DSA 8010 - continuous probability distributions

PDFs

Probability distributions

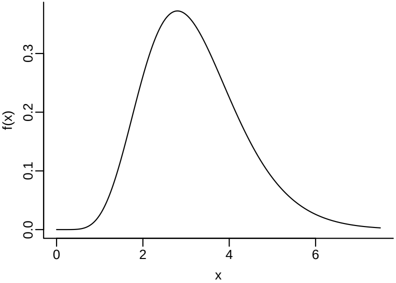
The *probability distribution* of a random variable is a function that determines the probability for any possible value.

Probability mass function (PMF). For discrete random variables. PMF is a formula, graph, or table that gives the probability that $X = x$ for every $x \in \mathcal{X}$.

Probability density function (PDF). For continuous random variables. A PDF is a formula for a curve. The probability of the random variable falling in a given interval is represented by an area under the curve.

Probability density function

- A probability density function for a R.V. X, denoted by $f(x)$, gives the height of a density curve for every $x \in \mathcal{X}$.
- $f(x)$ is not the $P(X = x)$.



- $P(a < X < b)$ is given by the area under the curve between a and b.

Probability density function

Probability density functions follow a few rules:

- Rule 1: the area under the curve is 1. ($\int_{x \in \mathcal{X}} f(x) = 1.$)
- Rule 2: $f(x) > 0$ for all $x \in \mathcal{X}$. (The density function is always nonnegative.)

Calculus recap

Integrals are used to find areas under a curve. Therefore, a probability for a continuous random variable is found by integrating the pdf:

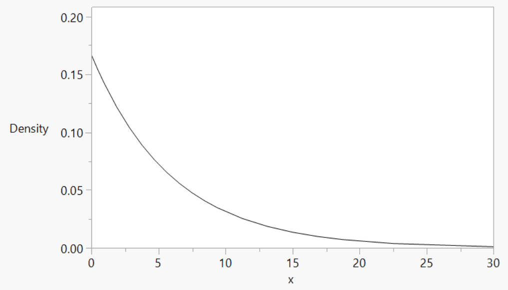
$$P(a < X < b) = \int_a^b f(x)dx$$

Probability density function

Example 1: Let X =wait time for bus. Assume that X has the following PDF:

$$f(x) = \frac{1}{6}e^{-x/6}, \quad 0 \leq x < \infty$$

PDF of X:

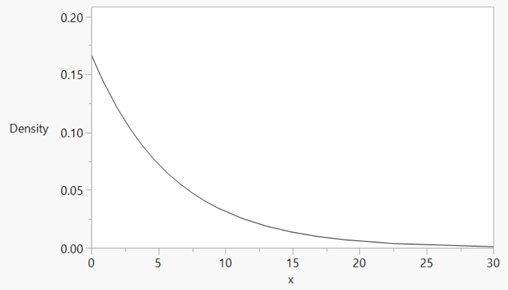


Probability density function

On the PDF below, shade the area corresponding to the following probabilities:

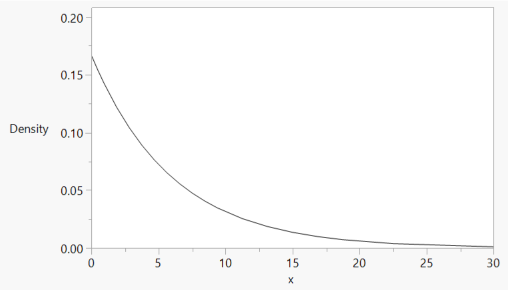
$P(10 \leq X \leq 15)$.

$P(X < 2.5)$.



Probability density function

Note: For any continuous RV, $P(X = x) = 0$.



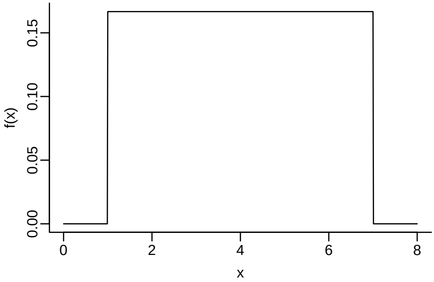
This means that $P(X \leq x) = P(X < x)$ for any x. Probabilities are the same whether strict inequalities or non-strict inequalities are used.

