STATISTICS
Lecture 2a

Limit theorems and sampling distributions

Paweł Strzelecki
Lucas van der Velde

Warsaw School of Economics
Winter Semester 2018

Lecture is based on J.T. Mc Clave, P.G. Benson,T. Sincich: Statistics for Business and Economics, 11th Edition, 2010

Introduction to inference analysis

- In inference analysis we make predictions about the population parameters based on the sample statistics.
- **A (population) parameter** is a numerical descriptive measure of a population. Because it is based on the entire population, its value is almost always unknown,
- **A sample statistic** is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.
- **A sampling distribution of a sample statistic** calculated from a sample of n measurements is the probability distribution of this statistics.

Introduction to inference analysis

- We will make inference about:
  - Population mean
  - Population proportion

- Hence we need to know the distribution of the sample mean and sample proportion.

Sampling distributions

Example 1

- Consider a population consisting of the measurements 0,3,12 and described by probability distribution shown here:

x	0	3	12
$p(x)$	$1/3$	$1/3$	$1/3$

What is the sampling distribution of the sample mean ?

What is the sampling distribution of the sample median ?

Example 1

All possible samples of n=3 (random with returns):

Possible Samples	$\bar{x}$	m	Probabili
0, 0, 0	0	0	$1/27$
0, 0, 3	1	0	$1/27$
0, 0, 12	4	0	$1/27$
0, 3, 0	1	0	$1/27$
0, 3, 3	2	3	$1/27$
0, 3, 12	5	3	$1/27$
0, 12, 0	4	0	$1/27$
0, 12, 3	5	3	$1/27$
0, 12, 12	8	12	$1/27$
3, 0, 0	1	0	$1/27$
3, 0, 3	2	3	$1/27$
3, 0, 12	5	3	$1/27$
3, 3, 0	2	3	$1/27$
3, 3, 3	3	3	$1/27$
3, 3, 12	6	3	$1/27$
3, 12, 0	5	3	$1/27$
3, 12, 3	6	3	$1/27$
3, 12, 12	9	12	$1/27$
12, 0, 0	4	0	$1/27$
12, 0, 3	5	3	$1/27$
12, 0, 12	8	12	$1/27$
12, 3, 0	5	3	$1/27$
12, 3, 3	6	3	$1/27$
12, 3, 12	9	12	$1/27$
12, 12, 0	8	12	$1/27$
12, 12, 3	9	12	$1/27$
12, 12, 12	12	12	$1/27$

Sampling distribution of mean:

$\bar{x}$	0	1	2	3	4	5	6	8	9	12
$p(\bar{x})$	$1/27$	$3/27$	$3/27$	$1/27$	$3/27$	$6/27$	$3/27$	$3/27$	$3/27$	$1/27$

Sampling distribution of median:

m	0	3	12
$p(m)$	$7/27$	$13/27$	$7/27$

Sampling distribution of a mean

⇨ If sample drawn from **normally distributed** population
- In which population standard deviation is known / unknown but sample is large (n>=30)
- In which population standard deviation is unknown and sample is small (n<30)

⇨ If sample drawn from population where the variable of interest is **not normally distributed**

Sampling distribution of a mean

Consider a random sample of n observations selected from a **normally distributed** population with mean μ and standard deviation σ

$$X \sim N(\mu,\sigma) \qquad \bar{x} \sim ?$$

$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} EX_i}{n} = \frac{n\sum_{i=1}^{n} EX}{n} = EX = \mu$$

$$D^2(\bar{x}) = D^2\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} D^2X_i}{n^2} = \frac{n\sum_{i=1}^{n} D^2X}{n^2} = \frac{D^2X}{n} = \frac{\sigma^2}{n}$$

$$\boxed{\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})}$$

If population standard deviation is unknown we can use sample standard deviation S unless the sample is less than 30 elements

Sampling distribution of a mean

If the sample is small the situation is more difficult.

A distribution for a standardised sample mean was derived for small samples drawn from a normal population with an uknown σ

This standardised mean has a t-student distribution:

$$Z = \frac{\bar{X} - \mu}{S} \sim t - Student$$ with n-1 degrees of freedom

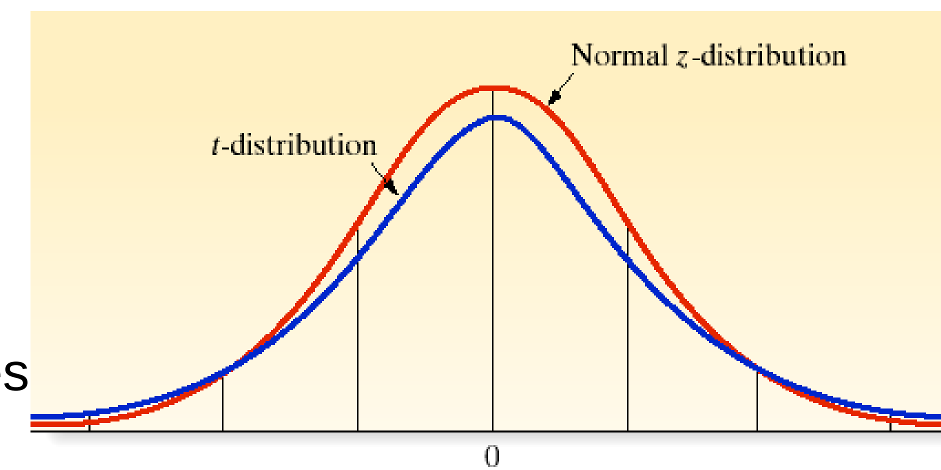
Student's t-distribution

WILLIAM S. GOSSET (1876–1937)
Student's *t*-Distribution

At the age of 23, William Gosset earned a degree in chemistry and mathematics at prestigious Oxford University. He was immediately hired by the Guinness Brewing Company in Dublin, Ireland for his expertise in chemistry. However, Gosset's mathematical skills allowed him to solve numerous practical problems associated with brewing beer. For example, Gosset applied the Poisson distribution to model the number of yeast cells per unit volume in the fermentation process. His most important discovery was that of the *t*-distribution in 1908. Since most applied researchers worked with small samples, Gosset was interested in the behavior of the mean in the small sample case. He tediously took numerous small sets of numbers, calculated the mean and standard deviation, obtained their *t*-ratio, and plotted the results on graph paper. The shape of the distribution was always the same—the *t*-distribution. Under company policy, employees were forbidden to publish their research results, so Gosset used the pen name *Student* to publish a paper on the subject. Hence, the distribution has been called Student's *t*-distribution.

For samples n>=30 t-distribution starts to overlap with the standard normal distribution (it can be approximated with a standard normal distribution).

