Regression Methods

Myrto Limnios

Autumn 2025 - Week 2

Reminder: Moment-generating function

Definition 1

The **moment-generating function (MGF)** *of a random vector* $Y_{n\times 1}$ *is*

$$M_Y(t) = \mathrm{E}(e^{t^{\mathrm{T}}Y}) = \mathrm{E}(e^{\sum_{j=1}^n t_j Y_j}), \quad t \in \mathcal{T} = \{t \in \mathbb{R}^n : M_Y(t) < \infty\},$$

and the **cumulant-generating function** *of* Y *is* $K_Y(t) = \log M_Y(t)$, $t \in \mathcal{T}$.

Then

- $0 \in \mathcal{T}$, so $M_Y(0) = 1$ and $K_Y(0) = 0$;
- if $\mathcal{T}$ contains an open set, then

$$\mu = \mathrm{E}(Y) = K_Y'(0) = \left.\frac{\partial K_Y(t)}{\partial t}\right|_{t=0}, \quad \Omega = \mathrm{var}(Y) = \left.\frac{\partial^2 K_Y(t)}{\partial t \partial t^{\mathrm{T}}}\right|_{t=0};$$

- if $\mathcal{A}, \mathcal{B}$ are disjoint subsets of $\{1, \ldots, n\}$ and $Y_{\mathcal{A}}$ denotes the sub-vector of Y containing $\{Y_j : j \in \mathcal{A}\}$, etc., then $Y_{\mathcal{A}} \perp\!\!\!\perp Y_{\mathcal{B}}$ if and only if

$$M_Y(t) = \mathrm{E}(e^{t_{\mathcal{A}}^{\mathrm{T}} Y_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}} Y_{\mathcal{B}}}) = M_{Y_{\mathcal{A}}}(t_{\mathcal{A}}) M_{Y_{\mathcal{B}}}(t_{\mathcal{B}}), \quad t \in \mathcal{T};$$

- the MGF of $Y_{\mathcal{A}}$ equals $M_Y(t)$ evaluated with $t_{\mathcal{B}} = 0$;
- if we recognise an MGF, then we know the probability distribution that gave it.

A random variable $Y_{n\times 1}$ with real components has the **multivariate normal distribution,** $Y \sim \mathcal{N}_n(\mu, \Omega)$, if $a^{\mathrm{T}}Y \sim \mathcal{N}(a^{\mathrm{T}}\mu, a^{\mathrm{T}}\Omega a)$ for every constant vector $a_{n\times 1}$, and then

1. (a) Ω is symmetric semi-positive definite with real components and

$$\mathrm{E}(Y) = \mu_{n\times 1}, \quad \mathrm{var}(Y) = \Omega_{n\times n}, \quad M_Y(t) = \exp(t^{\mathrm{T}}\mu + \tfrac{1}{2}t^{\mathrm{T}}\Omega t), \quad t \in \mathbb{R}^n,$$

   where we call μ the **mean vector** and Ω the **(co)variance matrix** of X;

2. (b) for any constants $a_{m\times 1}$ and $B_{m\times n}$, $a + BY \sim \mathcal{N}_m\left(a + B\mu, B\Omega B^{\mathrm{T}}\right)$;

3. (c) if $Y^{\mathrm{T}} = (Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}})$, where Y_1 is $m \times 1$, and μ and Ω are partitioned correspondingly, then the marginal and conditional distributions of Y_1 are also multivariate normal:

$$Y_1 \sim \mathcal{N}_m(\mu_1, \Omega_{11})$$

$$Y_1 \mid Y_2 = y_2 \sim \mathcal{N}_m\left\{\mu_1 + \Omega_{12}\Omega_{22}^{-1}(y_2 - \mu_2), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right\}$$

4. (d) $Y_1 \perp\!\!\!\perp Y_2$ iff $\Omega_{12} = 0$, and $a + BY \perp\!\!\!\perp c + DY$ iff $B\Omega D^{\mathrm{T}} = 0$;

5. (e) if $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $Y_{n\times 1} \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$; and finally,

6. (f) Y has a density on $\mathbb{R}^n$ iff Ω is positive definite (i.e., has rank n), and then

$$f(y; \mu, \Omega) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^{\mathrm{T}}\Omega^{-1}(y-\mu)\right\}, \quad y \in \mathbb{R}^n. \qquad (1)$$

Multivariate normal distribution

(a) Let e_j denote the n-vector with 1 in the jth place and zeros everywhere else.

