1 Reminder: χ^2_n = "sum of squares of n independent standard-normals"

Let $Z_1, Z_2, \ldots, Z_n$ be n independent random variables with the standard normal distribution $\mathcal{N}(0,1)$. Then, $X_n := Z_1^2 + Z_2^2 + \cdots + Z_n^2$ has the χ^2_n distribution, the Chi-square distribution of n degrees of freedom.

The probability density function of X_n is given by:

$$f_{X_n}(x) \;=\; \frac{1}{2^{n/2}\,\Gamma\!\left(\dfrac{n}{2}\right)}\, x^{(n/2)-1}\, e^{-x/2}, \quad \text{for } x \geq 0.$$

Remark 1.1
Note that

$$f_{\chi^2_1}(x) \;=\; \frac{1}{2^{1/2}\,\Gamma\!\left(\dfrac{1}{2}\right)}\, x^{(1/2)-1}\, e^{-x/2} \;=\; \frac{1}{\sqrt{2}\,\sqrt{\pi}}\, x^{-1/2}\, e^{-x/2} \;=\; \frac{1}{\sqrt{2\pi}\cdot x^{1/2}\cdot e^{x/2}}$$

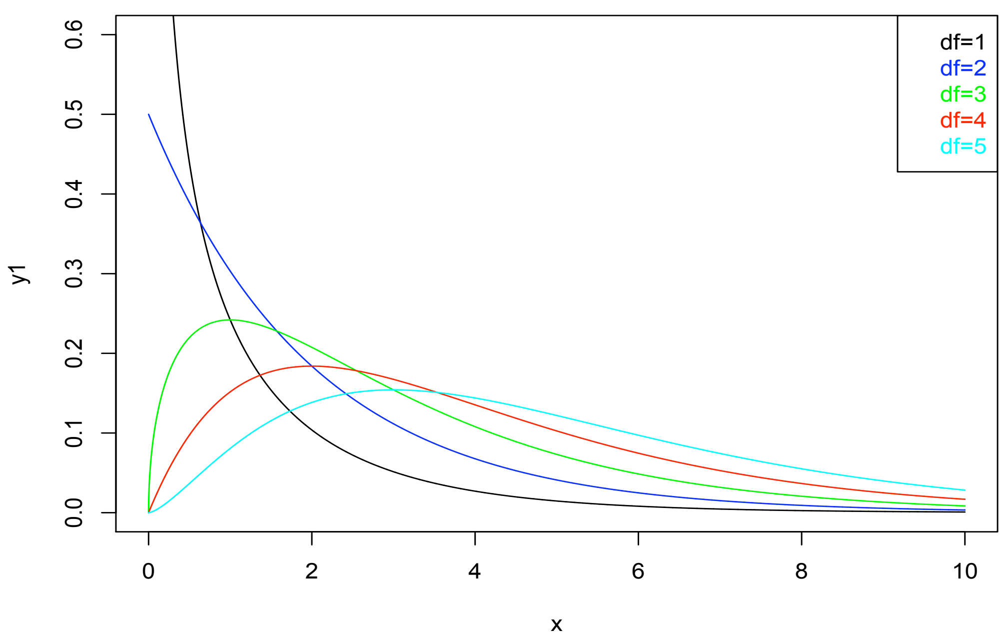
has a singularity at $x = 0$, and

$$f_{\chi^2_2}(x) \;=\; \frac{1}{2^{2/2}\,\Gamma\!\left(\dfrac{2}{2}\right)}\, x^{(2/2)-1}\, e^{-x/2} \;=\; \frac{1}{2\,\Gamma(1)}\, x^0\, e^{-x/2} \;=\; \frac{1}{2}\, e^{-x/2}$$

is simply an exponential decay in $x \geq 0$.

