Distributions of Sample Statistics
Lecture: 7

Hamid Raza
Assistant Professor in Economics
raza@business.aau.dk
Fib 2 - room 51

Aalborg University

Statistics - Statistik

Outline

Sampling distribution:

- The remainder of the course will develop various procedures for using statistical sample data to make inferences about statistical populations

Simple Random Sample:

- A simple random sample is chosen by a process that selects a sample of n objects from a population in such a way that each member of the population has the same probability of being selected

Sampling Distributions:

- The population mean μ, is a fixed (but unknown) number. We make inferences about the population mean by drawing a random sample from the population and computing the sample mean
- But for each sample we draw, there will be a different sample mean, and the sample mean can be treated as a random variable with a probability distribution.
- The distribution of possible sample means provides a basis for the population mean

Sampling distribution:

Sampling distribution: Example

- Assume this statistics class is our population and (for simplicity) it has 4 students.
- The age of the students is as follows:

```
age=c(18, 19, 22, 21 ) # Mikael, Tanja, Tine, Simon
```

- let's say we randomly select a sample of 2 students from the population:
- There is an equal chance that we might get any of the following combinations:

  $[18, 19], [18, 22], [18, 21], [19, 22], [19, 21], [22, 21]$
- The probability of choosing any of the above combination is 1/6

Sampling distribution:

Sampling distribution: Example

We now calculate the mean of each sample

Name	Age	Mean of sample
(Mikael, Tanja)	$[18, 19]$	18.5
(Mikael, Tine)	$[18, 22]$	20
(Mikael, Simon)	$[18, 21]$	19.5
(Tanja, Tine)	$[19, 22]$	20.5
(Tanja, Simon)	$[19, 21]$	20
(Tine, Simon)	$[22, 21]$	21.5

- Note: that the probability of each sample mean is 1/6
- The mean of all sample means is equal to the population mean, i.e., $(18.5 + 20 + 19.5 + 20 + 20.5 + 21.5)/6 = 20$

Sampling distribution:

Sampling distribution: Example

Two of the samples have the same mean = 20. This means the probability of obtaining a sample mean of 20 is 2/6

We now calculate the probability of sample means:

Mean	Probability
18.5	1/6
19.5	1/6
20.5	1/6
20	2/6
21.5	1/6

- The above table is the sampling distribution for the various sample means from the population
- We can also plot the probability function as will be shown

Sampling distribution:

Sampling distribution: Example in R

We can use R to directly calculate sampling distribution of means and plot the probability function

```
age=c(18, 19, 22, 21 ) # Mikael, Tanja, Tine, Simon
```

We now randomly select two samples from the population:

```
samples <- combn(age, 2) # this calculates all possible combinations when we randomly select 2 samples
```

We now calculate the mean of each sample:

```
samp_mean <- colMeans(samples)
```

We can calculate the frequency table:

```
x=table(samp_mean);       x


## samp_mean
## 18.5 19.5   20 20.5 21.5
##    1    1    2    1    1
```

We can calculate the relevant probabilites of the sample means:

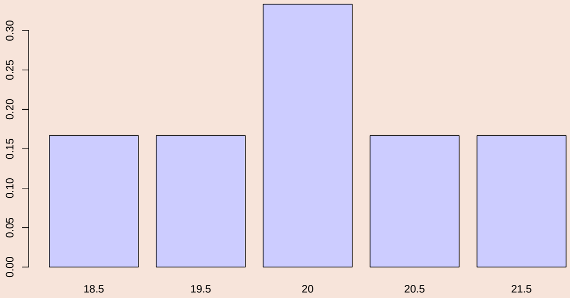
```
prop.table(x)


## samp_mean
##   18.5   19.5     20   20.5   21.5
## 0.1667 0.1667 0.3333 0.1667 0.1667
```

Sampling distribution:

Sampling distribution: Example in R

We can also plot the probability function:

```
barplot(prop.table(x), col = rgb(0.8, 0.8, 1))
```



Sampling distribution:

Some experiments in R:

Assume the population consists of the following:

```
set.seed(312); y=sample(1:30, 20)
```

Assume we perform two experiments:

- choose n=4 for each sample and see the distribution of samples means
- choose n=10 for each sample and see the distribution of samples means

