

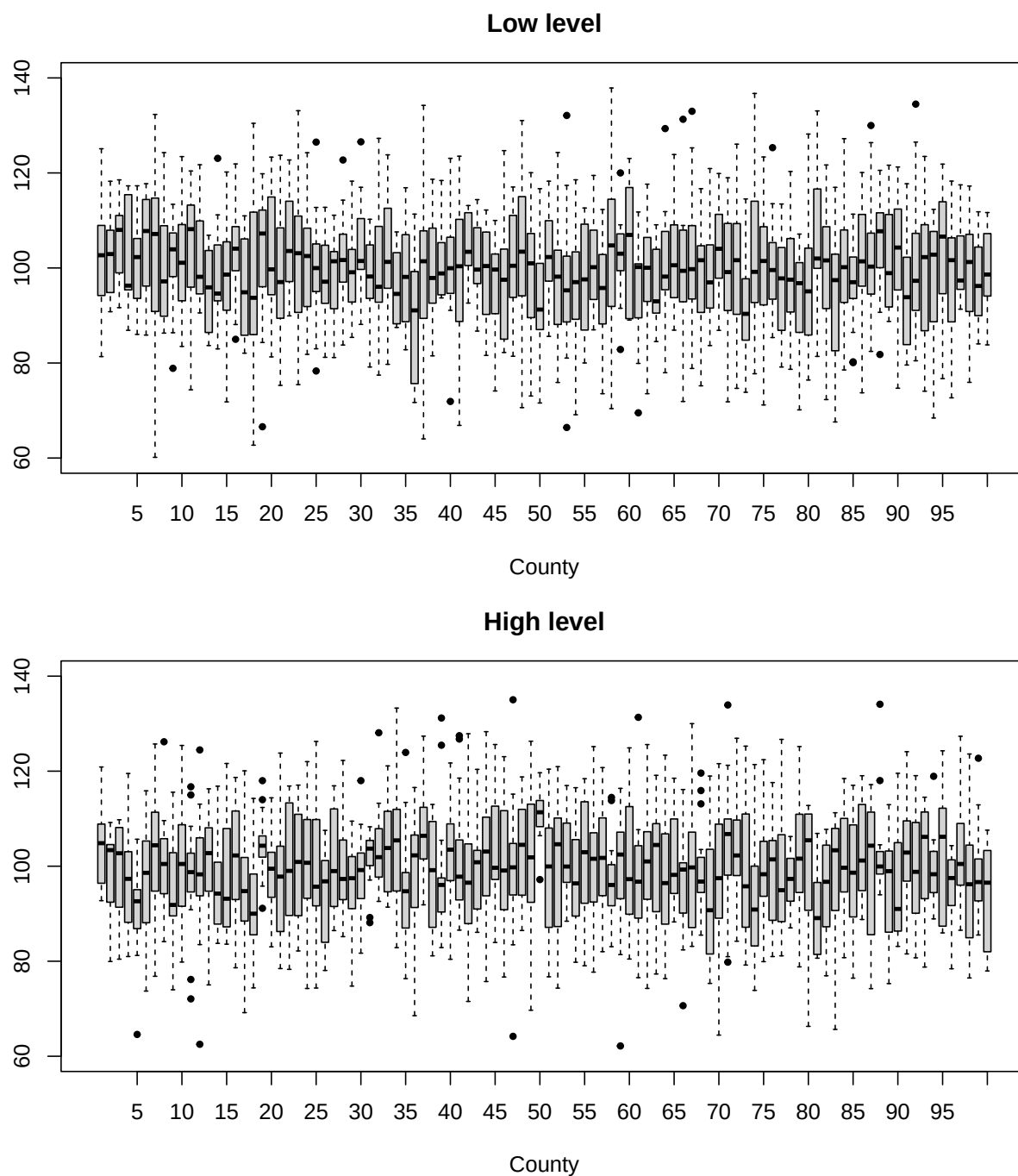
MATH 203: PRINCIPLES OF STATISTICS 1

FALL 2024

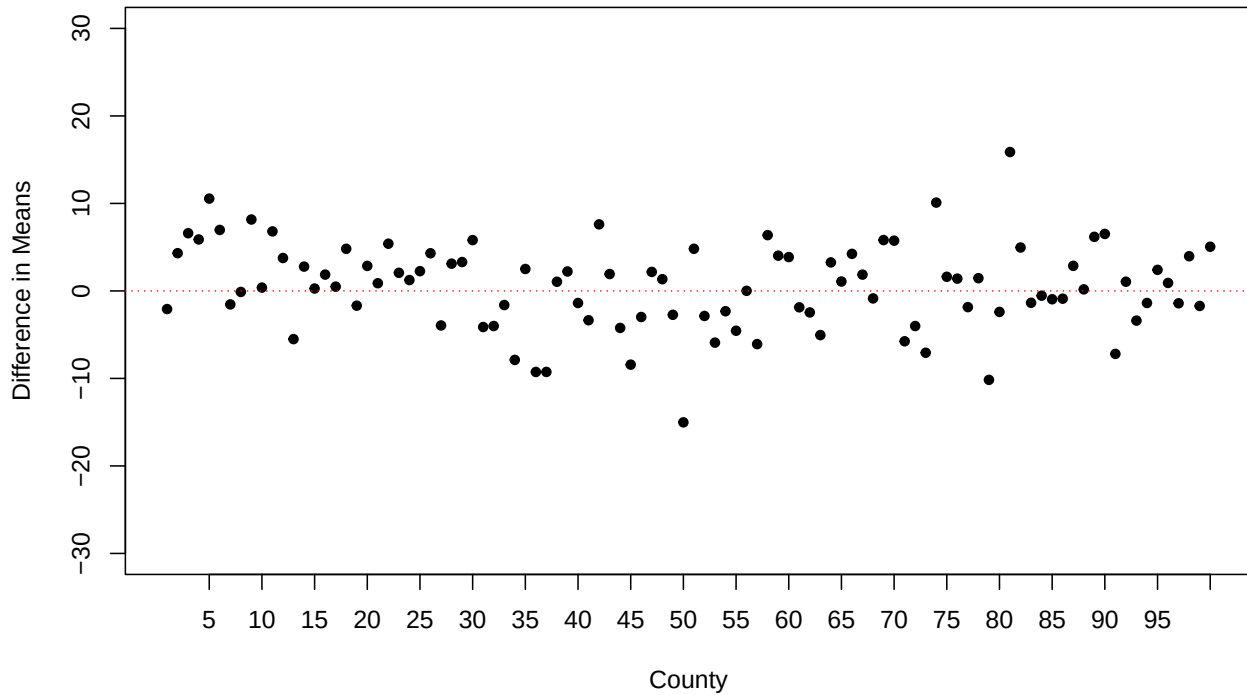
INTERPRETING THE INFERENCE

In each of 100 counties, data are collected measuring the IQ (measured by standard testing) of a small number of infants in the county, and at the same time, whether that infant lives in a household with a low or high level of an environmental factor.

The following plot displays boxplots of the IQ samples across the 100 counties.



