STAT2201

Analysis of Engineering & Scientific Data

Unit 6

Slava Vaisman

The University of Queensland
School of Mathematics and Physics

Statistical inference

- Let $\boldsymbol{X}_1, \ldots, \mathbf{X}_n \sim F(\boldsymbol{x})$ be a data drawn randomly from some **unknown** distribution F.

- Assume that the data is independent and identically distributed (i.i.d).
  1. $\boldsymbol{X}_i \sim F(\boldsymbol{x})$ for all $1 \leq i \leq n$
  2. $\boldsymbol{X}_i$s are independent

- Statistical Inference is the process of forming judgements about the parameters

A statistic (1)

- A *statistic* is any function of the observations in a random sample. Examples:

$$g(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$g(X_1, X_2, \ldots, X_n) = \max\{X_1, X_2, \ldots, X_n\}$$

- More examples.
  - Sample variance and sample standard deviation
  - Sample quantiles besides the median, (quartiles and percentiles)
  - Order statistics
  - Sample moments and functions

A statistic (2)

- The probability distribution of a statistic is called the sampling distribution.
- Note that $g(X_1, X_2, \ldots, X_n)$ is also a random variable!
- A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta} = g(X_1, X_2, \ldots, X_n)$.
- The statistic $\hat{\Theta}$ is called the point estimator.
- The most common statistic we consider is the sample mean, $\overline{X}$, with a given value denoted by $\overline{x}$. As an estimator, the sample mean is an estimator of the population mean, μ.

Normal, or Gaussian, Distribution

The normal (or Gaussian) distribution is the most important distribution in the study of statistics, engineering, and biology.

We say that a random variable has a normal distribution with parameters μ and σ^2 if its density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

- We write $X \sim \mathrm{N}(\mu, \sigma^2)$.
- The parameters μ and σ^2 turn out to be the expectation and variance of the distribution, respectively.
- If $\mu = 0$ and $\sigma = 1$ then

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R},$$

and the distribution is known as a **standard normal distribution**.

Properties of Normal Distribution

- If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

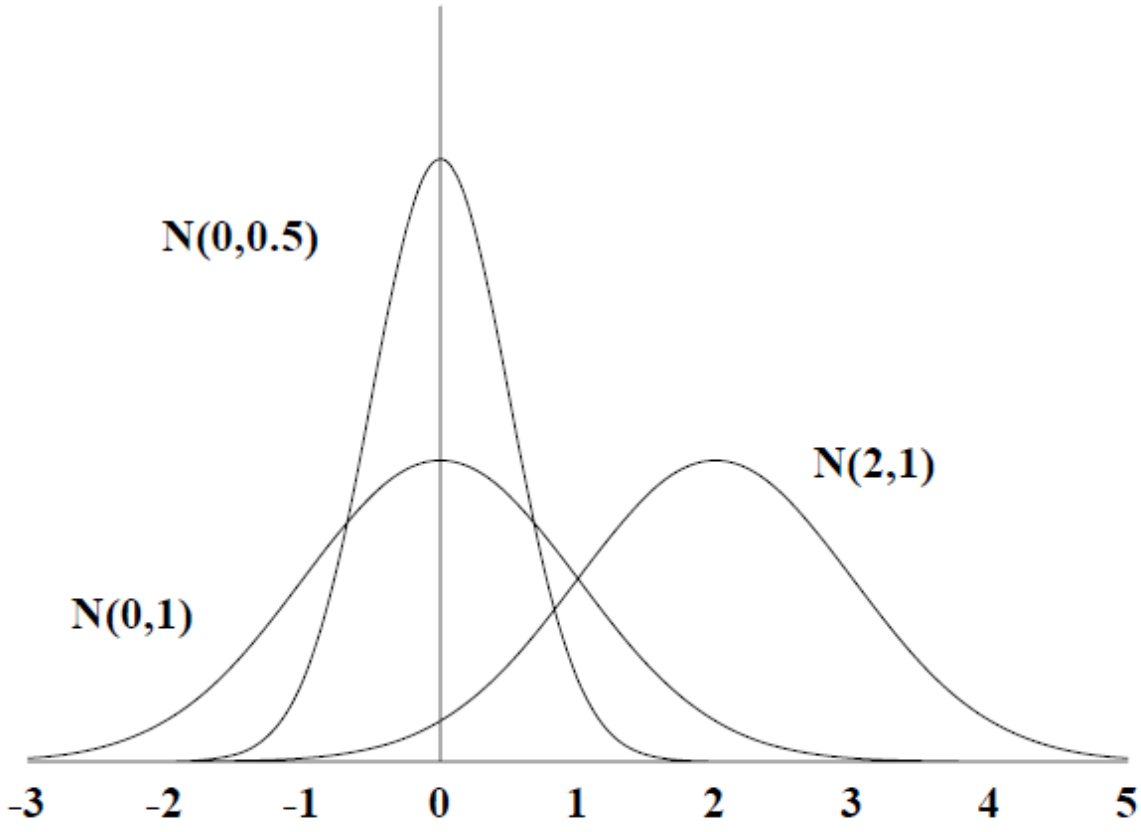
$$\frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

