Statistics and Data Analysis

Probability

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Road map

- **Random variables**.
- Expectation and variances.
- Continuous distributions.
- Normal distribution.

Random variables

- To describe a random event, we use random variables.
- A **random variable** (RV) is a variable whose outcomes are random.
- Examples:
  - The outcome of tossing a coin or rolling a dice.
  - The number of consumers entering a store at 7-8pm.
  - The temperature of a classroom at tomorrow noon.

Discrete and continuous random variables

- A random variable can be **discrete** or **continuous**.
- For a discrete random variable, its value is **counted**.
  - The outcome of tossing a coin.
  - The outcome of rolling a dice.
  - The number of consumers entering a store at 7-8pm.
- For a continuous random variable, its value is **measured**.
  - The temperature of this classroom at tomorrow noon.
  - The average studying hours of a group of 100 students.
- A discrete random variable has **gaps** among its possible values.
- A continuous random variable's possible values typically form an **interval**.

Discrete and continuous distributions

- How to describe a random variable?
  - Write down its **sample space**, which includes all the possible values.
  - For each possible value, write down the **likelihood** for it to occur.
- The two things together form a **probability distributions**, or simply distributions.
- Distributions may also be either discrete or continuous.
  - Let's start with discrete distributions.

Describing a discrete distribution

- For a discrete random variable, we may **list** all possible outcomes and their probabilities.
  - Let X be the result of tossing a fair coin:

x	Head	Tail
$\Pr(X = x)$	$\frac{1}{2}$	$\frac{1}{2}$

  - Let X be the result of rolling a fair dice:

x	1	2	3	4	5	6
$\Pr(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- The function $\Pr(X = x)$, sometimes abbreviated as $\Pr(x)$, for all $x \in S$, where S is the sample space, is called the **probability function** of X.
  - We have $\Pr(X = x) \in [0, 1]$ for all $x \in S$.
  - We have $\sum_{x \in S} \Pr(X = x) = 1$.

Example 1: coin tossing

- Let X_1 and X_2 be the result of tossing a fair coin for the first and second time, respectively.
- Let Y be the **number of heads** obtained by tossing a fair coin twice.
- What is the distribution of Y?
  - Possible values: 0, 1, and 2.
  - Probabilities: What are $\Pr(Y = 0)$, $\Pr(Y = 1)$, and $\Pr(Y = 2)$?
- We have:

y	0	1	2
$\Pr(Y = y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example 1: coin tossing

- What if the probability of getting a head is p?
- We have

$$\begin{aligned}\Pr(Y=2) &= \Pr((X_1,X_2)=(\text{Head},\text{Head})) = p^2,\\ \Pr(Y=0) &= \Pr((X_1,X_2)=(\text{Tail},\text{Tail})) = (1-p)^2, \text{ and}\\ \Pr(Y=1) &= \Pr((X_1,X_2)=(\text{H},\text{T})) + \Pr((X_1,X_2)=(\text{T},\text{H}))\\ &= p(1-p)+(1-p)p = 2p(1-p).\end{aligned}$$

- In summary:

y	0	1	2
$\Pr(Y=y)$	$(1-p)^2$	$2p(1-p)$	p^2

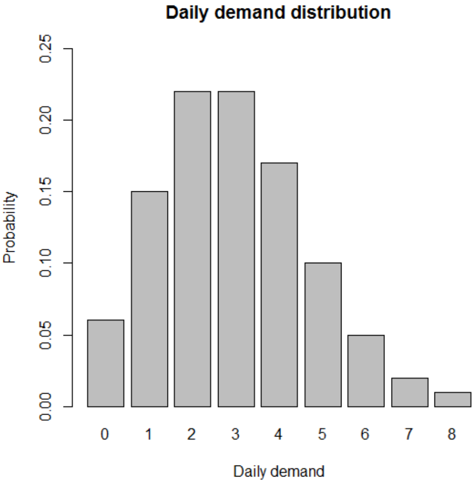
Example 2: inventory management

- Suppose that you sells apples.
  - The unit purchasing cost is \$2.
  - The unit selling price is \$10.
- Question: How many apples to prepare at the beginning of each day?
  - Too many is not good: **Leftovers** are valueless.
  - Too few is not good: There are **lost sales**.
- According to your historical sales records, you predict that tomorrow's demand is X, whose distribution is summarized below:

x	0	1	2	3	4	5	6	7	8
$\Pr(x)$	0.06	0.15	0.22	0.22	0.17	0.10	0.05	0.02	0.01

Daily demand distribution

- The probability distribution is depicted.
- This is a **right-tailed** (skewed to the right; positively skewed) distribution.
- The distribution of Y in Example 1 is **symmetric**.



Distributions of some events

x	0	1	2	3	4	5	6	7	8
$\Pr(x)$	0.06	0.15	0.22	0.22	0.17	0.10	0.05	0.02	0.01

- What is the minimum inventory level that can make the **probability of having shortage** lower than 20%?
  - This is the inventory level achieving a 80% **service level**.
  - If the inventory level is x, the service level is $\Pr(X \leq x)$.
  - As $F(x) = \Pr(X \leq x)$ is used often, it is given the name **cumulative distribution function** (cdf).
- The service level may be calculated for all x:
  - $F(1) = \Pr(X \leq 1) = \Pr(X = 0) + \Pr(X = 1) = 0.21$.
  - $F(3) = \Pr(X \leq 3) = \Pr(X = 0) + \cdots \Pr(X = 3) = 0.65$.
  - $F(4) = \Pr(X \leq 4) = \Pr(X = 0) + \cdots \Pr(X = 4) = 0.82$.

Road map

- Random variables.
- **Expectation and variances**.
- Continuous distributions.
- Normal distribution.

Expectation

- Consider a discrete random variable X with a sample space $S = \{x_1, x_2, ..., x_n\}$ and a probability function $\Pr(\cdot)$.
- The **expected value** (or mean) of X is

$$\mu = \mathbb{E}[X] = \sum_{i \in S} x_i \Pr(x_i).$$

- Intuition: For all the possible values, use their probabilities to do a weighted average.
- For the random outcome, if I may guess only one number, I would guess the expected value to minimize the average error.

Example 1: dice rolling

- Let X be the outcome of rolling a dice, then the probability function is $\Pr(x) = \frac{1}{6}$ for all $x = 1, 2, ..., 6$. The expected value of X is

$$\mathbb{E}[X] = \sum_{i=1}^{6} x_i \Pr(x_i) = \frac{1}{6}(1 + 2 + \cdots + 6) = 3.5.$$

- Let Y be the outcome of rolling an unfair dice:

y_i	1	2	3	4	5	6
$\Pr(y_i)$	0.2	0.2	0.2	0.15	0.15	0.1

  - The expected value of Y is

$$\begin{aligned}\mathbb{E}[Y] &= 1 \times 0.2 + 2 \times 0.2 + 3 \times 0.2 + 4 \times 0.15 + 5 \times 0.15 + 6 \times 0.1 \\ &= 3.15.\end{aligned}$$

  - Note that $3.15 < 3.5$, the expected value of rolling a fair dice. Why?

Conditional probability and expectation

- I sell orange juice everyday. Let D be the daily demand.
  - If it is sunny, I have $\Pr(D = 50|\text{sunny}) = \Pr(D = 250|\text{sunny}) = 0.5$.
  - If it is rainy, I have $\Pr(D = 10|\text{rainy}) = \Pr(D = 50|\text{rainy}) = 0.5$.
  - These are **conditional probabilities**.
- What is my expected daily demand given the weather condition?
  - We have $\mathbb{E}[D|\text{sunny}] = 150$ and $\mathbb{E}[D|\text{rainy}] = 30$.
  - These are **conditional expectations**.
- If with probability 70% it will be sunny tomorrow, what is my tomorrow expected demand?

$$\begin{aligned}\mathbb{E}[D] &= \Pr(\text{sunny})\mathbb{E}[D|\text{sunny}] + \Pr(\text{rainy})\mathbb{E}[D|\text{rainy}] \\ &= 0.7 \times 150 + 0.3 \times 30 = 114.\end{aligned}$$

- The two events are **dependent**, i.e., the realization of one event affects the distribution of the other. They are not **independent**.