- Then $Y_j = e_j^{\mathrm{T}} Y \sim N(\mu_j, \omega_{jj})$, giving the mean and variance of Y_j.
- Now $\mathrm{var}(Y_j + Y_k) = \mathrm{var}(Y_j) + \mathrm{var}(Y_k) + 2\mathrm{cov}(Y_j, Y_k)$, and

$$Y_j + Y_k = (e_j + e_k)^{\mathrm{T}} Y \sim \mathcal{N}(\mu_j + \mu_k, \omega_{jj} + \omega_{kk} + 2\omega_{jk}),$$

  which implies that $\mathrm{cov}(Y_j, Y_k) = \omega_{jk} = \omega_{kj}$. This gives the mean and covariance matrix of Y.
- Since $u^{\mathrm{T}} Y \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u)$, its MGF is $M_{u^{\mathrm{T}}Y}(t) = \mathrm{E}(e^{tu^{\mathrm{T}}Y}) = \exp(tu^{\mathrm{T}}\mu + \frac{1}{2}t^2 u^{\mathrm{T}}\Omega u)$. The MGF of Y is $M_Y(u) = \mathrm{E}(e^{u^{\mathrm{T}}Y}) = M_{u^{\mathrm{T}}Y}(1) = \exp(u^{\mathrm{T}}\mu + \frac{1}{2}u^{\mathrm{T}}\Omega u)$, for any $u \in \mathbb{R}^p$, as stated.

(b) The MGF of $a + BY$ equals

$$\begin{aligned}
\mathrm{E}\left[\exp\{t^{\mathrm{T}}(a + BY)\}\right] &= \mathrm{E}\left[\exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}Y)\}\right] \\
&= e^{t^{\mathrm{T}}a} M_Y(B^{\mathrm{T}}t) \\
&= \exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}\mu + \tfrac{1}{2}(B^{\mathrm{T}}t)^{\mathrm{T}}\Omega(B^{\mathrm{T}}t)\} \\
&= \exp\left\{t^{\mathrm{T}}(a + B\mu) + \tfrac{1}{2}t^{\mathrm{T}}(B\Omega B^{\mathrm{T}})t\right\},
\end{aligned}$$

which is the MGF of the $\mathcal{N}_m(a + B\mu, B\Omega B^{\mathrm{T}})$ distribution. Hence linear combinations of normal variables are themselves normal.

(c) Write $Y^{\mathrm{T}} = (Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}})$ and partition μ and Ω conformally. Then

$$M_Y(t) = \exp\left\{t_1^{\mathrm{T}}\mu_1 + t_2^{\mathrm{T}}\mu_2 + \tfrac{1}{2}\left(t_1^{\mathrm{T}}\Omega_{11}t_1 + 2t_1^{\mathrm{T}}\Omega_{12}t_2 + t_2^{\mathrm{T}}\Omega_{22}t_2\right)\right\}$$

and by setting $t_2 = 0$ and then $t_1 = 0$ we have

$$M_{Y_1}(t_1) = \exp\left(t_1^{\mathrm{T}}\mu_1 + \tfrac{1}{2}t_1^{\mathrm{T}}\Omega_{11}t_1\right), \quad M_{Y_2}(t_2) = \exp\left(t_2^{\mathrm{T}}\mu_2 + \tfrac{1}{2}t_2^{\mathrm{T}}\Omega_{22}t_2\right).$$

Hence the marginal distribution of Y_1 is $\mathcal{N}_m(\mu_1, \Omega_{11})$.
For the conditional distribution, note that $W = Y_1 - \Omega_{12}\Omega_{22}^{-1}Y_2$ is a linear combination of Y and

$$\mathrm{E}(W) = \mu_1 - \Omega_{12}\Omega_{22}^{-1}\mu_2, \quad \mathrm{var}(W) = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}, \quad \mathrm{cov}(W, Y_2) = \Omega_{12} - \Omega_{12}\Omega_{22}^{-1}$$

Hence $W \perp\!\!\!\perp Y_2$. As $Y_1 = W + \Omega_{12}\Omega_{22}^{-1}Y_2$ and conditioning on Y_2 does not change the distribution of W,

$$\mathrm{E}(Y_1 \mid Y_2 = y_2) = \mathrm{E}(W) + \Omega_{12}\Omega_{22}^{-1}y_2$$

$$\mathrm{var}(Y_1 \mid Y_2 = y_2) = \mathrm{var}(W + \Omega_{12}\Omega_{22}^{-1}y_2) = \mathrm{var}(W).$$

Putting the pieces together gives the stated conditional distribution.

