



Application of Equal Local Levels to Improve Q-Q Plot Testing Bands with R Package qqconf

Eric Weine
University of Chicago

Mary Sara McPeck 
University of Chicago

Mark Abney 
University of Chicago

Abstract

Quantile-quantile (Q-Q) plots are often difficult to interpret because it is unclear how large the deviation from the theoretical distribution must be to indicate a lack of fit. Most Q-Q plots could benefit from the addition of meaningful global testing bands, but the use of such bands unfortunately remains rare because of the drawbacks of current approaches and packages. These drawbacks include incorrect global type-I error rate, lack of power to detect deviations in the tails of the distribution, relatively slow computation for large data sets, and limited applicability. To solve these problems, we apply the equal local levels global testing method, which we have implemented in the R Package **qqconf**, a versatile tool to create Q-Q plots and probability-probability (P-P) plots in a wide variety of settings, with simultaneous testing bands rapidly created using recently-developed algorithms. **qqconf** can easily be used to add global testing bands to Q-Q plots made by other packages. In addition to being quick to compute, these bands have a variety of desirable properties, including accurate global levels, equal sensitivity to deviations in all parts of the null distribution (including the tails), and applicability to a range of null distributions. We illustrate the use of **qqconf** in several applications: assessing normality of residuals from regression, assessing accuracy of p values, and use of Q-Q plots in genome-wide association studies.

Keywords: Q-Q plots, equal local levels, Kolmogorov-Smirnov, GWAS, multiple testing, global testing, simultaneous region.

1. Introduction

Quantile-quantile (Q-Q) plots (Wilk and Gnanadesikan 1968) are a common statistical tool used for judging whether a sample comes from a specified distribution, and, perhaps most usefully, for visualizing the particular ways in which the sample might seem to deviate from that distribution. Despite their ubiquity, they are often difficult to interpret because it is

challenging to determine whether the extent of the observed deviation from the specified distribution is sufficient to indicate a lack of fit as opposed to just being due to sampling variability. To aid in interpretation, it is useful to put goodness-of-fit testing bands on Q-Q plots.

A few methods have been created toward this end (reviewed by [Aldor-Noiman, Brown, Buja, Rolke, and Stine \(2013\)](#)). Naively, one could use a pointwise testing band, an approach that is equivalent to conducting a level- α test on each order statistic of the sample. However, because of the large number of tests, the probability that at least one data point lies outside the band is far higher than α , so the pointwise approach does little to help with the problem of interpretability. To appropriately deal with this multiple testing problem, the Kolmogorov-Smirnov (KS) statistic ([Kolmogorov 1941](#); [Smirnov 1944](#)) is sometimes used to create a simultaneous testing band for a Q-Q plot. While this method controls type-I error, the KS test suffers from very low power under a variety of reasonable alternatives because it has low sensitivity to deviation in the tails of the null distribution ([Aldor-Noiman et al. 2013](#); [Berk and Jones 1979](#); [Mason and Schuenemeyer 1983](#)).

To overcome this problem, one can instead apply the equal local levels (ELL) global testing method to create simultaneous testing bands for Q-Q plots. The ELL global testing method was originally introduced by [Berk and Jones \(1979\)](#) (their M_n^+ is a one-sided ELL test statistic) and further developed by [Gontscharuk and colleagues \(Gontscharuk, Landwehr, and Finner 2015; Gontscharuk, Landwehr, Finner et al. 2016; Gontscharuk and Finner 2017\)](#) as an improvement over the higher criticism ([Donoho and Jin 2015](#)) and KS global testing methods. To conduct the ELL global test at level α , one conducts a “local” (or pointwise) test at level η on each order statistic of the sample and rejects the global test whenever at least one of the local tests is rejected, where the local level η must be chosen so that the global level of the test is the desired value α . The fact that the same local level is applied to each order statistic means that the ELL testing band can be viewed as impartial in its sensitivity to deviations from different parts of the null distribution, a sensible choice for use in a generic tool such as a Q-Q plot. In the specific context of assessing normality with a Q-Q plot, [Aldor-Noiman et al. \(2013\)](#) proposed to apply ELL to create two-sided testing bands by using simulation to determine the value of η needed in each case, a method they called “tail sensitive” (TS) because it has more sensitivity in the tails than KS. Through a series of examples and simulations, they effectively demonstrate the superiority of ELL testing bands over KS for detecting deviations from normality in a Q-Q plot in a variety of cases of interest. An advantage of the simulation-based approach to computing ELL bands is that it gives a straightforward way to incorporate the effects of parameter estimation. However, such an approach is arguably too slow to be conveniently applied to large, modern datasets. Considering that the Q-Q plot is meant to be a handy visualization tool and not an end-goal of analysis, it is important that the bands be virtually instantaneous to compute on a laptop or they are unlikely to be widely used. In our ELL implementation, instead of simulation, we use pre-computation based on fast algorithms, supplemented with asymptotic approximations for sample sizes over 100K.

Until now, available software for putting testing bands on Q-Q plots has been limited. The base R package `stats` ([R Core Team 2023](#)) provides functionality for creating a Q-Q plot to compare a sample against the normal distribution, and with a bit more difficulty, one can create Q-Q plots for other distributions, but it does not provide any way to put testing bands on those plots. The package `qqplotr` ([Almeida, Loy, and Hofmann 2018](#)) provides a number

of helpful additions to the base R functionality, including the ability to easily create Q-Q plots for a variety of reference distributions, the ability to create simultaneous testing bands using KS for a variety of reference distributions, and the ability to create simultaneous testing bands using TS only for the normal reference distribution. However, because it is based on simulation, the TS approach can be a bit slow, taking several minutes to produce bands for a sample size in the tens of thousands.

Our development of the R package **qqconf** is motivated by two major unmet needs in obtaining testing bands for Q-Q plots: (1) the need for ELL testing bands for non-normal distributions, particularly the uniform distribution, and (2) the practical need for greater speed in obtaining ELL testing bands for Q-Q plots in all cases, including normal. Regarding (1), in addition to testing for normality, important uses of Q-Q plots include assessing accuracy of p values (Section 3.2) and applications in genomics (Section 3.3) both of which involve assessing uniformity, so it would be extremely useful to have ELL simultaneous testing bands for Q-Q plots for the uniform case in particular, as well as for other non-normal distributions in general. Regarding (2), in light of the demonstrated superiority of ELL over other approaches for creating testing bands for Q-Q plots, one of our major software goals is to make creation of ELL testing bands (at least for $\alpha = 0.05$ or 0.01) so fast that this approach can confidently be used as the default for all Q-Q plots, without concern for taxing the casual user's patience or processing resources.

In what follows, we introduce the R package **qqconf** (Weine, McPeck, and Mark 2023), available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=qqconf>, for making Q-Q and probability-probability (P-P) plots. The `get_qq_band` function in **qqconf** can quickly provide ELL testing bands for comparing even very large samples to any reference distribution with a quantile function (e.g., `qnorm`, `qchisq`) implemented. In addition to these testing bands, which can be output for use with other plotting packages, **qqconf** provides a variety of plotting functionalities that allow the user to easily visualize where any deviation of the sample from the null distribution may occur. In Section 2, we introduce the methods required for computation of ELL testing bands for Q-Q plots. In Section 3, we demonstrate the functionality of **qqconf** in applications including assessing normality of residuals from regression (Section 3.1), assessing accuracy of p values (Section 3.2), and use of Q-Q plots in genome-wide association studies (Section 3.3).

2. Methods

2.1. Local levels for global hypothesis testing

Our ELL method for creating appropriate simultaneous testing bands for Q-Q plots can be viewed as an application of the following more general testing framework. Suppose we have real-valued observations

$$X_1, \dots, X_n \stackrel{iid}{\sim} F,$$

with order statistics $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, and we are interested in conducting the following hypothesis test at level α :

$$H_0: F = F_0 \text{ vs. } H_A: F \neq F_0,$$

where we refer to H_0 as the “global null hypothesis” and α as the “global level”, and where F_0 is a known continuous distribution on \mathbb{R}^1 (or on some finite or infinite sub-interval of \mathbb{R}^1 such as $(-1,1)$ or $(0,\infty)$). For simplicity, we start by assuming that all parameters of F_0 are known (we relax this assumption later). One approach to this hypothesis testing problem, referred to as “local levels” (Gontscharuk *et al.* 2016), is to conduct n separate (“local”) hypothesis tests, one on each of the order statistics $X_{(1)}, \dots, X_{(n)}$, where the test on the i th order statistic has level η_i (the i th local level). Then, one rejects the global null hypothesis if at least one of the n local tests results in a rejection. That is, we construct a set of intervals

$$(h_1, g_1), \dots, (h_n, g_n),$$

where $h_i < g_i$ for $1 \leq i \leq n$, and under the null hypothesis, $P(X_{(i)} \notin (h_i, g_i)) = \eta_i$, and we reject H_0 if

$$X_{(i)} \notin (h_i, g_i) \text{ for at least one value of } i \text{ such that } 1 \leq i \leq n. \quad (1)$$

In this general setting, the level α of the global test is determined by the vectors of lower and upper interval endpoints, (h_1, \dots, h_n) and (g_1, \dots, g_n) and the null cdf F_0 .

2.2. Two-sided ELL

For the Q-Q plot application, we want to create level- α testing bands that are “agnostic” to any alternative distribution. By this, we mean that we would like to design a local levels test such that, firstly, the global test applies equal scrutiny to each order statistic, i.e., we set the local levels to be equal:

$$\eta_1 = \eta_2 = \dots = \eta_n = \eta, \quad (2)$$

and, secondly, the local tests give equal weight to deviations of F from F_0 in either direction, i.e., we choose

$$h_i = F_{0i}^{-1}(\eta/2) \text{ and } g_i = F_{0i}^{-1}(1 - \eta/2), \quad (3)$$

where F_{0i} is the cdf of the i th order statistic under the null hypothesis, which is easily obtained from F_0 (see, e.g., Section 5.4 of (Casella and Berger 2002)). We refer to the global test derived from local levels under conditions (2) and (3) as the two-sided ELL test.

The main difficulty in applying the two-sided ELL test is in determining the local level η that will result in the desired global level α . One nice property of the two-sided ELL is that the local level η needed to achieve global level α depends only on α and on the sample size n , and not on F_0 at all. This can be seen by noting that under the null hypothesis,

$$F_0(X_1), \dots, F_0(X_n) \stackrel{iid}{\sim} U(0, 1). \quad (4)$$

Thus, without loss of generality, we can take the null distribution to be $U(0, 1)$ and determine the needed η and the interval endpoints (h_1, \dots, h_n) and (g_1, \dots, g_n) for this case. To convert back to the original scale, all that is needed is to apply F_0^{-1} to each of the resulting interval endpoints.

2.3. Calculation of the local level for two-sided ELL

Given the sample size n and the desired global level $0 < \alpha < 1$, we define $\eta_n(\alpha)$ to be the local level η that will result in global level α for the ELL test. Note that $\eta_n(\alpha)$ is a continuous,

monotone increasing function of α , and we denote its inverse by $\alpha_n(\eta)$. Given n and α , the basic approach to obtaining $\eta_n(\alpha)$ involves a binary search over $\eta \in (0, 1)$, where for each value of η , we obtain (h_1, \dots, h_n) and (g_1, \dots, g_n) via Equation 3, and then we calculate $\alpha_n(\eta)$, the probability of the event described in Equation 1, i.e., we find that probability that $(X_{(1)}, \dots, X_{(n)})$ falls outside the region $(h_1, g_1) \times \dots \times (h_n, g_n)$. Then we perform a binary search to find the η such that $\alpha_n(\eta) = \alpha$, the desired global level.

