Course Information

- Schedule and Homework
- Course Requirements
- Instructor's Information
- Links
- Main Page

Topics

1. Introduction
2. Summarizing Data: listing and grouping
3. Summarizing Data: Statistical Descriptions
4. Possibilities and Probabilities
5. Some rules of probability
6. Probability Distributions
7. Normal Distributions
8. Sampling Distributions
9. Problems of Estimation
10. Hypothesis Tests
11. Tests based on count data
12. Regression
13. Non-parametric Tests

Projects

1. Stock return distribution

- Forum

edit sec src prt

Site Manager

Chapter 7: Normal distribution

Continuous distributions

We saw in the previous chapter that when a random variable is a continuous random variable, the values it takes a values in a continuous interval, instead of discrete values.

For this reason, random variables that count items (e.g., number of successes) are not continuous.

Random variables that count for example time elapsed, on the contrary,
are continuous.

For a discrete random variable, the distribution is determined by the values that

$$f(k) = P(X = k) \quad (1)$$

takes for discrete values of k, for example
$k = 2, 3, 4, \ldots, 10, 11, 12$
when the procedure is tossing two dice and the random variable X describes the sum of the two sides that face upwards.

In the continuous case, $f(x)$ could take any value in an interval, not just integers or discrete (countable) values.

The conditions that we had for a function to be a discrete probability distribution, namely

1. $0 \leq f(k) \leq 1$
2. $\sum_k f(k) = 1,$

change slightly:

> A continuous probability distribution is the distribution

corresponding to a continuous random variable. The probability density function

$f(x) := P(X = x)$
satisfies the following properties:

1. $0 \leq f(x)$
2. Area under the function is 1, which can be written as $\int f(x)dx = 1$

Note that the conditions for a discrete probability distribution changed slightly in the continuous case: the requirement that the value
of the function is at most one is gone, and the sum has changed into the area (or integral).

For a continuous distribution, the function $f(x)$ is called **density function**.

Notice also that in the continuous case, the probability that the function is exactly equal to a specific value is zero, since a specific value has no area under the curve. Think for example about a waiting time in a line: what is the probability that you wait exactly 5.000000000000000000000000 minutes in line. Since, it is so unlikely that you wait exactly 5.000000000000000000000000 minutes in line (more likely it will be 5.0001 or 4.96 or 4.6, or 5.2, we say that P(X=5) = 0.

Some examples of continuous probability distributions follow.

Uniform distribution

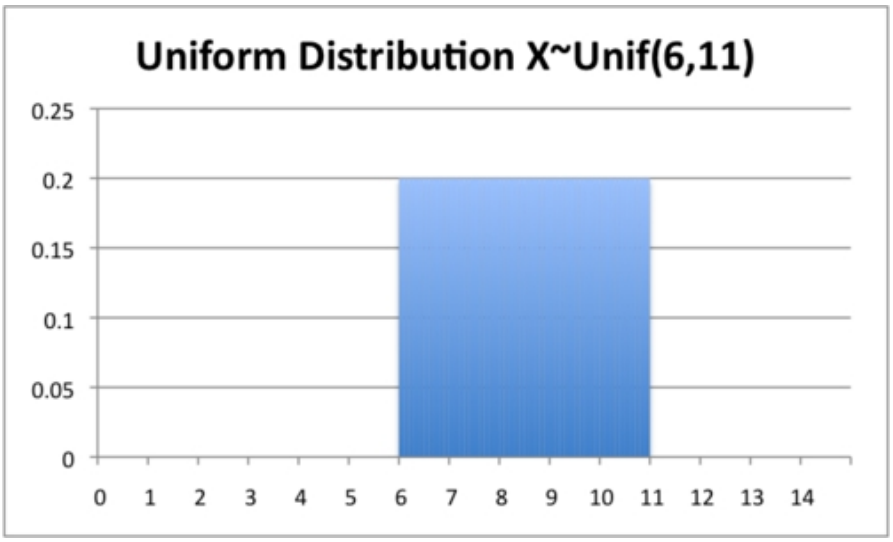
A uniform distribution is a continuous distribution that assigns the same probability to any subinterval of the same length in a given fixed interval. If X is a uniformly distributed random variable in the interval (a,b), then ***X~Unif(a,b)***,

Since the area under the curve has to be 1, the width is b-a, the height has to be 1/(b-a).

