Continuous Distributions: Normal Curve[1]

November 6, 2017

[1] HMS, 2017, v1.3

Chapter References

- Diez: Chapter 3.1, 3.2
- Navidi, Chapter 4.5, 4.7, 4.10

The Normal or Gaussian Distribution

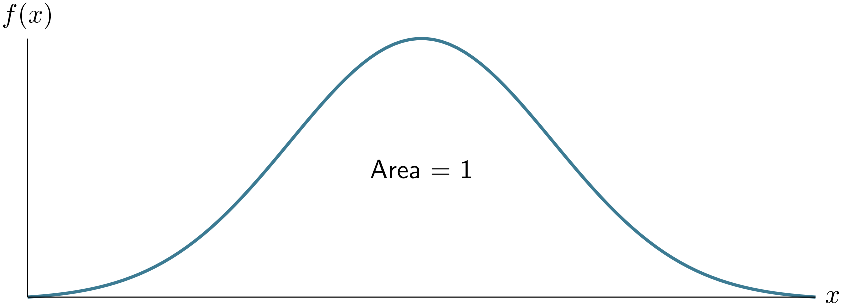
- This is the most commonly used distribution in statistics
- The probability density function of a normal random variable with mean μ and variance σ^2 is given by:

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

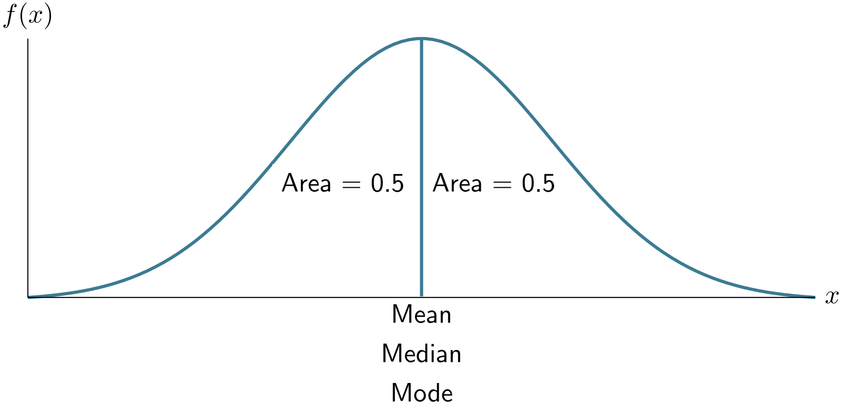
- The PDF is often abbreviated to:

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

The Normal or Gaussian Distribution



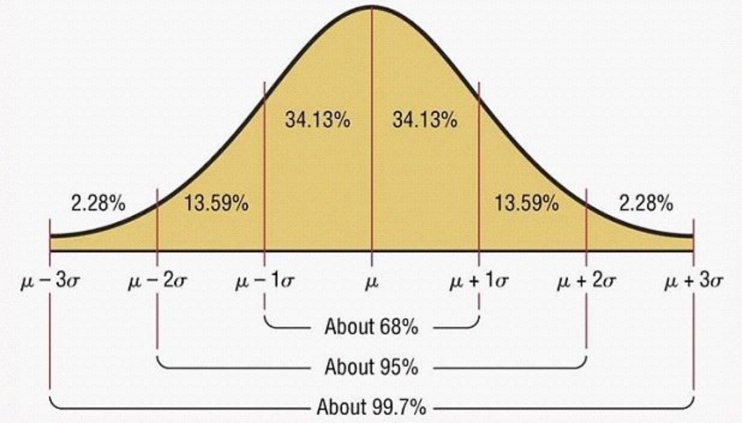
The Normal or Gaussian Distribution



The Normal or Gaussian Distribution

- The mean, median, and mode are equal.
- The normal curve is bell-shaped and symmetric about the mean.
- The total area under the curve is equal to one.
- The normal curve asymptotically approaches zero on either side of the mean.
- The inflection point occurs at -1 and $+1$ σ.

The Normal or Gaussian Distribution



The Normal or Gaussian Distribution

- About 68% of the probability is in the interval $\mu \pm \sigma$
- About 95% of the probability is in the interval $\mu \pm 2\sigma$
- About 99.7% of the probability is in the interval $\mu \pm 3\sigma$

Standard Normal Distribution: Z score

- Rescales the x-axis so that it is in units of standard deviation.

