from the joint p.m.f using the following formula

$$p_X(x_i) = \sum_{j=1}^{m} p(x_i, y_j)$$

$$p_Y(y_j) = \sum_{i=1}^{n} p(x_i, y_j)$$

Implementation in R

```r
# Example: Play Craps
# Step 1: Simulate Data first
set.seed(123456)
x<-sample(1:6,100000, replace = T)
y<-sample(1:6,100000, replace = T)

# Step 2: Calculate counts for each cell
mytable<-table(x,y)

# Alternatively, combine these two into a data frame
data<-as.data.frame(cbind(x,y))
mytable<-xtabs(~ x+y, data=data)

# Step 3: Cell proportions
prop.table(mytable)
```

```
##    y
## x          1       2       3       4       5       6
##   1 0.02814 0.02825 0.02822 0.02760 0.02852 0.02793
##   2 0.02783 0.02748 0.02867 0.02715 0.02775 0.02681
##   3 0.02760 0.02804 0.02762 0.02819 0.02737 0.02810
##   4 0.02762 0.02669 0.02715 0.02757 0.02848 0.02761
##   5 0.02881 0.02713 0.02756 0.02793 0.02869 0.02711
##   6 0.02730 0.02724 0.02759 0.02830 0.02762 0.02833
```

Implementation in R

```
# Step 4: Add Marginal Sums to Cell proportions
addmargins(prop.table(mytable))
```

```
##       y
## x             1       2       3       4       5       6     Sum
##   1   0.02814 0.02825 0.02822 0.02760 0.02852 0.02793 0.16866
##   2   0.02783 0.02748 0.02867 0.02715 0.02775 0.02681 0.16569
##   3   0.02760 0.02804 0.02762 0.02819 0.02737 0.02810 0.16692
##   4   0.02762 0.02669 0.02715 0.02757 0.02848 0.02761 0.16512
##   5   0.02881 0.02713 0.02756 0.02793 0.02869 0.02711 0.16723
##   6   0.02730 0.02724 0.02759 0.02830 0.02762 0.02833 0.16638
##   Sum 0.16730 0.16483 0.16681 0.16674 0.16843 0.16589 1.00000
```

Joint, Marginal, and Independence: The Case of Continuous Variables

It is essentially the same as the discrete case. The only difference is that we will replace discrete sets of values with continuous intervals, summation signs with integrals.

Discrete Case

1. p.m.f.: $p(x_i, y_j)$
2. CDF: $F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j)$
3. marginal p.m.f:

$$p_X(x_i) = \sum_{j=1}^{m} p(x_i, y_j)$$

$$p_Y(y_j) = \sum_{i=1}^{n} p(x_i, y_j)$$

Continuous Case

1. p.d.f.: $f(x, y)$
2. CDF: $F(x, y) = \int_c^y \int_a^x f(u, v) du dv$
3. marginal p.m.f:

$$f_X(x) = \int_a^b f(x, y) dy$$

$$f_Y(y) = \int_c^d f(x, y) dx$$

Joint Distribution, Marginal Distribution, and Independence

Question: How do we know whether or not two variables are related?

The need to detec and properly measure *association* and *dependence* is an essential task in economic model building and forecasting. This is the main activity in empirical economics whether one is searching for contemporaneous relations among economic variables, working with cross-section data or time series, or determining dynamic structure or any patterns.

Numerous diagnostic procedures, such as the DW test, LM tests, are used to examine model *residuals* for departure from *independence, i.i.d., martingale difference property, etc.*

Joint Distribution, Marginal Distribution, and Independence

We can at least define when two variables are NOT related, or independent of each other.

1. When two events are said to be **independent** of each other, what this means is that the probability that one event occurs in no way affects the probability of the other event occurring.
2. When two events are said to be **dependent**, the probability of one event occurring influences the likelihood of the other event (either smaller or larger than the likelihood of the other event alone!).

Joint Distribution, Marginal Distribution, and Independence

Mathematically, how do we express independence?

The following two conditions are equivalent:

$$F(x, y) = F_X(x)F_Y(y)$$
$$p(x_i, y_j) = p^*(x_i, y_j) = p_X(x_i)p_Y(y_j)$$

Example of Independence or Dependence

$X \backslash Y$	1	2	3	4	5	6	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	
$p(y_j)$							

Table: Dependence or Independence

Joint, Marginal, and Independence: The Case of Continuous Variables

$$X \perp Y$$

Discrete Case

1. Independence:

$$F(x, y) = F_X(x)F_Y(y)$$
$$p(x_i, y_j) = p_X(x_i)p_Y(y_j)$$

Continuous Case

1. Independence:

$$F(x, y) = F_X(x)F_Y(y)$$
$$f(x, y) = f(x)f(y)$$

Test of Independence

How do we test whether or not two (discrete) variables are indeed independent?

