Descriptive Statistics

Chunyen

2024/3/13

Introduction

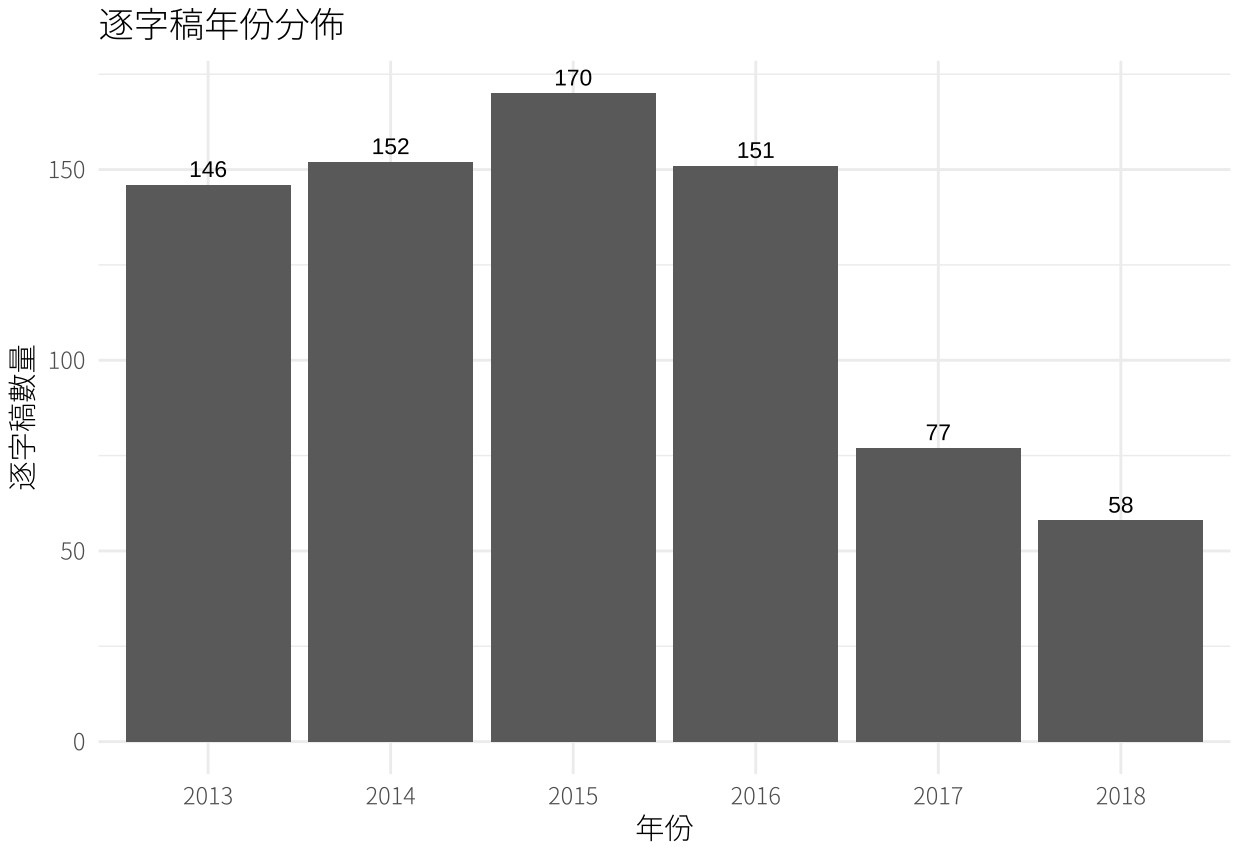
描述性統計自殺風險等級、逐字稿、參照句子

自殺風險等級

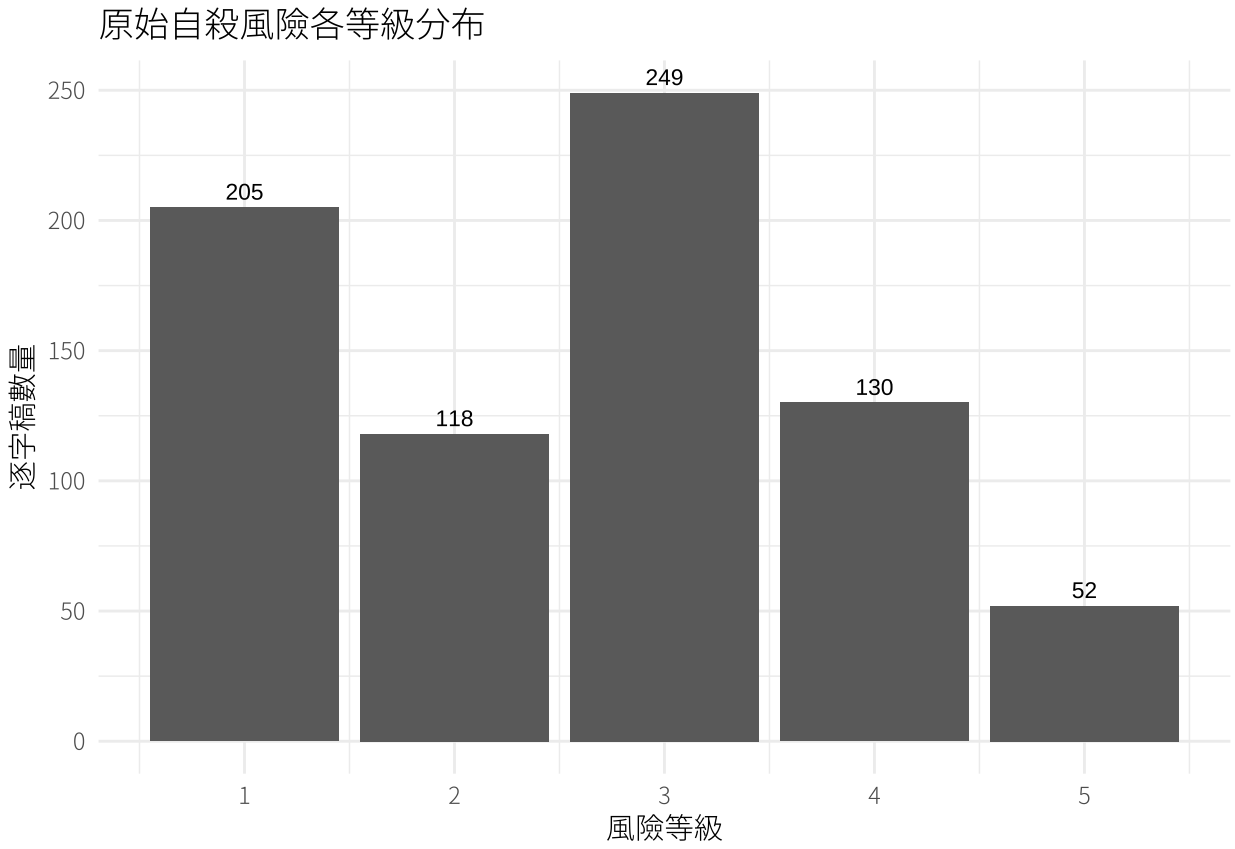
`risklevel.csv` 的架構

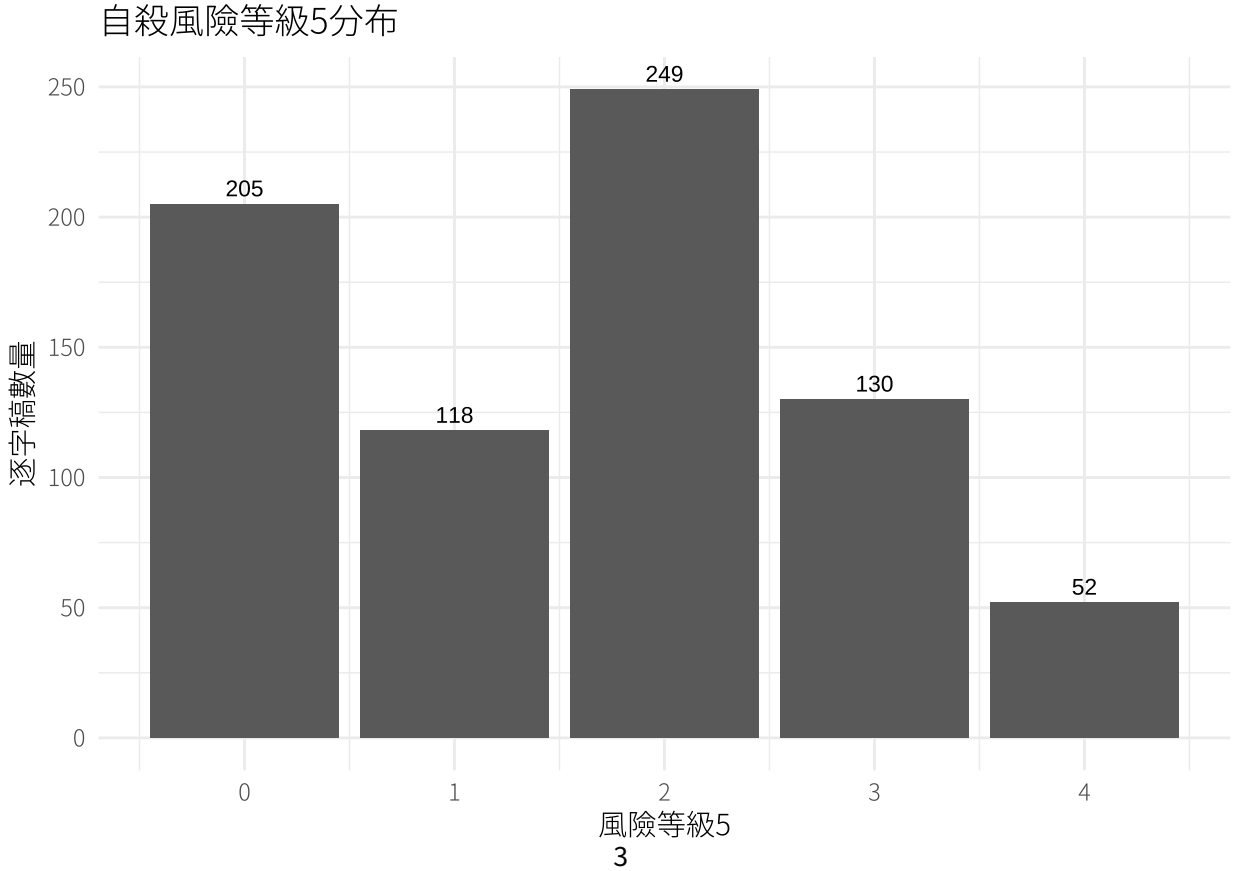
```
## spc_tbl_ [754 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ dialogue_id: num [1:754] 2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
##  $ risk_level : num [1:754] 2 2 2 1 5 1 1 4 2 2 ...