Example 2: Inventory decisions

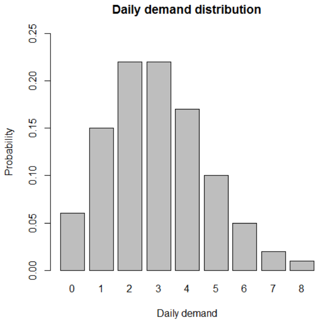
- Recall the inventory problem:
  - The unit purchasing cost is \$2.
  - The unit selling price is \$10.
  - The daily random demand's distribution is

x	0	1	2	3	4	5	6	7	8
$\Pr(x)$	0.06	0.15	0.22	0.22	0.17	0.10	0.05	0.02	0.01

  - How to find a **profit-maximizing** inventory level?
- For our example, at least we may try all the possible actions.
  - Suppose the stocking level is y, $y = 0, 1, ..., 8$, what is the **expected** profit $\pi(y)$?
  - Then we choose the stocking level with the highest expected profit.

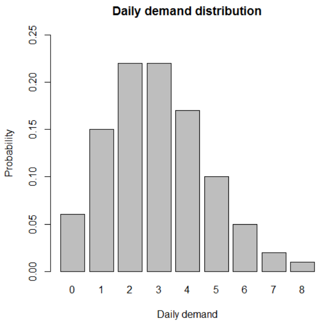
Expected profit function

- If $y = 0$, obviously $\pi(0) = 0$.
- If $y = 1$:
  - With probability 0.06, $X = 0$ and we lose $0 - 2 = -2$ dollars.
  - With probability 0.94, $X \geq 1$ and we earn $10 - 2 = 8$ dollars.
  - The expected profit is $(-2) \times 0.06 + 8 \times 0.94 = 7.4$ dollars, i.e., $\pi(1) = 7.4$.



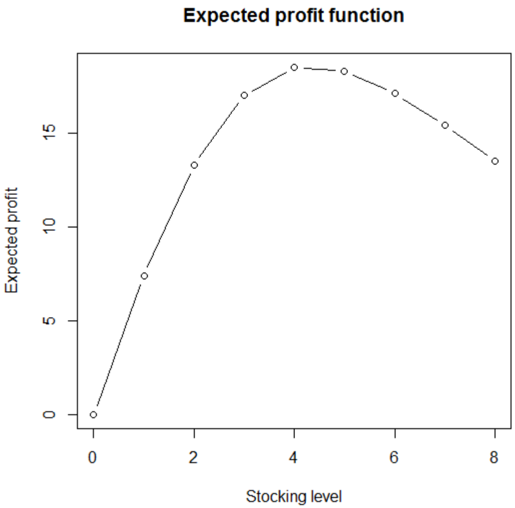
Expected profit function

- If $y = 2$:
  - With probability 0.06, $X = 0$ and we lose $0 - 4 = -4$ dollars.
  - With probability 0.15, $X = 1$ and we earn $10 - 4 = 6$ dollars.
  - With probability 0.79, $X \geq 2$ and we earn $20 - 4 = 16$ dollars.
  - The expected profit is $(-4) \times 0.06 + 6 \times 0.15 + 16 \times 0.79 = 13.3$ dollars, i.e., $\pi(2) = 13.3$.
- By repeating this on $y = 3, 4, ..., 8$, we may fully derive the expected profit function $\pi(y)$.



Optimizing the inventory decision

- The optimal stocking level is 4.
- What if the unit production cost is not \$2?