  Thus by subtracting the mean and dividing by the standard deviation we obtain a standard normal distribution. This procedure is called **standardization**.
- Standardization enables us to express the cdf of any normal distribution in terms of the cdf of the standard normal distribution.
- A trivial rewriting of the standardization formula gives the following important result: If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

$$X = \mu + \sigma Z, \quad Z \sim \mathrm{N}(0, 1).$$

- In other words, any Gaussian (normal) random variable can be viewed as a so-called affine (linear + constant) transformation of a standard normal random variable.

Normal Distribution



Sums of independent Random Variables

- The (probably most) celebrated theorem in probability: **the Central Limit Theorem** (CLT).
- Suppose, for example, that we weigh 20 randomly selected people. The average weight of the group is

$$\hat{w} = \frac{X_1 + \cdots + X_{20}}{20}.$$

- In general, let

$$X_1, X_2, \ldots, X_n$$

be independent and identically distributed random variables.

- For each n, let

$$S_n = X_1 + \cdots + X_n.$$

- Let $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ (assuming that these are finite).
- Note that $\mathbb{E}[S_n] = n\mu$, and $\mathrm{Var}(S_n) = n\sigma^2$.

Central Limit Theorem

The Central Limit Theorem states roughly that:

"The sum of a large number of iid random variables has approximately a Gaussian distribution."

More precisely, it states that, for all x,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x).$$

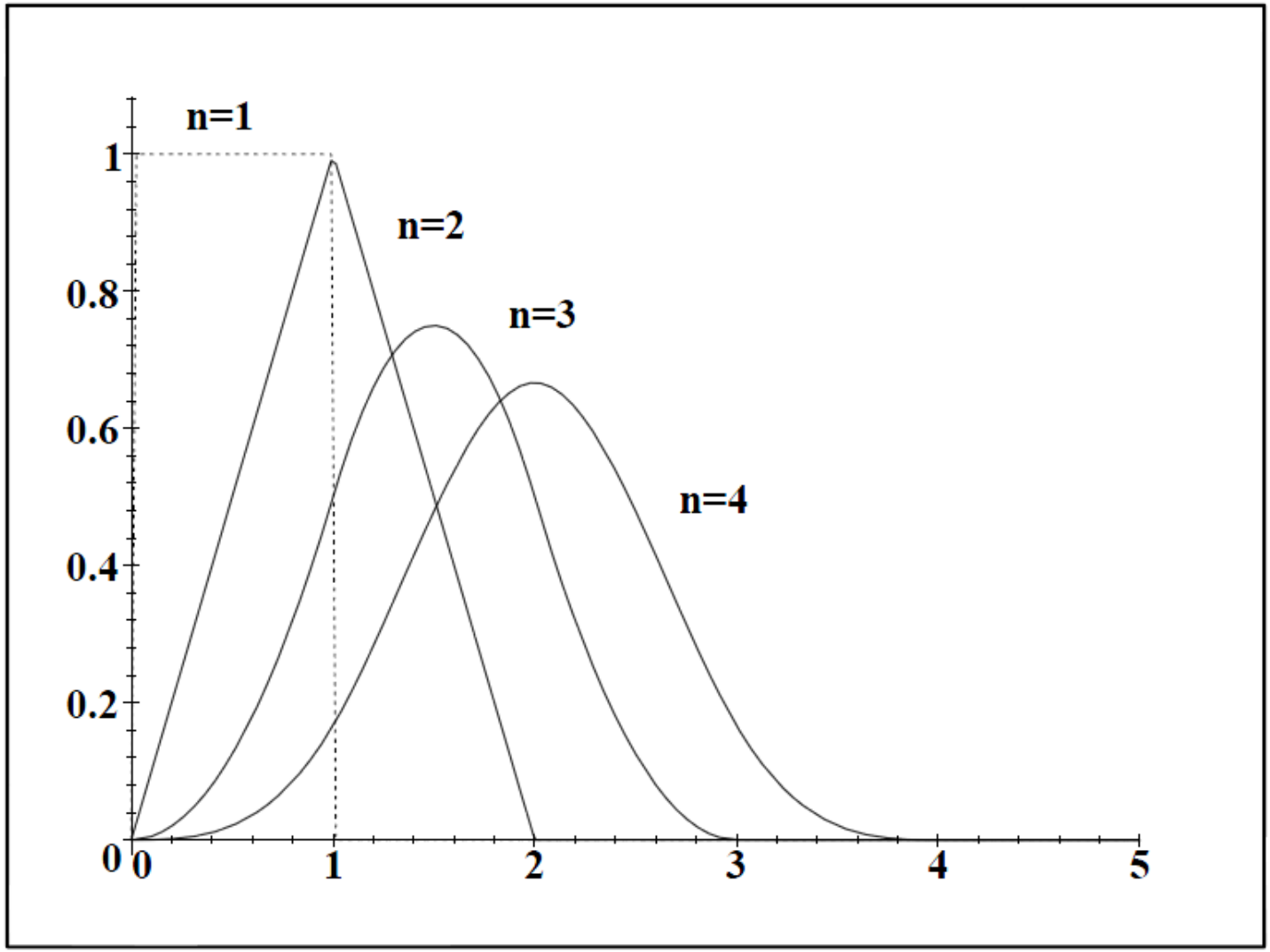
where Φ is the cdf of the standard normal distribution.