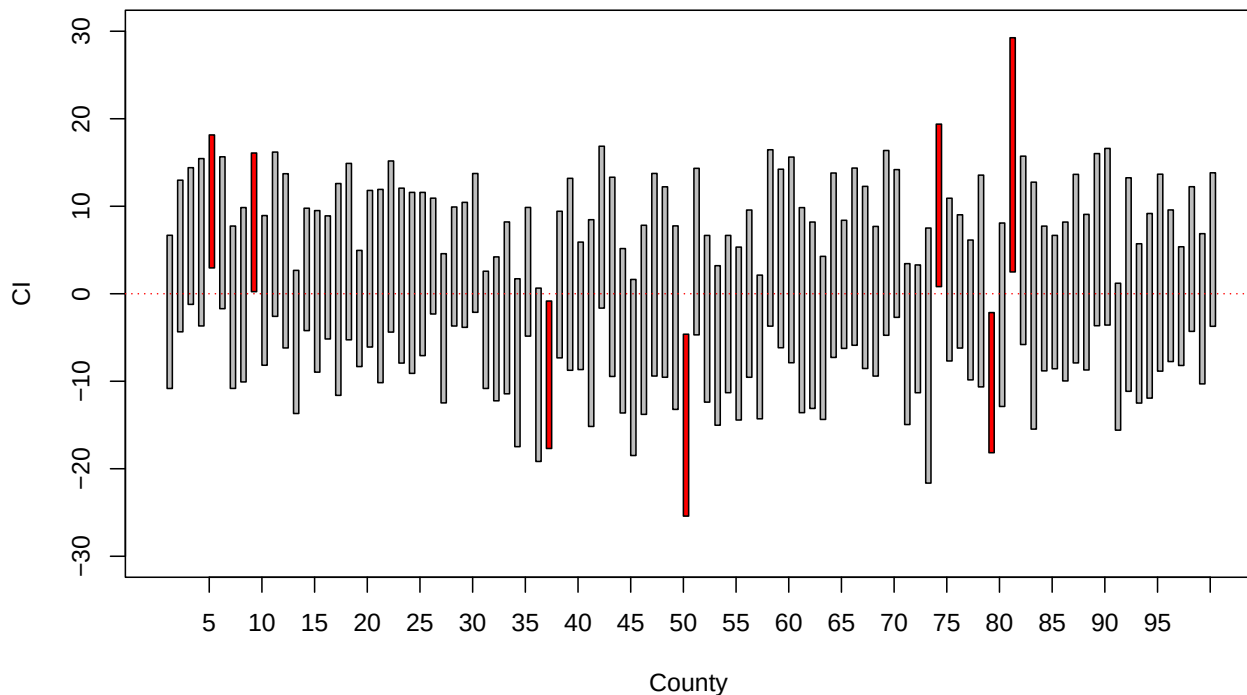
We can also plot the differences between the IQ means of the Low and High level households:



For a statistical assessment, we can construct 95% Confidence Intervals for the differences in means in each county, using the formula (in this small sample setting, under the assumption of equal population variances)

$$(\bar{x} - \bar{y}) \pm t_{\alpha/2, m+n-2} s_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

where m and n are the sample sizes for the Low and High level households.



The intervals coloured in red are those that exclude zero (that is, the interval lies entirely above or entirely below the zero line). There are seven such intervals out of 100. We can repeat also report the test statistics for tests of the hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

