BIOSTAT 701

Introduction to Statistical Theory and Methods I

Lynn Lin

Preparation

- **Normal distribution**: https://duke.zoom.us/rec/share/Ob9Mo-eJOFs-n_8SBx5Jor4RGZnDtm4RarnI3Dv3ywFpsG_mEndrB-v3gNNRzzDE.2TI-Sv-QZoiI6T-K?startTime=1648836761000
- **Student’s t distribution**: https://duke.zoom.us/rec/share/X0CUe6fqIAfKmXpmRUcatIhMEHph79kCOUdzgNbtYNsYvWCzJfi1iTWJCco5oeV8.pvdVGcpKruHkQTHs?startTime=1648837423000

Probability density function for continuous r.v.

- A function $f(x)$ is a probability density for a continuous random variable X if $P(X \in [a, b]) = \int_a^b f(x)dx$, i.e., $P(X \in [a, b])$ is the area under $f(x)$ over the interval $[a, b]$.
- **Remark**: Note that if X is continuous then $P(X = x) = 0$ for any given value x.
- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f(x)dx = 1$, i.e., the area under the entire graph of $f(x)$.

CDF for continuous r.v.

- The cumulative distribution function $F(x)$ for a continuous r.v. X is defined
- $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)d(y)$
- For each x, $F(x)$ is the area under the density curve to the left of x.

Expected value and variance

- If X is a continuous RV with density f(x), the expected value or the mean of X is defined as $\mu_X = E(X) = \int_{-\inf}^{\inf} xf(x)dx$

- The variance of X is defined as $\sigma_X^2 = V(X) = \int_{-\inf}^{\inf} (x - \mu_X)^2 f(x)dx$

Example

- The density of X is a constance 1 on [0,1] and 0 elsewhere. What is the mean and variance of X?

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 x\, dx = \frac{1}{2} x^2 \Big|_0^1 = \frac{1}{2}$$

