ESE

9. Data analysis and interpretation. P2

09.2 Data Interpretation.

09.2 Data Interpretation

- Statistics Basis of Statistical Tests
- Hypothesis test
- Tools for Hypothesis Tests
- Statistical errors
- Statistical power

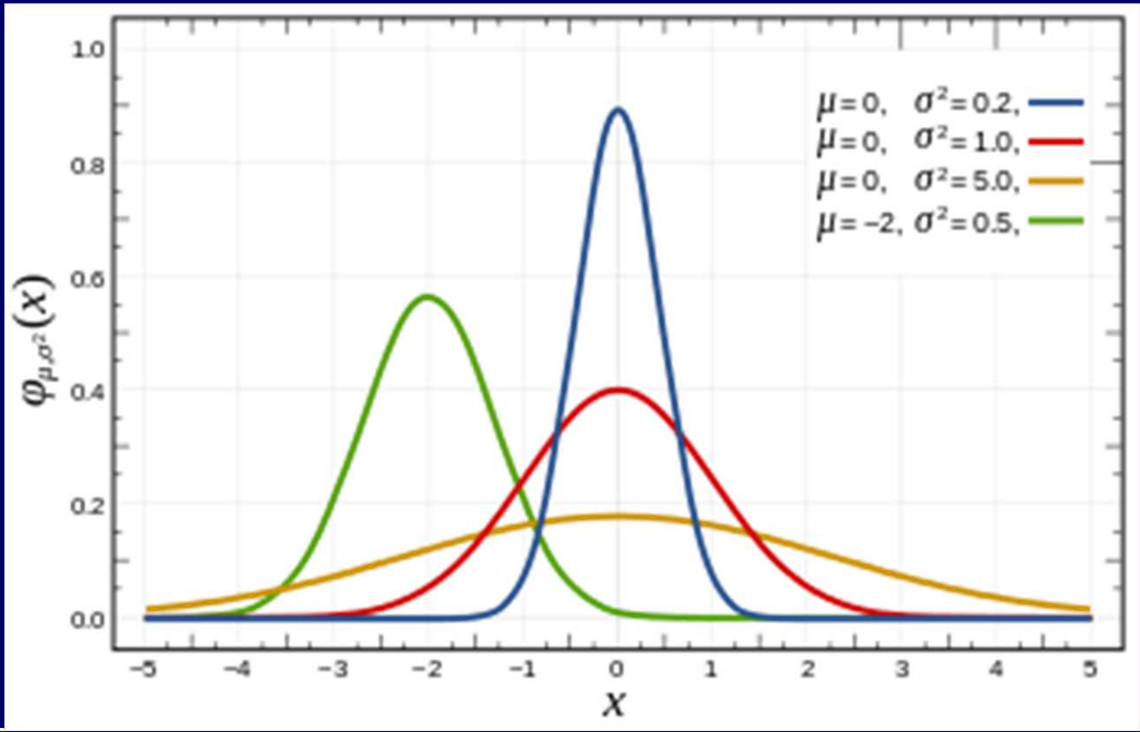
09.2.1 Statistical Basis of Hypothesis Testing

Formal Distributions and their Properties.

Normal Distribution, N(μ, σ^2)

