Sampling Distributions

Q: A company manufactures 100 Ohms resistors. A sample of 40 resistors from the assembly line is found to have a mean of 105 Ohms. How likely is the population mean (the mean of the probability density function) to be 100 Ohms?

A: In questions like this we need to make inferences about the population mean based on the sample mean. To do this, we need to know the probability distribution of the sample mean!!!

<u>Defn</u>: The probability distribution of a statistic is called a sampling distribution.

Sampling Distribution of Means

Given a sample with n observations: $X_1, X_2, \ldots, X_n$

Sample mean $\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$

$\bar{X}$ itself is a random variable, in fact, it is a linear combination of random variables $X_1 \ldots X_n$.

Now assume the observations were taken from a population with mean μ and standard deviation σ.

What is the mean of $\bar{X}$?

$$\mu_{\bar{X}} = \frac{1}{n}\underbrace{\left(\mu + \mu + \cdots + \mu\right)}_{n \text{ terms}} = \mu$$

What is the variance of $\bar{X}$?

$$\sigma^2_{\bar{X}} = \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right) = \frac{\sigma^2}{n}$$

Note: observations are independent of each other which allows us to use the simplified formulas from Chapter 4.

What do these results mean?

(1) If I take many samples from the population, each with n observations, the mean of the sample means will equal the population mean.

(2) If I take many samples from the population, each with n observations, the standard deviation of the sample means will equal $\sqrt{\frac{\sigma^2}{n}}$ (equivalently the variance of the sample means will be $\frac{\sigma^2}{n}$) where σ is the population ~~mean~~ standard deviation.

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and variance σ^2, then the limiting form of the probability distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

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

— o —

Notice that regardless of the population distribution $f(x)$, the central limit theorem states that the distribution of the sample mean is a normal distribution!!!

The normal distribution result of the central limit theorem is good if $n \geq 30$. For $n < 30$ it is only good if the population distribution is not too different from a normal distribution

Example: Manufacture resistors. Population mean 100 Ohms population standard deviation 20 Ohms. Find the probability that a random sample of 50 resistors will have a mean resistance of 101 Ohms or larger.

Solution: $\mu = 100 \quad \sigma = 20$

$$\mu_{\bar{X}} = 100 \qquad \sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{50}}$$

$$= 2.83$$

Since $n = 50 > 30$ we can use the Central limit theorem (even though we don't know the population distribution) to say that the sample mean $\bar{X}$ has a normal distribution with $\mu_{\bar{X}} = 100$ and $\sigma_{\bar{X}} = 2.83$

Then $P(\bar{X} > 101) = P\left(Z > \frac{101-100}{2.83}\right)$

$$= P(Z > 0.35)$$

$$= 1 - P(Z < 0.35)$$

$$= 1 - 0.6368 \qquad \text{From table A.3}$$

$$= 0.3632$$

Example: Manufacture light bulbs. Length of life approximately normally distributed with mean 800 hours and a standard deviation 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Solution: $\mu = 800 \quad \sigma = 40$, $\mu_{\bar{X}} = 800 \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$

Even though $n = 16 < 30$, the Central limit theorem can be used because it is stated that the population distribution is approximately normal.

$$P(\bar{X} < 775) = P\left(Z < \frac{775-800}{10}\right) = P(Z < -2.5)$$

$$= 0.0062 \quad \text{from Table A.3.}$$

Inference on the population mean: Given the probability we calculated in the previous example, we could ask the question how likely is the population mean to be really 800 hours? Not very likely since the prob value was so small. We will learn to answer this question in a formal manner when we discuss hypothesis testing.

Sampling Distribution of Difference of Two Means

Sometimes we are interested in comparing two populations, i.e is one manufacturing process better than the other (according to longer life expectancy or similar criterion)

Population 1	Population 2
μ_1, σ_1	μ_2, σ_2
$\downarrow$	$\downarrow$
Sample 1 with n_1 observations	Sample 2 with n_2 observations
$\bar{X}_1$: $\mu_{\bar{X}_1} = \mu_1$	$\bar{X}_2$: $\mu_{\bar{X}_2} = \mu_2$
$\sigma_{\bar{X}_1} = \sigma_1/\sqrt{n_1}$	$\sigma_{\bar{X}_2} = \sigma_2/\sqrt{n_2}$

What can we say about the sampling distribution of $\bar{X}_1 - \bar{X}_2$?

