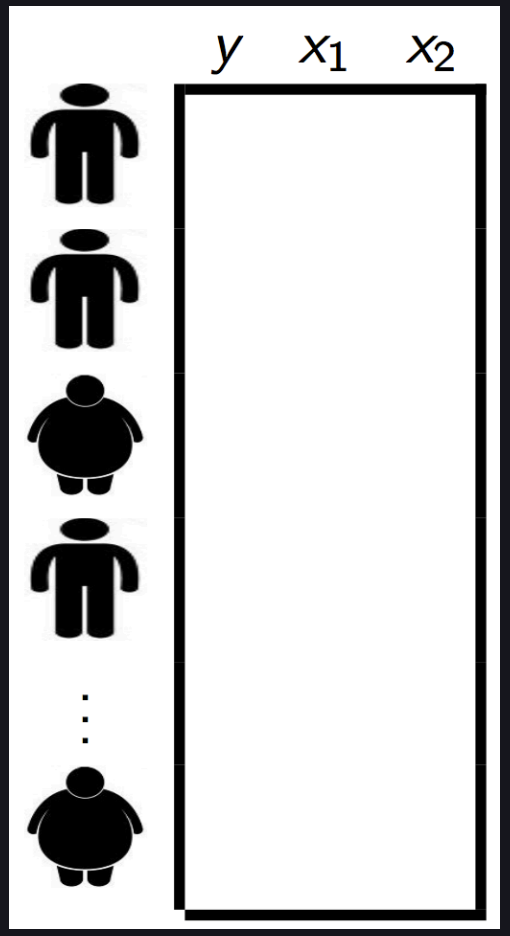
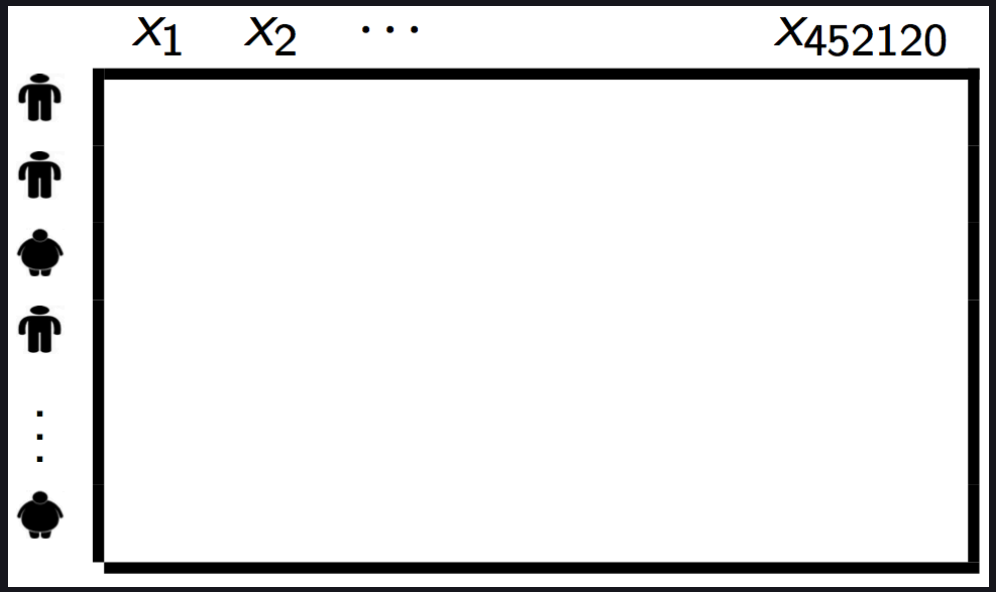
Multiple testing

Claus Thorn Ekstrøm
UCPH Biostatistics
ekstrom@sund.ku.dk

Data sizes. The $N \ll P$ problem

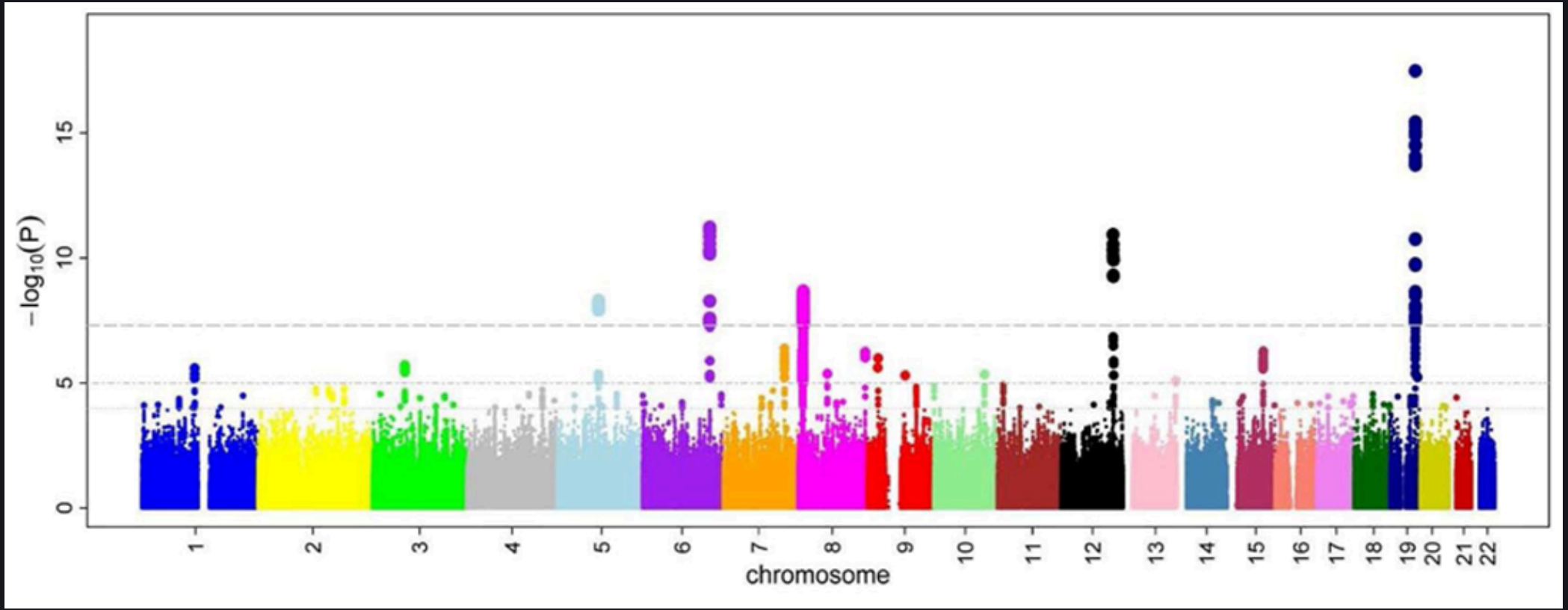




The "Big Data" revolution

1. "Big P small N" problem with many modern large-scale-datasets: registers, images, *-omics, ...
2. Need to reduce the dimension in some way
3. How do we evaluate significance when we have used the data for feature selection?
4. Multiple testing becomes an issue --- not just for high-dimensional data

Manhattan plot



Multiple comparison problems

Errors committed when testing a single null hypotheses, H_0

Analysis result	H0 true	H0 false
Reject	α	1-β
Don't reject	1-α	β

α is the significance level

$1 - \beta$ is the power

Multiple comparison problems

The family-wise error rate (FWER) is the probability of making at least one type I error (false positive).