##  $ risk3      : num [1:754] 0 0 0 0 2 0 0 2 0 0 ...
##  $ risk4      : num [1:754] 1 1 1 0 3 0 0 3 1 1 ...
##  $ risk5      : num [1:754] 1 1 1 0 4 0 0 3 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   dialogue_id = col_double(),
##   ..   risk_level = col_double(),
##   ..   risk3 = col_double(),
##   ..   risk4 = col_double(),
##   ..   risk5 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

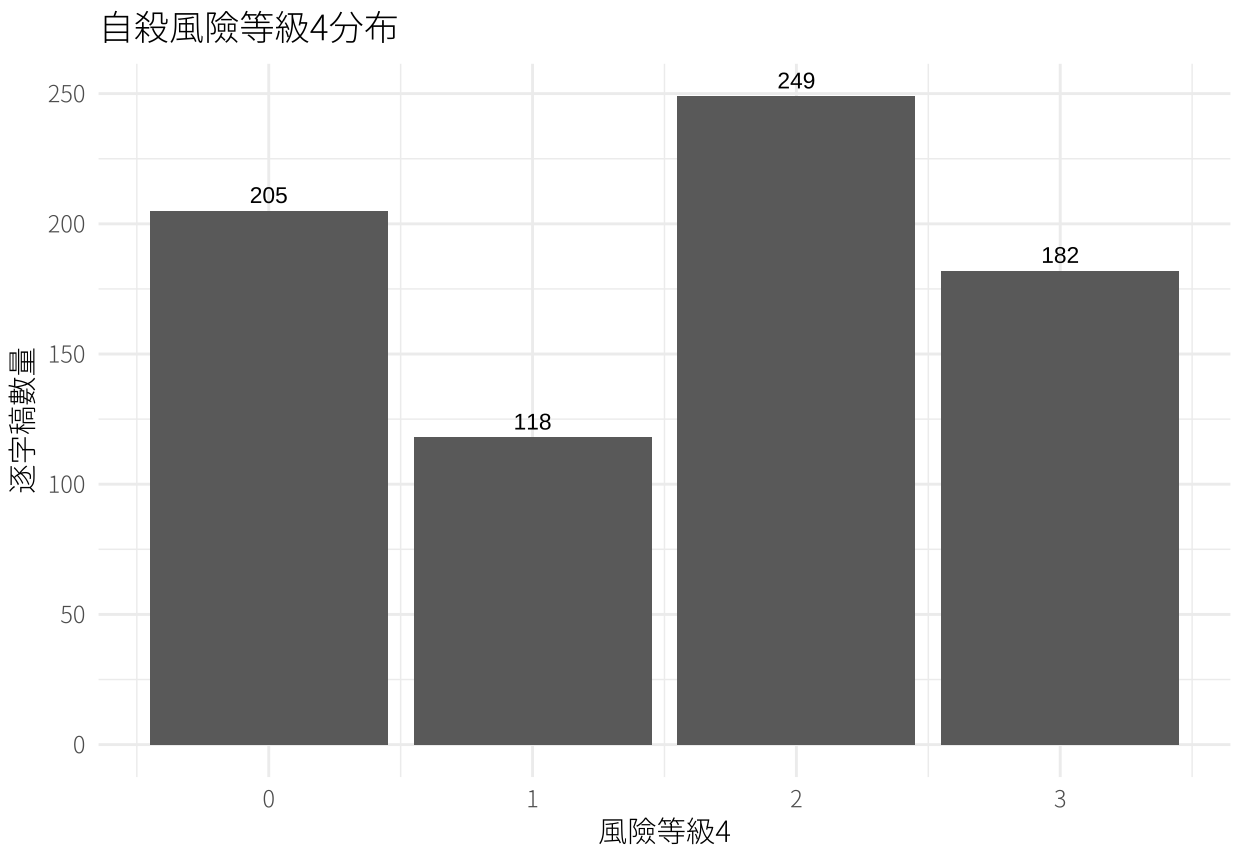
