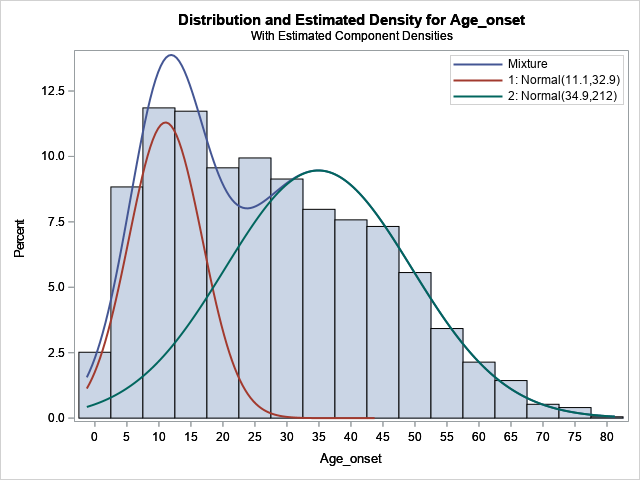
一.发病年龄双峰：

1、能否把整体的曲线和双峰放在一个图，如下图？

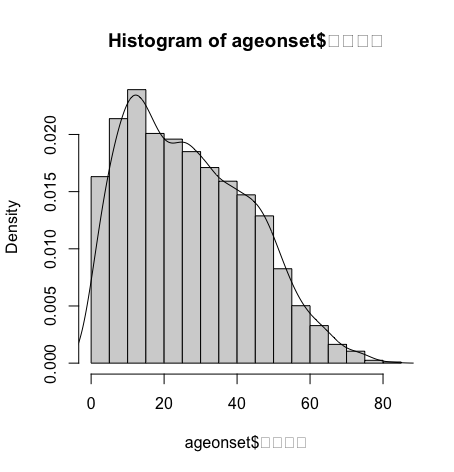
不好意思，我现在只能做出两个分曲线图。

整体的曲线图我一直查不到叫什么名字（英文）所以暂时没有头绪。

如果您可以给我一点类似的信息／线索，我可以再次尝试。

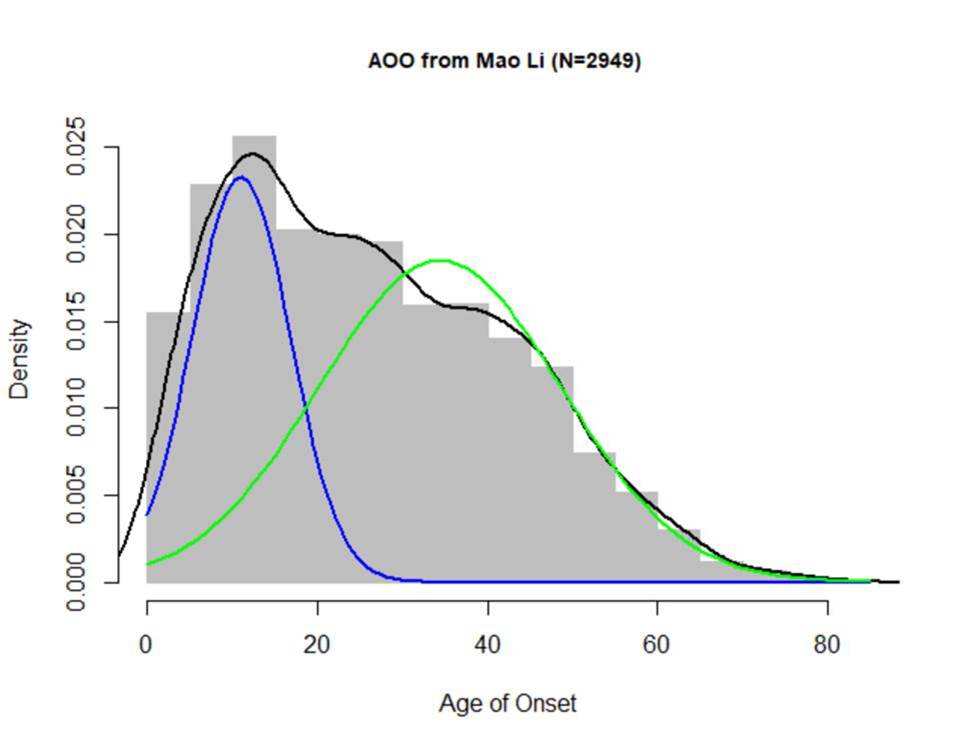


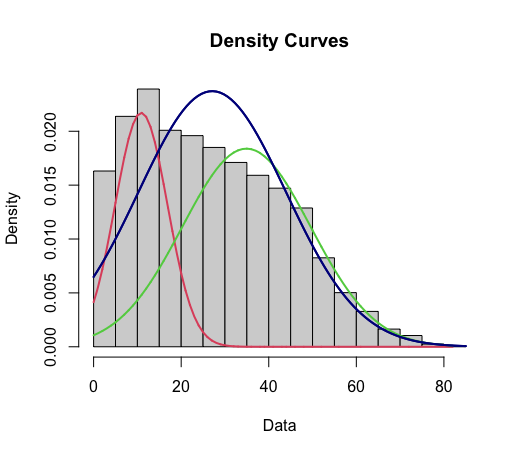
您上一次给我的文件“Analysis of Vitiligo Age-Of-Onset.docx”里面有一张图如下：



能否把这张图和双曲线图合并？

之前有老师帮我们用前期的3000例数据做出如下图，但现在不方便问她如何做的了。





上次文件的图和PDF里面的其实不是同一种绘图。

PDF论文里的是绘制模拟数据分布的图

上次文件的图只是把直方图（histogram）最中间点连起来，和之前3000数据做的是一种

二、发病年龄随年度变化

1. 现有数据是否能分析出2个分界点？如果不行的话，是不是因为80年以前的数据不够？

