Essential Mathematics for the Political and Social Research

JEFF GILL

Cambridge University Press

Lecture Slides, Chapter 8: Random Variables

Objectives

- This chapter describes the means by which we label and treat known and unknown values.
- We first talk here about the levels of measurement for observed values where the primary distinction is discrete versus continuous.
- Then we see that the probability functions used to describe the distribution of such variables preserves this distinction. Many of the topics here lead to the use of statistical analysis in the social sciences.

Two General Levels of Measurement

- *Discrete data* takes on a set of categorical values, and *continuous data* takes on values over the real number line (or some bounded subset of it).
- While discreteness requires countability, it can be infinitely countable, such as the set of positive integers.
- A continuous random variable takes on uncountably infinite values, even if only in some range of the real number line, like $[0:1]$, because any interval of the real line, finitely bounded or otherwise, contains an infinite number of rational and irrational numbers.
- To see why this is an uncountably infinite set, consider any two points on the real number line: it is always possible to find a third point between them.

Four Specific Levels of Measurement

- It is customary to divide levels of measurement into four types, the first two of which are discrete and the second two of which are either continuous or discrete.
- *Nominal* data are purely categorical in that there is no logical way to order a set of events. The classic example is religions of the world.
- *Ordinal* data are categorical (discrete) like nominal data, but with the key distinction that they can be ranked (i.e., ordered).
- *Interval* data are ordered but the spacing between categories is identical (like the positive integers).
- *Ratio* data are interval data where zero is a meaningful value so that ratios make sense.

Interval Data Notes

- Interval data can be discrete or continuous, but if they are measured on the real number line, they are continuous.
- Examples of interval measured data include
  - temperature measured in Fahrenheit or Celsius;
  - a “feeling thermometer” from 0 to 100 that measures how survey respondents feel about political figures;
  - size of legislature (it does not exist when $n = 0$);
  - time in years (0 AD is a construct).

Ratio data Notes

- Ratio measurement is useful because it allows direct *ratio* comparison.
- 10 Kelvin is twice as “warm” as 5 Kelvin.
- Other examples include
    - appropriations
    - votes
    - crime rates
    - war deaths
    - unemployment
    - group membership.

Distribution Functions

- Distribution functions are mathematical functions that describe the uncertainty of some key variable of interest in a formal way.
- We want some tool that says this variable “lives around here” and is widely or narrowly variable.
- Distribution functions are central in statistical and probabilistic analyses.
- These provide a description of how we believe that some data-generating process is operating.
- Thus they are central to *modeling*.

What Does *Random* Really Mean?

- Everyone is accustomed to the idea that some events are more likely to occur than others.
- Random simply means that a future outcome is unknown, but some possible outcomes are more likely than others.
- The key idea here is the expression of *relative difference* between the likelihood of events.
- *Probability* formalizes this notion by standardizing such comparisons to exist between zero and one inclusive, where zero means the event will absolutely not occur and one means that the event will certainly occur.
- Every other assignment of probability represents some measure of uncertainty and is between these two extreme values, where higher values imply a greater likelihood of occurrence.

Example: Measuring Ideology in the American Electorate

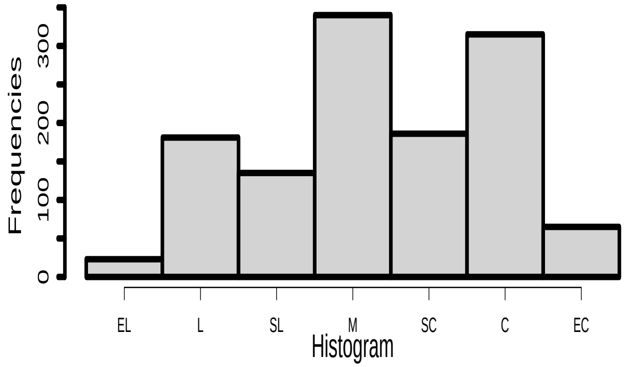
.

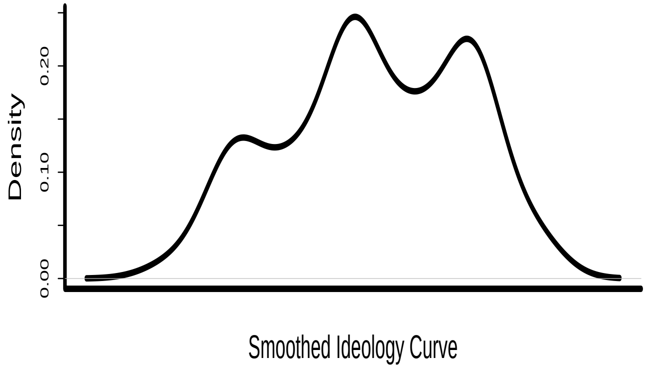
- Consider a question from the 2002 American National Election Study that asks respondents to identify their ideology on a seven-point scale that covers extremely liberal, liberal, slightly liberal, moderate, slightly conservative, conservative, and extremely conservative.

- 1245 in the survey placed themselves on this scale (or a similar one that was merged), and we will assume that it can be treated as an interval measure.

- The histogram shows ideology placements in the first panel, demonstrating the multimodality of the ideology placements with three modes at liberal, moderate, and conservative.

Example: Measuring Ideology in the American Electorate





Example: Measuring Ideology in the American Electorate

.

- The second panel is a "smoothed" version of the histogram, called a *density plot*.
- The y-axis is now rescaled because the area under this curve is normalized to integrate to one.
- The point of this density plot is to estimate an underlying probability structure that supposedly governs the placement of ideology.
- Thus we can consider this to be an empirically-driven version of a distribution function for ideology.

What Do You Mean There Are Two Interpretations of Probability?

- “Frequentists,” believe that probability statements constitute a long-run likelihood of occurrence for specific events.
- They believe that these are objective, permanent statements about the likelihood of certain events relative to the likelihood of other events.
- “Bayesians” or “subjectivists,” believe that all probability statements are inherently subjective in the sense that they constitute a certain “degree of belief” on the part of the person making the probability statement.
- What version do you think works best in the social sciences?

Back To Randomness and Variables

- Randomness means is that the outcome of some experiment (broadly defined) is not *deterministic*: guaranteed to produce a certain outcome.
- As soon as the probability for some described event slips below one or above zero, it is a random occurrence.
- If the probability of getting a jackpot on some slot machine is 0.001 for a given pull of the handle, then it is still a random event!
- Random variables describe unoccurred events abstractly for pedagogical purposes since it is often convenient to describe the results of an experiment *before it has actually occurred.*

Formal Definition

- A random variable, denoted with a capital Latin letter such as X or Y, is a function that maps the sample space on which it is "created" to a subset of the real number line, including possibly the whole real number line itself.
- There is a sample space that corresponds not to the physical experiment performed but to the possible outcomes of the random variable itself.
- Given an experiment flipping a coin 10 times ($n = 10$):
  - The random variable X is defined to be the number of heads in these 10 tosses.
  - The sample space of a single iteration of the experiment is $\{H, T\}$, and the sample space of X is $\{0, 1, 2, \ldots, 10\}$.
- Random variables provide the connection between events and probabilities because they give the abstraction necessary for talking about uncertain and unobserved events.

More On Coin Flipping

- the probability that X takes on each possible value in the sample space determined by 10 flips of a fair coin:

X	0	1	2	3	4	5	6	7	8	9	10
$p(X)$	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010

- Each possible event for the random variable X, from 0 heads to 10 heads, is paired with a specific probability value.

Probability Mass Functions

- When the state space is discrete, we can assign probability values to each single event.
- This is true even if the state space is countably infinite (discrete with an infinite number of distinct outcomes).
- For example, in the case of flipping a possibly unfair coin, we can assign a probability to heads, $p(H)$, and therefore a complementary probability to tails, $p(T)$.
- A*probability mass function* assigns probabilities to each unique event, such that the Kolmogorov Axioms still apply.
- It is common to abbreviate the expression "probability mass function" with "PMF" as a shorthand.
- We denote such PMF statements:

  $$f(x) = p(X = x),$$

  meaning that the PMF $f(x)$ is a function which assigns a probability for the random variable X equaling the specific numerical outcome x.
- This notation often confuses people on introduction because of the capital and lower case notation for the same letter.