(d) The joint MGF $M_Y(t)$ given in (c) factorises iff the variables are independent, and

$$M_Y(t) = M_{Y_1}(t_1)M_{Y_2}(t_2) \text{ for all } t \in \mathbb{R}^n \quad \Longleftrightarrow \quad \Omega_{12} = 0.$$

The variance matrix of

$$\begin{pmatrix} a \\ c \end{pmatrix} + \begin{pmatrix} B \\ D \end{pmatrix} Y$$

is

$$\begin{pmatrix} B\Omega B^{\mathrm{T}} & B\Omega D^{\mathrm{T}} \\ D\Omega B^{\mathrm{T}} & D\Omega D^{\mathrm{T}} \end{pmatrix},$$

so $a + BY \perp\!\!\!\perp c + DY$ iff $B\Omega D^{\mathrm{T}} = 0$.

(e) Each Y_j has mean μ and variance σ^2, and since they are independent, $\text{cov}(Y_j, Y_k) = 0$ for $j \neq k$. If $u \in \mathbb{R}^n$, then $u^{\mathrm{T}}Y$ is a linear combination of normal variables, with mean $\sum_{j=1}^n u_j\mu = u^{\mathrm{T}}\mu 1_n$ and variance $\sum_{j=1}^n u_j^2\sigma^2 = u^{\mathrm{T}}\sigma^2 I_n u$, so $Y \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$, as required.

(f) Since Ω is symmetric and positive semi-definite, the spectral theorem tells us that we may write $\Omega = ADA^{\mathrm{T}}$, where $D = \text{diag}(d_1, \ldots, d_n)$ contains the (real) eigenvalues of Ω, with $d_1 \geq \cdots \geq d_n \geq 0$, and A is a $n \times n$ orthogonal matrix, i.e., $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I_n$ and $|A| = 1$. The columns $A_1, \ldots, A_n$ of A are the eigenvectors corresponding to the respective eigenvalues,

$$\Omega = ADA^{\mathrm{T}} = \sum_{j=1}^n d_j a_j a_j^{\mathrm{T}},$$

with $|\Omega| = |ADA^{\mathrm{T}}| = |A| \times |D| \times |A^{\mathrm{T}}| = |D|$ and $\Omega^{-1} = AD^{-1}A^{\mathrm{T}}$ if the inverse exists.

- Now let $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ be a vector of independent standard normal variables, set $u \in \mathbb{R}^n$, and consider

$$u^{\mathrm{T}}(\mu + AD^{1/2}Z) = u^{\mathrm{T}}\mu + \sum_{j=1}^n Z_j u^{\mathrm{T}} a_j d_j^{1/2}.$$

This is a linear combination of normal variables, so it has a normal distribution, with mean $u^{\mathrm{T}}\mu$ and variance

$$\text{var}\left(u^{\mathrm{T}}\mu + \sum_{j=1}^n Z_j u^{\mathrm{T}} a_j d_j^{1/2}\right) = \sum_{j=1}^n d_j (u^{\mathrm{T}} a_j)^2 \text{var}(Z_j) = u^{\mathrm{T}} \left(\sum_{j=1}^n d_j a_j a_j^{\mathrm{T}}\right) u$$
$$= u^{\mathrm{T}}\Omega u,$$

so $X \stackrel{D}{=} \mu + AD^{1/2}Z \sim N_n(\mu, \Omega)$.

