Richard Gonzalez
Psych 613
Version 3.1 (Sep 2022)

LECTURE NOTES #2

Reading assignment

- Read MD ch 3 or G ch 6.

Goals for Lecture Notes #2

- Introduce decomposition of sum of squares
- Introduce the structural model
- Review assumptions and provide more detail on how to check assumptions

1. Fixed Effects One Way ANOVA

I will present Analysis of Variance (ANOVA) several different ways to develop intuition. The different methods I will present are equivalent—they are simply different ways of looking at and thinking about the same statistical technique. Each way of thinking about ANOVA highlights a different feature and provides different insight into the procedure.

(a) ANOVA as a generalization of the two sample *t* test

Example showing $t^2 = F$ (Figure 2-1). This is not a transformation of the data, but a relationship between two different approaches to the same test statistic.

ANOVA's null hypothesis

The null hypothesis for the one way ANOVA is:

$$\mu_i = \mu \quad \text{for all i} \tag{2-1}$$

In words, the means in each population i are all assumed to equal the same population mean μ. This is a generalization of the equality of two means that is tested by the two-sample t test. The alternative hypothesis is that at least one of the population group means is not equal to the average value μ. Rejection of the null doesn't point to a particular alternative as there are many possible patterns across the group means that

would fit the alternative hypothesis. We will return to this point in Lecture Notes 3 when we cover contrasts and post hoc tests.

(b) Structural model approach

Let each data point be denoted by Y_{ij}, where i denotes the group the subject belongs and j denotes that the subject is the "jth" person in the group.

Consider the following example from Kirk (1982). You want to examine the effects of sleep deprivation on reaction time. The subject's task is to press a key when he or she observes a stimulus light on the computer monitor. The dependent measure is reaction time measured in hundredths of a second. You randomly assign 32 subjects into one of four sleep deprivation conditions: 12 hrs, 24 hrs, 36 hrs, and 48 hrs. Your research hypothesis is that reaction time will slow down as sleep deprivation increases.

Here are the data:

	A_1 12 (hrs)	A_2 24 (hrs)	A_3 36 (hrs)	A_4 48 (hrs)
	20	21	25	26
	20	20	23	27
	17	21	22	24
	19	22	23	27
	20	20	21	25
	19	20	22	28
	21	23	22	26
	19	19	23	27
group mean	19.38	20.75	22.63	26.25
st. dv.	1.19	1.28	1.189	1.28

See Appendix 1 for exploratory data analysis of these data.

The "grand mean", i.e., the mean of the cell means, is denoted $\hat{\mu}$. For these data $\hat{\mu} = 22.25$. The grand mean is an unbiased estimate of the true population mean μ. I will use the symbols $\hat{\mu}$ and $\overline{Y}$ interchangeably for the sample mean. When the cells have equal sample sizes, the grand mean can also be computed by taking the mean of all data, regardless of group membership. This will equal the (unweighted) mean of the cell means. However, when the cells have different sample sizes, then the grand mean is computed by taking the mean of the cell means. There are other definitions of "grand mean" that we'll cover later, such as the sample-size weighted mean of cell means.

Figure 2-1: Example showing equivalence of *t* test and ANOVA with two groups with SPSS and R syntax.

I'll use the first two groups of the sleep deprivation data that we will discuss later.

data list file = "data.sleep" free/ dv group.

select if (group le 2).

TO SAVE SPACE I OMIT THE OUTPUT OF THE BOXPLOTS ETC

t-test groups=group
/variables=dv.

```
t-tests for independent samples of  GROUP

Variable         Number              Standard    Standard
                of Cases     Mean   Deviation      Error
-------------------------------
DV
      GROUP 1       8      19.3750      1.188       .420
      GROUP 2       8      20.7500      1.282       .453


                | Pooled Variance estimate | Separate Variance Estimate
                |                          |
    F   2-tail  |    t   Degrees of 2-tail |    t    Degrees of  2-tail
  Value  Prob.  |  Value   Freedom   Prob. |  Value    Freedom    Prob.
   1.16   .846  |  -2.23     14       .043  |  -2.23     13.92     .0430
```

NOW WE'LL DO A ONE WAY ANOVA

oneway variables=dv by group.

```
                                ANALYSIS OF VARIANCE

                                  SUM OF         MEAN             F      F
        SOURCE            D.F.    SQUARES       SQUARES          RATIO  PROB.
BETWEEN GROUPS               1       7.5625        7.5625        4.9532  .0430
WITHIN GROUPS               14      21.3750        1.5268
TOTAL                       15      28.9375
```

The value of the *t* test squared (-2.23^2) equals the value of F in the ANOVA table; the p-values are also identical (compare the two boxes).

R Version of ANOVA (data set has four groups, this example uses groups 1 and 2); same ANOVA F = 4.9532 as SPSS

```
setwd("/Users/gonzo/rich/Teach/Gradst~1/unixfiles/lectnotes/lect2")
data.sleep <- read.table("data.sleep", col.names = c("dv",
    "group"))
data.sleep <- data.sleep[data.sleep[, 2] < 3, ]
data.sleep[, "group"] <- factor(data.sleep[, "group"])
summary(aov(dv ~ group, data = data.sleep))

##             Df Sum Sq Mean Sq F value Pr(>F)
## group        1  7.562   7.562   4.953  0.043 *
## Residuals   14 21.375   1.527
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously, if we don't know how much sleep deprivation a particular subject had, our best prediction of his or her reaction time would be the grand mean. This is not a very precise prediction, but if it's the only thing we have we can live with it. The simple model of an individual data point when you don't know what condition subjects were assigned can be denoted as follows:

$$\mathrm{Y_{ij}} = \mu + \epsilon_{\mathrm{ij}} \tag{2-2}$$

where ϵ_{ij} denotes the error. The error term is assumed to be normally distributed with mean 0 and variance σ^2. In this simple model, we attribute the deviation between Y_{ij} and μ to the error ϵ. Maxwell and Delaney call Equation 2-2 the "reduced model."

However, if we know the amount of sleep deprivation for an individual, then we can improve our prediction by using the group mean instead of the grand mean μ. The group mean can itself be considered a deviation from the grand mean. Thus, the mean of a particular group has two components. One component is the grand mean μ itself, the other component is an "adjustment" that depends on the group the subjects was assigned. This adjustment will be denoted α_{i} to convey the idea that the term is adjusting the "ith" group. An example will make this clearer. The group mean for the 12 hour sleep deprivation group is 19.38. We can define this as:

$$\begin{aligned} \overline{Y}_{\mathrm{i}} &= \hat{\mu} + \hat{\alpha}_{\mathrm{i}} \\ 19.38 &= 22.25 + \hat{\alpha}_{\mathrm{i}} \\ -2.88 &= \hat{\alpha}_{\mathrm{i}} \end{aligned}$$

In words, everyone in the study has a "base" reaction time of 22.25 (the grand mean); being in the 12 hour sleep deprivation group "decreases" the reaction time by 2.88, yielding the cell mean 19.38. The estimate $\hat{\alpha}_{\mathrm{i}}$ represents the effect of being in treatment i, relative to the grand mean. The mean for group i, $\overline{\mathrm{Y}}_{\mathrm{i}}$, is an unbiased estimate of μ_{i}, the population mean of population i.

In one-way ANOVA's the α's are easy to estimate. Simply take the difference between the observed group mean and the observed grand mean, i.e.,

$$\overline{\mathrm{Y}}_{\mathrm{i}} - \overline{\mathrm{Y}} = \hat{\alpha}_{\mathrm{i}} \tag{2-3}$$

The $\hat{\alpha}$'s are constrained to sum to zero, i.e., $\sum \hat{\alpha}_{\mathrm{i}} = 0$. The treatment effect α_{i} is not to be confused with the Type I error rate, which also uses the symbol α.

Basic structural model

This gives us an interesting way to think about what ANOVA is doing. It shows that the one way ANOVA is fitting an additive model to data. That is,

$$\mathrm{Y_{ij}} = \mu + \alpha_{\mathrm{i}} + \epsilon_{\mathrm{ij}} \tag{2-4}$$

So under this model, a given subject's score consists of three components: grand mean, adjustment, and error. The grand mean is the same for all subjects, the treatment adjustment is specific to treatment i, and the error is specific to the jth subject in the ith treatment. Equation 2-4 is what Maxwell and Delaney call the full model.

Understanding the underlying model that is fit provides interesting insights into your data. This is especially true when dealing with more complicated ANOVA designs as we will see later. Note that if additivity does not make sense in your domain of research, then you should not be using ANOVA.

Most textbooks state that the ANOVA tests the null hypothesis that all the group means are equal. This is one way of looking at it. An equivalent way of stating it is that the ANOVA tests the null hypothesis that all the α_{i} are zero. When all the α_{i} are zero, then all the cell means equal the grand mean. So a convenient way to think about ANOVA is that it tests whether all the α_{i}s are equal to zero. The alternative hypothesis is that at least one of the αs are not equal to zero.

A measurement aside: the structural model used in ANOVA implies that the scales (i.e., the dependent variable) should either be interval or ratio because of the notion of additive effects and additive noise (i.e., additive α's and ϵ's).

(c) A third way to formulate ANOVA: Variability between group means v. variability within groups

i. Recall the definition for the estimate of the variance

$$\mathrm{VAR(Y)} = \frac{\sum(Y_{\mathrm{i}} - \overline{\mathrm{Y}})^2}{\mathrm{N} - 1} \tag{2-5}$$

The numerator of the right hand term is, in words, the sum of the squared deviations from the mean. It is usually just called "sum of squares". The denominator is called degrees of freedom. In general, when you compute an estimate of a variance you always divide the sums of squares by the degrees of freedom,

$$\mathrm{VAR(Y)} = \frac{\mathrm{SS}}{\mathrm{df}} \tag{2-6}$$

Another familiar example to drive the idea home. Consider the pooled standard deviation in a two sample *t* test (lecture notes #1). It can also be written as sums of squares divided by degrees of freedom

$$\text{pooled st. dev.} = \sqrt{\frac{\sum(Y_{1\mathrm{j}} - \overline{\mathrm{Y}}_1)^2 + \sum(Y_{2\mathrm{j}} - \overline{\mathrm{Y}}_2)^2}{(\mathrm{n}_1 - 1) + (\mathrm{n}_2 - 1)}} \tag{2-7}$$

$$= \sqrt{\frac{\text{SS from group means}}{\mathrm{df}}} \tag{2-8}$$

You can see the "pooling" in the numerator and in the denominator of Equation 2-7. We'll see that Equation 2-7, which was written out for two groups, can be generalized to an arbitrary number of groups. In words, the pooled standard deviation in

the classic t test is as follows: one simply takes the sum of the sum of squared deviations from each cell mean (the numerator), divides by the total number of subjects minus the number of groups (denominator), and then computes the square root.

Returning to ANOVA, consider the variance of all the data regardless of group membership. This variance has the sum of squared deviations from the grand mean in the numerator and the total sample size minus one in the denominator.

$$\text{VAR(Y)} = \frac{\sum_i \sum_j (Y_{ij} - \overline{Y})^2}{N - 1} \tag{2-9}$$

We call the numerator the sums of squares total (SST). This represents the variability present in the data with respect to the grand mean. This is identical to the variance form given in Equation 2-5.

SST can be decomposed into two parts. One part is the sums of squares between groups (SSB). This represents the sum of squared deviations between the group means and the grand mean (i.e., *intergroup* variability). The second part is the sum of squares within groups (SSW), or sum of squared errors. This represents the sum of squared deviations between individual scores and their respective group mean (i.e., *intragroup* variability). SSW is analogous to the pooled standard deviation we used in the *t* test, but it generalizes the pooling across more than two groups.

The decomposition is such that SST = SSB + SSW. In words, the total variability (as indexed by sum of squares) of the data around the grand mean (SST) is a sum of the variability of the group means around the grand mean (SSB) and the variability of the individual data around the group mean (SSW).

I will now show the derivation of the statement SST = SSB + SSW. We saw earlier that

$$\text{SST} = \sum \sum (Y_{ij} - \overline{Y})^2 \tag{2-10}$$

Decomposition of sums of squares

With a little high school algebra, Equation 2-10 can be rewritten as

$$\text{SST} = \sum_i n_i (\overline{Y}_i - \overline{Y})^2 + \sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2 \tag{2-11}$$

$$= \text{SSB} + \text{SSW} \tag{2-12}$$

(see, for example, Hays, section 10.8, for the proof).

A graphical way to see the decomposition is in terms of a pie chart where the entire pie represents SS total and the pie is decomposed into two pieces–SSB and SSW.

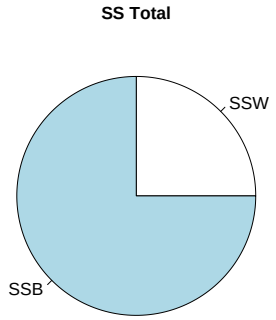


Figure 2-2: Pie chart depicting the ANOVA sum of squares decomposition.

Thus, the analysis of variance performs a decomposition of the variances (actually, the sums of squares). This is why it is called "analysis of variance" even though "analysis of means" may seem more appropriate at first glance.

ii. Degrees of freedom and ANOVA source table

For one-way ANOVA's computing degrees of freedom (df) is easy. The degrees of freedom corresponding to SSB is the number of groups minus 1 (i.e., letting T be the number of groups, T - 1). The degrees of freedom corresponding to SSW is the total number of subjects minus the number of groups (i.e., N - T). Of course, the degrees of freedom corresponding to SST is the total number of subjects minus one (N - 1). Thus, ANOVA also decomposes the degrees of freedom

$$df_{total} = df_{between} + df_{within} \quad (2\text{-}13)$$

$$N - 1 = (T - 1) + (N - T) \quad (2\text{-}14)$$

As always, a sums of squares divided by the degrees of freedom is interpreted as a variance (as per Equation 2-6). So, we can divide SSB and SSW by their appropriate degrees of freedom and end up with variances. These terms are given special names: mean squared between (MSB=$\frac{SSB}{df_{between}}$) and mean squared within (MSW=$\frac{SSW}{df_{within}}$).

The F test is defined as the ratio MSB/MSW where the degrees of freedom for the F test correspond to the degrees of freedom used in MSB and MSW. You reject the null hypothesis if MSB/MSW is greater than the critical F value.

ANOVA source table

To summarize all of this information, the results of an ANOVA are typically arranged in a "source table."

source	SS	df	MS	F
between		T - 1		
within		N - T		
total		N - 1		

The source table for the sleep deprivation example is below (the complete example is given in Appendix 1).

ANALYSIS OF VARIANCE

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F RATIO	F PROB.
BETWEEN GROUPS	3	213.2500	71.0833	46.5575	.0000
WITHIN GROUPS	28	42.7500	1.5268		
TOTAL	31	256.0000			

The conclusion we can draw from this p-value is that we can statistically reject the null hypothesis that the population means for the four groups are the same. This particular test does not provide information about which means are different from which means. We need to develop a little more machinery before we can perform those tests.

(d) Relating the structural model and the variance decomposition approaches to ANOVA

Residual defined

Let's rewrite the structural model as

$$\begin{aligned} \epsilon_{ij} &= Y_{ij} - \mu - \alpha_i \\ &= Y_{ij} - (\mu + \alpha_i) \\ &= Y_{ij} - \mu_i \\ \text{residual} &= \text{observed score} - \text{expected score} \end{aligned}$$

That is, each subject has a new score ϵ_{ij} that represents the deviation between the subject's observed score and the score hypothesized by the structural model.

MSW means the same as MSE for now

The variance of the ϵ_{ij} is equivalent to MSW (as long as you divide the sum of squares by $df_{error} = N - T$). See Appendix 2 for an example using the sleep deprivation data. MSW is an unbiased estimate of the population error variance σ^2_ϵ. This holds whether or not the null hypothesis is true. In other words, MSW is one way to estimate the variance of the residuals; it uses deviations between the observed data points and their respective group mean. A synonymous term for MSW is mean square error, MSE.

If the null hypothesis is true (i.e., no treatment effects), then the MSB term provides a different way to estimate the population variance of the residuals σ^2_ϵ.

However, if the null hypothesis is not true, then MSB estimates

$$\sigma_\epsilon^2 + \frac{\sum n_i \alpha_i^2}{T-1} \tag{2-15}$$

The term on the right is essentially the variability of the treatment effects because it is primarily driven by the αs from the structural model. Obviously, if the treatment effects are zero, the term on the right vanishes. (This would occur when the null hypothesis is exactly true.) If the group means are far apart (in terms of location), then the group means have a relatively high variability, which will inflate the MSB.

The ratio MSB/MSW is a test of how much the treatment effect has inflated the error term.

A little side note about what MSB and MSW estimate and how two different concepts provide an estimate of the same entity. Let's consider a case where you are concerned about a gas leak in your apartment or house. You call a technician to bring a sensor to take measurements. The technician reports the readings in the house suggest that there is gas inside. But should we be alarmed? Should the reading be compared to something else? It may be that the ambient air contains said gas, so if you take a reading outside you can measure the gas level in the ambient air. The measurement inside the house has potentially two sources contributing to the gas level: one source is that there may be a leaky gas pipe or leaky appliance and a second source is merely the same ambient air that is outside. The reading in the house reports an aggregate value and we can't tell by looking at the aggregate value the relative contribution of those two sources. The technician could take additional readings outside the house to measure the gas level in the ambient air (which has one potential source) and compare them to the reading in the house (which has two potential sources). The analogy to the source table terms is that MSW only has one source driving it's estimate of the variance of the residuals, but MSB has two sources, one from potential nonzero treatment effects (the αs) and a second source from the residuals. The way to see how MSB is influenced by the variance of the residuals is that even if there were no treatment effects the group means will not all be exactly the same in a sample. Just by sampling alone (say 8 subjects into 4 groups like in the sleep deprivation example) we will observe fluctuations in the group means. That fluctuation is driven by the variance of the residuals even if all the treatment effects α were all zero.

F test statistic defined

Putting all this together in the ANOVA application, the *F* test statistic is simply a ratio of two expected variances

$$F \sim \frac{\sigma_\epsilon^2 + \frac{\sum n_i \alpha_i^2}{T-1}}{\sigma_\epsilon^2} \tag{2-16}$$

In this context, the symbol $\sim$ means "distributed as". So, Expression 2-16 means that the ratio of variances on the right hand side is "distributed as" an F distribution. The

basic structure of how an F test is formed is, in words,

$$\text{F} \sim \frac{\text{(terms you don't care about)} + \text{(terms you do care about)}}{\text{(terms you don't care about)}} \quad (2\text{-}17)$$

In this example we don't care about the variance of the residuals but do care about the variability of the treatment effect αs. We'll talk more about this structure when we get to expected mean square terms. To connect back to the gas leak scenario in the previous paragraph, the terms you don't care would be analogous to the gas levels in the ambient air and terms you do care about would be analogous to the gas levels due to potential leaks in the house.

The F distribution has two types of degrees of freedom, one corresponding to the numerator term and another corresponding to the denominator term. I will give some examples of the F distribution in class. The F table appears in Maxwell and Delaney starting at page A-3. Here is an excerpt from those pages. Suppose we have an application with 3 degrees of freedom in the numerator, 28 degrees of freedom in the denominator, and wanted the Type I error cutoff corresponding to $\alpha = .05$. In this case, the tabled value for F for 3, 28 degrees of freedom for $\alpha = .05$ is 2.95. In the previous example the observed F was 46.56 (with 4 groups, so 3 & 28 degrees of freedom), which exceeds the tabled value of 2.95.

Excerpt from the F table ($\alpha = .05$)

df_{error}	$\cdots$	$df_{num} = 3$	$\cdots$	$df_{num} = \infty$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
28	$\cdots$	2.95	$\cdots$	1.65
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	$\cdots$	2.60	$\cdots$	1.00

Excel and *F* values

If you want to compute your own F tables, you can use a spreadsheet such as Microsoft's Excel. For instance, the excel function FINV gives the *F* value corresponding to a particular *p* value, numerator and denominator degrees of freedom. If you type "=FINV(.05,1,50)" (no quotes) in a cell of the spreadsheet, the number 4.0343 will appear, which is the *F* value corresponding to a *p* value of 0.05 with 1, 50 degrees of freedom. Check that you get 2.95 when you type =FINV(.05,3,28) in Excel. Newer versions of Excel (starting in 2010) have a new function called F.INV.RT for the right tail, so typing =F.INV.RT(.05,1,28) is the same as the old version FINV(.05,1,28). There is also an F.INV (a period between the F and the I) where one enters the cumulative probability rather than the right tail, so the command =F.INV(.95,1,28) is the same as =F.INV.RT(.05,1,28).

Figure 2-3: Hypothesis testing template for the F test in the oneway ANOVA

Null Hypothesis

- H_o: $\mu_i = \mu$
- H_a: $\mu_i \neq \mu$ (two-sided test)

Structural Model and Test Statistic

The structural model is that the dependent variable Y consists of a grand population mean μ, a treatment effect α, and random noise ϵ_{ij}. In symbols, for each subject j in condition i his or her individual observation Y_{ij} is modeled as $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$.

The test statistic is a ratio of the mean square between groups (MSB) and the mean square within groups (MSW) from the ANOVA source table

$$F_{\text{observed}} = \frac{\text{MSB}}{\text{MSW}}$$

Critical Test Value The critical test value will be the table lookup of the F distribution. We typically use $\alpha = 0.05$. The F distribution has two degrees of freedom. The numerator df is number of groups minus 1 (i.e., T-1) and the denominator df is total number of subjects minus number of groups (i.e., N-T).

Statistical decision If the observed F exceeds the critical value F_{critical}, then we reject the null hypothesis. If the observed F value does not exceed the critical value F_{critical}, then we fail to reject the null hypothesis.

R and *F* values

The command in R for generating F values is qf(). For example, if you enter qf(.95,3,28) corresponding to two-tailed α = .05, 3 degrees of freedom for the numerator, and 28 degrees of freedom for the denominator. R will return 2.95, just as the Excel example in the previous paragraph. Note that the qf command is defined by the cumulative probability, so we enter .95 rather than .05.

(e) Putting things into the hypothesis testing template

In Figure 2-3 I make use of the hypothesis testing template introduced in LN1. This organizes many of the concepts introduced so far in these lecture notes. We state the null hypothesis, the structural model, the computation for observed F, the critical F, and the statistical decision.

(f) Interpreting the F test from a One Way ANOVA

The F test yields an *omnibus* test. It signals that somewhere there is a difference between the means, but it does not tell you where. In research we usually want to be more specific than merely saying "the means are different." In the next set of lecture notes (LN3) we will discuss planned contrasts as a way to make more specific statements such as "the difference between the means of group A and B is significant, but there is insufficient evidence suggesting the mean of Group C differs from the means of both Group A and B." As we will see, planned contrasts also provide yet another way to think about ANOVA.

2. Power and sample size

power

Determining power and sample size needed to attain a given level of power is a nontrivial task requiring knowledge of noncentral distributions. Most people simply resort to rules of thumb. For most effect sizes in psychology 30-40 subjects per cell will do but much depends on the amount of noise in your data. You may need even larger cell sizes. Using such a heuristic commits us to a particular effect size that we are willing to deem statistically significant. In some areas where there is relatively little noise (or many trials per subject) fewer observations per cell is fine.

Different statistical tests have different computational rules for power. In general, if you are in doubt as to how to compute power (or needed sample size), a good rule of thumb is to consult an expert. A very preliminary discussion is given in Maxwell & Delaney, pages 120-. I'll discuss power in more detail in the context of planned contrasts in Lecture Notes 3. It doesn't make much sense to talk about power of an omnibus test if we are not advocating its use.

Power analyses can be misleading because they say something not just about the treatment effect but also about the specific context, experiment, subject pool, etc. Seems that psychological researchers would be better off worrying about reducing noise in their experiments and improving their measurement issues, rather than concerning themselves exclusively with sample sizes. Often, better designed studies and cleaner measurement have a more dramatic effect on statistical power than modest increases in sample size.

This discussion is moving in the direction of philosophy of science. If you are interested in learning about this viewpoint, see an article by Paul Meehl (1967, Philosophy of Science, 34, 103-115) where he contrasts data analysis as done by psychologists with data analysis as done in the physical sciences such as in physics.

3. Strength of Association

One question that can be asked is what percentage of the total variance is "accounted for" by

R^2

having knowledge of the treatment effects (i.e., group membership). This percentage is given by

$$R^2_{\text{Y.treatment}} = \frac{\text{SSB}}{\text{SST}} \tag{2-18}$$

This ratio is interpreted as the percentage of the variability in the total sample that can be accounted for by the variability in the group means. For the sleep deprivation data,

$$R^2_{\text{Y.treatment}} = \frac{213.25}{256.00}, \tag{2-19}$$

or 83% of the variability in the scores is accounted for by the treatments and the rest is error. In other words, R^2 denotes the percentage of total sum of squares corresponding to SSB. You can refer to the pie chart of sum of squares for this.

The phrase "percentage of variance accounted for" is always relative to a base model (or "reduced model"). In ANOVA, the base model is the simple model with only the grand mean other than the error term. Thus, R^2 in ANOVA refers to how much better the fit is by using individual group means (i.e., SSB) as opposed to the grand mean (i.e., SST). It does not mean that a theory predicts X amount of a phenomenon, all the phrase denotes is that a simple model of group means does better than a model of the grand mean. Most people don't know this when they use the phrase "percentage of variance accounted for".

A cautionary note about interpreting $R^2_{\text{Y.treatment}}$. This percentage estimate refers only to the sample for which it was based, the particular experimental procedures, and the particular experimental context. Generalizations to other domains, treatments, and subjects must be made with care. We will deal with the issue of interpreting R^2 in more detail when discussing regression.

Several people have suggested a different measure, usually denoted ω^2 (omega squared), which adjusts $R^2_{\text{Y.treatment}}$ to yield a better estimate of the true population value. We will not review such measures in this class (I personally don't think they are useful); they are discussed in Maxwell and Delaney if you want to learn about the details.

There are other measures of effect size that appear in the literature, such as Cohen's d. These measures are all related to each other. They are just on different metrics and it is possible to convert one into another with simple formulas. Throughout these lecture notes I'll use measures based on r to keep things simple. If a journal requires you to use a different measure of effect size, then it is a simple matter of converting r to that measure.

4. Computational formulae for the one way ANOVA.

Some people like to give computational formulae that are easier to work with than definitional formulae. In this course I will focus on the definitional formulae because they give insight

into the techniques. One can still do the computations by hand using the definitional formulae. In this class the formula will facilitate understanding, the statistical computer program will perform the computations.

FYI: Hand computation of a one-way ANOVA using definitional formulae is fairly easy. First, compute the sum of squared deviations of each score from the grand mean. This yields SST. Second, compute the sum of squared deviations of scores from their group mean. This yields SSW. Third, figure out the degrees of freedom. Now you have all the pieces to complete the ANOVA source table. Be sure you understand this.

5. SPSS Syntax

   The simple one-way ANOVA can be performed in SPSS using the ONEWAY command.

