Drawing conclusions from samples

↳ Imagine we want to average the weight of all the rats
↳ Solution: 01:
  ↳ Measure weights of all rats on earth
  ↳ Take average
↳ This is NOT possible.

↳ Solution: 02:
  ↳ We measure weights of a small set of rats at random
  ↳ We can say a great deal from sufficient sample.

What is population?

↳ It's the entire possible dataset $\{X\}$
↳ It has countable size N_p.
↳ Population mean popmean $(\{X\})$ is a number.
↳ Population std popstd $(\{X\})$ is a number.

What is a sample?

↳ It is a random subset of population and done with replacement.
↳ Sample size N << population size N_p
↳ Sample mean of a population is $X^{(N)}$ & is a random variable.

Sample mean of Population

- Imagine you want to understand something about a big group of things (a population), like the average height of all students in a school. Instead of measuring everyone, you randomly pick a smaller group (a sample) to measure.

- If you pick these students in a way where each one has an equal chance of being chosen (this is called "random sampling"), and the choice of one student doesn't affect the choice of another (this is called "independent sampling"), the sample is IID (independent and identically distributed).

- When you calculate the average (mean) of your sample, it will usually be close to the true average (mean) of the entire population, especially if you take a large enough sample. This happens because the randomness of how you pick and the independence between your choices balance things out.

$$X^{(N)} = \frac{1}{N}\left(x_1 + x_2 + \cdots + x_N\right)$$

$$\mathcal{E}\left[X^{(N)}\right] = \frac{1}{N}\left(\mathcal{E}\left[X^{(1)}\right] + \cdots + \mathcal{E}\left[X^{(N)}\right]\right)$$

$$= \frac{1}{N} N \cdot E[X^{(1)}]$$

$$= E[X^{(1)}]$$

↳ Expected value of multiple sample sets is equal to mean of one sample on average

↳ By this property, sample mean = population mean

Standard deviation of Sample mean:

$$\text{Var}[X^{(N)}] = \frac{\text{popvar}(\{X\})}{N}$$

$$\text{std}[X^{(N)}] = \frac{\text{pop std}(\{X\})}{\sqrt{N}}$$

↳ More samples, lesses the std$[X^N]$ i.e popmean estimate gets better.

↳ Improves slowly i.e to halve the std, we need to draw 4 times as many samples as $\sqrt{4} = 2$.

But we need popVar & popstd!!!

Unbiased Estimate of popsd & stderr.

↳ The unbiased estimate of popsd$(\{X\})$ is:

$$\text{stdunbiased}(\{X\}) = \sqrt{\frac{1}{N-1} \sum_{x_i \in \text{sample}} (x_i - \text{mean}(\{x_i\}))^2}$$

↳ std of mean of all samples

↳ So,

$$\text{std}[X^N] = \frac{\text{popsd}(\{X\})}{\sqrt{N}}$$

$$\frac{\text{pop std}(\{X\})}{\sqrt{N}} = \frac{\text{stdunbiased}(\{x\})}{\sqrt{N}} = \text{stderr}(\{x\})$$

Understanding sample accuracy:

↳ Stderr$(\{X\})$ tells about accuracy of sample mean as an estimator for the pop mean.

↳ Smaller stderr means more accuracy.

Improving estimates with more samples.

↳ As,

$$\text{stderr} \propto \frac{1}{\sqrt{N}}$$

increasing N better stderr

Example:

Sample	Result
1211	Epsy 49% : 51% hyde-smith

$$\text{std}_{\text{unbiased}} = \sqrt{\frac{1}{1211-1}\left[618(1-0.51)^2 + 593(0-0.51)^2\right.}$$

$$= 0.5$$

$$\text{Stderr} = \frac{0.5}{\sqrt{1211}} = 0.0144$$

↳ Sample mean is a RV & has it's own probability distribution, stderr is an estimate of sample mean's std deviation.

↳ When N is very large, sample mean is approaching a normal distribution.

$$\mu = \text{popmean}(\{x\}) \; ; \; \sigma = \text{stderr}(\{x\})$$

- The standard deviation measures the spread of data points around the mean of the sample or population. i.e. how much each individual observation in the dataset deviates from the mean.
- The standard error, measures the spread of the sample means around the true population mean. It reflects how much variability you would expect if you were to take many different samples from the same population..

Confidence intervals:

↳ Confidence interval for pop mean is defined by fraction.
↳ Given a percentage, find how many units of stderr it covers.

