BIM 105
Probability and Statistics for Biomedical Engineers

David M. Rocke

Department of Biomedical Engineering

$X_1, X_2, ..., X_n$ are random variables

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$$

$$\mu_Y = c_1 \mu_{X_1} + c_2 \mu_{X_2} + \cdots + c_n \mu_{X_n}$$

if the $\{X_i\}$ are all statistically independent, then

$$\sigma_Y^2 = c_1^2 \sigma_{X_1}^2 + c_2^2 \sigma_{X_2}^2 + \cdots + c_n^2 \sigma_{X_n}^2$$

$$\mu_{X_i} = \mu$$

$$\sigma_{X_i}^2 = \sigma^2$$

$$\bar{X} = (X_1 + X_2 + \cdots + X_n)/n = n^{-1} X_1 + n^{-1} X_2 + \cdots + n^{-1} X_n$$

$$E(\bar{X}) = n^{-1}\mu + n^{-1}\mu + \cdots + n^{-1}\mu = n(n^{-1}\mu) = \mu$$

$$V(\bar{X}) = n^{-2}\sigma^2 + n^{-2}\sigma^2 + \cdots + n^{-2}\sigma^2 = n(n^{-2}\sigma^2) = \sigma^2 / n$$

Types of Data and Distributions

- Discrete data are usually counts, and these can be of basically two types.
- One type is when we have a set of objects, and for each one the object is of one of two (or several) types.
  - Primary fibroblasts are to be re-programmed as stem cells (iPSC's). If we start with 100 cells, some will be successfully reprogrammed and some not. We can count the number out of 100 that are successfully re-programmed. This number is an integer between 0 and 100.
  - A medical device is tested after manufacture. In a sample of 25 devices from a lot of 5,000, we can count the number of defective ones. This number is an integer between 0 and 25.
  - A part for the medical device needs to fit in close tolerances. We can classify each part from a sample of 50 as good, too small, or too large. We obtain three numbers, each between 0 and 50, that add up to 50.

- Another type of count data is where there is no fixed upper limit.
  - The number of defects in a silicon wafer is a positive integer with no fixed upper limit.
  - If there are 800,000 spots on an Affymetrix microarray, and on the average 0.01% of them are defective, then the number of defective spots can be treated as a positive integer with no fixed upper limit. The upper limit is 800,000, but the average value is around 80 and is unlikely to be more than a few hundred.

Bernoulli Trials and the Binomial Distribution

- A series of Bernoulli trials is a series of statistically independent experiments where on each trial the probability of one of the outcomes (conventionally called "success") is the same probability p.
- If we have a series of Bernoulli trials of length n, and if the chance of success on each trial is p, then the total number of successes is a *Binomial* random variable.
- If X has a Binomial distribution with parameters n and p, then X is an integer between 0 and n.

Three Machines

- Machines 1, 2, and 3 are either up (operational) or down (not working).
- Each has independently a 10% chance of being down.
- There are eight elements in the sample space {UUU, UUD, UDU, UDD, DUU, DUD, DDU, DDD}
- The number of machines that are down is a binomial random variable with parameters n = 3 and p = 0.10.
- Here "success" is the machine is down.

$X \sim \text{Bin}(n, p) = \text{Bin}(3, 0.10)$

There are 8 possible outcomes.

One of the eight has $X = 0$, corresponding to none down.

$P(X = 0) = (0.9)^3 = 0.729$

One of the eight has $X = 3$, corresponding to all down.

$P(X = 3) = (0.1)^3 = 0.001$

Three of the eight have $X = 1$, corresponding to one down.

$P(X = 1) = 3(0.1)(0.9)^2 = 0.243$

Three of the eight have $X = 2$, corresponding to two down.

