The Normal Distribution

David Armstrong

UCI

Continuous Random Variables

- A continuous random variable can be an uncountable infinite possible values.
- Examples:

  X can be any number in the interval [0,1]. Thus $\mathbb{S}_X = [0, 1]$.

  Time it takes to drive to Las Vegas from UCI. Thus $\mathbb{S}_X = [3, \infty)$.

  Percent score on an Exam. Thus $\mathbb{S}_X = [0\%, 100\%]$.
- $f(x)$ is now called the *probability density function*. pdf
- $f(x)$ does not represent $P(X = x)$ anymore.
- $P(X = x) = 0$ for all x in $\mathbb{S}_X$.
- The probability that a continuous random variable is equal to a single fixed number is 0.
- A continuous random variable takes on an uncountably infinite number of possible values.
- For a **discrete** random variable X that takes on a finite or countably infinite number of possible values, we determined that P(X = x) for all of the possible values of X, and called it the probability mass function (pmf).
- With a continuous random variable, we can only calculate probabilities of intervals such as $P(a < X < b)$.
- The pdf $f(x)$ is an equation/curve used to calculate the probability of intervals and moments.
- The pdf can be quantifying something that is proportional to the probability.

Continuous Random Variables

Let X be a continuous random variable with support $\mathbb{S}_X$ and probability density function $f(x)$.

- For $f(x)$ to be a valid pdf, the following must hold.
  - $f(x) \geq 0$ for all x in $\mathbb{S}_X$.
  - $\int\limits_{\mathbb{S}_X} f(x)dx = 1$.
- $E(X) = \int\limits_{\mathbb{S}_X} x\, f(x)dx$.
- $VAR(X) = \int\limits_{\mathbb{S}_X} (x - E(X))^2 f(x)dx$.
  - Note: We can still use the previous equation for the variance: $VAR(X) = E(X^2) - [E(X)]^2$.

Continuous Random Variables

Let X be a continuous random variable with support $\mathbb{S}_X$ and probability density function $f(x)$.

A few things to note.

pdf $\rightarrow \int$ pdf

- $P(X = x) = 0$.
- $P(X \leq x) = P(X < x)$.
  - $P(X \leq x) = P(X < x) + P(X = x) = P(X < x)$.
  - Example $P(X < 50) = P(X \leq 50)$, since P(X=50) is 0.
- Also, probability of intervals can be written using cdf's.

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

The cumulative distribution function $F(x)$ is written as:

- $P(X < x) = P(X \leq x) = \int_{l}^{x} f(u)du$.
  - Where l is the lower bound of the support of X, $\mathbb{S}_X$ (commonly it is $-\infty$).
- Note that $P(X < x) = F(X) = F(X) - F(l)$ where $F(l) = 0$.
- As a result, $\frac{d}{dx}F(x) = f(x)$.
- The derivative of the cumulative distribution function (cdf) is the probability distribution function (pdf).

$F(x) =$ Area under curve

$f(x) =$ curve

$f'(x) =$ instant rate of change

set $= 0$

min

max

< 0 down ∩ $\leftarrow$

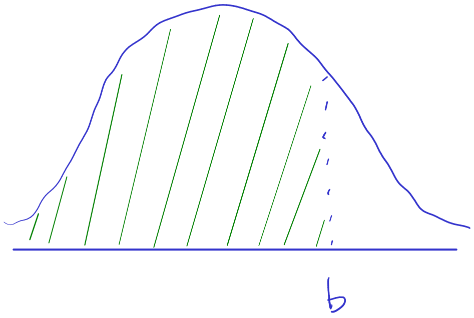
$f''(x) =$ concavity

> 0 up ∪ $\leftarrow$

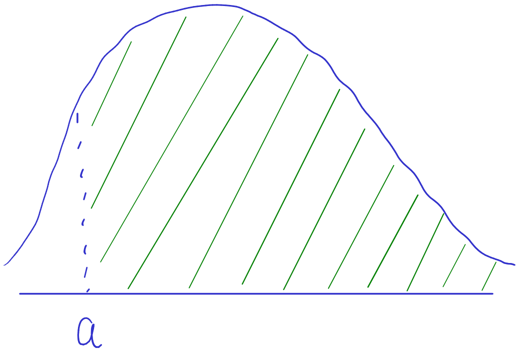
Continuous Random Variables

Assume a distribution follows a bell shaped curve. Sketch each of the following situations.