- If Ω has rank n, then $d_n > 0$. The change of variables $z \mapsto x = \mu + AD^{1/2}z$ has Jacobian

$$\left|\frac{\partial x}{\partial z}\right| = |AD^{1/2}| = |A||D|^{1/2} = 1 \times |D|^{1/2} = |\Omega|^{1/2} > 0.$$

Moreover $z = D^{-1/2}A^{\mathrm{T}}(x-\mu)$, and therefore $z^{\mathrm{T}}z = (x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)$. Hence using the joint density of Z, $f_Z(z) = (2\pi)^{-n/2}\exp(-\sum_{j=1}^n z_j^2/2)$,

$$\begin{aligned} f_X(x) = f_Z(z)|_{z=D^{-1/2}A^{\mathrm{T}}(x-\mu)} \left|\frac{\partial z}{\partial x}\right| \\ = (2\pi)^{-n/2} \exp\left(-\frac{z^{\mathrm{T}}z}{2}\right)\Bigg|_{z=D^{-1/2}A^{\mathrm{T}}(x-\mu)} |\Omega|^{-1/2}, \end{aligned}$$

which reduces to (1). If $d_n = 0$, then the Jacobian is zero, so the transformation $z \mapsto x$ is singular and X does not have a density on $\mathbb{R}^n$.

- Now suppose that $d_m > d_{m+1} = 0$, so just m eigenvalues of Ω are positive. Then

$$X = \mu + \sum_{j=1}^m Z_j a_j d_j^{1/2} \in \mathcal{S} = \mu + \mathrm{span}(a_1, \ldots, a_m),$$

where $\mathcal{S}$ is a hyperplane of dimension m passing through μ and generated by the vectors $a_1, \ldots, a_m$. In this case the previous argument shows that X has an m-dimensional Gaussian density on $\mathcal{S}$, but places no probability elsewhere.

χ^2 distribution

Definition 2

If $Y_j \overset{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$, *then* $W = Y_1^2 + \cdots + Y_\nu^2$ *has the* **non-central chi-square distribution with** ν **degrees of freedom (df) and non-centrality parameter** $\delta^2 = (\mu_1^2 + \cdots + \mu_\nu^2)/\sigma^2$; *we write* $W \sim \sigma^2 \chi^2_\nu(\delta^2)$. *Then*

$$M_W(t) = \exp\left(\frac{t\sigma^2\delta^2}{1-2t\sigma^2}\right)(1-2\sigma^2 t)^{-\nu/2}, \quad t < 1/(2\sigma^2).$$

If $\delta^2 = 0$ *and* $\sigma^2 = 1$ *then* W *has the (central)* **chi-square distribution with** ν **df**, *we write* $W \sim \chi^2_\nu$, *its MGF is* $M_W(t) = (1-2t)^{-\nu/2}$, *and its p-quantile is* $c_\nu(p)$.

Chi-square variables satisfy

- $\mathrm{E}(W) = \sigma^2(\nu + \delta^2)$, $\mathrm{var}(W) = 2\sigma^4(\nu + 2\delta^2)$;
- if $W_1 \sim \chi^2_{\nu_1} \perp\!\!\!\perp W_2 \sim \chi^2_{\nu_2}$, then $W_1 + W_2 \sim \chi^2_{\nu_1+\nu_2}$;
- $W \sim \chi^2_\nu$ implies that W has the gamma density

$$f(w) = \frac{\beta^\alpha w^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta w}, \quad w > 0, \quad \alpha, \beta > 0,$$

with $\alpha = \nu/2$ and $\beta = 1/2$.

