Some common problems in using statistical methodology

Snezana Kirin

Problem:

- In conducting and reporting of medical research, there are some common problems in using statistical methodology which may result in invalid inferences being made.
- We frequently hear in the news results of a research study that appears to contradict the results of a study published just a few years ago.
- Most of the problem comes from lack of understanding of statistical techniques, their proper use, and their limitations.

Uncertainty and probability

- We can't expect certainty - uncertainty is all around us.
- Uncertainty can often be "quantified" -- we can talk about degrees of certainty or uncertainty.
- This is the idea of probability:

a higher probability expresses a higher degree of certainty that something will happen.

Four perspectives on probability

There are four perspectives on probability that are commonly used:

1. Classical,
2. Empirical,
3. Subjective, and
4. Axiomatic.

- Using one perspective when another is intended can lead to misunderstandings and errors.
- Misunderstandings arising from lack of clarity about the reference category.

Misunderstandings about probability involving different uses of the word "risk“:

Usage word risk alone

- Brings up the question of reference category
- e.g. **"Your risk of dying of a heart attack is about 25%."**
- The risk may be higher or lower depending on other factors (gender, occupation, age, etc.) in addition to being a U.S. resident.

Relative risk

(risk ratio)

- This is a method of **comparing the risk of one group with the risk of another**
- One group might be people with a certain condition (or receiving a certain treatment) and the other group people without that condition (or not receiving the treatment). Relative risk is the ratio of the risks for the two groups. This immediately brings in two possible source of confusion:
- What are the two groups?
- Which group's risk is in the numerator and which group's risk is in the denominator?
- Sometimes only the relative risk is given. We also need to know the "absolute risk" (the risk of at least one of the groups involved) in order to interpret what the relative risk is telling us.

Good data are hard to find

- In fact, good data are notoriously hard to find.
- Just ask the statisticians at the World Health Organization (WHO).
- This group, representing 194 member countries, is tasked with analyzing the state of the world's health — a seriously consequent mission.
- Billions upon billions of dollars are allocated based on their findings, along with the work of other United Nations agenciessuch as UNICEF and the World Bank.
- Global policy shifts are set in motion based on their figures.
- And yet, a full two-thirds of deaths in the world are no registered. A third of all births worldwide are not registered. Here we are, trying to generate reports to make a global health estimate and we don't even know who's living or dying or where or why.

The future is completely uncertain...
...I am completely certain of this.

A conditional probability and diagnostic tests

- In practice, many probabilities we encounter are **conditional probabilities**, although that is **not always made explicit**.

- Diagnostic tests typically have two outcomes, labeled "**positive**" and "**negative**."

- Unfortunately, diagnostic tests are almost never perfect.

Mistakes in Thinking About Causation

Confusing correlation and causation

Example:
Consider elementary school students' weight and scores on a standard reading exam. They are correlated, but saying that higher weight *causes* higher reading scores is as absurd as saying that high reading scores cause larger shoe size.

- absurd to say that weight causes age.

Source of misleading research results is inadequate attention to the choice

- In most research, one or more ***outcome variables*** are **measured**. Statistical analysis is done on the outcome measures, and conclusions are drawn from the statistical analysis.
- ***One common source of misleading research results is giving inadequate attention to the choice of either outcome variables or summary statistics***.
- Making a good choice depends on the particulars of the context, including the research question.
- Although there are some guidelines**, there are no one-size-fits-all rules**. **So aspects of this topic can best be approached by examples**.

Sampling

- A randomised controlled trial (RCT) is a type of scientific (often medical) experiment, where the people being studied are randomly allocated one or other of the different treatments under study.
- **The RCT is the gold standard for a clinical trial**.

What is a Random Sample?

Sources of Confusion

- The word "random" in the phrase "random sample" does *not* have its ordinary, everyday meaning -- that is, does *not* refer to the first definition you would find in a dictionary.

Errors in Sampling

- One **common mistake** that arises from applying "random sample" is concluding that a sample is not random because it has a pattern.

- In fact, a random sample (using the technical meaning of the phrase) ***might*** **have a pattern** (or **it might not**).

- **In fact,** ***there is no way we can tell from looking at the sample whether or not it qualifies as a random sample*****.**

A simple random sample (SRS)

Population refers to the collection of people, animals, locations, etc. that the study is focusing on.