Paramter of the distribution: degrees of freedom



Sampling distribution of a mean

Consider a random sample of n observations selected from a population which is **not normally distributed or its distribution is unknown**

When n is sufficiently large (n≥100), the sampling distribution of $\bar{x}$ will be approximately a normal with mean μ and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\boxed{\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})}$$

or

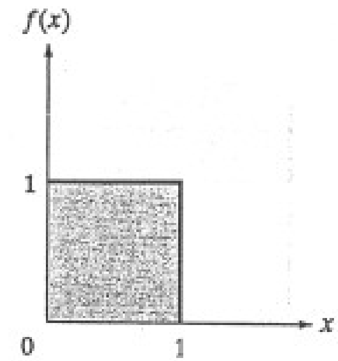
$$\boxed{\bar{x} \sim N(\mu, \frac{S}{\sqrt{n}})}$$

If population standard deviation is unknown

The larger the sample size, the better will be the normal approximation to the sampling distribution of $\bar{x}$

This theorem was first proved by Linderberg and Levy and it is called **Linderberg-Levy Limit Theorem**

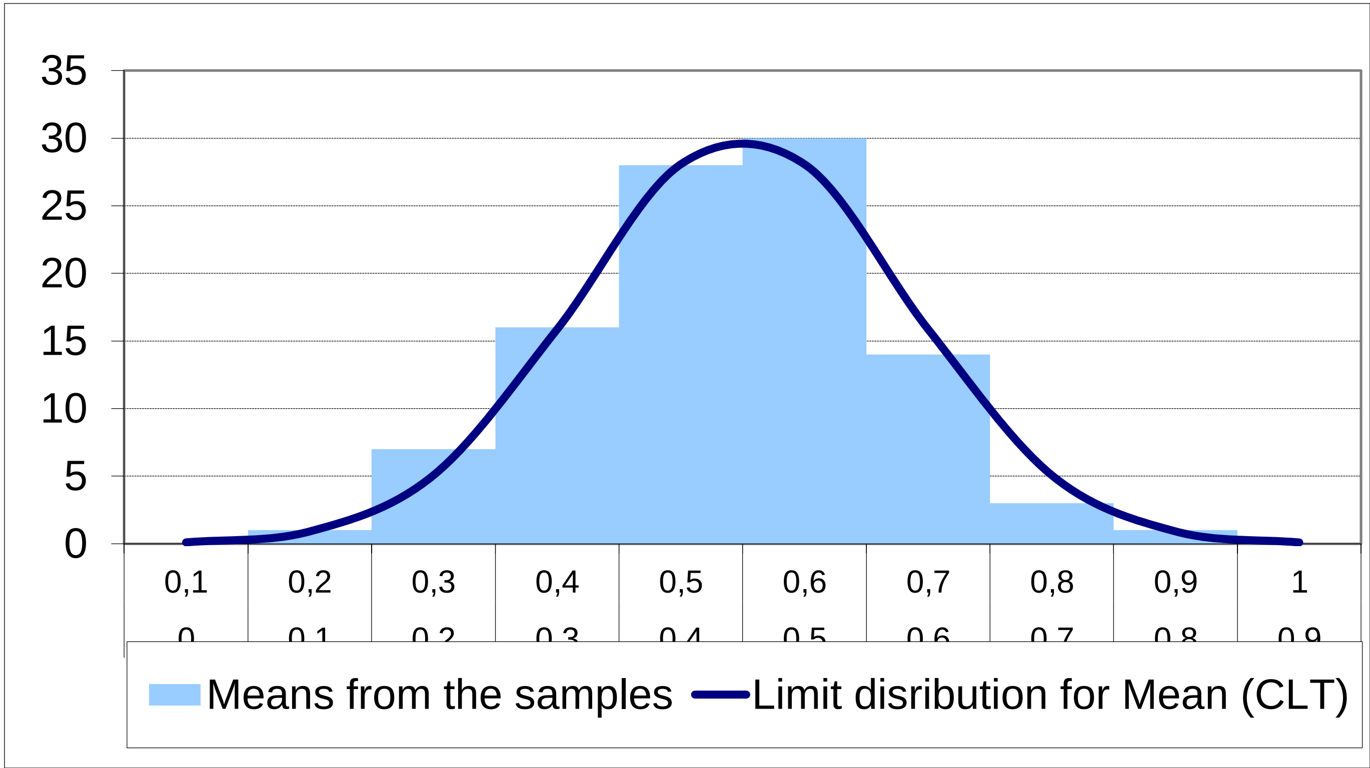
Ilustration of the Linderberg-Levy Limit Theorem



- In population continuous distribution - Uniform distribution from 0 to 1.

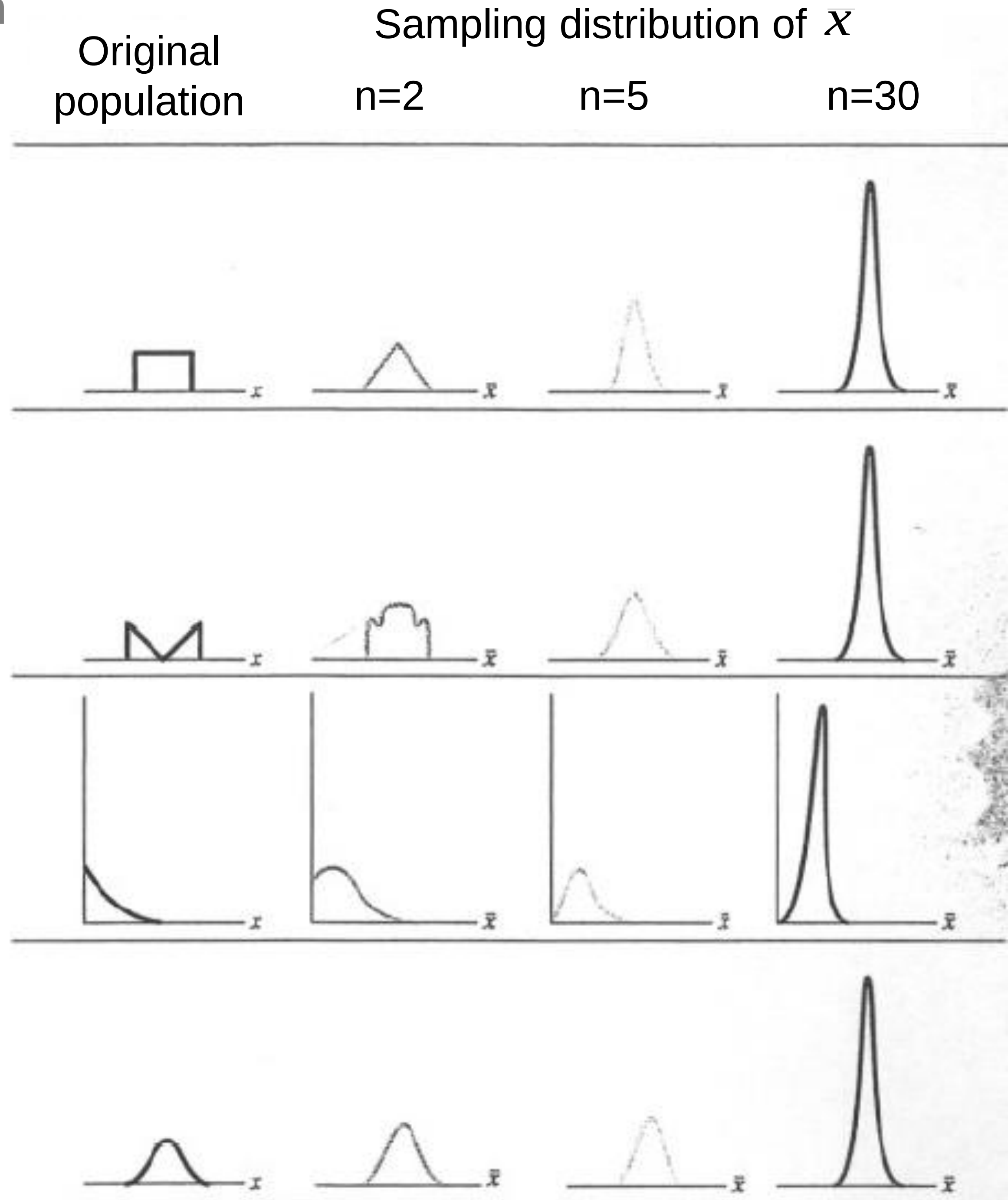
- Simulated distribution of sample mean if n=5

