## Sampling distributions

Let's draw all possible samples of size $n$ from a given population of size $N$. Then consider computing a statistic; the mean or a proportion or the standard deviation for each sample.

The probability distribution of this statistic is called a **sampling distribution**.

**Variability of a Sampling Distribution**

The variability of a sampling distribution is measured by its variance (or by its std. deviation).

This variability will depend on;

- $N$ : The number of observations in the population.
- $n$ : The number of observations in the sample.
- The method used to select the samples at random.

***Note:*** If $N$ is much larger than $n$, then$n/N$is fairly small and the sampling distribution has roughly the same sampling error, irrespective of whether sampling is done with or without replacement.

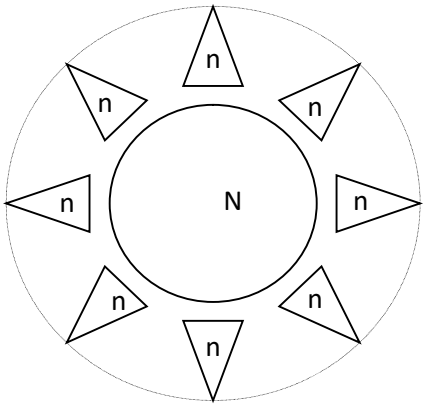
If sampling is done without replacement and the sample represents a significant fraction (say, 1/10) of the population size, the sampling error will be clearly smaller.

**The Central Limit Theorem**

The **Central Limit Theorem** (CLT) states that the probability distribution of any statistic (or the sampling distribution for any statistic) will be normal or nearly normal, if the sample size is "large enough". Thus the CLT permits approximate calculations for a variety of distributions.
Many statisticians say that a **sample size of 30 is "large enough"** as a rule of thumb.
These are some other instances in which the sample size can be considered as large enough.

- The population distribution is normal.
- The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
- The sample size is greater than 40, without outliers.

## 1. Sampling Distribution of the Mean




Let us take $\bar{x}$ as the mean of a sample of size $n$. Suppose there are$m$ number of such samples drawn from this large population.

If you take the average of the sample means by $\frac{\sum_{i=1}^{m} \bar{x}}{m}$; $\frac{\sum_{i=1}^{m} \bar{x}}{m} = \boldsymbol{\mu_{\bar{x}} = \mu}$ **(popl$^n$ mean)**

And, the standard error of the sampling distribution

$\sqrt{\sigma^2{}_{\bar{x}}} = \sigma_{\bar{x}} = \sqrt{\sigma^2(1/n - 1/N)} = \sqrt{\sigma^2(1/n)}$ as N→∝

Thus, $\boldsymbol{\sigma_{\bar{x}} = \sigma/\sqrt{n}}$

Therefore, we can specify the sampling distribution of the mean $\bar{x} \sim N(\mu_{\bar{x}}, \sigma^2{}_{\bar{x}})$as $\boldsymbol{\bar{x} \sim N(\mu\ , \sigma^2/n)}$; whenever two conditions are met:

- The population is normally distributed, or the sample size is sufficiently large.
- The population standard deviation σ is known.

## 2. Sampling Distribution of the Proportion

Let the probability of getting a success is P; and the probability of a failure is Q in a population. From this population of size $N$, suppose that we draw all possible samples of size $n$. And finally, within each sample, suppose that we determine the proportion of successes $p$ and failures $q$. In this way, we create a sampling distribution of the proportion.

Let us take p as proportion of successes in a sample of size $n$.

Suppose there are$m$ number of such samples drawn from this large population.

If you take the mean of the sample proportions by

$\frac{\sum_{i=1}^{m} p}{m}$; $\frac{\sum_{i=1}^{m} p}{m} = \boldsymbol{\mu_p = P}$ **(Population proportion of success)**

And, the standard error of the sampling distribution

$\sqrt{\sigma^2{}_p} = \sigma_p = \sqrt{\sigma^2(1/n - 1/N)} = \sqrt{PQ(1/n - 1/N)} = \sqrt{PQ/n}$as N→∝

Thus, $\boldsymbol{\sigma_p = \sqrt{PQ/n}}$

Therefore, we can specify the sampling distribution of the proportion $p \sim N(\mu_p, \sigma^2{}_p)$ as

$\boldsymbol{p \sim N(P, PQ/n)}$; whenever the sample size is sufficiently large and the population probability of success (P) is known.

**Example:**

1. Suppose that a biased coin has probability p=0.4of heads. In 1000 tosses, what is the probability that the number of heads exceeds 410?