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{when } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

Example:

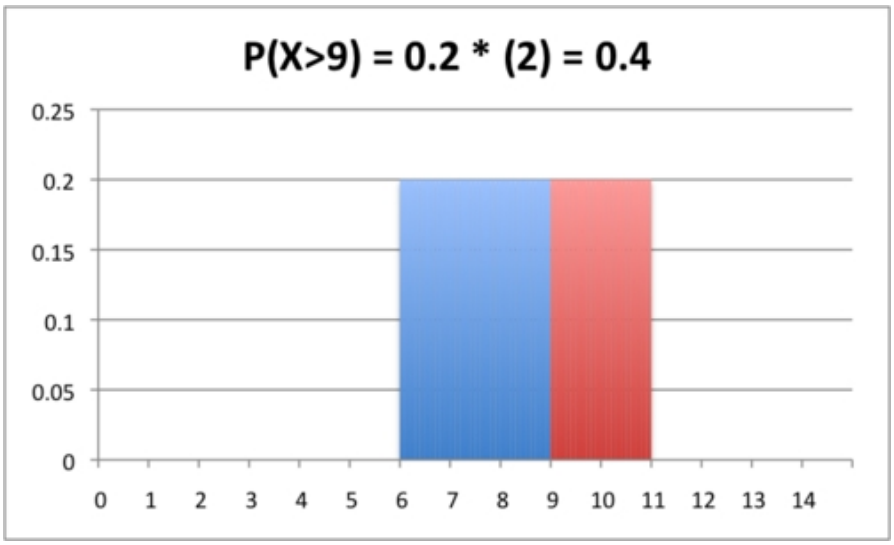
Suppose the waiting time at McDondals is uniformly distributed between 6 and 11 minutes at rush-hour, then the height of the function is 1/5=0.2, as can be seen in the graph below.



Then, the probability that the wait is longer than 9 minutes is given by:

P(X>9)= height * width = 0.2 * (11-9) = 0.2* 2 = 0.4

That area is represented in the graph below.



Remarks:

Note that in the case of a uniform distribution the area is easy to calculate geometrically, since we know the area of a rectangle.

In general, the area under a (continuous) curve $f(x)$ between $x=a$ and $x=b$ is given by the integral

$$P(a \leq X \leq b) = \int_a^b f(x)dx. \tag{3}$$

However, this is a topic in calculus 2, and even then, in the case of the normal distribution, the **density function**, which is the function that needs to be integrated

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

cannot be integrated in terms of functions that we can express in

terms of exponentials, polynomials trigonometric, logarithmic and rational functions. Therefore, we use a table to approximate the values of the area, as explained below.

Normal distribution

A random variable *X* is normally distributed with mean mu (μ) and standard deviation σ if the density function describing its probability distribution is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (5)$$

While we will not need to know this formula by hard, notice that when graphing this function as a function of *x*, two parameters have to be specified:

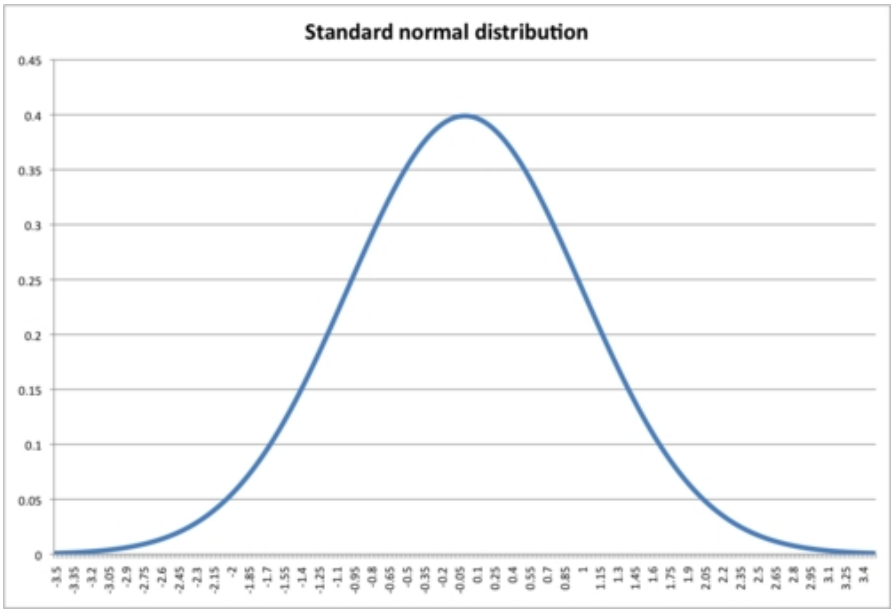
- μ: the mean
- σ: the standard deviation

Furthermore, for any μ and σ that we choose we obtain different values for $f(x)$.

