OVERVIEW OF TESTS OF SIGNIFICANCE

- Case I: One box, testing its average against an external standard.
  - Chapter 26.
- Case II: Two boxes, testing for the equality of their averages.
  - Chapter 27.
- Case III: One box, testing the percentage composition of the box against an external standard.
  - Chapter 28.
- Case IV: Two boxes, testing whether the percentage composition is the same for both boxes.
  - Chapter 28.

- The general methodology for Cases I and II is the same. The basic idea is to see how many SEs an observed result is from its expected value:

$$z = \frac{\text{observed result} - \text{expected result}}{\text{SE for the observed result}}.$$

  - Large values (positive or negative) of z are a sign that something is fishy.
  - Tests based on z are called z-tests.
- Cases III and IV are handled using the so-called χ^2-test.

THE EXPECTED VALUE AND SE FOR THE DIFFERENCE BETWEEN TWO CHANCE QUANTITIES

- Think about writing the difference between two random quantities (e.g., sample averages) in the form

$$\text{difference} = \text{expected value} + \text{chance error}.$$

How are the expected value for the difference and the SE for the difference (i.e., the likely size of the chance error) related to the expected values and SEs of the two individual quantities?

- The expected value for the difference is just the difference between the individual expected values.
  - This is true whether or not the quantities are independent.
- If the two quantities are *independent*, the SE for their difference is given by a square root law:

$$\text{SE for difference} = \sqrt{a^2 + b^2},$$

where a is the SE for the first quantity, and b the SE for the second quantity.

  - This is smaller than $a + b$, the sum of the individual SE's, because some cancellation of the chance errors in the individual quantities can take place, due to the independence.
    - Chance error in the difference = chance error in the first quantity − chance error in the second quantity.
  - The formula doesn't apply if the quantities are dependent.

- Reminder: The expected value of the difference

"first quantity − second quantity"

between two quantities equals

If the two quantities are independent, then the SE for the difference equals

- Two draws are made at random *with* replacement from the box

$$\boxed{\boxed{0}\ \boxed{0}\ \boxed{1}}\ .$$

Find the expected value and SE for the difference of the draws.

  - The box has average ________ and SD ________ = 0.471.
  - Each draw has expected value ________ and SE ________ .
  - The difference between the draws has expected value ________ and SE ____________.
    - The square root law applies, because the draws are ____________ .
  - The difference between the draws will be ________, give or take ________ or so.

- How could you check the final claim above?
  - By carrying out an ____________ study. I had the computer compute the difference between the draws 10,000 times. The 10,000 differences averaged out to $-0.0004 \approx 0$, and the RMS amount off from 0 equaled $0.668 \approx 2/3$. This confirms the theory.

- Two draws are made at random *without* replacement from the box

$$\boxed{\boxed{0}\ \boxed{0}\ \boxed{1}} \qquad \begin{array}{c}\text{average of box} = 1/3,\\ \text{SD of box} = \sqrt{2/9} = 0.471\ .\end{array}$$

True or false and explain: the difference between the draws will be about 0, give or take $\sqrt{(0.471)^2 + (0.471)^2} = 2/3$ or so.

  - The expected value is ____________, but the SE is ____________, because the draws are ____________.
  - I had the computer compute the difference between the draws 10,000 times. These 10,000 differences averaged out to $-0.002 \approx 0$, but the RMS amount off from 0 equaled 0.818, somewhat more than 2/3.

- When two quantities are negatively (respectively, positively) correlated, the SE for their difference will be ________ (respectively ________) than it would be for independence, because there tends to be ________ (respectively ________) cancellation of the individual chance errors.
  - Chance error in the difference = chance error in the first quantity − chance error in the second quantity.
  - If the two quantities are (strongly) negatively correlated, then the two chance errors tend to be of ________ signs, and ____________ takes place in forming the chance error for the difference.
  - If the two quantities are (strongly) positively correlated, then the two chance errors tend to be of the ________ sign, and ____________ takes place in forming the chance error for the difference.
  - In the example above, the two draws are ____________ correlated, because

COMPARING TWO INDEPENDENT SAMPLE PERCENTAGES

- Example *Cocaine usage*: The National Household Survey on Drug Abuse was conducted in 1974 and again in 1985. Among persons age 18 to 25, the percentage of current users of cocaine increased from 3.1% to 7.7%. Is this difference real, or a chance variation? You may assume that the results are based on two independent simple random samples, each of size 700.

