Intermediate Econometrics

Lecture 2

Vincenzo Gioia

16th October 2026

Statistical Model in Brief

Statistical Model: The Simplest One

- Independent observations (on n statistical units) about a random variable of interest
- The sample is composed by the random variables $Y_1, \ldots, Y_n$ that are independent and identically distributed according to a certain probability distribution (which is known up to an unknown parameter)

$$Y_i \sim f(\cdot, \theta), \quad \theta \in \Theta \subset \mathbb{R}^d, \quad d \geq 1, \quad i = 1, \ldots, n$$

Assumption

- The random variable is described by a probability mass function (discrete r.v.) or density function (continuous r.v.)
- θ: d-dimensional parameter assuming value in the paramteric space Θ

Statistical Model: The Normal Model

Normal/Gaussian Random Variable

A random variable X is said to be Normal (or Gaussian) with mean $E[X] = \mu$ and variance $V(X) = \sigma^2$ if the density function takes the functional form

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- We denote it with $X \sim \mathcal{N}(\mu, \sigma^2)$

Standard Normal

A particular normal random variable is the standard normal, that is a normal with mean 0 and variance 1

$$f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- We (usually) denote it using $Z \sim \mathcal{N}(0, 1)$ and it represents the core building for all the normal distributions

The Normal Distribution

Relationship between Standard Normal and a Generic Normal

- If $Z \sim \mathcal{N}(0,1)$ then $X = \mu + \sigma Z$ is distributed as $X \sim \mathcal{N}(\mu, \sigma^2)$
- Viceversa, if $X \sim \mathcal{N}(\mu, \sigma^2)$ then the r.v. $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

Cumulative Distribution Function

- $F_X(x) = P(X \leq x) = P(\mu + \sigma Z \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi(\frac{x-\mu}{\sigma})$ where $\Phi(\cdot)$ is the cumulative distribution function of Z (whose values are obtained numerically)

Quantiles of Order $p \in (0,1)$ of $X \sim \mathcal{N}(\mu, \sigma^2)$

- The smallest value x such that the probability of X being less than or equal to x is at least p $x_p = \inf\{x \in \mathbb{R} : F(x) \geq p\}$
- If F is continuous and strictly increasing, this reduces to $x_p = F^{-1}(p)$, i.e. the unique value satisfying $P(X \geq q_p) = p$.
- $x_p = \mu + \sigma z_p$ where $z_p = \Phi^{-1}(p)$ is the quantile of order p of Z

