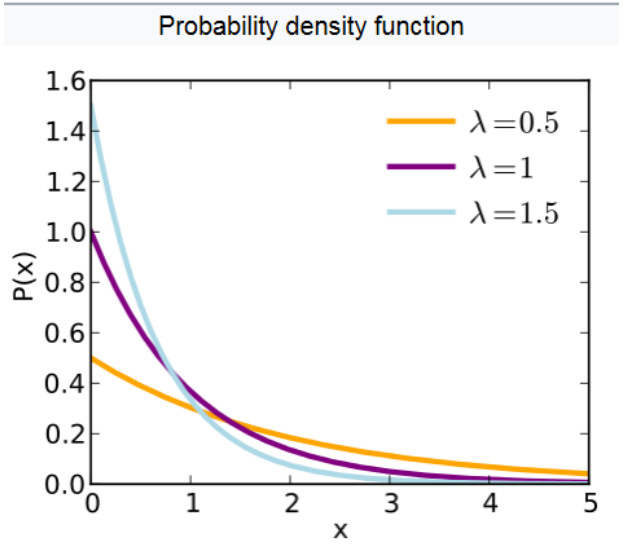
A few more notes on the CLT



Exponential distributions are moderately skewed

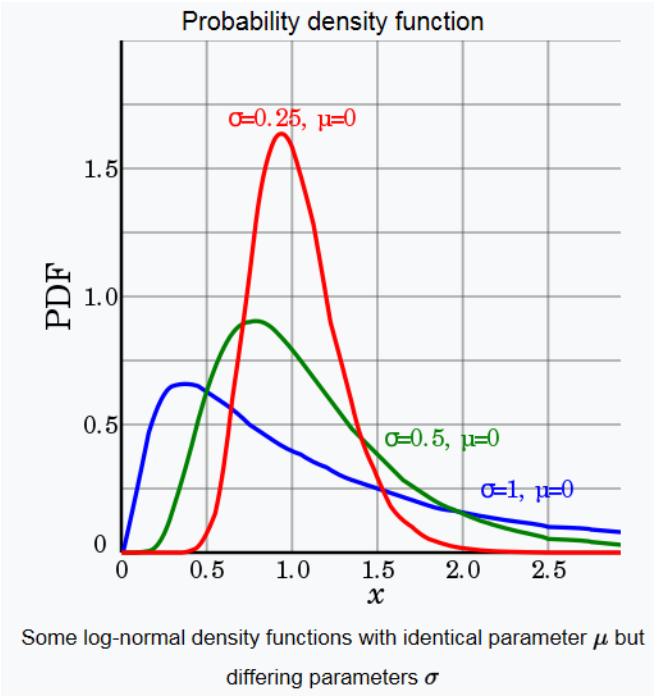
And every exponential distribution has the same skewness value of 2.

A distribution whose skewness varies with its variance is the lognormal distribution:

If $X \sim N(\mu, \sigma^2)$ then $Y = \exp(X) \sim \text{lognormal}(\mu, \sigma^2)$.