To calculate $\alpha_n(\eta)$, several recursive approaches have previously been developed (see [Shorack and Wellner \(2009\)](#)), as well as a fast Fourier transform (FFT) based approach ([Moscovich and Nadler 2016](#)). In **qqconf**, we apply the method of [Moscovich and Nadler \(2016\)](#), as implemented in [Moscovich \(2020b\)](#), which can be used by the ELL method to obtain simultaneous Q-Q plot testing bands at global level α for any n and α . In addition, **qqconf** offers a faster approximate approach specifically for the most commonly-used global levels of $\alpha = 0.05$ and 0.01 . To do this we have applied our own recursive formula (Appendix A) for obtaining $\alpha_n(\eta)$ in order to generate look-up tables for $\eta_n(\alpha)$ for $\alpha = 0.05$ and 0.01 with sample sizes n up to 1 million and 500K, respectively, where the tables are relatively dense for n up to 100K. If the user inputs $\alpha = 0.05$ or 0.01 with a value of n less than or equal to 100K, we either return back the pre-computed value of η if n happens to be a grid point, or we use linear interpolation if the value of n is between grid points, which leads to a highly accurate approximation. If the user inputs a value of n greater than 100K with $\alpha = 0.05$ or 0.01 , we use the asymptotic approximation given in Section 2.4. This allows **qqconf** to provide essentially instantaneous simultaneous testing bands for the cases $\alpha = 0.05$ and 0.01 for any reference distribution with quantile function implemented, with the FFT approach ([Moscovich and Nadler 2016](#)) used primarily for fast “on-the-fly” calculations with other choices of α .

2.4. Local level approximations in large samples

For sufficiently large values of the sample size n (or, equivalently, the number of local tests), it can be expedient to apply an accurate asymptotic approximation of $\eta_n(\alpha)$ in place of exact computation. Previous authors ([Gontscharuk and Finner 2017](#)) showed that an asymptotic approximation of $\eta_n(\alpha)$ is

$$\eta_{\text{asympt}} = \frac{-\log(1 - \alpha)}{2 \log(\log(n)) \log(n)}.$$

However, as they note, this approximation gives poor performance for n even as large as 10^4 (see Figure 1 of [Gontscharuk and Finner \(2017\)](#)). To improve this approximation, they propose to add a smaller order correction term, resulting in an approximation of the form

$$\eta_{\text{approx}} = \frac{-\log(1 - \alpha)}{2 \log(\log(n)) \log(n)} \left[1 - c_\alpha \frac{\log(\log(\log(n)))}{\log(\log(n))} \right], \quad (5)$$

where c_α is chosen empirically. For the values $\alpha = 0.01, 0.05$, and 0.1 they chose $c_\alpha = 1.6, 1.3$, and 1.1 , respectively. To select these c_α values, the authors calculated the values of $\eta_n(\alpha)$ to high precision on a grid of values up to $n = 10,000$.

We performed more extensive tests of these approximations for the cases $\alpha = 0.01$ and 0.05 . To do this, we calculated the values of $\eta_n(\alpha)$ with high precision on a grid of values up to $n = 500,000$ for $\alpha = 0.01$ and up to $n = 10^6$ for $\alpha = 0.05$. Based on our evaluation, we find that $c_\alpha = 1.3$ is satisfactory for $\alpha = 0.05$, but that $c_\alpha = 1.6$ for $\alpha = 0.01$ is not sufficiently accurate for our purposes. We instead found that $c_\alpha = 1.591$ led to better performance for

$\alpha = 0.01$. For example, for n in the range of 15K to 500K, the absolute relative error in the approximation based on $c_\alpha = 1.6$ is always more than 0.0067, while that based on $c_\alpha = 1.591$ is always less than 0.001.

We implement these asymptotic approximations in **qqconf** as part of our faster approximate approach specifically for $\alpha = 0.01$ and 0.05 with $n > 100K$, as described in Section 2.3. (Our package also implements the approximation given in Equation 5 for $\alpha = 0.1$ with $c_\alpha = 1.1$.)

2.5. One-sided ELL

In some instances, a one-sided version of ELL is of particular interest. For example, suppose X_1, \dots, X_n are p values, with X_i representing the p value of the i th hypothesis test, which has corresponding null hypothesis $H_0^{(i)}$, where X_1, \dots, X_n are assumed to be independent, with $X_i \sim U(0,1)$ if $H_0^{(i)}$ is true. Suppose we are interested in testing the global null hypothesis H_0 : all of $H_0^{(1)}, \dots, H_0^{(n)}$ are true against the alternative H_A : at least one of $H_0^{(1)}, \dots, H_0^{(n)}$ is false. Within the equal local levels framework, we would typically do this by assuming

$$X_1, \dots, X_n \stackrel{iid}{\sim} F,$$

and testing the null hypothesis $H_0: F(x) = x$ for all $x \in (0, 1)$ vs. the one-sided alternative $H_A: F(x) > x$ for at least one $x \in (0, 1)$. In this case, a one-sided test is commonly used because one is typically only interested in p values that are smaller than expected, not larger than expected. This is exactly the context considered in Berk and Jones (1979), in which the ideas behind ELL were first laid out.

More generally, one could test

$$H_0: F = F_0 \text{ for all } x \in \mathbb{R} \text{ vs. } H_A: F > F_0 \text{ for some } x \in \mathbb{R}.$$

In this context, a one-sided global test of H_0 based on local levels η_1, \dots, η_n would involve first constructing a set of lower bounds h_1, \dots, h_n , where

$$h_i = F_{0i}^{-1}(\eta_i), \tag{6}$$

and then rejecting if

$$X_{(i)} < h_i \text{ for at least one value of } i \text{ such that } 1 \leq i \leq n.$$

We define the one-sided ELL test with global level α to be the test of this type obtained by setting $\eta_1 = \dots = \eta_n = \eta$ and choosing η to obtain global level α .

Given the sample size n and the desired global level $0 < \alpha < 1$, we define $\eta'_n(\alpha)$ to be the local level η that will result in global level α for the one-sided ELL test. As in the two-sided case, we denote the inverse function of $\eta'_n(\alpha)$ by $\alpha'_n(\eta)$. Given n and α , we obtain $\eta'_n(\alpha)$ by a binary search over $\eta \in (0, 1)$, where for each value of η , we obtain (h_1, \dots, h_n) via Equation 6, and then we calculate $\alpha'_n(\eta)$, the probability that $(X_{(1)}, \dots, X_{(n)})$ falls outside the region $(h_1, 1) \times \dots \times (h_n, 1)$. Then we perform a binary search to find the η such that $\alpha'_n(\eta) = \alpha$, the desired global level.

To calculate $\alpha'_n(\eta)$, several approaches have previously been developed (Shorack and Wellner 2009; Moscovich 2020a). **qqconf** currently uses the method of Moscovich and Nadler (2016)

Sample size	Sample sd	Empirical type-I error (se) when using			True
		MAD	Q_n	S_n	
100	0.0011 (0.0003)	0.0963 (0.0030)	0.0278 (0.0016)	0.0427 (0.0020)	0.0522 (0.0022)
500	0.0035 (0.0006)	0.1015 (0.0030)	0.0260 (0.0016)	0.0443 (0.0021)	0.0514 (0.0022)
10,000	0.0111 (0.0010)	0.1040 (0.0031)	0.0327 (0.0018)	0.0498 (0.0022)	0.0480 (0.0021)

Table 1: Empirical type-I error at nominal level 0.05 for testing normality with different parameter estimation methods, based on 10^4 simulation replicates.

as implemented in [Moscovich \(2020b\)](#). We have also implemented two recursive approaches, described in [Appendix B](#): an exact version and an approximate version that is much faster and bounds the relative error in the reported global significance level to a tolerance set by the user.

2.6. Additional implementation issues

To create the “expected” quantiles for a Q-Q plot, we apply the inverse cdf F_0^{-1} to a set of probability points. For the normal distribution, it has been shown ([Blom 1958](#)) that the means of the order statistics of n i.i.d. draws are well-approximated by the above process when `ppoints(n)` is used to generate the probability points, while for the uniform distribution, the means of the order statistics are obtained exactly when `ppoints(n, a = 0)` is used. For other distributions, appropriate approximations to the means of the order statistics could be obtained on a case-by-case basis. (Because a P-P plot is basically a variation on a uniform Q-Q plot, the exact mean probability points for a P-P plot are obtained for all distributions by `ppoints(n, a = 0)`.) For creating the “expected” line in a Q-Q plot, we propose the medians of the order statistics as a useful alternative to their means. Exact medians of the order statistics for i.i.d. draws from any distribution can easily be obtained by applying the inverse cdf to `qbeta(0.5, 1:n, n:1)`. The resulting “expected” line is the unique line that is guaranteed to lie completely within the ELL band, regardless of the global level α or the distribution. All 3 of the above expected lines are options within `qqconf`.

The most commonly-encountered uses of Q-Q plots are to assess normality in various contexts and to assess uniformity of p values for a set of independent hypothesis tests, and we give examples of both in [Section 3](#). When assessing normality, typically the mean μ and standard deviation σ would not be known but would need to be estimated from the data in order to make either an “expected” line or any kind of testing band for a Q-Q plot. For example, in base R the function `qqline` makes an expected line that by default passes through the first and third quartiles, which is equivalent to estimating μ by the mid-quartile and σ by the inter-quartile range multiplied by 0.7413.

In `qqconf`, the default is to estimate μ by the median and σ by the estimator S_n of [Rousseeuw and Croux \(1993\)](#), where S_n is a highly robust scale estimator with very low gross-error sensitivity that is more efficient than median-absolute-deviation (MAD) and approximately unbiased even in small sample sizes. To validate this choice, we performed simulation studies under the null hypothesis of normality and assessed the type 1 error of the 5% rejection bounds generated by ELL, where we used one of 5 choices for (μ, σ) : (1) sample mean and sample s.d., (2) sample median and sample MAD, (3) sample median and Q_n , another estimator of σ discussed by [Rousseeuw and Croux \(1993\)](#), (4) sample median and S_n , and (5) the true values of μ and σ for comparison, and where these are denoted in [Table 1](#) by ‘sample sd’, ‘MAD’,

‘ Q_n ’, ‘ S_n ’ and ‘true’, respectively. We note that the entire simulation study is invariant to the choice of true mean and s.d., because these just become location and scale factors for all the data and the estimators and therefore cancel out in the type-I error assessment. The results for $n = 100, 500$ and 10^4 are given in Table 1, where we can see that using median and S_n gives type-I error very close to the nominal level, though slightly conservative for small sample sizes. Methods to handle parameter uncertainty with an exact calculation (as opposed to simulation) have been discussed in the context of the normal distribution (Rosenkrantz 2000), but a general method towards this end has not been developed. Use of Q-Q plots for distributions other than normal for which the parameters are unknown is rarer, and for those cases the current default in **qqconf** is maximum likelihood estimation, though the user can replace that with an estimate of their choice. Note that in applications such as assessing uniformity of p values (Section 3.2) or in the genomics example in Section 3.3, no parameter estimation is required.

3. Examples

One of the main advantages of the local levels method compared to other global testing approaches is that it can easily be used to put testing bands onto Q-Q plots by simply graphing each (h_i, g_i) interval. This allows us to examine how a dataset might deviate from some null distribution much better than simply applying a test that yields a binary conclusion. Below, we present a few examples where a Q-Q plot is useful, and where the local levels test seems ideal for assessing deviation from a global null hypothesis.

3.1. Assessing normality of residuals from regression

When performing an ordinary least squares (OLS) regression, it is common to assume that the error terms are independently drawn from a normal distribution, e.g., $Y = X\beta + \epsilon$, where $Y_{n \times 1}$ and $X_{n \times p}$ are observable, $\beta_{p \times 1}$ is an unknown parameter vector, and conditional on X , $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$ is assumed to satisfy

$$\epsilon_1, \dots, \epsilon_n \stackrel{iid}{\sim} N(0, \sigma^2). \quad (7)$$

After obtaining the OLS estimate $\hat{\beta}$ and the residual vector $r = Y - X\hat{\beta}$, we would like a Q-Q plot of the residuals with a normal reference distribution to aid in testing assumption (7) above. Without prior reason to believe that the errors may deviate from normality at any particular point in the distribution, it makes sense to use ELL bands in this case. This is very easy to do with **qqconf**, as we show below.

