Mathematics for Data Science

Lecture 6: Statistical Inference for Two Samples

Mohamad GHASSANY

EFREI Paris

Table of contents i

The safety of drinking water is a serious public health issue. An article in the *Arizona Republic* on May 27, 2001, reported on arsenic contamination in the water sampled from 10 communities in the metropolitan Phoenix area and 10 communities from rural Arizona. The data showed dramatic differences in the arsenic concentration, ranging from 3 parts per billion (ppb) to 48 ppb. This article suggested some important questions. Does real difference in the arsenic concentrations in the Phoenix area and in the rural communities in Arizona exist? How large is this difference? Is it large enough to require action on the part of the public health service and other state agencies to correct the problem? Are the levels of reported arsenic concentration large enough to constitute a public health risk?

Some of these questions can be answered by statistical methods. If we think of the metropolitan Phoenix communities as one population and the rural Arizona communities as a second population, we could determine whether a statistically significant difference in the mean arsenic concentration exists for the two populations by testing the hypothesis that the two means, say, μ_1 and μ_2, are different. This is a relatively simple extension to two samples of the one-sample hypothesis testing procedures of Chapter 9. We could also use a confidence interval to estimate the difference in the two means, say, $\mu_1 - \mu_2$.

The arsenic concentration problem is very typical of many problems in engineering and science that involve statistics. Some of the questions can be answered by the application of appropriate statistical tools, and other questions require using engineering or scientific knowledge and expertise to answer satisfactorily.

Inference on the Difference in Means of Two Normal Distributions, Variances Known

The previous two chapters presented hypothesis tests and confidence intervals for a single population parameter (the mean μ, the variance σ^2, or a proportion p). This chapter extends those results to the case of two independent populations.

The general situation is shown in Fig. 10-1. Population 1 has mean μ_1 and variance σ_1^2, and population 2 has mean μ_2 and variance σ_2^2. Inferences will be based on two random samples of sizes n_1 and n_2, respectively. That is, $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of n_1 observations from population 1, and $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample of n_2 observations from population 2.

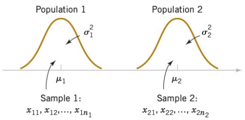

FIGURE 10-1 Two independent populations.

In this section, we consider statistical inferences on the difference in means $\mu_1 - \mu_2$ of two normal distributions where the variances σ_1^2 and σ_2^2 are known. The assumptions for this section are summarized as follows.

Assumptions for Two-Sample Inference

(1) $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
(2) $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
(3) The two populations represented by X_1 and X_2 are independent.
(4) Both populations are normal.

A logical point estimator of $\mu_1 - \mu_2$ is the difference in sample means $\bar{X}_1 - \bar{X}_2$. Based on the properties of expected values,

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

and the variance of $\bar{X}_1 - \bar{X}_2$ is

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Based on the assumptions and the preceding results, we may state the following.

The quantity

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (10\text{-}1)$$

has a $N(0, 1)$ distribution.

This result will be used to develop procedures for tests of hypotheses and to construct confidence intervals on $\mu_1 - \mu_2$. Essentially, we may think of $\mu_1 - \mu_2$ as a parameter θ where estimator is $\hat{\Theta} = \bar{X}_1 - \bar{X}_2$ with variance $\sigma^2_{\hat{\Theta}} = \sigma_1^2/n_1 + \sigma_2^2/n_2$. If θ_0 is the null hypothesis value specified for θ, the test statistic will be $(\hat{\Theta} - \theta_0)/\sigma_{\hat{\Theta}}$. Notice how similar this is to the test statistic for a single mean used in Equation 9-8 of Chapter 9.

HYPOTHESIS TESTS ON THE DIFFERENCE IN MEANS, VARIANCES KNOWN

We now consider hypothesis testing on the difference in the means $\mu_1 - \mu_2$ of two normal populations. Suppose that we are interested in testing whether the difference in means $\mu_1 - \mu_2$ is equal to a specified value Δ_0. Thus, the null hypothesis will be stated as $H_0 : \mu_1 - \mu_2 = \Delta_0$ Obviously, in many cases, we will specify $\Delta_0 = 0$ so that we are testing the equality of two means (i.e., $H_0 : \mu_1 = \mu_2$). The appropriate test statistic would be found by replacing $\mu_1 - \mu_2$ in Equation 10-1 by Δ_0: this test statistic would have a standard normal distribution under H_0.

