Inferential Statistics

The story so far..

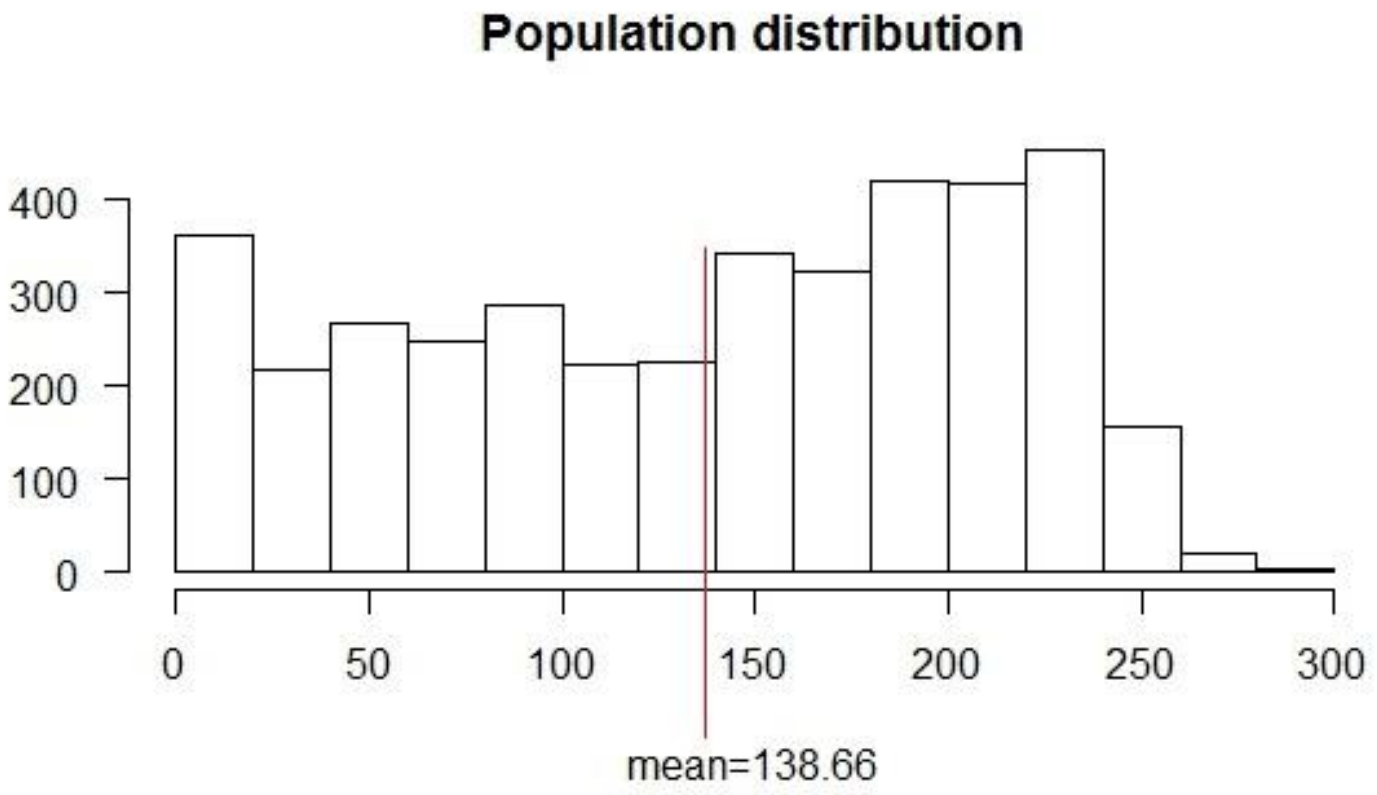
- Types of variables
- Scales of Measurement
- Central Tendency
- Spread of Data
- Basic Graphs
- Probability
- Probability Distributions