First, we generate data to perform a regression. Here, we generate each ϵ_i independently from a $t(3)$ distribution.

```
R> set.seed(20)
R> n <- 100
R> x <- runif(n)
R> eta <- rt(n, df = 3)
R> y <- x + eta
```

Then, we fit a regression with the simulated data

```
R> reg <- lm(y ~ x)
```

Figure 1 shows a Q-Q plot created with base R functionality using the functions `qqnorm` and `qqline`, as follows:

```
R> qqnorm_plot <- qqnorm(reg$residuals)
R> qqline(reg$residuals)
```

Clearly there is some indication of deviation from normality in Figure 1, but it can be hard to tell how significant the deviation is without a testing band. In Figure 2, we improve upon this by using `qq_conf_plot` to create a Q-Q plot with a 0.05-level ELL testing band as follows:

```
R> qq_conf_plot(obs = reg$residuals,
+   points_params = list(col = "blue", pch = 20, cex = 0.5))
```

In Figure 2, we can clearly see that both the left and right tails of the residuals go beyond the normal testing bounds, giving strong evidence that the errors were not generated from a normal distribution.

If the user prefers to use another plotting software, **qqconf** also provides a separate interface, `get_qq_band`, for obtaining the testing band itself. The `band_method` argument of `get_qq_band` allows for ELL, KS or pointwise bands to be created. The band computed by `get_qq_band` can easily be used with, e.g., base R's `qqnorm` as below, producing Figure 3. When adding a testing band to a Q-Q plot produced outside of **qqconf**, in order to have the correct type-I error for the band, it is essential that the same x -coordinates be used for plotting both the data points and the bound points for the band. This is accomplished in Figure 3 by use of `sort(qnorm_plot$x)` as the x -coordinates for the upper and lower bounds for the band in the code below, as these were the x -coordinates used by `qqnorm` to plot the data points. Figure 3 is generated as follows:

```
R> band <- get_qq_band(obs = reg$residuals)
R> plot(qnorm_plot, col = "blue", pch = 20, cex = 0.5,
+   xlab = "Expected quantiles", ylab = "Observed quantiles")
R> lines(sort(qnorm_plot$x), band$lower_bound, col = "red")
R> lines(sort(qnorm_plot$x), band$upper_bound, col = "red")
R> qqline(qnorm_plot$x, datax = TRUE, distribution =
+   function(p) qqnorm(p, mean = band$dparams$mean, sd = band$dparams$sd))
```

Moreover, if the user prefers to use **qqplotr** (Almeida *et al.* 2018), this can also be done easily, as shown in Figure 4. Again, it is critical that the same x -coordinates be used to plot both the points and the bounds of the band. This is accomplished in the code below for Figure 4 by setting `band_df$expected` equal to `build_plot$data[[1]]$x`, which puts the x -coordinates that will be used to plot the points into `band_df$expected`, and then using `x = expected` as an argument to `aes` in the call to `geom_ribbon` that creates the testing band. Figure 4 is generated as follows:

```
R> band_df <- data.frame(lower = band$lower_bound, upper = band$upper_bound,
+   obs = reg$residuals)
```

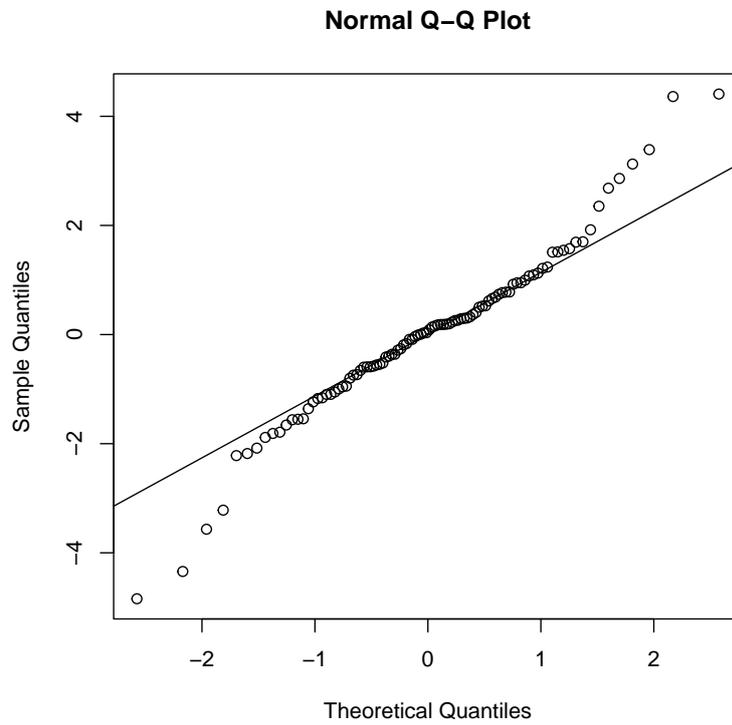
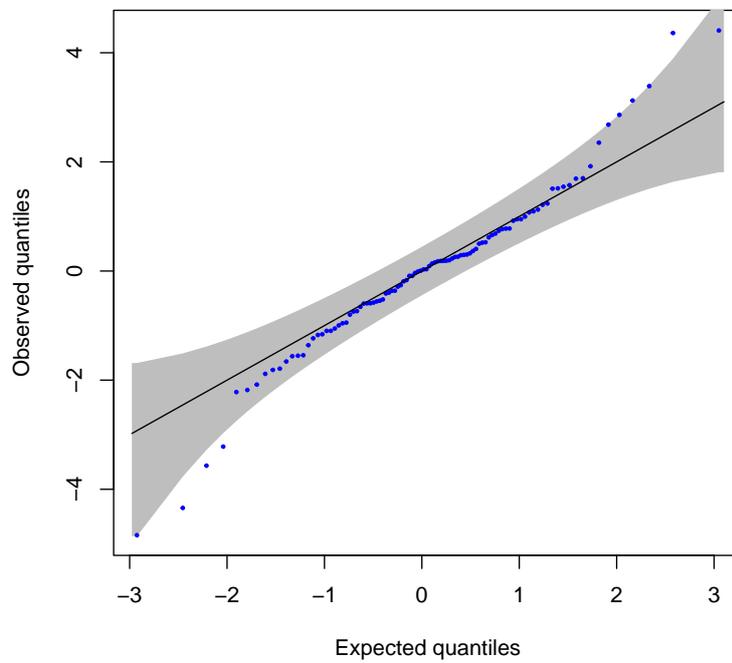


Figure 1: Q-Q plot for regression residuals with base R functionality.

Figure 2: Q-Q plot for regression residuals with ELL bounds using **qqconf**.

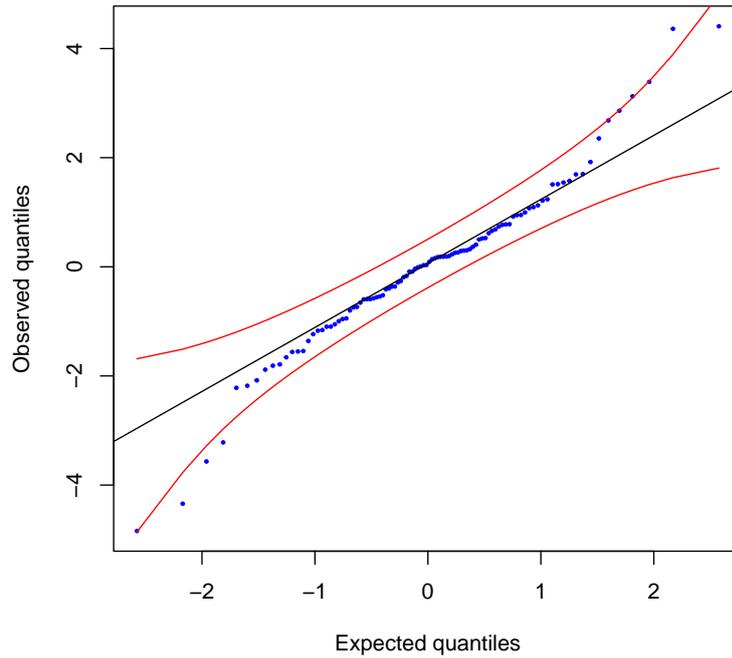


Figure 3: Q-Q plot for regression residuals with ELL bands from **qqconf** added to a plot made with base R **qqnorm**.

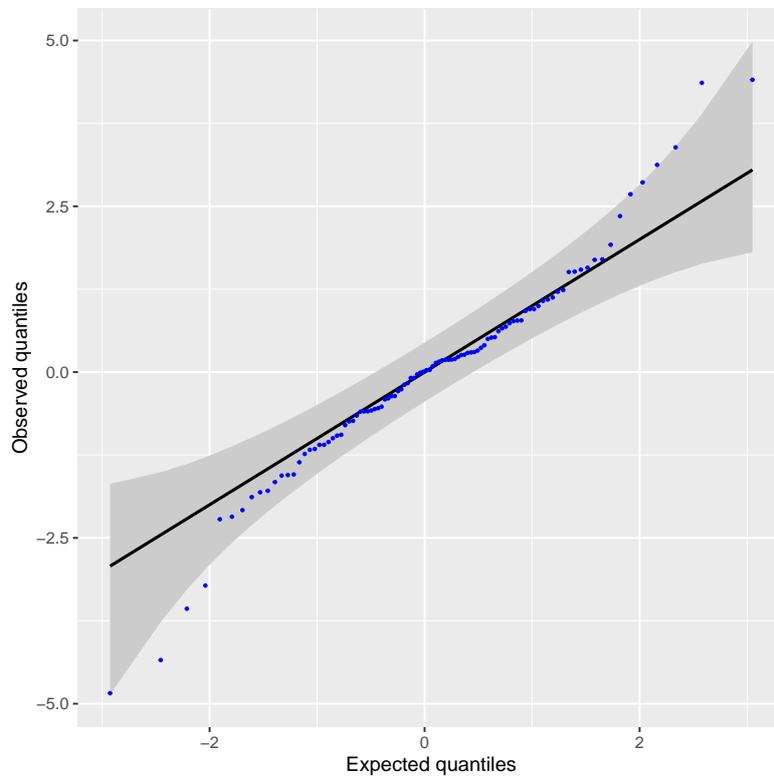


Figure 4: Q-Q plot for regression residuals with with ELL bands from **qqconf** added to a plot made with **qqplotr**.

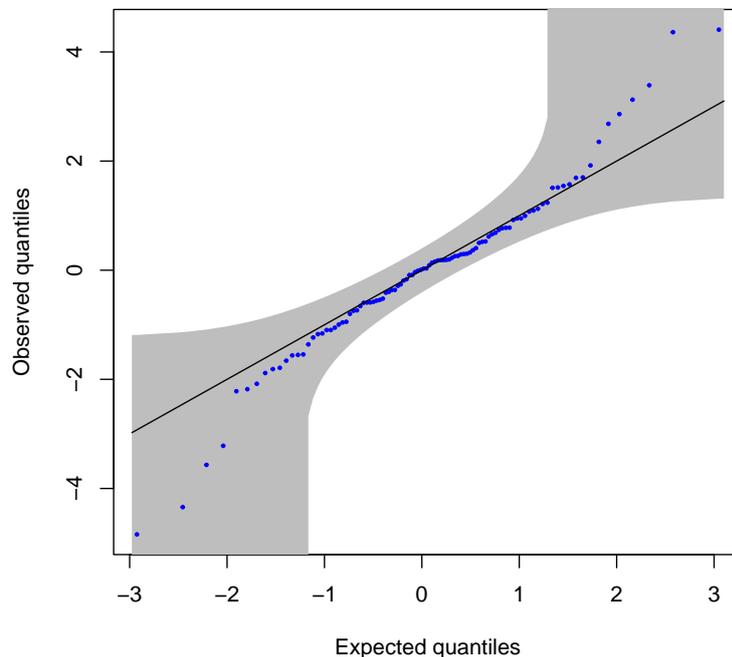


Figure 5: Q-Q Plot for regression residuals with KS bounds.

```
R> build_plot <- ggplot2::ggplot_build(
+   ggplot2::ggplot(band_df, ggplot2::aes(sample = obs)) +
+   qqplotr::stat_qq_point(dparams = band$dparams))
R> band_df$expected <- build_plot$data[[1]]$x
R> ggplot2::ggplot(band_df, ggplot2::aes(sample = obs)) +
+   ggplot2::geom_ribbon(
+     ggplot2::aes(ymin = lower, ymax = upper, x = expected),
+     fill = "grey80") +
+   qqplotr::stat_qq_line(dparams = band$dparams, identity = TRUE) +
+   qqplotr::stat_qq_point(dparams = band$dparams, color = "blue",
+     size = 0.5) +
+   ggplot2::xlab("Expected quantiles") +
+   ggplot2::ylab("Observed quantiles")
```

This example also highlights the advantages of the ELL method over KS. Because KS is much more sensitive to deviations in the center of the distribution than it is to deviations in the tails of the distribution, it does not yield a rejection of the null hypothesis in this case (Figure 5). We generate the KS bounds in Figure 5 by simply setting the `method` argument to "ks" in the `qq_conf_plot` function as follows:

```
R> qq_conf_plot(obs = reg$residuals, method = "ks",
+   points_params = list(col = "blue", pch = 20, cex = 0.5))
```

3.2. Q-Q plots for assessing accuracy of p values

Suppose we have devised a new testing procedure to test a null hypothesis H_0 with test statistic T , where we also specify a particular method to calculate or approximate p values. In such a situation it is important to perform some simulations under the null hypothesis and check that the resulting p value distribution is approximately uniform in the simulation experiment.

