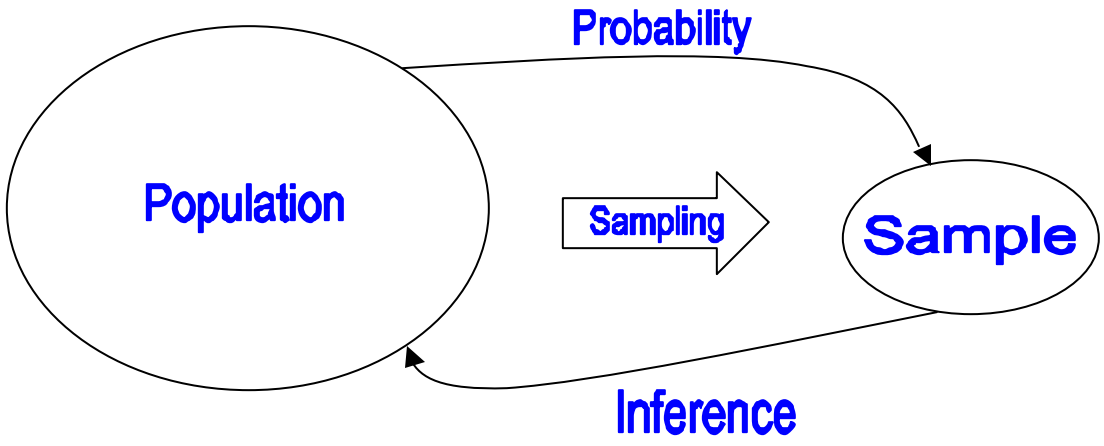
7. Theoretical Probability Distributions

First, a quick review.



Recall that a major objective of statistics is to make inference about a larger population from which our data (sample) was collected. Generally, in order to do this, you must make some assumptions about the population in order to effectively use your sample to infer something about the population. Luckily, many commonly observed random phenomenon are well described by relatively simple population models. These models are the assumptions which make inferences about the population easier (or even possible).

Some terminology

- A variable is used to describe any characteristic that can be measured or categorized
- A random variable is a variable whose value is governed by chance
- A discrete RV assumes only a finite or countable number of possible values--e.g. observe the blood type of a person randomly chosen from some population of individuals

- A continuous RV can take any value within an interval of values--e.g. observe cholesterol levels (μg/dl) from the population of women age 50 or older

7.1. Probability Distributions

We can describe a random variable by describing the probability with which it takes on possible values. That is, we can describe a random variable by describing its probability distribution--how likely certain values of the RV are to occur. In practice, most random variables can be described by some well known mathematical models for probability distributions. The nature of a probability distribution is different depending on whether it's a discrete RV distribution or a continuous RV distribution. In this chapter, we discuss 2 commonly used discrete distributions (binomial for a finite sum of binary (0-1) type of discrete data and Poisson for count data), and the normal distribution (very important continuous distribution)

- Usually use capital letters towards end of alphabet to denote random variable--e.g. X=number of individuals exposed to high levels of ozone per day.
- Usually use corresponding lower case letters to denote particular values of a random variable--e.g. X=x means the random variable takes on the value x.
- Note that X=x is an event, and we use our notation for probability presented in Chapter 6--e.g. P(X=2)=.25
- We've already seen probability distributions, but these we're empirical distributions (i.e., based on observations)--e.g. Table 7.1, page 163. We've worked with making statements of probability using these types of distributions in Chapter 6.

- The binomial, Poisson, and normal are theoretical distributions that serve to approximate (model) reality.

7.2. The binomial distribution

The binomial distribution is a discrete distribution. The assumptions underlying the use of the binomial distribution to model a random variable are:

- Fixed number of trials (i.e., "experimental runs", observations), n, are carried out. The outcome of each trial must be classified according to 1 of 2 mutually exclusive events (i.e., outcomes are binary or dichotomous)
  - standard terminology: "success" or "failure"
  - Bernoulli RV; often coded as Y = 1 for success, Y = 0 for failure
- The probability of success, p, remains constant from trial to trial (P("failure") = 1-p)
- Independent trials; the out come of any one trial does not affect the outcomes of any other trial

If we let X denote the number of successes in n trials, and let Y_i be the Bernoulli random variable corresponding to the ith trial with Y(success)=1 and Y(failure)=0. Then (assuming the above 3 bullets),

- X is called a binomial random variable with probability of success, p, and number of trials n: $X \sim Bin(n,p)$.
- Notice, with Y_i coded as 0 (for failure), and 1 (for success), then $X = \sum_{i=1}^{n} Y_i$.
- In short, a binomial random variable, X, is a sum of n independent Bernoulli random variables with constant success probability p.