$$f(X) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{else.} \end{cases}$$

$$\operatorname{var} X = E(X-E(X))^2 = EX^2 - [E(X)]^2$$

$$= \int_0^1 x^2 dx - \left(\frac{1}{2}\right)^2$$

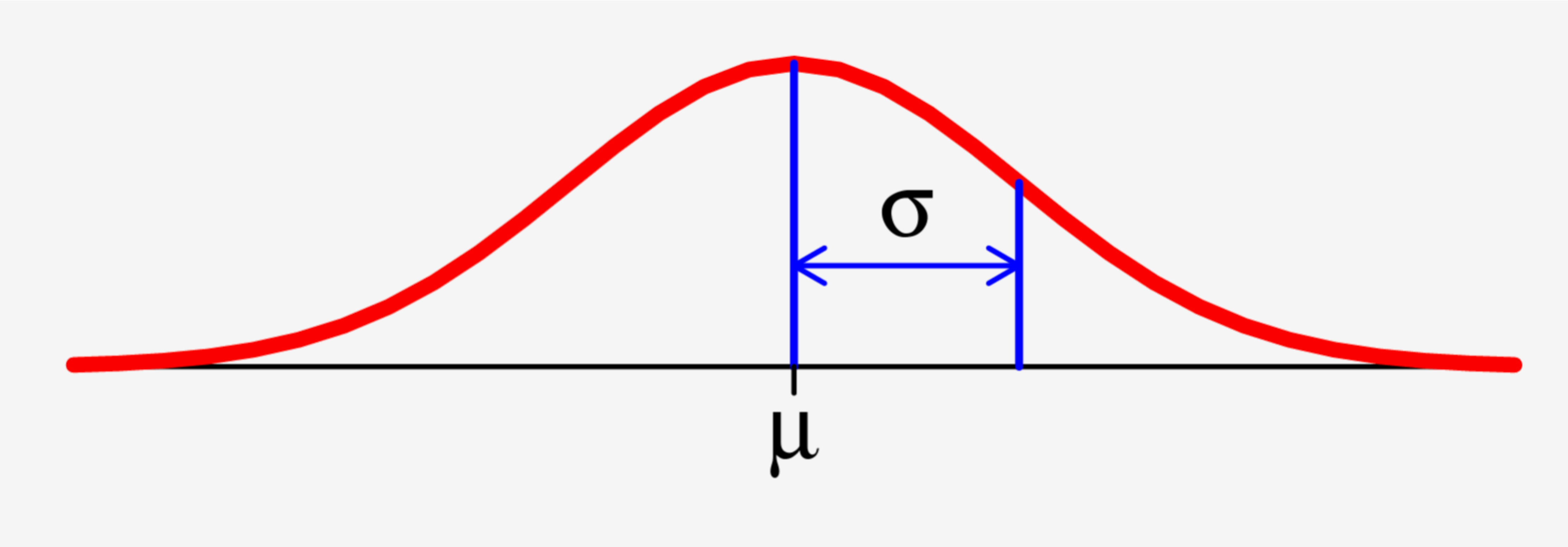
$$= \frac{1}{12} .$$

Normal distribution

- Normal distribution is a family of continuous distributions that can model many histograms of real-life data which are mound shape and symmetric (such as height, weight,...).
- A normal distribution has two parameters: 1. mean μ (center of the curve). 2. standard deviation σ (spread about the center).
- Thus, a random variable X which follows a normal distribution can be written as $X \sim N(\mu, \sigma)$