From what we learned in Chapter 4: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2}$

$$= \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Generalization of central limit theorem

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

standard normal variable. Good when both n_1 and $n_2 \geq 30$ or when the population dists. are approximately normal.

Example (8.9 in textbook) lifetime of product

Population 1: $\mu_1 = 6.5$ $\sigma_1 = 0.9$

Population 2: $\mu_2 = 6.0$ $\sigma_2 = 0.8$

Sample with $n_1 = 36$ observations from Population 1

" " $n_2 = 49$ observations " " 2.

What is the probability that a random sample of 36 TVs from population 1 will have a mean life that is at least 1 year longer than the sample of 49 TVs from population 2?

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = ?$$

Soln: Since both n_1 and $n_2 \geq 30$ the sampling distribution of $\bar{X}_1 - \bar{X}_2$ will be approximately normal with

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2 = 6.5 - 6.0 = 0.5$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2}$$

$$= \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}}$$

$$= 0.189$$

Therefore

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = P\left(Z \geq \frac{1.0 - 0.5}{0.189}\right)$$

$$= 1 - P(Z < 2.65)$$

Table A.3

$$= 1 - 0.996 = 0.004$$

Example (similar to example 8.8 textbook)

An electrical company is evaluating two different production methods for long lasting light bulbs. Call these methods A and B. The population distribution and the means for A and B are unknown, but we know that the standard deviation for both is 50 hours. Assuming that the mean life-time for both methods is the same, find $P(\bar{X}_A - \bar{X}_B \geq 15)$ when a random sample of $n_A = 100$ is taken from population A and a random sample of $n_B = 100$ is taken " " B.

Soln: Since both n_A and $n_B \geq$ ~~100~~ 30, the sampling distribution for $\bar{X}_A - \bar{X}_B$ will be approximately normal with

$$\mu_{\bar{X}_A} - \mu_{\bar{X}_B} = \mu_A - \mu_B = 0 \quad [\mu_A, \mu_B \text{ unknown but assumed to be equal}]$$

$$\sigma_{\bar{X}_A - \bar{X}_B} = \sqrt{\frac{50^2}{100} + \frac{50^2}{100}} = \sqrt{50} = 7.07$$

Therefore $P(\bar{X}_1 - \bar{X}_2 \geq 15) = P\left(Z \geq \frac{15-0}{7.07}\right)$

$$= 1 - P(Z < 2.12) = 1 - 0.983 \text{ (Table A.3)}$$

$$= 0.017.$$

This low probability suggests that the mean lifetime of the two populations likely are not the same based on these observations. Most likely A has a longer lifetime. Later we will use hypothesis testing to determine this.

Exercise 8.18 textbook

$$f(x) = \begin{cases} 1/3, & x = 2, 4, 6 \\ 0, & \text{elsewhere} \end{cases}$$

Random sample size $n = 54$

$$P(4.15 < \bar{X} < 4.35) = ?$$

First find population mean and standard deviation.

$$\mu = \sum_x x f(x) = 2 \times \frac{1}{3} + 4 \times \frac{1}{3} + 6 \times \frac{1}{3} = 4$$

$$\sigma^2 = E[X^2] - \mu^2 = \left(\sum_x x^2 f(x)\right) - 16$$

$$= 4 \times \frac{1}{3} + 16 \times \frac{1}{3} + 36 \times \frac{1}{3} - 16$$

$$= \frac{56}{3} - 16 = \frac{56 - 48}{3} = 8/3$$

Now the sample mean and variance

$$\mu_{\bar{X}} = \mu = 4 \qquad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{8/3}{54} = \frac{4}{81}$$

$$\sigma_{\bar{X}}^2 = 2/9$$

Finally $P(4.15 < \bar{X} < 4.35) = P\left(\frac{4.15 - 4}{2/9} < Z < \frac{4.35 - 4}{2/9}\right)$

$$= P(0.68 < Z < 1.58) = 0.9429 - 0.7517 \quad [\text{Table A3}$$

$$= 0.1912$$

Sampling Distribution of S^2

Remember S^2 is sample variance. This completely different than $\sigma_{\bar{X}}^2$ so don't get confused.

