INDUCED ENTRY DEMONSTRATION DESIGNS

$1 for $2 Benefit Offset: Exploring the Feasibility of Measuring Induced Entry and Evaluating Potential County-Level Project Designs

June 12, 2001

This paper is a working draft prepared jointly by L. Scott Muller, Robert Weathers, John Hennessey, and Fred Bellemore of the Office of Research, Evaluation and Statistics.

May 2001

Measuring Induced Entry Using County Level Data

The Ticket to Work and Work Incentives Improvement Act (TWWIIA) of 1999 requires SSA to design a title II disability $1-for-$2 benefit offset demonstration project that will evaluate, among other things, the effect of induced entry associated with the national implementation of such a program. The Office of the Chief Actuary (OCACT) estimated that there are a significant number of people who are eligible for the program based upon their medical condition but who choose to work under the existing program rules. Under a $1 for $2 above SGA, OCACT estimated that many of these people would leave their current situation (work, unemployment benefits, temporary disability benefits, workers compensation, etc.) and be "induced" to enter the DI program due to the $1-for-$2 benefit offset. Reliable measures of induced entry into the DI program based on observed behavioral responses from a demonstration project may not be possible due to both statistical issues and study effects.

The statistical issues arise because the induced entry effect size that leads to large program costs is relatively small. Identification of the effect size in a randomized experiment, even under the most generous parameters and sampling techniques, would require a demonstration much too large to be manageable. Quasi-experimental techniques, such as a state level design or a county level design where some states are provided the treatment and others serve as a comparison group, introduce a different set of problems. In a set of simulations using the actual data that would be used if SSA were to carry out such a design, it was determined that the existing variation in state level awards rates is likely to lead to misleading results. In some cases, the results indicated that such an analysis could imply "reduced entry" (negative coefficient on the treatment) in a situation where we artificially place an "induced entry" effect (true treatment effect is positive) into the data. This results from inadequate controls for (or specification of) within state (or county) time-varying unobserved factors which affect the state level awards rate. Even if quasi-experimental designs could statistically identify the impact of induced entry into the SSDI program resulting from a $1-for-$2 benefit offset, the results are likely to be biased by study effects. Study effects could arise in such a quasi-experimental design

from migration to locations where the $1-for-$2 is implemented, through the inadequate provision of information, through the decision to "apply before the $1-for-$2 is no longer available", and other factors. Assessing the impact of study effects is extremely difficult.

In this note, we describe the research undertaken to determine the feasibility of measuring induced entry utilizing a quasi-experimental design based upon county level data. Although the study effects associated with a county level demonstration project were expected to be more pronounced compared to a state level design (particularly problems with controlling information flow and migration into areas offering the benefit offset), it was hypothesized that a randomized quasi-experimental design based upon 3,000 counties may serve as a more effective means of identifying the treatment effect. That is, in a county level design, a larger number of counties could be randomly assigned to the treatment and comparison groups and the randomized assignment over a larger number of units is likely to serve as a more adequate control for unobserved factors that could potentially bias the results.

The Data

The analysis used county-level SSDI allowance rates, where the numerator (SSDI allowances) was obtained from SSA administrative records and the denominator (county population) was obtained from census reports. Since SSA administrative records do not include county identifiers, zip codes were used to identify counties. There is not a perfect match between zip code border and county borders so there is some discrepancy in the county-level counts of allowances. However, this is only data available for such an evaluation and it would be the data used in the actual evaluation of induced entry.

The Method

The method used in assessing the effectiveness of a county level design is similar to that employed in the state-level design, that is, a simulated demonstration was constructed with known treatment effects embedded in the data. Six years of county-level data on allowance rates were constructed and a sample of counties was randomly selected to be "demonstration" counties and a treatment effect was

embedded in the latter 3 years of the data. Regression analysis was used to control for time-invariant county-level differences, overall differences in awards rates over time, and to estimate the "treatment" effect. The estimated "treatment" effect was then compared to the known effect that was embedded in each "treatment" county's data.

As in the state-level analysis, several different treatment effects were simulated (no effect, 2% increase, 5% increase, and 10% increase). Furthermore, in order to assure an efficient design the demonstration project SSA would need to limit the number of sites in which the project was implemented. This raises two issues: (1) should small counties where few allowances would be expected (and thus little induced entry) be included in the demonstration project and (2) how many counties would be needed to be included in the project to assure sufficient power to detect the simulated effect. In order to address these issues the simulation was performed using different numbers of counties (25, 50, 100, or 200 counties) randomly selected from the following: all counties, counties over 25,000 population, and counties over 50,000 population. In the earlier examination of state-level models that focussed on selected groupings of states or matched pairs of states, the evaluation was limited to a single simulation for each grouping or pair of states. However, in the county level analysis, where treatment counties are randomly selected from all counties, the random selection and model estimation can be repeated to assess the variation in simulation results that occurs with different, repeated sample selection. In order to assess the consistency of results across random draws, and the "power" of the county level design to detect the "treatment effect", each simulation model that was estimated 200 times. Thus for each combination of treatment effect, size of counties, and number of counties 200 regressions were run.

Assessing the ability to measure induced entry

If the method of measuring induced entry using county-level data proved to be accurate one would obtain an estimated coefficient on the treatment variable that, when transformed into percentages, matched the embedded treatment effect and was statistically different from zero. However, that goal was unlikely to be achieved in every regression run. What interpretation would one use if an actual demonstration project obtained a particular

estimate, but the t-statistic proved the estimate was not different from zero? Would one declare that there was no induced entry effect? Would one utilize the transformed coefficient as our "best estimate" of induced entry? Would one offer policymakers an interval that indicated we were 95 percent confident that the true value was with the appropriate limits (which would run from some negative number to some positive number)?

It was evident that our assessment of the accuracy of the estimates required more than simply an estimate of the coefficient and the induced entry effect. The analysis had to take into account the range of the estimates, the significance of the estimates and the variance in the estimates. Thus, in order to assess the accuracy of each approach, 5 measures were examined:

1) mean estimated effect from the 200 trials;
2) the range (minimum and maximum) of the estimated effect;
3) the percentage of estimates falling with 1 percentage point (+/-) of the embedded effect;
4) the percentage of estimates that are statistically significant and fall with 1 percentage point (+/-) of the embedded effect;
5) the percentage of estimates whose 95 percent confidence interval fell within a certain number of percentages points (+ and -) of the embedded effect.

The first measure, the mean estimated effect from the 200 trials, provides an indication of whether the estimate approaches the true value with repeated trials, or whether there appears to be a bias in the measurement of induced entry. The range of the estimates (i.e., the minimum and maximum estimates) provides some context for how overstated or understated the actual effect might be given a single estimate from a demonstration project.

The third measures, the proportion of trials that resulted in estimates that were within 1 percentage point (+/-) of the true (embedded) effect, provides some idea of the ability of the county-level design to achieve an estimate that is close to the true effect. While 1 percentage point may be a 50% over or underestimate of the 2 percent embedded effect, or a 10% over or underestimate of the 10 percent embedded effect, this is a good measure of accuracy since a 1% percentage point increase in allowances represents a fixed level of increased program costs (attributable to error in measurement), regardless of the

overall level of induced entry. Because it was unclear how one should interpret estimated effects that are not statistically significant, a fourth measure the percentage of estimates that are both statistically significant and fall with 1 percentage point (+/-) of the embedded effect. In a sense this measure represents a power calculation, reflecting how often we were able to statistically discern an existing difference (1 percentage point) between the two groups (treatment and comparison counties). Absent study effects, one might press for a design that could achieve this measure in 80 or 90 percent of the trials which is analogous to traditional power levels.

Finally, a fifth measure was examined to assess how large the confidence intervals were on the estimates. This measure might be considered of lesser order than measure 4, but still providing some measure of how often the range of estimates implied by the confidence interval was close to the actual embedded effect. Larger confidence intervals clearly imply a larger potential error in cost estimates. The fifth measure is the percentage of estimates whose 95 percent confidence interval fell within a certain number of percentages points (+ and -) of the embedded effect. The fifth measure, in fact, is three measures with the confidence intervals specified as being with 2, 3 or 4 percentage points of the embedded effect.

Findings

The results of the simulation are shown in Table 1. Not surprisingly, the ability to accurately detect a treatment effect increased with the size of the embedded treatment effect, as well as with the number of counties selected for the treatment. Limiting the analysis to larger, more homogeneous counties also improved the ability to accurately detect the embedded treatment effect.

The issue, however, is whether quasi-experimental designs can detect the true effect accurately enough and with sufficient likelihood to justify the resources and expense to mount such a demonstration project. It is important to note that the smallest demonstration project (25 counties selected from all size counties) would involve offering the treatment to approximately 2.2 million persons and the largest demonstration project (200 counties selected from counties with populations in excess of 50,000) would involve offering over 52 million persons the treatment.

Thus any county level design would necessarily be large, expensive to undertake, and place the disability program at risk of substantial growth and increased program costs if there is even a modest induced entry effect.

The simulation results suggest that none of the designs that sample from all counties are sufficient to measure induced entry. Even the largest demonstrations (200 treatment counties and 17 million persons) with large embedded effects (5 or 10 percent increases in allowances) had barely a 50-50 chance of picking up a treatment effect that was significant and within one percentage point of the simulated treatment effect. The range of estimates obtained was large, even among designs with large number of counties.

The simulations performed on designs that used only counties with populations above 25,000 improved the results. The range of minimum and maximum treatment estimates decreased substantially from the results for all counties. Designs with 200 counties (nearly 34 million offered the treatment) began to approach the 80 percent target that would be analogous to conventional power calculations. Designs with 100 counties (nearly 17 million offered the treatment) were accurate over 50 percent of the time for simulated induced entry in the 5 to 10 percent range.

Finally, simulations based upon designs that selected from counties whose population exceeded 50,000 performed quite similarly to those based upon counties of 25,000 population, with slight improvement in results among the smaller county sample sizes. The simulations showed that there was little gain in the accuracy of results: in no case did the simulation achieve the 80 percent target. The number of persons being offered the treatment, however, rose by 50 percent. Selecting counties with populations exceeding 50,000 can be rejected in favor of selecting counties with populations 25,000+ for 3 reasons: there was little or no improvement to the estimates, the sample size (and risk to the agency) was considerably larger, and the sample would be less representative of the national experience than designs employing counties exceeding 25,000 population.

The question then becomes whether the results that may be obtained from designs sampling counties with populations

exceeding 25,000 are worth pursuing. Generally speaking, designs selecting 25 or 50 counties were not able to adequately perform and can not be considered. Designs utilizing samples of 100 or 200 were large (17 and 34 million offered the treatment, respectively), and none achieve the 80 percent rate that is generally viewed as the minimum acceptable power level for randomized field trials. The 200 county sample did approach this level, however the size of the sample and risk to the agency for costs associated with induced entry from a large sample is a concern. Beyond this, it is important to consider the difficulty county-level designs will have with study effects (e.g., in migration and out migration) that cannot be controlled in the analysis and may bias the results. Furthermore, the simulated treatment effect in this analysis is a constant fixed increase in allowances. When the demonstration project is placed in the field it is likely that the effect will not be so definitive and the models may have a greater difficulty detecting and measuring the induced entry effect than is suggested by this analysis.