The diagram below shows graphs of the probability density functions $f_{\chi^2_1}, \ldots, f_{\chi^2_5}$. It is generated with `R` with the following commands:

```
> x<-seq(0,10,0.001);
> y1<-dchisq(x,df=1); y2<-dchisq(x,df=2); y3<-dchisq(x,df=3); y4<-dchisq(x,df=4); y5<-dchisq(x,df=5);
> plot(x,y1,ylim=c(0,0.6),type="l"); > points(x,y2,type="l",col="blue"); points(x,y3,type="l",col="green");
> points(x,y4,type="l",col="red");points(x,y5,type="l",col="cyan");
> legend("topright",c("df=1","df=2","df=3","df=4","df=5"),text.col=c("black","blue","green","red","cyan"));
```



2 Reminder: t_n = "would-have-been standard-normal except for the estimated standard-deviation denominator"

- Let $X_1, X_2, \ldots, X_n$ be *i.i.d.* normal random variables with common mean μ and finite variance $\sigma^2 > 0$.

$$\overline{X}_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i\,, \qquad \text{and} \qquad S_n^2 \;:=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

  $\overline{X}_n$ is called the ***sample mean***, and S_n^2 is called the (***unbiased***) ***sample variance***. They are random variables and are unbiased estimators for μ and σ^2, respectively.

- Note that

$$T_{n-1} \;:=\; \frac{\overline{X}_n - \mu}{\sqrt{S_n^2\,/\,n}} \;=\; \frac{\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\left(\dfrac{(n-1)\,S_n^2}{\sigma^2}\right)\Big/\,(n-1)}}.$$

  We claim:

  - The numerator $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ of T_{n-1} has the standard normal distribution.
  - The term $\dfrac{(n-1)\,S_n^2}{\sigma^2} = \displaystyle\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2$ in the denominator of T_{n-1} is a random variable whose distribution is a χ^2 distribution with $(n-1)$ degrees of freedom.
  - The two random variables $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ and $\dfrac{(n-1)\,S_n^2}{\sigma^2}$ are independent of each other.

- **Definition 2.1**
  *Let Z be a standard normal random variable and X be a χ^2 random variable with n degrees of freedom. Suppse Z and X are independent. The **Student t distribution with n degrees of freedom** is the probability distribution of the following random variable:*

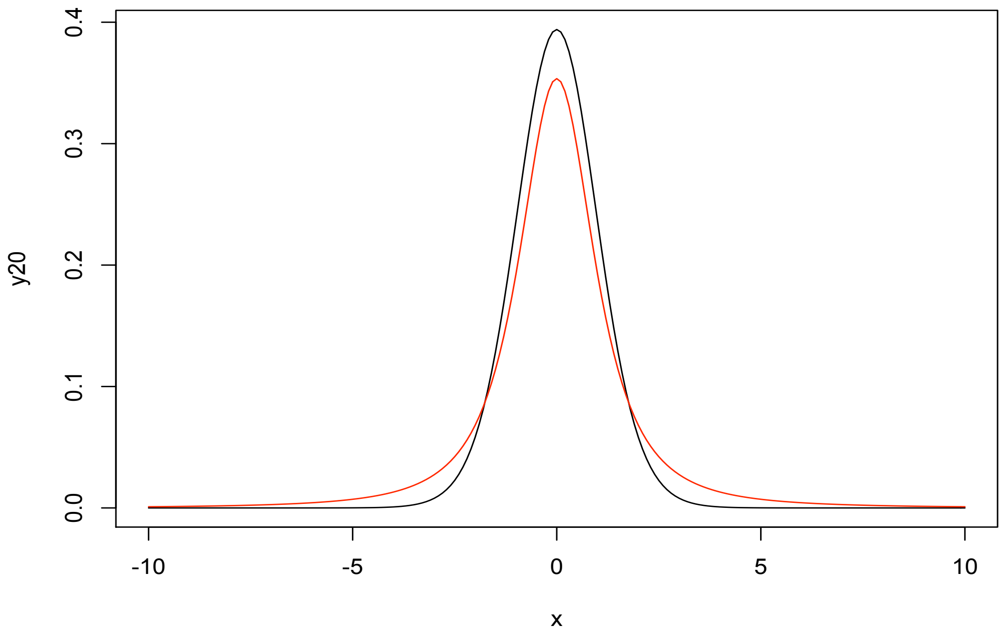
$$T_n \;:=\; \frac{Z}{\sqrt{X/n}}.$$

  *The random variable T_n is called the **Student's t ratio with n degrees of freedom**.*

