Scientific Computing – Statistics

Fabian Sinz
Dept. Neuroethology, University Tübingen
Bernstein Center Tübingen

10/21/2014

information

- Samuels, M. L., Wittmer, J. A., & Schaffner, A. A. (2010). Statistics for the Life Sciences (4th ed., p. 668). Prentice Hall.
- Zar, J. H. (1999). Biostatistical Analysis. (D. Lynch, Ed.)Prentice Hall New Jersey (4th ed., Vol. 4th, p. 663). Prentice Hall. doi:10.1037/0012764
- `http://stats.stackexchange.com`

Day 2 – errorbars, confidence intervals, and tests

Examples

- Before new drugs are given to human subjects, it is common practice to first test them in dogs or other animals. In part of one study, a new investigational drug was given to eight male and eight female dogs at doses of 8 mg/kg and 25 mg/kg. Within each sex, the two doses were assigned at random to the eight dogs. Many "endpoints" were measured, such as cholesterol, sodium, glucose, and so on, from blood samples, in order to screen for toxicity problems in the dogs before starting studies on humans. One endpoint was alkaline phosphatase level (or APL, measured in U/l). For females, the effect of increasing the dose from 8 to 25 mg/kg was positive, although small (the average APL increased from 133.5 to 143 U/l), but for males the effect of increasing the dose from 8 to 25 mg/kg was negative.

Examples

- Before new drugs are given to human subjects, it is common practice to first test them in dogs or other animals. In part of one study, a new investigational drug was given to eight male and eight female dogs at doses of 8 mg/kg and 25 mg/kg. Within each sex, the two doses were assigned at random to the eight dogs. Many "endpoints" were measured, such as cholesterol, sodium, glucose, and so on, from blood samples, in order to screen for toxicity problems in the dogs before starting studies on humans. One endpoint was alkaline phosphatase level (or APL, measured in U/l). For females, the effect of increasing the dose from 8 to 25 mg/kg was positive, although small (the average APL increased from 133.5 to 143 U/l), but for males the effect of increasing the dose from 8 to 25 mg/kg was negative.
- On 15 July 1911, 65-year-old Mrs. Jane Decker was struck by lightning while in her house. She had been deaf since birth, but after being struck, she recovered her hearing, which led to a headline in the New York Times, "Lightning Cures Deafness."

Examples

- Before new drugs are given to human subjects, it is common practice to first test them in dogs or other animals. In part of one study, a new investigational drug was given to eight male and eight female dogs at doses of 8 mg/kg and 25 mg/kg. Within each sex, the two doses were assigned at random to the eight dogs. Many "endpoints" were measured, such as cholesterol, sodium, glucose, and so on, from blood samples, in order to screen for toxicity problems in the dogs before starting studies on humans. One endpoint was alkaline phosphatase level (or APL, measured in U/l). For females, the effect of increasing the dose from 8 to 25 mg/kg was positive, although small (the average APL increased from 133.5 to 143 U/l), but for males the effect of increasing the dose from 8 to 25 mg/kg was negative.
- On 15 July 1911, 65-year-old Mrs. Jane Decker was struck by lightning while in her house. She had been deaf since birth, but after being struck, she recovered her hearing, which led to a headline in the New York Times, "Lightning Cures Deafness."
- Some research has suggested that there is a genetic basis for sexual orientation. One such study involved measuring the midsagittal area of the anterior commissure (AC) of the brain for 30 homosexual men, 30 heterosexual men, and 30 heterosexual women. The researchers found that the AC tends to be larger in heterosexual women than in heterosexual men and that it is even larger in homosexual men.

Samuels, Wittmer, Schaffner 2010

types of evidence

experiment
is better than
observational study
is better than
anecdotal evidence

Day 2 – errorbars, confidence intervals, and tests

sources of error in an experiment

Think about it for 2 min

If you repeat a scientific experiment, why do you not get the same result every time you repeat it?

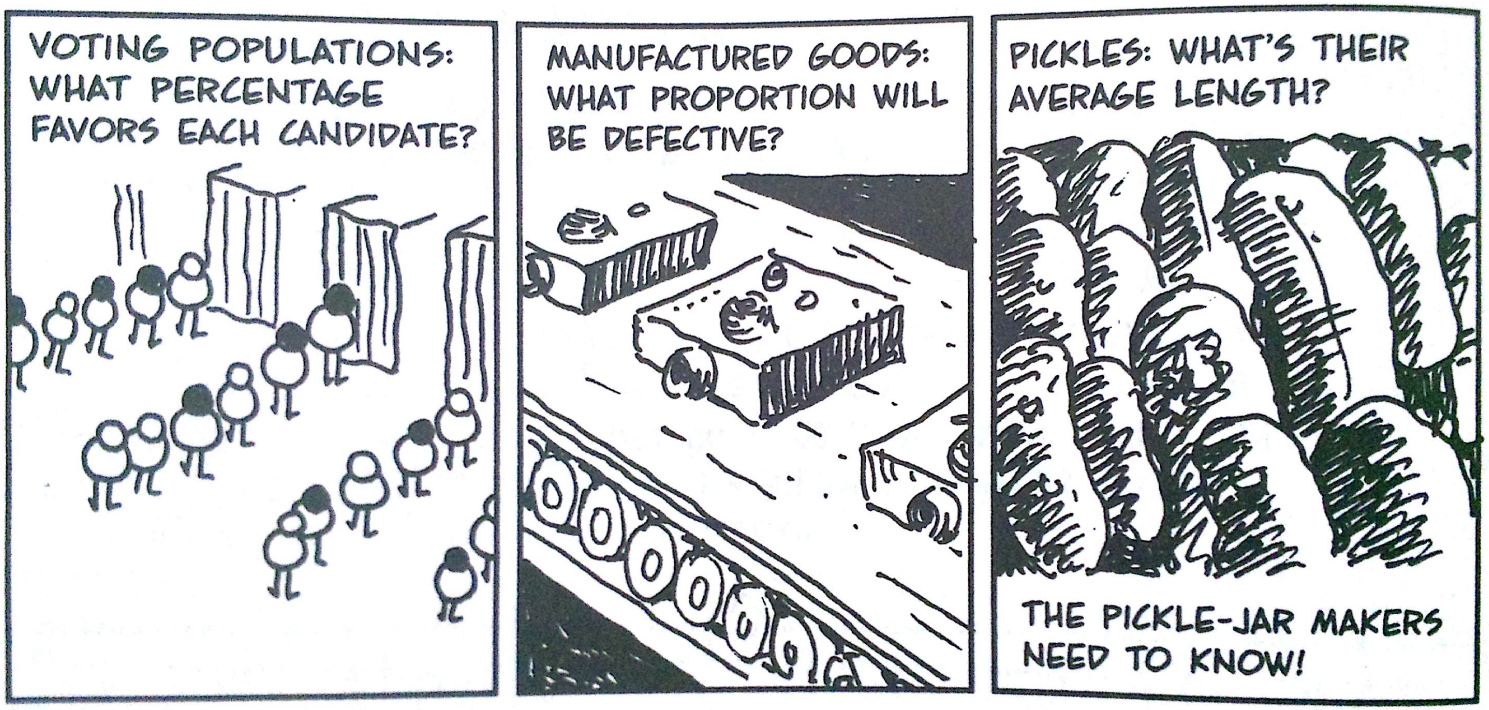
sources of error in an experiment

Think about it for 2 min

If you repeat a scientific experiment, why do you not get the same result every time you repeat it?

- sampling error (a finite subset of the population of interest is selected in each experiment)
- nonsampling errors (e.g. noise, uncontrolled factors)