Normal distribution

- The formula for $N(\mu, \sigma)$ curve is $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$



Normal distribution

PDF : $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

- In particular, N(0,1) is called “standard normal distribution”.
- Z is typically used to denote standard normal R.V.

Example

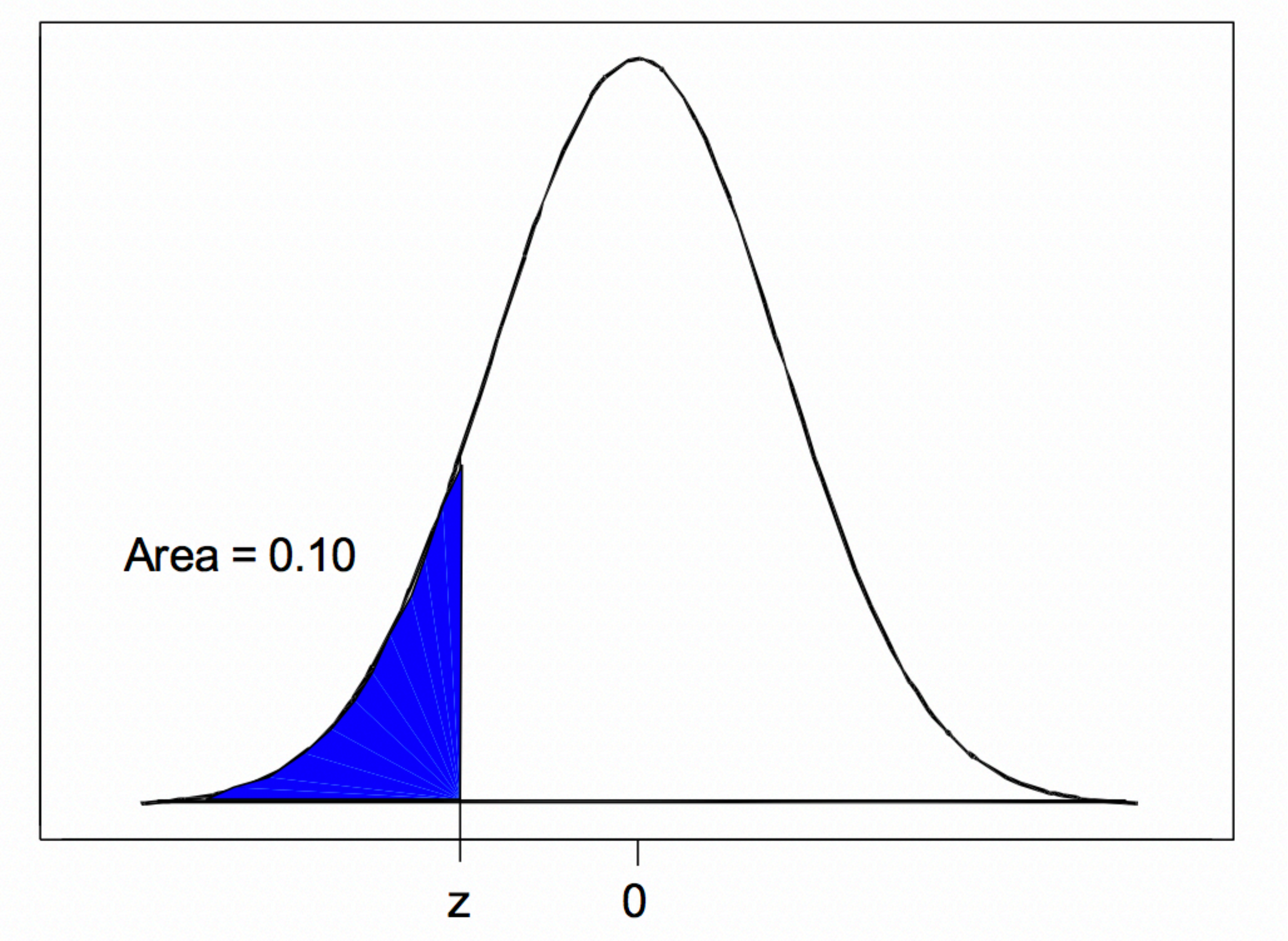
- Find the area under the standard normal curve to the right of 0.87.
- Answer: $P(Z > 0.87) = 1 - P(Z < 0.87) = 1 - 0.8078 = 0.1922$
- The R command pnorm() can find areas under the standard normal curve

