Mathematical Concepts — Lecture notes
MAT 117, Fall 2021 — D. Ivanšić

12.1 Sampling, Frequency Distributions

Statistics attempts to give us information about how often some feature happens in a large group of people or items, called a **population**.

Examples. We may like to know heights of males or females ages 15–25, what percentage of people are buying different brands of tablet computers, how favorable is the opinion of people toward a certain candidate for office, how long do various types of plants live, how successful various business are depending on their location, etc.

Generally, it is inconvenient to examine every member of the population, since typically there are too many in a population. Instead, we use **samples** of a population, subgroups of the population that we think will be representative of the population as a whole. Thus, it is good to use a **random sample**, one where every member of the population has an equal chance of being selected for the sample.

Example. Suppose the city of Murray is considering voting on whether to allow wine sales in grocery stores. We would like to anticipate the outcome of the vote, so we survey some Murray residents about their position towards wine sales in grocery stores. Are the following good **representative samples** of the whole population?

- Standing in front of Waterfield library, you interview random passers-by?
- You go to a retirement home in Murray and survey the people assembled for tea?
- You survey shoppers entering Wal-Mart.
- You subdivide the city into neighborhoods and survey random residents from each?

Thus, to get an accurate picture of the population as a whole, it is important to choose a sampling method well.

Once we have surveyed a sample, we wish to organize the data in a comprehensible way.

Example. The list below shows the final grades of 60 students in a college algebra class.

C, A, B, E, E, C, A, B, D, C, E, A, A, B, A, C, A, D, E, C, B, D, A, E, A, A, B, D, C, C, A, E, E, E, A, D, A, E, B, C, D, A, A, B, B, D, E, D, E, B, A, C, A, E, C, D, A, A, E, B

a) Construct a frequency distribution for the data.
b) Draw a histogram (bar graph) for the data and a frequency polygon.

Grade	Frequency
A	
B	
C	
D	
E	

Example. A "Consumer Reports" survey found the following yearly costs (in dollars) of running various refrigerators. It is not very informative to write the number of fridges that fall under every cost (it is small), so we subdivide the data into groups called **classes**.
a) Write the grouped frequency distribution table by dividing the data into classes of same width, where the first class is 33–38.
b) Draw a histogram representing the data.

34, 39, 35, 39, 36, 44, 40, 44, 45, 40, 44, 44, 41, 41,
40, 46, 60, 49, 53, 50, 52, 49, 53, 53, 61, 52, 60, 58, 55

Read about deceptions in visual displays of data in the book.

Mathematical Concepts — Lecture notes
MAT 117, Fall 2021 — D. Ivanšić

12.2 Measures of Central Tendency

We often try to capture data from a frequency distribution by using a single number that summarizes the data, usually referred to as an "average" (although this term frequently abused).

Examples. GPA, average monthly electric bill, median income in a city, etc.

The following are most often used as "averages":

midrange $= \dfrac{\text{lowest data value} + \text{highest data value}}{2}$

mean $= \overline{x} = \dfrac{\Sigma_i x_i}{n} = \dfrac{\Sigma_i x_i f_i}{n}$

median = the middle number of the data set when it is ordered in increasing order

Example. The number of people watching prime-time television during an average minute is given for each day of a week for a certain week.
a) Find the midrange of the data.
b) Find the median of the data.
c) Find the mean of the data.

Day	Viewers (mil)
Mon	95.3
Tue	94.9
Wed	93.2
Thu	93.7
Fri	80.9
Sat	78.9
Sun	91.6

Example. In the table below, letter grades from a previous example have been replaced by numbers.
a) Find the midrange of the grades.
b) Find the mode of the grades.
c) Find the median of the grades.
d) Find the mean of the grades.

Grade	Frequency
4	18
3	10
2	10
1	9
0	13

Example. The number of years of service, rounded to the nearest integer, of all Supreme Court justices who retired before 2021 is shown in the table. Using this information, estimate the mean years of service.

Years of service	Frequency	Representative value
0–5	15	
6–10	20	
11–15	16	
16–20	16	
21–25	13	
26–30	13	
31–35	12	
36–40	1	

Mathematical Concepts — Lecture notes
MAT 117, Fall 2021 — D. Ivanšić

12.3 Measures of Dispersion

Example. Students on spring break arriving at a resort were delighted to hear that the mean age of guests at the resort is 21.1. But they later found out the resort had 33 37-year-old parents and 40 8-year-old children. Show that the mean stated above is correct.

This example illustrates how the mean $\overline{x}$ often paints an incomplete picture. Data could vary from the mean by a lot, i.e. it could be spread away from the mean. To better understand data, we need measures of spread:

Range = highest value – smallest value

Standard deviation: $s = \sqrt{\dfrac{\Sigma_i (x_i - \overline{x})^2}{n-1}}$

Example. The number of people watching prime-time television during an average minute is given for each day of a week for a certain week.
Recall that the mean is $\overline{x} = 89.785714$.
a) Find the range of the data.
b) Find the standard deviation of the data.

Day	Viewers (mil)
Mon	95.3
Tue	94.9
Wed	93.2
Thu	93.7
Fri	80.9
Sat	78.9
Sun	91.6

Example. Grade distributions from two classes are listed below.
a) Find the mean for each distribution.
b) Draw a histogram for each distribution - in which one is the data more spread away from the mean? Make a guess which standard deviation will be larger.
c) Compute the standard deviations for each class and verify your guess.

Grade	Frequency
4	18
3	10
2	10
1	9
0	13

Grade	Frequency
4	10
3	18
2	12
1	13
0	7

12.4 The Normal Distribution

Example. Height measurement of 1 million women gave a histogram like below. The height of bars is percentage ("relative frequency") rather than frequency. The *areas* of rectangles are not equal to relative frequency, but we can make them so by scaling the heights of the rectangles.

