*NOTES FOR SUMMER STATISTICS INSTITUTE COURSE*

**COMMON MISTAKES IN STATISTICS –
SPOTTING THEM AND AVOIDING THEM**

**Day 4: Common Mistakes Based on Common
Misunderstandings about Statistical Inference**

MAY 23 – 26, 2011

Instructor: Martha K. Smith

**CONTENTS OF DAY 4**

## I. THE FILE DRAWER PROBLEM (PUBLICATION BIAS)

***Publication bias*** refers to the influence of the results of a study on whether or not the study is published.

Examples of how results might influence the publication decision:
- Whether or not the results are statistically significant.
- Whether or not the results are practically significant.
- Whether or not the results agree with the hopes or expectations of the researcher or sponsor.

Publication bias is also called the ***file drawer problem***, especially when the nature of the bias is that studies which fail to reject the null hypothesis (i.e., that do not produce a statistically significant result) are less likely to be published than those that do produce a statistically significant result.

Several studies (see Sterling et al 1995, Song et al 2009, and Hopewell et al 2009) have found evidence of publication bias in the research literature.

*Consequences of the File Drawer Problem:*

1. Investigators may spend unnecessary effort conducting research on topics that have already been well researched but not reported because results were negative.

2. Effects that are not real may appear to be supported by research.

- Recall that if a significance level of 0.05 is set, then in repeated studies, about 5% of studies of a situation where the null hypothesis is true will falsely reject the null hypothesis
- Thus, *if just (or even predominantly) the statistically significant studies are published, the published record misrepresents the true situation.*

3. Furthermore, papers that are published because of Type I errors, if underpowered, may show an exaggerated effect size, increasing the misrepresentation.
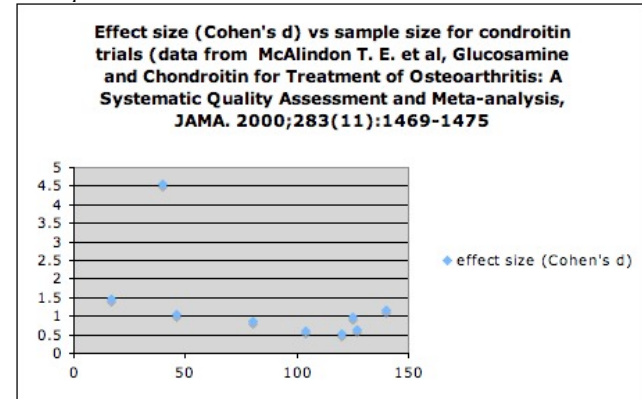- *Recall from the discussion of power*:
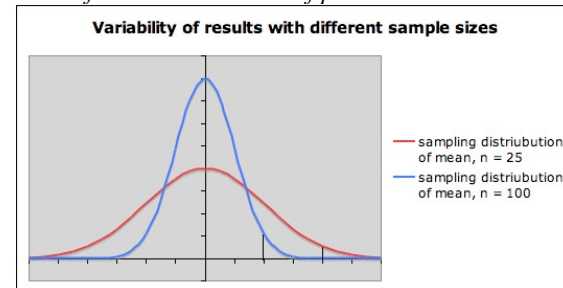
***Some Methods Proposed to Detect Publication Bias:***

1. Rosenthal (1979) proposed a method, based on probability calculations, for deciding whether or not a finding is "resistant to the file drawer threat."

- This method has become known as the *fail-safe file drawer (or FSFD) analysis.*

- It involves calculating a "fail-safe number" which is used to estimate whether or not the file-drawer problem is likely to be a problem for a particular review or meta-analysis.

- Scargle (2000) has criticized Rosenthal's method on the grounds that it fails to take into account the bias in the "file drawer" of unpublished studies, and thus can give misleading results.

2. Various types of plots have been used to try to detect publication bias. These plot some measure of precision against effect size, or vice-versa.

- *Example*:



Effect size (Cohen's d) vs sample size for condroitin trials (data from McAlindon T. E. et al, Glucosamine and Chondroitin for Treatment of Osteoarthritis: A Systematic Quality Assessment and Meta-analysis, JAMA. 2000;283(11):1469-1475

- Some such plots are called "funnel plots" because they typically have a funnel shape.

- However, Lau et al (2006) point out some problems in using these plots.

3. Research registries have been instituted in some areas.

  o For example, certain clinical trials are now required by law to be registered ClinicalTrials.gov.

*See the Appendix for suggestions for dealing with the File Drawer Problem.*

## II. MULTIPLE INFERENCE

*"Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present. Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading."*
    Professionalism Guideline 8, *Ethical Guidelines for Statistical Practice*, American Statistical Association, 1997

Performing more than one statistical inference procedure on the same data set is called ***multiple inference***, or ***joint inference***, or ***simultaneous inference***, or ***multiple testing***, or ***multiple comparisons***, or ***the problem of multiplicity***.

Performing multiple inference *without adjusting the Type I error rate accordingly* is a **common error** in research using statistics.
- For example, A. M. Strasak et al (2007) examined all papers from 2004 issues of the *New England Journal of Medicine* and *Nature Medicine* and found that 32.3% of those from *NEJM* and 27.3% from *Nature Medicine* were "Missing discussion of the problem of multiple significance testing if occurred."
- These two journals are considered the top journals (according to impact figure) in clinical science and in research and experimental medicine, respectively.

