USING SIMULATION TO UNDERSTAND PROBABILITY

DEEPAYAN SARKAR

Probability theory studies the behaviour of data generated from a completely known statistical model. Inference, on the other hand, starts with data and tries to infer plausible statistical models that *could have* generated that data. If you have data you want to "analyze", inference is probably what you want to do. However, to properly appreciate techniques used for inference, one must understand the associated probability theory. This is especially important for asymptotic (large-sample) statements such as the Central Limit Theorem. Modern statistical software come with a variety of tools to simulate random numbers, and these can be effectively used to understand abstract probabilistic statements.

Many quantitative statements made in inference say things about what would happen if "the exeriment was repeated a large number of times". One cannot of course do this in practice, but a random experiment can be repeated an arbitrary number of times on a computer. An important paradigm in inference (specifically hypothesis testing) is the following:

- Assume a parametric model for your data (with unknown parameters).
- Obtain a test statistic whose distribution *does not depend on the unknown parameter under the null model*.
- Determine how "probable" the observed value of the test statistic is under this *known* probability model.
- If the observed value was very unlikely, then the null model is in doubt.

Statisticians spend a lot of time trying to calculate exact theoretical distributions for various test statistics in various setups. But again, the ability to generate random numbers using modern software means you can explore the behaviour of test statistics even if you cannot compute the exact distribution. More importantly, you can explore what happens when your model assumptions (such as normality) are not exactly satisfied.

The Central Limit Theorem. If $X_1, X_2, \ldots, X_n$ are independent random variables from a common distribution, then the central limit theorem states that under certain conditions (e.g., finite mean μ and variance σ^2 of the X_i-s), the distribution of the standardized sample mean $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ "converges" to the standard normal distribution. More precisely,

$$\lim_{n\to\infty} P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq u\right) = P(Z \leq u) \tag{1}$$

where Z is a standard Normal random variable.

This result (which is especially impressive because of its generality) is perfectly satisfactory for probabilists, but practitioners would naturally ask "how large does n need to be for me to use this approximation?" Unfortunately, there is no *general* answer to that question; it depends on the context, i.e., the distribution of the X_i-s. Fortunately, it is easy enough to study the question for any *specific* distribution using simulation.

Example: the uniform distribution. Suppose our $X_1, X_2, \ldots, X_n$ are independent $U(0,1)$ random variables. Then clearly $\mu = \frac{1}{2}$, and

$$\sigma^2 = E(X_1^2) - \mu^2 = \int_0^\infty x^2 dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Let the standardized sample mean (of n observations) be

$$Z_n = \frac{\bar{X} - \frac{1}{2}}{\sqrt{\frac{1/12}{n}}}$$

We are then interested in knowing how large n needs to be before we can consider the Normal approximation reasonable.

In real life, we will usually have only one realization of the random variables $X_1, X_2, \ldots, X_n$. We can mimic this situation in R with

```
> n <- 10
> x <- runif(n)
```

We can compute the corresponding Z_n as

```
> (mean(x) - 0.5) / sqrt((1/12) / n)
[1] -1.643419
```

As this is not real life, we can repeat this pretend-experiment as many times as we want. There are several ways to do so (e.g., a `for` loop), but a useful R function for the task is `replicate()`, which will execute a statement multiple times, providing the results in a handy form.

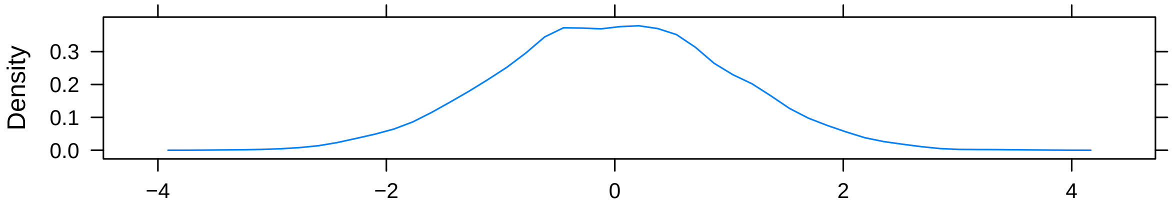
```
> zrep <- replicate(5000, (mean(runif(n)) - 0.5) / sqrt((1/12) / n))
```

