Learn Probability and Statistics
Last update: November 25, 2011

Timothy C. Haas
Lubar School of Business
University of Wisconsin-Milwaukee
haas@uwm.edu
Outline of this On-line Tutorial

1. Approach of this tutorial.

2. Goals of this tutorial.

3. Some references.

4. Learn basic probability.

5. Learn basic statistics.

6. Learn basic statistical hypothesis testing.

7. Learn basic regression analysis.

1 Approach of this Tutorial

One theory in education states that a concept is learned as a person is able to recognize it and use it in a variety of situations and contexts. To do this, it is essential to vary the words used to describe the concept and to give practice in use of the concept in apparently disparate situations.

A criticism of most online learning systems (hereafter, e-tutorials) is that they are boring. A person (hereafter, student) trying to learn a concept through an e-tutorial does not have benefit of peer pressure and encouragement of his/her classmates, and further, does not have the stimulation of face-to-face contact with an instructor as with a traditional course. Hence, the challenge to motivate thinking must come entirely from the online system. If the student finds the questions or problems of the e-tutorial too easy, the student will lose interest in learning. Hence, contrary to many e-tutorial approaches, this e-tutorial will strive to present only difficult, challenging problems for the student to solve. There are no fully worked out solutions to the exercises – only hints and numerical answers for the final value or values of an exercise. The reason for this restriction is that a student who is working through an unsupervised, unrestricted e-tutorial, has immediate access to exercise solutions – hence, students will jump immediately to these if just from mere curiosity. If a complete solution is read before attempting the exercise, the motivation to seriously think through to a solution will be killed.
Exercises should be self-checking, i.e., after a solution has been worked out, some mathematical or computational procedures should be applied to partially check the answer’s correctness. For example, any probability answer must be in the unit interval. Such diagnostics will be included in the exercise description.

Many e-tutorials use random generation of new data sets within exercises to supposedly keep the exercise from being predictable. This will be avoided here because emphasis will be on concept understanding and not on the mechanics of a data analysis.

Learning takes place through a serious effort given to solving the exercises. This is true of any quantitative subject regardless of the presentation medium. Hence, the completion of the exercises is viewed here as the student’s most important learning activity. Again, this is why there will be no fully-worked solutions available. In particular, the hints will be staged, i.e., there will be a sequence of hints that guide the student towards a solution of the exercise. Each subsequent hint will only be intelligible if the exercise has been worked up through all previous hints.

Another criticism of e-tutorials is that it is tedious to read through reams of online text. This e-tutorial will avoid this pitfall by keeping explanations and presentations of new material to the minimum number of words needed to convey the idea. Brevity will be a high priority of this e-tutorial.

Probability and statistics are intrinsically mathematical and symbol-based disciplines. In this author’s opinion, there is no use in sugar coating this fact, attempting to minimize it, or avoiding it altogether. This e-tutorial therefore, will use mathematical symbols as necessary and demand that the student become versed in the meanings of the symbols. Here, “demand” is operationalized by communicating exercises and their end-point solution exclusively through mathematical symbols. This e-tutorial can be used by any student who has had a course in pre-calculus.

2 Goals of this e-tutorial

A student who has completed this e-tutorial will REALLY understand:

1. The difference between a random variable and a datum.

2. The difference between an estimator and an estimate.

3. Probability distributions, cumulative distribution functions (CDFs), probability density functions (PDFs), quantiles, and realizations (deviates).
3 Some References

3.1 Taking Quantitative Lecture Notes

Taking lecture notes is hard work and requires concentration. If you are not used to taking lecture notes in a quantitative course, there are several good study guides on how to take lecture notes and do well in a quantitative course. Two of these are:


3.2 Introductory Statistics Texts


3.3 Free, Online Texts and Videos


http://davidakenny.net/statbook/kenny87.pdf

www.psychstat.missouristate.edu/sbk00.htm

www.sjsu.edu/faculty/gerstman/StatPrimer

www.khanacademy.org
4 Basic Probability

4.1 Definitions

1. An experiment is an activity that produces one or more outcomes. Experiments studied in probability theory have outcomes that are not completely predictable before conducting the experiment.

2. One performance of the experiment is called a trial.

3. Examples in business practice include (a) conducting a survey, (b) purchasing an investment instrument, (c) writing an invoice, (d) testing a product, (e) purchasing raw materials, (f) implementing a new business plan, or (g) announcing a new product.

4. Let \( E \) be a particular collection or set of outcomes of an experiment. \( E \) is called an event. Say that the experiment is performed \( n \) times (\( n \) trials).

5. A Sample Space, \( S \) is the set of all possible outcomes of an experiment.

6. The probability of \( E \) occurring on the next run of the experiment is

\[
\lim_{n \to \infty} \frac{\text{# of occurrences of } E}{n}
\]

This is called the long run frequency definition of probability.

7. Say that the experiment consists of randomly selecting \( n \) objects from \( N \) objects. Define an event, \( E \) to be a particular type of \( n \) objects.

8. Example: From a jury selection pool of 14 men and 13 women, there is 1 person who always votes for “acquit.” An event of interest is the chance or probability that a jury (12 persons) will be formed that contains this person – or equivalently the chance of selecting a guaranteed “hung” jury.

9. For the experiment of selecting a jury, the sample space is the list of all possible juries that can be formed from a particular group of 14 men and 13 women.

10. Events are built from combinations of the experiment’s outcomes. Set theory notation is used to express outcomes, event, and sample spaces.
11. **Example:** A copier has 2 points where it can jam. In one attempt to copy (a trial), either a jam occurs or not.

Notation:

- $O_1$: the outcome that a jam occurs at point 1
- $O_2$: the outcome that a jam occurs at point 2
- $E$: the event that the copier jams

Then $S = \{O_1, O_2, O_3\}$ where $O_3$ is the outcome that no jam occurs. $E = O_1 \cup O_2$ and is read “$O_1$ or $O_2$.”

12. The null set, $\phi$ is always in $S$, i.e., $\phi \in S$ (the symbol $\in$ means “is a member of”).

13. The **complement** of an event $E$, written $E^C$ is the set of all outcomes that are not in $E$. For the copier, $E^C = \{O_3, \phi\}$.

14. If two events or outcomes have no common members, they are said to be **disjoint** or mutually exclusive. With the copier, $O_1 \cap O_2 = \phi$ (read “$O_1$ and $O_2$”), and $E \cap O_3 = \phi$. So $O_1$ and $O_2$ are disjoint as are $E$ and $O_3$.

15. For *any* experiment, there are several Laws of Probability – all of which can be found from the following three Axioms of Probability:

   **Axiom 1:** If $E \subset S$, then $0 \leq P(E) \leq 1$ (the symbol $\subset$ means “is a subset of”).

   **Axiom 2:** $P(S) = 1$.

   **Axiom 3:** $E_1, E_2 \subset S$ and $E_1 \cap E_2 = \phi$, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

16. **Probability Laws:**

   (a) $P(\phi) = 0$

   (b) If $E \subset S$, then $P(E^C) = 1 - P(E)$.

   (c) If $E_1, E_2 \subset S$, then $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

17. For the copier experiment, say that $P(O_1) = .1$ and $P(O_2) = .2$. Then using Axiom 3,

   $$P(E) = P(O_1 \cup O_2) = .1 + .2 = .3.$$
18. A conditional probability can be computed with Bayes Theorem: If $A, B \subset S$ and $P(A) > 0$, then

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$ 

### 4.1.1 Exercises

1. Find $P(E^C)$ and $P(O_1 \cap O_2)$ for the copier experiment.
2. Promotions are supposed to be gender-blind. There are 4 male and 3 female district managers named Bill, Mack, Fred, Jim, Gale, Jen, and Lucy. Three positions that would be promotions for any of these managers are open. These 7 individuals are the only candidates to be considered for the 3 positions. The experiment is the filling of the three positions.

   1. List the sample space.
   2. Define a trial and all basic outcomes.
   3. Assuming no gender-bias, find the probability of no females being promoted.
   4. Find the probability of the complement of the event in part 3.
   5. Check: use a computer to randomly draw 3 people from this pool. Record if no females were drawn. Repeat these 2 steps 100 times. Approximate the probability of the event in part 3 by dividing the number of samples that have no females by 100, the number of samples.

### 4.2 Independence

1. If 2 events, $E_1$ and $E_2$ are completely unrelated, i.e., the occurrence of one contains no partial information about whether the other also occurred, the events are said to be independent and satisfy the probability theory definition of independence:

   $$P(E_1 \cap E_2) = P(E_1)P(E_2).$$

   In words, the probability of the intersection of $E_1$ and $E_2$ is equal to the product of their marginal probabilities.

2. Note that this definition implies $P(E_2|E_1) = P(E_2)$, i.e., the occurrence of $E_1$ does not change the chance that $E_2$ also occurred. Use Bayes Theorem to convince yourself of this result.
3. Example: In a nuclear power plant, cooling water stops flowing to the nuclear reactor only if both cooling water pumps fail. This is the design technique of built-in redundancy and is intended to make critical systems more reliable. Let $E_1$ be the event that pump #1 fails, $E_2$ that pump #2 fails, and $A$ the event that no cooling water flows to the reactor. Note that $A = E_1 \cap E_2$.

Pump manufacturer tests indicate that $P(E_1) = P(E_2) = .001$. If pump failures are independent, then

$$P(A) = P(E_1 \cap E_2) = P(E_1)P(E_2) = .001^2 = 10^{-6}$$

or 1 chance in a million.

Perhaps the failure of one pump overloads the other pump. In other words, pump failure is not independent, say

$$P(E_1|E_2) = P(E_2|E_1) = .2$$

or in words, the failure of one pump makes the failure of the other pump 200 times more likely than under the assumption of independence.

In this case,

$$P(A) = P(E_1 \cap E_2) = P(E_1|E_2)P(E_2) = .2 \times .001 = .0002$$

or 2 chances in 10,000 which again, is 200 times more likely than when independence is assumed.

### 4.2.1 Exercise

Three events happening together can ruin a certain insurance company. Each of these events has a 0.01 chance of happening. To be considered a blue chip stock, the chance of ruin must be $< 10^{-5}$.

1. Use Bayes Theorem to show that

$$P(E_1, E_2, E_3) = P(E_1|E_2, E_3)P(E_2|E_3)P(E_3).$$

This is called the recursive factorization of the joint event’s probability.

2. If $P(E_3) = .1$, find the region of values that $P(E_1|E_2, E_3)$ and $P(E_2|E_3)$ can jointly take on.

3. In part 2, what are the worst and best values of the 2 conditional probabilities from the standpoint the CEO of the insurance company?

4. Check: use a computer to approximate $P(E_1, E_2, E_3)$ using the worst case values found in part 2 and compare it to the exact value found using the recursive factorization.
4.3 Random Variables

1. A random variable (r.v.) is a function from the sample space \((S)\) into either integers (a discrete r.v.) or the real numbers (a continuous r.v.):

   **Example:**

   
   ![Diagram](image)

2. If \(A \subset S\) then \(X(A) = x\) means that the value of the r.v. \(X\) when event \(A\) occurs is \(x\).

3. Usually, the event in \(S\) that causes \(x\) is suppressed in textbooks. But every probability statement about a r.v., in full detail looks like:

   \[
P(X = x) = P(X(A) = x) = P(A).
   \]

4. A r.v. is indexing events in \(S\).

5. If there is an infinite number of events in \(S\), then \(X\) can take on an infinite number of different values.

6. A r.v. must associate exactly one of its values for each event in \(S\).

   **Example:** Define a r.v. \(X\) to be the number of good copies before a paper jam occurs. Also, define the experiment to be “making copies until a paper jam occurs.”

   **Note that this new experiment consists of repeated trials.**

   Let \(O_1\) be the event “jam at point 1,” \(O_2\) “jam at point 2,” \(O_3\) “no jam,” and the union of \(O_1\) and \(O_2\), \(E\) be “jam.” Say that \(P(O_1) = P(O_2) = .05\), trials of the “make one copy” experiment are independent, and \(E = O_1 \cup O_2\), as before. The
sample space (the set $S$) is:

$$S = \{E, O_3E, O_3O_3E, O_3O_3O_3E, \ldots\}$$

$$X = \{0, 1, 2, 3, \ldots\}$$

Therefore, $X$ can take on an infinite number of discrete values.

Consider the underlying events to figure out the probability of each value of $X$ occurring:

$$P(X = 0) = P(E) = .05 + .05 = .1$$

$$P(X = 1) = P(O_3, E) = P(O_3)P(E) = .9 \times .1 = .09$$

$$\vdots$$

Note that when the experiment is run, something must happen ($P(S) = 1$) and that something will be exactly one of the $X$ values occurring (because the events that $X$ is indexing are disjoint and exhaustive). Therefore, $\sum_{i=0}^{\infty} P(X = i) = 1$.

The function: $F_X(x) \equiv P(X = x)$ is called the probability mass function (PMF).

In this example, the PMF is:

$$P(X = 0) = .1 = .1$$

$$P(X = 1) = .9 \times .1 = .09$$

$$P(X = 2) = .9 \times .9 \times .1 = .081$$

$$P(X = 3) = .9 \times .9 \times .9 \times .1 = .0729$$

or $P(X = i) = .9^i \times .1$ for $i = 0, 1, 2, \ldots$. This r.v. is said to follow the Geometric distribution with parameter $q = 0.9$. A plot of this function appears in Figure 1.

7. In general, the PMF of a Geometric r.v. is $P(X = i) = q^i(1 - q)$, $i = 0, 1, 2, \ldots$ where $q \in (0, 1)$.

8. **Example:** A Bernoulli r.v. takes on the value 1 with probability $p$ and 0 otherwise. Hence, its PMF is

$$P(X = i) = \begin{cases} 
  p, & i = 1 \\
  1 - p & i = 0
\end{cases}$$
9. The most important function associated with a r.v. is the Cumulative Distribution Function (CDF), defined as: \( F_X(x) \equiv P(X \leq x) \).

For a discrete r.v., \( F_X(x) = \sum_{t \leq x} P(X = t) \).

For a continuous r.v., \( F_X(x) = \int_{-\infty}^{x} f_X(t)dt \) where \( f_X(t) \equiv \frac{dF_X(x)}{dx} \) evaluated at \( x = t \). \( f_X(t) \) is called the probability density function (PDF).

** Only continuous r.v.’s have probability density functions.

10. The basic properties of a CDF are:

(a) \( P(X > x) = 1 - F_X(x) = 1 - P(X \leq x) \)

(b) \( P(a < X < b) = F_X(b) - F_X(a) \)

(c) \( P(X = t) = F_X(t) - \lim_{y \uparrow t} F_X(y) \).

** Property C is called “continuous from the right” and is the only reason the CDF of a discrete r.v. can be used to compute PMF values.

** Some authors use the phrase “probability distribution function” to refer to a PDF and “probability function” to refer to a PMF.

11. The defining constant for the Geometric r.v. is \( q \) and for the Bernoulli, \( p \). These constants that identify a particular member of a distribution family are called parameters.

