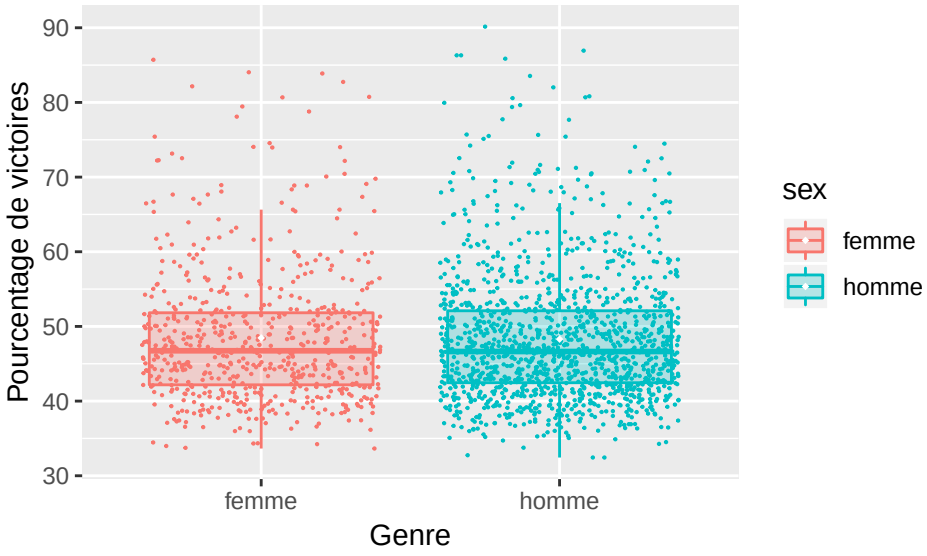
Projet STAT

You ZUO, Clément MONNOT

2019/5/26

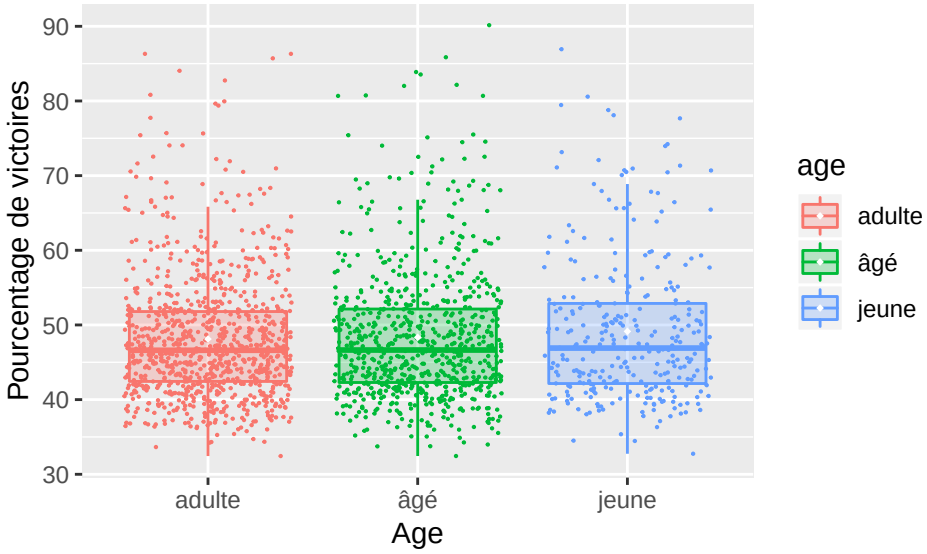
Statistiques Descriptives

<-1.a->

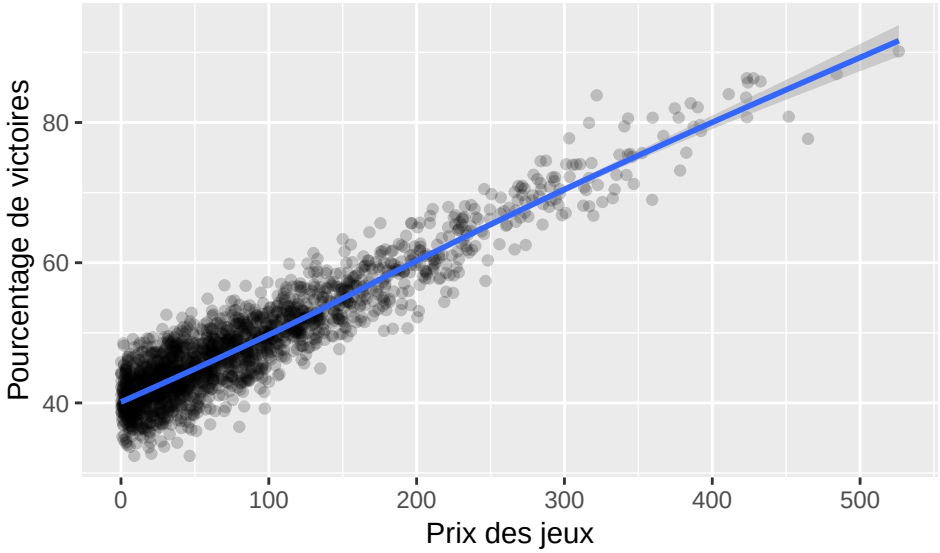


D'après le graphique ci-dessus, on peut voir que les pourcentages de victoires des hommes et des femmes sont assez