DATA ANALYSIS
Year 2019–2020

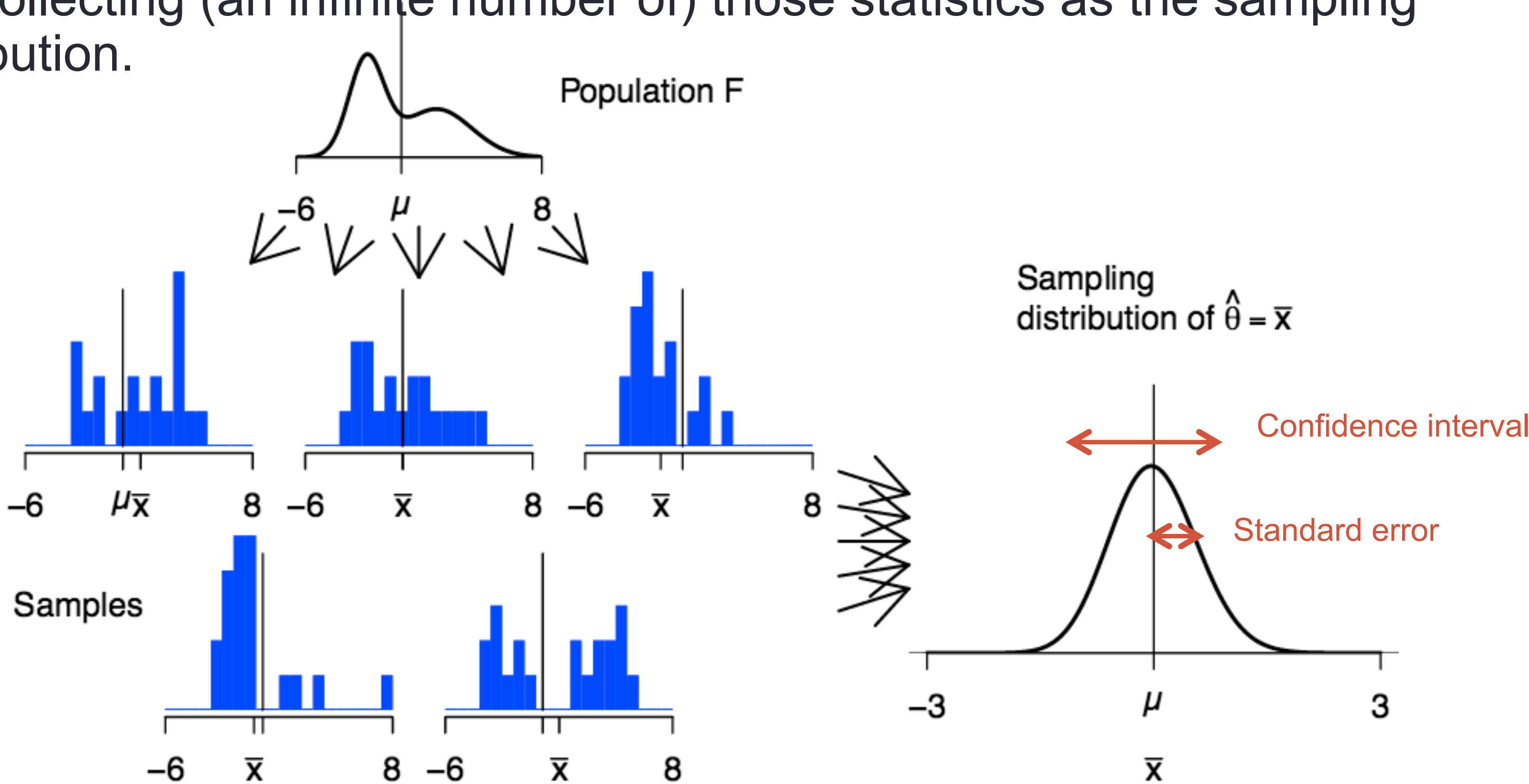
#4 Statistical Methods

Master M1–Marine Physics
Université de Bretagne Occidentale

Jonathan GULA
[gula@univ-brest.fr]

Estimating standard errors and confidence intervals

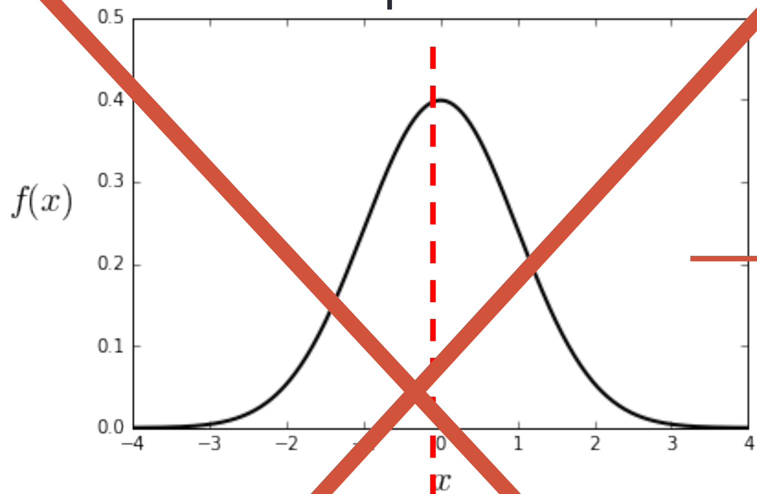
Ideal world: Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution.



