

Math 373 Lecture 24

Suppose you wish to measure the effect of smoking on heart disease. Don't just compare a sample of 40 smokers to a sample of 40 nonsmokers; rather select a sample of 40 pairs (A, B) such that A is a nonsmoker, B is a smoker but otherwise A and B are the same with respect to matters which affect heart disease, e.g., A and B have the same gender, same race, approximately the same height, weight and age. Such a paired difference experiment will more effectively separate the effects of smoking from the other factors which affect heart disease.

■ A long-term study compares the expected lifespan of nonsmokers vs. smokers with the intent of demonstrating that nonsmokers live longer. 25 pairs are selected. Each pair consists of a nonsmoker and a smoker who are otherwise matched w.r.t. race, gender, weight, exercise and dietary preferences. The average difference between the nonsmoker and smoker lifespans in the sample was $\bar{d} = \bar{x}_N - \bar{x}_S = 4$ years with std. dev. $s_d = 4$ years. Let $d = \mu_N - \mu_S$.

$H_0: \mu_N \leq \mu_S, \mu_N - \mu_S \leq 0, d \leq 0, H_a: \mu_N > \mu_S, \mu_N - \mu_S > 0, d > 0$.

$$df = 24, \quad SE = \frac{4}{\sqrt{25}} = \frac{4}{5} = .8, \quad t_\alpha = 1.711$$

Null region for d : $d \in (-\infty, 0]$.

The acceptance region the difference $\bar{d} = x_N - \bar{x}_S$:

$$\bar{d} \in (-\infty, 0 + 1.711 \times .8] = (-\infty, 1.37]$$

Does smoking significantly reduce lifespan? Yes.

Why?

$$\bar{d} = 4 \notin (-\infty, 1.37] = \text{acceptance for no signif. reduction.}$$

■ Find the minimum number of pairs of observations needed to estimate $\mu_1 - \mu_2$ with margin of error .5 years.

$$\text{Equation: } E = 1.96 \frac{s}{\sqrt{n}} \quad \text{Note: use 1.96, not a } t \text{ value.}$$

$$.5 = 1.96 \frac{4}{\sqrt{n}}, \quad \sqrt{n} = \frac{1.96 \times 4}{.5}, \quad n = 245.86 \rightarrow n = 246$$

Use the margin of error (which by definition is 1.96) not a t value.

Population variance and the χ^2 distribution

So far we have been mainly interested in a population's mean μ and have used the std. dev. s and variance s^2 mainly to determine the error in the estimate of μ . Sometimes though we are interested primarily in measuring the variance σ^2 . Annual income for example is correlated with IQ however the difference in income between males and females is due more to the difference in IQ variance (males have a larger variance; they dominate the extremes — both the homeless and the millionaires are mostly male) than to a difference between male and female IQs (there is none).

Suppose we select samples of size n from a normal population with mean μ and variance σ^2 . Different samples have different sample means \bar{x} and different sample variances s^2 . The sample means \bar{x} have a normal distribution with expected value μ and std. dev. SE.

What about the distribution of the sample variance s^2 ? The value of s^2 of the sample variance will vary around the population variance σ^2 . However the distribution is not normal, not even symmetric. Since the variance can never be negative, the distribution is skewed to the right.

We normalize a sample mean to a z-score by subtracting μ and dividing by SE. We “normalize” a sample variance to a *chi-square* variable by dividing by σ^2 and multiplying by $(n-1)$.

DEFINITION. $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ is the *chi-square* variable for the sample variance s^2 .

If $s = \sigma$, then $\chi^2 = \frac{(n-1)\sigma^2}{\sigma^2} = n-1$. This $n-1$ is the peak of the chi-square distribution; it is the “null value” for χ^2 . After normalizing the sample variance s^2 to χ^2 , we can use the chi-square probability distribution in Table 5 of Appendix I. As with the table for Student's t -distribution, Table 5 gives the values χ^2_α such that the tail to the right of χ^2_α has probability α . Since the distribution isn't symmetric, the table also includes the values needed to calculate left-hand tails. As with the t -distribution, the degrees of freedom for an n -element sample is $df = n-1$.

Since variances are nonnegative, the right-tailed* acceptance region is $[0, \chi^2_{1-\alpha}]$ rather than $(-\infty, \chi^2_{1-\alpha}]$. The left-tailed acceptance region is $[\chi^2_\alpha, \infty)$. The two-sided acceptance region is $[\chi^2_{1-\alpha/2}, \chi^2_{\alpha/2}]$. Acceptance regions always include the chi-square null value of $n-1$. The rejection regions are their complements. * right-tailed refers to the rejection region's tail.

■ A random sample of $n=25$ observations from a normal population produced a sample variance $s^2=10$. Does this provide sufficient evidence to indicate that $\sigma^2 < 15$?

State the null and alternate hypotheses regarding σ^2 .

$$H_a: \sigma^2 < 15, \chi^2 < 24. \quad H_0: \sigma^2 \geq 15, \chi^2 \geq 24. \quad df = 24.$$

Null region for σ^2 : $[15, \infty)$. Null region for χ^2 : $[24, \infty)$.

Recall that $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ and its null value is $df = n-1$.

$$\text{The } \chi^2 \text{ value of } s^2=10 \text{ is } \chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(24)10}{15} = 16.$$

Acceptance region for χ^2 : $\chi^2 \in [\chi^2_{1-\alpha}, \infty) = [13.85, \infty)$.

Is σ^2 significantly smaller than 15?

No. $\chi^2 = 16 \in [13.85, \infty) = \text{acceptance region for } H_0$.

The acceptance region for s^2 is “centered” around the null value $\sigma^2=15$.

$$13.85 \leq \chi^2 \Rightarrow 13.85 \leq \frac{(n-1)s^2}{\sigma^2} \Rightarrow 13.85 \leq \frac{24s^2}{15} \\ \Rightarrow \frac{13.85 \times 15}{24} \leq s^2 \Rightarrow s^2 \in [8.66, \infty)$$

The confidence region for σ^2 is “centered” around the measure value $s^2=10$.

$$13.85 \leq \chi^2 \Rightarrow 13.85 \leq \frac{(n-1)s^2}{\sigma^2} \Rightarrow 13.85 \leq \frac{24 \times 10}{\sigma^2} \\ \frac{1}{13.85} \geq \frac{\sigma^2}{24 \times 10} \Rightarrow \frac{24 \times 10}{13.85} \geq \sigma^2 \Rightarrow \sigma^2 \in [0, 17.33].$$

For the hypotheses $H_a: \sigma^2 \neq 15$ and $H_0: \sigma^2 = 15$, the acceptance region for χ^2 is $[12.40, 39.36]$.