Central Limit Theorem and Introduction to Inference

Sumanta Basu

Reading

- Reading: Textbook Sections 3.1, 4.1-4.2
- Recommended Exercise: 4.1, 4.5, 4.7, 4.9, 4.13

Prelim 1 will **NOT** include readings from Chapter 4 of textbook.

Central Limit Theorem (CLT)

Populations and samples

Goal: Draw scientific inference about a specific aspect of a target population from a *representative* sample

Example: 2012 Cherry Blossom 10-Mile Run

Target population: all runners who finished the run in 2012.

Research question: What is the average age of this population?

There were 16,924 people who finished… would be hard to ask them all!

Draw simple random sample (SRS).

Drawing sample

Options?

- before race starts, ask 50 people?
- send mass email with survey?
- give away energy bars to anyone who will answer your survey?
- be part of race and approach people you encounter along the way?
- stand by finish line for 30 minutes and ask whoever completes race during that time?

Drawing sample

None of those are good. Remember lectures on sampling!

None give a **SRS**.

Estimating mean age from sample

Let μ be the target population's mean age:

$$\mu = \frac{\text{age}_1 + \text{age}_2 + \cdots + \text{age}_{16,924}}{16,924}$$

This is unknown!

But let's select a SRS of size 10:

$$\bar{x} = \frac{31 + 30 + 48 + 41 + 30 + 33 + 25 + 28 + 33 + 39}{10} = 33.8$$

Do we "trust" our answer?

Is $\bar{x} = 33.8$ the same thing as μ?

The R in SRS is Random

When we draw an SRS, we are doing something **random**.

$\bar{x} = 33.8$ is a **realization** of this random process!

> **Key idea:** Imagine if we were to repeat this sampling process again and again.

SRS

Recall definition of **SRS:**

> Choose n units from target population at random so that each possible subset of size n is equally likely to be chosen.

Image we repeated experiment:

- draw 10 people
- calculate $\bar{x}$

We might get 34.0 and 29.3… different realizations of a **random variable**.

Sample mean from SRS

33.8, 34.0, 29.3 are *realizations* of a **random variable**, call it $\bar{X}_{10}$.

> $\bar{X}_{10}$ in words - *“draw a SRS of 10 people and average their ages”*

Why is it a random variable?

- a **random variable** is a numerical summary of a random outcome.
- “draw a SRS of 10 people” is a random experiment
- “average their ages” - sample mean of the ages of the random people drawn is a numerical summary

Sample mean as a random variable

We can write:

$$\bar{X}_{10} = \frac{X_1 + \cdots + X_{10}}{10}$$

X_i = age of the i th person drawn in SRS (also a random variable!)

Suppose my SRS gives

$$31\ 30\ 48\ 41\ 30\ 33\ 25\ 28\ 33\ 39$$

What is X_5?

Sample mean as a random variable

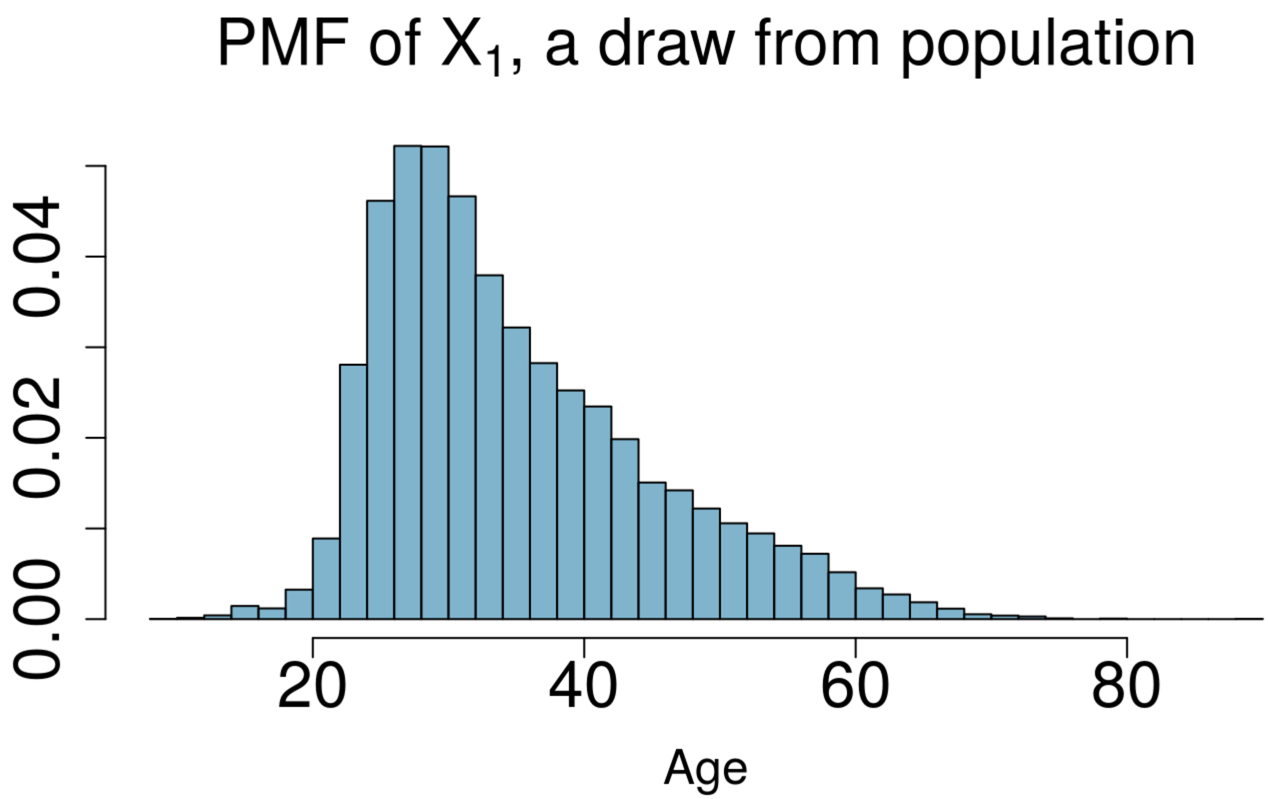
What is $\bar{X}_{10}$'s distribution?

If we had entire population of ages, we could simulate…

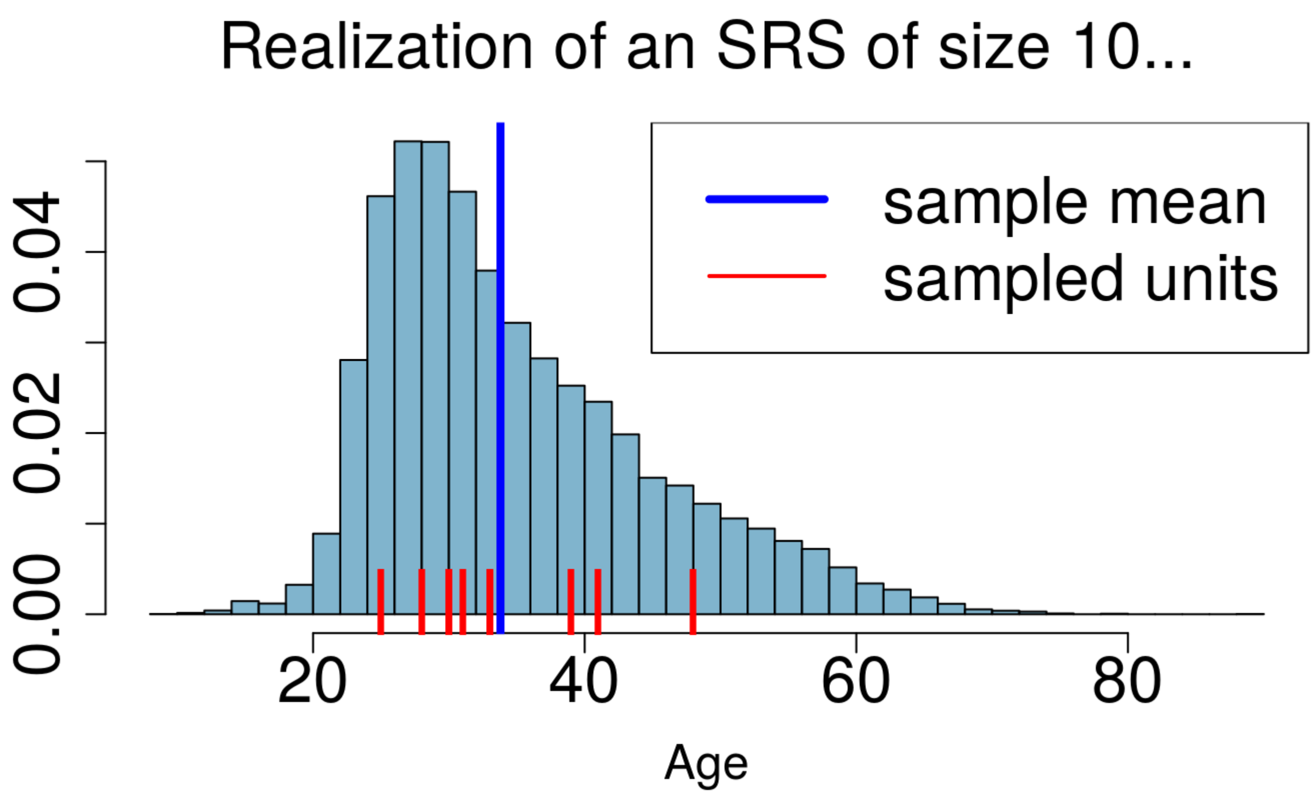
…actually in this case we do have the entire population's ages.