↳ For 95% of realized sample mean, the pop mean lies in:

For about 95% = [Sample mean − 2 stderr , Sample mean + 2 stderr]

For about 68 % = [Sample mean − stderr , Sample mean + stderr]

For about 99.7% = [Sample mean − 3 stderr , Sample mean + 3 stderr]

OR

↳ To calculate CI for a given confidence level, we find Area

$$CL = 1 - \alpha$$

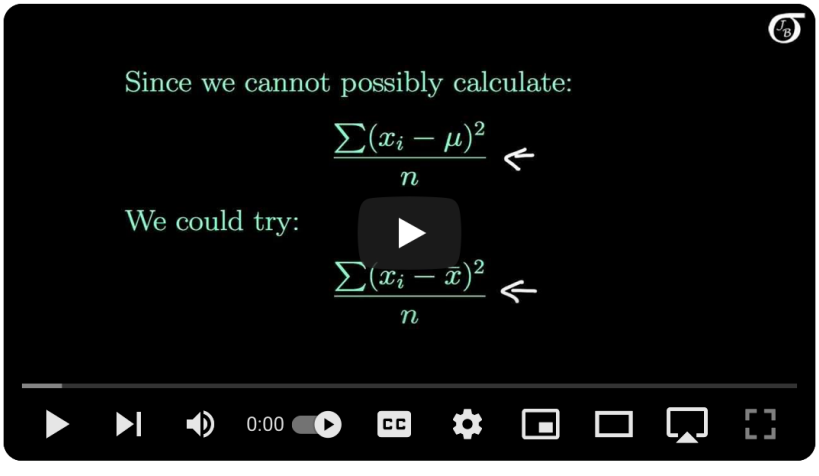
$$\alpha = 1 - CL$$

↳ α is area remaing in both left & right tails.
↳ $\frac{\alpha}{2}$ for each tail.

Therefore,

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Why do we divide by n−1?



The Sample Variance: Why Divide by n-1?

Why do we multiply z-score with std-err?

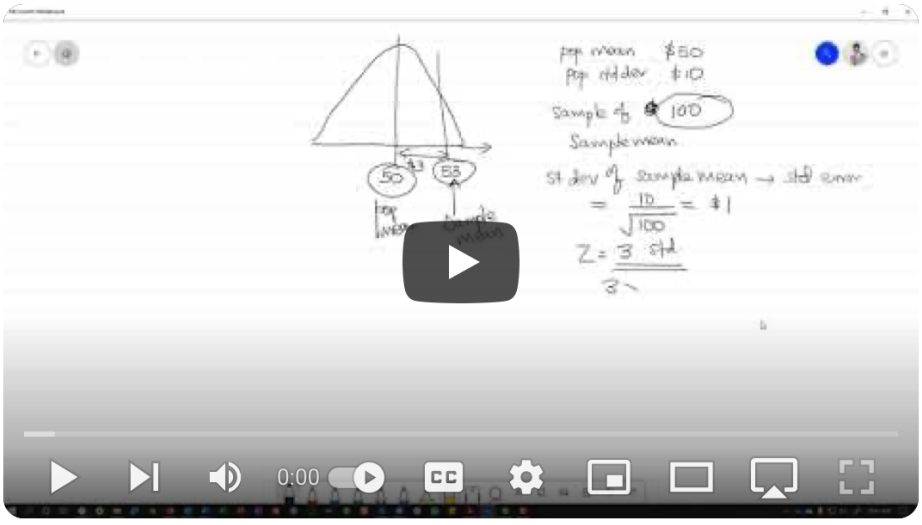
- **Repeated Sampling:** In theory, repeated sampling and averaging the sample means would give us the true population mean, but it's time-consuming.
- **Single Sample with Confidence Interval:** Instead of multiple samples, we use just one sample and construct

a confidence interval around its mean. This interval captures the uncertainty around the sample mean and provides a range where the true population mean is likely to fall.

- **Practical Balance:** This method is much more efficient, and the interval allows us to estimate the true mean without needing to take many samples.

2. Why Multiply by the Standard Error?

- The standard error quantifies the variability in the sample mean:
  - Even though you have a sample mean, it is not exact—it varies from sample to sample because it is based on a random subset of the population. Multiplying the critical value (z) by the standard error scales the variability appropriately for the confidence level, giving you the margin of error.
  - This margin of error reflects how much the sample mean could reasonably differ from the true population mean, given the data.