For m tests we have

$$FWER = P(\cup(p_i \leq \alpha))) = 1 - P(\text{no false positives}) = 1 - (1-\alpha)^m \leq m\alpha$$

where the third equality only holds under independence, but the inequality holds due to Boole's inequality.

Multiple comparison problems

Number of errors committed when testing m null hypotheses.

Analysis result	H_o true	H_o false	Total
Reject	V	S	R
Don't reject	U	T	m-R
Total	m_0	$m - m_0$	m

Here R, the number of rejected hypotheses/discoveries. V, S, U and T are unobserved. The FWER is

$$FWER = P(V > 0) = 1 - P(V = 0)$$

Bonferroni correction

The most conservative method but is free of dependence and distributional assumptions.

$$FWER = 1 - P(V = 0) = 1 - (1 - \alpha)^m \leq m\alpha$$

So set the significance level for each individual test at α/m.

In other words we reject the ith hypothesis if

$$mp_i \leq \alpha \Leftrightarrow p_i \leq \frac{\alpha}{m}$$

Sidak correction

$$1 - (1 - \alpha)^m = \alpha^* \Leftrightarrow \alpha = 1 - \sqrt[m]{1 - \alpha^*}$$

Slightly less conservative than Bonferroni (but not much). Requires independence!

Holm correction

1. Compute and order the individual p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Find $\hat{k} = \min\{k : p_{(k)} > \frac{\alpha}{m+1-k}\}$
3. If $\hat{k}$ exists then reject hypotheses corresponding to
   $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(\hat{k}-1)}$

Holm correction

Controls the FWER: Assume the (ordered) k is the first wrongly rejected true hypothesis. Then $k \leq m - (m_0 - 1)$.

Hypothesis k was rejected so

$$p_{(k)} \leq \frac{\alpha}{m + 1 - k} \leq \frac{\alpha}{m + 1 - (m - (m_0 - 1))} \leq \frac{\alpha}{m_0}$$

Since there are m_0 true hypotheses then (Bonferroni argument) the probability that one of them is significant is at most α so FWER is controlled.

Practical problems

- While guarantee of FWER-control is appealing, the resulting thresholds often suffer from low power.

  In practice, this tends to wipe out evidence of the most interesting effects

- FDR control offers a way to increase power while maintaining some principled bound on error

False discovery rate

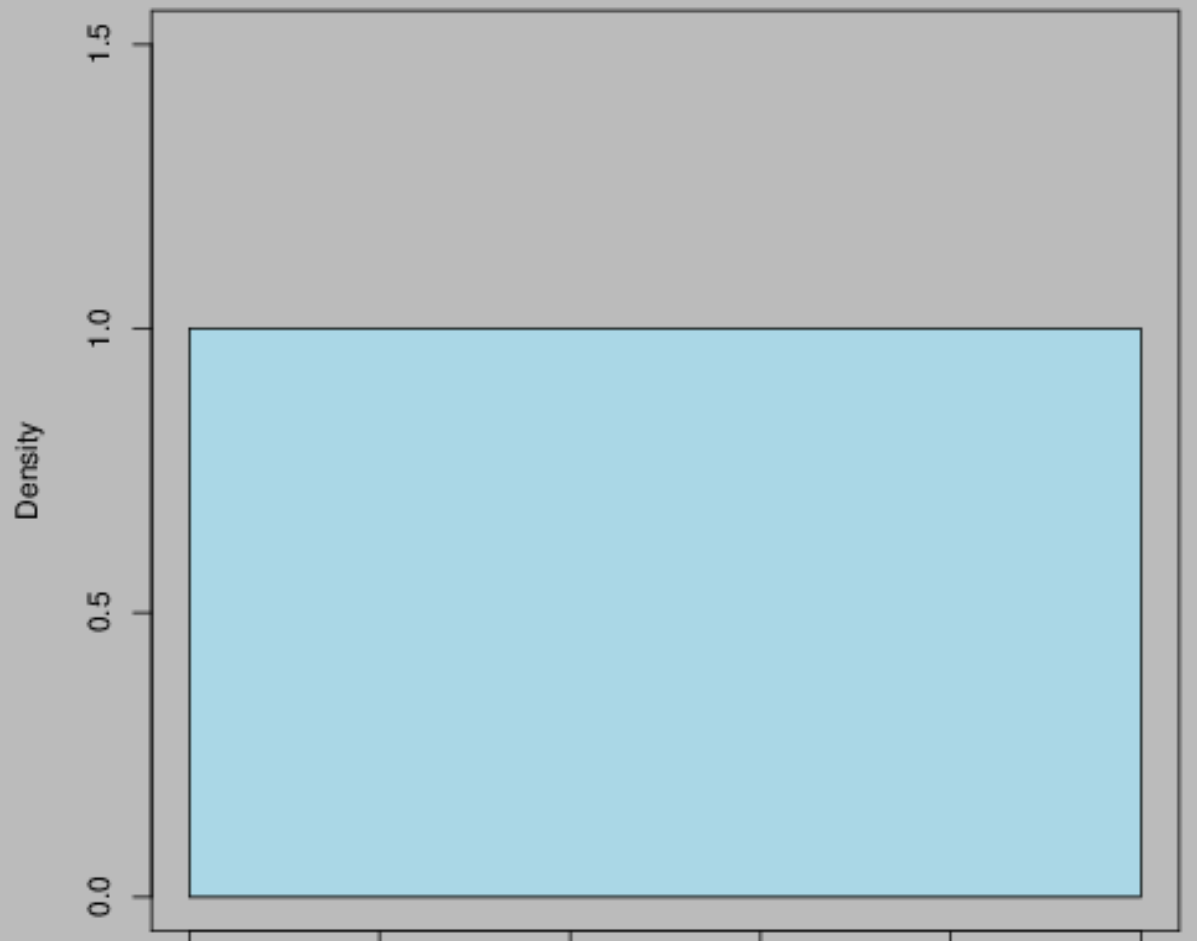
Number of errors committed when testing m null hypotheses.