逐字稿年份分佈

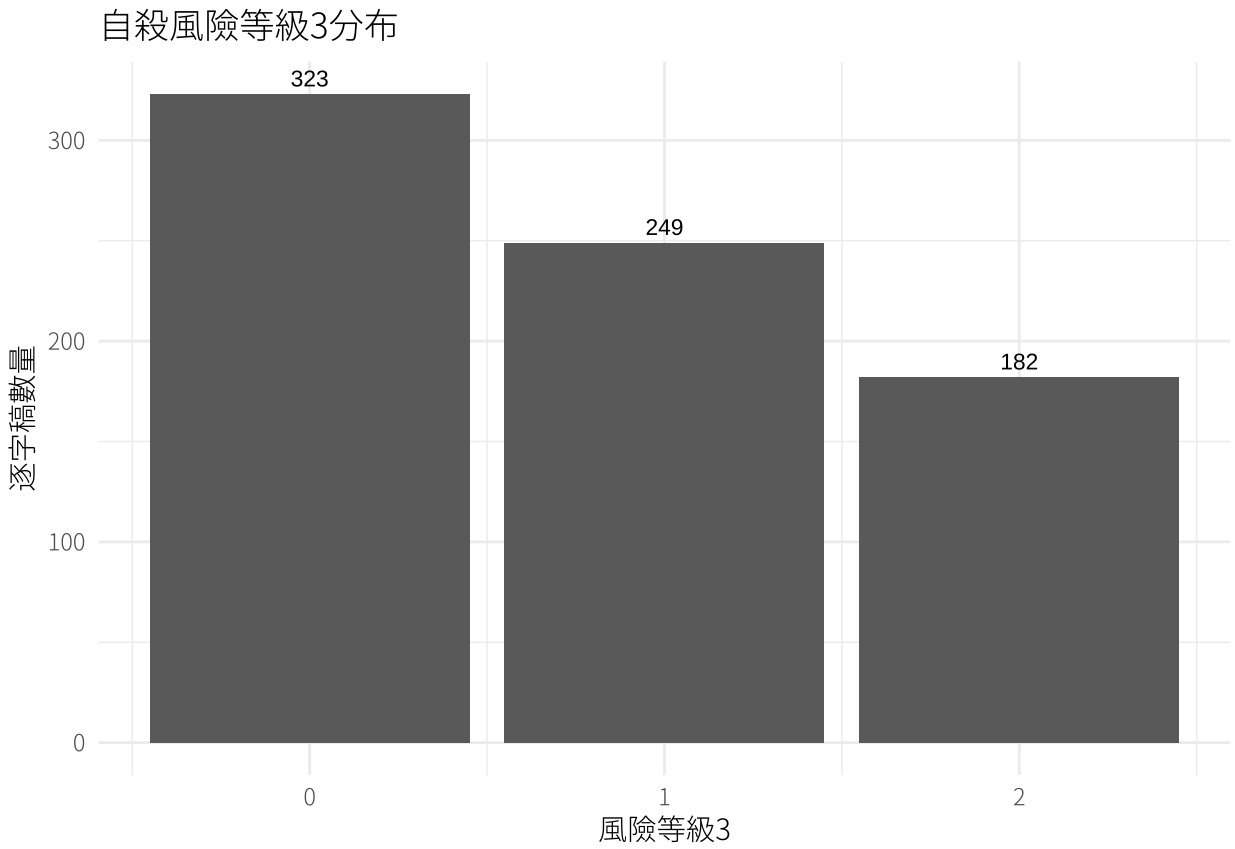

逐字稿年份分佈
逐字稿數量
150
100
50
0
146
152
170
151
77
58
2013
2014
2015
2016
2017
2018
年份

自殺風險各等級分布


原始自殺風險各等級分布
250
200
150
100
50
0
逐字稿數量
205
118
249
130
52
1
2
3
4
5
風險等級


自殺風險等級5分布
250
200
150
100
50
0
逐字稿數量
205
118
249
130
52
0
1
2
3
4
風險等級5


自殺風險等級4分布
250
200
150
100
50
0
逐字稿數量
205
118
249
182
0
1
2
3
風險等級4



逐字稿

只有求助者的發言語句

transcripts_emb_caller.csv 的架構

