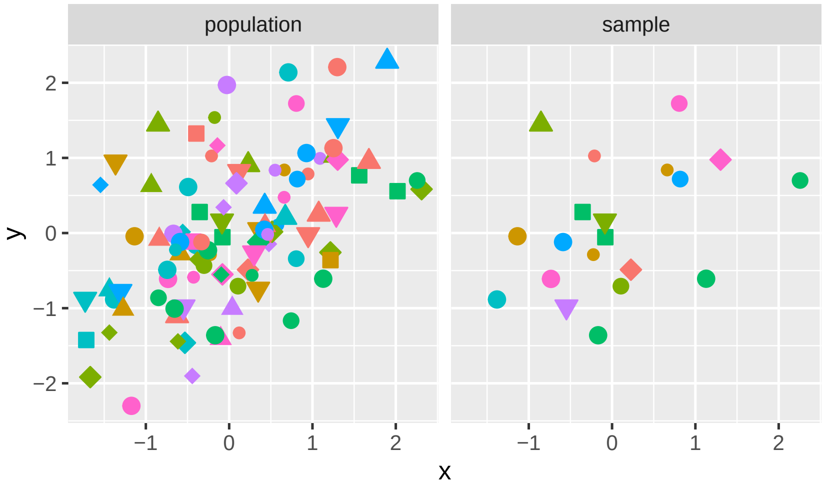
Bootstrapping basics

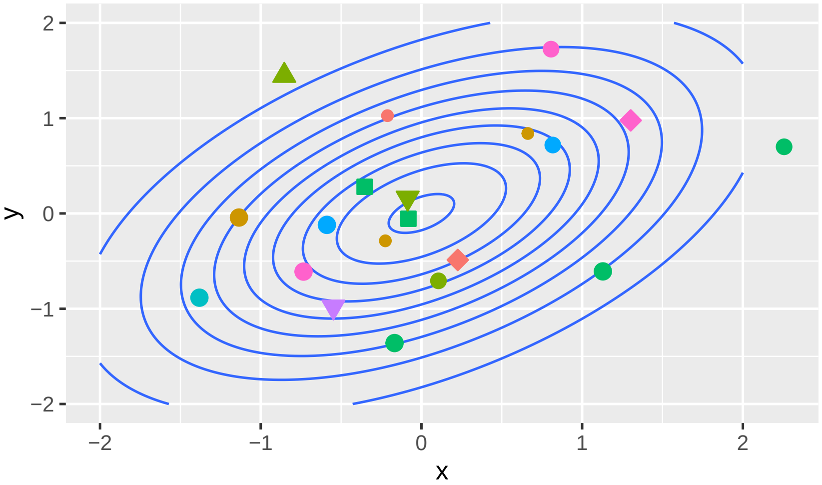
Michael Love

2023-06-01

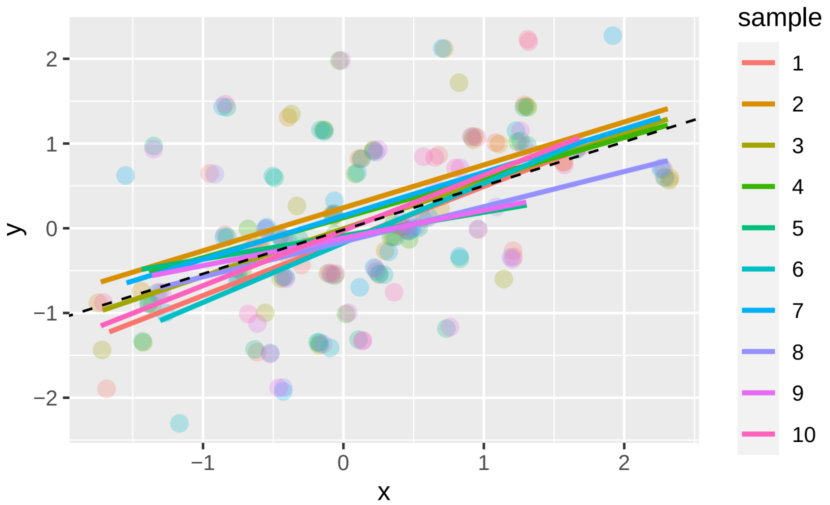
sampling variance



or when there is not a “population”

