STAT 210
Applied Statistics and Data Analysis
Week 4 - Summary

Joaquín Ortega
KAUST

Reminders

The deadline for sending an email with teams and subject for the project is Friday, Sept. 30.

The midterm evaluation for this course in on. Please give your opinion.

Videos 12 & 13: One sample problems

Speed of light experiment

- The speed of light experiment: Are the results consistent with the accepted value?
- Sample parameters:
  - sample mean $\hat{\mu}_n$,
  - variance s_n^2
  - standard deviation s_n.
- $\hat{\mu}_n = 909$, $\mu = 990$.
- **The sample mean is a random variable**: its distribution is known as the sampling distribution (for the empirical mean).

First alternative:

Assume that the data come from a normal distribution. Then the sampling distribution for $\hat{\mu}_n$ is also normal:

$$\hat{\mu}_n \sim N(\mu, \sigma^2/n).$$

Accepted value $\mu = 990$ Observed average $\hat{\mu}_n = 909$

Difference $= 81$

We want

$$P(|\hat{\mu}_n - \mu| \geq 81).$$

Problem: σ^2 is unknown.

The first alternative is to approximate the unknown value σ^2 by the sample estimate s_n^2 and use the normal distribution:

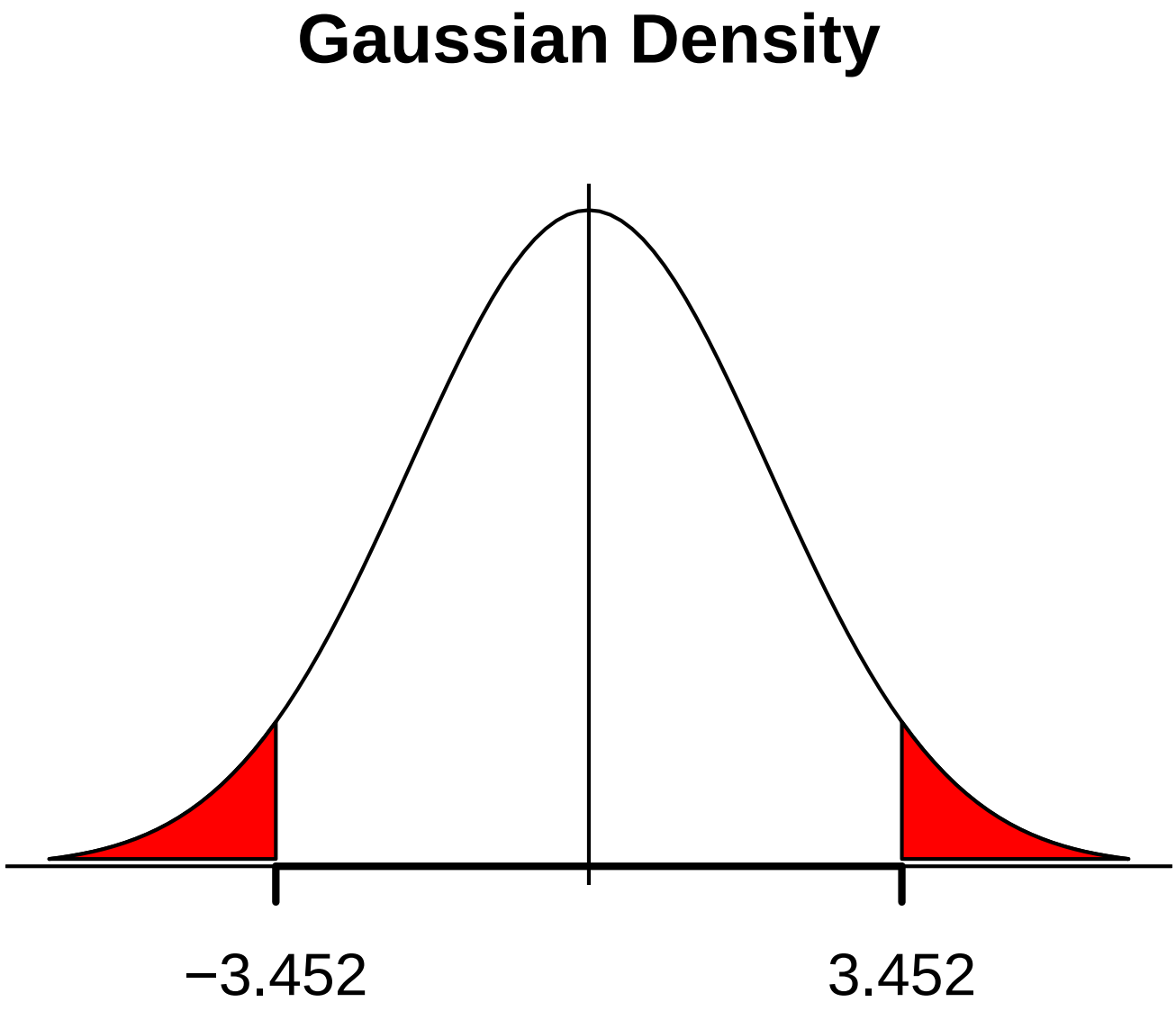
$$\hat{\mu}_n \approx N(\mu, s_n^2/n).$$

