

## 1. Random Numbers and Probability Distributions



### Let's roll some dice

Probability lessons from Casino Royale

#### Probability – the frequentist approach

- Probability is a measure between zero and one of the likelihood that an event might occur.
  - An event could be the likelihood of a stock market falling below or rising above a certain threshold.
- You are familiar with the weather forecast that often describes the likelihood of rainfall in terms of probability or chance.
  - You often hear the meteorologists explain that the likelihood of rainfall is, for instance, 45%. Thus, 0.45 is the probability that the event, rainfall, might occur.
- The probability associated with any outcome or event must fall in the zero and one (0–1) interval.
  - The probability of all possible outcomes must equate to one.

#### Random variables

- A random variable is a “quantity whose possible values depend, in some clearly-defined way, on a set of random events.”
  - It “is a function that maps outcomes (that is, points in a probability space).”
  - Rolling two dice can have one of 36 outcomes where each outcome could be considered a random outcome.
- A probability distribution is essentially a theoretical model depicting the possible values a random variable may assume along with the probability of occurrence.
  - We can define probability distributions for both discrete and continuous random variables.

Now, let us visit some basic definitions about probability, as it relates to the most commonly used concepts in statistics.

Essentially, **probability** is a measure between zero and one for the likelihood that something or some event might occur. For instance, you may hear that the stock markets, the chance of stock market's rising above some point, or falling below some point is x%, or you may hear that the chance for rain is 45% tonight. These are all coming from this very concept of probability. Essentially, probability is a measure between zero and one, so 45% would be 0.45, the discussion about probability is not complete without a discussion about random variables.

Essentially, **random variable** is a quantity whose possible values depend in some clearly defined way on a set of some random events. It's a function that maps out outcomes, that is, points in a probability space. So, probability space essentially is all possible outcomes, If you roll a die, it can have one out of six outcomes, so that's the probability space there. If you roll two dice, you can have one out of 36 outcomes where each outcome could be considered a random outcome.

And **probability distribution** is a theoretical model that depicts the possible values any random variable may assume along with the probability of its occurrence.

#### Casino Royale: Roll the Dice

A die has six faces, so rolling two dice can assume one of the 36 discrete outcomes:

- Each die can assume one of the six outcomes in a roll. Hence rolling two dice together will return one out of 36 outcomes.
- If one (1) comes up on each die, the outcome will be  $1 + 1 = 2$ , and the probability associated with this outcome is one out of thirty-six ( $1/36$ ) because no other combination of the two dice will return two (2).
- On the other hand, I can obtain three (3) with the roll of two dice by having either of the two dice assume one and the other assuming two and vice versa.
- Thus, the probability of an outcome of three with a roll of two dice is 2 out of 36 ( $2/36$ ).

#### 36 ways to feel lucky

|       | Column-1 | Column-2 | Column-3 | Column-4 | Column-5 | Column-6 |
|-------|----------|----------|----------|----------|----------|----------|
| Row-1 |          |          |          |          |          |          |
| Row-2 |          |          |          |          |          |          |
| Row-3 |          |          |          |          |          |          |
| Row-4 |          |          |          |          |          |          |
| Row-5 |          |          |          |          |          |          |
| Row-6 |          |          |          |          |          |          |

Figure 6.2 All possible outcomes of rolling two dice

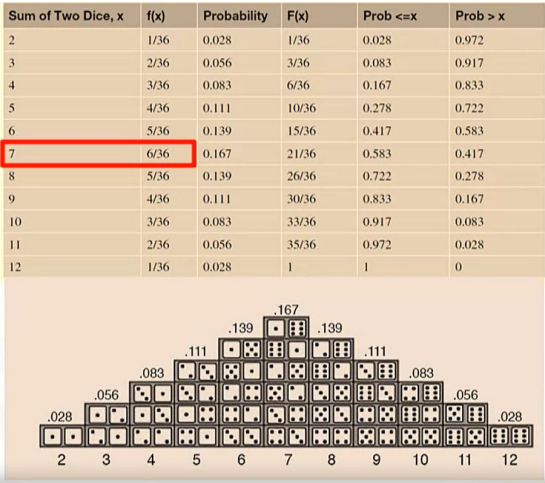
Source: <http://www.edcollins.com/backgammon/diceprob.htm>

We'll define this more with examples using two dice. So, consider two dice, a die has six faces, and if you roll two dice, it can assume one out of 36 discrete outcomes.

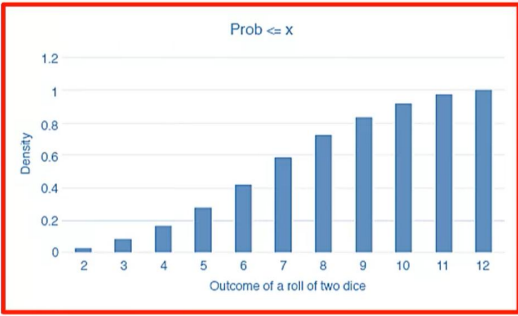
So, if you were to roll two dice, the probability that both die will one as the outcome will be  $1 + 1 = 2$ , and there's only one possibility of getting that and that's one out of 36. So, here we have two die, one black and one white, and if you were to look at the possibility of getting one on black and two on white and that's one outcome. So, that's  $1 + 2 = 3$  or you can have two on black and one on white, that's  $2 + 1 = 3$  again. So, there are two ways of getting three by rolling two dice, so the outcome or the probability is two out of 36 possible outcomes that are mapped out here.