Some examples:

1. **In a medical study, the population might be all adults over age 50 who have high blood pressure.**
2. **In another study, the population might be all hospitals in the country that perform heart bypass surgery.**

- Selecting a simple random sample in examples 1 and 2 is not easy.

Inappropriately Designating a Factor as Fixed or Random

- In Analysis of Variance and some other methodologies, there are two types of factors: ***fixed effect* and *random effect***.

- 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.

- **Random effect factor**: The factor has many possible levels, interest is in all possible levels, but only a random sample of levels is included in the data.

Additional Comments about Fixed and Random Factors

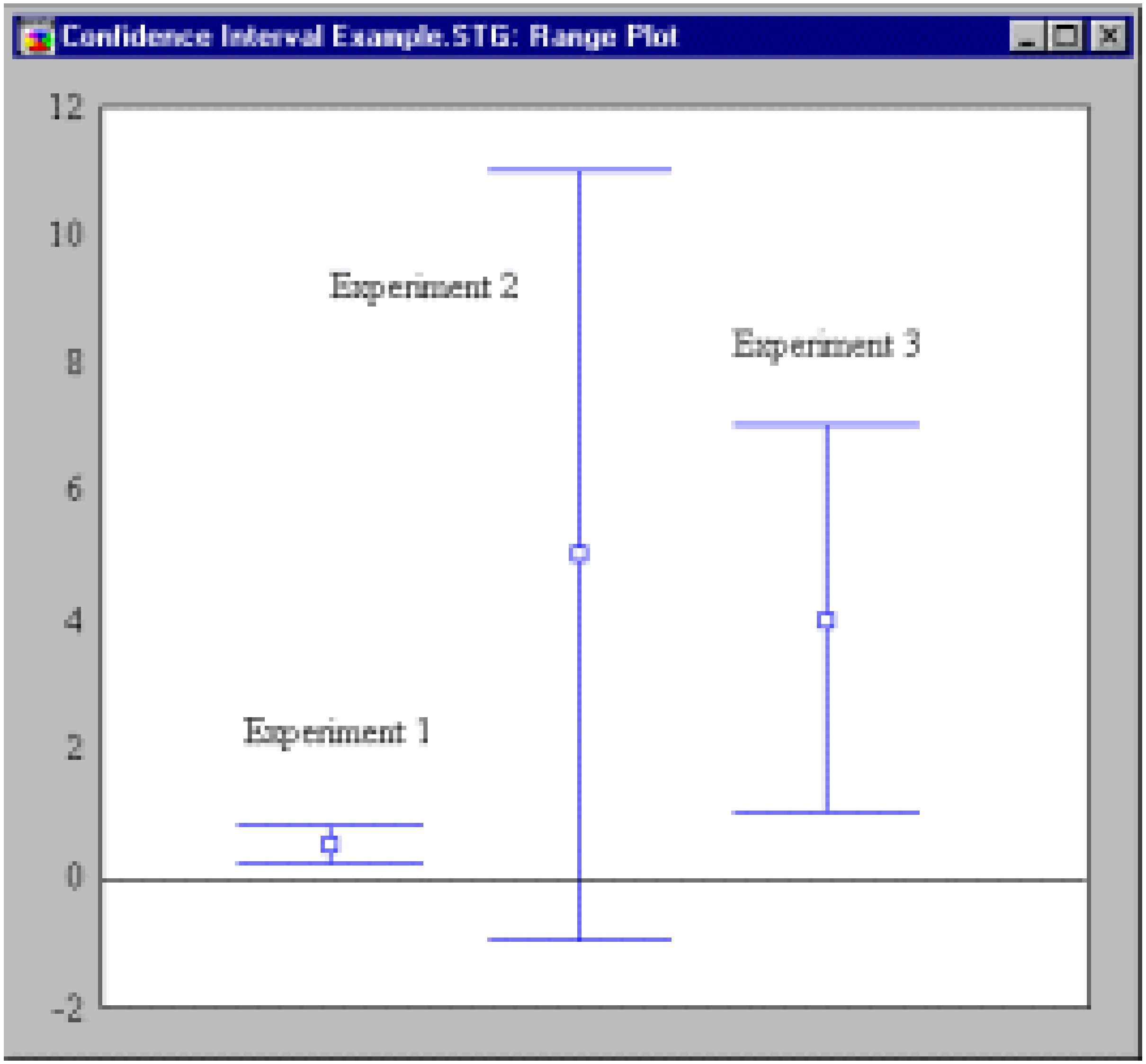
- The standard methods for analyzing random effects models assume that the random factor has infinitely many levels, but usually still work well **if the total number of levels of the random factor is at least 100 times the number of levels observed in the data.**

- Situations where the total number of levels of the random factor is less than 100 times the number of levels observed in the data require special "**finite population**" methods.

