DATA SCIENCE WITH R

DATA SCIENCE WITH R

HYPOTHESIS TESTING

Introduction to Hypothesis Testing

Basic Framework of a Hypothesis Test

Distance Measures

Central Limit Theorem

Types of Hypothesis Tests

Single Sample

Single Sample

The ***Central Limit Theorem*** *examples we looked at previously when we introduced hypothesis testing were not done quite correctly*

Single Sample

The ***Central Limit Theorem*** *examples we looked at previously when we introduced hypothesis testing were not done quite correctly*

Recalling the Quality Control example:

At a manufacturing unit, the mean weight of part A produced is expected to be 2.5 lbs, with a std dev of 0.12 lbs, and is normally distributed. A random sample of 45 units results in a mean weight of 2.68 lbs. Is there an issue?

Single Sample

The ***Central Limit Theorem*** *examples we looked at previously when we introduced hypothesis testing were not done quite correctly*

Recalling the Quality Control example:

At a manufacturing unit, the mean weight of part A produced is expected to be 2.5 lbs, with a std dev of 0.12 lbs, and is normally distributed. A random sample of 45 units results in a mean weight of 2.68 lbs. Is there an issue?

We had used –

Ho: No issue with process, unit weight is still 2.5 lbs

H1: Process is failing, weight per unit has increased

Single Sample

The ***Central Limit Theorem*** *examples we looked at previously when we introduced hypothesis testing were not done quite correctly*

<u>Recalling the Quality Control example</u>:

At a manufacturing unit, the mean weight of part A produced is expected to be 2.5 lbs, with a std dev of 0.12 lbs, and is normally distributed. A random sample of 45 units results in a mean weight of 2.68 lbs. Is there an issue?

We had used –

Ho: No issue with process, unit weight is still 2.5 lbs

H1: Process is failing, weight per unit has increased

Sig Level: 5%P- value: 1- norm.dist(2.68,2.5,0.12,true)

Single Sample

Instead use:

P = 1 – norm.dist(2.68, 2.5, (0.12/(45)^0.5), true)

Remember, the Central Limit theorem states that the distribution of sample means will follow a normal distribution with:

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Single Sample

Instead use:

P = 1 – norm.dist(2.68, 2.5, (0.12/(45)^0.5), true)

Remember, the Central Limit theorem states that the distribution of sample means will follow a normal distribution with:

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

So, the p value for this problem will be: a number very close to zero

Single Sample

Instead use:

P = 1 – norm.dist(2.68, 2.5, (0.12/(45)^0.5), true)

Remember, the Central Limit theorem states that the distribution of sample means will follow a normal distribution with:

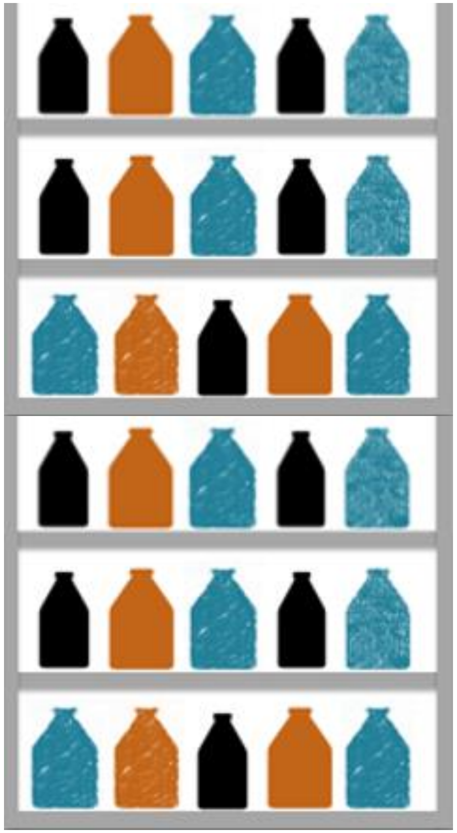
$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

