List of Definitions, Concepts, and Formulas for Math 125

Definitions and Concepts

Chapter 26

Tests of Significance:

Test of Significance (definition) A formal procedure for comparing observed data with a claim (called a hypothesis), the truth of which is being assessed. Usuually the null hypothesis is a statement about a box model for the whole population, and the results of the test are expressed in terms of a probability that measures how well the data and the claim agree.

Example:

According to one investigator's model, the data are like 625 draws made at random from a large box. The null hypothesis says that the average of the box equals 50. The alternative says that the average of the box is more than 50. In fact, the data averaged out to 52.9, and the SD was 30. Compute z and P. What do you conclude?

Answer:

SE for average $= 1.20$, so $z = (52.9 - 50)/1.20 \approx 2.42$ and P is approximately the area to the right of 2.42 under the normal curve. From the table, this is about 0.3%. The difference is hard to explain as chance variation. The alternative hypothesis is looking good.

Tests of Significance Using Zero-one Boxes:

Example:

In a test for ESP, a subject is confronted with "targets" numbered from one to five and asked to identify which of the 5 had been secretly and randomly selected as the winning target and is now being held up on the other side of a curtain. Suppose that in 500 trials, a subject correctly identifies 127 winners.
(A subject who has ESP would identify winning targets more often than would a person guessing randomly.)

(a) Set up the null hypothesis as a box model.

(b) The SD of the box is _____. Fill in the blank, using one of the options below, and explain briefly.

$$\sqrt{0.2 \times 0.8} \qquad\qquad \sqrt{0.254 \times 0.746}$$

(c) Make the z-test.

(d) What do you conclude?

Answer:

(a) Null hypothesis: the number of correct guesses is like the sum of 500 draws from a box with one ticket marked 1 and four 0's.

(b) $\sqrt{0.2 \times 0.8}$. The null hypothesis tells you what's in the box. Use it.

(c) $z \approx (127 - 100)/8.944 \approx 3.0$, and P is very small. (P is about 0.1355 of 1%.)
($\sqrt{500} \times$ SD of box $= 22.36 \times \sqrt{0.2 \times 0.8} = 22.36 \times 0.4 \approx 8.944$.)

(d) Whatever it was, it wasn't chance variation.

Example:

A coin is tossed 14 times, and it lands heads 12 times. Is the chance of heads equal to 50%? Or are there too many heads for that?

Answer:

The null hypothesis is that the coin is fair, i.e., the chance of heads is 1/2.

The alternative hypothesis is that the coin gets too many heads.

There are not enough draws to use the normal approximation; the chances must be calcuated from the binomial formula.

P is the chance of getting 12 or more heads, assuming $p = 1/2$.

$P = \mathrm{P}(k = 12) + \mathrm{P}(k = 13) + \mathrm{P}(k = 14)$.

These three probabilities are:

$\frac{14!}{12! \times 2!}(1/2)^{12}(1/2)^2 = \frac{91}{16384}$.

$\frac{14!}{13! \times 1!}2^{13}2^1 = \frac{14}{16384}$.

$\frac{14!}{14! \times 0!} = (1/2)^{14}(1/2)^0 = \frac{1}{16384}$.

The total is $106/16384 = 0.647\%$.

Since P is less than 1%, reject the null hypothesis and conclude that the coin gets too many heads.

Example:

In 2009, the New Jersey Nets basketball team lost the first 18 games of its sesaon. A diehard fan makes the following claim: "They are just as good as any other team; they just had bad luck."

Test the hypotheis that the team had a 50% chance of winning each game.

Answer:

The null hypothesis is that the chance of a win is 1/2.

The alternative hypothesis is that their chance of winning a game is less than 1/2.

There are not enough draws to use the normal approximation; the chances must be calcuated from the binomial formula.

P is the chance of getting 0 wins, assuming $p = 1/2$.

$P = \mathrm{P}(k = 0) = 1/(2^{18}) = 0.00038$ of 1%.

Since P is way less than 1%, reject the null hypothesis and conclude that the team did not have a 50% chance of winning each game.

And, say to that fan: "Teams do get bad luck, but they don't get that bad a losing streak as a random result. If it were just as a result of randomness and not an indication of a poor team, such an outcome would occur only once in over a quarter of a million years. Your assertion is unreasonable. I am forced to conclude that they are not as good as an average team."

