Comparing Two Means/Review of Basic Concepts of Statistical Inference

Lecture Notes II

Statistics 112, Fall 2002
Announcements

- We will finish the material in the first lecture notes at a later point in the semester. For right now, you are responsible for the material in Chapter 3.2.

- The first homework has been revised so that you should be able to complete it after today’s lecture. The original questions will be asked later in the semester.

- The TA Liang Wang’s office hour is Monday, 2-3 in Huntsman Hall 433. In addition, he will be available in the Stat Lab on Monday from 11-1.
Outline

- The problem of comparing the means of two populations.
- Review of basic concepts of statistical inference
  - Tests of significance
  - Confidence intervals
  - Power and inference as a decision (Chapter 6.4)
- Statistical inference for comparing the means of two populations.
- Reading for today’s lecture: Review Chapters 6.1-6.3 and 7.1. Read Chapter 6.4.
- Reading for Tuesday’s lecture: Chapter 7.2 and the part of Chapter 7.3 about robustness of normal inference procedures and the power of the two-sample $t$ test.
Comparing Two Population Means

- Many research questions can be formulated as comparisons of two population distributions. In particular, interest often focuses on comparing the means of two populations.

- Two Sample Experiments: Independent samples are taken from the two populations and we wish to make inferences about the difference in the means of the two populations on the basis of the sample. Examples:
  - How does the chance that a child receiving the Salk polio vaccination will develop polio compare to the chance that a child not receiving the vaccination will develop polio?
  - How does a bystander’s response to an emergency when alone compare to a bystander’s response when in a group of people (e.g., you are sitting in a waiting room for an interview and see white smoke begin pouring through a vent in the wall, do you report it)?
  - How does the mean IQ of adopted children compare to the mean IQ of children living with their natural parents?
  - How does employment at fast food restaurants change in a state that raises its minimum wage (NJ) compared to a state that does not change its minimum wage (PA)?
Among the standard personality inventories used by psychologists is the thematic apperception test (TAT). A subject is shown a series of pictures and is asked to make up a story about each one. Interpreted properly, the content of the stories can provide valuable insight into the subject’s mental well-being. The data on the next slide show the TAT results for 40 women, 20 of whom were the mothers of normal children and 20 the mothers of schizophrenic children. In each case the subject was shown the same set of ten pictures. The figures recorded were the numbers of stories (out of 10) that revealed a positive parent-child relationship, one where the mother was clearly capable of interacting with her child in a flexible, open-minded way.

Research Questions

The mean TAT score of the 20 mothers of normal children was 3.55. The mean TAT score of the 20 mothers was 2.10. Research questions:

- Does this provide convincing evidence that the population mean of TAT scores for mothers of normal children is different from the population mean of TAT scores for mothers of schizophrenics children.

- What range can we confidently assert that the differences in population means lies in?
Statistical Analysis of Research Questions

- Exploratory data analysis: Use graphical analysis and numerical summaries to look for outliers or interesting patterns in the data that may merit further investigation.

- Statistical inference.
  An inference is a conclusion that patterns in the data are present in some broader context.
  A statistical inference is an inference justified by a probability model linking the data to a broader context.

  - Do the two samples providing convincing evidence of a difference in population means? The method used to make statistical inferences for this question is called statistical hypothesis testing.

  - What range can we confidently assert that the differences in population means lies in? The method used to make statistical inferences for this question is called confidence intervals.
Graphical Methods for Comparing Two Samples

Look for outliers, patterns in the data.

- Relative frequency histograms.
- Back-to-back stem and leaf plots.
- Boxplots.
  - A central box spans the range between the 1st and 3rd quartiles (the “middle” of the data).
  - A line in the box marks the median.
  - Lines extending from the box show the largest and smallest points that are still within $1.5 \times IQR$ (interquartile range) of the central box.
  - Points that are more than $1.5 \times IQR$ of the central box are marked individually and should be investigated if possible.
- Normal quantile plots (to be discussed later).
Hypothesis Testing

- The goal of the research is presumably to make conclusions about the mean TAT scores for mothers of normal children and schizophrenic children in a broader population than just the mothers in this study (e.g., all children in the United States). Let $\mu_1$ and $\mu_2$ represent the means for mothers of normal and schizophrenic children in the broader population respectively. Question of research interest: does $\mu_1 = \mu_2$?

- Hypothesis Testing: The statement we are interested in testing is called the null hypothesis ($H_0$), e.g., $H_0 : \mu_1 = \mu_2$. The statement we hope or suspect is true is called the alternative hypothesis ($H_A$), e.g., $H_A : \mu_1 \neq \mu_2$.

- A test statistic measures the compatibility between the null hypothesis and the data. The test statistic is often chosen so that a large value provides evidence against the null hypothesis and a small value provides evidence for the null hypothesis test. A natural test statistic for this problem is

$$|z| = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma \bar{x}_1 - \bar{x}_2}$$
Suppose that the standard deviation of TAT scores for mothers of normal children in the broader population was known to be 1.88 and the TAT scores for mothers of schizophrenic children was known to be 1.55. Then for this sample,

\[
z = \frac{|3.55 - 2.1|}{\sqrt{\frac{1.88^2}{20} + \frac{1.55^2}{20}}} = 2.66
\]

What kind of evidence does \( z = 2.66 \) provide against the null hypothesis?
Tests of Significance

- In order to make statistical statements about how much
evidence a large value of a test statistic provides against a null
hypothesis, we need to formulate a probability model that links
the data to a broader context.