Bernoulli Trials

- The previous coin-flipping is actually much more general than it first appears.
- Suppose we are studying various political or social phenomenon such as whether a coup occurs, whether someone votes, cabinet dissolution or continuation, whether a new person joins some social group, if a bill passes or fails, etc.
- These can all be modeled as *Bernoulli outcomes* whereby the occurrence of the event is assigned the value "1," denoting success, and the nonoccurrence of the event is assigned the value "0," denoting failure.
- The value one occurs with probability p and the value zero occurs with probability $1 - p$.
- These outcomes form a partition of the sample space and are complementary.
- If x denotes the occurrence of the event of interest, then:

$$p(x) = p \qquad \text{and} \qquad p(x^{\complement}) = 1 - p.$$

- If we flip a coin 10 times and get 7 heads, then a reasonable estimate of p is 0.7.

Binomial Experiments

- The *binomial PMF* is an extension to the Bernoulli PMF whereby with multiple Bernoulli trials.
- This is historically called an experiment because it was originally applied to controlled tests.
- The random variable is no longer binary but instead is the sum of the observed binary events and is thus a count:

$$Y = \sum_{i=1}^{n} X_i.$$

- A complication to this Bernoulli extension is figuring out how to keep track of all of the possible sets of results leading to a particular sum.

Binomial Experiments

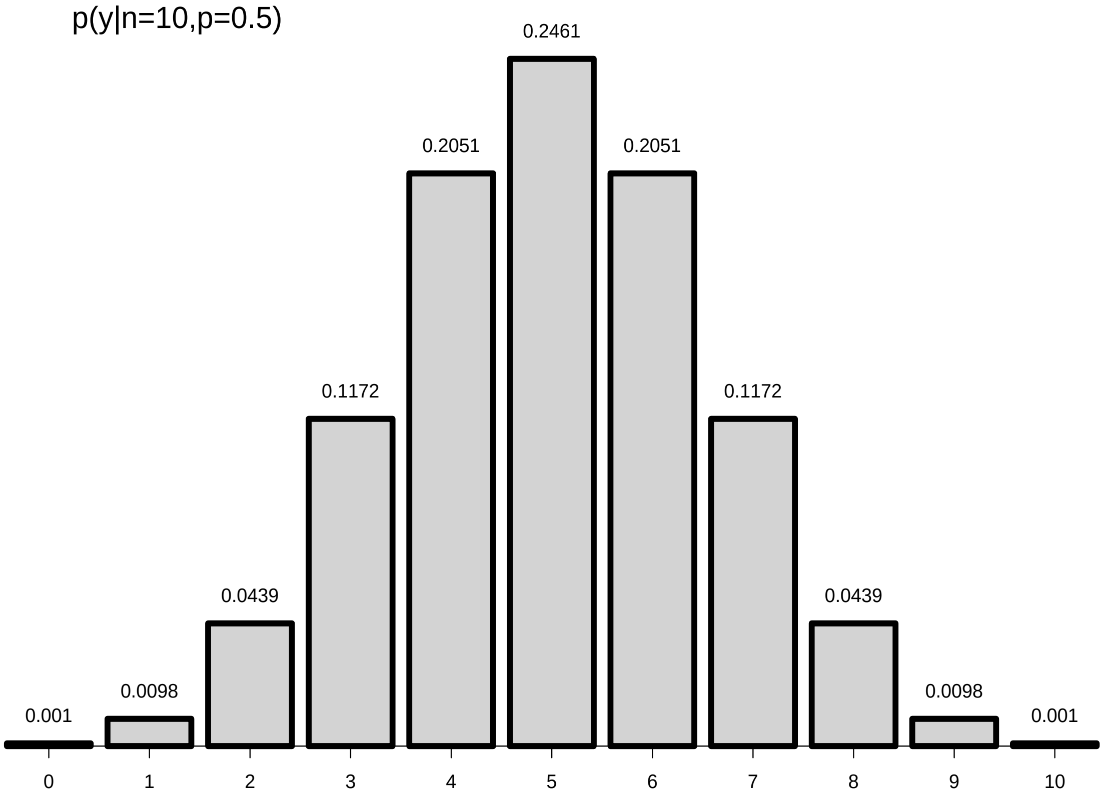
- Suppose we are studying three senior legislators who may or may not be retiring at the end of the current term.
- We believe that there is an underlying shared probability p governing their independent decisions and denote the event of retiring with R.
- We thus have a number of events E dictated by the three individual values and their ordering, which produce a sum bounded by zero (no retirements) and three (all retirements).
- The universe of events is summarized by:

E	Y	$p(Y=y)$	
$\{R^{\complement}, R^{\complement}, R^{\complement}\}$	0	$(1-p)(1-p)(1-p)$	$p^0(1-p)^{3-0}$
$\{R, R^{\complement}, R^{\complement}\}$	1	$(p)(1-p)(1-p)$	$p^1(1-p)^{3-1}$
$\{R^{\complement}, R, R^{\complement}\}$	1	$(1-p)(p)(1-p)$	$p^1(1-p)^{3-1}$
$\{R^{\complement}, R^{\complement}, R\}$	1	$(1-p)(1-p)(p)$	$p^1(1-p)^{3-1}$
$\{R, R, R^{\complement}\}$	2	$(p)(p)(1-p)$	$p^2(1-p)^{3-2}$
$\{R, R^{\complement}, R\}$	2	$(p)(1-p)(p)$	$p^2(1-p)^{3-2}$
$\{R^{\complement}, R, R\}$	2	$(1-p)(p)(p)$	$p^2(1-p)^{3-2}$
$\{R, R, R\}$	3	$(p)(p)(p)$	$p^3(1-p)^{3-3}$

Binomial Experiments

- The third column of the table gives the probabilities for each of these events, which is rewritten in the fourth column to show the structure relating Y and the number of trials, 3.
- The retirement decisions are assumed independent, so we can simply multiply the underlying individual probabilities to get the joint probability of occurrence.
- If we combine these by the term Y, we see that:
  - there is one way to get zero retirements with probability $(1-p)^3$,
  - three ways to get one retirement with probability $p(1-p)^2$,
  - three ways to get to two retirements with probability $p^2(1-p)$,
  - one way to get three retirements with probability p^3.
- This is choosing by unordered selection without replacement, so this is given by the expression $\binom{3}{y}$.

Example Binomial Probabilities



The Binomial Probability Mass Function (PMF)

- For this experiment with three trials we combine the choose part with the probability part:

$$p(Y = y) = \binom{3}{y} p^y (1-p)^{3-y}.$$

- More generally, the general form for n Bernoulli trials is:

$$p(Y = y|n,p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad n \geq y,\ n,y \in \mathcal{I}^+,\ p \in [0{:}1].$$

- This general form or any specific case has the shorthand $\mathcal{B}(n,p)$, for example $\mathcal{B}(10,5)$ for ten trials and five successes, and $\mathcal{B}(100,75)$ for 100 trials and 75 successes.

Example: Binomial Analysis of Bill Passage

- A given legislature has a 0.7 probability of passing routine bills (perhaps from historical analysis).
- If 10 routine bills are introduced in a given week, what is the probability that:
  - Exactly 5 bills pass?

$$
\begin{aligned}
p(Y = 5|n = 10, p = 0.7) &= \binom{10}{5}(0.7)^5(1-0.7)^{10-5} \\
&= (252)(0.16807)(0.00243) \\
&= 0.10292.
\end{aligned}
$$

  - Less than three bills pass?

$$
\begin{aligned}
p(Y < 3|10, 0.7) &= p(Y = 0|10, 0.7) + p(Y = 1|10, 0.7) + p(Y = 2|10, 0.7) \\
&= \binom{10}{0}(0.7)^0(1-0.7)^{10-0} + \binom{10}{1}(0.7)^1(1-0.7)^{10-1} + \binom{10}{2}(0.7)^2(1-0.7)^{10-2} \\
&= 0.0000059049 + 0.000137781 + 0.001446701 \\
&= 0.00159.
\end{aligned}
$$

Example: Binomial Analysis of Bill Passage

- continued...
  - Nine or less bills pass?

$$\begin{aligned}
p(Y \leq 9|10, 0.7) &= \sum_{i=1}^{9} p(Y = i|10, 0.7) = \sum_{i=1}^{10} p(Y = i|10, 0.7) - p(Y = 10|10, 0.7) \\
&= 1 - p(Y = 10|10, 0.7) \\
&= 1 - \binom{10}{10} (0.7)^{10} (1 - 0.7)^{10-10} \\
&= 1 - 0.02825 \\
&= 0.97175.
\end{aligned}$$

Poisson Counts

- Instead of counting the number of successes out of a fixed number of trials, we count the number of events without an upper bound.