Using this value,

$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \approx N(0,1).$$

With R, we can calculate the probability we are interested in. Observe that

$$\begin{aligned} P(|\hat{\mu}_n - \mu| \geq 81) &= P(\frac{|\hat{\mu}_n - \mu|}{s_n/\sqrt{n}} \geq \frac{81}{s_n/\sqrt{n}}) \\ &= P(|Z| \geq 3.452) \\ &= 2P(Z \leq -3.452) \\ &= 0.00056 \end{aligned}$$


Gaussian Density
−3.452
3.452

Thus, the probability of observing a difference as large or larger than the one we found is about 0.00056.

This is a small number, so it would make us doubt the validity of the accepted value.

We have done a test of hypothesis. Assuming that the sample is Gaussian, we have tested the null hypothesis that the mean of the common distribution is 990 versus the alternative that it is not:

$$H_0 : \mu = 990 \qquad \textit{vs.} \qquad H_A : \mu \neq 990$$

This test is frequently known as a z-test.

Second Alternative: Student's t distribution

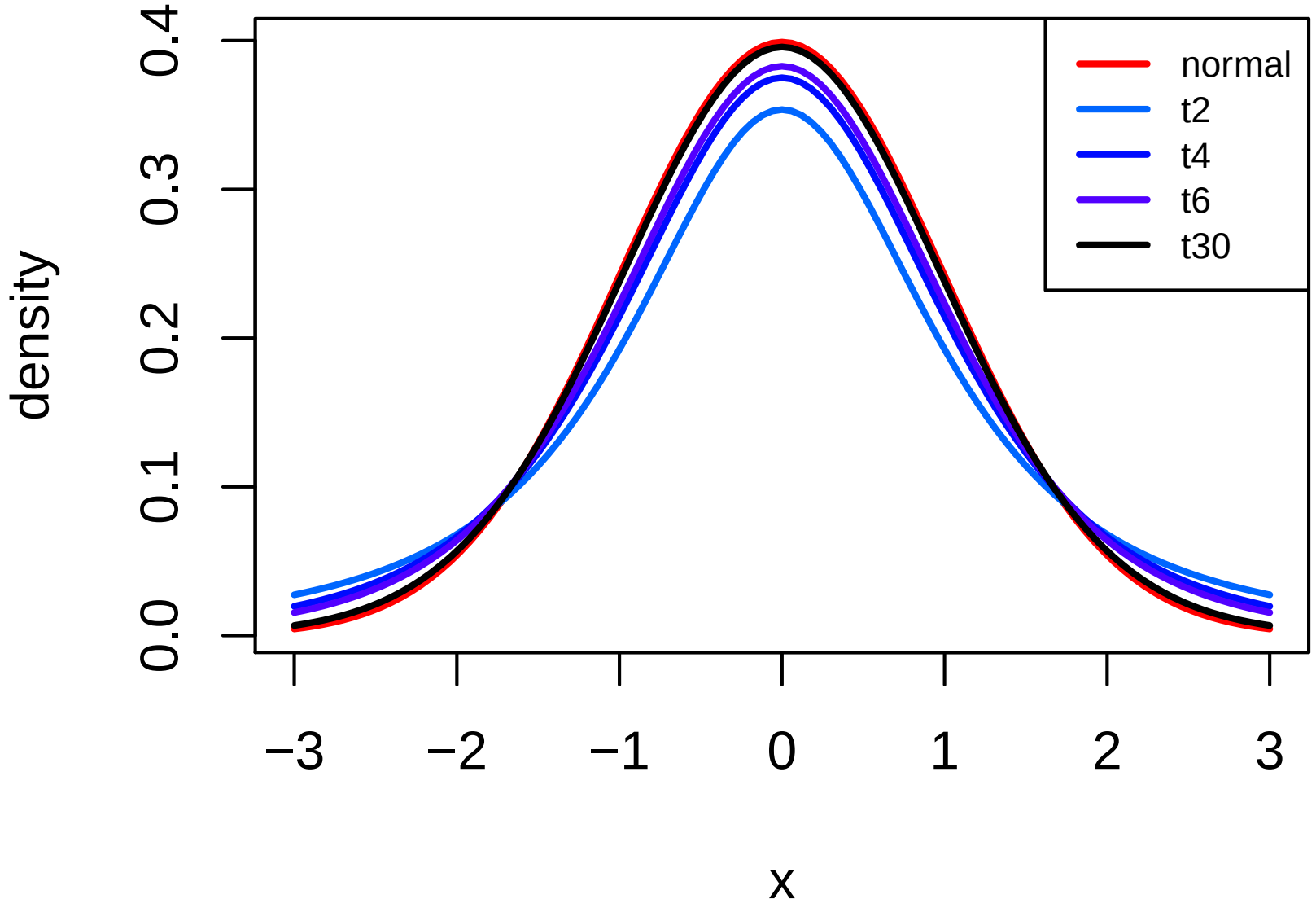
The t distribution is the sampling distribution of the empirical mean $\hat{\mu}_n$ when the data come from a normal distribution with unknown variance,

