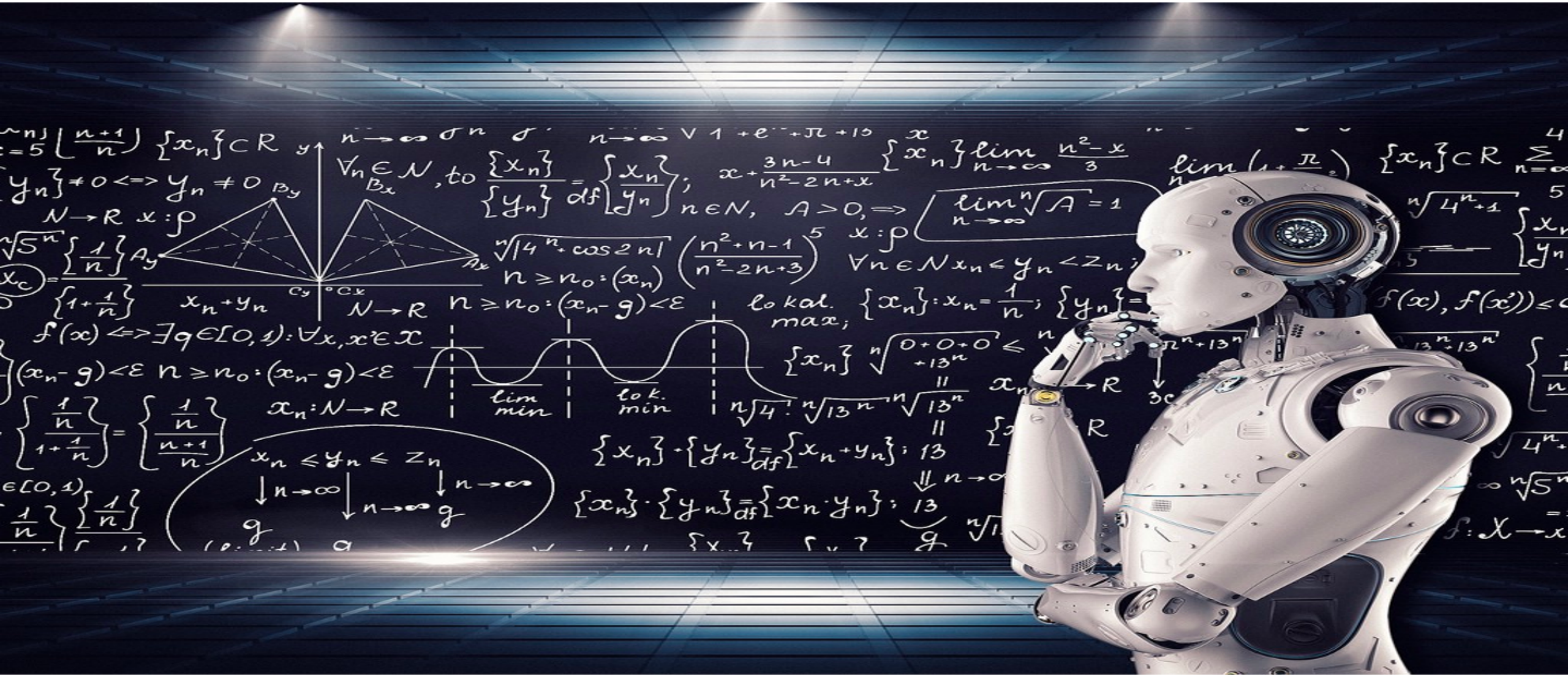
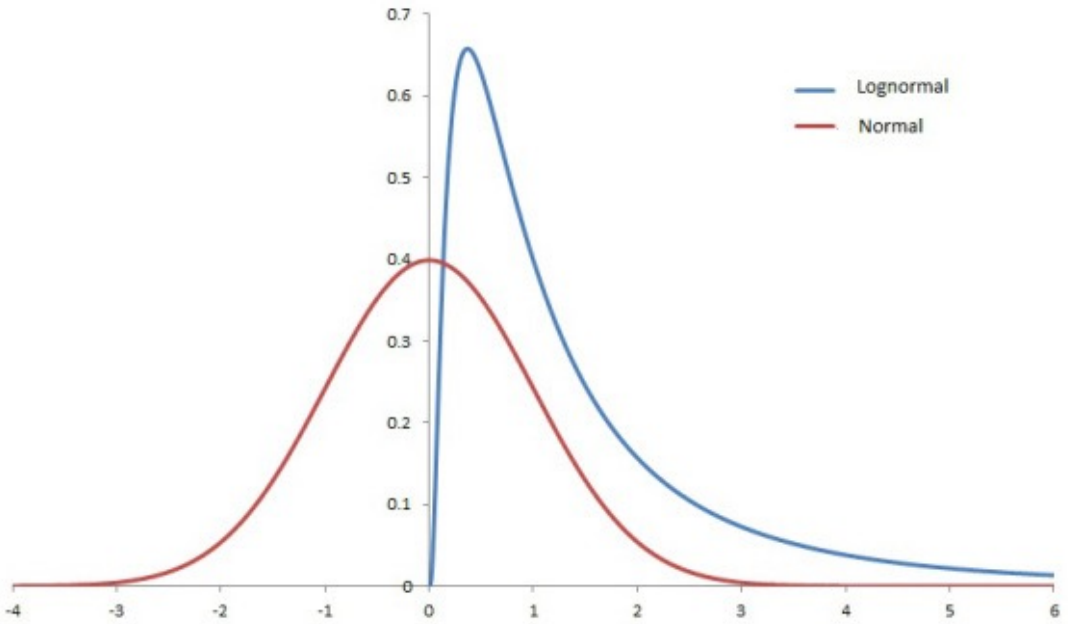
LogN, WRS, HG