```
## spc_tbl_ [206,355 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ dialogue_id  : num [1:206355] 2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
##  $ utterance_id : num [1:206355] 73 75 77 80 82 322 324 326 328 329 ...
##  $ utterance_len: num [1:206355] 2 2 2 3 2 2 1 5 13 19 ...
##  $ role         : chr [1:206355] "c" "c" "c" "c" ...
##  $ utterance    : chr [1:206355] "？" "…" " " "  " ...
##  $ emb          : chr [1:206355] "[ 6.58912212e-02 -1.36712044e-02  7.48135820e-02  4.55615669e-02\n
##  $ max_sim_ref1 : num [1:206355] 0.272 0.317 0.305 0.347 0.316 ...
##  $ avg_sim_ref1 : num [1:206355] 0.0587 0.1437 0.1413 0.1848 0.1552 ...
##  $ max_sim_ref2 : num [1:206355] 0.272 0.317 0.305 0.347 0.316 ...
##  $ avg_sim_ref2 : num [1:206355] 0.0484 0.1521 0.1605 0.1917 0.1578 ...
##  $ max_sim_ref3 : num [1:206355] 0.148 0.2 0.21 0.257 0.204 ...
##  $ avg_sim_ref3 : num [1:206355] 0.0701 0.1345 0.1201 0.1773 0.1524 ...
##  $ max_sim_ref4 : num [1:206355] 0.736 0.856 0.839 0.921 0.878 ...
##  $ avg_sim_ref4 : num [1:206355] 0.47 0.582 0.542 0.646 0.619 ...
##  - attr(*, "spec")=
##   .. cols(
```

