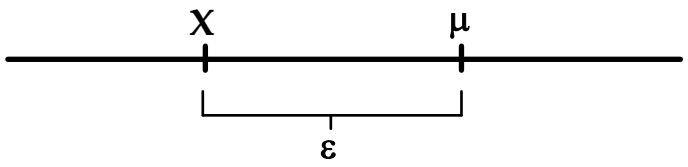
THE NORMAL DISTRIBUTION, THE LIMIT THEOREMS
AND SAMPLING DISTRIBUTIONS

Aim

- characterize the normal and the standard normal distribution
- formulate and use the limit theorems
- describe special distributions: χ^2, *Student*, *F*-distribution

Normal distribution $X \rightarrow N(\mu, \sigma^2)$

Observed Value = "True" Value + Measurement Error



Measurement error derives its properties of random variation from uncontrollable factors in the measurement process. The individual effect of any one of these factors is so small as to be undetectable. However, the accumulation of many such small perturbations results in an observable effect which we call measurement error.

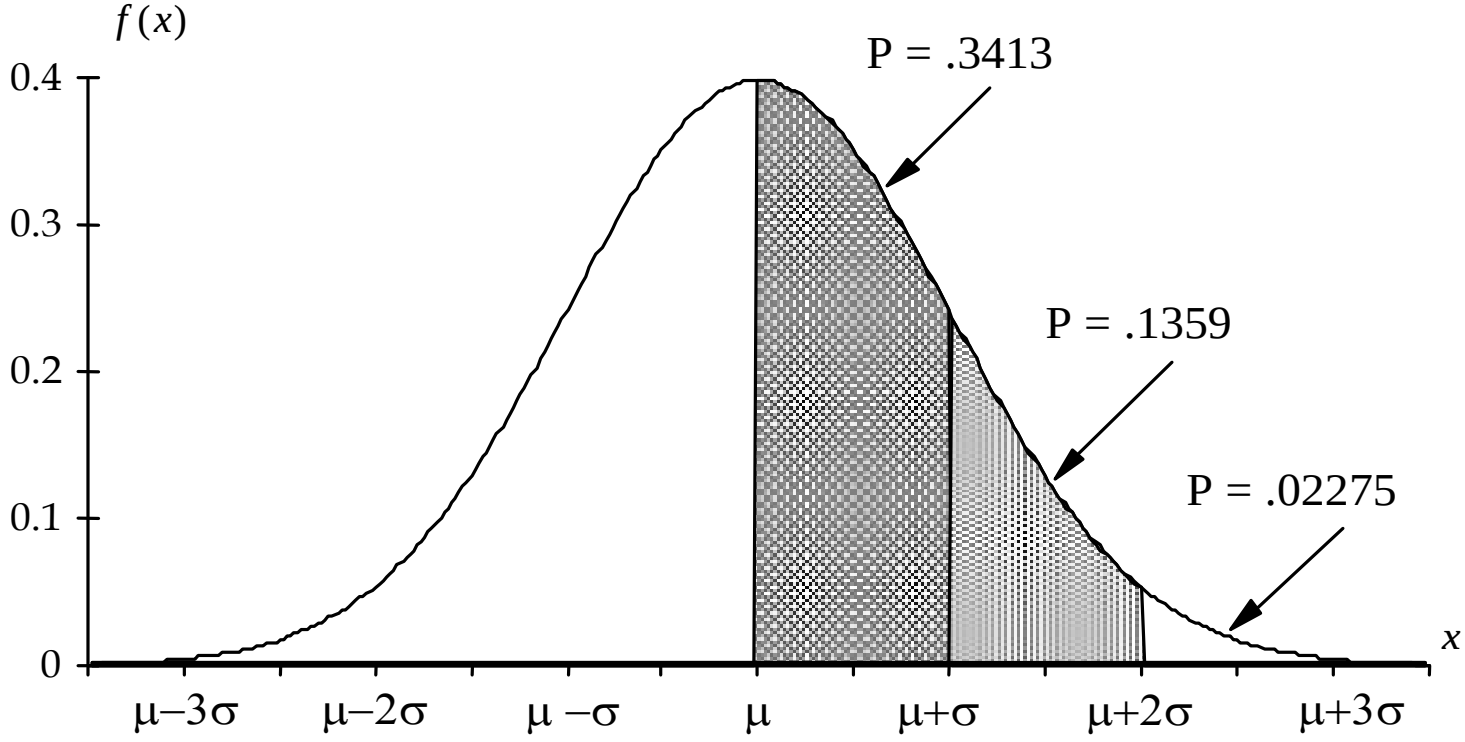
$$\varepsilon = \delta_1 + \delta_2 + \delta_3 + \ldots + \delta_k + ..$$

