4. Common Discrete Distributions

Readings: Rosner: 4.7 – 4.14
OpenIntro Statistics: 3.4, 3.5.2
R: dbinom (pmf), pbinom (cdf), qbinom (quantile), rbinom (PRNG), sum,
dpois (pmf), ppois (cdf), qpois (quantile), rpois (PRNG)
SAS functions: PROBBNML (cdf), RANBIN (PRNG), POISSON (cdf),
RANPOI (PRNG)

Homework: Homework 2 due by noon on September 17

Overview

A) Bernoulli distribution
B) Binomial distribution
C) Poisson distribution
D) Poisson approximation to the binomial distribution

Examples:

A. Suppose 10% of a specific population has Chronic Depression. A different population is of interest, and a random sample of 20 from this population is drawn. How many cases of Chronic Depression do we expect to see if the two populations are similar? Suppose we see 6 cases, what is the probability of 6 cases if a 10% prevalence holds? Of 6 or more cases if a 10% prevalence holds? Is this probability low enough to suggest that population of interest differs from the original population?

B. In the U.S., the probability of being involved in a motor vehicle accident in a given year is about 0.00024, or 24 per 100,000. In a sample of 10,000 people, what is the probability that nobody will be involved in an accident in the next year assuming the US rate holds? What's the probability that 7 or more will be involved in an accident?

C. The distribution of admissions for asthma to a certain ER has been found to have a mean of 0.8 admissions per day. What is the probability of having more than 3 admissions for asthma in a day? The probability of admitting none?

A) Bernoulli distribution – binary or dichotomous outcomes

A Bernoulli random variable has 2 possible outcomes (heads/tails, success/failure, live/die, hepatitis yes/no, accident yes/no…)

e.g. X=1 if adult current smoker
X=0 if not current smoker (mutually exclusive and exhaustive)

In 1987, 29% of US adults were current smokers (of cigars, pipes, cigarettes, etc.)
P(X=1) = p = _____
P(X=0) = 1-p = _____ *(note that* q=1-p *is also commonly used)*

Recall for discrete distributions (Lecture 3):
$E(X) = \sum_x x\, P(X=x) = (0)(1-p) + (1)(p) = p = \mu$
$V(X) = \sum_x (x-\mu)^2\, P(X=x) = (0-p)^2(1-p) + (1-p)^2 p = p^2 - p^3 + p - 2p^2 + p^3 = p - p^2 = p(1-p) = \sigma^2$

e.g. $\mu = p = 0.29$, $1-p = 0.71$, $\sigma^2 = p(1-p) = 0.29 \times 0.71 = 0.2059$

As you can see, the Bernoulli distribution is completely defined by the single parameter, *p*, i.e. once you know the value of *p*, you know all you need to know about the specific Bernoulli distribution.

B) Binomial – a collection of binary or dichotomous outcomes

A binomial random variable is a sum of *n* independent Bernoulli random variables (trials) with the same P(X=1) = p. This distribution helps us to make inferences about *proportions*. We can think of proportions as averages for Bernoulli r.v.

X: the random variable (r.v.) that counts the number of units in a sample of size *n* with a characteristic of interest
p: probability of observing the characteristic in an individual unit
q: 1-p, the probability of not observing the characteristic in an individual unit

Assumptions: A sample of *n* units is drawn from an infinite population *without replacement*; each unit in the population has the same probability *p* of having the characteristic of interest:

1) Fixed number *n* of trials, two possible outcomes on each trial
2) Trials are independent
3) *p* is the same for all trials

Sample space: {0, 1, 2, …, n }
Parameters: *n* = number of trials, *p* = P(success) = P(X=1)

e.g. Randomly select 3 adults from US
X = r.v. that counts the number who are smokers
With n = 3 the sample space is {0, 1, 2, 3} – mutually exclusive, exhaustive events
$P(X=0) = p^0(1-p)^3$ occurs in only 1 way
$P(X=1) = 3p(1-p)^2$ occurs in 3 ways
$P(X=2) = 3p^2(1-p)$ occurs in 3 ways
$P(X=3) = p^3(1-p)^0$ occurs in only 1 way
Do you see the assumption of *independence* at work here?