proches mais que la quantité de joueurs masculins est plus grande que celle des joueurs féminins.

<-1.b->

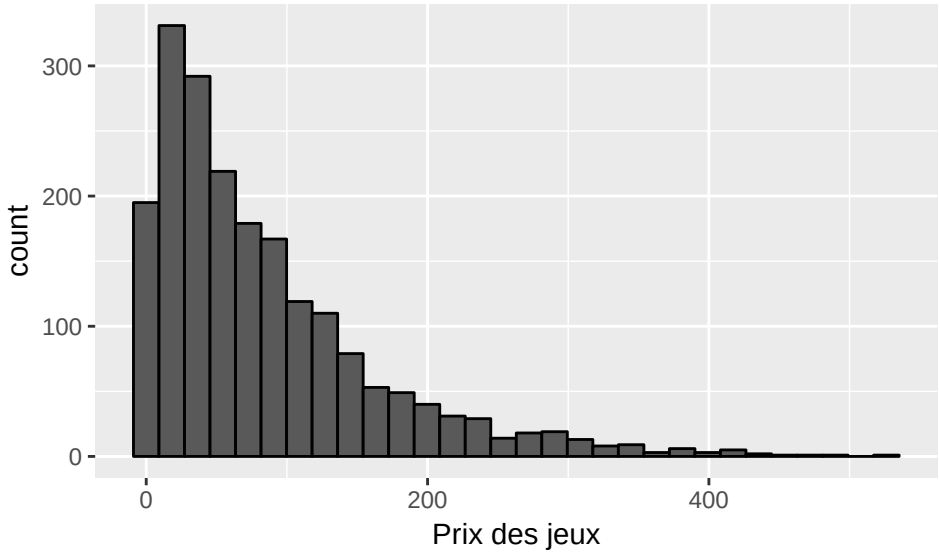


On observe une moyenne du pourcetage de victoires légèrement plus élevée chez les jeunes malgré une population plus faible par rapport aux adultes et aux personnées âgées.