12. The inverse function of the CDF is called the Inverse CDF or quantile function. For \( p \in (0, 1) \), this function defined to be the smallest \( x_p \) such that \( F_X(x_p) \geq p \). This

![Figure 1: The first 3 values of the Geometric r.v.’s PMF.](image-url)
function is denoted by $F_X^{-1}(p) = Q_X(p) = x_p$. Note that this function’s argument is a probability ($p$) and its return value is some value of the random variable. Only quantile functions of continuous random variables will be studied in this tutorial.

13. The Uniform$(a, b)$ distribution on $(a, b)$ has a PDF $1/(b - a)$ when $x \in (a, b)$ and 0 otherwise.

14. Another continuous probability distribution is the chi-squared with $q$ degrees of freedom. Standard notation for this distribution is: $Y \sim \chi^2_q$. The quantile function for this distribution is notated as $Q_{\chi^2_q}(p)$. One way to find these quantiles for any values of $q$ and $p$ is to use the Minitab statistical software package. For example, to find $Q_{\chi^2_5}(0.26)$, first enter 0.26 in the first row of column 1. Then, click **Calc > Probability distributions > chi-squared**. In the Dialog Box that appears, click **inverse probability**, enter 5 in the degrees of freedom box, and “C1” in the input column box.

15. The quantile function of the standard normal distribution is notated by $\Phi^{-1}(p)$.

### 4.3.1 Exercises

1. Let $X$ be the number of failed parts in a batch of $n$ parts. Describe the event underlying each value of $X$.

2. Say that $X$ is a Geometric r.v. with failure $q = 0.23$. Find $P(X \leq 4)$.

3. An ambulance company has a policy of always replacing an ambulance after 16 time units of use. The engines in these ambulances however, occasionally fail during a call. This engine model’s time-to-failure ($T$) density function up to a constant, $c$ is given by a smooth curve that goes through the points $(1, .1)$ $(2, .3)$ $(4, .5)$ $(6, .7)$ $(8, .4)$ $(11, .2)$ $(14, .1)$ $(15, .05)$ and $(16, 0)$. Use a computer to smooth a curve through these points in order to define the function $cf(t)$ where $t$ is the time-to-failure, and $f(t)$ is the PDF of $T$.

   1. Find $c$.

   2. Find an engine replacement time policy such that the chance of an in-service engine failure is $< .01$.

   3. Find and plot the CDF.

   4. Find and plot the quantile function.
5. Check: for any CDF, \( F_X(x) \), if \( Y \sim \text{Uniform}(0, 1) \) then the r.v. \( X = F_X^{-1}(Y) = Q_X(Y) \) has CDF \( F_X(x) \) (see for example, page 202, Theorem 12 in Mood, A. M, Graybill, F. A., and Boes, D. C. (1974), Introduction to the Theory of Statistics, Third Edition, New York: McGraw-Hill). Use this result to simulate 100 deviates from the time-to-failure distribution and use these simulated deviates to approximate the probability in part 2.

4.4 Expected Value (Mean) and Variance

1. The expected value of a discrete random variable is \( E[X] = \sum_u u P(X = u) \). The variance of a discrete random variable is \( \text{Var}[X] = \sum_u (u - E[X])^2 P(X = u) \).

2. The expected value of a continuous random variable is \( E[X] = \int x f_X(x) \, dx \). The variance of a continuous random variable is \( \text{Var}[X] = \int \left( x - E[X] \right)^2 f_X(x) \, dx \).

3. Example: If \( X \) is a Bernoulli r.v.,
   \[
   E[X] = 1P(X = 1) + 0P(X = 0) = 1p + 0(1 - p) = p.
   \]

4.5 Normal (Gaussian) Distribution

1. The continuous r.v. \( X \) is said to be distributed as normal with mean \( \mu \) and variance \( \sigma^2 \) if the PDF of \( X \) is:
   \[
   f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right], \quad -\infty < x < \infty
   \]
   where \( \sigma \equiv \sqrt{\sigma^2} \) is called the standard deviation.

2. This distribution is notated as: \( X \sim N(\mu, \sigma^2) \).

3. The \textit{Standard Normal} distribution is a r.v. \( Z \) such that \( Z \sim N(0, 1) \). \( Z \)'s PDF, CDF, and quantile function are illustrated below.
Figure 2: Standard normal density function.

Figure 3: Standard normal CDF.

Figure 4: Standard normal quantile function.
4. It is always true for the normal distribution that \( P(X \in \mu \pm \sigma) = .68 \) and \( P(X \in \mu \pm 2\sigma) = .95 \).

**Things to remember:**

1. The distribution is symmetric about the mean, \( \mu \).

2. The areas under the tails of a normal distribution are considered to be “about right” so that the term “heavy-tailed” refers to a distribution that has more tail area than a normal distribution with the same mean.

3. The normal distribution always allows infinitely negative and infinitely large values to be possible – is this a valid assumption for a real business process?

**4.5.1 Exercises**

1. By using several different values for \( \mu \), conduct a simulation experiment to see if the value of \( E[X] \) for \( X \sim N(\mu, 4) \) is \( \mu \).

2. Locate the formula for the variance of a Uniform\((a, b)\) distribution. Check this formula by simulating 100 deviates from a Uniform\((3, 10)\), computing the sample variance (see next Section), and comparing its value to the formula-based value.

**4.6 Approximations**

1. Below are exact and approximate formulas for hand-calculation of \( \chi^2_{\nu, \alpha} \) values, i.e., the \( \alpha \) quantile of the \( \chi^2 \) distribution having \( \nu \) degrees of freedom:

\[
Q_{\chi^2}(\alpha) = \begin{cases} 
 z^2_{(1-\alpha)/2} & \text{for } \nu = 1 \\
 -2 \ln(1 - \alpha) & \text{for } \nu = 2 \\
 \nu \left[ 1 - \frac{2}{9\nu} - z_{\alpha} \sqrt{\frac{2}{9\nu}} \right]^3 & \text{for } \nu > 2
\end{cases}
\]

where \( P(Z > z_{\alpha}) = \alpha \) and \( Z \) is the standard normal random variable. The third equation is most accurate when \( .001 < \alpha < .999 \) and is from Zar, J. H. (1978), “Approximations for the Percentage Points of the Chi-Squared Distribution,” *Applied Statistics*, 27(3): 280-290. This third equation has been modified from its form given in Zar (1978) so as to produce quantile approximations rather than upper-tail critical value approximations.
For example, if \( \nu = 5 \), the third equation gives 1.12799 for the .05 quantile (using 1.645 for \( z_\alpha \)), and 11.044 for the .95 quantile (using -1.645 for \( z_\alpha \)).

2. Hand-calculation formulas for the standard normal distribution are also available. Two particularly simple ones are:

\[
P(Z < z) = \Phi(z) = p \approx 1 - \frac{1}{2} \exp \left[ -\left( z^2 + 1.2z^{0.8} \right) / 2 \right]
\]

for \( z > 0 \), and

\[
z = \Phi^{-1}(p) \approx 0.2 + \left[ p^{0.14} - (1 - p)^{0.09} \right] / 0.1596
\]

5 Basic Statistics

5.1 Fundamental Characteristics and Definitions

1. Statistical statements about precision are all based on the theory of statistical inference which, in turn, is based on the theory of probability.

2. Probability is a branch of mathematics and hence is a form of *deductive reasoning*.

3. Statistical inference attempts to discover characteristics of an unobservable population based only on examination of a few members of the population. Hence, statistical inference is *inductive reasoning*.

4. A *statistic* is a quantity that is computable from a sample.

5. An *estimate* is a data-based guess about the value of a parameter.

6. A *prediction* is a data-based guess about the value of a r.v. that was not observed or before an observation on that r.v. is taken.

7. A statistical estimate is incomplete unless it carries with it a measure of uncertainty. Although the computation of the amount of uncertainty in a statistical estimate is sometimes difficult, providing an uncertainty measure helps a manager decide whether the estimate is too unreliable to be of any use.

8. A *simple random sample* requires that all subjects in a population have the same chance of being selected into a sample and having their personal value on a random variable, $X$ measured. If each subject has the same chance of being selected, then knowledge of one particular subject being selected contains no partial information about whether some other subject will be included in the sample. Hence, the joint distribution of $X$-values over the $n$ subjects of the sample is known: it is the product of the $n$ subject-specific $X$ PDFs or PMFs. These subject-specific $X$ distributions need to all be the same because the analyst is using just one distribution to explain where the sample was drawn from. This need is also met as long as each subject has the same chance of entering the sample. The reason is given in the following algorithm for simple random sampling.

(a) Order the x-values from all $N$ subjects in the population from smallest to largest and mark these values on the horizontal axis of $X$’s CDF.
(b) Mark the associated probability values, $p_1, \ldots, p_N$ associated with each of these $x$-values through the CDF of $X$. For example, the $i^{th}$ ordered $x$-value has associated probability value $p_i = F_X(x_{(i)})$.

(c) Select $n$ of these probability values uniformly. For each of these selected probabilities, find the corresponding $x$-value by traversing the CDF backwards, i.e., by finding the corresponding inverse CDF value.

Note that each probability uniquely identifies a subject. Therefore, instead of uniformly selecting probabilities, it suffices to uniformly select subjects, i.e., select subjects in such a way as to guarantee that all subjects have the same chance of entering the sample.

The set of $x$-values taken from the selected $n$ subjects is a set of $n$ deviates from the distribution $F_X(x)$ due to the Theorem mentioned in problem 5 of Exercises 4.3.1, above.

5.2 Histograms

1. A histogram is a bar chart where the bar *areas* stand for the frequency of a sample’s observations that fall into associated intervals.

2. Example: The salaries of 4 employees of ACME Jet Ski are 25K, 15K, 70K, and 80K.

   You decide to use 3 intervals (or *bins*) to build your histogram. These are: 0-30K, 30K-60K, and 60K-90K. You count how many observations belong to each of these intervals and come up with 2, 0, and 2 respectively.

   You create the histogram:

   Note that since one interval width is always 30K, the bar *areas* are proportional to the frequency. This is important because the eye responds to the bar’s area not just the height of the bar.

   In this example, there is considerable spread or *dispersion* and the *mean* salary, a measure of *central tendency* is not representative of a typical employee at ACME Jet Ski.

3. The shape of the histogram is the most important property to have knowledge of when judging the believability or reliability of subsequent statistical inference performed on the sample.
4. Histogram Terminology:

**Left skew:** there is more dispersion on the left side of the mode, i.e., the left tail is longer than the right tail (note that since the mode is not the median, it is okay that there are more observations to the left side of the mode).

**Right skew:** there is more dispersion on the right side of the mode.

**Bimodal:** Two modes.

**Symmetrical:** A mirror placed on the histogram at the median gives a reflection of one side that is identical to the other side.

5. Most statistical inference procedures assume that the population’s distribution is symmetric and unimodal.

6. Therefore, if the histogram shows skew or multiple modes, this assumption that you are making about the population you are observing may not be true.

7. A *relative frequency* histogram is a bar chart of the relative frequencies by interval instead of the raw frequencies. The sum of all bars then is always 1.0.

8. For ACME Jet Ski, the relative frequency bar heights would be 2/4, 0/4, and 2/4, respectively.
5.2.1 How Many Bins?

The histogram command in Minitab probably uses the rule \# of bins \((m) = \log_2 n + 1\) with bin width \((w)\) given by \(R/m\) (\(R\) is the sample range, see below). This rule assumes that the population’s distribution is normal (defined later) – but the main reason you want to plot a histogram in the first place is to explore the validity of this very strong assumption.

Use of the rule discovered by Freedman and Diaconis allows the histogram to be a good estimator of the PDF. This rule is: \(w = 2IQRn^{-1/3}\) and \(m = R/w\) (see Freedman, D. and Diaconis, P. (1981), “On the Histogram as a Density Estimator: \(L_2\) Theory,” Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57: 453-476 where \(IQR\) is the interquartile range (see below).

5.2.2 Ogive

An ogive plots the sum of all relative frequencies for bin intervals less than and including the current bin interval versus the upper boundary of the current bin. An ogive is an estimate of the population’s CDF.

Both the histogram and ogive are statistics.

5.3 Measures of Central Tendency

1. You have taken a sample: \(x_1, \ldots, x_n\). Three ways to describe this sample’s center are the mean, median, and mode and are defined as follows:

2. Sample Mean: \(\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\).

3. Sample Median: First sort the observations: \(x_1, \ldots, x_n\). Then, the median is \(x_{((n+1)/2)}\) if \(n\) is odd or the average of \(x_{(n/2)}\) and \(x_{(n/2+1)}\) if \(n\) is even. The median is less sensitive to very large or very small (extreme) values in the sample than the mean.

4. Sample Mode: The most frequent value in the sample. This measure is not very useful if \(n\) is small or if there are several different values that occur with the same frequency.
5.4 Measures of Dispersion

1. These statistics quantify the degree of a histogram’s spread. Three common measures are the sample standard deviation, the sample range, and the sample interquartile range and are defined as follows:

2. **Sample Variance:** \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \) which is approximately the mean of the squared deviations from the sample mean.

3. **Sample Standard Deviation:** \( s = \sqrt{s^2} \).

4. \( s \) is more easily interpreted than \( s^2 \) because \( s \) has the same units as individual observations in the sample.

5. **Sample Range:** \( R = x_{(n)} - x_{(1)} \). The range is very sensitive to any extreme observations.

6. **Sample Interquartile Range:** First compute the sample quartiles which are the values for which 25%, 50%, and 75% of the observations are less than. Call these values \( Q_1, Q_2, \) and \( Q_3 \). Note that the median is exactly \( Q_2 \). The interquartile range or \( IQR \) then, is \( Q_3 - Q_1 \).

7. Use the following formulas to find the quartiles: \( Q_1 = x_{([.25n]+1)}, Q_2 = x_{([.5n]+1)}, \) and \( Q_3 = x_{([.75n]+1)} \) where \([r]\) is the floor function – the integer part of the number \( r \).

5.5 Sample Quantiles and Q-Q Plots

1. The quartiles are actually the .25, .5, and .75 sample quantiles. The \( p^{th} \) sample quantile is the value for which 100\(p\)% of the sample is smaller. The quantiles for \( p = .1, .2, .3, \ldots, .9 \) are called the deciles, and the quantiles for \( p = .01, .02, \ldots, .99 \) are called the percentiles.

Hence, \( Q_1 \) is the 25\(^{th} \) percentile, \( Q_2 \) is the 50\(^{th} \) percentile, and \( Q_3 \) is the 75\(^{th} \) percentile.

Use the procedure given for finding quartiles to find percentiles or deciles.

2. You have a random sample of size \( n \) and you wonder if the distribution of the observable r.v.’s is normal. How can you check this assumption?
3. Compare the *expected* quantile values with the observed quantile values.

4. Let $x_{(1)}, \ldots, x_{(n)}$ be the sorted observations from a sample taken on the random variable $X$. $X_{(1)}, \ldots, X_{(n)}$ are called the **order statistics**. Let $k = \lfloor np \rfloor + 1$, $0 < p < 1$ where $\lfloor \cdot \rfloor$ is the floor function, i.e., $[r]$ is the integer part of the real number $r$ (for example $[7.77] = 7$, $[5.23] = 5$). Then $x_{(k)}$ is the $p^{th}$ sample quantile.

