In homework 10, we considered the `cgd` dataset from the `survival` package.

```
> data(cgd, package = "survival")
```

The dataset has 203 rows of data and 16 variables (See ?survival::cgd for details). Initially, we focused on the variables:

- `id` subject identification number.
- `tstart` and `tstop` the start and end of each time interval.
- `status` equals 1 if the interval ends with an infection.

Other variables of interest in the dataset are:

- `center` enrolling center
- `treatment` placebo or gamma interferon
- `age` age in years, at study entry
- `sex` sex
- `enum` observation number within subject

Following homework 10, the number of infections experienced by each patient can be summarized with `nInf`.

```
> table(nInf <- aggregate(status ~ id, cgd, sum)$status)

 0  1  2  3  4  5  7
84 27  9  5  1  1  1
```

We will use the `cgd` dataset and those variables to answer the following questions.

Preliminary study:

1. What is the sample size of this dataset? In other words, how many patients participated in this study?
2. Which patient has the most infection episodes?
3. Which enrolling center has the most patients?
4. What proportion of patients received placebo?
5. What are the average ages for male and female patients, respectively?
6. What is the average number of days for the first infection?
7. Create a subset of `nInf` that provides the number of infections experienced by each patient who was enrolled in the NIH center. Call this vector `nInfNIH` and print `table(nInfNIH)`.

Data analysis:

8. Use `nInfNIH` from # and hanging rootograms from the `vcd` package to determine the distribution (out of binomial, Poisson, and negative binomial) that fits `nInfNIH` the best. To receive full credit, display the rootograms, state the hypotheses, p-value, and explain your conclusion.

9. Let X be the random variable that represents the number of infections experienced by each patient who was enrolled in the NIH center. Use the best-fit distribution from # and its estimated parameter(s) to find the following quantities.

    a. $P(X = 1)$
    b. $P(X > 1)$

10. Use the best-fit distribution from # and its estimated parameter(s) to find the following quantities.

    a. $E(X)$

    2. The third quantile (Q3)

11. Suppose we aim to approximate the probabilities in #9 using a normal distribution. What would be the appropriate choice of the mean and standard deviation for the normal distribution?

12. What are the normal approximations to the following probabilities from #9 without continuity correction?

    a. $P(X = 1)$
    b. $P(X > 1)$

13. What are the normal approximations to the following probabilities from #9 with continuity correction?

    a. $P(X = 1)$
    b. $P(X > 1)$

14. Repeat the analysis in #, but this time, only consider patients who were 18 years or older at the time of study entry, regardless of the enrolling center.

1

. What is the sample size of this dataset? In other words, how many patients participated in this study?

```
> length(unique(cgd$id))
[1] 128
```

2

. Which patient has the most infection episodes?

```
> library(tidyverse)
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.1      v purrr   1.0.1
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.3.0      v stringr 1.5.0
v readr   2.1.3      v forcats 1.0.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
> cgd %>%
+   group_by(id) %>%
+   summarise(infections=sum(status)) %>%
+   filter(infections==max(infections))
# A tibble: 1 x 2
     id infections
  <int>      <int>
1     2          7
```

3

. Which enrolling center has the most patients?

```
> library(tidyverse)
> cgd %>%
+   group_by(center) %>%
+   summarise(n1_patients = length(unique(id))) %>%
+   filter(n1_patients== max(n1_patients))
# A tibble: 1 x 2
  center n1_patients
  <fct>        <int>
1 NIH             26
```

4

. What proportion of patients received placebo?

```
> cgd %>%
+   group_by(id) %>%
+   summarise(placebo=any(treat == "placebo")) %>%
+   pull(placebo) %>%
+   mean
[1] 0.5078125
```

5

. What are the average ages for male and female patients, respectively?

```
> library(dplyr)
> cgd %>%
+   group_by(id , sex) %>%
+   summarise(age= mean(age)) %>%
+   ungroup() %>%
+   group_by(sex) %>%
+   summarise(mean_age = mean(age))
`summarise()` has grouped output by 'id'. You can override using the `.groups`
argument.
# A tibble: 2 x 2
  sex    mean_age
  <fct>     <dbl>
1 male       13.5
2 female     19.8
```

