

More Tests for Averages: Comparing Two Samples

1

Where are we going?



Review of concepts of hypothesis testing

Comparing two independent samples

- SE of a difference
- Applications to percentages and continuous measurements

Randomized experiments

2

Review

A **significance test** is aimed at determining whether a result is real or could possibly be due to chance.

The **null hypothesis** says that the result is due to chance. A probability calculation is made under the null hypothesis via a box model.

A **test statistic** measures the difference between the data and what would be expected under the null hypothesis.

3

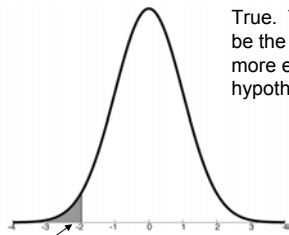
The **observed significance level**, or **P-value**, is the chance of getting a test statistic as or more extreme than the one observed if the null hypothesis were true.

Small P-values are evidence against the null hypothesis.

Sometimes a threshold (like 5%) is set in advance, and the null hypothesis is said to be **rejected** if the P-value is smaller than that threshold.

4

Question: If the null hypothesis is true, the chance of getting a P-value of .025 or smaller is .025. True or False?



True. The P-value is *defined* to be the chance of getting as or more extreme a result if the null hypothesis is true.

5

11. (Hard.) Discount stores often introduce new merchandise at a special low price in order to induce people to try it. However, a psychologist predicted that this practice would actually reduce sales. With the cooperation of a discount chain, an experiment was performed to test the prediction.¹⁰ Twenty-five pairs of stores were selected, matched according to such characteristics as location and sales volume. These stores did not advertise, and displayed their merchandise in similar ways.

A new kind of cookie was introduced in all 50 stores. For each pair of stores, one was chosen at random to introduce the cookies at a special low price, the price increasing to its regular level after two weeks; the other store in the pair introduced the cookies at the regular price. Total sales of the cookies were computed for each store for six weeks from the time they were introduced.

In 18 of the 25 pairs, the store which introduced the cookies at the regular price turned out to have sold more of them than the other store. Can this result be explained as a chance variation? Or does it support the prediction that introducing merchandise at a low price reduces long-run sales? (Formulate the null hypothesis as a box model; there is no alternative hypothesis about the box.)

6

Summary: 25 pairs of stores. One member (chosen randomly dropped the price). In 18 cases of 25, the store that did not drop the price sold more.

Null hypothesis: the difference is due to chance

Alternative hypothesis: there is some kind of effect.

Probability model for the null hypothesis: 25 draws with replacement from what kind of box?



7

Calculations under the assumption that the null hypothesis is true: what can we say about the number of times a store that did not drop the price sold more?

EV of # =

SE of # = 1

Test statistic:

8



P-value ~ 1.4%

Conclusion: such a result is quite unlikely if there is no effect. There is thus strong evidence against the null hypothesis.

9

Paired Samples

A key aspect of this design was that the stores were initially grouped into similar pairs. We used this in the analysis by just basing it on the outcomes of the 25 pairs. Thus we reduced the problem to that of comparing the results of a *single* sample to a standard. Then we used a test for a single sample.

Now suppose that the stores had not been paired and we had two independent samples? That's the kind of experiment that we will now be concerned with.

10

Example: Comparison of Polls

Public support for a ballot issue is measured by a simple random sample of size 1500 on August 1 and found to be 35%. On September 1 another independent random sample of size 1500 is taken, according to which the support has risen to 40%. How strong is the evidence that the support has actually increased? How strong is the evidence against the *null hypothesis* of no change?

11

How can we assess the plausibility of the null hypothesis? We need to calculate a p-value for the statistic

$$Z = \frac{\text{observed difference} - \text{expected value of difference if null true}}{\text{SE of difference if null is true}}$$

Diagram illustrating the components of the Z-statistic:

- Observed difference: Know this, 10%
- Expected value of difference if null true: Know this, 0%
- SE of difference if null is true: Don't know this

12

SE of a Difference

Model: Make **independent** draws from each of two boxes. From the draws from the first box you calculate a quantity X. From the draws from the second box you calculate a quantity Y.