So now we have 5000 realizations of the random variable Z_n, with the option of generating even more. Note that we cannot easily figure out the probability density function or distribution function of Z_n in closed analytic form. However, with the power to generate an arbitrary number of observations from this distribution, we *can* answer questions like

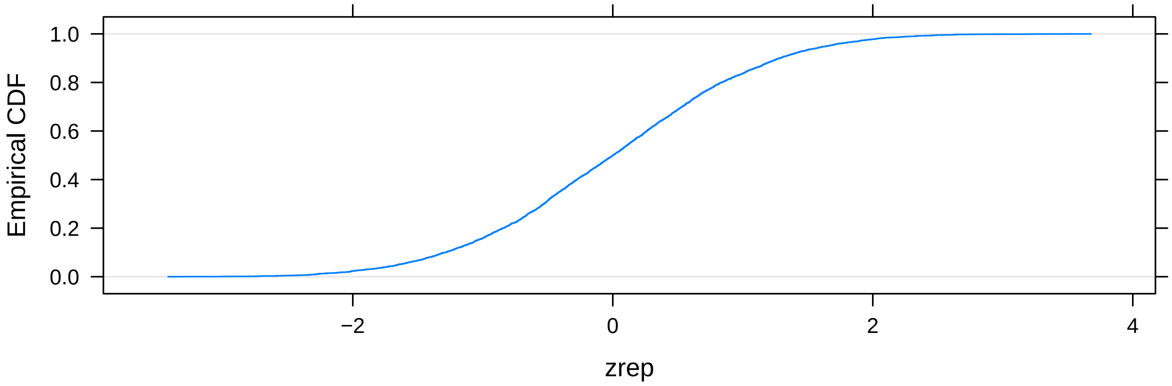
- What does the density of Z_n look like?
- What does the distribution function of Z_n look like?
- How close is the distribution of Z_n to $N(0, 1)$?

These questions can be answered by

```
> library(lattice)
> densityplot(~zrep, plot.points = FALSE, xlab = NULL)
```



```
> library(latticeExtra)
> ecdfplot(~zrep)
```



```
> ks.test(zrep, pnorm)

        One-sample Kolmogorov-Smirnov test

data:  zrep
D = 0.0106, p-value = 0.6282
alternative hypothesis: two-sided
```

The kernel density estimate and empirical CDF will of course still be random, but will approximate the true density and CDF better and better with more replications.

Notice that the Kolmogorov-Smirnoff test statistic gives you an approximation of

$$\max_{u\in\mathbb{R}} \left| P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le u\right) - P(Z \le u) \right|$$

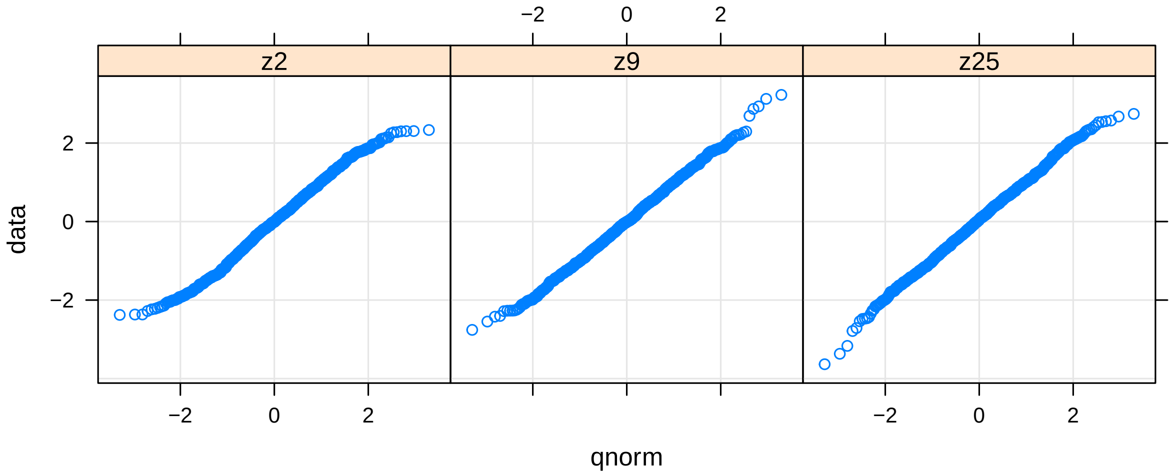
which is the maximum error you can make if you use the $N(0,1)$ distribution instead of the true distribution of Z_n (in the sense of the central limit theorem (1) above). This is an *approximation* in the sense that you are actually computing