Regardless of X_i's distribution, the sum behaves (approximately) as the Gaussian random variable!

Let us see the amazing CLT in action.

Central Limit Theorem

The next picture shows the pdf's of $S_1, \ldots, S_4$ for the case where the X_i have a U$[0, 1]$ distribution.



Central Limit Theorem for the mean

- Let $\overline{X} = \frac{S_n}{n}$.
- $\mathbb{E}[\overline{X}] = \mu$
- Var $(\overline{X}) = \frac{\sigma^2}{n}$
- Therefore,

$$\mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x).$$

Central Limit Theorem — summary

1. For the sum of i.i.d random variables S_n:

$$S_n \sim \mathrm{N}\left(n\mu, n\sigma^2\right).$$

2. For the mean of i.i.d random variables $\overline{X}$:

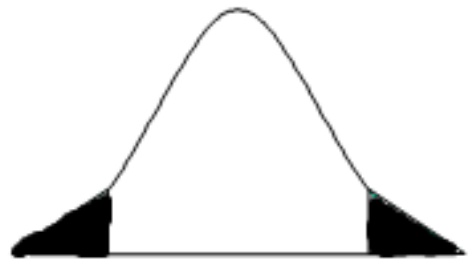
$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

The standard error of $\overline{X}$

- The standard error of $\overline{X}$ is given by $\sigma\sqrt{n}$.
- Note that In most practical situations σ is not known but rather estimated.
- The estimated standard error (SE) is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}}$$

- If $X \sim \mathrm{N}(0,1)$, the probability that X is between 0 ± 1 is about 0.68.



- What about $X \sim \mathrm{N}(\mu, \sigma^2)$?

Knowing the sampling distribution

Knowing the sampling distribution (or the approximate sampling distribution) of a statistic is the key for the two main tools of statistical inference that we study:

1. Confidence intervals — a method for yielding error bounds on point estimates.

2. Hypothesis testing — a methodology for making conclusions about population parameters.

Confidence intervals

The confidence interval

- A confidence interval estimate for μ **(the real mean)**, is an interval of the form

$$l \leq \mu \leq u,$$

  where the end-points l and u l and u are computed from the sample data $X_1, \ldots, X_n$.
- When we collect data, we can observe different $X_1, \ldots, X_n$, so these endpoints are values of random variables L and U, respectively.
- Suppose that

$$\mathbb{P}(L \leq \mu \leq U) = 1 - \alpha, \quad \alpha \in (0, 1).$$

- Then, the resulting confidence interval for μ is $l \leq \mu \leq u$, and the end-points or bounds l and u are called the **lower**- and **upper**-confidence limits (bounds), respectively, and $1 - \alpha$ is called the **confidence level**.