总结:

Lucky seven



Probability of less than equal to



| Sum of Two Dice, x | f(x) | Probability | F(x)  | Prob <=x | Prob > x |
|--------------------|------|-------------|-------|----------|----------|
| 2                  | 1/36 | 0.028       | 1/36  | 0.028    | 0.972    |
| 3                  | 2/36 | 0.056       | 3/36  | 0.083    | 0.917    |
| 4                  | 3/36 | 0.083       | 6/36  | 0.167    | 0.833    |
| 5                  | 4/36 | 0.111       | 10/36 | 0.278    | 0.722    |
| 6                  | 5/36 | 0.139       | 15/36 | 0.417    | 0.583    |
| 7                  | 6/36 | 0.167       | 21/36 | 0.583    | 0.417    |
| 8                  | 5/36 | 0.139       | 26/36 | 0.722    | 0.278    |
| 9                  | 4/36 | 0.111       | 30/36 | 0.833    | 0.167    |
| 10                 | 3/36 | 0.083       | 33/36 | 0.917    | 0.083    |
| 11                 | 2/36 | 0.056       | 35/36 | 0.972    | 0.028    |
| 12                 | 1/36 | 0.028       | 1     | 1        | 0        |

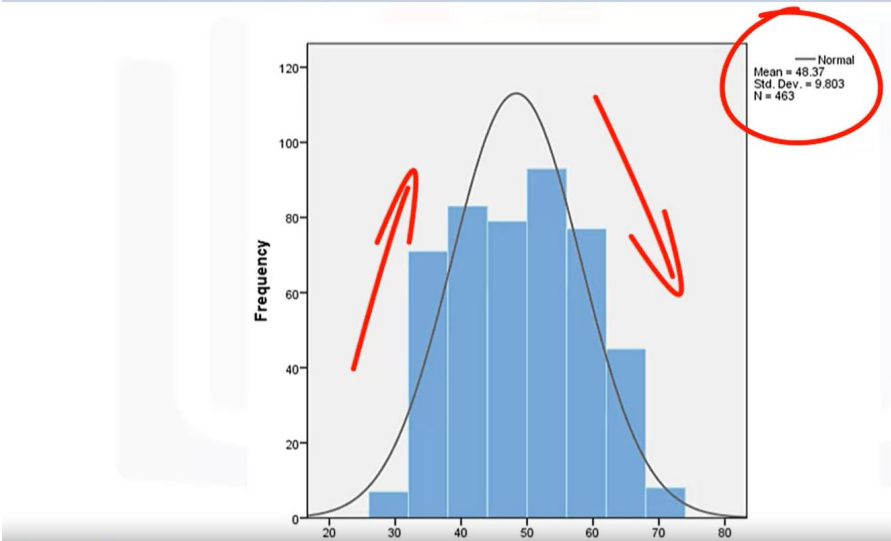
So, if you think about the sum of two dice being two, there's only one possibility out of 36. Getting a three, you have two possibilities, getting a four, you have three possibilities, getting a five, you have four possibilities and so on so forth.

The maximum most frequently possible number to have as a sum of two dice is 7, and the probability is 6 out of 36, which is 0.167.

And if you were to sum these probabilities up and that is 0.028 + 0.056, you get 0.084. So, if you sum up all these, they all sum up to one and the probability of getting six or greater than six is 0.58 or getting less than equal to six is 0.417. You sum these up, it's .1028 + 0.972 is 1, this plus this is 1, this plus this is 1 and obviously 1 + 0 is 1. The probability of getting 12, that is, both die show six is 1 out of 36 possible outcomes which is 0.028. So, the probability sums up to 1 and the probability of getting more than 12 is obviously 0 because the two dice can maximum produce this number.

So, nice way of looking at the way a probability distribution space is created by rolling two dice. And if I were to look at the probability of say, getting some number or less than some number, that's called the **cumulative distribution** function. If you were to simply chart this probability outcomes in a chart, you get this graph here.

Normally aging



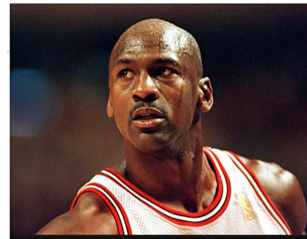
So, here we have age as our variable and I created a histogram of age, and then using the mean value of 48.37 which is the mean average age and a standard deviation of 9.8 years. I can fit a theoretical normal distribution with these two parameters.

## 2. State your hypothesis

# State Your Hypothesis

Murtaza Haider and Aije Egwaikhide

© IBM Corporation. All rights reserved.



Michael Jordan



Wilt Chamberlain

## The Basketball Giants

In this video, I will illustrate how to state your hypothesis when you're comparing the averages between two entity. We will use basketball as an example, using Michael Jordan and Wilt Chamberlain, the two highest scorers in the history of basketball, as examples.

### Jordan versus Chamberlain

Jordan's 30.12 points per game

Chamberlain's 30.06 points per game

If you were to recall, you would know that Michael Jordan, on average, scored 30.12 point in each game, and Chamberlain averages around 30.06 points per game. If you were to compare the two averages between Michael Jordan and Chamberlain, and even though they are very similar looking numbers, we need to set up a statistical hypothesis testing.

We are interested in comparing the average points scored by the two basketball players, and the comparison of means or averages is available in three flavors.

