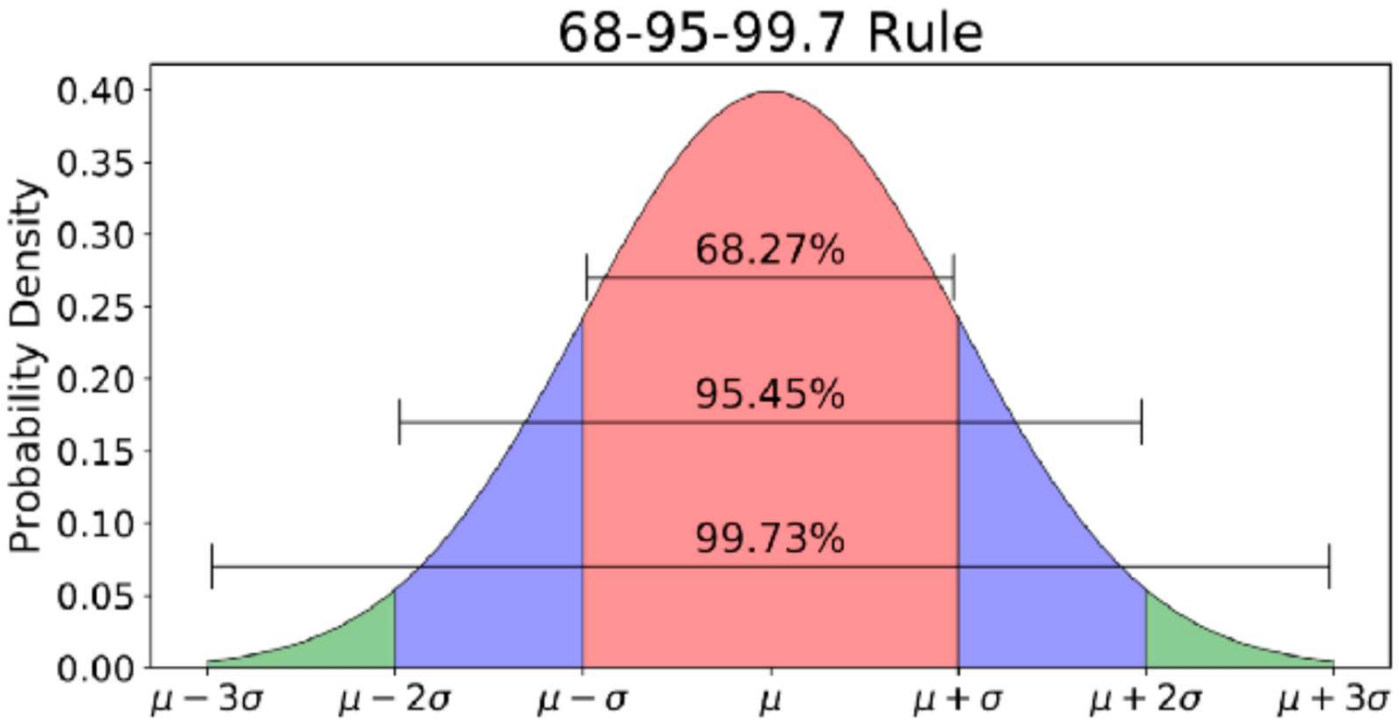
Statistical Inference

Sessions 7

Normal Distribution



Sampling Distribution

- A sampling distribution of the mean is the distribution of all possible sample means if one selected all possible samples of a given size.
- According to the Central Limit Theorem, the Sampling Distribution is centred around mean μ and the standard deviation (standard error) is given as:

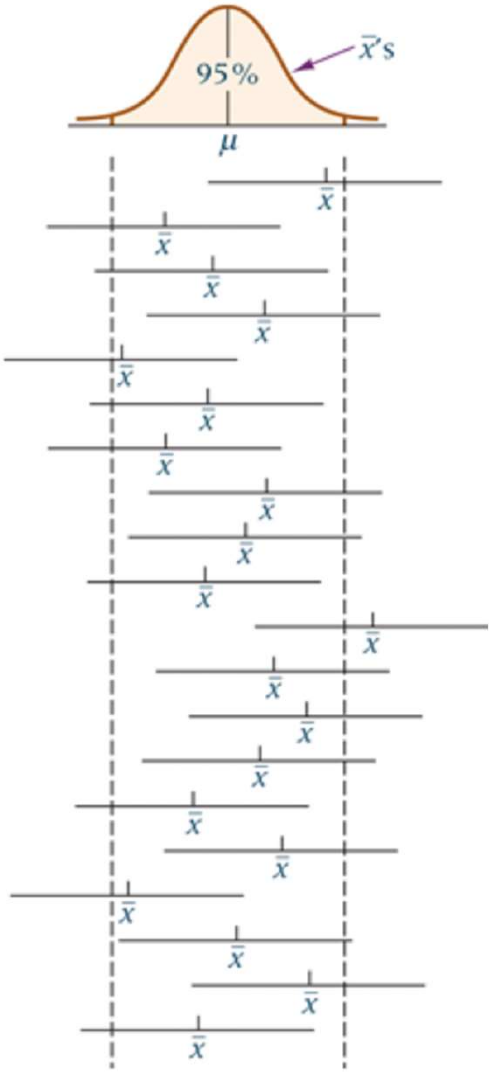
$$S_x = \frac{\sigma}{\sqrt{n}}$$

Application

Suppose a large cellular phone company in wanting to meet the needs of cell phone users hires a business analytics company to estimate the average number of texts used per month by users in the 35-to-54-years-of-age category. The analytics company studies the phone records of 85 randomly sampled users in the 35-to-54-years-of-age category and computes a sample monthly mean of 1300 texts. This mean ($\bar{X}$), which is a statistic, is used to estimate the population mean (μ), which is a parameter. Suppose the population standard deviation (σ) is 160. What is the monthly average number of texts sent by the target group in the population?

- Do we know the μ?
- But we know that $\bar{X}$ is normally distributed around μ
- We use this property to estimate the μ $\quad \mu = \bar{X} \pm z\frac{\sigma}{\sqrt{n}}$

Application

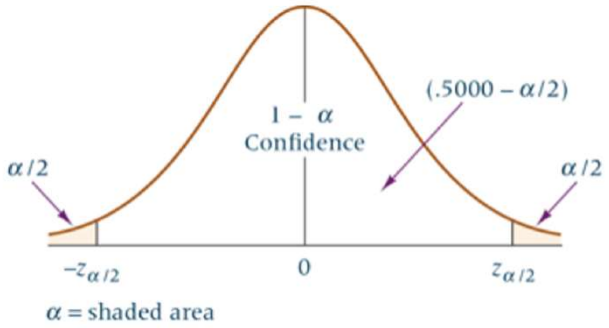


Application

- How much confidence do I need in my estimate? *(Managerial Decision)*
- Suppose 95% confidence
- For alpha = 0.05 → Confidence interval = 100(1 - α)%
- Find z $_{\alpha/2}$ from standard normal table / excel [=NORM.INV(0.025, 0, 1)]

$$\mu = 1300 \pm 1.96\frac{160}{\sqrt{85}}$$

$$1265.99 \leq \mu \leq 1334.01$$



- What does this mean: If 100 samples are selected and a 95% confidence interval is constructed around each, 95 such samples will contain μ
- What if you want to be more confident about your estimate?

From $\bar{X}$ to μ

- Estimating population mean μ from sample mean $\bar{X}$
- We know that the sample mean ($\bar{X}$) is normally distributed around population mean (μ):

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

- Rearranging:

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

- Since it is possible that error in estimation can be on either side:

$$\mu = \bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

- This is the confidence interval of the estimate.

Interval Estimates

- A **point estimate** is *a statistic taken from a sample that is used to estimate a population parameter*
- A point estimate is only as good as the representativeness of its sample.
- If other random samples are taken from the population, the point estimates derived from those samples are likely to vary.
- Because of variation in sample statistics, estimating a population parameter with an interval estimate is often preferable to using a point estimate.
- An **interval estimate** (confidence interval) is *a range of values within which the analyst can declare, with some confidence, the population parameter lies.*

Application

- A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India? A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation is 7.7 years. Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.
- A study is conducted in a company that employs 800 engineers. A random sample of 50 of these engineers reveals that the average sample age is 34.30 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years. Construct a 98% confidence interval to estimate the average age of all the engineers in this company. *(Use finite correction factor)*

Application

The owner of a large equipment rental company wants to make a rather quick estimate of the average number of days a piece of ditchdigging equipment is rented out per person per time. The company has records of all rentals, but the amount of time required to conduct an audit of *all* accounts would be prohibitive. The owner decides to take a random sample of rental invoices. Fourteen different rentals of ditchdiggers are selected randomly from the files, yielding the following data. She uses these data to construct a 99% confidence interval to estimate the average number of days that a ditchdigger is rented and assumes that the number of days per rental is normally distributed in the population.

