Sampling Distributions

The field of statistical inference consists of those methods used to make decisions or to draw conclusions about a population. These methods utilize the information contained in a sample from the population in drawing conclusions.

Statistical Inference may be divided into two major areas:

- **Parameter Estimation**
- **Hypothesis testing**

Suppose that we want to obtain a point estimate of a population parameter. We know that before the data is collected, the observations are considered to be **random variables,** say X_1, X_2,…,X_n. Therefore, any function of the observation, or any **statistic**, is also a **random variable**.

For example, the sample mean $\bar{X}$ and the sample variance S^2 are statistics and they are also random variables.

Since a statistic is a random variable, it has a probability distribution. We call the probability distribution of a statistic a sampling distribution.

In general, if X is a random variable with probability distribution f(x), characterized by the unknown parameter θ, if $X_1, X_2,\ldots,X_n$ is a random sample of size n from X the statistic

$$\hat{\Theta} = h(X_1, X_2, \ldots X_n)$$

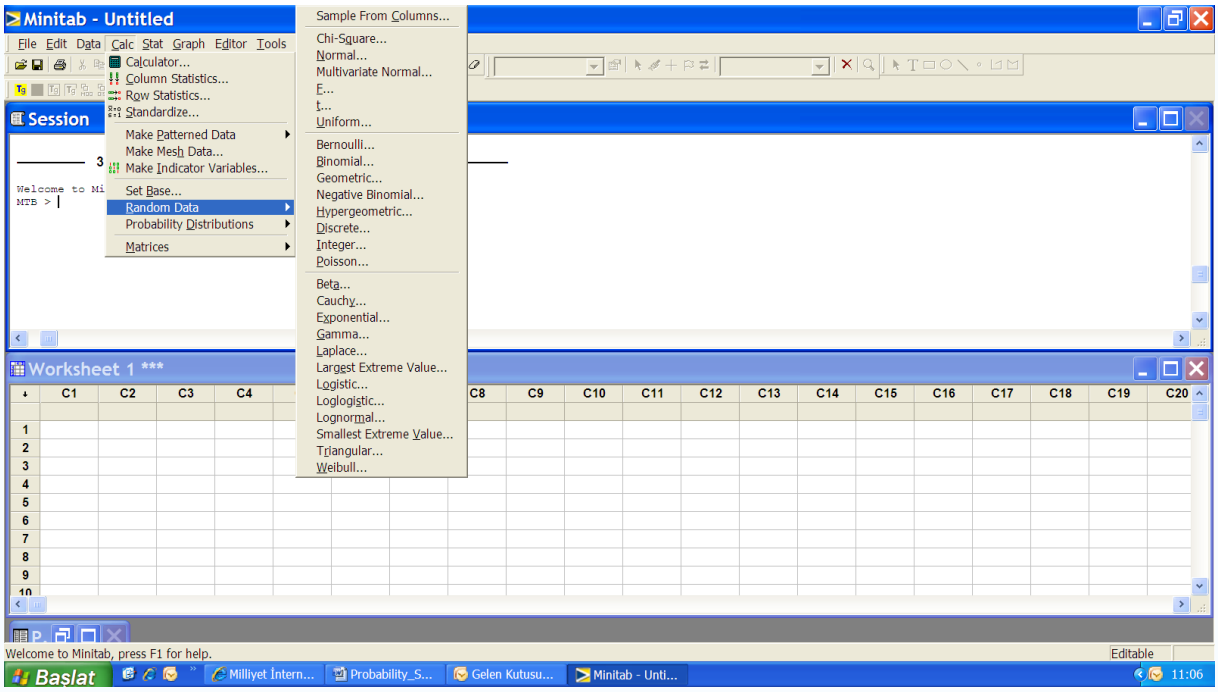
is called a point estimator of θ.

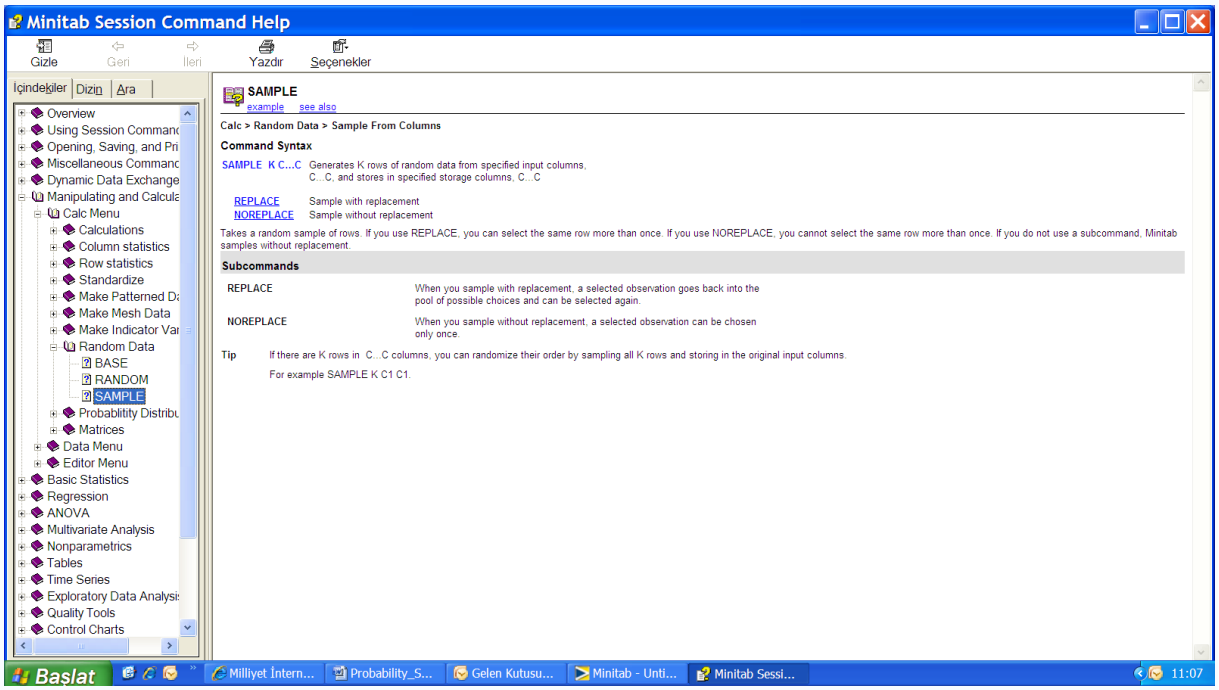
- **Note that $\hat{\Theta}$ is a *random variable* because it is a function of random variables**

After the sample has been selected, $\hat{\Theta}$ takes on a particular numerical value $\hat{\theta}$ called the point estimate of θ.