- **Theorem 2.2**
  The probability density function of the Student t distribution with n degrees of freedom is given by:

$$f_{T_n}(t) \;=\; \frac{\Gamma\!\left(\dfrac{n+1}{2}\right)}{\sqrt{n\,\pi}\,\Gamma\!\left(\dfrac{n}{2}\right)}\cdot\frac{1}{\left(1+\dfrac{t^2}{n}\right)^{(n+1)/2}}\,, \quad \text{for} \quad -\infty < t < \infty.$$

The diagram below shows the graphs of the probability density functions $f_{T_{20}}$ in black and f_{T_2} in red.



The above graph is generated with R with the following command:

```
> y20 = dt(x,df=20); y2 = dt(x,df=2); plot(x,y20,type="l"); points(x,y2,type="l",col="red");
```

3 Reminder: $\mathcal{F}_n^m$ = "ratio of two independent χ^2 random variables"

Definition 3.1
*Let $m, n \in \mathbb{N}$. Let $X_m \sim \chi_m^2$ and $X_n \sim \chi_n^2$ be independent χ^2 random variables with the indicated degrees of freedom. For $m, n \in \mathbb{N}$, the **F distribution with** m **and** n **degrees of freedom**, denoted by $\mathcal{F}_n^m$, is the probability distribution of the following random variable:*

$$F := \frac{X_m/m}{X_n/n}.$$

Theorem 3.2
The probability density function of the F distribution $\mathcal{F}_n^m$ with m and n degrees of freedom is given by:

$$f_{\mathcal{F}_n^m}(\zeta) \;=\; \left(m^{m/2} \cdot n^{n/2} \cdot \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\,\Gamma\left(\frac{n}{2}\right)} \right) \cdot \frac{\zeta^{(m/2)-1}}{(m\,\zeta + n)^{(m+n)/2}}, \quad \text{for } \zeta \geq 0.$$

Remark 3.3
The "F" in "F distribution" commemorates the renowned statistician Sir Ronald Fisher.
Let $T = \dfrac{Z}{\sqrt{X/n}}$ be a Student t ratio, i.e. Z and X are independent random variables with $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi_n^2$.
Then, $Z^2 \sim \chi_1^2$. Hence, $T^2 = \dfrac{Z^2}{X/n} = \dfrac{Z^2/1}{X/n} \sim \mathcal{F}_n^1$, the F distribution with $m = 1$ and n degrees of freedom.

4 Testing: $H_0 : \mu = \mu_0$ — the one-sample t-test for a normal distribution with unknown mean μ

See §2.

5 Testing: $H_0 : \mu_X = \mu_Y$ — the two-sample t-test for two normal distributions with unknown means μ_X and μ_Y but equal (but no-need-to-be-known) variances

Suppose $X_1, \dots, X_n, Y_1, \dots, Y_m$ are independent random variables. Suppose also $X_1, \dots, X_n \sim \mathcal{N}(\mu_X, \sigma_X^2)$, and $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.