$P(X = 2) = 3(0.1)^2(0.9) = 0.027$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

The symbol $\binom{n}{x}$ is the number of ways of choosing x positions in the sequence

out of the n positions in the sequence for the successes. For each one,

the chance of that sequence is the product with

x p's and $n - x$ $(1\text{-}p)$'s

Bin(5, .05)

x	P(X = x)	P(X ≤ x)
0	0.7737809	0.7737809
1	0.2036266	0.9774075
2	0.0214343	0.9988419
3	0.0011281	0.9999700
4	0.0000296	0.9999997
5	0.0000003	1.0000000

- Use the formula (for probability mass function)
- Use table A1 in Appendix A.
- Use a computer command.
- In Excel, `=Biomdist(1,5,0.05)`, `=Biomdist(1,5,0.05,True)`

The Binomial in MATLAB

```
>> binopdf(1,5,.05)

    0.2036

>> binocdf(1,5,.05)

    0.9774

>> binopdf(0:5,5,.05)

    0.7738    0.2036    0.0214    0.0011    0.0000    0.0000

>> binocdf(0:5,5,.05)

    0.7738    0.9774    0.9988    1.0000    1.0000    1.0000
```

The mean and variance of a binomial.

$$X = \text{Bin}(n, p) = Y_1 + Y_2 + \cdots + Y_n$$

where $Y_i \sim \text{Bin}(1, p)$

$$\mu_Y = (1)(p) + (0)(1-p) = p$$

$$E(Y^2) = (1^2)(p) + (0^2)(1-p) = p$$

$$\sigma_Y^2 = E(Y^2) - \mu_Y^2 = p - p^2 = p(1-p)$$

$$\mu_X = p + p + \cdots p = np$$

$$\sigma_X^2 = p(1-p) + p(1-p) + \cdots p(1-p) = np(1-p)$$

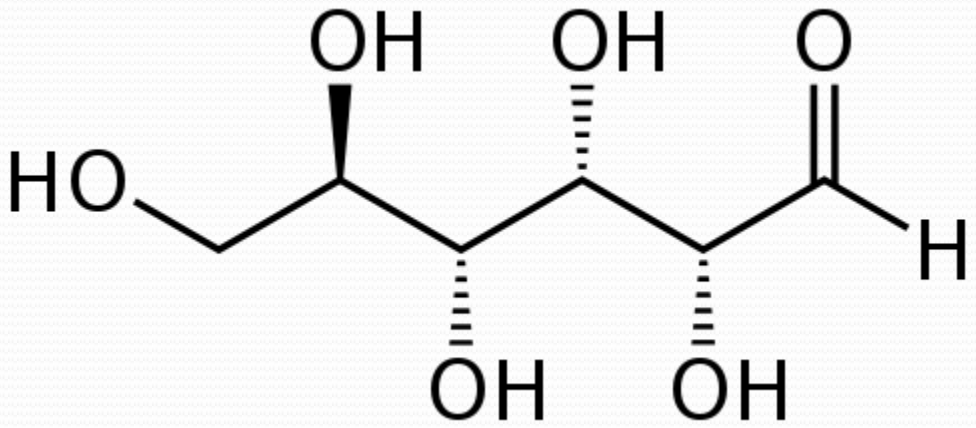
Estimating the binomial probability.