First, we can assume that the average points per game scored by the two players, Jordan and Chamberlain, are the same. That is, the difference between their mean scores is zero. If their averages are the same, average of one minus average of other should be zero, this becomes a null hypothesis. Let's say if  $\mu_j$  represents the average points per game scored by Michael Jordan, and  $\mu_c$  represents the average points per game scored by Wilt Chamberlain.

### Comparing basketball giants

- The comparison of means (averages) comes in three flavors.
- First, you can assume that the mean points per game scored by both Jordan and Chamberlain are the same.
- That is, the difference between the mean scores of the two basketball legends is zero.
  - This becomes our null hypothesis.
- Let  $\mu_j$  represent the average points per game scored by Jordan
- Let  $\mu_c$  represent the average points per game scored by Chamberlain.

#### The Null and Alternative hypotheses

- Null Hypothesis:  $H_0$
- Alternative Hypothesis:  $H_a$

$$H_0: \mu_j = \mu_c$$

The alternative hypothesis, denoted as  $H_a$ , is as follows:

$H_a: \mu_j \neq \mu_c$ ; their average scores are different.

#### Option 2: Jordan is better

- Null Hypothesis:  $H_0$
- Now let us work with a different null hypothesis and assume that Michael Jordan, on average, scored higher than Wilt Chamberlain did. Mathematically:

$$H_0: \mu_j \geq \mu_c$$

- Alternative Hypothesis:  $H_a$

$$H_a: \mu_j < \mu_c$$

#### Option 3: Chamberlain is better

- Null Hypothesis:  $H_0$
- Michael Jordan, on average, scored lower than Wilt Chamberlain did. Mathematically:

$$H_0: \mu_j \leq \mu_c$$

- Alternative Hypothesis:  $H_a$

$$H_a: \mu_j > \mu_c$$

- We can state the **null hypothesis** to be  $\mu_j$  equal  $\mu_c$ , that is the average scored by Jordan and average scored by Chamberlain are the same. The **alternative hypothesis** would be that no, these averages are not the same, they are different. The alternative hypothesis  $H_a$  compared to null hypothesis  $H_0$  or o. The alternative is that the averages are not the same, their average scores are different.
- Now, the other option, the second option, is to assume that Jordan scored better or higher. In that case, our null hypothesis is the average score by Michael Jordan is greater than or equal to the average score by Wilt Chamberlain. In this particular case, the alternative hypothesis would be different. It wouldn't be not equal to, but it would be less than. The alternative would be that the average scored by Jordan is less than that by Wilt Chamberlain.
- By the same account, our third option will be that the null hypothesis is that in fact Jordan scored less than Wilt Chamberlain and the alternative hypothesis will be the reverse of it, saying that no, Michael Jordan scored higher than Wilt Chamberlain.

In a nutshell, we have three options. We can say the averages are the same and the null would be, no, they are not the same, not equal. We can say the average is less than, the null is that Jordan's average is higher than Chamberlain, and the alternative would be the reverse of it. The third option is to say that the average scored by Jordan is less than average scored by Chamberlain, and the reverse of it will be the alternative hypothesis. So, three ways of defining a hypothesis.

总结:


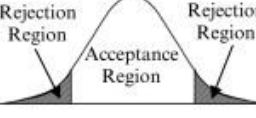



## Alpha (α) and P-value

**Alpha** and **p-value** are commonly used terms in statistical analysis. We hear these two terms quite often and will be hearing them until the end of this course. Both are used in hypothesis testing when we are trying to reject or fail to reject a given hypothesis.

Alpha (α) is also known as the significance level. It is the probability of rejecting the null hypothesis when the null hypothesis is true. The value used is often 5%. This means, that in a given population or sample if we computed a range of values in which the mean lies, 5 out of 100 times, this range of values will not contain the mean values but we might have said it does. The alpha value is the risk you are willing to accept if you are wrong, it signifies the rejection region and any value that falls inside this region will not be accepted. Depending on the use case, you may be willing to go as high as 20%. That means that you are okay with the consequences of falsely rejecting the null hypothesis 20% of the time. In a field like medicine, we will set the significance level as low as possible, e.g. 0.1% if there was potential harm to patients - in this case, we can't afford to leave any room for error.

We can see in this image the rejection regions for different kinds of tests.

| One-Tailed Test<br>(Left Tail)  | Two-Tailed Test  | One-Tailed Test<br>(Right Tail)  |
|---|--|--|
| $H_0 : \mu_X = \mu_0$<br>$H_1 : \mu_X < \mu_0$                                    | $H_0 : \mu_X = \mu_0$<br>$H_1 : \mu_X \neq \mu_0$                                  | $H_0 : \mu_X = \mu_0$<br>$H_1 : \mu_X > \mu_0$                                     |
|  |  |  |

**Note:** In cases of a Two-Tailed test, the rejection region on both tails add up to the total value of alpha i.e. if  $\alpha = 5\%$ , the rejection region on both tails will be 2.5% each.

