STAT 511

Lecture 18: Inferences Based on Two Samples Devore: Section 9.1-9.3

Prof. Michael Levine

March 29, 2019

z Tests and Confidence Intervals for a Difference Between Two Population Means

- An example of such hypothesis would be $\mu_1 - \mu_2 = 0$ or $\sigma_1 > \sigma_2$. It may also be appropriate to estimate $\mu_1 - \mu_2$ and compute its $100(1-\alpha)\%$ confidence interval
- Assumptions
  1. $X_1, \dots, X_m$ is a random sample from a population with mean μ_1 and variance σ_1^2
  2. $Y_1, \dots, Y_n$ is a random sample from a population with mean μ_2 and variance σ_2^2
  3. The X and Y samples are independent of one another

- The natural estimator of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$. To standardize this estimator, we need to find $E\,(\bar{X} - \bar{Y})$ and $V\,(\bar{X} - \bar{Y})$.
- $E\,(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$, so $\bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$.
- The proof is elementary: $E\,(\bar{X} - \bar{Y}) = E\,(\bar{X}) - E\,(\bar{Y}) = \mu_1 - \mu_2$

- The standard deviation of $\bar{X} - \bar{Y}$ is $\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
- The proof is also elementary:

$$V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

The standard deviation is the root of the above expression

The Case of Normal Populations with Known Variances

- As before, this assumption is a simplification.
- Under this assumption,

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \tag{1}$$

  has a standard normal distribution
- The null hypothesis $\mu_1 - \mu_2 = 0$ is a special case of the more general $\mu_1 - \mu_2 = \Delta_0$. Replacing $\mu_1 - \mu_2$ in (1) with Δ_0 gives us a test statistic.

- The following summary considers all possible types of alternatives:
  1. $H_a : \mu_1 - \mu_2 > \Delta_0$ has the P-value $1 - \Phi(z)$
  2. $H_a : \mu_1 - \mu_2 < \Delta_0$ has the P-value $\Phi(z)$
  3. $H_a : \mu_1 - \mu_2 \neq \Delta_0$ has the P-value equal to twice the area under the standard normal curve to the right of $|z|$.

Example I

- Analysis of a random sample of $m = 20$ specimens of cold-rolled steel gives the sample average yield strength $\bar{x} = 29.8$ ksi Another sample of $n = 25$ specimens of two-sided galvanized steel gives us $\bar{y} = 34.7$ ksi. The two variances are $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$ Note that $m \neq n$...it is not important now but will be later...
- The normality suggestion is based on some exploratory data analysis
- The hypotheses are $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 - \mu_2 \neq 0$

Example I

- The test statistic is

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = -3.66$$

- The corresponding P-value is $2[1 - \Phi(3.66)] \approx 0$ which implies rejection at *any* reasonable level.

Type II Error and the Choice of the Sample Size

- Consider the case of an upper-tailed alternative hypothesis $H_a : \mu_1 - \mu_2 > \Delta_0$.
- The rejection region is $\bar{x} - \bar{y} \geq \Delta_0 + z_\alpha \sigma_{\bar{X}-\bar{Y}}$. Therefore,

  $$P(\text{ Type II Error }) = P(\bar{X} - \bar{Y} < \Delta_0 + z_\alpha \sigma_{\bar{X}-\bar{Y}} \text{ when } \mu_1 - \mu_2 = \Delta^{'})$$
- Since $\bar{X} - \bar{Y}$ is normally distributed under the alternative $\mu_1 - \mu_2 = \Delta^{'}$ with mean $\Delta^{'}$ and standard deviation $\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$, we have

  $$\beta(\Delta^{'}) = \Phi\left(z_\alpha - \frac{\Delta^{'} - \Delta_0}{\sigma}\right)$$

- Similar results can be easily obtained for the other two possible alternatives. In particular, if $H_a : \mu_1 - \mu_2 < \Delta_0$, we have

$$\beta(\Delta^{'}) = 1 - \Phi\left(-z_\alpha - \frac{\Delta^{'} - \Delta_0}{\sigma}\right)$$

- If $\mu_1 - \mu_2 \neq \Delta_0$, the probability of Type II Error is

$$\Phi\left(z_{\alpha/2} - \frac{\Delta^{'} - \Delta_0}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\Delta^{'} - \Delta_0}{\sigma}\right)$$

Example

- Again, consider the steel data. Suppose that the probability of detecting a difference 5 between the two means should be .90. Can the .01 level test with $m = 20$ and $n = 25$ support this?
- For a two-sample test we have

$$\beta(5) = \Phi\left(2.58 - \frac{5-0}{1.34}\right) - \Phi\left(-2.58 - \frac{5-0}{1.34}\right) = .1251$$

- Because the rejection region is symmetric, we have $\beta(-5) = \beta(5)$, and, therefore, the probability of detecting a difference of 5 is $1 - \beta(5) = .8749$.
- We can conclude that slightly larger sample sizes are needed.