The confidence interval — intuition

- Suppose:

$$\mathbb{P}(L \leq \mu \leq U) = 1 - \alpha.$$

- Consider the following statements. What is your intuition about the α.

  1. "The average height in this class is between -10kg and 8000 kg"

  2. "The average height in this class is between 70kg and 72 kg"

The confidence interval for the mean (1)

- Recall that we know the sampling distribution of the mean:

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

- That is, for some positive scalar value $z_{1-\alpha/2}$, we have

$$\begin{aligned}\mathbb{P}\left(\overline{X} \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= \mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}\right) \\ &= \Phi(z_{1-\alpha/2})\end{aligned}$$

$$\begin{aligned}\mathbb{P}\left(\overline{X} \leq \mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= \mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq -z_{1-\alpha/2}\right) \\ &= \Phi(-z_{1-\alpha/2}) = 1 - \Phi(z_{1-\alpha/2})\end{aligned}$$

The confidence interval for the mean (2)

- From these equations, we have

$$\begin{aligned}\mathbb{P}\left(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ = \mathbb{P}\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ = \Phi(z_{1-\alpha/2}) - (1 - \Phi(z_{1-\alpha/2})) = 2\Phi(z_{1-\alpha/2}) - 1.\end{aligned}$$

- Recall that we want

$$\mathbb{P}\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

so, setting

$$\begin{aligned}1 - \alpha = 2\Phi(z_{1-\alpha/2}) - 1 = 2(1 - \Phi(-z_{1-\alpha/2})) - 1 \\ = 1 - 2\Phi(-z_{1-\alpha/2}) \Rightarrow \alpha = 2\Phi(-z_{1-\alpha/2}).\end{aligned}$$

The confidence interval for the mean (3)

- Therefore, a $100(1-\alpha)\%$ confidence interval on μ is given by

$$\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Since $\alpha = 2\Phi(-z_{1-\alpha/2})$, we can choose $z_{1-\alpha/2}$ as follows:

1. $99\% \Rightarrow \alpha = 0.01 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.005 \Rightarrow z_{1-\alpha/2} = 2.57$
2. $98\% \Rightarrow\Rightarrow \alpha = 0.02 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.01 \Rightarrow z_{1-\alpha/2} = 2.32$
3. $95\% \Rightarrow\Rightarrow \alpha = 0.05 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.025 \Rightarrow z_{1-\alpha/2} = 1.96$
4. $90\% \Rightarrow\Rightarrow \alpha = 0.1 \Rightarrow \Phi(-z_{1-\alpha/2}) = 0.05 \Rightarrow z_{1-\alpha/2} = 1.64$

The confidence interval for the mean — sample size

Confidence interval formulas give insight into the required sample size: If $\overline{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error $|\overline{x}-\mu|$ will not exceed a specified amount Δ when the sample size is not smaller than

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{\Delta}\right)^2,$$

since

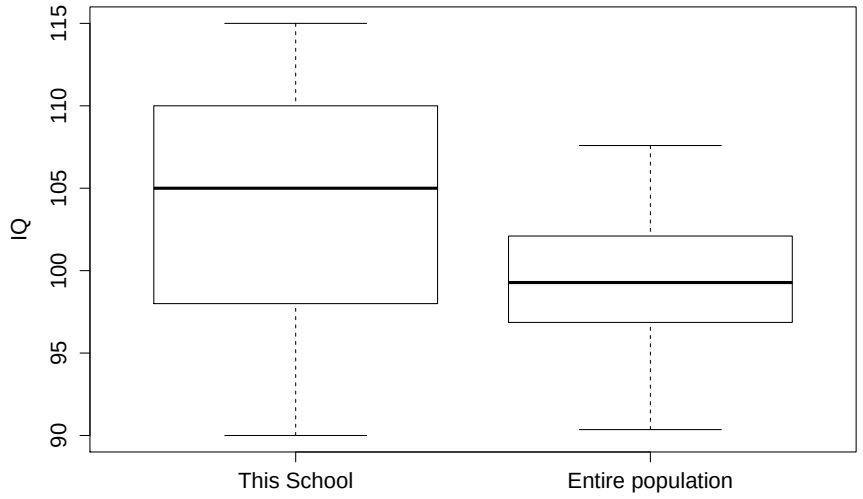
$$|\overline{x}-\mu| \leq \Delta \Rightarrow z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \Delta \Rightarrow n \geq \left(\frac{z_{1-\alpha/2}\sigma}{\Delta}\right)^2.$$

Hypothesis testing

Hypothesis testing — Choosing a school

A certain **(and not very cheap)** private school claims that its students have a higher IQ. The entire student population is known to have an IQ that is Gaussian distributed with mean 100 and variance 16.