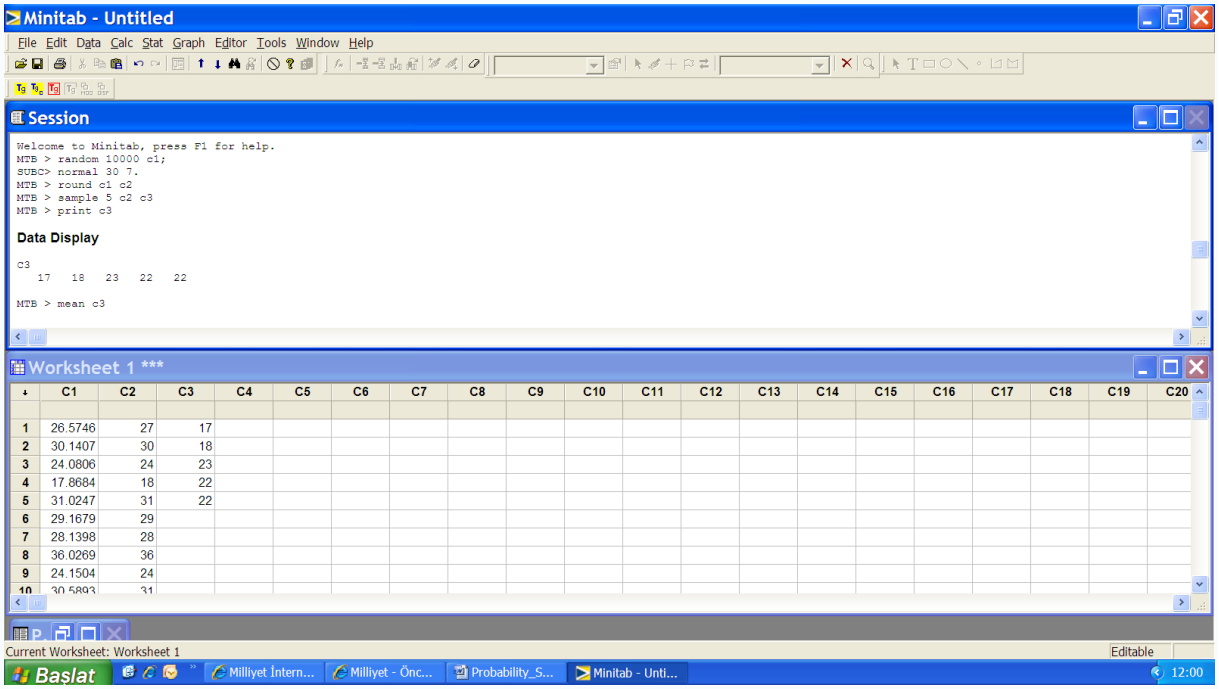
Definition: ***A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$. The statistic $\hat{\Theta}$ is called the point estimator.***

Definition: ***If a sample of n elements is selected from a population of N elements using a sampling plan in which each of the possible samples has the same chance of selection, then the sampling is said to be random and the resulting sample is a simple random sample.***


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools
Calculator...
Column Statistics...
Row Statistics...
Standardize...
Make Patterned Data
Make Mesh Data...
Make Indicator Variables...
Set Base...
Random Data
Probability Distributions
Matrices
Sample From Columns...
Chi-Square...
Normal...
Multivariate Normal...
F...
t...
Uniform...
Bernoulli...
Binomial...
Geometric...
Negative Binomial...
Hypergeometric...
Discrete...
Integer...
Poisson...
Beta...
Cauchy...
Exponential...
Gamma...
Laplace...
Largest Extreme Value...
Logistic...
Loglogistic...
Lognormal...
Smallest Extreme Value...
Triangular...
Weibull...
Session
Worksheet 1 ***
Welcome to Minitab, press F1 for help.
Editable
Başlat
Milliyet İntern...
Probability_S...
Gelen Kutusu...
Minitab - Unti...
11:06


Minitab Session Command Help
Gizle
Geri
İleri
Yazdır
Seçenekler
İçindekiler
Dizin
Ara
Overview
Using Session Command
Opening, Saving, and Pri
Miscellaneous Command
Dynamic Data Exchange
Manipulating and Calcula
Calc Menu
Calculations
Column statistics
Row statistics
Standardize
Make Patterned Da
Make Mesh Data
Make Indicator Var
Random Data
BASE
RANDOM
SAMPLE
Probablitity Distribu
Matrices
Data Menu
Editor Menu
Basic Statistics
Regression
ANOVA
Multivariate Analysis
Nonparametrics
Tables
Time Series
Exploratory Data Analysis
Quality Tools
Control Charts
SAMPLE
example see also
Calc > Random Data > Sample From Columns
Command Syntax
SAMPLE K C...C Generates K rows of random data from specified input columns, C...C, and stores in specified storage columns, C...C
REPLACE Sample with replacement
NOREPLACE Sample without replacement
Takes a random sample of rows. If you use REPLACE, you can select the same row more than once. If you use NOREPLACE, you cannot select the same row more than once. If you do not use a subcommand, Minitab samples without replacement.
Subcommands
REPLACE When you sample with replacement, a selected observation goes back into the pool of possible choices and can be selected again.
NOREPLACE When you sample without replacement, a selected observation can be chosen only once.
Tip If there are K rows in C...C columns, you can randomize their order by sampling all K rows and storing in the original input columns.
For example SAMPLE K C1 C1.
Başlat
Milliyet İntern...
Probability_S...
Gelen Kutusu...
Minitab - Unti...
Minitab Sessi...
11:07

As an example, suppose that the random variable X is normally distributed with an unknown mean μ. The sample mean is a point estimator of the unknown population mean μ.



After the sample has been selected, the numerical value $\bar{x}$ is the point estimate of population mean μ.

```
MTB > random 10000 c1;
SUBC> normal 30 7.
MTB > round c1 c2
MTB > sample 5 c2 c3
MTB > print c3
```

Data Display

C3

```
17   18   23   22   22
```

```
MTB > mean c3
```

Mean of C3

```
Mean of C3 = 20.4      >>>>>>>>>> normal 30 7.
```

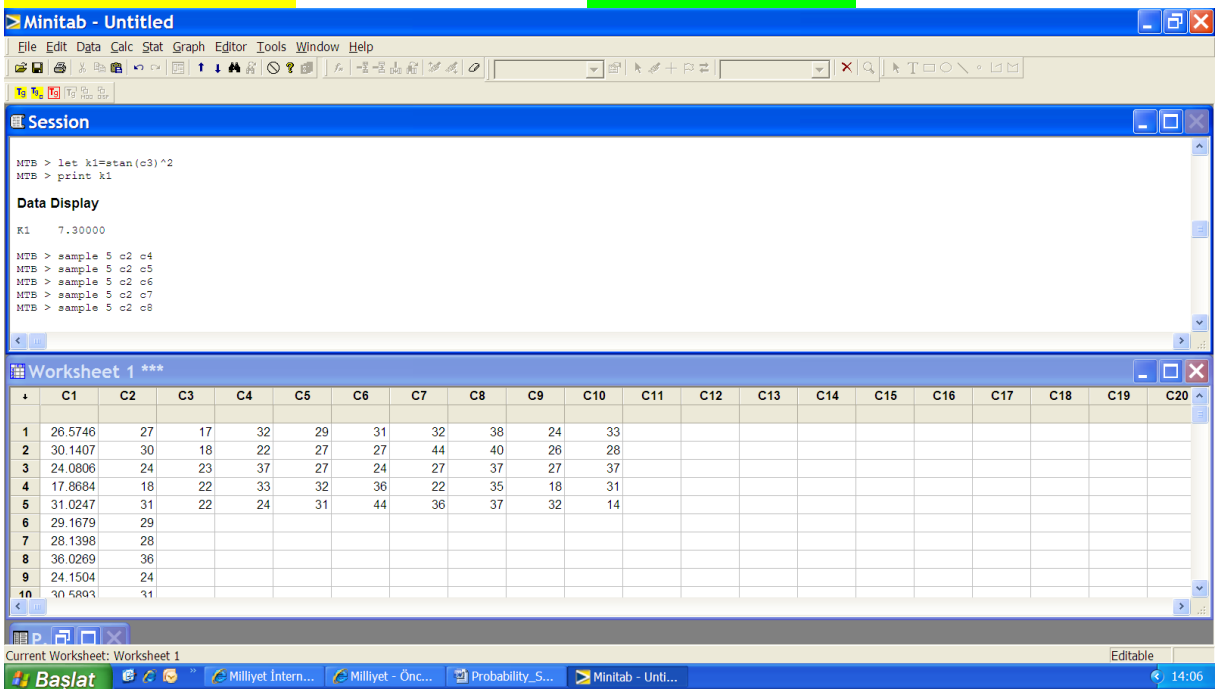
The point estimate of μ is $\bar{x} = \frac{17+18+23+22+22}{5} = 20.4$.