Why do we multiply Z score with Standard Error?

T-distribution:

What happens when N is small?

↳ A sample drawn (with $N < 30$) from a normally distributed population follows the student's t-distribution.
↳ Shape depends on df. → Degree of freedoms represent how many values are free to vary after estimating something like mean.
↳ When df → ∞, the distribution becomes normal distribution.

Why we can't use z-distribution?

- **Sampling Variability**
  - The sample standard deviation is calculated using a limited number of data points. When the sample size

N is small:
    - The values in the sample may not represent the full range of variability present in the population.
    - Smaller samples are more likely to either overestimate or underestimate the true population variability due to random chance.
    -
    - **For example:**
        - If you randomly draw a small sample, it might include only extreme values (overestimating) or only central values (underestimating) sample stdev.
        - With larger samples, the sample standard deviation stabilizes as it incorporates more of the population's variability.

- **Bias in the Estimation of Variance**
    - The formula for variance used in a sample divides by. N - 1 to correct for the tendency of small samples to underestimate the population variance. This correction accounts for the fact that the sample mean is an imperfect estimate of the population mean.

    - **Why does this happen?**
        - The sample mean is calculated based on the sample data, so it's closer to the sample values than the true population mean would be. This artificially reduces the variability within the sample.
        - For small samples, this effect is more pronounced, leading to an underestimation of the true variability.

- **Lack of Representation**
    - Small samples are less likely to represent the full diversity or spread of the population:
    - If the population has outliers or rare events, these are less likely to appear in a small sample, skewing the estimate of variability.
    - As the sample size increases, the probability of capturing the full spectrum of population variability increases, improving the accuracy of sample as an estimate for variance.

- **Sensitivity to Outliers**
    - Small samples are more sensitive to the presence of outliers:
    - A single extreme value in a small sample can disproportionately inflate the sample standard deviation.
    - In larger samples, the impact of outliers is diluted, making sample mean a more robust estimator of population stdev.

Mathematical Dependence on Sample Size
    - The formula for sample mean relies on dividing by N - 1, which makes it inherently more sensitive to small sample sizes. When N is small:
    - The denominator in the variance formula N - 1 is small, which amplifies the variability of sample mean across repeated samples.
    - This means sample mean fluctuates more from one small sample to another compared to larger samples.

How does t-distribution helps?

- The t-distribution adjusts for this extra uncertainty by having heavier tails than the normal distribution. This means:
  - For a given confidence level (e.g., 95%), the critical t-value will be larger than the corresponding z-value.
  - A larger t-value increases the margin of error, which widens the confidence interval to account for the additional uncertainty in small samples.

- The degree to which the t-distribution is wider depends on the degrees of freedom (df=n−1df=n−1):
  - For very small N, the t-distribution is much wider.
  - As N increases, the t-distribution converges to the z-distribution because the sample standard deviation becomes a more reliable estimate of σ.

So,

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{6}}$$

Example slide #47

$\bar{x} = 0.29$

stdunbiased = 0.0739

stderr = 0.0174

$\alpha = 1 - 0.99$
$= 0.01$

$\alpha/2 = 0.005$

$t_{0.005,17} = 2.898$

$[0.29 - 2.898 \times 0.0174 , 0.29 + 2.898 \times 0.0174]$

$[0.239, 0.34]$

Confidence interval for other statistics:

↳ We were able to produce convenient & useful estimates of standard error for sample means.

↳ Other statistics like median are also useful about drawing interesting conclusion about the population.

↳ It's often difficult to derive analytical expression in terms of stderr for a corresponding random variable.

↳ So, we use bootstrapping.

Bootstrapping

↳ It is a method for constructing confidence interval for other statistics using re-sampling of the sample dataset.

↳ It is essentially uniform random sampling with replacement on sample of size N.