Data: 3, 1, 3, 2, 5, 1, 2, 1, 4, 2, 1, 3, 1, 1 (Number of days each ditchdigger was used)

What if σ is not known?

- Estimating confidence intervals for t-distribution

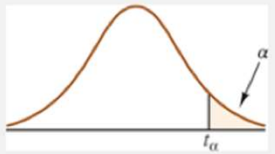
$$\mu = \bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

- Excel Formula: = T.INV(α, n-1)
- Example = T.INV(0.05, 5) = -2.015

 # Note that excel give values from -∞

TABLE A.6

Critical Values from the *t* Distribution



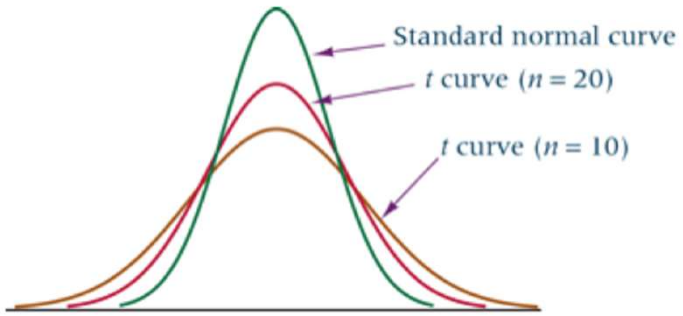
Critical Values of *t*

df	.100	.050	.025	.010	.005	.001
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686

What if σ is not known?

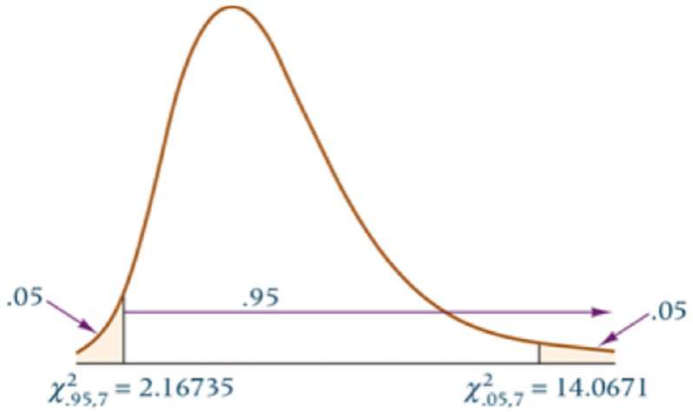
- Use t-distribution instead of normal distribution
- Assumption: Population is normally distributed – However t-distribution is relatively robust to this assumption (can be used despite some violation of assumption)
- Every sample size has different t-distribution
- For large samples t-distribution approximates normal distribution
- Formula: $t = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$

where s is the sample standard deviation



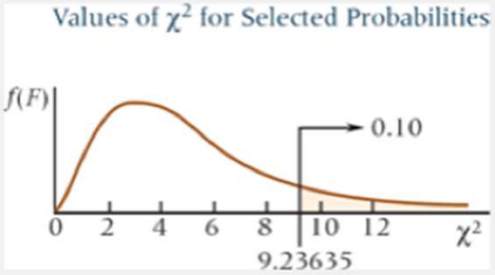
χ^2 Distribution

- A ratio of variances follows a χ^2 distribution if the population from which the values are drawn is normally distributed.
- Unlike the t-distribution, a χ^2 distribution lacks robustness, i.e. it is very sensitive to violation of assumptions.
- χ^2 distribution is non-symmetric and also varies by value of sample size and degrees-of-freedom
- χ^2 statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$
- Where, $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
- Degrees of freedom: *(n-1)*



χ^2 Distribution

The Chi-Square Table



Example: df (number of degrees of freedom) = 5, the tail above χ^2 = 9.23635 represents 0.10 or 10% of area under the curve.

