

# Testing for normality

Duncan Golicher

December 2, 2008

It is quite easy to run simulated sampling experiments using R. This can help to clarify the sort of statistical concepts we all have difficulty with. The practice of testing samples for normality has been compared with setting out in a rowing boat in order to test whether it is safe to launch an ocean liner (Box, 1953). Here is a demonstration of why Kolmogorov-Smirnov tests don't say what many people think they say.

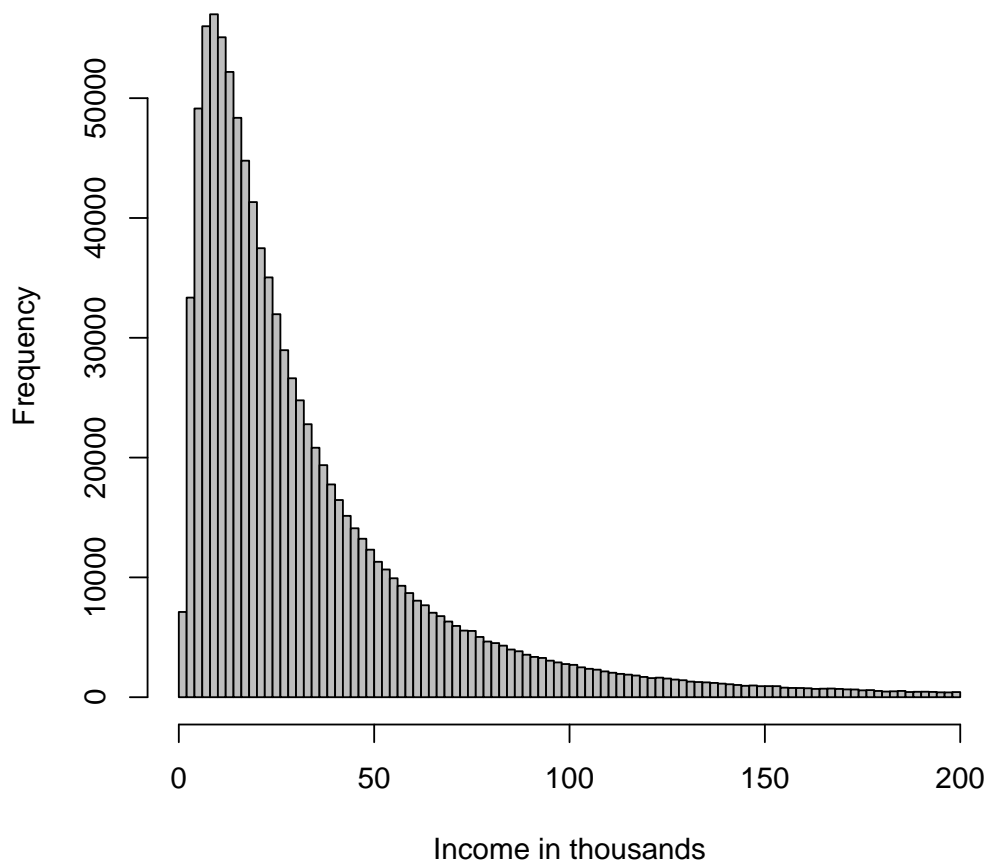
In R you need to install the package `nortest` first.

```
> install.packages("nortest")
```

We'll simulate a large set of numbers (one million) from a highly skewed population to produce a large but finite sampling frame that very closely approximates to an idealised statistical population. Say the UK national income figures. We can approximate these as a lognormal, but cut off the extreme tail.

```
> library(nortest)
> set.seed(1)
> pop <- rlnorm(1e+06, mean = log(23), sd = 1)
> pop <- pop[pop < 200]
> hist(pop, xlim = c(0, 200), breaks = 100, xlab = "Income in thousands",
+      col = "grey")
```