- Box model:

1974 box	1985 box
One ticket for each	One ticket for each
____________	____________
Tickets marked	Tickets marked
____________	____________
Percentage of 1's in box	Percentage of 1's in box
____________	____________

  - The percentage of current cocaine users in the sample for 1974 is like the ______________ among the 700 draws from the 1974 box. Similarly, for 1985.

- The percentage of current cocaine users (among all persons age 18 to 25) in 1974 is estimated as ______, give or take ______ or so:

$$\approx 0.7\%\,.$$

- Similarly, the percentage of current cocaine users in 1985 is estimated as ______, give or take 1% or so.

- The question is "Is the percentage of 1's in the boxes really different (the ________ hypothesis), or the same (the ________ hypothesis)?"

- 1974 usage: $3.1\% \pm 0.7\%$, 1985 usage: $7.7\% \pm 1.0\%$

- The difference (1985 usage - 1874 usage) in cocaine usage is estimated as ________, give or take ______ or so:

  - Because the two samples are ________, the SE for the difference in the sample percentages is given by the square root law:

$$\begin{pmatrix}\text{likely size of the}\\ \text{sampling error in the}\\ \text{difference } 7.7\% - 3.1\%\end{pmatrix} = \sqrt{\begin{pmatrix}\text{likely size of}\\ \text{the sampling}\\ \text{error in } 3.1\%\end{pmatrix}^2 + \begin{pmatrix}\text{likely size of}\\ \text{the sampling}\\ \text{error in } 7.7\%\end{pmatrix}^2}$$

$$= \sqrt{(\quad\ \%)^2 + (\quad\ \%)^2} \approx 1.2\%.$$

- How strong is the evidence against the null hypothesis of no change in cocaine usage between 1974 and 1985?

  - Test statistic:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for observed difference}}$$

$$= \frac{\quad\%-\quad\%}{\quad\%} = 3.8.$$

  - P-value:

  - If the null hypothesis is true, the probability histogram for z will follow the standard normal curve, approximately.

- How big was the increase in cocaine usage from 1974 to 1985?

  - 95% confidence interval is ________

    - With 95% assurance, we can assert that usage increased somewhere from ______% to ______%.

    - The null hypothesis says that the difference is 0%. On the basis of the CI, would you believe the null?

COMPARING TWO INDEPENDENT SAMPLE AVERAGES

• Freshmen at public universities work 12.2 hours a week for pay, on average, and the SD is 10.5 hours; at private universities, the average is 9.2 hours and the SD is 9.9 hours. Assume these data are based on two independent simple random samples, each of size 1,000. Is the difference between the averages due to chance? If not, what else might explain it?

• Box model:

  • There are two boxes, one for public institutions, the other for private ones. In the public box, there is one ticket for each freshman at a public university; a person's ticket shows the hours worked per week for pay. The average of the public box is the average hours worked per week for pay, by all freshman at public universities; this quantity is unknown (because we don't have access to all the tickets). 1,000 draws are made at random without replacement from the public box. The private box is set up the same way. There is no connection between the draws from the public and private boxes.

• Null and alternative hypotheses:

  • The null hypothesis says that the difference between the sample averages is due to chance; more precisely, it says that the average of the public box is the same as the average of the private box.

• The samples:

Schools	*Sample size*	*Average*	*SD*
Public	1,000	12.2 hrs/wk	10.5 hrs/wk
Private	1,000	9.2 hrs/wk	9.9 hrs/wk

• Estimates and their SEs:

  • μ_{pub} = (average of public box) is estimated as
$$12.2 \pm 10.5/\sqrt{1{,}000} = 12.2 \pm 0.33 \text{ hrs/wk}.$$

  • μ_{priv} = (average of private box) is estimated as
$$9.2 \pm 9.9/\sqrt{1{,}000} = 9.2 \pm 0.31 \text{ hrs/wk}.$$

  • $\Delta = \mu_{\text{pub}} - \mu_{\text{priv}}$ is estimated as
$$12.2 - 9.2 \pm \sqrt{(0.33)^2 + (0.31)^2} = 3 \pm 0.45 \text{ hrs/wk}.$$

• Test statistic and P-value:
$$z = \frac{\text{obs'd difference} - \text{exp'd difference}}{\text{SE for difference}} = \frac{3-0}{0.45} \approx 6.$$
$$P\text{-value} = \text{zilch}.$$

• Conclusion:

  • There is a real difference in the average hours per week worked for pay between freshman at public and private universities. With 95% assurance, we can state that on average freshman at public schools work somewhere from $3 - 2 \times 0.45 = 2.1$ to $3 + 2 \times 0.45 = 3.9$ more hours per week than do freshman at private schools. This is probably because the students at private universities tend to come from wealthier families and don't need to work as much for pay.

