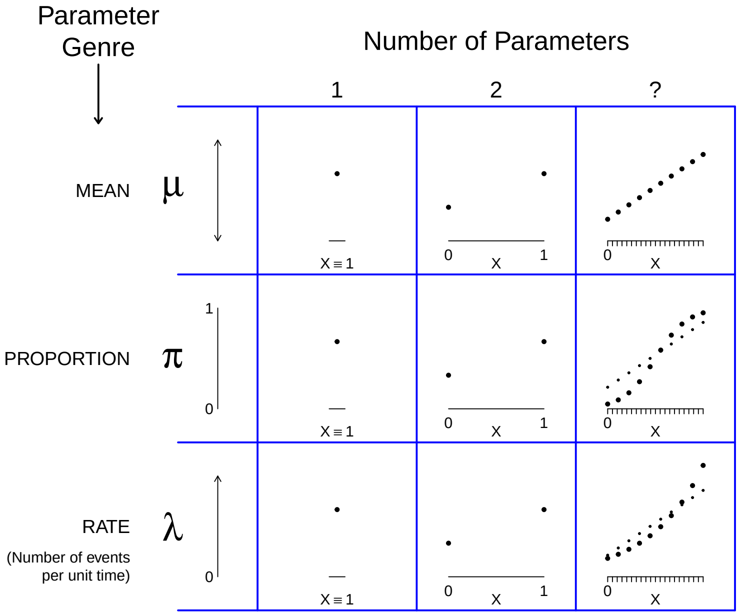
020 - Inference about a Population Rate (λ)

EPIB 607

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on October 28, 2021


Parameter Genre
Number of Parameters
1
2
?
MEAN
μ
PROPORTION
π
1
0
RATE
(Number of events per unit time)
λ
0
X ≡ 1
0
X
1
0
X

Motivating example

Poisson Model for Sampling Variability of a Count in a Given Amount of "Experience"

Inference regarding μ, based on observed count y

Inference regarding an event rate parameter λ, based on observed number of events y in a known amount of population-time (PT)

Test of $H_0 : \mu = \mu_0 \quad \Leftrightarrow \quad \lambda = \lambda_0$

Motivating example: HPV-16 Vaccine

The New England
Journal of Medicine

VOLUME 347 NOVEMBER 21, 2002 NUMBER 21

A CONTROLLED TRIAL OF A HUMAN PAPILLOMAVIRUS TYPE 16 VACCINE

LAURA A. KOUTSKY, PH.D., KEVIN A. AULT, M.D., COSETTE M. WHEELER, PH.D., DARRON R. BROWN, M.D., ELIAV BARR, M.D., FRANCES B. ALVAREZ, R.N., LISA M. CHIACCHIERINI, PH.D., AND KATHRIN U. JANSEN, PH.D., FOR THE PROOF OF PRINCIPLE STUDY INVESTIGATORS

Motivating example: HPV-16 Vaccine

- **Background**: $\approx 20\%$ of adults become infected with human papillomavirus type 16 (HPV-16), someof which progress to anogenital cancer.

Motivating example: HPV-16 Vaccine

- **Background**: $\approx 20\%$ of adults become infected with human papillomavirus type 16 (HPV-16), someof which progress to anogenital cancer.
- **Methods**:
  - Randomly assigned 2392 young women (females age 16-23) to receive three doses of placebo or HPV-16 virus-like-particle vaccine (40 µg per dose), given at day 0, month 2, and month 6.
  - Genital samples to test for HPV-16 DNA were obtained at enrollment, one month after the third vaccination, and every six months thereafter.
  - The primary end point was persistent HPV-16 infection, defined as the detection of HPV-16 DNA in samples obtained at two or more visits.

Motivating example: HPV-16 Vaccine

- **Background**: $\approx 20\%$ of adults become infected with human papillomavirus type 16 (HPV-16), someof which progress to anogenital cancer.
- **Methods**:
  - Randomly assigned 2392 young women (females age 16-23) to receive three doses of placebo or HPV-16 virus-like-particle vaccine (40 µg per dose), given at day 0, month 2, and month 6.
  - Genital samples to test for HPV-16 DNA were obtained at enrollment, one month after the third vaccination, and every six months thereafter.
  - The primary end point was persistent HPV-16 infection, defined as the detection of HPV-16 DNA in samples obtained at two or more visits.
- **Results**:
  - Median follow-up time of 17.4 months
  - Incidence of persistent HPV-16 infection:
    - Placebo: 3.8 per 100 woman-years at risk
    - Vaccine: 0 per 100 woman-years at risk

Table 3

Table 3. Efficacy Analyses of a Human Papillomavirus Type 16 (HPV-16) L1 Virus-like–Particle Vaccine.

Type of Analysis	End Point	HPV-16 Vaccine				Placebo				Observed Efficacy (95% CI)*	P Value
		No. of Women	Cases of Infection	Woman-Yr at Risk	Infection Rate per 100 Woman-Yr at Risk	No. of Women	Cases of Infection	Woman-Yr at Risk	Infection Rate per 100 Woman-Yr at Risk		
					%				%	%	
Primary per-protocol efficacy analysis†	Persistent HPV-16 infection	768	0	1084.0	0	765	41	1076.9	3.8	100 (90–100)	<0.001
Efficacy analysis including women with general protocol violations‡	Persistent HPV-16 infection	800	0	1128.0	0	793	42	1109.7	3.8	100 (90–100)	—§
Secondary per-protocol efficacy analysis†	Transient or persistent HPV-16 infection	768	6	1084.0	0.6	765	68	1076.9	6.3	91.2 (80–97)	—§

Question: For Primary and Secondary per-protocol efficacy analysis, calculate a 95% CI of infection rate per 100 woman-years at risk for vaccine and placebo group.

