Review of Probability: Part 2

Osman DOĞAN

Outline: Review of Probability - Part 2

- Review of Probability - Part 2:
  1. Normal distribution,
  2. Chi-squared distribution,
  3. t distribution
  4. F distribution
  5. Random sampling
  6. Sampling distribution of the sample average
  7. Large sample approximations to sampling distributions
  8. Scatterplots, the sample covariance and the sample correlation
- Readings:
  1. Stock and Watson (2020, Chapter 2 and Section 3.7 in Chapter 3).
  2. Hanck et al. (2021, Chapter 2).

Some common distributions

- In this course you will encounter four different probability distributions: normal, chi-squared, student t, and F.
- If the RV Y has a normal distribution with mean μ and variance σ^2, we will denoted this as $Y \sim N(\mu_Y, \sigma_Y^2)$.
- The pdf and the cdf of Y are denoted by $\phi_Y(y)$ and $\Phi_Y(y)$, respectively. These functions are

$$\phi_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2},$$

$$\Phi_Y(y) = P(Y \le y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \mathrm{d}y.$$

- If $Z \sim N(0,1)$, then we say that Z has a standard normal distribution.
- The normal cdf does not have a closed-form. We can use numerical methods to compute $\Phi_Z(z) = P(Z \le z)$.
- Values of $\Phi_Z(z)$ are tabulated in Appendix Table 1 (the Z-table) of our textbook for various z values.

Normal distribution

- Consider $N(0,1)$, $N(-2,1)$ and $N(0,9)$. The graphs of pdf and cdf of these distributions are shown in Figure 1.

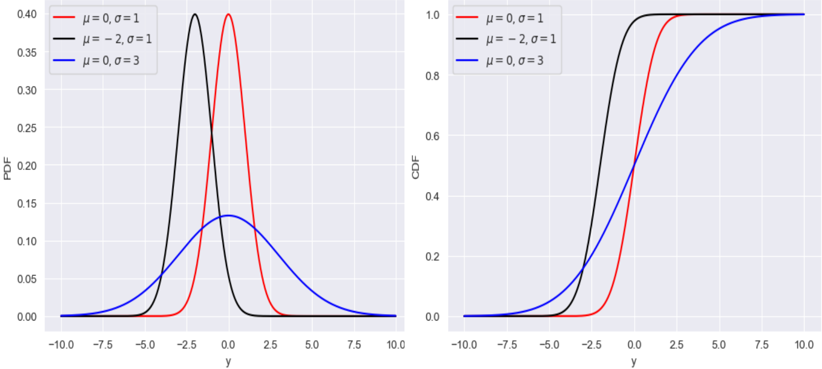


Figure 1: The pdf (left) and the cdf (right) of normal distributions

Normal distribution

Example 1

Assume $Y \sim N(1,4)$. Compute $P(Y \leq 2)$. We will standardize Y and then use the Z-table to compute $P(Y \leq 2)$.

$$P(Y \leq 2) = P\left(\frac{Y-\mu}{\sigma} \leq \frac{2-\mu}{\sigma}\right) = P(Z \leq \frac{2-1}{2}) = P(Z \leq 0.5)$$
$$= \Phi_Z(0.5) = 0.691.$$

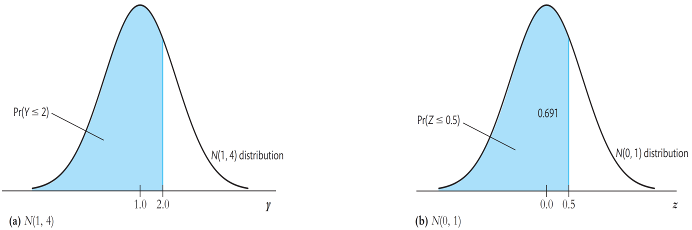


Figure 2: $P(Y \leq 2)$ and $P(Z \leq 0.5)$

Chi-square distribution

- The chi-squared distribution is used when testing certain types of hypotheses in econometrics.
- The chi-squared distribution that has m degrees of freedom, denoted by χ^2_m, is the distribution of the sum of m squared independent standard normal random variables.