statisticians are lazy

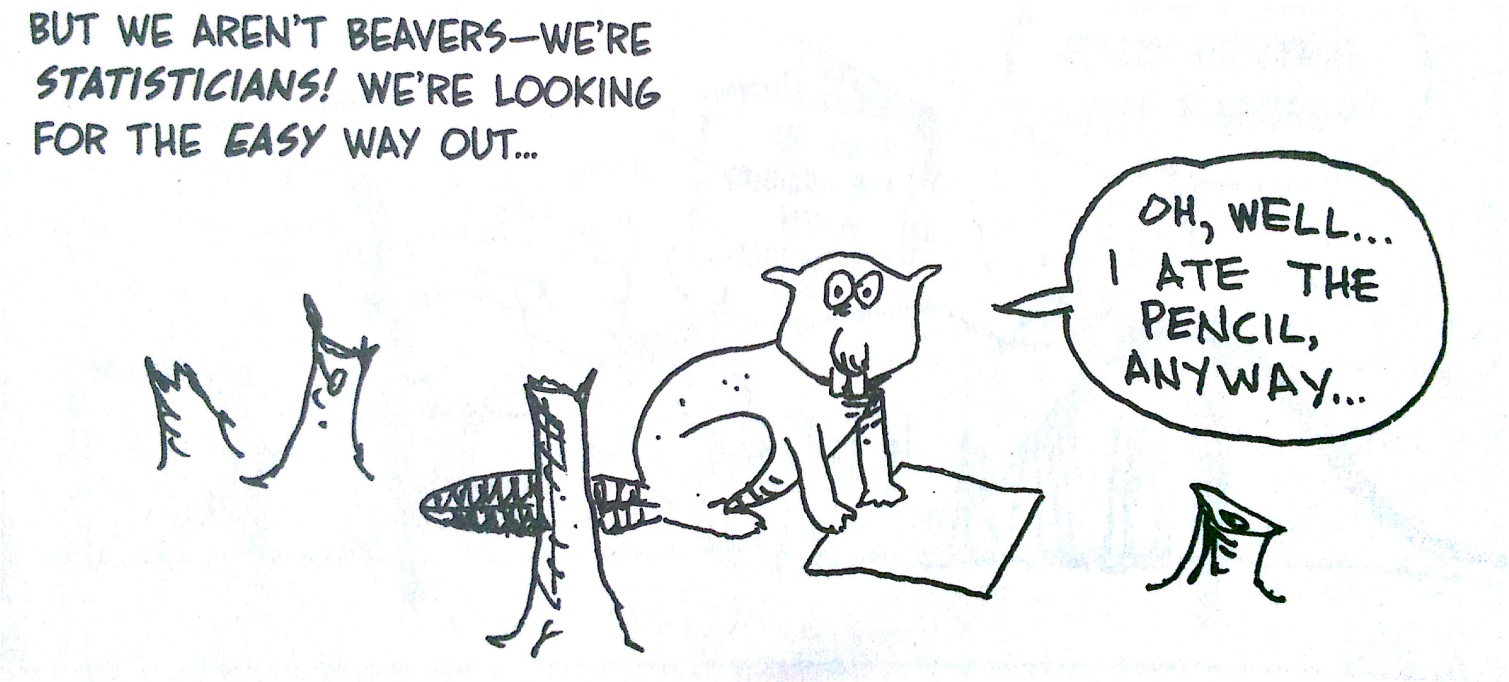


Larry Gonick, The Cartoon Guide to Statistics

statisticians are lazy

Larry Gonick, The Cartoon Guide to Statistics

statisticians are lazy

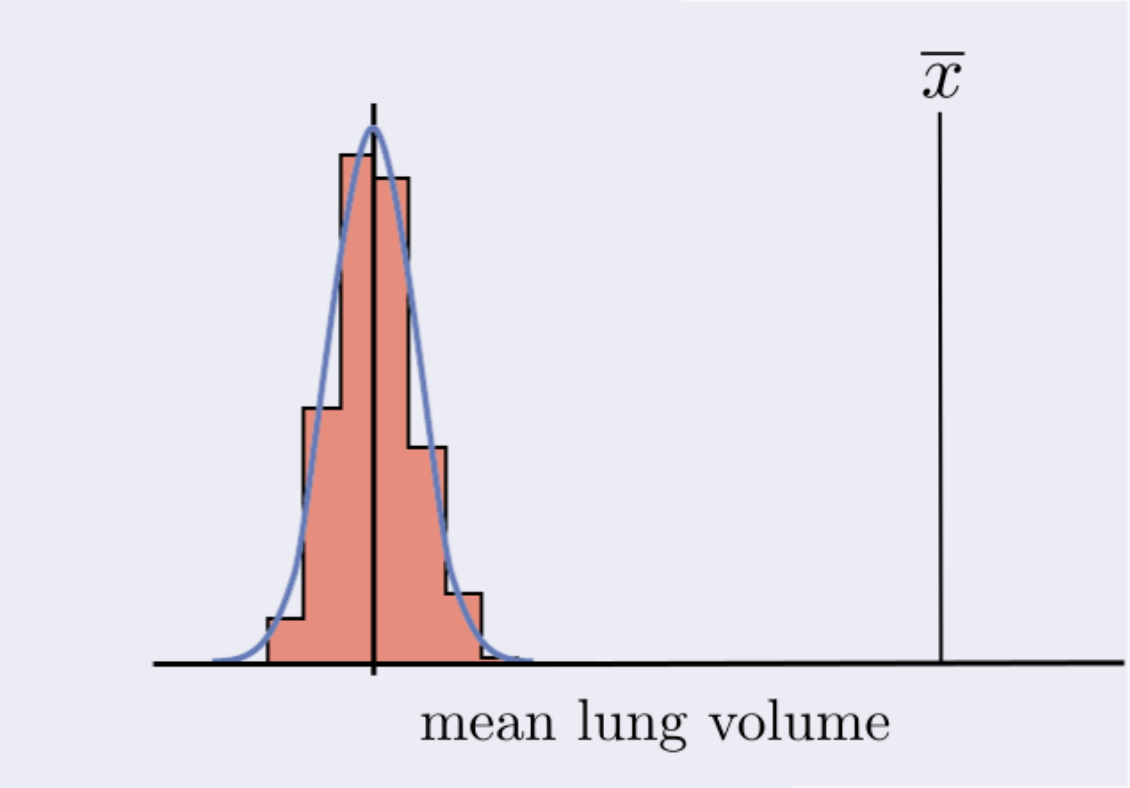


Larry Gonick, The Cartoon Guide to Statistics

illustrating examples

lung volume of smokers

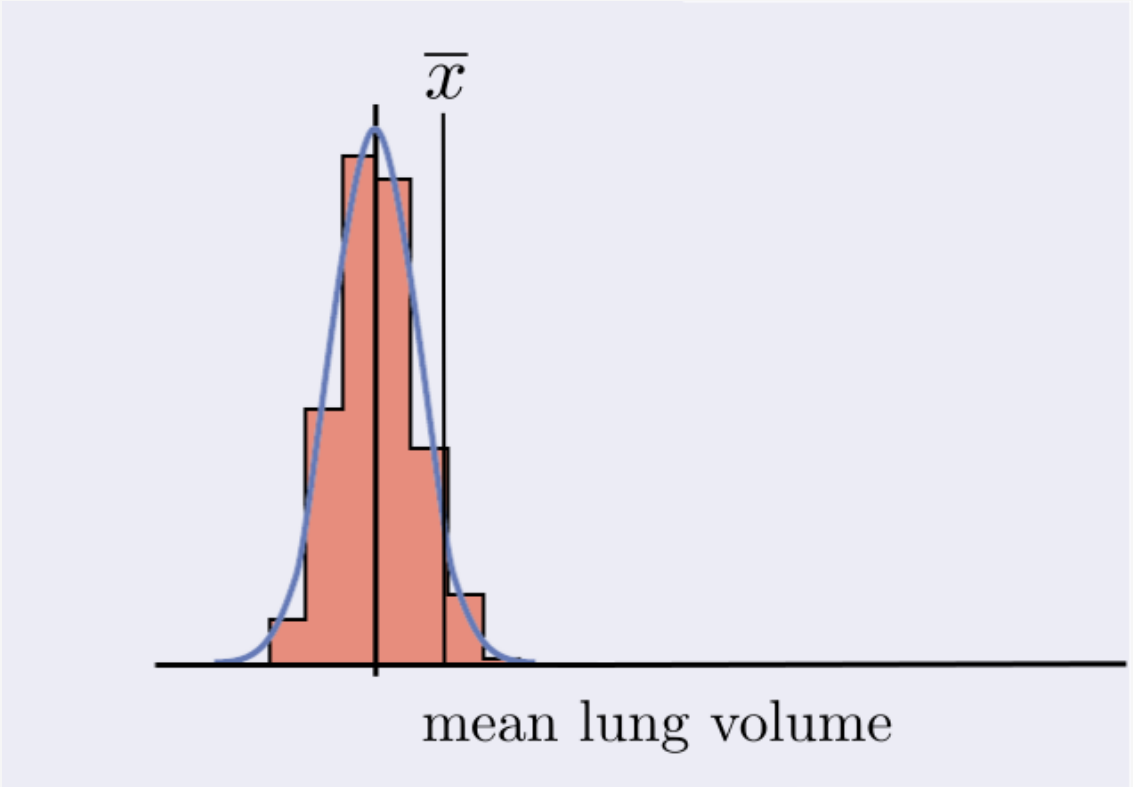
Assume you know the sampling distribution of the mean lung volume of smokers. Would you believe that the sample came from a group of smokers?



illustrating examples

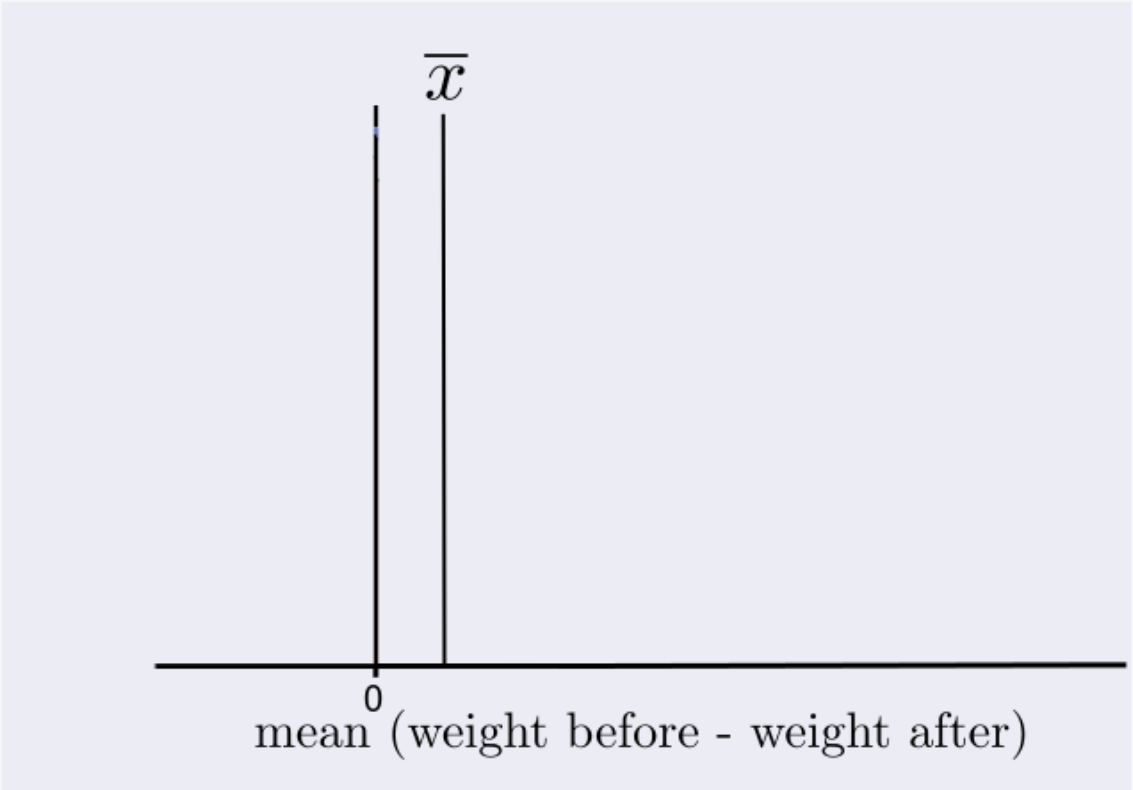
lung volume of smokers

What about now? How would the sampling distribution change if I change the population to (i) athletes, (ii) old people, (iii) all people?



illustrating examples

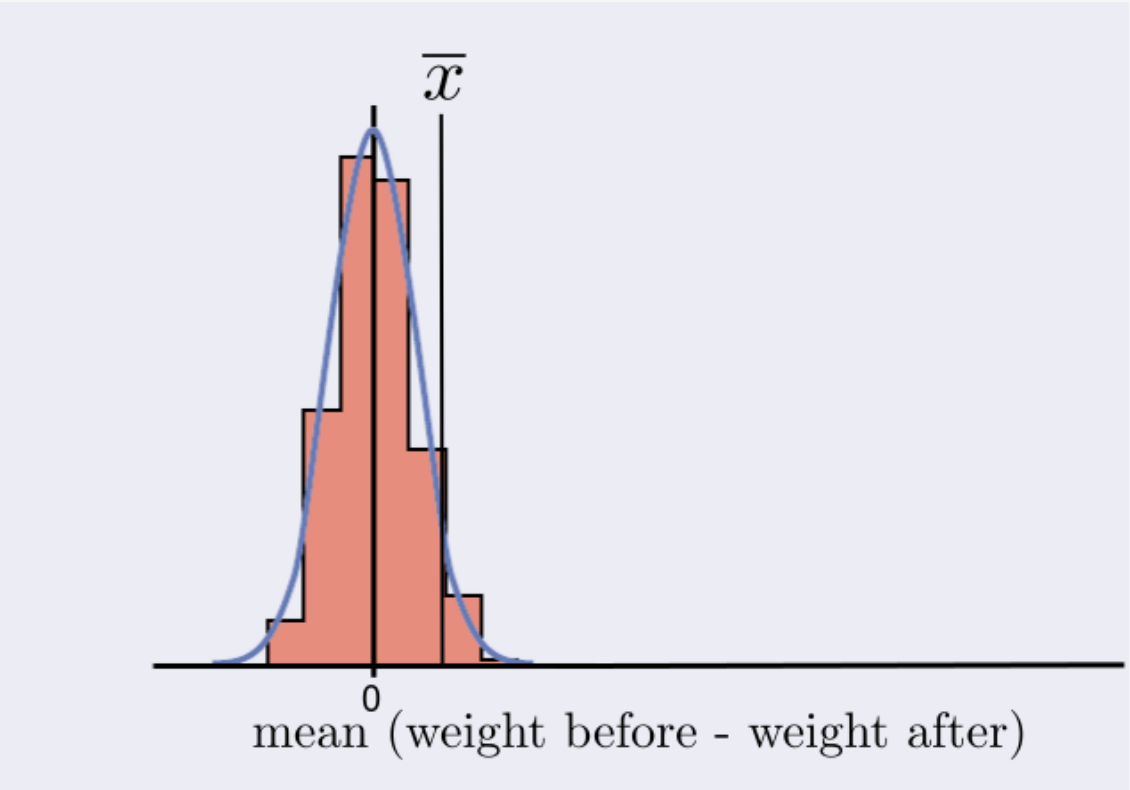
Is this diet effective?



