Null Distributions

Introduction

It's worth taking some time again to consider what is meant by the statement "My variable $\overline{X}$ has a normal distribution with mean value of μ and standard deviation (standard error when sampling distribution) of $\sigma/\sqrt{n}$." This statement can also be expressed notationally as

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

In short, it implies a few things:

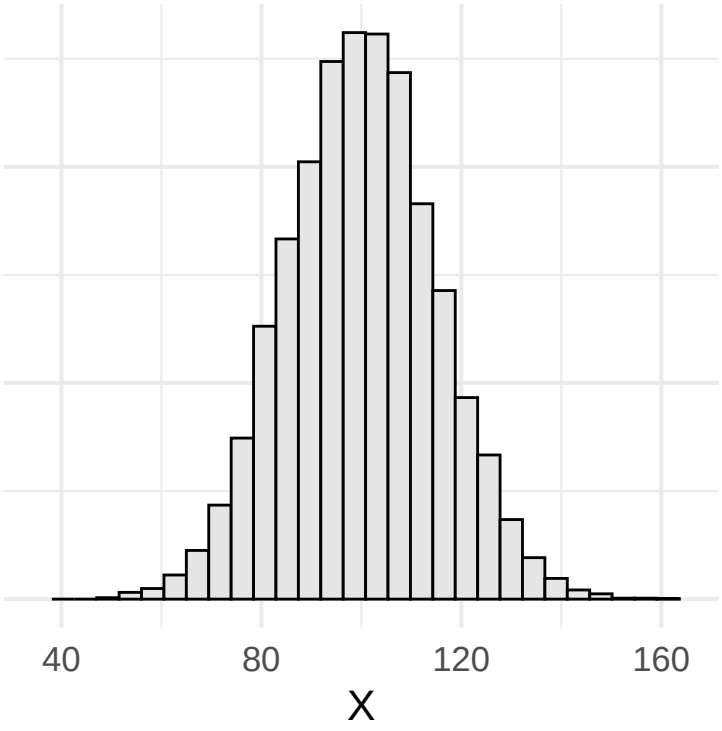
1. First, it implies that $\overline{X}$ follows a random process, meaning the process of collecting a sample and computing the value of $\overline{X}$ has a degree of randomness that inhibits me from knowing precisely what this value may be
2. Second, it implies that I understand structurally *how* this random process will behave. Because it is normally distributed, I know that values near μ will be more common than values further from μ, and from $\sigma/\sqrt{n}$, I have a sense of how tightly (or widely) I can expect them to be, in general. Are they all very close? Are they within $\pm$ 10? These are things I can determine from the standard error.

For example, assume that $\overline{X} \sim N(100, 15)$. We can simulate what potential values of $\overline{X}$ could be by drawing a couple of values from this distribution in R

```
## rnorm == Random NORMal
rnorm(n = 5, mean = 100, sd = 15)
```

```
## [1]  90.010  66.285  96.330 100.687 124.248
```

This is what is meant by our variable being a *random process*. Now, when I say that we understand structurally how it will look, what I mean is if we were to continue sampling from this distribution, I could create a histogram to visually represent what the distribution looks like

