Tuesday, September 6: Introduction

Welcome to Mathematics 343, Probability and Statistics. This handout has some important information for you as we begin this course.

Important Information

- Instructor: Randall Pruim
  Email: rpruim@calvin.edu (generally the best way to reach me)
  Office: North Hall 284
  Phone: 526-7113

- Office Hours (tentative): 10:30–11:30 am MTWF.
  I will usually be in my office at these times. Check with me in class to make sure. Other times can also be arranged, just catch me after class or send me an email to make an appointment.

- Course web page: http://www.calvin.edu/~rpruim/courses/m343/F05/
  Here you will find a course calendar, information about homework and exams, grading policies, materials from class (like this handout, R code from examples), links to the web, etc. You can also find this site from the Mathematics and Statistics web page (click on ‘courses’).


Assignments

1. For tomorrow:
   - Fill in and return the information sheet.
   - Buy the course text, read sections 2.1–2.2, and do the following problems from section 2.2: (1), 2, (3), [4], (5), 8, (11), (13), 15, 16, (17), 18, (19), 20
   - Come to class tomorrow prepared to ask any questions you have about this material.
   - Based on the cartoons (reverse side), your experience, and the reading, see if you can answer the following question: What is randomness? Include your answer on your homework sheet.
   - Visit the course web page. Look things over. Ask if you have questions.

2. For Friday:
   - Read L&M 1.1, 2.3
   - Summarize (briefly) the origins of modern statistics. Be sure to include the important people, places, times, and ideas. [You may present this information in any clear manner you like: outline, paragraph, annotated timeline, etc.]
   - What role do the authors claim that Christianity played in the development of the discipline of statistics? Summarize their argument in a paragraph or two.
   - Optional: Download and install the statistical package R (from http://www.r-project.org/).
Course Outline (Brief)

Mathematics 343 (Probability and Statistics) and 344 (Mathematical Statistics) form a two-course sequence providing a solid introduction to probability and statistics. The first semester (343) is primarily probability, but always with an eye toward statistical applications. Near the end of the semester we will get our first official taste of statistical inference.

In the second semester (344), we will learn about some of the most important statistical procedures. These are procedures that are widely used in a variety of disciplines. In addition, our study of these procedures will provide a basis from which one can learn about the many additional statistical methods that have been developed over the last 100 years or so, as well as to learn more about the particular procedures covered in these courses.

In outline form, this two-semester sequence looks something like this:

1. Introduction [chapter 1]
   - a bit of history
   - a preview of the kinds of questions statistics seeks to answer

   - basics of mathematical probability
   - important probability distributions and theory for statistics
   - Central Limit Theorem (with a name like that it must be important)

   - the two “biggies”:
     - confidence intervals (data-based estimation)
     - hypothesis testing (data-based decision making)
   - some statistical theory (e.g., likelihood-based statistics)
   - several examples of statistical inference in particular settings
     - will include chi-square procedures, regression, ANOVA, and some non-parametric methods

Dilbert cartoons here
Monday, September 13

Today we discussed **discrete** and **continuous probability functions** and saw several examples.

**Examples**

*How long until a head?* Geometric distribution

- similar to Example 2.4.1
- also gave proof of formula for geometric sums and series

*Old Men Dying* – Poisson distribution

The data come from Case Study 2.4.1 in the text. The function `Deaths()` is defined in `sep13.R`.

- Use Taylor Series to motivate probability function for **Poisson distribution**
- used trial and error to find a Poisson distribution that was a good fit for the data
- learned how a **probability histogram** can be used to represent a probability function graphically.

\[
\text{Deaths(c(0.75,0.78,1,1.5,2))}
\]