illustrating examples

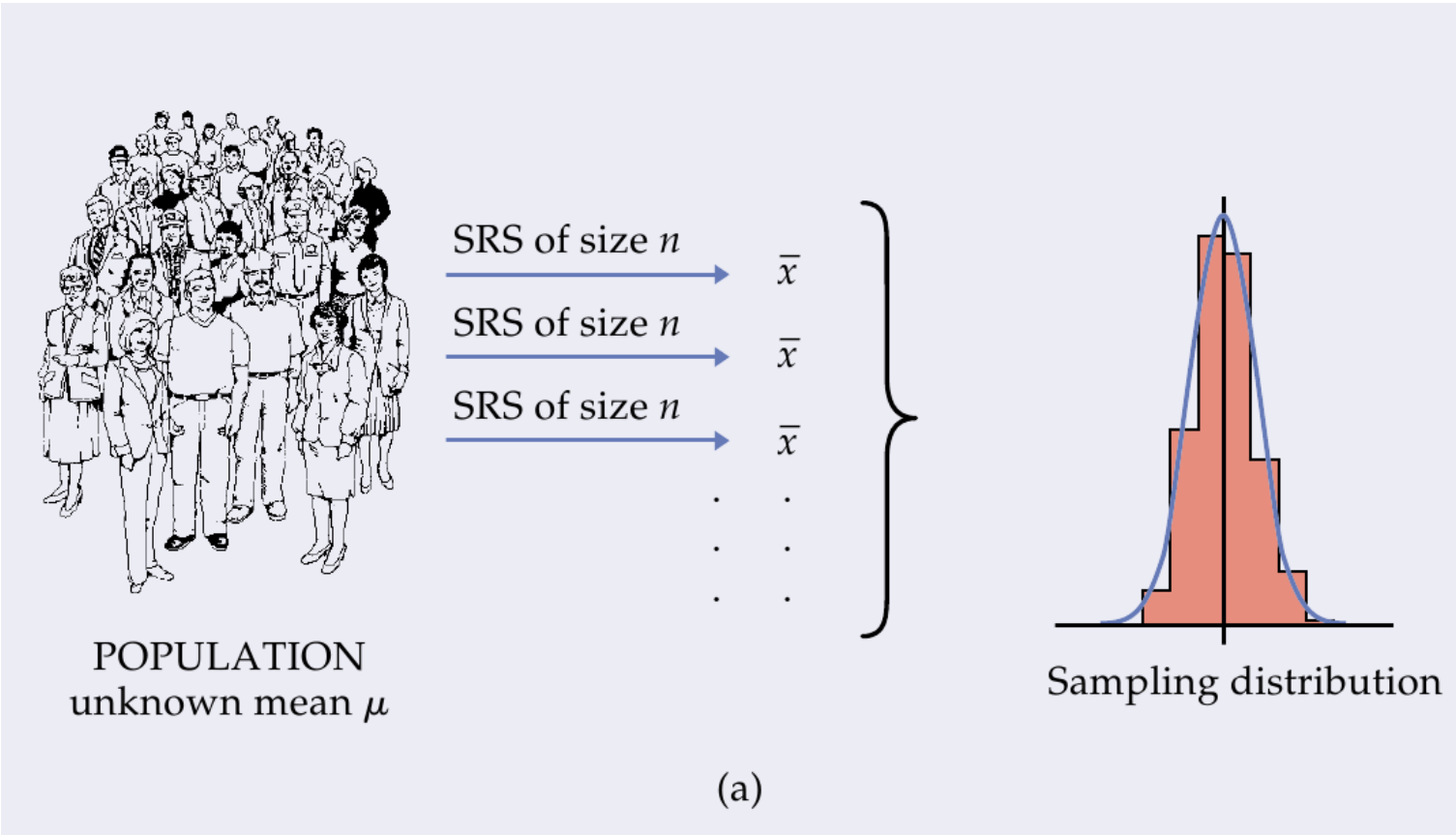
Is this diet effective?

What do you think now?



the (imaginary) meta-study

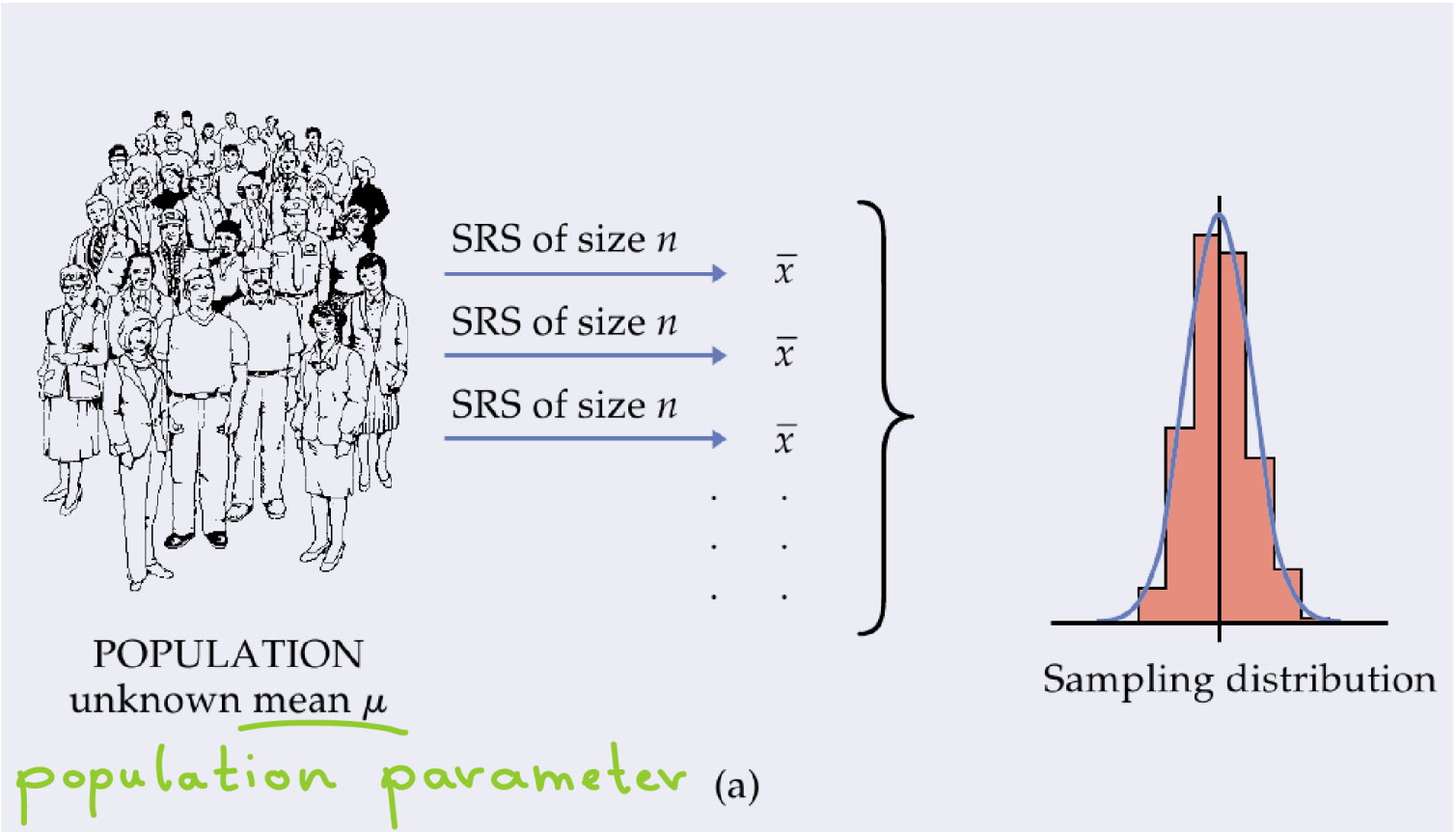
finite sampling introduces variation: the sampling distribution



