Chapter 5. Introduction to Probability Theory*

This chapter introduces some elementary probability theory which is the base of statistical inference in the next chapter. The foundations of modern probability theory were laid by Andrey Nikolaevich Kolmogorov (1903-1987) in 1933. Related materials are Chapter 1-5 of Casella and Berger (2002) and Appendix B of Hansen (2016). More advanced references include Ash (1972), Feller (1968, 1970), Billingsley (1995), Chung (2001), Dudley (2002) and Durrett (2010) among others.

1 Foundations

The set Ω of all possible outcomes of an experiment is called the **sample space** for the experiment. Take the simple example of tossing a coin. There are two outcomes, heads and tails, so we can write $\Omega = \{H, T\}$. If two coins are tossed in sequence, we can write the four outcomes as $\Omega = \{HH, HT, TH, TT\}$.

An **event** A is any collection of possible outcomes of an experiment. An event is a subset of Ω, including Ω itself and the null set $\emptyset$. Continuing the two coin example, one event is $A = \{HH, HT\}$, the event that the first coin is heads. We say that A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$. For example, the sets $\{HH, HT\}$ and $\{TH\}$ are disjoint.

A **probability function** P assigns probabilities (numbers between 0 and 1) to events A in Ω. This is straightforward when Ω is countable; when Ω is uncountable we must be somewhat more careful. A collection of sets, $\mathcal{F}$, is called a **sigma algebra** (or σ**-field**) if $\emptyset \in \mathcal{F}$, $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, and $A_1, A_2, \cdots \in \mathcal{F}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. We only define probabilities for events contained in $\mathcal{F}$. A simple example of $\mathcal{F}$ is $\{\emptyset, \Omega\}$ which is known as the **trivial** σ**-algebra**. For any sample space Ω, let $\mathcal{F}$ be the smallest sigma algebra which contains all of the open sets in Ω. When Ω is countable, $\mathcal{F}$ is simply the collection of all subsets of $\mathcal{F}$, including $\emptyset$ and Ω. When Ω is the real line, then $\mathcal{F}$ is the collection of all open and closed intervals. We call such an $\mathcal{F}$ the **Borel** σ**-algebra** associated with Ω, and a set $A \in \mathcal{F}$ a **Borel** (measurable) **set**.

We now can give the axiomatic definition of probability. Given Ω and $\mathcal{F}$, a probability function P satisfies $P(\Omega) = 1$, $P(A) \geq 0$ for all $A \in \mathcal{F}$, and if $A_1, A_2, \cdots \in \mathcal{F}$ are pairwise disjoint, then $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{n} P(A_i)$.

Some important properties of the probability function include the following

1. $P(\emptyset) = 0$

*Email: pingyu@hku.hk

2. $P(A) \leq 1$

3. $P(A^c) = 1 - P(A)$

4. $P(B \cap A^c) = P(B) - P(A \cap B)$

5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

6. If $A \subseteq B$ then $P(A) \leq P(B)$

7. Bonferroni's Inequality: $P(A \cap B) \geq P(A) + P(B) - 1$

8. Boole's Inequality: $P(A \cup B) \leq P(A) + P(B)$

2 Random Variables

A **random variable** (r.v.) X is a measurable function from a sample space Ω into the real line, i.e., for any Borel set A in $\mathbb{R}$, the inverse image of A is in $\mathcal{F}$. This induces a new sample space - the real line - and a new probability function on the real line. Typically, we denote r.v.'s by uppercase letters such as X, and use lowercase letters such as x for potential values and realized values. For a r.v. X we define its **cumulative distribution function** (cdf) as

$$F(x) = P(X \leq x).$$

Sometimes we write this as $F_X(x)$ to denote that it is the cdf of X.

We say that the r.v. X is **discrete** if $F(x)$ is a step function. In this case, X can take only a countable set of real numbers $x_1, \cdots, x_J$, where J can be ∞. The probability function for X takes the form of the **probability mass function** (pmf)

$$P(X = x_j) = p_j,\ j = 1, \cdots, J, \tag{1}$$

where $0 \leq p_j \leq 1$ and $\sum_{j=1}^{J} p_j = 1$. The cdf of X is $F(x) = \sum_{j=1}^{J} p_j 1(x_j \leq x)$, where $1(\cdot)$ is the indicator function which equals one when the event in the parenthesis is true and zero otherwise. A famous discrete r.v. is the Bernoulli r.v., where $J = 2$, $x_1 = 0$, $x_2 = 1$, $p_2 = p$ and $p_1 = 1 - p$. We say that the r.v. X is **continuous** if $F(x)$ is continuous in x. In this case $P(X = x) = 0$ for all $x \in \mathbb{R}$ so the representation (1) is unavailable. Instead, we represent the relative probabilities by the **probability density function** (pdf)