Log-normal (Galton) distributions

- A random variable Y is said to have a log-normal distribution if its log, log(Y), has a normal (Gaussian) distribution
- In other words – Y is log-normal if $Y = e^X$ for some Gaussian X.
  That is: $Y = e^{\mu+\sigma Z}$, where Z is standard normal

- Log-normals are always positive
- Can be useful in modelling intrinsically positive quantities
- Mean, mode and median are different from each other
- The log-normal distribution has a heavy right-side tail
- μ and σ are called the *location* and *scale* of Y
  They are NOT the mean and std of Y
- They are the mean and std of X = ln(Y)



The density of the log-normal distribution

- Let Y be a **standard** log-normal random variable. Let f(y) and F(y) be the PDF and CDF of a standard log-normal
- Let Z be standard normal and $\varphi(z)$ and $\Phi(z)$ denote the PDF and CDF of the standard normal distribution

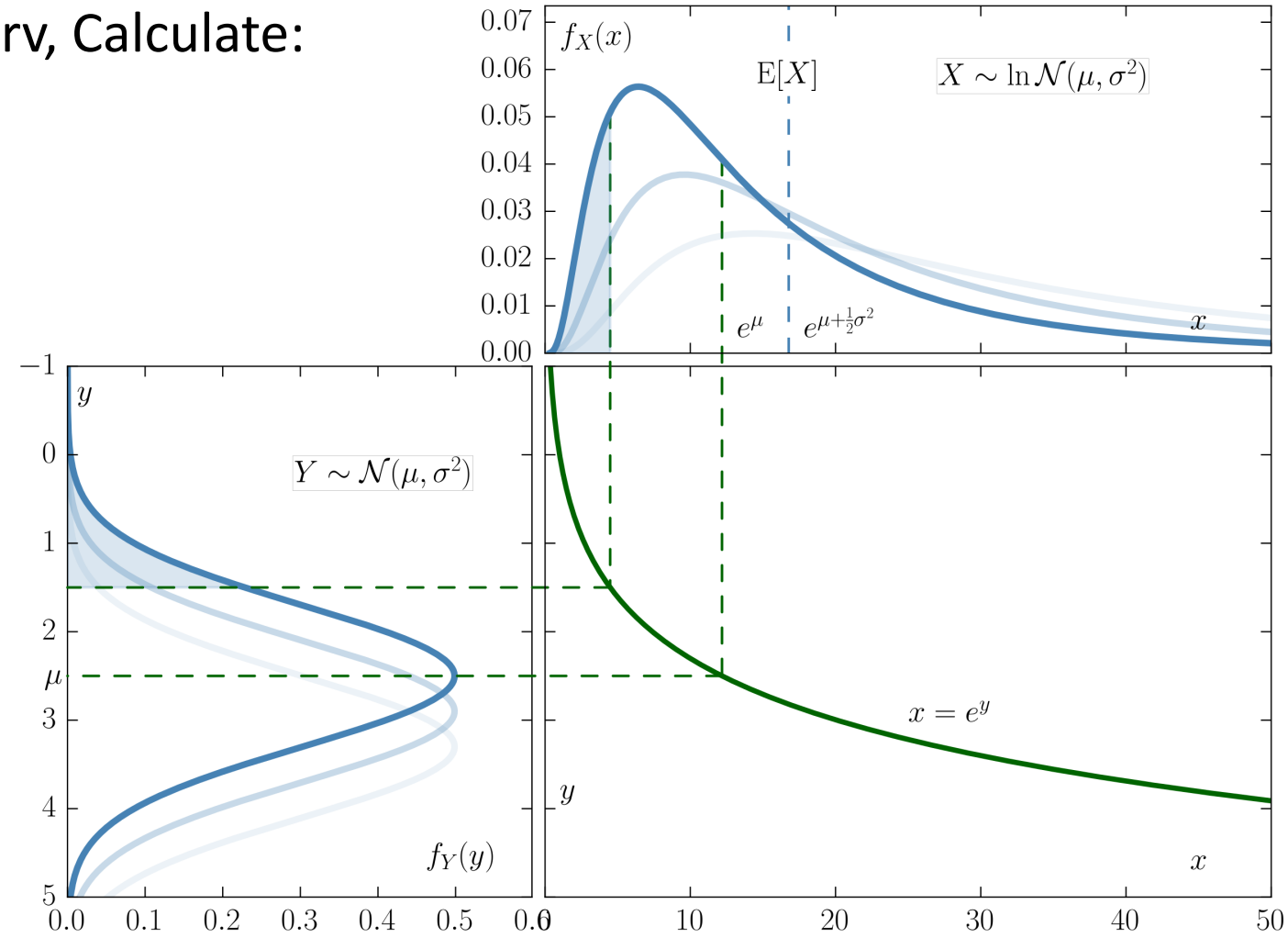
$$F(y) = P(Y \leq y) = P(\, e^{Z} \leq y) = P(\, Z \leq \ln y) \;=\; \Phi(\ln y)$$

- Therefore, since the PDF is the derivative of the CDF, we get:

$$f(y) = F'(y) = \frac{\Phi'(\ln y)}{y} = \frac{\varphi(\ln y)}{y}$$

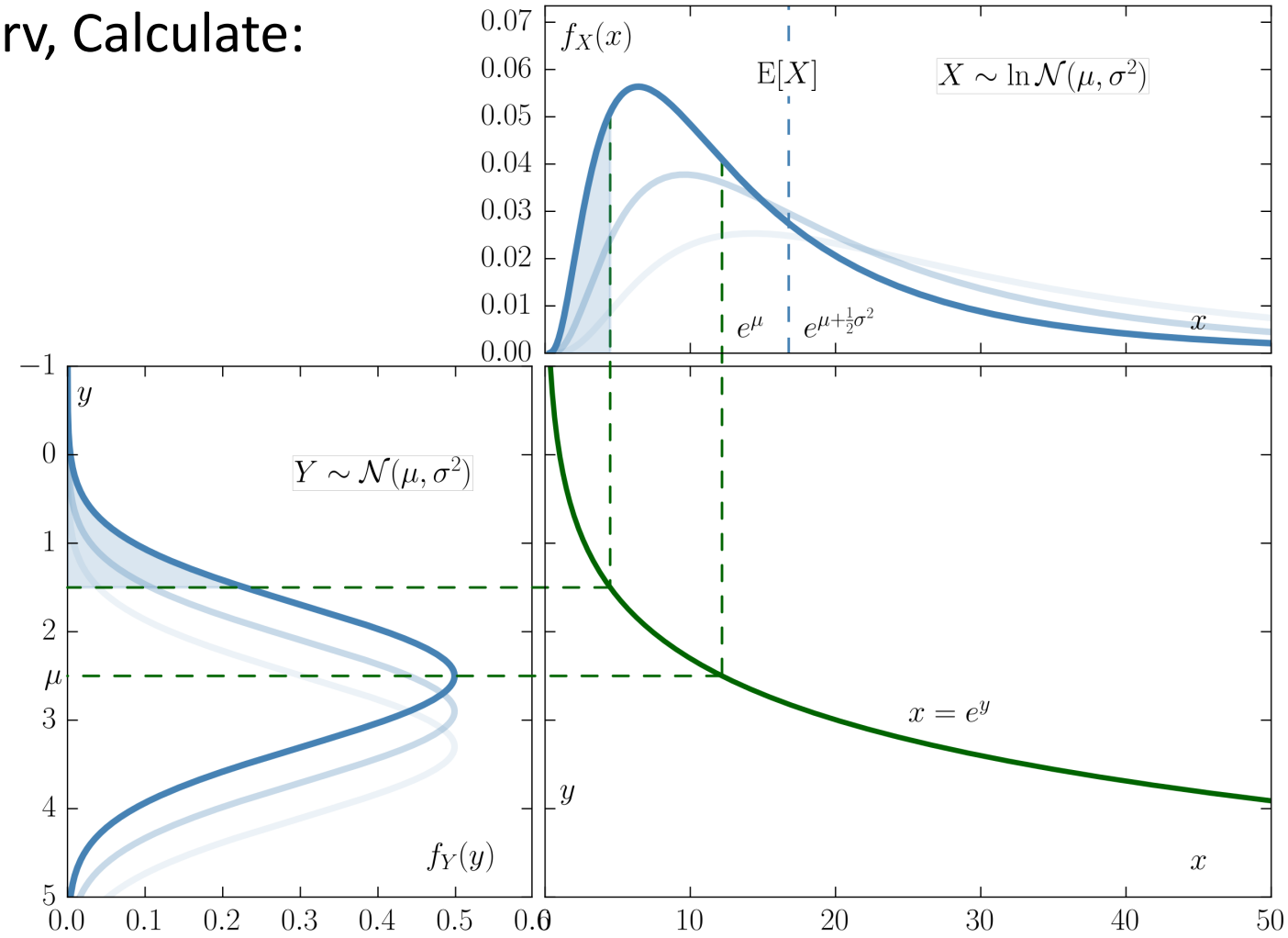
Median and expected value of a log-normal

