The $t$-statistic is a useful test statistic for questions regarding the mean of a population.
The basic assumption on the population is:

- $X_1, X_2, \ldots, X_n$ are a random sample from a normal population with mean $\mu$ and standard deviation $\sigma$.

Then we know that $\bar{x}$ has a normal distribution with the same $\mu$ for its mean, but with a standard deviation of $\sigma/\sqrt{n}$. Thus

$$Z = \frac{\bar{x} - \mu}{\sigma}\sqrt{n}$$

has a standard normal distribution.

To test for $\mu$, we could use this **if** we knew the true value of $\sigma$, which is unlikely, as we don't even know the true value of the population mean $\mu$. So we replace $\sigma$ with the sample standard deviation $s$:

$$T = \frac{\bar{x} - \mu}{s}\sqrt{n} = Z\frac{s}{\sigma}.$$

What then is the sampling distribution of $T$? It clearly isn't normal unless $s$, which is random, is equal to $\sigma$ which isn't random!

In 1908 Gossett characterized this distribution – mathematically and with a simulation. Here we will look at a simulation.

# 1 Simulation of $T$

We wish to focus on the differences between $Z$ and $T$. A simulation will maybe help.

First, run this command to download the simulation:

```
> source("http://www.math.csi.cuny.edu/verzani/classes/MTH214/sim-T.R")
```

The graphic shows the $T$ and $Z$ statistics for the same sample of size $n$. The $t$'s are red triangles, the $z$'s are black circles. The dashed line connecting statistics for the same sample. The difference is due to $s$ being different from $\sigma$. When $s < \sigma$ the line is red and the $t$ points contract towards 0, the $s > \sigma$ the line is blue and the $t$ points are bigger than the corresponding $z$.

There are $m$-samples represented in each figure. The boxplot summarizes the $T$ and $Z$ sampling distributions. By looking at the boxplots and the samples we can understand how the $t$ is both similar and different from the normal distribution.

1. When $n = 2$ is small, use a number of difference simulations to decide:

   (a) Are both symetric or is one skewed? (They are unimodal, but you can't tell from a boxplot.)

   (b) Do they have the same center?

   (c) Do they have the same spread?

   (d) Do they have the same tails?

2. Repeat with $n = 10$

3. Repeat with $n = 30$

4. Repeat with $n = 100$

5. Some books say that after $n = 30$ there is no difference, others use $n = 100$. From you simulation studies, what do you think?

6. The boxplot is not the best graphic to look at differences in distributions. What graphics might be better?

## 2 The t-test

Now load `pmg` and the built-in data set `Cars93`. Next, open the `Dynamic tests` dialog under the `tests` menu and select the `1-sample t-test`.

This dialog allows you do perform a t-test. In general performaing a significance test involves:

- specifying a null and alternative hypothesis;
- specifying a test statistic and verifying the population assumptions that allow the
- computing of a $p$-value.

We are focusing here on the one-sample $t$-test, so we've got the test statistic. In the dialog, one drags a variable to the `data:` line, selects the null hypothesis by clicking on the bold value beside the "not equal to" popup and changing the value. (Hit enter to finalize). Then adjust the alternative hypothesis by adjusting the popup.

For example, to test

$$H_0 : \mu = 2.5 \quad H_A : \mu > 2.5$$

For the Engine Size, we drag the EngineSize variable to the area, punch in 2.5 and then select "greater than." The p-value is found to be 0.0611, after rounding.

Easy? I hope so. This is the mechanical part of the class.

1. Is city mileage better than 20MPH? Do the test

   $$H_0 : \mu = 20 \quad H_A : \mu > 20$$

   Report the $p$-value and whether the difference is statistically significant at the 0.05 level.

2. Is highway mileage less than 30MPH? Do the test

   $$H_0 : \mu = 30 \quad H_A : \mu \leq 30$$

   Report the $p$-value and indicate if you would reject the null hypothesis at significance level of 0.05.

3. Is the mean vehicle weight different from 3000 pounds?

   $$H_0 : \mu = 3000 \quad H_A : \mu \neq 3000$$

   Report the $p$-value and indicate if you would reject the null hypothesis at the $\alpha = 0.05$ level.

4. Find a 95% confidence interval for the mean rear seat room.

5. For each of the 4 variables above, discuss how you checked the population assumptions needed so that the $p$-value is computed accurately.