03-Inferential Statistics Notes and Examples

Inferential Statistics

- Suppose you **randomly sampled 10 people** from the population of men in Mumbai, between the ages of 21 and 35 years and computed the **mean height** of your sample.
- You would **not expect** your sample mean to be equal to the mean of all men in Mumbai. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly.
- Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.
- **Inferential statistics** concerns generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter.
- In this example, the **sample statistics** are the sample means and the **population parameter** is the population mean

Sampling Distribution

- A **sampling distribution of sample means** is a distribution obtained by using the means computed from random samples of a specific size taken from a population
- Knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean.
- The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean
- **Sampling error** is the difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population

Central Limit Theorem

- *Given a population with a finite mean μ and a finite nonzero variance σ2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ2/N as N, the sample size, increases*
- What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases.

Sampling Distribution

Mean and Standard Deviation of Sample Means

- As the sample size n increases, the shape of the distribution of the sample means taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution

- The mean of the sample means equals the population mean

$$\mu_{\bar{X}} = \mu$$

- The standard deviation of the sample means is called the standard error of the mean

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}.$$

z - score

- The central limit theorem can be used to answer questions about sample means in the same manner that the normal distribution can be used to answer questions about individual values.
- A new formula must be used for the z values:

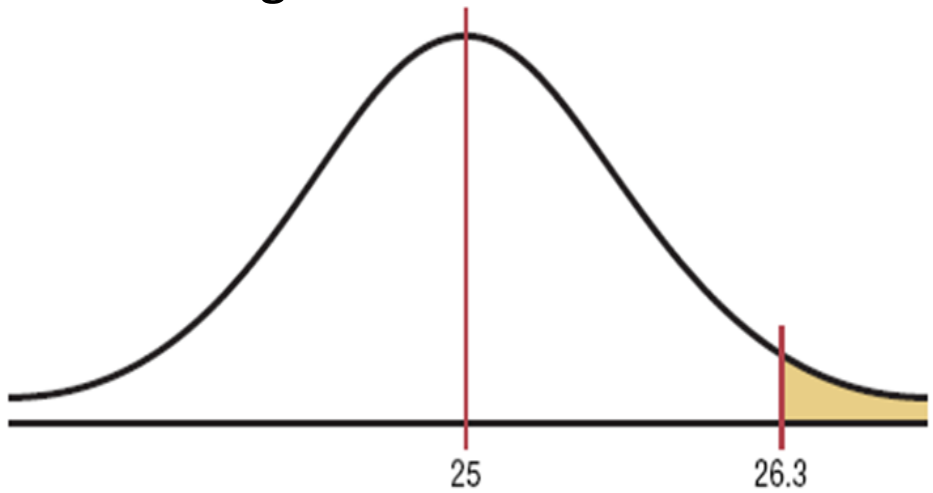
$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Example – Hours of Television

- A. C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week
- Assume the variable is normally distributed and the standard deviation is 3 hours
- If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours

- Since we are calculating probability for a sample mean

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{26.3 - 25}{3/\sqrt{20}} = 1.94$$

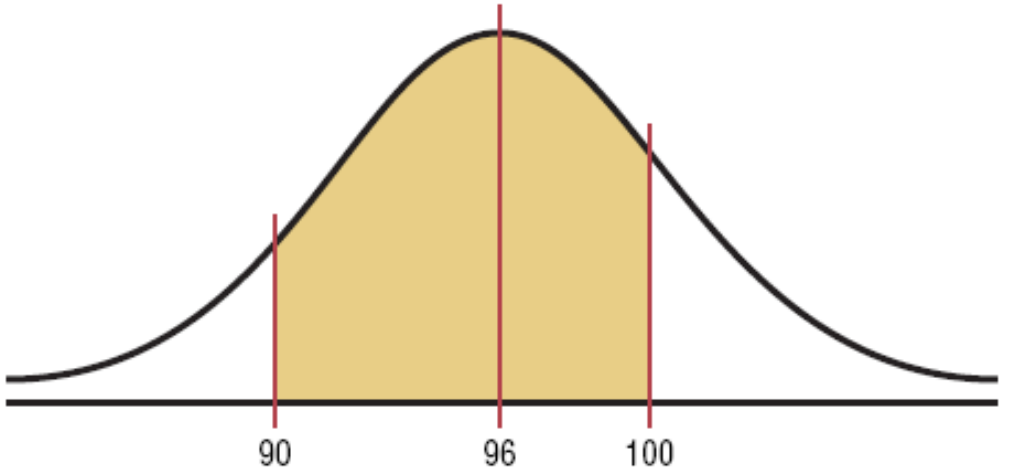


- For z=1.94, from the table, area under the curve=0.9738
- Thus the area is 1.0000 – 0.9738 = 0.0262.
- The probability of obtaining a sample mean larger than 26.3 hours is 2.62%.