- This can also model the length of time waiting for some prescribed event.

- If the probability of the event is proportional to the length of the wait, then the length of wait can be modeled with the *Poisson PMF*:

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y \in \mathcal{I}^{+},\ \lambda \in \mathfrak{R}^{+}.$$

- The single PMF parameter λ is called the intensity parameter and gives the expected number of events.

- λ is also assumed to be the variance of the number of events.

Poisson Counts of Supreme Court Decisions

- Recent Supreme Courts have handed down roughly 8 unanimous decisions per term.
- If we assume that $\lambda = 8$ for the next Court, then what is the probability of observing:

1. Exactly 6 decisions?

$$p(Y = 6|\lambda = 8) = \frac{e^{-8}8^{6}}{6!} = 0.12214.$$

2. Less than three decisions?

$$p(Y < 3|\lambda = 8) = \sum_{i=0}^{2} \frac{e^{-8}8^{y_i}}{y_i!} = 0.00034 + 0.00268 + 0.01073 = 0.01375.$$

3. Greater than 2 decisions?

$$p(Y > 2|\lambda = 8) = 1 - p(Y < 3|\lambda = 8) = 1 - 0.01375 = 0.98625.$$

Two Important Assumptions of the Poisson Model

- Events in different time periods are independent:
  - rates of occurrence in one time period are not allowed to influence subsequent rates in another.
- For small time periods, the probability of an event is proportional to the length of time passed in the period so far, and not dependent on the number of previous events in this period:
  - for some bounded slice of time, as the waiting time increases, the probability of the event increases.

Modeling Nineteenth-Century European Alliances

- McGowan and Rood (1975) look at the frequency of alliance formation from 1814 to 1914 in Europe between the "Great Powers:" Austria-Hungary, France, Great Britain, Prussia-Germany, and Russia.

- The mean number of alliances per year total is 0.545 (out of 55 alliances), which they used as their empirical estimate of λ.

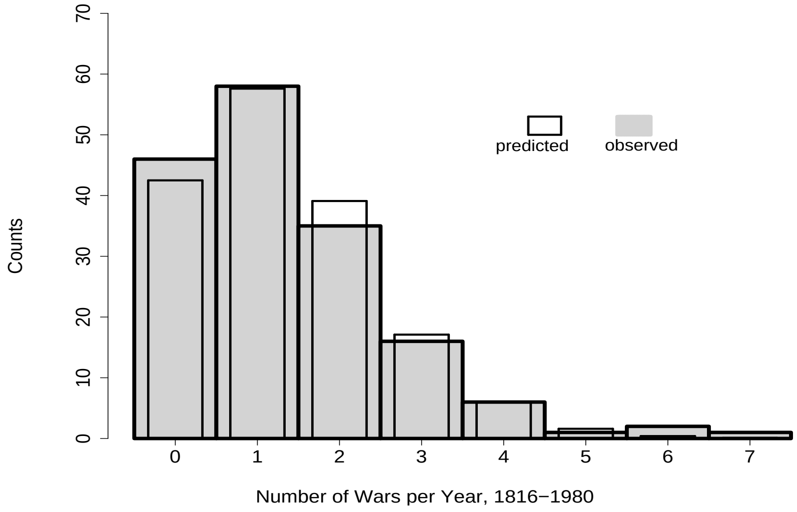
- Therefore we can compare observed events against predicted events:

Alliances/Year	$y = 0$	$y = 1$	$y = 2$	$y \geq 3$
Observed	61	31	6	3
Predicted	58.6	31.9	8.7	1.8

- $\lambda = 0.545$ is the intensity parameter for *five* countries to enter into alliances, so assuming that each country is equally likely, the intensity parameter for an individual country is $\lambda_i = 0.545/5 = 0.109$.

Poisson Process Model of Wars

- Houweling and Kuné (1984) looked at wars as discrete events in "Do Outbreaks of War Follow a Poisson-Process?"
- They compared 224 events of international and civil wars from 1816 to 1980 to that predicted by estimating the Poisson intensity parameter with the empirical mean: $\lambda = 1.35758$.
- Evidence from the figure indicates that the Poisson assumption fits the data well, but this is not so true when the wars were disaggregated by region.



The Cumulative Distribution Function: Discrete Version

- If X is a discrete random variable, then we can define the sum of the probability mass to the left of some point $X = x$: the mass associated with values less than X:

$$F(x) = p(X \leq x).$$

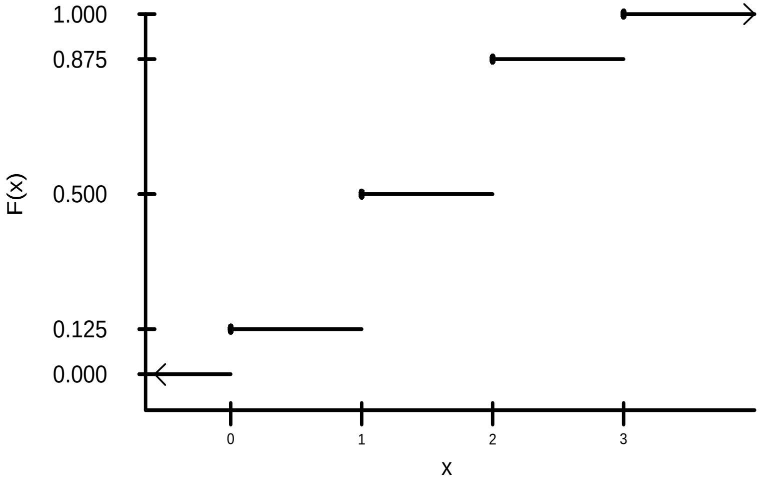
- This *cumulative distribution function* (CDF) is denoted with a capital “F” rather than the lower case notation given for the PMF.

- Sometimes the CDF notation is denoted $F_X(x)$, to remind us that this function corresponds to the random variable X.

- A CDF fully defines a probability function, as does a PMF: since we can readily switch between the two by noting the step sizes (CDF→PMF) or by sequentially summing (PMF→CDF), then the one we use is completely a matter of convenience.

Technical Definition of the Cumulative Distribution Function Definition.

- $F(x)$ is a CDF for the random variable X iff it has the following properties:
  - **bounds:** $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$,
  - **nondecreasing:** $F(x_i) \leq F(x_j)$ for $x_i < x_j$,
  - **right-continuous:** $\lim_{x \downarrow x_i} F(x) = x_i$ for all x_i defined by $f(x)$.
- This is illustrated here.



Probability Density Functions

- Consider a spinner sitting flat on a table measuring the direction of the spinner relative to some reference point in radians, which vary from 0 to 2π.

- There are infinity many outcomes possible because the spinner can theoretically take on any value on the real number line in $[0{:}2\pi]$, meaning a PMF is not the appropriate measure of probability here.

- For continuous random variables we replace the probability mass function with the *probability density function* (PDF).

- Like the PMF, the PDF assigns probabilities to events in the sample space, but because there is an infinite number of alternatives, we cannot say $p(X = x)$ and so just use $f(x)$ to denote the function value at x. The problem lies in questions such

- The solution lies in the ability to replace probabilities of specific events with probabilities of ranges of events.

