Computational Methods for Statistics (VU) (706.026)

Elisabeth Lex

ISDS, TU Graz

Jan 17, 2023

About today's class

- Definitions: population, random sample, sampled population, statistic
- Sampling distribution
- Law of large numbers + Central limit theorem
- Confidence intervals
- Bootstrap method to compute confidence intervals

Materials consist of slides and recommended readings.

Learning Goals

At the end of this unit, you will be able to:

- Define target population, random sample, sampled population, statistic.
- Explain the concept of a sampling distribution.
- Explain the difference between the distribution of a target population, the distribution of a sample, and the sampling distributions of a statistic.
- Create sampling distributions in Python.
- Explain the central limit theorem.
- Explain the bootstrap algorithm.

Repetition

- Motivation why computational methods are beneficial
- Inductive and deductive inference
- Sampling

Materials consist of slides and recommended readings.

Sampling

Random Sample

Definition 6 (Random sample)

Let the random variables $X_1, X_2, \ldots, X_n$ have a joint density $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ that factors as follows:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2) \cdot \cdots \cdot f(x_n),$$

where $f(\cdot)$ is the common density of each X_i. We then define $X_1, X_2, \ldots, X_n$ to be a *random sample* of size n from a population with density $f(\cdot)$. Thus, a random sample is a sequence of independent, identically distributed (*i.i.d.*) random variables.

Remark 9 (Sampling with/without replacement)

When sampling from a finite population, our definition requires to always sample with replacement as otherwise the drawings are not independent.

Random Sample

Example 3 (Rise of Skywalker)

We define X_i as 1 (ith person watched the movie) or as 0 (ith person did not watch the movie). If we sample people so that the variables $X_1, X_2, \ldots, X_n$ are independent and have the same density (all people have the same probability of watching the movie) then the sample is random.

Sampled Population

Definition 7 (Sampled population)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density $f(\cdot)$; then this population is called *sampled population*.

Remark 10 (Distinction between the sampled and the target population)

- With random samples we can only make valid probability statements about sampled population.
- Statements about target population are not valid.
- Unless the target population is also the sampled population.

Example 3 (Rise of Skywalker)

All people living in Austria form the *target population*. We draw a sample from Graz. Thus, Graz residents form the *sampled population*.

Methodology for inductive inference revisited

- Goal: study a population with density $f(\cdot;\theta)$.
- We know the form of the density, but it contains an unknown parameter θ.
- Take a random sample $X_1, X_2, \ldots, X_n$ of size n from $f(\cdot;\theta)$.
- We compute the value of some function $t(x_1, x_2, \ldots, x_n)$ to estimate θ.

Remark 11 (Terminology clarification)

- $X_i's$ are random variables (mathematical objects).
- $x_i's$ are realizations or data points (concrete observations).

Statistic

- The function, which we compute on some concrete realizations, is called **statistic**.

Definition 8 (Statistic)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from density $f(\cdot)$. Statistic is a function $T_n = t(X_1, X_2, \ldots, X_n)$.

Remark 12 (Properties of statistics)

- A function of a random variable is also a random variable. $\implies$ each statistic is a random variable.
- Each random variable has a density. $\implies$ each statistic has a density.
- For probabilistic properties, we study *sampling distribution* of T_n.
- For a concrete application, we study $t_n = t(x_1, x_2, \ldots, x_n)$.

Statistic

- Statistic cannot depend on unknown parameters θ if we have a density $f(\cdot;\theta)$.
- For example, if the random variable X has a density $f(\cdot;\mu,\sigma^2)$, where μ and σ^2 are unknown, then:
  - $X-\mu$ is not a statistic.
  - $\frac{X}{\sigma}$ is not a statistic.
  - X, $X+4$, X^2, X^3+lnX^2 are all statistics.

Remark 13 (Rule of thumb)

If you can write a computer function to compute a value only from your data, then it is a statistic.