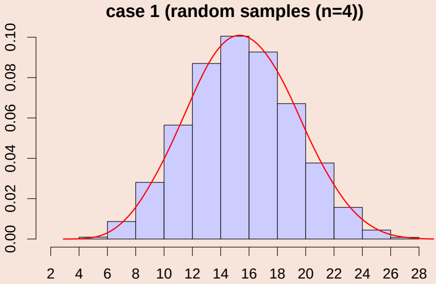
```
samp_1=combn(y, 4);    samp_1_m <- colMeans(samp_1)
samp_2=combn(y, 10);    samp_2_m <- colMeans(samp_2)
```

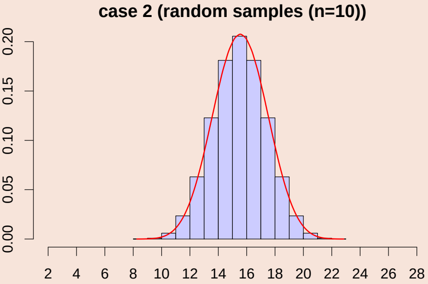
We now plot the distribution of our two experiments

Sampling distribution:

Some experiments in R:

```
hist(samp_1_m, col = rgb(0.8, 0.8, 1), xlim=c(2,28), xaxt='n', main="case 1 (random samples (n=4))", prob=T)
axis(side = 1, at = seq(2, 28, by = 2)); lines(density(samp_1_m), lwd=2, col="red")
hist( samp_2_m, col = rgb(0.8, 0.8, 1), xlim=c(2,28), xaxt='n', main="case 2 (random samples (n=10))", prob=T)
axis(side = 1, at = seq(2, 28, by = 2)); lines(density(samp_2_m), lwd=2, col="red")
```





Note: when the sample size increases, the distribution of the sample means becomes more concentrated around the population mean, i.e., it becomes more normally distributed and closer to the true mean

Sampling distribution:

Mean of sampling dist:

Note that the mean of the sample-means is equal to the poupulation mean:

```
mean(y);          mean(samp_1_m);        mean(samp_2_m)


## [1] 15.55
## [1] 15.55
## [1] 15.55
```

The idea that the mean of the sampling distribution of the sample means is the population mean can be mathematically expressed as follows:

$$E[\bar{X}] = E\Big[\frac{1}{n}(X_1 + X_2 + ... + X_n)\Big] = n\frac{\mu}{n} = \mu$$

So why do we care?

- In most applied statistical work, the populations are very large, and it is not practical or rational to construct the distribution of all possible samples of a given sample size like we did earlier
- But by using what we have learned about random variables, we can show that the sampling distributions for samples from all populations have characteristics similar to those shown above

Sampling distribution:

Variance of the distribution of sample means:

We now determine the variance of the distribution of sample means:

$$Var[\bar{X}] = Var\Big[\frac{1}{n}(X_1 + X_2 + ... + X_n)\Big]$$

$$Var[\bar{X}] = \left(\frac{1}{n}\right)^2 Var(X_1 + X_2 + ... + X_n)$$

Since, we have independent samples - $Var(X+Y) = Var(X) + Var(Y)$ - when X and Y are independent:

$$Var[\bar{X}] = \left(\frac{1}{n}\right)^2 \Big[Var(X_1) + Var(X_2) + ... + Var(X_n)\Big]$$

$$Var[\bar{X}] = \left(\frac{1}{n}\right)^2 \Big[\sigma^2_{X_1} + \sigma^2_{X_2} + ... + \sigma^2_{X_n}\Big]$$

$$Var[\bar{X}] = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

The variance of the sampling distribution of X decreases as the sample size n increases. In other words, larger sample sizes result in more concentrated sampling distributions (see the distribution earlier.