χ^2_ν densities

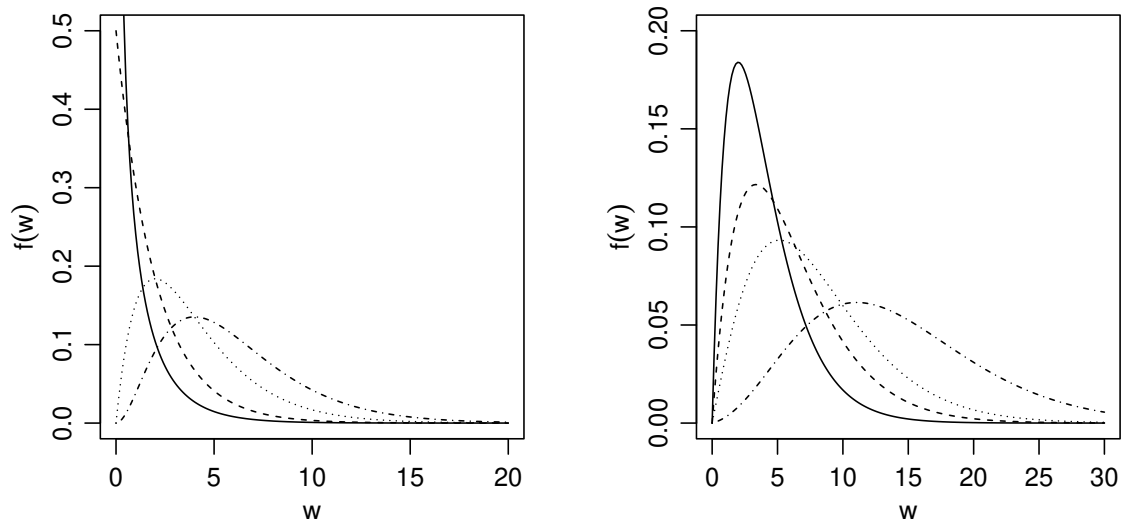


Figure: Left: central densities with $\nu = 1, 2, 4, 6$ (solid, large dashes, small dashes, dot-dash). Right: non-central densities with $\nu = 4$ and $\delta = 0, 2, 4, 10$ (solid, large dashes, small dashes, dot-dash).

Student t distribution

Definition 3

If $Z \sim \mathcal{N}(0,1) \perp\!\!\!\perp W \sim \chi^2_\nu$, *then* $T = Z/(W/\nu)^{1/2}$ *has the* **Student t distribution with ν df**, $T \sim t_\nu$, *and we write* $t_\nu(p)$ *for the corresponding p-quantile. The density function of* T *is*

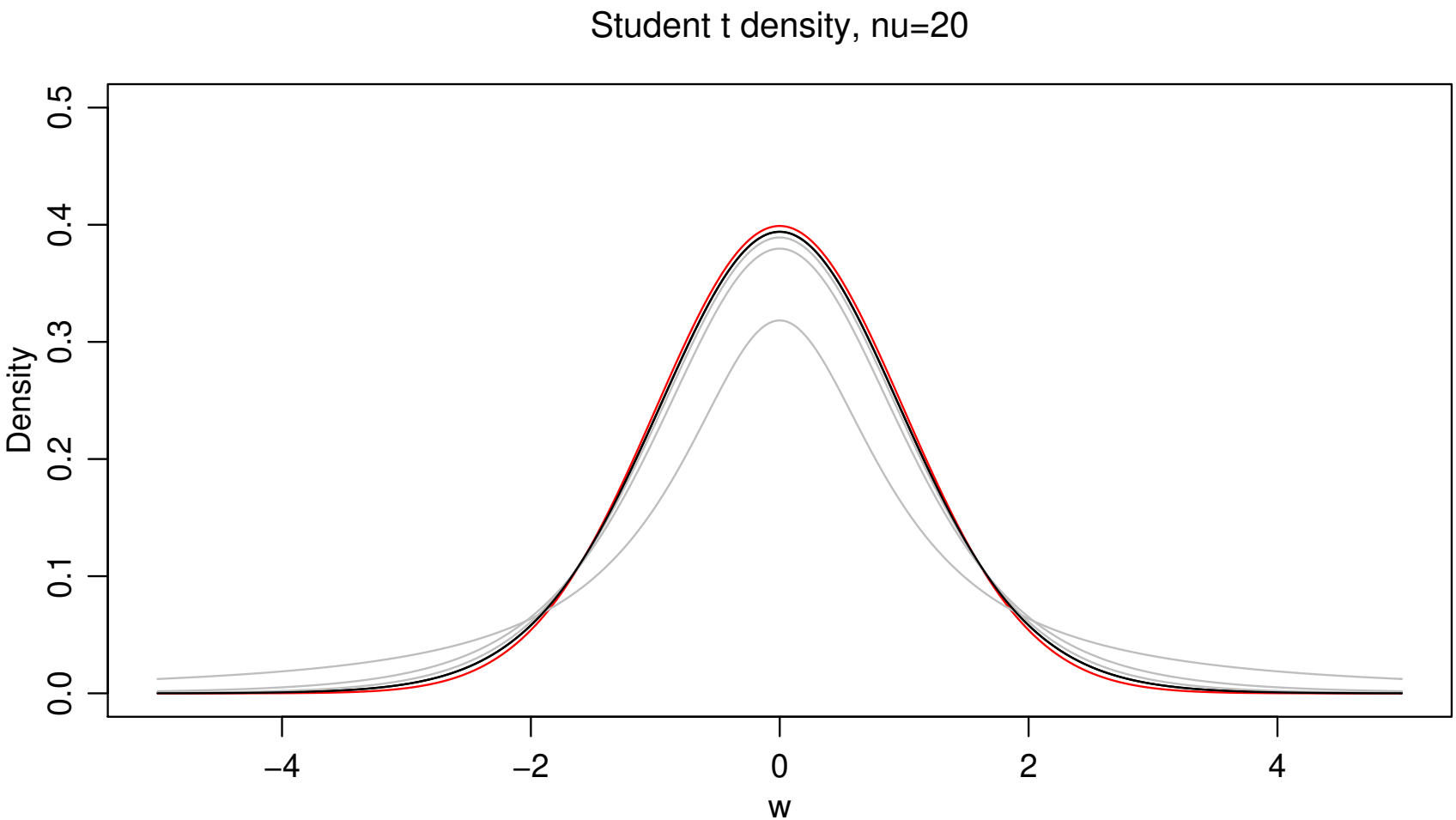
$$f_T(t) = \frac{\Gamma\{(\nu+1)/2\}}{\sqrt{\nu\pi}\Gamma(\nu/2)} \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \quad -\infty < t < \infty, \ \nu = 1, 2, \ldots.$$