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

- x is any point on the horizontal axis.
- The z-score moves the mean to zero.
- Normalizes σ so that the 68% mark is now at the x value of 1.0.
- The advantage of the z-score is that it makes it possible to compare different normal distributions.

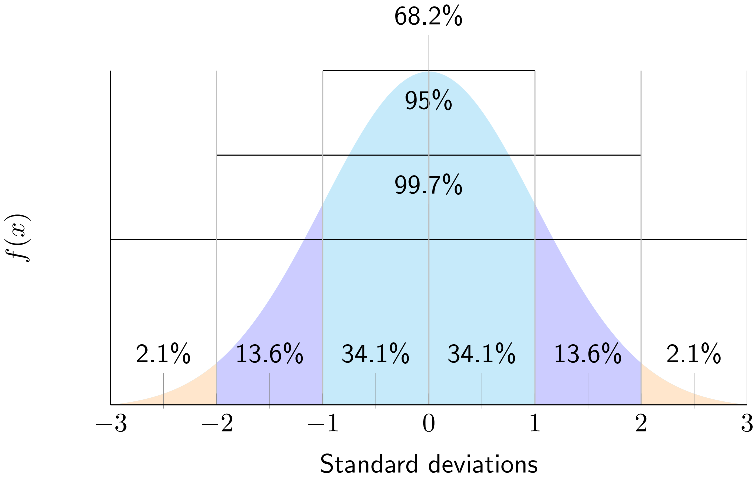
Standard Normal Distribution: Z score

- Note that when $x = \mu$ the z-score is zero:

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

- Given a normal curve $N(\mu, \sigma)$ the equivalent z-curve is denoted by $N(0, 1)$

Normalized Gaussian Distribution



Z score: Example

- During one week (Mon to Sun) the temperature fluctuated with temperatures, 62, 68, 52, 40, 78, 72 and 60.
- The mean temperature was 61.7 and standard deviation is computed to be 11.8
- Convert the daily temperatures to z-scores
- Applying $(x - \mu)/\sigma$ to each temperature yields:

  0.02, 0.53, -0.81, -1.8, 1.4, 0.86, and -0.14
- Which days are outside one standard deviation?

  Thursday and Friday

Z score: Going Backwards

- It is also possible to calculate a x value given a z-score:

$$x = \mu + z\sigma$$

Example: If a z-score is 2.5 and a population has a mean of 4.0 and standard deviation of 0.5, what is the value of x that corresponds to this z-score?

$$4.0 + 2.5 \times 0.5 = 5.25$$

Z score: Example

- Assume that a set of test scores has a mean of 150 and standard deviation of 25
- If a particular student had a score of 190, what is his/her z score?

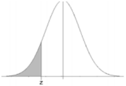
$$z = \frac{x - \mu}{\sigma}$$

- Therefore $z = (190 - 150)/25 = 1.6$
- What percentage of students have scores above this?
- We need to find the area above $z = 1.6$.

Z Score Table: Negative Values

Table A.2 in Navidi, Table B1 in Diez

Standard Normal Cumulative Probability Table



Cumulative probabilities for NEGATIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559

Z Score Table: Positive Values

Standard Normal Cumulative Probability Table

Cumulative probabilities for POSITIVE z-values are shown in the following table:

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	0.5319	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	0.7939	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	0.9788	0.9793	0.9798	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

Z score: Example

$$p = P(Z > +1.6) = 1 - P(Z \leq 1.6) = 1 - 0.9452 = 0.0548$$

That is 5.5%

Standard Normal Cumulative Probability Table

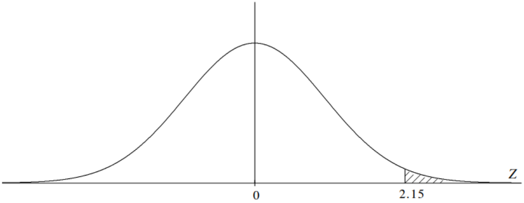
Cumulative probabilities for POSITIVE z-values are shown in the following table:

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	0.5319	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	0.7939	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	0.9788	0.9793	0.9798	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

Z score: Class Exercises

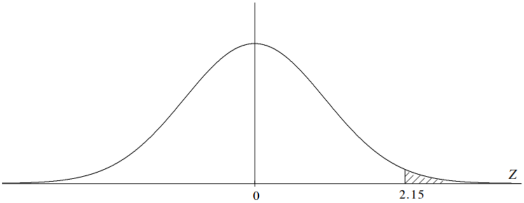
1. If $z = 2.15$, what is the area beyond z?
2. Find the area below z
3. What is the sum of the above two areas?
4. What proportion of the z scores are less than a z score of 1.58?
5. What is the area between the mean and 0.85 standard deviations below the mean?
6. What is the probability of obtaining a z score between 0.33 and 1.33?
7. What z score is exceeded by 10% of all scores?