```
##   ..   dialogue_id = col_double(),
##   ..   utterance_id = col_double(),
##   ..   utterance_len = col_double(),
##   ..   role = col_character(),
##   ..   utterance = col_character(),
##   ..   emb = col_character(),
##   ..   max_sim_ref1 = col_double(),
##   ..   avg_sim_ref1 = col_double(),
##   ..   max_sim_ref2 = col_double(),
##   ..   avg_sim_ref2 = col_double(),
##   ..   max_sim_ref3 = col_double(),
##   ..   avg_sim_ref3 = col_double(),
##   ..   max_sim_ref4 = col_double(),
##   ..   avg_sim_ref4 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

總字數、語句數

[1] “逐字稿共有 2842477 個字”

[1] “逐字稿共有 206355 個語句”

[1] “每個語句平均有 13.7746940951273 個字”

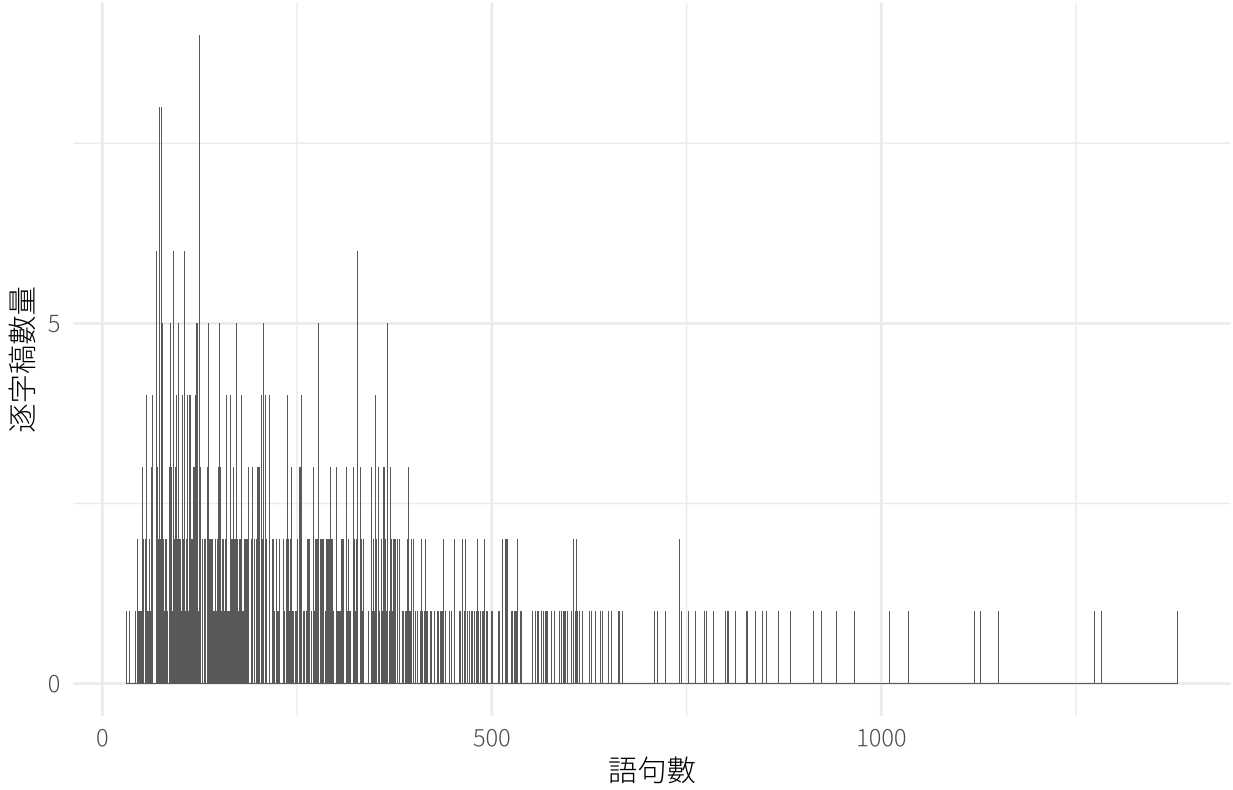
各發言語句中，發言的字數描述統計

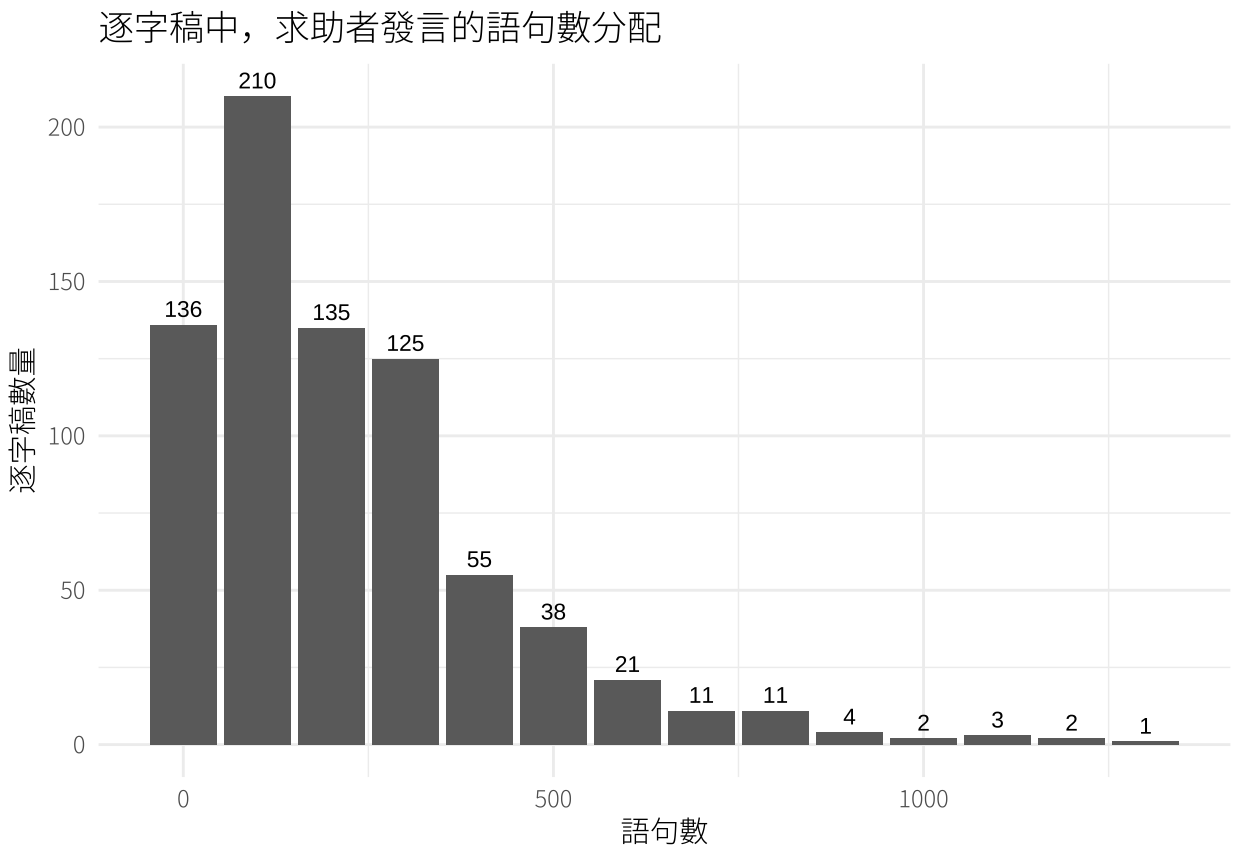
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
utterance_len	1	206355	13.77	17.52	9	10.64	10.38	1	946	945	6.54	146.98	0.04

各逐字稿中，發言的語句數分配

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
n	1	754	273.68	204.92	214.5	242.55	168.28	31	1380	1349	1.68	3.88	7.46