- To determine a sample size that satisfies $P($ Type II Error when $\mu_1 - \mu_2 = \Delta^{'}) = \beta$ we need to solve

$$\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} = \frac{(\Delta^{'} - \Delta_0)^2}{(z_\alpha + z_\beta)^2}$$

- For two equal sample sizes this yields

$$m = n = \frac{(\sigma_1^2 + \sigma_2^2)(z_\alpha + z_\beta)^2}{(\Delta^{'} - \Delta_0)^2}$$

Large-Sample Tests

- In this case, the assumption of normality for the data is unnecessary and variances σ_1^2, σ_2^2 need not be known
- This is because for large n the variable

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

is approximately standard normal

- Then, if the null hypothesis is $\mu_1 - \mu_2 = \Delta_0$, the test statistic

$$Z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

is approximately standard normal under the null hypothesis

- This test is usually appropriate if both $m > 40$ and $n > 40$

Example

- A company claims that its light bulbs are superior to those of its main competitor. If a study showed that a sample of $n_1 = 40$ of its bulbs has a mean lifetime of 647 hours of continuous use with a standard deviation of 27 hours , while a sample of $n_2 = 40$ bulbs made by its main competitor had a mean lifetime of 638 hours of continuous use with a standard deviation of 31 hours, does this substantiate the claim at the 0.05 level of significance?

- $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 - \mu_2 > 0$
- Calculations:

$$z = \frac{647 - 638}{\sqrt{\frac{27^2}{40} + \frac{31^2}{40}}} = 1.38$$

- Decision: H_0 cannot be rejected at $\alpha = 0.05$; the p-value is 0.0838

Confidence intervals for $\mu_1 - \mu_2$

- Since the test statistic Z that we just described is exactly normal when σ_1^2 and σ_2^2 are known,

$$P\left(-z_{\alpha/2} < Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

- The $100(1-\alpha)\%$ CI is easy to derive from this probability statement; it is

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sigma_{\bar{X}-\bar{Y}}$$

where $\sigma_{\bar{X}-\bar{Y}}$ is a square root expression.

- If both m and n are large, CLT implies that the normality assumption is not necessary and substitution of s_i^2 for σ_i^2, $i = 1, 2$ will produce an *approximately* $100(1-\alpha)\%$ CI
- More precisely, such an interval is

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

- Again, this result should be used only if both m and n exceed 40
- Note that this CI has a standard form of $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$

Example

- An experiment was conducted in which two types of engines, A and B, were compared. Gas mileage, in miles per gallon, was measured. 50 experiments were conducted using engine type A and 75 were done for engine type B. The gasoline used and other conditions were held constant. The average mileage for engine A was 36 mpg and the average for machine B was 42 mpg. Find an approximate 96% CI on $\mu_B - \mu_A$, where μ_A and μ_B are population mean gas mileage for machines A and B, respectively. Sample standard deviation are 6 and 8 for machines A and B, respectively.

- The point estimate of $\mu_B - \mu_A$ is $\bar{x}_B - \bar{x}_A = 42 - 36 = 6$. For $\alpha = 0.04$, we find the critical value $z_{.02} = 2.05$.
- Thus, the confidence interval is

$$6 \pm 2.05\sqrt{\frac{36}{50} + \frac{64}{75}} = (3.43, 8.57)$$

The Two-Sample t-test

- Assumptions:

  Both populations are normal, so that $X_1, \ldots, X_m$ is a random sample from a normal distribution and so is $Y_1, \ldots, Y_n$.

  The plausibility of these assumptions can be judged by constructing a normal probability plot of the x_is and another of the y_is.

- When the population distributions are both normal, the standardized variable

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

  has approximately t distribution with ν df

- ν can be estimated from data as

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

- ν has to be rounded down to the nearest integer...why not up?

- The *two-sample confidence interval for* $\mu_1 - \mu_2$ with confidence level $100(1-\alpha)\%$ is

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

  A one-sided confidence bound can also be calculated as described earlier.