Conclusion

This analysis provides strong evidence that even under the ideal circumstances (i.e., no study effects, a fixed constant effect) created in this simulation, it is unlikely that quasi-experimental designs based upon county-level analysis will be able to provide policymakers with adequate and accurate information on the potential for induced entry. Given the inability to assure that even a large, costly, and complex demonstration project could detect induced entry, the recommendation remains that SSA consider alternative approaches to directly measuring induced entry in the context of a demonstration project. Model based approaches using NSHA data, survey questions, and obtaining actuarial estimates are the suggested alternatives.

SUMMARY OF COUNTY LEVEL SIMULATIONS

Actual Effect Size	Number Treatme Counties	Population for Inclusion	Treatment County Population	Estimated Effect	Minimum Treatment Effect	Maximum Treatment Effect	Est Trtmt Within 1 Actual	% and Significant	Confidence Interval Within 2	Confidence Interval Within 3	Confidence Interval Within 4
0	25	All	2,217,059	0.4	-12.0	12.4	24.5	0.0	0.0	0.0	0.0
0	50	All	4,296,680	0.2	-7.6	9.0	34.0	0.0	0.0	0.0	0.0
0	100	All	8,597,392	0.2	-5.2	5.0	41.5	0.0	0.0	0.0	30.5
0	200	All	17,192,541	0.2	-3.7	3.8	52.5	0.0	0.0	41.5	78.0
2	25	All	2,223,300	2.4	-10.2	14.4	23.0	0.0	0.0	0.0	0.0
2	50	All	4,296,680	2.1	-5.8	11.0	34.5	0.0	0.0	0.0	0.0
2	100	All	8,597,392	2.1	-3.3	7.0	40.5	0.0	0.0	0.0	34.0
2	200	All	17,192,541	2.1	-1.7	5.8	54.0	19.5	0.0	41.0	79.0
5	25	All	2,217,059	5.3	-7.6	17.4	23.0	0.0	0.0	0.0	0.0
5	50	All	4,296,680	5.0	-3.0	14.0	35.5	31.5	0.0	0.0	0.0
5	100	All	8,597,392	5.1	-0.4	9.9	40.0	40.0	0.0	0.0	32.5
5	200	All	17,192,541	5.0	1.2	8.7	53.5	53.5	0.0	41.0	77.5
10	25	All	2,217,059	10.1	-3.2	22.4	21.5	21.5	0.0	0.0	0.0
10	50	All	4,296,680	9.8	1.5	18.9	32.0	32.0	0.0	0.0	0.0
10	100	All	8,597,392	9.9	4.4	14.9	39.5	39.5	0.0	0.0	32.0
10	200	All	17,192,541	9.9	6.1	13.7	57.0	57.0	0.0	42.0	74.5

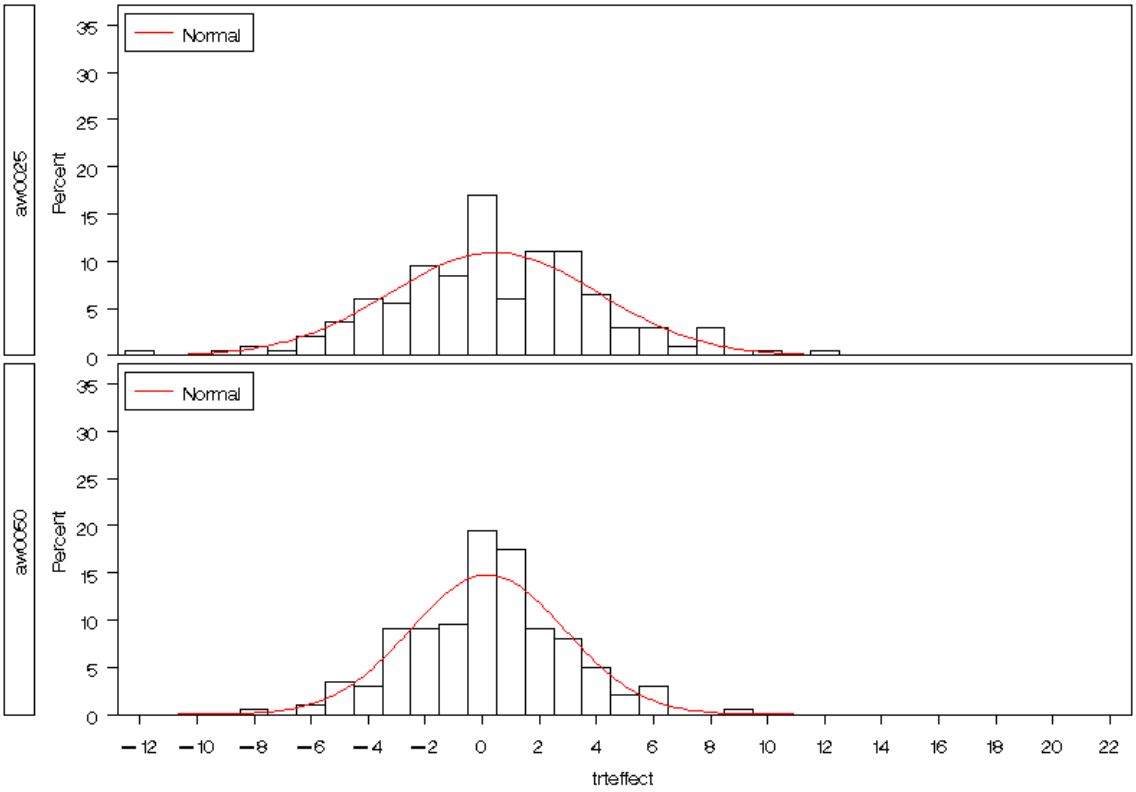
SUMMARY OF COUNTY LEVEL SIMULATIONS

Actual Effect Size	Number of Treatment Counties	Population for Inclusion	Treatment County Population	Estimated Effect	Minimum Treatment Effect	Maximum Treatment Effect	Est Trtmt Within 1 of Actual	% and Significant	Confidence Interval Within 2	Confidence Interval Within 3	Confidence Interval Within 4
0	25	> 25,000	4,194,927	0.0	-7.3	7.6	28.5	0.0	0.0	0.0	10.0
0	50	> 25,000	8,219,220	0.0	-5.5	4.3	44.5	0.0	0.0	17.0	57.0
0	100	> 25,000	16,783,359	0.0	-4.3	3.7	53.5	0.0	6.0	56.5	90.5
0	200	> 25,000	33,890,955	0.0	-2.3	2.7	75.5	0.0	53.5	90.0	99.5
2	25	> 25,000	4,194,927	2.0	-5.2	9.7	26.5	0.0	0.0	0.0	8.5
2	50	> 25,000	8,219,220	2.0	-3.6	6.4	43.5	6.0	0.0	19.5	56.5
2	100	> 25,000	16,783,359	2.0	-2.4	5.6	52.5	31.0	5.5	56.5	90.0
2	200	> 25,000	33,890,955	2.0	-0.4	4.6	75.5	65.0	54.5	90.0	100.0
5	25	> 25,000	4,194,927	4.9	-1.9	12.9	28.0	28.0	0.0	0.0	7.5
5	50	> 25,000	8,219,220	4.9	-0.8	9.6	43.5	43.5	0.0	19.0	56.0
5	100	> 25,000	16,783,359	4.9	0.4	8.5	55.0	55.0	7.0	57.5	90.0
5	200	> 25,000	33,890,955	4.9	2.4	7.5	77.0	77.0	50.5	89.5	100.0
10	25	> 25,000	4,194,927	9.7	3.5	18.1	28.0	28.0	0.0	0.0	8.0
10	50	> 25,000	8,219,220	9.7	3.9	14.9	40.5	40.5	0.0	16.5	56.0
10	100	> 25,000	16,783,359	9.7	5.0	13.4	56.0	56.0	7.5	60.5	89.0
10	200	> 25,000	33,890,955	9.6	7.1	12.2	71.0	71.0	47.0	89.5	99.0

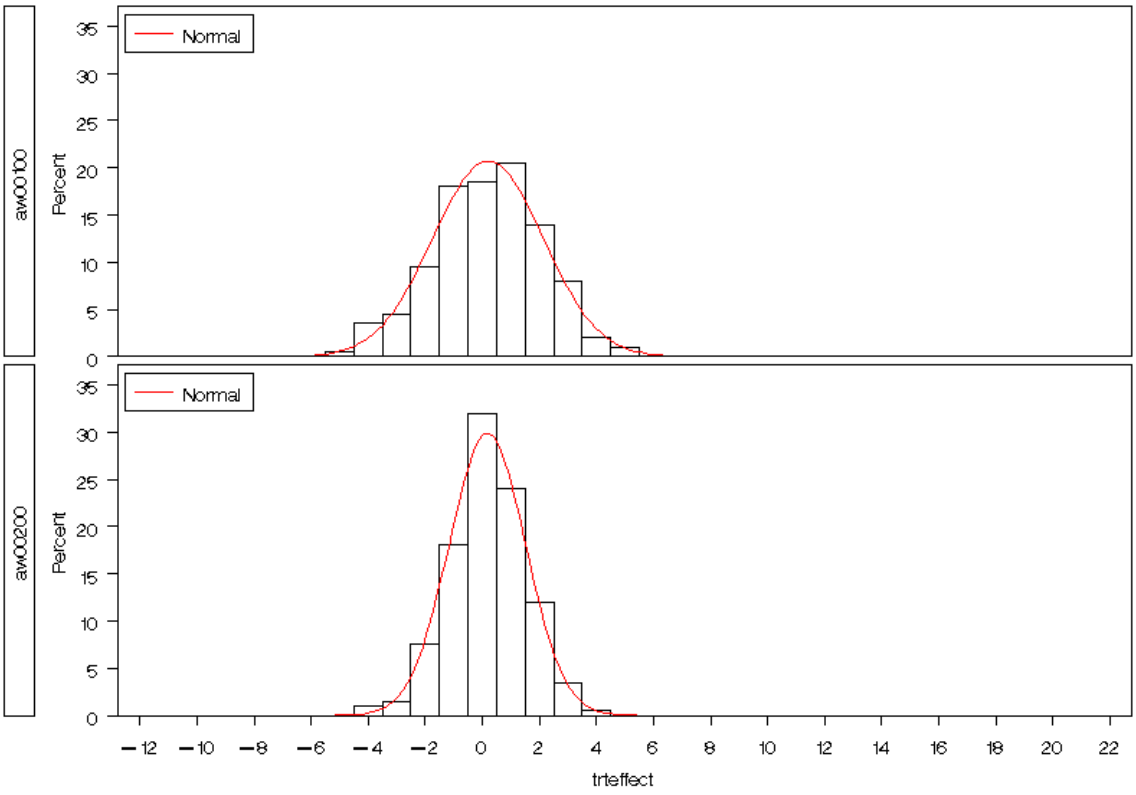
SUMMARY OF COUNTY LEVEL SIMULATIONS

Actual Effect Size	Number of Treatment Counties	Population for Inclusion	Treatment County Population	Estimated Effect	Minimum Treatment Effect	Maximum Treatment Effect	Est Trtmt Within 1 of Actual Trtmt	% Accurate and Significant	Confidence Interval Within 2	Confidence Interval Within 3	Confidence Interval Within 4
0	25	> 50,000	6,536,338	0.2	-3.9	6.0	35.5	0.0	0.0	0.0	31.0
0	50	> 50,000	12,776,948	0.2	-3.0	4.0	57.5	0.0	0.0	44.0	82.0
0	100	> 50,000	26,136,277	0.2	-2.9	3.5	68.0	0.0	28.0	83.0	98.5
0	200	> 50,000	52,280,207	0.0	-2.4	1.9	77.0	0.0	63.0	97.0	100.0
2	25	> 50,000	6,536,338	2.1	-2.2	7.8	38.0	0.0	0.0	0.0	33.0
2	50	> 50,000	12,776,948	2.1	-1.1	5.9	58.0	24.0	0.0	43.5	82.5
2	100	> 50,000	26,136,277	2.1	-1.0	5.5	68.0	48.5	26.5	83.5	98.5
2	200	> 50,000	52,280,207	2.0	-0.4	3.9	79.0	71.0	62.0	97.5	100.0
5	25	> 50,000	6,536,338	5.0	0.4	10.4	36.0	36.0	0.0	0.0	33.0
5	50	> 50,000	12,776,948	5.0	1.7	8.8	58.0	58.0	0.0	43.5	82.0
5	100	> 50,000	26,136,277	5.0	1.9	8.3	71.0	71.0	26.0	83.0	99.0
5	200	> 50,000	52,280,207	4.8	2.4	6.8	78.0	78.0	61.0	98.0	100.0
10	25	> 50,000	6,536,338	9.7	4.6	14.8	38.0	38.0	0.0	0.0	35.0
10	50	> 50,000	12,776,948	9.8	6.4	13.7	56.0	56.0	0.0	43.0	81.0
10	100	> 50,000	26,136,277	9.8	6.7	13.1	67.5	67.5	23.0	80.0	97.0
10	200	> 50,000	52,280,207	9.6	7.1	11.6	70.5	70.5	51.0	94.0	99.5

