The glue that binds statistical inference, tidy data, grammar of graphics, data visualisation and visual inference

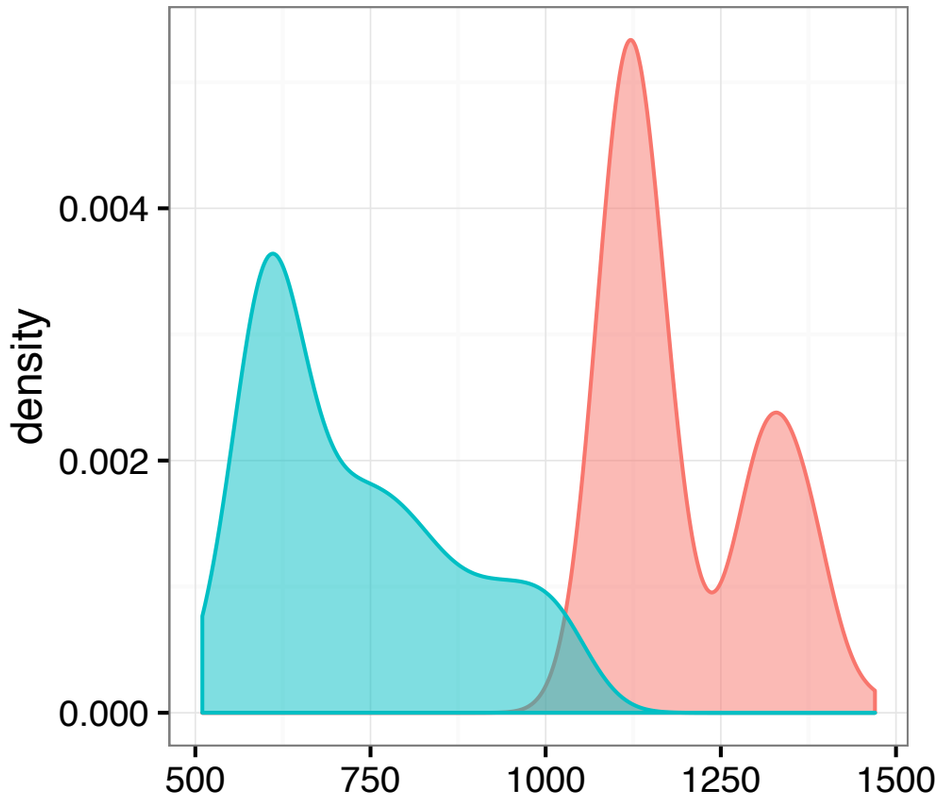
Di Cook
Monash University

If you have only this…

$\bar{X}_1 = -13.524$ $\bar{X}_2 = 24.166$

what can you say?

What about this?



What is a statistic?

- A statistic is a function on the values of items in a sample, e.g. for n iid random variables

$$\bar{X} = \sum_{i=1}^{n} X_i \qquad S_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- We study the behaviour of the statistic over all possible samples of size n.

Grammar of graphics

- The grammar of graphics is the mapping of (random) variables to graphical elements.
- Wilkinson (1999)
- Wickham (2008)'s ggplot2
- Enables comparison of different types of plots
- Enables graphics to be statistics

```
head(messy_data)
```

```
## # A tibble: 6 × 22
##    iso2  year  m_04 m_514 m_014 m_1524 m_2534 m_3544 m_4554 m_5564  m_65
##   <chr> <int> <int> <int> <int>  <int>  <int>  <int>  <int>  <int> <int>
## 1    AD  1989    NA    NA    NA     NA     NA     NA     NA     NA    NA
## 2    AD  1990    NA    NA    NA     NA     NA     NA     NA     NA    NA
## 3    AD  1991    NA    NA    NA     NA     NA     NA     NA     NA    NA
## 4    AD  1992    NA    NA    NA     NA     NA     NA     NA     NA    NA
## 5    AD  1993    NA    NA    NA     NA     NA     NA     NA     NA    NA
## 6    AD  1994    NA    NA    NA     NA     NA     NA     NA     NA    NA
## # ... with 11 more variables: m_u <int>, f_04 <int>, f_514 <int>,
## #   f_014 <int>, f_1524 <int>, f_2534 <int>, f_3544 <int>, f_4554 <int>,
## #   f_5564 <int>, f_65 <int>, f_u <int>
```

X_1 X_2 X_3 X_4 X_5

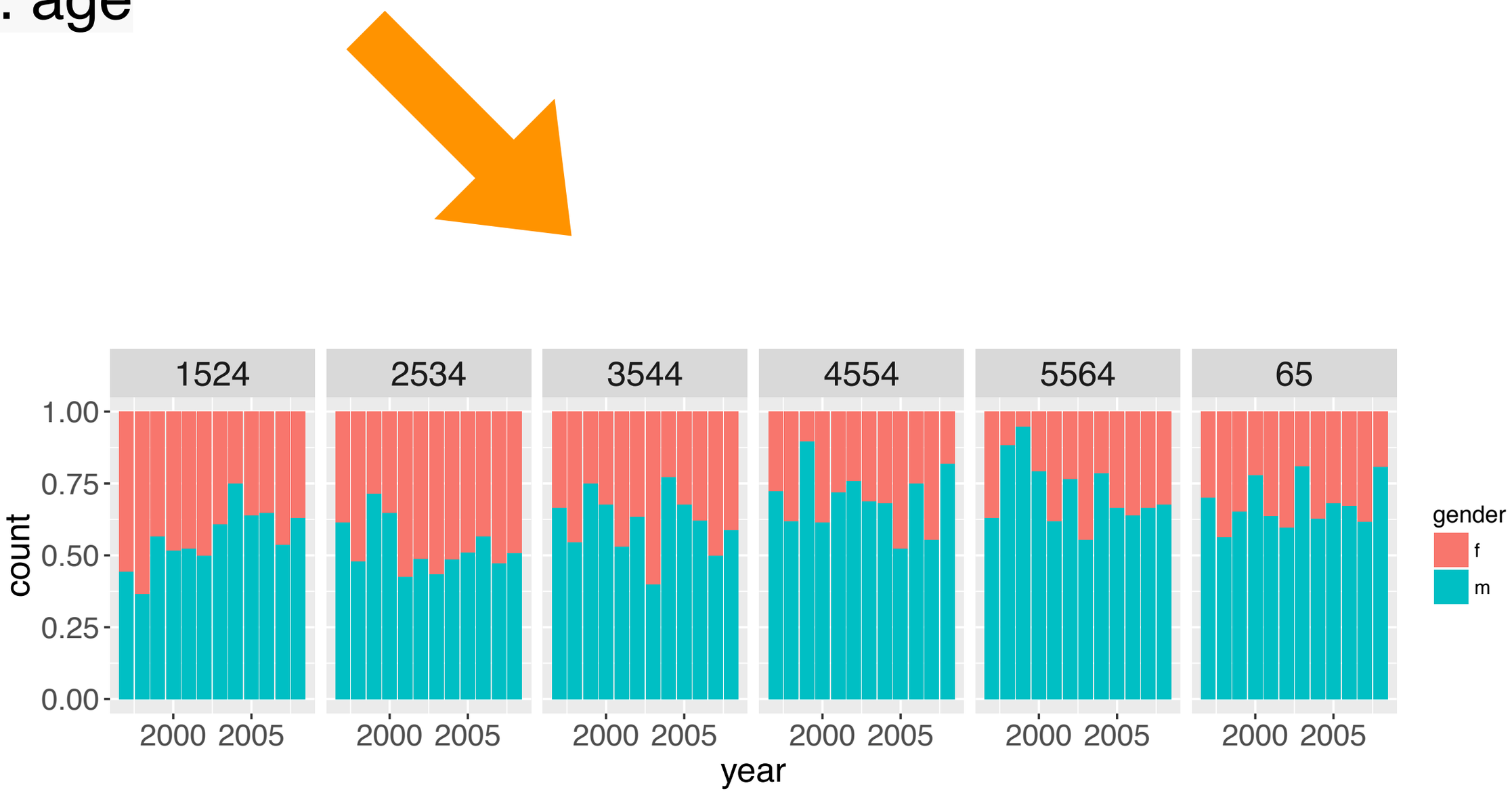
```
## # A tibble: 6 × 5
##    iso2  year gender   age count
##   <chr> <int>  <chr> <chr> <int>
## 1    AD  1996      m  1524     0
## 2    AD  1997      m  1524     0
## 3    AD  1998      m  1524     0
## 4    AD  1999      m  1524     0
## 5    AD  2000      m  1524     0
## 6    AD  2002      m  1524     0
```

```
## # A tibble: 6 × 5
##    iso2  year gender   age count
##   <chr> <int>  <chr> <chr> <int>
## 1    AD  1996      m  1524     0
## 2    AD  1997      m  1524     0
## 3    AD  1998      m  1524     0
## 4    AD  1999      m  1524     0
## 5    AD  2000      m  1524     0
## 6    AD  2002      m  1524     0
```

```
data: tidy_data
layer:
    mapping: x = year,
          y = count, fill = gender
    geom: fill-bar
facet: age
```