Central Limit Theorem

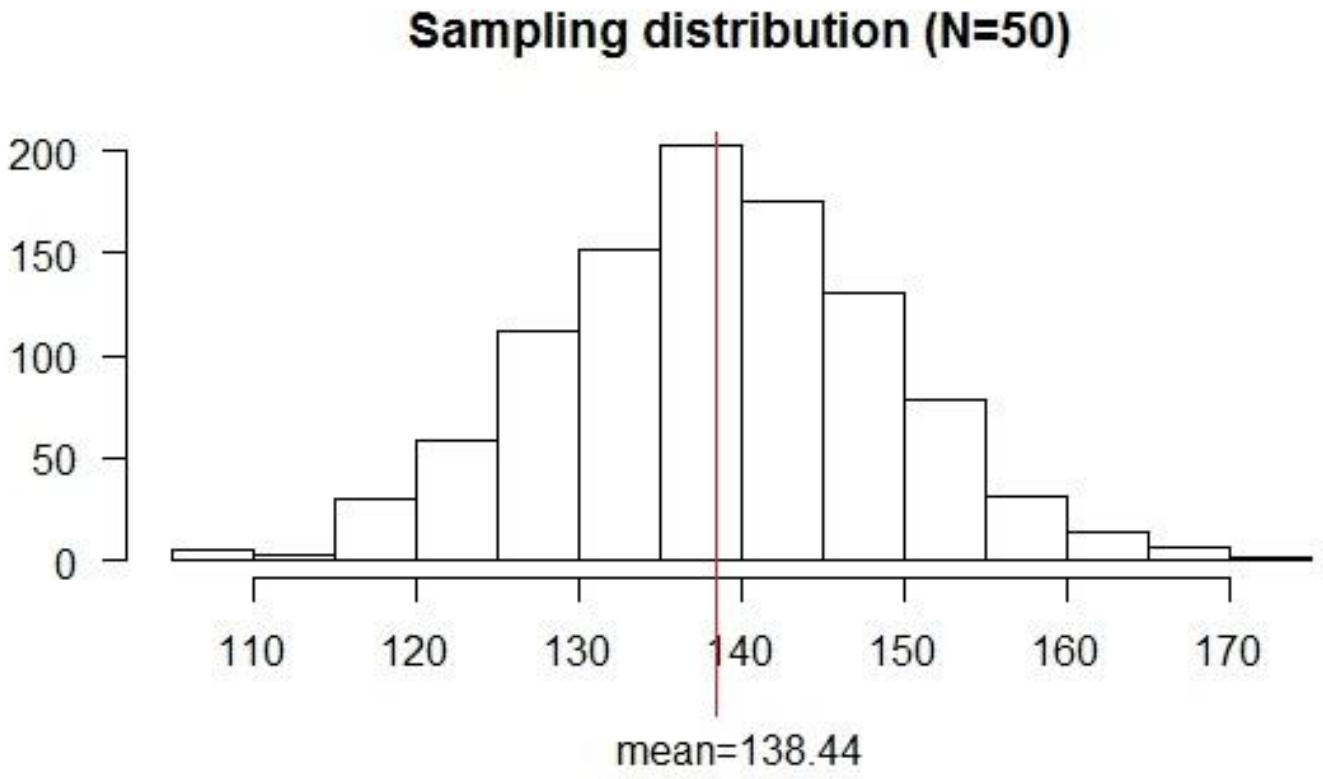
Central Limit Theorem

- Consider a distribution with mean μ and variance σ^2
- Let us take a sample of size N from the distribution
- Let us take many such samples of size N and make a sampling distribution
- The sampling distribution approaches normal distribution with mean μ and variance σ^2/N and N increases