Example →



Sampling distribution of a mean

Sampling distribution of mean – unknown distribution in population



If the sample is large enough the distribution of the mean doesn’t depend on distribution in population

Sampling distribution of a mean

Population distribution	Population standard deviation	Sample size	Sample statistic
normal	known	any	$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
normal	unknown	>=30	$\bar{x} \sim N(\mu, \frac{S}{\sqrt{n}})$
normal	unknown	< 30	$\frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t - Student$
any	known	>=100	$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

Sampling distribution of a sum

Population distribution	Population standard deviation	Sample size	Sample statistic
normal	known	any	$\sum X_i \sim N(\mu \cdot n, \sigma \cdot \sqrt{n})$
normal	unknown	>=30	$\sum X_i \sim N(\mu \cdot n, S \cdot \sqrt{n})$
any	known	>=100	$\sum X_i \sim N(\mu \cdot n, \sigma \cdot \sqrt{n})$

Example 1

A manufacturer of automobile batteries claims that the distribution of the lengths of life if its battery has a mean of 54 months and a standard deviation of 6 months. Suppose a consumer group decides to check the claim by purchasing a sample of 100 of these batteries and subjecting them to tests that determine battery life

a) Assuming that the manufacturer's claim is true, describe the sampling distribution of the mean lifetime of a sample of 50 batteries

b) Assuming that the manufacturer's claim is true, compute the probability that the sample mean battery life is at most 52.

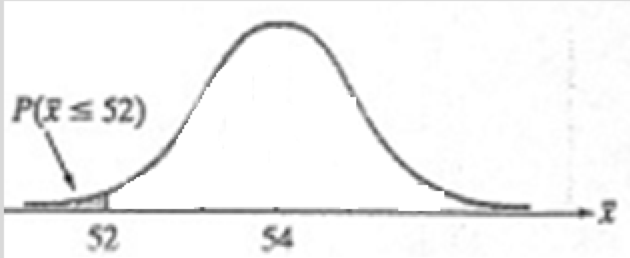
c) Assuming that the manufacturer's claim is true, describe the sampling distribution of the total life of all batteries in the sample. Compute the probability that the total life of all batteries in the sample is at most 5200.

Solution of a)

We have no information about the distribution, we only know the population mean and population standard deviation.
We can use Linderberg-Levy Limit Theorem to approximate the distribution of the sample mean:

$$\overline{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\overline{x} \sim N(54, \frac{6}{\sqrt{100}})$$

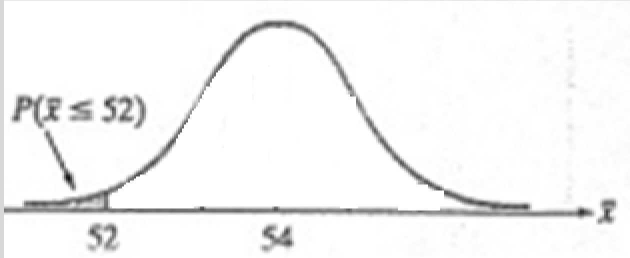


Example 1

b) Assuming that the manufacturer's claim is true, compute the probability that the sample mean battery life is at most 52.

Solution of b)

$$\bar{x} \sim N(54, 0.6)$$



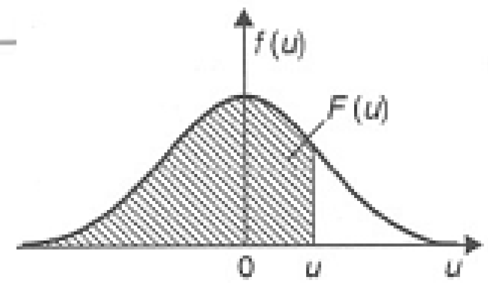
$$P(\bar{x} \leq 52) = P(Z \leq \frac{52-54}{0.6}) = P(Z < -3.33) = 1 - P(Z < 3.33) = 0.0004$$

The probability that the mean of the battery life in the sample of batteries purchased by the consumer group is less than 52 equals almost zero.