## Histogram of pop



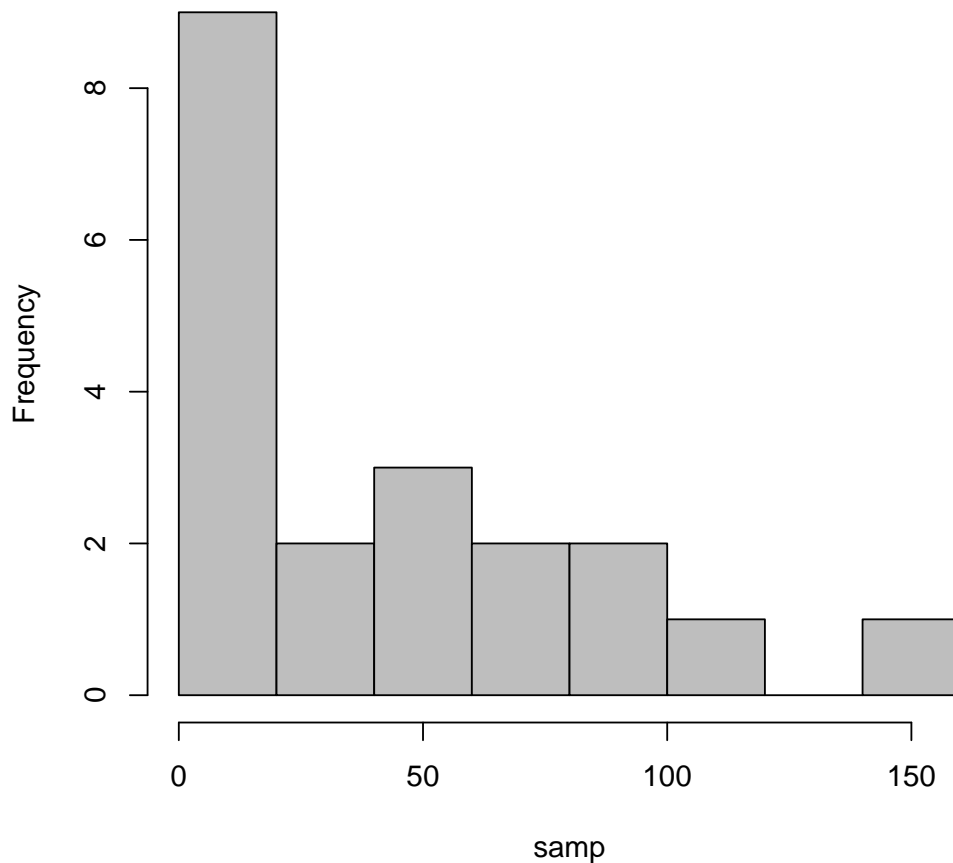
Now we'll take a sample of 20 people from this population and apply a KS test (or the adapted Lilliefors's version used in R (Lilliefors, 1967)).

```
> samp <- sample(pop, 20)
> hist(samp, col = "grey")
> lillie.test(samp)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: samp
D = 0.2474, p-value = 0.002330
```

## Histogram of samp



Great. The test tells us that the sample is *significantly non-normal* (sic), as it should. But it doesn't really of course. This is a misunderstanding. It tests the null hypothesis. There are often conceptual and practical problems with testing a null hypothesis of any kind (Johnson, 1999). The interpretation of a p-value for a null hypothesis test is like this.