- Let $\mathrm{Y} = e^{\mu+\sigma z}$ be a log-normal rv, Calculate:
  - Median(Y)
  - E(Y)



Median and expected value of a log-normal

- Let $Y = e^{\mu+\sigma z}$ be a log-normal rv, Calculate:
  - Median(Y) = e^{μ}
    - $P(Y \leq m) = 0.5$
    - $P\left(Z \leq \frac{\ln m - \mu}{\sigma}\right) = 0.5$
    - $\ln m = \mu \rightarrow m = e^{\mu}$
  - E(Y) = $e^{\mu+\frac{1}{2}\sigma^2}$



LogNormal question

- What is the mode of LogNormal distribution? Prove.

- $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
- Since $\ln x$ is strictly increasing, we can maximize $\ln f(x)$ to get the mode
- $g(x) = \ln\big(f(x)\big) = \ln\left(\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)\right) - \ln\big(x\sigma\sqrt{2\pi}\big) = -\frac{(\ln x - \mu)^2}{2\sigma^2} - \ln(x) - \ln\big(\sigma\sqrt{2\pi}\big)$
- $g'(x) = -\frac{\ln x - \mu}{\sigma^2 x} - \frac{1}{x} = -\frac{\ln x - \mu + \sigma^2}{\sigma^2 x} = 0 \rightarrow \ln x = \mu - \sigma^2 \rightarrow x = e^{\mu - \sigma^2}$

LogNormal question

- Let $X \sim LogNormal(\mu_X, \sigma_X^2)$, $Y \sim LogNormal(\mu_Y, \sigma_Y^2)$ be two independent LogNormal random variables. Let $Z = XY$.
- Express the CDF of Z in terms of Φ (the CDF of $N(0,1)$).

- $X = e^U \sim LogNormal(\mu_X, \sigma_X^2)$, $Y = e^V \sim LogNormal(\mu_Y, \sigma_Y^2)$,
- Therefore, using the fact that X and Y are independent to obtain the variance of $U + V$, we get:
- $Z = XY = e^{U+V} \sim LogNormal(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
  Denote $\mu_Z = \mu_X + \mu_Y$ and $\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2}$.
- We then have $CDF_Z(z) = P(Z \leq z) = P(e^W \leq z) = P(W \leq \ln z)$.
- Where $W \sim N(\mu_Z, \sigma_Z^2)$. The last step follows from monotonicity of the log function.
- Thus, $CDF_Z(z) = \Phi\left(\frac{\ln z - \mu_Z}{\sigma_Z}\right)$

LogNormal question

- Random variable R with a lognormal distribution
- 100,000 samples. 51,568 samples have value ≥ 50
- The empirical mean of the data is $\hat{\mu}_{LN} = 409$
- The empirical mode is 1
- What is the value r so that $P(R \leq r) = 0.2$?