Test of Independence

We know how the actual joint distribution look like?

We can present either the contingency table or the p.m.f table.

But how will this table look like if the two variables were not related at all?

If we can have this information, we can then compare these two tables!

Test of Independence

Expected Table: the Counterfactual Table in the case

$$E = \frac{\text{row total} \times \text{column total}}{\text{sample size}}$$

Example: Political Affiliation and Opinion on Tax Reform

Suppose that we have a random sample of 500 U.S. adults who are questioned regarding their political affiliation and opinion on a tax reform bill. The observed contingency table is given below.

	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

Table: Example: Observed Counts

Example: Political Affiliation and Opinion on Tax Reform

	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

Table: Observed Table

Example: How should we construct expected table?

	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

Table: Observed Table

	Favor	Indifferent	Opposed	Total
Democrat				285
Republican				215
Total	202	150	148	500

Table: Expected Table

Example: How should we construct expected table

1. Construct *expected* probability under **independence**: $p^*(x, y)$
   1.1 marginal distributions: $p(x)$ and $p(y)$
   1.2 $p^*(x, y) = p(x) \cdot p(y)$
2. Construct expected counts: $N \times p^*(x, y)$

Example: How should we construct expected probability table?

Under the null hypothesis of independence: $p(x, y) = p(x)p(y)$ [**How do we calculate the marginal distributions** $p(x)$ **and** $p(y)$**?**]

	Favor	Indifferent	Opposed	Total
Democrat				$\frac{285}{500}$
Republican				$\frac{215}{500}$
Total	$\frac{202}{500}$	$\frac{150}{500}$	$\frac{148}{500}$	$\frac{500}{500}$

Table: Expected (probability) Table

Example: How should we construct expected?

Under the null hypothesis of independence: $p(x, y) = p(x)p(y)$

	Favor	Indifferent	Opposed	Total
Democrat	$\frac{285 \cdot 202}{500 \cdot 500}$	$\frac{285 \cdot 150}{500 \cdot 500}$	$\frac{285 \cdot 148}{500 \cdot 500}$	$\frac{285}{500}$
Republican	$\frac{215 \cdot 202}{500 \cdot 500}$	$\frac{215 \cdot 150}{500 \cdot 500}$	$\frac{215 \cdot 148}{500 \cdot 500}$	$\frac{215}{500}$
Total	$\frac{202}{500}$	$\frac{150}{500}$	$\frac{148}{500}$	$\frac{500}{500}$

Table: Expected (probability) Table

Construct expected count table

	Favor	Indifferent	Opposed
Democrat	$\frac{285 \cdot 202}{500 \cdot 500} \cdot 500 = 115.14$	$\frac{285 \cdot 150}{500 \cdot 500} \cdot 500 = 85.5$	$\frac{285 \cdot 148}{500 \cdot 500} \cdot 500 = 84.$
Republican	$\frac{215 \cdot 202}{500 \cdot 500} \cdot 500 = 86.86$	$\frac{215 \cdot 150}{500 \cdot 500} \cdot 500 = 64.5$	$\frac{215 \cdot 148}{500 \cdot 500} \cdot 500 = 63.$
Total	202	150	148

Table: Expected Table

Example: Construct expected count table works?

	Favor	Indifferent	Opposed
Democrat	$\frac{285 \cdot 202}{500 \cdot 500} \cdot 500 = 115.14$	$\frac{285 \cdot 150}{500 \cdot 500} \cdot 500 = 85.5$	$\frac{285 \cdot 148}{500 \cdot 500} \cdot 500 = 84.$
Republican	$\frac{215 \cdot 202}{500 \cdot 500} \cdot 500 = 86.86$	$\frac{215 \cdot 150}{500 \cdot 500} \cdot 500 = 64.5$	$\frac{215 \cdot 148}{500 \cdot 500} \cdot 500 = 63.$
Total	202	150	148

Table: Expected Table

```
115.14/285
```

```
## [1] 0.404
```

```
86.86/215
```

```
## [1] 0.404
```

```
85.5/285
```

```
## [1] 0.3
```

```
64.5/215
```

```
## [1] 0.3
```

Example: Why this expected table works?