Similarly if the population variance σ^2 is also unknown, a point estimator for σ^2 is the sample variance S^2, and the numerical value of $S^2 = 7.3$.

```
MTB > let k1=stan(c3)^2
MTB > print k1
```

Data Display

```
K1     7.30000 >>>>>>>>>>>>>normal 30 7.
```




```
MTB > desc c3-c10
```

Descriptive Statistics: C3; C4; C5; C6; C7; C8; C9; C10

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
C3	5	0	20.40	1.21	2.70	17.00	17.50	22.00	22.50
C4	5	0	29.60	2.84	6.35	22.00	23.00	32.00	35.00
C5	5	0	29.20	1.02	2.28	27.00	27.00	29.00	31.50
C6	5	0	32.40	3.53	7.89	24.00	25.50	31.00	40.00
C7	5	0	32.20	3.77	8.44	22.00	24.50	32.00	40.00
C8	5	0	37.400	0.812	1.817	35.000	36.000	37.000	39.000
C9	5	0	25.40	2.27	5.08	18.00	21.00	26.00	29.50
C10	5	0	28.60	3.93	8.79	14.00	21.00	31.00	35.00

Variable	Maximum
C3	23.00
C4	37.00
C5	32.00
C6	44.00
C7	44.00
C8	40.000
C9	32.00
C10	37.00

In addition to simple random sampling, there are other sampling plans that involve randomization and therefore provide probabilistic basis for inference making.

When the population consists of two or more subpopulations, called strata, a sampling plan that ensures that each subpopulation is represented in the sample is called a stratified random sample.

Definition: ***Stratified random sampling involves selecting a simple random sample from each of given number of subpopulations, or strata.***

Another form of random sampling is used when the available sampling units are groups of elements, called clusters.

Definition: ***A cluster sample is a simple random sample of clusters from the available clusters in the population.***

The main difference between **cluster** sampling and **stratified sampling** is that in cluster sampling the cluster is treated as the sampling unit so analysis is done on a population of clusters (at least in the first stage). In stratified sampling, the analysis is done on elements within strata**. In stratified sampling, a random sample is drawn from each of the strata, whereas in cluster sampling only the selected clusters are studied.** The main objective of cluster sampling is to reduce costs by increasing sampling efficiency. This contrasts with stratified sampling where the main objective is to increase precision.

Estimation problems occur frequently in engineering. We often need to estimate

- **The mean μ of a single population**
- **The variance σ^2 (or standard deviation) of a single population**
- **The proportion p of items in a population that belong to a class of interest**
- **The difference in means of two populations, $\mu_1-\mu_2$**
- **The difference in two population proportions, p_1-p_2.**

Reasonable point estimates of these parameters are as follows:

- **For μ, the estimate is $\hat{\mu} = \bar{x}$ the sample mean.**
- **For σ^2, the estimate is $\hat{\sigma}^2 = S^2$, the sample variance.**
- **For p , the estimate is $\hat{p} = x/n$ sample proportion, where x is the number of items.**
- **For $\mu_1-\mu_2$, the estimate is $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$ the difference between the sample means of two independent random samples.**
- **For p_1-p_2, the estimate is $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions computed from two independent random samples.**

Definition: ***A statistic is any function of the observations in a random sample.***

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

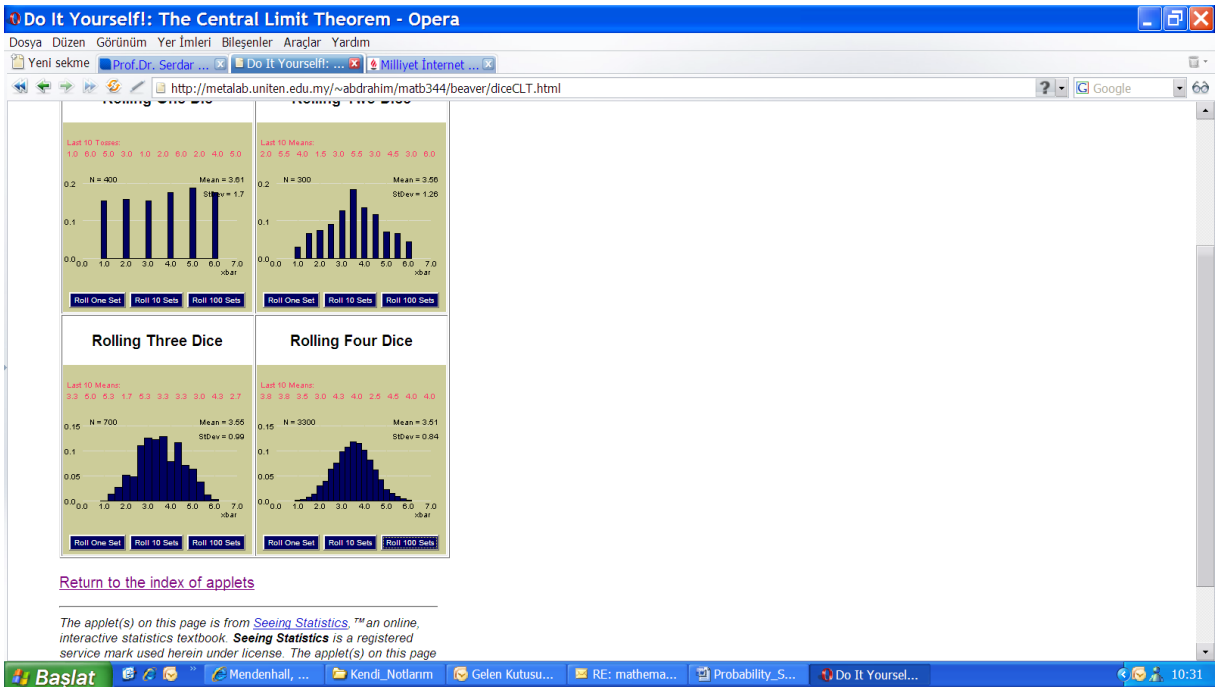
CENTRAL LIMIT THEOREM

If random samples of n observations are drawn from a nonnormal population with finite mean μ and finite variance σ^2, then, when n is large, the sampling distribution of the sample mean (sum) $\bar{x}$ is approximately normally distributed, with mean μ and standard deviation $\sigma/\sqrt{n}$.