Hesterberg et al., Bootstrap Methods and Permutation Tests

the (imaginary) meta-study

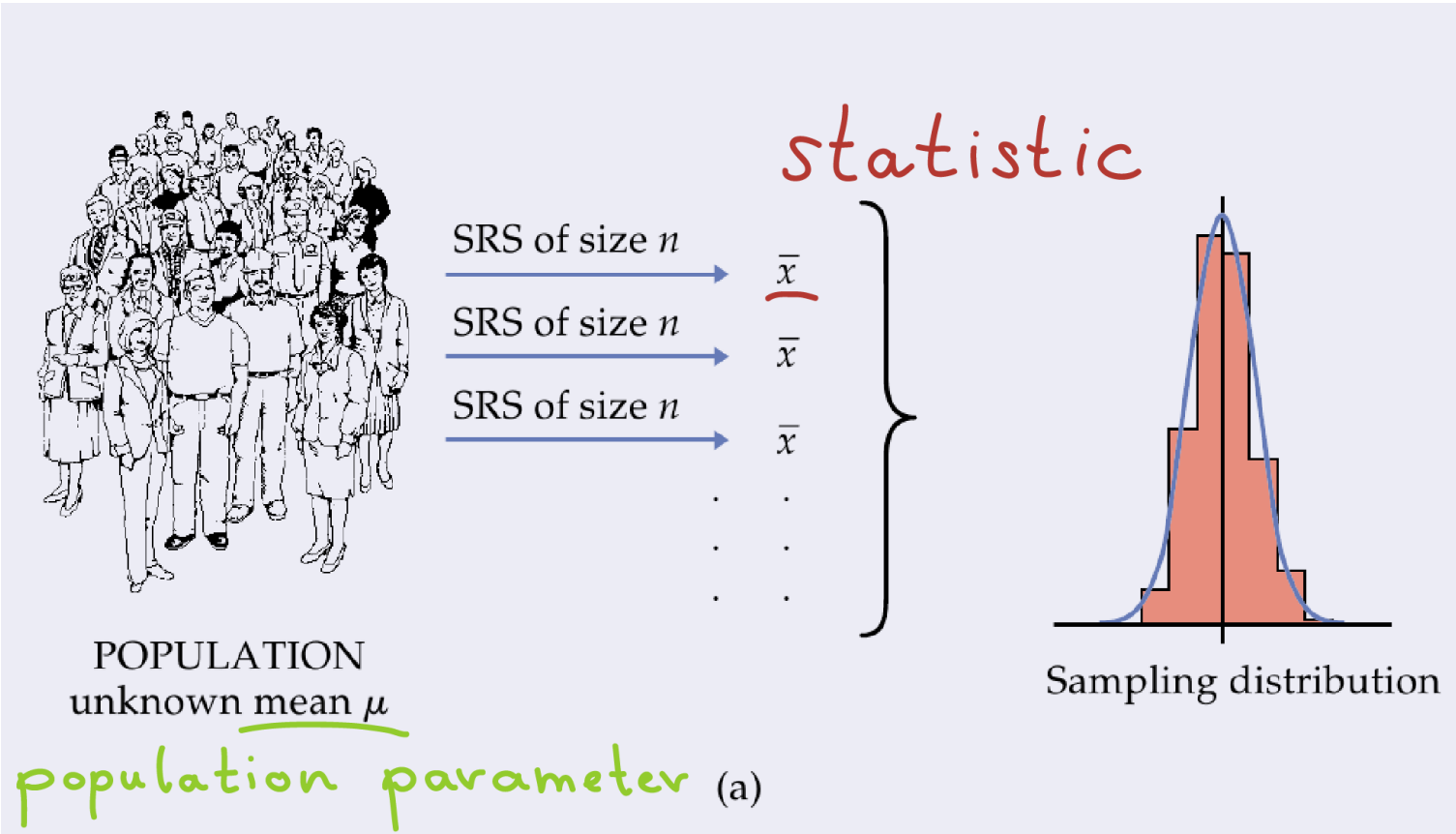
statistic vs. population parameter



Hesterberg et al., Bootstrap Methods and Permutation Tests

the (imaginary) meta-study

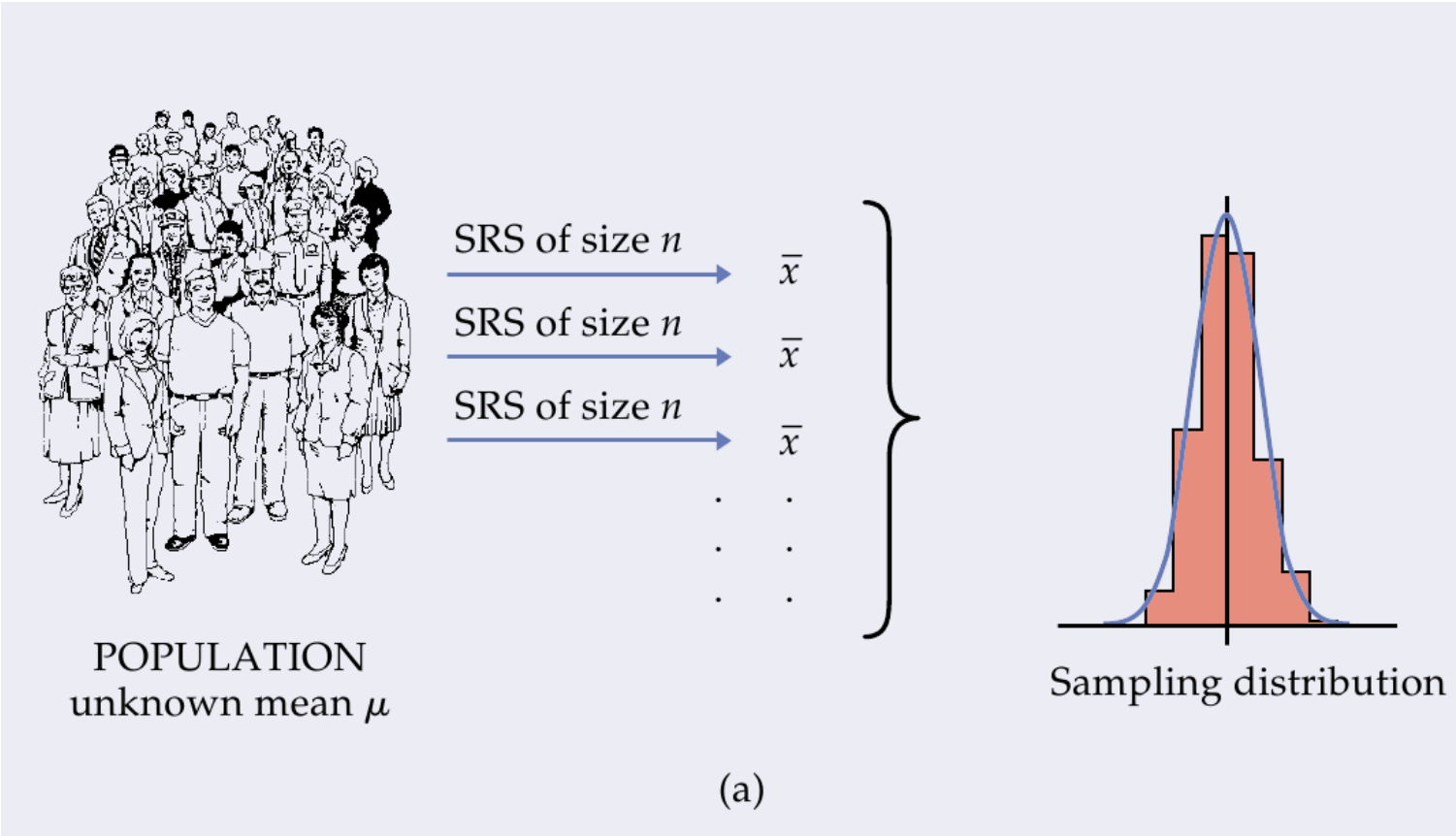
statistic vs. population parameter



Hesterberg et al., Bootstrap Methods and Permutation Tests

the (imaginary) meta-study

shat parts of this diagram do we have in real life?



Hesterberg et al., Bootstrap Methods and Permutation Tests

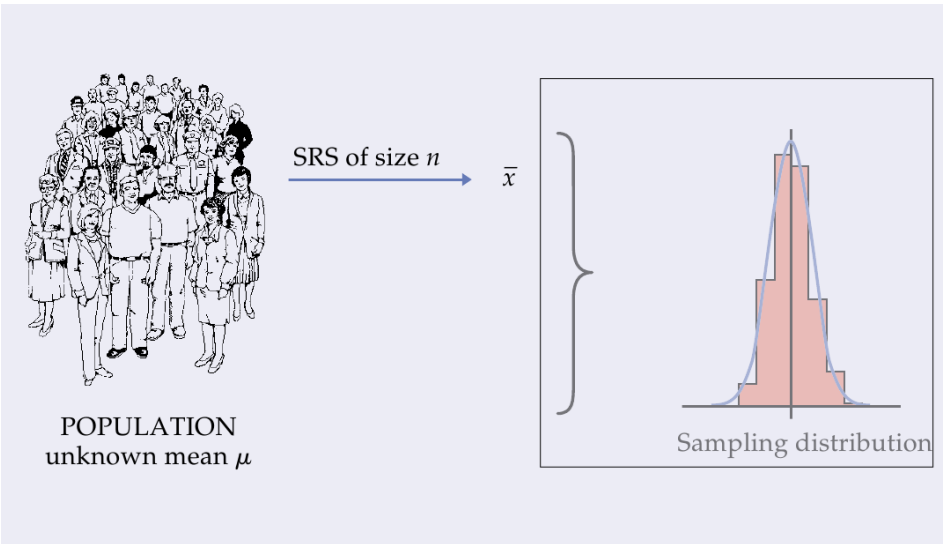
the (imaginary) meta-study

what parts of this diagram do we have in real life?

Hesterberg et al., Bootstrap Methods and Permutation Tests

the (imaginary) meta-study

what statistics does

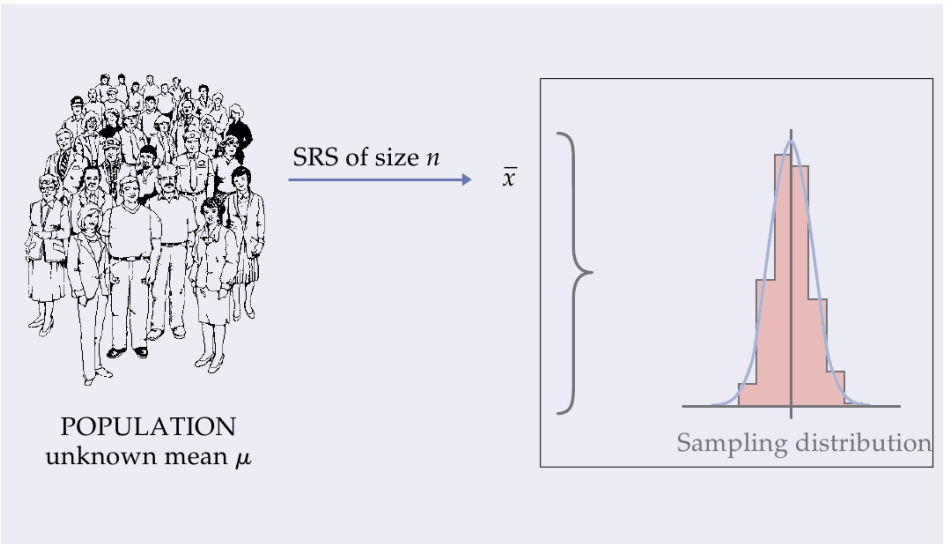


Hesterberg et al., Bootstrap Methods and Permutation Tests

- it assumes, derives, or simulates the sampling distribution

the (imaginary) meta-study

what statistics does

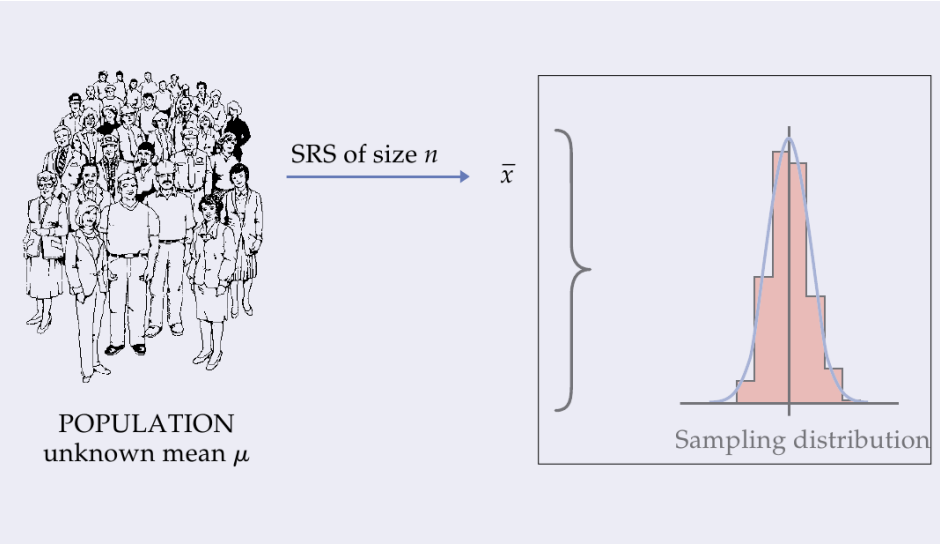


Hesterberg et al., Bootstrap Methods and Permutation Tests

- it assumes, derives, or simulates the sampling distribution
- the sampling distribution makes only sense if you think about it in terms of the meta study

the (imaginary) meta-study

what statistics does



Hesterberg et al., Bootstrap Methods and Permutation Tests

- it assumes, derives, or simulates the sampling distribution
- the sampling distribution makes only sense if you think about it in terms of the meta study
- the sampling distribution is the key to answering questions about the population from the value of the statistic

summary

- In statistics, we use finite samples from a population to reason about features of the population.

summary

- In statistics, we use finite samples from a population to reason about features of the population.
- The particular feature of the population we are interested in is called population parameter. We usually measure this parameter in our finite sample as well (statistic).

summary

- In statistics, we use finite samples from a population to reason about features of the population.
- The particular feature of the population we are interested in is called population parameter. We usually measure this parameter in our finite sample as well (statistic).
- Because of variations due to finite sampling the statistic almost never matches the population parameter.

summary

- In statistics, we use finite samples from a population to reason about features of the population.
- The particular feature of the population we are interested in is called population parameter. We usually measure this parameter in our finite sample as well (statistic).
- Because of variations due to finite sampling the statistic almost never matches the population parameter.
- Using the sampling distribution of the statistic, we make statements about the relation between our statistic and the population parameter.