Typically, the verification of type-I error rate is done using the following procedure:

1. Generate n simulated datasets under H_0 , and calculate T for each simulated dataset to obtain T_1, \dots, T_n .
2. Select a value of α , and for each of T_1, \dots, T_n , determine whether the null hypothesis is rejected at level α . Let N_α be the observed number of the n tests that are rejected at level α .
3. Let α^* denote the true probability of rejection under the above procedure. Test the null hypothesis $H_0: \alpha^* = \alpha$ by applying, e.g., a Z -test of proportions or an exact binomial test to the data N_α .

While the above procedure provides reliable information about the type-I error calibration for one level of α , it provides little information about the global calibration of p values. To obtain a useful visualization of the overall performance of the p value calculation method, we instead suggest the following procedure:

1. As above.
2. For each T_i , calculate the corresponding p value, p_i , to obtain p_1, \dots, p_n .
3. Make a Q-Q plot comparing p_1, \dots, p_n to a $U(0, 1)$ distribution, and apply the local levels procedure to create a simultaneous testing band for the null hypothesis that $p_1, \dots, p_n \stackrel{iid}{\sim} U(0, 1)$.

This allows us to easily visualize the global calibration of the p values with just one graph and diagnose any issues if they exist. In Step 3, one could use many different testing bands. However, in the calibration of p values, we typically do not have the expectation that our p values would be more likely to deviate from uniform in any particular region, and so it makes sense to use the local levels test because it is agnostic to the space of alternative distribution. Moreover, since it is generally most concerning if small p values are not calibrated (i.e., those in the lower tail of the uniform distribution), the local levels test is preferable to the standard KS test because it is much more sensitive in the tails (Aldor-Noiman *et al.* 2013).

χ^2 test for independence in a 2×2 table

We apply this approach to assess the calibration of p values from the Pearson χ^2 test for independence in a 2×2 table. A well-known rule of thumb is that the χ^2 test is appropriate as long as the expected cell count in each cell under the null hypothesis is at least 5. We fix the cell probabilities under the null hypothesis and consider two different cases: in scenario 1, the sample size is $s = 200$ and the condition in the rule of thumb holds (i.e., the minimum

expected cell count exceeds 5), and scenario 2, the sample size is only $s = 20$ and the condition in the rule of thumb does not hold. We use the local levels approach to generate simultaneous testing bands to assess the calibration of the p values from the Pearson χ^2 test for these two scenarios.

More specifically, in each scenario, we randomly generate $n = 1000$ 2×2 tables under the null hypothesis, where each table contains s observations, with $s = 200$ in scenario 1 and $s = 20$ in scenario 2. We generate the observations in each table as i.i.d. multinomial with given cell probabilities, which is natural since we want to investigate the effect of small expected cell counts. (Furthermore, the i.i.d. multinomial setting is one in which the asymptotic distribution for the test statistic is χ_1^2 .) For each table, the s observations are i.i.d. with probability $q_{i,j}$ of falling in cell (i, j) , for $i = 0, 1, j = 0, 1$, where $q_{1,1} = a \cdot b$, $q_{1,0} = a \cdot (1 - b)$, $q_{0,1} = (1 - a) \cdot b$, and $q_{0,0} = (1 - a) \cdot (1 - b)$, with $a = 0.15$ and $b = 0.4$. For each table, let $X_{i,j}$ denote the observed count in cell (i, j) . (If any table has $X_{0,0} + X_{0,1} = 0$ or $= s$, we discard the table and draw a new one, because that would imply that one of the rows of the table is empty, in which case the Pearson χ^2 test statistic is not defined. Similarly, if any table has $X_{0,0} + X_{1,0} = 0$ or $= s$, we discard the table and draw a new one.) For each of the tables in the resulting sample, a Pearson χ^2 test statistic T for independence is calculated. For each scenario, this results in $n = 1000$ test statistics, T_1, \dots, T_n , one for each table. From these, we obtain $n = 1000$ p values, p_1, \dots, p_n by applying the χ_1^2 approximation, i.e., $p_i = 1 - F(T_i)$ for $i = 1, \dots, n$, where F is taken to be the cdf of the χ_1^2 distribution.

Figure 6 shows the resulting Q-Q plots for scenarios 1 (in blue) and 2 (in vermillion), where the 45° line is shown as well as the testing band obtained from the equal local levels procedure for testing, at global level 0.05, the null hypothesis that p_1, \dots, p_n have the same distribution as n i.i.d. draws from $U(0,1)$. In Figure 6, the Q-Q plot for scenario 2 is made first, and then the Q-Q plot for scenario 1 is added to the same axes by setting the `add` argument of `qq_conf_plot` to `TRUE`, as follows:

```
R> pvals_scenario_1 <- scan("Data/pvals_scenario_1", quiet = TRUE)
R> pvals_scenario_2 <- scan("Data/pvals_scenario_2", quiet = TRUE)
R> par(pty = "s")
R> qq_conf_plot(obs = pvals_scenario_2, distribution = qunif,
+   points_params = list(
+     col = palette.colors(palette = "Okabe-Ito")["vermillion"],
+     type = "l"),
+   asp = 1)
R> qq_conf_plot(obs = pvals_scenario_1, distribution = qunif,
+   points_params = list(
+     col = palette.colors(palette = "Okabe-Ito")["blue"],
+     type = "l"),
+   add = TRUE)
R> legend("topleft", legend = c("s=200", "s=20"),
+   col = c(palette.colors(palette = "Okabe-Ito")["blue"],
+     palette.colors(palette = "Okabe-Ito")["vermillion"]),
+   lty = 1)
```

When assessing p values, typically the lower tail is of most interest, but this part of the plot is difficult to see when the plot axes are on the original scale. To focus the visualization on the

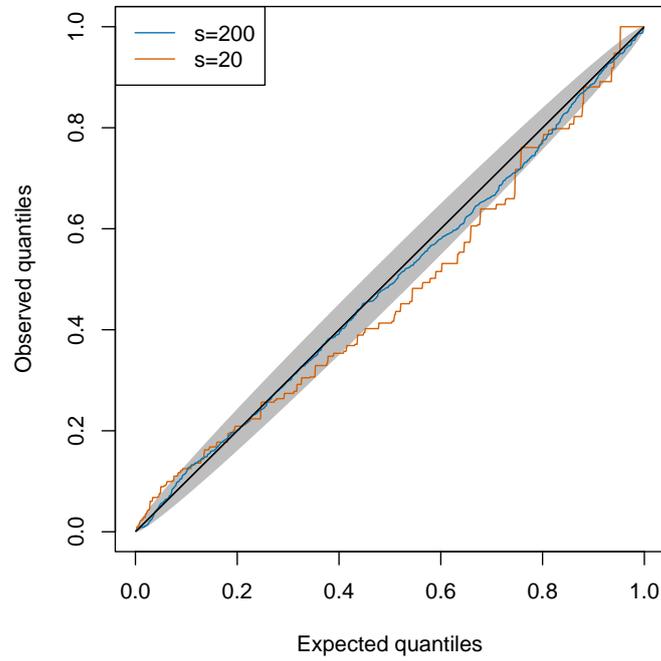


Figure 6: Q-Q plots for scenarios 1 ($s = 200$) and 2 ($s = 20$) with level 0.05 testing band and standard axes.

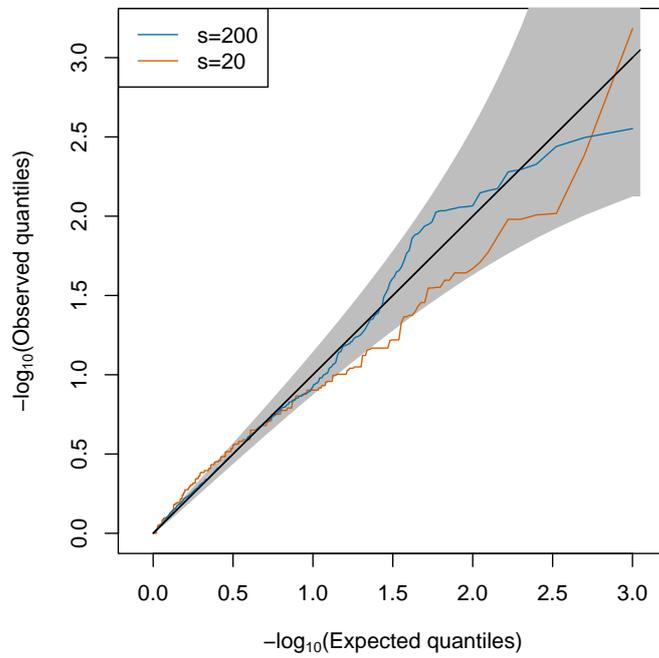


Figure 7: Q-Q plots for scenarios 1 ($s = 200$) and 2 ($s = 20$) with axes on the $-\log_{10}$ scale.

small p values we can plot the axes on the $-\log_{10}$ scale, as in Figure 7, by setting the `log10` argument of `qq_conf_plot` to `TRUE`. (Note that in Figure 7, small p values are to the top and right of the plot, so a curve that is too low is conservative, and too high is anti-conservative.) Figure 7 is generated as follows:

```
R> par(pty = "s")
R> qq_conf_plot(obs = pvals_scenario_2, distribution = qunif,
+   points_params = list(
+     col = palette.colors(palette = "Okabe-Ito")["vermillion"],
+     type = "l"),
+   log10 = TRUE, asp = 1)
R> qq_conf_plot(obs = pvals_scenario_1, distribution = qunif,
+   points_params = list(
+     col = palette.colors(palette = "Okabe-Ito")["blue"],
+     type = "l"),
+   log10 = TRUE, add = TRUE)
R> legend("topleft", legend = c("s=200", "s=20"),
+   col = c(palette.colors(palette = "Okabe-Ito")["blue"],
+     palette.colors(palette = "Okabe-Ito")["vermillion"]),
+   lty = 1)
```

From Figures 6 and 7, it can be seen that in scenario 1, when $s = 200$ and the smallest expected cell count is 12, there is no significant deviation of the p values from i.i.d. $U(0,1)$ under the null hypothesis. In contrast, in scenario 2, when $s = 20$ and the smallest expected cell count is 1.2, the χ_1^2 asymptotic distribution is not an accurate approximation to the sampling distribution of T . As a result, we can see in Figures 6 and 7 that the p values differ significantly from i.i.d. $U(0,1)$ under the null hypothesis, with small p values tending to be overly conservative, while the larger p values tend to be anti-conservative.

As a side note, to put our simulation results into the context of past theoretical work on the Pearson χ^2 test for independence, the following points from Lewis, Saunders, and Westcott (1984) are helpful: (1) Conservativeness or anti-conservativeness of the χ^2 approximation can vary with p -value, which is consistent with our simulation results. (2) Conservativeness or anti-conservativeness of the χ^2 approximation can be predicted theoretically based on the marginal totals, which vary across our simulations. However, if we analyze the expected marginal totals under scenario 2, then guideline II in Section 6 of Lewis *et al.* (1984) predicts that the χ^2 approximation will tend to be conservative in that setting, which agrees with what we observed in simulations for small p -values.