- **An interaction term involving both a fixed and a random factor should be considered a random factor.**

- A factor that is nested in a random factor should be considered random.

The confidence interval of differences between arithmetic mean



Experiment 1:

- large sample
- high precision
- little effect of the differences

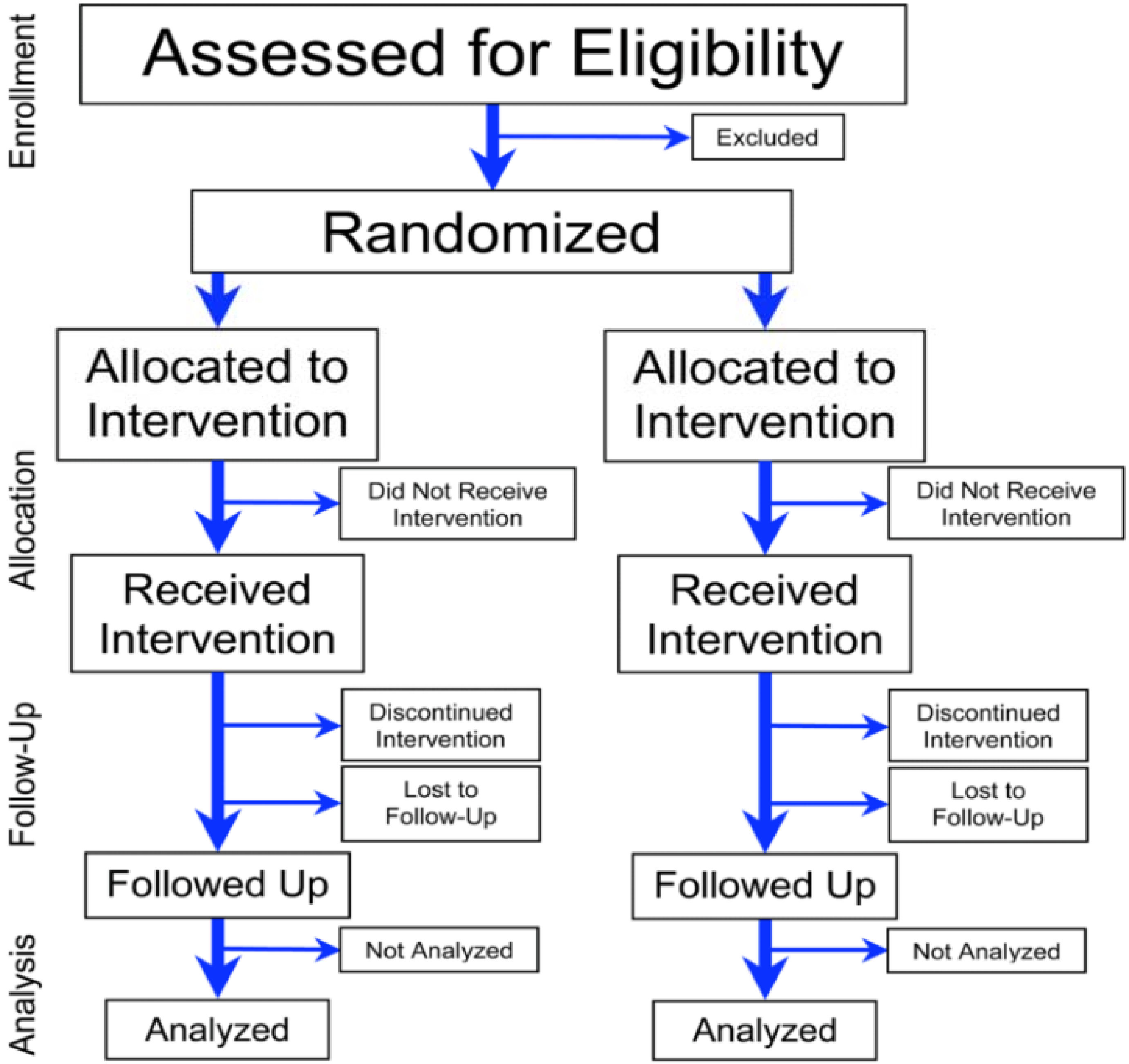
Experiment 2:

- too small the sample

Experiment 3:

- greater effect of differences
- less accurate

Source: http://www.statsoft.com/textbook/stpowan.html (06.06.2006)



Flowchart of four phases

1. **enrollment,**
2. **intervention allocation,**
3. **follow-up, and**
4. **data analysis)**

of a parallel randomized trial of two groups

modified from the CONSORT (Consolidated Standards of Reporting Trials) 2010 Statement

Why Is Random Sampling Important?

- **Randomly sampling will eliminate systematic bias.**

- **The *mathematical theorems which justify most frequentist statistical procedures apply only to random samples*.**

Over-Interpretation

Some common mistakes of this form include:

- **Extrapolation to a larger population than the one studied**
  - ***Example:***
    Running a psychology experiment with undergraduates enrolled in medicine classes and drawing a conclusion about people in general.
- **Extrapolation to a larger population occurs when the sampling method is biased.**
  - **Example:**
  - Results from a voluntary sample for an observational study cannot justifiably be extended to people who do not volunteer to participate.

Over-Interpretation

- **Using overly strong language in stating results**
  - Statistical procedures do not *prove* results.
  - They only give us information on whether or not the data support or are consistent with a particular conclusion.
  - **There is always uncertainty involved. Thus results need to be phrased in ways that acknowledge this uncertainty.**
- **Considering statistical significance but not practical significance**

Example:

  - Suppose that a well-designed, well-carried out, and carefully analyzed study shows that there is a statistically significant difference in life span between people engaging in a certain exercise regime at least five hours a week for at least two years and those not following the exercise regime. If the difference in average life span between the two groups is three days, ... well, so what?

Misinterpretations and misuses of p-values

p-value = the probability of obtaining a test statistic at least as extreme as the one from the data at hand, assuming:

- **the model assumptions** for the inference procedure used are all true, and
- **the null hypothesis is true**, and
- **the random variable is the same** (including the same population), and
- **the sample size is the same**.

Notice that this is a conditional probability:

- The probability that something happens, given that various other conditions hold.
- **One common misunderstanding is to neglect some or all of the conditions.**

Misinterpretations and misuses of p-values

- *Another* ***common misunderstanding*** *of p-values is the belief that the* ***p-value*** *is "the probability that the null hypothesis is true".*

- The basic assumption of hypothesis testing is that the null hypothesis is either true (in which case the probability that it is true is 1) or false (in which case the probability that it is true is 0).

Type I and II Errors and Significance Levels

Type I Error

Rejecting the *null* hypothesis when it is in fact true is called a *Type I error*.

Many people decide, before doing a hypothesis test, on a maximum p-value for which they will reject the null hypothesis. This value is often denoted α (alpha) and is also called the ***significance level***.

When a hypothesis test results in a p-value that is less than the significance level, the result of the hypothesis test is called *statistically significant*.

- **Common mistake: *Confusing statistical significance and practical significance*.**

Type I Error-example

- A large clinical trial is carried out to compare a new medical treatment with a standard one.
- The statistical analysis shows **a statistically significant difference in lifespan** when using the new treatment compared to the old one ⬅***statistical significance***
- But the increase in lifespan is at most three days, with average increase less than 24 hours, and with poor quality of life during the period of extended life. ⬅ ***practical significance***
- **Most people would not consider the improvement practically significant.**

Type I Error

- ***Caution***:

The larger the sample size, the more likely a hypothesis test will detect a small difference. Thus ***it is especially important to consider practical significance when sample size is large.***

Rephrasing using the definition of Type I error:

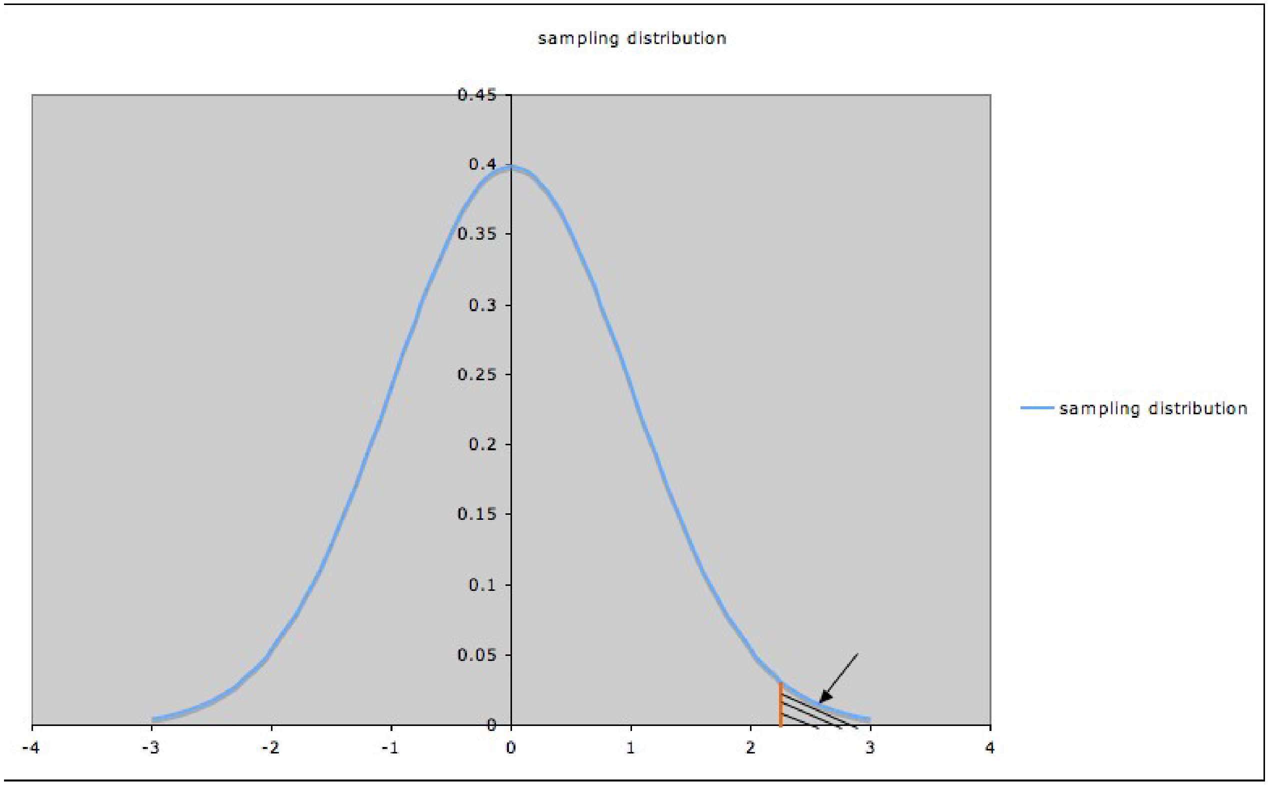
The significance level α *is the probability of making the wrong decision when the null hypothesis is true*.

Connection between Type I error and significance level:

A significance level α corresponds to a certain value of the test statistic, say t_α, represented by the orange line in the picture of a sampling distribution below (**the picture illustrates a hypothesis test with null hypothesis "μ = 0") and alternate hypothesis "μ > 0")**

Since the shaded area indicated by the arrow is the p-value corresponding to t_α, that p-value (shaded area) is α.

To have p-value less than α , a t-value for this test must be to the right of t_α.



So the probability of rejecting the null hypothesis when it is true is the probability that $t > t_\alpha$, which we saw above is α.

In other words, *the probability of Type I error is α.*

Pros and Cons of Setting a Significance Level:

- Setting a significance level (*before* doing inference) has the ***advantage*** that **the analyst is not tempted to chose a cut-off on the basis of what he or she hopes is true**.

- It has the ***disadvantage*** that **it neglects that some p-values might best be considered borderline**.

This is one reason why it is important to report p-values when reporting results of hypothesis tests.

It is also good practice to include confidence intervals corresponding to the hypothesis test.

Type II Error

- ***Not rejecting*** the *null* hypothesis when in fact the *alternate* hypothesis is true is called a ***Type II error***.

