
ACADGILD
Mastering Data
Science

Session 7 – Distributions and CLT

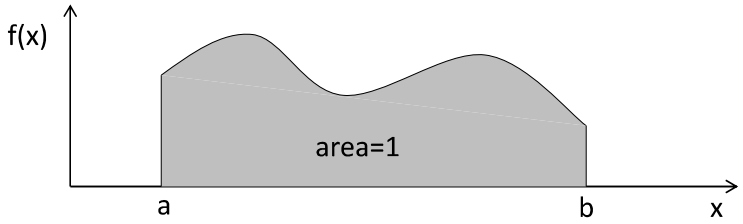
Agenda

1. Probability Function
2. Binomial Distribution
3. Binomial Random Variable
4. Poisson Distribution
5. Poisson Probability Distribution
6. The Normal Distribution
7. Standard Normal Distribution
8. Calculating Normal Probabilities
9. Using the Normal Table
10. Finding and Using the Values of Z
11. Central Limit Theorem
12. Sampling Distribution of the Sample Mean

Probability Density Function

A function f(x) is called a Probability Density Function over the range $a \le x \le b$ if it meets the following requirements:

1. $f(x) \ge 0$ for all x between a and b, and



2. The total area under the curve between a and b is 1.0

3.c. Discrete Probability distributions

I. Define and describe Discrete Probability distributions

1) Uniform
2) Bernoulli
3) Binomial
4) Poisson
5) Negative Binomial

Uniform distribution

Sample space $S = \{1, 2, 3, \ldots, k\}$.

Probability measure: equal assignment (1/*k*) to all outcomes, *ie* all outcomes are equally likely.

Random variable *X* defined by $X(i) = i$, $(i = 1, 2, 3, \ldots, k)$.

Distribution: $P(X = x) = \dfrac{1}{k}$ $\quad (x = 1, 2, 3, \ldots, k)$

Moments:

$$\mu = E[X] = \frac{(1 + 2 + \cdots + k)}{k} = \frac{\frac{1}{2}k(k+1)}{k} = \frac{k+1}{2}$$

$$E[X^2] = \frac{(1^2 + 2^2 + \cdots + k^2)}{k} = \frac{\frac{1}{6}k(k+1)(2k+1)}{k} = \frac{(k+1)(2k+1)}{6}$$

$$\Rightarrow \sigma^2 = \frac{k^2 - 1}{12}$$

For example, if X is the score on a fair die, $P(X = x) = \frac{1}{6}$ for $x = 1, 2, \ldots, 6$.

3.c. Discrete PDF- Bernoulli distribution

I. A Bernoulli trial is an experiment which has only two possible outcomes – s (“success”) and f (“failure”).

II. “success” and “failure” are mere labels and should not be taken literally. Instead we could have “yes” and “no” OR “true” and “false”

III. Sample space S = {s,f} .

IV. Probability measure:
   I. $P(\{s\}) = p$, $P(\{f\}) = 1 - p$ $0 < p < 1$

V. Random variable X defined by $X(s) = 1$, $X(f) = 0$.

VI. Distribution: $P(X = x) = p^x * (1-p)^{1-x}$, $x = 0, 1; 0 < p < 1$

VII. Expected values:
   I. Mean, $\mu = p$
   II. Variance, $\sigma^2 = p(1 - p)$

VIII. Examples:
   I. Tossing of a coin. “Head” corresponds to “success” and “Tail” corresponds to “failure”.
   II. Defaulting a home loan. “Default” corresponds to “success” and “Non-default” corresponds to “failure”.
   III. Auto insurance policy. “No claim” corresponds to “success” and “Claim” corresponds to “failure”.

Binomial Distribution

➢ The binomial distribution is the probability distribution that results from doing a "binomial experiment". Binomial experiments have the following properties:

1. Fixed number of trials, represented as n.
2. Each trial has two possible outcomes, a "success" and a "failure".
3. P(success)=p (and thus: P(failure)=1–p), for all trials.
4. The trials are independent, which means that the outcome of one trial does not affect the outcomes of any other trials.

Binomial Random Variable

➢ The binomial random variable counts the number of successes in n trials of the binomial experiment. It can take on values from 0, 1, 2, .., n. Thus, its a discrete random variable.

➢ To calculate the probability associated with each value we use combinatory:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for x=0, 1, 2, .., n

Binomial Random Variable

➢ The binomial random variable counts the number of successes in n trials of the binomial experiment. It can take on values from 0, 1, 2, .., n. Thus, its a discrete random variable.

➢ To calculate the probability associated with each value we use combinatory:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for x=0, 1, 2, .., n

3.c. Discrete PDF- Binomial distribution

What is the probability that at least 9 out of a group of 10 people who have been infected by a serious disease will survive, if the survival probability for the disease is 70%?

3.c. Discrete PDF- Binomial distribution

What is the probability that at least 9 out of a group of 10 people who have been infected by a serious disease will survive, if the survival probability for the disease is 70%?

The number of survivors is distributed binomially with parameters $n = 10$, and $p = 0.7$. If X is the number of survivors, then:

$$P(X \geq 9) = P(X = 9 \text{ or } 10) = \binom{10}{9} \times 0.7^9 \times 0.3 + \binom{10}{10} \times 0.7^{10} = 0.1493$$

Binomial Distribution

➢ As you expect, Statisticians have developed general formulas for the mean, variance, and standard deviation of a binomial random variable. They are:

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

$$\sigma = \sqrt{np(1-p)}$$

Poisson Distribution

➢ Named after Simeon Poisson, the Poisson distribution is a discrete probability distribution and refers to the number of events (a.k.a. successes) within a specific time period or region of space.

➢ For example,

- The number of cars arriving at a service station in 1 hour. (The interval of time is 1 hour)
- The number of flaws in a bolt of cloth. (The specific region is a bolt of cloth)
- The number of accidents in 1 day on a particular stretch of highway. (The interval is defined by both time, 1 day, and space and the particular stretch of highway.)

The Poisson Experiment

Similar to binomial experiment, a Poisson experiment has four defining characteristic properties:

1. The number of successes that occur in any interval is independent of the number of successes that occur in any other interval.
2. The probability of a success in an interval is the same for all equal-size intervals
3. The probability of a success is proportional to the size of the interval.
4. The probability of more than one success in an interval approaches 0 as the interval becomes smaller.

