W1211 Introduction to Statistics
Lecture 12

Wei Wang

Oct 15th, 2012

Normal Probability Plot

- Because of the important role that Normal Distribution plays in statistical inference, we often want to assess whether a given sample is roughly normal distributed. Normal Probability Plot is used for this purpose.

Normal Probability Plot

- Because of the important role that Normal Distribution plays in statistical inference, we often want to assess whether a given sample is roughly normal distributed. Normal Probability Plot is used for this purpose.
- The basic strategy is to compare sample features with population features. In probability plot, we are using sample percentile(quantile) and population percentile(quantile), so it is also known as Q-Q plot.

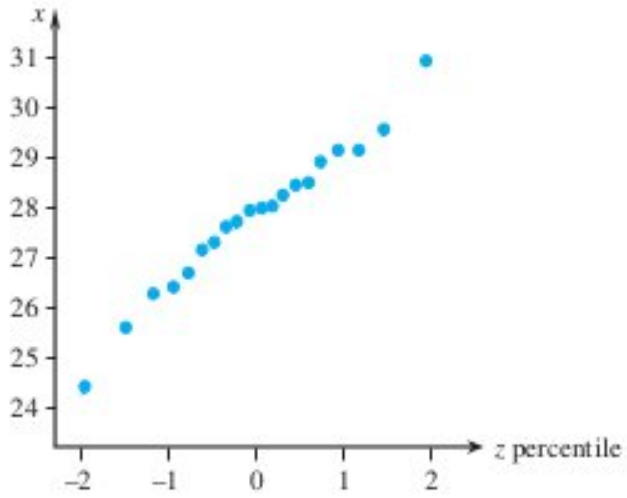
Normal Probability Plot

- Because of the important role that Normal Distribution plays in statistical inference, we often want to assess whether a given sample is roughly normal distributed. Normal Probability Plot is used for this purpose.
- The basic strategy is to compare sample features with population features. In probability plot, we are using sample percentile(quantile) and population percentile(quantile), so it is also known as Q-Q plot.
- The definition of a normal probability plot

> A plot of the n pairs
>
> ([100(i − .5)/n]th z percentile, ith smallest observation)
>
> on a two-dimensional coordinate system is called a **normal probability plot.** If the sample observations are in fact drawn from a normal distribution with mean value μ and standard deviation σ, the points should fall close to a straight line with slope σ and intercept μ. Thus a plot for which the points fall close to some straight line suggests that the assumption of a normal population distribution is plausible.