<table>
<thead>
<tr>
<th>num.deaths</th>
<th>num.days</th>
<th>prop.data</th>
<th>lambda=0.75</th>
<th>lambda=0.78</th>
<th>lambda=1</th>
<th>lambda=1.5</th>
<th>lambda=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>484</td>
<td>0.4416058</td>
<td>0.47236655</td>
<td>0.45840601</td>
<td>0.36787944</td>
<td>0.223130</td>
<td>0.135335</td>
</tr>
<tr>
<td>1</td>
<td>391</td>
<td>0.3567518</td>
<td>0.35427491</td>
<td>0.35755669</td>
<td>0.36787944</td>
<td>0.334695</td>
<td>0.270671</td>
</tr>
<tr>
<td>2</td>
<td>164</td>
<td>0.1496350</td>
<td>0.13285309</td>
<td>0.13944711</td>
<td>0.18393972</td>
<td>0.251021</td>
<td>0.270671</td>
</tr>
<tr>
<td>3</td>
<td>45</td>
<td>0.0410584</td>
<td>0.03321327</td>
<td>0.03625625</td>
<td>0.06131324</td>
<td>0.125511</td>
<td>0.180447</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>0.0100365</td>
<td>0.00622749</td>
<td>0.00706997</td>
<td>0.01532831</td>
<td>0.047067</td>
<td>0.090224</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.0090124</td>
<td>0.00093412</td>
<td>0.00110292</td>
<td>0.00306566</td>
<td>0.014120</td>
<td>0.036089</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.0000000</td>
<td>0.00013055</td>
<td>0.00016106</td>
<td>0.00059418</td>
<td>0.004456</td>
<td>0.016564</td>
</tr>
</tbody>
</table>

Our first continuous distribution – a uniform distribution


Tuesday, November 1

Review of Power Series

A power series (about \( t = 0 \)) is a series with the following form:

\[
A(t) = \sum_{k=0}^{\infty} a_k t^k
\]

1. A power series is a representation of a function since for different values of \( t \) the series might converge to different values (or fail to converge at all).

2. Convergence: For any power series there is a real number \( R \) (possibly \( R = \infty \)) such that
   - \( |t| < R \) \( \Rightarrow \) the power series converges;
   - \( |t| > R \) \( \Rightarrow \) the power series diverges;

   \( R \) is called the radius of convergence.

3. Derivatives: If a power series converges on \((-R, R)\), then it has derivatives of all orders on \((-R, R)\). The power series for the derivatives can be obtained by differentiating term by term. For example, the first derivative is

\[
\frac{dA}{dt} = a_1 + 2a_2 t + 3a_3 t^2 + \cdots = \sum_{k=1}^{\infty} k a_k t^{k-1} = \sum_{k=0}^{\infty} (k+1) a_{k+1} t^k.
\]

Repeated differentiation shows that

\[
A^{(k)}(0) = k! a_k \quad \text{i.e.,} \quad a_k = \frac{A^{(k)}(0)}{k!}.
\] (1)

- We can use this to obtain a power series representation of a function that has derivatives of all orders in an interval around 0 since the derivatives at 0 tell us the coefficients of the power series, although this doesn’t guarantee convergence. [Taylor Series]

4. Uniqueness: If two power series converge to the same sum on some interval \((-R, R)\), then all the coefficients are equal.

5. Multiplication: If \( A(t) = \sum a_k t^k \) and \( B(t) = \sum b_k t^k \), and both converge on some interval \((-R, R)\), then \( A(t) \cdot B(t) = \sum c_k t^k \) also converges on \((-R, R)\) where

\[
c_k = a_0 b_k + a_1 b_{k-1} + \cdots + a_k b_0 \quad k = 0, 1, 2, \ldots
\]

The sequence \( \{c_k\} \) is called the convolution of the sequences \( \{a_k\} \) and \( \{b_k\} \).
Generating Functions

In calculus, we used Taylor’s Theorem to find a power series for a given function (3 above). Now we want to do the reverse. Instead of using the power series to learn about the function it converges to, we want to use the function a power series converges to to learn about the sequence of coefficients of the series. In this case we refer to the power series function as the generating function of the sequence because we will see how to use it to generate the terms of the sequence.

Actually, there are two kinds of generating functions

1. the ordinary generating function for \( \{a_k\} \) (as described above):

\[
A(t) = \sum a_k t^k
\]

2. the exponential generating function for \( \{a_k\} \) is the ordinary generating function for \( \{\frac{a_k}{k!}\} \):

\[
B(t) = \sum \frac{a_k}{k!} t^k
\]

The difference is whether or not we consider \( \frac{1}{k!} \) in Equation 1 to be part of the sequence we are interested in or not.

Obtaining Generating Functions from a Sequence

Here are some examples.

1. Let \( \{a_k\} = 1, 1, 1, \ldots \)
   - ordinary: \( A(t) = 1 + t + t^2 + t^3 + \cdots = \frac{1}{1-t} \).
   - exponential: \( B(t) = 1 + t + \frac{1}{2!} t^2 + \frac{1}{3!} t^3 + \cdots = e^t \).

