Section III

Population distributions

The Normal (Gaussian) distribution model

(Application to Sensitivity & Specificity)

Binomial distribution

Poisson distribution

III – Distributions and the Gaussian Distribution

Notation - We will introduce two types of notation to distinguish between **samples** and **populations** (universes). A **population** or universe is a complete census of all members of a group. For example, if we want to know the average age of practicing pediatricians in the city of Los Angeles, we will know the this average exactly if we can get the birth date of all practicing pediatricians (say, from State records). However, more often we will not have a complete census, but will only have a sample. While the population may consist of 12,000 pediatricians, our sample may consist of far fewer.

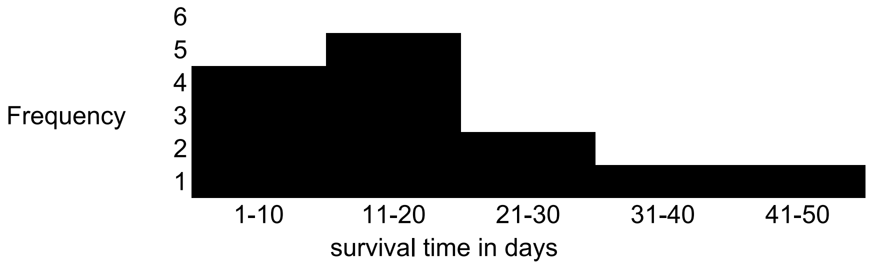
We generally use **Greek** letters to denote **population** quantities, called population parameters, and we use **Latin** letters to denote sample quantities, call **sample** statistics.

Notation in populations versus samples

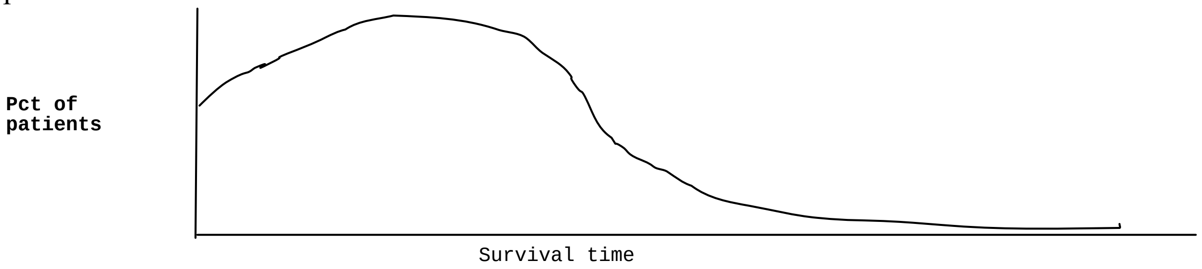
Quantity	population symbol	sample symbol
Mean	μ	$\bar{Y}$
Standard deviation	σ	S
Proportion	π	P
Correlation coeff	ρ	r
Slope	β	b
Intercept	α	a
Number of persons	N	n

Probability distributions

As we saw earlier, the distribution of sampled values can be represented by a histogram as in our stomach cancer survival data.



Suppose that the sample comes from a large population of patients. If we had the survival times of all patients in the population (i.e. all with stomach cancer) we could draw a histogram for the whole population. Since the population size is large, the class or bin sizes for the histogram could be small. We could also smooth over the steps of the histogram and produce a continuous curve. It might look like the picture below.

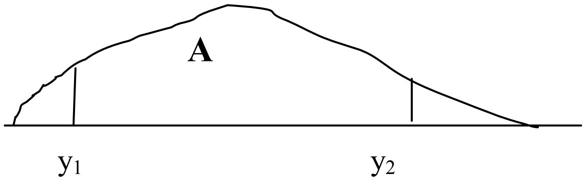


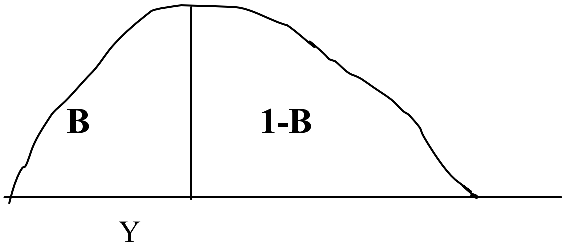
If the vertical axis of this curve is scaled so that the area under the curve is one, it is called a **probability distribution**. The continuous function **f(y)** sketched above is called the **probability density function**. Numerical quantities that have a probability distribution in a population are called **random variables**. So, in this example, we have a probability distribution of the random variable Y = survival time.

Properties of a probability distribution – area under the curve

In a probability distribution, the area under the curve between two values y_1 and y_2 equals the proportion of the population with values of Y between y_1 and y_2. Using "pr" to stand for this proportion (or this probability), one writes in symbols

$$\Pr(y_1 < X < y_2) = A$$

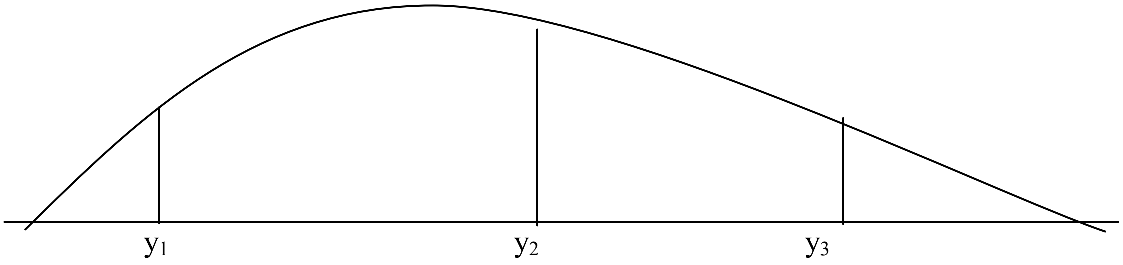




where A is the area shown below. By definition, if B = pr(Y < y), then y is called the 100th B percentile of the distribution. That is, y is the 100Bth percentile if y is the value that has 100 B percent of the area "behind" it.

The area A is also the probability that a randomly selected member of the population has a value of Y between y_1 and y_2. The probability of other events can be found by using the relationship between probability and area under the probability distribution. For example

$$\Pr(y_1 < Y < y_2) + \Pr(y_2 < Y < y_3) = \mathrm{pr}(y_1 < Y < y_3)$$



Population means and variances

The population mean μ and variance σ^2 of Y are defined by the probability distribution by the formulas

$$\mu = \int y\ f(y)\ dy \qquad\qquad \sigma^2 = \int (y - \mu)^2\ f(y)\ dy$$

The population SD is the root of the population variance. That is $\sigma = \sqrt{\sigma^2}$.
These definitions are analogous to the sample mean, variance and SD.

Standardized variables (Z scores)

If the mean μ is subtracted from all the Y values in the population, the resulting quantity has mean zero. Moreover, if we also divide Y-μ by the standard deviation σ, the resulting quantity has mean zero and standard deviation of 1.0. That is, the quantity Z, defined as

$$\mathbf{Z = (Y - \mu)/ \sigma \qquad (so \quad Y = \mu + Z\,\sigma\,)}$$

has mean zero and SD=1 over the entire population. Of course, in general, μ and σ are not usually known. Instead, in any sample, we only know $\bar{Y}$ and S.