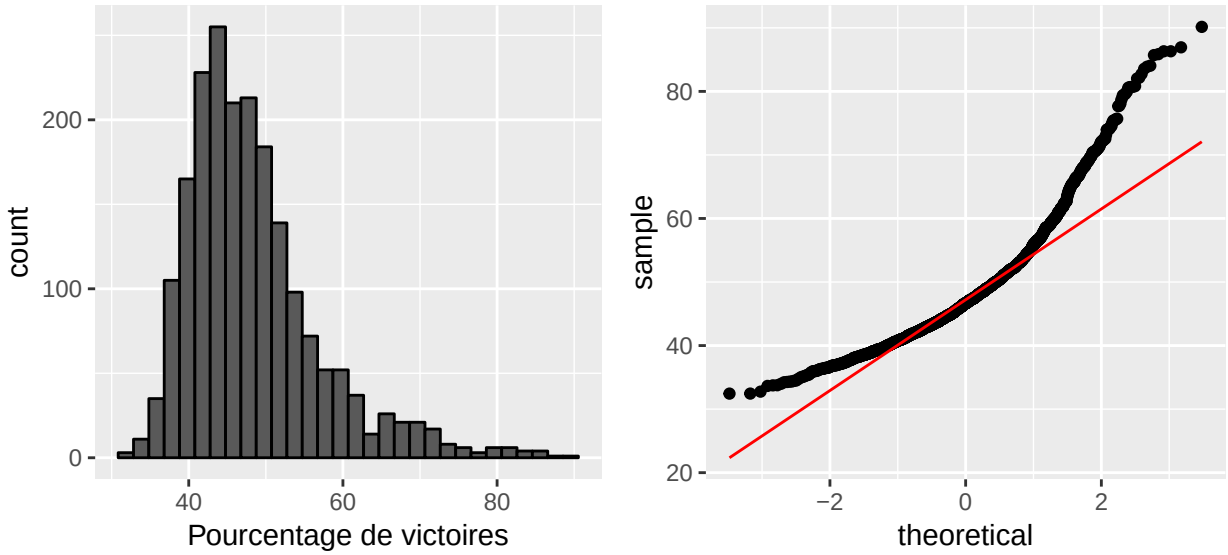
On peut voir que la pourcentage de victoires augmente de manière, a priori, proportiennelle par rapport aux prix des jeux.

<-1.c->



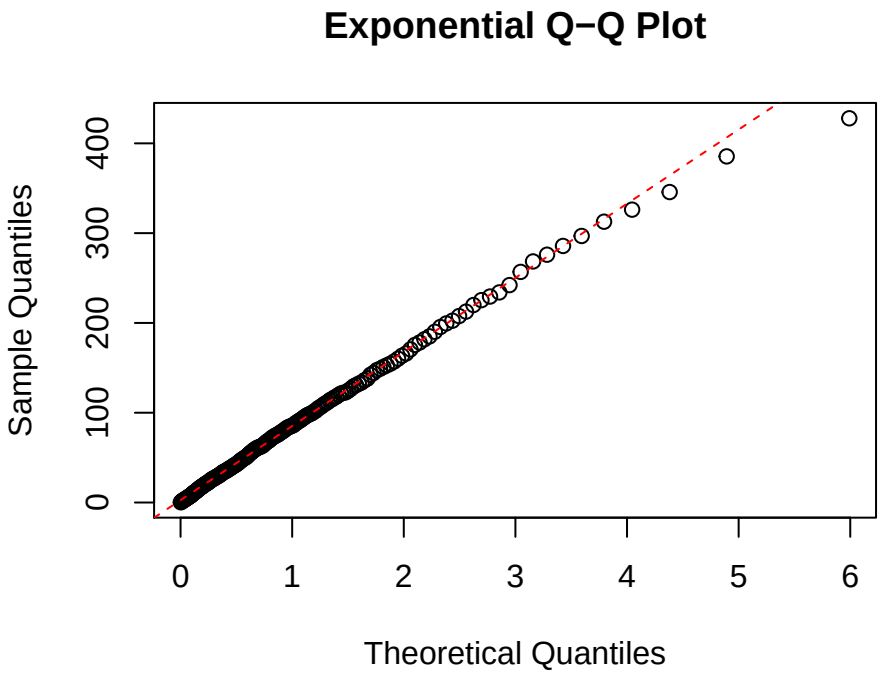
Selon le premier graphique, on peut supposer que le prix du jeu suit une loi exponentielle avec deux parametres car la distribution des échantillons montre une décroissance exponentielle à partir d'un certain point strictement positif.

