Introduction to Inference

Drawing Conclusions

1

- With an SRS of size n, from a N(p_0, SE), p = unknown. We want to test the hypothesis that p has a specified value p_0. The null hypothesis is

  H_0: p = p_0.

  The test is based on the sample proportion, $\hat{p}$.

  Normal calculations require standardized variables

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

It's important to note that we use p_0, not $\hat{p}$ to calculate SE_p

This **one-proportion z statistic** has the standard normal distribution when **H_0** is true. If the alternative is one-sided on the high side: Ha: p > p_0 then the P-value is the probability that a standard normal variable **Z** takes a value at least as large as the observed **z**. That is, **P = P(Z ≥ z)**

2

---

z Test for a Population Mean

- To test the hypothesis **H_0: p = p_0** based on an SRS of size n from a population with unknown proportion p and standard error **SE_p**, compute the **one-proportion z statistic**

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_0 q_0}{n}}}$$ ← **This is your Test statistic!!!**

In terms of a variable Z having the standard normal distribution, the P-value for a test of H_0 against

H_a: p > p_0 is P(Z ≥ z) **(one sided to the right)**

H_a: p < p_0 is P(Z ≤ z) **(one sided to the left)**

H_a: p ≠ p_0 is 2P(Z ≥ |z|) **(two sided)**

These p-values assume a normal distribution by the CLT and can be applied if the ***sample is large enough***

3

Tests with Fixed Significance Level, α

- A level of significance α says how much evidence we require. In terms of P-value, the outcome of a test is significant at level α if P-value ≤ α.

  α = 1 – C

  - Where C is the confidence

- **When Making a decision, use the following:**
  - **We reject H_0 if p-value ≤ α**
  - **We fail to reject H_0 if p-value > α**

4

Tests from CI's

- A two-sided test at significance level α can be carried out directly from a confidence interval with confidence level C = 1 - α.

  CI and Two-Sided Tests (p ≠ p_0)

- A level α two-sided significance test rejects a hypothesis H_0: p = p_0 exactly when the value p_0 falls outside a level 1 - α confidence interval for p.

5

Inference for a Population Proportion

Recall: Sample proportion,

$\hat{p}$ = the sample proportion

$$\hat{p} = \frac{\text{count of successes in the sample}}{\text{count of observations in the sample}}$$

p = the population proportion (the parameter)

SE_p = the standard error of the sampling distribution

$$\text{CI}: SE_p = \sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad \text{Significance Test}: SE_p = \sqrt{\frac{p_0 q_0}{n}}$$

Notice that we use a different *p* for *SE* depending on the type of inference being used!!!

6

Assumptions and Conditions for Inference about a Proportion

- **Independence Assumption**
  - Randomization condition
    - Were the data sampled at random or generated from a properly randomized experiment? Proper randomization is key – we really like to see an SRS if possible.
  - 10% condition
    - Are the samples drawn without replacement? If the sample exceeds 10%, chances are that some data may be in multiple samples.
  - **Plausible independence condition**
    - **Is there any reason to believe that the data values somehow affect each other? (Actually, this condition should always be checked even though we haven't specifically said this in the past.)**

7

Assumptions and Conditions for Inference about a Proportion

- **Large Enough Sample Assumption (or the Normality Assumption)**
  - Success/Failure condition
    - In order for us to use the CLT to assume the normality of the sampling distribution, we need to have a large enough sample. To check to see if we have a large enough sample, we use the success/failure condition.
    - We must expect to have at least 10 "successes" and at least 10 "failures." Recall that by tradition we arbitrarily label one alternative (usually the outcome being counted) as a "success" even if it's something bad (like a sick sea fan). The other alternative is the "failure."
    - This can be calculated using the formula:
      - For confidence intervals, we use: $n\hat{p} \geq 10$ or $n\hat{q} \geq 10$
      - For hypothesis tests, we use: $np_0 \geq 10$ or $nq_0 \geq 10$
  - **We always assume the Null-p to be true.**

8

Formulas for Inference of Proportions

- When the conditions are met, we either do a one-proportion confidence interval or a one-proportion z-test (notice that this is a z-test assuming that the sampling distribution is Normal).
- **Confidence Interval** – z^* is the upper (1 – C)/2

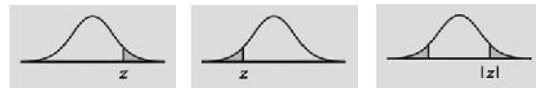
$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- For a **hypothesis test** where H_0: $p = p_0$, the z statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

9

H_a: $p > p_0$ is $P(Z \geq z)$ H_a: $p < p_0$ is $P(Z \leq z)$ H_a: $p \neq p_0$ is $2P(Z \neq |z|)$



One-sided to the right *One-sided to the left* *Two-sided*

Remember we determine the appropriate alternative hypothesis in context of the problem.