E.g. Monitor air pollution levels in the LA basin for a one week period. Let X be the number of days out of 7 on which the concentration of ozone exceeds the maximum allowable levels set by the EPA's National Ambient Air Quality Standards (NAAQS).

- Is it reasonable to model X as a binomial random variable (check the assumptions)?

E.g. Childhood lead poisoning is a public health concern in most urban areas of the US. In a certain area, 1 in 10 children has high blood lead levels (≥ 30 μg/dl).

- In a random sample of 3 children from this area, do the number of children with high blood lead levels follow a binomial distribution?

- What's the probability that exactly one child has a high blood lead level?

- Exactly 2?

- Exactly 3?

We have a nice formula to figure these probabilities: Binomial Probability Model (discrete probability distribution or probability mass function)

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x \in \{0,1,2,\ldots,n\}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)\cdots(2)1}{x(x-1)\cdots(2)1(n-x)(n-x-1)\cdots(2)1}$$

Note: "!" is "factorial" and 0! is defined as 1

A heuristic derivation of this formula using the blood lead level example above:

- P(X=0):

- P(X=1):

- P(X=2):

- P(X=3):

Some connections to things we might already know:

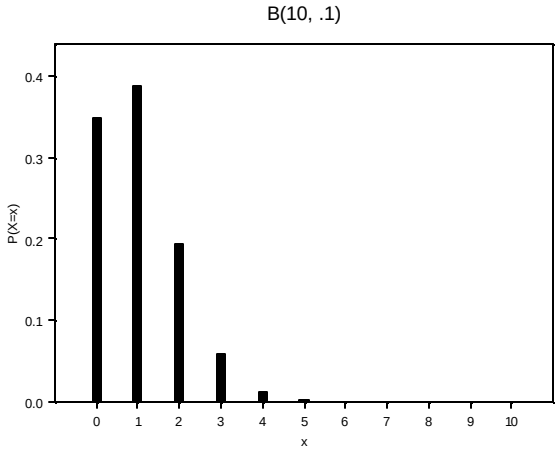
- Binomial Coefficient

$$(a+b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b^1 + \cdots + \binom{n}{n-1}a^1 b^{n-1} + \binom{n}{n}a^0 b^n$$
$$= \sum_{x=0}^{n}\binom{n}{x}a^{n-x}b^x$$

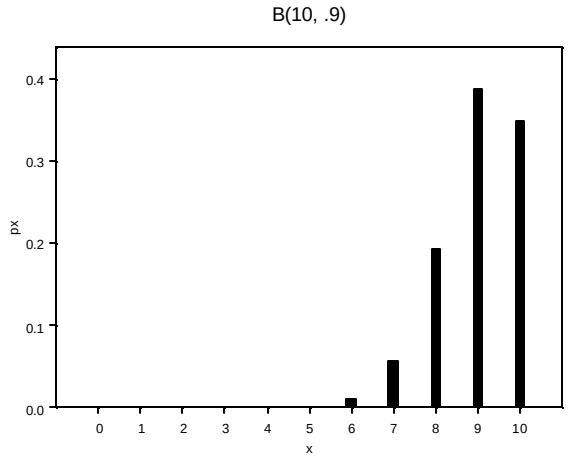
- Remember Pascal's Triangle?

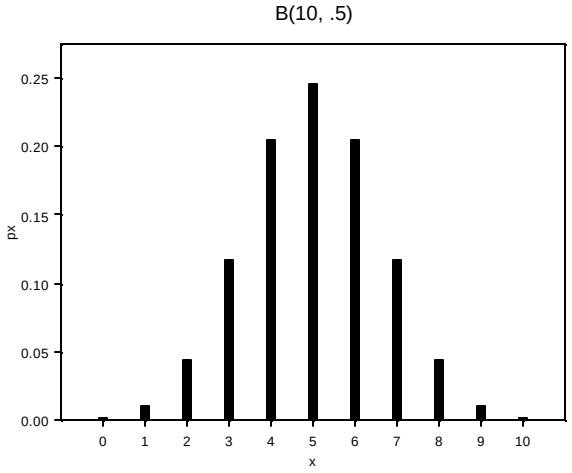
- How many ways to choose x objects from n, without replacement and without regard to the order in which the objects are selected (C_x^n)?