P-value is a calculated value and an output you get as part of conducting your hypothesis test. Depending on the test, the calculation will vary. The p-value can be interpreted as the probability of getting a result that is as extreme or more extreme when the null hypothesis is true i.e. the likelihood of observing that particular sample value if the null hypothesis were true. Therefore, if the p-value is smaller than your significance level, you will reject the null hypothesis. For example, we have a null hypothesis that:

Null Hypothesis:  $\mu = 100$

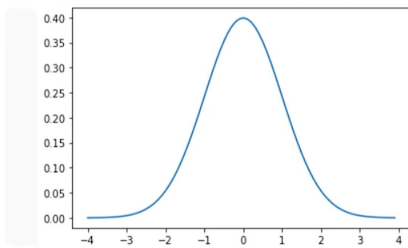
Alternative Hypothesis:  $\mu > 100$

If you conduct a test and you get a p-value of 0.02, this means that there is a 2% chance of obtaining a value of 100 or more than 100. If we picked a significance value of 5%, we will reject the null hypothesis, because 2% is less than 5%. If we picked a significance value of 1%, we will fail to reject the null hypothesis because 2% is greater than 1%.

To use both the p-value and significance level together, you have to decide on a value for alpha after you state your hypothesis. Suppose that is  $\alpha = 0.10$  (or 10%). You then collect the data and calculate the p-value. If the p-value is greater than alpha, you assume that the null hypothesis is true and you fail to reject. If the p-value is less than alpha, you assume that the null hypothesis is false and you reject it. In cases when the p-value and the significance levels are approximately equal e.g. a p-value of 0.11, it is your call to decide to reject or fail to reject or you could decide to resample and collect more data.

### 3. Normal Distribution

#### Normal Distribution



] Let me introduce you to normal distribution, which is one of the most commonly used distributions in statistical analysis, and even in everyday conversations. A large body of academic, scholarly and professional work rests on the assumption that the underlying data follows a normal distribution.

The defining characteristics of normal distribution is this bell shaped curve which you're familiar with from your textbooks.

Mathematically, normal distribution is presented by this equation. We can say that the normal distribution relies on three inputs and f stands for functions, so function of x, mu and sigma. X is your data, x is a random variable and it can attain any reasonable value. Mu stands for the mean. And sigma is standard deviation. And the mathematical formulation is right here, which is 1 divided by sigma times and then square root of 2 times pi, pi is 3.142 or 22/7. And then you have the exponential here. Do not forget this minus sign so exponential of this entity which is minus and in the numerator x- mean or x- mu d hold squared divided by 2 times sigma squared. So let me explain. 1 divided by the standard deviation and then square root of 2 times pi, 2 is known and pi's known. So is the value for exponential and what is not known is the sigma which you'd obtained from the data, that is standard deviation and the mean, which is also coming from data. So you have the mean and the standard deviation, and x is the random variable who's mean and standard deviation you're using. You put this all together and then you get the normal distribution, the bell shaped curve that you saw earlier.

#### The Standard Normal

- Mean ( $\mu = 0$ )
- Sigma ( $\sigma = 1$ )

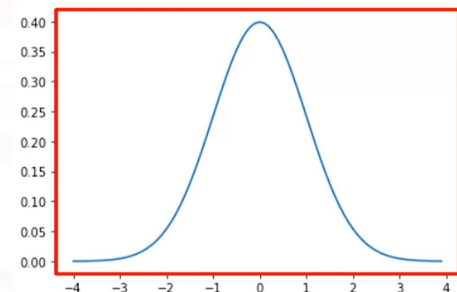
$$f(x, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)}$$