pdf :: $$\frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

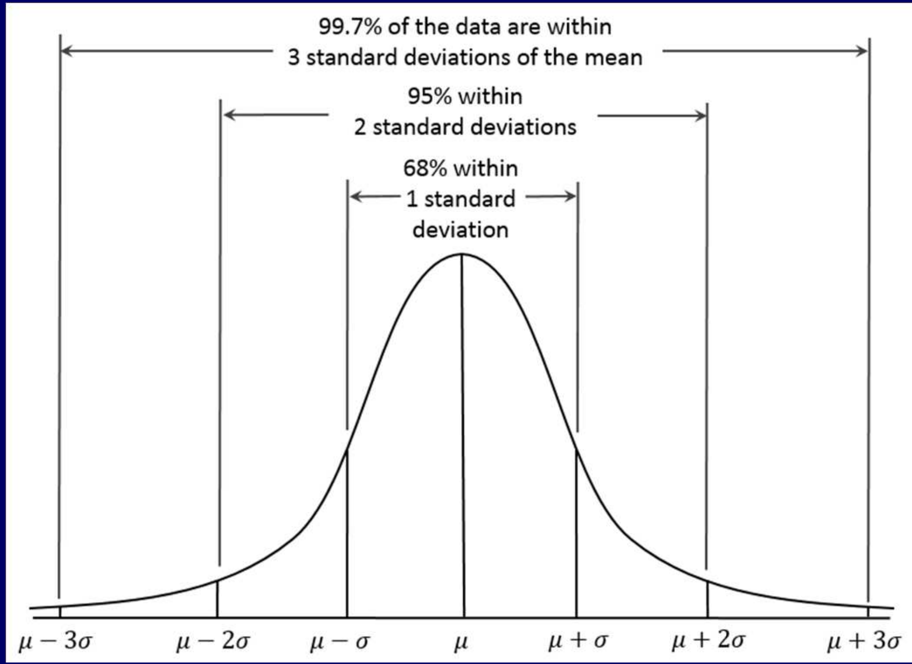
pdf



The 68-95-99.7 (Empirical) Rule for Normal Distribution

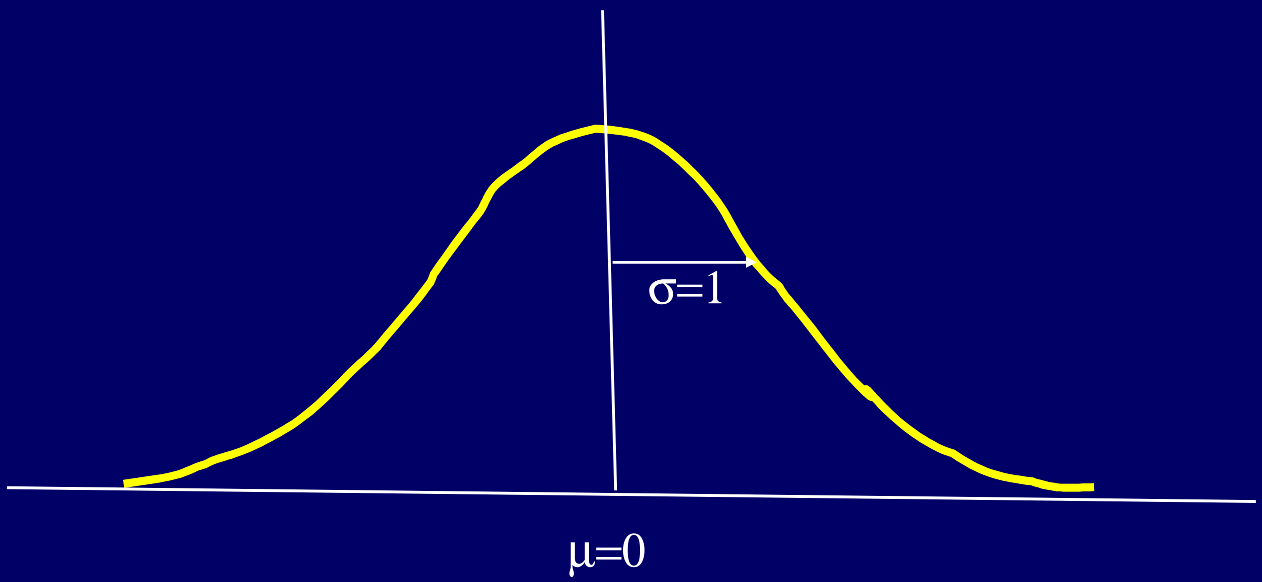
(or the 3-sigma rule)

About 68% of values drawn from a **normal distribution** are within one standard deviation σ away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations.