Uniform distribution

The uniform distribution gives equal density for any value inside a fixed interval.

If X has a Uniform(a, b) distribution, its pdf is

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



Uniform distribution

The uniform distribution gives equal density for any value inside a fixed interval.

If X has a Uniform(a, b) distribution, its pdf is

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

Set $a = 0$ and $b = 5$. Find and draw the pdf.

Uniform distribution

The uniform distribution gives equal density for any value inside a fixed interval.

If X has a Uniform(a, b) distribution, its pdf is

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

Set $a = 0$ and $b = 5$. Find $P(X < 3)$.

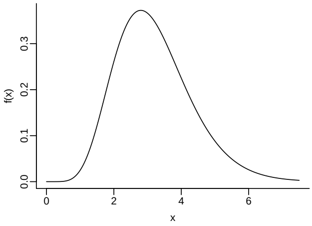
Expectation and variances

Expected value and variance

The mean or "expected value" of a continuous random variable X is defined as

$$E(X) = \int_{x \in \mathcal{X}} x f(x) dx.$$

- $E(X)$ can be thought of as the average value of the random variable. Sometimes μ will be used to denote $E(X)$.
- Think of $E(X)$ as being like the balancing point of the mass of the distribution.



Expected value and variance

The variance of a continuous random variable X is defined as

$$\text{Var}(X) = \int_{x \in \mathcal{X}} (x - E(X))^2 f(x) dx.$$

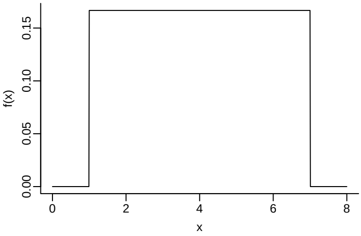
- Sometimes σ^2 will be used to denote Var(X).
- The square root of Var(X) is the standard deviation of X.

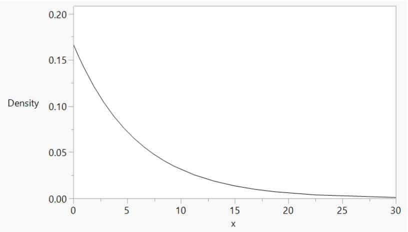
- Typically, with continuous random variables we use named families of distributions (e.g. Normal, chi-square. See later in the lecture) whose mean and variance are known functions of the parameters of the distribution.

Percentiles of random variables

The pth percentile of a probability distribution is a value x_0 that satisfies

$$P(X \leq x_0) = p.$$





Normal distribution

Normal distribution

Random variable. X continuous, $-\infty < X < \infty$.

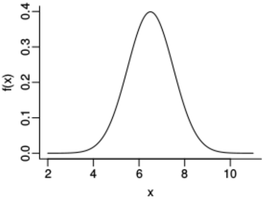
Parameters. μ, σ^2.

$E(X) = \mu$, $Var(X) = \sigma^2$.

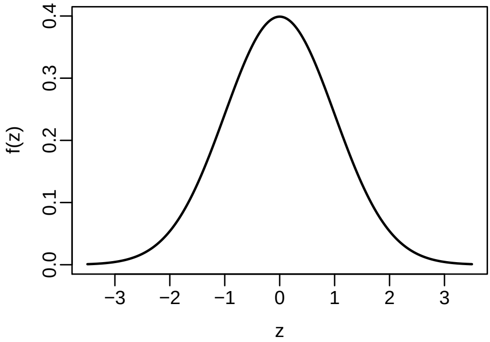
PDF. $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

Features. Symmetric about μ, bell-shaped.

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



Empirical rule for normal distributions.



If X has a $N(\mu, \sigma^2)$ distribution, then the following probabilities hold.

1. $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68$
2. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95$
3. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.7$

Empirical rule for normal distributions.

If X has a $N(\mu, \sigma^2)$ distribution, then the following probabilities hold.

1. $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68$

   There is a 68% chance that X is within one standard deviation of the mean.

2. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95$

   There is a 95% chance that X is within two standard deviations of the mean.

3. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.7$

   There is a 99.7% chance that X is within three standard deviations of the mean.

Example (empirical rule)

Assume that $X \sim N(4, 0.25^2)$. Then the probability that X is between __________ and __________ is about 0.95.