The approximation becomes more accurate as n becomes large `normal 30 7.`

Applet

http://metalab.uniten.edu.my/~abdrahim/matb344/beaver/diceCLT.html



```
random 1000 c1-c10 ;
SUBC> integer 1 6.
MTB > rsum c1-c10 c11
MTB > hist c11
```

Histogram of C11

```
MTB > random 100000 c1-c50;
SUBC> integer 1 6.
MTB > rmean c1-c50 c51
MTB > hist c51
```

Histogram of C51

```
MTB > desc c51
```

Descriptive Statistics: C51

```
Variable        N   N*     Mean    SE Mean    StDev  Minimum       Q1
Median       Q3
C51        100000    0   3.5006   0.000765   0.2420   2.5200   3.3400
3.5000   3.6600

Variable  Maximum
C51        4.6200
```

Mean and Variance

Suppose X is a discrete uniform random variable on the consecutive integers

$$a, a+1, a+2, \ldots, b, \quad for\ a \le b$$

The mean of X is

$$\mu = E(X) = \frac{a+b}{2}$$

The variance of X is

$$\sigma^2 = V(X) = \frac{(b-a+1)^2 - 1}{12}$$

$$\mu = E(X) = \frac{a+b}{2} = \frac{1+6}{2} = 3.5$$

C1	C2								c50	C51
1	3	3	3	1	4	1	3	.	.	3.00
2	4	4	2	1	3	5	3	.	.	3.64
3	6	2	3	3	4	5	4	.	.	3.28
4	2	5	3	6	3	5	6	.	.	3.88
5	5	6	2	1	4	5	4	.	.	3.62
4	1	4	2	2	6	2	5	.	.	3.70
1	5	2	4	5	3	6	3	.	.	3.74
3	5	6	1	3	3	5	5	.	.	3.54
2	2	3	6	1	5	5	4	.	.	3.12
1	2	6	3	2	5	4	6	.	.	3.36
4	2	2	2	4	3	4	2	.	.	3.34
3	3	3	4	4	3	2	3	.	.	3.02
4	1	6	4	5	3	4	2	.	.	3.44
6	3	2	3	4	5	3	2	.	.	3.52
4	6	1	5	5	3	6	5	.	.	3.76
2	4	2	6	6	1	4	4	.	.	3.04
3	6	2	3	2	1	2	6	.	.	3.62

$$\sigma^2 = V(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{(6-1+1)^2 - 1}{12} = \frac{35}{12} = 2.916667$$

$$\sigma = \sqrt{2.916667} = 1.7078$$

```
MTB > stdev c1
```

Standard Deviation of C1

```
Standard deviation of C1 = 1.70348
```

```
MTB > stdev c2
```

Standard Deviation of C2

```
Standard deviation of C2 = 1.70755
.
.
MTB > stdev c50
```

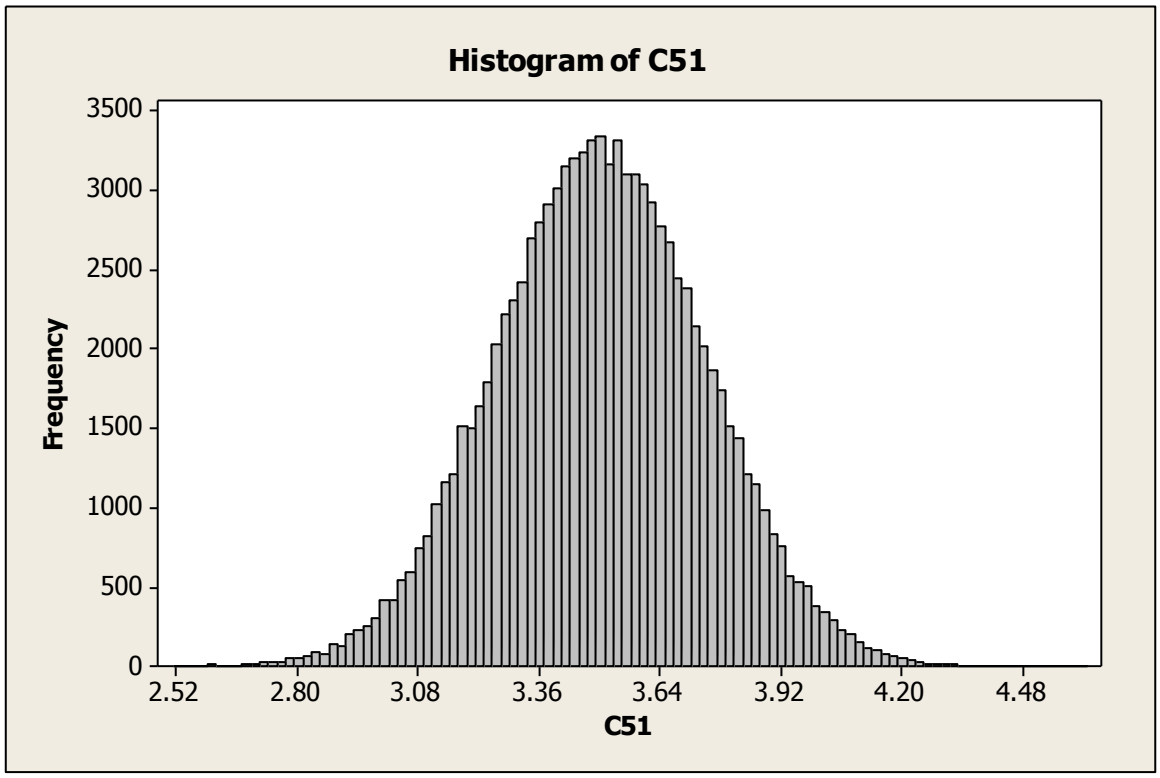
Standard Deviation of C50

```
Standard deviation of C50 = 1.70769
```