Variances and standard deviations

- Consider a discrete random variable X with a sample space $S = \{x_1, x_2, ..., x_n\}$ and a probability function $\Pr(\cdot)$.
- The expected value of X is $\mu = \mathbb{E}[X] = \sum_{i \in S} x_i \Pr(x_i)$.
- The **variance** of X is

$$\sigma^2 = \text{Var}(X) \equiv \mathbb{E}\big[(X - \mu)^2\big] = \sum_{i \in S} (x_i - \mu)^2 \Pr(x_i).$$

- The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

Example 1: dice rolling

- Let X be the outcome of rolling a dice, then the probability function is $\Pr(x) = \frac{1}{6}$ for all $x = 1, 2, ..., 6$.
  - The expected value of X is $\mu = \mathbb{E}[X] = 3.5$.
  - The variance of X is

$$\begin{aligned}\text{Var}(X) &= \sum_{i \in S} (x_i - \mu)^2 \Pr(x_i) \\ &= \frac{1}{6}\Big[(-2.5)^2 + (-1.5)^2 + \cdots + 2.5^2\Big] \approx 2.92.\end{aligned}$$

  - The standard deviation of X is $\sqrt{2.92} \approx 1.71$.

Example 1: dice rolling

- Let X be the outcome of rolling an unfair dice:

x_i	1	2	3	4	5	6
$\Pr(x_i)$	0.2	0.2	0.2	0.15	0.15	0.1

  - The expected value of X is $\mu = 3.15$.
  - The variance of X is

$$\begin{aligned}\text{Var}(X) &= \sum_{i \in S} (x_i - \mu)^2 \Pr(x_i) \\ &= (-2.15)^2 \times 0.2 + (-1.15)^2 \times 0.2 + (-0.15)^2 \times 0.2 \\ &\quad + 0.85^2 \times 0.15 + 1.85^2 \times 0.15 + 2.85^2 \times 0.1 \\ &\approx 2.6275.\end{aligned}$$

  - Note that $2.6275 < 2.92$, the variance of rolling a fair dice. Why?
  - The standard deviation of X is $\sqrt{2.6275} \approx 1.62$.

Example 2: investment decisions

- Let Green, Red, and White be three hypothetical **investments** with the following probability distributions for their yearly **gross returns**.

Probability	1/6	1/6	1/6	1/6	1/6	1/6
Green	0.8	0.9	1.1	1.1	1.2	1.4
Red	0.06	0.2	1	3	3	3
White	0.95	1	1	1	1	1.1

- Which one do you prefer?

Example 2: investment decisions

- For each investment, we may find its **mean** (expected value) and **standard deviation**.

Probability	1/6	1/6	1/6	1/6	1/6	1/6	Mean	SD
Green	0.8	0.9	1.1	1.1	1.2	1.4	1.083	0.195
Red	0.06	0.2	1	3	3	3	1.710	1.323
White	0.95	1	1	1	1	1.1	1.008	0.045

The mean measures the **expected return**. The standard deviation measures the **risk**.

  - We prefer high expected return and low risk.
- We may compare their volatility-adjusted returns $\mu - \frac{\sigma^2}{2}$:

$$\text{Green} > \text{White} > \text{Red } (1.064 > 1.007 > 0.835).$$

Road map

- Random variables.
- Expectation and variances.
- **Continuous distributions**.
- Normal distribution.

Continuous random variables

- Some random variables are **continuous**.
  - The value of a continuous random variable is **measured**, not counted.
  - E.g., the temperature of our classroom when the next lecture starts.
- For a continuous random variable, its possible values (sample space) typically form an **interval**.
  - Let X be the temperature (in Celsius) of our classroom when the next lecture starts. Then $X \in [0, 50]$.
- As another example, consider the number of courses taken by a student in this semester.
  - Let X_i be the number of courses taken by student i, $i = 1, 2, ..., n$.
  - Obviously, X_i is discrete.
  - However, their mean $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$ is (approximately) continuous!
  - Especially when n is large.
- We will often use a continuous random variable to approximate a discrete one.