Z score: Area beyond 2.15



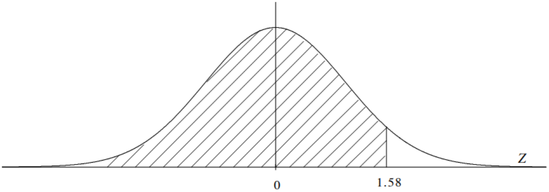
$$p = 0.0158$$

Z score: Area below 2.15



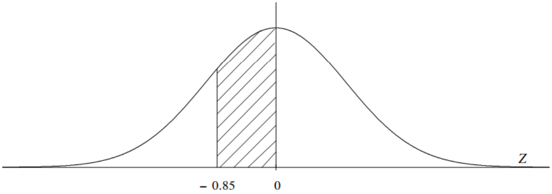
$$p = 1 - 0.0158$$

Z score: Area below 1.58



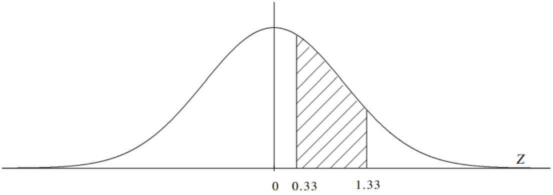
$$p = 0.9429$$

Z score: Area between 0 and -0.85



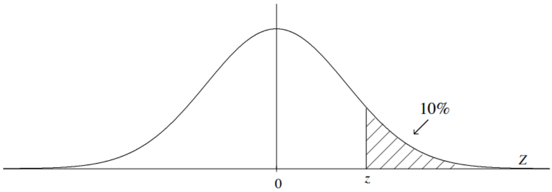
$$p = 0.3023$$

Z score: Area between 0.33 and 1.33



$$p = 0.2789$$

Z score: z score exceeded by 10% of all scores?



$z = 1.28$ this is the 90th percentile

Example

The lifetime of batteries is normally distributed. with mean 50 hours, and standard deviation 5 hours. Find the probability that a randomly chosen battery lasts between 42 and 52 hours.

Let X be the lifetime of a randomly chosen battery such that $X \sim N(50, 5^2)$[2]

[2] Sketch Curve and add 42, 52 shaded area. We need to compute the area.

Example

The lifetime of batteries is normally distributed. with mean 50 hours, and standard deviation 5 hours. Find the probability that a randomly chosen battery lasts between 42 and 52 hours.

$$z = \frac{42 - 50}{5} = -1.6 \qquad z = \frac{52 - 50}{5} = 0.4$$

Next use table to find bounded area between -1.6 and 0.4

$$0.6554 - 0.0548 = 0.6006$$

Example

How would you find the 40% percentile of the battery lifetimes?

Find the z-scores that bounds the area between the left and 0.4. This roughly equals $z = -0.25$. We want to know the lifetime at the z-score of -0.25:

$$-0.25 = \frac{x - 50}{5}$$

Solving for x yields: $x = 48.75$.

Example

A process manufactures artificial ball joints whose diameters are normally distributed with a mean of 2.505 cm and standard deviation of 0.008 cm. Specifications call for the diameter to be in the interval 2.5 ± 0.01 cm. What proportion of ball joints will meet the specification?

$$\mu = 2.505 \quad \sigma = 0.008$$

Example

Let X represent the diameter of a randomly chosen ball joint. Then $X \sim N(2.505, 0.008^2)$

First thing to do is compute z-scores for the interval 2.49 and 2.51:

$$z = \frac{2.49 - 2.505}{0.008} = -1.88 \qquad z = \frac{2.51 - 2.505}{0.008} = 0.63$$

Then look up the area that this represents

$$0.7357 - 0.031 = 0.7056$$

Linear Functions of Normal Variables

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

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

Linear Functions of Normal Variables: Example

A temperature is measured in $^\circ C$ with mean 40 $^\circ C$ and standard deviation 1 $^\circ C$. The measurement is converted to $^\circ F$ by the equation $F = 1.8C + 32$.

