Homogeneity of Variances
By Dr. Jon Starkweather, Research and Statistical Support consultant

This month we touch on a fundamental issue in statistical evaluation that often gets overlooked. Testing assumptions for parametric analysis *is* a fundamental step and a necessary one. For example, let's consider some of the simplest experimental design analysis available; the independent *t*-test and analysis of variance *F* test—testing for mean differences among independent groups. These tests have three key assumptions; normality, independence of observations, and **h**omogeneity **o**f **v**ariances (**HOV**). Generally speaking, experimental design dictates random sampling from a well defined population and random assignment to groups (aka. conditions, levels, etc.), both of which *should* help take care of assumptions mentioned. But, let's focus our attention on the third assumption (HOV), which needs to be (and can be) tested to ensure accurate or valid interpretation of the mean differences. Luckily, most statistical software packages offer a way to test for HOV (including PASW/SPSS). Generally, the Levene's test is used to statistically test the amount of difference between variances (of groups selected for a *t*-test or *F* test).

Means vs. variances, a Royal Rumble… Levene's test is testing for differences among our group's (2 or more) **variances**. A *t*-test is testing for differences among 2 group's **means**. An *F* test (one-way ANOVA) is testing for differences among more than 2 group's **means**. In these contexts, the independent variable is comprised of multiple groups; for the *t*-test, there are two groups; for the *F* test, there are more than two groups. Each group represents a treatment or lack of one in the case of a placebo. In essence, each group receives something different as stimulus, for example different drugs administered in each condition of an efficacy study.

Essentially, whether looking at 2 groups (*t*-test) or more than 2 groups (*F* test), we are ***concerned*** with the assumption of homogeneity of variances (among other assumptions). Recall that **variance is** a measure of dispersion, how much do the scores (of one group) VARY around the mean (whatever that mean happens to be). **Mean is** a measure of central tendency; arithmetic average. The HOV assumption states that our groups are similar in essence (similar variances), regardless of independent variable level (treatment or condition administered).

Providing some practical examples for the discussion.
Treatment administered is what each group experiences that is different, each level of the **independent variable**. Treatment in this sense represents your independent variable; that which is manipulated in the experiment and you are trying to establish that each level or group displays mean differences. The groups for our novel example will be **Zoloft vs. Xanax vs. Lithium vs. Placebo (a lack of treatment).** The **dependent variable** is that which is used to measure change in the independent variable, which for this example will be **Statistics Anxiety scores**

If we randomly sample introductory statistics class undergraduates at UNT, then randomly assign them to 4 treatment groups; we are ASSUMING the students are similar (homogeneous). Because they are all undergraduates taking intro stats at the same university and we randomly sampled and randomly assigned; thus equalizing individual differences (hair & eye color, etc.). However, if we realize after randomly assigning them to our groups that most of one group was made up of engineering majors and one group was almost completely made up of music majors

while the third was made up of primarily English majors and the fourth primarily physics majors; then we can see that the groups likely differ regardless of treatment administered.

Stated another way, we are likely to have significant group differences (and heterogeneous variances), because each group is different in essence and is likely to differ on our dependent variable (Statistics Anxiety scores). Already, you can imagine English and music majors having higher Statistics Anxiety scores than the physics and engineering majors, simply because they have different mathematics course requirements and likely different interests.

Why is this important? Well, if our groups are inherently different, then any differences we find in our dependent variable (the mean stats. anxiety scores) after administering treatments (the drugs) may have been due to the treatments OR the inherent differences of the groups (we wouldn't know which). In which case, whatever statistical test (t-test, or F test) results we find are of no practical validity. We cannot be confident that our Zoloft group displayed less Statistics Anxiety due to the Zoloft, because they may simply have been better or more relaxed with statistics due to a background heavy in mathematics.