5. **Result:** Let $q_p$ be the true $p^{th}$ quantile, i.e., $q_p = F_X^{-1}(p)$. If $k = \lfloor np \rfloor + 1$ and $n$ is large, then $X_{(k)}$ is approximately distributed as normal with expected value equal to $q_p$ and variance equal to $\left(\sqrt{p(1-p)/n}\right)/f_X(q_p)$. See Fisz, M. (1980), *Probability Theory and Mathematical Statistics*, third edition, Huntington, NY: Robert E. Krieger Publishing Co., pp. 374-383.

   This result suggests that we should plot the pairs $(q_p, x_{(i)})$ where $p_i = i/(n+1)$ and $i = 1, \ldots, n$ – and see if there is a one-to-one relationship.

6. **Example:** You have a sample consisting of the values $(-.7, .8, .25, -.3)$ and want to know whether $X_1$’s distribution is standard normal or not.

   **Step 1:** Sort the observations: $(-.7, -.3, .25, .8)$.

   **Step 2:** Compute the $n$ standard normal quantiles: $(-.84, -.25, .25, .84)$.

   **Step 3:** Draw the plot:
Figure 6: Normal Q-Q plot.
5.5.1 Exercise

Simulate a size-100 sample from a $N(5, 2)$ distribution.

1. Compute all statistics presented so far.

2. Derive all population values of these measures under this distribution.

3. Correct any disagreements.

5.6 Distribution of a Sample Statistic

1. Basic Setup in Statistics: Each observation in a randomly chosen sample of size $n$ is a realization from a particular random variable

   \[ X_1, \ X_2, \ldots \ X_n \]
   \[ \downarrow \quad \downarrow \quad \ldots \quad \downarrow \]
   \[ x_1, \ x_2, \ldots \ x_n \]

2. Almost all statistics are derived under the assumption that $X_1, X_2, \ldots, X_n$ each have exactly the same distribution but are pairwise independent, i.e., $X_i$ is independent of $X_j$ for any $i \neq j$. This assumption is concisely stated as “$X_1, \ldots, X_n$” are independently and identically distributed (i.i.d.).

3. A statistic is any function of the observable r.v.’s, $X_1, \ldots, X_n$.

4. The most well-known sample statistic is the sample mean: $\bar{X} = (1/n) \sum_{i=1}^{n} X_i$. Notice that all capital letters are used in this definition, i.e., $\bar{X}$ is a r.v. because it is a function of the observable r.v.’s, $X_1, \ldots, X_n$.

5. Having observed $X_1, \ldots, X_n$ as $x_1, \ldots, x_n$, a realization or observation on $\bar{X}$ can be had by computing $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$. $\bar{x}$ is lower-case because it is an actual number from a real sample.

   ** $\bar{x}$ is not a r.v., $\bar{X}$ is a r.v..

6. There are many sample statistics. Some are:

   a) Sample variance: $S^2 = (1/(n-1)) \sum_{i=1}^{n} (X_i - \bar{X})^2$

   b) Range $= X_{(n)} - X_{(1)}$

   c) Minimum $= X_{(1)}$
d) Maximum = $X_{(n)}$.

7. **Example**: Sample = \{2, 9\}. Then $\bar{x} = 5.5$ and

$$s^2 = \frac{1}{2-1} \left[ (2-5.5)^2 + (9-5.5)^2 \right]$$

$$= 24.505$$

and $s = 4.95$.

### 5.7 Central Limit Theorem

1. The sample mean statistic enjoys a property that is perhaps the most important result in statistics:

   If $n$ is very large, the true mean of $X_i$, $\mu_X$ is finite, the true variance of $X_i$, $\sigma^2_X$ is finite, and $X_1, \ldots, X_n$ are i.i.d.; then the distribution of $\bar{X}$ is approximately $N(\mu_X, \sigma^2_X/n)$ no matter what the distribution of $X_1$ is.

2. This statement implies that even if $X_1, \ldots, X_n$ are i.i.d. Bernoulli r.v.’s, as long as $n$ is large enough, $\bar{X}$ will have an approximately normal distribution.

3. This result “saves” alot of statistical analyses because even if the observable r.v.’s are not normally distributed, $\bar{X}$ may still be close to normal – thereby allowing conclusions that depend on the normality assumption to be viewed as reliable.

4. Remember that the Theorem does not claim that $\bar{X}$ will be exactly normal for a big enough $n$ – just approximately. For example, if $P(X_1 < 0) = 0$, then no matter how large $n$ is, $\bar{X} \geq 0$ and hence the distribution of $\bar{X}$ cannot be exactly normal because every normal distribution assigns some nonzero probability to negative values.

#### 5.7.1 Exercise

Simulate a size-100 sample from a Bernoulli(0.1) distribution. Compute $\bar{x}$ from this sample. Repeat these 2 steps 863 times. Construct the histogram of these 863 $\bar{x}$ values. What distribution’s PDF does this histogram look like?

### 5.8 Parameter Estimation

1. Recall that a parameter is a constant that is used to specify a particular probability distribution.
2. Parameter estimation is one aspect of statistical inference and is used to aid business decision making as follows:

(a) A business decision is influenced by a distribution’s parameter value.
(b) A sample is collected and then used to compute an estimate of the unknown parameter’s value.
(c) This estimate is used in place of the true but unknown parameter value to help make the business decision.

3. Example: If $X \sim N(\mu, \sigma^2)$, then $\mu$ and $\sigma^2$ are parameters. If $X \sim \text{Binomial}(n, p)$, then $n$ and $p$ are parameters.

4. Statistical inference uses a collection of different tools to make inferences about the probability distribution that generated the sample in-hand.

5. Example: You want to know if the new Beetle will sell well in the U.S. You estimate the proportion of people in the U.S. that will buy the Beetle by doing the following:

**Step 1:** You ask $n$ randomly chosen people if they will buy the new Beetle.

**Step 2:** Say that $m$ ($< n$) of these people said “yes.” You estimate the chance that the next randomly chose person will buy a Beetle with:

$$\hat{p} = \frac{m}{n}$$

** When you see the “hat” on top of a parameter, the parameter is being estimated with the statistic on the right hand side of the “=” sign. In this example, $m/n$ is the statistic being used to estimate the true but unknown value of $p$ where

$$p \equiv P(\text{randomly chosen person will buy a new Beetle})$$

To use actual numbers, say that only 10 people out of your sample of 100 people said they would buy a Beetle. Then $\hat{p} = .1$.

Reality Check: You just estimated that 10% of the U.S. population would buy a new Beetle – is this too good to be even close to the truth?

6. A statistic, being a random variable has a probability distribution. Often, a statistic’s distribution is complicated and difficult to find. Finding only the expected value (mean) and standard error (square root of the variance) is usually easier.
7. **Example:** If $X_1, \ldots, X_n$ are such that $E[X_i] = \mu_X$ for all $i$, then the expected value of the sample mean statistic, $\bar{X}$ is:

$$
E[\bar{X}] = E \left[ \frac{1}{n} \sum_{i=1}^{n} E[X_i] \right]
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} E[X_i]
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \mu_X
$$

$$
= \mu_X.
$$

In words, the expected value of $\bar{X}$ is $\mu_X$, the true mean of $X_i$. This is good because analysts use $\bar{X}$ to estimate $\mu_X$.

8. A statistic is said to be **unbiased** if the expected value of the statistic equals the parameter that it is being used to estimate.

9. **Example:** Variance of sample mean. Say that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.). Let the common variance of these random variables be $\sigma_X^2 < \infty$, i.e., $\text{Var}[X_1] = \ldots = \text{Var}[X_n] = \sigma_X^2$. Then,

$$
\text{Var}[\bar{X}] = \text{Var} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right]
$$

$$
= \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i]
$$

$$
= \frac{1}{n^2} n \sigma_X^2
$$

$$
= \frac{\sigma_X^2}{n}
$$

** So, the variance of $\bar{X}$ is going to 0 as $n$ becomes large ($n \to \infty$) – regardless of the true value of $\sigma_X^2$.

** This means that $\bar{X}$ becomes very efficient for estimating the true mean as your sample size, $n$ gets very large.

10. If you estimate $\mu_X$ with an observed value of $\bar{X}$ (denoted $\bar{x}$), you have computed a **point estimate** of $\mu_X$.

11. As mentioned above, it is usually difficult to find the probability distribution of a statistic. For many statistics however, an **approximation** based on the Central Limit Theorem can be made when the sample size, $n$ is large.
Example: $\bar{X}$ is approximately $\sim N(\mu_X, \sigma^2_X/n)$ as $n \to \infty$.

** Because of this result, if you have taken a large sample, your computed value of $\bar{x}$ will almost certainly be very close to the true value of $\mu_X$.

** This is a good thing because you can be fairly certain that the value of $\bar{x}$ is a reliable and accurate estimate of $\mu_X$.

5.9 Confidence Interval for $\mu$

1. For $Z \sim N(0, 1)$, denote the $1 - \alpha/2$ quantile with $z_{\alpha/2}$, i.e., $P(Z > z_{\alpha/2}) = \alpha/2$.

2. A $100(1 - \alpha)\%$ Confidence Interval (C.I.) for $\mu$ is derived with:

\[
1 - \alpha = P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma_X} < z_{\alpha/2})
= P(-\sigma_X z_{\alpha/2} + \bar{X} < \mu < \sigma_X z_{\alpha/2} + \bar{X}).
\]

$\mu$ is NOT random so for a given $\bar{x}$ (computed from a sample), you say “with confidence $100(1 - \alpha)\%$, $\mu \in \bar{x} \pm \sigma_X z_{\alpha/2}$ which is equal to $\mu \in \bar{x} \pm (\sigma_X/\sqrt{n})z_{\alpha/2}$ because $\sigma^2_X = \sigma^2_X/n$.

3. Another important statistic is the sample variance:

\[
S^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]

First, define a new probability distribution, the chi-squared with $q$ degrees of freedom. Standard notation for this distribution is: $Y \sim \chi^2_q$. Result: $E[Y] = q$.

Result: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu_X, \sigma^2_X)$, then $(n-1)S^2/\sigma^2_X \sim \chi^2_{n-1}$. Then we can say: $E[(n-1)S^2/\sigma^2_X] = n-1$ which implies $((n-1)/\sigma^2_X)E[S^2] = n-1$ because constants can be taken outside the integral. Therefore, $E[S^2] = \sigma^2_X$, i.e., $S^2$ is unbiased for $\sigma^2_X$.

A more complicated derivation shows: $\text{Var}[S^2] = 2\sigma^4_X/(n-1)$.

5.9.1 Exercise

Simulate a size-21 sample from a $N(2, 4)$ distribution. Compute the 80% C.I. for the mean using this sample. Repeat these steps 100 times. What percentage of these C.I.s contain the true mean 2? What should this percentage be?
5.10 Tolerance Intervals

1. A Tolerance Interval is an interval computed from a sample that contains $100(1-p)\%$ of the population with confidence level $100(1-\alpha)\%$. For given values of $p$ and $\alpha$, algorithms are available to compute the interval when the population is normally distributed.

2. Tolerance Intervals will not be discussed further in this Tutorial.

6 Hypothesis Testing

6.1 Definitions and Procedure

1. After parameter estimation, the second activity in statistical inference is hypothesis testing.

2. A scientific hypothesis is a researcher’s belief in how some phenomenon works and is usually a complicated collection of statements. A statistical hypothesis however, is much simpler. It is simply a statement about a parameter’s true (but unknown) value.

3. There are 2 kinds of statistical hypotheses: a null hypothesis, designated by $H_0$, and an alternative hypothesis, designated by $H_a$, $H_A$, or $H_1$.

These hypotheses are always concerning an unknown parameter, say $\theta$ and are written as: $H_0 : \theta = \theta_0$ (the null hypothesis is that the value of $\theta$ is $\theta_0$).

*** It is up to YOU to specify the hypothesized value, $\theta_0$.

and,

$H_1 : \theta \neq \theta_0$ (the alternative hypothesis is that $\theta$ does not equal $\theta_0$).

4. Example: $H_0 : \theta = 5$ versus $H_a : \theta \neq 5$.

5. There are 3 types of alternative hypotheses:

   2-sided, $H_a : \theta \neq \theta_0$,
   lower-tailed, $H_a : \theta < \theta_0$, and
   upper-tailed, $H_a : \theta > \theta_0$.

6. Example: You ask $n$ randomly-chosen people if they plan to buy a new Beetle in the next 6 months. Your research hypothesis is that less than half of the population
wants to buy a new Beetle. You represent each person’s answer with the value 0 if “no,” and 1 if “yes.” You have therefore observed $n$ Bernoulli r.v.’s. The probability of an answer being “yes” is $p$. Mathematically, you have observed $X_1, \ldots, X_n \sim i.i.d.$ Bernoulli($p$) r.v.’s and want to decide if the true value of $p$ is .5 or less than .5. If $p$ isn’t .5, you believe that the only other reasonable value for $p$ is some value less than .5. Your hypotheses are:

\[ H_0 : p = .5 \text{ versus } H_a : p < .5. \]

7. Notice that $H_0$ defines a single point value that $\theta$ can be; whereas $H_a$ defines an interval of hypothesized values.

8. To perform a hypothesis test, you do the following:

**Step 1:** Define $H_0$ and $H_a$ for the parameter you are interested in.

**Step 2:** Decide on how much risk you are willing to tolerate of making the mistake of rejecting $H_0$ when $H_0$ is true:

\[ \alpha = P(\text{rejecting } H_0 | H_0 \text{ is true}) \]

(the “|” symbol means “given”).

**Step 3:** Collect a sample of size $n$, compute the test statistic, and then see if the test statistic’s value falls in the critical region. If it does, decide that $H_0$ is false. If not, decide that there is not sufficient evidence to conclude that $H_0$ is false.

*** Note that this is NOT the same as saying $H_0$ is true!

*** Just because you decide that $H_0$ is false doesn’t necessarily mean that it actually is false – you may have made a mistake. On the other hand, if you fail to reject $H_0$ when $H_0$ really is false, you have made a different type of mistake.

*** These 4 outcomes are all the things that can happen when you conduct a “yes-no” or “reject-do not reject” hypothesis test.

9. Let $\theta_a$ be some value in the alternative hypothesis interval.

**Example:** If $H_0 : p = .5$ versus $H_a : p \neq .5$, then $p_a = .2$ is in $H_a$’s lower interval of $(0, .5)$. 

30
Your decision | State of Nature
---|---
You decide to reject $H_0$ | Type I error | no mistake
You decide there is not sufficient evidence to conclude $H_0$ is false. | no mistake | Type II error

Table 1: Possible decisions and mistakes when performing a hypothesis test.

10. Remember that $\alpha$ is the *chance* that you will commit a Type I error. In a similar manner, $\beta(\theta_a)$ is the *chance* that you will commit a Type II error when the true value of $\theta$ is $\theta_a$. We write:

$$
\beta(\theta_a) = P(\text{you fail to reject } H_0 | \theta = \theta_a \in H_a \text{ is true}).
$$

11. Like everything in life, there are *consequences* for making mistakes. Say that you can calculate the *cost* of making a Type I error and that this cost is $c_I$ dollars. Likewise, for $\theta_a \in H_a$, $c_{II}(\theta_a)$ is the *cost* of committing a Type II error when the true value of $\theta$ is $\theta_a$.