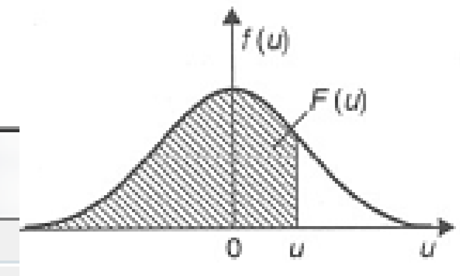
Cumulative normal distribution

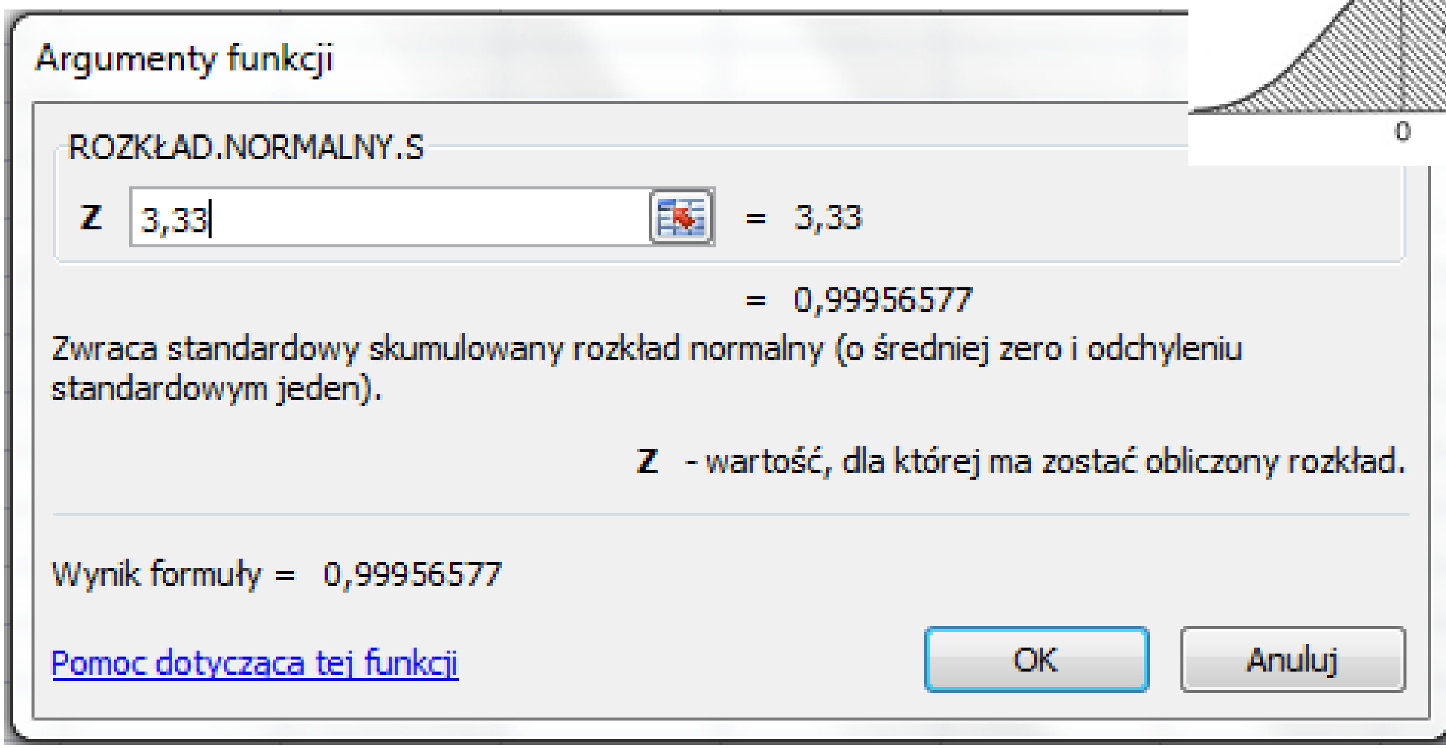
Table 1. Cumulative normal distribution

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07			
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179			
1,6	,94520	,94630	,94738	,94845	,94950	,95053	,95154	,95254			
1,7	,95543	,95637	,95728	,95818	,95907	,95994	,96080	,96164			
1,8	,96407	,96485	,96562	,96638	,96712	,96784	,96856	,96926			
1,9	,97128	,97193	,97257	,97320	,97381	,97441	,97500	,97558			
2,0	,97725	,97778	,97831	,97882	,97932	,97982	,98030	,98077	[illegible]	[illegible]	[illegible]
2,1	,98214	,98257	,98300	,98341	,98382	,98422	,98461	,98500	,98537	,98574	2,1
2,2	,98610	,98645	,98679	,98713	,98745	,98778	,98809	,98840	,98870	,98899	2,2
2,3	,98928	,98956	,98983	,$9^2$0097	,$9^2$0358	,$9^2$0613	,$9^2$0863	,$9^2$1106	,$9^2$1344	,$9^2$1576	2,3
2,4	,$9^2$1802	,$9^2$2024	,$9^2$2240	,$9^2$2451	,$9^2$2656	,$9^2$2857	,$9^2$3053	,$9^2$3244	,$9^2$3431	,$9^2$3613	2,4
2,5	,$9^2$3790	,$9^2$3963	,$9^2$4132	,$9^2$4297	,$9^2$4457	,$9^2$4614	,$9^2$4766	,$9^2$4915	,$9^2$5060	,$9^2$5201	2,5
2,6	,$9^2$5339	,$9^2$5473	,$9^2$5604	,$9^2$5731	,$9^2$5844	,$9^2$5975	,$9^2$6093	,$9^2$6207	,$9^2$6319	,$9^2$6427	2,6
2,7	,$9^2$6533	,$9^2$6636	,$9^2$6736	,$9^2$6833	,$9^2$6928	,$9^2$7020	,$9^2$7110	,$9^2$7197	,$9^2$7282	,$9^2$7365	2,7
2,8	,$9^2$7445	,$9^2$7523	,$9^2$7599	,$9^2$7673	,$9^2$7744	,$9^2$7814	,$9^2$7882	,$9^2$7948	,$9^2$8012	,$9^2$8074	2,8
2,9	,$9^2$8134	,$9^2$8193	,$9^2$8250	,$9^2$8305	,$9^2$8359	,$9^2$8411	,$9^2$8462	,$9^2$8511	,$9^2$8559	,$9^2$8605	2,9
3,0	,$9^2$8650	,$9^2$8694	,$9^2$8736	,$9^2$8777	,$9^2$8817	,$9^2$8856	,$9^2$8893	,$9^2$8930	,$9^2$8965	,$9^2$8999	3,0
3,1	,$9^3$0324	,$9^3$0646	,$9^3$0957	,$9^3$1260	,$9^3$1553	,$9^3$1836	,$9^3$2112	,$9^3$2378	,$9^3$2636	,$9^3$2886	3,1
3,2	,$9^3$3129	,$9^3$3363	,$9^3$3590	,$9^3$3810	,$9^3$4002	,$9^3$4230	,$9^3$4429	,$9^3$4623	,$9^3$4810	,$9^3$4991	3,2
3,3	,$9^3$5166	,$9^3$5335	,$9^3$5499	$\boxed{,9^3 5658}$	,$9^3$5811	,$9^3$5959	,$9^3$6103	,$9^3$6242	,$9^3$6376	,$9^3$6505	3,3
3,4	,$9^3$6631	,$9^3$6752	,$9^3$6869	,$9^3$6982	,$9^3$7091	,$9^3$7197	,$9^3$7299	,$9^3$7398	,$9^3$7493	,$9^3$7585	3,4



Computing the normal prbabilities in Excel





Excel also gives the probabilities P(Z<z), i.e. computed on the basis of the cumulative distribution function.

Example 1

c) Assuming that the manufacturer's claim is true, describe the sampling distribution of the total life of all batteries in the sample. Compute the probability that the total life of all batteries in the sample is at most 5200.

Solution of c)

$$\sum X_i \sim N(54 \cdot 100, 6 \cdot \sqrt{100})$$

$$\sum X_i \sim N(5400, 60)$$

$$P(\sum X_i \leq 5200) = P(Z \leq \frac{5200 - 5400}{60}) = P(Z < -3.33) = 1 - P(Z < 3.33) = 0.0004$$

The probability that the total life of all batteries in the sample is at most 5200 is at most 5200 equals almost zero.

Example 2

A soft-drink bottler purchases glass bottles from a vendor. The bottles are required to have an internal pressure of 150 pounds per square inch (psi) on average. One vendor claims that its bottles meet this standard. The prospective bottler strikes an agreement with the vendor that permits the bottler to sample 25 bottles to verify the vendor's claim. It turns out that the sample mean is 148.7 psi with a standard deviation of 3 psi.

Assuming the vendor's claim to be true what is the probability of obtaining a sample mean this far or father below the population mean? What does your answer suggest about the validity of vendor's claim? Assume that previous tests showed that the distribution of internal pressure of the bottle is normal.

Solution:

X – internal pressure of the bottle

$$X \sim N(150, \sigma)$$

$$\sigma = ?$$

$$n = 25 < 30$$

We cannot approximate the distribution of the sample mean with the nomrla curve.
We should then use the following formula:

$$\frac{\bar{x} - \mu}{S}\sqrt{n} \sim t - Student$$

Example 2

A soft-drink bottler purchases glass bottles from a vendor. The bottles are required to have an internal pressure of 150 pounds per square inch (psi) on average. One vendor claims that its bottles meet this standard. The prospective bottler strikes an agreement with the vendor that permits the bottler to sample 25 bottles to verify the vendor’s claim. It turns out that the sample mean is 148.7 psi with a standard deviation of 3 psi.

