Statistiek I

$t$-tests

John Nerbonne

CLCG, Rijksuniversiteit Groningen

http://www.let.rug.nl/nerbonne/teach/Statistiek-I/
Overview

1. Basics on $t$-tests
2. Independent Sample $t$-tests
3. Single-Sample $t$-tests
4. Paired Sample $t$-tests
5. Summary of $t$-tests
6. Discussion: Multiple Tests, Effect Size
To test an average or pair of averages when $\sigma$ is known, we use $z$-tests.

But often $\sigma$ is unknown, e.g., in specially constructed psycholinguistics tests, in tests of reactions of readers or software users to new books or products.

In general, $\sigma$ is known only for standard tests (IQ tests, CITO tests, ...).

$t$-Tests incorporate an “estimation” of $\sigma$ (based on the standard deviation SD of the sample in order to reason in the same way as in $z$-tests.)
Different \textit{t}-Tests

\textbf{Student’s \textit{t}-Test} (’Student’ pseudonym of Guiness employee without publication rights)

- three versions:
  - \textbf{independent samples} compares two means
determine whether difference is significant
  - \textbf{single sample} (estimate mean)
  - \textbf{paired data} compares pairs of values
  - example: two measurements on each of 20 patients

- population statistic $\sigma$ unnecessary
  of course, sample statistics need

- appropriate with numeric data “roughly normally distributed”
  see Mann-Whitney U-Test, Wilcoxon signed rank test for non-parametric fall-backs
The $t$ Statistic

$t$ statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Note that $s$ is used (not $\sigma$). Of course, $s$ should be a good estimate. Cf. $z$ test.

$n$ is (usually) the number of items in the sample

Always used with respect to a number of degrees of freedom, normally $n - 1$ (below we discuss exceptions)

To know the probability of a $t$ statistic we refer to the tables (e.g., M&M, Table E, Field, Table A2). We have to check on $P(t(df))$, where $dF$ is the degrees of freedom, normally $n - 1$. 
Given degrees of freedom, dF, and chance of \( t \leq p \), what \( t \) value is needed?

<table>
<thead>
<tr>
<th>dF/p</th>
<th>0.05</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>...</td>
<td>1.8</td>
<td>2.8</td>
</tr>
<tr>
<td>20</td>
<td>...</td>
<td>1.7</td>
<td>2.5</td>
</tr>
<tr>
<td>30</td>
<td>...</td>
<td>1.7</td>
<td>2.5</td>
</tr>
<tr>
<td>40</td>
<td>...</td>
<td>1.7</td>
<td>2.4</td>
</tr>
<tr>
<td>100</td>
<td>...</td>
<td>1.660</td>
<td>2.364</td>
</tr>
</tbody>
</table>

\[ z \quad \ldots \quad 1.645 \quad 2.326 \quad 3.091 \quad \ldots \]

Note comparison to \( z \). For \( n \geq 100 \), use \( z \) (differences negligible).

Compare \( t \)-tables. Be able to use table both to check \( P(t) \) and, for a given \( p \), to find min \( t \mid P(t) \leq p \)
Suppose you mistakenly used \( s \) in place of \( \sigma \) in a \( z \) test with 10 or 20 elements in the sample, what would the effect be?

<table>
<thead>
<tr>
<th>dF/p</th>
<th>0.05</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>...</td>
<td>1.8</td>
<td>2.8</td>
</tr>
<tr>
<td>20</td>
<td>...</td>
<td>1.7</td>
<td>2.5</td>
</tr>
</tbody>
</table>

\[ z \ldots 1.65\ 2.33\ 3.09\ \ldots \]

You would, e.g., treat a differences of +2.33\( s \) as significant at the level 0.01 level (only 1\% likely), when in fact you need to show differences of +2.8 or +2.5, respectively, to prove this level of significance.

Applying a \( z \) test using \( s \) instead of \( \sigma \) **overstates** the significance of the results (and increases the chance of a type I error, and a false positive decision on the \( H_a \)).
Independent Sample $t$-Tests

Two samples, unrelated data points (e.g., not before-after scores).