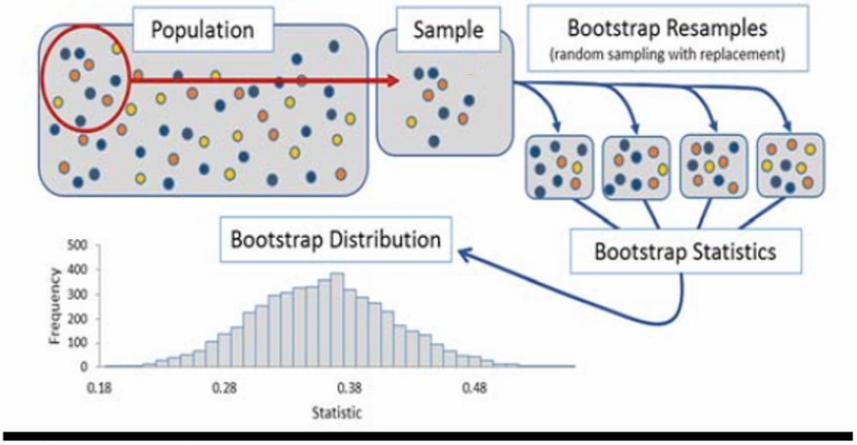
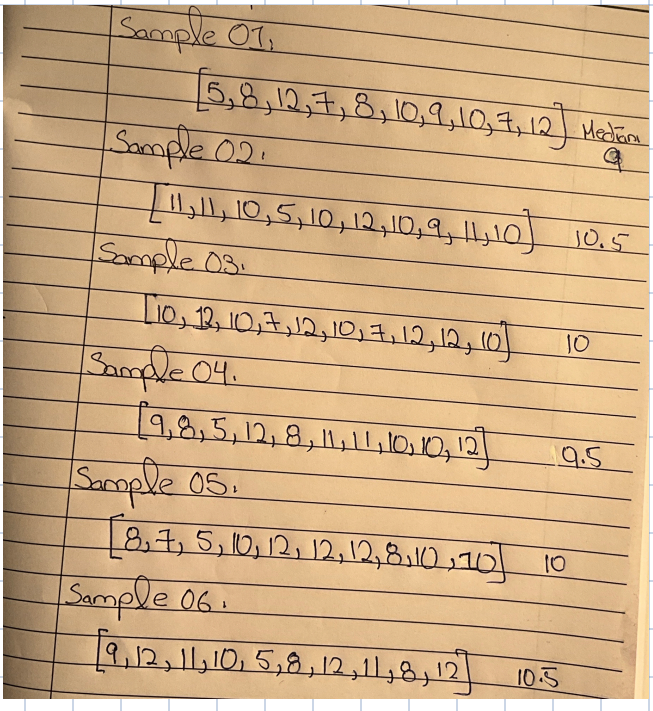


Figure 1. Summary of Bootstrapping Process

Example,



$\bar{X}_{median} = 9.91$

$std_{unbiased} = 0.58$

$std_{err} = 0.238$

$\alpha = 1 - 0.95$
$= 0.05$

$\frac{\alpha}{2} = 0.025$

$t_{0.025,5} = 2.57$

$[9.91 - (2.57 \times 0.238),\ 9.91 + (2.57 \times 0.238)]$

$[9.29, 10.52]$

Assignment 02 :

Task:01

$$Y = \begin{cases} X-1, & 1 \le x \le 11 \\ X+1, & 1 \le x \le 11 \end{cases}$$

$$p(R) = \frac{26}{52}$$

$$X = \begin{cases} 11, & 0.23 \\ 1, & 0.0769 \\ 2 & 0.0769 \\ 3 & 0.0769 \\ 4 & 0.0769 \\ 5 & 0.0769 \\ 6 & 0.0769 \\ 7 & 0.0769 \\ 8 & 0.0769 \\ 9 & 0.0769 \\ 10 & 0.0769 \end{cases}$$

$$p(2) = \frac{4}{52} = p(3) = \cdots$$

$$p(k \cup Q \cup J) = \frac{12}{52}$$

$$p(A) = \frac{4}{52}$$

Black 0.5 | Red 0.5

X	≤ X+1	≥ X−1
1	>= 2	>= 0
2	>= 3	>= 1
3	>= 4	>= 2
4	>= 5	>= 3
5	>= 6	>= 4
6	>= 7	>= 5
7	>= 8	>= 6
8	>= 9	>= 7
9	>= 10	>= 8
10	>= 11	>= 9
11	>= 12	>= 10