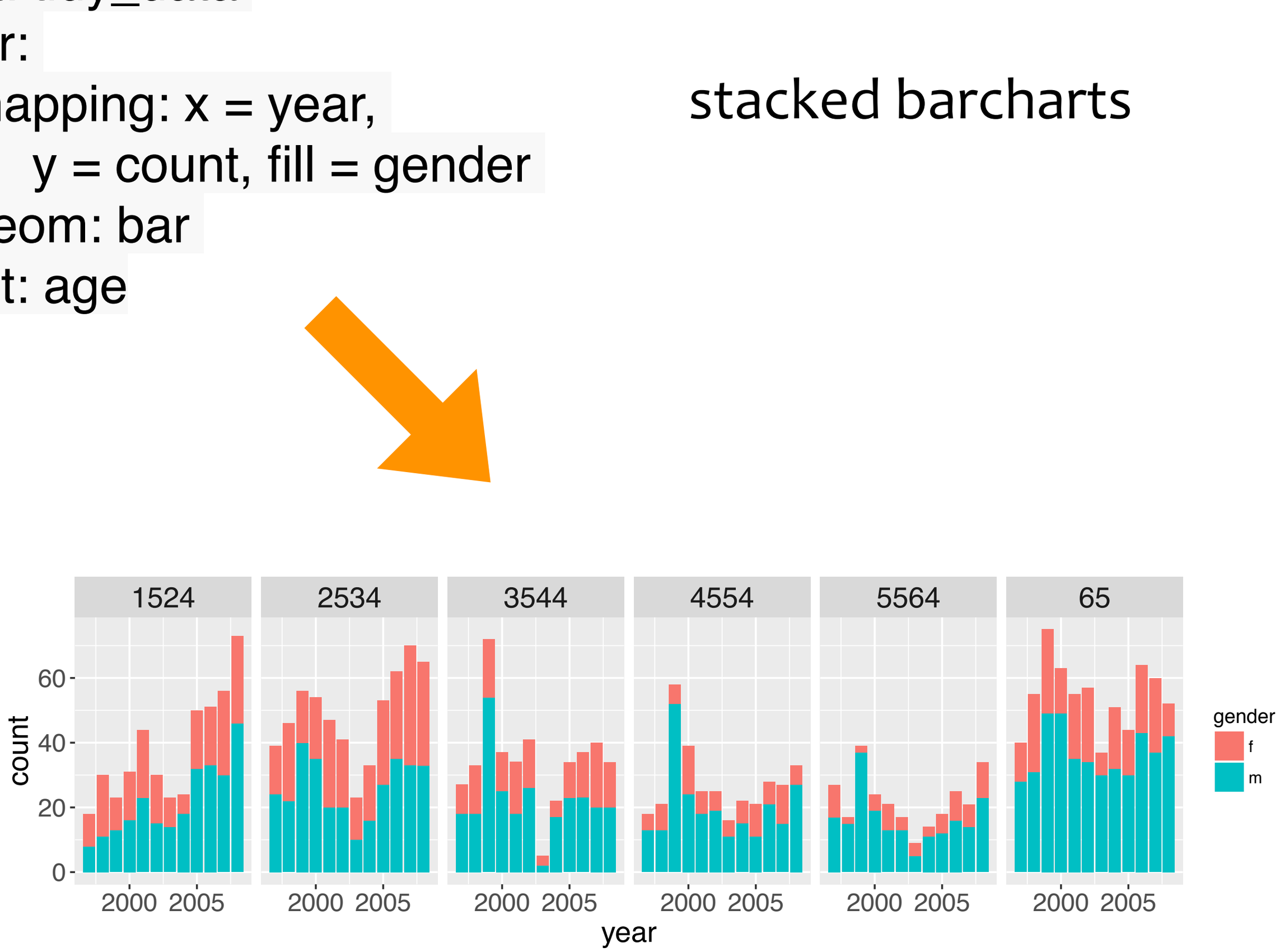
```
data: tidy_data
layer:
    mapping: x = year,
        y = count, fill = gender
    geom: fill-bar
facet: age
```