Astronomers were responsible for one of the earliest attempts to formally model the random variation inherent in the measurement process. The probability density function which was adopted at that time has been alternatively called the error function, the Gaussian curve, and today most commonly, the normal distribution.

Probability density function of X – normal RV

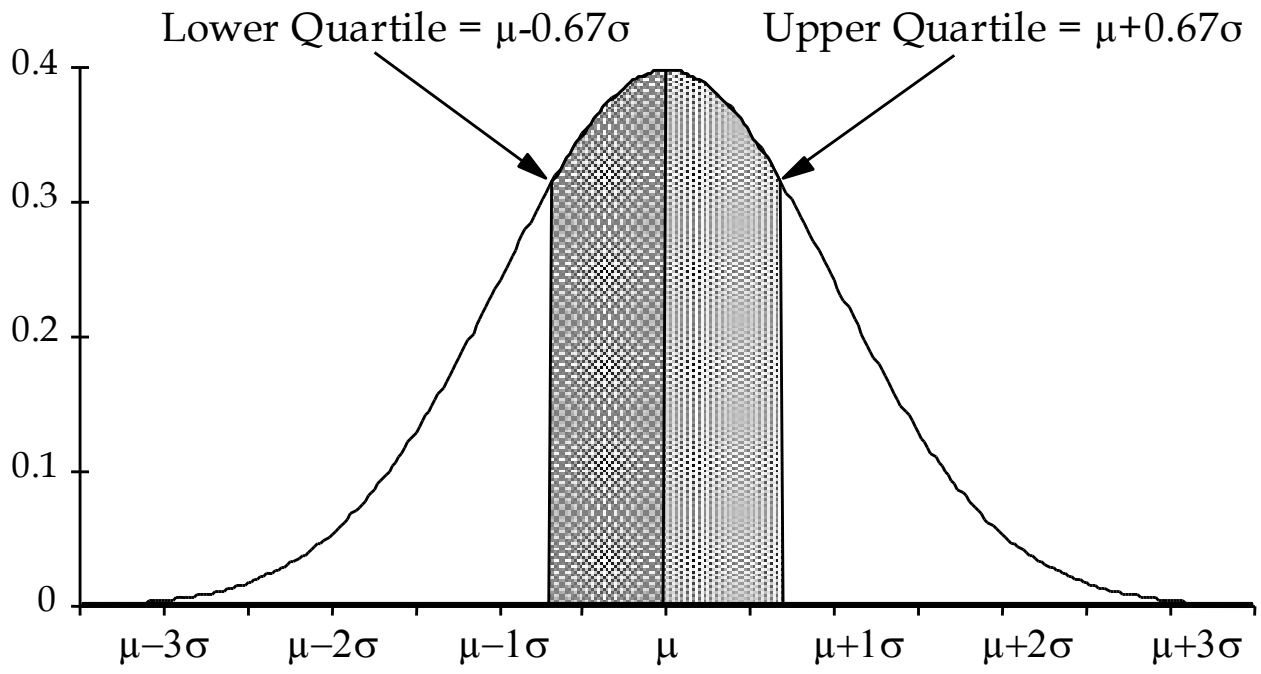
$$X \rightarrow N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; -\infty < x < +\infty \qquad \ldots \quad \text{pro } -\infty < \mu < \infty, \quad \sigma^2 > 0$$



Lower and Upper Quartile of the normal distribution

symmetric interval (μ +/- 0,67σ) divides the probability of the distribution into four equal parts of 25%.