Population of ages

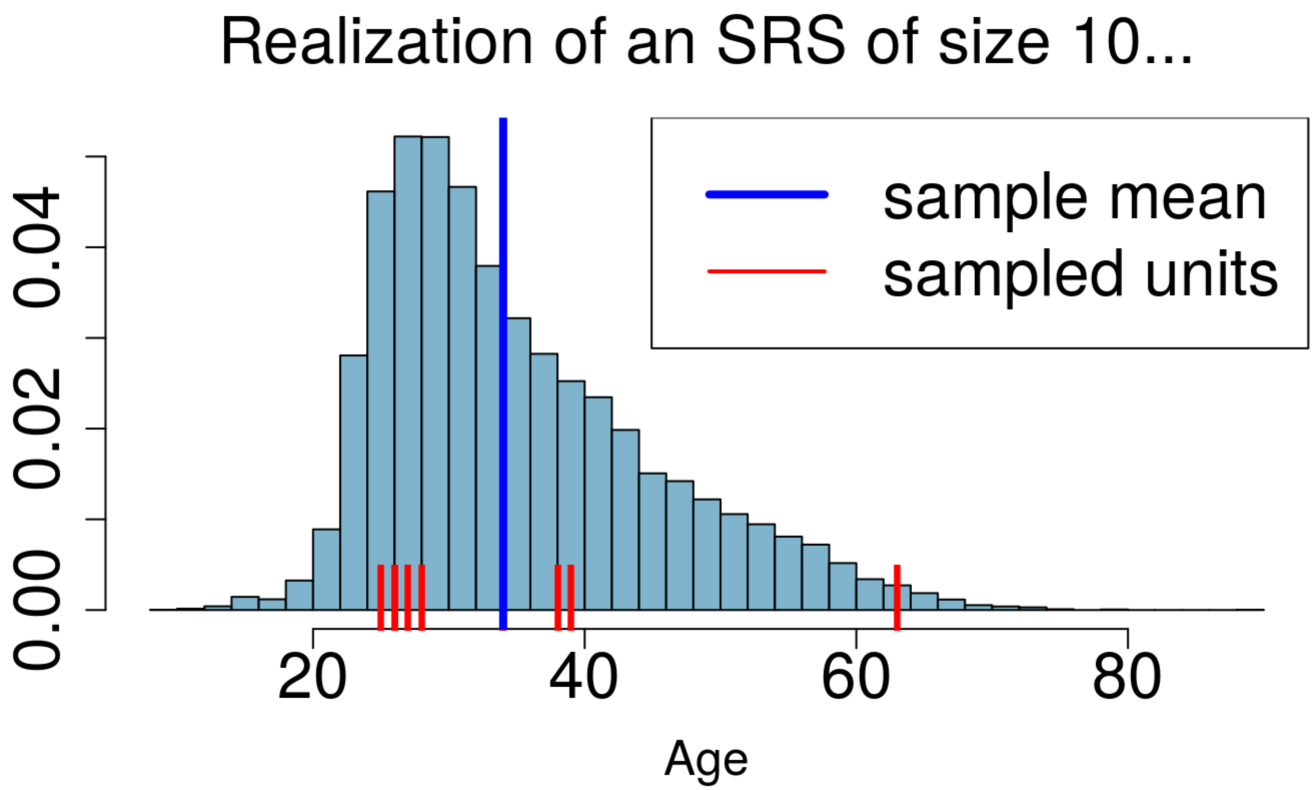
Drawing from this population



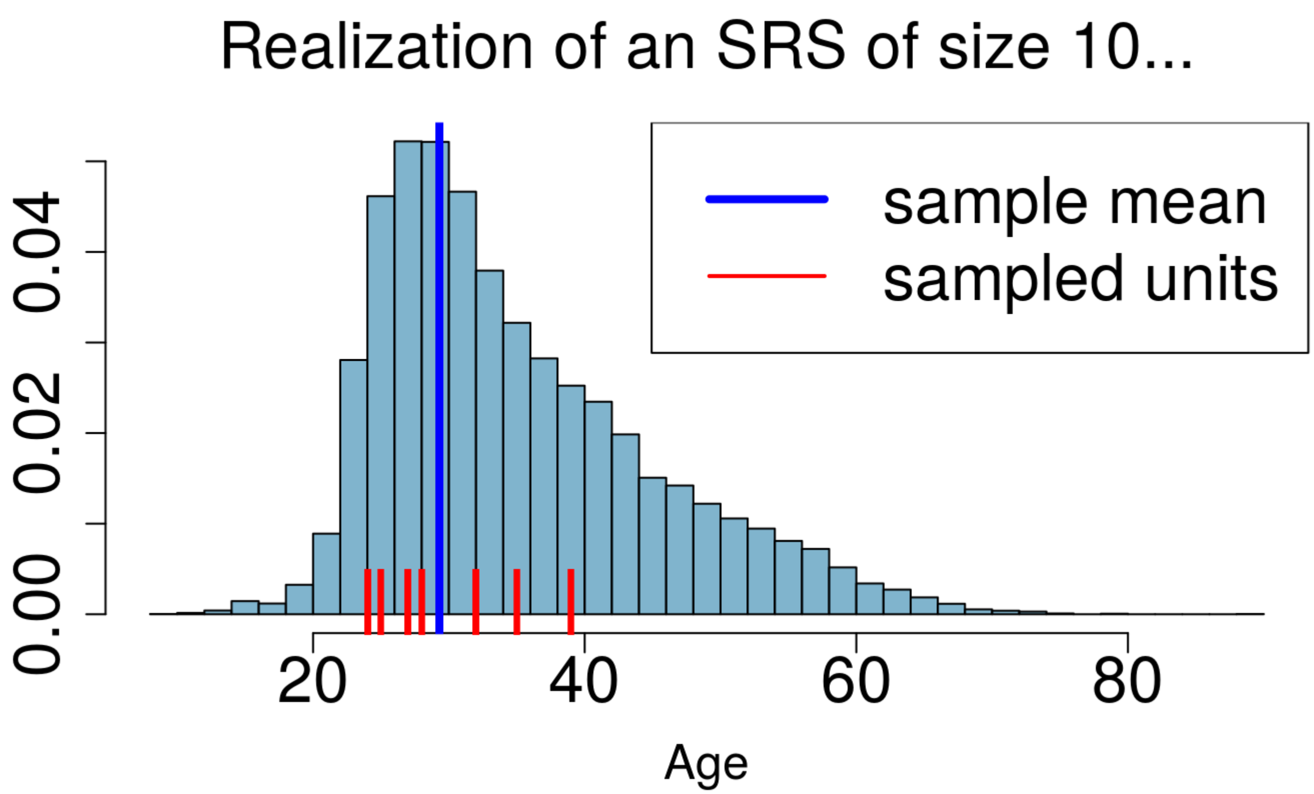
Drawing from this population



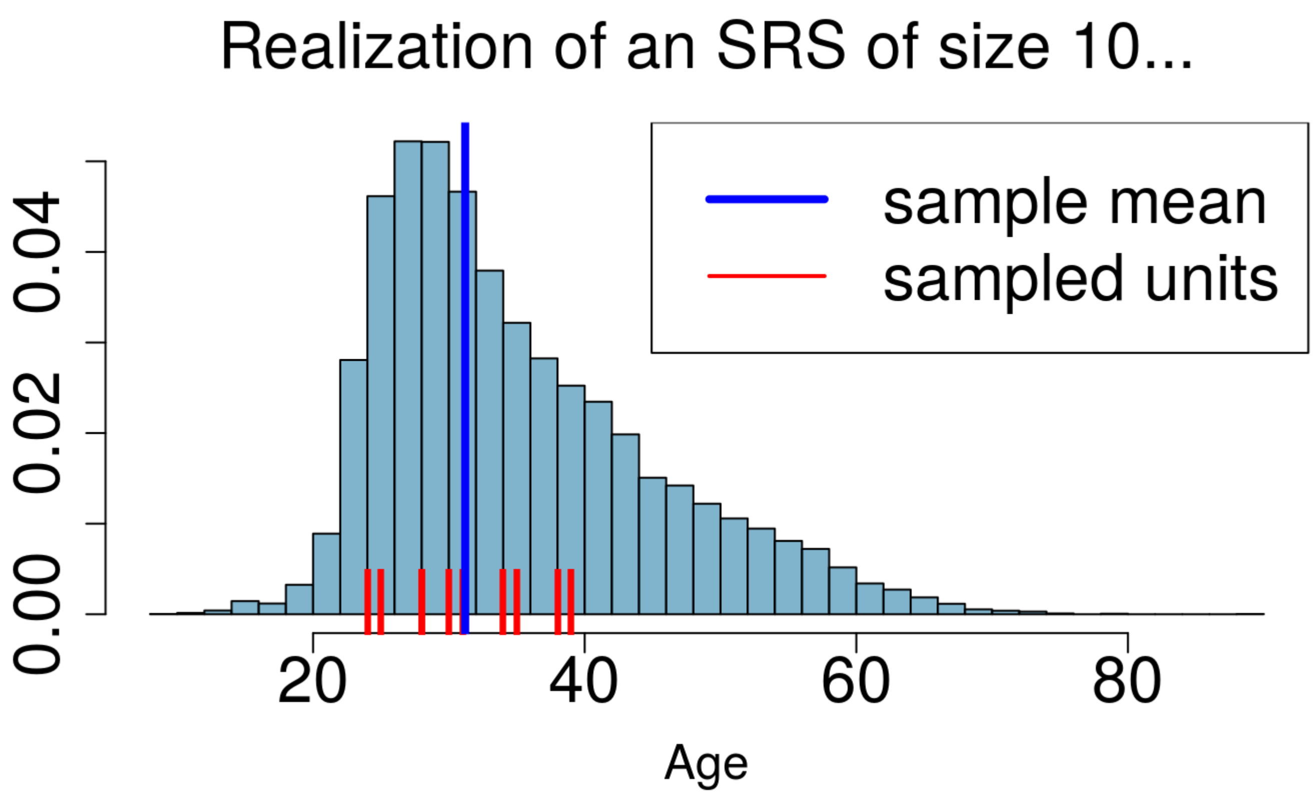
Drawing from this population



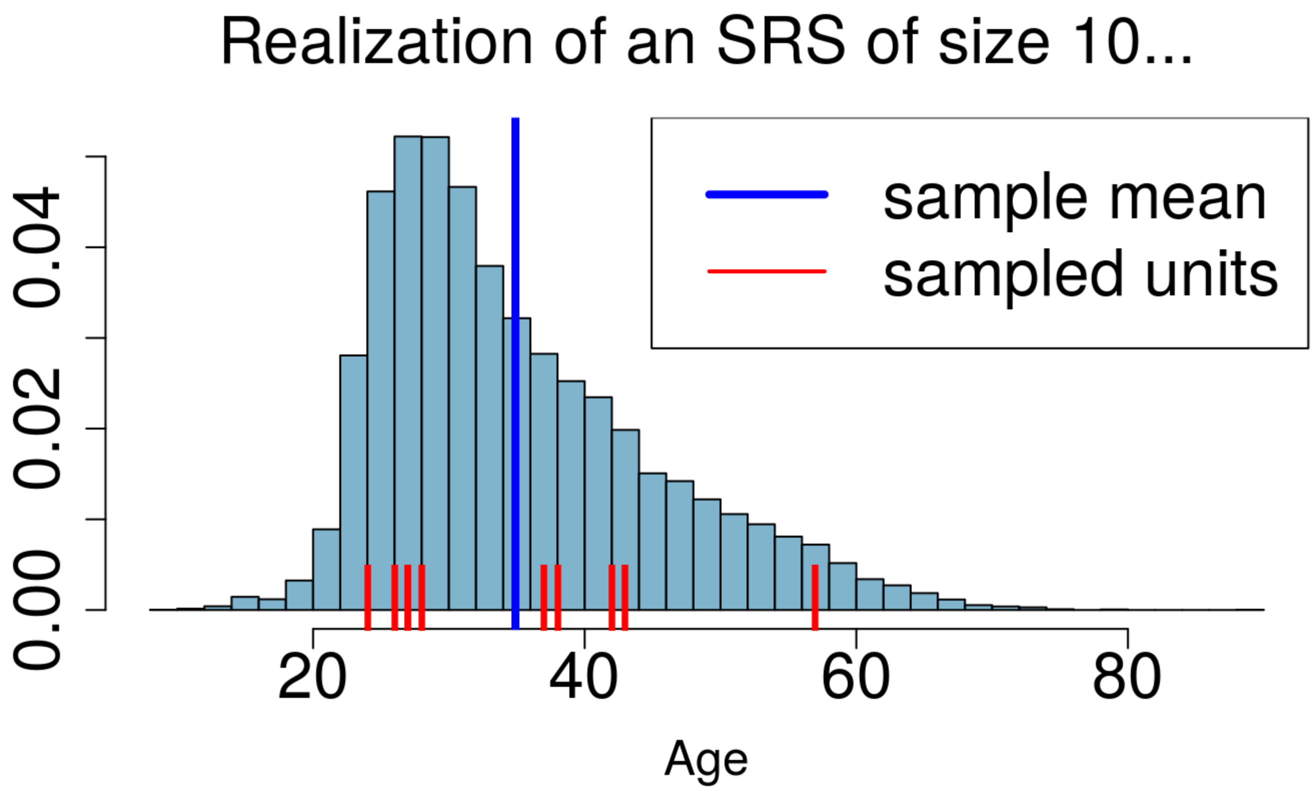
Drawing from this population



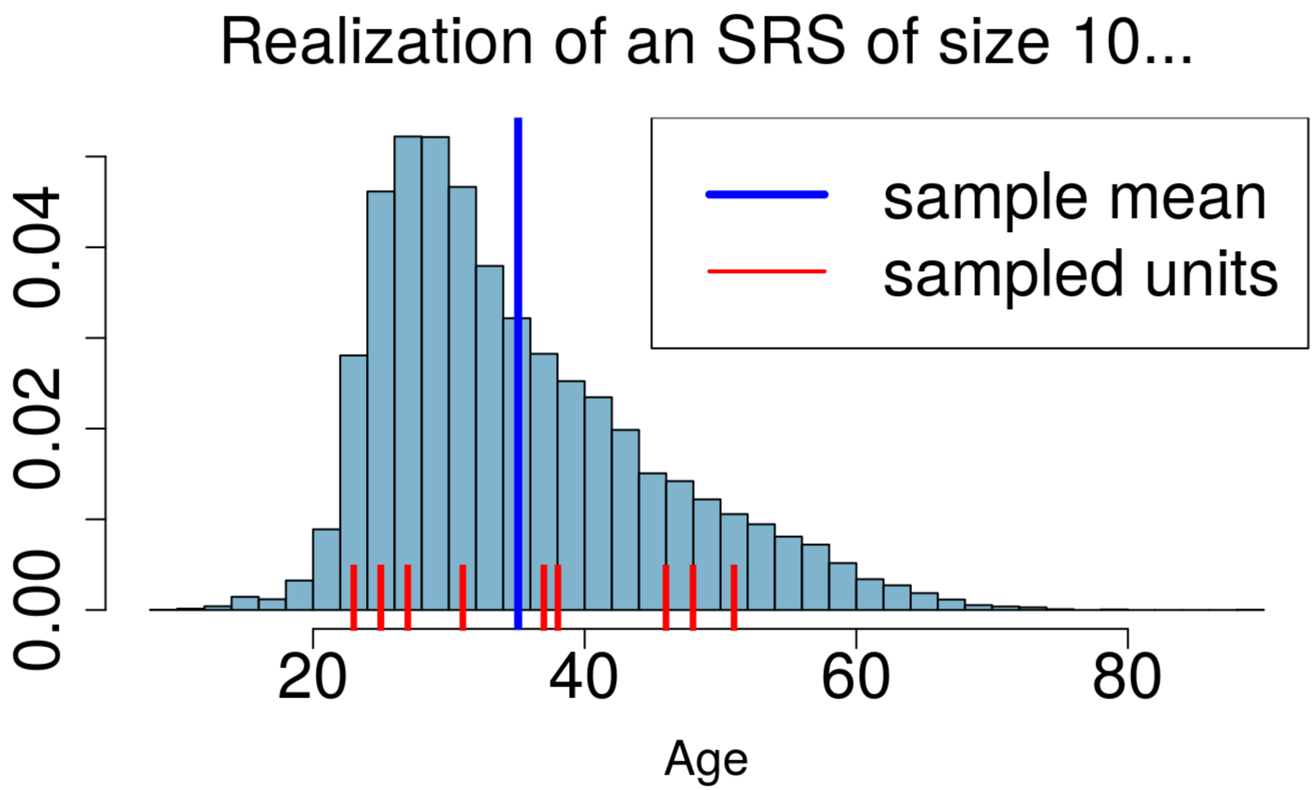
