MAT 125 Lecture 9. The Normal Distribution

Intro Stats, chapters 5 and 6

1 Overview

In Lecture 8, we discussed numerical summaries of the data which come down to defining some sort of representative value for the data set – an **average** and some sort of measure of the spread in the numbers – **the dispersion**. In today's lecture, we'll discuss theoretical distributions which model the phenomena which produced the data set. In particular, we will derive criteria to decide how likely a given value in a data set is to belong to the underlying distribution. Throughout this lecture, we will be using the dataset of president's ages when they were inaugurated, as we did in Lecture 8. In our sample of presidents' inaugural ages, we calculated a mean of 54.65 years and a standard deviation s of 6.33. You should note that s is the standard deviation of a **sample**. The true standard deviation of a **whole population** is denoted by σ (lowercase Greek letter "sigma"). Again, as the size of the sample approaches that of the whole population, s approaches σ. The standard deviation of a population is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \qquad (1)$$

Note that n denotes the size of the sample, and N denotes the size of the population. For the calculation of s we are introducing a correction because we use s as an estimate of σ, that's why we divide by n - 1 (instead of dividing by the number of all data points as we do for calculating σ).

2 The Normal Distribution

We can use the mean and the standard deviation to describe the distribution of our data set. If the data seem to "fall off" equally on either side of the mean, so that the data distribution looks like a bell, we can say that the data have a **normal distribution** with mean μ and standard deviation σ. The normal distribution is also called the **Gaussian distribution**, named for the mathematician Carl Friedrich Gauss. The data set on presidential ages is a good example of a distribution that is approximately normal (see Figure 1). If this distribution were truly normal, then as we increase the size of the sample, we expect the mean (54.65), the median (54.5), and the mode (54) to approach the same value, while the sample mean and standard deviation are expected to approach the true mean and standard deviation of the population (recall the law of Large Numbers from the probability lectures). Note that for the histogram, we summarized 5 years of age into bins whereas we used the individual values to calculate the measures of central tendency and the dispersion. We cannot, of course, increase the sample size without bound: there have been only a finite number of presidents and the data set in Table 1 is as big as it can get. For this and the many other such finite or small data sets, we are stuck with the inherent uncertainty in the numbers we derive for the mean and dispersion. It's the job of statistics to quantify how uncertain these numbers are.

A normal distribution can be described by the probability density function:

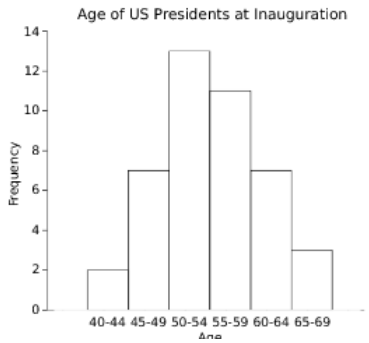


Figure 1. Histogram of US presidents' inaugural ages, in five-year bins.

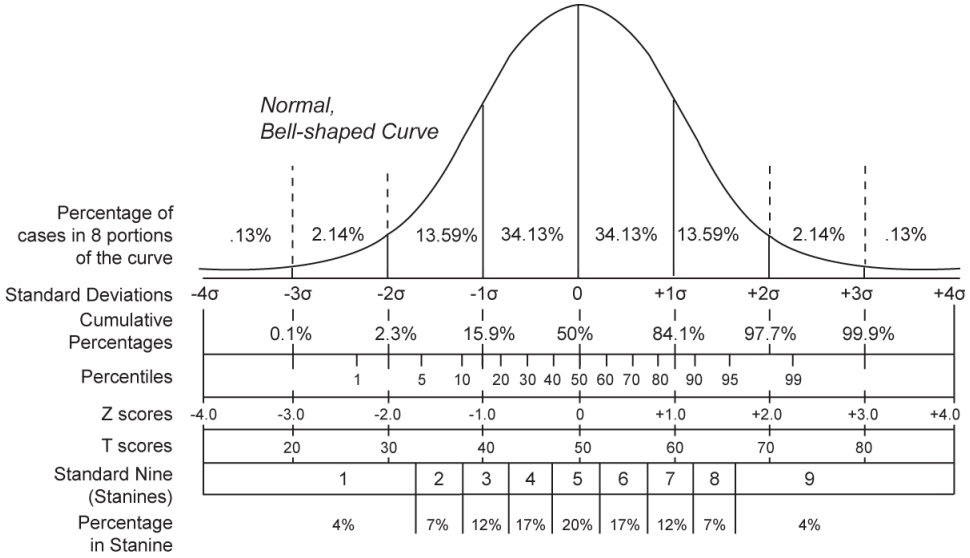


Figure 2. Normal distribution.