2. Let \( \{a_k\} = \{2^k\} = 1, 2, 4, \ldots \)
   - ordinary: \( A(t) = 1 + 2t + 4t^2 + 8t^3 + \cdots = 1 + 2t + (2t)^2 + (2t)^3 + \cdots = \frac{1}{1-2t} \).
   - exponential: \( B(t) = 1 + 2t + \frac{1}{2!} 2^2 t^2 + \frac{1}{3!} 2^3 t^3 + \cdots = e^{2t} \).

Obtaining Sequences from Generating Functions

Now suppose we know the generating function \( A(t) \), how do we get the sequence? We use Equation 1:

- ordinary: \( a_k = A^{(k)}(0) \).

- exponential \( \frac{a_k}{k!} = \frac{A^{(k)}(0)}{k!} \), so \( a_k = A^{(k)}(0) \).

So for example, if the (ordinary) generating function is \( A(t) = \frac{1}{1-3t} \), then \( A^{(k)}(t) = k!3^k (1-3t)^{-k-1} \), so \( A^{(k)}(0) = k!3^k \) and \( a_k = 3^k \). (This would be the exponential generating function for the sequence \( \{k!3^k\} \), which probably isn’t as interesting.)
Moment Generating Functions

We want to apply this to the sequence of moments of a pdf. In this case, the exponential generating function is easier to work with than the ordinary generating function. So we define the moment generating function of a pdf to be the exponential generating function of the sequence of moments \( \{\mu_k\} \):

\[
M_W(t) = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} t^k
\]

where we define \( \mu_0 = 1 \) for the purposes of the series above.

Now, if we can just get our hands on \( M_W(t) \), and if it is reasonably easy to differentiate, then we can get all of our moments via differentiation instead of summation or integration. What we need is a function \( f(t) \) such that \( f^{(k)}(0) = \mu_k \). (See Equation 1 again.) And the function is (drum roll please . . . ):

\[
M_W(t) = E(e^{tW})
\]

The easiest way to see this is to simply check that it works. If \( W \) is continuous, then we have (limits of integration omitted)

\[
E(e^{tW}) = \int e^{tw} f_W(w) \, dw
\]

\[
\frac{d}{dt} E(e^{tW}) = \frac{d}{dt} \int e^{tw} f_W(w) \, dw = \int \frac{d}{dt} e^{tw} f_W(w) \, dw = \int we^{tw} f_W(w) \, dw
\]

\[
\frac{d}{dt} E(e^{tW}) \bigg|_{t=0} = \int w e^{0w} f_W(w) \, dw = E(W) = \mu_1
\]

\[
\frac{d^2}{dt^2} E(e^{tW}) \bigg|_{t=0} = \int w^2 e^{tw} f_W(w) \, dw \bigg|_{t=0} = E(W^2) = \mu_2
\]

\[
\frac{d^i}{dt^i} E(e^{tW}) \bigg|_{t=0} = \int w^i e^{tw} f_W(w) \, dw \bigg|_{t=0} = E(W^i) = \mu_i
\]

For discrete random variables, we replace the integrals above with sums.

\[
E(e^{tW}) = \sum e^{tk} p_W(k)
\]

\[
\frac{d}{dt} E(e^{tW}) = \frac{d}{dt} \sum e^{tk} p_W(k) = \sum \frac{d}{dt} e^{tk} p_W(k) = \sum ke^{tk} p_W(k)
\]

\[
\frac{d}{dt} E(e^{tW}) \bigg|_{t=0} = \sum ke^{0k} p_W(k) = E(W) = \mu_1
\]

\[
\frac{d^i}{dt^i} E(e^{tW}) \bigg|_{t=0} = \sum k^i e^{tk} p_W(k) \bigg|_{t=0} = E(W^i) = \mu_i
\]
Example: Exponential Distributions

Let \( Y \) be a continuous random variable with pdf \( f_Y(y) = \lambda e^{-\lambda y}, \ y > 0 \). We can use the moment generating function to determine \( E(Y) \) and \( \text{Var}(Y) \).

\[
M_Y(t) = E(e^{tY}) = \int_0^\infty e^{ty} \lambda e^{-\lambda y} \, dy = \int_0^\infty \lambda e^{(t-\lambda)y} \, dy
\]

Now suppose \( t < \lambda \) and let \( u = (t-\lambda)y \), so \( du = (t-\lambda)dy \) and the integral becomes

\[
\int_{-\infty}^\infty \frac{\lambda}{t-\lambda} e^u \, du = \frac{\lambda}{\lambda-t} = \frac{1}{1-t/\lambda}
\]

\( (M_Y(y) \) doesn’t exist for \(|t| \geq \lambda \).)