Compares sample means, $\bar{x}_1$, $\bar{x}_2$, wrt significant difference.

$H_0$ is always $\mu_1 = \mu_2$, i.e., that two populations have the same mean. Two-sided alternative is $H_a : \mu_1 \neq \mu_2$ We use $\bar{x}_1$, $\bar{x}_2$ to estimate $\mu_1$, $\mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \cdot \sqrt{1/n_1 + 1/n_2}}$$

Degrees of freedom $dF = \text{Min}\{(n_1 - 1), (n_2 - 1)\}$ (using a smaller number makes showing significance harder, and therefore more reliable). **Notate bene**: S+ (and other statistical packages) often deviate from this conservative recommendation, using (something close to) $dF = (n_1 - 1) + (n_2 - 1) = n - 2$ (legitimate).

$t$ increases with large diff. in means, or with small standard deviations (like $z$).
Independent Sample $t$-Test: Assumptions

Assumptions of Independent Sample $t$-tests (M&M, §7.1.5; Field, §9.3.2)

1. Exactly two samples **unrelated**
2. Distribution roughly normal. Field is vague on this, but rules of thumb are:
   - normality required if $n < 15$
   - no large skew or outliers if $n < 40$

Equal variances (sd’s) often mentioned as condition, but easily corrected for (Welch test).

If three or more samples, use ANOVA (later in course, Stats II).

If distribution unsuitable, using Mann-Whitney (later in course).
Example: You wish to know whether there’s a difference in verbal reasoning in boys vs. girls. There are tests, but no published $\sigma$. You test for a difference in average ability in these two samples.

Assume two variables, \text{VReason, Sex}

<table>
<thead>
<tr>
<th></th>
<th>VReason</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>M</td>
</tr>
<tr>
<td>19</td>
<td>71</td>
<td>F</td>
</tr>
<tr>
<td>20</td>
<td>82</td>
<td>F</td>
</tr>
</tbody>
</table>

Two independent samples (no two scores from same person).
Example t-Test: One-Sided or Two-Sided?

**Example:** You wish to know whether there’s a difference in verbal reasoning in boys vs. girls. There are tests, but no published $\sigma$. You test for a difference in average ability in the two samples.

No question of one being *better* than the other.

This is a **two-sided question**.

Hypotheses:

\[ H_0 : \mu_m = \mu_f \]
\[ H_a : \mu_m \neq \mu_f \]

What would hypotheses be if we asked whether boys are better than girls?
Independent Sample $t$-Test: Normality Test

$n_1 = n_2 = 10$, so for $t$-test, distributions must be roughly normal. Are they?

Boys

Girls
Normal Quantile Plots

Plot expected normal distribution quantiles (x axis) against quantiles in samples. If distribution is normal, the line is roughly straight. Here: distribution roughly normal.

More exact techniques: check KOLMOGOROV-SMIRNOV GOODNESS OF FIT or SHAPIRO-WILKS TEST (both uses $H_0$: distribution normal). If rejected, alternative tests are needed (e.g., Mann-Whitney).
Visualizing Comparison

Box plots show middle 50% in box, median in line, 1.5 interquartile range (up to most distant).
Independent Sample $t$-Test: Example

But is difference statistically significant?

Invoke (in S+)

1. Compare samples $\rightarrow$ Two Samples $\rightarrow$ $t$-Test
2. Specify variable $V_{\text{Reason}}$, groups $\text{Sex}$

Calculates $t$, $dF$, \textbf{two-tailed} probability ($p$-value)
Results (in S+) 

Welch Modified Two-Sample t-Test 

data:  x: VReasn with Sex = M , and y: VReasn with Sex = F  
t = 1.7747, df = 17.726, p-value = 0.0931  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
  -1.166146  13.766146  
sample estimates:  
  mean of x mean of y  
    79.8    73.5  

Why “Welch”? — No assumption of equal variances.
Results (in $S+$)

Welch Modified Two-Sample t-Test

data:  x: VReasn with Sex = M , and y: VReasn with Sex = F

t = 1.7747, df = 17.726, p-value = 0.0931
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.166146   13.766146
sample estimates:
  mean of x  mean of y
    79.8       73.5