Now, by making classes smaller, we can get a finer picture of the data and adjust heights of rectangles again so that the *area* of the rectangles represents the relative frequency.

If we continue the process, the jagged top edges of the rectangles start to look like a curve. This curve is called the *distribution curve*, whose chief property is:

area under the curve **between two points**	=	**percentage of sample** **that falls between those points**

Distribution curves have varying shapes:

The most commonly occuring distribution curve is the *normal distribution curve*, or *bell curve*. Many physical measurements are distributed in this way: heights, weights, exam scores, product life lengths, etc.

Properties of the normal distribution:

Curve is symmetric around $\overline{x}$.

Mean and median are equal.

Area under curve is 1, so the areas
under the left and the right half are 0.5.

Mean and median are equal.

The 68-95-99.7 rule:

Approximately 68% of data items fall within
1 standard deviation of the mean.

Approximately 95% of data items fall within
2 standard deviations of the mean.

Approximately 99.7% of data items fall within
3 standard deviation of the mean.

Example. The heights of women are normally distributed with mean $\overline{x} = 65$ in and standard deviation $s = 3.5$ in.

2 standard deviations above the mean is:

1.5 standard deviation below the mean is:

Percentage of women with height between 61.5 and 72 inches is:

Percentage of women with height below 61.5 inches is:

Percentage of women with height above 72 inches is:

Definition. The z-score of a data item in a normal distribution is given by

$$z = \frac{x - \overline{x}}{s}$$

and it describes how many standard deviations above or below the mean the data item is.

Example. Find the z-scores of data items in the previous height distribution, $\overline{x} = 65$ in, $s = 3.5$ in.

z-score of 70 is:

z-score of 61 is:

Other than by a z-score, the position of a data item can also be given by its *percentile.*

Definition. A data item is in the n-th percentile if $n\%$ of items are below it.

Note: The median is in the 50th percentile.

Example. In the previous height distribution, $\overline{x} = 65$ in, $s = 3.5$ in, find the percentile of data item 68.5, and interpret what this means.

Suppose we were trying to determine which percentage of the population of Kentucky owns a dog, and we carry out a survey with a random sample of 500 households that says 62% of households own a dog.

If we did many surveys with a random sample of 500 households, we will probably get different percentages of dog ownership, but they would not be too far from each other. The true percentage $p\%$ of the households that own a dog would be somewhere near the ones we have obtained with surveys.

Mathematical theory says that survey results are normally distributed with mean $p\%$ and standard deviation $\frac{1}{2\sqrt{n}} \times 100\%$. Using the 68-95-99.7 rule, this means that there is a 95% probability that a single sample is within 2 standard deviations of the true percentage of households owning a dog. Since two standard deviations are $2 \cdot \frac{1}{2\sqrt{n}} = \frac{1}{\sqrt{n}} \times 100\%$, this means that there is a 95% probability (confidence) that the true percentage is within $\frac{1}{\sqrt{n}} \times 100\%$ of the survey result. This number is called the *margin of error*.

Example. Compute the margin of error for the above survey and state what this means.

Note: For 5% of surveys, the true percentage will be more than the margin of error away from the sample percentage. While this is small, it still happens once out of twenty times, on average.

12.5 Problem Solving with The Normal Distribution

Example. Scores on the verbal portion of an SAT exam have an approximately normal distribution with $\overline{x} = 505$, $s = 111$. Draw pictures for this normal distribution as well as the one for z-scores as you solve the problems. Use the table at the end of notes to find the probability values.

a) What percentile is the score 550?

b) What percentage of students scored between 520 and 620?

c) What is the probability a randomly selected student scored higher than 440?

d) What percentage of students scored between 300 and 400?

Table entry for *z* is the area under the standard normal curve to the left of *z*.

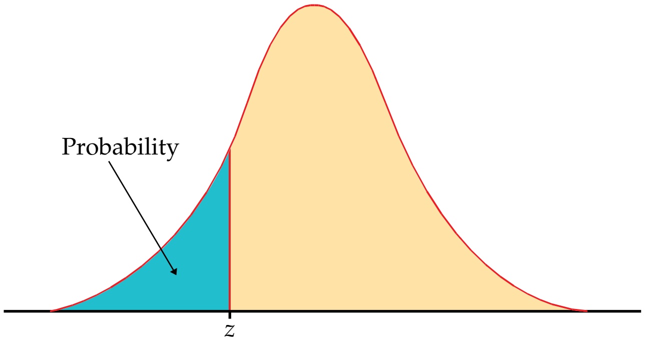

TABLE A

Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

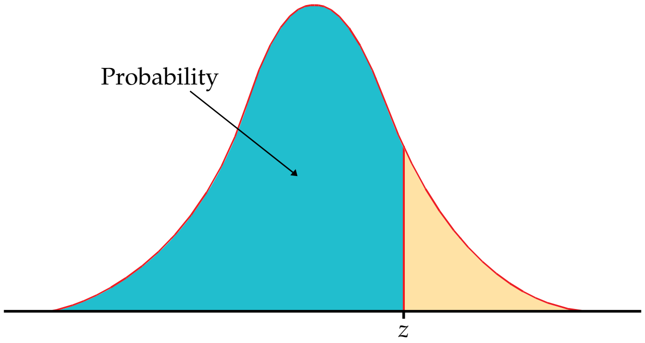

Table entry for *z* is the area under the standard normal curve to the left of *z*.

TABLE A

Standard normal probabilities (continued)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998