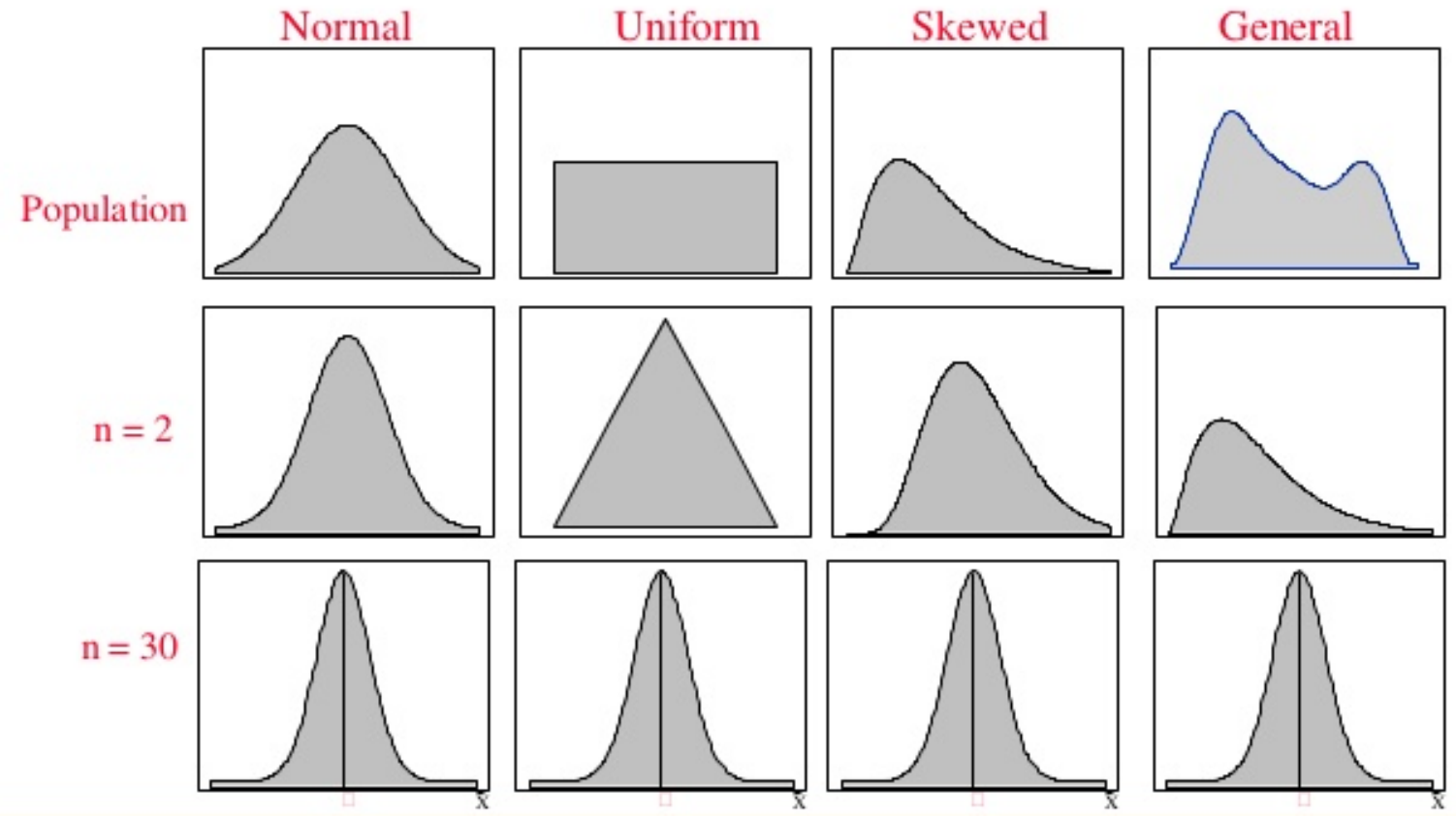
CLT - Illustration



CLT – Illustration (contd.)



CLT – Illustration (contd.)



http://www.slideshare.net/ShakeelNouman1/sampling-and-sampling-distributions

Implications of CLT

- Whenever we take a sample statistic from a sample of sample size N from a population with mean μ and variance σ^2, we can assume that:
  - The sample statistic is from the sampling distribution which is normally distributed
  - The mean of the sampling distribution is μ
  - The variance of the sampling distribution is $\sigma_m^2 = \sigma^2/N$
- This means that the mean of a sample statistic of sample size of N will be within 2 standard deviations ($2\sigma_m$) of original population mean μ

Statistical Hypothesis Testing

Statistical Hypothesis

- Statistical Hypothesis is an assumption about a population parameter
- The assumption may or may not be true
- Examples:
  - The mean length of the chocolate bars produced in the factory is 5 cm
  - The mean age of all learning data analytics course is 24.3 years
  - The proportion of women among learners of data analytics is 76 percent
  - The mean weight of cricket balls produced by two machines are equal

Null and Alternate Hypothesis

- Null hypothesis is the initial claim or the default position
  - It is denoted by H0
- Alternate hypothesis is the rival of the null hypothesis
  - It is denoted by H_1 or H_A
- For example, in the chocolate bars manufacturing case, the null hypothesis is that the mean length of the chocolate bars produced is 5cm. Alternate hypothesis is that the mean length is not 5 cm
  - H_0: $\mu = 5$
  - H_1: $\mu \neq 5$

Null and Alternate Hypothesis (contd.)

