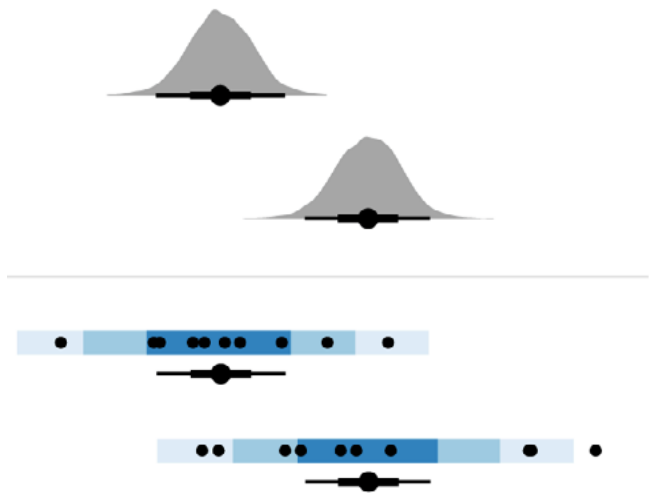
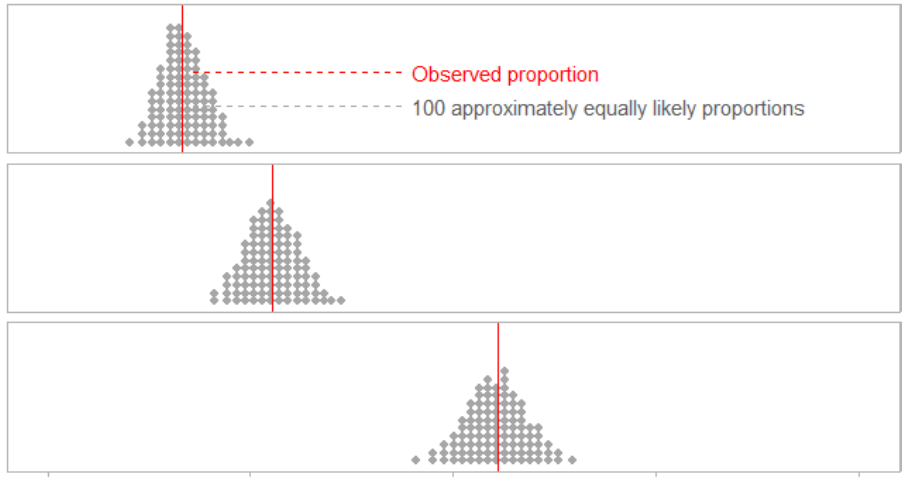
Visualizing distributions and uncertainty using ggdist

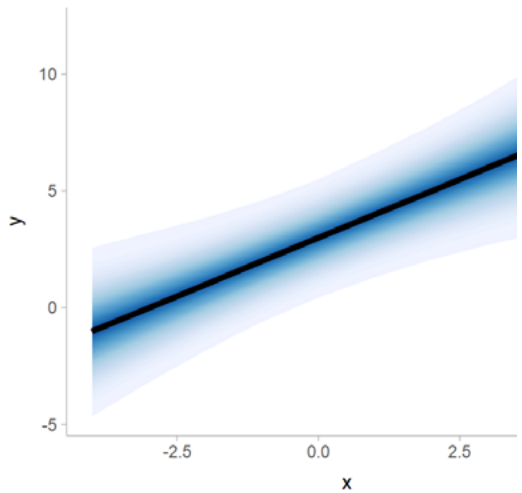
RStudio::conf(2022) talk

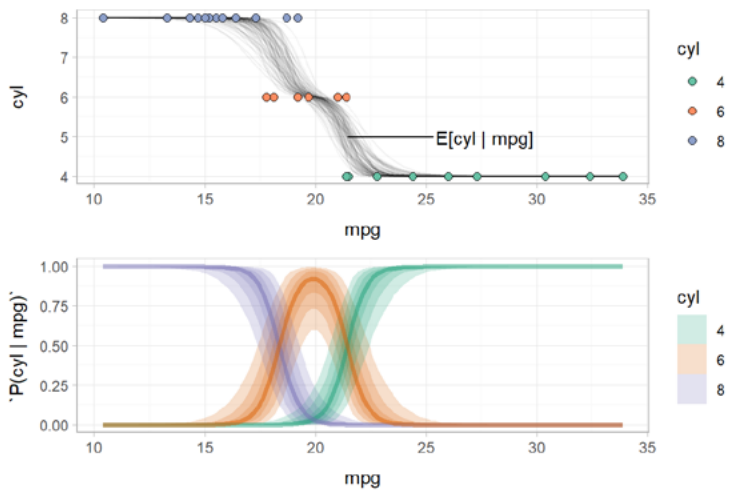
Matthew Kay
Assistant Professor
Depts of Computer Science and Communication
Northwestern University

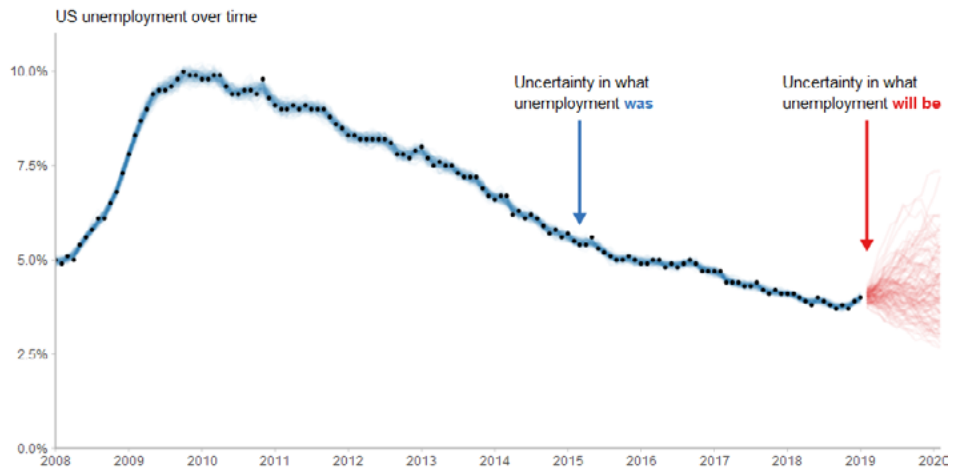
Visualizing distributions is a pain...

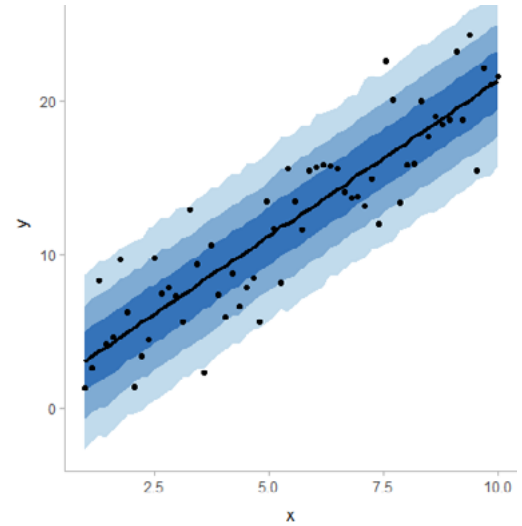



Observed proportion
100 approximately equally likely proportions


y
x


cyl
E[cyl | mpg]
mpg
`P(cyl | mpg)`
cyl
4
6
8


US unemployment over time
Uncertainty in what unemployment was
Uncertainty in what unemployment will be


y
x

1. Systematic way to think about building uncertainty visualizations

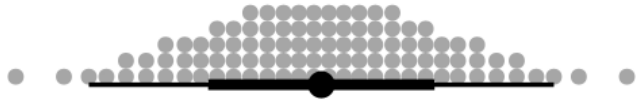
1. Systematic way to think about building uncertainty visualizations

2. A taste for how to do it using ggdist

geom_slabinterval()

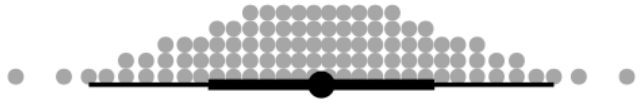
geom_slabinterval()

geom_dotsinterval()

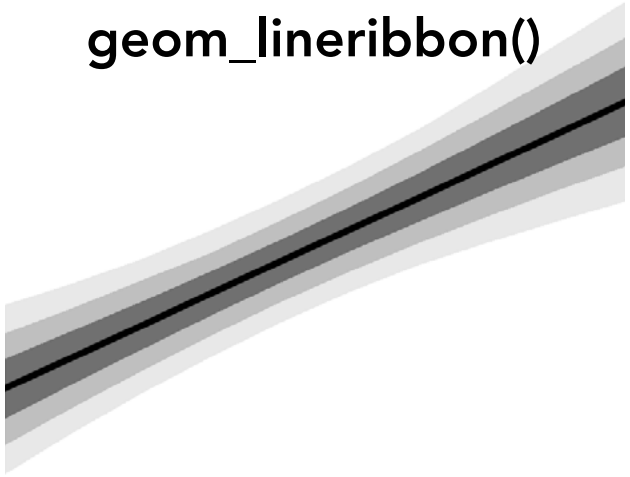


geom_slabinterval()

geom_dotsinterval()



geom_lineribbon()



Systematic uncertainty visualization...

Systematic uncertainty visualization...

1. Derive an uncertainty distribution

$f_X(x)$ *density function*

$F_X(x)$ *cumulative distribution function (CDF)*

$F_X^{-1}(p)$ *quantile function*

Systematic uncertainty visualization...

Bayesian

1. Derive a posterior distribution

   Let $x^{(1)}, \ldots, x^{(k)}$ be draws from $p(x \mid data)$

   $f_X(x)$ *kernel density estimator* of all $x^{(k)}$

   $F_X(x)$ *empirical CDF* of all $x^{(k)}$

   $F_X^{-1}(p)$ *empirical quantile function* of all $x^{(k)}$

Systematic uncertainty visualization...

Frequentist

1. Derive a confidence distribution

$\mathrm{t}(df, m, se)$	*scaled/shifted Student's t*
$f_X(x) = f_{\mathrm{t}}(x \mid df, m, se)$	*density*
$F_X(x) = F_{\mathrm{t}}(x \mid df, m, se)$	*CDF*
$F_X^{-1}(p) = F_{\mathrm{t}}^{-1}(p \mid df, m, se)$	*quantile function*

Systematic uncertainty visualization...