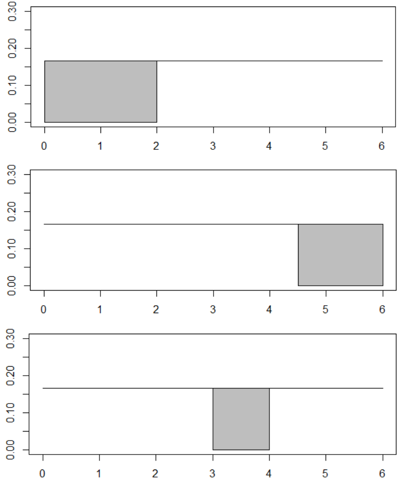
Continuous probability distribution

- Let X be a number randomly drawn from $[0, 6]$.
  - All values in $[0, 6]$ are equally likely to be observed.
- What is the probability of getting $X = 2$?
  - Because all the values (0, 1, 2.4, 3.657432, 4.44..., π, $\sqrt{2}$, etc.) may be an outcome, the probability of getting **exactly** $X = 2$ is **zero**.
  - In general, $\Pr(X = a) = 0$ for all $a \in \mathbb{R}$ as long as X is continuous.
- What is the probability of getting **no greater than** 2, $\Pr(X \leq 2)$?[1]

[1] Because $\Pr(X = 2) = 0$, we have $\Pr(X \leq 2) = \Pr(X < 2)$. In other words, "less than" and "no greater than" are the same regarding probabilities.

Continuous probability distribution

- Obviously, $\Pr(X \leq 2) = \frac{1}{3}$.
- Similarly, we have:
  - $\Pr(X \leq 3) = \frac{1}{2}$.
  - $\Pr(X \geq 4.5) = \frac{1}{4}$.
  - $\Pr(3 \leq X \leq 4) = \frac{1}{6}$.
- For a continuous random variable:
  - A **single value** has no probability.
  - An **interval** has a probability!



Uniform distribution

- The random variable X is very special:
  - All possible values are equally likely to occur.
- For a continuous random variable of this property, we say it follows a (continuous) **uniform distribution**.
  - When X is uniformly distributed in $[a, b]$, we write $X \sim \text{Uni}(a, b)$.
  - The likelihood of any possible value is $\frac{1}{b-a}$ (why)?
  - If a discrete random variable possesses this property (e.g., rolling a fair dice), we say it follows a discrete uniform distribution.
- When do we use a uniform random variable?
  - When we want to draw one from a population fairly (i.e., randomly).
  - When we collect a random sample from a population.

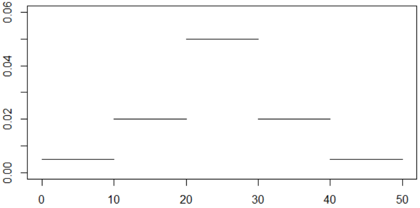
Non-uniform distribution

- Sometimes a continuous random variable is not uniform.
  - Let X be the temperature of the classroom when the next lecture starts.
  - We can say that $X \in [0, 50]$.
  - X is more likely to occur in $[20, 30]$ but less likely in $[10, 20]$ and $[30, 40]$. It is almost impossible for X to be in $[0, 10]$ and $[40, 50]$.
  - The likelihood of X in different intervals can be different.
- How to describe a continuous random variable with a non-uniform distribution? How to describe a continuous distribution?