Example: Sample Mean (is a Statistic)

Definition 9 (Sample mean)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the density $f(\cdot)$. *Sample mean* is defined as the average value:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

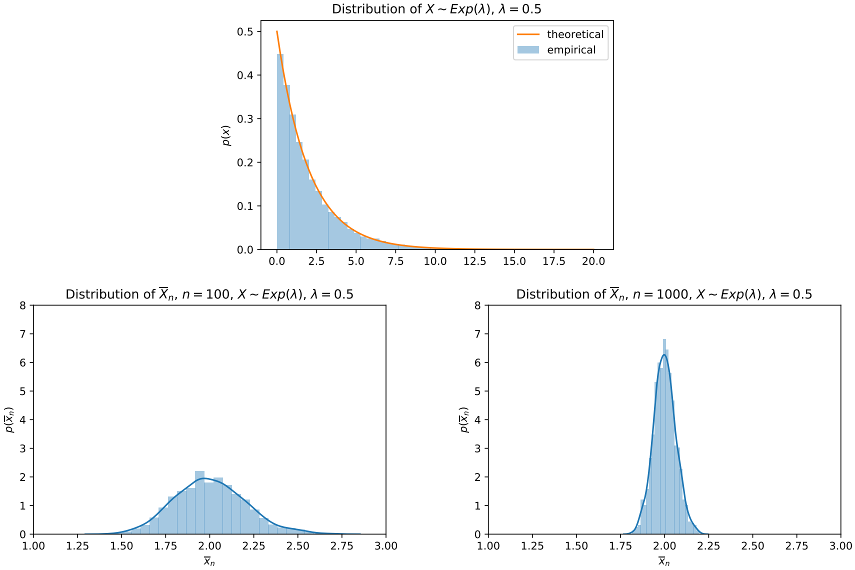
Example 4 (Sampling distribution of the sample mean)

Let $X_1, X_2, \ldots, X_n$ be a random sample with $X_i \sim Exp(\lambda)$. Estimate the sampling distribution of $\overline{X}_n$ by repeatedly drawing random numbers in python with $n = 100$ and $\lambda = 0.5$. Repeat the experiment for $n = 1000$.

Notebook 4 (Sample mean)

`sample_mean.ipynb`

Sampling Distribution of $\overline{X}_n$



Using Sample Mean for Estimation of θ

- We compute the value of some function $t(x_1, x_2, \ldots, x_n)$ to estimate θ
- For which parameter can we use $\overline{X}_n$?

Using Sample Mean for Estimation of θ

- We compute the value of some function $t(x_1, x_2, \ldots, x_n)$ to estimate θ
- For which parameter can we use $\overline{X}_n$?
- Expectation $E(X)$ (average or mean), which we will denote with μ
- What is the intuition behind using $\overline{X}_n$ as an estimator for μ?

Using Sample Mean for Estimation of θ

- We compute the value of some function $t(x_1, x_2, \ldots, x_n)$ to estimate θ
- For which parameter can we use $\overline{X}_n$?
- Expectation $E(X)$ (average or mean), which we will denote with μ
- What is the intuition behind using $\overline{X}_n$ as an estimator for μ?
- $E(X)$ is the average number of infinitely many values of X
- $\overline{X}_n$ is the average number of n values of X
- Theoretical result: **Law of large numbers**

Law of large numbers

Theorem 3 ((Weak) Law of large numbers)

Let $X_1, X_2, \ldots, X_n$ *be a random sample of size* n *from the density* $f(\cdot)$ *and let* $E(X_1) = \mu$ *and* $Var(X_1) = \sigma^2$ *be finite. Then:*

$$\lim_{n \to \infty} P(|\overline{X}_n - \mu| \geq \epsilon) = 0, \forall \epsilon > 0.$$

Remark 14 (Interpretation of WLLN)

The sampling distribution of $\overline{X}_n$ becomes more concentrated around the population mean μ as the sample size n gets large.

Law of large numbers

Proof.

We want to show:

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| \geq \epsilon) = 0, \forall \epsilon > 0.$$

We will use Chebyshev's inequality:

$$P(|X - E(X)| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}, \forall \epsilon > 0.$$

First, we compute $E(\overline{X}_n)$ and $Var(\overline{X}_n)$:

$$E(\overline{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{\cancel{n}}\cancel{n}\mu =$$

Law of large numbers

Proof contd.

$$\begin{aligned} Var(\overline{X}_n) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^{\cancel{2}}}\cancel{n}\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

We substitute $E(\overline{X}_n)$ and $Var(\overline{X}_n)$ in Chebyshev's inequality $P(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$ and compute limit when $n \to \infty$:

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

$\square$

Law of large numbers

Example 5 (Coin flips)

Consider flipping a coin with a probability of heads p, i.e., $X_1 \sim Bernoulli(p)$ and we have $E(X_1) = p$ and $Var(X_1) = p(1-p)$. We draw a random sample $X_1, X_2, \ldots, X_n$ by flipping the coin n times. According to the WLLN $\overline{X}_n$ converges to p in probability. Suppose $p = 0.5$. How large should n be so that $P(0.4 < \overline{X}_n < 0.6) \geq 0.9$?