A CASE OF TWO WRONGS MAKING A RIGHT

• A box contains 200 numbered tickets, which average out to 50; the SD is 30. 200 tickets are to be drawn at random *without* replacement. The difference between the average of the first 100 draws and the average of the second 100 draws will be about ________, give or take about ____________ or so:

- Each of the two sample averages will be about ________, give or take ____________ or so:

$$\text{SE}_{\text{with replacement}} =$$

$$\text{SE}_{\text{without replacement}} \approx$$

- The square root formula doesn't apply, because the two sample averages are *negatively correlated*; in fact,

$$(2^{\text{nd}} \text{ average}) = 2 \times (\text{average of box}) - (1^{\text{st}} \text{ average}).$$

- To proceed, note $(1^{\text{st}}\text{ average}) - (2^{\text{nd}}\text{ average}) = 2 \times \big((1^{\text{st}}\text{ average}) - (\text{average of box})\big)$, so

$$\begin{aligned} &\text{SE for the difference between the two averages} \\ &\quad = 2 \times (\text{SE for the first average}) \\ &\quad \approx 2 \times (3 \times \sqrt{1/2}) = \sqrt{3^2 \times 2^2/2} = \sqrt{3^2 + 3^2} \end{aligned}$$

- The final result is what you would have gotten if you:
  - computed the SE for each average on the basis of drawing *with* replacement;
  - combined the two SE's as though the two samples were *independent* i.e., by using the square root formula.
  - That's a case of two wrongs making a right!

$$\boxed{A|?}\ \boxed{A|?} \xleftarrow[\text{A sample}]{} \underline{\big|\ \boxed{A|B}\ \boxed{A|B}\ \ldots\ \boxed{A|B}\ \big|} \xrightarrow[\text{B sample}]{} \boxed{?|B}$$

EXPERIMENTS

• Here is a box model for a randomized controlled experiment comparing treatments A and B (e.g., a vaccine and a placebo).

• There is one ticket in the box for each subject in the experiment.

• Each ticket has two numbers: one shows what the subject's response to treatment A would be, the other what the response to treatment B would be.

- Only one of the two numbers can be observed.

• Some tickets are drawn at random *without* replacement from the box, and the responses to treatment A are observed.

• Then, a second sample is drawn at random *without* replacement from the remaining tickets, and the responses to treatment B are observed.

- The two samples are *dependent*.

• The difference between the two sample averages estimates the difference between

> the average of the "A" numbers on all the tickets in the box, i.e., what the average response to treatment A would be, if everybody were to get that treatment

and

> the average of the "B" numbers on all the tickets in the box.

• The SE for the difference between the two sample averages can be conservatively estimated as follows:

- compute the SEs for each average on the basis of drawing at random *with* replacement;
- combine the SEs as if the two samples were *independent*.

$$\boxed{\mathrm{A}|?}\ \boxed{\mathrm{A}|?} \xleftarrow[\text{A sample}]{} \boxed{\boxed{\mathrm{A}|\mathrm{B}}\ \boxed{\mathrm{A}|\mathrm{B}}\ \ldots\ \boxed{\mathrm{A}|\mathrm{B}}} \xrightarrow[\text{B sample}]{} \boxed{?|\mathrm{B}}$$

- "Conservatively" means that this procedure may overestimate the actual SE for the difference between the sample averages.
    - Confidence intervals computed using the conservative SE will have more than the stated confidence level.
    - P-values computed using the conservative SE will be larger than they actually are.
- The procedure yields nearly the right SE when:
    - the sample sizes are small compared to the number of tickets
        - then there is little difference between drawing with or without replacement, and the dependence between the averages is small too,

    or when

        - the "A" and "B" numbers on the tickets are highly correlated,
            - in particular, when the two numbers on each ticket are the same,
                - as they would be under the null hypothesis of no difference.
- The normal curve can be used to figure chances about the difference between the two sample averages provided the sample sizes are sufficiently large.
- The procedure applies to percentages (or rates), as well as to averages.
    - Classify and count; the "A" and "B" numbers are 0's and 1's.

