ECE 6960: Adv. Random Processes & Applications
Lecture Notes, Fall 2010

Lecture 10

Today: Hypothesis Testing: (1) Correlation (2) Distribution

Exam 1 on Thu. OH today: 3:30-5pm; Wed 9-10am.

1 Tests of Correlation

Here is the problem at hand. We measure two sets of data,

$$\begin{aligned} \mathbf{X} &= [X_1, \ldots, X_n]^T \\ \mathbf{Y} &= [Y_1, \ldots, Y_n]^T \end{aligned}$$

We often want to know, are these data correlated? Here is some more information on what this question means mathematically.

See Section 11.5 of Milton and Arnold, "Introduction to Probability and Statistics", McGraw-Hill, 2nd ed., 1990. Some things that will help you read this handout:

- The statistics folks use Greek letters for a concept (ρ and σ, for example). But then they use the capital Latin letters, R and S, to denote a random variable associated with what you get from some *calculation* of the value of the parameter from your (noisy) measurements.
- To elaborate on that last point:

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ S_{xy} &= \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ S_{xx} &= \sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned}$$

  Note that the mean is normalized by n, but S_{xy} and S_{xx} are not. This is just notation. As it says, they use S_{xy}/n as an estimate of the covariance of X and Y.
- The hat mark is used to denote an estimate.
- Note that there are other estimators for variance. One is $S_{xx}/(n+1)$, which is less biased than S_{xy}/n.

- The term "SSE" means sum of squared error. For linear regression, we assume that

$$Y_i = \beta_0 + \beta_1 X_i + N_i$$

where N_i is some zero mean (typically Gaussian) noise contribution with variance σ^2. So the SSE is

$$SSE = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

and the variance of the noise is estimated to be

$$S^2 = \hat{\sigma}^2 = SSE/(n-2).$$

1.1 Estimation of ρ

A main points of the handout is that you can estimate correlation coefficient from data,

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

An important question is then to test whether you have enough data to support the conclusion that $\rho \neq 0$.

Example: Ford Stock Price Changes

Consider the daily price for the stock of Ford Motor Company (F), for thirty years, Jan. 1977 through Sept. 2007, as shown in Figure 2. This data was recorded from `http://finance.yahoo.com/`, entering F and clicking 'get quotes', clicking 'Historical Data', and then clicking 'Download To Spreadsheet' (at the bottom). In there, you can delete any header data and save it to a text file, and then you can load it directly into Matlab.

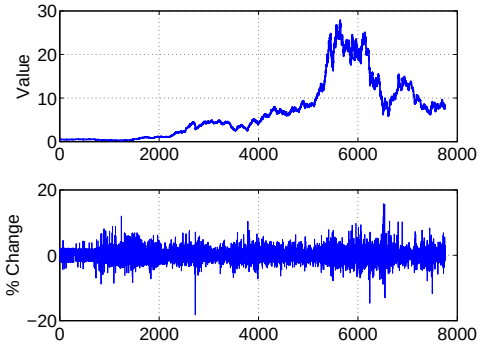


Figure 1: Ford stock value, and daily percent change.

If there is correlation in the price change, we could (theoretically) make money off of it - based on the change one day, we could

predict (and bet on) the change the next day. Over many many days, this might help.

Let X_i be the normalized daily change on day i and let Y_i be the normalized daily change on day $i+1$. Let $i = 1, \ldots, n$, where we have $n+1$ total samples of the stock price (here, $n = 7753$).

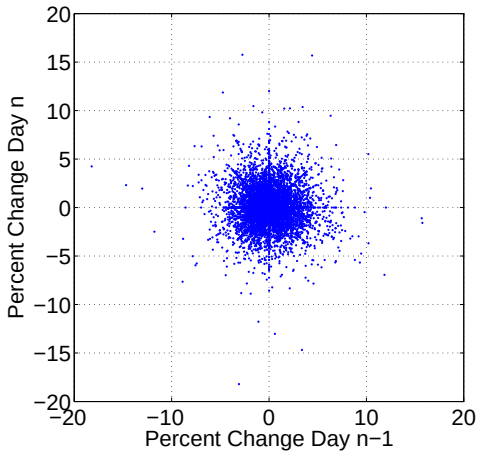


Figure 2: Daily percent change on a given day vs. the daily percent change the next day.