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

#### The Bell Curve

$$\mu = 0, \sigma = 1$$
$$-4 < x < 4$$

$$f(x, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 from scipy.stats import norm
4
5 # Plot between -4 and 4 with 0.1 steps.
6 x_axis = np.arange(-4, 4, 0.1)
7 # Mean = 0, SD = 1.
8 plt.plot(x_axis, norm.pdf(x_axis, 0, 1))
9 plt.show()
```

I also introduce you to the standard normal. And standard normal is when we say that x is a variable that has a mean 0 and standard deviation of 1. So what's mean 0 and standard deviation 1 look like? If you replace mu with 0 and standard deviation or sigma with 1, the equation reduce to this entity which is 1 / 2 times pi. Notice the sigma here which have great out a bit so that it doesn't interfere. Sigma is 1 so one times anything would be the same, so I've removed this. And then e to the power -(x- mu), but because mu is 0 so x- 0 is x, so ((x)2/2) times sigma squared, sigma, remember is 1, the square of 1 is 1, so 2 times 1. So removing sigma because it's 1 and anything multiplied with 1 is the same entity.

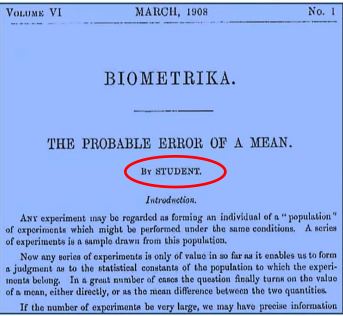
So how do I generate this normal density or a bell curve? Let's assume that the underlying variable has a mean 0 and the standard deviation of 1 and x varies between -4 and 4. So the mean is 0 and the minimum value is -4, the maximum value is 4. And I would substitute these -4 to 4 values in this equation. This is the only thing that's changing, the x here is the only entity that is changing, and let's see if this could generate the standard normal curve.

Let us do this in Python, we'll use the matplotlib function which you are already familiar with for the graphics, NumPy library, as well as the norm.pdf function in the SciPy stats library. In this example, I have used increments of 0.1. This will generate the standard normal curve that you hear about in statistics.

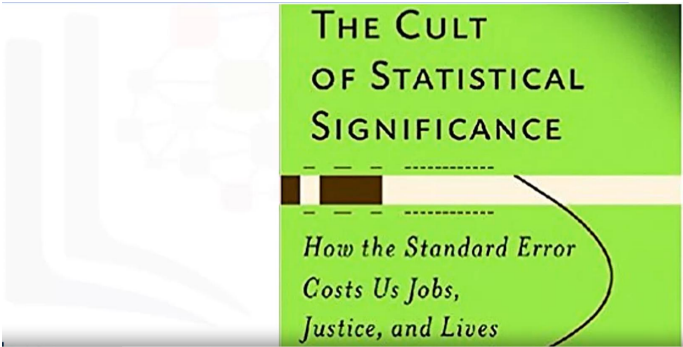
总结:

4. T distribution

William Sealy Gossett



Ziliak and McCloskey



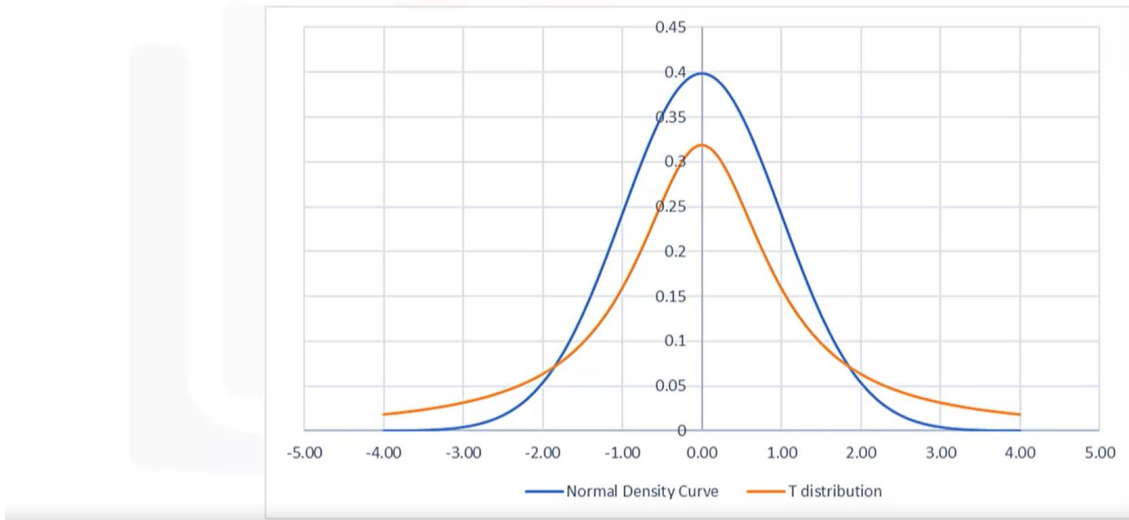
The other commonly used statistical distribution is known as the **Student's T-distribution**. **William Sealy Gosset** specified the T-distribution. In fact, he published a paper in Biometrika in 1908, and he published it under the pseudonym Student. He worked for the Guinness Brewery in Dublin, Ireland, where he worked with small samples of body. Mr. Gosset is the unsung hero of statistics. He published his work under a pseudonym because of the restrictions from his employer. Apart from his published work, his other contributions to statistical analysis are **equally significant**.

**The cult of statistical significance**, a must read book for anyone interested in Data Science, chronicles Mr. Gosset's work and how other influential statistician's of the time, namely Ronald Fisher and Egon Pearson, by way of their academic bonafide, ended up being more influential than the equally deserving Mr. Gosset.

t-Distribution

| cum. prob | t.50             | t.75  | t.80  | t.85  | t.90  | t.95  | t.975 | t.99  | t.995 | t.999  | t.9995 |
|-----------|------------------|-------|-------|-------|-------|-------|-------|-------|-------|--------|--------|
| one-tail  | 0.50             | 0.25  | 0.20  | 0.15  | 0.10  | 0.05  | 0.025 | 0.01  | 0.005 | 0.001  | 0.0005 |
| two-tails | 1.00             | 0.50  | 0.40  | 0.30  | 0.20  | 0.10  | 0.05  | 0.02  | 0.01  | 0.002  | 0.001  |
| df        |                  |       |       |       |       |       |       |       |       |        |        |
| 1         | 0.000            | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.31 | 636.62 |
| 2         | 0.000            | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3         | 0.000            | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4         | 0.000            | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173  | 8.610  |
| 5         | 0.000            | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893  | 6.869  |
| 6         | 0.000            | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208  | 5.959  |
| 7         | 0.000            | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785  | 5.408  |
| 8         | 0.000            | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501  | 5.041  |
| 9         | 0.000            | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297  | 4.781  |
| 10        | 0.000            | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144  | 4.587  |
| 11        | 0.000            | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025  | 4.437  |
| 12        | 0.000            | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930  | 4.318  |
| 13        | 0.000            | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852  | 4.221  |
| 14        | 0.000            | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787  | 4.140  |
| 15        | 0.000            | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733  | 4.073  |
| 16        | 0.000            | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686  | 4.015  |
| 17        | 0.000            | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646  | 3.965  |
| 18        | 0.000            | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610  | 3.922  |
