Technologies for Data Analysis Part II

Homework

due 8 June 2014

1 Nobel laureate age

The file Nobel_laureates_1940s_births.xlsx contains information about all awardees of the Nobel Prize in Physiology or Medicine who were born in the 1940s. For each of them, the file lists their year of birth and the age (in years) at which they were awarded the Prize.

Summary statistics

Compute and show some summary statistics. Is the mean or median a more appropriate measure?

Median: 55

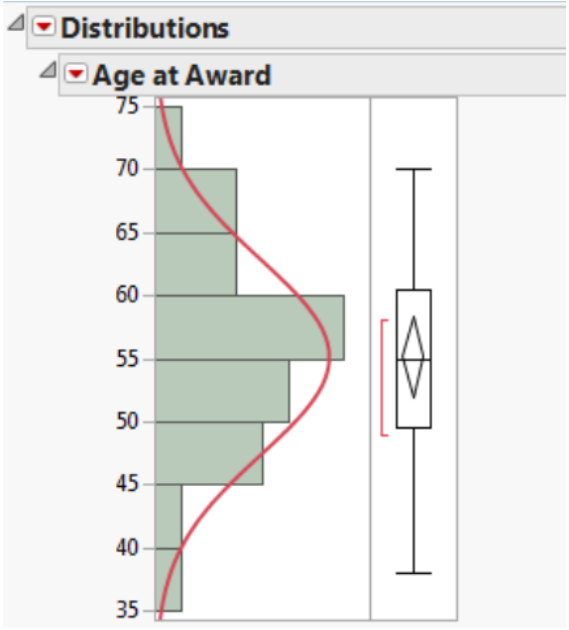
Mean: 55.16

Standard deviation: 7.73

The mean and median are very close and the data seems roughly normally distributed (see below). Both the mean and the median are appropriate choices.

Is your data normal?

Visualise the distribution of age at award. Does your data look normal?



The data looks normally distributed: The distribution is roughly bell-shaped. Mean and median almost coincide. Most of the data points come from the center of the distribution. There are enough data points to make the call.

Are Nobel laureates getting younger?

A similar analysis on Nobel laureates in Physiology or Medicine born in the 1960s has revealed an average age at award of 47 years with a standard deviation of 1.4. Editors at a large daily newspaper plan to run a story about this ("Are Nobel Laureates getting younger?") and ring you up for your expert opinion.

If you were to do a hypothesis test,

- **What is your hunch the test will show?**

  This is of course a subjective question. It looks at first glance as if Nobel Laureates are indeed getting younger (but see below for caveats!)

- **Would you choose a non-parametric or a parametric test?**

  The roughly normal distribution of ages among laureates born in the 1940s looks like a parametric test might be in order. But without seeing the distribution of laureates born in the 1960s, it is impossible to tell.

- **What is your Null Hypothesis?**

  That there is no difference in age at award between the two groups.

- **Would you choose a one- or two-tailed test?**

  Since the question is specifically whether they are getting younger, a one-tailed test would probably be most appropriate.

- **What significance level would you choose?**

  Again, this is open to debate. All of you choose 5 %, as is the convention in biology, and this seems like a reasonable choice.

- **What more information do you need?**

  At least the sample size and the shape of the distribution for the laureates born in the 1960. Better still, the entire dataset.

- **What might be a problem with this approach and how might you fix it?**

  This is a similar example to what we saw in class (the study on children hospitalised for a specific disease) where our Null Hypothesis is that two datasets are the same, but the method of data collection itself means that the datasets are different. To quote Vivian:

  > "...there simply has not been enough time between the 60s and now for there to be older laureates born in the 60s. i.e. perhaps there are scientists born in the 60s who will be awarded a Nobel in the future (say in the 2020s or 2030s), but this simply has not happened yet. The simple fact that only about 50 years has passed between the birth of laureates in the 60s, and the current year (2014) means that it is not possible for the mean to be any higher than $\sim$50."

  What could be done about this? One option would be to look not by decade of birth, but by decade of award. (I.e. how old were Nobel Laureates in the 2000s, compared to, for instance, the 1980s). This would eliminate the structural inequality in the two datasets and make an actual comparison possible.