Continuous Probability Distributions
Lecture: 5 + 6

Hamid Raza
Assistant Professor in Economics
raza@business.aau.dk

Aalborg University

Statistics - Statistik

Outline

Continuous probability distribution:

- we extend the probability concepts to continuous random variables and probability distributions. The concepts and insights for discrete random variables also apply to continuous random variables

Cumulative Distribution Function:

The cumulative distribution function, $F(x)$, for a continuous random variable X expresses the probability that X does not exceed the value of x, as a function of x:

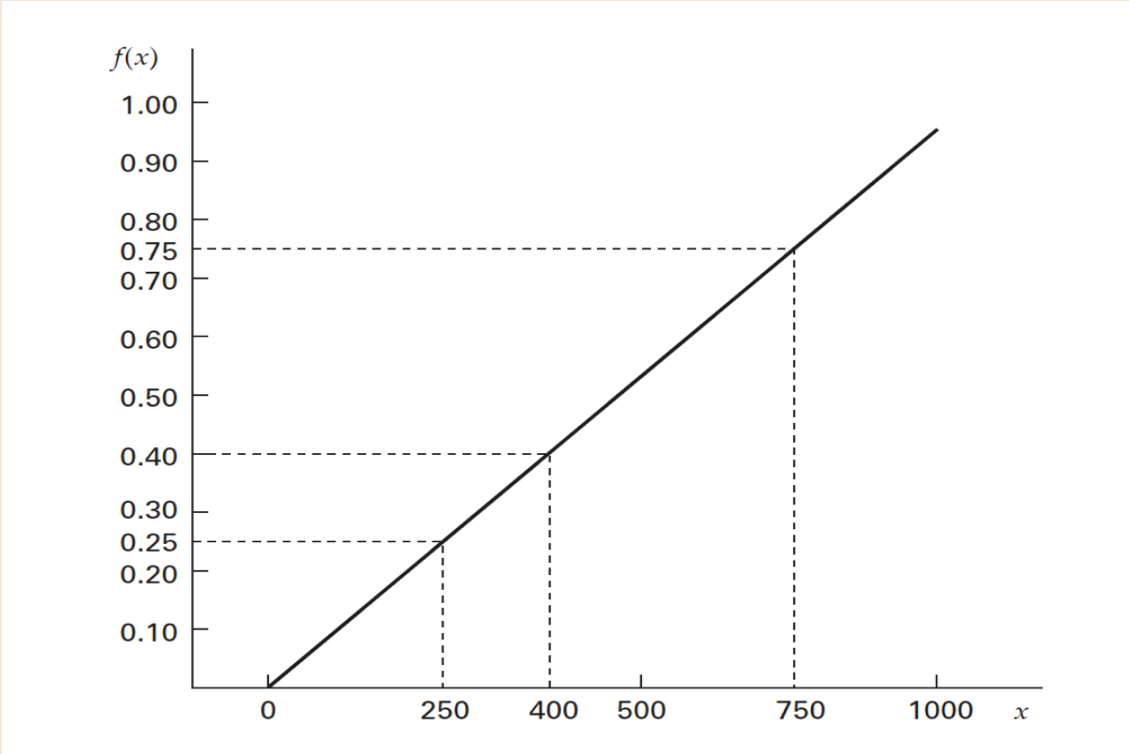
$$F(x) = P(X \leq x)$$

- Consider a petrol station that has a 1000 gallon storage
- The lowest limit it can sell is 0 and the upper limit is 1000 gallons
- Assume that selling any amount of petrol between 1 and 1000 is equally like Then, the distribution of X is said to follow a uniform probability distribution, with the probability of 0.001
- the cumulative distribution is: $F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 0.001x, & \text{if } 0 \leq x \leq 1000 \\ 1, & \text{if } x \geq 1000 \end{cases}$

Continuous probability distribution:

Cumulative Distribution Function (cont):

The function is graphed below:



- The probability of sales between 0 and 400 gallons is as follows:

$$P(X \leq 400) = F(400) = 0.001(400) = 0.4$$

- What is the probability of sales between 0 and 1000?

Continuous probability distribution:

Cumulative Distribution Function (cont):

- Can we calculate the probability of selling 400 gallons of petrol? No
- **Very Important:** In discrete probability cases, we were calculating the probability of a specific value that a discrete variable can take
- The probability of a specific value is 0 for continuous random variables, that concept is not directly relevant here
- Therefore, we use area and ranges instead of pin-pointing the probability of a value
- There is a simple formula that we can use to calculate probabilities for continuous random variable

Continuous probability distribution:

Probability of a Range Using a Cumulative Distribution Function:

Let X be a continuous random variable with a cumulative distribution function F(x), and let a and b be two possible values of X, with $a < b$

- The probability that X lies between a and b is as follows:

$$P(a \leq X \leq b) = F(b) - F(a)$$

- Example: For example, the probability of sales between 250 and 750 gallons is:

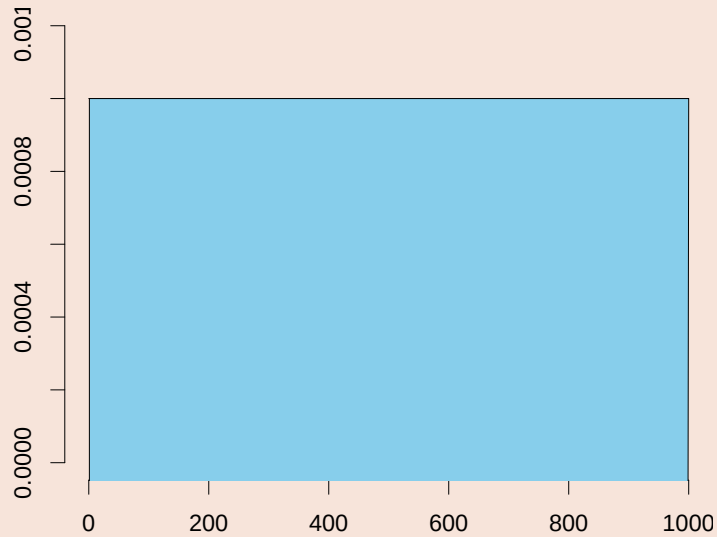
$$P(250 \leq X \leq 750) = 0.001(750) - 0.001(250) = 0.5$$

Continuous probability distribution in R:

Probability of a Range Using a Cumulative Distribution Function:

- R has built in package for different types of probability distributions
- In our example of sales of gallons, we used uniform distribution. We can easily visualize this distribution

```
x=seq(1, 1000, 1);          y=dunif(x, 0, 1000)
curve(dunif(x, 0, 1000), from = 0, to = 1000, bty="n", ylim=c(0.000, 0.0012))
```



Continuous probability distribution in R:

Probability of a Range Using a Cumulative Distribution Function:

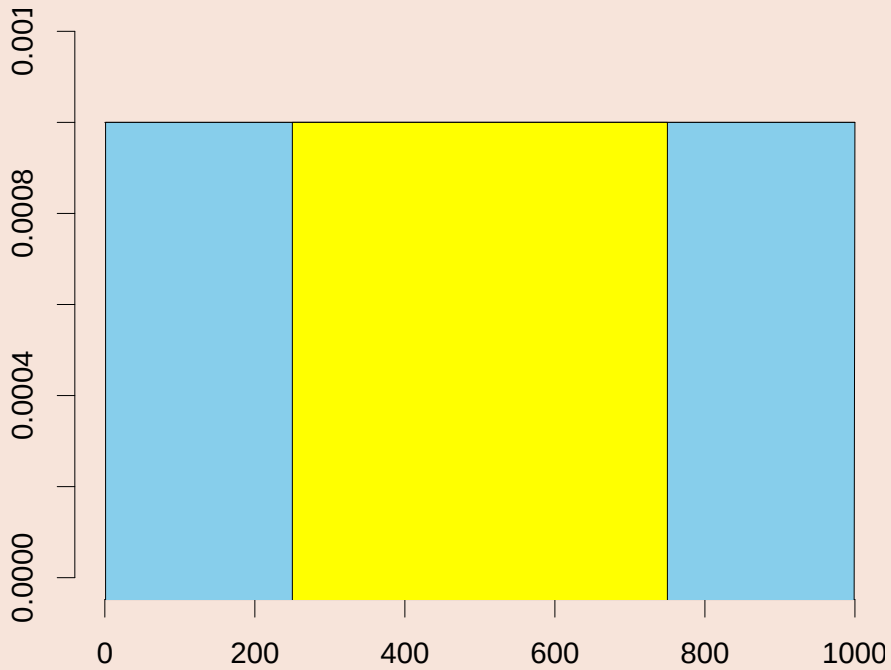
- We can calculate the probability of sales between 250 and 750 gallons and then simply use the formula: $F(b) - F(a)$:

```
punif(750 , min = 0 , max = 1000)   -  punif(250 , min = 0 , max = 1000)
```

```
## [1] 0.5
```

- we can graphically show the region between 250 and 750:

```
curve(dunif(x, 250, 750), from = 0, to = 1000, bty="n")
```



Outline

Continuous probability distribution

Probability density function

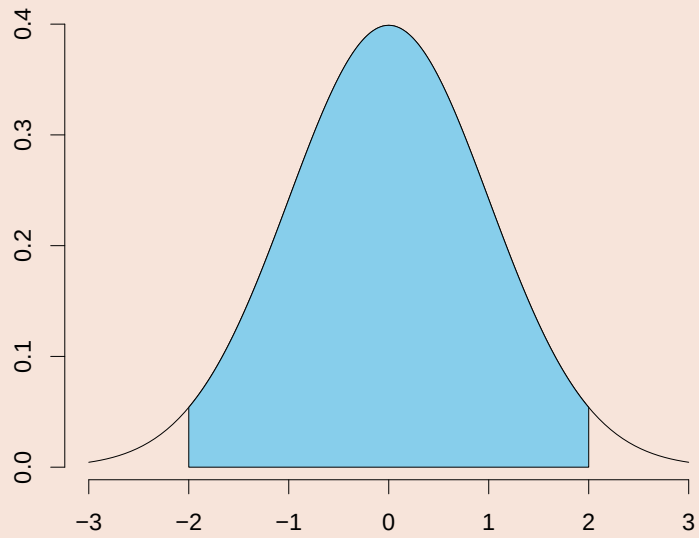
- Let X be a continuous random variable, and let x be any number lying in the range of values for the random variable
- The probability density function, $f(x)$, of the random variable is a function with the following properties:
  1. $f(x) > 0$ for all values of x
  2. The area under the probability density function, $f(x)$, over all values of the random variable, X within its range, is equal to 1
  3. Suppose that this density function is graphed. Let a and b be two possible values of random variable X, with $a > b$. Then, the probability that X lies between a and b is the area under the probability density function between these points
  $$P(a \leq X \leq b) = \int_a^b f(x)dx$$
  4. The cumulative distribution function, $F(x_0)$, is the area under the probability density function, $f(x)$, up to x_0,
  $$F(x_0) = \int_{x_m}^{x_0} f(x)dx$$

Continuous probability distribution

Probability density function

- Figure below shows the plot of a probability density function for a continuous random variable

```
curve(dnorm(x,0,1), bty="n", xlim=c(-3,3))
x <- c(-2,seq(-2,2,0.01),2)
y <- c(0,dnorm(seq(-2,2,0.01)),0)
polygon(x,y,col='skyblue')
```



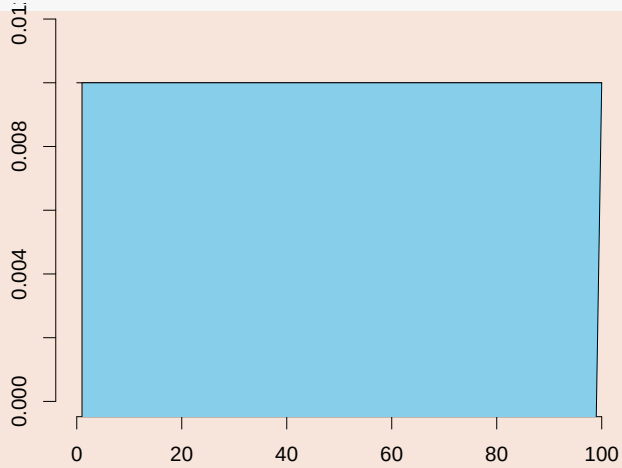
- Two possible values, -2 and +2 , are shown, and the shaded area under the curve between these points is the probability that the random variable lies in the interval between them

Continuous probability distribution

Uniform distribution:

- Figure below is a graph of the uniform probability density function over the range from 0 to 100. The probability density function

```
x=seq(1, 100, 1);          curve(dunif(x, 0, 100), from = 0, to = 100, bty="n", ylim=c(0, 0.012))
```



The probability density function (fx) of the continuous uniform distribution is:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

- For example, if we have a uniform distribution ranging from 0 to 100, then the probability density function is: $f(x) = \begin{cases} 0.01, & \text{if } 0 \leq x \leq 100 \\ 0, & \text{otherwise} \end{cases}$

Outline

Mean and Variance

Mean:

- X is a continuous random variable with density function $f(x)$
- The mean/expected value of X is given by:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

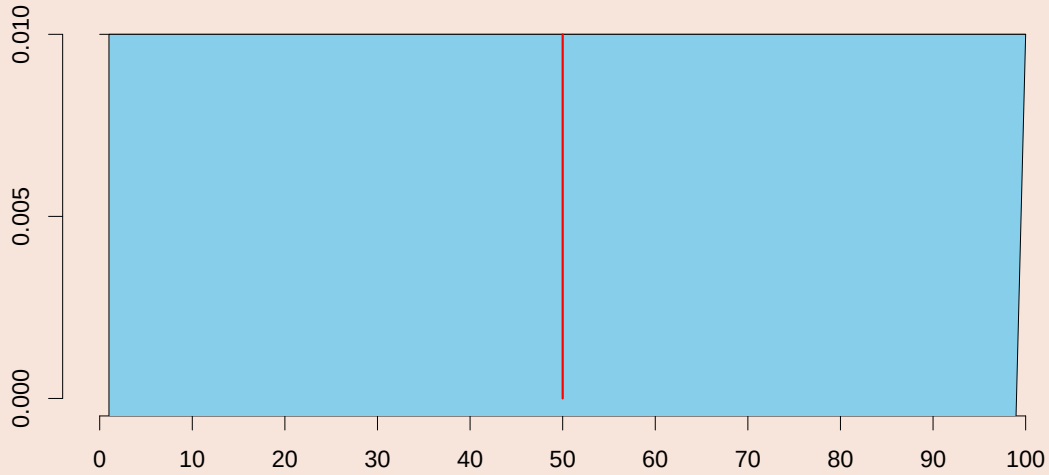
Note: the sum has changed to an integral

Mean and Variance

Mean:

- Figure below is a graph of the uniform probability density function over the range from 0 to 100. The probability density function

```
x=seq(1, 100, 1);                curve(dunif(x, 0, 100), from = 0, to = 100, bty="n", ylim=c(0, 0.012))
```



- The mean provides a measure of the center of the distribution
- For a uniform distribution defined over the range from a to b, we have the following results: $f(x) = \frac{1}{b-a}$, whereas mean is

$$\mu_X = \frac{a+b}{2}$$

Mean and Variance

Variance:

- X is a continuous random variable with density function $f(x)$
- variance of X is given by:

$$\sigma_X^2 = Var(X) = E[(X - \mu_x)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx$$

Note: the sum has changed to an integral

Mean and Variance

Rules for mean and variance:

The rules for mean and variance in the case of continuous random variables are the same as for discrete variables

- If g is a function and X is a continuous random variable, we get a new random variable $Y = g(X)$ with mean:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

- If g in a linear function such that $Y = g(X) = a + bX$, the mean and variance is calculated as follows:

$$\mu_Y = E(a + bX) = a + bE[X] = a + b\mu_X$$

The variance in this case will be:

$$\sigma_Y^2 = Var(Y) = E[a + bX] = b^2 Var(X) = b^2\sigma_X^2$$

Outline

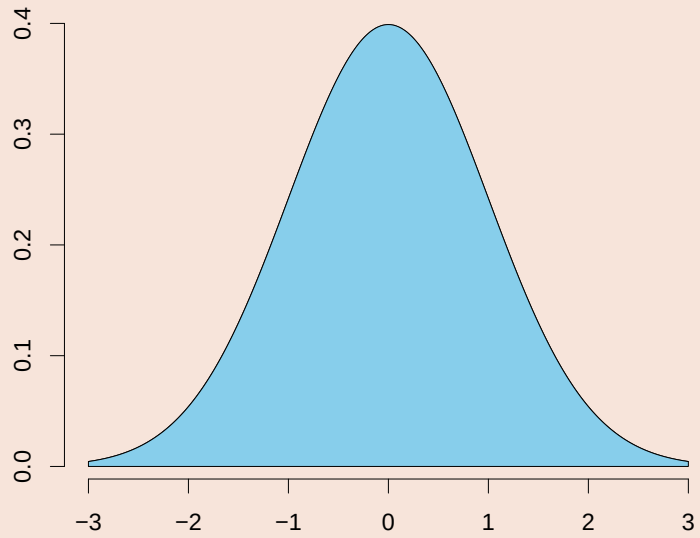
Continuous probability distribution

Normal Distribution:

Normal probability distribution, which is the continuous probability distribution used most often for economics and business applications

- Normal distribution refers to the following shape:

```
curve(dnorm(x,0,1), bty="n", xlim=c(-3,3))
```



Continuous probability distribution

Probability Density Function of the Normal Distribution:

The probability density function for a normally distributed random variable X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \qquad \text{for } -\infty < x < \infty$$

- where μ and σ^2 are any numbers such that $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$
- where e and π are physical constants, $e = 2.71828\ldots$, and $\pi = 3.14159$

Continuous probability distribution

Properties of the Normal Distribution:

- There is a family of normal distribution curves which are determined by 2 parameters (μ and σ):
  1. μ is the mean (expected value), which determines where the distribution is centered
  2. σ is the standard deviation, which determines the spread of the distribution about the mean
  3. The shape of the probability density function is a symmetric bell-shaped curve centered on the mean, μ
  4. If we know the mean and variance, we define the normal distribution by using the following notation:

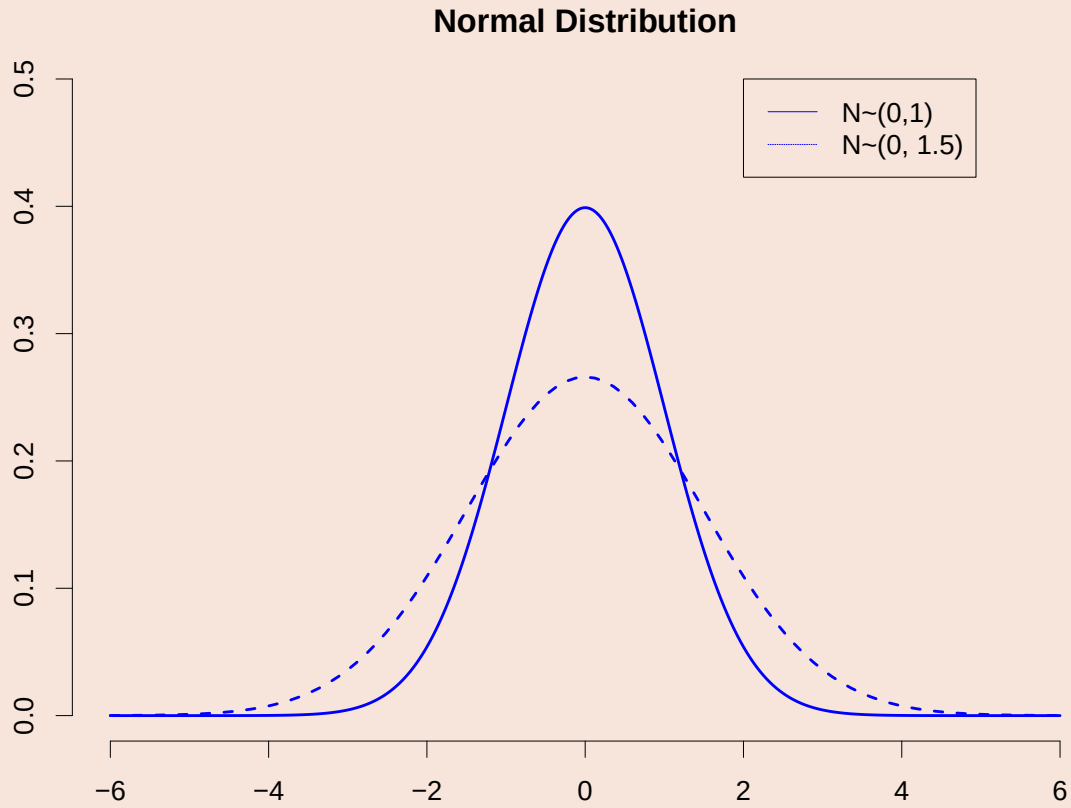
$$X \sim N(\mu, \sigma^2)$$

Continuous probability distribution

Properties of the Normal Distribution:

Two normal distributions with same mean (μ) and different standard deviation (σ):

```
x <- seq(-6,6,length=500)
plot(x,dnorm(x,mean=0,sd=1),type = "l",lty=1,lwd=3,col="blue")
curve(dnorm(x,0,1.5),add=TRUE,lty=2,col="blue", lwd=3)
```