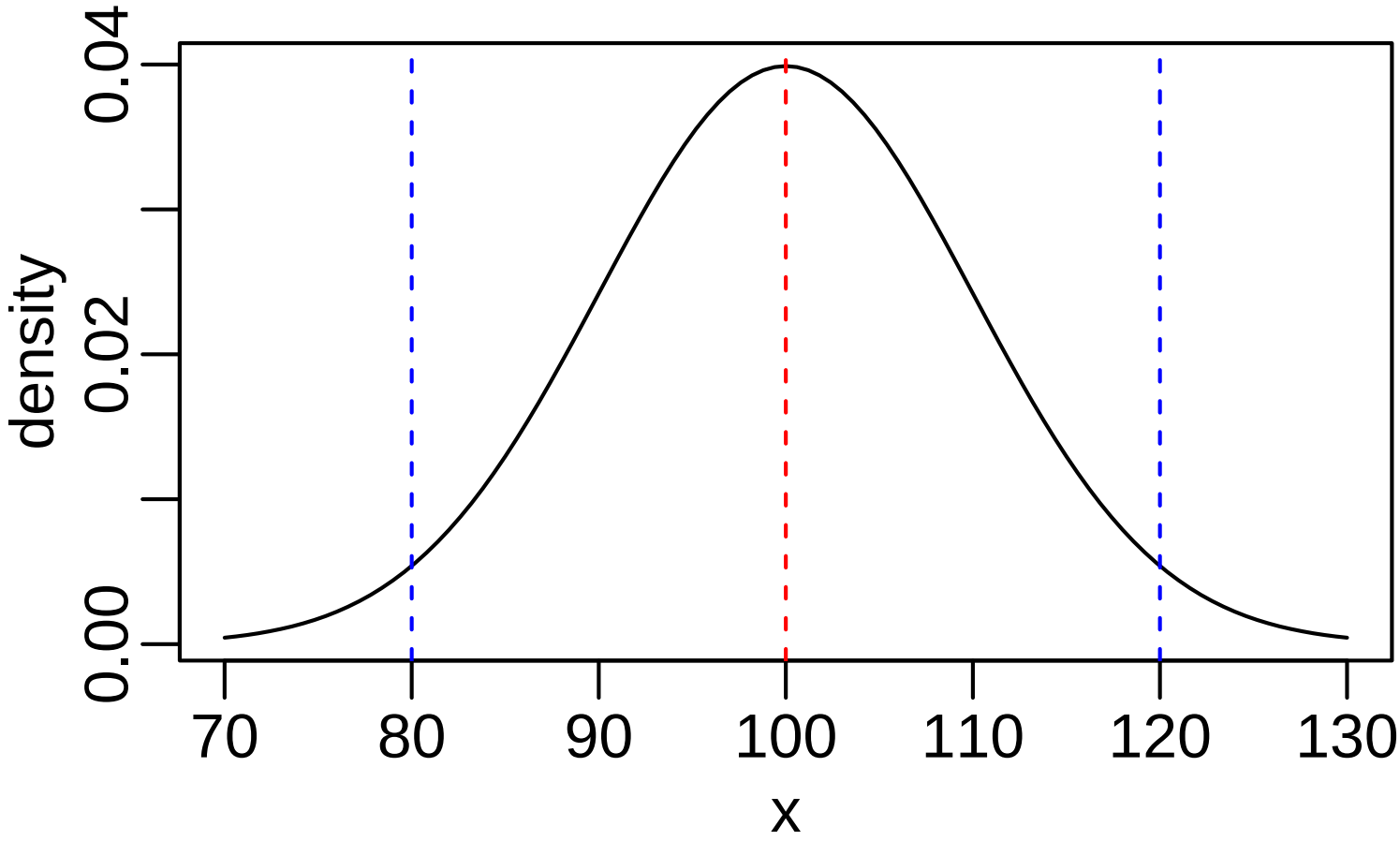
The Normal distribution

Characteristics of $f_X(x)$

- It is symmetric, with the line $x = \mu$ as its axis of symmetry
- It is increasing on the interval $(-\infty, \mu)$ and decreasing on the interval $(\mu, +\infty)$
- It has two inflection points at $x = \mu - \sigma$ and $x = \mu + \sigma$
- It is concave on the interval $(\mu - \sigma, \mu + \sigma)$ and convex elsewhere
- It has the x-axis as an asymptote

The Normal distribution

```
curve(dnorm(x, mean = 100, sd = 10), xlim = c(70, 130),
      ylab = "density")
abline(v = 100, lty = "dashed", col = "red")
abline(v = c(80, 120), lty = "dashed", col = "blue")
```

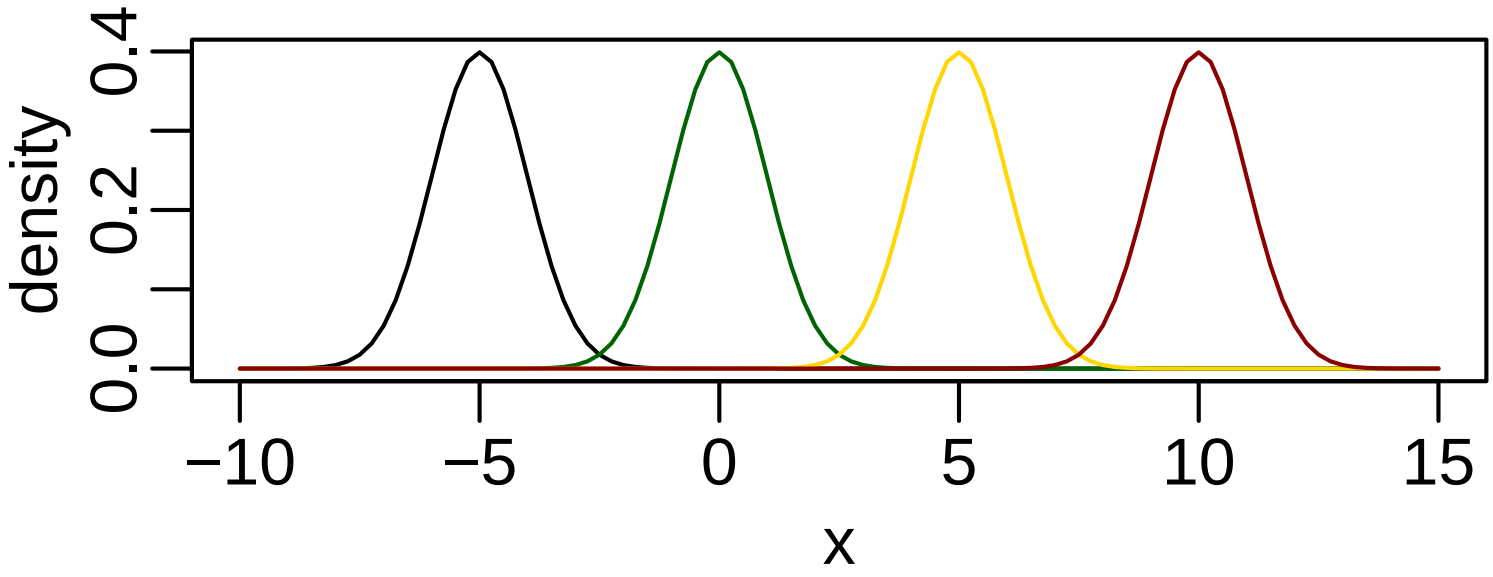


The Normal distribution

Location Shift

- $Z \sim \mathcal{N}(0, 1)$
- $X = \mu + Z \sim \mathcal{N}(\mu, 1)$
- Example considering $\mu \in \{-5, 0, 5, 10\}$

```
curve(dnorm(x, mean = -5, sd = 1), xlim = c(-10, 15), ylab = "density")
curve(dnorm(x, mean = 0, sd = 1), add = TRUE, col = "darkgreen")
curve(dnorm(x, mean = 5, sd = 1), add = TRUE, col = "gold")
curve(dnorm(x, mean = 10, sd = 1), add = TRUE, col = "darkred")
```