Normal Approximation Based CI for the Count

Primary analysis:

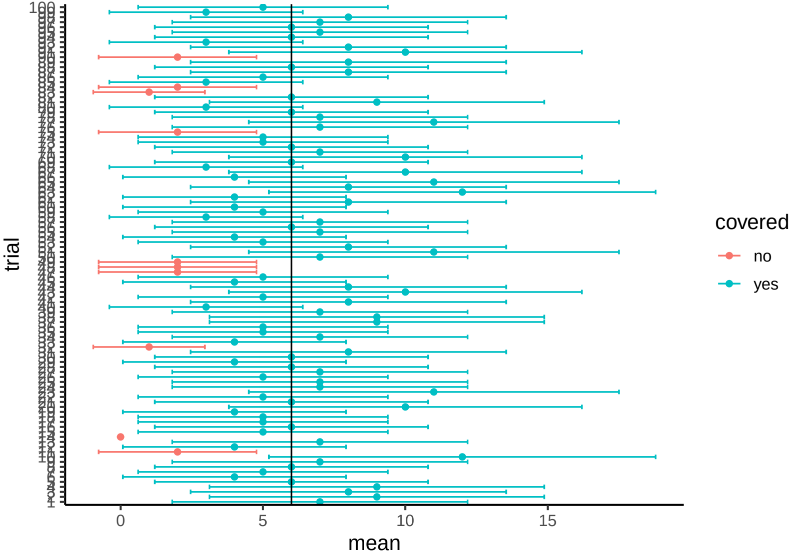
```r
# Vaccine group
qnorm(p = c(0.025, 0.975), mean = 0, sd = sqrt(0))

## [1] 0 0

# Placebo
qnorm(p = c(0.025, 0.975), mean = 41, sd = sqrt(41))

## [1] 28.45011 53.54989
```

Normal Approximation Based CI for the Count

Secondary analysis:

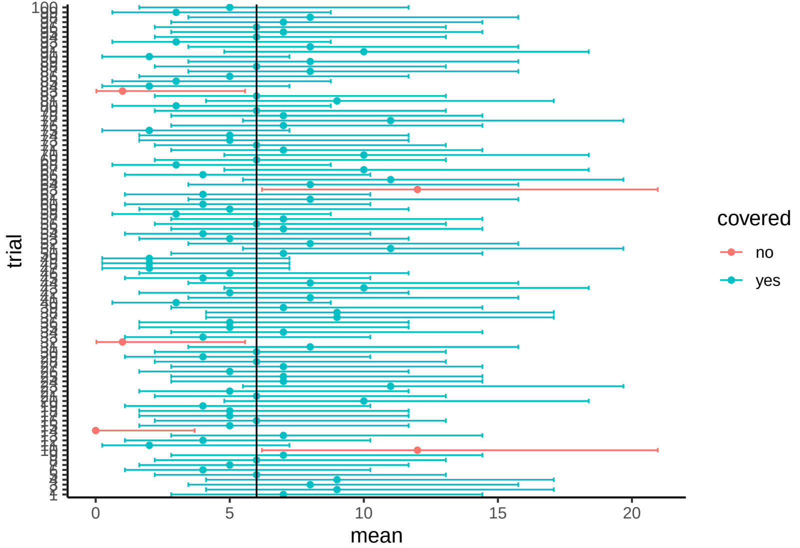
```
# Vaccine group
qnorm(p = c(0.025, 0.975), mean = 6, sd = sqrt(6))

## [1]  1.199088 10.800912

# Placebo
qnorm(p = c(0.025, 0.975), mean = 68, sd = sqrt(68))

## [1] 51.83772 84.16228
```

Coverage Probability of Normal Approx. - Truth is Poisson($\mu = 6$)



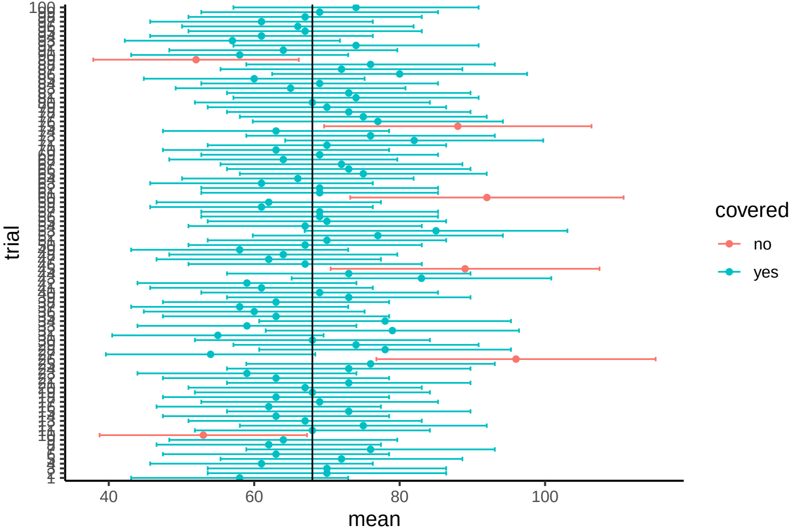
Each 95% CI was calculated using the Normal Approximation. Median CI width is 9.60

Coverage Probability of Exact Method - Truth is Poisson($\mu = 6$)



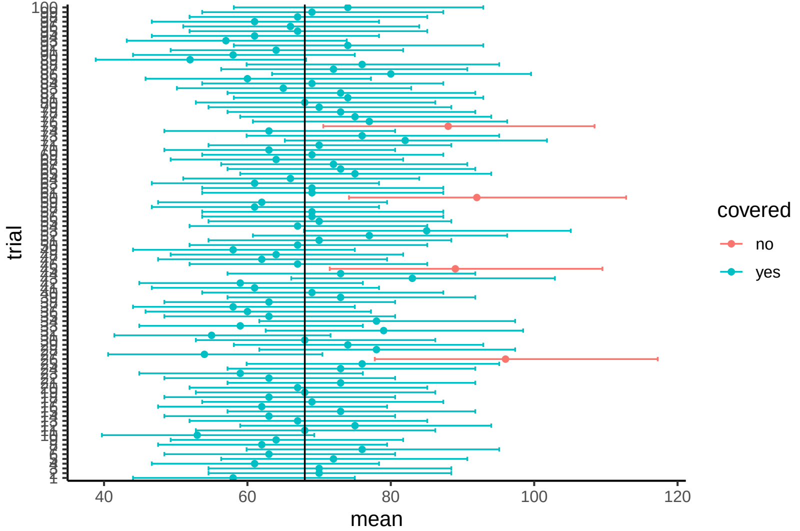
Each 95% CI was calculated using Poisson model. Median CI width is 10.86

Coverage Probability Normal Approx. - Truth is Poisson($\mu = 68$)



Each 95% CI was calculated using the Normal Approximation. Median CI width is 32.44

Coverage Probability Exact Method - Truth is Poisson($\mu = 68$)



Each 95% CI was calculated using Poisson model. Median CI width is 33.52

Motivating example

Poisson Model for Sampling Variability of a Count in a Given Amount of "Experience"

Inference regarding μ, based on observed count y

Inference regarding an event rate parameter λ, based on observed number of events y in a known amount of population-time (PT)

Test of $H_0 : \mu = \mu_0 \quad \Leftrightarrow \quad \lambda = \lambda_0$

The Poisson Distribution

- The (infinite number of) probabilities $P_0, P_1, ..., P_y, ...$, of observing $Y = 0, 1, 2, \ldots, y, \ldots$ events in a given amount of "experience."