Find z-value

$E[z]=0$ $Var[z]=1$

- The above example is to find area (probabilities) given the z-value. Next, we will show how to find the z-value given the area (probabilities).

- Example: Find the 10-th percentile of the standard normal distribution.



General normal distribution

- Find the probability given the range of observations: The weight of 10 year old girls are known to be normally distributed with mean 70 pounds and standard deviation of 13 pounds. Obtain the percentage of 10 year old girls with weight between 60 pounds and 90 pounds.

$$Z = \frac{X - \mu}{\sigma}$$

$$X = 60. \quad Z_1 = \frac{60-70}{13} = -\frac{10}{13}$$

$$X = 90 \quad Z_2 = \frac{90-70}{13} = \frac{20}{13}$$

$$P(Z \leq -0.77) \approx 0.2206$$

$$P(Z \leq 1.54) \approx 0.9382$$

$$\therefore P(60 \leq X \leq 90)$$

$$= P(Z \leq 1.54) - P(Z \leq -0.77)$$

$$= 0.7176$$

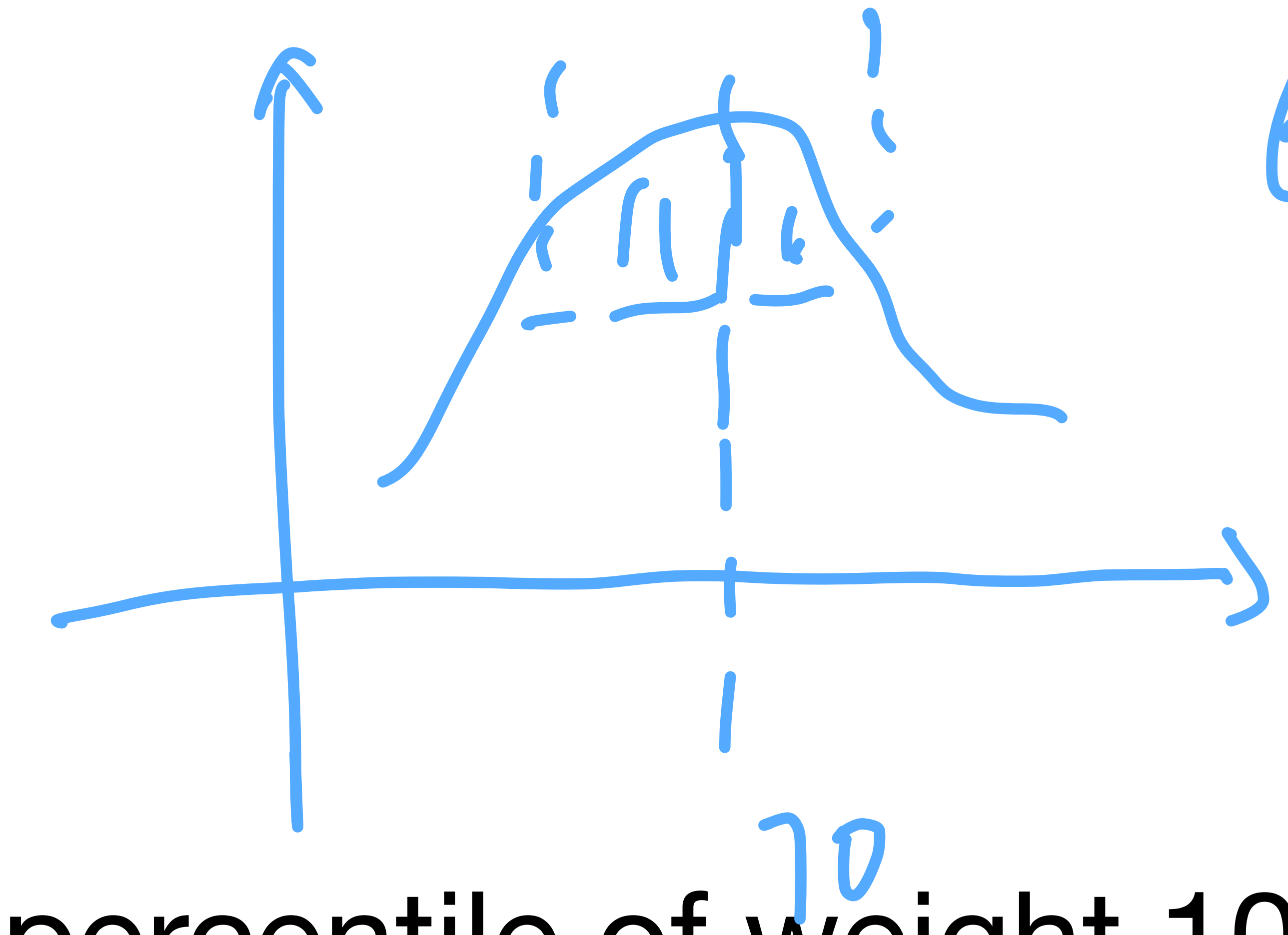
Empirical rule

- Recall that the empirical rule we mentioned states that if the set of measurements follow a bell shaped distribution, then
- $\bar{x} \pm s$ contains about 68% of the data
- $\bar{x} \pm 2s$ contains about 95% of the data
- $\bar{x} \pm 3s$ contains about all data

Validate this rule

- $P(-1 < Z < 1) = P(Z < 1) - P(Z < -1) = 0.8413 - 0.1587 = 0.6826$
- $P(-2 < Z < 2) = P(Z < 2) - P(Z < -2) = 0.9772 - 0.0228 = 0.9544$
- $P(-3 < Z < 3) = P(Z < 3) - P(Z < -3) = 0.9987 - 0.0013 = 0.9974$

Example



$\sigma = 13$

- Find the 60-th percentile of weight 10 years old girls given that the weight is normally distributed with mean 70 pounds and standard deviation of 13 pounds.

$$\text{qnorm}(0.6, \mu=0, \sigma=1) = 0.253$$

$$z = \frac{x-\mu}{\sigma} \qquad 0.253 = \frac{x-70}{13}$$

$$x = 73.29$$

$$\text{qnorm}(0.6, 70, 13)$$

Example

$x = z\sigma + \mu$

$X = 99.1F$

$X = 98.2 + 1.28 \times 0.73$

$z_{0.90} = 1.28$

- Body temperatures of healthy humans are distributed nearly normally with mean 98.2 F and standard deviation 0.73 F. What is the cutoff for the lowest 3% of human body temperatures?

$\mu = 98.2$ $\sigma = 0.73$

$P(X \geq x) = 0.10 \Leftrightarrow P(X \leq x) = 0.90$

- What is the cutoff for the highest 10% of human body temperatures?