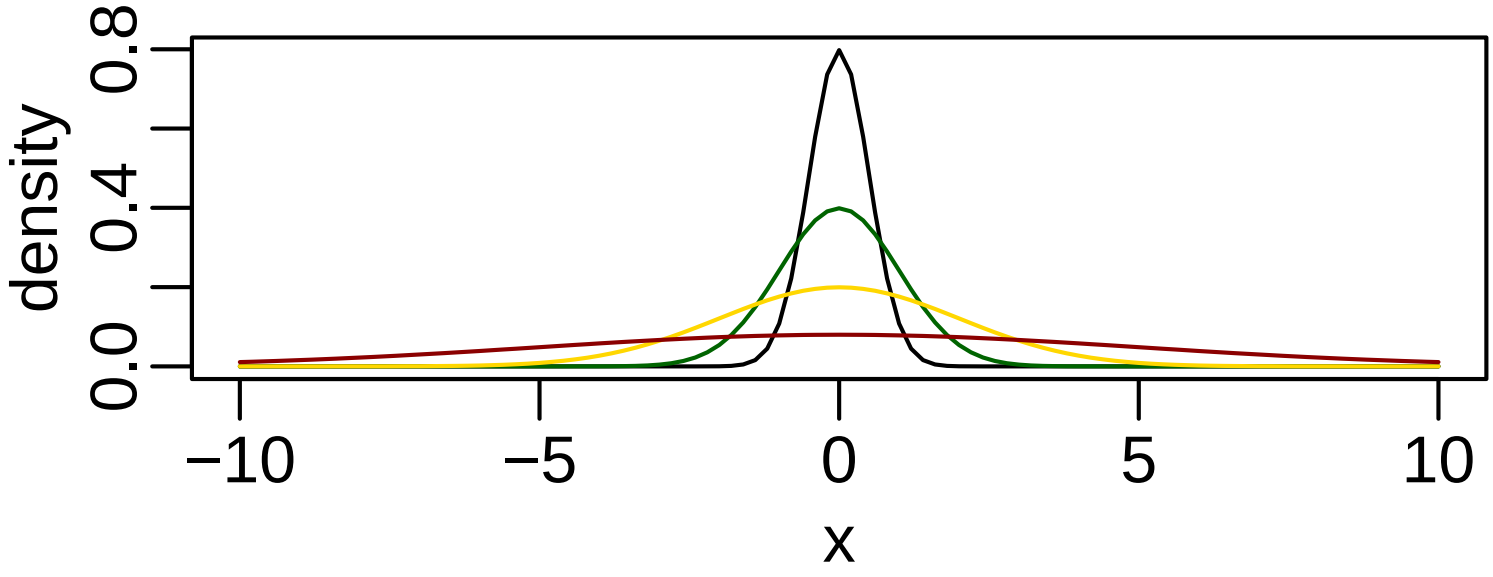


The Normal distribution

Scaling

- $Z \sim \mathcal{N}(0, 1)$
- $X = \sigma Z \sim \mathcal{N}(0, \sigma^2)$
- Example considering $\sigma \in \{0.5, 1, 2, 5\}$

```
curve(dnorm(x, mean = 0, sd = 0.5), xlim = c(-10, 10), ylab = "density"
curve(dnorm(x, mean = 0, sd = 1), add = TRUE, col = "darkgreen")
curve(dnorm(x, mean = 0, sd = 2), add = TRUE, col = "gold")
curve(dnorm(x, mean = 0, sd = 5), add = TRUE, col = "darkred")
```