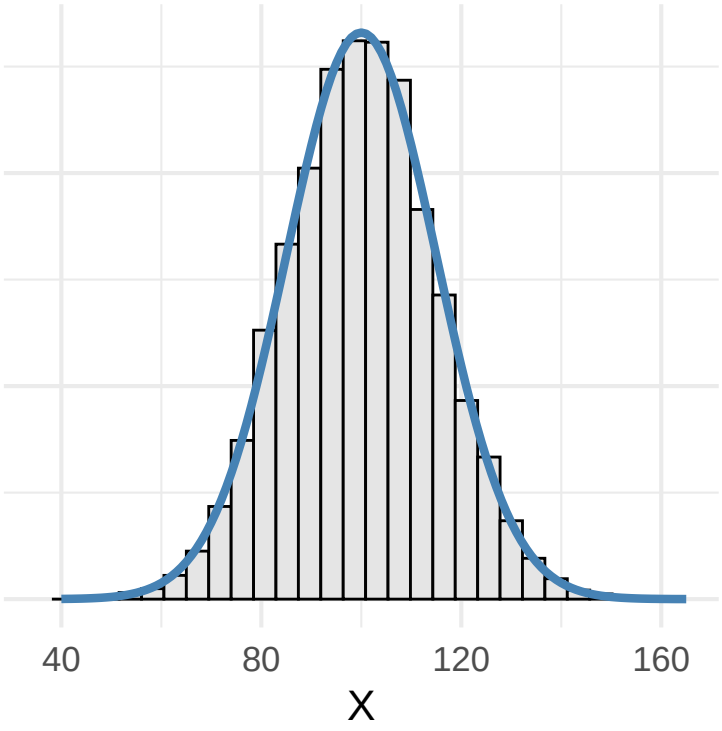


This is known as an *empirical distribution* since it is a distribution constructed from empirical evidence. If all I had was data and nothing else, this is what I would be left with. If I wanted to find the quantiles of this data, I would have to take the quantiles from the data directly

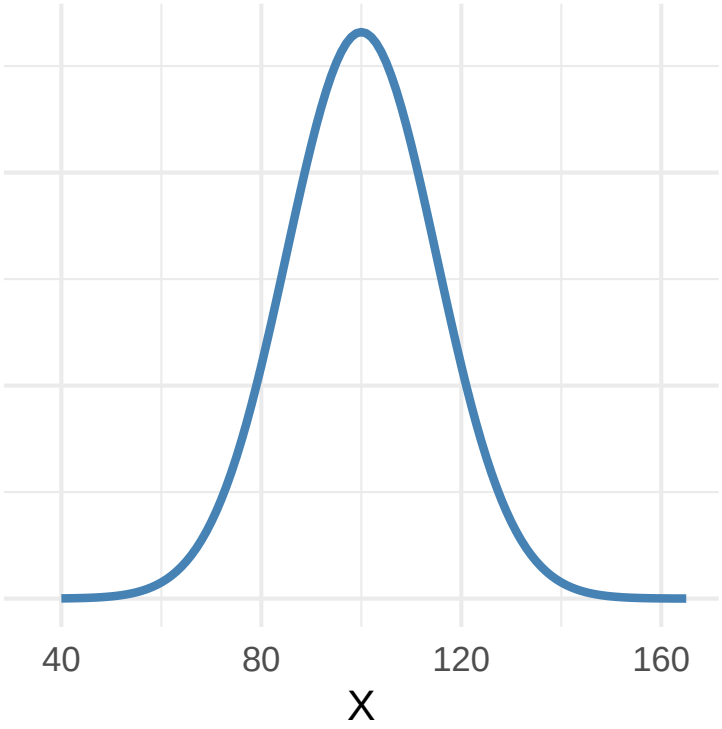
```
quantile(x, probs = c(0.1, 0.9))
```

```
##     10%     90%
##  80.878 119.283
```

Now, because of the Central Limit Theorem, I know that this follows a *theoretical distribution*, namely, the normal distribution. This means that, instead of using my data directly, I can estimate my distributional parameters, μ and $\sigma/\sqrt{n}$. From this, I can construct my theoretical distribution which, as we can see, lines up nicely with what we found empirically



Once we know that our theoretical distribution is a good match, we can do away with all of the data itself and *only* deal with the theoretical distribution