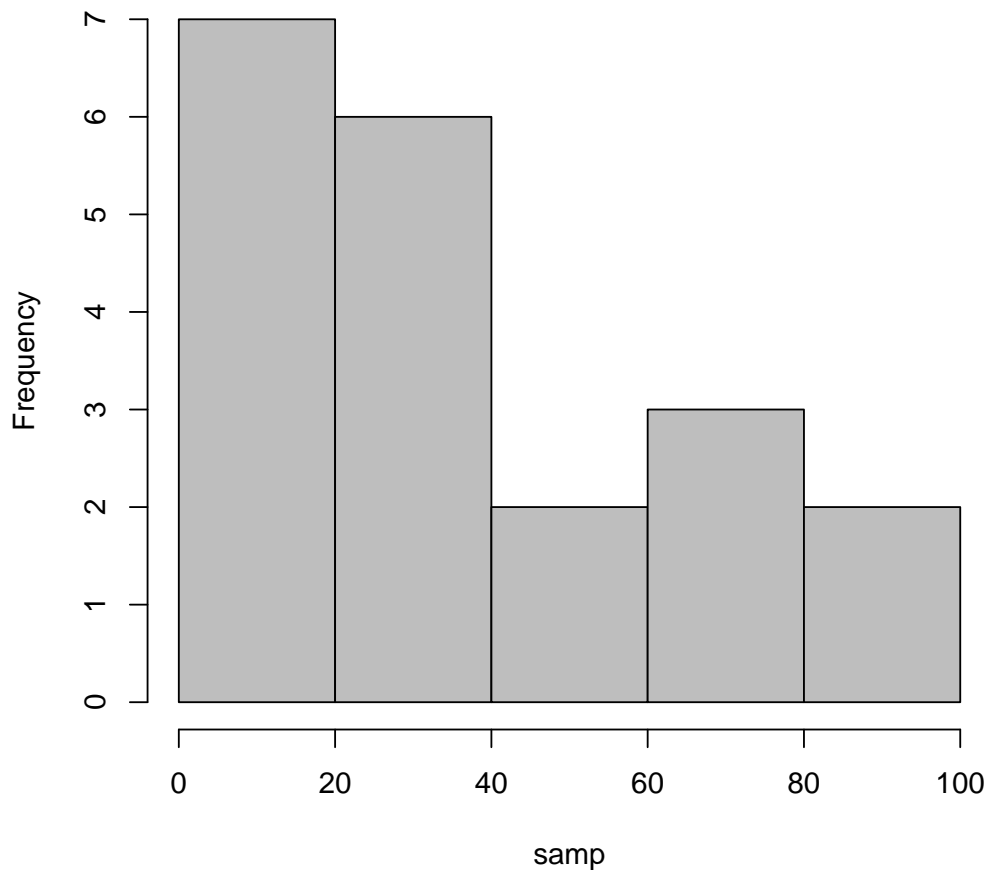
*How probable is it that this data, or data more extreme, could have been obtained from a normal distribution?*  
So let's do it over again.

```
> samp <- sample(pop, 20)
> hist(samp, col = "grey")
> lillie.test(samp)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: samp
D = 0.1354, p-value = 0.438
```

## Histogram of samp



Oops, the second result doesn't look right somehow. **If the p-value is  $>0.05$  (conventionally) we don't reject the null hypothesis.** So the correct interpretation of the test is that we do not have enough evidence to conclude that the data weren't drawn from a normal distribution. **In reality they weren't**, and we know that because we designed the population ourselves. Another common, but clearly incorrect interpretation is that the data themselves form a normal distribution. They **don't** and **the test is not designed to tell us this**. Asking if the data are normal is a meaningless question. Data are never normal and never can be, because normality is just a (very convenient) theoretical mathematical construct. For inference to be valid even the underlying population doesn't really have to be exactly normal. That would be expecting far too much. An acceptable approximation is good enough<sup>1</sup>. In this case the population distribution is **clearly not an acceptable** approximation to normal, it is far too skewed. So we really should not be led to assume that is normal very often or we will be led astray. However, maybe this was just a one off.

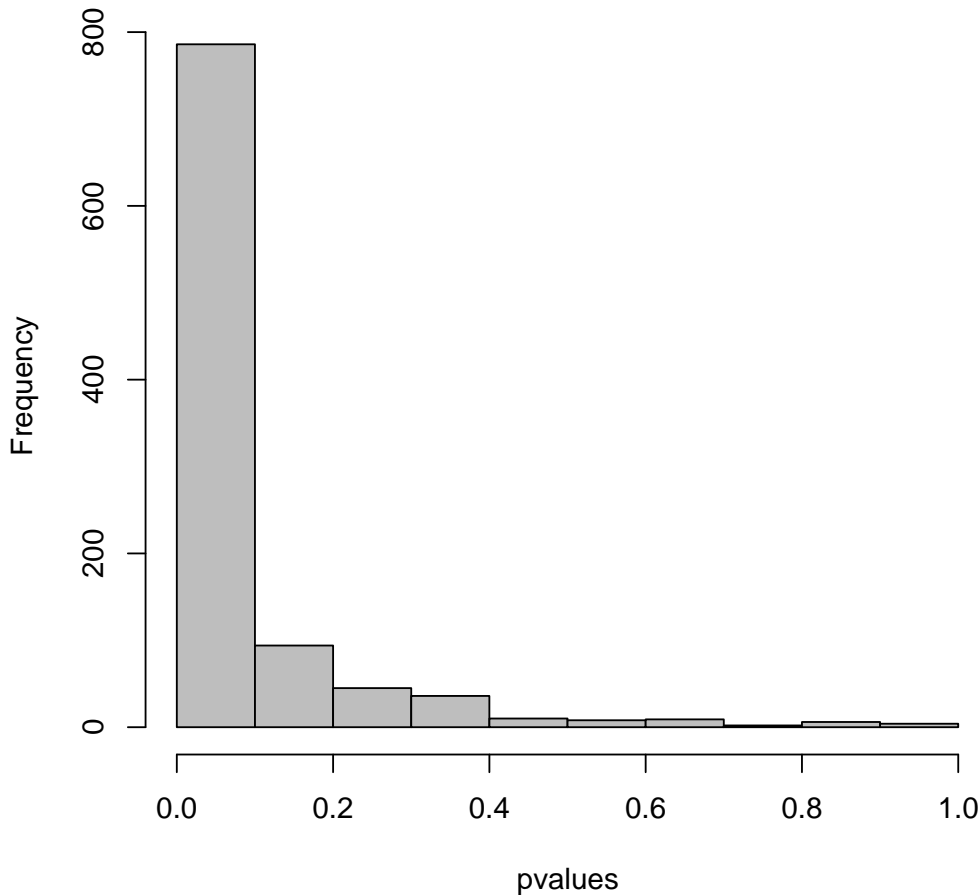
We can try the sampling experiment another 1000 times and see how well the test works by getting 1000 p-values. According to most peoples intuition 1 in 20 of these should be  $>0.05$

```
> pvalues <- replicate(1000, lillie.test(sample(pop, 20))$p.value)
> hist(pvalues, col = "grey")
```

