CS 215- Data Interpretation and Analysis
(Post Midsem)

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Lecture-4

Towards Hypothesis Testing

16oct23

Recap

Z-score table

Standard Normal Probabilities

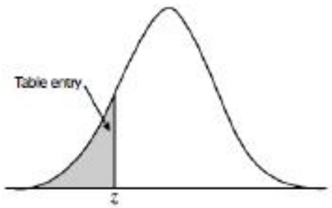


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

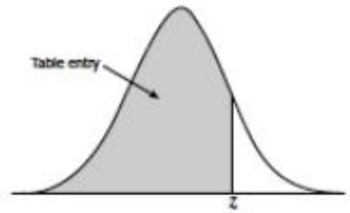
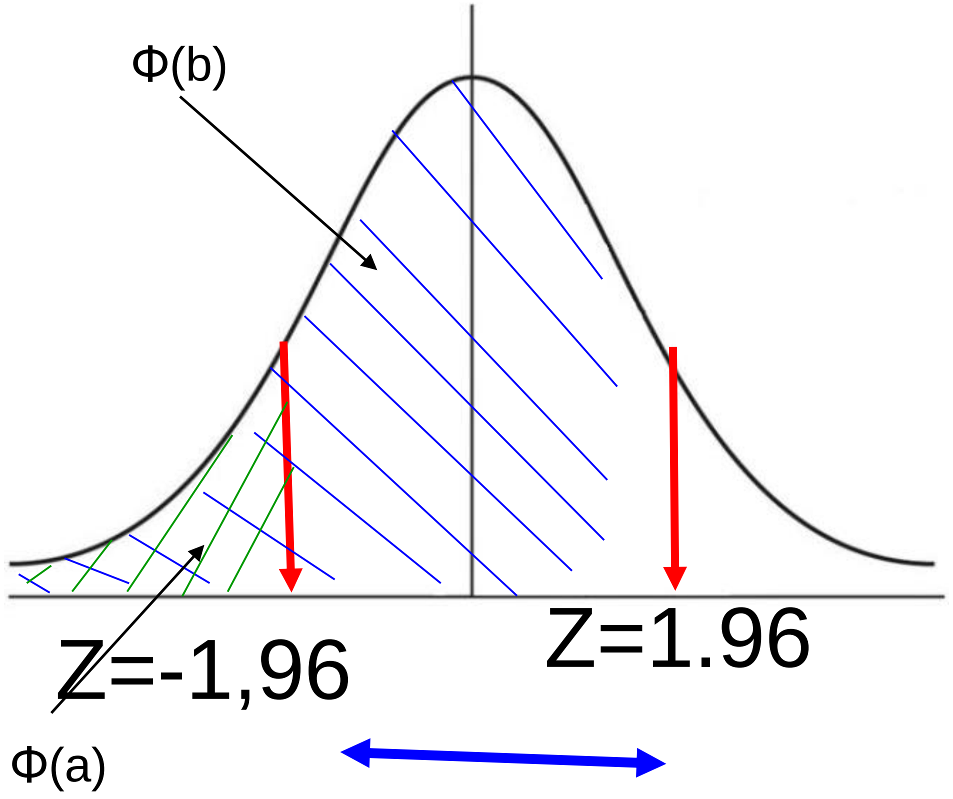


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

The 95% confidence interval



$-1.96 <= Z <= +1.96$

- $\Phi(1.96)=0.9750$

- By symmetry

$\Phi(-1.96)= 1-\Phi(1.96)$

$\rightarrow \Phi(1.96)-\Phi(-1.96)= 2.\Phi(1.96)-1= 2 \text{ X } 0.975-1.0=1.95-1.0=0.95$

Statement of Central Limit Theorem

- Let $X_1, X_2, X_3, \ldots, X_n$ be n independent random variables, each with mean μ and variance σ^2
- Also let

$$S_n = X_1 + X_2 + X_3, \ldots + X_n$$

- Then,

the following is **standard normal**

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Mathematical adjustment

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \; gives$$

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\dfrac{S_n}{n} - \mu}{\dfrac{\sigma\sqrt{n}}{n}} = \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$$

Eqv Statement of CLT

Let $X_1, X_2, X_3, \ldots, X_n$ be n independent random variables forming a sample from a population with mean μ and variance σ^2.

