B34.UC2

Numerical Computation and Statistics in Engineering

Unit 3: Sampling

Introduction

In this unit we will mainly talk about *infinite* populations. However, most of the techniques and results will hold for (large) finite populations.

We are interested in taking random samples of a population to determine properties like the (unknown) mean or (unknown) standard deviation of the population. The process of taking random samples from a population is known as *sampling*.

In general, we want to have the following three properties of an estimator (say for the population mean) when sampling:

- unbiasedness,
- consistency, and
- efficiency.

Introduction

Unbiasedness refers to the fact that the average over *all* possible sample means (of a given size n) is equal to the population mean.

An estimator is *consistent* if, as the sample size increases, the difference between estimate and true population value (here mean) approaches zero. (For example, the formula for the standard deviation with $n-1$ in the denominator is unbiased and consistent, the one with n in the denominator is consistent, but biased.)

Efficiency, the last desirable property of an estimator, refers to the precision of the sample.

Basic Definitions

If $x_1, \ldots, x_n$ are *independent* and *identically distributed* random variables then they form a *random sample* from the population.

Example. A machine manufactures conductors. Each week *one* sample of fifty conductors is taken and the capacity is measured. If x_i is the average of the measures in week i, then the x_i form a random sample from the population. □

If $x_1, \ldots, x_n$ are a random sample then

$$\bar{x} = \frac{1}{n} \sum_i x_i \qquad \text{and} \qquad s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

are called the *sample mean* and *sample variance*. As before, the *sample standard deviation* is the square root of the sample variance.

The *standard error* of a statistics is the standard deviation of its sampling distribution.

Distribution of the Mean

If $x_1, \ldots, x_n$ are a random sample from an infinite population with mean μ and standard deviation σ, then

$$E(\bar{x}) = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \, .$$

This result says that if we sample then the expected value of these samples is the actual mean of the population, i.e., the correct value, and, the standard deviation of the sample decreases if the sample size increases: The more samples we take the more we can be assured that $\bar{x}$ is close to μ. Thus, the estimator $\bar{x}$ for the population mean is unbiased and consistent.

Together with Chebycheff's Theorem (see Unit 1) the previous result can be rephrased as follows:

> For every positive c, the probability that $\bar{x}$ will take a value in the interval $[\mu - c, \mu + c]$ is at least $1 - \frac{\sigma^2}{nc^2}$.

The Central Limit Theorem

Of more importance (both theoretically and practically) is the following version of the *Central Limit Theorem*:

> If n is sufficiently large (in practice $n \geq 30$) the random variable $\bar{x}$ can be approximated with a *normal* probability distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ (regardless of the actual shape of the sampled population (!)).

If the distribution of the population is symmetric then taking samples of size $n \geq 25$ is enough. If the distribution of the population is *normal* then $\bar{x}$ has normal distribution too, regardless of the size of n.

The Central Limit Theorem

Example. A coffee vending machine fills cups with coffee with mean 150 milliliter and standard deviation 15 milliliter. What is the probability that the average amount of coffee in a random sample of size 40 is at least 155 milliliters?

The distribution of $\bar{x}$ (the average amount in the sample of 40 cups of coffee) has sample mean $\mu_{\bar{x}} = 150$, and standard deviation $\sigma_{\bar{x}} = \frac{15}{\sqrt{40}}$, and this distribution is approximately normal. Using the standardized normal distribution $z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$ the corresponding z-value is $\frac{155-150}{\frac{15}{\sqrt{40}}} = \frac{\sqrt{40}}{3} \approx 2.10817$. Thus, tables show that

$$P(\bar{x} \geq 155) = P(z \geq 2.108) \approx 0.0175\,.$$

The table gives $P(0 \leq z \leq 2.10) = 0.4821$, and $P(0 \leq z \leq 2.11) = 0.4826$. With a linear approximation we find that $P(0 \leq z \leq 2.10817) \approx 0.4821 + 0.81(0.4826 - 0.4821) = 0.4825$. Thus, $P(z \geq 2.10817)$ is approximately $1 - P(z \leq 2.10817) \approx 1 - (.5 + 0.4825) = 0.0175$. □

The Chi-Square Distribution

The importance of the chi-square distribution results from the following fact:

> If x has a standard normal distribution then x^2 has chi-square distribution (with $\nu = 1$ degree of freedom) with density function
>
> $$\rho(x) = \begin{cases} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} x^{-\frac{1}{2}} e^{-\frac{x}{2}} & \text{if } x > 0, \\ 0 & \text{else,} \end{cases}$$
>
> with mean 1 and standard deviation $\sqrt{2}$.