---

<sup>1</sup>Which of course begs the question how close to normal is acceptable?

## Histogram of pvalues



OK. Clearly there are more small p-values than larger ones. The test is doing something right. However does it tell us what we expect? The proportion of the p-values considered “significant” at the 0.05 level is only 0.69

```
> sum(pvalues > 0.05)/1000
```

```
[1] 0.31
```

```
> sum(pvalues < 0.05)/1000
```

```
[1] 0.69
```

The problem with many peoples incorrect intuitive interpretation (the test should say “non normal” 95% of the time) is that they are confusing a p-value with the power of a test to reject the null. When samples are small or the population distribution not far from normal the power of a KS test is very low. When samples are large the power becomes greater. However if you have a large sample you can get a very good idea of the distribution of the population anyway. The test doesn’t tell you anything new. The outcome of relying on a test for normality in this case is that 31% of the time you would go ahead with inference on the mean, even though the underlying population is such that you should probably be comparing medians for most practical purposes. And of course in the converse case, of the null being true, in around one in twenty cases you would be told that the data didn’t fit the assumption for your analysis, even though the population did.

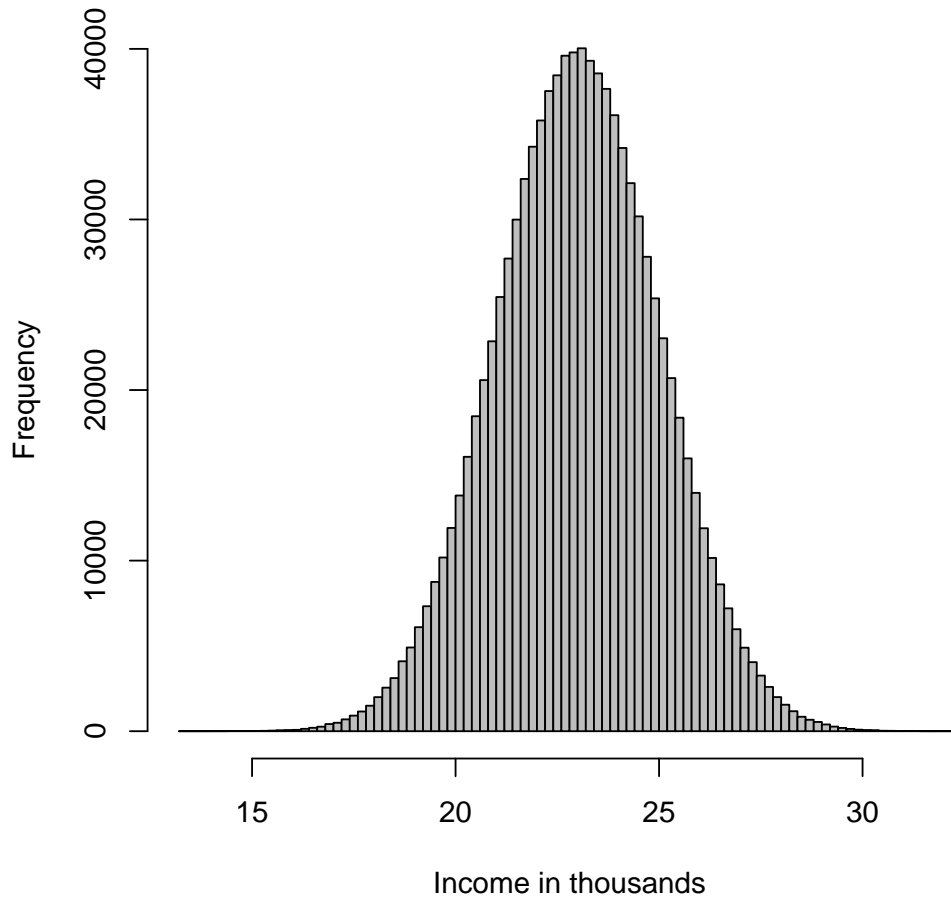
Just to confirm this we’ll get a large set of numbers from a genuinely normal population.

```
> set.seed(1)
```

```
> pop <- rnorm(1e+06, mean = 23, sd = 2)
```

```
> hist(pop, breaks = 100, xlab = "Income in thousands", col = "grey")
```