- $Mode(R) = e^{\mu-\sigma^2} = 1 \rightarrow \mu = \sigma^2$
- $\hat{\mu}_{LN} = e^{\mu+\frac{\sigma^2}{2}} = 409 \rightarrow \mu + \frac{\sigma^2}{2} = 6$
- $\mu + \frac{\mu}{2} = 6 \quad \rightarrow \quad \mu = 4, \sigma = 2$
- $P(R \leq r) = 0.2$ is equivalent to $r = e^s$ so that s satisfies $P(X \leq s) = 0.2$ where $X \sim N(4,2^2)$
- Therefore $r = e^{4+2\Phi^{-1}(0.2)} = 10.14$

LogNormal question

- You can uniformly reduce all values by a constant factor.

  A process that will reduce all values by a factor of $\alpha < 1$ ($R_{new} = \alpha R$) will cost $3000 + \frac{1000}{\alpha^2}$.

  What is the minimal cost required to get that at most 10% of values are larger than 100?

- $R_{new} = \alpha R = e^{\ln(\alpha)} e^{X} = e^{X+\ln(\alpha)}$, where $X \sim N(4,2^2)$
  - The new RV is lognormal with the underlying μ shifted by $\ln(\alpha)$ and with no change in the underlying σ
- Let α^* be the optimal α to fulfill the requirement
- Let R^* be the R_{new} obtained by using α^*. That is: $R^* = e^{X+\ln(\alpha^*)}$.
- We set $P(R^* \leq 100) = 0.9$ which is equivalent to $P(X + \ln(\alpha^*) \leq \ln(100)) = 0.9$.
- Now, this is equivalent to $\Phi\left(\frac{\ln(100)-4-\ln(\alpha^*)}{2}\right) = 0.9$, and thus: $\frac{\ln(100)-4-\ln(\alpha^*)}{2} = \Phi^{-1}(0.9)$
- $\ln(\alpha^*) = -1.96 \rightarrow \alpha^* = 0.14$
- $Cost = 3000 + \frac{1000}{0.14^2} = 54000$

Wilcoxon Rank Sum test

- Nonparametric alternative to the non-matched (separate populations) t-test.
- We are comparing numbers, or measured quantities, obtained for two different populations.
- Individuals are assumed to be sampled randomly and independently.
- We have two vectors of measured values – one for each population.

Wilcoxon Rank Sum test

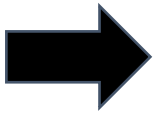
- Consider two sample sets of independently acquired observations from two different labels/populations.
- Example: safety test results for cars manufactured in the Randomistan VW factory vs those made in the German factory.

German factory	Stochastic Heights factory
3.2	3.7
4.5	8.5
8.1	6.1
9.9	9.3
4.1	9.1
7.3	4.3
5.2	7.2
6.0	
nG = 8	nR = 7

Wilcoxon Rank Sum test

- Null assumption:
  When considering samples from both factories then all rank configurations are equiprobable.

German factory	Stochastic Heights factory
3.2	3.7
4.5	8.5
8.1	6.1
9.9	9.3
4.1	9.1
7.3	4.3
5.2	7.2
6.0	
nG = 8	nR = 7



9.9	G
9.3	R
9.1	R
8.5	R
8.1	G
7.3	G
7.2	R
6.1	R
6.0	G
5.2	G
4.5	G
4.3	R
4.1	G
3.7	R
3.2	G

N = nG + nR = 15

Wilcoxon Rank Sum test

- Null model, abstracted:
  All N choose B binary configurations in {0,1}^(N,B)
  are equiprobable.
- Let T = sum of the ranks of the entries labeled 1.

- {0,1}^(N,B): All binary vectors with N elements out of which B are 1s

9.9	G	1	0
9.3	R	2	1
9.1	R	3	1
8.5	R	4	1
8.1	G	5	0
7.3	G	6	0
7.2	R	7	1
6.1	R	8	1
6.0	G	9	0
5.2	G	10	0
4.5	G	11	0
4.3	R	12	1
4.1	G	13	0
3.7	R	14	1
3.2	G	15	0

$N = nG + nR = 15$ $N = 15, B = 7, T = 50$

Wilcoxon Rank Sum test

- Under the null model we have

$$E(T) = 56$$

- The result here points to better (lower numbers) ranks for R.
- But is it <u>significant</u>?