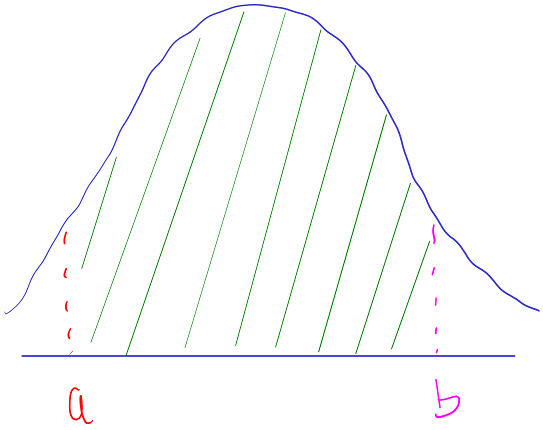
- $P(\underline{X < b}) = \int\limits_{x<b} f(x)dx.$



- $P(X > a) = \int\limits_{a<x} f(x)dx.$



- $P(a < X < b) = \int_a^b f(x)dx.$ $= F(x)\Big|_a^b = F(b) - F(a)$

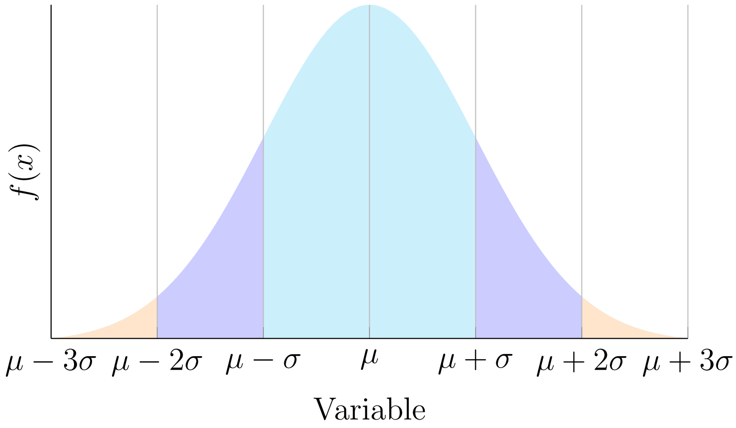


Normal Distribution

- Most common.
- Symmetric, unimodal, bell curve
- Gaussian distribution was named after Frederic Gauss, the first person to formalize its mathematical expression.
- $\mathbb{S}_X = (-\infty, \infty)$
- Say X follows a Normal distribution with parameters μ and σ.

  The location parameter is μ in $(-\infty, \infty)$

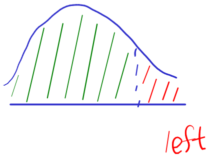
  The scale parameter is σ in $[0, \infty)$ sigma
- We write: $X \sim Normal(\mu, \sigma)$. standard deviation
- Examples:
  - SAT scores
  - Heights of US adult males
  - The amount of time teenagers spend on the internet
  - Weights of babies



The Normal Distribution

- Denoted $X \sim Normal(\mu, \sigma)$
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$.
  - We use this distribution for continuous variables that follow a bell shape curve.
- $\mathrm{F}(X) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$
- $\mathrm{E}(X) = \mu$
- $\mathrm{VAR}(X) = \sigma^2$
- Standardized score $Z = \dfrac{\overset{\text{observed value}}{x} - \overset{E(x)}{\mu}}{\underset{\text{std dev}}{\sigma}}$

R Code



To get the area to the left of a Normal(0,1) variable:
$pnorm(x, \mu = 0, \sigma = 1)$

To get the area to the right of a Normal(0,1) variable:
$1 - pnorm(x, \mu = 0, \sigma = 1)$

To get the area between two values c and d (c < d):
$pnorm(d, \mu = 0, \sigma = 1) - pnorm(c, \mu = 0, \sigma = 1)$