3.3. Q-Q plots for p values from genome-wide association studies

The goal of a genome-wide association study (GWAS) is to identify genetic variants that influence a trait (where a trait is commonly a disease or some other measured variable such as blood pressure or blood glucose level). For each individual in the sample, trait data are collected as well as genetic data on a large number of single-nucleotide polymorphisms (SNPs) throughout the genome. Based on these data, a statistical test is typically performed for each SNP to assess whether it is associated with the trait, resulting in a large number of p values, one for each SNP. Then a very stringent multiple testing correction is applied in order to

declare a result for a SNP to be significant. As part of the data analysis, a Q-Q plot is commonly presented to visualize and assess the distribution of genome-wide p values.

The implicit null hypothesis being assessed in such a Q-Q plot is H_0 : none of the tested SNPs is associated with the trait. If the SNPs could be assumed to be independent, then this null hypothesis would correspond to the p values being i.i.d. $U(0,1)$ random variables. One could argue for either a one-sided or two-sided alternative. The one-sided alternative would be that there is an excess of small p values (which is equivalent to $F(x) > x$ for some $x \in (0,1)$ in the notation of Section 2.5), which would be biologically interpretable as indicating that at least one SNP was associated with the trait. The two-sided alternative would simply be that the distribution is non-uniform. While an excess of large p values would not have any particular biological interpretation, it could indicate a problem with the data analysis, e.g., use of an inappropriate statistical test or the unexpected failure of assumptions underlying the test used.

In fact, there is local correlation among genome-wide SNPs, which decays very rapidly with distance as a result of genetic recombination. Typically, some “pruning” is done on the genome-wide SNPs prior to analysis so that the remaining SNPs in each small, local region are less correlated. While the remaining SNPs have a local correlation structure, this has only a weak effect at a genome-wide scale, and a Q-Q plot with appropriate simultaneous bounds can still provide a valuable visualization tool to assess the extent and type of deviation from the null. When substantial correlation remains, an alternative is to create the testing band based on an “effective number” of independent SNPs, `neff`. This could be done by setting `n=neff` in `get_qq_band`. (In this case, different x -coordinates would obviously need to be used for plotting the bounds of the band than for plotting the points.)

In GWAS, if the p values deviate from the null, it can be very useful to view graphically how they deviate. For instance, if a few tests yield unusually small p values but the p values from the bulk of the tests look relatively uniform, this suggests that the genetic variation affecting the trait of interest is likely driven by a relatively small number of genetic variants. If, however, there are some small deviations from uniformity throughout the p value distribution, this could indicate that the trait of interest is affected by a large number of genetic variants that all play some small part in a complex biological process, or it could potentially indicate that there are some confounding variables that are not controlled for.

Either of the above two scenarios, representing very different alternative distributions, could commonly arise in a GWAS, so the ELL method is a desirable choice for a putting testing bands on the Q-Q plot because it is agnostic to the choice of alternative distribution. Moreover, since small p values are often of great interest in GWAS because they can indicate the genetic variants that have the greatest influence on the trait, the use of ELL testing bands is far superior to use of KS testing bands because of the comparatively greater tail sensitivity of the former.

Application of equal local levels to Creutzfeld-Jakob Disease

We downloaded the p values from a GWAS of Creutzfeld-Jakob disease (CJD) in a sample of 4,110 cases and 13,569 controls (Jones *et al.* 2020). Tests of association between risk for the disease and genetic variants were done at 6,314,492 SNPs. Major genetic risk loci were found on chromosomes 1 and 20. Here, we remove those chromosomes from the results in order to be able focus on parts of the genome where we remain uncertain about to what

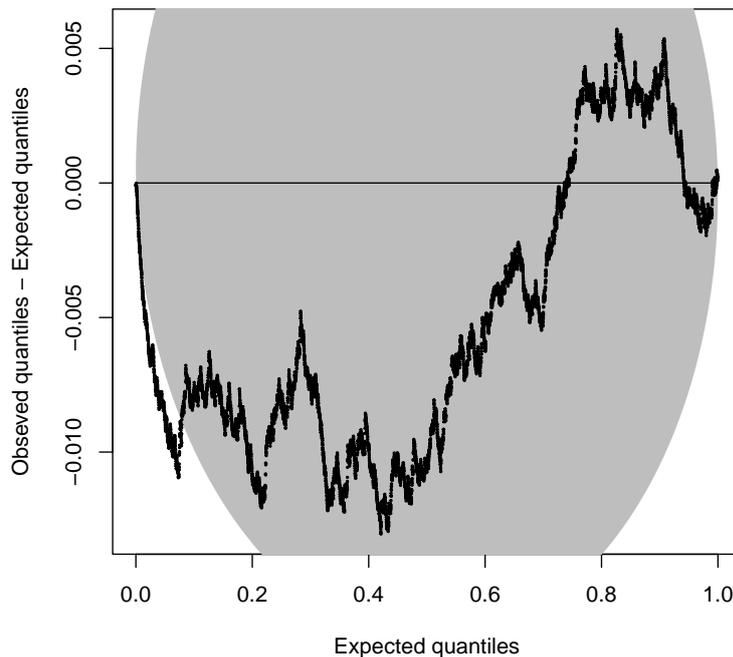


Figure 8: Differenced Q-Q plot of CJD GWAS p values.

extent risk variants are present. For convenience, we subsample the remaining SNPs to 10,000 approximately evenly spaced SNPs, which also helps ensure that correlation between SNPs is minimized.

```
R> cjd_df <- read.table("Data/cjd_sample.txt", header = TRUE)
```

We then make Q-Q plots of these 10,000 p values with 0.05-level testing bands. Note that for large datasets, a Q-Q plot with standard axes is undesirable, because, e.g., the 0.05-level testing band becomes extremely close to the diagonal as n grows, so generally all the interesting information in the plot is more-or-less collapsed onto the diagonal, rendering it less effective as a visual tool. For better visualization in a large dataset, we recommend instead plotting the difference between the observed and expected quantiles versus the expected quantiles, which we call a “differenced” Q-Q plot (alternatively called a “detrended” Q-Q plot by Almeida *et al.* 2018). Such a plot can easily be created by setting the `difference` argument to `TRUE` in `qq_conf_plot`. Figure 8, which depicts the differenced Q-Q plot for the CJD data, is produced as follows:

```
R> qq_conf_plot(obs = cjd_df[,3], distribution = qunif,
+   points_params = list(pch = 21, cex = 0.2), difference = TRUE)
```

In a GWAS the lower tail of the p value distribution would typically represent the most important genetic variants, so to highlight this region, we can use the `log10` argument to plot the axes on the $-\log_{10}$ scale for either a standard Q-Q plot (Figure 9) or for a differenced Q-Q plot (Figure 10). To create Figure 9 we use

```
R> qq_conf_plot(obs = cjd_df[,3], distribution = qunif,
+   points_params = list(pch = 21, cex = 0.2), log10 = TRUE)
```

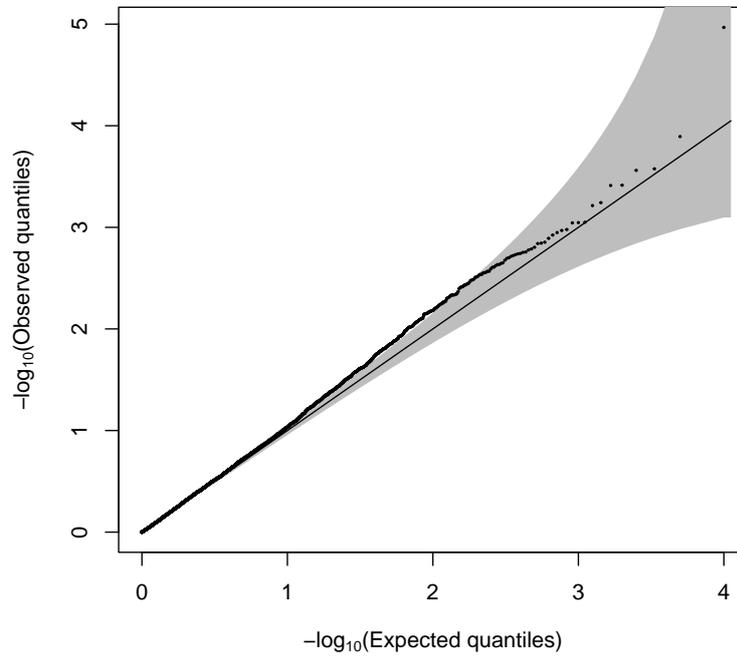


Figure 9: Q-Q plot of CJD GWAS p values, with axes on the $-\log_{10}$ scale.

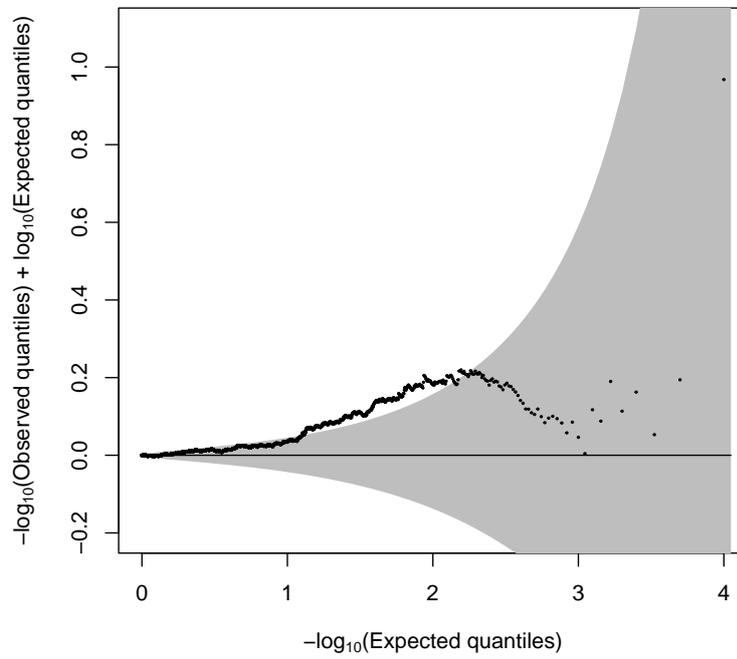


Figure 10: Differenced Q-Q plot of CJD p values, with quantiles on the $-\log_{10}$ scale.

and to create Figure 10 we use

```
R> qq_conf_plot(obs = cjd_df[,3], distribution = qunif,
+   points_params = list(pch = 21, cex = 0.2), difference = TRUE,
+   log10 = TRUE, ylim = c(-0.2, 1.1))
```

From the Q-Q plots, we can see that there is an excess of moderately small p values, indicating that the test statistics do not follow the null distribution. The type of deviation observed is suggestive of a large number of sub-significant signals, likely representing genetic variants that each contribute a small amount to the trait. It is a common phenomenon in GWAS of complex traits to have many small-effect SNPs whose signals do not become significant except in very large sample sizes.

4. Discussion

A Q-Q plot can be extremely valuable as a visualization tool for understanding the extent and type of deviation of a data set from a given reference distribution. A crucial part of the interpretation of a Q-Q plot is the ability to distinguish run-of-the-mill sampling variability from meaningful deviation, and this can be accomplished by adding an appropriate testing band to a Q-Q plot. ELL testing bands have been shown to be a notable improvement over other available methods such as KS, but previously available software has been limited to the normal distribution and is somewhat slow because it uses simulation to create the bands. To address the need for rapid generation of testing bands for Q-Q plots for a variety of reference distributions, we have developed **qqconf**, an R package for creating Q-Q plots, which is available on CRAN. A notable feature of **qqconf** is the option to quickly and easily add a simultaneous testing band based on ELL to a Q-Q or P-P plot, for any reference distribution with a quantile function implemented. We show how **qqconf** can easily be used to output bands for use in other plotting packages. For the most common testing levels of 0.05 or 0.01, generation of testing bands with `get_qq_band` in **qqconf** is so fast that one can confidently generate such bands as a default when creating Q-Q plots.

qqconf makes various accommodations for large data sets, including (1) use of pre-computed and/or asymptotic values for even faster implementation of testing bands in the function `get_qq_band`; and (2) the option to easily display Q-Q and P-P plots on the difference scale for better visualization of large data sets. For applications in genomics (Section 3.3) and assessing accuracy of p values (Section 3.2), one is particularly interested in visualizing deviations in the tail of the distribution. In such cases, it is particularly informative to view the Q-Q or P-P plot on a log scale (where the details of the log transformation depend on which tail is of interest). **qqconf** gives the option to easily generate such a log-transformed Q-Q or P-P plot to focus on deviation in either the left or right tail.