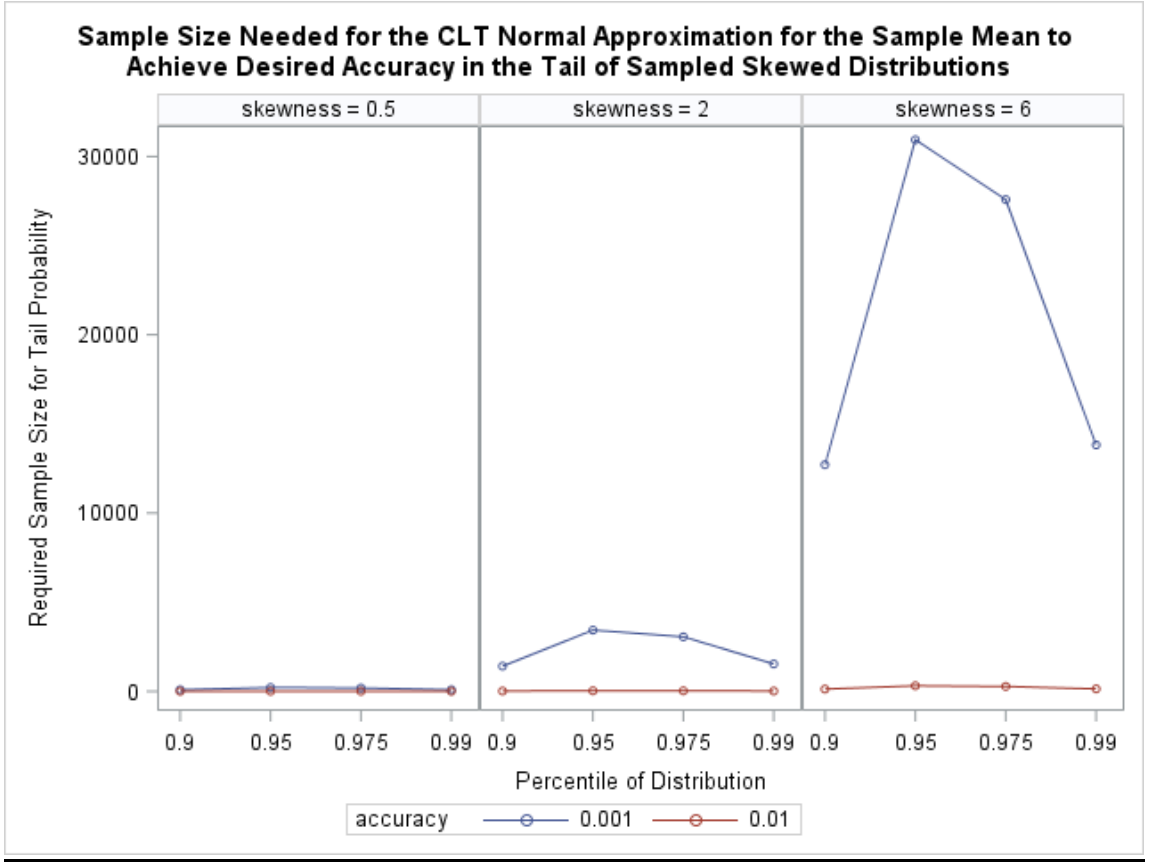
Skewness for a lognormal distribution is a function of its variance σ^2: $\sqrt{e^{\sigma^2}-1}(2+e^{\sigma^2})$

For a lognormal with $\mu=0$ and $\sigma=1$, skewness ≈ 6



Some log-normal density functions with identical parameter μ but differing parameters σ

Let's take a look at the graph below to see how greater skewness affects the sample size calculation we did last time …



So, when would we want a high level of accuracy in applying the CLT?

Compelling argument from Hesterberg (2008)

Confidence interval coverage at each extreme can be off by (much) more than 10% when the underlying distribution being sampled is highly skewed and the sample size is only slightly larger than 30, i.e.

$$P(\mu \leq \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) \text{ is not } \leq (+/-1.1)*\alpha/2$$

and/or

$$P(\mu \geq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) \text{ is not } \leq (+/-1.1)*\alpha/2$$

Bottom line: If we use the CLT to obtain normal or even t-distribution confidence intervals for population means (and other parameters) in order to draw conclusions when the distribution being sampled is very skewed, we can be wrong more often than we realize!

Lower bound of CI:

e.g. quality control: we want to be assured that *at least* a certain threshold is achieved (a minimum quantity exists)

Upper bound of CI:

e.g. ensure that a toxic substance *does not exceed* a maximum

Another note: Illustration of the mean as center of gravity of a distribution

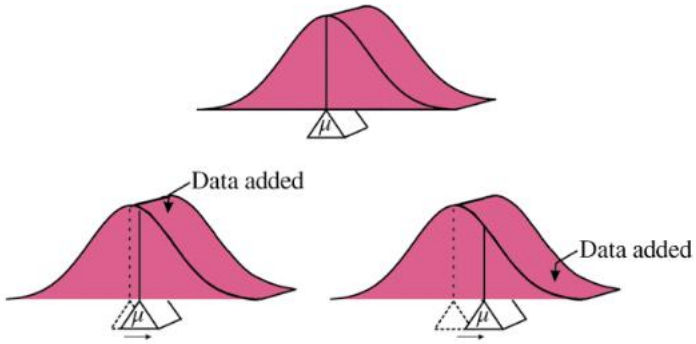


FIGURE 2.7 Illustration of the mean being the center of gravity of a distribution of data. Like a center of gravity, the position of the mean is affected more by data added farther from it than it is by data added closer to it.