```
ONEWAY dv BY group.
```

   More examples are given in the appendices.

6. R Syntax

   The R command aov() performs basic ANOVA. For example, a one-way ANOVA is called by

```
out <- aov(dv ~ group)
summary(out)
```

   where dv is the dependent variable and group is the grouping variable. The grouping variable should be defined as a factor such as

```
group <- factor(group)
```

   This attaches additional structure to the grouping variable that we will discuss later. Failure to define a factor will lead to incorrect results, particularly when there are three or more levels of the grouping factor.

7. What to report in a paper?

Turn out, not much of what we presented so far. The results of the F test itself are generally not useful. But the source table, which provides the decomposition of variance and the observed F, provides useful ingredients to a few more analysis. It is those additional analyses that we will typically be reporting in our papers.

8. Examining assumptions revisited: boxplots and normal-normal plots

The omnibus F test makes the same three assumptions we covered in LN1: independence, normality and equality of variances. I'll show how to use the boxplots for more than two groups and add some new plots to the mix. The spread-and-level plot will guide the selection of transformations and normal probability plots will assist in determining violations of normality.

(a) Boxplots defined

boxplot

a plot of the median, 1st & 3rd quartiles, and "whiskers." (see LN1)

(b) Getting a feel for sampling variability—simulations with boxplots

The lesson here is that sample size matters. If cell sizes are large (say 100 in each cell) and the population variances are equal, then the boxplots look pretty much the same across groups and minor deviations would not be tolerated (i.e., would lead to a rejection of the equal variance assumption). However, with small sample sizes in each cell (say 10), then one should be more tolerant of moderate discrepancies because they naturally occur even when sampling from a distribution where all groups have the same population variance.

(c) Examining the equality of variance assumption

i. Examples of unequal variances.

I generated three samples of size 20 using these parameters:

A. normal with mean 10 and sd 4

B. normal with mean 40 and sd 16

C. normal with mean 300 and sd 64

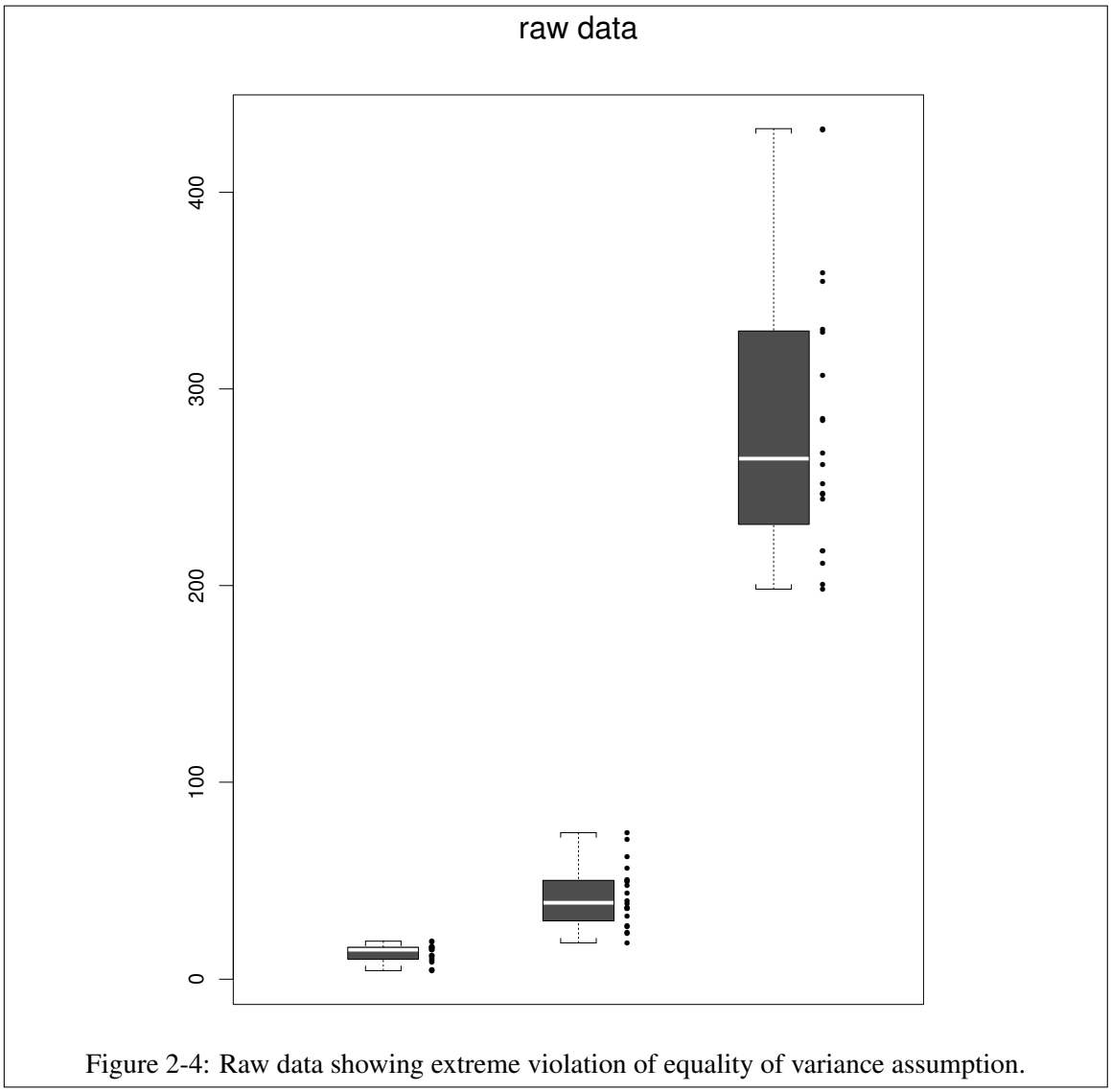


Figure 2-4: Raw data showing extreme violation of equality of variance assumption.

These data are shown in boxplots. Figure 2-4 shows a clear violation of the assumption of equal variances. Figure 2-5 shows how a log transformation of the data improves the status of the assumption.

ii. The family of power transformations.

power transformations

Consider the family of transformations of the form X^p. This yields many simple transformations such as the square root (p = .5), the reciprocal (p = -1), and the log (when p = 0, see footnote[1]).

[1]Only for those who care.... That p = 0 yields the log transformation is not easy to prove unless you know some

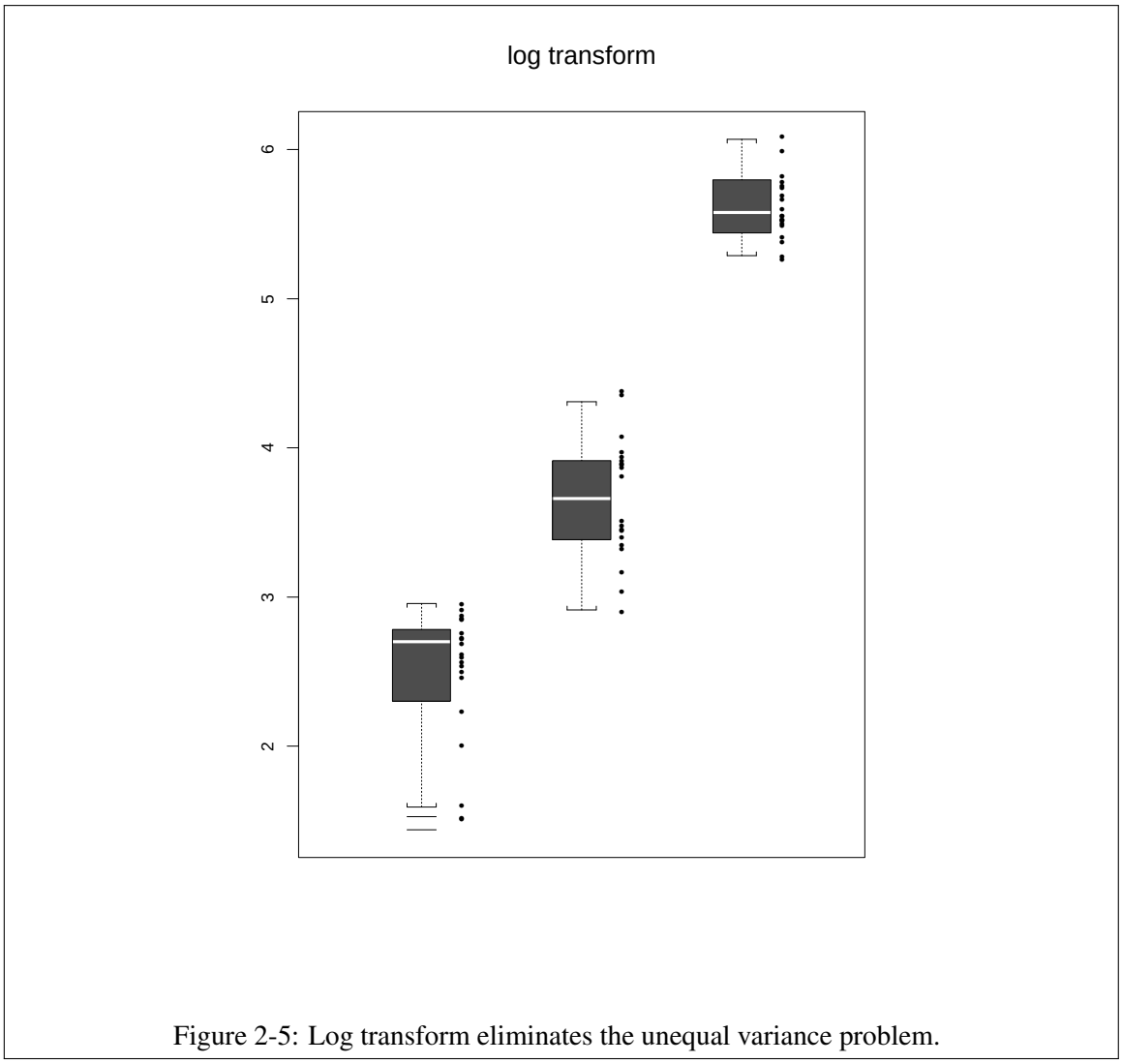


Figure 2-5: Log transform eliminates the unequal variance problem.

Figure 2-6 shows the "ladder of power transformations". Note how varying the exponent p produces different curves, and these curves are continuous variations of the continuously varying exponent p.

The SPSS syntax to compute these power transformations (i.e,. x^p) is as follows: for p not equal to 0,

