Probability Distributions

II

Multinomial Distribution - Motivation

Suppose we modified assumption (1) of the binomial distribution to allow for more than two outcomes.

For example, suppose that for the family with parents that are heterozygote carriers of a recessive trait, we are interested in knowing the probability of

$\mathbf{Q_1}$: One of their n=3 offspring will be unaffected (AA), 1 will be affected (aa) and one will be a carrier (Aa),

$\mathbf{Q_2}$:All of their offspring will be carriers,

$\mathbf{Q_3}$:Exactly two of their offspring will be affected (aa) and one will be a carrier.

Multinomial Distribution - Motivation

For each child, we can represent these possibilities with three indicator variables for the *i*-th child as

$Y_{i1} = 1$ if unaffected (AA), & 0 otherwise
$Y_{i2} = 1$ if carrier (Aa), & 0 otherwise
$Y_{i3} = 1$ if affected (aa), & 0 otherwise

Notice only one of the three Y_{i1}, Y_{i2}, Y_{i3} can be equal to 1, so $\Sigma_j Y_{ij} = 1$.

For the binomial distribution with 2 outcomes, there are 2^n unique outcomes in n trials. In the family with n=3 children, there are $2^3 = 8$ unique outcomes.

For the multinomial distribution with n trials and only 3 outcomes, the number of unique outcomes is 3^n. For our small family, that's 3^3=27 outcomes.

Possible Outcomes

Combinations: As with the binomial, there are different ways to arrange possible outcomes from a total of *n* objects (trials) if order doesn't matter. For the multinomial distribution, the combinations are summarized as

$$C_k^n = \frac{n!}{k_1!k_2...k_J!}$$

where the k_j (j=1,2,…,J) correspond to the totals for the different outcomes.

E.g. (*n*=2 offspring)

Child number

1	2	Outcomes
AA	AA	2 unaffected, 0 carrier, 0 affected
AA	Aa	1 unaffected, 1 carrier, 0 affected
Aa	AA	1 unaffected, 1 carrier, 0 affected
AA	aa	1 unaffected, 0 carrier, 1 affected
aa	AA	1 unaffected, 0 carrier, 1affected
Aa	Aa	0 unaffected, 2 carrier, 0 affected
aa	Aa	0 unaffected, 1 carrier, 1 affected
Aa	aa	0 unaffected, 1 carrier, 1 affected
aa	aa	0 unaffected, 0 carrier, 2 affected

For the case of n=2 offspring (i.e., trials), what are the probabilities of these outcomes?

E.g. (n=2, k_1=unaffected, k_2=carrier, k_3=affected)

Child number

1	2	Outcomes	# ways
p_1	p_1	$k_1=2, k_2=0, k_3=0$	1
p_1	p_2	$k_1=1, k_2=1, k_3=0$	2
p_2	p_1	$k_1=1, k_2=1, k_3=0$	
p_1	p_3	$k_1=1, k_2=0, k_3=1$	2
p_3	p_1	$k_1=1, k_2=0, k_3=1$	
p_2	p_2	$k_1=0, k_2=2, k_3=0$	1
p_3	p_2	$k_1=0, k_2=1, k_3=1$	2
p_2	p_3	$k_1=0, k_2=1, k_3=1$	
p_3	p_3	$k_1=0, k_2=0, k_3=2$	1

For each possible outcome, the probability $\Pr[Y_1=k_1, Y_2=k_2, Y_3=k_3]$ is

$$p_1^{k1} p_2^{k2} p_3^{k3}$$

There are $\frac{n!}{k_j!}$ sequences for each probability, so in general…

Multinomial Probabilities

What is the probability that a multinomial random variable with **n** trials and success probabilities p_1, p_2, ..., p_J will yield exactly k_1, k_2, ...k_J successes?

$$\mathrm{P}(Y_1 = k_1, Y_2 = k_2, \ldots, Y_J = k_J) = \frac{n!}{k_1!k_2!\ldots k_J!} p_1^{k_1} p_2^{k_2} \cdots p_J^{k_J}$$

Assumptions:

1) J possible outcomes – only one of which can be a success (1) a given trial.

2) The probability of success for each possible outcome, p_j, is the same from trial to trial.

3) The outcome of one trial has no influence on other trials (independent trials).

4) Interest is in the (sum) total number of "successes" over all the trials.

k_1	k_2	k_3	k_4	· · ·	k_{J-1}	k_J

$n = \Sigma_j k_j$ is the total number of trials.

Multinomial Random Variable

A <u>multinomial random variable</u> is simply the total number of successes in *n* trials.

