t-Tests, Type I and II Errors Supplement

**Independent Samples T-tests (Between Subjects t-test)** are done when a researcher wishes to compare two distinct group or subgroup means. For example, the researcher may wish to compare the means of the male and female participants in a group. In this case, gender would become the independent variable being tested. This is the most widely used statistical tests, not just in nursing, but in all disciplines that use quantitative analysis.

**Paired or Dependent Samples T-tests (Within Subjects t-test)** are used when a researcher wishes to compare two means collected from the same sample, but at different times. An example would be to compare pretest and post test scores in a class.

T tests were specifically developed to use with smaller samples, usually of 30 or less in each group. In order to calculate the t statistic, you must calculate the mean, standard deviation (SD or S) and the variance ($S^2$).

**Assumptions and Rules when using the t-test**

- The T test assumes that the research involves **the estimation of at least one variable**.
- It also assumes that **the groups are independent** and that **nothing in one group helped to determine who was in the second group**.
- If the groups are related, and you still want to compare differences, a **variation** of the t test, called a **paired t** or a **correlated t test** can be used.
- Variables must have been measured at the **interval or ratio level**.
- It is also assumed that the variable being studied is **equally distributed** or **homogenous** in the overall population.
- The sample was selected **randomly**

**Type I and II errors**

- **Type I error**
  
  Reject a null hypothesis that is really true (with tests of difference this means that you say there was a difference between the groups when there really was not a difference). The probability of making a Type I error is the alpha level you choose. If you set your probability (alpha level) at $p < 0.05$, then there is a 5% chance that you will make a Type I error. You can reduce the chance of making a Type I error by setting a smaller alpha level ($p < .01$). The problem with this is that as you lower the chance of making a Type I error, you increase the chance of making a Type II error.

- **Type II error**
Fail to reject a null hypothesis that is false (with tests of differences this means that you say there was no difference between the groups when there really was one)

- **Non directional (two-tailed)**

  **Research Question:** Is there a (statistically) significant difference between males and females with respect to math achievement?

  **H0:** There is no (statistically) significant difference between males and females with respect to math achievement.

  **HA:** There is a (statistically) significant difference between males and females with respect to math achievement.

- **Directional (one-tailed)**

  **Research Question:** Do males score significantly higher than females with respect to math achievement?

  **H0:** Males do not score significantly higher than females with respect to math achievement.

  **HA:** Males score significantly higher than females with respect to math achievement.

The basic idea for calculating a t-test is to find the difference between the means of the two groups and divide it by the STANDARD ERROR (OF THE DIFFERENCE) -- which is the standard deviation of the distribution of differences.

**CONFIDENCE INTERVAL** for a two-tailed t-test is calculated by multiplying the CRITICAL VALUE times the STANDARD ERROR and adding and subtracting that to and from the difference of the two means.

**Inferential Statistics: What Inference Is and How It Works**

An inference is the act of arriving at a conclusion that has some probability of being true, based on available facts. In statistics, an inference is a conclusion made about a population based on facts obtained from a sample. Inferential statistics are used when two or more data sets are compared.
• Two groups of subjects, one of which has received a research intervention
• Two groups of subjects, one of which possesses a certain characteristic
• One group of subjects at time A and time B

Inferential statistics applied to any of these cases allow the researcher to decide whether the two measurements differ.

Describing the relationship between two variables allows the researcher to make predictions.

Predictions Made Based On Variable Relationships

• Which patients are more or less likely to develop a certain complication, depending on the length of their hospitalization
• How many teaching sessions will be necessary in order for a patient to master self-care skills, depending on overall level of functioning
• The time interval during which a postsurgical patient is most likely to experience postoperative diuresis, depending on age

Making Inferences: Mom and Vacuuming

A stay-at-home mother moves the portable television out of the living room and into the breakfast room, and then notices a month later that she can get by with vacuuming crumbs out of the living room carpet weekly instead of every other day. She infers that:

• The children are watching TV as they graze, scattering crumbs while transfixed by the tube
• Keeping the TV in the breakfast room will continue to decrease the frequency with which she has to vacuum the living room

Inferential Statistics: Making Inferences

With a deck of playing cards, the hand dealt is a sample of the entire deck (population). If a person were dealt four cards and two of them were hearts, the person might come to the erroneous conclusion—the erroneous inference—that a standard deck of cards contains 50% hearts. An estimate of the population, given almost any sample, will almost always be slightly off. Statistics define the margin of error and provide decision rules that make good inferences—not great ones, but good ones.
Inferential Statistics: P-values

Some chance of error always exists. In research, all statistical tests are conducted with a certain probability of error, and that amount of error is selected by the researcher. This probability of error is expressed as "p (probability) < .xx".

In nursing research, the usual probability of statistical error is set at 5% (expressed as p < .05), although some studies use a lower margin of error. This means that there is less than a 5% chance of error when the statistics reveal that a statistically significant difference exists. Statistical significance is the measure of how large a chance there is of Type I error—a measure of how "true" the results are likely to be.

A statistical test conducted at the .05 level of significance has a confidence level, then, of 1 minus alpha = 95%. Therefore, one can be 95% confident that the result is correct.

