A Justification for the Standard Deviation in a Two – Sample Independent T - Test

In class, I seem to remember quite a bit of dry – heaving when you all were presented with the test statistic for the two – sample t – test (for unmatched samples):

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Now, I'm pretty sure the dry – heaving was due to the denominator, not the numerator. Here's how I need you to look at that test statistic: it's a *z* – score. Well, actually, it's a ***t*** – score, but let's not pick nits: it's a way of turning raw data (the data in sets 1 and 2) into standardized data by subtracting the assumed mean (0) from the difference in the sample means and dividing by the assumed standard deviation (which we're going to explore here).

First off, I'd like you to square the denominator. That'll leave you with the ***variance***:

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

Now why on earth would we do that? Well, as it turns out, the variance is easier to deal with than the standard deviation, since the sample variance (as you may or may not remember) is an unbiased estimator of the population variance. Once we've dealt with the variance, we can simply take the square root again to arrive at the sample standard deviation.

So, we have two independent data sets. These independent data sets have their own unique sample variances, and we call them $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. However, to deal with the data sets as one entity, we need to combine the variances. From the previous paragraph, it appears as though that means we need to add them...but why does ***that*** make sense?

In fact, it doesn't make ***any*** sense to me! We're looking at the ***difference*** of the two data sets 1 and 2, not the ***sum***. Why shouldn't we subtract the data sets' variances instead?

OK, OK, OK....let's break it down to the idea of what a variance is...if you remember from MTH 243, the idea of variance is just that: how much data varies. But varies from ***what***? Well, we defined it to be variance from the average, or expected value. So, what we need to find is, on average, how much each data point differs from the data set's average. Then, since we're dealing with variances, we square the result. In formal language (which is easier to use, as you will see), this is $\mathrm{VAR}[X] = \mathrm{E}\{[X - \mathrm{E}(X)]^2\}$.

Where our case is more interesting is this: our data points are actually ***differences***. So, our variance formal language will look like this:

$$VAR[X_1 - X_2] = E\left\{\left[(X_1 - X_2) - E(X_1 - X_2)\right]^2\right\}$$

Now, though this looks totally vomit – inducing, it's really just a little algebraic relationship that we can now play with. For starters, look the at the ***E***[***X***$_1$ – ***X***$_2$] part. It's asking this question: if I randomly select a data point from data set 1, and subtract a data point from set 2, what difference do I ***expect***? Well, I typically expect the average, don't I? And what is that average of that difference? It's just $\bar{x}_1 - \bar{x}_2$. You knew that from the test statistic! So, let's substitute that back into the previous equation:

$$VAR\left[X_1 - X_2\right] = E\left\{\left[\left(X_1 - X_2\right) - E\left(X_1 - X_2\right)\right]^2\right\} = E\left\{\left[\left(X_1 - X_2\right) - \left(\bar{x}_1 - \bar{x}_2\right)\right]^2\right\}$$

Now, if you like algebra, here comes the good part (and if you don't, I'd stop reading now!):

$$\begin{aligned} VAR\left[X_1 - X_2\right] &= E\left\{\left[\left(X_1 - X_2\right) - \left(\bar{x}_1 - \bar{x}_2\right)\right]^2\right\} \\ &= E\left\{X^2_1 - 2X_1X_2 + X^2_2 - 2X_1\bar{x}_1 + 2X_2\bar{x}_1 + 2X_1\bar{x}_2 - 2X_2\bar{x}_2 + \bar{x}^2_1 - 2\bar{x}_1\bar{x}_2 + \bar{x}^2_2\right\} \\ &= E\left\{X^2_1 - 2X_1\bar{x}_1 + \bar{x}^2_1 + X^2_2 - 2X_2\bar{x}_2 + \bar{x}^2_2 - 2X_1X_2 + 2X_2\bar{x}_1 + 2X_1\bar{x}_2 - 2\bar{x}_1\bar{x}_2\right\} \\ &= E\left\{\left[X_1 - \bar{x}_1\right] + \left[X_2 - \bar{x}_2\right] - 2X_1X_2 + 2X_2\bar{x}_1 + 2X_1\bar{x}_2 - 2\bar{x}_1\bar{x}_2\right\} \end{aligned}$$