```
COMPUTE newx = x**p.
EXECUTE.
```

for p equal to 0,

```
COMPUTE newx = ln(x).
EXECUTE.
```

Transformations in R are computed more directly as in

```
newx <- x^2
newxln <- ln(x)
```

iii. Spread and level plots for selecting the exponent

spread and level plot

This plot exploits the relationship between measures of central tendency and measures of variability. By looking at, say, the relationship between the median and the interquartile ranges of each group we can "estimate" a suitable transformation in the family of the power functions. The spread and level that SPSS generates plots the log of the medians and the log of the interquartile ranges.

In SPSS, the plot is generated by giving the */plot=spreadlevel* subcommand to the *EXAMINE* command. That is,

```
EXAMINE variable = listofvars BY treatment
 /plot spreadlevel.
```

calculus. Here's the sketch of the proof. Consider this linear transformation of the power function

$$\frac{X^p - 1}{p} \tag{2-20}$$

We want to prove that the limit of this function as p approaches 0 is log(X). Note that this limit is indeterminate, so apply L'Hopital's rule and solve for the limit. When p is negative, the scores X^p are reversed in direction; thus, a side benefit Equation 2-20 is that it yields transformed data in the correct order.

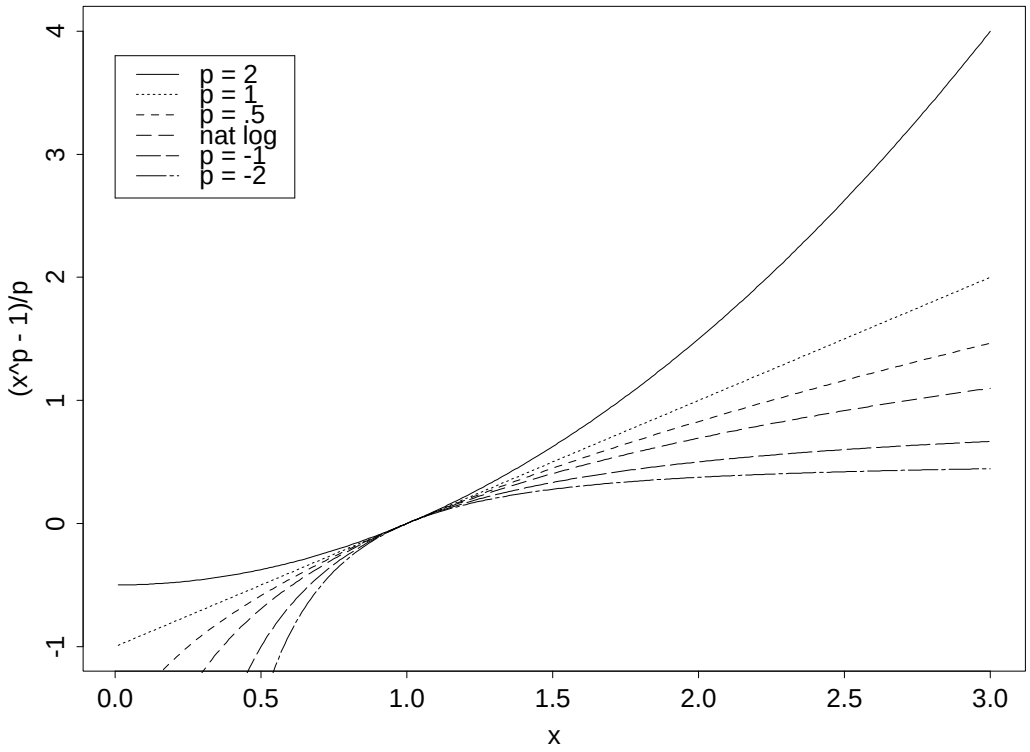


Figure 2-6: Ladder of power transformations.

The program finds the line of best fit (i.e., the regression line) through the points in that plot. If you limit yourself to power transformations, then the slope of the regression line is related to the power that will "best" transform the data (more precisely, power $= 1 -$ slope). For example, if you find that the slope of the regression line is 0.5, then that suggests a square root transformation will probably do the trick (i.e., $1 - 0.5 = 0.5$). If you run the spread and level plot in SPSS on the sleep deprivation example it will suggest a transformation of -.095, or roughly a log transformation. More examples in the appendix.

The spread and level plot works well when the means and variances have a simple pattern (e.g,. the variances increase linearly with means). If you the data don't exhibit a simple pattern, then the spread and level plot will likely suggest strange transformations (e.g., p = -8.54). In cases where the patterns are not simple, then the spread and level plot should not be trusted.

In R the spread and level plot is available through the car package

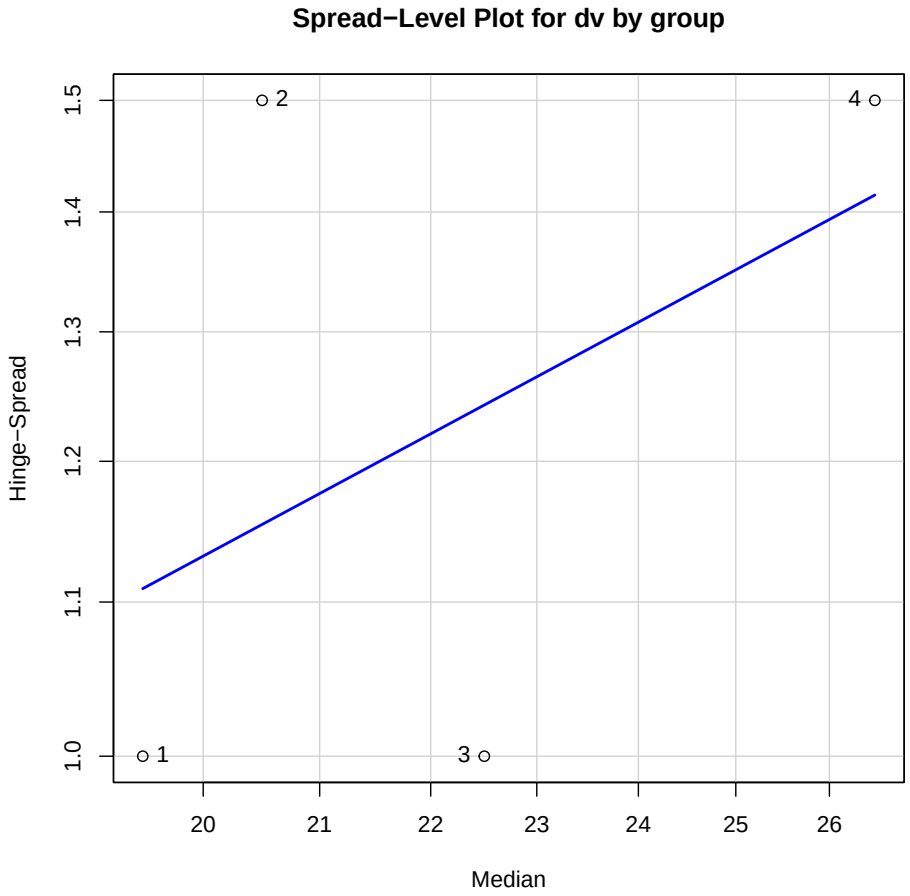
```
library(car)
spreadLevelPlot(dv ~ group)
```

This function complains if there are negative values or zeros in the dependent variable, and will add a constant to eliminate them. Also, the spreadLevelPlot command in the car package has some idiosyncratic definitions, for example, using a robust estimator for the regression to compute slope and using an old approach for computing the quantiles (see the help for the R quantile function for a list of the various approaches). These differences can sometimes produce suggested transformations that given by other programs. You can avoid the robust regression line estimate by specifying the argument as shown below but changing the approach to compute the quantiles is not currently possible within the car package's implementation of the spread and level plot.

```
spreadLevelPlot(dv ~ group, robust.line=F)
```

For the sleep deprivation data we have

```
library(car)
#reread data because earlier I deleted two groups
data.sleep <- read.table("data.sleep",
                col.names=c("dv","group"))
data.sleep[,"group"] <- factor(data.sleep[,"group"])
spreadLevelPlot(dv ~ group, data=data.sleep)
```




```
##   LowerHinge Median UpperHinge Hinge-Spread
## 1       19.0   19.5       20.0          1.0
## 2       20.0   20.5       21.5          1.5
## 3       22.0   22.5       23.0          1.0
## 4       25.5   26.5       27.0          1.5
##
## Suggested power transformation:  0.206773
```

If you round down, the R command suggests a log transform just like SPSS. The number is a little different than SPSS because this R command uses a different way to compute the IQR (that is, it uses the hinge method). You'll see that the IQRs computed by R are 1, 1.5, 1, and 1.5, but SPSS reports 1, 1.75, 1, 1.75 (run EXAMINE command to find out). This is due to different ways of interpolating when computing the IQR; see the help file for the quantile command in R for details. If you use type=6 as in with(data.sleep,by(dv,group,IQR,type=6)) you reproduce the SPSS IQRs, and then can reproduce the SPSS spread and level plot if you do the appropriate regression on the logs. This is just a tiny point, and usually doesn't

make much difference in the suggested transformation.

The SPSS version of the spread and level plot (which uses some different computational approaches such as not using a robust regression approach and using a different approach to compute quantiles) reports a slope of 1.09 so a suggested power for transformation of -.09. While not close to the car package's suggested power of .21 from the spreadLevelPlot command, both are estimates are rounded to same nearest rung on the ladder of power transformations leading to a suggested log transformation.

See Box and Cox (1964, *Journal of the American Statistical Society, Series B, 26*, 211-243) for a derivation of why the slope of a spread and level plot tells you something about which transformation to use. They make their arguments from both maximum likelihood and Bayesian perspectives. A classic article in statistics! Note that the Box & Cox derivation is for a spread and level plot that uses log means and log standard deviations. The version that SPSS uses (log medians and log interquartile ranges) comes from arguments made by Tukey (1977) about the robustness of medians and interquartile ranges.

These procedures are merely guidelines—you should double-check the result of the transformation to make sure the transformation had the desired effect. In my opinion nothing beats a little common sense and boxplots—usually within a couple of tries you have a transformation that works at minimizing the violation to the equality of variances assumption.

(d) more examples given in class

Bank data, Norusis, SPSS Base manual (Appendix 3)

Helicopter data, Kutner et al (Appendix 4)

Population in cities from 16 different countries—from the boxplot chapter in Hoaglin, Mosteller & Tukey's Understanding Robust and Exploratory Data Analysis (UREDA) (Appendix 5)

Sometimes the spread and level plot doesn't produce a good result (as seen in the helicopter data). The spread & level plot is merely a tool. When it yields strange results, the plot should be ignored. Recall that complicated patterns between means and variances cannot usually be corrected with a simple power transformation. You should always check that the transformation produced the desired result.

9. Checking the normality assumption

histogram

One way to check for normality is to examine the histogram. This is not necessarily the best thing to do because there is a lot of freedom in how one constructs a histogram (i.e., the sizes of the intervals that define each "bin").

boxplot

Another device is the trusty boxplot, which will tell you whether the distribution is symmetric (a necessary condition of normality). A suggestion of symmetry occurs if the median falls in the middle of the "box" and the two whiskers are similar in length.

normal probability plot

A more sophisticated way to check normality is through a quantile-quantile plot. This is what SPSS calls the normal plot. The logic of this plot is quite elegant. Take your sample and for each point calculate the percentile. For example, one data point might be at the 50th percentile (i.e., the median). Then find the z scores from a normal distribution ($\mu = 0, \sigma = 1$) that correspond to the each percentile. For example, if one data point had a percentile score of 95, then it corresponds to a z score of 1.645. Finally, plot the raw data points (y-axis) against their corresponding z scores (x-axis). If the points fall on a straight line, then the distribution is consistent with a normal distribution.

The way to calculate the percentile for purposes of this plot is to order your data from least to greatest so that the first data point is the least, ..., the nth data point is the greatest. The percentile for the data point in the ith position is given by

$$\frac{\mathrm{i} - 0.5}{\mathrm{n}} \tag{2-21}$$

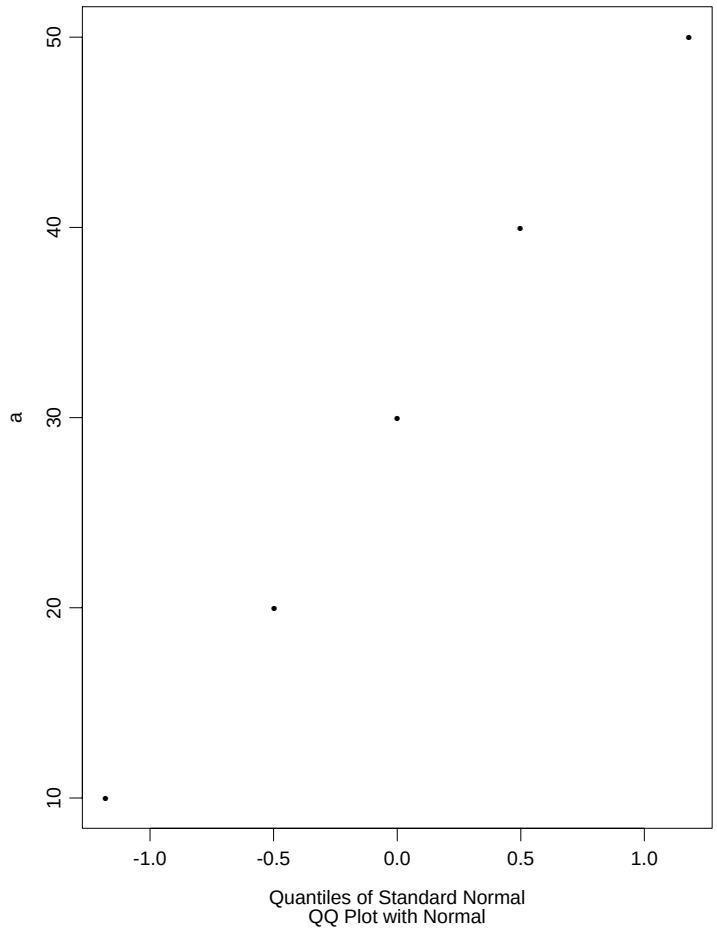
An example will illustrate. Suppose you observe these five data points: 30, 20, 40, 10, and 50. Create the following table:

raw data	ordered	percentile	z score
10	1	0.1	-1.28
20	2	0.3	-0.52
30	3	0.5	0.00
40	4	0.7	0.52
50	5	0.9	1.28

The normal plot for these data appear in Figure 1. These five points fall on a straight line so a normal distribution cannot be ruled out.

The normal plot is useful in many situations. We will use it frequently when we look at residuals in the context of regression.

Figure 2-7: Illustration of QQ plot on five data points



The quantile-quantile plot can be generalized to any theoretical or empirical distribution. For example, you can check whether your distribution matches a χ^2 with df=3. Or, you can plot one observed distribution against another observed distribution to see whether the two distributions are similar. Unfortunately, the canned quantile-quantile plot that appears in SPSS only allows one distribution and compares it only to the normal distribution. The technique is more general than the specific SPSS implementation.

SPSS has the quantile-quantile plot (aka normal plot) in several of its commands. Today we'll discuss the subcommand in **EXAMINE**. Just enter the following:

```
EXAMINE  variables = listofvariables
     /PLOT NPPLOT.
```

Appendix 6 shows an example.

To get separate normal plots for each cell enter:

```
EXAMINE variables = listofvariables BY treatment
     /PLOT NPPLOT.
```

It is okay to put all three plots (boxplot, spreadlevel, and normal probability plot) on one SPSS line as in:

```
EXAMINE variables = listofvariables BY treatment
     /PLOT BOXPLOT SPREADLEVEL NPPLOT.
```

Assumptions should usually be checked cell by cell (rather than for all cells combined).

SPSS also prints out the "detrended" plot. This is simply a plot of the residuals.

$$\begin{aligned} \text{residual} &= \text{observed} - \text{expected} \\ &= \text{raw score} - \text{linear model} \end{aligned}$$

If the straight line fits the data, then the detrended plot will look like a horizontal band (see the last appendix for an example). The intuition is that if the linear model captured all or most of the systematic variance, then just noise is left over. Noise is just as likely to be negative as positive (and it should not be related, under the assumption of independence, to the level

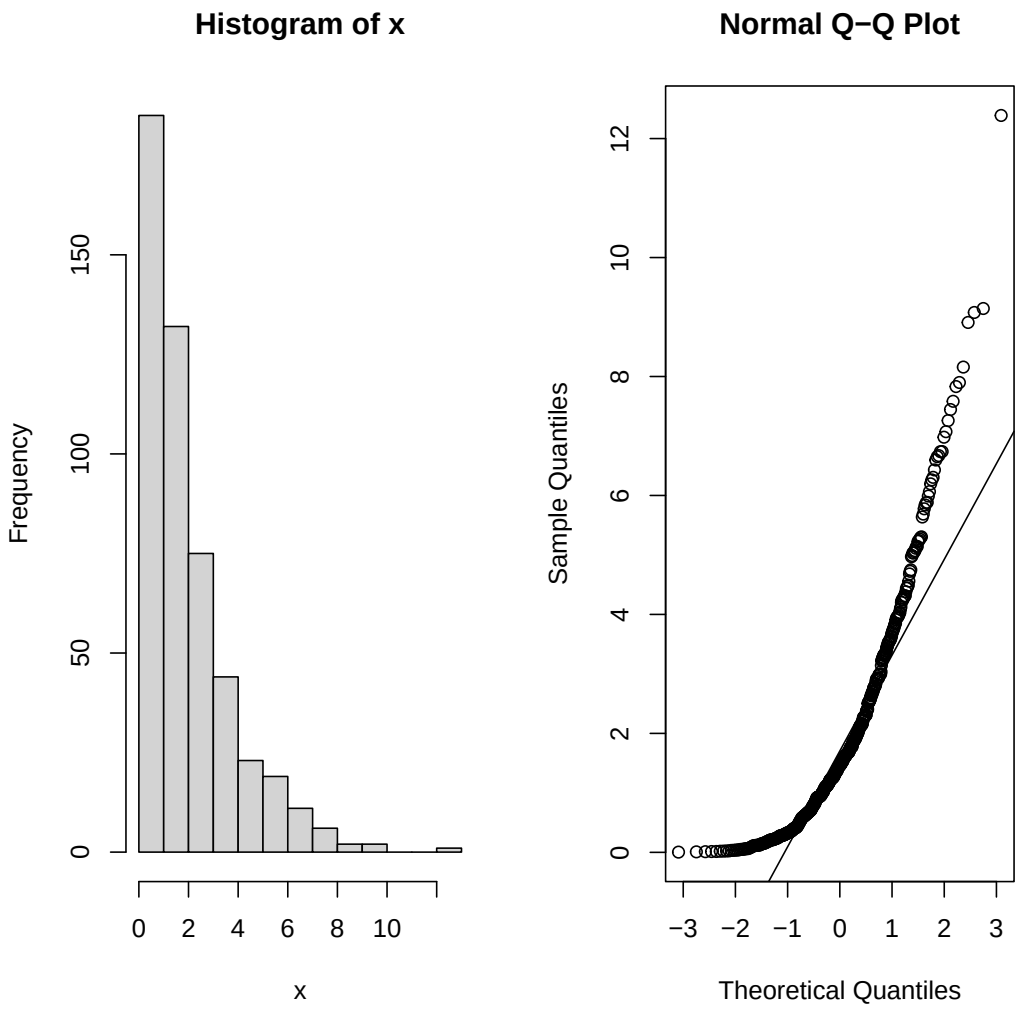
of the score). Some people like these detrended plots because it is easier to detect visually a discrepancy from a horizontal line than to detect discrepancy from a line with a non-zero slope.

There is also a command in SPSS called PPLOT, which allows you to compare your data to different distributions not just the normal. To get this plot for each of your groups you need to split your data into groups and get SPSS to run the command separately by group.

R quantile plot

In R the quantile-quantile plot is produced by the qqnorm() and qqline() commands. The former sets up the plot, the latter prints the line.

```
# generate some non-normal data for illustration
x <- rchisq(500, 2)
par(mfcol = c(1, 2))
hist(x)
qqnorm(x)
qqline(x)
```

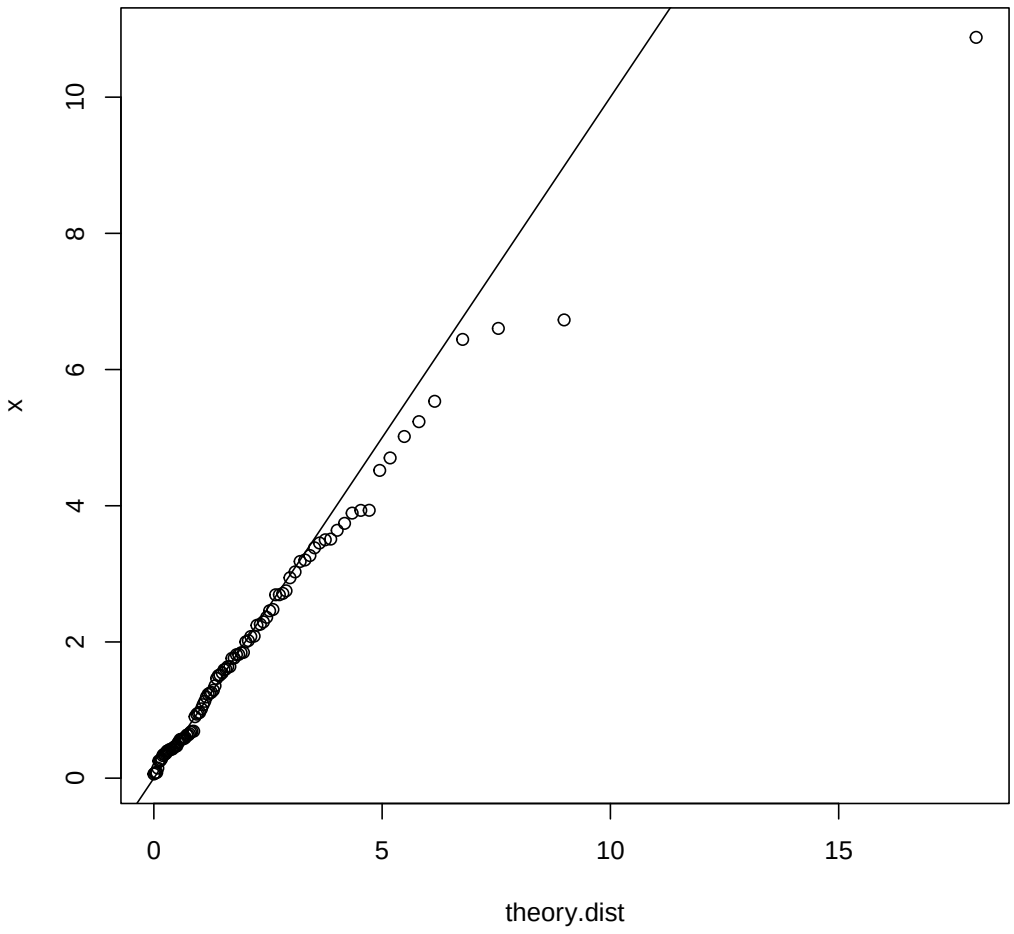



```
par(mfcol = c(1, 1))
```

As the name suggests, qqnorm and qqline compare the data distribution against the normal distribution. There is a more general function called qqplot, which compares data to an arbitrary distribution of your choosing. For example,

```
# some data but let's pretend we don't know the
# distribution we hypothesize a theoretical
# distribution of chisq with 2 df so set up a
# theoretical dataset of 10,000 compare data to the
# theoretical dist
x <- rchisq(100, 2)
theory.dist <- rchisq(10000, 2)
```

```
qqplot(theory.dist, x)
# draw line to guide the eye
abline(0, 1)
```



Unfortunately, the qq family of functions in R is somewhat limited to only do one group at a time. If you have many groups you can do a for loop or use a combination of with and by commands. With says which data set to use to find variable names, by repeats the command for each level of group. Since there are two functions to produce the plot (one for the normal qq and another to draw the line), the by command defines an internal function to run both qqnorm and qqline separately on each group. One could do the brute force method of saving each group into a separate data frame and then doing the qqnorm plot separately for each group, but the with/by combination is more efficient and easier to read.

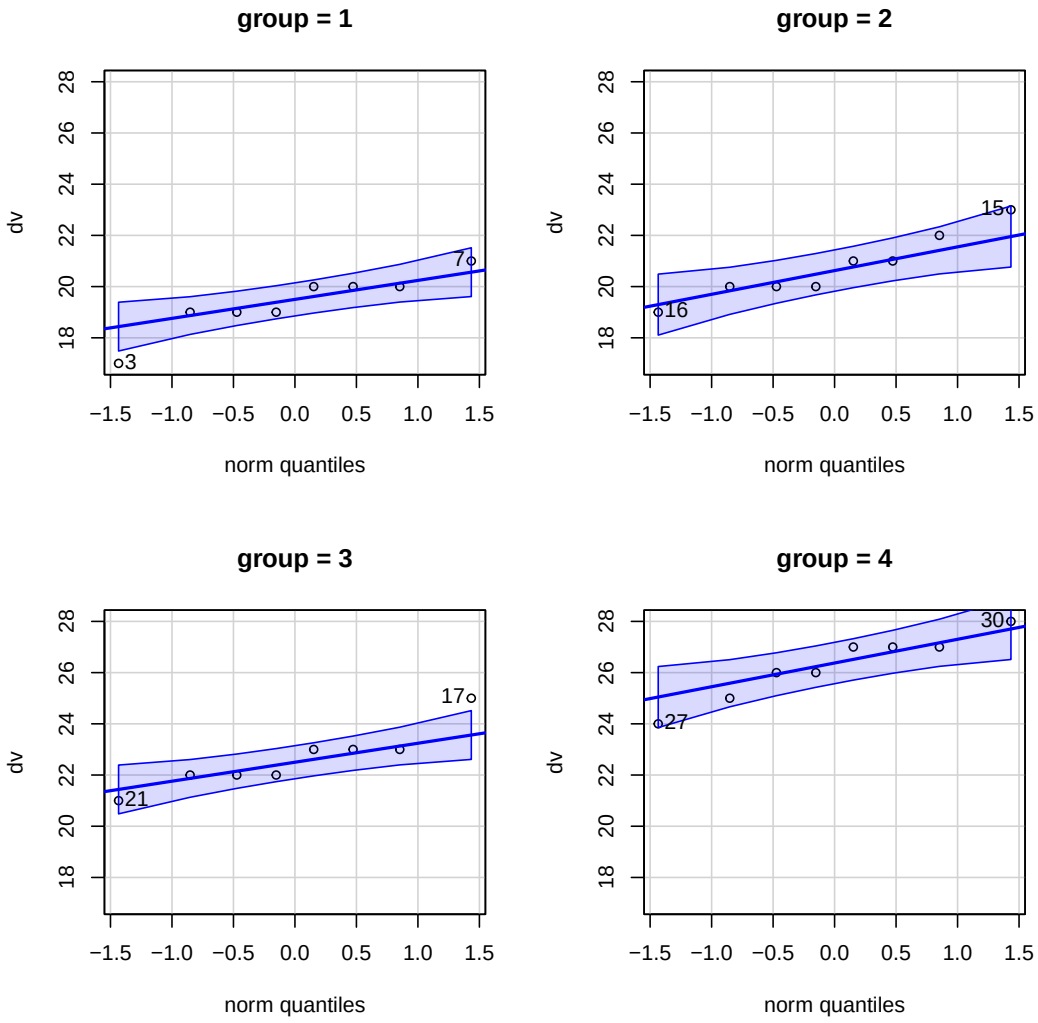
```
with(data, by(dv, group, function(i) {
    qqnorm(i)
    qqline(i)
}))
```

and, to put the four plots on a 2x2 grid, one could type this command prior to the line above in order to force plots to appear on a grid

```
par(mfcol = c(2, 2))
```

Or if you don't want to deal with this with/by combination or a for loop, the car package has a convenient qqPlot function (capital P) that accepts the formula syntax more commonly used throughout R. The general syntax is (see Appendix 1 for another example):

```
qqPlot(dv ~ group, data = data.sleep)
```



transformations may backfire

Transformations can also be used to remedy violations of normality. But be careful because a transformation used to remedy one violation (e.g., to fix normality) may backfire and create problems for another assumption (e.g,. equal variances).

10. Getting a feel for sampling variability–simulations involving normal plots.

The lesson here is that with large samples (say 250+) you can detect violations of normality and you should not be very tolerant of minor deviations from the straight line. However, with relatively small sample sizes (say 20) it is difficult to distinguish data that came from a normal distribution from data that came from a mildly skewed distribution.

Here is the rub: we know from the central limit theorem that as sample size increases, the

sampling distribution of the mean approaches a normal distribution. Thus, the cases where we can be confident that we have found a violation of normality (a normal probability plot with large N), are the very cases where violations of normality probably don't matter much in terms of the effects on the t or F tests. The cases where we can't be so confident that we have found a violation of normality (small N) are the very cases where the t or F test can break down if we violate normality. Simulations suggest that one needs to violate normality in a very major way before seeing adverse effects on the results of the t and F tests. [Note that the equality of variance assumption can create havoc even for moderately large N.]

11. Robustness

Maxwell and Delaney make an excellent point about violations of assumptions. The real issue about violations is whether they have an adverse effect on the results of the inferential tests. Will the width of our confidence intervals be wrong? Will the p-value of the t or F test be wrong? For the case of normality we can appeal to the central limit theorem for large N. But what about small N? What about violations of the equality of variance assumption?

Here we turn to simulation studies where we purposely create violations and find ways of measuring how "wrong" the t, the F and the CI turn out to be. Maxwell and Delaney review this work and you can read their take on it.

My take is that people who review the simulation work see what they want to see. The results are mixed. Sometimes violations matter, sometimes they don't. Some people conclude that because there are cases where the ANOVA behaves well in the face of violations, we don't need to worry so much about violations; other people conclude that because there are cases where ANOVA doesn't behave well under violations, then we always need to worry about violations.

For me the issue is that for a specific dataset I don't know what situation it falls in (i.e., I don't know if that dataset will be in the "assumption violation probably doesn't matter" camp or the "assumption violation probably matters" camp). Because I don't know, then the best thing I can do is play it safe. Check the data to see if there are any gross discrepancies. If there are major violation assumptions, then I worry and try something remedial (a nonparametric test, a Welch test if the problem is equal variances, a transformation). I'm not so worried about close calls because they probably won't make much of a difference in my final results.

12. Explaining transformations in results sections

A transformation will add a sentence or two to your results section. You need to tell the reader that you checked the assumptions, that you had reason to believe violations were present, and that you believed it was necessary to transform the data.

It would be nice to report measures of central tendency (e.g., mean, median, etc.) and measures of variability (e.g., standard deviation, IQR) for both raw scores and transformed scores. But, if you have many measures, then you might want to trim down and only report the transformed scores. Hopefully, psychology journals will start publishing boxplots soon. Major journals such as *Science, Journal of the American Statistical Association, American Statistician*, and *New England Journal of Medicine* have started publishing papers with boxplots.

Some people will report the transformed means that have been "transformed back" into the original scale. For example, suppose you have the data 1, 2, 3, 4, and 5. You decide that a transformation is necessary and you use the square root. The mean on the square root scale is 1.676. The inferential test is computed on the transformed scores. Some people will transform the mean back to its original scale when reporting the cell means and cell standard deviations. In this example one would square the mean on the square root scale and report a mean of 2.81. So, they calculate their p values on the transformed score but report means that have been transformed back to the original scale. Note that the mean on the original scale is 3, which is different from the re-transformed mean of 2.81 on the square root scale.

You should not feel ashamed about having to report violations of assumptions. Actually, it reflects favorably on you because it shows how careful you are at data analysis. Remember that statistical assumptions are creations to make some derivations tractable (e.g., recall our discussion of the two-sample t test during the first Psych 613 lecture). If your data do not satisfy the assumptions of a statistical test, it does not mean you have inherently bad data. It simply means that one particular tool cannot be used because the conditions that make it possible to use that tool do not hold. The critical point is how you handle assumption violations. Do you do a transformation? a nonparametric test? a test like Welch's that relaxes some of the assumptions? All three of these options are not always available in all applications of inferential tests. Careful consideration of all your options and careful examination of statistical assumptions are the most important aspects of data analysis.

Pre-registration

These "best practices" on assumption checking make it difficult to write out a pre-registration plan. Some pre-registration forms, such as one on the Open Science Framework, have a series of questions for how you will check assumptions, what you will do if you suspect assumption violations, etc. Other pre-registration forms don't specify these details. If you believe in pre-registration, you should outline all your processing steps, including those related to assumption checking. Ideally, in the pre-registration write out all the code you will use and include comments of what you will examine in the output to decide on assumption violations. Also, state what you will do if there are assumptions violations (e.g., if I see evidence of outliers, I will conduct a nonparametric test and will report both the classic results and the nonparametric test in the report).

Appendix 1

Example of One Way ANOVA: SPSS and R

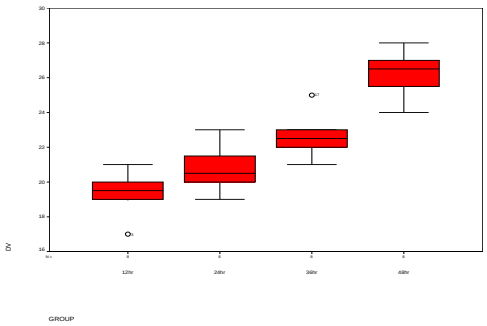
Basic ANOVA in SPSS