Now we can determine the moments of \( Y \) by differentiating:

- \( M_Y'(t) = \lambda(-1)(\lambda - t)^{-2}(-1) = \lambda(\lambda - t)^{-2} \), so
  - \( E(Y) = \mu_1 = M_Y'(0) = 1/\lambda \).

- \( M_Y''(t) = \lambda(-2)(\lambda - t)^{-3}(-1) = 2\frac{\lambda}{(\lambda-t)^3} \)
  - \( E(Y^2) = \mu_2 = M_Y''(0) = 2/\lambda^2 \).

- \( \text{Var}(Y) = \mu_2 - \mu_1^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2 \).
Wednesday, November 2

Review of Moment Generating Functions

• A moment generating function $M_W(t)$ is the exponential generating function of the sequence of moments of the pdf for $W$.

• $M_W(t) = E(e^{tW})$, which gives us a way to compute the moment generating function in many important cases.

• The moment generating function may not exist (because the moments don’t exist or because the power series doesn’t converge).

• The moments of $W$ can be recovered from the mgf using the identity

$$M^{(k)}(0) = \mu_k = E(W^k)$$

Example: Binomial Distributions

Let $X \sim \text{Bin}(n, p)$. Then $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$.

$$E(e^{tX}) = \sum e^{tk} p_X(k) = \sum e^{tk} \binom{n}{k} p^k (1-p)^{n-k}$$
$$= \sum \binom{n}{k} e^{tk} p^k (1-p)^{n-k}$$
$$= \sum \binom{n}{k} (e^t p)^k (1-p)^{n-k}$$
$$= (pe^t + 1 - p)^n$$

for all values of $t$.

Now we can find $E(X)$ by differentiating:

• $M'_X(t) = n(pe^t + 1 - p)^{n-1}(pe^t)$

• $M'_X(0) = n(p + 1 - p)^{n-1}(p) = np$, as we already knew.
Moment Generating Functions for Independent Sums

An especially important situation in statistics is the sum of two or more independent (often iid) random variables. Notice that if $X$ and $Y$ are independent, then

$$E(e^{t(X+Y)}) = E(e^{tX+Y}) = E(e^{tX} \cdot e^{tY}) = E(e^{tX}) \cdot E(e^{tY})$$

This means that the moment generating function for a sum is the product of moment generating functions. This fact will allow us to get moment generating functions easily in many situations of interest.

Example: Sum of Two Binomials

Suppose $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = (pe^t + 1 - p)^n \cdot (pe^t + 1 - p)^m = (pe^t + 1 - p)^{n+m}$$

Notice that this is the mgf for a binomial random variable with parameters $n + m$ and $p$. In fact, the distribution of $X + Y$ is binomial with those parameters, and this is an instance of the following general result (which we won’t prove):

**Uniqueness Property for Moment Generating Functions**

If $W_1$ and $W_2$ have the same moment generating function on an interval containing 0, then $W_1$ and $W_2$ have the same pdf, i.e., they are the same random variable.

This means that if we can recognize the mgf, we know what the distribution is.

Moment Generating Functions for Linear Transformations

If $a$ and $b$ are constants, then

- $M_{aW}(t) = E(e^{atW}) = M_W(at)$.
- $M_{W+b}(t) = E(e^{t(W+b)}) = E(e^{bt} \cdot e^{tW}) = e^{bt} M_W(t)$.
- so $M_{aW+b}(t) = e^{bt} M_W(at)$.