In our survival data sample, if the true mean is μ=20 days and the true σ is σ=14 days, the standardized values are as below. Y= original data in day, Z is in SD units from μ.

Y= 4, 6, 8, 8, 12, 14, 15, 17, 19, 22, 24, 34, 45
Z=(Y-20)/14= -1.14,-1.0, -0.86,-0.86,-0.57, -0.43, -0.36, -0.21, -0.07, 0.14, 0.29, 1.0,1.79

If the true mean is 17.5 days, the same as the sample mean, and the true SD is 11.68 days, the same as the sample SD, then the Z scores are

Y= 4, 6, 8, 8, 12, 14, 15, 17, 19, 22, 24, 34, 45
Z=(Y-17.5)/11.68 -1.16,-0.99, -0.82, -0.82, -0.47,-0.30, -0.22, -0.05, 0.13, 0.38, 0.55, 1.41, 2.35

The Gaussian ("Normal") distribution

The "bell shaped" or Gaussian distribution is a special density function (f(y)) that plays a central role as a statistical model for data. Mathematically, the definition of the Gaussian density function with mean μ and SD of σ is given by

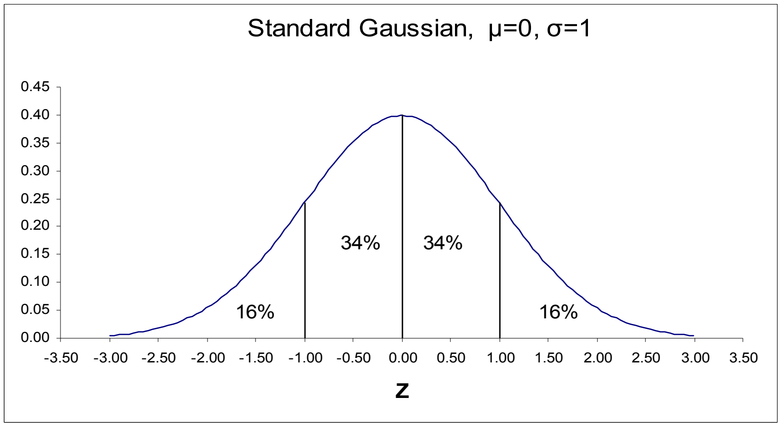
Rel frequency= $f(y) = (1/\sqrt{[2\pi\,\sigma^2]}\,)\exp[\,-(1/2\sigma^2)\,(y-\mu)^2]$ (don't memorize this formula)

The "standard" Gaussian is the Gaussian with μ=0 and σ=1. Setting μ=0 and σ=1 in the above formula gives the standard Gaussian density function

$f(z) = (1/\sqrt{[2\pi]}\,)\exp[-z^2/2]$ (don't memorize this formula either)

Area under the curve for the standard Gaussian

Mathematically, the area from –infinity to Z is given by the integral of f(z). However, this integral has no closed form.



Thus, tables are given for the area below Z .

Value of Z	area below Z (=percentile/100)
-2	0.0228 = 2.28%
-1.5	0.0668 = 6.68%
-1	0.1587 = 15.87%
0	0.5000 = 50.0%
+1	0.8413 = 84.13%
+1.5	0.9332 = 93.32%
+2	0.9772 = 97.72%

Using the standard Gaussian to calculate area for any Y assumed to have a Gaussian distribution.

If we want areas and/or percentiles for a variable Y with **<u>known</u>** mean μ and **<u>known</u>** SD σ, one can first calculate the standardized value

$$Z = (Y - \mu)/ \sigma$$

and then look up the area (percentile) corresponding to Z.

Example: Assume Y has a Gaussian distribution with mean μ=20 and SD σ = 14. What proportion of the Y values will be less than 35. This is the same as the probability that Y will be less than 35.

Answer: Z = (35-20)/14 = 1.07. Looking up 1.07 on the Gaussian table gives an area of (about) 0.85. Thus, 35 is about the 85th percentile. That is, about 85% of the area is less than 35. Equivalently, the probability of being 35 or less is 0.85. The probability of being more than 35 is 1-0.85=0.15.

Example: What is the upper quartile (75th percentile) value Y for a Gaussian distribution with mean μ=8 and SD σ = 5.

Answer: The area of 0.75 corresponds to the standardized Z value of Z= 0.674. We know Z, we now need Y. So, 0.674 = (Y – 8)/5. Therefore, Y = 5 (0.674) + 8 = 11.37.

Using the Gaussian

To a very good approximation, test scores on the National Boards (i.e SAT) have a Gaussian distribution with population mean, μ=500 and population SD, σ=100

Q: What is the 80th percentile (a value called Y_{80})?

A: $Y_{80} = 500 + Z_{80} \times 100 = 500 + .842 \times 100 = 584$

Q: What percentiles do scores of 700, 500 and 450 correspond to?

A:

X	Z	area (percentile)
700	(700-500)/100 = 2	.9772 (97.72%)

500	(500-500)/100 =	0	.5000	(50.0 %)
450	(450-500)/100 =	-.5	.3085	(30.85%)

The Standard Cumulative Gaussian Distribution
Z is in SD units below or above the mean

Z	area behind z	Z	area behind z	Z	area behind z
	(percentile/100)		(percentile/100)		(percentile/100)
-3.00	**0.0013**				
-2.95	0.0016	-0.95	0.1711	1.05	0.8531
-2.90	0.0019	-0.90	0.1841	1.10	0.8643
-2.85	0.0022	-0.85	0.1977	1.15	0.8749
-2.80	0.0026	-0.80	0.2119	1.20	0.8849
-2.75	0.0030	-0.75	0.2266	1.25	0.8944
-2.70	0.0035	-0.70	0.2420	1.30	0.9032
-2.65	0.0040	-0.65	0.2578	1.35	0.9115
-2.60	0.0047	-0.60	0.2743	1.40	0.9192
-2.55	0.0054	-0.55	0.2912	1.45	0.9265
-2.50	0.0062	-0.50	0.3085	1.50	0.9332
-2.45	0.0071	-0.45	0.3264	1.55	0.9394
-2.40	0.0082	-0.40	0.3446	1.60	0.9452
-2.35	0.0094	-0.35	0.3632	1.65	0.9505
-2.30	0.0107	-0.30	0.3821	1.70	0.9554
-2.25	0.0122	-0.25	0.4013	1.75	0.9599
-2.20	0.0139	-0.20	0.4207	1.80	0.9641
-2.15	0.0158	-0.15	0.4404	1.85	0.9678
-2.10	0.0179	-0.10	0.4602	1.90	0.9713
-2.05	0.0202	-0.05	0.4801	1.95	0.9744
-2.00	**0.0228**	**0.00**	**0.5000**	**2.00**	**0.9772**
-1.95	0.0256	0.05	0.5199	2.05	0.9798
-1.90	0.0287	0.10	0.5398	2.10	0.9821
-1.85	0.0322	0.15	0.5596	2.15	0.9842
-1.80	0.0359	0.20	0.5793	2.20	0.9861
-1.75	0.0401	0.25	0.5987	2.25	0.9878
-1.70	0.0446	0.30	0.6179	2.30	0.9893
-1.65	0.0495	0.35	0.6368	2.35	0.9906
-1.60	0.0548	0.40	0.6554	2.40	0.9918
-1.55	0.0606	0.45	0.6736	2.45	0.9929
-1.50	0.0668	0.50	0.6915	2.50	0.9938
-1.45	0.0735	0.55	0.7088	2.55	0.9946
-1.40	0.0808	0.60	0.7257	2.60	0.9953
-1.35	0.0885	0.65	0.7422	2.65	0.9960
-1.30	0.0968	0.70	0.7580	2.70	0.9965
-1.25	0.1056	0.75	0.7734	2.75	0.9970
-1.20	0.1151	0.80	0.7881	2.80	0.9974
-1.15	0.1251	0.85	0.8023	2.85	0.9978
-1.10	0.1357	0.90	0.8159	2.90	0.9981
-1.05	0.1469	0.95	0.8289	2.95	0.9984
-1.00	**0.1587**	**1.00**	0.8413	**3.00**	**0.9987**