In statistics we use the following more general fact:

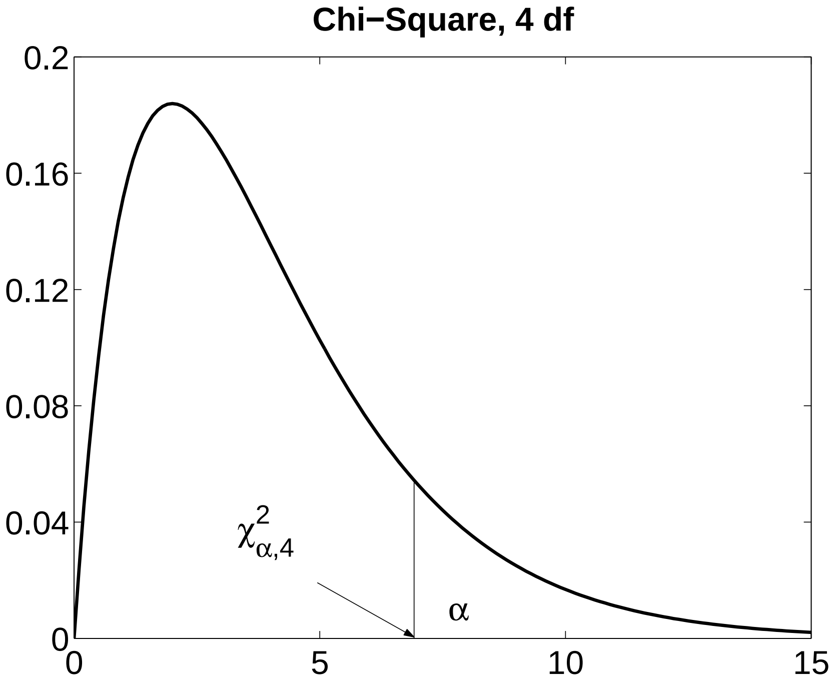
The Chi-Square Distribution

If $\bar{x}$ and s are the mean and standard deviation of a random sample of size n from a normal population with mean μ and standard deviation σ, then

- $\bar{x}$ and s^2 are independent, and
- the random variable $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom.

Tables of the chi-square distribution show, for given degree of freedom ν, the value of the random variable (here $\frac{(n-1)s^2}{\sigma^2}$, but often denoted $\chi^2_{\alpha,\nu}$), such that

$$P(\chi^2 \geq \chi^2_{\alpha,\nu}) \geq \alpha\,.$$



The Chi-Square Distribution

Example. Suppose the thickness of some semi-conductor part (with normal distribution) is critical, and suppose that the accepted variation around the mean is at most one standard deviation $\sigma = 0.6 \cdot 10^{-3}$ cm. Random samples of size 20 are taken each week to monitor the manufacturing process.

The machine is to be readjusted if the probability that s^2 will take a value greater than or equal to the observed value is 0.01 or less. What can we conclude if we find that in a sample $s = 0.84 \cdot 10^{-3}$ cm?