12. **Example:** Say that the Food and Drug Administration (FDA) will jerk a new headache tablet from the market if more than 10% of the users have an adverse reaction.

You have spent lots of money developing this product and would lose all this money if you don’t market the drug. On the other hand, your company’s reputation could be permanently damaged if the FDA pulls one of your products.

Your hypothesis test is to decide if $p = P(\text{a user will have an adverse reaction}) < .1$ versus $p \geq .1$. You represent these hypotheses with: $H_0 : p = .1$ versus $H_a : p > .1$.

You predict your costs to be:

$c_I = \text{cost of rejecting } H_0 \text{ when } H_0 \text{ is true}$

$= \text{cost of not marketing the (actually) safe-enough product}$

$= \$1 \text{ million}$.

whereas,

$c_{II}(.3) = \text{cost of failing to detect that the true value of } p \text{ is } .3$

$= \text{cost of having the FDA accuse your firm of placing a drug on the market that}$
made 30% of the users sick
= $10 million (due to reduced stock value, loss of sales, etc.).

Here, the cost of a Type II error at \( p = .3 \) is 10 times as expensive as the cost of a Type I error.

13. What most statistics texts fail to mention is how to select a value for \( \alpha \) that will allow you to minimize the costs of making either a Type I or Type II error.

One procedure that sets \( \alpha \) so that these costs are minimized is as follows.

Case 1: You’ve already collected your sample of size \( n \).

**Step 1:** Decide on a plausible point in the alternative hypothesis space that is minimally different from \( \theta_0 \), call this point \( \theta_a \). This value should be the smallest value of \( \theta \) that is practically or scientifically significantly different from \( \theta_0 \).

**Step 2:** Using your knowledge of the consequences of committing a Type I error or a Type II error, decide on a value for the ratio \( c_{II}(\theta_a)/c_I \).

**Note that you do not need to know actual dollar figures to do this – just the relative magnitudes.**

**Step 2:** Use the test’s power curve (see below) to find the smallest \( \alpha \) and \( \beta(\theta) \) values such that:

\[
\frac{\alpha}{\beta(\theta_a)} = \frac{c_{II}(\theta_a)}{c_I}.
\]


In the above Example, because \( c_{II}(.3) \) is 10 times as big as \( c_I \), you want the chance of a Type II error to be one-tenth the chance of a Type I error. Therefore, \( \alpha = 10\beta(\theta_a) \) for the Example.

By following this procedure, you can make a decision to either reject or not reject \( H_0 \) and be confident that the risks that you run in making either type of error are in proportion to the costs of those errors.

Case 2:
You are asked to determine a sample size, \( n \) that will minimize the total cost of performing the hypothesis test.

**Step 1:** There are 3 costs: the cost of committing a Type I error, the cost of a Type II error, and the cost of collecting the sample. Therefore, the total cost is:

\[
E[C] = c_I \alpha + c_{II}(\theta_a) \beta(\theta_a) + c_{\text{sample}}
\]

assuming equal chances of both \( H_0 \) being true and the point in \( H_A \) being true. In this Case then, you do need to know actual dollar figures.

**Step 2:** Using the power curve again, find \( n, \alpha, \beta(\theta_a) \) such that \( E[C] \) is minimized.

*** Note that \( n, \alpha, \text{ and } \beta(\theta) \) are not independent of each other, i.e., if you know any 2 of these, you can compute the third.

So actually, just hunt for \( n \) and \( \alpha \) that will minimize \( E[C] \) (this is equivalent to doing Case 1 for different values of \( n \)).

14. In general, \( n \uparrow \) implies \( \alpha, \beta(\theta_a) \downarrow \).

For a fixed \( n \), \( \alpha \downarrow \) implies \( \beta(\theta_a) \uparrow \) and \( \alpha \uparrow \) implies \( \beta(\theta_a) \downarrow \).

*** \( \alpha + \beta(\theta_a) \) DO NOT have to equal 1.0 !!! ***

15. Because \( \beta(\theta_a') < \beta(\theta_a'') \) when \( |\theta_a' - \theta_0| > |\theta_a'' - \theta_0| \), all you really need to do is specify the smallest deviation from \( \theta_0 \) that you want to be able to detect (smallest practical or scientifically significant deviation). Use this smallest deviation to determine the \( \theta_a \) value to use in computing \( \alpha \) and \( \beta(\theta_a) \) for which

\[
\frac{\alpha}{\beta(\theta_a)} = \frac{c_{II}(\theta_a)}{c_I}
\]

By following this procedure, you are assured of performing the experiment in such a manner that minimizes all three sources of cost: cost of collecting the data, cost of committing a Type I error, and the cost of committing a Type II error.

*** Note that

\[
\beta(\theta_0) = P(\text{not rejecting } H_0|H_0 \text{ is true}) = 1 - \alpha
\]

This means that an upper bound on \( \beta(.) \) is always \( \beta(\theta_0) \).
16. The *power* of a hypothesis test is the test’s ability to detect the fact that $\theta = \theta_a$ when conducted at the level $\alpha$:

$$
\text{power}(\theta_a) = P(\text{rejecting } H_0 | \theta = \theta_a \in H_a) = 1 - \beta(\theta_a).
$$

17. **Example continued:** Referring to the drug testing example, remember that $H_0 : p = .1$ versus $H_a : p > .1$. You collect a sample of 10 users and record a “1” if a user got sick and a “0” if the user did not get sick.

You find that 3 of the 10 users got sick.

To perform the test, you do the following:

**Step 1:** Because $n$ is not very big, you decide to use the sum of the observed Bernoulli variables as your test statistic: $T(X_1, \ldots, X_n) = T(X_i) = \sum_{i=1}^{n} X_i$.

**Step 2:** You look up the distribution of $T(X_i)$ when $H_0$ is true and find it to be Binomial($n = 10$, $p = .1$).

**Step 3:** You find the set of values of $T(X_i)$ that are *unlikely* to happen if $H_0$ is true but *likely* to happen if some point in $H_a$ is true. By “unlikely” you mean “with probability $\alpha$.”

<table>
<thead>
<tr>
<th>$t(x_i)$</th>
<th>$P(T(X_i) = t(x_i))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.3487</td>
</tr>
<tr>
<td>1</td>
<td>.3874</td>
</tr>
<tr>
<td>2</td>
<td>.1937</td>
</tr>
<tr>
<td>3</td>
<td>.0574</td>
</tr>
<tr>
<td>4</td>
<td>.0112</td>
</tr>
<tr>
<td>5</td>
<td>.0015</td>
</tr>
<tr>
<td>6</td>
<td>.0001</td>
</tr>
<tr>
<td>7-10</td>
<td>.0000</td>
</tr>
</tbody>
</table>

Table 2: PMF of test statistic under $H_0 : p = .1$. 


Figure 7: PMF and rejection region for adverse-effect hypothesis test.
Say that you have arbitrarily set $\alpha$ to .05. To find your upper-tail critical region, you need to find a value $t$ such that $P(T \geq t|H_0$ is true) = .05 but when you use the Minitab commands:

MTB> cdf;
SUBC> binomial 10 .1.

you discover that $P(T \geq 3) = .0702$ and $P(T \geq 4) = .0128$.
So, you decide to be conservative and reject $H_0$ if $T(X_i)$ is observed to be the value 3 or greater.

**Step 4:** Your observed value of $T(X_i)$ is 3. This value is in the critical region so you reject $H_0$ and decide to not market the new drug.

### 6.2 The z-test for hypotheses about $\mu$

1. Say that you observe $X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$ where you happen to know the value of $\sigma^2$. You want to test a set of hypotheses about $\mu$.

2. The best test statistic to use is the z-transformed sample mean:

$$T(X_i) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

where $H_0 : \mu = \mu_0$.

Then, if $H_0$ is true, $T(X_i) \sim N(0,1)$.

3. **Example:** you want to test $H_0 : \mu = 4$ versus $H_a : \mu \neq 4$. Do the following:

**Step 1:** You set $\alpha$ to .05. the critical region will be the set of values of $T(X_i)$ that are unlikely if $\mu = 4$ but likely if $\mu \neq 4$. Since $T = (\bar{X} - 4)/(\sigma/\sqrt{n})$, if $\mu = E[\bar{X}]$ is not 4, then $T(X_i)$ will usually be either negative (for $\mu < 4$) or positive (for $\mu > 4$). Therefore, the critical region will actually be 2 open intervals:

**Step 2:** The z-test has maximum power when $|c_1| = |c_2|$ (symmetric). So, from the Normal Table, you find $c_1 = -1.96$ and $c_2 = 1.96$.

Your decision rule then is to reject $H_0$ if the observed value of $T(X_i)$, written $t(x_i)$ is either $< -1.96$ or $> 1.96$ since in either case, $t(x_i)$ will be in the critical region.
Step 3: Now, you collect data and find that $n = 16$, $\bar{x} = 5.0$ and you know already that $\sigma^2 = 25$. Then,

$$t = \frac{\bar{x} - 4}{5/4} = \frac{5.0 - 4.0}{5/4} = .8.$$ 

Step 4: You note that .8 is not in the critical region so you do not reject $H_0$.

Say that instead of the 2-sided test, you wanted to perform a 1-sided test: $H_a : \mu > 4$. Now, the critical region is only 1 interval – the positive side of $T(X_i)$.

For example, if $\mu$ were actually 10, $E[T|H_0$ is true] = $E[(\bar{X} - 4)/(5/4)]$ which equals $(4/5)(10 - 4)$ or 4.8 which is much bigger than 0.

This new, 1-sided critical region is $c = \{t : t > 1.645\}$ and makes the 1-sided test more sensitive to $\mu$ being $> 4$ simply because $1.645 < 1.96$ (more opportunities to reject $H_0$).

On the other hand, the test is now completely blind to the possibility of $\mu$ being less than 4.

![Critical regions for the 2-sided z-test at $\alpha = .05$.](image)
Figure 9: A 1-sided z-test critical region (upper-tail). The region’s left boundary is at 1.645 which is the $1 - \alpha$ quantile.
6.3 P-values

1. Say that you don’t like the idea of being forced to either reject \( H_0 \) or not reject \( H_0 \) but just want to report how well the data agrees with the null hypothesis. The \( p \)-value is one way to measure such agreement.

2. Definition: The \( p \)-value is the probability of obtaining the observed value of the test statistic or a value even less likely when \( H_0 \) is true.

3. The idea is to report how typical the observed value of the test statistic is if \( H_0 \) is true.

   For a 1-sided test, the \( p \)-value is the area under the PDF starting from the observed value of the test statistic and going into the critical region (even less likely values of the test statistic under \( H_0 \)). For a 2-sided test, the observed test statistic value and its negative are used to find the \( p \)-value, see the Figure.

4. If the test rejects at \( \alpha \), then the \( p \)-value is \(< \alpha \).

5. Note that the \( p \)-value is actually a statistic because it depends on the test statistic’s value. Hence, you can’t compute the \( p \)-value until after you have collected a data set.

\[
p\text{-value} = \begin{align*}
1 - F_T(t(x_i)), & \quad (1\text{-sided test (upper tail)}) \\
F_T(t(x_i)), & \quad (\text{lower-tailed test}) \\
2F_T(t(x_i)), & \quad (t(x_i) < 0, \ 2\text{-sided test}) \\
2(1 - F_T(t(x_i))), & \quad (t(x_i) > 0, \ 2\text{-sided test})
\end{align*}
\]
6. **Example:** Calculating $\beta(\theta_a)$ for the z-test:

Let $X_1, \ldots, X_{25} \sim \text{i.i.d. } N(\mu, 9)$. $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. You decide to use $\alpha = .01$. You think that if $H_0$ is false, $\mu$ might plausibly be 1.0.

Under $H_0$, the test statistic is:

$$T(X_i) = \frac{\bar{X} - 0}{3/5} = \frac{\bar{X}}{3/5} = \frac{5}{3} \bar{X},$$

but since the true value of $\mu$ under the considered point in $H_a$ is $\mu = 1$, the test statistic's distribution is shifted:

![Figure 11: Shift in $T(X_i)$'s distribution under $H_a$. All area under the shifted PDF and the arrow's shaft gives $\beta(1)$.

$$E[T(X_i)|\mu = 0] = 0 \text{ but } E[T(X_i)|\mu = 1] = E[(\bar{X} - 0)/(3/5)] \text{ which equals } 5/3.$$\]

$$\beta(1.0) = P(\text{not rejecting } H_0|\mu = 1 \in H_a \text{ is true})$$

$$= P(T(X_i) < 2.33|\mu = 1)$$

$$= P(\frac{\bar{X} - 0}{3/5} < 2.33|\mu = 1)$$

$$= P(\bar{X} < 2.33(3/5)|\mu = 1) \text{ (first isolate } \bar{X})$$

$$= P(\bar{X} < 1.398|\mu = 1)$$

$$= P(\frac{\bar{X} - 1}{3/5} < \frac{1.398 - 1}{3/5}|\mu = 1)$$

(40)

$$= P(Z < .663|\mu = 1)$$

$$= .74$$

and hence $\text{power}(1) = 1 - \beta(1) = 1 - .74 = .26.$
Therefore, there is only a 26% chance that the test will be able to detect the truth when the truth is that $\mu = 1$.

7. Power means that the test “does the right thing” which is rejecting $H_0$ when $H_0$ is false.

8. In the example, $P(\text{Type I error}) = .01$ but $P(\text{Type II error}) = .74$. Given these probabilities, the use of $\alpha = .01$ implies that the cost of a Type I error is 74 times as expensive as the cost of a Type II error.

9. The need to balance $\alpha$ with $\beta(\theta_a)$ becomes very important when $n$ is large. For example, if $\alpha$ is fixed at (say) .05, then, as $n \to \infty$, $\text{power}(\theta_a) \uparrow 1.0$. In other words, unless $\alpha$ is adjusted down for a large sample size, an $n$ can be found such that the test will almost certainly reject $H_0$ no matter what the value of $\theta_a$ (the truth), e.g. even when $|{(\theta_0 - \theta_a)}/\theta_0| << .00001!!$

6.4 Power Curves

1. A power curve is simply the plot of $\text{power}(\theta)$ over all values of $\theta$. The so-called “Operating Characteristic” curve or $OC$ curve is equal to 1 minus the power curve.

