sign up log in tour help

t-test on highly skewed data

I have a data set with tens of thousands of observations of medical cost data. This data is highly skewed to the right and has a lot of zeros. It looks like this for two sets of people (in this case two age bands with > 3000 obs each):

```
 Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.0      0.0      0.0   4536.0    302.6 395300.0
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.0      0.0      0.0   4964.0    423.8 721700.0
```

If I perform Welch's t-test on this data I get a result back:

```
Welch Two Sample t-test

data:  x and y
t = -0.4777, df = 3366.488, p-value = 0.6329
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2185.896  1329.358
sample estimates:
mean of x mean of y
 4536.186  4964.455
```

I know its not correct to use a t-test on this data since its so badly non-normal. However, if I use a permutation test for the difference of the means, I get nearly the same p-value all the time (and it gets closer with more iterations).

Using perm package in R and permTS with exact Monte Carlo

```
    Exact Permutation Test Estimated by Monte Carlo

data:  x and y
p-value = 0.6188
alternative hypothesis: true mean x - mean y is not equal to 0
sample estimates:
mean x - mean y
      -428.2691

p-value estimated from 500 Monte Carlo replications
99 percent confidence interval on p-value:
 0.5117552 0.7277040
```

Why is the permutation test statistic coming out so close to the t.test value? If I take logs of the data then I get a t.test p-value of 0.28 and the same from the permutation test. I thought the t-test values wold be more garbage than what I am getting here. This is true of many

other data sets I have like this and am wondering why the t-test appears to be working when it shouldn't.

My concern here is that the individual costs are not i.i.d. There are many sub-groups of people with very different cost distributions (women vs men, chronic conditions etc) that seem to voilate the iid requirement for central limit theorem, or should I not worry about that?

t-test skewness permutation

edited Sep 13 '13 at 15:10

asked Sep 12 '13 at 20:53
Chris
84 6

How does it happen that both the minimum value *and* the median of your data is zero? – Alecos Papadopoulos Sep 12 '13 at 21:42

More than half the values are zero, indicating half the people had no medical care that year. – Chris Sep 12 '13 at 22:08

And why do you think that the permutation test should be different? (if both groups have a similarly non-normal distribution) – FairMiles Sep 12 '13 at 23:06

Keep in mind that i.i.d. is two separate assumptions. The first is 'independent'. The second is 'identically distributed'. You seem to be suggesting that the observations are not 'identically distributed'. This should not affect the answers provided so far, as we can still assume that all the observations are from one big mixture of distributions. But if you think that the observations are not independent, that is a much different and potentially more difficult issue. – zkurtz Sep 13 '13 at 19:21

2 Answers

Neither the t-test nor the permutation test have much power to identify a difference in means between two such extraordinarily skewed distributions. Thus they both give anodyne p-values indicating no significance at all. The issue is not that they seem to agree; it is that because they have a hard time detecting any difference at all, they simply cannot disagree!

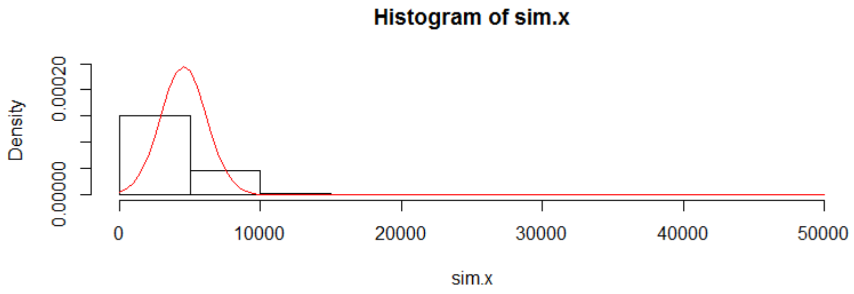
For some intuition, consider what would happen if a change in a *single* value occurred in one dataset. Suppose that the maximum of 721,700 had not occurred in the second data set, for instance. The mean would have dropped by approximately 721700/3000, which is about 240. Yet the difference in the means is only 4964-4536 = 438, not even twice as big. That suggests (although it does not prove) that *any* comparison of the means would not find the difference significant.

We can verify, though, that the t-test is not applicable. Let's generate some datasets with the same statistical characteristics as these. To do so I have created mixtures in which

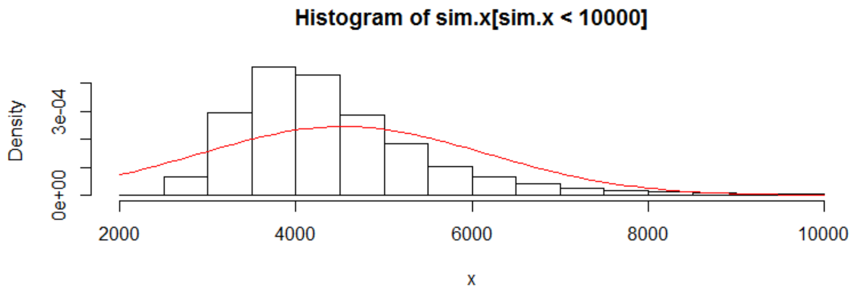
- $5/8$ of the data are zeros in any case.
- The remaining data have a lognormal distribution.
- The parameters of that distribution are arranged to reproduce the observed means and third quartiles.