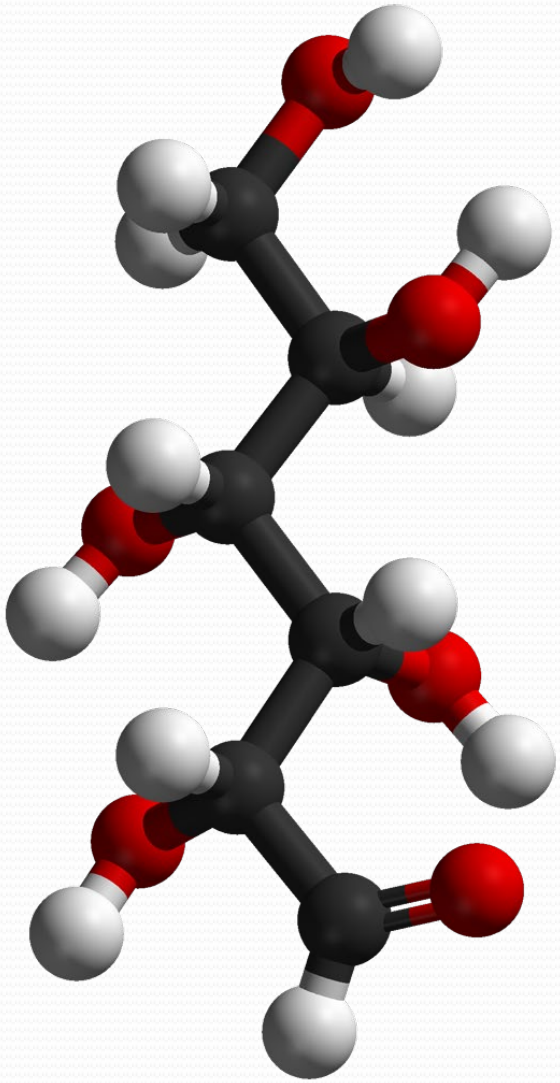
$$\mu_X = np$$
$$\sigma_X^2 = np(1-p)$$

$$\hat{p} = X / n$$
$$\mu_{\hat{p}} = p$$
$$\sigma_{\hat{p}}^2 = np(1-p)/n^2 = p(1-p)/n$$
$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

$\hat{p}$ estimates p with uncertainty $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$

Glucose





Isotopes and Molecular Weight

- Glucose is a sugar with molecular formula $C_6H_{12}O_6$.
- In round numbers, its molecular weight is 180 Daltons if all the atoms are the most common isotopes of the elements.
- Carbon-13 (^{13}C) instead of the more common carbon-12 (^{12}C) occurs at a rate of 1.1% and weighs 1 Dalton more. What is the probability distribution of the molecular weights of glucose?
- Using Bin(6, 0.011),
  - The chance of zero ^{13}C is 0.9358 (180 Daltons)
  - The chance of one ^{13}C is 0.0624 (181 Daltons), and
  - the chance of two ^{13}C is 0.0017 (182 Daltons).
  - There is almost no chance of three or more (about 1 in 40,000).

- Oxygen-17 (^{17}O) occurs with a probability of 0.00038, so the chance of any ^{17}O occurring in a molecule of glucose is 0.0023
- ^{2}H occurs with a probability of 0.00015, so the chance of any ^{2}H occurring is 0.0018.
- These calculations are important in mass spectrometry to find sub-peaks of a main peak of a compound to be identified.
- A small protein might have 100 amino acids and about 400 carbon atoms. The most frequent number of ^{13}C atoms is 4, but numbers from 0 to 9 all can occur. 2, 3, 4, 5, 6 all occur at > 10%. Thus, the main peaks are 2 to 6 Daltons above the nominal molecular weight.

Acceptance Sampling

- A supplier of parts for a medical device has certified that no more than 1% of the parts will fail to meet the specifications.
- We want to test a random sample of each shipment of size n and reject the shipment if there are more than x defectives out of n.
- If the fraction of defectives is actually 1%, then we would like to reject the shipment no more than 5% of the time.
- If the fraction of defectives is actually 2%, then we would like to reject the shipment at least 95% of the time.
- What should we choose as n and x?

Choose n and x so that when $p = 0.01$ then

$P(X \le x) > 0.95$

and so that when $p = 0.02$ then

$P(X \le x) \le 0.05$

and make n as small as possible.