$$\max_{u\in\mathbb{R}} \left| \hat{P}_n\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le u\right) - P(Z \le u) \right|$$

where $\hat{P}_n$ is the probability distribution corresponding to the empirical CDF $\hat{F}_n$. It is possible to show that they are the same asymptotically. You can already see the pointwise convergence (for any fixed u) using the Weak Law of Large Numbers: notice that $\hat{F}_n(u)$ is nothing but the proportion of successes in n tosses of a coin with probability of head $p = P(Z \le u)$.

The Q-Q plot. Having said all that, the simplest and usually the most effective way to judge the normality of a distribution is to plot the Normal Q-Q plot of data from that distribution.

```
> zrep2 <- replicate(1000, (mean(runif(2)) - 0.5) / sqrt((1/12) / 2))
> zrep9 <- replicate(1000, (mean(runif(9)) - 0.5) / sqrt((1/12) / 9))
> zrep25 <- replicate(1000, (mean(runif(25)) - 0.5) / sqrt((1/12) / 25))
> zrep.df <- make.groups(z2 = zrep2, z9 = zrep9, z25 = zrep25)
> qqmath(~data | which, zrep.df, distribution = qnorm, type = c("g", "p"))
```



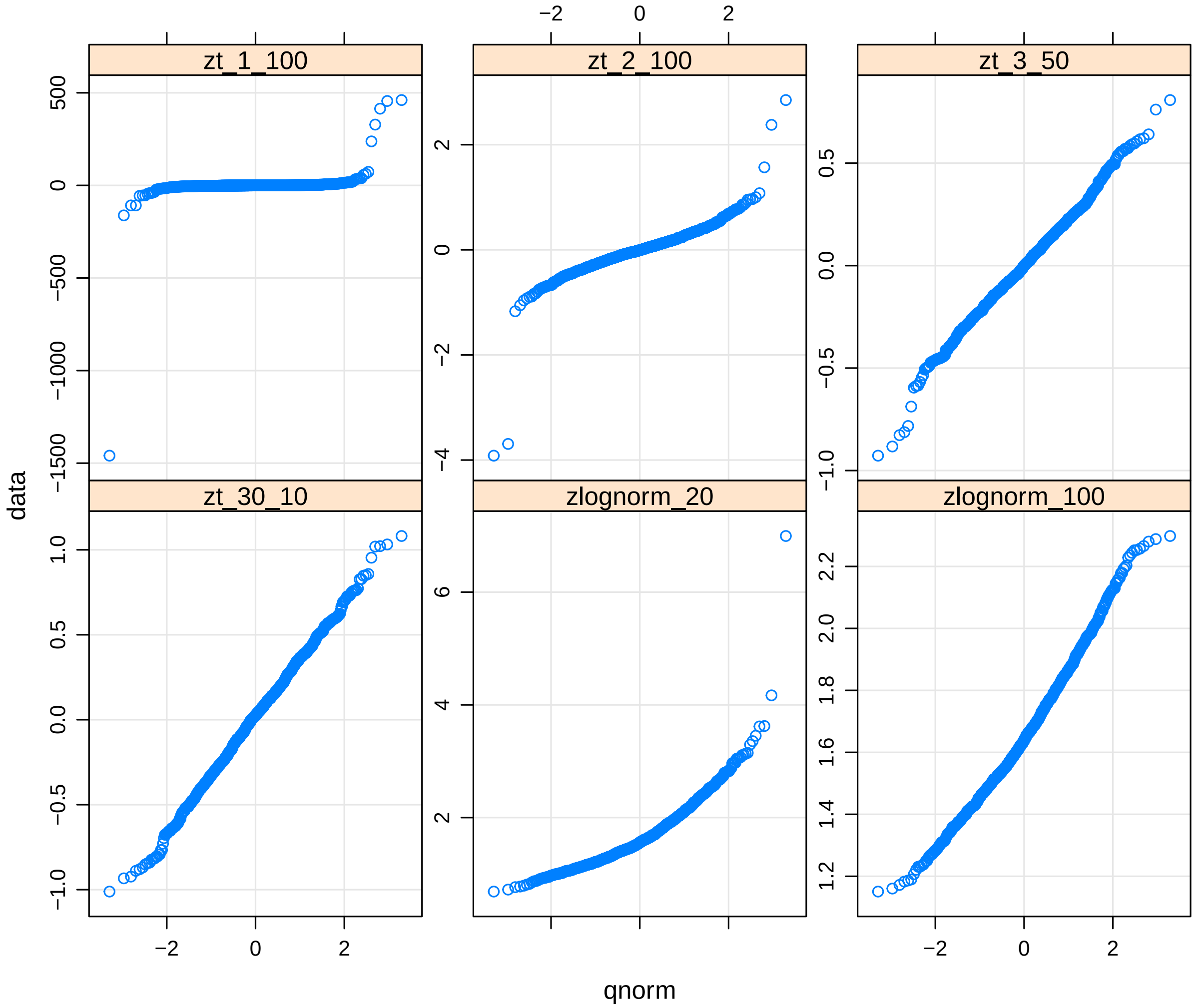
Departure from normality is suggested by *systematic* deviations from a straight line in the Q-Q plot. The reason Q-Q plots are so effective is that the human eye finds it easier to detect departure from straight lines than from curves.