$X \sim N(\mu = 98.2, \sigma = 0.73)$

$P(X \leq x) = 0.03$

$Z = \frac{X - \mu}{\sigma}$

$X = \mu + z \cdot \sigma$

$P(Z \leq z) = 0.03$

$= 98.2 + (-1.88) \times 0.73$

$z_{0.03} \approx -1.88$

$X \approx 98.2 - 1.37$

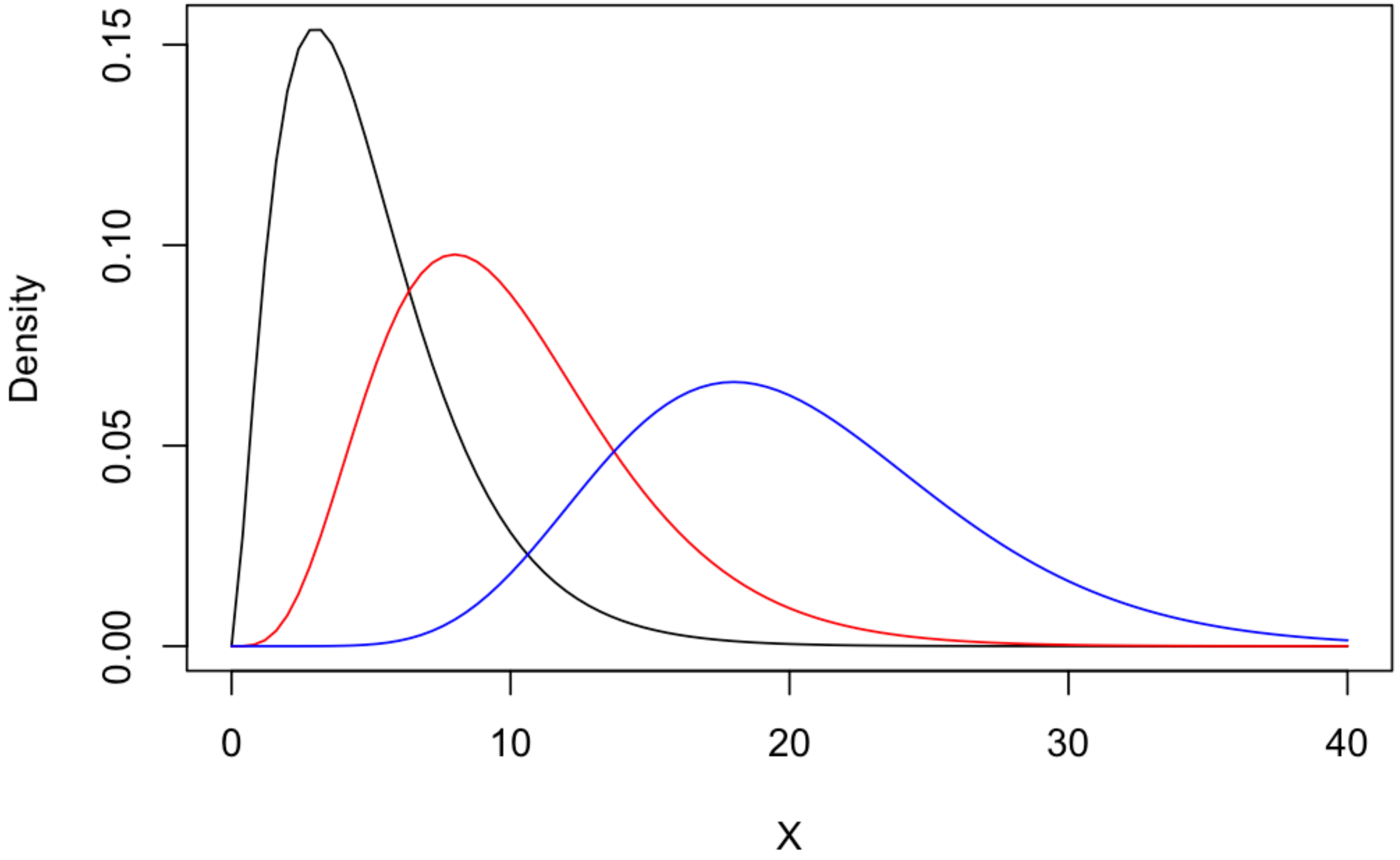
$= 96.8F$

Chi-Square distribution

- χ^2-distribution is related to the normal distribution
- Suppose $Z \sim N(0,1)$, then $Z^2 \sim \chi_1^2$
- Suppose $Z_1, Z_2, \ldots, Z_p \sim N(0,1)$ and are mutually independent; then $\sum_{i=1}^{p} Z_i^2 \sim \chi_p^2$, p is the degrees of freedom (df)
- PDF: $f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} e^{-x/2}$

Chi-Square distribution

- χ^2-distributions will have different shapes depending on the number of degrees of freedom



```
curve(dchisq(x, df = 5), from = 0, to = 40,
ylab="Density", xlab="X")
curve(dchisq(x, df = 10), from = 0, to = 40,
add=TRUE, col="red")
curve(dchisq(x, df = 20), from = 0, to = 40,
add=TRUE, col="blue")
```

Chi-Square distribution

- If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are mutually independent, then $\sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma})^2 \sim \chi_n^2$
- If we let $S = \sum (X_i - \mu)^2 / n$, then $nS^2/\sigma^2 \sim \chi_n^2$

Chi-Square distribution

- First, they are all bounded on the left side by zero.
- Second, they are not symmetrical although they become more symmetrical as the df increases.