The Normal Rules of Thirds

- **1/2 of the distribution is within 2/3 σ of the mean μ**
- **2/3 of the distribution is within σ of the mean**
- **virtually all of the distribution is within 3 σ of the mean μ, 99%**
- **more than 95% of the distribution is within 2 σ of the mean μ**

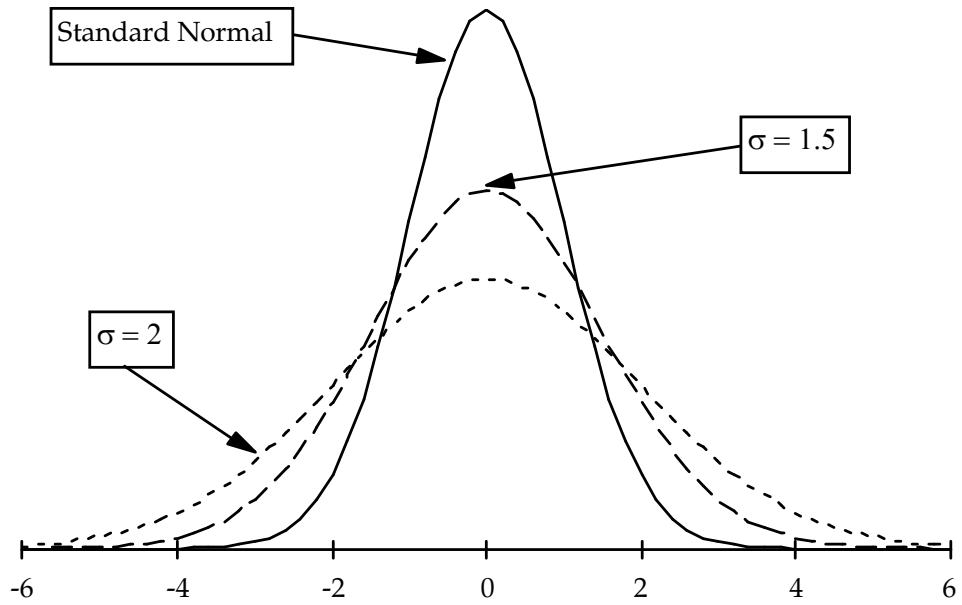
Expected value and Variance of the normal distribution:

$$E(X) = \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \mu$$

$$D(X) = E[(X - E(X))^2] = \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \sigma^2$$

Distribution function of the normal RV:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2} dt} dt$$



Standard normal distribution

- Normal distribution with μ = 0 a $\sigma^2 = 1$ … $Z \rightarrow N(0,1)$,

  standard normal density:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}; -\infty < z < +\infty$$

Distribution function Φ:

$$\Phi(z) = P(Z < z) = \int_{-\infty}^{z} \varphi(u)du = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

- Symmetric around μ = 0

$$\varphi(-z) = \varphi(z); \quad -\infty < z < +\infty \ , \ \Phi(-z) = 1 - \Phi(z); \quad -\infty < z < +\infty$$

Standardization - relation between normal and standard normal distribution

X is normal RV $N(\mu, \sigma^2)$, we can define a related standard normal random variable

$$Z = \frac{X-\mu}{\sigma}$$

Z has the **standard normal distribution**.

$$X \ldots N(\mu, \sigma^2) \;\Rightarrow\; Z = \frac{X-\mu}{\sigma} \;,\; Z \ldots \Phi(0, 1)$$

Derivation:

The distribution function of X can therefore be computed from the derived random variable:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}dt} dt = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \Phi(\frac{x-\mu}{\sigma})$$

but $$\Phi(z) = P(Z < z) = \int_{-\infty}^{z} \varphi(u)du = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

Solved example

X ... $N(2, 25)$, determine $P(2 < X < 8)$

Solution:

$P(2 < X < 8) = F(8) - F(2) = \Phi(\frac{8-2}{\sqrt{25}}) - \Phi(\frac{2-2}{\sqrt{25}}) = \Phi(1.2) - \Phi(0) = 0.885 - 0.5 = 0.385$

We can compute by the use of suitable software or tables of normal RV:

$\Phi(1.2) = 0.885$ $\Phi(0) = 0.5$

Limit Theorems

Definitions of the basic notions

Convergence in probability:

$$\forall \varepsilon > 0,\ \lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1 \quad \Rightarrow \quad X_n \xrightarrow{p} X$$

a sequence of random variables $\{X_n\}$ converges ***in probability*** *to random variable X*

Convergence in distribution:

$\{F_n(x)\}$ ... is a sequence of distribution functions corresponding to random variables $\{X_n\}$

The sequence $\{X_n\}$ converges towards X ***in distribution***, if:

$$\lim_{n\to\infty} F_n(x) = F(x),$$

$F(x)$ ... asymptotical distribution function

Law of large numbers:

$\{X_n\}$ ... is a sequence of independent random variables, each having a mean $EX_n = \mu$ and a variance $DX_n = \sigma^2$.

Define a new random variable as: $\overline{X_n} = \frac{1}{n}.\sum_{j=1}^{n} X_j$, $\quad n \in N$

The sequence $\left\{\overline{X_n}\right\}$ **converges in probability to μ**: $\overline{X_n} \xrightarrow{p} \mu$.

****

Bernoulli theorem: $\{X_n\}$ ... is a sequence of the binomial independent random variables with parameters $n=1$ a $p \in (0,1)$
(so-called alternative random variable, let $X_n = 0, 1$; $P(X_n = 1) = p$ a $P(X_n = 0) = 1-p$). Then we know that

$$\overline{X_n} = \frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow{p} p$$

Note:

$\overline{X_n}$ represents a relative frequency of the event occurrence in the sequence n trials. That is why we can estimate a probability of any occurrence by relative frequency of this event occurrence in the sequence n trials when we have a great number of the trials.

Central limit theorem

Lindeberg's theorem:

Let $X_1, X_2, \ldots, X_n \ldots$ be a sequence of independent random variables, $n \to \infty$.
$X_i \ldots$ have the same probability distribution, $EX_i = \mu$, $DX_i = \sigma^2$.
$\Rightarrow$

1. $Y_n = \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$ has an **asymptotic normal distribution** *N(0,1)*
2. $\boldsymbol{S_n}$ has an **asymptotic normal distribution**, $S_n = \sum_{i=1}^{n} X_i \to N(n\mu, n\sigma^2)$
3. Sample mean has an **asymptotic normal distribution**:

   $\overline{X_n} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, i.e.

$$\overline{X_n} \to N(\mu, \frac{\sigma^2}{n})$$

Central limit theorem

Let X_i ... has the alternative distribution thus binomial Bi(1, p)

$$S_n = \sum_{i=1}^{n} X_i \quad \rightarrow \quad S_n \text{ ... Bi}(n, p); \qquad \mathrm{E}S_n = np; \qquad \mathrm{D}S_n = np(1-p)$$

Then for large n we have CLT: $S_n = \sum_{i=1}^{n} X_i \rightarrow N(np, np(1-p))$

We use the CLT for computation the following probabilities:

$$P(k_1 < S_n < k_2) = F(k_2) - F(k_1) = \Phi\left(\frac{k_2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - np}{\sqrt{np(1-p)}}\right)$$

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Applications of the CLT – Normal Approximations to the binomial distribution

Taking *n* Bernoulli trials, at which *k* successes occur,

sample proportion is: $$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{k}{n}$$

CLT says: $$\hat{p} \rightarrow N(p, \frac{p(1-p)}{n})$$

So that we are able to compute the following probabilities:

$$P(p_1 < \hat{p} < p_2) = F(p_2) - F(p_1) = \Phi\left(\frac{p_2 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) - \Phi\left(\frac{p_1 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$F(x) = \Phi(\frac{x-\mu}{\sigma})$$

Applications of the CLT – Normal Approximations to the Poisson distribution

Time horizon is *(0,t)* , suppose λt large.