- These probabilities, $P(Y = k) \rightarrow$ `dpois()`, are governed by a single parameter, the mean $E[Y] = \mu$ which represents the expected **number** of events in the amount of experience actually studied.

- We say that a random variable $Y \sim \text{Poisson}(\mu)$ distribution if

$$P(Y = k) = \frac{\mu^k}{k!} e^{-\mu}, \quad k = 0, 1, 2, \ldots$$

The Poisson Distribution

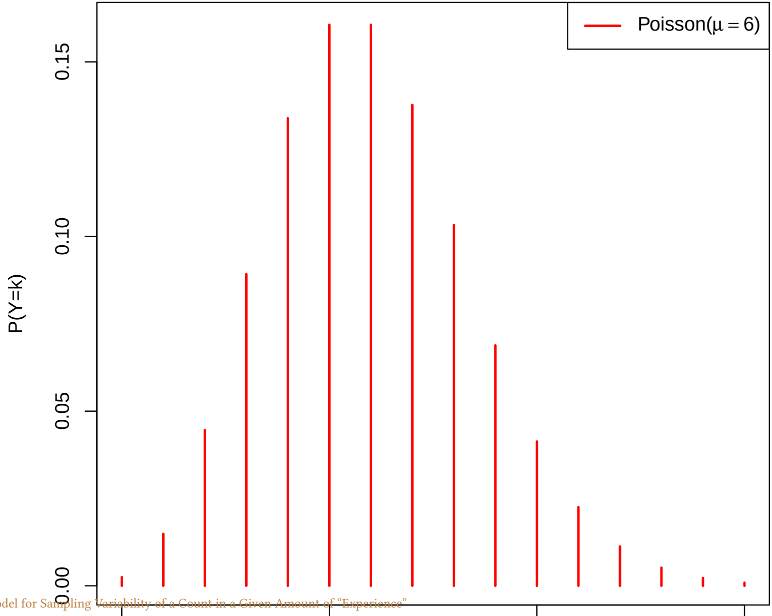
- The (infinite number of) probabilities $P_0, P_1, ..., P_y, ...,$ of observing $Y = 0, 1, 2, \ldots, y, \ldots$ events in a given amount of "experience."
- These probabilities, $P(Y = k) \rightarrow$ `dpois()`, are governed by a single parameter, the mean $E[Y] = \mu$ which represents the expected **number** of events in the amount of experience actually studied.
- We say that a random variable $Y \sim \text{Poisson}(\mu)$ distribution if

$$P(Y = k) = \frac{\mu^k}{k!} e^{-\mu}, \quad k = 0, 1, 2, \ldots$$

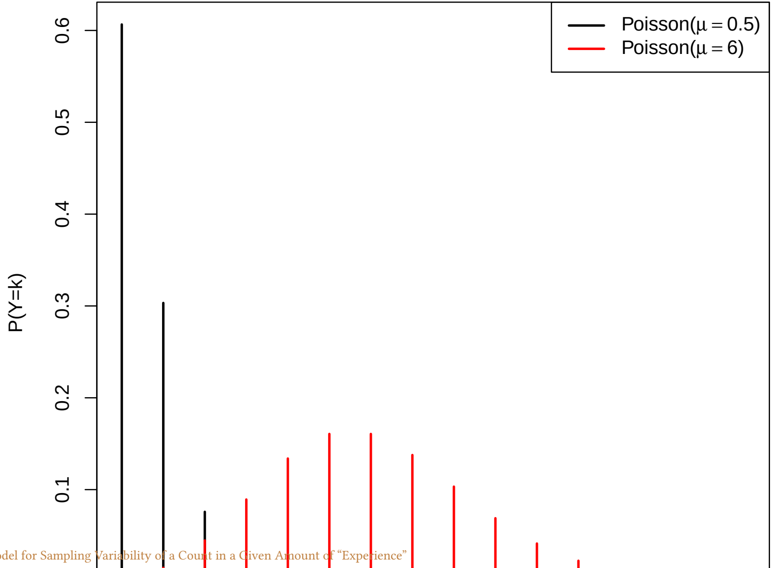
- Note: in `dpois()` μ is referred to as `lambda`
- Note the distinction between μ and λ
  - μ: expected **number** of events
  - λ: **rate** parameter

The probability mass function for $\mu = 6$

```
dpois(x = 0:15, lambda = 6)
```



The probability mass function



The Poisson Distribution: what it is, and features

- $\sigma_Y^2 = \mu \ \rightarrow \ \ \sigma_Y = \sqrt{\mu}.$

The Poisson Distribution: what it is, and features

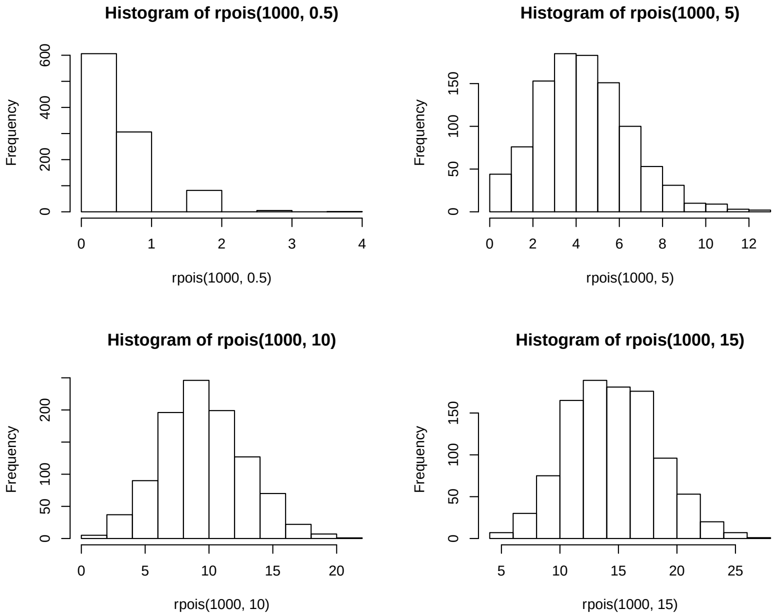
- $\sigma_Y^2 = \mu \;\rightarrow\; \sigma_Y = \sqrt{\mu}$.