Then, $\mathrm{Var}\left(\,\overline{X} - \overline{Y}\,\right) = \mathrm{Var}\left(\,\overline{X}\,\right) + \mathrm{Var}\left(\,\overline{Y}\,\right) = \dfrac{1}{n}\mathrm{Var}(X) + \dfrac{1}{m}\mathrm{Var}(Y) = \dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}$. Hence,

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}} \quad \sim \quad \mathcal{N}(0,1).$$

And,

$$\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)^2 + \sum_{i=1}^{m}\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right)^2 \quad \sim \quad \chi^2_{n-1+m-1} \;=\; \chi^2_{n+m-2}$$

Hence,

$$\frac{\dfrac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}}}{\left(\dfrac{1}{n+m-2}\left\{\displaystyle\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)^2 + \sum_{i=1}^{m}\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right)^2\right\}\right)^{1/2}} \quad \sim \quad t_{n+m-2}$$

Now, **suppose further** $\sigma_X^2 = \sigma_Y^2$. Then,

$$H_0 \,:\, \mu_X = \mu_Y \quad \Longrightarrow \quad \frac{1}{\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} \cdot \frac{\overline{X} - \overline{Y}}{\left(\dfrac{1}{n+m-2}\left\{\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2\right\}\right)^{1/2}} \quad \sim \quad t_{n+m-2}$$

6 Testing: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ — the F-test and ANOVA (analysis of variance)

Suppose we are given $k \in \mathbb{N}$, and $n_1, n_2, \dots, n_k \in \mathbb{N}$. Let $n := n_1 + \cdots + n_k$. Suppose further we are given a doubly indexed set of **independent** random variables

$$\{\; Y_{ij} \mid 1 \le i \le k, 1 \le j \le n_i \;\}.$$

Suppose that each Y_{ij} has the form:

$$Y_{ij} \;=\; \mu_i + \epsilon_{ij},$$

where $\mu_i \in \mathbb{R}$ is constant, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. The *between-cluster sum of squares* SSB and the *within-cluster sum of squares* SSW are defined as follows:

$$\begin{aligned}
\mathrm{SSB} \;&:=\; \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{Y}_i - \overline{Y}\right)^2 \;=\; \sum_{i=1}^{k} n_i\left(\overline{Y}_i - \overline{Y}\right)^2, \\
\mathrm{SSW} \;&:=\; \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2,
\end{aligned}$$

where

$$\overline{Y}_i \;:=\; \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\,, \quad \overline{Y} \;:=\; \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \;=\; \frac{1}{n}\sum_{i=1}^{k} n_i\,\overline{Y}_i\,.$$

Note that

$$\frac{\mathrm{SSW}}{\sigma^2} := \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{Y_{ij}-\overline{Y}_i}{\sigma}\right)^2$$

has a χ^2 distribution of $\sum_{i=1}^{k}(n_i-1) = n-k$ degrees of freedom.

Theorem 6.1 *SSB and SSW are independent random variables.*

Theorem 6.2 *If $\mu_1 = \mu_2 = \cdots = \mu_k$, then*

- $\dfrac{\mathrm{SSB}}{\sigma^2}$ *has a χ^2 distribution of $k-1$ degrees of freedom.*

$$\frac{\mathrm{SSB}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{k} n_i\left(\overline{Y}_i-\overline{Y}\right)^2 = \sum_{i=1}^{k}\frac{1}{\sigma^2/n_i}\left(\overline{Y}_i-\overline{Y}\right)^2 = \sum_{i=1}^{k}\left(\frac{\overline{Y}_i-\overline{Y}}{\sigma/\sqrt{n_i}}\right)^2$$

- *Consequently,*

$$\frac{SSB/(k-1)}{SSW/(n-k)} = \frac{\dfrac{1}{k-1}\displaystyle\sum_{i=1}^{k} n_i\left(\overline{Y}_i-\overline{Y}\right)^2}{\dfrac{1}{n-k}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_i\right)^2} \sim \mathcal{F}_{n-k}^{k-1}$$

7 Testing: $H_0 : \sigma_X^2 = \sigma_Y^2$ — the (two-sample) F-test for two normal distributions

Suppose $X_1,\ldots,X_n,Y_1,\ldots,Y_m$ are independent random variables. Suppose also $X_1,\ldots,X_n \sim \mathcal{N}(\mu_X,\sigma_X^2)$, and $Y_1,\ldots,Y_m \sim \mathcal{N}(\mu_Y,\sigma_Y^2)$.