X and Y might be:

- Totals of tickets
- Averages of tickets
- The number of times something happens
- The percentage of the time something happens

13

Fact, to be
taken on
Authority:



In this setup, if X has an SE(X) and Y has an SE(Y), then, X-Y has an SE equal to

$$SE(X-Y) = \sqrt{SE(X)^2 + SE(Y)^2}$$

Aside: the quantity $SE(X)^2$ is called the "variance" of X.

14

Example:

August 1, 35% of 1500

SE =

Sept 1, 40% of 1500

SE =

SE of difference =

15

Null Hypothesis: no change

test statistic

$$Z = \frac{\text{observed difference} - \text{expected value of difference if null true}}{\text{SE of difference if null is true}}$$

$$Z = \quad .$$

P-value?

16

More Examples

1. A coin is tossed 100 times. The number of heads is counted. It is then tossed another 100 times and the number of tails is counted.

SE of # heads =

SE of # tails =

SE of # heads minus # tails =

17

2. A coin is tossed 100 times. The number of heads is counted and the number of tails is counted.

SE # heads =

SE # tails =

SE of # heads minus # tails =

18

3. 25 items are sampled from a production one day and they have an average of 100 and an SD of 5. 25 items are sampled the next day and they have an average of 90 and an SD of 8.

SE of first day average =

SE of second day average =

SE of first day average minus second day average =

Z=

19

Example

Company makes plastic grocery bags. They had been having problems with bags tearing. Bags sampled from two production runs and their tensile strengths measured.

Run 1: 32 bags, average = 102.33, SD = 14.06

Run 2: 40 bags, average = 118.19, SD = 24.44

Is there a significant difference between the average tensile strengths of the two runs?

20

Run 1: 32 bags, average = 102.33, SD = 14.06

Run 2: 40 bags, average = 118.19, SD = 24.44

SE of average from run 1:

SE of average from run 2:

SE of difference:

Test statistic:

p-value: \approx

Conclusion:

21

Randomized Experiments: Example

An experiment was done using male bank supervisors who were attending a workshop. They had to make decisions about items in their in-baskets. Unknown to the bank supervisors, the investigators imbedded their experiments in the in-baskets.



22



Two personnel files were prepared. They were identical except that one was labeled as belonging to a male employee and one as belonging to a female employee. By random selection, 24 of the supervisors got the "male" and 24 got the "female" file. They had to decide whether to promote the employee or hold the file and interview additional candidates.

23

The Results

	Male	Female
Promote	21	14
Hold	3	10

Males: $21/24 = 87\%$ promoted

Females: $14/24 = 58\%$ promoted

Is there sex discrimination?

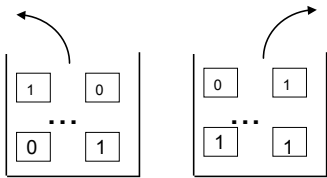
24

The null hypothesis: "The unfortunate difference in promotion rates is just due to chance. We are deeply committed to diversity in our workforce."

Alternative hypothesis: those bank supervisors are sexist pigs!



Our previous approach to assessing plausibility of the null hypothesis: 24 draws with replacement from each of two 0-1 boxes with unknown fractions of 0's and 1's



The null hypothesis is that the fractions are equal.

A model closer to how the experiment was actually done:

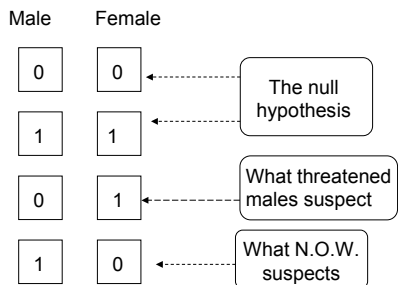
There are 48 supervisors in all. Each has a potential response to a file labeled female and a file labeled male. Think of each supervisor as holding two tickets:

what he would do with the file labeled female

what he would do with the file labeled male

We get to see **one** of these two.

If 1 = promote and 0 = hold, there are four possibilities:



28

Under the null hypothesis the box has 48 ticket-pairs which are either

M	F
0	0
1	1

24 are drawn without replacement and the left one is read. Then other 24 are drawn and the right one is read.

29

This is a very different model

- The "male" draws are made without replacement. So they are dependent.
- The "female" draws are made without replacement, so they are dependent
- The averages of the "female" draws and the "male" draws are dependent.

So it looks like the box model we used earlier is not appropriate and would give the wrong SE to compare the male and female draws.

