Usando dados da RAIS e Análise de Sobrevivência para entender desemprego

Pedro Cavalcante

2018-10-07

Negros estão mais sujeitos à rotatividade de trabalhos? Se sim, isso se explica por variáveis observáveis como escolaridade ou não? E mulheres? Essas são questões muito comuns entre economistas do trabalho e podem ser atacadas de várias maneiras. Uma delas, que eu acho particularmente interessante, é com Análise de Sobrevivência.

Análise de Sobrevivência é um termo bem amplo para descrever modelos que servem para explorar tempo até que um evento de interesse aconteça. Até onde eu sei esse tipo de técnica nasceu em pesquisas clínicas, para melhor entender efeitos de certos tratamentos contra câncer. Hoje é aplicado por cientistas sociais em análise de eventos, por engenheiros para entender melhor falha e confiabilidade de sistemas e por economistas, principalmente para estudar desemprego.

Curvas de Kaplan-Meier, um pouco de teoria

A função de sobrevivência, doravante $S(t)$, é um mapa que relaciona momento de tempo t à probabilidade de *não* acontecimento de um evento. A função *hazard* - acho que "risco" seja uma tradução apropriada? - relaciona a probabilidade de um evento acontecer no momento t. Esse evento pode ser morte do paciente, uma revolução, falha de um sistema mecânico ou, no nosso caso, desemprego.

Uma das ferramentas iniciais de Análise de Sobrevivência é a Curva de Kaplan-Meier. A Curva KM tem a seguinte forma funcional:

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

Onde n_i é o número de empregados até t_i, d_i é o número de demissões em t_i. Antes de computar isso, vamos explorar nossa amostra.

Amostra

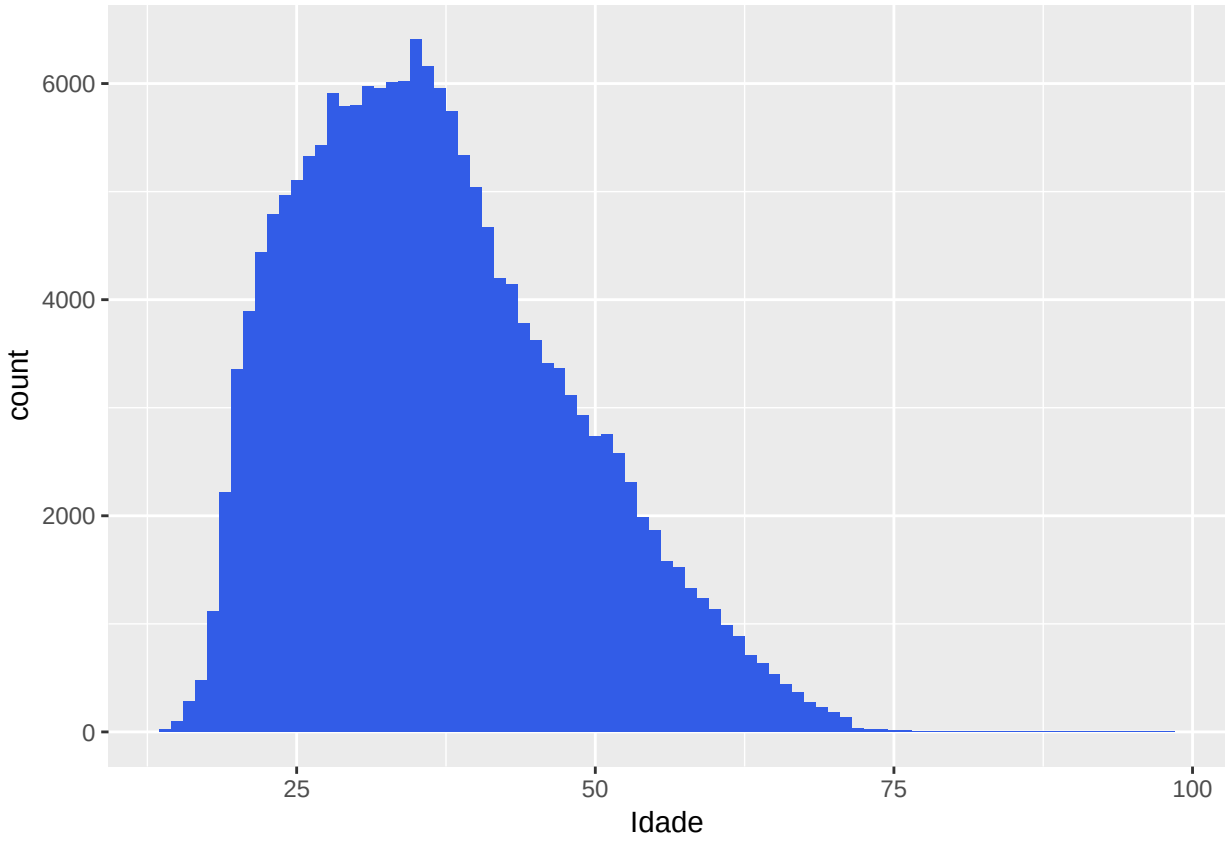
Vamos usar dados anonimizados da RAIS de 2017, mais especificamente do Acre. Já tive o trabalho de limpa-los e deixei o arquivo `.Rds` disponível no repositório do AZUL no github. Você pode puxar os dados diretamente do repositório para o R e deixo como exercício ao leitor o código que faz isso (se quiser o código porque não está conseguindo eu estou sempre disponível).

Vamos explorar a amostra.