Constructing a tree to identify the probability of each event:

X ~ Bin (n, p): means X is distributed as a binomial r.v. with parameters *n* and *p*

Calculating Binomial Probabilities: P (X=x) for x=0, …, *n*
For very small n, we can construct a tree (like on the previous slide) but there are 2^n branches at the last step, which quickly becomes unwieldy as *n* increases.

To obtain the probabilities we need to work out:

1) First *x* trials give "success" (outcome of interest)
2) Number of ways of getting *x* successes in *n* trials

The result is a binomial probability mass function (pmf), one of the well-used theoretical discrete probability distributions:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0,1,\ldots,n \text{ and } 0 \le p \le 1$$

$\binom{n}{x}$ is called the *binomial coefficient* and may also be denoted ${}_nC_x$

$\binom{n}{x}$ **=** number of ways of choosing *x* objects from a total of *n* without regard to order

$\binom{n}{x} = \frac{n!}{(n-x)!x!}$, where $n! = n \times (n-1) \times \ldots \times 2 \times 1$ and $0! = 1$

e.g. Randomly select 3 adults from US where X = number who are smokers

Recall from previous slide: $P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}=\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}$

P(X=0) =

P(X=1) =

P(X=2) =

P(X=3) =

e.g. probability and binomial coefficient motivation: 5 adults drawn from large healthy population. How many had at least 1 cold in past year? (Need to *assume* these adults are independent of each other … in large population.)

$p = 0.20$ probability of 1+ colds
$q = 1-p = 0.8 =$ probability of no colds

Let's name these adults: A, B, C, D, E
Possible values of $X = \{0, 1, 2, 3, 4, 5\}$

What is the probability that <u>3 or more of the 5 adults</u> had 1+ colds in last year: $P(X \geq 3)$?

<u>Let's first consider P(X = 3)</u>:
1) First 3 trials give success: $p\ p\ p\ q\ q = p^3q^2 = (0.2)^3(0.8)^2$
This gives the probability for one specific ordering of the outcomes, e.g. A+, B+, C+, D-, E-

2) But, 3 individuals having 1+ colds can occur in more than one way…

To determine the number of ways that 3 out of 5 adults had 1+ colds and 2 out of 5 adults had 0, let's think about ordering and consider the motivation for the binomial coefficient…

Motivation for the binomial coefficient: $\binom{n}{x}$

5! = "5 factorial"
= number of *permutations* (order matters) of 5 individuals into 5 slots: <u>5</u> <u>4</u> <u>3</u> <u>2</u> <u>1</u>
= 5 x 4 x 3 x 2 x 1
= 120

5!/2! = number of permutations or orderings of 5 individuals into only 3 slots: $\underline{5}\ \underline{4}\ \underline{3}$
These slots represent the 3 affected individuals. But these slots are still ordered and we don't really care about that.

3! = number of permutations or orderings of 3 individuals into 3 affected slots: $\underline{3}\ \underline{2}\ \underline{1}$

What we want then are the number of *combinations* (order doesn't matter) of individuals, 3 of whom are affected and 2 of whom are not. So, the number of *combinations* of 5 individuals, 3 of whom are affected is: $\frac{(5!/_{2!})}{3!} = \frac{5!}{2!3!}$. These combinations are mutually exclusive and exhaustive. For P(X = 3):

$$P(X = 3) = \binom{5}{3}(0.2)^3(0.8)^2 = \frac{5!}{2!\,3!}(0.2)^3(0.8)^2 = \frac{120}{12}(0.2)^3(0.8)^2 = 0.0512$$

What about P(3 or more adults with 1+ colds)?