10

Example

- In sports, everyone has heard of the "Home Field Advantage" where teams tend to win more often when they play at home. If there were no home field advantage, teams would win 50% of the games at home. In 2002, the MLB played 2425 regular season games. It turns out that the home team won 1314 games at home, or about 54.2% of the time. Could this deviation from 50% be explained by natural sampling variability (by chance) or is there evidence to suggest that there really is a home field advantage, at least in professional baseball?
- Perform a hypothesis test to determine if there is a home field advantage.
- **What do you think??? Is it better to win more than 50% of the games at home?** More likely to win…

11

Example

Step 1: Identify population **Parameter**, state the null and alternative **Hypotheses**, determine what you are trying to do (and determine what the question is asking).

We want to know whether the home team in professional baseball is more likely to win. The **parameter** of interest is the proportion of home team wins. The population are home teams in MLB. With no advantage, we would expect that the proportion of home wins to be .50.

$$H_0: p = 0.50$$
$$H_A: p > 0.50$$

12

Example

Step 2: Verify the **Assumptions** by checking the conditions

Independence Assumption

•**Randomization condition:** Although we have all the data for 2002, we are interested in more than just that year, so it is not a randomized set of data. Although not randomized, 2002 can be a representative sample of recent and future games played.

•**10% condition:** We can reasonably assume that we observed fewer than 10% of all recent and future games played

•**Plausible independence condition:** Generally, the outcome of one game has no effect on the outcome of another game. But this may not always be true. For example, if a key player is injured, the probability that a team wins may be reduced. However, independence is roughly true.

13

Example

Step 2: Verify the **Assumptions** by checking the conditions

Normal Assumption (Large Enough Sample Assumption)

•**Success/Failure condition:**

$$np_0 \geq 10 \quad \text{and} \quad nq_0 \geq 10$$
$$np_0 = 2425(.5) = 1212.5 \geq 10$$
$$nq_0 = 2425(.5) = 1212.5 \geq 10$$

14

Example

Step 3: If conditions are met, **Name the inference procedure,** find the **Test statistic**, and **Obtain the p-value** in carrying out the inference:

Since the conditions are satisfied, it is appropriate to model the sampling distribution with the Normal distribution with mean p_0 and SE (p-hat).

- We will use a one-proportion z-test **(this is the name of the inference)**

$$SE_{\hat{p}} = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(.5)(.5)}{2425}} \approx .01015$$

$\hat{p} = 0.542$ (this was given in the beginning)

15

Example

Step 3: If conditions are met, **Name the inference procedure,** find the **Test statistic**, and **Obtain the p-value** in carrying out the inference:

Test Statistic: $z = \frac{\hat{p} - p_0}{SE(\hat{p})} = \frac{0.542 - 0.50}{0.01015} \approx 4.14$

p-value: $P(z > 4.14) \approx .000017$

Try to use the calculator to get the same data

GO to
- **STAT**
- **TESTS**
- **5: 1-PropZTest … enter po, x, n, and tail (>, <, or ÷**

16

Example

Step 3: If conditions are met, **Name the inference procedure,** find the **Test statistic**, and **Obtain the p-value** in carrying out the inference:

Test Statistic: $z = \frac{\hat{p} - p_0}{SE(\hat{p})} = \frac{0.542 - 0.50}{0.01015} \approx 4.14$

p-value: $P(z > 4.14) \approx .000017$

17

Example

Step 3: If conditions are met, **Name the inference procedure,** find the **Test statistic**, and **Obtain the p-value** in carrying out the inference:

Test Statistic: $z = \frac{\hat{p} - p_0}{SE(\hat{p})} = \frac{0.542 - 0.50}{0.01015} \approx 4.14$

p-value: $P(z > 4.14) \approx .000017$

Right Tail: normalcdf(z, 99)

Try to use the calculator to get the same data using tests

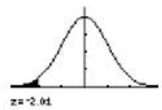
GO to
- **STAT**
- **TESTS**
- **5: 1-PropZTest … enter p_o, x, n, and tail (>, <, or ÷)**

18

TI-Calc: Finding the p-value

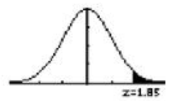
Left Tail: normalcdf(-99, -z)

p-value = 0.0222



Right Tail: normalcdf(z, 99)

p-value = 0.0322



Two-tailed

Do the same as with a right tailed or left-tailed test but **multiply your answer by 2.** Just recall that for a two-tailed test that:

- The p-value is the area to the left of the test statistic if the test statistics is on the left.
- The p-value is the area to the right of the test statistic if the test statistic is on the right.

19

Example

<u>M</u>ake a decision (reject or fail to reject H_0). **<u>S</u>tate your conclusion** in context of the problem using p-value.

The very small p-value says that if the true proportion of home team wins were 0.50, then the observed value of 0.542 or larger would occur in less than 1 out of 10,000 seasons. We reject the null hypothesis **(this is our decision)**. There is very strong evidence, p-value = 0.000017, that the true proportion of home team wins is more than 50% and there is a home field advantage **(conclusion)**.

20

Sample size for desired Margin of Error.

- To determine the sample size n that will yield a level C confidence interval for a population proportion p with specified margin of error m, set the following expression for the margin of error to be less than or equal to m, and solve for n:

$$z^* \sqrt{\frac{p^* q^*}{n}} \leq m,$$

where p* is a guessed value for the sample proportion. The margin of error will be less than or equal to m if you take the guess p* to be 0.5.

21