Properties of N(μ, σ^2)

Y && N(μ, σ^2) → Z:: (Y-μ)/σ



Standard Normal Distribution, N(0,1)

Central Limit Theorem

Let Y_1, Y_2, … Y_n be n random, independent and identically distributed variables ($\mu_1=\mu_2$; $\sigma_1^2=\sigma_2^2$…)*:

$$\rightarrow (\Sigma_{\forall i=1..n} Y_i - n\mu)/(\sigma^2 \sqrt{n}) \text{ is } N(0,1).$$

In other words:

$$\sqrt{n}\,(S_n - \mu) \rightarrow N(0,1)$$

where $S_n = \Sigma_{\forall i=1..n} Y_i /n$.

* E.g.: n sources of independent additive errors.

Consequences of the Central Limit Theorem

The central limit theorem implies that certain relevant distributions can be approximated by the normal distribution, including the binomial, the Poisson, the chi-squared and the Student's t distributions.

Consequences of the Central Limit Theorem. Binomial D.

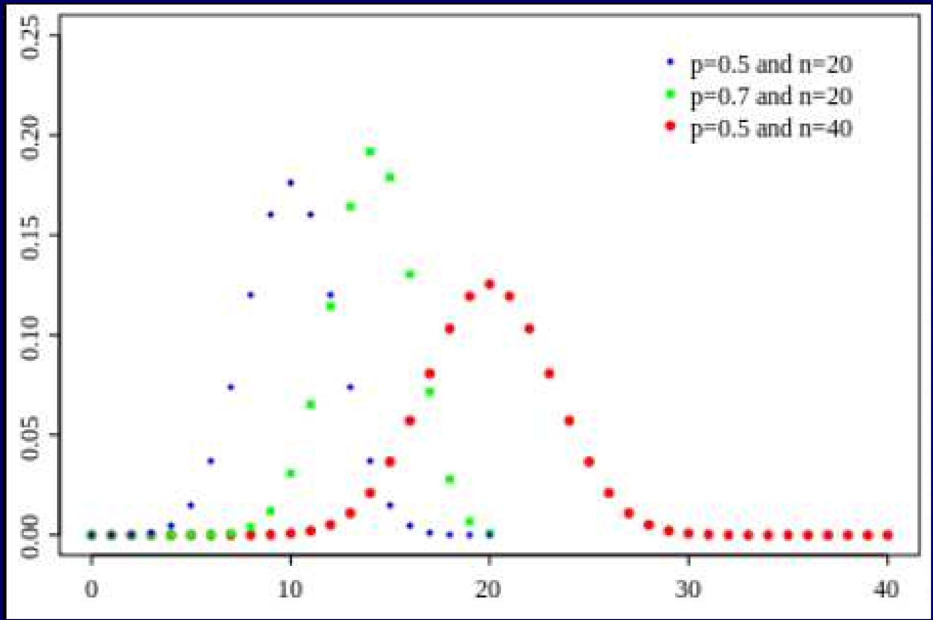
- The binomial distribution is the discrete probability distribution of the number of successes in a sequence of *n* independent yes/no experiments, each of which yields success with probability *p*. The probability of getting exactly *k* successes in *n* trials is given by:

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

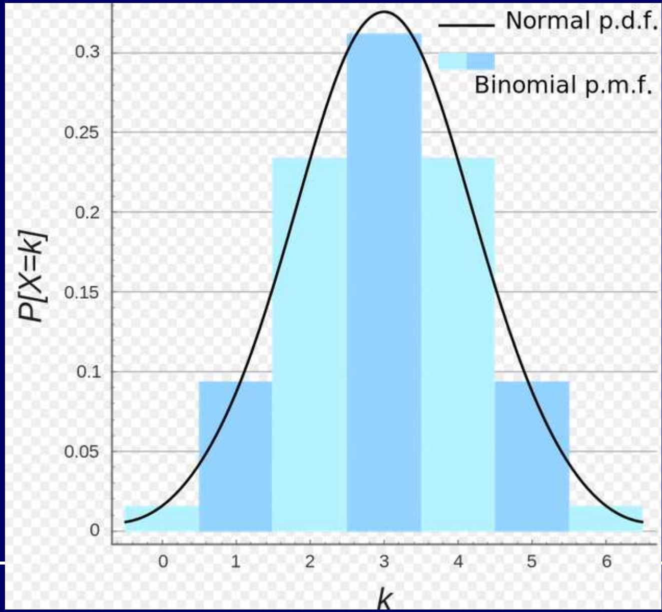
Consequences of the Central Limit Theorem



- The binomial distribution is the basis for the popular binomial test of statistical significance.