Tests on the Difference in Means, Variances Known

Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (10\text{-}2)$$

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$, $P = 2\left[1 - \Phi\left(\lvert z_0 \rvert\right)\right]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Example 10-1 **Paint Drying Time** A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

We apply the seven-step procedure to this problem as follows:

1. **Parameter of interest:** The quantity of interest is the difference in mean drying times, $\mu_1 - \mu_2$, and $\Delta_0 = 0$.
2. **Non hypothesis:** $H_0: \mu_1 - \mu_2 = 0$, or $H_0: \mu_1 = \mu_2$.
3. **Alternative hypothesis:** $H_1: \mu_1 > \mu_2$. We want to reject H_0 if the new ingredient reduces mean drying time.
4. **Test statistic:** The test statistic is

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad \text{where } \sigma_1^2 = \sigma_2^2 = (8)^2 = 64 \text{ and } n_1 = n_2 = 10.$$

5. **Reject H_0 if:** Reject $H_0: \mu_1 = \mu_2$ if the P-value is less than 0.05.
6. **Computations:** Because $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, the test statistic is

$$z_0 = \frac{121 - 112}{\sqrt{\dfrac{(8)^2}{10} + \dfrac{(8)^2}{10}}} = 2.52$$

7. **Conclusion:** Because $z_0 = 2.52$, the P-value is $P = 1 - \Phi(2.52) = 0.0059$, so we reject H_0 at the $\alpha = 0.05$ level.

Practical Interpretation: We conclude that adding the new ingredient to the paint significantly reduces the drying time. This is a strong conclusion.

Inference on the Difference in Means of two Normal Distributions, Variances Unknown

We now consider tests of hypotheses on the difference in means $\mu_1 - \mu_2$ of two normal distributions where the variances σ_1^2 and σ_2^2 are unknown. A t-statistic will be used to test these hypotheses. As noted earlier and in Section 9-3, the normality assumption is required to develop the test procedure, but moderate departures from normality do not adversely affect the procedure. Two different situations must be treated. In the first case, we assume that the variances of the two normal distributions are unknown but equal; that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. In the second, we assume that σ_1^2 and σ_2^2 are unknown and not necessarily equal.

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Suppose that we have two independent normal populations with unknown means μ_1 and μ_2, and unknown but equal variances, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We wish to test

$$H_0: \mu_1 - \mu_2 = \Delta_0$$
$$H_1: \mu_1 - \mu_2 \neq \Delta_0 \qquad (10\text{-}11)$$

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample of n_1 observations from the first population and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample of n_2 observations from the second population. Let $\bar{X}_1$, $\bar{X}_2$, S_1^2, and S_2^2 be the sample means and sample variances, respectively. Now the expected value of the difference in sample means $\bar{X}_1 - \bar{X}_2$ is $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$, so $\bar{X}_1 - \bar{X}_2$ is an unbiased estimator of the difference in means. The variance of $\bar{X}_1 - \bar{X}_2$ is

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

It seems reasonable to combine the two sample variances S_1^2 and S_2^2 to form an estimator of σ^2. The **pooled estimator** of σ^2 is defined as follows.

Pooled Estimator of Variance

The **pooled estimator** of σ^2, denoted by S_p^2, is defined by

$$S_p^{\,2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \qquad (10\text{-}12)$$

Now we know that

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a $N(0, 1)$ distribution. Replacing σ by S_p gives the following.

Given the assumptions of this section, the quantity

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (10\text{-}13)$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Tests on the Difference in Means of Two Normal Distributions, Variances Unknown and Equal*

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$.

Test statistic:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{10-14}$$

Alternative Hypotheses	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert t_0 \rvert$ and probability below $-\lvert t_0 \rvert$	$t_0 > t_{\alpha/2, n_1+n_2-2}$ or $t_0 < -t_{\alpha/2, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha, n_1+n_2-2}$

Example 10-5 **Yield from a Catalyst** Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently used; but catalyst 2 is acceptable. Because catalyst 2 is cheaper, it should be adopted, if it does not change the process yield. A test is run in the pilot plant and results in the data shown in Table 10-1. Figure 10-2 presents a normal probability plot and a comparative box plot of the data from the two samples. Is there any difference in the mean yields? Use $\alpha = 0.05$, and assume equal variances.