```
First, we examine the describe statistics and boxplot. Then we present
the ANOVA.
```

```
data list free /dv group.
begin data
20 1
20 1
17 1
19 1
20 1
19 1
21 1
19 1
21 2
20 2
21 2
22 2
20 2
20 2
23 2
19 2
25 3
23 3
22 3
23 3
21 3
22 3
22 3
23 3
26 4
27 4
24 4
27 4
25 4
28 4
26 4
27 4
end data.

value labels group 1 '12hr' 2 '24hr' 3 '36hr' 4 '48hr'.

    examine dv by group
     /plot = boxplot .
```

```
    DV
By  GROUP          1.00  12hr

Valid cases:          8.0   Missing cases:         .0   Percent missing:       .0


Mean        19.3750  Std Err      .4199  Min         17.0000  Skewness    -.9698
Median      19.5000  Variance    1.4107  Max         21.0000  S E Skew     .7521
5% Trim     19.4167  Std Dev     1.1877  Range        4.0000  Kurtosis    1.8718
                                         IQR          1.0000  S E Kurt    1.4809
```

```
    DV
By  GROUP          2.00  24hr

Valid cases:         8.0   Missing cases:        .0   Percent missing:      .0


Mean       20.7500  Std Err      .4532  Min        19.0000  Skewness     .6106
Median     20.5000  Variance    1.6429  Max        23.0000  S E Skew     .7521
5% Trim    20.7222  Std Dev     1.2817  Range       4.0000  Kurtosis    -.0212
                                        IQR         1.7500  S E Kurt    1.4809

    DV
By  GROUP          3.00  36hr

Valid cases:         8.0   Missing cases:        .0   Percent missing:      .0


Mean       22.6250  Std Err      .4199  Min        21.0000  Skewness     .9698
Median     22.5000  Variance    1.4107  Max        25.0000  S E Skew     .7521
5% Trim    22.5833  Std Dev     1.1877  Range       4.0000  Kurtosis    1.8718
                                        IQR         1.0000  S E Kurt    1.4809

    DV
By  GROUP          4.00  48hr

Valid cases:         8.0   Missing cases:        .0   Percent missing:      .0


Mean       26.2500  Std Err      .4532  Min        24.0000  Skewness    -.6106
Median     26.5000  Variance    1.6429  Max        28.0000  S E Skew     .7521
5% Trim    26.2778  Std Dev     1.2817  Range       4.0000  Kurtosis    -.0212
                                        IQR         1.7500  S E Kurt    1.4809
```



HERE IS THE ANOVA

```
 oneway dv by group
  /statistics = all .
      Variable  DV
   By Variable  GROUP

                    ANALYSIS OF VARIANCE
```

```
                                SUM OF          MEAN            F      F
        SOURCE           D.F.  SQUARES        SQUARES         RATIO  PROB.

BETWEEN GROUPS              3    213.2500       71.0833      46.5575  .0000

WITHIN GROUPS              28     42.7500        1.5268

TOTAL                      31    256.0000
```

```
                                    STANDARD   STANDARD
GROUP        COUNT        MEAN     DEVIATION     ERROR     MINIMUM     MAXIMUM     95 PCT CONF INT FOR MEAN

Grp 1            8     19.3750      1.1877       .4199     17.0000     21.0000      18.3820 TO     20.3680
Grp 2            8     20.7500      1.2817       .4532     19.0000     23.0000      19.6784 TO     21.8216
Grp 3            8     22.6250      1.1877       .4199     21.0000     25.0000      21.6320 TO     23.6180
Grp 4            8     26.2500      1.2817       .4532     24.0000     28.0000      25.1784 TO     27.3216

TOTAL           32     22.2500      2.8737       .5080     17.0000     28.0000      21.2139 TO     23.2861

          FIXED EFFECTS MODEL       1.2356       .2184                             21.8026  TO     22.6974

         RANDOM EFFECTS MODEL                   1.4904                             17.5069  TO     26.9931

RANDOM EFFECTS MODEL - ESTIMATE OF BETWEEN COMPONENT VARIANCE          8.6946
```

```
Tests for Homogeneity of Variances

      Cochrans C = Max. Variance/Sum(Variances) =  .2690, P = 1.000 (Approx.)
      Bartlett-Box F =                             .025 , P =  .995
      Maximum Variance / Minimum Variance         1.165
```

Some notes about this output. The CI's in the printout are based on the individual cell standard deviations, not the pooled standard deviation. The CI around the "TOTAL" mean is the CI around the grand mean using the standard deviation computed from the entire sample (i.e., N - 1). The 1.2356 in the fixed effect model is the sqrt of MSW (i.e., $\sqrt{1.5268}$), the standard error of that estimate is $1.2356/\sqrt{32}$ (the denominator contains the total sample size). Thus, the fixed effect is based on the pooled MSW. The CI given in the fixed effects column is the CI around the grand mean, using the square root of the pooled MSE term as the standard deviation estimate and having N - T degrees of freedom (rather than N - 1).

Basic ANOVA in R

```
data.sleep <- read.table("data.sleep", col.names = c("dv",
    "group"))
data.sleep[, "group"] <- factor(data.sleep[, "group"], levels = 1:4,
    labels = c("12hr", "24hr", "36hr", "48hr"))
# here is a print out of the data
data.sleep


##    dv group
```

```
## 1  20  12hr
## 2  20  12hr
## 3  17  12hr
## 4  19  12hr
## 5  20  12hr
## 6  19  12hr
## 7  21  12hr
## 8  19  12hr
## 9  21  24hr
## 10 20  24hr
## 11 21  24hr
## 12 22  24hr
## 13 20  24hr
## 14 20  24hr
## 15 23  24hr
## 16 19  24hr
## 17 25  36hr
## 18 23  36hr
## 19 22  36hr
## 20 23  36hr
## 21 21  36hr
## 22 22  36hr
## 23 22  36hr
## 24 23  36hr
## 25 26  48hr
## 26 27  48hr
## 27 24  48hr
## 28 27  48hr
## 29 25  48hr
## 30 28  48hr
## 31 26  48hr
## 32 27  48hr
```

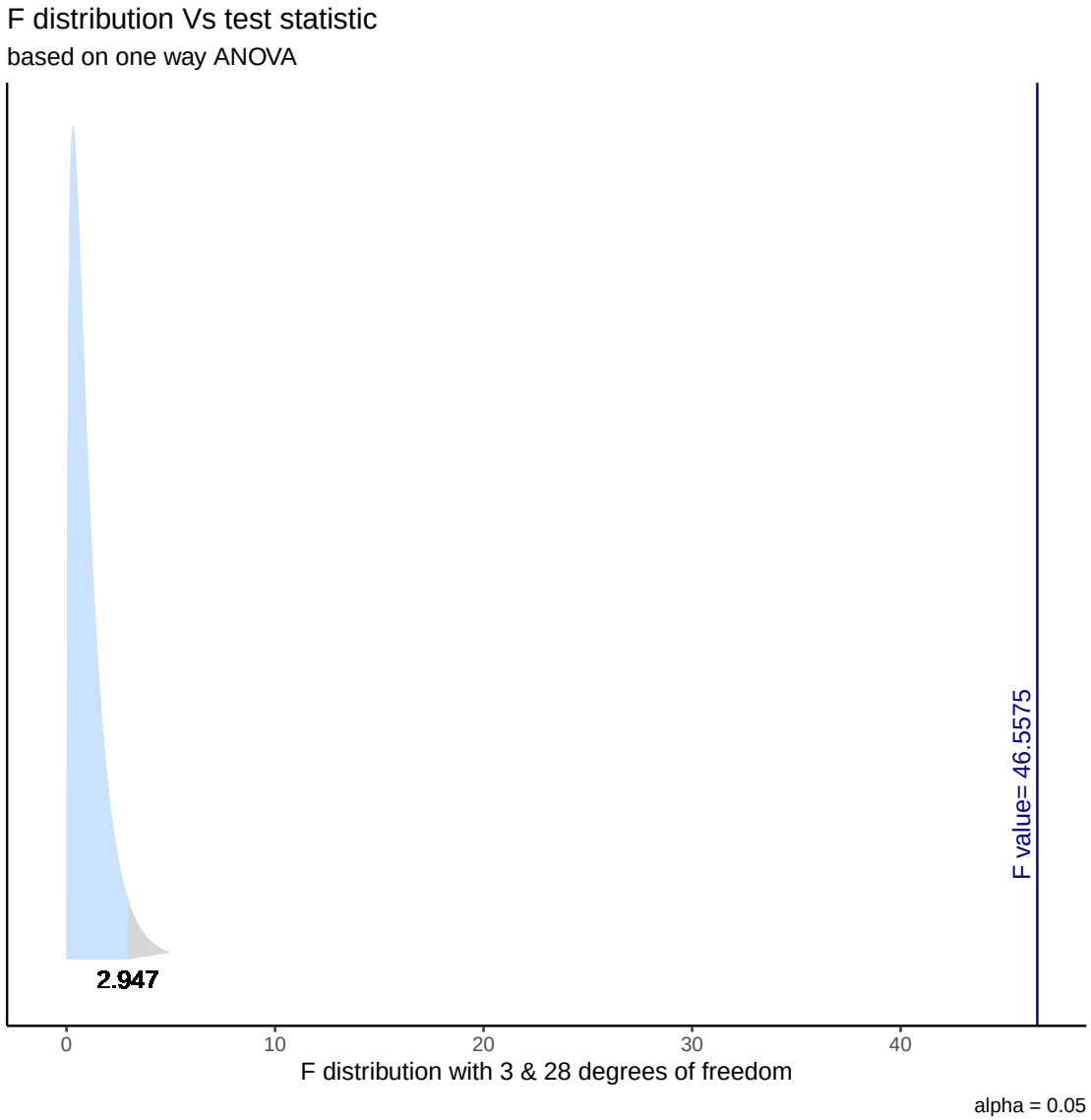
```
summary(aov(dv ~ group, data = data.sleep))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## group        3 213.25   71.08   46.56 5.22e-11 ***
## Residuals   28  42.75    1.53
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is an analogous theoretical plot to do what we saw in LN1 for the two sample t-test in the

gginference package. It plots the theoretical F distribution and overlays the F observed. In this example, the F observed of 46.6 is much greater than the critical value of 2.95 (or equivalently, that the p-value for this observed test is much smaller than the usual .05 criterion). The blue region represents 95% of the area to the left of the critical F and the gray region is 5% to the right of the critical F. The plot didn't draw the very thin band of gray that extends way out to 46.6 and beyond.

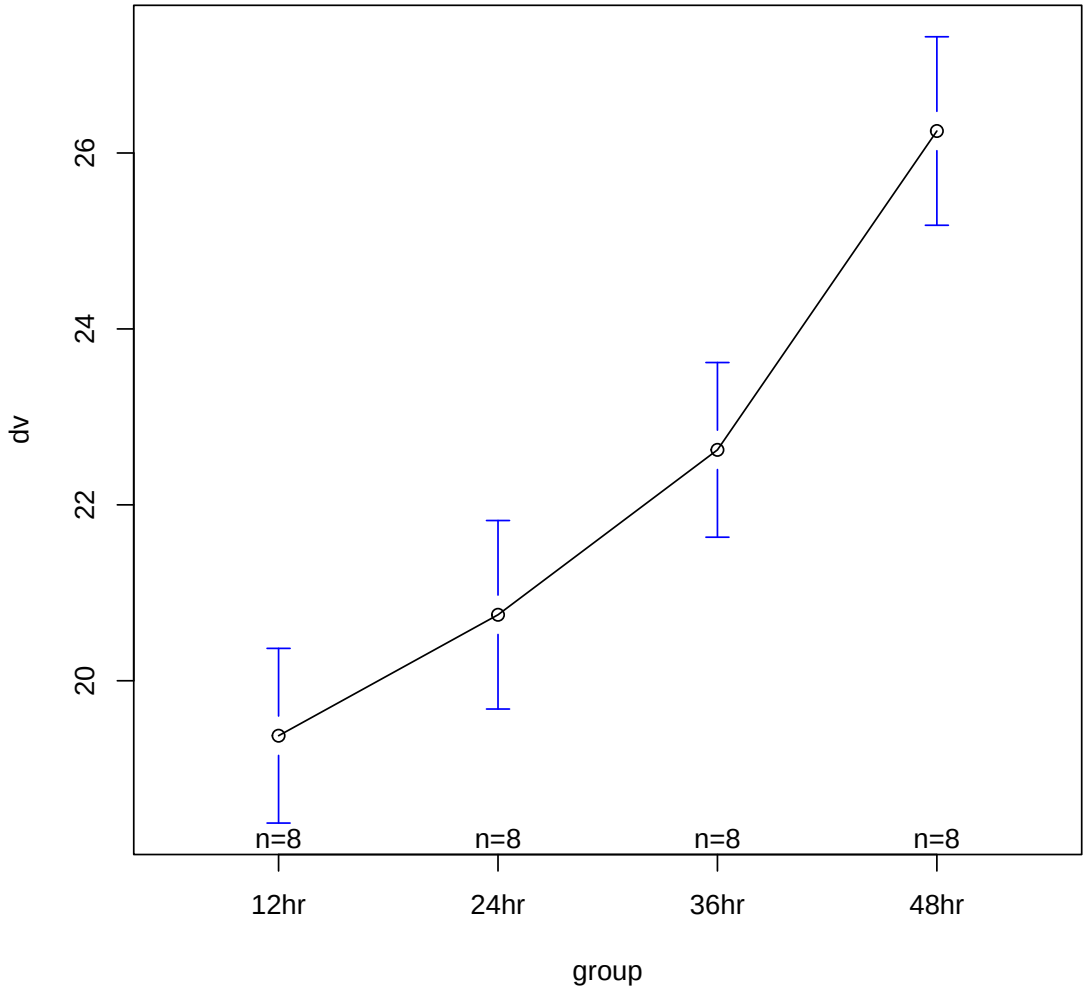
```
library(gginference)
ggaov(aov(dv ~ group, data = data.sleep))
```



There are several ways to plot confidence intervals around means in R. LN1 mentions several; ggplot is probably the best approach but that requires learning more about R. Here I'll illustrate the plotmeans() command in the package gplots. You can look at the help file for plotmeans() for