1. Derive an uncertainty distribution

$f_X(x)$ *density function*

$F_X(x)$ *cumulative distribution function (CDF)*

$F_X^{-1}(p)$ *quantile function*

Systematic uncertainty visualization...

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties
   onto visual channels (aesthetics)

Systematic uncertainty visualization...

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

$f_X(x)$ → *thickness*

Systematic uncertainty visualization...

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

   $f_X(x)$ → *thickness*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

$f_X(x)$ → *thickness*

$F_X^{-1}(0.975)$ → x_{min}

$F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(thickness = stat(pdf))
)
```

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

   $f_X(x)$ → *thickness*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(thickness = stat(pdf)),
  side = "top"
)
```

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

   $f_X(x)$ → *thickness*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(thickness = stat(pdf)),
  side = "bottom"
)
```

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

   $f_X(x)$ → *thickness*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(thickness = stat(pdf)),
  side = "both"
)
```

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

   $f_X(x)$ → *thickness*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(thickness = stat(cdf)),
  side = "both"
)
```

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

   $F_X(x)$ → *thickness*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(alpha = stat(pdf)),
  side = "both"
)
```

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

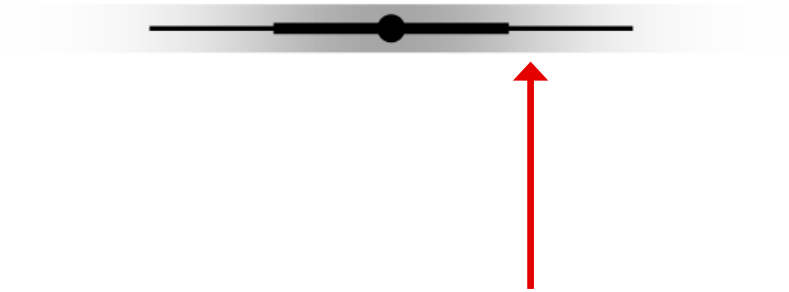
   $f_X(x)$ → *alpha*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(alpha = stat(pdf)),
  side = "both"
)
```



true gradients! (R >= 4.1)

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

   $f_X(x)$ → *alpha*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