- The two-sample t-test for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$ is conducted using the test statistic

$$t = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

Alternative hypothesis	P-values for approximate level α test
$H_a : \mu_1 - \mu_2 > \Delta_0$	Area under t_ν curve to the right of t
$H_a : \mu_1 - \mu_2 < \Delta_0$	Area under the t_ν curve to the left of t
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	Twice the area under t_ν curve to the right of $\|t\|$

Example

- The void volume within a textile fabric affects comfort flammability and insulation properties. The following is the summary information on air permeability for two different fabric types:

Fabric Type	Sample Size	Sample Mean	Sample Standard Deviation
Cotton	10	51.71	.79
Triacetate	10	126.14	3.59

- We assume that porosity distributions for both types of fabric are normal; then, the two-sample t-test(CI) can be used. Note that we do not assume anything about variances of the two populations concerned...
- The number of df is

$$\nu = \frac{\left(\frac{.6241}{10} + \frac{12.881}{10}\right)^2}{\frac{(.6241/10)^2}{9} + \frac{(12.881/10)^2}{9}} = 9.87$$

and we use $\nu = 9$

- The resulting CI is

$$51.71-136.14\pm(2.262)\sqrt{\frac{.6241}{10}+\frac{12.8881}{10}}=(-87.06,-81.80)$$

- Conclusion...

Pooled t test

- A simpler alternative test is available when it is known that $\sigma_1^2 = \sigma_2^2$.
- In this case, standardizing $\bar{X} - \bar{Y}$ we have

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

which has a standard normal distribution

- Instead of unknown $\sigma_1^2 = \sigma_2^2$ we use the weighted average

$$S_p^2 = \frac{m-1}{m+n-2} S_1^2 + \frac{n-1}{m+n-2} S_2^2$$

  In this case, both samples contribute equally to the common variance estimate.

- Substituting S_p^2 instead of σ_2^2 gives us a t distribution with $m+n-2$ degrees of freedom. This can serve as a basis for CI's and tests analogous to the one we described in the previous section.

Remarks

- Traditionally, this test has been recommended as the first to use when comparing two different means. It has a number of advantages over the two-sample t test: it is a likelihood ratio test, it is an **exact** test and it is easier to use!
- However, this test has a major problem: it is not robust to the violation of equality of variance assumption. When $\sigma_1^2 = \sigma_2^2$, its gains in power are small when compared to the two-sample t-test. That is why today it is often recommended to use the two-sample t test in most cases. It is especially true when the sample sizes are different.

- It may seem to be a plausible idea that one could first test a hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ and then choose the type of the t test based on the outcome.
- Unfortunately, the most common type of test used for this purpose (we will consider it at the very end of the course) is very sensitive to the violation of normality assumption and often not very reliable as a result.
- Yet another warning concerns normality of the data. If the distribution of the data is strongly asymmetric, both of these tests will prove unreliable. The alternative is to use a special class of tests that do not use any distribution assumptions at all (so-called nonparametric tests).

Analysis of Paired Data

- The data consists of n independently selected *pairs* (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) with $E\,X_i = \mu_1$ and $E\,X_i = \mu_2$. The differences $D_i = X_i - Y_i$ are assumed to be normally distributed with mean value $\mu_D = \mu_1 - \mu_2$ and variance σ_D^2. The last requirement is usually the consequence of X's and Y's being normally distributed themselves

Example

- Trace metals in drinking water affect the water flavor; moreover, the high concentrations can be a health hazard. Six river locations in South India were selected and the concentration of zinc in mg/L determined for both surface water and bottom water at each location. Presumably, there is some connection between surface water and bottom water concentrations...

The Paired t test

- The test considered is $H_0 : \mu_D = \Delta_0$ where $D = X - Y$
- The test statistic is
$$t = \frac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$$
where $\bar{d}$ and s_D are the sample mean and standard deviation of d_i's
- Note that the old method of computing the variance of the difference does not work anymore since X and Y are NOT independent

- The **differences** themselves are independent. Thus, hypotheses about $\mu_D = \mu_1 - \mu_2$ can be tested using a one-sample t-test with D_i's as data
-
    - $H_0 : \mu_D = \Delta_0$
    - Test statistic value is

$$t = \frac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$$

where $\bar{d}$ and s_D are the sample mean and standard deviation of the d_i's.

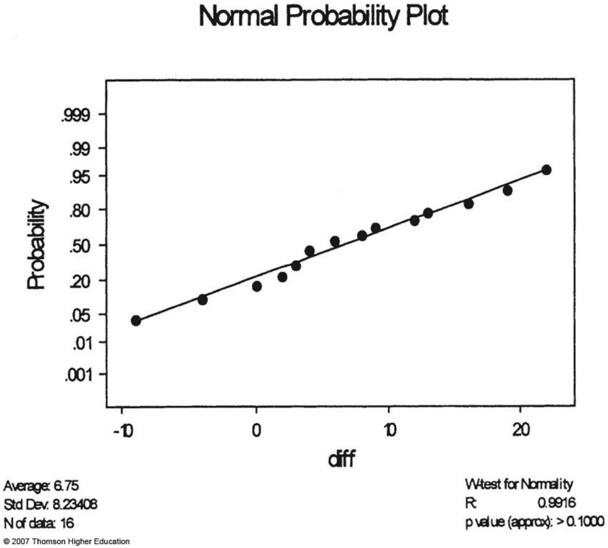
- Possible alternatives are $H_a : \mu_D > \Delta_0$, $H_a : \mu_D < \Delta_0$ and $H_a : \mu_D \neq \Delta_0$.
- Their corresponding P-values are the area under t_{n-1} curve to the right of t, the area under t_{n-1} curve to the left of t, and twice the area under the t_{n-1} curve to the right of $|t|$