Other distributions. We can repeat this whole procedure for distributions other than $U(0,1)$. We do not actually need to compute and standardize by the mean and standard deviation to use Q-Q plots, as a Normal Q-Q plot will still be linear as long the data follows *some* Normal distribution (μ and σ will change the intercept and slope of the line[1]). Here are some examples.

```
> ## A more flexible sample mean generator
>
> meanGenerator <- function(replications = 1000, n, rdist = runif, ...)
  {
      rsampleMean <- function()
      {
          mean(rdist(n, ...)) ## ... can supply extra parameters
      }
      replicate(replications, rsampleMean())
  }
> zt_1_100 <- meanGenerator(1000, n = 100, rt, df = 1) ## Cauchy
> zt_2_100 <- meanGenerator(1000, n = 100, rt, df = 2)
> zt_3_50 <- meanGenerator(1000, n = 50, rt, df = 3)
> zt_30_10 <- meanGenerator(1000, n = 10, rt, df = 30)
> zlognorm_20 <- meanGenerator(1000, n = 20, rlnorm)
> zlognorm_100 <- meanGenerator(1000, n = 100, rlnorm)
```

[1] Note that Q-Q plots can be drawn with non-Normal reference distributions as well, but this location-scale invariance property does not hold in general.

```
> qqmath(~data | which,
         data = make.groups(zt_1_100, zt_2_100, zt_3_50, zt_30_10,
                            zlognorm_20, zlognorm_100),
         scales = list(y = "free"), as.table = TRUE, type = c("g", "p"))
```