Exponential PDF

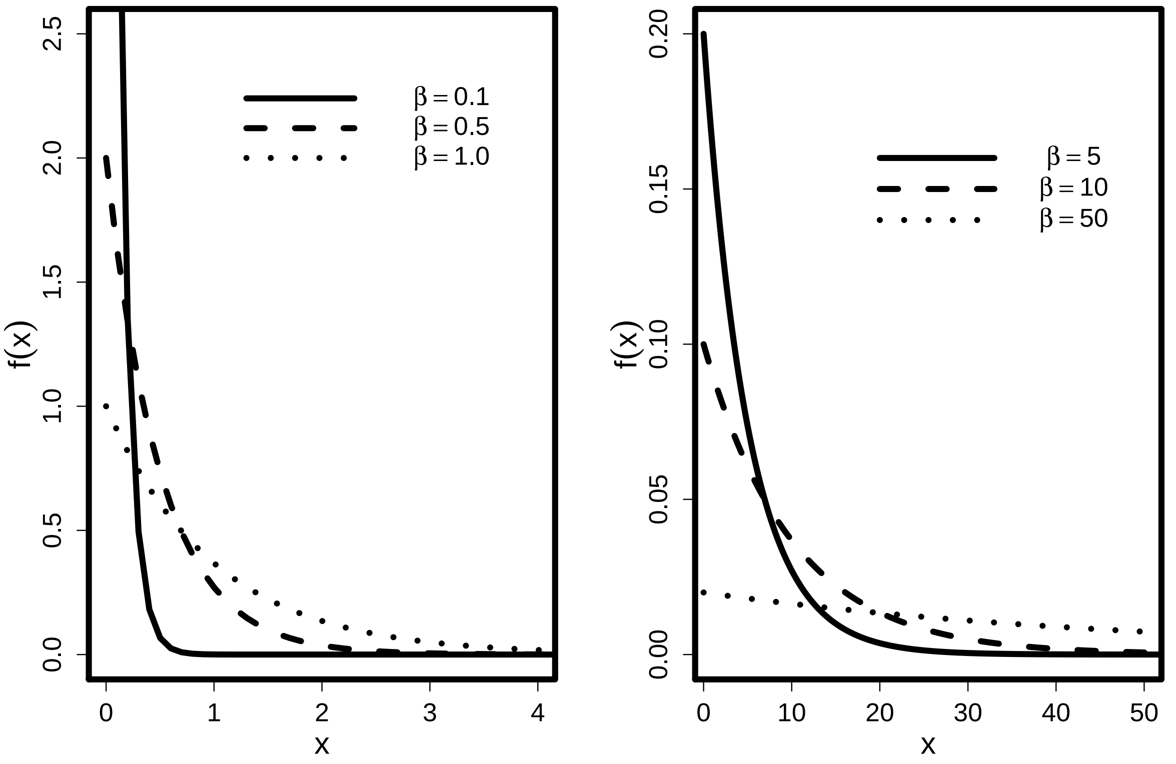
- The *exponential PDF* is a very general and useful functional form that is often used to model durations (how long "things last").

- The PDF is:
$$f(x|\beta) = \frac{1}{\beta}\exp\left[-\frac{x}{\beta}\right], \qquad 0 \leq x < \infty, \quad 0 < \beta,$$
where, similar to the Poisson PMF, the function parameter (β here) is the mean or expected duration.

- The following figure gives six different parameterizations in two frames.

- β is called a *scale parameter*: it affects the scale (extent) of the main density region.

Exponential PDF



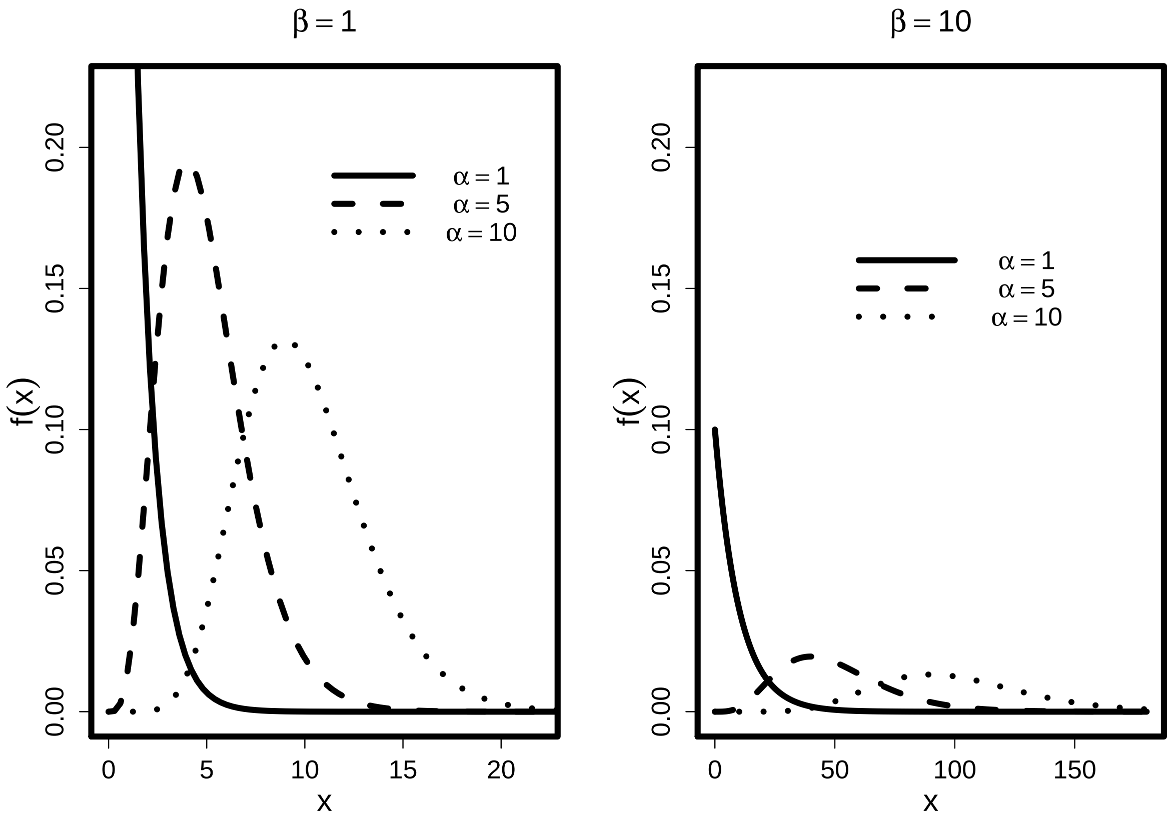
Gamma PDF

- the exponential distribution is a special case of the even more flexible gamma distribution.
- This adds a *shape parameter* that changes the "peakedness" of the distribution: how sharply the density falls from a modal value.
- The gamma PDF is given by

  $$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp\left[-\frac{x}{\beta}\right], \quad 0 \le x < \infty, \quad 0 < \alpha, \beta,$$

  where α is the new shape parameter, and the mean is now $\alpha\beta$.
- Note the use of the gamma function.
- The figure shows different forms based on varying the α and β parameters where the y-axis is fixed across the two frames to show a contrast in effects.

Gamma PDF Forms



Chi-Square PDF

- An important special case of the gamma PDF is the χ^2 distribution, which is used in many statistical tests, including the analysis of tables.
- The χ^2 distribution is a gamma distribution where $\alpha = \frac{df}{2}$ and $\beta = 2$, and *df* is a positive integer value called the *degrees of freedom*.
- The χ^2 PDF is given by:

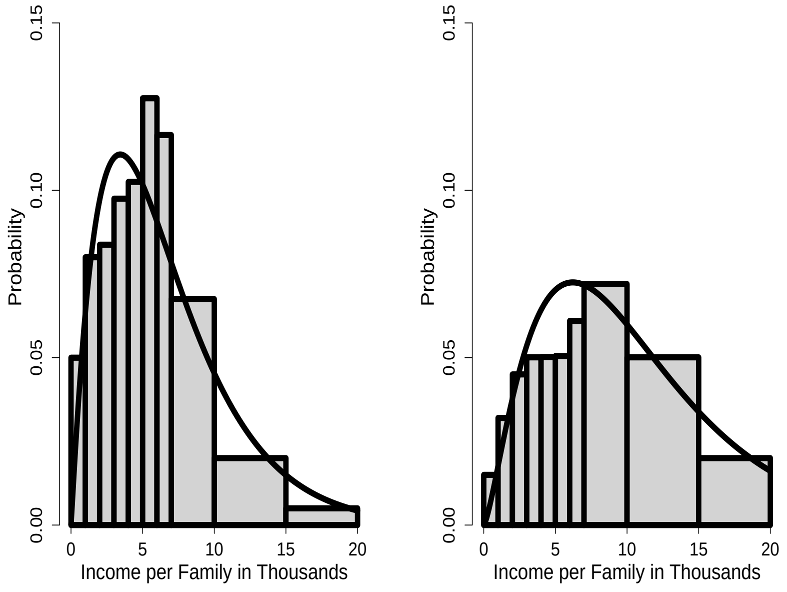
$$f(x|df) = \frac{1}{\Gamma\left(\frac{df}{2}\right) 2^{\frac{df}{2}}} x^{\frac{df}{2}-1} \exp\left[-\frac{x}{2}\right], \quad 0 \leq x < \infty, \quad 0 < df$$

Characterizing Income Distributions

- The gamma distribution is well suited to describing data that have a mode near zero and a long right (positive) skew.

- Pareto (1897) first noticed that income in societies, no matter what kind of society, follows this pattern, and this effect is sometimes called *Pareto's Law.*

- Examples fitting gamma distributions to income:



Normal PDF

- By far the most famous probability distribution is the *normal PDF*, sometimes also called the *Gaussian PDF*.