From *Introduction to Biostatistical Applications*, R.P. Hirsch, Wiley, 2016

3. Random Variables and Distributions
Expected Value and Variance
Properties and Types of Estimators

Readings: Rosner: Ch. 4 and 5; Chihara and Hesterberg: Ch. 6
OpenIntro Statistics: 2.2, 2.5, 2.6

Homework: Homework 2 due by noon on September 17

Overview

A) Definitions and Notation
B) Discrete: Probability Mass Function
C) Discrete: Expected Value (Mean), Variance, SD
D) Continuous: Probability Density Function
E) Continuous: Expected Value (Mean), Variance, SD
F) Joint, Marginal, Conditional Distributions
G) Properties and types of estimators

A) Definitions and Notation

Random variables and probability distributions are the theoretical or mathematical representations of data values and frequency distributions.

Random variable (r.v.): Quantity that takes on different values or sets of values with various probabilities. A numerical function that assigns a number to each possible outcome (i.e. point in the sample space) of a random trial.
Convention: capital letters: X, Y, etc. for r.v.; small letters for values of r.v., *x*, *y*, etc.

Discrete random variable: can only take on a finite or countable number of values
e.g. number of children in family, number of cases of disease

Continuous random variable: can take on any value in an interval
e.g. height, weight, food intake

Probability Distribution: describes the probabilities for each outcome for a discrete random variable (probability mass function, *pmf*) or the probabilities of values in a range for a continuous random variable (probability density function, *pdf*).

Probability mass functions, probability density functions and cumulative distribution functions are useful as tools for describing the frequency distributions of random variables for the entire population of interest and as tools for providing probability statements about events involving random variables.

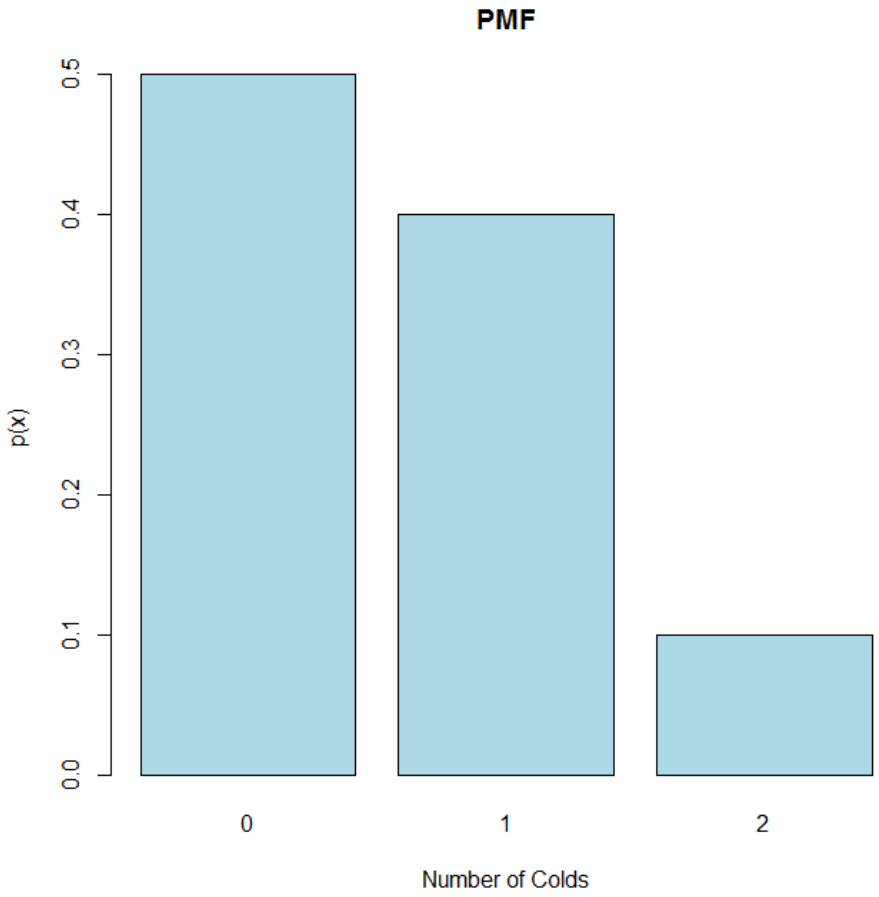
B) Discrete r.v.: Probability Mass Function

Let X be a discrete random variable and let *x* represent the values that X can take on.