Standard Normal Distribution

If $\mu = 0$ and $\sigma = 1$, then we say that the random variable follows a **standard normal distribution** and then it is often denoted by ***Z***, instead of ***X***.

The graph representing the probability density in this case is given below:



As mentioned above, the area above the *x*-axis and under the curve has to be one, and the area under the curve represents the probability.

For example, P(-2<X<2) is the area under the curve between x=-2 and x=2. Since the standard deviation is 1, this represents the probability that a normal distribution is between 2 standard deviations away from the mean, and we know from before that this is 0.95.

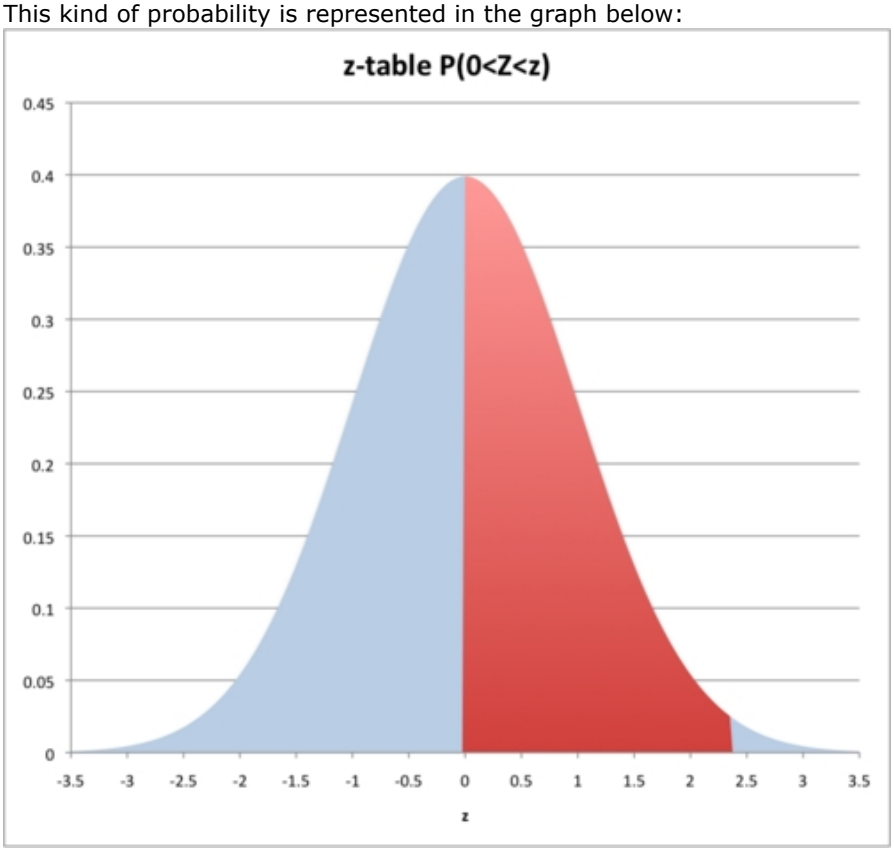
z-score table

In general, we use the table (often called z-score table) given in the link below to calculate probabilities for the standard normal distribution:
Standard normal distribution table

The table that you can download from the link above give you probabilities (areas under the standard normal curve) of the form

$$P(0 \leq Z \leq z) \qquad (6)$$

This kind of probability is represented in the graph below:



where the area in red represents the probability that Z is between z=0 and z=2.34, or

P(0<Z<2.34)

To read this probability (area) from the table in the link above (or here again for convenience), split z into two parts, its units and tenths, i.e.

> 2.3 (which determines the row to read from)

and its hundredths, i.e.

> 0.04 (which determines the column to read from)

All the values in the middle of the table are probabilities (areas under the curve):

z	**...**	**.03**	**.04**	**.05**	**...**
...	...	...	**...**	...	...
2.3	**...**	**.4901**	***.4904***	**.4906**	**...**

We find that

P(0<Z<2.34) = 0.4904.

To find probabilities like

P(Z>1.28) we exploit the fact that the probability under the curve is 1, and that on the positive side the area is 0.5.