Degrees of Freedom	Area in Upper Tail									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.0000393	0.0001571	0.0009821	0.0039322	0.0157907	2.7055	3.8415	5.0239	6.6349	7.8794
2	0.010025	0.020100	0.050636	0.102586	0.210721	4.6052	5.9915	7.3778	9.2104	10.5965
3	0.07172	0.11483	0.21579	0.35185	0.58438	6.2514	7.8147	9.3484	11.3449	12.8381
4	0.20698	0.29711	0.48442	0.71072	1.06362	7.7794	9.4877	11.1433	13.2767	14.8602
5	0.41175	0.55430	0.83121	1.14548	1.61031	9.2363	11.0705	12.8325	15.0863	16.7496
6	0.67573	0.87208	1.23734	1.63538	2.20413	10.6446	12.5916	14.4494	16.8119	18.5475
7	0.98925	1.23903	1.68986	2.16735	2.83311	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.34440	1.64651	2.17972	2.73263	3.48954	13.3616	15.5073	17.5345	20.0902	21.9549
9	1.73491	2.08789	2.70039	3.32512	4.16816	14.6837	16.9190	19.0228	21.6660	23.5893
10	2.15585	2.55820	3.24696	3.94030	4.86518	15.9872	18.3070	20.4832	23.2093	25.1881
11	2.60320	3.05350	3.81574	4.57481	5.57779	17.2750	19.6752	21.9200	24.7250	26.7569

Population Variance

The Labour Statistics data on the hourly compensation costs for production workers in manufacturing for various countries show that the average hourly wage for a production worker in manufacturing is Rs. 217.8. Suppose the business council wants to know how consistent this figure is. It randomly selects 25 production workers in manufacturing from across the country and determines that the sample standard deviation of hourly wages for such workers is Rs.11.2. Use this information to develop a 95% confidence interval to estimate the population variance for the hourly wages of production workers in manufacturing. Assume that the hourly wages for production workers across the country in manufacturing are normally distributed.

Estimate population variance using variance of sample data.

Population Variance

The Labour Statistics data on the hourly compensation costs …

- $n = 25, s = 11.2, \alpha = 0.05, \chi^2_{0.025,24} = 39.3641, \chi^2_{\{0.975,24)} = 12.40115$

Confidence Interval of estimated variance:

=CHISQ.INV($\alpha/2$, df)

=CHISQ.INV(1- $\alpha/2$, df)

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2),n-1}}$$

$$\frac{(25-1)11.2^2}{39.3641} \leq \sigma^2 \leq \frac{(25-1)11.2^2}{12.40115}$$

$$76.48 \leq \sigma^2 \leq 242.76$$

With 95% confidence the Business Council can conclude that the variance in the hourly wage rate is between 76.48 and 242.76

Application

A manufacturing plant produces steel rods. During one production run of 20,000 such rods, the specifications called for rods that were 46 centimeters in length and 3.8 centimeters in width. Fifteen of these rods comprising a random sample were measured for length; the resulting measurements are shown here. Use these data to estimate the population variance of length for the rods. Assume rod length is normally distributed in the population. Construct a 99% confidence interval. Discuss the ramifications of the results.

44 cm	47 cm	43 cm	46 cm	46 cm
45 cm	43 cm	44 cm	47 cm	46 cm
48 cm	48 cm	43 cm	44 cm	45 cm

Sample Size

Suppose an analyst wants to estimate the average monthly expenditure on bread by a family in Amritsar. She wants to be 90% confident of her results. How much error is she willing to tolerate in the results? Suppose she wants the estimate to be within Rs. 10 of the actual figure (error) and the standard deviation of average monthly bread purchases is Rs. 40. What is the sample size estimation for this problem? The value of *z* for a 90% level of confidence is 1.645.

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

This can be simplified as:

$$n = \left(\frac{z_{\alpha/2}\sigma}{x - \mu}\right)^2$$

The denominator $(x - \mu)$ is the extent of error that the analyst is willing to tolerate

Thus, $n = \frac{1.645^2 * 40^2}{10^2} = 43.29$ or a sample size of 44 will be required.

How would you reduce the sample size?

Application

Suppose you want to estimate the average age of all Boeing 737-600 airplanes now in active domestic circuit in India. You want to be 95% confident, and you want your estimate to be within one year of the actual figure. The 737-600 was first placed in service about 23 years ago, but you believe that no active 737-600s in the domestic fleet are more than 20 years old. How large of a sample should you take?

Note: σ is not known. In such cases it can be approximated as σ = (1/4)Range = 5