Analysis result	H_o true	H_o false	Total
Reject	V	S	R
Don't reject	U	T	m-R
Total	m_0	$m-m_0$	m

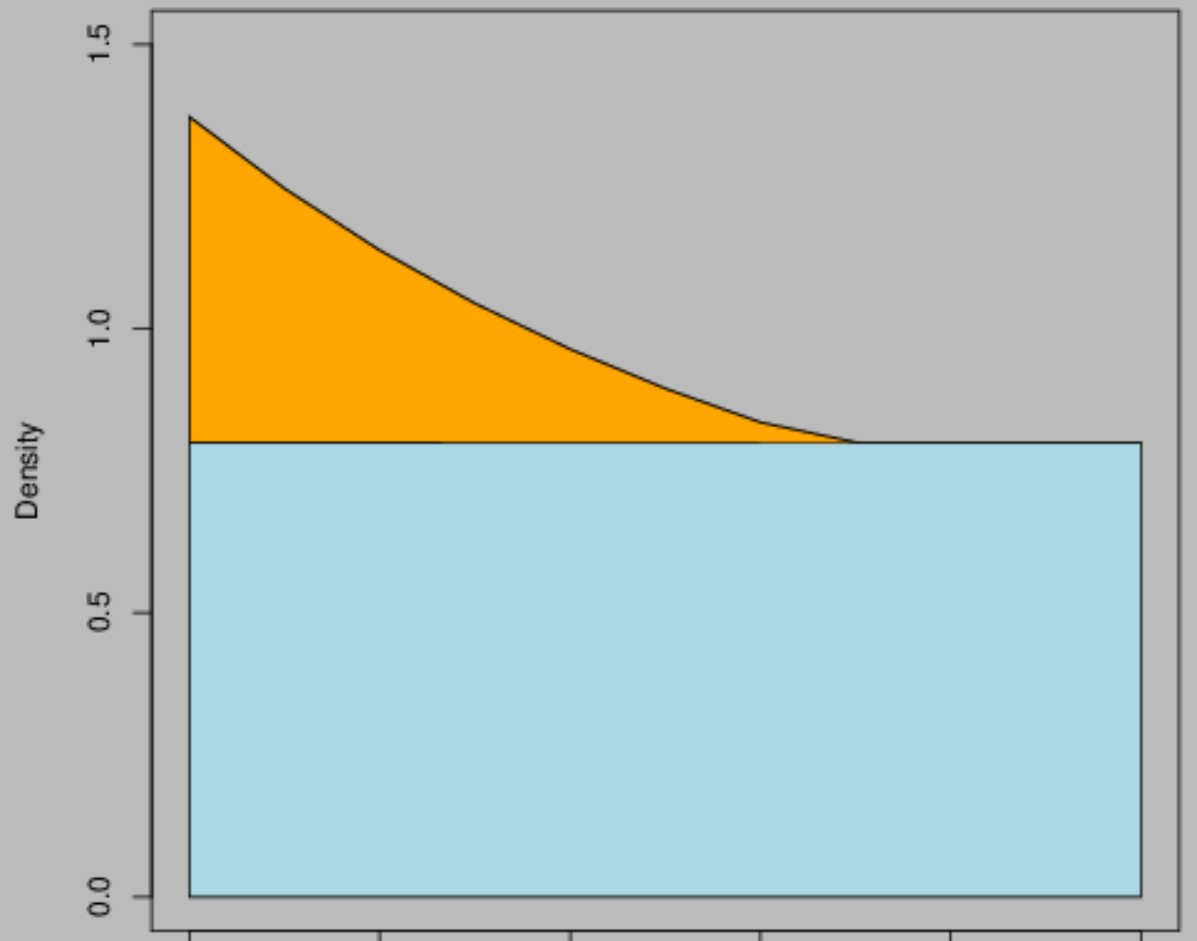
Proportion of false discoveries is $Q = \frac{V}{R}$. [Set to 0 for $R = 0$]