Law of large numbers

Solution.

We have $E(\overline{X}_n) = p = 0.5$ and $Var(\overline{X}_n) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{0.25}{n}$. From Chebyshev's inequality:

$$\begin{aligned} P(0.4 < \overline{X}_n < 0.6) &= P(|\overline{X}_n - 0.5| < 0.1) = 1 - P(|\overline{X}_n - 0.5| \geq 0.1) \\ &\geq 1 - \frac{0.25}{0.1^2 n} = 1 - \frac{25}{n}. \end{aligned}$$

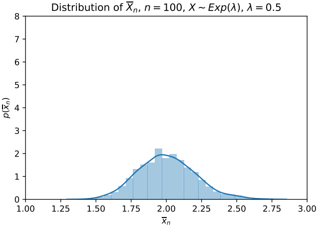
For 0.9 we get $n = 250$.

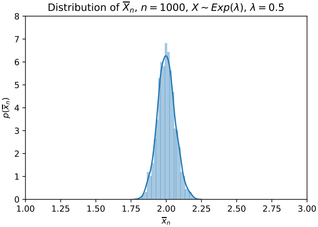
Notebook 5 (WLLN for Bernoulli r.v.)

`wlln_bernoulli.ipynb`

Central Limit Theorem

- WLLN says that the sampling distribution of $\overline{X}_n$ concentrates near μ
- To approximate probability statements about $\overline{X}_n$ we need further properties.
- Intuition from our example: the distribution of $\overline{X}_n$ is approx. Normal
- Theoretical result: **Central Limit Theorem**





Central Limit Theorem

Theorem 4 (Central Limit Theorem)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the density $f(\cdot)$ and let $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$ be finite. Then:

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

In other words, the CDF $F_n(z)$ of Z_n tends to CDF of a standard normal r.v. for $n \to \infty$:

$$\lim_{n\to\infty} F_n(z) = \lim_{n\to\infty} P(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$$

Remark 15 (Interpretation of CLT)

Probability statements about $\overline{X}_n$ can be approximated using a Normal distribution.

Central Limit Theorem

- We will skip the proof
- Proof involves e.g. characteristic functions

Example 6 (Software bugs)

Suppose we have a source code for a software program consisting of $n = 100$ files of code. Let X_i be the number of errors in the i^{th} file. Suppose that the $X_i \sim Pois(1)$ and that they are independent. Let $Y = \sum_{i=1}^{n} X_i$ be the total number of errors in the software program. Use the central limit theorem to approximate $P(Y \leq 80)$.

Central Limit Theorem

Solution.

For a $X \sim Poisson(\lambda)$ we have $E(X) = \lambda$ and $Var(X) = \lambda$. We are interested in $Y = \sum_{i=1}^{n} X_i = n\overline{X}_n$. We know from CLT that $\overline{X}_n \approx N(\mu, \frac{\sigma^2}{n})$. Thus, we have:

$$Y = n\overline{X}_n \approx N(n\mu, n\sigma^2)$$

For $\lambda = 1$ we obtain $Y \approx N(100, 100)$ and:

$$P(Y \leq 80) \approx \Phi\left(\frac{80 - 100}{10}\right) = 0.028.$$

Notebook 6 (CLT for Poisson r.v.)

`clt_poisson.ipynb`

Central Limit Theorem

Remark 16 (Main result of CLT revisited)

The CLT does not say anything about the form of the original density function $f(\cdot)$. Whatever the distribution function (given a finite variance), the sample mean is approximately normally distributed for large samples. In other words, given a random sample from any $f(\cdot)$ with finite σ^2:

$$\overline{X}_n \approx N(\mu, \frac{\sigma^2}{n}).$$

Readings

Fahrmeir et. al, Statistik: der Weg zur Datenanalyse (Chapters 1.4, 7)

Wasserman, All of Statistics (Chapter 5)

Sampling Distribution of $\overline{X}_n$ for $X \sim N(\mu, \sigma^2)$

Theorem 5 (Sampling distribution of $\overline{X}_n$ for Normal distribution)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n for $X_1 \sim N(\mu, \sigma^2)$. Then:

$$\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n}).$$

Proof.