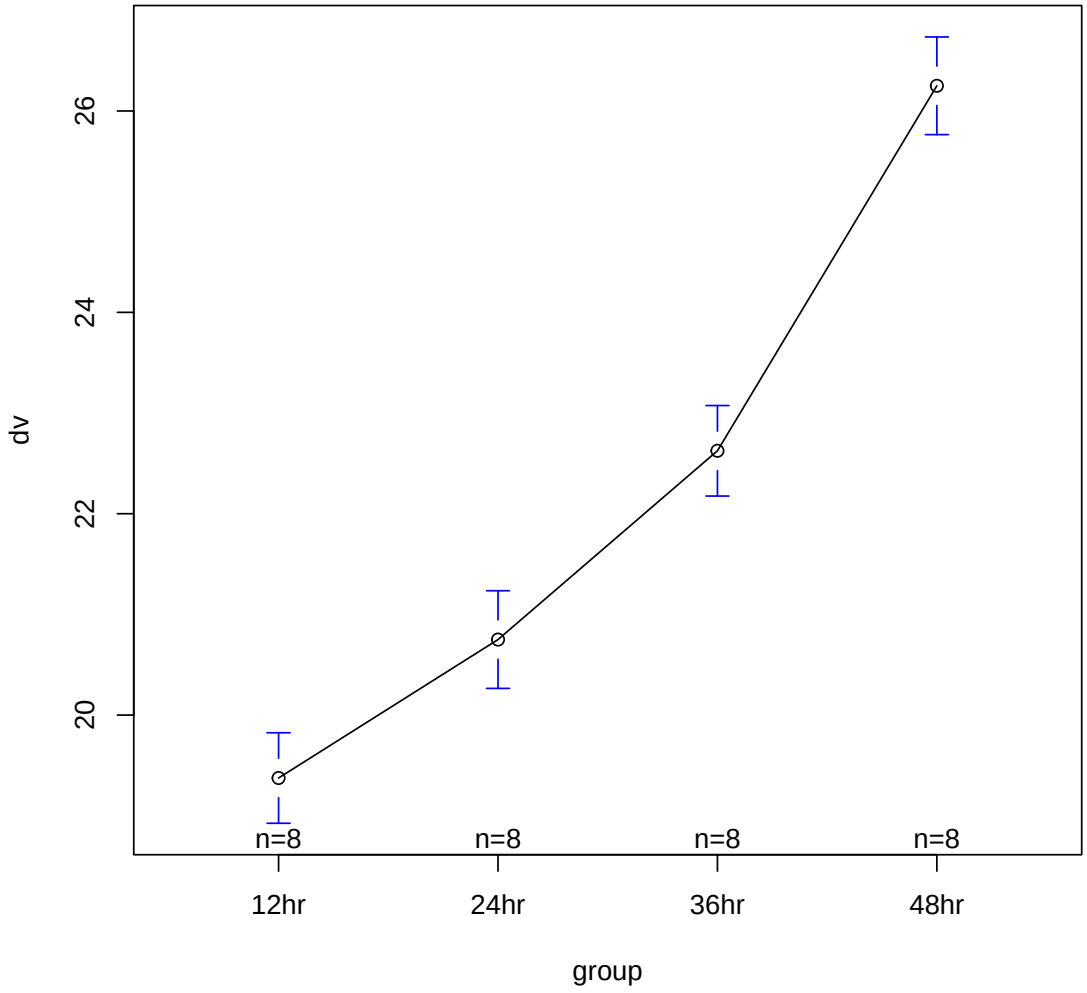
additional features.

```
library(gplots)
plotmeans(dv ~ group, data = data.sleep)
```



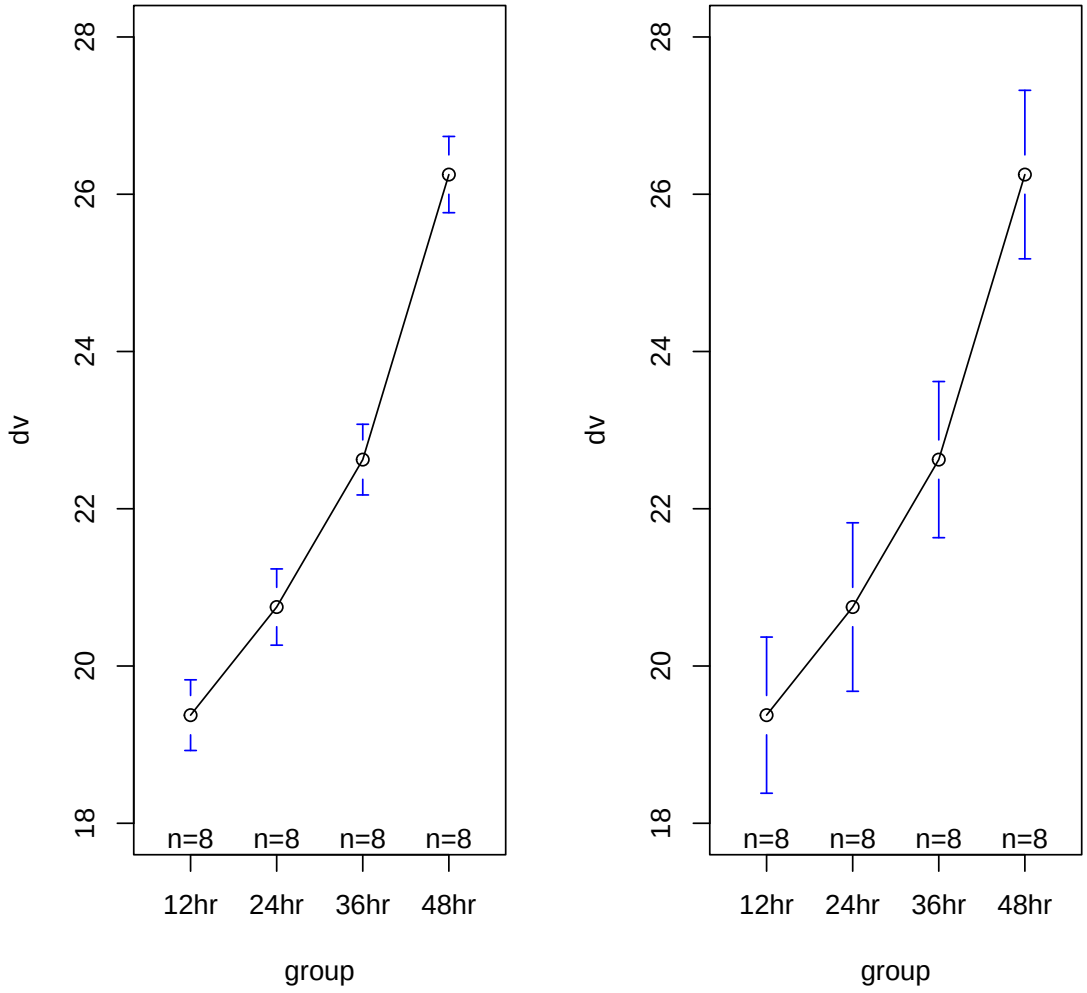
To plot plus/minus one standard error just set the CI to be 68% (which is the CI corresponding to plus/minus one standard error).

```
plotmeans(dv ~ group, data = data.sleep, p = 0.68)
```



Note that the y axis change slightly from the .95 to the .68 CI plot so if you want to put them side to side to compare the width of the intervals (in order to get an intuition of what plus/minus 1 vs plus/minus 2 looks like) be sure to set the ylim to be equal in both plots (e.g., add the argument ylim=c(18,28), or in both plotmeans calls, to set the y axis scale to range from, say, 18 to 28).

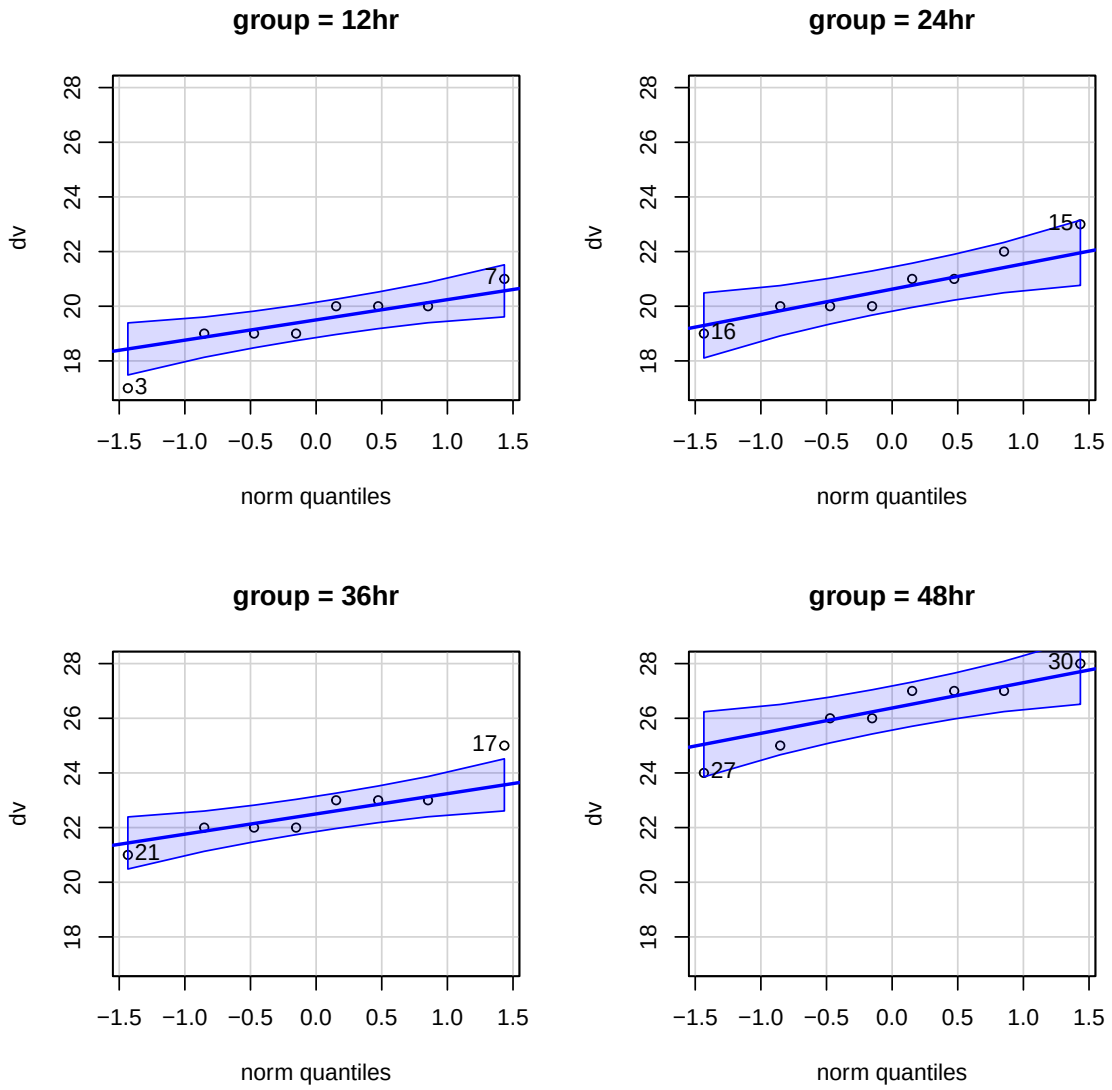
```
par(mfcol = c(1, 2))
plotmeans(dv ~ group, data = data.sleep, p = 0.68, ylim = c(18,
    28))
plotmeans(dv ~ group, data = data.sleep, ylim = c(18, 28))
```




```
# revert back to one plot per page
par(mfcol = c(1, 1))
```

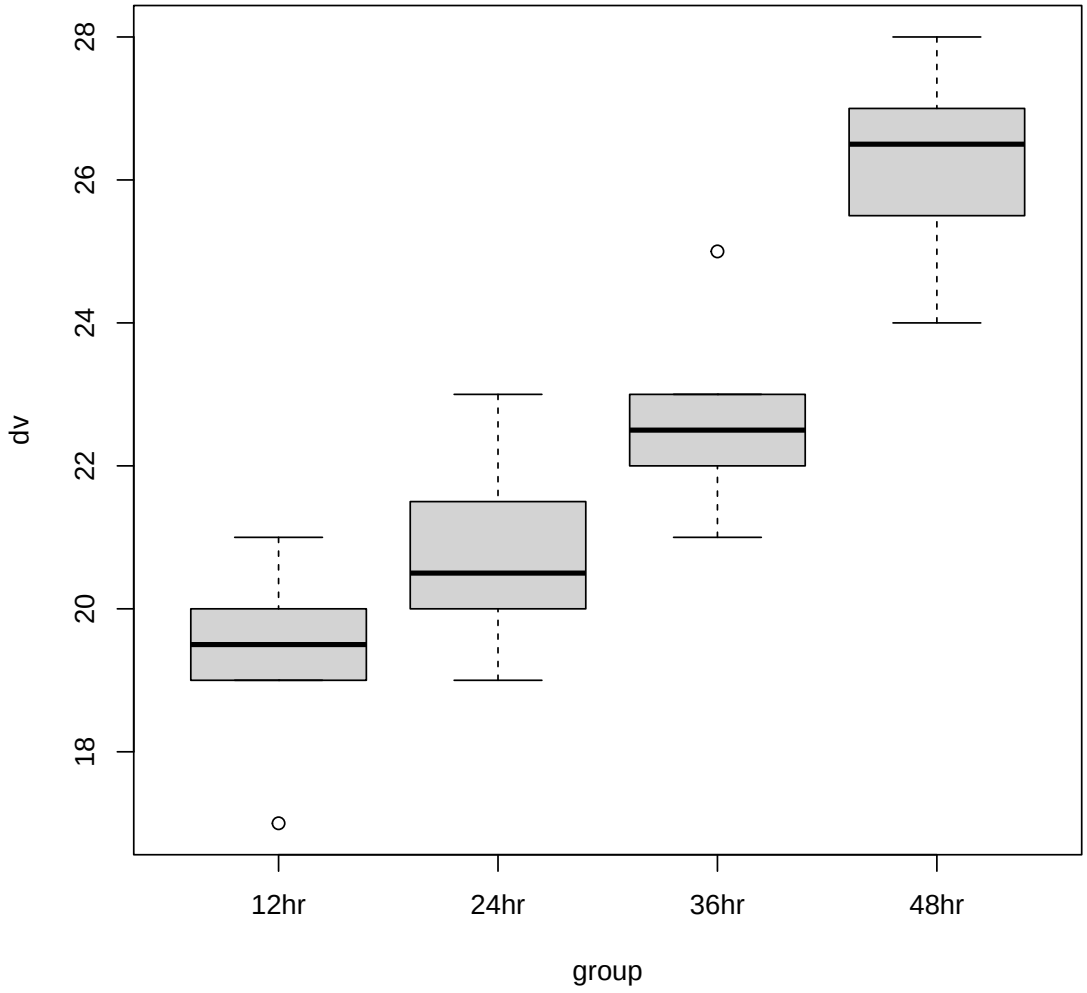
The car package has a nice qqPlot function for qqplots for each group using the formula notation

```
qqPlot(dv ~ group, data = data.sleep)
```



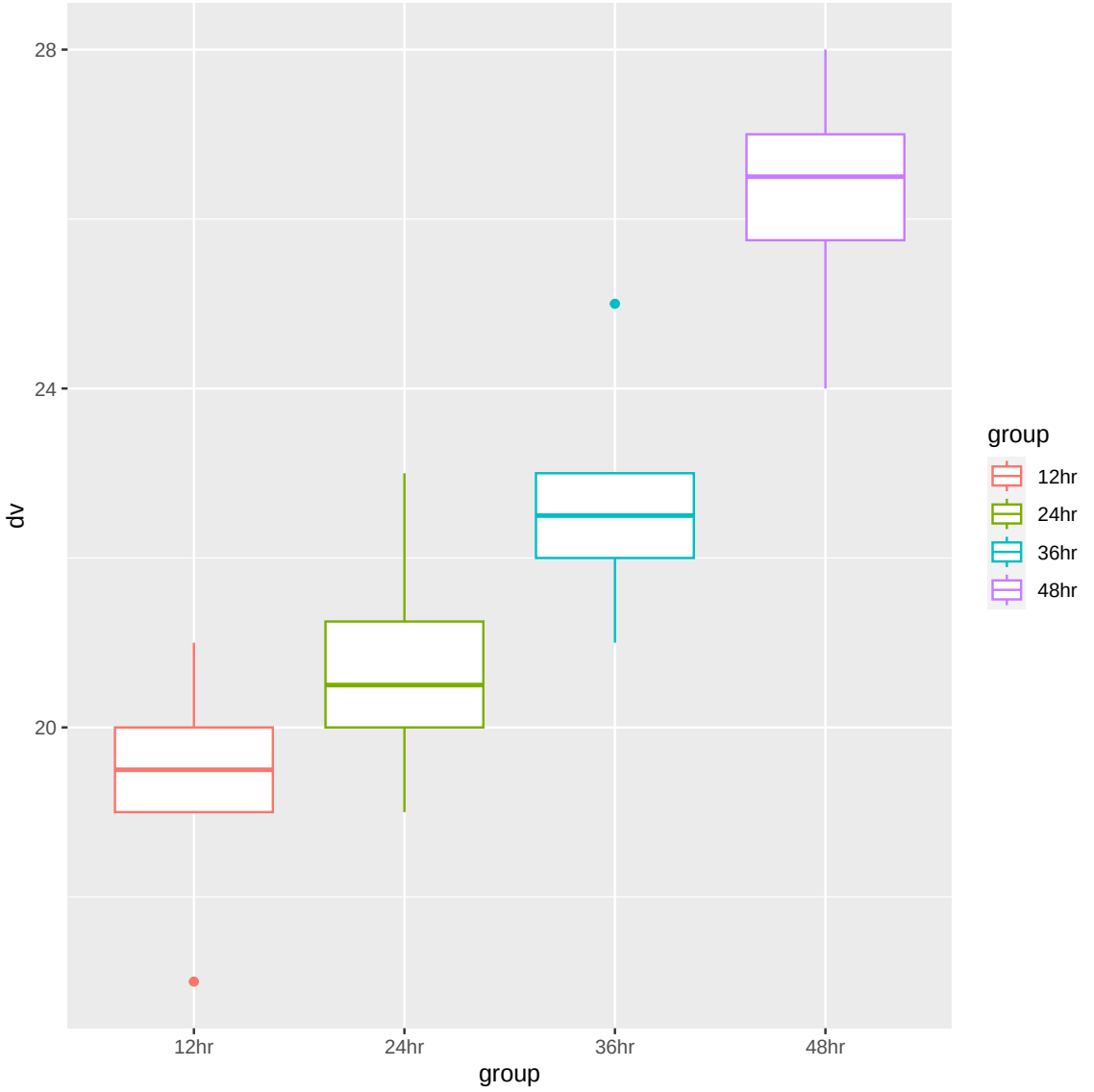
Boxplots can be done easily through

```
boxplot(dv ~ group, data = data.sleep)
```



or with the fancier ggplot2 package

```
library(ggplot2)
ggplot(data.sleep, aes(x = group, y = dv, color = group)) +
    geom_boxplot()
```