<-1.d->



Selon la tendance de sa distribution, on a supposé qu'il obéissait à la distribution normale, mais après le test du quantile-quantile, il constate qu'il n'obéit pas à la distribution normale.

Estimation



Afin de vérifier si l'échantillon obéit à la distribution exponentielle, on a effectué un test quantile-quantile. Comme le montre la figure ci-dessus, à l'exception des derniers points qui s'écartent de la ligne standard, les autres valeurs sont une bonne indication que notre échantillon obéit à la distribution exponentielle.

<-2.b->

Pour les échantillons $(X_1, ..., X_n)$ qui suivent une loi exponentielle la vraisemblance du modèle est

$$\mathcal{L}(\lambda, x_0|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_n) = \prod_{i=1}^{n} \lambda\ e^{-\lambda(x_i - x_0)} 1_{[x_0, +\infty[}(x_i)$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n}(x_i - x_0)} \prod_{i=1}^{n} 1_{[x_0, +\infty[}(x_i)$$

la log-vraisemblance du modèle est

$$log\mathcal{L}(\lambda, x_0|x_1, ..., x_n) = nlog(\lambda) - \lambda \sum_{i=1}^{n}(x_i - x_0) + \sum_{i=1}^{n} log(1_{[x_0, +\infty[}(x_i))$$

<-2.c->

Pour le paramètre λ, soit

$$\frac{\partial log\mathcal{L}(\lambda, x_0|x_1, ..., x_n)}{\partial \lambda} = 0$$

Alors, on a

$$\frac{n}{\lambda} - \sum_{i=1}^{n}(x_i - x_0) = 0$$

donc, finalement, l'estimateur de λ est

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}(x_i - x_0)}$$

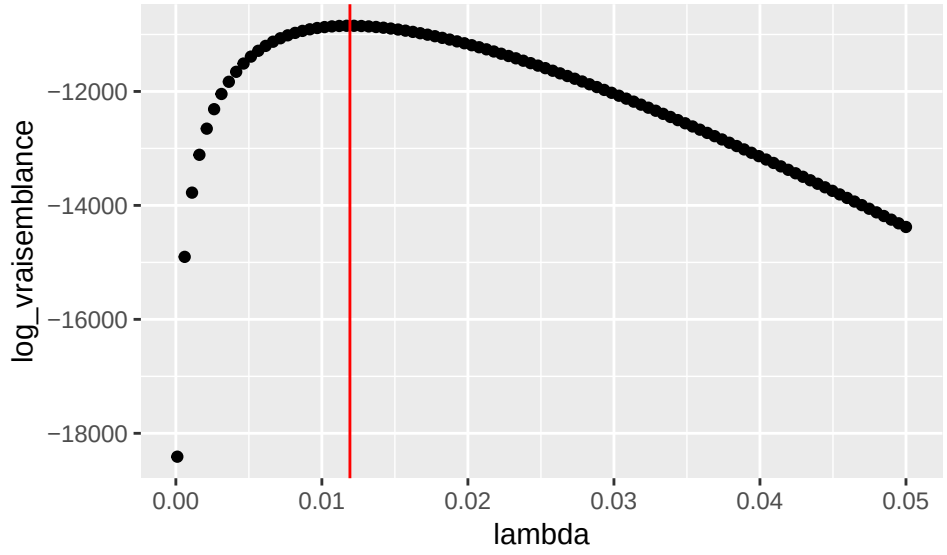
Pour le paramètre x_0, $\mathcal{L}(\lambda, x_0|x_1, ..., x_n)$ est maximum si

$$\prod_{i=1}^{n} 1_{[x_0, +\infty[}(x_i) = 1$$

donc $\forall i = 1, ..., n\ x_i > x_0$ d'où l'estimateur de x_0 :

$$\hat{x}_0 = min_i(x_i)$$

On va tracer la log-vraisemblance de l'échantillon en fonction de la valeur du paramètre λ