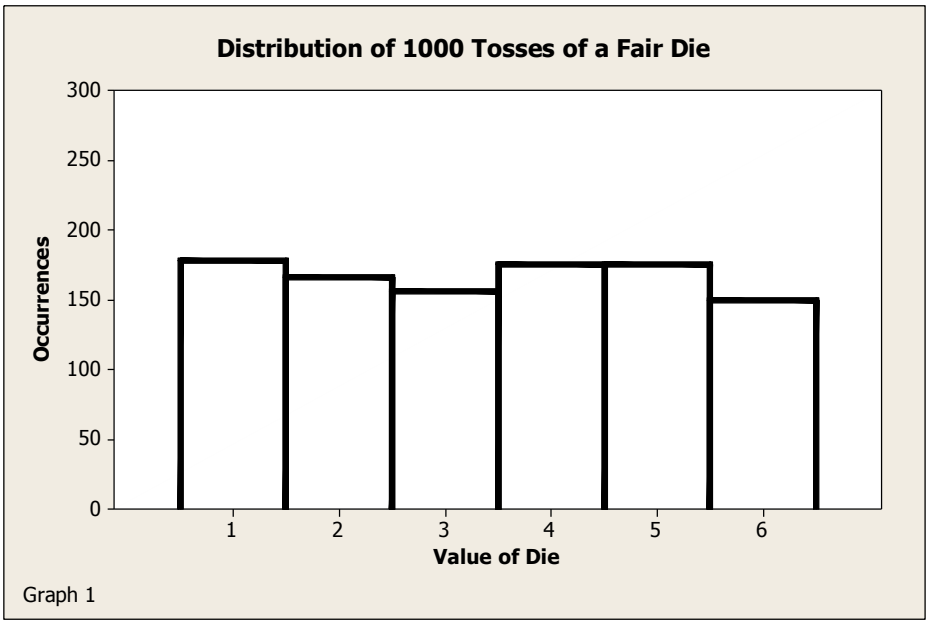
n=50

$$\sigma / \sqrt{50} = 1.7078 / \sqrt{50} = 0.24152$$

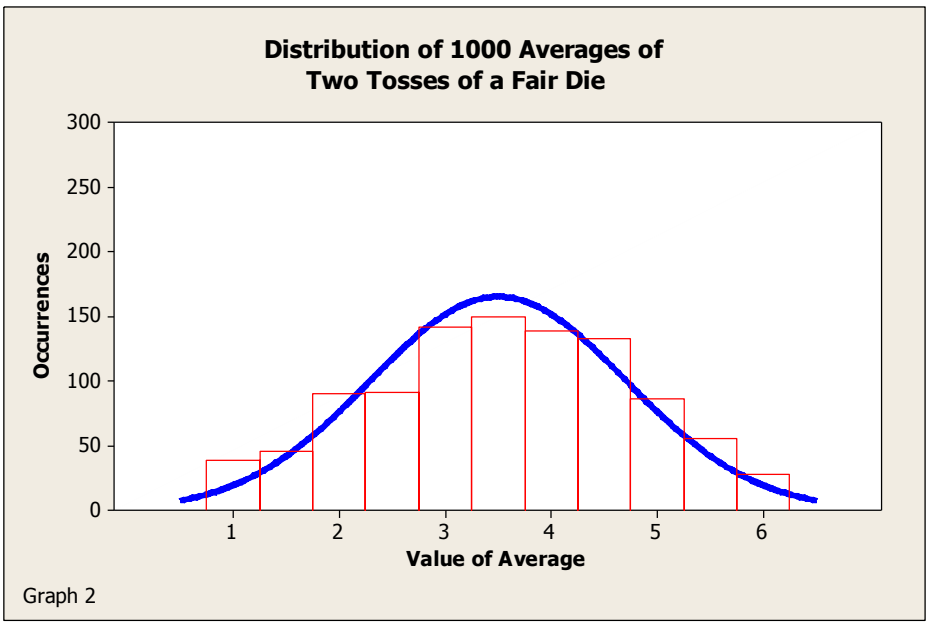
$$\sigma / \sqrt{100000} = 0.24152 / \sqrt{100000} = 0.000765$$