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

In a truly normal distribution, you can use the mean and the standard deviation to compute the amount of your data that lies between two values, or the amount of data that lies above or below a certain point (see Figure 2).

A normal distribution represents all of the possible values for a random variable, and their probabilities. If you were to add up the probabilities of all of the possible values, you would find that they add up to 1 (or 100%): the area enclosed by the curve, or under the curve, represents 1 or 100% (see Figure 2), since the sum of all the probabilities has to add up to 1. This means that if you choose a point along the distribution, the probability of all possible values above or below that value must be a fraction of 100% (or a fraction of 1). Likewise, the probability of all possible values between two points is also a fraction of 100%.
Since the area under the curve is always the same, distributions with large values of σ will have smaller values of the peak probability, the center value (called the *amplitude*) while distributions

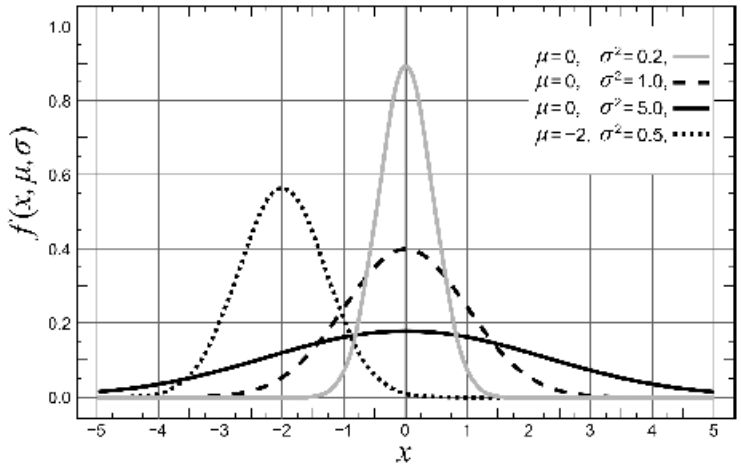


Figure 3. Normal distributions for different standard deviations.

with small values of σ have large amplitudes (see Figure 3).

The quantities that define the normal distribution are then (see Figure 2 and equation 2):

(i) the *mean*, μ;

(ii) the *dispersion* or standard deviation σ

(iii) the *amplitude* $\dfrac{1}{\sigma\sqrt{2\pi}}$
This is the highest value of the probability curve, its value when $x = \mu$.
Proving this is a bit beyond what's taught in this course and requires calculus. Note that the amplitude is smaller if the dispersion is larger.

3 The z-score

Let's assume that the data in our example are normally distributed and you want to know how many data points lie below age 55. First, you would calculate a **z-score**, a representation of how many standard deviations above or below the mean a certain data point lies. The z-score is defined as:

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

(note again that you have to know the true population mean, μ, and standard deviation, σ, to calculate the z-score) The z-score is a measure of how likely a given measured data point is, i.e. what is the probability of measuring this value *given that* it is drawn from a population of mean μ and dispersion σ (Bayes again! Scientists are always asking questions like: given that my theory for the structure of the universe predicts a certain formula, for example that the distances between galaxies has a certain average value and standard deviation, how probable is this value for the

distance between two galaxies that I've just measured?). By extension, the z-score for a given data point can also be used to measure the probability that that data point actually belongs to the population described by μ and σ.

In our example of the presidential age distribution, assuming a normal distribution and that we sampled the whole population, the population standard deviation σ using formula (5) is 6.26, so the z-score for age 55, z_{55}, would be (55 years - 54.65 years)/(6.26 years) = 0.06. You can then look up this z-score in Table 1 to see how likely it is.