E.g. Blood lead levels continued. Here assume n=10 children p=0.1



Other p's:





Using Table A.1 (pp A-1 - A-5)

E.g. Blood lead levels (n=10, p=.1)

E.g. Using S-Plus functions pbinom(x,n,p) and dbinom(x,n,p)

Some Numerical Summaries of the Binomial Distribution

- Expectation or Expected Value or Population Mean
  - $E[X] = \sum_{x=0}^{n} xP(X = x) = np$ (proof?)
  - Looks like a weighted mean of x values with weight P(X=x)—something akin to grouped mean back in Chapter 3

E.g. Blood lead levels.

- Variance and Standard Deviation
  - $Var[X] = \sum_{x=0}^{n} (x - E[x])^2 P(X = x) = np(1-p)$
  - s.d.[X] =

E.g. Blood lead levels

7.3. Poisson

The Poisson is another discrete probability distribution. It is commonly used to model the probabilities of observing the a number (counts) of occurrences of an event within some interval (of time, space, area, volume, etc.). Assumptions underlying the Poisson

- Probability that a single event occurs within an interval in proportional to the length (area, volume) of that interval.
- An infinite number of occurrences of the event are theoretically possible (no fixed n like with binomial).
- Events occur independently both within the same interval and between intervals.

E.g. Number of patients visiting the emergency room within a 24 hour period. Assumptions?

E.g. Number of cells per square centimeter in a petri dish. Assumptions?

Poisson Probability Distribution Model

$$P(X = x) = \frac{e^{-l}\, l^{x}}{x!} \quad x \in \{0,1,2,\ldots\}$$

$$E[X] = \sum_{x=0}^{\infty} xP(X = x) = l$$

$$Var[X] = \sum_{x=0}^{\infty} (x - E[X])^2 P(X = x) = l$$

- Sometimes denoted $X \sim P(\lambda)$ or $X \sim Pois(\lambda)$
- λ is often called the rate parameter or intensity parameter—rate of occurrence of events per unit interval

- E.g. Suppose it is known that patients arrive a the hospital emergency room at rate of 12 patients per 24 hours. $\lambda=12$ patients per day or $\lambda=0.5$ patients per hour, with corresponding random variable X=number of patients observed in 24 hours in the first case, and X=number of patients observed in an hour in the latter.)
- P(X=0):

- P(X > 5):

Table A.2 (pp. A-6 – A-8)

Splus functions dpois() and ppois()

Poisson Approximation to the Binomial

- For n large and p small, the Poisson($\lambda=np$) distribution can be used to approximate binomial(n,p)

See plots p175 POB.

7.4. The Normal Distribution

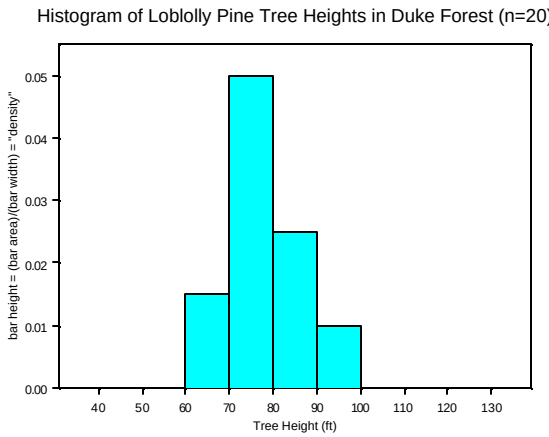
The normal distribution (also Gaussian or bell-curve) is definitely the most important distribution in statistics (see Chapter 8). Unlike the binomial and Poisson distributions (discrete), the normal is a continuous distribution (i.e., the random variable with a normal distribution is takes on values in $(-\infty,\infty)$. Because the binomial and Poisson random variables assumed a finite or countable number of possible values, it made sense to talk about the probability of observing a single numerical values, e.g. P(X=x). With continuous distributions, we talk of probabilities within intervals. For continuous distributions, P(X=x)=0. We use a function called a probability density function (pdf) for continuous distributions. Discrete models that give P(X=x) are sometimes called probability mass functions (pmf) since they actually give probability "masses" at a point. But, the pdf gives the "density" or "probability per x value". Since we're interested in probabilities of intervals, we "sum" (integrate) the area under the pdf between interval endpoints to get the probability.