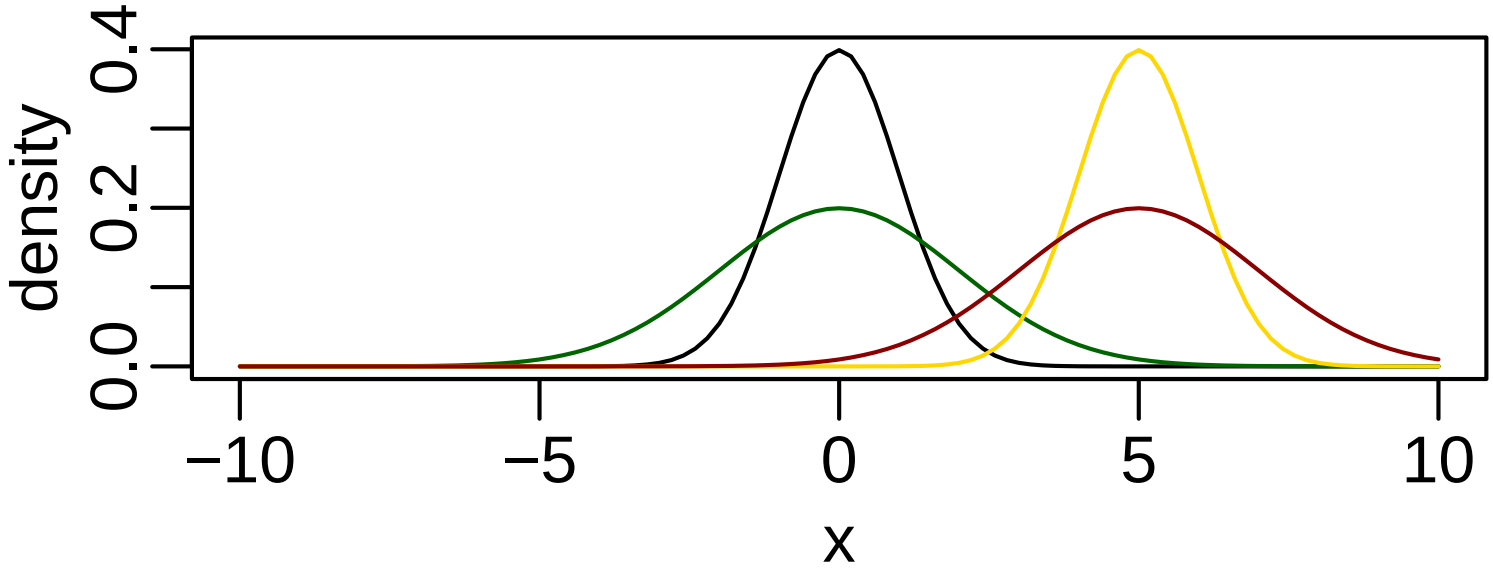


The Normal distribution

Location Shift and Scaling

- $Z \sim \mathcal{N}(0, 1)$
- $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$
- Example using the combination of $\mu \in \{0, 5\}$ and $\sigma \in \{1, 2\}$

```
curve(dnorm(x, mean = 0, sd = 1), xlim = c(-10, 10), ylab = "density")
curve(dnorm(x, mean = 0, sd = 2), add = TRUE, col = "darkgreen")
curve(dnorm(x, mean = 5, sd = 1), add = TRUE, col = "gold")
curve(dnorm(x, mean = 5, sd = 2), add = TRUE, col = "darkred")
```

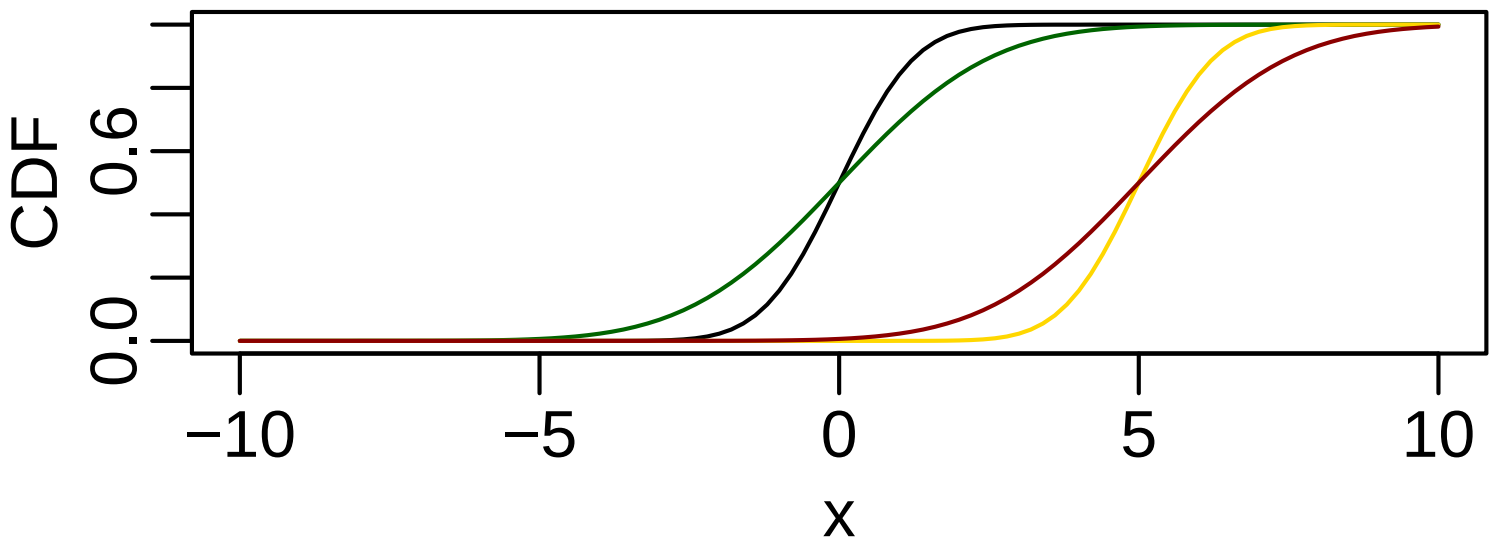


The Normal distribution

The Cumulative Distribution Function $F_X(x)$

- $Z \sim \mathcal{N}(0, 1)$
- $X = \mu + \sigma Z \sim \mathcal{N}(0, \sigma^2)$
- Example using the combinations of $\mu \in \{0, 5\}$ and $\sigma \in \{1, 2\}$

```
curve(pnorm(x, mean = 0, sd = 1), xlim = c(-10, 10), ylab = "CDF")
curve(pnorm(x, mean = 0, sd = 2), add = TRUE, col = "darkgreen")
curve(pnorm(x, mean = 5, sd = 1), add = TRUE, col = "gold")
curve(pnorm(x, mean = 5, sd = 2), add = TRUE, col = "darkred")
```