Example: family of 3 offspring.

	child 1		child 2		child 3		Total
Q_1:	1 0 0	+	0 0 1	+	0 1 0	=	1 1 1
Q_2:	0 1 0	+	0 1 0	+	0 1 0	=	0 3 0
Q_3:	0 0 1	+	0 1 0	+	0 0 1	=	0 1 2

Multinomial Probabilities - Examples

Returning to the original questions:

Q₁: One of *n*=3 offspring will be unaffected (AA), one will be affected (aa) and one will be a carrier (Aa) (recessive trait, carrier parents)?

Solution: For a given child, the probabilities of the three outcomes are:

p_1 = Pr[AA] = 1/4,
p_2 = Pr[Aa] = 1/2,
p_3 = Pr[aa] = 1/4.

We have

$$P(Y_1 = 1, Y_2 = 1, Y_3 = 1) = \frac{3!}{1!1!1!} p_1^1 p_2^1 p_3^1$$

$$= \frac{(3)(2)(1)}{(1)(1)(1)} \left(\frac{1}{4}\right)^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^1$$

$$= \frac{3}{16} = 0.1875.$$

Exercises

Q_2: What is the probability that all three offspring will be carriers?

Q_3: What is the probability that exactly two offspring will be affected and one a carrier?

Solutions

Q_2: What is the probability that all three offspring will be carriers?

$$P(Y_1 = 0, Y_2 = 3, Y_3 = 0) = \frac{3!}{0!3!0!} p_1^0 p_2^3 p_3^0$$

$$= \frac{(3)(2)(1)}{(3)(2)(1)} \left(\frac{1}{4}\right)^0 \left(\frac{1}{2}\right)^3 \left(\frac{1}{4}\right)^0$$

$$= \frac{1}{8} = 0.125.$$

Q_3: What is the probability that exactly two offspring will be affected and one a carrier?

$$P(Y_1 = 0, Y_2 = 1, Y_3 = 2) = \frac{3!}{0!1!2!} p_1^0 p_2^1 p_3^2$$

$$= \frac{(3)(2)(1)}{(2)(1)} \left(\frac{1}{4}\right)^0 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^2$$

$$= \frac{3}{32} = 0.09375.$$

Example - Mean and Variance

It turns out that the (marginal) outcomes of the multinomial distribution are binomial. We can immediately obtain the means for each outcome (i.e., the j[th] cell)

MEAN: $E[k_j] = E\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} E[Y_{ij}]$

$$= \sum_{i=1}^{n} p_j = np_j$$

VARIANCE:

$$V[k_j] = V\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} V[Y_{ij}]$$

$$= \sum_{i=1}^{n} p_j(1-p_j) = np_j(1-p_j)$$

COVARIANCE:

$$Cov[k_j, k_{j'}] = -np_j p_{j'}$$

Multinomial Distribution Summary

Multinomial

1. Discrete, bounded
2. Parameters - $n, p_1, p_2, \ldots, p_J$
3. Sum of n independent outcomes
4. Extends binomial distribution
5. Polytomous regression, contingency tables

Continuous Distributions

Continuous Distributions

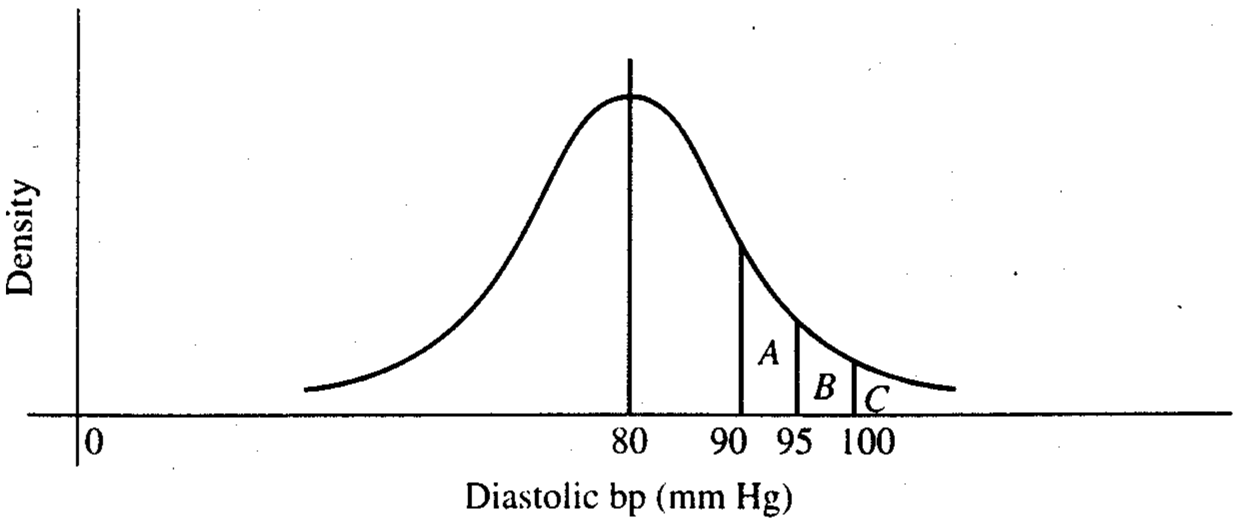
For measurements like height and weight which can be measured with arbitrary precision, it does not make sense to talk about the probability of any single value. Instead we talk about the probability for an **interval**.