100% charts



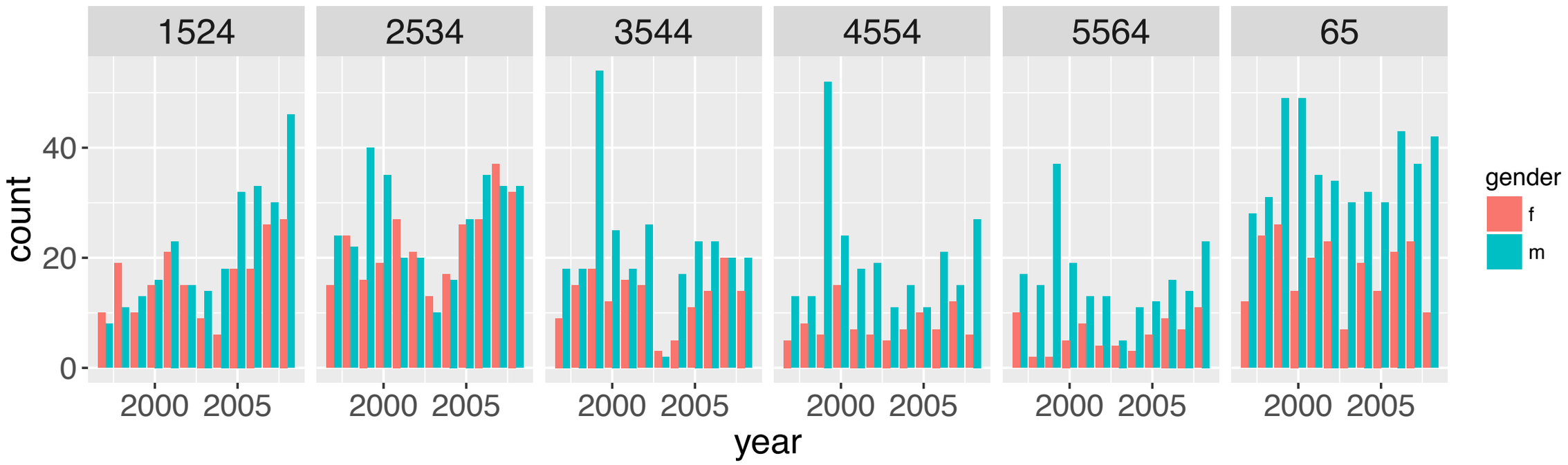
```
data: tidy_data
layer:
    mapping: x = year,
        y = count, fill = gender
    geom: bar
facet: age
```

stacked barcharts



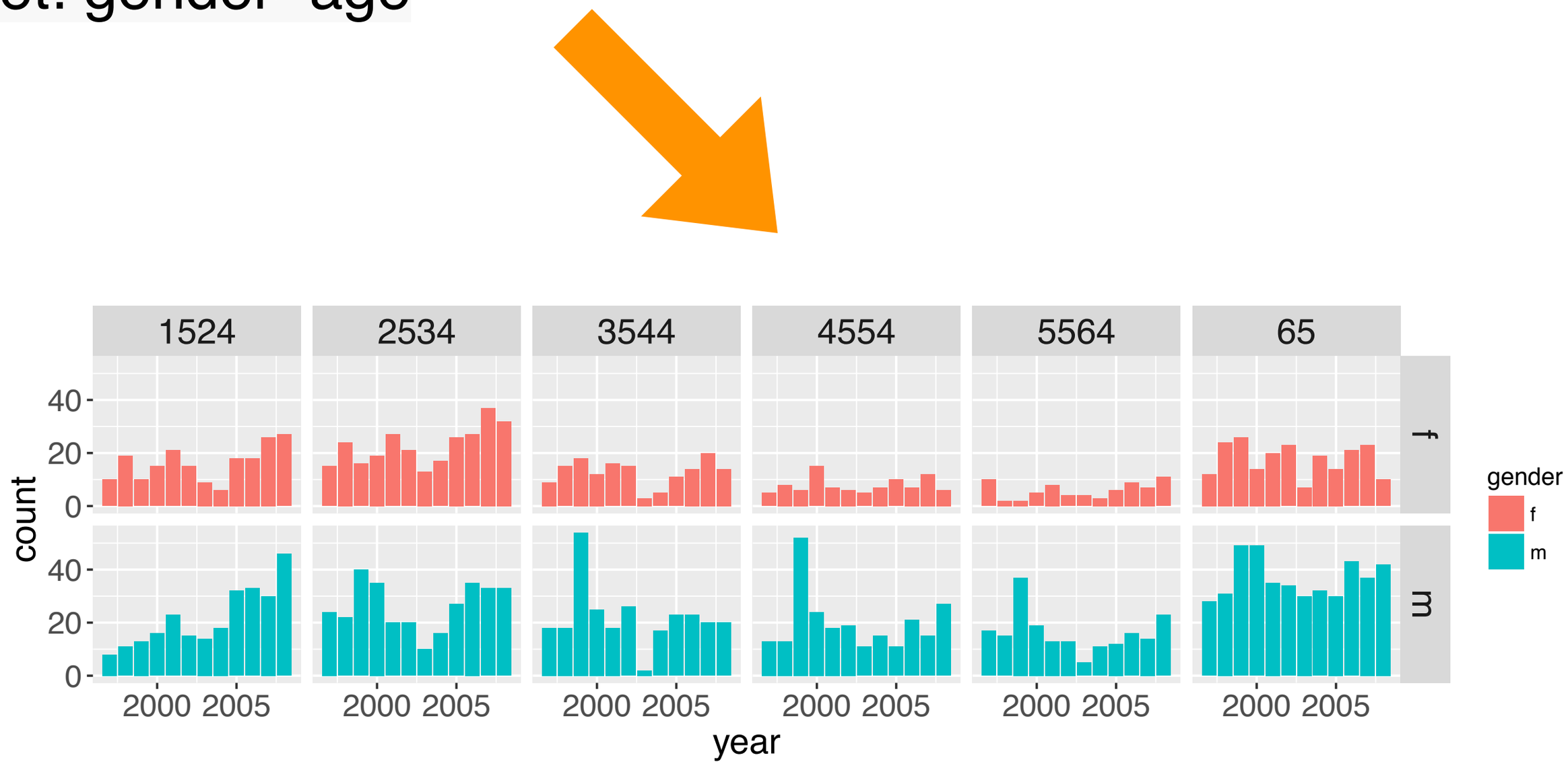
```
data: tidy_data
layer:
    mapping: x = year,
        y = count, fill = gender
    geom: dodge-bar
facet: age
```