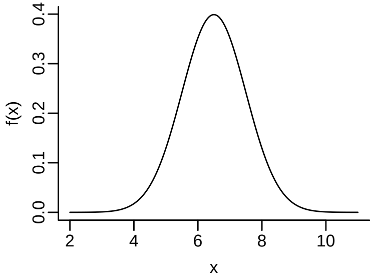
Assume that $X \sim N(50, 10^2)$. Then the probability that X is between 20 and 80 is about __________.

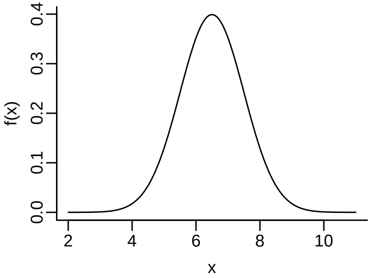
Normal probabilities

The plots below show the pdf of the $N(6.5, 1^2)$ distribution.
Sketch the quantities

1. $P(X > 9)$ and
2. $P(4 < X < 8)$

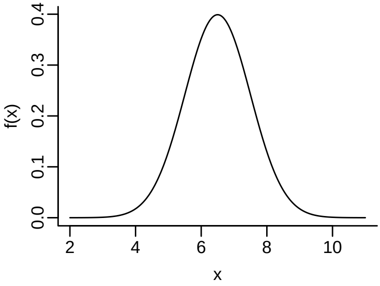
and guess their numeric values.

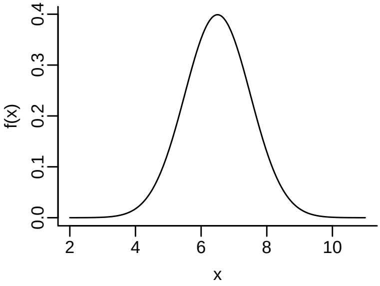




Normal percentiles

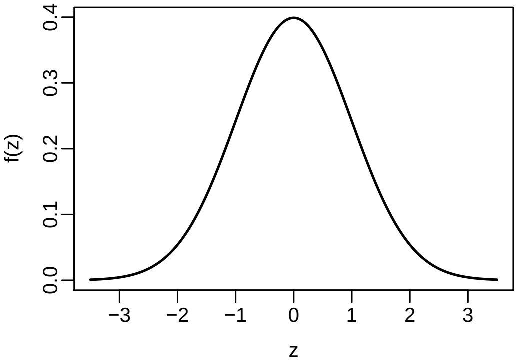
The plots below show the pdf of the $N(6.5, 1^2)$ distribution. Sketch the (approximate) locations of the 95th and 40th percentiles and guess their numeric values.





Standard normal distribution.

The standard normal distribution is a normal distribution with a mean of 0 and standard deviation of 1.

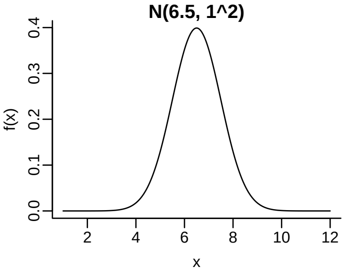


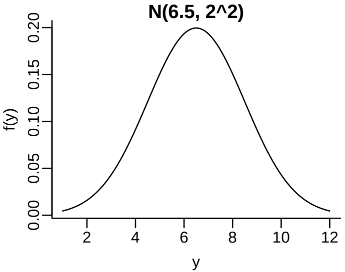
Standard normal distribution.

- Any normal RV can be converted to a standard Normal RV.

  If X has a N(μ, σ^2), then $Z = (X - \mu)/\sigma$ has a N(0, 1) distribution.

- Z indicates how many standard deviations X is above or below the mean. It is sometimes called a *z-score* or *standard score* and it measures how "unusual" X is, relative to its mean and standard deviation.

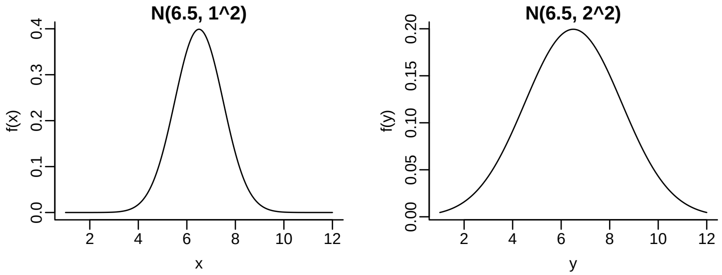




Z scores

Use a Z score to determine which is more unusual:

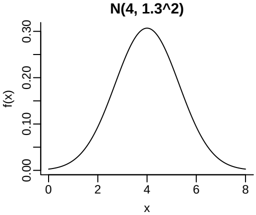
1. $P(X > 8)$ or
2. $P(Y > 9.5)$.