$$P[\text{weight} = 70.000\text{kg}] \approx 0$$

$$P[69.0\text{kg} \leq \text{weight} \leq 71.0\text{kg}] = 0.08$$

For discrete random variables we had a probability mass function to give us the probability of each possible value. For continuous random variables we use a **probability density function** to tell us about the probability of obtaining a value within some interval.

E.g. Rosner - diastolic blood pressure in 35-44 year-old men (figure 5.1)



For any interval, the **area** under the curve represents the probability of obtaining a value in that interval.

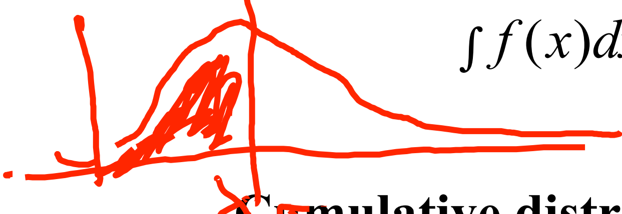
Probability density function

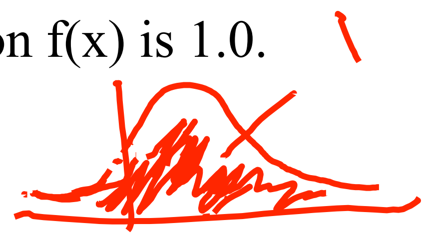
1. A function, typically denoted f(x), that gives probabilities based on the **area** under the curve.

2. $f(x) \geq 0$

3. Total area under the function f(x) is 1.0.

$$\int f(x)dx = 1.0$$



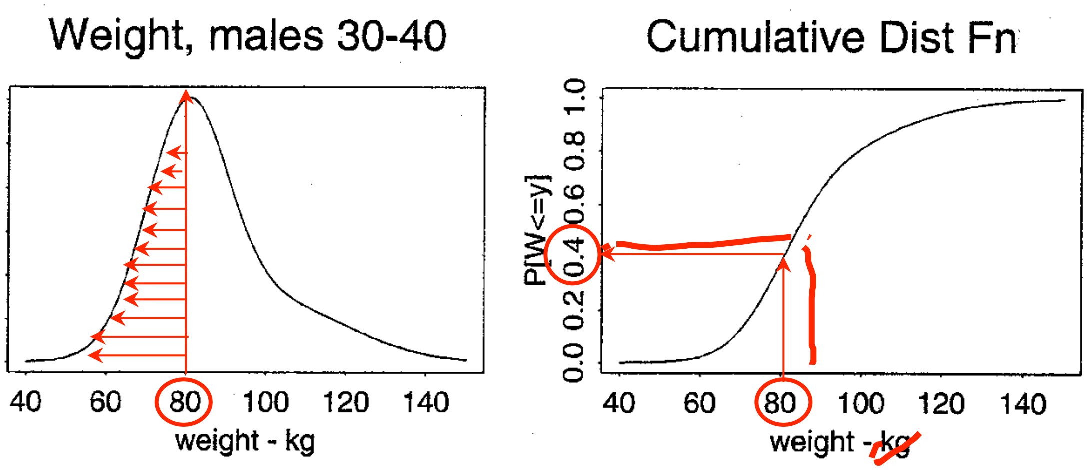


Cumulative distribution function

The <u>cumulative distribution function</u>, F(t), tells us the total probability less than some value t.

$$F(t) = P(X \leq t)$$

This is analogous to the cumulative relative frequency.



Prob[wgt < 80] = 0.40

Area under the curve

Pr(random man's weight ≤ 85) ≅ 0.42

Normal Distribution

- A common probability model for continuous data
- Can be used to characterize the Binomial or Poisson under certain circumstances
- Bell-shaped curve

⇒ takes values between $-\infty$ and $+\infty$

⇒ symmetric about mean

⇒ mean=median=mode

- Examples

  birthweights

  blood pressure

  CD4 cell counts (perhaps transformed)

Normal Distribution

Specifying the mean and variance of a normal distribution completely determines the probability distribution function and, therefore, all probabilities.