$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} denotes a t distribution with $n-1$ degrees of freedom.

For $n \geq 30$, the t distribution is very similar to the normal distribution but for n small there are important differences.

Student's t distribution



Thus, *under the null hypothesis*, i.e., assuming the null hypothesis is true, we have that

$$\frac{\hat{\mu}_n - 990}{s_n/\sqrt{n}} \sim t_{n-1}.$$

We want

$$\begin{aligned} P(|\hat{\mu}_n - 990| \geq 81) &= P(\frac{|\hat{\mu}_n - 990|}{23.46} \geq \frac{81}{23.46}) \\ &= P(|t_{19}| \geq 3.542) \\ &= 2P(t_{19} \leq -3.542) \\ &= 0.00267 \end{aligned}$$

The function `t.test` in `R` carries out this test.

```
t.test(mich.exp1, mu=990)
```

```
##
##  One Sample t-test
##
## data:  mich.exp1
## t = -3.4524, df = 19, p-value = 0.002669
## alternative hypothesis: true mean is not equal to 990
## 95 percent confidence interval:
##  859.8931 958.1069
## sample estimates:
## mean of x
##       909
```

Third Alternative: Non-parametric test

Some tests are **distribution-free**, i.e., they do not make distributional assumptions.

They are known as **non-parametric tests** because they are not based on the assumption of a parametric family of distributions.

Many non-parametric methods are based on **order statistics** and **ranks**.

Assume you have a sample $x_1, x_2, \ldots, x_n$ and that all values are different. The **order statistics** for this sample are the ordered values:

$$x_{(1)} < x_{(2)} < \cdots < x_{(n-1)} < x_{(n)}.$$

The **rank** is the position that a particular value has in the ordered sample.

Assume that your sample comes from a continuous distribution that is symmetric with respect to its average value μ.

By symmetry, it is equally likely that values will be above or below the mean.

Also, positive and negative differences of the same magnitude have the same probability of occurring.