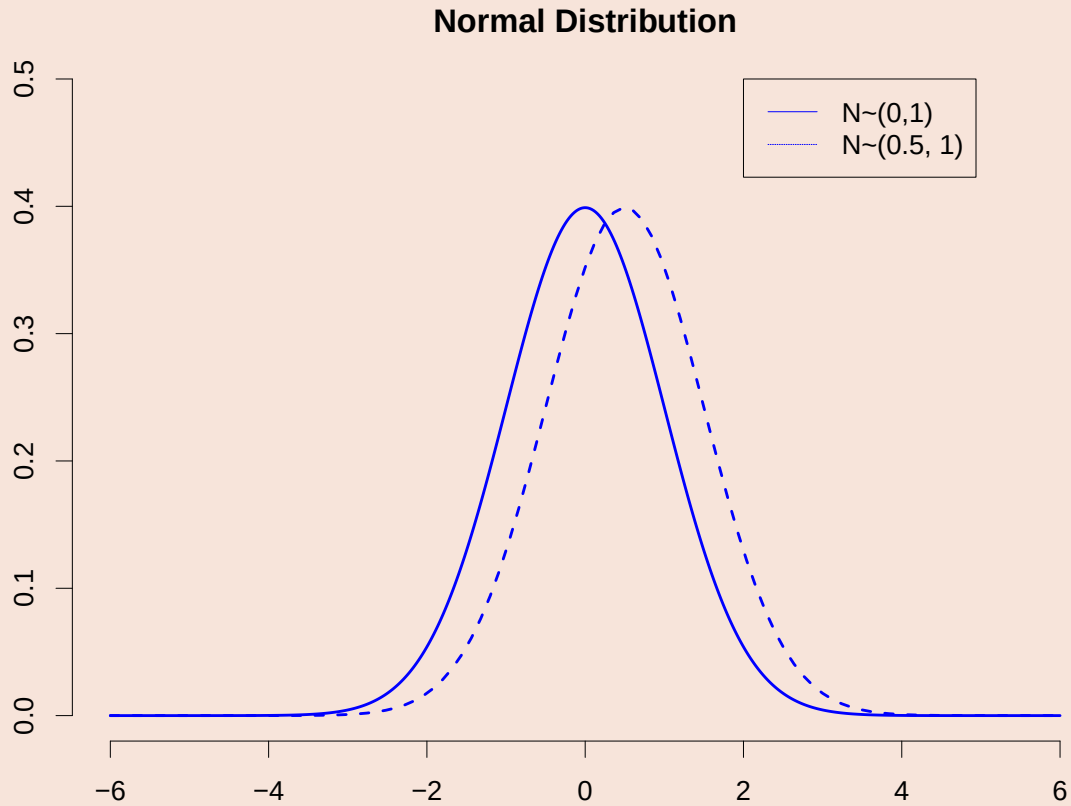


Continuous probability distribution

Properties of the Normal Distribution:

Two normal distributions with different mean (μ) and equal standard deviation (σ):

```
x <- seq(-6,6,length=500)
plot(x,dnorm(x,mean=0,sd=1),type = "l",lty=1,lwd=3,col="blue")
curve(dnorm(x,0.5,1),add=TRUE,lty=2,col="blue", lwd=3)
```

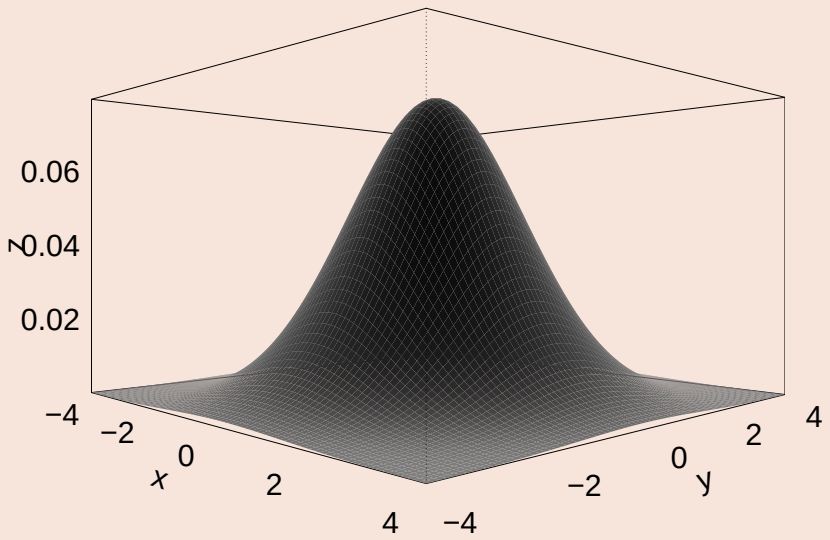


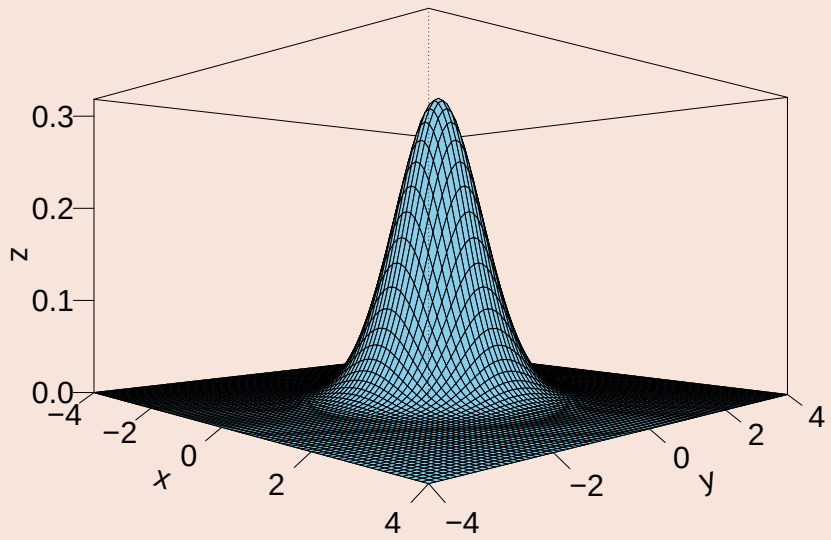
Continuous probability distribution

Properties of the Normal Distribution:

You can create some 3D fancy plots in R as well:

```
library(mnormt)
mu <- c(0,0);                                    sigma <- matrix(c(2,0,0,2),2,2)
x<-seq(-4,4,0.1);                                y<-seq(-4,4,0.1)
f<-function(x,y){dmnorm(cbind(x,y), mu, sigma)}; z<-outer(x,y,f)
persp(x,y,z, box=T, ticktype="detailed", theta=47, phi=6, expand = 0.65)
```





Continuous probability distribution

Cumulative Distribution Function of the Normal Distribution:

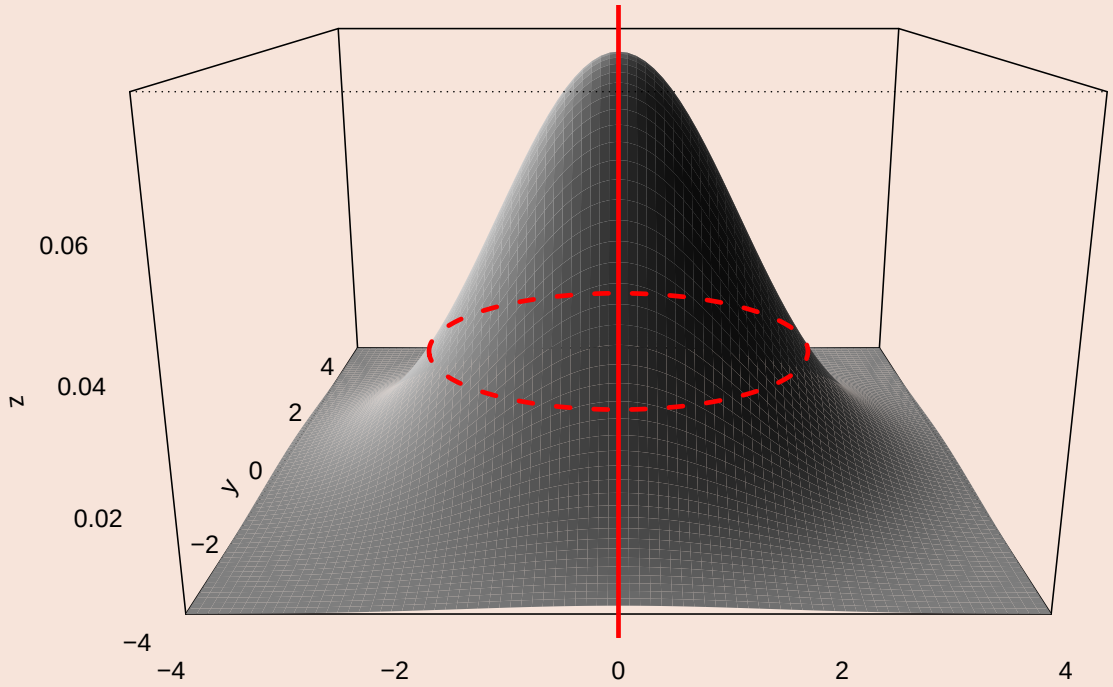
Then the cumulative distribution function of the normal distribution is the same as for any continous random variable:

$$F(x) = P(X \leq x_0)$$

The above equations just shows the area under the normal probability density function to the left of x_0

Continuous probability distribution

Properties of the Normal Distribution:



$\mu = 0$ in this plot; $P(X) \leq \mu$ is the region to the left of the distribution

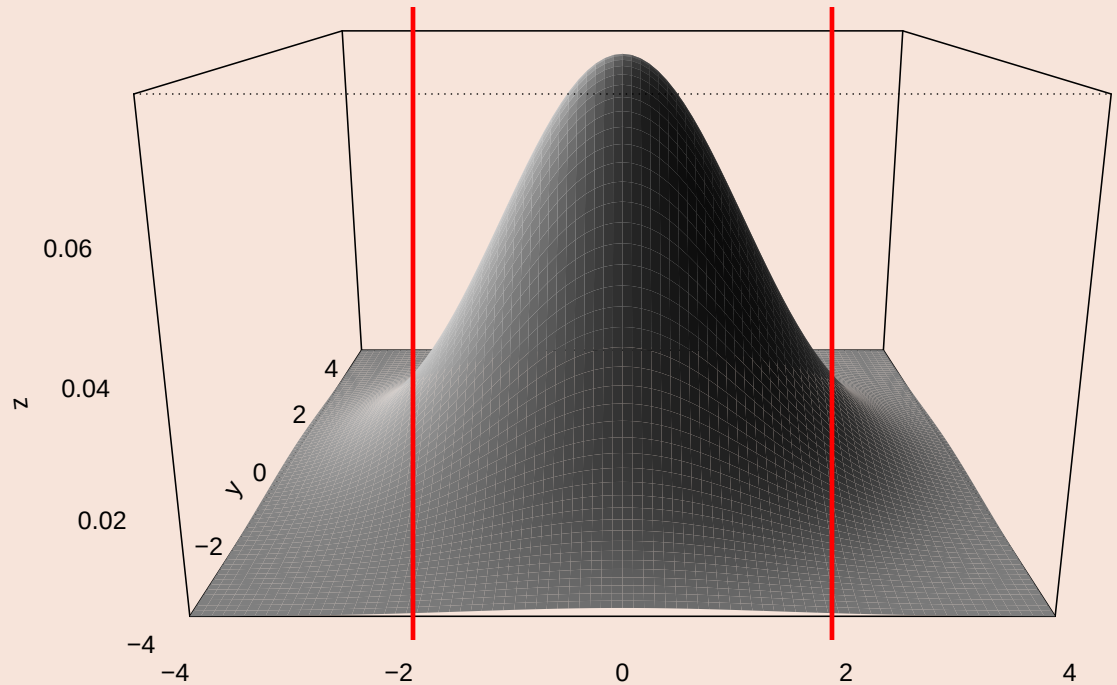
Like any proper density function, the total area under the curve is 1

Continuous probability distribution

Range Probabilities for Normal Random Variables:

Let X be a normal random variable with cumulative distribution function $F(x)$, and let a and b be two possible values of X, with $a < b$ Then,

$$P(a < X < b) = F(b) - F(a)$$



Continuous probability distribution

Standard Normal distribution:

We say that a random variable Z follows the standard normal distribution, when it has a mean $(\mu = 0)$, and $(\sigma = 1)$, such that:

$$Z \sim N(0, 1)$$

Normal z-score:

- We can obtain probabilities for any normally distributed random variable by first converting the random variable to the standard normally distributed random variable, Z
- To transform a normally distributed variable X to a standard normal distribution, we calculate its Z-score using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

The above gives us Z-score

Continuous probability distribution

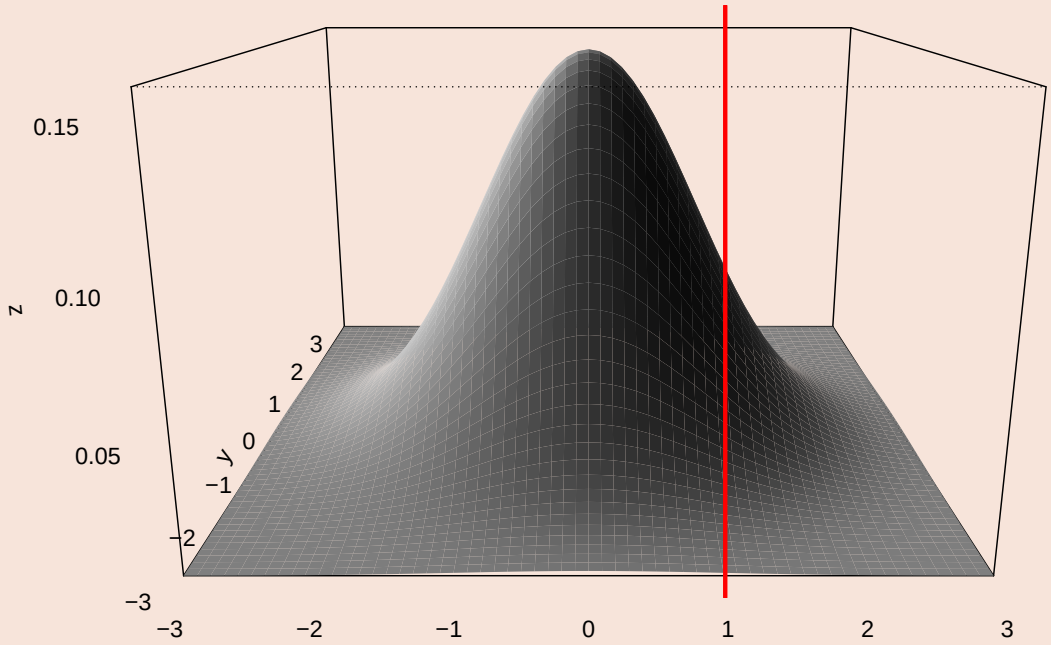
Normal z-score (cont):

- Z counts the number of standard deviations that the observation lies away from the mean, where a negative value tells that we are below the mean
- After calculating Z-score, we can use the standard normal table to compute probabilities of any random normally dist. variable X
- Table that indicates the probability for various intervals under the standard normal distribution can be found in your statistics book

Continuous probability distribution

Normal z-score (cont):

- But we can also use R to calculate probabilities for various intervals in a standard probability distribution. Consider the standard normal