Standard deviation is simply the square root of variance

Outline

Central Limit Theorem:

Earlier, we learned that the sample mean $\bar{X}$ for a random sample of size n drawn from a population with a normal distribution with mean μ and variance σ^2, is also normally distributed with mean μ and variance σ^2/n

Central Limit Theorem:

- The central limit theorem shows that the mean of a random sample, drawn from a population **with any probability distribution**, will be approximately:
  - normally distributed with mean μ and variance σ^2/n, given a large-enough sample size
- Let $X_1, X_2, ..., X_n$ be a set of n independent random variables having identical distributions with mean μ, variance σ^2/n, and $\bar{X}$ as the mean of these random variables
  - As n becomes large, the central limit theorem states that the distribution of

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

approaches a standard normal distribution

Central Limit Theorem:

A related and important result is the law of large numbers

Law of large numbers:

The law of large numbers, which concludes that given a random sample of size n from a population, the sample mean will approach the population mean as the sample size n becomes large, regardless of the underlying probability distribution

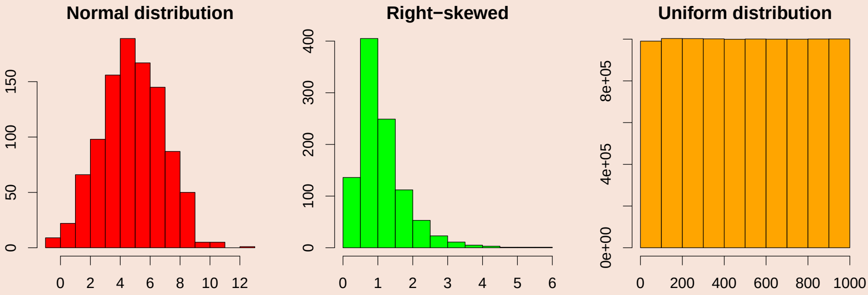
- The standard normal distribution can be used to obtain probability values for many observed sample means

Central Limit Theorem in R:

Assume, we have three different population dist. as follows::

```
set.seed(213)
pop_1 <- rnorm(1000, mean=4.9, sd = 2.1);      pop_2 <- rf(1000, 10, 20)
pop_3 <- runif(10000000, 1, 1000);
```

```
hist(pop_1, col="red", main = "Normal distribution")
hist(pop_2, col="green", main = "Right-skewed")
hist(pop_3, breaks=10, col="orange", main = "Uniform distribution")
```

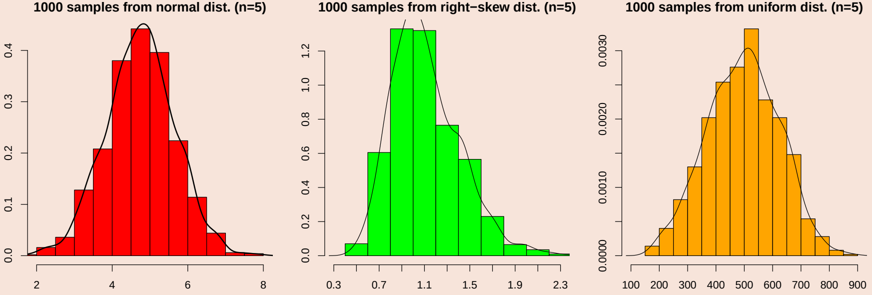


Central Limit Theorem in R:

Next, we obtained 1,000 random samples from the above three distributions using sample sizes n = 5:

```
x_samp5 = replicate(1000, mean(sample(pop_1, 5, replace = T)))
y_samp5 = replicate(1000, mean(sample(pop_2, 5, replace = T)))
z_samp5 = replicate(1000, mean(sample(pop_3, 5, replace = T)))
```

```
hist(x_samp5, prob=T, col="red", main="1000 samples from normal dist. (n=5)"); lines(density(x_samp5))
hist(y_samp5, prob=T, col="green", main="1000 samples from right-skew dist. (n=5)"); lines(density(y_samp5))
hist(z_samp5, prob=T, col="orange", main="1000 samples from uniform dist. (n=5)"); lines(density(z_samp5))
```

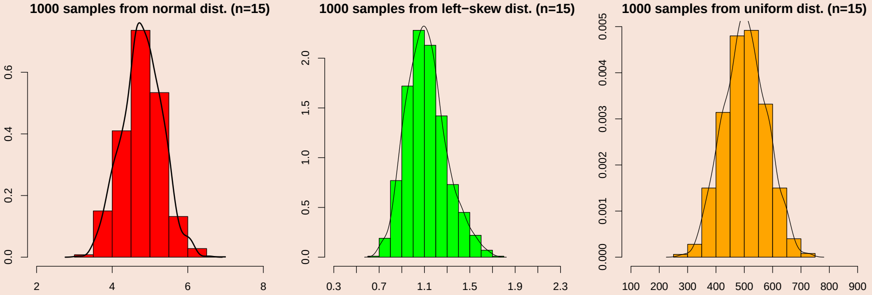


Central Limit Theorem in R:

Next, we obtained 1,000 random samples from the above three distributions using sample sizes n = 15:

```
x_samp15 = replicate(1000, mean(sample(pop_1, 15, replace = T)))
y_samp15 = replicate(1000, mean(sample(pop_2, 15, replace = T)))
z_samp15 = replicate(1000, mean(sample(pop_3, 15, replace = T)))
```

```
hist(x_samp15, prob=T, col="red", main="1000 samples from normal dist. (n=15)"); lines(density(x_samp15))
hist(y_samp15, prob=T, col="green", main="1000 samples from right-skew dist. (n=15)"); lines(density(y_samp15))
hist(z_samp15, prob=T, col="orange", main="1000 samples from uniform dist. (n=15)"); lines(density(z_samp15))
```



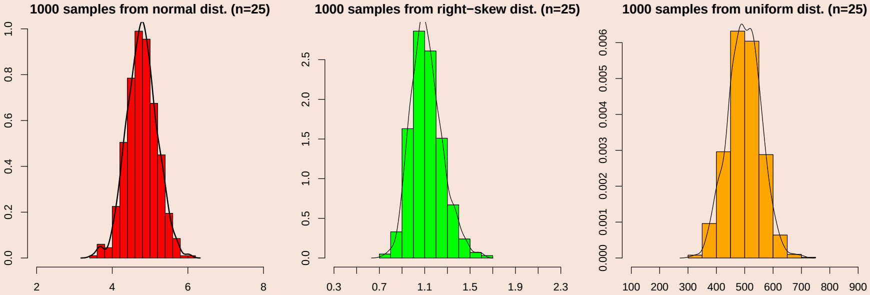
Note: The dist. approach a standard normal dist. in all cases when n is increased

Central Limit Theorem in R:

Next, we obtained 1,000 random samples from the above three distributions using sample sizes n = 25:

```
x_samp25 = replicate(1000, mean(sample(pop_1, 25, replace = T)))
y_samp25 = replicate(1000, mean(sample(pop_2, 25, replace = T)))
z_samp25 = replicate(1000, mean(sample(pop_3, 25, replace = T)))
```

```
hist(x_samp25, prob=T, col="red", main="1000 samples from normal dist. (n=25)"); lines(density(x_samp25))
hist(y_samp25, prob=T, col="green", main="1000 samples from right-skew dist. (n=25)"); lines(density(y_samp25))
hist(z_samp25, prob=T, col="orange", main ="1000 samples from uniform dist. (n=25)"); lines(density(z_samp25))
```



Note: Even when the distribution of parent population is highly skewed, the sampling distribution of sample means closely approximates a normal distribution when n increases

Acceptance Interval:

- An acceptance interval is an interval within which a sample mean has a high probability of occurring, given that we know the population mean and variance
- If the sample mean is within that interval, then we can accept the conclusion that the random sample came from the population with the known population mean and variance.
- Assuming that we know the population mean μ and variance σ^2, then we can construct a symmetric acceptance interval:

$$\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \quad \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$\pm z_{\alpha/2}$ represents the critical value, and $1 - \alpha$ represents the confidence level (CL)

Acceptance Interval:

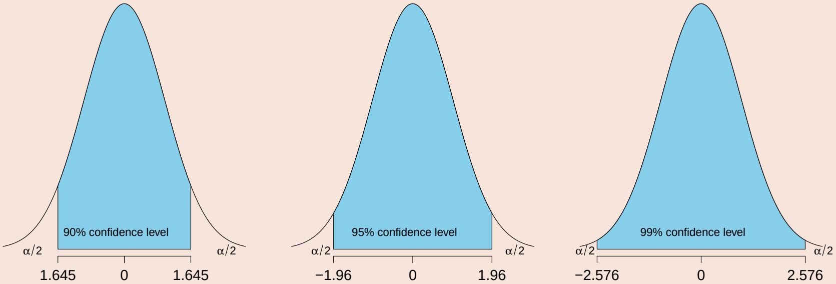
$CL = 1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
90	0.1	0.05	1.645
95	0.05	0.025	1.96
99	0.01	0.005	2.576

```r
qnorm(1-0.05, mean = 0, sd=1);    qnorm(1-0.025, mean = 0, sd=1);    qnorm(1-0.005, mean = 0, sd=1)
```

```
## [1] 1.645
## [1] 1.96
## [1] 2.576
```

