Peter Junglas

Is body height really normally distributed?

Abstract. Analysing data sets of body heights one finds that for extensive amounts of data goodness-of-fit tests generally reject the hypothesis of normally distributed values. Numerical experiments quickly reveal that this is due to the rounding of the measured results. Further numerical studies show how this affects several tests and how one could cope with this problem. This can help to teach students how to utilise an "experimental" approach to mathematical questions. Furthermore it points out the difference between mathematical concepts like "real numbers" and actual data.

Introduction

Since statistics is applied almost everywhere, it is taught in many disciplines. The general approach usually is a mathematical one: Statistics is based on probability theory, its foundation is clear, its assertions and methods are proven.

Nevertheless statistics usually has a bad public reputation, that has not become better since Mark Twains dictum: "There are three kinds of lies: lies, damn lies, and statistics." [1]. This is probably mainly due to flawed applications, results and statements made by people, who are better trained in the application area at hand than in mathematical intricacies. Common problems are wrong assumptions, such as data being normally distributed or independent, wrong conclusions, e. g. to mix up correlation and causality, or starting with wrong or biased data in the first place.

Therefore it is of paramount importance in teaching statistics to show not only correct methods, but to make students aware of the many pitfalls and errors that come up in everyday applications. The main purpose of this paper it to point out a special complication that is often overlooked and seldom taught, namely the consequences of using data that has been rounded. This will be done by presenting some simple numerical experiments that shed light on the problem. Hopefully this can motivate students to come up with experiments of their own to cope with the concrete challenges they will face in everyday statistical analysis.

Analysis of body heights

The analysis starts with two data sets of German female body heights containing $N = 300$ and $N = 3000$ entries. To get a first impression of their distribution, one can plot a simple histogram. This leads to the obvious assumption that the data is normally distributed, which can be tested qualitatively with a Q-Q plot. Fig. 1 shows the plots for $N = 300$, the other data set leads to similar results.

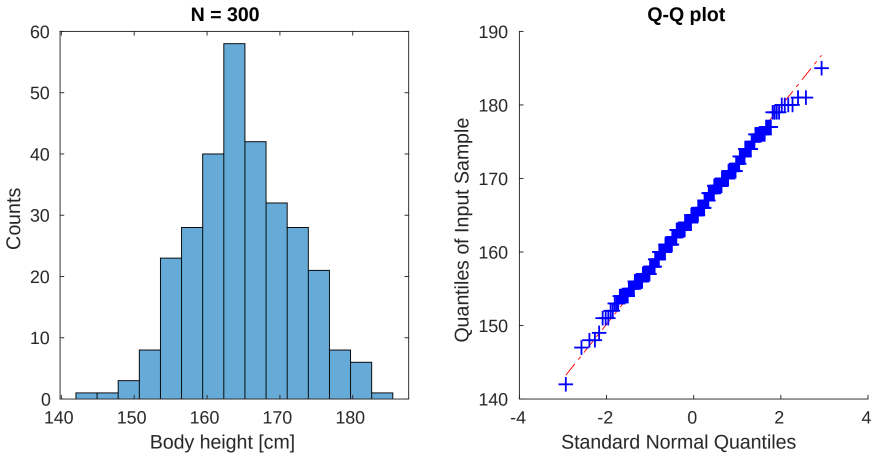

Fig. 1: Histogram and Q-Q plot for $N = 300$.

A simple way to test the normality assumption is the well-known χ^2 goodness of fit test. To use it one has to partition the range of data values into non-overlapping intervals that span the whole range. One then counts the amount of data values in each interval ("bin") and compares these numbers with the expected mean values under the hypothesis of normal distribution.