2. Example continued: Previously, the power for $\mu_a = 1.0$ was found to be .26. The general formula for any $\mu_a \in H_a$ is:

$$\text{power}(\mu_a) = 1 - \beta(\mu_a)$$

$$= 1 - P(\frac{\bar{X} - 0}{3/5} < 2.33| \mu = \mu_a)$$

$$= 1 - P(\bar{X} < 1.398| \mu = \mu_a)$$

$$= 1 - P(\frac{\bar{X} - \mu_a}{3/5} < \frac{1.398 - \mu_a}{3/5}| \mu = \mu_a)$$

$$= 1 - P(Z < 2.33 - \frac{5}{3}\mu_a| \mu = \mu_a)$$

Say that you collected a sample of size 100.

$$\text{power}(\mu_a) = 1 - P(\frac{\bar{X} - 0}{3/10} < 2.33| \mu = \mu_a)$$

$$= 1 - P(\bar{X} < .699| \mu = \mu_a)$$

$$= 1 - P(\frac{\bar{X} - \mu_a}{3/10} < \frac{.699 - \mu_a}{3/10}| \mu = \mu_a)$$

$$= 1 - P(Z < 2.33 - 3.33\mu_a)$$
<table>
<thead>
<tr>
<th>$\mu_a$</th>
<th>$power(\mu_a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1 - P(Z &lt; 2.33) = 1 - .99 = .01 = \alpha$</td>
</tr>
<tr>
<td>.5</td>
<td>$1 - P(Z &lt; 1.49) = 1 - .933 = .067$</td>
</tr>
<tr>
<td>1</td>
<td>$.26$ (from previous)</td>
</tr>
<tr>
<td>1.5</td>
<td>$1 - P(Z &lt; -1.17) = 1 - .433 = .567$</td>
</tr>
<tr>
<td>2</td>
<td>$1 - P(Z &lt; -1.0) = .841$</td>
</tr>
<tr>
<td>3</td>
<td>$1 - P(Z &lt; -2.67) = .996$</td>
</tr>
<tr>
<td>4</td>
<td>$1 - P(Z &lt; -4.33) \approx 1.0$</td>
</tr>
</tbody>
</table>

Table 3: Power calculations for the Example.

<table>
<thead>
<tr>
<th>$\mu_a$</th>
<th>$power(\mu_a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.01</td>
</tr>
<tr>
<td>.5</td>
<td>$1 - P(Z &lt; .665) = .253$</td>
</tr>
<tr>
<td>1.0</td>
<td>$1 - P(Z &lt; -1) = .841$</td>
</tr>
<tr>
<td>1.5</td>
<td>$1 - P(Z &lt; -2.665) = .9962$</td>
</tr>
<tr>
<td>2.0</td>
<td>$1 - P(Z &lt; -4.33) \approx 1.0$</td>
</tr>
</tbody>
</table>

Table 4: Power calculations when $n = 100$.

3. In general, the power curve becomes steeper near $\theta_0$ as $n$ gets large.

Figure 12: Power curves for $n = 25$ and $100$. 

42
6.4.1 Exercise

1. Find a general formula for the Type II error rate of a 1-sided, lower-tailed z-test. Your answer should be a standard normal probability lookup of a constant that is a function of $\mu_0$, $\sigma^2$, $\mu_a$, and $\alpha$.

6.5 A test for $\mu$ when $\sigma^2$ is unknown: the $t$-test

1. The test statistic is:

$$T(X_i) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

where $t_{n-1}$ is the Student’s $t$ distribution with $n-1$ degrees of freedom (d.o.f.). Some percentile values are given statistics texts. Remember that $S^2 =$ sample variance $= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

2. For a 2-tailed alternative hypothesis, the critical region looks very similar to that of the $z$-test:

![Figure 13: 2-tailed t-test critical region. The sample size is 6. The standard normal PDF is super-imposed for comparison purposes.](image-url)
6.6 Tests for Paired Data

1. Say that you have measured the same subject before and after some “treatment” was applied to the subject. You did this to \( n \) subjects and want to see if the mean response has changed due to the treatment.

2. In this case, you can’t use the 2-sample t-test because the assumption that the two populations (here, before and after) are independent is false because you went back to the same subjects after the treatment. At least you have independence across subjects – what to do?

3. Create a new variable called the “difference”: \( D_i = Y_{1i} - Y_{2i}, \ i = 1, \ldots, n \) and test for a zero mean using a 1-sample t-test.

   **Note:** if \( Y_{11}, \ldots, Y_{1n} \sim N(\mu_1, \sigma_1^2) \) and \( Y_{21}, \ldots, Y_{2n} \sim N(\mu_2, \sigma_2^2) \), then \( D_1, \ldots, D_n \sim \text{i.i.d. } N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - \text{Cov}(Y_{11}, Y_{21})). \)

4. \( \text{Cov}(Y_{11}, Y_{21}) \) is the *covariance* between the random variables \( Y_{11} \) and \( Y_{21} \). Covariance is a measure *association* between two random variables and is defined as:

\[
\text{Cov}(Y_{11}, Y_{21}) \equiv \mathbb{E}[Y_{11}, Y_{21}] - \mu_1\mu_2.
\]

5. Fortunately, you don’t need to estimate \( \text{Cov}(Y_{11}, Y_{21}) \) because all you need is an estimate of \( \sigma_D^2 \equiv \sigma_1^2 + \sigma_2^2 - \text{Cov}(Y_{11}, Y_{21}) \) and *that* is provided by the sample variance of the \( D_i \) observations:

\[
S_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2
\]

6. So, all you have to do is compute

\[
T(D_i) = \frac{\bar{D} - \mu_D^{(0)}}{S_D/\sqrt{n}}
\]

which is distributed as \( t_{n-1} \) when the null hypothesis that \( \mu_1 - \mu_2 = \mu_D^{(0)} \) is true.

7. **Example:** You want to know if the Tylenol-tampering deaths is associated with the sales of Tylenol. You collect the number of Tylenol sales from 5 local drugstores before and after the Tylenol deaths:

\[
\text{MTB} > \text{set c1}
\]
\[
\text{DATA}> 151 200 40 67 119
\]
<table>
<thead>
<tr>
<th>Drugstore</th>
<th>Before incident sales</th>
<th>After incident sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>151</td>
<td>110</td>
</tr>
<tr>
<td>B</td>
<td>200</td>
<td>207</td>
</tr>
<tr>
<td>C</td>
<td>40</td>
<td>32</td>
</tr>
<tr>
<td>D</td>
<td>67</td>
<td>71</td>
</tr>
<tr>
<td>E</td>
<td>119</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 5: Drugstore data.

DATA> end
MTB > set c2
DATA> 110 207 32 71 50
DATA> end
MTB > let c3 = c1 - c2
MTB > ttest 0 c3;
SUBC> alternative 1.

TEST OF MU = 0.0 VS MU G.T. 0.0

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>MEAN</th>
<th>STDEV</th>
<th>SE MEAN</th>
<th>T</th>
<th>P VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3</td>
<td>5</td>
<td>21.4</td>
<td>32.7</td>
<td>14.6</td>
<td>1.46</td>
<td>0.11</td>
</tr>
</tbody>
</table>
6.7 Testing for Equal Variances

1. Back to the case of two independent samples. Just how different do $\sigma_1^2$ and $\sigma_2^2$ have to be before you should switch to the approximate t-test?

2. One way to answer this question is to test for whether the two variances are equal as follows. Let $H_0: \sigma_1^2 = \sigma_2^2$ versus either $H_A: \sigma_1^2 > \sigma_2^2$ or $H_A: \sigma_1^2 \neq \sigma_2^2$.

   The test statistic is: $T(Y_{1i}, Y_{2i}) \equiv S_1^2 / S_2^2$ and is distributed as $F_{n_1-1, n_2-1}$ when $H_0$ is true.

3. $F_{\nu_1, \nu_2}$ is the $F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom. Again, Sir Ronald Fisher invented this distribution and hence is honored by the “F” designation.

4. Probabilistically, the $F$-distribution is the ratio of two independent $\chi^2$ random variables, each divided by their respective degrees of freedom:

   $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$ and $X_1, X_2$ are independent, then $F \equiv (X_1/\nu_1)/(X_2/\nu_2) \sim F_{\nu_1, \nu_2}$

5. The $F$ test is derived from the Theorem: If $X_1, \ldots, X_{n_1}$ i.i.d. $\sim N(\mu_X, \sigma^2)$ and $Y_1, \ldots, Y_{n_2}$ i.i.d. $\sim N(\mu_Y, \sigma^2)$ and the $X$’s and independent of the $Y$’s, then $S_X^2 / S_Y^2 \sim F_{n_1-1, n_2-1}$.
7 Basic Regression Analysis

7.1 Definitions and Fundamental Characteristics

1. Due to historical reasons, when a statistician models the relationship between a dependent variable, $Y$ and a continuous explanatory variable, $X$ the terminology is that $Y$ is being regressed onto $X$. The statistical method for doing this is called regression.

2. $X$ is also referred to as the independent variable or simply the regressor.

3. The word “relationship” is not quite as strong as saying $X$ is causing $Y$ but rather $X$ and $Y$ are associated.

4. A business decision maker is interested in regression because it can help determine if there is any relationship between two variables, and if so, what kind of relationship it is.

   Business relationships that could be explored are:

   (a) Are interest rates associated with inflation rates?
   (b) Over what range does product price correlate negatively with sales?
   (c) Does a stock’s risk predict the stock’s return?
   (d) Does age predict how many sick days an employee will take?

   From a management perspective, the independent variable is under management’s control and the dependent variable is a proxy for firm profit.

5. The data to be collected for a regression study is composed of pairs of observations on the explanatory and dependent variables. For such a sample of size $n$, we write $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

6. Let $g(X)$ be the function that maps an $X$ value to a $Y$ value, i.e., $Y = g(X)$. $g(X)$ could be a lot of things such as a straight line: $\beta_0 + \beta_1 X$ or a quadratic polynomial: $\beta_0 + \beta_1 X + \beta_2 X^2$, etc.

   Actually, $g(X)$ is called the deterministic component of the statistical model relating $X$ to $Y$. The complete statistical model is: $Y = g(X) + \epsilon$ where $\epsilon$ is assumed to be a random variable which is the stochastic component of the model. The random variable $\epsilon$ is also called the error term.
A good way to get a guess about what function to use in the regression analysis is to draw a **scatterplot**, which is just a plot of the \( x, y \) pairs.

7. **Example:** You want to know if advertising budget is positively associated with sales. Since advertising budget can be controlled, you set up the variables \( Y_i = \) sales for month \( i \), and \( X_i = \) advertising budget for month \( i \). You collect four months of \( x, y \) pairs and plot them:

Your manager looks at this plot and states “there is a point of diminishing returns to spending on advertising.”

8. If \( g(X) \) is a straight line, the complete model of the relationship between \( X \) and \( Y \) is: \( Y = \beta_0 + \beta_1 X + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \). Remember that the equation of a straight line, \( \beta_0 \) is called the “y-intercept” and \( \beta_1 \) is called the “slope.”

We assume with this model that \( \beta_0, \beta_1, \) and \( \sigma^2 \) are unknown parameters (constants). Since all we have from the process that generated the data set are the pairs, \((x_1, y_1), ..., (x_n, y_n)\) we need to figure out a way to **estimate** these parameters.

The notation for these estimated parameters will be \( \hat{\beta}_0, \hat{\beta}_1, \) and \( s^2 \), respectively.

9. How could we estimate these parameters? One not-very-good way is to always use 5 as the estimate. Sometimes you will be exactly correct but often you will be ridiculously wrong.

The estimation method we will use is called the method of **Least Squares**. The idea is to minimize the sum of the squared distances between the “fitted” model (straight line for now) and the data pairs.

How could you find the values of \( \hat{\beta}_0, \hat{\beta}_1 \) that would minimize this sum? Gauss figured out the equations hundreds of years ago. They are: \( \hat{\beta}_1 = s_{XY}/s_X^2 \) and \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \).

The notation, \( s_{XY} \) stands for

\[
\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).
\]

We call the equation, \( \hat{\beta}_0 + \hat{\beta}_1 x \) the “fitted regression line” and use it to **predict** the \( Y \) values. In other words, we “predict” that \( Y_i \) is \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \) at the \( x_i \) value of the explanatory variable, \( X \).
But we know what the value of $Y$ is at $x_i$, it is $y_i$. The mistake that our fitted model makes at the $x_i$ value is the difference: $y_i - \hat{y}_i$. Since we used estimated values for the slope and intercept, we call this the **predicted error** at $x_i$ and write $\hat{e}_i = y_i - \hat{y}_i$.

10. **Example continued:** You observe advertizing budget and total sales figures for four months: (1,10), (2,25), (3,35), and (5, 30). Regression output from the Minitab statistics package follows.
MTB : read 'mtbex1.dat' c1-c2.
4 ROWS READ

ROW  C1   C2
   1   1   10
   2   2   25
   3   3   35
   4   5   30

MTB : name c1 'advertiz'
MTB : name c2 'sales'
MTB : regress c2 on 1 predictor in c1

The regression equation is
sales = 12.4 + 4.57 advertiz

 Predictor  Coef  Stdev  t-ratio  p
Constant  12.429  9.650  1.29  0.327
advertiz  4.571  3.090  1.48  0.277

s = 9.142  R-sq = 52.2%  R-sq(adj) = 28.4%

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>182.86</td>
<td>182.86</td>
<td>2.19</td>
<td>0.277</td>
</tr>
<tr>
<td>Error</td>
<td>2</td>
<td>167.14</td>
<td>83.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>350.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MTB : plot c2 c1
11. The regression model: \( Y = \beta_0 + \beta_1 x + \epsilon, \epsilon \sim N(0, \sigma^2) \) implies that \( Y \), the dependent variable you are observing, is a random variable. So, \( Y \) must have a distribution, an expected value, and a variance just like any other random variable. Small problem: at what value of \( x \) are we talking about? It turns out that \( Y \) is distributed normally with a mean that depends on \( x \):

\[
Y(x) \sim N(\beta_0 + \beta_1 x, \sigma^2).
\]

The notation for \( Y \)'s expected value at \( x \) is: \( E[Y|x] = \beta_0 + \beta_1 x \). In words: “the expected value of \( Y \) given \( x \).” Another notation is: \( \mu_{Y|x} \).

12. Notice this model does not give any distribution to the \( x \) values. So, \( s_X \) and \( s_{XY} \) are not estimates of \( \sigma_X^2 \) or \( \text{Cov}(X,Y) \). They are just notation for more complicated formulas.

13. If you collect another data set of \( n \) observed \((x, y)\) pairs and compute \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), they may not be the same as the ones above.

If you collect many different data sets, each of size \( n \) and compute \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) each time, you could construct a histogram of \( \hat{\beta}_0 \) values and a histogram or \( \hat{\beta}_1 \) values. So actually, \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are sample statistics, i.e., r.v.s with distributions:

\[
\hat{\beta}_0 \sim N(\beta_0, \sigma^2 \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}) = N(\beta_0, \sigma^2_{\hat{\beta}_0})
\]

and

\[
\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{(n-1)s_X^2}) = N(\beta_1, \sigma^2_{\hat{\beta}_1}).
\]

14. Estimates of \( \sigma_{\hat{\beta}_0}^2 \) and \( \sigma_{\hat{\beta}_1}^2 \) are computed by replacing \( \sigma^2 \) with \( s^2 \) (the MSE in the regression output’s ANOVA table) in the above formulas. Doing that allows you to compute confidence intervals for \( \beta_0 \) and \( \beta_1 \) and to also perform hypothesis tests.