The **normal probability density function** is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

where

$\pi \approx 3.14$ (a constant)

Notice that the normal distribution has two parameters:

μ = the mean of X

σ = the standard deviation of X

We write $X \sim N(\mu , \sigma^2)$. The **standard normal** distribution is a special case where $\mu = 0$ and $\sigma = 1$.

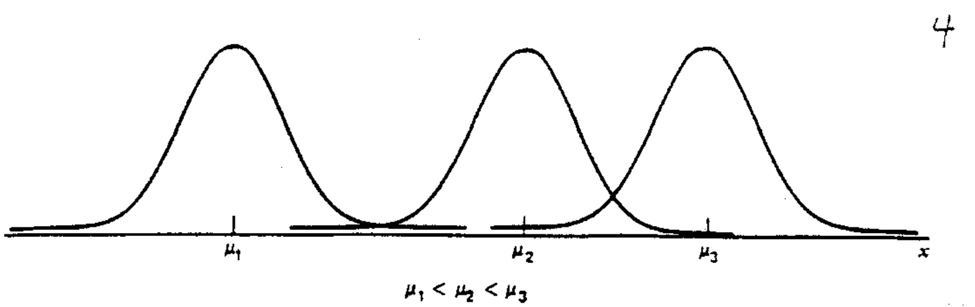


FIGURE 3.6.3
Three Normal Distributions with Different Means

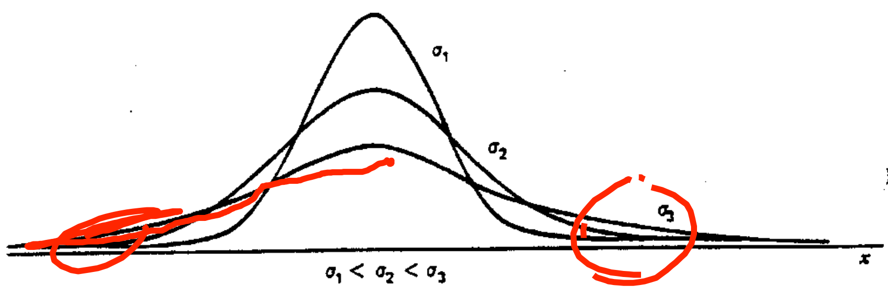


FIGURE 3.6.4
Three Normal Distributions with Different Standard Deviations

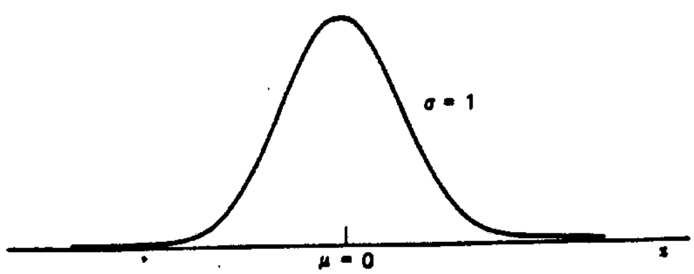
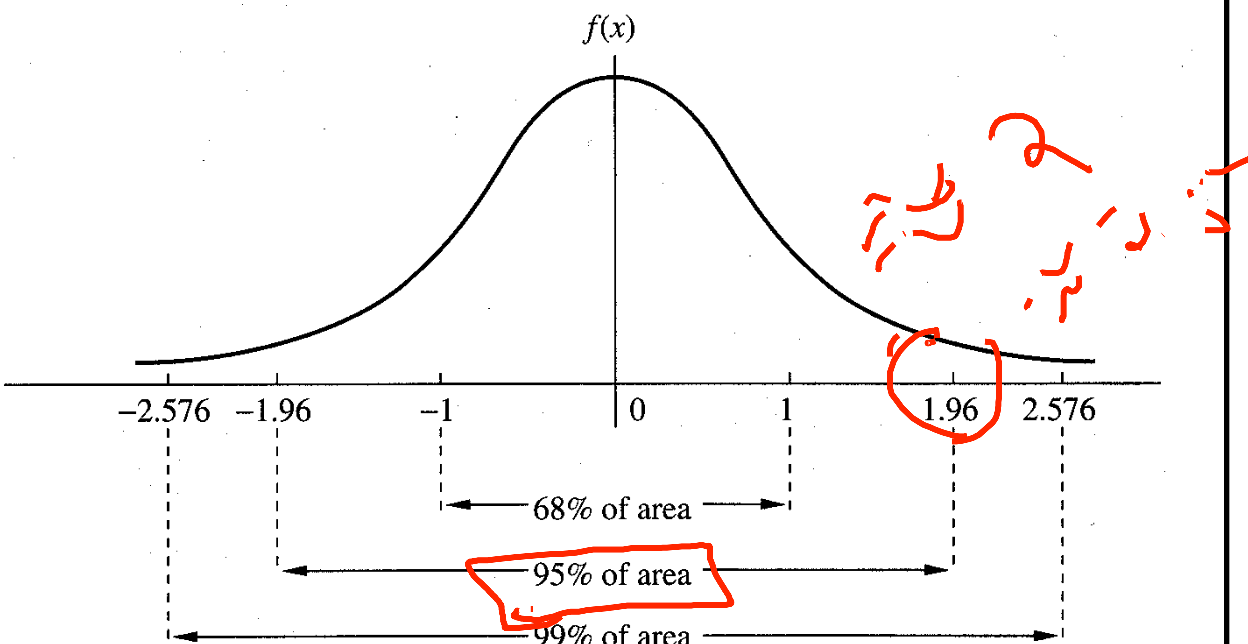