## Histogram of pop



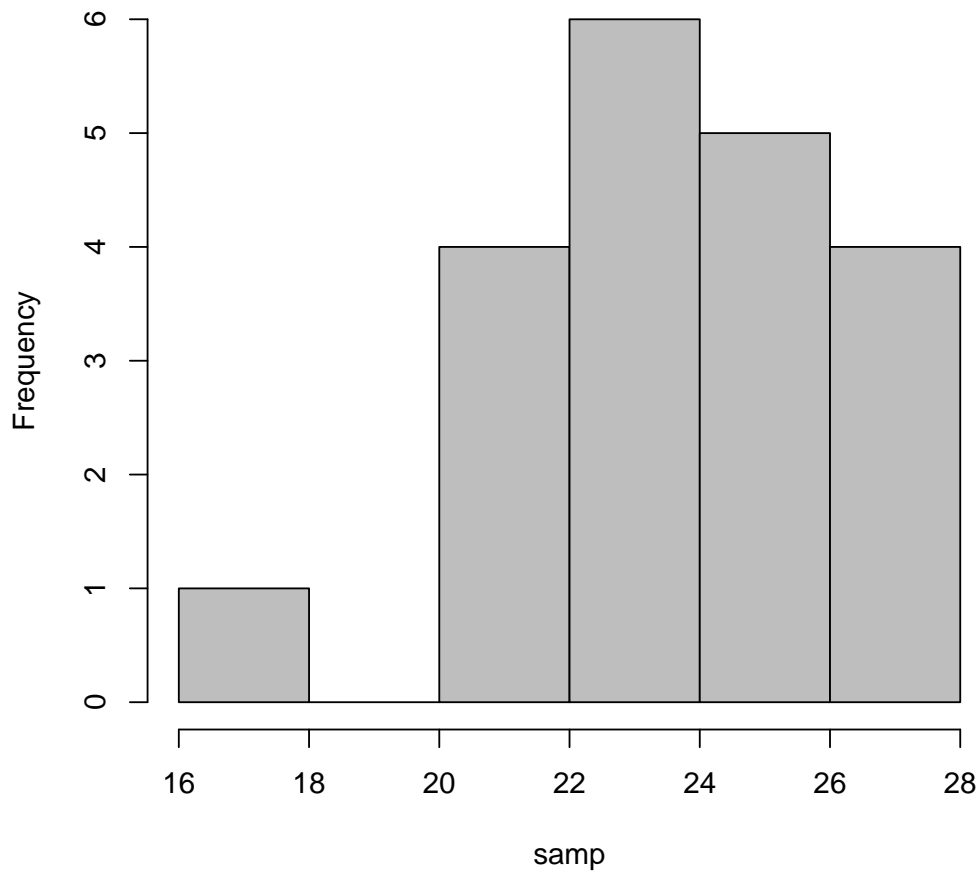
Now take a sample of 20 people from this population and apply a KS test.

```
> samp <- sample(pop, 20)
> hist(samp, col = "grey")
> lillie.test(samp)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: samp
D = 0.0996, p-value = 0.869
```

## Histogram of samp



The histogram doesn't look very normal but remember that the test asks.

