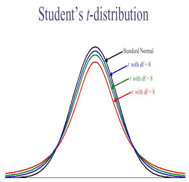

Sampling distribution

The key to statistical inference is knowing the sampling distribution of some statistic. For sample proportions, $\hat{p}$, we have from the normal approximation to the binomial the following three facts:

1. The center, or expected value, of $\hat{p}$ is p, the population parameter
2. The spread, or standard deviation, of $\hat{p}$ is $\sqrt{p(1-p)/n}$. This is the population standard deviation divided by $\sqrt{n}$
3. The shape, or distribution, of the sampling distribution of $\hat{p}$ is normal. This allows us to use the normal tables to compute probabilities.

Is the same sort of thing true for $\bar{x}$? We'll see in this project, the answer is yes, and no.

For instance, Figure 1 shows the density of a bimodal population. Thirty five random samples of size 10 are shown as gray diamonds. For each sample, the sample mean is plotted as a black triangle at the bottom of the graph. For these sample means a scaled density estimate is shown. By looking at many samples we can see clearly that the sample mean has a distribution that is different in shape and scale from the population mean.

Question 1: Is the center of the population, the same as the center of the sampling distribution?

Question 2: Is the spread of the population, the same as the spread of the sampling distribution?

Question 3: Is the shape of the population, the same as the shape of the sampling distribution?

For $\bar{x}$ we will see three things:

1. The center of $\bar{x}$ is also the population center, μ.
2. The spread of $\bar{x}$ is $\sigma/\sqrt{n}$, where σ is the spread (standard deviation) of the population
3. The shape of $\bar{x}$, *for large enough* n is the normal distribution

As such, the z-scores for $\bar{x}$

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

Should be a standard normal for large enough n.

To investigate, we use a function that first needs to be downloaded:

```
> source("http://www.math.csi.cuny.edu/verzani/R/make.z.R")
```

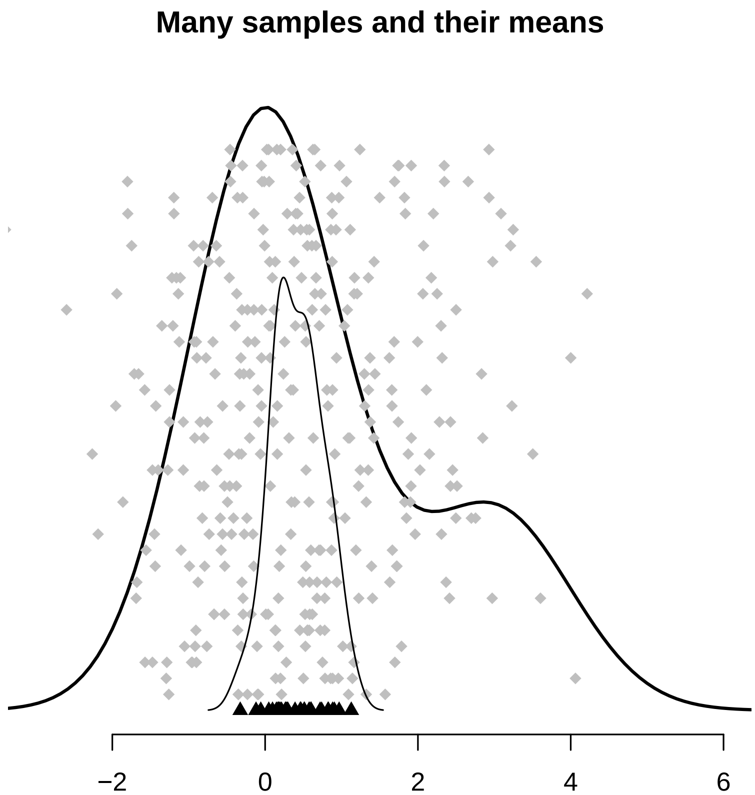

Figure 1: An illustration of many samples taken from the bimodal distribution. Each row of diamonds is a new sample. The triangles at the bottom are the sample means, and the smaller density is a density estimate for the sample means shrunk to fit the diagram.

This downloads two function, `make.z()` and `make.t()`. These create random sample of the scaled statistic for different values of n.

Question 4: Run the command `make.z(1)`. What is output? 500 samples of Z for $n = 1$. How would you want to summarize this data?

Question 5: Run the command `make.z(1)` and store the results into a variable `res`. Make a histogram, a boxplot, and a density plot of `res`. (The latter is done with `plot(density(res))`.)

Does this data look "bell shaped," Does 95% of the data look like it is between -2 and 2?

Question 6: Run the command `make.z(10)` and store the results into a variable `res`. Make a histogram, a boxplot, and a density plot of `res`. (The latter is done with `plot(density(res))`.)

Does this data look "bell shaped," Does 95% of the data look like it is between -2 and

2?

(When the population is normal, like this one, Z always has the same shape.)

When we have different populations, the story is different. Only as n gets big enough do we get a bell shape.

Question 7: The command `make.z(1, family="exp")` will show a different population. Run this command, store the results and make the three graphs above. Does the data look "bell-shaped?"

Question 8: The command `make.z(10, family="exp")` will show a different population. Run this command, store the results and make the three graphs above. Does the data look "bell-shaped?"

Question 9: The command `make.z(100, family="exp")` will show a different population. Run this command, store the results and make the three graphs above. Does the data look "bell-shaped?"

In summary, as n gets large the sampling distribution of $\bar{x}$ becomes normal. If the population is normal, then $n = 1$ is large. When the population is not normal, large values of n are needed.

1 The T statistic

The T statistic uses s instead of σ in the denominator

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Does this make any difference? By subtracting μ we have a center of 0, but when we divide by $s/\sqrt{n}$ (the standard error), instead of $\sigma/\sqrt{n}$ (the standard deviation) the spread is different, and also then the shape.

How different? The `make.t()` function can show us.

First, when the population is normal, like the default usage.

Question 10: The following command will produce 3 boxplots for different values of n. Is there a difference? Explain what it is.

```
> boxplot(make.t(2), make.t(5), make.t(10))
```

Question 11: Repeat the above, only use values of 20, 50 and 100. Are the differences so dramatic?

Question 12: Let's see that T and Z are indeed different. This command will produce two boxplots for a value of $n = 4$

```
> boxplot(make.t(4), make.z(4))
```

What are the differences?

Question 13: Do the differences go away? Make boxplots comparing T and Z for $n = 10, 20, 30, 50$ Is there always a difference? In the book we use a t table, but after $n = 30$ (29 degrees of freedom), it uses the same set of numbers. Explain why this is possible.

Question 14: Is this summary correct: For a normal population and small n (less than 30) the distribution of T and Z are different, as the tails of T are longer. For larger n, the distributions are approximately the same. Does the population affect the distribution of T? We say above that as n gets large, the distribution of Z becomes bell shaped. Is the same true for T? What about when n is small, is the distribution of T the same as when the population is normal?

Question 15: The command

```
> boxplot(make.t(5, family = "exp"), make.t(5))
```

will compare distributions for a skewed population (the left boxplot) and a normal population. Make the graph. Do the populations look different? How so.

Question 16: Repeat the above using $n = 10, 20, 30, 50$, and 100. Do the two boxplots ever look like they come from the same sampling distribution?