Example – Vehicle Age

- The average age of a vehicle registered in the United States is 8 years, or 96 months
- Assume the standard deviation is 16 months
- If a random sample of 36 vehicles is selected, find the probability that the mean of their age is between 90 and 100 months.
- Since the sample is 30 or larger, the normality assumption is not necessary

$$z = \frac{90 - 96}{16/\sqrt{36}} = -2.25 \qquad z = \frac{100 - 96}{16/\sqrt{36}} = 1.50$$



- Table gives us areas 0.9332 and 0.0122, respectively.
- The desired area is 0.9332 - 0.0122 = 0.9210.
- Thus, the probability of obtaining a sample mean between 90 and 100 months is 92.1%.

Statistical Inference and Estimation

- Statistical inference is the act of generalizing from the data (“sample”) to a larger phenomenon (“population”) with calculated degree of certainty.
- The act of generalizing and deriving statistical judgments is the process of inference
- The two common forms of statistical inference are:
  - Estimation
  - Null hypothesis tests of significance (NHTS) – Hypothesis Testing
- There are two forms of estimation
  - Point estimation (maximally likely value for parameter)
  - Interval estimation (also called confidence interval for parameter)

Parameters and Point Estimates

- Both Estimation and Hypothesis Testing are used to infer parameters.
- A parameter is a statistical constant that describes a feature about a phenomena/population
- Examples of parameters include:
  - Expected value μ (also called “the population mean”)
  - Standard deviation σ (also called the “population standard deviation”)
  - Binomial probability of “success” p (also called “the population proportion”)
- Point estimates are single points that are used to infer parameters directly. For example
  - Sample mean x (“x bar”) is the point estimator of μ
  - Sample standard deviation s is the point estimator of σ
  - Sample proportion $\hat{p}$ (“p hat”) is the point estimator of p

Population Parameters – Point Estimates

- One of the major applications of statistics is estimating population parameters from sample statistics
- For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium
- Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition.
- This value of 0.53 is called a **point estimate** of the population proportion. It is called a point estimate because the estimate consists of a single value or point

Interval Estimates

- An interval estimate of a parameter is an interval or a range of values used to estimate the parameter.
- Interval estimates are called **Confidence Intervals**
- A Confidence Interval will have a lower limit and an upper limit – and are called the confidence limits.

Confidence Intervals

- Say you were interested in the mean weight of 10-year-old girls living India. Since it would have been impractical to weigh all the 10-year-old girls in India, you took a sample of 25 and found that the mean weight was 40 kgs.
- This sample mean of 40 is a point estimate of the population mean.
- A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far this sample mean may be from the population mean.
- For example, can you be confident that the population mean is within 2 kgs of 40? You simply do not know
- Confidence intervals provide more information than point estimates.
- Confidence intervals for means are intervals constructed using a procedure that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time.
- These intervals are referred to as 95% and 99% confidence intervals respectively

Confidence Intervals

- Point estimates are usually supplemented by interval estimates called confidence intervals.
- Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time.
- For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \pi < 0.60$.
- The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal.
- The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%.

Confidence Interval of the mean

- Confidence Interval = Observed mean ± Margin of error
- Margin of error = (Confidence Coefficient) × Standard Error of the mean

$$\overline{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

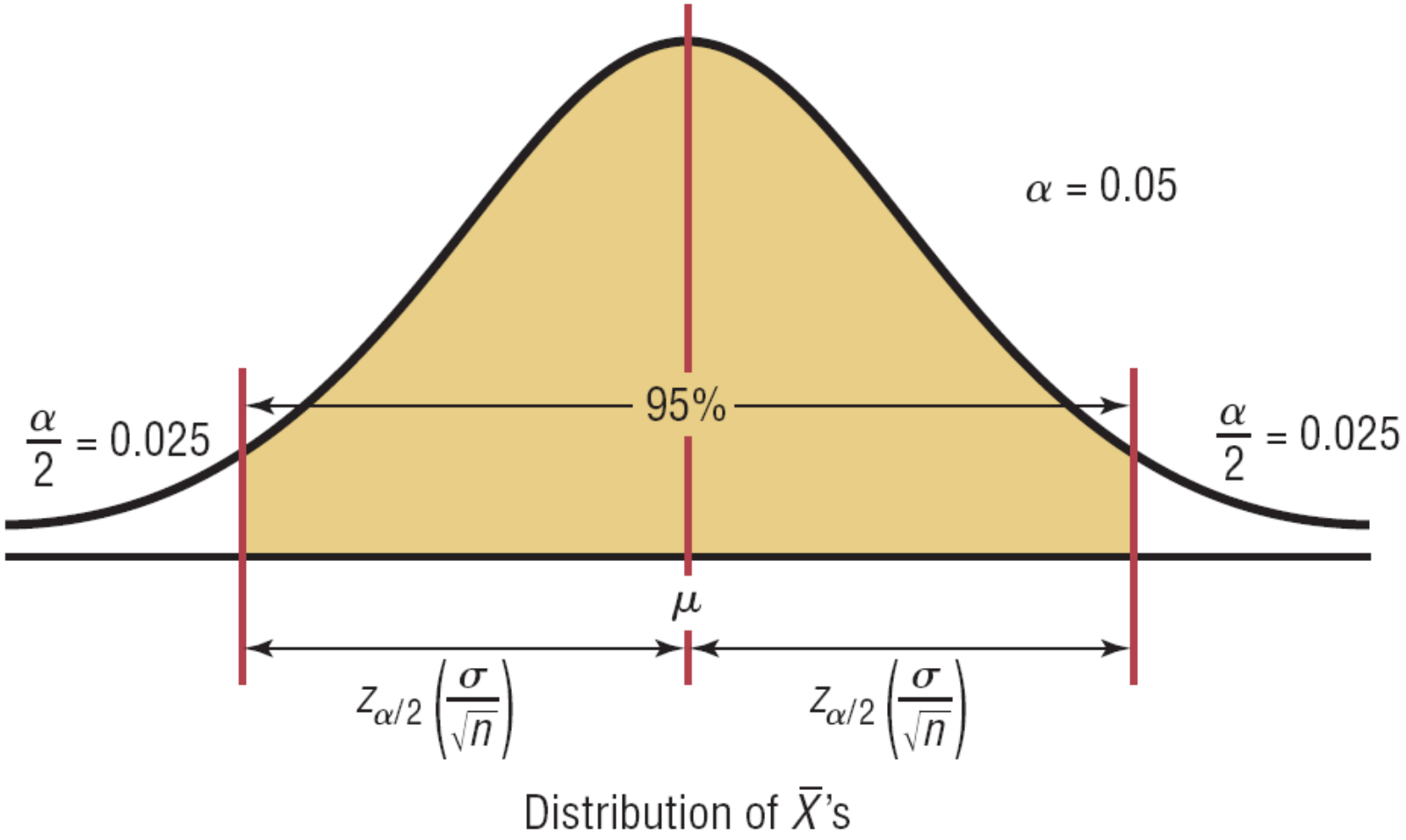
For a 90% confidence interval: $z_{\alpha/2} = 1.65$

