MATH 321: Mathematical Statistics Lectures

Colton Gearhart

April 22, 2024

Contents

Test 1

Contents

Lecture 14 – Bivariate Distributions

MATH 321: Mathematical Statistics

Lecture 14: Bivariate Distributions

Chapter 4: Bivariate Distributions (4.1 and 4.4)

Introduction

- In previous chapters, we have discussed probability models and computation of probability for events involving only one random variable. These are called **univariate models**.
- In this chapter, we will discuss models that involve more than one random variable called **multivariate models**.
- Examples:
  - Univariate: The body weights of several people in the population is measured.
  - Multivariate: Temperature, height, and blood pressure, in addition to weight, are measured. These observations on different characteristics could also be modeled as observations on different random variables.
- We need to know how to describe and use probability models that deal with more than one random variable at a time.

  The basic setting for multivariate random variables is the same as those for univariate random variables.
- Definition of a random variable:
  - A random variable is a function from the sample space S to $\mathbb{R}$.

    Univariate:

    Multivariate:

  - We could extend this for an n-dimensional random vector:

– We will mainly discuss **bivariate models**, involving two random variables. With each point in a sample space, we associate an ordered pair of of numbers $(x, y) \in \mathbb{R}^2$, where $\mathbb{R}^2$ denotes the plane $\Longrightarrow$ Possible range is the (x, y) coordinate plane.

Joint probability function for discrete random variables

Example

- An investor owns two assets. They are interested in the value of each of them during one year. It is not enough to know the separate probability distributions, they want to know how the two assets behave together.
- This requires a joint probability distribution for X and Y, which can be specified in a table:

y \ x	90	100	110
0	.05	.27	.18
10	.15	.33	.02

Definition

- The random vector (X, Y) is called a **discrete random vector** if it has only a _____ number of possible values (i.e. _____ range).

- Definition: Let (X, Y) be a discrete bivariate random vector. Then the **joint probability mass function (joint pmf)** is defined as $f(x, y) = P(X = x, Y = y)$ for all $(x, y) \in \mathbb{R}^2$ and has properties
  1. $f_{X,Y}(x, y)$ for all x, y.
  2. $\sum\sum f_{X,Y}(x, y) =$

     where the sum is over all value (x, y) that are assigned nonzero probabilities.
  3. Let A be any subset of $\mathbb{R}^2$, then

     $P((X, Y) \in A) = \sum\sum f(x, y)$.

- The joint pmf of (X, Y) completely defines the probability distribution of the random vector (X, Y), just as the pmf of a discrete univariate random variable does.

Examples

1. Consider the experiment of tossing two fair 3-sided die. Let X = sum of two dice and Y = | difference of the two dice |.

   (a) Find the sample space and the range of (X, Y).

   (b) Find the joint pmf of (X, Y).

   Recall the **sample point method** for probabilities: $f(x, y) = \frac{\#\text{ outcomes in } (x,y)}{\text{total } \#\text{ outcomes}}$

   (c) Find $P(3Y \geq X)$.

2. Joint probability functions for discrete random variables are often given in tables, but they may also be given in formulas.

   An analyst is studying traffic accidents in two adjacent towns. The random variables X and represents the number of accidents in a day in towns X and Y, respectively. The joint probability function for X and Y is given by:

   $f(x, y) = \frac{e^{-2}}{x!y!}$ for $x = 0, 1, 2, \ldots$ and $y = 0, 1, 2, \ldots$.

   Find $P(X = 1, Y < 2)$.

Marginal distributions for discrete random variables

Motivation and example

- Even if we are considering a probability model for a random vector (X, Y), there may be probabilities of interest that involve only one of the random variables in the vector. For example, $P(X = 2)$.
- $\{X = x\}$ suggests that Y _____ as long as the condition on X is met. This corresponds to the joint event:

  $\{X = x\} =$
- Once we know the joint distribution, it is really easy to find the probabilities for individual values of X and Y.

  Joint pmf table: Marginals =
- Back to the investor example, if they want to know how each individual asset is behaving:

y \ x	90	100	110
0	.05	.27	.18
10	.15	.33	.02

Definition

- Let (X, Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then, the **marginal pmfs** of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$ are given by

$$f_X(x) = \sum_y f_{X,Y}(x, y) \qquad \text{and} \qquad f_Y(y) = \sum_x f_{X,Y}(x, y)$$

Example

- Given the joint random vector (X, Y), let

$$f(x, y) = \begin{cases} c(x + y) & \text{for } x = 1, 2, 3 \text{ and } y = 1, 2 \\ 0 & \text{elsewhere} \end{cases}$$

  (a) Find c so that $f(x, y)$ is a valid pmf.

(b) Find $f_X(x)$ and $f_Y(y)$ (NOTE: should be functions of ONLY x and ONLY y, respectively).

(c) Find $P(X \leq 2)$.

Probabilities for only one random variable $\Longrightarrow$ Find marginal first.

Summary

- The joint pmf of (X, Y) has more information about the distribution of (X, Y) than the marginal pmfs of X and Y alone.

  This is because it contains information about the relationship between X and Y. And can easily find the marginal distributions from the joint distribution.

Joint and marginal distributions for continuous random variables

Joint distribution definition

- We can also consider random vectors whose components are continuous random variables. The probability distribution of a continuous random vector is usually described using a pdf, as in the univariate case.
- The definitions are really the same except that integrals replace summations.
- Definition: The **joint probability density function (joint pdf)** is a function $f(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ such that
  1. $f_{X,Y}(x, y)$ for all x, y.
  2. $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y\,\mathrm{d}x = 1$ (switch order of integration)
  3. For $A \subset \mathbb{R}^2$

     $$P((X, Y) \in A) = \int\int_A f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int\int_A f(x, y)\,\mathrm{d}y\,\mathrm{d}x$$

Probabilities

- For a continuous random variable X with pdf $f_X(x)$ and $A = [a, b]$:

- Finding $P((X, Y) \in A)$ can be a bit more complicated because it involves double integration.

  For a bivariate continuous random vector (X, Y) with joint pdf $f(x, y)$ and $A = [a, b] \times [c, d]$:

- Examples:

  1. Suppose $f(x, y) = 2 - 1.2x - 0.8y \quad$ for $0 \leq x \leq 1, 0 \leq y \leq 1$.

     (a) Verify this is a valid pdf.

     (b) Rectangular region: Find $P(0.50 \leq X \leq 1, 0.50 \leq Y \leq 1)$.

(c) More general region:
Find $P(X+Y>1)$, $\quad f(x,y)=2-1.2x-0.8y \quad$ for $0\le x\le 1, 0\le y\le 1$

(1) Draw range $(\mathcal{X},\mathcal{Y})$ (2) Draw and shade conditions on (x,y)

(3) Set bounds (3) Set bounds (again)

Steps

1) Draw the region where the density function is positive (the range of X and Y).
2) Shade the region of the probability we want (make $Y<f(X)$ or $Y>f(X)$).
3) Choose the order of integration and set the bounds of the integrals. Start with the outside integral, then the inside integral.

   Suppose we choose x as the outside integral. Find the interval of x in the shaded region. Then find the interval of y in the shaded region as a function of x

("moving x").

2. Suppose $f(x,y) = 3x \quad$ for $0 \leq y \leq x \leq 1$.

   Find $P(0 \leq X \leq 0.5, Y > 0.25)$.

   (1) Draw range $(\mathcal{X}, \mathcal{Y})$ (2) Draw and shade conditions on (x, y)

   (3) Set bounds

   (3) Set bounds (again)

3. Suppose $f(x, y) = 1 \quad$ for $0 \leq x \leq 1, 0 \leq y \leq 1$.

   Find $P(XY < 0.5)$.

   – NOTE: Constant density $\Longrightarrow$ Uniform distribution (i.e. flat surface)

   Probability of a uniform distribution is just a _____

   $\mathbb{R}^1 : \text{Prob} = \frac{\text{length of interval of interest}}{\text{length of entire interval}}$

   $\mathbb{R}^2 : \text{Prob} = \frac{\text{Area of interest}}{\text{Total area}}$

Marginal distributions

- Definition: Let $f(x, y)$ be the joint pdf for the bivariate continuous random vector (X, Y). Then the **marginal pdfs** are defined by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x$$

- Examples:
  1. Continuing previous example 1: $f(x, y) = 2 - 1.2x - 0.8y$ for $0 \le x \le 1, 0 \le y \le 1$.
     (a) Find $f_X(x)$ Should be a function of ONLY x.

     (b) Find $f_Y(y)$. Should be a function of ONLY y.

  2. Suppose $f(x, y) = 1/2$ for $0 \le x \le y \le 2$.
     (a) Find $f_X(x)$.

     (b) Find $f_Y(y)$.

Expected values of functions of random variables

Introduction

- Many practical applications require the study of a function of two or more random variables. For example, for the investor that owns two assets, they may wish to find the distribution of $g(X, Y) = X + Y$, which gives the total value of the two assets.
- Other common functions include (where we also need fancier techniques to find the distribution functions):

  $XY \rightarrow$ Requires a bivariate transformation in the continuous case.

  $\min(X, Y)$ and $\max(X, Y) \rightarrow$ Order statistics.
- For now, we are not going to find distributions of functions of random variable. Rather, we are going focus on expectations of functions of random vectors $g(X, Y)$, which can be computed from the original joint distribution.

  These can be computed just as with univariate random variables.