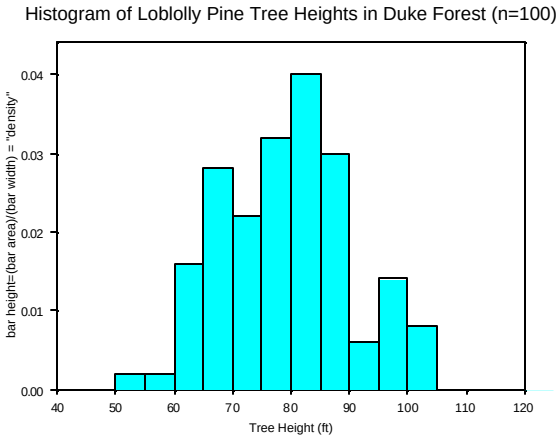
The Normal pdf

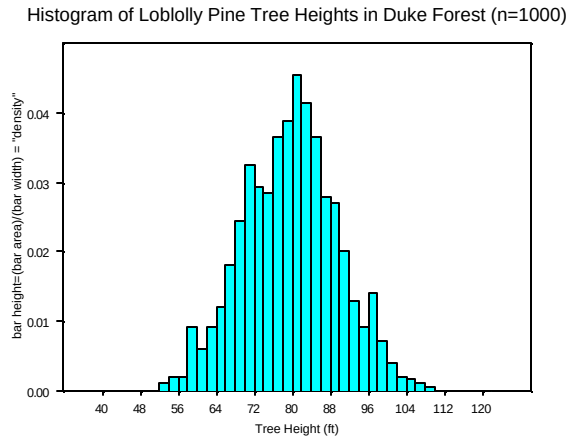
- $f(x) = \frac{1}{\sqrt{2p}\,s} e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2} \quad x \in (-\infty,\infty)$
- $E[X]=\mu$ (mean median and mode)—location parameter
- $Var[X]=\sigma^2$--dispersion or scale parameter σ
- Often denoted $X\sim N(m, s^2)$
- $P(X \le x) = \int_{-\infty}^{x} f(t)dt$

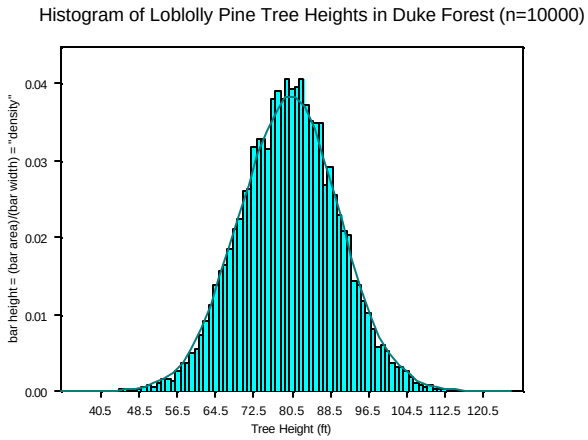
- Luckily, we don't have to do integration, it's already done for us in Table A.3 (p. A-9)—more later.
- Approximates binomial for large n
- Approximates Poisson for large λ

Relation of pdf to density histogram using $N(80,10^2)$:





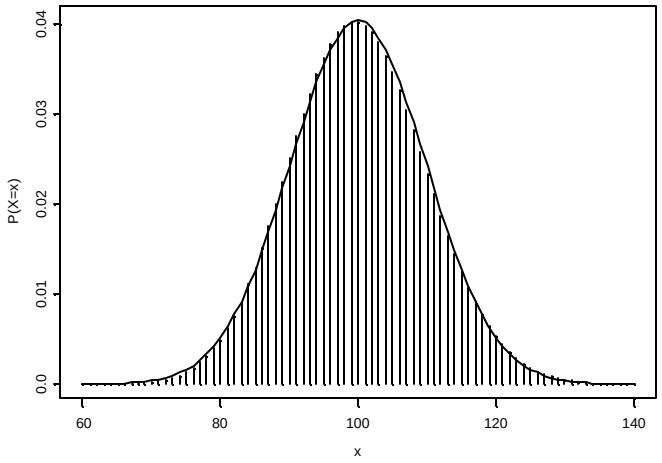




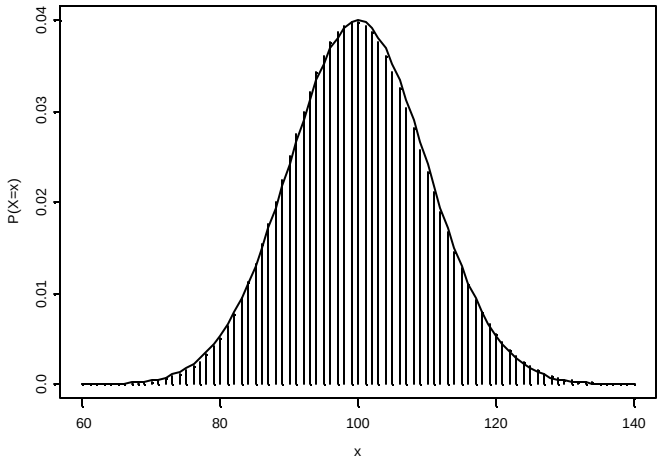
Properties of density curves

- always nonnegative
- total area under curve = 1
- area under curve between any two points on the horizontal axis is the proportion (probability) of values falling in that interval