Chapter 27

Example:

A researcher claims that SAT mathematics scores are lower among New Hampshire students than among Maine students. In his survey, New Hampshire students only averaged 469 on the test, and their SD was 83; while Maine students averaged 490 on the test with the same SD of 78. Can this difference be explained as a chance variation? You may assume that the researcher took a simple random sample of 400 New Hampshire students who took the SAT, and an independent simple random sample of 250 Maine students who took the SAT. Formulate the null and alternative hypotheses in terms of a box model before answering the question.

Answer:

This will require a two-sample z-test for difference in averages.

Null: both populations (New Hampshire and Maine) have the same average mathematics SAT scores.

Alternative: Maine students have a higher average on the mathematics SAT.

First find the sample errors for the population averages. Bootstrap the sample SD and use it in the formula instead of the unknown population SD.

SE for sample average = population SD/$\sqrt{n}$

For New Hampshire: SE avg = 83/20 = 4.15.

For Maine: SE avg = 78/15.811 = 4.93

Next find the standard error for the difference in sample averages:

$\sqrt{a^2 + b^2}$, where a and b are the SEs for the sample averages.

This gives $\sqrt{4.15^2 + 4.93^2} = \sqrt{17.22 + 24.30} = \sqrt{41.52} = 6.44$.

The sample sizes are large, so use the normal approximation.

z = (observed diff in averages − expected diff in averages)/SE diff in averages = $((-21) - 0)/6.44 = -3.26$.

P is equal to about 0.06%, using a one-sided test.

This is a very small value of P.

This difference cannot be explained as a chance variation. Conclude that Maine students had a higher average mathematics SAT score at the time of the survey.

Example:

The Gallup poll asks respondents how they would rate the honesty and ethical standards of people in different fields—very high, high, averge, low, or very low. The percentage who rated clergy "very high or high" dropped from 60% in 2000 to 54% in 2005. This may have been due to scandals involving sex abuse; or it may have been a chance variation. (You may assume that in each year, the results are based on independent simple random samples of 1,000 persons in each year.)

(a) Should you make a one-sample z-test or a two-sample z-test? Why?

(b) Formulate the null and alternative hypotheses in terms of a box model. Do you need one box or two? Why? How many tickets go into each box? How many draws? What do the tickets show? What do the null and alternative hypotheses say about the box(es)?

(c) Can the difference between 60% and 54% be explained as a chance variation? Or was it the scandals? Or something else?

Answer:

Model: There are two samples, you need to make a two-sample z-test.
Model: There are two boxes. The 2005 box has a ticket for each person in the population, marked 1 for those who would rate clergymen "very high or high," and 0 otherwise. The 2005 data are like 1000 draws from the 2005 box. The 2000 box is set up the same way. The null hypothesis says that the percentage of 1's in the 2005 box is the same as in the 2000 box. The alternative hypothesis says that the percentage of 1's in the 2005 box is smaller than the percentage of 1's in the 2000 box.

The SD of the 2005 box is estimated from the data as $\sqrt{0.54 \times 0.46} \approx 0.50$. On this basis, the SE for the 2005 number is $\sqrt{1000} \times 0.50 \approx 16$: the number of respondents in the sample who rate clergymen "very high or high" is 540, and the chance error in that number is around 16. Convert the 16 to percent, relative to 1000. The SE for the 2005 percentage is estimated as 1.6%. Similarly, the SE for the 2000 percentage is about 1.5%.

The SE for the difference is computed from the square root law (p. 502) as $\sqrt{1.6^2 + 1.5^2} \approx 2.2\%$. The observed difference is $54\% - 60\% = -6\%$. On the null hypothesis, the expected difference is 0%. So $z = (\text{obs} - \text{exp})/\text{SE} = -6/2.2 \approx -2.7$, and $P \approx 3/1000$. The difference is real. What the cause is, the test cannot say.

Comment. Either a one-sided or a two-sided test can be used. Here, the distinction is not so relevant: for discussion, see chapter 29.

Chapter 28

Example:

(a) One day, upon tossing a single die 60 times, I got:

5 ones, 7 twos, 17 threes, 16 fours, 8 fives, and 7 sixes.

Compute χ^2 and find P for this experiment.

(b) Another day, upon tossing the same single die 600 times, I got:

90 ones, 110 twos, 100 threes, 80 fours, 120 fives, and 100 sixes.

Compute χ^2 and find P for this experiment.

(c) Now, compute the pooled χ^2 using the combined degrees of freedom, and find the pooled P-value.

Is the die biased, based on the combined evidence?