Notation

- Sometimes it is convenient to replace the symbols X and Y representing random variables by X_1 and X_2.
- This is particularly true in situations in which we have more than two random variables. Both will be used from here on out.

Expected values of a function of a random variable

- Definition: Let $g(X, Y)$ be a function of a bivariate random vector (X, Y).
  (a) If X and Y are discrete with joint pmf $f(x, y)$,

  $$E[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y)$$

  (b) If X and Y are continuous with joint pdf $f(x, y)$,

  $$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

Examples

1. Back to the investor with two asset random variables X and Y:

y \ x	90	100	110
0	.05	.27	.18
10	.15	.33	.02

   Find $E(X+Y)$ and $E(XY)$.

2. Suppose $f(x,y) = 3x \quad$ for $0 \le y \le x \le 1$.
   Find $E(X^2Y^2)$.

Special expectations

- Just like with probabilities, even if we are working with random vectors, we may only be interested in expectations for a single variable.
- Definitions:
  (a) Let (X_1, X_2) be a bivariate discrete random vector with joint pmf $f(x_1, x_2)$.
    i) If $g(X_1, X_2) = X_1$, then

    ii) If $g(X_1, X_2) = (X_1 - \mu_1)^2$, then

    iii) If $g(X_1, X_2) = \mathrm{e}^{tX_1}$, then

    Results: The mean μ_i, variance σ_i^2 and mgf $M_{X_i}(t)$ can be computed from the joint distribution (pmf / pdf) $f(x_1, x_2)$ or the marginal distribution (pmf / pdf) $f_{X_i}(x_i)$ for $i = 1, 2$.

  (b) Same results hold in the continuous case, just replace the summation with integration.

Examples

1. Continuing with the investor with two asset random variables X and Y:

y \ x	90	100	110
0	.05	.27	.18
10	.15	.33	.02

   (a) Let $g_1(X,Y) = X$. Find $E[g_1(X,Y)]$ using the joint pmf and then using the marginal pmf.

   (b) Let $g_2(X,Y) = Y$. Find $E[g_2(X,Y)]$.

   (c) Compare $E(X+Y)$ and $E(XY)$ to their "intuitive answers".

2. Suppose $f(x,y) = 1/2 \quad$ for $0 \le x \le y \le 2$ and $g(X,Y) = Y$.
   Find $E[g(X,Y)]$ both ways (joint pdf and marginal pdf).

Expected value of $X + Y$ and XY

- Theorem: **Expected value of a sum of two random variables**.
  - Let (X, Y) be a bivariate random vector and function $g(X, Y) = X + Y$.
    $E(X + Y) = E(X) + E(Y)$
    Note: This result always holds (regardless of independence).
  - Proof:

    A similar proof is used for continuous random variables, again just replace $\sum$ with $\int$.
  - We could generalize this theorem to the following:

    If $g_1(X, Y)$ and $g_2(X, Y)$ are two functions and a, b and c are constants, then
    $E[ag_1(X, Y) + bg_2(X, Y) + c] = aE[g_1(X, Y)] + bE[g_2(X, Y)] + c$
  - $E(X + Y)$ is a special case with $g_1(X, Y) = X$, $g_2(X, Y) = Y$, $a = b = 1$, $c = 0$.
- Products of random variables are not so simple.
  The expected value of XY does *not* always equal the product of the expected values.
- Theorem: **Expected value of a product of two random variables**.
  - Let (X, Y) be a bivariate random vector and function $g(X, Y) = XY$.
    If X and Y are independent, $E(XY) = E(X) \cdot E(Y)$

– Notes:

This theorem may fail to hold if X and Y are not independent. There are examples of random variables X and Y which are not independent but the results of this theorem are still true.

It is still important to be able to compute $E(XY)$ directly when X and Y are not known to be independent, and because it is used in the calculation of covariance, which is also where we will see why the theorem doesn't go both ways.

Lecture 15 – Conditional Distributions

MATH 321: Mathematical Statistics

Lecture 15: Conditional Distributions

Chapter 4: Bivariate Distributions (4.3)

Introduction

- Oftentimes, two random variables (X, Y) are related. Knowing about the value of X gives us some information about the value of Y, even if it doesn't tell us the value Y exactly (can find $E(Y \mid X = x)$, but not the exact value of $Y \mid X = x$).
- Example: Study hours X and Test grade Y.

  $P(Y > 90 \mid X = 1 \text{ hrs})$ $\quad$ $P(Y > 90 \mid X = 5 \text{ hrs})$
- Sometimes, knowledge about X gives us no information about Y.

Discrete conditional distributions

Conditional pmf

- Recall the conditional probability of events: $P(B \mid A) = \dfrac{P(\quad\quad)}{P(\quad)}$
- Events in a conditional distribution.
  - Suppose that X and Y are discrete random variables. The conditional event of $Y = y$ given $X = x$ is

    where _____ is the conditioning event (i.e. the given event),

    and _____ is the event of interest (i.e. the event whose probability we want to know).

- Definition: Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$.

  (a) For any x such that $P(X = x) = f_X(x) > 0$ $(x \in \mathcal{X})$, the **conditional pmf of Y given that $X = x$** is the function of y denoted by $f(y \mid x)$ and defined by

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

  (b) For any y such that $P(Y = y) = f_Y(y) > 0$ $(y \in \mathcal{Y})$, the **conditional pmf of X given that $Y = y$** is the function of x denoted by $f(x \mid y)$ and defined by

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

Probabilities

- Once we have the conditional pmf, we can find probabilities as expected.
  For $A \subset \mathbb{R}^2$,
  $$P(X \in A \mid Y = y) = \sum_{x \in A} P(X = x \mid Y = y) =$$
  (just flip for $y \mid x$)
- We can also show that the conditional pmf is indeed a valid pmf.
  Proof, need to show:
  1. $f(x \mid y) \geq 0$ for all x.
  2. $\sum_x f(x \mid y) = 1$.

Examples

1. Interpreting distributions:
   - Let $X =$ GPA and $Y =$ study hours per day.
     If we are given the joint pmf $f(x, y) = P(X = x, Y = y) \longrightarrow$
     then we can find the following:
     i) $f_X(x) =$ $\longrightarrow$ Probability student has
     ii) $f_Y(y) =$ $\longrightarrow$ Probability student has
     iii) $f(x \mid y) =$ $\longrightarrow$ Probability student has
     iv) $f(y \mid x) =$ $\longrightarrow$ Probability student has

2. Define the joint pmf of (X, Y) by:

   $f(0,10) = f(0,20) = 2/18, \quad f(1,10) = f(1,30) = 3/18,$

   $f(1,20) = 4/18, \quad$ and $\quad f(2,30) = 4/18.$

   (a) Compute the conditional pmf of Y given X for each of the possible values of X.

   (b) Find $(X = 2, Y > 20)$

   _____ event

   (c) Find $P(X < 1)$

   _____ event

   (d) Find $P(Y > 10 \mid X = 0)$

   _____ event

3. In a previous example, we had the joint pmf

$$f(x,y) = \frac{x+y}{21} \quad \text{for } x = 1,2,3 \text{ and } y = 1,2.$$

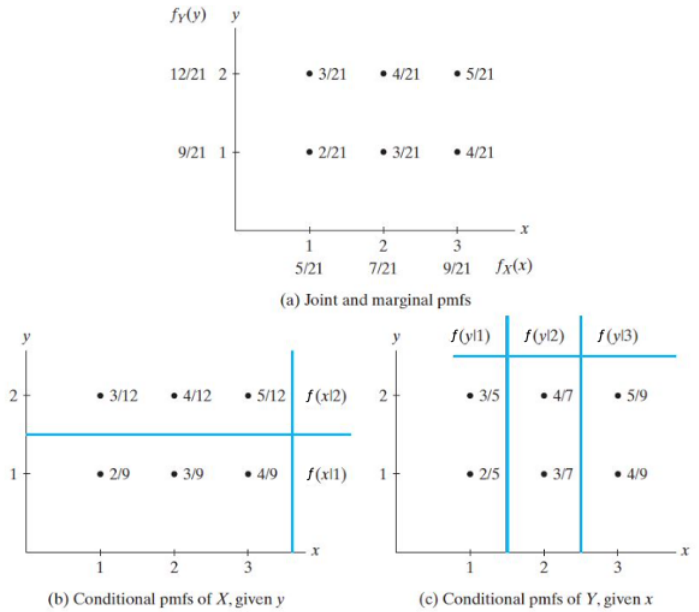
And we found the marginal distributions:

$$f_X(x) = \frac{2x+3}{21} \quad \text{for } x = 1,2,3$$

$$f_Y(y) = \frac{3y+6}{21} = \frac{y+2}{7} \quad \text{for } y = 1,2$$

Find $f(x \mid y)$ and $f(y \mid x)$.

Plots of ranges with corresponding probabilities for all distributions:



Conditional random variable

Understanding conditional random variables

- $Y \mid X = x$ is a random variable about Y having the conditional pmf of $f(y \mid x)$.