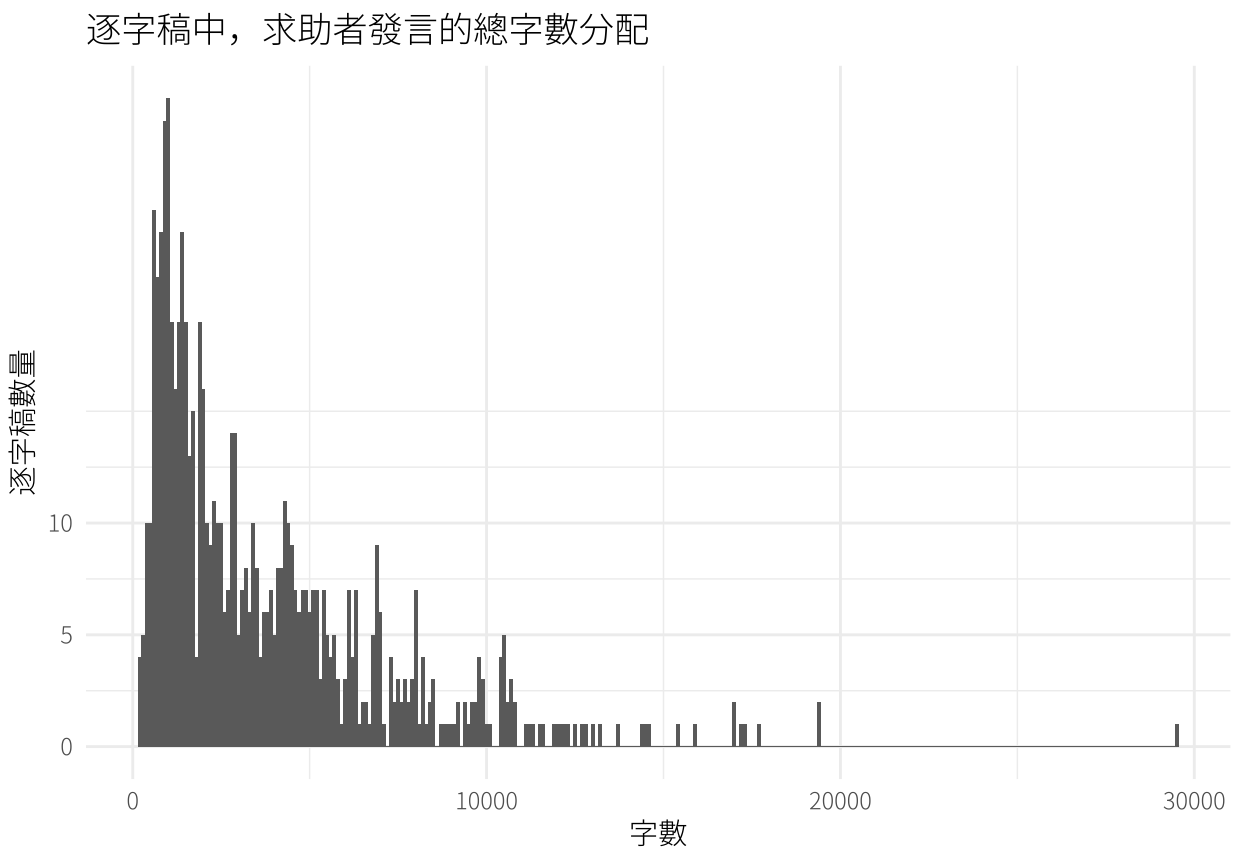
逐字稿中，求助者發言的語句數分配

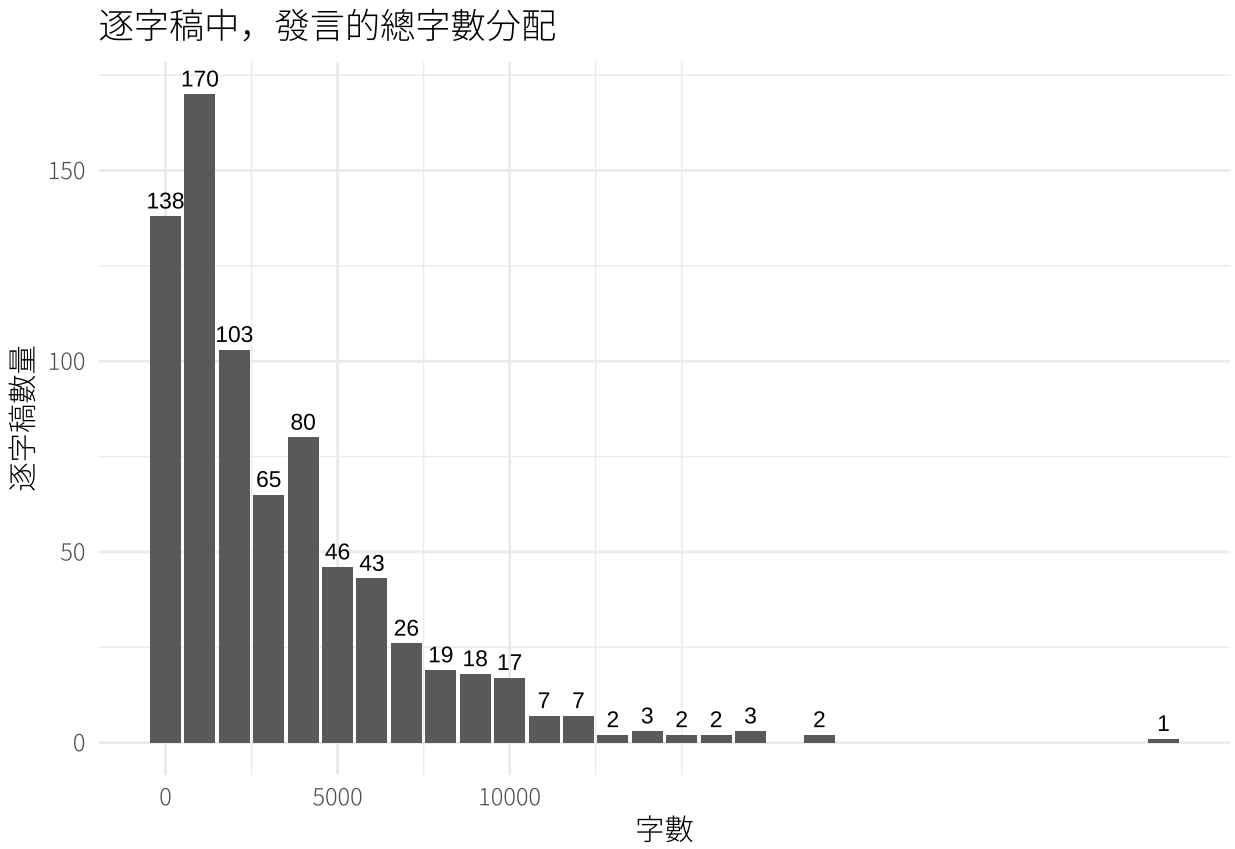




各逐字稿中，發言的總字數分配

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
n	1	754	3769.86	3444.3	2699.5	3195.55	2522.64	169	29522	29353	1.97	6.25	125.43