15. A 100(1 - \( \alpha \))% confidence interval for \( \beta_0 \) is:

\[
\hat{\beta}_0 \pm t_{\alpha/2, n-2} s_{\hat{\beta}_0}
\]
and for $\beta_1$:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}s_{\hat{\beta}_1}$$

16. **Example continued:** To find 95% confidence intervals for $\beta_0$ and $\beta_1$, first look up $t_{0.025,2}$ and find the value 4.303. Then,

$$\beta_0 \in (12.429 \pm 4.303 \times 9.65) = (-29.0, 54.0)$$

with 95% confidence and

$$\beta_1 \in (4.571 \pm 4.303 \times 3.09) = (-8.73, 17.87)$$

with 95% confidence.

So, zero advertizing could produce negative sales – very strange. Also, an increase in advertizing could cause sales to decline, also very strange.

17. In the last example, the confidence intervals were so wide they included silly values of $\beta_0$ and $\beta_1$. This is because the normal distribution allows extremely negative and extremely positive error values.

More complicated methods are needed to restrict the C.I.'s to reasonable ranges.

18. You can also test hypotheses about the parameters. The first one to do is whether there is a straight line relationship between $X$ and $Y$.

Since $E[Y|x] = \beta_0 + \beta_1 x$, $X$ is associated with $Y$ only if $\beta_1 \neq 0$. So, the hypothesis, $H_0 : \beta_1 = 0$ is equivalent to the hypothesis that there is no straight line relationship.

The different kinds of relationships are:
Figure 15: Positive relationship.

Figure 16: No relationship.

Figure 17: Negative relationship.
The test statistic for this test is \( \hat{\beta}_1 / S_{\hat{\beta}_1} \) and is distributed as \( t_{n-2} \) if \( H_0 \) is true.

19. **Example continued:** \( \hat{\beta}_1 = 4.571 \), \( s_{\hat{\beta}_1} = 3.09 \), and \( t_{0.025,2} = 4.303 \). The test statistic value then, is 1.48.

Unless you can provide a reason for why \( \beta_1 \) could never be negative (or positive), you should perform a two-tailed test.

20. You can also test hypotheses about \( \beta_0 \). Because \( E[Y|0] = \beta_0 \) is usually not very useful for describing the relationship between \( X \) and \( Y \), these tests are not often performed.

21. The Analysis of Variance (ANOVA) table for simple linear regression is:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.o.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>( \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 )</td>
<td>( \text{SSR/d.o.f.} )</td>
<td>( \text{MSR/MSE} )</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>( n-2 )</td>
<td>( \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 )</td>
<td>( \text{SSE/d.o.f.} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n-1 )</td>
<td>( \sum_{i=1}^{n}(y_i - \bar{y})^2 )</td>
<td>( \text{SST} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 18: Illustration of SST.
Figure 19: Illustration of SSR.

Figure 20: Illustration of SSE.
22. Roughly, if SSR/SSE is big, most of the variation is “explained” by the regression line.

This suggests another test for a “significant” regression: If $\beta_1 = 0$, then MSR/MSE $\sim F_{1,n-2}$.

If $\beta_1$ is zero, then SSR will be much smaller than SSE, if $\beta_1$ is not zero, the SSR will be bigger than SSE.

This fact means that you conduct a upper-tailed F-test with the statistic MSR/MSE.

23. If there is no error in observing $Y$ and the relationship between $X$ and $Y$ is a straight line, then all of the observations would line up perfectly.

![Figure 21: Perfect straight line relationship.](image)

This means that SSR = SST and SSE = 0 since SSR + SSE = SST. As the errors increase (more noise), SSE increases and SSR decreases.

This fact can be used to construct a measure of how well the fitted line fits the data. Define $R^2$ to be $\frac{SSR}{SST} = 1 - \frac{SSE}{SST}$. If $R^2$ is close to one, the sum of squares
explained by the fitted model is the majority of the total sum of squares.

24. **Example continued:** $R^2 = 52.2\%$, MSR/MSE = 2.19 which gives a p-value of 0.277.

### 7.1.1 Exercises

1. Why are $\hat{\beta}_0$ and $\hat{\beta}_1$ **unbiased** estimators of $\beta_0$ and $\beta_1$?

2. At $\alpha = 0.05$, would you reject or not reject $H_0 : \beta_1 = 0$?

3. What are the smallest and largest values that $R^2$ can be?

4. If $R^2$ is close to one, is it reasonable to think that there is so little noise?

5. At what $\alpha$-level would you reject the hypothesis that $\beta_1 = 0$ based on the F-test?

### 7.2 Prediction and Estimation

1. Other than finding C.I.’s for the parameters and testing for a relationship between $X$ and $Y$, is there anything else that the fitted model is good for?

   Yes! For a particular $X$ value, you can **estimate** $E[Y|x] = \beta_0 + \beta_1 x$ and, for a value of $X = x_0$ which you did not observe, you can **predict** $Y = \beta_0 + \beta_1 x_0 + \epsilon$.

   The estimate of the expected value of $Y$ at $X = x$ is written $\hat{\mu}_{Y|x}$. The prediction of $Y$ at $X = x_0$ is written $\hat{Y}_p$.

   For the regression model that assumes independent and identically distributed errors, $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$ and $\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_0$, i.e., the estimate and prediction share the same function.

   This can cause some problems with interpretation if you aren’t real careful.

2. **Prediction** means that the next time you observe $Y$ when the $X$ variable equals $x_0$, you **predict** that the value will be $\hat{y}_p$.

   On the other hand, **estimation** means that your best guess of the true, unknown deterministic part of the regression model at $X = x_0$, $\beta_0 + \beta_1 x_0$, is $\hat{\mu}_{Y|x_0}$.

   Let’s say you have inside information so that you know the true values of $\beta_0$ and $\beta_1$. You then know $\beta_0 + \beta_1 x_0$ exactly, in other words, there would be no uncertainty in your estimate of $\mu_{Y|x_0} = \beta_0 + \beta_1 x_0$.  

57
Even with this inside information, you still couldn’t predict $Y$ at $X = x_0$ with complete certainty because you don’t know what the value of the error term, $\epsilon$ will be.

Therefore, the variance of the prediction error at $x_0$ is larger than the variance of the estimation error at $x_0$. This variance is:

$$
\begin{align*}
\text{Var}(Y_{x_0} - \hat{Y}_p) &= \text{Var}(\beta_0 + \beta_1 x_0 + e_0 - (\hat{\beta}_0 + \hat{\beta}_1 x_0)) \\
&= \text{Var}(\mu_{Y|x_0} + e_0 - \hat{\mu}_{Y|x_0}) \\
&= \text{Var}(\hat{\mu}_{Y|x_0}) + \text{Var}(e_0) \\
&= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n - 1)s_X^2} \right] + \sigma^2 \\
&= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n - 1)s_X^2} \right].
\end{align*}
$$

3. Let $\sigma^2_p \equiv \text{Var}(Y_{x_0} - \hat{Y}_p)$. Since you rarely know $\sigma^2$, you estimate $\sigma^2_p$ with

$$
s^2_p \equiv s^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n - 1)s_X^2} \right].
$$

You can calculate a $100(1-\alpha)$% C.I. for $y_{x_0}$ with $s_p$. It is: $\hat{y}_p \pm s_p t_{\alpha/2,n-2}$.

Remember, if you have already observed $Y$ at $X = x_0$, you don’t need this C.I. since you know $y_{x_0}$.

4. From above, we see that $s_p$ is the sample standard deviation of $Y_{x_0} - \hat{Y}_p$.

5. The variance of the model’s expected value estimator at $x_0$ is

$$
\sigma^2_m = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n - 1)s_X^2} \right].
$$

Since again, you almost never know $\sigma^2$, you estimate this variance with

$$
s^2_m = s^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n - 1)s_X^2} \right].
$$

And a C.I. for $\mu_{Y|x_0}$ is $\hat{\mu}_{Y|x_0} \pm t_{\alpha/2,n-2}s_m$. 

58
6. The width of the estimation C.I. is \( 2s_m t_{\alpha/2,n-2} \) and \( 2s_p t_{\alpha/2,n-2} \) is the width of the prediction interval. 

Since \( s_m \) and \( s_p \) get larger as \( x_0 \) gets farther from \( \bar{x} \), the widths of these C.I.s get bigger as \( x_0 \) moves farther away from \( \bar{x} \).

**Moral:** The most precise estimates and predictions are made close to \( \bar{x} \).

Also, \( s_m < s_p \) so the prediction interval is always larger than the estimation interval – you are always less certain about predicting a realization of \( Y_{x_0} \) than you are estimating \( \mu_{Y|x_0} \).

### 7.2.1 Exercise

1. You collect 10 observation pairs and compute \( s^2 \), \( \bar{x} \), and \( s^2_X \) to all be equal to 1. For a prediction at \( X = -5000 \), will \( s_p^2 \) be large or small?

### 7.3 Original Scale Prediction

If a log transformation has been applied to the response variable, \( Y \), and a prediction is needed in the original scale, a nearly unbiased prediction is: 

\[
\hat{y} = \exp(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + .5s^2) \]

where \( b_0, b_1, \ldots, b_k \) are the fitted regression coefficients, \( x_1, \ldots, x_k \) are the values of the predictors at which a prediction is desired, and \( s^2 \) is the squared standard error of the regression or, equivalently, the mean squared error in the ANOVA table of the regression. See Miller (1984) for nearly unbiased predictors for the transformations \( Y^{1/N}, N > 0 \), and \( Y^{-1} \).

The prediction interval (confidence interval of the true value of \( Y \) at the predictor variable values) is simply found by exponentiating the prediction interval found in the log-scale. This log-scale interval is computed by Minitab when a predict statement is given to the regress command.


### 7.4 Misuse of Regression Models

1. There are several ways to misuse a fitted regression model.

2. **Misuse #1:** Concluding a causal relationship exists between \( X \) and \( Y \) when only association has been shown.
3. Example:

Figure 22: Scatterplot of reading ability versus shoe size.

Does this mean that shoe size causes reading ability? No! An unobserved intervening variable, “age” is causing shoe size and reading ability to increase.

Figure 23: Causal diagram of age, reading ability, and shoe size.

Shoe size is simply associated with reading ability.

4. You need to do the following to establish causality:

   (a) You observe the association in different situations and data sets.
(b) After removing or holding constant possible intervening variables, the association is still present.

(c) A plausible mechanism or explanation of the presumed causality based on known mechanisms can be argued.

5. **Misuse #2:** Predicting $Y$ outside the range of data taken on the predictor variable, $X$.

6. **Example:** You collect sales data every 3 months on start-up firms:

![Figure 24: Start-up firm sales versus number of months since start-up.](image)

At least half of all start-ups fail in the first year, the survivors rarely become large corporations.

**Moral:** Use your fitted regression model to predict outside the range of your predictor variable observations only if you are certain the fitted relationship continues to be the same. This assumption can only be checked by collecting data in this new range.

### 7.4.1 Exercises

1. Maybe big-footed kids do read better. How would you collect data and analyze it with regression analysis so as to take into account the effect of age?
2. Assuming you believe this example, what do you estimate the average start-up sales will be after 48 months?
8 Influence Diagram (ID) Definitions

1. An ID is a graphical representation of a multivariate probability distribution. The representation is in the form of a mathematical construction called a graph.

2. A graph is a pair \( G = (V, E) \) where \( V \) is a finite set of vertices and \( E \) is a subset of the \( V \times V \) ordered pairs of distinct edges. For \( \alpha, \beta \in V \), if the edge \((\alpha, \beta)\) is in \( E \) but \((\beta, \alpha)\) is not in \( E \), then the edge is directed; otherwise, the edge is undirected.

3. If \( \beta \rightarrow \alpha \) then \( \beta \) is a parent of \( \alpha \) and \( \alpha \) is a child of \( \beta \). In a graph consisting of only directed edges, vertices having no parents are roots, and vertices having no children are terminal.

4. A path of length \( n \) from \( \alpha \) to \( \beta \) is a sequence \( \alpha = \alpha_0, \alpha_1, ..., \alpha_{n-1}, \alpha_n = \beta \) of distinct vertices such that \((\alpha_{i-1}, \alpha_i)\) is in \( E \) for \( i = 1, ..., n \).

5. A cycle of length \( n \) is a path wherein \( \alpha_0 = \alpha_n \). The cycle is directed if it contains at least one directed edge. A graph whose edge set contains only directed edges is called a directed graph. A graph for which there are no directed cycles is called an acyclic graph.
6. Kiiveri et al. (1984): the product of the unconditional and conditional distributions of a set of random variables indexed by the vertex set of a directed acyclic graph defines a joint probability distribution as long as this product is positive for all joint events.

7. Random variables indexed by root vertices are assigned unconditional distributions, all other random variables are assigned conditional distributions for each possible event of the variables indexed by that variable’s parent or parents. This joint distribution \( P \) then, is said to recursively factorize according to the graph \( G \).

8. The ID model architecture helps to clarify variable dependencies, random components, and possible system control points. By conditioning on desired sets of control variables, the effect of different interventions can be explored.

9. Previous work on including quantitative variables in an ID have focused on gaussian variables. The most developed of these models is the Conditional Gaussian model of Lauritzen and Wermuth (1989).

10. **Example:** If the vertices in the above graph index the random variables \( X_1 \), \( X_2 \), and \( X_3 \), then

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \\
\times P(X_2 = x_2 | X_1 = x_1) P(X_1 = x_1).
\]

11. Two advantages of this graph theoretic representation of a joint distribution:

   1. The above results allow the construction of a large, complicated joint distribution by specifying a set of much simpler, local unconditional and conditional distributions.

   2. The directed graph provides in one picture, all probabilistic dependencies.

12. An ID represents random variables as circles, decisions as squares, and value variables as diamonds. An ID having only random variables is often called a Bayesian Belief Network (BBN).

13. Hereafter, any variable represented within an ID will be referred to as a *node*.

**Environmental BBN Example**
1. SHADE: amount of shade present in an aspen stand – takes on the values shade, (0 to 25% full sunlight), and full sunlight (90 to 100% sun).

2. MOISTURE: soil moisture of the stand – takes on the values dry, (5 to 25% soil moisture content by volume), and moist (25 to 50% moisture content).

3. SITECNDS: qualitative measure of physiographic site conditions for aspen sucker growth response – takes on the values unfavorable, and favorable conditions.

4. Causal relationships represented: amount of shade and amount of soil moisture independently influence the site conditions for sucker growth.

5. Unconditional distributions for root nodes and conditional distributions for other nodes that need to be specified:

\[
P(\text{SHADE}),
\]
\[
P(\text{MOISTURE}), \text{ and}
\]
\[
P(\text{SITECNDS}|\text{SHADE}, \text{MOISTURE}).
\]

6. Substantive science represented by the distributions (see Haas (1991a)):

1. Aspen sucker growth is vigorous under the conditions of full sunlight and moist soil.

2. Suckers will tolerate dry conditions in full sunlight but heavy shade will severely restrict sucker growth no matter what the moisture content is.