  The conditional random variables $Y \mid X = 0$ and $Y \mid X = 1$ have different pmfs.
- **The conditional pmf $f(y \mid x)$ is determined by _____ and thus _____ behaves like a parameter** (e.g. Geometric(p)),

Relationship between joint pmf and conditional pmfs

- The following theorem contains the relationship between the joint pmf of X and Y and the two conditional pmfs $f(y \mid x)$ and $f(x \mid y)$.
- Theorem: For bivariate random vector (X, Y) with joint pmf $f(x, y)$ and x and y such that $f_X(x) > 0$ and $f_Y(y) > 0$,

  $f(x, y) = f_Y(y) \cdot f(x \mid y) = f_X(x) \cdot f(y \mid x)$

Continuous conditional distributions

Conditional pdf

- If X and Y are continuous random variables, then $P(X = x) =$, for every value of x
  $\Longrightarrow$ Can't use $\dfrac{f(x, y)}{P(X = x)}$ because it is undefined.

  To define a conditional probability distribution for Y given $X = x$ when X and Y are both continuous is analogous to the discrete case with pdfs replacing pmfs.

- Definition: Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$.

  (a) Given x such that $f_X(x) > 0$, $\qquad f(y \mid x) = \dfrac{f(x, y)}{f_X(x)}$

  (b) Given y such that $f_Y(y) > 0$, $\qquad f(x \mid y) = \dfrac{f(x, y)}{f_Y(y)}$

Example

- In a previous example, we had the joint pdf

  $f(x, y) = 1/2 \quad$ for $0 \le x \le y \le 2$.

  And we found the marginal distributions:

  $f_X(x) = (2 - x)/2 \quad$ for $0 \le x \le 2 \qquad$ and $\qquad f_Y(y) = y/2 \quad$ for $0 \le y \le 2$

  (a) For $0 \le x < 2$, find the conditional pdf $f(y \mid x)$.

  - NOTE: The range of $Y \mid X = x$ often depends on x. To help, you should draw the range of X and Y just like when finding joint probabilities.
  - For $0 \le x < 2$,

    $$f(y \mid x) =$$

    Conditioned on $X = x$, we see that $Y \mid X = x \sim$

  (b) Find the distribution of $Y \mid X = 1$

  (we have a specific "parameter" value now).

  (c) For $0 < y \le 2$, find the conditional pdf $f(x \mid y)$.

  - For $0 < y \le 2$,

    $$f(x \mid y) =$$

  - Given $Y = y$, we see that $X \mid Y = y \sim$

  (d) Find the distribution of $X \mid Y = 1.5$

  (e) Find the conditional probability that $X \le 1/2 \mid Y = 1.5$.

Expected value of a conditional random variable

Conditional expectations and when to use which density

- In addition to their usefulness for calculating probabilities, the conditional pmfs and pdfs can also be used to calculate expected values.

  Just remember that $f(y \mid x)$ as a function of y is a pmf or pdf; so use it in the same way that we have previously used unconditional pmfs or pdfs.

- Suppose, we have $f(x, y)$, $f_X(x)$, $f_Y(y)$, $f(y \mid x)$ and $f(x \mid y)$. What density function should we use to compute the following?

  1. $E(X) =$

  2. $E(Y^2) =$

  3. $E(Y - Y) =$

  4. $E(X^2\, Y) =$

  5. $E(Y \mid X = 2) =$

  6. $E(Y^2 \mid X = 3) =$

  7. $E(X \mid X = 3) =$

  8. $E(X + Y^2 \mid Y = 3) =$

  9. $E(XY \mid X = 3) =$

Conditional expected values

- Definition: Let $g(Y)$ be a function of Y, then the **conditional expected value of $g(Y)$ given that $X = x$** is denoted by $E[g(Y) \mid X = x]$ and is given by

$$E[g(Y) \mid x] = \sum g(y) f(y \mid x) \qquad \text{and} \qquad E[g(Y) \mid x] = \int_{-\infty}^{\infty} g(y) f(y \mid x)\, \mathrm{d}y$$

  in the discrete and continuous cases, respectively.

- Conditional mean and variance definitions (assuming X and Y are discrete):

  (i) If $g(Y) = Y$, then the **conditional mean of Y given $X = x$** is

  (ii) If $g(Y) = (Y - \mu_{Y|X})^2$, then the **conditional variance of Y given $X = x$** is

Examples

1. In a previous example, we had the joint pmf

   $f(x, y) = \dfrac{x + y}{21}$ for $x = 1, 2, 3$ and $y = 1, 2$.

   And we found the conditional distribution:

   $f(x \mid y) = \dfrac{x + y}{3y + 6}$ for $x = 1, 2, 3$ when $y = 1, 2$.

   (a) Find $\mu_{X|1}$.

   (b) Find $\sigma^2_{X|1}$.

2. For $0 < x \leq 1$, the conditional pdf of $Y \mid X = x$ is $f(y \mid x) = \dfrac{2y}{x^2}$ $0 \leq y \leq x$.

   Note: For this example, the range of _____ depends on _____. So the density as well as the range change when _____ is given.

   (a) Find $E(Y \mid X = x)$.

(b) Find the conditional variance $V(Y \mid X = 0.5)$.

Understanding conditional expectation

- $E(X)$, $E(Y)$, $E(XY)$ are _____ → Center is _____.
- How about $E(Y \mid X = x)$?

  Let's compare the following two conditional expectations.

$$E(Y \mid X = 1/2) = \int_{-\infty}^{\infty} y \qquad \mathrm{d}y$$

$$E(Y \mid X = 1) = \int_{-\infty}^{\infty} y \qquad \mathrm{d}y$$

- The conditional expectation depends only on the _____ which is determined by the value of _____. Consequently, the conditional expected value, $E(Y \mid X = x)$, is determined by the value of _____.

  In other words, as _____ changes, $E(Y \mid X = x)$ changes. Thus, $E(Y \mid X = x)$ is a function of _____.

- What if x is not specified like $E(Y \mid X)$? Then $E(Y \mid X)$ is a function of random variable of X, and thus it is a _____.

  When x is not specified, replace x by X. Then $E(Y \mid X) =$

- **Why is conditional expectation important?**
  - **Regression Analysis**. The main purpose of regression analysis is to identify _____, which explains the mean behavior of Y given X.
  - In regression analysis, we usually assume that Y and X have a **linear relationship**, that is $E(Y \mid X) = \beta_0 + \beta_1 X$.
  - We will study this more later.

Lecture 16 – Independence and the Correlation Coefficient

MATH 321: Mathematical Statistics

Lecture 16: Independence and the Correlation Coefficient

Chapter 4: Bivariate Distributions (4.1, 4.2, and 4.4)

Independence for random variables

Definition

- Events A and B are independent if and only if

- The definition of independence for two discrete random variables relies on this multiplication rule.

  Applying the idea of independence between two events to random variables, we say that X and Y are independent random variable if and only if the events _____ and _____ are independent for all $x \in R$ and all $y \in R$.

- Definition: Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called **independent random variables** if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

- If X and Y are not independent, they are said to be _____.

Checking independence

- In general, if given $f(x, y)$ and checking to see if X and Y are independent (in notation: $X \perp\!\!\!\perp Y$), we have to find $f_X(x)$ and $f_Y(y)$ and then multiply and check.
- Example: Define the joint pmf of (X, Y) by

y \ x	0	1
0	1/4	1/4
1	1/4	0
2	0	1/4

(a) Are X and Y independent?

(b) Are the events $\{X = 0\}$ and $\{Y = 0\}$ independent of each other?

- Interpreting independent random variables.
  - We just saw that the random variables can be _____ even when specific _____ are _____.

    This goes back to the definition of independence. In order for the entire random variables to be independent $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ _____ pairs (x, y).
  - When X and Y are independent, observing $Y = y$ does not alter the probability model for X. Similarly, observing $X = x$ does not alter the probability model for Y.

    Therefore, learning that $Y = y$ provides no information about X and learning that $X = x$ provides no information about Y.
- More examples: Determine if X and Y independent in each of the following scenarios.

  1.
  $f(x, y) = 1/2$ for $0 \le x \le y \le 2$

  $f_X(x) = (2 - x)/2$ for $0 \le x \le 2$

  $f_Y(y) = y/2$ for $0 \le y \le 2$

  2.
  $f(x, y) = 4xy$ for $0 \le x \le 1$, $0 \le y \le 1$

  $f_X(x) = 2x$ for $0 \le x \le 1$

  $f_Y(y) = 2y$ for $0 \le y \le 1$

- What causes difference between the above examples?

- We just saw that the range of (X, Y) plays a crucial role in determining when random variables are independent. We can also look at the functional form of the joint pmf / pdf when checking independence.

- **Independence theorem**: X and Y are independent random variables if and only if

$$f(x,y) = g(x) \cdot h(y), \qquad a \le x \le b,\ c \le y \le d$$

where $g(x)$ is a nonnegative function of x alone and $h(y)$ is a nonnegative function of y alone.

Note that $g(x)$ and $h(y)$ do not themselves need to be density functions.

- Checking independence by inspection.
  - Let $f(x,y) = 2x$ for $0 \le x \le 1$, $0 \le y \le 1$.
  - Process:
    1. Is the range rectangular?
    2. Find any $g(x)$ and $h(y)$ such that $f(x,y) = g(x) \cdot h(y)$

    ** Conclusions:

Range	Range	Range
Separable	Separable	Separable

Rectangular range is a _____, but not _____ condition for independence.

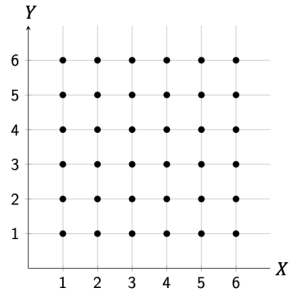
- More examples: Determine if X and Y independent in each of the following scenarios.