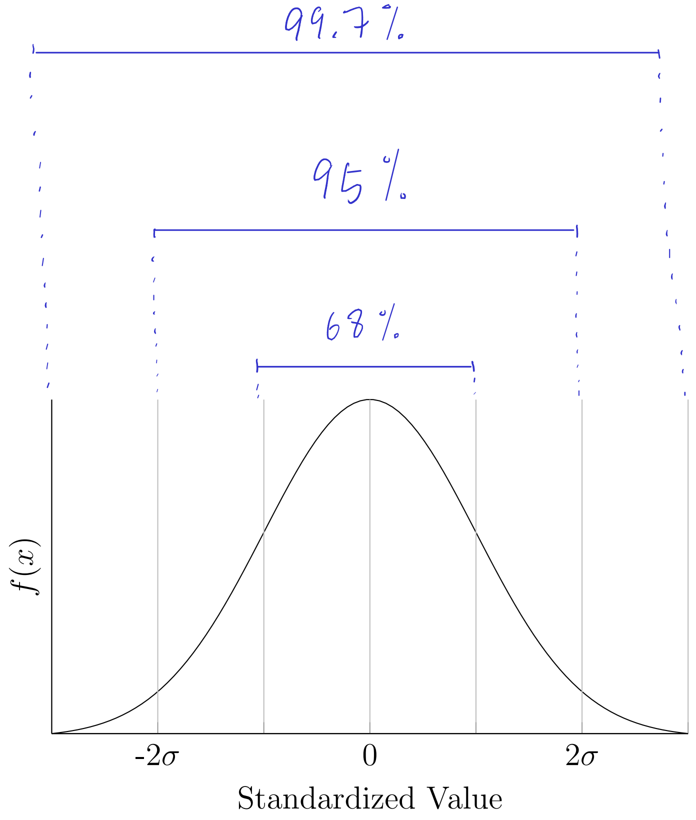
To get the value of x related to the lower tail (α):
$qnorm(\alpha, \mu = 0, \sigma = 1)$

To get the value of x related to the upper tail:
$qnorm(1 - \alpha, \mu = 0, \sigma = 1)$

Empirical Rule

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without R or Z-table. 68% will fall within one standard deviation of the mean, 95% within two standard deviations of the mean, and 99.7% within three standard deviations of the mean.

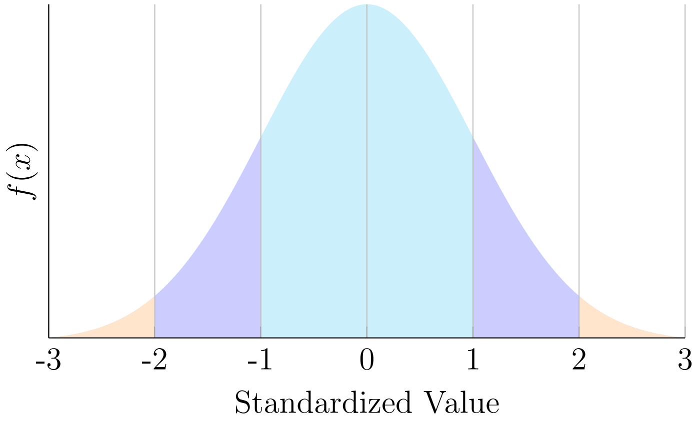
68% — 95% — 99.7%



Standard Normal Distribution

$$Z = \frac{x - \mu}{\theta}$$

- Denoted $Z \sim Normal(\mu = 0, \sigma = 1)$
- $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$.
  - We use this distribution to standardize the values of continuous variables that follow a bell shape curve.
- $\mathrm{F}(Z) = P(Z \leq z) = \int_{-\infty}^{z} f(z)dz$
- $\mathrm{E}(Z)$=0
- $\mathrm{VAR}(Z) = 1$



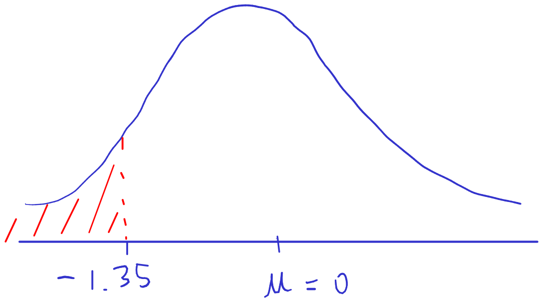
Example: What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region?

$Z \sim N(\mu = 0, \theta = 1)$

Be sure to draw a graph.

$P(Z < -1.35)$ $= 0.0885$

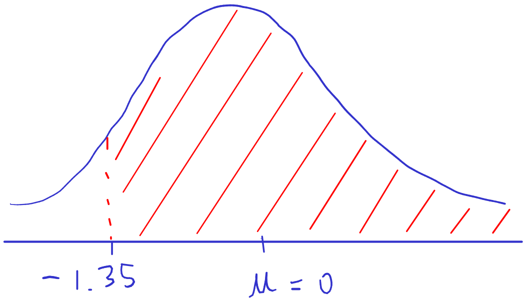
pnorm(-1.35, 0, 1)