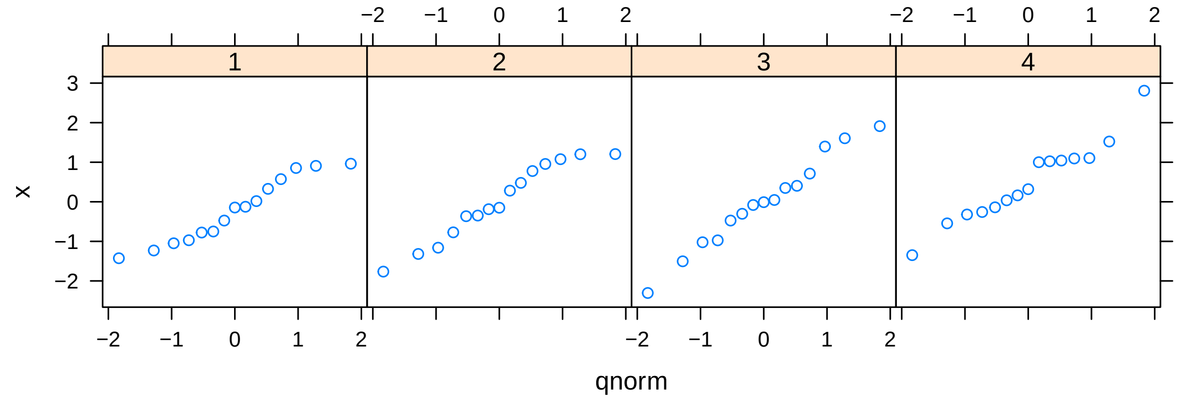


Note the bad behaviour of t_1 (which does not in fact satisfy the conditions required by the central limit theorem because its mean does not exist), t_2 (which has a finite mean but infinite variance, also not satisfying the conditions of the CLT), and the log-Normal distribution (which is skewed).

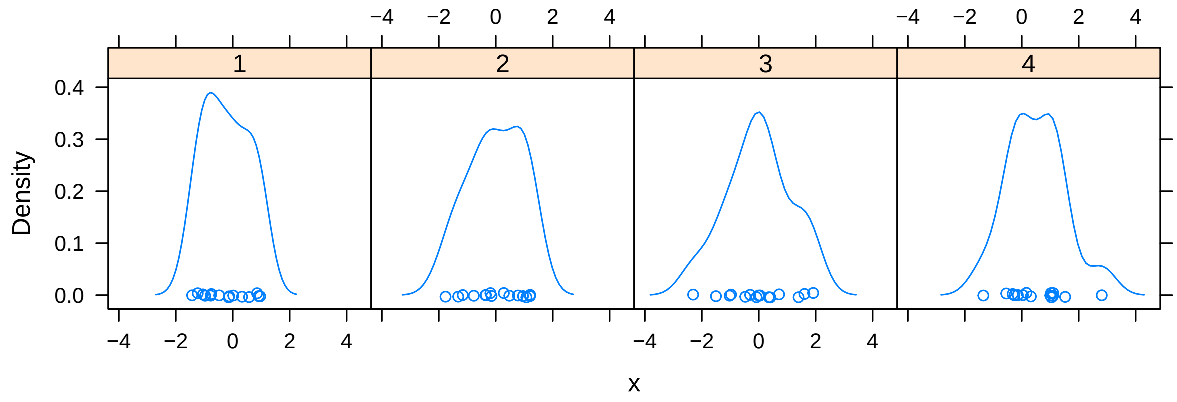
Q-Q plots for small samples

One should not come away with the impression that Q-Q plots (or any other graphical method or test) can be used to easily judge Normality. Such judgments are very difficult to make for small samples. Here are some examples of Q-Q plots and kernel density plots for small Normal samples.

```
> x <- rnorm(15 * 4)
> g <- gl(4, 15) ## An artificial grouping variable
> qqmath(~x | g)
```



