Inferential Statistics using the Coefficient of Variation

Michael Lloyd, Ph.D.
Professor of Mathematics

Abstract
Recall that the Coefficient of Variation $CV = 100s/\bar{x}$ is a unit-less measure of spread with respect to the center of a data set. The distributions of the standard deviation s and CV will be derived and statistical inference involving the CV are discussed.

Green lynx spider attacking a wasp

Consider the following example: "A biologist who studies spiders believes that not only do female green lynx spiders tend to be longer than their male counterparts, but also that the lengths of the female spiders seem to vary more than those of the male spiders. We shall test whether this latter belief is true."

Here are data for the lengths of lynx spiders in millimeters:

Lengths of males
5.20 4.70 5.75 7.50 6.45 6.55 4.70 4.80 5.95
5.20 6.35 6.95 5.70 6.20 5.40 6.20 5.85 6.80
5.65 5.50 5.65 5.85 5.75 6.35 5.75 5.95 5.90
7.00 6.10 5.80

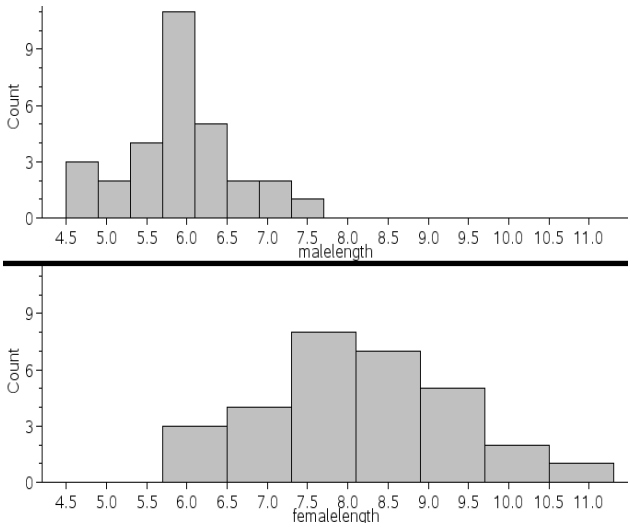
Lengths of females
8.25 9.95 5.90 7.05 8.45 7.55 9.80 10.80 6.60
7.55 8.10 9.10 6.10 9.30 8.75 7.00 7.80 8.00
9.00 6.30 8.35 8.70 8.00 7.50 9.50 8.30 7.05
8.30 7.95 9.60

We will assume that these are independent simple random samples, and that the shapes of the accompanying histograms are approximately Normal. Thus, we are justified in applying the 2-sample F-test for comparing standard deviations. The hypotheses are

$$H_0: \sigma_M = \sigma_F \text{ and } H_a: \sigma_M < \sigma_F$$

where M is the male length and F is the female length variable, respectively.

The F statistic is $s_F^2/s_M^2 = 3.21$ with a degrees of freedom of (29,29) and a p-value of 0.0012.



The small p-value is evidence that the variance in the lengths of the females is greater than that of the males. However, doing a 2-sample t-test will provide evidence that female lynx spiders are longer on average than male lynx spiders. Since the female spiders tend to be longer, it would be more appropriate to compare the sex difference in length for the lynx spider using a statistic that measures variation relative to length. The coefficient of variation is such a statistic and is defined to be $CV = 100\frac{s}{\bar{x}}$. For the above data,

CV_M = 100*.663/5.92 = 11.2 and CV_F = 100*1.19/8.15 = 14.6, so $CV_M < CV_F$,

but is this difference significant? We will need to determine the distribution of the CV random variable, so for simplicity the 100 in the formula will be dropped and we redefine $CV = \frac{s}{\bar{x}}$.