FIGURE 3.6.5
The Unit Normal Distribution


f(x)
−2.576
−1.96
−1
0
1
1.96
2.576
68% of area
95% of area
99% of area

Normal Distribution - Calculating Probabilities

Example: Rosner 5.20

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

X = serum cholesterol in an individual.

$\mu =$ 219

$\sigma =$ 50

$$P[x < 200] = \int_{-\infty}^{200} \underbrace{\frac{1}{50\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-219)^2}{50^2}\right)}_{f(x)} dx$$

negative values for cholesterol - huh?

Standard Normal Distribution - Calculating Probabilities

First, let's consider the **standard normal** - N(0,1). We will usually use Z to denote a random variable with a standard normal distribution. The density of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right)$$

and the **cumulative distribution** of Z is:

$$P(Z \le x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right) dz$$

Any computing software can give you the values of f(z) and Φ(z)

Standard Normal Distribution - Calculating Probabilities

Google

cdf normal distribution calculator

All | Videos | Images | News | Shopping | More | Settings | Tools

About 69,900 results (0.84 seconds)

Free Cumulative Distribution Function (CDF) Calculator for the Normal ...
www.danielsoper.com/statcalc/calculator.aspx?id=53
This **calculator** will compute the **cumulative distribution function** (CDF) for the **normal distribution** (i.e., the area under the **normal distribution** from negative infinity ...

Cumulative Distribution Function (CDF) Calculator for the Normal Distribution

This calculator will compute the cumulative distribution function (CDF) for the normal distribution (i.e., the area under the normal distribution from negative infinity to x), given the upper limit of integration x, the mean, and the standard deviation.

Please enter the necessary parameter values, and then click 'Calculate'.

Mean: 0

Standard deviation: 1

x: 0.5

Calculate!

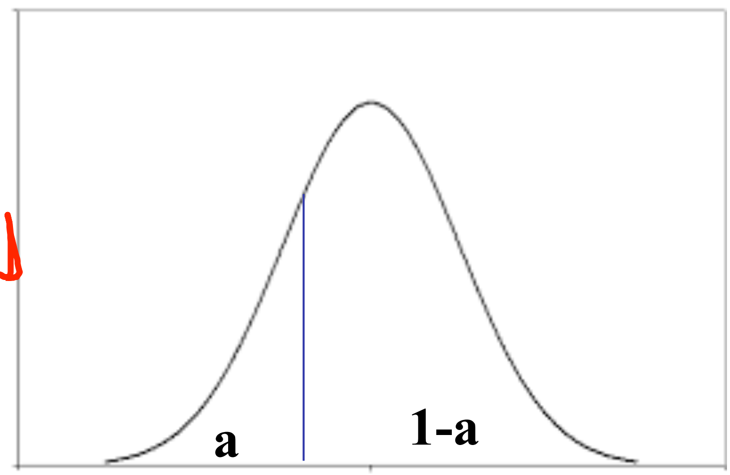
Cumulative distribution function: 0.69146246