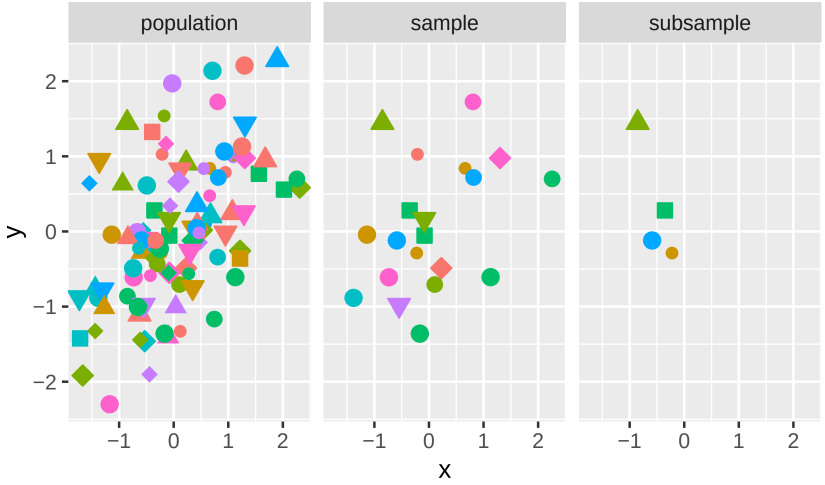


estimating a parameter



How can we assess the effect of sampling variance on parameter estimates?

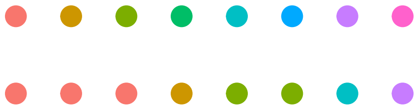
idea: sub-sampling a sample



what is good about the sub-sample? what is a problem?

idea of bootstrap

- instead of sub-sample, take a sample of the same size
- sample each observation, then put it back ("with replacement")



idea of bootstrap

Real World	Bootstrap World
$P \to X_n$	$\widehat{P}_n \to X_n^*$
$\hat{\theta}_n = f(X_n)$	$\hat{\theta}_n^* = f(X_n^*)$

from "Introduction to the Bootstrap" Efron & Tibshirani (1993)

bootstrap often used for the variance of an estimator

From Yen-Chi Chen (UW) notes:

- Sample $X_n^{*(1)}, X_n^{*(2)}, \dots, X_n^{*(B)}$
- Obtain $\widehat{\theta}_n^{*(1)}, \widehat{\theta}_n^{*(2)}, \dots, \widehat{\theta}_n^{*(B)}$
- Sample variance of these bootstrap estimates $= \widehat{\mathrm{Var}}_B(\hat{\theta}_n^*)$

Want:

$$\widehat{\mathrm{Var}}_B(\hat{\theta}_n^*) \approx \mathrm{Var}(\hat{\theta}_n)$$

consistency of the bootstrap variance

From Yen-Chi Chen (UW) notes:

Want: $\widehat{\mathrm{Var}}_B(\hat{\theta}_n^*) \approx \mathrm{Var}(\hat{\theta}_n)$

For large B, we have:

$$\widehat{\mathrm{Var}}_B(\hat{\theta}_n^*) \approx \mathrm{Var}(\hat{\theta}_n^* | \widehat{P}_n)$$

Need to show:

$$\mathrm{Var}(\hat{\theta}_n^* | \widehat{P}_n) \approx \mathrm{Var}(\hat{\theta}_n)$$

Sketch: for a given estimator, need to show that the variance of the functional of the empirical density $\widehat{P}_n$ converges in probability to the variance of the functional of the original density P.