We will assume for the remainder of this paper that $X_1, X_2, ..., X_n$ is a simple random sample from $N(\mu, \sigma)$ where $\mu > 0$. Recall the following theorem:

Theorem $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ and $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X - X_j\right)^2$ are independent, and $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

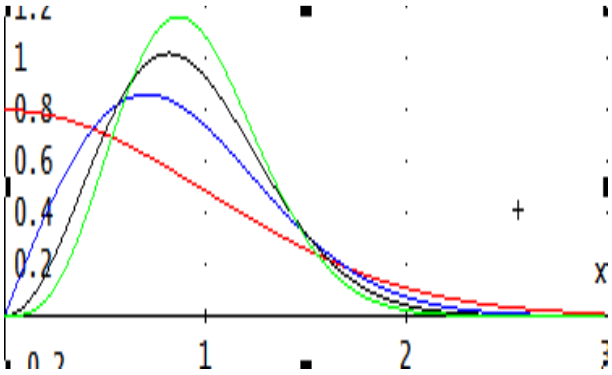
We will derive the distribution of the standard deviation. Recall that the probability density function of $U \sim \chi^2(n-1)$ is $f_U(u) = \frac{u^{\frac{n-3}{2}} e^{-\frac{u}{2}}}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}}, u > 0, n = 2,3,\ldots$. Write $S = \sqrt{\frac{\sigma^2}{(n-1)} \cdot \frac{(n-1)S^2}{\sigma^2}} = \frac{\sigma}{\sqrt{n-1}}\sqrt{U}$. Apply the change-of-variable technique to obtain

$$f_S(s) = f_U(u)\frac{du}{ds}$$

$$= \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}}\left[\frac{(n-1)s^2}{\sigma^2}\right]^{\frac{n-3}{2}} \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right]\frac{d}{ds}\left[\frac{(n-1)s^2}{\sigma^2}\right]$$

$$= \frac{s^{n-2}}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-3}{2}}}\left[\frac{(n-1)}{\sigma^2}\right]^{\frac{n-1}{2}} \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right], s > 0$$

Here are graphs of some probability distribution functions for the standard deviation and a formula for its mean and variance:



Graph of S pdf for n=2,3,4,5 using σ=1

$$E[S]=\frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)}\sigma\to\sigma^{-}\ \text{as}\ n\to\infty$$

$$Var[S]=\left[1-\frac{2\Gamma\left(\frac{n}{2}\right)^2}{(n-1)\Gamma\left(\frac{n-1}{2}\right)^2}\right]\sigma^2\to 0\ \text{as}\ n\to\infty$$

Mean and Variance of S

To find a formula for the distribution of the CV, consider the joint distribution of $(\overline{X},S)$. By independence, the joint probability density function is $f_{\overline{X}}f_S$ with support $\mathrm{R}\times\mathrm{R}^{+}$. The probability density function for S was derived earlier, and the probability mass function for $\overline{X}$ is easily computed to be $f_{\overline{X}}(x)=\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{n(x-\mu)^2}{2\sigma^2}\right)$, $x\in\mathrm{R}$. Therefore, the cumulative distribution function for CV is

$$F_{CV}(c)=P\left[\frac{S}{\overline{X}}\le c\right]=P\left[\left(S\le c\overline{X}\right)\cap\left(\overline{X}>0\right)\right]+P[\overline{X}<0]$$

$$=\int_0^{\infty}\int_0^{cx}f_{\overline{X}}(x)f_S(s)dsdx+P[\overline{X}<0],\ c>0$$

(If n is sufficiently large, then $P[CV<0]\approx 0$.)

Differentiate the cumulative distribution function to obtain the probability density function:

$$f_{CV}(c)=F_{CV}^{'}(c)=\int_0^{\infty}f_{\overline{X}}(x)f_S(cx)\frac{d(cx)}{dc}dx+0$$

$$=\int_0^{\infty}\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\exp\left[-\frac{n(x-\mu)^2}{2\sigma^2}\right]\frac{(cx)^{n-2}x}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-3}{2}}}\left[\frac{n-1}{\sigma^2}\right]^{\frac{n-1}{2}}\exp\left[-\frac{(n-1)(cx)^2}{2\sigma^2}\right]dx$$