1）没有数据的年份：1953-1961， 1963-1969，1974，1976-1978

2）拥有一个数据的年份：1952，1962, 1973，1975，1979，1983，1984，1991

3）拥有两个数据的年份: 1970, 1982 , 1984

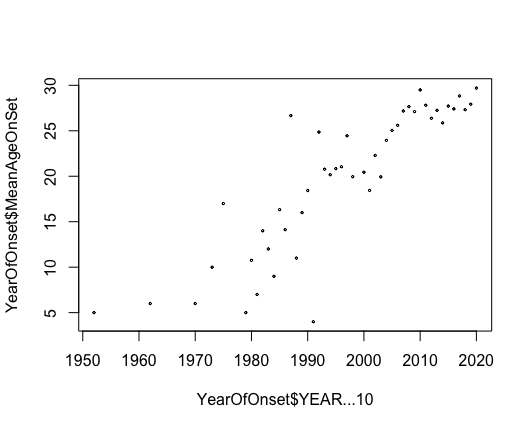
4) Central Limit Theorem(中心极限定理): distribution of sample means approximates a normal distribution as the sample size gets larger. Sample size equal to greater than 30 are considered sufficient for CLT to hold. (在统计学中一般默认抽取样本的数量要大于等于30, 数据会呈现正态分布)

综上，1990年以前的数据都非常缺乏，平均发病年龄在数据样本量（当年的发病年龄数据样本量)大于30才有普遍代表意义（simple, random sample). (也就是说至少要有当年30个病人的发病年龄然后再计算当年平均发病年龄才有代表意义)