95.45% of the area is between Z= -2 and Z=2

95% of the area is between Z=-1.96 and Z=1.96

So, for example, a person with a score of 700 is (approximately) in the 97th percentile <u>if</u> Board scores follow a Gaussian (or normal) distribution. In fact, board scores are nearly normal so this approximation is not perfect, but is not bad.

Example to be done in class: Anesthesia (i.e. Halothane)

To put adults to sleep the average dose of Halothane needed is μ= 50 mg per kg of body weight (in one minute) with a standard deviation of σ = 10 mg per kg of body weight (in one minute).

However, the average <u>lethal</u> dose is μ=110 mg with a standard deviation of σ = 20 mg.

Q: What dose will put 90% to sleep? What percent might die as a result of this dose?

A: (In class)

Prediction intervals & "normal" clinical range

When μ and σ are known population values and the data is assumed to have a **Gaussian** distribution, the interval formed by

$$(\mu - Z\sigma\ ,\ \ \mu + Z\sigma)$$

is called a **<u>prediction interval</u>**. If Z > 0 is the kth Gaussian percentile, the interval formed above contains the middle 2k-100 percent of the distribution. Thus, this interval is called a 2k-100 percent prediction interval for patients. For example, if Z=1.96, the k= 97.5th percentile, then the interval formed by (μ - 1.96σ, μ + 1.96σ) is a 2(97.5)-100 = 95% prediction interval (not a 97.5% prediction interval) since the <u>middle</u> 95% of the population's values (X) are within these bounds. The 95% prediction interval is also sometimes called the "normal" clinical range. If Z=1.645, (the 95th percentile), the interval (μ-1.645σ, μ+1.645σ) is a 90% prediction interval (not 95%). Note that a prediction interval gives the Gaussian theory based percentage of <u>individual</u> patient values that lie within the specified bounds. **DO NOT** confuse this prediction interval with confidence intervals which will be studied later.

Crude **Rule of thumb** (for quantities with a **<u>Gaussian</u>** distribution)

The middle 95% of patient values (most of the values) are approximately in the "normal" range of

$$(\mu - 2\ \sigma,\quad \mu + 2\sigma)$$

$$\text{therefore } \sigma \approx \text{ ``range'' } / \ 4$$

Where "range" excludes all "unusual" patients (patients below the 2.5th percentile and those above 97.5th percentile)

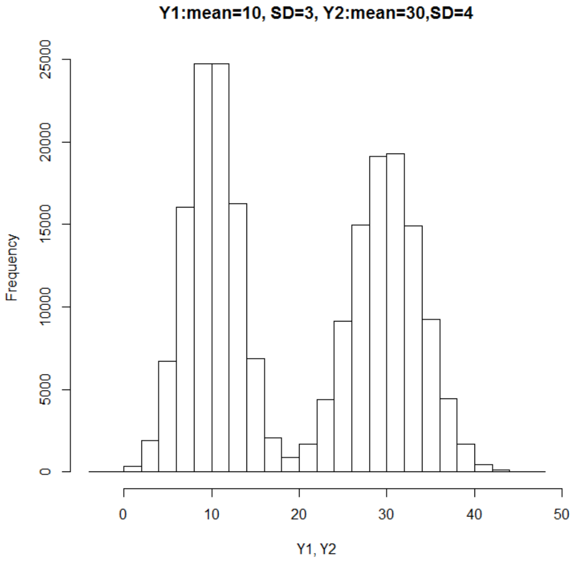
Differences and sums of normally distributed variables

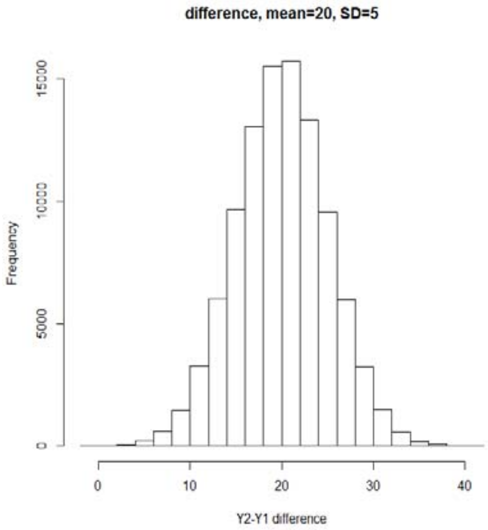
If Y_1 and Y_2 each have a normal distribution and are independent with means and SDs as given below,

variable	mean	SD
Y_1	μ_1	σ_1
Y_2	μ_2	σ_2

Then	mean	SD
diff= d= Y_1-Y_2	μ_1-μ_2	sqrt($\sigma_1^2 + \sigma_2^2$)
sum= Y_1+Y_2	μ_1+μ_2	sqrt($\sigma_1^2 + \sigma_2^2$)

The difference and the sum have normal distributions as well.





Computing false positive and negative rates from the Gaussian.
Computing **specificity** and **sensitivity** for a diagnostic test.

Let Y = Serum Creatinine in mg/dl

In normal adults	$\mu = 1.1$ mg/dl $\sigma = 0.2$ mg/dl
In (at least one type of) renal disease	$\mu = 1.7$ mg/dl $\sigma = 0.4$ mg/dl

We suspect renal disease **if** Y > 1.6 mg/dl - the **cutoff or threshold** value (Test is **positive** for disease if Y > 1.6)

Assuming a Gaussian distribution in both populations

Q: What is probability (Pr) of a **false positive**? (Patient tests positive but is normal)

Pr($Y > 1.6$ given μ=1.1, $\sigma = 0.2$)

$Z = (1.6 - 1.1)/0.2 = 2.5$

from Gaussian Table, area above 2.5 is .0062 or .62%.

Therefore, **specificity =** probability of a true negative = 1-.0062 = 0.9938 or about 99% specific.

Specificity = Probability test is negative given that the patient is normal.

Q: What is probability of a **false negative** ?
(Patient tests negative even though diseased)

Pr($Y < 1.6$ given μ=1.7, $\sigma = 0.4$)

$Z = (1.6 - 1.7)/0.4 = -0.25$

from Gaussian Table, area below -.25 is .4013 or about 40%

Therefore **sensitivity** = probability of a true positive = 1-0.4013 = 0.5987 or about 60% sensitive.

Sensitivity = Probability that the test is positive given that the
patient actually has disease.