1. Find the probability that of the next 120 births, no more than 40% will be boys. Assume equal probabilities for the births of boys and girls. Assume also that the number of births in the population (N) is very large, essentially infinite.

**Exercises:**

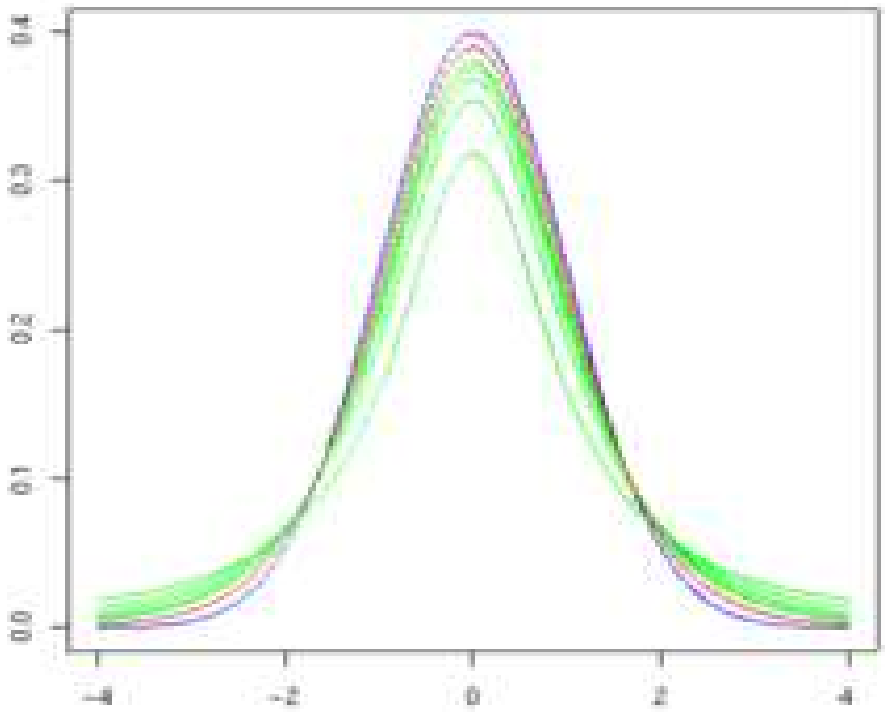
1. A true-false examination has 48 questions. Jane has probability 3/4 of answering a question correctly. Ama just guesses on each question. A passing score is 30 or more correct answers. Compare the probability that Jane passes the exam with the probability that Ama passes it. Jane's score has distribution B(48,0.75), so the probability that Jane's score is 30 or more is 1-P(X<=29) = 0.9627. In case your calculator doesn't give an answer, you will have to use a normal approximation to the Binomial distribution (based on the Central Limit Theorem)

2. A restaurant feeds 400 customers per day. On the average 20 percent of the customers order apple pie.
   (a) Give a range for the number of pieces of apple pie ordered on a given day such that you can be 95 percent sure that the actual number will fall in this range.
   (b) How many customers must the restaurant have, on the average, to be at least 95 percent sure that the number of customers ordering pie on that day falls in the 19 to 21 percent range?

3. A rookie is brought to a baseball club on the assumption that he will have a 0.3 batting average. (Batting average is the ratio of the number of hits to the number of times at bat.) In the first year, he comes to bat 300 times and his batting average is 0.267. Assume that his at bats can be considered Bernoulli trials with probability 0.3 for success. Could such a low average be considered just bad luck or should he be sent back to the minor leagues?

### 3) Student's t Distribution

A particular form of the t distribution is determined by its **degrees of freedom**. The "degrees of freedom" refers to the number of independent observations in a set of data.




In general, the degrees of freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself (which, in sample variance, is one, since the sample mean is the only intermediate step).

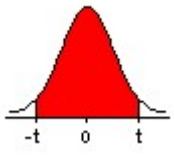
Lane, David M.."Degrees of Freedom".*HyperStatOnline*.StatisticsSolutions.http://davidmlane.com/hyperstat/A42408.html. Retrieved 2008-08-21.

Suppose we have a simple random sample of size $n$ drawn from a Normal population with mean $\mu$ and standard deviation $\sigma$. Let $\bar{x}$denote the sample mean and $s$, the sample standard deviation.

**Then the quantity $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a *t* distribution with *n*-1 degrees of freedom.**

The $t$ score produced by this transformation can be associated with a unique cumulative probability. This cumulative probability represents the likelihood of finding a sample mean less than or equal to $\bar{x}$, given a random sample of size *n*.