### The Problem

*Recall*: If you perform a hypothesis test using a certain significance level (we'll use 0.05 for illustration), and if you obtain a p-value less than 0.05, then there are three possibilities:

1. The model assumptions for the hypothesis test are not satisfied in the context of your data.
2. The null hypothesis is false.
3. Your sample happens to be one of the 5% of samples satisfying the appropriate model conditions for which the hypothesis test gives you a Type I error – i.e., you falsely reject the null hypothesis.

Now suppose you are performing *two* hypothesis tests, *using the same data* for both.

- Suppose that in fact all model assumptions are satisfied and *both* null hypotheses are true.

- *There is in general no reason to believe that the samples giving a Type I error for one test will also give a Type I error for the other test.*

- See Jerry Dallal's Simulation

- So we need to consider the *joint Type I error rate*

***Joint Type I error rate***: This is the probability that a randomly chosen sample (of the given size, satisfying the appropriate model assumptions) will give a Type I error for *at least one* of the hypothesis tests performed.

The joint Type I error rate is also known as the ***overall Type I error rate***, or ***joint significance level***, or the ***simultaneous Type I error rate***, or the ***family-wise error rate*** (***FWER***), or the ***experiment-wise error rate***, etc.
  • The acronym FWER is becoming more and more common, so will be used in the sequel, often along with another name for the concept as well.

An *especially serious* form of neglect of the problem of multiple inference is the one alluded to in the quote from the ASA ethics page:
  • Trying several tests and reporting just one significant test, without disclosing how many tests were performed or correcting the significance level to take into account the multiple inference.
  • *Don't do it!*

***Multiple inference with confidence intervals***

The problem of multiple inference also occurs for confidence intervals.
  • In this case, we need to focus on the *confidence level*.
  • *Recall*: A 95% confidence interval is an interval obtained by using a procedure that, for 95% of all suitably random samples, of the given size from the random variable and population of interest, produces an interval containing the parameter we are estimating (assuming the model assumptions are satisfied).
  • In other words, the procedure does what we want (i.e. gives an interval containing the true value of the parameter) for 95% of suitable samples.
  • *If we are using confidence intervals to estimate two parameters, there is no reason to believe that the 95% of samples for which the procedure "works" for one parameter* (i.e. gives an interval containing the true value of the parameter) *will be the same as the 95% of samples for which the procedure "works" for the other parameter*.
  • If we are calculating confidence intervals for more than one parameter, we can talk about the ***joint (****or **overall** or **simultaneous** or **family-wise** or **experiment-wise)** **confidence level**.
  • For example, a group of confidence intervals (for different parameters) has an ***overall 95% confidence level*** (or ***95% family-wise confidence level***, etc.) if the intervals are calculated using a procedure which, for 95% of all suitably random samples, of the given size from the population of interest, produces for *each* parameter an interval containing that parameter (assuming the model assumptions are satisfied).

*What to do about it*

Unfortunately, there is no simple formula to cover all cases:
- Depending on the context, the samples giving Type I errors for two tests might be the same, they might have no overlap, or they could be somewhere in between.
- Various techniques for bounding the FWER (joint Type I error rate) have been devised for various special circumstances.
- Only two fairly general methods (Bonferroni and False Discovery Rate) will be discussed here.
- For more information on other more specialized methods, see, e.g., Hochberg and Tamhane (1987) and Miller (1981)

*Bonferroni method*:

Fairly basic probability calculations show that *if the <u>sum</u> of the Type I error rates for different tests is less than α, then the overall ("family-wise") Type I error rate (FWER) for the combined tests will be at most α.*

- So, for example, if you are performing five hypothesis tests and would like an FWER (overall significance level) of at most 0.05, then using significance level 0.01 for *each* test will give an FWER (overall significance level) of at most 0.05.
- Similarly, if you are finding confidence intervals for five parameters and want an overall confidence level of 95%, using the 99% confidence level for each confidence interval will give you overall confidence level at least 95%. (Think of confidence level as 1 - $\alpha$.)

The Bonferroni method can be a used as a fallback method when no other method is known to apply.

- However, if a method that applies to the specific situation is available, it will often be better (less conservative) than the Bonferroni method, so calculate by both methods and compare.

The Bonferroni method is also useful for apportioning the overall Type I error between different types of inference.

- For example, if three confidence intervals and two hypothesis tests are planned, and an overall Type I error rate of .05 is desired, then using 99% confidence intervals and individual significance rates .01 for the hypothesis tests will achieve this.

- This can also be used to apportion Type I error rate between *pre-planned inference* (the inference planned as part of the design of the study) and "data-snooping" inferences (inferences based on looking at the data and noticing other things of interest; more below).

- However, *this apportioning should be done <u>before</u> analyzing the data.*

Whichever method is used, *it is important to make the calculations based on the number of tests that have been done, <u>not</u> just the number that are reported.* (See Data Snooping below for more discussion.)

***False discovery rate***:

An alternative to bounding Type I error was introduced by Benjamini and Hochberg (1995): bounding the *False Discovery Rate*.

> The ***False Discovery Rate*** (FDR) of a group of tests is the *expected value of the ratio of falsely rejected hypotheses to all rejected hypotheses*.

("Expected value" refers to the mean of a distribution. Here, the distribution is the sampling distribution of the ratio of falsely rejected hypotheses to all rejected hypotheses tested.)

*Note:*
- The family-wise error rate (FWER) focuses on the possibility of making *any* error among all the inferences performed.
- The false discovery rate (FDR) tells you what *proportion* of the *rejected* null hypotheses are, *on average*, really true.
- Bounding the FDR rather than the FWER may be a more reasonable choice when many inferences are performed, especially if there is little expectation of harm from falsely rejecting a null hypothesis.
- Thus it is increasingly being adopted in areas such as micro-array gene expression experiments or neuro-imaging.