```
stat_slabinterval(
  aes(alpha = stat(pdf)),
  side = "both"
)
```

1. Derive an uncertainty distribution

   $f_X(x)$ *density function*

   $F_X(x)$ *cumulative distribution function (CDF)*

   $F_X^{-1}(p)$ *quantile function*

2. Map distribution properties onto visual channels (aesthetics)

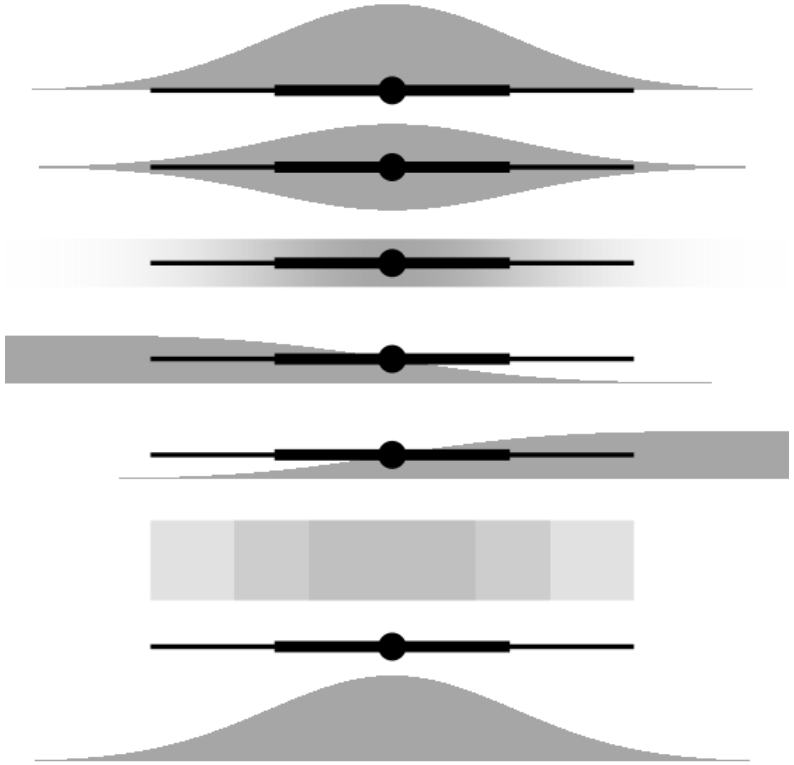
   $f_X(x)$ → *alpha*

   $F_X^{-1}(0.975)$ → x_{min}

   $F_X^{-1}(0.025)$ → x_{max}

Systematic uncertainty visualization...

`stat_slabinterval` family



1. Derive an uncertainty distribution

   $f_X(x)$ *density function*
   $F_X(x)$ *cumulative distribution function (CDF)*
   $F_X^{-1}(p)$ *quantile function*

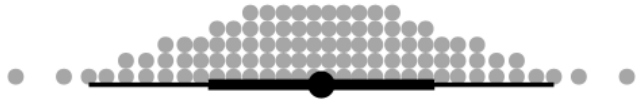
2. Map distribution properties onto visual channels (aesthetics)