$$f(x) = \frac{d}{dx} F(x)$$

so that

$$F(x) = \int_{-\infty}^{x} f(u) du$$

and

$$P\left(a \leq X \leq b\right) = \int_a^b f(u)du.$$

These expressions make sense only if $F(x)$ is differentiable. While there are examples of continuous r.v.'s which do not possess a pdf, these cases are unusual and are typically ignored. A function $f(x)$ is a pdf iff $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{-\infty}^{\infty} f(x)dx = 1$. One famous continuous r.v. is the normal r.v. which will be discussed in Section 6. Another famous continuous r.v. is the standard **uniform** r.v. whose pdf is $f(x) = 1\left(0 \leq x \leq 1\right)$, i.e., X can occur only on $[0,1]$ and occurs uniformly. We denote $X \sim U[0,1]$. A generalization is the uniform r.v. on $[a,b]$, $a < b$, whose pdf is $f(x) = 1\left(a \leq x \leq b\right)/(b-a)$. We denote $X \sim U[a,b]$.

3 Expectation

For any real function g, we define the **mean** or **expectation** $E\left[g(X)\right]$ as follows. If X is discrete,

$$E\left[g(X)\right] = \sum_{j=1}^{J} g(x_j)p_j,$$

and if X is continuous

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

To cover both the discrete and continuous case (and even more general cases), we often write $E\left[g(X)\right]$ as a Riemann-Stieltjes integral $\int g(x)dF(x)$ (why intuitively correct? see Chapter 1).

Since $E\left[a + bX\right] = a + b \cdot E\left[X\right]$, we say that expectation is a linear operator.

For $m > 0$, we define the mth **moment** of X as $E\left[X^m\right]$, the mth **central moment** as $E\left[(X - E[X])^m\right]$, the mth **absolute moment** of X as $E\left[|X|^m\right]$, and the mth **absolute central moment** as $E\left[|X - E[X]|^m\right]$.

Two special moments are the **mean** $\mu = E[X]$ and **variance** $\sigma^2 = E\left[(X-\mu)^2\right] = E\left[X^2\right] - \mu^2$. We call $\sigma = \sqrt{\sigma^2}$ the **standard deviation** of X. We can also write $\sigma^2 = Var(X)$. For example, this allows the convenient expression $Var(a + bX) = b^2 Var(X)$.

For a r.v. X, the **standardized r.v.** is defined as

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

Let $a = -\mu/\sigma$ anb $b = 1/\sigma$, we have

$$E\left[Z\right] = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\mu = 0,$$

and

$$Var\left(Z\right) = \frac{Var\left(X\right)}{\sigma^2} = 1.$$

This transformation is frequently used in statistical inference.

Skewness is a measure of asymmetry of a distribution. For a r.v. X, its skewness is defined as $E\left[Z^3\right] = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$. If X has a symmetric distribution about μ, then its skewness is zero. If X has a long right tail, then its skewness is positive, and X is called **positive-** (or **right-**) **skewed**. If X has a long left tail, then its skewness is negative, and X is called **negative-** (or **left-**) **skewed**. **Kurtosis** is a measure of the heavy-tailedness of a distribution. For a r.v. X, its kurtosis is defined as $E\left[Z^4\right] = \frac{E\left[(X-\mu)^4\right]}{\sigma^4}$. The normal distribution has kurtosis 3: $E\left[Z^4\right] = 3$. If the kurtosis of a distribution is greater than 3, then it is called **heavy-tailed** or **leptokurtoic**.

4 Multivariate Random Variables

A pair of bivariate r.v.'s (X, Y) is a function from the sample space into $\mathbb{R}^2$. The joint cdf of (X, Y) is

$$F(x, y) = P\left(X \leq x, Y \leq y\right).$$

If F is continuous, the joint probability density function is

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

For a Borel measurable set $A \subset \mathbb{R}^2$,

$$P\left((X, Y) \in A\right) = \int \int_A f(x, y) dx dy.$$

For any measurable function $g(x, y)$,

$$E\left[g(X, Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

The **marginal distribution** of X is

$$F_X(x) = P(X \leq x) = \lim_{y \to \infty} F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y) dy dx,$$

so the marginal density of X is

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Similarly, the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