- Should we try to place our child in this school?
- Is the observed result *significant* **(can be trusted?)**, or due to a *chance*?



Example (Medical treatment)

Consider an experimental medical treatment, in which 14 subjects were randomly assigned to control or treatment group. The survival times (in days) are shown in the table below.

	Data	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12, 18	43.285

- Did the treatment prolong the survival?
- Is the observed result *significant*, or due to a *chance*?

Making an error in this example, can have much more serious consequences when placing a child in an average school.

Example (Tossing a coin)

I take a coin, toss it 10 times, and tell you the number of heads.

- Is this a fair coin?
- Is the observed result *significant*, or due to a *chance*?

Example (Testing an Improved Battery)

A manufacturer claims that its new improved batteries have a much longer lifetime. The old batteries are known to have a lifetime that is Normally distributed with mean 150 and variance 16. We measure the lifetime of nine batteries and obtain a sample mean of 155 hours.

- Is this new battery superior to the previous version?
- Is the observed result *significant*, or due to a *chance*?

The framework

- Note that all the above examples are some-that similar.
- Specifically, we observed a system (school, or medical treatment, or coin toss, or electric battery),
- and asked ourself the following questions:
  1. Is the observed data is due to **chance**, or,
  2. due to **effect**?

For example,

1. Is the observed IQ in the school is due to "chance", or
2. the observed IQ in the school is due to "effect"; that is. one should definitely prefer this school!

Can you provide a similar statement for the medical, coin toss (you observed 7 heads out of 10 tosses), and battery experiments?

The framework

- To conclude, regardless the nature of our experiment, we always ask the same question:

The question

Is the observed *data* is due to **chance**, or due to **effect**?

- This question brings us to a formulation of hypothesis. Specifically, given a data, our first task is to formulate two hypotheses.

The research hypotheses

1. The *null* hypothesis H_0, which stands for our initial assumption about the data.
2. The *alternative hypothesis* H_1, (sometimes called H_A).

Setting the Hypothesis

Note that the *null* and the *alternative* hypotheses are two mutually exclusive statements!

Example (Criminal Trial)

- H_0 : Defendant is **not guilty**.
- H_1 : Defendant is **guilty**.

Example (Choosing a school)

1. H_0 : The observed IQ in the school is due to "chance".
2. H_1 : The observed IQ in the school is due to "effect". (One should definitely prefer this school!)

Setting the Hypothesis

Example (Medical treatment)

1. H_0 : The observed data is due to "chance", that is, the treatment does not prolong the survival.
2. H_1 : The observed data is due to "effect". (One should definitely consider this treatment!)

Example (Coin toss with 7 out of 10 heads)

1. H_0 : The observed data is due to "chance", that is, the coin is fair.
2. H_1 : The observed data is due to "effect"; that is, the coin is biased.

Can you provide H_0 and H_1 for the battery experiment?

Hypothesis testing

Hypothesis testing

The general idea of hypothesis testing involves the following steps.

1. Collecting data.
2. Formulating the H_0 and the H_1 hypotheses.
3. Based on the data, decide whether to reject or not reject the initial hypothesis H_0.

- Sometimes, we alternate the first and the second steps.
- The first and the second steps look manageable.
- The third step looks like the most interesting (critical) one.

At this stage, suppose that we performed a test and made a decision regarding the **null** hypothesis.

Making an error

Regardless of the procedure in the third step, we either

1. reject H_0, or
2. do not reject H_0.

This, can lead to an error, which is summarized in the table below.

	True state	
Decision	H_0 true	H_1 true
Retain H_0	OK	Type II error (false negative)
Reject H_0	Type I error (false positive)	OK

Making an error

Definition (Significance level of the statistical test)

The probability of a type I error is called the **significance level of the test** and is denoted by α. (It is common to set the significance level to 0.05, that is, accepting to have a 5% probability of incorrectly rejecting the null hypothesis.)

$$\alpha = \mathbb{P}(\text{type I error}) = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$$

Definition (Power of the statistical test)

The probability of a type II error is called the **power of the test** and is denoted by β. β is the probability of making type II error.

$$\beta = \mathbb{P}(\text{type II error}) = \mathbb{P}(\text{retain } H_0 \mid H_1 \text{ is true})$$

Hypothesis testing

We wish: α is low and power $(1 - \beta)$ as high as can be.