Table 1 shows the fraction of the distribution between the mean ($z = 0$) and the z-score you have computed. (**This table is also given in the textbook: Table Z in Appendix D**.) The value for a z-score of 0.06 in our table corresponds to a fraction of 0.024 (2.4%). Thus, if we want to compute the amount of data lying below age 55, we add 0.024 to 0.50, to get 0.524 or 52.4% (recall: 52% is the percentile rank of 55 years, and 55 years is the 52nd percentile). We can use 0.50 because for a normal distribution) the amount of data that lies below the mean is 50%.

Likewise, if we want to know the amount of data that lies above age 55, we subtract 0.024 from 0.50, to get 0.476 or 47.6%.

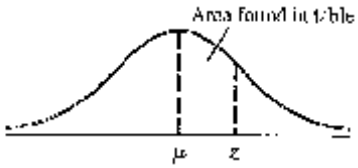
If we want to know the amount of data that lies between ages 45 and 55, we would first compute the z-scores for each value: z_{45} is -1.54, corresponding to an area of 0.438 between -1.54 and 0 (since the curve is symmetrical, we can look up the value for $z = 1.54$); z_{55} is 0.06, corresponding to an area of 0.024 between 0.06 and 0. Because 45 lies below the mean and 55 lies above the mean, we can simply add the two areas together. Thus, the space between 45 and 55 represents 0.462 or 46.2% of the data. If both values were on the same side of the mean, we would subtract these values from each other to yield a positive fraction.

4 The Binomial Distribution

Many natural phenomena are described by normal distributions. How does this distribution come about? Well, this is a bit beyond this course, but you can get an idea by going back to the binomial probability formula (Lecture 6). If you carry out n independent trials of a process with two outcomes, success and failure, the probability of getting exactly k successes is:

$$P(k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \tag{1}$$

where p is the probability of a successful outcome in a single trial. If you plot *P(k)* on the y-axis and k (from 1 to n) on the x-axis, you will see a distribution that looks similar to a normal distribution, becoming more so the larger the value of n. Remember the experiment we did by tossing coins in the first class, where we ended up plotting a histogram of the number of times a sequence of n heads (or tails) in a row was tossed as a function of n? We got a curve that was pretty close to a normal distribution. Both the normal and binomial distributions are characterized by having the same mean, median and mode; see Figure 4.



The column under A gives the area under the entire curve that is between $z = 0$ (or the mean) and a positive value of z