The false discovery rate is $FDR = E(Q) = E(\frac{V}{R})$

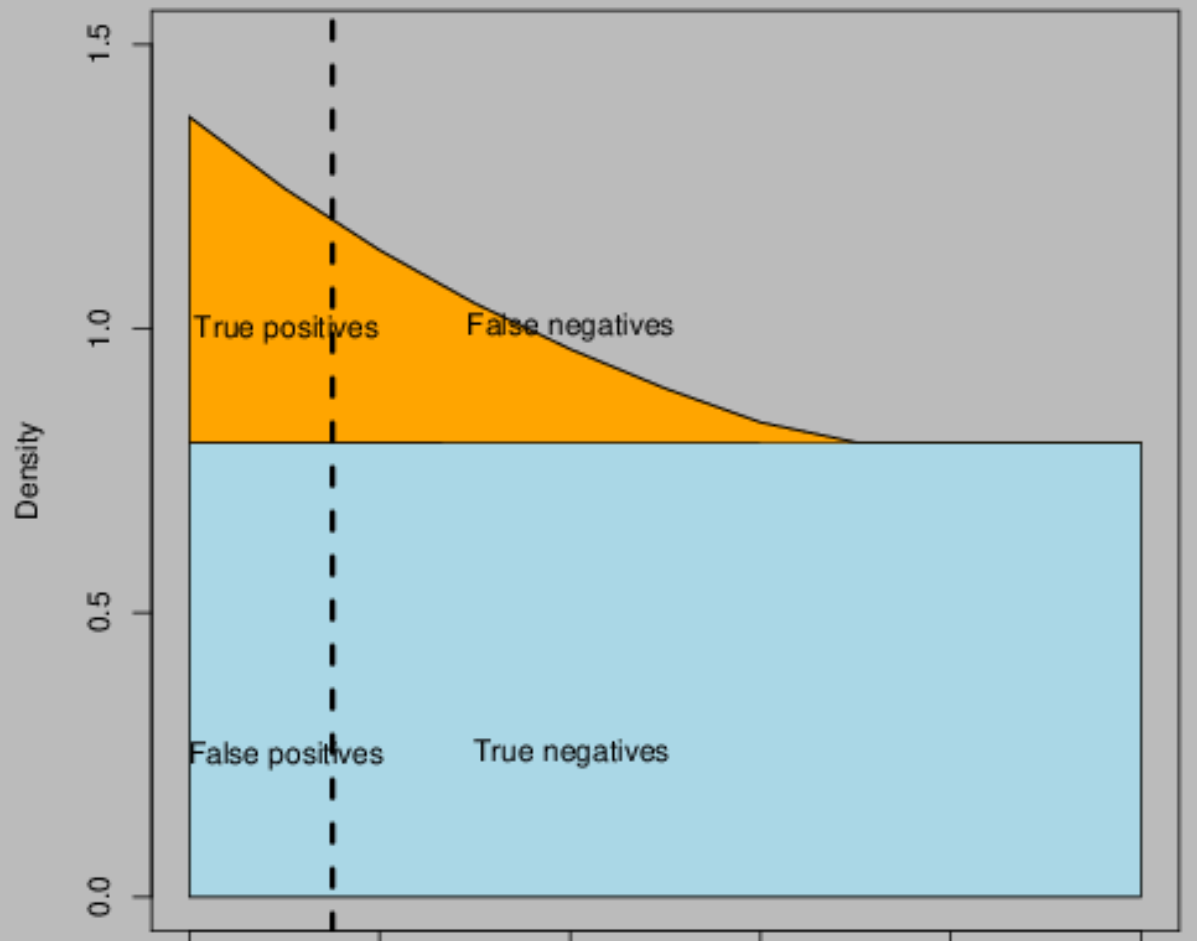
Estimating FDR



Estimating FDR



Estimating FDR



Estimating FDR — BH step-up

Benjamini-Hochberg step-up procedure to control the FDR at α.

1. Compute and order the individual p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Find $\hat{k} = \max\{k : \frac{m}{k} \cdot p_{(k)} \leq \alpha\}$
3. If $\hat{k}$ exists then reject hypotheses corresponding to $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(\hat{k})}$

Estimating FDR — BH step-up

p-values

$$\begin{aligned}
\tilde{p}_{(1)} &= \min\{\tilde{p}_{(2)}, mp_{(1)}\} \\
&\vdots \qquad\qquad \vdots \\
\tilde{p}_{(m-1)} &= \min\{\tilde{p}_{(m)}, \tfrac{m}{m-1}p_{(m-1)}\} \\
\tilde{p}_{(m)} &= p_{(m)}
\end{aligned}$$

Note that each p_i is smaller or equal to the criterium in Holm's method so controls the FWER.

Estimating FDR — BH step-up

If iid of the m_0 tests (and all tests independent) and ordered so the m_0 true tests comes first. Control FDR at level q:

$$
\begin{aligned}
E(V/R) &= \sum_{r=1}^{m} E[\frac{V}{r} 1_{R=r}] = \sum_{r=1}^{m} \frac{1}{r} E[V 1_{R=r}] \\
&= \sum_{r=1}^{m} \frac{1}{r} E[\sum_{i=1}^{m_0} 1_{p_i \leq \frac{qr}{m}} 1_{R=r}] = \sum_{r=1}^{m} \frac{m_0}{r} [1_{p_1 \leq \frac{qr}{m}} 1_{R=r}] = \cdots \\
&= \sum_{r=1}^{m} \frac{m_0}{r} [\sum_{i=1}^{m_0} 1_{p_1 \leq \frac{qr}{m}} 1_{R=r}] \\
&= q \frac{m_0}{m} \leq q
\end{aligned}
$$

q values

The q-value is defined to be the FDR analogue of the p-value.

$$q \text{ value}(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t)$$

The q-value of an individual hypothesis test is the minimum FDR at which the test may be called significant.

q values

- When all m null hypotheses are true then FDR control is equivalent to FWER control.
- FDR approach generally gives more power than FWER control and fewer Type I errors than uncorrected testing.
- The FDR bound holds for certain classes of dependent tests. In practice, it is quite hard to "break"

Exercises

Bootstrapping: evaluating complex methods

When we have complex data (or perhaps just big data combined with simple methods) and non-parametric methods then we still with to evaluate them.

How stable are the results?