Poisson Distribution

➢ The Poisson random variable is the number of successes that occur in a period of time or an interval of space in a Poisson experiment.

➢ For example, on average, 96 trucks arrive at a border crossing every hour.

➢ For example, the number of typographic errors in a new textbook edition averages 1.5 per 100 pages.

Poisson Probability Distribution

➢ The probability that a Poisson random variable assumes a value of x is given by:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!} \quad for \quad x = 0,\ 1,\ 2,\ldots$$

where μ is the mean number of successes in the interval

e – natural logarithm base

E(X) = V(X) = μ

Poisson Probability Distribution

Example

➢ The number of typographical errors in new editions of textbooks varies considerably from book to book. After some analysis it concludes that the number of errors is Poisson distributed with a mean of 1.5 per 100 pages. The instructor randomly selects 100 pages of a new book. What is the probability that there are no typos?

That is, what is P(X=0) given that μ = 1.5?

$$P(0) = \frac{e^{-\mu}\mu^{x}}{x!} = \frac{e^{-1.5}1.5^{0}}{0!} = .2231$$

"There is about a 22% chance of finding zero errors"

Poisson Distribution

➢ The probability of success is proportional to the size of the interval.

➢ Thus knowing an error rate of 1.5 typos per 100 pages, we can determine a mean value for a 400 page book as:

μ = 1.5 (4) = 6 typos/ 400 pages

μ = 1.5 (4) = 6 typos / 400 pages

3.c. Discrete PDF- Negative- Binomial distribution

Discrete probability distribution of the number of failures (r) in a sequence of Bernoulli trials before a specified (non-random) number k of success occurs

Special generalized case of the Poisson distribution

- Intensity rate (λ) is no longer taken to be constant (Assumed to follow a Gamma Distribution)
- Two-parameter distribution
- Provides additional flexibility in fitting data
- Parameter uncertainty maybe high with less data points (typical of scenario where annual frequency data points maybe 3-6)
- Advantages
- Allows modelling of the frequency dependence due to the assumption that occurrence of operational losses may be affected by some external factor

Negative Binomial Distr. – 2 Param Distr.

The probability of getting exactly r failures before k successes is given by the probability mass function:

for k = 0, 1, 2, ..., n,

where
${}^{n}C_{k} = n!/\{(n-k)!*k!\}$

Discrete (Counting) Distribution – Popular for Modeling Frequency of Losses
Variance > Mean, useful if variance of operational loss frequency is greater than mean

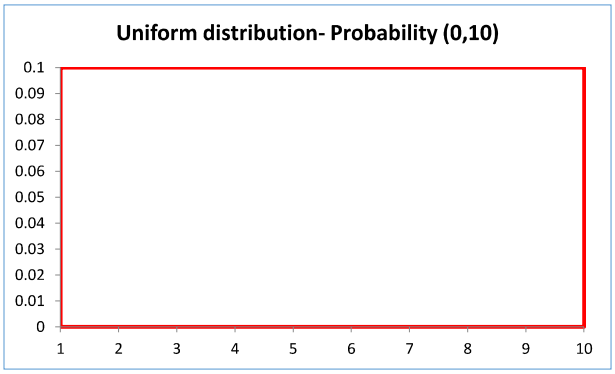
3.c. Continuous Probability Distribution

Define and describe Continuous Probability distributions

1) Uniform
2) Gamma
3) Normal
4) Lognormal
5) Exponential
6) Gamma
7) Chi-square
8) t- distribution
9) F- distribution

3.c. Continuous PDF- Uniform distribution

- Assigns equal probability to all values between its minimum and maximum values.
- Random variable X takes a value between two number a and b (say).
- Probability density function: $f_X(x) = 1/(b-a)$, $a < x < b$
- Denoted as $X \sim U(a, b)$
- Expected values:
  - Mean, $\mu = (a + b)/2$
  - Variance, $\sigma^2 = (b - a)^2/12$
- Example: Assigning equal probability of default to a portfolio of credit card holders.



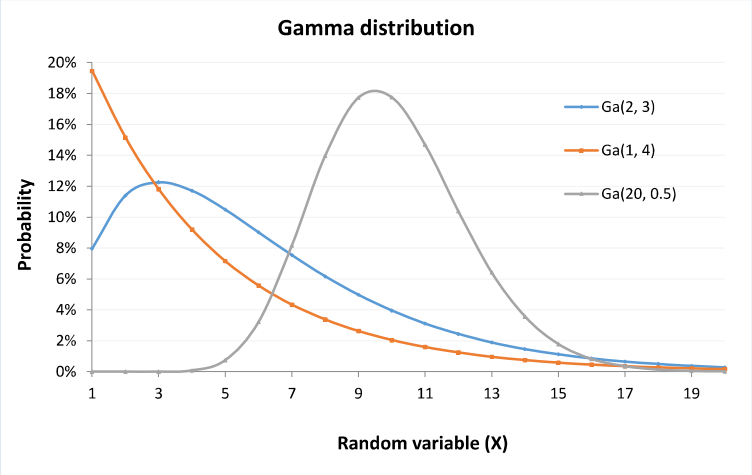
3.c. Continuous PDF- Gamma distribution

- Gamma family if distributions is a positively-skewed distribution explained by two parameters "α" and "λ" (say).
- It is bounded at zero and can take various shapes depending on values of parameters.
- Random variable X takes a non-zero positive value.
- Probability density function: $f_X(x) = (\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}) / \Gamma(\alpha)$, $x > 0$
- Denoted as $X \sim Gamma(\alpha, \lambda)$
- Expected values:
  - Mean, $\mu = \alpha / \lambda$
  - Variance, $\sigma^2 = \alpha / \lambda^2$
- Special cases:
  - Exponential distribution when $\alpha = 1$: $f_X(x) = \lambda e^{-\lambda x}$, $x > 0$
  - Chi-square distribution with $\alpha = 2v$ (v any positive integer) and $\lambda = 1/2$
- Example:
  - Used the predict claim amount in Auto insurance.
  - Used the predict loss amount in bank loan defaults

3.c. Continuous PDF- Gamma distribution

- Plotting PDFs for different Gamma distributions using MS Excel.