three types of bootstrapping

Consider regression:

$$Y = X\beta + \varepsilon \quad (1)$$
$$\varepsilon \sim N(0, \sigma^2) \quad (2)$$

- estimate $\hat{\sigma}^2$
  - simulate new errors $\varepsilon^* \sim N(0, \hat{\sigma}^2)$
  - simulate new data via $X\hat{\beta} + \varepsilon^*$
- resample residuals $\hat{\varepsilon}$ with replacement
  - simulate new data via $X\hat{\beta} + \hat{\varepsilon}^*$
- resample cases entirely

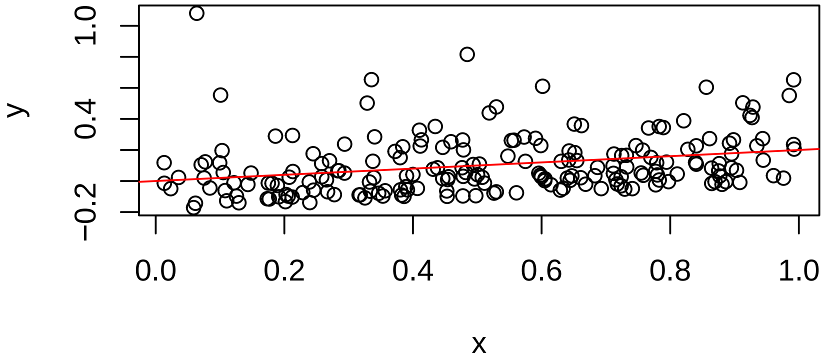
what do we assume in these three types?

example: line with non-normal errors

```r
set.seed(1)
n <- 200
x <- runif(n)
eps <- rexp(n, 5)
eps <- eps - mean(eps)
slope <- .2
y <- slope * x + eps
dat <- data.frame(x,y)
```

example: line with non-normal errors

```
plot(x,y)
abline(0, slope, col="red")
```



simple bootstrapping

```r
library(dplyr)
library(broom)
set.seed(5)
boots <- replicate(1000, {
  idx <- sample(n, replace=TRUE)
  coef(lm(y ~ x, data=dat[idx,]))[2]
})
sd(boots)
```

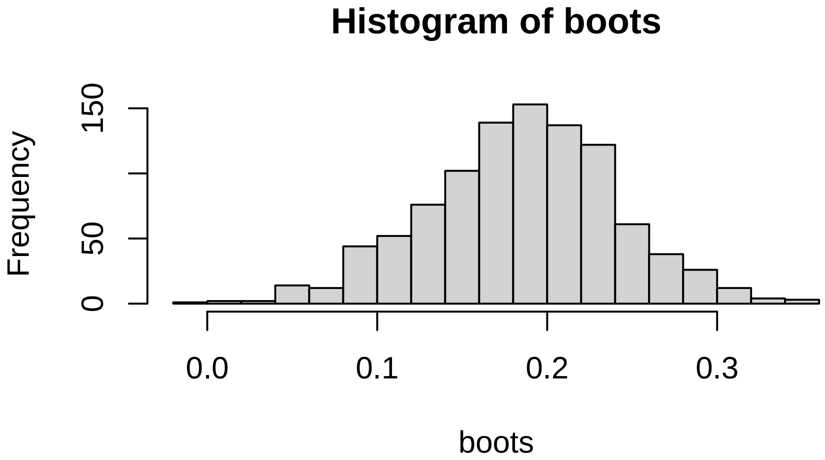
```
[1] 0.05373745
```

```r
fit <- lm(y ~ x, data=dat)
fit %>% tidy() %>%
  filter(term=="x") %>% pull(std.error)
```

```
[1] 0.04845154
```