z	A	z	A	z	A	z	A	z	A	z	A	z	A	z	A	z	A
.00	.000	.37	.144	.74	.270	1.11	.367	1.48	.431	1.85	.468	2.22	.487	2.59	.495	2.96	.499
.01	.004	.38	.148	.75	.273	1.12	.369	1.49	.432	1.86	.469	2.23	.487	2.60	.495	2.97	.499
.02	.008	.39	.152	.76	.276	1.13	.371	1.50	.433	1.87	.469	2.24	.488	2.61	.496	2.98	.499
.03	.012	.40	.155	.77	.279	1.14	.373	1.51	.435	1.88	.470	2.25	.488	2.62	.496	2.99	.499
.04	.016	.41	.159	.78	.282	1.15	.375	1.52	.436	1.89	.471	2.26	.488	2.63	.496	3.00	.499
.05	.020	.42	.163	.79	.285	1.16	.377	1.53	.437	1.90	.471	2.27	.488	2.64	.496	3.01	.499
.06	.024	.43	.166	.80	.288	1.17	.379	1.54	.438	1.91	.472	2.28	.489	2.65	.496	3.02	.499
.07	.028	.44	.170	.81	.291	1.18	.381	1.55	.439	1.92	.473	2.29	.489	2.66	.496	3.03	.499
.08	.032	.45	.174	.82	.294	1.19	.383	1.56	.441	1.93	.473	2.30	.489	2.67	.496	3.04	.499
.09	.036	.46	.177	.83	.297	1.20	.385	1.57	.442	1.94	.474	2.31	.490	2.68	.496	3.05	.499
.10	.040	.47	.181	.84	.300	1.21	.387	1.58	.443	1.95	.474	2.32	.490	2.69	.496	3.06	.499
.11	.044	.48	.184	.85	.302	1.22	.389	1.59	.444	1.96	.475	2.33	.490	2.70	.497	3.07	.499
.12	.048	.49	.188	.86	.305	1.23	.391	1.60	.445	1.97	.476	2.34	.490	2.71	.497	3.08	.499
.13	.052	.50	.192	.87	.308	1.24	.393	1.61	.446	1.98	.476	2.35	.491	2.72	.497	3.09	.499
.14	.056	.51	.195	.88	.311	1.25	.394	1.62	.447	1.99	.477	2.36	.491	2.73	.497	3.10	.499
.15	.060	.52	.199	.89	.313	1.26	.396	1.63	.449	2.00	.477	2.37	.491	2.74	.497	3.11	.499
.16	.064	.53	.202	.90	.316	1.27	.398	1.64	.450	2.01	.478	2.38	.491	2.75	.497	3.12	.499
.17	.068	.54	.205	.91	.319	1.28	.400	1.65	.451	2.02	.478	2.39	.492	2.76	.497	3.13	.499
.18	.071	.55	.209	.92	.321	1.29	.402	1.66	.452	2.03	.479	2.40	.492	2.77	.497	3.14	.499
.19	.075	.56	.212	.93	.324	1.30	.403	1.67	.453	2.04	.479	2.41	.492	2.78	.497	3.15	.499
.20	.079	.57	.216	.94	.326	1.31	.405	1.68	.454	2.05	.480	2.42	.492	2.79	.497	3.16	.499
.21	.083	.58	.219	.95	.329	1.32	.407	1.69	.455	2.06	.480	2.43	.495	2.80	.497	3.17	.499
.22	.087	.59	.222	.96	.332	1.33	.408	1.70	.455	2.07	.481	2.44	.493	2.81	.498	3.18	.499
.23	.091	.60	.226	.97	.334	1.34	.410	1.71	.456	2.08	.481	2.45	.493	2.82	.498	3.19	.499
.24	.095	.61	.229	.98	.337	1.35	.412	1.72	.457	2.09	.482	2.46	.493	2.83	.498	3.20	.499
.25	.099	.62	.232	.99	.339	1.36	.413	1.73	.458	2.10	.482	2.47	.493	2.84	.498	3.21	.499
.26	.103	.63	.236	1.00	.341	1.37	.415	1.74	.459	2.11	.483	2.48	.493	2.85	.498	3.22	.499
.27	.106	.64	.239	1.01	.344	1.38	.416	1.75	.460	2.12	.483	2.49	.494	2.86	.498	3.23	.499
.28	.110	.65	.242	1.02	.346	1.39	.418	1.76	.461	2.13	.483	2.50	.494	2.87	.498	3.24	.499
.29	.114	.66	.245	1.03	.349	1.40	.419	1.77	.462	2.14	.484	2.51	.494	2.88	.498	3.25	.499
.30	.118	.67	.249	1.04	.351	1.41	.421	1.78	.463	2.15	.484	2.52	.494	2.89	.498	3.26	.499
.31	.122	.68	.252	1.05	.353	1.42	.422	1.79	.463	2.16	.485	2.53	.494	2.90	.498	3.27	.500
.32	.126	.69	.255	1.06	.355	1.43	.424	1.80	.464	2.17	.485	2.54	.495	2.91	.498	3.28	.500
.33	.129	.70	.258	1.07	.358	1.44	.425	1.81	.465	2.18	.485	2.55	.495	2.92	.498	3.29	.500
.34	.133	.71	.261	1.08	.360	1.45	.427	1.82	.466	2.19	.486	2.56	.495	2.93	.498	3.30	.500
.35	.137	.72	.264	1.09	.362	1.46	.428	1.83	.466	2.20	.486	2.57	.495	2.94	.498	3.31	.500
.36	.141	.73	.267	1.10	.364	1.47	.429	1.84	.467	2.21	.487	2.58	.495	2.95	.498	3.32	.500