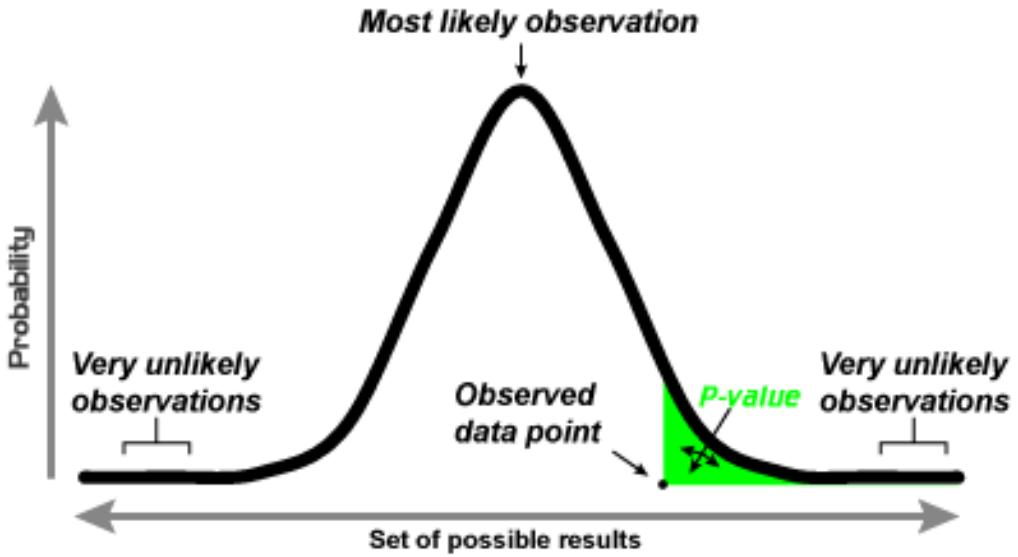
- In the example about the proportion of data analytics learners, the null hypothesis is that the proportion of women among data analytics learners is 0.76
- The alternate hypothesis is that the proportion of women is not 0.76
  - H_0: P = 0.76
  - H_1: P ≠ 0.76

Null and Alternate Hypothesis (contd.)

- We try to prove the alternate hypothesis
- If the alternate hypothesis is proved, we reject the null hypothesis
- If the alternate hypothesis is disproved, we fail to reject the null hypothesis
- If the alternate hypothesis is disproved, we don't say we accept the null hypothesis instead we say there is not enough evidence to reject the null hypothesis

p-value

- p-value is the probability of obtaining a sample statistic equal to or more extreme than the observed value given that the null hypothesis is true



https://en.wikipedia.org/wiki/File:P-value_Graph.png

p-value (contd.)

- Let us design an experiment to decide whether a coin is fair or biased
  - H0: Coin is fair ($p(H) = 0.5$)
  - H1: Coin is biased ($p(H) \neq 0.5$)
- Let us start the coin and find the p-value i.e. the probability of getting the result in the observation

Sample Value	p-value (Probability)
1 Head	0.5
2 Heads	0.25
3 Heads	0.125
4 Heads	0.0625
5 Heads	0.03125
6 Heads	0.015625

Significance Level

- Significance Levels (α) refers to the predefined probability to compare the p-value
- If the p-value is less than α, we reject the null hypothesis
- If the p-value is not less than α, we fail to reject the null hypothesis
- We say that we have achieved a statistically significant result if the p-value is less than α
- Typical significance level is 0.05 (5%)
- Sometimes .01 (1%) is used as a significance level

