Robustness

The hypothesis testing and confidence interval procedures we've looked at are mathematically justified based on certain *assumptions*.

What if these assumptions don't hold? They almost never will, exactly.

We must rely on the procedures being at least somewhat *robust* — that they work fairly well even when the assumptions aren't quite true.

We have to use our own judgement in deciding whether the assumptions are close enough to being true for our purposes. There are no absolute rules to rely on!

Robustness of t Test Procedures

The one-sample t test and confidence intervals will be correct if either

- The data is exactly normally distributed.
- The number of observations is very large.

Experience shows that they are *approximately* correct when

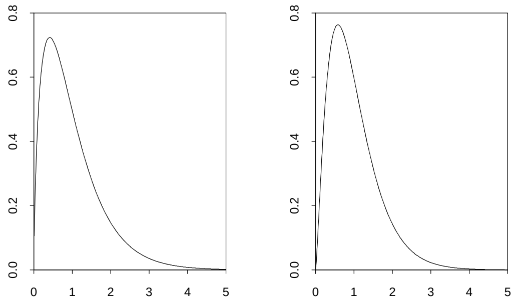
- the observations are close to being normally distributed.
- $n > 15$, and the distribution is moderately non-normal (eg, somewhat skewed).
- $n > 40$ or so, even if the distribution is strongly skewed, or there are some extreme observations (but not really huge ones).

Reliability of two-sample t procedures depends on similar considerations. In addition, it helps if

- The two distributions have similar shapes.
- The two samples are about the same size.

Demonstration of Robustness

Let's look at what happens when we use a two-sample t test (without pooling the variance) for samples from two populations with the following density functions:

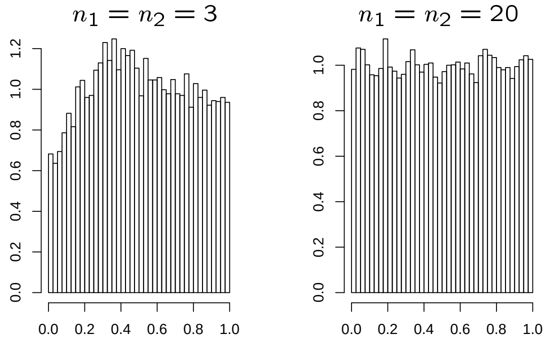


These distributions are fairly strongly skewed. They both have mean of one, so the null hypothesis of equal means is true.

How Correct are the P-Values?

Since the null hypothesis is true here, the P-values should be uniformly distributed between 0 and 1 — ie, the probability of getting a P-value less than α should equal α.

To test this, I simulated 20000 data sets, with sample sizes of $n_1 = n_2 = 3$ and $n_1 = n_2 = 20$. I got the following histograms:



The fraction rejecting at $\alpha = 0.05$ was 0.033 when $n_1 = n_2 = 3$ and 0.051 when $n_1 = n_2 = 20$.