As with the FWER, there are various methods of actually bounding the false discovery rate.
- References: Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), and Benjamini and Yekutieli (2005)

***Subtleties and controversies***

Bounding the overall Type I error rate (FWER) will reduce the power of the tests, compared to using individual Type I error rates.

- Some researchers use this as an argument against multiple inference procedures.
- The counterargument is the argument for multiple inference procedures to begin with: Neglecting them will produce excessive numbers of false findings, *so that the "power" as calculated from single tests is misleading.*
  - See Maxwell and Kelley (2011) and Maxwell (2004) for more details.
- Bounding the False Discovery Rate (FDR) will usually give higher power than bounding the overall Type I error rate (FWER).

Consequently, it is important to consider the particular circumstances, as in considering both Type I and Type II errors in deciding significance levels.

- *In particular, it is important to consider the consequences of each type of error in the context of the particular research.*

Examples:

1. A research lab is using hypothesis tests to screen genes for possible candidates that may contribute to certain diseases.
   - Each gene identified as a possible candidate will undergo further testing.
   - If the results of the initial screening are not to be published except in conjunction with the results of the secondary testing, *and* if the secondary screening is inexpensive enough that many second level tests can be run, then the researchers could reasonably decide to ignore overall Type I error in the initial screening tests, since there would be no harm or excessive expense in having a high Type I error rate.
   - However, if the secondary tests were expensive, the researchers would reasonably decide to bound either family-wise Type I error rate or False Discovery Rate.

2. Consider a variation of the situation in Example 1:
   - The researchers are using hypothesis tests to screen genes as in Example 1, but plan to publish the results of the screening *without* doing secondary testing of the candidates identified.
   - In this situation, ethical considerations would warrant bounding either the FWER or the FDR -- *and* taking pains to emphasize in the published report that these results are just of a preliminary screening for possible candidates, and that these preliminary findings need to be confirmed by further testing.

### III. DATA SNOOPING

***Data snooping*** refers to statistical inference that the researcher decides to perform *after* looking at the data (as contrasted with *pre-planned* inference, which the researcher plans *before* looking at the data).

Data snooping can be done:

- professionally and ethically, or

- misleadingly and unethically, or

- misleadingly out of ignorance.

Misleading data snooping out of ignorance is a **common error** in using statistics. The problems with data snooping are essentially the problems of multiple inference.

- One way in which researchers unintentionally obtain misleading results by data snooping is in failing to account for *all* of the data snooping they engage in.

- In particular, *in accounting for Type I error when data snooping, you need to count not just the actual hypothesis tests performed, but <u>also all comparisons looked at when deciding which post hoc (i.e., not pre-planned) hypothesis tests to try.</u>*

*Example*:  A group of researchers plans to compare three dosages of a drug in a clinical trial.
- There is no pre-planned intent to compare effects broken down by sex, but the sex of the subjects is recorded.
- The researchers have decided to have an overall Type I error rate of 0.05, allowing 0.03 for the pre-planned inferences and 0.02 for any data snooping they might decide to do.
- The pre-planned comparison shows no statistically significant difference between the three dosages when the data are not broken down by sex.
- However, since the sex of the patients is known, the researchers decide to look at the outcomes broken down by combination of sex and dosage, notice that the results for women in the high-dosage group look much better than the results for the men in the low dosage group, and perform a hypothesis test to check that out.
- *In accounting for Type I error, the researchers need to take the number of data-snooping inferences performed as 15, <u>not</u> one*.
  - o The reason is that *they have looked at fifteen comparisons*:  there are $3 \times 2 = 6$ dosage$\times$sex combinations, and hence $(6 \times 5)/2 = 15$ pairs of dosage$\times$sex combinations.
  - o Thus the significance level for the post hoc test should not be 0.02, but (if the Bonferroni method is used) 0.02/15.

*See the Appendix for suggestions on data snooping professionally and ethically.*

## IV: USING AN INAPPROPRIATE METHOD OF ANALYSIS

*"Assumptions behind models are rarely articulated, let alone defended. The problem is exacerbated because journals tend to favor a mild degree of novelty in statistical procedures. Modeling, the search for significance, the preference for novelty, and the lack of interest in assumptions -- these norms are likely to generate a flood of nonreproducible results."*
   David Freedman, *Chance* 2008, v. 21 No 1, p. 60

*Recall*: Each frequentist inference technique (hypothesis test or confidence interval) involves *model assumptions*.
- Different techniques have different model assumptions.
- The validity of the technique depends (to varying extents) on whether or not the model assumptions are true for the context of the data being analyzed.
- Many techniques are *robust* to departures from at least some model assumptions.
    - o This means that if the particular assumption is not too far from true, then the technique is still approximately valid.
    - o Illustration: Rice Virtual Lab in Statistics Robustness Simulation

*Thus, when using a statistical technique, it is important to ask:*

- What are the model assumptions for that technique?
- Is the technique robust to some departures from the model assumptions?
- What reason is there to believe that the model assumptions (or something close enough, if the technique is robust) are true for the situation being studied?

***Neglecting these questions is a <u>very common mistake</u> in using statistics.***

- Sometimes researchers check only some of the assumptions, perhaps missing some of the most important ones.

Unfortunately, the model assumptions vary from technique to technique, so there are few if any general rules. One general rule of thumb, however is:

*Techniques are least likely to be robust to departures from assumptions of independence.*

- *Recall*: Assumptions of independence are often phrased in terms of "random sample" or "random assignment", so these are very important.