Beyond the Q-Q plot application, ELL is a generic global testing method, and the problem of determining, for a given number of local tests n and a given global testing level α , the appropriate local level $\eta_n(\alpha)$ for an ELL test can arise in other applications, particularly in genomics. The **qqconf** package contains generic functions (`get_bounds_two_sided`, `get_level_from_bounds_two_sided`, `get_bounds_one_sided`, and `get_level_from_bounds_one_sided`) to quickly obtain $\eta_n(\alpha)$ for both one-sided and two-sided testing problems, using the FFT method (Moscovich and Nadler 2016) as implemented in Moscovich

(2020b). For two-sided ELL in the cases of $\alpha = 0.05$ and 0.01 , we have used the method of Appendix A to generate extensive look-up tables for n as large as 1 million and 500K, respectively, and this permits quick access to these values or quick linear interpolation to approximate between the grid points in cases where the grid is not saturated (e.g., near the largest values of n). In addition, we have refined and applied previous asymptotic approximations for two-sided ELL, which can be confidently used for data sets of size 100K or larger.

In practice, we find that Q-Q plots are most often used either for distributions with known parameters, such as $U(0,1)$ or χ^2 with known degrees of freedom, or for the normal distribution with unknown parameters. **qqconf** provides extremely accurate testing bands for all such cases. For the case of non-normal, non-uniform reference distributions with unknown parameters, if the quantile/cdf/density functions are implemented in R, then by default **qqconf** will use maximum likelihood estimation to estimate the parameters (though the user can easily substitute estimates of their choice) and then form the testing band by taking these estimates as known values. In sufficiently large data sets, standard asymptotic theory ensures that the parameter estimates will be close to the true values, and this method will work well. In small sample sizes, use of maximum likelihood estimation in this context (non-normal distributions with unknown parameters) tends to lead to overly conservative testing bands. However, we have not yet identified a substantive application that requires a non-normal reference distribution with unknown parameters, so we do not know if this is of sufficient interest to warrant further extensions. If a need for this were identified, then two possible approaches to making the testing bands less conservative for that situation would be (1) for each distribution of interest, identify or develop a parameter estimation method that can be shown to generate bands with the appropriate global level (as we have already done for the normal distribution) or (2) extend the simulation-based approach of Aldor-Noiman *et al.* (2013) to the distribution of interest, where this could also involve choosing an appropriate estimation method as in (1). A different potential extension of arguably greater interest is to dependent rather than i.i.d. data, and for the case of multivariate normal data where the covariance structure is known or could be estimated, a simulation-based approach along the lines of Akinbiyi (2020) could be developed.

Acknowledgments

This study was supported by National Institutes of Health grant R01 HG001645 (to M.S.M.).

References

- Akinbiyi T (2020). *Equal Local Levels: A Global Testing Approach with Application to Trans eQTL Detection*. Ph.D. thesis, The University of Chicago, Department of Statistics.
- Aldor-Noiman S, Brown LD, Buja A, Rolke W, Stine RA (2013). “The Power to See: A New Graphical Test of Normality.” *The American Statistician*, **67**(4), 249–260. doi:10.1080/00031305.2013.847865.
- Almeida A, Loy A, Hofmann H (2018). “**ggplot2** Compatible Quantile-Quantile Plots in R.” *The R Journal*, **10**(2), 248–261. doi:10.32614/rj-2018-051.

- Berk RH, Jones DH (1979). “Goodness-of-Fit Test Statistics That Dominate the Kolmogorov Statistics.” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **47**(1), 47–59. doi:10.1007/bf00533250.
- Blom G (1958). *Statistical Estimates and Transformed Beta Variables*. John Wiley & Sons.
- Casella G, Berger RL (2002). *Statistical Inference*. Duxbury.
- Donoho D, Jin J (2015). “Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects.” *Statistical Science*, **30**(1), 1–25. doi:10.1214/14-sts506.
- Gontscharuk V, Finner H (2017). “Asymptotics of Goodness-of-Fit Tests Based on Minimum P-Value Statistics.” *Communications in Statistics - Theory and Methods*, **46**(5), 2332–2342. doi:10.1080/03610926.2015.1041985.
- Gontscharuk V, Landwehr S, Finner H (2015). “The Intermediates Take It All: Asymptotics of Higher Criticism Statistics and a Powerful Alternative Based on Equal Local Levels.” *Biometrical Journal*, **57**(1), 159–180. doi:10.1002/bimj.201300255.
- Gontscharuk V, Landwehr S, Finner H, *et al.* (2016). “Goodness of Fit Tests in Terms of Local Levels with Special Emphasis on Higher Criticism Tests.” *Bernoulli*, **22**(3), 1331–1363. doi:10.3150/14-bej694.
- Jones E, Hummerich H, Viré E, Uphill J, Dimitriadis A, Speedy H, Campbell T, Norsworthy P, Quinn L, Whitfield J, *et al.* (2020). “Identification of Novel Risk Loci and Causal Insights for Sporadic Creutzfeldt-Jakob Disease: a Genome-Wide Association Study.” *The Lancet Neurology*, **19**(10), 840–848. doi:10.1016/s1474-4422(20)30273-8.
- Kolmogorov A (1941). “Confidence Limits for an Unknown Distribution Function.” *The Annals of Mathematical Statistics*, **12**(4), 461–463. doi:10.1214/aoms/1177731684.
- Lewis T, Saunders IW, Westcott M (1984). “The Moments of the Pearson Chi-Squared Statistic and the Minimum Expected Value in Two-Way Tables.” *Biometrika*, **71**(3), 515–522. doi:10.1093/biomet/71.3.515.
- Mason DM, Schuenemeyer JH (1983). “A Modified Kolmogorov-Smirnov Test Sensitive to Tail Alternatives.” *The Annals of Statistics*, **11**(3), 933–946. doi:10.1214/aos/1176346259.
- Moscovich A (2020a). “Fast Calculation of p -Values for One-Sided Kolmogorov-Smirnov Type Statistics.” *arXiv 2009.04954*, arXiv.org E-Print Archive. doi:10.48550/arXiv.2009.04954.
- Moscovich A (2020b). **Crossprob**. C++ package, URL <https://github.com/mosco/crossing-probability>.
- Moscovich A, Nadler B (2016). “Fast Calculation of Boundary Crossing Probabilities for Poisson Processes.” *Statistics and Probability Letters*, **123**, 177–182. doi:10.1016/j.spl.2016.11.027.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

- Rosenkrantz WA (2000). “Confidence Bands for Quantile Functions: A Parametric and Graphic Alternative for Testing Goodness of Fit.” *The American Statistician*, **54**(3), 185–190. doi:[10.2307/2685588](https://doi.org/10.2307/2685588).
- Rousseeuw PJ, Croux C (1993). “Alternatives to the Median Absolute Deviation.” *Journal of the American Statistical Association*, **88**(424), 1273–1283. doi:[10.1080/01621459.1993.10476408](https://doi.org/10.1080/01621459.1993.10476408).
- Shorack GR, Wellner JA (2009). *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics.
- Smirnov NV (1944). “Approximate Laws of Distribution of Random Variables from Empirical Data.” *Uspekhi Matematicheskikh Nauk*, **10**, 179–206.
- Weine E, McPeck MS, Mark A (2023). **qqconf**: *Creates Simultaneous Testing Bands for QQ-Plots*. R package version 1.3.1, URL <https://CRAN.R-project.org/package=qqconf>.
- Wilk MB, Gnanadesikan R (1968). “Probability Plotting Methods for the Analysis of Data.” *Biometrika*, **55**(1), 1–17. doi:[10.1093/biomet/55.1.1](https://doi.org/10.1093/biomet/55.1.1).

A. Recursion to compute global level of two-sided ELL

First, note that because of property (4), without loss of generality, we can assume $X_1, \dots, X_n \stackrel{iid}{\sim} U(0, 1)$, and we denote the resulting probability distribution by P_0 . The goal is then to calculate the following probability:

$$\alpha = P_0\left(\bigcup_{i=1}^n \{X_{(i)} \notin (h_i, g_i)\}\right) = 1 - P_0\left(\bigcap_{i=1}^n \{X_{(i)} \in (h_i, g_i)\}\right) = 1 - P_0\left(\bigcap_{i=1}^n \{X_{(i)} \in (h_i, g_i]\}\right).$$

Let b_1, \dots, b_{2n} be the sorted values of $h_1, \dots, h_n, g_1, \dots, g_n$ in ascending order. We also define $b_0 = 0$ and $b_{2n+1} = 1$. We divide the interval (b_0, b_{2n+1}) into $2n + 1$ bins, where bin 1 is $B_1 = (b_0, b_1]$, bin 2 is $B_2 = (b_1, b_2]$, \dots , and bin $2n + 1$ is $B_{2n+1} = (b_{2n}, b_{2n+1})$. Let $N_j = \sum_{i=1}^n \mathbb{1}(X_i \in B_j)$ denote the random variable that counts the number of X 's falling into bin j , for $1 \leq j \leq 2n + 1$, and let $S_k = \sum_{j=1}^k N_j$ be the k th partial sum of the N 's, for $1 \leq k \leq 2n + 1$. We make the following key observation:

$$\{X_{(i)} \in (h_i, g_i] \text{ for } i = 1, \dots, n\} = \{l_k \leq S_k \leq u_k \text{ for } k = 1, 2, \dots, 2n\},$$

where for $1 \leq k \leq 2n$,

$$u_k = \begin{cases} 0, & \text{if } k = 1 \\ \sum_{i=1}^{k-1} \mathbb{1}(b_i \in \{h_1, \dots, h_n\}), & \text{otherwise} \end{cases}$$

$$l_k = \sum_{i=1}^k \mathbb{1}(b_i \in \{g_1, \dots, g_n\})$$

Note that $u_{2n} = l_{2n} = n$ always holds. Here, u_k is the number of order statistics whose lower interval end points are to the left of bin k , so it is an upper bound on the number of X_i 's that could occur in $\cup_{j=1}^k B_j$. Similarly, l_k is the number of order statistics whose upper interval endpoints are to the left of bin $k + 1$, so it is a lower bound on the number of X_i 's that could occur in $\cup_{j=1}^k B_j$. Thus, if we define $\Lambda = \{(m_1, \dots, m_{2n}) \in \{0, \dots, n\}^{2n} \text{ s.t. } l_k \leq s_k \leq u_k \text{ for } 1 \leq k \leq 2n, \text{ where } s_k = \sum_{i=1}^k m_i \text{ for } 1 \leq k \leq 2n\}$, then

$$P_0(\cap_{i=1}^n \{X_{(i)} \in (h_i, g_i]\}) = \sum_{(m_1, \dots, m_{2n}) \in \Lambda} P_0(N_j = m_j \text{ for } j = 1, \dots, 2n)$$

$$= \sum_{(m_1, \dots, m_{2n}) \in \Lambda} \binom{n}{m_1, \dots, m_{2n}} \prod_{j=1}^{2n} (b_j - b_{j-1})^{m_j},$$

a sum of probabilities of multinomial events, where m_j is the number of X_i 's that fall into bin j . To calculate the needed probability, we define

$$c_j^{(k)} = P_0(S_k = j \text{ and } l_q \leq S_q \leq u_q \text{ for } q = 1, \dots, k-1), \text{ for } k = 1, \dots, 2n \text{ and } j = 0, \dots, n.$$

$$\text{Then } c_j^{(k)} = P_0\left(\bigcup_{m=l_{k-1}}^{\min(j, u_{k-1})} \{S_{k-1} = m \text{ and } N_k = j - m \text{ and } l_q \leq S_q \leq u_q \text{ for } q = 1, \dots, k-2\}\right)$$

$$\begin{aligned}
&= \sum_{m=l_{k-1}}^{\min(j, u_{k-1})} P_0(S_{k-1} = m \text{ and } N_k = j - m \text{ and } l_q \leq S_q \leq u_q \text{ for } q = 1, \dots, k-2) \\
&= \sum_{m=l_{k-1}}^{\min(j, u_{k-1})} c_m^{(k-1)} \cdot P_0(N_k = j - m | S_{k-1} = m) \\
&= \sum_{m=l_{k-1}}^{\min(j, u_{k-1})} c_m^{(k-1)} \cdot P(B = j - m), \text{ where } B \sim \text{Binomial}\left(n - m, \frac{b_k - b_{k-1}}{1 - b_{k-1}}\right),
\end{aligned}$$

which gives an easily computed recursive formula, where the initialization is $c_0^{(1)} = (1 - b_1)^n$. In the case of general vectors $((h_1, \dots, h_n)$ and $(g_1, \dots, g_n))$, subject only to $0 \leq h_i < g_i \leq 1$ for $1 \leq i \leq n$, we could use the recursion to obtain $c_n^{(2n)}$, and then obtain the global level $\alpha = 1 - c_n^{(2n)}$. For the special case in which $(h_1, g_1), \dots, (h_n, g_n)$ are derived from two-sided ELL, i.e., Equations 2 and 3 hold, then as a result of the symmetry in the problem, for each $1 \leq j \leq n$, we need only calculate $c_j^{(k)}$ for $k = 1, \dots, n+1$ instead of $k = 1, \dots, 2n$ and then use

$$1 - \alpha = P_0\left(\bigcap_{i=1}^n X_{(i)} \in (h_i, g_i)\right) = \sum_{j=l_n}^{u_n} c_j^{(n)} \cdot \frac{c_{n-j}^{(n+1)}}{\binom{n}{j} b_n^j (1 - b_n)^{n-j}}.$$