Drawing from this population



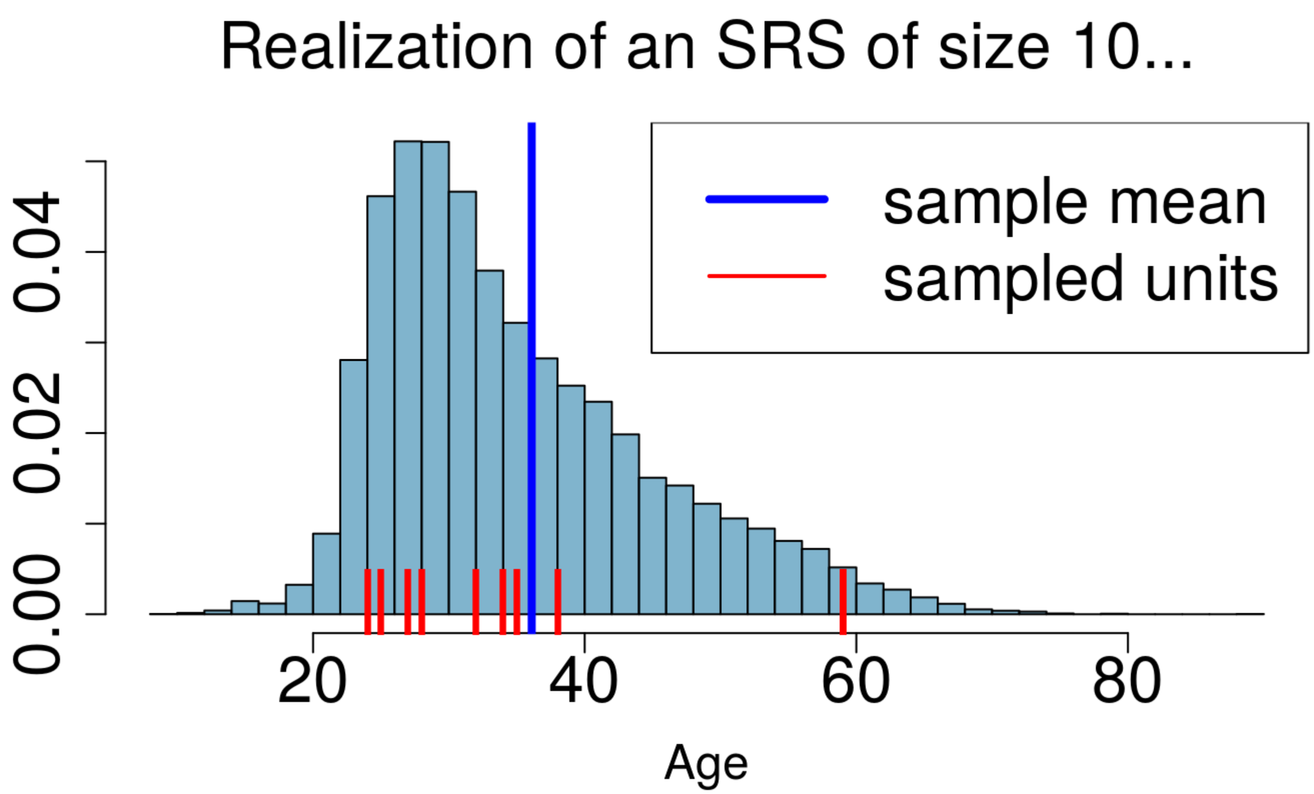
Drawing from this population



Drawing from this population

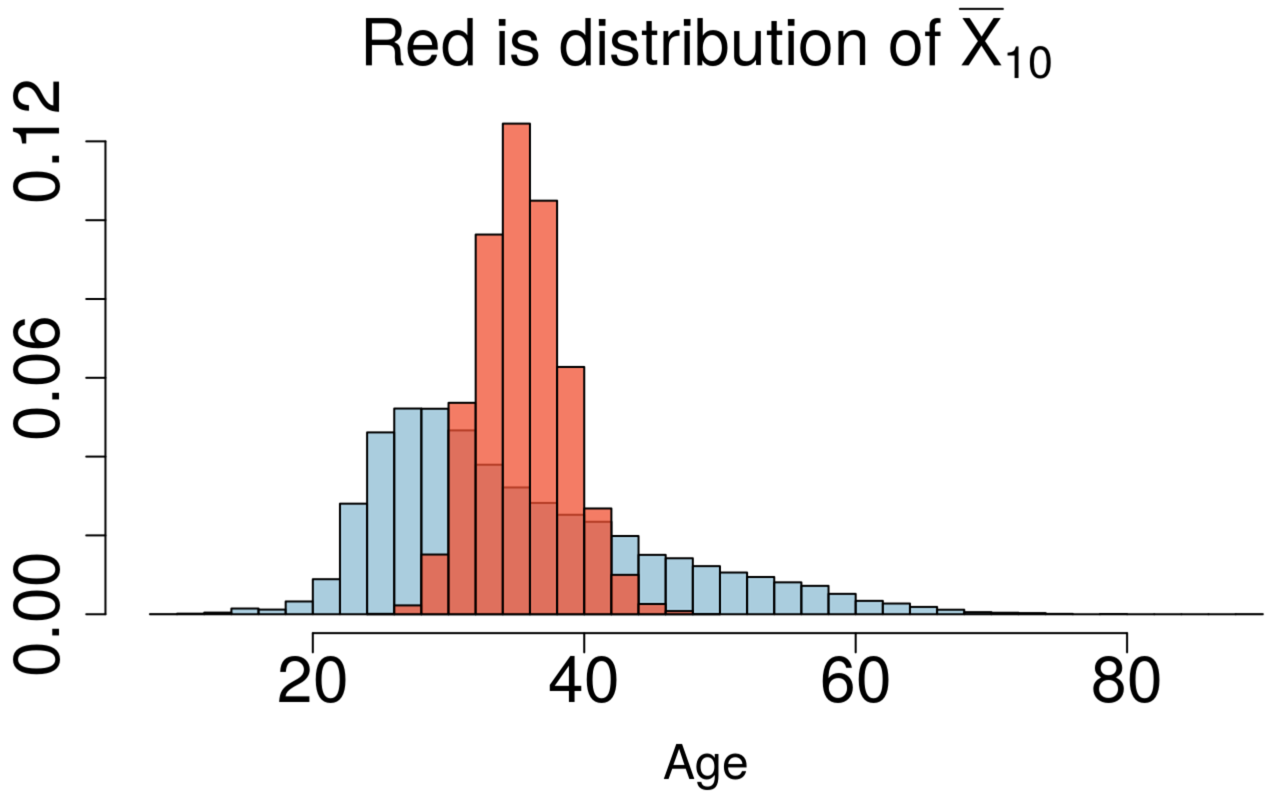


Drawing from this population

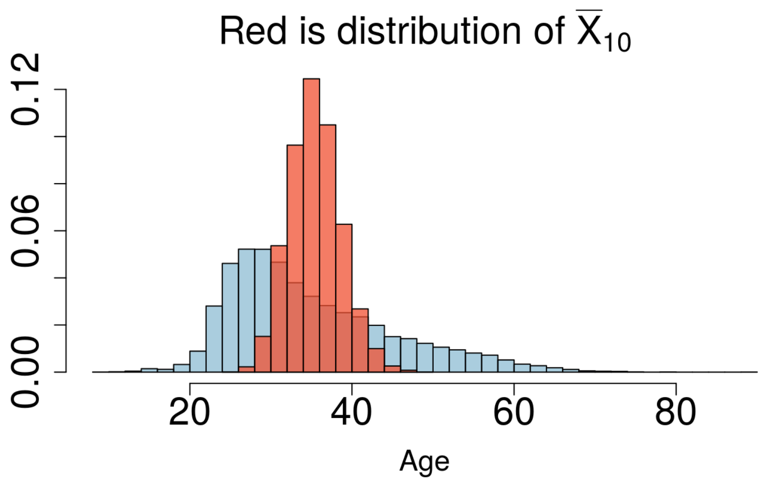


Sample mean's distribution

By Monte Carlo simulation, we can get $\bar{X}_{10}$'s distribution:

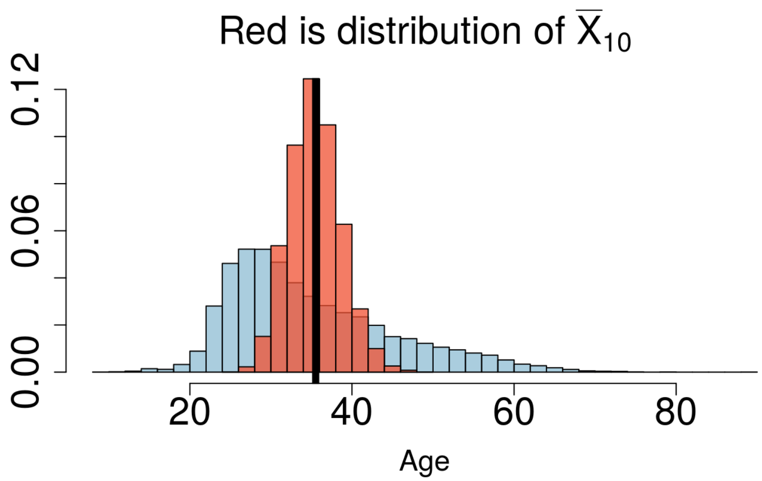


Comments



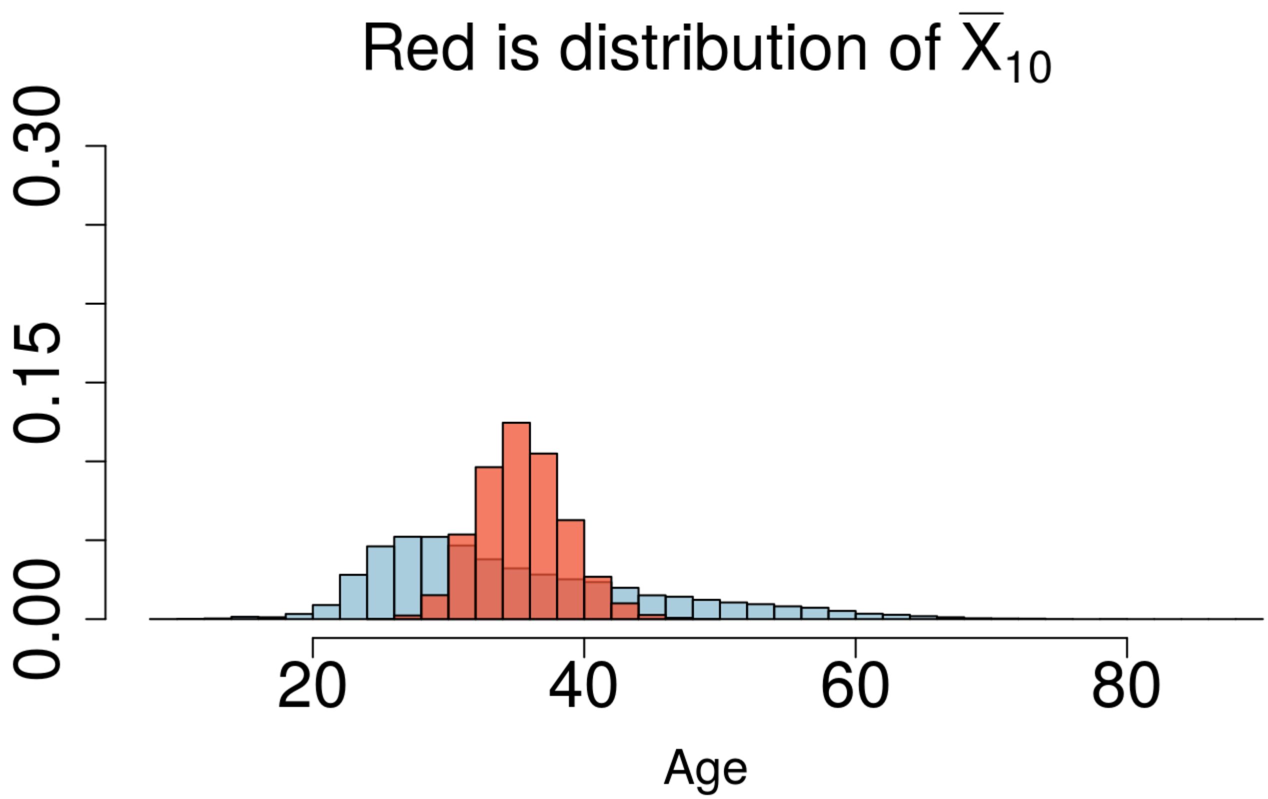
- when you compute a statistic from a sample using SRS, that is a *realization of a random variable*
- Red shows the distribution of the statistic $\bar{X}_{10}$
- this distribution is called the **sampling distribution** of the statistic
- could only simulate sampling distribution since we had entire population's data (not realistic)

Observations

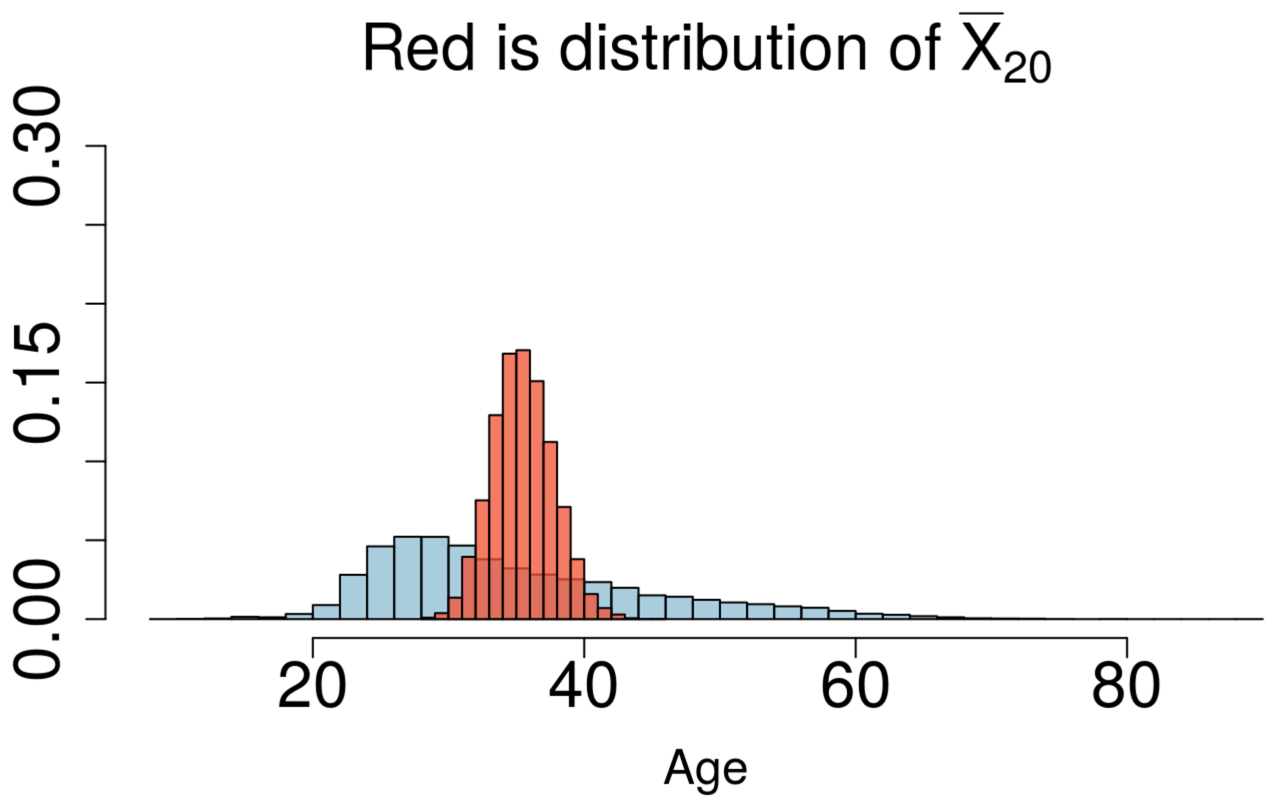


- **sampling distribution** of $\bar{X}_{10}$ seems to be centered at population mean, μ (shown in black)
- variance of $\bar{X}_{10}$ is smaller than variance of population
- shape looks normal!!
- that's surprising since population distribution was not!

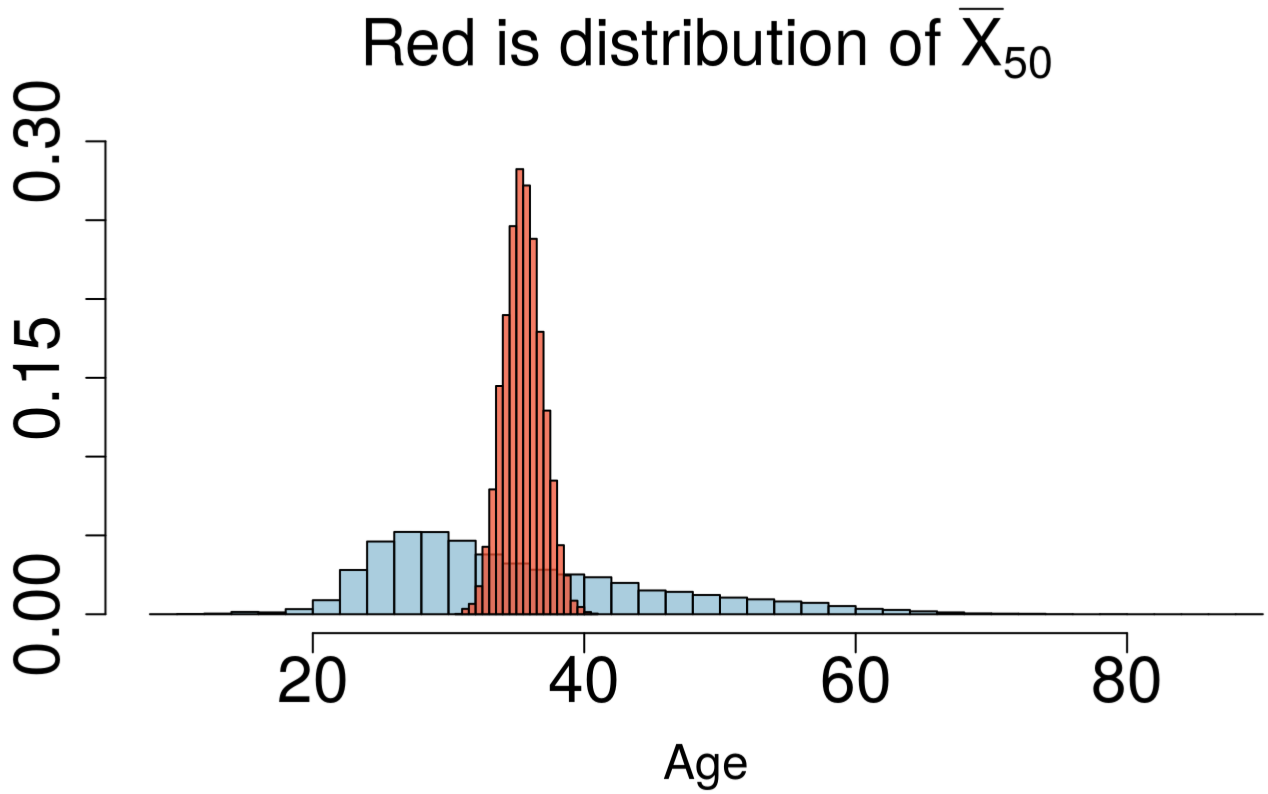
Effect of sample size?



