Lecture 05 Part 01 - Standard Scores, the Normal Distribution,

Data Analysis in CJ (CJUS 6103)

Outline

Today we will be talking about a few things, chief among them are:

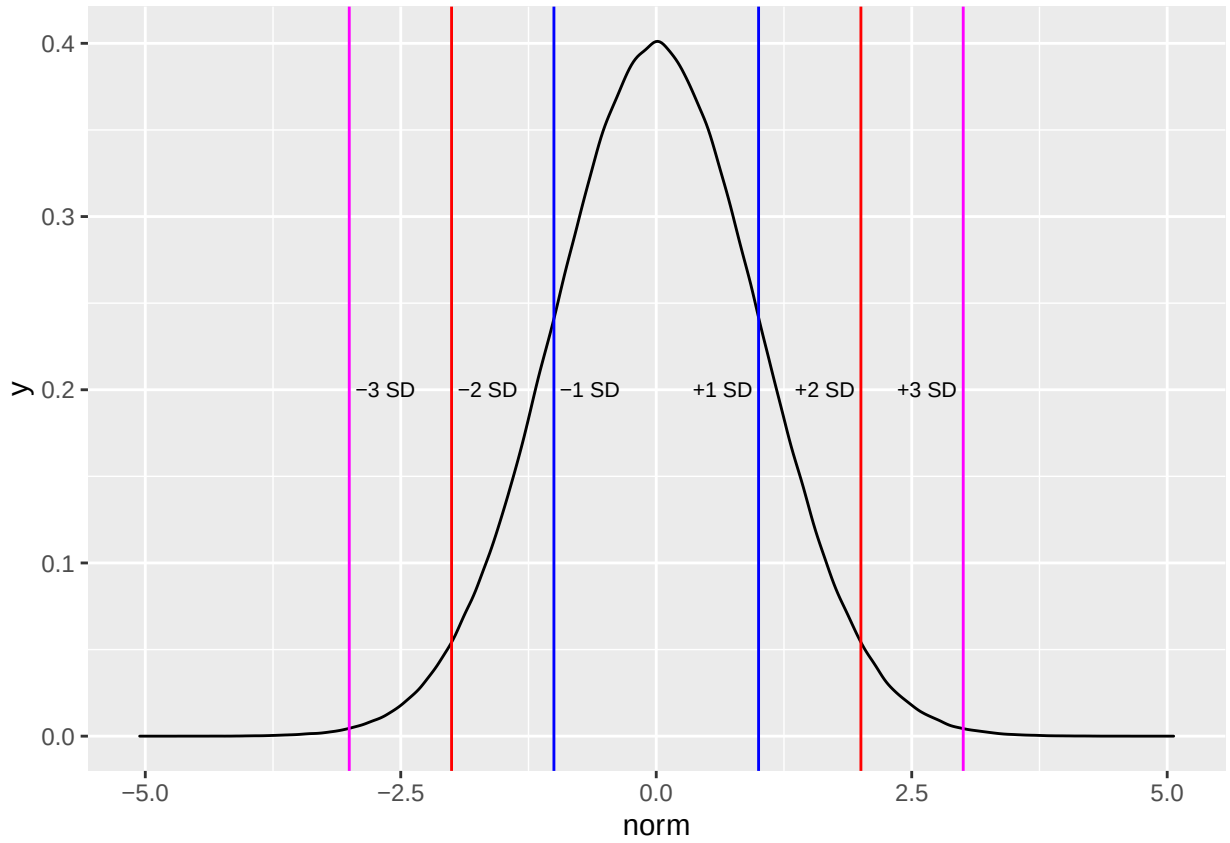
I. Characteristics of the Normal Distribution
II. Standard scores and the Normal Distribution
III. Sample, population, and sampling distributions
IV. Point and interval estimation
V. Logic of single sample inference

Characteristics of the Normal Distribution

The normal distribution, as is the case with any probability distribution, is defined by two parameters: μ (population mean) determines the location of the distribution and σ (population standard deviation) determines the shape of the distribution.

The normal distribution has three important features.