The moment-generating function (alternative specification of the density $f(\cdot)$) of a r.v. X is: $M_X(t) = E(e^{tX}), t \in \mathbb{R}$. For $X \sim N(\mu, \sigma^2)$: $M_X(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$. For $\overline{X}_n$ we have:

$$M_{\overline{X}_n}(t) = E(e^{t\overline{X}_n}) = E(e^{\frac{t}{n}\sum_{i=1}^{n} X_i}) = E(\prod_{i=1}^{n} e^{\frac{t}{n}X_i})$$

Sampling Distribution of $\overline{X}_n$ for $X \sim N(\mu, \sigma^2)$

Proof contd.

Because $X_i's$ are i.i.d and their joint density factors in the product of the individual densities, the expectation of their products factors in the product of expectations:

$$
\begin{aligned}
M_{\overline{X}_n}(t) &= \cdots = \prod_{i=1}^{n} E(e^{\frac{t}{n}X_i}) \\
&= \prod_{i=1}^{n} M_{X_i}(\frac{t}{n}) = \prod_{i=1}^{n} e^{\frac{t\mu}{n} + \frac{1}{2}\left(\frac{\sigma t}{n}\right)^2} \\
&= e^{t\mu + \frac{1}{2}\frac{(\sigma t)^2}{n}},
\end{aligned}
$$

which is the moment-generating function for $N(\mu, \frac{\sigma^2}{n})$. □

Radar Guns

Example 7 (Radar guns)

Suppose n radar guns are set up along a stretch of road to catch people driving over the speed limit. Each radar gun is known to have a normal measurement error $N(0, \sigma^2)$. For a car passing at speed μ, let $\overline{X}_n$ be the average of the n readings. Simulate and plot the distribution of $\overline{X}_n$ for $\mu = 45km/h$, $n = 2$ radars and $\sigma = 5km/h$. Now suppose that error of the each radar gun is distributed uniformly with $E(X) = 0$ and $Var(X) = \sigma^2$. Using the same values for μ, n and σ as previously simulate and plot the distribution of $\overline{X}_n$ with the uniform measurement error. Compare and discuss the results from the two experiments.

Radar Guns

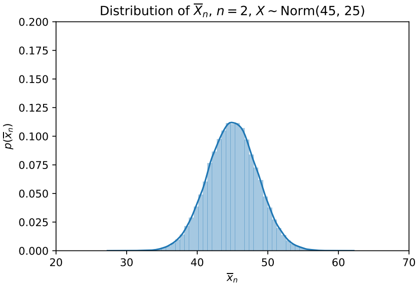
Solution.

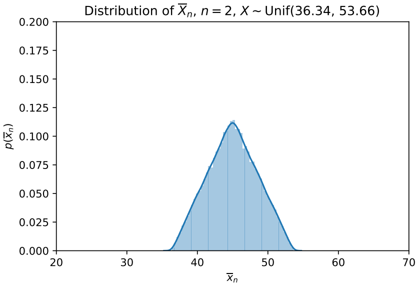
For a r.v. $X \sim Unif(a, b)$ we first need to compute a and b. We have $E(X) = 0$ and $Var(X) = \sigma^2$:

$$\begin{aligned} E(X) &= \frac{1}{2}(a+b) = 0 \implies a = -b \\ Var(X) &= \frac{1}{12}(b-a)^2 = \frac{1}{12}(2b)^2 = \sigma^2 \\ &\implies b = \sqrt{3}\sigma, a = -\sqrt{3}\sigma \end{aligned}$$

Thus, for a normal measurement error we draw random samples of size 2 from $N(45, 25)$ and for the uniform measurement error we simulate from $Unif(45 - 5\sqrt{3}, 45 + 5\sqrt{3})$.

Radar Guns





Notebook 7 (Radar Guns)

`radar_guns.ipynb`

Sample Variance

- In many applications we do not know μ and σ^2
- To estimate μ we compute $\overline{X}_n$
- We estimate σ^2 with *sample variance* S_n^2

Definition 10 (Sample variance)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the density $f(\cdot)$. *Sample variance* is defined as:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

CLT with S_n^2

Theorem 6 (CLT with sample variance)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the density $f(\cdot)$ and let $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$ be finite. Then:

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \approx N(0,1).$$

Proof.