Let's compute the correlation coefficient using $\mathbf{X}$ and $\mathbf{Y}$.

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-0.184}{\sqrt{3.38 \cdot 3.38}} = -0.0543.$$

1.2 Testing for $\rho \neq 0$

We want to test the following hypotheses:

- H_0: (X_i, Y_i) are Gaussian with $\rho = 0$.
- H_1: (X_i, Y_i) are Gaussian with $\rho \neq 0$.

We'd love to be more general than Gaussian but this is what analysis is available. (Analysis project?)

You can compute a statistic they call T_{n-2},

$$T_{n-2} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

If the data under H_0 is approximately Gaussian, then the r.v. T_{n-2} is distributed (approximately) as the T distribution with "$n-2$ degrees of freedom". This distribution is also called the "Student-t" distribution. If the data has $\rho \neq 0$, then T_{n-2} will be very positive or negative.

From our detection theory knowledge, we then test

- H_0: $T_{n-2} \sim$ Student-t with $n-2$ d.f.
- H_1: T_{n-2} is not Student-t with $n-2$ d.f.

Note that this is not a "simple" test at all - we only are saying that the distribution under H_1 has high probability far away from the origin. Because of that, we're going to decide H_0 when T_{n-2} is close to zero:

$$|T_{n-2}| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

We want a $P_{FA} = \alpha$. Letting $t_\nu(x)$ be the CDF of the Student-t r.v. with ν d.f., we have that

$$\begin{aligned}
\alpha &= 1 - \int_{-\gamma}^{\gamma} f_T(t)dt \\
&= 1 - t_{n-2}(\gamma) + t_{n-2}(-\gamma) \\
&= 1 - t_{n-2}(\gamma) + [1 - t_{n-2}(\gamma)] \\
&= 2 - 2t_{n-2}(\gamma) \\
\gamma &= t_{n-2}^{-1}(1 - \alpha/2)
\end{aligned}$$

This inverse of the Student-t CDF can be done in Matlab with the `tinv(1-alpha/2,n-2)` command. When n is very high (more than a couple hundred) the Student-t CDF becomes approximately the standard normal CDF.

In summary, you calculate T_{n-2} and if it is greater than a threshold, you can reject H_0 with a given P_{FA} reliability level.

The handout provides another method for finding a confidence interval for ρ, but we haven't talked about confidence intervals in here. It might be useful for a project.

1.2.1 T distribution side note

William Sealy Gosset published "The probable error of a mean" anonymously under the name "Student", while he was employed at the Guinness Brewery in Dublin [2].

We proved that the following quantity

$$Z = \frac{\overline{X} - \mu_X}{\sigma/\sqrt{n}}$$

is normally distributed with mean 0 and variance 1. Gosset studied a related quantity,

$$T = \frac{\overline{X} - \mu_X}{\frac{S_{xx}}{n-1}/\sqrt{n}}.$$

While similar to Z, the variance $\frac{S_{xx}}{n-1}$ is estimated. Thus $\frac{S_{xx}}{n-1}$ has a χ^2_n distribution (which introduces uncertainty into the denominator). Gosset's work showed that T has the pdf

$$f_T(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}(1 + t^2/\nu)^{-(\nu+1)/2},$$

with $\nu = n - 1$ and where Γ is the Gamma function.

1.3 Example

Continuing the Ford stock price example, test whether the estimated ρ is significantly different from zero. Here, we use $\alpha = 0.0001$ (!) to find that $\gamma = t_{n-2}^{-1}(1 - \alpha/2) = 3.9$. Next,

$$T_{n-2} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = -4.79$$

Since $|T_{n-2}| = 4.79 > 3.9$, we can say with very high confidence that the percent change in stock price on day i is negatively correlated with the percent change in stock price on day $i + 1$.