Why “Welch”? — No assumption of equal variances.
Results (in SPSS) (slightly diff. data)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys/Girls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal Reas.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys</td>
<td>10</td>
<td>79,8</td>
<td>7,96</td>
<td>2,52</td>
</tr>
<tr>
<td>Girls</td>
<td>10</td>
<td>73,5</td>
<td>7,43</td>
<td>2,35</td>
</tr>
</tbody>
</table>

Independent Samples Test

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>Sig.</th>
<th>t</th>
<th>df</th>
<th>Sig.</th>
<th>Mean Diff.</th>
<th>SE</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal var.</td>
<td>,053</td>
<td>,821</td>
<td>1,83</td>
<td>18</td>
<td>,084</td>
<td>6,3</td>
<td>3,44</td>
<td>-93</td>
<td>13,5</td>
</tr>
<tr>
<td>No Equal var.</td>
<td>1,83</td>
<td>17,9</td>
<td>,084</td>
<td>6,3</td>
<td>3,44</td>
<td>-93</td>
<td>13,5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

John Nerbonne
Results

Levene’s test | t-test for Equality of Means

| Equal var.   | 0.053 | 0.821 | 1.83 18 | 0.084 | 6.3 | 3.44 | 3.93 13.5 |
| No Equal var.| 1.83 17.9 | 0.084 | 6.3 | 3.44 | 3.93 13.5 |

Note that some tables give values for one-tailed tests (Moore & McCabe; Field gives both one-tailed and two-tailed probabilities). One-tailed values ≈ 1/2 those of SPSS.

Degrees of Freedom ≈ \((n_1 - 1) + (n_2 - 1)\) (less conservative than book)

Equal Variance vs. No Equal Variance usually unimportant — even when variances are very different.

How do you understand the 95% CI = (-0.9, 13.5)?
### Effect Size

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Reas.</td>
<td>Boys</td>
<td>10</td>
<td>79.8</td>
<td>7.96</td>
<td>2.52</td>
</tr>
<tr>
<td></td>
<td>Girls</td>
<td>10</td>
<td>73.5</td>
<td>7.43</td>
<td>2.35</td>
</tr>
</tbody>
</table>

**Cohen's d**: \( \frac{(m_1 - m_2)}{sd_{all}} = \frac{(79.8 - 73.5)}{10.9} \approx 0.6sd \)

Cohen's d: How many standard deviations apart are the two means?

For interpretable differences, effect size may be reported practically, e.g. in money spent, survival rates, time saved, etc.
Reporting Results

We tested whether boys and girls differ in verbal reasoning. We selected 10 healthy individuals of each group randomly, and obtained their scores on *<Named> Verbal Reasoning Assessment*. We identify the hypotheses:

\[ H_0 : \mu_m = \mu_f \text{ (male and female the same)} \]
\[ H_a : \mu_m \neq \mu_f \text{ (male and female different)} \]

Since \( \sigma \) is unknown for this test, we applied a two-sided \( t \)-test after checking that the distributions were roughly normal. We discuss one outlier below.

Although the difference was large (Cohen’s \( d = 0.6 \) sd, see box-and-whiskers plots), it turned out not to reach significance, \( t(18) = 1.83, p = 0.084 \), so we did **not** reject the null hypothesis \( p \leq 0.05 \). We retain \( H_0 \).

Questions

1. Given the small sample, and low outlier in the higher group, we might confirm \( H_a \) by recalculating, eliminating this individual. Should we?
2. Have we reported effect size?
We tested whether boys and girls differ in verbal reasoning. We selected 10 healthy individuals of each group randomly, and obtained their scores on <Named> Verbal Reasoning Assessment. We identify the hypotheses:

\[ H_0 : \mu_m = \mu_f \] (male and female the same)
\[ H_a : \mu_m \neq \mu_f \] (male and female different)

Since \( \sigma \) is unknown for this test, we applied a two-sided \( t \)-test after checking that the distributions were roughly normal. We discuss one outlier below.

Although the difference was large (Cohen’s \( d = 0.6 \) sd, see box-and-whiskers plots), it turned out not to reach significance, \( t(18) = 1.83, p = 0.084 \), so we did **not** reject the null hypothesis \( p \leq 0.05 \). We retain \( H_0 \).

**Questions**

1. Given the small sample, and low outlier in the higher group, we might confirm \( H_a \) by recalculating, eliminating this individual. Should we?
2. Have we reported effect size?
Reporting Results — Guidelines

1. State the issue in terms of the populations (not merely the samples). 
   Formulate $H_0$ and $H_a$.

2. State how your hypothesis is to be tested, how samples were obtained, 
   what procedures (test materials) were used to obtain measurements.

3. Identify the statistical test to be used, why.

4. Illustrate your research question graphically, if possible. 
   —For example, with box plots, as above.

5. Present the results of the study on the sample, their significance level, 
   and an effect size.

6. State conclusions about the hypotheses.

7. Discuss and interpret your results.

Practice this in laboratory exercises!
One-Sided $t$-Test

If testing directional hypothesis, e.g., that boys are \textbf{better} than girls in 3-dim. thinking, then one can divide 2-tailed prob. obtained above by 2. (Since $0.09/2 < 0.05$, you could conclude immediately that the null hypothesis is rejected at the $p = 0.05$-level.)

But you can avoid even this level of calculation, by specifying the one-sided hypothesis in S+.

\begin{verbatim}
Welch Modified Two-Sample t-Test
data:  x: VReasn with Sex = M , and y: VReasn with Sex = F
t = 1.7747, df = 17.726, p-value = 0.0466
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval: 
  0.1392472    NA
...
\end{verbatim}
Single Sample $t$-Tests

Moore & McCabe introduce paired sample $t$-tests using single sample $t$-tests (not focus here, but useful below)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where $dF = n - 1$. Recall that $t$ increases in magnitude with large diff. in means, or with small standard deviations.

Use e.g. to test whether $\mu$ has a particular value, $\mu_0$.

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_a : \mu \neq \mu_0$$

Then $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, where in general, scores of large magnitude (positive or negative) indicate large differences (reason to reject $H_0$).
Single Sample $t$-Tests

**Example:** Claim that test has been developed to determine “EQ” (Emotional IQ). Test shows that $\mu = 90$ (in general population), no info on $\sigma$. We test:

$$H_0 : \mu = 90$$
$$H_a : \mu \neq 90$$

Measure 9 randomly chosen Groningners ($df = n - 1 = 8$). Result: $ar{x} = 87.2, s = 5$ (Could the restriction to Groningen be claimed to bias results? ;)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{87.2 - 90}{5/\sqrt{9}} = \frac{-2.8}{1.6} = -1.75$$

*Two-tailed chance of* $t = -1.75$ ($df = 8$, Field, Tbl A.2) $\gg 0.05$, Cohen’s $d = -2.8/5 = -0.56$ (but note that we’re interested mostly in the size, not the sign).
Interpreting Single Sample $t$-Test

$H_0 : \mu_{EQ} = 90$
$H_a : \mu_{EQ} \neq 90$

The value $-1.75$ falls within the central 90% of the distribution.

Result: insufficient reason for rejection. Retain $H_0$.

Discussion: The small sample gives insufficient reason to reject the claim (at $p = 0.05$ level of significance) that the test has a mean of 90.
Single Sample $t$-Tests

**Example 2:** Confidence Intervals

As above, check system for “EQ”. Measure $\bar{x}$ and derive 95% confidence interval.

1. Measure 9 randomly chosen Groningers ($df = n - 1 = 8$). Result: $\bar{x} = 87.2$, $s = 5$

2. Find (in table) $t^*$ such that $P(|t(8)| > t^*) = 0.05$, which means that $P(t(8) > t^*) = 0.025$. Result: $t^* = 2.3$

3. Derive 95%-Confidence Interval

$$= \bar{x} \pm t^*(s/\sqrt{n})$$

Practice in SPSS Lab 2!
Calculating Confidence Intervals with $t$

$P(t(8) > 2.3) = 0.025$ The $t$-based 95%-confidence interval is

$$= \bar{x} \pm t^*(s/\sqrt{n})$$

$$87.2 \pm 2.3(5/\sqrt{9})$$

$$87.2 \pm 3.7 = (83.5, 90.9)$$

Subject to same qualifications as other $t$-tests.

- Sensitive to skewness and outliers (as is mean!) if $n < 40$. Look at data!
- Only use when distribution approximately normal when $n < 15$. Look at Normal-Quantile Plot.
Paired $t$-Tests

More powerful application of $t$-test possible if data comes in pairs. Then pairwise comparison of data points possible (instead of just mean).

Conceptually, this is very clever. All the sources of individual variation are effectively cancelled by examining data in pairs!

Naturally, many sources remain, e.g. measurement error, variation within a given individual.
**Paired $t$-Tests**

We examine the difference between scores. We can check the hypothesis that the scores are from the same populations by check whether the average differences tend to be zero.

\[ H_0 : \mu(x_i - y_i) = 0 \quad \text{and} \quad H_a : \mu(x_i - y_i) \neq 0 \]

This can be regarded as a **single sample** of differences. We get the calculations (in some statistics packages), by calculating a set of differences, then applying a single-sample $t$-test to check the hypothesis that the mean is zero.

Some packages have built-in paired $t$-tests, e.g., SPSS.
### Paired t-Test: Example

**Example:** Suppose you suspect that the test for verbal reasoning which prove boys better is flawed. You find a second. Now you look at results of *both tests* on 15 subjects.

<table>
<thead>
<tr>
<th>subj</th>
<th>test1</th>
<th>test2</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.6</td>
<td>7.3</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>10.2</td>
<td>9.1</td>
<td>1.1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>15</td>
<td>8.4</td>
<td>7.5</td>
<td>0.9</td>
</tr>
<tr>
<td>m</td>
<td>6.85</td>
<td>6.23</td>
<td>0.42</td>
</tr>
<tr>
<td>sd</td>
<td>3.2</td>
<td>2.94</td>
<td></td>
</tr>
</tbody>
</table>

If we (incorrectly) applied an independent-sample *t*-test, w. hypotheses:

\[
H_0 : \mu_{test1} = \mu_{test2} \\
H_a : \mu_{test1} \neq \mu_{test2}
\]

Then \( t \approx 0.83, p \approx 0.2 \), retaining \( H_0 \) at \( \alpha = 0.05 \).

No proof that tests are different.
Paired $t$-Tests

Paired $t$-test appropriate when there are two related samples (e.g., two measurements $x, y$ of the same people, or same cases). This is called PAIRED DATA, or REPEATED MEASURES. Then we examine differences between the two elements of the pairs:

$$H_0 : \mu(x_i - y_i) = 0 \text{ and } H_a : \mu(x_i - y_i) \neq 0$$

1. The differences between scores should be normally distributed in small samples ($n \leq 15$), and should have large skew or outliers in modestly sized samples ($n \leq 40$)
2. Paired $t$-test is inappropriate when scores differ in scale, e.g., when one score is %-percentage, and other in $[0, 600]$. Consider then REGRESSION.
Paired \( t \)-Test: Application & Results

We assume two sets of data, \( X, Y \). Let \( \delta_i \) be \( x_i - y_i \).

\[
t = \frac{\bar{\delta}}{s_\delta / \sqrt{n}}
\]

For data in last slide (p.24), \( s_\delta \approx 0.4 \)

\[
t = \frac{0.4}{0.4 / \sqrt{15}}
\]
\[
= \frac{0.4}{0.4 / 3.9}
\]
\[
= \frac{0.4}{0.1}
\]
\[
= 4
\]

For \( dF = 14, \ p \approx 0.000 \) (Table, Field, p.803) Cohen’s \( d = \bar{\delta}/s_\delta = 0.4/0.4 = 1 \) Large!
Paired $t$-Test: Interpretation

We reject the null hypothesis at $p \leq 0.001$-level. The tests do not yield the same results.

**Discussion:** The second test yields consistently lower scores.

Note: We were **not** able to prove the two tests different using the $t$ test for independent samples. But since we have two sets of test results for the same people, we can examine the data in pairs.

**General lesson:** More sophisticated statistics allow more sensitivity to data. Using the paired data, we could prove that the tests were different.

Using the independent-samples test commits *error of second sort*: null hypothesis false, but not rejected.
z- vs. $t$-tests, including paired data

- **$z$ test**
  - $\sigma$ known
    - 2 grp
  - $\sigma$ unknown
    - 3 grp

- **$t$-test**
  - compare
    - 2 averages
  - different subjects
  - same subjects
    - paired $t$-test
  - unrelated samples

- **$t$-test**
  - Independent Sample $t$-tests
  - Single-Sample $t$-tests
  - Paired Sample $t$-tests
  - Summary of $t$-tests
  - Discussion: Multiple Tests, Effect Size
Nonparametric Alternatives to paired $t$-Tests

If distribution nonnormal, recommended alternative is the WILCOXON SIGNED RANK TEST (treated later in this course).

If distribution nonnormal and asymmetric, we can note the sign of the differences and use those in the SIGN TEST, based on the BINOMIAL DISTRIBUTION (later in this course), which is sensitive only to a single difference, e.g. improvement vs. non-improvement in scores.

*Sign test* (Field, p.555; M&M, p.409) — sometimes necessary when small samples too skewed for a $t$-test
**t-Tests: Summary**

Simple $t$ statistic:

$$t = \frac{m_1 - m_2}{s/\sqrt{n}}$$

- for numeric data, compares means of two groups, determines whether difference is significant
- population statistics ($\sigma$) unnecessary, of course, sample statistics need
- three applications:
  - independent samples: compares two means
  - single sample (e.g., to estimate mean, or check hypothesis about mean)
  - paired: compares pairs of values
    - example: two measurements on each of 20 patients
t-Tests: summary

Assumptions with all t-tests

1. Exactly two samples *unrelated measurements*
2. Distribution roughly normal if $n < 15$
   No large skew or outliers if $n < 40$

Additionally

1. Pay attention to whether variance is equal (reported in SPSS automatically via Levene’s test)
2. Report effect size using Cohen’s $d = (m_1 - m_2)/s$

Nonparametric fallbacks:

- Independent samples $\rightarrow$ Mann-Whitney
- Paired $t$-test $\rightarrow$ Wilcoxon signed rank test
Multiple Tests

Applying multiple tests risks finding apparent significance through sheer chance.

**Example:** Suppose you run three tests, always seeking a result significant at 0.05. The chance of finding this in one of the three is Bonferroni’s **family-wise** \( \alpha \)-level

\[
\alpha_{FW} = 1 - (1 - \alpha)^n
\]

\[
= 1 - (1 - .05)^3
\]

\[
= 1 - (.95)^3
\]

\[
= 1 - .857 = 0.143
\]

To guarantee a family-wise alpha of 0.05, divide this by number of tests

**Example:** 0.05/3 = 0.017 (set \( \alpha \) at 0.1)

**Analysis of Variance** is better in these cases (topic later).
Effect Size and Sample Size

Statistical significance obtains when an effect is unlikely to have arisen by chance. Very small differences may be significant when samples are large, i.e., these small differences are (probably) not due to chance.

As we saw in discussion of \( z \), a difference of two standard errors or more (\( z \geq 2 \) or \( z \leq -2 \)) is likely to arise in less than 5% of the time due to chance.

<table>
<thead>
<tr>
<th>diff (in ( \sigma )'s)</th>
<th>( n )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>40,000</td>
<td>0.05</td>
</tr>
<tr>
<td>0.1</td>
<td>400</td>
<td>0.05</td>
</tr>
<tr>
<td>0.25</td>
<td>64</td>
<td>0.05</td>
</tr>
<tr>
<td>0.37</td>
<td>30</td>
<td>0.05</td>
</tr>
<tr>
<td>0.5</td>
<td>16</td>
<td>0.05</td>
</tr>
</tbody>
</table>

We recommend samples of “about 30” because small effect sizes (under 0.4\( \sigma \)) are uninteresting, unless differences are important (e.g., health).

Watch out for effect size when reading research reports!
Next Topic

\[ \chi^2 \] tests of independence