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}{\frac{S_n/\cancel{\sqrt{n}}}{\sigma/\cancel{\sqrt{n}}}} = \frac{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}{S_n/\sigma}$$

We know that $\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$ and that $S_n^2/\sigma^2 \xrightarrow{P} 1$ for large n due to the consistency of S_n^2 estimator. It follows that $\frac{\overline{X}_n-\mu}{S_n/\sqrt{n}} \approx N(0,1)$ □

Readings

Fahrmeir et. al, Statistik: der Weg zur Datenanalyse (Chapters 9.4)

Wasserman, All of Statistics (Chapter 6)

Point Estimators

- $\overline{X}_n$ and S_n^2 are examples for *point estimators*
- They are also statistics and have probability density functions
- What is the probability that such point estimates equal the true value of parameter?

Point Estimators

- $\overline{X}_n$ and S_n^2 are examples for *point estimators*
- They are also statistics and have probability density functions
- What is the probability that such point estimates equal the true value of parameter?
- This probability is 0!
- The probability that a continuous r.v. equals any value is 0!
- **Always accompany point estimates with some measure of error!**

Confidence Intervals

- Informally, our estimates will take the form of e.g., $\overline{X}_n \pm T_n$
- As previously T_n is a statistic and hence a r.v.
- We obtain in this way an interval that itself is a r.v.
- Continuous intervals will have probabilities different than 0
- We will call such intervals *confidence intervals*

Radar Guns Revisited

Example 8 (Radar guns revisited)

Suppose $n = 4$ radar guns are set up along a stretch of road to catch people driving over the speed limit. Each radar gun is known to have a normal measurement error $N(0, \sigma^2)$, $\sigma = 5km/h$. For a car passing at speed μ four readings are $(45.71, 47.41, 40.95, 50.65)$. Compute a random interval that covers the true unknown car speed μ with probability of 0.95.

Radar Guns Revisited

Solution.

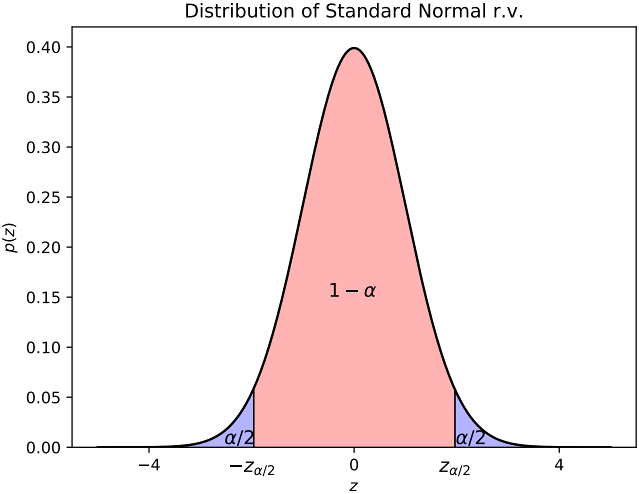
We know that $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ since $X_1 \sim N(\mu, \sigma^2)$. Therefore:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Recall that $z_{\alpha/2}$ is the Z-value such that the area to the right of it under the standard normal curve is $\alpha/2$: $P(Z \geq z_{\alpha/2}) = \alpha/2$.
Due to symmetry $-z_{\alpha/2}$ is the Z-value such that the area to the left of it under the standard normal curve is $\alpha/2$: $P(Z \leq -z_{\alpha/2}) = \alpha/2$.
From $P(Z \geq z_{\alpha/2}) = \alpha/2$ and $P(Z \leq -z_{\alpha/2}) = \alpha/2$ we conclude that:

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Radar Guns Revisited



Radar Guns Revisited

Solution contd.

Continuing with $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$:

$$\begin{aligned}
P(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) &= 1 - \alpha \\
P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}_n - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(-\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha
\end{aligned}$$

For our values, we get $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, $\overline{X}_n = 46.18$ and the 95% confidence interval is $(41.28, 51.08)$.

Notebook 8 (Radar Guns Revisited)

`radar_guns_revisited.ipynb`

Confidence Intervals

- Last example introduced Z-score confidence intervals
- However, we typically define confidence intervals more generally

Definition 11 (Confidence Interval)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from density $f(\cdot)$. A $1-\alpha$ *confidence interval* for a parameter θ is an interval $C_n = (T_{n,1}, T_{n,2})$ with $T_{n,1} = t_1(X_1, X_2, \ldots, X_n)$ and $T_{n,2} = t_2(X_1, X_2, \ldots, X_n)$ such that:

$$P(\theta \in C_n) \geq 1 - \alpha.$$

Remark 17 (Interpretation of confidence intervals)

C_n traps θ with probability $1-\alpha$. **Important:** C_n is random, θ is fixed.