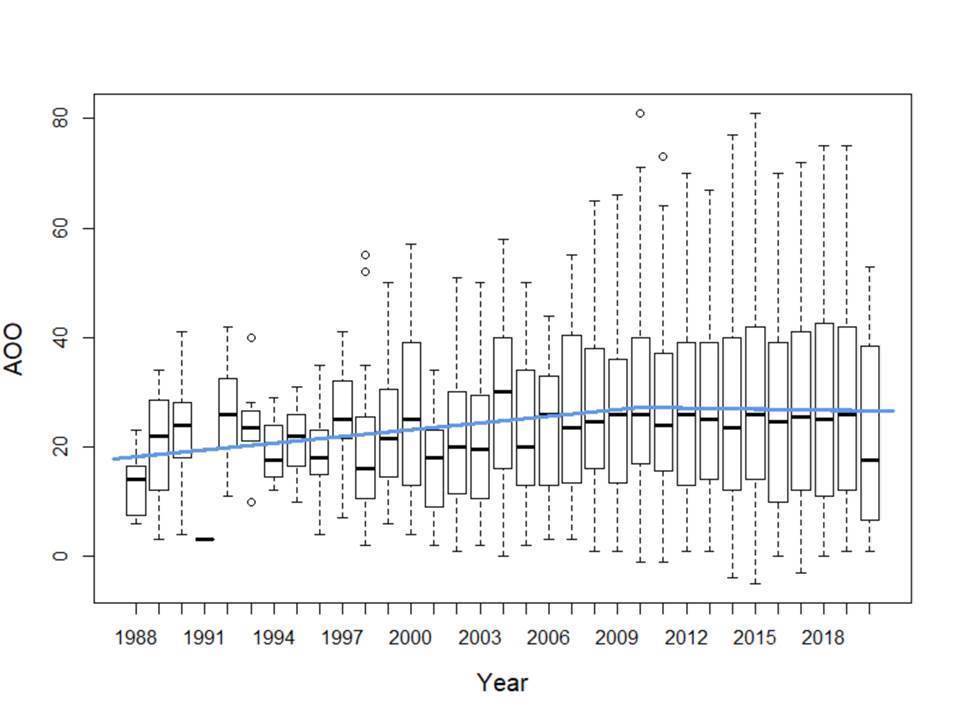
在所给的数据中，基本只有2000年以后的平均发病年龄才能满足这个要求。某些年份只有一份数据，只能按照唯一的发病年龄算作当年的平均发病年龄（参见2,），这其实**是不能满足在统计学中做回归曲线的前提的**，因为我们不能保证：

* 样本是随机抽取的（did we use simple random sample?）
* 样本之间是独立且没有关联性的（are the subject independent?）

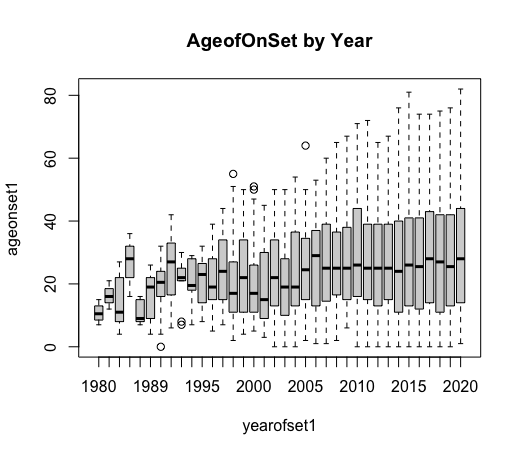
参见scatter plot, 我们可以明显看出和2000年以后的数据相比，1990年以前的平均发病年龄是不规律分布的，因为很多这些平均发病年龄是没有代表意义的。尤其某些特殊的年份。（举例：比如只有一份发病年龄数据的年份， 1991）



我知道早期的年份很多数据不足，只能确定2002年这个分界点也可以。能否这样处理：剔除只有1、2个数据的年份，剩下的年度数据做成下图类似的趋势线和箱状图，并计算各段slope，SE和P值？（这是以前的老师用3000例数据做的）



* The breakpoint is estimated at 2010 with SE = 2.6.
* The slope on the 1988 – 2010 segment is 0.4063 (SE=0.1271, P= 0.00140)
* The slope on the 2010 – 2020 segment is not significantly different from 0: estimate=-0.083745, SE=0.12544, P=0.5044224)