Consequences of the Central Limit Theorem

- The binomial distribution $B(n, p)$ is *approximately normal* with mean $n \bullet p$ and variance $n \bullet p(1-p))$ for large n and for p not too close to zero or one.

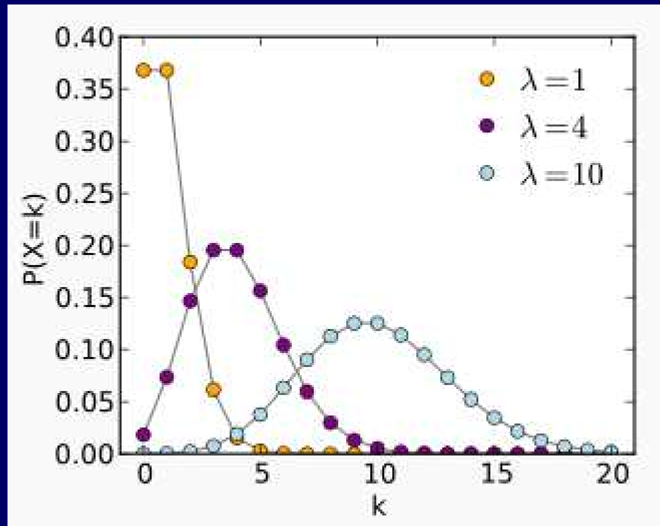


n=6
p=0.5

Consequences of the Central Limit Theorem. Poisson D.

- The Poisson distribution with parameter *λ* $\frac{\lambda^k e^{-\lambda}}{k!}$ is approximately normal *N(λ, λ)* (with mean *λ* and variance *λ*), for large values of *λ*.

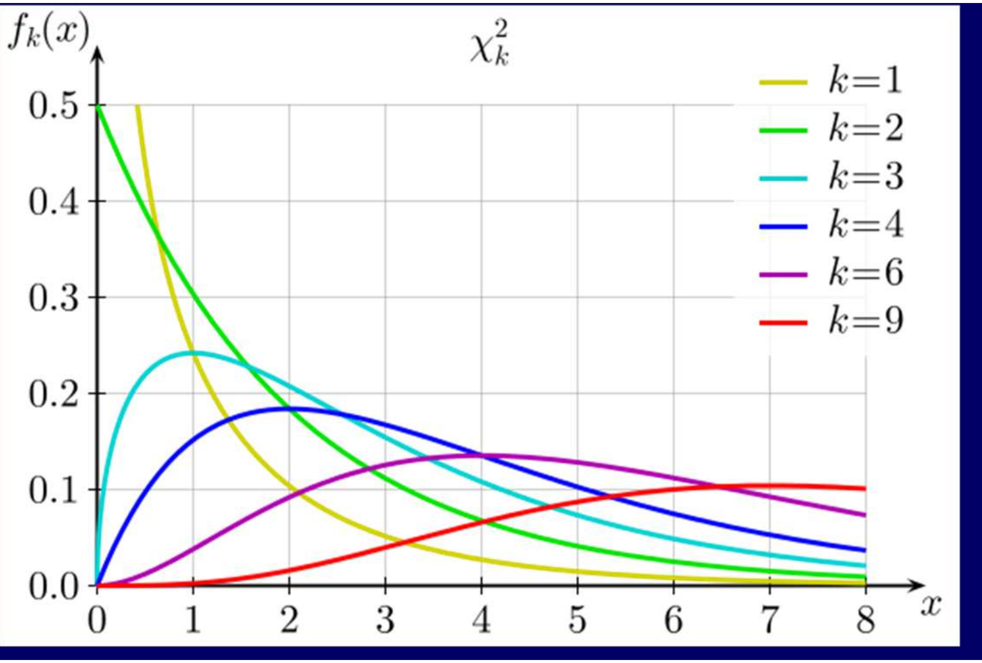
- k:: (integer) number of occurrences. The connecting lines are only guides for the eye.
- λ:: expected value



