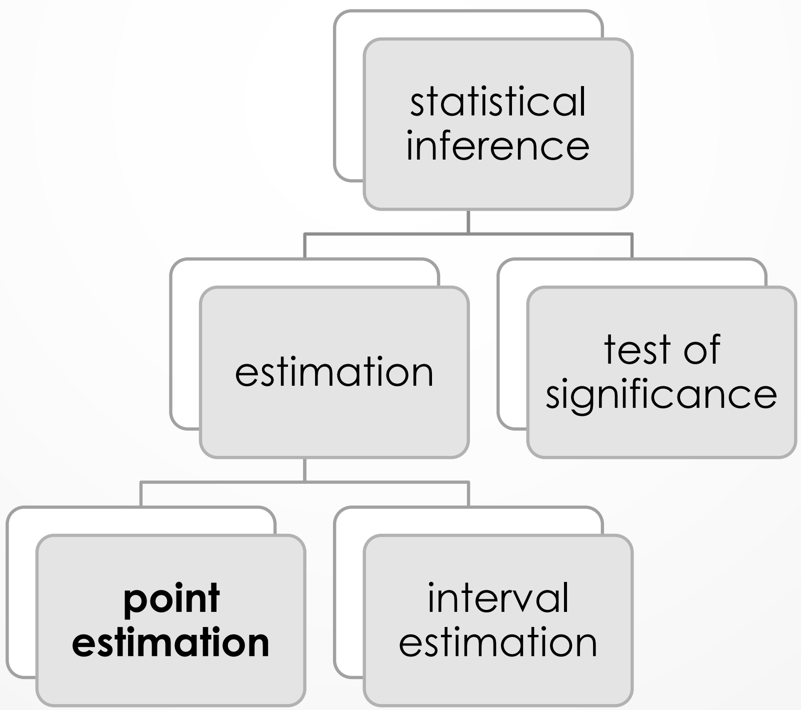
Introduction to probability, statistics and data handling

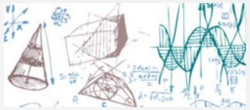
Tomasz Szumlak, Agnieszka Obłąkowska-Mucha
Faculty of Physics and Applied Computer Science

AGH UST Krakow

Statistical Inference

- The statistical inference consists in arriving at (quantitative) conclusions concerning a **population** where it is impossible or impractical to examine the entire set of observations that make up the population. Instead, we depend on a **subset** of observations - a **sample**.





Statistical Sample and Population

- Sample posses a property X (our RV); $X \rightarrow f(x, \boldsymbol{\lambda})$ (probability density function), $\boldsymbol{\lambda}$ – set of parameters of the population to be determined from the sample (e.g. $\mu, \sigma,$ etc.).
- Any function of the random variables constituting a random sample that is used for **estimation** of unknown distribution parameters $\boldsymbol{\lambda}$ is called a **statistic** $\boldsymbol{S}$:

$$S = S(X_1, X_2, \dots, X_n)$$

$$\lambda_i = E[S(X_1, X_2, \dots, X_n)] \equiv \hat{S}$$

We say: the estimated value of a statistic $\hat{S}$ is said to be estimator of the parameter λ ; the estimation is carried out on the basis of an n-element **sample**.

do we know any statistic?

Parameter estimation

The parameters of a pdf are any constants that characterize it,

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v. parameter

i.e., θ indexes a set of hypotheses.

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

$$\boxed{\widehat{\theta}(\vec{x})}$$

← estimator written with a hat

Sometimes we say 'estimator' for the function of $x_1, \ldots, x_n$;
'estimate' for the value of the estimator with a particular data set.