- One exception is that, for large enough populations, sampling *without* replacement is good enough, even though "independent" technically means sampling *with* replacement.

- Variance estimates depend strongly on the assumption of independence, so results can be very misleading when observations are not independent.

*Note*: Many techniques are most robust to violations of normality assumptions, at least if the sample size is large and the distribution is not strongly skewed or multimodal.

- This is because test statistics are often sums or linear combinations, which by the Central Limit Theorem tend to be approximately normally distributed. (More below)

***How do I know whether or not model assumptions are satisfied?***

Unfortunately, there are no one-size-fits-all methods, but here are some rough guidelines:

1. When selecting samples or dividing into treatment groups, be very careful in randomizing *according to the requirements of the method of analysis to be used.*

2. Sometimes (not too often) model assumptions can be justified plausibly by well-established facts, mathematical theorems, or theory that is well supported by sound empirical evidence.

- Here, "well established" means *well established by sound empirical evidence and/or sound mathematical reasoning.*

- This is *not* the same as "well accepted," since sometimes things may be well accepted without sound evidence or reasoning.

- More below.

3. Sometimes a rough idea of whether or not model assumptions might fit can be obtained by plotting the data or residuals obtained from a tentative use of the model.

- Unfortunately, these methods are typically better at telling you when the model assumption does *not* fit than when it does.

- Some examples, guidelines, and cautions below.

- But always remember "The Big Picture":

*Robert Kass' Big Picture of Statistical Inference*

In Kass (2011, p. 6, Figure 1), Robert Kass has proposed the following diagram to depict the "big picture" in using statistics:
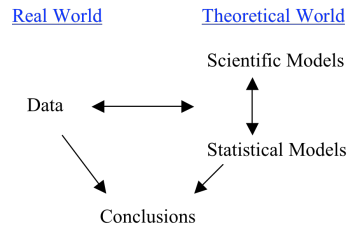


Points this picture is intended to show include:
- Both statistical and scientific models are abstractions, living in the "theoretical" world, as distinguished from the "real" world where data lie.
- Conclusions straddle these two worlds: conclusions about the real world typically are indirect, via the scientific models.
- "When we use a statistical model to make a statistical inference we implicitly assert that the variation exhibited by data is captured reasonably well by the statistical model, so that the theoretical world corresponds reasonably well to the real world." (p. 5)
- Thus "careful consideration of the connection between models and data is a core component of … the art of statistical practice…" (p. 6)

For a recent accessible discussion of problems with model assumptions in a topic of current wide interest (value-added models in education), see Wainer (2011).

**V. EXAMPLES OF CHECKING MODEL ASSUMPTIONS USING WELL-ESTABLISHED FACTS OR THEOREMS**

*Recall*: Here, "well established" means *well established by empirical evidence and/or sound mathematical reasoning*.

- This is *not* the same as "well accepted," since sometimes things may be well accepted without sound evidence.

**1. Using laws of physics**

Hooke's Law says that when a weight that is not too large (below what is called the "elastic limit") is placed on the end of a spring, the length of the (stretched) spring is approximately a linear function of the weight.

- This tells us that if we do an experiment with a spring by putting various weights (below the elastic limit) on it and measuring the length of the spring, we are justified in using a linear model,

    Length = A×Weight + B

## 2. Using the Central Limit Theorem

One form of The Central Limit Theorem says that for most distributions, a linear combination (e.g., the sum or the mean) of a large enough number of independent random variables is approximately normal.

- Thus, if a random variable in question is the sum of independent random variables, then it is usually safe to assume that it is approximately normal.

- For example, adult human heights (at least if we restrict to one sex) are the sum of many heights: the heights of the ankles, lower legs, upper legs, pelvis, many vertebrae, and head.

  - Empirical evidence suggests that these heights vary roughly independently (e.g., the ratio of height of lower leg to that of upper leg varies considerably).

  - Thus it is plausible by the Central Limit Theorem that human heights are approximately normal.

  - This in fact is supported by empirical evidence.

- *Caution*: "Most" is not "all." There are some distributions for which the central limit theorem is not valid. One notable exception is distributions which are "heavy-tailed" (also called *leptokurtic*). Such distributions occur in certain situations, such as seed dispersal in biology.

- The Central Limit Theorem can also be used to reason that some distributions are approximately *lognormal* -- that is, that the logarithm of the random variable is normal.

  - For example, the distribution of a pollutant might be determined by successive independent dilutions of an original emission.

  - This translates into mathematical terminology by saying that the amount of pollution (call this random variable Y) in a given small region is the *product* of independent random variables.

  - Thus logY is the *sum* of independent random variables.

  - If the number of successive dilutions is large enough, the reasoning above shows that logY is approximately normal, and hence that Y is approximately lognormal.

# VI. USING PLOTS TO CHECK MODEL ASSUMPTIONS

**Overall Cautions**:

1. Unfortunately, these methods are typically better at telling you when the model assumption does *not* fit than when it does.

2. There is inherently an element of subjectivity in using model-checking plots.

   o Some people are more likely than others to "see things that aren't really there."

   o Buja et al (2009) have recently proposed some protocols for taking this into account.

3. Different techniques have different model assumptions, so additional model checking plots may be needed.

   o Be sure to consult a good reference *for the particular technique* you are considering using.

**General Rule of Thumb**:

- First check any independence assumptions,

- then any equal variance assumption,

- then any assumption on distribution (e.g., normal) of variables.

*Rationale*: Techniques are usually least robust to departures from independence, and most robust to departures from normality.
- See van Belle (2008), pp. 173 - 177 and the references given there for more detail.