side-by-side bar charts



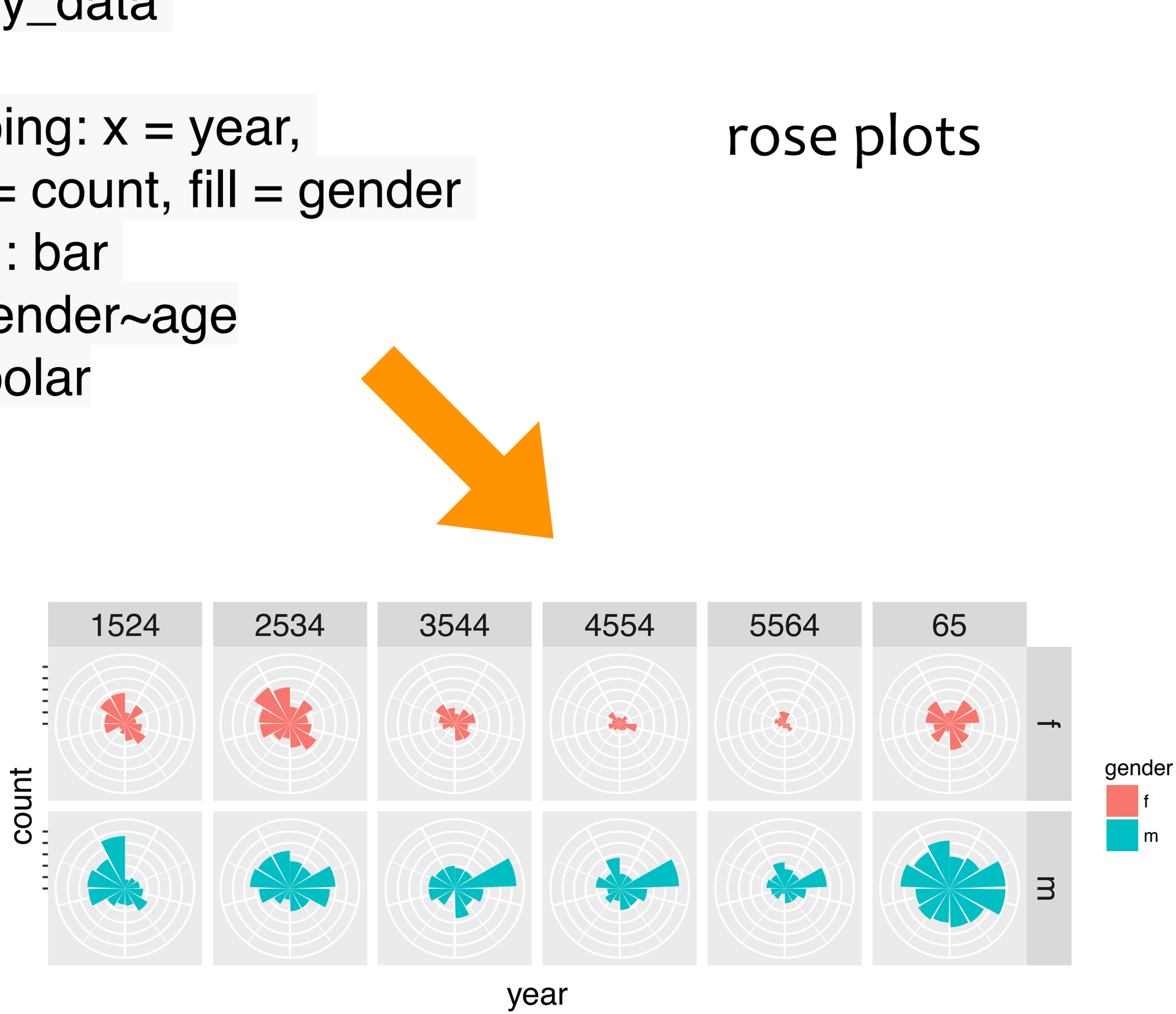
```
data: tidy_data
layer:
    mapping: x = year,
         y = count, fill = gender
    geom: bar
facet: gender~age
```

bar charts



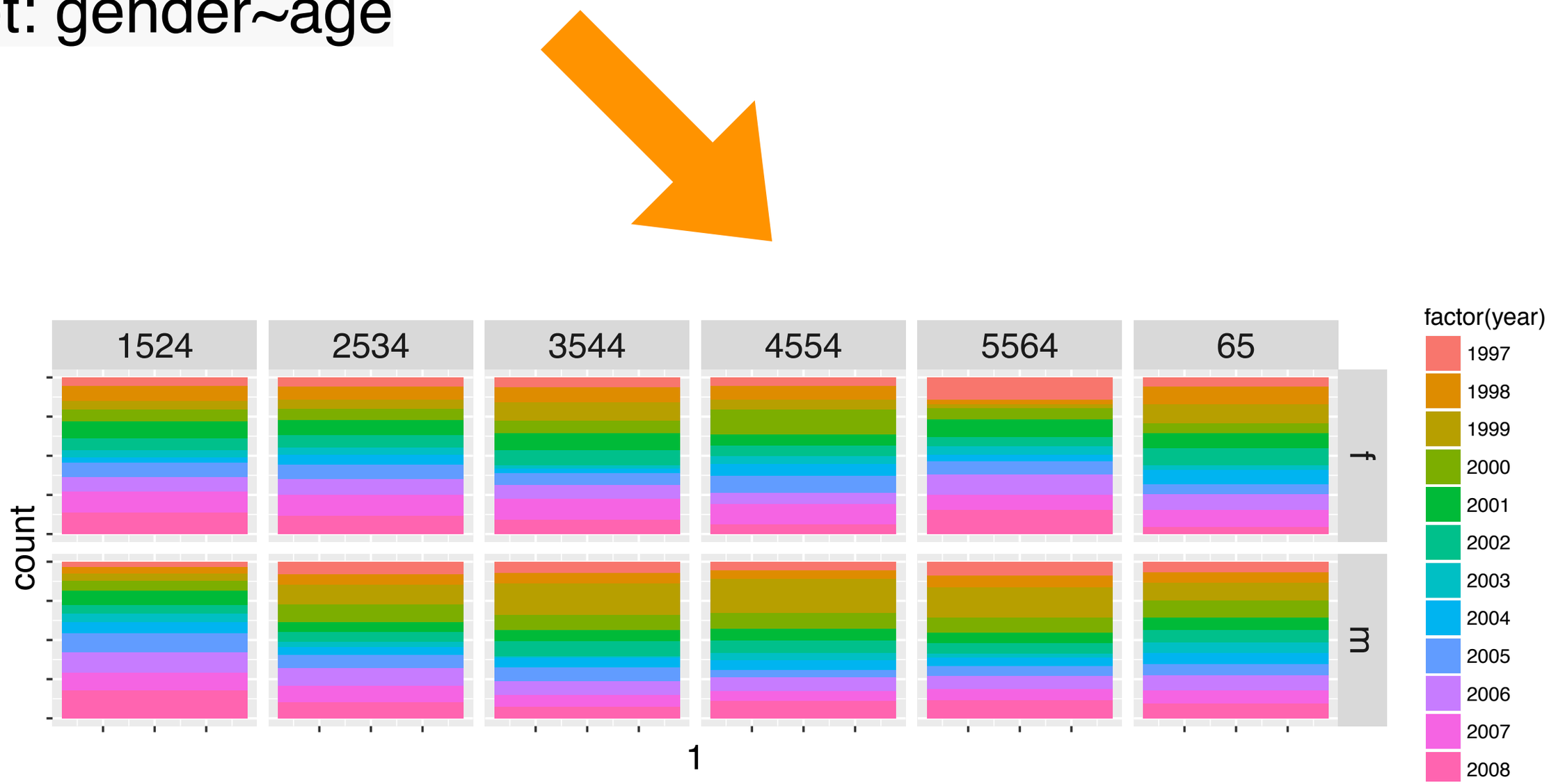
```
data: tidy_data
layer:
    mapping: x = year,
         y = count, fill = gender
    geom: bar
facet: gender~age
coord: polar
```