Probability distribution of X is p(*x*) = P(X = *x*)
e.g. X = number of colds in a year caught by healthy adult = 0, 1, 2

Number of Colds:	0	1	2
P(X=*x*)	0.5	0.4	0.1

Note: All probabilities are nonnegative and the sum over all mutually exclusive and exhaustive values of the r.v. X is 1.

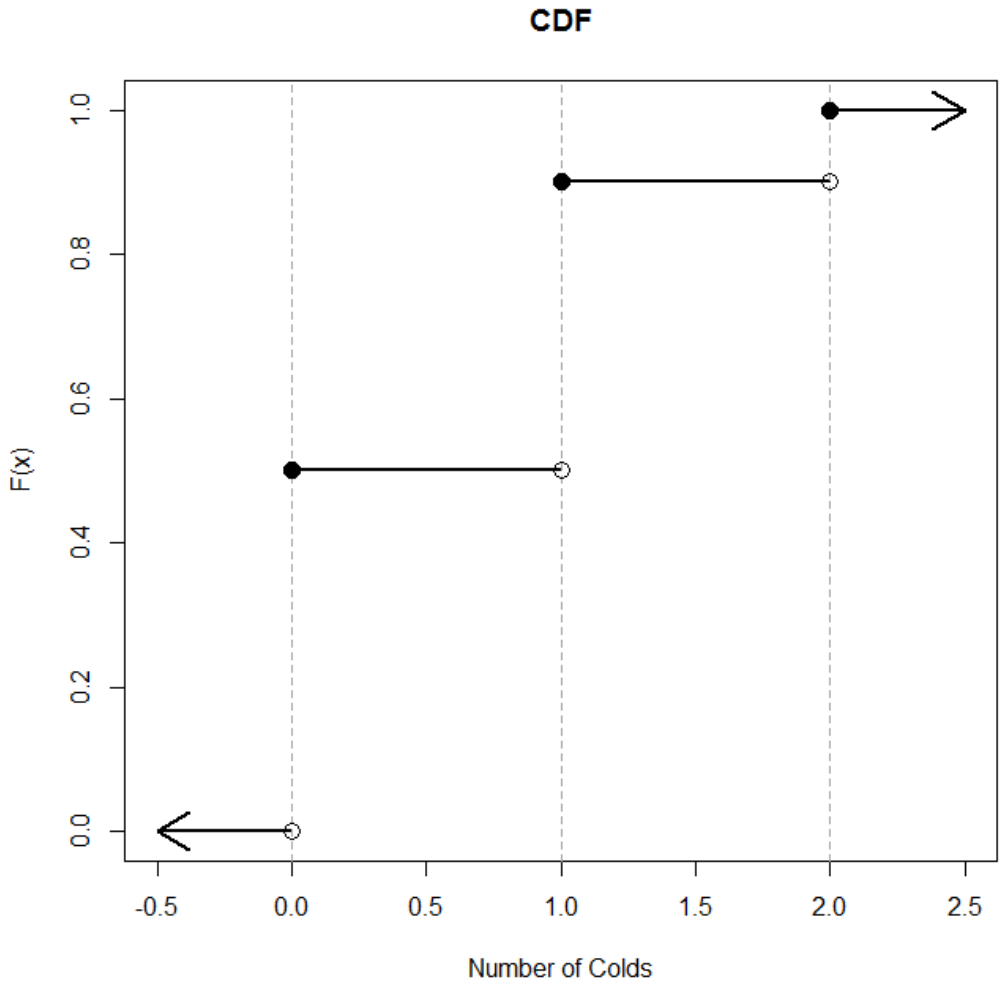


Cumulative Distribution Function (CDF): cumulative probability distribution of X

$F_x(x) = P(X \leq x) = \sum_{x=0}^{k} P(X = x)$ (accumulate probabilities from lowest to highest)

monotone ↑, $F(-\infty) = 0$, $F(\infty) = 1$

F(0) = 0.5 (number of colds <= 0)
F(1) = 0.5 + 0.4 (number of colds <=1)
F(2) = 0.5 + 0.4 + 0.1 = 1.0