For example, if we choose $n = 500$, then we can pick $x = 9$ and

$P(X \le 9 \mid p = 0.01) = 0.9689$

$P(X \le 9 \mid p = 0.02) = 0.4567$

so 500 is not large enough

If we choose $n = 2000$, then we can pick $x = 28$ and

$P(X \le 28 \mid p = 0.01) = 0.9664$

$P(X \le 28 \mid p = 0.02) = 0.0282$

so 2000 is large enough, possibly larger than needed.

```
>> binocdf(0:20,500,.01)

  Columns 1 through 13

    0.0066    0.0398    0.1234    0.2636    0.4396    0.6160    0.7629    0.8677    0.9329
    0.9689    0.9868    0.9948    0.9981

  Columns 14 through 21

    0.9994    0.9998    0.9999    1.0000    1.0000    1.0000    1.0000    1.0000

>> binoinv(0.95,500,.01)

     9

>> binocdf(0:20,500,.02)

  Columns 1 through 13

    0.0000    0.0005    0.0026    0.0098    0.0281    0.0652    0.1276    0.2175    0.3305
    0.4567    0.5830    0.6979    0.7935

  Columns 14 through 21

    0.8667    0.9186    0.9530    0.9743    0.9866    0.9934    0.9969    0.9986

>> binocdf(9,500,.02)

    0.4567
```

```
>> binoinv(0.95,1000,.01)
    15
>> binocdf(15,1000,.01)
    0.9521
>> binocdf(15,1000,.02)
    0.1539
>> binoinv(0.95,2000,.01)
    28
>> binocdf(28,2000,.01)
    0.9664
>> binocdf(28,2000,.02)
    0.0282
```

```
>> binoinv(0.95,1625,.01)
    23
>> binocdf(23,1625,.01)
    0.9584
>> binocdf(23,1625,.02)
    0.0498
>> binoinv(0.95,1624,.01)
    23
>> binocdf(23,1624,.01)
    0.9587
>> binocdf(23,1624,.02)
    0.0502
```

The Multinomial Distribution

- Suppose we have more than two outcomes. Say a medical device voltage is a) in spec; b) too high; or c) too low. If we test n devices, then we have three random variables:
  - X_a = number in spec
  - X_b = number with voltage too high
  - X_c = number with voltage too low.
- If the device tests are statistically independent with constant probabilities of the three outcomes, then the three numbers have a multinomial distribution with parameters n and (p_a, p_b, p_c).

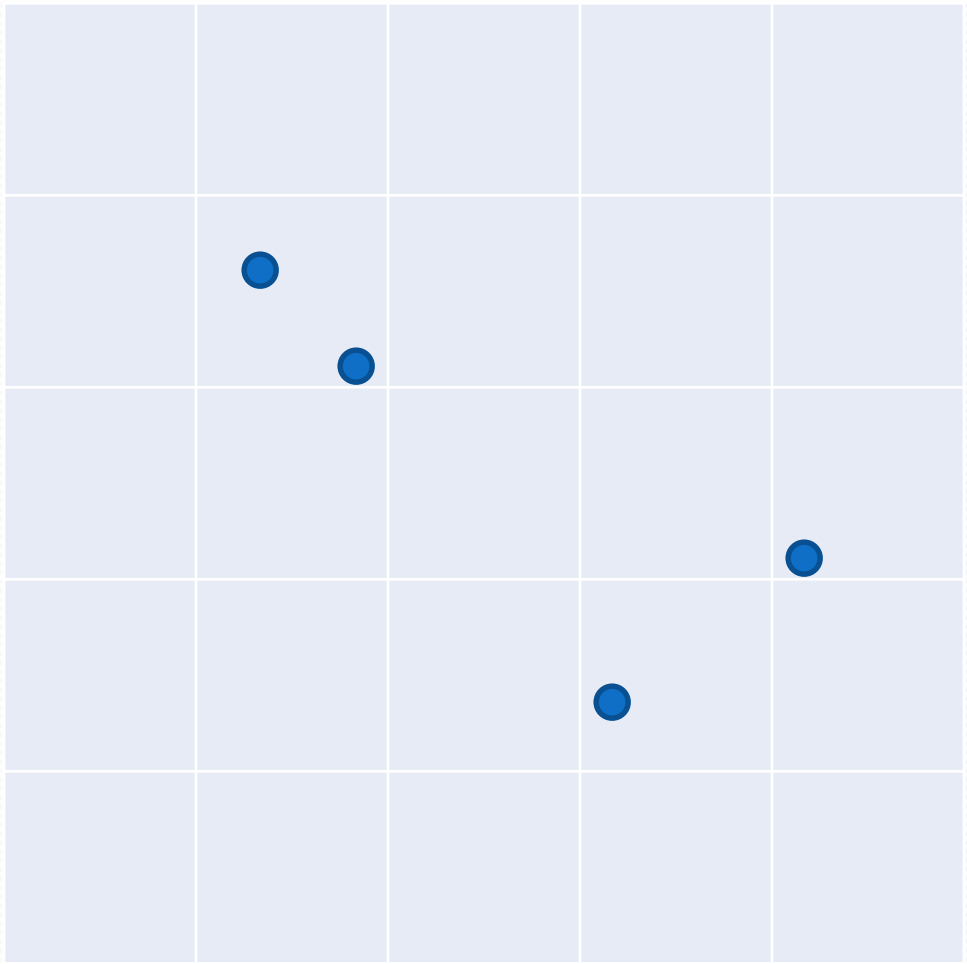
- The multinomial is defined by the number of trials, and the vector of probabilities of each outcome.
- For example, if we test 100 devices, and the probabilities of in spec, too high, and too low, are (0.90, 0.08, and 0.02), and if the tests are presumed to be statistically independent, then this has a multinomial distribution.
- Each of the three counts is binomial [Bin(100, 0.90), Bin(100, 0.08), and Bin(100, 0.02)], so the means and variances are the same as the binomial.
- But the counts are correlated (the more devices there are that are in spec, the less likely it is that the count of too high is large).