- The PDF is:

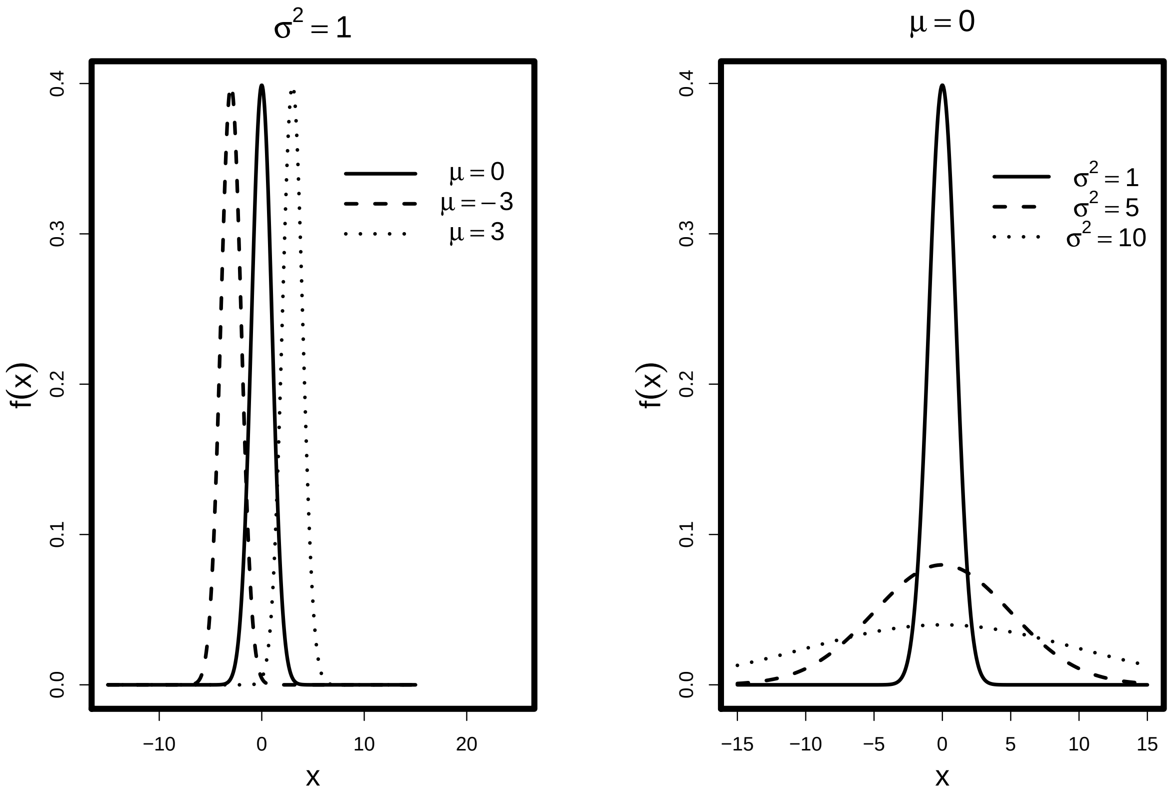
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \qquad -\infty < x, \mu < \infty, 0 < \sigma^2,$$

  where μ is the mean parameter and σ^2 is the dispersion (variance) parameter.

- The two terms completely define the shape of the particular normal form where μ moves the modal position along the x-axis, and σ^2 makes the shape more spread out as it increases.

- The normal distribution is a member of the *location-scale family* of distributions because μ moves only the location (and not anything else) and σ^2 changes only the scale (and not the location of the center or modal point).

Normal PDF Illustration



The Standard Normal Distribution

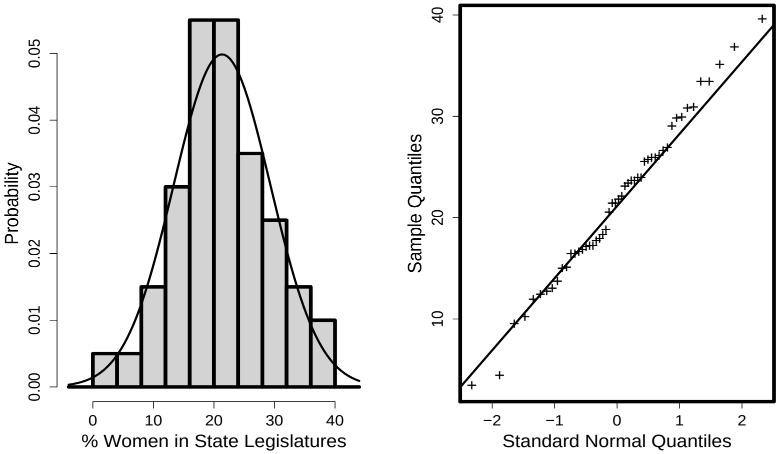
- A normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called a *standard normal* and is of great practical as well as theoretical significance.
- The PDF for the standard normal simplifies to

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right], \qquad -\infty < x < \infty.$$

- Due to the location-scale characteristic, any other normal distribution can be transformed to a standard normal and then back again to its original form:
  - Suppose $x \sim \mathcal{N}(\mu, \sigma^2)$,
  - then $y = (x - \mu)/\sigma^2 \sim \mathcal{N}(0, 1)$,
  - return to x by substituting $x = y\sigma^2 + \mu$.
- This means is that textbooks need only include one normal table (the standard normal) for calculating tail values.

Levels of Women Serving in U.S. State Legislatures

- The first panel below shows a histogram of the percent of women in legislatures for the 50 states with a normal distribution ($\mu = 21, \sigma = 8$) superimposed:



- The normal curve appears to match well the distribution given in the histogram.
- The second panel of the figure is a “qqplot” that plots the data against standard normal quantiles (a set of ordered values from the standard normal PDF of length equal to the evaluated vector).
- The fit is quite close with just a little bit of deviation in the tails.

The Cumulative Distribution Function: Continuous Version

- If X is a continuous random variable, then we can also define the sum of the probability mass to the left of some point $X = x$: the density associated with all values less than X.
- The function:
$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(x)dx$$
defines the cumulative distribution function (CDF) for the continuous random variable X.
- It is always a smooth curve monotonically nondecreasing from zero to one.

The Standard Normal CDF: Probit Analysis

.

- The CDF of the standard normal is abbreviated $\Phi(X)$ for $\mathcal{N}(X \leq x|\mu = 0, \sigma^2 = 1)$ (the associated PDF notation is $\phi(X)$).
- While people may make dichotomous choices (vote/not vote, purchase/not purchase, etc.), the underlying mechanism of decision is really a smooth, continuous preference or utility function that describes more subtle thinking.
- The positive/action choice is labeled as “1” and the opposite event as “0,”
- If there is some interval measured variable X that affects the choice, then $\Phi(X) = p(X = 1)$ is called the *probit model*.
- Higher levels of X are assumed to push the subject toward the “1” decision, and lower levels of X are assumed to push the subject toward the “0” decision (the opposite effect can easily be modeled as well).

The Standard Normal CDF: Probit Analysis

.

- Consider the dichotomous choice outcome of voting for a Republican congressional candidate against an interval measured explanatory variable for political ideology.

- There is also a second variable indicating whether the respondent owns a gun.

- A probit model is specified for these data as:

$$p(Y_i = 1) = \Phi(IDEOLOGY_i + GUN_i).$$

- $IDEOLOGY_i$ is the political ideology value for individual i.

- GUN_i is a dichotomous variable equaling one for gun ownership and zero otherwise.