Then the sample mean is normally distributed with mean μ and variance σ^2/n.

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

MGF

Moment Generating Function

$M_X(t)=E(e^{tX})$,

X is a random Variable and

$f(x_j)=P(X=x_j)$

$$M_X(t)=\sum_{j=1}^{n} e^{tx_j} f(x_j)$$

for discrete distribution

$$M_X(t)=\int_{-\infty}^{+\infty} e^{tx} f(x)dx$$

for continuous distribution

Proof regarding n^{th} derivative and n^{th} moment

$$M_X'(t) = \frac{d}{dt}E(e^{tX})$$
$$= E\left[\frac{d}{dt}(e^{tX})\right]$$
$$= E[Xe^{tX}]$$
$$= E(X)$$
$$= M_X'(0)$$

$$M_X''(t) = \frac{d}{dt}M_X'(t)$$
$$= \frac{d}{dt}E(Xe^{tX})$$
$$= E\left[\frac{d}{dt}(Xe^{tX})\right]$$
$$= E[X^2e^{tX}]$$
$$= E(X^2); \quad at \quad t = 0$$
$$\therefore \mathrm{var}(X) = M_X''(0) - [M_X'(0)]^2$$

Uniqueness Theorem

- Suppose X and Y are random variables having moment generating functions $M_X(t)$ and $M_Y(t)$ respectively.

- Then X and Y have the same probability distribution if and only if $M_X(t)=M_Y(t)$ identically.

Standard Normal Distribution, *N(0,1)* and its PDF

Normal:

$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma} \right)$$

Standard normal:

$$P(Z = y) = \frac{1}{\sqrt{2\pi}} \exp\left(- y^2 / 2 \right)$$

MGF of *N(0,1)*

$$MGF = \int_{-\infty}^{+\infty} e^{ty} \frac{1}{\sqrt{2\pi}} e^{\left(-y^2/2\right)} dy$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\left(-y^2/2+ty\right)} dy$$

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y^2-2yt+t^2)} . e^{\frac{t^2}{2}} dy$$

$$= e^{\frac{t^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-t)^2} dy$$

$$= e^{\frac{t^2}{2}}$$

Proof of CLT

- To prove that

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- Is standard normal, we will show that

$$M_{S_n^*}(t) = M_Z(t)$$

- i.e., the moment generating function of S_n* is equal to the moment generating function of standard normal r.v.

Proof: MGF

$$E(e^{tS_n^*}) = E[e^{t(S_n - n\mu)/\sigma\sqrt{n})}]$$

$$= E[e^{t(\sum_{i=1}^{n} X_i - n\mu)/\sigma\sqrt{n})}]$$

$$= E[e^{\sum_{i=1}^{n} t(X_i - \mu)/\sigma\sqrt{n})}]$$

$$= E[\prod_{i=1}^{n} (e^{t(X_i - \mu)/\sigma\sqrt{n})})]$$

$$= \prod_{i=1}^{n} E[e^{t(X_i - \mu)/\sigma\sqrt{n})}]$$

$$= \{E[e^{t(X_i - \mu)/\sigma\sqrt{n})}]\}^n$$

Proof: cntd.

$$\{E[e^{t(X_i-\mu)/\sigma\sqrt{n})}]\}^n$$

$$now,$$

$$e^{t(X_i-\mu)/\sigma\sqrt{n})}=[1+\frac{t(X_i-\mu)}{\sigma\sqrt{n}}+\frac{t^2(X_i-\mu)^2}{2\sigma^2 n}+\ldots],$$

$$by\ Taylor\ series\ \exp ansion$$

Proof: working with *E*

$$E[1+\frac{t(X_i-\mu)}{\sigma\sqrt{n}}+\frac{t^2(X_i-\mu)^2}{\sigma^2 n}+...],$$

$$=E(1)+\frac{tE(X_i-\mu)}{\sigma\sqrt{n}}+\frac{t^2E(X_i-\mu)^2}{2\sigma^2 n}+...]$$