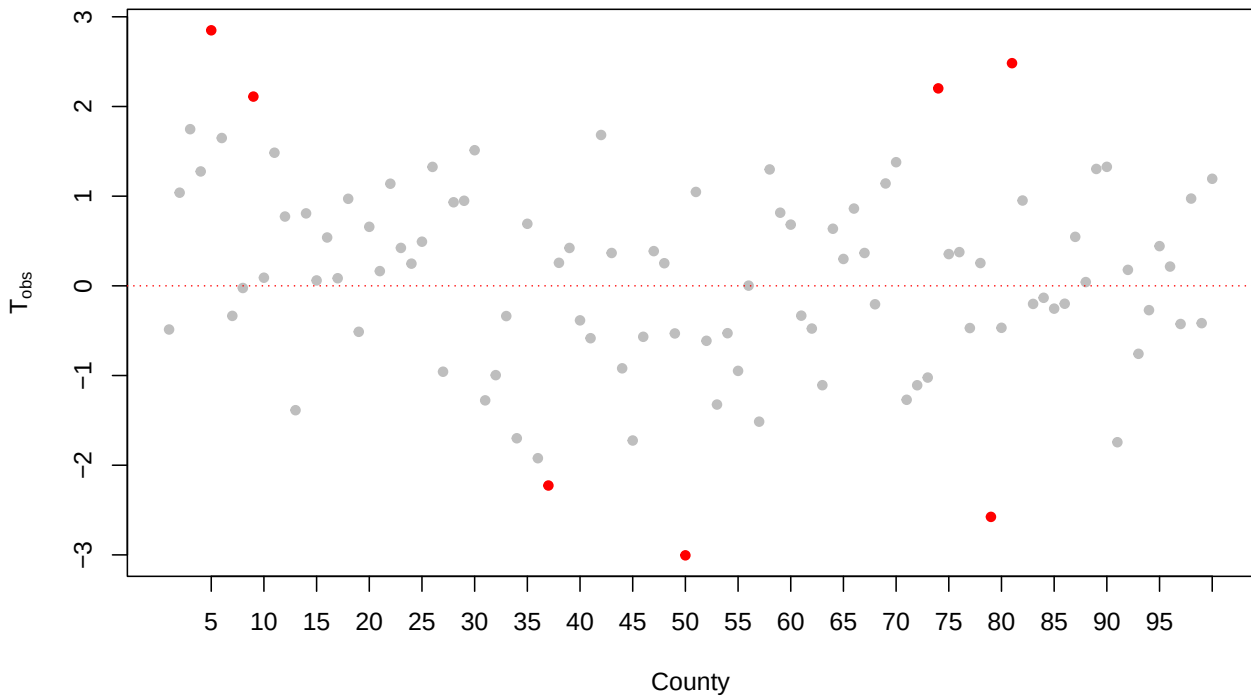
assuming a common population variance σ for the two populations using the test statistic

$$T_{obs} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

plotting T_{obs} in red for the statistics indicating rejection of H_0 in favour of H_a . The Rejection Region in these two-sided tests take the form

$$\{|T| \geq t_{\alpha/2, m+n-2}\} = \{T \leq -t_{\alpha/2, m+n-2}\} \cup \{T \geq t_{\alpha/2, m+n-2}\}$$

(the union of the left-tail region and the right-tail region) and we may test using $\alpha = 0.05$.



We see the same seven counties identified as leading to a rejection of the null hypothesis. It might be considered more relevant to consider the one-sided alternative hypothesis

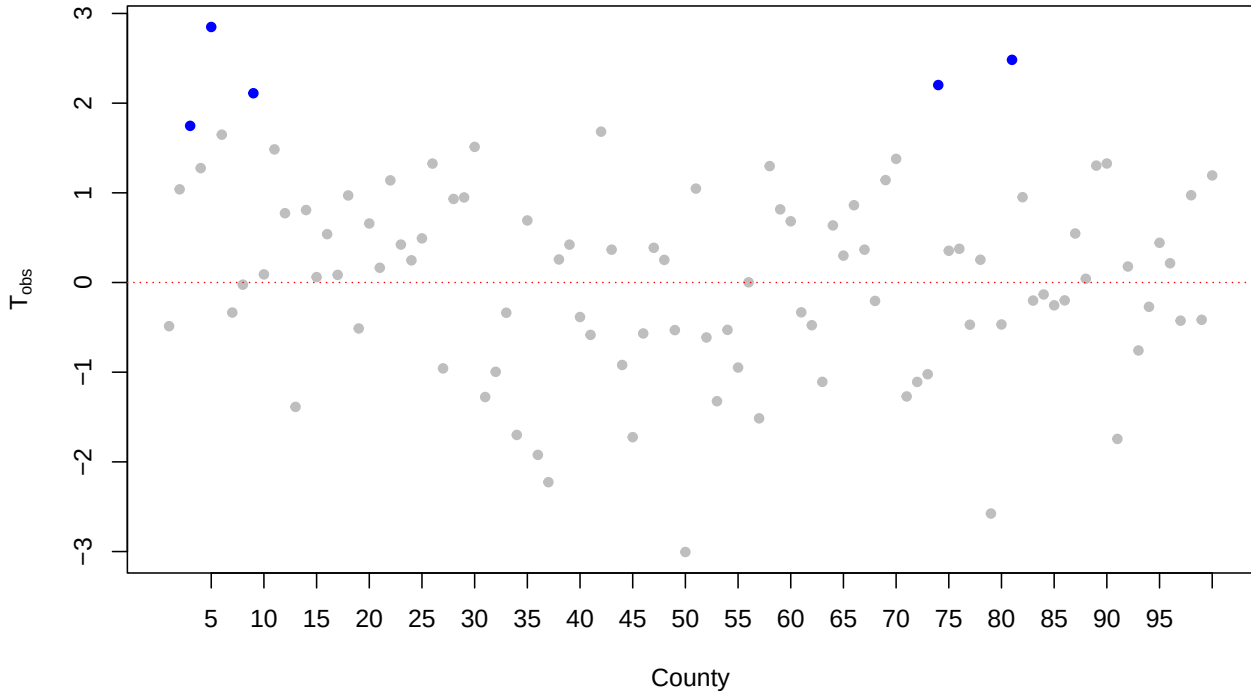
$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