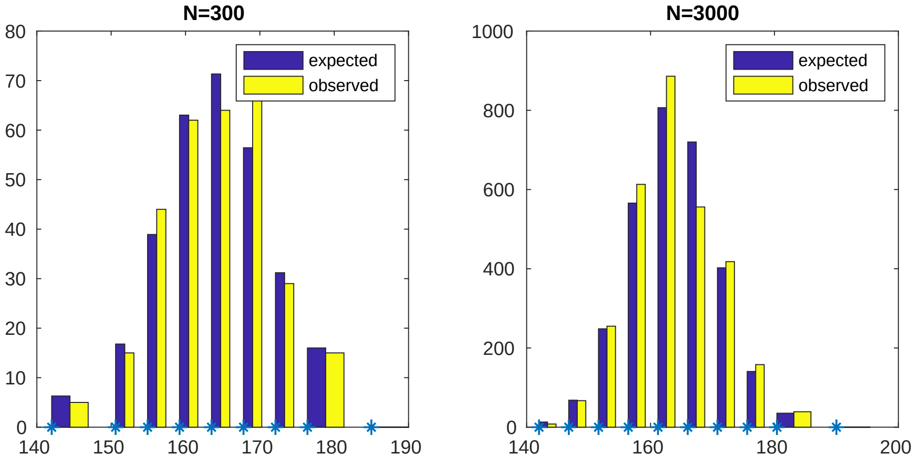

Fig. 2: Expected and observed bin counts for both data sets.

The actual choice of intervals is not really important for the test as long as the following rules of thumb are fulfilled: All bins are non-empty and at least 80% of the bins contain more than 5 data elements [2]. For concreteness in the following we will use Matlabs function `chi2gof` that chooses apropriate intervals automatically. Corresponding results are shown in Fig. 2.

These results are now combined in a variable T that quantifies the difference between expected and observed values. From it one computes the p-value p of the test, i. e. the probability that normally distributed data values have a T value at least as large as the observed one. For the two data sets one gets

$$\begin{aligned} p(300) &= 0.470 \\ p(3000) &= 0.266 \cdot 10^{-3} \end{aligned}$$

While the first result is in good agreement with the assumption of normally distributed data, the second one definitely is not. For any reasonable confidence level one would reject the hypothesis. This is in contrast to common expectation about body heights and to the qualitative results of the histogram and Q-Q plot.

After some carefully guided discussion in class, students can come up with the suggestion that the rounding of the numbers to integer centimeter values might be the reason for the strange result. To validate this idea, one can perform a simple numerical experiment: One uses a random number generator to create normally distributed values similar to the given data (adopting $\mu = 165$, $\sigma = 6.9$), rounds them to integers and applies the χ^2 test. This basically reproduces the p-values from above, which confirms the significance of the rounding effect. In hindsight this seems obvious: Rounded values always come from a discrete distribution, which in our case we assume to be a rounded version of the normal distribution.

For small data sets the χ^2 test cannot discriminate between rounded and unrounded values, but with growing data size N the p-value decreases. To study this behaviour in more detail we repeat our experiment with varying N. Since the p-values largely fluctuate, each experiment is repeated three times and the mean values are plotted over N. They decrease drastically, therefore a semilogarithmic plot is used that hints at an exponential decrease of p (Fig. 3).

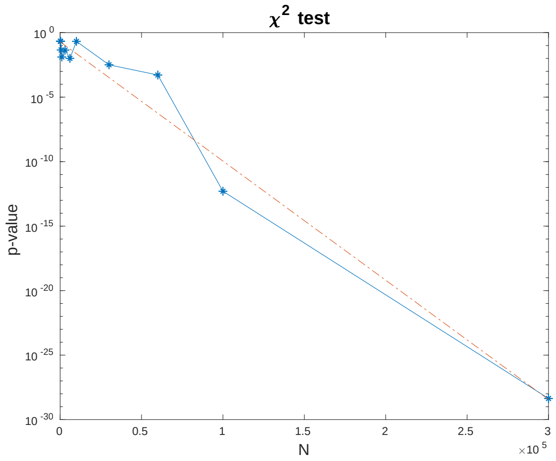

Fig. 3: p-values using the χ^2 test.