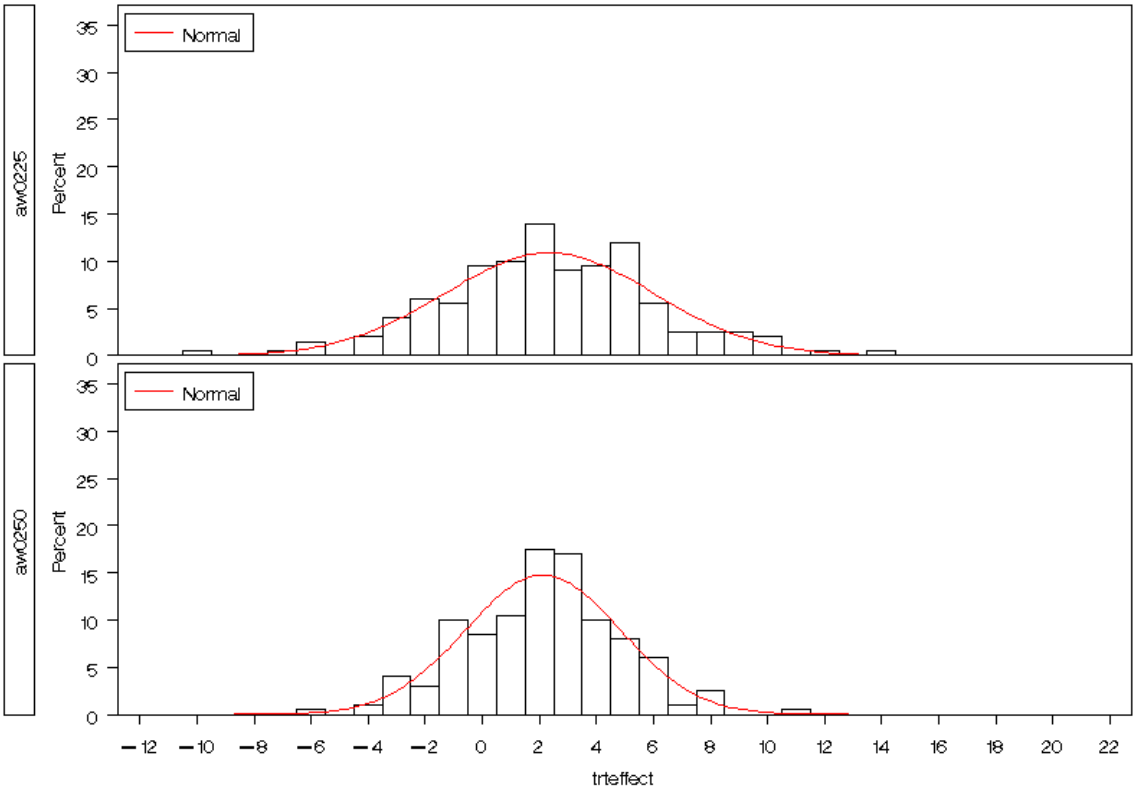
DISTRIBUTION OF TREATMENT EFFECTS



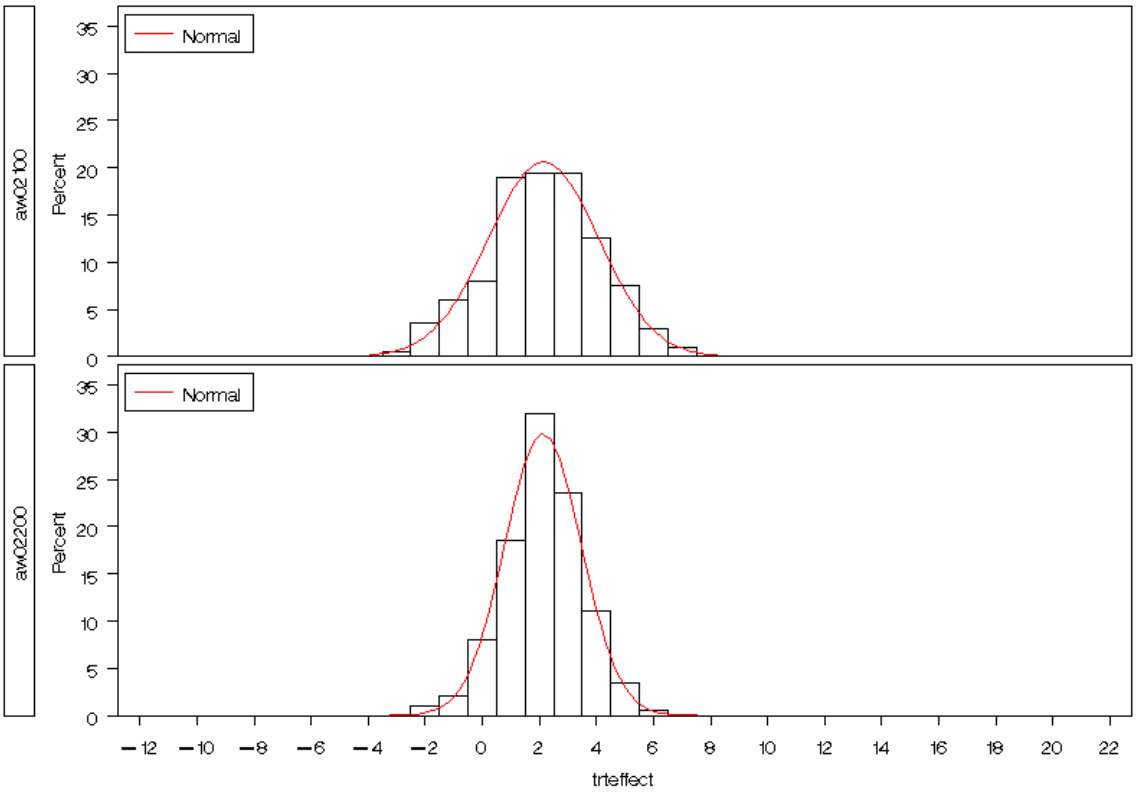
DISTRIBUTION OF TREATMENT EFFECTS



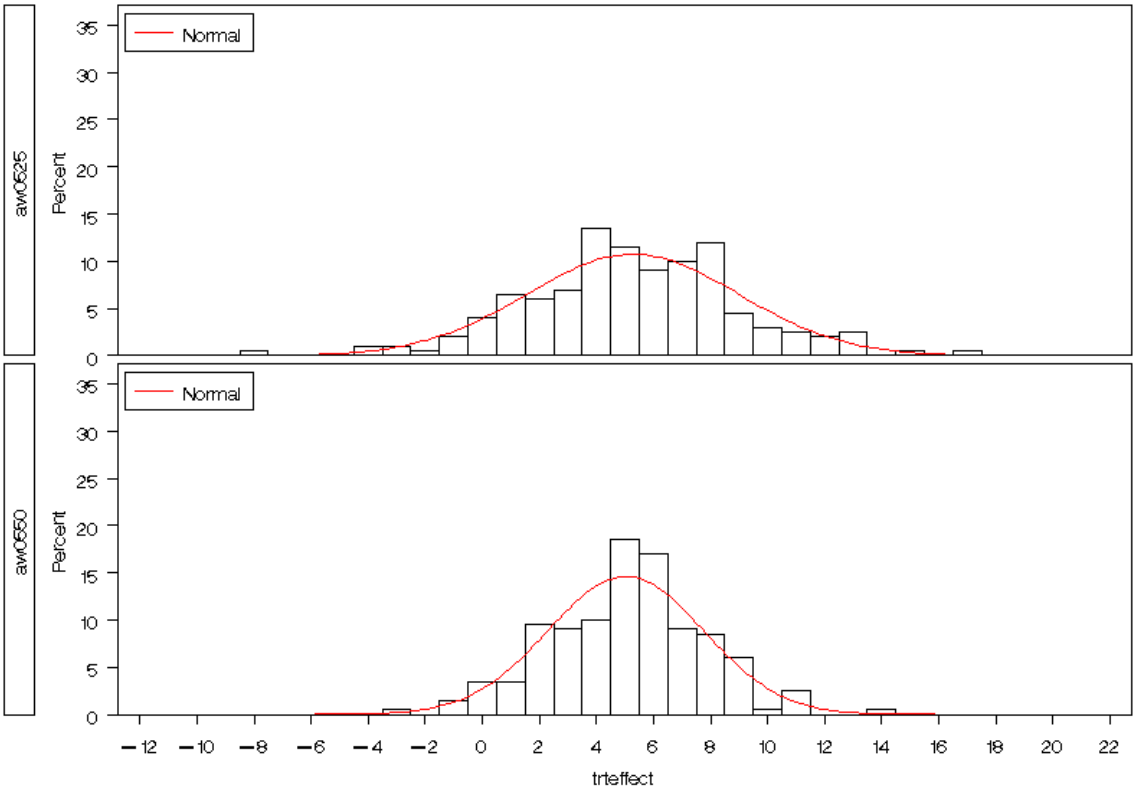
DISTRIBUTION OF TREATMENT EFFECTS



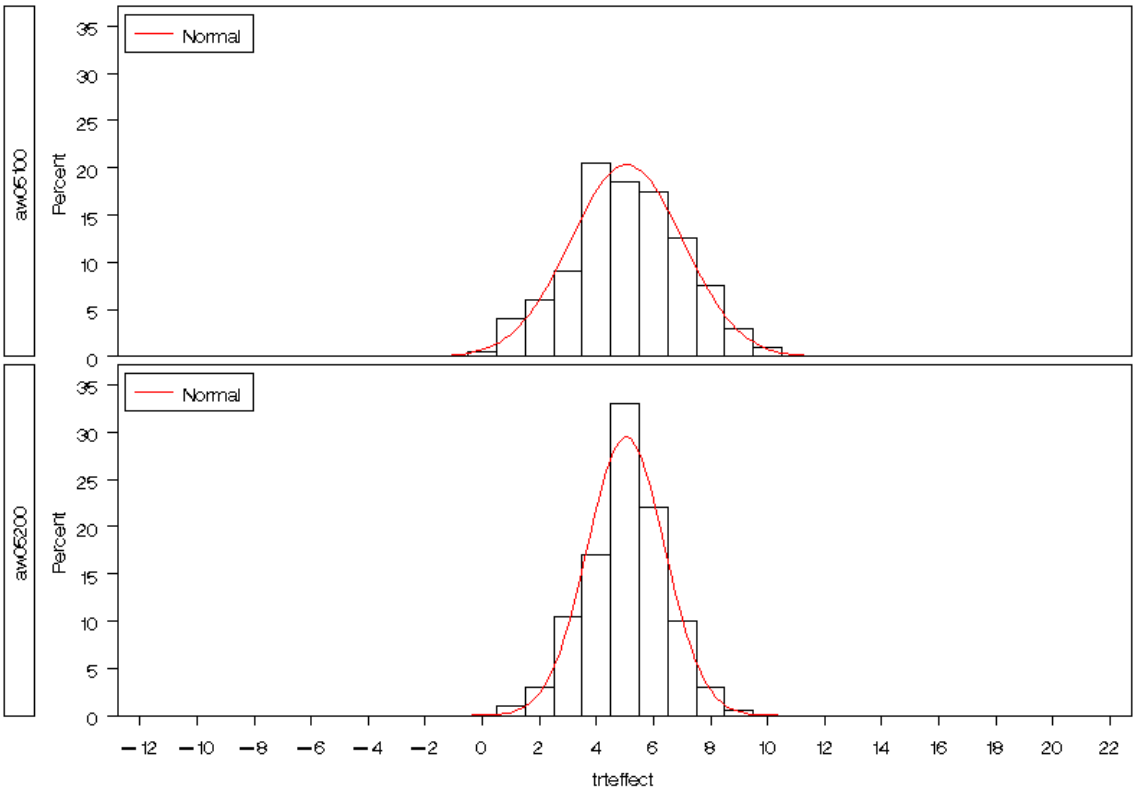
DISTRIBUTION OF TREATMENT EFFECTS



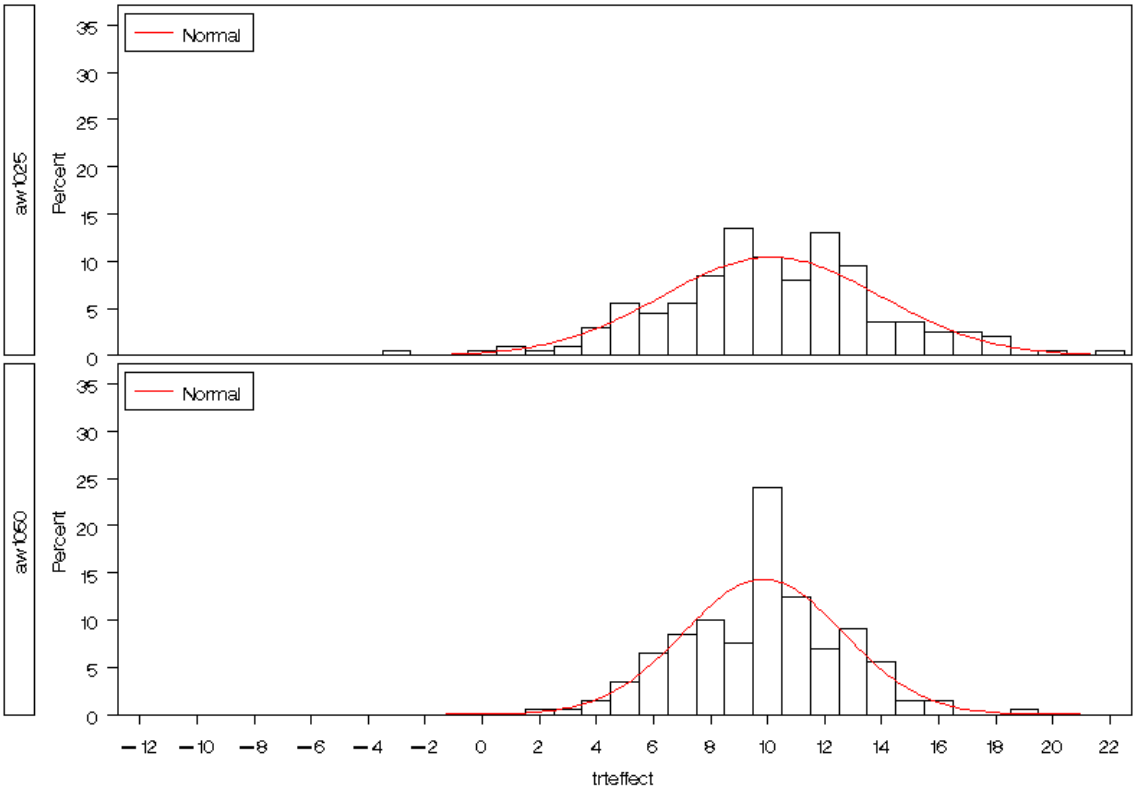