TABLE • 10-1 Catalyst Yield Data, Example 10-5

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

The solution using the seven-step hypothesis-testing procedure is as follows:

1. **Parameter of interest:** The parameters of interest are μ_1 and μ_2, the mean process yield using catalysts 1 and 2, respectively, and we want to know if $\mu_1 - \mu_2 = 0$.
2. **Null hypothesis:** H_0: $\mu_1 - \mu_2 = 0$, or H_0: $\mu_1 = \mu_2$
3. **Alternative hypothesis:** H_1: $\mu_1 \neq \mu_2$
4. **Test statistic:** The test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

5. **Reject H_0 if:** Reject H_0 if the P-value is less than 0.05.
6. **Computations:** From Table 10-1, we have $\bar{x}_1 = 92.255$, $s_1 = 2.39$, $n_1 = 8$, $\bar{x}_2 = 92.733$, $s_2 = 2.98$, and $n_2 = 8$. Therefore

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(7)(2.39)^2 + 7(2.98)^2}{8 + 8 - 2} = 7.30$$

$$s_p = \sqrt{7.30} = 2.70$$

and

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{2.70\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.70\sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$

7. **Conclusions:** Because $|t_0| = 0.35$, we find from Appendix Table V that $t_{0.40,14} = 0.258$ and $t_{0.25,14} = 0.692$. Therefore, because $0.258 < 0.35 < 0.692$, we conclude that lower and upper bounds on the P-value are $0.50 < P < 0.80$. Therefore, because the P-value exceeds $\alpha = 0.05$, the null hypothesis cannot be rejected.

Practical Interpretation: At the 0.05 level of significance, we do not have strong evidence to conclude that catalyst 2 results in a mean yield that differs from the mean yield when catalyst 1 is used.

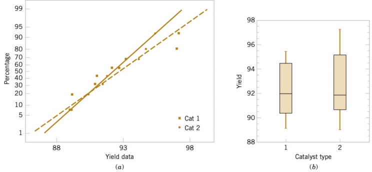

FIGURE 10-2 Normal probability plot and comparative box plot for the catalyst yield data in Example 10-5. (a) Normal probability plot. (b) Box plots.

Case 2: $\sigma_1^2 \neq \sigma_2^2$

In some situations, we cannot reasonably assume that the unknown variances σ_1^2 and σ_2^2 are equal. There is not an exact t-statistic available for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ in this case. However, an approximate result can be applied.

Case 2: Test Statistic for the Difference in Means, Variances Unknown and Not Assumed Equal

If $H_0: \mu_1 - \mu_2 = \Delta_0$ is true, the statistic

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (10\text{-}15)$$

is distributed approximately as t with degrees of freedom given by

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \quad (10\text{-}16)$$

If v is not an integer, round down to the nearest integer.

Therefore, if $\sigma^2_1 \neq \sigma^2_2$, the hypotheses on differences in the means of two normal distributions are tested as in the equal variances case except that T_0^* is used as the test statistic and $n_1 + n_2 - 2$ is replaced by v in determining the degrees of freedom for the test.

The pooled t-test is very sensitive to the assumption of equal variances (so is the CI procedure in section 10-2.3). The two-sample t-test assuming that $\sigma_1^2 \neq \sigma_2^2$ is a safer procedure unless one is very sure about the equal variance assumption.

Inference on the Variances of Two Normal Distributions

F DISTRIBUTION

Suppose that two independent normal populations are of interest when the population means and variances, say, μ_1, σ_1^2, μ_2, and σ_2^2, are unknown. We wish to test hypotheses about the equality of the two variances, say, $H_0: \sigma^2_1 = \sigma^2_2$. Assume that two random samples of size n_1 from population 1 and of size n_2 from population 2 are available, and let S_1^2 and S_2^2 be the sample variances. We wish to test the hypotheses

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2 \qquad (10\text{-}26)$$

The development of a test procedure for these hypotheses requires a new probability distribution, the F distribution. The random variable F is defined to be the ratio of two independent chi-square random variables, each divided by its number of degrees of freedom. That is,

$$F = \frac{W/u}{Y/v}$$

where W and Y are independent chi-square random variables with u and v degrees of freedom, respectively. We now formally state the sampling distribution of F.

Let W and Y be independent chi-square random variables with u and v degrees of freedom, respectively. Then the ratio

$$F = \frac{W/u}{Y/v} \qquad (10\text{-}27)$$

[redacted] is said to follow the F distribution with u degrees of freedom in the numerator and v degrees of freedom in the denominator. It is usually abbreviated as $F_{u,v}$.

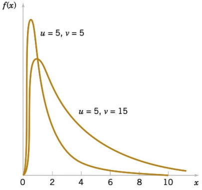

FIGURE 10-4 Probability density functions of two F distributions.