Example 2

Let Z_1, Z_2 and Z_3 be independent standard normal random variables. Define $Y = Z_1^2 + Z_2^2 + Z_3^2$. Then, $Y \sim \chi^2_3$.

Normal distribution

- Consider χ^2_3, χ^2_4 and χ^2_{10}. The graphs of pdf and cdf of these distributions are shown in Figure 3.

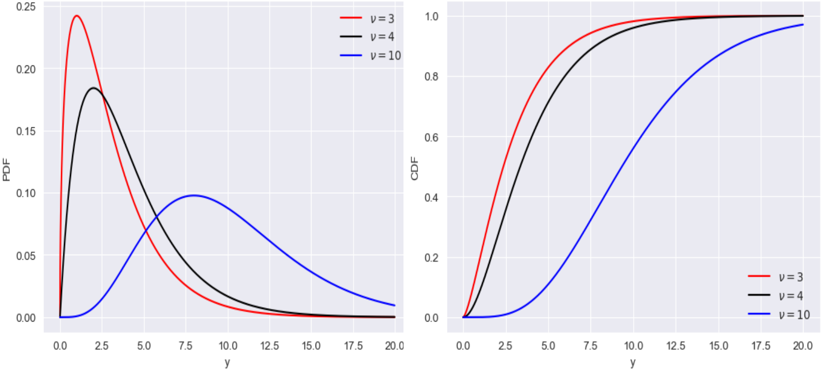


Figure 3: The pdf (left) and the cdf (right) of chi-squared distributions

Student t distribution

- The student t distribution with m degrees of freedom, denoted by t_m, is defined to be the distribution of the ratio of a standard normal random variable to the square root of an independently distributed chi-squared random variables with m degrees of freedom divided by m.

- Let $Z \sim N(0,1)$ and $W \sim \chi^2_m$. Assume that Z and W are independent. Then, we have
$$\frac{Z}{\sqrt{W/m}} \sim t_m.$$

- The t distribution has a bell shape similar to that of the normal distribution, but it has more mass in the tails; that is, it is a "fatter" bell shape than the normal.

- When m is 30 or more, the t distribution is well approximated by the standard normal distribution, and the t_∞ distribution equals the standard normal distribution.

Student t distribution

- In Figure 4, we compare the pdf and cdf of t_1 and t_{10} with that of $N(0,1)$.

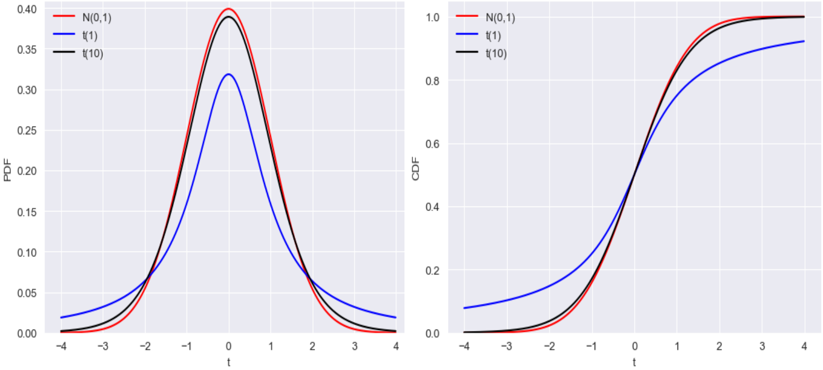


Figure 4: The pdf and cdf of t with the standard normal distribution

F distribution

- The F distribution with m and n degrees of freedom is denoted by $F_{m,n}$.
- Let $W \sim \chi^2_m$ and $V \sim \chi^2_n$. Assume that W and V are independent. Then, we have
$$\frac{W/m}{V/n} \sim F_{m,n}.$$
- In econometrics, an important special case of the F distribution arises when the denominator degrees of freedom is large enough that the $F_{m,n}$ can be approximated by $F_{m,\infty}$.
- $F_{m,\infty}$ distribution is the distribution of a chi-squared random variable with m degrees of freedom divided by m, i.e., $F_{m,\infty} = W/m$, where $W \sim \chi^2_m$.