If the sample values are $x_1, \ldots, x_n$ define

$$d_i = x_i - \mu_0$$

To compute the test statistic, follow these steps:

1. Take the absolute value of the d_i's.
2. Order these values and assign the ranks. If there are ties, use the midranks (average rank of the tied values).
3. Multiply the rank values obtained in step 2 by the original signs of the d_i's.
4. Sum the positive values in step 3 and denote the result by t^+.

t^+ is a sum of **ranks**, not of values, and is (the value of) the test statistic (we denote the corresponding r. v. by T^+).

If there are n values in the sample, the possible values of T^+ are between 0 (if all the values in the sample are negative) and

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

if all values are positive.

If the hypothesis of symmetry is valid, we would not expect very small or very large values of T^+.

If we observe either of these situations, we will reject the null hypothesis.

The distribution of T^+ is challenging to calculate, particularly if there are ties in the sample.

There is a normal approximation to the distribution that is frequently helpful.

The test can be carried out in `R` with the command `wilcox.test`.

```
wilcox.test(mich.exp1,mu=990)
```

```
## Warning in wilcox.test.default(mich.exp1, mu =
## 990): cannot compute exact p-value with ties

##
##  Wilcoxon signed rank test with continuity
##  correction
##
## data:  mich.exp1
## V = 22.5, p-value = 0.00213
## alternative hypothesis: true location is not equal to 990
```

Video 14: Pointwise estimation

Concepts Introduced:

Parametric family of distributions and parametric models

Random sample

Statistic and estimator

Sampling distribution

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E[\hat{\mu}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$$

$$Var(\hat{\mu}_n) = \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)\Big]$$

Since the variables in the sample are independent, $Cov(X_i, X_j) = 0$ if $i \neq j$, and

$$Var(\hat{\mu}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n}\sigma^2$$

If $X \sim N(\mu, \sigma^2)$ then $\hat{\mu}_n \sim N(\mu, \sigma^2/n)$.

Central Limit Theorem

Let $X_n, n \geq 1$ be a sequence of independent random variables having the same distribution with mean μ and variance σ^2. Then,

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

converges to a standard normal distribution, as $n \to \infty$. What this means is that for any number x, and n large,

$$P\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\Big) \approx \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

We denote this by

$$\bar{X}_n \approx N(\mu, \sigma^2/n)$$

Summary

Let $X_1, \ldots, X_n$ be iid rv's with mean μ and variance σ^2.

1. The sampling distribution of $\bar{X}_n$ has mean μ .
2. The standard deviation of the sampling distribution, known as the standard error, is $\sigma/\sqrt{n}$.
3. When n is large, the sampling distribution of $\bar{X}_n$ approaches the normal distribution, regardless of the distribution of the population.
4. When the population distribution is normal, so is the sampling distribution for any value of n.
5. If the variance is not known, and we estimate it from the sample, the sampling distribution for the normalized sample mean is t_{n-1}.
6. When n is small, and the population distribution is not normal, we cannot assume that the sampling distribution is normal.

A Rule of Thumb

For a sample size $n > 30$, the sampling distribution for the mean for any population with mean μ and variance σ^2 can be approximated by a normal distribution with mean μ and variance σ^2/n.

When n is small, and the population is not normal, all we can say is that

1. The mean of the sampling density of the mean equals μ, the mean of the population.
2. The variance of the sampling distribution of the mean is σ^2/n.

One way to proceed in this case is to use the bootstrap, a technique that will study later in this course.

Video 15: Interval estimation

Let α be a number in the interval $(0, 1)$, and let $Z \sim N(0, 1)$. Let z_α be the real number defined by the relation