X	Ga(2, 3)	Ga(1, 4)	Ga(20, 0.5)
1	7.96%	19.47%	0.00%
2	11.41%	15.16%	0.00%
3	12.26%	11.81%	0.00%
4	11.72%	9.20%	0.08%
5	10.49%	7.16%	0.75%
6	9.02%	5.58%	3.23%
7	7.54%	4.34%	8.17%
8	6.18%	3.38%	13.98%
9	4.98%	2.63%	17.73%
10	3.96%	2.05%	17.77%
11	3.12%	1.60%	14.71%
12	2.44%	1.24%	10.40%
13	1.90%	0.97%	6.44%
14	1.46%	0.75%	3.56%
15	1.12%	0.59%	1.79%
16	0.86%	0.46%	0.82%
17	0.65%	0.36%	0.35%
18	0.50%	0.28%	0.14%
19	0.37%	0.22%	0.05%
20	0.28%	0.17%	0.02%

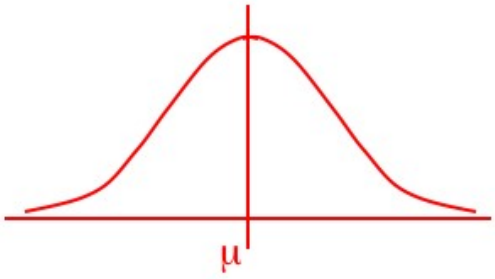


Normal Distribution

➢ The normal distribution is the most important of all probability distributions. The probability density function of a normal random variable is given by:

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

➢ It looks like bell shaped, symmetrical around the mean, μ



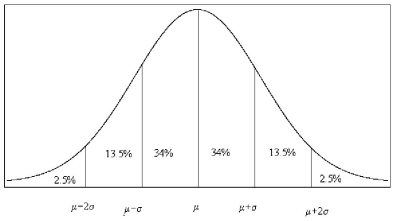
Normal Distribution

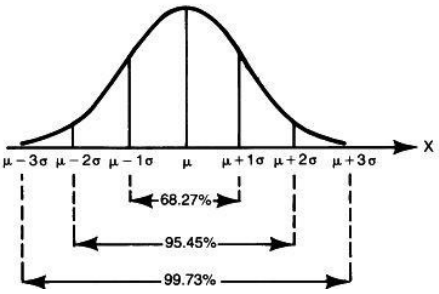
➢ The normal distribution is completely defined by two parameters Standard Deviation and Mean.

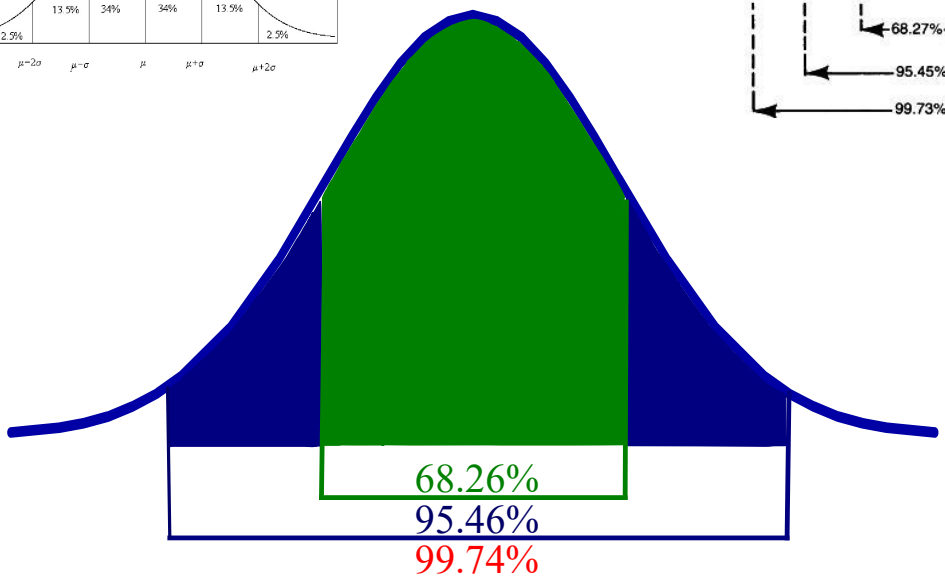
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

➢ The normal distribution is bell shaped and symmetrical about the mean.

Normal Distribution





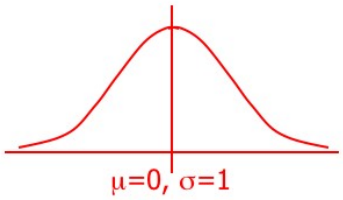


Standard Normal Distribution

➢ A normal distribution whose mean is zero and standard deviation is 1 is called the standard normal distribution.

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

➢ As we shall see shortly, any normal distribution can be converted to a standard normal distribution with simple algebra. This makes calculations much easier.

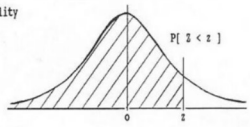


STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[\,Z < z\,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}z^2)\, dz$



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	0.5159	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	0.7854
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	0.8804	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	0.9773	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	0.9865	0.9868	0.9871	0.9874	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	0.9924	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	0.9968	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	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

What is the area to the left of Z=1.51 in a standard normal curve?