Remaining questions:

- Is the data Gaussian?
- Has the model changed over the past 30 years?
- Given my model and today's price change, how much should I expect the stock price to drop/climb tomorrow?
- How is the price change of Ford tomorrow related to other stocks, or other information that I have access to today?

2 Empirical Distributions

The way we're typically told to look at a distribution of a bunch of experimental data is to plot a histogram. For example, I generated 100 i.i.d. Gaussian random variables $X_1, \ldots, X_{100}$ using Matlab's `randn` function and then plotted the histogram, shown in Figure 3(a). Sure, it kind of "looks Gaussian", but how can you really tell from this plot?

2.1 Empirical CDF

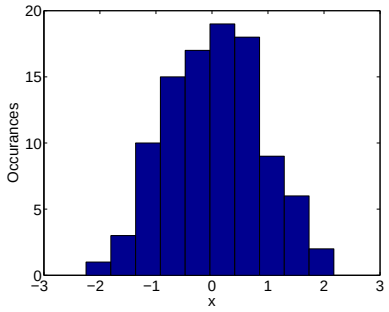
Let $X_1, \ldots, X_n$ be the recorded data. Define the empirical distribution function F_n for the n observations as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}$$

where $I_{X_i \leq x}$ is the indicator function. Basically, we count the fraction of measurements which are less than x. In other words,

$$F_n(X_i) = \frac{j}{n}$$

where j is its order (from smallest to largest) as mentioned earlier.

(a) 

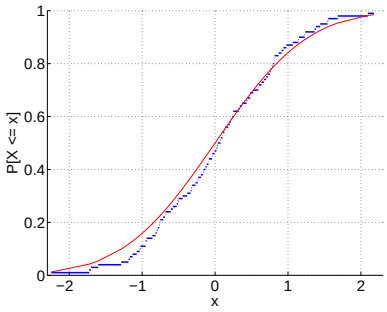
(b) 

Figure 3: (a) Histogram and (b) Empirical CDF, of data generated from `randn`.

2.2 Quantile-Quantile Plot

A good statistician would plot a quantile-quantile (QQ) plot rather than a histogram. A QQ plot compares the CDF of one set of data either to: (1) another set of data, or (2) an analytical CDF.

Here's how a QQ plot is generated.

1. Order (sort) the probabilities from smallest to largest. We denote $X_{(j)}$ to be the jth smallest measured value of the set of measurements $\{X_1, \ldots, X_n\}$.

2. Determine the analytical model to compare with the data. Let $F_{X_i}(x) = F_X(x)$ for all i be the CDF of data.

3. For each $j = 1, \ldots, N$, find a z_j such that

$$F_X(z_j) = \frac{j - 0.5}{n}$$

4. Plot values of $X_{(j)}$ vs. values of z_j.

Note that $(j - 0.5)/n$ maps the 'order' to a number between 0 and 1. The value $z_j \; s.t. \; F_X(z_j) = \frac{j-0.5}{n}$ is an ideal: z_j is where the jth smallest value of a sample of n values should be, according to the model F_X.

Thus if the data follows the model, $z_j = X_{(j)}$, and the data will fall on a straight line with slope 1. People also plot this straight line as a reference on the QQ plot.

Matlab's `qqplot` function compares a data set to the standard normal distribution. You can modify this function if comparing to a non-Gaussian distribution.

For example, Figure 4(i) shows the QQ plot of the same data from Figure 3, compared to the standard Normal distribution (zero-mean, unit-variance Gaussian distribution).

3 Tests of Distribution

We want to test

- $H_0 : X_1, \ldots, X_n$ come from distribution with CDF $F_X(x)$.
- $H_1 : X_1, \ldots, X_n$ do not come from distribution with CDF $F_X(x)$.

We are willing to accept a probability of false alarm of α. Here are two tests that will achieve such a false alarm rate. Note these are not "most powerful" tests – we don't claim to have a MP or UMP test of distribution. These are two tests that are general (may be used regardless of distribution) and commonly used. Even if they are not the most powerful for a given n, we can still use them to detect a different distribution with a desired P_{FA}.