- Probability model: The TAT scores of mothers of normal
children and mothers of schizophrenic children are
\( N(\mu_1, \sigma_1^2) \) and \( N(\mu_2, \sigma_2^2) \) in the broader population. The
two samples are independent samples from these populations
of size \( n_1 \) and \( n_2 \) respectively.

- Once we have a probability model, we can figure out the
sampling distribution of the test statistic. The test statistic \( Z \)
(the random variable representing the test statistic for a
random sample) has the sampling distribution
\( N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2) \).

- \( p \)-value: What is the probability that the test statistic would be
as large as it is in the actual sample under repeated samples if
the null hypothesis were true? This probability is called the
\( p \)-value and a small \( p \)-value provides evidence against the null
hypothesis. For our setting, \( Z \) is \( N(0, 1) \) under the null
hypothesis and the \( p \)-value is \( P(|Z| \geq |z|) \).
• For the schizophrenia example, 
  \[ p = P(\lvert Z \rvert \geq 2.66) = 0.008. \]

• The use of the \( p \)-value as a measure of the evidence the data provides against the null hypothesis is called a significance test.

• Logic of significance tests: When we observe a large value of the test statistic \( \lvert Z \rvert \) (hence we have a low \( p \)-value), then either
  
  – The null hypothesis is true but we have observed something unusual.
  
  – The null hypothesis is not true.

  Because the test statistic \( \lvert Z \rvert \) is more likely to take on a large value if a member of the alternative hypothesis is true, e.g., \( \mu_1 - \mu_2 = 0.5 \) than if the null hypothesis \( \mu_1 - \mu_2 = 0 \) is true, a small \( p \)-value provides evidence against the null hypothesis.

• If the \( p \)-value is as small or smaller than \( \alpha \), we say that the data are statistically significant at level \( \alpha \).

• Traditionally, being statistically significant at level 0.05 or 0.01 is regarded as being decisive, but it is better to view the \( p \)-value as being a continuous scale of evidence against the null hypothesis.
• Statistical significance: A small $p$-value, such as the $p = 0.004$ for the TAT study, indicates that there is good evidence that the null hypothesis ($\mu_1 = \mu_2$) is not true, i.e., that there is an effect of having a schizophrenic child on a mother's TAT score. But it does not indicate how large that effect is. If $\mu_1 - \mu_2 = 0.001$, it may not be of much interest to psychologists studying how mothers with schizophrenic children are affected.

• A sufficiently large sample will declare very small effects statistically significant.

• Practical significance: This refers to whether the effect is of scientific interest. An effect size of $\mu_1 - \mu_2 = 0.001$ might not be considered practically significant.

• A sufficiently large sample will declare very small effects statistically significant.

• A confidence interval is much more informative about the practical significance of an effect found in the data.
Example: Two samples are taken to compare the mean SAT score of students using a coaching program ($\mu_1$) with students not using an SAT coaching program ($\mu_2$). The standard deviation of SAT scores is known to be 100 for students using the coaching program and students not using the coaching program. Consider the $p$-value for the test of $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 > \mu_2$ in each of the following situations.

- Samples of size 100 are taken. $\bar{x}_1 = 478$ and $\bar{x}_2 = 475$. The test statistic is $z = \frac{478 - 475}{100/\sqrt{100}} = 0.3$ and the $p$-value is $P(Z \geq 0.3) = 0.38$.

- Samples of size 1000 are taken. $\bar{x}_1 = 478$ and $\bar{x}_2 = 475$. The test statistic is $z = \frac{478 - 475}{100/\sqrt{1000}} = 0.95$ and the $p$-value is $P(Z \geq 0.95) = 0.17$.

- Samples of size 10,000 are taken. $\bar{x}_1 = 478$ and $\bar{x}_2 = 475$. The test statistic is $z = \frac{478 - 475}{100/\sqrt{10000}} = 3$ and the $p$-value is $P(Z \geq 3) = 0.001$. 
Confidence Intervals

• What range can we confidently assert that the differences in population means of TAT scores between mothers with normal and schizophrenic children lies in? We are interested in whether the range only includes differences that are of practical significance.

• The purpose of a confidence interval is to provide such a range.

• Any confidence interval has two parts: an interval computed from the data and a confidence level. The interval often has the form

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

• The confidence level states the probability that the method will give a correct answer (i.e., include the true difference in population means). If you use 95% confidence intervals often, in the long run, 95% of your intervals will contain the true difference in population means.
As with tests of significance, forming a confidence interval requires a probability model that relates the specific data obtained to a sampling mechanism. For the probability model stated before, a level \( C \) confidence interval for \( \mu_X - \mu_Y \) is

\[
(\bar{X} - \bar{Y}) \pm z^* \sqrt{\frac{\sigma^2_X}{n_x} + \frac{\sigma^2_Y}{n_y}}
\]

where \( z^* \) satisfies \( P(-z^* \leq Z \leq z^*) = C/100 \) for a \( N(0, 1) \) random variable \( Z \). You can obtain \( z^* \) from the bottom row in Table D.