3. If the joint pdf is e^{xy} with a rectangle range.

4. If the joint pdf is e^{x+y} with a rectangle range.

5. If the joint pdf is $\log(x+y)$ with a rectangle range.

6. Let (X,Y) be the numbers on die 1 and die 2, respectively, when a pair of fair 6-sided dice are thrown.

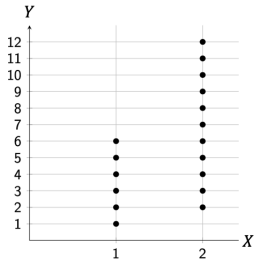


In the discrete case, "rectangular" means that the range of (X, Y) must equal the product set (aka cartesian product) of the individual ranges of X and Y: $\mathcal{X} \times \mathcal{Y} = (\mathcal{X}, \mathcal{Y})$

y \ x	x_1	x_2	x_3
y_1	(x_1, y_1)	(x_2, y_1)	(x_3, y_1)
y_2	(x_1, y_2)	(x_2, y_2)	(x_3, y_3)
y_3	(x_1, y_3)	(x_2, y_2)	(x_3, y_3)

In other words, need _____ $f(x, y) > 0$ for all possible (x, y) combos.

7. A fair coin is tossed. If heads is tossed then one fair 6-sided die is thrown and if tails is tossed two fair 6-sided dice are thrown. Let $X = 1$ for heads and $X = 2$ for tails and let Y be the total number of dots on the dice.



8. Define the joint pmf of (X, Y) by

y \ x	-1	0	1
-1	1/18	1/9	1/6
0	1/9	0	1/6
1	1/6	1/9	1/9

NOTE: If have pmf table (and not equations) and have rectangular range $\Longrightarrow$ Need to check $f(x, y) = f_X(x) \cdot f_Y(y)$ for ALL pairs.

9. Let $f(x, y) = \dfrac{x + y}{21}$ for $x = 1, 2, 3$ and $y = 1, 2$.

Conditional distributions and independence

- Recall if events A and B are independent, then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$. Let's see if this holds for distributions as well.
- Example: An analyst is studying traffic accidents in two adjacent towns. The random variables S and T represents the waiting time between accidents in towns X and Y, respectively. The joint probability function for S and T is given by:

  $f(s,t) = \mathrm{e}^{-(s+t)}$ for $s \geq 0$ and $t \geq 0$.

  (a) Find the marginal distributions $f_S(s)$ and $f_T(t)$.

  (b) Find the conditional distributions $f(s \mid t)$ and $f(t \mid s)$.

- Theorem: If X and Y are independent,

$$f(x \mid y) = f_X(x) \qquad \text{and} \qquad f(y \mid x) = f_Y(y)$$

- Proof:

Using independence

- We can always find the marginal distributions from the joint distribution, but the converse is not always true (so the joint distribution has more information).
- But when X and Y are _____, the joint distribution and marginal distributions contain the equal amount of information about X and Y.

  That means we can also find the _____ distribution from _____ distributions.

- Example: Suppose that $X \sim \text{Exp}\,(\lambda = 1)$ and $Y \sim \text{Uniform}\,(0, 1)$ and $X \perp\!\!\!\perp Y$.

  Find the joint pdf of X and Y.

- Theorem: Let X and Y be independent random variables.

  (a) For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$,

  $$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

  That is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.

  (b) Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then

  $$E[g(X) \cdot h(Y)] = E[g(X)] \cdot E[h(Y)]$$

  Proof for $E(XY)$ for the discrete case:

- When applying this theorem, a bivariate question reduces to a univariate question, making it a lot simpler.

  Example: Let X and Y be independent exponential $(\lambda = 1)$ random variables.

  1. Find $P(X \geq 4, Y < 3)$.

  2. Find $E(X^2Y)$ and $E(X + Y)$.

Covariance

Introduction

- One of main purposes of studying bivariate random vectors is to study the dependence between two random variables.

  Recall that the advantage of using the joint pmf / pdf over the respective marginal distributions is that it usually contains additional information about interaction between the two random variables.

  We can use this joint distribution to check how much two random variables change together.

- More specifically, covariance is how we will study this dependence. Covariance is a special expected value with two very useful applications.

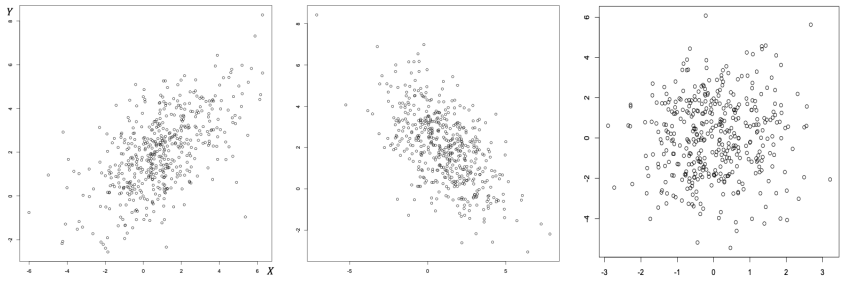
  1. Finding the variance of $X + Y$.

  2. Quantify linear dependence.

- There are two different scenarios that we can calculate covariance for: on a sample of data and on a probability distribution.

  Covariance is easier to conceptualize if we have a sample of data. So we will start with this. But technically we are still in a probability context, so then we will shift to working with distributions (i.e. population information) rather than data points.

Visualizing dependence



- In the first plot above:
  - Large values ($> \mu$) of X mainly correspond with _____ values of Y.

    Small values ($< \mu$) of X mainly correspond with _____ values of Y.
  - In this case, the two random variables are _____ dependent / correlated.

    In statistics, correlated = linear relationship.

- In the second plot above:
  - Large values of X mainly correspond with _____ values of Y.

    Small values of X mainly correspond with _____ values of Y.
  - In this case, the two random variables are _____ dependent / correlated.

- In the third plot above:
  - Large values of X mainly correspond with _____ values of Y.

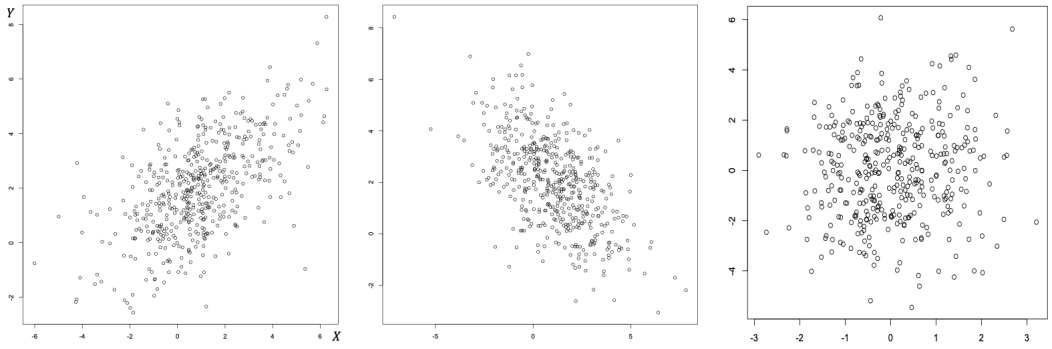
    Small values of X mainly correspond with _____ values of Y.
  - In this case, the two random variables appear to _____ be dependent / correlated.

Quantifying dependence

- So we know how dependence shows up in a scatterplot, so then how to we quantify (measure) it?

  On the plots above, we were looking at where X and Y were in relation to their respective means and then also how these related (interacted) with each other.

  In statistics / modeling, interaction means multiplying terms. So here is the function that we will start with:



- If X and Y are positively dependent (plot 1), then

  $$(X - \mu_X)(Y - \mu_Y)$$

  will be mostly _____.

  For example, height and weight. Most of people taller than the average weigh more than the average. Most of people shorter than the average weigh less than the average.

- If X and Y are negatively dependent (plot 2), then

  $$(X - \mu_X)(Y - \mu_Y)$$

  will be mostly _____.

  For example, stress (e.g. on a mechanical part or system) and time to failure. More stress often results in shorter time to failure and less stress often leads to longer time to failure.

- If X and Y appear to not be dependent (plot 3), then

  $$(X - \mu_X)(Y - \mu_Y)$$

  the positives and negatives will almost _____ (spread evenly in all "quadrants").

- So our function "results" match our visual explanation of dependence, but also want to emphasize "mostly negative and mostly positive". Some values could have the different sign from the most of others.

  We want to measure **overall tendency**. So we define the _____ of $(X - \mu_X)(Y - \mu_Y)$ as a measure of **linear dependence** between X and Y.

Definition, theorems and properties of covariance

- Definition: The **covariance** of X and Y is the number defined by

$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

- Notes:
  - Covariance is a measure of how much two random variables change together.
  - If the $\text{Cov}(X,Y) > 0$, then X and Y are _____ correlated.
  - If the $\text{Cov}(X,Y) < 0$, then X and Y are _____ correlated.
  - If the $\text{Cov}(X,Y) = 0$, then X and Y are _____, which means there is **no linear dependence** between X and Y.
  - Covariance can measure only the **linear** dependence between two random variables.

    May not pick up on non-linear relationships (i.e. curvature).
- Properties / theorems of covariance:

(i) **Calculation**: If (X,Y) is discrete, then

$$E[(X-\mu_X)(Y-\mu_Y)] =$$

Examples: Calculate $\text{Cov}(X,Y)$ for the following two joint pmfs.