- Approximated by $\mathcal{N}(\mu, \sqrt{\mu})$ when $\mu >> 10$

The Poisson Distribution: what it is, and features

- $\sigma_Y^2 = \mu \ \rightarrow \ \ \sigma_Y = \sqrt{\mu}$.
- Approximated by $\mathcal{N}(\mu, \sqrt{\mu})$ when $\mu >> 10$
- Open-ended (unlike Binomial), but in practice, has finite range.
- Poisson data sometimes called "numerator only": (unlike Binomial) may not "see" or count "non-events"

Normal approximation to Poisson is the CLT in action



How it arises

- Count of events or items that occur randomly, with low homogeneous intensity, in time, space, or 'item'-time (e.g. person–time).
- Binomial(n, π) when $n \to \infty$ and $\pi \to 0$, but $n \times \pi = \mu$ is finite.
- $Y \sim Poisson(\mu_Y)$ if time (T) between events follows an $T \sim \text{Exponential}(\mu_T = 1/\mu_Y)$.
  http://www.epi.mcgill.ca/hanley/bios601/Intensity-Rate/Randomness_poisson.pdf
- As sum of ≥ 2 *independent* Poisson random variables, with same **or different** μ's:
  $Y_1 \sim \text{Poisson}(\mu_1)\ \ Y_2 \sim \text{Poisson}(\mu_2) \Rightarrow Y = Y_1 + Y_2 \sim \text{Poisson}(\mu_1 + \mu_2)$.

Poisson distribution as a limit

The rationale for using the Poisson distribution in many situations is provided by the following proposition.

Proposition 1 (Limit of a binomial is Poisson).

Suppose that $Y \sim$ Binomial(n, π). If we let $\pi = \mu/n$, then as $n \to \infty$, Binomial$(n, \pi) \to$ Poisson(μ). Another way of saying this: for large n and small π, we can approximate the Binomial(n, π) probability by the Poisson$(\mu = n\pi)$.

Poisson approximation to the Binomial

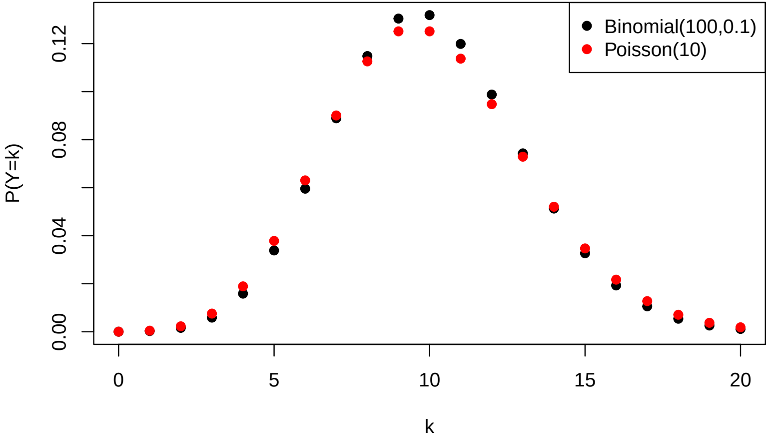


Figure: Probability mass funtion for Bin(n=100,0.1) and Poisson(10)

Examples

- numbers of asbestos fibres
- deaths from horse kicks*
- needle-stick or other percutaneous injuries
- bus-driver accidents*
- twin-pairs*
- radioactive disintegrations*
- flying-bomb hits*
- white blood cells
- typographical errors
- cell occupants – in a given volume, area, line-length, population-time, time, etc. [1]

[1] * included in http://www.epi.mcgill.ca/hanley/bios601/Intensity-Rate/

Motivating example

Poisson Model for Sampling Variability of a Count in a Given Amount of "Experience"

Inference regarding μ, based on observed count y

Inference regarding an event rate parameter λ, based on observed number of events y in a known amount of population-time (PT)

Test of $H_0 : \mu = \mu_0 \quad \Leftrightarrow \quad \lambda = \lambda_0$

Confidence interval for μ

- If the CLT hasn't kicked in, then the usual CI might not be appropriate:

$$\text{point-estimate} \pm z^{\star} \times \text{standard error}$$

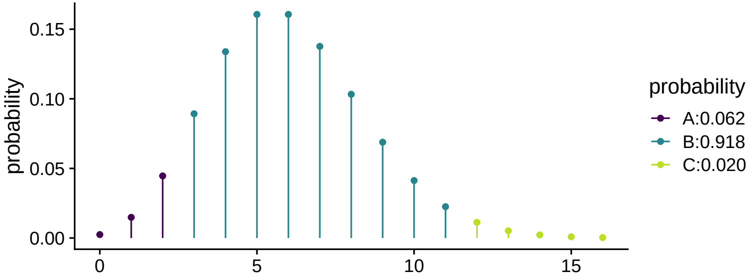
Confidence interval for μ

- If the CLT hasn't kicked in, then the usual CI might not be appropriate:

$$\text{point-estimate} \pm z^{\star} \times \text{standard error}$$

- `qpois` function doesn't work either:

```
# middle area is not 95%
mosaic::xqpois(c(0.025, 0.975), lambda = 6)
```



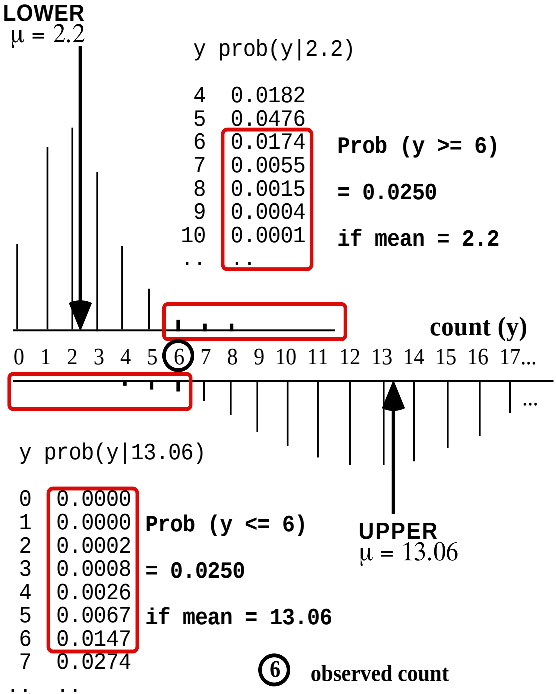

```
## [1]  2 11
```

Confidence interval for μ

- Similar to the binomial (Clopper-Pearson CI), we consider a *first-principles* $100(1-\alpha)\%$ CI $[\mu_{LOWER},\ \mu_{UPPER}]$ such that

$$P(Y \geq y \,|\, \mu_{LOWER}) = \alpha/2 \quad \text{and} \quad P(Y \leq y \,|\, \mu_{UPPER}) = \alpha/2.$$

- For example, the 95% CI for μ, based on $y = 6$, is $[\underline{2.20}, \underline{13.06}]$.