$$=1+0+\frac{t^2}{2n}+...$$

As n tends to infinity…

$$E(e^{tS_n^*}) = (1+\frac{t^2}{2n}+\ldots)^n$$

$$Study\ \ L_n = (1+\frac{t^2}{2n}+\ldots)^n,\ as\ n->\infty$$

$$\log L_n = n\log(1+\frac{t^2}{2n}+\ldots)$$

$$= \frac{\log(1+\frac{t^2}{2n}+\ldots)}{1/n}$$

$$Both\ num\ and\ \ denom \rightarrow 0,\quad as\ n->\infty$$

As n tends to infinity…

$$take\ derivative\ of\ numerator\ and$$

$$numerator\ as\ per\ L'Hospital\ rule$$

$$= \frac{\dfrac{(-\dfrac{t^2}{2n^2})}{(1+\dfrac{t^2}{2n}+\ldots)}}{-1/n^2} = \frac{\dfrac{t^2}{2}}{(1+\dfrac{t^2}{2n}+\ldots)}$$

$$= \frac{t^2}{2},\ as\ n \to \infty$$

$$same\ as\ the\ mgf\ of\ Z$$

Interval Estimate

Sample Mean and Population Mean

- $X_1, X_2, X_3, \ldots, X_n$ is a sample from a normal distribution having unknown mean μ and known variance σ^2.
- Maximum likelihood point estimator of μ is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$\bar{X}$

- We know that $\bar{X}$ is normally distributed with mean μ and known standard deviation $\sigma/\sqrt{n}$
- So the following is standard normal distribution:

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

95% confidence interval

$$P\left[-1.96 < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right] = 0.95$$

$$\Rightarrow P\left[-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$$\Rightarrow P\left[\bar{X}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

A manufacturing situation

Suppose that a machine part manufacturer has made parts with the dimensions as given below:

(a) 9 pieces of the machine part

(b) dimensions are respectively

5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5

Suppose somehow it is known that IF the parts could be measured on the whole population sample by sample, the variance of the measurements would be 4

(*artificial? Yes, but useful for concept building*)

95% confidence interval for μ

5+8.5+12+15+7+9+7.5+6.5+10.5=81

$\bar{X}$=81/9=9

It follows that under the assumption that the values are independent, a 95% confidence interval for μ is

[9-1.96.(2/3), 9+1.96.(2/3)]

=**(7.6, 10.31)**

Interpretation of the observation

Based on the

(a) observation (9 samples)

(b) knowledge obtained somehow that variance is 4

We reach the 95% confidence interval as (7.69, 10.31)

Qualitatively

- If the manufacturer says, “I can ensure a part dim of 15”, we cannot trust!!
- If the maker says, “I assure dim of 10”, we CAN trust
- Trust with 95% confidence

End recap

What if the sample size increases? 95% confidence interval for μ

Let the mean remain the same: 9, also the s.d is 2. But *n=100*

$$P\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

Then a 95% confidence interval for μ is

[9-1.96.(2/10), 9+1.96.(2/10)]=**(8.6, 9.4)**

As the sample size increases, the interval around the sample mean is tighter

Two sided and one sided confidence intervals

- What we saw is 2 sided confidence interval

- Similarly, one sided upper and lower confidence intervals

95% one sided intervals

- Upper

$$\left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right)$$

- Lower

$$\left(-\infty, \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\right)$$

Towards test of hypothesis

Problem statement *(Sheldon M. Ross, PSES, 2004)*

All cigarettes presently on the market have an average nicotine content of at least 1.6mg per cigarette. A firm that produces cigarettes claims that it has discovered a new way to cure tobacco leaves that will result in the average nicotine content of a cigarette being less than 1.6 mg. To test this claim, a sample of 20 of the firms cigarettes were analysed. If it is known that the standard deviation of a cigarette’s nicotine content is 0.8 mg., what conclusions can be drawn at the 5% level of significance if the average nicotine content of the 20 cigarettes is 1.54?