逐字稿中，求助者發言的總字數分配
逐字稿數量
10
5
0
0
10000
20000
30000
字數



只有接線員的發言語句

transcripts_emb_operator.csv 的架構

```
## spc_tbl_ [224,417 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ dialogue_id  : num [1:224417] 2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
##  $ utterance_id : num [1:224417] 1 3 5 7 9 11 13 17 19 21 ...
##  $ utterance_len: num [1:224417] 8 10 12 6 41 3 3 1 16 8 ...
##  $ role         : chr [1:224417] "r" "r" "r" "r" ...
##  $ utterance    : chr [1:224417] " ?    " "     ?" "      ?" "   ?" ...
##  $ emb          : chr [1:224417] "[ 6.05873056e-02  1.02354422e-01 -3.47988075e-03  6.18398283e-03\n
##  $ max_sim_ref1 : num [1:224417] 0.261 0.227 0.245 0.255 0.312 ...
##  $ avg_sim_ref1 : num [1:224417] 0.1245 0.0648 0.0958 0.1136 0.18 ...
##  $ max_sim_ref2 : num [1:224417] 0.261 0.227 0.198 0.255 0.259 ...
##  $ avg_sim_ref2 : num [1:224417] 0.1217 0.0721 0.0868 0.1272 0.1835 ...
##  $ max_sim_ref3 : num [1:224417] 0.216 0.181 0.245 0.189 0.312 ...
##  $ avg_sim_ref3 : num [1:224417] 0.1275 0.0568 0.1058 0.0985 0.1762 ...
##  $ max_sim_ref4 : num [1:224417] 0.658 0.508 0.524 0.588 0.436 ...
##  $ avg_sim_ref4 : num [1:224417] 0.429 0.185 0.217 0.258 0.212 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   dialogue_id = col_double(),
##   ..   utterance_id = col_double(),
##   ..   utterance_len = col_double(),
```

```