simple bootstrapping

```
hist(boots, breaks=20)
```



using boot

provides some extra bells and whistles re: stratified data

```
library(boot)
get_slope <- function(data, idx) {
  coef(lm(y ~ x, data=data[idx,]))[2]
}
set.seed(5)
boots2 <- boot(dat, get_slope, R=1000)
```

using boot

```r
boots2
```

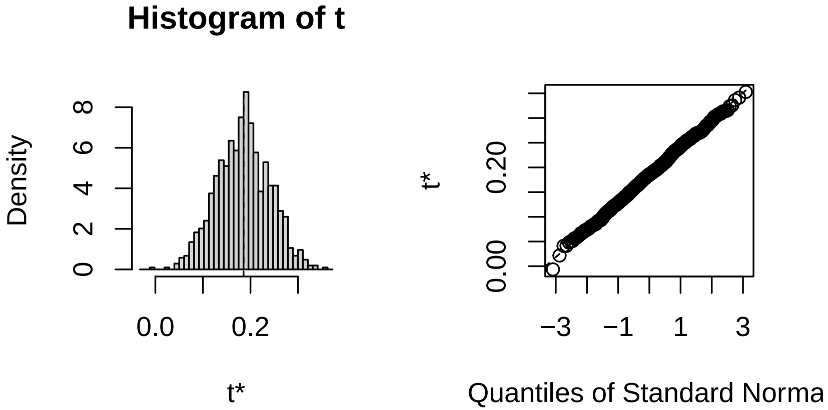
```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = dat, statistic = get_slope, R = 1000)


Bootstrap Statistics :
     original          bias    std. error
t1* 0.1854471 -0.0008217989  0.05544021
```

using boot

```
plot(boots2)
```



additional options from car

```r
library(car)
set.seed(5)
boots3 <- Boot(fit, method="residual")
```

additional options from car

```
boots3
```

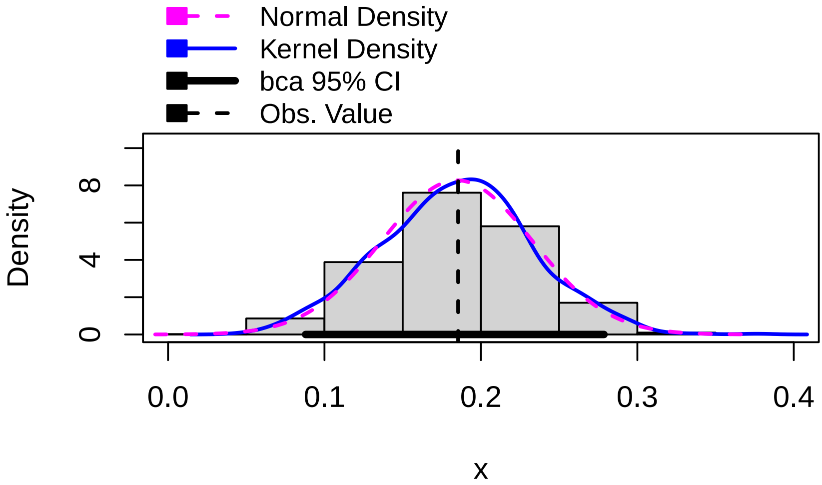
```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot::boot(data = dd, statistic = boot.f, R = R, .fn = f, p
    ncpus = ncores, cl = cl2)


Bootstrap Statistics :
       original          bias    std. error
t1* 0.007533451  0.0001025333  0.02810170
t2* 0.185447090 -0.0007108046  0.04823957
```

additional options from car

```
hist(boots3, parm="x")
```



additional options from car

bca = bias-corrected and accelerated, Efron and Tibshirani (1993)

considers:

- proportion of $\widehat{\theta}^*_n < \widehat{\theta}$
- skewness of the distribution of $\widehat{\theta}^*_n$

```
confint(boots3)
```

```
Bootstrap bca confidence intervals

                   2.5 %    97.5 %
(Intercept) -0.04416639 0.0648779
x            0.08790722 0.2787100
```