1. Symmetrical – a vertical line through the center of the distribution (at μ) produces equal halves.
2. Unimodal – only one peak to the distribution (i.e., one mode).
3. Constant area – no matter the shape of the distribution (i.e., regardless of its standard deviation), there is a constant probability under the curve between the mean and any given distance from the mean measured in standard deviation units.



This last feature is an especially important one. We know that the area that lies under the normal curve contains 100 percent of all cases, and we know that the vertical line representing the mean splits the distribution into two equal halves. We also know that 68.26 (2×.3413) percent of all cases lie within one standard deviation of the mean; 95.44 (2×.4772) percent of all cases lie within two standard deviations of the mean; and 99.74 (2×.4987) percent of all cases lie within three standard deviations of the mean.

Standard Scores and the Normal Distribution

One disadvantage of the normal distribution is that no two are exactly alike. This is because each distribution is defined by its mean and standard deviation. Thus, it is difficult to compare individual scores within and across distributions. Lucky for us, all we have to do is take a simple transformation that will allow us to do this. We have to standardize our data in this way:

$$z = \frac{x - \mu}{\sigma}$$

This formula produces a z-score for individual cases in a distribution. The resulting z-score tells us the number of standard deviation units that a particular case lies from the mean. The sign tells us whether a case lies above or below the mean, and the magnitude tells us how many standard deviation units lie between this observation and the mean. Standardizing scores within any distribution results in the standard normal distribution, or the z-distribution, with mean 0 and standard deviation 1. By standardizing, we are creating a universal (i.e., standard) metric that can be compared across different scales of measurement.

The nice thing about standard scores is that we can use a z-table to determine the probability that lies between any two values. This means that we can calculate the probability of observing events.

z-Score	Area between μ and z	Area beyond z
0.00	.0000	.5000
0.50	.1915	.3085
1.00	.3413	.1587
1.50	.4332	.0668
2.00	.4772	.0228
2.50	.4938	.0062
3.00	.4987	.0013

Another representation of the z-table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.00	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.10	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.20	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.30	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.40	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319

Standard Scores Example - Highway Patrol

Let's say that the highway patrol has collected data on the speed of cars traveling past a certain checkpoint. They find that the speed of cars traveling past the checkpoint is normally distributed, with average speed limit $\mu = 60.3$, and standard deviation $\sigma = 6.5$.

1) What is the probability of observing a car going **more** than 70 miles per hour?

$$p(x > 70) \Rightarrow \frac{x - \mu}{\sigma} = \frac{70 - 60.3}{6.5} = 1.492 \Rightarrow p(z > 1.49) = 0.068$$

2) What is the probability of observing a car going **less** than 53 miles per hour?

$$p(x < 53) \Rightarrow \frac{x-\mu}{\sigma} = \frac{53-60.3}{6.5} = -1.123 \Rightarrow p(z > -1.123) = 0.131$$

3) What is the probability of observing a car going **between** 61 and 63 miles per hour?

$$\text{x} = 61 => \frac{61-60.3}{6.5} = 0.108;\ \text{x} = 63 => \frac{63-60.3}{6.5} = 0.415$$
$$\Rightarrow p(61 < \text{x} < 63) = p(0.108 < z < 0.415) = p(z > 0.108) - p(z > 0.415) = 0.118$$

4) What is the probability of observing a car going **between** 60 and 65 miles per hour?

$$\text{x} = 60 => \frac{60-60.3}{6.5} = -0.046;\ \text{x} = 63 => \frac{65-60.3}{6.5} = 0.723$$

$$\Rightarrow p(60 < \text{x} < 65) = p(-0.046 < z < 0.723) = p(z > -0.046) - p(z > 0.723) = 0.283$$

Last answer will be slightly off (0.001) due to rounding error - using the qnorm function (demonstrated below) will fix that.

A Note on the pnorm() Function

In the above calculations I use an R function called **pnorm**. This function returns the lower or upper-tail probability of finding some particular z-score, given a particular mean and standard deviation for a distribution. Instead of relying on inexact tables, it's more accurate to calculate a probability value in R using this function.