95% one sided intervals

- Upper

$$\left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}, \infty\right) = \left(1.54 - 1.645\frac{0.8}{\sqrt{20}}, \infty\right) = \left(1.24, \infty\right)$$

- Lower

$$\left(-\infty, \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\right) = \left(-\infty, 1.54 + 1.645\frac{0.8}{\sqrt{20}}\right) = \left(-\infty, 1.84\right)$$

Calculate the probability of >=1.6mg

- Z value for 1.6mg

$$\left(\bar{X}-1.6\right)/\frac{\sigma}{\sqrt{n}}=\left(1.54-1.6\right)/\frac{0.8}{\sqrt{20}}=-0.335$$

- Probability>=1.6mg=z-score from -0.335 to +infinity
- = from z-score table, 1.0-0.37=0.63
- There is 63% probability that the nicotine content is >=1.6mg
- This value is less than 95%

Analysis

1. There is 95% probability that the nicotine content will lie in the range $(-\infty, 1.84)$, i.e., can go up to 1.84
2. There is 95% probability that the nicotine content will lie in the range $(1.24, +\infty)$, i.e. can be at least 1.24
3. There is 63% probability that the nicotine content is $>=1.6$mg

We note these observations and do not reach any conclusion yet

Terminology for Test of hypothesis

Terminology

Null and Alternative Hypothesis

- H_0: Null Hypothesis → the hypothesis we want to reject

- H_A or H_1: Alternative Hypothesis → opposite of H_0

- We use the sample statistics, trying to reject H_0

H_0 and H_A for manufacturing-part problem

Dimensions

5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5

Variance: 4 (somehow known)

Sample mean: 9

Company “claims” av. dimension as 15

H_0: Dim>=15, H_A: Dim < 15 (one sided)

Type I and Type II error

- Type I: incorrectly reject H_0, when it should have been accepted.

- Type II: incorrectly accept H_0 when it should have been rejected.

More on H_0

- The data would be unlikely to occur if the null hypothesis were true.

- In logical form:

$$N_H \;\; |\text{-} \sim D$$

- Where N_H is the proposition “null hypothesis true” and D is the proposition “Data occurs”

Digression: Hypothesis Testing in Logic

Using Predicate Calculus

Himalayan Club example

- Introduction through an example *(Zohar Manna, 1974):*
  - Problem: A, B and C belong to the Himalayan club. Every member in the club is either a mountain climber or a skier or both. A likes whatever B dislikes and dislikes whatever B likes. A likes rain and snow. No mountain climber likes rain. Every skier likes snow. *Is there a member who is a mountain climber and not a skier?*
- Given knowledge has:
  - Facts
  - Rules

Example contd.

- Let *mc* denote mountain climber and *sk* denotes skier. Knowledge representation in the given problem is as follows:
  1. *member(A)*
  2. *member(B)*
  3. *member(C)*
  4. $\forall x[member(x) \rightarrow (mc(x) \vee sk(x))]$
  5. $\forall x[mc(x) \rightarrow \sim like(x, rain)]$
  6. $\forall x[sk(x) \rightarrow like(x, snow)]$
  7. $\forall x[like(B, x) \rightarrow \sim like(A, x)]$
  8. $\forall x[\sim like(B, x) \rightarrow like(A, x)]$
  9. *like(A, rain)*
  10. *like(A, snow)*
  11. *Question:* $\exists x[member(x) \wedge mc(x) \wedge \sim sk(x)]$
- We have to infer the 11^{th} expression from the given 10.
- Done through Resolution Refutation.

Club example: Inferencing

1. *member(A)*
2. *member(B)*
3. *member(C)*
4. $\forall x[member(x) \rightarrow (mc(x) \vee sk(x))]$
   - Can be written as $[member(x) \rightarrow (mc(x) \vee sk(x))]$
   - $\sim member(x) \vee mc(x) \vee sk(x)$
5. $\forall x[sk(x) \rightarrow lk(x, snow)]$
   - $\sim sk(x) \vee lk(x, snow)$
6. $\forall x[mc(x) \rightarrow \sim lk(x, rain)]$
   - $\sim mc(x) \vee \sim lk(x, rain)$
7. $\forall x[like(A, x) \rightarrow \sim lk(B, x)]$
   - $\sim like(A, x) \vee \sim lk(B, x)$