Notebook 9 (Interpretation of confidence interval)

`ci_interpretation.ipynb`

T-Score Confidence Interval

- In case that σ^2 is unknown we can replace it with S_n^2 for large n
- We know that for large n: $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \approx N(0,1)$
- In practice, already for $n \geq 30$ we can compute Z-score confidence intervals
- But, what do we do for small $n < 30$?

T-Score Confidence Interval

- In case that we have a normal population, e.g., physical, biological systems, measurements, etc.
- If $X_1 \sim N(\mu, \sigma^2)$ then $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$
- t distribution with $n-1$ degrees of freedom
- For normal population and small n we can use $t_{\alpha/2}$ instead of $z_{\alpha/2}$
- We obtain then $1-\alpha$ T-score confidence interval: $\overline{X}_n \pm t_{\alpha/2}\frac{S_n}{\sqrt{n}}$

Radar Guns Revisited Once More

Example 9 (Radar guns revisited once more)

Suppose $n = 4$ radar guns are set up along a stretch of road to catch people driving over the speed limit. Each radar gun is known to have a normal measurement error $N(0, \sigma^2)$ with σ^2 unknown. For a car passing at speed μ four readings are $(45.71, 47.41, 40.95, 50.65)$. Compute a random interval that covers the true unknown car speed μ with probability of 0.95.

Radar Guns Revisited Once More

Solution.

For our values, we get $\alpha = 0.05$, $S_n = 4.04$, $t_{\alpha/2} = 3.18$, $\overline{X}_n = 46.18$ and the 95% confidence interval is $(39.74, 52.62)$.

Notebook 5 (Radar Guns Revisited Once More)

`radar_guns_revisited.ipynb`

Non-Normal Population and Alternative Estimators

- For $n < 30$ and non-normal populations we use T-score confidence intervals with extreme caution
- Alternatively, we can use non-parametric or computational methods
- Our discussion so far concentrated on $\overline{X}_n$
- But what happens if we want to estimate alternative parameters, e.g. median md
- There are theoretical results for the sampling distribution of the sample median but they are much more involved than results for $\overline{X}_n$
- In many of these cases we can resort to **computer simulation**

Simulation a.k.a. Monte Carlo Method

- In some of our python notebooks we already used **simulation**
- This method is also called **Monte Carlo method**
- Why is it ok to use simulation to estimate quantities of interest?

Simulation a.k.a. Monte Carlo Method

- In some of our python notebooks we already used **simulation**
- This method is also called **Monte Carlo method**
- Why is it ok to use simulation to estimate quantities of interest?
- (Weak) Law of large numbers

$$\overline{X}_b = \frac{1}{b}\sum_{i=1}^{b} X_i \xrightarrow{P} E(X) \text{ as } b \to \infty.$$

- For large enough b we can use $\overline{X}_b$ to approximate $E(X)$
- **Good thing**: in simulation we can make b as large as we like

Simulation a.k.a. Monte Carlo Method

- **Even better**: simulation also works in a more general case
- If h is any function with finite mean:

$$\frac{1}{b}\sum_{i=1}^{b} h(X_i) \xrightarrow{P} E(h(X)) \text{ as } b \to \infty.$$

- For example, to estimate the variance $Var(X)$:

$$\frac{1}{b}\sum_{i=1}^{b} (X_i - \overline{X}_b)^2 \xrightarrow{P} Var(X) \text{ as } b \to \infty.$$

- Prerequisite for simulation is that we can draw random samples from some target distribution $F(\cdot)$
- In some applications we know $F(\cdot)$ but generally we do not know it

Readings

Wasserman, All of Statistics (Chapter 8)

Givens and Hoeting, Computational Statistics (Chapter 9)

Chernick, Bootstrap Methods: A Guide for Practitioners and Researchers (various Chapters)

Bootstrap Principle

- The **bootstrap** is a simulation method for estimating variation of point estimates and computing confidence intervals
- Let $T_n = t(X_1, X_2, \ldots, X_n)$ be a statistic, with $X_1 \sim F(\cdot)$
- Suppose we want to know $Var_F(T_n)$
- Basic bootstrap idea has the following three steps:
  1. Estimate $F(\cdot)$ from a random sample $X_1, X_2, \ldots, X_n$ to obtain empirical distribution function $F_n(\cdot)$
  2. Simulate from $F_n(\cdot)$ to obtain b *bootstrap samples*
  3. Approximate $Var_F(T_n)$ with $Var_{F_n}(T_n)$

