# Statistics for Data Science

## Wintersemester 2023/24

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Second lecture, October 23, 2023

# Random variables

# Random variables

Let $(\Omega, \mathbb{P})$ be a probability space and let $E$ be a set.

**Definition**

A random variable (RV) $X$ with values in $E$ is a function $X \colon \Omega \to E$.

*Remark.* The set $E$ is called the outcome or target space.

- When $E \subset \mathbb{R}$, we say that $X$ is a real-valued random variable.
- When $E \subset \mathbb{R}^n$, $n \geq 2$, we call $X$ a vector-valued random variable.
- When $E$ is countable, we call $X$ a discrete random variable.

In practice, $\omega$ is usually not observed directly and analysis is based on the observed random variable $X(\omega)$. Physically, one can think of a realization $X(\omega)$ of a random variable for some $\omega \in \Omega$ as some measurement, or observation performed on a system.

Statistical analysis is based on the *pushforward measure* $B \mapsto \mathbb{P}(X^{-1}(B))$, also called the *probability distribution* or *law* of $X$, not on $\mathbb{P}$. Note that here $X^{-1}(B) := \{\omega \in \Omega \mid X(\omega) \in B\}$ is the preimage of $B$ under the mapping $X$.

# Example, two dice

As an example of a random variable, consider the sum:

$$X\colon \{(1,1),(1,2),\dots,(6,6)\} \to \{2,\dots,12\},\ X(\omega) = \omega_1 + \omega_2.$$

The identity function $Y(\omega_1, \omega_2) = (\omega_1, \omega_2)$ also defines a random variable. Since $Y\colon \Omega \to \mathbb{R}^2$, this random variable is vector-valued.

Let $(\Omega, \mathbb{P})$ be a probability space and $E$ a set. A random variable $X\colon \Omega \to E$ induces a probability measure $P_X$ on $E$, defined by

$$P_X(B) := \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) \quad \text{for all subsets } B \subset E,$$

which is called the probability distribution (or law) of $X$.

*In other words, a random variable $X$ connects an event $B \subset E$ with a corresponding event $X^{-1}(B) \subset \Omega$ and assigns the probability of $X^{-1}(B)$ to $B$.*

Often, we shall simply denote

$$\{X \in B\} := \{\omega \in \Omega \mid X(\omega) \in B\},$$

and write

$$P_X(B) = \mathbb{P}(X \in B).$$

Two random variables $X$ and $Y$ with the same target space $E$ are said to be equal in law if they have the same probability function, i.e.,

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \quad \text{for all subsets } B \subset E.$$

Usually, we are ultimately interested in the laws of random variables, rather than the random variables *per se*.

### Example

Two players play Heads and Tails on a fair coin. The coin is thrown 10 times, the gain of player 1 is the total number of Heads, while the gain of player 2 is the total number of Tails. This situation is modeled by introducing $\Omega = \{H, T\}^{10}$ endowed with the uniform distribution, and defining random variables $X$ and $Y$ by

$$X(\omega) = \#\{i = 1, \dots, 10 \mid \omega_i = H\}, \quad Y(\omega) = \#\{i = 1, \dots, 10 \mid \omega_i = T\}$$

for all $\omega \in \{H, T\}^{10}$. Then $X + Y = 10$. Clearly $X$ and $Y$ are not equal, however they have equal distribution: for all $k$,

$$\mathbb{P}(X = k) = \frac{1}{2^{10}}\binom{10}{k} = \frac{1}{2^{10}}\binom{10}{10-k} = \mathbb{P}(X = 10 - k) = \mathbb{P}(Y = k).$$

This implies that $X$ and $Y$ are equal in distribution.

# Probability mass function

Let $(\Omega, \mathbb{P})$ be a probability space. Let $X \colon \Omega \to E$ be a discrete random variable (recall that this means that $E$ is countable). Then, for all $B \subset E$, we can write

$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x), \tag{1}$$

where $p_X(x) := \mathbb{P}(X = x)$, $x \in E$. We call $p_X$ the probability mass function (PMF) of $X$.