Type I and Type II Errors

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

http://whatilearned.wikia.com/wiki/File:Type_I_and_Type_II_Error_Table.jpg

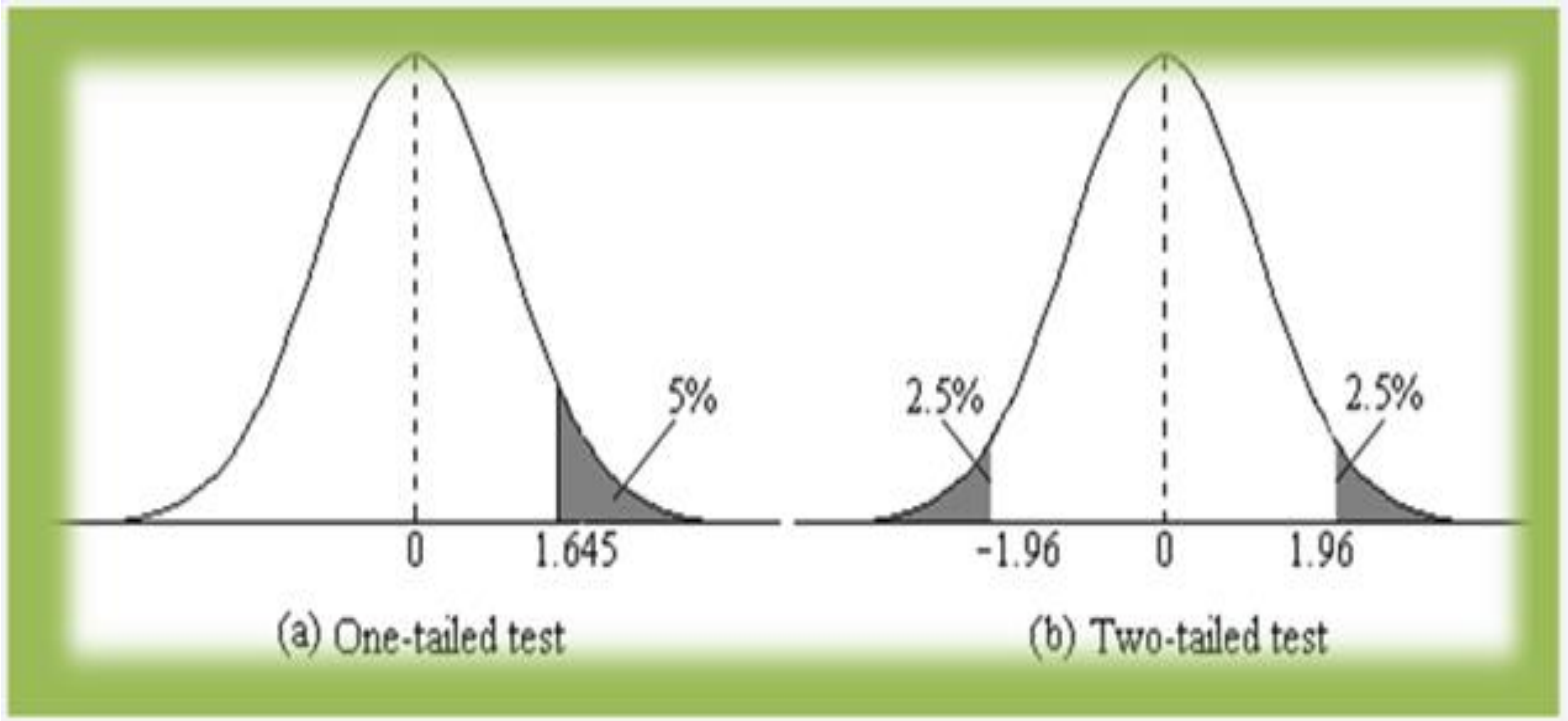
One-tailed and two-tailed tests

- One-tailed test allows you to determine whether a sample statistic is greater than or less than a certain value
- Here we are interested in one direction
- The direction has to be chosen before the test
- Example:
  - A chips manufacturer claims that the average amount of saturated fat in a packet of chips is no more than 10g
  - H_0: $\mu = 10$, H_1: $\mu > 10$
  - We are interested only in one direction i.e. whether the fat content is more than 10 g

One-tailed and two-tailed tests (contd.)

- Two-tailed test allows you to determine whether a sample statistic is not equal to a certain value
- Here we are interested in both the directions i.e. the sample statistic is greater or less than a certain value
- Example:
  - A cricket ball manufacturer says the average weight of the ball produced by them is 161 gm
  - H_0: $\mu = 161$, H_1: $\mu \neq 161$
  - We are interested in both directions i.e. whether the ball is heavier or lighter than 161

One-tailed and two-tailed tests (contd.)



One-sample, two-sample and paired tests

- When a sample statistic is compared with a single value, one-sample test is used
  - e.g. Average weight of cricket ball produced is 161 g
  - H_0: $\mu = 161$, H_1: $\mu \neq 161$
- When two sample values are compared and the two samples are independent of each other, two-sample test is used
  - e.g. Lets say there are two machines in the company which produces cricket balls and we need to find whether the weight of the cricket balls produced by these two machines differ
  - H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$

One-sample, two-sample and paired tests (contd.)