Statistical Sample and Population

- We start with two estimators:

  - estimator of a **mean value**
  - estimator of a **variance**

  we want to estimate μ and σ^2 of a **population** with a use of **sample**

- Later we will develop methods for the estimation of unknown parameter of a model (linear, or any other) based on samples (method of momets, method of least squares, maximum likelihood estimation)

Point estimation

- Let's think about the following: we are looking at some phenomena (took a data sample), now what we like to do is to try describe the data using a model (have we already discussed any models?)
- Using the statistics lingo we would say: we want to estimate the parameters for the hypothesised population model
- **As usual there are a lot of methods, we are going to have a look at a few of them**
- Estimators should have specific features (we will discuss it today)

BUT

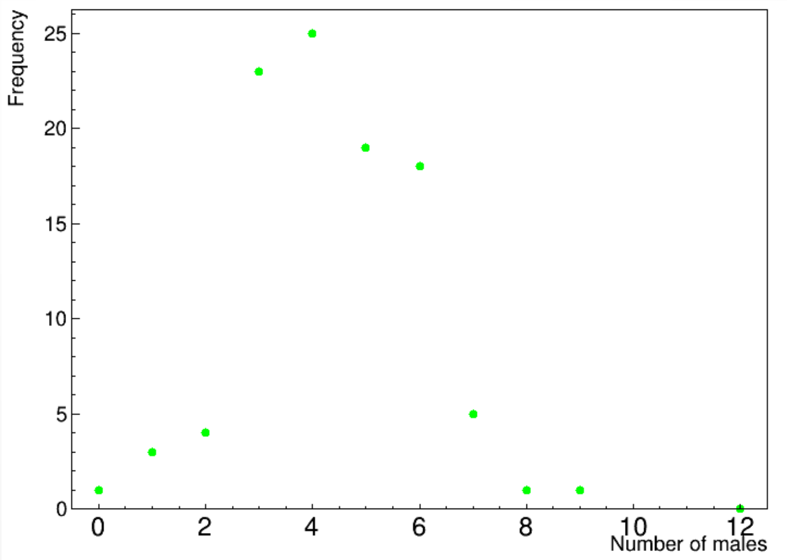
- Let's start with some **examples** first!

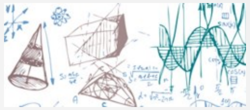
Number of males in a queue

- An experiment has been conducted in London Tube to check the number of males in each of 100 queues all of length 10. The results obtained were as follows

Counts	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	3	4	23	25	19	18	5	1	1	0

- And the plot





Number of males in a queue

- Can you tell what is the underlaying **parent distribution**?
- Well, one could prove that the **binominal** one fits quite good $\mathcal{B}(n, p)$, $n = 10$ being the length of the queue and $\boldsymbol{p}$ the proportion of males (check this on your own)
- We could estimate the p using the collected sample

$$\frac{\#males}{\#all\ passangers} = \frac{1 \cdot 0 + 3 \cdot 1 + \cdots + 1 \cdot 9 + 0 \cdot 10}{1000} = \frac{435}{1000} = \mathbf{0.435}$$

- **What would be the weak point of this assumption**?
- Can we actually come up with a generic strategy to say, the value of a parameter of interest is this and that?
- Yes! We can! We need to perform an experiment and run an analysis
- Another question would be how reliable this estimate is (but we leave it for the next lectures)

Estimators

- Consider the following: to check the water for contamination by a micro-organism a number of samples were taken, the results are summarised as follow

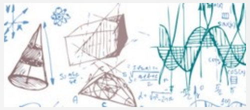
Counts	0	1	2	3	4	5	6	7	8	>9
Frequency	53	25	13	2	2	1	1	0	1	0

- One can assume that the data follow the Poisson distribution with an unknown parameter μ (each water sample is an independent observation on the same random variable!)
- For these particular data, we can estimate the μ as:

$$\bar{x} = \frac{0 \cdot 53 + 1 \cdot 25 + \cdots + 8 \cdot 1}{58 + 25 + \cdots + 1} = \frac{84}{103} = 0.816$$

$$\{X_1, X_2, \dots, X_{103}\} \rightarrow X \equiv Poisson(\mu)$$

$$\bar{X}_{(1)} = \frac{X_1 + X_2 + \cdots + X_{103}}{103} \rightarrow \quad \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$