<-2.d->

```
##
## Call:
## stats4::mle(minuslogl = (logvrsb_lambda), start = list(lambda = 1e-04),
##     method = "L-BFGS-B", lower = 1e-04, upper = 0.05)
##
## Coefficients:
##     lambda
## 0.01192053
```

<-2.e->

On suppose que le pourcentage de victoires d'un joueur suit une loi Normale $\mathcal{N}(40 + \frac{P}{10}, \sigma^2)$, où P est le prix de son jeu. Pour les échantillons du pourcentage de victoires d'un joueur $V_1, V_2, ..., V_n$ la vraisemblance théorique du modèle du pourcentage de victoires est

$$\mathcal{L}(\sigma|v_1, ..., v_n) = \prod_{i=1}^{n} f(v_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(v_i - 40 - \frac{p_i}{10})^2}{2\sigma^2})$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(v_i - 40 - \frac{p_i}{10})^2)$$

la log-vraisemblance est

$$log\mathcal{L}(\sigma|v_1, ..., v_n) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(v_i - 40 - \frac{p_i}{10})^2$$

<-2.f->

Pour déterminer le maximum de vraisemblance du paramètre σ, on a

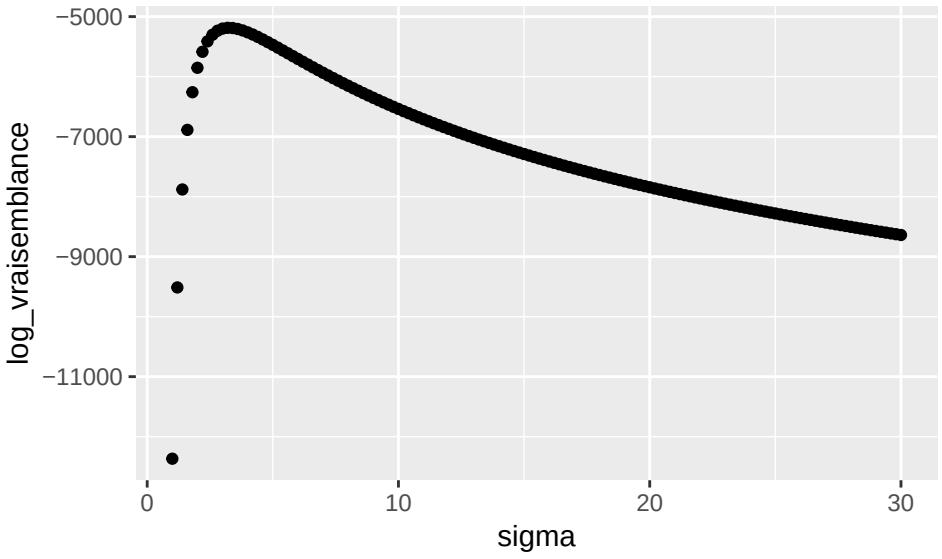
$$\frac{\partial log\mathcal{L}(\sigma|v_1, ..., v_n)}{\partial\sigma^2} = 0$$

donc on a

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(v_i - 40 - \frac{p_i}{10})^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(v_i - 40 - \frac{p_i}{10})^2$$

On va tracer la log-vraisemblance de l'échantillon en fonction de la valeur du paramètre σ



Intervalles de confiance

<-3.a->

A la question précédente, on a vérifié que les échantillons suivent une loi exponentielle $\mathcal{E}(\lambda, x_0)$. Dans ce cas là, on a une statistique $\gamma_n = \lambda n \bar{X}_n$ qui suit une loi $\Gamma(1, n)$ donc selon la fonction de densité de $\gamma(x)$ on peut obtenir

$$P_\lambda([\gamma_{\frac{\alpha}{2}}(n) < \lambda n(\bar{X}_n - x_0) < \gamma_{1-\frac{\alpha}{2}}(n)]) \longrightarrow 1 - \alpha$$