The notation $t_\alpha$ represents the t-score that has a cumulative probability of (1 - α).




Example: $t_{0.05}$ = 2.92, then $t_{0.95}$ = -2.92 for *df*=3

**Properties of the t Distribution**

- The **mean** of the distribution is equal to **0** .
- The **variance** is equal to ***v* / ( *v* - 2 )**, where *v* is the degrees of freedom and $v \geq 2$.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the standard normal distribution.

**When to use the t Distribution**

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). i.e. when the population size is large but the sample sizes are small and the standard deviation of the population is unknown t-Distribution can be applied.

*Example*: $\boldsymbol{t = (p - P)/\sqrt{(PQ\ /\ n)}}$ **has a *t* distribution with *n*-1 degrees of freedom**

**When not to use the t-distribution**

The t distribution should *not* be used with small samples from **populations that are not approximately normal.**

**Example:**

1. A random sample of 12 observations from a normal population with mean 48 produced the following
   Estimates: $\bar{x} = 47.1$and $s^2 = 4.7$. Find the probability of getting a sample of the same size with its mean less than or equal to the population mean.

2. The MD of Orrange light bulb manufactures claims that an average of their light bulbs lasts 300 days. An investigator randomly selects 15 bulbs for testing and those bulbs last an average of 290 days, with a standard deviation of 50 days. Assuming MD's claim as true, determine the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

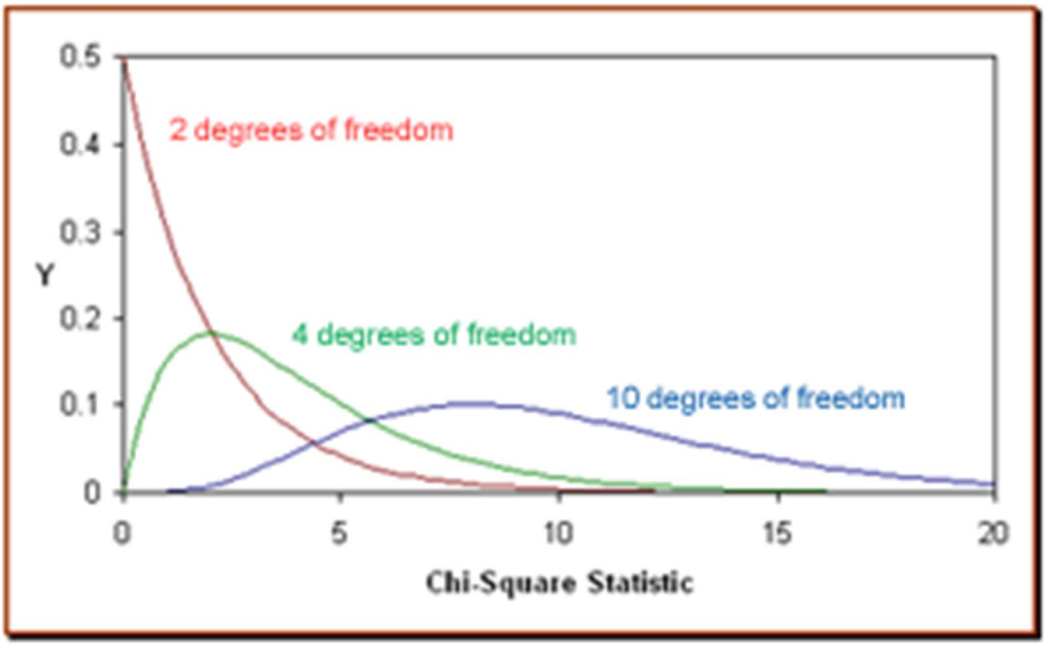
## 4. Chi-square Distribution

**The chi-square statistic can be calculated from a sample of size $n$drawn from a population, which is normal**, using the following equation:

$$\chi^2 = {(n-1)s^2}/{\sigma^2}$$

When sampling is done for an infinite number of times, and by calculating the chi-square statistic for each sample, the sampling distribution for the chi-square statistic can be obtained. It is then called the chi-square distribution.

The **chi-square distribution** also depends on the degrees of freedom; $(n-1)$.



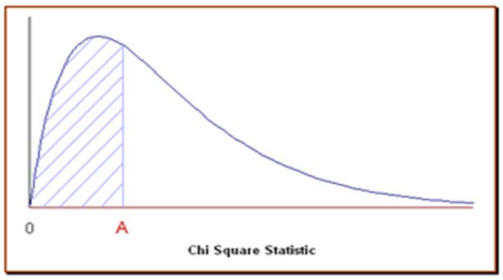

**Properties of the chi-square distribution:**

- The mean of the distribution is equal to the number of degrees of freedom: $\boldsymbol{\mu = \nu}$.
- The variance is equal to two times the number of degrees of freedom: $\boldsymbol{\sigma^2 = 2\nu}$
- When the degrees of freedom are greater than or equal to 2, the maximum value for $f(x), the\ pdf\ of\ chi-square$ occurs.