   $f_X(x)$ → *alpha*

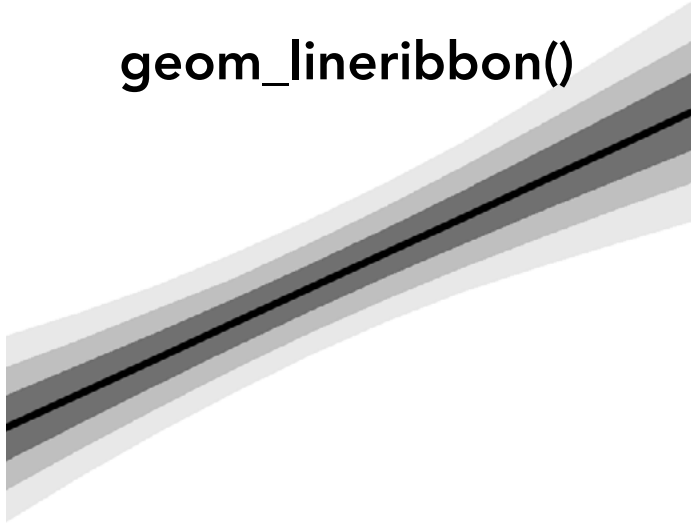
   $F_X^{-1}(0.975)$ → x_{min}
   $F_X^{-1}(0.025)$ → x_{max}

geom_slabinterval()

geom_dotsinterval()



geom_lineribbon()



Matthew Kay
@mjskay · Apr 22

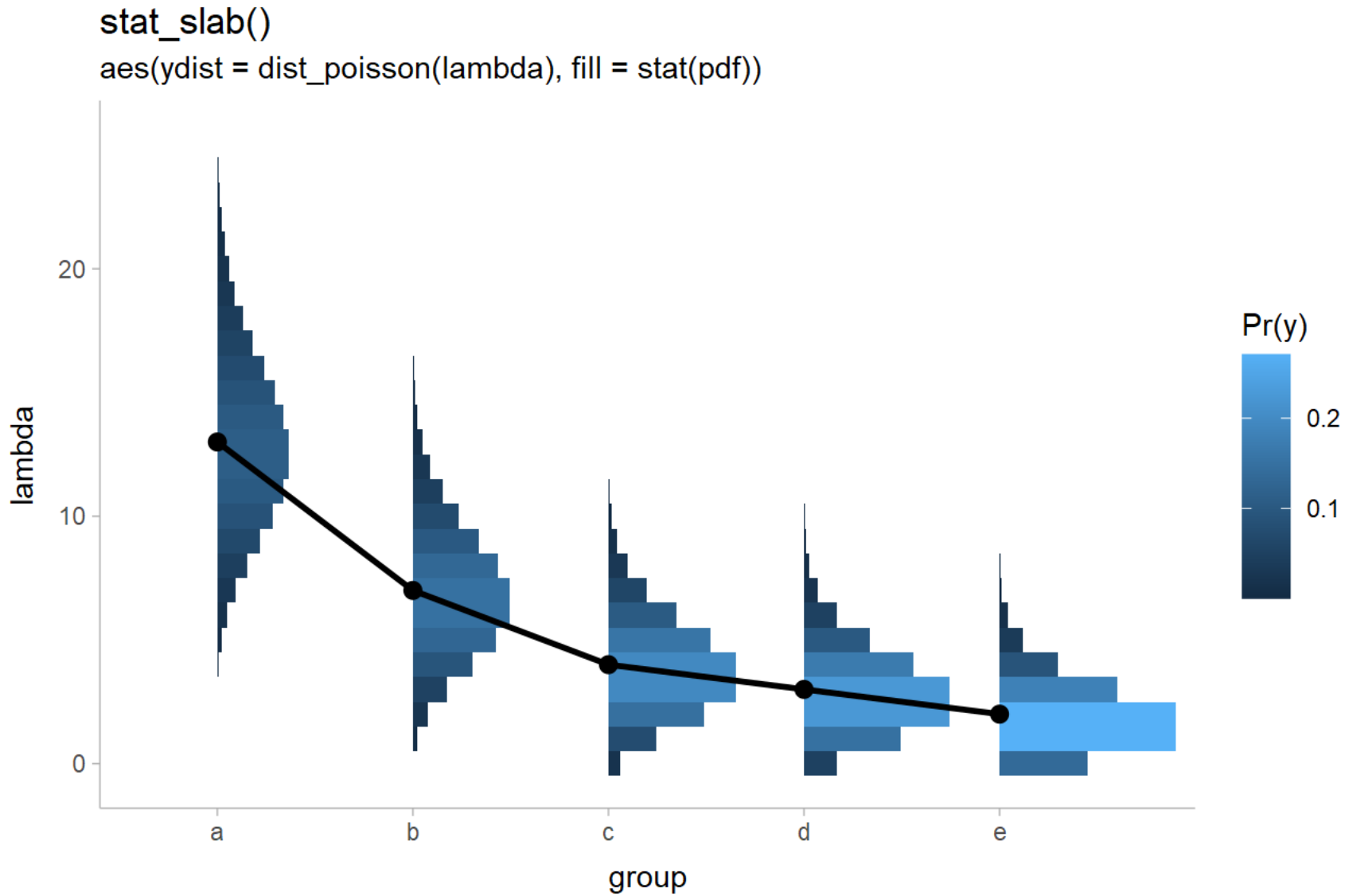
My #rstudioconf talk on {ggdist} got accepted as a lightning talk! Get ready for a whirlwind tour of as much cool stuff about distributional visualization as I can fit in 5 minutes! It's gonna be great!!

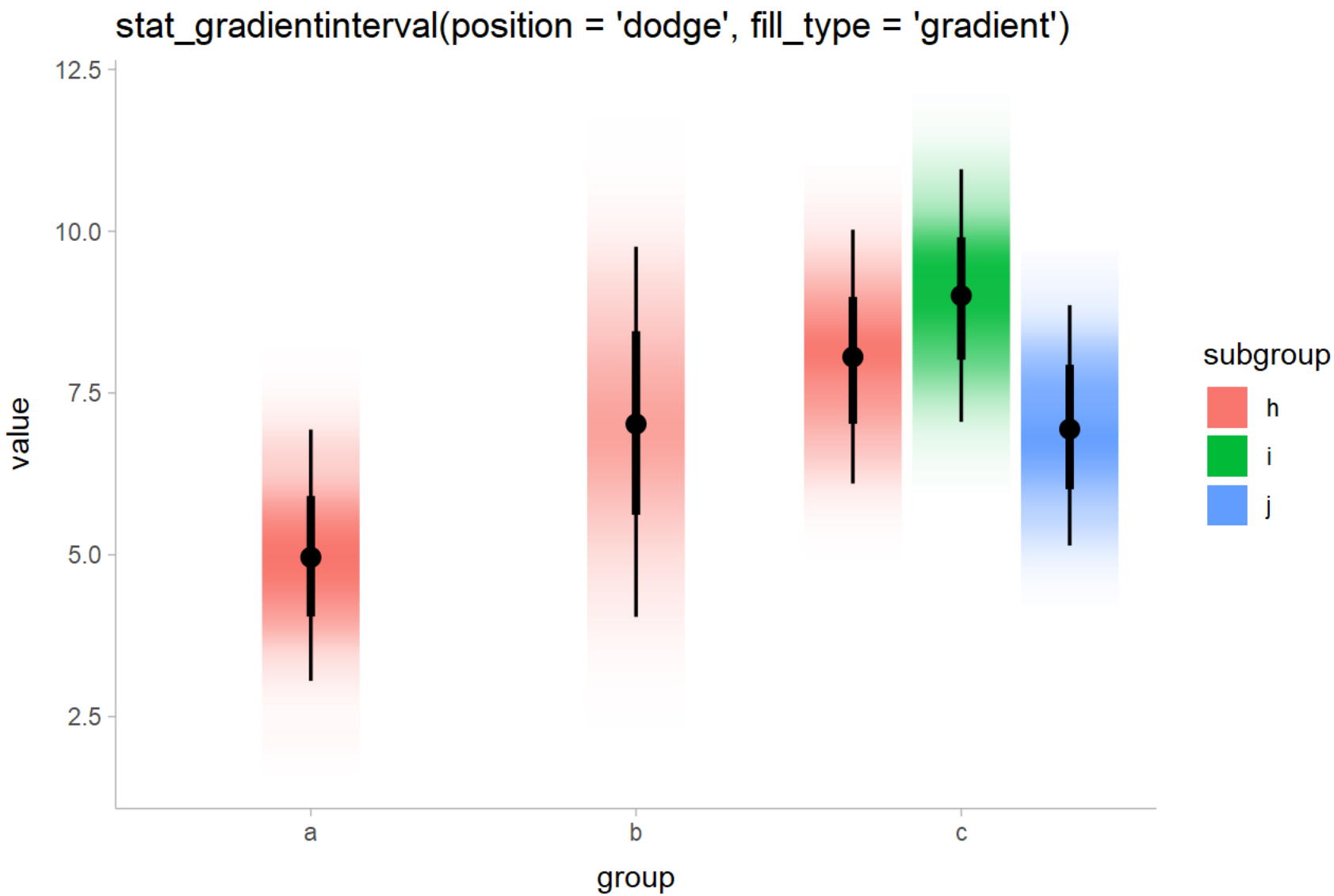
4 13 163

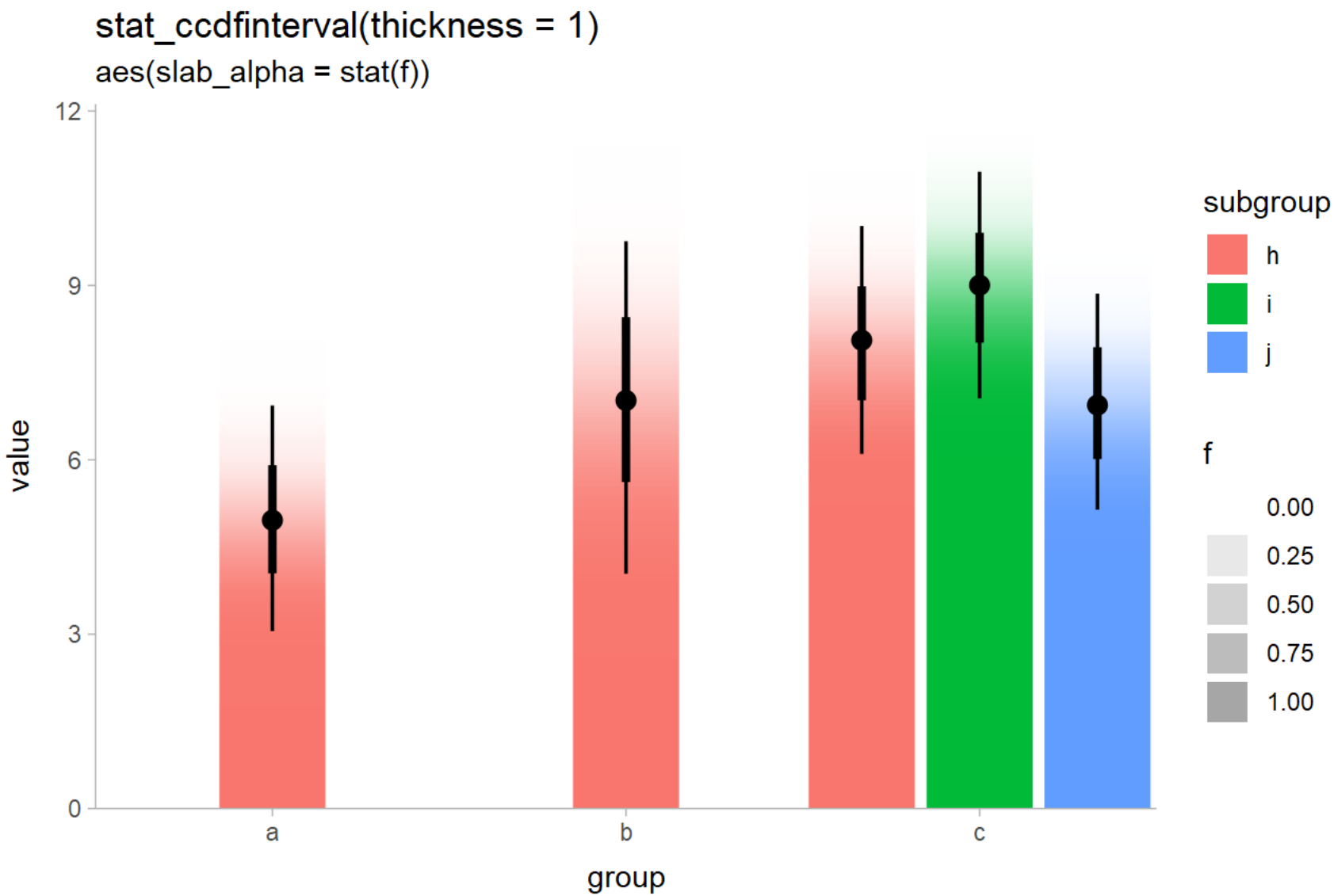
Matthew Kay
@mjskay

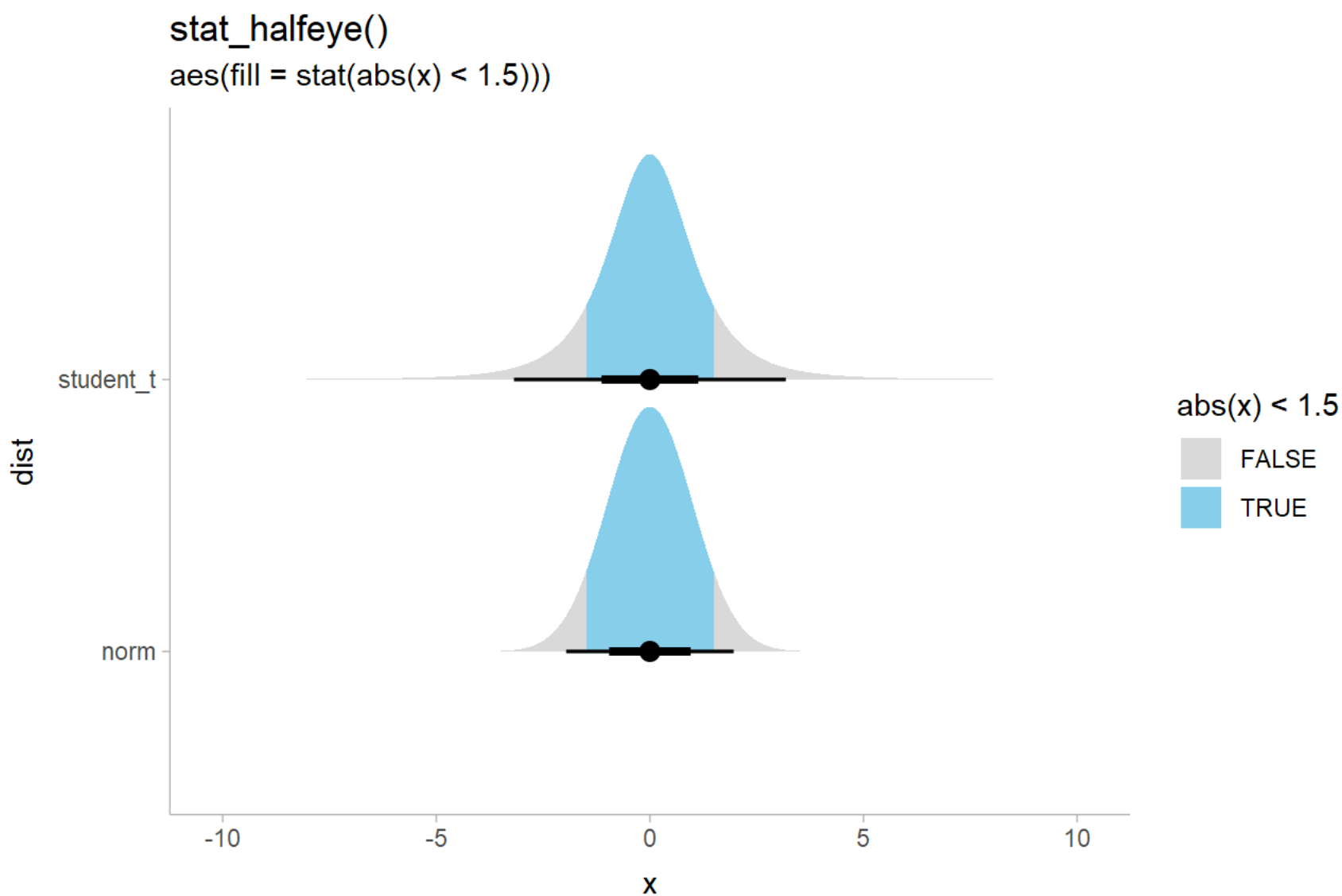
it will be like a twitter thread directly into your skull

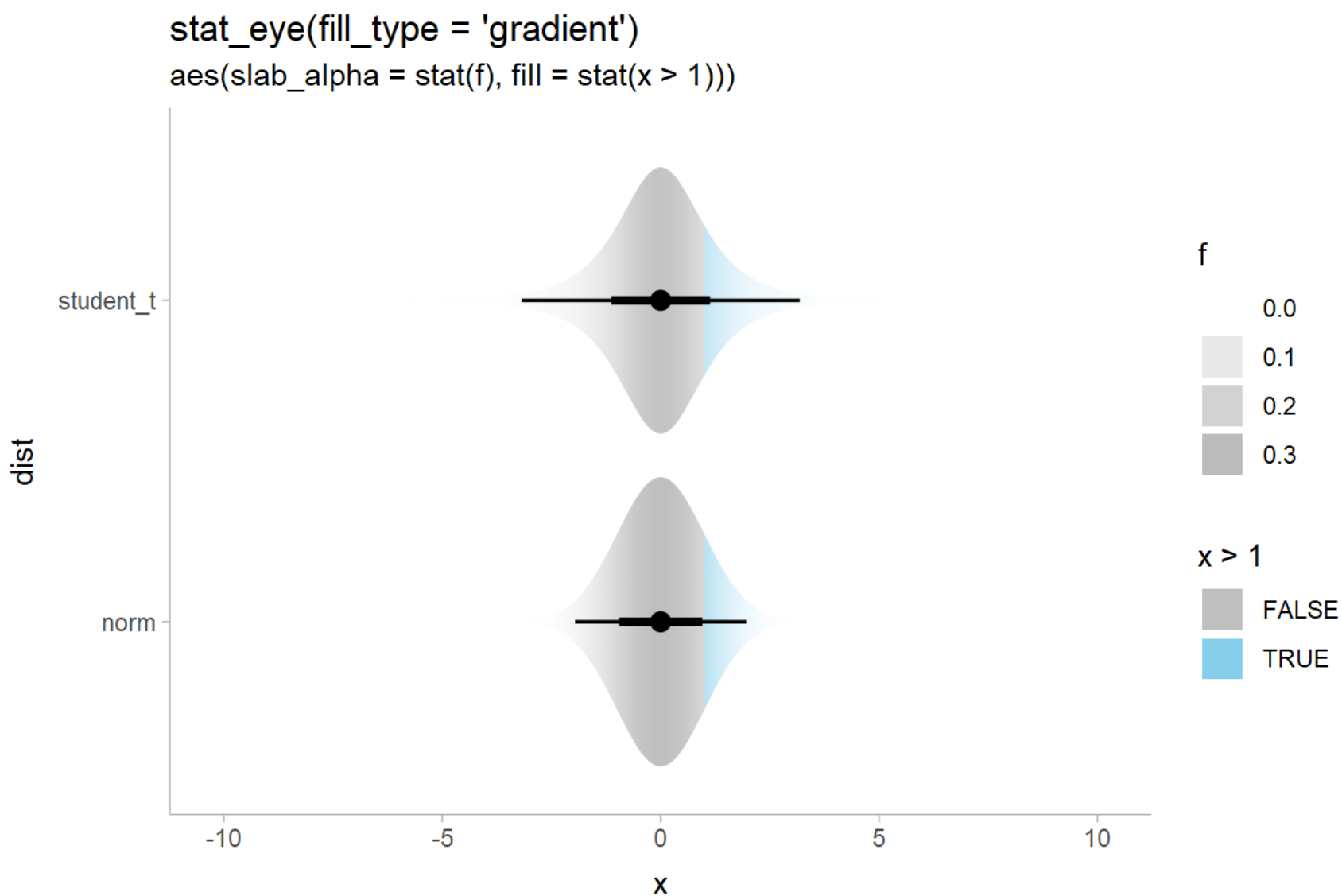