$$P(Z > z_\alpha) = \alpha.$$

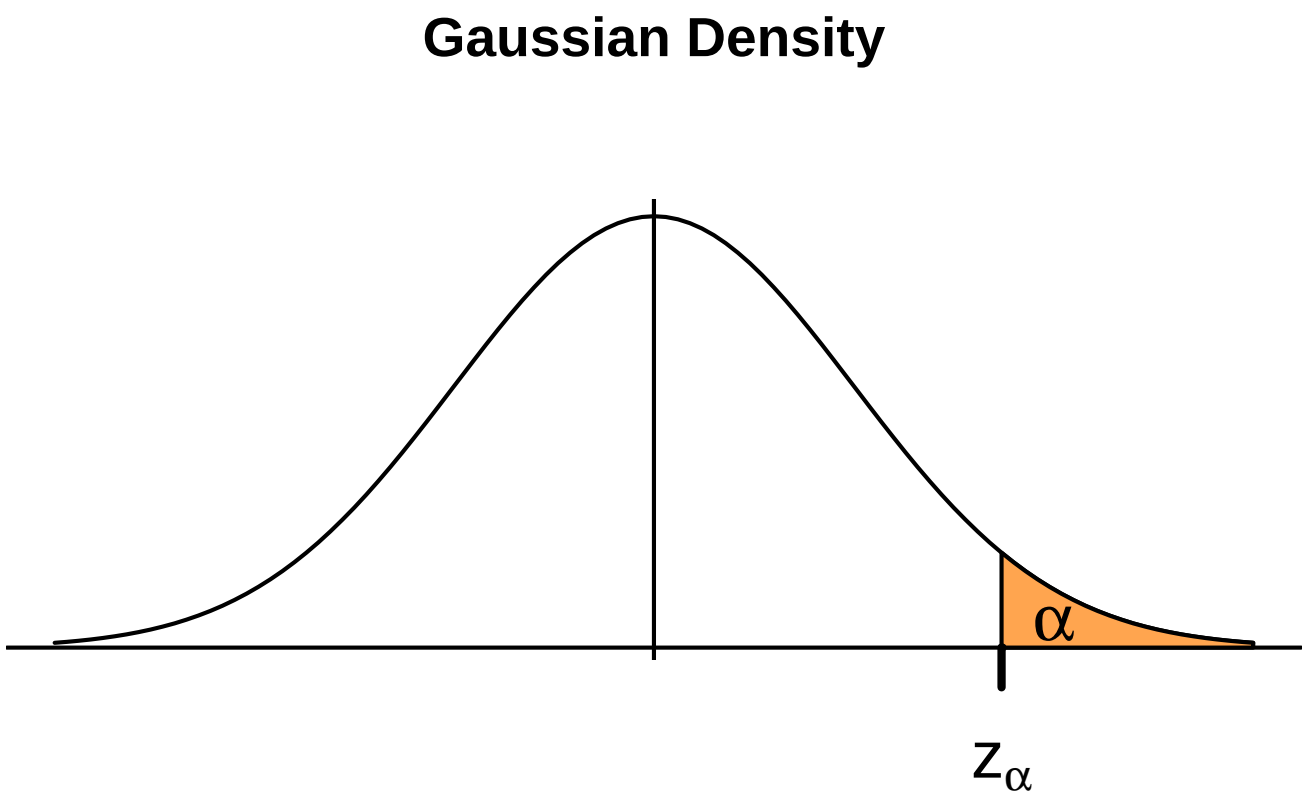


Figure 1: Definition of z_α for a Gaussian distribution

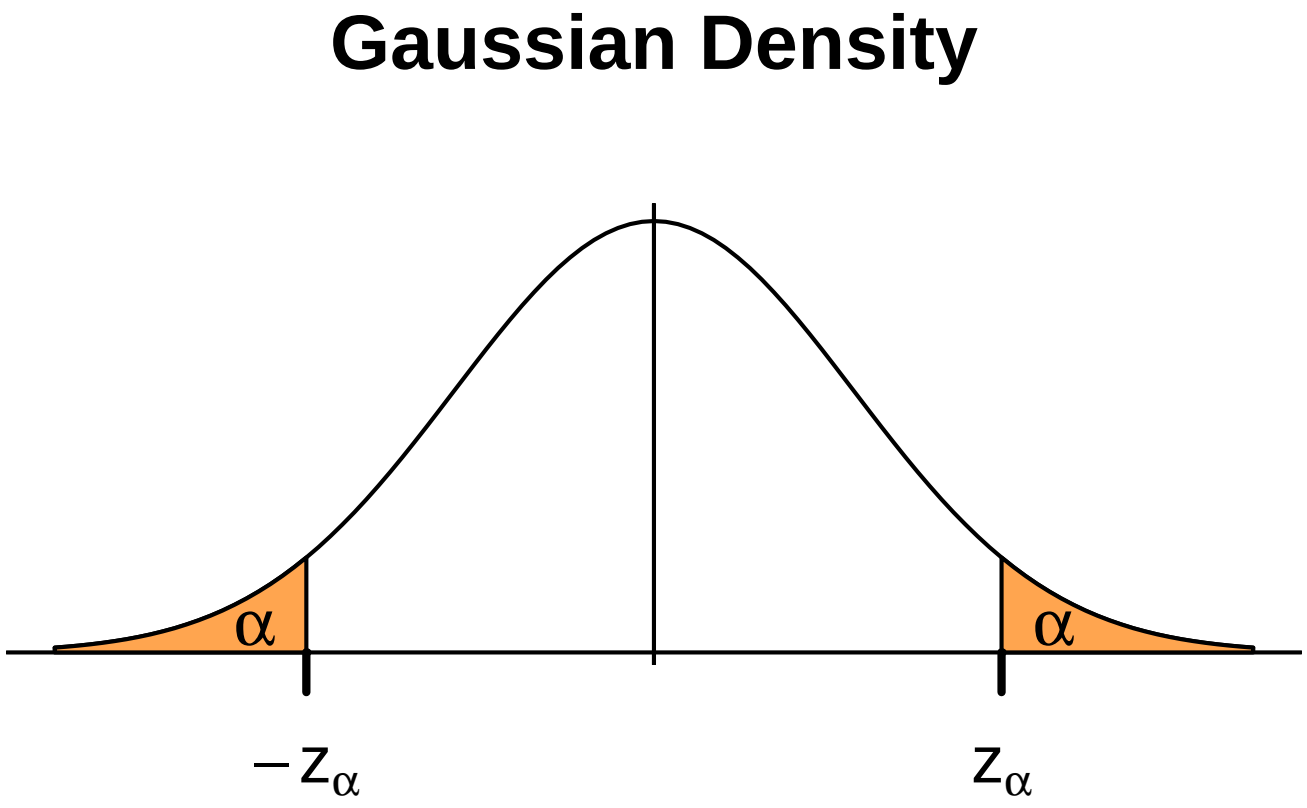


Figure 2: Definition of z_α and $-z_\alpha$ for a Gaussian distribution

If $Z \sim N(0,1)$,

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

We know that if the sample comes from a normal distribution,

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

Therefore, replacing Z by $\sqrt{n}(\hat{\mu}_n - \mu)/\sigma$ in (1),

$$P\Big(\Big|\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma}\Big| \leq z_{\alpha/2}\Big) = 1 - \alpha \tag{2}$$

After some manipulation of the inequality in the expression above we get that

$$P\left(\hat{\mu}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{3}$$

This equation says that, with probability $1 - \alpha$, the interval

$$I(\alpha; \sigma, n) = \left[\hat{\mu}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

contains the actual value of the parameter μ.

We say that $(1 - \alpha)$ is the **confidence level** of the interval, frequently expressed as a percentage $100(1 - \alpha)\%$. The smaller α, the higher the confidence level.

The extremes of the interval are random and the probability statement we just made applies to them and not to the parameter, which has a fixed (but unknown) value.

The width of the interval is

$$\left|I(\alpha;\sigma,n)\right| = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{4}$$

and the smaller the width, the sharper our estimate is.

For fixed sample size, the precision (width) of the interval, and the confidence level go in opposite directions. If we want to reduce one of them, we have to increase the other.

It is possible to use equation (4) to determine the sample size required to have a given width and confidence level, as long as we know the standard deviation σ or express the desired width in terms of σ.

Confidence intervals when the variance is unknown

Based on

$$\frac{\hat{\mu}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1} \tag{5}$$

where t_{n-1} denotes a t distribution with $n-1$ degrees of freedom.

Let $T_n \sim t_n$. We define $t_{\alpha,n}$ to be the real number that satisfies

$$P(T_n > t_{\alpha,n}) = \alpha.$$

By symmetry we have that $P(T_n < -t_{\alpha,n}) = \alpha$.

Following a similar argument as before, from equation (5) we get that

$$I^*(\alpha; n) = \left[\hat{\mu} - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}\right]$$

is a confidence interval of level $1-\alpha$ for the mean. One-sided confidence intervals can be obtained similarly.