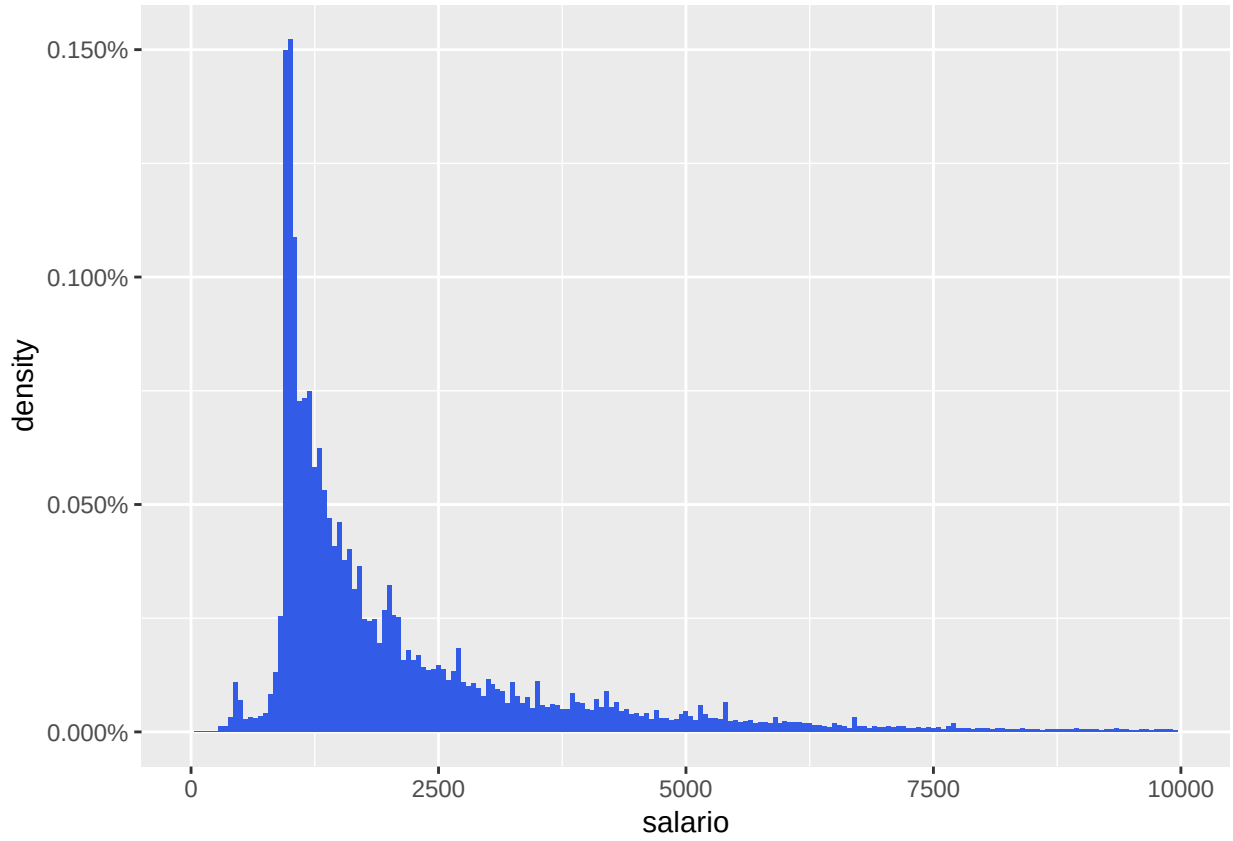
```
library(ggplot2)
library(dplyr)
library(scales)

dim(dados)
```

```
## [1] 177358     54
```

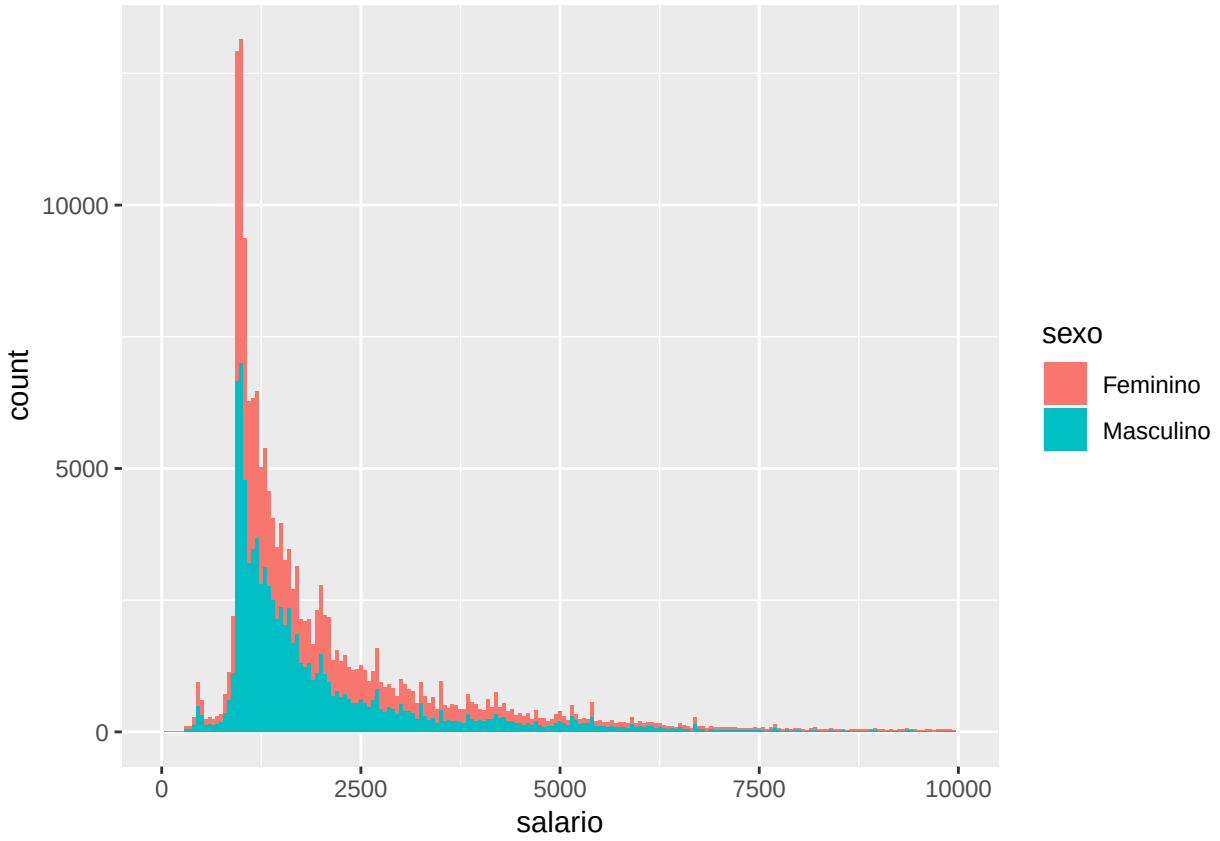
```
dados %>%
  ggplot(aes(x = Idade)) +
  geom_histogram(fill = "#325ce7", binwidth = 1)
```