LET'S DO THIS

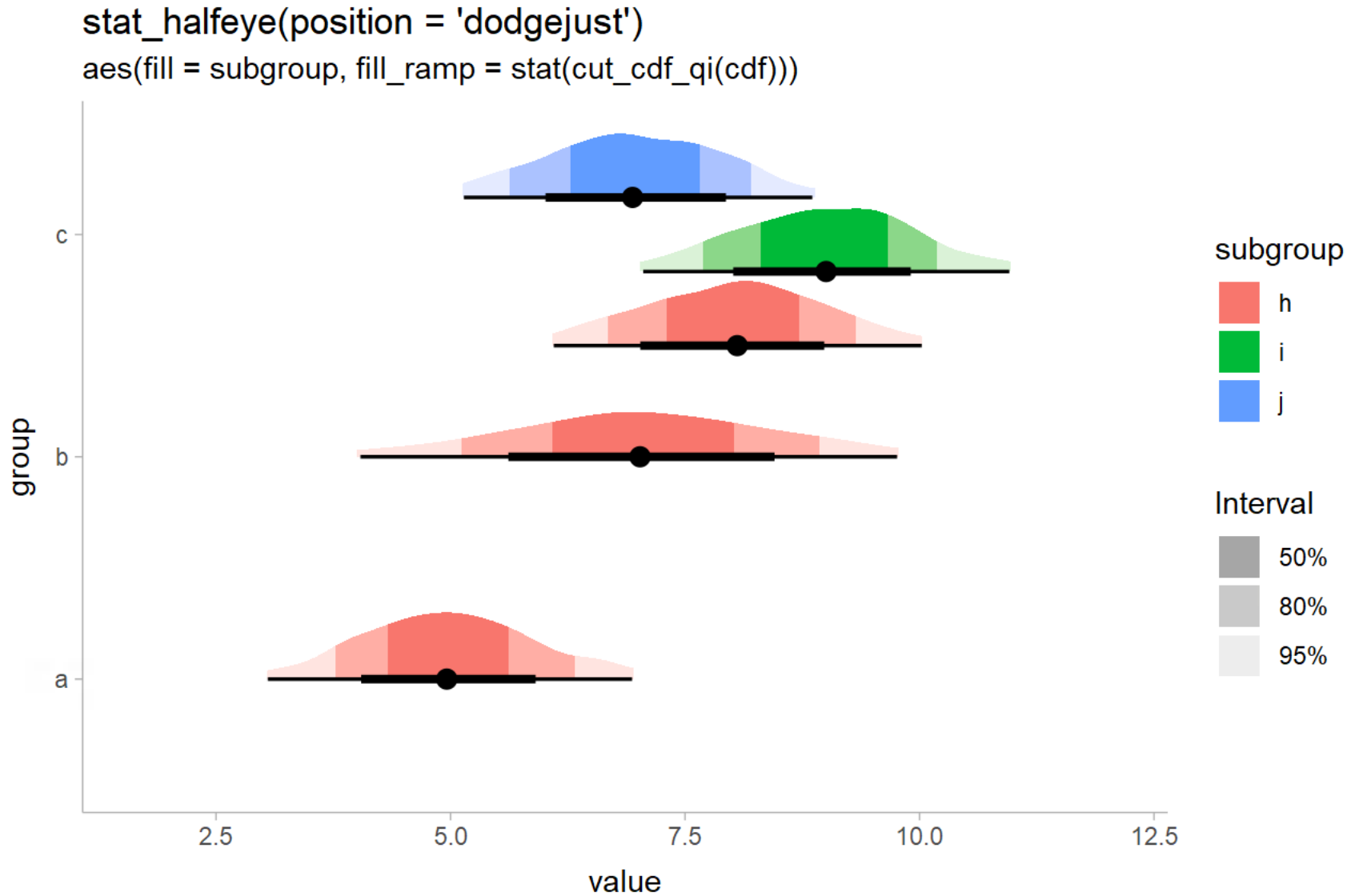

stat_slab()
aes(ydist = dist_poisson(lambda), fill = stat(pdf))
lambda
20
10
0
a
b
c
d
e
group
Pr(y)
0.2
0.1

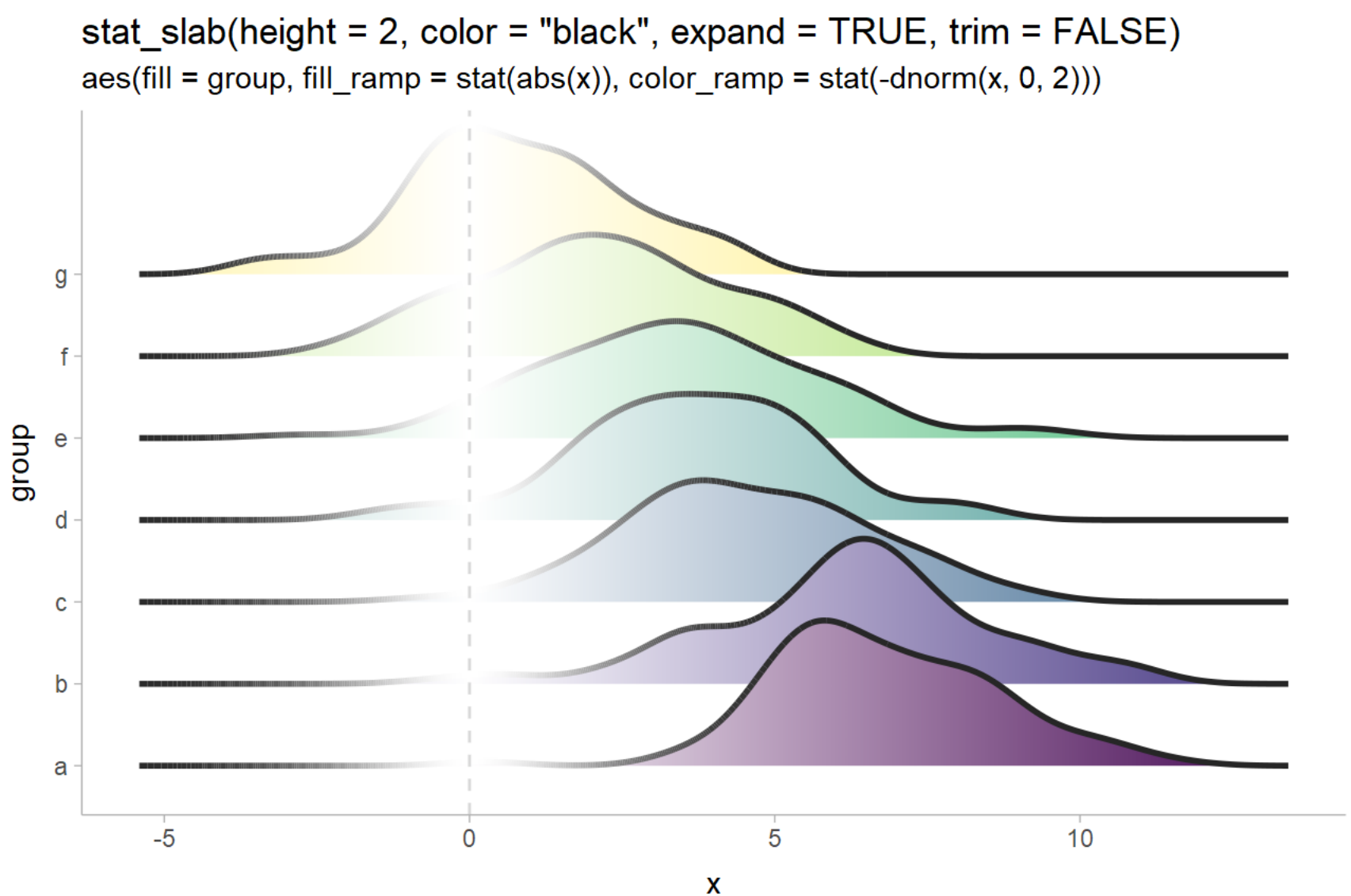

stat_gradientinterval(position = 'dodge', fill_type = 'gradient')
12.5
10.0
7.5
5.0
2.5
value
a
b
c
group
subgroup
h
i
j


stat_ccdfinterval(thickness = 1)
aes(slab_alpha = stat(f))
12
9
6
3
0
value
a
b
c
group
subgroup
h
i
j
f
0.00
0.25
0.50
0.75
1.00

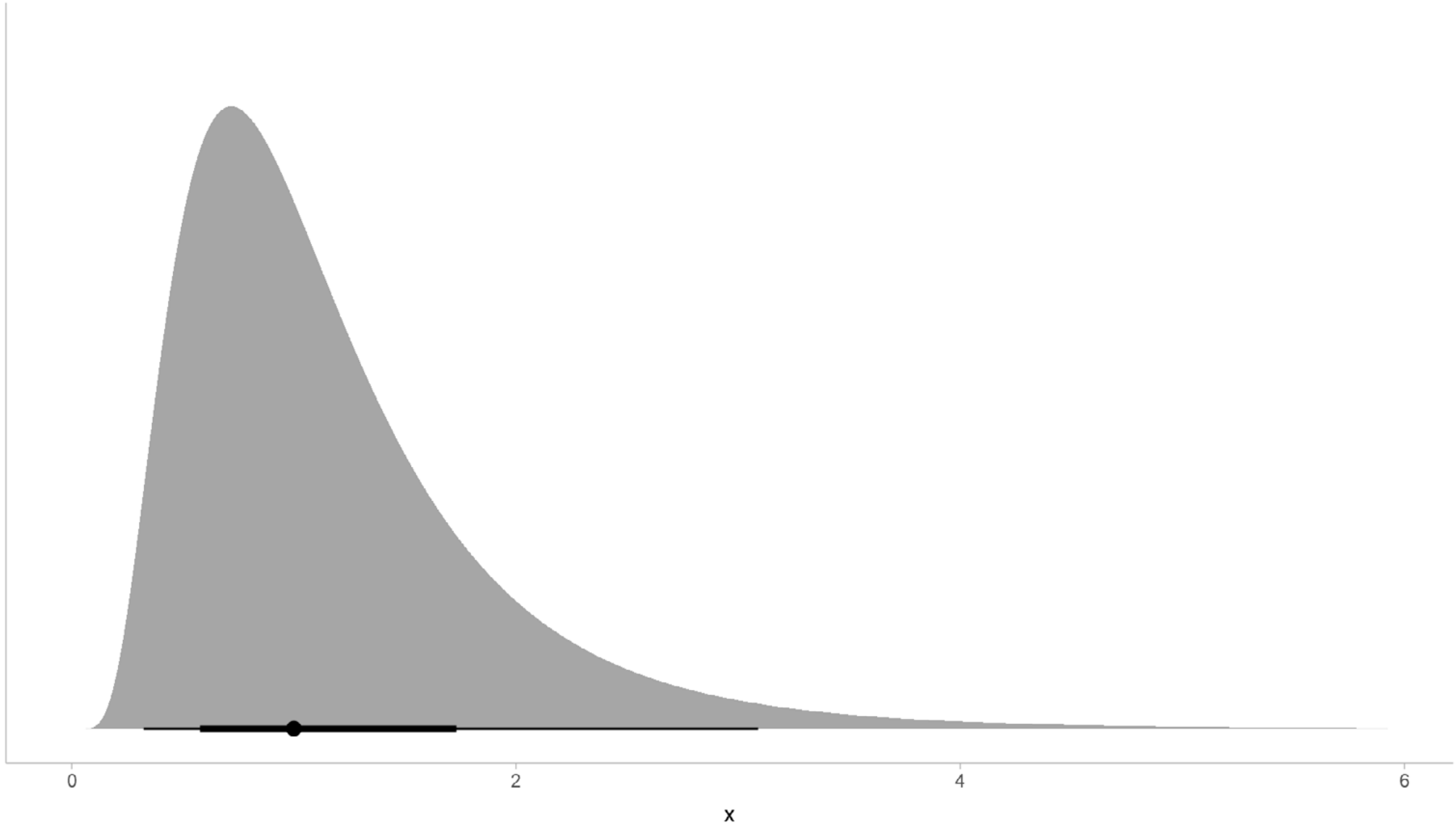

stat_halfeye()
aes(fill = stat(abs(x) < 1.5)))
student_t
norm
dist
-10
-5
0
5
10
x
abs(x) < 1.5
FALSE
TRUE


stat_eye(fill_type = 'gradient')
aes(slab_alpha = stat(f), fill = stat(x > 1)))
student_t
norm
dist
-10
