Learning Objectives: Correlation Tests

- **Calculate Pearson correlation coefficient**
- **Calculate Spearman correlation coefficient**
- **Calculate Kendall correlation coefficient**
- **Perform a Chi-Square test on a data set(s)**

definition

Assumptions

- Learners are comfortable creating vectors and data frames with numerical elements.
- Learners are comfortable applying statistical functions on vectors and data frames.

Limitations

- This section will cover only the most commonly used correlation tests.

Correlation Coefficients

Before we begin, let's open up the `correlation.r` file within RStudio. See instructions below:

info

Open the `correlation.r` file

Within RStudio, open the `correlation.r` file by selecting: `File` —> `Open File...` —> `code` —> `describe` —> `correlation.r`

Pearson, Spearman, and Kendall Correlation Coefficients

Pearson's, Spearman's, and Kendall's Correlation Coefficients are used to determine how strongly two variables or pieces of data are associated. Depending on your needs, you may opt to use one over the other. The following breaks down some of the similarities and differences between the three correlation coefficients (source: datascience.stackexchange.com):

Con of E Cor Coe

cor Spe and cor

* N par cor are pov bec use inf

in t
calc
Pea
corr
use
infc
abo
mea
dev
fro
mea
non
par
corr
use
ord
infc
and
pai

* In
of n
par
corr
it's
tha
Y va
be c
or c
and
app
nor
dist
for
are
req
in t
Pea
corr
ass
dist
of X
sho
nor
dist
and
con

* Co

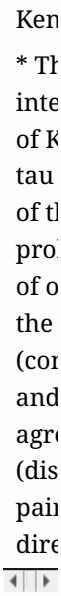
coe
onl
line
(Pe
mo
(Sp
and
rela

###
Spe
cor
Ken
cor

* In
nor
Ken
cor
mo
and
tha
Spe
cor
me
Ken
cor
pre
wh
are
sam
som

* K
cor
has
com
com
com
wit
log
Spe
cor
wh
the
size

* Sp
rho
larg



Parametric vs. Non-Parametric

The term **parametric** often refers to data that is **normally** distributed. Normally distributed means that there are assumptions about the mean and the variance of the data that we can make. On the other hand, it is more difficult to determine the true mean or variance regarding data is non-parametric. All of the comparison tests you learned previously assume that the data is parametric. In regards to the correlation coefficient tests, **Pearson** is parametric while **Spearman and Kendall** are non-parametric.

Determining If Data Is Parametric or Not

To determine whether the data is normally distributed or not, you really have to **look** at the data itself. If **most** of the data points (mode of the data) fall into the middle of the data range, we can more safely assume that the data is parametric. We sometimes refer to parametric data as looking like a **bell curve**. On the other hand, if the data points seem very randomly distributed (high level of variance) or if the data tends to lean to the minimum or maximum value then it is more likely that the data is non-parametric. We sometimes refer to parametric data as being **skewed** or leaning to one side.

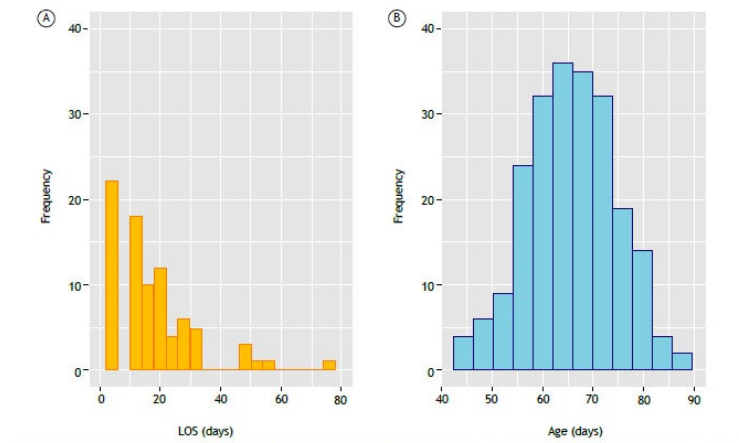


Figure 1. In A, hospital length of stay (LOS) of patients admitted for COPD exacerbations. The data clearly have a non-normal distribution and are skewed to the right. In B, age distribution of the same group of patients. The data are normally distributed (N = 200 patients).

.guides/img/corr/para-nonpara-models

Source

The figure above shows an example of non-parametric data on the left (labeled `A`) and an example of parametric data on the right (labeled `B`).

Syntax

The basic syntax for all three tests to calculate the correlation coefficient is:

```
cor(x, y, method = c("pearson", "kendall", "spearman"))
```

To compute both the correlation coefficient as well as the **significance level** (or p-value) of the correlation, you can use:

```
cor.test(x, y, method = c("pearson", "kendall", "spearman"))
```

In both cases,
* `x` and `y` represent numeric vectors
* `method` represents the correlation method
* Choose `"pearson"`, `"kendall"`, or `"spearman"`

Examples

If I want to determine the correlation coefficient between the height and weight measurements for a population of people, I can perform the three tests as follows:

Pearson

```r
d <- read.csv("data/biostats.csv")

h <- d$Height..in.
w <- d$Weight..lbs.
df <- data.frame(h, w)

print(cor.test(h, w, method = c("pearson")))
```

Result:

```
    Pearson's product-moment correlation

data:  h and w
t = 7.1566, df = 16, p-value = 2.28e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6853287 0.9518602
sample estimates:
      cor
0.8729047
```

Again, note that Pearson assumes that the data is **normally distributed**.

