1 Random sampling

Definition 1 *A population consists of the totality of the observations with which we are concerned*

Definition 2 *A sample is a subset of a population.*

In the field of statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. Therefore, we must depend on a subset of observations from the population to help us make **inferences** concerning that same population.

Definition 3 *A sample is a subset of a population.*

To eliminate any possibility of **bias** in the sampling procedure, it is desirable to choose a random sample in the sense that the observations are made independently and at random.

2 Some important statistics

Definition 4 *Any function of the random variables constituting a random sample is called a statistic.*

- Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Sample median: $\widetilde{X} = \begin{cases} x(n+1)/2, \text{ if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), \text{ if } n \text{ is even.} \end{cases}$

- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

The computed value of S^2 for a given sample is denoted by s^2.

Theorem 5 *If S^2 is the variance of a random sample of size n, we may write*

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$

- Sample standard deviation: $S = \sqrt{S^2}$

3 Sampling Distibutions

Let us consider a soft-drink machine designed to dispense, on average, 240 milliliters per drink. A company official who computes the mean of 40 drinks obtains $\overline{x} = 236$ milliliters. On the basis of this value, she decides that the machine is still dispensing drinks with an average content of $\mu = 240$ milliliters. The 40 drinks represent a sample from the infinite population of possible drinks that will be dispensed by this machine.The company official made the decision that the soft-drink machine dispenses drinks with an average content of 240 milliliters, even though the sample mean was 236 milliliters, because he knows from sampling theory that, if $\mu = 240$ milliliters, such a sample value could easily occur. In fact, if she ran similar tests, say every hour, she would expect the values of the statistic $\overline{x}$ to fluctuate above and below $\mu = 240$ milliliters. Only when the value of $\overline{x}$ is **substantially** different from 240 milliliters will the company official initiate action to adjust the machine.

Since a statistic is a random variable that depends only on the observed sample, it must have a probability distribution.

Definition 6 *The probability distribution of a statistic is called a sampling distribution.*

4 Sampling Distribution of Means and the Central Limit

Theorem 7 *If $X_1, X_2, ..., X_n$ are independent (?) random variables having normal distributions with means $\mu_1, \mu_2, ..., \mu_n$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$, respectively, then the random variable $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ has a normal distribution with mean*

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

and variance

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

Suppose that a random sample of n observations is taken from a normal population with mean μ and variance σ^2. Each observation X_i, $i = 1, 2, ..., n$, of the random sample will then have the same normal distribution. Hence, from Theorem 7, we conclude that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

has a normal distribution with mean

$$\mu_{\overline{X}} = \frac{1}{n}\{\mu + \mu + ... + \mu\} = \sum_{i=1}^{n} \mu = \mu$$

and variance

$$\sigma_{\overline{X}}^2 = \frac{1}{n^2}\left\{\sigma^2 + \sigma^2 + ... + \sigma^2\right\} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

Theorem 8 (Central Limit Theorem) *If $\overline{X}$ is the mean of a random sample of size n taken from a population*

with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$\overline{Z} = \frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $N(0, 1)$.

The normal approximation for $\overline{X}$ will generally be good if $n \geq 30$.

Example 9 *An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to* 800 *hours and a standard deviation of* 40 *hours. Find the probability that a random sample of* 16 *bulbs will have an average life of less than* 775 *hours.*

Solution 10 *The z values corresponding to $x_1 = 778$ and $x_2 = 834$ are*

$$z_1 = \frac{778 - 800}{40} = -0.55 \text{ and } z_2 = \frac{834 - 800}{40} = 0.85.$$

Hence,

$$\begin{aligned}\Pr(778 &< X < 834) = P(-0.55 < Z < 0.85)\\ &= P(Z < 0.85) - P(Z < -0.55)\\ &= 0.8023 - 0.2912 = 0.5111.\end{aligned}$$

Example 11 *Traveling between two campuses of a university in a city via shuttle bus takes, on average,* 28 *minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers* 40 *times. What is the probability that the average transport time was more than* 30 *minutes?*

Solution 12 *In this case,* $\mu = 28$ *and* $\sigma = 3$*. We need to calculate the probability* $Pr(\overline{X} > 30)$ *with* $n = 40$*. Since the time is measured on a continuous scale to the nearest minute, an* $\overline{x}$ *greater than* 30 *is equivalent to* $\overline{x} \geq 30.5$*. Hence,*

$$\begin{aligned}Pr(\overline{X} &> 30) = \Pr\left(\frac{\overline{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right)\\ &= \Pr\left(Z \geq 3.16\right) = 0.0008.\end{aligned}$$

There is only a slight chance that the average time of one bus trip will exceed 30 minutes. An