##   ..   role = col_character(),
##   ..   utterance = col_character(),
##   ..   emb = col_character(),
##   ..   max_sim_ref1 = col_double(),
##   ..   avg_sim_ref1 = col_double(),
##   ..   max_sim_ref2 = col_double(),
##   ..   avg_sim_ref2 = col_double(),
##   ..   max_sim_ref3 = col_double(),
##   ..   avg_sim_ref3 = col_double(),
##   ..   max_sim_ref4 = col_double(),
##   ..   avg_sim_ref4 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

總字數、語句數

[1] “逐字稿共有 2634172 個字”

[1] “逐字稿共有 224417 個語句”

[1] “每個語句平均有 11.7378451721572 個字”

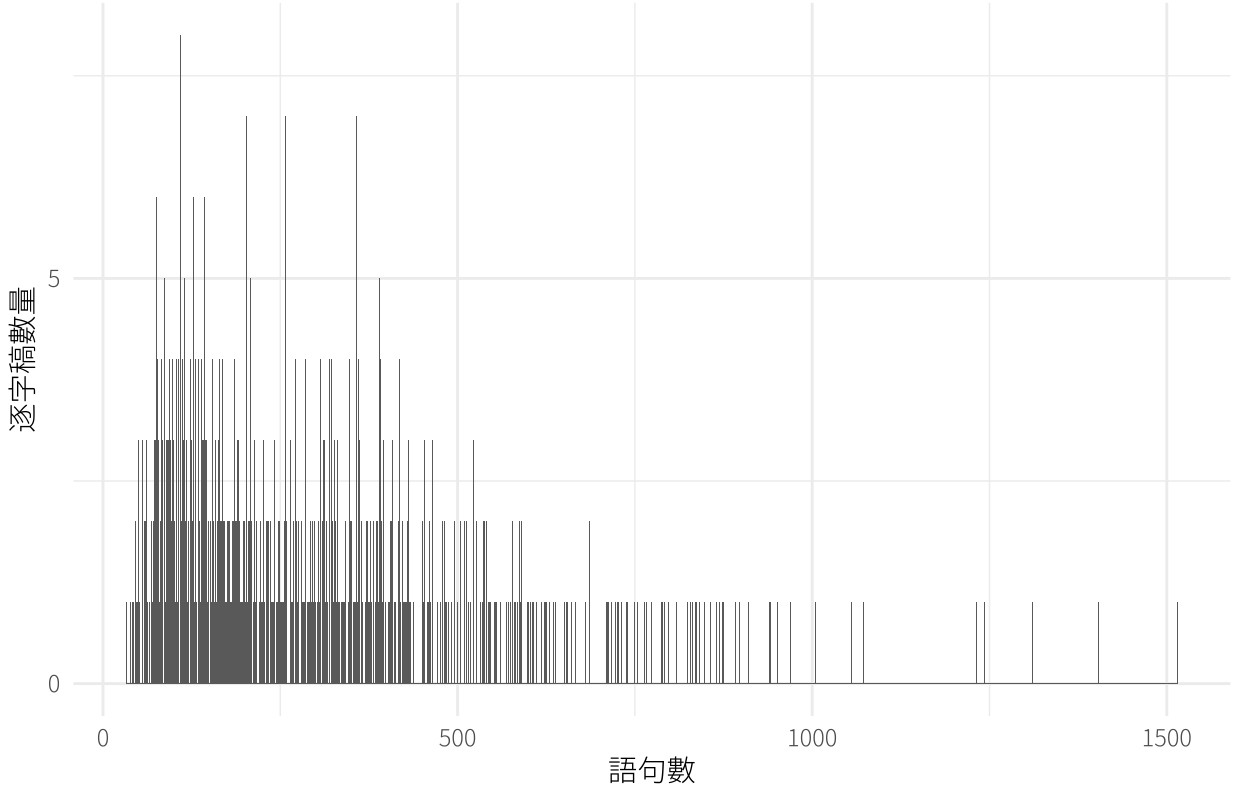
各發言語句中，發言的字數描述統計

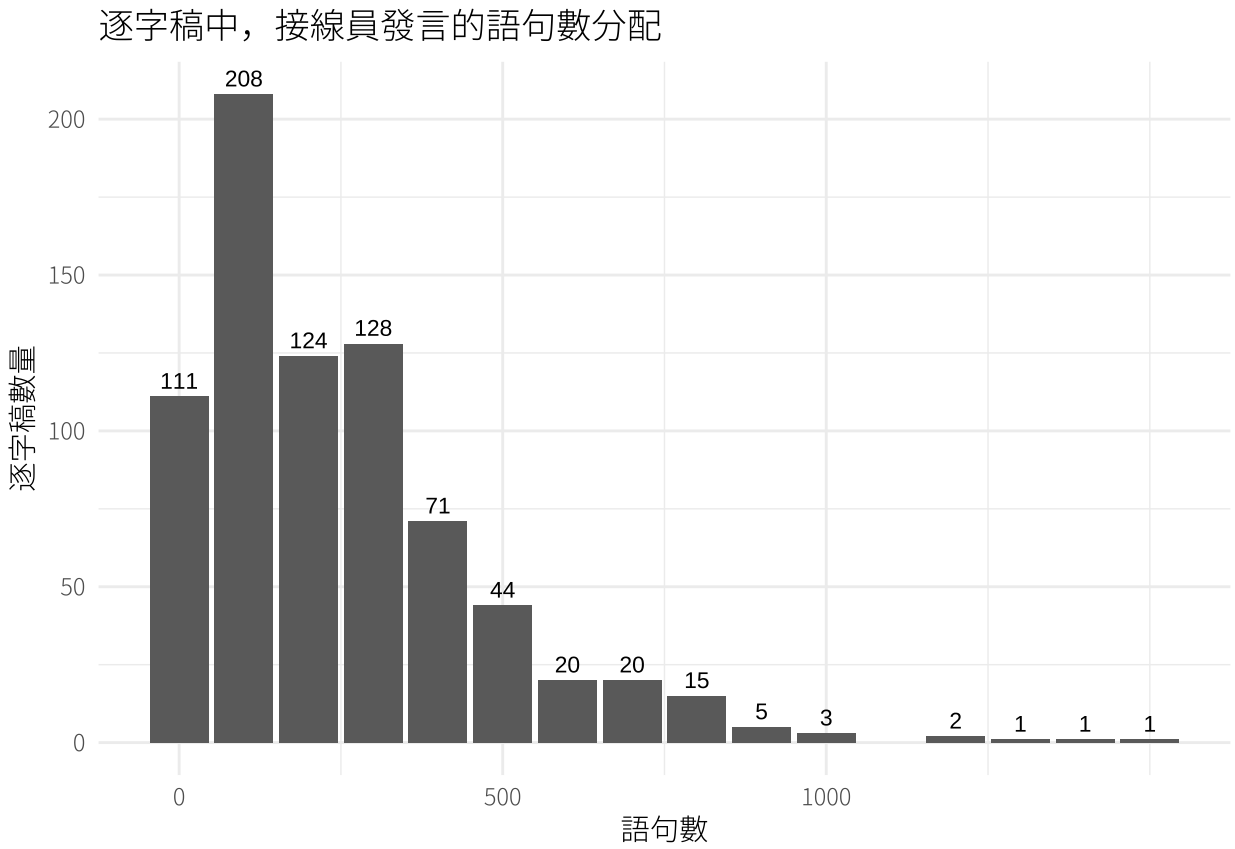
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
utterance_len	1	224417	11.74	19.04	5	7.84	5.93	1	957	956	6.94	137.91	0.04

各逐字稿中，發言的語句數分配

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
n	1	754	297.64	218.71	242	265.13	187.55	33	1515	1482	1.6	3.56	7.96

逐字稿中，接線員發言的語句數分配

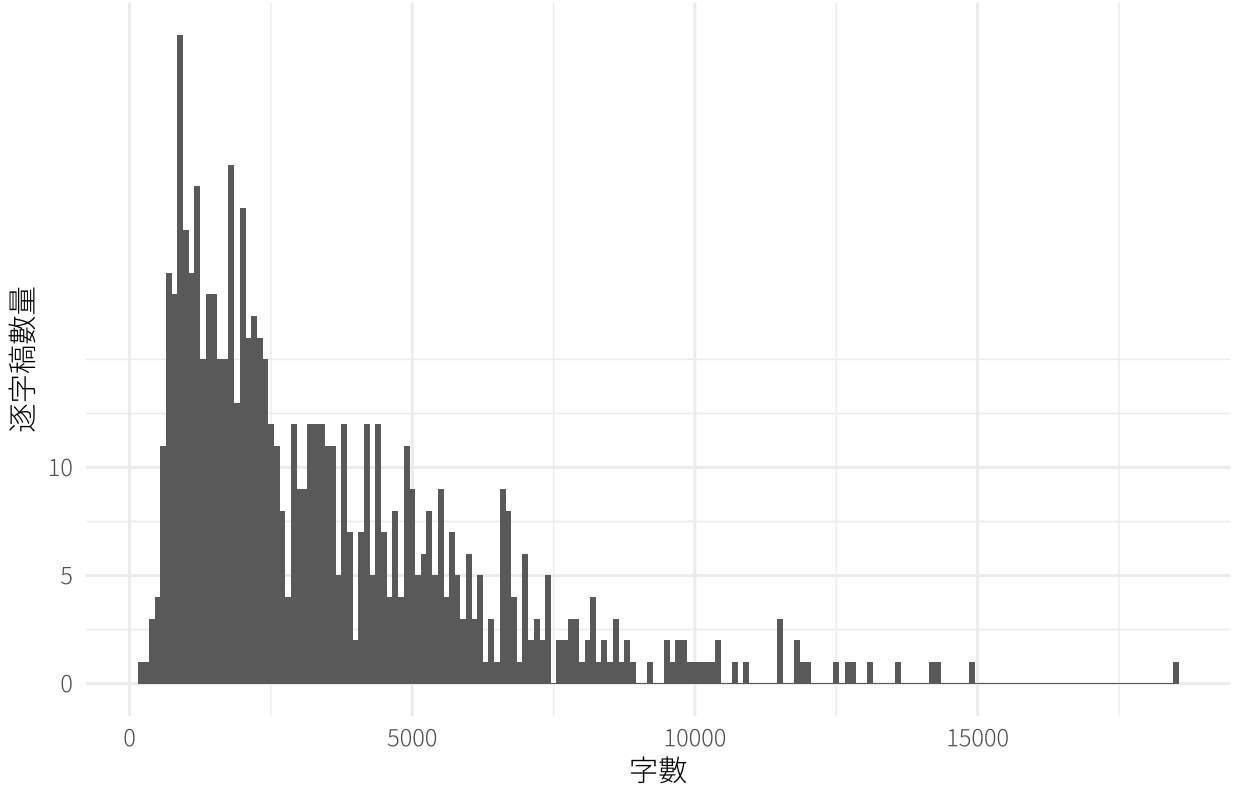


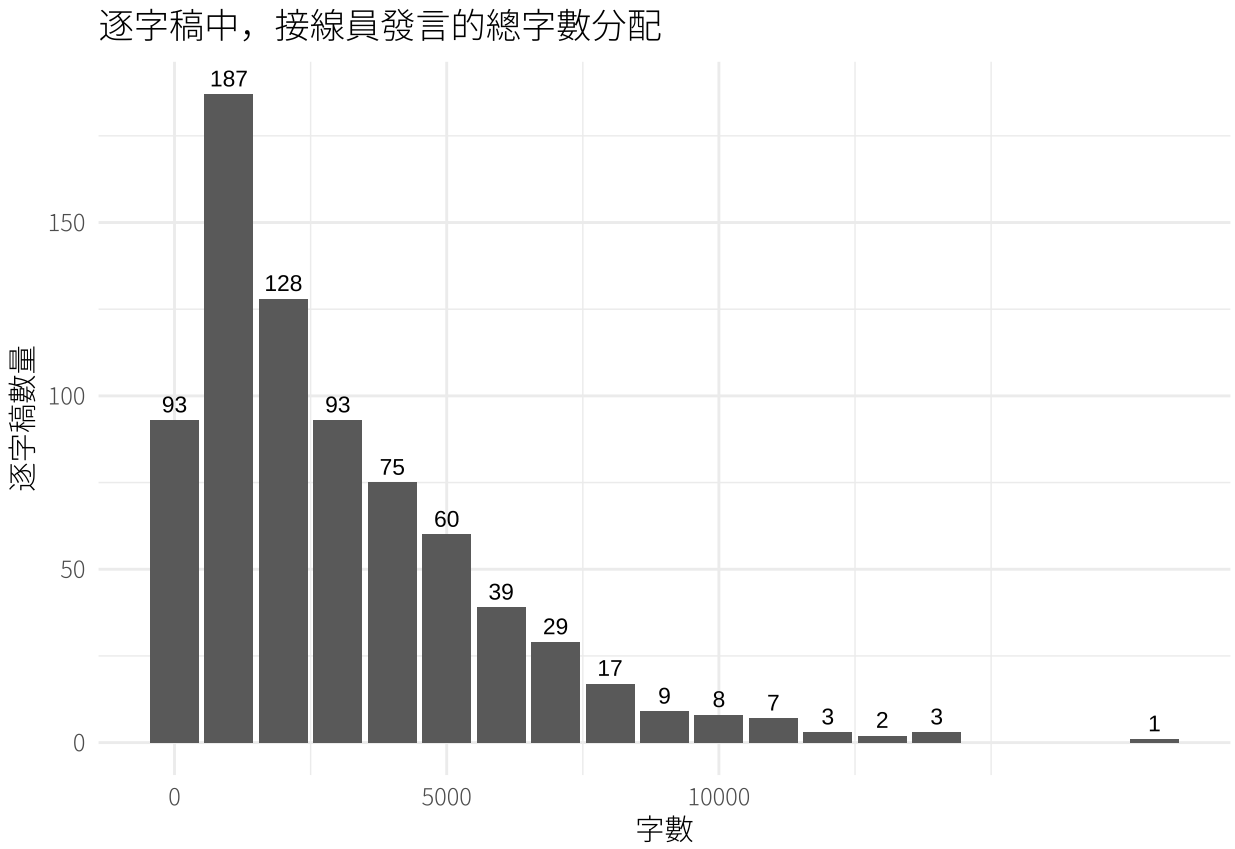


各逐字稿中，發言的總字數分配

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
n	1	754	3493.6	2675.87	2654	3097.95	2208.33	245	18458	18213	1.52	2.91	97.45

逐字稿中，接線員發言的總字數分配





參照句子

- 合併量表（自殺危險程度量表＋自殺意念量表）
- 自殺意念量表（BSS）
- 自殺危險程度量表
- 自殺辭典

題數

[1] “合併量表題數：40 題” [1] “自殺意念量表：21 題” [1] “自殺危險程度量表：19 題” [1] “自殺詞典：411 個詞”

字元數

[1] “合併量表總字元數：627 字” [1] “自殺意念量表總字元數：294 字” [1] “自殺危險程度量表總字元數：333 字” [1] “自殺詞典總字元數：1217 字”