# Suggestions and Guidelines for

# Checking Specific Model Assumptions

## Checking for Independence

Independence assumptions are usually formulated in terms of error terms rather than in terms of the outcome variables.
- For example, in simple linear regression, the model equation is
$$Y = \alpha + \beta x + \varepsilon,$$
  where Y is the outcome (response) variable and $\varepsilon$ denotes the error term (also a random variable).
- It is the error terms that are assumed to be independent, not the values of the response variable.

We *do not know* the values of the error terms $\varepsilon$, so we can only plot the residuals $e_i$ (defined as the observed value $y_i$ minus the fitted value, according to the model), which approximate the error terms.

*Rule of Thumb*: To check independence, plot residuals against:

- Any time variables present (e.g., order of observation)

- Any spatial variables present,

- Any variables used in the technique (e.g., factors, regressors)

A pattern that is not random suggests lack of independence.

*Rationale*: Dependence on time or on spatial variables is a common source of lack of independence, but the other plots might also detect lack of independence.

*Comment*s:

1. Because time or spatial correlations are so frequent, it is important when making observations to *record any time or spatial variables that could conceivably influence results.*
  • This not only allows you to make the residual plots to detect possible lack of independence, but also allows you to change to a technique incorporating additional time or spatial variables if lack of independence is detected in these plots.

2. Since it is known that the residuals sum to zero, they are *not* independent, so the plot is really a very rough *approximation*.

3. Some models only require that errors are uncorrelated, not independent; model checks are the same as for independence.

*See the Appendix for some suggestions for checking model assumptions of equal variance and of normality.*

## Checking for Linearity

When considering a *simple linear regression model*, it is important to check the linearity assumption -- i.e., that the *conditional means* of the response variable are a linear function of the predictor variable.

Graphing the response variable vs. the predictor can often give a good idea of whether or not this is true.

However, one or both of the following refinements may be needed:

1. Plot residuals (instead of response) vs. predictor.
  • A non-random pattern suggests that a simple linear model is not appropriate; you may need to transform the response or predictor, or add a quadratic or higher term to the mode.

2. Use a scatterplot smoother such as lowess (also known as loess) to give a visual estimation of the conditional mean.
  • Such smoothers are available in many regression software packages.
  • *Caution*: You may need to choose a value of a smoothness parameter. Making it too large will over smooth; making it too small will not smooth enough.

When considering a *linear regression with just two terms*, plotting response (or residuals) against the two terms (making a three-dimensional graph) can help gauge suitability of a linear model, especially if your software allows you to rotate the graph.

*Caution*: It is *not* possible to gauge from scatterplots whether a linear model in *more than two predictors* is suitable.
- One way to address this problem is to try to transform the predictors to approximate multivariate normality.
  - See, e.g., Cook and Weisberg (1999), pp. 324 – 329.
- This will ensure not only that a linear model is appropriate for all (transformed) predictors together, but also that a linear model is appropriate even when some transformed predictors are dropped from the model.

*Note*: It is a **common mistake** to assume that if a linear model fits with all predictors included, then a linear model will still fit when some predictors are dropped. (*See Appendix for example.*)

# VII. SOME SPECIFIC SITUATIONS WHERE MISTAKES INVOLVING MODEL ASSUMPTIONS ARE COMMON

A. Comparing groups in studies with drop-outs (Intent-to-treat analysis)

B. Using a two-sample test comparing means when cases are paired (and generalizations)

C. Not distinguishing between fixed and random factors

D. Analyzing data without regard to how they were collected

E. Pseudoreplication

F. Mistakes in regression

For more discussion of some inappropriate methods of analysis, see:

- The Appendix

- Harris et al (2009)

- The Common Mistakes in Using Statistics website at http://www.ma.utexas.edu/users/mks/statmistakes/TOC.html

**A. Intent to Treat Analysis: Comparing Groups when there are Dropouts**

*The Problem*: In many forms of comparison of two treatments involving human subjects (or animals or plants), there are subjects who do not complete the treatment.

- They may die, move away, encounter life circumstances that take priority, or just decide for whatever reason to drop out of the study or not do all that they are asked.

- It is tempting to just analyze the data for those completing the protocol, essentially ignoring the dropouts. ***This is usually a serious mistake***, for two reasons:

1. In a good study, subjects should be randomized to treatment.

   o Analyzing the data for only those who complete the protocol *destroys the randomization, so that model assumptions are not satisfied*.

   o To preserve the randomization, outcomes for *all* subjects *assigned* to each group (whether or not they stick with the treatment) need to be compared. This is called **intent-to-treat** (or intention-to-treat, or ITT) analysis.

2. Intent-to-treat analysis is usually more informative for consumers of the research.

- For example, in studying two drug treatments, dropouts for reasons not related to the treatment can be expected to be, on average, roughly the same for both groups.

- But if one drug has serious side-effects that prompt patients to discontinue use, that would show up in the drop-out rate, and be important information in deciding which drug to use or recommend.

Reason 1 (and sometimes also reason 2) also applies when treatments are applied to animals, plants, or even objects.

For more information on intent-to-treat analysis, see Freedman (2005, pp. 5, 15), Freedman (2006), van Belle (2008, pp. 156 – 157), and Moher et al (2010)

**B. Using a Two-Sample Test Comparing Means when Cases Are Paired (and similar problems)**

One of the model assumptions of the two-sample t-tests for means is that the observations *between groups*, as well as within groups, are independent.

- Thus if samples are chosen so that there is some natural pairing, then the members of pairs are not independent, so the two-sample t-test is *not* appropriate.