(also see notes at end of section 2)

Data Transformations, log transformation

One might think that the Gaussian model is not very useful since it is limited only to data that have symmetric, unimodal distributions. It is not a good model for skewed distributions.

However, a skewed distribution can often be made symmetric via a **scale transformation**, usually just called a **transformation**. Below is a figure showing the distribution of serum bilirubin in normal adults, a quantity that is useful in assessing liver function. As is obvious from the figure, bilirubin has a skewed distribution. However, the next figure shows the distribution of the **same** data on a logarithmic scale. In this example, the distribution of **log** bilirubin is much closer to a symmetric one. Therefore, using Gaussian theory and the mean and standard deviation to summarize the distribution of log bilirubin is more meaningful. That is, the Gaussian model is at least approximately correct on this transformed scale.

On the log 10 scale, the mean log bilirubin value is 1.55 log μmol/L. The **antilog** of this value is $10^{1.55}$ = 35.5 μmol/L. This antilog of the mean of the log values is called the **geometric mean.** Notice that it is quite a bit lower than the arithmetic mean of 64.3 μmol/L and close to the median of 34.7 μmol/L.

Similarly, if we attempt to compute a "normal" clinical range for the middle 95% of patients using Gaussian theory (Z=2) and the mean and SD of the original untransformed data we get the absurd range of 64.3 μmol/L +/- 2 x 104.3 μmol/L or (-144.3 μmol/L, 272.9 μmol/L). Of course, bilirubin cannot have negative values!

However, if we compute a normal clinical range on the **log** data using Gaussian theory, we obtain 1.55 log μmol/L +/- 2 x 0.456 log μmol/L or (0.64 log μmol/L, 2.46 log μmol/L) in log 10 units. To express this in the original μmol/L units, we take the antilog of each range endpoint and obtain ($10^{0.64}$, $10^{2.46}$) or (4.3 μmol/L, 290 μmol/L) as our approximate "normal" range. These values more closely agree with the impression we get from the figure of range where the middle 95% of the bilirubin values lie.

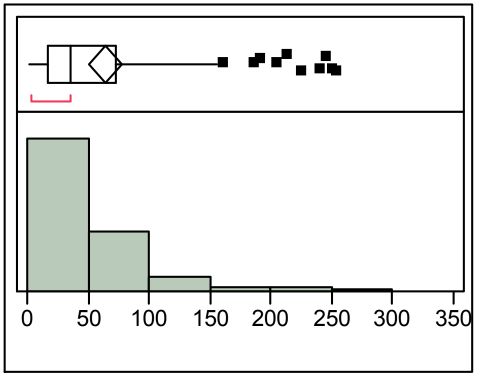
While this example is obvious when the data histogram is given, it is less obvious when the histogram is not given and only the means and standard deviations are reported.

Many types of continuous data follow a normal distribution on the log scale (log normal distribution) including bacterial growth and proliferation measures such as CFU (colony forming units), antibody or antigen titers (IgA .. IgG etc), pH, response to sound or other neurological stimuli (dB), most steroids and hormones (Estrogen, Testosterone …), cytokines (IL-1, MCP-1, …) and liver function measures (Bilirubin, Creatinine) to name a few. When no transformation can be found, one should quote medians and ranges, not means and SDs.

The distribution of Ratios usually more closely follow the Gaussian on the log scale including ORs, RRs and HRs. On the log scale 100/1 (log=2) is symmetric with 1/100 (log=-2).

Ratio:	100/1,	10/1,	1/1,	1/10,	1/100
Log ratio:	2,	1,	0,	-1,	-2

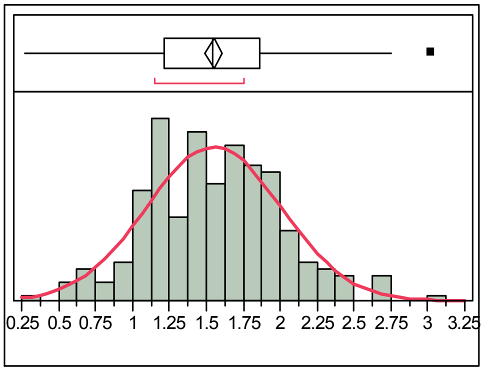
Biulirubin (μmol/L)



100.0%	maximum	1041.6
97.5%		367.2
90.0%		129.3
75.0%	quartile	73.4
50.0%	median	34.7
25.0%	quartile	16.2
10.0%		10.3
2.5%		4.6
0.0%	minimum	1.9

Mean	64.297
Std Dev	104.300
Std Err Mean	7.097
n	216

log 10 Bilirubin (log μmol/L)



100.0%	maximum	3.0177
97.5%		2.5479
90.0%		2.1116
75.0%	quartile	1.8659
50.0%	median	1.5402
25.0%	quartile	1.2096
10.0%		1.0127
2.5%		0.6607
0.0%	minimum	0.2741

Mean	1.5496
Std Dev	0.4561
Std Err Mean	0.0310
n	216

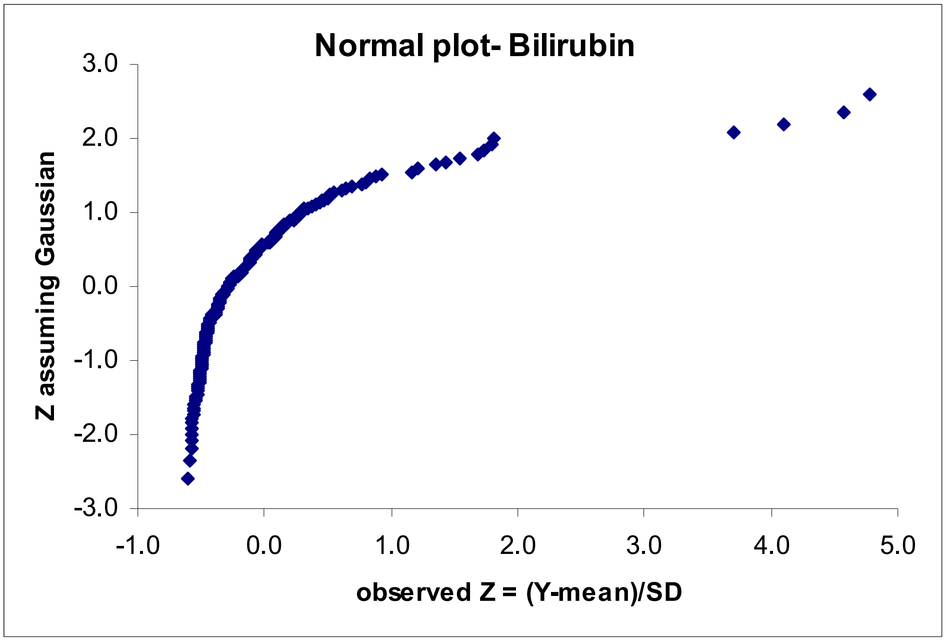
Geometric mean: $10^{1.55}$ = 35.5 μmol/L (not 64.3)