Table 1. z scores

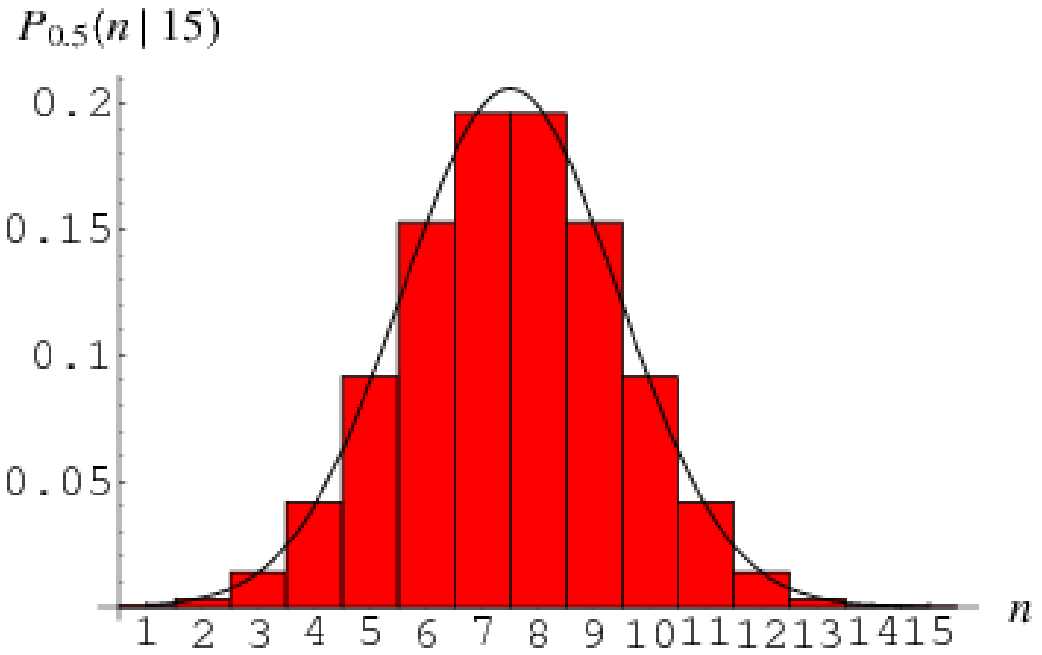


Figure 4. Binomial distribution (histogram) and normal distribution (smooth curve).

5 Other distributions

Finally, let's discuss cases in which the distribution does not have an equal mean, median, and mode. If the majority of presidents began to be inaugurated in their 30s and early 40s, and only very few in their 50s and 60s, it would create a considerable tail to the right of the most frequent value. This distribution would be called a **right skewed or positively skewed distribution**. We would expect both the mean and the median to be pushed to the right of the most frequent value, the mode, and we would expect the mean to be larger than the median (the mean is the measure of central tendency most sensitive to skewness; see Figure 5). On the other hand, if the majority of presidents began to be inaugurated in their 60s and late 50s, and only very few in their 30s and 40s, it would create a considerable tail to the left of the most frequent value. This distribution is called a **left skewed or negatively skewed distribution**. We would expect both the mean and the median to be pushed to the left of the most frequent value, the mode, and we would expect the mean to be smaller than the median (see Figure 4).

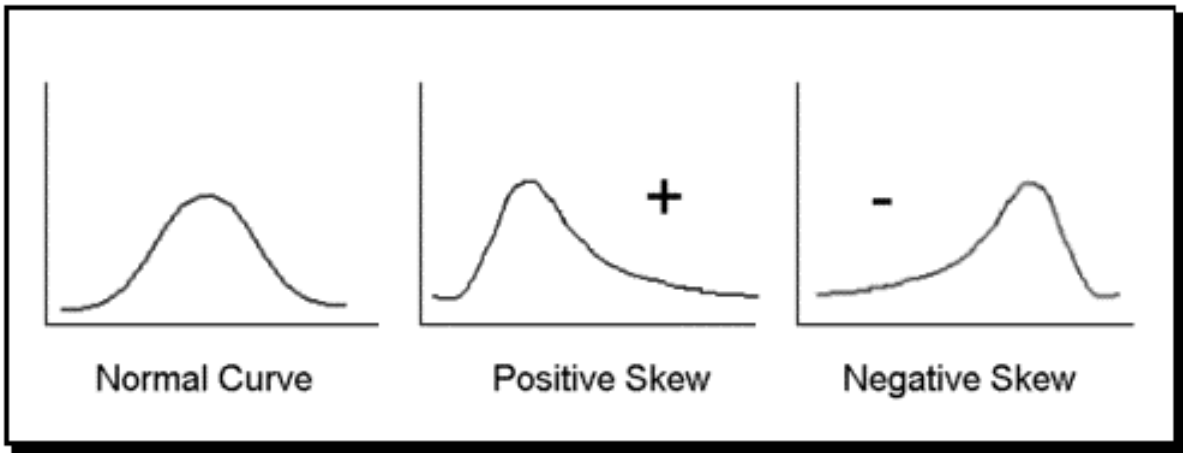


Figure 5. Left-and right-skewed distributions

There are a few other common distributions. If the probability of any value occurring is nonzero and equal to the full range of possible values, this is a **rectangular (or uniform) distribution** - see Figure 6.