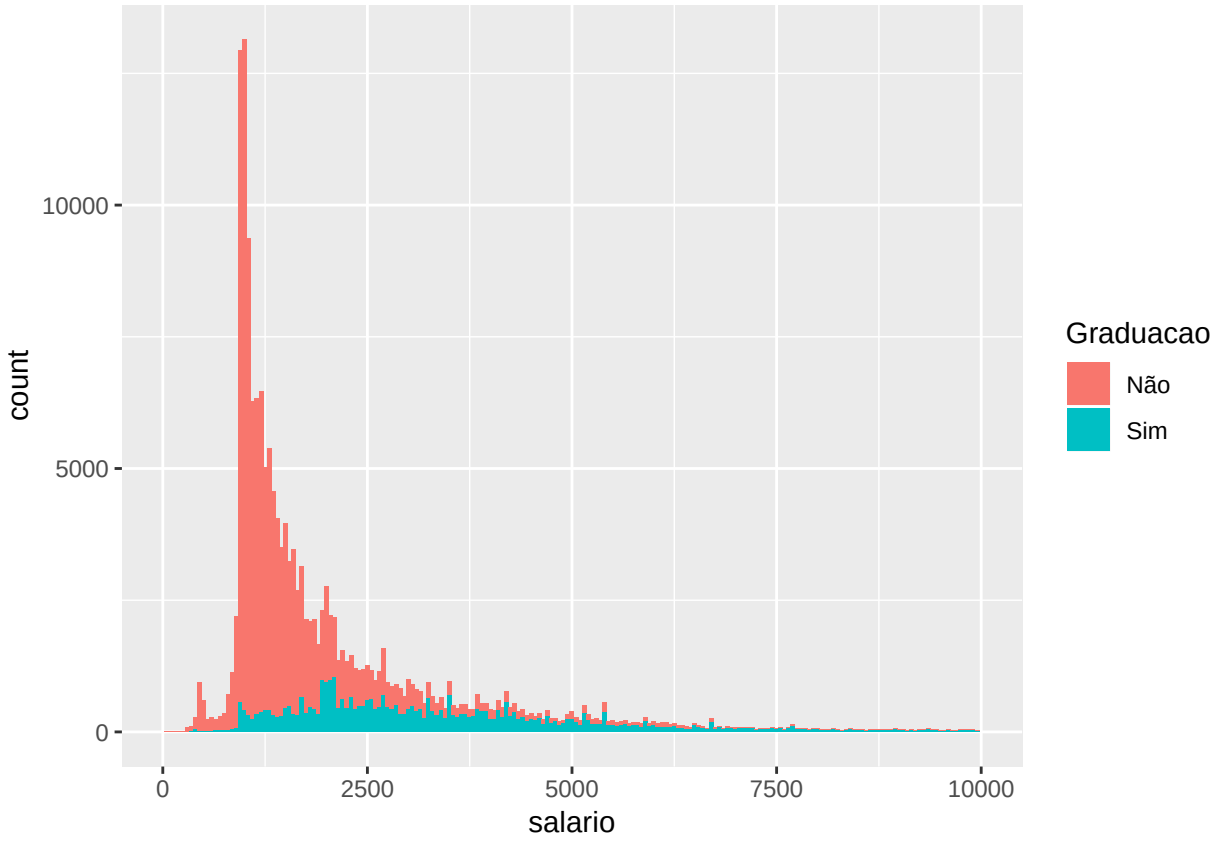

```
dados %>%
  ggplot(aes(x = salario)) +
  geom_histogram(aes(y=..density..), fill = "#325ce7", binwidth = 50) +
  scale_y_continuous(labels = percent) +
  xlim(0, 10000)
```