- The example below provides a situation where the concept of Type II error is important:
  In a t-test for a sample mean μ, with
  - **null hypothesis "μ = 0" and**
  - **alternate hypothesis "μ > 0",**
- we may talk about the **Type II error relative to the *general alternate hypothesis* "μ > *0*"**, or may talk about the **Type II error relative to the *specific alternate hypothesis* "μ > *1*"**.
- Note that the specific alternate hypothesis is a special case of the general alternate hypothesis.
- **In practice, people often work with Type II error relative to a *specific* alternate hypothesis.**
- In this situation, the probability of Type II error relative to the specific alternate hypothesis is often called **β**. In other words, **β** is the probability of making the wrong decision when the *specific alternate* hypothesis is true.

Type I and II Errors

- The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test (which equals 1−β).
- What we actually call type I or type II error depends directly on the null hypothesis.
- Negation of the null hypothesis causes type I and type II errors to switch roles.
- **The goal of the test is to determine if the null hypothesis can be rejected.**
- **A statistical test can either reject or fail to reject a null hypothesis, but never prove it true.**

Considering both types of error together:

The following table summarizes Type I and Type II errors:

	Truth		
		Null Hypothes is True	Null Hypothes is False
Decision (based on sample)	Reject Null	*Type I Error*	*Correct Decision*
	Fail to reject	*Correct Decision*	*Type II Error*

Some illustrations

- Usually **a type I error** leads one to conclude that a supposed effect or relationship exists when in fact it doesn't.
  - Examples of type I errors include:
    - a test that shows a patient to have a disease when in fact the patient does not have the disease,
    - a fire alarm going off indicating a fire when in fact there is no fire or
    - an experiment indicating that a medical treatment should cure a disease when in fact it does not.
- A **type II error** (or **error of the second kind**) is the failure to reject a false null hypothesis.
  - Examples of type II errors would be:
    - a blood test failing to detect the disease it was designed to detect, in a patient who really has the disease;
    - a fire breaking out and the fire alarm does not ring or a clinical trial of a medical treatment failing to show that the treatment works when really it does.

How to make decisin what significance level to use

- **It should be done *before* analyzing the data -- preferably before gathering the data.**

- The choice of significance level should be based on the consequences of Type I and Type II errors.

- **If the consequences of a type I error are serious or expensive, then a very small significance level is appropriate.**

What significance level to use - example

Example 1: Two drugs are being compared for effectiveness in treating the same condition. **Drug 1 is very affordable, but drug 2 is extremely expensive.**

- – The null hypothesis is "**both drugs are equally effective**," and
- – the alternate is "**Drug 2 is more effective than Drug 1**.“

- In this situation, a **Type I error would be deciding that Drug 2 is more effective**, when in fact it is no better than Drug 1, but would cost the patient much more money. That would be undesirable from the patient's perspective, so **a small significance level is warranted**.

- **If the consequences of a Type I error are not very serious (and especially if a Type II error has serious consequences), then a larger significance level is appropriate.**

Common mistake:

- Neglecting to think adequately about possible consequences of Type I and Type II errors (and deciding acceptable levels of Type I and II errors based on these consequences) before conducting a study and analyzing data.
- Sometimes there may be serious consequences of each alternative, so **some compromises or weighing priorities may be necessary**.
- **The trial analogy illustrates this well:**
  - **Which is better or worse, imprisoning an innocent person or letting a guilty person go free?**
  - ***This is a value judgment;***
  - ***value judgments are often involved in deciding on significance levels.***
  - ***Trying to avoid the issue by always choosing the same significance level is itself a value judgment.***

Common mistake:

- **Claiming that an alternate hypothesis has been "proved" because it has been rejected in a hypothesis test.**
- This is an instance of the common mistake of expecting too much certainty.
- There is always a possibility of a Type I error; the sample in the study might have been one of the small percentage of samples giving an unusually extreme test statistic.
- **This is why replicating experiments (i.e., repeating the experiment with another sample) is important.**
- **The more experiments that give the same result, the stronger the evidence.**
- **There is also the possibility that the sample is biased or the method of analysis was inappropriate**; either of these could lead to a misleading result.

Power of a statistical procedures

- *"... power calculations ... in general are more delicate than questions relating to Type I error."*

- Power is the probability that a randomly chosen sample satisfying the model assumptions will detect a difference of the specified type when the procedure is applied, if the specified difference does indeed occur in the population being studied.

- *For a confidence interval procedure*, power can be defined as the probability that the procedure will produce an interval with a half-width of at least a specified amount.

- In many real-life situations, there are reasonable conditions that we would be interested in being able to detect, and others that would not make a practical difference.

Examples:

- Some differences are of no practical importance -- for example, a medical treatment that extends life by 10 minutes is probably not worth it.