$$Pr(Z \leq 0.5) = 0.69146$$

Facts about probability distributions

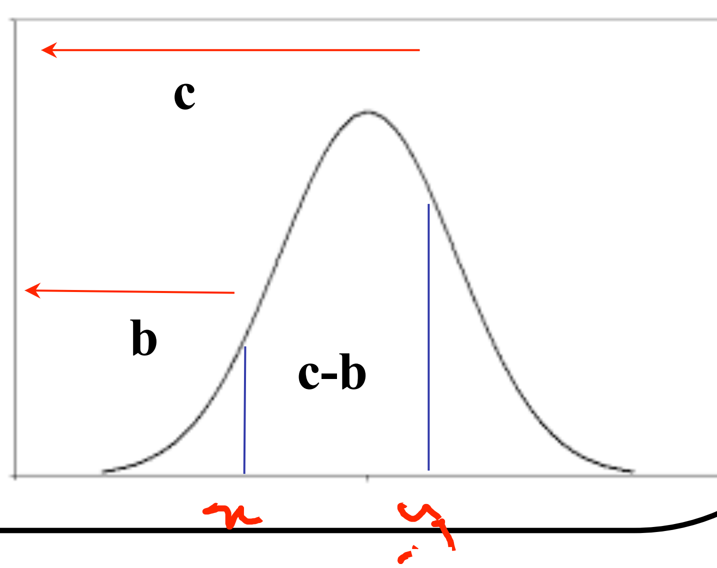
$P[Z \leq z] = a$

$\Rightarrow P[Z > z] = 1\text{-}a$



$P[Z \leq x] = b,\ P[Z \leq y] = c$

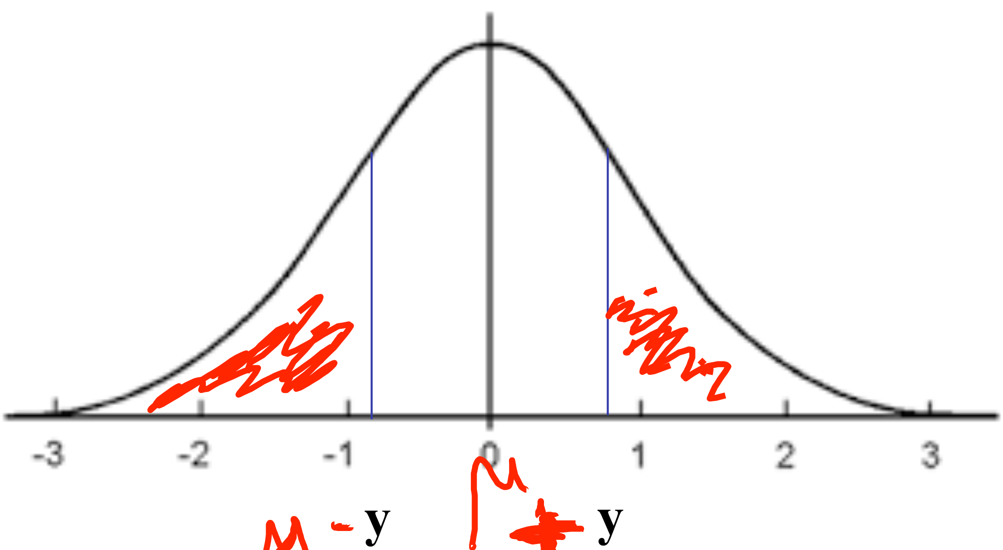
$\Rightarrow Pr[x < Z \leq y\ ] = c\text{-}b$



Facts about the standard normal distribution

Because the N(0,1) distribution is symmetric around 0,

$$\Pr[Z \le -y] = \Pr[Z \ge y]$$



Exercises

$P[Z \leq 1.65] =$

$P[Z \geq 0.5] =$

$P[-1.96 \leq Z \leq 1.96] =$

$P[-0.5 \leq Z \leq 2.0] =$

$Z \sim N(0,1)$

Solutions to Exercises

$P[Z \leq 1.65] = 0.9505$

$P[Z \geq 0.5] = 1 – 0.6915 = 0.3085$

$P[-1.96 \leq Z \leq 1.96] = 0.975 – 0.025 = 0.95$

$P[-0.5 \leq Z \leq 2.0] = 0.9772 – 0.3085=0.6687$

Converting to Standard Normal

This solves the problem for the N(0,1) case. Do we need a special table for every (μ,σ)?

No!

Define: $X = \mu + \sigma Z$ where $Z \sim N(0,1)$

1. $E(X) = \mu + \sigma E(Z) = \mu$
2. $V(X) = \sigma^2 V(Z) = \sigma^2$.
3. X is normally distributed!

Linear functions of normal RV's are also normal.

If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$
then

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

Converting to Standard Normal

How can we convert a $N(\mu,\sigma^2)$ to a standard normal?

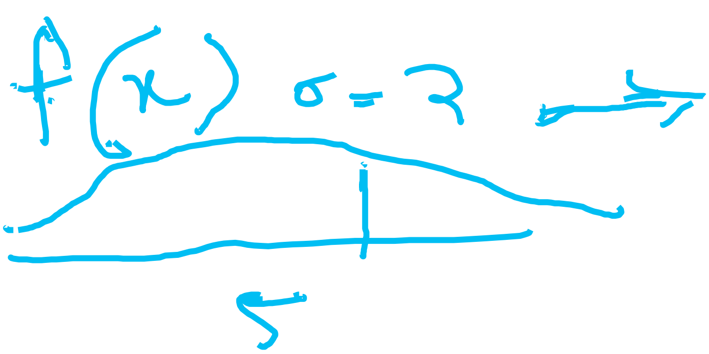
Standardize:

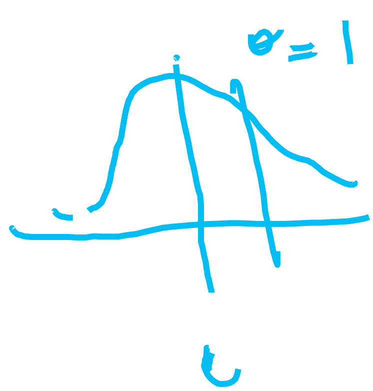
$$Z = \frac{X - \mu}{\sigma}$$

What is the mean and variance of Z?