平均发病年龄

Group.1 x

1 1980 10.75000

2 1985 16.33333

3 1986 14.12500

4 1987 26.66667

5 1988 11.00000

6 1989 16.00000

7 1990 18.42857

8 1992 24.85714

9 1993 20.77778

10 1994 20.16667

11 1995 20.83333

12 1996 21.04762

13 1997 24.44828

14 1998 19.95238

15 1999 23.66667

16 2000 20.44444

17 2001 18.40000

18 2002 22.28571

19 2003 19.93103

20 2004 23.94286

21 2005 25.04167

22 2006 25.60000

23 2007 27.17647

24 2008 27.65217

25 2009 27.11236

26 2010 29.49306

27 2011 27.81022

28 2012 26.37674

29 2013 27.25311

30 2014 25.86851

31 2015 27.71655

32 2016 27.40577

33 2017 28.83177

34 2018 27.31933

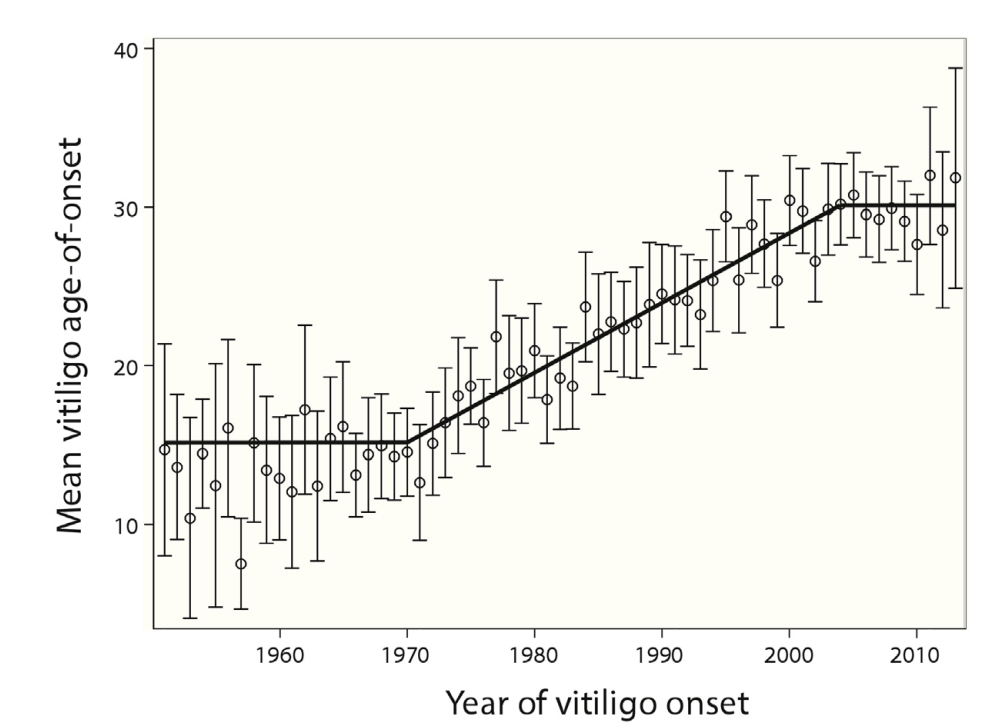
35 2019 27.92387

36 2020 29.69733

|  |  |  |  |
| --- | --- | --- | --- |
|  | Intercept | SE | P |
| 1980-2002 | 0.28387 | 0.09706 | 0.0132 |
| 2002-2020 | -0.31225 | 0.65512 | 0.0118 |

*请注意！： 按照这个方法做出来的趋势线 不等于 回归曲线*

2、能否按下图表示出每年的SD和变化折线？



因为无法满足做回归的前提，而且我也不知道这个图的英文名，所以这个图也没无法画出来。每条竖线代表的不是SD而是 [ 95% confidence interval ], 如果您需要指定年段的95% confidence interval，我可以帮您用软件计算出来（比如有实际意义的2000年以后的数据）。

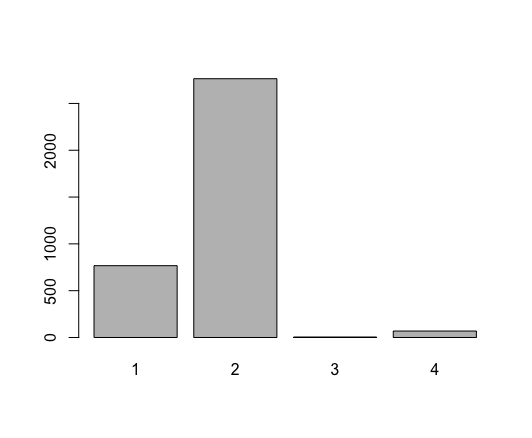
能否软件画上图的95% confidence interval，然后手动添加趋势线？

抱歉，这个我还真不会，我只会求解95% CI，

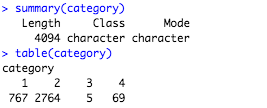
那就请帮我把有意义的年份的95%CI算一下吧，谢谢！

三、其他分析

1、发病类型比例随年度变化趋势

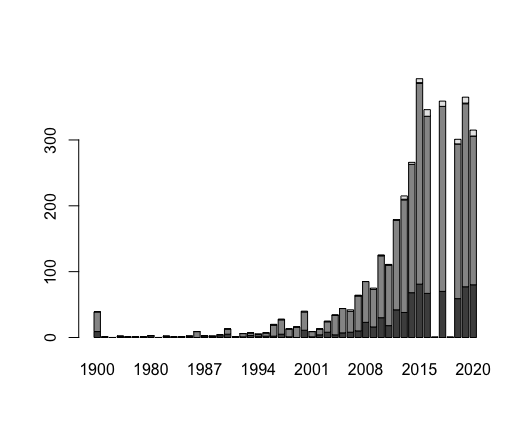
发病类型即表格中的西医分型，主要想看看各分型的比例（主要是1型和2型）是否随年度有变化规律？

