Elementary Statistics Lecture 6

Statistical Inference I

Chong Ma

Department of Statistics
University of South Carolina

Outline

Motivation-Defective Chips

A supplier of electronic chips for tablets claims that only 4% of his chips are defective. A manufacture tests 500 randomly selected chips from a large shipment from the supplier for potential defects. The manufacture will return the entire shipment if he finds more than 5% of the 500 sampled chips to be defective. Assume that the manufacture actually finds 5.6% of these chips are truly defective, is there anyway that you could convince the manufacture not returning the shipment on behalf of the supplier?

Questions

- Is the supplier's statement is true?
- What's the population proportion of defective chips in the entire shipment?

Statistical Inference

Statistical inference has two main purposes, one is for estimation of population parameters and the other is for testing hypotheses about the parameter values.

- Statistical inference methods use probability calculations that assume that the data were gathered with a random sample or a randomized experiment.
- The probability calculations refer to a sampling distribution of a statistic which is often approximately normal.

Outline

Point and Interval Estimates of Population Parameters

Point Estimate: **A single number that is our best guess for the parameter.** It is usually not sufficient, because it doesn't tell us how close the estimate is likely to fall to the parameter.

Interval Estimate: **An interval of numbers that is believed to contain the actual value of the parameter.** It is more useful, because it helps us gauge the accuracy of the point estimate.

Point Estimator

A point estimator is usually a statistic having a sampling distribution. A good point estimator has two desirable properties.

Property 1 **Unbiased**. A good point estimator has a sampling distribution that is centered at the parameter it tries to estimate.

Property 2 **Small standard deviation**. A good point estimator has a small standard deviation compared to other estimators.

Outline

Confidence Interval(CI)

An interval estimate indicates precision by giving an interval of numbers around the point estimate.

Confidence Interval

A confidence interval is an interval containing the most believable values for a parameter with a certain degree of confidence. It is composed by

$$[\text{point estimate} \pm \text{margin of error}]$$

The **margin of error** measures how accurate the point estimate is likely to be in estimating a parameter. It is usually a multiple of the standard deviation of the sampling distribution of the point estimate, when the sampling distribution is normal.

Assumption

- Data obtained by randomization.
- A large enough sample size such that the number of successes and the number of failures, that is, $n\hat{p}$ and $n(1-\hat{p})$ are both at least 15.

Confidence Interval at confidence level h

$$[\hat{p} \pm z_{\frac{1+h}{2}} se] = [\hat{p} \pm z_{\frac{1+h}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Where

Confidence level(h)	Error Probability	z-score	CI
0.90	0.10	1.645	$\hat{p} \pm 1.645se$
0.95	0.05	1.96	$\hat{p} \pm 1.96se$
0.99	0.01	2.58	$\hat{p} \pm 2.58se$

Properties of CI

Given the point estimate for the parameter is fixed, the cover of the confidence interval is usually determined by the margin of error, which is farther determined by the confidence level and sample size.

The **margin of error** for a confidence interval:

- Increases as the confidence level increases.
- Decreases as the sample size increases.

Interpretation of CI

For example, assume $[a, b]$ is a 95% confidence interval for population proportion p. The interpretation is **We are 95% confidence that the population proportion p is between a and b.**

Question: what does it mean to say that we have "95% confidence"?

- The confidence refers not to a probability for the population proportion p but rather to a probability that applies to the confidence interval **method** in its relative frequency sense.
- In the long run about 95% of those intervals would give correct results, containing the population proportion. About 5% of the time the confidence interval doesn't contain p.
- In frequentist statistics, you probably cannot say that the probability of 95% that p falls between a and b, because p is considered as a fix number and then probabilities can apply to statistics, not to parameters.