Let

$$S_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \quad\text{and}\quad S_Y^2 := \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2.$$

Recall that

$$\frac{(n-1)S_X^2}{\sigma_X^2} = \frac{1}{\sigma_X^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \sim \chi_{n-1}^2 \quad\text{and}\quad \frac{(m-1)S_Y^2}{\sigma_Y^2} = \frac{1}{\sigma_Y^2}\sum_{i=1}^{m}\left(Y_i-\overline{Y}\right)^2 \sim \chi_{m-1}^2.$$

Consequently,

$$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} = \frac{\frac{1}{m-1}\sum_{j=1}^{m}\left(Y_j-\overline{Y}\right)^2\Big/\sigma_Y^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\Big/\sigma_X^2} \sim \mathcal{F}_{n-1}^{m-1}$$

Therefore,

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad\Longrightarrow\quad \frac{S_Y^2}{S_X^2} = \frac{\frac{1}{m-1}\sum_{j=1}^{m}\left(Y_j-\overline{Y}\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \sim \mathcal{F}_{n-1}^{m-1}$$

8 χ^2 Goodness-of-fit test — testing goodness-of-fit of a probability model via an induced multinomial model

Suppose $\mathbf{X} = (X_1,X_2,\ldots,X_k)$ has a multinomial distribution with number-of-trials parameter n and probability parameter $\mathbf{p} = (p_1,p_2,\ldots,p_k)$, with $p_i > 0$, for each $i = 1,2,\ldots,k$. In other words, $\mathbf{X} \sim \mathrm{Multinomial}(n,\mathbf{p})$, or equivalently, $(X_1,X_2,\ldots,X_k) \sim \mathrm{Multinomial}(n,(p_1,p_2,\ldots,p_k))$. This simply means:

$$\mathrm{Prob}(\,X_1 = m_1,\, X_2 = m_2,\,\ldots,\, X_k = m_k\,) = \frac{n!}{m_1!\,m_2!\,\cdots\,m_k!}\,p_1^{m_1}\,p_2^{m_2}\,\cdots\,p_k^{m_k}.$$

Note that the random variables $X_1,X_2,\ldots,X_k$ are subject to the restriction: $\sum_{i=1}^{k} X_i = n$.

Theorem 8.1 *The random variable*

$$D := \sum_{i=1}^{k} \frac{(X_i - n\,p_i)^2}{n\,p_i}$$

has ***approximately*** *a* χ^2_{k-1} *distribution.*
(That the degree of freedom is $k-1$*, rather than* k*, is a manifestation of the restriction that* $\sum_{i=1}^{k} X_i = n$*.)*

Suppose $\mathbf{X} = (\,X_1, X_2, \ldots, X_k\,)$ has a multinomial distribution with number-of-trials parameter n and probability parameter $\mathbf{p}(\Theta) = (p_1(\Theta), p_2(\Theta), \ldots, p_k(\Theta))$, with $p_i(\Theta) > 0$, for each $i = 1, 2, \ldots, k$, where $\Theta = (\theta_1, \theta_2, \ldots, \theta_s) \in \mathbb{R}^s$ is a vector parameter of the multinomial model. Let $\widehat{p}_1 := p_1\left(\widehat{\Theta}_{\mathrm{MLE}}\right)$, $\widehat{p}_2 := p_2\left(\widehat{\Theta}_{\mathrm{MLE}}\right)$, $\ldots$, $\widehat{p}_k := p_k\left(\widehat{\Theta}_{\mathrm{MLE}}\right)$.

Theorem 8.2 *The random variable*

$$D_1 := \sum_{i=1}^{k} \frac{(X_i - n\,\widehat{p}_i)^2}{n\,\widehat{p}_i}$$

has ***approximately*** *a* χ^2_{k-s-1} *distribution.*

References