*Properties.* The PMF $p_X$ of a discrete random variable $X$ is

- non-negative $p_X(x) \geq 0$ for all $x \in E$;
- normalized $\sum_{x \in E} p_X(x) = 1$.

In consequence, $0 \leq p_X(x) \leq 1$ for all $x \in E$.

- The law of a discrete random variable $X$ with countable target space $E$ is uniquely determined by its PMF. This is a consequence of the fact that, by (1),

  $$P_X(B) := \mathbb{P}(X \in B) = \sum_{x \in B} p_X(x),$$

  meaning that the PMF *determines* the law of $X$ completely.

# Probability density function

### Definition

A function $f : \mathbb{R} \to \mathbb{R}$ is called a probability density function (PDF) if the following conditions hold:

- $f(x) \geq 0 \quad$ for all $x \in \mathbb{R}$;
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$

A real-valued random variable $X$ is said to be a continuous random variable if there exists a PDF $f_X : \mathbb{R} \to \mathbb{R}$ such that, for all $a \leq b$, there holds

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x. \tag{2}$$

Then we call $f_X$ the probability density function (PDF) of $X$.

Equation (2) implies for any (measurable) subset $A \subset \mathbb{R}$ that

$$P_X(A) := \mathbb{P}(X \in A) = \int_A f_X(x)\,\mathrm{d}x,$$

meaning that the PDF $f_X$ *determines* the law of $X$ completely.

*Remark.* One may think of the PDF as a "continuous" version of the PMF. However, the PMF and PDF are two quite different types of functions.

- The PMF of a *discrete random variable* $X$ can take values between $[0, 1]$, i.e.,
$$\mathbb{P}(X = x) = p_X(x) \in [0, 1].$$
- For a *continuous random variable* $X$, there *always* holds
$$\mathbb{P}(X = x) = \int_x^x f_X(y)\,\mathrm{d}y = 0.$$

# Examples of discrete random variables

**Example**

Let $p \in (0,1)$. Let $X$ be a random variable with values in $E = \{0,1\}$ and with PMF given by

$$p_X(x) = \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1. \end{cases}$$

Then we say that $X$ is a Bernoulli random variable with parameter $p$, and we write

$$X \sim \mathrm{Ber}(p).$$

A Bernoulli random variable with parameter $p$ represents the result of throwing a coin that falls on Heads with probability $p$ and Tails with probability $1-p$ ($p = 1/2$ is the coin is fair).

### Example

Let $p \in (0,1)$ and $n \geq 1$ an integer. Let $X$ be a random variable with values in $\{0, \ldots, n\}$ and with PMF given by

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, \ldots, n\}.$$

Then we say that $X$ is a binomial random variable with parameters $n$ and $p$, and we write

$$X \sim \mathrm{Bin}(n, p).$$

This corresponds to the probability of the number of times a coin lands on Heads in $n$ tosses of a coin, with $p$ denoting the probability of a coin landing on Heads.

Example

Let $p \in (0, 1)$. Let $X$ be a random variable with values in $\mathbb{N}$ and with PMF given by

$$p_X(x) = (1-p)^{x-1}p, \quad x \geq 1.$$

Then we say that $X$ is a geometric random variable with parameter $p$, and we write

$$X \sim \mathrm{Geo}(p).$$

This corresponds with the probability of hitting Heads for the first time, when the probability of hitting Heads is equal to $p$.

That is,

$$\mathbb{P}(X = k) = p_X(k) = (1-p)^{k-1}p$$

denotes the probability of hitting Tails for the first $k-1$ rounds and hitting heads on the $k^{\mathrm{th}}$ round.