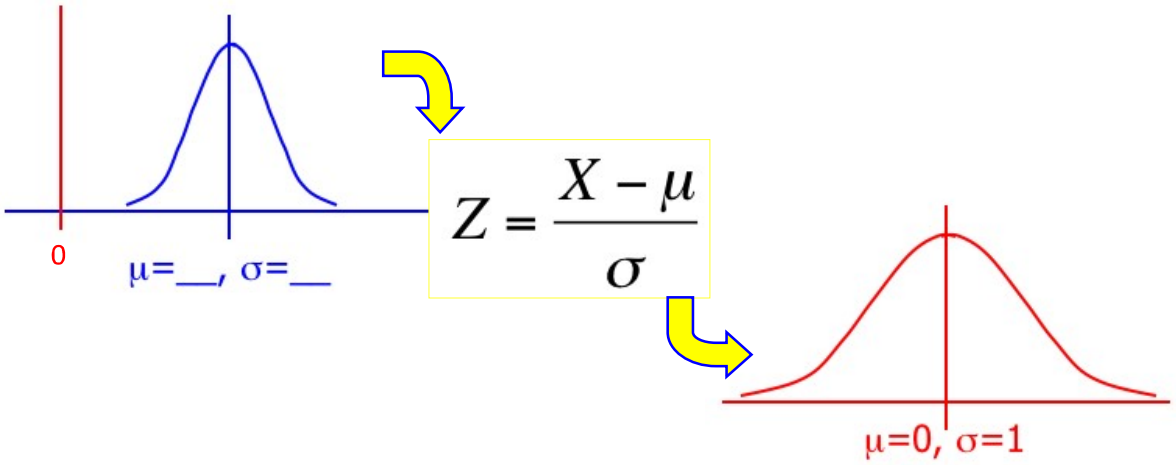
Z=1.51

Z=1.51

Area is 93.45%

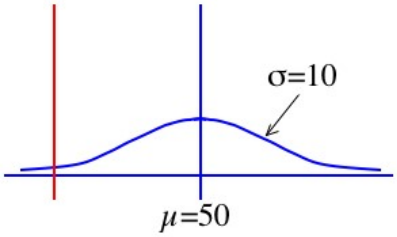
Calculating Normal Probabilities

➢ We can use the following function to convert any normal random variable to a standard normal random variable



Calculating Normal Probabilities

Example: The time required to build a computer is normally distributed with a mean of 50 minutes and a standard deviation of 10 minutes.

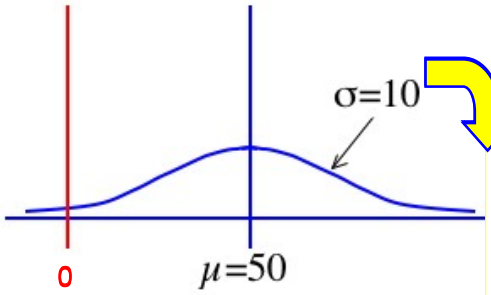


➢ What is the probability that a computer is assembled in a time between 45 and 60 minutes?

➢ Algebraically speaking, what is P(45 < X < 60)?

Calculating Normal Probabilities

P(45 < X < 60)?



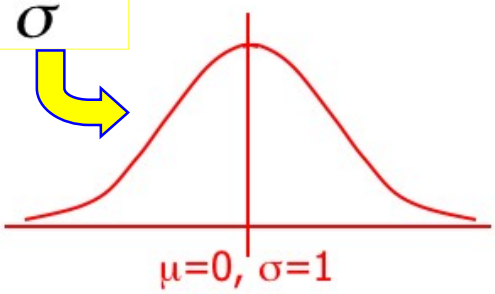
Mean of 50 minutes and a standard deviation of 10 minutes…

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

$$P(45 < X < 60) =$$

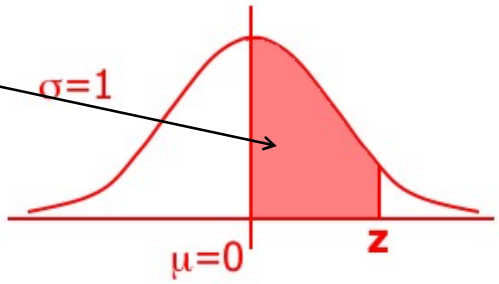
$$P\left(\frac{45-50}{10} < \frac{X-\mu}{\sigma} < \frac{60-50}{10}\right) =$$

$$P(-.5 < Z < 1)$$



Calculating Normal Probabilities

We can use z-table probabilities $P(0 < Z < z)$



We can break up $P(-.5 < Z < 1)$ into:

$P(-.5 < Z < 0) + P(0 < Z < 1)$

The distribution is symmetric around zero, so we have:

$P(-.5 < Z < 0) = P(0 < Z < .5)$

Hence: $P(-.5 < Z < 1) = P(0 < Z < .5) + P(0 < Z < 1)$

Calculating Normal Probabilities

How to use z-table?

This table gives probabilities **P(0 < Z < z)**
First column = integer + first decimal
Top row = second decimal place

P(0 < Z < 0.5)

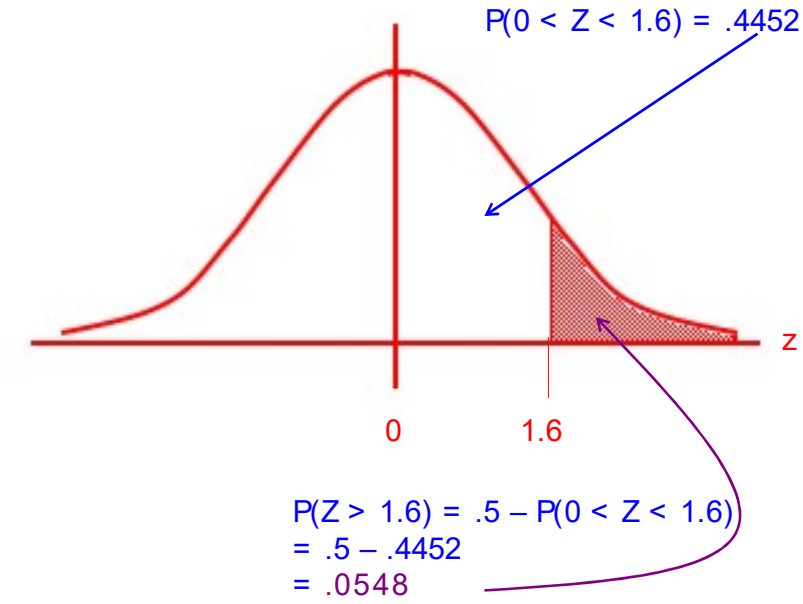
P(0 < Z < 1)