Answer:

(a) $\chi^2 = 13.2$, 5 deg.fr., $P \approx 2.2\%$. (b) $\chi^2 = 10$, 5 deg.fr., $P \approx 7.5\%$.
(c) Pooled $\chi^2 \approx 13.2 + 10 = 23.2$, $5 + 5 = 10$ deg.fr., $P \approx 1\%$;
conclude that the die is biased.

Calculations: (1st day) $\chi^2 = \frac{(-5)^2+(-3)^2+7^2+6^2+(-2)^2+(-3)^2}{10} = 13.2.$

(2nd day) $\chi^2 = \frac{(-10)^2+10^2+0^2+(-20)^2+20^2+0^2}{100} = 10.$

Chapter 29

Example:

State lotteries offer a customer the opportunity to bet on the outcome of a 3-digit number beween 000 and 999. It is assumed that all 1,000 numbers have approximately equal chance.

An investigator runs 51 tests: one test for each of the games in the 50 states and the District of Colombia. Three of these tests yield significant values of P: those in Tennessee, Oregon, and Oklahoma.

The investigator claims that the games in those three states are not random. Is this a valid conclusion?

Answer:

No.

This is a classic case of data mining. The tests were not singled out in advance. With a P value of 5%, about 5% of fair games will produce a result that appears to be not random. Remember, the P value is the chance of getting an extreme result when the null hypothesis is true. So about 5% of 51, 2 or 3 tests, should return significant results even if all 51 games are fair.

The error of data mining is naming the three jurisdictions after running the tests, deciding which tests to emphasize only after seeing the results.

The correct conclusion is that—generally speaking—the lottery games are random.

If further investigation were desired, he should run 51 more tests on new data and see if those 3 states end up with significant results.

Example:

A coin is tossed 400 times and lands heads 219 times. The investigator wants to decide if the chance of heads is 50%. He concludes—based on $P = 3\%$—that the coin is not fair.

Is this correct?

Answer:

This is not correct. The alternative hypothesis should be that the chance of heads is not 50%. This investigator used the alternative hypothesis that the chance of heads is greater than 50%. It was not known until the data were obtained whether the coin gets too many heads or too few heads. So a two-sided test was appropriate.

Using both tails of the normal distribution, $P = 6\%$, and the conclusion that the coin is fair was the correct one.

Calculation:

EV sum $= 400 \times 1/2 = 200$; SE sum $= \sqrt{400} \times 1/2 = 20 \times 1/2 = 10$;

$z = (218.5 - 200)/10 = 1.85$.

With 94% of the area in the middle, each tail has 3% of the area.

Example:

A study was done in 1962 showing that among *all* 3,600 seniors graduating from public high schools in the city of Rochester, New York, the average number of state capitals a student could name correctly was 17.1. Incidentally the SD of the number of state capitals that a student could name correctly was 6.5.

In 1990 a simple random sample of 169 seniors graduating from public high schools in Rochester, New York, showed that the average number of state capitals that a student in the sample could name correctly was 16.3 with an SD of 7.8.

Does this show that the average number all graduating seniors could name correctly has gone down? Or can the difference be explained as a chance variation?

Is it appropriate to make a two-sample z-test? If it isn't appropriate, explain why not.

Finally, formulate the null and alternative hypotheses and decide.

Answer:

A one-sample z should be used. The 1962 data were for the whole population, there is no need to find the SE standard error for the sample. The 1990 data were for a sample average. In that case, we find use the bootstrap and estimate the standard error for the sample average.

The null hypothesis is that the population average in 1990 is equal to 17.1. (The average has stayed the same, not gone down.)

The alternative hypothesis is that the population average in 1990 is less than 17.1.

The SE for the average in 1990 is (SD of box)/$(\sqrt{n}) = 7.8/\sqrt{169} = 7.8/13 = 0.6$.

$z = (16.3 - 171.)/0.6 = 0.8/0.6 = 1.33$, use 1.35.

$P = 9\%$. The difference can be explained as chance variation. It does not appear that the number has gone down.

Formulas

Tests of Significance: The Null and the Alternative

To make a test of significance, the null hypothesis has to be formulated as a statement about a box model. Usually, the alternative does too.

- The *null hypothesis* says that an observed difference just reflects chance variation.
- The *alternative hypothesis* says that the observed difference is real.

The null hypothesis expresses the idea that an observed difference is due to chance. To make a test of signficance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box; it says that the difference is real.

Testing Hypotheses: The Observed Significance Level

A test statistic is used to measure the difference between the data and what is expected on the null hypothesis.

z says how many SEs away an observed value is from its expected value, where the expected value is calculated using the null hypothesis. The formula is

$$z = \frac{\text{observed value} - \text{expected value}}{\text{standard error}}.$$

The observed significance level—often called the P-value—is the chance of getting a test statistic as extreme as or more extreme than the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller the chance is, the stronger the evidence against the null.