For a 95% confidence interval: $z_{\alpha/2} = 1.96$

For a 99% confidence interval: $z_{\alpha/2} = 2.58$

- Relationship between α and the confidence level
  - The Greek letter α(alpha) represents the total area in both tails of the standard normal distribution curve.
  - The stated confidence level is the percentage equivalent of $1 - \alpha$
  - When 95% confidence interval is to be found , $\alpha = 0.05$ since $1 - \alpha = 1 - 0.05 = 0.95$ or 95%

95% confidence interval of the mean



Example: Age of Automobiles

- A survey of 50 adults found that the mean age of a person's primary vehicle is 5.6 years. Assuming the standard deviation of the population is 0.8 year, find the 99% confidence interval of the population mean.

  - $\overline{X} = 5.6$
  - $\sigma = 0.8$
  - For 99% Confidence Level, α= 0.005 on each side of the bell curve
  - z = -2.58 on left side and +2.58 on right side (from the table)
  - $Confidence\ Interval = \ 5.6 \pm 2.58\left(\frac{0.8}{\sqrt{50}}\right) = 5.6 \pm 0.4 = (5.2,6.0)$
  - One can be 99% confident that the mean age of all primary vehicles is between 5.2 and 6.0 years, based on a sample of 50 vehicles.

z-value in python

- import scipy.stats as st
- area=(1+conf_level)/2
- z=st.norm.ppf(area)

Hypothesis Testing

- Researchers are interested in answering many types of questions. For example,
  - Does a new machine produce parts that are less faulty?
  - Does a new medication lower blood pressure?
  - Does the public prefer a certain color in a new fashion line?
  - Do seat belts reduce the severity of injuries?
- These types of questions can be addressed through statistical hypothesis testing, which is a decision-making process for evaluating claims about a population.

Hypothesis Testing

- Hypothesis testing refers to the process of making inferences or *educated guesses about a particular parameter*.
- A hypothesis consists of a “Null Hypothesis” and an “Alternative Hypothesis”
  - **Null Hypothesis (H0)** – a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.
  - **Alternative Hypothesis (H1)** – a statistical hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters.

Hypothesis Testing

- When a hypothesis test proves the "alternative hypothesis," then the original hypothesis (the "null hypothesis") is overturned or "rejected."
- You must decide the level of statistical significance in your hypothesis, as you can never be 100 percent confident in your findings

Hypothesis testing

two-tailed

- A medical researcher is interested in finding out whether a new medication will have any undesirable side effects on the pulse rate of the patients.
- Will the pulse rate ***increase, decrease, or remain unchanged*** after a patient takes the medication?
- The researcher knows that the mean pulse rate for the population under study is 82 beats per minute.
- The hypotheses for this situation are

$$H_0 : \mu = 82 \qquad H_1 : \mu \neq 82$$

- This is called a two-tailed (or non directional)hypothesis test.

Hypothesis Testing

one sided right-tailed

- A researcher thinks that if expectant mothers use vitamins, the birth weight of the babies will increase. The average birth weight of the population is 8.6 pounds

- The hypotheses for this situation are
  - $\mathbf{H_0: \mu = 8.6}$
  - $\mathbf{H_1: \mu > 8.6}$

- This is called a ***right-tailed*** (one directional)hypothesis test.