6

. What is the average number of days for the first infection?

```
> cgd %>%
+   filter(status==1) %>%
+   group_by(id) %>%
+   filter(tstart == min(tstart)) %>%
+   ungroup() %>%
+   mutate(days= tstop-tstart) %>%
+   summarise(mean(days))
# A tibble: 1 x 1
  `mean(days)`
         <dbl>
1         160.
```

7

. Create a subset of `nInf` that provides the number of infections experienced by each patient who was enrolled in the NIH center. Call this vector `nInfNIH` and print `table(nInfNIH)`.

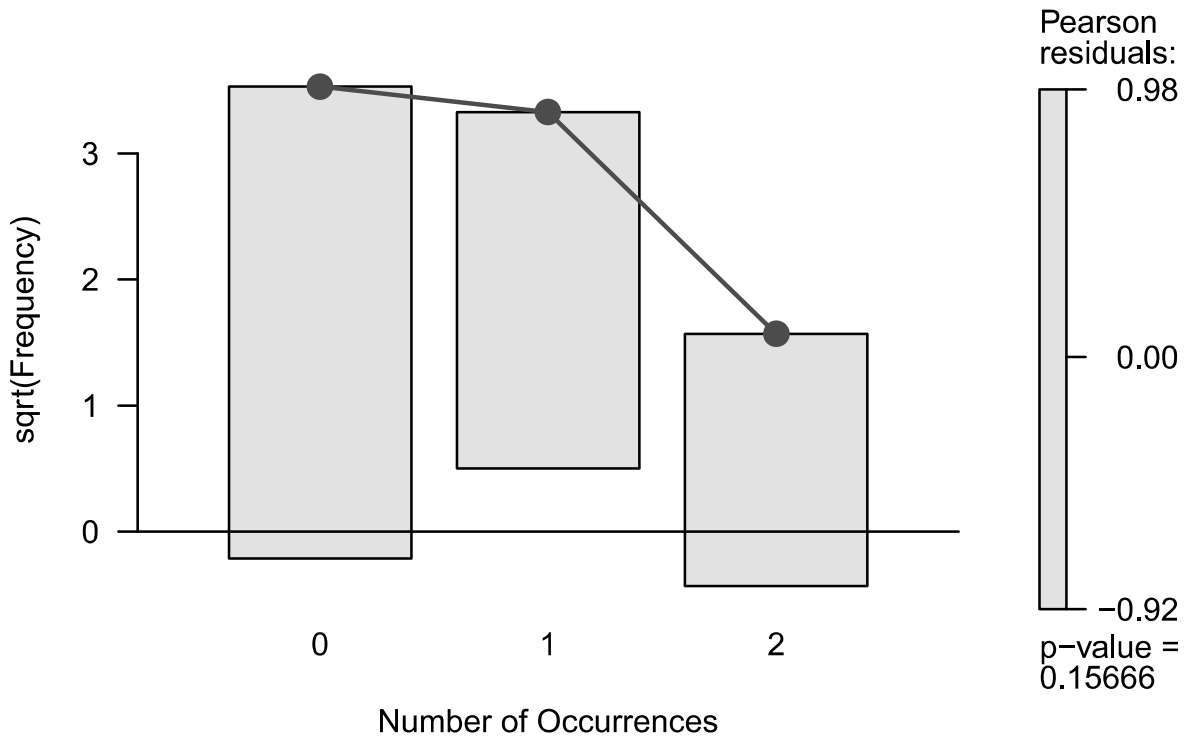
```
> table(nInfNIH <- aggregate(status ~ id, cgd, sum, subset=center=="NIH")$status)

 0  1  2
14  8  4
```

Data analysis:

8. Use `nInfNIH` from # and hanging rootograms from the `vcd` package to determine the distribution (out of binomial, Poisson, and negative binomial) that fits `nInfNIH` the best. To receive full credit, display the rootograms, state the hypotheses, p-value, and explain your conclusion.