A more precise example (stay with me now…)
Imagine administering 150 mg. of Xanax (a commonly prescribed anti-anxiety drug) to one group and administering a placebo (inert pill) to the other group (2 groups only = t-test). Our dependent variable is Stats. Anxiety scores again. Well, does each individual respond to the same dosage (say 150 mg.) of any drug differently? Consider something as simple as body weight, more body weight = more volume of drug required to have the desired effect (makes sense doesn't it?). Obviously its more complex than that, physiology, liver functioning, tolerance, etc. each plays a part. BUT we all understand that six shots of vodka for me (approx. 180 lbs.) is going to have a different effect than for my fraternal twin (approx. 140 lbs.; usually on the floor drooling on himself after six shots!). SO; if each person in our Xanax group reacts differently to the 150 mg. Xanax, then that group is likely to have more variance than our placebo group. Which is likely to have very little variance because they were administered no active drug…therefore, each person's weight, physiology, liver functioning, tolerance, etc. will not matter in the placebo group—regardless of mean Statistics Anxiety score!

You can now see that when looking at the variances of statistics anxiety scores we might find differences based on body weight, not necessarily on who got the Xanax and who didn't. REMEMBER, our t-tests or F tests are testing for differences among the group **means**. Levene's test is testing for differences among group **variances**.

Another way of **Looking** at it.
Consider a few distributions each with different variance:
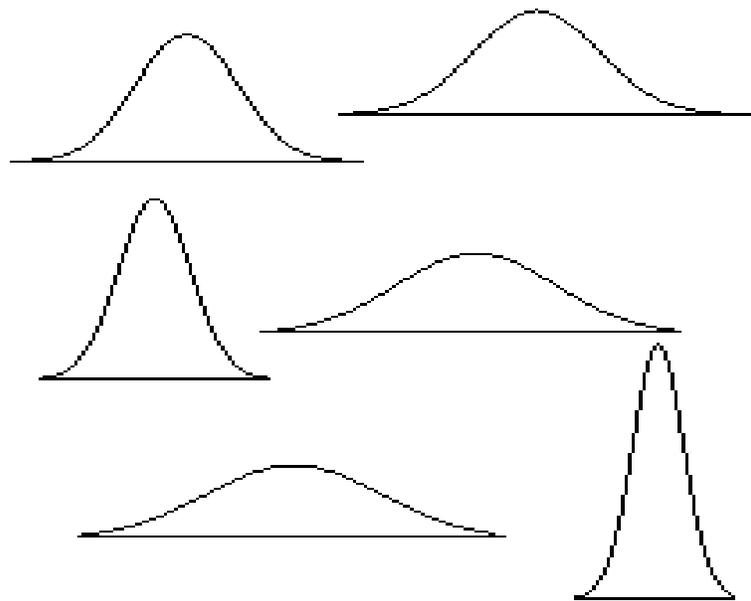
Imagine each of these represents one of our groups; Zoloft, Xanax, Mountain Dew, coffee, alcohol and placebo… You can see it makes no difference what mean (stats. anxiety score) happens to be under the middle of each distribution; they are different from one another in their variance. Inherently different groups! Stated another way; each group responded to their respective treatment differently. Some group's participants were more similar (low variability or a narrow distribution), while others were more different (high variability or a wide distribution). How does this relate to the Levene's test of the HOV assumption?

Recall; the **H**omogeneity **O**f **V**ariances assumption stipulates that our groups have similar variances; similar reactions to the treatment/condition/drug they received. If this assumption holds then we know that whatever test result (*t*-test or *F* test) we find is attributable to the different treatment (drug) each group received (treatment effects, not confounds). Furthermore, recall that Levene's test is testing whether or not the variances of our groups are statistically different. We generally use the .05 probability level (or "*Sig*." value) to determine statistical significance; so, if Levene's test shows a "Sig." value of less than (<) .05; then we conclude that the variances are significantly different; meaning our statistical test (*t*-test or *F* test) is invalid and we can't make conclusive inferences from it. Likewise, if Levene's test shows a "Sig." value of greater than (>) .05; then we conclude the variances are NOT significantly different---which is what we **want** to see so that we can have confidence in the validity our *t*-test or *F* test result.

Additional discussion of Heterogeneity of Variance:
Bryk, A. & Raudenbush, S. (1988). Heterogeneity of variance in experimental studies: A
    challenge to conventional interpretations. *Psychological Bulletin, 104*(3), 396 – 404.
    DOI: 10.1037/0033-2909.104.3.396

Until next time, you don't need a weatherman to know which way the wind blows…