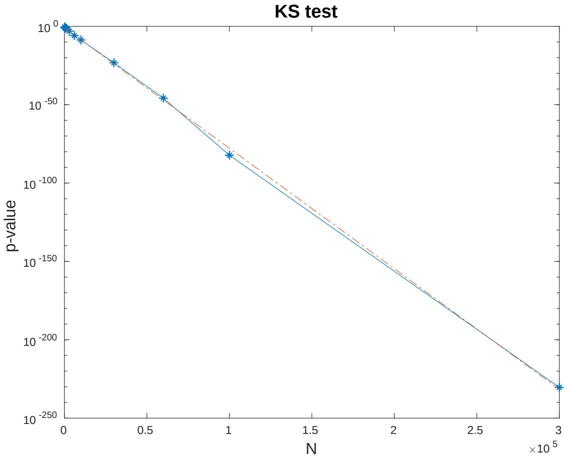

Fig. 4: p-values using the KS test.

With the freedom to choose the bin sizes and positions, the χ^2 test has a lot of parameters, which complicates the analysis. Therefore the Kolmogorov-Smirnov test ("KS test") is used to check our findings. Without any parameters it directly compares the empirical and given distribution functions and leads to a complicated, but well known test statistics. The results in Fig. 4 show that the p-values again decrease exponentially with N, but much faster than before. Apparently the KS test discriminates better between continuous and rounded values.

Choosing the bins in the χ^2 test

Since the χ^2 test is very common in many applications, we will now study, how the sizes and positions of the bins influence the ability to discriminate between exact and rounded values. We create rounded normal values as before and apply the χ^2 test, but this time we define the bins ourselves instead of relying on the choice of the Matlab routine.

We begin by using bins of identical width w, positioned such that one bin is centered around the mean value, and vary w. The resulting p-values as function of w are plotted for two data sizes in Fig. 5. One finds that the p-values are much larger at integer widths; this effect is more pronounced for odd width and for large N.

Next we fix the bin width at $w = 5$ and vary the position of the bins. Initially their edges are at integer values, than we shift them by a certain amount s and plot the p-values over s (Fig. 6). They are now minimal at integer, maximal at half-integer shifts.

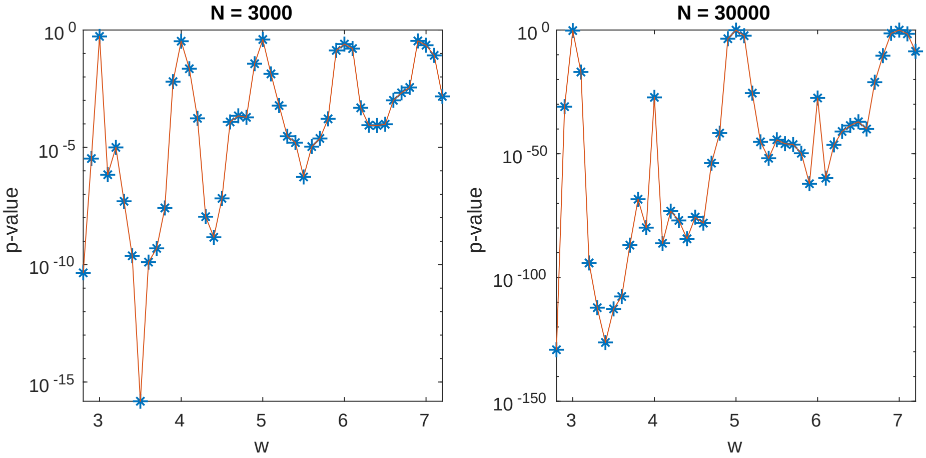

Fig. 5: p-values for different bin widths.