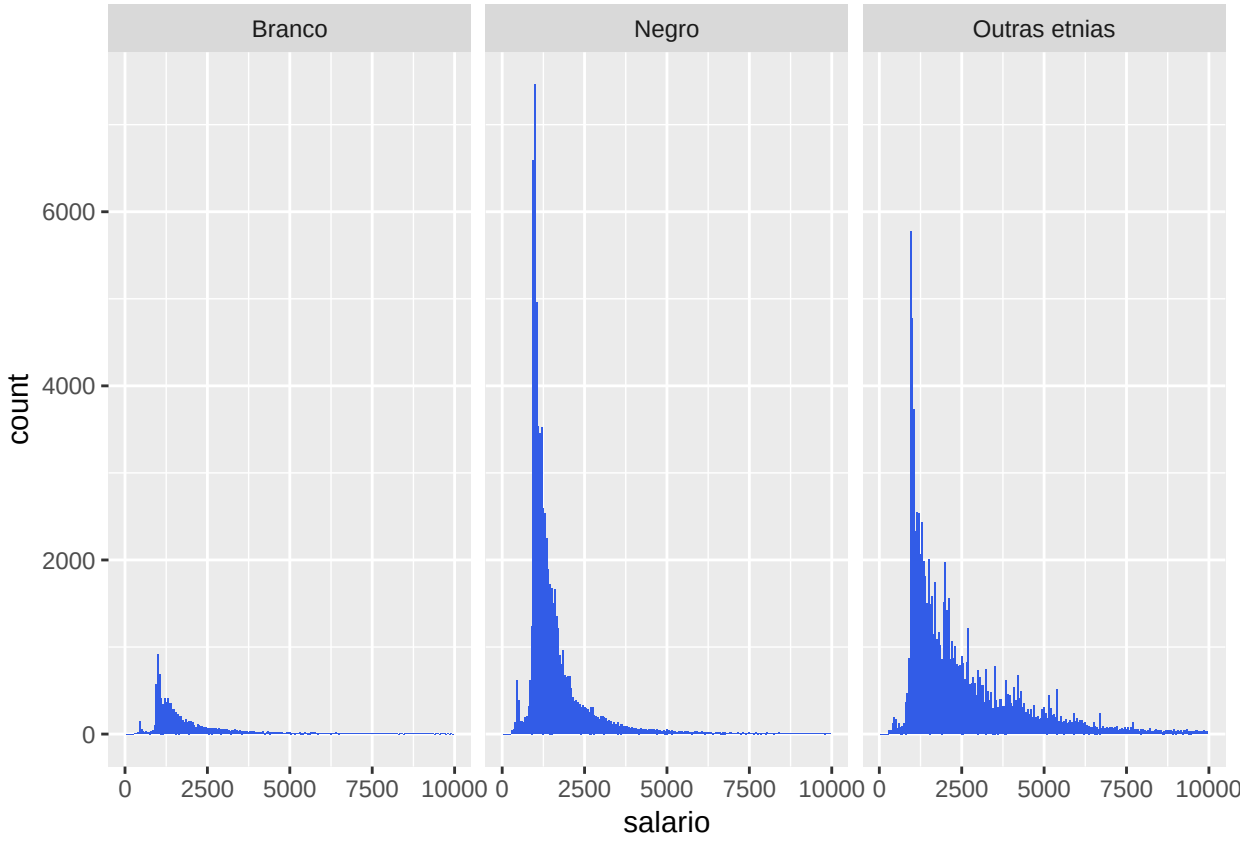

```
dados %>%
  ggplot(aes(x = salario, fill = sexo)) +
  geom_histogram(binwidth = 50) +
  xlim(0, 10000)
```



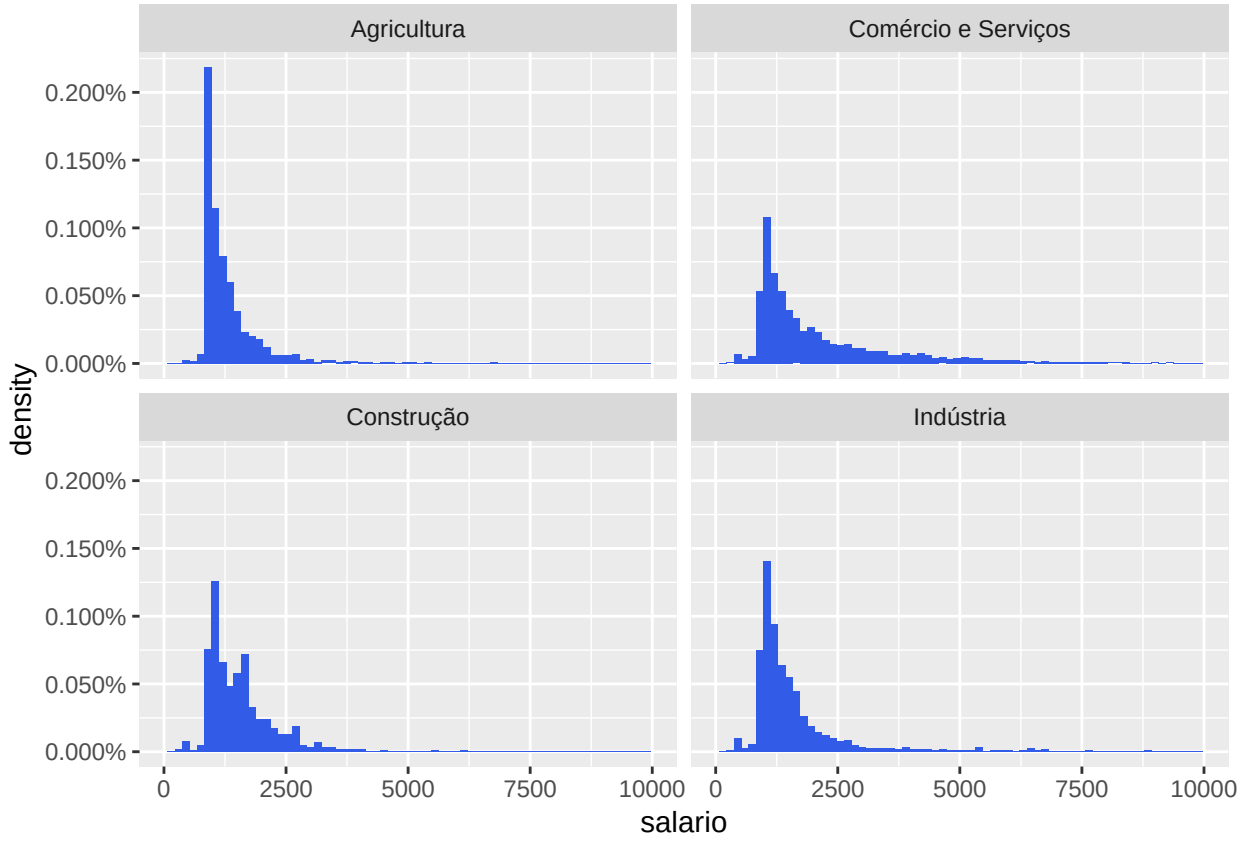

```
dados %>%
  ggplot(aes(x = salario, fill = Graduacao)) +
  geom_histogram(binwidth = 50) +
  xlim(0, 10000)
```




```
dados %>%
  ggplot(aes(x = salario)) +
  geom_histogram(fill = "#325ce7", binwidth = 50) +
  xlim(0, 10000)+
  facet_wrap(~etnia)
```