Examples

1. Let $X \sim \text{Bin}(n, p)$ and let $Y = 2X + 3$, then the mgf for $Y$ is

$$e^{3t} M_X(2t) = e^{3t} \cdot (pe^{2t} + 1 - p)^n$$

2. Let $X$ be an exponential random variable with parameter $\lambda$, i.e., $M_X(t) = \frac{\lambda}{\lambda - t}$. Let $Y = kX$ for some constant $k > 0$. Then

$$M_Y(t) = M_X(kt) = \frac{\lambda}{\lambda - kt} = \frac{\lambda/k}{\lambda/k - t}$$

Notice that is the mgf for an exponential function with parameter $\lambda/k$. So all the exponential distributions are simply rescalings of each other.
Friday, November 4

Modeling with the Poisson Distribution

The Poisson distribution is a good model for a wide variety of situations. Here are some examples where a Poisson model has been used:

- The number of men killed by being kicked by a horse in a given Prussian army corps in a given year.
- The number of alpha particles given off by a radioactive source and detected by a Geiger counter in a fixed length of time.
- The number of fumbles in NCAA football games.
- The number of cars that pass a given point on a road in a fixed length of time under conditions where traffic is light (but not when traffic is heavy).
- The number of telephone calls coming in to a given exchange between 8:00 and 8:15am.
- The number of freak accidents (like falling in the tub) that an insurance company will have to pay a claim for in a given month.

So why does the Poisson distribution model these kinds of situations? And how do we know which value of $\lambda$ to use? The answer comes from investigating something called a Poisson process. In a Poisson process we are interested in counting the number of “occurrences” (times something happens) in a situation where the following properties hold:

- Our random variable counts the number of occurrences that take place in a fixed amount of time or space. (We’ll use time as we describe what follows to make the description simpler.)

- If we divide our total time into $n$ very small equally-sized intervals, then as $n$ gets large, the following are very nearly true:
  - There is either 0 or 1 occurrence in each small interval. (Not enough time for 2 or more.)
  - The number of occurrences in each small interval is independent of the others.
  - The probability of an occurrence is proportional to the size of the interval, namely $\lambda/n$ for some constant $\lambda$.

If these conditions are (approximately) met, we can work out some probabilities. Let $X$ be the number of occurrences, then

$$P(X = k) \approx \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$= \frac{n!}{n!(n-k)!} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$\rightarrow e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{as } n \rightarrow \infty$$
Poisson Approximation to Binomial – The Law of Small Numbers

Notice that the derivation above can be interpreted to say that the Poisson distribution is a good approximation of the binomial distribution provided \( n \) is large, \( p \) is small, and \( \lambda = np \). In fact, the Poisson distribution is a kind of limit of the binomial distribution if we let \( n \) increase to infinity and adjust \( p \) so that \( pn \) is a constant (namely \( \lambda \)) for all \( n \).

Historically, this is how the Poisson distribution first arose (in 1898 by L. v. Bortkiewicz in a paper entitled Das Gesetz der kleinen Zahlen or The Law of Small Numbers.) Binomial probabilities are difficult to compute when \( n \) is large because of the large factorials and exponents involved. (Imagine trying to compute the probability of getting between 5 and 10 successes in 1000 tries when the probability of success in each try is 1/100.) In this case, it is much easier to compute the corresponding Poisson probability.

Relationship between Poisson and Exponential Distributions

The exponential distribution can be derived from what we know about the Poisson distribution. Let’s suppose we are in a situation where the Poisson distribution is a good model for the number of occurrences of some event in a given amount of time or space. A Poisson random variable \( X \) counts the number of occurrences in a fixed amount of time. As we saw earlier, the pdf for \( X \) is

\[
    f(x; \lambda) = \begin{cases} 
    \frac{e^{-\lambda x} \lambda^x}{x!} & \text{if } x \text{ is a non-negative integer} \\
    0 & \text{otherwise}
    \end{cases}
\]

The parameter \( \lambda = E(X) \) and is sometimes called the \textit{intensity parameter} and gives the “average” number of occurrences.

A related variable is the continuous rv \( Y \) which measures the amount of time until the next occurrence of an event. If events occur at an average rate of \( \lambda \) events per unit time, then the number of events in the time interval \([0, t] \) (now until \( t \) time units from now) is a Poisson rv with expected value \( \lambda t \). We can use this to determine the cdf and pdf for \( Y \) as follows:

\[
    F_Y(t) = P(Y \leq t) = 1 - P(Y > t)
    = 1 - P(X = 0)
    = 1 - e^{-\lambda t} \frac{\lambda^0}{0!}
    = 1 - e^{-\lambda t}
\]

Notice that the exponent is \(-\lambda t\) because \( X \) is Poisson with intensity parameter \( \lambda t \).