What is the distribution of F?

$$\mu_c = 40, \quad \text{therefore} \quad \mu_F = 1.8(40) + 32 = 104$$

$\sigma_C^2 = 1$, therefore $\sigma_F^2 = 1.8^2(1) = 3.24$. Therefore:

$$F \sim N(104, 3.24)$$

Linear Combinations of Normal Random Variable

If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, etc. are **independent** normal random variables and let c_1, c_2, etc be constants then:

$$c_1X_1 + c_2X_2 + \ldots \sim N(c_1\mu_1 + c_2\mu_2 + \ldots,\ c_2^2\sigma_1^2 + c_2^2\sigma_2^2 + \ldots)$$

Linear Combinations of Normal Random Variable

If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, etc. are **independent** normal random variables and let c_1, c_2, etc be constants then:

$$c_1 X_1 + c_2 X_2 + \ldots \sim N(c_1\mu_1 + c_2\mu_2 + \ldots,\ c_2^2\sigma_1^2 + c_2^2\sigma_2^2 + \ldots)$$

Read example 4.51 in Navidi

Also:

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

How do we know if the Data is Normally Distributed?

Use Probability Plots (Neither textbooks give a great explanation of these but see Diez Sec 3.2 and Navidi 4.10)

There are a number of ways to do it but one way is the so-called Q-Q plot. Suppose we have a population that is normally distributed: $X \sim N(\mu, \sigma^2)$. Then:

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

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

The above equation is a straight line. We can exploit this to see if a sample X is normally distributed by seeing how well it conforms to a straight line. We can plot Z versus the data points X.

How do we know if the Data is Normally Distributed?

Procedure:

1. Rank your data (of which there are n) from smallest to largest value
2. Compute the percentile value for the ith data point using:

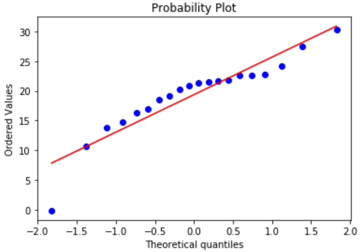
$$p_i = \frac{i - 0.5}{n}$$

   e.g if we have 20 data points, then the percentile for the 10th data point is $(10 - 0.5)/20 = 0.475$. Consider this value as the area under a normal curve. This value will have a corresponding z value.
3. Compute the z score for each computed percentile and plot the data (X) on the y axis and z score (Z) on the x-axis.
4. A straight line indicates normality.

Using Python to compute Q-Q plots

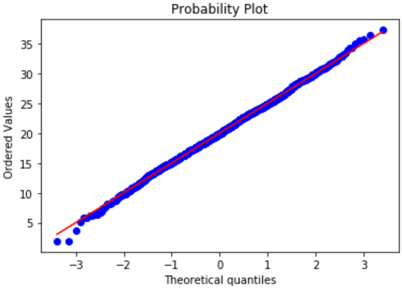
```
import numpy as np
import pylab
import scipy.stats as stats

measurements = np.random.normal(loc = 20, scale = 5, size=20)
stats.probplot(measurements, plot=pylab)
pylab.show()
```

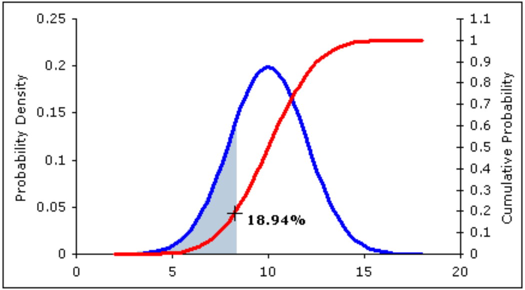


Using Python to compute Q-Q plots

2000 data points samples from a Normal distribution:



CDF: Cumulative Distribution



Central Limit Theorem

The central limit theorem is one of the most important results in statistics.

Many statistical methods rely on this theorem.

Central Limit Theorem

What does it say?

If we draw a large enough sample (≥ 30) from a population, then the distribution of the sample mean is approximately normal, **no matter** what the underlying distribution of the population data.

Central Limit Theorem

In other words:

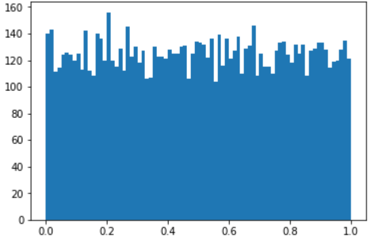
1. Draw many samples from a population.
2. Compute the mean, $\overline{X}$, for each sample.
3. The distribution of the means will be normal.