(a)

y \ x	1	2
1	4/8	1/8
2	1/8	2/8

x	y	f(x,y)	(x - mu_x)(y - mu_y)	g(x,y)f(x,y)
1	1	0.5	0.141	0.0703
2	1	0.2	-0.234	-0.0469
1	2	0.125	-0.234	-0.0293
2	2	0.25	0.391	0.0977
			covariance =	0.0918
mu_x =	1.375			
mu_y =	1.375			

(b)

y \ x	1	2
1	1/16	6/16
2	7/16	2/16

(ii) Recall how with variance we had a definition and an alternate form that is easier to calculate by hand: $V(X) = E[(X - \mu_X)^2] = E(X^2) - [E(X)]^2$. We have a similar theorem for covariance:

Alternate calculation for covariance: For any random variables X and Y,

$$\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$$

Proof:

Example: Back to the investor with two asset random variables X and Y. We have previously calculated each of these pieces and thus:

The _____ covariance means that as one investment performs above average, the other tends to perform _____ average.

(iii) **Variance is a special case of covariance**:

$$V(X) = \text{Cov}(X, X)$$

Proof:

(iv) Order in covariance does not matter (i.e. **symmetric**).

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

Proof:

(v) Covariance of a random variable and a constant is zero.

$$\text{If } c \text{ is a constant, then } \text{Cov}(X, c) = 0$$

Proof:

(vi) Can factor out coefficients in covariance.

$$\text{Cov}(aX, bY) = ab \cdot \text{Cov}(X, Y)$$

Proof:

(vii) Can factor out coefficients, but added constants disappear.

$$\text{Cov}(aX + c, bY + d) = ab \cdot \text{Cov}(X, Y)$$

Like variance, the location shift does not influence covariance, which means it does not impact the _____ between X and Y.

Example: Suppose investment X is now performing 1.2 times better than previously and they added 5 to investment Y. Find the new covariance of the two investments.

(viii) **Distributive property** of covariance.

$$\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$$

Proof:

Independence and covariance

- Theorem: If X and Y are independent random variables, then

$$\text{Cov}(X, Y) = 0$$

- Proof:

- Independent vs uncorrelated:
  - If X and Y are independent, there does not exist **any type of dependence** between X and Y.
  - If X and Y are uncorrelated (i.e. $\text{Cov}(X, Y) = 0$), there is no _____ dependence between X and Y.

    But, there may be some other type of relationship.
  - Thus, independent random variables _____ be uncorrelated, but uncorrelated random variables _____ be independent.

    Summary:

- Example: Let $f(x, y) = 1/3$ for $(x, y) = (0, 1), (1, 0), (2, 1)$.

  (a) Are X and Y independent?

  (b) Are X and Y uncorrelated?

Interpreting covariance

- We have already shown that we can determine the **direction** of the relationship based on the sign of the covariance, this is a useful interpretation.

  However, we cannot use covariance to measure the **strength** of the relationship.

  This is because its value depends on the scale of measurement.

- Example demonstrating this: Suppose (X, Y) = (income, savings) in dollars and $(X', Y') = (100X, 100Y)$ = (income, savings) in cents. Further, let $\text{Cov}(X, Y) = 3$.

  $\text{Cov}(X', Y') =$

  We _____ say that the linear dependence between income and savings in cents is stronger than that in dollars because _____, all we know is _____.

- Because of this, covariance cannot be used as an absolute measure of linear dependence (i.e. a single measure that tells us everything, direction and strength).

  So how can we improve it?

Correlation

Definition

- Motivation: The problem can be eliminated by _____ the covariance value.
- Definition: As an absolute measure of dependence, the **correlation coefficient** of X and Y is the number defined by

$$\rho_{XY} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Continuing previous example:

$$\text{Corr}(X', Y') = \frac{\text{Cov}(X', Y')}{\sqrt{V(X')V(Y')}} =$$

  Thus, the correlation is unit-free and _____ by change in scale or location.

- Back to investor example: It can be shown that $V(X) = 40$ and $V(Y) = 25$ and we previously found $\text{Cov}(X, Y) = -13$. Thus

Properties of correlation

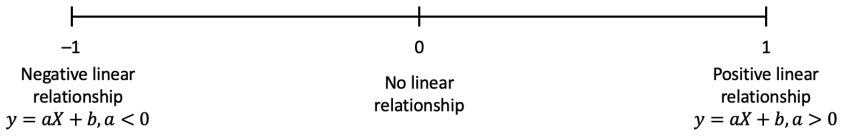
- We are going derive these!
- Covariance measures linear dependence, so lets see what happens to correlation when there is a perfect linear relationship, i.e. $Y = aX + b$.
  - Sidenote: This is called a deterministic (or functional) relationship / model as opposed to stochastic (or probabilistic or statistical), which is when there is some randomness involved.

  If $Y = aX + b$,

  Thus when X and Y are linearly related, the correlation coefficient is _____ when the slope of the straight line is positive and _____ when the slope is negative.
- To see what might happen when X and Y are not linearly related, we can look at the extreme case when X and Y are independent and have no systematic dependence.

  Obviously $\rho_{XY} =$ whenever $\text{Cov}(X, Y) =$. Keep in mind that variables can be uncorrelated without being independent.
- Putting these two pieces together, it can be shown that for any random variables X and Y,

$$\leq \rho_{XY} \leq$$



  - Possible values of ρ_{XY} lie on a continuum between -1 and 1.
  - Values of ρ_{XY} close to ± 1 are interpreted as an indication of a high level of linear association between X and Y.
  - Values of ρ_{XY} near 0 are interpreted as implying little or no linear relationship between X and Y.

- Formally, we can state these properties in the following theorem:

  Theorem: For any random variable X and Y,

  (i) $-1 \leq \rho_{XY} \leq 1$

  The sign represents the direction of linear dependence between X and Y.

  $|\rho_{XY}|$ represents the magnitude of dependence.

  (ii) $\rho_{XY} = 1$ if and only if there exist numbers $a > 0$ and b such that $P(Y = aX + b) = 1$.

  X and Y have perfect positive correlation. As X increases, Y _____.

  (iii) $\rho_{XY} = -1$ if and only if there exist numbers $a < 0$ and b such that $P(Y = aX + b) = 1$.

  X and Y have perfect negative correlation. As X increases, Y _____.

  (iv) When $\rho_{XY} = 0$, X and Y are _____.

Examples

1. Is the dependence between X_1 and X_2 stronger than the dependence between Y_1 and Y_2?

   (a) $\text{Cov}(Y_1, Y_2) = 0.4$, $\text{Cov}(X_1, X_2) = 0.6$

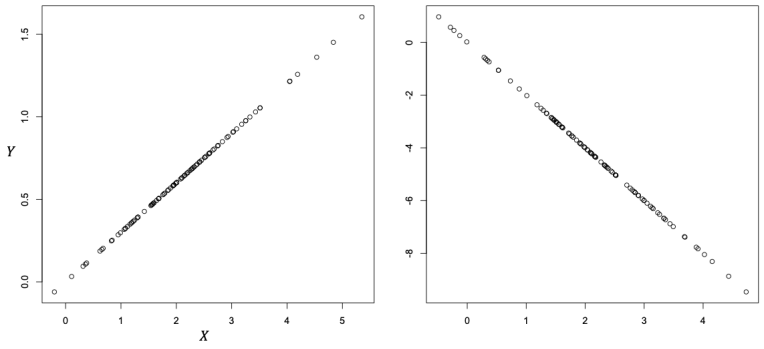
   (b) $\text{Cov}(Y_1, Y_2) = 0.4$, $\text{Cov}(X_1, X_2) = -0.6$

   (c) $\text{Corr}(Y_1, Y_2) = 0.4$, $\text{Corr}(X_1, X_2) = 0.6$

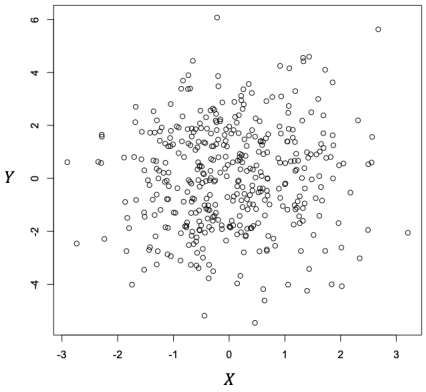
   (d) $\text{Corr}(Y_1, Y_2) = 0.4$, $\text{Corr}(X_1, X_2) = -0.6$

2. Find ρ_{XY}:

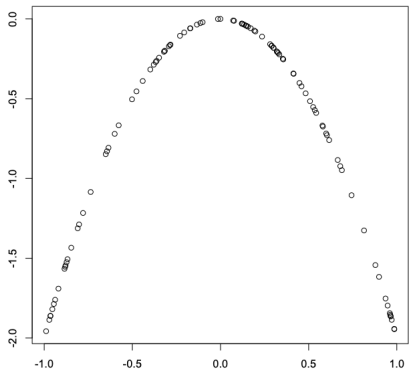
   (a) $\text{Corr}(X, Y) =$ (b) $\text{Corr}(X, Y) =$



(c) $\text{Corr}(X, Y) =$



(d) $\text{Corr}(X, Y) =$



Variance of $X + Y$

Derivation

- We stated at the start that $V(X + Y) \neq V(X) + V(Y)$. Now that we understand covariance, we can see why!
- Let X and Y be two random variables:

- Theorem: **Variance of a sum of two random variables**

$$V(X + Y) =$$

- Examples:
  1. Back to the investor with two assets X and Y, find the $V(X + Y)$:

2. Let $f(x,y) = x + \frac{3}{2}y^2 \quad$ for $0 \le x \le 1, 0 \le y \le 1$. Find the variance of $X + Y$.

Variance of $X + Y$ when independent

- Derivation: If X and Y are independent, then

- Theorem: **Variance of a sum of two independent random variables**

  If $X \perp\!\!\!\perp Y$, then $V(X+Y) =$

  $\Longrightarrow$ Can only JUST add variances when independent.
- Example: Let $X \sim \text{Exp}\,(\lambda = 1)$ and $Y \sim \text{Exp}\,(\lambda = 3)$. If $X \perp\!\!\!\perp Y$, find $V(X+Y)$.