1	0
2	1
3	1
4	1
5	0
6	0
7	1
8	1
9	0
10	0
11	0
12	1
13	0
14	1
15	0

$$N = 15, B = 7, T = 50$$

Wilcoxon Rank Sum test

- We want to compute

$$P(T \leq 50)$$

- under the null model.

1	0
2	1
3	1
4	1
5	0
6	0
7	1
8	1
9	0
10	0
11	0
12	1
13	0
14	1
15	0

$$N = 15, B = 7, T = 50$$

Wilcoxon Rank Sum test

- Normal approximation (Wilcoxon 1947) for calculating $P_{Null}(T \leq 50)$?
- When B and N-B are sufficiently large and depending on the desired accuracy
- Let

$$\mu_T = \frac{B(N+1)}{2}$$

- and

$$\sigma_T = \sqrt{\frac{B(N-B)(N+1)}{12}}$$

- then

$$Z(T) = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$$

Wilcoxon Rank Sum test

- Back to Randomistan VW factory: N = 15, B = 7, T = 50
- Using the normal approximation even though the numbers are not sufficiently large:

$$\mu_T = \frac{7(15+1)}{2} = 56 \quad \sigma_T = \sqrt{\frac{7(15-7)(15+1)}{12}} \approx 8.64$$

$$Z \approx \frac{50-56}{8.64} = -0.7$$

and therefore

$$P_{Null}(T \leq 50) \approx 0.24$$

and we **do not have sufficient confidence for rejecting the hypothesis** that the German factory is as good as the one in Stochastic Heights.

WRS question

- Given a binary vector $v \in \{0,1\}^{(N,B)}$.
- Recall that this notation means:

$$v = (v_1, \dots v_N), \qquad v_i \in \{0,1\}, \qquad \sum_{r=1}^{N} v_r = B$$

- Let $S(v)$ be the sum of squares of the ranks of the entries equal to 1:

$$S(v) = \sum_{r=1}^{N} r^2 v_r$$

- Assume that V is uniformly drawn from $\{0,1\}^{(N,B)}$
- Consider the random variable $S = S(V)$

WRS question

- What is the minimal value $S_{\min}$ that S can attain?
- What is $P(S = S_{min})$?
- *This is attained where all the 1s are in the first B ranks:*

$$S_{min} = \sum_{r=1}^{B} r^2 \, , P(S = S_{min}) = \frac{1}{\binom{N}{B}}$$

- What is the maximal value $S_{\max}$ that S can get?
- What is $P(S = S_{max})$?
- *This is attained where all the 1s are in the last B ranks:*

$$S_{max} = \sum_{r=N-B+1}^{N} r^2 \, , P(S = S_{max}) = \frac{1}{\binom{N}{B}}$$

- Let $N = 15, B = 5$. What is $P(S = 66)$?
- *This is attained only when ranks 1,2,3,4 and 6 are the five 1s:*

$$66 = 1 + 4 + 9 + 16 + 36 \rightarrow P(S = 66) = \frac{1}{\binom{N}{B}}$$

WRS question

- Derive the formula for $E[S]$ for general N, B.

- Under the null hypothesis that V is uniformly drawn from $\{0,1\}^{(N,B)}$:

$$S = \sum_{i=1}^{B} R_i^2$$

- where $R_i \sim U(1, N), i = 1, \ldots, B$ are discrete random variables representing the ranks of the B occurrences of 1. Note that the variables $R_1, \ldots, R_B$ are not independent. However, by linearity of expectations:

$$E[S] = E\left[\sum_{i=1}^{B} R_i^2\right] = \sum_{i=1}^{B} E\left[R_i^2\right]$$

- Let $R_i = R \sim U(1, N)$:

$$E[R] = \frac{N+1}{2}, Var(R) = \frac{N^2-1}{12}$$

$$E[R^2] = Var(R) + E[R]^2 = \frac{N^2-1}{12} + \left(\frac{N+1}{2}\right)^2 = \frac{N^2-1+3(N^2+2N+1)}{12} = \frac{4N^2+6N+2}{12} = \frac{2N^2+3N+1}{6}$$
$$= \frac{(N+1)(2N+1)}{6}$$