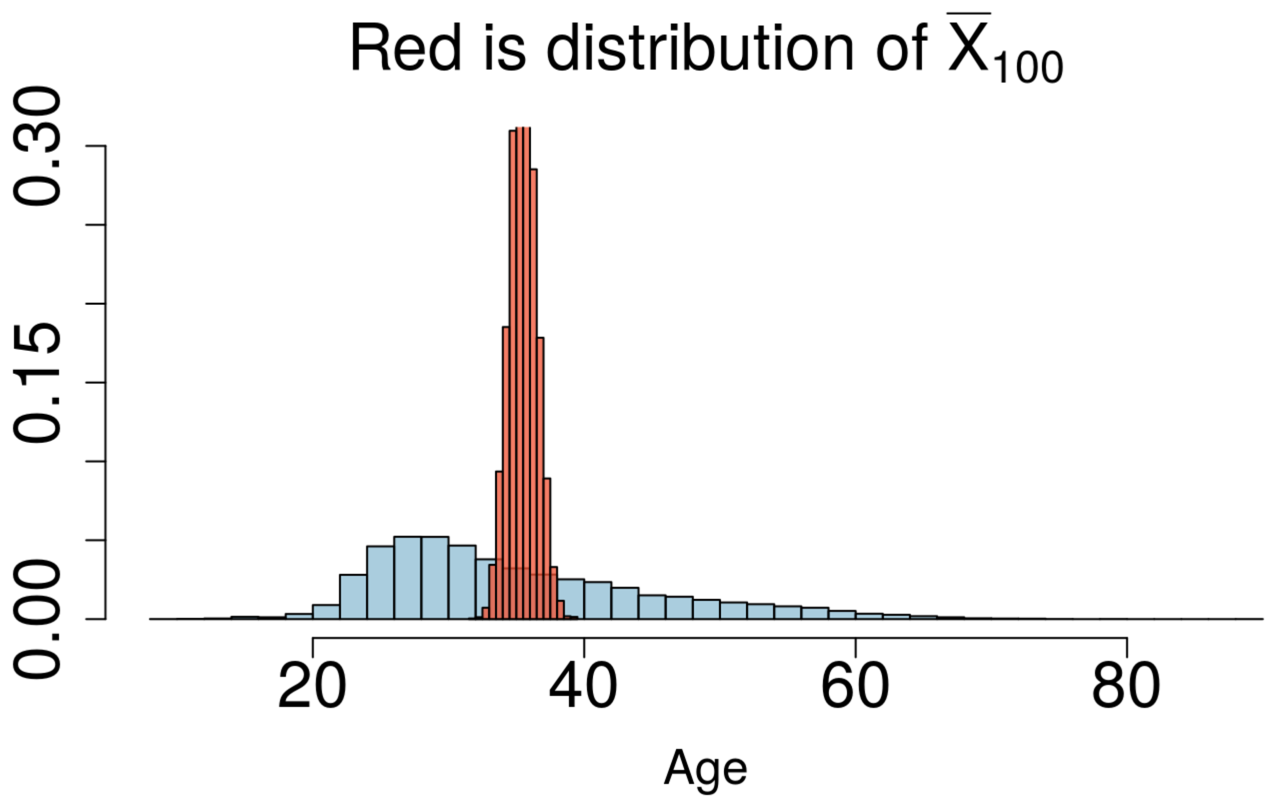
Effect of sample size?



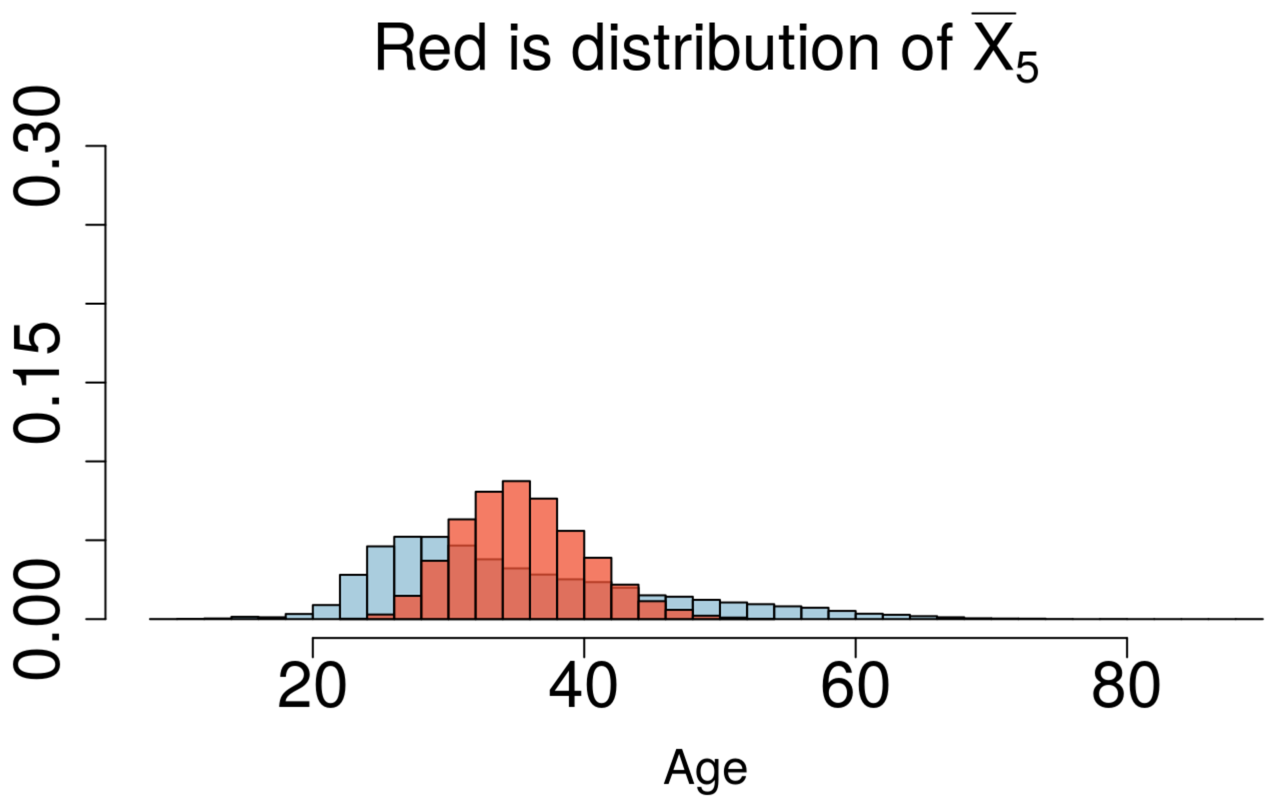
Effect of sample size?



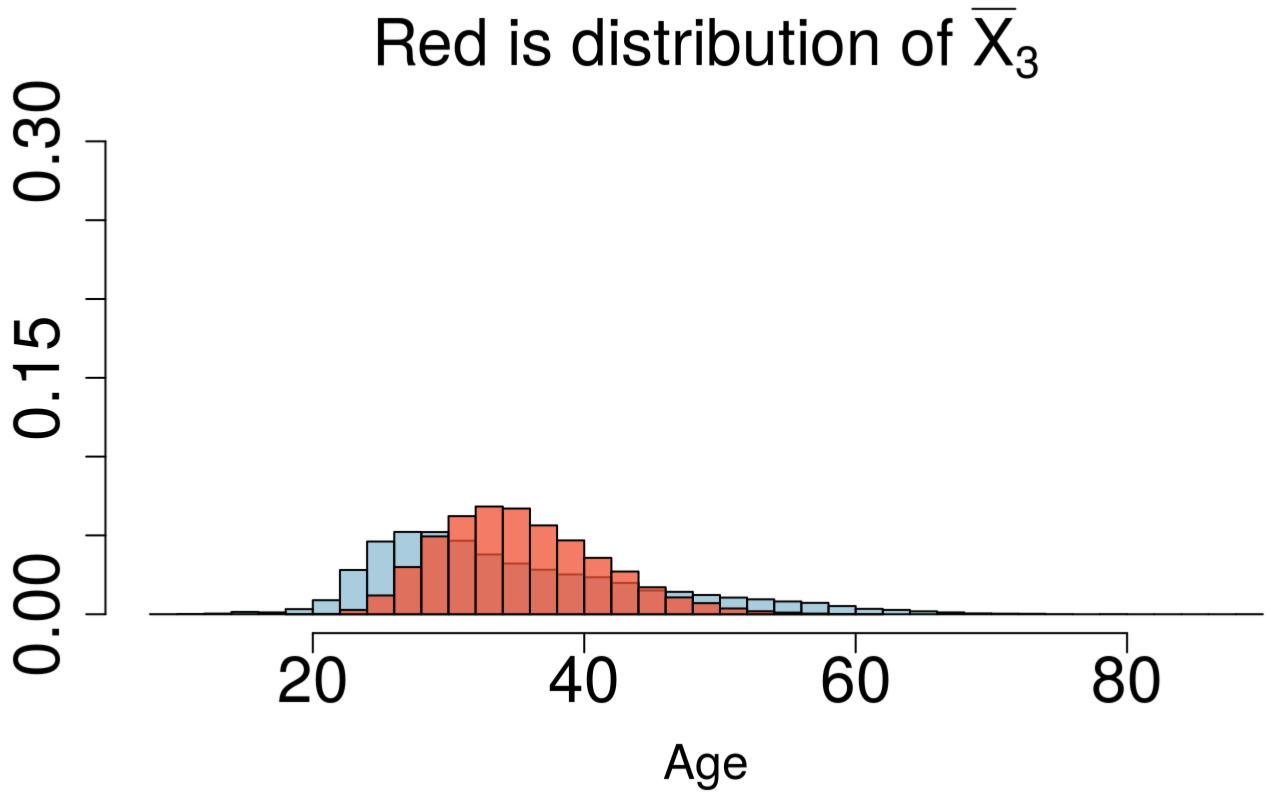
Effect of sample size?



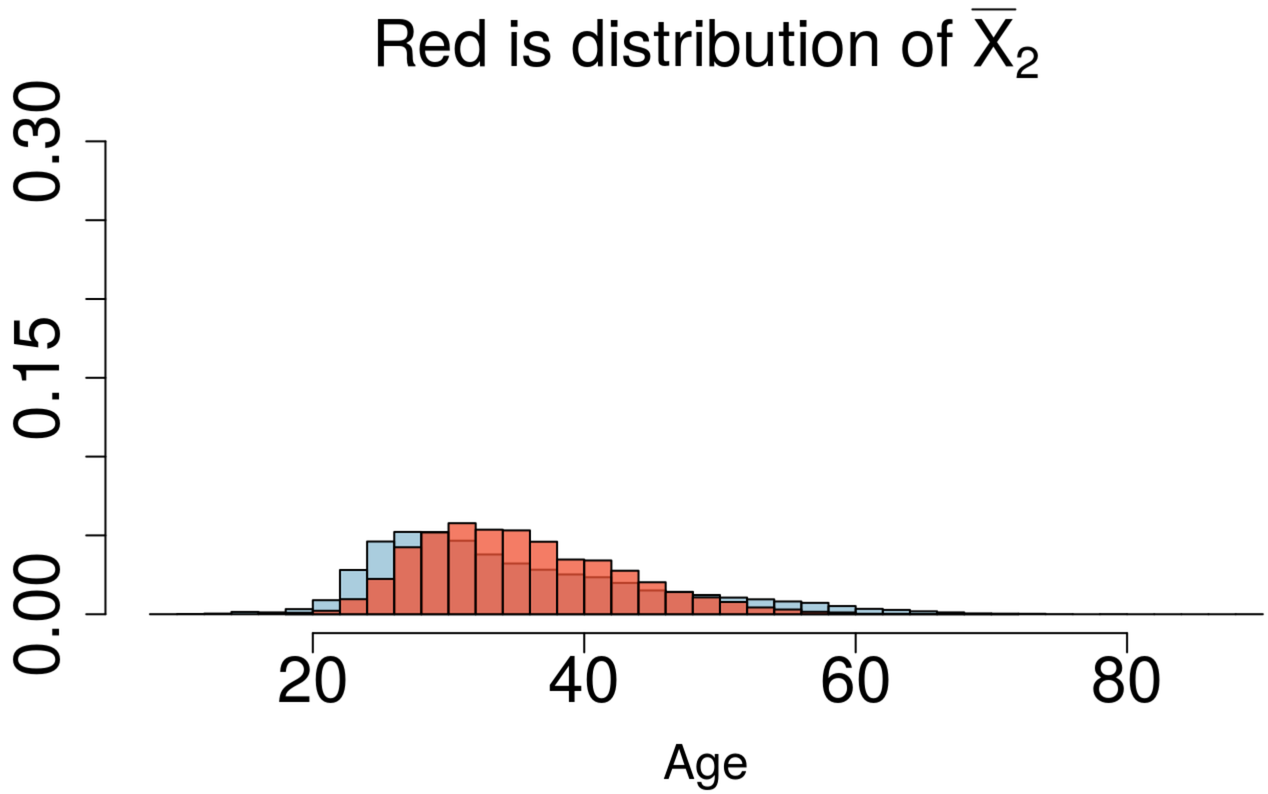
Effect of sample size?