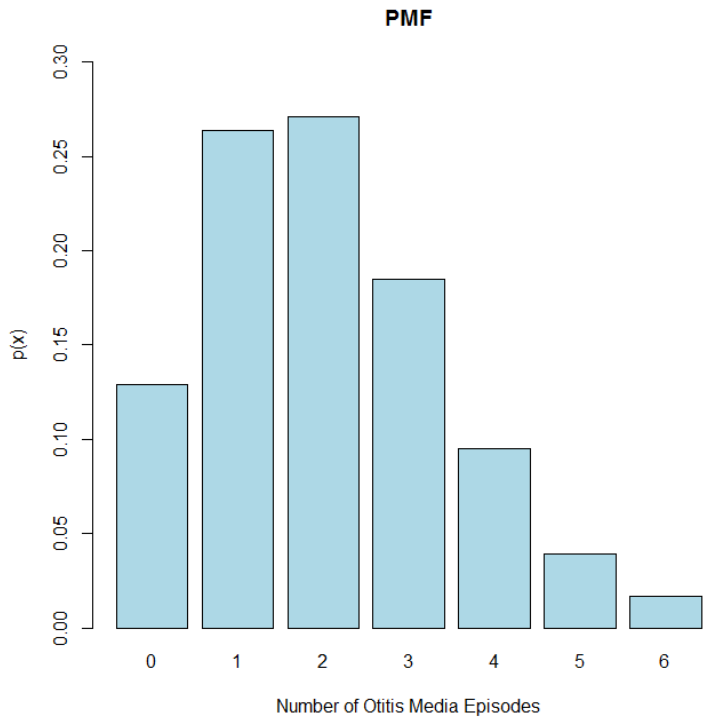


e.g. X = number of episodes of otitis media (disease of middle ear) in first 2 years of life:

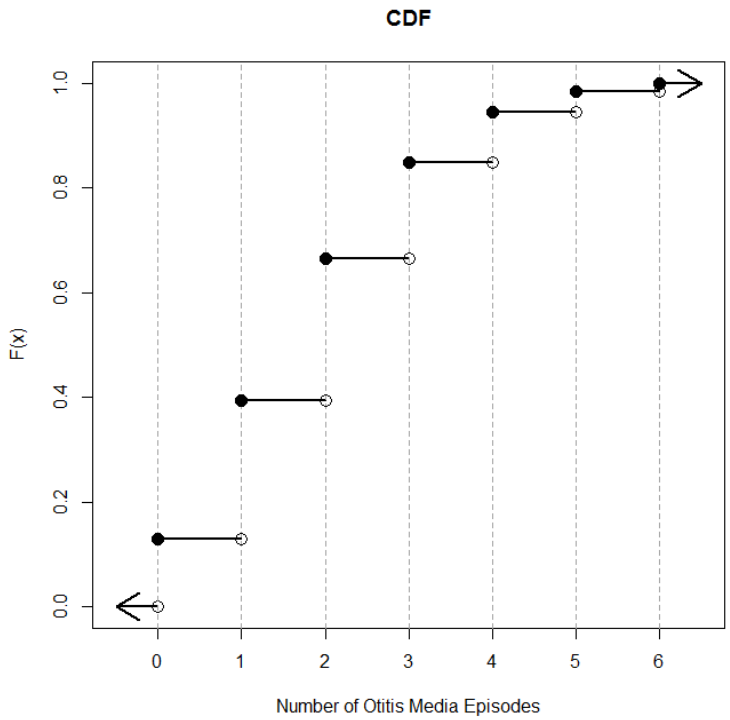
x	0	1	2	3	4	5	6
P(X=x)	0.129	0.264	0.271	0.185	0.095	0.039	0.017

Check: $\Sigma\, p_i = 1$

PMF: P(X = x)



CDF: P(X $\leq$ x)



C) Discrete: Expected Value (Mean), Variance, SD

There are summary values for random variables just like those for a sample of data. For a discrete r.v. X:

Expected Value (Expectation, Mean)

$$E(X) = \sum_{possible\ x} x\,P(X=x) = \mu$$

$$= \sum_{x} x\,p(x) = \mu$$

x represents possible values

$P(X=x)$, $p(x)$ represent the weight (probability) for a given x

$\boldsymbol{E(X)}$ is the balance point (center of gravity in physics) on the graph

e.g. X = number of colds in a year caught by healthy adult = 0, 1, 2

Number of Colds:	0	1	2
P(X=x)	0.5	0.4	0.1

$\mu = E(X) = 0(0.5) + 1(0.4) + 2(0.1) = 0.6$ colds per year

Variance

$\text{Var}(X) = V(X) = \sigma^2 = E[(X - E(X))^2] = E[(X - \mu)^2]$

For discrete variables:

$$V(X) = \sum_{possible\ x} (x-\mu)^2\, P(X=x) = \sigma^2$$
$$= \sum_{x} (x-\mu)^2\, p(x) = \sigma^2$$

$$SD(X) = \sqrt{\sum_{possible\ x} (x-\mu)^2\, P(X=x)} = \sqrt{V(X)} = \sigma$$

Number of Colds:	0	1	2		Recall:
P(X=*x*)	0.5	0.4	0.1		E(X)=0.6

$V(X) = \sigma^2 = (0 - 0.6)^2(0.5) + (1 - 0.6)^2 (0.4) + (2 - 0.6)^2 (0.1) = 0.44$

$SD(X) = \sigma = \sqrt{0.44} = 0.66$ colds per year

Computational formula: $V(X) = E[X^2] - (E[X])^2$

$E(X^2) = 0^2 (0.5) + 1^2(0.4) + 2^2 (0.1) = 0.8$

$V(X) = 0.8 - (0.6)^2 = 0.44$

D) Continuous r.v.: the Probability Density Function (pdf)

For a random variable measured on a continuous scale, there are an infinite number of possible values between any 2 adjacent points on the scale. Thus the probability of observing any specific value on the scale is 0, or P(X = *x*) = 0.

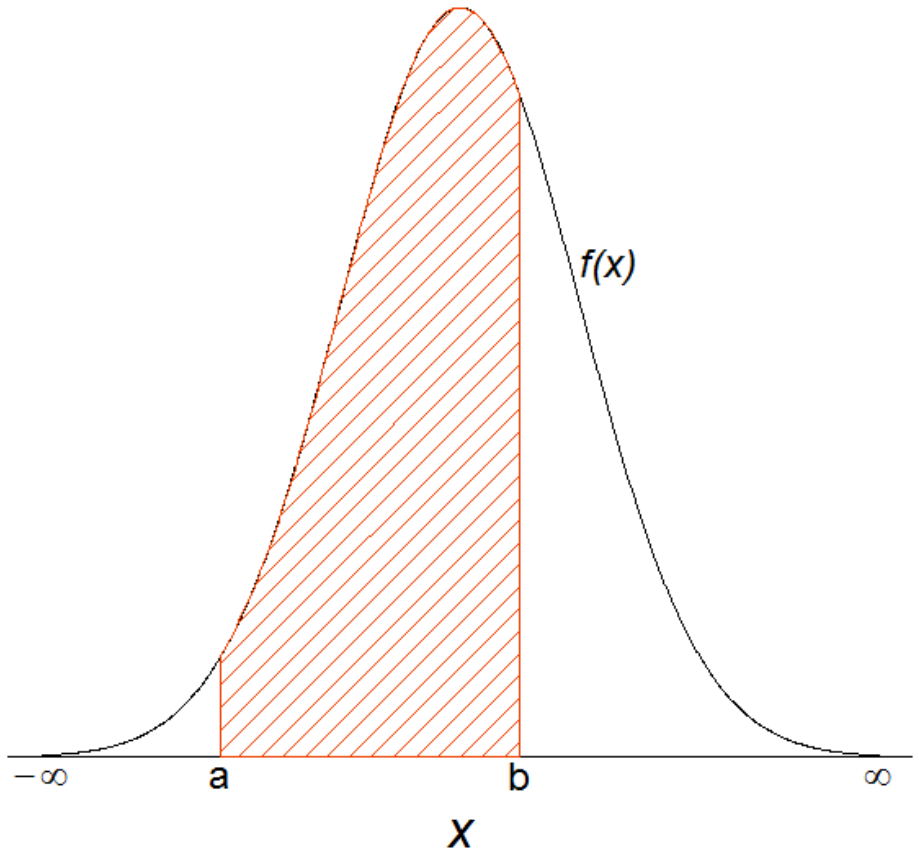
To work with probabilities for continuous r.v., we must instead consider an interval of values on the scale, e.g. (a, b). The idea of a probability mass function does not apply to continuous scale r.v. Instead we denote the function that assigns probability to values of the r.v. as *f(x)dx*. This is called the *probability density function* (pdf).