*How probable is it that this data, or data more extreme, could have been obtained from a normal distribution?*

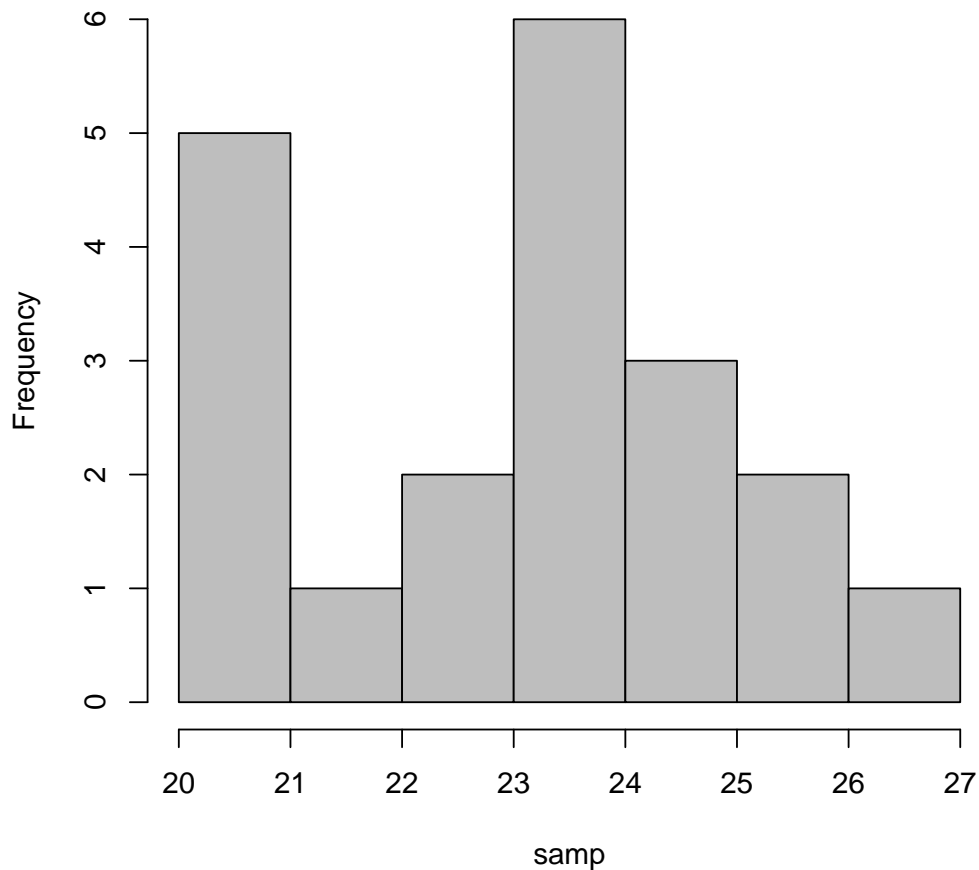
Try again.

```
> samp <- sample(pop, 20)
> hist(samp, col = "grey")
> lillie.test(samp)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: samp
D = 0.1246, p-value = 0.5728
```