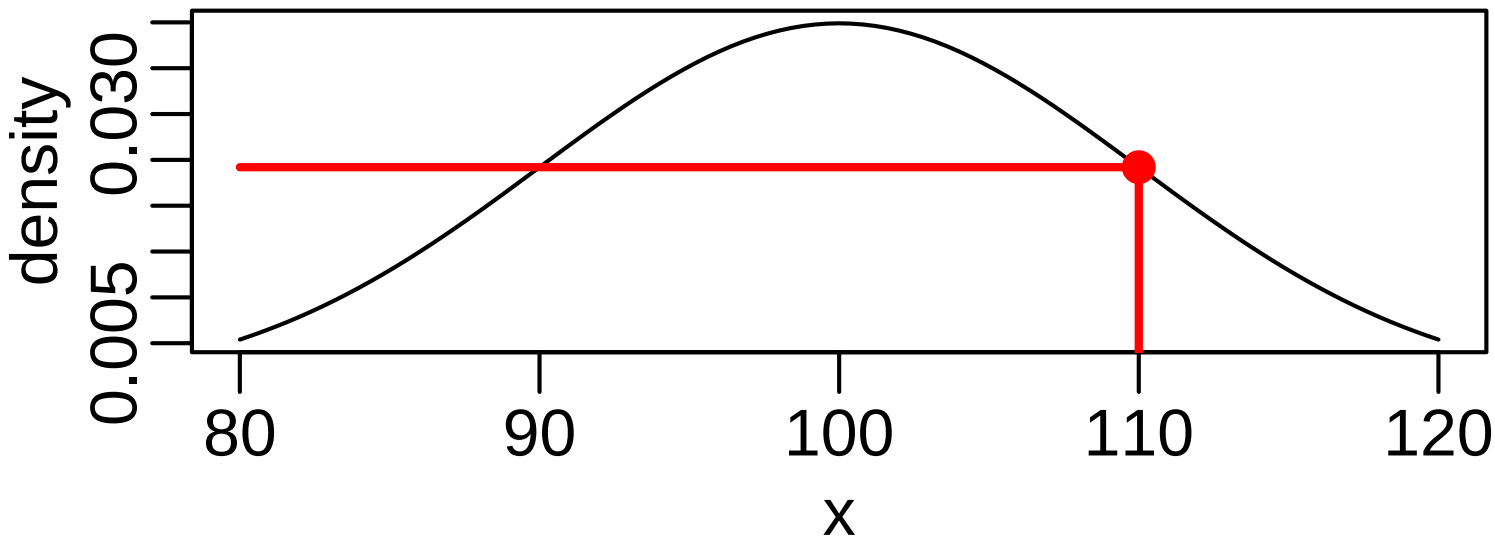


The Normal distribution

The Density Function $f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

```
curve(dnorm(x, mean = 100, sd = 10), xlim = c(80,120), ylab = "density"
dnorm(110, mean = 100, sd = 10)
y <- dnorm(110, mean = 100, sd = 10)
segments(x0 = 110, y0 = 0, x1 = 110, y1 = y, col = "red", lwd = 2)
segments(x0 = 80, y0 = y, x1 = 110, y1 = y, col = "red", lwd = 2)
points(110, y, col = "red", pch = 19)
```

```
## [1] 0.02419707
```

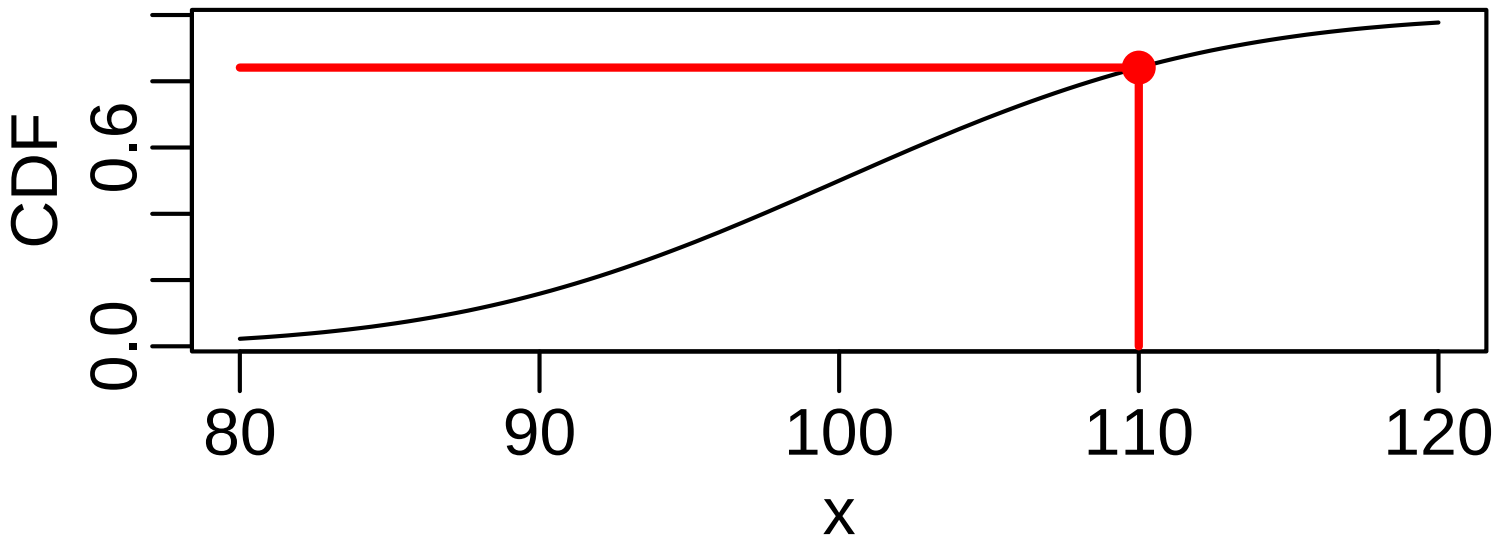


The Normal distribution

The Cumulative Distribution Function $F_X(x; \mu, \sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

```
curve(pnorm(x, mean = 100, sd = 10), xlim = c(80,120), ylab = "CDF")
pnorm(110, mean = 100, sd = 10)
y <- pnorm(110, mean = 100, sd = 10)
segments(x0 = 110, y0 = 0, x1 = 110, y1 = y, col = "red", lwd = 2)
segments(x0 = 80, y0 = y, x1 = 110, y1 = y, col = "red", lwd = 2)
points(110, y, col = "red", pch = 19)
```