Lecture 17 – Several Random Variables

MATH 321: Mathematical Statistics

Lecture 17: Several Random Variables

Chapter 5: Distributions of Functions of Random Variables (5.3 and 5.4)

Multivariate distributions

Introduction

- Now we are extending bivariate distributions to multivariate distributions.

  The good news is that the jump from 2 random variables to 3 or 4 or n random variables is much easier than the jump from 1 to 2.

- The concepts such as marginal and conditional distributions generalize from the bivariate to the multivariate setting.

  We will start by giving these generalizations, then demonstrating via examples.

- A note on notation: Boldface letters are used to denote multiple variates. Write $\mathbf{X}$ to denote $X_1, \ldots, X_n$ and $\mathbf{x}$ to denote the sample $x_1, \ldots, x_n$.

Definitions and theorems

- Joint distributions and probabilities.
  - The random vector $\mathbf{X} = (X_1, \ldots, X_n)$ has a range that is a subset of $\mathbb{R}^n$ (n dimensions).
  - If $\mathbf{X} = (X_1, \ldots, X_n)$ a discrete random vector (the range is countable), then the **joint pmf** of $\mathbf{X}$ is the function defined by

    $$f(\mathbf{x}) = f(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n) \text{ for each } (x_1, \ldots, x_n) \in \mathbb{R}^n$$

  - Finding probabilities: Then, for any $A \subset \mathbb{R}^n$,

    $$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x})$$

  - If $\mathbf{X} = (X_1, \ldots, X_n)$ a continuous random vector, then the **joint pdf** of $\mathbf{X}$ is the function $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ that satisfies

    $$P(\mathbf{X} \in A) = \int \cdots \int_A f(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int \cdots \int_A f(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

- Expected values.
  - Let $g(\mathbf{x})$ be a real-valued function defined on the range of $\mathbf{X}$. The **expected value** of $g(\mathbf{X})$ is

    Discrete: $E[g(\mathbf{X})] = \sum_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x})f(\mathbf{x})$

    Continuous: $E[g(\mathbf{X})] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(\mathbf{x})f(\mathbf{x})\mathrm{d}x_1\cdots\mathrm{d}x_n$

- These and other definitions are analogous to the bivariate definitions, except now the sums or integrals are over the appropriate subset of $\mathbb{R}^n$ rather than $\mathbb{R}^2$.
- Marginal distributions.
  - The **marginal pdf or pmf** of any subset of the coordinates of $(X_1,\ldots,X_n)$ can be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates.
  - Thus for example, the marginal distribution of $(X_1,\ldots,X_k)$ the first k coordinates of $(X_1,\ldots,X_n)$ is given by the pdf or pmf:

    Simple case: $n=5$, $k=2$

  - Even though these marginal distributions can themselves be multivariate, they are still called marginal because they have less variables than the joint distribution.
- Conditional distributions.
  - The **conditional pmf or pdf** of a subset of the coordinates of $(X_1,\ldots,X_n)$ given the value of the remaining coordinates is obtained by dividing the joint pdf or pmf by the marginal pdf or pmf of the remaining coordinates.

    $$f(x_{k+1},\ldots,x_n \mid x_1,\ldots,x_k) =$$

Example

1. Let $n=4$ and

   $$f(x_1,x_2,x_3,x_4) = \frac{3}{4}(x_1^2+x_2^2+x_3^2+x_4^2), \qquad 0<x_i<1, \quad i=1,2,3,4$$

   (a) Verify $f(x_1,x_2,x_3,x_4)$ is a valid pdf.

   (b) NOTE: Probabilities ALWAYS need ALL integrals.

   Find $P(X_1<1/2, X_2<3/4, X_4>1/2)$.

(c) Find the marginal pdf of (X_2, X_3).

Now any probability or expected value that involves only X_1 and X_2 can be computed using this marginal pdf.

(d) Find $E(X_2X_3)$.

(e) Find the conditional pdf $f(x_1, x_4 \mid x_2, x_3)$.

(f) Find $P(X_1 > 3/4, X_4 < 1/2 \mid X_2 = 1/3, X_3 = 2/3)$.

Independence

- Generally, we will be working with independent random variables. This is a very common assumption in probability and statistics that each observation from a random experiment is independent.

  Lets see how this impacts the definitions and theorems we just presented.

- Joint distributions:
  - Definition: Let random variables $X_1, \ldots, X_n$ have joint pdf (or pmf) $f(x_1, \ldots, x_n)$ and let $f_{X_i}(x_i)$ be the marginal pdf (or pmf) of X_i. Then $X_1, \ldots, X_n$ are **mutually independent random variables** if, for every $(x_1, \ldots, x_n)$, the joint pdf (or pmf) can be written as

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

  - Notes
    - We keep the subscripts on X_i because the marginal distributions can be different.
    - Mutual independence (strongest form) $\Longrightarrow$ Pairwise independence AND all possible subsets are independent
  - Example: Let X_1, X_2, X_3 be (mutually) independent exponential random variables with parameters $\lambda_1 = 3, \lambda_2 = 5, \lambda_3 = 1$, respectively.
    (a) Find the joint pdf of X_1, X_2, X_3.

    (b) Find $P(1 < X_1 < 4, X_2 > 3, X_3 \leq 2)$

- If $X_1, \ldots, X_n$ are mutually independent, then knowledge about the values of some coordinates gives us no information about the values of other coordinates. Mutually independent random variables have many nice properties.
- Conditional distributions.
  - If $X_1, \ldots, X_n$ are mutually independent, we can show that the conditional distribution of any subset of the coordinates, given the values of the rest of the coordinates, is the same as the marginal distribution of the subset.
  - Example: Let $X_1, \ldots, X_4$ be mutually independent random variables. Show $f(x_3, x_4 \mid x_1, x_2) = f(x_3)f(x_4)$.

- Expected value.
  - Let $X_1, \ldots, X_n$ be mutually independent random variables. Let $g_1, \ldots, g_n$ be real-valued functions such that $g_i(x)$ is a function only of x_i, $i = 1, \ldots, n$. Then

    $$E[g_1(X_1) \cdots g_n(X_n)] =$$

  - Example: Let X_1, X_2, X_3 be independent exponential random variables with parameters $\lambda_1 = 3, \lambda_2 = 5, \lambda_3 = 1$, respectively.

    Find $E[(2X_1)(X_2 + 1)(X_3)]$.

Linear functions of random variables

Introduction and definition

- Definition: A **linear function (combination) of random variables** consists of n random variables $X_1, \ldots, X_n$ and n coefficient $a_1, \ldots, a_n$

  $$a_1X_1 + a_2X_2 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i$$

- Why is the linear function of random variables important?

  Most of estimators of parameters are linear functions of random variables.

  1. The estimator of the population mean $\mu = E(X)$ is

  2. The estimator of the population variance $\sigma^2 = V(X)$ is

- In order to study the properties of estimators, it is necessary to know how to compute the expected value and variance of linear functions of random variables.

  We will learn how to find their distributions soon (start of MATH 321 topics).

Expected value and variance of linear functions of random variables

- We will start by demonstrating these in the simplest cases (small n and no coefficients, i.e. all $a_i = 1$), then generalize.
- Recall for when $n = 2$.

  $E(X + Y) =$

  $V(X + Y) =$

  - Now generalizing with constants a and b.

    $E(aX + bY) =$

    $V(aX + bY) =$

- Now for $n = 3$.

  $E(X + Y + Z) =$

  $V(X + Y + Z) =$

- The general pattern should be easy to see. Now we can extend this to n (still with no coefficients).

  Theorem: **Mean and variance of $\boldsymbol{X_1 + \cdots + X_n}$**

  $E\left(\sum_{i=1}^{n} X_i\right) =$

  $V\left(\sum_{i=1}^{n} X_i\right) =$

- Finally, in general we have the following theorem:

  (i) **Expected value of a linear function of random variables**

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$$

  (ii) **Variance of a linear function of random variables**

$$V[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = \sum_{i=1}^{n} a_i^2 V(X_i) + 2\sum_{i<j} a_i a_j \text{Cov}(X_i, X_j)$$

  If $X_1, \ldots, X_n$ are mutually independent (or uncorrelated),

$$V[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = \sum_{i=1}^{n} a_i^2 V(X_i)$$

- Easy way to understand and remember the variance of linear functions of random variables.
  - Lets look at the result when we square simple linear functions and expand:

$(X_1 + X_2)^2 = (X_1 + X_2)(X_1 + X_2) =$

$(X_1 + X_2 + X_3)^2 =$

$(a_1X_1 + a_2X_2 - a_3X_3)^2 =$

  - Why is this useful? Replace all quadratic (squared) terms with variances and cross (interaction) terms with covariances.