DISTRIBUTION OF TREATMENT EFFECTS



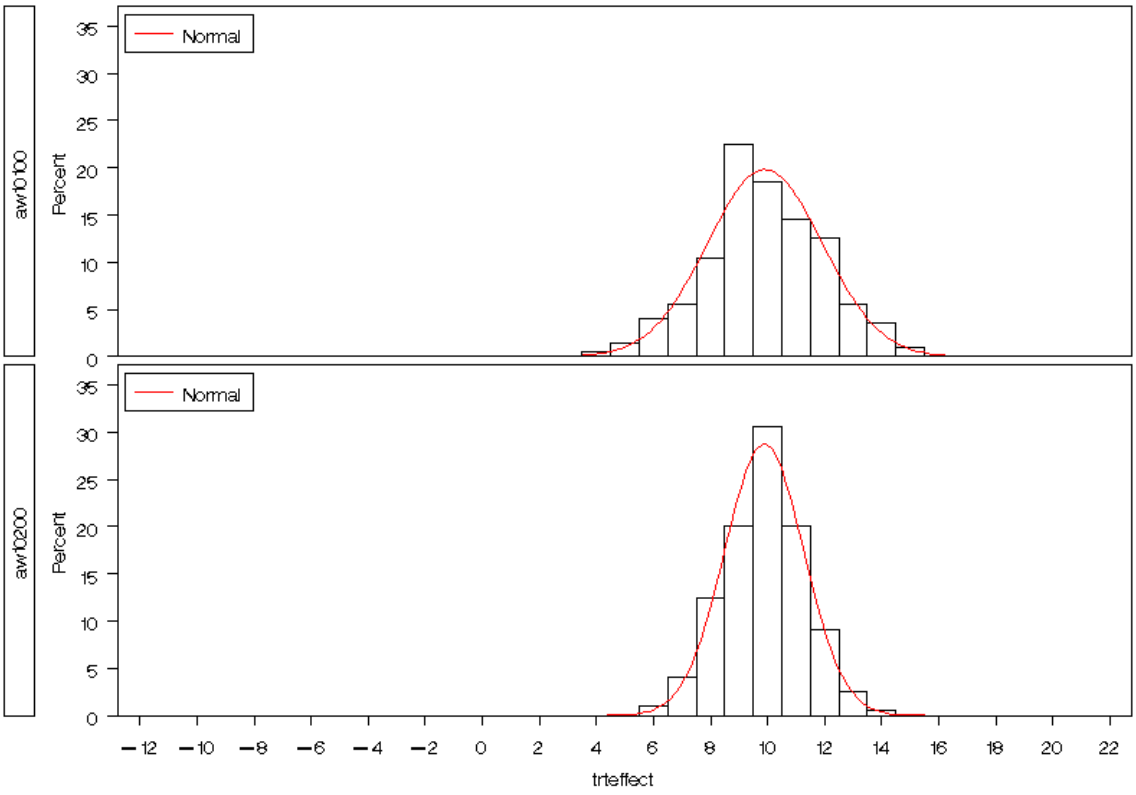
DISTRIBUTION OF TREATMENT EFFECTS



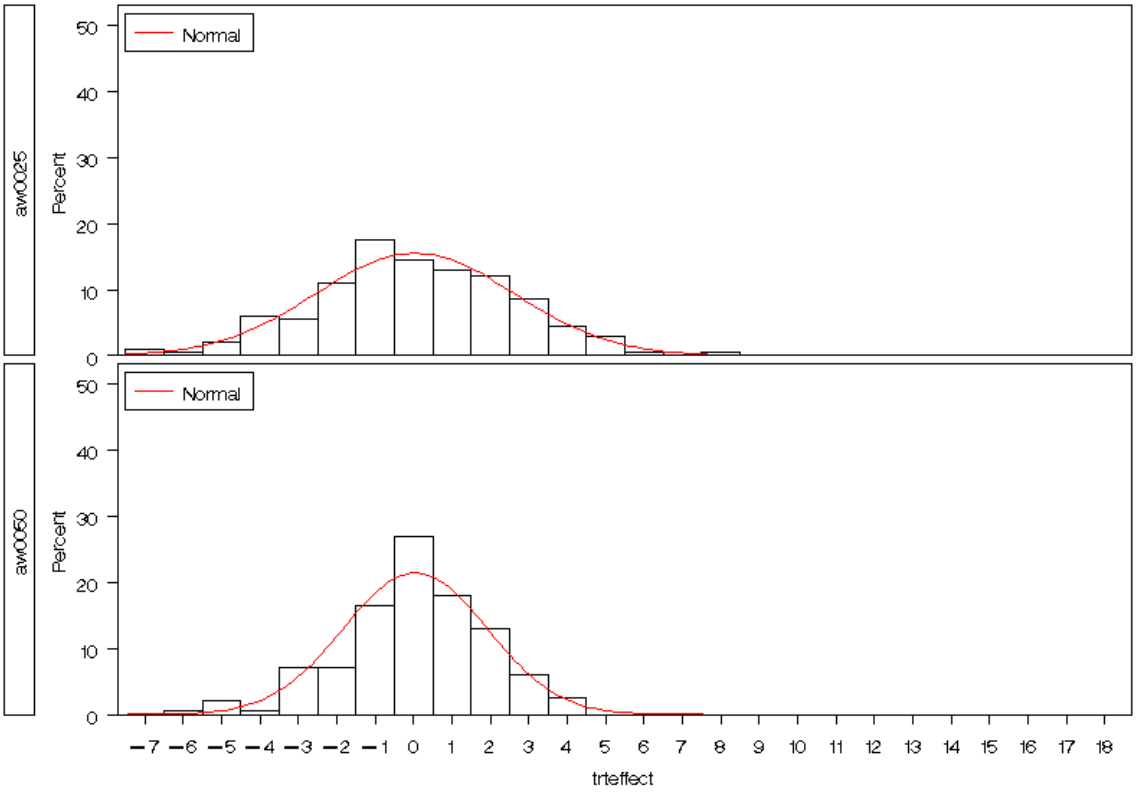
DISTRIBUTION OF TREATMENT EFFECTS



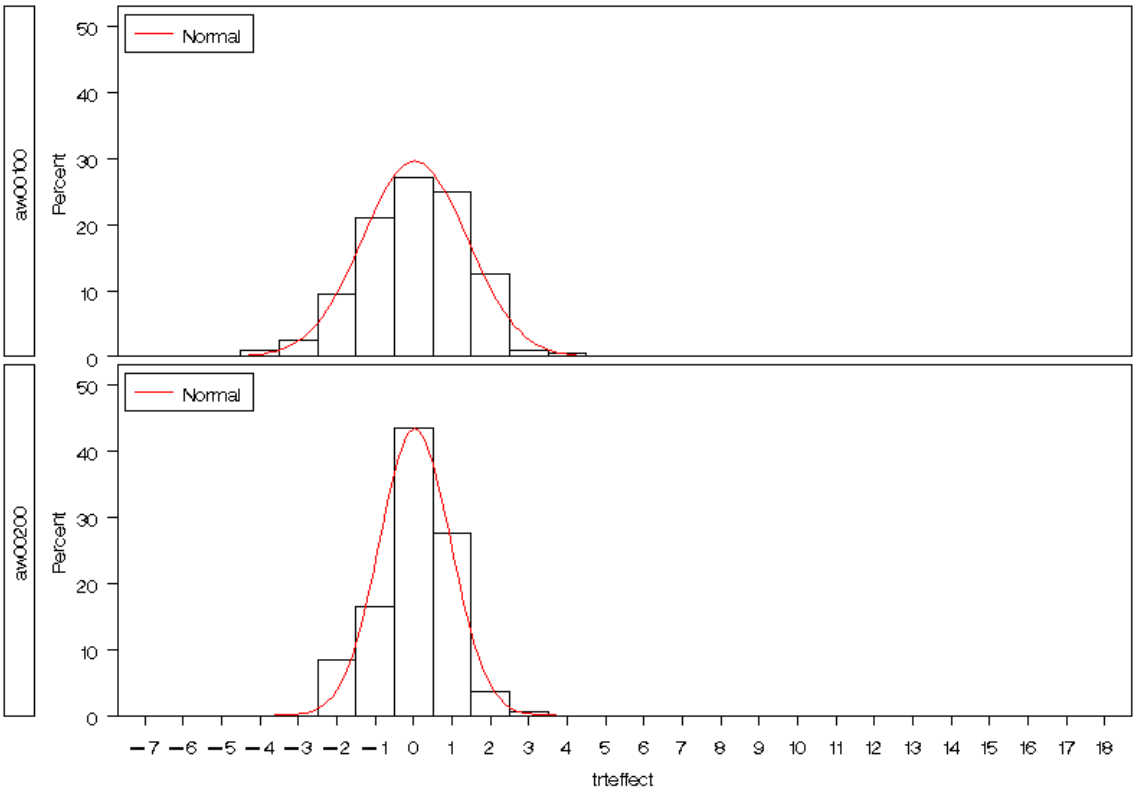
DISTRIBUTION OF TREATMENT EFFECTS



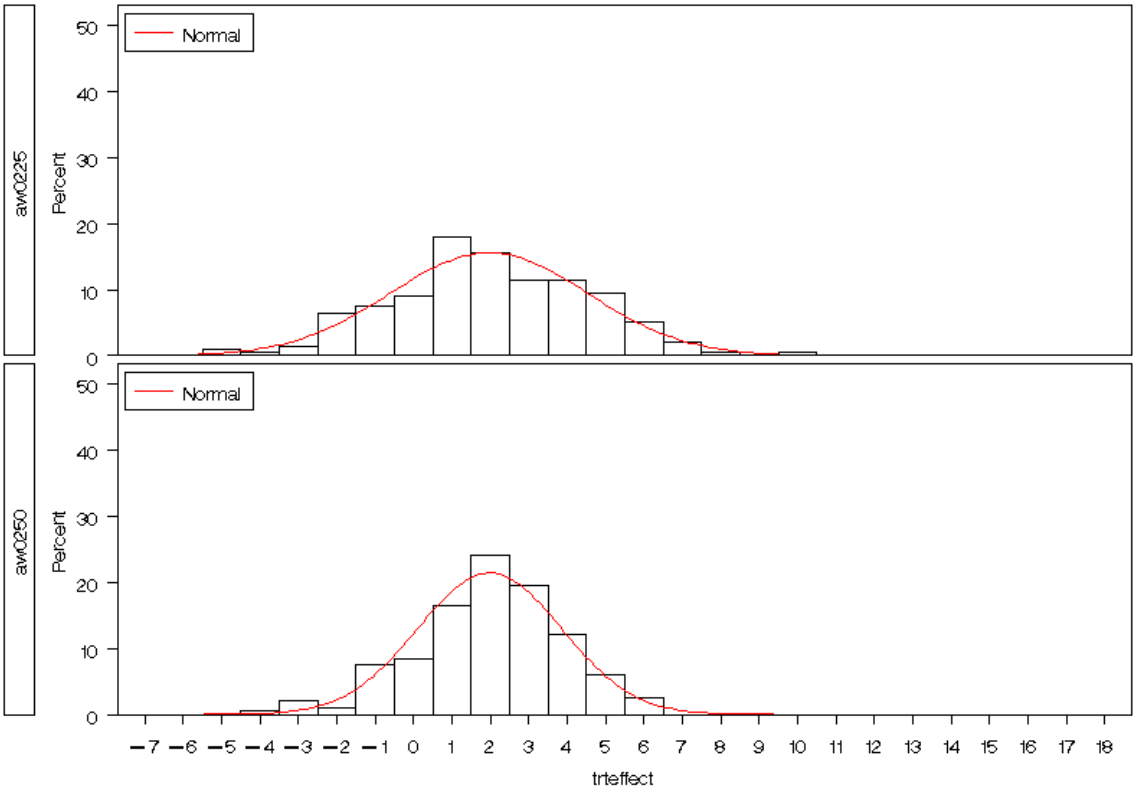