for which the the Rejection Region take the form

$$\{T \geq t_{\alpha, m+n-2}\}$$

(the right-tail region only).



The statistics plotted in blue indicated rejection of the null hypothesis in favour of the alternative hypothesis, again using $\alpha = 0.05$. We see that five tests lead to rejection.

Note: In fact, these data are simulated under the null hypothesis with equal variances:

$$\text{Low } X_1, \dots, X_m \sim \text{Normal}(100, 12^2)$$

$$\text{High } Y_1, \dots, Y_n \sim \text{Normal}(100, 12^2)$$

The observed variation in the test statistic, and the rejections of H_0 that we see, are merely the result of random variation, and there are in fact no systematic differences between the two populations.

These results remind us of the logic being appealed to in the hypothesis testing framework: after having formed the hypotheses and chosen a suitable test statistic T , we

- predict how T **should** behave if the null hypothesis is **true** by forming the null distribution;
- compute the observed test statistic T_{obs} ;
- check whether the observed result is **surprising**, compared with our prediction: does T_{obs} lie in a low probability region of the null distribution?
- If the result is sufficiently surprising we **reject** the null hypothesis in favour of the alternative: the degree of ‘surprise’ we are willing to tolerate is measured by α , as

$$\alpha = \Pr(T \text{ lies in R. R.} | H_0 \text{ is true}).$$

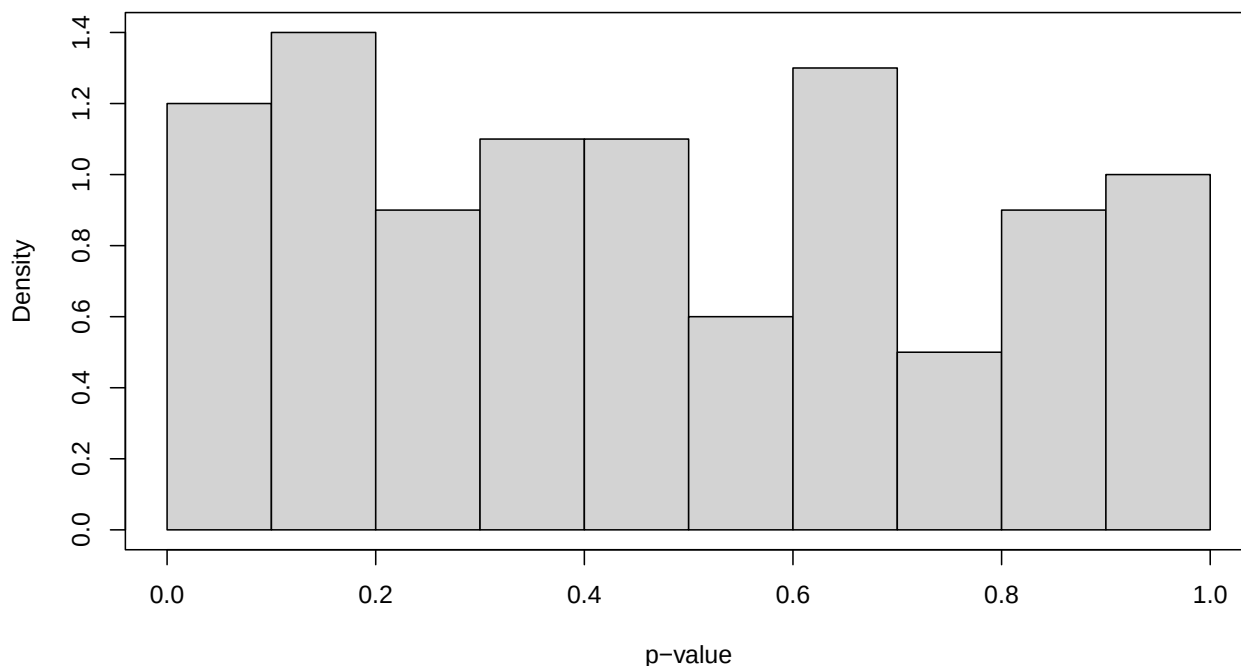
- even if H_0 is true, we **expect** to see T_{obs} in the Rejection Region $100\alpha\%$ of the time.