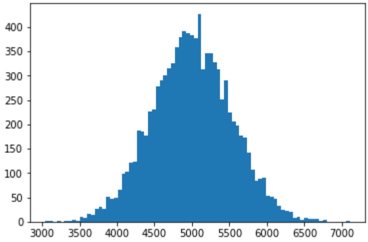
In particular:

- $\mu_{\overline{X}} = \mu$ The mean of the means will equal the population mean
- $\sigma^2_{\overline{X}} = \sigma^2/n$ The variance of the means will equal the population
- variance divided by the sample size.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Central Limit Theorem





Question

Assume we sample from a normal distribution. Will the standard error be smaller or bigger than the population standard deviation?

Standard Error

- The standard error estimates variability between samples - accuracy
- The standard deviation estimates variability within a single sample - precision

Standard Error

The standard error also indicates that if we want to reduce the error of our estimate of the mean by a factor of n, we have to gather n^2 times more data.

Standard Error: Example

- Three measurements of gene expression yield the values 1.34, 3.23, and 2.11 (unit: mg hr^{-1} L^{-1})
- Find the mean and standard deviation of these values.
- Find the standard deviation of the mean
- Note that the three values represent a sample of expression values.

Since we don't have the population standard deviation we must use the sample standard deviation to estimate the standard error.

$$\sigma = 0.95 \qquad \text{Mean} = 2.1 \qquad \text{SE} = \sigma/\sqrt{3} = 0.548$$

Standard Error: Example

- The cholesterol levels in a sample of 25 men aged 55 to 60 was found to have a mean of 5.2 mmol L^{-1}. The standard deviation of the same was 0.6 mmol L^{-1}.
- Compute the standard error of the means from the sample?

$$\text{SE} \sim 0.6/\sqrt{5} = 0.12 \text{ mmol L}^{-1}$$

Standard Error: Example

- The cholesterol levels in a sample of 25 men aged 55 to 60 was found to have a mean of 5.2 mmol L^{-1}. The standard deviation of the same was 0.6 mmol L^{-1}.
- Compute the standard error of the means from the sample?

$$\text{SE} \sim 0.6/\sqrt{5} = 0.12 \text{ mmol L}^{-1}$$

Sources of Error

- Human error, usually random
- Variation in subject, usually random
- Systematic error, vias in instrument, not random - can be corrected

Due to the central limit theorem, many errors are normally distributed.

Simulation as a Way to Estimate Error

Consider two resistors connected in parallel: $X = 100\ \Omega$ and $Y = 25\ \Omega$.

The total resistance is given by:

$$R = \frac{XY}{X+Y} = 2500/125 = 20\ \Omega$$

Resistances are sold according to tolerances, therefore assume $X \sim N(100, 10^2)$ and $Y \sim N(25, 2.5^2)$ - roughly 10% tolerance level

The specification for the circuit states that $19 < R < 21$. What is the probability that an assembly will meet this specification?

To compute the probability we will use simulation.

Simulation as a Way to Estimate Error

Sample N resistors each for X and Y. That is

```
X = np.random.normal(Mean, SD, sizeOfSample)
```

```
Xn = np.random.normal(100, 10, 10)
Yn = np.random.normal(25, 2.5, 10)
```

```
Rn = (Xn * Yn)/(Xn + Yn)
```

Simulation as a Way to Estimate Error

Sort the values in Rn:

```
Rn = np.sort (Rn)
pylab.hist (Rn, 50)

count = 0.0
for i in Rn:
    if i >= 19 and i <= 21:
        count = count + 1

print "Probability = ", count/N
Probability =  0.4560181
```

Simulation as a Way to Estimate Error

One step further:

```
mean = np.mean (Rn)
stddev = np.std (Rn)
```

What Happens if you Don't Know the Underlying Distribution?

Assume we don't know how the resistor values actually vary? Instead someone measures 8 resistors values of each type X and Y:

`21, 26, 24, 27, 28, 23, 26, 23` $\overline{x} = 24.75$

`105, 96, 101, 98, 103, 99, 105, 96` $\overline{x} = 100.4$

What Happens if you Don't Know the Underlying Distribution?

Using our samples we create new synthetic samples X_1, $X_2, \ldots$, and $Y_1, Y_2, \ldots$ by randomly picking out values (with replacement) to generate new samples, eg

`26, 23, 27, 23, 26, 23, 28, 21`

Do this 1000 times. Then use the samples sets to compute the assembly resistancd as before.