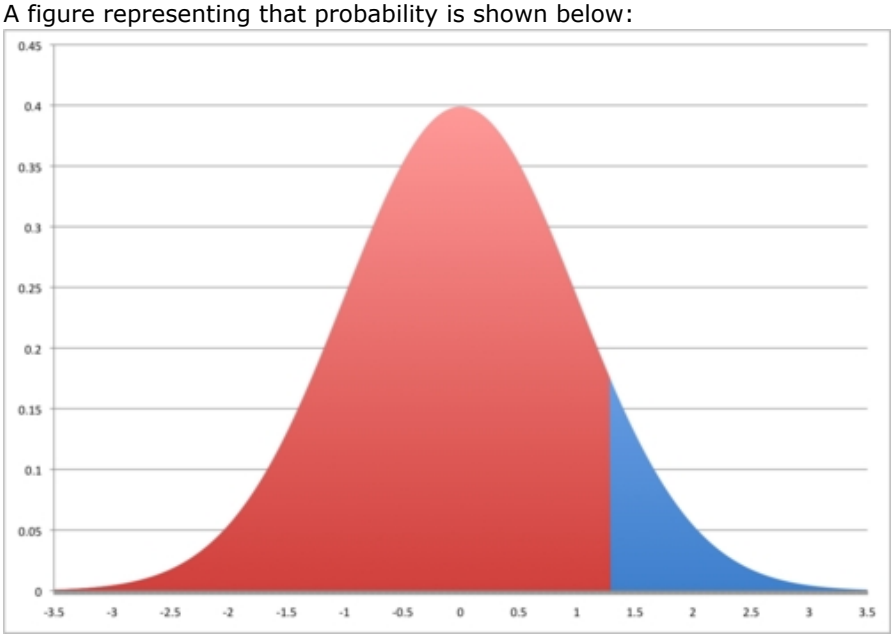
Therefore, we find in the table that
P(0<Z<1.28) = 0.4162
Indeed:

z	**...**	**.07**	**.08**	**.09**	**...**
...	...	...	**...**	...	...
1.2	**...**	**.3979**	***.3997***	**.4014**	**...**

Therefore,

P(Z>1.28)

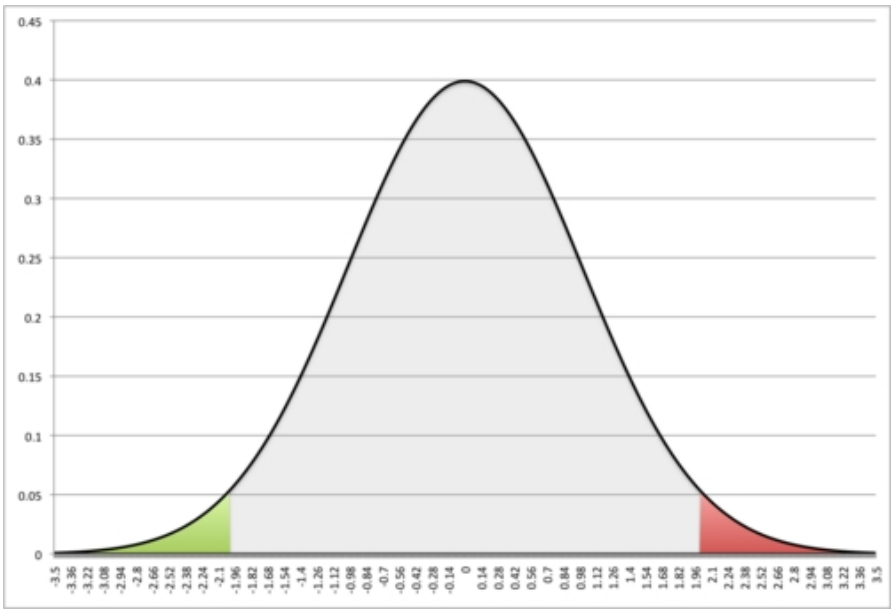
=0.5 - P(0<Z<1.28) = 0.5 - 0.4162 = 0.0838.
A figure representing that probability is shown below:



Let us now try to find

P(Z<-2.01)

Again, we can exploit the symmetry of the normal distribution and find the area P(Z>2.01). Both these probabilities are equal, because of symmetry, as can be seen in the graph below:



Therefore, reading from the table we obtain:

z	.00	.01	.02	...
...	...	...	...	...
2.0	**.4772**	***.4777***	**.4783**	**...**

Therefore,
P(Z<-2.01) = P(Z>2.01) = 0.5-P(0<Z<2.01) = 0.5 - 0.4777 = 0.0233

Percentiles or z-scores

If instead of looking for the probability, we look for the value of z for a given **percentile** or probability (or area), the problem is typically called finding the z-score.