**各分型的比例（主要是1型和2型）:**

****

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |  |
| Count | 767 | 2764 | 5 | 69 | Total: 4094 |
| Percentage | 18.73% | 67.51% | 0.12% | 1.68% | Total: 100% |

**是否随年度有变化规律？**



我想看的是每年的1型和2型所占当年发病病例的百分比是否随年度有变化规律。能否把每年的1型和2型的百分比做个图？

上图就是表示1型和2型所占当年发病病例的百分比随年份的变化规律，我不清楚您要什么样的图，饼图还是？？可以再具体一些。

其实很简单，就做一个1型占当年的百分比随时间变化的柱状图即可

X轴为时间，Y轴为1型的比例

列个表格每年的百分比：

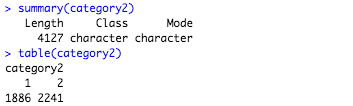
|  |  |  |
| --- | --- | --- |
|  | **1** | **2** |
| **1952** |  | **100.00%** |
| **1953** |  |  |
| **1954** |  |  |
| **1955** |  |  |
| **1956** |  |  |
| **1957** |  |  |
| **1958** |  |  |
| **1959** |  |  |
| **1960** |  |  |
| **1961** |  |  |
| **1962** |  |  |
| **1963** |  |  |
| **1964** |  |  |
| **1965** |  |  |
| **1966** |  |  |
| **1967** |  |  |
| **1968** |  |  |
| **1969** |  |  |
| **1970** |  | **100.00%** |
| **1971** |  |  |
| **1972** |  |  |
| **1973** |  | **100.00%** |
| **1974** |  |  |
| **1975** |  | **100.00%** |
| **1976** |  |  |
| **1977** |  |  |
| **1978** |  |  |
| **1979** |  | **100.00%** |
| **1980** | **33.00%** | **67.00%** |
| **1981** |  |  |
| **1982** |  | **100.00%** |
| **1983** |  | **100.00%** |
| **1984** |  | **100.00%** |
| **1985** | **100.00%** |  |
| **1986** |  | **100.00%** |
| **1987** | **67.00%** | **33.00%** |
| **1988** |  | **100.00%** |
| **1989** | **25.00%** | **75.00%** |
| **1990** | **38.46%** | **61.54%** |
| **1991** |  | **100.00%** |
| **1992** | **16.67%** | **83.33%** |
| **1993** |  |  |
| **1994** | **40.00%** | **60.00%** |
| **1995** | **28.57%** | **71.43%** |
| **1996** | **10.00%** | **90.00%** |
| **1997** | **17.86%** | **82.14%** |
| **1998** | **7.69%** | **92.31%** |
| **1999** |  | **100.00%** |
| **2000** | **27.50%** | **45.00%** |
| **2001** | **11.11%** | **88.89%** |
| **2002** | **30.77%** | **69.23%** |
| **2003** | **32.00%** | **64.00%** |
| **2004** | **11.76%** | **88.24%** |
| **2005** | **14.29%** | **75.51%** |
| **2006** | **19.05%** | **76.19%** |
| **2007** | **15.63%** | **82.81%** |
| **2008** | **27.06%** | **61.18%** |
| **2009** | **21.33%** | **76.00%** |
| **2010** | **24.00%** | **75.20%** |
| **2011** | **16.22%** | **82.88%** |
| **2012** | **23.46%** | **75.98%** |
| **2013** | **17.67%** | **79.53%** |
| **2014** | **25.56%** | **73.31%** |
| **2015** | **20.61%** | **77.61%** |
| **2016** | **19.31%** | **77.52%** |
| **2017** | **19.44%** | **78.06%** |
| **2018** | **19.60%** | **78.07%** |
| **2019** | **21.10%** | **76.16%** |
| **2020** | **25.40%** | **71.75%** |