Estimating standard errors and confidence intervals

Reality: The problem is that we have only ONE sample

True pdf of X

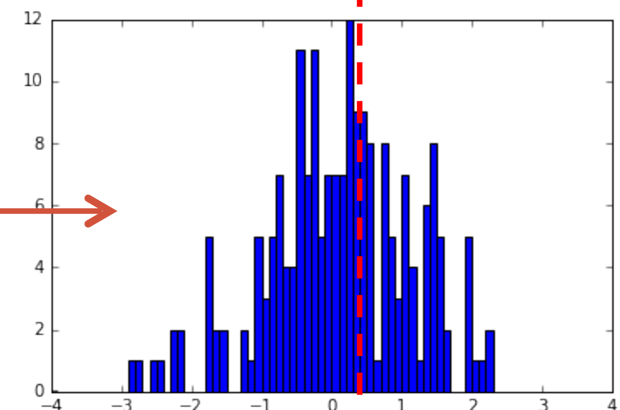


True population mean:

$$\mu = \int x f(x) dx$$

Sample with N values

x_k for $k = 1..N$



Mean estimator (= sample mean)

$$\overline{x} = \frac{1}{N} \sum_k x_k$$

For example if you want to know the mean (works the same for any other statistics), you can only compute one sample mean (a-priori different than the true mean).

Estimating standard errors and confidence intervals

Reality: The problem is that we have only ONE sample

For example if you want to know the mean, you cannot do better than the sample mean. This is the best estimator.

But you also want to be able to quantify "***how far your sample mean is from the true mean?***"

This is what a **standard error** and a **confidence interval** will tell you.

The confidence interval defines the degree of certainty that a given quantity θ (mean or any other statistics) will fall between specified lower and upper bounds θ_L and θ_U :

$$P(\theta_L < \theta < \theta_U) = 1 - \alpha$$

Where α is the level of significance (or confidence level), and $100(1-\alpha)$ is the percent significance level.

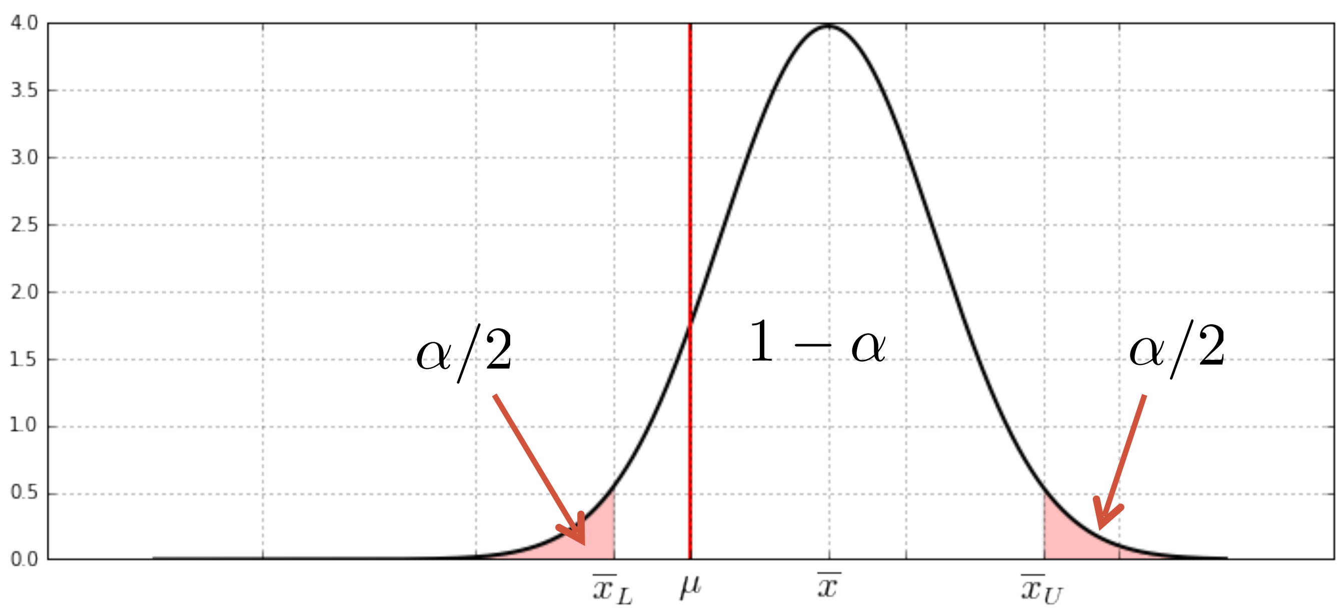
A typical value is $\alpha = 0.05$, which means $100(1-\alpha) = 95\%$, and corresponds to the 95 % confidence interval.

Estimating standard errors and confidence intervals

For example, let's say you compute the sample mean $\overline{x}$

The confidence interval will correspond to the lower and upper bounds, $\overline{x}_L, \overline{x}_U$ such that:

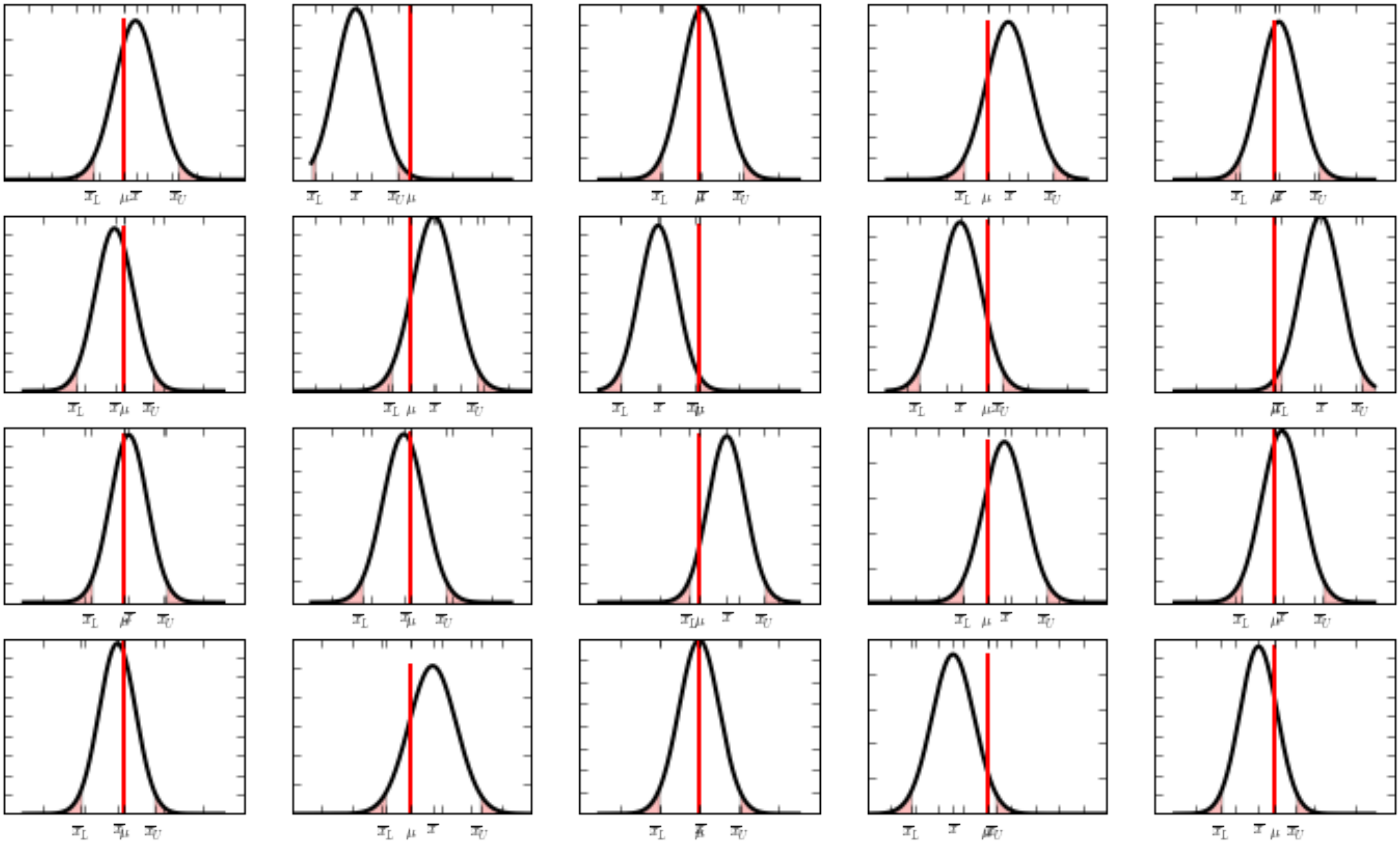
$$P(\overline{x}_L < \mu < \overline{x}_U) = 1 - \alpha$$



Which means that the interval $(\overline{x}_L, \overline{x}_U)$ has a probability $1 - \alpha$ of containing the true mean μ.

In other words = *we are 95% confident that the true value of the parameter is in the confidence interval…*

Estimating standard errors and confidence intervals



If you get Ns=100 samples and construct Ns=100 confidence intervals at 95%, the true mean will be included in 95 confidence intervals out of the 100 (on average)

Estimating standard errors and confidence intervals

Question:
How do we compute the standard error and confidence interval (CI)?

There are different ways of doing that, we'll talk about 2 common ways in oceanography:

1. **Using the Central Limit Theorem** (for the mean only)

2. **Resampling methods** (bootstrap and Jackknife methods, for any statistics)

1. CI for the mean (using the CLT)

Central Limit Theorem:

Let $X_i, i = 1..\mathrm{Ns}$ be a sequence of independent random variables (each containing N values) drawn from distributions with mean μ and variance σ^2 . Then as *Ns* becomes large, the distribution of the mean values X_i of each sample $\hat{\mu}_i$ approaches the normal distribution with mean μ and variance σ^2/N .

$$\hat{\mu}(x) \sim \mathcal{N}(\mu, \sigma/\sqrt{N})$$

1. CI for the mean (using the CLT)

Central Limit Theorem:

Let $X_i, i = 1..\mathrm{Ns}$ be a sequence of independent random variables (each containing N values) drawn from distributions with mean μ and variance σ^2. Then as *Ns* becomes large, the distribution of the mean values X_i of each sample $\hat{\mu}_i$ approaches the normal distribution with mean μ and variance σ^2/N.

$$\hat{\mu}(x) \sim \mathcal{N}(\mu, \sigma/\sqrt{N})$$ = standard error

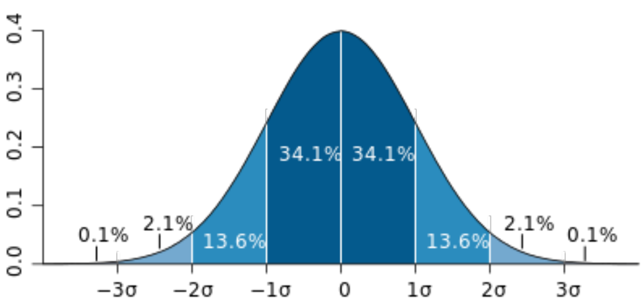
1. CI for the mean (using the CLT)

So the standard error is for the mean is $\dfrac{\sigma}{\sqrt{N}}$

And the $100(1-\alpha)$ percent confidence interval for the population mean is given by:

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Where $z_{\alpha/2}$ are the values giving the $100(1-\alpha)$ percent confidence interval for a standard normal distribution. The $z_{\alpha/2}$ can be recomputed using the theoretical normal distribution function or they are directly read on a table