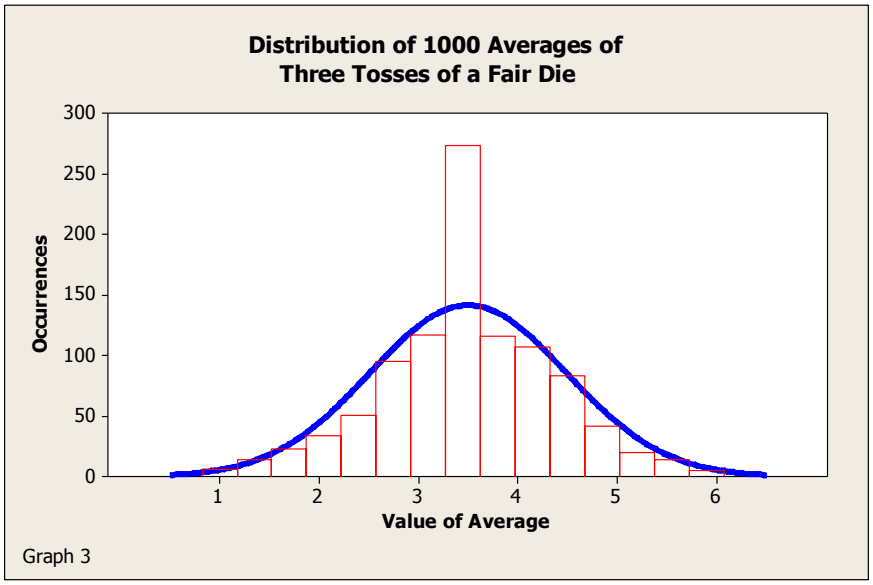
Minitab: Central Limit Theorem



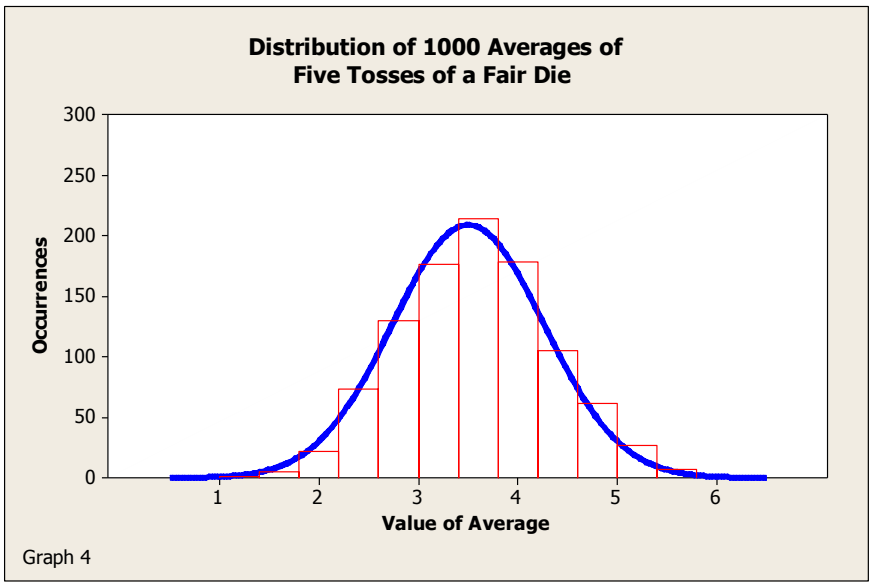
Graph 1



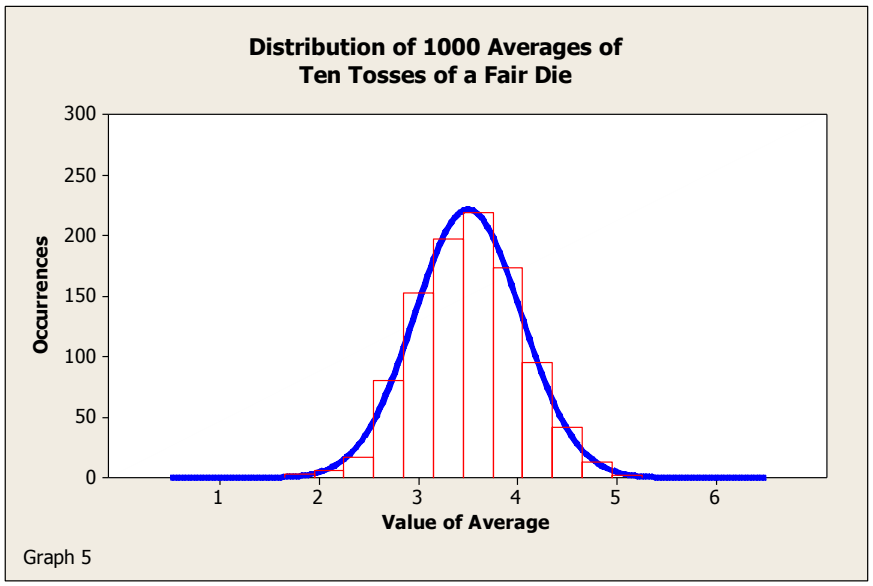
Graph 2



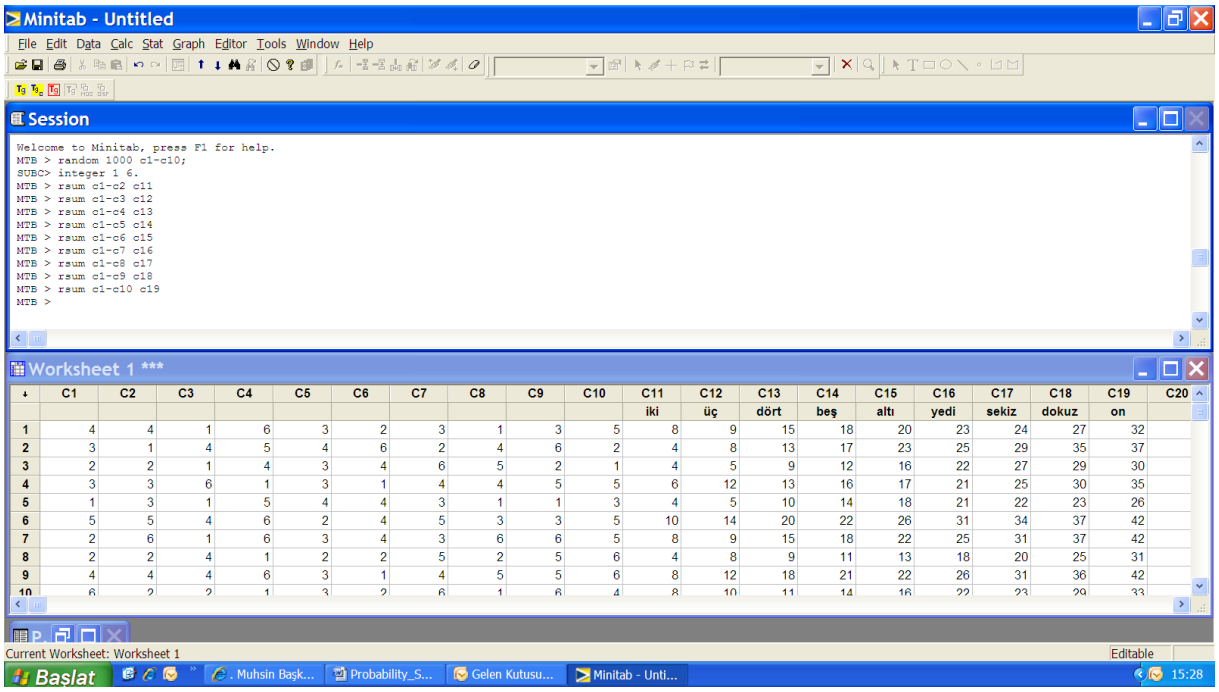
Graph 3



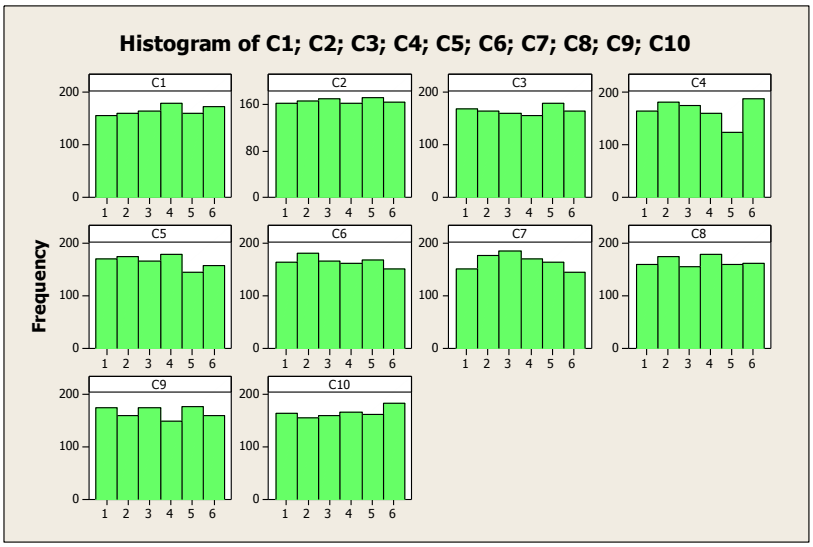
Graph 4

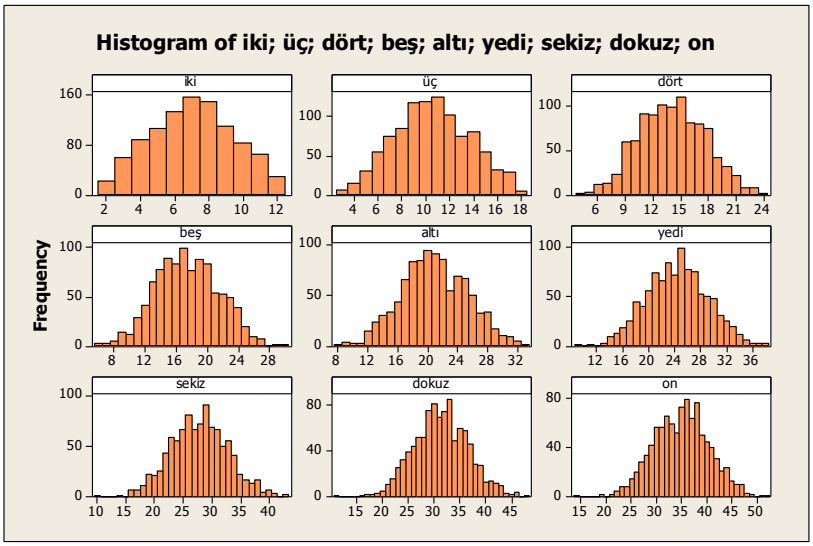


Graph 5



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
											iki	üç	dört	beş	altı	yedi	sekiz	dokuz	on	
1	4	4	1	6	3	2	3	1	3	5	8	9	15	18	20	23	24	27	32	
2	3	1	4	5	4	6	2	4	6	2	4	8	13	17	23	25	29	35	37	
3	2	2	1	4	3	4	6	5	2	1	4	5	9	12	16	22	27	29	30	
4	3	3	6	1	3	1	4	4	5	5	6	12	13	16	17	21	25	30	35	
5	1	3	1	5	4	4	3	1	1	3	4	5	10	14	18	21	22	23	26	
6	5	5	4	6	2	4	5	3	3	5	10	14	20	22	26	31	34	37	42	
7	2	6	1	6	3	4	3	6	6	5	8	9	15	18	22	25	31	37	42	
8	2	2	4	1	2	2	5	2	5	6	4	8	9	11	13	18	20	25	31	
9	4	4	4	6	3	1	4	5	5	6	8	12	18	21	22	26	31	36	42	
10	6	2	2	1	3	2	6	1	6	4	8	10	11	14	16	22	23	29	33	