The r.v.'s X and Y are defined to be **independent** if $f(x, y) = f_X(x) f_Y(y)$. If X and Y are

independent, then

$$\begin{aligned} E\left[g(X)h(Y)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy \\ &= E\left[g(X)\right]E\left[h(Y)\right], \end{aligned} \tag{2}$$

if the expectations exist. For example, if X and Y are independent then

$$E\left[XY\right] = E\left[X\right]E\left[Y\right].$$

The **covariance** between X and Y is

$$Cov(X,Y) = \sigma_{XY} = E\left[\left(X - E\left[X\right]\right)\left(Y - E\left[Y\right]\right)\right] = E\left[XY\right] - E\left[X\right]E\left[Y\right],$$

whose unit is the unit of X times the unit of Y. The **correlation** between X and Y is

$$Corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

The Cauchy-Schwarz Inequality implies that

$$|\rho_{XY}| \leq 1.$$

The correlation is a measure of *linear* dependence, free of units of measurement.

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$. The reverse, however, is not true. For example, if $E\left[X\right] = 0$ and $E\left[X^3\right] = 0$, then $Cov(X, X^2) = 0$.

A useful fact is that

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

An implication is that if X and Y are independent, then

$$Var(X+Y) = Var(X) + Var(Y),$$

the variance of the sum is the sum of the variances.

A $k \times 1$ random vector $\boldsymbol{X} = (X_1, \cdots, X_k)'$ is a function from Ω to $\mathbb{R}^k$. Let $\boldsymbol{x} = (x_1, \cdots, x_k)'$ denote a vector in $\mathbb{R}^k$. The vector $\boldsymbol{X}$ has the distribution and density functions

$$\begin{aligned} F(\boldsymbol{x}) &= P\left(\boldsymbol{X} \leq \boldsymbol{x}\right), \\ f(\boldsymbol{x}) &= \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F(\boldsymbol{x}). \end{aligned}$$

For a measurable function $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^s$, we define the expectation

$$E\left[\mathbf{g}(\boldsymbol{X})\right] = \int_{\mathbb{R}^k} \mathbf{g}(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x},$$

where the symbol $d\boldsymbol{x}$ denotes $dx_1 \cdots dx_k$. In particular, we have the $k \times 1$ multivariate mean

$$\boldsymbol{\mu} = E\left[\boldsymbol{X}\right]$$

and $k \times k$ covariance matrix

$$\boldsymbol{\Sigma} = E\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)'\right] = E\left[\boldsymbol{X}\boldsymbol{X}'\right] - \boldsymbol{\mu}\boldsymbol{\mu}'.$$

If the elements of $\boldsymbol{X}$ are mutually independent, then $\boldsymbol{\Sigma}$ is a diagonal matrix and

$$Var\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} Var\left(X_i\right).$$

5 Conditional Distributions and Expectation

The **conditional density** of Y given $\boldsymbol{X} = \boldsymbol{x}$ is defined as

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

if $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$.

The **conditional mean** or **conditional expectation** is the function

$$m(\boldsymbol{x}) = E\left[Y|\boldsymbol{X} = \boldsymbol{x}\right] = \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) dy.$$

The conditional mean $m(\boldsymbol{x})$ is a function, meaning that when $\boldsymbol{X}$ equals $\boldsymbol{x}$, then the expected value of Y is $m(\boldsymbol{x})$. Similarly, we define the **conditional variance** of Y given $\boldsymbol{X} = \boldsymbol{x}$ as

$$\begin{aligned}
\sigma^2(\boldsymbol{x}) &= Var\left(Y|\boldsymbol{X} = \boldsymbol{x}\right) \\
&= E\left[\left(Y - m(\boldsymbol{x})\right)^2 \middle| \boldsymbol{X} = \boldsymbol{x}\right] \\
&= E\left[Y^2 \middle| \boldsymbol{X} = \boldsymbol{x}\right] - m(\boldsymbol{x})^2.
\end{aligned}$$

If Y and $\boldsymbol{X}$ are independent, then $E\left[Y|\boldsymbol{X} = \boldsymbol{x}\right] = E\left[Y\right]$ and $Var\left(Y|\boldsymbol{X} = \boldsymbol{x}\right) = Var\left(Y\right)$. (why?)