**Example 1**: A random sample of heterosexual married couples is chosen. Each spouse of each pair takes a survey on marital happiness. The intent is to compare husbands' and wives' scores.

- The two-sample t-test would compare the *average* of the husband's scores with the *average* of the wives' scores.
- However, it is not reasonable to assume that the samples of husbands and wives are not independent -- some factors influence a particular husband's score are likely to influence his wife's score, and vice versa.
- Thus the independence assumption *between* groups for a two-sample t-test is violated.
- In this example, we can instead consider the individual differences in scores for each couple: (husband's score) - (wife's score). If the questions of interest can be expressed in terms of these differences, then we can consider using the one-sample t-test (or perhaps a non-parametric test if the model assumptions of that test are not met).

**Example 2**: A test is given to each subject before and after a certain treatment. (For example, a blood test before and after receiving a medical treatment; or a subject matter test before and after a lesson on that subject)

- This poses the same problem as Example 1: The "before" test results and the "after" test results for each subject are *not independent*, because they come from the same subject.

- The solution is the same: analyze the *difference in scores*.

- Example 2 is a special case of what is called *repeated measures*: some measurement is taken more than once on the same unit.

  o Because repeated measures on the same unit are not independent, the analysis of such data needs a method that takes this lack of independence into account.

  o There are various ways to do this; just which one is best depends on the particular situation.

**Similar Problem:** *Hierarchical (multilevel) situations may violate model assumptions of independence*

*Example*: Researchers are studying how well scores on a standardized eighth grade math exam predict performance on an Algebra I end-of-course exam for ninth-grade students.

- They have data from an entire school district.

- They propose to analyze it by simple linear regression.

- However, standard regression methods of inference assume that observations are uncorrelated, whereas observations from students in the same school can be expected to be correlated.

- Instead, the researchers need to use a multilevel (also called hierarchical) model that takes into account that observations from the same school may be correlated.

**C. Inappropriately Designating a Factor as Fixed or Random**

In Analysis of Variance and Multilevel Modeling, there are two types of factors: *fixed effect* and *random effect*.

Which type is appropriate depends on the context of the problem, the questions of interest, and how the data is gathered. Here are the differences:

***Fixed effect factor***: Data has been gathered from *all the levels of the factor that are of interest*.

*Example*: The purpose of an experiment is to compare the effects of three specific dosages of a drug on the response.
- "Dosage" is the factor.
- The three specific dosages in the experiment are the levels.
- There is no intent to say anything about other dosages.
- Therefore this is a fixed factor.

***Random effect factor***:

- The factor has *many possible levels*.

- *All* possible levels are of interest.

- Only a *random sample of levels* is included in the data.

*Example*: A large manufacturer of widgets is interested in studying the effect of machine operator on the quality final product. The researcher selects a random sample of operators from the large number of operators at the various facilities that manufacture the widgets and collects data on just these operators.
  - The factor is "operator."
  - Each operator is a level of the factor.
  - Since interest is not just in the operators for whom data is gathered, this is a random factor.
  - The analysis will not estimate the effect of each of the operators in the sample, but will instead estimate the variability attributable to the factor "operator".

*(See Appendix for more discussion and a possible confusion involving terminology)*

***The appropriate statistical analysis depends on whether the factor is treated as fixed or as random***. That is, fixed and random effects require different models

- Consequently, inferences may be incorrect if the factor is classified inappropriately.

- Mistakes in classification are most likely to occur when more than one factor is considered in the study.

*Example*: Two surgical procedures are being compared.
  - Patients are randomized to treatment.
  - Five different surgical teams are used.
  - To prevent possible confounding of treatment and surgical team, each team is trained in both procedures, and each team performs equal numbers of surgery of each of the two types.
  - Since the purpose of the experiment is to compare the *procedures*, the intent is to generalize to other surgical teams.
  - *Thus surgical team should be considered as a random factor, not a fixed factor*.

*Comments*:

- This example can help understand why inferences might be different for the two classifications of the factor: Asserting that there is a difference in the results of the two procedures *regardless of the surgical team* is a stronger statement that saying that there is a difference in the results of the two procedures *just for the teams in the experiment.*
- Technically, the levels of the random factor (in this case, the five surgical teams) used in the experiment should be a random sample of all possible levels.
  - o This is in practice usually impossible, so the random factor analysis is usually used if there is reason to believe that the teams used in the experiment could reasonably be a random sample of all surgical teams who might perform the procedures.
  - o However, this assumption needs careful thought to avoid possible bias.
  - o For example, the conclusion would be sounder if it were limited to surgical teams which were trained in both procedures in the same manner and to the same extent, and who had the same surgical experiences, as the five teams actually studied.