-5
0
5
10
x
f
0.0
0.1
0.2
0.3
x > 1
FALSE
TRUE

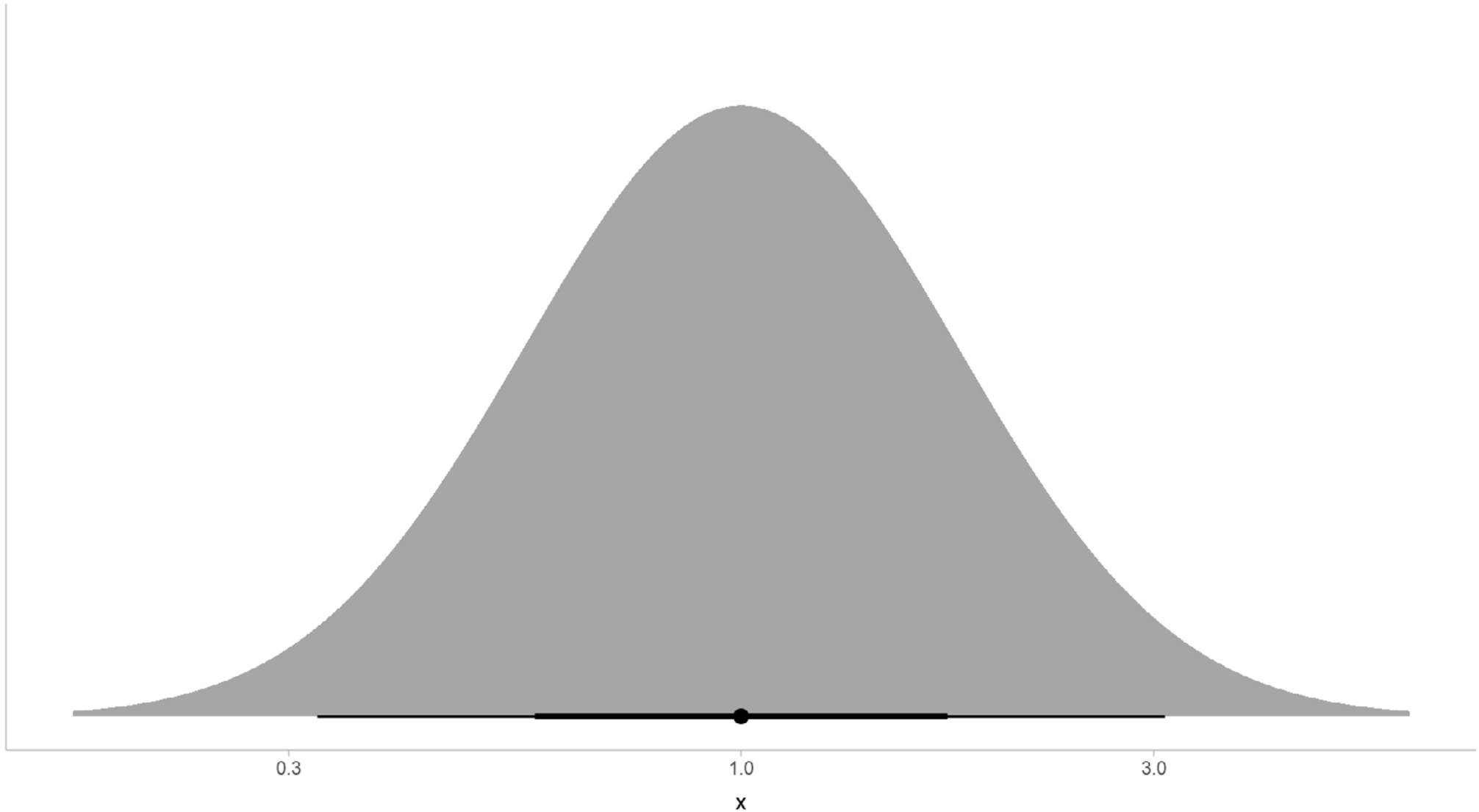

stat_halfeye(position = 'dodgejust')
aes(fill = subgroup, fill_ramp = stat(cut_cdf_qi(cdf)))
group
c
b
a
2.5
5.0
7.5
10.0
12.5
value
subgroup
h
i
j
Interval
50%
80%
95%

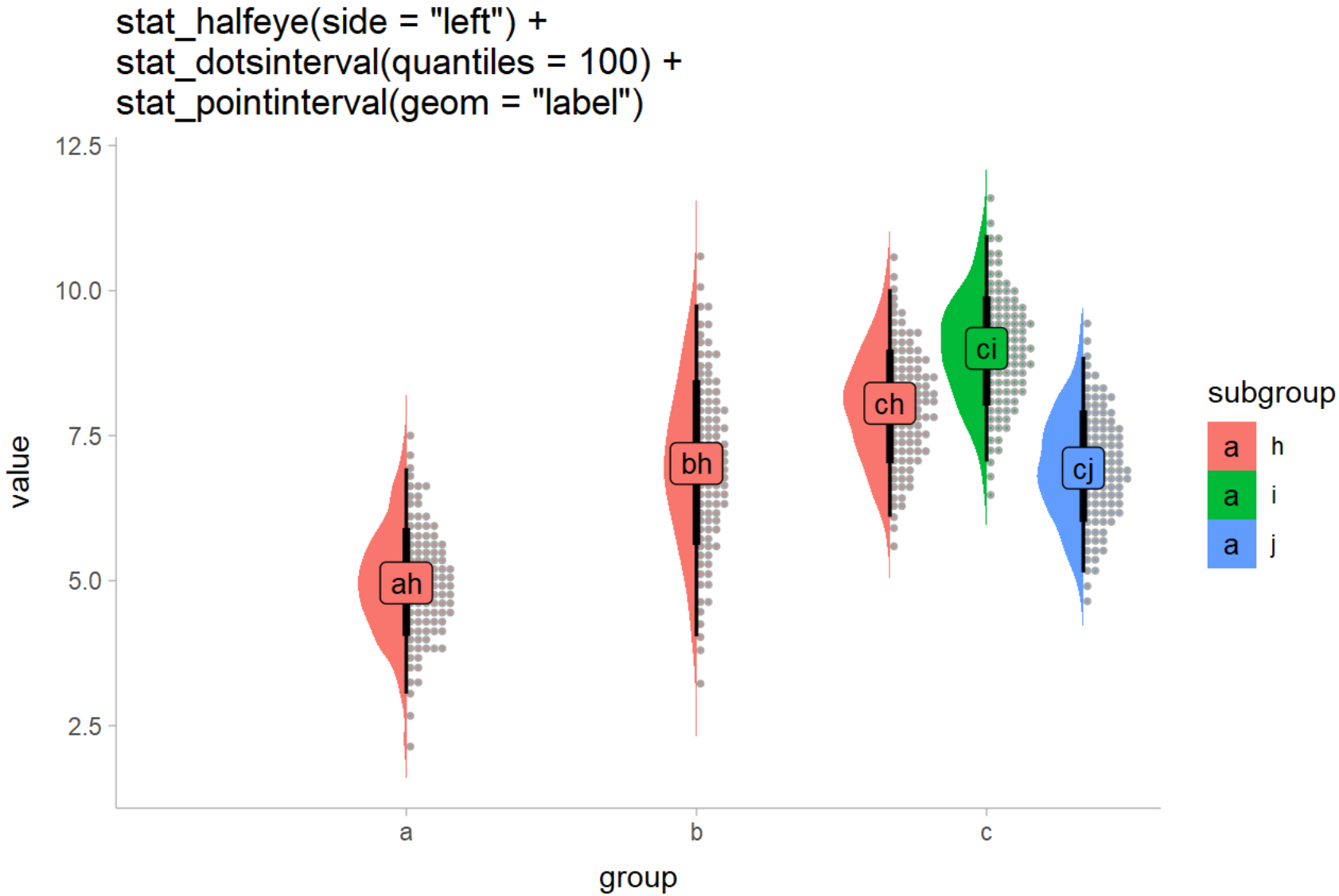

stat_slab(height = 2, color = "black", expand = TRUE, trim = FALSE)
aes(fill = group, fill_ramp = stat(abs(x)), color_ramp = stat(-dnorm(x, 0, 2)))
group
g
f
e
d
c
b
a
-5
0
5
10
x

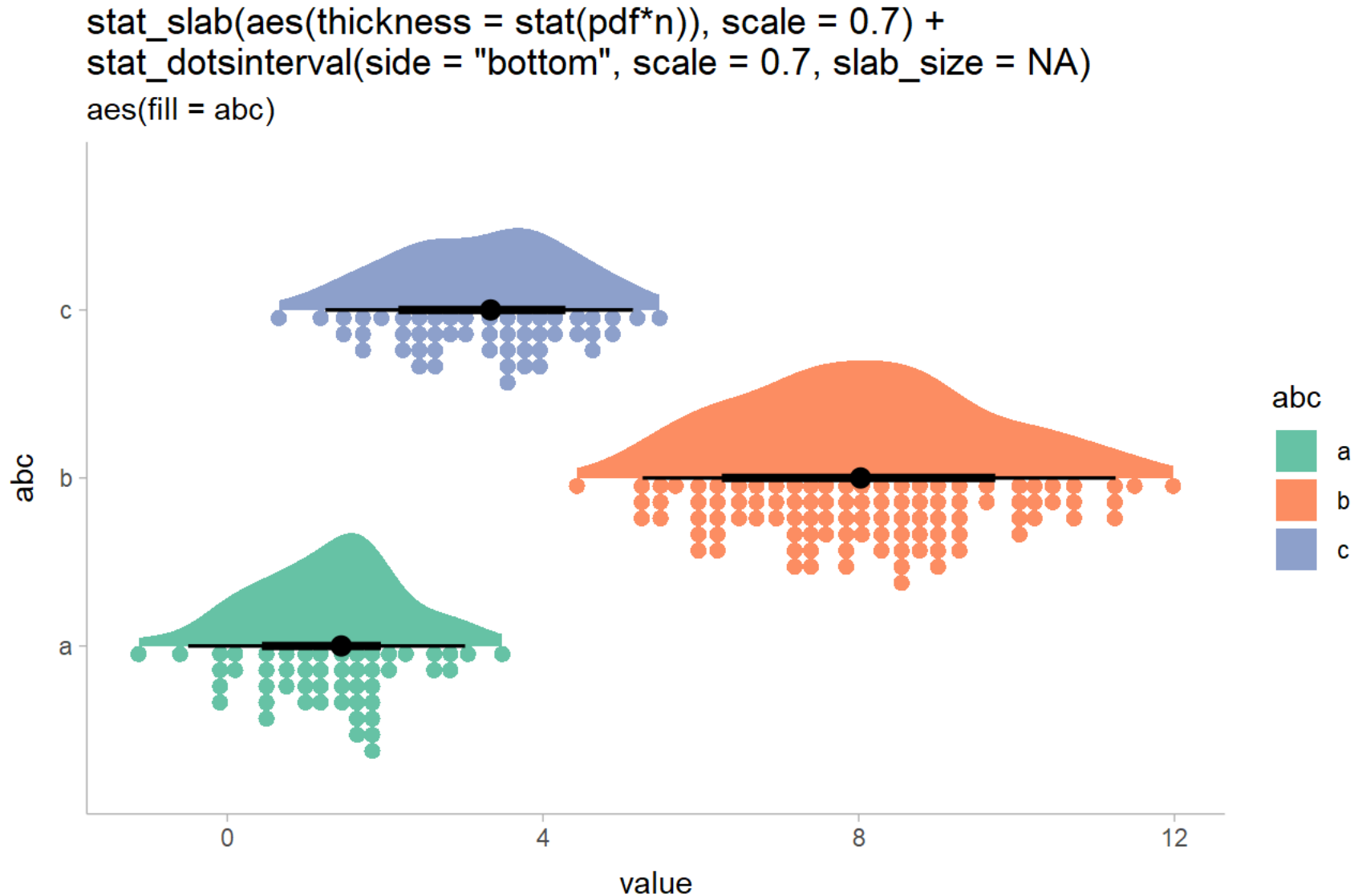
geom_halfeye()
aes(dist = dist_wrap("lnorm", 0, log(10)/4))



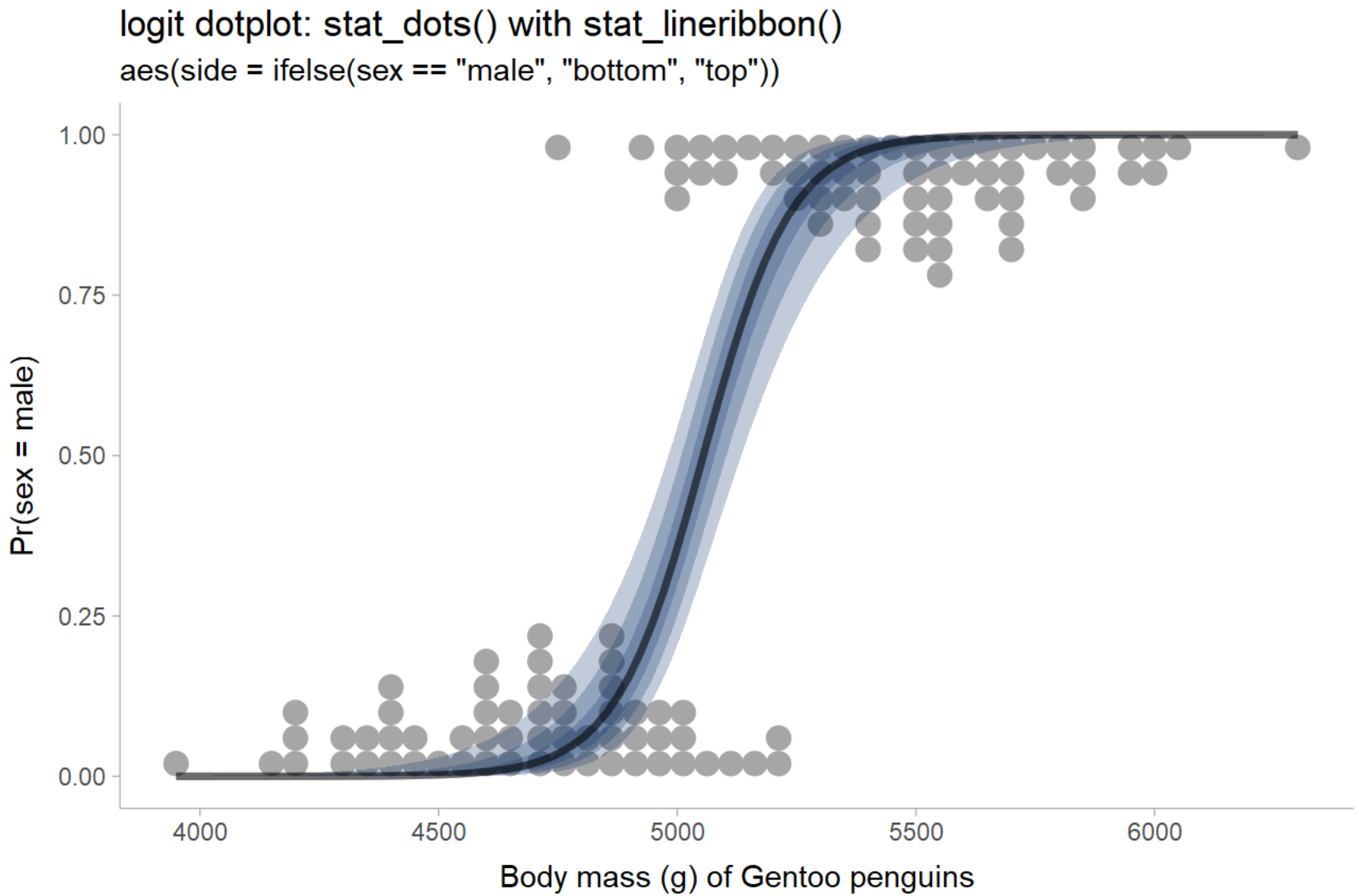
geom_halfeye()
aes(dist = dist_wrap("lnorm", 0, log(10)/4))
scale_x_log10()

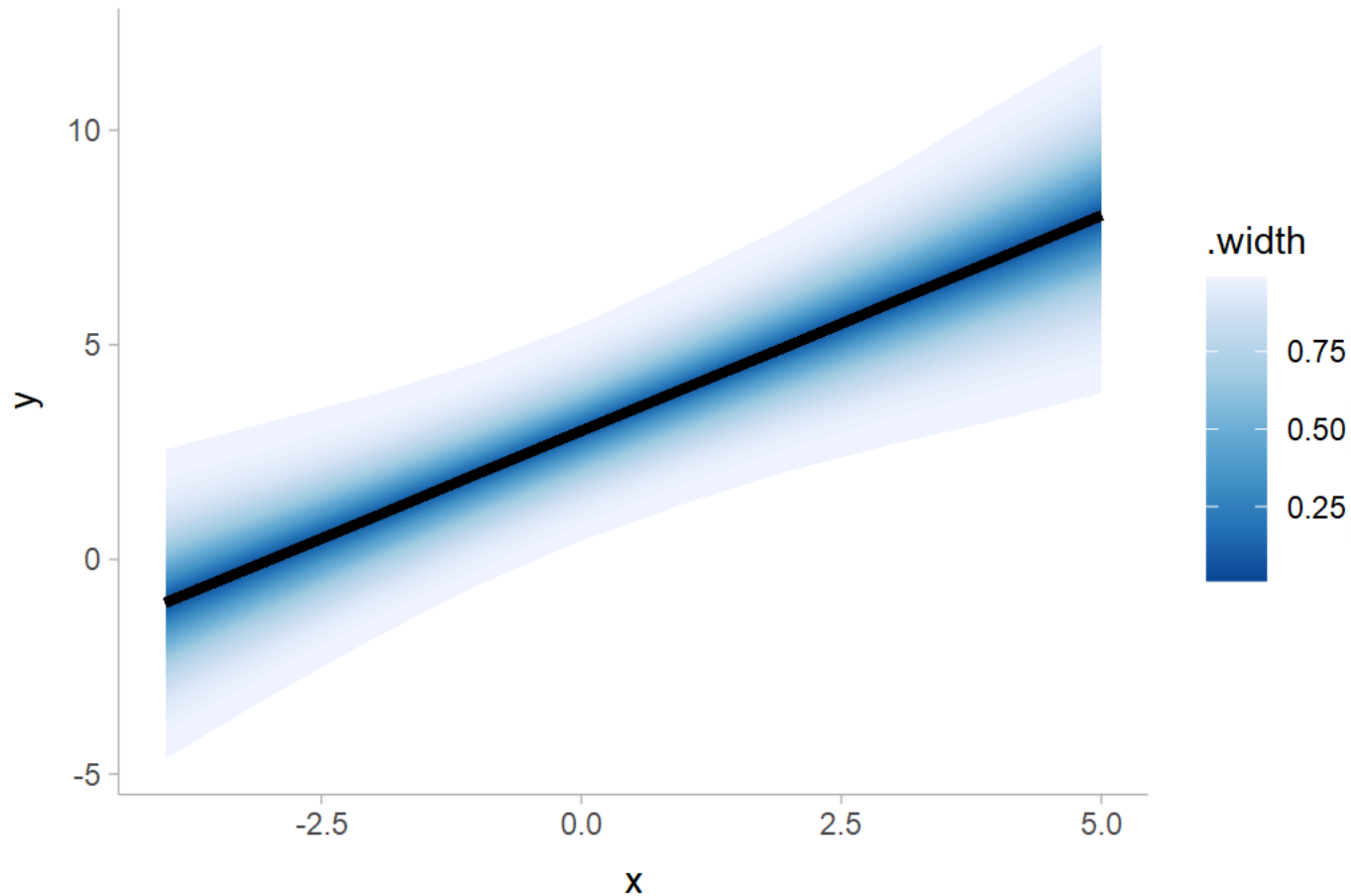