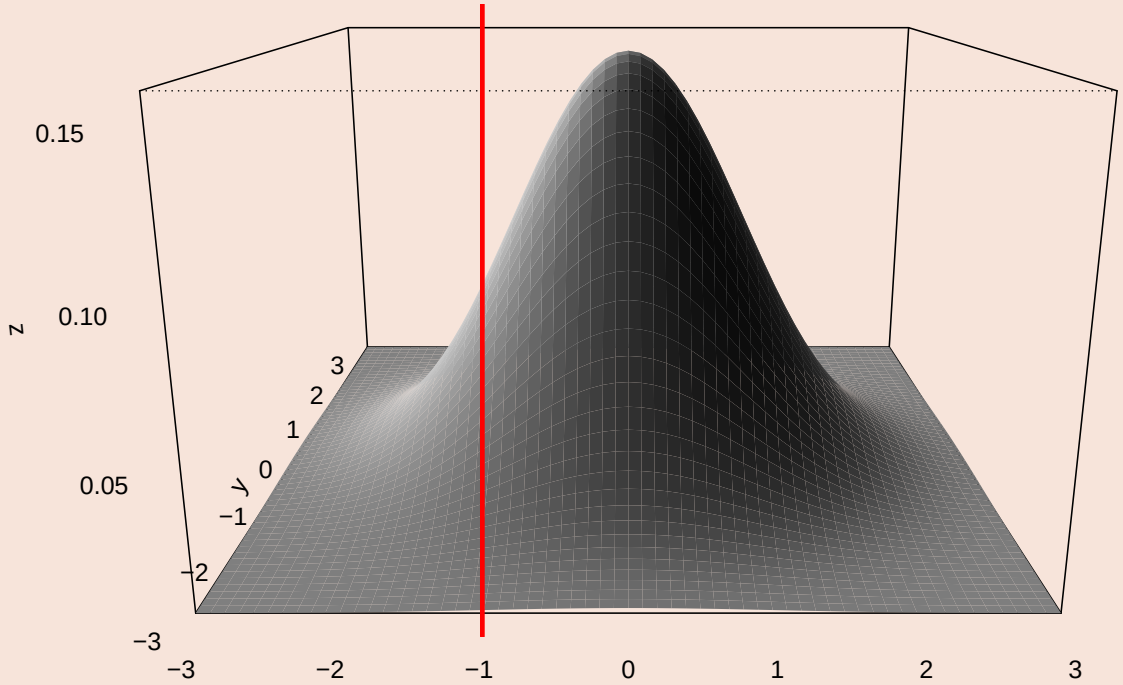
distribution:
What is the probability associated with the area left to 1?

```
pnorm(1, mean = 0, sd = 1)

## [1] 0.8413
```

Continuous probability distribution

Normal z-score (cont):



- What is the probability associated with the area left to -1?

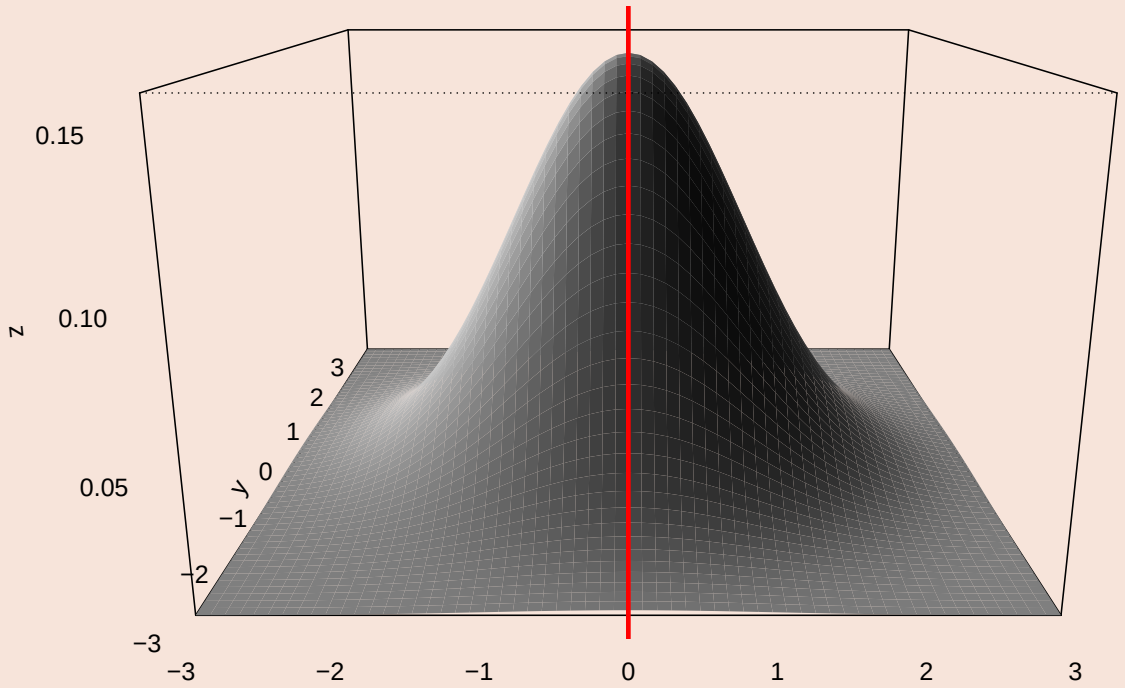
```
pnorm(-1, mean = 0, sd = 1)
```

```
## [1] 0.1587
```

Note that the area right to the -1 is 0.84 due to symmetric shape

Continuous probability distribution

Normal z-score (cont):



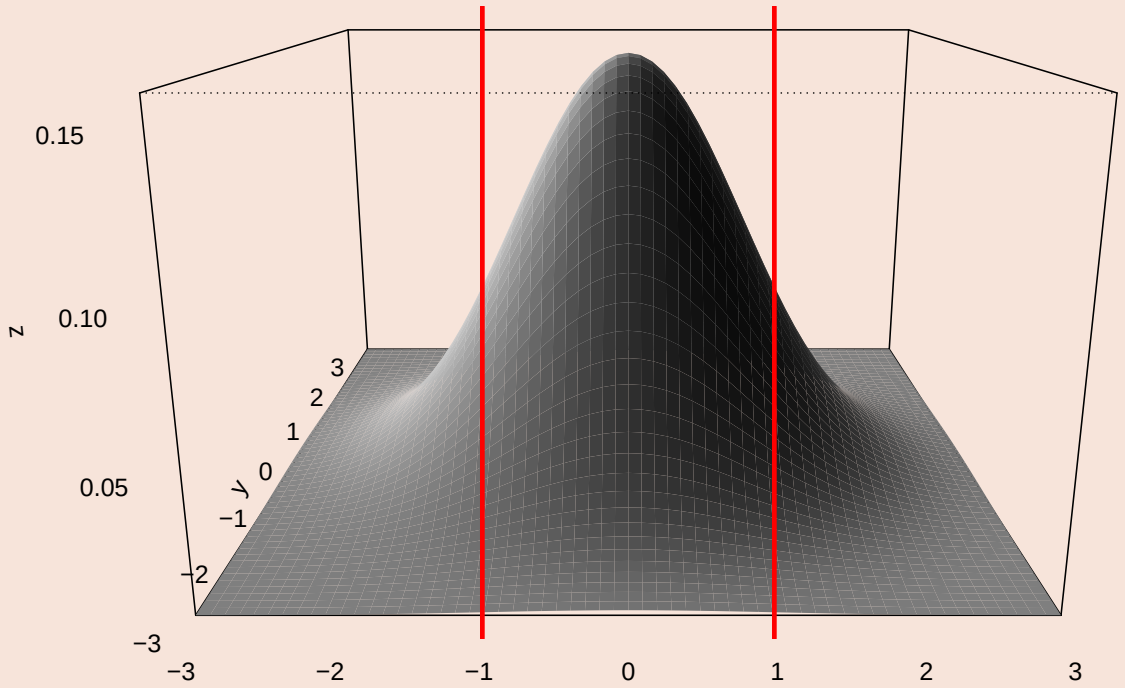
- What is the probability associated with the area left to mean $\mu = 0$?

```
pnorm(0, mean = 0, sd = 1)
```

```
## [1] 0.5
```