5 Sampling Distribution of the Difference between Two Means

A scientist or engineer may be interested in a comparative experiment in which two manufacturing methods, 1 and 2, are to be compared. The basis for that comparison is $\mu_1 - \mu_2$, the difference in the population means. Suppose that we have two populations, the first with mean μ_1 and variance σ_1^2, and the second with mean μ_2 and variance σ_2^2. Let the statistic $\overline{X}_1$ represent the mean of a random sample of size n_1 selected from the first population, and the statistic $\overline{X}_2$ represent the mean of a random sample of size n_2 selected from the second population, independent of the sample from the first population. What can we say about the sampling distribution of the difference $\overline{X}_1 - \overline{X}_2$ for repeated samples of size n_1 and n_2? According to Theorem 8, the variables $\overline{X}_1$ and $\overline{X}_2$ are both approximately normally distributed with means μ_1 and μ_2 and variances σ_1^2/n_1 and σ_2^2/n_2, respectively. This

approximation improves as n_1 and n_2 increase. We can conclude that $\overline{X}_1 - \overline{X}_2$ is approximately normally distributed with mean

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_{\overline{X}_1} - \mu_{\overline{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

The Central Limit Theorem can be easily extended to the two-sample, two-population case.

Theorem 13 *If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is approximately normally distributed with mean and variance given by*

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is approximately a standard normal variable.

If both n_1 and n_2 are greater than or equal to 30, the normal approximation for the distribution of $\overline{X}_1 - \overline{X}_2$ is good. Two independent experiments are run in which two different types of paint are compared.

Example 14 *Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be* 1.0*. Assuming that the mean drying time is equal for the two types of paint, find* $P(\overline{X}_A - \overline{X}_B > 1.0)$*, where* $\overline{X}_A$ *and* $\overline{X}_B$ *are average drying times for samples of size* $n_A = n_B = 18$*.*

Solution 15 *From the sampling distribution of* $\overline{X}_A - \overline{X}_B$*, we know that the distribution is approximately normal with mean* $\mu_{\overline{X}_A - \overline{X}_B} = \mu_A - \mu_B = 0$ *and variance* $\sigma^2_{\overline{X}_A - \overline{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = 1/9$*. Corresponding to the value* $\overline{X}_A - \overline{X}_B = 1.0$*, we have*

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3$$

so

$$\Pr(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

Example 16 *The television picture tubes of manufacturer* A *have a mean lifetime of* 6.5 *years and a standard deviation of* 0.9 *year, while those of manufacturer* B *have a mean lifetime of* 6.0 *years and a standard deviation of* 0.8 *year. What is the probability that a random sample of* 36 *tubes from manufacturer* A *will have a mean lifetime that is at least* 1 *year more than the mean lifetime of a sample of* 49 *tubes from manufacturer* B*?*

Solution 17 *We are given the following information:*

Population 1	*Population 2*
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

If we use, the sampling distribution of $\overline{X}_1 - \overline{X}_2$ *will be approximately normal and will have a mean and standard deviation*

$\mu_{\overline{X}_1 - \overline{X}_2} = 6.5 - 6.0$ *and* $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189$
Hence,

$$\begin{aligned} \Pr(\overline{X}_1 - \overline{X}_2 &\geq 1.0) = P(Z > 2.65) = 1 - P(Z < 2.65) \\ &= 1 - 0.9960 = 0.0040. \end{aligned}$$

6 Sampling Distribution of S^2

Theorem 18 *If* $X_1, X_2, ..., X_n$ *an independent random sample that have the same standard normal distribution*

then $X = \sum_{i=1}^{n} X_i^2$ *is chi-squared distribution, with* $\nu = n$ *degrees of freedom.*

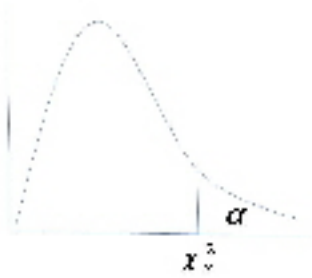
Theorem 19 *The mean and variance of the chi-squared distribution* χ^2 *with* ν *degrees of freedom are* $\mu = \nu$ *and* $\sigma^2 = 2\nu$.

Theorem 20 *If* S^2 *is the variance of a random sample of size* n *taken from a normal population having the variance* σ^2, *then the statistic*

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

has a chi-squared distribution with $\nu = n - 1$ *degrees of freedom.*

Table A.5 gives values of χ^2_α for various values of α and ν. Hence, the χ^2 value with 7 degrees of freedom, leaving an area of 0.05 to the right, is $\chi^2_{0.05} = 14.067$. Owing to lack of symmetry, we must also use the tables to find $\chi^2_{0.95} = 2.167$ for $\nu = 7$.



Example 21 *For a chi-squared distribution, find*
(a)$\chi^2_{0.025}$ *when* $\nu = 15$*;*
(b)$\chi^2_{0.01}$*when* $\nu = 7$*;*
(c)$\chi^2_{0.05}$ *when* $\nu = 24$*.*

Solution 22 *(a) 27.488.(b) 18.475.(c) 36.415*

For a chi-squared distribution X, find χ^2_α such that
(a) $P(X > \chi^2_\alpha) = 0.99$ when $\nu = 4$;
(b) $P(X > \chi^2_\alpha) = 0.025$ when $\nu = 19$;
(c) $P(37.652 < X < \chi^2_\alpha) = 0.045$ when $\nu = 25$.