By the addition rule of probabilities for mutually exclusive events (see Lecture SA3 Intro Probability.pdf in the “What to know – Statistics” section of the Week 0 “Packet for Enrolled Students” on Canvas):

$$
\begin{aligned}
P(X \geq 3) &= P(X = 3) \quad + P(X = 4) \quad + P(X = 5) \\
&= \binom{5}{3}(0.2)^3(0.8)^2 + \binom{5}{4}(0.2)^4(0.8)^1 + \binom{5}{5}(0.2)^5(0.8)^0 \\
&= 0.0512 \quad + 0.0064 \quad + 0.00032 \\
&= 0.05792
\end{aligned}
$$

```
# R code to obtain probabilities
dbinom( 3, 5, 0.2) #P(X=3 w/colds | n=5adults, prob cold=0.2)
[1] 0.0512

dbinom (4, 5, 0.2) #P(X=4 w/colds | n=5adults, prob cold=0.2)
[1] 0.0064

dbinom (5, 5, 0.2) #P(X=5 w/colds | n=5adults, prob cold=0.2)
[1] 0.00032

# Or you can use one line of code to incorporate the 3 lines above
sum(dbinom (3:5, 5 ,0.2)) #P(X>=3 w/ colds | n=3 adults, prob cold=0.2)
[1] 0.05792
```

Properties of the Binomial:

For $X \sim \text{Bin}(n,p)$: $X = \sum_{i=1}^{n} X_i$ where $X_i \overset{ind}{\sim} \text{Bern}(p)$,

Note: *ind* is abbreviation for independent; E(X$_i$)=p; V(X$_i$)=p(1-p)

Mean: E(X) = np

$E[X] = E[\sum_{i=1}^{n} X_i] = nE[X_i] = np$ (using sum of Bernoulli random variables to prove E(X))

$E[X] = \mu = \sum_{x=0}^{n} xP(X=x) = \sum_{x=0}^{n} x \binom{\boldsymbol{n}}{\boldsymbol{x}} p^x(1-p)^{n-x} = np$ (using defn. of E(X) for discrete r.v.)

Variance: V(X) = np(1-p)

$$V[X] = V\left[\sum_{i=1}^{n} X_i\right] \overset{ind}{=} \sum_{i=1}^{n} V[X_i] = nV[X_i] = np(1-p)$$

$V[X] = \sigma^2 = \sum_{x=0}^{n}(x-\mu)^2 P(X=x) = \sum_{x=0}^{n}(x-np)^2 \binom{\boldsymbol{n}}{\boldsymbol{x}} p^x(1-p)^{n-x} = np(1-p)$

Standard Deviation: s.d.(X) = $\sqrt{np(1-p)}$

As you can see, the binomial distribution is completely defined by its two parameters, *n* & *p*.

e.g. For X = number with 1+ colds, n = 5 and p = 0.2

μ = 5(0.2) = 1

σ^2 = 5(0.2)(0.8) = 0.8

Comments:

Verify using:
https://www.stat.berkeley.edu/~stark/Java/Html/BinHist.htm

- For p < 0.5, the distribution is skewed right
- For p > 0.5, the distribution is skewed left
- V(X) is greatest when p = 0.5
- When p = 0 or p = 1, V(X) = 0
- When *n* is "large" and *p* is "not too small or too large", the distribution of X is nearly symmetric and the 68% and 95% rules hold "fairly" well (see Lecture SA2 Descriptive.pdf in Canvas "Packet for Enrolled Students"):, but be sure to keep in mind here and below (***) what we've discussed regarding symmetry, the CLT, and accuracy of the normal approximation!

Binomial Proportions

We will often be most interested in the proportion of "successes", *p*: $\hat{p} = \frac{X}{n}$ is the estimate.