- In cases such as these, neglecting power could result in one or more of the following:
  - Doing much more work than necessary
  - Obtaining results which are meaningless,
  - Obtaining results that don't answer the question of interest.

Power of a statistical procedure-examples

- *For a hypothesis test*, power can be defined as the probability of rejecting the null hypothesis under a specified condition.

- *Example*: For a one-sample t-test for the mean of a population, with null hypothesis H_0: $\mu = 100$, you might be interested in the probability of rejecting H_0 when $\mu \geq 105$, or when $|\mu - 100| > 5$, etc.
- As with **Type II error**, we need to think of power in terms of ***power against a specific alternative*** rather than against a general alternative.

Power of a statistical procedure-conclusion

- The general phenomenon: ***the farther an alternative is from the null hypothesis, the higher the power of the test to detect it.***

- For most tests, it *is* possible to calculate the power against a specific alternative, at least to a reasonable approximation, if relevant information (or good approximations to them) is available.

- **It is *not* usually possible to calculate the power against a general alternative, since the general alternative is made up of infinitely many possible specific alternatives.**

Power and Type II Error

- Recall that the **Type II Error rate β** of a test against a specific alternate hypothesis test is represented in the diagram above as the area under the sampling distribution curve for that alternate hypothesis and to the *left* of the cut-off line for the test. Thus

  (Power of a test against a specific alternate hypothesis) + β = total area under sampling distribution curve = 1, so: **Power = 1 - β**

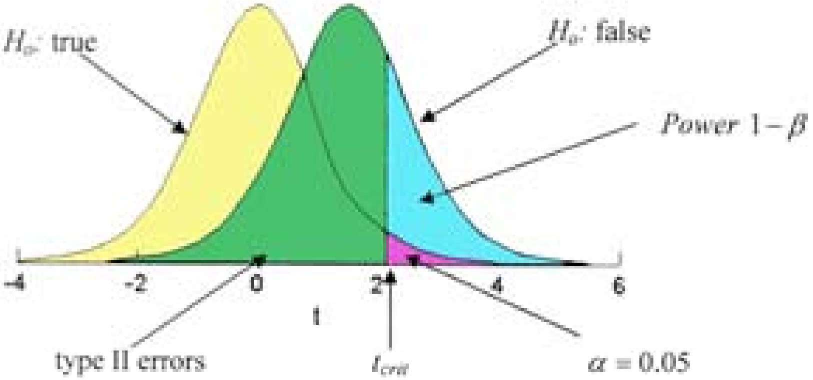
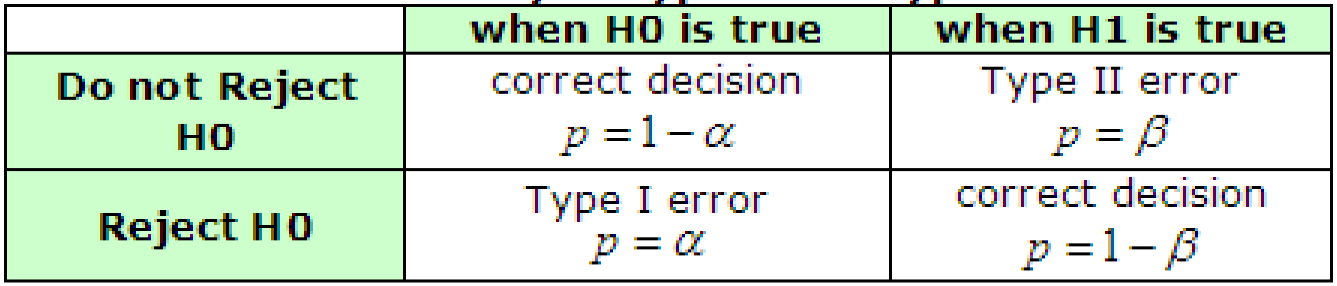


Table 1: Summary of Type I and Type II Errors



	when H0 is true	when H1 is true
Do not Reject H0	correct decision $p = 1 - \alpha$	Type II error $p = \beta$
Reject H0	Type I error $p = \alpha$	correct decision $p = 1 - \beta$

Factors that Affect the Power of a Statistical Procedure

- The power of a statistical procedure depends on the specific alternative chosen (for a hypothesis test) or a similar specification, such as **width of confidence interval** (for a confidence interval).