Minitab - Untitled

File Edit Data Calc Stat Graph Editor Tools Window Help

Session

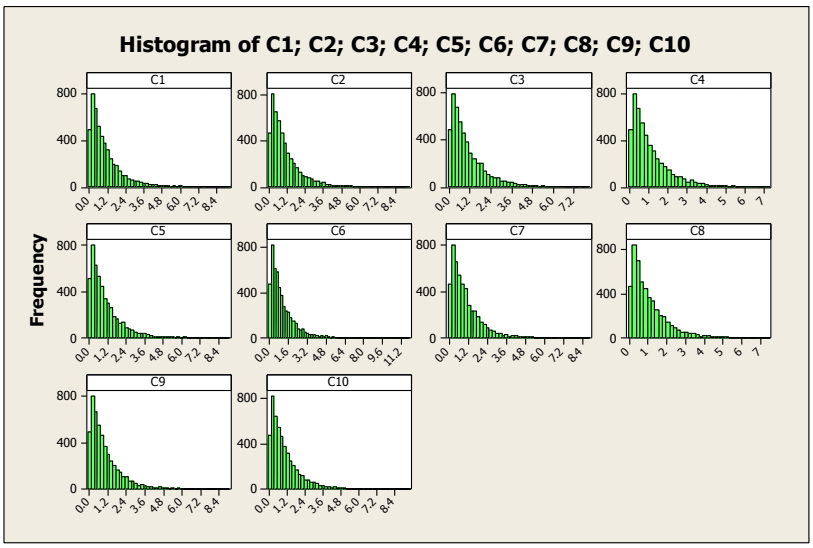
```
Welcome to Minitab, press F1 for help.
MTB > random 5000 c1-c10;
SUBC> exponential 1.
MTB > rsum c1-c2 c11
MTB > rsum c1-c3 c12
MTB > rsum c1-c4 c13
MTB > rsum c1-c5 c14
MTB > rsum c1-c6 c15
MTB > rsum c1-c7 c16
MTB > rsum c1-c8 c17
MTB > rsum c1-c9 c18
MTB > rsum c1-c10 c19
MTB >
```

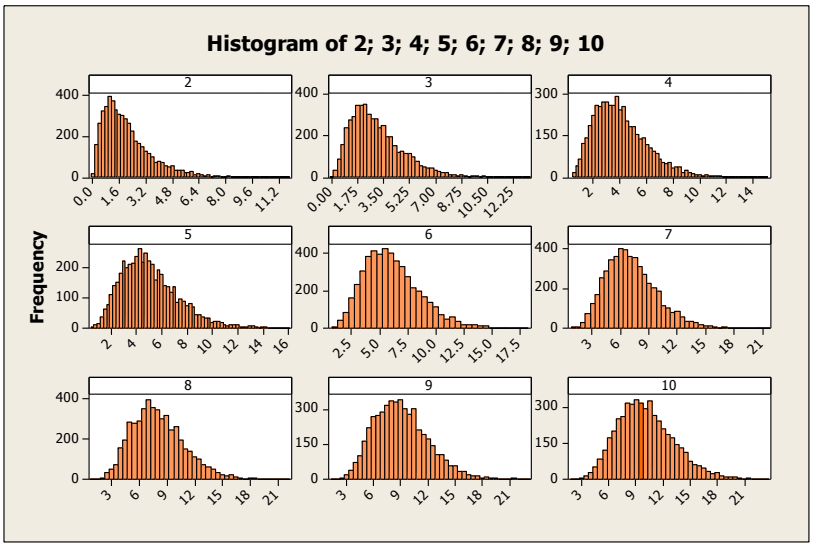
Worksheet 1 ***

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
											2	3	4	5	6	7	8	9	10	
1	0.14470	0.13403	0.45219	2.51926	0.93079	0.6851	0.83708	1.02191	0.39381	0.94141	0.2787	0.7309	3.2502	4.1810	4.8661	5.7031	6.7251	7.1189	8.0603	
2	0.74940	1.05974	1.61262	1.99371	0.92854	0.7591	0.12771	2.50083	0.15402	0.38152	1.8091	3.4218	5.4155	6.3440	7.1031	7.2308	9.7316	9.8856	10.2672	
3	0.69264	0.68466	0.34648	0.06069	0.53654	0.0274	2.11530	0.70528	0.79354	0.45097	1.3773	1.7238	1.7845	2.3210	2.3484	4.4637	5.1690	5.9626	6.4135	
4	0.82752	2.12913	1.70485	1.51827	0.10800	0.3862	2.31475	0.16537	1.39195	0.09310	2.9566	4.6615	6.1798	6.2878	6.6740	8.9887	9.1541	10.5460	10.6391	
5	4.24948	1.22451	0.04028	0.13104	0.05408	0.7141	0.34491	0.07411	1.09719	0.94548	5.4740	5.5143	5.6453	5.6994	6.4135	6.7585	6.8326	7.9298	8.8752	
6	2.10616	1.39540	1.28278	2.85782	2.19210	0.0081	1.00617	1.16521	0.09448	0.49029	3.5016	4.7843	7.6422	9.8343	9.8424	10.8486	12.0138	12.1083	12.5985	
7	0.15264	0.85021	0.34532	0.11640	0.26137	0.1917	0.11018	0.78406	0.09485	0.69055	1.0029	1.3482	1.4646	1.7259	1.9176	2.0278	2.8119	2.9067	3.5973	
8	0.20670	0.27270	0.64104	0.69355	0.47419	1.9369	0.84922	0.89485	1.06475	0.06063	0.4794	1.1204	1.8140	2.2882	4.2251	5.0743	5.9691	7.0339	7.0945	
9	1.71834	0.43748	0.86760	0.50202	1.26254	0.0252	1.10003	0.13291	0.88600	1.62148	2.1558	3.0234	3.5255	4.7880	4.8132	5.9132	6.0462	6.9322	8.5536	
10	3.89349	0.26858	1.23687	0.86639	3.28080	0.6272	0.10728	0.64424	0.72562	1.35453	4.1621	5.3989	6.2653	9.5461	10.1734	10.2806	10.9249	11.6505	13.0050	