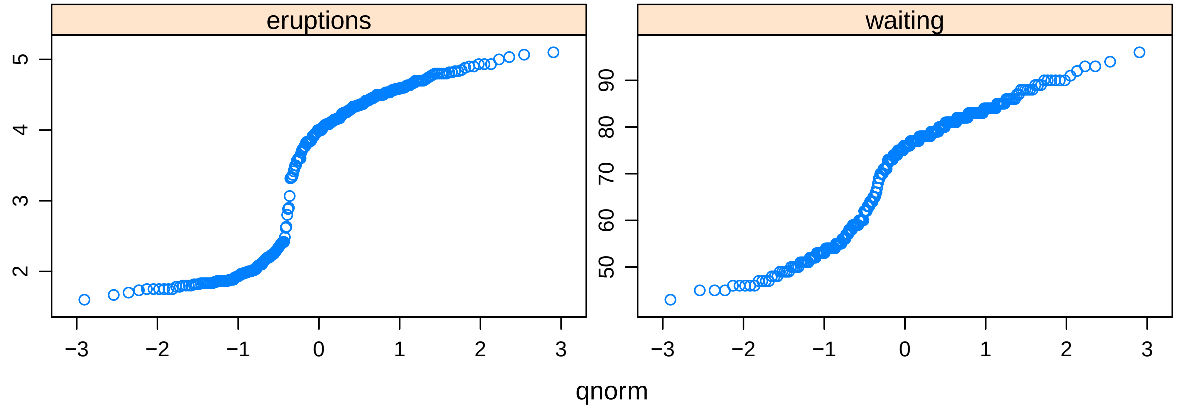
```
> densityplot(~x | g)
```



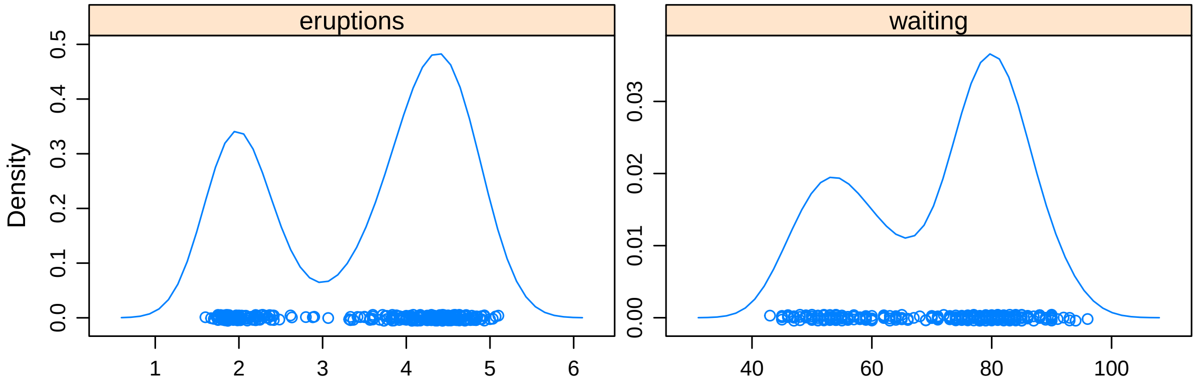
You might think that using the Kolmogorov-Smirnoff test is the solution, but that will have the problem of *low power* in small samples; that is, you will usually not reject the null hypothesis of Normality even when it is not true.

Detecting bimodality. Q-Q plots are also not very good at detecting bimodality; density plots are usually more effective for that.

```
> qqmath(~ eruptions + waiting, faithful, outer = TRUE, scales = "free", ylab = NULL)
```