Assuming the vendor’s claim to be true what is the probability of obtaining a sample mean this far or father below the population mean? What does your answer suggest about the validity of vendor’s claim? Assume that previous tests showed that the distribution of internal pressure of the bottle is normal.

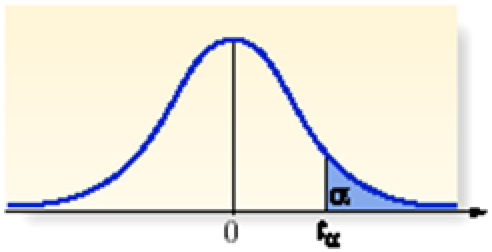
Solution:

X – internal pressure of the bottle

$$P(\bar{x}-\mu \geq 1.3) = P\left(\frac{\bar{x}-\mu}{S}\sqrt{n} \geq \frac{1.3}{3}\cdot\sqrt{25}\right) = P\left(t \geq \frac{1.3}{3}\cdot\sqrt{25}\right) =$$

$$= P\left(t \geq 2.16\right) = \ldots$$

T-table (one-sided)



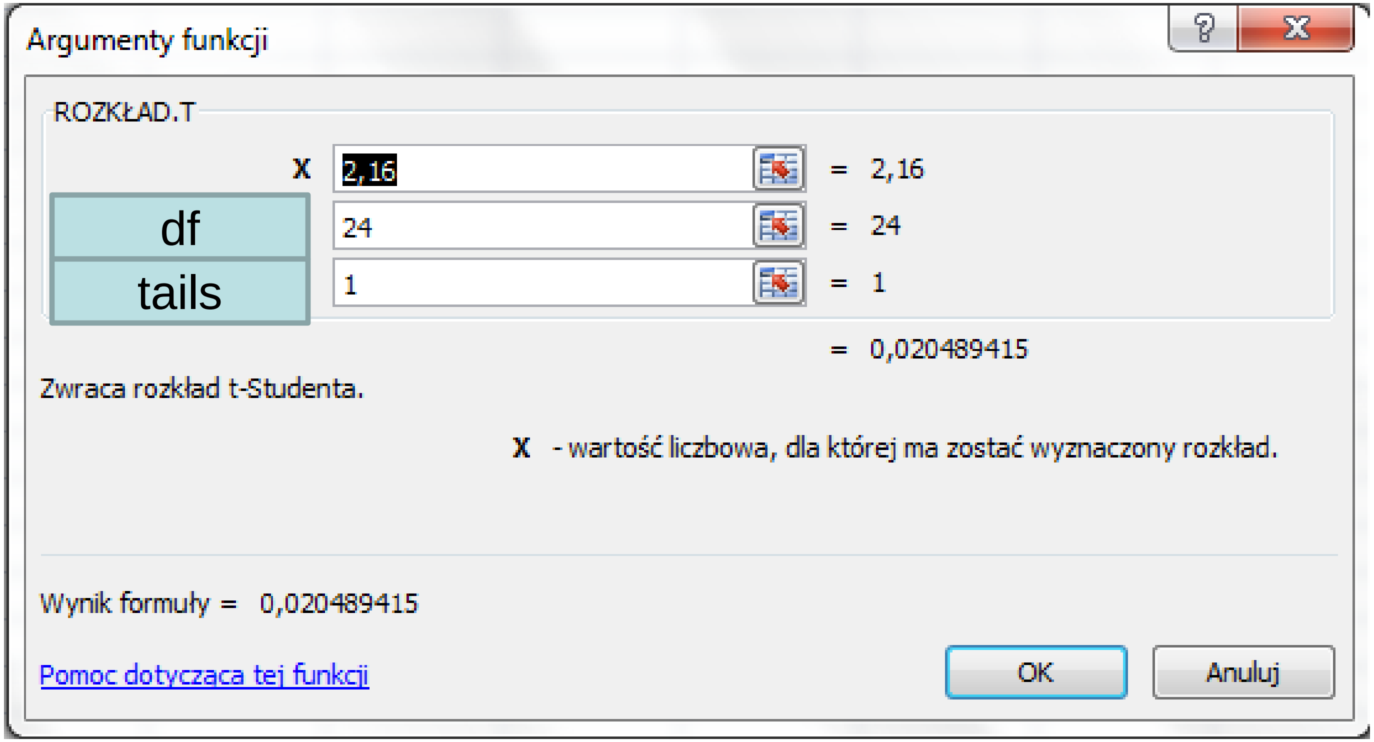
$$P\left(t \geq 2.16\right) \approx 0.025$$

Notice that t-table does not operate on cumulative distribution function by contrast to the table for normal distibution.

Table 3 Critical Values of t

Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.657	318.31
2	1.886	2.920	4.303	6.965	9.925	22.326
3	1.638	2.353	3.182	4.541	5.841	10.213
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	$\boxed{2.064}$	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421

T-table (one-sided) in Excel



$$P(t \geq 2.16) = 0.02$$

Notice that t-table does not operate on cumulative distribution function by contrast to the table for normal distibution (also in Excel!!!)

Sampling distribution of a proportion

⇨We will limit our considerations to large samples (100 or more elements)

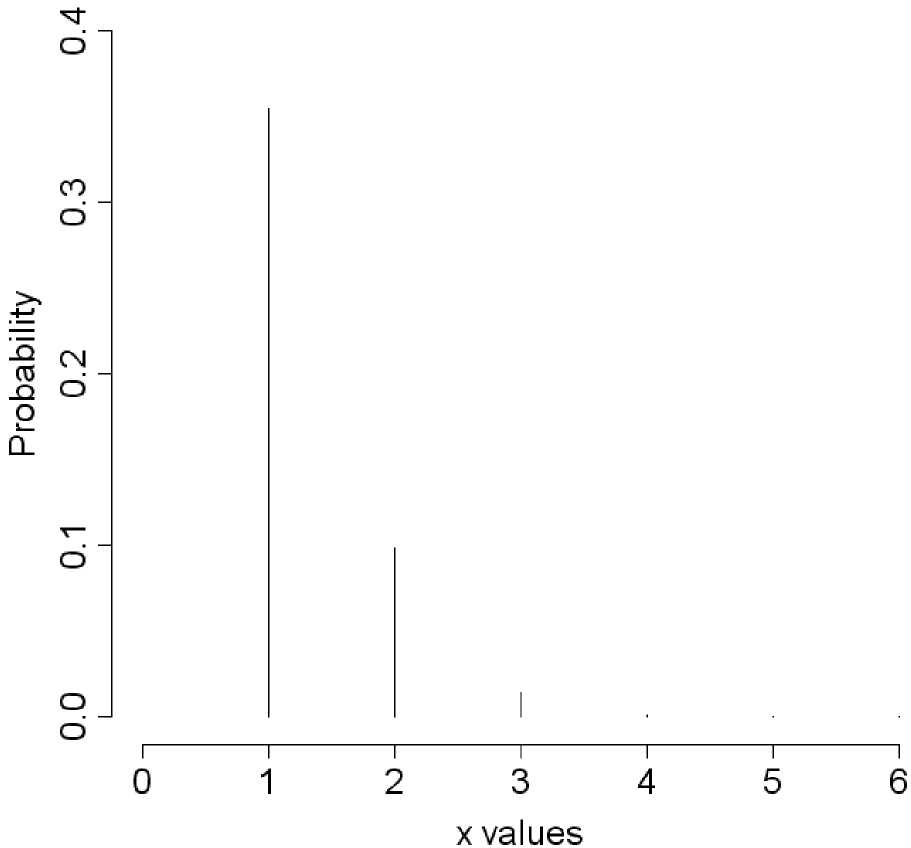
De-Moivre-Laplace limit theorem

$\hat{p} = \frac{X}{n}$ sample proportion

- Where X has a binomial distribution
- When n is large (n>=100) it is cumbersome to calculate the probabilities for the sample space
- In such situation we can approximate the # successes with a normal probability distribution with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$