Bootstrap Principle

$$\begin{array}{rccccc} \text{Real} & F & \Longrightarrow & X_1, X_2, \ldots, X_n & \Longrightarrow & T_n = t(X_1, X_2, \ldots, X_n \\ \text{Bootstrap} & F_n & \Longrightarrow & X_1^*, X_2^*, \ldots, X_n^* & \Longrightarrow & T_n^* = t(X_1^*, X_2^*, \ldots, X_n^* \end{array}$$

- To estimate $F_n(\cdot)$ we put $1/n$ probability over each data point from $X_1, X_2, \ldots, X_n$
- Therefore, drawing an observation from $F_n(\cdot)$ is equivalent to drawing one point at random from the original random sample
- **Important:** we always draw at random with replacement
- **Important:** the size of a single bootstrap sample equals the size of the original random sample n

Bootstrap Example

Example 10 (6-sided die)

Suppose we roll a 6-sided die $n = 10$ times and get the following data, written in increasing order:

$$1, 1, 2, 3, 3, 3, 4, 5, 6, 6.$$

Imagine writing these values on 10 pieces of paper, putting them in a hat and drawing one at random. Then, e.g., probability of drawing a 3 is $3/10$ and of 4 is $1/10$. We can put the full *empirical distribution* P_{10} in a probability table:

x	1	2	3	4	5	6
$P_{10}(x)$	$2/10$	$1/10$	$3/10$	$1/10$	$1/10$	$2/10$

Bootstrap Example

Example 10 (6-sided die contd.)

Notice the difference between the true distribution P and the empirical distribution P_{10}.

x	1	2	3	4	5	6
$P(x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$
$P_{10}(x)$	$2/10$	$1/10$	$3/10$	$1/10$	$1/10$	$2/10$

Remark: By the WLLN we know that $P_n(x) \xrightarrow{P} P(x), \forall x, n \to \infty$. Since in practice we often do not know P we resample from P_n:

1. We draw from P_n with replacements! We draw a piece of paper from the hat at random, then we **put back** this piece of paper in the hat.
2. We draw in this way n times to obtain a single bootstrap sample
3. We repeat steps 1 and 2 b times to obtain b bootstrap samples

Bootstrap Example

Example 10 (6-sided die contd.)

For example, for $b = 6$ we may obtain:

$$
\begin{aligned}
X_1, X_2, \ldots, X_{10} &: 1, 1, 2, 3, 3, 3, 4, 5, 6, 6. \\
X^*_{1,1}, X^*_{2,1}, \ldots, X^*_{10,1} &: 1, 1, 1, 3, 3, 4, 4, 5, 5, 6. \\
X^*_{1,2}, X^*_{2,2}, \ldots, X^*_{10,2} &: 1, 2, 3, 3, 3, 3, 4, 4, 5, 6. \\
X^*_{1,3}, X^*_{2,3}, \ldots, X^*_{10,3} &: 2, 2, 3, 3, 3, 3, 5, 5, 5, 6. \\
X^*_{1,4}, X^*_{2,4}, \ldots, X^*_{10,4} &: 1, 2, 3, 4, 4, 4, 4, 6, 6, 6. \\
X^*_{1,5}, X^*_{2,5}, \ldots, X^*_{10,5} &: 1, 1, 2, 3, 3, 3, 4, 4, 5, 6. \\
X^*_{1,6}, X^*_{2,6}, \ldots, X^*_{10,6} &: 1, 2, 2, 3, 3, 3, 4, 6, 6, 6.
\end{aligned}
$$

Bootstrap Algorithm

1. Estimate F_n from $X_1, X_2, \ldots, X_n$ by putting $1/n$ probability over each data point (this is the maximum likelihood estimator for F)
2. Draw $X_1^*, X_2^*, \ldots, X_n^* \sim F_n$
3. Compute $T_{n,i}^* = t(X_1^*, X_2^*, \ldots, X_n^*)$
4. Repeat steps 2 and 3 b times
5. Approximate variability of T_n by the variability of T_n^*

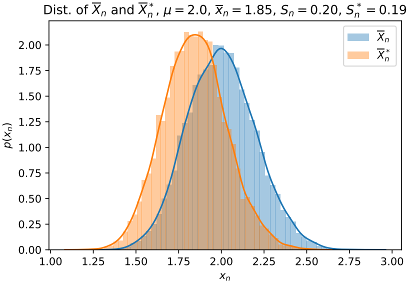
Bootstrap: Why Does It Work?