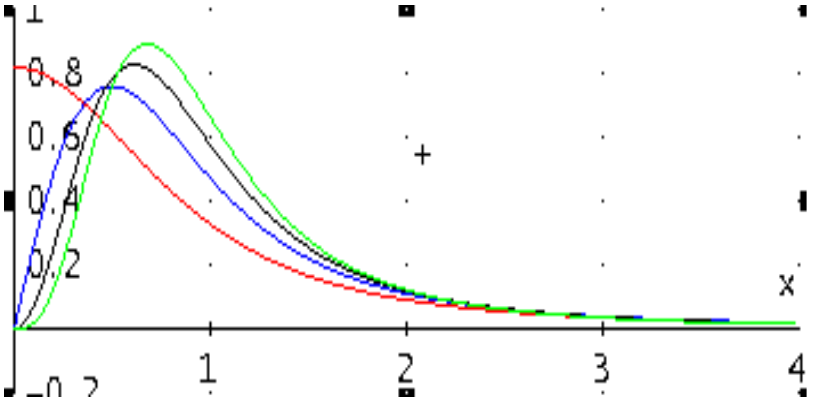
$$=\left(\frac{n}{\pi}\right)^{\frac{1}{2}}\frac{c^{n-2}(n-1)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)\sigma^n 2^{\frac{n-2}{2}}}\exp\left[-\frac{n\mu^2}{2\sigma^2}\right]\int_0^{\infty}x^{n-1}\exp\left[-\frac{(n-1)c^2+n}{2\sigma^2}x^2+\frac{n\mu x}{\sigma^2}\right]dx,$$

$$c>0$$

For the case $n = 2$, this formula simplifies to $f_{CV}(c) =$

$$\frac{\sqrt{2}\cdot e^{-\mu^2/\sigma^2}\cdot\left(\sqrt{2}\cdot\sqrt{\pi}\cdot\mu\cdot e^{2\cdot\mu^2/(\sigma^2\cdot(c^2+2))}\cdot\left(\mathrm{ERF}\left(\frac{\sqrt{2}\cdot\mu}{\sigma\cdot\sqrt{(c^2+2)}}\right)+1\right)+\sigma\cdot\sqrt{(c^2+2)}\right)}{\pi\cdot\sigma\cdot(c^2+2)^{3/2}}$$

where $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,dt$. I could not find general formula for all *n*, but the computer algebra system Derive can find a formula for small *n*. Here are graphs of the CV for $n = 2,3,4,5; \mu = 1; \sigma = 1$. The areas under these curves are .921, .958, .977, .987, respectively. These areas are less than 1 because μ and n are small and the $P[CV < 0]$ term in the derivation of the probability density function was ignored. However, note that the area under the probability density function appears to converge to 1 rapidly as n converges to infinity.

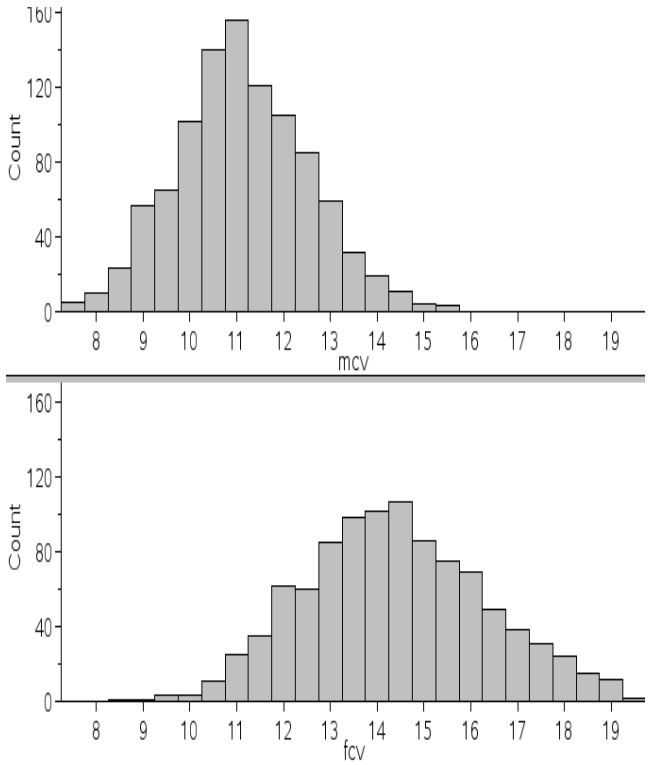


Since it appeared hopeless to find a formula for the distribution of the CV for $n = 30$, I attempted to directly compute the probability using $P[CV_M < CV_F] = \int_0^\infty \int_0^{c_M} f_M(c_M) f_F(c_F)\,dc_F\,dc_M$. However, both Derive and Maple were unable to compute f_M or f_F because n=30 was apparently too large. For example,