Now, those two bracketed items in the last line are just the variances of data set 1 and 2, respectively, so

$$\begin{aligned} VAR\left[X_1 - X_2\right] &= E\left\{\left[X_1 - \bar{x}_1\right] + \left[X_2 - \bar{x}_2\right] - 2X_1X_2 + 2X_2\bar{x}_1 + 2X_1\bar{x}_2 - 2\bar{x}_1\bar{x}_2\right\} \\ &= VAR(X_1) + VAR(X_2) + E\left\{-2X_1X_2 + 2X_2\bar{x}_1 + 2X_1\bar{x}_2 - 2\bar{x}_1\bar{x}_2\right\} \end{aligned}$$

We're almost there, believe it or not...we just need to legally get rid of the expected quantity in the brackets at the end of the expression. I'll start by factoring a little:

$$\begin{aligned} VAR\left[X_1 - X_2\right] &= VAR(X_1) + VAR(X_2) + E\left\{-2X_1X_2 + 2X_2\bar{x}_1 + 2X_1\bar{x}_2 - 2\bar{x}_1\bar{x}_2\right\} \\ &= VAR(X_1) + VAR(X_2) - 2E\left\{X_1X_2 - X_2\bar{x}_1 - X_1\bar{x}_2 + \bar{x}_1\bar{x}_2\right\} \end{aligned}$$

Now, I'm going to play with that last term a bit. I'll start by taking the expectation across all four of its terms:

$$VAR\left[X_1 - X_2\right] = VAR(X_1) + VAR(X_2) - 2\left\{E\left(X_1X_2\right) - E\left(X_2\bar{x}_1\right) - E\left(X_1\bar{x}_2\right) + E\left(\bar{x}_1\bar{x}_2\right)\right\}$$

Now, realizing that the expectation of an average is just the average itself (it's not really "expected"...it's known):

$$\begin{aligned}VAR\left[X_1-X_2\right]&=VAR(X_1)+VAR(X_2)-2\left\{E\left(X_1X_2\right)-E\left(X_2\bar{x}_1\right)-E\left(X_1\bar{x}_2\right)+E\left(\bar{x}_1\bar{x}_2\right)\right\}\\&=VAR(X_1)+VAR(X_2)-2\left\{E\left(X_1X_2\right)-\bar{x}_1E\left(X_2\right)-\bar{x}_2E\left(X_1\right)+\bar{x}_1\bar{x}_2\right\}\\&=VAR(X_1)+VAR(X_2)-2\left\{E\left(X_1X_2\right)-\bar{x}_1\bar{x}_2-\bar{x}_2\bar{x}_1+\bar{x}_1\bar{x}_2\right\}\\&=VAR(X_1)+VAR(X_2)-2\left\{E\left(X_1X_2\right)-\bar{x}_1\bar{x}_2\right\}\\&=VAR(X_1)+VAR(X_2)-2E\left(X_1X_2\right)+2\bar{x}_1\bar{x}_2\end{aligned}$$

So, we now have a unique looking little quantity there at the third term, no? It's asking me, "if I multiply two random data points, one from data set 1, and one from data set 2, what can I expect their product to be?" Well, here's where the fact that the data sets are independent[1] comes in: since the variables from data set 1 and data set 2 don't depend on one another, it must be that $E\left(X_1X_2\right)=\bar{x}_1\bar{x}_2$. This means the following:

$$\begin{aligned}VAR\left[X_1-X_2\right]&=VAR(X_1)+VAR(X_2)-2E\left(X_1X_2\right)+2\bar{x}_1\bar{x}_2\\&=VAR(X_1)+VAR(X_2)-2\bar{x}_1\bar{x}_2+2\bar{x}_1\bar{x}_2\\&=VAR(X_1)+VAR(X_2)\end{aligned}$$

Hell YEAH! We just proved that, so long as two data sets are independent, the variance of the difference of the data sets is found by just adding the individual variances together[2]. So, you should feel much better about seeing that lovely denominator. Here's how its variance shakes out (with a little help from the CLT):

$$\begin{aligned}VAR\left[x_1-x_2\right]&=VAR(x_1)+VAR(x_2)\\&=\frac{s_1^{\ 2}}{n_1}+\frac{s_2^{\ 2}}{n_2}\end{aligned}$$

And so the standard deviation will just be

$$\begin{aligned}STDEV\left[x_1-x_2\right]&=\sqrt{VAR\left[x_1-x_2\right]}\\&=\sqrt{VAR(x_1)+VAR(x_2)}\\&=\sqrt{\frac{s_1^{\ 2}}{n_1}+\frac{s_2^{\ 2}}{n_2}}\end{aligned}$$

Now don't you feel better? I know ***I*** do!

[1] This whole argument hinges on this fact, just so you know!

[2] What's pretty kooky is that $VAR\left[X_1+X_2\right]=VAR(X_1)+VAR(X_2)$, too.