Hypothesis Testing

one sided left-tailed

- A contractor wishes to lower heating bills by using a special type of insulation in houses. If the average of the monthly heating bills is $78, her hypotheses about heating costs with the use of insulation are

- The hypotheses for this situation are

$$\boldsymbol{H}_0 : \ \boldsymbol{\mu} = 78 \qquad \boldsymbol{H}_1 : \ \boldsymbol{\mu} < 78$$

- This is called a ***left-tailed*** (one directional) hypothesis test.

Hypothesis Testing

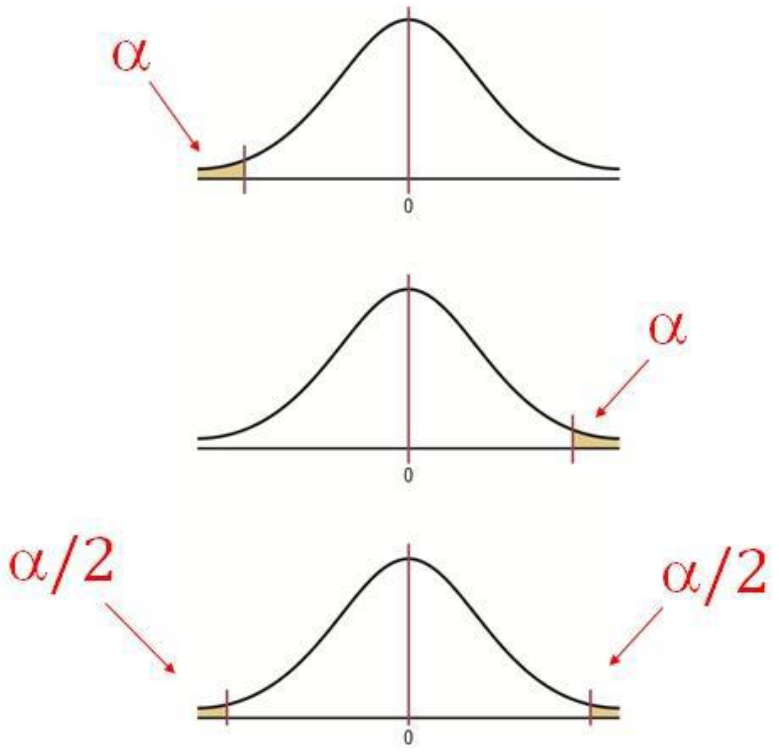
- After stating the hypotheses, the researcher designs the study
  - Select the correct statistical test
  - Choose an appropriate **level of significance (α)**
  - Formulate a plan for conducting the study.
- **Statistical Test** – uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected.
- **Test Value (test statistic)** – the numerical value obtained from a statistical test.
- Typical significance levels are:
  - 0.10, 0.05, and 0.01
  - Corresponding to confidence levels 0.90, 0.95, 0.99

Hypothesis Testing – Critical Value

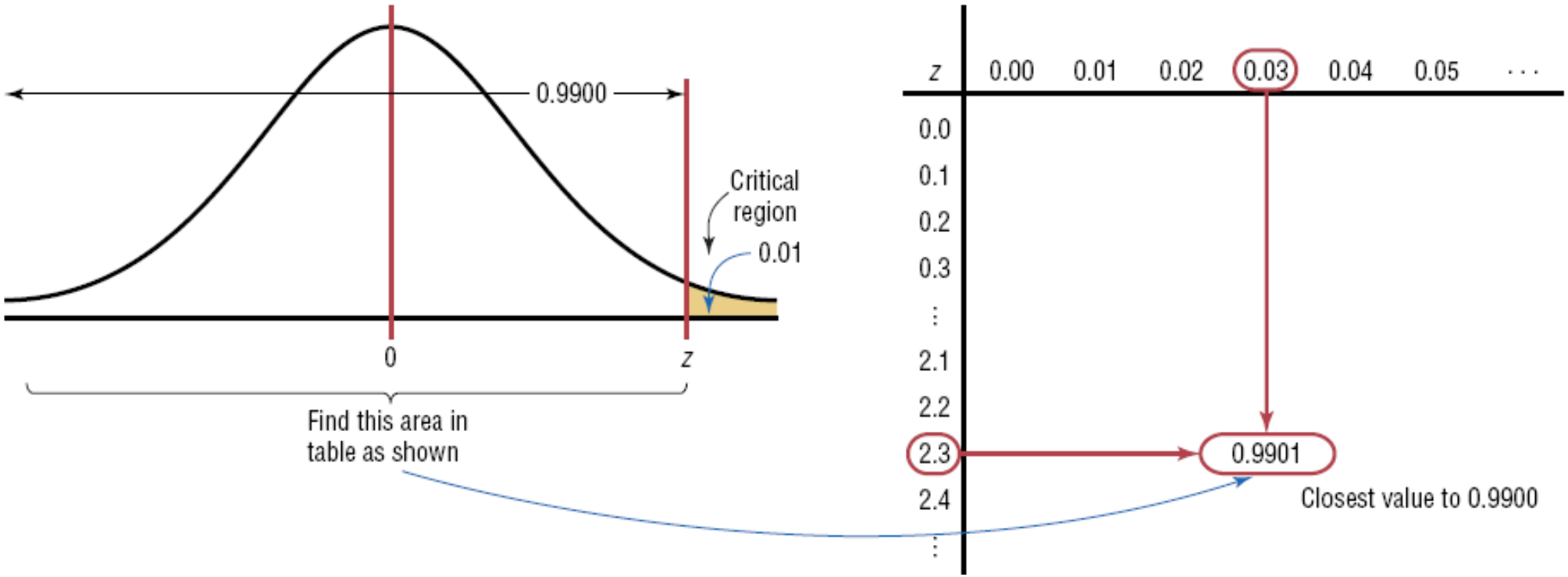
- **Critical or Rejection Region** – the range of values for the test value that indicate a significant difference and that the null hypothesis should be rejected.
- **Non-critical or Non-rejection Region** – the range of values for the test value that indicates that the difference was probably due to chance and the null hypothesis should not be rejected
- **Critical Value (CV)** – separates the critical region from the non-critical region, i.e., when we should reject H_0 from when we should not reject H_0.