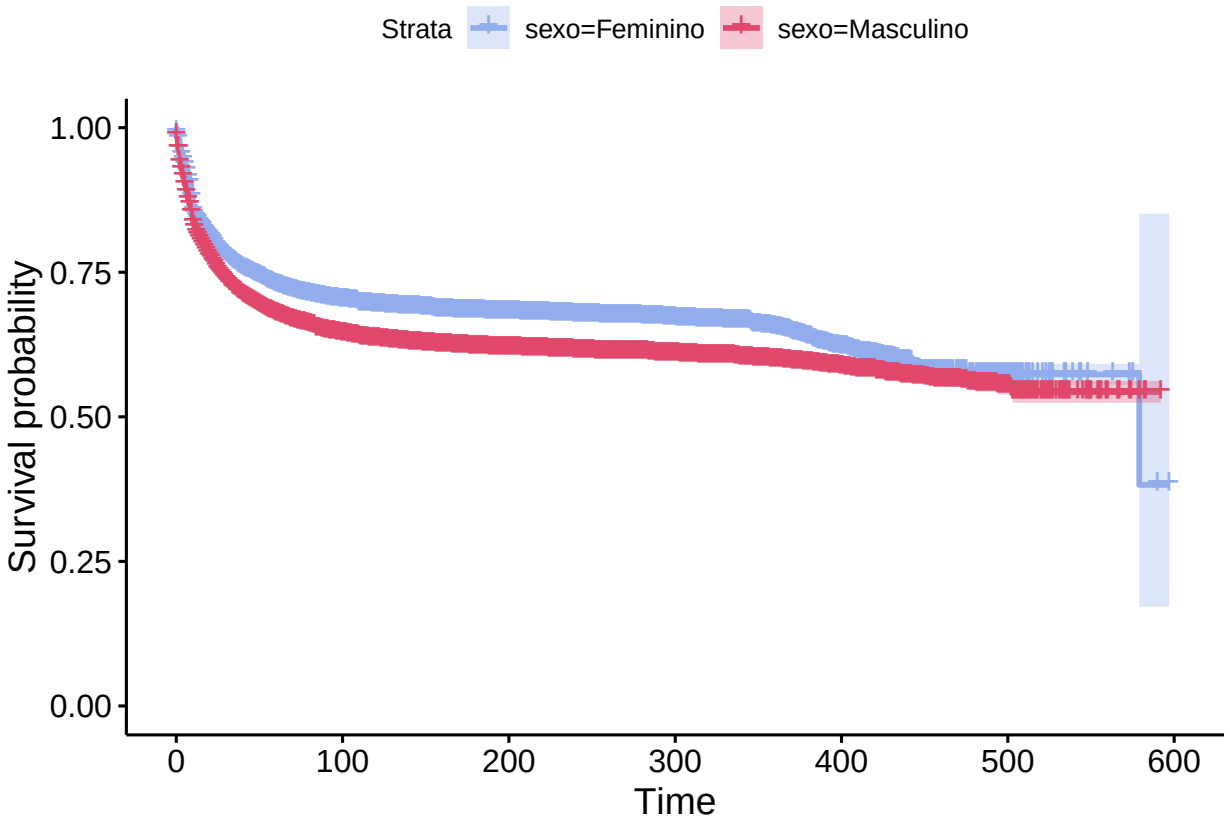



```
dados %>%
  ggplot(aes(x = salario)) +
  geom_histogram(aes(y=..density..), fill = "#325ce7", binwidth = 150) +
  scale_y_continuous(labels = percent) +
  xlim(0, 10000)+
  facet_wrap(~setor)
```