2、初发部位（暴露/非暴露部位）比例随年度变化趋势

表格中“基本信息\_初发部位”这部分来区分暴露部位和非暴露部位，其中“头、面、颈、手”属于暴露部位，其余属于非暴露部位。需要处理一下原始数据。看暴露部位/非暴露部位的比例有没有随年度变化的规律？

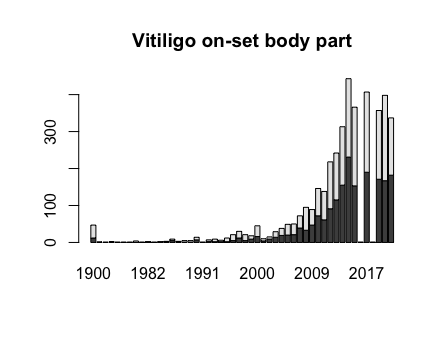
* Value contains “头、面、颈、手” = 1
* Value does not contain “头、面、颈、手”= 2





|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 |  |
| Count | 1886 | 2241 | Total:4127 |
| Percentage | 45.69% | 54.30% | Total:99.99% |

看暴露部位/非暴露部位的比例有没有随年度变化的规律？



我想看的是每年的曝光部位和非曝光部位所占当年发病病例的百分比是否随年度有变化规律。能否把每年的曝光部位和非曝光部位的百分比做个图？

上图就是表示每年的曝光部位和非曝光部位的百分比随年份的变化规律，我不清楚您要什么样的图，饼图还是？？可以再具体一些。

其实很简单，就做一个曝光部位占当年的百分比随时间变化的柱状图即可

X轴为时间，Y轴为曝光部位的比例

列个表格每年的百分比：

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1952 | 100.00% |  |
| 1953 |  | 100.00% |
| 1954 |  |  |
| 1955 |  |  |
| 1956 |  |  |
| 1957 |  |  |
| 1958 |  |  |
| 1959 |  |  |
| 1960 |  |  |
| 1961 |  |  |
| 1962 |  |  |
| 1963 |  |  |
| 1964 |  |  |
| 1965 |  |  |
| 1966 |  |  |
| 1967 |  |  |
| 1968 |  |  |
| 1969 |  |  |
| 1970 | 100.00% |  |
| 1971 |  |  |
| 1972 |  |  |
| 1973 |  | 100.00% |
| 1974 |  |  |
| 1975 |  | 100.00% |
| 1976 |  |  |
| 1977 |  |  |
| 1978 |  |  |
| 1979 |  | 100.00% |
| 1980 | 25.00% | 75.00% |
| 1981 |  | 100.00% |
| 1982 | 50.00% | 50.00% |
| 1983 |  | 100.00% |
| 1984 | 50.00% | 50.00% |
| 1985 | 66.67% | 33.33% |
| 1986 | 55.56% | 44.44% |
| 1987 | 66.67% | 33.33% |
| 1988 | 20.00% | 80.00% |
| 1989 | 20.00% | 80.00% |
| 1990 | 50.00% | 50.00% |
| 1991 |  | 100.00% |
| 1992 | 42.86% | 57.14% |
| 1993 | 22.22% | 77.78% |
| 1994 | 33.33% | 66.67% |
| 1995 | 16.67% | 83.33% |
| 1996 | 23.81% | 28.57% |
| 1997 | 40.00% | 60.00% |
| 1998 | 23.81% | 76.19% |
| 1999 | 50.00% | 50.00% |
| 2000 | 35.56% | 64.44% |
| 2001 | 40.00% | 60.00% |
| 2002 | 60.00% | 40.00% |
| 2003 | 48.28% | 51.72% |
| 2004 | 50.00% | 50.00% |
| 2005 | 40.82% | 59.18% |
| 2006 | 42.00% | 58.00% |
| 2007 | 54.17% | 45.83% |
| 2008 | 34.74% | 65.26% |
| 2009 | 52.81% | 47.19% |
| 2010 | 49.32% | 50.68% |
| 2011 | 44.20% | 55.80% |
| 2012 | 41.74% | 58.26% |
| 2013 | 47.52% | 52.48% |
| 2014 | 49.52% | 50.48% |
| 2015 | 52.14% | 47.86% |
| 2016 | 41.69% | 58.31% |
| 2017 | 46.57% | 53.43% |
| 2018 | 47.90% | 52.10% |
| 2019 | 41.96% | 58.04% |
| 2020 | 54.01% | 45.99% |