F distribution

- Consider $F_{1,10}$, $F_{5,50}$ and $F_{10,100}$. The graphs of pdf and cdf of these distributions are shown in Figure 5.

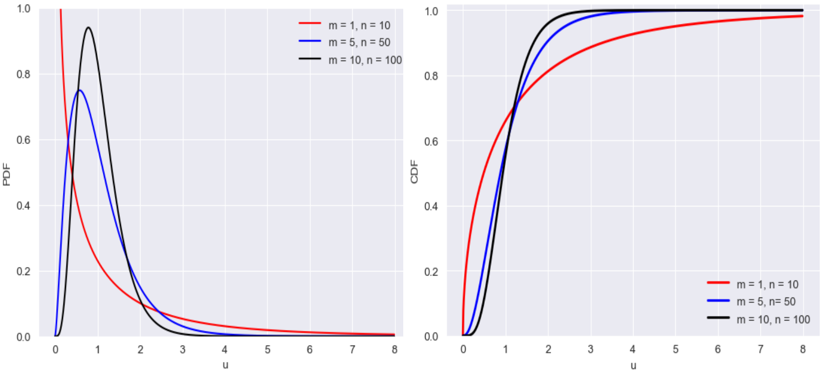


Figure 5: The pdf (left) and cdf (right) of F distribution

Computing probabilities

Example 3

Compute the following probabilities

(a) If Y is distributed t_{15}, find $P(Y \leq 1.75)$.

(b) If Y is distributed t_{90}, find $P(-1.99 \leq Y \leq 1.99)$.

(c) If Y is distributed $N(0,1)$, find $P(-1.99 \leq Y \leq 1.99)$.

(d) Why are the answers to (b) and (c) approximately the same?

(e) If Y is distributed $F_{7,4}$, find $P(Y \geq 4.12)$.

(f) If Y is distributed χ^2_1, find $P(Y \leq 1.02)$.

Computing probabilities

Listing 1: Solution of Example 3 with R

```
(a) pt(1.75,df=15,lower.tail = TRUE) = 0.9497299

(b) pt(1.99,df=90,lower.tail = T)-pt(-1.99,df=90,lower.tail = T) = 0.9503742

(c) pnorm(1.99,mean=0,sd=1,lower.tail=T)-pnorm(-1.99,mean=0, sd=1,lower.tail=T)=
    0.9534091

(e) pf(4.12, df1=7, df2=4, lower.tail = F) = 0.09471335

(f) pchisq(1.02,df=1,lower.tail = T)= 0.687481
```

Listing 2: Solution of Example 3 with Python

```
import scipy.stats as stats

stats.t.cdf(1.75, df=15) # (a)
stats.t.cdf(1.99, df=90) - stats.t.cdf(-1.99, df=90) # (b)
stats.norm.cdf(1.99) - stats.norm.cdf(-1.99) # (c)
1 - stats.f.cdf(4.12, dfn=7, dfd=4) # (e)
stats.chi2.cdf(1.02, df=1) #(f)
```

Random Sampling

- Almost all the statistical and econometric procedures used in this course involve averages or weighted averages of a sample of data.
- Because the sample is often drawn randomly from a larger population, the sample average itself becomes a random variable.
- It takes different values from one sample to the next if we draw many samples randomly from the population.
- Then, to characterize the sample average, we need to figure out its probability distribution, i.e., its sampling distribution.
- In this course, we will assume that all samples are drawn using simple random sampling.

Random Sampling

- In a simple random sample, n objects are selected at random from a population, and each object is equally likely to be drawn.
- The n observations in the sample are denoted $Y_1, Y_2, \ldots, Y_n$.
- Before the sample is drawn, we do not know this values, hence $Y_1, Y_2, \ldots, Y_n$ are random variables.
- Because of simple random sampling, Y_i's are independently distributed, i.e., Y_i provides no information about Y_j for $i \neq j$.
- Because Y_i's are randomly drawn from the same population, they are also identically distributed.
- Hence, due to simple random sampling, $Y_1, Y_2, \ldots, Y_n$ are independently and identically distributed (i.i.d.).