The machine is re-adjusted if

$$P(S^2 \geq (0.84 \cdot 10^{-3})^2) \leq 0.01\,.$$

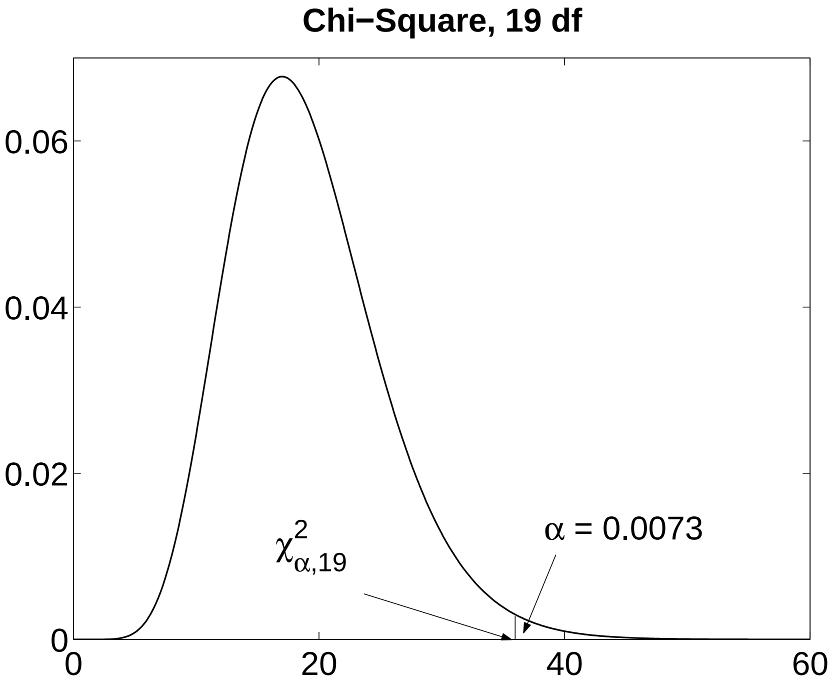
The probability on the left is that of $P(\frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)s^2}{\sigma^2})$, which has a chi-square distribution with $20 - 1$ degrees of freedom.

The Chi-Square Distribution

For $n = 20$, $s = 0.84 \cdot 10^{-3}$ and $\sigma = 0.6 \cdot 10^{-3}$ we find $\frac{(n-1)s^2}{\sigma^2} = 37.24$, and

$$P(\frac{(n-1)S^2}{\sigma^2} \geq 37.24) \approx 1 - 0.9926 = 0.0073\,,$$

so that the machine has to be re-adjusted. (The value was calculated using MATLAB.) □



The Chi-Square Distribution

In practice we have to solve such a question using tables:

In our example we want $P(\frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)s^2}{\sigma^2}) \leq 0.01$, for $\nu = 19$, $\alpha = 0.01$. From the table we find that his is the case if

$$\frac{(n-1)s^2}{\sigma^2} \geq 36.1908\,.$$

Since we found that $\frac{(n-1)s^2}{\sigma^2} = 37.24 \geq 36.1908$ the machine has to be re-adjusted.

The Student's t-Distribution

We already saw that for random samples from normal populations with mean μ and standard deviation σ that the random variable $\bar{x}$ has a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, that is, $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.

We cannot apply this knowledge in practice since usually σ, the standard deviation of the population, is unknown.

Hence we replace σ by its estimation s which we get from the random sample. The probability distribution of the random variable

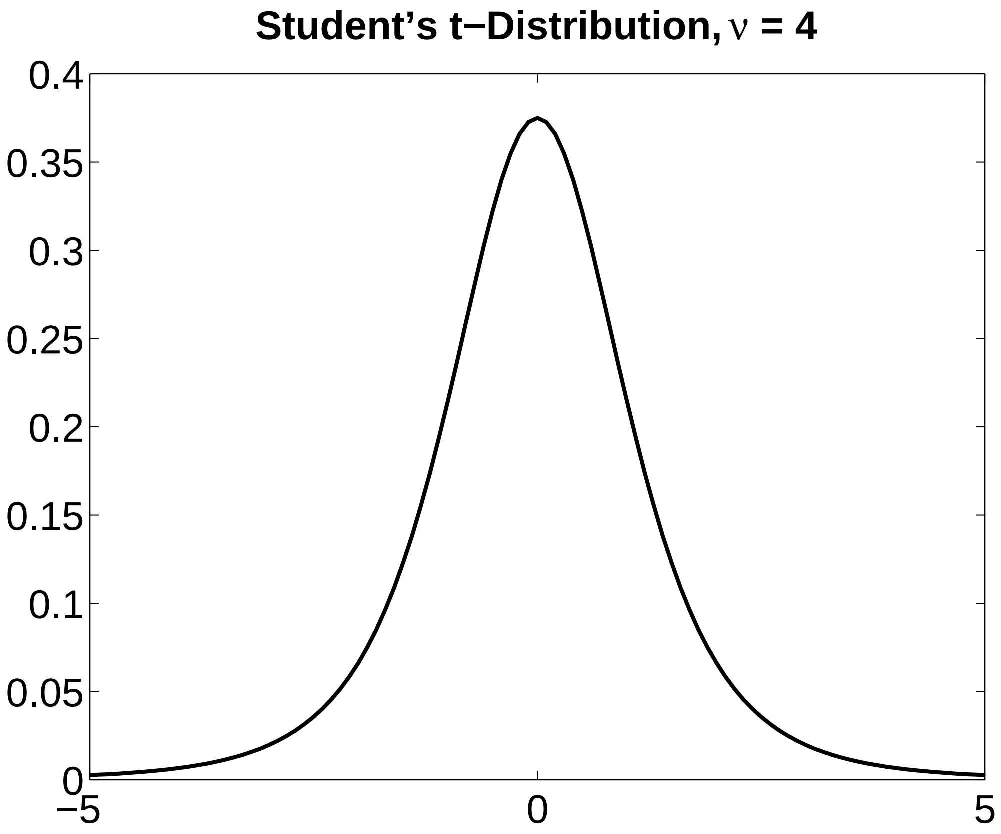
$$\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

