THE GLOBE AND MAIL

Nearly half B.C. residents support Northern Gateway: poll

Lauren Krugel
Calgary— The Canadian Press
Published Thursday, Jan. 05, 2012

Nearly half of British Columbians support Enbridge Inc. ENB-T's proposed Northern Gateway project, which would pass through the province, according to an Ipsos Reid poll conducted on behalf of the pipeline company. The survey, released Thursday, suggests 48 per cent of B.C. residents back the controversial $5.5-billion project – 14 per cent of those strongly. On the flip side, 32 per cent oppose it – 13 per cent strongly.

The poll of 1,000 British Columbians, taken between Dec. 12 and Dec. 15, has a margin of error of plus or minus 3.1 percentage points 19 times out of 20.

Reason 3:
The Central Limit Theorem

The properties of normal distributions can be used even when outcomes are not normally distributed.

How can we be sure survey results match the true values in the population?

The Normal Distribution

Understanding | Z-Scores | Applicability | Confidence Intervals

Terminology

Population

$$z = \frac{x - \mu}{\sigma}$$

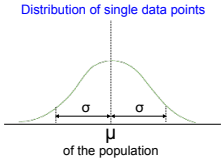
Sample

$$t = \frac{x - \bar{x}}{s}$$

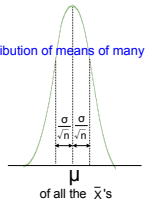
The Central Limit Theorem and Repeated Sampling

The properties of the normal distribution can be used even when populations are not normally distributed. This is because if one takes all the means from a large quantity of random samples from a non-normal distribution, those means will still be distributed normally, even though the underlying population is not.

Distribution of single data points



Distribution of means of many samples



Note how: if n=1 sample, the samples are single data points and $\sigma = \frac{\sigma}{\sqrt{n}}$

Terminology

Population

$$z = \frac{x - \mu}{\sigma}$$

Sample

$$t = \frac{x - \bar{x}}{s}$$

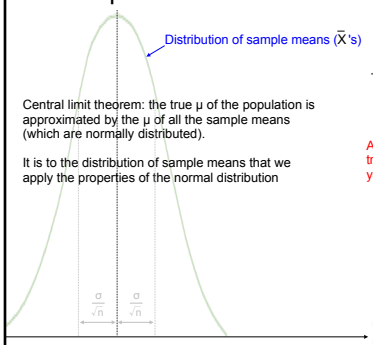
Confidence Intervals

A confidence interval calculates the range the true mean of the population falls within, based on a degree of likelihood you specify.

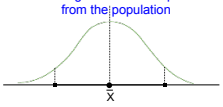
True but unknown value of μ is equal/close to:

μ of X̄'s

Distribution of sample means (X̄'s)



A single random sample from the population



Central limit theorem: the true μ of the population is approximated by the μ of all the sample means (which are normally distributed).

It is to the distribution of sample means that we apply the properties of the normal distribution

A confidence interval tells you how likely the true population mean falls within the range you've specified using the calculation:

$$\bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

where t^* is found by taking the t score corresponding to the level of confidence you want to have.

Confidence Intervals: Example

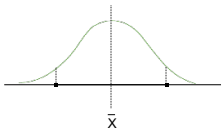
A confidence interval calculates the range the true mean of the population falls within, based on a degree of likelihood you specify.

1) You decide to test a paint to see how long it takes to dry. You paint 25 equal strips and discover that the average drying time is $\bar{x}$ = 13 minutes with a sample standard deviation of s = 3 minutes

Find a 95% confidence interval for μ, the true average drying time for this product.

$$\bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

"Margin of Error"



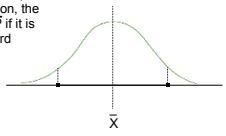
Confidence	T-score range (Number of Standard Deviations)
95%	± 1.96s
90%	± 1.64s
80%	± 1.28s
70%	± 1.04s
60%	± 0.84s

Confidence Intervals for rates

A confidence interval calculates the range the true mean of the population falls within, based on a degree of likelihood you specify.

2) Often we are looking for a confidence interval for a sample or population that measures the rate at which something occurs, such as approval ratings or rates of success. In this situation, the rate of approval is the rate or probability of success P or $\bar{P}$ if it is for the sample. The formula is the same except for standard deviation which is shown below.

$$\bar{P} \pm t \cdot \frac{\sqrt{\bar{P} \cdot \bar{Q}}}{\sqrt{n}}$$



Can we be 95% confident that in both polls the conservatives are truly ahead? What is the size of the ranges of the two 95% confidence intervals? Why?

Party	Gallup	Ipsos
Conservative	37%	36%
Liberals	35%	34%
Respondents	900	100

Exercises

3) A manufacturer of mixed nuts promises there will be at least 20% cashews in every can. If not, they will refund the $8 retail price. A consumer-research agency tests 150 cans of nuts and finds a mean of 22% cashews with a standard deviation of 1.5%. The proportions of cashews are normally distributed.

a) What percent of the samples have less than 20% cashews?
b) Based on the sample, what proportion of cans have between 15 and 30 percent cashews?
c) Can the firm be 95% confident that its guarantee is sound?
d) Based on the sample, what is the expected guarantee cost of refunding cans that don't meet the promise? Assume all customers will ask for their money back.

4) In a recent survey, 42% of high school graduates indicated that they expected to earn over $100,000 per year by the time they retire. 127 students were surveyed. Determine a 95% confidence interval for this rate.

5) A study of 120 patients suffering from low-back pain reported that the mean duration of the pain was 17.6 months, with a standard deviation of 5.1 months. Assuming that the duration of this problem is normally distributed in the population, determine a 99% confidence interval for the mean duration of low-back pain in the population. How large would you sample need to be to have a margin of error of 1 month or less?

6) A social scientist wants to estimate the average salary of office managers in a large city. She wants to be 95% confident that her estimate is correct. Assume that the salaries are normally distributed and that σ = $1,050. How large a sample must he take to obtain the desired information and be accurate within $200?

7)a) What is the sample standard deviation from the poll in the Globe and Mail article on the first slide of this note?
b) A survey of 625 people finds an approval rating for the Prime Minister of 56%. What is the standard deviation and the margin of error if you want a 95% confidence interval?

Hypothesis Testing

Continuous Distributions

Hypothesis testing tests the strength of an assertion based on a sample and threshold level you specify.

Statisticians and scientists often need to quantify the **reliability** of research studies and their conclusions. Using random sampling, you can test the strength of a statement or **hypothesis** by evaluating it in terms of its likelihood **using the properties of the normal distribution.**

8) A study concludes that Canadians are getting more obese. Critical to this conclusion was a finding in the study that indicated the average weight of Canadian adults was 102 kg with a standard deviation of 8 kg.

To test his claim, you randomly sample 100 Canadians and find that their average weight is 100 kg. This is less than the study, so...

Exercises

Continuous Distributions

9) A machine makes steel bearings with a mean diameter of 39mm and a standard deviation of 3mm. The bearing diameters are normally distributed. A quality-control technician found that in a sample of 50 bearings the mean diameter was 44mm. Test the significance with an alpha of 0.01. Decide if the machine needs to be adjusted or not.

10) Goodyear - a tire manufacturer - testifying at an environmental inquiry claims that its baseline tires last 100,000 km but most customers are able to get 110,000 km from them with a standard deviation of 19,500 km. If you randomly select 90 Goodyear tires coming to a waste facility and they have an average wear of 97,000 km, test the validity of Goodyear's assertion at the 1% level of significance.

11) Advertisers are concerned about falling TV viewership since fewer people means TV is less of an audience for their commercials. The broadcasters assure marketing firms that a television viewer still watches approximately 2.7 hours of TV per day with a standard deviation of .7 hours. You randomly sample 60 people and found average viewing time was 2.5 hours.
a) Give a 95% confidence interval for the population's true average viewing time.
b) Can you reject the broadcaster's hypothesis at the 1% level of significance?

12) Many studies have determined that patients suffering from low-back pain reported that the mean duration of the pain was 17.6 months, with a standard deviation of 9.3 months. You select a random group of 113 patients suffering from low-back pain, and introduce them to a new treatment plan. Patients who undergo this treatment had an average duration of 15.5 months. Can you be certain that the new treatment plan improves patient recovery, and isn't just a statistical sampling anomaly resulting in faster recovery?

Hypothesis Testing

Binomial Distributions

Hypothesis testing tests the strength of an assertion based on a sample and threshold level you specify.

Many situations involving **rates** are in fact **binomial distributions**. Because the **properties** of these distributions **do not change** between samples or large populations, the approach differs slightly. Moreover, you can use the normal to estimate, or use a spreadsheet to solve more accurate using the actual binomial calculations.

13) Before a recent advertising campaign, a children's breakfast cereal held 8% of the market. After the campaign, 18 families out of a sample of 200 families indicate they purchased the cereal. Was the advertising campaign a success?

Because someone buys or does not buy this cereal, it's binomially distributed. Ideally, use your spreadsheet to answer. If not, use the normal to estimate. Since n•P > 5 and n•Q > 5, it's okay to use the normal to estimate. Thus, μ = 16, and $\sigma = \sqrt{(n \bullet P \bullet Q)} \doteq 3.837$

Exercises

Binomial Distributions

14) A drug company tested a new drug on 250 pigs with swine flu. Historically, 20% of pigs contracting swine flue die from the disease. Of the 250 pigs treated with the new drug, 215 recovered. Test the significance of this result with an alpha of 0.01.

15) An insurance company claims that 38% of automobile accidents occur within 5 km of home. The company examined 400 recent accidents and found that 120 occurred within 5 km of the driver's home. Does this result support or refute the company's claim? Test with a significance of 0.01.

16) A new drug will not be considered for acceptance by Health Canada unless it causes serious side effects in less than 0.01% of the population. In a trial with 80,000 people, 9 suffered serious side effects. Test the significance of this result with alpha = 0.01. Do you recommend that this drug be accepted? Explain why or why not.

17) Do question #7 first. If you randomly selected 100 B.C. residents, and asked them if they approve of the Gateway pipeline now and 40% agree that they do (on some level), can you conclude that attitudes in the province towards the pipeline have changed since the article was written?

Answer Clues

1) 11.82 to 14.18 minutes

2) Gallop: Con. 33.8 to 40.2
Lib. 31.8 to 38.1

Ipsos: Con. 26.6 to 45.4
Lib. 24.7 to 43.3

3a) 9.1%
b) All of them
c) Yes μ = 21.76 to 22.24
d) $0.73 per can sold

4) 33.4% to 50.6%

5) 16.4 to 18.8

6) 106 people

7a) s=50%;
b) s=49.6%, margin of error = ±3.9%

9) P(z ≥ 11.79) = 0.0000 is < 1% so machine needs to be adjusted

10) P(z ≤ -1.46) = 0.0722 is > 1%, so you can't reject

11a) 2.52 and 2.88 hours

11b) P(z ≤ -2.21) = 1.34% which is > 1% so cannot reject H_0

12) P(z ≤ -2.4) = 0.82% which is < 1%, so reject H_0

14) P(z ≥ -2.37) = 0.88% is < 1% so reject H_0

15) P(z ≤ -3.30) is < 5% evidence refutes company's claim with high certainty

16) P(z ≥ 0.35) is > 0.01 so recommend.

17) P(z ≤ -1.6) = 5.48%, so not extremely rare. Depends on alpha.

Practice Assignment


Breast Cancer is the most wide-spread form of cancer among women. The American Center for Disease Control (CDC) collects data concerning all types of diseases and illnesses. The actual data for breast cancer occurrence in 2007 is 120.4 cases per 100,000 women, with a standard deviation of 0.31 (data collection centers cover 151.6 million Americans).

a) What is the probability that as a woman, you will get breast cancer in your lifetime? Can you calculate this from the data, why or why not?
b) What is the probability that in a community of 150,000 women, more than 185 will get breast cancer?
c) What is the probability that **exactly** 180 will get breast cancer? Why is this single value such a high probability?
d) If you have a close group of a dozen girlfriends who all met in University, what is the probability that someone will get breast cancer this year?
e) If you are a 17 year old female, and will live to the age of 85, do you expect to get breast cancer (assume the rate remains constant for your entire life)?
f) You manage the cancer ward in the region's hospital and it services a population of 1,000,000 Americans. Estimate the budget you must allocate to the ward to handle breast cancer treatment if the yearly cost of treatment is $275,000 per patient. You wish to be confident that you will have allocated enough money to handle all treatment 95% of the time.
e) 22.2 white American women per 100,000 die from breast cancer each year, with a sample standard deviation of 0.1. Conversely, 31.4 black American women die with a standard deviation of 0.41. Is this a fluke and the real average for black women is closer to white women? Find a 95% confidence interval to be sure. Surmise on reasons for the dramatic difference in the two s-values.

Source: U.S. Cancer Statistics Working Group. *United States Cancer Statistics: 1999–2007 Incidence and Mortality Web-based Report* . Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; 2010.