Given a sample: 5, 11, 9, 5, 10, 15, 6, 10, 5, 10

$$\bar{X} = \frac{1}{10}(5+11+9+5+10+15+6+10+5+10) = 8.4$$

and $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{9}\left[(5-8.6)^2 + (11-8.6)^2 + \cdots + (10-8.6)^2\right]$$

$$= 10.933$$

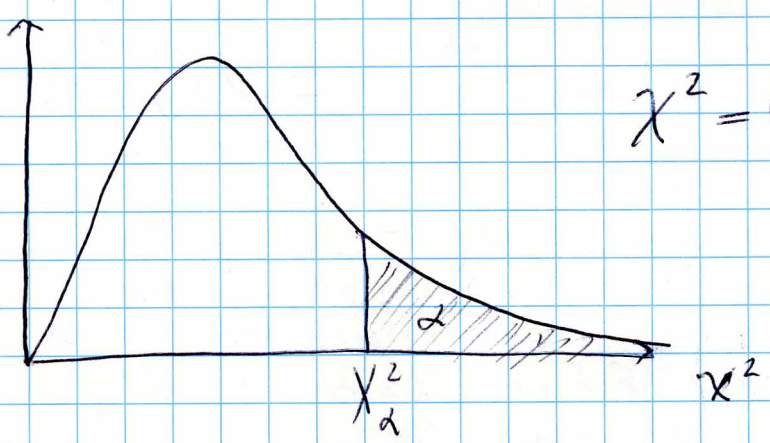
Now we will talk about the probability distribution for S^2. Before we were talking about " " " for $\bar{X}$.

Sampling distribution of S^2 is used in studying variability.

---

* A random variable of the form $Y = \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$ has a χ^2 (chi-squared) distribution with n degrees of freedom if the X_i are normally distributed.

* The mathematical expression for the χ^2 distribution is complicated (Section 6.8 textbook), the shape looks like



$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

χ^2_α represents the χ^2 value above which we find an area of α. (Table A.5)

Sample $= X_1, \ldots, X_n$ population mean μ, standard dev σ

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} \left((X_i - \bar{X}) + (\bar{X} - \mu) \right)^2$$

$$= \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (\bar{X} - \mu)^2 + \sum_{i=1}^{n} 2(\bar{X} - \mu)(X_i - \bar{X})$$

$$= \underbrace{\sum_{i=1}^{n} (X_i - \bar{X})^2}_{\text{this is } (n-1)s^2} + n(\bar{X} - \mu)^2 + \cancel{2(\bar{X} - \mu) \underbrace{\sum_{i=1}^{n} (X_i - \bar{X})}_{\emptyset \text{ since } \sum_{i=1}^{n} X_i = \bar{X} n}}$$

Divide both sides by σ^2

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 = \frac{(n-1)s^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2 / n}$$

$$\underbrace{\frac{(n-1)s^2}{\sigma^2}}_{\substack{\chi^2 \text{ with} \\ n-1 \text{ degrees} \\ \text{of freedom.}}} = \underbrace{\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2}_{\substack{\chi^2 \text{ with } n \\ \text{degrees of} \\ \text{freedom}}} - \underbrace{\frac{1}{\sigma^2 / n} (\bar{X} - \mu)^2}_{\substack{\chi^2 \text{ with} \\ 1 \text{ degrees of} \\ \text{freedom}}}$$

<u>Theorem</u>: 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

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

has a χ^2 distribution with $v = n - 1$ degrees of freedom.

Example: (Example 8.10 textbook)

Car batteries supposed to last 3 years on average with standard deviation of 1 year. If a sample of 5 batteries are found to have life-times: 1.9, 2.4, 3.0, 3.5 and 4.2 years, is the claim that the population standard deviation is 1 year valid? (Assume the battery lifetimes are normally distributed)

Solution

firs find S^2 value.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{where} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{X} = \frac{1}{5}(1.9 + 2.4 + 3.0 + 3.5 + 4.2) = 3.0$$

$$S^2 = \frac{1}{4}\left[(1.9-3)^2 + \cdots + (4.2-3)^2\right] = 0.815$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{4 \times 0.815}{1} = 3.26$$