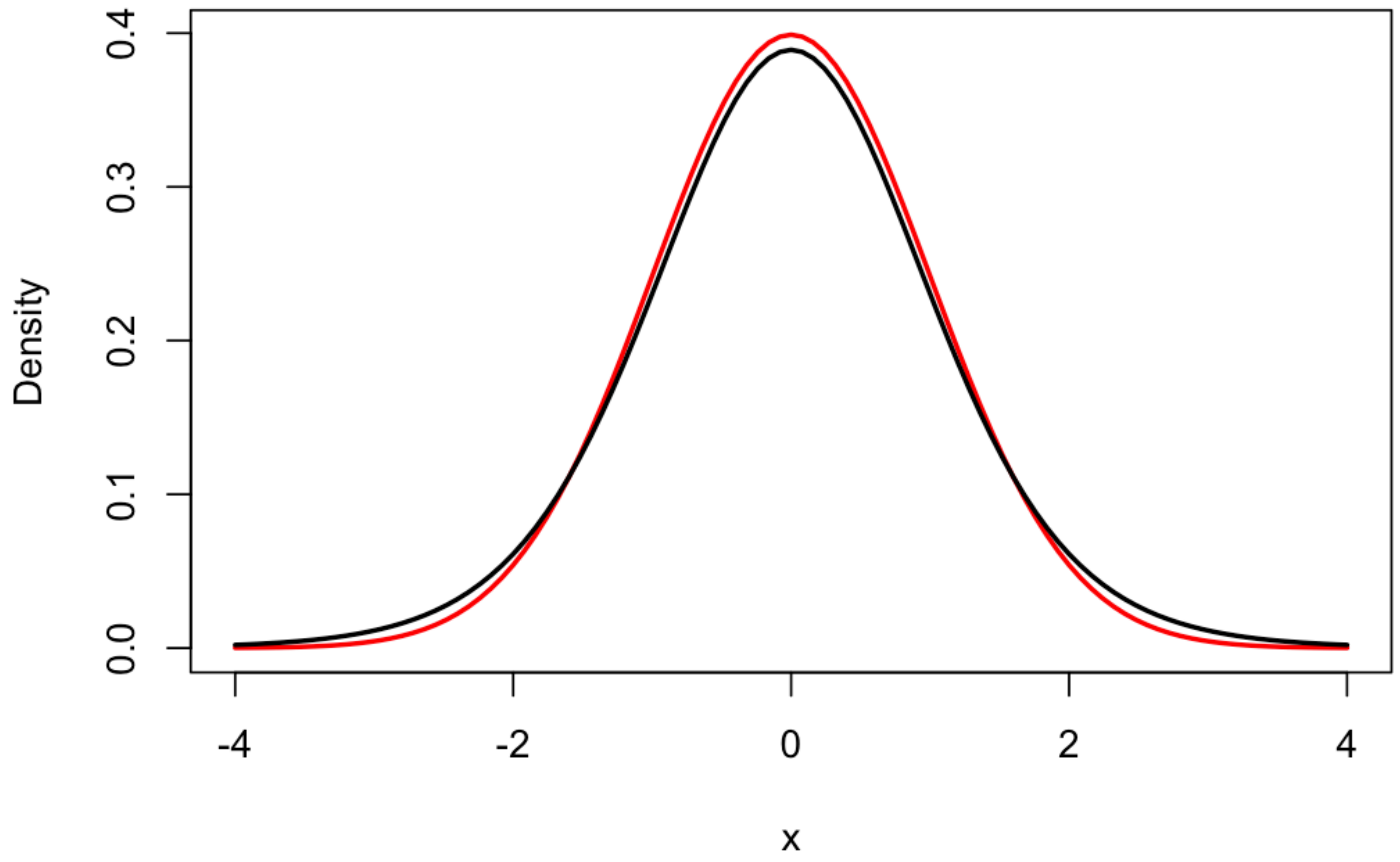
The Student's t-distribution

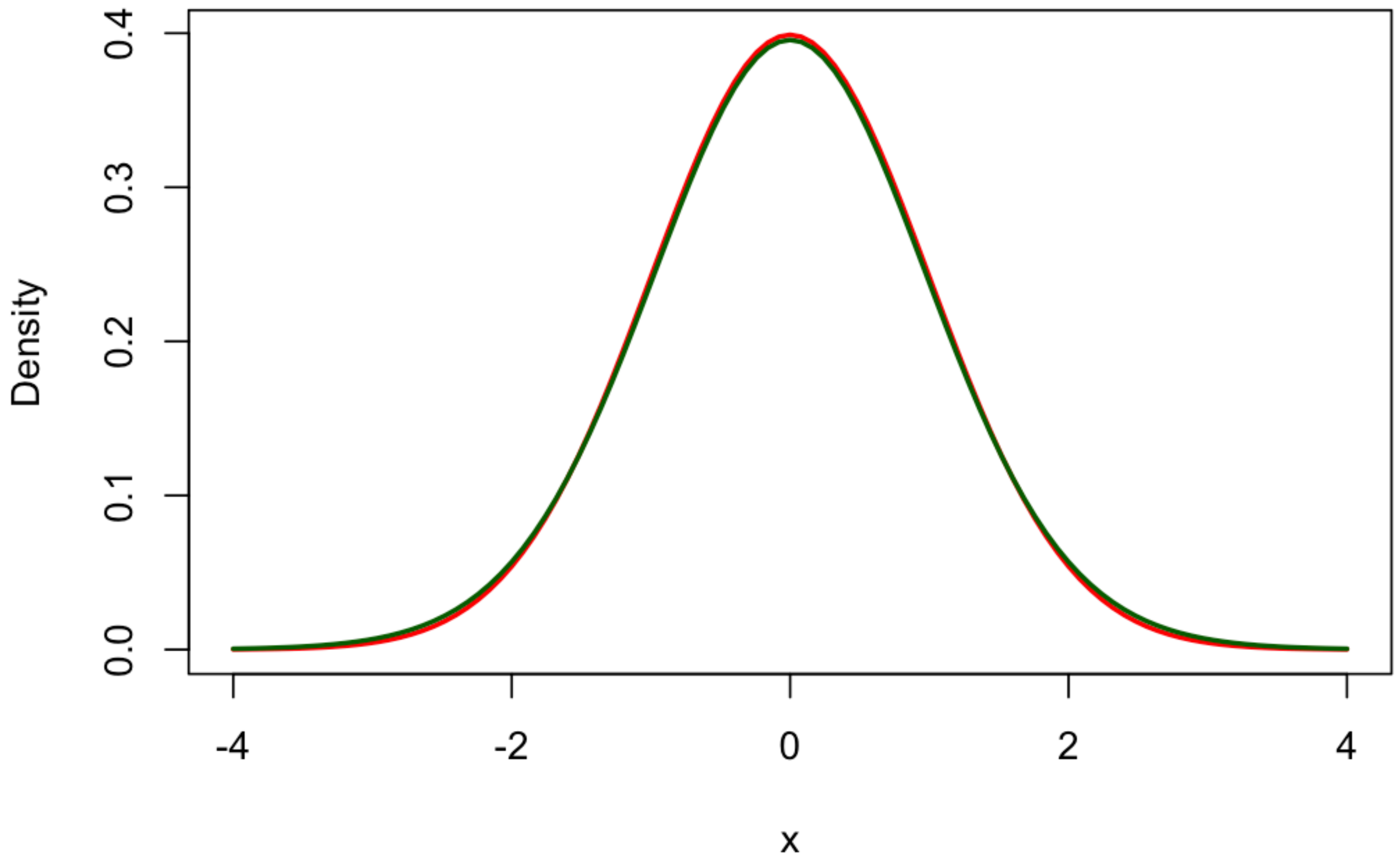
- Suppose $Z \sim N(0,1)$, $X^2 \sim \chi^2_n$, and Z and X^2 are independent. Then, $\dfrac{Z}{\sqrt{X^2/n}} \sim t_n$, i.e., the t distribution with n df.

The Student’s t-distribution

- The normal and t-distributions are similar:
  - Both are symmetric around mean
  - Both have positive support over the entire real line
  - As the df goes up, the t-distribution converges to the normal distribution
- However, the tails of the t-distribution are thicker than those of the normal distribution. And the difference can be large with small df.

