Statistical Inference and Hypothesis Testing in Data Science Applications

DTSA 5003 offered on Coursera

by the University of Colorado, Boulder

Instructor: J.N. Corcoran

The Normal Distribution

The continuous random variable X is said to have a **normal distribution** with mean μ and variance σ^2 if X has probability density function

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

for $-\infty < x < \infty$. Here μ and σ^2 are parameters with $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

The symbols μ and σ^2 are typically reserved for the mean and variance of a distribution. Fortunately this works out well here since the mean of this distribution (also known as the expected value of X) is defined and denoted as

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

If you were to work this out for the given normal distribution pdf, you will indeed get μ.

The variance of the distribution is defined and denoted as

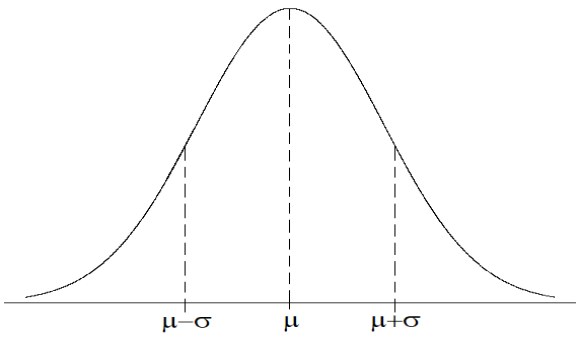
$$Var[X] = \mathsf{E}[(X-\mu)^2] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2.$$

One could work out

$$\mathsf{E}[X^2] = \int_{-\infty}^{\infty} x^2\, f(x)\, dx$$

and put everything together to show that $Var[X] = \sigma^2$.

The plot of this pdf looks like



With just a little calculus, one can see that the points $\mu \pm \sigma$ are where the pdf changes concavity. σ and the square root of the variance σ^2, is known as the *standard deviation* of the distribution. It is routine (though maybe not easy) to show by integration that 99.7% of the area under the pdf is within 3 standard deviations of the mean μ.

If X is a random variable with this distribution, we write

$$X \sim N(\mu, \sigma^2).$$

Linear Combinations of Normal Random Variables are Normal

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. This means that they are independent and identically distributed and we write

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2).$$

Using moment generating functions, one can show (see DTSA 5002 and possibly DTSA 5001) that, for any constants $a_1, a_2, \ldots, a_n$ (that are not all zero) and any constant b that

$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b$$

has a normal distribution. That is to say that any linear combination of normal random variables is again normal. (For the record, it is also true that (almost) any linear combination of normal random variables is again normal even if the X_i are not iid. We say "almost" to avoid situations like $X_2 = -X_1$ for which $X_1 + X_2 = 0$ is clearly not a normal random variable. In this course we will only work with the iid case.)

What is the mean and variance of the given linear combination? Since expectation is a linear operator, we have

$$\mathsf{E}\left[\sum_{i=1}^{n} a_i X_i + b\right] = \mathsf{E}\left[\sum_{i=1}^{n} a_i X_i\right] + b = \sum_{i=1}^{n} \mathsf{E}[a_i X_i] + b = \sum_{i=1}^{n} a_i \mathsf{E}[X_i] + b = \sum_{i=1}^{n} a_i \mu + b = \mu \sum_{i=1}^{n} a_i + b.$$

To run the sum through a variance, we need independence of the X_i:

$$Var\left[\sum_{i=1}^{n} a_i X_i + b\right] = Var\left[\sum_{i=1}^{n} a_i X_i\right] \overset{indep}{=} \sum_{i=1}^{n} Var[a_i X_i] = \sum_{i=1}^{n} a_i^2 Var[X_i] = \sum_{i=1}^{n} a_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} a_i^2.$$

Thus, we have that

$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \sim N\left(\mu \sum_{i=1}^{n} a_i + b, \sigma^2 \sum_{i=1}^{n} a_i^2\right).$$

A linear combination of particular importance to us is the sample mean:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n).$$

Note that the sample mean has the same mean as the original X_i but a smaller variance. Why does this make sense? Simply put, averages do not vary as much as individual values. Suppose there are 2,000 Calculus I students in a given semester at a particular university. If we consider all scores for the first midterm, these grades may vary quite a bit, from single digit grades to a full 100 points. On the other hand, if the students are grouped into 50 recitation sections each of size 40 and we considered the average score for each section, it is much less likely to see average values down by 0 and up by 100. You would not expect these sample means to vary as much as the individual scores.

The Standard Normal Distribution

The standard normal distribution is the normal distribution with mean 0 and variance 1. Statisticians typically use the letter Z for a standard normal random variable.

$$Z \sim N(0, 1).$$

(Note, however, that Z is not a very special symbol and it's possible that you'll come across a random variable Z that is not meant to have the $N(0, 1)$ distribution!)

The standard normal probability density function (pdf) is

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

for $-\infty < x < \infty$. While it is possible to show that this integrates to 1 using polar coordinates, integration over subsets of the real line can not be done in closed-form. In particular, there is no closed form expression for the cumulative distribution function (cdf) of the standard normal distribution.