  The following factors also influence power:

1. **Sample Size**
   Power depends on sample size. **Other things being equal, larger sample size yields higher power.**

2. **Variance**
   Power also depends on variance: smaller variance yields higher power.

- **Variance can sometimes be reduced by using a better measuring instrument, restricting to a subpopulation, or by choosing a better experimental design.**

Using an Inappropriate Method of Analysis

- Each 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.

What are the model assumptions for certain statistical technique?

- Many techniques are *robust* to departures from at least some model assumptions.
- This means that if the particular assumption is not too far from true, then the technique is still approximately valid.
- *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 common mistake in using statistics***.
- Sometimes researchers check only some of the assumptions, perhaps missing some of the most important ones.

General rules

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.**

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

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 that can help sometimes:
  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**.

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

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

Statistical study design
POWER ANALYSIS

- How large the sample is required for accurate and reliable statistical inference?

- How likely is it that a statistical test detected an effect in a specific situation?

Specific suggestions for planning research:

- **Decide what questions you will be studying.**
  Trying to study too many things at once is likely to create problems with multiple testing, so it may be wise to limit your study.
- **If you will be gathering data, think about how you will gather *and* analyze it *before* you start to gather the data**.
- **Read reports on related research, focusing on problems that were encountered and how you might get around them and/or how you might plan your research to fill in gaps in current knowledge in the area.**
- **If you are planning an experiment, look for possible sources of variability and design your experiment to take these into account as much as possible.**
- **The design will depend on the particular situation.**
- The literature on design of experiments is extensive; consult it.
- Remember that the design affects what method of analysis is appropriate.

Specific suggestions for planning research:

- **If you are gathering observational data, think about possible confounding factors and plan your data gathering to reduce confounding.**
- Be sure to record any time and spatial variables present, or **any other variables that might influence outcome**, whether or not you initially plan to use them in your analysis.
- **Also think about any factors that might make the sample biased.**
- You may need to limit your study to a smaller population than originally intended.
- **Think carefully about what measures you will use.**
- If your data gathering involves asking questions, put careful thought into choosing and phrasing them. Then check them out with a test-run and revise as needed.

Specific suggestions for planning research:

- Think about how you will randomize (for an experiment) or sample (for an observational study).
- Think about whether or not the model assumptions of your intended method of analysis are likely to be reasonable.
  - If not, revise either your plan for data gathering or your plan for analysis, or both.
- Conduct a pilot study to trouble shoot and obtain variance estimates for a power analysis.
  - Revise plans as needed.
- Decide on appropriate levels of Type I and Type II error, taking into account consequences of each type of error.
- Plan how to deal with multiple inferences, including "data snooping" questions that might arise later
- Do a power analysis to estimate what sample size you need to detect meaningful differences.
- Take into account any relevant considerations such as multiple inference, Intent-to-Treat analysis or how you plan to handle missing data.
  - Revise plans as needed.

Specific suggestions for analyzing data

- Before doing any formal analysis, ask whether or not the model assumptions of the procedure are plausible in the context of the data.
- Plot the data (or residuals, as appropriate) as possible to get additional checks on whether or not model assumptions hold.
- If model assumptions appear to be violated, consider transformations of the data, or use alternate methods of analysis as appropriate.
- **If more than one statistical inference is used, be sure to take that into account by using appropriate methodology for multiple inference.**
- **If you use hypothesis tests, be sure to calculate corresponding confidence intervals as well.**
- But be aware that there may also be other sources of uncertainty not captured by confidence intervals.
- Keep careful records of decisions made in data cleaning and in using software.
- **Until a happier future arrives, imperfections in models require further thought, and routine disclosure of imperfections would be helpful.**

Aim for transparency and reproducibility

- Include enough detail so the reader can critique both the data gathering and the analysis.
- Look for and report possible sources of bias or other sources of additional uncertainty in results.
- Include enough detail so that another researcher could replicate both the data gathering and the analysis.
- **Include discussion of why the analyses used are appropriate**
  - i.e., why model assumptions are well enough satisfied for the robustness criteria for the specific technique, or whether they are iffy.
- **If you do hypothesis testing, be sure to report p-values (rather than just phrases such as "significant at the .05 level") and also give confidence intervals.**
- **If you have built a model, be sure to explain the decisions that went into the selection of that model**

THANK YOU FOR YOUR ATTENTION!