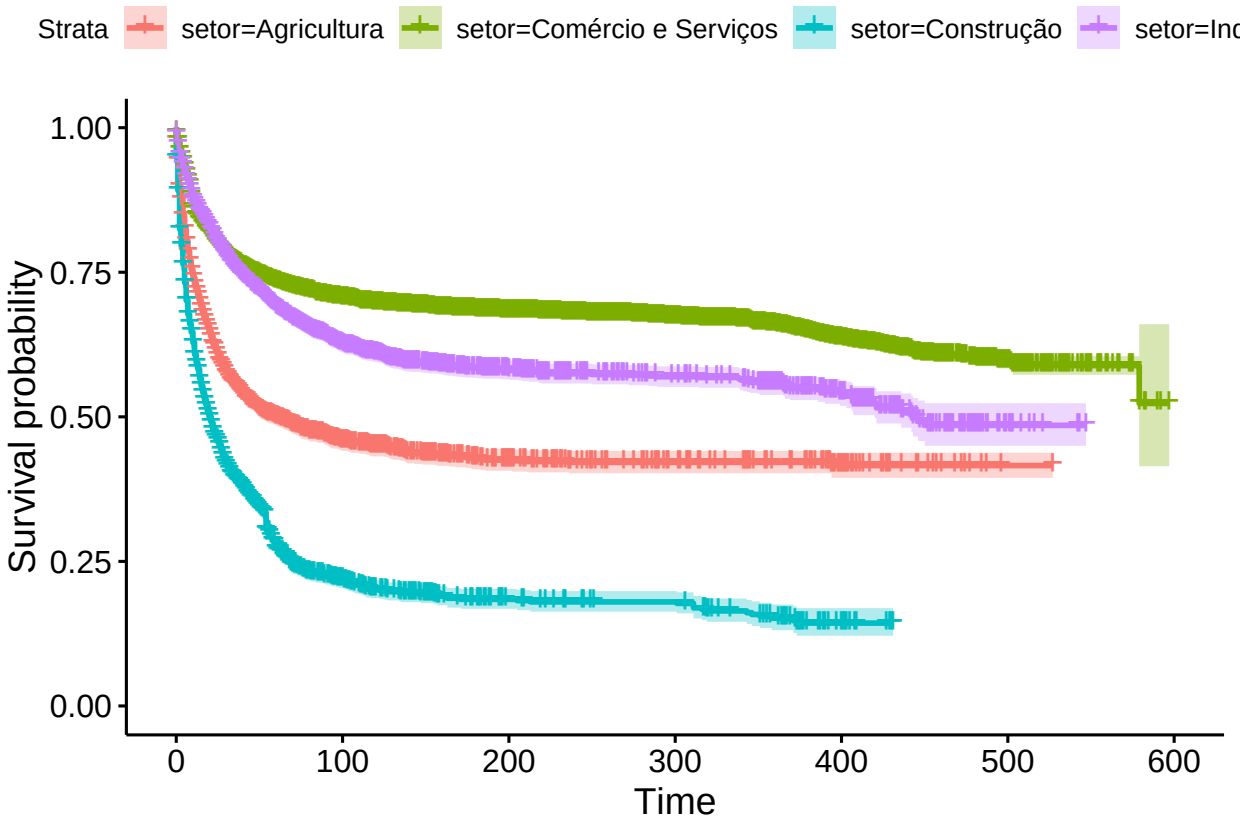


Agora podemos começar a brincar mais e tentar encaixar curvas de sobrevivência aqui. Temos ferramentas para estima-las no pacote `survival` e podemos visualiza-las com o pacote `survminer` que implementa uma viz baseada em `ggplot2`.