- When two sample values being compared are dependent on each other, paired test should be used
  - e.g. Let us say we want to study whether a particular drug reduces blood glucose level.
  - We take 100 persons and record their blood glucose levels before and after taking the drug.
  - Now we have to compare blood glucose levels before taking the drug and after taking the drug

One-sample, two-sample and paired tests (contd.)

	Blood Glucose (mg/dL)	
	Before Drug	After Drug
Person 1	102	92
Person 2	127	108
Person 3	113	110
Person 4	106	117
Person 5	133	108
Person 6	103	114
Person 7	132	82
Person 8	137	106
Person 9	95	117
Person 10	130	88
Person 11	136	111
Person 12	130	119
Person 13	102	120
Person 14	122	83
Person 15	123	91
Person 16	140	84
Person 17	98	87

One-sample, two-sample and paired tests (contd.)

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$
- Here in each observation, the blood glucose levels before and after are related to each other
- If the blood glucose level is relatively higher before drug, the blood glucose level after drug will also be relatively higher
- Since each pair of observation is dependent on the other, we use paired test here

Steps in hypothesis testing

- State the null hypothesis H_0 and the alternate hypothesis H_1
- Select the appropriate level of significance α
- Calculate the test statistic, p-value
- Compare test statistic with alpha
- Make a decision about null hypothesis H_0

One Sample Tests

One sample tests

- One sample t-test
- One sample z-test
- Chi-square test for variance
- One sample proportion test

One sample t-test

- Used to test population mean μ
- Population standard deviation is unknown
- Sample size is small (n < 30)
- Example:
  - Given below is the weight of 8 sample cricket balls produced by a factory
  - 160.39, 160.64, 160.13, 160.53, 160.86, 161.89, 160.51, 162.43
  - With this sample data, we want to check whether the average weight of the cricket ball produced by the factory is equal to 161 g

One sample z-test

- Used to test population mean μ
- Population standard deviation is known
- If the sample size is large ($n \geq 30$), population standard deviation is approximately equal to the sample standard deviation ($\sigma \approx s$)
- Example:
  - Suppose we need to find out whether the mean height of 12-year old boys in India is 150cm
  - We take 1000 random samples from all over India

Chi-square test for variance

- Used to test the population variance (or standard deviation)
- Sample variance is calculated and compared with the hypothetical value of population variance
- Example:
  - Let us assume that, in the previous example, the population standard deviation of height of boys aged 12 is 2.6 cm
  - Sample variance of 1000 boys aged 12 years can be calculated from the data
  - From this, hypothesis on population variance can be tested

One sample proportion test

- Used to test population proportion or probability
- Example:
  - We are interested in finding out whether a candidate in an election will get more than 55% of votes
  - Null hypothesis is that the population proportion who would vote for our candidate is 0.55
  - H_0: $p > 0.55$, H_1: $p \leq 0.55$
  - We take a random sampling of 1000 voters and 513 of them said they will vote for the candidate

Two Sample Tests

Two sample tests

- Two sample t-test
- Two sample z-test
- Paired t-test
- Two sample proportion test
- F-test

Two sample t-test

- Used to test whether there is any significant difference between two population means
- Example:
  - We want to find whether there is any significant difference between the mean weight of cricket balls produced by two machines in a factory
  - Samples from each machine is taken
  - Sample 2 from Machine 1: 159.17, 159.86, 159.77, 161.19, 159.25, 159.31, 161.75, 160.45, 162.80, 159.54
  - Sample 2 from Machine 2: 154.47, 159.66, 161.11, 160.82, 161.49, 159.16, 158.89, 163.73, 160.77, 158.20

Two sample z-test

- Used to test whether there is any significant difference between two population means
- Example:
  - Suppose we want to find out whether there is any significant difference between the mean heights of 12 year old boys and 12 year old girls in the country
  - 1000 random sample heights of both 12 year old boys and girls are taken across the country

Paired t-test

- Used to test whether there is any significant difference between two population means where the samples are dependent on each other
- Example:
  - We are interested in finding whether a particular drug changes the blood glucose level
  - We take blood glucose level of 25 persons and administer the drug to them
  - Blood glucose level is taken after the drug is administered