6 Normal and Related Distributions

The **standard normal** density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), -\infty < x < \infty.$$

It is conventional to write $X \sim N(0,1)$, and to denote the standard normal density function by $\phi(x)$ and its distribution function by $\Phi(x)$. The latter has no closed-form solution. The normal density has all moments finite. Since it is symmetric about zero all odd moments are zero. By iterated integration by parts, we can also show that $E\left[X^2\right] = 1$ and $E\left[X^4\right] = 3$. In fact, for any positive integer m, $E\left[X^{2m}\right] = (2m-1)!! = (2m-1)(2m-3)\cdots 1$. Thus $E\left[X^4\right] = 3$, $E\left[X^6\right] = 15$, $E\left[X^8\right] = 105$, and $E\left[X^{10}\right] = 945$.

If Z is standard normal and $X = \mu + \sigma Z$, then X has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty.$$

which is the **univariate normal density**. The mean and variance of the distribution are μ and σ^2, and it is conventional to write $X \sim N(\mu, \sigma^2)$.

For $\boldsymbol{x} \in \mathbb{R}^k$, the **multivariate normal density** is

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \boldsymbol{x} \in \mathbb{R}^k.$$

The mean and covariance matrix of the distribution are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. When $\boldsymbol{X}$ has the multivariate normal density, we call $\boldsymbol{X}$ follows the multivariate normal distribution and write $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

If $\boldsymbol{X} \in \mathbb{R}^k$ is multivariate normal and the elements of $\boldsymbol{X}$ are mutually uncorrelated, then $\boldsymbol{\Sigma} = \text{diag}\left\{\sigma_j^2\right\}$ is a diagonal matrix. In this case the density function can be written as

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{\sqrt{2\pi}\sigma_1\cdots\sigma_k} \exp\left(-\frac{(x_1-\mu_1)^2/\sigma_1^2 + \cdots + (x_k-\mu_k)^2/\sigma_k^2}{2}\right) \\ &= \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right), \end{aligned}$$

which is the product of marginal univariate normal densities. This shows that if $\boldsymbol{X}$ is multivariate normal with uncorrelated elements, then they are mutually independent.

Theorem 1 *If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a k-dimensional multivariate normal vector, and $\boldsymbol{Y} = \mathbf{a} + \mathbf{B}\boldsymbol{X}$, where $\mathbf{B}$ is an $m \times k$ matrix with full row rank, then $\boldsymbol{Y} \sim N(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$ is a m-dimensional multivariate normal vector.*

Theorem 2 *Let $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_r)$. Then $Q = \boldsymbol{X}'\boldsymbol{X}$ is distributed chi-square with r degrees of freedom, written as χ_r^2.*

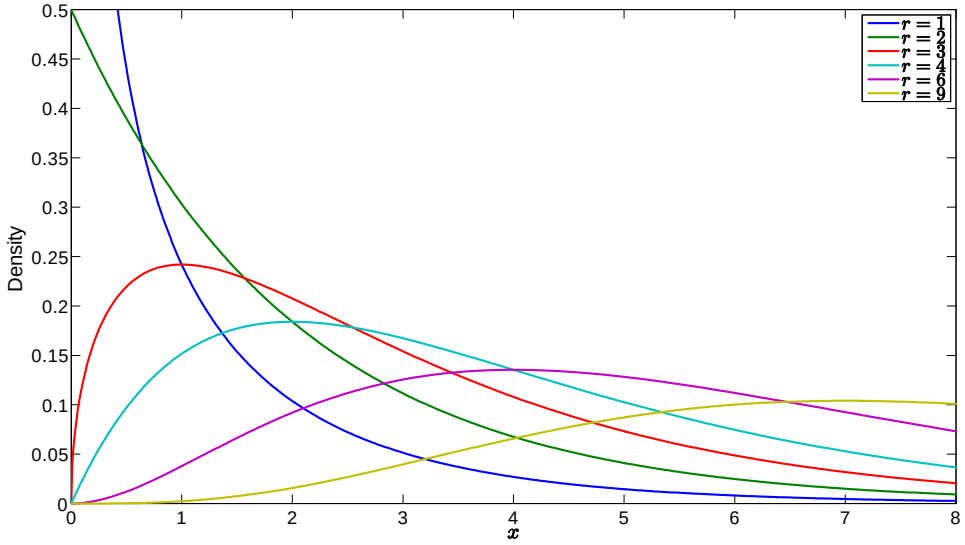

Figure 1: Chi-Square Density for $r = 1, 2, 3, 4, 6, 9$

Theorem 3 *If* $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{A})$ *with* $\mathbf{A} > 0$, $q \times q$, *then* $\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z} \sim \chi^2_q$.