Critical value

- **One-tailed test** – indicates that the null hypothesis should be rejected when the test value is in the critical region on one side.
  - **Left-tailed test** – when the critical region is on the left side of the distribution of the test value.
  - **Right-tailed test** – when the critical region is on the right side of the distribution of the test value.
- **Two-tailed test** – the null hypothesis should be rejected when the test value is in either of two critical regions on either side of the distribution of the test value

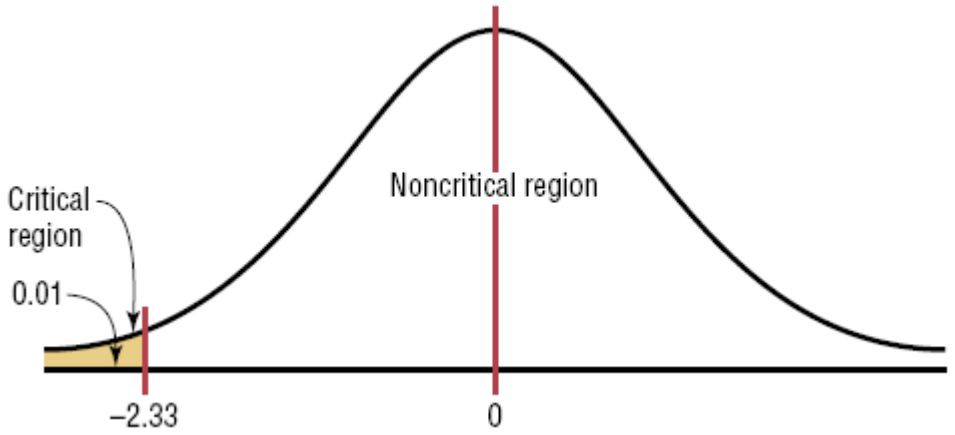


Finding the Critical Value for $\alpha = 0.01$ (Right-Tailed Test)



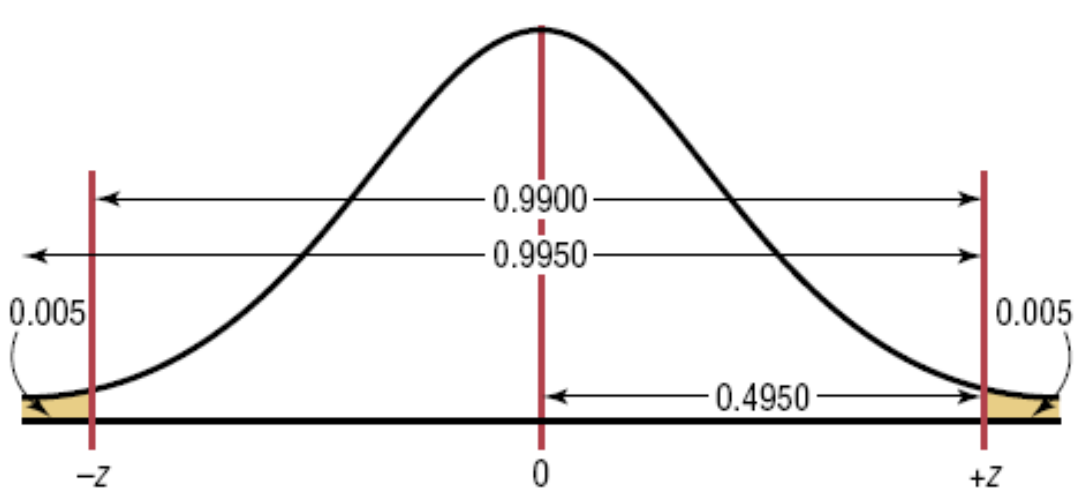
- $z = 2.33$ for $\alpha = 0.01$ (Right-Tailed Test)