χ^2 Distribution (k DoF)

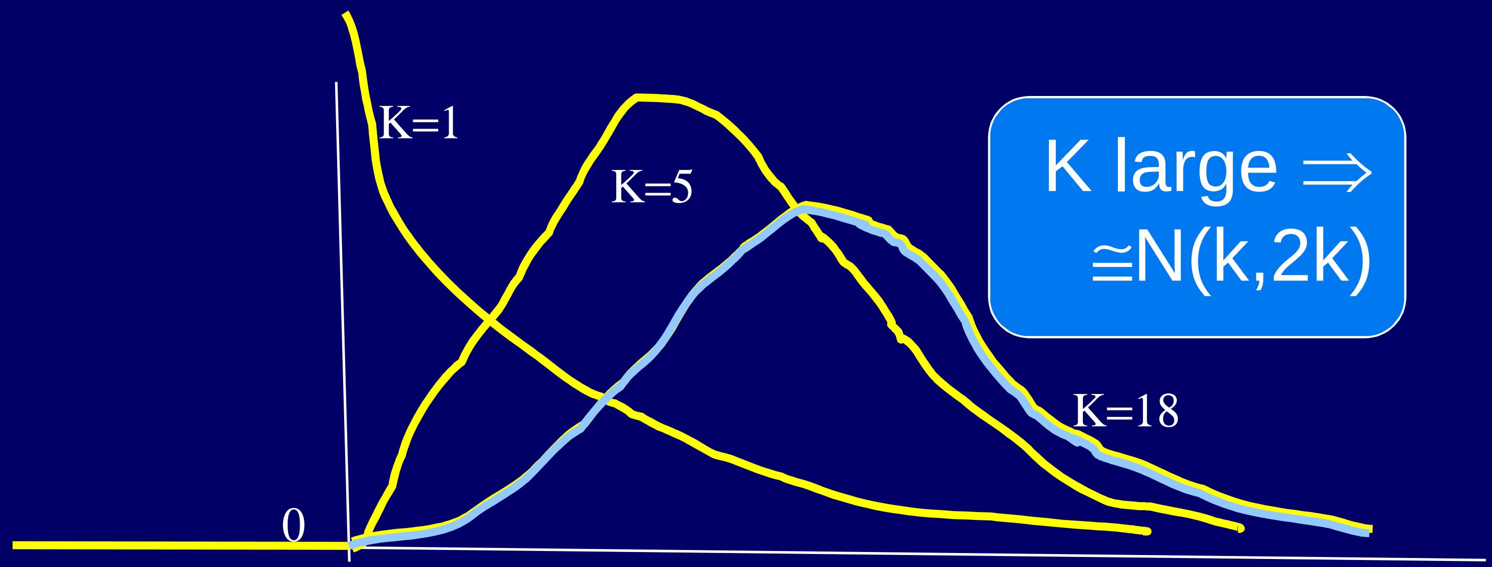
pdf ::

$$f(x;\,k) = \begin{cases} \dfrac{x^{(k/2-1)}e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$



Consequences of the Central Limit Theorem. Chi-squared D.

- The chi-squared distribution $\chi^2(k)$ is approximately normal with mean *k* and variance 2*k*, for large *k*.



χ^2 Distribution Theorem

$$Z_1, Z_2, \ldots Z_k: \forall_i, Z_i \sim NID(0,1)$$

$$\rightarrow$$

$$X = Z_1^2 + Z_2^2 + \ldots + Z_k^2 \cong \chi^2$$

χ^2 Distribution – Example

Let samples Y_1, Y_2, … Y_n come randomly from any $N(\mu, \sigma^2)$: →

$$SS/\sigma^2 = (\Sigma_{\forall i(}(Y_i - \underline{Y})^2/\sigma^2 \cong \chi_{n-1}^2$$