3. Under heavy shade and dry soil, sucker growth is all but curtailed.

\[
P(\text{favorable}|\text{shade, dry}) = 0.05
\]
\[
P(\text{unfavorable}|\text{shade, dry}) = 1.0 - 0.05
\]
\[
P(\text{favorable}|\text{shade, moist}) = 0.1
\]
\[
P(\text{unfavorable}|\text{shade, moist}) = 1.0 - 0.1
\]
\[
P(\text{favorable}|\text{full sun, dry}) = 0.7
\]
\[
P(\text{unfavorable}|\text{full sun, dry}) = 1.0 - 0.7
\]
\[
P(\text{favorable}|\text{full sun, moist}) = 0.9
\]
\[
P(\text{unfavorable}|\text{full sun, moist}) = 1.0 - 0.9.
\]
7. For example, the 0.05, 0.95 distribution is intended to represent very little conditional belief in a \textit{favorable} site condition given shade and dry soil.

8. These numbers are the IDs \textit{parameters} and are stored in a Conditional Probability Table (CPT).

9. Rationale for root node distributions.

\textsc{Shade}: assume that this particular stand has been recently clearcut, but that due to a vigorous shrub and forb population, the belief that full sunlight reaching aspen suckers is only .8, implying a .2 belief that the suckers are in shade.

\textsc{Moisture}: based on soil moisture meter measurements and an examination of the types of forbs and shrubs growing on the stand, a belief of .9 can be assigned to the soil moisture value of 0-25\% soil moisture; i.e., there is a .9 belief on the part of the individual examining the stand that the soil moisture is less than 25\%. This implies that there is only .1 belief that this is a moist site.

10. Combining the above conditional and unconditional distributions gives the joint distribution for this three-node BBN:

\[
P(\text{SITECNDS}, \text{Shade}, \text{Moisture}) = P(\text{SITECNDS}|\text{Shade}, \text{Moisture}) \times P(\text{Shade})P(\text{Moisture}).
\]

11. For example, the overall or marginal belief in the site condition value of \textit{favorable} implied by the above unconditional and conditional distributions is found by computing:

\[
P(\text{SITECNDS} = \text{favorable}) = \sum_{x,y} P(\text{SITECNDS} = \text{favorable}, \text{Shade} = x, \text{Moisture} = y)
\]

\[
= P(\text{SITECNDS} = \text{favorable}|\text{full sun, moist soil}) \times P(\text{full sun})P(\text{moist soil})
\]

\[
+ P(\text{SITECNDS} = \text{favorable}|\text{shade, moist soil}) \times P(\text{shade})P(\text{moist soil})
\]
\[ + P(\text{SITECNDS} = \text{favorable}|\text{full sun, dry soil}) \\
\times P(\text{full sun})P(\text{dry soil}) \\
+ P(\text{SITECNDS} = \text{favorable}|\text{shade, dry soil}) \\
\times P(\text{shade})P(\text{dry soil}) \]

and is

\[
= (0.9 \times 0.8 \times 0.1) + (0.1 \times 0.2 \times 0.1) \\
+ (0.7 \times 0.8 \times 0.9) + (0.05 \times 0.8 \times 0.9) \\
= 0.614.
\]

Since SITECNDS is dichotomous,

\[
P(\text{SITECNDS} = \text{unfavorable}) \\
= 1 - P(\text{SITECNDS} = \text{favorable}) \\
= 0.386.
\]

12. Overall, the model gives a .61 belief in this particular set of site conditions being favorable for aspen growth.

13. This belief can be interpreted as a judgment of the site’s physiographic favorability for aspen sucker growth derived from expert opinion of aspen sucker growth sensitivities and the particulars of the stand under consideration.

14. The value can be used as input to management decisions concerning what sites should be singled out as having high potential for aspen sucker growth.

15. See Haas (1991a) for a more complete model of aspen stand sucker response.

16. Finding all marginal probabilities of a BBN is called “solving the BBN.”

17. Finding the optimal decision represented by an ID requires the expected value of the utility node be computed for each combination of decision nodes. The combination that maximizes this expected value is the optimal decision. Performing these computations is called “solving the ID.”

18. Solving a BBN or ID is NP-hard.
19. Exact methods based on graph manipulations are used in many software systems, e.g. Hugin and Netica.

20. Simulation can also be used to approximately solve a BBN or ID. logic sampling (Henrion 1988) is the simplest of these methods. Analytica and my system (id) uses simulation to solve BBNs and IDs.

Modeling Causal Relationships
with an ID (Pearl 1995)

A mathematical language to express causal relationships can be derived from properties of an ID.

\(d\)-separation: Let \(X\), \(Y\), and \(Z\) be three disjoint subsets of nodes in a directed acyclic graph \(G\), and let \(p\) be any path between a node in \(X\) and a node in \(Y\). Then \(Z\) is said to block \(p\) if there is a node \(w\) on \(p\) satisfying one of the following two conditions:

(i) \(w\) has converging arrows along \(p\), and neither \(w\) nor any of its descendants are in \(Z\), or (ii) \(w\) does not have converging arrows along \(p\), and \(w\) is in \(Z\). Further, \(Z\) is said to \(d\)-separate \(X\) from \(Y\), in \(G\), written \((X \perp Y | Z)_G\), if and only if \(Z\) blocks every path from a node in \(X\) to a node in \(Y\).

Causal Effect: Given two disjoint sets of nodes, \(X\) and \(Y\), the causal effect of \(X\) on \(Y\), denoted \(P(Y | \tilde{x})\), is a function from \(X\) to the space of distributions on \(Y\). For each realization \(x\) of \(X\), \(P(Y | \tilde{x})\) gives the probability of \(Y = y\) induced by deleting from the graph all nodes that are parents of nodes in \(X\) and substituting \(x\) for \(X\) in the remainder.

the \(\tilde{x}\) notation means that these values for \(X\) are intentionally fixed.

Identifiability: The causal effect of \(X\) on \(Y\) is said to be identifiable if the quantity \(P(Y | \tilde{x})\) can be computed uniquely from any positive distribution of the observed nodes that is compatible with \(G\).

Theorem: Computing \(P(Y = y | \tilde{x})\).

Case 1: If a set of nodes \(Z\) satisfies a condition called the back-door criterion relative to \((X, Y)\), then the causal effect of \(X\) on \(Y\) is identifiable and is given by the formula

\[
P(Y = y | \tilde{x}) = \sum_{\tilde{z}} P(Y = y | \tilde{x}, \tilde{z}) P(Z = \tilde{z}).
\]
A set of nodes $Z$ satisfies the back-door criterion relative to an ordered pair of nodes $\{X_i, X_j\}$ in a directed acyclic graph $G$ if: (i) no node in $Z$ is a descendant of $X_i$ and (ii) $Z$ blocks every path between $X_i$ and $X_j$ which contains an arrow into $X_i$. If $X$ and $Y$ are two disjoint sets of nodes in $G$, $Z$ satisfies the back-door criterion relative to $(X, Y)$ if it satisfies it relative to any pair $(X_i, X_j)$ such that $X_i \in X$ and $X_j \in Y$.

**Back-Door Criterion Example**

\{X_3, X_4\} and \{X_4, X_5\} satisfy the Back-Door criterion for the ordered pair $X_i, X_j$.

\{X_4\} does NOT because $X_4$ does not block the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$ because there are converging arrows onto $X_4$ along this path.

**Case 2:** (Front-door) If a set of nodes $Z$ satisfies the following conditions relative to an ordered pair of nodes $(X, Y)$: (i) $Z$ intercepts all directed paths from $X$ to $Y$, (ii) there is no back-door path between $X$ and $Z$, and (iii) every back-door path between $Z$ and $Y$ is blocked by $X$. Then the causal effect of $X$ on $Y$ is identifiable and is given by the formula

\[
P(Y = y | \tilde{x}) = \sum_{\tilde{z}} P(Z = \hat{z} | \tilde{x}) \sum_{\hat{z}'} P(Y = y | \hat{x}', \hat{z}) P(X = \hat{x}') .
\]
Front-Door Criterion Example

$U$ is unobserved.

$X_2$ satisfies the Front-Door criterion for the ordered pair $X_1, X_3$.

1. The Theorem allows many apparently useless nonexperimental samples to be used for discovering causal relationships.

2. Ecosystem managers often have many such data sets.

**Advantages of IDs in Natural Resource Management**

1. Domain experts can quickly learn how to build complex stochastic models of ecosystems.

2. The graphical representation allows more intuitive model construction and more effective communication of the model.

3. Direct assignment of the CPT’s allows stochastic models to be built of processes that lack other mathematical models such as systems of differential equations.

4. Uncertainty about data, relationships, and outcomes can be made an integral part of the model from the beginning.

5. Interventions and management activities can be explicitly represented with decision nodes.
6. The value or utility of different future ecosystem states (e.g. extinction/non-extinction) can be modeled with utility nodes.

Two Applications of BBNs to Ecosystem Management

Waterbody Eutrophication

1. Reckhow (1999) describes an ID for managing eutrophication (and hence the possibility of a fishkill) of an estuary.

2. Percent-Forested-Buffer is the management plan decision node and Onset-of-an-Anoxia-Condition (lack of oxygen in the water) is the utility node.

Wildlife Population Viability

1. Marcot et al. (2000) describes a BBN approach to modeling fish and wildlife population viability as affected by land management and development to support management decisions by the USDA Forest Service as part of the Interior Columbia Basin Ecosystem Management Project (ICBEMP).

2. Key Environmental Correlates (KEC’s) are defined by experts to be the characteristics most important to the success of the population.

3. Proxy nodes computed from GIS maps are used to “cause” (indicate) the degree to which these KEC’s are present.

4. Habitat suitability nodes are “caused” by these KEC’s and population response is “caused” by habitat suitability.

ID Parameter Estimation (ID Learning I)

1. Estimating ID parameters (conditional probability tables) from data is the more-understood aspect of “learning” an ID.

2. Currently, the most computationally efficient frequentist method for estimating the parameters of an influence diagram consisting solely of qualitative chance nodes from an incomplete sample is the EM(η) (for Expectation – Maximization with learning rate η) method of Bauer et al. (1997). Zhang (1996) is also able to speed up EM convergence.
3. Also for discrete IDs, the most efficient Bayesian method appears to be the Bound and Collapse (BC) method of Ramoni and Sebastiani (1997) although the Heckerman et al. (1994) method may be almost as efficient.

4. Cheesman and Stutz (1997) give a Bayesian method for fitting a mixed ID (some discrete and some continuous-valued nodes) with missing data (equivalently, latent nodes). A classification task is assumed in that each observation is assumed to have come from a particular class. Neither these class assignments nor the number of these classes are known and are to be estimated from the data.

5. A software system (Autoclass) is available and can be viewed as a Bayesian data mining system.

   **All of these Bayesian methods use a Dirichlet prior distribution.** This is appropriate since ID parameters are probabilities and a Dirichlet distribution is a multivariate Beta distribution wherein each component has support (0, 1).

**ID Model Selection**

*(ID Learning II)*

1. Construction of a BBN shares the *knowledge acquisition* bottleneck common to any effort to build a decision support system.

2. A solution would be to automate the construction of the BBN by applying a BBN structure learning algorithm to a set of observations.

3. To do this, one must evaluate the “goodness” of a large number of candidate graphs (models) and decide on a structure that best represents dependencies in the data. In other words, one must first define a *score function* that gives a score to each model and then use this score function to search the space of possible models and then select the model with the highest score value.

4. Also called “structure learning.” Equivalent to model selection in statistics.

5. Structure learning is much more difficult than estimating parameters due to the combinatoric explosion of the number of candidate models.
6. Many score functions have been devised, both frequentist and Bayesian, see Castillo et al. (1997, ch. 11).

7. If prior probabilities on candidate models can be obtained, use the Bayesian score function. Let $S$ be a sample, $G$ be a candidate network having $p$ nodes, and $\theta$ be the parameters defining the conditional and unconditional distributions in $G$. Then,

$$\text{Bayesian Score} \equiv \log(P(G)) + \log(P(S|G, \hat{\theta}))$$

$$- .5 \cdot \text{Sizeof}(G) \log(\text{Sizeof}(S)).$$

The third term penalizes for complex models.

8. If prior probabilities are difficult to justify, use the Minimum Description Length (MDL) criterion (see Castillo et al. (1997, p. 509). Let $v_i$ be the number of discrete values for node $i$ and $c_i$ be the number of unique combinations of values of node $i$’s parents. Then,

$$\text{MDL Score} \equiv \sum_{i=1}^{p} \sum_{j=1}^{v_i} \sum_{k=1}^{c_i} N_{ijk} \log(N_{ijk}/N_{ik})$$

where for each node $i$, $N_{ik}$ is the number of observations node $i$ with parent-value combination $k$ and $N_{ijk}$ is the number of observations on node $i$ with value $j$ under parent-value combination $k$.

9. **Recommendation:** use the K2-AS method of Provan and Singh (1997) because of its emphasis on finding a model that both has a high score and high predictive accuracy.

10. Symbol Definitions:

(a) Consider $p$ discrete-valued nodes, $X_1, \ldots, X_p$, each measured on an experimental unit and $n$ experimental units observed to make up the sample, $S$.

(b) Partition $S$ into a training, testing, and evaluation subsamples, $S_{\text{train}}$, $S_{\text{test}}$, and $S_{\text{eval}}$, respectively.

(c) Let $Z$ be the set of the nodes. Let $\Delta$ be a subset of $Z$.

11. Definition of conditional independence test:

For nodes $a$ and $b$, find the set of nodes $C_{ab}$ for which $a$ is independent of $b$ given values on all nodes in $C_{ab}$. Do this by performing a $\chi^2$ test for conditional independence of $a$ and $b$ given a macro node $a_m$ where $a_m$ is a discrete node having values that are all the unique combinations of values of the nodes in $C_{ab}$.  

73
12. Definition of the structure learning algorithm, K2:

(a) Given an ordering of the nodes, start with the first node. Temporarily make each remaining node a parent of $X_0$. Assign as a permanent parent to $X_0$, the node that makes the 2-node network have the largest score function. The score function used is the joint probability of the sample which is proportional to the probability of the model given the sample.

(b) Repeat the above procedure on every node in the network as long as the score function of the new network is bigger than the previous network.

(c) Do a greedy search for parents by only adding one candidate parent node at a time.
13. Definition of CB: Use the K2 learning algorithm with the node ordering found with conditional independence tests.

14. Definition of K2-AS: (K2 with Attribute (node) Selection):

(a) Variable Selection Phase:
   i. Divide the sample into $S_{train}$, $S_{test}$, and $S_{pred}$.
   ii. Add $X$ to $\Delta$ if this node creates a CB-built network using $X \cup \Delta$ nodes that has a uniquely higher predictive accuracy than any other node in $Z - \Delta$. Use $S_{train}$ to build each trial network and $S_{test}$ to compute the predictive accuracy of each of the networks.
   iii. Continue to add nodes in this manner to $\Delta$ until no one node creates a network that has a uniquely higher predictive accuracy.

(b) Network Learning Phase:
   Use only the nodes in $\Delta$ to create a network with CB using $S_{train}$.

(c) Predictive Accuracy Phase:
   Compute the predictive accuracy of the phase 2 network using $S_{pred}$.


**ID Sensitivity Analysis**

1. Three main types of sensitivity analysis:

   (a) Parameter sensitivity due to either (a) expert misspecification or (b) variability of estimated parameter values.
   (b) Sensitivity of marginal probabilities to addition or deletion of dependency links in the ID.
   (c) Variability of marginal probability distribution of outcome nodes.