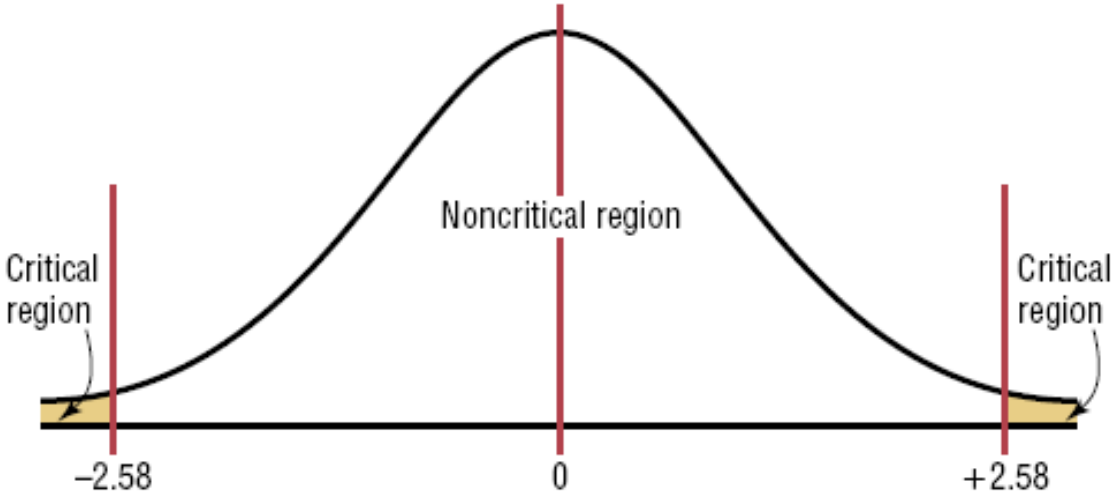
Finding the Critical Value for $\alpha = 0.01$ (Left-Tailed Test)



- z = -2.33 for $\alpha = 0.01$ (Left-Tailed Test)

Finding the Critical Value for $\alpha = 0.01$ (Two-Tailed Test)





$z = \pm 2.58$ for $\alpha = 0.01$ (Two-Tailed Test)

Hypothesis Test - Steps

- Step 1 State the hypotheses and identify the claim.
- Step 2 Find the critical value(s) from the appropriate table.
- Step 3 Compute the test value.
- Step 4 Make the decision to reject or not reject the null hypothesis.
- Step 5 Summarize the results.

Hypothesis testing - example

- It has been reported that the average credit card debt for college seniors is $3262
- The student senate at a large university feels that their seniors have a debt much less than this, so it conducts a study of 50 randomly selected seniors .
- It finds that the average debt is $2995, and the population standard deviation is $1100
- Let’s conduct the test based on Significant value α=0.05.

Hypothesis testing - example

- Step 1: H_0: $\mu = \$3262$ H_1: $\mu < \$3262$
- Step 2: Find the critical value(s) from the table.
  - Left-tailed test, $\alpha=0.05$ => z will be negative and have probability 0.05 underneath it
  - Z = -1.65
- Step 3: Compute the test value z $= \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
  - Z = (2995 -3262)/1100/√50 = -1.716
- Step 4: Make the decision to reject or not reject H_0
  - Since this is a left-tailed test, our rejection region consists of values of Z that are smaller than our critical value of Z = -1.65
  - Since our test value (-1.716) is less than our critical value (-1.65), we reject the null hypothesis
- Step 5: Summarize the results.
  - We have evidence to support the student senate claim that the university's seniors have credit card debt that is less than the reported average debt.
  - This is based on a Type I error rate of 0.05. This means we falsely make the claim above 5% of the time.

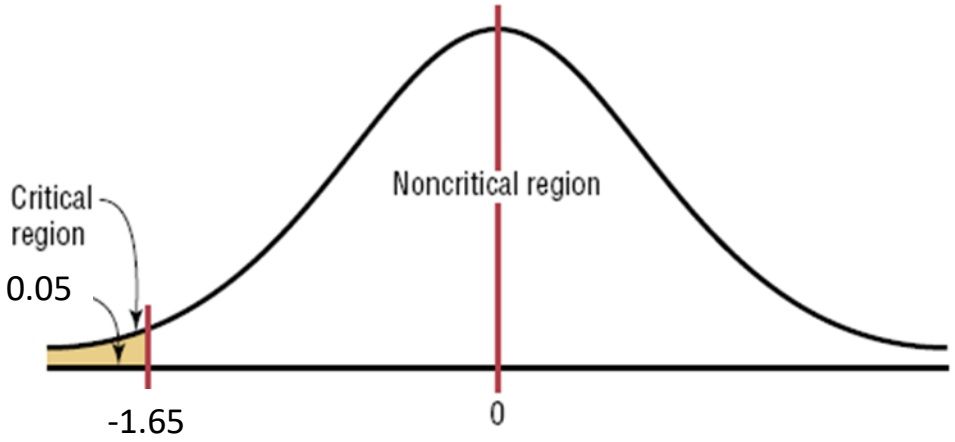


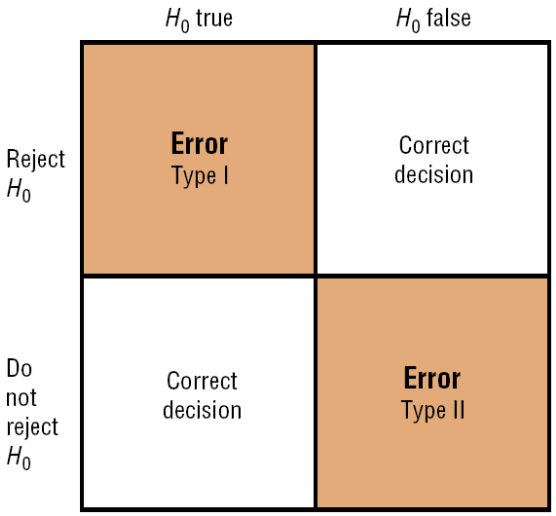
Table E The Standard Normal Distribution