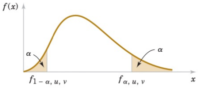

FIGURE 10-5 Upper and lower percentage points of the F distribution.

Table VI contains only upper-tailed percentage points (for selected values of $f_{\alpha,u,v}$ for $\alpha \le 0.25$) of the F distribution. The lower-tailed percentage points $f_{1-\alpha,u,v}$ can be found as follows.

Finding Lower Tail Points of the *F*-Distribution

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}} \qquad (10\text{-}30)$$

For example, to find the lower-tailed percentage point $f_{0.95,5,10}$, note that

$$f_{0.95,5,10} = \frac{1}{f_{0.05,10,5}} = \frac{1}{4.74} = 0.211$$

HYPOTHESIS TESTS ON THE RATIO OF TWO VARIANCES

A hypothesis-testing procedure for the equality of two variances is based on the following result.

Distribution of the Ratio of Sample Variances from Two Normal Distributions

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample from a normal population with mean μ_1 and variance σ_1^2, and let $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample from a second normal population with mean μ_2 and variance σ_2^2. Assume that both normal populations are independent. Let S_1^2 and S_2^2 be the sample variances. Then the ratio

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

has an F distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.

This result is based on the fact that $(n_1 - 1)\, S_1^2 / \sigma_1^2$ is a chi-square random variable with $n_1 - 1$ degrees of freedom, that $(n_2 - 1)\, S_2^2 / \sigma_2^2$ is a chi-square random variable with $n_2 - 1$ degrees of freedom, and that the two normal populations are independent. Clearly, under the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$, the ratio $F_0 = S_1^2 / S_2^2$ has an F_{n_1-1, n_2-1} distribution. This is the basis of the following test procedure.

Tests on the Ratio of Variances from Two Normal Distributions

Null hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

Test statistic: $$F_0 = \frac{S_1^2}{S_2^2} \quad (10\text{-}31)$$

Alternative Hypotheses	Rejection Criterion
$H_1: \sigma_1^2 \neq \sigma_2^2$	$f_0 > f_{\alpha/2, n_1-1, n_2-1}$ or $f_0 < f_{1-\alpha/2, n_1-1, n_2-1}$
$H_1: \sigma_1^2 > \sigma_2^2$	$f_0 > f_{\alpha, n_1-1, n_2-1}$
$H_1: \sigma_1^2 < \sigma_2^2$	$f_0 < f_{1-\alpha, n_1-1, n_2-1}$

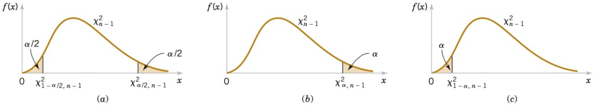

FIGURE 10-6 The F distribution for the test of $H_0: \sigma_1^2 = \sigma_2^2$ with critical region values for (a) $H_1: \sigma_1^2 \neq \sigma_2^2$. (b) $H_1: \sigma_1^2 > \sigma_2^2$. and (c) $H_1: \sigma_1^2 < \sigma_2^2$.

The critical regions for these fixed-significance-level tests are shown in Figure 10-6. Remember that this procedure is relatively sensitive to the normality assumption.

Example 10-13 **Semiconductor Etch Variability** Oxide layers on semiconductor wafers are etched in a mixture of gases to achieve the proper thickness. The variability in the thickness of these oxide layers is a critical characteristic of the wafer, and low variability is desirable for subsequent processing steps. Two different mixtures of gases are being studied to determine whether one is superior in reducing the variability of the oxide thickness. Sixteen wafers are etched in each gas. The sample standard deviations of oxide thickness are $s_1 = 1.96$ angstroms and $s_2 = 2.13$ angstroms, respectively. Is there any evidence to indicate that either gas is preferable? Use a fixed-level test with $\alpha = 0.05$.