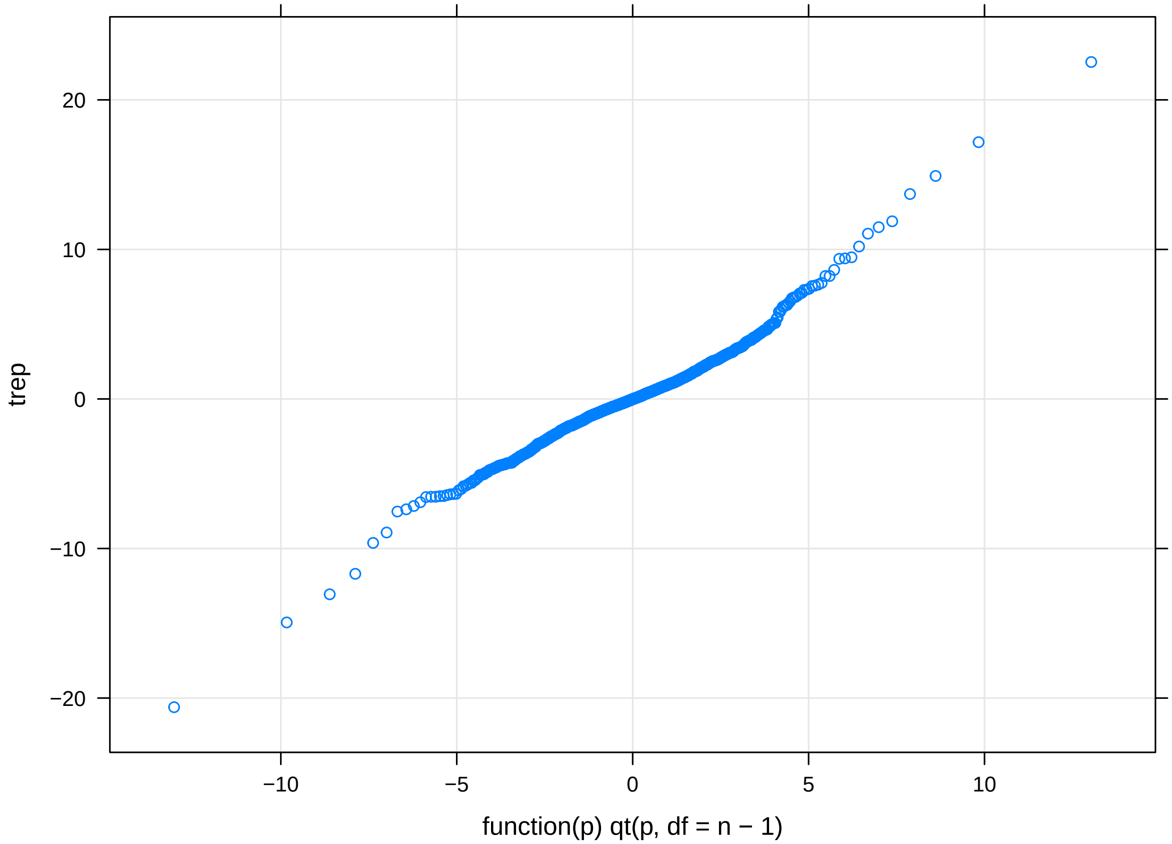
```
> densityplot(~ eruptions + waiting, faithful, outer = TRUE, scales = "free", xlab = NULL)
```



Q-Q plots with non-normal distributions

Let us revisit the example where $X_1, X_2, \ldots, X_n$ are independent $U(0,1)$ random variables. But instead of the sample mean, let us look at the t-statistic for testing $H_0\colon \mu = 0.5$.

```
> n <- 5
> trep <-
      replicate(5000,
            {
                x <- runif(n)
                s <- sd(x)
                t <- sqrt(n) * (mean(x) - 0.5) / s
            })
> qqmath(~trep, distribution = function(p) qt(p, df = n - 1),
         type = c("p", "g"))
```



Is this a good fit?

Exercise 1. *Check what the Q-Q plot would have been like in the "perfect" case, by repeating the experiment using any Normal distribution for your sample, rather than the uniform used here.*

Exercise 2. *Repeat the experiment above using different sample sizes* `n` *and different non-normal distributions (other than the uniform) to explore whether the t-test gives good results when the underlying distribution is not Normal. This is the study of* robustness *of the t-test against departures from Normality.*