Day 2 – errorbars, confidence intervals, and tests

illustrating example

As part of a study of the development of the thymus gland, researcher weighed the glands of 50 chick embyos after 14 days of incubation. The following plot depicts the mean thymus gland weights in (mg):

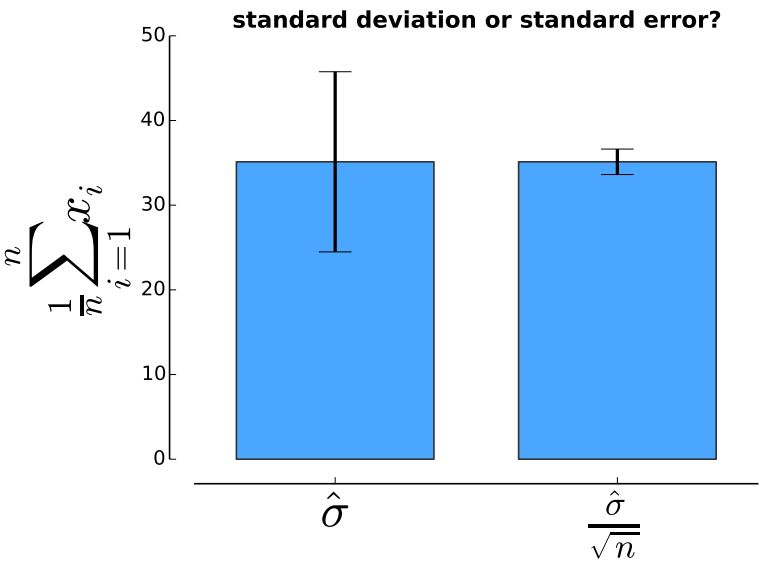
modified from SWS exercise 6.3.3.

illustrating example

As part of a study of the development of the thymus gland, researcher weighed the glands of 50 chick embyos after 14 days of incubation. The following plot depicts the mean thymus gland weights in (mg):

modified from SWS exercise 6.3.3.

Which of the two bar plots is the correct way of displaying the data?

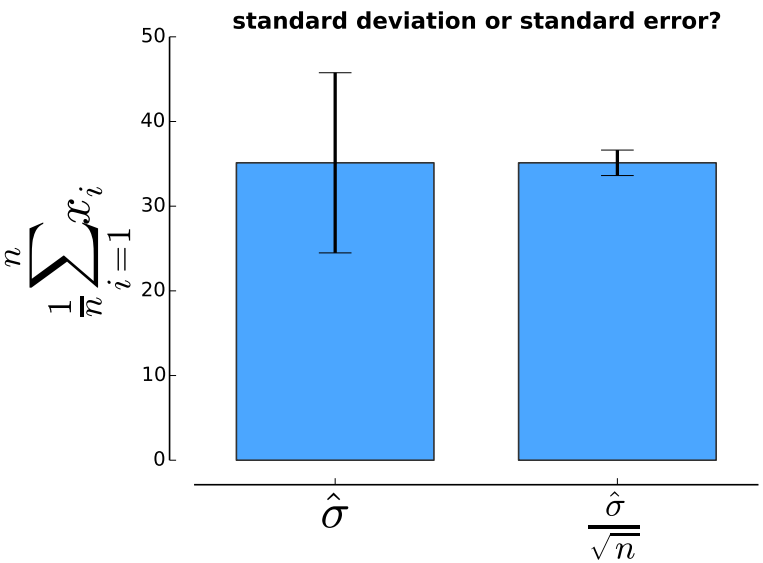


illustrating example

As part of a study of the development of the thymus gland, researcher weighed the glands of 50 chick embyos after 14 days of incubation. The following plot depicts the mean thymus gland weights in (mg):

modified from SWS exercise 6.3.3.

Which of the two bar plots is the correct way of displaying the data?



That depends on what you want to say

- To give a measure of variability in the data: use the standard deviation $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$
- To make a statement about the variability in the mean estimation: use standard error $\frac{\hat{\sigma}}{\sqrt{n}}$

standard error

bootstrapping

standard error vs. standard deviation

- Download the dataset `thymusglandweights.dat` from Ilias
- Write a program that loads the data into matlab, extracts the the first 80 datapoints, and repeat the following steps $m = 500$ times:
  1. draw 80 data points from x with replacement
  2. compute their mean and store it

  Look at the standard deviation of the computed means.
- Compare the result to the standard deviation of the original 80 data points and the standard error.

standard error

```
load thymusglandweights.dat

n = 80;
m = 500;
x = thymusglandweights(1:n);


mu = zeros(m,1);
for i = 1:m
    mu(i) = mean(x(randi(n,n,1)));
end
disp(['bootstrap standard error: ', num2str(std(mu))]);
disp(['standard error: ', num2str(std(x)/sqrt(n))]);
```

standard error

bootstrapping

- The sample standard error $\frac{\hat{\sigma}}{\sqrt{n}}$ is an estimate of the standard deviation of the means in repeated experiments which is computed form a single experiment.
- When you want to do statistical tests on the mean, it is better to use the standard error, because one can eyeball significance from it

  Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. The Journal of Cell Biology, 177(1), 7–11.

- Bootstrapping is a way to generate an estimate of the sampling distribution of any statistic. Instead of sampling from the true distribution, it samples from the empirical distribution represented by your dataset.

  Efron, B., & Tibshirani, R. J. (1994). An Introduction to the Bootstrap. Chapman and Hall/CRC

standard error of the median?

What kind of errorbars should we use for the median?

It depends again:

Descriptive statistics

- As a descriptive statistic one could use the median absolute deviation: the median of the absolute differences of the datapoints from the median.
- Alternatively, one could bootstrap a standard error of the median.

standard error of the median?

What kind of errorbars should we use for the median?

It depends again:

Descriptive statistics

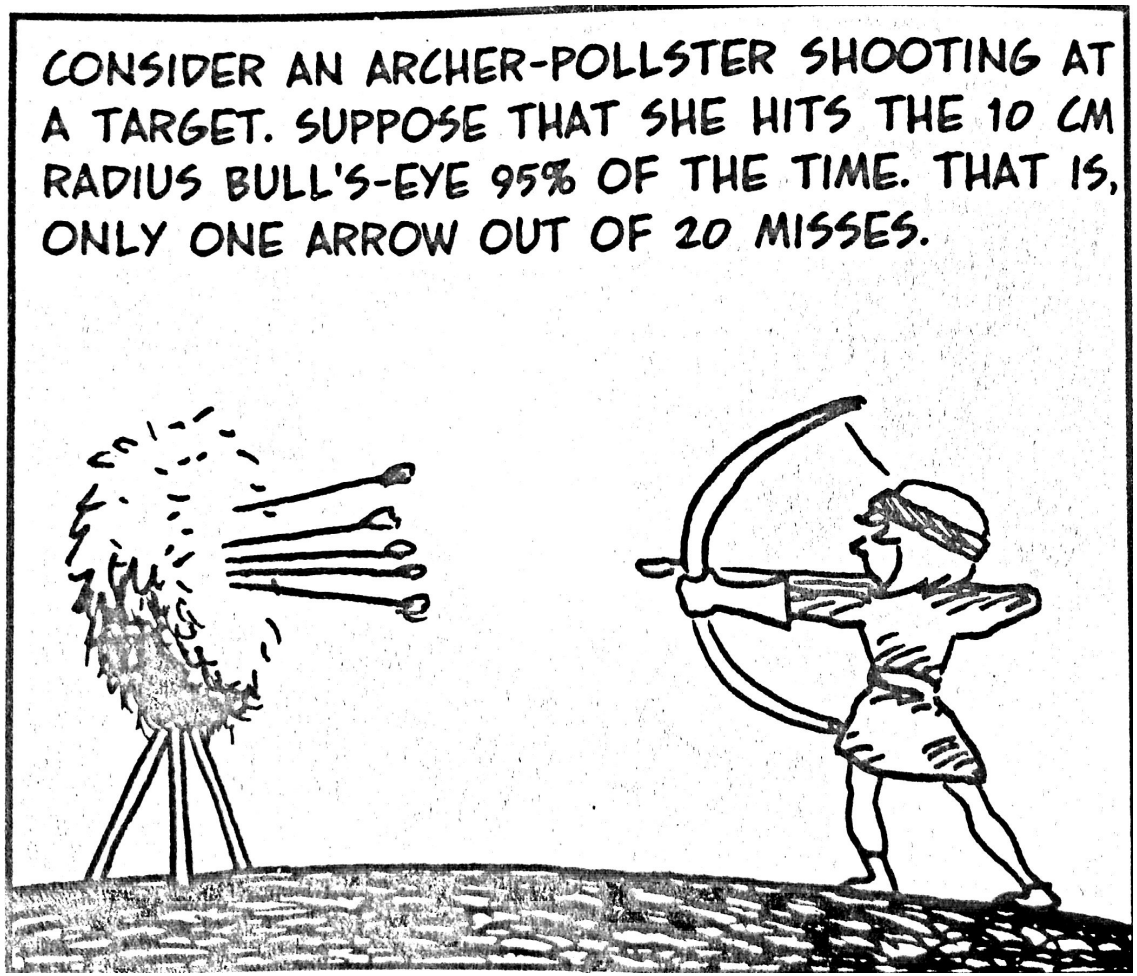
- As a descriptive statistic one could use the median absolute deviation: the median of the absolute differences of the datapoints from the median.
- Alternatively, one could bootstrap a standard error of the median.

Inferential statistics

- For inferential statistics one should use something that gives the reader information about significance.
- Here, confidence intervals are a better choice.

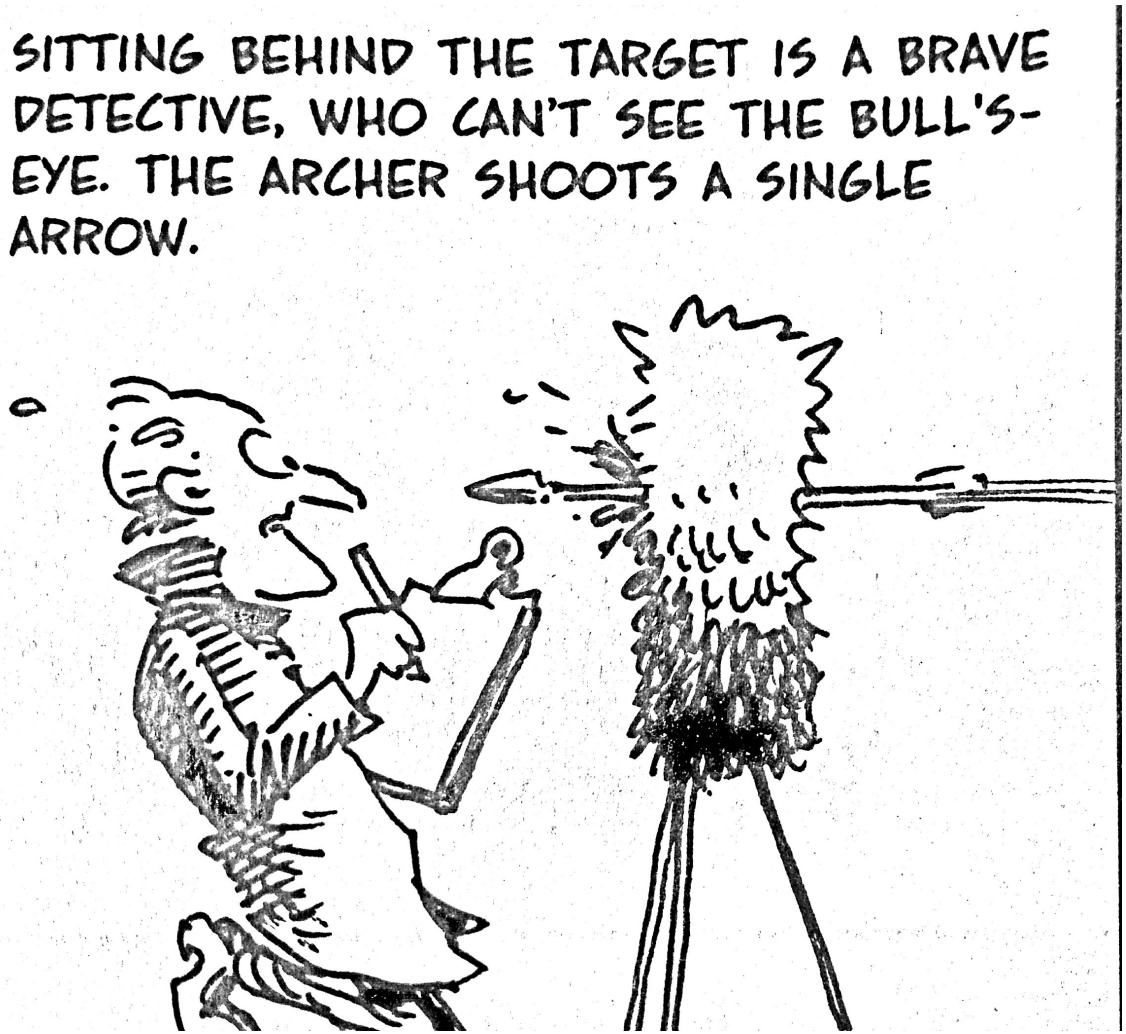
Day 2 – errorbars, confidence intervals, and tests