Random Sampling

- Since $Y_1, Y_2, \ldots, Y_n$ are randomly drawn using simple random sampling, the sample average $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is also random.
- The question is what can we say about the (sampling) distribution of $\bar{Y}$?
- We saw that $Y_1, Y_2, \ldots, Y_n$ i.i.d. due to simple random sampling. Assume that $Y_i \sim (\mu_Y, \sigma_Y^2)$. Note that μ_Y and σ_Y^2 are unknown constants.
- Using the fact that $\mathrm{cov}(Y_i, Y_j) = 0$ for all $i \neq j$, we have

$$\mathrm{E}(\bar{Y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu_Y = \mu_Y,$$

$$\mathrm{var}(\bar{Y}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\left(Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 = \frac{\sigma_Y^2}{n}.$$

1. $\bar{Y}$ is an unbiased estimator of μ_Y, e.i., $\mathrm{E}(\bar{Y}) = \mu_Y$.
2. $\mathrm{var}(\bar{Y})$ is inversely proportional to n.
3. The spread of the sampling distribution is proportional to $1/\sqrt{n}$

Random Sampling

- Notice that we only assumed the mean and the variance of Y exist and we are able characterize the mean and the variance the (sampling) distribution of $\bar{Y}$.
- If you are willing to make stronger assumptions, say $Y_i \sim N(\mu_Y, \sigma_Y^2)$, then you know exactly the distribution of $\bar{Y}$, $N(\mu_Y, \sigma_Y^2/n)$.
- Unfortunately, if the distribution of Y is not normal, then in general the exact sampling distribution of $\bar{Y}$ is very complicated and depends on the distribution of Y.
- We will instead try to approximate the sampling distribution of $\bar{Y}$.
- The approximate approach uses approximations to the sampling distribution that rely on the sample size being large.
- The large-sample approximation to the sampling distribution is often called the asymptotic distribution-"asymptotic" because the approximations become exact in the limit that $n \to \infty$.

Law of large numbers

Definition 1 (Convergence in probability)

The sample average $\bar{Y}$ converges in probability to μ_Y (or, equivalently, $\bar{Y}$ is a consistent estimator of μ_Y) if the probability that $\bar{Y}$ is in the range $(\mu_Y - c)$ to $(\mu_Y + c)$ becomes arbitrarily close to 1 as n increases for any constant $c > 0$. The convergence of $\bar{Y}$ to μ_Y in probability is written $\bar{Y} \xrightarrow{p} \mu_Y$.

Theorem 1 (Law of large numbers)

If $Y_1, \ldots, Y_n$ are independently and identically distributed with $\mathrm{E}(Y_i) = \mu_Y$ and $\mathrm{var}(Y_i) = \sigma_Y^2 < \infty$, then $\bar{Y} \xrightarrow{p} \mu_Y$.

Central limit theorem

- The central limit theorem (CLT) says that, under general conditions, the distribution of $\bar{Y}$ is well approximated by a normal distribution when n is large.

Theorem 2 (Central limit theorem)

If $Y_1, \ldots, Y_n$ are independently and identically distributed with $\mathrm{E}(Y_i) = \mu_Y$ and $\mathrm{var}(Y_i) = \sigma_Y^2 < \infty$, then $\frac{\bar{Y}-\mu_Y}{\sqrt{\mathrm{var}(\bar{Y})}} = \frac{\bar{Y}-\mu_Y}{\sigma_Y/\sqrt{n}}$ is well approximated by $N(0,1)$.

- Note that the CLT implies that $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n)$ when n is large.
- Figure 6 shows the sampling distributions of the sample average of n Bernoulli random variables.
- It is easy to see that, if n is large enough, the distribution of $\bar{Y}$ is well approximated by a normal distribution.

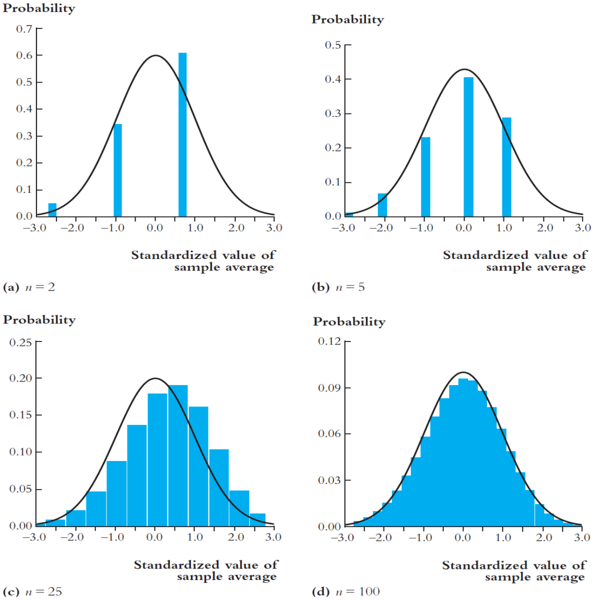


Figure 6: The sampling distributions of the sample average of n Bernoulli random variables

Summary: Sampling distribution of $\bar{Y}$

- Assume that $Y_1, \ldots, Y_n$ are independently and identically distributed with $\mathrm{E}(Y_i) = \mu_Y$ and $\mathrm{var}(Y_i) = \sigma_Y^2 < \infty$. Then,

  1. The exact sampling distribution of $\bar{Y}$ has mean μ_Y and variance σ_Y^2/n.
  2. Other than its mean and variance, the exact distribution of $\bar{Y}$ is complicated and depends on the distribution of Y (the population distribution).
  3. When n is large the sampling distribution simplifies:
     1. $\bar{Y} \xrightarrow{p} \mu_Y$ (law of large numbers),
     2. $\frac{\bar{Y}-\mu_Y}{\sqrt{\mathrm{var}(\bar{Y})}} = \frac{\bar{Y}-\mu_Y}{\sigma_Y/\sqrt{n}}$ is approximately $N(0,1)$ (CLT).

Example 4

$Y_1, \ldots, Y_n$ are i.i.d. Bernoulli random variables with $p = 0.4$. Let $\bar{Y}$ denote the sample mean.

(a) Use the central limit to compute approximations for (i) $P(\bar{Y} \geq 0.43)$ when $n = 100$, and (ii) $P(\bar{Y} \leq 0.37)$ when $n = 400$.

(b) How large would n need to be to ensure that $P(0.39 \leq \bar{Y} \leq 0.41) = 0.95$?

Solution to Example 4.

Note that if $Y_i \sim$ Bernoulli$(p = 0.4)$, then $\mathrm{E}(Y_i) = 1 \times p + 0 \times (1-p) = 0.4$ and Var$(Y_i) = p(1-p) = 0.24$. Recall that $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n)$ when n is large by the CLT.

(a) When $n = 100$, we have $\bar{Y} \sim N(0.4, 0.24/100)$. Then,

$$P(\bar{Y} \geq 0.43) = \texttt{pnorm}(0.43, \texttt{mean} = 0.4, \texttt{sd} = \texttt{sqrt}(0.0024), \texttt{lower.tail} = \texttt{F}) = 0.270.$$

When $n = 400$, we have $\bar{Y} \sim N(0.4, 0.24/400)$. Then,

$$P(\bar{Y} \leq 0.37) = \texttt{pnorm}(0.37, \texttt{mean} = 0.4, \texttt{sd} = \texttt{sqrt}(0.24/400), \texttt{lower.tail} = \texttt{T}) = 0.110.$$

(b) Note that

$$P(0.39 \leq \bar{Y} \leq 0.41) = P\left(\frac{0.39 - 0.4}{\sqrt{0.24/n}} \leq \frac{\bar{Y} - 0.4}{\sqrt{0.24/n}} \leq \frac{0.41 - 0.4}{\sqrt{0.24/n}}\right)$$
$$= P\left(\frac{-0.01}{\sqrt{0.24/n}} \leq Z \leq \frac{0.01}{\sqrt{0.24/n}}\right)$$

Since $P(-1.96 \leq Z \leq 1.96) = 0.96$, we must have $\frac{-0.01}{\sqrt{0.24/n}} \leq -1.96$ and $\frac{0.01}{\sqrt{0.24/n}} \geq 1.96$. These inequalities suggest that $n \geq 9220$.

$\square$

Example 5

In any year, the weather can inflict storm damage to a home. From year to year, the damage is random. Let Y denote the dollar value of damage in any given year. Suppose that in 95% of the years $Y = 0$ dollars but in 5% of the years $Y = 20.000$ dollars.

(a) What are the mean and standard deviation of the damage in any year?

(b) Consider an "insurance pool" of 100 people whose homes are sufficiently dispersed so that, in any year, the damage to different homes can be viewed as independently distributed random variables. Let Y denote the average damage to these 100 homes in a year. (i) What is the expected value of the average damage? (ii) What is the probability that average damage exceeds 2000 dollars?

Solution to Example 5.

(a) Since $P(Y=0)=0.95$ and $P(Y=20.000)=0.05$, we have

$$\begin{aligned}
\mu_Y &= 0\times P(Y=0)+20.000\times P(Y=20.000)=1000,\\
\sigma_Y^2 &= (0-1000)^2\times P(Y=0)+(20.000-1000)^2\times P(Y=20.000)\\
&= (-1000)^2\times 0.95+19000^2\times 0.05=1.9\times 10^7.
\end{aligned}$$

(b) Note that $\mu_{\bar{Y}}=\mu_Y=1000$ and $\sigma_{\bar{Y}}^2=\sigma_Y^2/n=1.9\times 10^5$. Then, by the CLT, we have $\bar{Y}\sim N(1000, 1.9\times 10^5)$. Thus,

$$P(\bar{Y}>2000)=\texttt{pnorm}(2000,\texttt{mean}=1000,\texttt{sd}=\texttt{sqrt}(1.9*10^5),\texttt{lower.tail}=\texttt{F})=0.0109.$$

$\square$

Scatterplots, sample covariance and sample correlation

- Three ways to summarize the relationship between X and Y are: the scatterplot, the sample covariance, and the sample correlation coefficient.
- The relationship between average district income and average test scores in 420 CA school districts.

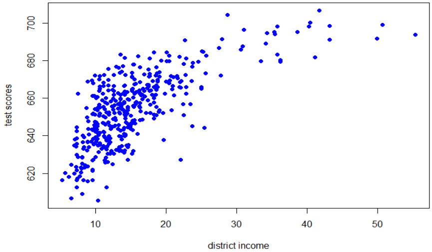


Figure 7: Scatter plot of average test score and average district income

- What sign would you expect for the correlation between income and test scores?

Scatterplots, sample covariance and sample correlation

- The covariance and correlation are derived from the joint probability distribution of the random variables X and Y.
- Because the population distribution is unknown, in practice we do not know the population covariance or correlation.
- The population covariance and correlation can, however, be estimated by taking a random sample of n members of the population and collecting the data (X_i, Y_i), $i = 1, 2, \ldots, n$.
- The sample covariance, denoted s_{XY}, is

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

- The sample correlation coefficient, or sample correlation, denoted r_{XY}, is

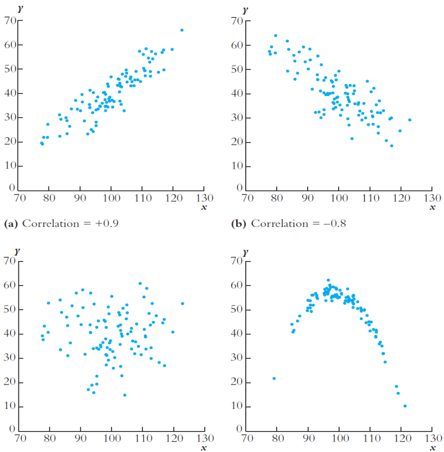
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

- These statistics are also consistent, i.e., $s_{XY} \xrightarrow{p} \sigma_{XY}$ and $r_{XY} \xrightarrow{p} \text{corr}(X_i, Y_i)$.

Scatterplots, sample covariance and sample correlation

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets

The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between *X* and *Y*. In Figure 3.3c, *X* is independent of *Y* and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.



(a) Correlation = +0.9

(b) Correlation = −0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

Bibliography I

Hanck, Christoph et al. (2021). *Introduction to Econometrics with R*. URL: https://www.econometrics-with-r.org/index.html.

Stock, James H. and Mark W. Watson (2020). *Introduction to Econometrics*. Fourth. Pearson.