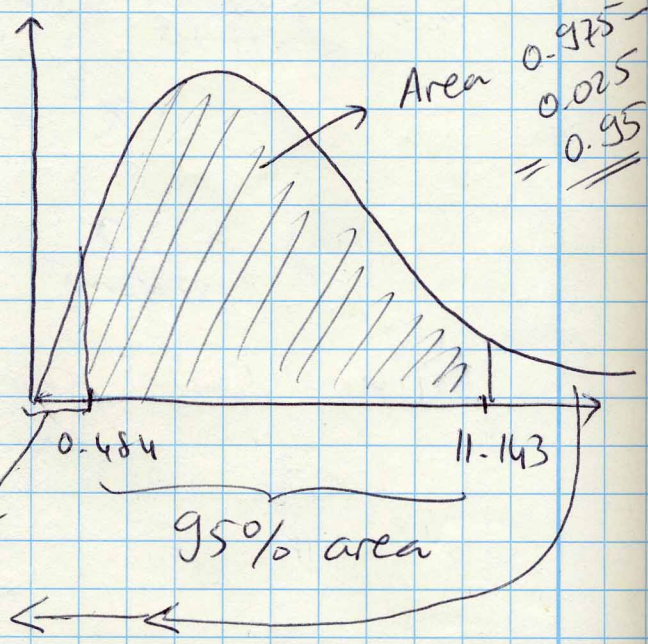
Since $n = 5$, χ^2 has $v = n-1 = 4$ degrees of freedom.

From Table A.5 row $v = 4$ we see that

$$\chi^2_{0.025} = 11.143 \quad \text{and} \quad \chi^2_{0.975} = 0.484$$

This means that 0.95 area under the χ^2 curve falls between 0.484 and 11.143.

Since our χ^2 value for this sample $\chi^2 = 3.26$ falls within this range it is reasonable.



If the computed χ^2 fell here it would be a sign that $\sigma = 1$ is likely wrong.

t - Distribution

We saw how to use the normal distribution to compute probabilities about the sample mean when we know (or can compute) the population mean and variance. In some experiments we might know the population mean but not the variance. In these cases the sample variance can be substituted for the population variance and the resulting sampling distribution is called the t-distribution.

Remember $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

σ known

Normal distribution $n \geq 30$

or

any n but population distribution known to be close to normal

$T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$

σ unknown

T-distribution.

For $n \geq 30$ T-distribution close to normal distribution

However, if n small S^2 can fluctuate significantly from sample to sample

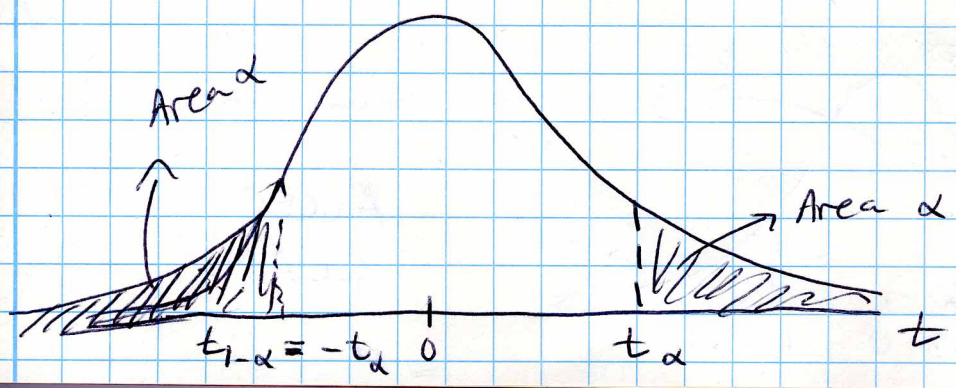
---

Let $X_1, X_2, \ldots, X_n$ be independent random variables that are normally distributed with mean μ. Let

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

then $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ has a t-distribution with

$\nu = n - 1$ degrees of freedom.