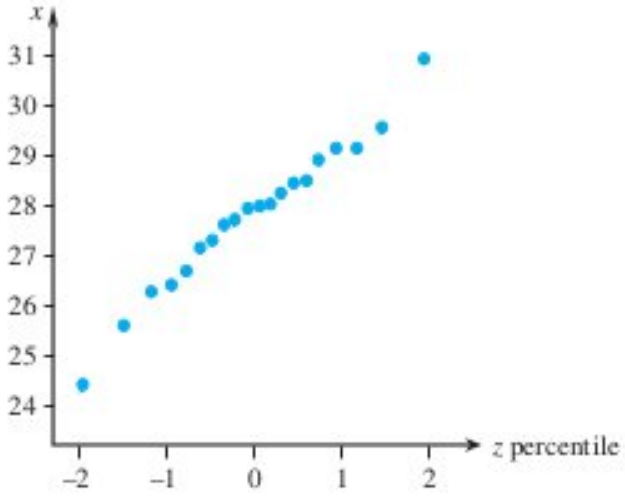
Examples of Normal Probability Plot

- A Normal Sample



Examples of Normal Probability Plot

- A Normal Sample



- Two Non-normal Samples

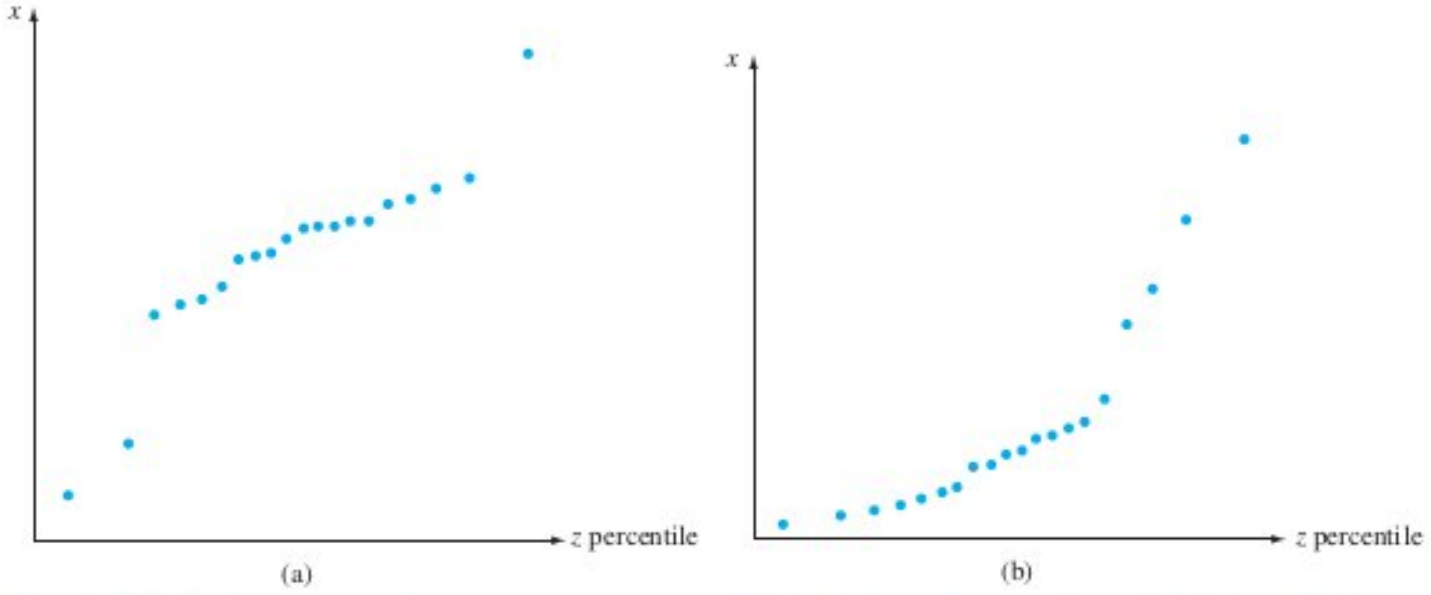


Figure 4.37 Probability plots that suggest a nonnormal distribution: (a) a plot consistent with a heavy-tailed distribution; (b) a plot consistent with a positively skewed distribution

Joint Distribution

- How can we model two rv's using probability models? For example, if we are interested in both weight and height.

- Is it enough if we just use a normal model for weight and another normal model for height?

- We need to introduce joint probability distribution in order to model multiple rv's.

Joint PMF

- Let X and Y be two discrete rv's defined on the sample space. The joint probability mass function $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X=x, Y=y).$$

- As in the single rv case, we must have $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$.

Marginal PMF

- The marginal probability mass functions of X and Y, denoted by $p_X(x)$ and $p_Y(y)$, respectively, are given by

$$p_X(x) = \sum_y p(x,y) \quad p_Y(y) = \sum_x p(x,y)$$

Ex.

p_{ij}	1	2	3	$p_X(x)$
1	4/9	2/9	0	> 2/3
2	1/9	1/9	1/9	> 1/3
	v|	v|	v|	
$p_Y(y)$	5/9	1/3	1/9	

- Notice that the marginal probability mass functions are automatically proper pmf's. (why?)

Two continuous rv's

- We would like to extend the same ideas to the continuous case. Let X and Y be continuous rv's. A joint probability density function $f(x, y)$ for these two variables is a function satisfying $f(x, y) \geq 0$ and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

- The marginal probability density function of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \quad \text{for } -\infty < y < \infty$$

Remarks

- In the continuous case, roughly speaking, $f(x, y)dxdy$ can be treated as $P(X=x,Y=y)$.
- $$\mathrm{P}(a < \mathrm{X} < b, c < \mathrm{Y} < d) = \int_a^b \int_c^d f(x,y)dxdy$$
- As in the discrete case, $f_X(x)$ and $f_Y(y)$ calculated from the joint distribution are automatically proper pdf's.
- Marginal distributions are, in fact, the distributions of the marginal random variables when they are treated as univariate random variables.

