Frequentist Inference, pt. I

ECON 3640–001

Marcio Santetti
Spring 2022

Motivation

Housekeeping

Notes based on `Keller (2009)`:

- Chapter **10**, sections `10.1` and `10.2`.

Motivation

When studying ***Bayesian*** inference, all our questions were answered through **entire probability distributions**.

In the case of ***frequentist*** inference, we appeal to the concept of **sampling distributions**.

But how to **link** sampling distributions to actual inference procedures?

Point and Interval estimators

Point and Interval estimators

The goal of statistical inference is to approximate a **population parameter** through a **sample statistic**.

For instance, the **estimator** of the population mean (μ) is the sample mean, $\bar{x}$.

Once it is computed, the specific value of $\bar{x}$ is an **estimate** of the population mean.

Careful! An **estimator** is a formula/procedure one follows to obtain a measure of a parameter of interest. An **estimate** is a specific value calculated using an estimator.

A sample statistic may be used to represent a population parameter in two ways:

1. Point estimator;
2. Interval estimator.

Point and Interval estimators

A **point** estimator draws inferences about a population by estimating the value of an unknown parameter using a **single** value or point.

In the first part of the course, we spent some time using such estimators.

- E.g., *sample mean, sample variance, sample standard deviation, sample median*....

These estimators, however, are **fragile**.

Recall that $P(X = x) = 0$ for continuous random variables!

Furthermore, how does *varying the sample size* reflects how good/bad a point estimator is?

Point and Interval estimators

Interval estimators draw inferences about a population by estimating the value of an unknown parameter using an *interval*.

This way, representing a parameter of interest through an interval is *better suited* when we don't know the whole population.

And the sample size (n) **does matter** here!

Point and Interval estimators

The selection of the sample statistic to be used as an estimator, however, depends on the **characteristics** of that statistic.

These are:

1. Unbiasedness;
2. Consistency;
3. Efficiency.

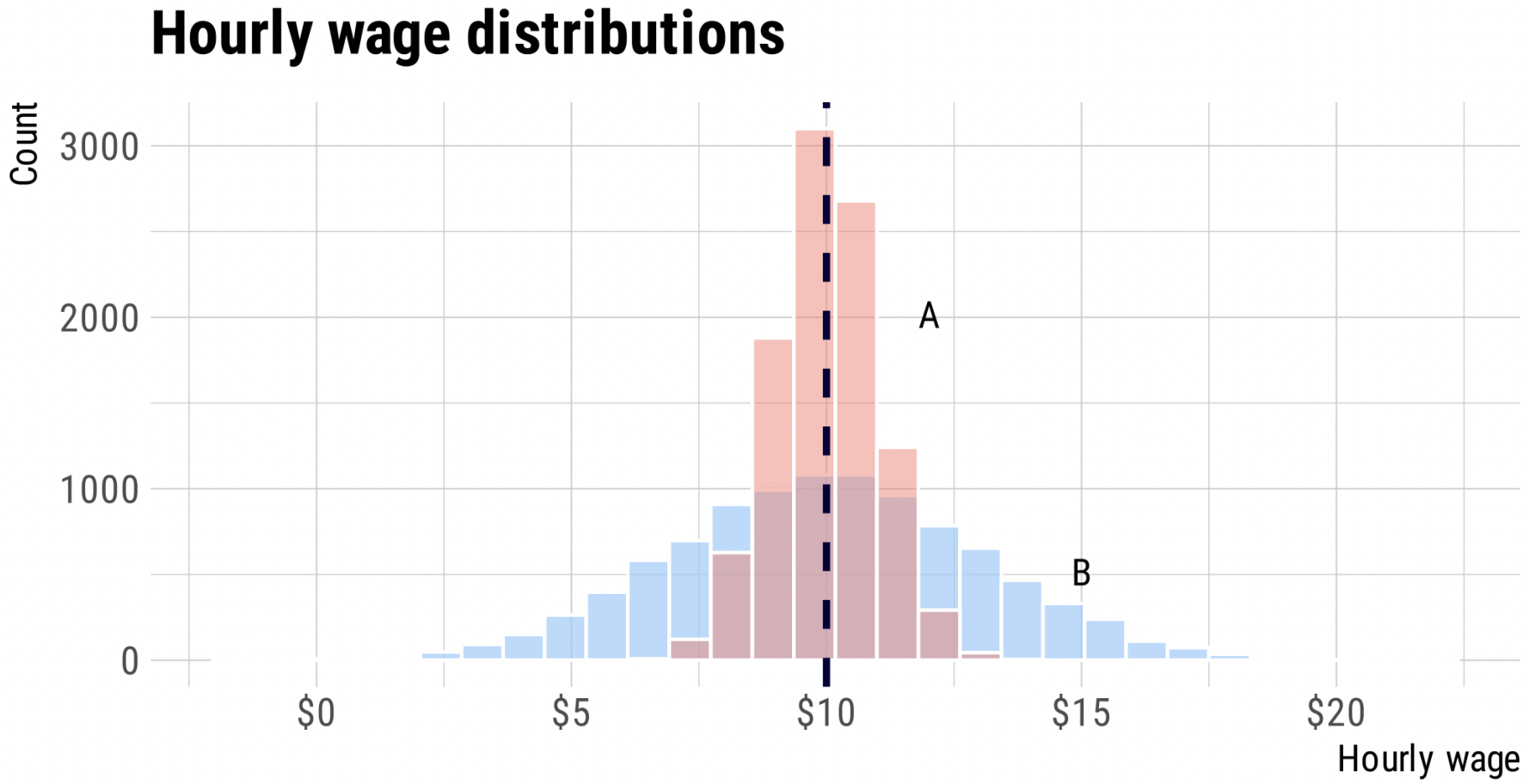
Point and Interval estimators

An **unbiased** estimator of a population parameter is an estimator whose *expected value is equal to that parameter*.

Unbiasedness implies that if one takes an **infinite** number of samples and calculate the value of the estimator in each sample, the *average value* of the estimators would *equal* the parameter.

If there are two unbiased estimators of a parameter, the one whose variance is smaller is said to have **relative efficiency**.

Point and Interval estimators



Point and Interval estimators

Intuition behind unbiasedness:

- If one were to take an **infinite** amount of samples from the population and compute the value of the estimator (mean, median, variance, standard deviation...) in each sample, the **average** value would equal the parameter.
- $E(\bar{X}) = \mu$.

A third desirable propery of an estimator is **consistency**.

It simply implies that, for an unbiased estimator, the difference between the estimator and the parameter grows smaller *as the sample size grows larger*.

- Therefore, consistency is a "large sample" property.

Building interval estimators

Building interval estimators

Now that we know where **sampling distributions** come from and the desired **properties** of an estimator, we may move on to one inferential tool within frequentist inference:

- **Confidence intervals**

Assume the (unrealistic) scenario in which we are curious about a population's **mean value** (μ), and its **standard deviation** (σ) is known.

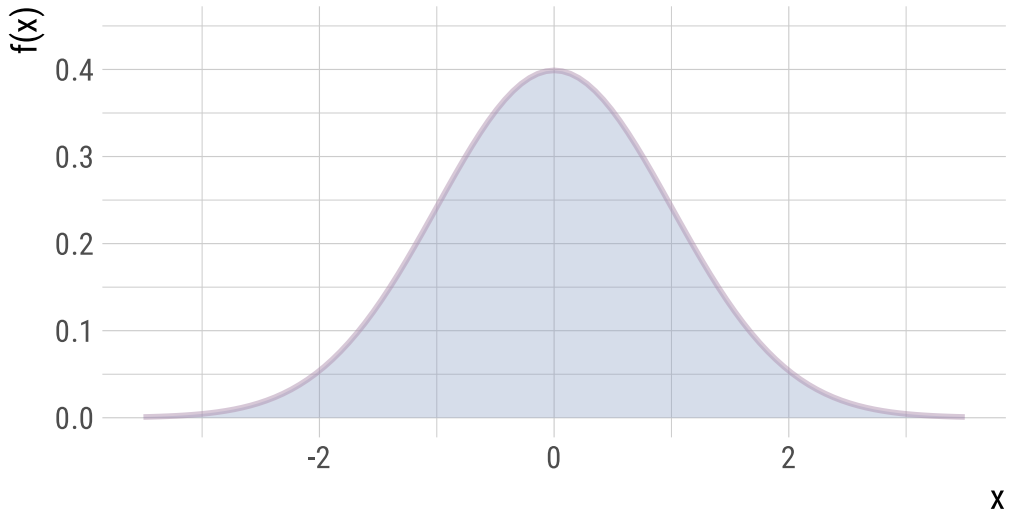
So, the task is to draw a random sample n from the population N and obtain its sample mean, $\bar{x}$.

From the **Central Limit Theorem**, if a random variable X is normally distributed, its mean $\bar{X}$ will **also** be normally distributed (or at least approximately).

Building interval estimators

As we've already studied, the **fundamental** parameters of the Normal distribution are the population **mean** and **standard deviation**.

$$X \sim \mathcal{N}(\mu, \sigma)$$



Building interval estimators

Given that the Normal distribution is **symmetric** about its **mean**, one very useful transformation we can apply to a normally distributed random variable is **standardization**.