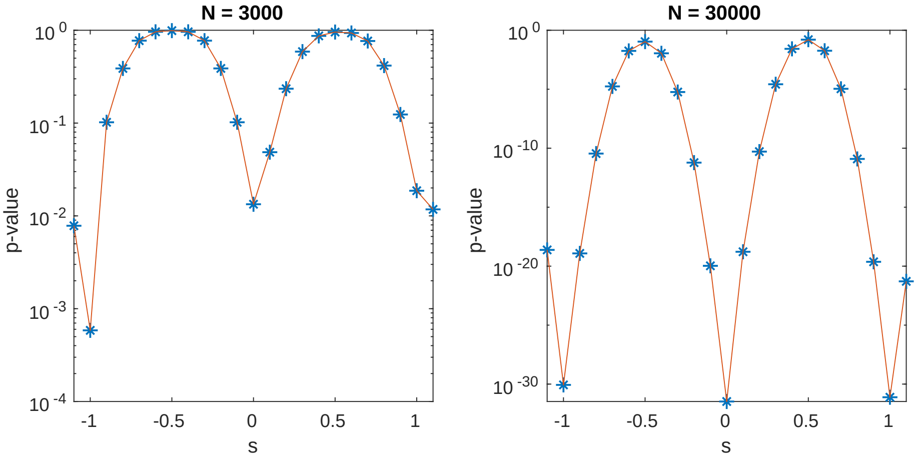

Fig. 6: p-values for different bin positions.

Again, looking at the results, their explanation is easy: The p-values are maximal for bins of the type $[n_1 + \frac{1}{2}, n_2 + \frac{1}{2}]$ with $n_1, n_2 \in \mathbb{N}$. These bins are invariant under rounding, i. e. the expected bin counts with or without rounding are the same. In these cases the χ^2 test can't detect rounding and confirms the normal distribution of the (original) values. This as well explains the minimality of the odd integer widths: Since the mean value is $\mu = 169$, the corresponding boundaries of the centered bins lie on half-integer values.

In summary we suggest the following approach: If one wants to detect the difference between exact and rounded values, one should use bin boundaries at integers – or choose the KS test. If one is only interested in the distribution of the (theoretical) real values and wants to get rid of rounding effects, one should use bin boundaries at half-integers.

Varying precision

An interesting question is, how many data points one needs to detect the rounding for given measurement precision. For definiteness we will use a confidence level of $\alpha = 0.01$, choose a precision and experimentally determine the needed size N at which the p-value is smaller than 0.01.

The experiment starts by creating normally distributed numbers, but the rounding is now done using a scale w:

$$x_{round} := w\,\mathrm{round}(x/w)$$

Here smaller values of w mean higher measurement precision. We use the KS test to measure the p-value as a function of the data size N for a set of fixed scale values w. The results are shown in Fig. 7, where again the p-values shown are mean values of three runs.