$$Fm := \int_0^\infty 1.084400797\cdot10^{-506}\cdot s^{28}\cdot x^{29}\cdot e^{-x^2\cdot(32.96318745\cdot s^2+34.09984909)}\cdot(1.755435499\cdot10^{175}\,x)\,dx$$

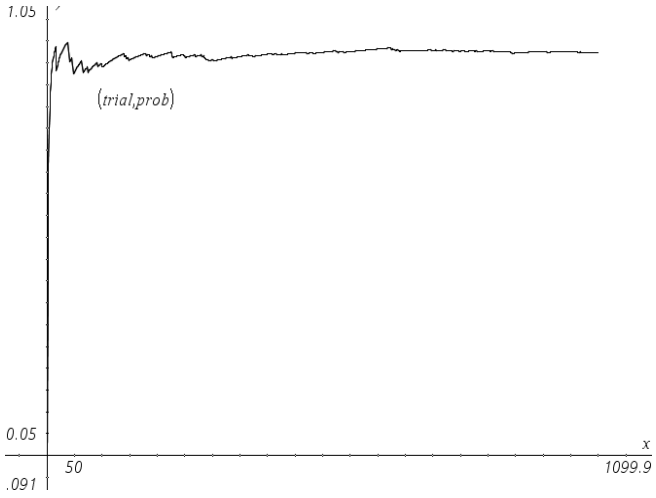
could not be evaluated. My last approach was to try a simulation which was successful.

One thousand samples of size 30 were simulated using the accompanying TI-nspire program. The parameters μ_M=5.92, σ_M=0.663, μ_F=8.15, σ_F=1.18 were estimated from the sample statistics. Here are the histograms for the simulated coefficients of variation:




```
simcv                                                    1/
Define simcv(num)=
Prgm
Local c,dat,n
fcv:=newList(num)
mcv:=newList(num)
prob:=newList(num)
trial:=newList(num)
c:=0
For n,1,num
  trial[n]:=n
  dat:=randNorm(5.91667,.663238,30)
  mcv[n]:=100·stDevSamp(dat)/mean(dat)
  dat:=randNorm(8.15333,1.18743,30)
  fcv[n]:=100·stDevSamp(dat)/mean(dat)
  If mcv[n]<fcv[n]
    c:=c+1
  prob[n]:=c/n
EndFor
```

The variable mcv is the male CV, and the variable fcv is the female CV, respectively. The accompanying diagram is a typical Law-of- Large Numbers graph showing the empirical probability verses the number of trials generated using the above program. This probability estimated using $n = 1000$ gives $P[CV_M < CV_F] \approx 0.923$. This probability is less than 0.95, so this provides evidence that the female spider CV does not appear to be significantly larger than the male spider CV.



Conclusion

Unless the distributions can be modeled with simpler functions, it is impractical to directly do statistical inference using the CV. Also, perhaps simulations should not be overlooked as an important tool.

It appears that the mean of CV is approximately σ/μ, and its variance approaches zero as n approaches infinity. Further investigation would be to provide rigorous proofs or numerical evidence for these conjectures.

References

- *Probability and Statistical Inference* by Hogg and Tanis , 7th edition ©2006 Pearson.
- http://creatures.ifas.ufl.edu/beneficial/green_lynx04.htm

Biography

Michael Lloyd received his B.S in Chemical Engineering in 1984 and accepted a position at Henderson State University in 1993 after earning his Ph.D. in Mathematics from Kansas State University. He has presented papers at meetings of the Academy of Economics and Finance, the American Mathematical Society, the Arkansas Conference on Teaching, the Mathematical Association of America, and the Southwest Arkansas Council of Teachers of Mathematics. He has also been an AP statistics consultant since 2001 and a member of the American Statistical Association.