The computation of p-values does not change this interpretation. For example, for the one-sided alternative hypothesis, we compute the p-value as

$$\text{p-value} = P_0(T \geq T_{obs})$$

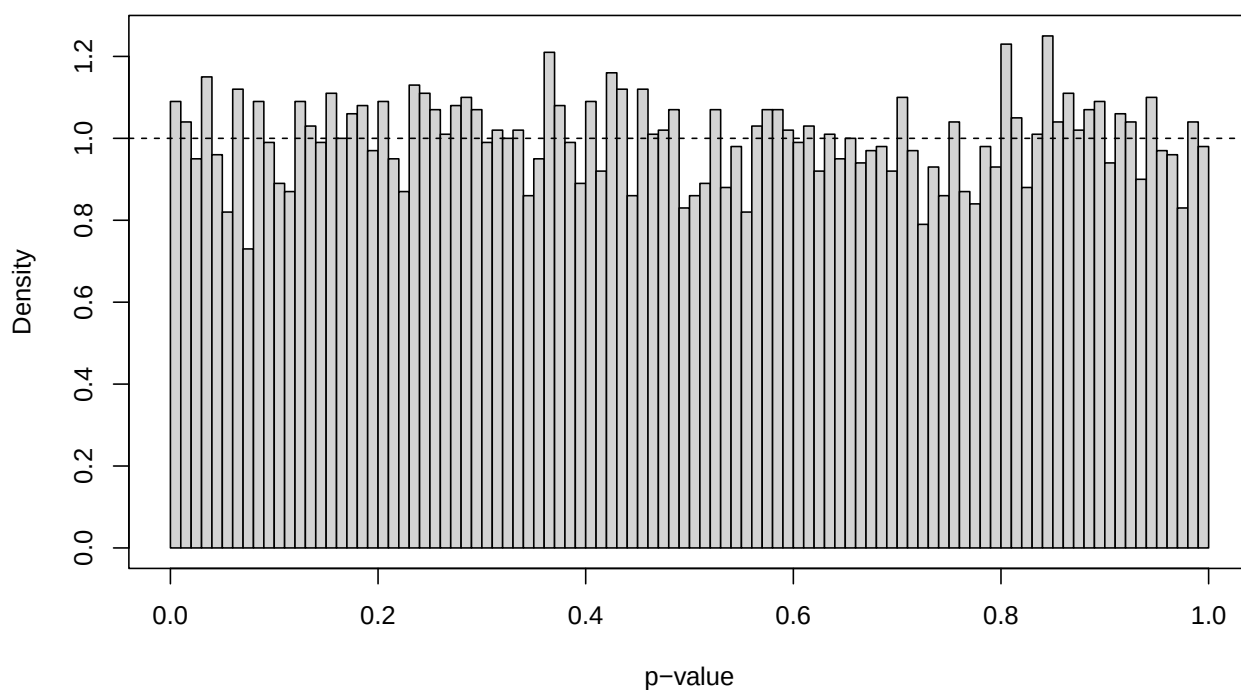
and we can plot the 100 p-values in a histogram.