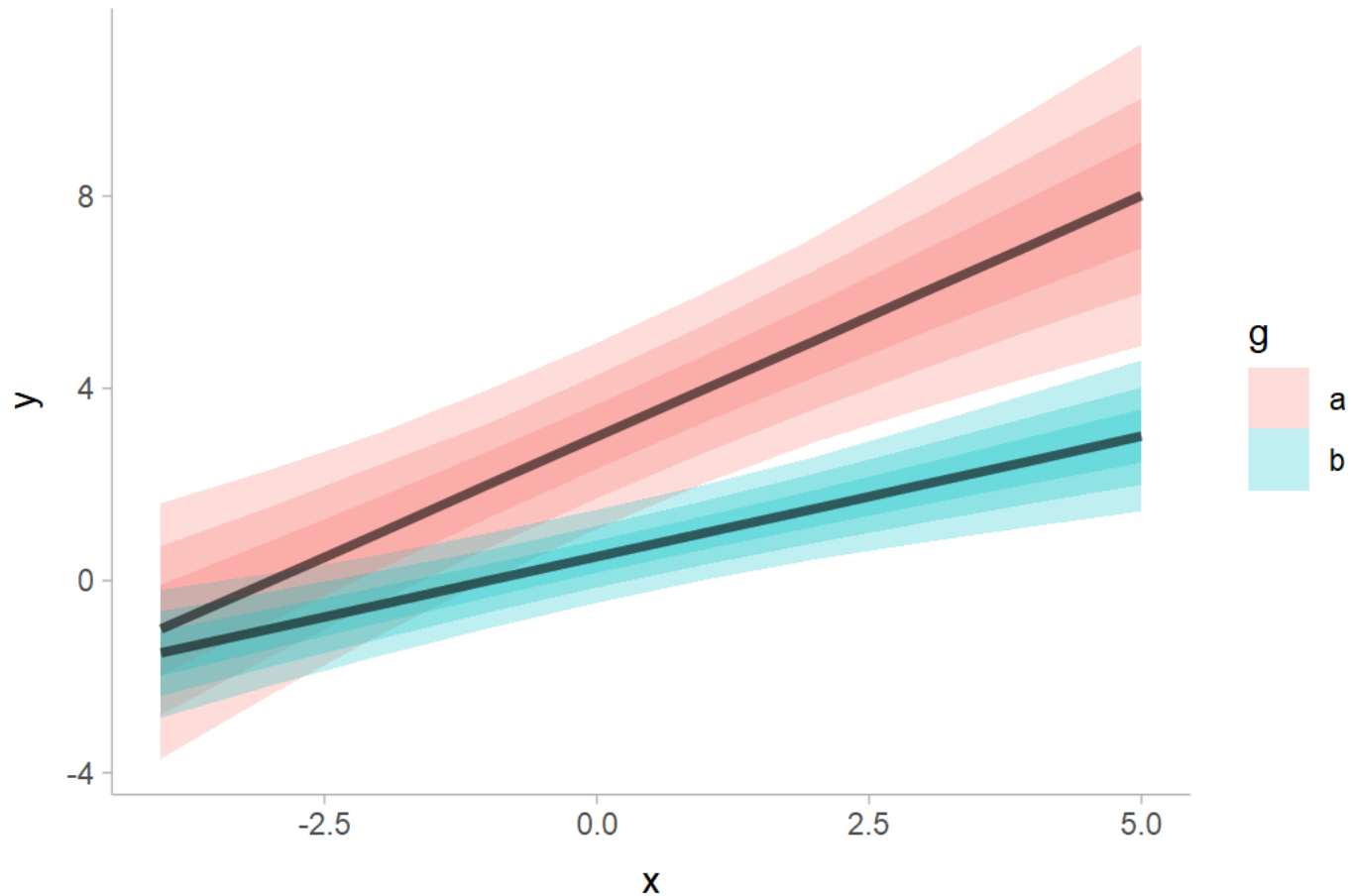

stat_halfeye(side = "left") +
stat_dotsinterval(quantiles = 100) +
stat_pointinterval(geom = "label")
12.5
10.0
7.5
5.0
2.5
value
ah
bh
ch
ci
cj
a
b
c
group
subgroup
a h
a i
a j

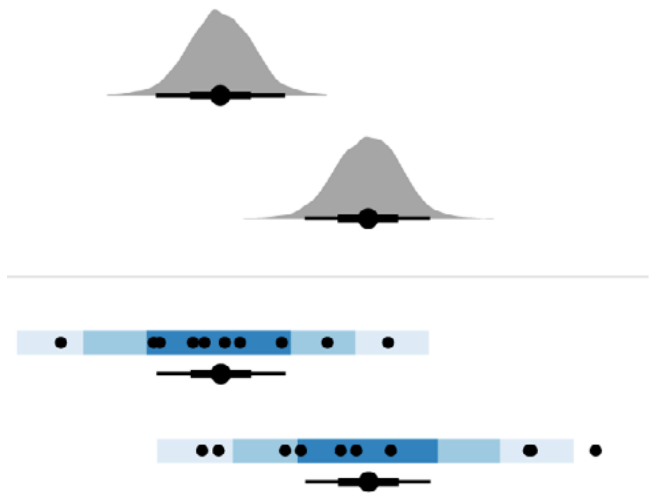

stat_slab(aes(thickness = stat(pdf*n)), scale = 0.7) +
stat_dotsinterval(side = "bottom", scale = 0.7, slab_size = NA)
aes(fill = abc)
abc
c
b
a
0
4
8
12
value
abc
a
b
c

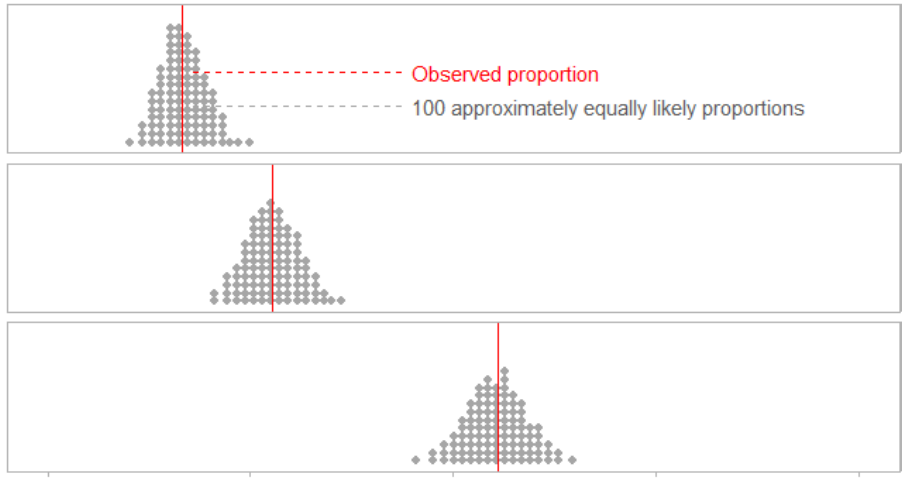
dotsinterval sizing goes here

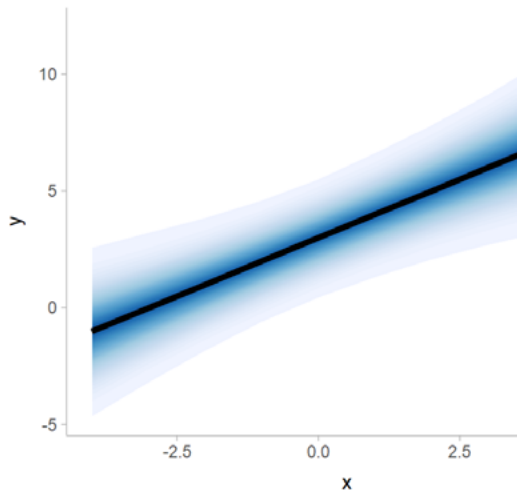

logit dotplot: stat_dots() with stat_lineribbon()
aes(side = ifelse(sex == "male", "bottom", "top"))
1.00
0.75
0.50
0.25
0.00
Pr(sex = male)
4000
4500
5000
5500
6000