Interpretation of CI

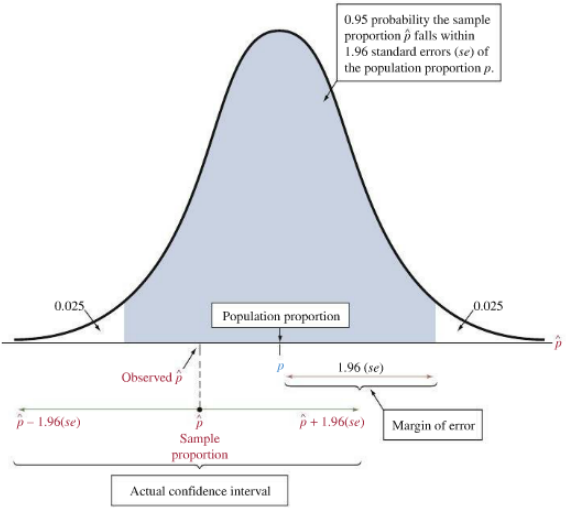


Figure 1: Sampling Distribution of Sample Proportion $\hat{p}$

Interpretation of CI

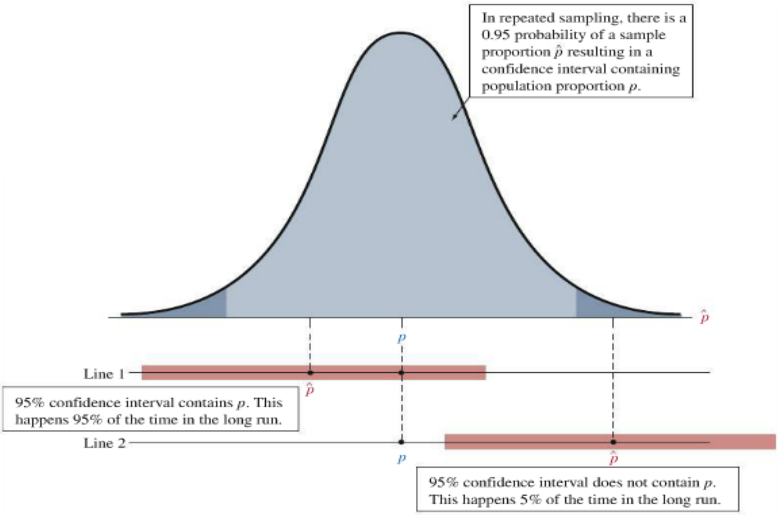


Figure 2: Sampling Distribution of Sample Proportion $\hat{p}$

Flu Shot

In a clinical study, 3900 subjects were vaccinated with a vaccine manufactured by growing cells in fertilized chicken eggs. Over a period of roughly 28 weeks, 24 of these subjects developed the flu.

a Find the point estimate of the population proportion that were vaccinated with the vaccine but still developed the flu.

b Find the standard error of this estimate.

c Find the margin of error for a 95% confidence interval.

d Construct 95% confidence interval for the population proportion. Interpret the interval.

Flu Shot

a Find the point estimate of the population proportion that were vaccinated with the vaccine but still developed the flu.

$$\hat{p} = 24/3900 = 0.0062$$

b Find the standard error of this estimate.

$$se = \sqrt{\hat{p}(1-\hat{p})/n} = 0.0013$$

c Find the margin of error for a 95% confidence interval.

$$1.96se = 1.96 * 0.0013 = 0.0025$$

d Construct 95% confidence interval for the population proportion. Interpret the interval.

$$[\hat{p} \pm 1.96se] = [0.0037, 0.0086]$$

Interpretation: We are 95% confidence that the population proportion vaccinated with the vaccine will still develop the flu is between 0.37% and 0.86%.

Outline

Significance Test

Step 1: Assumption

Step 2: Hypothesis

Step 3: Test Statistics

Step 4: P-Value

Step 5: Conclusion

The assumption commonly pertains to the method of data production(randomization), the sample size and the shape of the population distribution.

About a Population Proportion
Step 1: Assumption

- The variable is categorical.
- The data are obtained using randomization.
- The sample size is sufficiently large that the sampling distribution of sample proportion $\hat{p}$ is approximately normal, which holds when the *expected numbers of successes and failures are both at least 15*, that is, $np \geq 15$ and $n(1-p) \geq 15$.

H_0 : The parameter takes a particular value, usually representing *no effect*.

H_a : The parameter falls in some alternative range of values, usually representing an effect of some type.

About a Population Proportion
Step 2: Hypothesis Test

Null Hypothesis	Alternative Hypothesis
$H_0 : p = p_0$	$H_a : p > p_0$ (one sided)
$H_0 : p = p_0$	$H_a : p < p_0$ (one sided)
$H_0 : p = p_0$	$H_a : p \neq p_0$ (two sided)

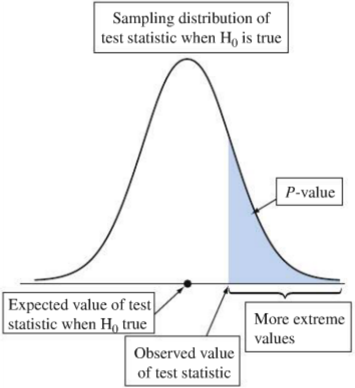
- Describe how far that point estimate falls from the parameter value given in the null hypothesis H_0.
- Measured by the number of standard errors between the point estimate and the parameter.

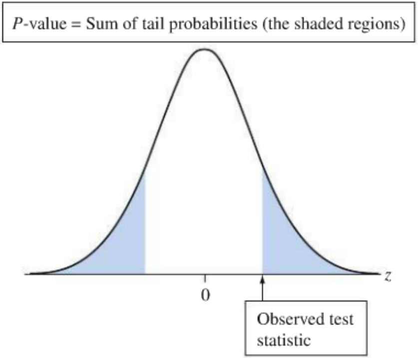
About a Population Proportion
Step 3: Test Statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where z-score measures the number of standard deviations between the sample proportion and the null hypothesis value p_0.

The **P-value** is the probability that the test statistic equals the observed value or a value even more extreme, presuming that the null hypothesis H_0 is true.





About a Population Proportion
Step 4: P-value

The **P-value** describes how unusual the data would be if H_0 were true, that is, the population proportion is p_0.

Null Hypothesis	Alternative Hypothesis	P-value
$H_0 : p = p_0$	$H_a : p > p_0$ (one sided)	$P(Z > z_{obs})$
$H_0 : p = p_0$	$H_a : p < p_0$ (one sided)	$P(Z < z_{obs})$
$H_0 : p = p_0$	$H_a : p \neq p_0$ (two sided)	$P(\|Z\| > \|z_{obs}\|) = 2P(Z > \|z_{obs}$

Smaller p-values give stronger evidence against H_0. How small would the p-value need to be to reject H_0? It is arbitrary but commonly we select the cutoff point $\alpha = 0.05$ that is called the **significance level**. That is, we decide to reject H_0 if p-value ≤ 0.05.

Significance level
: A number such that we reject H_0 if the p-value is less than or equal to that number.

Statistical significance
: The result of the test is called statistically significant when the data provide sufficient evidence to reject H_0 and support H_a.

P-value	Decision about H_0
≤ 0.05	Reject H_0
> 0.05	Do not reject H_0

Misinterpretations of Results of Significane Tests

- **"Do not reject H_0" does not mean "Accept H_0"** If your p-value is above the preselected significance level α, you cannot conclude that H_0 is correct. We can never accept a single value. A test merely indicates whether a particular parameter value is plausible. The population parameter might have many plausible values besides the number in H_0.
- **The p-value cannot be interpreted as the probability that H_0 is ture.** We are calculating probabilities about test statistic values, not about the parameter.
- **Some tests may be statistically significant just by chance.** If you run 100 times, even if all the null hypotheses are correct, you would expect to get p-values of 0.05 or less about $100 * 0.05 = 5$ times.

Defective Chips

A supplier of electronic chips for tablets claims that only 4% of his chips are defective. A manufacture tests 500 randomly selected chips from a large shipment from the supplier for potential defects. The manufacture will return the entire shipment if he finds more than 5% of the 500 sampled chips to be defective. Assume that the manufacture actually finds 5.6% of these chips are truly defective, is there anyway that you could convince the manufacture not returning the shipment on behalf of the supplier?

Questions

- Is the supplier's statement is true?
- What's the 95% confidence interval for the population proportion of defective chips in the entire shipment?

Defective Chips

Step 1: Assumption	1. Categorical Variable
	2. $np = 500 * 4\% = 20 \geq 15, n(1-p) = 480 \geq 15$
Step 2: Hypothesis test	$H_0 : p = 0.04$ Vs. $H_a : p > 0.04$
Step 3: Test statistic	$z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.016}{0.0102} = 1.57$
Step 4: P-value	$P(Z > z_{obs}) = P(Z > 1.57) = 0.058$
Step 5: Conclusion	At the significance level 0.05, do not reject H_0. We don't have sufficient evidence to conclude that the population proportion of defective chips in the entire shipment is greater than 4%.

Defective Chips

Note that $\hat{p} = 0.056$, the number of defective chips in the sample is $n\hat{p} = 500 * 0.056 = 28 > 15$, the number of non-defective chips is $n\hat{p} = 500 * 0.944 = 472 > 15$. The 95% confidence interval is

$$\begin{aligned}[\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}] &= [0.056 \pm 1.96\sqrt{\frac{0.056 \times 0.944}{500}}] \\ &= [0.056 \pm 0.02] \\ &= [0.036, 0.076]\end{aligned}$$

Interpretation: We have 95% confidence that the population proportion of defective chips in the shipment is between 3.6% and 7.6%. Because 4% falls in the 95% confidence interval, 4% is a plausible value for the population proportion of defective chips. The conclusion here coincides with the conclusion from the significance test.

Practice Problems for Exam III

Page 287-289 6.35, 6.45, 6.46, 6.49

Page 327-329 7.31, 7.33, 7.49, 7.50, 7.51

Page 379-381 8.69, 8.74, 8.75, 8.76, 8.77

Page 438-441 9.71, 9.72, 9.73, 9.99