1. $E(Z) = (1/\sigma)E(X - \mu) = 0$
2. $V(Z) = (1/\sigma^2)V(X) = 1$

$\frac{X - 5}{2}$





Normal Distribution - Calculating Probabilities

Return to cholesterol example (Rosner 5.20)

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

$$
\begin{aligned}
P(X < 200) &= P\left(\frac{X-\mu}{\sigma} < \frac{200-219}{50}\right) \\
&= P\left(Z < \frac{200-219}{50}\right) \\
&= P(Z < -0.38) \\
&= P(Z > 0.38) \quad \textit{from Table 3, column (b)} \\
&= 0.3520.
\end{aligned}
$$

Normal Approximation to Binomial

Example

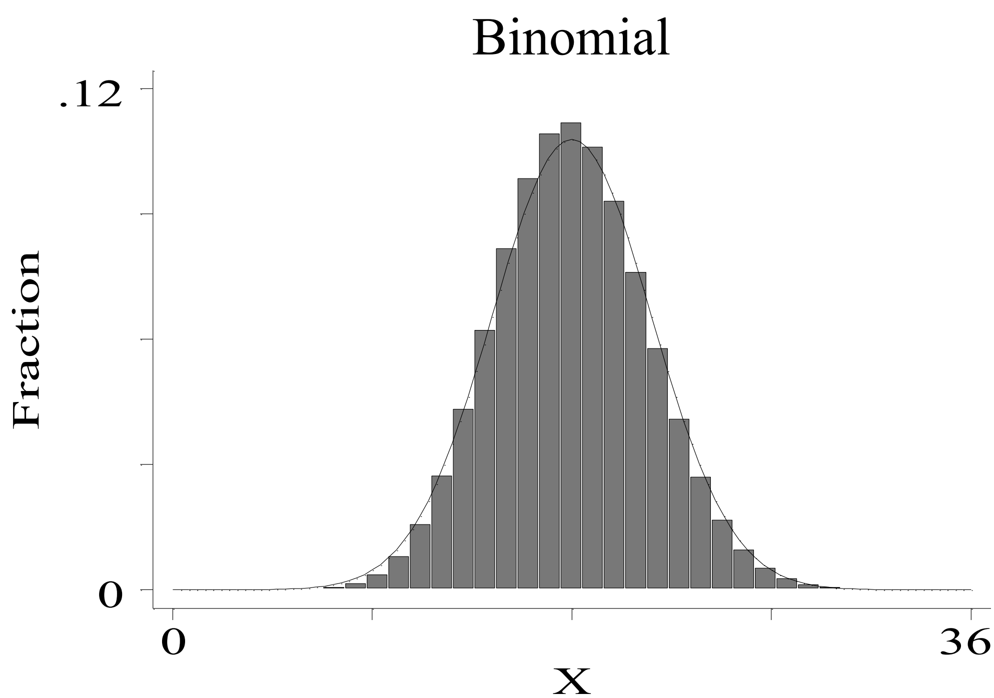
Suppose the prevalence of HPV in women 18 -22 years old is 0.30. What is the probability that in a sample of 60 women from this population 9 or fewer would be infected?

Random variable?

Distribution?

Parameter(s)?

Question?



graph X [weight=PX] if (X<37), hist bin(37) normal gap(3) yscale(0,.12)

Bin(60, .3)

Normal Approximation to Binomial

Binomial

- When **np(1-p)** is “large” the normal may be used to approximate the binomial.
- X ~ bin(n,p)

  E(X) = np

  V(X) = np(1-p)
- X is approximately N(np,np(1-p))

Normal Approximation to Binomial

<u>Example</u>

Suppose the prevalence of HPV in women 18 -22 years old is 0.30. What is the probability that in a sample of 60 women from this population that 9 or less would be infected?

Random variable?

$\Rightarrow$ X = number infected out of 60

Distribution?