- For the TAT study, a 95% confidence interval is

\[
(3.55 - 2.1) \pm 1.96 \sqrt{\frac{1.88^2}{20} + \frac{1.55^2}{20}} = (0.38, 2.52)
\]

- For the SAT example with 10,000 students, a 95% confidence interval is

\[
(478 - 475) \pm 1.96 \sqrt{100^2/10000} = (1.04, 5.96)
\]

Interpretation: “Yes, the mean score is higher after coaching but only by a small amount.”
Power of a Test

- Suppose we make a decision about whether to adopt the null or alternative hypothesis by choosing a fixed significance level \( \alpha \) and seeing whether the \( p \)-value is less than or equal to \( \alpha \). For example, the FDA decides whether to allow drugs to be sold based on such a criterion and for better or worse, journals often decide what to publish on this basis.

- The power of a test against a particular alternative is the probability that a fixed level \( \alpha \) significance test will reject \( H_0 \) when the particular alternative is true.

- Suppose that the \( p \)-value is greater than \( \alpha \) so that we do not reject the null hypothesis. If the power of the test against an alternative of interest is small, then we have learned little. Even if that alternative is true, then we would have a small chance of getting a \( p \)-value less than \( \alpha \).

- In designing a study, it is important to choose the sample size to be large enough so that the power against alternatives of interest is large.
Computing the Power

- State $H_0$, the particular alternative $H_a$ we want to detect and the significance level $\alpha$.
- Find the values of the test statistic that will lead us to reject $H_0$.
- Calculate the probability of observing these values of the test statistic when the alternative is true.
- Example: Consider the SAT example with sample size 100, $H_0 : \mu_1 = \mu_2$ and $H_A : \mu_1 - \mu_2 = 10$ and $\alpha = 0.05$. The $p$-value for the test statistic

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/100 + \sigma_2^2/100}}$$

is $P(Z \geq z)$ where $Z \sim N(0, 1)$ so that we will reject $H_0$ if $z \geq 1.645$ or equivalently if $\bar{x}_1 - \bar{x}_2$ is greater than

$$1.645\sqrt{100^2/100 + 100^2/100} = 23.26$$

Under $H_a$, $\mu_1 - \mu_2 \sim N(10, 100^2/100 + 100^2/100)$. **
The power is

\[
P(\bar{x}_1 - \bar{x}_2 \geq 23.26 \text{when } \mu_1 - \mu_2 = 10) = \]

\[
P\left(\frac{\bar{x}_1 - \bar{x}_2 - 10}{\sqrt{\sigma_1^2/100 + \sigma_2^2/100}} \geq \frac{23.26 - 10}{14.14}\right) = \]

\[
P(Z \geq 0.94) = 0.17
\]
We made the assumption that we knew the variance of the populations of TAT scores for mothers of normal and schizophrenic children. But in most two sample problems, we do not know the variance of the two populations.

The $t$-test statistic and associated $t$-confidence intervals provide valid inferences for the two-sample problem in which the populations are assumed to be normal $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Suppose first that we know that $\sigma_1^2 = \sigma_2^2 = \sigma$ although we do not know $\sigma$. The $t$-test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s_p$ is an estimate of $\sigma$,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Under the null hypothesis $H_0 : \mu_1 = \mu_2$, the random variable $T$ has a $t$-distribution with $n_1 + n_2 - 2$ degrees of freedom.

For the TAT study,

$$s_p^2 = \frac{19(1.88^2) + 19(1.55^2)}{20 + 20 - 2} = 2.97$$
and 
\[ t = \frac{3.55 - 2.1}{2.97 \cdot 5 \sqrt{\frac{1}{20} + \frac{1}{20}}} = 2.66 \]

- One-sided significance test: \( p \)-value equals \( P(T \geq t) \) when \( T \) has the \( t \)-distribution with \( n_1 + n_2 - 2 \) degrees of freedom.

- For the TAT study, \( p \)-value equals 0.006, slightly higher than if we assumed that we knew the variances of the population of TAT scores for mothers with normal and schizophrenic children were equal to their sample variances.

- Confidence interval: Under the same probability model, a level \( C \) confidence interval for \( \mu_1 - \mu_2 \) is

\[ (\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \]

where \( t^* \) satisfies \( P(-T \leq t^* \leq T) = C/100 \) \((T \sim t(n_1 + n_2 - 2))\).

- For the TAT study, a 95% confidence interval for \( \mu_1 - \mu_2 \) is

\[ (3.55 - 2.1) \pm 2.02 \times 2.97 \cdot 5 \sqrt{\frac{1}{20} + \frac{1}{20}} = (0.35, 2.55) \]