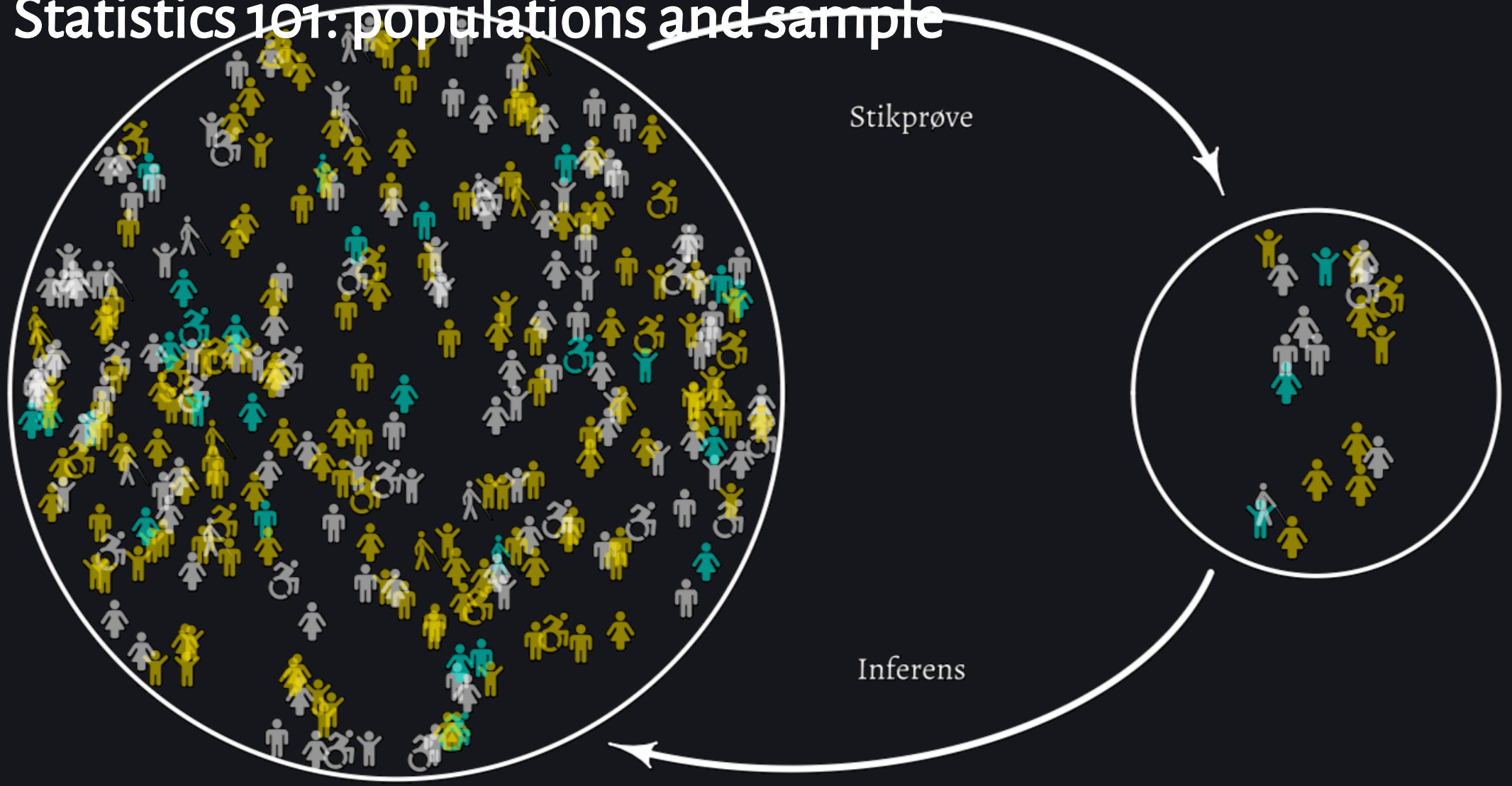
The bootstrap/jackknife procedures

Whenever we provide an estimate (mean, proportion, …) we *also* want to infer its precision!

We may or may not be able to formulate a full parametric (or semi-parametric model).

The bootstrap procedure allows us to estimate the standard error even in complicated situations or for non-standard statistics.