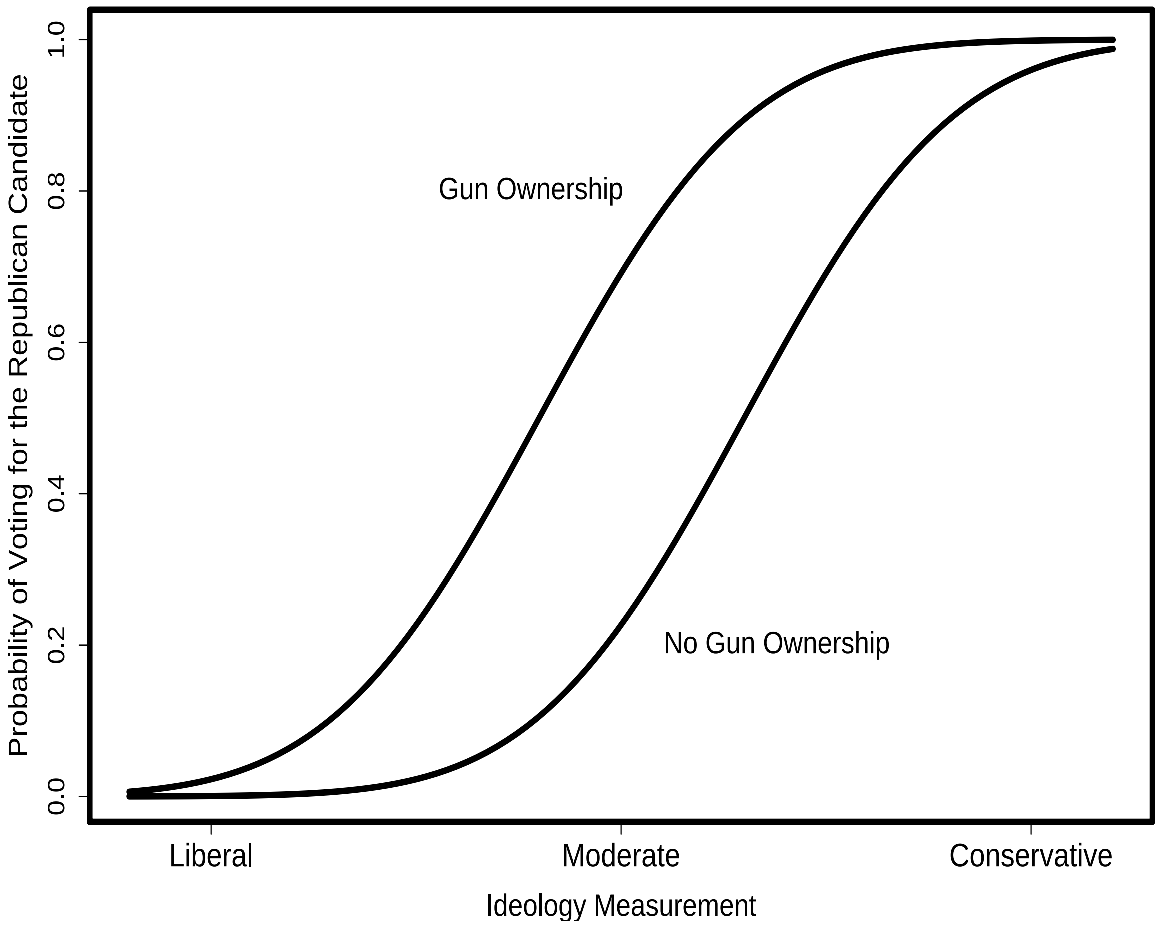
The Standard Normal CDF: Probit Analysis

.

- In the following figure gun owners and nongun owners are separated.
- Gun ownership shifts the curve affecting the probability of voting for the Republican candidate by making it more likely at more liberal levels of ideology.
- For very liberal and very conservative respondents, gun ownership does not really affect the probability of voting for the Republican.
- For respondents without a strong ideological orientation, gun ownership matters considerably: a difference of about 50% at the center.

The Standard Normal CDF: Probit Analysis

.



The Uniform Distributions

- The uniform distribution (flat) models equal probabilities for both discrete and continuous assumptions.
- The form is:

  k-Category Discrete Case (PMF):

$$p(Y=y|k)=\begin{cases}\frac{1}{k}, & \text{for } y=1,2,\ldots,k\\ 0, & \text{otherwise;}\end{cases}$$

  Continuous Case (PDF):

$$f(y|a,b)=\begin{cases}\frac{1}{b-a}, & \text{for } a\leq y\leq b\\ 0, & \text{otherwise.}\end{cases}$$

- Each outcome has equal individual probability (PMF) or equal density (PDF).

The Uniform Distributions

- The discrete case specifies k outcomes (hence the conditioning on k in $p(Y = y|k)$) that can be given any range.

- Greater ranges make $\frac{1}{k}$ smaller for fixed k), and the continuous case just gives the bounds (a and b), which are often zero and one.

- The uniform distribution is sometimes used to reflect great uncertainty about outcomes, although it is definitely saying something specific about the probability of events.

- The continuous case with $a = 0$ and $b = 1$ is particularly useful in modeling probabilities.

Measures of Central Tendency: Mean, Median, and Mode

- The first and most useful step in summarizing observed data values is determining its *central tendency*: a measure of where the “middle” of the data resides on some scale.

- There is more than one definition of what constitutes the center of the distribution of the data, the so-called average.

- The first choice for an average is the mean: for n data points, $x_1, x_2, \ldots, x_n$, the mean is :

  $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

  where the bar notation is universal for denoting a mean average.

Measures of Central Tendency: Mean, Median, and Mode

- The median average has a different characteristic; it is the point such that as many cases are greater as are less: For n data points $x_1, x_2, \ldots, x_n$, the median is X_i such that $i = \lceil n/2 \rceil$ (even n) or $i = \frac{n+1}{2}$ (odd n).

- This definition suits an odd size to the dataset better than an even size, but in the latter case we just split the difference and define a median point that is halfway between the two central values.

- The median is defined as

$$M_x = X_i \colon \int_{-\infty}^{x_i} f_x(X)dx = \int_{x_i}^{\infty} f_x(X) = \frac{1}{2}$$

where $f_x(X)$ denotes the *empirical distribution* of the data, that is, the distribution observed rather than that obtained from some underlying mathematical function generating it.

Measures of Central Tendency: Mean, Median, and Mode

- The mode is the most frequently observed value.
- Since all observed data are countable, and therefore discrete, this definition is workable for data that are also continuously measured.
- The mode is:

$$m_x = X_i \colon n(X_i) > n(X_j) \; \forall j \neq i,$$

where the notation "$n()$" means "number of" values equal to the X in the cardinality sense.

Employment by Race in Federal Agencies

.

Table 1: Percent Employment by Race, 1998

Agency	Black	Hispanic	Asian	White
Agriculture	10.6	5.6	2.4	81.4
Commerce	18.3	3.4	5.2	73.1
DOD	14.2	6.2	5.4	74.3
Army	15.3	5.9	3.7	75.0
Navy	13.4	4.3	9.8	72.6
Air Force	10.6	9.5	3.1	76.8
Education	36.3	4.7	3.3	55.7
Energy	11.5	5.2	3.8	79.5
EOP	24.2	2.4	4.2	69.3
HHS	16.7	2.9	5.1	75.4
HUD	34.0	6.7	3.2	56.1
Interior	5.5	4.3	1.6	88.6
Justice	16.2	12.2	2.8	68.9
Labor	24.3	6.6	2.9	66.5
State	14.9	4.2	3.7	77.1
Transportation	11.2	4.7	2.9	81.2
Treasury	21.7	8.4	3.3	66.4
VA	22.0	6.0	6.7	65.4
GSA	28.4	5.0	3.4	63.2
NASA	10.5	4.6	4.9	80.1
EEOC	48.2	10.6	2.7	38.5

Source: Office of Personnel Management

Employment by Race in Federal Agencies

.

- The mean values by racial group are $\bar{X}_{\text{Black}} = 19.43$, $\bar{X}_{\text{Hispanic}} = 5.88$, $\bar{X}_{\text{Asian}} = 4.00$, and $\bar{X}_{\text{White}} = 70.72$.
- The median values differ somewhat: $M_{\text{Black}} = 16.2$, $M_{\text{Hispanic}} = 5.2$, $M_{\text{Asian}} = 3.4$, and $M_{\text{White}} = 73.1$.
- Cases where the mean and median differ noticeably are where the data are skewed (asymmetric) with the longer "tail" in the direction of the mean.
- These data do not have a modal value in unrounded form, but we can look at modality through a *stem and leaf plot*, which groups data values by leading digits and looks like a histogram that is turned on its side.
- Unlike a histogram, though, the bar "heights" contain information in the form of the lower digit values.