**Example**

Let $\lambda > 0$. Let $X$ be a random variable with values in $\mathbb{N}_0$ and with PMF given by

$$p_X(x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x \geq 0.$$

We then say that $X$ is a Poisson random variable with parameter $\lambda$, and we write

$$X \sim \mathrm{Poisson}(\lambda).$$

Poisson random variables can be used to model the count of rare events such as nuclei decaying in a radioactive sample.

# Examples of continuous real-valued random variables

**Definition**

Let $a < b$. Let $X$ be a real-valued continuous random variable with PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise}, \end{cases} \qquad x \in \mathbb{R}.$$

We then say that $X$ is a uniform random variable in $[a, b]$, and we write

$$X \sim \mathcal{U}(a, b).$$

### Definition

Let $\lambda > 0$. Let $X$ be a real-valued continuous random variable with PDF

$$f_X(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \qquad x \in \mathbb{R}.$$

We then say that $X$ is an exponential random variable with parameter $\lambda$, and we write

$$X \sim \mathrm{Exp}(\lambda).$$

### Example

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. Let $X$ be a real-valued continuous random variable with PDF given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

We then say that $X$ is a Gaussian random variable with parameters $\mu$ and $\sigma^2$, and we write

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

The parameter $\mu$ is called the mean and $\sigma$ is called the standard deviation of $X$.

# Cumulative distribution function

The cumulative distribution function (CDF) of a real-valued random variable $X$ is the function $F_X : \mathbb{R} \to [0,1]$ given by

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \le x\}) . \qquad \text{(or shortly} = \mathbb{P}(X \le x))$$

Note that the CDF is defined for any random variable taking values in $\mathbb{R}$, whether discrete or continuous.

**Proposition**

*Let $F_X : \mathbb{R} \to [0,1]$ be the CDF of a real-valued random variable $X$. Then*

- *$F_X$ is non-decreasing: if $a \le b$, then $F_X(a) \le F_X(b)$.*
- *$F_X$ is right-continuous: for all $a \in \mathbb{R}$,*

$$F_X(a) = \lim_{x \to a+} F_X(x).$$

- *$F_X(-\infty) := \lim_{x \to -\infty} F_X(x) = 0$ and $F_X(\infty) := \lim_{x \to \infty} F_X(x) = 1$.*

One can read off relevant information on the distribution of $X$ from its CDF.

**Lemma**

*Let $F_X : \mathbb{R} \to [0,1]$ be the CDF of a real-valued random variable $X$. Then*

- *For any real numbers $a < b$,*

$$\mathbb{P}(a < x \leq b) = F_X(b) - F_X(a).$$

- *For any $a \in \mathbb{R}$,*

$$\mathbb{P}(X > a) = 1 - F_x(a).$$

- *For any $x \in \mathbb{R}$,*

$$\mathbb{P}(X = x) = F_X(x) - \lim_{y \to x-} F_X(y).$$

*Remark.* In particular, if $X$ is a continuous random variable, we have $F_X(x) = \lim_{y \to x-} F_X(y)$ for all $x \in \mathbb{R}$; no jumps occur. For a discrete random variable, the situation is different: $F_X$ is then a pure-jump function, meaning that it increases purely through jumps.

# Relationship between the CDF and PMF (discrete case)

### Proposition

*Let $X$ be a discrete random variable taking values in a countable subset $E$ of $\mathbb{R}$. Denoting the PMF of $X$ by $p_X$ and its CDF by $F_X$, we have*

$$F_X(a) = \sum_{\substack{x \in E \\ x \leq a}} p_X(x) \quad \textit{for all } a \in \mathbb{R},$$

$$p_X(x) = F_X(x) - \lim_{y \to x-} F_X(y).$$

*Proof.* By the definition of the PMF, there holds

$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x) \quad \text{for all subsets } B \subset E.$$

Setting $B = \{x \in E \mid x \leq a\}$ yields the first relation.