is the t-*distribution* with $\nu = n - 1$ degrees of freedom, with density function

$$\rho(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \text{for } -\infty < t < \infty.$$

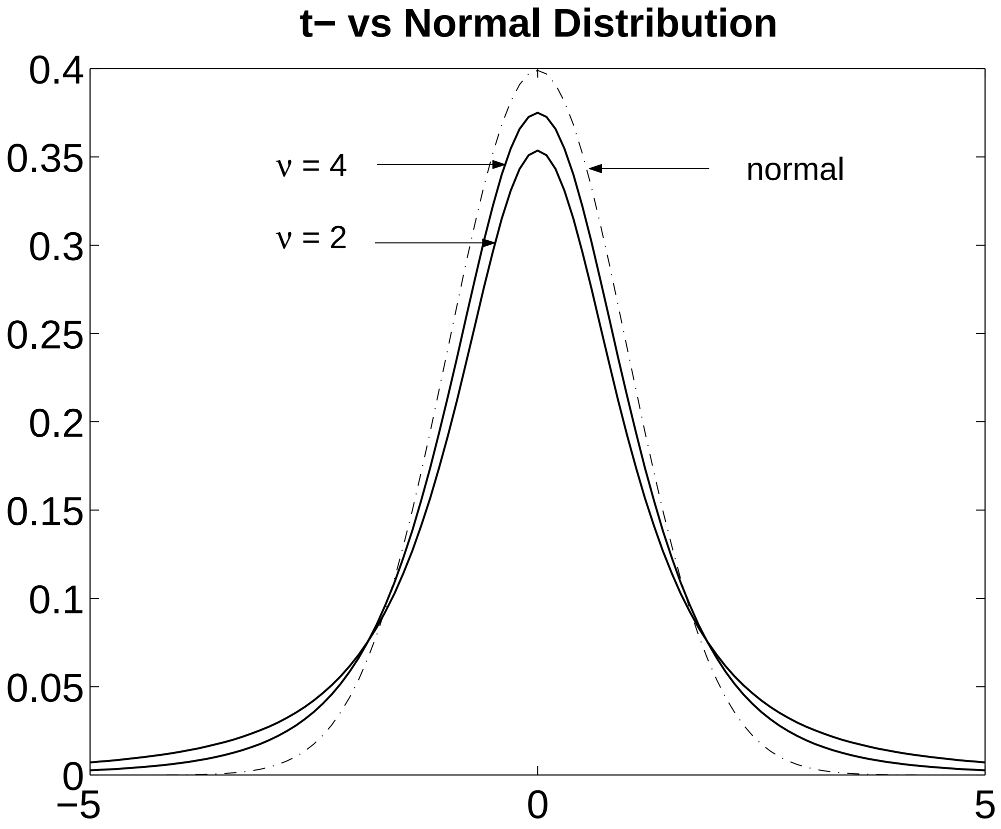
The Student's t-Distribution

W. S. Gossett discovered this distribution through his work at the Guinness brewery. At that time the brewery did not allow its staff to publish, so Gossett used the pseudonym Student, hence the name Student's t-distribution.