| 19        | 0.000            | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579  | 3.883  |
| 20        | 0.000            | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552  | 3.850  |
| 21        | 0.000            | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527  | 3.819  |
| 22        | 0.000            | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505  | 3.792  |
| 23        | 0.000            | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485  | 3.768  |
| 24        | 0.000            | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467  | 3.745  |
| 25        | 0.000            | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450  | 3.725  |
| 26        | 0.000            | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435  | 3.707  |
| 27        | 0.000            | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421  | 3.690  |
| 28        | 0.000            | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408  | 3.674  |
| 29        | 0.000            | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396  | 3.659  |
| 30        | 0.000            | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385  | 3.646  |
| 40        | 0.000            | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307  | 3.551  |
| 60        | 0.000            | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232  | 3.460  |
| 80        | 0.000            | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195  | 3.416  |
| 100       | 0.000            | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174  | 3.390  |
| 1000      | 0.000            | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098  | 3.300  |
| Z         | 0.000            | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090  | 3.291  |
|           | 0%               | 50%   | 60%   | 70%   | 80%   | 90%   | 95%   | 98%   | 99%   | 99.8%  | 99.9%  |
|           | Confidence Level |       |       |       |       |       |       |       |       |        |        |

# Comparing Normal and T-Distribution



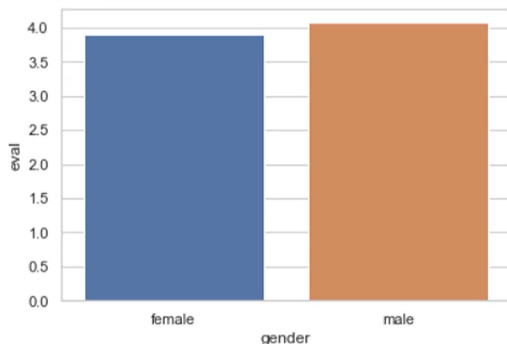
The normal distribution describes the mean for the population, whereas the T-distribution describes the mean of samples drawn from a population.

The T-distribution for each sample could be different and the T-distribution resembles the normal distribution for large sample sizes.

Here, I present normal distribution which is drawn in blue, and the T-distribution with a degree of freedom of one, **as the degrees of freedom increase, the T-distribution curve becomes more similar to the normal distribution**. In statistical analysis, several statistical tests rely on T-distribution. For instance, a comparison of means tests use the T-distribution. And it's also known as the **T-test**.

We have been working with a dataset comprising teaching evaluations of instructors from University of Texas. And I will illustrate the use of T-tests or T-distribution with the question of, does instructor evaluation score differ by gender? Do males and females get different teaching evaluations from students? Now, if I were to take the same dataset and compute the means and standard deviations. I can test this statistically.

## Is the Difference Statistically Significant?



## T-test

### Testing for statistical significance

- First the Assumptions
  - Scale of measurement
  - Simple Random Sample
  - Bell-shaped distribution
  - Homogeneity of variance

I have computed the visual representation of the average teaching evaluation score for male and female instructors. The blue bar represents the average teaching evaluation value for females. The orange bar represents the average teaching evaluation value for males. By eyeballing it, it is around four and slightly less for females. **Now it's a small difference between males and females. The question is, is this difference statistically significant?** To use a T-test, you have to make some assumptions are met.

- The first assumption is that the scale of measurement applied to the data collected **follows a continuous or ordinal scale**.
- The second assumption is that the data is collected from a **representative randomly selected portion of the total population**.
- The third assumption is **the Data when plotted, will follow a normal distribution. (bell-shaped)**
- And the final assumption is **homogeneity of variance**. To avoid the test statistics to be biased towards larger sample sizes. There's a test for this, which will be discussed later.



# T-test

Testing for statistical significance

- First the Assumptions
  - Scale of measurement
  - Simple Random Sample
  - Bell-shaped distribution
  - Homogeneity of variance
- State your hypothesis
  - Null hypothesis:  $\mu_1 = \mu_2$  ("there is no difference in evaluation scores for male and females")
  - Alternative hypothesis:  $\mu_1 \neq \mu_2$  ("there is a difference in evaluation scores between male and females")
  - alpha ( $\alpha$ ) level = 0.05

Before we go perform the test in Python. First, we will state our hypothesis.

- The null hypothesis is as follows. There is no difference in evaluation scores for males and females.
- The alternate hypothesis is there is a difference in evaluation scores between males and females. Then set alpha level to 0.05.

## T-test in python