- As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

**Cumulative Probability of the Chi-Square Distribution**

The chi-square distribution is constructed so that the total area under the curve is equal to 1. The probability that the value of a chi-square statistic will fall between 0 and $A$; $P(\chi^2_{n-1} \leq A)$ is illustrated by the following diagram.




Using the following Chi-Square Distribution table, one can find thecritical$\chi^2$value, when the probability of exceeding the critical value is given.

| Degrees of Freedom | Probability | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.95 | 0.90 | 0.80 | 0.70 | 0.50 | 0.30 | 0.20 | 0.10 | 0.05 | 0.01 | 0.001 |
| 1 | 0.004 | 0.02 | 0.06 | 0.15 | 0.46 | 1.07 | 1.64 | 2.71 | 3.84 | 6.64 | 10.83 |
| 2 | 0.10 | 0.21 | 0.45 | 0.71 | 1.39 | 2.41 | 3.22 | 4.60 | 5.99 | 9.21 | 13.82 |
| 3 | 0.35 | 0.58 | 1.01 | 1.42 | 2.37 | 3.66 | 4.64 | 6.25 | 7.82 | 11.34 | 16.27 |
| 4 | 0.71 | 1.06 | 1.65 | 2.20 | 3.36 | 4.88 | 5.99 | 7.78 | 9.49 | 13.28 | 18.47 |
| 5 | 1.14 | 1.61 | 2.34 | 3.00 | 4.35 | 6.06 | 7.29 | 9.24 | 11.07 | 15.09 | 20.52 |
| 6 | 1.63 | 2.20 | 3.07 | 3.83 | 5.35 | 7.23 | 8.56 | 10.64 | 12.59 | 16.81 | 22.46 |
| 7 | 2.17 | 2.83 | 3.82 | 4.67 | 6.35 | 8.38 | 9.80 | 12.02 | 14.07 | 18.48 | 24.32 |
| 8 | 2.73 | 3.49 | 4.59 | 5.53 | 7.34 | 9.52 | 11.03 | 13.36 | 15.51 | 20.09 | 26.12 |
| 9 | 3.32 | 4.17 | 5.38 | 6.39 | 8.34 | 10.66 | 12.24 | 14.68 | 16.92 | 21.67 | 27.88 |
| 10 | 3.94 | 4.86 | 6.18 | 7.27 | 9.34 | 11.78 | 13.44 | 15.99 | 18.31 | 23.21 | 29.59 |
| | Nonsignificant | | | | | | | | Significant | | |

**Example:** My Cell company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 5 minutes. Suppose the manufacturing department runs a quality control test. They randomly select 10 batteries. The standard deviation of the selected batteries is 6 minutes.

a) What is the chi-square statistic which represents this test?

b) What is the probability that the standard deviation of any sample of size 10 would be greater than 6 minutes?

## 5. F Distribution

The distribution of all possible values of the ***f* statistic** is called an **F distribution**, with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. The ***f* statistic**, also known as an ***f* value**, is a random variable that has an F distribution.

How to compute an ***f* statistic**:

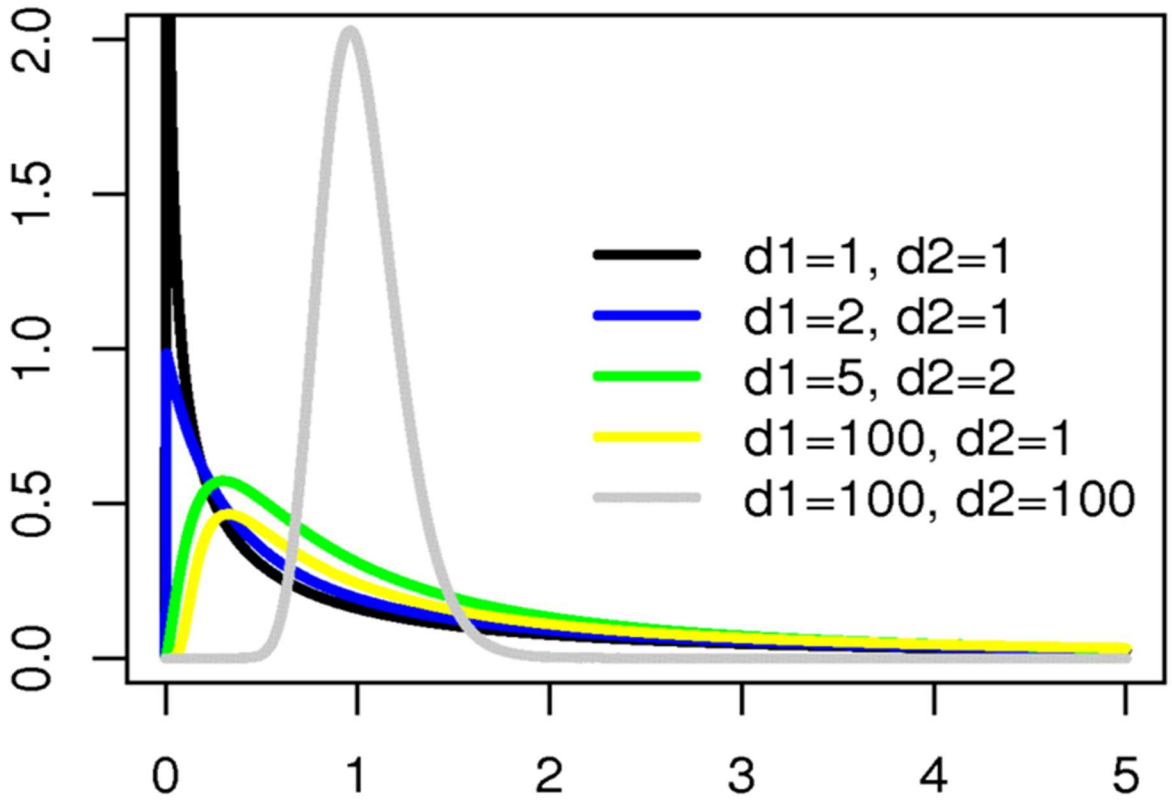
- Select a random sample of size $n_1$ from a normal population, having a standard deviation equal to $\sigma_1$.
- Select an independent random sample of size $n_2$ from a normal population, having a standard deviation equal to $\sigma_2$.
- The *f* statistic is the ratio of $s_1^2/\sigma_1^2$ and $s_2^2/\sigma_2^2$.

The following equations are commonly used in equivalent to an *f* statistic:

$$f(v_1, v_2) = [\, s_1^2/\sigma_1^2 \,] / [\, s_2^2/\sigma_2^2 \,]$$
$$f(v_1, v_2) = [\, s_1^2 . \sigma_2^2 \,] / [\, s_2^2 . \sigma_1^2 \,]$$

$$f(v_1, v_2) = [\, \chi^2_1 / v_1 \,] / [\, \chi^2_2 / v_2 \,]$$
$$f(v_1, v_2) = [\, \chi^2_1 . v_2 \,] / [\, \chi^2_2 . v_1 \,]$$

The curve of the F distribution depends on the degrees of freedom, $v_1$ and $v_2$.




**Properties of the F distribution:**

- The mean of the distribution is equal to $v_2 / ( v_2 - 2 )$ for $v_2 > 2$.
- The variance is equal to $[\, 2v_2^2( v_1 + v_2 - 2 ) \,] / [\, v_1( v_2 - 2 )^2( v_2 - 4 )]$ for $v_2 > 4$.

**Cumulative Probability of the F Distribution**

This cumulative probability represents the likelihood that the *f* statistic is less than or equal to a specified value.

F-distribution table can be used to find the value of an *f* statistic having a cumulative probability of (1 - α); represented by $f_{\alpha}$.

Thus, $f_{0.05}(v_1, v_2)$ refers to value of the f statistic having a cumulative probability of (1-0.05)= 0.95, with $v_1$and $v_2$degrees of freedom.

**Example:**

Suppose a sample of 11 of cows was selected at random from a population of them having the population standard deviation of their weight is 5 kg and the estimated sample sd is 4.5 kg. Another sample of size 7 of bulls was taken in a similar way with their population sd is 3.5 kg and sample sd is 4 kg.

a) Compute an f-statistic.

b) Determine the associated cumulative probability by finding an approximate f-value to the above answer from the f-tables available for different significance levels ($\alpha$).

c) Interpret the probability you found.

Reference for f-table: http://www.socr.ucla.edu/Applets.dir/F_Table.html

***Upgrade your knowledge by:***

- Finding the *pdf* 's of the above sampling distributions.
- Studying the patterns of *cdf* 's of the above sampling distributions.