$Q_1.\ P(X <= 2) = 0.0769 + 0.0769$
$= 0.1538$

$Q_2.\ P(X >= 10) = 0.0769 + 0.23$
$= 0.3069$

$Q_3.\ P(X >= Y) = 0.5$

$Q_5.\ P(Y >= 12) = \frac{3 \times 2}{52}$
$= 0.1154$

$Q_4.\ Y - X$

$X+1 - X$: 1

$X-1 - X$: −1

$-1 \le Y - X \le 1$

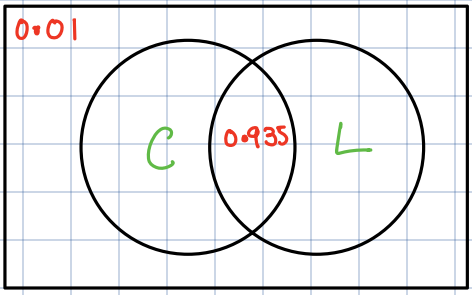
Task:03

$P(C \cap L) = \frac{187}{200}$, $P(\overline{C \cup L}) = \frac{2}{200}$

Q1. $P(C \cup L) = 1 - P(\overline{C \cup L})$

$= 0.99$

Q2.



$P(\overline{L}) = x$ $P(\overline{C}) = \frac{x}{4}$

$P(C \cup L) = P(C) + P(L) - P(C \cap L)$

$0.99 = (1 - \frac{x}{4}) + (1 - x) - 0.93$

$1.92 = \frac{4-x}{4} + 1 - x$

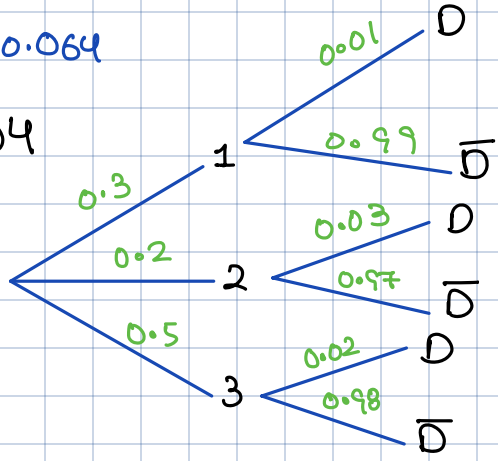
$= \frac{4 - x + 4 - 4x}{4}$

$7.68 = -5x + 8$

$+0.32 = +5x$

$x = 0.064$

Task :04



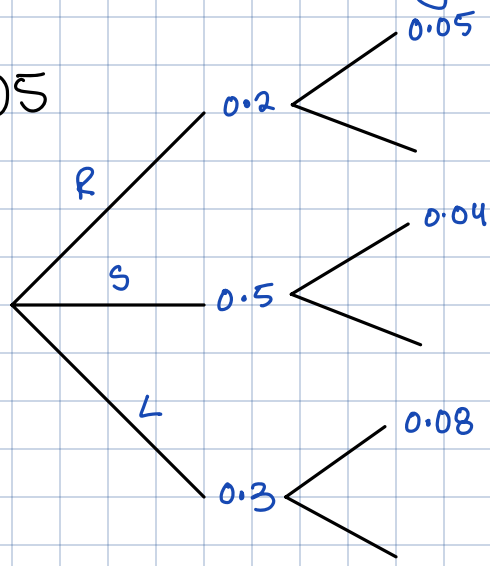
$$P(D) = 0.3 \times 0.01 + 0.2 \times 0.03 + 0.5 \times 0.02$$
$$= 0.019$$

$$P(D|P_1) = \frac{0.3 \times 0.01}{0.019} \qquad P(D|P_2) = \frac{0.2 \times 0.03}{0.019}$$
$$= 0.157 \qquad = 0.31$$

$$P(D|P_3) = \frac{0.5 \times 0.02}{0.019}$$
$$= 0.52$$

↳ Plan 3 was most likely

Task: 05



(a) $P(F) = 0.2 \times 0.05 + 0.5 \times 0.04 + 0.3 \times 0.08$
$= 0.054$

(b) $P(F|L) = \frac{0.3 \times 0.08}{0.054}$

$= 0.444$

Task: 06

$f(x,y)$

y \ x	1	2	3	
1	0.05	0.05	0.1	0.2
3	0.05	0.1	0.35	0.5
5	0	0.2	0.1	0.3
	0.1	0.35	0.55	