Some remarks

- In most hypothesis tests used in practice (and in this course), a specified level of type I error, α is predetermined (e.g. $\alpha = 0.05$) and the type II error is not directly specified.

- The probability of making a type II error also depends on the sample size n - increasing the sample size results in a decrease in the probability of a type II error.

- The population (or natural) variability (e.g. described by σ) also affects the power.

The formal hypothesis testing framework - rejection region, test statistics, and critical value

- Let X be a random variable such that $\mathcal{X}$ is the range of X.
- The hypothesis testing is performed via finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**.
- Specifically, if

$$\begin{cases} X \in R \Rightarrow & \text{reject the null hypothesis } H_0 \\ X \notin R \Rightarrow & \text{do not reject the null hypothesis.} \end{cases}$$

- In many cases, the rejection region R takes the form of

$$R = \{x : T(x) > c\},$$

where T is some **test statistic** and c is called a **critical value**.

Back to the school example

Example (Choosing a school)

Recall that the total population IQ is distributed according to $\mathrm{N}(100, 16)$, and suppose that we gathered some data $X_1, \ldots, X_n$ from this private school.

- A reasonable **test statistics** $T(X_1, \ldots, X_n)$, can be:

$$T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i - 100 = \overline{X} - 100.$$

- Intuitively, we should reject the null hypothesis is $\overline{X} - 100$ is large.
- To do so, we should define **large**. Specifically, we need to specify the **critical value** c, (say $c = 4$?), such that the rejection region is defined via:

$$R = \left\{ X_1, \ldots, X_n : \overline{X} - 100 > c \right\}.$$

Finding the critical value

So, what is the appropriate critical value c?

Recall that the Type I error (false positive), happens when we reject H_0 when it is in fact true.

Definition (A reminder: Significance level of the statistical test)

The probability of a type I error is called the **significance level of the test** and is denoted by α.

That is, c will be a function of the significance level α that is defined by us!

Intuitively, the critical value c should depend on the test's significance level. The larger is c, the smaller is α. In particular, recall the school rejection region

$$R = \left\{ X_1, \ldots, X_n \, : \, \overline{X} - 100 > c \right\}.$$

Example (Finding critical value)

- Let $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, ($\sigma$ is known).
- We would like to test $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$. Therefore, $\Theta = [\mu_0, \infty), \quad , \Theta_0 = \{\mu_0\}, \quad , \Theta_1 = (\mu_0, \infty)$.
- We choose the test statistics T to be $T = \overline{X}$, and, we define the rejection region to be

$$R = \left\{x_1, \ldots, x_n \, : \, \overline{X} > c\right\}.$$

- Finally, we set the significance level of the test to be α.
- Here is some calculus:

$$\underbrace{\alpha = \mathbb{P}_{\mu_0}(\overline{X} > c)}_{\text{Type I error}} = \mathbb{P}_{\mu_0}\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} > \frac{\sqrt{n}(c - \mu_0)}{\sigma}\right)$$

$$= \mathbb{P}\left(Z > \frac{\sqrt{n}(c - \mu_0)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right) = \alpha.$$

Example (Finding critical value cnt.)