This simply implies transforming a variable such that it follows a **Standard Normal** distribution.

- The Standard Normal distribution implies a mean of 0 and a standard deviation (variance) of 1.
- $X \sim \mathcal{N}(0, 1)$

To **standardize** a random variable, we apply the following formula:

$$z = \frac{x - \mu}{\sigma}$$

Building interval estimators

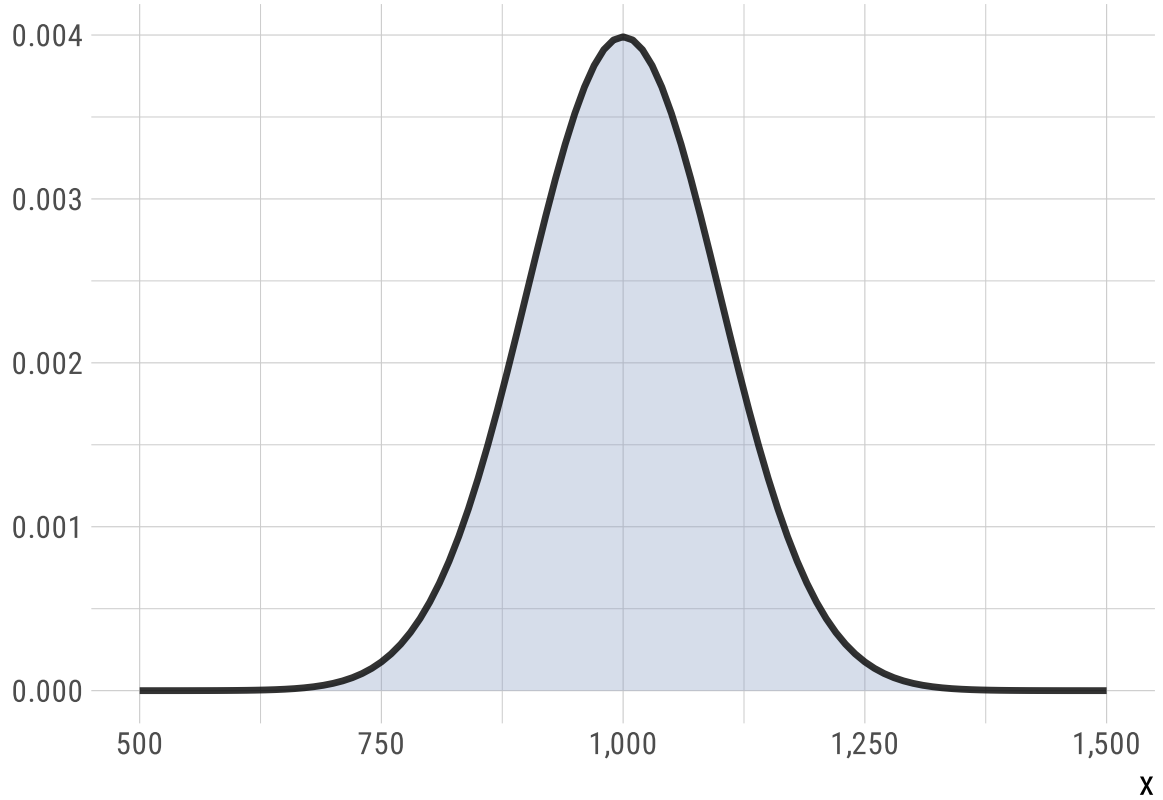
A quick example:

Suppose the daily demand for gasoline at a station is a normally distributed random variable with a mean of 1,000 and a standard deviation of 100 gallons.

The owner opened this station today and noted that there is exactly 1,100 gallons in storage. The next delivery will only happen tomorrow. She would like to know the probability that she will have enough regular gasoline to satisfy today's demand.