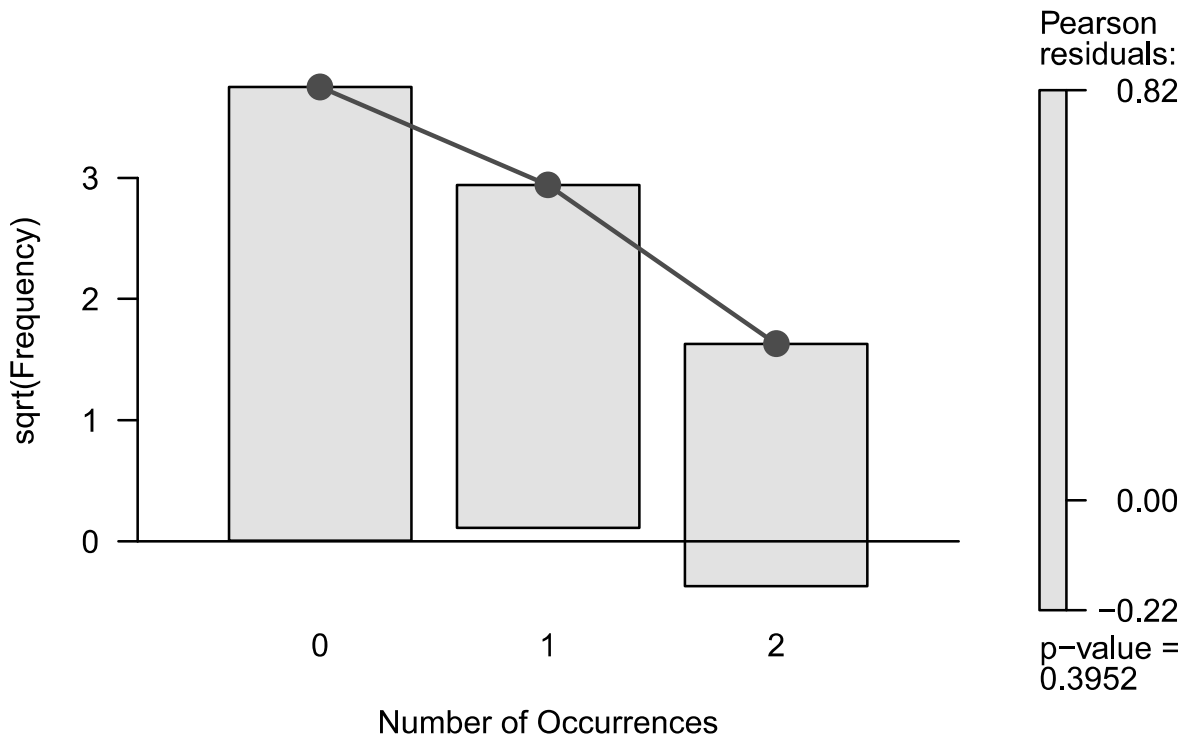
H0 : The given data follows a Poisson/Binomial/Negative binomial distribution HA:The given data does not follows a Poisson/Binomial/Negative binomial distribution

```
> library(grid)
> library(vcd)
> binom_fit  = goodfit(nInfNIH, type="binomial")
Warning in goodfit(nInfNIH, type = "binomial"): size was not given, taken as
maximum count
> summary(binom_fit)

     Goodness-of-fit test for binomial distribution

                       X^2 df  P(> X^2)
Likelihood Ratio 1.936794  1 0.1640171
> rootogram(binom_fit, shade = TRUE)
```