LOWER
μ = 2.2
y prob(y|2.2)
4 0.0182
5 0.0476
6 0.0174
7 0.0055
8 0.0015
9 0.0004
10 0.0001
.. ..
Prob (y >= 6)
= 0.0250
if mean = 2.2
count (y)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17...
...
y prob(y|13.06)
0 0.0000
1 0.0000
2 0.0002
3 0.0008
4 0.0026
5 0.0067
6 0.0147
7 0.0274
.. ..
Prob (y <= 6)
= 0.0250
if mean = 13.06
UPPER
μ = 13.06
6 observed count

Confidence interval for μ

- For a given confidence level, there is one CI for each value of y.
- Each one can be worked out by trial and error, or – as has been done for the last 80 years – directly from the (exact) link between the tail areas of the Poisson and **Gamma** distributions.
- These CI's – for y up to at least 30 – were found in special books of statistical tables or in textbooks.
- As you can check, z-based intervals are more than adequate beyond this y. **Today**, if you have access to `R` (or `Stata` or `SAS`) you can obtain the first principles CIs directly **for *any* value of** y.

80%, 90% and 95% CI for mean count μ if we observe 0 to 30 events in a certain amount of experience

y	95%		90%		80%	
0	0.00	3.69	0.00	3.00	0.00	2.30
1	0.03	5.57	0.05	4.74	0.11	3.89
2	0.24	7.22	0.36	6.30	0.53	5.32
3	0.62	8.77	0.82	7.75	1.10	6.68
4	1.09	10.24	1.37	9.15	1.74	7.99
5	1.62	11.67	1.97	10.51	2.43	9.27
6	2.20	13.06	2.61	11.84	3.15	10.53
7	2.81	14.42	3.29	13.15	3.89	11.77
8	3.45	15.76	3.98	14.43	4.66	12.99
9	4.12	17.08	4.70	15.71	5.43	14.21
10	4.80	18.39	5.43	16.96	6.22	15.41
11	5.49	19.68	6.17	18.21	7.02	16.60
12	6.20	20.96	6.92	19.44	7.83	17.78
13	6.92	22.23	7.69	20.67	8.65	18.96
14	7.65	23.49	8.46	21.89	9.47	20.13
15	8.40	24.74	9.25	23.10	10.30	21.29
16	9.15	25.98	10.04	24.30	11.14	22.45
17	9.90	27.22	10.83	25.50	11.98	23.61
18	10.67	28.45	11.63	26.69	12.82	24.76
19	11.44	29.67	12.44	27.88	13.67	25.90
20	12.22	30.89	13.25	29.06	14.53	27.05
21	13.00	32.10	14.07	30.24	15.38	28.18
22	13.79	33.31	14.89	31.41	16.24	29.32
23	14.58	34.51	15.72	32.59	17.11	30.45
24	15.38	35.71	16.55	33.75	17.97	31.58

95% CI for mean count μ with q function

- To obtain these in `R` we use the natural link between the Poisson and the *gamma* distributions.[2]

- In `R`, e.g., the 95% limits for μ based on $y = 6$ are obtained as

```
qgamma(p = c(0.025,0.975), shape = c(6, 7))
## [1]  2.201894 13.059474
```

- More generically, for *any* y, as

```
qgamma(p = c(0.025,0.975), shape = c(y, y+1))
```

[2] details found here

95% CI for mean count μ with canned function

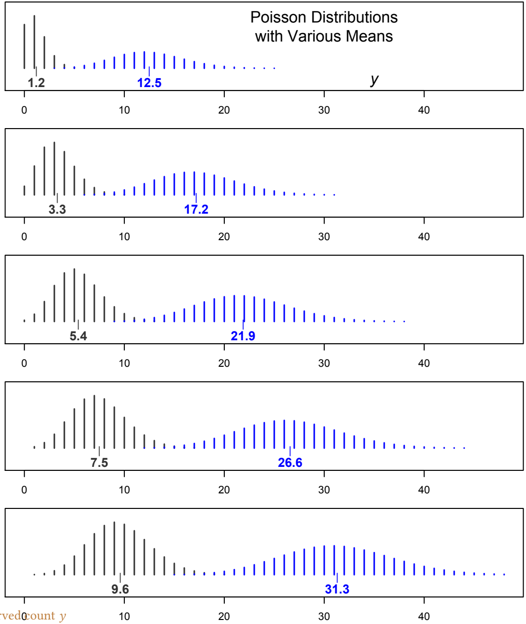
- These limits can also be found using the canned function in R

```
stats::poisson.test(6)
```

```
## Exact Poisson test with 6 time base: 1
## number of events = 6, time base = 1, p-value = 0.0005942
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##   2.201894 13.059474
## sample estimates:
## event rate
##          6
```

z-based confidence intervals

once μ is in the upper teens, the Poisson $\rightarrow$ the Normal



z-based confidence intervals

- Thus, a plus/minus CI based on SE = $\hat{\sigma} = \sqrt{\hat{\mu}} = \sqrt{y}$, is simply

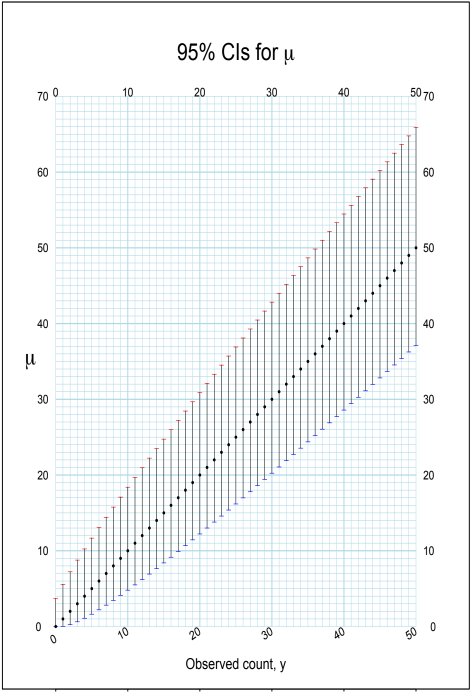
$$[\mu_L,\ \mu_U] = y \pm z^{\star} \times \sqrt{y}.$$