In the next topic, we will look at some specific types of statistical inferential tests.

ERROR

Type I error: The null hypothesis has been INCORRECTLY rejected.

Type II error: The null hypothesis has been INCORRECTLY accepted.

Type I and II Errors: Type I Error and Type II Error: How They Relate

Error happens. It is tempting to conclude that using a more stringent level of significance, say a lower level like p < .01, would be the best solution. This essentially makes it much harder to arrive at significant results. In the real world, this means conducting a larger study with more subjects. However, using a lower level of significance isn't the answer to avoiding all types of error. The .05 level of significance is usually acceptable for nursing research, which typically has small budgets and small numbers of qualified and willing subjects to draw upon.
Null Hypothesis

<table>
<thead>
<tr>
<th>Null hypothesis is not rejected</th>
<th>Null hypothesis is rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis is true</td>
<td>Correct decision!</td>
</tr>
<tr>
<td>Null hypothesis is false</td>
<td>Type I error</td>
</tr>
<tr>
<td>Type II error</td>
<td>Correct decision!</td>
</tr>
</tbody>
</table>

In statistical analysis, two types of errors are possible:

- **Type I error**: The null hypothesis has been INCORRECTLY rejected.
- **Type II error**: The null hypothesis has been INCORRECTLY accepted.

Statistical testing uses the null hypothesis—the opposite of the working hypothesis that the researcher has in mind. The null hypothesis is used because research can never prove that something is true; it can only disprove, or prove that something is false.

Because of this, a null hypothesis is created, expressing that there is no difference:

- Between groups
- Between time A and time B

This way, if the null hypothesis is disproved, the researcher knows that there is a difference between these two factors.

Type I error is what the probability value represents. With a p < .05 level of significance, there is a 5% chance of concluding that the null hypothesis is incorrect when it really is correct.

When using the .01 level of significance, researchers minimize Type I error but increase Type II error. When using the .05 level of significance, the risk of a Type I error decreases and the risk of a Type II error increases.

The p-level is also called the "alpha" and represents the chance of Type I error. Concluding that the null hypothesis is incorrect when it really is correct is an example of Type I error.

**Type I and II Errors: Effect Size**

The effect size is the magnitude of the difference in a variable, in comparison with the standard deviation, when that variable is measured under two conditions:

- In experimental or quasi-experimental research, the effect size is the amount of change an independent
variable produces in a dependent variable in comparison with the variable's standard deviation

- In nonexperimental research, the effect size is the change in value of one variable, in comparison with its standard deviation, that occurs when a different variable's value changes

Sad but true fact: As the level of significance becomes smaller and smaller, minimizing Type I error, the likelihood of Type II error increases, and vice versa.

**Effect Size**

Correlational research measures the difference in number of daytime awakenings for shift workers in the presence of "white noise" produced by a fan and in the absence of "white noise." Daytime sleep in shift workers without a fan may have an average of 4.15 awakenings in eight hours. If the standard deviation is 1.8, then the decrease in "with fan" sleep awakenings from a mean of 4.15 to a mean of 2.95 would be 1.2 divided by 1.8 = 0.67. This would be somewhere between a moderate effect and a large effect.

**Type II Error**

A medication for arthritis decreases average daily pain from 5 to 4.3. This is a small effect size. The researcher will use a .05 level of significance and many subjects to try to detect this little effect.

A medication for panic attacks decreases average weekly frequency from 21.7 to 3. This is a large effect size. The researcher may choose to use a level of significance lower than .05, and not very many subjects will be required to detect this huge effect.

**Effect Size and Type II Error**

Type II error occurs when the statistical test doesn't reach significance and the null hypothesis is not rejected, even though the null hypothesis is actually false. There is a difference, but the statistical test didn't detect that difference.

The researcher can minimize Type II error in the following ways:

- Use more research subjects
- Use a higher (less stringent) level of significance

Type II error is more likely when the amount of change the researcher is measuring is small. This means that when the effect size is tiny, the researcher must use a large number of subjects to detect this little difference and will rarely use any lower level of significance than .05.
Type I and II Errors: Power Analysis

Power analysis is a mathematical computation that reveals the likelihood of detecting a significant difference (if one exists) with a given effect size, sample size, and level of significance. Use of a power analysis can assist a researcher in determining the ideal sample size to use at a given level of significance with an established effect size.

Power should be at least 0.8 to detect a difference when it does exist. Type II error occurs when use of a too-small sample leads to the erroneous conclusion that no difference exists. This is referred to as "The Nursing Error," because so many nursing studies contain samples too small to detect a difference. If the sample is too small, power will fall below the acceptable 0.8 level.

Relative Age of Students Who Become Mothers in High School

A nurse working in an urban high school has the impression that the students who become mothers while in high school are generally a bit more than half a year older than they used to be. The nurse finds, upon review of records she has maintained, that the average age of students who gave birth in the past six school terms has been 16.1 years. She calculates the standard deviation and finds that it is 0.7 years. The nurse decides to perform a t-test to determine whether the mean now is significantly different from the mean previously observed. However, the nurse now has only 12 pregnant students, as of December, or students who have become mothers this academic year. Based on past experience, she anticipates another 10 between now and June. Is a sample size of 22 large enough to show statistical significance at an acceptable power level, and can the nurse complete this research during the academic year or will she have to make it a two-year study?