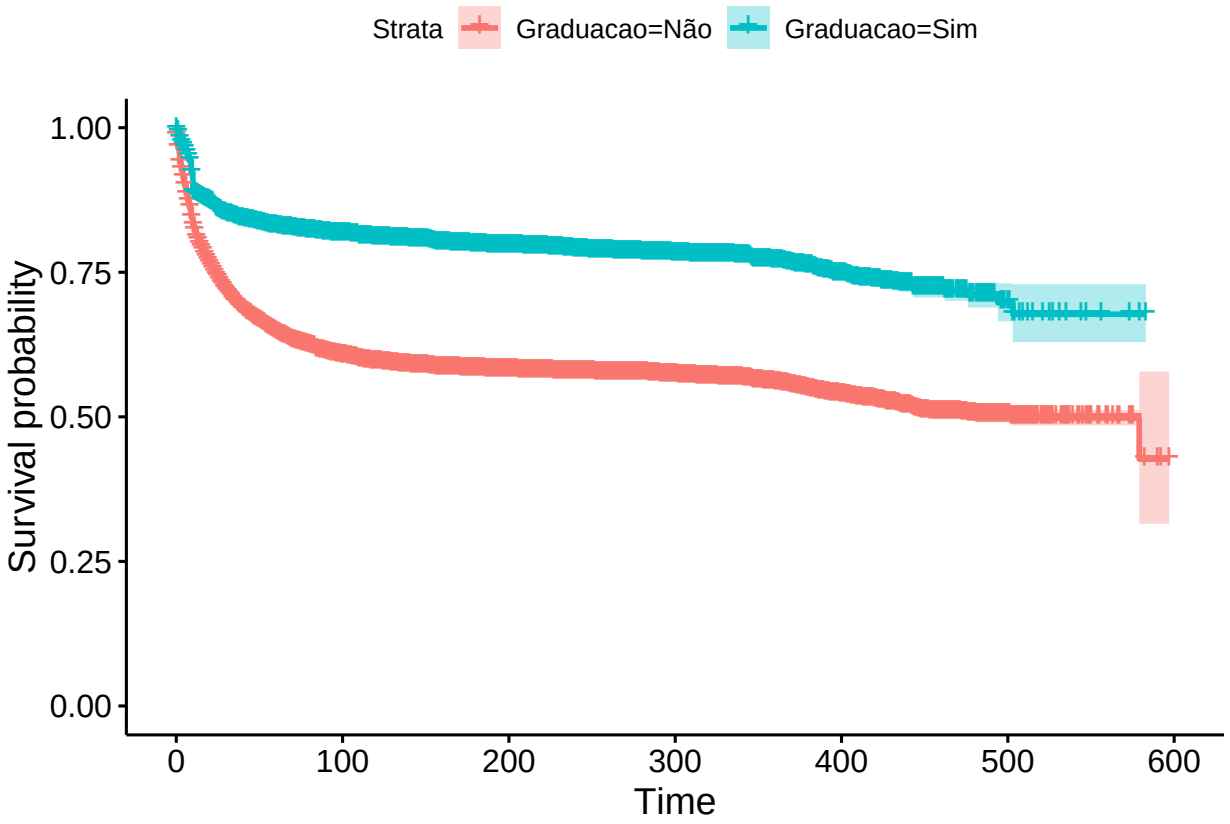
```
library(survival)

fit_sexo = survfit(Surv(tempo_emprego, demissao) ~ sexo, data = dados)
fit_setor = survfit(Surv(tempo_emprego, demissao) ~ setor, data = dados)
fit_ensinosuperior = survfit(Surv(tempo_emprego, demissao) ~ Graduacao, data = dados)
fit_etnia = survfit(Surv(tempo_emprego, demissao) ~ etnia, data = dados)

library(survminer)

ggsurvplot(fit_sexo, conf.int = TRUE,
          palette = c("#91aded", "#e1476b"))
```



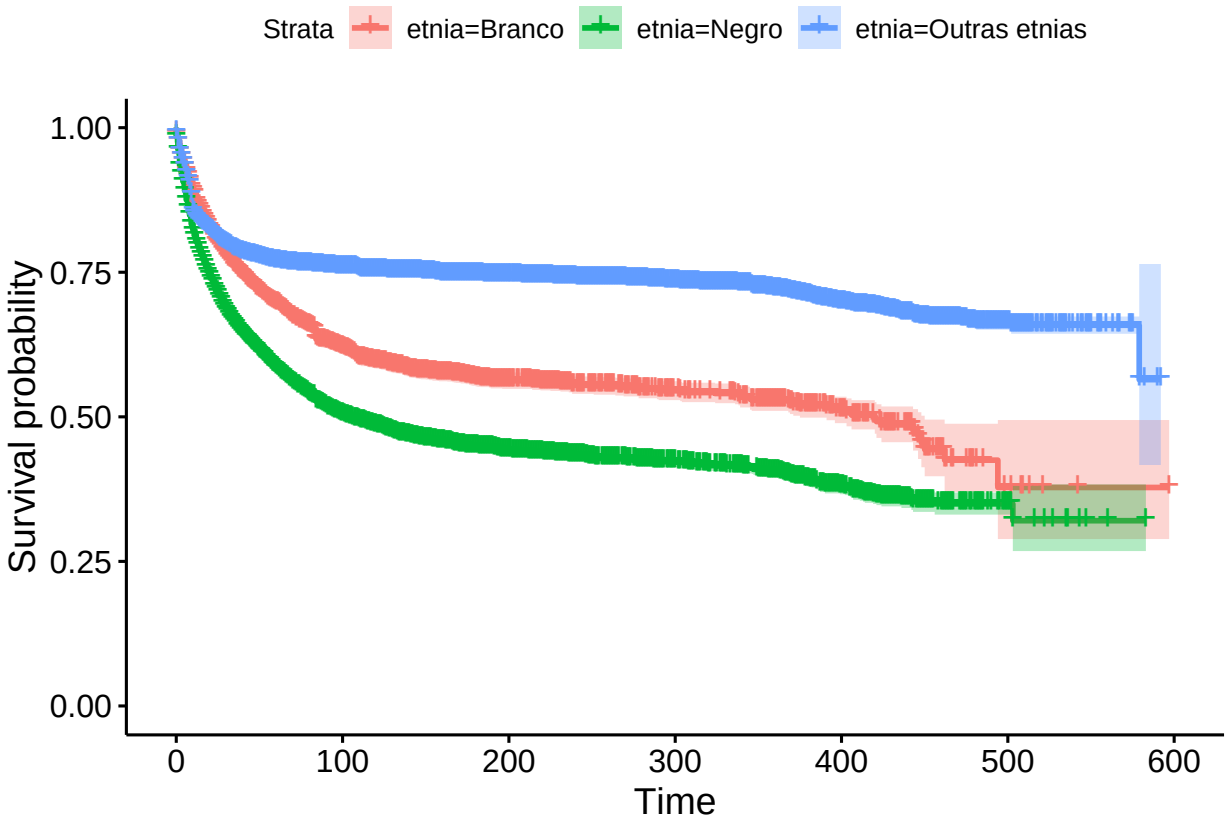

```
ggsurvplot(fit_setor, conf.int = TRUE)
```




```
ggsurvplot(fit_ensinosuperior, conf.int = TRUE)
```




```
ggsurvplot(fit_etnia, conf.int = TRUE)
```



Como esperado, trabalhadores agrícolas tem empregos mais curtos, diplomas de ensino superior normalmente levam a empregos mais longos e negros têm rotatividade maior.