Employment by Race in Federal Agencies

Black, the decimal point is 1 digit to the right of the |

```
0|6
1|111123455678
2|2244
2|8
3|4
3|6
4|
4|8
```

Asian, the decimal point is at the |

```
1|6
2|47899
3|12334778
4|29
5|124
6|7
7|
8|
9|8
```

Hispanic, the decimal point is at the |

```
 2|49
 3|4
 4|233677
 5|0269
 6|0267
 7|
 8|4
 9|5
10|6
11|
12|2
```

White, the decimal point is 1 digit to the right of the |

```
3|9
4|
5|66
6|356799
7|3345577
8|00119
```

Breakdown Bound

- One way to consider the utility of the three different averages is to evaluate their resistance to large outliers.
- The *breakdown bound* is the proportion of data values that can become unbounded (go to plus or minus infinity) before the statistic of interest becomes "unbounded" itself.
- The mean has a breakdown bound of 0 because even one value of infinity will take the sum to infinity.
- The median is much more resistant because almost half the values on either side can become unbounded before the median itself goes to infinity.
- The mode is much more difficult to analyze in this manner as it depends on the relative placement of values:
  - It is possible for a high proportion of the values to become unbounded provided a higher proportion of the data is concentrated at some other point.
  - If these points are more spread out, however, the infinity point may become the mode and thus the breakdown bound lowers.
  - Due to this uncertainty, the mode cannot be given a definitive breakdown bound value.

Measures of Dispersion: Variance, Standard Deviation, and MAD

- The second most important and common data summary technique is calculating a measure of spread: how dispersed are the data around a central position?
- Often a measure of centrality and a measure of spread are sufficient to give researchers a very good, general view of the behavior of the data.
- This is particularly true if we know something else, such as that the data are unimodal and symmetric.

Measures of Dispersion: Variance, Standard Deviation, and MAD

- The most useful and common measure of dispersion is the *variance*:
  - For n data points $x_1, x_2, \ldots, x_n$, the variance is given by:

$$\text{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

  - The fraction, $\frac{1}{n-1}$, is slightly surprising given the more intuitive $\frac{1}{n}$ for the mean.
  - Without the -1 component the statistic is *biased*: not quite right on average for the true underlying population quantity.

- A second closely related quantity is the *standard deviation*, which is simply the square root of the variance:

$$SD(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Since the variance and the standard deviation give the same information, the choice of which one to use is often circumstantial.

Measures of Dispersion: Variance, Standard Deviation, and MAD

- The *median absolute deviation* (MAD) is the median of the absolute deviations from the data median:
  $$MAD(X) = median(|x_i - median(x)|),$$
  for $i = 1, 2, \ldots, n$.

- The variance (and therefore the standard deviation) is very sensitive to large outliers, more so even than the mean due to the squaring.

- Conversely, the MAD obviously uses medians which are far more resistant to atypical values.

Example: Employment by Race in Federal Agencies, Continued

- Calculate here the three described measures of dispersion for the racial groups.
- It is important to remember that none of these three measures is necessarily the "correct" view of dispersion in the absolute sense.
- For the race data there are:

Table 2: Measures of Dispersion, Race in Agencies

	Black	Hispanic	Asian	White
variance	107.20	6.18	3.16	122.63
standard deviation	10.35	2.49	1.78	11.07
MAD	5.60	1.00	0.60	6.60

Correlation and Covariance

- Sometimes we care about how to variables "move" or more precisely vary together.

- For example we expect income and education to vary in the same direction: higher levels of one are associated with higher levels of the other, and conversely income and prison time to vary negatively together.

- This means that if we look at a particular case with a high level of education, we expect to see a high income.

- Note the careful use of the word "expect" here, meaning that we are allowing for cases to occur in opposition to our notion without necessarily totally disregarding the general idea.

- This means that we allow for exceptions to the income/education relationship (professors and plumbers).

Covarying

- *Covariance* is a measure of variance with two paired variables.
- Positive values mean that there is positive varying effect between the two, and negative values mean that there is negative varying effect: High levels of one variable are associated with low levels of another.
- For two variables of the same length,

  $$x_1, x_2, \ldots, x_n \qquad \text{and} \qquad y_1, y_2, \ldots, y_n,$$

  the covariance is given by

  $$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$
- If there is no relationship between two variables, then it seems reasonable to expect a covariance near zero (but not vice-versa).

Covarying

- What happens if we calculate the covariance of some variable with itself? Let's take $\text{Cov}(X, Y)$ and substitute in $Y = X$:

  $$\begin{aligned}\text{Cov}(X, X) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \text{Var}(X).\end{aligned}$$

  so the covariance is a direct generalization of the variance where we have two variables instead of one to consider.

- But there is one problem with the covariance: We do not have a particular scale for saying what is large and what is small for a given dataset.

Pearson's Product Moment Correlation Correlation

- To solve this *scaling problem*, we the covariance in units of the standard deviation of X and Y:

$$\text{Cov}(X,Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

- This scales this statistic to be bounded by $[-1{:}1]$. That is,
- Notice that if we re-performed our trick of substituting $Y = X$ (or equivalently $X = Y$), then the statistic would be equal to one: Y covaries *exactly* as X covaries. On the other hand, if we
- Substituting $Y = -X$ (or conversely $X = -Y$), produces a correlation equal to negative one, meaning that Y covaries in exactly the opposite manner as X. Since these are the limits of the ratio,
- Any value inbetween ±1 represents lesser degrees of absolute scaled covariance.

An Ethnoarchaeological Study in the South American Tropical Lowlands

- Siegel (1990) looked at the relationship between the size of buildings and the number of occupants in a South Amerindian tropical-forest community located in the upper Essequibo region of Guyana.

- The main tool employed by Siegel was a correlational analysis between floor area of structures and occupational usage.

- There are four types of structures: residences, multipurpose work structures, storage areas, and community buildings.

- In these tribal societies it is common for extended family units to share household space including kitchen and storage areas but to reserve a component of this space for the nuclear family.

- Thus there is a distinction between *households* that are encompassing structures, and the individual *residences* within.

An Ethnoarchaeological Study in the South American Tropical Lowlands

- The table of correlation coefficients is given between between the size of the floor area for three definitions of space and the family unit for the village of Shefariymo where **Total** is the sum of **Multipurpose** space, **Residence** space, and **Storage** space:

Table 3: Correlation Coefficient of Family Unit vs. Space

Structure Type	Family Unit	Cases	Correlation
Multipurpose	Nuclear	16	0.137
Residence	Nuclear	24	0.662
Total	Nuclear	24	0.714
Multipurpose	Extended	12	0.411
Residence	Extended	11	0.982
Total	Extended	11	0.962

- We see therefore: a positive but weak relationship between the size of the nuclear family and the size of the multipurpose space (0.137), a relatively strong relationships between the size of these same nuclear families and the size of their residences (0.662) and the total family space (0.714), the size of the extended family is almost perfectly correlated with residence size and total size.

Expected Value

- Expected value is a probability-weighted average over possible events.
- For example with a fair coin, there are 5 *expected* heads in 10 flips.
- This does not mean that 5 heads will necessarily occur, but that we would be inclined to bet on 5 rather than any other number.
- With interval measured data you *never* get the exactly the expected value in a given experiment because the probability of any one point on the real number line is zero.

Expected Value

- The discrete form of the expected value of some random variable X is

  $$E[X] = \sum_{i=1}^{k} X_i p(X_i),$$

  for k possible events in the discrete sample space $\{X_1, X_2, \ldots, X_k\}$.

- The continuous form is exactly the same except that instead of summing we need to integrate:

  $$E[X] = \int_{-\infty}^{\infty} X p(X) dX,$$

  where the integral limits are adjusted if the range is restricted for this random variable.

Expected Value Example

- Suppose someone offered you a game that consisted of rolling a single die with the following payoffs:

die face	1	2	3	4	5	6
X, in dollars	0	1	1	1	2	2

- Would you be inclined to play this game if it costs \$2?
- The expected value of a play is calculated as

$$\begin{aligned} E[X] = \sum_{i=1}^{6} X_i p(X_i) &= \frac{1}{6}(0) + \frac{1}{6}(1) + \frac{1}{6}(1) + \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(2) \\ &= \$1.167 \text{ (rounded).} \end{aligned}$$

- Therefore it would not make sense to pay \$2 to play this game.

Expected Values of Functions of a Random Variable

- Discrete and continuous forms are given by

  $$E[f(X)] = \sum_{i=1}^{k} f(X_i)p(X_i), \qquad E[f(X)] = \int_{-\infty}^{\infty} f(X)p(X)dX.$$

  For instance, if we have the function $f(X) = X^2$, then, given a continuous random variable, $E[X^2] = \int X^2 p(X)dX$.

- The calculation of expected value for vectors and matrices is only a little bit more complicated because we have to keep track of the dimension.

- A $k \times 1$ vector $\mathbf{X}$ of discrete random variables has the expected value $E[\mathbf{X}] = \sum \mathbf{X}p(\mathbf{X})$.