rose plots



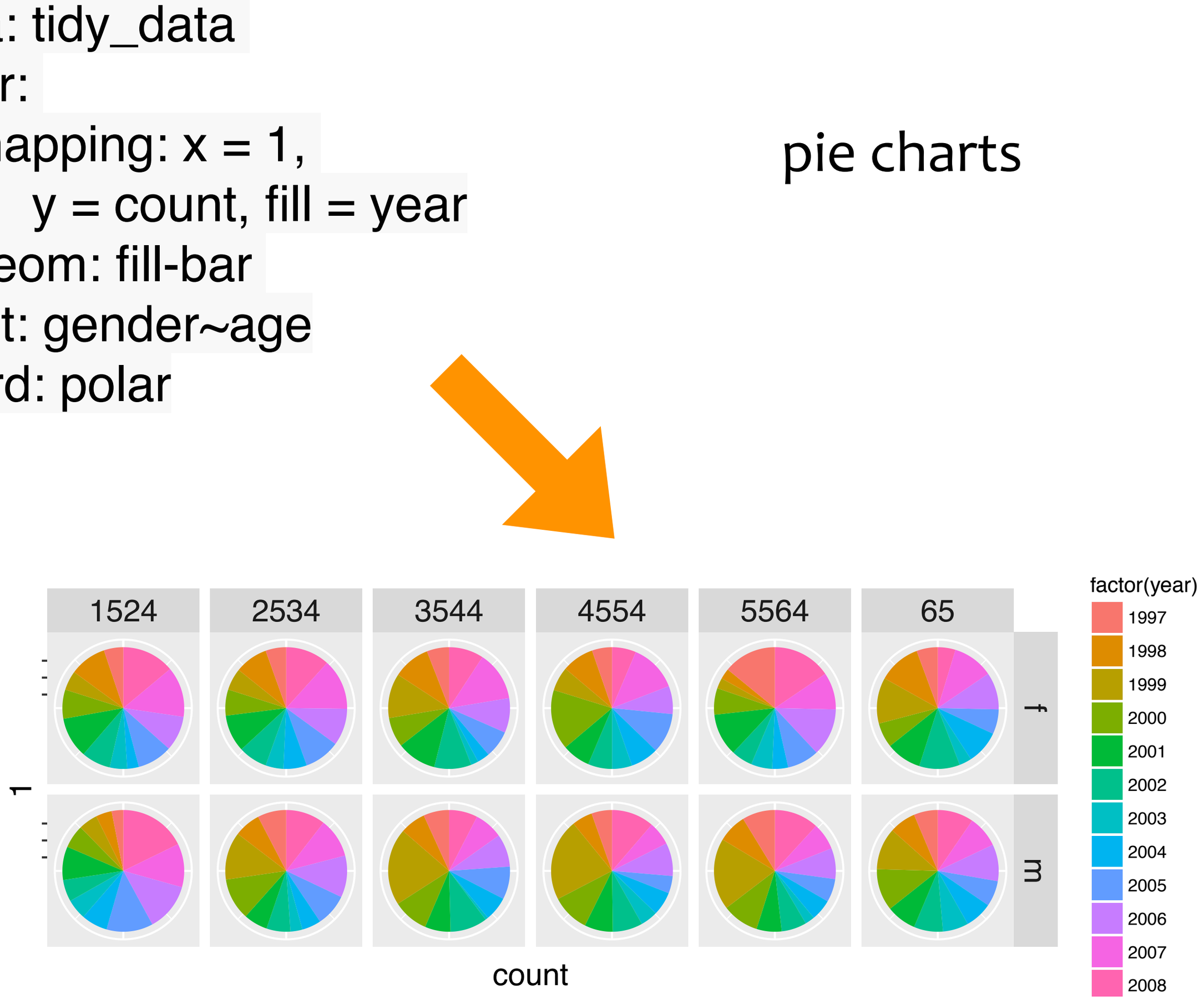
```
data: tidy_data
layer:
    mapping: x = 1,
        y = count, fill = year
    geom: fill-bar
facet: gender~age
```

100% charts



```
data: tidy_data
layer:
    mapping: x = 1,
         y = count, fill = year
    geom: fill-bar
facet: gender~age
coord: polar
```

pie charts



http://knowyourmeme.com/memes/morphing/photos/sort/favorites

Hypothesis testing

two hypotheses H_0 and H_1

prior probabilities $P(H_0) = p_0$ and $P(H_1) = p_1$

random variable (or the random vector) Y

distribution of Y $f_Y(y|H_0)$, and $f_Y(y|H_1)$

posterior probabilities of H_0 and H_1

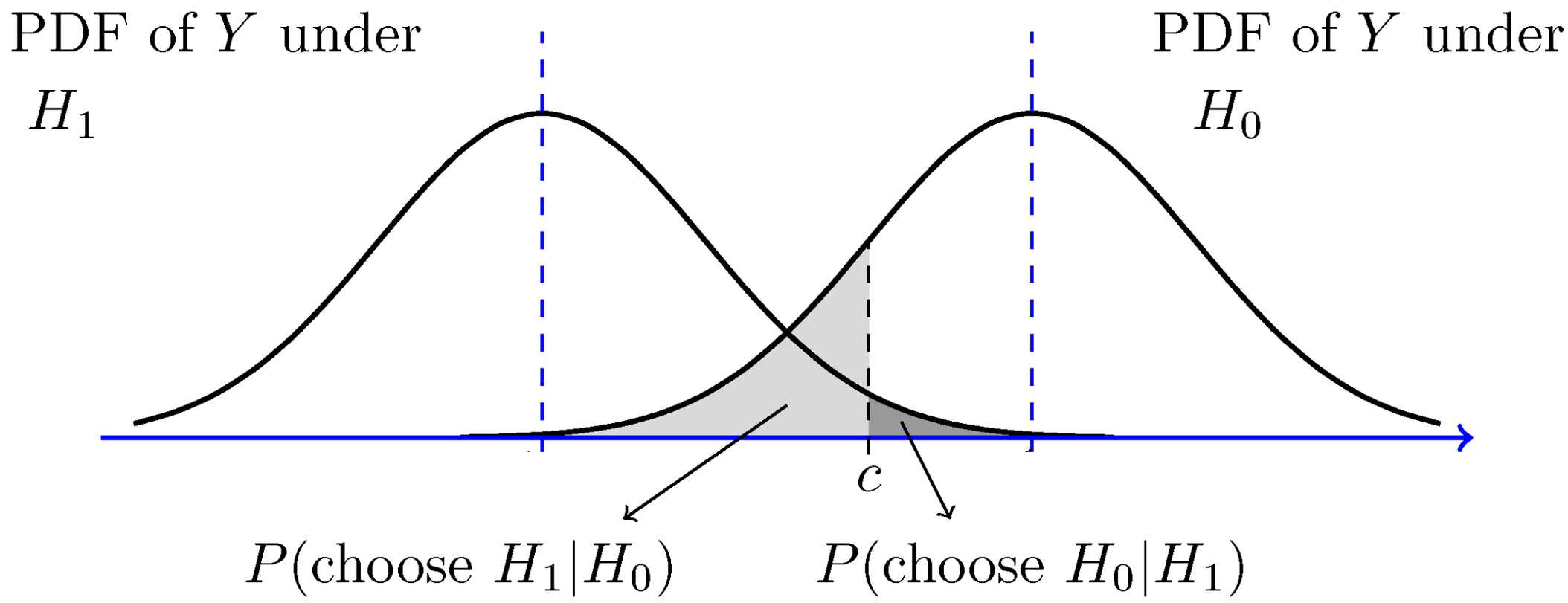
$$P(H_0|Y = y) = \frac{f_Y(y|H_0)P(H_0)}{f_Y(y)},$$