Example: Suppose that we were looking for the value of z so that the area under the curve, to the left of z is 0.90. That value that we find is called a **critical value**, and is sometimes denoted as $\mathbf{z_{90}}$, indicating that 90% of the area is to the left of that value. To find the exact value of $\mathbf{z_{90}}$, we would have to look in the center of the table were the table for the value 0.40, since the table only covers areas from 0 to a value of z, and we would find that the table changes from 0.3997 to 0.4014 precisely between 1.28 and 1.29.

Similarly, if we are looking for the value of z such that the area under the curve to the left of z is 0.10, then by analogy with the paragraph above we are looking for the critical value $\mathbf{z_{10}}$, indicating that 10% of the area is to the left of this value. This value should be negative since the area to its left is less that 0.50. By the symmetry of the normal density curve, the area between 0 and $\mathbf{z_{10}}$ is the same as the area between 0 and $\mathbf{z_{90}}$, namely, 0.40. Hence $\mathbf{z_{10}}$ = - $\mathbf{z_{90}}$ = - 1.285.

z-score Linear Interpolation

To be very precise, the value is not 1.28 nor 1.29, so we would have to (linearly) interpolate, namely, the value 0.4000 is closer to 0.3997 than to 0.4014, but how much closer?

To determine that we need to do the following calculation:
(0.4-0.3997)/(0.4014-.3997)=0.1764706
which tells us that the exact value is approximately 1.28 + 0.1764706* (0.01) = 1.282.

The following z-score are often looked for and will be used later:

P(Z<z)	0.80	0.90	0.95	0.99
z	0.841	1.282	1.645	2.326

Standarization

In most cases in which the normal distribution plays a role, the mean is not zero and the standard deviation is not 1. Luckily, one can transform any normal distribution with a certain mean μ and standard deviation σ into a standard normal distribution, by the *z-score conversion formula*

$$z = \frac{x - \mu}{\sigma} \tag{7}$$

Example

Using an example we have seen before, assume that the height of women in the US is normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches, find

1. the probability that a randomly selected woman is taller than 70.4 inches (5 foot 10.4 inches).
2. the probability that a randomly selected woman is between 60.3 and 65 inches tall.

1. Since the height of women follows a normal distribution but not a standard normal, we need to standardize first. Since x=70.4, the mean μ = 64 inches and the standard deviation is σ = 2.5 inches, we need to calculate z:

$$z = \frac{x - \mu}{\sigma} = \frac{70.4 - 64}{2.5} = \frac{6.4}{2.5} = 2.56 \tag{8}$$

Therefore, the probability P(X>70.4) is equal to P(Z>2.56), where X is the normally distributed height with mean μ = 64 inches and the standard deviation σ = 2.5 inches (X~N(64,2.5) for short), and Z is a standard normal distribution (Z~N(0,1)).

Therefore using the table

P(X>70.4) = P(Z>2.56) =0.5 - 0.4948 = 0.0012.

2. For the second problem we have to values of x to standarize, x_1=60.3 and x_2=65. Standarizing these values we obtain:

z_1=-1.48 and z_2=0.40.

Notice that the first value is negative, which means that it is below the mean.

Therefore,

P(60.3<X<65) = P(-1.48<Z<0.40)= P(Z<0.40)-P(Z<-1.48) = (0.5+0.1554)-(0.5-0694) =0.6554-0.0694=0.5860.

Percentiles or x-score examples

If instead of looking for the probability, we look for the value of x for a given percentile (= probability or area), the problem can be called finding the x-score.

Example: Suppose that for the example above we try to find the height corresponding to the 90th percentile. In other words we are trying to find the height that corresponds to being taller than 90 percent of the women (in this case).
the area under the curve, to the left of z is 0.90. That value is then called a **critical value**, and is sometimes denoted as $\mathbf{z_{90}}$, indicating that 90% of the area is to the left of that value. To find the exact value of $\mathbf{z_{90}}$, we would have to look in the center of the table were the table for the value 0.40, since the table only covers areas from 0 to a value of z, and we would find that the table changes from 0.3997 to 0.4014 precisely between 1.28 and 1.29.