Continuous probability distribution

Normal z-score (cont):



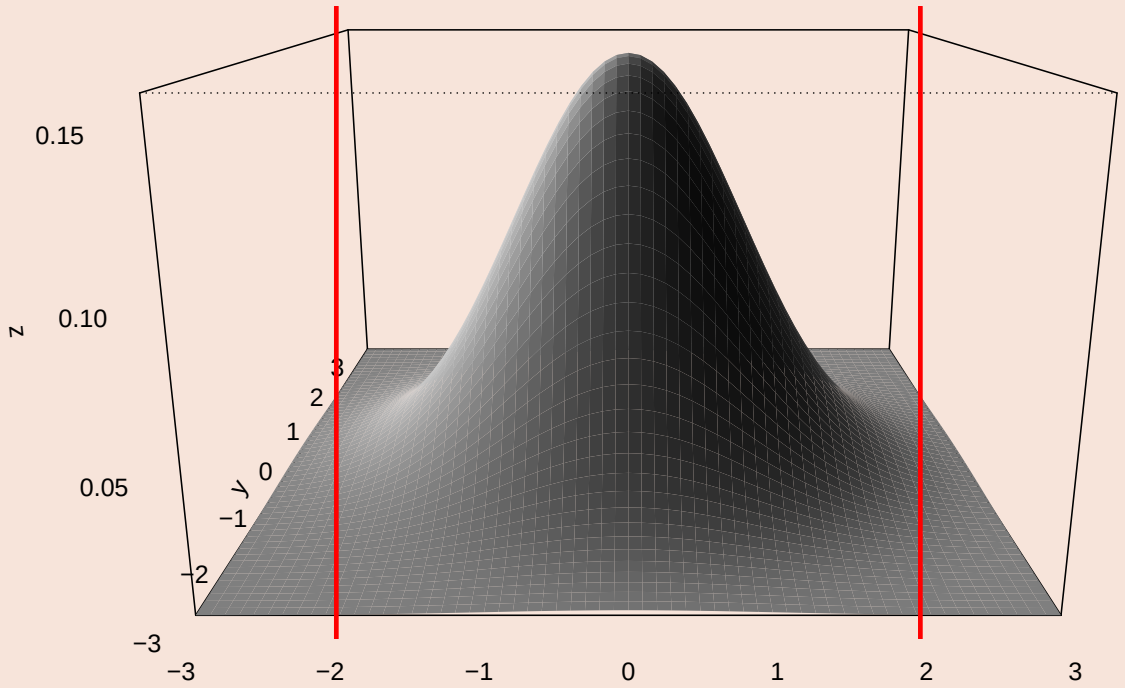
- What is the probability associated with the area between +1 and -1?

```
pnorm(1, mean = 0, sd = 1) - pnorm(-1, mean = 0, sd = 1)
```

```
## [1] 0.6827
```

Continuous probability distribution

Normal z-score (cont):



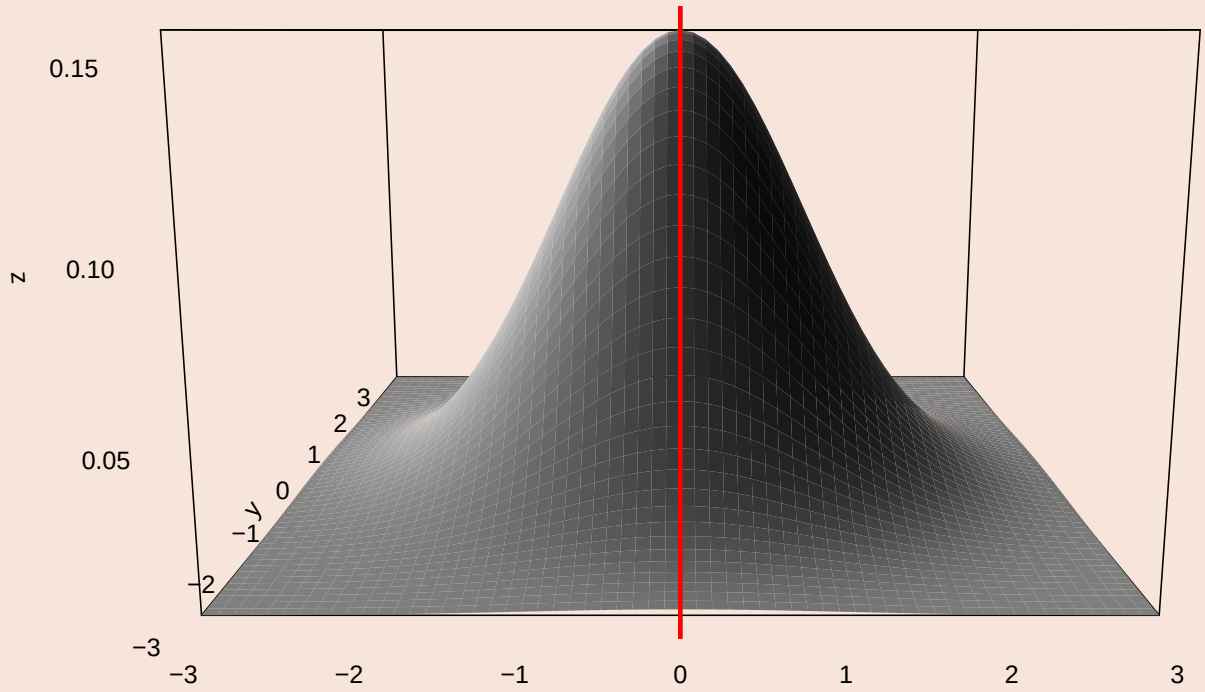
- What is the probability associated with the area between +2 and -2?

```
pnorm(2, mean = 0, sd = 1) - pnorm(-2, mean = 0, sd = 1)
```

```
## [1] 0.9545
```

Continuous probability distribution

Normal z-score (cont):



- What is the probability associated with the area between +3 and -3?

```
pnorm(3, mean = 0, sd = 1) - pnorm(-3, mean = 0, sd = 1)

## [1] 0.9973
```