Bin(5000,0.02) (vertical lines) with N(100, 9.899) pdf



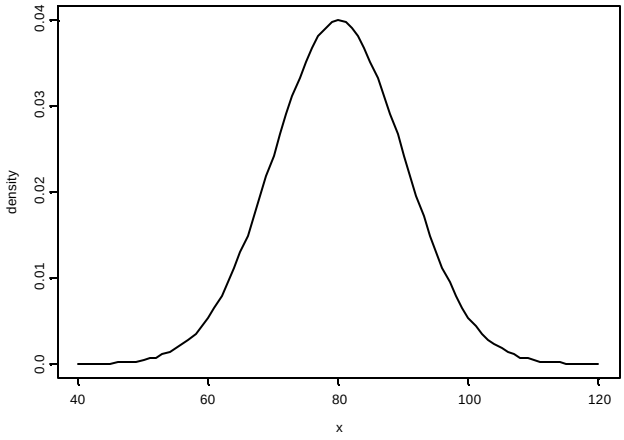
Pois(100) (vertical lines) overlaid with N(100,10)

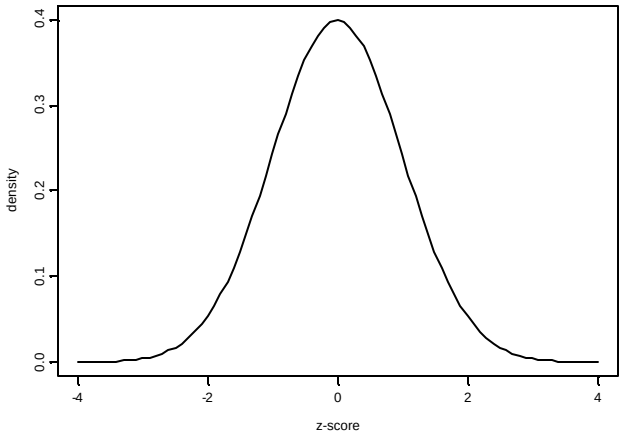


Calculating Normal Probabilities

- We get different normal pdfs for different mean, μ, and standard deviation, σ.
- Useful result
  - If $X \sim N(\mu, \sigma^2)$, then $Z=(X-\mu)/\sigma \sim N(0,1)$
  - Z is a standard normal RV and N(0,1) is the standard normal distribution
  - a realization (a particular value) of Z is denoted z, and is called a z-score, $z=(x-\mu)/\sigma$

Calculate some probabilities for $X \sim N(80,10^2)$ using Table A.3. Make sure you can do things like $P(X \leq x)$, $P(X \geq x)$, $P(x_1 \leq X \leq x_2)$, $P(Z \geq z)$, $P(Z \leq -z)$, $P(|Z| \geq z)$,.....





E.g. Suppose that diastolic blood pressure (X) in hypertensive women in centered about 100 mm Hg and has a standard deviation of 16 mm Hg and is normally distributed.

- For a randomly selected hypertensive woman, what is the probability that her diastolic BP is below 90 mm Hg? (Be systematic about these calculations. Write down all that you know, what you need, etc. Also, drawing a picture of the X pdf and Z pdf is not a bad idea, especially when you're first getting the hang of these sorts of calculations.)

- $P(X \geq 124)$?

- P(96£X£104)?

E.g. Remember the empirical rule for symmetric, unimodal distributions? Suppose $X \sim N(\mu, \sigma^2)$.

- $P(|X-\mu| \pounds \sigma) = P(\mu-\sigma \pounds X \pounds \mu+\sigma) =$
- $P(|X-\mu| \pounds 2\sigma) = P(\mu-2\sigma \pounds X \pounds \mu+2\sigma) =$
- $P(|X-\mu| \pounds 3\sigma) = P(\mu-3\sigma \pounds X \pounds \mu+3\sigma) =$

E.g. Using Table A.3 "in reverse": find the 90th, 95th, and 99th percentile (quantile) of a standard normal distribution (i.e., what are the z-scores corresponding to these percentiles?).