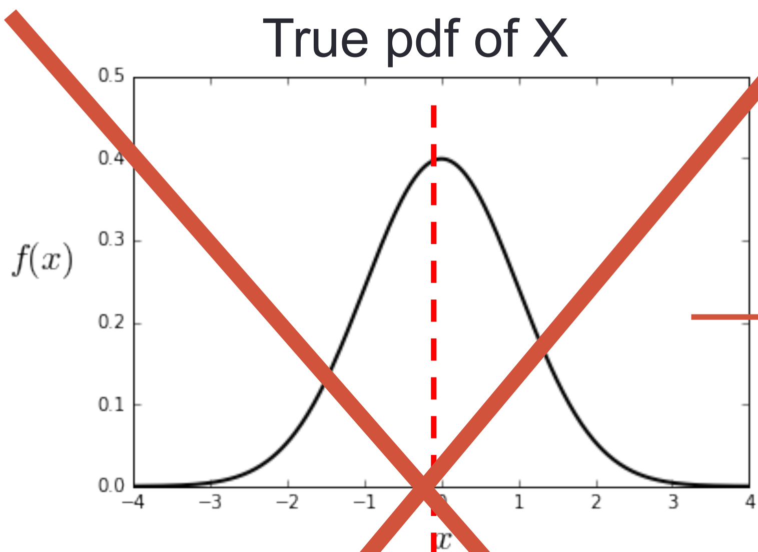
$100(1-\alpha)$	$z_{\alpha/2}$
99%	2.576
98%	2.326
95%	1.96
90%	1.645

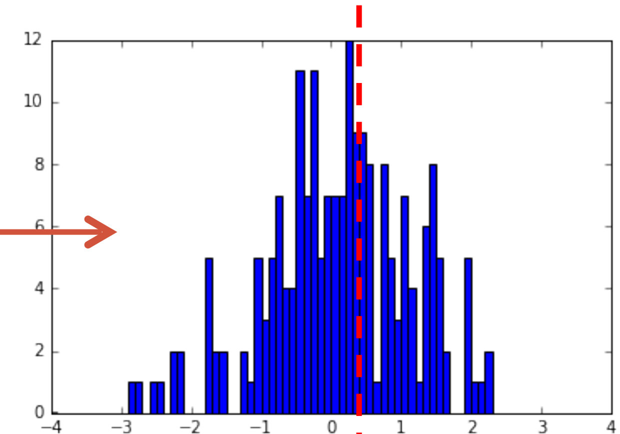
1. CI for the mean (using the CLT)

Problem: Most of the time we don't know the true standard deviation σ !!!
We only know the standard deviation estimated from of our sample = s.

Sample with N values

$$x_k \text{ for } k = 1..N$$





$$\mu = \int x f(x) dx$$

$$\sigma^2 = \int (x - \mu)^2 f(x)\, dx$$

$$\overline{x} = \frac{1}{N} \sum_k x_k$$

$$\boxed{s^2 = \frac{1}{N-1} \sum_k (x_k - \hat{\mu})^2}$$

1. CI for the mean (using the CLT)

If we don't know the true standard deviation we can estimate the standard error of the mean using: $\frac{s}{\sqrt{N}}$

(Where σ has been replaced by the sample standard deviation s)

And the $100(1-\alpha)$ percent confidence interval for the population mean is given by:

$$\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{N}} < \mu < \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{N}}$$

(Where σ has been replaced by the sample standard deviation s and the $z_{\alpha/2}$ have been replaced by $t_{\alpha/2}$)

Where $t_{\alpha/2}$ are the values giving the $100(1-\alpha)$ percent confidence interval for a student's t-distribution with $N-1$ degrees of freedom.

1. CI for the mean (using the CLT)

The $t_{\alpha/2}$ can be recomputed but most of the time they are directly read on a table

(given the number of degrees of freedom df = N-1 and the desired confidence interval)

Note that a student's t-distribution converge toward a normal distribution when N becomes large.

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

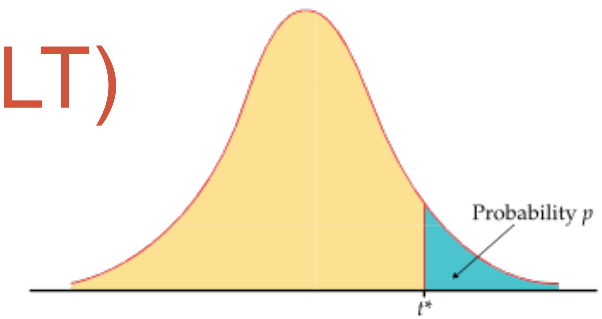


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

1. CI for the mean (using the CLT)

Summary

a. σ is know

The standard error is $\frac{\sigma}{\sqrt{N}}$

the $100(1-\alpha)$ percent confidence interval for the population mean is given by:

$$\overline{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

b. σ is not know

The standard error is $\frac{s}{\sqrt{N}}$

the $100(1-\alpha)$ percent confidence interval for the population mean is given by:

$$\overline{x} \pm t_{\alpha/2} \frac{s}{\sqrt{N}}$$

Estimating standard errors and confidence intervals

Question:
How do we compute the standard error and confidence interval (CI)?

There are different ways of doing that, we'll talk about 2 common ways in oceanography:

1. **Using the Central Limit Theorem** (for the mean only)

2. **Resampling methods** (bootstrap and Jackknife methods, for any statistics)

2. Bootstrapping method

The CLT is very convenient to estimate the standard error of the mean.

But in a more general case we want to compute different statistics (median, standard deviation, kurtosis, percentiles, …) and know how much they will vary from one sample to another (standards errors and confidence intervals.)