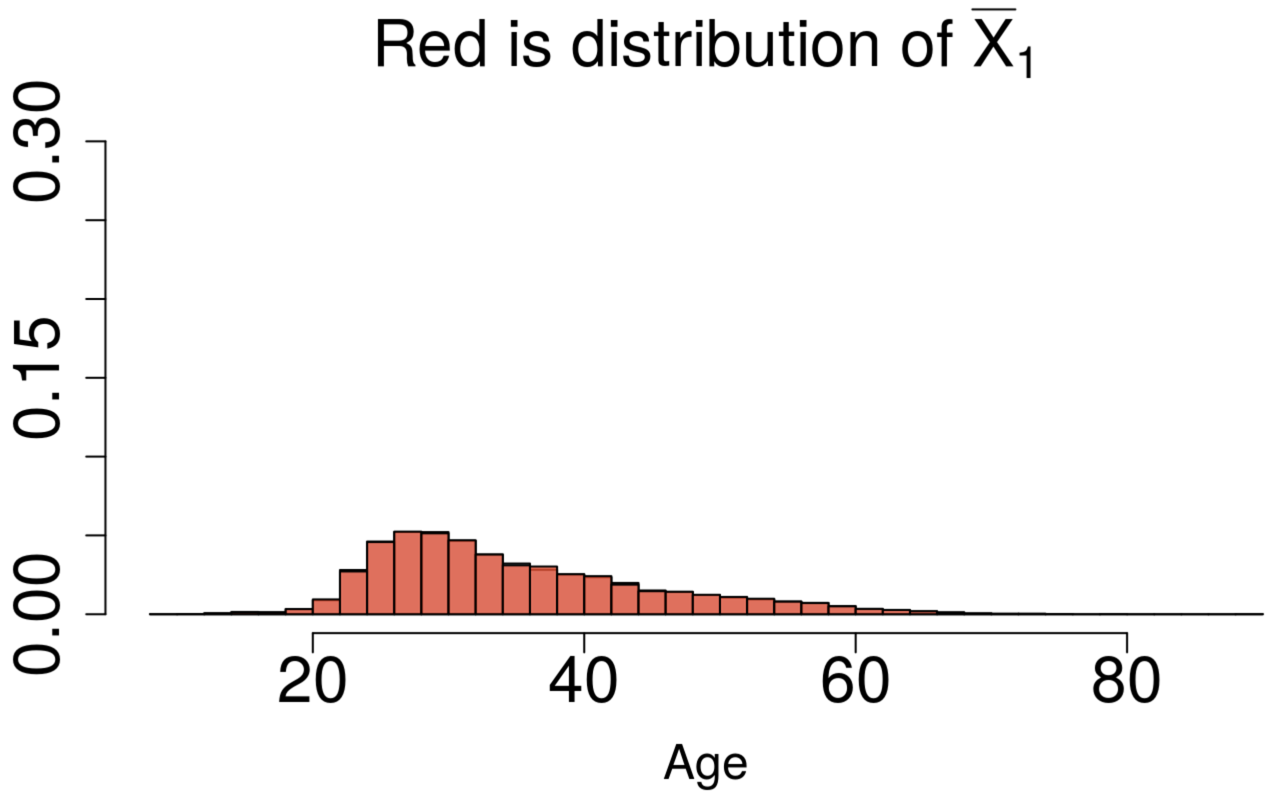
Effect of sample size?



Effect of sample size?



Effect of sample size?



What we've observed

Let $\bar{X}_n$ be the sample mean of a SRS of size n from a population with mean μ and standard deviation σ. Then, $\bar{X}_n$ is a random variable with

$$E(\bar{X}_n) = \mu$$

and

$$\mathrm{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

And what's more…

When n is "large enough", $\bar{X}_n$ is approximately normal!!

$$\bar{X}_n \text{ is approximately } N(\mu, \sigma/\sqrt{n})$$

Central Limit Theorem

If $X_1, \ldots, X_n$ are independent draws from a distribution with mean μ and standard deviation σ, then for large n, the sample mean $\bar{X}_n$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- remarkable since individual X_i 's don't have to look at all like a normal distribution
- how large should n be? Depends, but if distribution of X_i 's is not strongly skewed, say $n \geq 30$

Example

Suppose $X_1, \ldots, X_n$ are independent coin flips, i.e., $X_i \sim \text{Bernoulli}(p)$.

The sample proportion, sometimes written $\hat{p}_n$, is just $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Recall $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$.

CLT tells us

$$\hat{p}_n \approx N(p, \sqrt{p(1-p)/n})$$

Galton Bean Machine

A dramatic video of a beautiful phenomenon

Prelim 1 will cover topics only upto this point, not the next slides

Inference

Sampling distribution: A probabilistic description of how the observed values of a numerical summary statistic (e.g., sample mean) behave under repeated SRS.

This concept underlies all basic statistical inference procedures – its importance cannot be overstated!

In practice: we only collect one sample.

Question: how can we combine the information from a single SRS about a population parameter with our knowledge of sampling distributions in order to perform statistical inference?

Two primary goals

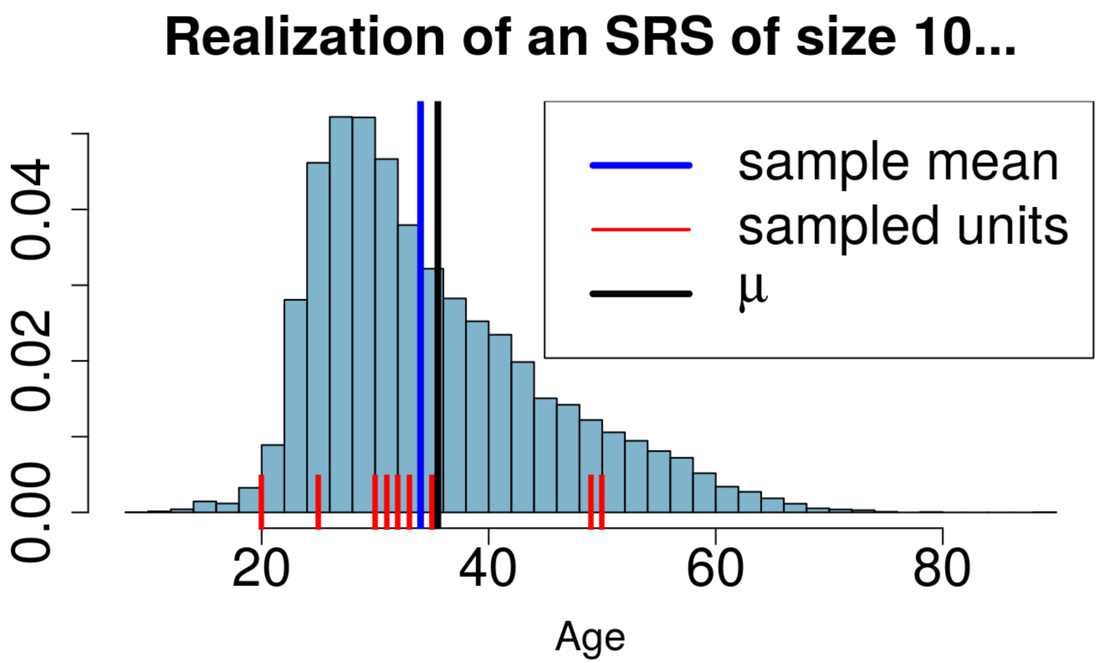
1. A **confidence interval** - a range of plausible values for a (population) parameter, based on the data obtained from our observed sample.

2. A **hypothesis (or significance) test** - an assessment of whether the observed value of a statistic computed using the sample data is consistent with or divergent from some hypothesized value of the (population) parameter.

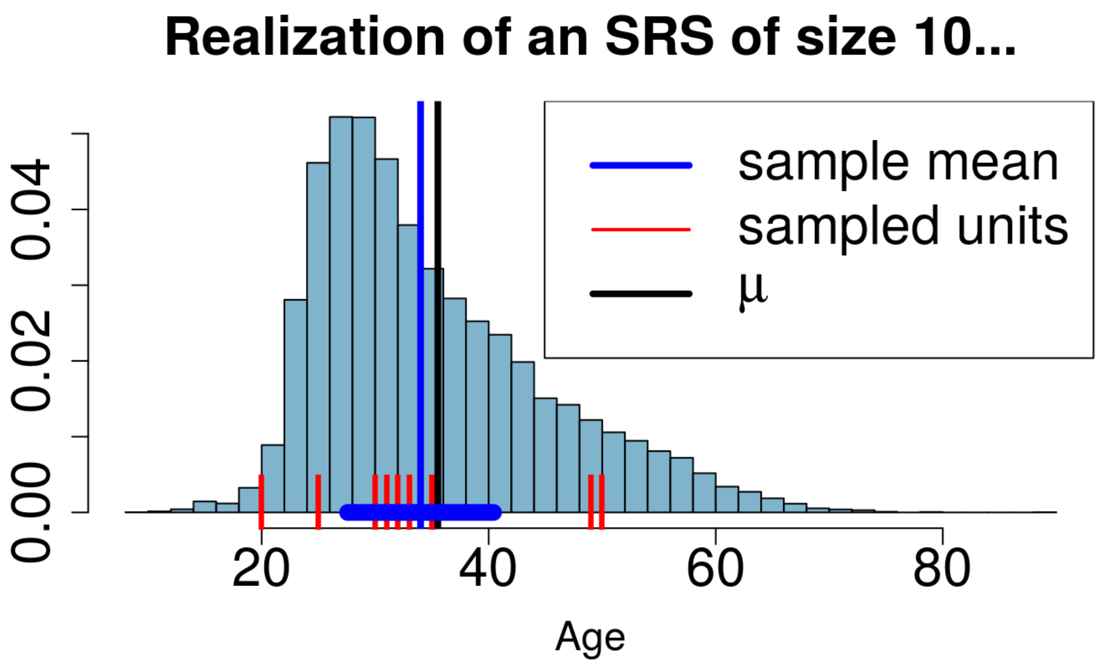
Note: these get at *"what is μ?"* better than just reporting a single **point estimate** (e.g., $\bar{x} = 33.8$)

Confidence Interval (CI)