Example

- Musculoskeletal neck-and-shoulder disorders are common among people who perform repetitive tasks using visual display units. A sample of $n = 16$ subjects is used to obtain data on whether more varied work conditions would have any impact on arm movement.
- Each observation is the amount of time (proportion of the total time observed) spent with the arm elevation below 30 degrees. The two measurements from each subject were obtained 19 months apart. Work conditions changed during this period. Does the data suggest that the true average time spent with arm elevation below 30 degrees differs after the change?
- The formal hypothesis is $H_0 : \mu_D = 0$ vs. $H_a : \mu_D \neq 0$.

Example

- The normal probability plot is



Average: 6.75
Std Dev: 8.23408
N of data: 16

W-test for Normality
R: 0.9916
p value (approx): > 0.1000

© 2007 Thomson Higher Education

- The box plot suggests that we have > 0

A confidence interval for μ_D

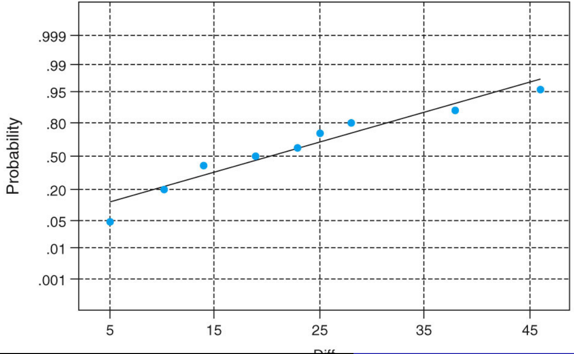
- The paired t confidence interval for μ_D is

$$\bar{d} \pm t_{\alpha/2,n-1} \cdot s_D/\sqrt{n}$$

- Note that for large n this interval is valid without any restrictions on the distribution of differences. The same is not true if n is relatively small

Example

- Adding computerized medical images to a database is potentially useful for physicians but the issue of efficiency of access needs to be investigated. The time to retrieve an image from a library of slides vs. retrieving the same image from a computer database for 13 computer-proficient medical professionals has been recorded.
- The normal probability plot



Example

- The needed $\bar{d} = 20.5$ sec and $s_D = 11.96$ while the number of df is $n - 1 = 12$.
- The 95% confidence interval is

$$\bar{d} \pm t_{\alpha/2,n-1} \cdot \frac{s_D}{\sqrt{n}} = 20.5 \pm (2.179) \cdot \frac{11.96}{\sqrt{13}} = (13.3, 27, 7)$$

- The confidence interval is rather large - the consequence of large standard deviation. However, 0 lies outside the confidence interval suggesting $\mu_D > 0$.

Paired Data and Two-Sample t Procedures

- The main difference between the paired data t test and the standard t test lies in how we estimate $V(\bar{X}-\bar{Y})$. In the independent case we have $V(\bar{X}-\bar{Y}) = V(\bar{X}) + V(\bar{Y})$ but...
- In the paired data case

$$V(\bar{X}-\bar{Y}) = V\left(\frac{1}{n}\sum D_i\right) = \frac{V(D_i)}{n} = \frac{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}{n}$$

- In the above, $\rho = Corr(X, Y) = Cov(X, Y)/[\sqrt{V(X) \cdot V(Y)}]$; in general,

$$V(X \pm Y) = \sigma_1^2 + \sigma_2^2 \pm 2\rho\sigma_1\sigma_2$$

  In the independence case, $\rho = 0$ and, therefore, $V(X \pm Y) = V(X) + V(Y)$.
- Thus, using a regular t-test in paired data means overestimating the variance of $\bar{X} - \bar{Y}$, and consequently, underestimating the significance of the data

Pros and Cons of Pairing

- For great heterogeneity and large correlation within experimental units, the loss in degrees of freedom will be compensated for by an increased precision associated with pairing (use pairing).
- If the units are relatively homogeneous and the correlation within pairs is not large, the gain in precision due to pairing will be outweighed by the decrease in degrees of freedom (use independent samples).