Because $S^2 = SS/(n-1)$:

$$S^2 \sim [\sigma^2/(n-1)]\, \chi_{n-1}^2$$

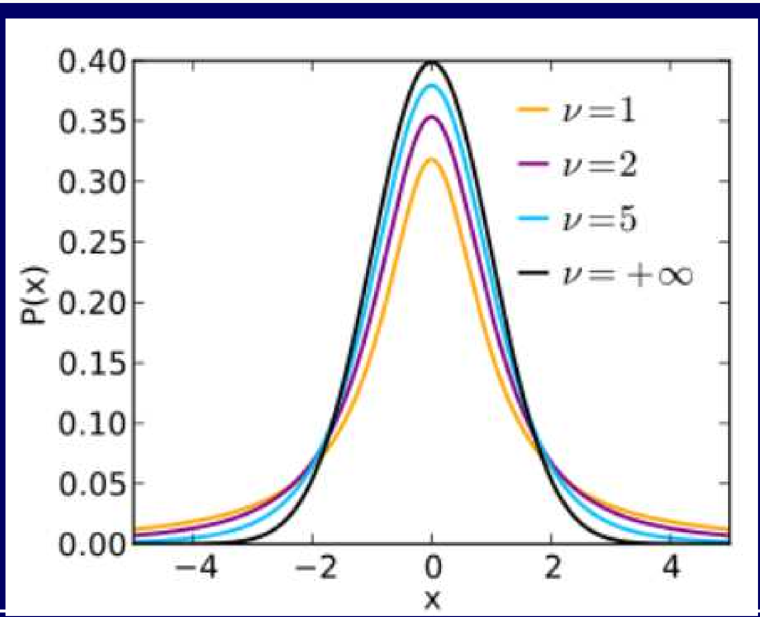
I.e.: It is χ-square distributed the variance of $N(\mu, \sigma^2)$ distributed samples.

t Student distribution (ν DoF)

f:: $\forall x \in (-\infty, +\infty), f(x) =$

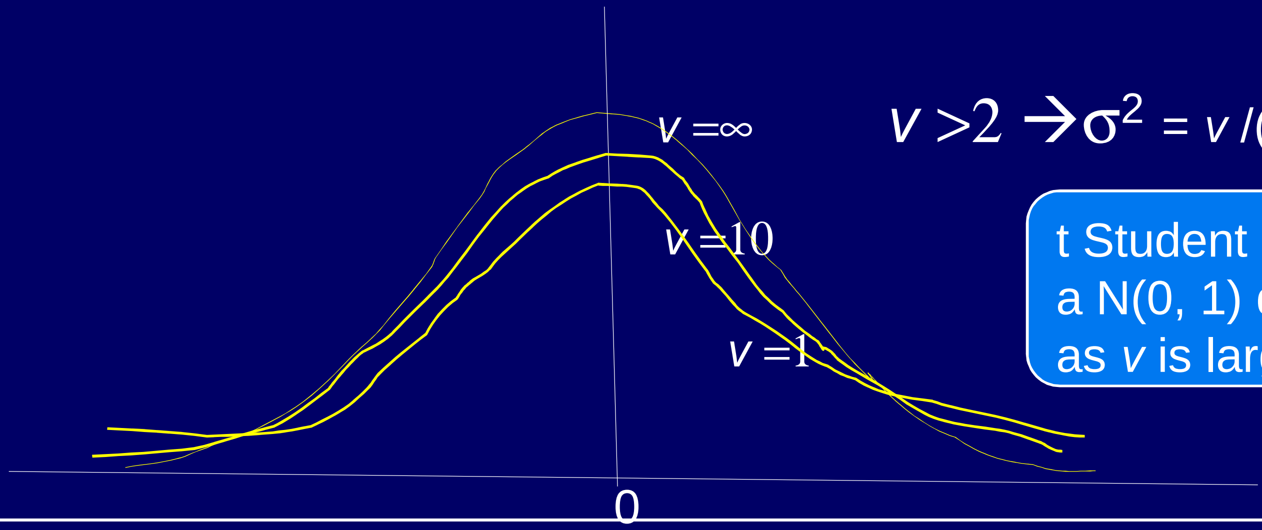
$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

pdf



Consequences of the Central Limit Theorem. Student t-D.