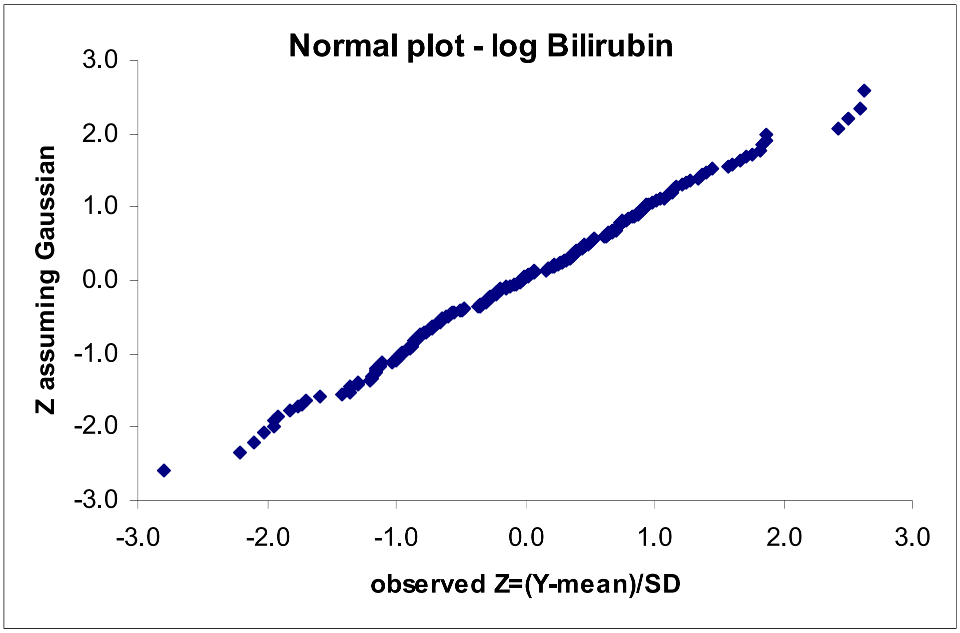
“normal range (mean +/- 2 SD)

$$10^{[1.55 - 2(0.456)]} = 10^{0.64} = 4.3 \ \mu\text{mol/L}$$
$$10^{[1.55 + 2(0.456)]} = 10^{2.46} = 289.7 \ \mu\text{mol/L}$$

Normal probability plot–a distribution diagnostic plot

After the data is sorted from lowest to highest and the percentiles computed, the observed Z value is plotted versus the expected Z value assuming the percentiles come from a Gaussian distribution.





If the data distribution is Gaussian, the plot is a straight line.

Data distributions that tend to be Gaussian on the log scale

Growth measures - bacterial CFU
Ab or Ag titers (IgA, IgG, …)
pH
Neurological stimuli (dB, Snellen units)
Steroids, hormones (Estrogen, Testosterone)
Cytokines (IL-1, MCP-1, …)
Liver function (Bilirubin, Creatinine)
Hospital Length of stay (can be Poisson)

The distribution of ratios is much closer to Gaussian on the log scale

The "inverse" of 3/1 is 1/3. This is symmetric only on the log scale

Original: 100/1, 10/1, 1/1, 1/10, 1/100
Log: 2, 1, 0, -1, -2

true for OR, RR and HR

Measures of growth & proliferation have distribution closer to the Gaussian on the log scale

“Quick” Probability Theory- Terms

Mutually exclusive events (“or”) - levels of one variables

When events are **mutually exclusive**, their probabilities add. For example, for blood type, one must be either type A, type B, type AB or type O. These four categories are exclusive and exhaustive. If 50% of the population is type O and 20% is type A, then the probability of being type A or type O is 50%+20%=70%.

Independent events (“and”) - different variables

If two events are **independent**, their probabilities multiply. For example, if 5% of pregnant women have gestational diabetes, and 8% have preclampsia **and the two events are independent**, then the probability of having both diabetes and preeclamsia is 5% x 8% = 0.4%. This multiplication rule **requires** independence. It is often misused as independence is lacking.

Conditional probability

The probability of an event changes if it is made conditional on another event. For example, the probability (prevalence) of tuberculosis in the general adult population is only about 0.1%. But, in Vietnamese immigrants the probability is 4%. That is, **conditional** on being a Vietnamese immigrant, the probability of TB is 4%.

Bayes’ rule for computing conditional probability

Probability of B given A = $P(B \mid A)$ =

Joint Probability of A and B / Probability of A = $P(A \cap B) / P(A)$

= [Probability of A given B x Probability of B] / Probability of A

$$= [\ P(A|B)\,P(B)\ ] \ / \ P(A)$$

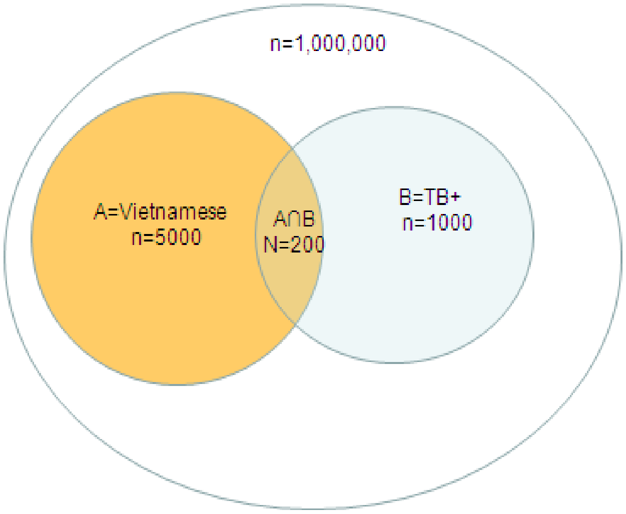
Example: In a population of 1,000,000 persons, 5,000 are Vietnamese immigrants (0.5%). In these 1,000,0000, 1000 (0.1%) have TB. Of all 1000 who have TB, 20% are Vietnamese immigrants (we only know ethnicity in TB cases). What is the probability of TB conditional on being a Vietnamese immigrant? A = Vietnamese, B = have TB

P(A) = 5,000/1,000,000 = 5/ 1000 = 0.5% = 0.005 = probability of being Vietnamese in pop
P(B) = 0.1% = 0.001 = Probability of TB in population
P(A|B) = 20% = 0.20 = Conditional Probability of being Vietnamese given that one has TB

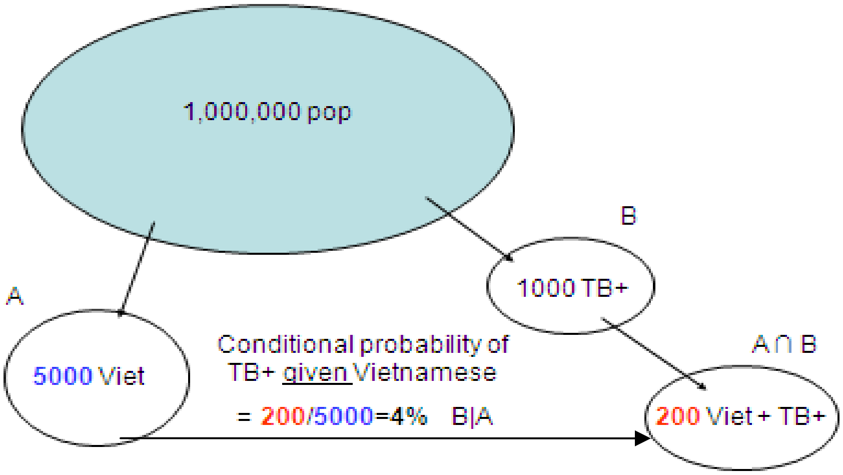
P(B|A) = (0.20 x 0.001) / 0.005 = 0.04 = 4% = Probability of TB given Vietnamese

If A and B are independent $P(B|A) = P(B)$

Conditional Probability



Bayes rule (graph)



Conditional probability of TB+ <u>given</u> Vietnamese

= 200/5000=4%   B|A

Bayesian vs Frequentist

The Bayesian approach is to compute

Prob(hypothesis|data) =

$$\frac{\text{Prob(data|hypothesis)}}{\text{Prob(data)}} \;\text{x}\; \underline{\text{ P(hypothesis)}}$$

= Data Likelihood x prior probability

If data (evidence) refutes a hypothesis then
Prob(data | hypothesis)=0 so
Prob(hypothesis | data)=0

The frequentist approach is to compute

Prob(data+|hypothesis)= p value

The p value is the probability of the observed data or something more extreme than the observed data (“data+”) under the (null) hypothesis.