Probability density functions

- We use a **probability density function** (pdf) $f(x)$ to describe the likelihood of each possible value. Larger $f(x)$ means **higher** likelihood.
- For X, let its pdf be

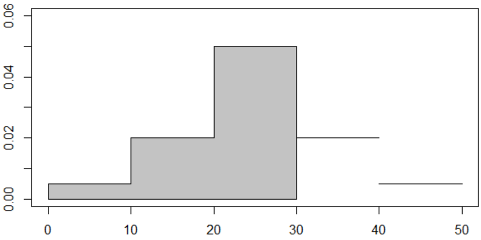
$$f(x) = \begin{cases} 0.005 & \text{if } x < 10 \\ 0.02 & \text{if } 10 \leq x < 20 \\ 0.05 & \text{if } 20 \leq x < 30 \\ 0.02 & \text{if } 30 \leq x < 40 \\ 0.005 & \text{if } 40 \leq x \end{cases}.$$



- The higher the pdf, the more likely the outcome is there.

Cumulative distribution functions

- The concept of **cumulative distribution function** (cdf) still applies to continuous distributions.
- Given the pdf $f(x)$, its cdf is $F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(v)dv$, which is the **area below the pdf** from $-\infty$ to x.
  - The "sum" of the likelihood of all values between 0 to x is the probability.
- $\Pr(X \leq 30) = \int_{0}^{30} f(v)dv = 10 \times 0.005 + 10 \times 0.02 + 10 \times 0.05 = 0.75.$

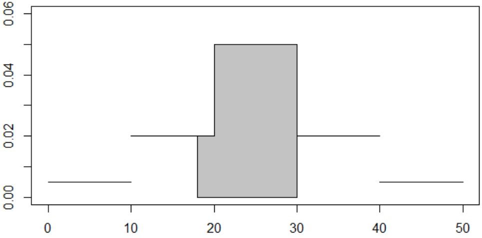


Cumulative distribution functions

- For any given region $[a, b]$, we then have

$$\Pr(a \leq X \leq b) = \Pr(X \leq b) - \Pr(X \leq a) = F(b) - F(a).$$

- E.g., $\Pr(18 \leq X \leq 30) = F(30) - F(18) = 0.75 - 0.21 = 0.54$.



- In most cases, we let statistical software do the calculations. All we need to know is **what to calculate**.

Road map

- Random variables.
- Expectation and variances.
- Continuous distributions.
- **Normal distribution**.

Central tendency

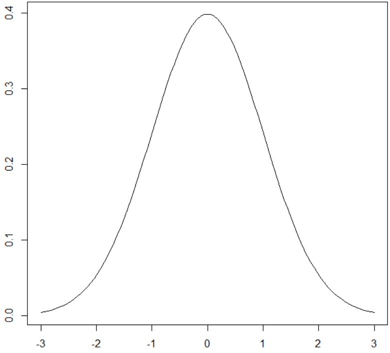
- In practice, typically data do not spread uniformly.
- Values tend to be **close to the center**.
  - Natural variables: heights of people, weights of dogs, lengths of leaves, temperature of a city, etc.
  - Performance: number of cars crossing a bridge, sales made by salespeople, consumer demands, student grades, etc.
  - All kinds of errors: estimation errors for consumer demand, differences from a manufacturing standard, etc.
- We need a distribution with such a central tendency.

Normal distribution

- A random variable X following a **normal distribution** with mean μ and standard deviation σ if its pdf is

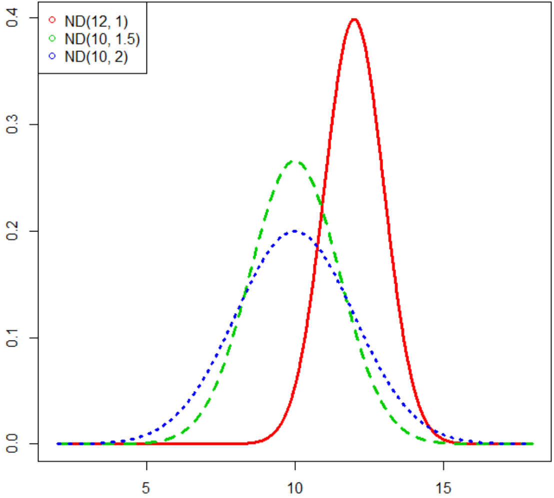
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  for all $x \in (-\infty, \infty)$.
  - If a random variable follows the normal distribution, most of its "normal values" will be close to the center.
  - We write $X \sim \text{ND}(\mu, \sigma)$.
- It is **symmetric** and **bell-shaped**.