30

The Right Box Model for the Null

There are 48 tickets, either labeled [0 0] or [1 1]

	Male	Female
Promote	21	14
Hold	3	10

How many [1 1] tickets?

How many [0 0] tickets?

31

Box Model: 35 [1 1] tickets and 13 [0 0] tickets.

Draw 24 *without replacement* and call them "male." Count how many 1's you get.

In the actual experiment, the number was 21. Is this surprisingly large? Could it be due to chance variation?

In principle, can work out the probabilities for drawing without replacement.

The P-value is about .025.

32

So, with some work, we can figure out how to calculate probabilities for drawing with replacement from this box



Miraculously, it turns out that the results you get using the *wrong* box model are approximately correct anyhow.

33

The Calculations

	Male	Female
Promote	21	14
Hold	3	10

Males: $21/24 = 87\%$ promoted

Females: $14/24 = 58\%$ promoted

34

The wrong box model. Draw independently 24 times from each. Estimated box SDs are

The estimated SEs of the percents are then

SE for box 1 =

SE for box 2 =

35

The SE of the difference is:

The observed difference is:

The expected difference if the null hypothesis is true is:

36

$$Z = \frac{\text{observed difference} - \text{expected value of difference if null true}}{\text{SE of difference if null is true}}$$

P-value = ____

The samples are a little small to use the normal curve. (The right model gave a p-value = .025.)

Conclusion: the difference is unlikely to have occurred by chance. It looks very much like discrimination.

37

Randomized Experiments with Quantitative Responses

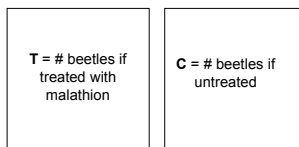
Example: Study of leaf beetle damage to oats. 27 small plots were randomly divided into 13, which served as controls, and 14 to which malathion was applied. The numbers of leaf beetles per stem were then counted.

	Average	SD
Control	3.47	1.21
Malathion	1.36	.52

38

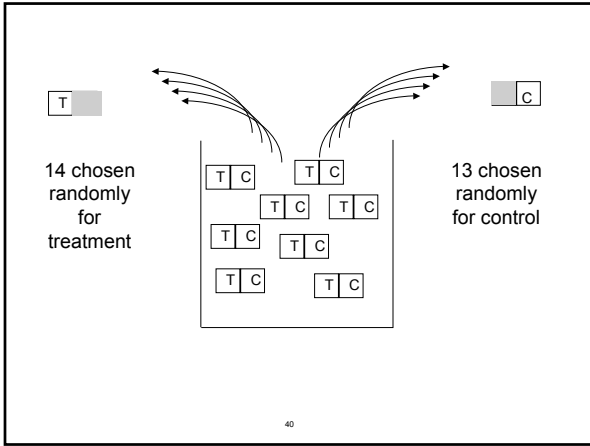
A Box Model

Each of the 27 plots has a ticket with two values on it



treatment value control value

39



The 14 treatment values and the 13 control values are not independent simple random sample (draws with replacement)

- The 14 treatment values are drawn without replacement, so there is dependence among them.
- The 13 control values are drawn without replacement, so there is dependence among them.
- There is dependence between the treatment and control values, since they are drawn without replacement from the same box

This chance model is quite different than the one for which we know how to calculate an SE for a difference. That one is based on independent SRS from each of two populations (draws with replacement from each of two boxes).

So how can we calculate a standard error for the difference in order to do a significance test?

A Miracle Occurs!



We can do the math for the correct box model and discover that the SE as calculated from the wrong model is about right.

	Average	SD	n
Control	3.47	1.21	13
Malathion	1.36	.52	14

SE of control average =


SE of treatment average =

SE of difference =

Null hypothesis: Response is the same for both treatments. Malathion has no effect.

Test statistic:

Conclusion??

 The New England Journal of Medicine

A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee
J. Bruce Moseley, M.D., Kimberly O'Malley, Ph.D., Nancy J. Petersen, Ph.D., Terri J. Menke, Ph.D., Baruch A. Brody, Ph.D., David H. Kuykendall, Ph.D., John C. Hollingsworth, Dr.P.H., Carol M. Ashton, M.D., M.P.H., and Nelda P. Wray, M.D., M.P.H.

