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: $\overline{x}_B \approx \mu_B \pm SE_B = \mu_B \pm \dfrac{\sigma_B}{\sqrt{m}}$

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

$$\Rightarrow \qquad SE_{diff} = \sqrt{SE_A^2 + SE_B^2} \qquad \Leftarrow$$

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

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

$\Rightarrow$ The difference of the averages is likely to be $\approx 0 \pm 0.585$.

(*) *The difference between sample averages, drawn independently, at random with (or without)replacement from two boxes* ***follows the normal curve*** *(approximately) (if the numbers of draws are large enough).*

$\Rightarrow$ There is a a 95% chance that the difference of the sample averages will fall in the range $(-1.17, 1.17)$.

Two-Sample tests of significance

Two-Sample tests of significance

Problem: How do we determine whether the difference between the averages (or percentages) of samples drawn from two different populations is due to chance or due to a difference between the populations 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.

Box model: 2004-box: one ticket for every 17-year old (in 2004) with that student's score on the ticket. 1978-box: one ticket for every 17-year old (in 1978) with that student's score on the ticket. μ_{2004} = average of 2004-box and μ_{1978} = average of 1978-box.

(*) $H_0 : \mu_{2004} - \mu_{1978} = 0$ (the two averages are the same)

(*) $H_0 : \mu_{2004} - \mu_{1978} \neq 0.$ (the two averages are *not* the same)

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

(*) The observed value of the test statistic is

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

(*) ***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).***

⇒ The P-value can be read off the normal table in this case:

P-value $= P(|z| \geq 5.466) \approx 0.$

(*) ***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 (statistically) significant*.

Comments:

(*) 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.

(*) ***Statistical Significance*** means that the observed difference between the sample averages is not (likely to be) explained by chance error. *Nothing else.*

Example. A researcher studying the media consumption habits of U.S. adults believes that women watch more hours of cooking shows than men. To test this hypothesis, she...

(*) Formulates her hypotheses:

$H_0: \ \mu_w - \mu_m = 0,$ (μ_w and μ_m are the average numbers of hours

$H_A: \ \mu_w - \mu_m > 0.$ of cooking shows watched per week by women and men, respectively.)

(*) Determines the test statistic and its probability distribution:

$$z = \frac{\overbrace{(\overline{w} - \overline{m}) - 0}^{\text{observed diff.} - H_0\text{-expected diff.}}}{SE_{\text{diff}}} = \frac{\overline{w} - \overline{m}}{\sqrt{SE_w^2 + SE_m^2}},$$

(*) $\overline{w}$ = the average of a simple random sample of women,

(*) $\overline{m}$ = the average of a simple random sample of men.

(*) If the samples are independent of each other and both sufficiently large then z follows the normal distribution (approximately).

(*) Collects the data:

She surveys a simple random sample of 1225 U.S. men and a simple random sample of 1444 U.S. women. The men surveyed watched an average of 7.86 hours per week of cooking shows, with an SD of 3.8 hours per week. The women watched an average of 8.13 hours per week of cooking shows, with an SD of 2.7 hours per week.

(*) Calculates (the observed value of) the test statistic, and finds the *p*-value:

$$z^* = \frac{\overline{w} - \overline{m}}{\sqrt{SE_w^2 + SE_m^2}} = \frac{8.13 - 7.86}{\sqrt{\frac{2.7^2}{1444} + \frac{3.8^2}{1225}}} \approx 2.08$$

The *p*-value is equal to the area under the normal curve to the right of $z^* = 2.08$, which is about 1.88%.

The difference in averages is (statistically) significant (but not quite *highly* significant). It is not likely explained by chance, so there is (likely) a difference in viewing habits of men and women, when it comes to cooking shows.

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.

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

Alternative hypothesis: Using a calculator does have an effect on the students' work.

In terms of the relevant parameters:

$H_0 : \%_{pp} - \%_c = 0,$

$H_A : \%_{pp} - \%_c \neq 0,$

Where $\%_{pp}$ = percentage of correct answers among all 13-year olds who don't use a calculator, and $\%_{pp}$ = percentage of correct answers among all 13-year olds who do use a calculator

(*) The *observed difference*: $23.6\% - 7.2\% = 16.4\%$

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

(*) The standard errors are

$$SE_c = \frac{\sqrt{0.072 \times 0.918}}{\sqrt{250}} \times 100\% \approx 1.626\%$$

and

$$SE_{\text{pp}} = \frac{\sqrt{0.236 \times 0.764}}{\sqrt{250}} \times 100\% \approx 2.686\%.$$

(*) The standard error for the difference is

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

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

(*) The P-value is the area under the normal curve outside the interval $(-5.2, 5.2)$: $p \approx 0$.

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

Comments:

- The samples in a controlled experiment are drawn *without replacement*, and usually represent a *significant proportion* of the 'population' (the box of test subjects). 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 receive the same information in one of two forms.

Form A:

> *Of* 100 *people having surgery,* 10 *will die during surgery,* 32 *will die within one year and* 66 *will die within five years.*
>
> *Of* 100 *people having radiation, none will die during treatment,* 23 *will die within one year and* 78 *will die within five years.*

Form B:

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

Test of significance:

Hypotheses:

Null: The difference in the forms has no effect on the percentage of doctors favoring surgery...

$$H_0 : \quad \%_A - \%_B = 0$$

Alternative: The difference in the forms does have an effect on the percentage of doctors favoring surgery...

$$H_1 : \quad \%_A - \%_B \neq 0$$

(*) $\%_A$ = percentage of doctors who read form A that favor surgery.

(*) $\%_B$ = percentage of doctors who read form B that favor surgery.

Test statistic:

$$z = \frac{\text{observed difference} - H_0\text{-expected difference}}{SE_{\text{diff}}}$$

P-value: Area under normal curve outside of $(-z, z)$.

Observed percentages: $\%_A = 50\%$ and $\%_B = 84\%$.

Standard errors:

$$SE_{\%_A} = \frac{\sqrt{0.5 \times 0.5}}{\sqrt{80}} \times 100\% \approx 5.6\%$$

and

$$SE_{\%_B} = \frac{\sqrt{0.84 \times 0.16}}{\sqrt{87}} \times 100\% \approx 3.9\%.$$

Test statistic:

$$z = \frac{(84\% - 50\%) - 0\%}{\sqrt{(5.6\%)^2 + (3.9\%)^2}} \approx 4.98$$

P-value: ... $p \approx 0$

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

(*) In this case, we might also say that the difference is also *significant* in the (non-statistical) sense: *the way that information is presented can have a substantial effect on the decisions that doctors make in situations like this.*