Histogram of p-values from 100 counties



The p-values are **uniformly** distributed across the interval $(0, 1)$. The uniformity is easier to see if we have $N = 10000$ counties:

Histogram of p-values from 10000 counties

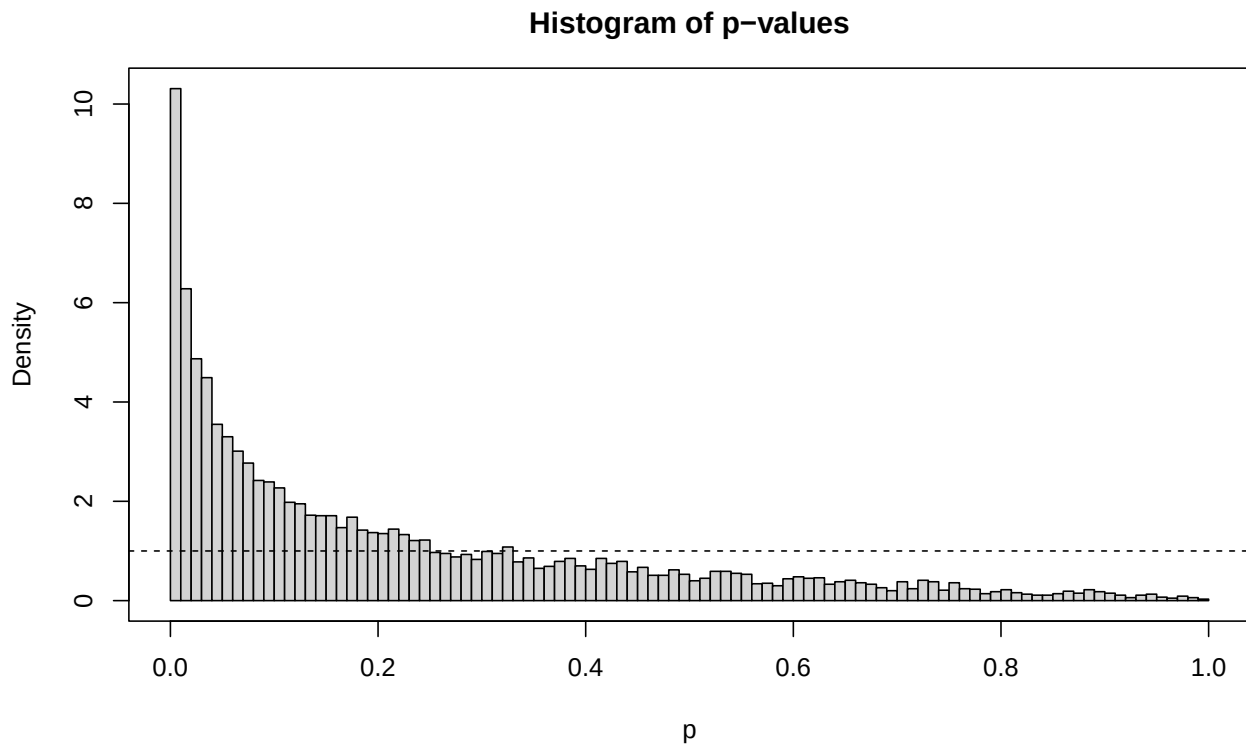


Contrast this with the case where there is in fact a systematic difference between Low and High level households: if we simulate

$$\text{Low } X_1, \dots, X_m \sim \text{Normal}(100, 12^2)$$

$$\text{High } Y_1, \dots, Y_n \sim \text{Normal}(95, 12^2)$$

the p-value histogram appears as follows:



If we use a p-value threshold of $\alpha = 0.05$, it turns out that 2950 of the tests lead to p-values below α and a (correct) rejection of the null hypothesis.