P(–.5 < Z < 1) = .1915 + .3414 = .5328

z	.00	.01	.02	.03
0.0	.0000	.0040	.0080	.0120
0.1	.0398	.0438	.0478	.0517
0.2	.0793	.0832	.0871	.0910
0.3	.1179	.1217	.1255	.1293
0.4	.1554	.1591	.1628	.1664
0.5	.1915	.1950	.1985	.2019
0.6	.2257	.2291	.2324	.2357
0.7	.2580	.2611	.2642	.2673
0.8	.2881	.2910	.2939	.2967
0.9	.3159	.3186	.3212	.3238
1.0	.3413	.3438	.3461	.3485
1.1	.3643	.3665	.3686	.3708
1.2	.3849	.3869	.3888	.3907

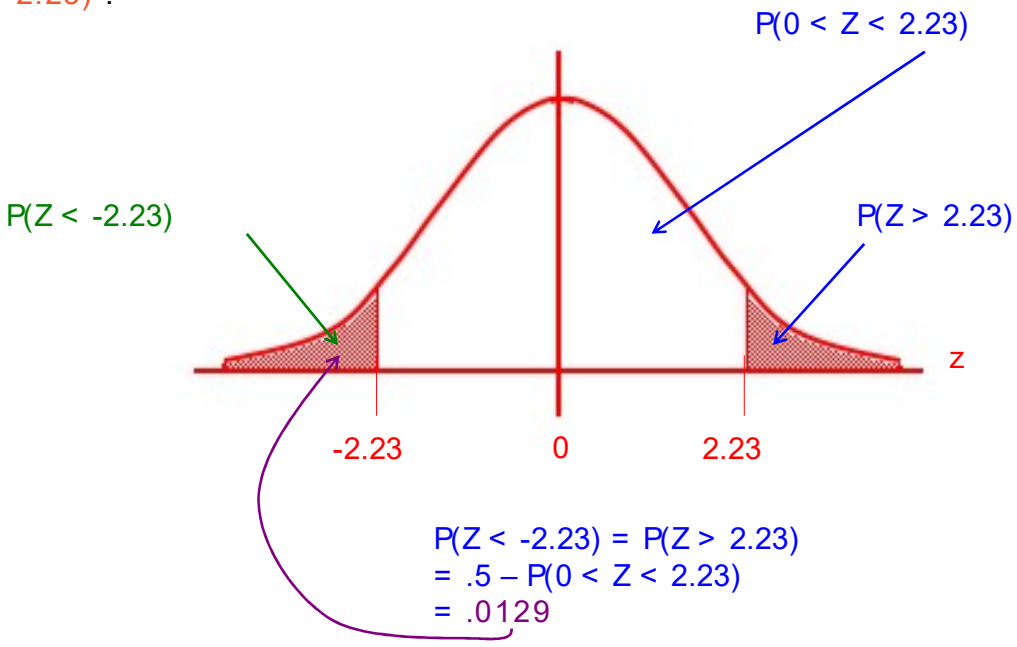
Using the Normal Table

What is P(Z > 1.6) ?



Using the Normal Table

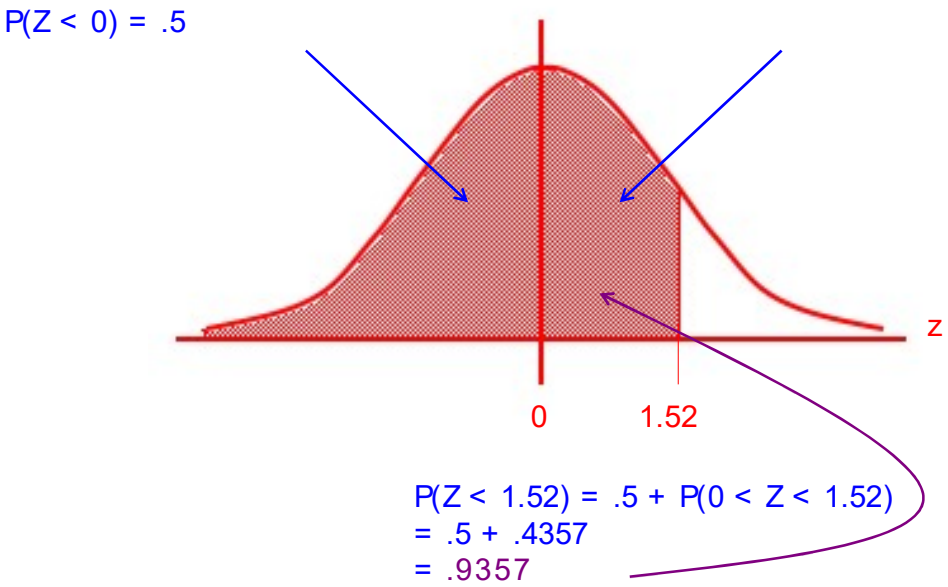
What is P(Z < -2.23) ?



P(Z < -2.23) = P(Z > 2.23)
= .5 – P(0 < Z < 2.23)
= .0129

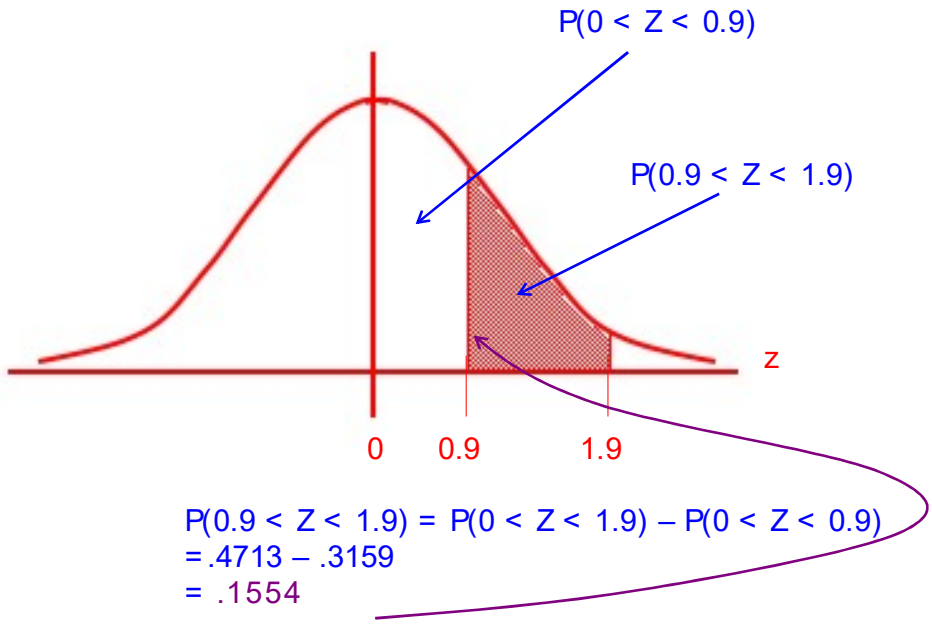
Using the Normal Table

What is P(Z < 1.52) ?