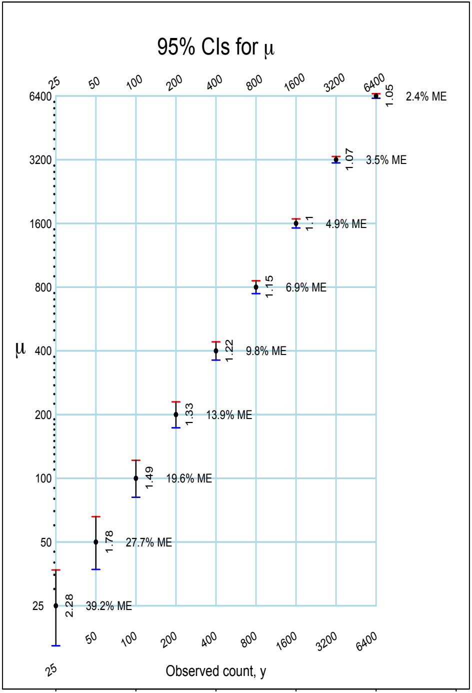
- Equivalently we can use the q function:

$$qnorm(p = c(0.025, 0.975), mean = y, sd = \sqrt{y})$$

- From a single realization y of a $N(\mu, \sigma_Y)$ random variable, we can't estimate **both** μ and σ_Y: for a SE, we would have to use *outside* information on σ_Y.

- In the Poisson(μ) distribution, $\sigma_Y = \sqrt{\mu}$, so we calculate a "model-based" SE.


95% CIs for μ
0
10
20
30
40
50
60
70
μ
Observed count, y


95% CIs for μ
25
50
100
200
400
800
1600
3200
6400
μ
Observed count, y
1.05
2.4% ME
1.07
3.5% ME
1.1
4.9% ME
1.15
6.9% ME
1.22
9.8% ME
1.33
13.9% ME
1.49
19.6% ME
1.78
27.7% ME
2.28
39.2% ME

Note

How is it that one can form a CI for μ from a single observation y?

Note

How is it that one can form a CI for μ from a single observation y?

- If we had a single realization y of a $\mathcal{N}(\mu, \sigma_Y)$ random variable, we could not, from this single y, estimate both μ and σ_Y
- However, the *Poisson*(μ) distribution is different in that $\sigma_Y = \sqrt{\mu}$ so we can calculate a **model-based** standard error from this relationship between the mean and the variance

Motivating example

Poisson Model for Sampling Variability of a Count in a Given Amount of "Experience"

Inference regarding μ, based on observed count y

Inference regarding an event rate parameter λ, based on observed number of events y in a known amount of population-time (PT)

Test of $H_0 : \mu = \mu_0 \quad \Leftrightarrow \quad \lambda = \lambda_0$

Rates are better for comparisons

year	deaths (y)
1971	33
2002	211

Table: Deaths from lung cancer in the age-group 55-60 in Quebec in 1971 and 2002

Rates are better for comparisons

year	deaths (y)
1971	33
2002	211

Table: Deaths from lung cancer in the age-group 55-60 in Quebec in 1971 and 2002

A researcher asks: Is the situation getting worse over time for lung cancer in this age group?

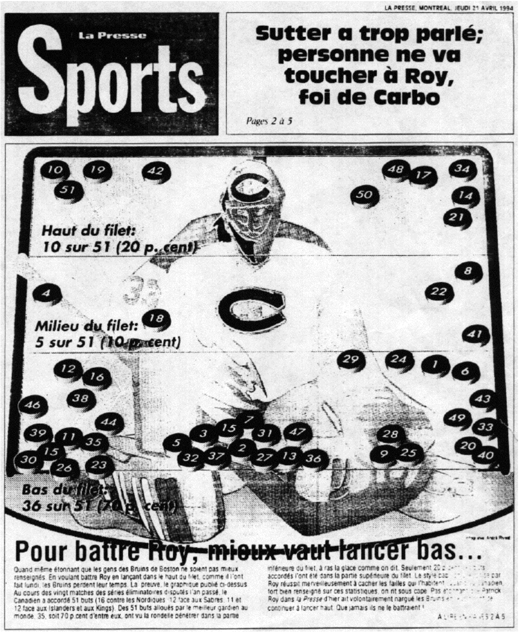
Rates are better for comparisons

year	deaths (y)
1971	33
2002	211

Table: Deaths from lung cancer in the age-group 55-60 in Quebec in 1971 and 2002

A researcher asks: Is the situation getting worse over time for lung cancer in this age group?

Your reply: **What's the denominator??**



LA PRESSE, MONTRÉAL, JEUDI 21 AVRIL 1994

La Presse

Sports

Sutter a trop parlé; personne ne va toucher à Roy, foi de Carbo

Pages 2 à 5

Haut du filet: 10 sur 51 (20 p. cent)

Milieu du filet: 5 sur 51 (10 p. cent)

Bas du filet: 36 sur 51 (70 p. cent)

Pour battre Roy, mieux vaut lancer bas...

Quand même étonnant que les gens des Bruins de Boston ne soient pas mieux renseignés. En voulant battre Roy en lançant dans le haut du filet, comme il l'ont fait lundi, les Bruins perdent leur temps. La preuve, le graphique publié ci-dessus. Au cours des vingt matches des séries éliminatoires disputés l'an passé, le Canadien a accordé 51 buts (16 contre les Nordiques, 12 face aux Sabres, 11 et 12 face aux Islanders et aux Kings). Des 51 buts alloués par le meilleur gardien au monde, 35, soit 70 p.cent d'entre eux, ont vu la rondelle pénétrer dans la partie inférieure du filet, à ras la glace comme on dit. Seulement 20 [illegible] accordés l'ont été dans la partie supérieure du filet. Le style [illegible] par Roy réussit merveilleusement à cacher les failles qui l'habitent. [illegible] Canadien, fort bien renseigné sur ces statistiques, on rit sous cape. Pas [illegible] Patrick Roy dans *la Presse* d'hier ait volontairement nargué les Bruins [illegible] continuer à lancer haut. Que jamais ils ne le battraient !