```
1 import scipy.stats
```

```
1 scipy.stats.ttest_ind(ratings_df[ratings_df['gender'] == 'female']['eval'],  
2 ratings_df[ratings_df['gender'] == 'male']['eval'])
```

```
Ttest_indResult(statistic=-3.249937943510772, pvalue=0.0012387609449522217)
```

alpha ( $\alpha$ ) level = 0.05

To do this in Python, we will use the T-test independent sample in the Scipy stats function. The function takes in the two samples it is trying to test. The statistical difference of means for, in our example is the female's evaluation scores versus all the male's evaluation scores. It will return a T-statistic and a P-value. **Since the P-value is less than 0.05, the alpha level, we reject the null hypothesis as there is enough evidence that there is a statistical difference in teaching evaluations based on gender.**



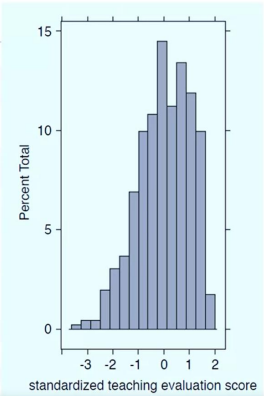
5. Probability of Getting a High or Low Teaching Evaluation

Probability of Getting a High or Low Teaching Evaluation

Standardization

z = (x - mu) / sigma

z = (4.5 - 3.998) / 0.554 = 0.906

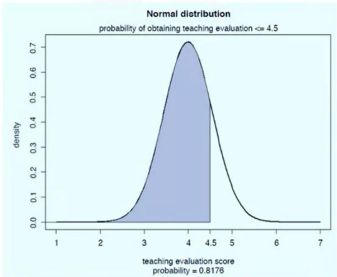


© IBM Corporation. All rights reserved.

Let me illustrate how to obtain the probability of getting a high or low teaching evaluation score from our dataset.

First, an important concept is the **standardization** of a variable such that it returns a dataset with a mean of zero and a standard deviation of one. I use the formula in equation shown here, where the standardization is taking a variable X and subtracting from it the average value mu, then dividing it by the standard deviation so that if the teaching evaluation score of an instructor on a scale of one to five is 4.5, we subtract the average teaching evaluation of 3.998 from it and divide it by the standard deviation, which is 0.554, resulting in a Z score of 0.906. If we were to just display the data as a histogram, you would see that it has a mean around zero. And the spread is shown on a scale where the X axis varies from minus three to two.

Normal Distribution Table



Probability Content from -∞ to Z

| Z   | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5315 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7938 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9789 | 0.9795 | 0.9799 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9969 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9983 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |

Standardization

z = (x - mu) / sigma

z = (4.5 - 3.998) / 0.554 = 0.906

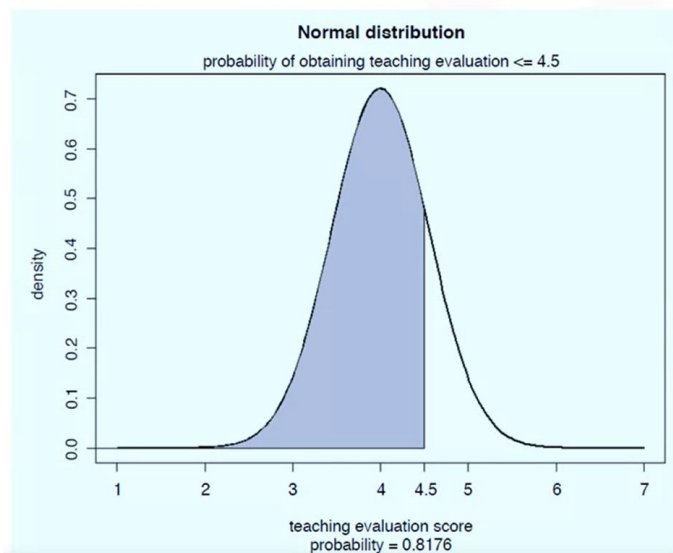
z | 0.00 | 0.01

|     |        |        |
|-----|--------|--------|
| 0.0 | 0.5000 | 0.5040 |
| 0.1 | 0.5398 | 0.5438 |
| 0.2 | 0.5793 | 0.5832 |
| 0.3 | 0.6179 | 0.6217 |
| 0.4 | 0.6554 | 0.6591 |
| 0.5 | 0.6915 | 0.6950 |
| 0.6 | 0.7257 | 0.7291 |
| 0.7 | 0.7580 | 0.7611 |
| 0.8 | 0.7881 | 0.7910 |
| 0.9 | 0.8159 | 0.8186 |

In a case where you do not have access to a computer with statistical software, you can still compute probabilities from a probability table using a simple and standard normal table found in statistics textbooks or downloaded online. A copy of such a table is on the right. Notice that the normal distribution graph to the left is grayed out in some parts. That grayed-out area represents the probability of getting some value Z, in this case Z. This value of Z or less than, we will need to first standardize the variable to determine the probability of a teaching evaluation score higher than 4.5 or less than 4.5.

Let's say, we have a dataset where the average teaching evaluation is 3.998 and the standard deviation is 0.554. And we are interested in determining the probability of getting a teaching evaluation score of 4.5 or less. So, from the table that I showed in the last slide, we can determine this. If we were to standardize the data, it becomes 0.906 because the accuracy of this table is only good for two decimal places. So 0.906 effectively becomes 0.91. We get a 0.8186 value here, hence, the probability of obtaining a teaching evaluation score of 4.5 or less is 0.8186.

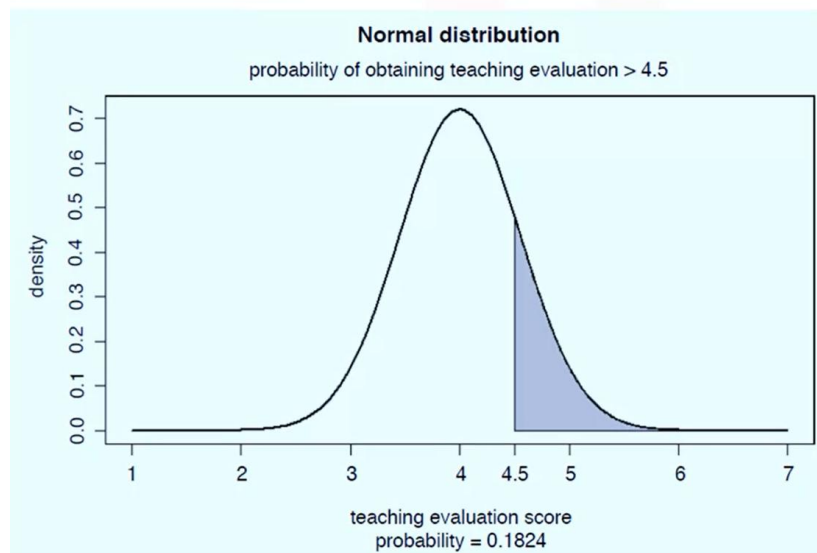
## Probability $\leq 4.5$



If you were to look at this graphic, you will see that I have plotted the area under the curve by shading it gray. That's the area that depicts the probability of an instructor receiving a teaching evaluation of less than or equal to 4.5. And that probability is 0.8176 or 81.76 percent.

Now, what will be the probability of receiving a teaching evaluation score of greater than 4.5. In fact, you can see from the next graphic that the probability is the reverse of one that we saw earlier. And hence, the probability of obtaining a teaching evaluation score of greater than 4.5 is 18.24 percent, which is the area shaded in gray. The reason for this is because the area under the normal distribution curve is equal to one. So one minus 0.8176 will give you the area for evaluation scores greater than 4.5.

## Probability $> 4.5$



## Python Syntax

Find the mean and Standard deviation