2. Parameter sensitivity has been studied in Pearl (1988), Haas (1991a), and Laskey (1993).

3. Pearl (1986) notes that the natural tendency of the effect of a node to be attenuated on nodes multiple links removed and cites this characteristic as an example of a BBNs stability:
The addition of any new node \( x_i \) to the network requires only that the expert identify a set \( S_i \) of nodes which “directly influence” \( x_i \), locally assess the strength of this relation and make no commitment regarding the effect of \( x_i \) on other nodes, outside \( S_i \). Even though each judgment is performed locally, their sum total is guaranteed to be consistent. This model-building process permits people to express qualitative relationships perceived to be essential, and the network preserves these qualities, despite sloppy assignments of numerical estimates.

4. Haas (1991a) derives sensitivity ratios for particular cases. For example: \( \beta \) parameterizes a parent of \( X_2 \) in a 2-node network. Let \( P(X_1 = 0) = \beta_1 \), \( P(X_2 = 0|X_1 = 0) = \beta_2 \), and \( P(X_2 = 0|X_1 = 1) = \beta_3 \). It is desired to study the effect of changes in \( \beta_1 \) on the marginal probability of \( X_2 \) \( (P(X_2 = 0)) \). Note,

\[
\begin{align*}
P(X_2 = 0) &= P(X_2 = 0|X_1 = 0)P(X_1 = 0) \\
&\quad + P(X_2 = 0|X_1 = 1)P(X_1 = 1) \\
&= \beta_2\beta_1 + \beta_3(1 - \beta_1) \\
&= \beta_1(\beta_2 - \beta_3) + \beta_3 \\
&= \beta_1a_2 + \beta_3
\end{align*}
\]

where \(-1 < a_2 < 1\). Hence, \( \delta(\epsilon) = \epsilon a_2 \). The sensitivity of the BBN to small errors in a parameter, now one dependency link removed, are less than the changes in the parameter itself.


### Drawbacks to IDs

1. Large number of parameters: can over-parameterize, expensive to estimate.

2. Difficult to represent time.

3. Difficult to represent feedback.

4. Currently, the only well-known extension to BBNs that allow continuous-valued nodes is the mixed, conditional gaussian BBNs of Lauritzen. These networks suffer from being highly dependent on a very specific assumption: conditional multivariate normality.
5. Although BBNs are ideal for structuring knowledge domains where there are few mathematical models, little work has been done on extending BBNs to model non-gaussian and/or stochastic model-generated random nodes (but see below).

**IDs for Complex Stochastic Systems**

1. **statistical models:** stochastic models built solely for purposes of prediction and not derived from physical process relationships

2. **mechanistic models:** models based at least in part on scientific theory of how the phenomenon’s state nodes are related to each other.

3. Mechanistic models can be based on systems of Ordinary Differential Equations (ODE)’s or *Stochastic* Differential Equations (SDE)’s.

4. The conditional distribution of a set of dependent nodes can be a parametric multivariate distribution such as a multivariate Beta, or implicit as outputs from a complex multivariate stochastic model.

5. An ID provides a unified structure into which models of quantitative and/or qualitative system state nodes can be placed.

6. Most mathematical mechanistic models have time explicitly present in the model description, e.g. the symbol \( t \) in a system of ODE’s or SDE’s.

7. Because of this convention, time is explicitly represented in the proposed IDs as a decision node.

8. Random model parameters or random independent nodes that are not time dependent are root nodes that generate a random value at \( t = 0 \).

ODE or SDE system output nodes are dependent ID nodes that produce a random value conditional on both the given end-time value (\( t_{\text{end}} \)) and all decision node values.

To obtain a realization of the model over the interval \((0, t_{\text{end}})\), the ID is sampled at times \( t = 0, \ t = t_1, \ldots, t = t_{\text{end}} \). Other nodes may have time-dependent parameters and are therefore descendants of the time node.
Species Viability ID: The Cheetah in Kenya

The cheetah management ID consists of three main sections

1. Decision nodes for representing management options, subregions or areas of Kenya, and a time value at which ID outputs are desired.

2. A vector SDE model of cheetah population size represented by a set of chance and deterministic nodes.

3. Presence/absence and associated utility represented by a chance and value node, respectively.

Probabilistic Representation of Geographic Characteristics

Areas of the study region that are large enough to have the potential of sustaining the species under favorable conditions are identified.

These areas should be as homogeneous in terms of vegetation and climate as possible. Strict homogeneity is not required however, because the ID uses r.v.’s to represent within-area heterogeneity.

The Gross (1998) regions (Marsabit, Eastern, Samburu, Tsavo surroundings, Masailand, Laikipia, and Nakuru) are used here with the addition of regions for the densely settled western districts (Western), the central farmlands (Central), Turkana (Turkana), and coastal areas (Coastal).

Each area’s climate is modeled as a discrete, time-independent r.v. taking on values very-arid, arid, semi-arid, and non-arid (from Gros (1998)).

Land-use values of nomadic-camel, nomadic-cattle, ranching, and farming, are also modeled as time-independent discrete random nodes (also from Gros (1998)).

Probabilistic Representation of Carrying Capacity

A region’s cheetah carrying capacity is a deterministic function of herbivore biomass:

\[ K_t \equiv \text{nearest integer}(\beta_{K_t}^{(0)} + \beta_{K_t}^{(1)}B_t). \]
A single birth-death model for the meta-population size of cheetah-prey herbivores at time $t$ is used to represent the population dynamics of the prey populations. Let $k_0$ be the herbivore carrying capacity at time $t_0$. The SDE for $B_t$ with random carrying capacity is

$$dB(t) = B(t)(k_0 - B(t))dt + B(t)\sigma dW_t^{(B)}$$

where $dW_t^{(B)}$ is a zero mean, unit variance white noise process.

$B_t$ is influenced by the nodes $t$, $C$, and $m$. $U$ could also be an influence but was not used here due to the increased complexity.

This *logistic growth* SDE is added to the system of SDE's (see below) and this system of four equations is solved numerically.

**Cheetah Population Dynamics Model**

Wells et al. (1998) give a DE model of species population size (count) as a function of time. The authors define the following parameters:

- $f \in (0, 1)$: instantaneous birth rate,
- $r \in (0, 1)$: instantaneous mortality rate (death rate),
- $c \in (0, 1)$: proportion of the $N$ animals that meet over a short period of time. This implies that $cN$ is the number of meetings over a short period of time.
- $P$: the probability that any one meeting does not result in a litter.
- $k$: carrying capacity of the environment in terms of maximum number of animals that can be supported.

The model is

$$\frac{dN}{dt} = f(1 - P^cN)N - rN - (f - r)\frac{N^2}{k}.$$

$N_0$, the initial count of the species within an area is modeled as an area-dependent parameter. The initial time at which this count exists is $t_0$.

There are several sources of uncertainty when using (8) to predict population size through time. These are:

1. heterogeneity of land use, vegetation, and climate within an area,
2. the partial effect that herbivore density has on carrying capacity,
3. the partial effect that poaching and pest hunting has on an area’s death rate, and
4. the impact that unpredictable forces such as droughts have on birth and death rates.

Uncertainty effects on birth and death rates are modeled by making these rates stochastic processes defined by the solutions of their governing, independent SDEs as follows.

Let \( \mathbf{W}_t \equiv (W_t^{(f)}, W_t^{(r)}, W_t^{(N)})' \) be a vector of three independent Wiener processes.

Also, let
\[
\begin{align*}
    a_f(X_t) &\equiv -(\alpha_f + \beta_f^2 X_t)(1 - X_t^2) \text{ (drift)}, \\
    b_f(X_t) &\equiv \beta_f(1 - X_t^2) \text{ (diffusion)}, \\
    f_t &\equiv U(X_t) \equiv (1 + X_t)/2.
\end{align*}
\]

The distribution of \( f_t \) at \( t \) is the solution to the SDE:
\[
df_t = (1/2) a_f(U^{-1}(f_t))dt + (1/2) b_f(U^{-1}(f_t))dW_t^{(f)}.
\]

This SDE was chosen because its solution is bounded between 0 and 1 making \( f_t \) a well-defined birth rate.

A similar development leads to the death rate SDE:
\[
dr_t = (1/2) a_r(U^{-1}(r_t))dt + (1/2) b_r(U^{-1}(r_t))dW_t^{(r)}.
\]

The tendency of more females to have litters within protected areas (Gros 1998) is represented by having the parameter \( \alpha_f \) be conditional on the area’s status.

Similarly, to represent the effect of poaching and pest hunting on \( r_t \), \( \alpha_r \) is conditional on the degree of poaching and pest hunting.

The variability of the sample paths of \( f_t \) and \( r_t \) are controlled by the parameters \( \beta_f \) and \( \beta_r \), respectively.

All other effects of uncertainty (e.g. model inadequacy, age-dependent parameters) on the within-area cheetah count differential \( (dN_t) \) are represented by the derivative of a Wiener process — accomplished by converting (8) to an SDE:
\[
\begin{align*}
    dN_t &= \left[ f_t(1 - P^{(N_t)})N_t - r_t N_t - (f_t - r_t) \frac{N_t}{k} \right] dt \\
    &\quad + \beta_N dW_t^{(N)}.
\end{align*}
\]
where $P$, $c$, $N_0$, and $\beta_N$ are fixed parameters, and $k$ is a random parameter.

Conditional on $k$, the random vector $(f_t, r_t, N_t)'$ is the solution to this nonlinear vector SDE.

If $f_t$, $r_t$, and $k$ are fixed, the distribution of $N_t$ is the solution of an SDE with constant coefficients. This is the causal model represented by the ID in terms of Pearl’s definition of causality.

The solution that is found when $k$ is a random node, and $f_t$ and $r_t$ are stochastic processes corresponds to finding the joint probability distribution of $(k, f_t, r_t, N_t)'$ for given values of $t$, area, and management option.

Defining the viability model as an ID then, is critical to conveying causal structure to all involved parties. Such specific statements concerning causality cannot be deduced from examination of the vector SDE alone.

A species is defined to be viable within a particular area if the expected count within that area is nonzero at a distant future time point.

The ID gives a probability distribution of species count within each area at a given time and given management option.

**Fraction-of-Area-Detected Node**

The final output node, $D_t$ measures the fraction of a region’s area over which cheetah have been detected. Let $ra$ be a region’s surface area and $d = N_t/ra$, i.e., the density of cheetah in the region.

An observation on $D_t$ can be computed from maps of cheetah presence/absence by district: divide by $ra$, the sum of all areas of districts in the region on which cheetah have been detected.

$D_t$ is a deterministic function of $N_t$ and $ra$: Let $\xi$ be the minimum cheetah density that results in a cheetah detection report. Let $\rho$ be a cheetah density above which cheetah are certain to be reported. Then

$$D_t = \begin{cases} 0, & d < \xi \\ (d - \xi)/(\rho - \xi), & d \in (\xi, \rho) \\ 1, & d > \rho. \end{cases}$$
Note that it is possible for $N_t$ to be positive but $D_t$ to be zero, i.e., $\xi$ can be interpreted as the minimum density detection limit.

The loss node is a deterministic function of $D_t$. This loss function is arrived at through public debate and reflects a composite of the ecological and economic values that have been expressed by all stakeholders.

The loss function used here represents the values (a) high loss if cheetah go extinct, and (b) bounded economic loss from cheetah predation on livestock:

$$L = \begin{cases} 
10^4, & D_t < .15 \\
1, & .15 < D_t < .5 \\
5, & .5 < D_t < .8 \\
10, & .8 < D_t.
\end{cases}$$

The value of $m$ that minimizes the expected value of this loss node is chosen for implementation.

**ID Parameter Estimation**

1. **Problem:** Huge environmental statistical models are difficult to estimate and validate because data sets are often not large enough or complete enough to allow the use of standard statistical methods.

2. **Proposal:** fix some of the parameters at values based on the substantive literature and then use the available data to estimate the balance of the parameters. This approach was used by Speed (1993) to estimate a large state-space time series model of salmon population in the Pacific Northwest.

3. **Consistency Analysis** (Haas 1997, 2001) extends this idea of fixing some of the parameters to substantive theory-justified values (called the **hypothesis values**) by producing parameter estimates such that the resultant fitted or consistent distribution deviates minimally from the hypothesis distribution while continuing to be consistent with the available data.

4. In situations in which it is difficult to justify a particular prior distribution on model parameters, a Bayesian approach requires more from the analyst than the analyst can justify. In these situations, CA then, is one way to incorporate theoretical knowledge into parameter estimation without inheriting some of the criticisms of the application of Bayesian methods to the assessment of ecological models (see Dennis (1996)).
Assessment of Fit

1. Because this is a temporal model and is intended to be used to predict future cheetah viability, some assessment of its prediction skill is of interest.

2. The statistic used here is the root mean squared error of predictions one time step ahead (RMSPE). For each prediction, the model is fitted with consistency analysis using all data that is strictly earlier than the current prediction time.

3. When the sample contains unknown errors however, the RMSPE should not be viewed as the ultimate criterion for model selection. Rather, it should be examined in light of known weaknesses in the sample.

4. Separate RMSPE values were computed for herbivore counts and detection fraction predictions over the observation times.

5. For $c_H = .5$, these calculations yielded $3757.5$, and $.727$ for $B_t$ and $D_t$, respectively.

6. For $c_H = 0$, these values are $3716.9$, and $.720$.

Use of the Consistent Influence Diagram for Cheetah Management

1. Now that the influence diagram has been fitted to both the sample and prior knowledge, it can be used to aid cheetah management decisions.

2. To illustrate the graphical outputs that are possible from the influence diagram, Figure 5 gives the mean $D_t$ values by region for the conditions $m = do\ nothing$ and $t = 2020$. Cheetah presence is low in regions of heavy farming and higher in less cultivated areas.

Selection of Optimal Management Option

1. A hypothetical decision selection exercise is described to illustrate how the estimated influence diagram is used within the EMS to manage cheetah viability.

2. The best management option for the Central region is to be found for a 20 year planning horizon. Hence, the decision nodes are set as follows: $t = 2020$, $q = Central$, and $m$ set in turn to each of its values.
3. A Monte Carlo estimate of the expected loss under each option is computed from 100 simulated realizations drawn from the influence diagram.

4. The option that minimizes the expected loss is increase anti-poaching enforcement.

5. The standard deviations indicate this expected loss is significantly lower at the $\alpha = .1$ level than those of the other options.

**Conclusions**

1. An ID allows more accurate representation of expert opinions in a decision support system than many other methods due to the modular nature of the IDs knowledge base.

2. The marginal distribution of the IDs output nodes can be used to approximate the marginal expected value, e.g. $N_t$.

3. The effect of interventions and management activities can be studied using either (a) best-available qualitative knowledge incorporated into a discrete or conditional gaussian ID, or (b) best-available population dynamics or other quantitative models incorporated into an extended ID as in Haas (2001).

4. Theoretically sound statistical estimation of an ID's parameters can be performed with an incomplete and/or small sample using either maximum likelihood, Bayesian, or a mixed method as in CA.

**References Not Included in Haas (2001)**


