

Description

`ustrlen(s)` returns the number of Unicode characters in the Unicode string *s*. An invalid UTF-8 sequence is counted as one Unicode character. Note that any Unicode character besides ASCII characters (0–127) takes more than 1 byte in UTF-8 encoding, for example, “é” takes 2 bytes.

`ustrninvalidcnt(s)` returns the number of invalid UTF-8 sequences in *s*. An invalid UTF-8 sequence can contain one byte or multiple bytes.

When *s* is not a scalar, functions return element-by-element results.

Syntax

real matrix `ustrlen(string matrix s)`

real matrix `ustrninvalidcnt(string matrix s)`

Remarks and examples

`ustrlen(s)`, when *s* is a binary string (a string containing null terminator `char(0)`), returns the overall length of the Unicode string. Note that null terminator `char(0)` is a valid Unicode code point.

Use `udstrlen()` to obtain the length of a string in display columns. Use `strlen()` to obtain the length of a string in bytes. See [\[U\] 12.4.2.2 Displaying Unicode characters](#).

Conformability

`ustrlen(s)`, `ustrninvalidcnt(s)`:

<i>s</i> :	$r \times c$
<i>result</i> :	$r \times c$

Diagnostics

`ustrlen(s)` and `ustrninvalidcnt(s)` return negative error codes if an error occurs.

Also see

[M-5] **strlen()** — Length of string in bytes

[M-5] **udstrlen()** — Length of Unicode string in display columns

[M-4] **String** — String manipulation functions

[U] **12.4.2.2 Displaying Unicode characters**

Stata, Stata Press, Mata, NetCourse, and NetCourseNow are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow is a trademark of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.



For suggested citations, see the FAQ on [citing Stata documentation](#).