The Student's t-Distribution

The t-distribution is similar to the normal distribution, but is *wider* than the latter, i.e., has more tail. This is due to the fact that σ, the true population standard deviation, is only estimated. For larger ν the t-distribution becomes closer to the normal distribution.



The Student's t-Distribution

Example. The output of an old line printer is analyzed for a couple of days and it is found that the printer prints about 45 characters per second, with sample deviation 2 characters per second.

What is the probability that the sample mean of a random sample of 60 seconds will be between 44.5 and 45.3 characters per second?

Here $\bar{x} = 45$, $s = 2$, and $n = 60$. For $z = \frac{x-\bar{x}}{s/\sqrt{n}}$, which has a t-distribution, we find using MATLAB that

$$\begin{aligned} P(44.5 \leq x \leq 45.3) &= P(-1.9365 \leq z \leq 1.1619) \\ &= P(z \leq 1.1619) - P(z \leq -1.9365) \\ &= 0.8750 - 0.0288 \ = \ 0.8462. \end{aligned}$$

The Student's t-Distribution

Approximating the t-distribution by the normal distribution we find that

$$\begin{aligned} P(z \leq 1.1619) &= 0.5 + (0.3770 + .19(0.3790 - 0.3770)) \\ &= 0.8774\,, \end{aligned}$$

$$\begin{aligned} P(z \leq -1.9365) &= P(z \geq 1.9365) \\ &= 1 - P(z \leq 1.9365) \\ &= 1 - (0.5 + (0.4732 + 0.65(0.4732 - 0.4726))) \\ &= 0.0264\,, \end{aligned}$$

and thus $P(-1.9365 \leq z \leq 1.1619) = 0.851$. □

Sampling From Finite Populations

If $\bar{x}$ is the mean of a random variable of size n of a *finite* population of size N with mean μ and standard deviation σ then

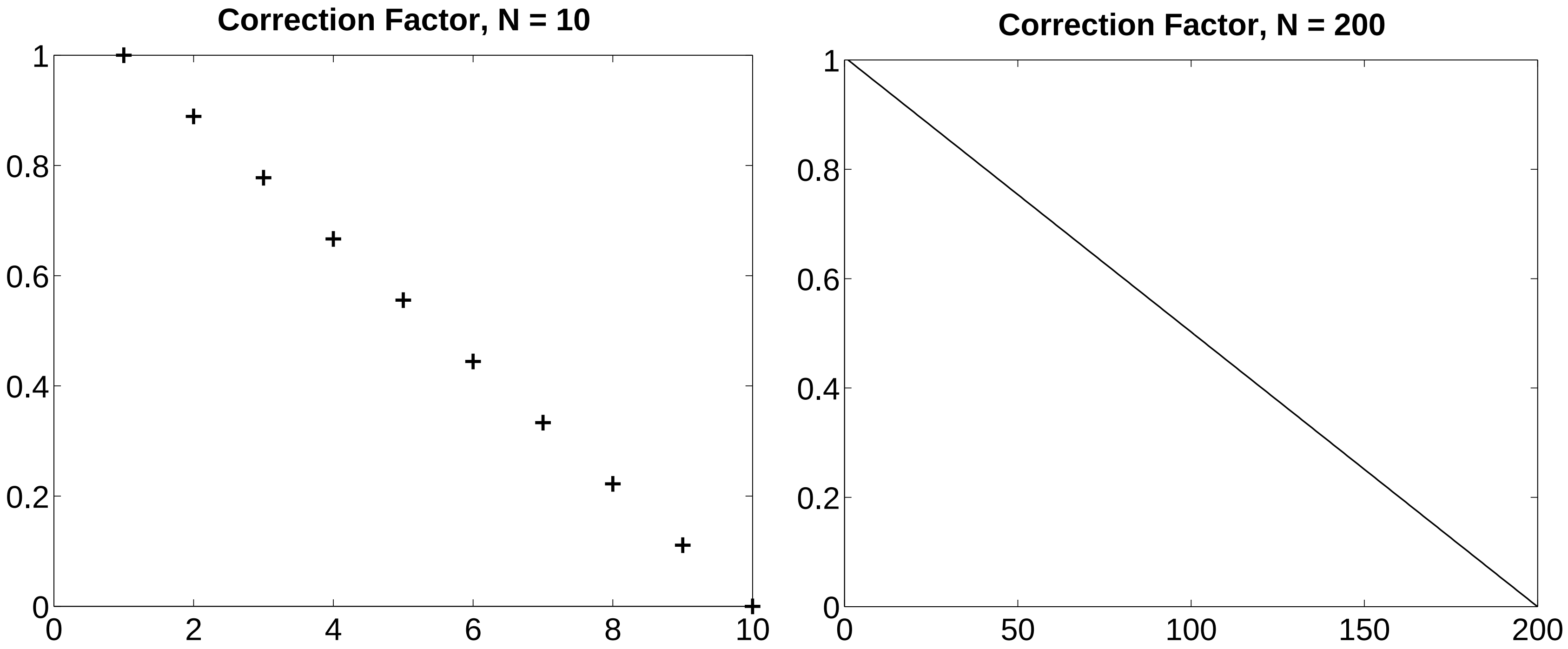
$$E(\bar{x}) = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\,.$$