From the power analysis that the nurse performs online, she learns that for a power level of 0.8 (considered acceptable), an anticipated difference in means of 0.5, and a standard deviation of 0.7, she will need at least 18 subjects in order to detect a difference between the current students and the former students, if there really is a difference. The nurse concludes that she will most likely be able to complete her research within this academic year.

The t-Test

$t$-Tests compare one set of data with another to determine whether the two sets are from the same population or whether they are sufficiently different to be able to conclude that the two data sets are from different populations. $t$-Tests can be used to compare two groups, to compare one newly emergent group with a large known population, or to compare one group at two different times to determine whether:

- The research intervention makes a difference in a certain measurement
**Intervention Makes a Difference**

The experimental group receives hyperbaric treatments five times a week, and the control group receives hyperbaric treatments twice a week. A *t*-test reveals that wound healing occurs significantly more rapidly in the experimental group.

- Whatever happened between time A and time B makes a difference in a certain measurement

**Between Time A and Time B Makes a Difference**

Persons with seasonal allergies are measured for performance on tests of concentration at time A. Then they are given a new medication for allergy and measured for performance on tests of concentration at time B. A paired *t*-test reveals that performance is significantly worse at time B. (The new medication, although it may help with allergies, also interferes with performance on tests of concentration.)

- A data set is part of a large population or whether it is significantly different

**Data Set**

The duration of an average human pregnancy is 267 days. A new method of in vitro fertilization is developed that combines two ova instead of using the traditional ovum-sperm combination. Pregnancies that result from this new method have a duration of pregnancy that is shorter than average. *T*-tests reveal that while pregnancies are shorter than average, this is not a large enough difference to be statistically significant.

- If two parts of a group are different after the research question has been answered

*t*-Tests that decide whether a data set is significantly different from a certain population, in which the standard deviation and mean are known, can be especially useful for tracking changing trends in vital statistics.

**Two Parts of a Group Are Different**

After the research question, "Does premedication with diphenhydramine have an additive effect in decreasing pain when used with procedural sedation in the emergency room?" is answered, the researcher performs a *t*-test to determine whether there is a disproportionate number of burn patients in the control group as compared with the experimental group. (Burn patients may experience more pain than others requiring procedural sedation.) The *t*-test shows that even though there is a disproportionate number of burn patients in the control group, their pain scores do not differ significantly from those of other emergency room patients.

**The *t*-test: Example**

A researcher working with the national health service in Egypt wonders whether the average age at death among the adult population has changed in the town of Ahwr over the past year. He
checks the database and finds that the average age at death in Ahwr two years ago was 60.6, three years ago was 61.4, and four years ago was 60.2.

His database provides him with the standard deviations for these data: 15.7, 15.4, and 16.4 for the three years, respectively. View the data set below.

Data Set for $t$-test


He compares this past year's data with data of two years ago.

The formula for the $t$-test is:

\[
\frac{\text{mean of the sample} - \text{population mean}}{\text{standard deviation of the sample divided by the square root of } n}
\]

If the calculated value of $t$ is below -1.9867 or greater than 1.9867, the researcher will conclude that the past year's average age at death for adults is significantly different from previous years and that age at death is indeed changing.

\[
\frac{\text{mean of the sample} - \text{population mean}}{\text{standard deviation divided by the square root of } n}
\]

\[
= \frac{64.652 - 60.6}{16.6/(89)^{1/2}}
\]

\[
= \frac{4.052}{1.760} = 2.303
\]

The researcher can conclude that the average age at death for adults in Ahwr is indeed changing, and it is increasing.

$t$-tests

Six months after a poison gas attack, survivors are tested for anemia. Fifteen male human subjects are randomly selected for participation in the study from a group of 500 survivors. Mean hematocrit for males in the human population is 46.

Data set:

40.4, 42.3, 43.1, 44.1, 44.2, 44.9, 45.3, 45.6, 46.1, 46.8, 47.2, 47.2, 48.4, 50.5, 51.9
A t-test is conducted:

- .05 level of significance
- 14 degrees of freedom
- 3.00 standard deviation
- square root of 15 = 3.87

Remember that the formula is:

\[
\frac{\text{mean of the sample} - \text{population mean}}{\text{standard deviation of the sample divided by the square root of "n"}}
\]

If the \( t \) value is < -2.1448 or > 2.1448, the sample is significantly different from the population. Obtaining a value that extreme would mean that this sample of men exposed to poison gas has a significantly different hematocrit than do normal men, and so poison gas affects hematocrit 6 months after exposure.

If the \( t \) value falls between -2.1448 and 2.1448, the sample is not significantly different from the population. Obtaining a value within this range would mean that the sample of men exposed to poison gas has roughly the same hematocrit as do normal men, and so poison gas does not affect hematocrit 6 months after exposure.