Two sample proportion test

- Used to test whether the difference in two proportions is significant
- Example:
  - Two machines produce cricket balls in a factory, some of which are defective
  - We are interested in testing whether there is a significant difference between the proportion of defective balls produced by the 2 machines
  - Null hypothesis is that the proportion of defective cricket balls is the same
  - H_0: $p_1 = p_2$, H_1: $p_1 \neq p_2$
  - Machine 1 gave 31 defective balls out of 250 and machine 2, 42 out of 300

F-test

- Used to test the variance of two samples
- Samples should be normally distributed
- Example:
  - Let us consider the mean height of 12 year old boys and girls
  - We know that the height of a population is normally distributed
  - We are interested in the variance in height among the boys and girls

Multi Sample Tests

Multi sample tests

- ANOVA (Analysis of Variance)
- Chi-square test for independence

ANOVA

- Also known as Analysis of Variance
- Used to compare three or more means
- Example:
  - For example, in the height of 12 year old boys case let us say we need to find whether there is any difference between the mean of boys from different states
- For single step multiple comparison, Tukey HSD test is used

Chi-square test of independence

- Used to evaluate the relationship between variables i.e. whether variables are independent or there is a dependency between them
- Example:
  - Consider the following table which gives the colour preference for boys and girls

	Pink	Blue	White	
Boys	20	120	100	240
Girls	140	40	80	260
	160	160	180	500

  - We are interested in knowing whether there is a relationship between colour preference and gender

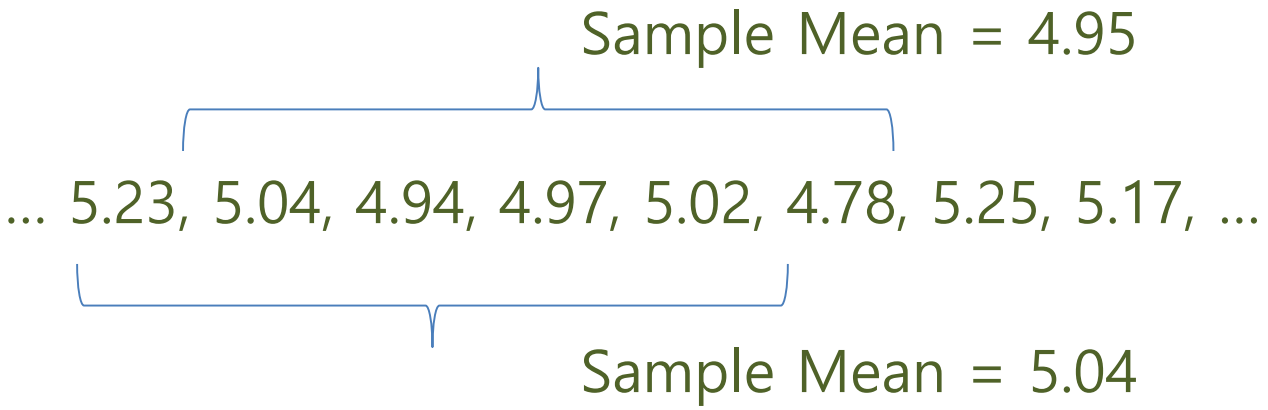
Chi-square test of independence (contd.)

- Here we have two variables gender and colour preference
- The null hypothesis:
- H_0: The variables gender and colour are independent of each other
- The alternate hypothesis:
- H_1: The variables gender and colour are dependent on each other

Confidence Intervals

Confidence Interval for Means

- Consider a machine in a chocolate factory which produces bars of chocolates
- We are interested in the average length of each bar of chocolate



- How confident or sure are we about sample average being equal to the true average of the chocolate bars?

Confidence Interval for Means (contd.)

- Confidence Intervals helps us describe the amount of uncertainty associated with a sample estimate
- It is an interval estimate combined with a probability statement