Properties:

- the mean and variance exist only for $\nu \geq 2$ and $\nu \geq 3$ respectively, and then
$$\mathrm{E}(T) = 0, \quad \mathrm{var}(T) = \frac{\nu}{\nu - 2};$$
- with $\nu = 1$ we have the **Cauchy density**,
$$\frac{1}{\pi(1+t^2)}, \quad -\infty < t < \infty,$$
and then T has no moments;
- as $\nu \to \infty$, the limiting distribution of T is $\mathcal{N}(0,1)$; usually the approximation is 'good enough' for $\nu > 25$ (say).

Student t densities

Student t density functions with $\nu = 1, 5, 10, 20$ (black, $\nu = 20$), and the standard normal density (red):



F distribution

Definition 4

If $W_1, W_2 \overset{\text{ind}}{\sim} \chi^2_{\nu_1}, \chi^2_{\nu_2}$, *then*

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has the F **distribution with** ν_1 **and** ν_2 **df**: *we write* $F \sim F_{\nu_1,\nu_2}$.
The density function is

$$f_F(u) = \frac{\Gamma\left(\frac{1}{2}\nu_1 + \frac{1}{2}\nu_2\right) \nu_1^{\nu_1/2} \nu_2^{\nu_2/2}}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)} \frac{u^{\frac{1}{2}\nu_1 - 1}}{(\nu_2 + \nu_1 u)^{(\nu_1+\nu_2)/2}}, \quad u > 0,\ \nu_1, \nu_2 = 1, 2, \ldots,$$

and the p-quantile is written $F_{\nu_1,\nu_2}(p)$.

Computation

- Quantiles of the $\mathcal{N}(\mu, \sigma^2)$, χ^2_ν, t_ν, F_{ν_1,ν_2} distributions can be found in tables, or in environments such as `R` (see `http://www.r-project.org/`), where they can also be simulated.
- Examples:

```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.2.1  (2005-12-20 r36812)
...
> qnorm(0.025)     # this is a comment; access normal quantiles
[1] -1.959964      # the [1] means this is the first  element
 of a vector
> ?qnorm           # help on use of function qnorm()
> qchisq(0.025,df=3)  # chi-squared quantiles, nu=3
[1] 0.2157953
> qt(0.025,df=3)      # t quantiles, nu=3
[1] -3.182446
> qf(0.025,df1=3,df2=4)  # F quantiles, nu1=3, nu2=4
[1] 0.06622087
```

Statistical models

- Least squares fitting gives a deterministic description of the variation in some numbers y in terms of other numbers X.
- A **statistical model** is a description of data y in terms of a collection of probability distributions on the sample space for y.
- We distinguish
  - **primary** aspects of a model, which specify what questions we aim to answer, from
  - **secondary** aspects, which complete the model, indicate what analysis might be suitable, and determine the precision of conclusions.
- Often the primary aspects are embodied in one or more **parameters** of the model.
- (Almost) all models are **tentative**, and we must check that they are reasonable.