Using the Normal Table

What is $P(0.9 < Z < 1.9)$?



$P(0.9 < Z < 1.9) = P(0 < Z < 1.9) - P(0 < Z < 0.9)$
$= .4713 - .3159$
$= .1554$

Finding Values of Z

➢ The other values of Z are:

- Z.05 = 1.645
- Z.01 = 2.33

Using the Values of Z

Because $z_{.025} = 1.96$ and $-z_{.025} = -1.96$, it follows that we can state:

$P(-1.96 < Z < 1.96) = .95$

Similarly

$P(-1.645 < Z < 1.645) = .90$

If $X \sim N(25,36)$, calculate:

(i) $P(X < 28)$

(ii) $P(X > 30)$

(iii) $P(X < 20)$

(iv) $P(|X - 25| < 4)$

Solution

(i) $P(X < 28) = P\left(Z < \frac{28-25}{\sqrt{36}}\right) = P(Z < 0.5) = 0.69146$

The following answers use interpolation on the tables:

(ii) $P(X > 30) = P(Z > 0.833) = 1 - P(Z < 0.833) = 1 - 0.79758 = 0.20242$

(iii) $P(X < 20) = P(Z < -0.833) = 1 - P(Z < 0.833) = 1 - 0.79758 = 0.20242$

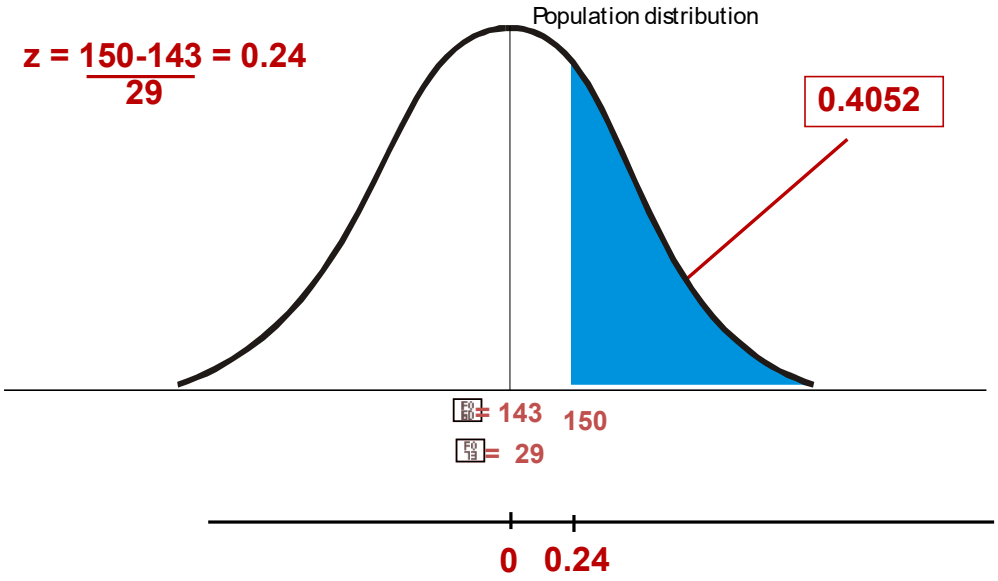
(iv) We need to simplify the expression involving the absolute value:

$$\begin{aligned} P(|X - 25| < 4) &= P(-4 < X - 25 < 4) \\ &= P(21 < X < 29) \\ &= P(X < 29) - P(X < 21) \\ &= P(Z < 0.667) - P(Z < -0.667) \\ &= P(Z < 0.667) - [1 - P(Z < 0.667)] \\ &= 0.49522 \end{aligned}$$

3.c. Continuous PDF- Normal Distribution (Contd.)

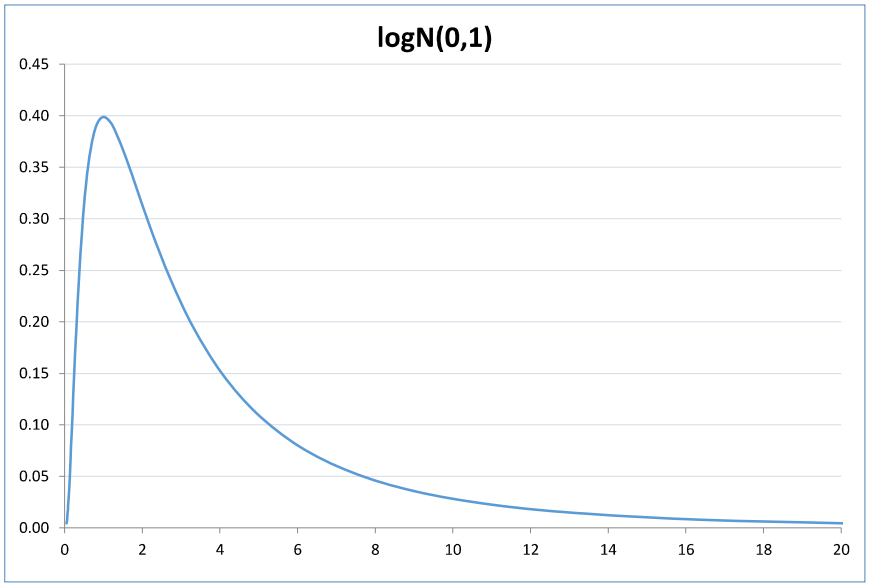
Given the population of women has normally distributed weights with a mean of 143 lbs and a standard deviation of 29 lbs,

1. if one woman is randomly selected, find the probability that her weight is greater than 150 lbs.