So the cdf for \( Y \) is \( F_Y; \lambda) = 1 - e^{-\lambda t} \). We can get the pdf for \( Y \) by differentiating:

\[
    \text{pdf for } Y = f_Y(t; \lambda) = \frac{d}{dt} f(t; \lambda) = \lambda e^{-\lambda t}
\]

The choice of variable \( t \) (which reminds us of \textit{time}) is, of course, arbitrary and may be replaced by the more usual \( x \) or \( y \) if we prefer.
What We Already Know About Poisson and Exponential Distributions

<table>
<thead>
<tr>
<th>distribution</th>
<th>pdf</th>
<th>mean</th>
<th>variance</th>
<th>mgf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$e^{-\lambda \frac{k!}{k}}$</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
<td>$e^{-\lambda + \lambda e^\lambda}$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\lambda e^{-\lambda x}$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda^2}$</td>
<td>$\frac{\lambda}{\lambda - t}$</td>
</tr>
</tbody>
</table>

Examples

Example 1: At the Bank

Suppose that customers at a bank arrive independently at an average rate of 10 per hour and that the Poisson (and exponential) distributions are good models for the number of arrivals in a fixed time (and the time until the next arrival). How do we determine the probability that no customers arrive in the next ten minutes?

- Using the exponential distribution.
  If we measure time in hours, then we are looking for $P(Y > \frac{1}{6})$, where $Y$ is exponentially distributed with $\lambda = 10 =$ the expected number of customers per hour.\(^1\) Using the results above we see that

  $$P(Y > \frac{1}{6}) = 1 - P(Y \leq \frac{1}{6}) = 1 - (1 - e^{-\lambda \frac{1}{6}}) = e^{-\lambda \frac{1}{6}} = e^{-10 \times \frac{1}{6}} \approx 0.188$$

  So nearly 20% of the time there will be a 10 minute gap between customers.

- Using the Poisson distribution.
  On average there will be $10/6 = 5/3$ customers in a 10 minute time span. So we are interested in $P(X = 0)$ where $X$ is a Poisson random variable with intensity parameter $5/3$. So

  $$P(X = 0) = e^{-5/3} \frac{(5/3)^0}{0!} = e^{-5/3} \approx 0.188$$

---

\(^1\)A couple notes here: First, $E(Y) = \frac{1}{\frac{1}{6}} = 3$ makes sense. If customers arrive at a rate of 10 per hour, we expect a customer about every tenth of an hour (every 6 minutes).

Second, we could use minutes instead of hours throughout. This would be modeled by a rv $Y_2$ that is exponentially distributed with $\lambda = \frac{1}{6}$ and we would want to determine $P(Y_2 \geq 10)$. You should be able to see pretty easily that we get the same probability either way.
Tuesday, November 29

Confidence Intervals: General Framework

A confidence interval for a parameter \( \theta \) is an interval \((A, B)\) where \(A\) and \(B\) are functions of

- the data, and
- a confidence level \( L = 1 - \alpha \),

such that if we consider \(A\) and \(B\) to be random variables (random because they depend on the data, which in turn depend on random sampling), then

\[
P(\theta \in (A, B)) = L = 1 - \alpha .
\]

It is important to remember what is random and what is not in the probability statement above.

- the data are random (depends on sampling)
- the interval is random (depends on the data)
- the parameter \( \theta \) IS NOT RANDOM, it is an unknown constant

This means that the probability statements are about the interval, not about the parameter. For example, if we compute many 95% confidence intervals, roughly 95% of these intervals will correctly contain the parameter \( \theta \), and roughly 5% will not. Of course, we won’t know which confidence intervals are correct.

The key to forming confidence intervals is to understand the sampling distribution, the distribution of some sample statistic (\( \hat{\theta} \)). As we shall see, if \( \hat{\theta} \) is an unbiased estimator for \( \theta \), then smaller values of \( \text{Var}(\hat{\theta}) \) will produce smaller confidence intervals since \( \hat{\theta} \) is more likely to be close to its mean \( \theta \).