We can use generic methods known as **resampling methods** such as jackknifing and **bootstrapping.**

2. Bootstrapping method

Bootstrapping is a very simple and efficient method to estimate the sampling distribution of almost any statistic using random sampling methods

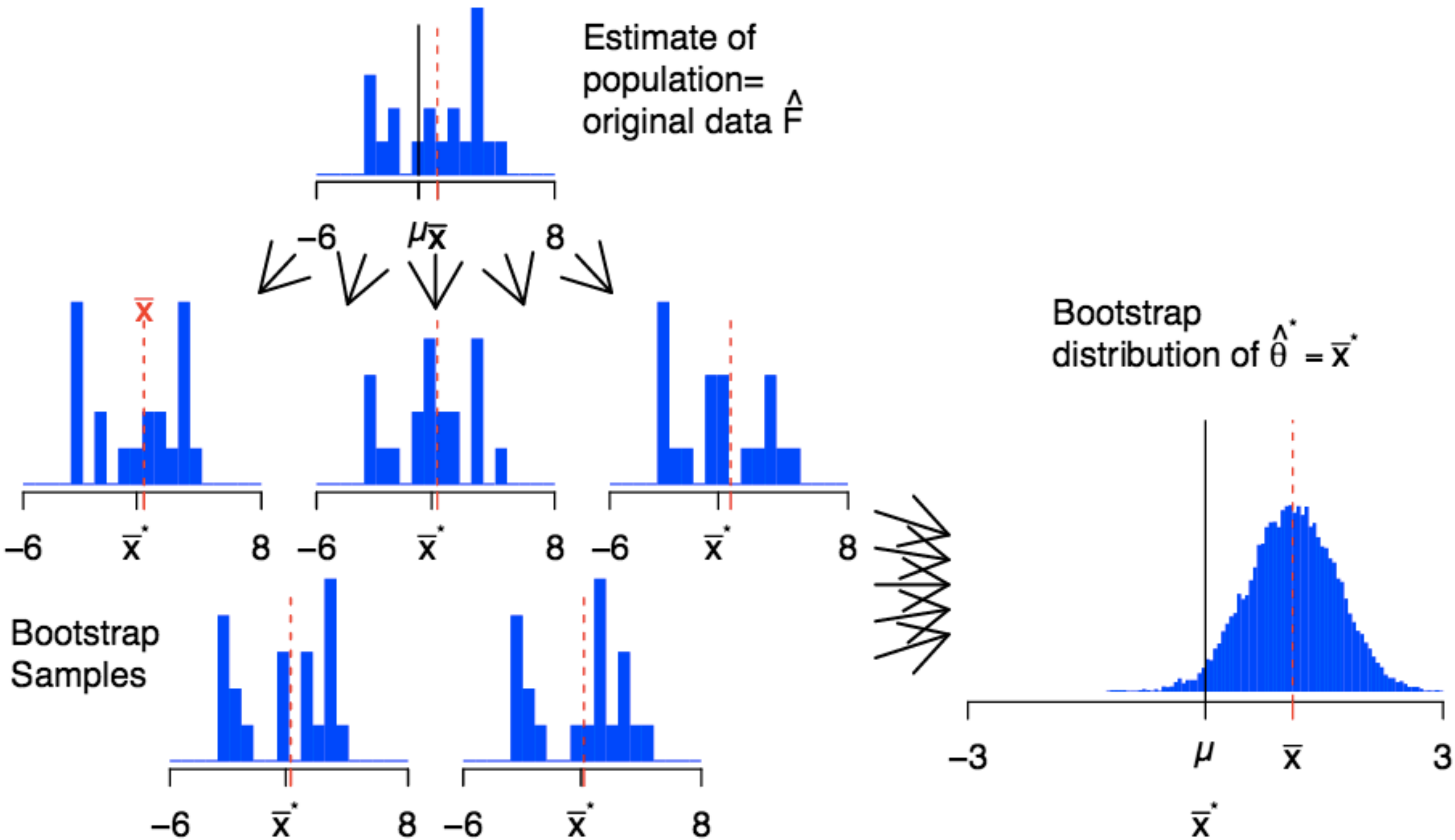
The idea is to:

- Treat the sample as the true population.
- **Sample with replacement** your actual distribution *M* times.
- Compute the statistic of interest on each “re-sample”.

The name bootstrap comes from the idea of “lifting yourself up by your own bootstraps" [the loop at the top of tall boots].

2. Bootstrapping method

- Bootstrap world. The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics.