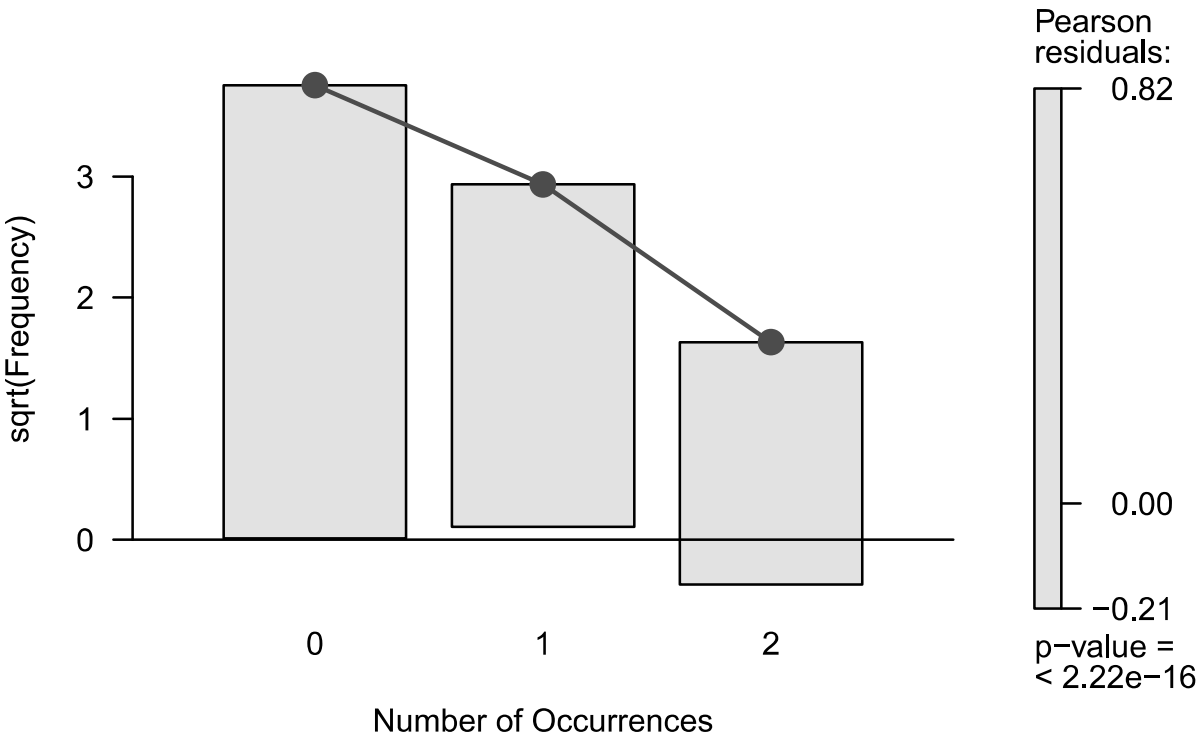



```
>
> pois_fit=goodfit(nInfNIH, type="poisson")
> summary(pois_fit)

     Goodness-of-fit test for poisson distribution
```

```
                      X^2 df  P(> X^2)
Likelihood Ratio 1.915432  1 0.1663611
> rootogram(pois_fit, shade = TRUE)
```




```
>
> nbinom_fit=goodfit(nInfNIH, type="nbinomial")
> summary(nbinom_fit)

     Goodness-of-fit test for nbinomial distribution

                     X^2 df P(> X^2)
Likelihood Ratio 1.93409  0        0
> rootogram(nbinom_fit, shade = TRUE)
```



Conclusion: The Poisson model yields the highest p-value(p-value=0.3952) indicating that Poisson distribution fits the data the best.

9. Let X be the random variable that represents the number of infections experienced by each patient who was enrolled in the NIH center. Use the best-fit distribution from # and its estimated parameter(s) to find the following quantities.

    a. $P(X = 1)$

```
> dpois(1, pois_fit$par$lambda)
[1] 0.3325742
```

b. $P(X > 1)$

```
> 1 - ppois(1, pois_fit$par$lambda)
[1] 0.1269929
```

10. Use the best-fit distribution from # and its estimated parameter(s) to find the following quantities.

    a. $E(X)$

```
> pois_fit$par$lambda
[1] 0.6153846
```

2. The third quantile (Q3)

```
> qpois(0.75, pois_fit$par$lambda)
[1] 1
```

11

. Suppose we aim to approximate the probabilities in #9 using a normal distribution. What would be the appropriate choice of the mean and standard deviation for the normal distribution?

```
> mean = lambda = 0.6153846
> #sd = sqrt(lambda)= 0.7844645
> sigma=sqrt(0.6153846)
```