$V(X_1 + X_2 + X_3) =$

$V(a_1X_1 + a_2X_2 - a_3X_3) =$

– If we have more than 3 random variables, this still works!

$$(a_1X_1 + a_2X_2 + \cdots + a_nX_n)^2 = \sum_{i=1}^{n} a_i^2X_i^2 + 2\sum_{i<j} a_ia_jX_iX_j$$

$$V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = \sum_{i=1}^{n} a_i^2V(X_i) + 2\sum_{i<j} a_ia_j\text{Cov}(X_i, X_j)$$

– Example: Let X_1, X_2 and X_3 be random variables, where $V(X_1) = 1$, $V(X_2) = 3$, $V(X_3) = 5$, $\text{Cov}(X_1, X_2) = -0.4$, $\text{Cov}(X_1, X_3) = 0.5$, $\text{Cov}(X_2, X_3) = 2$.

Find $V(3X_1 - X_2 + 2X_3)$.

Mgf of sums of independent random variables

Introduction

- In some applications, it is sufficient to know the mean and variance of a linear combination of random variables, say, Y. This is what we learned last section (5.3).

  However, it is often helpful to know exactly how Y is distributed (pmf / pdf / mgf). The easiest way to do this is via moment generating functions.

- Recall the definition: The moment generating function (mgf) of random variable X (or the distribution of X), denoted $M_X(t)$, was defined by the following in the univariate case

$$M_X(t) = \underbrace{E(\mathrm{e}^{tx})}_{\text{In general}} \rightarrow \underbrace{\sum_x \mathrm{e}^{tx} f(x)}_{\text{Discrete}} \quad \underbrace{\int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\,\mathrm{d}x}_{\text{Continuous}}$$

  Additionally, the mgf of a random variable uniquely determines its distribution (i.e. no two random variables with "different" distributions share the same pdf).

Mgf of sums of independent random variables

- Theorem: Let X and Y be independent random variables with mgfs $M_X(t)$ and $M_Y(t)$. Then the mgf of the random variable $S = X + Y$ is given by

$$M_S(t) = M_X(t) \cdot M_Y(t)$$

- Proof:

- Example 1: $X \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y \sim \text{Normal}(\mu_2, \sigma_2^2)$ and $X \perp\!\!\!\perp Y$. Find the distribution of $S = X + Y$.

- Note: Whenever finding the distribution of a sum random variables (e.g. $X + Y$), always start with mgfs. It is usually to use the mgf rather than doing transformations using the pmf / pdf.

- Now we can extend the previous theorem to a sum of n random variables:

  Theorem: Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $Y = X_1 + \cdots + X_n$.

  $M_Y(t) =$

  In particular, if $X_1, \ldots, X_n$ all have the same distribution with mgf $M_X(t)$, then

$$M_Y(t) =$$

- More examples:

  2. Suppose X_1 and X_2 are independent Poisson random variables with means λ_1 and λ_2, respectively. Find the distribution of $X_1 + X_2$.

     Recall that the mgf of a Poisson(λ) distribution is $M_X(t) = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}$.

3. Suppose X_1 and X_2 are *iid* Bernoulli random variables ($M_X(t) = (1-p) + pe^t$). Find the distribution of $X_1 + X_2$.

4. The same logic can be used for *iid* geometric distributions ($M_X(t) = \frac{pe^t}{1-qe^t}$) and *iid* exponential distributions ($M_X(t) = \frac{\beta}{\beta - t}$).

5. Suppose $X_1, \ldots, X_n$ are mutually independent random variables, and $X_i \sim \text{Gamma}(\alpha_i, \beta)$. Find the distribution of $Y = X_1 + \cdots + X_n$.

   Recall that the mgf of a Gamma(α, β) distribution is $M_X(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$.

- In general, we can say extend the previous examples and state the following results, which all match our previous explanations / interpretations of the relationships between these distributions:
  - Poisson:

    If $X_1, \ldots, X_n \overset{\perp\!\!\!\perp}{\sim} \text{Poisson}(\lambda_i)$, then $Y = X_1 + \cdots + X_n \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_n)$.

  - Bernoulli:

    $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, then $Y = X_1 + \cdots + X_n \sim \text{Binomial}(n, p)$.

  - Geometric:

    If $X_1, \ldots, X_r \overset{iid}{\sim} \text{Geometric}(p)$, then $Y = X_1 + \cdots + X_r \sim \text{Negative Binomial}(r, p)$.

  - Exponential:

    $X_1, \ldots, X_\alpha \overset{iid}{\sim} \text{Exponential}(\lambda)$, then $Y = X_1 + \cdots + X_\alpha \sim \text{Gamma}(\alpha, \beta)$.

  - Gamma:

    $X_1, \ldots, X_n \overset{\perp\!\!\!\perp}{\sim} \text{Gamma}(\alpha_i, \beta)$, then $Y = X_1 + \cdots + X_n \sim \text{Gamma}(\alpha_1 + \cdots + \alpha_n, \beta)$.

- Extension of previous theorem to sums of linear combinations of random variables:

  Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Let $Y = (a_1X_1+b_1)+\cdots+(a_nX_n+b_n)$. Then the mgf of Y is

$$M_Y(t) = \left(e^{t\sum b_i}\right) M_{X_1}(a_1t) \cdots M_{X_n}(a_nt)$$

  Proof:

- Example: Let $X_1 \sim \text{Normal}\,(\mu = 5, \sigma^2 = 4)$, $X_2 \sim \text{Normal}\,(\mu = 3, \sigma^2 = 8)$, and $X_1 \perp\!\!\!\perp X_2$. Find the distribution of $Y = 3X_1 + 2X_2 - 1$.

  Recall $M_X(t) = \exp\left[\mu t + \frac{\sigma^2 t^2}{2}\right]$.

- Important result from this:

  Theorem: Let $X_1, \ldots, X_n$ be mutually independent random variables with $X_i \sim \text{Normal}\,(\mu_i, \sigma_i^2)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Then,

$$Y = \sum_{i=1}^{n}(a_iX_i + b_i) \sim \text{Normal}\left(\mu = \sum_{i=1}^{n}(a_i\mu_i + b_i),\ \sigma^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

  $\Longrightarrow$ Sum of normal random variables is _____ normal.

Test 2

Contents

Lecture 1 – Random Samples and Common Statistics

MATH 321: Mathematical Statistics

Lecture 1: Random Samples and Common Statistics

Chapter 5: Distributions of Functions of Random Variables (5.5)

MATH 320 vs MATH 321

Relationship between Probability and Statistics

- We studied **Probability** in MATH 320 and are going to study **Statistical Inference** in MATH 321.

  So what is the difference between them?

- In a probability problem, the properties of the population are assumed known, and we use these to infer properties of the sample.

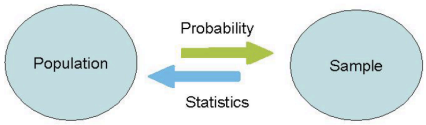


Figure: The reverse actions of Probability and Statistics

- Whereas statistics is concerned with learning (inferring) population properties from sample information (which is the opposite of probability).
- In spite of this difference, statistical inference itself would not be possible without probability because it is based on probability calculations.
- Example:
  - Suppose we know 75% of batteries last longer than 1500 hours. We want to know the chance that in a sample of 30 batteries at least 20 will last more than 1500 hours.
    - What is known (in other words, fixed)?

      This is _____ information

      $\Longrightarrow$ This a _____ question.
    - What is unknown (in other words, variable)?
    - We answer _____ question using the _____.
  - The most important thing when solving probability questions is the **distribution**. If we know this, we can answer any questions in probability.

  - Suppose that in a sample of 30 batteries, only 20 are found to last more than 1500 hours. We want to know if that is enough evidence to conclude that the proportion of all batteries that last more than 1500 hours is less than 75%.
    * What is known (in other words, fixed)?

      This is _____ information

      $\Longrightarrow$ This a _____ question.
    * What is unknown (in other words, variable)?
    * We answer _____ question using the _____.
  - Whether in a probability or statistics context, we are always looking for the **distribution**.

- What we are going to study in MATH 321:
  - Statistics (and their properties)
  - The distributions of statistics
  - Principles we would like statistics to have
  - Statistical inference (confidence intervals and hypothesis tests)
  - Other applications, such as regression

Example process

1. Collect data, $x = \{1510, 1700, 1400, \ldots\}$ 30 observations
2. Transform data to 0s and 1s (if $x > 1500 \to 1$, else 0)
3. Summarize with a statistic (which is a random variable)
4. Find distribution
5. Compute stuff (expected value, variance, probabilities, etc.)

Basic concepts of random samples

Random sample

- Motivation: Often, the data collected in an experiment consist of several observations on a variable of interest.

  In this section, we present a model for data collection that is often used to describe this situation, a model referred to as **random sampling**.

- Definition: The random variables $X_1, \ldots, X_n$ are called a **random sample** of size n from the population $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$.

  In other words, $X_1, \ldots, X_n$ are **independent and identically distributed (*iid*)** random variables with pdf or pmf $f(x)$.

- Lets build up to the model (joint pdf or pmf) that we will be working with from now on starting with what we learned in the multivariate setting:

  0. Joint distribution

  1. Mutually independent random variables:

  2. Identically distributed:

  3. Parametric family:

- In a statistical setting, if we assume that the population we are observing is a member of a specific parametric family, but the true parameter value is unknown, then a random sample from this population has a joint pdf or pmf of the above form with the value of θ unknown.