So,

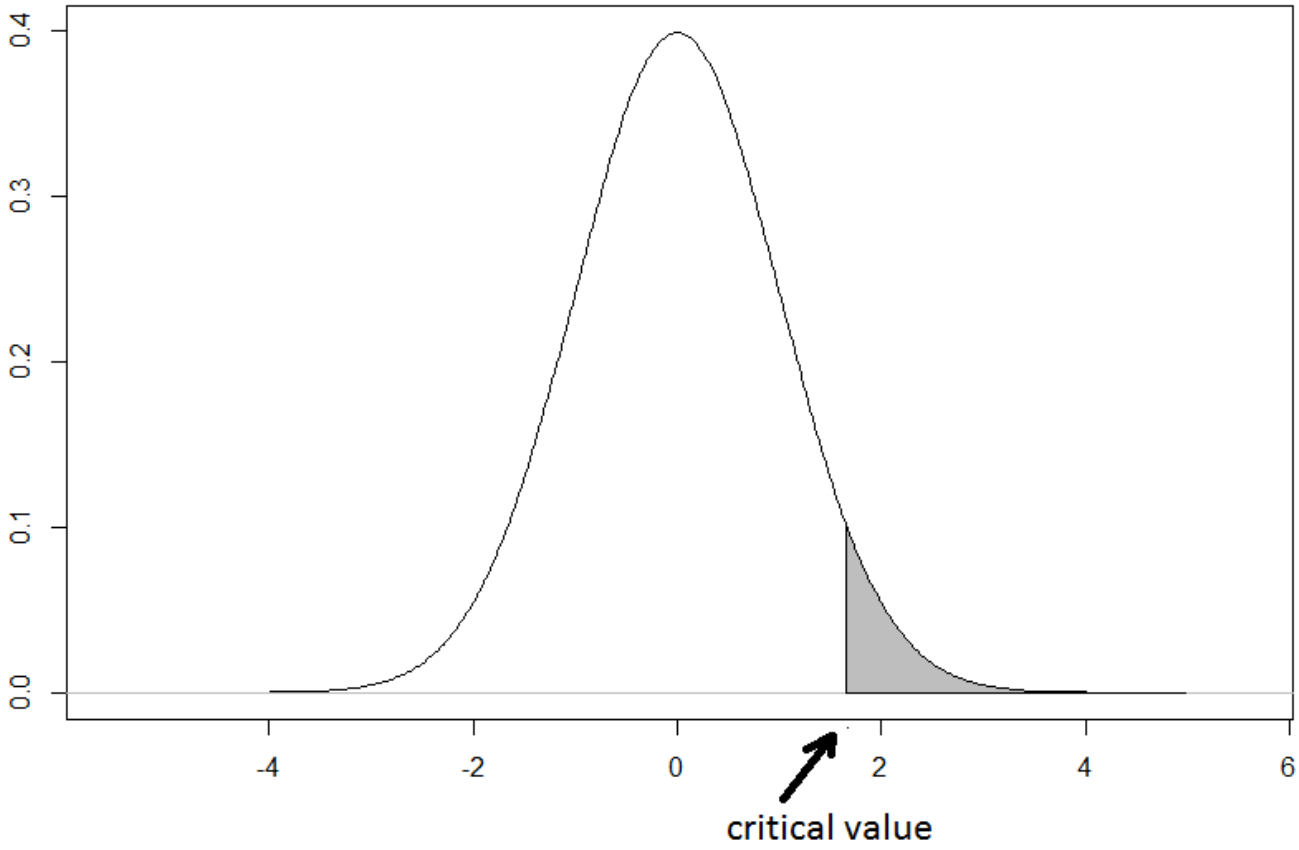
$$\mathbb{P}_{\mu_0}(\overline{X} > c) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right) = \alpha.$$

Therefore, the critical value c is:

$$c = \mu_0 + \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

Note that $\Phi(1-\alpha)$ is monotonically increasing function, that is,

The critical value c grows as α decrease! (As expected!)