Point estimate

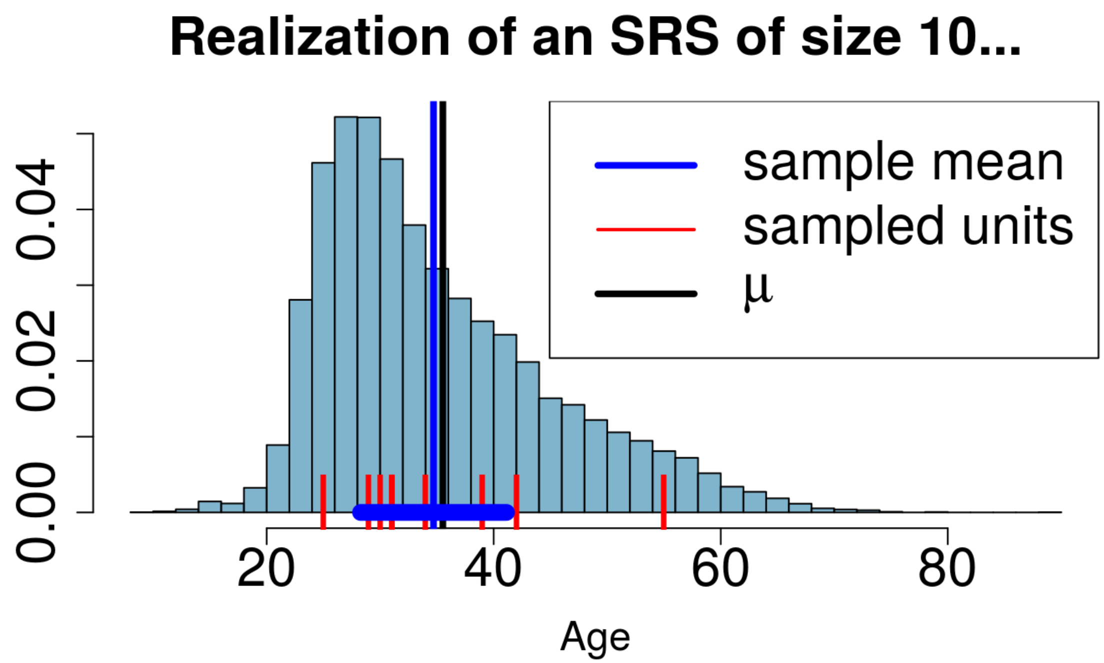


Interval estimate



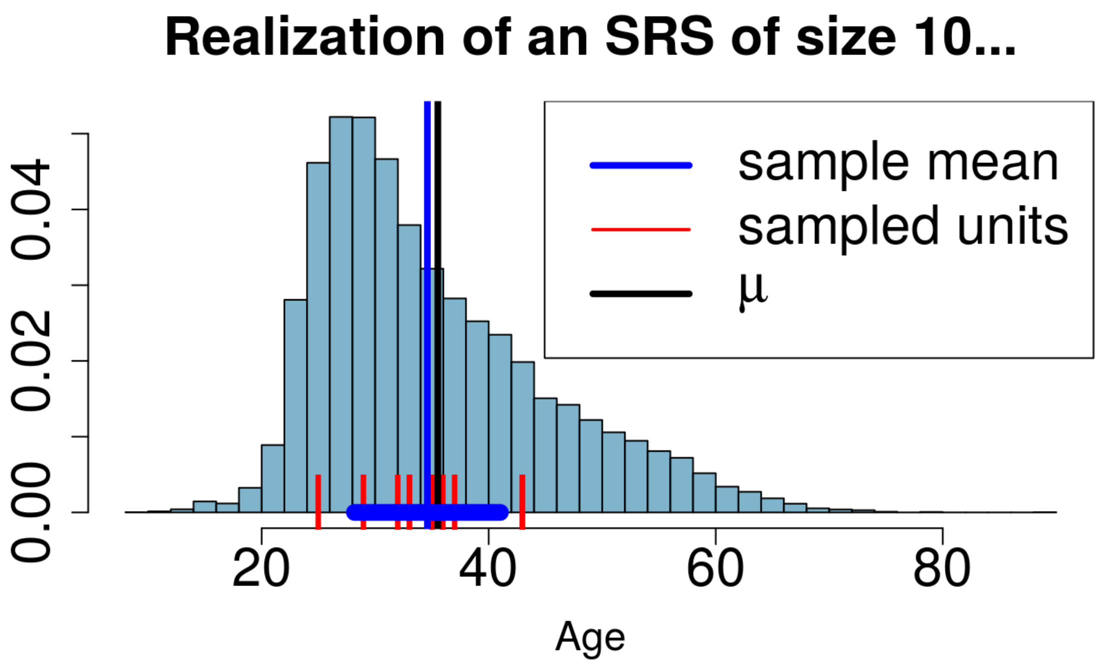
- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random… remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Interval estimate



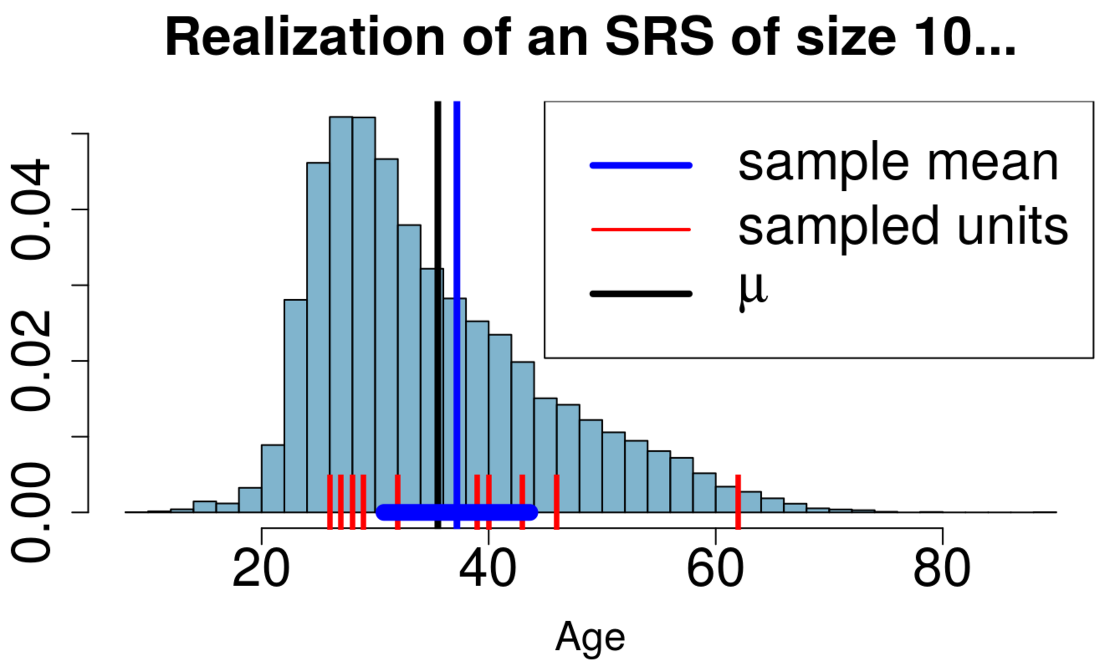
- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random… remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Interval estimate



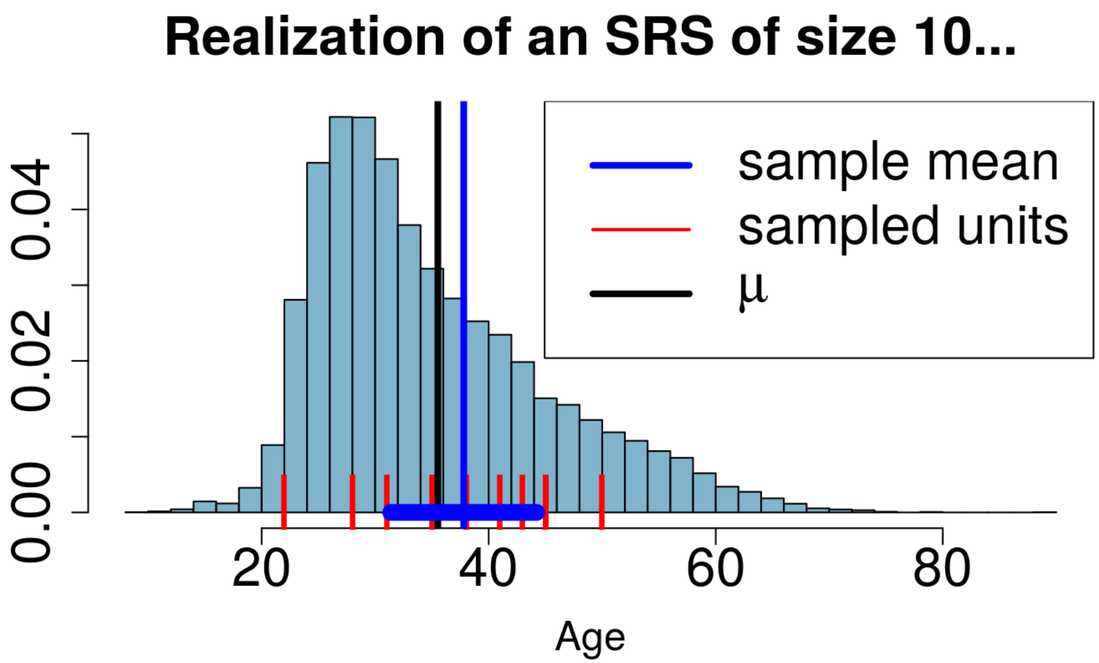
- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random… remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Interval estimate



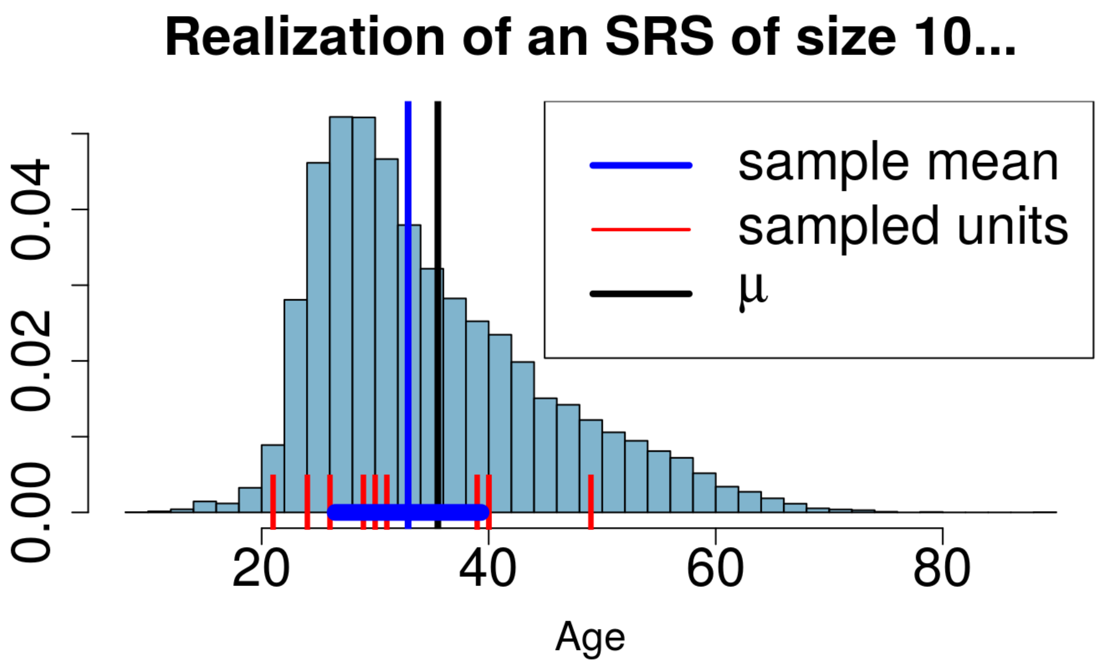
- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random… remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Interval estimate