This claim was proved by de-Moivre and Laplace and is called now the **de Moivre-Laplace limit theorem**

Binomial distribution if n large



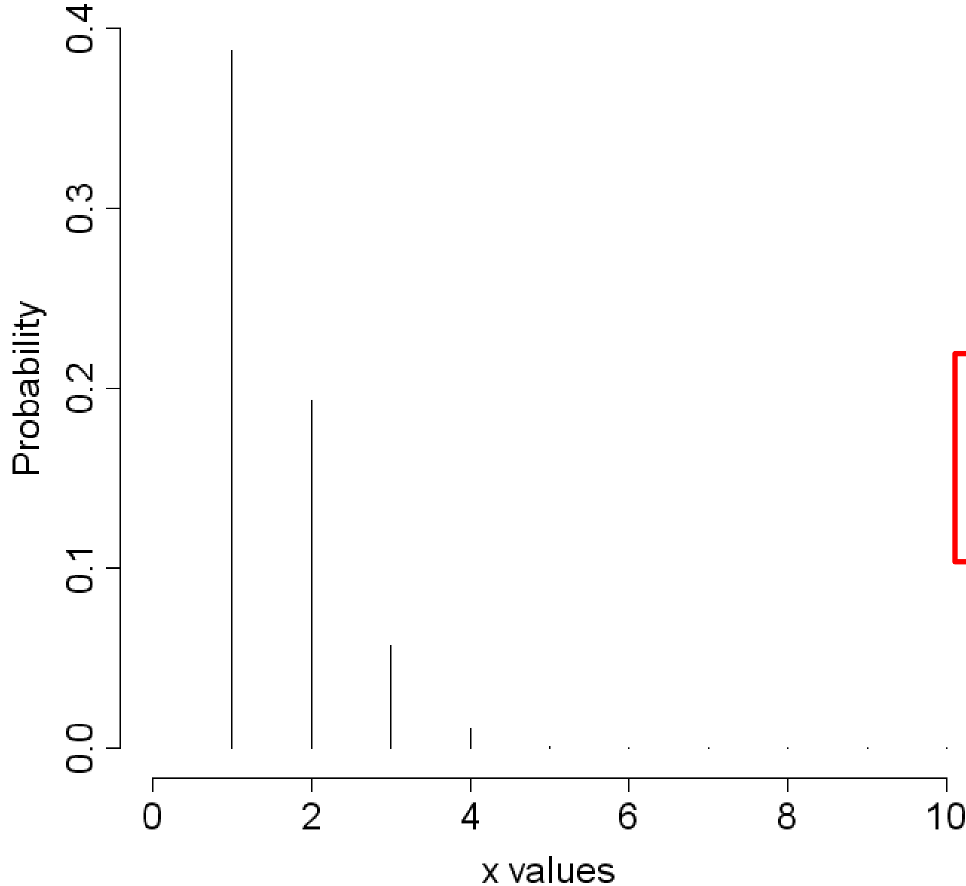
n=6, p=0.1

$$\mu = 6 \cdot 0.1 = 1$$

$$\sigma = \sqrt{10 \cdot 0.1 \cdot 0.9} = 0.73$$

$$\boxed{\begin{aligned} \mu - 3\sigma &= -1.6 \\ \mu + 3\sigma &= 2.8 \end{aligned}}$$

Binomial distribution if n large



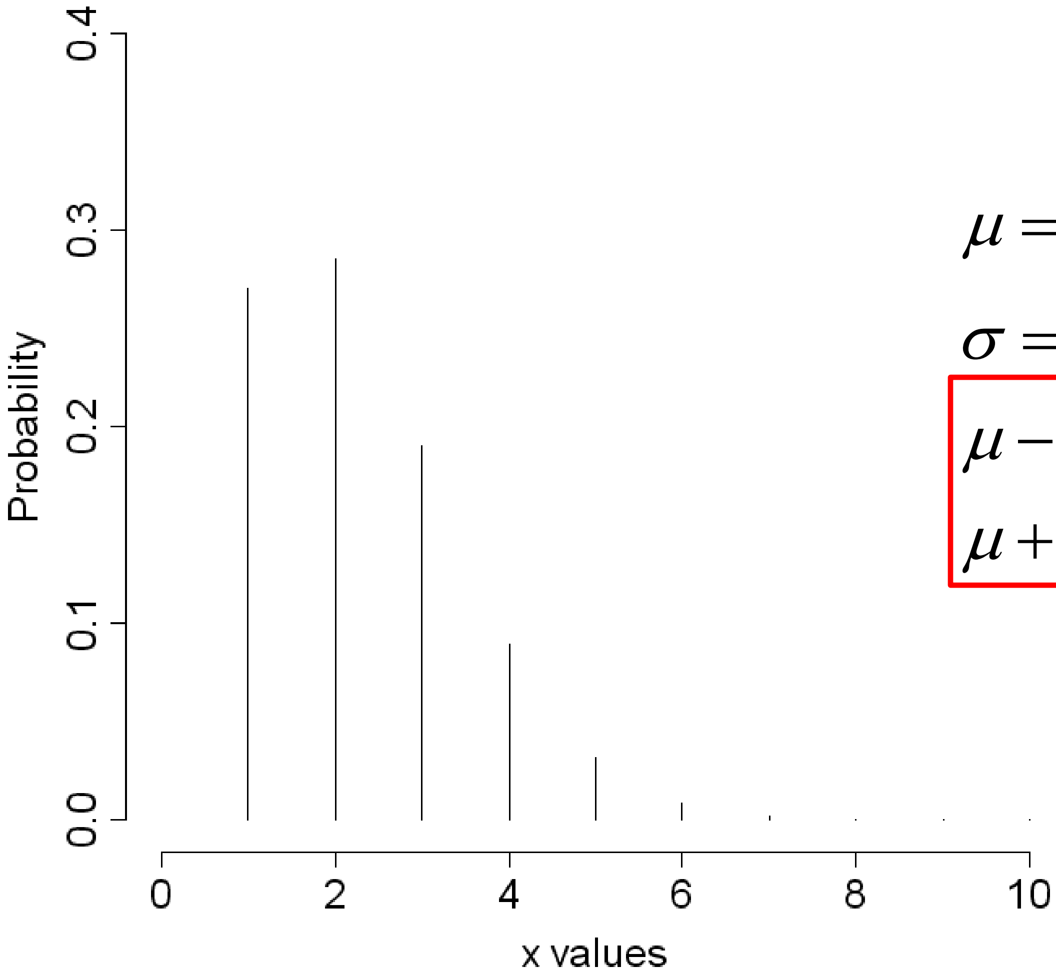
n=10, p=0.1

$$\mu = 10 \cdot 0.1 = 1$$

$$\sigma = \sqrt{10 \cdot 0.1 \cdot 0.9} = 0.94$$

$$\boxed{\begin{aligned} \mu - 3\sigma &= -1.84 \\ \mu + 3\sigma &= 3.84 \end{aligned}}$$