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

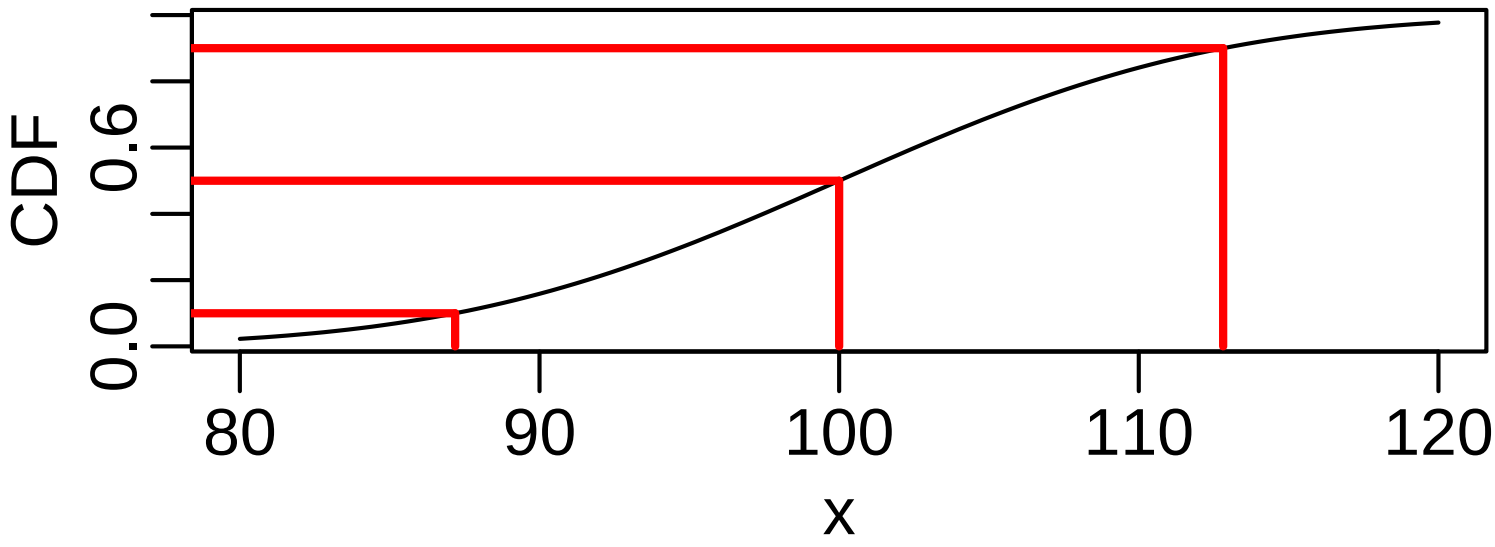


The Normal distribution

Quantiles of Order $p \in (0,1)$ of $X \sim \mathcal{N}(\mu, \sigma^2)$

```
curve(pnorm(x, mean = 100, sd = 10), xlim = c(80,120), ylab = "CDF")
qnorm(c(0.1,0.5,0.9), mean = 100, sd = 10)
x <- qnorm(c(0.1,0.5,0.9), mean = 100, sd = 10)
segments(x0 = c(0,0,0), y0 = c(0.1, 0.5, 0.9), x1 = x, y1 = c(0.1, 0.5,
segments(x0 = x, y0 = 0, x1 = x, y1 = c(0.1, 0.5, 0.9), col = "red", lw
```

```
## [1]  87.18448 100.00000 112.81552
```