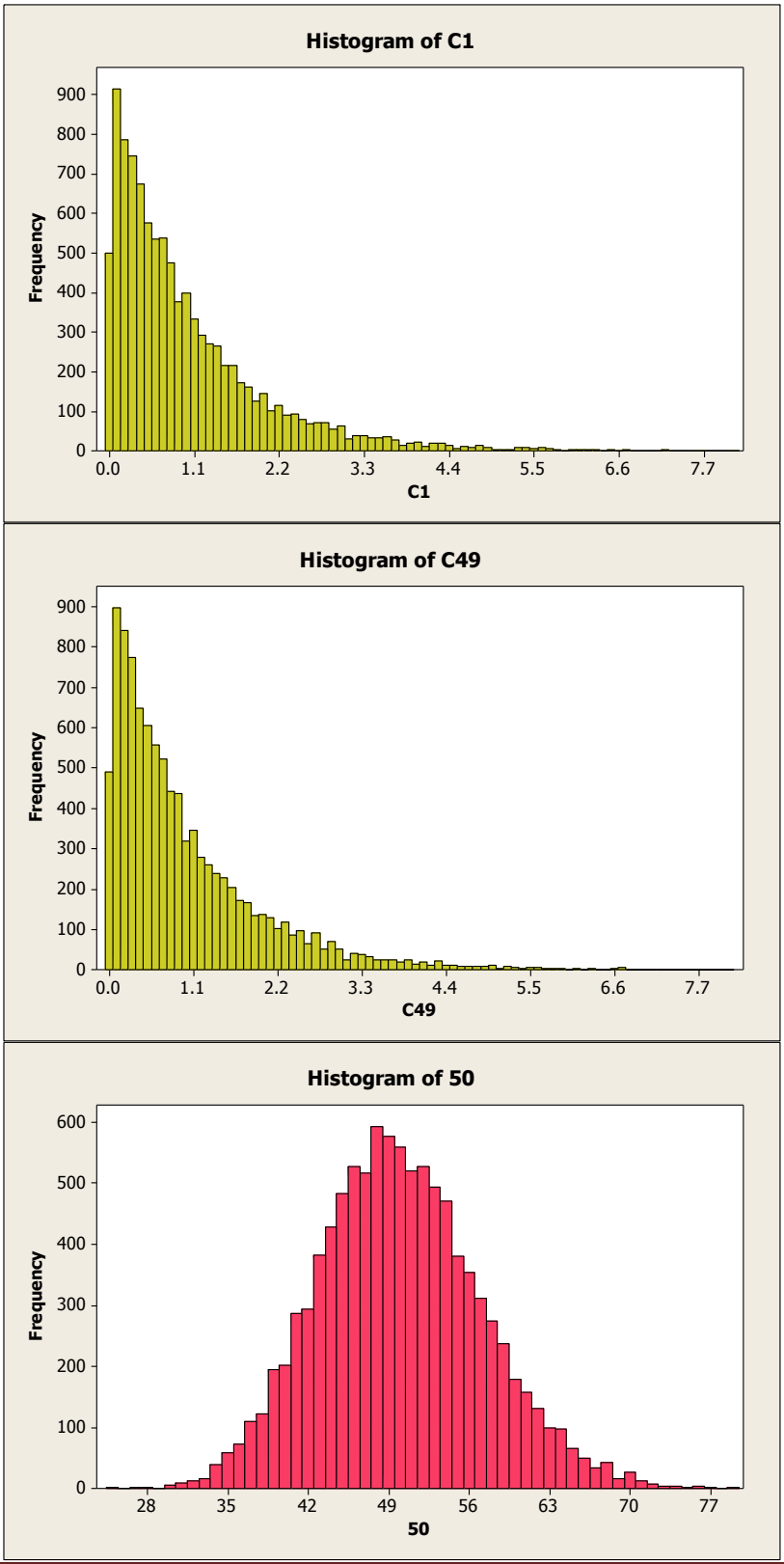
Current Worksheet: Worksheet 1 Editable

Başlat | Milliyet İntern... | 10 bin kişi ile... | Probability_S... | Gelen Kutusu... | Minitab - Unti... | 16:06

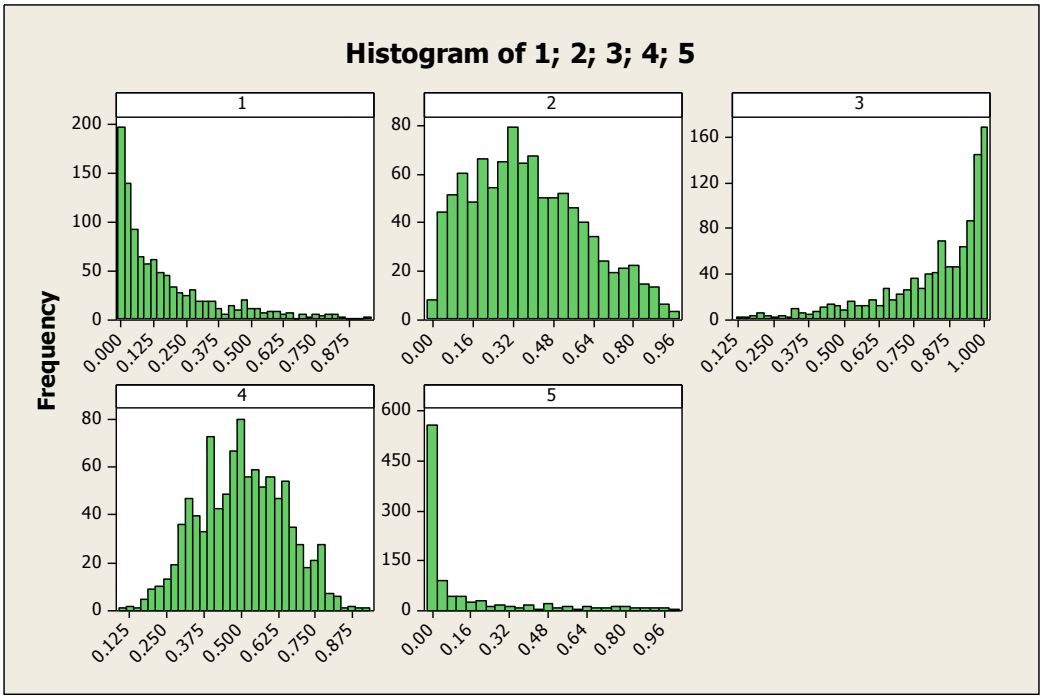


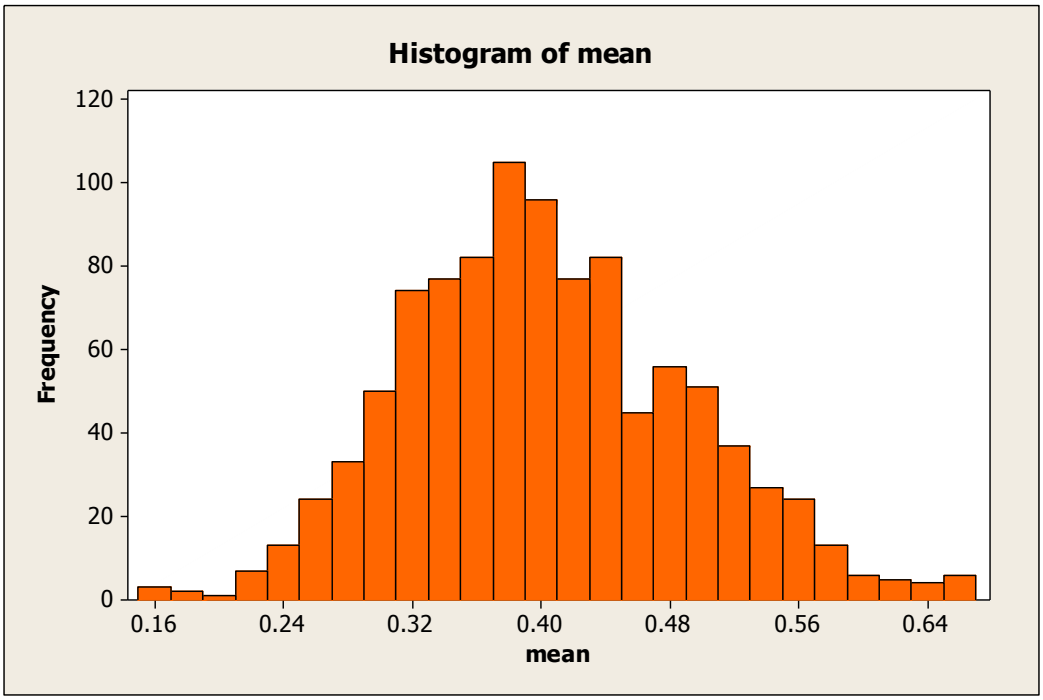


```
MTB > random 10000 c1-c50;
SUBC> expo 1.
MTB > rsum c1-c50 c51
MTB >
```



Different distributions





- **If the sampled population is normal, then the sampling distribution of $\bar{x}$ will also be normal, no matter what sample size we choose.**
- **When the sampled population is approximately symmetric, the sampling distribution of $\bar{x}$ becomes approximately normal for relatively small values of n. (Dice Example)**
- **When the sampled population is skewed, the sample size n must be larger with n at least 30 before the sampling distribution of $\bar{x}$ becomes approximately normal.**

If n ≥30, the normal approximation will satisfactory regardless of the shape of the population.