$P(Z > -1.35)$ $= 1 - P(Z \leq -1.35) = 1 - 0.0885$

$= 0.9915$

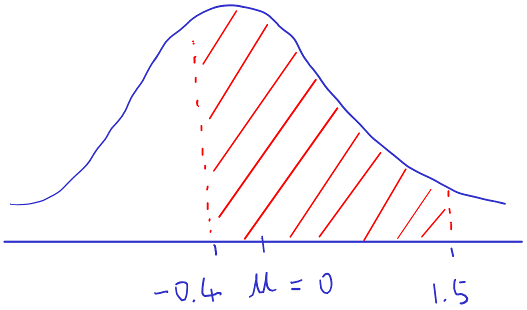
1 - pnorm(-1.35, 0, 1)



$P(-0.4 < Z < 1.5)$ $= P(Z < 1.5) - P(Z < -0.4)$

$= 0.9332 - 0.3446 = 0.5886$

pnorm(1.5, 0, 1) - pnorm(-0.4, 0, 1)



Example: The distribution of SAT and ACT scores are both nearly normal.

	SAT	ACT
Mean	1500	21
SD	300	5

observed score → x ; expected score → μ ; std dev → θ

$$z = \frac{x - \mu}{\theta}$$

- Suppose Ann scored 1700 on her SAT and Tom scored 24 on his ACT. Who performed better?

$X \sim N(\mu = 1500, \theta = 300)$

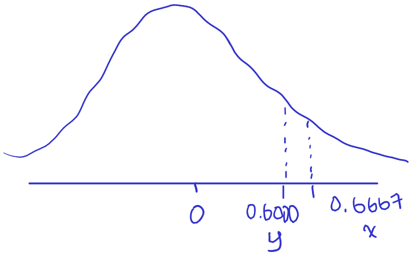
Ann $x = 1700$

$$z = \frac{x - \mu}{\theta} = \frac{1700 - 1500}{300} = 0.6667$$

$Y \sim N(\mu = 21, \theta = 5)$

Tom $y = 24$

$$z = \frac{y - \mu}{\theta} = \frac{24 - 21}{5} = 0.6000$$



we are looking for higher score, thus Ann performed better.

- What is the probability someone scores above a 1550 on their SAT?

$X \sim N(1500, 300)$

$$P(X > 1550) = 1 - P(X \leq 1550)$$
$$= 1 - \text{pnorm}(1550, 1500, 300) = 0.4338$$

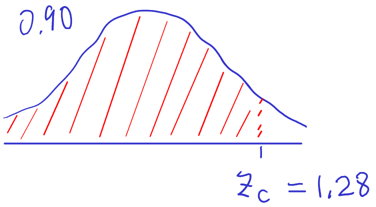
$$P(X > 1550) = 1 - P(X \leq 1550)$$
$$= 1 - P(X - \mu \leq 1550 - \mu)$$
$$= 1 - P\left(\frac{X - \mu}{\theta} \leq \frac{1550 - \mu}{\theta}\right)$$
$$= 1 - P\left(Z \leq \frac{1550 - 1500}{300}\right)$$
$$= 1 - P(Z \leq 0.17) = 1 - 0.5675 = 0.4325$$

$1 - \text{pnorm}(0.17, 0, 1)$

$$z = \frac{x-\mu}{\theta} \qquad x_c = z_c(\theta) + \mu$$

Example: The length of time required to complete a college test is found to be normally distributed with mean 50 minutes and standard deviation 12 minutes.

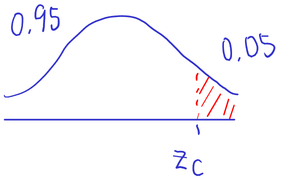
a. When should the test be terminated if we wish to allow sufficient time for 90% of the students to complete the test?



$$z_c = \text{qnorm}(0.90, 0, 1) = 1.28$$

$$\begin{aligned} x_c &= z_c(\theta) + \mu \\ &= 1.28(12) + 50 \\ &= 65.36 \end{aligned}$$

$$\text{qnorm}(0.90, 50, 12) = 63.3786$$



$$1 - \text{qnorm}(0.95, 0, 1)$$

b. What proportion of students will finish the test between 30 and 60 minutes?

$$\begin{aligned} P(30 < x < 60) &= P(x \leq 60) - P(x < 30) \\ &= \text{pnorm}(60, 50, 12) - \text{pnorm}(30, 50, 12) \\ &= 0.7499 \end{aligned}$$

c. What proportion of students will finish faster than 45 minutes?

$$\begin{aligned} P(x < 45) &= \text{pnorm}(45, 50, 12) \\ &= 0.3385 \end{aligned}$$