$E[\hat{p}] = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np}{n} = p$. Therefore, $\hat{p}$ is an unbiased estimator of *p.*

$$V[\hat{p}] = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}, SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}} = s.e.[\hat{p}]$$

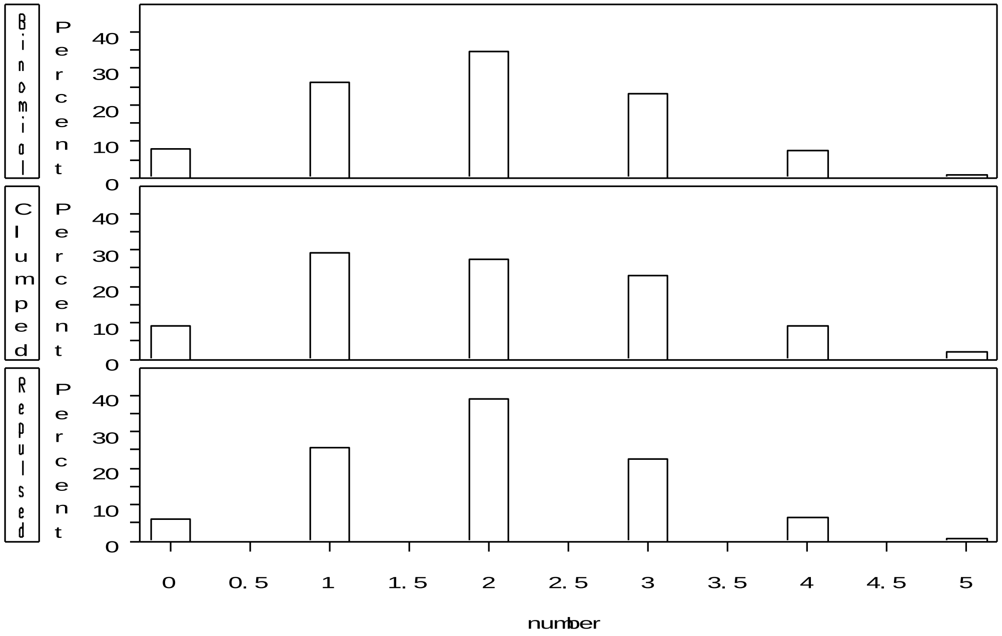
***A common rule of thumb: When np≥10 and n(1-p)≥10, the distribution of $\hat{p}$ is roughly symmetric. CAUTION!: See Lecture 2

How do we decide whether a binomial distribution or model is a good fit to a data sample?

1) Check assumptions – *n* independent trials, homogeneous *p*
2) Chi-square goodness of fit tests – fyi: https://www.youtube.com/watch?v=O7wy6iBFdE8
3) Check for departures – e.g. clumping (due to sometimes due to contagion) or repulsion – may be due to lack of independence of observations

Clumping: more observations at either extreme vs. in the center of the distribution than expected. This could be due to nonrandom sampling, infection, or heterogeneity among individuals. Sometimes called *extra-binomial variation* or *overdispersion*.

Repulsion: more observations in the center vs. the extremes of the distribution than expected. This occurs when some sort of compensatory mechanism develops or when saturation occurs. Sometimes called *underdispersion*. Less common than overdispersion.



Jakob Bernoulli (Basel, December 27, 1654 - August 16, 1705), also known as Jacob, Jacques or James Bernoulli, was a Swiss mathematician and scientist and the older brother of Johann Bernoulli.

His masterwork was *Ars Conjectandi* (the Art of Conjecturing), a groundbreaking work on probability theory. It was published eight years after his death in 1713 by his nephew Nicholas. The terms Bernoulli trial and Bernoulli numbers result from this work, and are named after him.

C) Poisson distribution – useful for rare outcomes in an (essentially) infinite population

The Poisson distribution is often, but not exclusively, used for modeling counts of relatively *rare* events – e.g. number of cases of cancer in Colorado in 2005, radioactive counts per unit of time, number of plankton per aliquot of seawater, number of raisins per cookie…

Unlike the binomial distribution, no upper bound is assumed for the sample space in the Poisson distribution, so the counts are assumed to occur in infinite (or very large) populations.