DISTRIBUTION OF TREATMENT EFFECTS



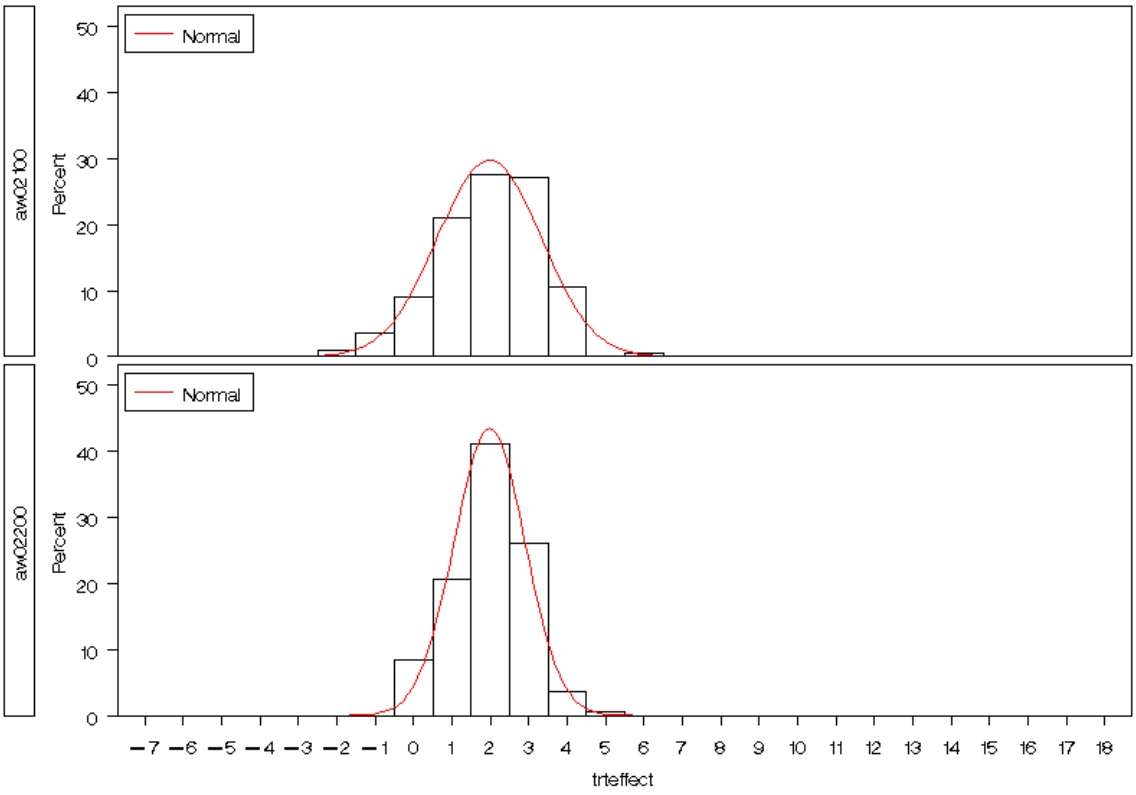
DISTRIBUTION OF TREATMENT EFFECTS



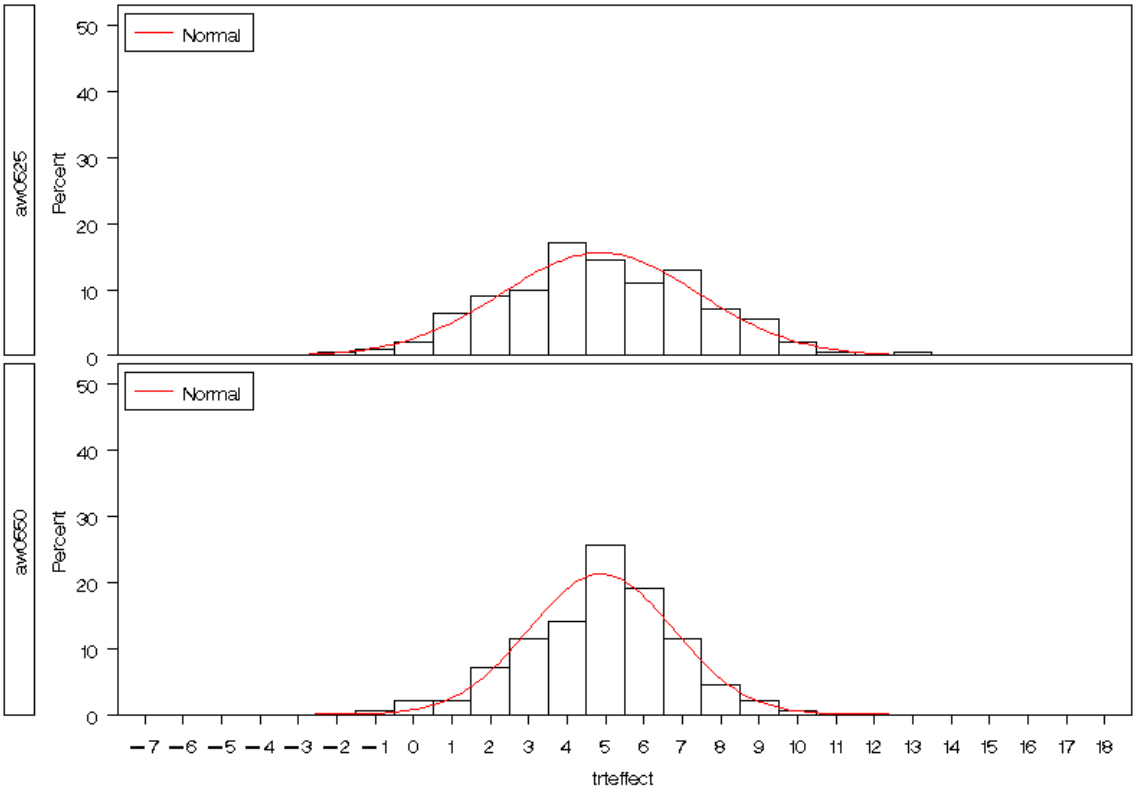
DISTRIBUTION OF TREATMENT EFFECTS



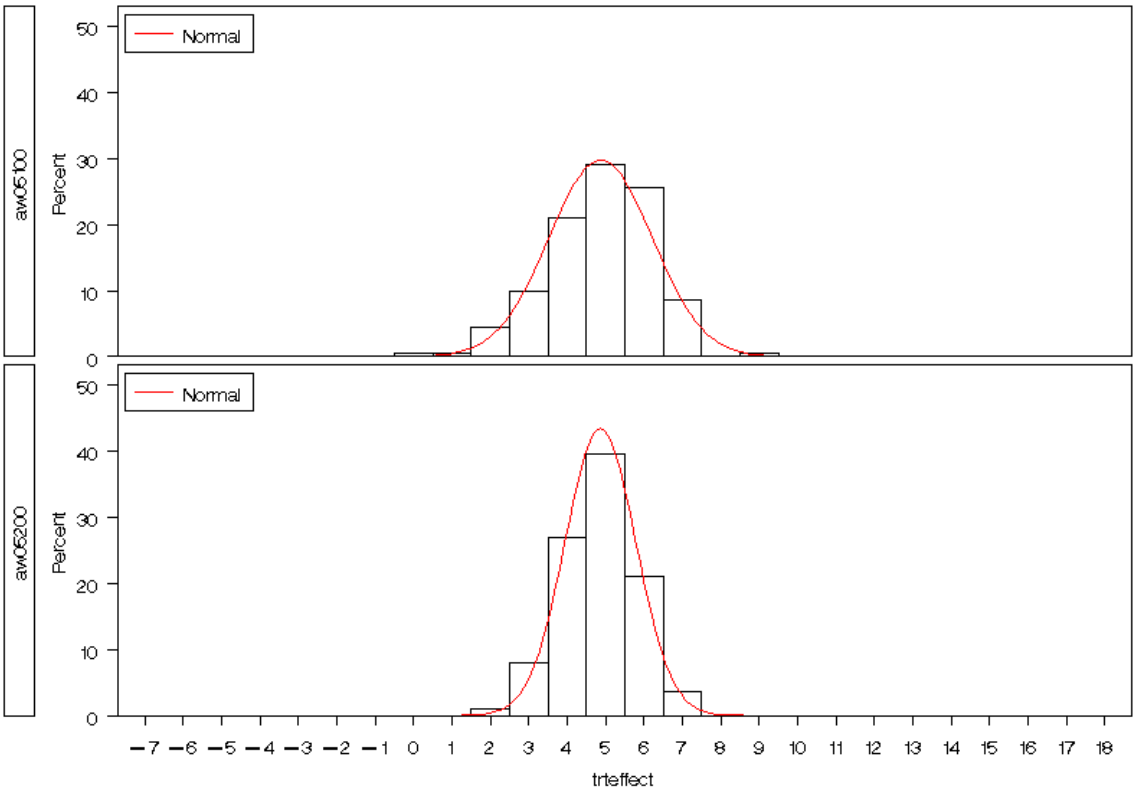
DISTRIBUTION OF TREATMENT EFFECTS

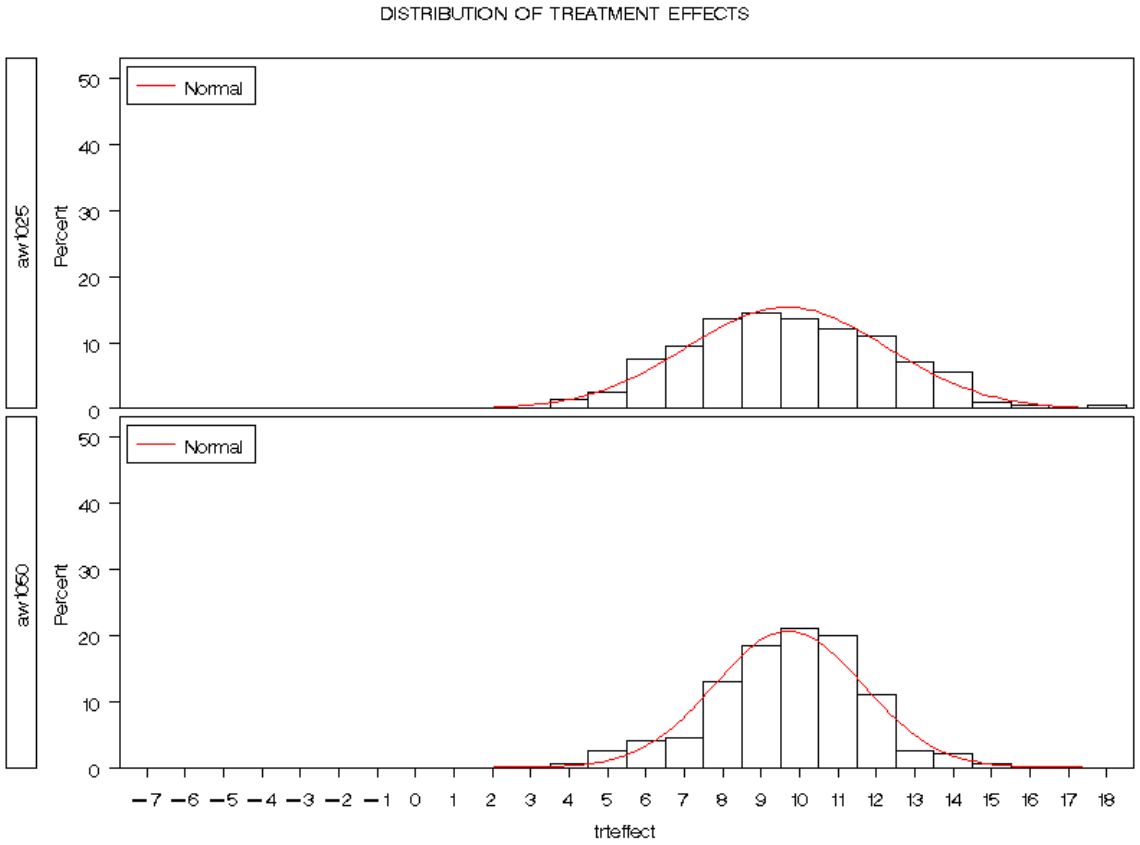


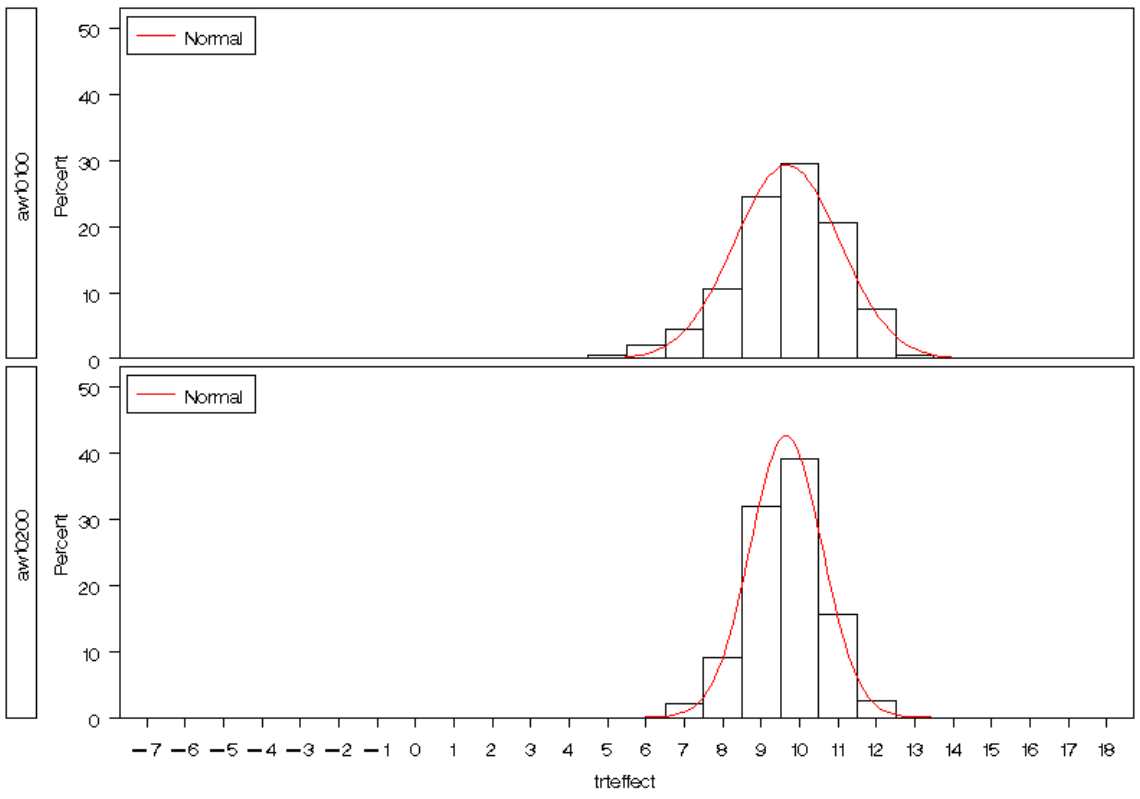