Spearman

```r
d <- read.csv("data/biostats.csv")

h <- d$Height..in.
w <- d$Weight..lbs.
df <- data.frame(h, w)

print(cor.test(h, w, method = c("spearman")))
```

Result:

```
    Spearman's rank correlation rho

data:  h and w
S = 143.59, p-value = 7.263e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8518208

Warning message:
In cor.test.default(h, w, method = c("spearman")) :
  Cannot compute exact p-value with ties
```

Kendall

```
d <- read.csv("data/biostats.csv")

h <- d$Height..in.
w <- d$Weight..lbs.
df <- data.frame(h, w)

print(cor.test(h, w, method = c("kendall")))
```

Result:

```
    Kendall's rank correlation tau

data:  h and w
z = 4.0373, p-value = 5.407e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.7092245

Warning message:
In cor.test.default(h, w, method = c("kendall")) :
  Cannot compute exact p-value with ties
```

Spearman and Kendall, on the other hand, **do not assume normal distribution** of data, thus you see the warning message `Cannot compute exact p-value with ties`.

cor, rho, and tau

Unlike comparison tests where we are **comparing** data sets, we only try to find whether an **association** exists within two data sets in correlation tests. Thus, we focus on the following correlation values: `cor` for Pearson, `rho` for Spearman, and `tau` for Kendall.

important

IMPORTANT

Correlation coefficient values such as `cor`, `rho`, and `tau` range from `-1` to `1`. We can determine correlation between the data sets using these rules:
* The closer the coefficient is to `0`, the less association there is between the data sets. A coefficient of exactly `0` means there is **no association** between them.
* The closer the coefficient is to `-1`, the more **negatively associated** the data sets are. A coefficient of exactly `-1` means there is a **very strong negative association** between them.
* The closer the coefficient is to `1`, the more **positively associated** the data sets are. A coefficient of exactly `1` means there is a **very strong positive association** between them.

The closer the coefficients are to `-1` or `1` means there is a very **strong association**. On the other hand, closer to `0` means a very **weak** association. A **strong positive** correlation (closer to `1`) means that as one variable increases, the other also increases and as one variable decreases, the other does as well. A **strong negative** correlation (closer to `-1`) means that as one variable increases, the other decreases and vice versa.

Looking at our correlation coefficients `cor`, `rho`, and `tau`, we have `0.8729047`, `0.8518208`, and `0.7092245`. All of which are closer to a value of `1`. This means that there is a **strong positive** association between our data sets `h` and `w` which stand for `height` and `weight`. This strong positive correlation suggests that as height increases, so does weight.

Chi-Square

Chi-Square

A Pearson's Chi-Square test is used to determine if two independent sets of data are **associated** with each other. The Chi-Square test is best used for categorical data such as gender/sex, day of the week, etc. This is determined using the **p-value** that is calculated from the test. A p-value of less than or equal to `0.05` means the data sets are **associated** with each other while a higher p-value means the data sets are **not associated** with each other. Like the Spearman and Kendall correlation coefficient tests, the Chi-Square test **does not** assume that the data is parametric (normally distributed).

The basic syntax to perform a Chi-Square test is:

```
chisq.test(table(x, y))
```

Where:
* `x` and `y` represent vectors holding categorical data

For example, if we want to know if gender and the passing of a math test are associated with each other, we first read that data:

```
d <- read.csv("data/gender-test.csv")

gender <- d$Gender
test <- d$Test
df <- data.frame(gender, test)

print(df)
```

Result:

```
   gender       test
1       M          P
2       M          P
3       M          P
4       M          P
5       M          P
6       M          P
7       M          P
```

```
8       M           P
9       M           P
10      M           P
11      M           P
12      M           P
13      M           P
14      M           P
15      M           P
16      M           P
17      M           P
18      M           P
19      M           N
20      M           N
21      M           N
22      M           N
23      M           N
24      M           N
25      M           N
26      F           P
27      F           P
28      F           P
29      F           P
30      F           P
31      F           P
32      F           P
33      F           P
34      F           P
35      F           P
36      F           P
37      F           P
38      F           P
39      F           P
40      F           P
41      F           P
42      F           P
43      F           P
44      F           P
45      F           P
46      F           N
47      F           N
48      F           N
49      F           N
50      F           N
```

- `M` and `F` represent “male” and “female” respectively
- `P` and `N` represent “passed” and “not passed” respectively

Note that a `table` works similarly to a data frame, the difference is that a table will include counts or how many times an event happens instead of listing all of those events themselves. Use the following syntax to print a table that includes `gender` and `test`:

```
print((table(gender, test)))
```

Result:

```
      test
gender          N          P
     F          5         20
     M          7         18
```

From the table, we can determine that `20` females passed the test compared to `18` males and `5` females did not pass the test compared to `7` males. Based on this data, it might be tempting to say that if you are a female you have a higher chance to pass the test than if you are a male. This is where the usefulness of a Chi-Square test comes into play because we can test if this association is even significant.

Chi-Square Test

To perform a Chi-Square test on the data, use the following:

```
d <- read.csv("data/gender-test.csv")

gender <- d$Gender
test <- d$Test

print(chisq.test(gender, test))
```

Result:

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  gender and test
X-squared = 0.10965, df = 1, p-value = 0.7405
```

Since the p-value of `0.7405` is **not** less than or equal `0.05`, we **fail to reject** the null hypothesis meaning the calculations show that there is **no significant association** between gender and passing a math test.