{infer} package

S, H, G, C = Specify, Hypothesize, Generate, Calculate

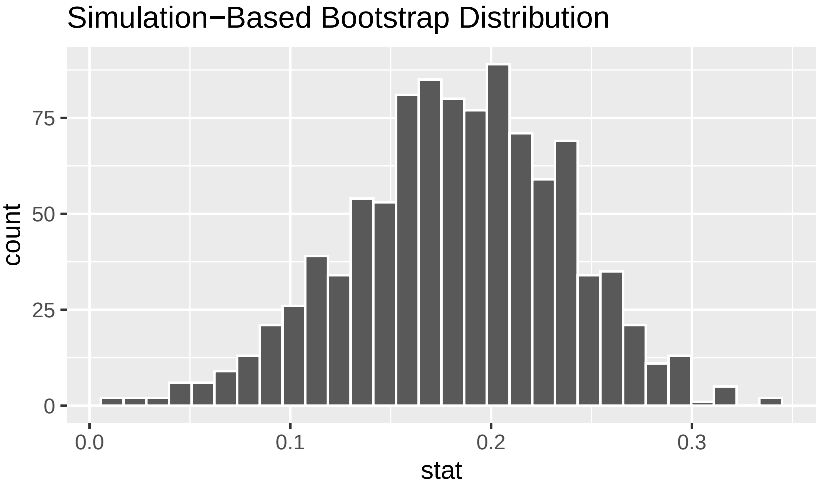
```
library(infer)
set.seed(5)
perm <- dat %>% specify(y ~ x) %>%
  hypothesize(null="independence") %>%
  generate(reps=1000, type="permute") %>%
  calculate(stat="slope")
```

bootstrapping a statistic

```r
library(infer)
set.seed(5)
boot <- dat %>% specify(y ~ x) %>%
  generate(reps=1000, type="bootstrap") %>%
  calculate(stat="slope")
```

bootstrapping a statistic

```
visualize(boot, bins=30)
```



confidence intervals

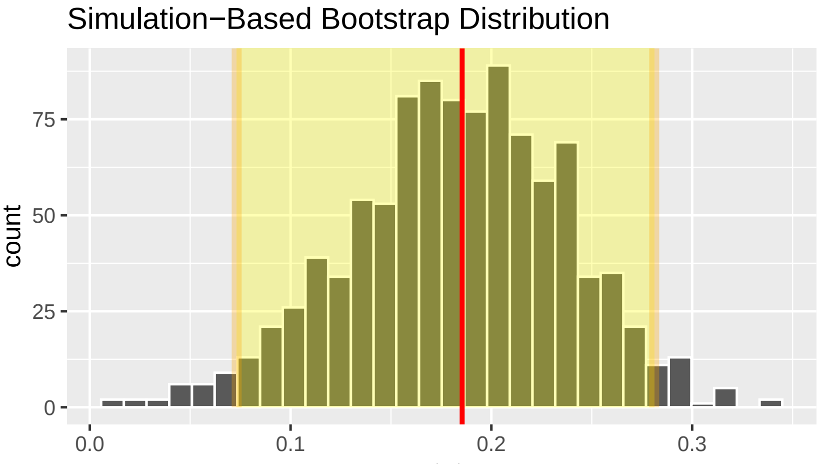
```
ci <- get_ci(boot)
ci
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1   0.0731    0.281
```

```
obs_beta <- dat %>%
  specify(y ~ x) %>%
  calculate(stat="slope")
```

visualize

```
visualize(boot, bins=30) +
  shade_confidence_interval(
    ci, alpha=.3, fill="yellow", color="orange") +
  geom_vline(xintercept=obs_beta$stat,
             color="red", linewidth=1)
```



going further

Bootstrapping Regression Models in R An Appendix to An R Companion to Applied Regression, 3rd ed. John Fox & Sanford Weisberg

(can find PDF online)