Estimators

- Let’s set a **generic procedure** using this simple example
- First, we **pick the parameter** to be estimated
- Next, we need to **collect data** and **compute a sampling statistics** using a formula corresponding to the parameter we are interested in
- In our example that is a **sample mean**

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- This, in turn, we call an **estimator** of true parameter, in our case this would be: $\mu \rightarrow \bar{X} = \hat{\mu}$ (we use the caret symbor "^")
- Remember – the estimator is a random variable, for different sample we are going to get different value
- The estimator will follow its own distribution – **sampling distribution of the estimator**

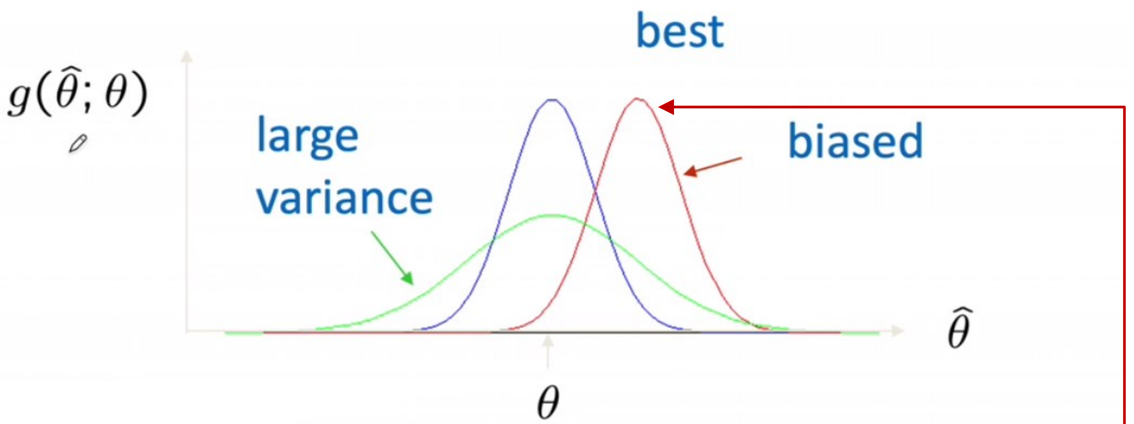
A big question

- So, we collected the data – we are going to be interested in a procedure, which basing on the observed variation gives the best value (we could also ask about the range of values) for the corresponding underlaying model parameter(s)
- Again, using the stat lingo we want to get **the best possible estimate of the value of the parameter**(s)
- That is what the **point estimation** is all about
- BTW, it may also be useful to estimate the range of „good" parameter values – that is yet another story called **estimation with confidence** – we are going to look at this next time!

Not quite like that…

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

Estimators

- Consider the following: to check the water for contamination by a micro-organism a number of samples were taken, the results are summarised as follow

Counts	0	1	2	3	4	5	6	7	8	>9
Frequency	53	25	13	2	2	1	1	0	1	0

- One can assume that the data follow the **Poisson distribution** with an unknown parameter μ (each water sample is an independent observation on the same random variable!)
- For these particular data, we can estimate the μ as:

$$\bar{x} = \frac{0 \cdot 53 + 1 \cdot 25 + \cdots + 8 \cdot 1}{58 + 25 + \cdots + 1} = \frac{84}{103} = 0.816$$

$$\{X_1, X_2, \ldots, X_{103}\} \rightarrow X \equiv Poisson(\mu)$$

$$\bar{X}_{(1)} = \frac{X_1 + X_2 + \cdots + X_{103}}{103} \rightarrow \quad \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Estimators

- Let’s set a **generic procedure** using this simple example
- First, we **pick the parameter** to be estimated
- Next, we need to **collect data** and **compute a sampling statistics** using a formula corresponding to the parameter we are interested in
- In our example that is a **sample mean**