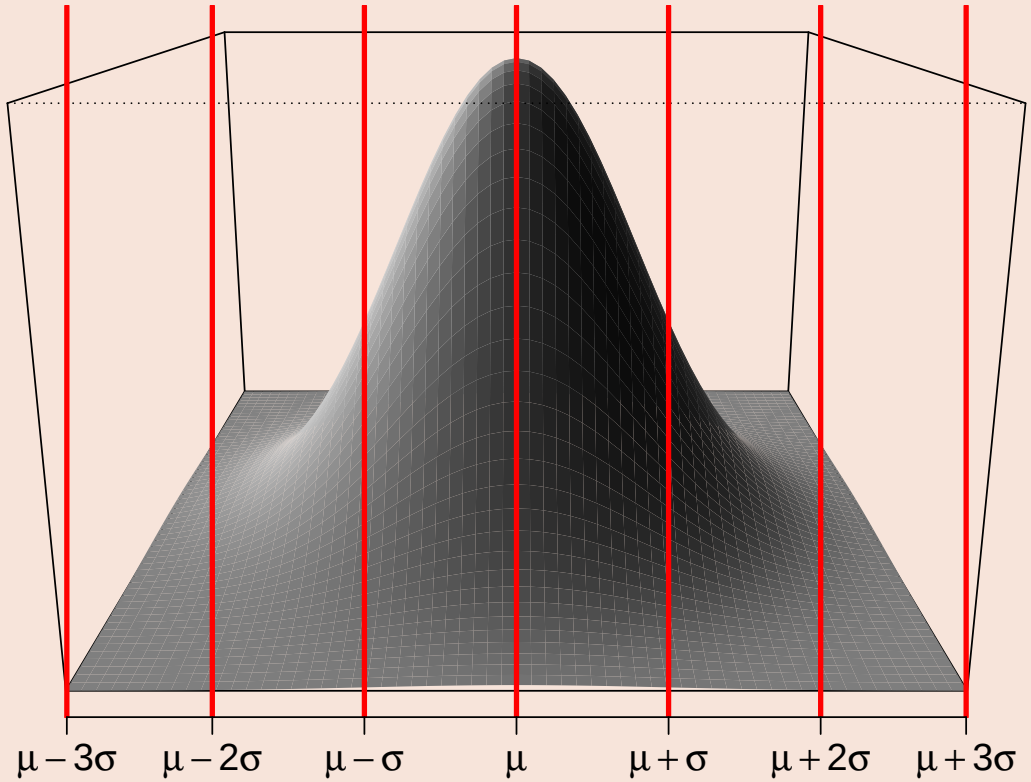
Continuous probability distribution

Normal z-score (cont):

- In a standard normal distribution $+1$ refers to the standard deviation above the mean whereas -1 referes to one standard deviation below the mean
- Z score between $+1$ and -1 covers 68% of the area
- Z score between $+2$ and -2 covers 95% of the area
- Z score between $+3$ and -3 covers 99.7% of the area

Continuous probability distribution

Normal z-score (cont):



$\mu \pm \sigma \approx 68\%$ $\mu \pm 2\sigma \approx 95\%$ $\mu \pm 3\sigma \approx 99.7\%$

Continuous probability distribution

Example:

- A company produces lightbulbs whose life follows a normal distribution, with a mean of 1,200 hours and a standard deviation of 250 hours. If we choose a lightbulb at random, what is the probability that its lifetime will be between 900 and 1300 hours?
- Solution: First calculate the corresponding z-score of the upper limit (1300 hours) and lowe limit (900 hours)

```
z_low = (900 - 1200)/250;          z_high= (1300 - 1200)/250
z_low;  z_high

## [1] -1.2
## [1] 0.4
```

Now we simply compute the probability:

```
pnorm(0.4, mean = 0, sd=1) - pnorm(-1.2, mean = 0, sd=1)

## [1] 0.5404
```

NOTE: we can use even a more direct method by providing relevant mean and standard deviation of a distribution to R:

```
pnorm(1300, mean = 1200, sd=250)  - pnorm(900, mean = 1200, sd=250)

## [1] 0.5404
```

Continuous probability distribution

Example:

We can also calculate the associated area, if probability for a distribution is given:

- Re-consider the bulb examples. We calculate the probability that the life time of the bulb is between 0 and 900

```
pnorm(900, mean = 1200, sd=250)

## [1] 0.1151
```

- Assume we are instead given a probability 0.1151, and we are asked to calculate lifetime of bulbs a this probability

```
qnorm(0.1151, mean = 1200, sd=250)

## [1] 900
```

- You can see using 0.11 probability, we can calculate the corresponding lifetime of bulbs. In other words, this would be the region to the left of 900 hours on data distribution

Mini-Exercise

Question:

Life span of people in a Aalborg is normally distributed with a mean 82 years and standard deviation 6. What is the probability that a randomly selected person from this class in Aalborg will live more than 90 years?

Outline

Joint Distribution

Joint Distribution:

X and Y are continuous random variables:

- The marginal cumulative distribution function for X is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- The marginal cumulative distribution function for Y is:

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(t)dt$$

- $F(x, y)$ is the joint cummulative distribution function for X and Y:

$$F(x, y) = P(X \leq x \cap Y \leq y)$$

- The random variables X and Y are independent if and only if:

$$F(x, y) = F(x).F(y)$$

Mean, Variance and Covariance

Mean:

$$E[X+Y] = E[X] + E[Y] = \mu_X + \mu_Y$$

Assume $W = aX + bY + c$:, and a, b and c are some constants, then the mean of W is:

$$E[W] = E[aX + bY + c] = aE[X] + bE[Y] + c = a\mu_X + b\mu_Y + c$$

Statistical independence and Mean of a random variable:

If X and Y are independent variables:

$$E(XY) = E(X)E(Y) = \mu_X \mu_Y$$

Mean, variance and covariance of joint probability dist.

Variance of variables with join probability dist (cont):

$$Var(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2Cov(X,Y)$$

Assume $W = aX + bY + c$:, and a, b and c are some constants, then the Variance of W is:

$$\sigma_W^2 = Var(aX + bY + c) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abCov(X,Y)$$

Variance and statistical independence:

X and Y are statistically independent variable:

Then $Var(X+Y) = Var(X) + Var(Y)$

Mean, variance and covariance of joint probability dist.

Covariance of variables with joint prob dist:

Assume $W = aX + bY$ and a, b are some constants, then the Covariance of X and Y is:

$$Cov(aX, bY) = ab.Cov(X, Y)$$

Also:

$$Cov(a + X, b + Y) = Cov(X, Y)$$

Note: that adding constants to X and Y does not affect the covariance between them

Covariance and statistical independence:

X and Y are statistically independent variable:

Then $Cov(X, Y) = 0$

Outline

Correlation

Correlation:

X and Y are random variables:

The correlation between X and Y is:

$$\delta = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

The correlation is the covariance divided by the standard deviations of the two random variables

The correlation coefficient, provides a measure of the strength of the linear relationship between two random variables, with the measure being limited to the range from -1 to $+1$

$+1$ means perfect positive correlation, -1 means perfect negative correlation

A correlation closer to 0 means weaker correlation, and 0 means no correlation

Correlation

Correlation in R:

We can very easily calculate mean, variannce, covariance, correlation in R:

```
x=rnorm(100, 0, 2)
y=rnorm(100, 1, 1)
var(x) # variance of x
```

```
## [1] 3.718
```

```
mean(y) # mean of y
```

```
## [1] 0.9346
```

```
sd(x) # standard deviation of x
```

```
## [1] 1.928
```

```
cov(x,y) # covariance between x and y
```

```
## [1] 0.0785
```

```
cor(x,y) # correlation between x and y
```

```
## [1] 0.04319
```

Outline

Exercise

Exercise:

- This is time for you to start dealing with real world data.
- Download quarterly GDP and unemployment rate data for Denmark from 1995–2016.
  - Data can be downloaded Statistics Denmark
- Import the data in R
- Calculate GDP growth. Plot the distribution of GDP growth and fit a probability density curve. Does it look like a normal distribution?
- Calculate the mean, variance and the standard deviation of both GDP growth and unemployment rate
- Calculate the covariance and correlation of unemployment rate and GDP growth. Is the correlation positive or negative?
- Show the relationship between GDP growth and unemployment in a scatter plot (with a line).