The function takes the following general form: pnorm(x, mean=0, sd=1, lower.tail=TRUE/FALSE). Where x is the z-score you have obtained, mean=0 and sd=1 specify a standard normal distribution, and lower.tail=TRUE/FALSE tells R to return the lower (TRUE) or upper (FALSE) tail probability. Lower tail probabilities are to the LEFT of your z-score and upper tail probabilities are to the RIGHT of your z-score.

Back to Highway Patrol Example

5) We can also use algebra to work backwards from particular probabilities to obtain raw scores using the following equation:

$$x = (z * \sigma) + \mu$$

Suppose the county wants to aggressively enforce speed limits against the top 10% of speeders. Which speeds should they target to do so?

$$p(z > 1.282) = 0.1 => 1.282 = \frac{x-60.3}{6.5} => x = (1.282 * 6.5) + 60.3 = 68.63$$

Additional Example - Graduating with Honors

In order to graduate with honors, students must be in the top 2% (Summa Cum Laude), 3% (Magna Cum Laude), or 5% (Cum Laude) of their graduating class. Suppose that GPAs are normally distributed with mean $\mu = 2.60$ and standard deviation $\sigma = .65$. What GPA must a student have to graduate with each of these three honors?

Summa Cum Laude: $p(z > 2.054) = 0.02 => 2.054 = \frac{\text{x} - 2.60}{.65} => \text{x} = (2.054 * 0.65) + 2.60 = 3.935$

Magna Cum Laude: $p(z > 1.881) = 0.03 => 1.881 = \frac{\text{x} - 2.60}{.65} => \text{x} = (1.881 * 0.65) + 2.60 = 3.823$

Cum Laude: $p(z > 1.645) = 0.05 => 1.645 = \frac{\text{x} - 2.60}{.65} => \text{x} = (1.645 * 0.65) + 2.60 = 3.669$

A Note on the qnorm() Function

In the above calculations I use an R function called **qnorm**. This function returns a z-score that is associated with some lower/upper tail probability, given a particular mean and standard deviation for a distribution. Like the **pnorm** function, we can use the **qnorm** function to obtain more exact z-scores than we would if we were relying on z-score tables alone.

The function takes the following general form: qnorm(p, mean=0, sd=1, lower.tail=TRUE/FALSE). Where p is the probability value you are interested in, mean=0 and sd=1 specify a standard normal distribution, and lower.tail=TRUE/FALSE tells R to return a z-score that leaves that value of probability to its left (TRUE) or right (FALSE). Because the normal distribution is symmetrical, this last part simply applies the right sign (postive/negative) to the z-score. You can test this out by holding all but the lower.tail option constant and shifting between lower.tail=TRUE and lower.tail=FALSE - the absolute value of the z-score remains the same, only its sign changes.

Logic of Sampling

The use of standard scores brings us closer to being able to conduct hypothesis tests using a sample mean ($\overline{x}$) as an estimate for the population mean (μ). Our goal is to be able to use the sample mean as a **best guess** for the population mean.

Recall that one of the properties of a statistic, such as a sample mean, is that although it is empirical (i.e., it can be measured) and known (we actually collect data from a sample), it is not fixed. This means that there is variation in the value of the mean from one sample to the next. Since we know this to be the case, we cannot simply use the quantities we obtain from our sample in place of the quantities we want to estimate for the population. Rather, we have to resort to what is called a sampling distribution. A sampling distribution is simply a particular type of probability distribution.

Let's do a thought experiment. Let's say we draw a sample of size 100 from the population. We know that the sample has mean $\overline{x}$ with standard deviation s, and that the population has mean μ with standard deviation σ. The problem is that we want to know μ but cannot determine this value exactly since it is not possible to gather data on every person in the population. So we want to use our sample mean as an estimate for the population mean.