- Finally:

$$E[S] = B \cdot E[R^2] = \frac{B(N+1)(2N+1)}{6}$$

Hypergeometric

- Suppose we are interested in the number of defectives in a sample of size n units drawn from a lot containing N units, of which α are defective
- Each object has same chance of being selected, then the probability that the first drawing will yield a defective unit = $\frac{\alpha}{N}$
- But for the second drawing

$$P = \begin{cases} \frac{\alpha - 1}{N - 1} & if\ first\ unit\ is\ defective \\ \frac{\alpha}{N - 1} & if\ first\ unit\ is\ not\ defective \end{cases}$$

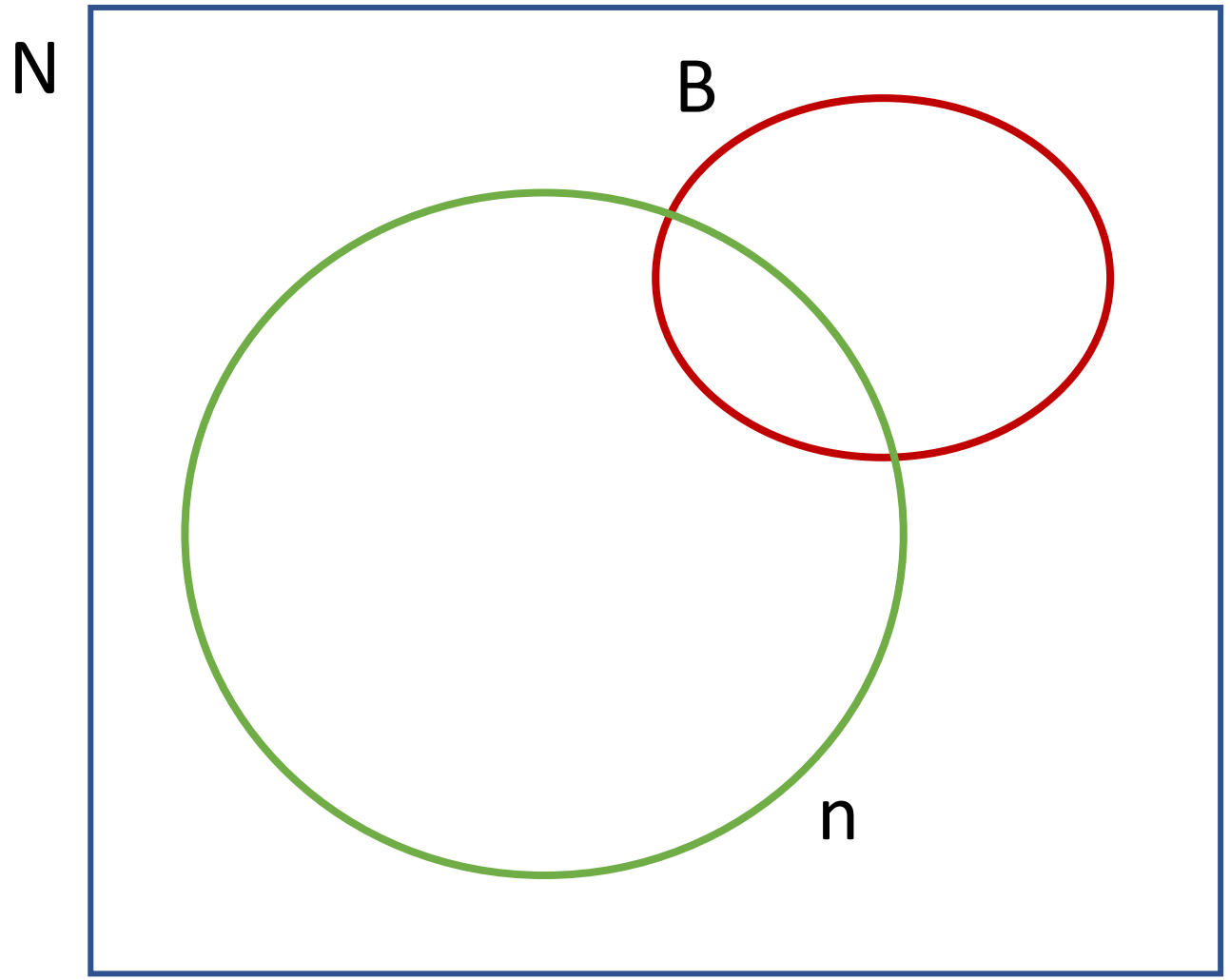
Hypergeometric

- The trials here are not independent and hence the assumption underlying the **binomial** distribution is not fulfilled and therefore, we cannot apply binomial distribution here
- Binomial distribution would have been applied if we do sampling with replacement. If each unit selected from the sample would have been replaced before the next one is drawn