A physics problem

The following assignment problem has been posed to you by Dr. Sabyasachi Chatterjee. Consider a gas column of length L, with N_0 hydrogen atoms per unit volume and at equilibrium at temperature T, moving with a net drift velocity $\bar{v}$, such that the number of atoms per unit volume having the velocity in the x-direction lying between v and $v + \mathrm{d}v$ is given by

$$N(v)\ \mathrm{d}v = \frac{N_0}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(v-\bar{v})^2}{2\sigma^2}\right\}\ \mathrm{d}v$$

If a photon of frequency ω traveling in the x-direction is incident on this system, an atom 'sees' the frequency to be doppler-shifted as

$$\omega_{obs} = \omega(1 + v/c)$$

and with ω_0 as the natural frequency of absorption, the atom can absorb the photon if $\omega_{obs} = \omega_0$, so that the observed absorption spectrum looks like

$$A(\omega)\ \mathrm{d}\omega = L\ \alpha\ \left[\int_{-\infty}^{+\infty} \delta(\omega(1+\frac{v}{c}) - \omega_0)N(v)\mathrm{d}v\right]\mathrm{d}\omega$$

α being an absorption cross-section. Alternatively, using $\omega = 2\pi c/\lambda$, where λ is the wavelength of light, we have the absorption spectrum as

$$A(\lambda)\ \mathrm{d}\lambda \approx \frac{LN_0\alpha}{\sqrt{2\pi|\Delta\lambda|^2}} \exp\left[-\frac{(\lambda-\bar{\lambda})^2}{2|\Delta\lambda|^2}\right]\mathrm{d}\lambda$$

where $\bar{\lambda} = \lambda_0(1+\bar{v}/c), \sigma^2 = k_BT/M$ where M is the mass of the hydrogen atom, $\lambda_0 = 2\pi c/\omega_0$, and $|\Delta\lambda|^2 = \lambda_0^2\ \sigma^2/c^2$.

Generate 'fake data' for $T = 6000K, \bar{v} = 5\text{km/sec}, \lambda_0 = 1216\ \overset{\circ}{\mathrm{A}}$ (Lyman-α), $1026\ \overset{\circ}{\mathrm{A}}$ (Lyman-β). Check if the data can be fitted to

(1) $\bar{v} = 3.5\text{km/sec}, T = 7000K$
(2) $\bar{v} = 5\text{km/sec}, T = 5000K$

The situation can be made more complex if we have two gas clouds in our line of sight with column lengths L_1 and L_2, densities N_1 and N_2, and temperatures T_1 and T_2. The absorption coefficients must then read

$$A(\lambda)\ \mathrm{d}\lambda = A_1(\lambda)\ \mathrm{d}\lambda + A_2(\lambda)\ \mathrm{d}\lambda$$

with

$$A_1(\lambda) = \frac{\alpha N_1 L_1}{\sqrt{2\pi|\Delta\lambda_1|^2}} \exp\left[-\frac{(\lambda-\bar{\lambda}_1)^2}{2|\Delta\lambda_1|^2}\right]$$

$$A_2(\lambda) = \frac{\alpha N_2 L_2}{\sqrt{2\pi|\Delta\lambda_2|^2}} \exp\left[-\frac{(\lambda-\bar{\lambda}_2)^2}{2|\Delta\lambda_2|^2}\right]$$

$$|\Delta\lambda_1|^2 = \lambda_0^2\sigma_1^2/c^2\ , |\Delta\lambda_2|^2 = \lambda_0^2\sigma_2^2/c^2$$

$$\sigma_1^2 = k_BT_1/M\ , \sigma_2^2 = k_BT_2/M$$

$$k_B = 1.38045 \times 10^{-16}\ \text{erg}/K$$

$$M = 1.6726 \times 10^{-24}\ \text{erg}/K$$

$$c = 3.0 \times 10^{10}\ \text{cm/sec}$$

$$1\ \overset{\circ}{\mathrm{A}} = 10^{-8}\ \text{cm}$$

In the last case, depending upon the choice on $N_1, N_2, \bar{v}_1, \bar{v}_2, T_1, T_2$ it may not be possible to distinguish whether there is one gas column or two.

,