## Histogram of samp

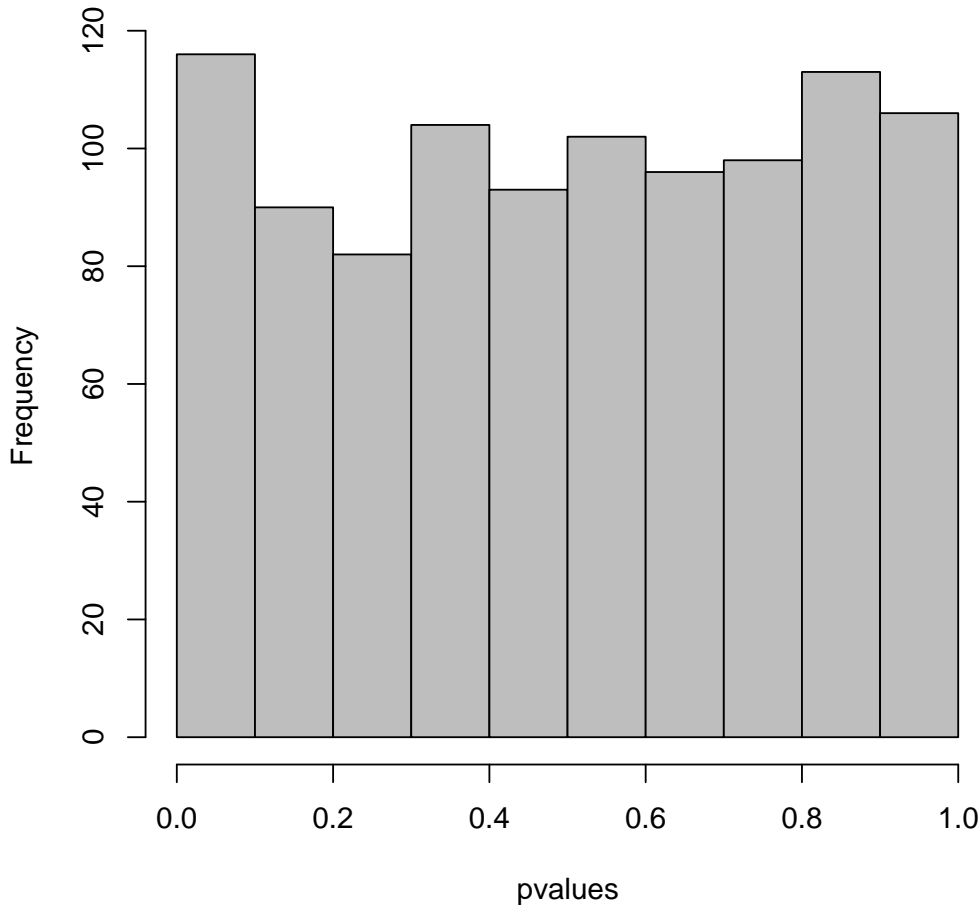


It is interesting to note that neither histogram looks normal. However the test says that there is nothing wrong with the assumption that the sample **could** have been drawn from a normal population. This is correct. It **could**, and in fact we know that it **was**. Small samples rarely look normal even when the population is perfectly normal, as in this case it is.

We again will run the sampling experiment another 1000 times.

```
> pvalues <- replicate(1000, lillie.test(sample(pop, 20))$p.value)
> hist(pvalues, col = "grey")
```

## Histogram of pvalues



Now the p-values form a more or less uniformly distributed random variable on the interval 0-1. The proportion of the p-values considered “significant” at the 0.05 level in this experiment is close to 5%. Because the experiment was only run 1000 times it is not exactly 0.05, it is in fact 0.06 . However a larger sampling experiment would be closer.

```
> sum(pvalues > 0.05)/1000
```

```
[1] 0.94
```

```
> sum(pvalues < 0.05)/1000
```

```
[1] 0.06
```

Tutors do tend to understand these issues, but a problem still arises for research practice. Many students justifiably don’t get it at first take. There is something in null hypothesis testing that is so counter intuitive that is actually very unfair to expect them to. If they take their incorrect intuitive interpretation of the tests literally they begin to worry about whether “their data will be vaild”. Pragmatic solutions to these sort of problems in an applied context have been proposed by ecologists over the years, eg (Johnson, 1999; Osenberg et al., 2002; Stewart-Oaten, 1995; Stewart-Oaten et al., 1996). The simplest in my view is to insist that students try to have a good idea in their heads of the most likely distribution for the data before beginning work.

## References

Box, G. (1953). Non-normality and tests of variance. *Biometrika*, 40(3-4):318–335.

Johnson, D. (1999). The insignificance of significance testing. *Journal of Wildlife Management*, 63:763–772.

- Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.
- Osenberg, C., St. Mary, C., Schmitt, R., Holbrook, S., Chesson, P., and Byrne, B. (2002). Rethinking ecological inference: density dependence in reef fishes. *Ecology Letters*, 5(6):715–721.
- Stewart-Oaten, A. (1995). Rules and judgments in statistics: three examples. *Ecology*, 76(6):2001–2009.
- Stewart-Oaten, A., Schmitt, R., and Osenberg, C. (1996). Goals in environmental monitoring. *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats*, pages 17–27.