Solution 23 *(a)* $\chi^2_\alpha = \chi^2_{0.99} = 0.297$*.(b)* $\chi^2_\alpha = \chi^2_{0.025} = 32.852$*.(c)* $\chi^2_{0.05} = 37.652$*. Therefore,* $\alpha = 0.05 - 0.045 = 0.005$*. Hence,* $\chi^2_\alpha = \chi^2_{0.005} = 46.928$*.*

7 t-Distribution

Theorem 24 *Let Z be a standard normal random variable and ν a chi-squared random variable with ν degrees of freedom. If Z and ν are independent, then the distribution of the random variable T, where*

$$T = \frac{Z}{\sqrt{\nu/\nu}}$$

This is known as the t-distribution with ν degrees of freedom.

Corollary 25 *Let $X_1, X_2, ..., X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Then the random variable $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $\nu = n - 1$ degrees of freedom.

Example 26 *The t-value with* $\nu = 14$ *degrees of freedom that leaves an area of* 0.025 *to the left, and therefore an area of* 0.975 *to the right, is*

$$t_{0.975} = -t_{0.025} = -2.145$$

Example 27 *Find* $\Pr(-t_{0.025} < T < t_{0.05})$.

Solution 28 *Since* $t_{0.05}$ *leaves an area of* 0.05 *to the right, and* $-t_{0.025}$ *leaves an area of* 0.025 *to the left, we find a total area of* $1 - 0.05 - 0.025 = 0.925$ *between* $-t_{0.025}$ *and* $t_{0.05}$. *Hence* $\Pr(-t_{0.025} < T < t_{0.05}) = 0.925$.

Example 29 *Find* k *such that* $\Pr(k < T < -1.761) = 0.045$ *for a random sample of size* 15 *selected from a normal distribution with* $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$.

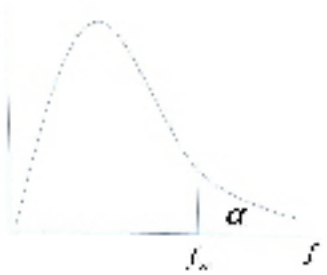
Solution 30 *From Table A.4 we note that* 1.761 *corresponds to* $t_{0.05}$ *when* $\nu = 14$. *Therefore,*$-t_{0.05} = -1.761$. *Since* k *in the original probability statement is to the left of* $-t_{0.05} = -1.761$, *let* $k = -t_\alpha$. *Then, by using figure, we have*

$$0.045 = 0.05 - \alpha, \text{ or } \alpha = 0.005.$$

Hence, from Table A.4 with $\nu = 14$,
$k = -t_{0.005} = -2.977$ *and* $\Pr(-2.977 < T < -1.761) = 0.045$.

8 F-Distribution

The statistic F is defined to be the ratio of two independent chi-squared random variables, each divided by its number of degrees of freedom.



Theorem 31 *The random variable*

$$F = \frac{U/\nu_1}{V/\nu_2}$$

*where U and V are independent random variables having chi-squared distributions with ν_1 and ν_2 degrees of freedom, respectively,is the F-***distribution** *with ν_1 and ν_2 degrees of freedom (d.f.).*

Writing $f_\alpha(\nu_1, \nu_2)$ for f_α with ν_1 and ν_2 degrees of freedom, we obtain

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}$$

Thus, the f-value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right, is $f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246$.

8.1 The F-Distribution with Two Sample Variances

Suppose that random samples of size n_1 and n_2 are selected from two normal populations with variances σ_1^2 and σ_2^2, respectively. From Theorem 16, we know that

$$\chi_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ and } \chi_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

are random variables having chi-squared distributions with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom. Furthermore, since the samples are selected at random, we are dealing with independent random variables. Then, using Theorem 24 with $\chi_1^2 = U$ and $\chi_2^2 = V$, we obtain the following result.

Theorem 32 *If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then*

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

has an F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

8.2 Example

For an F-distribution, find
(a) $f_{0.05}$ with $\nu_1 = 7$ and $\nu_2 = 15$;
(b) $f_{0.05}$ with $\nu_1 = 15$ and $\nu_2 = 7$:
(c) $f_{0.01}$ with $\nu_1 = 24$ and $\nu_2 = 19$;
(d) $f_{0.95}$ with $\nu_1 = 19$ and $\nu_2 = 24$;
(e) $f_{0.99}$ with $\nu_1 = 28$ and $\nu_2 = 12$.

Solution 33 *(a) 2.71.(b) 3.51.(c) 2.92.(d) $1/2.11 = 0.47$.(e) $1/2.90 = 0.34$.*