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05
−3.4	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0005	.0005	.0005	.0004	.0004	.0004
−3.2	.0007	.0007	.0006	.0006	.0006	.0006
−3.1	.0010	.0009	.0009	.0009	.0008	.0008
−3.0	.0013	.0013	.0013	.0012	.0012	.0011
−2.9	.0019	.0018	.0018	.0017	.0016	.0016
−2.8	.0026	.0025	.0024	.0023	.0023	.0022
−2.7	.0035	.0034	.0033	.0032	.0031	.0030
−2.6	.0047	.0045	.0044	.0043	.0041	.0040
−2.5	.0062	.0060	.0059	.0057	.0055	.0054
−2.4	.0082	.0080	.0078	.0075	.0073	.0071
−2.3	.0107	.0104	.0102	.0099	.0096	.0094
−2.2	.0139	.0136	.0132	.0129	.0125	.0122
−2.1	.0179	.0174	.0170	.0166	.0162	.0158
−2.0	.0228	.0222	.0217	.0212	.0207	.0202
−1.9	.0287	.0281	.0274	.0268	.0262	.0256
−1.8	.0359	.0351	.0344	.0336	.0329	.0322
−1.7	.0446	.0436	.0427	.0418	.0409	.0401
−1.6	.0548	.0537	.0526	.0516	.0505	.0495

Statistical Test – Type I & Type II error

- Based on the statistical test, a decision is made to reject H0 or not to reject H0 based on the data obtained from a sample
- There are two types of errors that we could make
  - Type I error – reject H0 when H0 is true.
  - Type II error – do not reject H0 when H0 is false
- The maximum probability of committing a type I error is called **level of significance** and is symbolized by α (alpha).
  - $P(\text{Type I error} \mid H0 \text{ is true}) = \alpha$
  - For example, when a = 0.10, there is a 10% chance of rejecting a true null hypothesis
- Likewise
  - P(Type II error | H1 is false) = β (beta).
  - 1- β is called statistical power



p-Value Method for Hypothesis Testing

- The p-Value (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true
- If the null hypothesis was true, a small p-value means that there is a low probability of seeing a point estimate as the one we saw.
  - We interpret this as strong evidence in favor of the alternative.
- **If *P*-value < α(usually 0.05) then reject H_0 , else retain H_0**

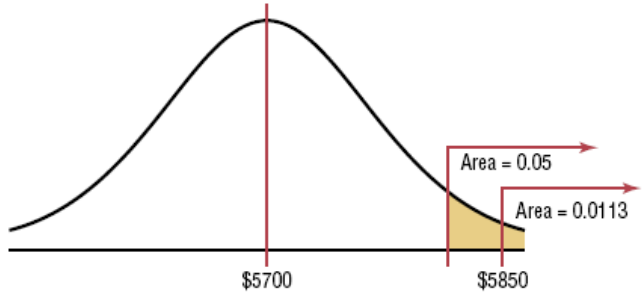
p-value method for Hypothesis Testing

- P values evaluate how well the sample data support the argument that the null hypothesis is true
- It measures how compatible your data are with the null hypothesis. How likely is the effect observed in your sample data if the null hypothesis is true?
  - High P values: your data are **likely supporting** the null hypothesis
  - Low P values: your data are **unlikely to support** the null hypothesis
- A low P value suggests that your sample provides enough evidence that you can reject the null hypothesis for the entire population
- **If *P*-value < α(usually 0.05) then reject H_0 , else retain H_0**

p-value method: example

- A researcher wishes to test the claim that the average cost of tuition and fees at a four-year public college is greater than $5700
- She selects a random sample of 36 four-year public colleges and finds the mean to be $5950
- The population standard deviation is $659. Is there evidence to support the claim at a 0.05? Use the P-value method.

- Step 1: State the hypotheses and identify the claim.
  - H_0: μ = $5700 and H_1: μ > $5700 (claim)
- Step 2: Compute the test value $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{5950 - 5700}{659/\sqrt{36}} = 2.28$
- Step 3: Find the P-value.
  - From the table, for z = 2.28, the area is 0.9887
  - p-value (area of tail) = 1 - 0.9887 = 0.0113



- Step 4: Make the decision.
  - Since the P-value is less than 0.05, the decision is to reject the null hypothesis.
- Step 5: Summarize the results.
  - There is enough evidence to support the claim that the tuition and fees at four-year public colleges are greater than $5700.
- Note: If α = 0.01, the null hypothesis would not be rejected.

