Re: Standard Deviation & the 68−95−99.7 rule

Source: http://sci.tech−archive.net/Archive/sci.math/2007−12/msg04456.html

- *From*: Ray Vickson <RGVickson@xxxxxxx>
- *Date*: Fri, 21 Dec 2007 16:47:06 −0800 (PST)

On Dec 21, 3:25 pm, Maya <maya_s...@xxxxxxxxxxx> wrote:

> On Dec 21, 3:12 pm, Virgil <Vir...@xxxxxxx> wrote:
>
> > In article
> > <befa0c10−aa71−454d−8b95−60c00767b...@xxxxxxxxxxxxxxxxxxxxxxxxxxxx>,
> >
> > Maya <maya_s...@xxxxxxxxxxx> wrote:
> >
> > > On Dec 21, 1:58 pm, Maya <maya_s...@xxxxxxxxxxx>
> > > wrote:
> >
> > > > On Dec 21, 1:02 pm,
> > > > "FredJeffr...@xxxxxxxxx"
> > > > <FredJeffr...@xxxxxxxxx>
> > > > wrote:
> >
> > > > > On Dec 21, 11:39 am, Maya
> > > > > <maya_s...@xxxxxxxxxxx>
> > > > > wrote:
> >
> > > > > > At the
> > > > > > bottom of
> > > > > > the intro to
> > > > > > the
> > > > > > Wikipedia
> > > > > > entry on the
> > > > > > 68−95−99.7
> > > > > > rule, it
> > > > > > states:

"This rule is often used to quickly get a rough estimate of something's probability, given its standard deviation."

What an awful sentence.

What " thing's " probability could I estimate, given the thing's standard deviation? Let's say I have this data set: {6, 6, 8, 8} . It's standard deviation is 1. So, given its "1", I can estiate the probability of ..... what?

http://en.wikipedia.org/wiki/68−95−99.7_rule

You also need the mean, in the case of your data set 7. So the 68−95−99.7 rule says that about 68% of observations

will be within 1
of 7 (between 6 and 8), 95%
within 2 of 7 (between 5 and
9) and 99.7%
within 3 of 7 (between 4 and
10) IF your data set were
distributed
normally.

Say you have a normally
distributed data set with
mean 7 and standard
deviation 1. Pick an element
at random from your set.
The probability
of that element's being
between 6 and 8 is 68%, the
probability of its
being between 5 and 9 is
95%, etc.

There is a better example at
the bottom of this
page:http://www−stat.stanford.edu/~naras/jsm/NormalDensity/NormalDensity
tml

Thanks Fred.

This stuff seems to be going in a circle. The
Empirical Rule applies
only to Normal Distributions. So I can
ascertain some things about the
data points in a normal distribution by
applying the empirical rule,

No. If you know the distribution is normal, or nearly so, you can get exact probabilities, so there is no need at all for an empirical rule (although it may still be useful if you lack access to a scientific calculator or statistical tables, etc.). The so−called "empirical rule" is /derived/ from exact calculations on normal distributions.

> but I should only apply the empirical rule if
> I'm first sure that the
> data set is a normal distribution!

If the distribution is a bit different from normal and not too badly skewed, the empirical rules may at least give you figures in the right ballpark. (Of course, it is hard to say exactly what is meant by "a bit different" and "not too badly ... ".) Typically, these things are tested experimentally, perhaps using simulation

> I'm trying to find a real−world use for
> Standard Deviation

Well, for one thing, it can give you bounds that are often useful and are certainly better than nothing. Chebychev's Inequality says that / for any kind of distribution at all/ (discrete, continuous,..., anything) that has finite mean m and finite standard deviation s, the probability that the random variable X deviates from m by more than k standard deviations is <= 1/k^2. So, the probability that you are more than 1 standard deviation from m is <= 1 (not very useful!), that you are more than 2s away from m is <= 1/4, that you are more than 3s away from m is <= 1/9, etc. In other words, for any distribution at all, the chance of being within 2s of m is at least 3/4, of being within 3s of m is at least 8/9, etc. In the special case of the normal distribution we can give better answers, but the bounds are sometimes good enough in applications. Also, in some types of applied models, quantities like system costs, etc., may be quadratic functions of the random variable, in which case evaluating the expected cost needs only the mean and the variance. Also, the standard deviation is, in general, a type of "spread" measure that is often used in real−world applications. For example, Markowitz won the Nobel Prize in Economics for his work on mean−variance investment−portfolio analysis. These concepts are widely used in designing mutual funds and other types of investment instruments.

> and the
>
> > Empirical Rule, but so far it seems the only
> > uses are to tell me
> > things about a data set if and only if I
> > already know those very
> > things about the data set are already true.−
> > Hide quoted text −

− Show quoted text −

> I thought the result of calculating the Standard Deviation of a data
> set would tell me whether the data set's distribution is Normal,
> Continuous, or Discrete?

Absolutely not. If you give me the mean and the standard deviation I can construct several different discrete and several different continuous distributions that have that mean and variance.

> If it can't tell me that, then what good is it to know that data points are either: 1)close to the mean, or 2)not so close to the mean ?

It depends on the application: who wants to know the answer to that question, and why is it important to them?

> Among other things, this sort of information is used in statistical hypothesis testing, q.v.− Hide quoted text −
>
> − Show quoted text −

> IN what way? What would be a simple example of the good of knowing only the SD without knowing which type of distribution the data set is?

See some of the reasons I listed above.

R.G. Vickson

.