Altering normal distributions

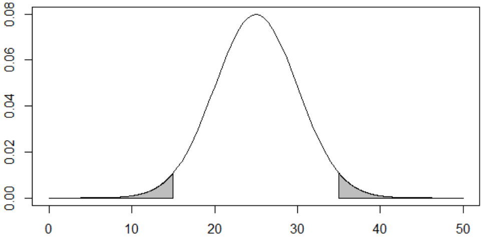
- Increasing the expected value μ shifts the curve to the right.
- Increasing the standard deviation σ makes the curve flatter.



Example 1: classroom temperature

- Let X be the room temperature when the next lecture starts.
- Suppose that $X \sim \text{ND}(25, 5)$.
- Suppose that the lecture must be canceled if $X < 15$ or $X > 35$.
- The probability for the lecture to be canceled is

$$\begin{aligned}\Pr(X < 15 \text{ or } X > 35) &= \Pr(X < 15) + \Pr(X > 35) \\ &= 2\Pr(X < 15) \approx 5\%.\end{aligned}$$



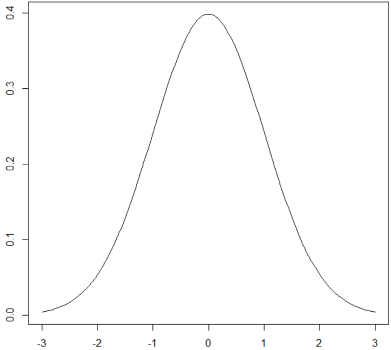
Standard normal distributions

- The **standard normal distribution** is a normal distribution with $\mu = 0$ and $\sigma = 1$.
- All normal distributions can be transformed to the standard normal distribution.

Proposition 1

If $X \sim \text{ND}(\mu, \sigma)$, *then* $Z = \frac{X-\mu}{\sigma} \sim \text{ND}(0, 1)$.

- This transformation is called **standardization**.



Equivalence among normal distributions

- Consider a normal random variable $X \sim \text{ND}(\mu, \sigma)$.
- For a value x, we define its **z-score** as $z = \frac{x-\mu}{\sigma}$.
  - It measures how far this value is from the mean, using the standard deviation as the unit of measurement.
  - E.g., if $z = 2$, the value is 2 standard deviations above the mean.
  - We say that x is **two-sigma above the mean**.
- Suppose that $X \sim \text{ND}(100, 20)$ and $Y \sim \text{ND}(90, 10)$.
  - For a value x to be two-sigma above the mean of X, $x = 140$.
  - For a value y to be two-sigma above the mean of Y, $y = 110$.
  - The standardization of normal distribution implies that

$$\begin{aligned}\Pr(X \geq 140) &= \Pr\left(\tfrac{X-100}{20} \geq \tfrac{140-100}{20}\right) = \Pr(Z \geq 2) \\ &= \Pr\left(\tfrac{Y-90}{10} \geq \tfrac{110-90}{10}\right) = \Pr(Y \geq 110).\end{aligned}$$

- "k-sigma away from the mean" is equivalent for **all** normal distribution!

The three-sigma rule for detecting outliers

- Recall our classroom temperature example:
  - $X \sim \text{ND}(25, 5)$ and $\Pr(X < 15) + \Pr(X > 35) \approx 5\%$.
  - For a normally distributed data set, the probability of being two-sigma away from the mean is 5%.
  - For a normally distributed data set, the probability of being two-sigma above (below) the mean is 2.5%.
- Recall our three-sigma rule for **detecting outliers**.
  - For any normal distribution, the probability of being three-sigma away from the mean is only 0.25%.
  - That is why the distance of three σs is suggested.