DISTRIBUTION OF TREATMENT EFFECTS

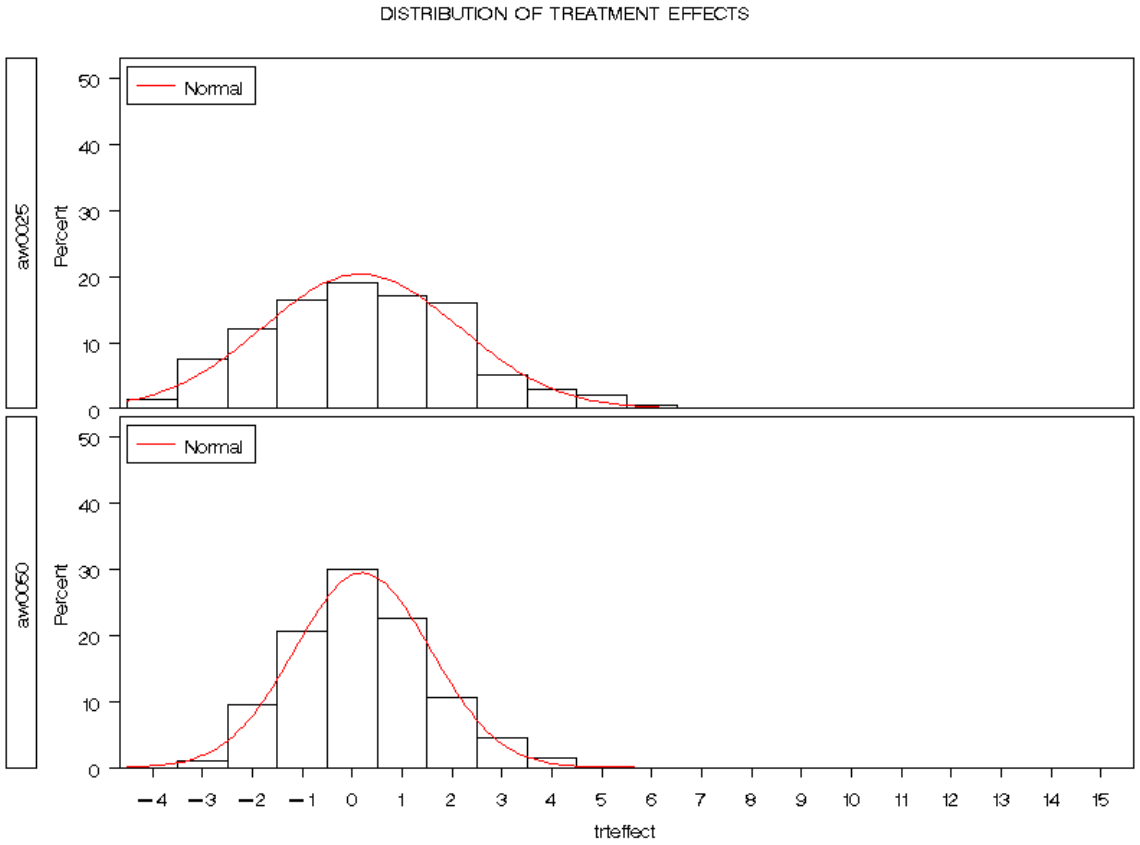


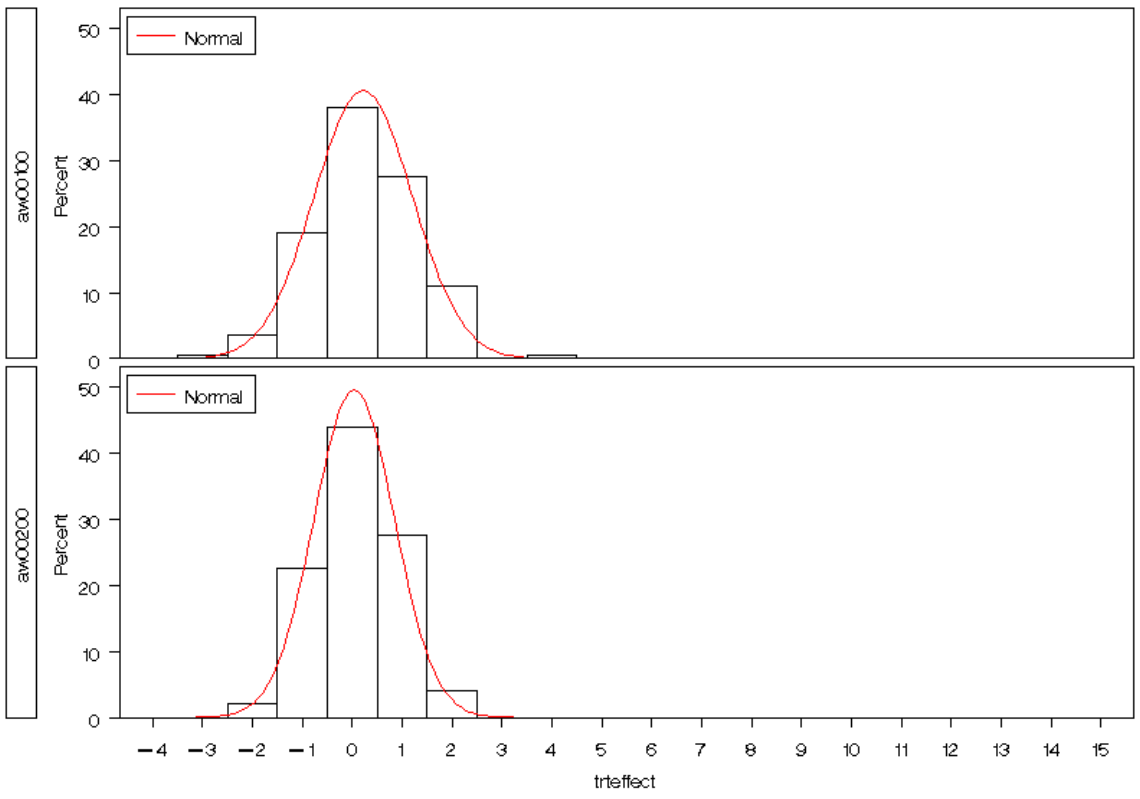
DISTRIBUTION OF TREATMENT EFFECTS



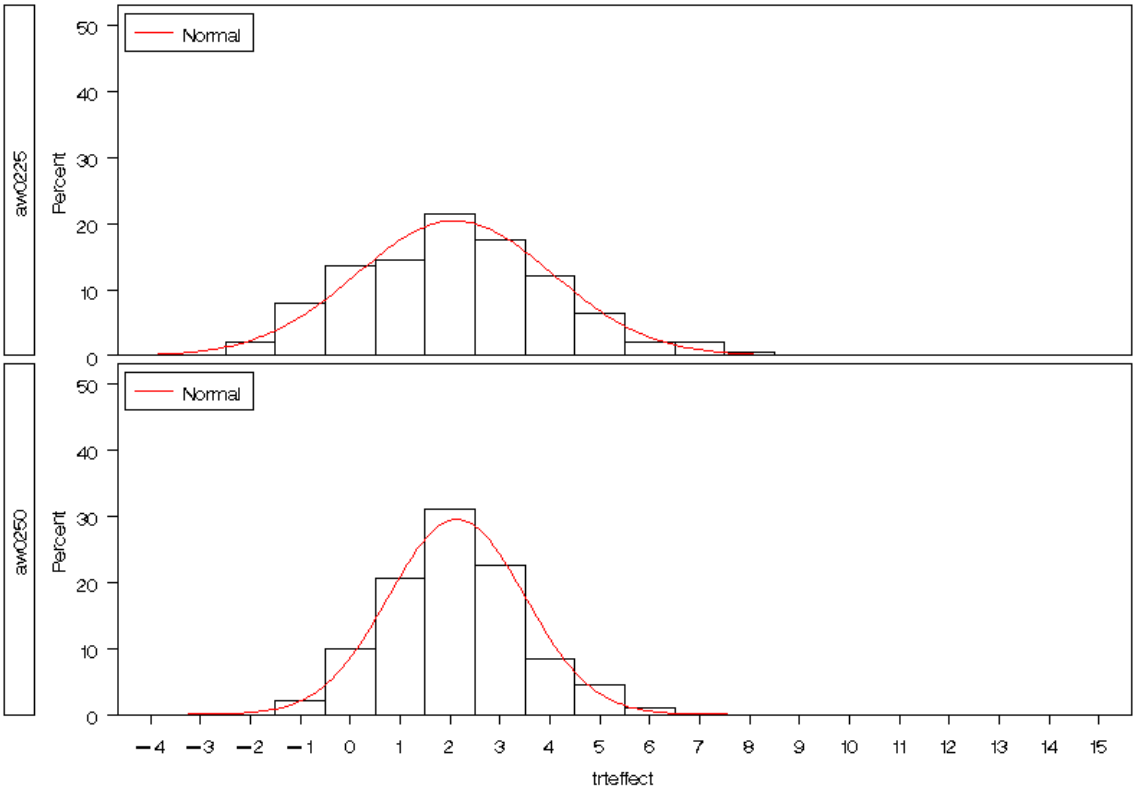

DISTRIBUTION OF TREATMENT EFFECTS
Normal
Percent
awr1025
awr1050
trteffect


DISTRIBUTION OF TREATMENT EFFECTS
Normal
Percent
awr0100