confidence intervals



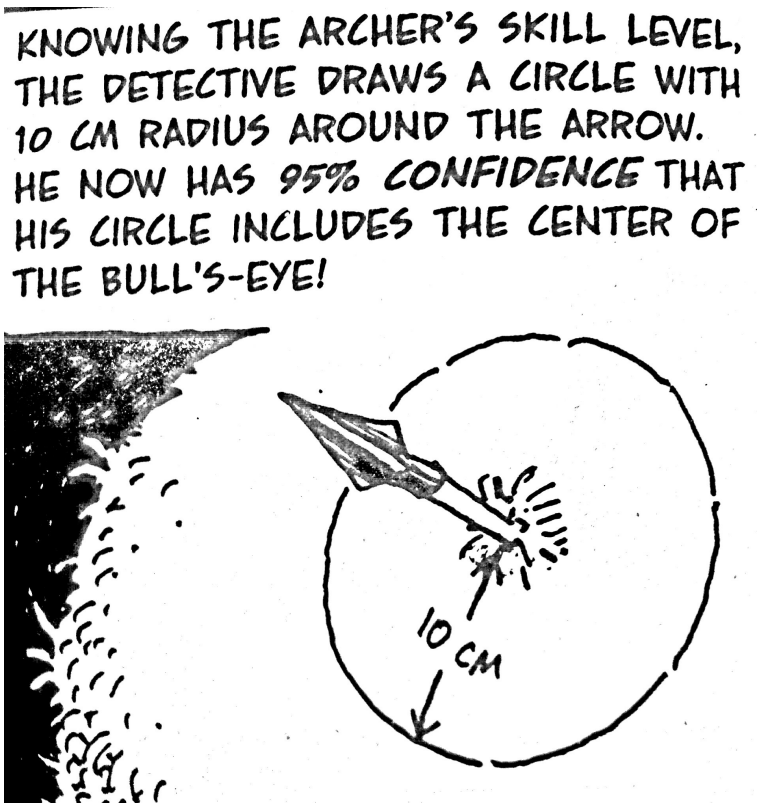
Larry Gonick, The Cartoon Guide to Statistics

confidence intervals



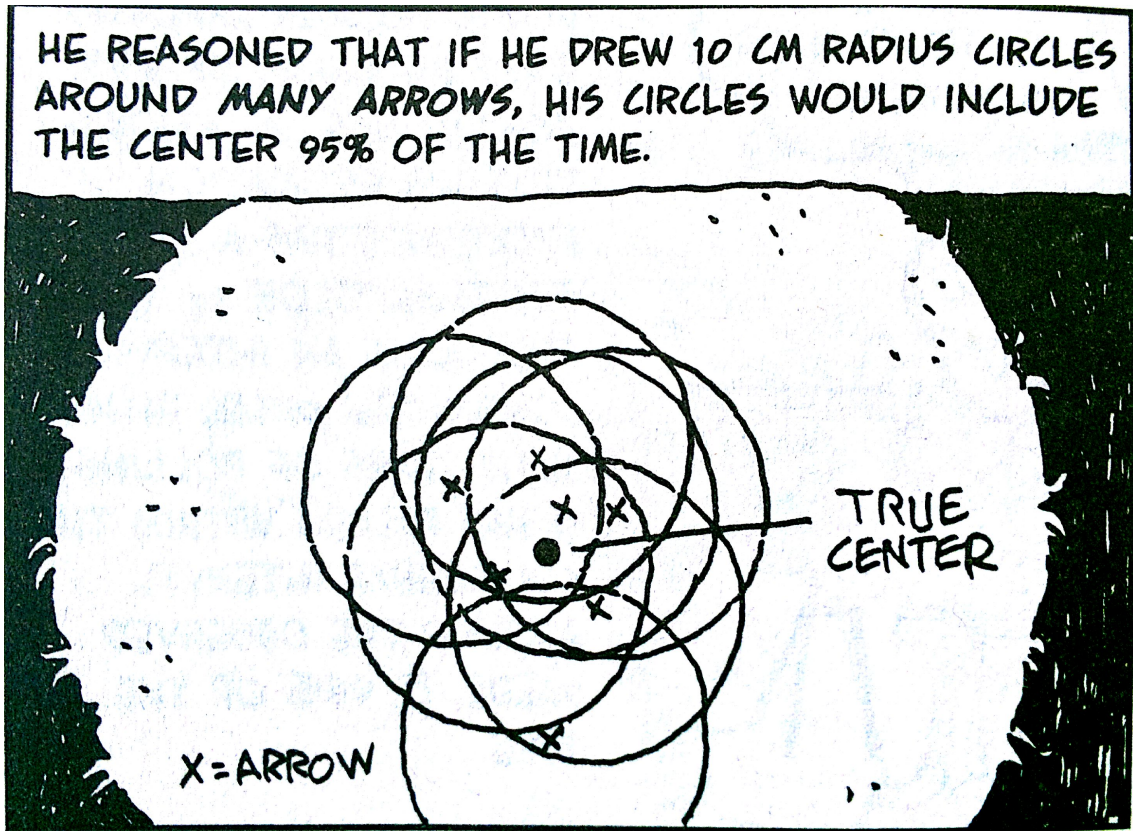
Larry Gonick, The Cartoon Guide to Statistics

confidence intervals



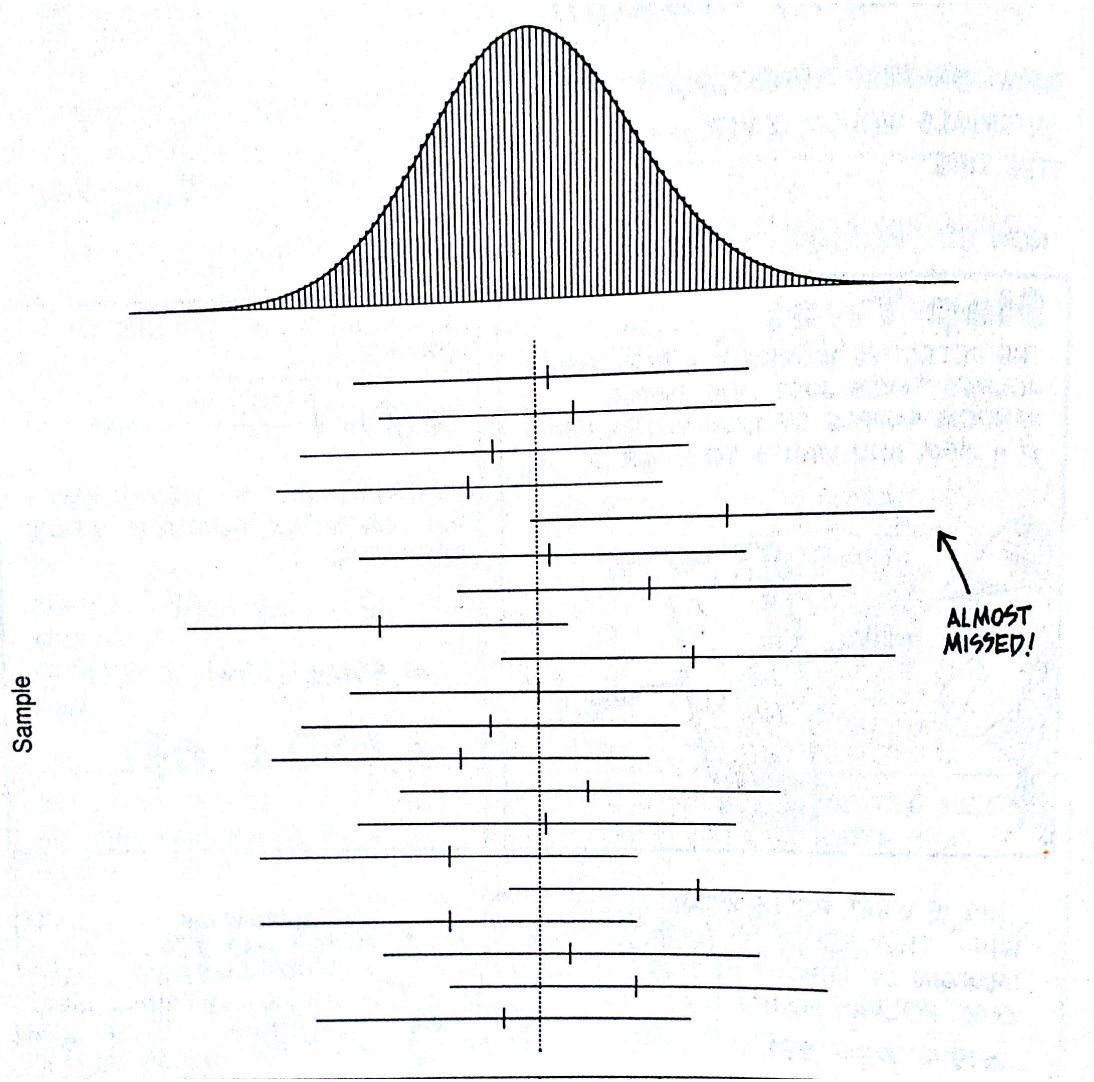
Larry Gonick, The Cartoon Guide to Statistics