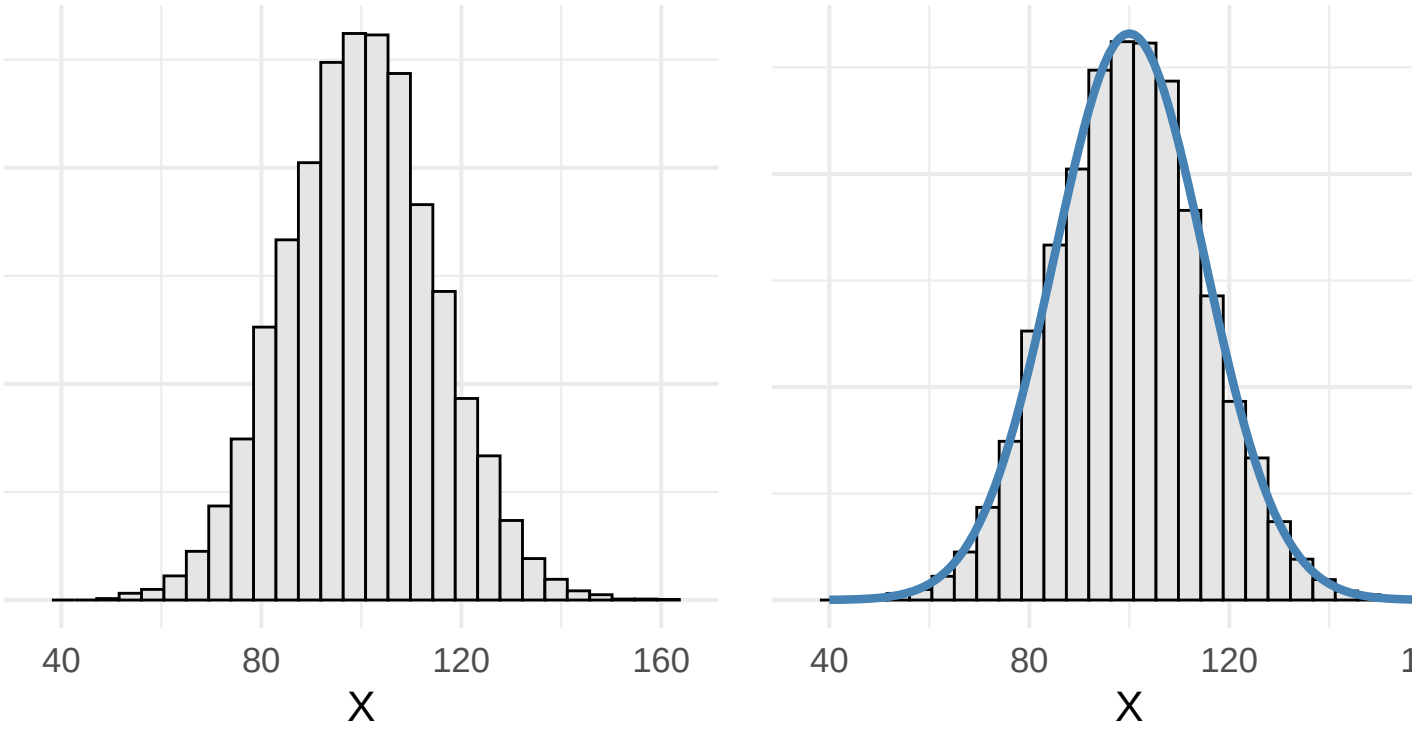


To find the quantiles of a theoretical distribution, I can use my `qnorm` function, which should match the values I found empirically

```r
qnorm(c(0.1, 0.9), mean = mean(x), sd = sd(x))
```

```
## [1]  80.849 119.224
```

Together, the process from empirical looks like this:



In practice, however, we only have a single $\overline{X}$, drawn from our empirical distribution. We know that, were we to continue sampling, we could construct a histogram like the one on the left above from which we could compute our quantiles directly. As we are left with only a single sample, however, we instead utilize the CLT, giving us a theoretical distribution from which we can instead perform our calculations

Most of what we will be discussing will be in this context: we have a single value, say, $\overline{x}$ or $\hat{p}$ or others, and we have to decide what exactly the distribution it follows should be. As we will rarely know the true value μ, we will instead have to try candidate values. The purpose of this lab will be to illustrate how different candidates for μ, denoted μ_0 change both our sampling distribution and how our value $\overline{X}$ relates to it.

Walkthrough Together

Suppose that we know for certain that $\mu = 50$ and $\sigma/\sqrt{n} = 5$, with

$$\overline{X} \sim N(50, 5)$$

Part 1 Draw the sampling distribution for $\overline{X}$ with tick marks at μ as well as tick marks at each standard deviation

Part 2 Write out the sampling distribution for $\overline{X} - \mu$. Draw the sampling distribution, again with tick marks at the mean and at each standard deviation

Part 3 Write out the sampling distribution for $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. Draw the sampling distribution, again with tick marks at the mean and at each standard deviation.