The Kolmogorov-Smirnov (KS) statistic is given by

$$D_n = \max_i |F_n(X_i) - F_X(X_i)|.$$

The Cramér-von-Mises (CVM) statistic is given by

$$W^2 = \frac{1}{12n} + \sum_{j=1}^{n} \left[\frac{j - 0.5}{n} - F_X(X_{(j)}) \right]^2.$$

In both tests, the null hypothesis is rejected if statistic (D_n or W^2) are too high.

Both test the 'distance' between the CDF $F_X(x)$ and the empirical CDF. H_0 is rejected if the distance is too great.

The KS test has an analytical threshold as a function of α to use in the test. The Matlab function kstest will produce this value for you and run the test. The data in Figure 3 was run through a KS test and had a $p = 0.74$. This means that any $\alpha < p$ would have 'passed' the KS test, *i.e.*, kept the null hypothesis.

The CVM test uses results specific to a distribution, so you'll need to get a value out of a table. The Matlab function `mtest` works for tests of Gaussianity. The CVM test run on the data in Figure 3 decided to keep the null hypothesis.

3.1 Research Example

I have collected data (about 900 data points) for two r.v.s R and T [1], data available on `http://span.ece.utah.edu/data-and-tools`. These are the noise or error in received signal strength (RSS) measurements and in time-of-arrival (TOA) measurements. Essentially, RSS and TOA follow a physically-based model; the errors are the difference between the physical (deterministic) model and the measured data. The standard assumption in the literature is that R is

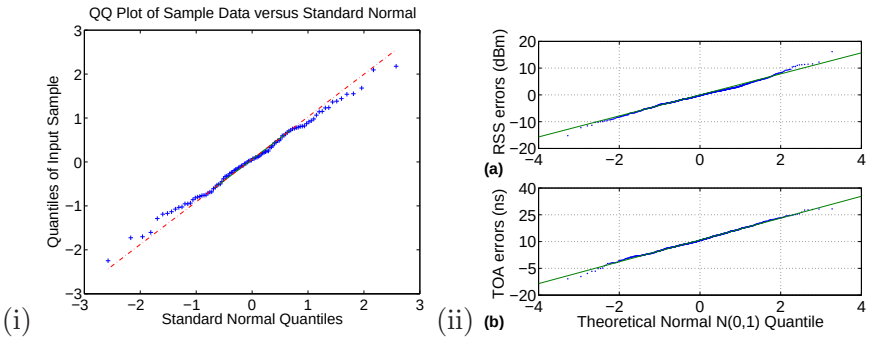


Figure 4: QQ Plot of (i) data generated from `randn`, and measured (ii-a) dB RSS and (ii-b) TOA data, compared to the standard Normal distribution.

log-normally distributed and T is Gaussian-distributed. So, I have tested $10\log_{10} R$ and T for Gaussianity. Their Q-Q plots are in Figure 4(ii).

Using a Kolmogorov-Smirnov (KS) test, I tested the hypothesis:

- $H_0 : R_{i,j} \sim \mathcal{N}(\bar{r}^R, S_R^2)$
- $H_1 : R_{i,j}$ is not distributed as given.

where $\bar{r}^R$ is the sample mean of $R_{i,j}$ and S_R^2 is the sample variance. (Problem?)

An identical test was conducted on $T_{i,j}$ for the TOA measurements.

For the RSS and TOA residuals, the KS tests yield p-values of 0.09 and 0.50, respectively. In both cases, we would decide to accept H_0 at a level of significance of $\alpha = 0.05$.

However, the low p-value for the RSS data indicated that the log-normal model may not fully characterize the data. In fact, if we had supposed H_0 to be a 2-component Gaussian mixture distribution (with parameters estimated from $R_{i,j}$ via the MLE), the KS test would have yielded a p-value of 0.88.

References

[1] N. Patwari, A. O. Hero III, M. Perkins, N. Correal, and R. J. O'Dea. Relative location estimation in wireless sensor networks. *IEEE Trans. Signal Process.*, 51(8):2137–2148, Aug. 2003.

[2] Wikipedia. Student's t-distribution. Online: `http://en.wikipedia.org/wiki/Student's_t-distribution`, accessed 28 Sept. 2010.