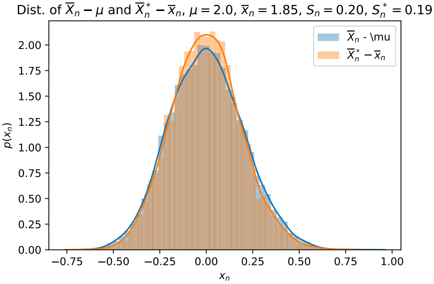
- Distribution of T_n is centered at θ (the parameter that T_n estimates)
- Likewise, the distribution of T_n^* is centered at t_n (the concrete realization of T_n for the original sample)
- If t_n and θ differ significantly than the two distributions are also significantly different
- But the two distributions of the variation around the center are approximately equal: $T_n - \theta$ and $T_n^* - t_n$
- Alternatively, the bootstrap works because
  1. $F_n \approx F$
  2. The amounts of variation in the estimates with F_n and F are similar

Notebook 7 (Bootstrap Principle)

`edf.ipynb`

Bootstrap: Why Does it Work?





Bootstrap Confidence Intervals

- Let us define two new r.v. $\delta_n = T_n - \theta$ and $\delta_n^* = T_n^* - t_n$
- Bootstrap principle states that $\delta_n \approx \delta_n^*$
- If we would know the distribution of δ_n then $1-\alpha$ confidence intervals for T_n are:

$$
\begin{aligned}
P(\delta_{n_{\alpha/2}} < T_n - \theta < \delta_{n_{1-\alpha/2}}) &= 1-\alpha \\
P(\delta_{n_{\alpha/2}} - T_n < -\theta < \delta_{n_{1-\alpha/2}} - T_n) &= 1-\alpha \\
P(T_n - \delta_{n_{1-\alpha/2}} < \theta < T_n - \delta_{n_{\alpha/2}}) &= 1-\alpha
\end{aligned}
$$

- Since we do not know the distribution of δ_n but we can compute the distribution of δ_n^* we replace the former with the latter to obtain *bootstrap confidence intervals*:

$$P(T_n - \delta^*_{n_{1-\alpha/2}} < \theta < T_n - \delta^*_{n_{\alpha/2}}) = 1-\alpha$$

Radar Guns Again

Example 11 (Radar guns again)

Suppose $n = 4$ radar guns are set up along a stretch of road to catch people driving over the speed limit. For a car passing at speed μ four readings are $(45.71, 47.41, 40.95, 50.65)$. Compute a random interval that covers the true unknown car speed μ with probability of 0.95.

Notebook 5 (Radar Guns Again)

`radar_guns_revisited.ipynb`

Bootstrapping Alternative Statistics

- Bootstrap is not restricted to estimating confidence intervals of the sample mean
- In fact, bootstrap can be used to estimate the distribution of almost any statistics

Example 12 (Skewness of the game sales)

The skewness is a measure of asymmetry of a distribution. A normal distribution, which is symmetric, has skewness 0. Estimate the confidence interval for the skewness of game sales in North America using the Game Sales dataset from Kaggle:
`https://www.kaggle.com/gregorut/videogamesales`.

Notebook 8 (Skewness of the game sales)

`bootstrap_skewness.ipynb`

Bootstrapping Alternative Statistics

Example 13 (Bioequivalence)

This example is from Efron and Tibshirani (inventors of bootstrap), An Introduction to the Bootstrap. When drug companies introduce new medications, they need to show bioequivalence, i.e., that the new drug is not substantially different than the current one. Here are data on eight subjects who used hormone infusing drug. Each subject received three treatments: placebo, old, new. Let $X =$old-placebo and $Y =$new-old. The bioequivalence is given if $|\theta| \leq 0.2$, where:

$$\theta = \frac{E_F(Y)}{E_F(X)}.$$

Answer the bioeqivalence question with bootstrap. The data is given in the table on the next slide.

Bootstrapping Alternative Statistics

Example 13 (Bioequivalence contd.)

placebo	old	new
9243	17649	16449
9671	12013	14614
11792	19979	17274
13357	21816	23798
9055	13850	12560
6290	9806	10157
12412	17208	16570
18806	29044	26325

Notebook 9 (Bioequivalence)

`bioequivalence.ipynb`

Summary

- Sampling
- Random sample, sampled population, statistic, sampling distribution.
- Sampling distribution of the statistic: tells us how close the statistic is to the parameter.
- Law of large numbers and Central Limit Theorem
- Confidence intervals
- Bootstrapping