*(See Appendix for additional comments.)*

**D. Analyzing Data without Regard to How They Were Collected**

Using a two-sample t-test when observations are paired (see above) is one example of this. Here is another:

*Example*: [See Potcner and Kowalski (2004) for data and details.] An experiment was conducted to study the effect of two factors (pretreatment and stain) on the water resistance of wood.
- Two types of pretreatment and four types of stain were considered.
- For reasons of practicality and economy, the experiment was conducted with a *split-plot design* as follows:
  - o Six entire boards were the *whole plots*.
  - o One pretreatment was applied to each board, with the two pretreatments randomly assigned to the six boards (three boards per pretreatment).
  - o Then each pre-treated board was cut into four smaller pieces of equal size (these were the *split-plots*).
  - o The four pieces from each entire board were randomly assigned to the four stains.
  - o The water resistance of each of the 24 smaller pieces was measured; this was the response variable.
- The following chart shows the p-values of the three significance tests involved if the correct split-plot analysis is used, and if an incorrect analysis (assuming a crossed design, with the 6 treatment combinations randomly assigned to the 24 smaller pieces of wood, with 4 small pieces per treatment combination) is used.
- Note that the conclusions from the two analyses would be quite different!

| p-values | Correct (Split Plot) Analysis | Incorrect (Crossed Design) Analysis |
|---|---|---|
| **Interaction** | 0.231 | 0.782 |
| **Pretreatment** | 0.115 | 0.002 |
| **Stain** | 0.006 | 0.245 |

Some of the many considerations to take into account in deciding on an appropriate method of analysis include:

- The sampling method
- Whether or not there was blocking in an experimental design
- Whether factors are nested or crossed
- Whether factors are fixed or random
- Pseudoreplication (See below)

## E. PSEUDOREPLICATION

The term *pseudoreplication* was coined by Hurlbert (1984,p. 187) to refer to

> "the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent."

His paper concerned ecological field experiments, but pseudoreplication can occur in other fields as well.

In this context, *replication* refers to having more than one experimental (or observational) unit with the same treatment. Each unit with the same treatment is called a *replicate*.

> *Note*: There are other uses of the word replication -- for example, repeating an entire experiment is also called replication; each repetition of the experiment is called a replicate. This meaning is related to the one given above: If each treatment in an experiment has the same number r of replicates (in the sense given above), then the experiment can be considered as r replicates (in the second sense) of an experiment where each treatment is applied to only one experimental unit.

Heffner et al (1996, p. 2558) distinguish a pseudoreplicate from a *true replicate*, which they characterize as

> "the smallest experimental unit to which a treatment is independently applied."

Most models for statistical inference require true replication.
- *True* replication permits the estimation of *variability within a treatment*.
- Without estimating variability within treatments, it is impossible to do statistical inference.

*Example*: Consider comparing two drugs by trying drug A on person 1 and drug B on person 2.
- Drugs typically have different effects in different people.
- So this simple experiment will give us no information about generalizing to people other than the two involved.
- But if we try each drug on several people, then we can obtain some information about the *variability* of each drug, and use statistical inference to gain some information on whether or not one drug might be more effective than the other on average.

True replicates are often confused with repeated measures or with pseudoreplicates. The following illustrate some of the ways this can occur.

***Examples***:

1. Suppose a blood-pressure lowering drug is administered to a patient, and then the patient's blood pressure is measured twice.
- This is a *repeated measure*, not a replication.
- It can give information about the *uncertainty in the measurement process*, but *not* about the *variability in the effect of the drug*.
- On the other hand, if the drug were administered to two patients, and each patient's blood pressure was measured once, we can say the treatment has been replicated, and the replication may give some information about the variability in the effect of the drug.

2. A researcher is studying the effect on plant growth of different concentrations of $CO_2$ in the air.
- He needs to grow the plants in a growth chamber so that the $CO_2$ concentration can be set.
- He has access to only two growth chambers, but each one will hold five plants.
- However, since the five plants in each chamber share whatever conditions are in that chamber besides the $CO_2$ concentration, and in fact may also influence each other, they are not independent replicates but are pseudoreplicates.
- The growth chambers are the experimental units: the treatments are applied to the growth chambers, not to the plants independently.

3. Two fifth-grade math curricula are being studied.
- Two schools have agreed to participate in the study.
- One is randomly assigned to use curriculum A, the other to use curriculum B.
- At the end of the school year, the fifth-grade students in each school are tested and the results are used to do a statistical analysis comparing the two curricula.
- There is no true replication in this study; the students are pseudo-replicates.
- The schools are the experimental units; they, not the students, are randomly assigned to treatment.
- Within each school, the test results (and the learning) of the students in the experiment are not independent; they are influenced by the teacher and by other school-specific factors (e.g., previous teachers and learning, socioeconomic background of the school, etc.).

*Consequences of doing statistical inference using pseudoreplicates rather than true replicates*

*Variability will probably be <u>underestimated</u>.* This will result in

- Confidence intervals that are too small.
- An inflated probability of a Type I error (falsely rejecting a true null hypothesis).

### Comments

- Note that in Example 2, there is no way to distinguish between effect of treatment and effect of growth chamber; thus the two factors (treatment and growth chamber) are *confounded*. Similarly, in Example 3, treatment and school are confounded.
- Example 3 may also be seen as applying the two treatments to two different *populations* (students in one school and students in the other school)
- Observational studies are particularly prone to pseudoreplication.
- Regression can sometimes account for lack of replication, provided data are close enough to each other.
  - The rough idea is that the responses for nearby values of the explanatory variables can give some estimate of the variability.
  - However, having replicates is better.

*(See Appendix for suggestions on dealing with pseudoreplication.)*

## F. MISTAKES IN REGRESSION

*There are many common mistakes involved in regression!*

Only one will be discussed here; some others will be listed at the end of these notes, with a web reference to more discussion.

### Overfitting

> *With four parameters I can fit an elephant and with five I can make him wiggle his trunk.*
> John von Neumann

If we have n distinct x values and corresponding y values for each, it is possible to find a curve going exactly through all n resulting points (x, y); this can be done by setting up a system of equations and solving simultaneously.
- But this is *not* what regression methods typically are designed to do.
- Most regression methods (e.g., least squares) estimate *conditional means* of the response variable given the explanatory variables.
- They are not expected to go through all the data points.