Example

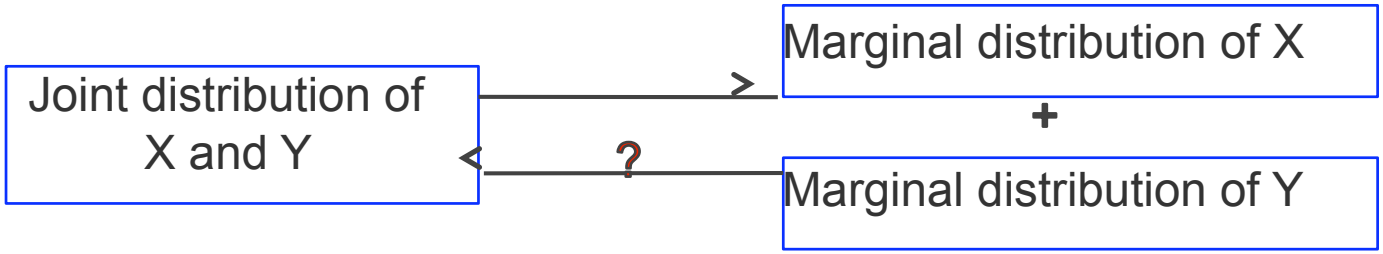
Ex. Suppose the joint pdf of the pair (X, Y) is given by

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

1. Show that this is a proper joint pdf.
2. What is P(0 ≤ X ≤ 1/4, 0 ≤ Y ≤ 1/4)?
3. What is P(0 ≤ Y ≤ 1/4)

Joint and Marginal

- Now we have



- In general, we CANNOT go the other way around. Further information about the dependence structure of X and Y is needed to determine the joint distribution.

Example

Ex. Consider the following two joint distributions of X and Y.

p_{ij}	0	1
0	$3/10$	$3/10$
1	$3/10$	$1/10$

p_{ij}	0	1
0	$9/25$	$6/25$
1	$6/25$	$4/25$

It is easy to see that the marginal distributions of X and Y are the same in both cases. P(X=0) = P(Y=0) = 3/5; P(X=1) = P(Y=1) = 2/5.

This is the example that *different* joint distributions may have the *same* marginal distributions.

Independent rv's

- Recall the definition of independence of two random events A and B.

$$P(A \cap B) = P(A)\,P(B)$$

- We say two random variables X and Y are independent if and only if

$$P(X=x, Y=y) = P(X=x)\,P(Y=y), \text{ for any } x \text{ and } y.$$

- More specifically, two random variables X and Y are said to be independent if for every pair x and y values,

$$p(x, y) = p_X(x)\,p_Y(y), \text{ when X and Y are discrete;}$$

or

$$f(x, y) = f_X(x)\,f_Y(y), \text{ when X and Y are continuous.}$$

Ex. The second case of the previous example.

Multiple Random Variables

- If $X_1, X_2, \ldots, X_n$ are all discrete random variables, the joint pmf of the variables is the function

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

If the variables are continuous, the joint pdf of $X_1, X_2, \ldots, X_n$ is the function $f(x_1, x_2, \ldots, x_n)$ such that for any n intervals $[a_1, b_1], \ldots, [a_n, b_n]$,

$$\mathrm{P}(a_1 \leq \mathrm{X}_1 \leq b_1, \ldots, a_n \leq \mathrm{X}_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

- What should be the regularity conditions for $p(x_1, x_2, \ldots, x_n)$ and $f(x_1, x_2, \ldots, x_n)$?

- How do get the marginal distributions of $X_1, X_2, \ldots$ by using $p(x_1, x_2, \ldots, x_n)$ and $f(x_1, x_2, \ldots, x_n)$?