It turns out in these simulations that the maximum values are not far from the reported maxima, either.

Let's replicate the first dataset 10,000 times and track its mean. (The results will be almost the same when we do this for the second dataset.) The histogram of these means estimates the sampling distribution of the mean. The t-test is valid when this distribution is approximately Normal; the extent to which it deviates from Normality indicates the extent to which the Student t distribution will err. So, for reference, I have also drawn (in red) the PDF of the Normal distribution fit to these results.



We can't see much detail because there are some whopping big outliers. (That's a manifestation of this sensitivity of the means I mentioned.) There are 123 of them--1.23%--above 10,000. Let's focus on the rest so we can see the detail and because these outliers may result from the assumed lognormality of the distribution, which is not necessarily the case for the original dataset.



That is still strongly skewed and deviates visibly from the Normal approximation, providing sufficient explanation for the phenomena recounted in the question. It also gives us a sense of how large a difference in means could be detected by a test: it would have to be around 3000 or more to appear significant. Conversely, **the actual difference of 428 might be detected provided you had approximately $(3000/428)^2 = 50$ times as much data (in each group).** Given 50 times as much data, I estimate the power to detect this difference at a significance level of 5% would be around 0.4 (which is not good, but at least you would have a chance).

Here is the `R` code that produced these figures.

```
#
# Generate positive random values with a median of 0, given Q3,
# and given mean. Make a proportion 1-e of them true zeros.
#
rskew <- function(n, x.mean, x.q3, e=3/8) {
  beta <- qnorm(1 - (1/4)/e)
  gamma <- 2*(log(x.q3) - log(x.mean/e))
  sigma <- sqrt(beta^2 - gamma) + beta
  mu <- log(x.mean/e) - sigma^2/2
  m <- floor(n * e)
  c(exp(rnorm(m, mu, sigma)), rep(0, n-m))
}
#
# See how closely the summary statistics are reproduced.
# (The quartiles will be close; the maxima not too far off;
# the means may differ a lot, though.)
#
set.seed(23)
x <- rskew(3300, 4536, 302.6)
y <- rskew(3400, 4964, 423.8)
summary(x)
summary(y)
#
# Estimate the sampling distribution of the mean.
#
set.seed(17)
sim.x <- replicate(10^4, mean(rskew(3367, 4536, 302.6)))
hist(sim.x, freq=FALSE, ylim=c(0, dnorm(0, sd=sd(sim.x))))
curve(dnorm(x, mean(sim.x), sd(sim.x)), add=TRUE, col="Red")
hist(sim.x[sim.x < 10000], xlab="x", freq=FALSE)
curve(dnorm(x, mean(sim.x), sd(sim.x)), add=TRUE, col="Red")
#
# Can a t-test detect a difference with more data?
#
set.seed(23)
n.factor <- 50
z <- replicate(10^3, {
  x <- rskew(3300*n.factor, 4536, 302.6)
  y <- rskew(3400*n.factor, 4964, 423.8)
  t.test(x,y)$p.value
})
hist(z)
mean(z < .05) # The estimated power at a 5% significance level
```

answered Sep 13 '13 at 16:47

whuber ♦
76k 8 126 264

When n is large (like 300, even far less than 3000), the t-test is essentially the same as the z-test. That is, the t-test becomes nothing more than an application of the central limit theorem, which says that the MEAN for each of your two groups is almost exactly normally distributed (even if the observations underlying the two means are very far from being normally distributed!). This is also the reason that your typical t-table does not bother to show values for n greater than 1000 (for example, this t-table). Thus, I am not surprised to see that you are getting such well-behaved results.

Edit: I seem to have underestimated the extremity of the skewness and its importance. While my point above has merit in less extreme circumstances, for this question I would like to point out that I think **whuber**'s answer to the question is much better.

edited Sep 13 '13 at 17:13

answered Sep 13 '13 at 1:39

zkurtz
510 1 9

2 When skewness is extreme--as the quoted statistics attest--we have no assurance that the sampling distribution of the mean of 300 or even 3000 samples will be anywhere near Normal. *That* is why the OP is surprised. You

counter that by saying you are not surprised, but that appears to come down to one person's intuition compared to another's. What *objective* argument can you supply *for these data* demonstrating that 300 (or 3000) is a large enough sample for the t-test to work well? – whuber ♦ Sep 13 '13 at 15:34

Great point. I admit, if the data is sufficiently skewed, my argument fails. So the question to me is, exactly how skewed is the data, and is there a formal result out there relating the skewness to the required sample size. – zkurtz Sep 13 '13 at 16:36

I have posted an answer to that question. We know (at least approximately) how skewed the data are based on the summary statistics in the question. That skew is so strong that neither 300, nor 3000, nor even 30,000 observations per group will make the sampling distribution of the mean "almost exactly normal." You probably need around 300,000 or so before that claim becomes plausible. Thus we must seek a different explanation for why the two tests agree. Mine is that *neither* is "well-behaved" rather than that both are well-behaved. – whuber ♦ Sep 13 '13 at 16:51