STAT 2
Lecture 30:
The z-test

Recall: confidence interval for a percentage

- I survey a simple random sample of 1000 voters. 357 of them support the Center Party. What is the 95% confidence interval for the percentage of all voters who support the Center Party?

Recall: confidence interval for a percentage

- For a SRS, estimated percentage is 357/1000*100% = 35.7%
- Estimate of standard error = sqrt(0.357*0.643/1000)*100% = 0.73%
- 95% confidence interval is 35.7 +/- 2*0.73 = 34.2% to 37.2%

Recall: other confidence levels

If we wanted to find a 90% confidence interval:

- Look up 5th and 95th percentiles in the standard normal table (since 90% of normal data is between these boundaries); get $z = +/- 1.65$
- CI ranges from 1.65 SEs below the estimate to 1.65 SEs above it

Recall: confidence interval for an average

- I survey a simple random sample of 1000 adults. They have an average income of $32000 a year, with a SD of $11000 a year. What is the 95% confidence interval for the average annual income of all adults?

Recall: confidence interval for an average

- Estimated average = \$32000
- Estimate of SE
  = \$11000/sqrt(1000) = \$350
- 95% CI is \$32000 +/- 2*\$350
  = \$31300 to \$32700

Recall: confidence interval from a series of measurements

- I make 100 independent measurement of the density of a bar of gold. The average of the density measurements is 19320 kg/m^3, while the SD of these measurements is 500 kg/m^3. What is the 99% CI for the density of the gold bar?

Recall: confidence interval from a series of measurements

- Estimate of density = 19320 kg/m^3
- Estimate of SE = 500/sqrt(100) = 50 kg/m^3
- $0.5^{th}/99.5^{th}$ percentile of normal gives z = +/- 2.58
- CI is 19320 +/- 2.58*50 = 19190 kg/m^3 to 19450 kg/m^3

Recall: other confidence intervals

- We can turn the CI for a percentage into the CI for the number (with some characteristic) by dividing by 100% and multiplying by the population size
- We turn the CI for an average into a CI for the total by multiplying by population size

Why might estimates be inaccurate?

- No true value: making a single estimate doesn't make sense
- Bias: expected value will not be equal to true value
- Not SRS/dependent/heteroscedastic/not like draws from box: hard to estimate SD and SE
- Outliers: estimate of SD may be poor, mean might not be appropriate
- Small sample: estimate of SD may be inaccurate, CLT may not hold

Today

- What is a hypothesis test?
- The z-test
- The P-value and statistical significance

I

The basic test: tickets in boxes

Example: tickets in a box

- I think the average of the tickets in a box is 10
- My friend thinks it's not 10
- To see who's right, we draw 100 tickets from the box: these have mean 45 and SD 10

Example: tickets in a box

Let's suppose the box average is really 50.

- The expected sample average is 50
- Standard error of sample average is estimated as 10/sqrt(100) = 1
- We observed the sample average was 45

Example: tickets in a box

- If the box average is really 50, then 45 is 5 standard errors below what we expect
- It's unlikely we'd get 5 SEs below the expected value just by chance
- Therefore, we have evidence the box average is not 50

Example: another box

- My friend and I argue about another box. Shew think the average of the tickets in the box is 0; I think it's not 0.
- We draw 100 tickets from the box; these have mean -5 and SD 50.

Example: another box

Suppose the box average is really 0.

- The expected sample average is 0
- Standard error of sample average is estimated as 50/sqrt(100) = 5
- Observed sample average = -5

Example: another box

- If the box average is really 0, then -5 is 1 SE below what we expect
- It's quite possible we'd be 1 SE below the expected value just by chance
- Therefore, we have insufficient evidence to show the box average is not 0

Example: another box

Note: This doesn't prove the box average is 0: it might be -5, or -10, or +2. Again, all we can say is that I can't show the box average isn't zero.

- The only way to find the true box average is to look at all the tickets, which is impractical

II

Hypotheses

Hypotheses

To formalise the problem, we state two hypotheses (theories):

- The *null hypothesis*: the box average is 0
- The *alternative hypothesis*: the box average is not 0

Testing hypotheses

- In a *hypothesis test*, we first assume the null hypothesis
- We then look at the data and see if it's consistent with the null hypothesis: are any differences just due to chance?
- If not, and the difference can be explained using the alternative hypothesis, we reject the null and accept the alternative

Example

I think a coin is biased, so I'm going to toss it many times and count the number of heads and tails

- Null hypothesis: for each toss, P(head) = P(tail) = ½
- Alternative: P(head) and P(tail) are not equal

Example

I wish to test a vaccine for cooties. The vaccine is given to a treatment group, while a placebo is given to a control group.

- Null hypothesis: which group you're in doesn't affect your chance of getting cooties
- Alternative: you're less likely to get cooties in the treatment group

How do we determine whether the data are consistent with the null?

- *Before* we collect the data, we choose a test statistic. This is what we will use to determine whether or not the data are consistent with the null hypothesis.
- Choose a test statistic that has a distribution we know, or can estimate. If we get an extreme value of that statistic, reject the null.