So, the p value for this problem will be: a number very close to zero

So we REJECT the null hypothesis, since p value < Sig level

Single Sample

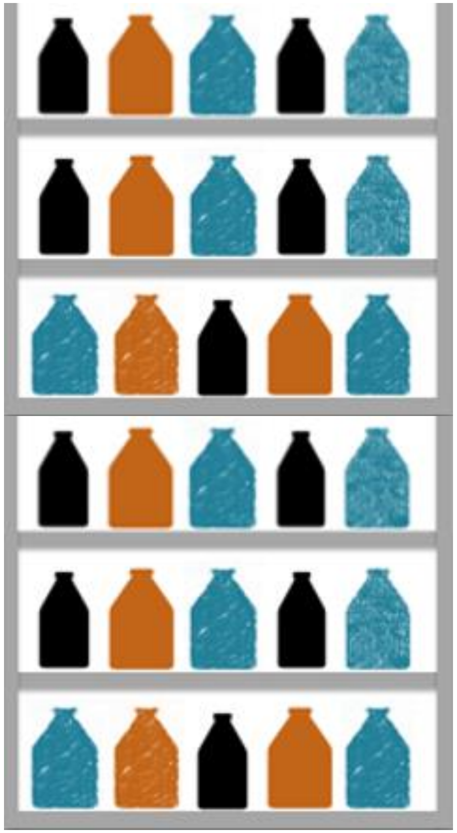
Inventory Optimization



Single Sample

Inventory Optimization

To optimize inventory costs for your retail chain, you look at shelf stable beverages:

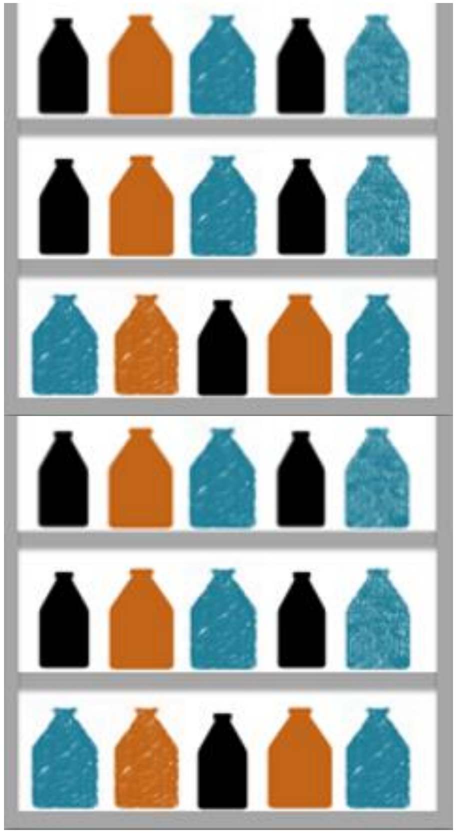


Single Sample

Inventory Optimization

To optimize inventory costs for your retail chain, you look at shelf stable beverages:

- Historical data shows average daily sales for this category is 310, with a std deviation of 85.

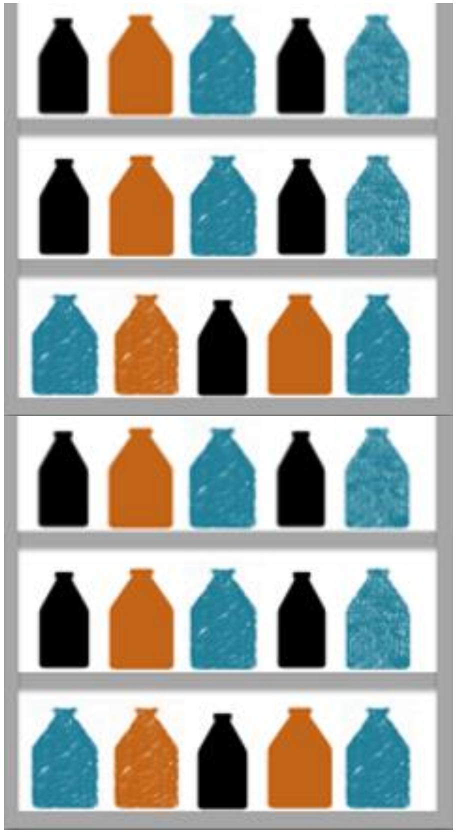


Single Sample

Inventory Optimization

To optimize inventory costs for your retail chain, you look at shelf stable beverages:

- Historical data shows average daily sales for this category is 310, with a std deviation of 85.
- Taking a current sample of the last 45 days to validate, you find average daily sales are 338.