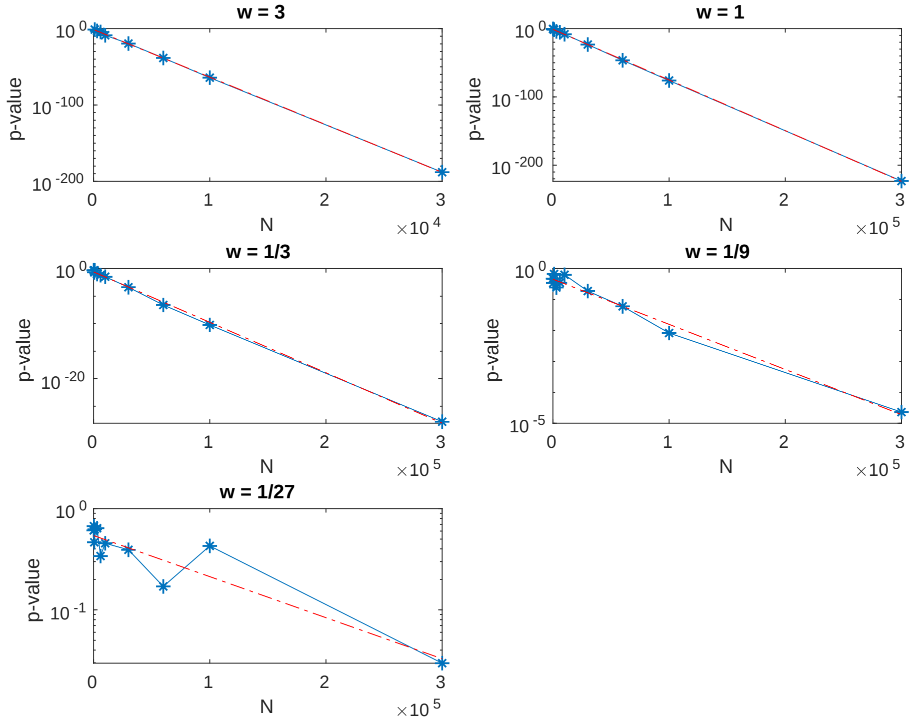

Fig. 7: p-values over N for varying scale w.

Once more we can see the exponential decline, i. e. approximately we have

$$\log(p) = a\,N + b$$

The parameters a, b of corresponding fitting lines are computed and the lines shown in the figure. We now use them to compute the size N_0, at which the p-value reaches the confidence level α and get

w	N_0	$w^2 \cdot N_0$
3.00	122	1098
1.00	1279	1279
0.33	15358	1706
0.11	113500	1401
0.04	427332	586

The last row is inaccurate, since the data size N used in the experiments was not large enough. Considering this and the large scale of values, the third column convincingly reveals the relationship

$$N_0 \sim \frac{1}{w^2}.$$

Therefore increasing the precision by a factor 10, one needs 100 times as much data to see the rounding effect in a goodnes-of-fit test.

We have only used a fixed data distribution, especially a fixed standard variation σ. It is well known that the relevant parameter describing the importance of rounding is σ/w [3]. This can be seen experimentally or by a simple scaling argument.

Smoothing the distribution

A completely different way to look at rounding comes from signal theory, where it is called “quantization” [4]. The basic idea is to look at the effects of rounding as a kind of noise. One can proof that under certain conditions this noise has special properties, which might allow to replace the actual discrete distribution function of the rounded values by a continuous one. This makes its further use in statistical computations much simpler. The necessary conditions are not fulfilled exactly in our case, but at least approximately.

We begin by writing

$$X_s = X + Y$$

where X describes our (original) data which is normally distributed,

$$X \sim N(\mu, \sigma^2),$$

and X_s is the rounded data,

$$X_s = \text{round}(X).$$

The difference Y is now interpreted as noise, which has approximately a uniform distribution: $Y \sim U(-0.5, 0.5)$. Of course X and Y are not independent, but they are at least approximately uncorrelated.

To check whether these approximations are valid in our example, we test them experimentally. We create normally distributed random values X, round them to get X_s and compute $Y = X - X_s$. We then test for the distribution of Y using a χ^2 test, compute the correlation coefficients and get

N	p-value	$\rho(X, Y)$
3000	0.6165	-0.0077
30000	0.5279	0.0085
300000	0.4733	0.0003

So the assumptions about Y seem to be reasonable here. If we further assume that X and Y are approximately independent, we can use their known distribution functions to compute the density function f_s of X_s via convolution and get

$$f_s(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} dz\, e^{-\dfrac{(x - z - \mu)^2}{2\sigma^2}}$$

This is the standard Gaussian function smoothed over an interval of width one, which coincides with the probabilities of X_s at integer x. Of course it can be easily computed numerically, as well as the corresponding cumulative distribution function F_s. For the values of our standard example f_s is shown in Fig. 8.