$\hat{\lambda} = \frac{X}{t}$ is number of events *X* per unit of time (rate of occurrence of events)

We know: $E(\frac{X}{t}) = \lambda$; $D(\frac{X}{t}) = \frac{\lambda}{t}$ $\hat{\lambda} = \frac{X}{t} \to N(\lambda, \frac{\lambda}{t})$

Using CLT we can compute:

$$P(g_1 < \frac{X}{t} < g_2) = F(g_2) - F(g_1) = \Phi\left(\frac{g_2 - \lambda}{\sqrt{\frac{\lambda}{t}}}\right) - \Phi\left(\frac{g_1 - \lambda}{\sqrt{\frac{\lambda}{t}}}\right)$$

$$P(k_1 < X < k_2) = F(k_2) - F(k_1) = \Phi\left(\frac{k_2 - \lambda t}{\sqrt{\lambda t}}\right) - \Phi\left(\frac{k_1 - \lambda t}{\sqrt{\lambda t}}\right) \cong \Phi\left(\frac{k_2 + 0{,}5 - \lambda t}{\sqrt{\lambda t}}\right) - \Phi\left(\frac{k_1 - 0{,}5 - \lambda t}{\sqrt{\lambda t}}\right)$$

Special sampling distribution

Chi-square distribution

if $Z_1, Z_2, \ldots, Z_n$ are independent RV, each has normal distribution $N(0,1)$, then the RV:

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2$$ is Chi-square RV with n degrees of freedom (n is parameter)

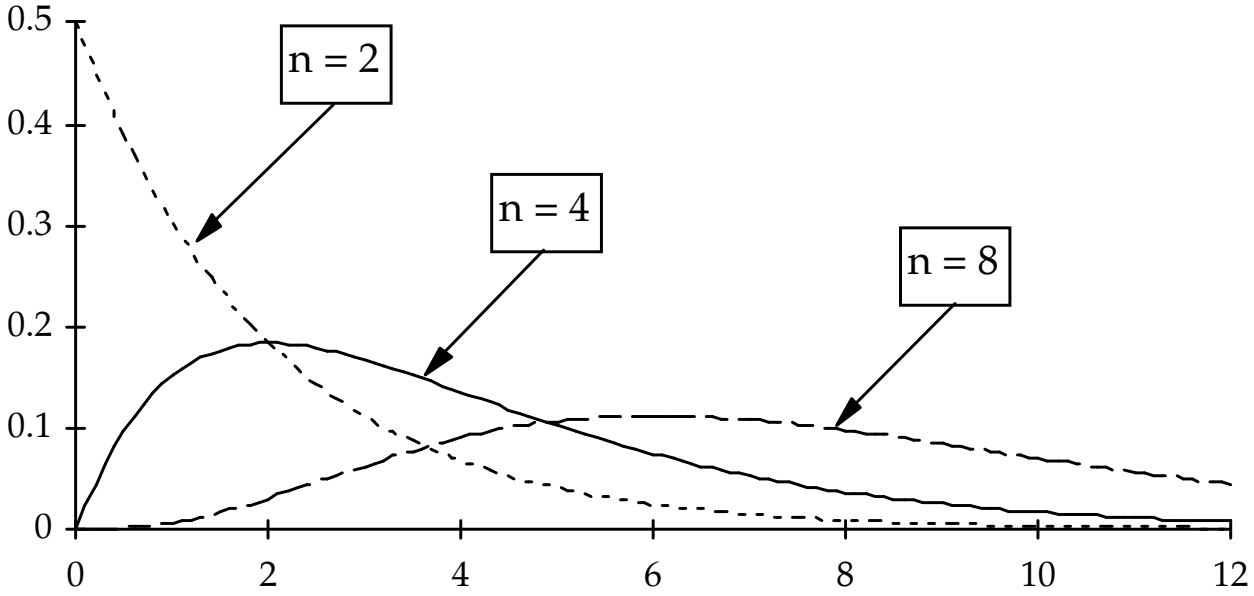
Prob. density function: $$f_\chi(x) = \frac{x^{\left(\frac{n}{2}-1\right)} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$$, $\Gamma(t)$ is a gamma function.

