A tale of two Boxes.

Box A	Box B
Avg= μ_A	Avg= μ_B
SD= σ_A	SD= σ_B

(*) n draws are made at random with replacement from Box A and m draws are made at random with replacement from Box B

(*) The draws from the two boxes are made independently.

What can we expect the ***difference*** of the two sample averages to be?

(*) Sample average from box A: $\overline{x}_A \approx \mu_A \pm SE_A = \mu_A \pm \dfrac{\sigma_A}{\sqrt{n}}$

(*) Sample average from box B: $\overline{x}_B \approx \mu_B \pm SE_B = \mu_B \pm \dfrac{\sigma_B}{\sqrt{m}}$

$$\overline{x}_A - \overline{x}_B \approx \overbrace{\mu_A - \mu_B}^{\text{expected value}} \pm \overbrace{\sqrt{SE_A^2 + SE_B^2}}^{\text{chance error}} \quad \checkmark$$

Example: Box A has an average of 15 with an SD of 6 and Box B also has an average of 15, but with an SD of 9. If 200 tickets are drawn at random with replacement from Box A and 500 tickets are drawn at random with replacement from Box B, what is the likely size of the difference between the two sample averages?

- The expected difference is $15 - 15 = 0$.
- $SE_A = \dfrac{6}{\sqrt{200}}$ and $SE_B = \dfrac{9}{\sqrt{500}}$.
- The standard error for the difference is

$$SE_{\text{diff}} = \sqrt{SE_A^2 + SE_B^2} = \sqrt{\frac{36}{200} + \frac{81}{500}} \approx 0.585.$$

- The difference of the average is likely to fall in the range 0 ± 0.585.

(*) The difference between sample averages, drawn independently, at random with replacement from two boxes *approximately follows the normal curve* (if the numbers of draws are large enough).

- A 95%-confidence interval for the difference is given by $(-1.17, 1.17)$.

Two-Sample tests of significance

Problem: How do we determine whether the difference between the averages of two different samples is due to chance or due to a difference between the 'boxes' from which the samples were drawn?

Example. The NAEP (National Assessment of Educational Progress) administered tests in mathematics to a nation-wide sample of 17-year-olds in 1978 and then again in 2004. The average scores in the two samples were 300 in 1978 and 307 in 2004.

Is the seven-point difference in average scores due to chance? If not, what conclusions can we draw from these statistics?

(*) We can assume that the two samples are ***independent***: the first sample had no effect on how the second sample was selected.

(*) For simplicity's sake,* I will pretend that these were simple random samples of 10,000 students each, with

- $SD_{1978} = 100$,
- $SD_{2004} = 80$.

(*) Neglecting correction factors (why?), we find that

- $SE_{1978} = 1.0$,
- $SE_{2004} = 0.8$.

---

* In fact, the sampling procedure was more complicated, the samples were larger and I made up the SDs. But the assumption that the samples are independent is still valid and the average scores and standard errors are correct. The data came from the NAEP website:

http://nces.ed.gov/nationsreportcard.

To decide whether the difference between the averages is statistically significant, we perform a test of significance.

(*) **Null Hypothesis:** There is no overall difference between 1978 students and 2004 students. Any observed difference in sample averages is due to chance.

(*) **Alternative Hypothesis:** The difference is ***not*** due to chance. The NAEP test is detecting an actual difference between the students in 2004 and their 1978 counterparts.

(*) The null hypothesis gives us the expected value of the difference. The ***expected difference*** is 0.

(*) The alternative hypothesis says that the difference between the averages is $\neq 0$.

(*) The ***observed difference*** is $307 - 300 = 7$.

(*) The observed value of the test statistic is

$$\frac{\text{observed difference} - \text{expected difference}}{\text{SE for the difference}} = \frac{7-0}{\sqrt{1^2 + 0.8^2}} \approx 5.466.$$

(*) ***The probability histogram for the difference of the averages of two independent simple random samples is approximately normal, (if the sample sizes are both sufficiently large).***

(*) This means that the P-value can be read off the normal table in this case:

P-value $= P(|z| \geq 5.466) \approx 0$. (Two-sided test, why?)

(*) ***Conclusion:*** Reject the null hypothesis. The difference between the 1978 and 2004 performances of 17-year-olds on the NAEP mathematics exams is almost certainly ***not*** explained by chance. The difference in the scores is *highly significant.*

Comments:

(*) It is tempting to say that 2004 students are ‘better at math’ than 1978 students, but that conclusion is not warranted on the basis of this significance test alone.

(*) It is important to remember that *highly significant* doesn’t necessarily mean that the difference in average scores is either particularly *big* or *important*.

- The 7-point difference in average scores is *not* big on the scale of the scores themselves.
- It is also not clear that these findings reveal important changes in the quality of education or mathematical ability over the 26-year time frame covered by the study.