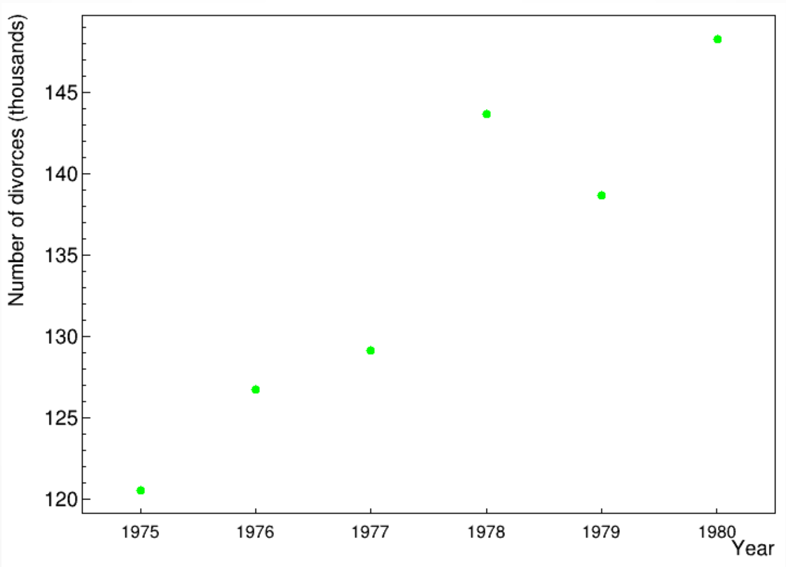
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- This, in turn, we call an **estimator** of true parameter, in our case this would be: $\mu \to \bar{X} = \hat{\mu}$ (we use the caret symbor "^")
- Remember – the estimator is a random variable, for different sample we are going to get different value
- The estimator will follow its own distribution – **sampling distribution of the estimator**

More than one way…

- Lets inspect the following data regarding the number of divorces in different years in some country in Europe

Year	1975	1976	1977	1978	1979	1980
# divorces (10^3)	120.5	126.7	129.1	143.7	138.7	148.3



- Interesting…, very tempting to fit a model right away.

More than one way…

- From the plot we could conclude, that the **true underlaying distribution** describing the data can be represented by a **linear model**
- From the data we also conclude that **the slope** of the line is positive – ok, the task is then to **estimate this slope**, α, and then we could predict the annual rate of increase of divorces
- But how do we do that? <u>It is not so obvious like the water example</u>(??)
- Consider this:
  - $\hat{\alpha}_1$- join the first and the last point
  - $\hat{\alpha}_2$ - join the mid-points P_1P_2 and P_5P_6
  - $\hat{\alpha}_3$ - join the centroid of the first triplet and the second one
- Mind you, these are all sensible options!

we will continue the discussion how to obtain the best estimators in a few weeks!

Summary so far

- A generic „algorithm” for point estimation task would be:
- **Collect the data** and understand it
- Come up with a **model**, this will specify a **parameter** or many parameters that we need to make an estimate
- For a given parameter(s) we need an **estimator(s)** (typically we will concentrate on the mean value or variance, however we also can tackle more ambitious cases – e.g., divorces)
- Work out the **estimate of the parameter** – this is a random variable and will be different for different data sets
- Finally, analyse the **sampling distribution of the estimator** to make a judgement of its usefulness
- We are looking for **unbiased** (expectation value) and **efficient** estimators (variance)

Estmator Wish List

- We are looking for the best estimator (but what does „best" mean?
- In the best of all possible worlds, we could find an estimator $\hat{\mu}$ for which $\hat{\mu} = \mu$ in all samples. But this does not exist, sometimes $\hat{\mu}$ will be too small, fort other samples too big.
- Let's write (in general): $\hat{\theta} = \theta + error\ of\ estmation.$ Therefore the best estimator $\hat{\theta}$:
  - has small estimator errors: the mean squared error RMS $E\left[\left(\hat{\theta} - \theta\right)^2\right]$ shoud be the smallest
  - should be **unbiased** $E\left[\left(\hat{\theta}\right)\right] = \theta$
  - should have small variance $VAR\left[\left(\hat{\theta}\right)\right]$

We are looking for **unbiased** (expectation value) and **efficient** estimators (variance).