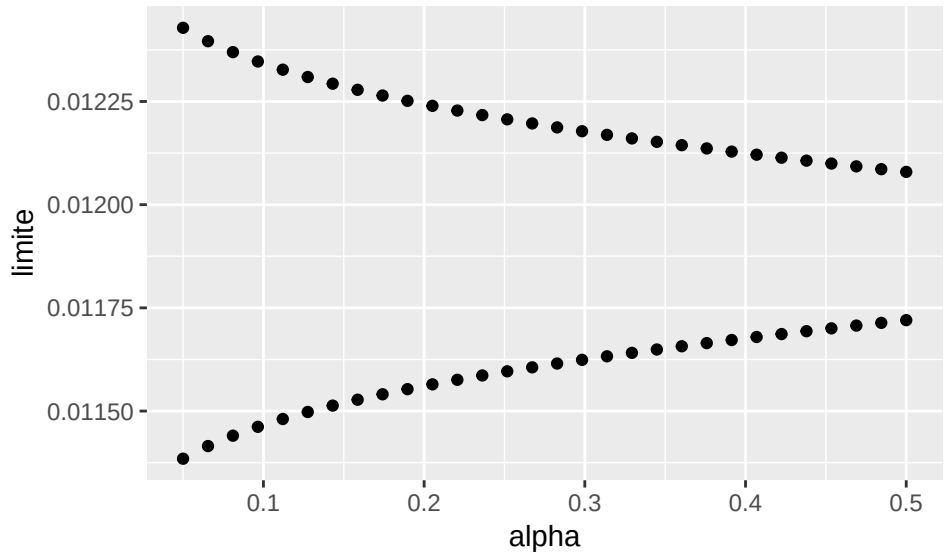
lorsque n tend vers l'infini.

Donc on a l'intervalle de confiance asymptotique pour λ qui est

$$[\frac{\gamma_{\frac{\alpha}{2}}(n)}{n(\bar{X}_n - x_0)}, \frac{\gamma_{1-\frac{\alpha}{2}}(n)}{n(\bar{X}_n - x_0)}]$$

<-3.b->

En faisant varier $\alpha \in \{5\%, ..., 50\%\}$, on obtient



On remarque que plus α est petit, plus l'intervalle de confiance augmente, c'est-à-dire, plus on a de chance que λ soit dans l'intervalle de confiance.

Tests

<-4.a->

Soit la moyenne du pourcentage de victoires des hommes μ_h et en fixant $\alpha = 5\%$, on fait donc les hypothèses suivantes

$$\begin{cases} H_0 : \ \mu_h \leq 50\% \\ H_1 : \ \mu_h > 50\% \end{cases}$$

On a des hypothèses similaires pour les échantillons féminins.

$$\begin{cases} H_0 : \ \mu_f \leq 50\% \\ H_1 : \ \mu_f > 50\% \end{cases}$$

Les échantillons du pourcentage de victoires que nous avons testés auparavant ne suivent pas une distribution normale, on ne peut donc pas utiliser le test de Student. Donc on choisit d'utiliser le test de Wilcoxon.

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  vic_homme[, "victory"]
## V = 303400, p-value < 2.2e-16
## alternative hypothesis: true location is less than 50
```

D'après le resultat, on peut voir que $p - value < \alpha$, ce qui signifie qu'on peut rejeter l'hypothèse H_0.

On fait la même chose pour des échantillons du pourcentage de victoires des femmes

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  vic_femme[, "victory"]
## V = 65284, p-value = 6.334e-14
## alternative hypothesis: true location is less than 50
```

D'après le resultat, on peut voir que $p-value < \alpha$, ce qui signifie qu'on peut rejeter l'hypothèse H_0. Donc le pourcentage de victoires obtenue par les femmes et les hommes sont, en moyenne, tous les deux inférieurs à 50%.