A small value of P indicates that an explanation saying that this is chance variation is unreasonable. We cannot accept the model stated in the null.

A large value of P could very well be due to chance variation and we accept the null as a reasonable model.

The P-value of a test is the chance of getting a big test statistic—assuming the null hypothesis to be right. P is not the chance of the null hypothesis being right.

Making a Test of Significance

Based on some available data, the investigator has to—

- translate the null hypothesis into a box model for the data;
- define a test statisic to measure the difference between the data and what is expected on the null hypothesis;
- compute the observed significance level P.

The choice of test statistic depends on the model and the hypothesis being considered.

Making the Decision

Many statisticians have a dividing line that indicates how small the observed significance level has to be before an investigator should reject the null hypothesis.

- If P is less than 5%, the result is called *statistically significant* and the null hypothesis is rejected.
- If P is greater than 5%, we accept the null and state that chance error is a reasoanable explanation for this result.

Tests of Significance Using Zero-one Boxes (Chapter 26, section 5)

If a problem involves classifying and counting, the z-test can be used. The box will contain 0's and 1's.

Once the null hypothesis has been translated into a box model, it is easy to test using

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

More Tests for Averages

The Standard Error for a Difference

The expected value for the difference of two independent quantities is $b - a$, where

- b is the expected value of the quantity from which the other is being subtracted;
- a is the expected value of the quantity which is being subtracted from the other.

The standard error for the difference of two independent quantities is $\sqrt{a^2 + b^2}$, where

- a is the SE for the first quantity;
- b is the SE for the second quantity.

Comparing Two Sample Averages

To test whether two populations (boxes) have the same average, one sets up the null hypothesis to be that the averages of the two boxes are equal. On this basis, the difference between the sample averages is expected to be 0, and the observed difference just reflects the luck of the draw. The alternative hypothesis says that the average of one box is smaller that the average of the other box. The two-sample z-statistic is computed as

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

The expected difference will in this case be equal to 0.

The two sample z-statistic is computed from—

- the sizes of the two samples,
- the averages of the two samples,
- the SDs of the two samples.

The test assumes two independent simple random samples.

Comparing Two Sample Percentages

The same procedure can be used to test whether the percentages of two boxes are equal.

The Chi-Square Test (χ^2-Test)

The χ^2-test is used to check whether a box model for classification involving more than two categories is appropriate in view of certain observed data. This is an approximation to the actual probabilities and may be trusted when each expected frequency in the table is 5 or more.

The χ^2-statistic is obtained by evaluating

$$\chi^2 \;=\; \text{sum of} \;\; \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

There is one term in the sum for each line in the table listing observed and expected frequencies.

This statistic measures the distance between observed and expected frequencies.

For the χ^2-test, P is approximately equal to the area to the right of the observed value for the χ^2-statistic, under the χ^2-curve with the appropriate number of degrees of freedom.

When the model is fully specified (no parameters to estimate),

$$\text{degrees of freedom} \;=\; \text{number of terms in } \chi^2 \;-\; \text{one.}$$

- The χ^2-test says whether the data are like the result of drawing at random from a box whose contents are given.

- The z-test says whether the data are like the result of drawing at random from a box whose average is given.

The ingredients of the χ^2-test are the basic data, chance model, frequency table, χ^2-statistic, degrees of freedom, and the observed significance level.

Independent Experiments

With independent experiments, the results can be pooled by adding up the separate χ^2-statistics; the degrees of freedom add up too.

Testing Independence

When testing independence in an $m \times n$ table (with no other constraints on the probabilities), there are $(m-1) \times (n-1)$ degrees of freedom.

A Closer Look at Tests of Significance

Was the Result Signficant?

Investigators should summarize the data, say what test was used, and report the P-value instead of just comparing P to 5% or 1%.

Data Snooping

Data-snooping makes P-values hard to interpret.

Was the Result Important?

The P-value of a test depends on the sample size. With a large sample, even a small difference can be "statistically significant," that is, hard to explain by the luck of the draw. This doesn't necessarily make it important. Conversely, an important difference may not be statistically significant if the sample is too small.

The Role of the Model

To make sense out of a test of significance, a box model is needed.

If a test of significance is based on data for the whole population, watch out.

If a test of significance is based on a sample of convenience, watch out.

Does the Difference Prove the Point?

A test of significance does not check the design of the study.