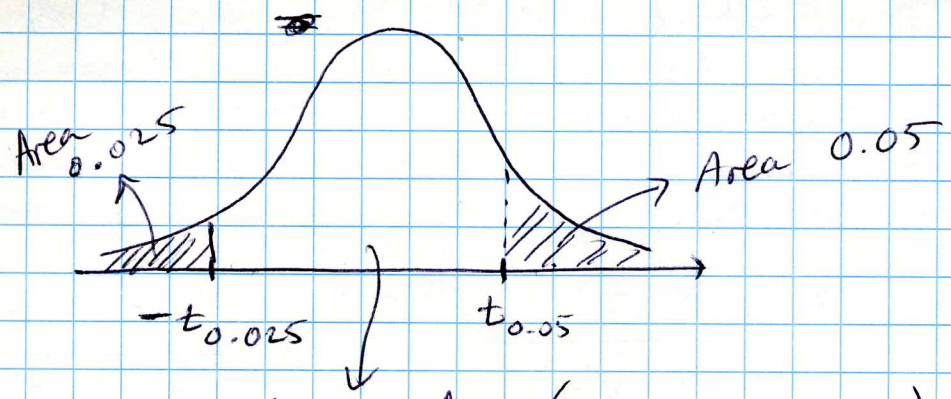


$t_{0.95} = -t_{0.05}$

$t_{0.99} = -t_{0.01}$

$t_{1-\alpha} = -t_\alpha$

Example : $P(-t_{0.025} < T < t_{0.05})$



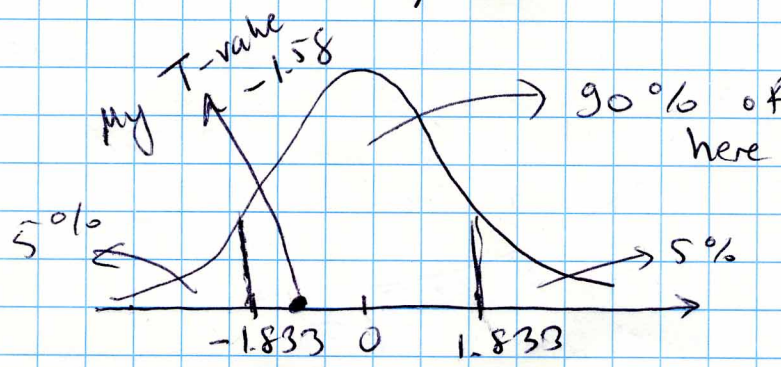
$$\text{Area} = 1 - (0.05 + 0.025) = 0.925$$

Note: t-distribution symmetric about mean of 0. In other words, the t-value leaving an area of $1-\alpha$ to the right and therefore an area of α to the left is equal to the negative t-value that leaves an area α to the right.

Example : An ISP claims that the mean connection speed provided to my house is 5 Mbps. To check this claim I measure the connection speed at 10 different occasions and find a mean connection speed of 4.5 Mbps and a sample standard deviation of 1.0 Mbps. I will be satisfied if I can show that the ISP's claim is true if the t-value for my sample falls between $-t_{0.05}$ and $t_{0.05}$. Should I accept their claim?

$$T = \frac{4.5 - 5}{1.0/\sqrt{10}} = -1.58$$

For $v = n - 1 = 9$
$t_{0.05} = 1.833$
so
$-t_{0.05} = -1.833$



ACCEPT THE CLAIM.

Example: ISP claims mean connection speed of 5 MBPS. On 8 occasions I measure: 4.4, 5.5, 4.1, 5.1, 4.2, 2.6, 4.1, 3.8

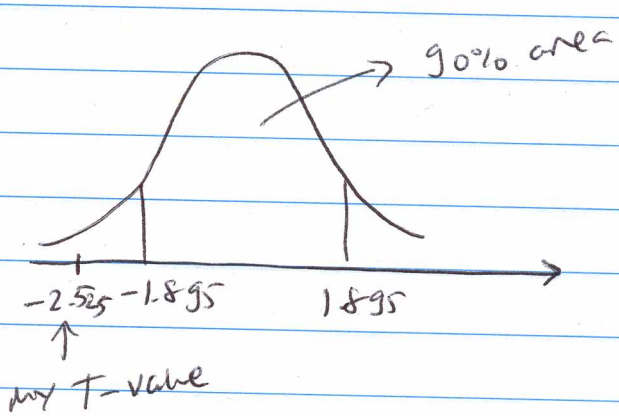
Does the t-value fall between $-t_{0.05}$ and $t_{0.05}$?

$$\bar{X} = 4.225 \qquad S = 0.8681$$

$$T = \frac{4.225 - 5}{0.8681/\sqrt{8}} = -2.525$$

Since $n = 8$, $V = 7$ so we use that row in Table A.4.

$$t_{0.05} = 1.895 \text{ so } -t_{0.05} = -1.895$$



Falls outside acceptable range. In fact, it even falls outside the range $-t_{0.025}$ to $t_{0.025}$ (95% area)

Note: in these two examples we assumed that the connection speeds were normally distributed which is necessary for using the t-distribution.