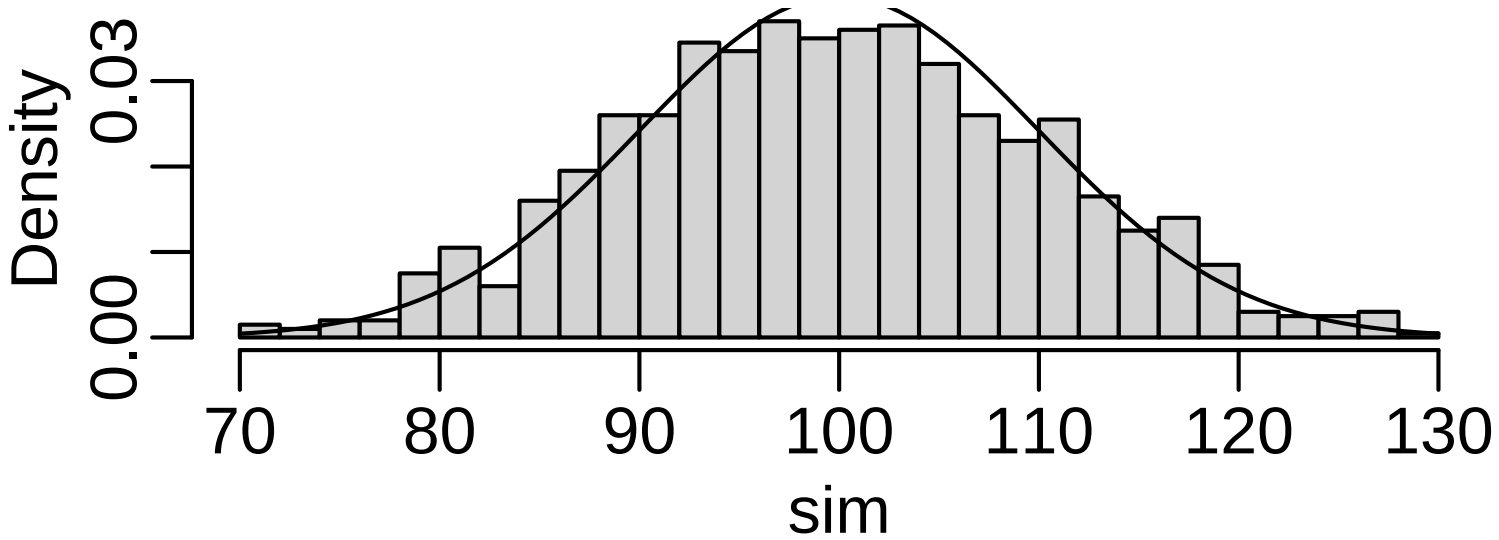


The Normal distribution

Random Generation

- We can generate a number of random values from the $\mathcal{N}(\mu, \sigma^2)$

```
set.seed(14)
sim <- rnorm(1000, mean = 100, sd = 10)
hist(sim, breaks = 40, freq = FALSE)
curve(dnorm(x, mean = 100, sd = 10), add = TRUE)
```



Statistical model: The Normal Model

The Normal (or Gaussian) Model

- We assume that $Y_1, \ldots, Y_n$ are random variables independent and identically distributed according to the $\mathcal{N}(\mu, \sigma^2)$ distribution
- Parameter: $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$
- The most popular estimators of μ and σ^2 are respectively

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

  that is the sample mean and the sample variance, respectively
- The estimators above are unbiased and consistent

The Normal Model: Inferential Results

Inference on $\theta = (\mu, \sigma^2)$

- $\overline{Y}$ and S^2 are independent
- $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, where χ^2_ν is the Chi-square distribution with ν degrees of freedom

Pivotal Quantity (Focus on μ)

- $T = \frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ where t_ν denotes the Student's t distribution with ν degrees of freedom

Confidence Interval for μ at level $1-\alpha$

- For $0 < \alpha < 1$: $\Pr(\mathrm{t}_{\mathrm{n}-1;\alpha/2} \leq \mathrm{T} \leq \mathrm{t}_{\mathrm{n}-1;1-\alpha/2}) = 1-\alpha$
- The realization of the confidence interval is

$$IC_\mu^{1-\alpha} = (\bar{y} - t_{n-1;1-\alpha/2}s/\sqrt{n}, \bar{y} + t_{n-1;1-\alpha/2}s/\sqrt{n})$$

where $-t_{n-1;1-\alpha/2} = t_{n-1;\alpha/2}$ and $t_{n-1;1-\alpha/2}$ represent the quantiles of order $\alpha/2$ and $1-\alpha/2$ of a Student's t distribution with $n-1$ degrees of freedom, respectively.

Statistical Model: Inferential Results

Confidence Intervals: Exercise

- Let us to generate $B = 1000$ samples of size $n = 10$ from a $N(\mu = 5, \sigma^2 = 4)$ and fix $\alpha = 0.05$
- Compute the CI for each replication
- Count how many intervals includes the true parameter

```
mu <- 5; sigma <- 2; B <- 1000; n <- 10; alpha <- 0.05
CI <- matrix(0, B, 2)
l <- rep(0, B)

set.seed(23)
q0975 <- qt(1 - alpha/2, n - 1)

for (i in 1 : B){
 x <- rnorm(n, mu, sigma)
 CI[i, ] <- mean(x) + c(-q0975, q0975) * sd(x)/sqrt(n)
 l[i] <- (mu > CI[i,1] & mu < CI[i,2])
}
```