While a generic cdf is usually denoted by a capital F, the standard normal cdf is typically denoted using a captial "phi":

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \, dx.$$

Again, we will not be able to evaluate $\Phi(z)$ in closed-form so we will have to use numerical approximations to the integral. These are often tabulated in a "z-table". In this course, we will use R (software) to evaluate the standard normal cdf. For example,

$$\Phi(1.8) = P(Z \leq 1.8) \approx 0.9640697$$

can be gotten by typing "pnorm(1.8)".

Standardizing and Probabilities for a Normal Random Variable

Let $X \sim N(\mu, \sigma^2)$. As a simple linear combination, we know that

$$\frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$$

has a normal distribution. Note that

$$\mathsf{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}\mathsf{E}[X-\mu] = \frac{1}{\sigma}(\mathsf{E}[X]-\mu) = \frac{1}{\sigma}(\mu-\mu) = 0.$$

Also note that

$$Var\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma^2}Var[X-\mu] = \frac{1}{\sigma^2}Var[X] = \frac{1}{\sigma^2}\cdot\sigma^2 = 1.$$

Thus, we have show that

$$X \sim N(\mu, \sigma^2) \qquad \Rightarrow \qquad \frac{X-\mu}{\sigma} \sim N(0,1).$$

This is a transformation that turns any normal random variable into s standard normal random variable. As such, we say that we are "standardizing X".

Similarly, we can go the other way to show that

$$Z \sim N(0,1) \qquad \Rightarrow \qquad \sigma Z + \mu \sim N(\mu, \sigma^2).$$

Computing Probabilities Examples:

1. Suppose that $X \sim N(4,3)$. What is $P(X \leq 4.2)$?
   We're going to standardize X symbolically on the left side of the inequality and numerically on the right side.

   $$P(X \leq 4.2) \overset{standardize}{=} P\left(\frac{X-\mu}{\sigma} \leq \frac{4.2-4}{\sqrt{3}}\right) = P\left(Z \leq \frac{4.2-4}{\sqrt{3}}\right)$$

   where $Z \sim N(0,1)$. In R this is

   ```
   > pnorm((4.2-4)/sqrt(3))
   [1] 0.5459637
   ```

   There is actually an R command to compute this probability directly without first turning X into a standard normal random variable.

   ```
   > pnorm(4.2,4,sqrt(3))
   [1] 0.5459637
   ```

   Historically though, people had to carry around printed tables with estimated probabilities and no one wanted to carry around a table for every μ and σ^2 combination! Even though we can easily call up non-standard normal probabilities in R, we will still standardize everything as it will be necessary for more theoretical problems.

2. Suppose that $Z \sim N(0,1)$. What is $P(Z \leq -1.13)$?
   Again, we can go directly to R with

```
> pnorm(-1.13)
[1] 0.1292381
```

   however, for some of the more theoretical problems in this course, it will be convenient for us to be able to use the symmetry of the $N(0,1)$ distribution about 0. We have

$$P(Z \leq -1.13) = P(Z \geq 1.13) = 1 - P(Z < 1.13) \overset{continuous}{=} 1 - P(Z \leq 1.13)$$

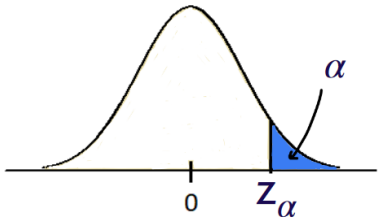
```
> 1-pnorm(1.13)
[1] 0.1292381
```

3. Suppose that $X_1, X_2, \ldots, X_5$ is a random sample from the $N(1,3)$ distribution. Write $P(\overline{X} > 0.57)$ in terms of the standard normal cdf.
   The standardization is always the same– we subtract the mean and divide by the standard deviation. Here, however, we must make sure to use the mean and standard deviation for the sample mean. These are $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. We have

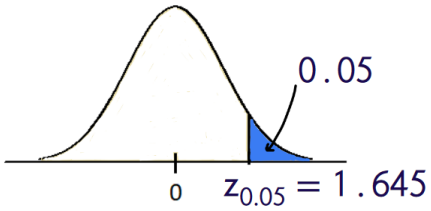
$$\begin{aligned} P(\overline{X} > 0.57) &= P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{0.57-1}{\sqrt{3}/\sqrt{5}}\right) = P\left(Z > \frac{0.57-1}{\sqrt{3}/\sqrt{5}}\right) \\ &= 1 - P\left(Z \leq \frac{0.57-1}{\sqrt{3}/\sqrt{5}}\right) = 1 - \Phi\left(\frac{0.57-1}{\sqrt{3}/\sqrt{5}}\right). \end{aligned}$$

Critical Values

Critical values for distributions are numbers that cut off specified areas under pdfs. For the $N(0,1)$ distribution, we will use the notation z_α to denote the value that cuts off area α to the right as depicted here. (Be careful when reading other references. Some authors use the notation z_α to denote the value that cuts off area α to the left!)



The R command to get a critical value that cuts off a specified area to the left is "qnorm". For example, $z_{0.05} \approx 1.645$



and this can be found using R as follows.

```
> qnorm(0.95)
[1] 1.644854
```