Statistics 101: populations and sample

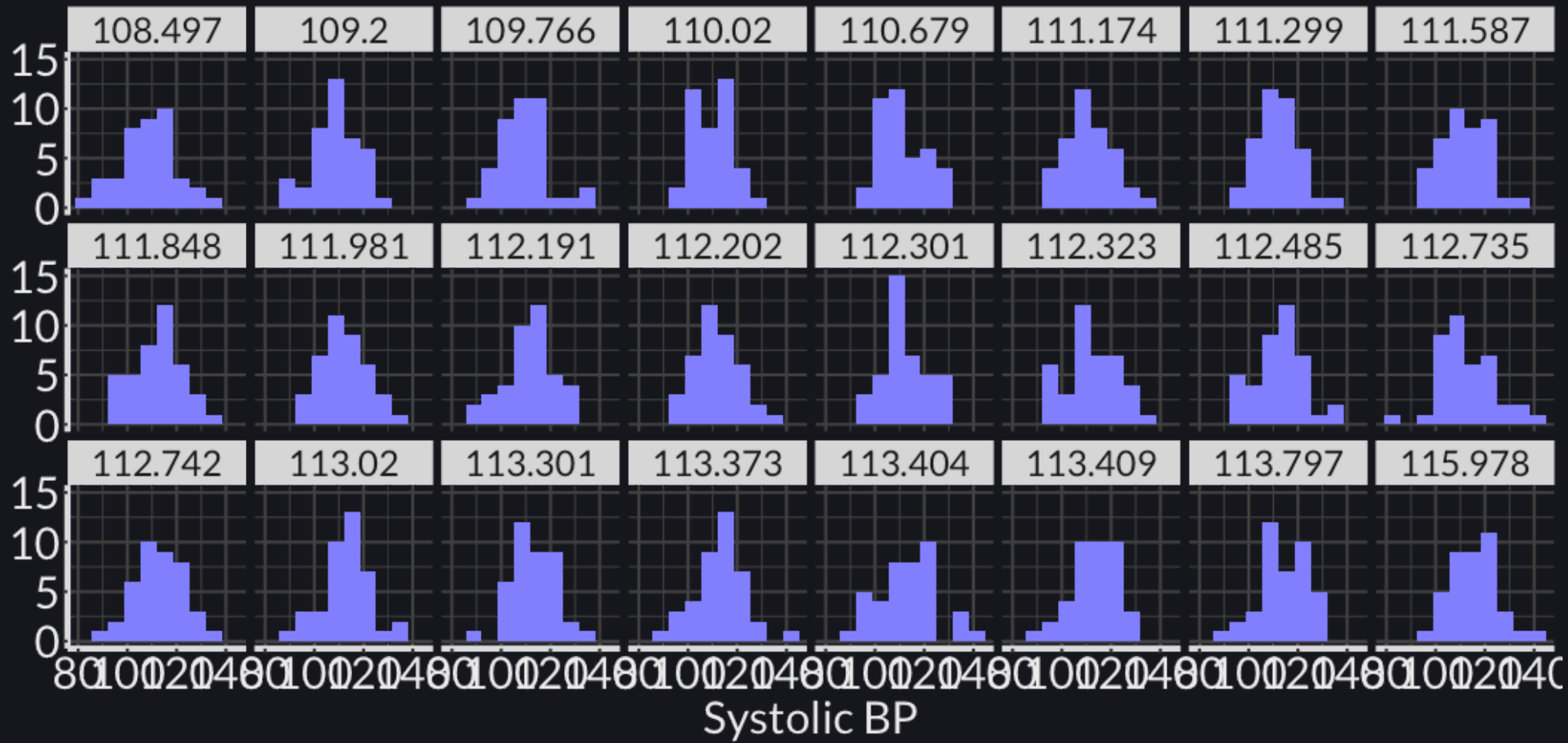


Statistics 101: multiple samples

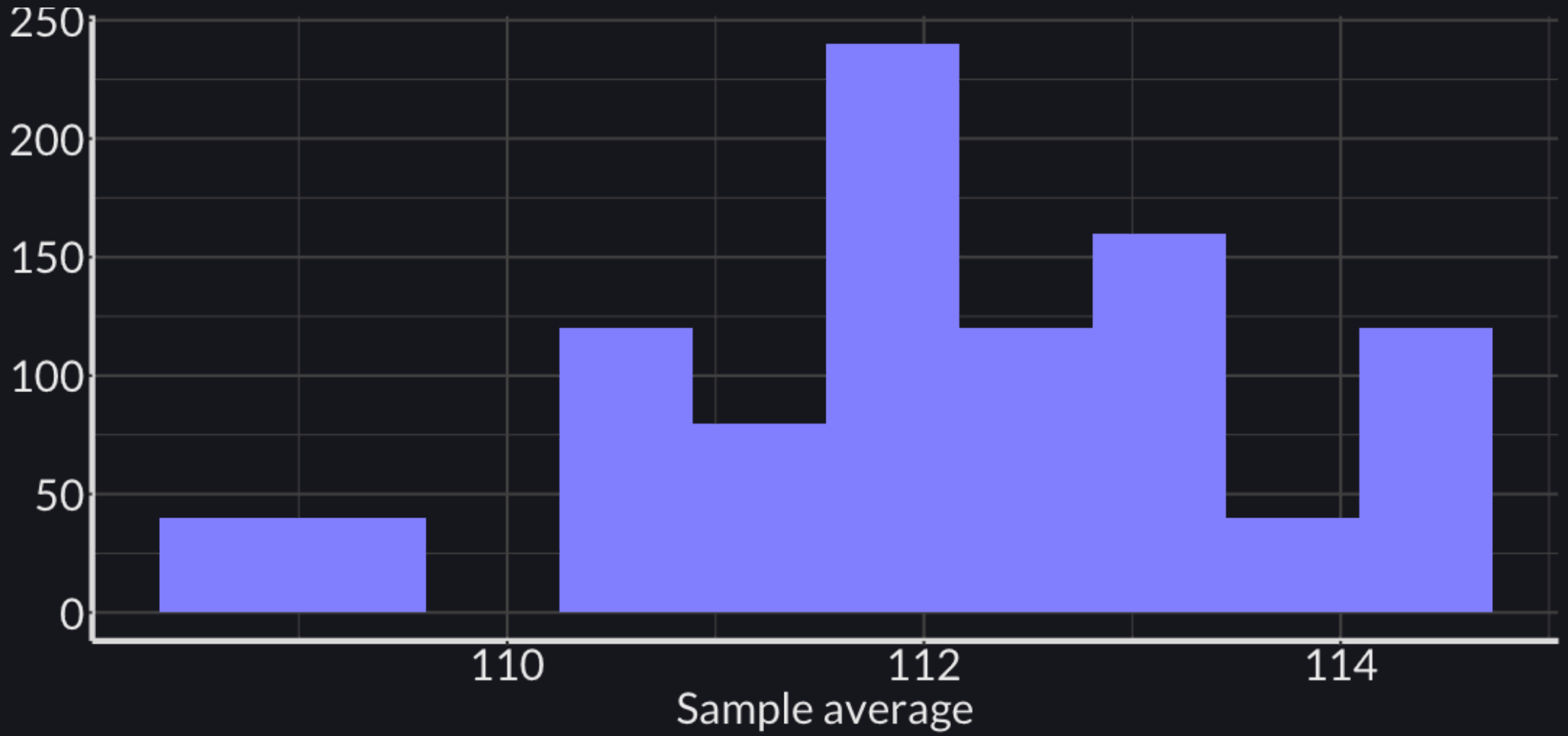
Different samples will result in different outcomes.

If we had the means to produce several samples we would know the sampling distribution.