Statistical Model: Inferential Results

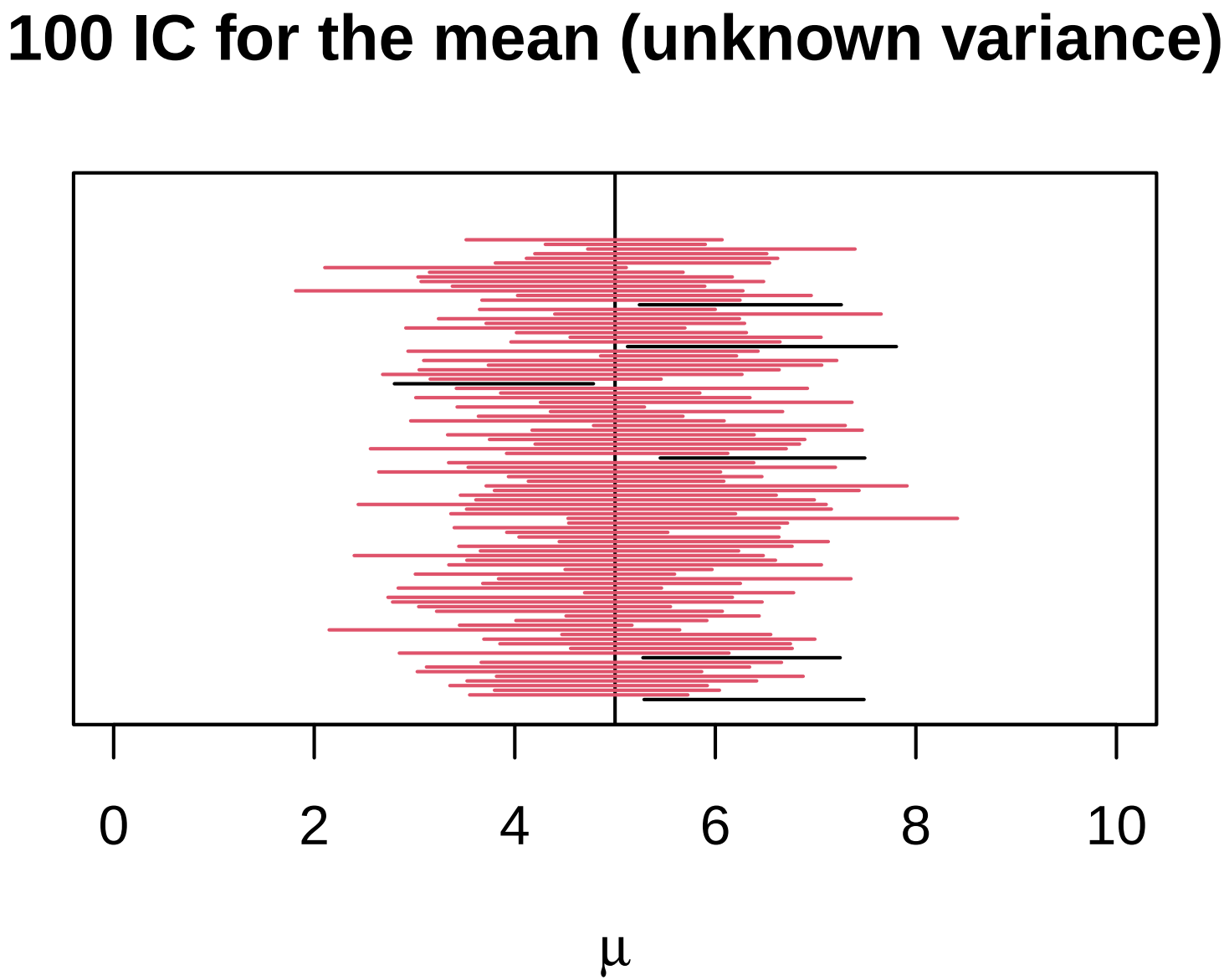
```r
#Empirical coverage probability
mean(l)
```

```
## [1] 0.956
```

```r
# Plot of the (first 100) CIs
plot(1, xlim = c(0,10), ylim = c(0,11), type = "n",
     xlab = expression(mu), ylab = "", yaxt  = "n",
     main = paste("100 IC for the mean (unknown variance)")
abline(v = mu)

d <- 0
for (i in 1 : 100){
  d <- d + 0.1
  lines(seq(CI[i, 1], CI[i, 2], length = 100), rep(d, 100)
}
```

Statistical Model: Inferential Results



The Normal Model: Inferential Results

Hypothesis test

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

- The reject region of level α associated to the test is

$$R_\alpha = \left\{ (y_1, \ldots, y_n) : \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| > t_{n-1;1-\alpha/2} \right\}$$

- The p-value (the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value) is

$$p - value = \Pr\left(|T| > \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \right)$$

The Normal Model: Inferential Results

Example

To check whether the declared weight on a package corresponds to the truth, a pasta manufacturer analyzes the weight of a sample of $n = 10$ packages of pasta. The packages have a declared weight of $500g$, and it is reasonable to assume that the weight of a package is a normally distributed random variable. To verify that the machine does not produce packages with a different weight, it is checked whether the sample mean weight deviates too much from 500; if it does, production is suspended.
To decide when the sample mean weight is too far from 500, a 5% rejection region is used. The hypothesis system is:

$$\begin{cases} H_0 : \mu = 500 \\ H_1 : \mu \neq 500 \end{cases}$$

The sample has given the following measurements:

497.0621, 497.7690, 497.2631, 498.6688, 494.6227,

496.2185, 498.2635, 502.7199, 497.4339, 494.2883

The Normal Model: Inferential Results

```
y <- c(497.0621, 497.7690, 497.2631, 498.6688, 494.6227,
       496.2185, 498.2635, 502.7199, 497.4339, 494.2883)
t_obs <- (mean(y) - 500)/(sd(y)/sqrt(length(y)))
curve(dt(x, 9), xlim = c(-5, 5), ylim = c(0, 0.4),
      main = "Rejection region", col = "blue",
      lwd = 2, xlab = "",  ylab = expression(t[9]),  yaxs="i")
cord.x0 <- c(-5, seq(-5, qt(0.025, 9), 0.01),  qt(0.025, 9))
cord.y0 <- c(0, dt(seq(-5,qt(0.025, 9), 0.01), 9), 0)
cord.x1 <- c(qt(0.975, 9),seq(qt(0.975, 9), 5, 0.01), 5)
cord.y1 <- c(0, dt(seq(qt(0.975, 9), 5, 0.01), 9), 0)
polygon(cord.x0, cord.y0, col = "darkred", border = NA )
polygon(cord.x1, cord.y1, col = "darkred", border = NA )

abline(v = t_obs, lty = 2, lwd = 2, col = "red")
text(0, 0.2, paste("Accept", expression(H0)))
text(c(-2.7,2.7), 0.08, paste("Reject", expression(H0)), col = "darkred
text(as.double(t_obs) - 0.15, 0.02, "t", col = "red", cex = 1.2)
```

The Normal Model: Inferential Results