For the second relation, we note that

$$\{X = x\} = \bigcap_{n \geq 1} E_n,$$

where the sets $E_n := \{X \in (x - \frac{1}{n}, x]\}$ form a decreasing sequence of events $E_{n+1} \subset E_n$ for $n \geq 1$. In this case, there holds

$$\begin{aligned} \mathbb{P}\Big(\bigcap_{n \geq 1} E_n\Big) &= \lim_{n \to \infty} \mathbb{P}(E_n) \\ &= \lim_{n \to \infty} \big(F_X(x) - F_X\big(x - \tfrac{1}{n}\big)\big) \\ &= F_X(x) - \lim_{y \to x-} F_X(y), \end{aligned}$$

as desired. □

# Relationship between the CDF and PDF (continuous case)

**Proposition**

*Let $X$ be a continuous real-valued random variable. Denoting the PDF of $X$ by $f_X$, and its CDF by $F_X$, we have*

$$F_X(a) = \int_{-\infty}^{a} f_X(y)\,\mathrm{d}y \quad \textit{for all } a \in \mathbb{R}.$$

*In addition, if $F_X$ is differentiable at $x \in E$, we have*

$$f_X(x) = F_X'(x).$$

*Proof.* For the first statement, note that for all $u < a$ there holds

$$F_X(a) - F_X(u) = \mathbb{P}(X \in (u, a]) = \mathbb{P}(X \in [u, a]) = \int_{u}^{a} f_X(y)\,\mathrm{d}y,$$

where we used the fact that $\mathbb{P}(X = u) = 0$ since $X$ is a continuous random variable. Letting $u \to -\infty$ and recalling $F_X(-\infty) = 0$, we obtain $F_X(a) = \int_{-\infty}^{a} f_X(y)\,\mathrm{d}y$. The second statement follows from the fundamental theorem of calculus ($F_X$ is the antiderivative of $f_X$). $\square$

### Proposition

*The probability distribution of a real-valued random variable is uniquely determined by its CDF.*

*Proof.* We give a proof in the discrete case. Let $X$ and $Y$ be two real-valued random variables with the same CDF:

$$F_X(x) = F_Y(x) \quad \text{for all } x \in \mathbb{R}.$$

Then by the previous discussion,

$$p_X(x) = F_X(x) - \lim_{y \to x-} F_X(y) = F_Y(x) - \lim_{y \to x-} F_Y(y) = p_Y(x).$$

Thus $X$ and $Y$ have the same PMF, meaning that $X$ and $Y$ are equal in law. □

# Quantile function

**Definition (Revised 30.10.2023)**

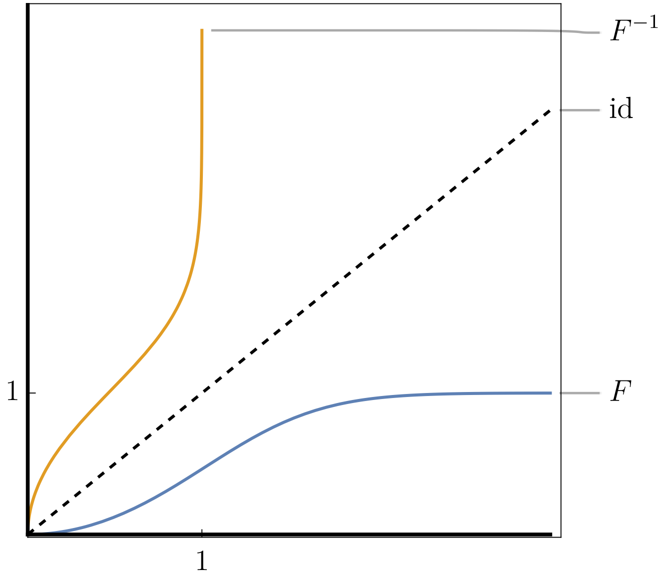
Let $X$ be a real-valued random variable with CDF $F$. The generalized inverse $F^{-1}\colon (0,1) \to \mathbb{R}$,