```
1 eval_mean = round(ratings_df['eval'].mean(), 3)
2 eval_sd = round(ratings_df['eval'].std(), 3)
3 print(eval_mean, eval_sd)
```

3.998 0.555

Using the norm.cdf package in scipy.stats, find the probability value.

```
1 import scipy.stats
2 prob0 = scipy.stats.norm.cdf((4.5 - eval_mean)/eval_sd)
3 print(1 - prob0)
```

0.1828639734596742

Z- scores are always to the left of the curve so we remove from one to get the opposite side

Let me illustrate the example of getting a teaching evaluation score of greater than 4.5. In Python, when we use the norm dot cdf function in the scipy.stats package. After finding the mean standard deviation, we plug it into the function with the X value of 4.5. And we will get the area to the left, which is the less than 4.5 area. Because we want the area to the right of 4.5, that is, the probability of greater than 4.5, we will remove the value from one as indicated here.

3. If a negatively skewed distribution (i.e. skewed to the left) has a median of 50, which of the following statements are true? (Select all that apply)

- ☐ Mean is greater than 50
- ☐ None of the above
- ☒ Mode is greater than 50

✓ 正确

Correct! Mean tends to move towards the tail of the data and mode does the opposite

- ☒ Mean is less than 50

✓ 正确

Correct! Mean tends to move towards the tail of the data and mode does the opposite

6. What is the area under a conditional Cumulative Density Function?

- ☐ 0.5
- ☐ 2
- ☐ 0
- ☒ 1

✓ 正确

Correct! The area under a Cumulative Density Function is calculated by adding the individual probabilities. This must always be equal to 1



# Standard Normal Table

You can use the z-table to find a set of “less-than” probabilities for a wide range of z-values. To use the z-table to find probabilities::

1. Go to the row that represents the first digit and the first digit after the decimal point of your z-value. (e.g 0.9 in 0.93)
2. Go to the column that represents the second digit after the decimal point of your z-value. (e.g 3 in 0.93, this will be 0.03 in the column)
3. Intersect the row and column from Steps 1 and 2.

For example, suppose you want to find the probability of z-score less than 1.44 denoted as  $p(Z < 1.44)$ . Using the second Z-table below, find the row for 1.4 and the column for 0.04. Intersect that row and column to find the probability: 0.92507. Therefore  $p(Z < 1.44) = 0.92507$ .

The area under any normal curve (including the standardized normal curve) is 1, that means that,  $p(Z < 1.44) + p(Z > 1.44) = 1$ . Therefore, the probability of z-score greater than 1.44 i.e.  $p(Z > 1.44) = 1 - p(Z < 1.44)$  which equals  $1 - 0.92507$  which equals 0.07493.

Suppose you want to look for  $p(Z < -1.44)$ . You find the row for  $-1.4$  and the column for 0.04. Intersect the row and column and you find 0.07493. That means  $p(Z < -1.44) = 0.07493$ . This happens to be the same as the value of  $p(Z > +1.44)$ . This is because the normal distribution is symmetric. So the tail of the curve below  $-1.44$  representing  $p(Z < -1.44)$  looks exactly like the tail above 1.44 representing  $p(Z > +1.44)$ .

You can also do a reverse lookup, assuming you are told that the age of grade 1 kindergarten pupils in Kampala city is normally distributed with a mean of 6 years. If only 0.71% (same as 0.0071) of the pupils are 8 years and above, what is the z-score?

We can look within the negative box for the z-score closest to 0.0071. In this case -2.45, but because we are looking at the right-tailed test, we get rid of the negative sign i.e. z-value = 2.45 (Note: if it was a left-tailed test e.g. 0.71% of them were less than 3 years old, we will use the value as-is) OR

We can look within the positive table, for the value that corresponds to  $1 - 0.0071 = 0.99286$  (we do this because the standard normal table always gives you values to the left, so to get values to the right, you will have to remove from 1). The value that corresponds to this will be 2.45.

To make things easier, you can also use a "z-score from p-value" or a "p-value from z-score" calculator online.