Mean and Variance:

$$E(\chi_n^2) = n$$

$$D(\chi_n^2) = 2n$$

The density function χ^2 of distribution



The density function for various values of the parameter *n*

Student distribution (*t* distribution)

Random variable

$$t_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

has a Student distribution with *n* degrees of freedom (*n* is parameter)

Density function:

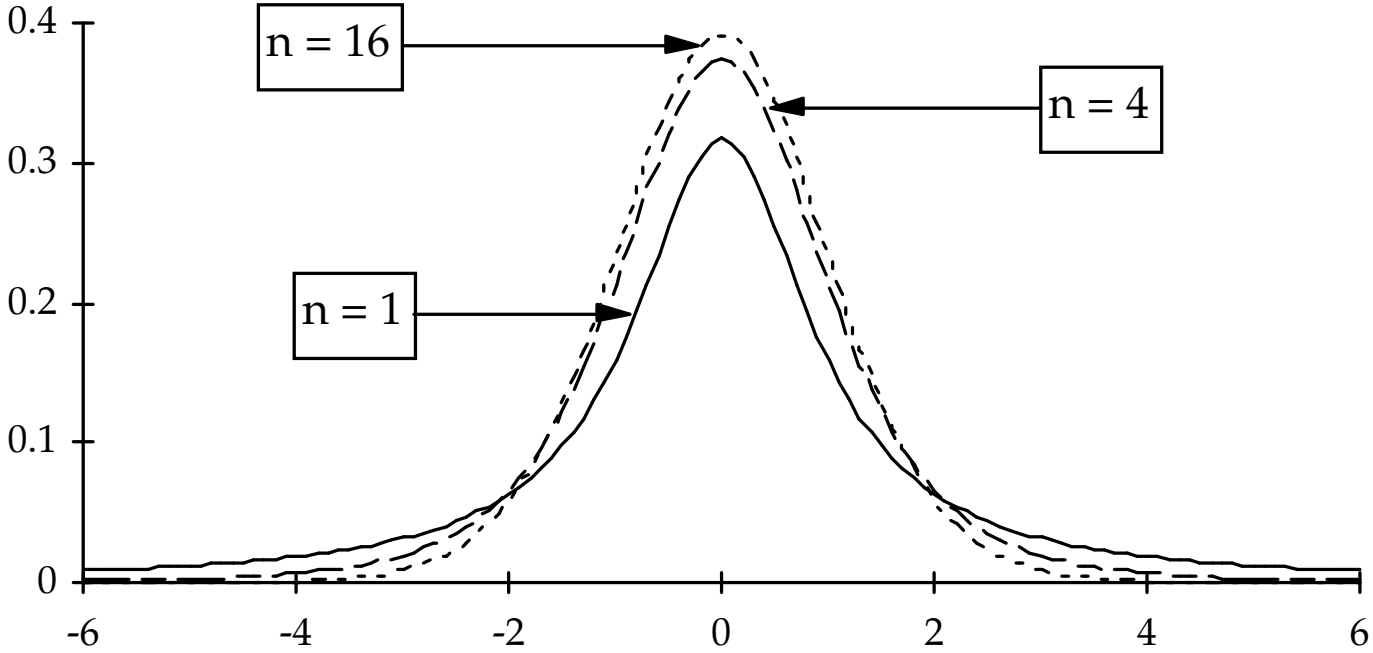
$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\ \Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Mean and Variance:

$$E(t_n) = 0$$

$$D(t_n) = \frac{n}{n-2}$$

The density function Student (*t*) distribution



Fisher-Snedecor‘s distribution - F distribution

Random variable

$$F_{n,m} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}}$$

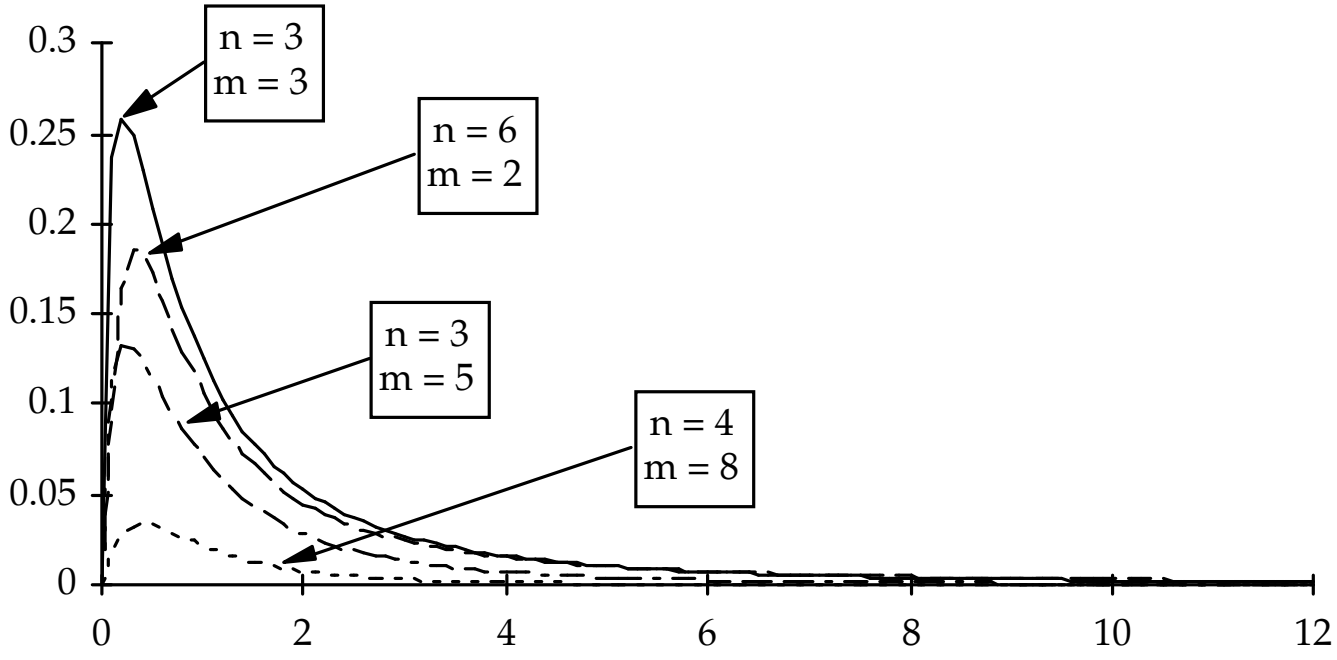
has a Fisher-Snedecor‘s distribution - F distribution with n and m degrees of freedom (n,m are parameters)

Density function

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right) n^{\frac{n}{2}} m^{\frac{m}{2}} x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)(m+nx)^{\frac{n+m}{2}}}$$

Mean and Variance:

$$E(F_{n,m}) = \frac{m}{m-2} \qquad D(F_{n,m}) = \frac{2m^2\left(1+\frac{m-2}{n}\right)}{(m-2)^2(m-4)}$$

The density function of F distribution at different degrees of freedom

m and *n*: