START CLASSNOTES: Mon 9/20/21

HANDOUT #5: PARAMETRIC SUMMARIES OF POPULATIONS AND PROCESSES

1. Summaries of Center/Location in a Distribution
   (a) Population/Process Mean (μ)
   (b) Population/Process Median ($\tilde{\mu}$)
   (c) Population/Process Quartiles ($Q(.25), \tilde{\mu}, Q(.75)$)
   (d) Population/Process Trimmed Mean ($\mu_{(\alpha)}$)
2. Summaries of Level of Dispersion/Variability in a Distribution
   (a) Population/Process Range (R)
   (b) Population/Process Semi-interquartile Range (SIQR)
   (c) Population/Process Standard Deviation (σ)
   (d) Population/Process Median Absolute Deviation (MAD)
3. Summaries of Shape/Tail Weight in a Distribution
   (a) Population/Process Skewness (β_1)
   (b) Population/Process Kurtosis (β_2)
4. Do Mean and Standard Deviation Summarize Distribution?
5. How Are Mean and Median Related?
6. Correlation and AutoCorrelation ($\rho_{y,x}, \rho_k$)

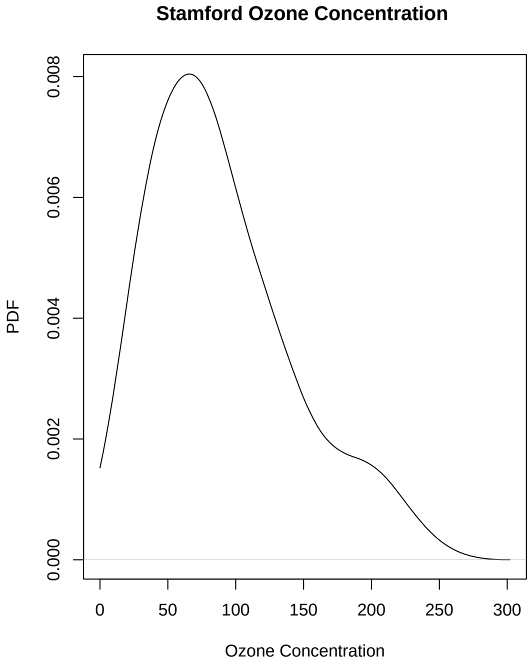
Supplemental Reading:

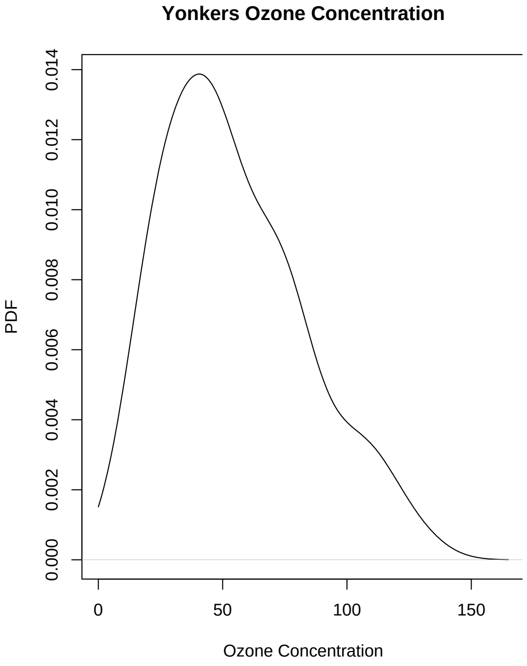
- Chapter 2 and Sections 4.1, 4.2, and 4.3.4 in the Tamhane/Dunlop book

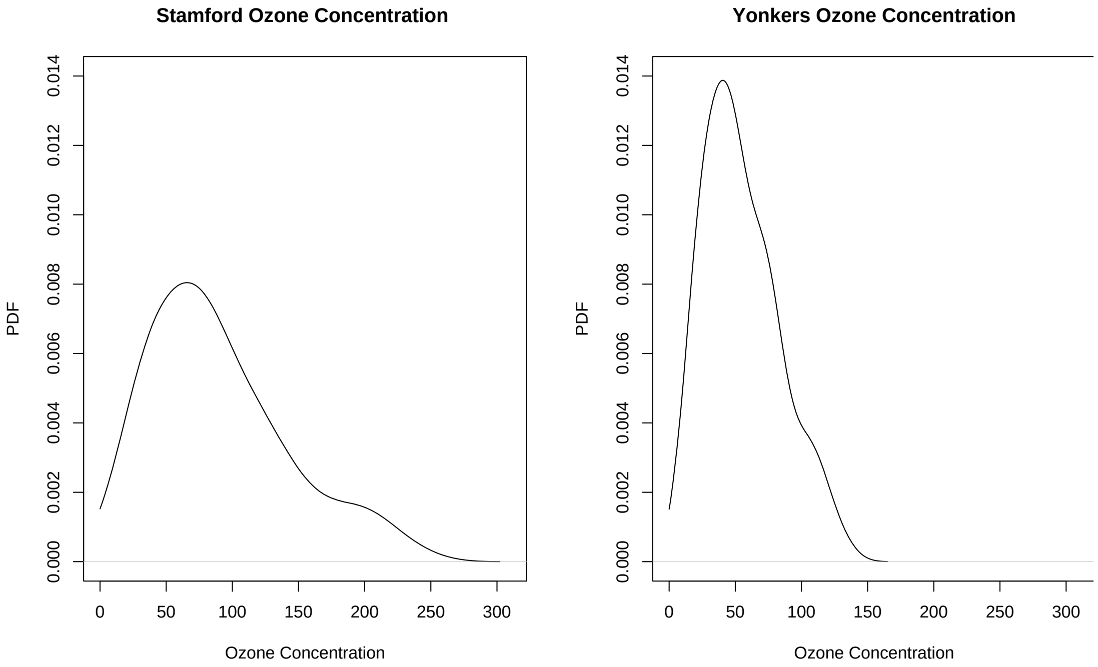
Comparing Several Population Distributions

A process or population is completely described by its cdf or quantile function or pdf. However, it is often very difficult to compare several populations or processes using their pdfs or cdfs. Also, its difficult to precisely quantify the differences in several pdfs or cdfs. The distribution of maximum ozone concentration in two cities, Stamford, Connecticut and Yonkers, New York over the summer months are given on the next two pages. In the first set of graphs, we plotted the pdfs with the default settings for the scale on the Y-axis and X-axis. From these two plots the distribution of maximum ozone would appear to be nearly the same in the two cities. When we set the range of the scales to be the same for the two pdfs, a very large difference in the two distributions is apparent. However, if we were EPA regulators or state environmental monitors, how can we describe the distribution of maximum ozone concentration or state precisely the difference in the distributions for the two cities? Furthermore, if we wanted to quantify changes in ozone concentrations in the two cities after new environmental controls on the emissions from automobiles or industrial plants were implemented, would it be possible to make precise statements about any such changes that the public or congress would understand using just the plots of the distributions?

Thus, we will define several parameters which will summarize the information contained in the cdf or pdf concerning the distribution of values in a population or process. These numerical summaries will not completely describe the differences in several populations or processes but may provide an adequate description of changes or differences in many situations.







Parametric Summarization of a Population Distribution

Definition A *parameter* is a numerical characteristic of a population or process.
It may be a functional of the cdf or a constant contained within the formula for the cdf.

Example 1 The location, scale, or shape parameters in a cdf:

The three parameter Weibull distribution has a location parameter θ, a scale parameter β and shape parameter γ:

$$F(y;\theta,\beta,\gamma) = \begin{cases} 0 & \text{if } y < \theta \\ 1 - e^{-((y-\theta)/\beta)^{\gamma}} & \text{if } y \geq \theta \end{cases}$$

$$\text{with } -\infty < \theta < \infty, \quad \beta > 0, \quad \gamma > 0.$$

The three quantities θ, β, and γ are parameters.

Example 2 The median or IQR of the expenses per household in Louisiana resulting from the lack of a rapid local/state/federal response to problems caused by Katrina.
Both the median and IQR would be parameters associated with the distribution of household damage expenses.

Example 3 The mean and standard deviation of the mileage (MPG) of a new hybrid automobile.
The mean and standard deviation of MPG would be used to compare MPG of hybrids to conventional automobiles.

We will now define a variety of parameters which will attempt to describe different aspects of a distribution. Many of these parameters will be associated with a functional of the cdf called the *expected value* of a function of the r.v Y, **E[h(Y)]**:

Definition: Expected Value of h(Y): The expected value of a function, h(Y) of a r.v. Y having cdf F_Y and pdf f_Y is defined as

For a continuous strictly increasing cdf:

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y)f(y)dy$$

For discrete cdf:

$$\mu = E[h(Y)] = \sum_{i=1}^{\infty} h(y_i)f(y_i)$$

From the above formulas, the expected value of a random variable is a weighted average of the values of the random variables with weights being the pdf of the random variable.

EXAMPLE Let Y be the number of defective parts in a randomly selected container of 10 parts. The pmf for Y is given in the following table. The cost per container of repairing defective parts is given by $C = h(Y) = 5Y + 3$. The average cost of repairs is then obtained by weighting the values of $h(Y)$ by the corresponding values of the pmf and then summing these values.

y	0	1	2	3	4	5	6	7	8	9	10	
c	3	8	13	18	23	28	33	38	43	48	53	Total
$f(y)$	.68	.18	.05	.03	.015	.014	.012	.008	.007	.003	.001	1.000
$y * f(y)$	0	0.180	0.100	0.090	0.060	0.070	0.072	0.056	0.056	0.027	0.010	0.721
$c * f(y)$	2.040	1.440	0.650	0.540	0.345	0.392	0.396	0.304	0.301	0.144	0.053	6.605

From the above table we have that the average number of defects per container is $E(Y) = .721$ and the average cost of repairing the parts in a container is $E(C) = 6.605$.

As is often true for random variables taking on only integer values, $E(Y)$ is not one of the possible values of Y and $E(C)$ is not one of the possible values for C.

Note if we just took the average number of defects per container, $E(Y)$, .721 and used it in the cost formula we would obtain $E(C) = 5(.721)+3 = 6.605$.

Optional Material: Riemann vs Riemann-Stieltjes Integral

Many parameters are defined by specifying the function h(y):

Mean value of Y, $\mu = E[Y]$, has $h(y) = y$,

Variance of Y, $\sigma^2 = E[(Y - \mu)^2]$, has $h(y) = (y - \mu)^2$.

It is somewhat cumbersome having two separate definitions for $E[h(Y)]$ depending on whether the distribution of Y is discrete or continuous. The two definitions can be consolidated through the use of a generalization of the Riemann Integral, the Riemann-Stieltjes Integral:

Let g be a bounded function on $[a, b)$.

Let $P_n = a = x_{0,n} < x_{1,n} < \cdots < x_{n-1,n} < x_{n,n} = b$ be a partition of $[a, b)$ with mesh size,

$$||P_n|| = \max_{1 \le i \le n} |x_{i,n} - x_{i-1,n}|.$$

$||P_n||$ is the maximum gap in P_n.

Riemann Integral: Let P_n be a sequence of partitions of $[a, b)$ for which each term in the sequence is a refinement of its predecessor and for which $\lim_{n\to\infty} ||P_n|| = 0$. Let $x^*_{i,n} \epsilon [x_{i-1,n}, x_{i,n})$ and $R(P_n) = \sum_{i=1}^{n} g(x^*_{i,n})(x_{i,n} - x_{i-1,n})$ be an approximating sum for the area under the curve $g(\cdot)$ between $[a, b)$. If $\lim_{n\to\infty} R(P_n) = R$ for all such sequences of partitions then $R = \int_a^b g(x)dx$ is the Riemann Integral of g over $[a, b)$.

Generalization of Riemann Integral is Riemann-Stieltjes Integral of g with respect to F:

generalization of Riemann integral

Riemann-Stieltjes Integral of g wrt F: Let F be a cdf and g be a continuous, real valued function. Let P_n be a sequence of partitions of $[a, b)$ for which each term in the sequence is a refinement of its predecessor and for which $\lim_{n\to\infty} ||P_n|| = 0$. Let $x^*_{i,n} \epsilon [x_{i-1,n}, x_{i,n})$ and

$$RS(P_n) = \sum_{i=1}^{n} g(x^*_{i,n}) \left[F(x_{i,n}) - F(x_{i-1,n})\right]$$

be an approximating sum. If $\lim_{n\to\infty} RS(P_n) = RS$ for all such sequences of partitions then $RS = \int_a^b g(x)dF(x)$ is the Riemann-Stieltjes Integral of g w.r.t. F over $[a, b)$.

We will not need the full power of the Riemann-Stieltjes integral but a few examples will illustrate its usefulness in defining population parameters and later in defining sample estimators of these parameters.

Example 1 Let F be a continuous strictly increasing cdf with pdf f. Then

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

Example 2 Let F be a cdf of a discrete r.v. Y with pmf f and jumps in F at $y_1, y_2, \ldots$ of size $p_i = f(y_i)$. Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)dF(x) = \sum_{i=1}^{\infty} g(y_i)f(y_i) = \sum_{i=1}^{\infty} p_i g(y_i)$$

Example 3 Let F be a cdf with jumps in F at $Y_1, Y_2, \ldots, Y_n$ of size $1/n$. That is, F is the empirical cdf. Then

$$\int_{-\infty}^{\infty} g(y)dF(y) = \sum_{i=1}^{n} \frac{1}{n} g(Y_i) = \frac{1}{n} \sum_{i=1}^{n} g(Y_i)$$

In particular, if $g(y) = y$ then

$$\int_{-\infty}^{\infty} g(y)dF(y) = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$$

Note that F is the empirical distribution function (edf) for the n data values $Y_1, Y_2, \ldots, Y_n$

Parametric Summaries of the Center and Dispersion of a Distribution

We will define several parameters associated with a distribution which describe the “average” value in the distribution or the amount of “dispersion” in the distribution. That is, how spread out the values are about the ”center” of the distribution. Let Y be a r.v. which represents a random selected value from a population or process. Suppose Y has cdf F and pdf(pmf) f.

Definition: The kth Moment of Y The kth moment of Y is

$$\begin{aligned} m_k = E[Y^k] &= \int_{-\infty}^{\infty} y^k dF(y) \\ &= \int_{-\infty}^{\infty} y^k f(y) dy \quad \text{if F is continuous} \\ &= \sum_{i=1}^{\infty} y_i^k f(y_i) \quad \text{if F is discrete} \end{aligned}$$

We usually denote m_1 as μ.

Definition: The kth Central Moment of Y about the center μ

The kth central moment of Y is

$$\begin{aligned} \mu_k = E[(Y-\mu)^k] &= \int_{-\infty}^{\infty} (y-\mu)^k dF(y) \\ &= \int_{-\infty}^{\infty} (y-\mu)^k f(y) dy \quad \text{if F is continuous} \\ &= \sum_{i=1}^{\infty} (y_i-\mu)^k f(y_i) \quad \text{if F is discrete} \end{aligned}$$

where $\mu = m_1$ is called the mean or expected value of the r.v. Y.

Note: $\mu_1 = 0$

Mean of a Population or Distribution

Several key parameters used in describing distributions are either central moments or are functions of several central moments. We will now define two of these parameters, the mean and standard deviation. These are the two most widely used parameters (correctly used or not) in describing the center and dispersion in distributions.

Definition: The expected value or mean value of a r.v. Y or of its distribution F is defined as

$$\mu = E[Y] = m_1$$

For a r.v. Y having pdf $f(y)$,

$$\mu = E[Y] = \int_{-\infty}^{\infty} y f(y) dy$$

Example Suppose Y has pdf $f(y) = \frac{1}{\beta} e^{-y/\beta}$ for $y \geq 0$ and 0 otherwise.

$$E[Y] = \int_{-\infty}^{\infty} y f(y) dy = \int_{-\infty}^{0} y(0) dy + \int_{0}^{\infty} y \frac{1}{\beta} e^{-y/\beta} dy = 0 - \beta e^{-y/\beta}|_0^\infty = \beta$$

This is a "weighted" average value of the occurrence of the values of Y in the population or process where the weights are the values of the pdf. The idea of weighting is more clearly observed in the case of a discrete distributions.

For example, suppose Y has k possible values $y_1, y_2, \cdots, y_k$ and pmf f, where $f(y_i) = Pr[Y = y_i]$.

Then, μ is simply the weighted average value of the $y_i s$, with weights $f(y_i)$:

$$\mu = E[Y] = \sum_{i=1}^{k} y_i f(y_i).$$

In particular, for a population containing N units with distinct values of Y: $y_1, y_2, \cdots, y_N$ occuring with frequencies: $f_i = 1/N$, we have

$$\mu = \sum_{i=1}^{N} y_i f_i = \sum_{i=1}^{N} y_i 1/N = \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{Y}$$

Thus μ is our common notation for the mean of a population.

For some continuous, heavy-tailed distributions μ does not exist because the integral is indeterminate. For the Cauchy distribution,

$$\mu = E[Y] = \int_{-\infty}^{\infty} y f(y) dy = \int_{-\infty}^{\infty} y \left(\pi\theta_2 \left[1 + \left(\frac{y - \theta_1}{\theta_2} \right)^2 \right] \right)^{-1} = \infty - \infty \neq 0$$

Standard Deviation of a Population or Distribution

The most widely used measure of how spread out the values of Y are about the measure of the center μ is the standard deviation of Y:

Definition: The Variance of a r.v. Y or of its distribution F is defined as

$$\sigma^2 = Var(Y) = E\left[(Y-\mu)^2\right] = \int_{-\infty}^{\infty} (y-\mu)^2 dF(y)$$

σ^2 is the weighted average squared distance of the values of Y about the mean μ. Its units are the square of the units of Y.

When Y has a pdf $f(y)$, $\sigma^2 = \int_{-\infty}^{\infty}(y-\mu)^2 f(y)dy$

A little algebra demonstrates that $\sigma^2 = E\left[(Y-\mu)^2\right] = E[Y^2] - (\mu)^2 = m_2 - m_1^2 = \mu_2$

Example Suppose Y has pdf $f(y) = \frac{1}{\beta}e^{-y/\beta}$ for $y \geq 0$ and 0 otherwise.

$\mu = E[Y] = \beta$ and

$$E[Y^2] = \int_0^{\infty} y^2 \frac{1}{\beta} e^{-y/\beta} dy = 2\beta^2 \;\Rightarrow\; \sigma^2 = E[Y^2] - \mu^2 = 2\beta^2 - (\beta)^2 = \beta^2$$

For any continuous distributions in which the mean does not exist, then the variance would also not exist.

For example, the Cauchy distribution.

For the t-distribution with df $= 2$, $E[Y] = 0$ but $E[Y^2] = \infty$.

For the t-distribution with df $= 1$, the mean and variance do not exist (t with df=1 is a Cauchy distribution).

In particular, for a population containing N units with distinct values of Y: $y_1, y_2, \cdots, y_N$ occurring with frequencies: $f_i = 1/N$, we have

$$\sigma^2 \;=\; \sum_{i=1}^{N}(y_i-\mu)^2 f_i \;=\; \sum_{i=1}^{N}(y_i-\mu)^2 1/N \;=\; \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$$

Note that when we are describing the population variance, the divisor is N and not $N-1$, as would be in the sample standard deviation when μ is unknown.

Definition: The Standard Deviation of a r.v. Y or of its distribution F is defined as

$$\sigma = \sqrt{Var(Y)} = \sqrt{E\left[(Y-\mu)^2\right]}$$

The standard deviation has the same units as Y and hence is more easily interpreted as a measure of the dispersion of the population values about the mean than is the variance which has units the square of the units of the r.v.

Example 1 For the example on page 4 concerning Y, the number of defective parts in a randomly selected container of 10 parts. We have the pmf for Y is given in the following table:

y	0	1	2	3	4	5	6	7	8	9	10	
$f(y)$	.68	.18	.05	.03	.015	.014	.012	.008	.007	.003	.001	1.000

The expected value of Y is given by

$$\mu = E[Y] = \sum_{y=0}^{10} yf(y) \Rightarrow$$

$$\mu = 0(.68) + 1(.18) + 2(.05) + 3(.03) + 4(.015) + 5(.014) + 6(012) + 7(.008) + 8(.007) + 9(.003) + 10(.001) = 0.721$$

That is, the average number of defectives per container is 0.721.

A simplified formula for computing variance is given by

$$\underbrace{\sigma^2}_{} = E[(Y-\mu)^2] = E[Y^2 - 2Y\mu + \mu^2] = E[Y^2] - 2\mu E[Y] + \mu^2] = \underbrace{E[Y^2] - \mu^2}_{} \qquad \ast$$

For our defectives example, we have

$$E[Y^2] = 0(.68)+1(.18)+4(.05)+9(.03)+16(.015)+25(.014)+36(012)+49(.008)+64(.007)+81(.003)+100(.001) = 2.855$$

Therefore, the standard deviation has value,

$$\sigma = \sqrt{Var(Y)} = \sqrt{E[Y^2] - \mu^2} = \sqrt{2.855 - (.721)^2} = 1.528$$

Example 2 Let X have an exponential pdf: $f(x) = \frac{1}{\beta}e^{-x/\beta}$ for $x > 0$

$$\begin{aligned}
\mu &= E[X] = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x = \int_0^{\infty} x\frac{1}{\beta}e^{-x/\beta}\mathrm{d}x = -xe^{-x/\beta}|_0^{\infty} - \beta e^{-x/\beta}|_0^{\infty} = \beta \\
E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x = \int_0^{\infty} x^2\frac{1}{\beta}e^{-x/\beta}\mathrm{d}x = 2\beta^2 \\
\sigma^2 &= Var(X) = E[X^2] - (\mu)^2 = 2\beta^2 - (\beta)^2 = \beta^2 \;\Rightarrow \\
\sigma &= \beta
\end{aligned}$$

Note: For the exponential distribution, $\sigma = \beta = \mu$

How completely do the pair (μ, σ) describe a family of distributions?

One answer to the above question is to examine the proportion of the population values falling within $k\sigma$ units of μ, $p(k)$. That is, determine whether or not $p(k)$ varies depending on

1. the values of (μ, σ) or
2. the family of distributions or
3. is it a constant for all distributions?

We will answer the above questions by examining:

$$p(k) = P[Y \text{ is within } k\sigma \text{ units of } \mu] \Rightarrow$$

$$p(k) = P[|Y - \mu| \leq k\sigma] = P[\mu - k\sigma \leq Y \leq \mu + k\sigma] = F[\mu + k\sigma] - F[\mu - k\sigma]$$

General Bound: Chebyshev's Inequality If Y is a r.v. with mean μ, $0 < \sigma < \infty$ and $k > 1$ then

$$p(k) = P[|Y - \mu| \leq k\sigma] \geq 1 - \frac{1}{k^2}$$

How sharp is this bound? That is, how close is the true probability of this event to the value given by the bound, $1 - \frac{1}{k^2}$?

Case 1: Location-Scale Family Suppose the mean and standard deviation (μ, σ) of the r.v. Y are also location-scale parameters in a family of distributions. Then, with $Z = (Y - \mu)/\sigma$,

$$P[|Y - \mu| \leq k\sigma] = P[|Y - \mu|/\sigma \leq k] = P[|Z| \leq k] = F_Z[k] - F_Z[-k]$$

The proportion of the distribution within k standard deviations of the mean is the same value for every member of the family. It does not depend on the values of (μ, σ). However, the proportion does vary from family to family because the cdf of Z, F_Z would be different for each family of distributions, that is, the proportion for the normal family of distributions is different from the value for the Cauchy or Double Exponential family of distributions.

Case 2: General Family of Distributions In general, the exact proportion will vary greatly within even the same family and will depend on the values of (μ, σ) when (μ, σ) are not location-scale parameters for the population.

We will illustrate these ideas with a few examples.

The following table contains the proportion of the distribution within k standard deviations of the population mean for various values of k and distributions:

$$p(k) = P[|Y - \mu| \leq k\sigma] = F_Y(\mu + k\sigma) - F_Y(\mu - k\sigma)$$

By Chebyshev inequality, $p(k) \geq 1 - \frac{1}{k^2}$

	k						
Distribution	.5	1	1.25	1.5	2	2.5	3
Bound($1 - \frac{1}{k^2}$)	-3.0	0	.36	.556	.750	.840	.889
Normal	.383	.683	.789	.866	.955	.988	.9973
Cauchy	.295	.500	.570	.626	.705	.758	.7950
ChiSq(df=1)	.3970	.880	.904	.923	.950	.967	.978
ChiSq(df=5)	.3820	.724	.848	.916	.955	.978	.987
ChiSq(df=10)	.3823	.701	.815	.893	.959	.980	.991
ChiSq(df=50)	.3828	.686	.794	.871	.956	.986	.9954

From the above table we note the following:

1. If $k \leq 1$, the Chebyshev bound is not useful, $1 - \frac{1}{k^2} \leq 0$.

2. The Chebyshev bound is not very close to the true proportion for the normal distribution or the Chi-square distribution. Not surprising because the bound must be true for all distributions having a finite standard deviation.

3. The Cauchy distribution has proportions which are in fact smaller than the values specified by the Chebyshev bound. Does this fact demonstrate that the Chebyshev bound is not valid?

4. The proportion of the chi-square distribution that is within 1.0 standard deviations of its mean varies widely from 0.880 to 0.686 as the df increase from 1 to 50.

df	1	5	10	50	75	100	3000
$\mu = df$	1	5	10	50	75	100	3000
$\sigma = \sqrt{2df}$	1.414	3.162	4.472	10	12.2474	14.1421	77.4597
$P[\|Y - \mu\| \leq \sigma]$	.8798	.7236	.7007	.6860	.6849	.6843	.6827

5. The proportion of the chi-square distribution that is within 1 standard deviations of its mean is close to the proportion for the normal distribution, .6827, when $df \geq 3000$.

Thus, we observe that the pair (μ, σ) do not adequately describe a distribution. We need to know much more about the distribution in order to provide a complete picture of the distribution, for example, the peakedness of the distribution or the tail behavior of the distribution. In an attempt to provide this information, we define two more parameters based on the central moments.

STOP: 9/20/21

Skewness and Kurtosis

Definition: **The skewness** of a r.v. Y or its distribution F is defined as

$$\beta_1 = \frac{E[(Y-\mu)^3]}{\sigma^3} = \frac{\mu_3}{(\mu_2)^{3/2}}.$$

We note the following properties of β_1:

1. The distribution of Y is said to be symmetric about θ if

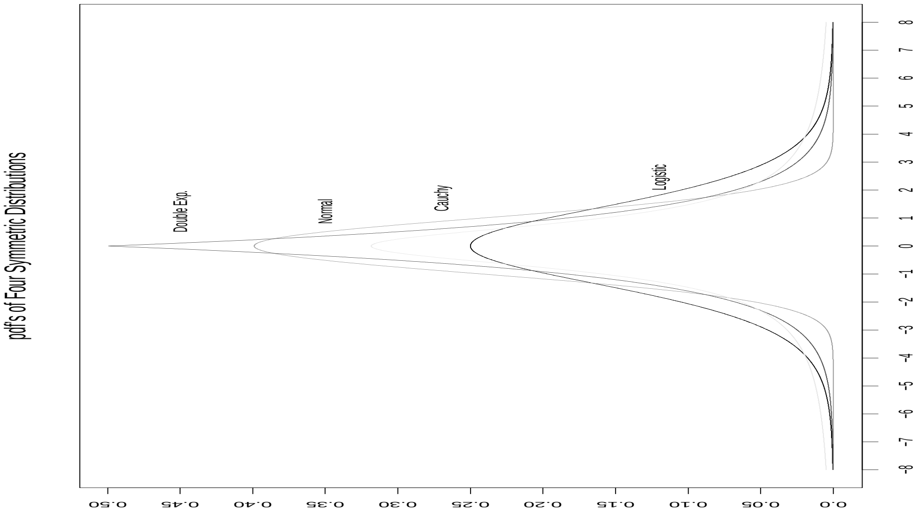
$$P[Y \le \theta - y] = P[Y \ge \theta + y] \ \text{ for all } \ y,$$

that is,

$$F(\theta - y) = 1 - F((\theta+y)^-) \ \text{ for all } \ y,$$

If F is symmetric with pdf f then $F(\theta - y) = 1 - F(\theta + y)$ for all y $\Rightarrow$

$$f(\theta - y) = f(\theta + y) \ \text{ for all } \ y.$$



- If the distribution of Y is symmetric about θ then $\beta_1 = 0$ and $\theta = \mu$ provided μ exists.

2. The converse of the above is not true in general. That is, $\beta_1 = 0$ does not necessarily imply that the cdf is symmetric See Ord (1968) *Annals of Mathematical Statistics*, 39, pp. 1513-1516. This article shows for the discrete t-distribution that $\beta_1 = 0$ but the distribution is not symmetric.

3. Right skewed distributions have $\beta_1 \ge 0$ and left skewed distributions have $\beta_1 \le 0$.

4. Although, $\beta_1 = 0$ does not imply symmetry in the distribution. The following is true

- If $\beta_1 \ne 0$, then the pdf f is not symmetric.

Definition: The kurtosis of a r.v. Y or its distribution F is defined as

$$\beta_2 = \frac{E[(Y-\mu)^4]}{\sigma^4} = \frac{\mu_4}{(\mu_2)^2} \quad \text{scaled 4th moment}$$

What does kurtosis actually measure. The following statement is taken from our textbook:

The kurtosis measures the tail-heaviness (the amount of probability in the tails) of the distribution. For the normal distribution, $\beta_2 = 3$. The normal distribution is considered a light-tailed distribution because the probability in its tails beyond, say, three standard deviations from the mean is negligible (.0027). Thus, depending on whether $\beta_2 > 3$ or $\beta_2 < 3$, a distribution is heavier tailed or lighter tailed than the normal distribution.

The above interpretation of kurtosis is often stated. However, this interpretation is very inaccurate. There are many articles addressing the interpretation of kurtosis. Two of these articles are

Kurtosis: A Critical Review, by K. Balanda and H. MacGillivray, *The American Statistician*, May 1988, Vol. 42, pp. 111-120.

The Meaning of Kurtosis: Darlington Reexamined, by J. Moors, *The American Statistician*, November 1986, Vol. 40, p. 283.

A few comments will be extracted from these articles:

- From Moors: *A valid interpretation (of kurtosis) may be formulated as follows: kurtosis measures dispersion around two values $\mu \pm \sigma$, it is an inverse measure of the concentration in these two points. High kurtosis, therefore, may arise in two situations:*

1. *concentration of probability mass near μ, corresponding to a peaked unimodal distribution and*

2. *concentration of probability mass in the tails of the distribution.*

- From Balanda and MacGillivray: Because of the "averaging" nature of moments, however, the relationship of β_2 to shape is far from clear;. . .

1. An error commonly associated with kurtosis is that the sign of $\beta_2 - 3$ compares the value of the density (function) at the center with that of the corresponding normal density.

2. The value of β_2 is affected by so many different aspects of a distribution that . . . a given value of β_2 can correspond to several different distributional shapes.

3. Figure 2 contains a number of standardized symmetric densities with $\mu = 0, \sigma = 1, \beta_2 = 3$ (the value for the normal distribution). Although Curve 3 has finite support (and thus short tails) it is a good approximation to the normal distribution. Curve 4 is bimodal whereas Curve 2, although it has infinite support and is unimodal, is considerably more peaked than the standard normal distribution. Thus, there is no logical connection between the value of the density (function) of the standardized distribution at the center and the sign of $\beta_2 - 3$.

Curve 1 - Normal Distribution Curve 3 - TukeyLambda[L=.135] Distribution
Curve 2 - TukeyLambda[L=5.2] Distribution Curve 4 - Double Gamma Distribution

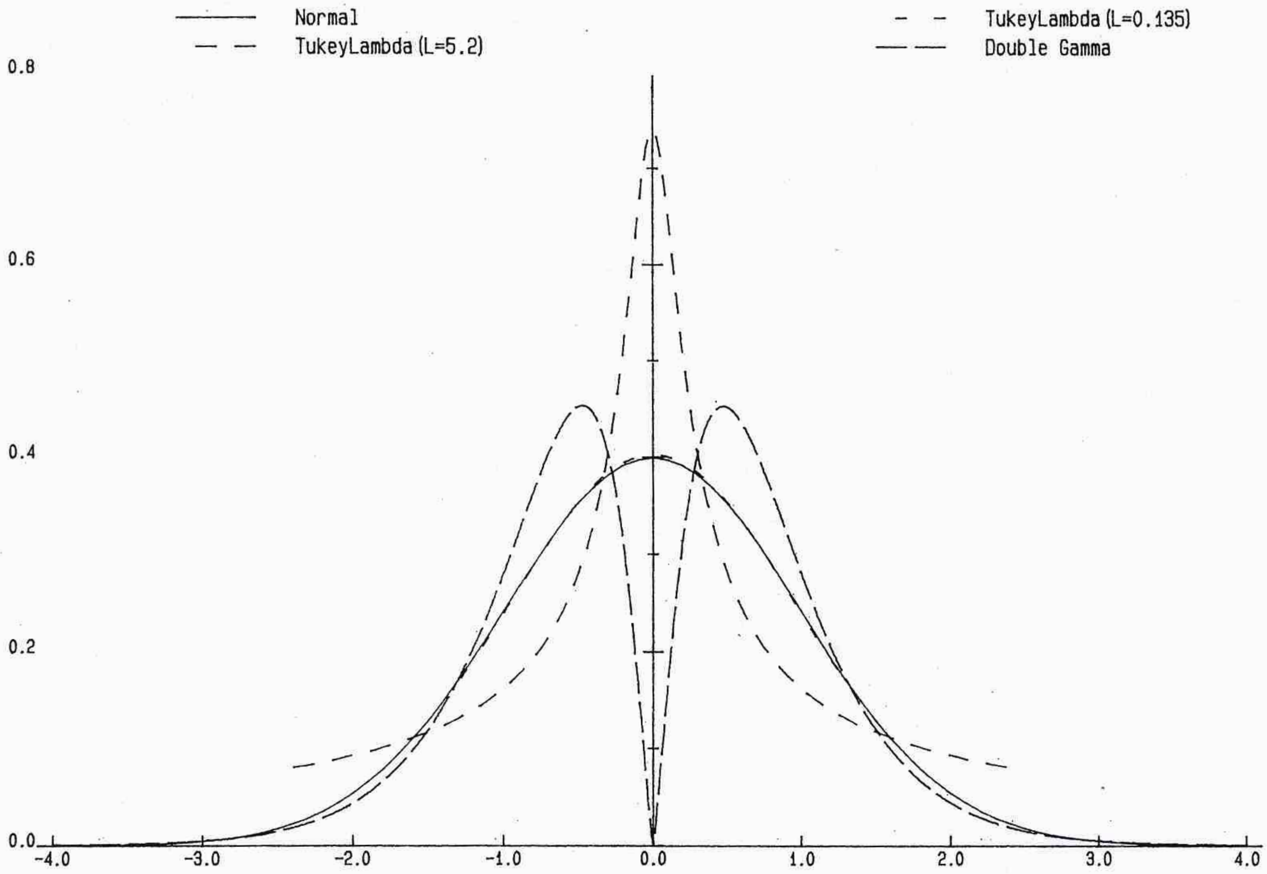


Figure 2. Standardized Symmetric Densities With γ_2 = 0: Standard Normal Distribution; Symmetric Tukey Lambda Distribution With λ .135; Symmetric Tukey Lambda Distribution With λ = 5.2; Double Gamma Distribution With α = $(1 + 13^{1/2})/2$.

The following discussion was sent to me but I do not have the reference:

There are two fallacies that are commonly associated with the shape parameters, skewness and kurtosis. The first fallacy is that two distributions having the same mean, standard deviation, skewness, and kurtosis have the same shape. In fact, Karl Pearson stated in the early 1900's that knowing $\mu, \sigma, \beta_1, \beta_2$ would completely describe a distribution. The second fallacy is that a distribution with a skewness parameter of zero will be symmetric. That these are indeed fallacies will be illustrated by the following examples. Consider the following two pdfs:

$$f_1(y) = \begin{cases} 0.6391 + 1.0337y & \mathit{if} \quad -0.0091 < y < 0.5387 \\ 1.7527 - 1.0337y & \mathit{if} \quad 0.5387 < y < 1.0864 \\ 0 & \mathit{if} \quad y \leq -0.0091 \ \text{ or } \ y \geq 0.5387 \end{cases}$$

$$f_2(y) = \begin{cases} (18.1484)(.0629)y^{-19.1484}\left[1 + y^{-18.1484}\right]^{-1.0629} & \mathit{if} \quad y > 0 \\ 0 & \mathit{if} \quad y \leq 0 \end{cases}$$

The two pdfs have the same values for the parameters:

$$\mu = .5387 \qquad \sigma = .2907 \qquad \beta_1 = 0 \qquad \beta_2 = 2$$

The pdfs are graphed below. Thus, even though the two pdfs have the same values for mean, standard deviation, skewness and kurtosis, they definitely do not have the same shape. Also, pdf f_2 has skewness coefficient equal to 0 when it is obvious that f_2 is nonsymmetric.

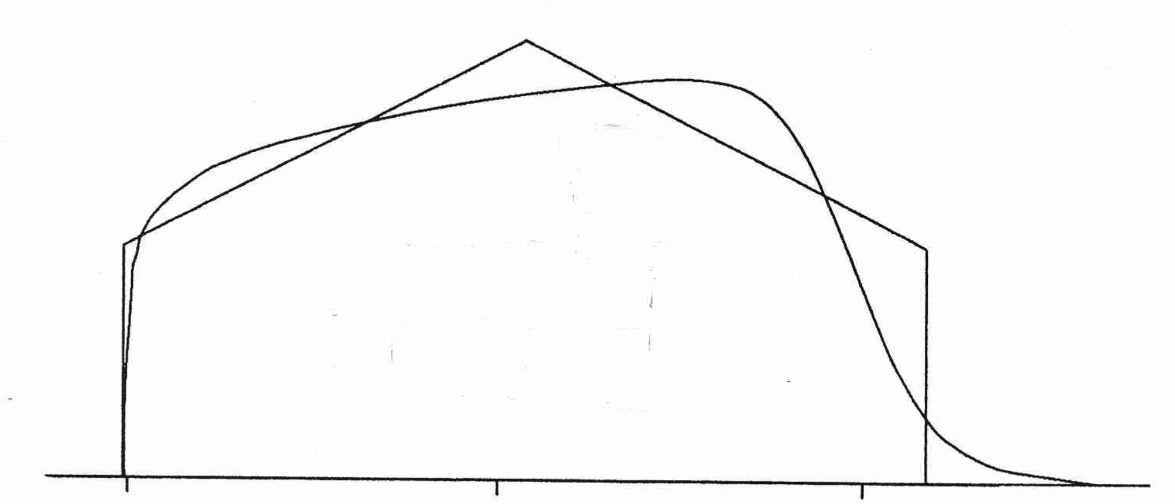


Figure 13.7: The Usefulness of Skewness and Kurtosis

Casella-Berger on page 64 display two pdf's which have all their moments, m_r for $r = 1, 2, 3, \cdots$, equal, but the two pdf's are very different. The moments of a distribution, m_r, uniquely determine the distribution, only if the Carleman's Condition holds: $\sum_{r=1}^{\infty} m_{2r}^{-1/2r} = \infty$

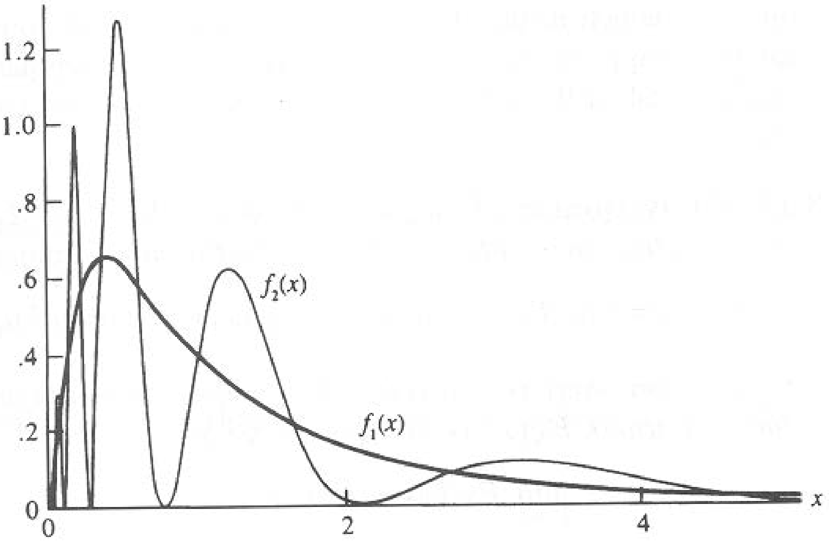


Figure 2.3.2. *Two pdfs with the same moments:* $f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}$ *and* $f_2(x) = f_1(x)[1 + \sin(2\pi \log x)]$

The following table contains values for the four parameters μ, σ, β_1, and β_2 for a wide variety of distributions.

TABLE 4.1
Some frequently encountered continuous probability distribution functions

Distribution	Probability Density Function		Mean	Variance
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	$-\infty < \mu < \infty$ $\sigma > 0$	μ	σ^2
χ^2	$f(x) = \frac{x^{(n/2)-1}e^{-(x/2)}}{2^{n/2}\Gamma(n/2)}$	$x > 0$ $n > 0$	n	$2n$
t	$f(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{[(n+1)/2]}$	$n > 0$	0	$\frac{n}{n-2}$ $(n>2)$
F	$f(x) = \frac{n^{n/2}m^{m/2}}{\beta(n/2, m/2)} x^{(n-2)/2}(m+nx)^{-(n+m)/2}$	$x > 0$ $m,\ n > 0$	$\frac{m}{m-2}$ $(m > 2)$	$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ $(m > 4)$
Gamma	$f(x) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1}e^{-\lambda x}$	$x > 0$ $\lambda, r > 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
Exponential	$f(x) = \lambda e^{-\lambda x}$	$\lambda > 0$ $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Weibull	$f(x) = \alpha\beta x^{\beta-1}\exp(-\alpha x^{\beta})$	$x > 0$ $\alpha, \beta > 0$	$\alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right)$	$\alpha^{-\frac{2}{\beta}}\left\{\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma^2\left(1+\frac{1}{\beta}\right)\right\}$
Lognormal	$f(x) = \frac{1}{\beta\sqrt{2\pi}} x^{-1}\exp\left[-(\ln x - \alpha)^2/2\beta^2\right]$	$x > 0$ $\beta > 0$	$\exp\left(\alpha + \frac{\beta^2}{2}\right)$	$\exp(2\alpha + \beta^2)\left[\exp(\beta^2) - 1\right]$
Beta	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$0 < x < 1$ $\alpha, \beta > 0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Uniform	$f(x) = \frac{1}{b-a}$	$a \leq x \leq b$ $-\infty < a < b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Distribution	β_1	β_2
Normal	0	3
χ^2	$\frac{8}{n}$	$3\left(\frac{4}{n} + 1\right)$
t	0	$\frac{6}{n-4} + 3 \qquad (n > 4)$
F	$\frac{8(2n+m-2)^2(m-4)}{n(m-6)^2(n+m-2)} \qquad (m > 6)$	$\frac{(m-2)^3(m-4)(n+6)(n+4)(n+2)}{4(m-6)(m-8)(n+m-2)^2n^2} - \frac{8(m-4)(2n+m-2)}{(m-6)(n+m-2)} - \frac{3n(m-4)}{(n+m-2)} - \frac{n^2(m-4)}{4(n+m-2)^2}$ $(m > 8)$
Gamma	$\frac{4}{r}$	$\frac{6}{r} + 3$
Exponential	4	9
Weibull	$\left\{\Gamma\left(1+\frac{3}{\beta}\right) - 3\Gamma\left(1+\frac{1}{\beta}\right)\Gamma\left(1+\frac{2}{\beta}\right) + 2\Gamma^3\left(1+\frac{1}{\beta}\right)\right\}^2 \ast$	$\Gamma\left(1+\frac{4}{\beta}\right) - 4\Gamma\left(1+\frac{1}{\beta}\right)\Gamma\left(1+\frac{3}{\beta}\right) + 6\Gamma^2\left(1+\frac{1}{\beta}\right)\Gamma\left(1+\frac{2}{\beta}\right) - 3\Gamma^4\left(1+\frac{1}{\beta}\right) \ast$
	$\ast \left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma^2\left(1+\frac{1}{\beta}\right)\right]^3$	$\ast \left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma^2\left(1+\frac{1}{\beta}\right)\right]^2$
LogNormal	$\left[\exp\left(\beta^2\right) - 1\right]\left[\exp\left(\beta^2\right) + 2\right]^2$	$\left[\exp\left(\beta^2\right) - 1\right]\left[\exp\left(3\beta^2\right) + 3\exp\left(2\beta^2\right) + 6\exp\left(\beta^2\right) + 6\right] + 3$
Beta	$\frac{4(\beta-\alpha)^2(\alpha+\beta+1)}{\alpha\beta(\alpha+\beta+2)^2}$	$\frac{3(2\alpha^2 + \alpha^2\beta - 2\alpha\beta + \alpha\beta^2 + 2\beta^2)(\alpha+\beta+1)}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$
Uniform	0	1.8

From: Statistical Analysis for Engineers and Scientists by J.W. Barnes

Alternative Measures of Location/Center in a Distribution

When the population/process distribution has very heavy tails, the mean μ may not exist, for example, the Cauchy distribution. Also, when the distribution is heavily right or left skewed, the mean μ may be a very poor representation of the "center" or "typical value" of the population or process distribution. If both of these cases, we will seek an alternative to the mean μ.

The median $\tilde{\mu}$ is the most commonly used alternative to the μ as a representation of the center of a population distribution. To illustrate its performance relative to the mean for a highly skewed distribution we will examine the lognormal distribution:

Mean and Median of Lognormal Distribution

To illustrate the distribution of a population about its mean and median, we will next calculate the proportion of the values of a random variable Y less than its Mean μ_Y and the proportion less than its Median $\tilde{\mu}_Y$, when the distribution of Y is Lognormal distribution with parameters $\theta_1 = 0$, and θ_2, for various values of θ_2.

Let X have a $N(\theta_1, \theta_2^2)$ distribution then $\theta_1 = \mu_X = \tilde{\mu}_X$

- $Y = e^X$ has a lognormal distribution

- $\mu_Y = e^{\theta_1 + \frac{1}{2}\theta_2^2}$ (Using mgf of X)

- $\tilde{\mu}_Y = e^{\theta_1}$ This follows from

  $$.5 = P[Y \le \tilde{\mu}_Y] = P[X \le log(\tilde{\mu}_Y)] \quad \Rightarrow \quad log(\tilde{\mu}_Y) = \tilde{\mu}_X = \theta_1$$

- Set $\theta_1 = 0$, then $\mu_Y = e^{\frac{1}{2}\theta_2^2}$ and $\tilde{\mu}_Y = e^0 = 1$. We then have that

  $$P\left(Y \le \mu_Y\right) = P\left(Z \le \frac{\theta_2}{2}\right),$$

  where Z is N(0,1).

  This follows from

  $$P(Y \le \mu_Y) = P(log(Y) \le log(\mu_Y)) = P\left(X \le \frac{\theta_2^2}{2}\right) = P\left(\frac{X}{\theta_2} \le \frac{\theta_2}{2}\right)$$

- By the definition of $\tilde{\mu}_Y$, $P(Y \le \tilde{\mu}_Y) = 0.5$

For the lognormal distribution with parameters, θ_1 and θ_2, the skewness parameter is given by

$$\beta_1 = \left(e^{\theta_2^2} - 1\right)\left(e^{\theta_2^2} + 2\right)^2$$

θ_2	β_1	μ_Y	$\tilde{\mu}_Y$	$P(Y \leq \mu_Y)$	$P(Y \leq \tilde{\mu}_Y)$
0.50	3.06e+00	1.1	1	0.599	0.5
0.75	1.06e+01	1.3	1	0.646	0.5
1.00	3.82e+01	1.6	1	0.691	0.5
1.25	1.72e+02	2.1	1	0.734	0.5
1.50	1.12e+03	3.0	1	0.773	0.5
1.75	1.11e+04	4.6	1	0.809	0.5
2.00	1.71e+05	7.3	1	0.841	0.5
2.25	4.01e+06	12.5	1	0.870	0.5
2.50	1.39e+08	22.7	1	0.894	0.5
2.75	7.14e+09	43.8	1	0.915	0.5
3.00	5.32e+11	90.0	1	0.933	0.5
3.25	5.77e+13	196.6	1	0.948	0.5
3.50	9.12e+15	457.1	1	0.960	0.5
3.75	2.09e+18	1131.4	1	0.970	0.5
4.00	7.01e+20	2980.9	1	0.977	0.5
4.25	3.41e+23	8360.3	1	0.983	0.5
4.50	2.41e+26	24959.2	1	0.988	0.5
4.75	2.49e+29	79320.3	1	0.991	0.5
5.00	3.73e+32	268337.2	1	0.994	0.5

From the table, we can observe

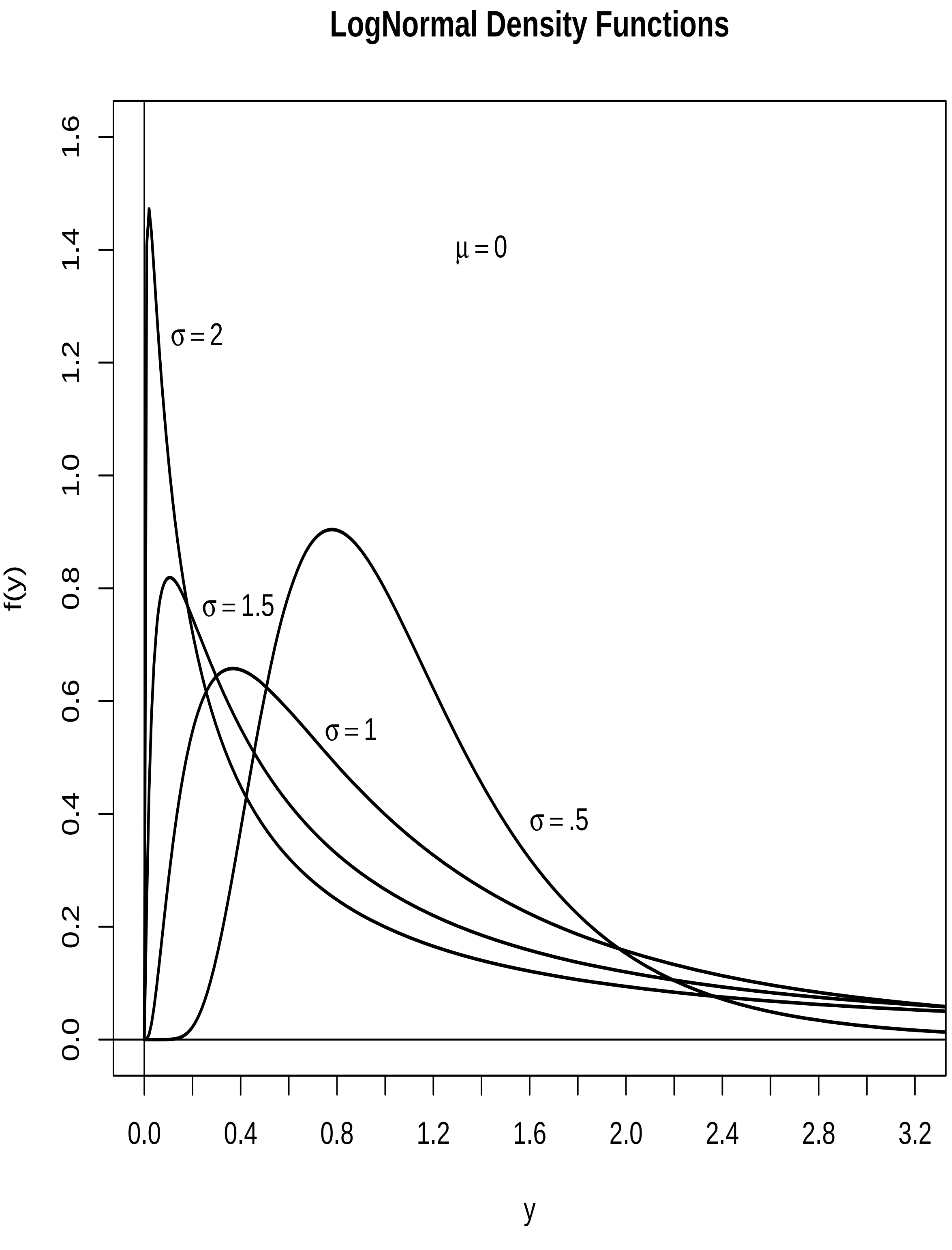
- For highly skewed distributions, the mean μ is not an appropriate representation of the center of the distribution. Even though there is only a small percent of the distribution in the tail of a highly skewed distribution, these few very large values (relative to the rest of the distribution), cause the mean of the distribution to be very large.

- The median does not have this flaw. Whereas, the mean is the weighted average of all values for the r.v., the median is not affected by the extreme values for the r.v.

- The median is the true middle value dividing the distribution into two equal parts irregardless of the size of the extremes in the population/process.

- The median $\tilde{\mu}$ is the more appropriate parameter in these situations for representing the center of the distribution or the typical value for the population/process.

- Suppose we have a population distribution which is highly right skewed. Based on a random sample of n observation, we want to estimate the population total, $T = \sum_{i=1}^{N} Y_i$. Should we use $\widehat{T} = N\widehat{\mu}$ or $\widehat{T} = N\tilde{\mu}$, mean or median as the "typical value" ?


LogNormal Density Functions
μ = 0
σ = 2
σ = 1.5
σ = 1
σ = .5
f(y)
y
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2

In general we will make use of all three quartiles when describing the "center" of a population distribution:

$$Q_1 = Q(.25) \qquad \tilde{\mu} = Q_2 = Q(.5) \qquad Q_3 = Q(.75)$$

Note that the middle 50% of the distribution falls between Q_1 and Q_3, with 25% between Q_1 and $\tilde{\mu}$ and 25% between $\tilde{\mu}$ and Q_3.

How is the median and quartiles related to the mean?

From our results from the lognormal distribution it would appear there is no relationship between the mean and median. However, we know that

$$\mu_Y = e^{\theta_1 + \frac{1}{2}\theta_2^2} = \tilde{\mu}_Y e^{\frac{\theta_2^2}{2}}$$

For the lognormal distribution with parameter $\theta_1 = 0$, as the lognormal distribution became more skewed the mean of the distribution increased to a very large value whereas as the median remained constant. However, the standard deviation of the lognormal also was increasing very rapidly as can be seen from

$$\sigma_Y = e^{\theta_1}\sqrt{e^{\theta_2^2} - 1}$$

In fact, we have shown that for Y distributed logNormal(θ_1, θ_2),

$$\mu_Y = \tilde{\mu}_Y e^{\theta_2^2/2}$$

For the lognormal distribution, the mean of Y is related to both the median and the θ_2 parameter.

We can establish the following relationship between any percentile $Q(p)$ and the pair (μ, σ) for any distribution which has $|\mu| < \infty$ and $\sigma < \infty$:

$$|Q(p) - \mu| \;\leq\; \sigma \max\left\{\sqrt{\frac{1-p}{p}}, \;\; \sqrt{\frac{p}{1-p}}\right\}.$$

In particular, the median (p=1/2) satisfies

$$|\tilde{\mu} - \mu| \leq \sigma$$

That is, for any distribution for which the mean and standard deviation exist, the median and mean differ by at most one standard deviation.

From this statement we cannot infer that the mean will necessarily be *close* to the median.

Why?

Recall the lognormal example in which σ can be very large for large values of θ_2.

Trimmed Mean

When a distribution is very heavy-tailed, extremes to the central values can occur with a reasonably high frequency (Cauchy, logistic, t with small df). In these situations, the mean μ can be drawn a considerable distance from the median $\tilde{\mu}$.

An alternative to the mean, other than the median, which reduces the influence of extremes in the distribution is the trimmed mean. In heavy-tailed distributions, we "trim" off the extreme values prior to averaging the values in the population/process.

Definition: The $\alpha-$**Trimmed Mean** $\mu_{(\alpha)}$ is the mean of the population after excluding the smallest and largest $100\alpha\%$ of the distribution.

Let Y have pdf f and quantile function Q. After trimming off the smallest and largest $100\alpha\%$ of the distribution, it is necessary to renormalize the area under the pdf so that the total area is 1.

This results in the following pdf for the $\alpha-$trimmed distribution:

$$f_{(\alpha)}(y) = \begin{cases} \frac{1}{1-2\alpha} f(y) & if \quad Q(\alpha) \le y \le Q(1-\alpha) \\ 0 & if \quad y < Q(\alpha) \\ 0 & if \quad y > Q(1-\alpha) \end{cases}$$

$\frac{1}{1-2\alpha}$ is a normalizing constant which guarantees that the total area under $f_{(\alpha)}(y)$ is 1.

Using this pdf for the trimmed distribution, we obtain the following formula for the $\alpha-$trimmed mean:

$$\mu_{(\alpha)} = \int_{-\infty}^{\infty} y f_{(\alpha)}(y) dy = \frac{1}{1-2\alpha} \int_{Q(\alpha)}^{Q(1-\alpha)} y f(y) dy.$$

In the limit we obtain both the mean and median from the $\alpha-$trimmed mean:
As the amount of trimming is reduced, the trimmed mean converges to the population mean:

$$\lim_{\alpha \to 0} \mu_{(\alpha)} = \mu$$

As the amount of trimming is increased to 50%, the trimmed mean converges to the population median. (See Assignment 4 for a proof).

$$\lim_{\alpha \to .5} \mu_{(\alpha)} = \tilde{\mu} \quad \ast$$

Measures of Dispersion/Variability in a Population/Process Distribution

In a manufacturing process, a product is produced having a nominal physical characteristic of size θ_o:

- Producing ball bearings with a nominal diameter of 5 cm ($\theta_o = 5cm$)
- Producing an antibiotic with a nominal weight of 10 mg ($\theta_o = 10mg$)
- Producing gasoline with a nominal amount of an additive of .8% ($\theta_o = .8\%$)

In such situations, we want to determine

1. If the "average" value of the physical characteristic equals the nominal value and
2. How near are the values of the physical characteristic in a daily production run to this "average" value.

For example, suppose our daily production of ball bearings comes from two processes:

Process 1: $\mu = 5$cm with 15% of output less than 4.8 cm and 15% of output greater than 5.2 cm.

Process 2: $\mu = 5.002$cm with .1% of output less than 4.8 cm and .12% of output greater than 5.2 cm.

Which process is doing a better job of producing ball bearings with a diameter of 5 cm?

We need to evaluate both the location and the dispersion of the measured characteristic in a population/process.

There are a number of possible methods to measure the dispersion of the population/process about a central value. We will now define a few of these parameters.

1. **Definition:** **The Range** is the distance from the smallest possible value to the largest possible value in the population.

   For most distributions, the range is not very useful because it is generally infinity. Also, it depends on just two values and does not describe to any degree the clumping or lack of clumping of values about the measure of the center of the distribution.

   normal distribution has Range = 2∞,   Weibull has Range = ∞,

   Beta on [0,1], has Range =1.

2. **Definition:** **The Standard Deviation** is a measure of the concentration of the values about the population/process mean μ:

$$\sigma = \sqrt{E\left[(Y-\mu)^2\right]} = \sqrt{\int_{-\infty}^{\infty} (y-\mu)^2 dF(y)}$$

   σ is the "weighted average" distance of values about μ

   For very heavy-tailed distributions μ and/or σ may not exist (Cauchy and t with df$\leq$ 2)

   Highly skewed and heavy-tailed distributions result in highly inflated values of σ which may not represent a high concentration of values about μ with just a small proportion of values extreme to μ.

3. **Definition:** **The Interquartile Range (IQR)** measures the distance to cover the middle 50% of distribution

$$IQR = [(Q(.75) - Q(.25)]$$

4. **Definition:** **The Semi-Interquartile Range (SIQR)** measures the spread of the middle 50% of distribution about the median $\tilde{\mu}$:

$$SIQR = \frac{1}{2}[(Q(.75) - \tilde{\mu}) + (\tilde{\mu} - Q(.25)] = \frac{1}{2}(D_2 + D_1) = \frac{1}{2}IQR$$

   SIQR is the average distance from the lower quartile to the median, D_1 and from median to the upper quartile, D_2.

   For very skewed distributions and very heavy-tailed distributions, the pair $(\tilde{\mu}, SIQR)$ is often a better representation of center, variability than is (μ, σ). However, it only is dealing with the middle 50% of the data (from Q(.25) to Q(.75)) and thus is trimming 25% of the data in both tails prior to measuring spread. Too insensitive to tail behavior.

```
  |<- - - - - - -  D1 - - - - - - ->|<- - - D2 - - ->|
  |----------------------------------|----------------|
Q(.25)                             Q(.5)         Q(.75)
```

An alternative to the SIQR for very skewed distributions and very heavy-tailed distributions is **MAD**:

5. **Definition: The Median Absolute Deviation about the Median**

$$MAD = median\{|Y - \tilde{\mu}|\}$$

MAD measures the dispersion of the data about the median by taking the median absolute distance from the values in the population/process to the median of the population/process.

MAD is not affected by extremes in the population/process

Several versions of MAD

(a) Mean absolute deviation about the mean:

$$MAD_1 = E[|Y - \mu|]$$

Eliminates the problem in σ due to squaring distance for extreme observations but still is taking average over all values in population/process and still uses μ as the measure of center

(b) Mean absolute deviation about the median:

$$MAD_2 = E[|Y - \tilde{\mu}|]$$

Eliminates the problem of using μ as the measure of center but still is taking average over all values in population/process

(c) Median absolute deviation about the median:

$$MAD = Median|Y - \tilde{\mu}|$$

Eliminates the problem of using μ as the measure of center and eliminates problem of including the extremes of the population/process in computing the average distance

Some further observations about MAD.

1. Let Y be a r.v. with median $\tilde{\mu}$ and let $W = |Y - \tilde{\mu}|$.

   Then MAD is just the median of the distribution of W.

2. For a symmetric distribution with $\mu < \infty$ we have

$$\mu = \tilde{\mu} \quad \text{and} \quad MAD = SIQR$$

3. For a $N(\mu, \sigma^2)$ distribution,

$$\mu = \tilde{\mu} \quad \text{and} \quad MAD = 0.6745\sigma$$

   Normal distribution is symmetric $\Rightarrow$ MAD = SIQR

   $Q(.25) = \mu + Q_Z(.25)\sigma = \mu - .6745\sigma$

   $Q(.75) = \mu + Q_Z(.75)\sigma = \mu + .6745\sigma$

   MAD = SIQR = $\frac{1}{2}\left(Q(.75) - Q(.25)\right) = .6745\sigma$

4. Often MAD will be defined as

$$MAD = median|Y - \tilde{\mu}|/.6745$$

   For a normal distribution MAD and σ are the same. This is the definition used in R.

5. For the logNormal(θ_1, θ_2) distribution, $\sigma = e^{\theta_1}\sqrt{e^{\theta_2^2}\left[e^{\theta_2^2} - 1\right]}$

   For $\theta_1 = 0$ and $MAD = median(|Y - \tilde{\mu}|)/0.6745$, we have

	θ_2					
	.5	.75	1.0	1.5	2.0	2.5
σ	.604	.1.15	2.16	8.97	54.1	517.5
MAD	.485	.70	.89	1.17	1.34	1.42

Recommendations on the Selection of Measures of Center/Dispersion about Center

Why not just use the pair $(\tilde{\mu}, MAD)$ for all distributions?

1. When the population/process distribution is not heavily skewed or too heavily tailed (near normal in shape), (μ, σ) provide a more complete picture of the distribution.

   Furthermore, when we use data to estimate these parameters, the sample counterparts of (μ, σ) are more efficient estimators than are the the sample counterparts of $(\tilde{\mu}, MAD)$

2. When the population total $T = \sum_{i=1}^{N}$ is the parameter of interest to the researcher, the sample estimator $\hat{T} = N\hat{\mu}$ is a better estimator of T than is the estimator $\hat{T} = N\hat{\tilde{\mu}}$.

   When estimating the population total we will in most cases want the effect of extremes to be a part of the estimator.

   Example Let T = daily amount of pollutants discharged in a river from a chemical plant. The days in which there was a large discharge would have a large impact on the yearly total. Thus, in estimating the year total discharge, YT = $\sum_{i=1}^{365} T_i$ we would choose between 365 times the average daily discharge and 365 times the median daily discharge. However, $365\hat{\tilde{\mu}}_T << 365\hat{\mu}_T$. Thus, we could greatly underestimate YT if we used median in place of mean. However, if we wanted to obtain information on the "typical" daily discharge the median may be a more appropriate representative than the mean.

3. For highly skewed and/or heavy-tailed distributions use $(\tilde{\mu}, MAD)$.

   For Cauchy and t with $df \leq 2$, μ and σ do not exist but $(\tilde{\mu}, MAD)$ always exist.

4. If $(\theta_1,\ \theta_2)$ are location-scale parameters for a distribution, is the following true?

$$\mu = \theta_1 \quad \text{and} \quad \sigma = \theta_2$$

   a. If the distribution is Cauchy$(\theta_1,\ \theta_2)$, then μ and σ do not exist but $(\theta_1,\ \theta_2)$ are valid location-scale parameters

   b. If the distribution is Uniform on $(\theta + a,\ \theta + b)$ then θ is a location parameter but

$$\mu = \theta + \frac{a+b}{2} \neq \theta \quad \text{and} \quad \sigma = \sqrt{(b-a)^2/12}$$

   c. If the distribution is logistic$(\theta_1,\ \theta_2)$ then $(\theta_1,\ \theta_2)$ are valid location-scale parameters.

$$\mu = \theta_1 \quad \text{but} \quad \sigma = \frac{\pi\theta_2}{\sqrt{3}} \neq \theta_2$$

Measures of Association Amongst Vectors of R.V.s

Suppose we are dealing with a population/process where we measure several variables on each unit in the population/process or we observe a single characteristic of the unit across time or space. In these situations we want to determine the degree to which these variables are related.

> Example 1: Suppose a company is developing a new production process for producing an alloy used in golf clubs. The company would be interested in measuring:
>
> (a) X_1 - Rockwell hardness of alloy
> (b) X_2 - Tensile strength of alloy
> (c) X_3 - Smoothness of surface of alloy
> (d) X_4 - % Carbon in alloy

> Example 2: Suppose a new therapy of treating patients with high blood pressure is under evaluation. The medical doctors would be interested in measuring on each patient:
>
> (a) W_1 - Age
> (b) W_2 - Blood pressure prior to starting therapy
> (c) W_3 - Blood pressure at conclusion of therapy
> (d) W_4 - BMI
> (e) W_5 - Cholestrol level
> (f) W_6 - Hours per week of exercise
> (g) W_7 - Yes or No for a stroke

Example 3: Suppose we are investigating how long a carcinogen in an industrial discharge remains in the atmosphere

D_0 - Amount of carcinogen in air sample immediately prior to discharge

D_1 - Amount of carcinogen in air sample 1 minute after discharge

D_2 - Amount of carcinogen in air sample 2 minutes after discharge

.

.

.

D_{150} - Amount of carcinogen in air sample 150 minutes after discharge

In each of the three examples the researchers would be interested in measuring the degree of association between the vector of r.v.s.

Definition: **The Correlation** between two random variables Y and W having means and standard deviations: $\mu_Y, \mu_W, \sigma_Y, \sigma_W$, respectively, is given by

$$\rho_{Y,W} = Corr(Y,W) = \frac{E[(Y-\mu_Y)(W-\mu_W)]}{\sigma_Y \sigma_W} = \frac{Cov(Y,W)}{\sigma_Y \sigma_W}$$

$* \ cov(Y,W) = E(YW) - E(Y)E(W)$

1. Correlation is a unit-free measure of the linear relationship between the two variables.

2. $-1 \leq \rho_{Y,W} \leq 1$

3. $\rho_{Y,W} = \pm 1$ implies $Y = \beta_o + \beta_1 W$ where the sign of β_1 is the same as the sign of $\rho_{Y,W}$

4. $\rho_{Y,W}$ has the limitation of only measuring linear relationship.

   Thus, higher order relationships may not be detected

   Example: Let X have a symmetric distribution with mean μ and variance σ^2.

   Let $Y = (X-\mu)^2$. Thus, $E[Y] = \sigma^2$.

   $Corr(X,Y) = 0$ because

   $E[(X-\mu)(Y-\sigma^2)] = E[(X-\mu)^3] - \sigma^2 E[(X-\mu)] = 0 - 0 = 0$

   Thus, Y and X are uncorrelated but they are perfectly related by $Y = (X-\mu)^2$, a nonlinear relationship.

5. $\rho_{Y,W}$ only measures linear relationships between two of the many variables under study.

   Thus, may fail to detect nonlinear relationships that exist between several of the variables simultaneously. For example, $W_3 = W_1 e^{W_2}$

6. If X and Y are independent, $Corr(X, Y) = 0$:

   $$E[(X - \mu_X)(Y - \mu_Y)] = E[(X - \mu_X)]E[(Y - \mu_Y)] = (E(X) - \mu_X)(E(Y) - \mu_Y)] = (0)(0)$$

- The converse is not true. That is, $Corr(X, Y) = 0$ does NOT imply that X and Y are independent.

  Counter Example

  Let X have a pdf $f(x - \theta)$ which is symmetric distribution about 0 with $\mu_X = E[X] = \theta$

  Let $Y = I(|X - \theta| < 2)$, then X and Y are not independent:

  $$P[X > \theta + 3] \ > 0 \ \text{ but } \ P[X > \theta + 3|Y = 1] = 0$$

  However,

  $\mu_Y = E[Y] = P[|X - \theta| < 2] = \int_{\theta-2}^{\theta+2} f(x - \theta)dx = \int_{-2}^{2} f(t)dt$

  $$\begin{aligned} E[XY] = \int_{-\infty}^{\infty} xI(|x - \theta| < 2)f(x - \theta)dx &= \int_{\theta-2}^{\theta+2} xf(x - \theta)dx \\ &= \int_{-2}^{+2} (t + \theta)f(t)dt \\ &= \int_{-2}^{+2} tf(t)dt + \theta \int_{-2}^{+2} f(t)dt \\ &= 0 + \theta E[Y] = E[X]E[Y] \end{aligned}$$

We used the fact that $f(x - \theta)$ was symmetric about 0 to conclude that $\int_{-2}^{+2} tf(t)dt = 0$
Therefore, $E[XY] = E[X]E[Y]$ which implies $Cov(X, Y) = 0$ which implies $Corr(X, Y) = 0$
However, X and $Y = I(|X - \theta| < 2)$ are obviously not independent.

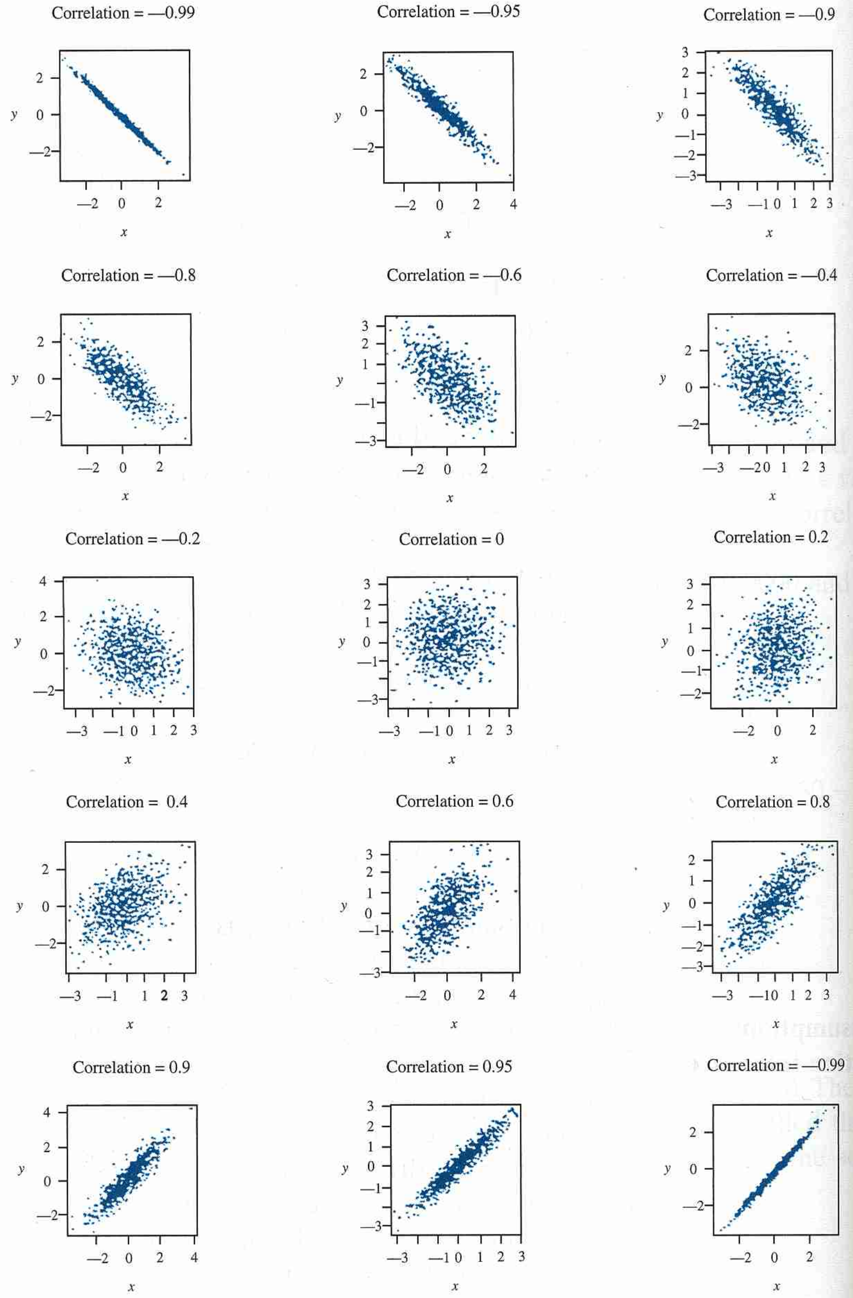
Special Case: If (X, Y) have a bivariate normal distribution, then

$Corr(X, Y) = 0$ implies X and Y are independent

The following scatter plots from *An Introduction to Statistical Methods and Data Analysis* will give an indication of what correlation is measuring.

FIGURE 11.20

Samples of size 1,000 from the bivariate normal distribution



Correlation = —0.99

Correlation = —0.95

Correlation = —0.9

Correlation = —0.8

Correlation = —0.6

Correlation = —0.4

Correlation = —0.2

Correlation = 0

Correlation = 0.2

Correlation = 0.4

Correlation = 0.6

Correlation = 0.8

Correlation = 0.9

Correlation = 0.95

Correlation = —0.99

START ON: 9/24/21 (Friday)

Time Series Data

When we are observing a physical characteristic over time, we are interested in the degree to which these measurements are associated. One measure of this association is the autocorrelation:

Definition: The AutoCorrelation of Order k in a series of stationary random variables: $X_t : t = 1, 2, 3, \ldots$ having the same mean μ and standard deviation σ is given by

$$\rho_k = \frac{E[(X_t - \mu)(X_{t-k} - \mu)]}{\sigma^2} \quad \text{for} \quad k = 1, 2, \ldots$$

↳ correlation btwn obs at time t & obs at time k.

ρ_k measures the degree of linear relationship in the random variable X over time or space.

ρ_1 the 1st order autocorrelation is the most widely used of these correlations.

A very simple, but widely used model for correlated over time or space observations is the $AR(1)$ model:

$$X_t = \theta + \rho X_{t-1} + e_t,$$

where $e_t s$ are iid with $E[e_t] = 0, Var(e_t) = \sigma_e^2$, $e_t s$ are independent of the $X_t s$ and $|\rho| < 1$.

Under this model, we can show that:

- $X_t s$ have mean:

$$\mu = E[X_t] = \theta + \rho E[X_{t-1}] + E[e_t] = \theta + \rho\mu + 0 \quad \Rightarrow \quad \mu = \frac{\theta}{1-\rho} \quad *$$

- $X_t s$ have variance:

$$\sigma_X^2 = Var(X_t) = Var(\theta + \rho X_{t-1} + e_t) = \rho^2 Var(X_{t-1}) + Var(e_t) = \rho^2\sigma_X^2 + \sigma_e^2 \Rightarrow \underbrace{\sigma_X^2 = \sigma_e^2/(1-\rho^2)}_{} \; *$$

- $X_t s$ are not independent:

$$\begin{aligned}
Cov(X_t, X_{t-1}) &= E[(X_t - \mu)(X_{t-1} - \mu)] \\
&= E[(\theta + \rho X_{t-1} + e_t - \mu)(X_{t-1} - \mu)] \\
&= \theta E[X_{t-1} - \mu)] + \rho E[(X_{t-1})^2] - \rho\mu E[X_{t-1}] + E[(e_t - \mu)(X_{t-1} - \mu)] \\
&= 0 + \rho(\sigma^2 + \mu^2) - \rho\mu^2 + E[e_t - \mu]E[X_{t-1} - \mu] \\
&= 0 + \rho(\sigma^2 + \mu^2) - \rho\mu^2 + 0 \\
&= \rho\sigma^2 > 0 \ \Rightarrow \ X_t \ \text{ and } X_{t-1} \text{ are Not independent}
\end{aligned}$$

- $\rho_k = Corr(X_t, X_{t-k}) = \rho^k \to 0$ as $k \to \infty$ because $|\rho| < 1$

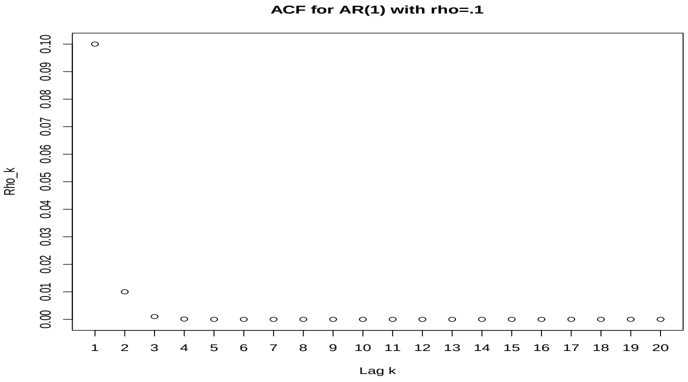
The plots on the next pages display the autocorrelation function (acf) for an AR(1)

with $\rho = .9, .6, .3, .1$.

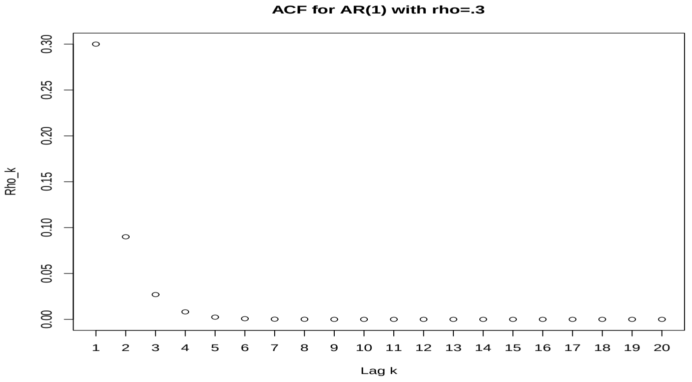
Also, there are corresponding plots of the time series obtained by simulating 300 observations from an AR(1)