Controlled Experiments

Question: Is the observed difference between the control group and the treatment group due to chance, or is the 'treatment' having an effect?

Example. (Problem 6, page 519) During the 1983 NAEP mathematics survey, a group of five hundred 13-year-olds from the same school district were asked to solve the following word problem:

> *An army bus holds 36 soldiers. If 1,128 soldiers are being bused to a training site, how many buses are needed?*

Half the students were randomly selected to use calculators and the other half used pencil and paper. Eighteen of the calculator group (7.2%) and fifty-nine of the pencil-and-paper group (23.6%) found the right answer. Can the difference in the percentages of correct answers be explained by chance? Or did the calculators have a negative effect on the students' work?

We answer this question by thinking of it as a ‘two-sample’ problem. In this case though, both samples are drawn from the same box: the five hundred students in the experiment. Nonetheless, we will proceed as if the samples were selected independently of each other from two separate boxes, with replacement.

We have the following null and alternative hypotheses

Null Hypothesis: Using a calculator had no effect on the students’ work. Any difference in the sample percentages was due to chance variation.

Alternative hypothesis: Using a calculator had a negative effect on the students’ work.

On to the calculations...

(*) The *observed difference* between the percentages of success in the pencil-and-paper and the calculator samples is $23.6\% - 7.2\% = 16.4\%$

(*) The *expected difference*, predicted by the null hypothesis, is 0%.

(*) The standard errors are $SE_{\text{calc}} = \frac{\sqrt{0.072 \times 0.918}}{\sqrt{250}} \times 100\% \approx 1.626\%$ and $SE_{\text{pnp}} = \frac{\sqrt{0.236 \times 0.764}}{\sqrt{250}} \times 100\% \approx 2.686\%$.

(*) The standard error for the difference of the sample percentages is

$$SE_{\text{diff}} \approx \sqrt{(1.626\%)^2 + (2.686\%)^2} \approx 3.14\%.$$

(*) The test statistic is $z = \frac{\text{observed}-\text{expected}}{\text{standard error}} = \frac{16.4\%}{3.14\%} \approx 5.2$.

(*) The test statistic follows the normal distribution in this case, so the P-value is $P(z \geq 5.2) \approx 0$

(*) **Conclusion:** We reject the null hypothesis and conclude that calculators had a negative effect on the students' work.

Note: 1128/36=31.333, which is what many students gave as their answer, especially in the calculator group. Another common answer was 31. The correct answer is 32.

Comments:

- The samples in a controlled experiment are drawn *without replacement*, and usually represent a significant proportion of the population. The correction factors for standard errors in these cases are considerably less than 1, and not using them ***inflates*** the estimates of the standard errors for the individual sample statistics.

- The samples in a controlled experiment are also *dependent*. An individual assigned to the control group is ***not*** assigned to the treatment group. The standard error for the difference between two statistics (averages, percentages, etc.) coming from *dependent* samples is *higher* than the standard error for independent samples. So combining the standard errors of the two samples as if they were independent ***lowers*** the estimated value of the SE of the difference.

- The effects of these two errors ***offset each other***, resulting in a ***slightly conservative*** estimate for the standard error of the difference—it is a *little bit* larger than the true SE.

Example. (Kahneman and Twersky)

Doctors deciding how to treat lung cancer received information in one of two forms.

Form A:

> *Of* 100 *people having surgery,* 10 *will die during treatment,* 32 *will have died by the end of one year and* 66 *will have died by the end of five years. Of* 100 *people having radiation, none will die during treatment,* 23 *will have died by the end of one year and* 78 *will have died by the end of five years.*

Form B:

> *Of* 100 *people having surgery,* 90 *will survive the treatment,* 68 *will have survive one year or longer and* 34 *will have survive five years or longer. Of* 100 *people having radiation, all will survive the treatment,* 77 *will survive one year or longer and* 32 *will survive five years or longer.*

(*) 167 doctors were randomized into two groups: 80 received Form A and 87 received Form B.

	Form A	Form B
Favored surgery	40	73
Favored radiation	40	14
Total	80	87
Percent favoring surgery	50%	84%

(*) Is the difference between the percentages favoring surgery due to chance?

$$SE_A = \frac{\sqrt{0.5 \times 0.5}}{\sqrt{80}} \times 100\% \approx 5.6\%, SE_B = \frac{\sqrt{0.84 \times 0.16}}{\sqrt{87}} \times 100\% \approx 3.9\%$$

$$z = \frac{(84\% - 50\%) - 0\%}{\sqrt{(5.6\%)^2 + (3.9\%)^2}} \approx 5... \Rightarrow \ p \approx 0$$

Conclusion: The difference between the percentages favoring surgery is not due to chance error. Something about the way the information was presented affected the doctors' decisions.