$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

- $s/\sqrt{n}$ is called the standard error

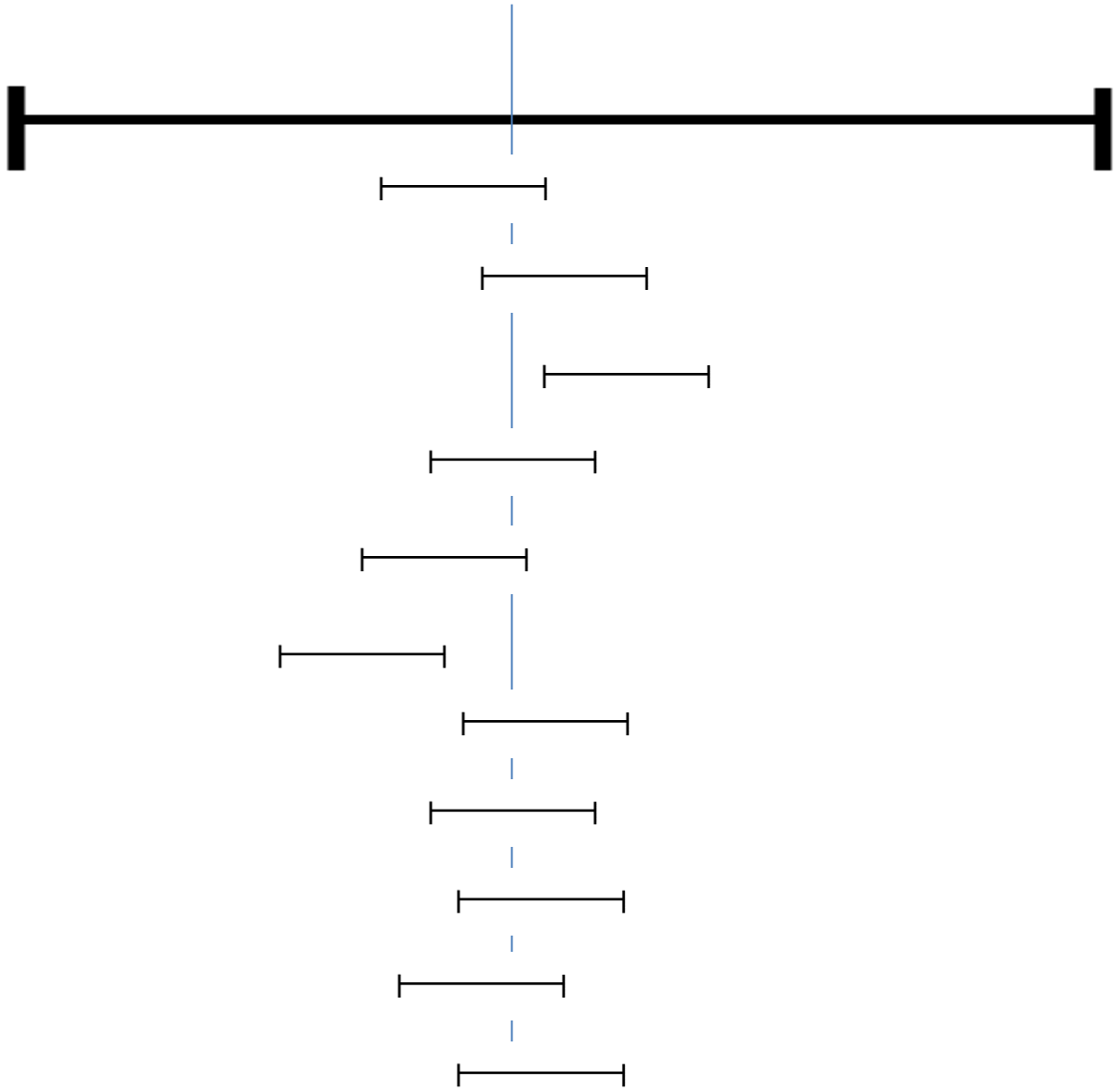
T-Distribution Table

	Probability (P)												
n	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Interpreting Confidence Intervals

- Confidence Intervals communicate how accurate our estimate is likely to be
- A 95% confidence interval means that if we use the same sampling method to select different samples and compute the interval estimate for each sample, we would expect the true population estimate to fall within the interval estimates 95% of the time

Interpreting Confidence Intervals (contd.)



Confidence Interval for Proportion

- So far we have seen confidence intervals for mean
- Consider the following case:
  - A pre-poll survey was conducted to find out whether people would vote for a party or not
  - Response was coded as 1 if the respondent say 'Yes' and 0 if he says 'No'
  - 1000 responses were collected
  - The responses were: …1011100110…
  - 582 responses were 1
  - The probability of 1 from the sample, $\hat{p}(1) = 0.582$

Confidence Interval for Proportion (contd.)

- The aim is to find the probability of people voting for a party
- The population is all the people in the state who are eligible to vote
- The sample size is 1000
- The sample probability we have got is 0.582
- If we take another sample, we will get a different value
- So there is an uncertainty in the sampling
- Can we express the probability as a confidence interval like we did for mean?