INTRODUCTION TO STATISTICS

Session 9

June 12th 2017

Madoka Takeuchi

Review
Population/ parameters

- A **population** is any large collection of objects or individuals
- A **parameter** is a descriptive summary describing the entire population.
  - Most of the time it is nearly impossible to find the real value of the parameter since we can not summarize the entire population→ We estimate the population parameter from the sample statistic
- A **sample** is a representative group drawn from the population.
- A **statistic** is a descriptive summary describing the sample.

Confidence intervals and hypothesis tests

- There are two ways to estimate/ learn about the population parameter-we have to perform statistical inference:
  1. Confidence intervals
     - The true mean or proportion for the population exists but we don't know it!
     - Using sample statistics, can get an estimate of where the population parameter is expected to be.
       - "we can be 95% confident that the proportion of students in the class are between 19 and 22 years old"
  2. Hypothesis test
     - "There is statistical evidence to conclude that the mean blood pressure of adults is approx 120/80."

Confidence Intervals

- Want to estimate the population mean but it is nearly impossible so obtain the mean of a sample randomly selected from the population.
  - There will be a range of values of the sample mean depending on the sample selected and can be relatively confident that the the population mean will be in the range- this **range** is called the confidence interval

i.e. A student was interested in whether or not students at SFC thought smoking on campus should be banned. The student randomly polled 100 students and 51% thought smoking should be banned. However the students reports that there is a margin of error of 5%.

- The Confidence interval for the unknown population proportion is 51% ±5 %

- we can be confident that between 46% and 56% of the students at SFC think smoking should be banned on campus

The margin of error will decrease as the the sample size is increased

Steps to calculate Confidence Intervals

The general formula for the confidence interval is

Sample estimate of the mean $\pm$ margin of error

1. Identify the sample statistic of interest that will estimate the population parameter- sample mean, sample proportion
2. Select the confidence interval- 90,95,99%
3. Find the margin of error
4. Specify the confidence interval

Confidence interval = sample statistic +Margin of error

Specific level of confidence

The general formula for the confidence interval does not give any specific information of the interval- want to be more specific

- define an interval that will contain the mean $\underline{X}$ proportion of the time- usually 95, 99%
- If repeated samples were taken and the 95% confidence interval is computed for each sample, 95% of the intervals would contain the population mean. Naturally, 5% of the intervals would not contain the population mean.
- It is natural to interpret a 95% confidence interval as an interval with a 0.95 probability of containing the population mean- FALSE
  - i.e the 95% CI is (10, 50)
    - Interpertation- there is reason to believe that the population mean lies between 10 and 50 in 95% of the CIs

Margin of error

-Range of values above and below the sample statistic is called the **margin of error**.

-margin of error can be defined by either of the following equations.

- Margin of error = Critical value x Standard deviation of the statistic
- Margin of error = Critical value x Standard error of the statistic

- Standard deviation vs standard error
  - The standard error is the sample Standard deviation

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Critical value

- We learned that the distribution of a statistic will be normal if the sample size is large- as a rough reference, 30 is considered large enough.
  - BUT if the data is skewed or has outliers, then a larger sample size is needed.
- When we have a normal distribution, the critical value is expressed as either a t score or z score-
  - When the sample size is small (<30) use t score, other wise use the z score

Steps to find the critical value

1. Compute α → 1-(desired confidence interval/100)
2. Compute the critical probability (p*)→
   - (p*)= 1-α/2

Confidence Interval	α	Critical probability (1-α/2)	α/2
90	0.1	0.95	0.05
95	0.05	0.975	0.025
99	0.01	0.995	0.005

Critical value- Z-score

3. Express the critical value as either a z or t score

- Z score- find the z score with the cumulative probability equal to the critical probability (p*)- use Z score table

90%=1.64 95%=1.96 99%=2.58

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0..6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

- We now have the critical value so can calculate the margin of error and the confidence interval
- The confidence interval is the sample statistic $\pm$ margin of error

Example

- Nine hundred people were randomly selected for a national survey. Among survey participants, the mean age was 57, and the standard deviation was 10. What is the margin of error, assuming a 95% confidence level
  - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 0.95 = 0.05
  - Find the critical probability (p*): p* = 1 - α/2 = 1 - 0.05/2 = 0.975
    - Z-score is 1.96
  - find the standard error of the mean, using the following equation: SE_x = s / sqrt(n) = 10 / sqrt(900) = 10 / 30 = 0.333
  - Margin of error = Critical value x Standard error
    = 1.96 * 0.333
    = 0.65268
  - CI = 57$\pm$0.65268
    = (56.34732, 57.65268)

Exercise(2)

- Want to estimate the average weight of an adult male in Fujisawa. We draw a random sample of 100 men from a population of 1,000,000 and weigh them. We find that the average man in our sample weighs 75 kg, and the standard deviation of the sample is 15 kg. What is the 95% confidence interval.

Answer

- Identify a sample statistic- mean weight (75)
- Select a confidence level. In this case, the confidence level is defined for us in the problem. 95% confidence level.
- Find Margin of error-
  - Find standard error. The standard error (SE) of the mean is: SE = s / sqrt(n) = 15 / sqrt(100) = 15/10 = 1.5
  - Find critical value- large sample size- use Z-score
  - Critical value = 1.96
  - Compute margin of error (ME): ME = critical value * standard error = 1.96 * 1.5 = 2.94
- Specify the confidence interval. The range of the confidence interval is *sample statistic* $\pm$ *margin of error*.

95% confidence interval is 75 $\pm$ 2.94

Introduction to hypothesis testing

- Hypothesis testing/significance testing is the use of statistics to determine the probability that a given hypothesis is true.

General idea of hypothesis testing-

- Make an initial assumption.
- Collect evidence (data).
- Based on the available evidence (data), decide to reject or not reject the initial assumption.

Example

- A researcher wants to know if the mean body temperature is really 98.6 degrees- he/she speculates that the temperature is not 98.6
  - Initial assumption- the mean body temperature is 98.6
  - Collect data- the researcher randomly samples 100 people and measures the body temp- find the mean to be 98.5
  - Make conclusion-
    - If the researcher concludes that the mean body temperature is not 98.6, then the initial assumption is rejected
    - If the researcher concludes that the mean body temperature is 98.6, then the initial assumption is NOT rejected.

Null and Alternative Hypothesis

- Null Hypothesis(H_0)- effect, relationship does not exist
  - Reflects no change, nothing of interest is occurring
  - "devil's advocate"- assumes whatever you are trying to prove did not happen
  - In the previous example- H_0: the mean body temperature is 98.6

- Alternative Hypothesis (H_A)- there is an observed effect
  - Opposite of the null hypothesis- what you are trying to prove
  - In the previous example- H_A: the mean body temperature is not 98.6

Exercise (3)

- An inventor has developed a new efficient car engine. He claims that the engine will run continuously for 100 hours on 10 gallons of regular gasoline. The inventor has a stock of 50 engines. To collect evidence, he selects a simple random sample of 5 engines for testing. The engines run for an average of 98 hours
  - What is the null and alternative hypothesis?