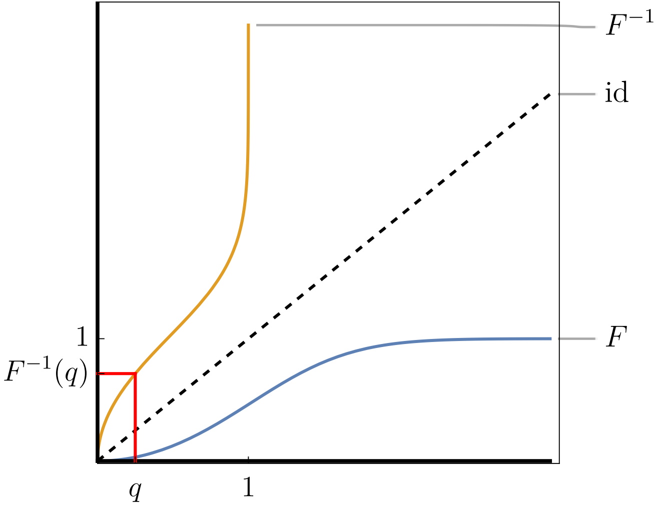
$$F^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \geq q\}, \quad q \in (0,1), \tag{3}$$

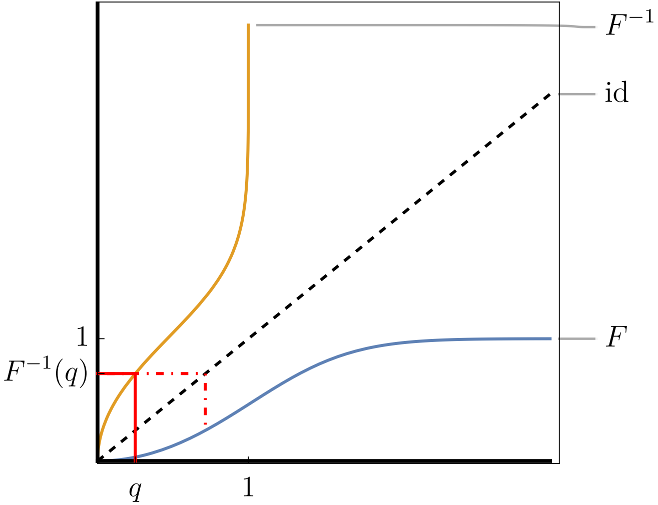
is called the quantile function of $X$.

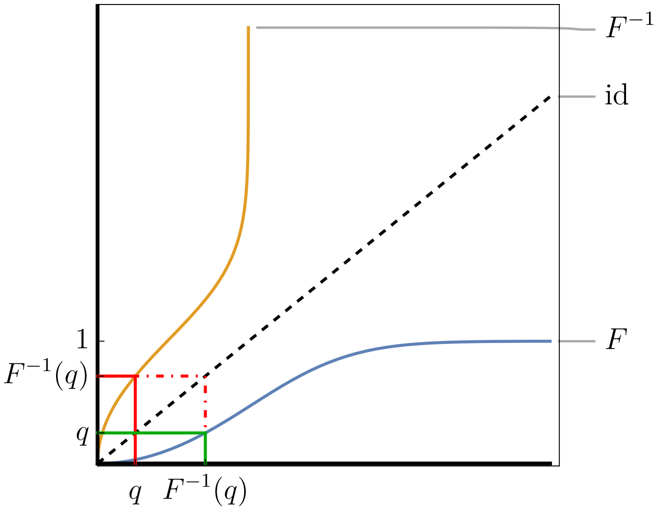
- If $F$ is strictly increasing, then the quantile function is the inverse function of $F$.
- For example, the CDF and inverse CDF of a Bernoulli random variable $X \sim \mathrm{Ber}(\frac{1}{2})$ are $F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$ and $F^{-1}(q) = \begin{cases} 0 & \text{if } 0 < q \leq \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} < q < 1. \end{cases}$

*Remark.* Another way to define the quantile function is $Q(q) = \inf\{x \in \mathbb{R} \mid F(x) > q\}$, $q \in (0,1)$. For the Bernoulli random variable $X \sim \mathrm{Ber}(\frac{1}{2})$, we would have $Q(q) = \begin{cases} 0 & \text{if } 0 < q < \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq q < 1 \end{cases}$ (note the difference in the semiopen intervals).