3.c. Continuous PDF- Lognormal Distribution

- Plotting PDFs for lognormal (0,1) distributions using MS Excel.



The Distribution of Sample Means (cont.)

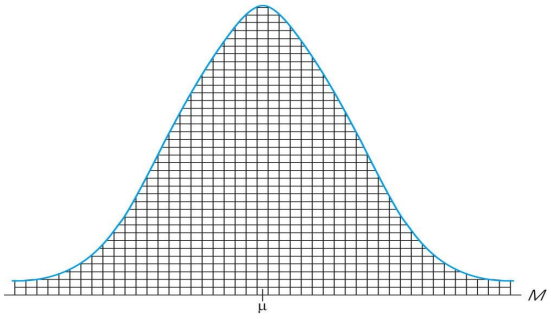
The concept of the distribution of sample means and its characteristics should be intuitively reasonable:

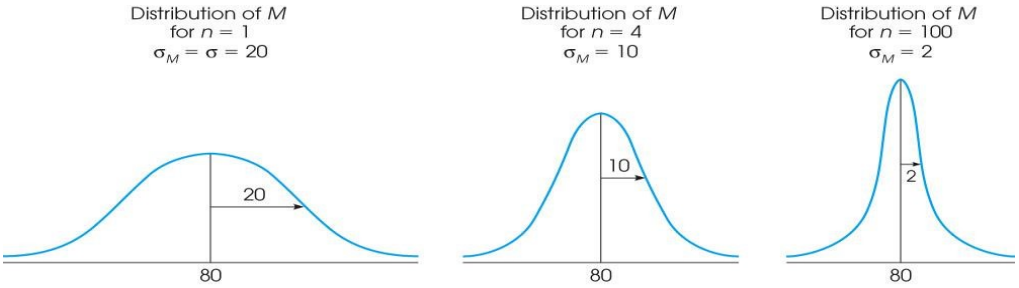
1. You should realize that sample means are variable. If two (or more) samples are selected from the same population, the two samples probably will have different means.

2. Although the samples will have different means, you should expect the sample means to be close to the population mean. That is, the sample means should "pile up" around μ. Thus, the distribution of sample means tends to form a normal shape with an expected value of μ.

3. You should realize that an individual sample mean probably will not be identical to its population mean; that is, there will be some "error" between M and μ. Some sample means will be relatively close to μ and others will be relatively far away. The standard error provides a measure of the standard distance between M and μ.

The Distribution of Sample Means (cont.)





A key Distinction

Population Distribution – distribution of all individual scores in the population

Sample Distribution – distribution of all the scores in your sample

Sampling Distribution – distribution of all the possible sample means when taking samples of size n from the population. Also called "the distribution of sample means".

Central Limit Theorem

➢ The sampling distribution of the mean of a random sample drawn from any population is approximately normal for a sufficiently large sample size.

➢ The larger the sample size, the more closely the sampling distribution of 'X' will resemble a normal distribution.

Central Limit Theorem

- If the population is normal, then 'X' is normally distributed for all values of n.
- If the population is not-normal, then 'X' is approximately normal only for larger values of n.
- In many practical situations, a sample size of 30 may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of X.

Sampling Distributions of the Sample Mean

1. $\mu_{\bar{x}} = \mu$

2. $\sigma_{\bar{x}}^2 = \sigma^2 / n \quad and \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$

3. If X is normal, X is normal. If X is not normal then X is approximately normal for sufficiently large sample sizes.

Note:

➢ The definition of "sufficiently large" depends on the extent of non-normality of X.

➢ For example, heavily skewed; multimodal

For Example,

The foreman of a bottling plant has observed that the amount of soda in each "32-ounce" bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of 0.3 ounce.

If a customer buys one bottle, what is the probability that the bottle will contain more than 32 ounces?

Standard Error

The "standard error" of the mean is:

The standard deviation of the distribution of sample means.

The standard error measures the standard amount of difference between x-bar and [illegible] that is reasonable to expect simply by chance.

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

Practical Rules Commonly Used

1. For samples of size n larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets better as the sample size n becomes larger.

2. If the original population is itself normally distributed, then the sample means will be normally distributed for <u>any sample size.</u>

	small n	large n
normal population	$\overline{X}$ is normal	$\overline{X}$ is normal
non-normal population	$\overline{X}$ is nonnormal	$\overline{X}$ is normal

Example:

Given the population of women has normally distributed weights with a mean of 143 lbs and a standard deviation of 29 lbs,

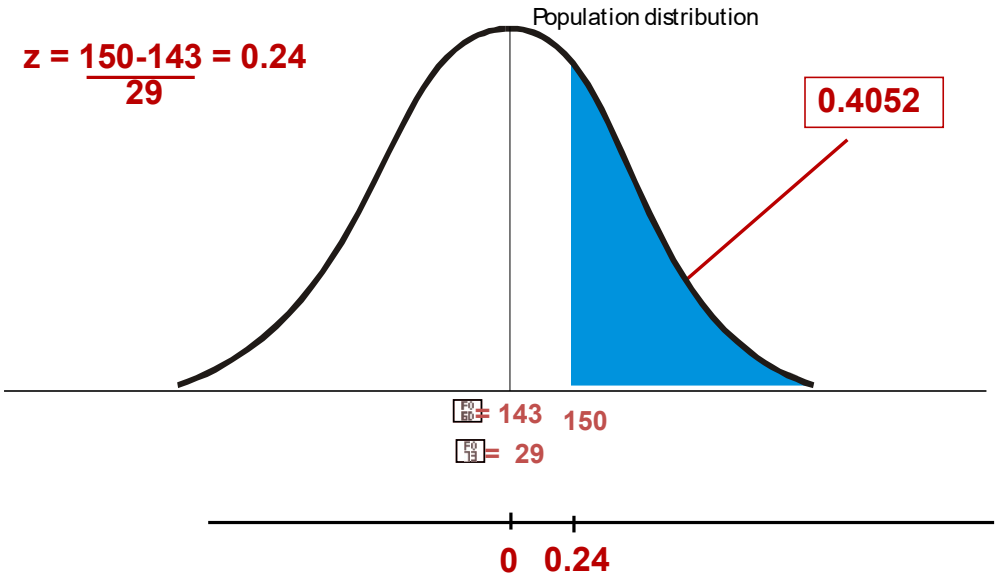
1. if one woman is randomly selected, find the probability that her weight is greater than 150 lbs.