P(a $\leq$ X $\leq$ b) = area under the pdf, *f(x)dx*, from a to b = $\int_a^b f(x)dx$

Note: f(*x*) $\geq$ 0 for all *x,* and $\int_{-\infty}^{\infty} f(x)dx = 1$

The pdf does not give probability values but, rather, tells what *set* of values is most likely:

$$P(a \leq X \leq b) = P(X \leq b) - P(X < a) = F(b) - F(a)$$



Cumulative Distribution Function: CDF

For a continuous r.v.: $\int_{-\infty}^{a} f(x)dx$ is the area under the curve from -∞ to *a*. We call this the *cumulative distribution function* (cdf) of X evaluated at the value *a*, F(*a*).

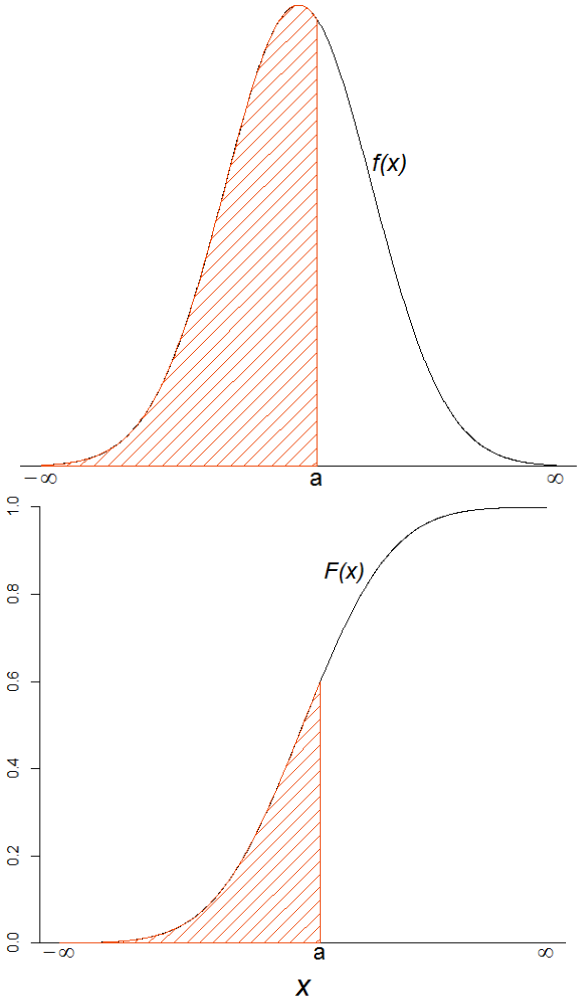
The cdf is a monotone increasing function.

F(-∞) = 0

F(+∞) = 1

CDF and PDF relationship: $f(x) = \frac{dF(x)}{dx}$

Every time we execute a statistical test or determine a p-value (level of significance), we will be using the cdf of a relevant continuous or discrete random variable.



E) Continuous: Expected Value (Mean), Variance, SD

Expected Value: Observations weighted by density function over the range of X:

$$\boldsymbol{E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{all\ x} xf(x)dx = \mu}$$

Variance:

$$\boldsymbol{V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{all\ x} (x-\mu)^2 f(x)dx = \sigma^2}$$

Standard Deviation:

$$\boldsymbol{s.d.(X) = \sqrt{V(X)} = \sigma}$$

F) Joint, Marginal and Conditional Distributions of r.v.

A respiratory disease and smoking example:

	Non-Smoker X = 0	Smoker X = 1	Total
No respiration problem Y = 0	.50	.30	.80
Respiration problem Y = 1	.05	.15	.20
Total	.55	.45	1.00

This is a distribution defined by two r.v. – i.e. it's a *bivariate* distribution:
X = 1 if smoker, 0 if non-smoker
Y = 1 if respiratory problems, 0 if not

Joint distribution: $P(X = x \text{ and } Y = y) = P(X = x \cap Y = y)$

$P(X = 0 \cap Y = 1)$ = P (non-smoker and respiration problem) = 0.05

Marginal distributions: P(X = *x*); P(Y = *y*)

	Non-Smoker X = 0	Smoker X = 1	Total
No respiration problem Y = 0	.50	.30	.80
Respiration problem Y = 1	.05	.15	.20
Total	.55	.45	1.00