ABSTRACT

Background Many patients report symptomatic relief after undergoing arthroscopy of the knee for osteoarthritis, but it is unclear how the procedure achieves this result. We conducted a randomized, placebo-controlled trial to evaluate the efficacy of arthroscopy for osteoarthritis of the knee.

Methods A total of 180 patients with osteoarthritis of the knee were randomly assigned to receive arthroscopic débridement, arthroscopic lavage, or placebo surgery. Patients in the placebo group received skin incisions and underwent simulated débridement without insertion of the arthroscope. Patients and assessors of outcome were blinded to the treatment-group assignment. Outcomes were assessed at multiple points over a 24-month period with the use of five self-reported scores — three on scales for pain and two on scales for function — and one objective test of walking and stair climbing.

Scores on Pain after 18 Months

Placebo Group: (n = 52) average = 55.6 SD = 23.6

Debridement group (n = 51) average = 50.7 SD = 24.4

SE of Placebo Group:

SE of Treatment Group:

SE of Difference:

Null hypothesis: treatment has no effect

Test statistic:

P-value:

46

Where have we been? A review of z-tests



47

$$Z = \frac{\text{observed value} - \text{expected value under null hypothesis}}{\text{SE of value under null hypothesis}}$$

value might be: average, total, percentage, difference of average, total, percentage

In order for the test to be meaningful, it has to be based on a chance model.

48

When Does a Z-Test Apply?

- You have to have a chance model corresponding to drawing tickets from boxes with replacement. A simple random sample, for example.
- The sample size must be large enough for the Central Limit Theorem to apply.

Which Z-Test?

- **One sample test:** compare results of a single sample to a standard. Matched pairs are a special case of this.
- **Two sample test:** compare results of two independent samples to each other. This test can also be used in randomized comparisons of treatment and control (27.3 & 27.4 of text).

1. From each of several litters, two rats are taken. An experiment is done in which one is randomly selected for treatment and the other used for control. After the experiment the averages and SDs of the two groups are found. Can the two sample z-test be used to test for significance?

2. A simple random sample of 100 people from a population was asked to (1) rate their confidence in lawyers on a 0-6 scale, and (2) rate their confidence in doctors on a 0-6 scale. The averages and SDs of the two ratings are found. Can the z-test be used to assess whether the two are significantly different?

52

3. In a study published in the *Journal of Human Ergology*, 28 middle aged college faculty who had volunteered for a fitness program were divided into low and high fitness groups based on a physical exam. They then took the Cattell Sixteen Factor Personality Inventory. Their scores on the "ego strength" personality factor were as follows:

	Average	SD
Low fitness (14)	4.64	.69
High fitness (14)	6.43	.43

Can the two sample z-test be used to see if the results are significantly different?

53

4. Last year, 62% of the residents of a large apartment complex were satisfied with the performance of the complex manager. In a simple random sample of 75 residents taken from among all residents this year, 40 were satisfied with the manager's performance. Are fewer residents satisfied this year than last year? Or could the difference be due to chance?

54

5. According to the Census, the average annual income of heads of households in City A is \$20,000 with and SD of \$10,000. In a simple random sample of 900 households taken from City B, the average income of heads of households was \$25,000, with and SD of \$12,000. Is the difference in averages real? Or is this chance variation?

56

6. A simple random sample of 1000 SAT math scores taken in 1967 showed an average of about 488 points, with an SD of about 100 points. An independent simple random sample of 4000 SAT math scores taken in 1987 showed an average of about 475 points, with an SD of about 100 points. Has there been a decline in SAT math scores? Or is this just chance variation?

56

7. An investigator believes that among married couples in her city, the husbands have a higher average educational level than the wives. In order to test this hypothesis, she proposes the following study:

the population: all married couples in her city

the sample: a simple random sample of 500 couples from this population

the data: for each sample family, the educational level of the husband and of the wife

the test: the usual two-sample z-test for the difference between the average educational levels of husbands and wives.

Will the test be valid?

57

8. There are 300 workers in an office building, and they all have colds. 150 workers are picked at random and given a well-known night-time cough medicine. The other 150 are given no medication at all. The variable of interest is the percent in each group that feel better after two days. Can you use the usual two-sample z-test to compare the percents? Or would it not be valid?