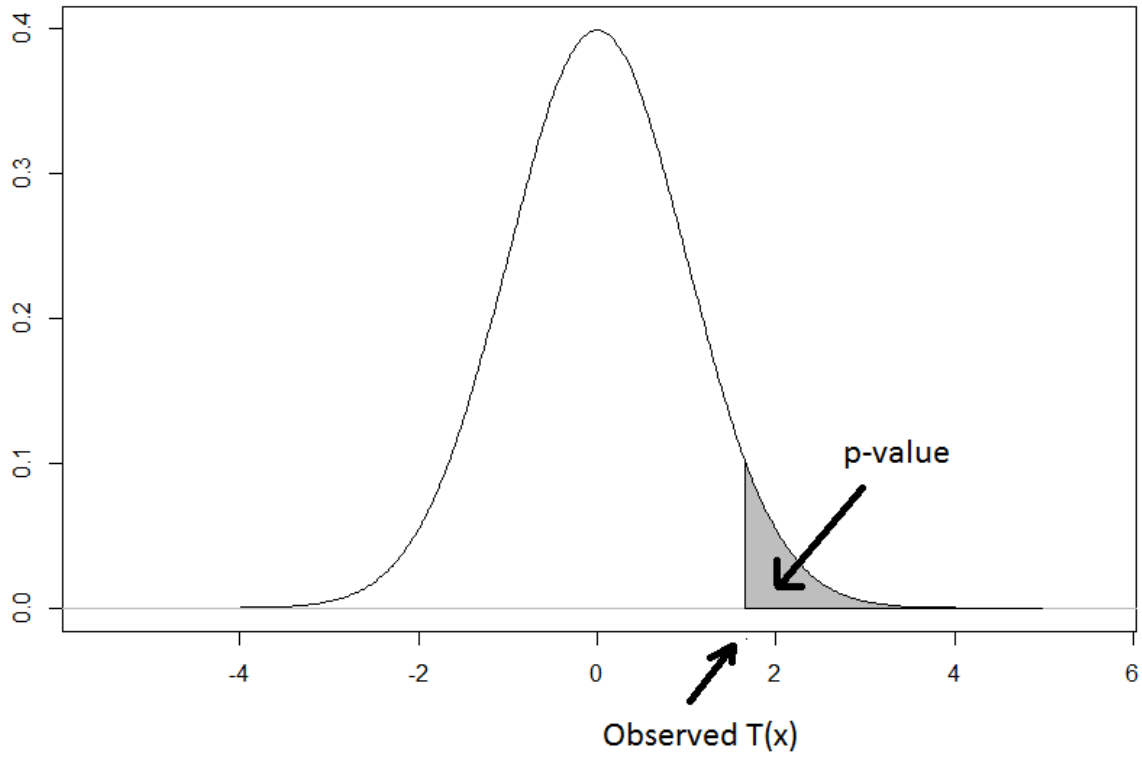


- The area of the shaded area is α!
- So, if the observed test statistics falls into the shaded area, we **reject the null hypothesis**.

Equivalent approach: *p*-value

Definition (*p*-value)

The *p-value* is the probability that under the null hypothesis, the random test statistic takes a value as extreme as or more extreme than the one observed.



Equivalent approach: p-value

- Critical region and p-values are essentially the same.
- However, it is easier to work with p-values; we will see why.
- The general statistical test procedures using p-values is as follows.

1. Formulate a statistical model for the data.
2. Give the null and alternative hypotheses (H_0 and H_1).
3. Choose an appropriate test statistic.
4. Determine the distribution of the test statistic under H_0.
5. Evaluate the outcome of the test statistic.
6. Calculate the p-value.
7. Accept or reject H_0 based on the p-value.

Equivalent approach: p-value

In the last step, if we reject H_0 for p-value less than α, we are back again to the statistical significance!

An easy to remember rule is:

> p-value low $\Rightarrow$ H_0 must go!

p-value	evidence
< 0.01	very strong evidence against H_0
$0.01 - 0.05$	moderate evidence against H_0
$0.05 - 0.10$	suggestive evidence against H_0
> 0.1	little or no evidence against H_0

Example

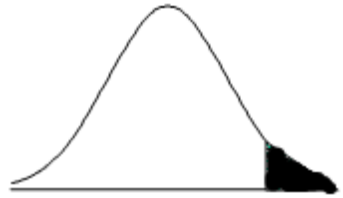
- Let $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, ($\sigma$ is known).
- We would like to test $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$. Therefore, $\Theta = [\mu_0, \infty), \quad , \Theta_0 = \{\mu_0\}, \quad , \Theta_1 = (\mu_0, \infty)$.
- We choose the test statistics T to be $T = \overline{X}$.
- Under H_0, T is distributed $\mathrm{N}(\mu_0, \sigma^2/n)$.
- Calculate the p-value:

$$\mathbb{P}_{H_0}(T(X) > \overline{X}) = \cdots = \mathbb{P}\left(Z > \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right).$$

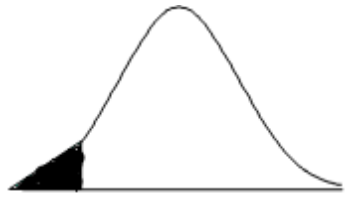
- If $\mathbb{P}_{H_0}(T(X) > \overline{X})$ is less than the test significance level α, we reject the null hypothesis.
- It can be shown that this is absolutely identical to the usage of the critical value and the rejection region.

Types of tests

- **Right one-sided test:** where H_0 is rejected for the p-value defined by $\mathbb{P}_{H_0}(T \geq t)$.



- **Left one-sided test:** where H_0 is rejected for the p-value defined by $\mathbb{P}_{H_0}(T \leq t)$.



- **Two-sided test:** where H_0 is rejected for the p-value defined by $\mathbb{P}_{H_o}(T \geq t) + \mathbb{P}_{H_o}(T \leq -t) = 2\mathbb{P}_{H_o}(T \geq t)$.