confidence intervals



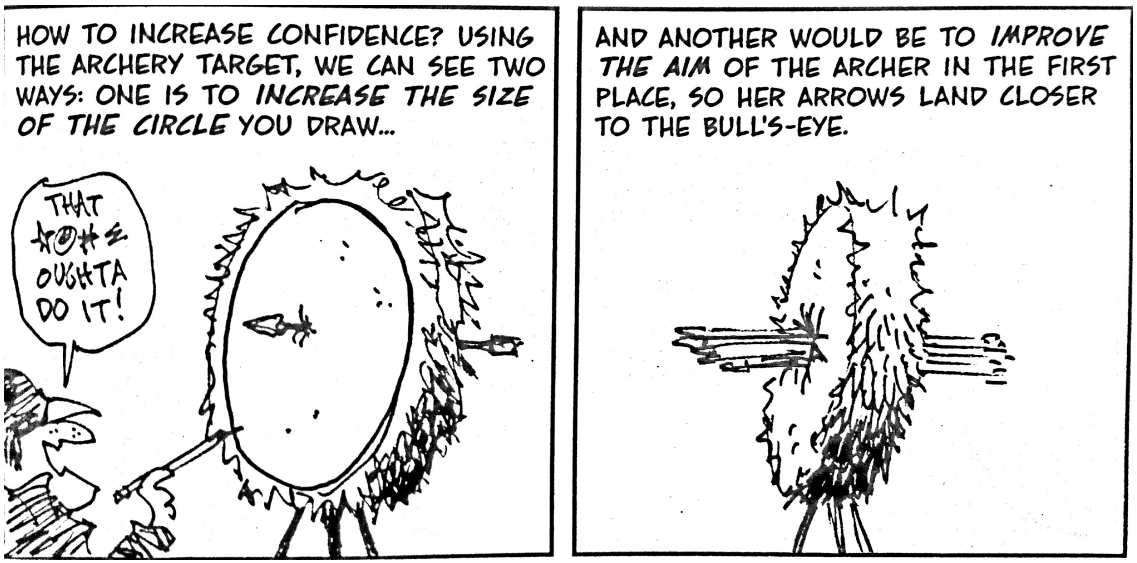
Larry Gonick, The Cartoon Guide to Statistics

confidence intervals



Larry Gonick, The Cartoon Guide to Statistics

confidence intervals



Larry Gonick, The Cartoon Guide to Statistics

confidence intervals for the median

Confidence interval

A confidence $(1-\alpha)\cdot 100\%$ interval for a statistic $\hat{\theta}$ is an interval $\hat{\theta} \pm a$ such that the population parameter θ is contained in that interval $(1-\alpha)\cdot 100\%$ of the experiments.
An alternative way to put it is that $(\hat{\theta}-\theta) \in [-a, a]$ in $(1-\alpha)\cdot 100\%$ of the cases.

If we knew the sampling distribution of the median $\hat{m}$, could we generate a e.g. a 95% confidence interval?

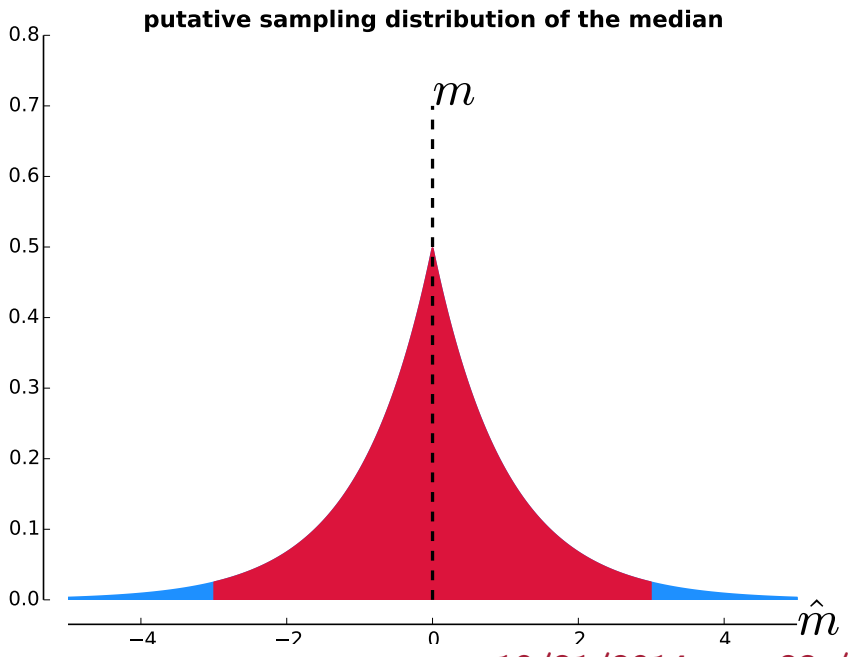
confidence intervals for the median

Confidence interval

A confidence $(1-\alpha)\cdot 100\%$ interval for a statistic $\hat{\theta}$ is an interval $\hat{\theta} \pm a$ such that the population parameter θ is contained in that interval $(1-\alpha)\cdot 100\%$ of the experiments.
An alternative way to put it is that $(\hat{\theta} - \theta) \in [-a, a]$ in $(1-\alpha)\cdot 100\%$ of the cases.

If we knew the sampling distribution of the median $\hat{m}$, could we generate a e.g. a 95% confidence interval?

Yes, we could choose the interval such that $\hat{m} - m$ in that interval in 95% of the cases.



confidence intervals for the mean via bootstrapping

how to get the sampling distribution

bootstrapping a confidence interval for the mean

- Use the same dataset as before.
- Bootstrap 500 means.
- Plot their distribution.
- Compute the 2.5% and the 97.5% percentile of the 500 means.
- Mark them in the plot.

These two numbers give you $\hat{m} - a$ and $\hat{m} + a$ for the 95% confidence interval.

confidence intervals for the median

```
1  load thymusglandweights.dat
2  n = 80;
3  x = thymusglandweights(1:n);
4
5  m = 500;
6  me = zeros(m,1);
7  for i = 1:m
8      me(i) = mean(x(randi(n,n,1)));
9  end
10
11 disp(['bootstrap quantiles: ' , num2str(quantile(me,0.025)), ' ' ,num2str(
       quantile(me,1-0.025))]);
```

confidence intervals

Notice the theme!

1. choose a statistic
2. get a the sampling distribution of the statistic (by theory or simulation)
3. use that distribution to reason about the relation between the true population parameter (e.g. m) and the sampled statistic $\hat{m}$

This is the scaffold of most statistical techniques. Try to find it and it can help you understand them.

confidence interval for the mean

Let's search the pattern in the normal way of computing a confidence interval for the mean

- If the $x_1, ..., x_n \sim \mathcal{N}(\mu, \sigma)$ are Gaussian, then $\hat{\mu}$ is Gaussian as well
- What is the mean of $\hat{\mu}$? What is its standard deviation?

confidence interval for the mean

Let's search the pattern in the normal way of computing a confidence interval for the mean

- If the $x_1, ..., x_n \sim \mathcal{N}(\mu, \sigma)$ are Gaussian, then $\hat{\mu}$ is Gaussian as well
- What is the mean of $\hat{\mu}$? What is its standard deviation?

  $\langle \hat{\mu} \rangle_{x_1,...,x_n} = \mu$ and $\text{std}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$

confidence interval for the mean

Let's search the pattern in the normal way of computing a confidence interval for the mean

- If the $x_1, ..., x_n \sim \mathcal{N}(\mu, \sigma)$ are Gaussian, then $\hat{\mu}$ is Gaussian as well
- What is the mean of $\hat{\mu}$? What is its standard deviation?
  $\langle \hat{\mu} \rangle_{x_1,...,x_n} = \mu$ and $\text{std}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$
- The problem is, that $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ depends on unknown population parameters.

confidence interval for the mean

Let's search the pattern in the normal way of computing a confidence interval for the mean

- If the $x_1, ..., x_n \sim \mathcal{N}(\mu, \sigma)$ are Gaussian, then $\hat{\mu}$ is Gaussian as well
- What is the mean of $\hat{\mu}$? What is its standard deviation?
  $\langle \hat{\mu} \rangle_{x_1,...,x_n} = \mu$ and $\text{std}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$
- The problem is, that $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ depends on unknown population parameters.
- However,

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \sim \text{t-distribution with } n-1 \text{ degrees of freedom}$$

- Therefore,

$$P\left(t_{2.5\%} \leq \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \leq t_{97.5\%}\right) = P\left(t_{2.5\%}\frac{\hat{\sigma}}{\sqrt{n}} \leq \hat{\mu} - \mu \leq t_{97.5\%}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

confidence interval for the mean

Bootstrapping a confidence interval for the mean

Extend your script to contain the analytical confidence interval using

$$P\left(t_{2.5\%} \leq \frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}} \leq t_{97.5\%}\right) = P\left(t_{2.5\%}\frac{\hat{\sigma}}{\sqrt{n}} \leq \hat{\mu}-\mu \leq t_{97.5\%}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

Hint: Use the function `tinv(0.025, n-1)` to get the value of $t_{2.5\%}$ and similar for $t_{97.5\%}$.

solution

```
1  load thymusglandweights.dat
2  n = 80;
3  x = thymusglandweights(1:n);
4
5  m = 500;
6  me = zeros(m,1);
7  for i = 1:m
8      me(i) = mean(x(randi(n,n,1)));
9  end
10
11 t025 = tinv(0.025, n-1);
12 t975 = tinv(0.975, n-1);
13
14 se = std(x)/sqrt(n);
15
16 disp(['bootstrap quantiles: ' , num2str(quantile(me,0.025)), ' ' ,num2str(
       quantile(me,1-0.025))]);
17 disp(['analytical CI: ' , num2str(mean(x)+t025*se), ' ' , num2str(mean(x)+
       t975*se)]);
```

Day 2 – errorbars, confidence intervals, and tests

ingredients into a test

- **What is the goal of a test?**

ingredients into a test

- **What is the goal of a test?**

  Check whether a measured statistic looks different from what you would expect if there was no effect.

ingredients into a test

- **What is the goal of a test?**

  Check whether a measured statistic looks different from what you would expect if there was no effect.
- **What are the ingredients into a test?**

ingredients into a test

- **What is the goal of a test?**

  Check whether a measured statistic looks different from what you would expect if there was no effect.
- **What are the ingredients into a test?**

  a test statistic (e.g. the mean, the median, ...) and a null distribution

ingredients into a test

- **What is the goal of a test?**

  Check whether a measured statistic looks different from what you would expect if there was no effect.
- **What are the ingredients into a test?**

  a test statistic (e.g. the mean, the median, ...) and a null distribution
- **What is a null distribution?**

ingredients into a test

- **What is the goal of a test?**

  Check whether a measured statistic looks different from what you would expect if there was no effect.
- **What are the ingredients into a test?**

  a test statistic (e.g. the mean, the median, ...) and a null distribution
- **What is a null distribution?**

  The sampling distribution of the statistic in case there is no effect (i.e. the Null hypothesis is true).

how tests work

1. Choose a statistic.
2. Get a null distribution.
3. Compare your actually measure value with the Null distribution.

Example: one sample test

step 2: get a Null distribution

Assume that the expected weight of a thymus gland from the literature is 34.3g. We want to test whether the mean of our thymus gland dataset is different from the expectation in the literature. Comparing a statistic of a dataset against a fixed value is called one sample test.

Example: one sample test

step 2: get a Null distribution

Assume that the expected weight of a thymus gland from the literature is 34.3g. We want to test whether the mean of our thymus gland dataset is different from the expectation in the literature. Comparing a statistic of a dataset against a fixed value is called one sample test.

- **How could we simulate the distribution of the data if the mean was really 30g?**

Example: one sample test

step 2: get a Null distribution

Assume that the expected weight of a thymus gland from the literature is 34.3g. We want to test whether the mean of our thymus gland dataset is different from the expectation in the literature. Comparing a statistic of a dataset against a fixed value is called one sample test.