- $z_{0.05} = \pm 1.645$ refers to the area equal to ± 1.645 standard deviations from the mean of a normal dist. (and $z_{0.05} \pm 1.645$ covers 90% of the distribution around the mean - (which means our confidence level is 90%)
- $z_{0.025} = \pm 1.96$ refers to the area equal to ± 1.96 standard deviations from the mean of a normal dist. (and $z_{0.025} \pm 1.96$ covers 95% of the distribution around the mean - (which means 95% CL)
- $z_{0.005} = \pm 2.576$ refers to the area equal to ± 2.576 standard deviations from the mean of a normal dist. (and $z_{0.005} \pm 2.576$ covers 99% of the distribution around the mean - (which means 99% CL)

Acceptance Interval:



What is $\alpha/2$ in the above three charts?
what is $z_{\alpha/2}$ in the above three charts?

Outline

Distribution of sample proportion:

- We consider a given characteristic (e.g. smoker/non-smoker) and note 1 if an individual has this characteristic and 0 otherwise.
- The (unknown) proportion of ones in the population is denoted P. We have a sample of 0 and 1 values. The number of ones (successes) is:

$$X = X_1 + X_2 + .. + X_n$$

- X is binomially distributed: Recall:
  - The mean of binomian dist. is: $E[X] = nP$
  - The variance of binomial dist. is: $Var[X] = nP(1 - P)$
- The sample proportion is:

$$\hat{p} = X/n = (X_1 + X_2 + ... + X_n)/n$$

- Due to the CLT, when n increases the sample proportion ($\hat{P}$) approximately follows a normal distribution
- **Rule of thumb: the approximation is good if** $np(1 - p) > 5$
- The sample proportion has mean $E(\hat{p}) = nP/n = P$ and variance $Var(\hat{p}) = nP(1 - P)/n^2 = P(1 - P)/n$

Outline

Sample Variance:

- Recall the sample variance is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- This is a random quantity which gives a new value for every sample from the population
- Mathematical theory tells us that the mean of sample variances is equal to the variance of population:

$$E(s^2) = \sigma^2$$

- If we assume that the underlying population distribution is normal, then it can be shown that the sample variance is:

$$Var(s^2) = \frac{2\sigma^4}{n-1}$$

Chi-Square Distribution:

- If we can assume that the underlying population distribution is normal, then it can be shown that the sample variance and the population variance are related through a probability distribution known as the **chi-square distribution**
- Given a random sample of n observations from a normally distributed population whose population variance is σ^2 and whose resulting sample variance is s^2, it can be shown that:

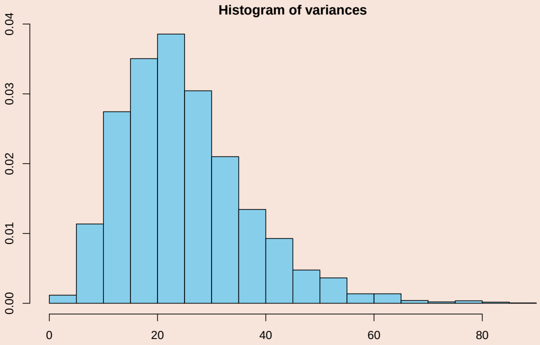
$$\chi^2_{(n-1)} = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$$

  has a distribution known as the **chi-square (χ^2) distribution** with $n-1$ degrees of freedom
- The chi-square family of distributions is used in applied statistical analysis because it provides a link between the sample and the population variances

Chi-Square Distribution (cont):

- In R, we can show the distribution of sample variance s^2 in a population of 100000 individuals:
  - Population mean 50 and population variance 25 ($\sigma = 5$)
  - The sample size is 10 (we draw a new random sample 5000 times)

```
pop <- rnorm (100000 , mean = 50 , sd = 5 )
variances <- replicate (5000 , var (sample (pop, 10 )))
hist (variances, prob=TRUE, col="skyblue" )
```