To show this, we first define the following values for $k = 1, \dots, 2n$:

$$\begin{aligned}
\tilde{u}_k &= \begin{cases} 0, & \text{if } k = 2n \\ \sum_{i=k+1}^{2n} \mathbb{1}(b_i \in \{g_1, \dots, g_n\}), & \text{otherwise} \end{cases} \\
\tilde{l}_k &= \sum_{i=k}^{2n} \mathbb{1}(b_i \in \{h_1, \dots, h_n\}) \\
T_k &= \sum_{j=k+1}^{2n+1} N_j = n - S_k
\end{aligned}$$

Now, we make the following observations: (a) With two-sided ELL, $g_i = 1 - h_{n+1-i}$ for $i = 1, \dots, n$ because $F_{\text{Beta}(i, n+1-i)}^{-1}(1 - \frac{\eta}{2}) = 1 - F_{\text{Beta}(n+1-i, i)}^{-1}(\frac{\eta}{2})$. (b) $b_k = 1 - b_{2n+1-k}$ for $k = 1, \dots, 2n$ by (a). (c) $u_k = \tilde{u}_{2n+1-k}$ and $l_k = \tilde{l}_{2n+1-k}$, for $k = 1, \dots, 2n$ by (a) and (b). (d) The random vector (N_1, \dots, N_k) has the same distribution as $(N_{2n+1}, \dots, N_{2n+2-k})$ for $k = 1, \dots, 2n+1$. This follows from the fact that the vector (X_1, \dots, X_n) has the same distribution as $(1 - X_1, \dots, 1 - X_n)$ (since each X_i is independent uniform) and (b). (e) The random vector (S_1, \dots, S_k) has the same distribution as $(T_{2n}, \dots, T_{2n+1-k})$ for $k = 1, \dots, 2n$. This follows from (d). (f) $c_j^{(k)} = P_0(T_{2n+1-k} = j \text{ and } \tilde{l}_r \leq T_r \leq \tilde{u}_r \text{ for } r = 2n+2-k, \dots, 2n)$, which follows from (c) and (e). (g) Conditional on S_k , the random vector $(X_{(S_k+1)}, \dots, X_{(n)})$ is distributed as the order statistics of $n - S_k$ i.i.d. draws from $U(b_k, 1)$. (h) The random vector (S_1, \dots, S_r) and the random vector (T_r, \dots, T_n) are independent conditional on S_r . This follows directly from (g). Combining the above results, we can write

$$\{X_{(i)} \in (h_i, g_i) \text{ for } i = 1, \dots, n\} = \{l_k \leq S_k \leq u_k \text{ for } k = 1, \dots, 2n\}.$$

Also, observe that for any $2 \leq r \leq 2n - 1$, we have $\{l_k \leq S_k \leq u_k \text{ for } k = 1, \dots, 2n\} =$

$$\{l_k \leq S_k \leq u_k \text{ for } k = 1, \dots, r \text{ and } T_r = n - S_r \text{ and } \tilde{l}_k \leq T_q \leq \tilde{u}_k \text{ for } q = r, \dots, 2n\}.$$

Thus, we can write $\sum_{j=l_r}^{u_r} P(S_r = j \text{ and } l_k \leq S_k \leq u_k \text{ for } k = 1, \dots, r-1)$

$$\begin{aligned} & \cdot I(\tilde{l}_r \leq n - j \leq \tilde{u}_r) \cdot P(\tilde{l}_q \leq T_q \leq \tilde{u}_r \text{ for } q = r+1, \dots, 2n | T_r = n - j) \\ &= \sum_{j=l_r}^{u_r} c_j^{(r)} \cdot I(\tilde{l}_r \leq n - j \leq \tilde{u}_r) \cdot P(\tilde{l}_q \leq T_q \leq \tilde{u}_r \text{ for } q = r+1, \dots, 2n | T_r = n - j) \\ &= \sum_{j=l_r}^{u_r} c_j^{(r)} \cdot I(\tilde{l}_r \leq n - j \leq \tilde{u}_r) \cdot \frac{P(\tilde{l}_q \leq T_q \leq \tilde{u}_r \text{ for } q = r+1, \dots, 2n \text{ and } T_r = n - j)}{T_r = n - j} \\ &= \sum_{j=l_r}^{u_r} c_j^{(r)} \cdot \frac{c_{n-j}^{(2n+1-r)}}{\binom{n}{j} b_r^j (1-b_r)^{n-j}} \end{aligned}$$

Now, if we let $r = n$ above, then we get

$$P_0\left(\bigcap_{i=1}^n X_{(i)} \in (h_i, g_i)\right) = \sum_{j=l_n}^{u_n} c_j^{(n)} \cdot \frac{c_{n-j}^{(n+1)}}{\binom{n}{j} b_n^j (1-b_n)^{n-j}}$$

as claimed.

For the calculation of general boundary crossing probabilities, this type of algorithm requires $O(n^3)$ operations. However, for the ELL boundary crossing problem, based on experiments involving a dense grid of values of n between 10 and 50,000 and $\alpha = 0.05$, we find that the number of recursive steps required is approximately $8n^2$. While each recursive step itself requires calculating a binomial probability which is $O(n)$ due to the calculation of the binomial coefficient multiplied by a quantity with powers as large as n , these calculations can be memoized with $O(n^2)$ cost. Thus, in our specific context of ELL-based boundaries and for the range of sample sizes in our application, we find that the computational time is approximately a constant multiple of n^2 .

To find the value of η for which the local level is α , we perform a binary search over the range $(\eta_{lower}, \eta_{upper})$, where $\eta_{upper} = \alpha$ and $\eta_{lower} = \frac{\alpha}{n}$, which is the lower bound given by the Bonferroni correction. For the given choice of P_0 , note that F_{0i} is Beta($i, n - i + 1$), which is used to obtain h_i and g_i via Equation 3.

B. Recursions to compute global level of one-sided ELL

We describe an exact recursion to calculate $\alpha'_n(\eta)$ as well as an approximation, also recursive, which is much faster and bounds the relative error in the reported global significance level to a tolerance set by the user. Again, because of the property in Equation 4, without loss of generality, we assume that under the null hypothesis, $X_1, \dots, X_n \stackrel{iid}{\sim} U(0, 1)$. Given a proposed set of lower bounds h_1, \dots, h_n , where $h_1 < \dots < h_n$, the goal is to calculate the following probability

$$\alpha = P_0\left(\bigcup_{i=1}^n \{X_{(i)} < h_i\}\right) = 1 - P_0\left(\bigcap_{i=1}^n \{X_{(i)} \geq h_i\}\right) = 1 - P_0\left(\bigcap_{i=1}^n \{X_{(i)} > h_i\}\right),$$

where P_0 is the probability under the null hypothesis H_0 : X_1, \dots, X_n i.i.d. $U(0, 1)$.

Similar to the two-sided case, we divide the interval $[0, 1]$ into $n + 1$ bins. First, we define $h_0 = 0$ and $h_{n+1} = 1$. Now, suppose bin 1 is $B_1 = (h_0, h_1]$, bin 2 is $B_2 = (h_1, h_2]$, \dots , and bin $n + 1$ is $B_{n+1} = (h_n, h_{n+1})$. Let $N_j = \sum_{i=1}^n \mathbf{1}(X_i \in B_j)$ denote the random variable that counts the number of X 's falling into bin j , for $1 \leq j \leq n + 1$, and let $S_k = \sum_{j=1}^k N_j$ be the k th partial sum of the N 's, for $1 \leq k \leq n + 1$. Similar to the two-sided case, we observe that the following two events are the same:

$$\{X_{(i)} > h_i \text{ for } i = 1, \dots, n\} = \{S_k \leq k - 1 \text{ for } k = 1, 2, \dots, n\},$$

Thus, if we define $\Lambda = \{(m_1, \dots, m_n) \in \{0, \dots, n\}^n \text{ s.t. } w_n = n \text{ and } w_k \leq k \text{ for } 1 \leq k \leq n, \text{ where } w_k = \sum_{i=1}^k m_i \text{ for } 1 \leq k \leq n\}$, then

$$\begin{aligned} P_0(\cap_{i=1}^n \{X_{(i)} > h_i\}) &= \sum_{(m_1, \dots, m_n) \in \Lambda} P_0(N_1 = 0 \text{ and } N_j = m_{j-1} \text{ for } j = 2, \dots, n + 1) \\ &= \sum_{(m_1, \dots, m_n) \in \Lambda} \binom{n}{m_1, \dots, m_n} \prod_{j=1}^n (h_{j+1} - h_j)^{m_j}, \end{aligned}$$

a sum of probabilities of multinomial events, where m_j is the number of X_i 's that fall into bin $j + 1$.

B.1. Recursion for exact calculation

For exact calculation of the needed probability, we define, for $k = 1, \dots, n + 1$,

$$c_j^{(k)} = P_0(S_k = j \text{ and } S_l \leq l - 1 \text{ for } l = 1, \dots, k - 1), \text{ for } j = 0, \dots, k - 1, \text{ and } c_k^{(k)} = 0.$$

Then for $0 \leq j \leq k - 1$,

$$\begin{aligned} c_j^{(k)} &= P_0(S_k = j \text{ and } S_q \leq q - 1 \text{ for } q = 1, \dots, k - 1) \\ &= P_0\left(\bigcup_{m=0}^j \{S_{k-1} = m \text{ and } N_k = j - m \text{ and } S_q \leq q - 1 \text{ for } q = 1, \dots, k - 2\}\right) \\ &= \sum_{m=0}^j P_0(S_{k-1} = m \text{ and } N_k = j - m \text{ and } S_q \leq q - 1 \text{ for } q = 1, \dots, k - 2) \\ &= \sum_{m=0}^j c_m^{(k-1)} \cdot P_0(N_k = j - m | S_{k-1} = m) \\ &= \sum_{m=0}^j c_m^{(k-1)} \cdot P(B = j - m), \text{ where } B \sim \text{Binomial}\left(n - m, \frac{h_k - h_{k-1}}{1 - h_{k-1}}\right), \end{aligned}$$

which gives an easily computed recursive formula, where the initialization is $c_0^{(1)} = (1 - h_1)^n$. Then the global level α is equal to $1 - c_n^{(n+1)}$.

B.2. Recursion for fast approximation with error control

For sufficiently large k , the terms of the sum for small i in the update step

$$c_j^{(k)} = \sum_{i=0}^j c_j^{(k-1)} \cdot P(B = j - i), \text{ where } B \sim \text{Bin}\left(n - i, \frac{h_k - h_{k-1}}{1 - h_{k-1}}\right),$$

become negligible, and as k gets large, we speed up the algorithm by dropping negligible terms while bounding the relative error in the final calculation of the global level. Because all terms in the sum are positive, the approximation will be less than or equal to the true global level, and below we define the error to be the global level minus the approximation, which will always be nonnegative. This will lead to a slightly conservative ELL test, but with the relative error in the level guaranteed to be bounded by an arbitrary pre-specified amount.