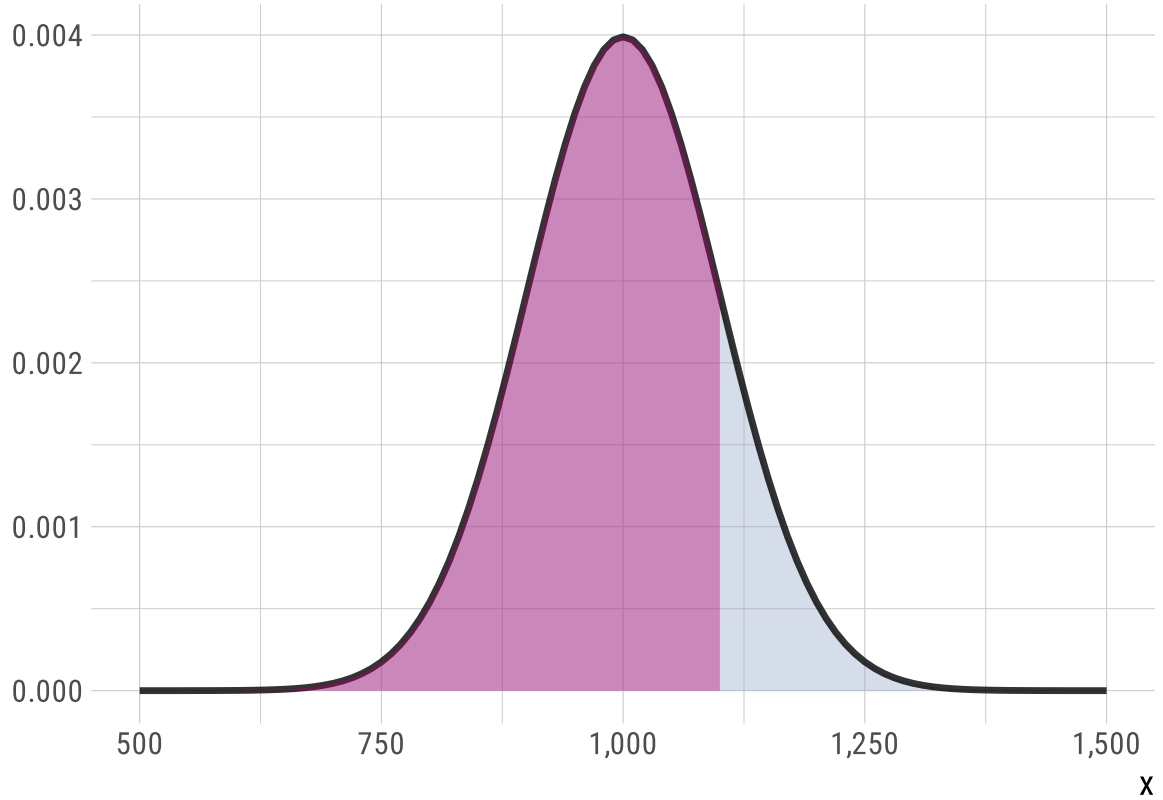
Building interval estimators

Visually:



Building interval estimators

And the area we are interested in is:



Building interval estimators

Let us standardize our random variable *X*:

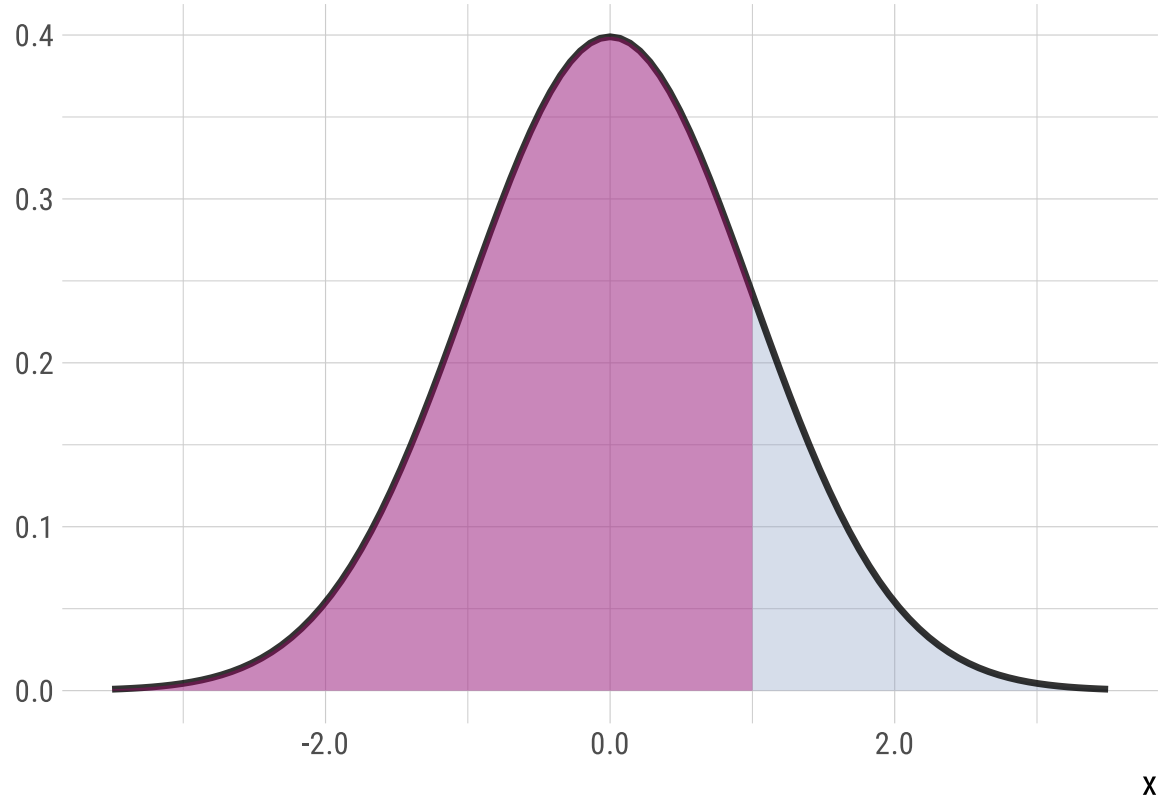
$$z = \frac{x - \mu}{\sigma} = \frac{1,100 - 1000}{100} = 1.00$$