Now, imagine drawing another sample of size 100 from the same population. You can be sure that the mean of this sample will not be the same as the mean from the first sample. Continue drawing samples of size 100 from the same population, and calculate a mean for each sample. Pretty soon, we find that we are beginning to have a distribution of sample means. We keep drawing an infinite number of samples and calculating a mean for every sample. The distribution of these sample means is called a sampling distribution. This sampling distribution is important because we know that the mean of the sampling distribution (i.e., the

mean of the sample means) is μ. The standard deviation of this distribution is called a standard error, and is calculated as:

$$\frac{\sigma}{\sqrt{n}}$$

Distribution	Properties	Mean	Standard Deviation
Sample	Empirical, known	$\overline{x}$	s
Population	Empirical, unknown	μ_x	σ_x
Sampling	Theoretical, known	$\mu_{\overline{x}} = \mu_x$	$\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$

So why is the sampling distribution important to us? The reason is that the sampling distribution provides a necessary link between our sample and the population that we are trying to know something about. We can imagine our sample mean as but one mean from a theoretical distribution of all possible sample means.

With a sampling distribution, we can then use the laws of probability to determine the probability of obtaining our particular sample mean. One important property of a sampling distribution is that as the sample size increases, the standard error decreases, meaning that we have greater precision in estimating the true value of μ when n gets larger and larger. In other words, the probability distribution gets tighter around the mean.

Another property of a sampling distribution is that it is always normally distributed if the population characteristic is normally distributed. This means that we can take advantage of the properties of the normal distribution to determine probabilities. The obvious problem with this is that few variables are normally distributed in the population, and even fewer that are of interest to criminologists. But all is not lost.

Central Limit Theorem

Central limit theorem – if an infinite number of samples of size n are drawn from the population, the sampling distribution will approach normality as the sample size becomes infinitely large, even if the characteristic is not normally distributed in the population.

This theorem is essential for criminologists, since it implies that even though a variable like *number of arrests* is highly skewed in the population, we can still assume that the sampling distribution of sample means will be normal when we have a large sample. This means that we can still employ the standard normal probability distribution to conduct hypothesis tests and estimate confidence intervals.

Standard Score for a Sample Mean

The sampling distribution is a probability (i.e., theoretical) distribution that forms a necessary link between the observed sample statistic and the unknown population parameter. In other words, it is essential for statistical inference. Even though we do not observe this distribution, we know its properties, specifically its mean and standard deviation. Moreover, the central limit theorem tells us that with large n, the sampling distribution is approximately normal, even though the characteristic is not normally distributed in the population. The advantage to normality is that we can resort to using the standard normal distribution to estimate probabilities.

Suppose that we have a sample mean, and want to compare it to a population mean. Just like we cannot compute probabilities when we are dealing with raw scores, we cannot compute probabilities when we are dealing with sample means. We need to transform a mean and standard deviation from different samples into a common metric. We thus rely on the standard normal or z-distribution. Recall the z-score formula that we have seen so far.

$$z = \frac{x - \mu}{\sigma}$$

Notice that with this formula, we are standardizing a raw score with respect to a mean in standard deviation units from a population. We want to retain the same logic for the present problem, but instead use a sample-mean analog to this raw score formula. The first case is when σ is known, and the second is when σ is unknown.

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

Notice here that we are standardizing a sample mean with respect to a (hypothesized) population mean in standard error units. The z-score formula is thus general. The specific formula that we use depends on what distribution we are dealing with: sample, population, or sampling; and in the latter case on whether the population standard deviation is known.

New Examples - Household Income

Suppose that household income in the U.S. has mean \$32,000 and standard deviation \$5,000. Answer the following probability questions for a sample of 30 individuals.

$p(\overline{x} < \$30,000) = p\Big(z < \frac{30000-32000}{5000/\sqrt{30}}\Big) = p(z < -2.191) = 0.014$

$p(\overline{x} > \$33,000) = p\Big(z < \frac{33000-32000}{5000/\sqrt{30}}\Big) = p(z > 1.095) = 0.137$

$p(31500 < \overline{x} < 32500) = p\Big(\frac{31500-32000}{5000/\sqrt{30}} < z < \frac{32500-32000}{5000/\sqrt{30}}\Big) = p(-0.548 < z < 0.548) = 0.416$