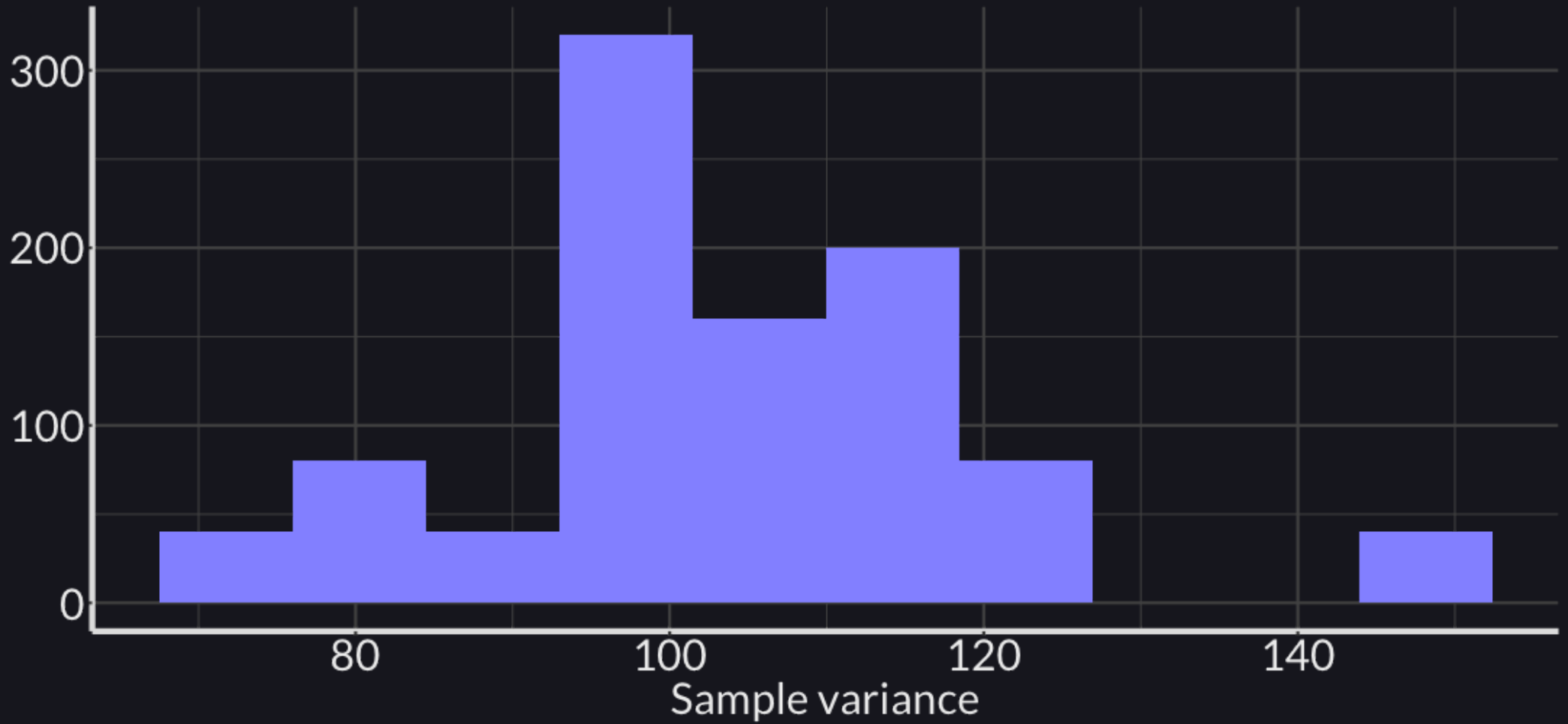
Sample variation



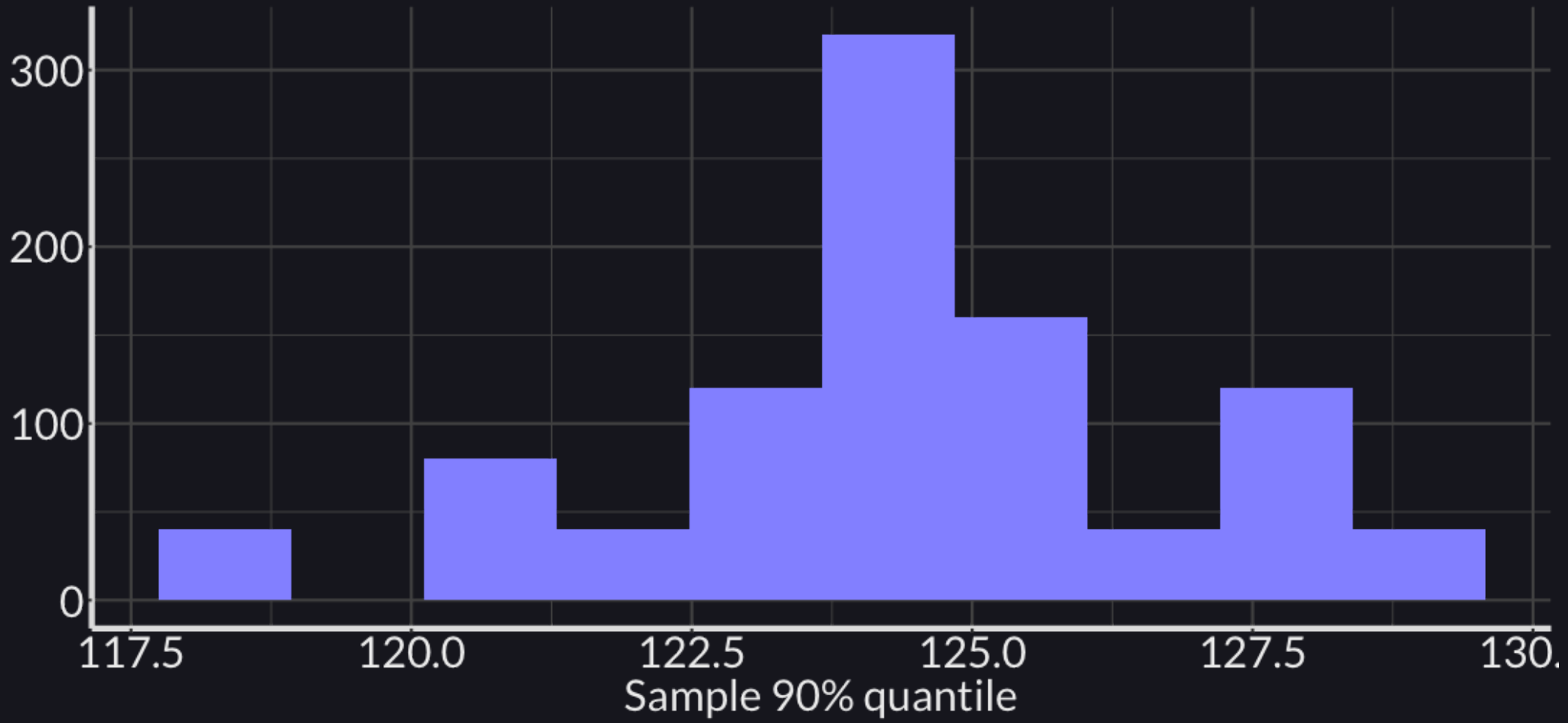
Sample variation



Sample variation



Sample variation



Resampling

Have observations $x_i \sim F$ and an estimate

$$\hat{\theta} = s(x)$$

for some estimation algorithm.

Want the SE of $\hat{\theta}$.

The jackknife estimate

Estimate $\hat{\theta}$ N times - once with each observation removed.

$$\hat{\theta}_{(-i)} = s(x_{(-i)})$$

Then the jackknife estimator is

$$SE(\hat{\theta}) = \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}(\hat{\theta}_{(-i)} - \hat{\bar{\theta}}_{(-)})^2}$$

Reduces to the standard error if $\hat{\theta}$ is a sample average.

The jackknife estimate

- Non-parametric, no assumptions on F, samples of size $N - 1$
- In general: the jackknife standard error is upwardly biased.
- Only to be used with smooth, differentiable statistic

```
x <-c(8.26, 6.33, 10.4, 5.27, 5.35, 5.61, 6.12, 6.19, 5.2,
      7.01, 8.74, 7.78, 7.02, 6, 6.5, 5.8, 5.12, 7.41, 6.52, 6.21,
      12.28, 5.6, 5.38, 6.6, 8.74)
CV <- function(x) sqrt(var(x))/mean(x)
CV(x)
```

```
[1] 0.2524712
```

```
library("bootstrap") ; res <- jackknife(x, CV)
```

The jackknife estimate

```
res
```

```
$jack.se
[1] 0.05389943

$jack.bias
[1] -0.009266436

$jack.values
 [1] 0.2563873 0.2565586 0.2384298 0.2507329 0.2513200
 [6] 0.2530603 0.2557374 0.2560293 0.2501992 0.2580969
[11] 0.2541045 0.2577524 0.2581067 0.2551946 0.2571038
[16] 0.2541711 0.2495662 0.2581975 0.2571609 0.2561093
[21] 0.2020978 0.2529980 0.2515338 0.2573745 0.2541045

$call
jackknife(x = x, theta = CV)
```

Estimating bias

When $\hat{\theta}$ is unbiased then

$$E(\hat{\bar{\theta}}) = \frac{1}{N}\sum_{i=1}^{N} E(\hat{\theta}_{(-i)}) = \theta$$

But if the procedure has bias

$$E(\hat{\theta}) = \theta + \frac{a}{N} + \frac{b}{N^2} + \text{rest}$$

then we can estimate the size of the bias from jackknife results.

Estimating bias

$$E(\hat{\tilde{\theta}} - \hat{\theta}) = \frac{a}{N(N-1)} + \text{rest}$$

Thus,

$$\text{bias}_{\text{jack}} = (N-1)(\hat{\tilde{\theta}} - \hat{\theta}) = \frac{a}{N}.$$

Furthermore, the bias-corrected jackknife estimate,

$$\hat{\theta}_{\text{jack}} = \hat{\theta} - \text{bias}_{\text{jack}}$$

is an unbiased estimate of θ up to second order.

Improvement on the jackknife

Instead of removing 1 observation at a time, remove d. Then there are $\binom{N}{d}$ sets

$$SE = \sqrt{\frac{N-d}{d\binom{N}{d}} \sum((\hat{\theta}_{(Z)} - \hat{\bar{\theta}}_{(-)})^2}$$

… or use the bootstrap!

Nonparametric bootstrap

If we could draw extra samples from the population it would be easy!

Use the sample as the population and generate "fake samples"

Population $\rightarrow$ Sample $\rightarrow$ "Fake sample"

Nonparametric bootstrap

Get a random bootstrap sample from the sample *with replacement*

$$x^* = (x_1^*, x_2^*, \ldots, x_N^*)$$

Then we can get

$$\hat{\theta}^* = s(x^*)$$

Do that B times and get information about the full distribution.