They are not the same and both can be useful.

Binomial distribution for binary data – 0 or 1 per patient

The other extreme from a continuous outcome is a binary (positive or negative) outcome. Sometimes we wish to know the probability of "y" positive responders out of "n" people where "π" is the probability of a positive response in any one person (so [1-π] is the probability of a negative response) and the response in any person is independent of the response in any other person. In a set of n people (n independent events), there are y positive responders and n-y negative responders.

Examples:

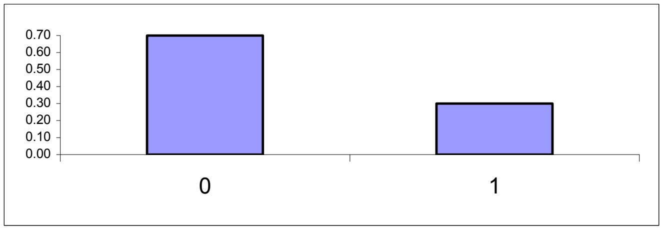
Binomial Distribution

Y= number of positive ("yes") responses out of n

	population	
pos	**π =**	**0.3**
neg	**(1-π)=**	**0.7**

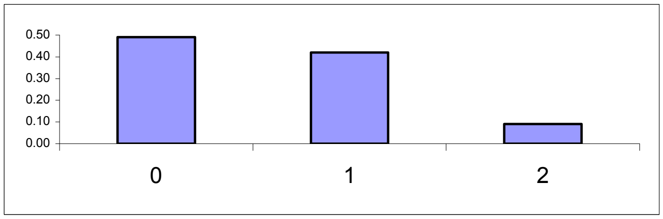
n=1

Y	probability
0	0.700
1	0.300
total	1.00



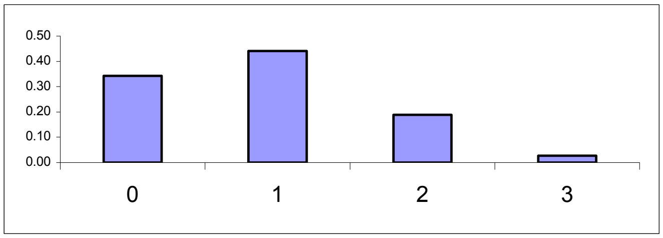
n=2

Y	probability
0	0.49
1	0.42
2	0.09
total	1.000



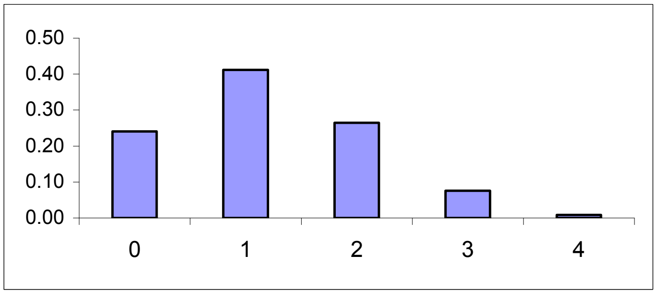
n=3

Y	probability
0	0.343
1	0.441
2	0.189
3	0.027
total	1.000



n=4

Y	probability
0	0.2401
1	0.4116
2	0.2646
3	0.0756
4	0.0081
total	1.000



For **n=2**, there are three possible values of Y: 0, 1 or 2, each occurring with the probability below

Y	probability
0=0+0	$(1-\pi)\,(1-\pi) = (1-\pi)^2$
1=1+0 or 0+1	$\pi\,(1-\pi) + (1-\pi)\,\pi = 2\,\pi\,(1-\pi)$
2=1+1	$(\pi)(\pi) = \pi^2$
Total	1.0

In general, the Binomial formula below can be used to compute the probability of "**y**" positive patients out of **n** patients. The full formula is

Probability of "y" responders out of "n" patients

$$= n!/[y!(n-y)!]\ \pi^y\ (1-\pi)^{(n-y)}$$

$$[\ a! = a\,(a-1)\,(a-2)\,(a-3)\ \ldots\ (3)\,(2)\,(1)\ ]$$

The expected (mean) number of responses = πn

The SD of the number of responses out of n patients is $SD=\sqrt{[n\,\pi\,(1-\pi)]}$

Examples: Q: What are the expected number of cases of Herpes in 50 teens if the prevalence is 4%?

A: $\pi= 0.04$, $n=50$. We expect $50 \times 0.04 = 2$ cases,
($SD=\sqrt{50 \times 0.04 \times 0.96} = 1.4$, but responses don't have Gaussian distribution)

Q: What is the probability of observing exactly 5 Herpes cases in 50 teens if the Herpes prevalence is 4%.

A: Probability = $50!/(5!\ 45!)\ (0.04)^5\ (0.96)^{45} = 0.034561=3.4\%$

Q: What is the probability of observing 5 or fewer cases?
(Plug in 5,4,3,2,1,0 and add) A: probability = 0.98559.
What is the probability of 6 or more cases = 1-0.98559=0.01441.

Can compute with "=BINOMDIST(y,n,π,0)" in EXCEL.

Special case – a “fair” coin

When $\pi = 0.5$, the probability of “y” successes (“heads”) out of n is just
Probability = $\mathbf{n!/[y!(n-y)!] / 2^{n}}$

Example : n=3 (flip 3 fair coins), $2^3 = 8$ possibilities

0+0+0=0=y
0+0+1=1=y
0+1+0=1=y
1+0+0=1=y
0+1+1=2=y
1+0+1=2=y
1+1+0=2=y
1+1+1=3=y

y	**freq**	**prob**
0	**1**	**1/8**
1	**3**	**3/8**
2	**3**	**3/8**
3	**1**	**1/8**
total	**8**	**8/8**

Pascal’s triangle - computing frequencies and probabilities

n	y: 0 to n successes	2^n
-	1	
1	1 1	2
2	1 2 1	4
3	1 3 3 1	8
4	1 4 6 4 1	16
5	1 5 10 10 5 1	32

For n=5,	y	freq	probability
	0	1	1/32
	1	5	5/32
	2	10	10/32
	3	10	10/32
	4	5	5/32
	5	1	1/32

Hypothesis testing- Binomial case

Question: How likely is y=7 success out of n=10 coins if the coins are fair? The coins are fair if π=0.50.