O teste log-rank

Podemos nos perguntar, no entanto, se existe significância estatística nessas diferenças. Existem evidências para apoiar a tese de que duas curvas de sobrevivência são *de fato* diferentes? Podemos usar um teste não-paramétrico, o log-rank para responder essa pergunta. Ele é interessante porque não depende de hipóteses sobre a distribuição das curvas de sobrevivência. O procedimento é - tendo como hipótese nula que as duas curvas são iguais - comparar o número observado de eventos em cada grupo com o esperado caso a hipótese nula valesse. O pacote "survival"' traz uma implementação desse teste.

```
survdiff(Surv(tempo_emprego, demissao) ~ sexo, data = dados)
```

```
## Call:
## survdiff(formula = Surv(tempo_emprego, demissao) ~ sexo, data = dados)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## sexo=Feminino   80793    19606    22066       274       532
## sexo=Masculino 96565    26517    24057       251       532
##
##  Chisq= 532  on 1 degrees of freedom, p= <2e-16
```

O p-valor do teste, menor que 2^{-16}, nos diz que podemos rejeitar a hipótese nula com considerável confiança, encontramos evidências estatisticamente significantes de que de fato as curvas de sobrevivência de homens e mulheres são diferentes. Convido o leitor a repetir o teste com outros recortes em mente.

O modelo de Riscos Proporcionais de Cox

Tendo $h(t)$ como o risco no momento t, $h_o(t)$ como o risco base no período, β como um vetor de parâmetros e x como um vetor de k variáveis explicativas, esse modelo exprime a função risco da seguinte maneira:

$$h(t) = h_0(t) \times e^{\sum_{i=1}^{k} \beta_i x_i}$$

Podemos estimar o vetor β com regressão linear se aplicarmos logarítimos no modelo, que passa a ser:

$$\ln h(t) = \ln h_0(t) + \sum_{i=1}^{k} \beta_i x_i$$

Esse modelo é dito de riscos proporcionais porque a forma funcional que assumimos implica que curvas de riscos de indivíduos são múltiplos umas das outras e que, portanto,não se cruzam. Nesse modelo existe independência temporal na razão de risco de quaisquer dois indivíduos da amostra. Podemos estimar os parâmetros facilmente com `survival::coxph`.

```
cox = coxph(Surv(tempo_emprego, demissao) ~ homem + horas + Idade + branco + negro + industria + agricul

summary(cox)
```

```
## Call:
## coxph(formula = Surv(tempo_emprego, demissao) ~ homem + horas +
##     Idade + branco + negro + industria + agricultura + CNPJ +
##     firma_grande + servicos + salario, data = dados)
##
##   n= 177358, number of events= 46123
##
##                    coef  exp(coef)   se(coef)       z Pr(>|z|)
## homem         4.935e-03  1.005e+00  9.959e-03   0.496   0.6202
## horas        -1.402e-02  9.861e-01  7.075e-04 -19.822  < 2e-16 ***
## Idade        -5.357e-02  9.478e-01  5.367e-04 -99.804  < 2e-16 ***
## branco       -1.354e-01  8.733e-01  1.975e-02  -6.856 7.09e-12 ***
## negro         1.896e-02  1.019e+00  1.086e-02   1.746   0.0808 .
## industria    -1.336e+00  2.630e-01  2.409e-02 -55.444  < 2e-16 ***
## agricultura  -4.521e-01  6.363e-01  3.443e-02 -13.131  < 2e-16 ***
## CNPJ          3.218e-01  1.380e+00  3.669e-02   8.770  < 2e-16 ***
## firma_grande -5.104e-01  6.002e-01  1.182e-02 -43.192  < 2e-16 ***
## servicos     -1.213e+00  2.973e-01  1.507e-02 -80.488  < 2e-16 ***
## salario      -2.093e-04  9.998e-01  4.039e-06 -51.805  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## homem           1.0049     0.9951    0.9855    1.0248
## horas           0.9861     1.0141    0.9847    0.9874
## Idade           0.9478     1.0550    0.9468    0.9488
## branco          0.8733     1.1450    0.8402    0.9078
## negro           1.0191     0.9812    0.9977    1.0411
## industria       0.2630     3.8030    0.2508    0.2757
## agricultura     0.6363     1.5716    0.5948    0.6807
## CNPJ            1.3796     0.7249    1.2838    1.4824
## firma_grande    0.6002     1.6660    0.5865    0.6143
## servicos        0.2973     3.3634    0.2887    0.3062
```

```
## salario         0.9998     1.0002    0.9998    0.9998
##
## Concordance= 0.736  (se = 0.001 )
## Rsquare= 0.175   (max possible= 0.998 )
## Likelihood ratio test= 34183  on 11 df,   p=<2e-16
## Wald test            = 29993  on 11 df,   p=<2e-16
## Score (logrank) test = 33199  on 11 df,   p=<2e-16
```

Se defirnimos a Razão de Risco r_i de i-ésima covariada como $r_i := e_i^{\beta}$, então $r_i > 1$ implica que a i-ésima covariada leva a um aumento no nível de risco e $r_i < 1$ a uma diminuição. A tabela acima então deve ser capaz de responder algumas das perguntas que fizemos no primeiro parágrafo.