Now, the random variable *z* follows a Standard Normal distribution!

And we are able to ask *P(z < 1.00)*, instead of *P(X < 1,100)*.

Building interval estimators

Visually:



Building interval estimators

The values of *z* specify the **location** of the corresponding value of *X*, our random variable.

A value of *z = 1.00* corresponds to a value of *X* that is **1.00 standard deviation above the mean**.

Another *advantage* of standardizing a random variable to a *z* value is that it automatically centers the population mean (μ) to **zero**.

So what is this probability?

```
pnorm(q = 1, mean = 0, sd = 1)
```

```
#> [1] 0.8413447
```

Which is the same as

```
pnorm(q = 1100, mean = 1000, sd = 100)
```

```
#> [1] 0.8413447
```

Building interval estimators

Now, we are ready to build a **confidence interval** for a **population mean** of interest.

Applying the standardization process to a sample mean $\bar{X}$, we have

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

And we want to get here:

$$P(\bar{x} - z_{\frac{\alpha}{2}}\, \sigma/\sqrt{n} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\, \sigma/\sqrt{n}) = 1 - \alpha$$

Building interval estimators

Again:

$$P(\bar{x} - z_{\frac{\alpha}{2}} \sigma/\sqrt{n} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \sigma/\sqrt{n}) = 1 - \alpha$$

The **left-hand side** is a probability statement, which considers the probability with which the **population mean** (μ) lies between two values of the **sample mean**: one lower value, **corrected** downwards by a **standard error** ($z_{\alpha/2}$), multiplied by the **sampling distribution**'s standard deviation ($\sigma/\sqrt{n}$); and one upper value, **corrected** upwards by the same factor.

In other words, the above says that, with **repeated sampling** from this population, the proportion of values of $\bar{X}$ for which the interval $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}; \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}]$ **includes** the population mean μ is **equal to** $1 - \alpha$.

Building interval estimators

This form of probability statement is called the confidence interval estimator of μ.

The left part of the inequality on the left-hand side is known as the **lower confidence limit** (LCL); while the right portion of the inequality is the **upper confidence limit** (UCL).

The right-hand side, $1 - \alpha$, is the **confidence level** assumed for the confidence interval.

- The latter is usually pre-specified, and represents the probability that the interval includes the actual value of μ.

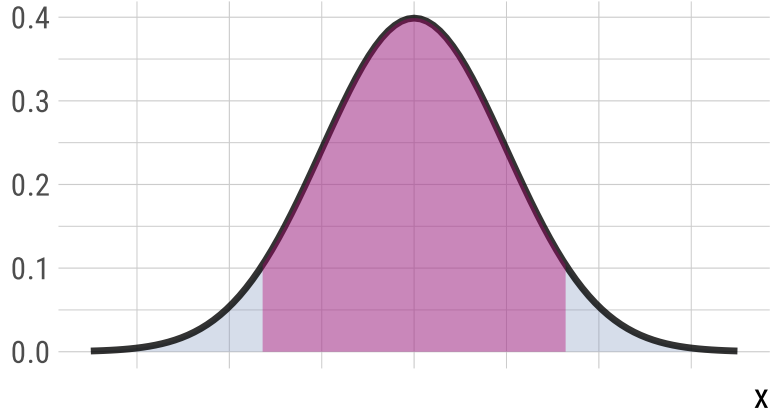
Building interval estimators

An example:

A computer store manager would like to estimate, with 95% of confidence, the optimum average inventory level. She also knows that the overall standard deviation is 75 computers.

She has used a random sample of 25 periods, calculating a sample mean of 370.16 computers.

What is the **area** we are interested in?



Next time: Hypothesis testing