Part 4 Assume that we have collected a sample and found that $\overline{x} = 45$. Indicate where on each of the sampling distributions above this value falls, first with $\overline{x}$, then with $\overline{x} - 50$ and finally with $(\overline{x} - 50)/5$. What does the value represent in the graph in Part 3?

Part 5 Suppose now that we did not actually know that $\mu = 50$ and instead, we hypothesized that the true value of the mean was $\mu_0 = 55$. This is our *null hypothesis*. Under the null, write out the hypothesized sampling distribution of $\overline{X}$ and, just as in Part 1, draw the sampling distribution, labeling the appropriate marks. On top of this, in a dashed line, draw again the true sampling distribution.

Part 6 Repeat part two, this time writing out and drawing the sampling distribution for $\overline{X} - \mu_0$. Again use a dashed line to draw the original sampling distribution, $\overline{X} - \mu$.

Part 7 Finally, repeat part 3, diving by the standard error.

Part 8 Just as we did in Part 4, label in each stage where our observed $\overline{x} = 45$ would fall. In the final plot from Part 7, what does this value represent? Has $\overline{x}$ changed? In this case, what does the t-statistic represent?

Part 9 What would happen to our test statistic if we repeated this with $\mu_0 = 47.5$?

Question 1

Suppose that we sample from a population where the true mean is $\mu = 100$. We do not know σ, but we have estimated that $\hat{\sigma} = 10$. Further, suppose that we have collected a sample of size $n = 25$ and found that $\overline{x} = 97$.

Part A Suppose that we hypothesize that $\mu_0 = 95$. Draw the sampling distribution for $\overline{X}$ creating tick marks and labeling the mean and each standard deviation. Indicate where $\overline{x}$ falls in this distribution

Part B Draw now the sampling distribution for $\overline{X} - \mu_0$. Again, label the ticks and mark the location of $\overline{x} - \mu_0$

Part C Draw the sampling distribution for $\frac{\overline{X}-\mu}{\hat{\sigma}/\sqrt{n}}$. Again, mark where our observed data falls.

Part D What distribution does our statistic follow in Part C? Look at your table of critical values for the appropriate distribution. Does our observed test statistic fall within 95% of the expected values when the null hypothesis is true?

Question 2

Repeat Question 1, continuing to assume that $\mu_0 = 95$ and that $\overline{x} = 97$, but this time assume that our sample size is $n = 100$

Part A Write out the sampling distribution of $\overline{X}$ and draw it, labeling the ticks as necessary. Mark $\overline{x}$

Part B Draw now the sampling distribution for $\overline{X} - \mu_0$. Again, label the ticks and mark the location of $\overline{x} - \mu_0$

Part C Draw the sampling distribution for $\frac{\overline{X}-\mu}{\hat{\sigma}/\sqrt{n}}$. Again, mark where our observed data falls.

Part D Using your table of critical values, does our observed test statistic fall within 95% of the expected values when the null distribution is true? (For $n \geq 40$ you may use the standard normal as your approximation).

Question 3

Consider the results you found in Question 1 and Question 2. Explain in words how $\overline{x}$ and μ_0 were the same for each, yet our conclusions were different. In particular, address:

- How does increasing n impact our certainty of $\overline{x}$?
- What impact does this have on my evidence in asserting that $\mu_0 = 95$?

Question 4

Suppose $\overline{x} = 24$, $\hat{\sigma} = 5$ and $n = 25$. Would $\overline{x}$ be in a 95% CI around $\mu_0 = 20$? What about $\mu_0 = 27$? For which of these hypothesis does our observed data ($\overline{x}$) provide more evidence *against*? Justify your answer

Question 5

Consider two scenarios:

Scenario 1 In a sample with $n = 20$, we find that $\hat{\sigma} = 2$ and $\overline{x} = 15.916$. Our null hypothesis is $\mu_0 = 15$

Scenario 2 In a sample with $n = 40$, we find that $\hat{\sigma} = 5$ and $\overline{x} = 126.62$. Our null hypothesis is $\mu_0 = 125$

From this, answer the following:

Part A: Which of these two scenarios shows the greatest difference between the observed and hypothesized mean?

Part B Which of these scenarios do you think offers more compelling evidence that the null hypothesis is false? To do so, for each scenario, consider the range of values that we would expect 95% of our observations to fall if the null hypothesis were true.