Confidence Intervals: Proportions

- Parameter of interest: \( p \), a proportion
- Estimator: \( \hat{p} = X/n \) \([X \text{ is sample count; } n \text{ is sample size}]\)
- Sampling distributions:
  - \( X \sim \text{Bin}(n, p) \approx N(np, np(1-p)) \)
  - \( \hat{p} = X/n \approx N(p, p(1-p)/n) \)
  - \( \hat{p} - p \approx N(0, p(1-p)/n) \)

The last distribution above can be used to get confidence intervals for \( p \). If we want a \( 1 - \alpha \) confidence interval, we can use the critical value \( z_{\alpha/2} \) (chosen so \( P(Z \geq z_{\alpha/2}) = \alpha/2 \)). The set of \( p \)'s in our confidence interval are the \( p \)'s such that

\[
-z_{\alpha/2} \sqrt{p(1-p)/n} \leq \hat{p} - p \leq z_{\alpha/2} \sqrt{p(1-p)/n}
\]

We would like to express this as an interval of values of \( p \). There are two ways to do this:
1. The traditional, less exact, and somewhat simpler way:

We use the fact that \( p(1 - p) \approx \hat{p}(1 - \hat{p}) \). Then our inequality becomes:

\[
|p - \hat{p}| \leq z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

i.e.,

\[
\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

so our confidence interval is

\[
\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right)
\]

which is often abbreviated as

\[
\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

or

\[
\hat{p} \pm z_{\alpha/2} SE
\]

where SE standards for **standard error** and is an approximation to the standard deviation of the sampling distribution for \( \overline{Y} \).

2. A more exact, but somewhat messier way.

Note that

\[
-z_{\alpha/2} \sqrt{p(1 - p)/n} \leq \hat{p} - p \leq z_{\alpha/2} \sqrt{p(1 - p)/n} \quad \Rightarrow \quad |\hat{p} - p| \leq z_{\alpha/2} \sqrt{p(1 - p)/n}
\]

\[
\Rightarrow (\hat{p} - p)^2 \leq z_{\alpha/2}^2 p(1 - p)/n
\]

This is quadratic in \( p \), so we can solve

\[
(\hat{p} - p)^2 = (z_{\alpha/2})^2 p(1 - p)/n
\]

using the quadratic formula to find the largest and smallest values of \( p \) that satisfy the inequalities. I’ll spare you the (straightforward) algebra and show you the result: The endpoints of the interval are

\[
\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm \frac{z_{\alpha/2}}{2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}
\]

\[
1 + \frac{z_{\alpha/2}^2}{n}
\]

Notice that the limit as \( n \to \infty \) of this expression is the simpler confidence interval we gave above. Thus for large values of \( n \), the two confidence intervals will be nearly identical. For smaller values of \( n \), method 2 performs better in the sense that confidence intervals formed this way have an actual confidence level nearer to the nominal level.
Example

1. A public opinion poll is conducted. 132 out of 350 people surveyed say they approve of the job the president is doing. Give a 95% confidence interval for the president’s job approval rating.

Answer: \( \hat{p} = \frac{132}{350} = 0.37714, \) \( n = 350, \) so \( SE = \sqrt{\frac{(0.377)(1-0.377)}{350}} \approx 0.026 \) and our confidence interval is

\[ 0.377 \pm 1.96(0.026) = 0.377 \pm 0.051 = (0.326, 0.428) \]

A newspaper would probably report this as a job approval rating of 38% with a margin of error of ±5%.

Confidence Intervals: Sampling from Normal Distributions

We can play the same game when sampling from a normal population. Recall that

\[ \bar{Y} \sim N(\mu, \frac{\sigma^2}{n}) \]

where \( \mu \) and \( \sigma^2 \) are the mean and the variance of the population.\(^2\) Thus

\[ \mu - \bar{Y} \sim N(0, \frac{\sigma^2}{n}) \]

so

\[ P \left( -\frac{z_{\alpha/2} \sigma}{\sqrt{n}} \leq \mu - \bar{Y} \leq \frac{z_{\alpha/2} \sigma}{\sqrt{n}} \right) = 1 - \alpha \]

that is,

\[ P \left( \bar{Y} - \frac{z_{\alpha/2} \sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{z_{\alpha/2} \sigma}{\sqrt{n}} \right) = 1 - \alpha \]

so given some data our confidence interval is

\[ \left( \bar{y} - \frac{z_{\alpha/2} \sigma}{\sqrt{n}}, \bar{y} + \frac{z_{\alpha/2} \sigma}{\sqrt{n}} \right) \]

which we can also write as

\[ \bar{y} \pm \frac{z_{\alpha/2} \sigma}{\sqrt{n}}, \]

or

\[ \boxed{\bar{y} \pm z_{\alpha/2}SD} \]

where \( SD \) is the standard deviation of the sampling distribution.