2. if 36 different women are randomly selected, find the probability that their mean weight is greater than 150 lbs.

Probability and the Distribution of Sample Means (contd.)

Given the population of women has normally distributed weights with a mean of 143 lbs and a standard deviation of 29 lbs,

1. if one woman is randomly selected, find the probability that her weight is greater than 150 lbs.



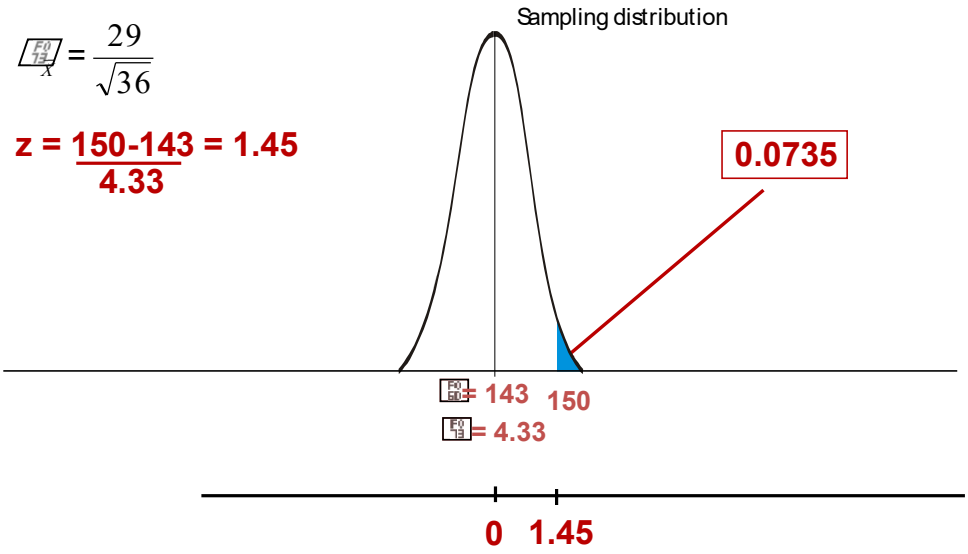
Probability and the Distribution of Sample Means (contd.)

Given the population of women has normally distributed weights with a mean of 143 lbs and a standard deviation of 29 lbs,

2. if 36 different women are randomly selected, find the probability that their mean weight is greater than 150 lbs.

$$\sigma_{\bar{x}} = \frac{29}{\sqrt{36}}$$

$$z = \frac{150-143}{4.33} = 1.45$$



Probability and the Distribution of Sample

Given the population of women has normally distributed weights with a mean of 143 lbs and a standard deviation of 29 lbs,

1. if one woman is randomly selected, find the probability that her weight is greater than 150 lbs. $P(X > 150) = .41$

2. if 36 different women are randomly selected, find the probability that their mean weight is greater than 150 lbs. $P(\bar{X} > 150) = .07$

Example

It is assumed that the number of claims arriving at an insurance company per working day has a mean of 40 and a standard deviation of 12. A survey was conducted over 50 working days. Calculate the probability that the sample mean number of claims arriving per working day was less than 35.

Solution

Using the notation given in Core Reading, $\mu = 40$, $\sigma = 12$, $n = 50$.

The Central Limit Theorem states that $\overline{X} \sim N(40, 12^2/50)$.

We want $P(\overline{X} < 35)$:

$$P(\overline{X} < 35) \approx P\left(Z < \frac{35 - 40}{\sqrt{12^2/50}}\right)$$

$$= P(Z < -2.946) = 1 - P(Z < 2.946) = 1 - 0.99839 = 0.00161$$

3.d. The Central Limit Theorem

The cost of repairing a vehicle following an accident has mean \$6,200 and standard deviation \$650. A study was carried out into 65 vehicles that had been involved in accidents. Calculate the probability that the total repair bill for the vehicles exceeded \$400,000.

The cost of repairing a vehicle following an accident has mean \$6,200 and standard deviation \$650. A study was carried out into 65 vehicles that had been involved in accidents. Calculate the probability that the total repair bill for the vehicles exceeded \$400,000.

Using the notation given in Core Reading, we have $\mu = 6,200$, $\sigma = 650$, $n = 65$. Also let $Z \sim N(0,1)$.

We want the probability that the total repair bill, T is greater than 400,000. The Central Limit Theorem states that:

$$T \approx N(65 \times 6200,\ 65 \times 650^2) = N(403000,\ 5240^2)$$

So the probability is found as follows:

$$P(T > 400,000) \approx P\left(Z > \frac{400,000 - 403,000}{5,240}\right)$$

$$= P(Z > -0.572) = P(Z < 0.572) = 0.71634$$

3.d. The Central Limit Theorem

- Introduction:
- It is perhaps one of the most important result in statistics
- It provides the basis for large-sample inference about a population mean when the population distribution is unknown.
- It also provides the basis for large-sample inference about a population proportion, for example, in opinion polls and surveys.
- Definition:
- If $X_1, X_2, \ldots, X_n$ is a sequence of independent, identically distributed (iid) random variables with finite mean μ and finite (non-zero) variance σ^2 then the distribution of $(\langle X \rangle - \mu)/(\sigma/\sqrt{n})$ approaches the standard normal distribution, $N(0,1)$, as $n \to \infty$
- μ is the population mean from which $X_1, X_2, \ldots, X_n$ have been extracted.
- $\langle X \rangle$ is the sample mean calculated as $\langle X \rangle = (1/n) \sum_{i=1}^{i=n} X_i$
- For large n, $(\langle X \rangle - \mu)/(\sigma/\sqrt{n})$ and $(\sum Xi - n\mu)/(\sqrt{(n\sigma^2)})$ has N(0, 1) distribution
- OR
  - $\langle X \rangle \sim N(\mu, \sigma^2/n)$
  - $\sum Xi \sim N(n\mu, n\sigma^2)$

Email us - support@acadgild.com