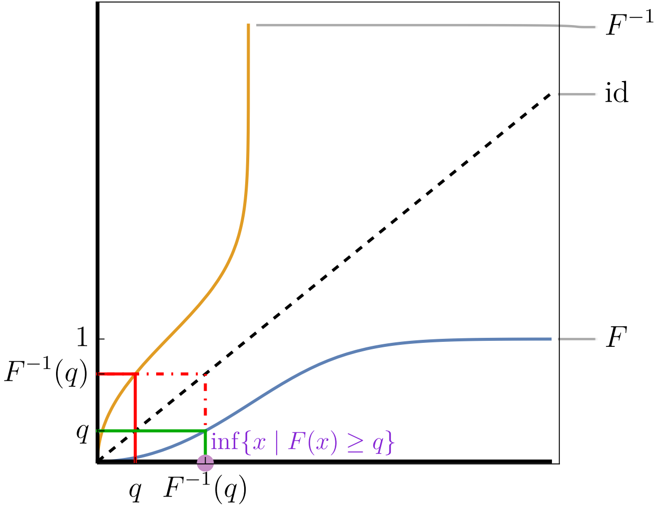

$F^{-1}$
id
$F$
1
1


$F^{-1}$
id
$F$
1
$F^{-1}(q)$
$q$
1


$F^{-1}$
id
$F$
1
$F^{-1}(q)$
$q$
1



$F^{-1}$
id
$F$
1
$F^{-1}(q)$
$q$
$q$
$F^{-1}(q)$




"Find the smallest value of $x$ such that $F(x) \geq q$."

## Proposition

*Let $X$ be a real-valued random variable with CDF $F_X$. Then*

1. *For all $q \in (0,1)$, $F_X(F_X^{-1}(q)) \geq q$.*
2. *If $X$ is a continuous random variable, then $F_X(F_X^{-1}(q)) = q$ for all $q \in (0,1)$.*

*Proof.* (1) Let $q \in (0,1)$. Since $F_X^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \geq q\}$ by definition, we can find a sequence $(a_n)_{n\geq 1}$ of real numbers such that $F_X(a_n) \geq q$ and $a_n \searrow F_X^{-1}(q)$. By the right-continuity of $F_X$, there holds

$$F_X(F_X^{-1}(q)) = \lim_{n\to\infty} F_X(a_n) \geq q.$$

(2) It suffices to prove the inequality $F_X(F_X^{-1}(q)) \leq q$ by (1). Assume to the contrary that $F_X(F_X^{-1}(q)) > q$. Since $F_X$ is the CDF of a continuous random variable, it is continuous. By continuity of $F_X$, there exists $a \in (-\infty, F_X^{-1}(q))$ such that $F_X(a) > q$, which contradicts the definition of $F_X^{-1}$. $\square$

# CDF of a normal random variable

**Example**

The CDF of a normal random variable $X \sim \mathcal{N}(0,1)$ is often denoted by $\Phi$,

$$\Phi(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t, \quad x \in \mathbb{R}.$$

Typical values to remember:

$$\Phi(1.645) = \mathbb{P}(X \leq 1.645) \approx 0.95,$$
$$\Phi(1.960) = \mathbb{P}(X \leq 1.960) \approx 0.975.$$

In this case the CDF $\Phi$ is injective and the quantile function, denoted by $\Phi^{-1}$, coincides with its inverse. The above equalities can be recast as

$$\Phi^{-1}(0.95) \approx 1.645,$$
$$\Phi^{-1}(0.975) \approx 1.960.$$