$\Rightarrow$ Binomial

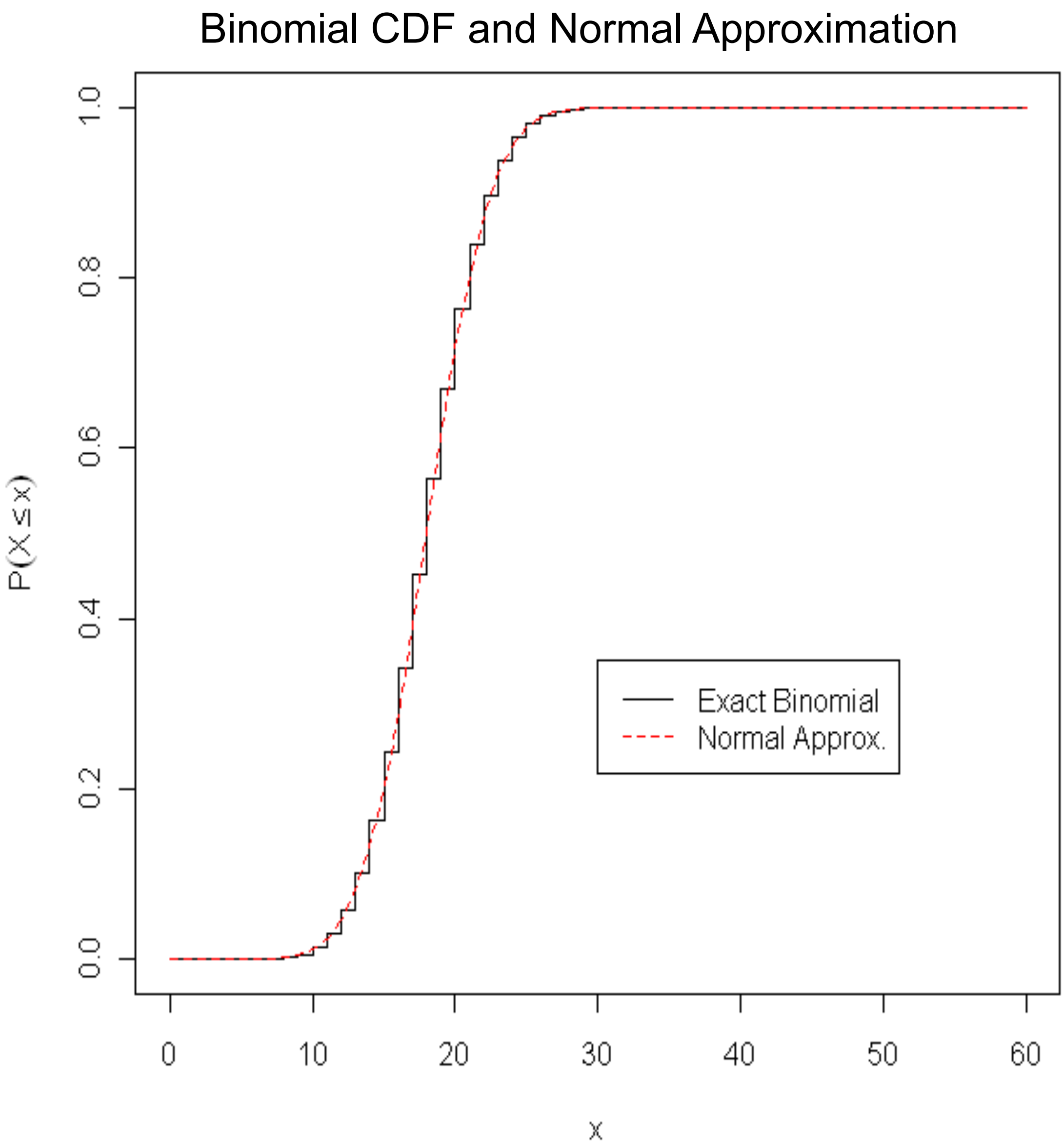
Parameter(s)?

$\Rightarrow n = 60, \; p = .30$

Question?

$\Rightarrow P(X \leq 9) =$

$\Rightarrow$ normal approx. =



$P(X \leq 9, n{=}60, p{=}.3) = 0.0059$

$P\left(\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{9-60*.3}{\sqrt{60*.3*.7}}\right) = P(Z \leq -2.535) = 0.0056.$

$$X \sim Bin(60, 0.3)$$

$$\text{so } \mu = np = 60 \times 0.3 = 18,$$

$$\sigma^2 = np(1-p) = 60 \times 0.3 \times 0.7 = 12.6$$

$$\rightarrow X \sim N(18, \sigma^2 = 12.6) \text{ approximately}$$

so

$$Pr(Binomial(60, 0.3) \leq 9) \approx Pr(Normal(18, 12.6) \leq 9.5)$$

$$= 0.0083 \text{ from website with normal CDF function}$$