Q2. $P(y) = 0.2, 0.5, 0.3$

Q1. $P(x) = 0.1, 0.35, 0.52$

Q3. $P(Y=3|X=2) = \frac{P(Y=3 \cap X=2)}{P(X=2)}$

$= \frac{0.1}{0.35}$

$= 0.28$

Task: 08

$P(A \cap B) = (0.7)^2$

$P(C \cap D \cap E) = (0.8)^3$

$P(A \cap B \cap C \cap D \cap E) = (0.7)^2 \times (0.8)^3 = 0.25$

Q1. $P((A \cap B) \cup (C \cap D \cap E)) = (0.7)^2 + (0.8)^3 - 0.25$ $P(A') + P(\text{system works}) - P(A' \cup \text{system works})$

$= 0.75$

Q2. $P(\overline{A} \mid \text{system\_works}) = \frac{P(\overline{A} \cap \text{system\_works})}{P(\text{system\_works})}$

$= \frac{0.3 \times (0.8)^3}{0.7511}$

$= 0.2$

Total master keys = 8

40% are open

1 master key will open a house at any time

P(gets into house) = P(house is open) + p(house locked but opened by key)

$= 0.4 + 0.375 \times 0.6 = 0.625$

$P(\text{key works}) = \frac{7C2 \times 1C1}{8C3}$

$= 0.375$

$= \frac{3}{56}$

Task :09

$P(A) = 2P(B)$, $P(C|A) = \frac{2}{7}$, $P(C|B) = \frac{4}{7}$

$$P(C|A \cup B) = \frac{P[(A \cup B) \cap C]}{P(A \cup B)}$$

$$= \frac{P(B \cap C) + P(C \cap A)}{3P(B)}$$

$P(C|A) = \frac{P(A \cap C)}{P(A)}$ $P(C|B) = \frac{P(B \cap C)}{P(B)}$ $P(A \cup B) = P(A) + P(B)$

$= 2P(B) + P(B)$

$= 3P(B)$

$P(A \cap C) = P(A) \times P(C|A)$ $P(B \cap C) = P(C|B) \times P(B)$

$= 2P(B) \times \frac{2}{7}$ $= \frac{4}{7} P(B)$

$= \frac{4}{7} P(B)$

$$= \frac{\frac{4}{7} P(B) + \frac{4}{7} P(B)}{3P(B)}$$

$$= \frac{1.14P(B)}{3P(B)}$$

$$= \frac{1.14}{3}$$

$= 0.38$

Task :07

4 dimes & 2 nickels

1. 3 dimes, 0 nickels 2. 2 dimes, 1 nickel 3. 1 dime, 2 nickels

$$\frac{4C3 \times 2C0}{6C3} = \frac{4}{20}$$

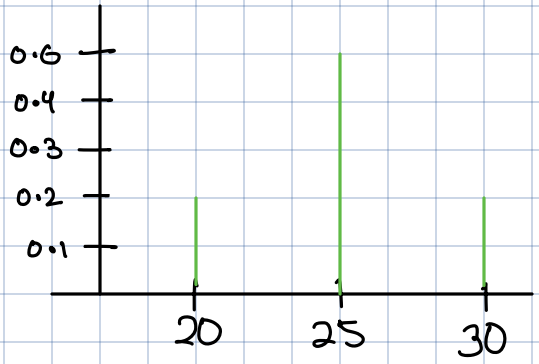
$$= \frac{4C2 \times 2C1}{6C3} = \frac{12}{20}$$

$$= \frac{4C1 \times 2C2}{6C3} = \frac{4}{20}$$

$T = 10 \times 3 = 30$

$= 10 + 10 + 5 = 25$

$= 10 + 5 + 5 = 20$



Task: 02

- Each player has 40 deck
- $L_1 = 10$, $L_2 = 20$

Q1.

$S = 30$

$$P(S=0) = P(L_1=0) \times P(L_2=0)$$

$$= \frac{30C7}{40C7} \times \frac{20C7}{40C7}$$

$$= 0 + 0$$

Q2. $D = -10$

$$P(D=0) = \sum_{k=0}^{7} P(L1=k) \times P(L2=k)$$

$$= \sum_{k=0}^{7} \frac{10Ck \times 30C(7-k)}{40C7} \times \frac{20Ck \times 20C[7-k)}{40C7}$$

Q3. $P(L_1 = k) = \frac{10Ck \times 30C(7-k)}{40C7}$, $k = 0, \ldots, 7$

Q4. $P(L_1 \mid L_t = 10) = P(L_1 = k)$

Q5. $P(L_1 \mid L_t = 5) = \frac{5Ck \; 35C(7-k)}{40C7}$