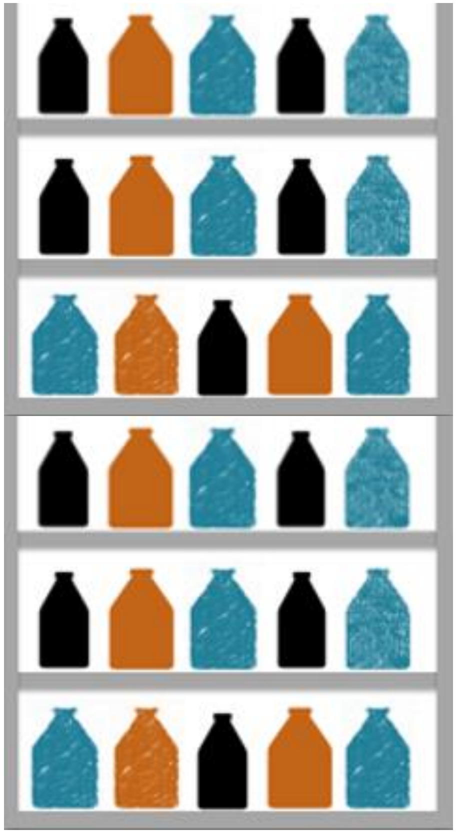


Single Sample

Inventory Optimization

To optimize inventory costs for your retail chain, you look at shelf stable beverages:

- Historical data shows average daily sales for this category is 310, with a std deviation of 85.
- Taking a current sample of the last 45 days to validate, you find average daily sales are 338.
- Should you increase inventory levels?

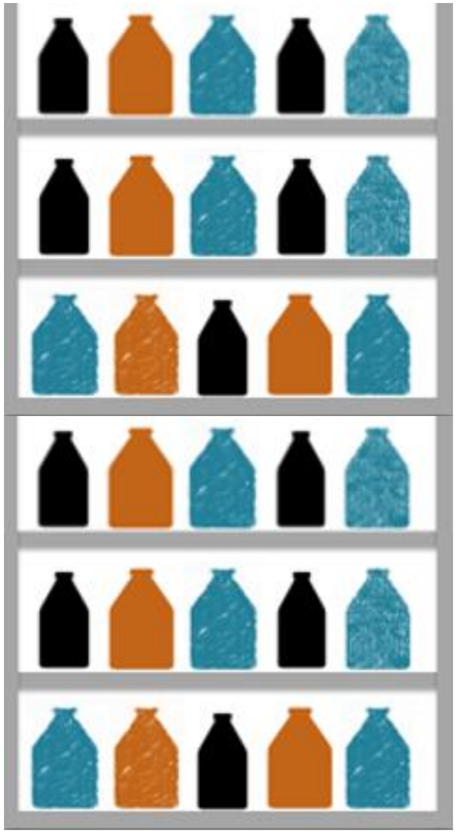


Single Sample

Inventory Optimization

To optimize inventory costs for your retail chain, you look at shelf stable beverages:

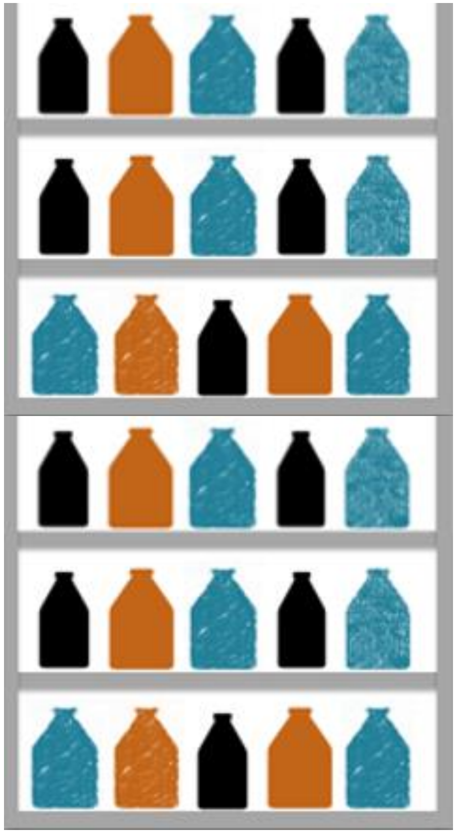
- Historical data shows average daily sales for this category is 310, with a std deviation of 85.
- Taking a current sample of the last 45 days to validate, you find average daily sales are 338.
- Should you increase inventory levels?
- CLT – Test distribution will be?