The seven-step hypothesis-testing procedure may be applied to this problem as follows:

1. **Parameter of interest:** The parameters of interest are the variances of oxide thickness σ_1^2 and σ_2^2. We will assume that oxide thickness is a normal random variable for both gas mixtures.
2. **Null hypothesis:** H_0: $\sigma_1^2 = \sigma_2^2$
3. **Alternative hypothesis:** H_1: $\sigma_1^2 \neq \sigma_2^2$
4. **Test statistic:** The test statistic is given by Equation 10-31:

$$f_0 = \frac{s_1^2}{s_2^2}$$

5. **Reject H_0 if:** Because $n_1 = n_2 = 16$ and $\alpha = 0.05$, we will reject $H_0 : \sigma_1^2 = \sigma_2^2$ if $f_0 > f_{0.025,15,15} = 2.86$ or if $f_0 < f_{0.975,15,15} = 1/ f_{0.025,15,15} = 1/2.86 = 0.35$. Refer to Figure 10-6(a).
6. **Computations:** Because $s_1^2 = (1.96)^2 = 3.84$ and $s_2^2 = (2.13)^2 = 4.54$, the test statistic is

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{3.84}{4.54} = 0.85$$

7. **Conclusion:** Because $f_{0.975,15,15} = 0.35 < 0.85 < f_{0.025,15,15} = 2.86$, we cannot reject the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ at the 0.05 level of significance.

Practical Interpretation: There is no strong evidence to indicate that either gas results in a smaller variance of oxide thickness.

Inference on Two Population Proportions

LARGE-SAMPLE TESTS ON THE DIFFERENCE IN POPULATION PROPORTIONS

Suppose that two independent random samples of sizes n_1 and n_2 are taken from two populations, and let X_1 and X_2 represent the number of observations that belong to the class of interest in samples 1 and 2, respectively. Furthermore, suppose that the normal approximation to the binomial is applied to each population, so the estimators of the population proportions $P_1 = X_1/n_1$ and $P_2 = X_2/n_2$ have approximate normal distributions. We are interested in testing the hypotheses

$$H_0: p_1 = p_2 \qquad H_1: p_1 \neq p_2$$

The statistic

Test Statistic for the Difference of Two Population Proportions

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \qquad (10\text{-}34)$$

is distributed approximately as standard normal and is the basis of a test for $H_0: p_1 = p_2$. Specifically, if the null hypothesis $H_0: p_1 = p_2$ is true, by using the fact that $p_1 = p_2 = p$, the random variable

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

is distributed approximately $N(0, 1)$. A pooled estimator of the common parameter p is

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

The **test statistic** for $H_0: p_1 = p_2$ is then

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

This leads to the test procedures described as follows.

Approximate Tests on the Difference of Two Population Proportions

Null hypothesis: $H_0: p_1 = p_2$

Test statistic: $$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \qquad (10\text{-}35)$$

Alternative Hypothesis	*P*-Value	Rejection Criterion for Fixed-Level Tests
$H_1: p_1 \neq p_2$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$. $P = 2\left[1 - \Phi(\lvert z_0 \rvert)\right]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: p_1 > p_2$	Probability above z_0. $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: p_1 < p_2$	Probability below z_0. $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Example 10-16 **St. John's Wort** Extracts of St. John's Wort are widely used to treat depression. An article in the April 18, 2001, issue of the *Journal of the American Medical Association* ("Effectiveness of St. John's Wort on Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. John's Wort with a placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo. After eight weeks, 19 of the placebo-treated patients showed improvement, and 27 of those treated with St. John's Wort improved. Is there any reason to believe that St. John's Wort is effective in treating major depression? Use $\alpha = 0.05$.

The seven-step hypothesis testing procedure leads to the following results:

1. **Parameter of interest:** The parameters of interest are p_1 and p_2, the proportion of patients who improve following treatment with St. John's Wort (p_1) or the placebo (p_2).
2. **Null hypothesis:** H_0: $p_1 = p_2$
3. **Alternative hypothesis:** H_1: $p_1 \neq p_2$
4. **Test statistic:** The test statistic is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where $\hat{p}_1 = 27/100 = 0.27$, $\hat{p}_2 = 19/100 = 0.19$, $n_1 = n_2 = 100$, and

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{19+27}{100+100} = 0.23$$

5. **Reject H_0 if:** Reject H_0: $p_1 = p_2$ if the P-value is less than 0.05.
6. **Computation:** The value of the test statistic is

$$z_0 = \frac{0.27 - 0.19}{\sqrt{0.23(0.77)\left(\frac{1}{100}+\frac{1}{100}\right)}} = 1.34$$

7. **Conclusion:** Because $z_0 = 1.34$, the P-value is $P = 2\left[1 - \Phi(1.34)\right] = 0.18$, so, we cannot reject the null hypothesis.

Practical Interpretation: There is insufficient evidence to support the claim that St. John's Wort is effective in treating major depression.

Content is copied from book: Applied Statistics and Probability for Engineers