Theorem 4 *Let* $Z \sim N(0,1)$ *and* $Q \sim \chi^2_r$ *be independent. Then* $T_r = Z/\sqrt{Q/r}$ *is distributed as student's* t *with* r *degrees of freedom.*

Remark 1 $T_\infty \sim N(0,1)$. $E[T_r] = 0$ *for* $r > 1$ *and* $Var(T_r) = r/(r-2)$ *for* $r > 2$.

Theorem 5 *Let* $P \sim \chi^2_q$ *and* $Q \sim \chi^2_r$ *be independent. Then* $F_{q,r} = \frac{P/q}{Q/r}$ *is distributed as Fisher's* F *distribution with* q *and* r *degrees of freedom.*

Remark 2 $qF_{q,\infty} \sim \chi^2_q$.

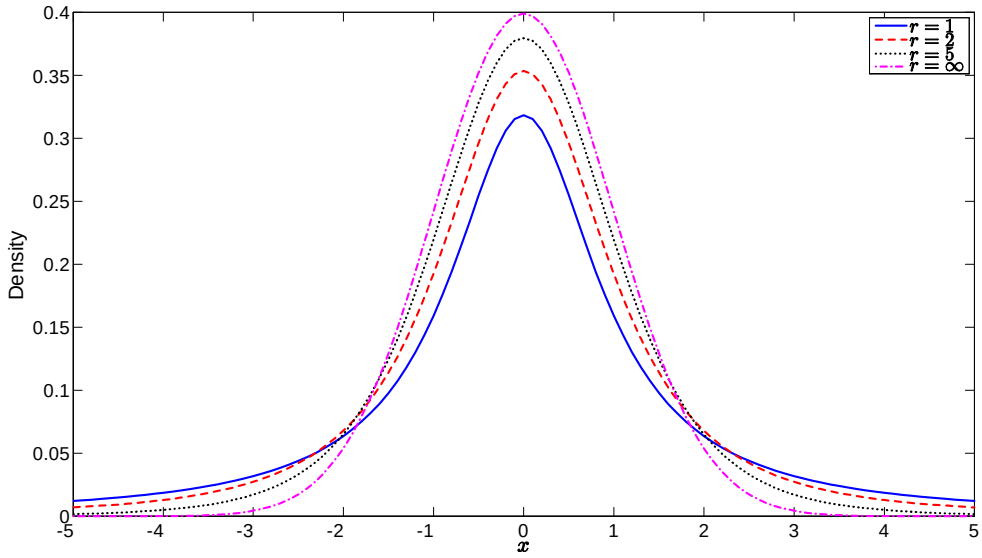

Figure 2: Density of the t Distribution for $r = 1, 2, 5, \infty$

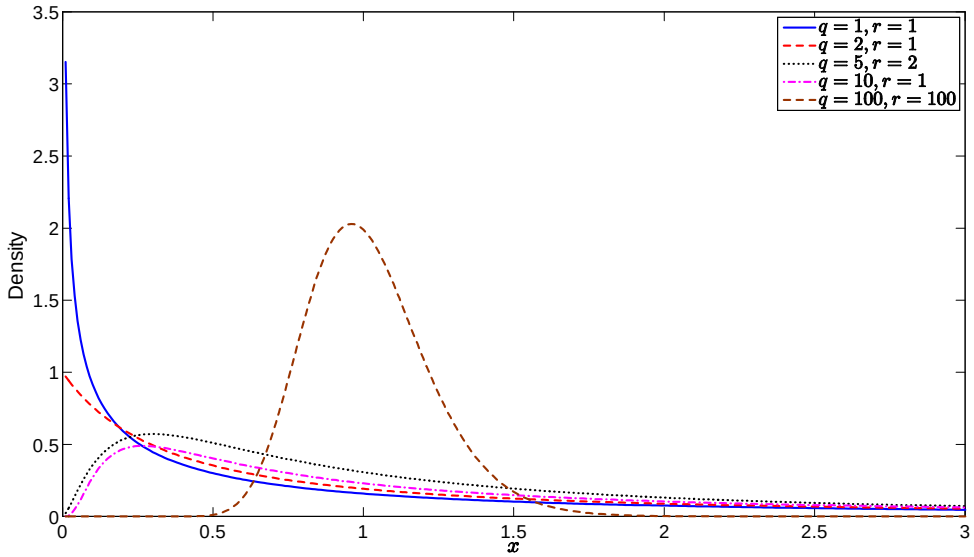

Figure 3: Density of the F Distribution for $(q, r) = (1, 1), (2, 1), (5, 2), (10, 1), (100, 100)$