#Sigma= 0.7844645 #mean = lambda = 0.6153846

12. What are the normal approximations to the following probabilities from #9 without continuity correction?

a. $P(X = 1)$

```
> pnorm(1, pois_fit$par$lambda, sqrt(pois_fit$par$lambda))- pnorm(1, pois_fit$par$lambda, sqrt(pois_fit$pa
[1] 0
```

b. $P(X > 1)$

```
> pnorm(1, pois_fit$par$lambda, sqrt(pois_fit$par$lambda), lower.tail = F)
[1] 0.3119642
```

13

. What are the normal approximations to the following probabilities from #9 with continuity correction?

a. $P(X = 1)$

```
> pnorm(1.5,pois_fit$par$lambda, sqrt(pois_fit$par$lambda))- pnorm(0.5, pois_fit$par$lambda, sqrt(pois_fit
[1] 0.4287382
```

b. $P(X > 1)$

```
> pnorm(1.5,pois_fit$par$lambda, sqrt(pois_fit$par$lambda),lower.tail= F)
[1] 0.1297301
```

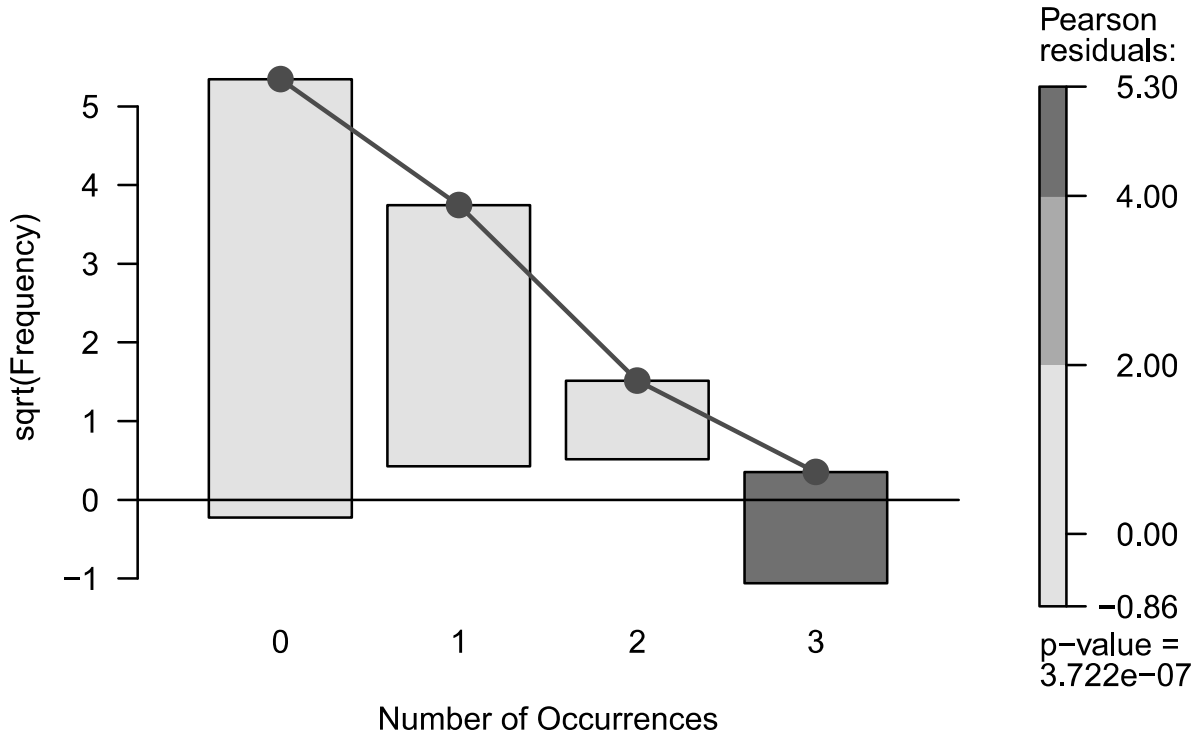
14

. Repeat the analysis in #, but this time, only consider patients who were 18 years or older at the time of study entry, regardless of the enrolling center.