Sampling distribution

- Any sample statistics is a function of R.Vs and is therefore itself a random variable – that is absolutely critical to remember!
- The probability distribution of a sample statistics is called **the sampling distribution** of this statistics (sorry for complicated circular sentences…)
  - A recipe to get such distribution would be as follow: we should draw all possible samples of size n from a population, next we should compute the statistics at hand, thus, obtaining the distribution of this statistics. We call it the sampling distribution
- It is perfectly ok to compute the mean, variance, standard deviation and other moments for the sampling distribution!
- To make it a bit more comprehensible, let’s consider the sample mean. Let $X_1, X_2, \cdots, X_n$ be independent, identically distributed RVs. The mean of the sample is another R.V. defined as follow:

$$\boldsymbol{\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{\sum_{i/1}^{i/n} X_i}{n}}$$

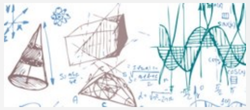
Estimator for the mean

Parameter: $\mu = E[x] = \langle x \rangle = \int_{-\infty}^{\infty} x f(x)\, dx$

Estimator: $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \overline{x}$ ('sample mean')

We find: $b = E[\widehat{\mu}] - \mu = 0$

$$V[\widehat{\mu}] = \frac{\sigma^2}{n} \qquad \left(\sigma_{\widehat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$



Sampling dist. of means

- **Theorem 1**. The mean of the sample means is a consistent etimator of μ:

$$\boxed{E[\bar{X}] = \mu_{\bar{X}} = \mu}$$

where μ is the mean of the population. So, we say, that the expected value of the sample mean is the population mean – **how interesting**!

- **Theorem 2**. If a population is infinite and the sampling is random, or if a population is finite and sampling is with replacement, then the variance of the distributions of the sample means, denoted by $\sigma_{\bar{X}}$, is:

$$\boxed{E[(\bar{X} - \mu)^2] = \sigma_{\bar{X}}^2 = \frac{1}{n}\sigma^2}$$

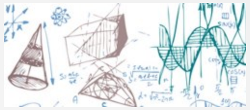
Sampling dist. of means

- **Theorem 3**. If the population is not infinite (of size N) or is the sampling is done without replacement, then the variance should be evaluated using:

$$\sigma'^2_{\bar{X}} = \frac{1}{n}\sigma^2\left(\frac{N-n}{N-1}\right), N \to \infty: \sigma'^2_{\bar{X}} \to \sigma^2_{\bar{X}}$$

- **Theorem 4**. If the population from which we draw samples is normally distributed with mean μ and variance σ^2, then the sample mean is also normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\frac{\sigma^2}{n}}$
- **Theorem 5**. Let's assume that the population from which samples are drawn has mean μ and variance σ^2. The population **may or may not be normally distributed**. The standardised variable associated with $\bar{X}$ can be written as:

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



Estimator for sample variance

- If $\{X_1, X_2, \cdots X_n\}$ denote R.Vs for a random sample of size n, the R.V. giving the variance of the sample (the sample variance) is defined as:

$$S^2 = \frac{\mathbf{1}}{\boldsymbol{n}}[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2]$$

- We already know, that $E[\bar{X}] = \mu$, is this the same for $E[S^2] = \sigma^2$?
  - A little digression – whenever the expected value of a statistics **is equal** to the corresponding **population parameter**, we call this statistics **an unbiased estimator**. Its value is then an unbiased estimate of the respective parameter
- Unfortunately, it can be proved that for the sample variance, we have:

$$E[S^2] = \mu_{S^2} = \frac{n-1}{n}\sigma^2$$

- However, an unbiased variance estimator is easy to find:

$$\boxed{\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{\mathbf{1}}{\boldsymbol{n} - \mathbf{1}}[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2]}$$