$$P(H_1|Y = y) = \frac{f_Y(y|H_1)P(H_1)}{f_Y(y)}.$$

we choose H_0 if and only if $P(H_0|Y = y) \geq P(H_1|Y = y)$

https://www.probabilitycourse.com/chapter9/9_1_8_bayesian_hypothesis_testing.php

Errors

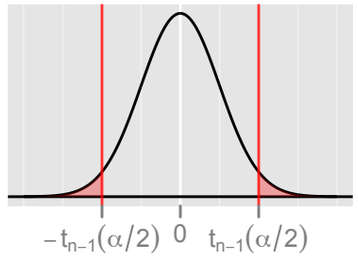
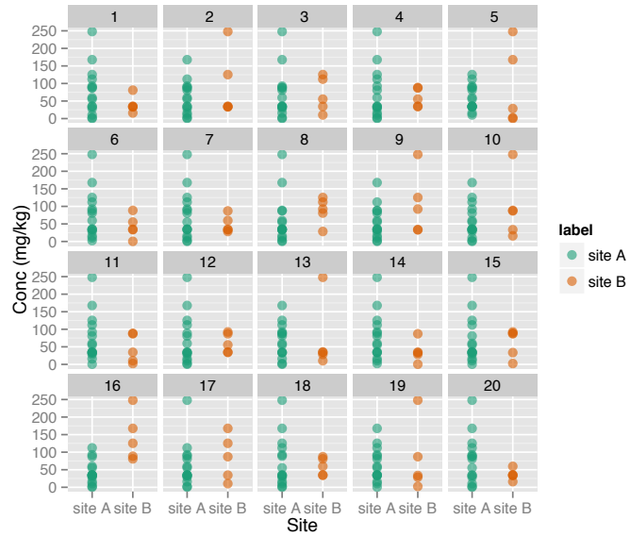


Type I
alpha

Type II
beta

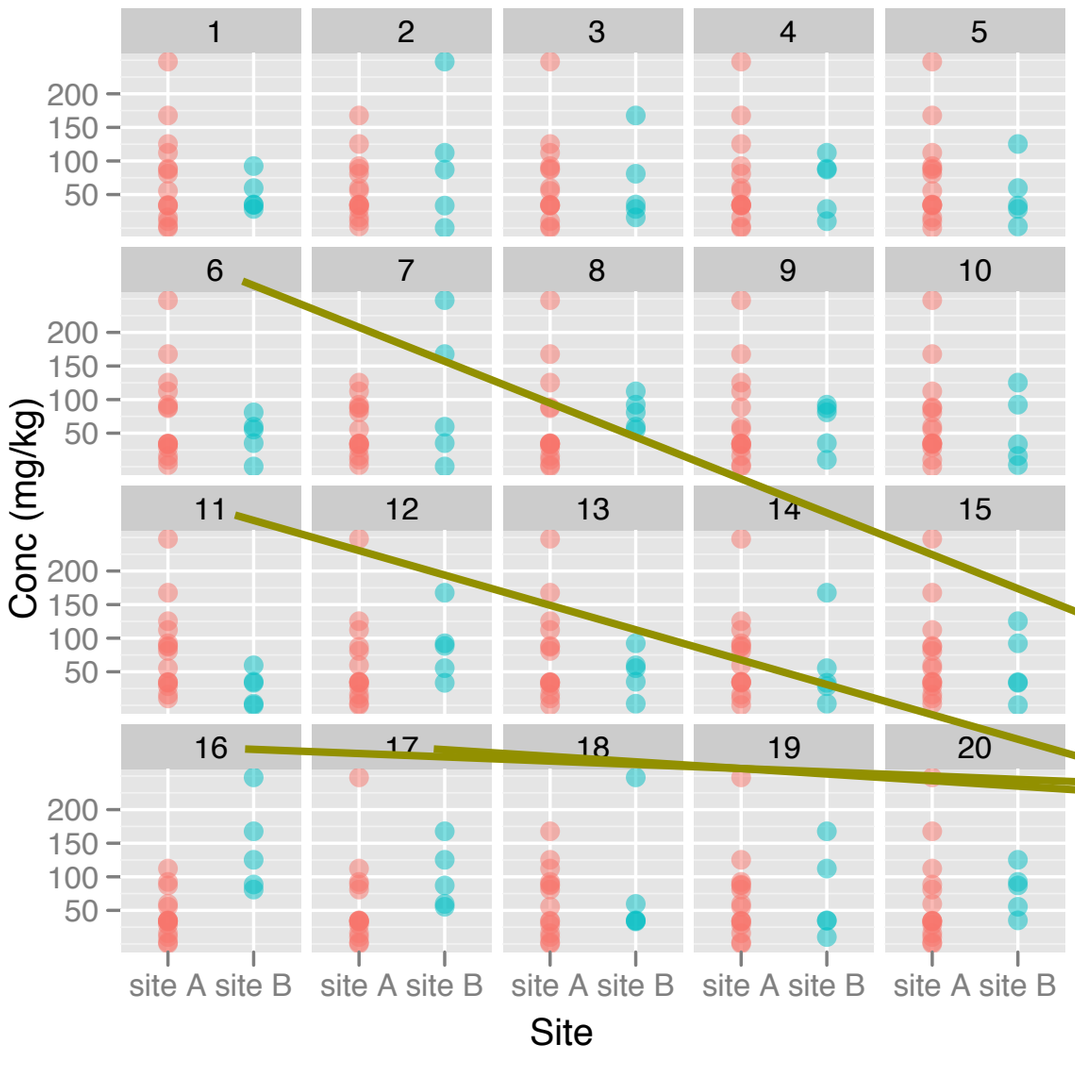
Power = $P(\text{choose } H_1 | H_1)$ = 1-beta