Probability of 7 success out of 10 =

$$10!/(7! \text{ x } 3!) / 2^{10} = 120/1024 = 0.1172$$

How likely is 7 or more successes out of n=10?

y	probability
7	120/1024 = 0.1172
8	45/1024 = 0.0439
9	10/1024 = 0.0098
10	1/1024 = 0.0010
total	176/1024= 0.1719

Question: How likely is observing 70 successes (y=70) out of n=100 if the coins are fair?

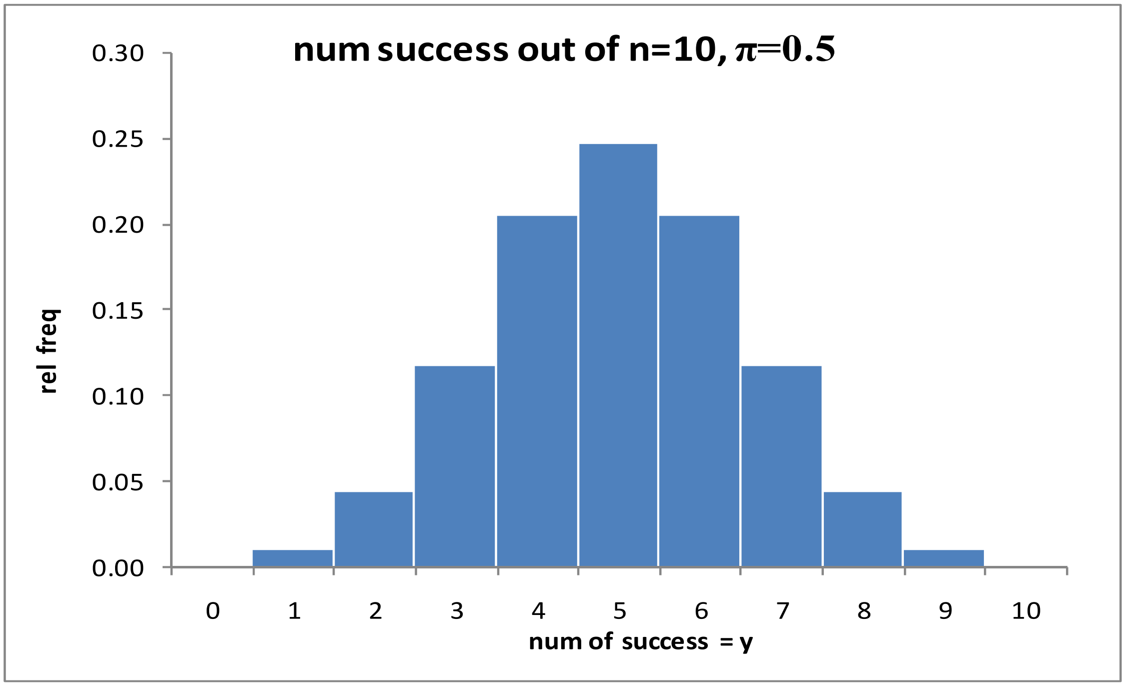
$$\text{Prob}(y=70) = [100!/(70!\ 30!)] / 2^{100} = 2.32 \text{ x } 10^{-5}$$

How likely is it to observe 70 or more successes out of 100?

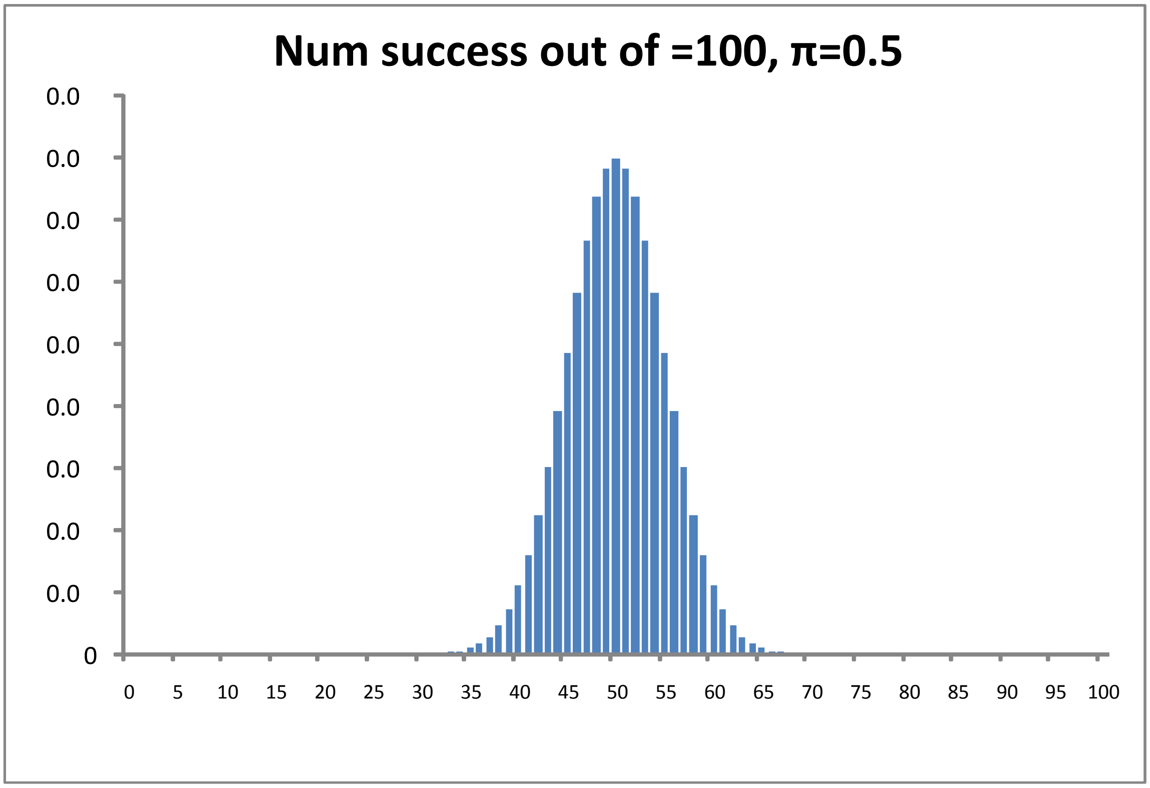
$$= \text{prob}(y=70) + \text{prob}(y=71) + \ldots + \text{prob}(y=100) = 3.93 \text{ x } 10^{-5}$$

This is a simple example of hypothesis testing. The probability of observing y=70 or more successes out of n=100 under the "null hypothesis" that the true population π=0.5 is called a one sided **p value.**

In both cases, the sample proportion is 0.7. But if n=10 the p value is 0.1719. If n=100, the p value is 3.93 x 10^{-5}.