awr0200
trteffect

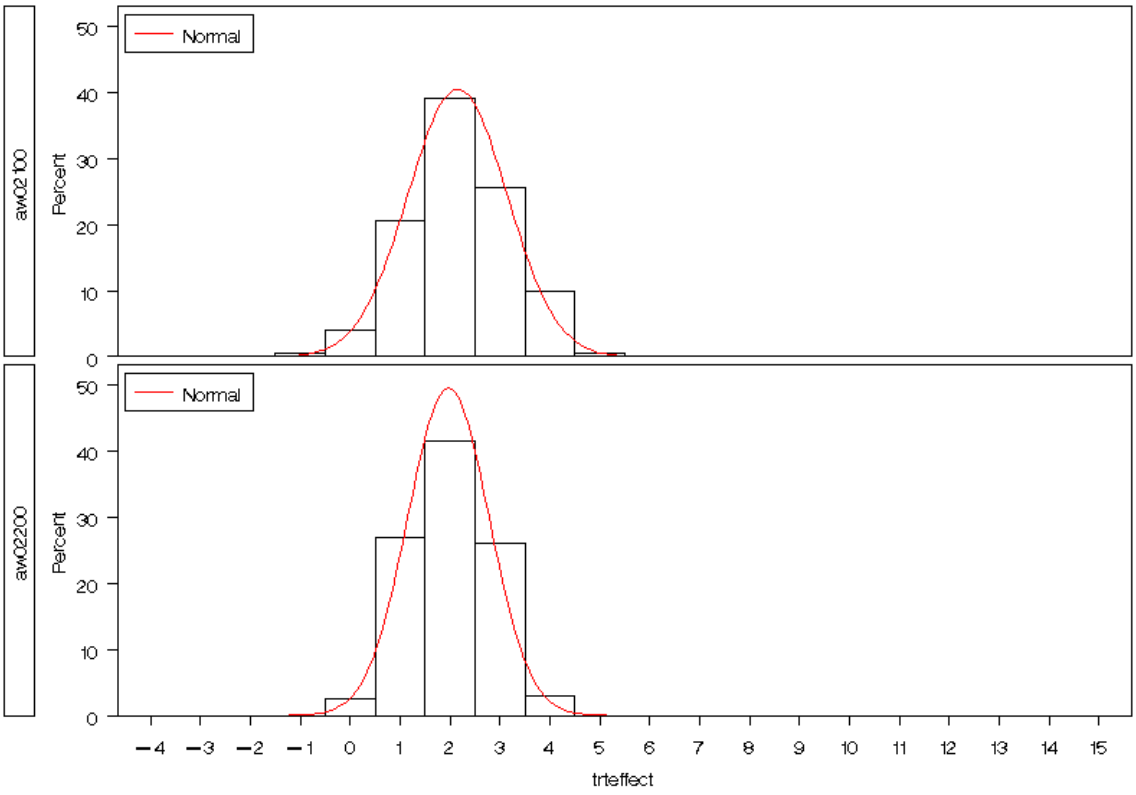

DISTRIBUTION OF TREATMENT EFFECTS
aw0025
aw0050
Normal
Percent
trteffect


DISTRIBUTION OF TREATMENT EFFECTS
aw0100
aw0200
Normal
Percent
trteffect

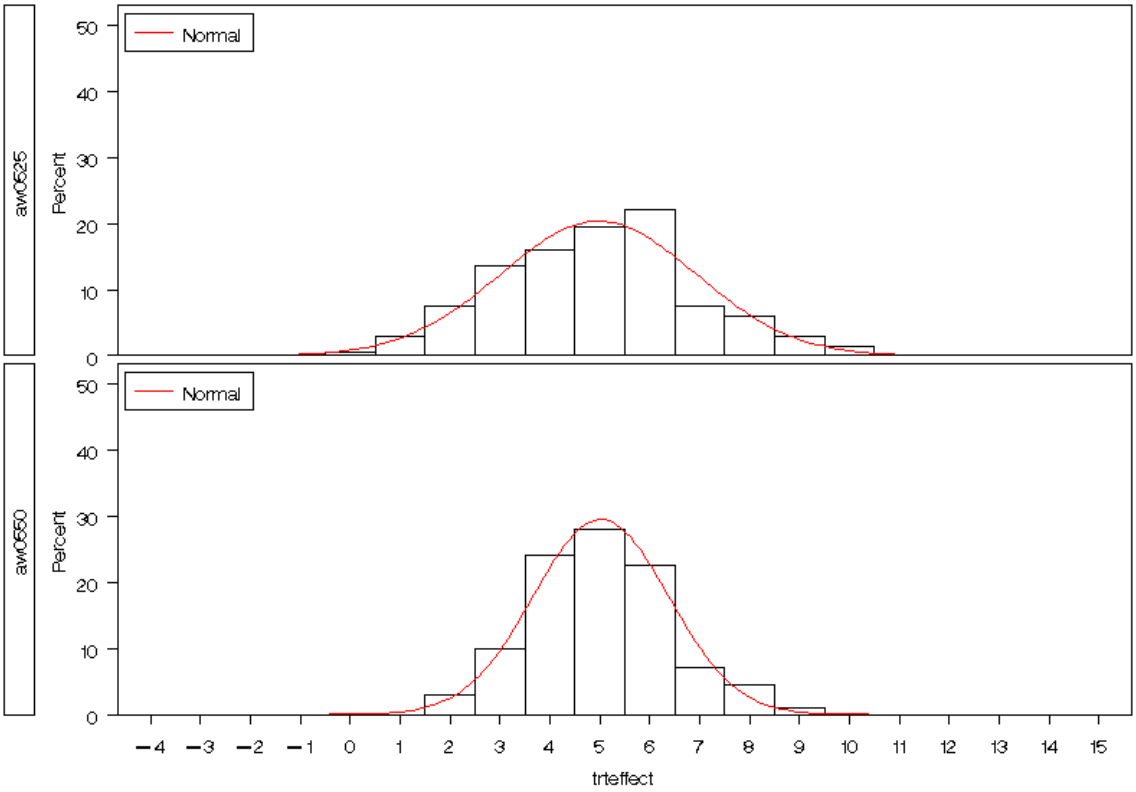
DISTRIBUTION OF TREATMENT EFFECTS



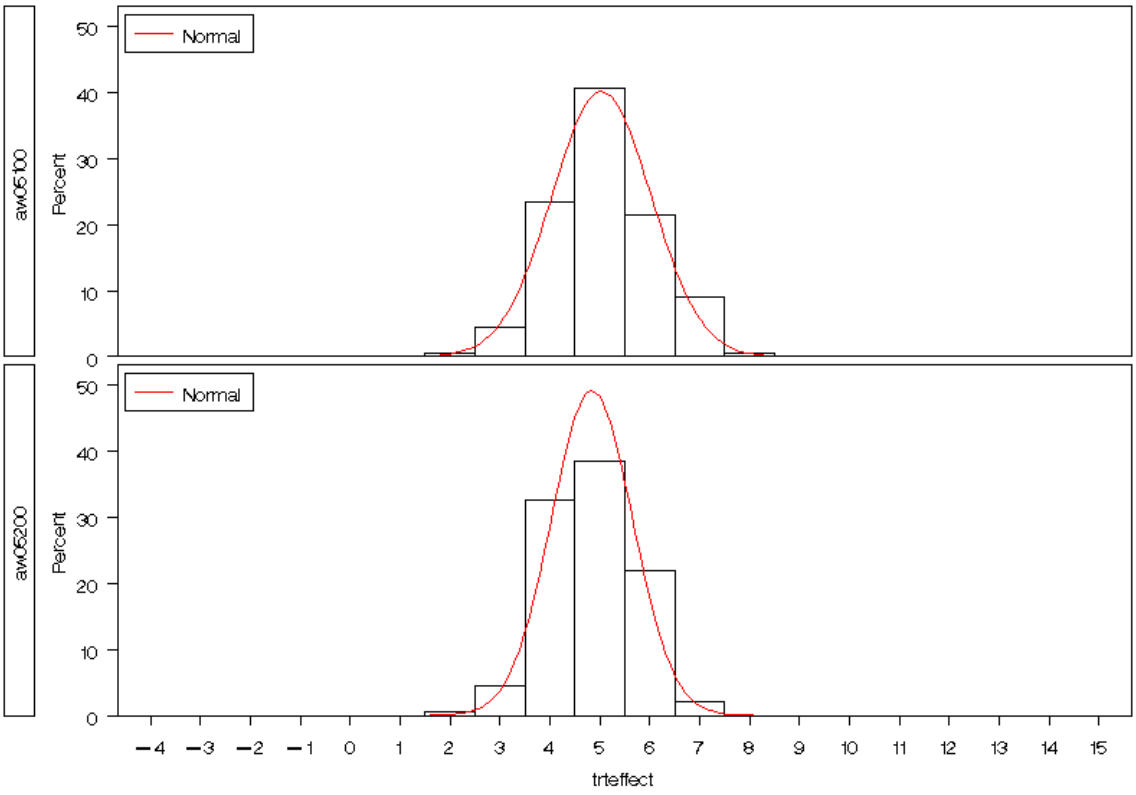