8. $\forall x[\sim lk(A,x) \rightarrow lk(B,x)]$

- $lk(A,x) \vee lk(B,x)$

9. $lk(A, rain)$

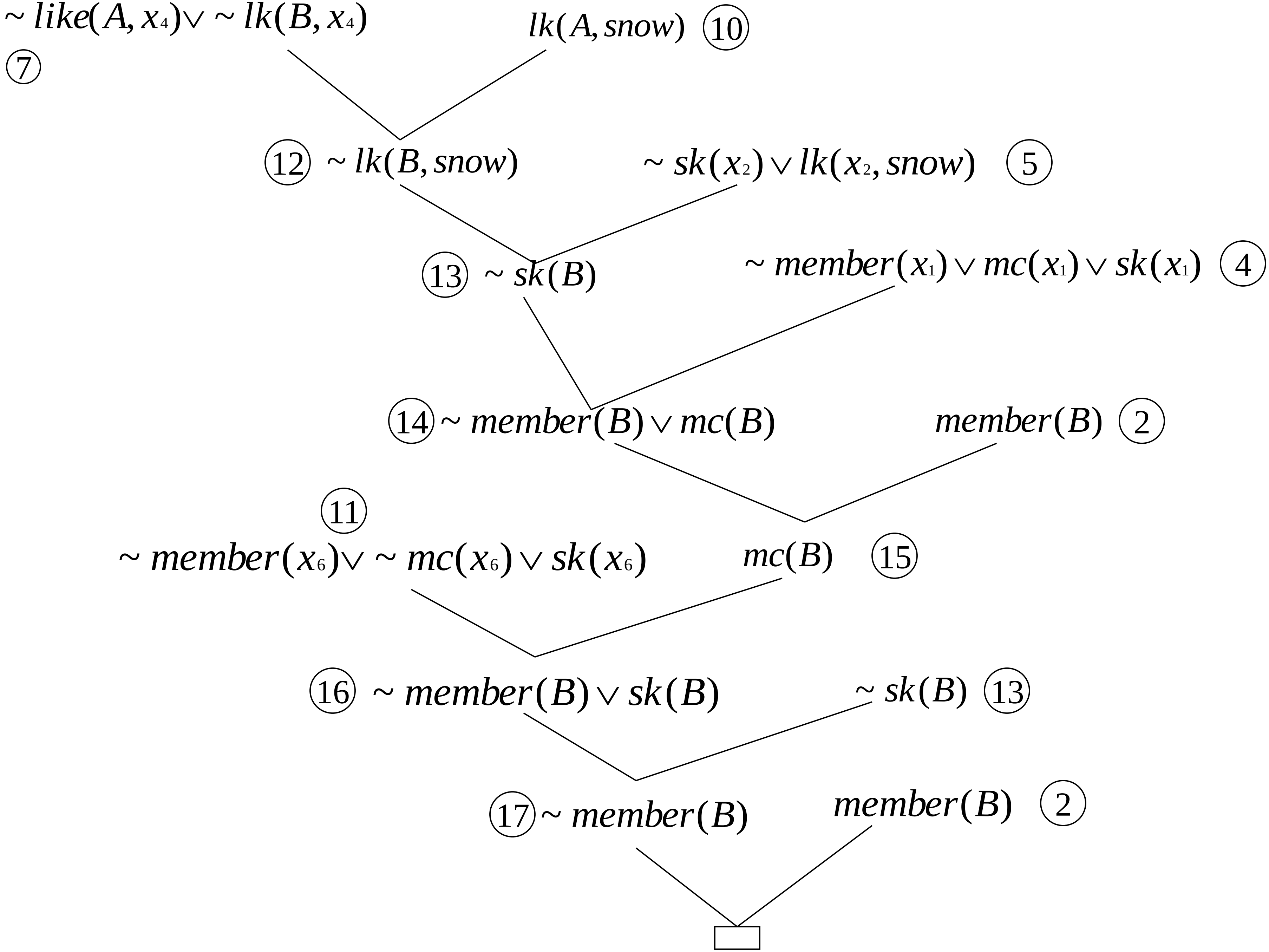
10. $lk(A, snow)$

11. $\exists x[member(x) \wedge mc(x) \wedge \sim sk(x)]$

- Negate– $\forall x[\sim member(x) \vee \sim mc(x) \vee sk(x)]$

- Now standardize the variables apart which results in the following

1. $member(A)$
2. $member(B)$
3. $member(C)$
4. $\sim member(x_1) \vee mc(x_1) \vee sk(x_1)$
5. $\sim sk(x_2) \vee lk(x_2, snow)$
6. $\sim mc(x_3) \vee \sim lk(x_3, rain)$
7. $\sim like(A, x_4) \vee \sim lk(B, x_4)$
8. $lk(A, x_5) \vee lk(B, x_5)$
9. $lk(A, rain)$
10. $lk(A, snow)$
11. $\sim member(x_6) \vee \sim mc(x_6) \vee sk(x_6)$



~ like(A, x4) ∨ ~ lk(B, x4) 7
lk(A, snow) 10
12 ~ lk(B, snow)
~ sk(x2) ∨ lk(x2, snow) 5
13 ~ sk(B)
~ member(x1) ∨ mc(x1) ∨ sk(x1) 4
14 ~ member(B) ∨ mc(B)
member(B) 2
11
~ member(x6) ∨ ~ mc(x6) ∨ sk(x6)
mc(B) 15
16 ~ member(B) ∨ sk(B)
~ sk(B) 13
17 ~ member(B)
member(B) 2

Null Hypothesis: H_0

- H_0: The club does NOT have any member that is a mountain climber (MC) and not a skier (SK)

- Key question: Under H_0, is the observation valid?

- In other words: is the hypothesis consistent with the data?

Methodology

- If Hypothesis not consistent with data, hypothesis must be rejected
- Data cannot be rejected
- Data is GOLD!

Data for Himalayan Club Example

(1) A, B and C belong to the Himalayan club.

(2) Every member in the club is either a mountain climber or a skier or both.

(3) A likes whatever B dislikes and

(4) dislikes whatever B likes.

(5) A likes rain and snow.

(6) No mountain climber likes rain.

(7) Every skier likes snow

Null Hypothesis for Himalayan Club Example

- H_0: *There is NOT a single member who is a mountain climber and not a skier*
- H_0 inconsistent with data

- So must be rejected
- Methodology: Logical Inferencing-Resolution-Refutation