- Example *Alexithymia.* Children who grow up in alcoholic homes are likely to have difficulty identifying and expressing their feelings and needs. Adults who are unable to identify or express feelings are said to be "alexithymic." Such individuals may have difficulty in relating to other people and in developing stable and satisfying relationships.
- Psychologists have noted that feelings involve both affective (subjective experience of emotion) and cognitive (intellectual understanding) components.
- In this example we compare two styles of group therapy for treating alexithymic adult children of alcoholics. After agreeing to participate in the study, subjects were allocated to one of two treatment groups using a randomization procedure:
    - "psychodrama" (affective emphasis) 25 subjects.
        - Focus on experience.
        - Reaction to events.
        - Group members act out key situations in their lives.
    - "choices" (cognitive emphasis) 25 subjects.
        - Focus on intellectual understanding.
        - Improve interpersonal communication skills.
        - Lectures plus group exercises.
- The extent of alexithymia was measured before and after eight weeks of group therapy, and a measure of improvement was obtained. Values of 0 mean no improvement, and negative values indicate worsening.

- Results:
  - "Psychodrama": 25 subjects, average 8, SD 12.
  - "Choices": 25 subjects, average -0.5, SD 6.
- Is the difference between the averages statistically significant?
- Box model:
  - There is one ticket in the box for each of the 50 subjects in the study. Each ticket has two numbers: A – what the subject's improvement would be under psychodrama, B – what the subject's improvement would be under choices. 25 draws are made at random w/o replacement and the A numbers on the tickets are revealed. Then the remaining tickets are drawn and their B's are revealed.
- The difference between what the average improvement under "psychodrama" would be for all 50 subjects and what the average improvement under "choices" would be for all 50 subjects is estimated as __________.
- To put a standard error on the difference we may pretend that we have two independent samples drawn at random with replacement.
  - The SE for the "psychodrama" average is estimated as __________ $= 2.4$.
  - The SE for the "choices" average is estimated as __________ $= 1.2$.
  - The SE for the difference is (conservatively) estimated as __________ ≈ 2.7.

- Difference in average improvement $= 8 - (-0.5) = 8.5$. Estimated SE for the difference $= 2.7$.
- The null hypothesis says that there is _______ difference between the two therapies.
  - On this basis, the expected value for the difference between the sample averages is _______.
- The test statistic is

$$= 3.1\ .$$

- The P-value is less than _______%. The difference between the therapies is __________.
- How big is the difference?
  - An approximate 95% confidence interval for the difference in average response to psychodrama and the average response to choices, for all 50 subjects, is $8.5 \pm 2 \times 2.7$.
- How important is the difference?
  - Does an 8 point improvement mean that those treated are really functioning better? That's for the psychologist to decide.
- Can the results be extrapolated beyond the 50 subjects in the study?
  - Yes, if the 50 subjects were a SRS from larger population (then the box model would have 1 ticket for each of the

people in that population, and not all the tickets would be drawn). If the 50 subjects are not such a SRS, the book would argue "No" — you can't draw conclusions from a sample of convenience (Sneer). But there are statisticians who would argue a qualified "yes", that the chance process that got the 50 subjects into the study is sufficiently like a SRS that the CI's and P-values are about right. Of course, you have to be careful. These subjects were willing to participate in 8 weeks of group therapy (they wanted help) and may be different from those who would not want to participate. Also, if there were siblings in the study, the computed SEs might be too small (that's the case for cluster samples).

SUMMARY

• The standard error for the difference of two independent quantities is $\sqrt{a^2 + b^2}$, where: a is the SE of the first quantity; and b is SE of the second quantity. The formula does not apply to the difference of two dependent quantities.

• Suppose two independent and reasonably large simple random samples are taken from two separate boxes. The null hypothesis says that the two boxes have the same average, so the expected difference between the sample averages is 0. The appropriate test statistic is

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference.}}$$

Tests based on z are called *two-sample z-tests*.

• This procedure can handle situations which involve classifying and counting, by putting 0's and 1's in the boxes.

• This procedure can also be used to compare treatment and control averages or rates in an experiment. Suppose there is a box of tickets. Each ticket has two numbers: one shows what the response would be to treatment A; the other, to treatment B; only one of the numbers can be observed. Some tickets are drawn at random w/o replacement from the box, and the responses to treatment A are observed. Then, a second sample is drawn at random w/o replacement from the remaining tickets. In the second sample, the responses to treatment B are observed. The SE for the difference between the two sample averages can be conservatively estimated as follows: (i) compute the SEs for the averages on the basis of drawing at random with replacement; (ii) combine the SEs as if the two samples were independent.