Hypergeometric

$$HG(N, B, n, b) = \frac{\binom{N-B}{n-b}\binom{B}{b}}{\binom{N}{n}}$$

- The parameters of hypergeometric distribution are the sample size n, the "population" size N, and the number of "successes" in the population B
- What if n is small compared to N? The composition of the population is not seriously affected by drawing the sample and the binomial distribution with parameters n and $P = \frac{\alpha}{N}$ will yield a good approximation



Hypergeometric – Mean

$$\mu = \frac{B}{N} n$$

- Proof:

$$\mu = \sum_{b=0}^{n} b \cdot HG(N, B, n, b) = \sum_{b=1}^{n} b \cdot \frac{\binom{N-B}{n-b}\binom{B}{b}}{\binom{N}{n}}$$

$$\binom{B}{b} = \frac{B!}{b!\,(B-b)!} = \frac{B}{b}\frac{(B-1)!}{(b-1)!\,(B-b)!} = \frac{B}{b}\binom{B-1}{b-1}$$

$$\mu = \sum_{b=1}^{n} B \cdot \frac{\binom{N-B}{n-b}\binom{B-1}{b-1}}{\binom{N}{n}} = \frac{B}{\binom{N}{n}} \sum_{b=1}^{n} \binom{N-B}{n-b}\binom{B-1}{b-1}$$

Hypergeometric – Mean

- Let $b-1=y$

$$\mu = \frac{B}{\binom{N}{n}} \sum_{y=0}^{n-1} \binom{N-B}{n-1-y} \binom{B-1}{y}$$

- Using the identity $\sum_{r=0}^{k} \binom{s}{k-r}\binom{m}{r} = \binom{m+s}{k}$,

  where $k=n-1, m=B-1, r=y, s=N-B$, we get:

$$\mu = \frac{B}{\binom{N}{n}} \binom{N-1}{n-1} = B\,\frac{n!\,(N-n)!}{N!}\,\frac{(N-1)!}{(n-1)!\,(N-n)!} = \frac{B}{N}n$$

- Variance proof in google…

HG question 1

- A company (the producer) supplies microprocessors to a manufacturer (the consumer) of electronic equipment.
  The microprocessors are supplied in batches of 50.
  The consumer regards a batch as acceptable if there are not more than 5 defective microprocessors in the batch.
  Rather than test all the microprocessors in the batch, 10 are selected at random and tested.

- Find the probability that out of a sample of 10, d = 0, 1, 2, 3, 4, 5.

$$\frac{\binom{45}{10-k}\binom{5}{k}}{\binom{50}{10}}$$

HG question 1

- Suppose that the consumer will accept the batch provided that not more than m defectives are found in the sample of 10.
  - Find the probability that the batch is accepted when there are 5 defectives in the batch.

$$\sum_{k=0}^{m} P(X=k) = \sum_{k=0}^{m} \frac{\binom{45}{10-k}\binom{5}{k}}{\binom{50}{10}}$$

Where $m \leq 5$

  - Find the probability that the batch is rejected when there are 3 defectives in the batch.

$$1 - \sum_{k=0}^{m} P(X=k) = \sum_{k=0}^{m} \frac{\binom{47}{10-k}\binom{3}{k}}{\binom{50}{10}}$$

Where $m \leq 5$

HG question 2

- An advertising company has 12 women and 8 men. Suppose the company needs to select a team of 5 members to work on a commercial for a new hybrid car.
- If the members of the team are selected at random, what is the probability that 3 women and 2 men will be selected?

$$\frac{\binom{12}{3}\binom{8}{2}}{\binom{20}{5}}$$

- What is the probability that women will constitute a majority in the team?

$$\frac{\binom{12}{3}\binom{8}{2}}{\binom{20}{5}} + \frac{\binom{12}{4}\binom{8}{1}}{\binom{20}{5}} + \frac{\binom{12}{5}\binom{8}{0}}{\binom{20}{5}}$$