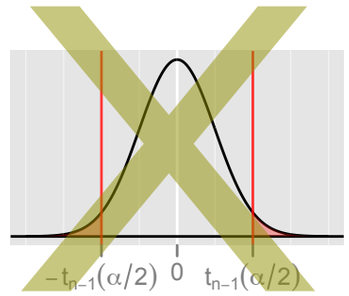
Visual inference

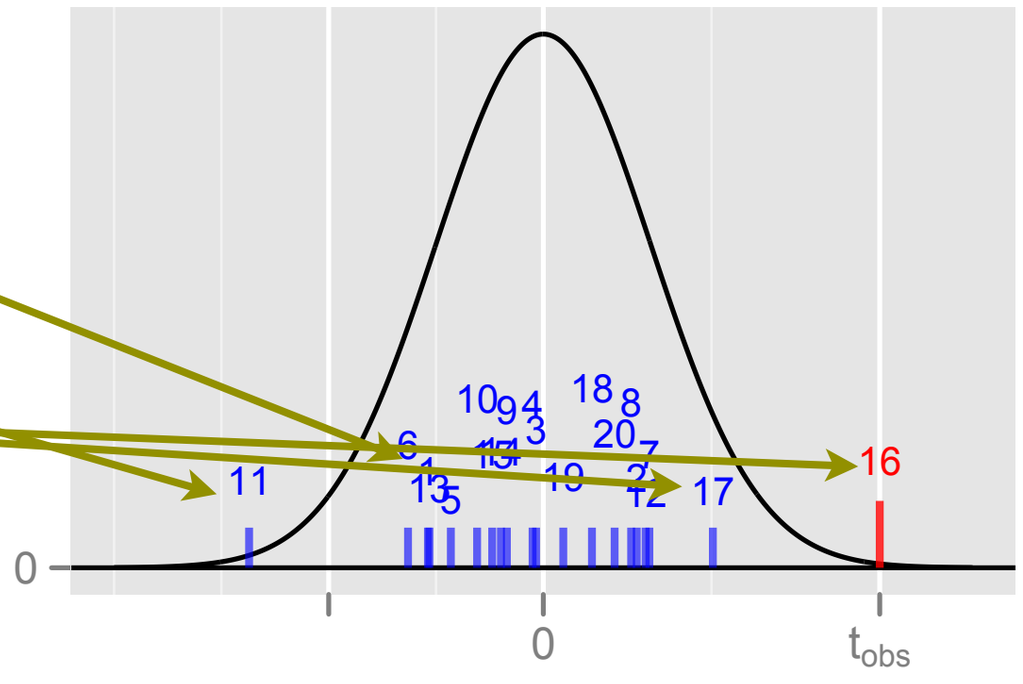
	Mathematical Inference	Visual Inference
Hypothesis	$H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_o : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	↓	↓
Test Statistic	$T(y) = \frac{\overline{y}_1 - \overline{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$V(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t);$ 	$f_{V(y)}(t);$ 
	↓	↓
Reject H_o if	observed T is extreme	observed data plot is identifiable

Null distribution

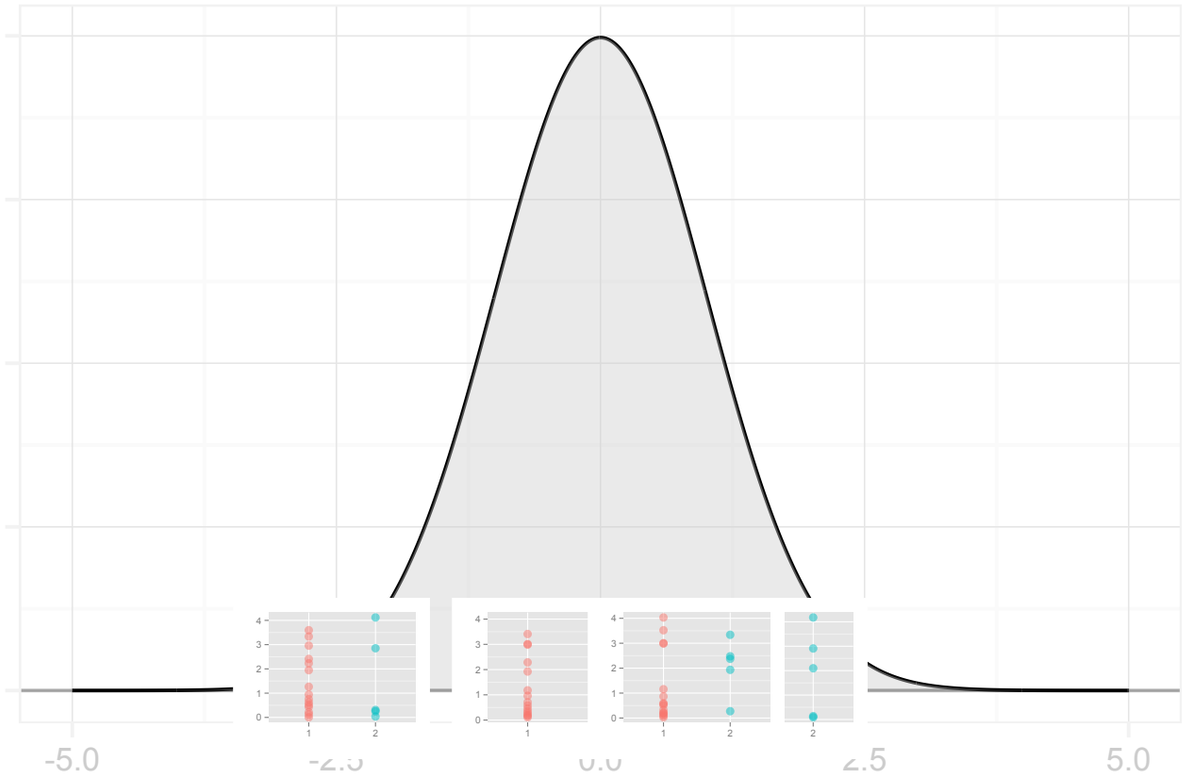
Sampling distribution comparison is against a finite set





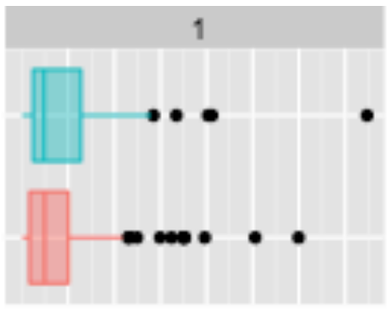


All we have is (m-1) representatives from whatever that distribution is.