Body mass (g) of Gentoo penguins

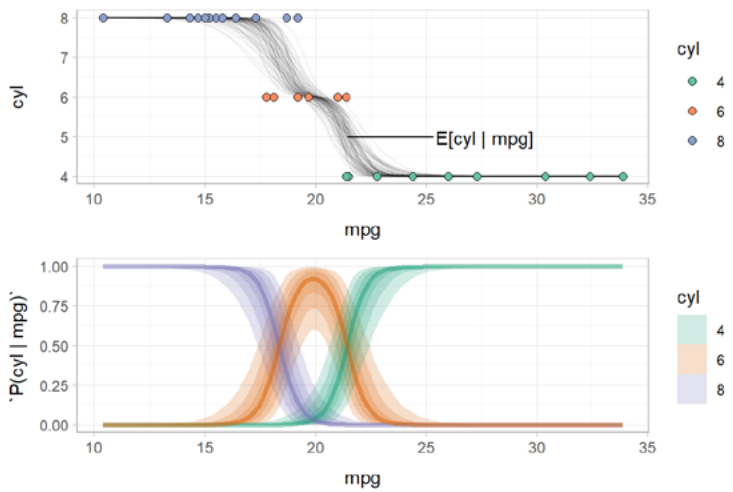

10
5
0
-5
y
-2.5
0.0
2.5
5.0
x
.width
0.75
0.50
0.25

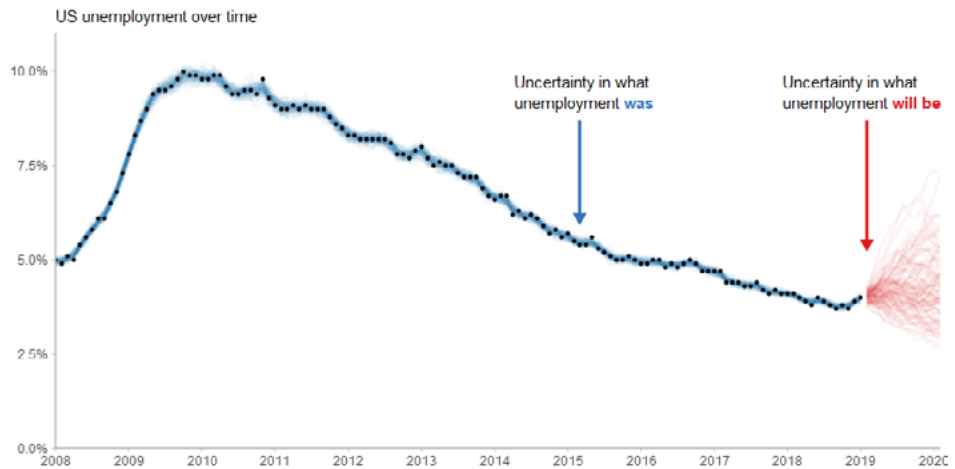

8
4
0
-4
y
-2.5
0.0
2.5
5.0
x
g
a
b

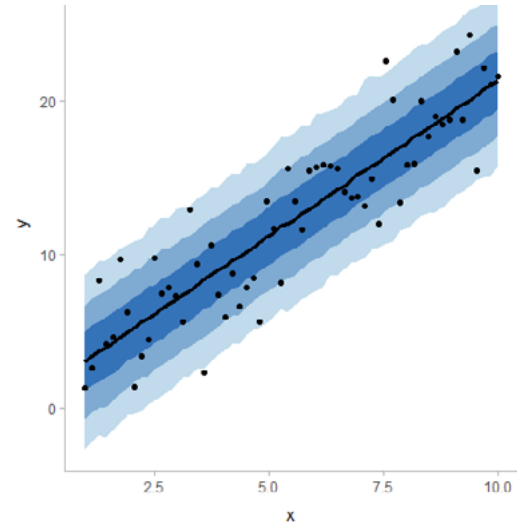



Observed proportion
100 approximately equally likely proportions


y
x

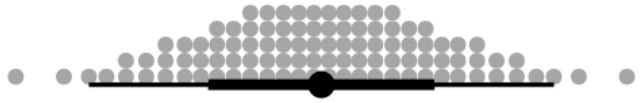

cyl
E[cyl | mpg]
mpg
`P(cyl | mpg)`
cyl
4
6
8


US unemployment over time
Uncertainty in what unemployment was
Uncertainty in what unemployment will be

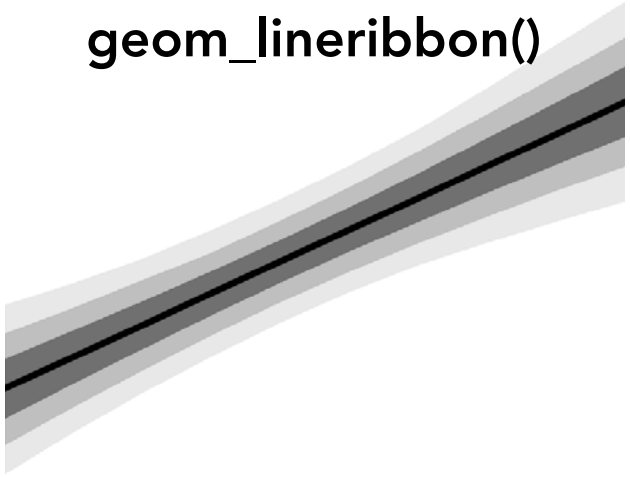

y
x

geom_slabinterval()

geom_dotsinterval()



geom_lineribbon()



Give ggdist a try!

And for examples from this talk, check out the vignettes at:
https://mjskay.github.io/ggdist/