<-4.b->

Nous souhaitons comparer le pourcentage de victoires des femmes avec celui des hommes pour voir si ils ont des différences plus évidentes. Donc on propose un test d'indépendance: Soit(S, V) un couple de v.a dans $\{s_0, s_1\} \times \{v_1, ..., v_n\}$, où S signifie le genre et V le pourcentage de victoires. Soit $((S_i; V_i); i \leq i \leq n)$ un n-échantillon de (S, V). On note $p_{ij} = P_{H_0}(S_1 = s_i, V_1 = v_j)$, et les marginales

$$p_{i.} = \sum_{j=v_1}^{v_n} p_{ij}, \quad p_{.j} = \sum_{i=s_1}^{s_2} p_{ij}$$

On souhaite vérifier si les variables S et V sont indépendantes. On fait les hypothèses suivantes

$$\begin{cases} H_0 : \ p_{ij} = p_{i.}p_{.j} \\ H_1 : \ p_{ij} \neq p_{i.}p_{.j} \end{cases}$$

On note les occurrences

$$N_{ij} = \sum_{k=1}^{n} 1_{S_k=s_i, V_k=v_j}; \ N_{i.} = \sum_{k=1}^{n} 1_{S_k=s_i}; \ N_{.j} = \sum_{k=1}^{n} 1_{V_k=v_j}$$

La statistique du test est:

$$D_n = \sum_{i=1}^{2}\sum_{j=1}^{n} D_{ij} = \sum_{i=1}^{2}\sum_{j=1}^{n} \frac{(N_{ij} - n\hat{p}_{i.}\hat{p}_{.j})^2}{n\hat{p}_{i.}\hat{p}_{.j}} = \sum_{i=1}^{2}\sum_{j=1}^{n} \frac{(N_{ij} - \frac{N_{i.}N_{.j}}{n})^2}{\frac{N_{i.}N_{.j}}{n}}$$

Sous H_0, D_n suit une loi de $\mathcal{X}^2_{(n-1)(2-1)}$, donc on a la zone de rejet

$$\mathcal{W} = \{D > F^{-1}_{\mathcal{X}^2_{(n-1)(2-1)}}(1-\alpha)\}$$

```
##
##  Pearson's Chi-squared test
##
## data:  tmp
## X-squared = 1997, df = 1996, p-value = 0.4895
```

$p-valeur > \alpha$ donc on ne peut pas rejeter l'hypothèse H_0, c'est-à-dire qu'il n'existe pas une grande différence entre le pourcentage de victoires des hommes avec celui des femmes.

<-4.c->

On suppose que les observations du pourcentage de victoires des jeux chers (dont la valeur est supérieure au quantile 70%) $v_{1_c}, ..., v_{n_c}$ sont issus d'une distribution normale avec un écart-type $\sigma_0 = 5$. On le vérifie par le test de $\mathcal{X}^2$. On a les hypothèses

$$\begin{cases} H_0 : \ \sigma^2 = \sigma_0^2 \\ H_1 : \ \sigma^2 \neq \sigma_0^2 \end{cases}$$

ici, la moyenne μ est inconnue donc on la remplace par la moyenne empirique $\bar{v}_{n_c}$ Sous l'hypothèse H_0, on sait que

$$\mathcal{X}^2 = \frac{(n_c - 1)S^{*2}_n}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(v_{n_c} - \bar{v}_{n_c})^2}{\sigma_0^2}$$

la zone de rejet

$$\mathcal{W} = \{(v_{1_c}, ..., v_{n_c}); \mathcal{X}^2 \geq \mathcal{X}^2_{1-\frac{\alpha}{2}}(n_c - 1) \; ou \; \mathcal{X}^2 \leq \mathcal{X}^2_{\frac{\alpha}{2}}(n_c - 1)\}$$

En calculant la valeur de notre statistique selon nos observations d'échantillons, on obtient

$$\mathcal{X}^2 = 1683.108$$

En prenant $n_c = 599$, on obtient le quantile de $\mathcal{X}^2(598) = 1883$ qui est 1, c'est à dire qu'on doit rejeter l'hypothèse à partir de $\alpha = 0$.