with $\rho = .9, .6, .3, .1$ using the model

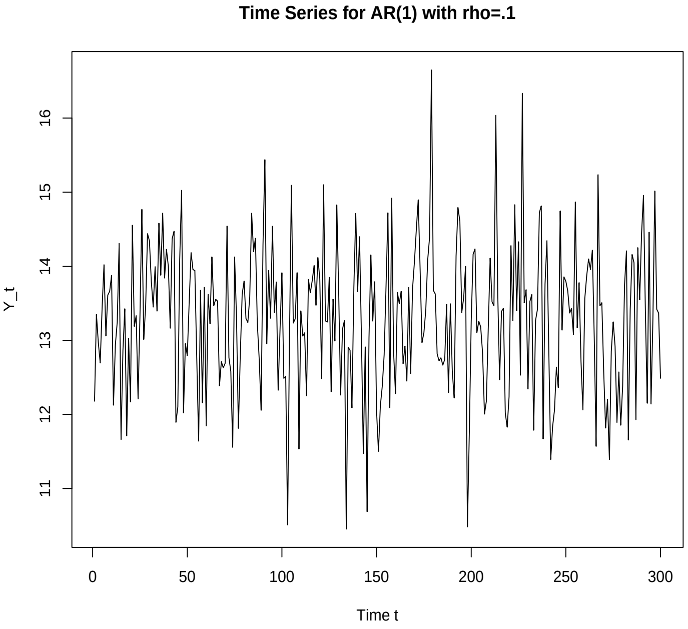
$$X_t = 12 + \rho X_{t-1} + e_t.$$



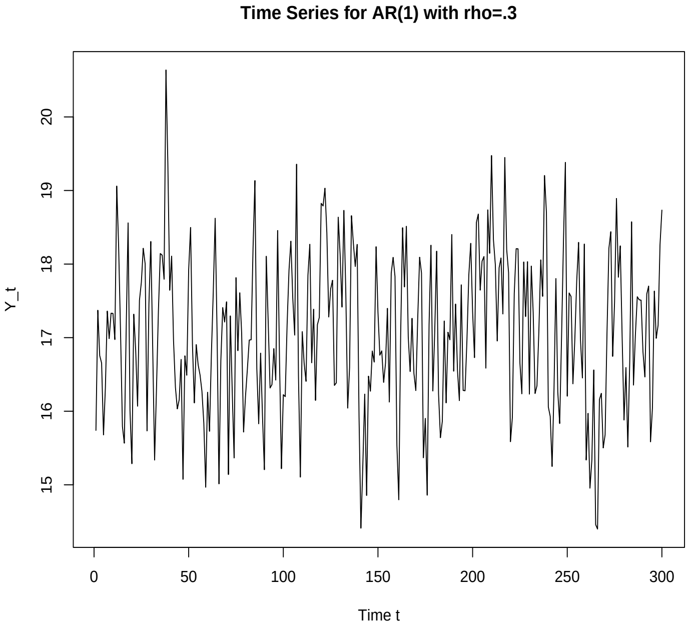
ACF for AR(1) with rho=.1



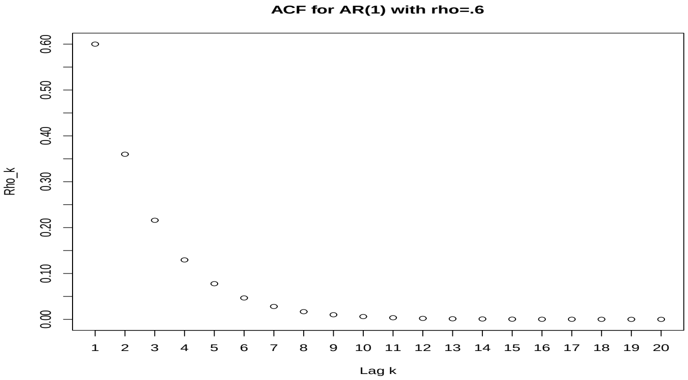
ACF for AR(1) with rho=.3

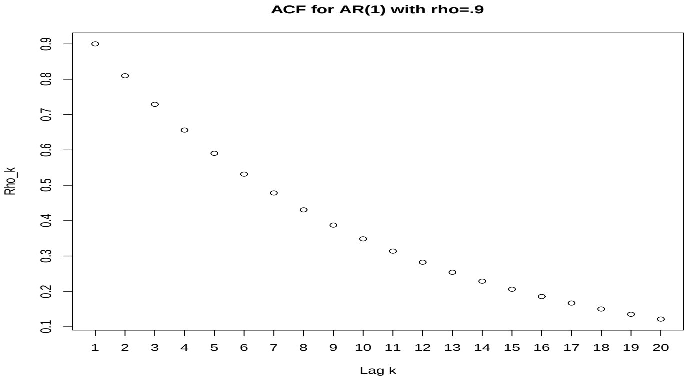


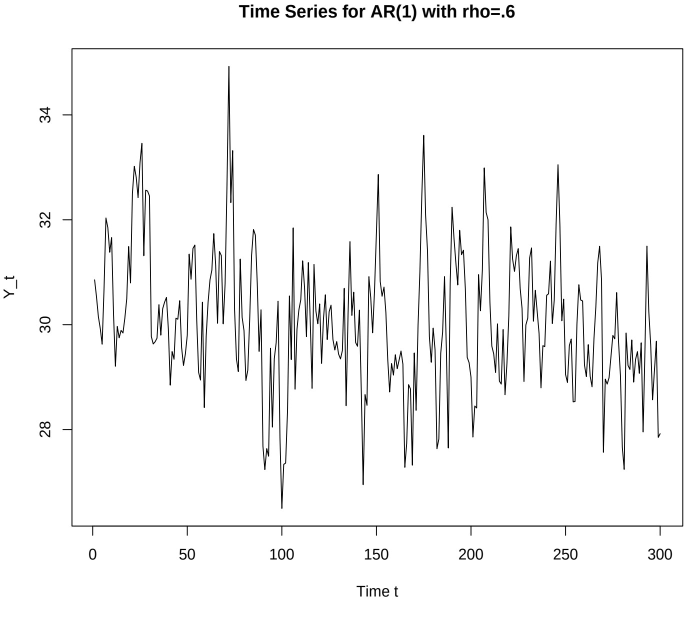
Time Series for AR(1) with rho=.1

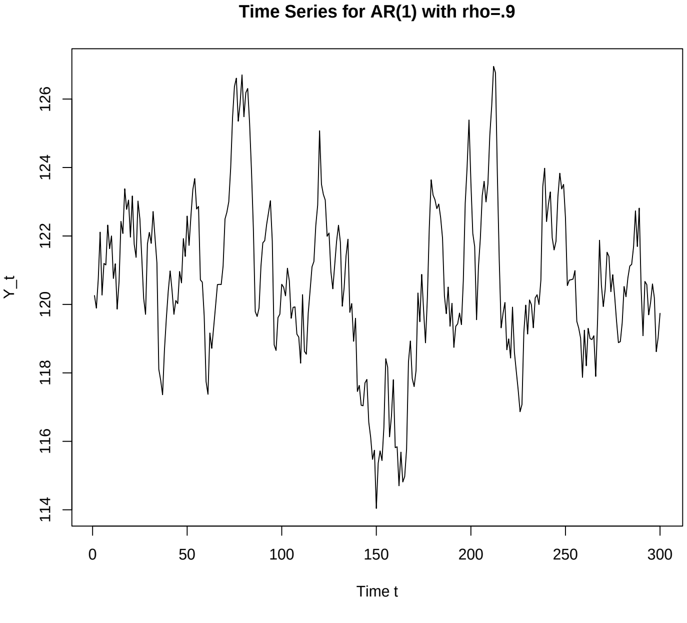


Time Series for AR(1) with rho=.3


ACF for AR(1) with rho=.6
Rho_k
0.00
0.10
0.20
0.30
0.40
0.50
0.60
Lag k
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


ACF for AR(1) with rho=.9
Rho_k
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Lag k
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


Time Series for AR(1) with rho=.6
Y_t
28
30
32
34
Time t
0
50
100
150
200
250
300


Time Series for AR(1) with rho=.9
Y_t
114
116
118
120
122
124
126
Time t
0
50
100
150
200
250
300

The plots on the previous pages were of stationary time series, that is, time series in which the mean and variance remained constant over time.

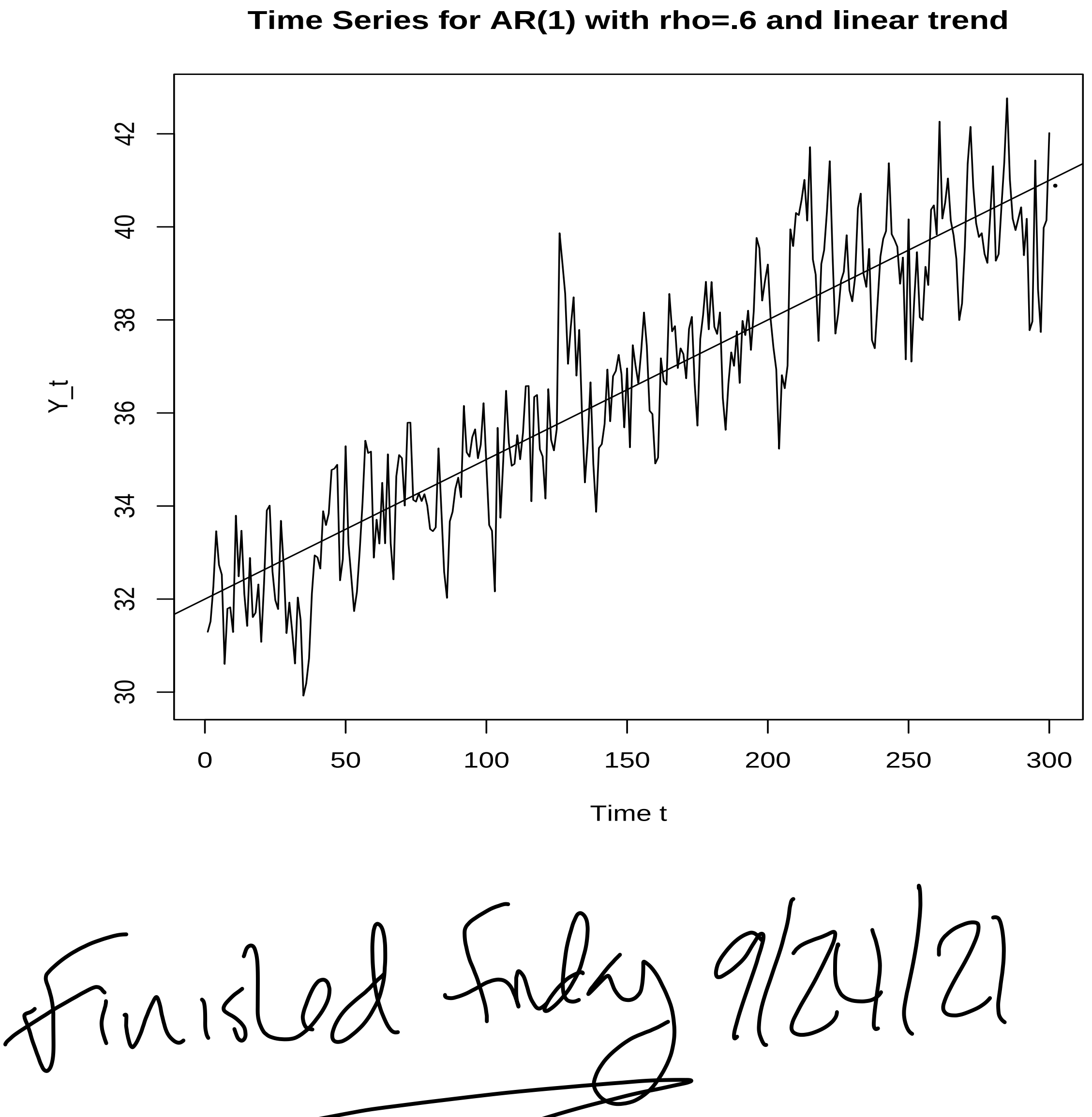
$$\mu_t = E[X_t] = \mu \quad \text{and} \quad \sigma_t^2 = Var(X_t) = \sigma^2 \quad \text{for} \quad t = 1, 2, 3, 4, \ldots$$

In many applications, the time series will not be stationary but will have a trend in its mean and/or variance.

In the plot given below we can observe that there is an increase in the mean of X_t with increasing t. This type of behavior often occurs when we are studying such measures as monthly sales or temperature or many other physical processes.

The process generating the data is given by

$$Y_t = 2 + .03t + .6Y_{t-1} + e_t \quad \text{where } e_t \text{ are iid N(0,1) r.v.s}$$



Finished Friday 9/24/21