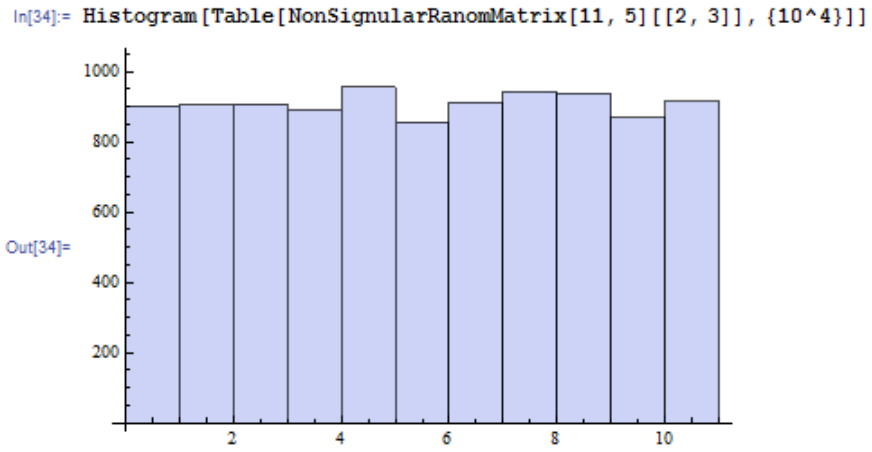


Figure 6. Uniform distribution

If the probability of a value occurring is greater than that of all values smaller than that value – in other words, the probability rises as the value rises — this is a **J-shaped distribution**. If the probability of a value occurring is smaller than that of all values smaller than that value — in other words, the probability falls as the value rises – this is also a **J-shaped distribution** – see Figure 7.

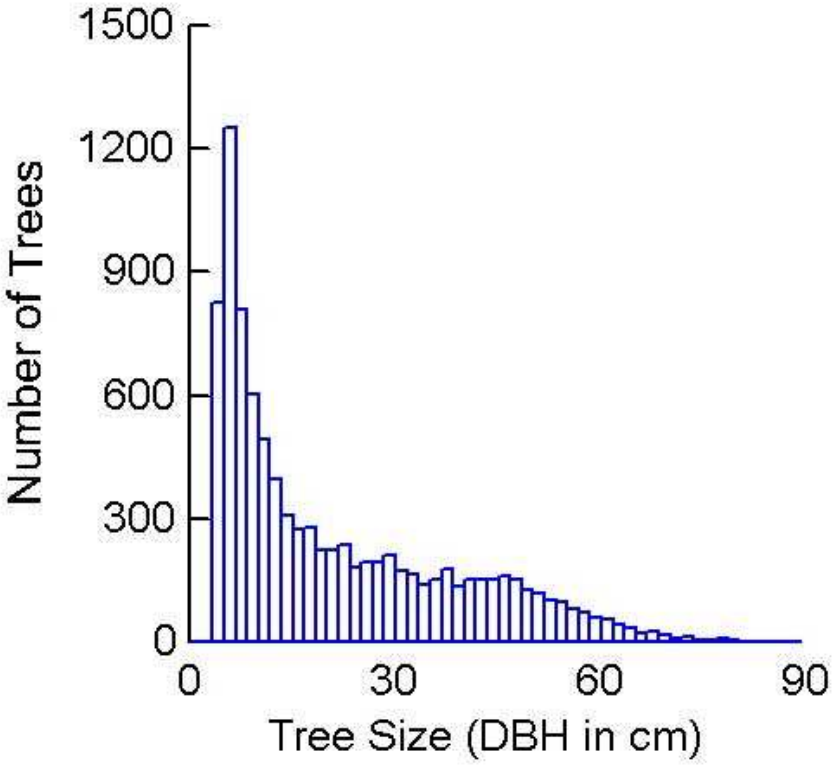


Figure 7. J- distribution

Finally, if the data set has two modes, the distribution is **bimodal**, as we saw in the lecture 8.