After all this work it's time to try out our new density function and test, whether it really is a good approximation for the distribution of rounded values. Therefore we once again create rounded random numbers and perform a χ^2 test, this time using f_s instead of the normal distribution function. We use edges on integral boundaries to make sure that the test is sensitive to rounding, and get

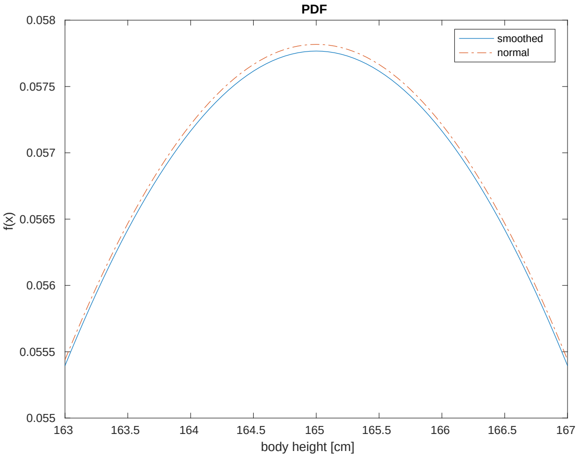

Fig. 8: Normal and smoothed normal density functions.

N	p-value (normal)	p-value (f_s)
3000	0.0044	0.0046
30000	3.7098e-42	4.9609e-42
300000	0	0

The "smoothed" distribution function is as bad as the original one. Apparently the whole idea is completely useless for a goodness-of-fit test.

Conclusions

Though it is long known that rounding can have an effect on statistical evaluations, research on precise ways to cope with it is rather recent. Several rules of thumb existed that have been shown to be wrong in [3] and replaced by better ones. Using maximum-likelihood methods one can find correct statistical estimates of rounded data, e. g. confidence intervals, especially for small data sets and small values of σ/w [5].

Our experiments have shown that rounding can have an effect even for large data sets, at least for the results of a goodness-of-fit test. If one is interested in the distribution of the "real" (i.e. unrounded) values, one can use qualitative tools like a Q-Q plot or use adapted methods like a χ^2 test with properly chosen intervals.

Furthermore there are a few general lessons that students can learn from this study: First of all numerical experiments are a valuable tool to find facts, test ideas and identify relevant parameters. Second, statistical analysis is not always done best with a huge amount of data – sometimes this might lead to fine grained details one is not interested in. Third, "reasonable" assumptions or approximations might be wrong, therefore one has always to question everything – even if it has been used for decades.

Finally we have to answer the initial question: Is body height really normally distributed? The immediate answer "no, since it cannot be negative" is irrelevant, because the probabilty of negative heights is much too small in our example to be of any consequences. That a body height has to be measured and therefore has always rounded values, is more relevant for our question, but on the other hand one can usually increase the precision, until this point doesn't really matter. The true answer is, that a "body height" doesn't exist! What could it possibly refer to: the position of the "highest" electron of your body? According to Heisenberg's uncertainty principle this concept makes no sense.

One has always to keep in mind that "real numbers" are a mathematical abstraction, while in statistics one is generally talking about the real world. Whether a mathematical model fits the situation at hand, is generally an important question, not only because of rounding.

Literaturverzeichnis

[1] **Twain, M.**: *Autobiography Vol. I.* Ed. A. B. Paine, Harper Brothers New York and London (1924).

[2] **Ross, S. M.**: *Introduction to Probability and Statistics for Engineers and Scientists.* Academic Press London, 5th ed. (2014).

[3] **Tricker, A.; Coates, E.; Okell, E.**: *The Effect on the R Chart of Precision of Measurement.* J. Qual. Technol., 3, **30**, 232–239 (1998).

[4] **Widrow, B.; Kollar, I.; Liu, M.-C.**: *Statistical theory of quantization.* IEEE Trans. Instrum. Meas., 2, **45**, 353–361 (1996).

[5] **Vardeman, S. B.; Lee, C.-S.**: *Likelihood-based statistical estimation from quantized data.* IEEE Trans. Instrum. Meas., 1, **54**, 409–414 (2005).

Author

Prof. Dr. rer. nat. Peter Junglas
Private Hochschule für Wirtschaft und Technik Vechta/Diepholz
Schlesierstraße 13a
D-49356 Diepholz
E-Mail: peter@peter-junglas.de