Intuitively, because this table implies that the share of democrats in favor of the reform is the same as the republicans in favor of the reform. The share for each category does not vary across party affiliations!

We construct it that way!

Test of Independence

How will you develop the test?

Test of Independence

$$H_0 : Independent$$
$$H_A : Not \quad Independent$$

Test Statistic:

$$\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-1)(m-1)}$$

In this example,

$$\frac{(138-115.14)^2}{115.14} + \frac{(83-85.50)^2}{85.50} + \frac{(64-84.36)^2}{84.36} + \frac{(64-86.86)^2}{86.86} + \frac{(67-64.50)^2}{64.50} + \frac{(84-63.64)^2}{63.64} = 22.152 \sim \chi^2_{(2-1)(3-1)=2}$$

Test of Independence

```
1-pchisq(22.152,df=2)

## [1] 1.547941e-05
```

These two tables are very different.

What is your conclusion?

Test of Independence

Conclusion:

Given the p-value is very small (smaller than commonly used criteria such as $0.01, 0.05, 0.10$), we consider it is highly likely under the null hypothesis (that these two tables are equal to each other)

We reject the null hypothesis

The observed table is not the same as the expected table

It is not independent!

Implementation in R: A Worked Example

Question: Are the students smoking habit indepedent of their exercise levels?

```
library(MASS)
mytable <- table(survey$Smoke, survey$Exer)
mytable

##
##         Freq None Some
##    Heavy    7    1    3
##    Never   87   18   84
##    Occas   12    3    4
##    Regul    9    1    7

chisq.test(mytable)

## Warning in chisq.test(mytable):  Chi-squared approximation may be
incorrect

##
##  Pearson's Chi-squared test
##
## data:  mytable
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

Implementation in R: A Worked Example

The warning message found in the solution above is due to the small cell values in the contingency table.

We can combine those small cells, for example,

```
mytable<-cbind(mytable[,"Freq"],mytable[,"None"]+ mytable[,"Some"])

mytable

##       [,1] [,2]
## Heavy    7    4
## Never   87  102
## Occas   12    7
## Regul    9    8

chisq.test(mytable)

##
##  Pearson's Chi-squared test
##
## data:  mytable
## X-squared = 3.2328, df = 3, p-value = 0.3571
```

How to visualize independence?

Before you even go into formal testing, can you propose a way to visualize independence with a plot? (**Hint: Have you used QQ plot before?**)
For example, Bhattaharya-Hellinger-Matusita Entropy Distance

$$\frac{1}{2}\int\int \left(f(x,y)^{\frac{1}{2}} - (f(x)f(y))^{\frac{1}{2}}\right)^2 dxdy$$

This is more involved because we need to estimate $f(x,y), f(x), f(y)$. Nonparametric density estimation is beyond the scope of this course. But one can discretize the continuous variables to make it a discrete variable and use the method above.

Test of Independence: Additional Tests

Extra Reading: I will leave this for you to learn outside classroom. There are alternative tests of independence, for example,

1. **Fishers Exact Test**: In R, `fisher.test(mytable)`
2. **Cochran-Mantel-Haenszel Test:** In R, `mantelhaen.test(mytable)`

Another Application: Super Bowl and White Jerseys

The team that has worn white jerseys in the Super Bowl has a 33-19 all-time record, including 12 wins in the last 13 years.

	Win	Lose	
White	33	19	51
Non-White	19	33	51
	51	51	102

Table: Super Bowl and White Jerseys

Super Bowl and White Jerseys

```
mytable <- matrix(c(33,19,19,33),ncol=2,byrow=TRUE)
mytable<-as.table(mytable)
mytable
```

```
##    A  B
## A 33 19
## B 19 33
```

```
chisq.test(mytable)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mytable
## X-squared = 6.5, df = 1, p-value = 0.01079
```

Super Bowl and White Jerseys

Question: What will happen if we exclude the last 12 years? Or, if we simply examine the last 12 years?

Visualization of Dependence

Can you think of a way to visualize the independence?

Scatter plot of joint probability and the product of marginal probabilities and a 45 degree straight line.

What can we learn?

Dependence or Independence is just again about relationship, not a causal one.

R or econometrics cannot tell us whether or not there exists a causal relationship.

In predictions, this is not what we care about, either. Later we can see that this has predictive power, but we need further assumptions or more information in order to make any causal claims.

Summary

1. We are interested in joint distribution when there are more than one variable of interest.
2. We define marginal distribution and many useful concepts based on the joint distribution.
3. We define independence based on joint and marginal distributions.
4. We discuss tests of independence.