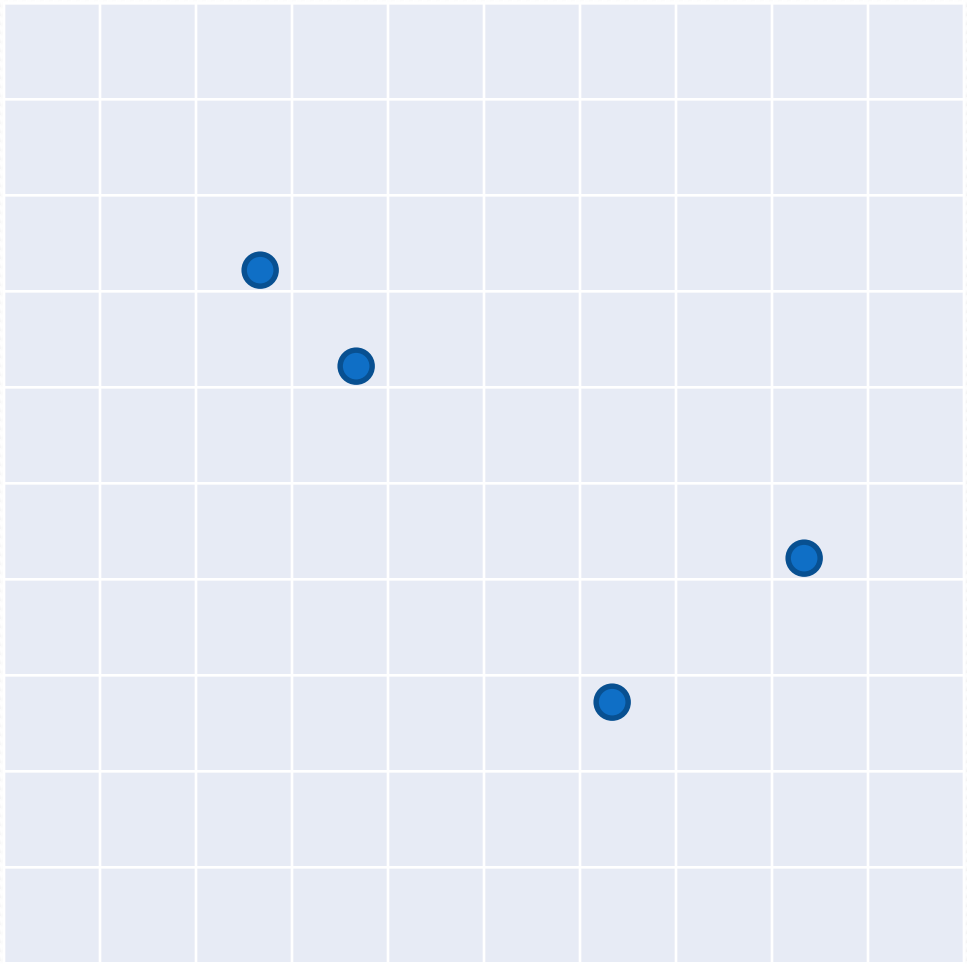
The Poisson Distribution

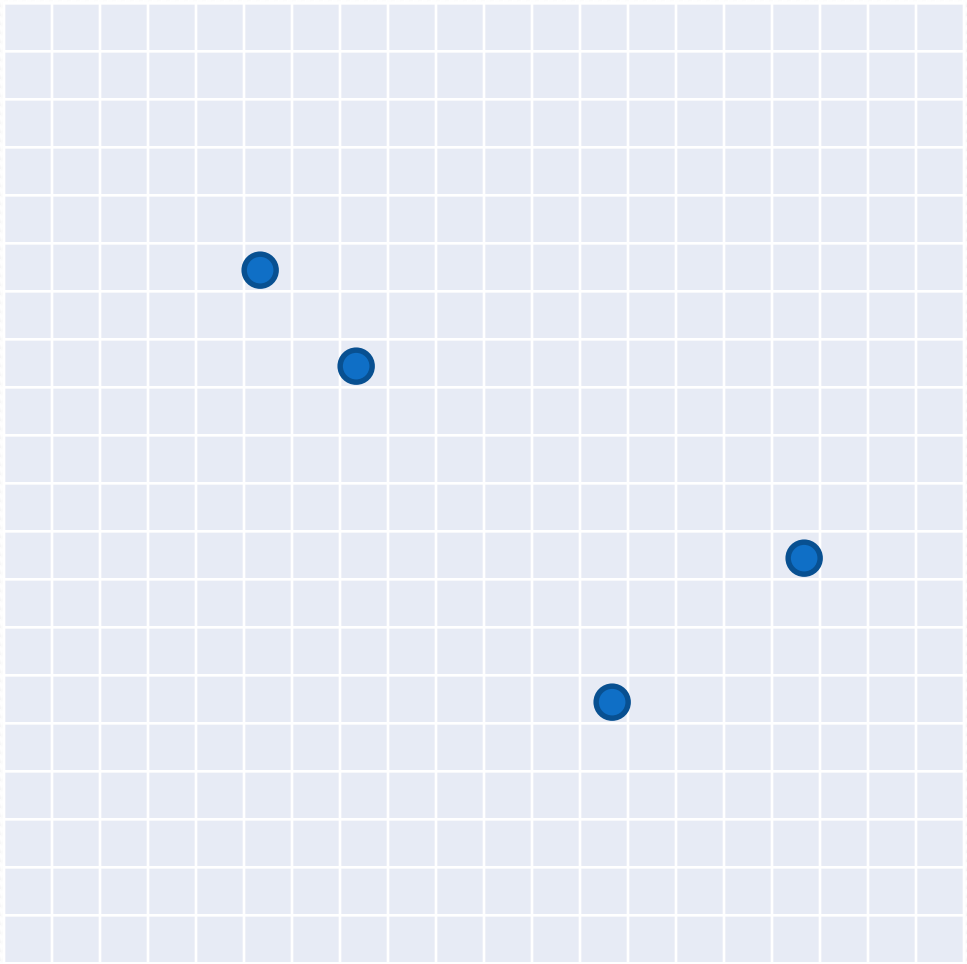
- The Poisson arises in contexts in which the thing being counted has no fixed upper limit.
- Siméon-Denis Poisson (1781–1840) was a French mathematician. (*pwa-so*^{n}) definitely not poison!
- These are usually point events in time or space.
  - Count the repairs made in a year on a five-year old drill press.
  - Count the defects in an optical fiber cable of length 1000m.
  - Count the defects in a silicon wafer of surface area of 180cm^2.