Finding normal probabilities

Suppose that Y has a $N(4, 1.3^2)$ distribution. What is the probability that Y is less than 3?

Picture:



Analytical expression

$$\int_{-\infty}^{3} \frac{1}{\sqrt{2\pi \cdot 1.3^2}} e^{\left(-\frac{1}{2\cdot 1.3^2}(x-4)^2\right)} dy$$

Finding normal probabilities in R

Assume that $X \sim N(a, b^2)$. The following R functions will perform calculations related to the normal distribution.

`pnorm(x,a,b)`	Returns $P(X \leq x)$.
`qnorm(p,a,b)`	Returns the $p \times 100$th percentile of the distribution; that is, the value x such that $P(X \leq x) = p$.
`rnorm(n,a,b)`	Generates `n` random variables from the normal distribution.
`dnorm(x,a,b)`	Returns the probability density at x.

Note that the b value in the functions is the standard deviation, not the variance of the Normal distribution.

Example: finding normal probabilities

Assume that X has a $N(4, 1.3^2)$ distribution. Find $P(X \leq 2)$ and $P(X > 2)$.

Example: finding normal probabilities

Assume that X has a $N(0, 2.7^2)$ distribution. Find $P(-1 < X < 3)$

Example: finding normal probabilities

Assume that X has a $N(4, 1.3^2)$ distribution. Find the 75th percentile.

Other continuous families

Parameters of distributions

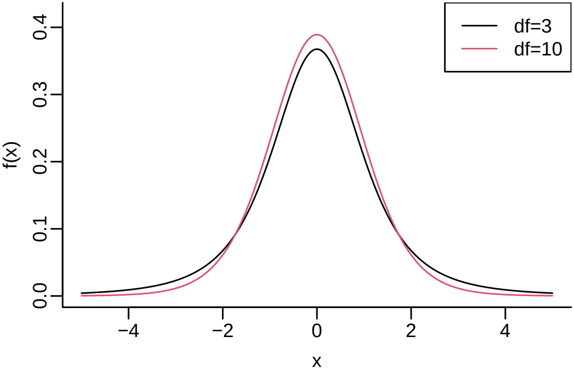
Most probability distributions, discrete or continuous, have parameters, or fixed numbers that determine the features of the distribution. For example, the Bernoulli distribution has parameter π which determines how likely successes are to occur. The Normal distribution has μ, which determines the location of the bell curve, and σ, which determines how spread out it is.

In probability calculations, we assume we know the parameters. In data applications, the parameters are usually unknown.

Student's t

Random variable. Continuous T; $\infty < T < \infty$. Symmetric shape.

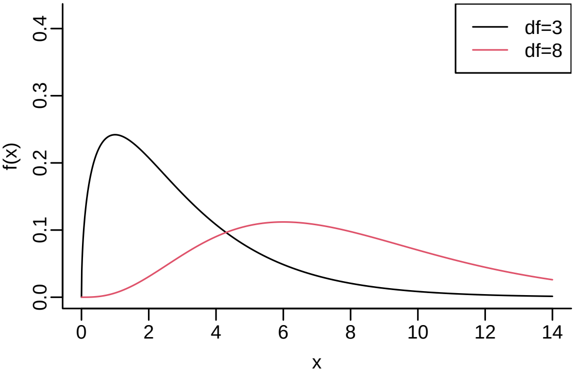
Parameters. Degrees of freedom, or df, which controls how "fat" the tails are.



Chi-square

Random variable. Continuous X; $0 < X < \infty$. Right-skewed shape.

Parameters. Degrees of freedom, or df, which controls how fair the tails extends to the right.



Gamma

Random variable. Continuous X; $0 < X < \infty$. Can be skewed in either direction.

Parameters. a and b, which control the mean, variability, and direction of the skew.