Single Sample

Therefore:

1. Set up hypothesis

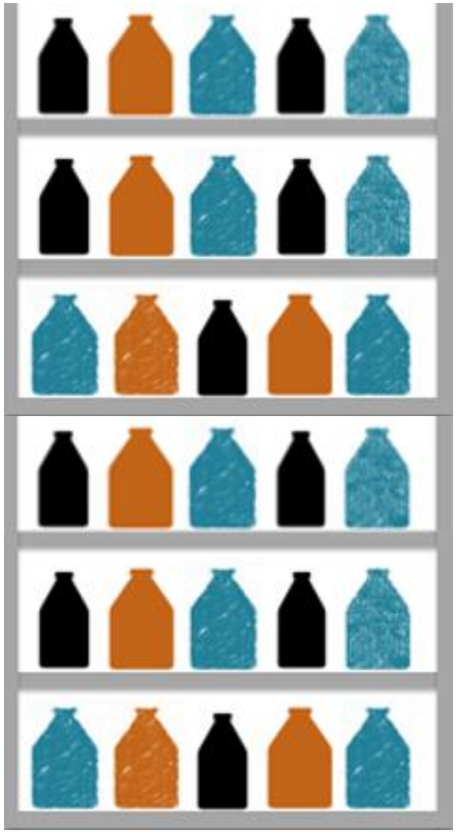


Single Sample

Therefore:

1. Set up hypothesis

2. Decide on a significance level



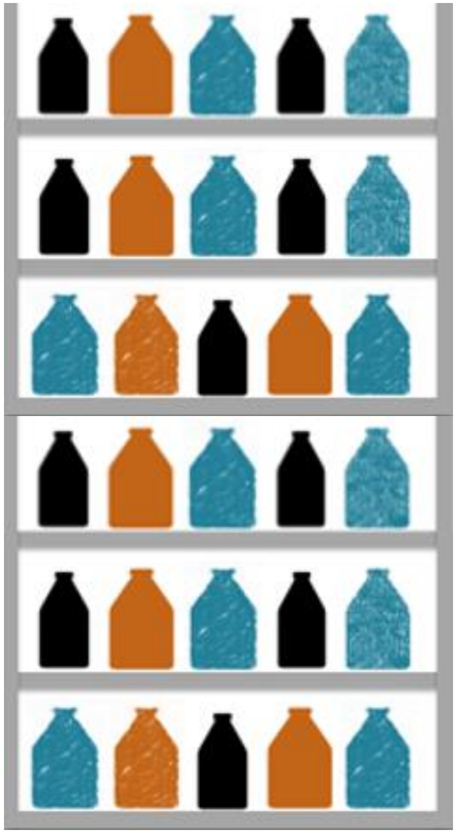
Single Sample

Therefore:

1. Set up hypothesis

2. Decide on a significance level

3. Calculate p-value of outcomes more extreme than observed



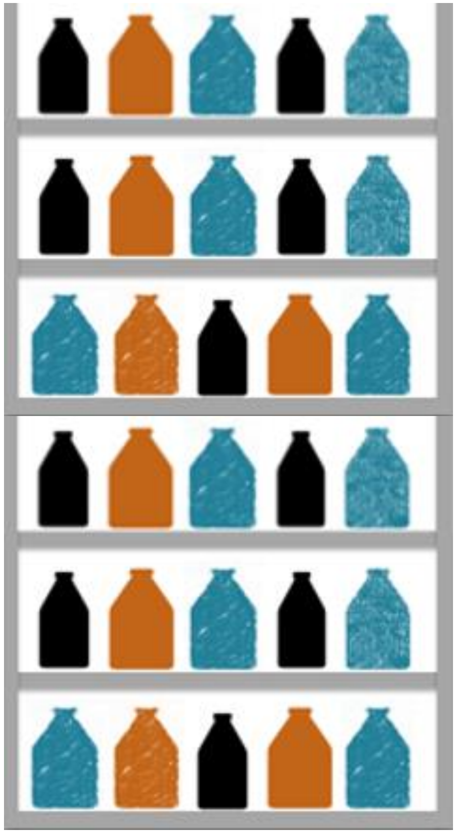
Single Sample

Therefore:

1. Set up hypothesis

2. Decide on a significance level

3. Calculate p-value of outcomes more extreme than observed
    1. Use normal distribution function



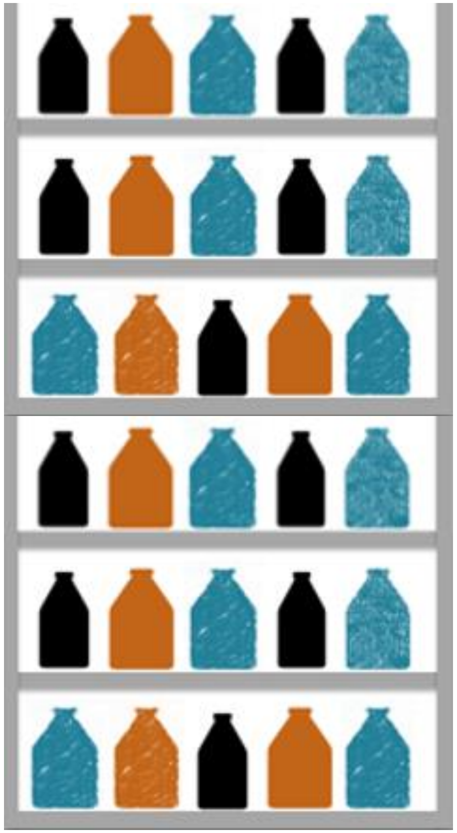
Single Sample

Therefore:

1. Set up hypothesis

2. Decide on a significance level

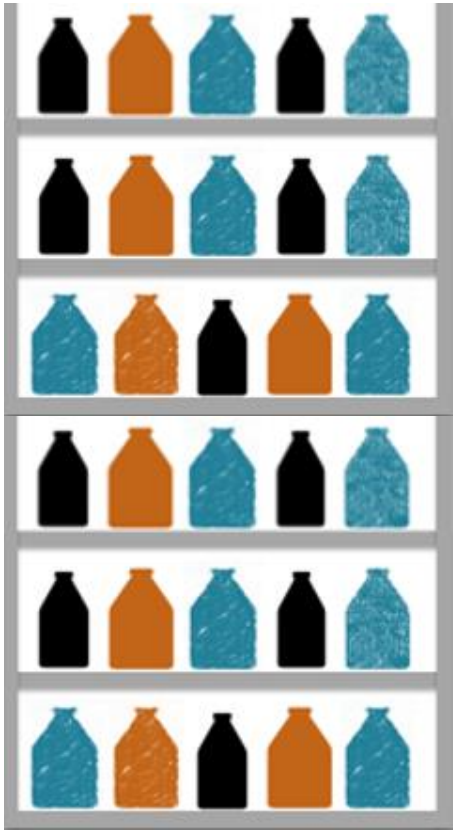
3. Calculate p-value of outcomes more extreme than observed
    1. Use normal distribution function
    2. Use a table



Single Sample

<u>Therefore:</u>

1. Set up hypothesis
2. Decide on a significance level
3. Calculate p-value of outcomes more extreme than observed
    1. Use normal distribution function
    2. Use a table
4. Compare to significance level / critical value and come to a conclusion



Single Sample

1. **Set up hypothesis**

   Null Hypothesis?

   HO: No difference in average daily units sold

Single Sample

1. **<u>Set up hypothesis</u>**

   Null Hypothesis?

   HO: No difference in average daily units sold

   Alternate Hypothesis:

   H1: There is a change in average daily units sold

Single Sample

1. **Set up hypothesis**

   Null Hypothesis?