P(X = 0) = P(non-smoker) = 0.55

P(Y = 0) = P(no respiration prob) = 0.80

Conditional distributions:

$$P(X = x \mid Y = y) = \frac{P(X=x \cap Y=y)}{P(Y=y)}$$

$$P(Y = 1 \mid X = 0) = P(\text{resp prob given non-smoker}) = \frac{P(X=0 \cap Y=1)}{P(X=0)} = \frac{0.05}{0.55} = \mathbf{0.091}$$

$$P(Y = 1 \mid X = 1) = P(\text{resp prob given smoker}) = \frac{P(X=1 \cap Y=1)}{P(X=1)} = \frac{0.15}{0.45} = \mathbf{0.333}$$

Independence of two r.v.: X and Y are independent *iff* (if and only if)

$P(X = x \cap Y = y) = P(X = x) \times P(Y = y)$, or

$P(Y = y \mid X = x) = P(Y = y)$

$P(X = 1 \cap Y = 1) = 0.15$

$P(X = 1) \times P(Y = 1) = (0.45) \times (0.20) = 0.09$

Therefore, X and Y are *not* independent.

Roughly, two random variables are independent if knowing the value of one does not change the probability distribution of the other. Which sequences $\{X_1, X_2, \ldots, X_n\}$ are independent?

1. X_i = high temperature in Denver on day *i*

2. X_i = color of car *i* in a row of parked cars

3. X_i = 1 if has flu, 0 if not for people working in an office building

4. X_i = religion of person *i*, where people are selected randomly from a phone book?

5. X_i = religion of person *i*, where people are selected in alphabetical order from a phone book?

G. Properties and Types of Estimators

Properties

Unbiasedness – not sample size dependent
e.g. E(X) = population mean μ, E($\bar{X}$) = μ, regardless of the sample used to obtain $\bar{X}$

Example: Let X_1, X_2, X_3 be a random sample of size 3 from any distribution with mean parameter μ. X – heights of sons whose fathers are over 5'10"

Estimator 1 = X_1, Estimator 2 =($X_1 + X_2$)/2, Estimator 3 = ($X_1 + 2X_2$)/3.

Are these estimators unbiased? (fill-in during class …)

Consistency – sample size dependent unbiasedness, an estimator converges *in probability* to the true population parameter – asymptotic result, i.e. approaches true population parameter as sample size gets large

Median consistent estimator of the mean (for symmetric distributions) – Homework 2

Mean Square Error = Bias^2 + Variance – good for comparing biased estimators to each other, tradeoff can be important

Efficiency – variability (and power; more on power later)

Example: Let X_1, X_2, X_3 be a random sample of size 3 from any distribution with variance parameter σ^2. X – heights of sons whose fathers are over 5'10"

Estimator 1 = X_1, Estimator 2 =$(X_1 + X_2)/2$, Estimator 3 = $(X_1 + 2X_2)/3$.

Which estimator is most efficient? (i.e. has the smallest variance) (fill-in during class …)

Types of Estimators

Population parameters – mean, variance, proportions, etc.

Method of Moments (MoM) Estimators – based on powers of X_i

e.g. Sample mean is a function of X_i^1: $\bar{X} = \sum_{i=1}^{n} \frac{X_i^1}{n}$

Sample variance is a function of X_i^2: $s^2 = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{n}$

Sample skewness is a function of X_i^3: $\sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^3}{n}$

etc.

Maximum Likelihood Estimators

If $X_1, X_2, \ldots, X_n$ follow the same distribution $f_X(x; \theta)$, then the likelihood function *L* for a sample of n independent and identically distributed observations, $x_1, x_2, \ldots, x_n$: $L \propto \prod_{i=1}^{n} f_X(x_i; \theta)$, where θ is a population parameter(s) that define the distribution.

By maximizing the function L with respect to θ (Steps: Take first derivative with respect to θ and set equal to 0; Solve using numerical methods, sometimes closed form is possible; Check solution is a maximum by taking second derivative with respect to θ) …we obtain $\hat{\theta}$ ("theta-hat"), the value of the population parameter that makes the data most likely to have been observed. $\hat{\theta}$ is known as the maximum likelihood estimator (MLE) of θ.

Example: Sample of size n, X – height of U.S. women over age 20, mean μ, variance σ^2

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}; \ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \hat{\mu})^2}{n}$$

Notes:

- MLE are sometimes not unbiased but they are usually consistent and they converge to the true population parameter faster than MoM estimators
- MLE often have the smallest variance compared with other estimators, like MoM estimators

Estimators for Regression Models

Ordinary Least Squares (OLS) for linear models – identical to MLE (will see later – Lecture 22)

For regression models that are not linear, MLE and weighted variants of OLS tend to have best properties