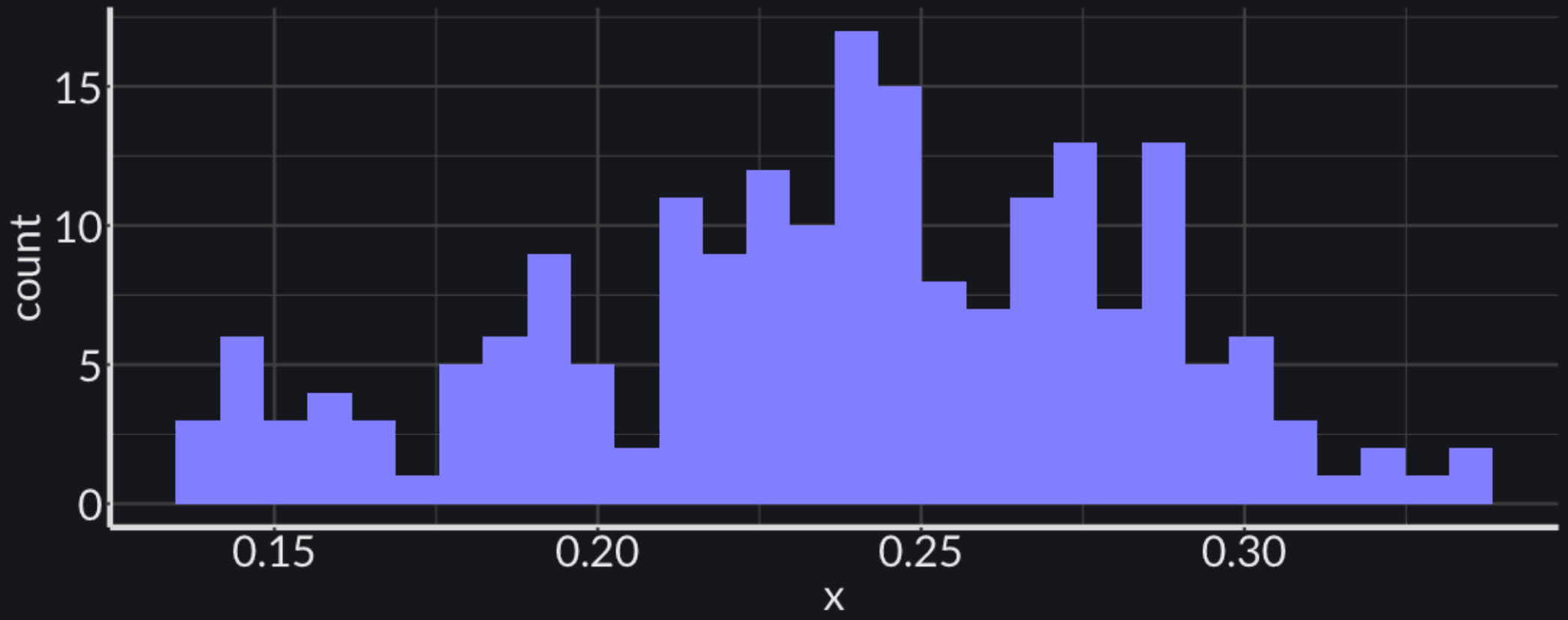
Jackknife vs bootstrap

- Jackknife provides stable results (will always get the same result) whereas bootstrap varies.
- Jackknife only estimates the variance of the point estimator whereas the bootstrap provides information on the distribution.

$$SE = \sqrt{\frac{\sum(\hat{\theta}^{*b} - \hat{\theta}^{*-})^2}{B-1}}$$

Nonparametric bootstrap in R

```
results <- bootstrap(x, 200, CV)
```



Parametric bootstrap

The nonparametric bootstrap made no assumptions about the distribution. Use distribution information if known.

- Fit model to data
- Draw B samples of random numbers from the fitted model
- Use those for bootstrap

Useful for small sample sizes (assuming the model holds), difficult evaluations, … Sampling from the "wrong" distribution and forgetting the uncertainty. Retains the information in the explanatory variables but needs the error distribution.

Parametric bootstrap - resample residuals

- Fit model to data, keep predictions $\hat{y}_i$ and compute a vector of residuals, $\hat{\epsilon}_i = y_i - \hat{y}_i$.
- Create new sets of observations $y^* = \hat{y}_i + \hat{\epsilon}_j$ using a random residual.
- Refit the model using the new set of response variables, and compute the statistic
- Do that B times

Retains the information in the explanatory variables. What to resample?

Rough R code

```r
x <- c(5, 9, 8, 4, 7, 4, 2)
# Non-parametric bootstrap
x.star <- sample(x, replace = TRUE)
```

```r
# Parametric bootstrap for assumed Gaussianity
x.star <- rnorm(length(x), mean = mean(x), sd = sd(x))
```

```r
# Mean approximates the mean for Gaussian distribution for residuals
resids <- x - mean(x)
x.star <- mean(x) + sample(resids, replace=TRUE)
```

What to do?

Depends on the situation.

- The structure of the data might make some options easier.
- Belief about the parametric model would improve efficiency.
- Belief about the bias of the estimate would influence the choice.

Bootstrap confidence intervals

Standard 95% confidence intervals

$$\hat{\theta} \pm 1.96SE$$

Could get that directly from the bootstrap results.

```
mean(results$thetastar) + c(-1.96, 1.96)*sd(results$thetastar)
```

```
[1] 0.1497948 0.3262128
```

Only really works if the distribution is symmetric

Bootstrap percentile confidence intervals

Generate the "full" distribution. Cut off 2.5% at each end. Use an improvement that depends on the precision of the percentiles.

```
bcanon(x, 2000, CV)
```

```
$confpoints
      alpha bca point
[1,] 0.025 0.3035158
[2,] 0.050 0.3307189
[3,] 0.100 0.3595159
[4,] 0.160 0.3875672
[5,] 0.840 0.6210273
[6,] 0.900 0.6629274
[7,] 0.950 0.7128380
[8,] 0.975 0.7218803

$z0
```

Bootstrap percentile confidence intervals

Can also use the t distribution

```
boott(x, CV)
```

```
$confpoints
          0.001      0.01     0.025      0.05       0.1
[1,] 0.2793437 0.3234873 0.3348129 0.3488559 0.3805315
          0.5       0.9      0.95     0.975     0.99
[1,] 0.47362 0.6681084 0.7423936 0.7955465 1.383483
         0.999
[1,] 1.596856

$theta
NULL

$g
NULL
```