- **How could we simulate the distribution of the data if the mean was really 30g?**

  Bootstrapping.

generating a null distribution

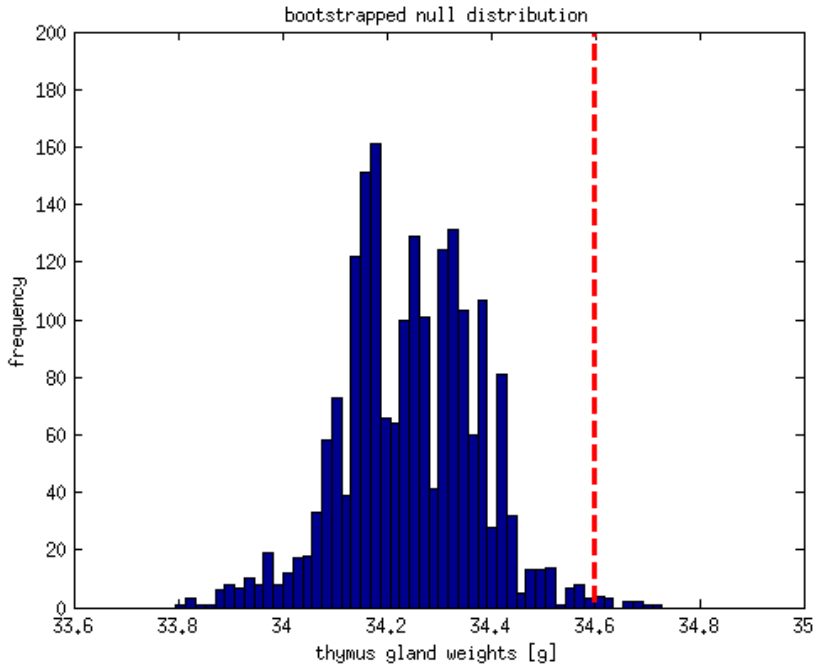
- Write a matlab program that bootstraps 2000 means from the thymus gland dataset.
- How can we adjust the data that it has mean 34.3g (remember, we want to simulate the null distribution)?
- Plot a histogram of these 2000 means.
- Also indicate the actual mean of the data.

Example: one sample test

step 3: compare the actual value to the Null distribution

The question we want to answer in this step is:

Does the actually measure value look like it came from the Null distribution?



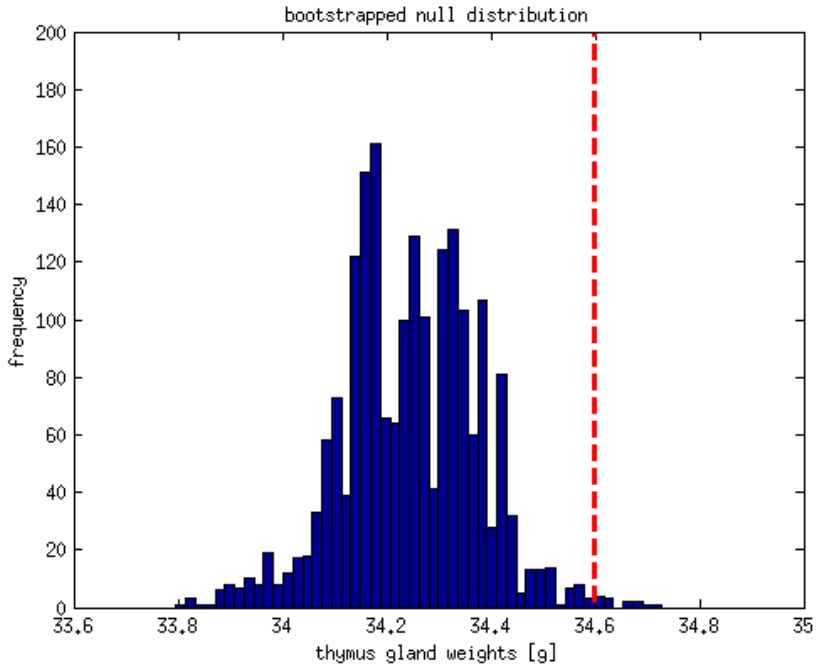
How could we do this in our bootstrapping example?

Example: one sample test

step 3: compare the actual value to the Null distribution

The question we want to answer in this step is:

Does the actually measure value look like it came from the Null distribution?



How could we do this in our bootstrapping example?

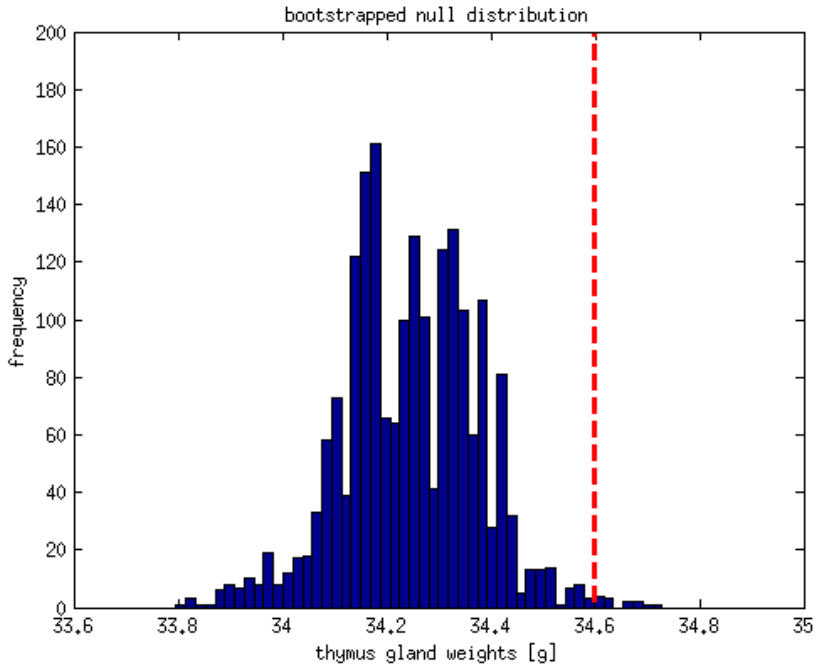
- Set a threshold.

Example: one sample test

step 3: compare the actual value to the Null distribution

The question we want to answer in this step is:

Does the actually measure value look like it came from the Null distribution?



How could we do this in our bootstrapping example?

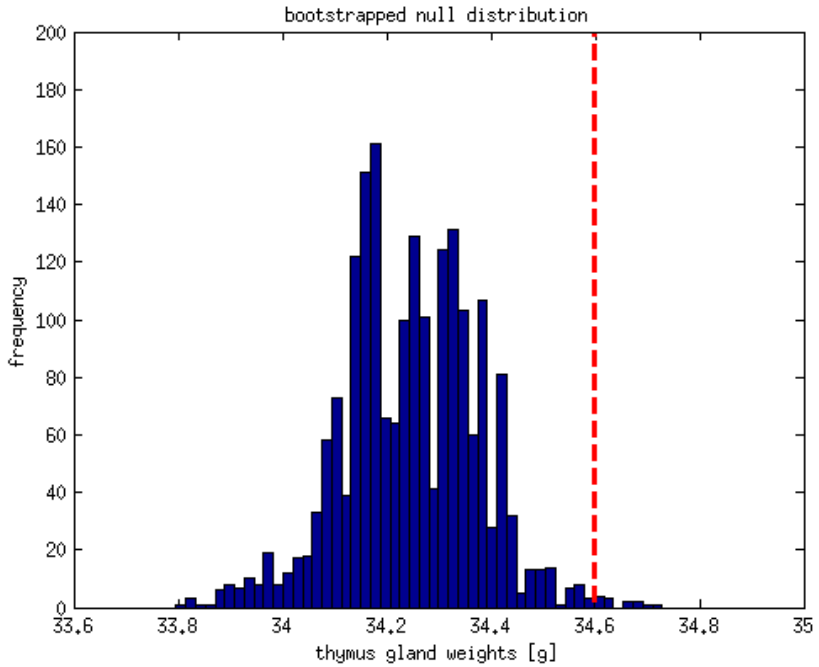
- Set a threshold. How do we choose the threshold?

Example: one sample test

step 3: compare the actual value to the Null distribution

The question we want to answer in this step is:

Does the actually measure value look like it came from the Null distribution?



How could we do this in our bootstrapping example?

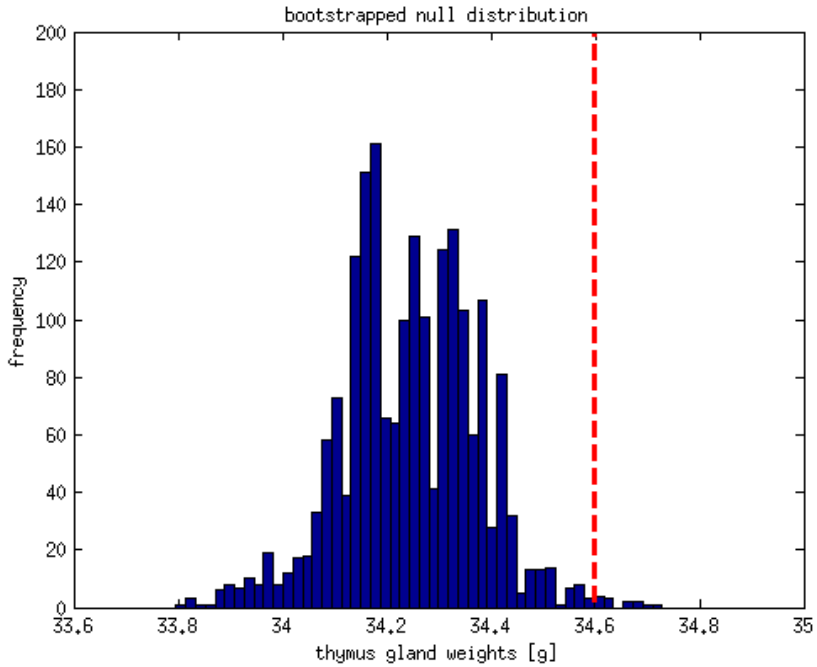
- Set a threshold. How do we choose the threshold? Via type I error.

Example: one sample test

step 3: compare the actual value to the Null distribution

The question we want to answer in this step is:

Does the actually measure value look like it came from the Null distribution?



How could we do this in our bootstrapping example?

- Set a threshold. How do we choose the threshold? Via type I error.
- Specify the type I error if we used the actual measured value as threshold (p-value). Why is that a reasonable strategy?

Example: one sample test

step 3: compare the actual value to the Null distribution

type I error and p-value

Extend the script such that it

- computes the 5% significance boundaries from the distribution and plot it into the histogram.
- computes a p-value.

two sample test

permutation test

Brain Weight In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. Brain weights are given in gram. The data can be downloaded from Ilias (example 002 from yesterday).

How could we determine (similar to bootstrapping) whether the mean brain weight of males and females are different?

- What do we use as a statistic?
- How do we simulate the null distribution?

two sample test

permutation test

Brain Weight In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. Brain weights are given in gram. The data can be downloaded from Ilias (example 002 from yesterday).

How could we determine (similar to bootstrapping) whether the mean brain weight of males and females are different?

- What do we use as a statistic?

  The difference of the means of the two groups.
- How do we simulate the null distribution?

two sample test

permutation test

Brain Weight In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. Brain weights are given in gram. The data can be downloaded from Ilias (example 002 from yesterday).

How could we determine (similar to bootstrapping) whether the mean brain weight of males and females are different?

- What do we use as a statistic?

  The difference of the means of the two groups.
- How do we simulate the null distribution?

  Shuffle the labels "male" and "female", compute difference in means of two groups, and repeat.

That's it.