Binomial distribution if n large



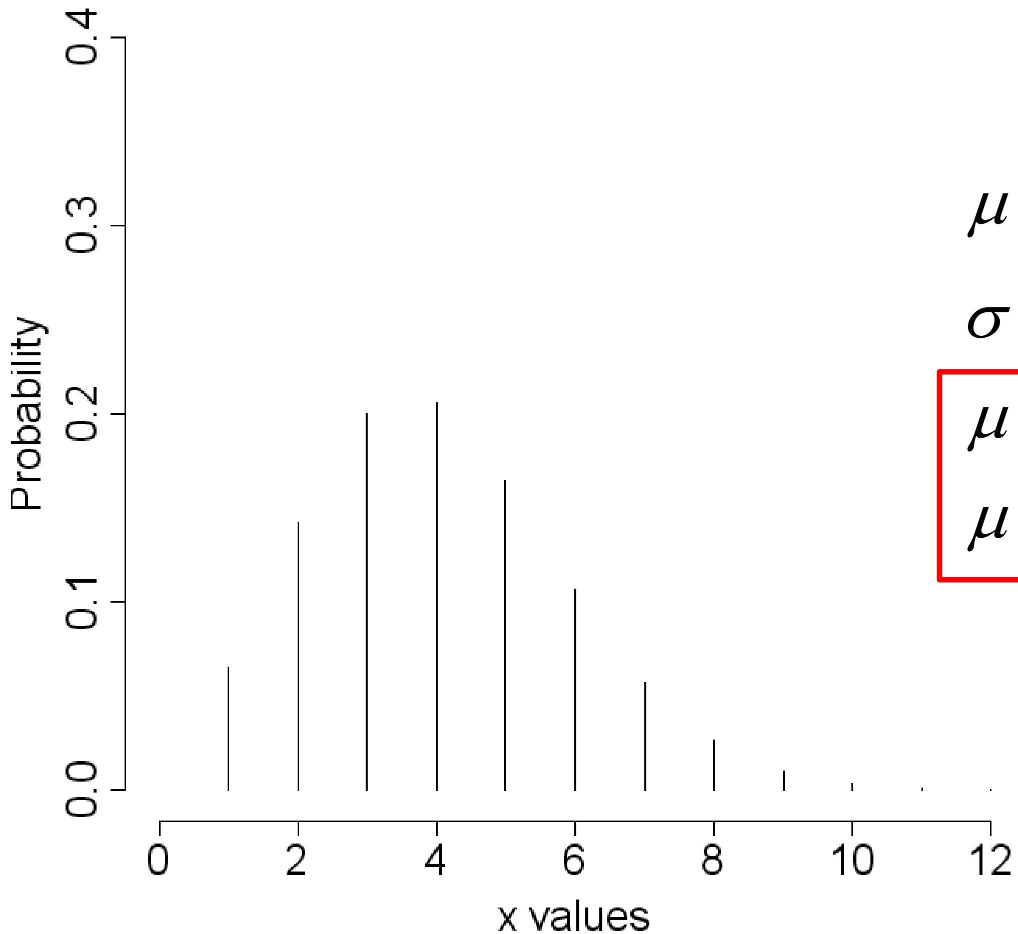
n=20, p=0.1

$$\mu = 20 \cdot 0.1 = 2$$

$$\sigma = \sqrt{20 \cdot 0.1 \cdot 0.9} = 1.34$$

$$\boxed{\begin{aligned} \mu - 3\sigma &= -2.02 \\ \mu + 3\sigma &= 6.02 \end{aligned}}$$

Binomial distribution if n large



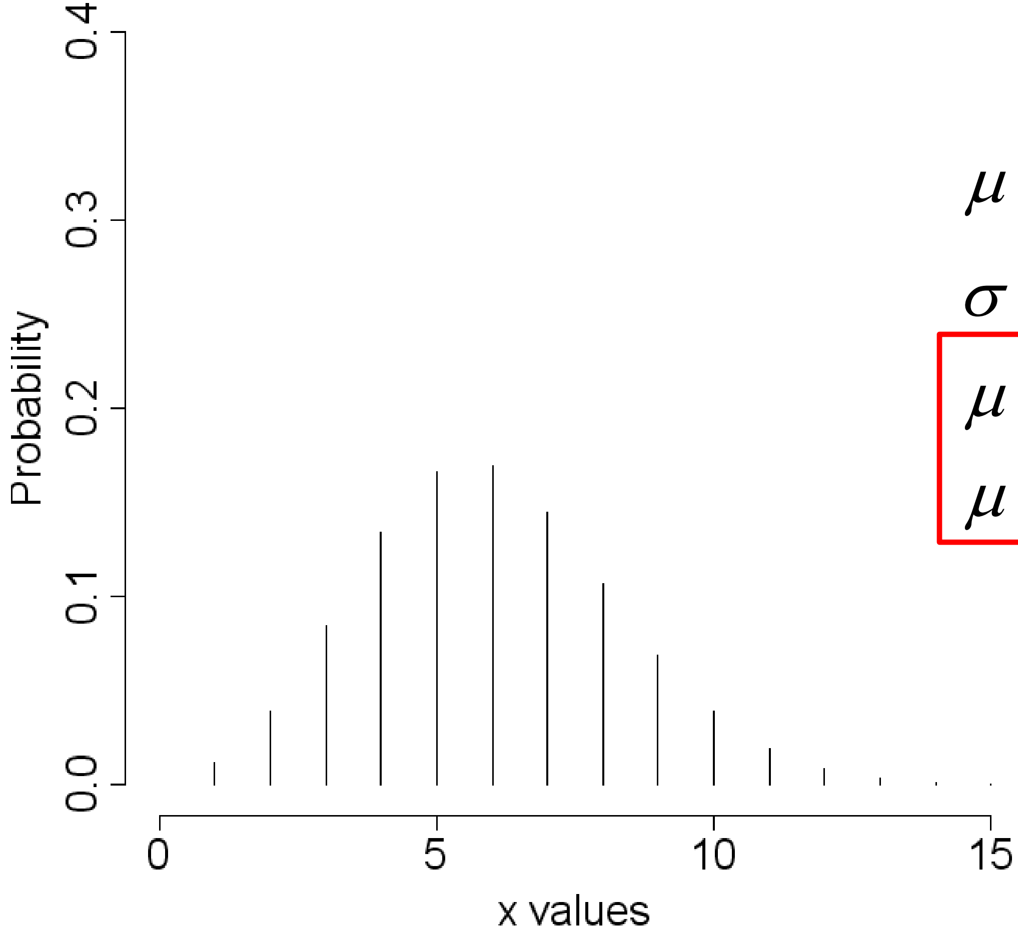
n=40, p=0.1

$$\mu = 40 \cdot 0.1 = 4$$

$$\sigma = \sqrt{40 \cdot 0.1 \cdot 0.9} = 1.90$$

$$\boxed{\begin{aligned} \mu - 3\sigma &= -1.70 \\ \mu + 3\sigma &= 9.70 \end{aligned}}$$