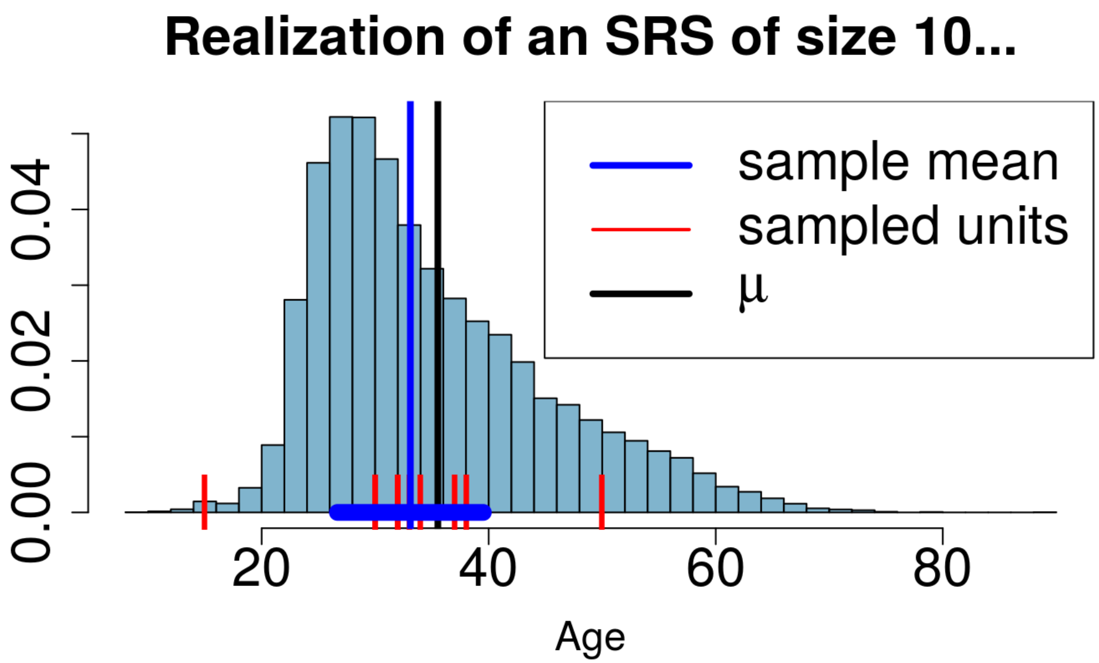
- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random… remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Interval estimate



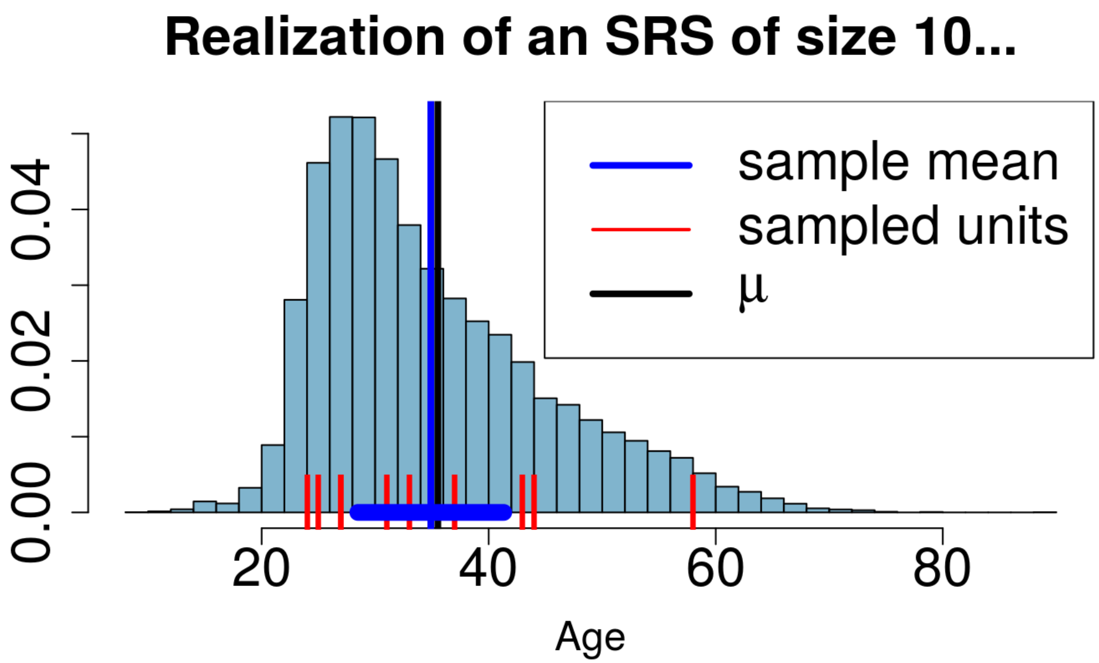
- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random... remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Interval estimate



- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random… remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Interval estimate



- interval is calculated based on sample (centered at $\bar{X}_n$)
- thus it is random... remember we're looking at just one realization
- interval here is $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$

Question

What is the probability that $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$ includes μ?

```
## [1] 0.9583
```

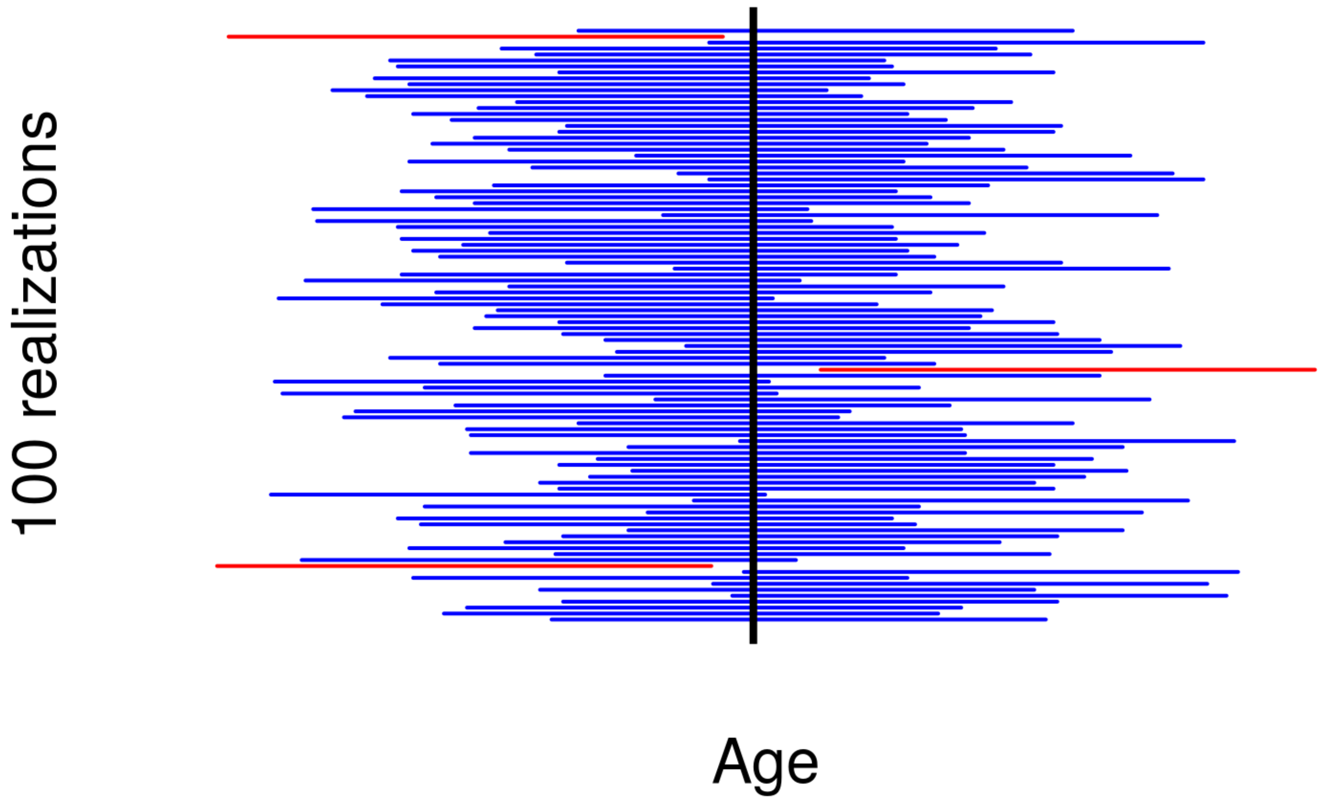
(used Monte Carlo to answer this.)

What happens if interval narrower?

$$P(\bar{X}_n - 2.0 \leq \mu \leq \bar{X}_n + 2.0]) = ?$$

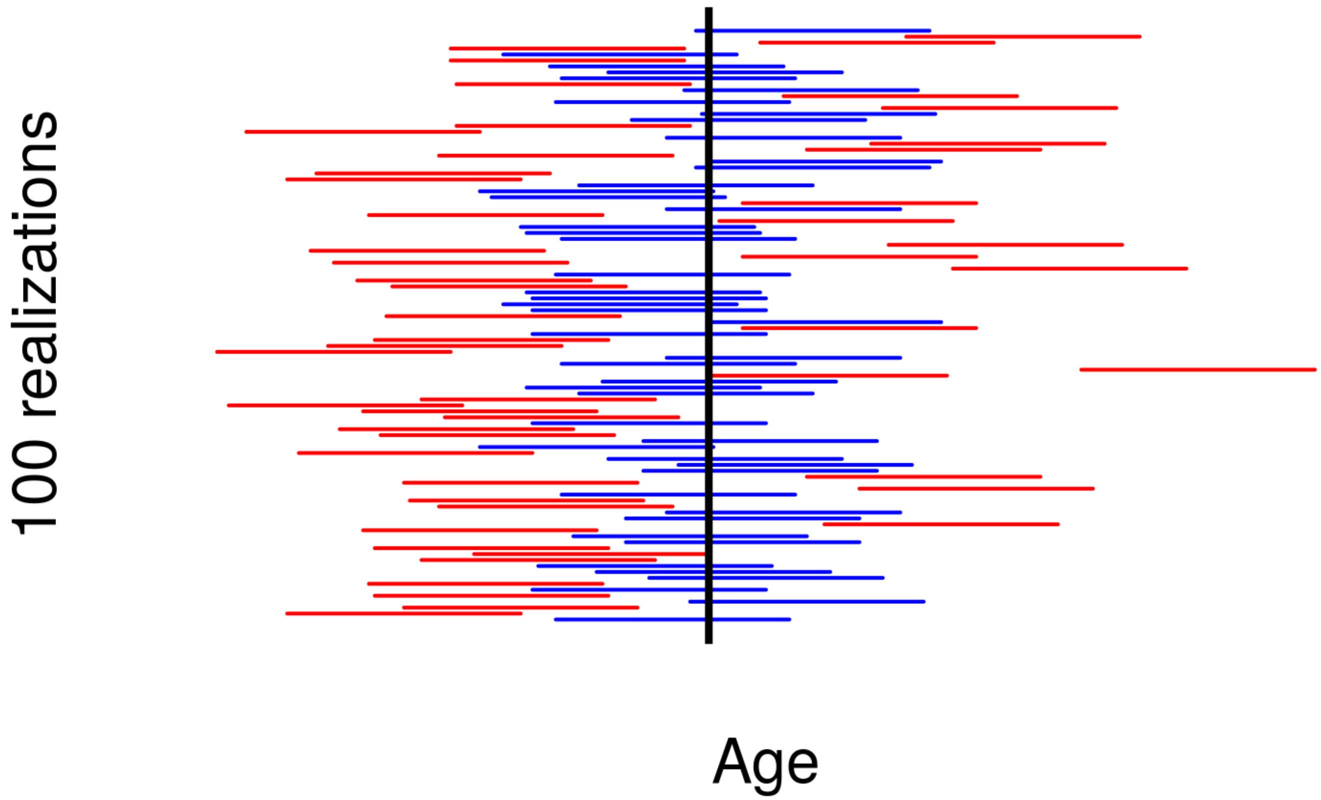
```
## [1] 0.4608
```

Monte Carlo simulation in a picture



- 100 realizations of the random interval $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$
- 97 out of 100 “cover” μ

Monte Carlo simulation in a picture



- 100 realizations of the random interval $[\bar{X}_n - 2.0, \bar{X}_n + 2.0]$
- 48 out of 100 “cover” μ

Confidence Interval

The random interval $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$ is called a 96% **confidence interval**. Here, 96% is said to be its **confidence level**.

$$P([\bar{X}_n - 3.2, \bar{X}_n + 3.2] \text{ includes } \mu) = 96\%$$