Example: coin tossing

- I will toss a coin 100 times, and use the number of heads as my test statistic (since we can use the binomial distribution or the normal approximation to find the probabilities of each number of heads)
- Null hypothesis: P(head) = P(tail) = ½, which means the number of heads has a binomial distribution with $n = 100$, $p = 0.5$
- Alternative: P(head) and P(tail) are not equal

Example: coin tossing

- I get 35 heads

From binomial distribution:

- P(35 or fewer heads) = 0.18%
- P(35 or fewer tails) = 0.18%

35 heads is very unlikely to happen by chance, so reject the null. The coin is biased.

III

The z-test

The z-statistic

The most widely used test statistic simply finds how many standard errors you are from the value expected under the null:

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

This statistic can be used for *z-tests*

The z-test

- If the observed value is normally distributed, then under the null hypothesis, the z-statistic has a *standard* normal distribution
- So the further the z-statistic is from zero, the stronger the evidence that the null hypothesis is false

Example

- I find a z-statistic of 2.5
- If the z-statistic really has a standard normal distribution, what's the chance of a random z-statistic being so far from zero?

Example

- Look at normal table:

P(Z < -2.5) = 0.62%

P(Z > 2.5) = 100 – 99.38 = 0.62%

The probability is 1.24%.

Such an extreme *z*-statistic is unlikely to happen by chance: reject the null hypothesis.

The P*-value*

- For the previous test, 1.24% is called the (two-sided) *P-value*
- The *P*-value of the test is the chance of getting a test statistic as extreme, or more extreme, than the one we observed, *if* the null hypothesis is true

Note: NOT the probability the null hypothesis is true

Example

I find a z-statistic of 0.6. What's the P-value?

- From normal table: $P(Z < -0.6) = P(Z > 0.6) = 27.43\%$
- Two-sided P-value is 54.86%

This z-statistic is not at all unusual. There's no evidence that we should reject the null hypothesis.

Example

I find a *z*-statistic of -0.02. What's the *P*-value?

- From normal table: $P(Z < -0.02) = P(Z > 0.02) = 49.20\%$
- Two-sided *P*-value is 98.4%

This *z*-statistic is close to zero, but that's consistent with the null. We do not reject the null hypothesis.

When can we use the z-test?

- Whenever the *z*-statistic has a standard normal distribution
- This occurs for z-statistics based on the sample percentage or average, provided the sample size is large. (There are other tests for moderate-sized samples.)

IV

Statistical significance

Testing statistical significance

There are several parts to any significance test:

- State hypotheses (if in doubt, relate to box models)
- Select test statistic
- Get data
- Calculate test statistic
- Find P-value
- Draw conclusions

Is the difference significant?

In old school statistics:

- A *P*-value of less than 5% is statistically significant: the difference can't be explained by chance, so reject the null hypothesis
- A *P*-value of more than 5% is not statistically significant: the difference can be explained by chance. We don't have enough evidence to reject the null hypothesis.

Is the difference significant?

In contemporary statistics, there's a growing trend to state the *P*-value and then draw conclusions based on the situation.

- Advantages: more informative, avoid an arbitrary cut-off
- Disadvantages: subjective, no one understands P-values (excluding you)

Note

If you reject the null hypothesis whenever the *P*-value is less than 5%, then *if* the null hypothesis is true, you will have about a 5% chance of rejecting it

- A *Type I error* occurs when you reject the null when it's in fact true

Example: hypotheses

I think that Berkeley students study, on average, 20 hours a week. My friend thinks the true average is much less than this.

- Null hypothesis: average hours studied per week = 20
- Alternative hypothesis: average hours studied per week < 20

One-sided and two-sided tests

My friend is willing to let me win even if the average number of hours studied is much more than 20.

- We perform a *one-sided*, or *one-tailed*, test: we will only reject the null if the average is low, not high

Example: test statistic

- I decide to survey a simple random sample of 200 Berkeley students
- My test statistic will be the z-statistic based on the average number of hours studied of the sample. I can use this because the sample average for a large sample is approx. normally distributed.

Example: the data

- My sample has a mean of 19 hours and an SD+* of 9 hours
- Estimated SE of sample average = 9/sqrt(200) = 0.636
- z-statistic = (19 – 20)/0.636 = -1.57

*it's slightly better to use the SD+ when using the sample SD to stand in for the population SD

Example: P-*value*

- From tables, $P(Z < -1.57) = 5.82\%$

We don't need $P(Z > 1.57)$ because it's a one-tailed test, so the *P*-value is 5.82%

- If we reject strictly if the *P*-value is less than 5%, we can't reject here: the difference isn't significant at the 5% level

How do we interpret this?

- I would say the difference was not statistically significant at the 5% level, and keep believing the true average is 20 hours
- My friend would say I got lucky: that the true value is less than 20 hours, and though the *P*-value was low, the test didn't quite pick it up. He might suggest another survey with a larger sample size: this would reduce the SE and make it more likely the test would pick up the difference

How do we interpret this?

- I would reply that even if there were a difference, it would be small: it may or may not be statistically significant, but it isn't *practically* significant
- The tests we've seen don't test for a practically significant difference (which is usually subjective)

Tomorrow

- Much more on interpreting tests
- Student's *t*-test