A LIRE [illegible] 2 A 5

Rates are better for comparisons

- So far, we have focused on inference regarding μ, the expected **number** of events in the amount of experience actually studied.
- However, for comparison purposes, the frequency is more often expressed as a **rate**, **intensity** or **incidence density (ID)**.

Rates are better for comparisons

- So far, we have focused on inference regarding μ, the expected **number** of events in the amount of experience actually studied.

- However, for comparison purposes, the frequency is more often expressed as a **rate**, **intensity** or **incidence density (ID)**.

year	deaths (y)	person-time (PT)	rate ($\hat{\lambda}$)
1971	33	131,200 years	25 per 100,000 women-years
2002	211	232,978 years	91 per 100,000 women-years

Table: Deaths from lung cancer in the age-group 55-60 in Quebec in 1971 and 2002

Rates are better for comparisons

- The *statistic*, the empirical rate or empirical incidence density, is

$$rate = \hat{ID} = \hat{\lambda} = y/\mathrm{PT}.$$

- where y is the observed number of events and PT is the amount of Population-Time in which these events were observed.

- We think of $\hat{ID}$ or $\hat{\lambda}$ as a point estimate of the (theoretical) Incidence Density *parameter*, ID or λ.

CI for the rate parameter λ

- To calculate a CI for the ID parameter, we **treat the PT <u>denominator</u> as a constant**, and the **<u>numerator</u>, y, as a Poisson random variable**, with expectation $E[y] = \mu = \lambda \times PT$, so that

$$\lambda = \mu \div \mathrm{PT}$$
$$\hat{\lambda} = \hat{\mu} \div \mathrm{PT}$$
$$= y \div \mathrm{PT}$$

$$\boxed{\text{CI for } \lambda = \{\text{CI for } \mu\} \div \mathrm{PT}.} \tag{1}$$

CI for the rate parameter λ

- $y = 211$ deaths from lung cancer in 2002 leads to a 95% CI for μ:

```r
qgamma(p = c(0.025, 0.975), shape = c(211, 212))
```

```
## [1] 183.4885 241.4725
```

CI for the rate parameter λ

- $y = 211$ deaths from lung cancer in 2002 leads to a 95% CI for μ:

```r
qgamma(p = c(0.025, 0.975), shape = c(211, 212))
## [1] 183.4885 241.4725
```

- From this we can calculate the 95% CI **per 100,000 WY** for λ using a PT=232978 years:

```r
qgamma(p = c(0.025, 0.975), shape = c(211, 212)) / 232978 * 1e5
## [1]  78.75788 103.64607
```

CI for the rate parameter λ

- $y = 211$ deaths from lung cancer in 2002 leads to a 95% CI for μ:

```
qgamma(p = c(0.025, 0.975), shape = c(211, 212))
## [1] 183.4885 241.4725
```

- From this we can calculate the 95% CI **per 100,000 WY** for λ using a PT=232978 years:

```
qgamma(p = c(0.025, 0.975), shape = c(211, 212)) / 232978 * 1e5
## [1]  78.75788 103.64607
```

- $y = 33$ deaths from lung cancer in 131200 women-years in 1971 leads to a 95% CI per 100,000 WY for λ of

```
qgamma(c(0.025,0.975), c(33,34)) / 131200 * 1e5
## [1] 17.31378 35.32338
```

CI for the rate parameter λ using canned function

```
stats::poisson.test(x = 33, T = 131200)

## Exact Poisson test with 33 time base: 131200
## number of events = 33, time base = 131200, p-value < 2.2e-16
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##  0.0001731378 0.0003532338
## sample estimates:
##   event rate
## 0.0002515244
```

Motivating example

Poisson Model for Sampling Variability of a Count in a Given Amount of "Experience"

Inference regarding μ, based on observed count y

Inference regarding an event rate parameter λ, based on observed number of events y in a known amount of population-time (PT)

Test of $H_0 : \mu = \mu_0 \quad \Leftrightarrow \quad \lambda = \lambda_0$

Statistical evidence and the p-value

Recall:

- P-Value = Prob[y or more extreme | H_0]
- With 'more extreme' determined by whether H_{alt} is 1-sided or 2-sided.
- For a **formal test**, at level α, compare this P-value with α.

Example: Cancers surrounding nuclear stations

- Cancers in area surrounding the Douglas Point nuclear station

Example: Cancers surrounding nuclear stations

- Cancers in area surrounding the Douglas Point nuclear station
- Denote by $\{CY_1, CY_2, \ldots\}$ the numbers of Douglas Point child-years of experience in the various age categories that were pooled over.
- Denote by $\{\lambda_1^{Ont}, \lambda_2^{Ont}, \ldots\}$ the age-specific leukemia incidence rates during the period studied.

Example: Cancers surrounding nuclear stations

- Cancers in area surrounding the Douglas Point nuclear station
- Denote by $\{CY_1, CY_2, \ldots\}$ the numbers of Douglas Point child-years of experience in the various age categories that were pooled over.
- Denote by $\{\lambda_1^{Ont}, \lambda_2^{Ont}, \ldots\}$ the age-specific leukemia incidence rates during the period studied.
- If the underlying incidence rates in Douglas Point were the same as those in the rest of Ontario, the ***E***xpected total number of cases of leukemia for Douglas Point would be

$$E = \mu_0 = \sum_{ages} CY_i \times \lambda_i^{Ont} = 0.57.$$

Example: Cancers surrounding nuclear stations

- Cancers in area surrounding the Douglas Point nuclear station
- Denote by $\{CY_1, CY_2, \dots\}$ the numbers of Douglas Point child-years of experience in the various age categories that were pooled over.
- Denote by $\{\lambda_1^{Ont}, \lambda_2^{Ont}, \dots\}$ the age-specific leukemia incidence rates during the period studied.
- If the underlying incidence rates in Douglas Point were the same as those in the rest of Ontario, the ***E***xpected total number of cases of leukemia for Douglas Point would be

$$E = \mu_0 = \sum_{ages} CY_i \times \lambda_i^{Ont} = 0.57.$$

The actual total number of cases of leukemia ***O***bserved in Douglas Point was

$$O = y = \sum_{ages} O_i = 2.$$

Example: Cancers surrounding nuclear stations

- Cancers in area surrounding the Douglas Point nuclear station
- Denote by $\{CY_1, CY_2, \dots\}$ the numbers of Douglas Point child-years of experience in the various age categories that were pooled over.
- Denote by $\{\lambda_1^{Ont}, \lambda_2^{Ont}, \dots\}$ the age-specific leukemia incidence rates during the period studied.
- If the underlying incidence rates in Douglas Point were the same as those in the rest of Ontario, the ***E***xpected total number of cases of leukemia for Douglas Point would be

$$E = \mu_0 = \sum_{ages} CY_i \times \lambda_i^{Ont} = 0.57.$$

The actual total number of cases of leukemia ***O***bserved in Douglas Point was

$$O = y = \sum_{ages} O_i = 2.$$

Age *Standardized Incidence Ratio (SIR)* = $O/E = 2/0.57 = 3.5$.