For example, with one explanatory variable X (e.g., height) and response variable Y (e.g., weight), if we fix a value x of X, we have a *conditional distribution of Y given X = x* (e.g., the conditional distribution of weight for people with height x).
- This conditional distribution has an expected value (population mean), which we will denote $E(Y|X = x)$ (e.g., the mean weight of people with height x).
- This is the *conditional mean of Y given X = x*. It depends on x -- in other words, $E(Y|X = x)$ is a mathematical function of x.

In least squares regression (and most other kinds of regression), *one of the model assumptions is that the conditional mean function has a specified form*.

- Then we use the data to find a function of x *that approximates the function E(Y|X = x)*.
- This is different from, and subtler (and harder) than, finding a curve that goes through all the data points.

***Example***: To illustrate, I have used simulated data:

- Five points were sampled from a joint distribution where the conditional mean $E(Y|X = x)$ is known to be $x^2$, and where each conditional distribution $Y|(X = x)$ is normal with standard deviation 1.
- I used least squares regression to estimate the conditional means by a quadratic curve $y = a + bx + cx^2$. That is, I used least squares regression, with

  $E(Y|X=x) = \alpha + \beta x + \gamma x^2$

  as one of the model assumptions, to obtain estimates a, b, and c of $\alpha$, $\beta$, and $\gamma$ (respectively), based on the data.
- There are other ways of expressing this model assumption, for example,

  $y = \alpha + \beta x + \gamma x^2 + \varepsilon,$

  or

  $y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$

The graph below shows:

- The five data points in *red* (one at the left is mostly hidden by the green curve)
- The curve $y = x^2$ of true conditional means (*black*)
- The graph of the calculated regression equation (in *green*).

Scatterplot of y vs x, resp vs expl, fit vs x
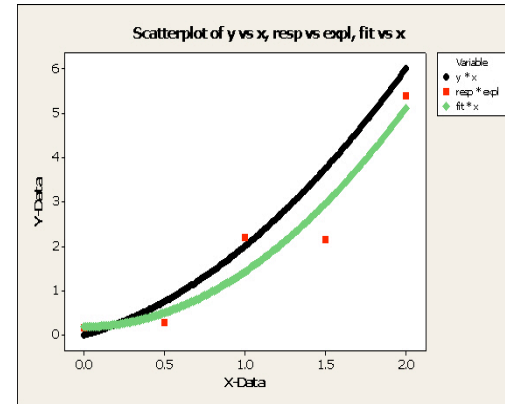
Note that:

- The points sampled from the distribution do *not* lie on the curve of means (black).

- The green curve is not exactly the same as the black curve, but is close.

- In this example, the sampled points were mostly below the curve of means.

- Since the regression curve (green) was calculated using just the five sampled points (red), the red points are more evenly distributed above and below it (green curve) than they are in relation to the real curve of means (black).

*Note*: In a real world example, we would *not* know the conditional mean function (black curve) -- and in most problems, would not even know in advance whether it is linear, quadratic, or something else.

- Thus, *part of the problem of finding an appropriate regression curve is figuring out what kind of function it should be*.

Continuing with this example, if we (naively) try to get a "good fit" by trying a quartic (fourth degree) regression curve -- that is, using a model assumption of the form

$$E(Y|X=x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4,$$

we get the following picture:
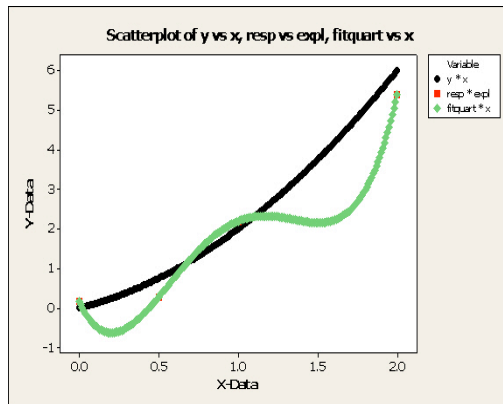
You can barely see any of the red points in this picture.

- That's because they're all on the calculated regression curve (green).

- We have found a regression curve that fits all the data!

- But it is *not* a good regression curve -- because what we are really trying to estimate by regression is the *black curve* (curve of conditional means).

- We have done a rotten job of that; we have made the mistake of *over-fitting*. We have fit an elephant, so to speak.

If we had instead tried to fit a cubic (third degree) regression curve -- that is, using a model assumption of the form

$$E(Y|X=x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

we would get something more wiggly than the quadratic fit and less wiggly than the quartic fit.

- However, it would still be over-fitting, since (by construction) the correct model assumption for these data would be a quadratic mean function.

*See the Appendix for suggestions on trying to avoid overfitting.*

***Other Common Mistakes in Using Regression***

*For further discussion of these mistakes, see links from*
**http://www.ma.utexas.edu/users/mks/statmistakes/regression.html**

Using Confidence Intervals when Prediction Intervals Are Needed.

Over-interpreting High $R^2$

Mistakes in Interpretation of Coefficients
- Interpreting a coefficient as a rate of change in Y instead of as a rate of change in the conditional mean of Y.
- Not taking confidence intervals for coefficients (i.e., uncertainty of estimation of coefficients) into account
- Interpreting a coefficient that is not statistically significant
- Interpreting coefficients in multiple regression with the same language used for a slope in simple linear regression.
- Neglecting the issue of multiple inference when dealing with more than one coefficient in the same data set.

Mistakes in Selecting Terms
- Assuming linearity is preserved when variables are dropped. (*See also Appendix.*)
- Problems involving stepwise model selection procedures.