Area above y=7 is 0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.1719

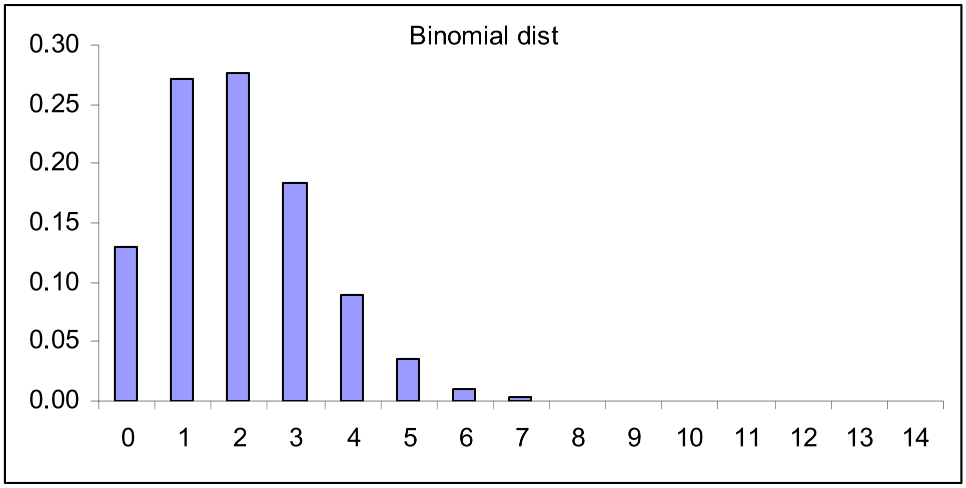


Area above y=70 is very small

Gaussian approximation to the Binomial

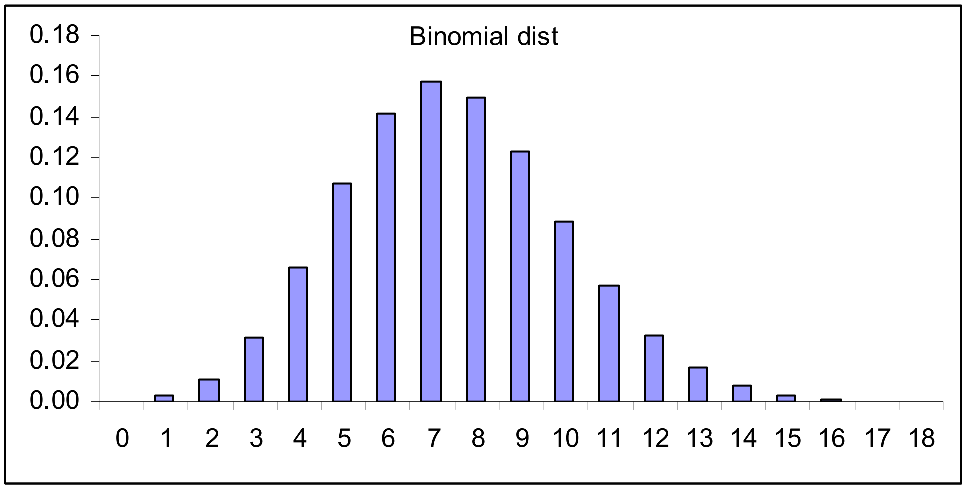
In general, the binomial distribution is not necessarily symmetric

π = 0.04, n=50, mean =(0.04)(50)=2, SD = 1.4



But, if π is not too close to 0 or 1.0 and n is not too small, the distribution of y, the number (out of n) with a positive response, can be approximated by a Gaussian with mean **nπ** and standard deviation √ **[n π (1-π)]**

π = 0.15, n= 50, mean =0.15(50)=7.5, SD = 2.5



Actual 2.5th percentile between 2 and 3 Gaussian = 7.5 – 2 (2.5) = 2.5
Actual 97.5th percentile between 12 and 13 Gaussian = 7.5 + 2(2.5) = 12.5

<u>Poisson distribution for count data</u>

For a patient, y is a positive integer: 0,1,2,3,…

Probability of "y" responses (or events) given mean μ

$$= (\mu^{y} e^{-\mu}) / (y!)$$

(Note: $\mu^{0}=1$ by definition)

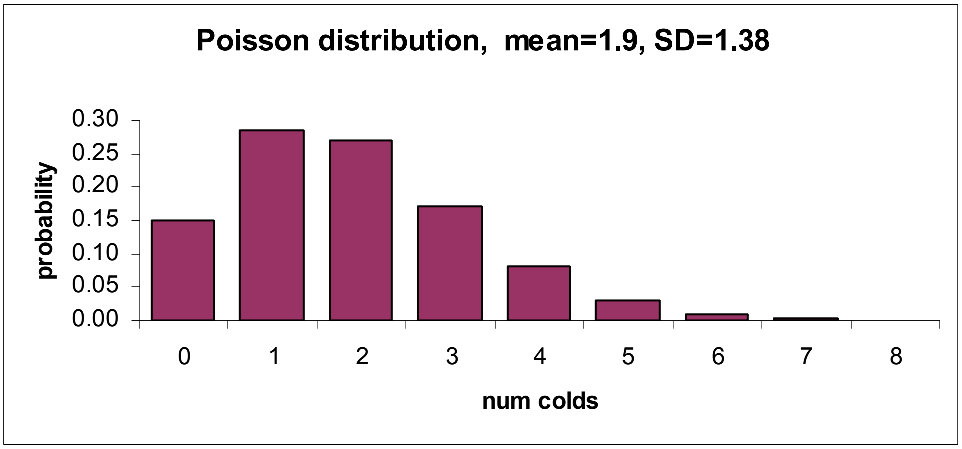
For Poisson, if mean=μ then SD=√μ

Examples: number if colds per season, number of firings of a neuron in 30 seconds (firing rate).

Q: If the average number colds in a single winter is μ=1.9, what is the probability that a given patient will have 4 colds in one winter?

A: $(1.9)^{4}\ e^{-1.9}/4x3x2x1 = 0.0812 \approx 8\%$.

What is the probability of 4 or more (find for 0-3, subtract from 1), prob=12%



Note: y! = y (y-1) (y-2) (y-3) … (2) (1)

Example: 5! = 5 x 4 x 3 x 2 x 1 = 120

Poisson process

Mean rate of events is h/time unit (Hazard rate). In T time units, we expect μ=hT events on average. Can substitute in above formula to get probability of "y" events in T time units.

Example: Cancer clusters

Q: Given a cancer rate of 3/1000 per year, what is the expected number of cases in 2 years in a population of 1500?
A: Rate in 2 years is 2 x (3/1000) =h= 6/1000. Expected is μ=hT= 6/1000 x 1500 = 9 cases.

Q: What is the probability of observing exactly 15 cases?
A: μ=9, Probability =$(9^{15}\ e^{-9})/15!$ = 0.019431≈ 2%.

Q: What is the probability of observing 15 or more cases in 1500 persons?
A: Plug in 0,1,2, …14 and add to get Q= probability of 14 or less. Probability is 1-Q = 1-0.958534 = 0.041466 ≈ 4%.

Can compute this with "=Poisson(y,μ,0)" in EXCEL.

T does not have to be time. For example, in studying traffic accidents, investigators are interested in the number of accidents per mile of freeway. For T miles of freeway, the number of accidents may follow a Poisson distribution with mean hT.

****

Summary: descriptive statistics for Normal, Binomial & Poisson

n = sample size

Distribution	mean	variance	SD	SE
Normal	μ	σ^2	σ	$\sigma/\sqrt{n}$
Binomial	π	$\pi(1-\pi)$	$\sqrt{}\pi(1-\pi)$	$\sqrt{}\pi(1-\pi)/n$
Poisson	μ	μ	$\sqrt{\mu}$	$\sqrt{\mu/n}$

$SD=\sqrt{\text{variance}}$, $SE=SD/\sqrt{n}$