Inference for graphics

- Choice of plot implicitly sets H_0, H_1
- Generically, we are thinking H_0: no pattern, H_1: pattern, but the choice of plot makes this much more explicit



What is the question?

Is there a difference between the two groups

Ho: no difference, H1: difference

What is the data?

Two variables: V1, V2; V1 is categorical

What is the mapping?

x=V2, y=V1, colour=V1

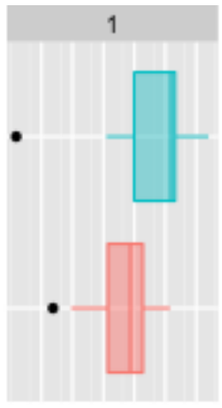
geom=boxplot

What is a null generating mechanism?

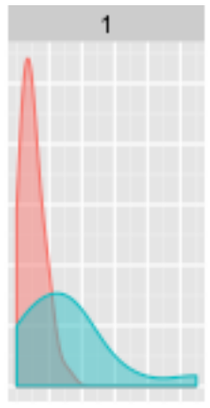
permute the values of V1, relative to V2

boxplot_3_135_135_12_13

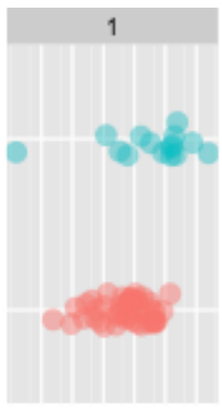
Which is the best design?



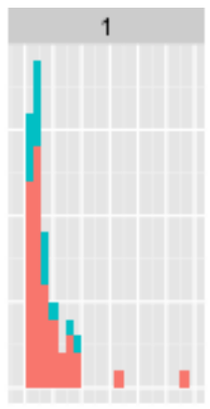
mapping:
x = V_2
y = V_1
geom: boxplot



mapping:
x = V_1
color = V_2
geom: density

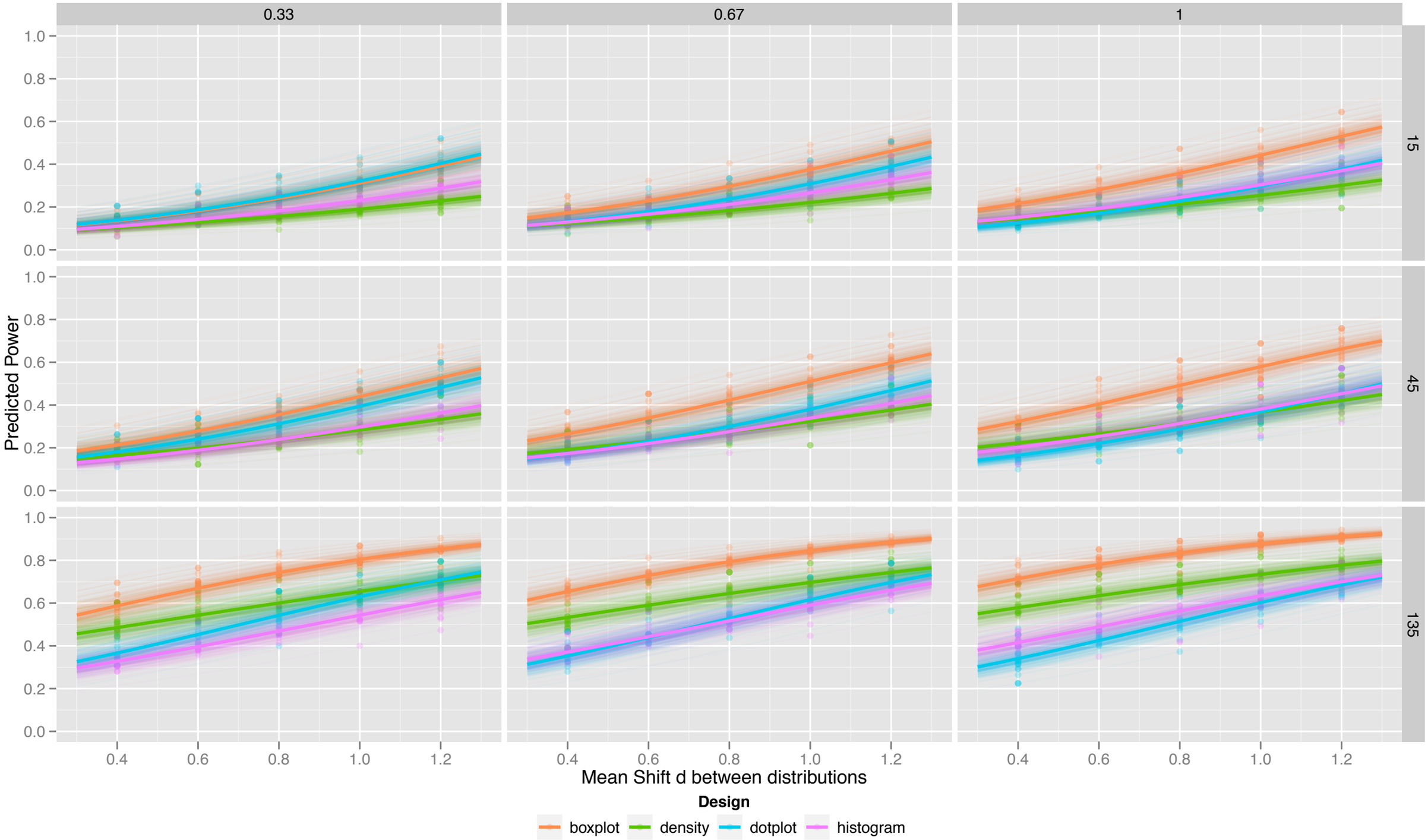


mapping:
x = V_2
y = V_1
geom: dotplot



mapping:
x = V_1
color = V_2
geom: histogram

Compute power for each design.
The design which allows reader to detect the difference more frequently is the most powerful statistic.


0.33
0.67
1
15
45
135
Predicted Power
1.0
0.8
0.6
0.4
0.2
0.0
0.4
0.6
0.8
1.0
1.2
Mean Shift d between distributions
Design
boxplot
density
dotplot
histogram

Rejecting Ho and power of a plot

- Assume the data is not different from the null data. Compute the probability of x/k observers or more selecting the data plot. If this is small, reject Ho.
- Power of the statistic is the proportion of people who select the data plot from the rest. If the data is really different, how well can people detect it.

Your turn

- With your neighbour(s) pick one of the plots that you have brought
- Determine

1. what the null hypothesis and alternative is,
2. what the underlying data is,
3. how variables are mapped to graphical elements, and
4. what a null generating mechanism could be
5. What are possible alternative designs?

The newbie probabilist:

What do you mean the probability of getting heads on the coin toss is 0.5? I just tossed it and I got a head. The probability was 1!