More on H_0

- Maximum Likelihood in action
- Move the ball to the “court” of observations
- Formulate H_0 in such a way that high probability of H_0 makes the data probability low

p-value

- The **p-value**, or probability **value**, tells you how likely it is that your data could have occurred under the null hypothesis. ... The **p-value** is a proportion:
- **p-value** of 0.05 **means** that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true.

Nicotine Problem

Problem statement *(Sheldon M. Ross, PSES, 2004)*

All cigarettes presently on the market have an average nicotine content of at least 1.6mg per cigarette. A firm that produces cigarettes claims that it has discovered a new way to cure tobacco leaves that will result in the average nicotine content of a cigarette being less than 1.6 mg. To test this claim, a sample of 20 of the firms cigarettes were analysed. If it is known that the standard deviation of a cigarette's nicotine content is 0.8 mg., what conclusions can be drawn at the 5% level of significance if the average nicotine content of the 20 cigarettes is 1.54?

Solution to the Nicotine problem

- First decide H_0
- Requirement: the probability of rejecting H_0 when it is true will never exceed α
- We should test
- H_0: $\mu>=1.6$ versus H_1: $\mu<1.6$

Nicotine problem cntd.

Value of test statistic is

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} = \frac{\sqrt{20}(1.54-1.6)}{0.8}$$

$$= -0.336$$

So the p-value is given by

p-value=P[Z<-0.336], Z~N(0,1)

=0.368

Conclusion from the nicotine problem

- 0.368 > 0.05
- Foregoing data do not enable us to reject at the 0.05 percent level of significance the hypothesis that "mean nicotine content exceeds 1.6"
- In other words, the evidence though supporting the cigarette producer's claim is not strong enough to prove the claim