As k increases, we specify a schedule for checking whether there are sufficiently small terms that can be dropped. We begin checking at $k = \text{first_check}$, and after that, we check whenever k is a multiple of check_interval . At a given checkpoint, the decision of whether additional terms can be dropped is based only on the current values of the recursive variables $c_i^{(k)}$, $1 \leq i \leq k - 1$ and the current value of $\text{accumul_err_upper_bnd}$, which is an upper bound on the error in the global level due to the terms that have already been dropped. The checkpoints at which additional terms of the sum are chosen to be dropped are termed “drop points,” and we label these d_1, \dots, d_w , where $\text{first_check} \leq d_1 < \dots < d_w < n$. We end up with a modified recursion with initialization $\tilde{c}_0^{(1)} = (1 - h_1)^n$ and $\tilde{c}_1^{(1)} = 0$, and with update step:

$$\tilde{c}_j^{(k)} = \sum_{i=\text{skip}_k+1}^j \tilde{c}_i^{(k-1)} \cdot P(B = j - i), \text{ where } B \text{ is Binomial}(n - i, \frac{h_k - h_{k-1}}{1 - h_{k-1}}),$$

for $\text{skip}_k + 1 \leq j \leq k - 1$ and $\tilde{c}_k^{(k)} = 0$, where $\text{skip}_k = -1$ for $1 \leq k \leq d_1$, $\text{skip}_k = T_i$ for $d_i < k \leq d_{i+1}$ and $1 \leq i \leq w - 1$, and $\text{skip}_k = T_w$ for $d_w < k \leq n$. In other words, at drop point d_1 , terms of the sum indexed 0 through T_1 are dropped, and for $2 \leq m \leq w$, at drop point d_m additional terms indexed by $T_{m-1} + 1$ through T_m are dropped (terms 0 through T_{m-1} having already been dropped at previous drop points). At drop point d_1 , the value T_1 is chosen to be the largest value of T such that

$$T < d_1 \text{ and } \frac{\sum_{j=0}^T \tilde{c}_j^{(d_1)}}{1 - \sum_{l=0}^{d_1-1} \tilde{c}_l^{(d_1)}} \leq \text{max_rel_err}, \quad (8)$$

and for $2 \leq m \leq w$, at drop point d_m , the value T_m is chosen to be the largest value of T such that

$$d_m > T > T_{m-1} \text{ and } \frac{e_{m-1} + \sum_{j=T_{m-1}+1}^T \tilde{c}_j^{(d_m)}}{1 - e_{m-1} - \sum_{l=T_{m-1}+1}^{d_m-1} \tilde{c}_l^{(d_m)}} \leq \text{max_rel_err}, \quad (9)$$

where we define e_m to be the value of $\text{accumul_err_upper_bnd}$ after recursion step $k = d_m$, which is given by

$$e_1 = \sum_{j=0}^{T_1} \tilde{c}_j^{(d_1)} \text{ and for } m > 1, e_m = \sum_{j=0}^{T_1} \tilde{c}_j^{(d_1)} + \sum_{i=2}^m \sum_{j=T_{i-1}+1}^{T_i} \tilde{c}_j^{(d_i)}.$$

A given checkpoint k becomes a drop point if and only if there is some T satisfying the corresponding constraints (either Equation 8 or 9).

We first show that e_m is an upper bound on the actual accumulated error a_m that will be incurred in the calculation due to all terms dropped prior to drop point d_{m+1} . We define E to be the event $\{S_k \leq k - 1 \text{ for } k = 1, 2, \dots, n\}$. Then, as noted in Section 2, $P(E) = c_n^{(n+1)}$, and the global level is $1 - P(E)$.

Lemma: $a_i \leq e_i$ for $i = 1, \dots, w$.

Proof: Note that at drop point d_1 , dropping terms of the sum indexed 0 through T_1 is equivalent to adding an extra requirement that $S_{d_1} > T_1$, so that instead of $P(E)$, we will be calculating $P(E \cap \{S_{d_1} > T_1\})$. Therefore, the actual accumulated error in the global level that will be incurred by this is $a_1 = P(E \cap \{S_{d_1} \leq T_1\})$, which is bounded above by $P(S_{d_1} \leq T_1 \text{ and } S_l \leq l - 1 \text{ for } l = 1, \dots, d_1 - 1) = \sum_{j=0}^{T_1} c_j^{(d_1)} = \sum_{j=0}^{T_1} \tilde{c}_j^{(d_1)} = e_1$.

At the induction step, we assume that $a_{m-1} \leq e_{m-1}$. Now consider the actual accumulated error a_m due to the terms dropped at drop points d_1, \dots, d_m . By similar logic as above,

$$\begin{aligned} a_m &= P(E \cap [\cap_{j=1}^m (S_{d_j} > T_j)]^c) = P(E \cap \{[\cap_{j=1}^{m-1} (S_{d_j} > T_j)]^c \cup (S_{d_m} > T_m)^c\}) \\ &= P(E \cap [\cap_{j=1}^{m-1} (S_{d_j} > T_j)]^c) + P(E \cap [\cap_{j=1}^{m-1} (S_{d_j} > T_j)] \cap (S_{d_m} > T_m)^c) \\ &= a_{m-1} + P(E \cap [\cap_{j=1}^{m-1} (S_{d_j} > T_j)] \cap (S_{d_m} \leq T_m)) \leq a_{m-1} + \sum_{j=T_{m-1}+1}^{T_m} \tilde{c}_j^{(d_m)} \\ &\leq e_{m-1} + \sum_{j=T_{m-1}+1}^{T_m} \tilde{c}_j^{(d_m)} = e_m, \end{aligned}$$

where the 2nd and 3rd equalities are based only on elementary set theory. \square

We now prove that this algorithm guarantees relative error of no more than `max_rel_err` in the calculated global level. First we note that $P(E) = c_n^{(n+1)}$ satisfies

$$c_n^{(n+1)} \leq \sum_{j=0}^{k-1} c_j^{(k)}, \text{ for all } 1 \leq k \leq n+1,$$

where this useful inequality follows directly from the definition of $c_j^{(k)}$ in Section 2. A consequence is that $P(E) \leq \sum_{j=0}^{d_1-1} c_j^{(d_1)} = \sum_{j=0}^{d_1-1} \tilde{c}_j^{(d_1)}$. Second, we note that by similar reasoning, for $m \geq 2$,

$$\begin{aligned} P(E) &= P(E \cap [\cap_{j=1}^m (S_{d_j} > T_j)]^c) + P(E \cap [\cap_{j=1}^m (S_{d_j} > T_j)]) \\ &= a_{m-1} + P(E \cap [\cap_{j=1}^m (S_{d_j} > T_j)]) \leq a_{m-1} + \sum_{j=T_{m-1}+1}^{d_m-1} \tilde{c}_j^{(d_m)} \leq e_{m-1} + \sum_{j=T_{m-1}+1}^{d_m-1} \tilde{c}_j^{(d_m)}. \end{aligned}$$

Therefore, the actual relative error in the global level incurred at drop point d_1 , which is $a_1/(1 - P(E))$, satisfies

$$\frac{a_1}{1 - P(E)} \leq \frac{e_1}{1 - \sum_{j=0}^{d_1-1} \tilde{c}_j^{(d_1)}} = \frac{\sum_{j=0}^{T_1} \tilde{c}_j^{(d_1)}}{1 - \sum_{j=0}^{d_1-1} \tilde{c}_j^{(d_1)}} \leq \text{max_rel_err}$$

by Equation 8. Furthermore, the actual relative error in the global level that will be incurred in the calculation due to all terms dropped at or before d_m is

$$\begin{aligned} a_m/(1 - P(E)) &\leq \frac{e_m}{1 - e_{m-1} - \sum_{l=T_{m-1}+1}^{d_m-1} \tilde{c}_j^{(d_m)}} = \frac{e_{m-1} + \sum_{j=T_{m-1}+1}^{T_m} \tilde{c}_j^{(d_m)}}{1 - e_{m-1} - \sum_{l=T_{m-1}+1}^{d_m-1} \tilde{c}_j^{(d_m)}} \\ &\leq \text{max_rel_err} \end{aligned}$$

by Equation 9. \square

The resulting algorithm is given below.

Algorithm 1 Calculate approximate global level α for one-sided ELL from proposed lower bounds using speedup.

Input: Vector of lower bound values where these must be increasing (h_1, \dots, h_n) , first value of k for which to check for skipping `first_check`, interval for which to check for skipping `check_interval`, maximum allowed relative error in global level calculation `max_rel_err`.
`get_level_from_bounds_one_sided` $((h_1, \dots, h_n), \text{first_check}, \text{check_interval}, \text{max_rel_err})$

```

1:  $c_0^{(1)} \leftarrow (1 - h_1)^n$ 
2:  $c_1^{(1)} \leftarrow 0$ 
3: accumul_err_upper_bnd  $\leftarrow 0$ 
4: skip  $\leftarrow -1$ 
5: for  $k = 2, \dots, n$  do
6:   for  $j = \text{skip} + 1, \dots, k - 1$  do
7:      $c_j^{(k)} \leftarrow 0$ 
8:     for  $m = \text{skip} + 1, \dots, j$  do
9:        $c_j^{(k)} \leftarrow c_j^{(k)} + c_m^{(k-1)} \cdot \text{dbinom}(x = j - m, \text{size} = n - m, \text{prob} = \frac{h_k - h_{k-1}}{1 - h_{k-1}})$ 
10:    end for
11:  end for
12:   $c_k^{(k)} \leftarrow 0$ 
13:  if  $(k > \text{first\_check} \text{ and } k \% \text{check\_interval} == 0)$  or  $k == \text{first\_check}$  then
14:    available_err  $\leftarrow \text{max\_rel\_err} - (1 + \text{max\_rel\_err}) \cdot \text{accumul\_err\_upper\_bnd}$ 
15:    calculated\_total\_prob  $\leftarrow 0$ 
16:    for  $j = \text{skip} + 1, \dots, k - 1$  do
17:      calculated\_total\_prob  $\leftarrow \text{calculated\_total\_prob} + c_j^{(k)}$ 
18:    end for
19:    available_err  $\leftarrow \text{available\_err} - \text{max\_rel\_err} \cdot \text{calculated\_total\_prob}$ 
20:    proposed_err  $\leftarrow c_{\text{skip}+1}^{(k)}$ 
21:    proposed\_skip  $\leftarrow \text{skip} + 1$ 
22:    while proposed_err  $\leq \text{available\_err}$  do
23:      proposed\_skip  $\leftarrow \text{proposed\_skip} + 1$ 
24:      proposed_err  $\leftarrow \text{proposed_err} + c_{\text{proposed\_skip}}^{(k)}$ 
25:    end while
26:    accumul_err_upper_bnd  $\leftarrow \text{proposed_err} - c_{\text{proposed\_skip}}^{(k)}$ 
27:    skip  $\leftarrow \text{proposed\_skip} - 1$ 
28:  end if
29: end for
30:  $c_n^{(n+1)} \leftarrow 0$ 
31: for  $l = \text{skip} + 1, \dots, n - 1$  do
32:    $c_n^{(n+1)} \leftarrow c_n^{(n+1)} + c_l^{(n)}$ 
33: end for
34:  $\alpha \leftarrow 1 - c_n^{(n+1)}$ 
35: return  $\alpha$ 
end

```

Affiliation:

Eric Weine, Mary Sara McPeek
Department of Statistics
University of Chicago
5747 S Ellis Ave, Chicago, IL, United States of America
E-mail: ericweine15@gmail.com, mcpeek@uchicago.edu
URL: <https://www.stat.uchicago.edu/~mcpeek/>

Mark Abney
Department of Human Genetics
University of Chicago
920 E 58th St., Chicago, IL, United States of America
E-mail: abney@uchicago.edu
URL: <https://hgen.uchicago.edu/program/faculty/mark-abney>