The Poisson and the Binomial

- Suppose we have a silicon wafer in which we count defects.
- Suppose that the average number of defects per wafer is 1.
- What is the chance that there is no defect?







Divide the wafer into 100 sections and compute the binomial distribution.

If the average number of defects per wafer is 1, and there are 100 sections then the binomial would have $p = 1/100 = 0.01$. The chance of no defects under the binomial is

$(0.99)^{100} = 0.3660$

The only difference really is that under the Poisson, there could be more than one defect in a given section, which should be a rare event if the chance that there is one defect is 1%.

If we divide the wafer into 1000 sections, then $p = 1/1000 = 0.001$. The chance of no defects under the binomial is

$(0.999)^{1000} = 0.3677$

$(0.9999)^{10000} = 0.3679$

$e^{-1} = 0.3679$

Average number of defects per plate is λ.

Divide the plate in n equal sections, and then think of n getting larger.

The number of defects per section is approximately Bernoulli so long as no section holds 2 or more defects, which will be unlikely if n is much larger than λ. For the whole disk,

$$P(X=0)=\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$$

Because

$$n\ln(1-\lambda/n)=n\left[-\lambda/n+\lambda^2/2n^2-\cdots\right]$$
$$\approx n\left[-\lambda/n\right]=-\lambda$$

$$P(X=x)\approx\binom{n}{x}\left(1-\frac{\lambda}{n}\right)^{n-x}(\lambda/n)^x$$
$$=\frac{n(n-1)-\cdots(n-x+1)}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}(\lambda/n)^x$$
$$\to\frac{n^x}{x!}e^{-\lambda}(1)\frac{\lambda^x}{n^x}$$
$$=e^{-\lambda}\frac{\lambda^x}{x!}$$

For the silicon wafer, now suppose that $\lambda = 0.5$ so

$$P(X=0) = e^{-0.5} = 0.606531$$

$$P(X=1) = e^{-\lambda}\frac{\lambda^1}{1!} = (0.5)e^{-0.5} = 0.303265$$

$$P(X=2) = e^{-\lambda}\frac{\lambda^2}{2!} = \frac{0.5^2}{2}e^{-0.5} = \frac{1}{8}e^{-0.5} = 0.075816$$

$$P(X=3) = e^{-\lambda}\frac{\lambda^3}{3!} = \frac{0.5^3}{6}e^{-0.5} = \frac{1}{48}e^{-0.5} = 0.012636$$

If we divide the wafer into 100 section, then the number of defects in each section is Poisson with $\lambda = 0.005$

$$P(X=0) = e^{-0.005} = 0.995012$$

$$P(X=1) = e^{-\lambda}\frac{\lambda^1}{1!} = (0.005)e^{-0.005} = 0.004975$$

$$P(X \le 1) = 0.995012 + 0.004975 = 0.999987$$

For 1000 sections, $\lambda = 0.0005$ and

$$P(X \le 1) = 0.999500125 + 0.000499750 = 0.999999875$$

The Mean and Variance of a Poisson

A Poisson random variable X with parameter λ is approximately binomial with parameters n and $p = \lambda / n$ for large n

$\mu_X \approx np = n(\lambda / n) = \lambda$

$\sigma_X^2 \approx np(1-p) = n(\lambda / n)(1-\lambda / n) \approx \lambda$

For a Poisson random variable, the mean and the variance are both λ

X estimates λ with uncertainty $\sqrt{\lambda} \approx \sqrt{X}$

Rates and Means

- Suppose that a flask has a suspension of cells with an average density of 12 cells per cc.
- A sample of 3 cc is taken.
- The number of cells in the sample could be modeled as a Poisson random variable with mean $\lambda = 36$.
- The variance is 36 and the standard deviation is 6.
- The *rate* is 12 cells/cc.
- The *mean* of the Poisson random variable is (3)(12) = 36.