Q: Is the $O = 2$ significantly higher than $E = 0.57$

Question:

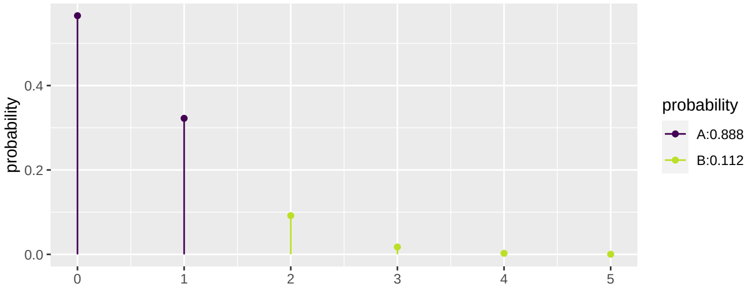
- Is the $y = 2$ cases of leukemia observed in the Douglas Point experience statistically significantly higher than the $E = 0.57$ cases "expected" for this many child-years of observation if in fact the rates in Douglas Point and the rest of Ontario were the same?

- Or, is the $y = 2$ observed in this community compatible with $H_0 : y \sim \text{Poisson}(\mu = 0.57)$?

A: Is the $O = 2$ significantly higher than $E = 0.57$

- Answer: Under H_0, the age-specific numbers of leukemias $\{y_1 = O_1,\ y_2 = O_2,\ \ldots\}$ in Douglas Point can be regarded as independent Poisson random variables, so their sum y can be regarded as a single Poisson random variable with $\mu = 0.57$.

```
mosaic::xppois(1, lambda = 0.57, lower.tail = FALSE)
```




```
## [1] 0.1121251
```

95% CI for the SIR by hand

- To get the CI for the SIR, divide the CI for Douglas Point μ_{DP} by the null $\mu_0 = 0.57$ (Ontario scaled down to the same size and age structure as Douglas Point.) We treat it as a constant because the Ontario rates used in the scaling are measured with much less sampling variability that the Douglas Point ones.

95% CI for the SIR by hand

- To get the CI for the SIR, divide the CI for Douglas Point μ_{DP} by the null $\mu_0 = 0.57$ (Ontario scaled down to the same size and age structure as Douglas Point.) We treat it as a constant because the Ontario rates used in the scaling are measured with much less sampling variability that the Douglas Point ones.
- The $y = 2$ cases translates to
  - 95% CI for $\mu_{DP} \rightarrow [0.24, 7.22]$
  - 95% CI for the SIR $\rightarrow$ [0.24/0.57, 7.22/0.57]=[0.4, 12.7].

95% CI for the SIR using canned function

- We can *trick* `stats::poisson.test` to get the same CI by putting time as 0.57:

```
stats::poisson.test(x=2,T=0.57)

## Exact Poisson test with 2 time base: 0.57
## number of events = 2, time base = 0.57, p-value = 0.1121
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##   0.4249286 12.6748906
## sample estimates:
## event rate
##   3.508772
```

Session Info

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.13.so

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] DT_0.16           mosaic_1.7.0      Matrix_1.3-2      mosaicData_0.20.1
 [5] ggformula_0.9.4   ggstance_0.3.4    lattice_0.20-41   kableExtra_1.2.1
 [9] socviz_1.2        gapminder_0.3.0   here_0.1          NCStats_0.4.7
[13] FSA_0.8.30        forcats_0.5.1     stringr_1.4.0     dplyr_1.0.7
[17] purrr_0.3.4       readr_1.4.0       tidyr_1.1.4       tibble_3.1.5
[21] ggplot2_3.3.5     tidyverse_1.3.0   knitr_1.36

loaded via a namespace (and not attached):
 [1] fs_1.5.0           lubridate_1.7.9    webshot_0.5.2      httr_1.4.2
 [5] rprojroot_2.0.2    latex2exp_0.4.0    backports_1.2.1    utf8_1.2.2
 [9] R6_2.5.1           DBI_1.1.1          colorspace_2.0-2   withr_2.4.2
[13] tidyselect_1.1.1   gridExtra_2.3      leaflet_2.0.3      curl_4.3.2
[17] compiler_4.1.1     cli_3.0.1          rvest_1.0.0        pacman_0.5.1
[21] xml2_1.3.2         ggdendro_0.1.22    labeling_0.4.2     mosaicCore_0.8.0
[25] scales_1.1.1       digest_0.6.28      foreign_0.8-81     rmarkdown_2.11.3
[29] rio_0.5.16         pkgconfig_2.0.3    htmltools_0.5.2    highr_0.9
[33] dbplyr_1.4.4       fastmap_1.1.0      htmlwidgets_1.5.3  rlang_0.4.12
[37] readxl_1.3.1       rstudioapi_0.13    farver_2.1.0       generics_0.1.0
[41] jsonlite_1.7.2     crosstalk_1.1.1    zip_2.2.0          car_3.0-9
[45] magrittr_2.0.1     Rcpp_1.0.7         munsell_0.5.0      fansi_0.5.0
[49] abind_1.4-5        lifecycle_1.0.1    stringi_1.7.5      carData_3.0-4
[53] MASS_7.3-53.1      plyr_1.8.6         grid_4.1.1         blob_1.2.1
[57] ggrepel_0.8.2      crayon_1.4.1       cowplot_1.1.0      haven_2.3.1
[61] splines_4.1.1      hms_1.1.1          pillar_1.6.4       reprex_0.3.0
[65] glue_1.4.2         evaluate_0.14      data.table_1.14.2  modelr_0.1.8
[69] vctrs_0.3.8        tweenr_1.0.1       cellranger_1.1.0   gtable_0.3.0
[73] polyclip_1.10-0    assertthat_0.2.1   TeachingDemos_2.12 xfun_0.26
[77] ggforce_0.3.2      openxlsx_4.1.5     broom_0.7.9        viridisLite_0.4.0
```