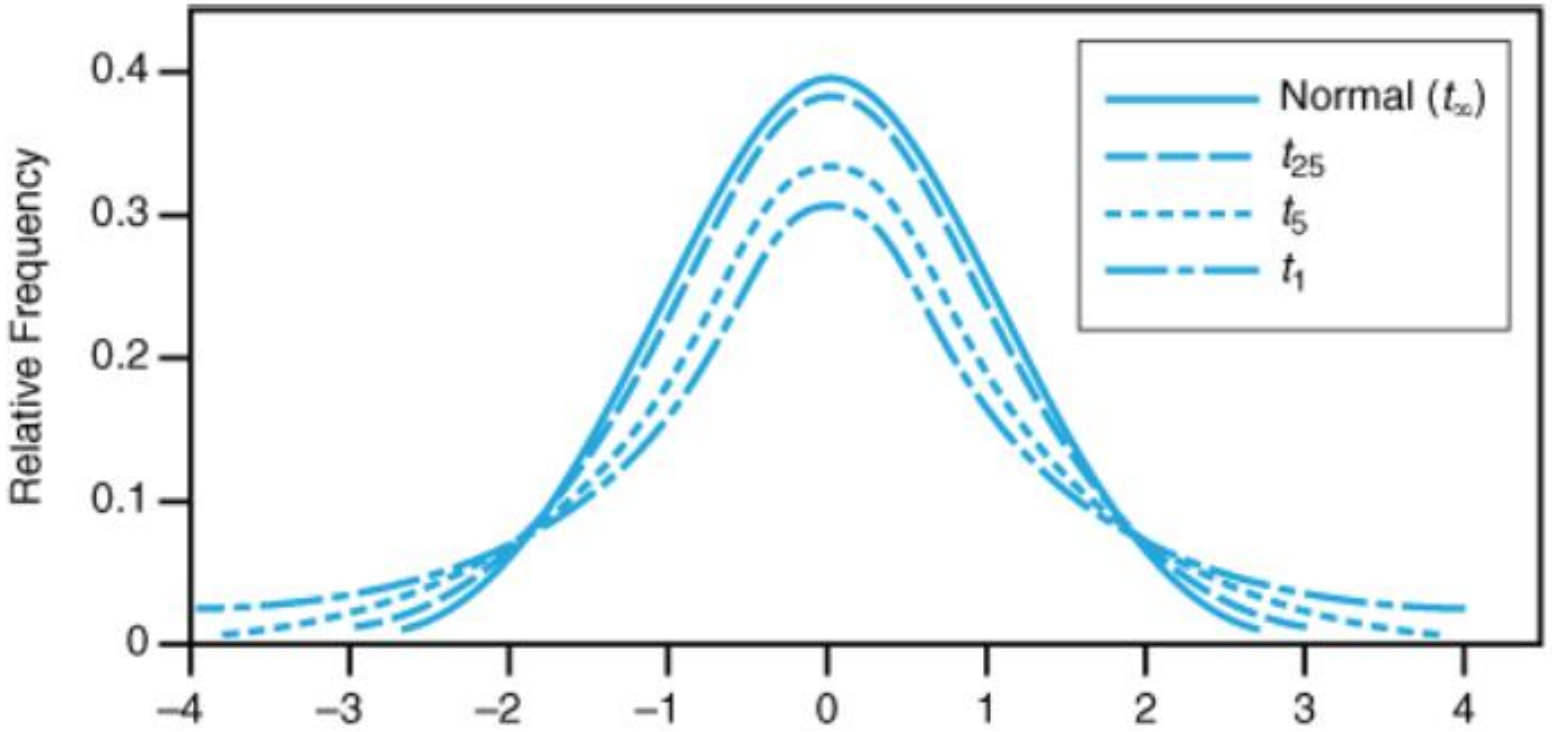
t-test for a mean

- When the population standard deviation (σ) is not known we use the sample deviation (s) in it's place.
- When s is used, especially when the sample size is small (less than 30) the means are no longer normally distributed.
- Now we use the **t-distribution** (also called Student's t-distribution)

t-distribution

- t-distribution is similar to standard normal distribution (bell curve, symmetric about mean, never touches x-axis)
- t-distribution:
  - The variance is greater than 1.
  - The t distribution is actually a family of curves based on the concept of degrees of freedom, which is related to sample size
  - The degrees of freedom for a confidence interval for the mean d.f.=n-1, where n is the sample size
- As the sample size increases, the t distribution approaches the standard normal distribution
- When $n \geq 30$ and population standard deviation is known we can use either *t* or *z* distribution.
- But if $n<30$ and population standard deviation is unknown, we use t distribution.

t-distribution with 1,5,25 d.f.



Confidence Interval for the Mean When σ Is Unknown & *n< 30*

$$\overline{X} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

t-table

- Find the critical *t* value for α=0.01 with sample size of 13 for a left-tailed test
  - Left tailed means the critical t value will be negative
  - n=13 means the degrees of freedom are n-1 = 12
  - The critical value is -2.681

Table F The *t* Distribution

d.f.	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977

t-value in Python

- from scipy.stats import t
- area=(1+conf_level)/2
- df=n-1 # degree of freedom
- t=t.ppf(area, df)

Normal Distribution (68-95-99.7 Rule)

For a normal distribution:

- 68% of the observations are within +/- one standard deviation of the mean
- 95% are within +/- two standard deviations
- and 99.7% are within +- three standard deviations