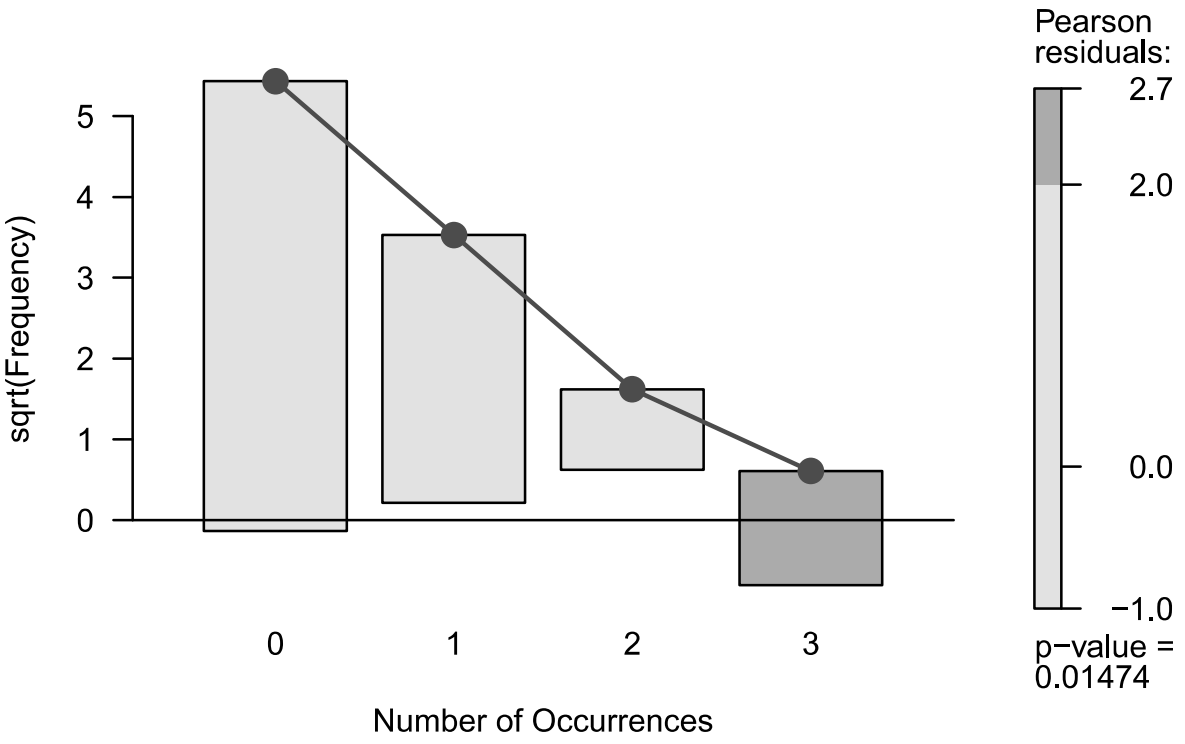
```
> table(nInf18 <- aggregate(status ~ id, cgd, sum, subset=age>=18)$status)

 0  1  2  3
31 11  1  2
```

H0 : The given data follows a Poisson/Binomial/Negative binomial distribution HA:The given data does not follows a Poisson/Binomial/Negative binomial distribution

```
> library(grid)
> library(vcd)
> binom_fit  = goodfit(nInf18, type="binomial")
Warning in goodfit(nInf18, type = "binomial"): size was not given, taken as
maximum count
> rootogram(binom_fit, shade = TRUE)
```




```
>
> pois_fit=goodfit(nInf18, type="poisson")
> rootogram(pois_fit, shade = TRUE)
```




```
>
> nbinom_fit=goodfit(nInf18, type="nbinomial")
> rootogram(nbinom_fit, shade = TRUE)
```