- The Student's t-distribution $t(x)$ is approximately normal with mean 0 and variance 1 when v is large.



$v > 2 \rightarrow \sigma^2 = v/(v-2)$

t Student is as near to a N(0, 1) distribution as *v* is large

Skip

t Student distribution Theorem

It has the distribution *t Student with k DoF,* the random variable:

$$T_k = Z/(\chi_k^2/k)$$

with Z and χ_k^2 independent random variables with distributions *N(0, 1)* and χ square, respectively.

Example of t Student distribution.

Let the sample Y_1, Y_2, … Y_n be ~ $N(\mu, \sigma^2)$.

It is t distributed with n-1 *DoF*

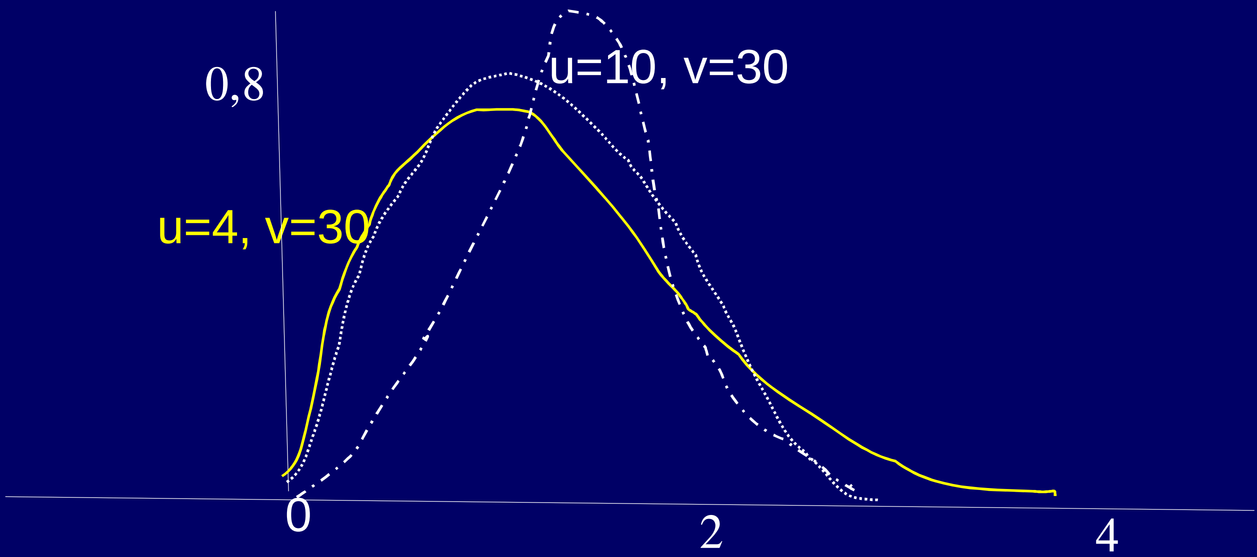
$$T = \frac{\underline{Y}-\mu}{S\sqrt{n}}$$

F distribution (DoF *u* and *v*)

The probability distribution h of a variable x is F distributed with u Degrees of Freedom of numerator and v Degrees of Freedom of denominator when:

$$h:: \forall x \in (-\infty, +\infty),\ h(x) = \frac{\Gamma(^{u+v}{}_{2})(^{u}{}_{v})^{u/2}\, x^{u/2-1}}{\Gamma(^{u}{}_{2})\, \Gamma(^{v}{}_{2})[(^{u}{}_{v})x+1]^{(u-v)/2}}$$

F distribution (DoF *u* and *v*)



Example of F distribution

Let $Y_{1,1}$, $Y_{1,2}$, … $Y_{1,n1}$ be $\sim N(\mu_1, \sigma^2)$, and

$Y_{2,1}$, $Y_{2,2}$, … $Y_{2,n2}$ be $\sim N(\mu_2, \sigma^2)$:

i.e., let us consider two normal populations and two random samples from these population; then;

the rate of the two variances is F distributed:

$$S_1^2/S_2^2 \approx F_{n1-1,\, n2-1}$$