This distribution helps us to make inferences about *rates*. Rates are usually average occurrences of events over time or of objects over space, such as the number of events per person-time, or the density of objects in a specified area.

e.g. number accidents in Denver in one hour
number of bacterial colonies on an agar culture dish
number of raisins in an otherwise great cookie

In contrast to the binomial distribution, the Poisson distribution is completely defined by one parameter, λ, where $\lambda > 0$.

Let λ = the rate of occurrence of events per unit time, or the density of objects per unit area, or unit time. For a period of time *t* or an area *A*, we can interpret λt or λA as the *average* or *expected number* of events or of objects, μ, for the Poisson distribution.

Poisson probabilities:

For X ~ Poisson (λ), the probability of *x* events occurring (in a time period of length *t* or in an area *A*) with parameter λ is the probability mass function:

$$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x=0,1,2,\ldots$$

e = Euler's number (a constant) = $1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{\infty}=2.71828\ldots$

e is the base for the natural log

e.g. X = number of accidents in Denver in one hour

If X ~ Poisson (λ=2), i.e. an average of 2 accidents per hour:

$P(0) = \frac{2^0 e^{-2}}{0!} = 0.135$

$P(1) = \frac{2^1 e^{-2}}{1!} = 0.271$

$P(2) = \frac{2^2 e^{-2}}{2!} = 0.271$, etc.

```
# R code for these probabilities

dpois(0, 2) # P(X=0 accidents in 1 hr | avg 2/hr)
[1] 0.1353353

dpois(1, 2) # P(X=1 accidents in 1 hr | avg 2/hr)
[1] 0.2706706

dpois(2, 2) # P(X=2 accidents in 1 hr | avg 2/hr)
[1] 0.2706706
```

Assumptions of Poisson processes

1) The probability that a single event occurs in a small interval is proportional to the length/size of the interval Δt; i.e. P(1 event) = $\lambda\Delta t$ (i.e. longer intervals are more likely to have an event).

2) The probability of observing 0 events over Δt is approximately $1 - \lambda\Delta t$.

3) Probability of observing more than 1 event over this time interval is essentially 0.

4) *Stationarity*: Number of events per unit time is the same throughout the entire time interval t. *Thorough mixing*: Number of objects per unit area (volume) is the same throughout the entire area (volume).

5) *Independence*: Events occur independently within an interval and between intervals.

The Poisson distribution, like the binomial, doesn't lend itself well to modeling epidemics of infectious diseases, or other phenomena where "clumping" can occur. But there are ways to incorporate or allow for overdispersion when modeling these – "mixed models"

Properties of a Poisson distribution

An interesting and characteristic property of the Poisson distribution is that its mean and variance are the same:

Mean: E(X) = λ

$$E(X) = \mu = \sum_x xP(X = x) = \sum_x x\frac{\lambda^x e^{-\lambda}}{x!} = \lambda$$

Variance: V(X) = λ

$$V(X) = \sigma^2 = \sum_x (x - \mu)^2 P(X = x) = \sum_x (x - \mu)^2 \frac{\lambda^x e^{-\lambda}}{x!} = \lambda$$

Standard Deviation: s.d.(X) = $\sqrt{\lambda}$

The Poisson distribution is a positively skewed distribution. It is “roughly symmetric” for values of $\mu \geq 10$. (Same caution as above for “roughly symmetric binomial”.)

e.g. Patients entering waiting line at pharmacy

X = number of patients entering per hour

λ = 2 patients per 15 minutes, t = one hour

λt = 2x4 = 8 patients per hour

P(X=0): $\frac{8^0 e^{-8}}{0!} = 0.00034$

If we expect 8 per hour, the probability of not seeing any is 0.00034.

```
dpois(0, 8) # P(X=0 arrivals in 1 hr | avg 8/hr)
[1] 0.0003354626
```