Examples

2. Suppose we sample 4 values from a normal population with variance \( \sigma^2 = 9. \) Our data are 5, 10, 11, 12. What is the 95% confidence interval computed from this data?

Answer: \( z_{0.025} = 1.96, \) and \( \bar{y} = 9.5, \) so our confidence interval is

\[ 9.5 \pm 1.96 \left( \frac{3}{\sqrt{4}} \right) = 9.5 \pm 2.94 = (6.56, 12.44) \]

\(^2\)Recall too that the **Central Limit Theorem** tells us that this will be approximately correct for any distribution provided \( n \) is large enough.
3. Suppose we sample 50 values from a normal population with variance $\sigma^2 = 9$. The mean of our sample is 9.5. What is the 90% confidence interval computed from this data?

Answer: $z_{0.05} = 1.645$, so our confidence interval is

$$9.5 \pm 1.645 \left( \frac{3}{\sqrt{50}} \right) = 9.5 \pm 0.698 = (8.8021, 10.198)$$

4. Suppose we know that the population variance is $\sigma^2 = 9$. We want to compute a 95% confidence interval with margin of error at most 0.25. How large must the sample be?

Answer: $z_{0.025} = 1.96$, so our confidence interval is

$$9.5 \pm 1.96 \left( \frac{3}{\sqrt{n}} \right)$$

We want the margin of error, $1.96 \left( \frac{3}{\sqrt{n}} \right)$ to be at most 0.25, so we solve

$$0.25 = 1.96 \left( \frac{3}{\sqrt{n}} \right)$$

for $n$, and find that $n$ must be at least $\left( \frac{3(1.96)}{0.25} \right)^2 \approx 554$ (rounding up to nearest integer). So we need a sample of size 554 or larger.

Notes

- Notice that the forms of the confidence intervals for means and for proportions are very similar (see the boxed expressions).

- Notice that the width of a confidence interval depends on
  - sample size: larger $n$ $\implies$ smaller interval
  - population variance: smaller variance $\implies$ smaller interval
  - confidence level: smaller confidence $\implies$ smaller interval

- The confidence interval for means that we have just presented is almost never used in practice because we almost never know $\sigma^2$ for the population.

You can probably guess what we do instead: we estimate $\sigma^2$ (using our unbiased estimator). The resulting sampling distribution is no longer normal, however. We will learn about the distribution we need to handle this more applicable case (called Student’s $t$ distribution) next semester.
Asymmetric Confidence Intervals

All of the confidence intervals we have done so far have been symmetric in the sense that they were based using the same critical value $z_{\alpha/2}$ to get both the lower and upper end points of the confidence interval. This isn’t required. For example, we could get an unusual 95% confidence interval as follows.

$$
\bar{y} + z_{0.01}SD \leq \mu \leq \bar{y} + z_{0.04}SD
$$

instead of

$$
\bar{y} + z_{0.025}SD \leq \mu \leq \bar{y} + z_{0.025}SD
$$

The important thing is that the two tail probabilities together add to 5% so that the middle portion is 95%. The example above would be a bit strange, perhaps, but there is one common use of asymmetric confidence intervals:

$$
\bar{y} - z_0SD \leq \mu \leq \bar{y} + z_\alpha SD
$$

or

$$
\bar{y} - z_\alpha SD \leq \mu \leq \bar{y} + z_0 SD
$$

produce what are called one-sided confidence intervals. Note that we are abusing notation here a little bit by letting $z_0 = \infty$.

Example

5. Suppose we sample 50 values from a normal population with variance $\sigma^2 = 9$. The mean of our sample is 9.5. Calculate the two one-sided 95% confidence intervals based on this data.

Answer: $z_{0.05} = 1.645$, so our confidence intervals are

$$
\left(9.5 - 1.645 \left(\frac{3}{\sqrt{50}}\right), \infty\right) = (8.8021, \infty)
$$

$$
\left(-\infty, 9.5 + 1.645 \left(\frac{3}{\sqrt{50}}\right)\right) = (-\infty, 10.198)
$$

Compare this with Example 3 from yesterday. What do you notice? Why?