The Student’s t-distribution



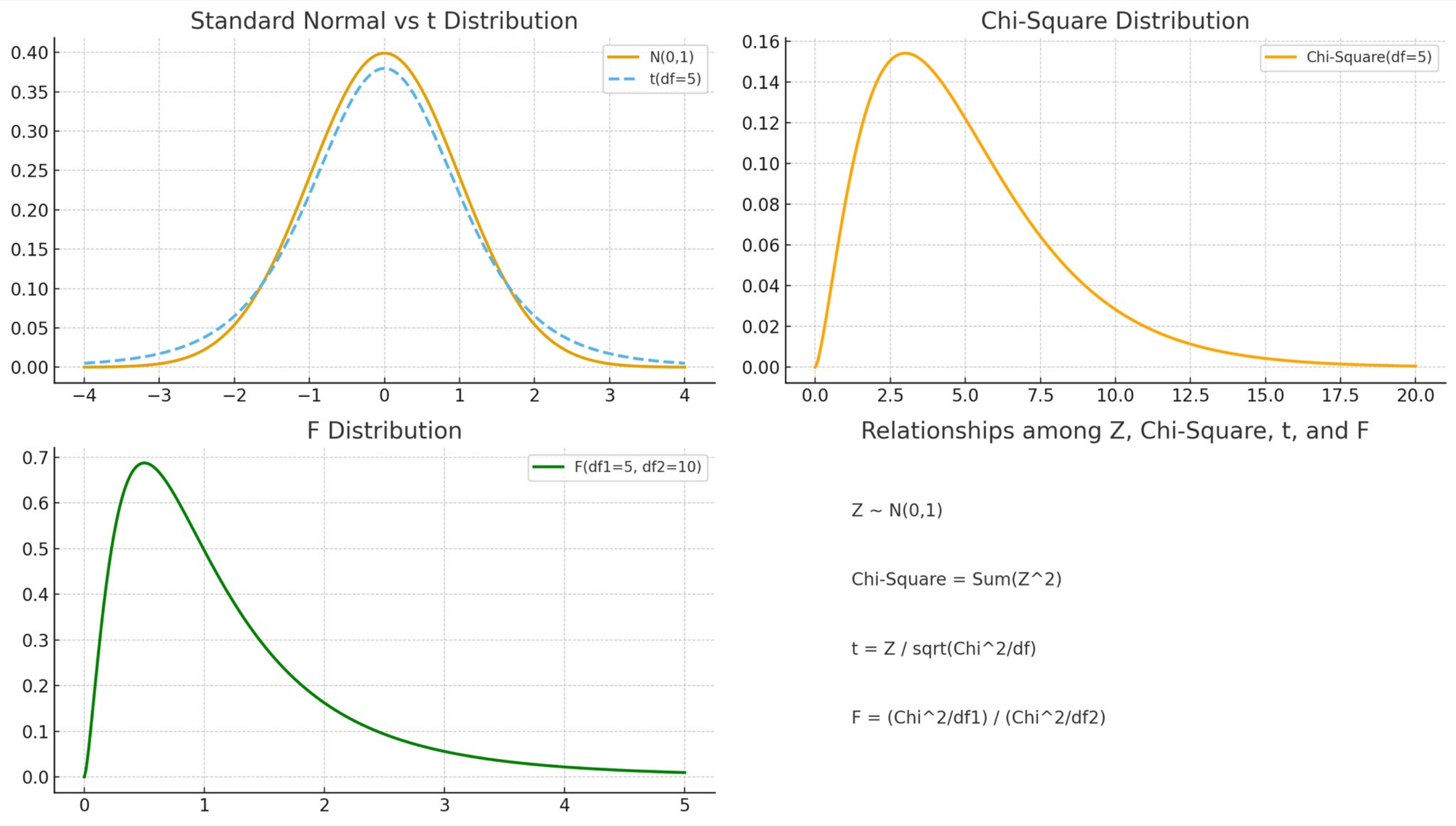


```
# Code for the left side plot
curve(dnorm(x), -4, 4, col = "red",ylab="Density",lwd=2)
curve(dt(x, df = 10), add = TRUE, col="black",lwd=2)

# Code for the right side plot
curve(dnorm(x), -4, 4, col = "red",ylab="Density",lwd=2)
curve(dt(x, df = 30), add = TRUE, col="darkgreen",lwd=2
```

F distribution

- If χ_a and χ_b are independent χ^2 RVs with a and b df, respectively.
- Then: $\dfrac{\chi_a/a}{\chi_b/b} \sim F_{a,b}$, which is an F distribution with df (a,b).
- As with the Chi-square and t-distributions, the F distributions all have different shapes according to the number of degrees of freedom.
- If $X \sim t_b$, then $X^2 \sim F(1,b)$



Standard Normal vs t Distribution

Chi-Square Distribution

F Distribution

Relationships among Z, Chi-Square, t, and F

Z ~ N(0,1)

Chi-Square = Sum(Z^2)

t = Z / sqrt(Chi^2/df)

F = (Chi^2/df1) / (Chi^2/df2)

Exponential distribution

- The exponential distribution is the continuous analog of the geometric distribution.
- It is often used to model the time to first event.
- $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$
- Or equivalently, $f(x) = \lambda e^{-\lambda x} 1_{(0,\infty)}(x)$
- λ is the rate parameter.

Indicator function

- The indicator function of a set A, denoted by $1_A(x)$, is the function

$$1_A x = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Exponential distribution

- The exponential distribution "lacks memory", and is a natural choice when the past behavior of a system provides no information about future behavior.
  - If $X \sim Exp(\lambda)$, then $P(X > x + c \,|\, X > c) = P(X > x)$.
- One application of the exponential distribution is in survival analysis when the hazard of death is constant.