P(X≤5) = P(X=0) + P (X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)

$$= \frac{8^0 e^{-8}}{0!} + \frac{8^1 e^{-8}}{1!} + \frac{8^2 e^{-8}}{2!} + \frac{8^3 e^{-8}}{3!} + \frac{8^4 e^{-8}}{4!} + \frac{8^5 e^{-8}}{5!}$$

= 0.00034 + 0.00268 + 0.01073 + 0.02863 + 0.05725 + 0.09160 = 0.1912

```
sum(dpois(0:5, 8)) # P(X<=5 arrivals in 1 hr | avg 8/hr)
[1] 0.1912361
```

P(X≥10) = 1 - P(X< 10)

= 1 - [P(X≤5) + P(X=6) + P(X=7) + P(X=8) + P(X=9)]

= 1 - [0.1912 + 0.1221 + 0.1396 + 0.1396 + 0.1241]

= 1 - 0.7166 = 0.2834

```
1-ppois(9, 8) # P(X>=10 arrivals in 1 hr | avg 8/hr)
[1] 0.2833757
```

Probability calculations like this can be used to determine if the pharmacy should add another service window or not.

D) Poisson approximation to the binomial distribution:

The Poisson distribution can be used to approximate a binomial distribution under the right conditions, i.e. when *n* is large and *p* is small. Thus, the Poisson serves as a *limiting* distribution for the binomial.

The distribution of X can be approximated by a Poisson with mean $\mu = \lambda = np$.

For *n* large, *p* small:

$$V(X) = np(1-p) \to np \text{ since } 1-p \to 1$$
$$E(X) = np \to \lambda$$

Some rules of thumb for approximating the binomial by the Poisson:

1. when $n \geq 100$ and $p \leq 0.01$ (Rosner)
2. when $p < 0.1$ and $np < 5$ (Sokal and Rohlf)
3. Stigler (2013) Digital Approximation of the Binomial by the Poisson, *American Statistician* - a seriously amusing view of just how good (or not) the Poisson approximation to the binomial can be. You can find it on Canvas in the Paper Repository.

e.g. 1000 females age 40-49 with a maternal history of breast cancer are followed for 1 year. Four cases of breast cancer were observed in this group of women.

In the general population of women age 40-49, the incidence of breast cancer regardless of family history is 1/1000 per year = 0.001 = *p*

In one year, 1 case would be expected for these 1000 females: *np* = 1000(0.001) = 1 case of breast cancer (*do we meet our rules of thumb to approximate the binomial by the Poisson?*)

In this study 4 cases were observed after 1 year of follow-up, does this give evidence that these females are at risk due to a genetic component? That is, are these women different from the general population? If we consider X=4 cases of breast cancer to be an extreme result (since we expected only 1), what's the probability of seeing this or anything more extreme?

$P(X \geq 4) = P(4 \text{ or more breast cancer cases in 1 year} \mid \text{we expect } 1/1000)$

Binomial: $P(X \geq 4) = P(X=4) + P(X=5) + \ldots + P(X=1000)$

$$= \binom{1000}{4}(0.001)^4(0.999)^{996} + \binom{1000}{5}(0.001)^5(0.999)^{995} + \cdots + \binom{1000}{1000}(0.001)^{1000}(0.999)^0$$

…or we could calculate 1 - P(X < 4).

Notice that the conditions for the Poisson approximation to the binomial hold ($p < 0.01$ and $n > 100$; $p < 0.1$ and $np < 5$). This binomial distribution has $\mu = np = 1000(0.001) = 1$. And this becomes the mean $\mu=\lambda$ for the approximating Poisson distribution.

For the Poisson:

$P(X \geq 4) = \sum_{x=4}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3)]$ (the additive rule for mutually exclusive events, again!)