Estimator for the variance

Parameter: $\sigma^2 = V[x] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$

Estimator: $\widehat{\sigma^2} = \frac{1}{\boxed{n-1}} \sum_{i=1}^{n} (x_i - \overline{x})^2 \equiv s^2$ ('sample variance')

We find:

$b = E[\widehat{\sigma^2}] - \sigma^2 = 0$ (factor of *n*-1 makes this so)

Point estimators - summary

- Sample mean $\bar{X}$ is the point estimator of parameter μ:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1..n} X_i$$

- The unbiased estimator for variance is:

$$\hat{S}^2 = \frac{1}{n-1}\left[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right] = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

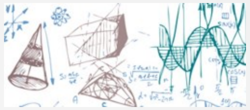
- The estimator of the correlation (X, Y) is:

$$r(X,Y) = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{yy}}}$$

$$S_{XX} = \sum (X_i - \bar{X})^2$$

$$S_{YY} = \sum (Y_i - \bar{Y})^2$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$



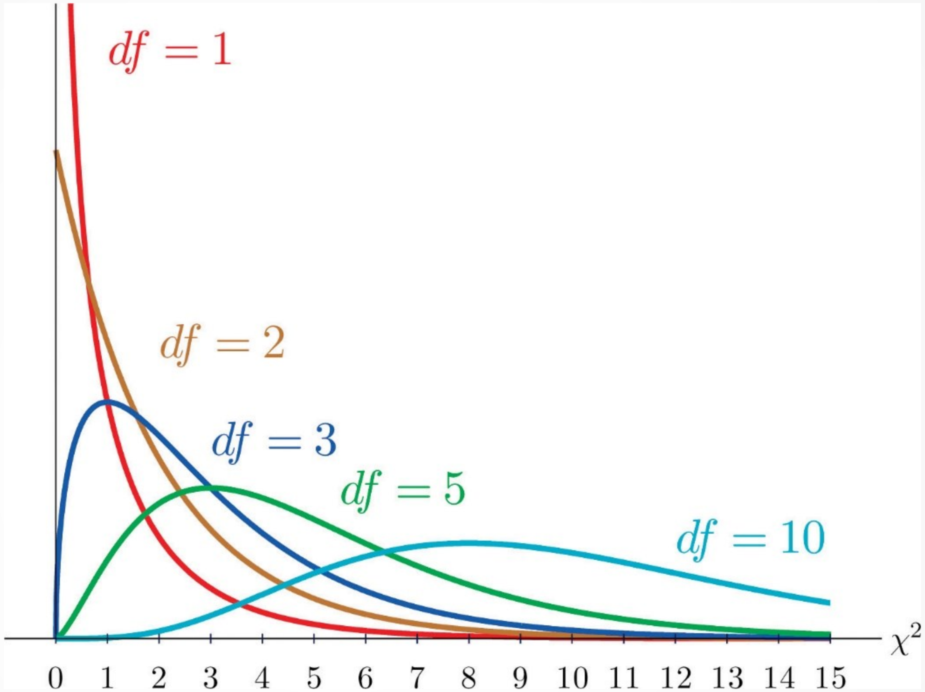
Sampling dist. of variances

- In order to create the sampling distribution of variances, we take all the possible samples of size n, that can be drawn from a population and calculate their variances
- One change is, that instead of looking directly at the distribution of the sample variance, we look at the R.V.:

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{\sigma^2}$$

- **Theorem 6**. If a random samples of size n are taken from a population having a normal distribution, than the sampling variable $\frac{nS^2}{\sigma^2}$ has a χ^2 distribution with $n-1$ degrees of freedom

χ^2 distribution



- This is another very popular distribution in Statistics!
- The mathematical formula describing it is quite complex, again we are going to use tabulated values when solving problems!