- For the expected value of a function of the continuous random vector it is common to use the *Riemen-Stieltjes integral* form:

  $$E[f(\mathbf{X})] = \int f(\mathbf{X})dF(\mathbf{X}),$$

  where $F(\mathbf{X})$ denotes the joint distribution of the random variable vector $\mathbf{X}$.

Conditional Expectation

- In much statistical work expected value calculations are "conditional" in the sense that the average for the variable of interest is taken conditional on another.

- For instance, the discrete form for the expected value of Y given a specific level of X is

$$E[Y|X] = \sum_{i=1}^{k} Y_i p(Y_i|X).$$

Sometimes expectations are given subscripts when there are more than one random variables in the expression and it is not obvious to which one the expectation holds:

$$\text{Var}_x[E_y[Y|X]] = \text{Var}_x\left[\sum_{i=1}^{k} Y_i p(Y_i|X)\right].$$

Some Handy Expectation Properties and Rules

- Since expectation is a summed quantity, many of these rules are obvious, but some are worth thinking about.

- Let X, Y, and Z be random variables defined in $\mathfrak{R}$ (the real number line), whose expectations are finite.

- Finite Expectation Properties for X, Y, Z:
  - ▷ $E[a + bX] = a + bE[X]$, for constants a, b.
  - ▷ $E[X + Y] = E[X] + E[Y]$.
  - ▷ $E[X + Y|Z] = E[X|Z] + E[Y|Z]$.
  - ▷ $E[Y|Y] = Y$.
  - ▷ $E[E[Y|X]] = E[Y]$, "double expectation" (also Y: $E[E[f(Y)|X]] = E[f(Y)]$).
  - ▷ If $X \geq Y$, then $E[X] \geq E[Y]$ with probability one.
  - ▷ $|E[X|Y]| \leq E[|X||Y]$ with probability one.
  - ▷ If X and Y are independent, then $E[XY] = E[X]E[Y]$ (also holds under the weaker assumption of uncorrelatedness).

Variance Properties and Expectation

- Finite Variance Properties for X and Y (without assuming independence or uncorrelatedness):
  - $\text{Cov}[X, Y] = E[XY] - E[X]E[Y]$.
  - $\text{Var}[X] = E[X^2] - (E[X])^2$.
  - $\text{Var}[X|Y] = E[X^2|Y] - (E[X|Y])^2$.
  - $\text{Cov}[X, Y] = \text{Cov}[X, E[Y|X]]$.
  - $\text{Var}[Y] = \text{Var}_x[E_y[Y|X]] + E_x[\text{Var}_y[Y|X]]$, “decomposition.”

Inequalities Based on Expected Values, All Values Assumed Finite

- *Chebychev's Inequality.* If $f(X)$ is a positive and nondecreasing function on $[0, \infty]$, then for all (positive) values of k
$$p(f(X) > k) \leq E[f(X)/k].$$
- A more common and useful form of Chebychev's inequality involves μ and σ, the mean and standard deviation of X For k greater than or equal to 1:
$$p(|X - \mu| \geq k\sigma) \leq 1/k^2.$$
  To relate these two forms, recall that μ is the expected value of X.
- *Markov Inequality.* Similar to Chebychev's Inequality:
$$P[|X| \geq k] \leq E[|X|^\ell]/k^\ell.$$

Inequalities Based on Expected Values, All Values Assumed Finite

- *Jensen's Inequality.* If $f(X)$ is a concave function (open toward the x-axis, like the natural log function), then
$$E[f(X)] \leq f(E[X]).$$
Conversely, if $f(X)$ is a convex function (open away from the x-axis, like the absolute value function), then
$$E[f(X)] \geq f(E[X]).$$

- *Minkowski's Inequality.* For $k > 1$,
$$(E[|X+Y|^k])^{1/k} \leq (E[|X|^k])^{1/k} + (E[|Y|^k])^{1/k}.$$

Inequalities Based on Expected Values, All Values Assumed Finite

- For $\frac{1}{k} + \frac{1}{\ell} = 1$, *Hölder's Inequality*
$$|E[XY]| \le E[|XY|] \le (E[|X|]^k)^{1/k}(E[|Y|^\ell])^{1/\ell}.$$

- *Schwarz Inequality.* An interesting special case of Hölder's Inequality where $k = \ell = \frac{1}{2}$:
$$E[|XY|] \le \sqrt{E[X^2]E[Y^2]}$$

- *Liapounov's Inequality.* For $1 < k < \ell$,
$$(E[|X|]^k)^{1/k} \le (E[|X|]^\ell)^{1/\ell}.$$

Inequalities Based on Expected Values, All Values Assumed Finite

- *Cramer-Rao Inequality.* Given a PDF or PMF conditional on a parameter vector, $f(\mathbf{X}|\boldsymbol{\theta})$, define the *information matrix* as

  $$I(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log(f(\mathbf{X}|\boldsymbol{\theta}))'\frac{\partial}{\partial\boldsymbol{\theta}}\log(f(\mathbf{X}|\boldsymbol{\theta}))\right],$$

  and define the vector quantity

  $$\alpha = \frac{\partial}{\partial\boldsymbol{\theta}}E[f(\mathbf{X}|\boldsymbol{\theta})].$$

  Then

  $$\text{Var}(f(\mathbf{X}|\boldsymbol{\theta})) \geq \alpha I(\boldsymbol{\theta})\alpha.$$

- *Berge Inequality.* Suppose $E[X] = E[Y] = 0$, $\text{Var}[X] = \text{Var}[Y] = 1$, and $\sigma^2 = \text{Cov}[X,Y]$. (these are *standardized* random variables), then for (positive) k

  $$P[\max(|X|,|Y|) \geq k] \leq (1 + \sqrt{1 - \text{Cov}(X,Y)^2})/k^2.$$

Moments of a Distribution

- Most (but not all) distributions have a series of *moments* that define important characteristics of the distribution.

- The first moment, which is the mean or expected value of the distribution.

- The general formula for the kth moment is based on the expected value:

$$m_k = E[X^k] = \int_X x^k dF(x)$$

  for the random variable X with distribution $f(X)$ where the integration takes place over the appropriate support of X.

- It can also be expressed as

$$m_k = \int_X e^{kx} dF(x).$$

- An important theory says that a distribution function is “determined” by its moments of all orders (i.e., all of them), and some distributions have an infinite number of moments defined.

Central Moments

- The kth *central moment* is (often called just the "kth moment")
$$m'_k = E[(X - m_1)^k] = \int_X (x - m_1)^k dF(x).$$
where m_1 is the mean.

- The most obvious and important central moment is the variance:
$$\sigma^2 = E[(X - \bar{X})^2].$$

Central Moments

- This calculation for the exponential PDF is

$$
\begin{aligned}
\text{Var}[X] &= E[(X - E[X])^2] \\
&= \int_0^\infty (X - E[X])^2 f(x|\beta)dx \\
&= \int_0^\infty (X - \beta)^2 \frac{1}{\beta} \exp[-x/\beta]dx \\
&= \int_0^\infty (X^2 - 2X\beta + \beta^2) \frac{1}{\beta} \exp[-x/\beta]dx \\
&= \int_0^\infty X^2 \frac{1}{\beta} \exp[-x/\beta]dx \\
&\qquad + \int_0^\infty 2X \exp[-x/\beta]dx + \int_0^\infty \beta \exp[-x/\beta]dx \\
&= (0 - 2\beta^2) + (2\beta^2) + (\beta^2) \\
&= \beta^2,
\end{aligned}
$$

where we use integration by parts and L'Hospital's Rule to do the individual integrations.

Cauchy PDF Example

- The Cauchy PDF has no finite moments at all, even though it is "bell shaped" and looks like the normal.
- Definition:

$$f(x|\beta) = \frac{1}{\beta}\frac{1}{1+(x-\beta)^2}, \qquad -\infty < x, \beta < \infty.$$

- Calculating the first moment with integration by parts gives

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} \frac{1}{\beta}\frac{x}{1+(x-\beta)^2}dx \\ &= \left[x\mathrm{arctan}(x-\beta) - (x-\beta)\mathrm{arctan}(x-\beta) + \frac{1}{2}\log(1+x^2) \right]\Bigg|_{-\infty}^{\infty} \\ &= \beta\mathrm{arctan}(x-\beta)\Bigg|_{-\infty}^{\infty} + \frac{1}{2}\log(1+x^2)\Bigg|_{-\infty}^{\infty}. \end{aligned}$$

- The first term above is finite because $\mathrm{arctan}(\pm\infty) = \pm\frac{1}{2}\pi$.
- The second term is clearly infinite.