These formulae are similar to those for infinite populations, except for the *finite population correction factor*

$$\sqrt{\frac{N-n}{N-1}}\,.$$

If N is large relative to n then this correction factor is close to 1 and, indeed, the distribution of $\bar{x}$ is then approximated by the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Sampling From Finite Populations


Correction Factor, N = 10

Correction Factor, N = 200

Normal as Approximation to the Binomial Distribution

Recall that for a binomial random variable the success probability was

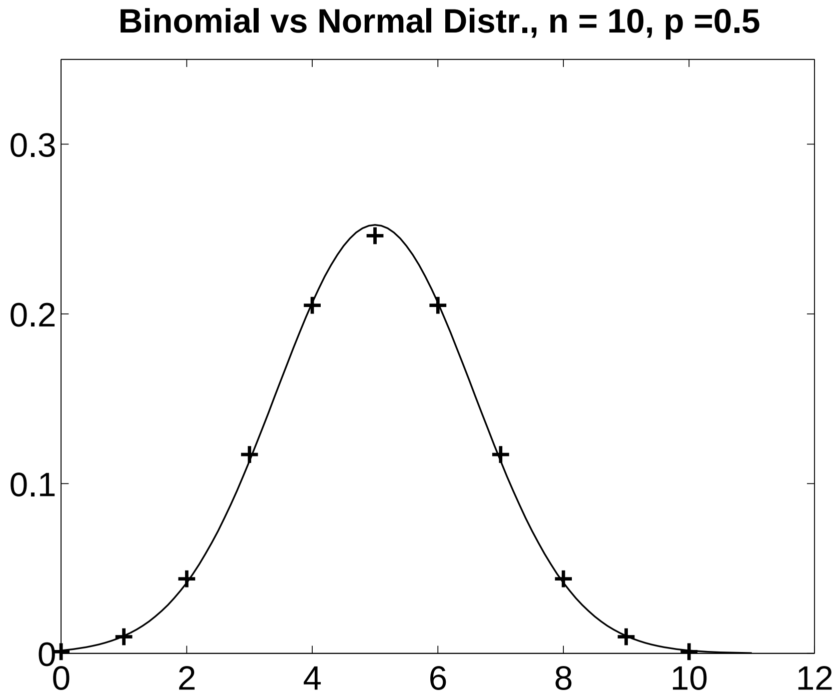
$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} ,$$