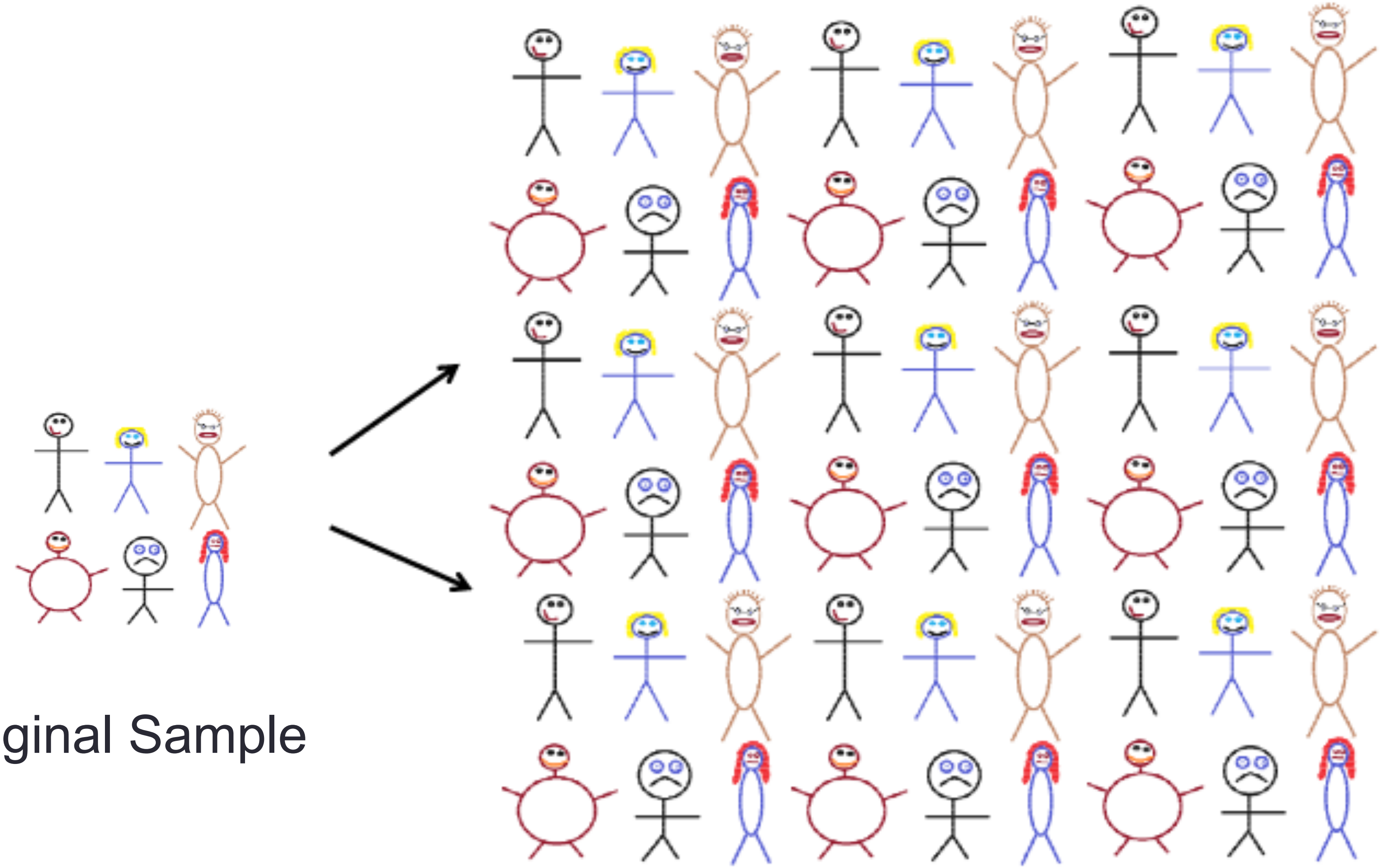


2. Bootstrapping method

Sampling with replacement:

Suppose we have a random sample of 6 people

2. Bootstrapping method



Original Sample

A simulated “population” to sample from

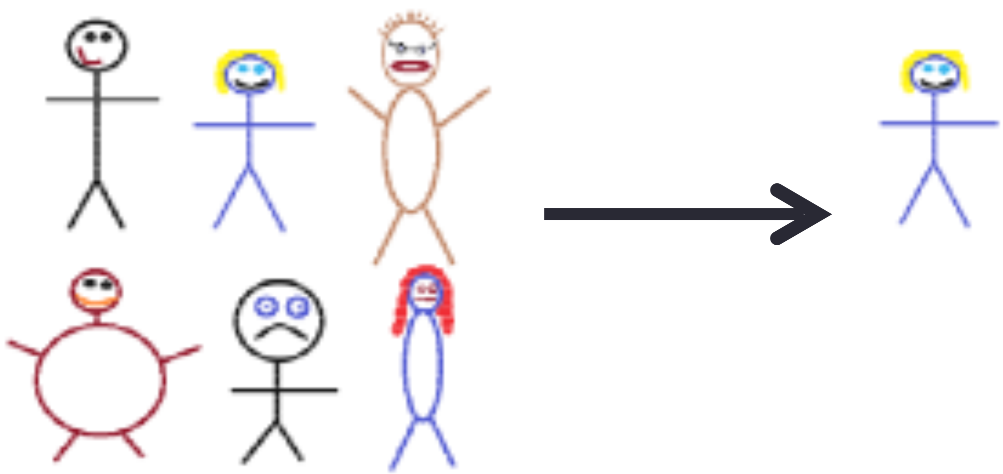
2. Bootstrapping method

Sampling with replacement:

Imagine putting papers in a hat. If you **sample with replacement**, you would choose one paper, put the paper back in the hat, and then choose another paper.

2. Bootstrapping method

Sampling with replacement:

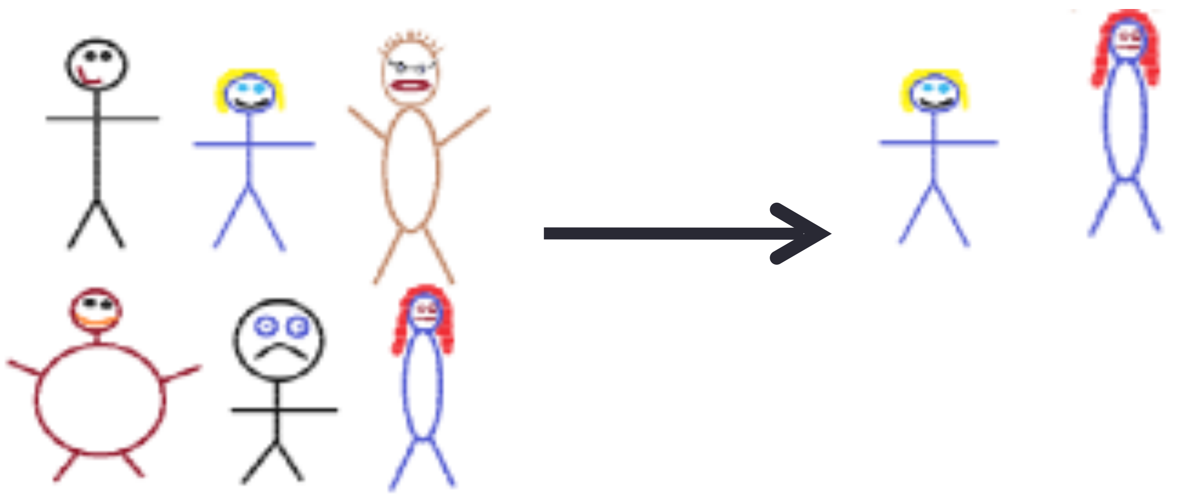


Original Sample   Bootstrap Sample

A bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Bootstrapping method

Sampling with replacement:

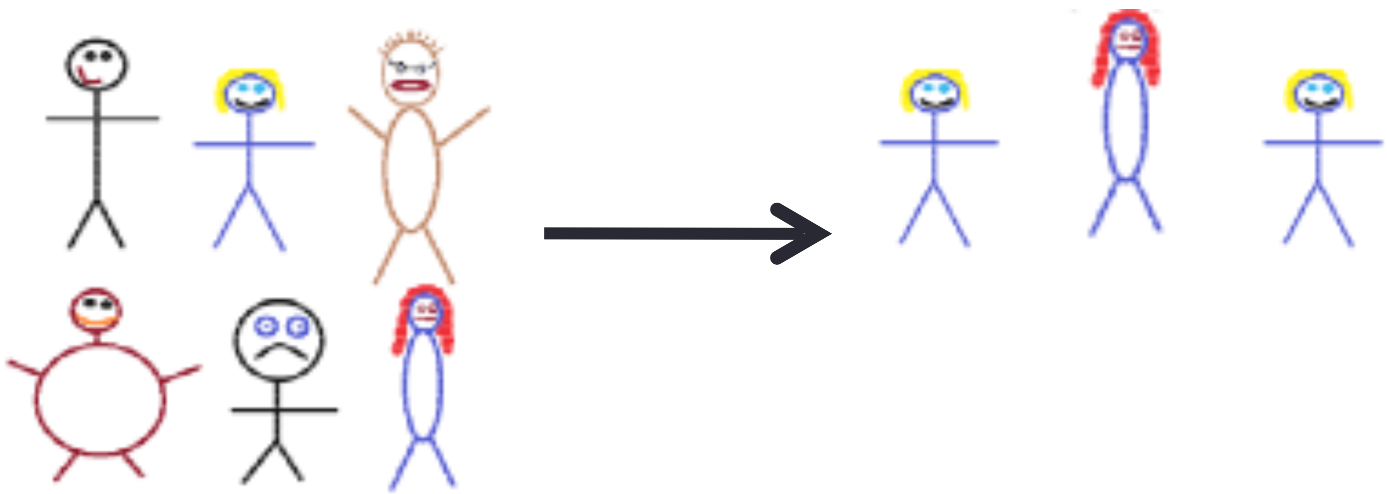


Original Sample Bootstrap Sample

A bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Bootstrapping method

Sampling with replacement:

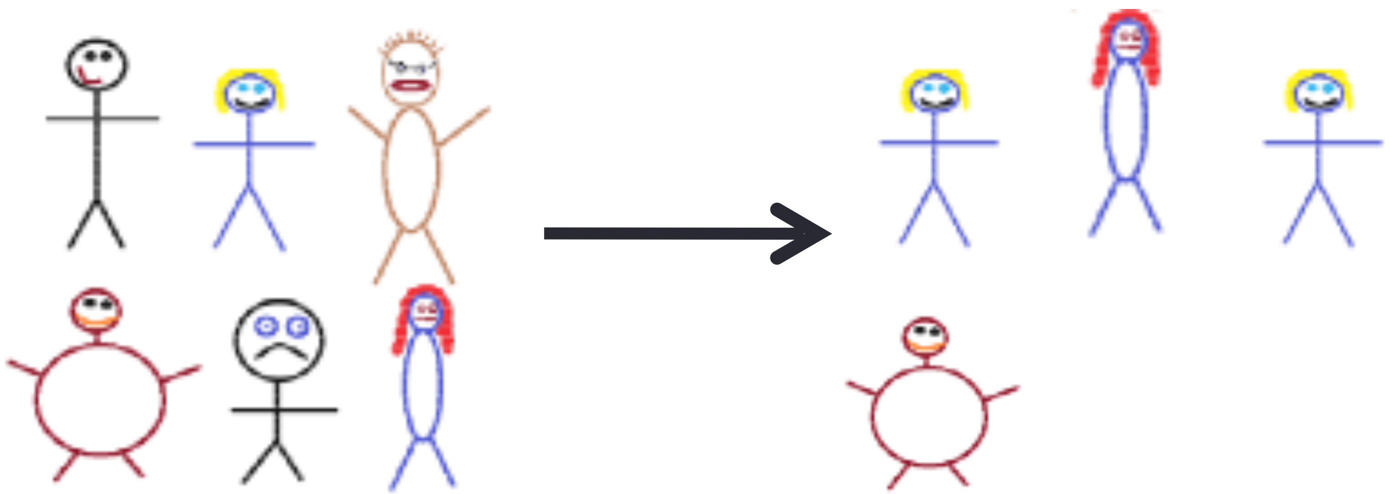


Original Sample Bootstrap Sample

A bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Bootstrapping method

Sampling with replacement:

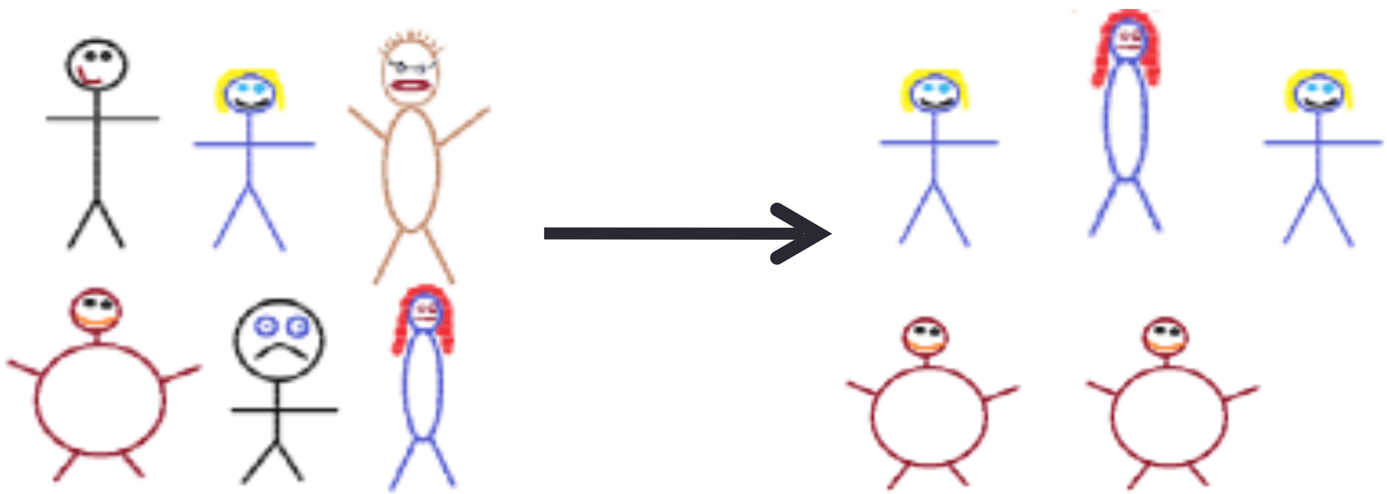


Original Sample Bootstrap Sample

A bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Bootstrapping method

Sampling with replacement:

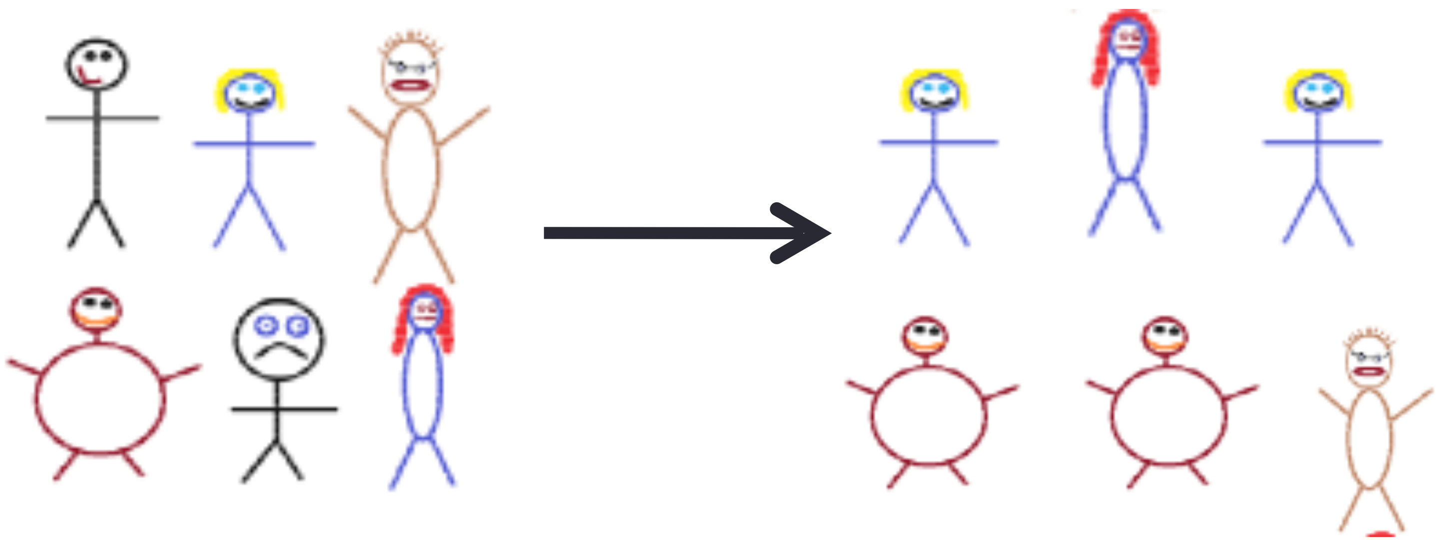


Original Sample Bootstrap Sample

A bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Bootstrapping method

Sampling with replacement:



Original Sample

Bootstrap Sample

A bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Bootstrapping method

- A ***bootstrap statistic*** is the statistic computed on a bootstrap sample
- A ***bootstrap distribution*** is the distribution of many bootstrap statistics

- The variability of the bootstrap statistics is similar to the variability of the sample statistics
- The standard error of a statistic can be estimated using the standard deviation of the bootstrap distribution!
- Confidence intervals can be created using the standard error or the percentiles of a bootstrap distribution

2. Bootstrapping method

Number of Bootstrap Samples

- When using bootstrapping, you may get a slightly different confidence interval each time. This is fine!
- The more bootstrap samples you use, the more precise your answer will be.
- Increasing the number of bootstrap samples will not change the standard error or interval
- In real life, you probably want to take 10,000 or even 100,000 bootstrap samples

Estimating standard errors and confidence intervals

- See TD4 – Confidence intervals