$= 1 - \left[\frac{1^0 e^{-1}}{0!} + \frac{1^1 e^{-1}}{1!} + \frac{1^2 e^{-1}}{2!} + \frac{1^3 e^{-1}}{3!}\right] = 0.019$

```
# binomial probability
bprob_x_ge_4 <- 1-pbinom(3, 1000, 0.001) # P(X>=4 w/ BrCa | n=1000 women, probBrCa = 0.001)
bprob_x_ge_4
[1] 0.01892683
# Poisson probability
pprob_x_ge_4 <- 1-ppois(3, 1) # P(X>=4 w/ BrCa | avg of 1 among 1000 women)
pprob_x_ge_4
[1] 0.01898816
```

We would expect an outcome as or more extreme less than 2 out of 100 times. This suggests that observing 4 or more cases of breast cancer in this group of women is *unlikely if* the general population incidence rate applies.

Conclusion: Therefore, the women with a maternal history of breast cancer do appear to be at higher risk than the general population. We have now made an *inference* about this population of women with respect to the general population.

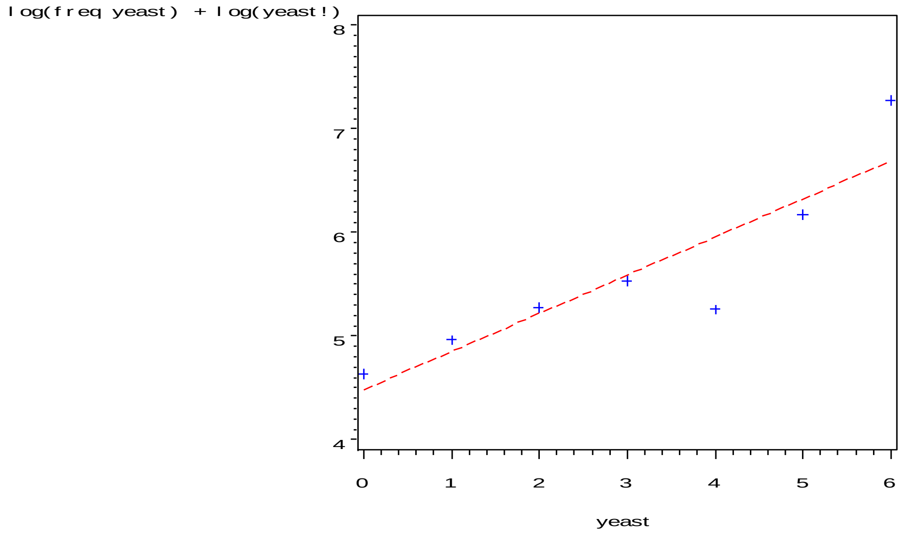
How do we decide whether a Poisson distribution or model is a good fit to the data?

1. Check assumptions
2. Chi-square goodness of fit tests
3. Look at $s^2/\bar{X}$ - should be close to 1; departures – e.g. clumping (due to sometimes due to contagion) or repulsion – may be due to lack of independence of observations, or heterogeneity within the population
4. Poissonness plot – Hoaglin, 1980. *Amer. Stat*, 34:146-149. (see pdf document on Canvas Paper Repository)

The FREQ Procedure - Observed Frequency of Yeast Cells in 400 Squares; W.S. Gosset (Student)

```
                          Cumulative  Cumulative
yeast   Freq   Percent    Frequency   Percent
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
  0      103    25.75        103       25.75
  1      143    35.75        246       61.50
  2       98    24.50        344       86.00
  3       42    10.50        386       96.50
  4        8     2.00        394       98.50
  5        4     1.00        398       99.50
  6        2     0.50        400      100.00
```

The plot below should show a straight-line relationship. Does the distribution look Poisson?



Siméon Denis Poisson

Born: 21 June 1781 in Pithiviers, France
Died: 25 April 1840 in Sceaux (near Paris), France

http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Poisson.html

Classic example of use of the Poisson distribution:
http://www.mun.ca/biology/scarr/smcPoisson_distributions.html