28
24
20
dv
12hr
24hr
36hr
48hr
group
group
12hr
24hr
36hr
48hr

Appendix 2

Example showing how "variance of residuals" (with corrected df) is equal to MSW in the ANOVA source table.

```
RAW DATA:  SLEEP DEPRIVATION DATA

  ROW   12hr   24hr   36hr   48hr

    1     20     21     25     26
    2     20     20     23     27
    3     17     21     22     24
    4     19     22     23     27
    5     20     20     21     25
    6     19     20     22     28
    7     21     23     22     26
    8     19     19     23     27

ONEWAY ANOVA ON RAW SCORES (JUST FOR REFERENCE)

 ANALYSIS OF VARIANCE
 SOURCE     DF        SS        MS        F        p
 FACTOR      3    213.25     71.08    46.56    0.000
 ERROR      28     42.75      1.53
 TOTAL      31    256.00

 POOLED STDEV =    1.236

SUBTRACT GROUP MEAN FROM EACH OBSERVED SCORE (CREATE NEW VARIABLES)
  compute 12hrE = 12hr - 19.38.
  compute 24hrE = 24hr - 20.75.
  compute 36hrE = 36hr - 22.63.
  compute 48hrE = 48hr - 26.25.
  execute.
PRINT RESIDUALS

  ROW     12hrE  24hrE     36hrE  48hrE
    1   0.62000   0.25   2.37000  -0.25
    2   0.62000  -0.75   0.37000   0.75
    3  -2.38000   0.25  -0.63000  -2.25
    4  -0.38000   1.25   0.37000   0.75
    5   0.62000  -0.75  -1.63000  -1.25
    6  -0.38000  -0.75  -0.63000   1.75
    7   1.62000   2.25  -0.63000  -0.25
    8  -0.38000  -1.75   0.37000   0.75


DESCRIPTIVE STATISTICS ON THE RESIDUALS

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
 RESIDUAL       32   -0.002    0.000   -0.002    1.174    0.208

              MIN      MAX       Q1       Q3
 RESIDUAL   -2.380    2.370   -0.720    0.718
```

The standard deviation of the residual scores is 1.174, so the variance is 1.378. But, this variance was computed by SPSS using N - 1 in the denominator. We need N - T. So, multiply 1.378 by the "correction factor" $\frac{N - 1}{N - T}$ to yield 1.53–the MSE from the ANOVA source table.

Just to check our understanding let's treat the residuals as the dependent variable and run an ANOVA. What do you expect the source table to look like? Try to think this through before looking at the answer below.

An ANOVA on the residuals yields MSB = 0 because the "effects" (that is, the treatment effects represented by each α) have been subtracted out so there isn't any SSB. The MSW = 1.53 is the same as the MSW for the raw data.

```
ONE WAY ANOVA ON THE RESIDUALS

 ANALYSIS OF VARIANCE
 SOURCE     DF        SS        MS        F        p
 FACTOR      3      0.00      0.00     0.00    1.000
 ERROR      28     42.75      1.53
 TOTAL      31     42.75

 POOLED STDEV =    1.236
```

Appendix 3

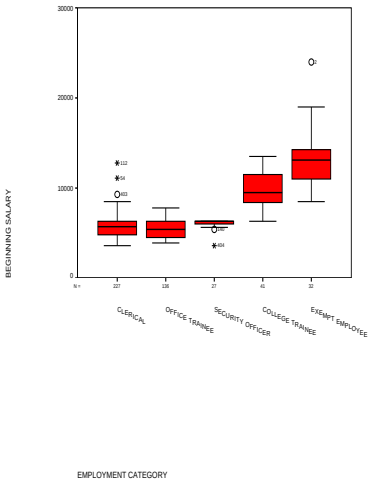
A more complete example of the bank data used by Norusis in the SPSS Base Manual.

select if (JOBCAT LE 5).

examine variables = salbeg by jobcat /plot=boxplot .

```
    SALBEG    BEGINNING SALARY
By  JOBCAT    1         CLERICAL

Valid cases:       227.0   Missing cases:        .0   Percent missing:      .0


Mean      5733.947  Std Err    84.4228  Min       3600.000  Skewness    1.2506
Median    5700.000  Variance   1617876  Max       12792.00  S E Skew     .1615
5% Trim   5661.713  Std Dev   1271.957  Range     9192.000  Kurtosis    4.4695
                                       IQR       1500.000  S E Kurt     .3217

    SALBEG    BEGINNING SALARY
By  JOBCAT    2         OFFICE TRAINEE

Valid cases:       136.0   Missing cases:        .0   Percent missing:      .0


Mean      5478.971  Std Err    80.3222  Min       3900.000  Skewness     .3660
Median    5400.000  Variance  877424.1  Max       7800.000  S E Skew     .2078
5% Trim   5440.490  Std Dev   936.7092  Range     3900.000  Kurtosis    -.9385
                                       IQR       1800.000  S E Kurt     .4127


    SALBEG    BEGINNING SALARY
By  JOBCAT    3         SECURITY OFFICER

Valid cases:        27.0   Missing cases:        .0   Percent missing:      .0


Mean      6031.111  Std Err   103.2483  Min       3600.000  Skewness   -3.8758
Median    6300.000  Variance  287825.6  Max       6300.000  S E Skew     .4479
5% Trim   6125.309  Std Dev   536.4938  Range     2700.000  Kurtosis   17.2035
                                       IQR       300.0000  S E Kurt     .8721

    SALBEG    BEGINNING SALARY
By  JOBCAT    4         COLLEGE TRAINEE

Valid cases:        41.0   Missing cases:        .0   Percent missing:      .0


Mean      9956.488  Std Err   311.8593  Min       6300.000  Skewness     .1221
Median    9492.000  Variance   3987506  Max       13500.00  S E Skew     .3695
5% Trim   9954.374  Std Dev   1996.874  Range     7200.000  Kurtosis   -1.1850
                                       IQR       3246.000  S E Kurt     .7245

    SALBEG    BEGINNING SALARY
By  JOBCAT    5         EXEMPT EMPLOYEE

Valid cases:        32.0   Missing cases:        .0   Percent missing:      .0


Mean      13258.88  Std Err   556.1423  Min       8496.000  Skewness    1.4015
Median    13098.00  Variance   9897415  Max       24000.00  S E Skew     .4145
5% Trim   13010.25  Std Dev   3146.016  Range     15504.00  Kurtosis    3.2323
```

```
                    IQR       3384.000   S E Kurt      .8094
```



The boxplot suggests a violation of the equality of variance assumption. The sample sizes are large enough that if the assumption was met we wouldn't expect to see this degree of deviation, so we conclude that the equality of variance assumption is suspect. I'll first perform an ANOVA on the raw data and then check out a possible transformation.

oneway salbeg by jobcat
/statistics all .

```
   Variable  SALBEG     BEGINNING SALARY
By Variable  JOBCAT     EMPLOYMENT CATEGORY

                                ANALYSIS OF VARIANCE

                                 SUM OF         MEAN            F      F
       SOURCE            D.F.    SQUARES       SQUARES       RATIO  PROB.
BETWEEN GROUPS              4   2230311013   557577753.4    266.5954  .0000
WITHIN GROUPS             458   957895695.7  2091475.318
TOTAL                     462   3188206709

                                 STANDARD   STANDARD
GROUP       COUNT        MEAN    DEVIATION      ERROR     MINIMUM     MAXIMUM     95 PCT CONF INT FOR MEAN

Grp 1         227   5733.9471   1271.9574    84.4228   3600.0000  12792.0000    5567.5907 TO   5900.3036
Grp 2         136   5478.9706    936.7092    80.3222   3900.0000   7800.0000    5320.1181 TO   5637.8231
Grp 3          27   6031.1111    536.4938   103.2483   3600.0000   6300.0000    5818.8812 TO   6243.3410
Grp 4          41   9956.4878   1996.8740   311.8593   6300.0000  13500.0000    9326.1966 TO  10586.7790
Grp 5          32  13258.8750   3146.0158   556.1423   8496.0000  24000.0000   12124.6153 TO  14393.1347

TOTAL         463   6570.3801   2626.9527   122.0848   3600.0000  24000.0000    6330.4697 TO   6810.2905

          FIXED EFFECTS MODEL   1446.1934    67.2103                          6438.3013  TO   6702.4589

         RANDOM EFFECTS MODEL              1583.1406                          2174.9485  TO  10965.8117
```

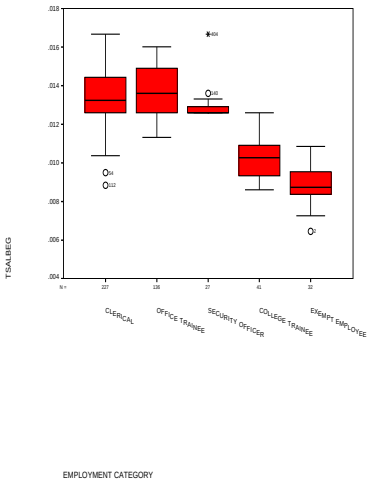
```
RANDOM EFFECTS MODEL - ESTIMATE OF BETWEEN COMPONENT VARIANCE   7300834.4825
```

It seems from running the spread and level plot that a -.5 slope (equivalent to the reciprocal of the square root transformation) will help with the equality of variance assumption.

compute tsalbeg = 1/sqrt(salbeg).
execute.

examine variables = tsalbeg by jobcat /plot = boxplot .

```
    TSALBEG
By  JOBCAT    1         CLERICAL

Valid cases:       227.0   Missing cases:        .0   Percent missing:      .0


Mean         .0134  Std Err      .0001  Min          .0088  Skewness     .1002
Median       .0132  Variance     .0000  Max          .0167  S E Skew     .1615
5% Trim      .0134  Std Dev      .0014  Range        .0078  Kurtosis    -.0697
                                       IQR          .0018  S E Kurt     .3217


    TSALBEG
By  JOBCAT    2         OFFICE TRAINEE

Valid cases:       136.0   Missing cases:        .0   Percent missing:      .0


Mean         .0137  Std Err      .0001  Min          .0113  Skewness    -.0638
Median       .0136  Variance     .0000  Max          .0160  S E Skew     .2078
5% Trim      .0137  Std Dev      .0011  Range        .0047  Kurtosis   -1.2928
                                       IQR          .0023  S E Kurt     .4127


    TSALBEG
By  JOBCAT    3         SECURITY OFFICER

Valid cases:        27.0   Missing cases:        .0   Percent missing:      .0


Mean         .0129  Std Err      .0002  Min          .0126  Skewness    4.4368
Median       .0126  Variance     .0000  Max          .0167  S E Skew     .4479
5% Trim      .0128  Std Dev      .0008  Range        .0041  Kurtosis   21.3123
                                       IQR          .0003  S E Kurt     .8721


    TSALBEG
By  JOBCAT    4         COLLEGE TRAINEE

Valid cases:        41.0   Missing cases:        .0   Percent missing:      .0


Mean         .0102  Std Err      .0002  Min          .0086  Skewness     .3008
Median       .0103  Variance     .0000  Max          .0126  S E Skew     .3695
5% Trim      .0101  Std Dev      .0010  Range        .0040  Kurtosis    -.8333
                                       IQR          .0016  S E Kurt     .7245

    TSALBEG
By  JOBCAT    5         EXEMPT EMPLOYEE

Valid cases:        32.0   Missing cases:        .0   Percent missing:      .0


Mean         .0088  Std Err      .0002  Min          .0065  Skewness    -.2254
Median       .0087  Variance     .0000  Max          .0108  S E Skew     .4145
5% Trim      .0089  Std Dev      .0009  Range        .0044  Kurtosis     .2456
                                       IQR          .0012  S E Kurt     .8094
```



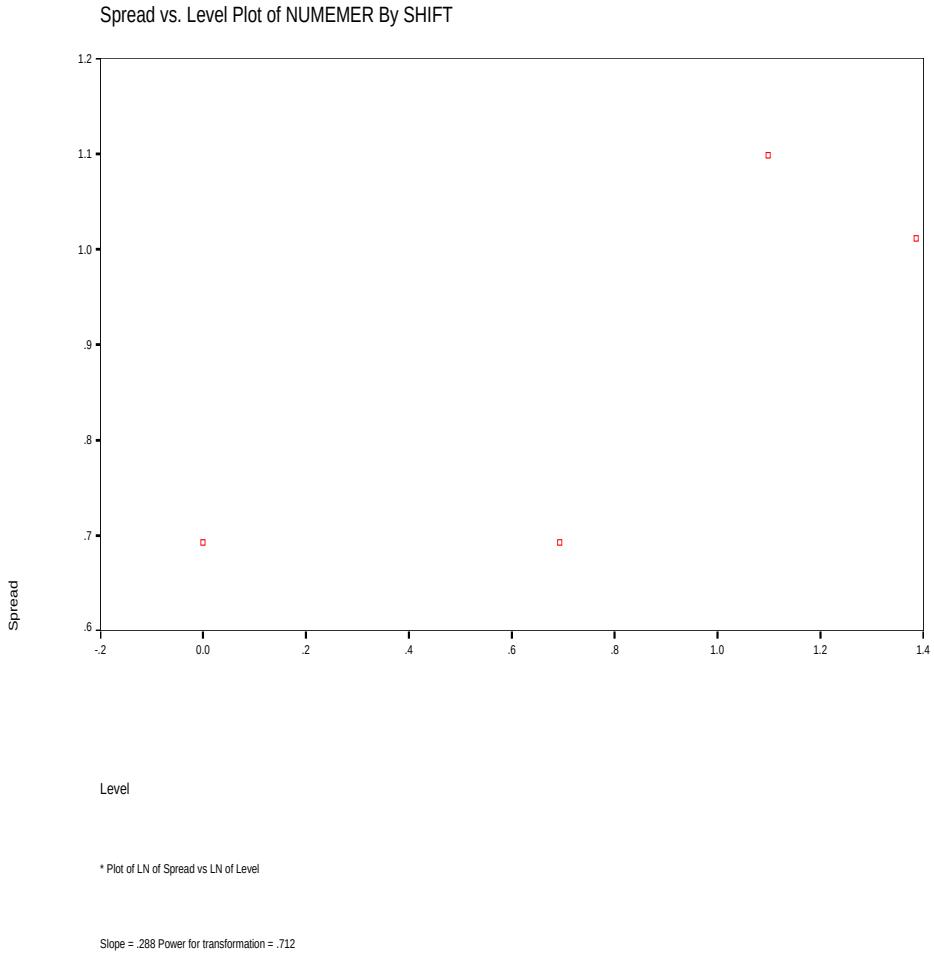
The transformation didn't do a perfect fix but with one exception (the middle group) the boxplots look better than the original data.

oneway tsalbeg by jobcat
/statistics all .

```
     Variable  TSALBEG
  By Variable  JOBCAT     EMPLOYMENT CATEGORY

                              ANALYSIS OF VARIANCE

                                SUM OF          MEAN             F      F
       SOURCE           D.F.    SQUARES        SQUARES          RATIO  PROB.
BETWEEN GROUPS             4        .0010          .0002     155.9406  .0000
WITHIN GROUPS            458        .0007          .0000
TOTAL                    462        .0017


                                STANDARD   STANDARD
GROUP       COUNT       MEAN    DEVIATION      ERROR     MINIMUM     MAXIMUM     95 PCT CONF INT FOR MEAN

Grp 1         227       .0134       .0014      .0001       .0088       .0167       .0132 TO       .0136
Grp 2         136       .0137       .0011      .0001       .0113       .0160       .0135 TO       .0138
Grp 3          27       .0129       .0008      .0002       .0126       .0167       .0126 TO       .0132
Grp 4          41       .0102       .0010      .0002       .0086       .0126       .0098 TO       .0105
Grp 5          32       .0088       .0009      .0002       .0065       .0108       .0085 TO       .0092

TOTAL         463       .0129       .0019      .0001       .0065       .0167       .0127 TO       .0130

         FIXED EFFECTS MODEL       .0012      .0001                              .0127  TO       .0130

         RANDOM EFFECTS MODEL                 .0010                              .0100  TO       .0158

RANDOM EFFECTS MODEL - ESTIMATE OF BETWEEN COMPONENT VARIANCE         0.0000
```

Appendix 4: Helicopter data

```
examine variables = numemer by shift
    /statistics=none
    /plot=spreadlevel .
```



Thus the spread and level plot suggests somewhere between a square root transformation and the identity. Most likely a transformation is not needed in this case–not just because the spread and level plots says so but it is also obvious from the boxplots. First, let's look at the raw helicopter data, then the transformed (square root).