   HO: No difference in average daily units sold

   Alternate Hypothesis:

   H1: There is a change in average daily units sold

2. **Decide on a significance level**

Single Sample

1. **Set up hypothesis**

   Null Hypothesis?

   HO: No difference in average daily units sold

   Alternate Hypothesis:

   H1: There is a change in average daily units sold

2. **Decide on a significance level**

   5%

Single Sample

1. **<u>Set up hypothesis</u>**

   Null Hypothesis?

   HO: No difference in average daily units sold

   Alternate Hypothesis:

   H1: There is a change in average daily units sold

2. **<u>Decide on a significance level</u>**

   5%

3. **<u>Calculate p-value of outcomes more extreme than observed</u>**

Single Sample

1. **Set up hypothesis**

   Null Hypothesis?

   HO: No difference in average daily units sold

   Alternate Hypothesis:

   H1: There is a change in average daily units sold

2. **Decide on a significance level**

   5%

3. **Calculate p-value of outcomes more extreme than observed**

   a) Use normal distribution function: norm.dist(338,310,(85/(45)^0.5, true) =0.136

Single Sample

1. **<u>Set up hypothesis</u>**

   Null Hypothesis?

   HO: No difference in average daily units sold

   Alternate Hypothesis:

   H1: There is a change in average daily units sold

2. **<u>Decide on a significance level</u>**

   5%

3. **<u>Calculate p-value of outcomes more extreme than observed</u>**

   a) Use normal distribution function: norm.dist(338,310,(85/(45)^0.5, true) =0.136

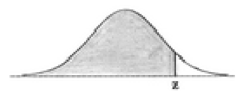
   b) Use a table

Single Sample

To use a table:

- Calculate std distance
  Z = (338-310)/(85/(45)^0.5)
  = 2.209

Tables of the Normal Distribution



Probability Content
from -oo to Z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986

Single Sample

To use a table:

- Calculate std distance

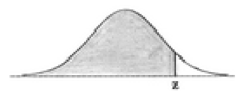
  Z = (338-310)/(85/(45)^0.5)

  = 2.209

- Look up the cumulative prob of < = Z in the table

  = 0.9864

Tables of the Normal Distribution



Probability Content
from -oo to Z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986

Single Sample

To use a table:

- Calculate std distance

  Z = (338-310)/(85/(45)^0.5)

  = 2.209

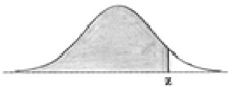
- Look up the cumulative prob of < = Z in the table

  = 0.9864

- We need prob of > Z (outcomes more extreme than observed)

  = 1- 0.9864 = 0.136

Tables of the Normal Distribution



Probability Content
from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986

Single Sample

<u>To use a table:</u>

- Calculate std distance

  Z = (338-310)/(85/(45)^0.5)

  = 2.209

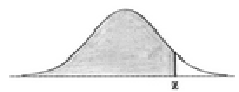
- Look up the cumulative prob of < = Z in the table

  = 0.9864

- We need prob of > Z (outcomes more extreme than observed)

  = 1- 0.9864 = 0.136

Tables of the Normal Distribution



Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986

Single Sample

P – value:

- Using excel: 1- 0.9864 = 0.013
- Using the table: 1- 0.9864 = 0.013

Single Sample

<u>P – value:</u>

- Using excel: 1- 0.9864 = 0.013
- Using the table: 1- 0.9864 = 0.013

Conclusion?

P value < significance level : reject null hypothesis

Single Sample

P – value:

- Using excel: 1- 0.9864 = 0.013
- Using the table: 1- 0.9864 = 0.013

Conclusion?

P value < significance level : reject null hypothesis

i.e. : reject the null that average units have not changed: Accept alternate.

Single Sample

P – value:

- Using excel: 1- 0.9864 = 0.013
- Using the table: 1- 0.9864 = 0.013

Conclusion?

P value < significance level : reject null hypothesis

i.e. : reject the null that average units have not changed: Accept alternate.

Business recommendation? Increase inventory levels

Coming Up

Types of Hypothesis Tests:

Population Distribution Not Normal

THANK YOU