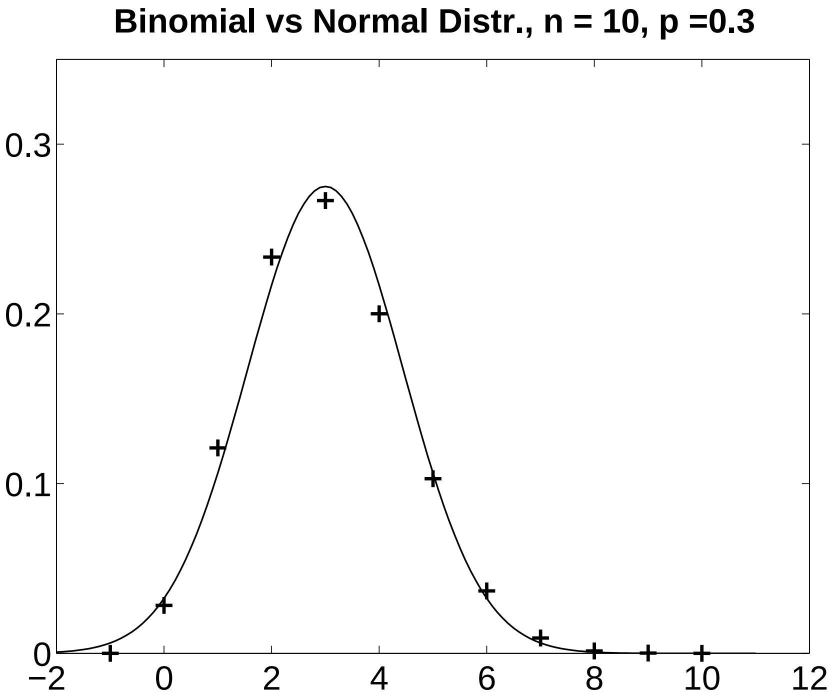
where n is the number of trials (or observations), and p is the success probability in one trial. We found that $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

The normal distribution approximates reasonably well the binomial distribution, even for small n ($n = 10$) when p is close to 0.5, and the distribution is symmetric around $\mu = np$. When p differs from 0.5 the binomial distribution is skewed, but the skewness disappears for large n. In general, the approximation is good for n large enough so that

$$0 \leq \mu - 2\sigma, \ \mu + 2\sigma \leq n .$$

Normal as Approximation to the Binomial Distribution


Binomial vs Normal Distr., n = 10, p =0.5


Binomial vs Normal Distr., n = 10, p =0.3

Note that, for example in the second case, $\mu = 3$ and $\sigma = 1.45$, so that $\mu - 2\sigma = 0.01 \geq 0$, and $\mu + 2\sigma = 5.9 \leq 10$.

Normal as Approximation to the Binomial Distribution

The following are known as *continuity correction* for the normal approximation: If x is a binomial random variable with parameters n and p, and if $z = \frac{x-\mu}{\sigma}$, then z has approximately standard normal distribution, and

- $P(x \leq a) \approx P(z \leq \frac{(a+0.5)-\mu}{\sigma})$,
- $P(x \geq a) \approx P(z \geq \frac{(a-0.5)-\mu}{\sigma})$,
- $P(a \leq x \leq b) \approx P(\frac{(a-0.5)-\mu}{\sigma} \leq z \leq \frac{(a+0.5)-\mu}{\sigma})$.

Normal as Approximation to the Binomial Distribution

Example. In quality control we randomly check 200 items if they meet the specifications. Suppose that the lot is accepted if the failure rate is below 6%. If, unknown to the quality control engineer, the failure rate is 8%, what is the probability that the lot is accepted?

In this example $n = 200$, $p = 0.08$, and we are looking for the probability that $P(x \leq 0.06 \cdot 200) = P(x \leq 12)$. Using the approximations above this is roughly

$$\begin{aligned} P(z \leq \frac{12.5 - 200 \cdot 0.08}{\sqrt{200 \cdot 0.08 \cdot 0.92}}) &= P(z \leq -0.9123) \\ &= P(z \geq 0.9123) \\ &= 0.5 - P(0 \leq z \leq 0.9123) \\ &\approx 0.5 - (0.3186 + 0.23(0.3213 - 0.3186)) \\ &= 0.180779\,. \end{aligned}$$

MATLAB gives as exact value 0.1821. □

Summary

- Sampling is about taking random sampling from (usually infinite) populations. Sampling is often used to calculate estimators for population parameters.
- The Central Limit Theorem states that for large sample size n the sample mean is approximately normally distributed with mean the true population mean, and standard deviation the true standard deviation of the population divided by the square root of the sample size.
- The random variable $\frac{(n-1)s^2}{\sigma^2}$, with s the sample error and σ the population standard deviation has a chi-square distribution with $(n-1)$ degrees of freedom.
- Usually, the true population standard deviation is not known. In this case the sample mean has a Student's t-distribution with mean the true population mean, and standard deviation the sample deviation divided by the square root of the sample size.