DISTRIBUTION OF TREATMENT EFFECTS

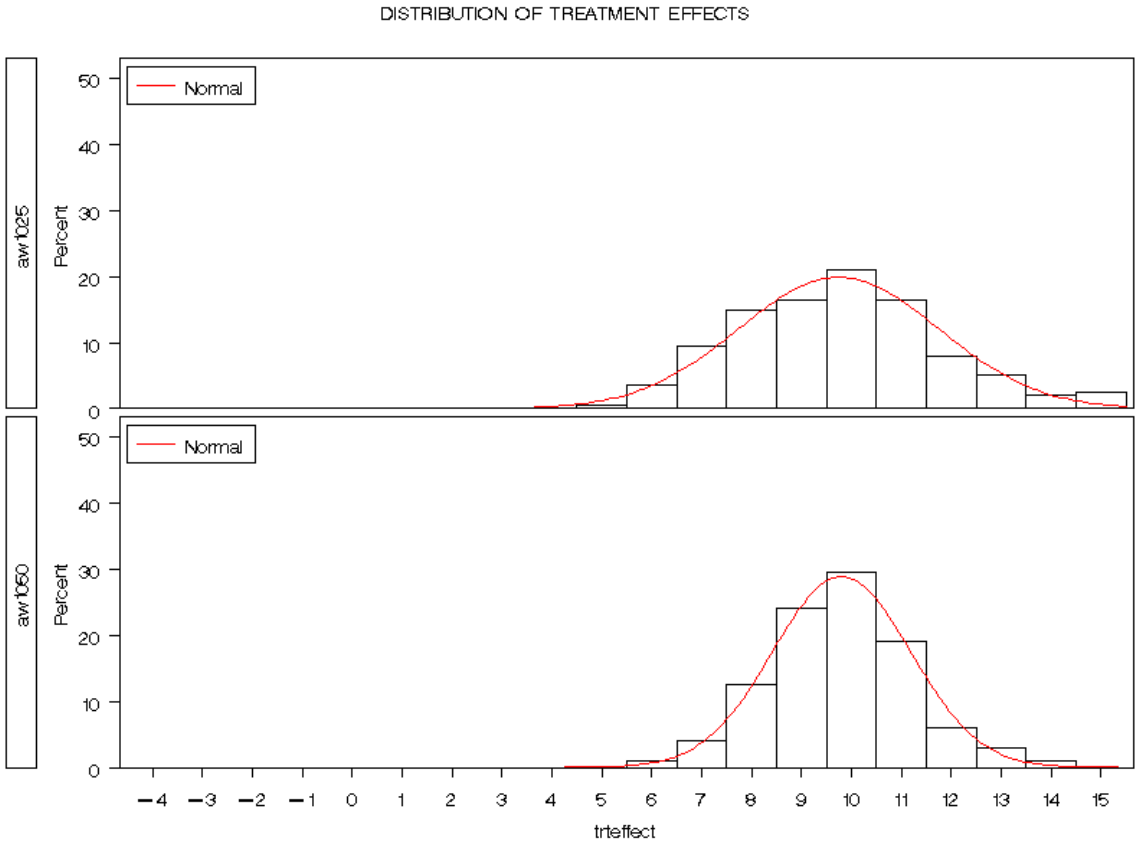


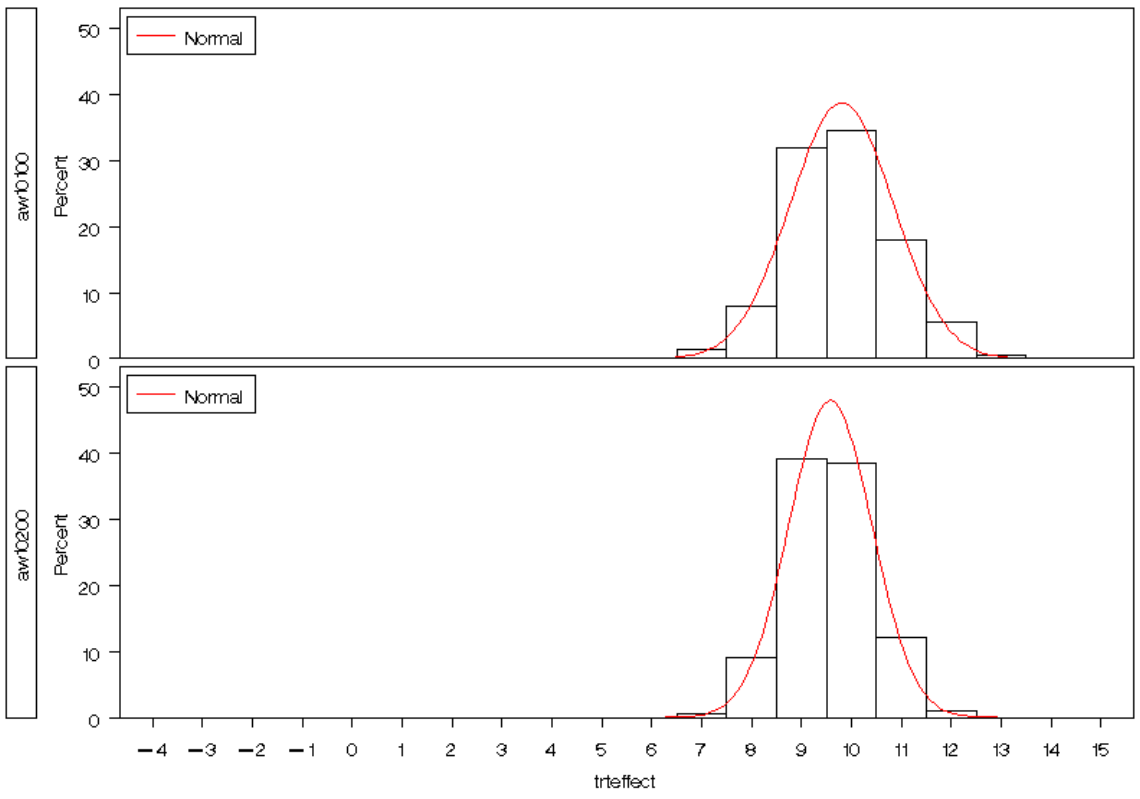
DISTRIBUTION OF TREATMENT EFFECTS



DISTRIBUTION OF TREATMENT EFFECTS




DISTRIBUTION OF TREATMENT EFFECTS
awr1025
awr1050
Normal
Percent
trteffect


DISTRIBUTION OF TREATMENT EFFECTS
awr0100
awr0200
Normal
Percent
trteffect