Polymerase Chain Reaction (PCR)

- The most definitive test for viral diseases (HIV, Hepatitis A, B, C, Zika, Ebola) involves searching for the actual virus in the blood. A faster, but less accurate test is based on antibodies, which can measure past infections that are not necessarily current ones.
- PCR involves a process of multiple cycles. On each cycle the number of copies of the viral DNA (or RNA which has been reverse-transcribed to DNA) is approximately doubled. This can be done for up to 40 cycles ($2^{40} = 10^{12}$ or 1 trillion)

- If there are approximately 25 copies of the virus in the starting sample, then the amplification will yield a detectable signal.
- Suppose we use 10 μL of extracted DNA for a test, and this comes from a larger sample from a patient. If the average concentration of copies of the viral DNA is 25/10 μL, then a sample will often have too few copies.
- How high does the copy number per μL need to be so that the sample has almost always 25 copies or more?
- Specifically, so that the chance is at least 95%.
- If X is a Poisson random variable with parameter λ, how big does λ need to be so that $P(X \leq 24) < 0.05$?

```
>> poisscdf(24,25)     (24=x, 25=λ)
    0.4734
>> poisscdf(24,40)
    0.0045
>> poisscdf(24,30)
    0.1572
>> poisscdf(24,35)
    0.0324
>> poisscdf(24,34)
    0.0460
>> poisscdf(24,33)
    0.0642
```

The copy number in the source needs to be at least 34 per 10 µL to insure 25 copies per 10 µL sample.

poisspdf() gives individual values of the probability mass function

The concentration of particles in a suspension is 4 per mL. The suspension is thoroughly agitated, and then 2 mL is withdrawn. Let X represent the number of particles that are withdrawn. Find

a. $P(X = 6)$

b. $P(X \leq 3)$

c. $P(X > 2)$

d. μ_X

e. σ_X

The concentration of particles in a suspension is 4 per mL.

The suspension is thoroughly agitated, and then 2 mL is withdrawn.

Let X represent the number of particles that are withdrawn. Find

a. P(X = 6)

$\lambda = (4/\text{mL})(2\text{mL}) = 8$

$$P(X=6) = e^{-8}\frac{8^6}{6!} = (3.3546\times10^{-4})\frac{262144}{720} = (3.3546\times10^{-4})(364.09) = 0.12214$$

b. P(X $\leq$ 3)

c. P(X > 2)

d. $\mu_X = 8$

e. $\sigma_X = \sqrt{8} = 2.828$

0	1	2	3	4	5	6	7
0.0003	0.0027	0.0107	0.0286	0.0573	0.0916	0.1221	0.1396

$$P(X \le 3) = 0.0003 + 0.0027 + 0.0107 + 0.0286 = 0.0424$$

$$P(X > 2) = 1 - P(X \le 2) = 1 - 0.0003 + 0.0027 + 0.0107 = 1 - 0.0138 = 0.9862$$

$$P(X \ge 2) = 1 - P(X \le 1) = 1 - 0.0003 + 0.0027 = 1 - 0.0030 = 0.9970$$

```
>> poisscdf(0:6,8)

ans =

    0.0003    0.0030    0.0138    0.0424    0.0996    0.1912    0.3134
```

Other Discrete Probability Distributions

- If X_1, X_2, … are independent Bernoulli random variables (1/0 with probability $p/(1-p)$), then the sum of n of them, for fixed n, is a binomial random variable
- If instead, we continue observing the sequence until the sum reaches x, the length of the sequence is a negative binomial random variable.
- This has mean $px/(1-p)$ and variance $px/(1-p)^2$
- When we divided up the silicon disk into many equal parts, we assumed that the chance of a defect per unit area was the same. If instead, this varied (maybe higher on the edges?), we can still have the same mean λ but the variance is then larger than λ.
- Perhaps remarkably, this can often be modeled as a negative binomial (which has a variance larger than its mean). This has applications in next generation sequencing, especially RNA-Seq.