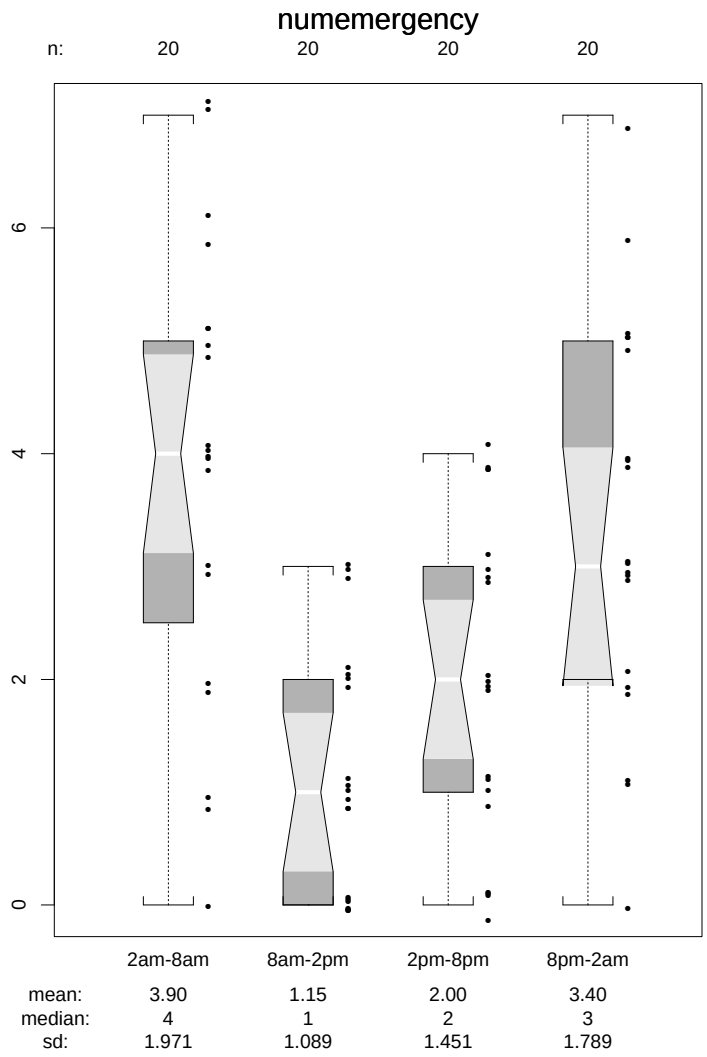


Figure 2-8: Raw helicopter data. Transformation doesn't seem to be needed here. This is a different form of the boxplot, called a notched boxplot, where the light gray notch depicts plus/minus one standard error around the median and the dark gray still represents the usual middle 50%. The standard error around the median is typically defined as 1.5*IQR/n

.

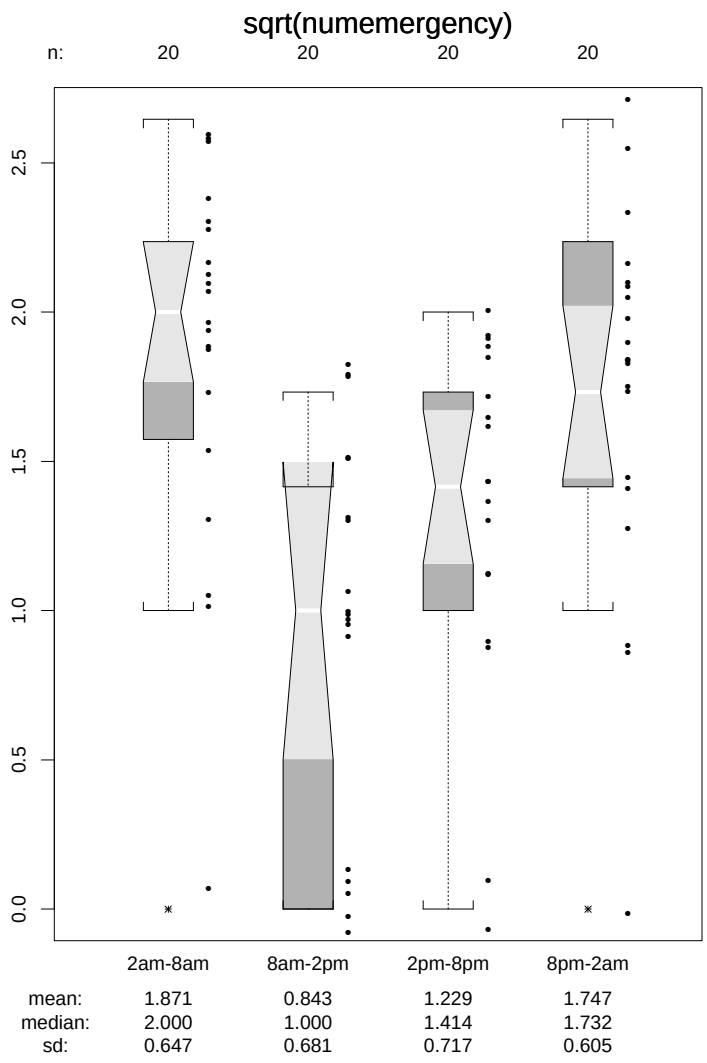
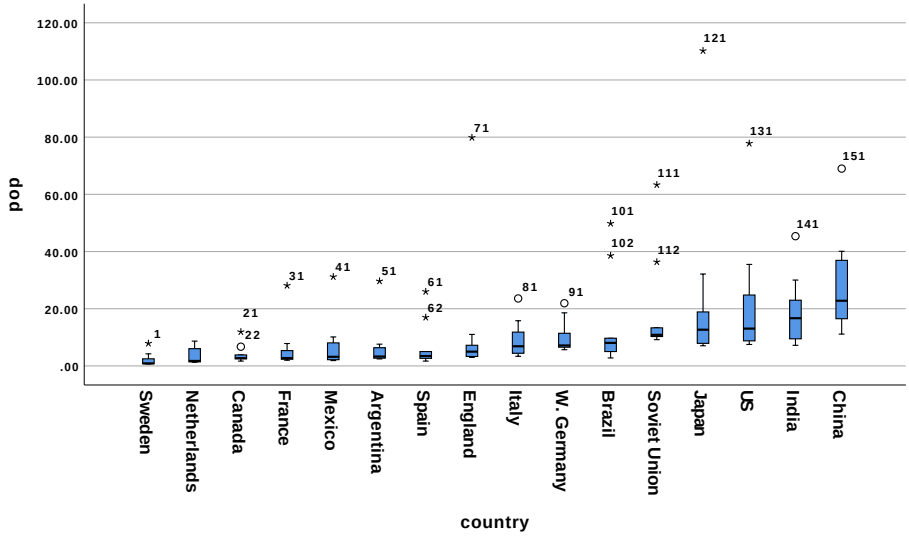


Figure 2-9: Square root of the helicopter data. The spread and level plot suggests that if a transformation is needed, then it might be the square root given that we limit ourselves to the "power family". It isn't clear that the transformed boxplot is any better than the original. So, we'd opt for not performing a transformation for these data.

Appendix 5: Population data

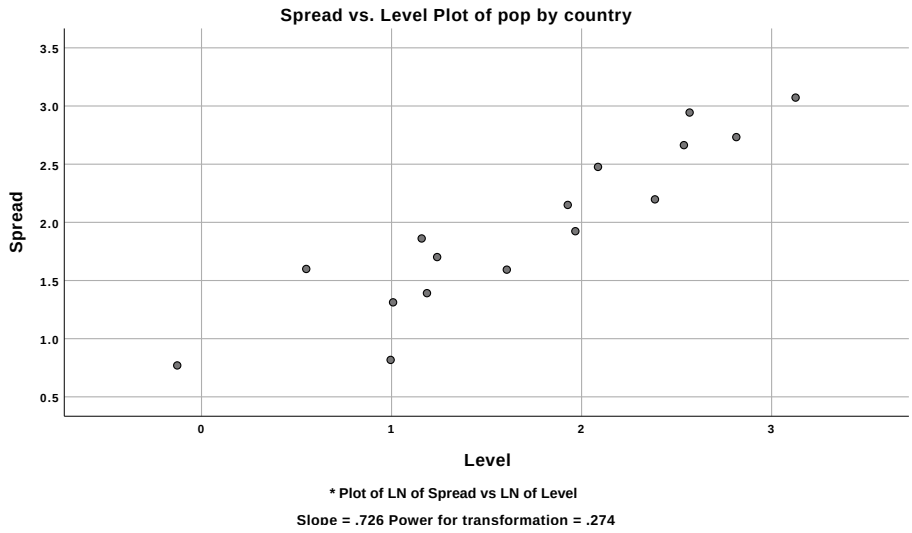
Here is an example of something that clearly violates the equal variance assumption and the sample size is large.



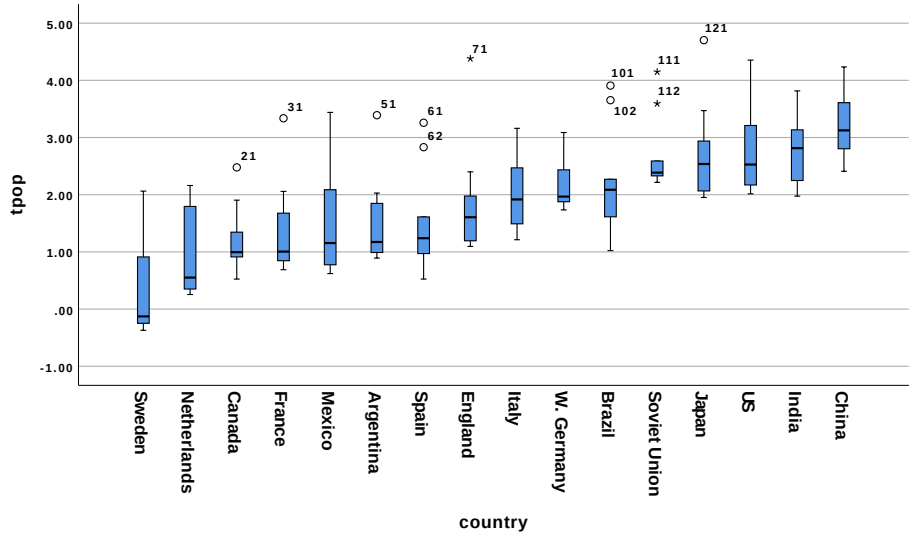
The spread and level plot

```
examine variables = pop by country
  /statistics=none
  /plot=spreadlevel .
```

```
* Plot of LN of Spread vs LN of Level.
Slope =    .726          Power for transformation =    .274
Test of homogeneity of variance             df1       df2         Significance
Levene Statistic               2.4580        15       144             .0031
```



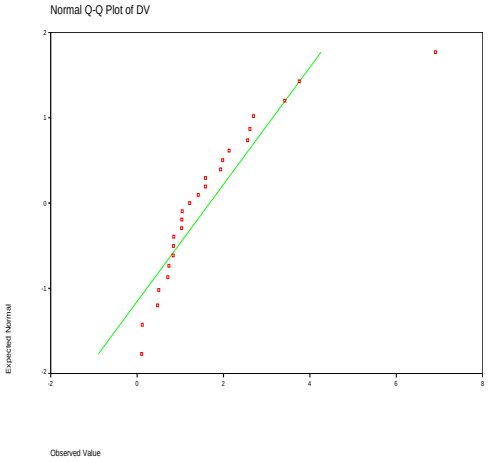
Based on the spread and level results I decided to do a log transformation and I replot the boxplot on the transformed data to double check the result is better than in the raw data.

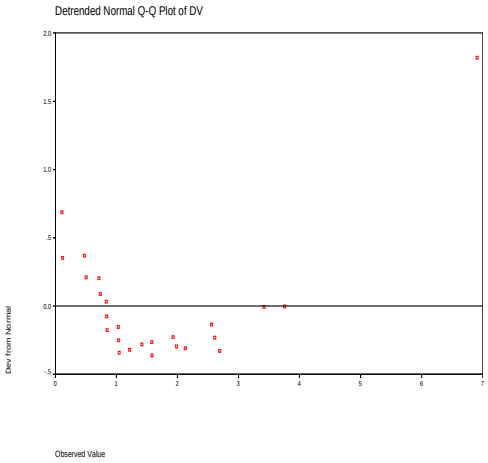


Appendix 6: Transformations and Symmetry

Showing the use of transformation for satisfying the assumption of normality. I generated some random data. A normal plot follows. I know that such a distribution can be transformed, or re-expressed, into something that looks more normal by performing a log transformation. A second normal plot shows the "improvement" on the log scale.

```
 data list free / dv.
begin data.
 1.0355169
 2.6118367
 0.8492587
 3.4192738
 2.1297903
 1.9830919
 3.7556537
 0.5049452
 1.2166372
 1.0311578
 1.5835213
 0.8411258
 1.0442388
 0.1184379
 0.4746945
 0.7125102
 1.5810627
 0.1060652
 2.5585323
 0.8339119
 1.4155780
 2.6909101
 6.9099978
 1.9278372
 0.7376325
end data.

examine variables = dv
  /plot = boxplot npplot .
      DV
 Valid cases:        25.0   Missing cases:        .0   Percent missing:      .0

 Mean        1.6829  Std Err      .2914  Min          .1061  Skewness    2.1090
 Median      1.2166  Variance    2.1221  Max         6.9100  S E Skew     .4637
 5% Trim     1.5151  Std Dev     1.4568  Range       6.8039  Kurtosis    6.0512
                                         IQR         1.5584  S E Kurt     .9017
```

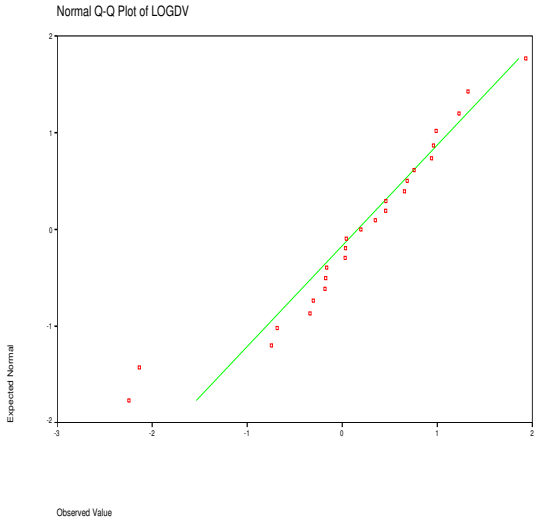


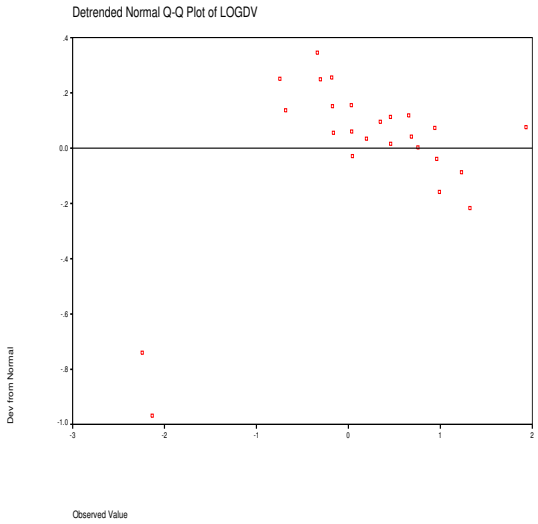


compute logdv = ln(dv).
execute.

examine variables = logdv
/plot = boxplot npplot .

```
LOGDV
 Valid cases:        25.0   Missing cases:        .0   Percent missing:      .0

 Mean          .1630  Std Err      .1918  Min        -2.2437  Skewness    -.8914
 Median        .1961  Variance     .9199  Max         1.9330  S E Skew     .4637
 5% Trim       .2039  Std Dev      .9591  Range       4.1767  Kurtosis    1.4631
                                         IQR         1.0907  S E Kurt     .9017
```


Normal Q-Q Plot of LOGDV
Expected Normal
Observed Value
2
1
0
-1
-2
-3
-2
-1
0
1
2


Detrended Normal Q-Q Plot of LOGDV
Dev from Normal
Observed Value
.4
.2
0.0
-.2
-.4
-.6
-.8
-1.0
-3
-2
-1
0
1
2