Binomial distribution if n large



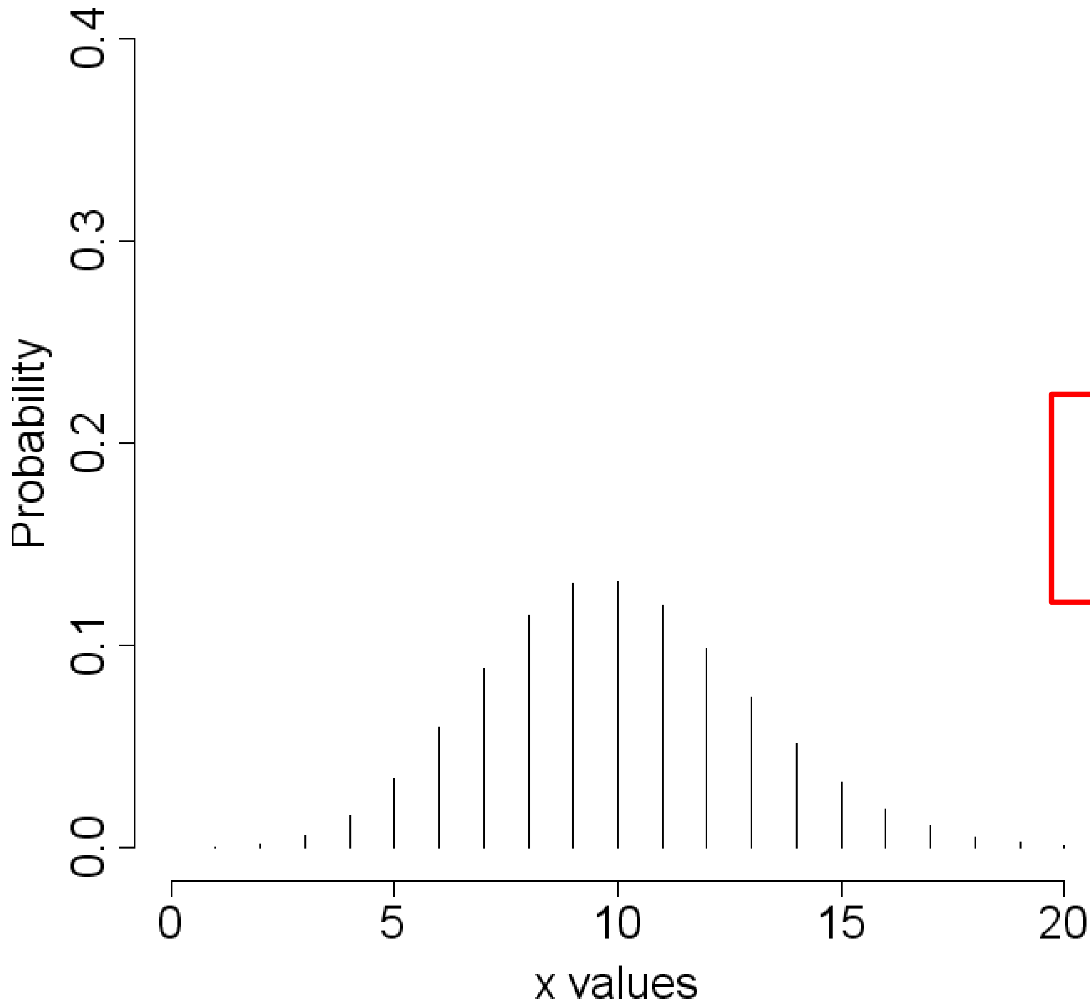
n=60, p=0.1

$$\mu = 60 \cdot 0.1 = 6$$

$$\sigma = \sqrt{60 \cdot 0.1 \cdot 0.9} = 2.32$$

$$\boxed{\begin{aligned} \mu - 3\sigma &= -0.97 \\ \mu + 3\sigma &= 13 \end{aligned}}$$

Binomial distribution if n large



n=100, p=0.1

$$\mu = 100 \cdot 0.1 = 10$$

$$\sigma = \sqrt{100 \cdot 0.1 \cdot 0.9} = 3$$

$$\boxed{\begin{aligned} \mu - 3\sigma &= 1 \\ \mu + 3\sigma &= 19 \end{aligned}}$$

Example 3

In order to test the quality of production a manufacturer of calculators chooses 200 calculators from day’s production and determines x, the number of defectives. Up to a 6% rate of defectives is acceptable. Find the probability that 20 or more defectives are observed.

Example 3

In order to test the quality of production a manufacturer of calculators chooses 200 calculators from day's production and determines x, the number of defectives. Up to a 6% rate of defectives is acceptable. Find the probability that 20 or more defectives are observed.

Solution:

- X is binomial with n=200 and p=0.06.
- The parameters of the distribution are:

$$\mu = np = 200(.06) = 12$$

$$\sigma = \sqrt{npq} = \sqrt{200(.06)(.94)} = \sqrt{11.28} = 3.36$$

- n>=100

We can use normal approximation

Example 3

- In order to test the quality of production a manufacturer of calculators chooses 200 calculators from day's production and determines x, the number of defectives. Up to a 6% rate of defectives is acceptable. Find the probability that 20 or more defectives are observed.

Solution:

$$P(X \geq 20) =$$

Example 3

- In order to test the quality of production a manufacturer of calculators chooses 200 calculators from day's production and determines x, the number of defectives. Up to a 6% rate of defectives is acceptable. Find the probability that 20 or more defectives are observed.

Solution:

$$P(X \geq 20) = 1 - P(X < 20) = 1 - P(X < \frac{20-12}{3.36}) =$$

$$= 1 - P(X < 2.38) = 1 - 0.9914 = 0.0086$$

- This means that the probability of producing more than 20 defectives among 200 is very small.
- If the manufacturer were likely to observe more than 20 defectives it would mean that the rate of defectives exceeds the acceptable level of 6%.

Sampling distribution of a proportion

- So we learnt that if n (n>=100) is large and X has a binomial distribution we can use a normal approximation for X:

$$X \sim N(np, \sqrt{np(1-p)})$$

- Similarly we can derive a distribution for a sample proportion:

$$\hat{p} = \frac{X}{n}$$

$$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

Example 4

In order to test the quality of production a manufacturer of calculators chooses 200 calculators from day's production and determines x, the number of defectives. Up to a 6% rate of defectives is acceptable. Find the probability that 10% or more defectives are observed in the sample.

Solution:

- The parameters of the distribution are:

$$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

$$\hat{p} \sim N(0.06, \sqrt{\frac{0.06(1-0.06)}{200}})$$

$$P(X \geq 0.1) = 1 - P(X < 0.1) = 1 - P(X < \frac{0.1 - 0.06}{0.0167}) =$$

$$= 1 - P(X < 2.38) = 1 - 0.9914 = 0.0086$$