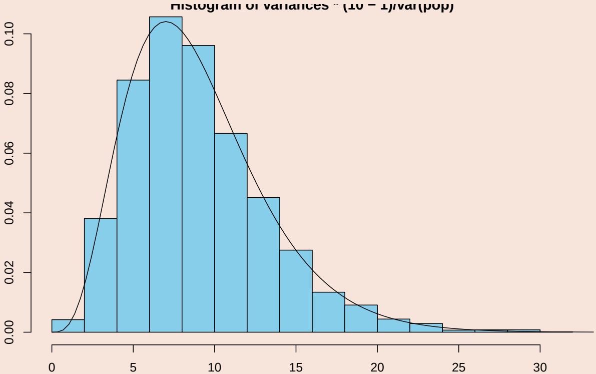


- The measured variance should be around the true value of 25 with some uncertainty. We notice that the variation around the mean is skewed to the right

Chi-Square Distribution (cont):

Let us look at the distribution of $s^2(n-1)/\sigma^2$ in R:

```
hist(variances*(10-1)/var(pop), prob=TRUE, col="skyblue")
curve(dchisq(x, df = 9), from = 0, to = 35, add = TRUE)
```



The above is chi-square (χ^2) distribution with $n-1$ degrees of freedom (df):

(χ^2) is characterised by degrees of freedom, i.e., many degrees of freedom, the values are larger and vice versa

(χ^2) is always positive and right skewed

Example:

- Let the population have variance $\sigma^2 = 25$ and mean $\mu = 50$ as before
- What is the probability that a sample of 10 people will have a sample variance above 40?

$$P(s^2 > 40) = P[s^2(n-1)/\sigma^2 > 40(9)/25]$$

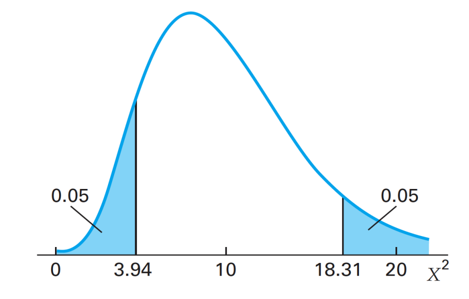
$$P(s^2 > 40) = P[\chi^2_9 > 14.44] = 1 - P[\chi^2_9 < 14.44]$$

```r
1 - pchisq(14.4, df = 9)
```

```
## [1] 0.1088
```

Chi-Square Distribution (cont)

- Table for the distribution of the chi-square random variable is available in standard statistics text books
- Table 7 in the appendix (Newbold et al) the degrees of freedom are noted in the left column and the critical values of K for various probability levels are indicated in the other columns.
- For example, for 10 degrees of freedom the lower interval is 3.940 and for the upper 0.05 interval the value of K is 18.307.




```
qchisq(0.05, df = 10);                    qchisq(0.95, df = 10)

## [1] 3.94
## [1] 18.31
```

Outline

Exercise

- Create a population, which is normally distributed with the mean and standard deviation of your choice (as I have done on slide no. 28)
- Draw a new random sample 1000 times (the sample size should be 30, i.e., n=30) and calculate sample variances
- Prove that $E(s^2) = \sigma^2$ and $Var(s^2) = Var(s^2) = \frac{2\sigma^4}{n-1}$ where s^2 refers to the sample variances, and σ^2 refers to the population variance
- Plot the dist. of $s^2(n-1)/\sigma^2$. Does it look like a chi-square distribution?
- Calculate the degrees of freedom in this example
- For the degrees of freedom that you have calculated, what is the chi-square value at the lower 0.01 interval and the upper 0.05 interval.
- What is the probability that a sample (out of 1000) will have a variance below 35?
- What is the probability that a sample (out of 1000) will have a variance above 15?
- Repeat question no. 2 but this time draw a new random sample 1000 times (the sample size should be 5, i.e., n=5).
- Plot the dist. of $s^2(n-1)/\sigma^2$. Does it look like a chi-square distribution? Comment on how the distribution has changed when we decreased the degrees of freedom.