  By considering different possible values of θ, we can study how a random sample would behave for different populations.

- Example: $X_1, \ldots, X_n$ correspond to the times until failure (measured in years) for n identical circuit boards that are put on test and used until they fail. Find the joint pdf of the random sample $X_1, \ldots, X_n$.

Scenario

- Suppose we collect information on 30 circuit boards, say $x_1, \ldots, x_{30}$.

  When reporting this information, do we ever report the entire set of observations?

- Instead, we report _____.

  In doing so, we are summarizing the information in a sample by determining a few key features of the sample values. This is usually done by computing _____ (functions of the sample).

- These statistics define a form of _____. There are advantages and consequences of this.

  Whether or not statistics are "good enough" is a topic that will be left for grad school, but we will work with the most common ones that have good properties.

Sums of random variables from random samples

Statistics (estimators)

- Definition: Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued functions whose domain includes the sample space of $(X_1, \ldots, X_n)$.

  Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution of Y**.

- Breakdown of definition:
  - Statistic $Y = T(X_1, \ldots, X_n)$ is a function. All functions have a domain (set of inputs) and codomain (set of outputs).

    The domain of Y is the sample space of $X_1, \ldots, X_n$ and the codomain of Y depends on the statistics.
  - Example: Roll 3 die (X_1, X_2, X_3)

    Some examples of statistics: Mean, variance, median, min, max, etc.

    Usually we are interested in one of these at a time, but not always (e.g. $(\bar{X}, S^2)$ is a two-dimensional statistic (a vector)).
  - The statistic $Y = T(X_1, \ldots, X_n)$ is a _____ because it is a function of random variables.

- Miscellaneous notes:
  - Statistics and _____ are exactly the same thing.
  - It's called a sampling distribution because it is derived from a random sample.
  - The definition of a statistic is very broad. The only restriction is that a statistic can not be a function of the _____ (because it is unknown).
  - It is not necessary that $X_1, \ldots, X_n$ be *iid* to define a statistic. Even if they are dependent and/or not identically distributed, $Y = T(X_1, \ldots, X_n)$ is still called a statistic, but is is much more difficult to deal with (of course beyond the scope of this class).

- Ultimately, we want to find the _____, which is usually tractable (able to be found) because of its simple probability structure (*iid*). But we will start with _____ of two common statistics.

Sample mean and variance

- Definition: The **sample mean** is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Definition: The **sample variance** is the statistic defined by

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

  The **sample standard deviation** is the statistic defined by $S = \sqrt{S^2}$.

- Theorem: Let $X_1, \ldots, X_n$ be a random sample of size n from a population (of any distribution) with mean μ and variance $\sigma^2 < \infty$. Then

  (a) $E(\bar{X}) = \mu$ (mean of sample means)

  (b) $V(\bar{X}) = \frac{\sigma^2}{n}$ (variance of sample means)

  (c) $E(S^2) = \sigma^2$ (mean of sample variance)

  Some proofs (a) and (b):

- Notes:
  - It can be shown why we use $n-1$ in the definition of S^2 rather than just n (i.e. it is necessary to be unbiased, which means the statistic's expected value is equal to the parameter it is estimating).
  - The statistic $\bar{X}$ is an unbiased estimator of μ, and S^2 is an unbiased estimator of σ^2. We discuss this more later.

Sampling distribution of $\bar{X}$

- Now we would like to study the sampling distribution of $\bar{X}$.

  Recall whenever finding the distribution of a sum of random variables, we want to use the mgf technique.

- Theorem: Let $X_1, \ldots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is

$$M_{\bar{X}}(t) = [M_X(t/n)]^n$$

  Note this theorem is a combo of previous theorems where we found the mgf of *iid* $X_1 + \cdots + X_n$ and also now with all coefficients $a_i = 1/n$.

- Examples:
  1. Continuing circuit board failure time example: Find the distribution of $\bar{X}$ the mean time until failure (measured in years) for n identical circuit boards that are put on test and used until they fail (recall $M_X(t) = \lambda/(\lambda - t)$).

  2. Suppose $X_i \overset{iid}{\sim}$ Binomial $(p = 0.6, n = 10)$, for $i = 1, \ldots, 5$.

     Find $E(\bar{X})$, $V(\bar{X})$, and the distribution of $\bar{X}$ (recall $M_X(t) = (q + pe^t)^n$).

Sampling from the normal distribution

Introduction

- This section deals with the properties of sample quantities drawn from a normal population – still one of the most widely used statistical models.
- In practice, often
  1. We assume population has a normal distribution (e.g. heights, test scores, errors in measurements, etc.).
  2. Then sample from population and form statistics in order to estimate the mean and variance of the population.
- Thus we wish to know the distribution of these statistics which will allow us to determine accuracy, form confidence intervals, hypothesis testing, etc.

  Sampling from a normal population leads to many useful properties of sample statistics and also to many well-known sampling distribution.

Theorem

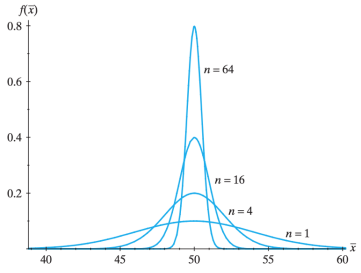
- Let $X_1, \ldots, X_n$ be a random sample of size n from a Normal (μ, σ^2) distribution. Then
  (a) $\bar{X}$ and S^2 are independent random variables.
  (b) $\bar{X}$ has a Normal $(\mu, \dfrac{\sigma^2}{n})$ distribution.
  (c) $\dfrac{(n-1)}{\sigma^2} S^2$ has a chi squared distribution with $n-1$ degrees of freedom (df).
- Notes / proofs:
  (a) $\bar{X} \perp\!\!\!\perp S^2 \rightarrow$ This is just a necessary fact when deriving (c), we will not prove this.
  (b) $\bar{X} \sim \text{Normal}\,(\mu, \dfrac{\sigma^2}{n})$

  We know this from prior theorems (parameters: mean and variance of sample means for normal), distribution (sum of normals = normal).

  Example: Let $X_1, \ldots, X_n$ be a random sample from $N(\mu = 50, \sigma^2 = 16)$.

  See the effect of n in the distributions.

  $P(49 < \bar{X}_{64} < 51) \approx 0.9545$
  $P(49 < X < 51) \approx 0.1974.$



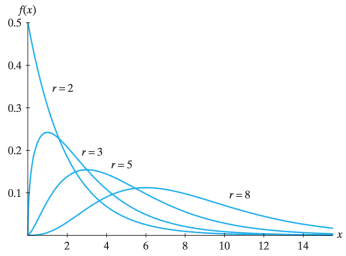
pdfs of means of samples from $N(50, 16)$

(c) $\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2(n-1) \rightarrow$ This is a new distribution, so lets learn about this before motivating the idea behind this part of the theorem.

Chi-square distribution

Definition

- The chi-square distribution is a special case of the gamma distribution that plays an important role in statistics.
- If $X \sim \chi^2$ with r degrees of freedom (often written χ^2_r or $\chi^2(r)$), then
  (i) Pdf:
  $$f(x) = \frac{1}{\Gamma(\frac{r}{2})2^{r/2}}\, x^{\frac{r}{2}-1}\, \mathrm{e}^{-\frac{x}{2}}, \qquad x \geq 0$$
  (ii) Mean and variance: (scale gamma)
  $$E(X) = \qquad\qquad V(X) =$$
  (iii) Mgf:
  $$M_X(t) =$$
- Notes about the chi-square distribution.
  - Special case of gamma: Chi-square is more commonly represented as the **scale parameterization** of the gamma (not the version we used).
    Scale $\theta = \frac{1}{\text{rate}} = \frac{1}{\beta}$
    This changes the pdf, expected value, variance and mgf slightly.
  - Pdf: Just the (scale) gamma density function with $\alpha = r/2, \;\; \theta = 2$.
  - Mean and variance: Just the mean and variance of (scale) gamma with the specific parameter values above.
  - Characteristics: Right-skewed density function. Unbounded support.
  - Probabilities: Just like the gamma distribution, probabilities need to be found using software. There is also a table of probabilities (just like a Z-table), because it is used in statistical tests.
  - Parameter: The degrees of freedom r must be a positive integer when used in the gamma distribution.
    Here is how it affects the shape (recall α is the shape parameter) of the density curve: Larger values of r shifts the probability to the right.



Important facts about the chi-square random variables

1. If $Z \sim \text{Normal}\,(0,1)$, then $Z^2 \sim \chi^2\,(1)$.
   - So the square of a standard normal random variable follows a chi squared distribution with 1 degree of freedom.
   - We will not prove this, but this means we can start with a normal random variable X with mean μ and standard deviation σ and end up with a chi-square random variable:

2. If $X_1, \ldots, X_n$ are mutually independent and $X_i \sim \chi^2\,(r_i)$ for $i = 1, \ldots, n$, then $Y = X_1 + \cdots + X_n \sim \chi^2\,(r_1 + \cdots + r_n)$.
   - The degrees of freedom are additive.
   - Proof:

   - Result / extension of this: If $X_1, \ldots, X_n$ are mutually independent random variables with $X_i \sim \text{Normal}\,(\mu_i, \sigma_i)$ for $i = 1, \ldots, n$, then

     This is one way to think about what the parameter r represents:

     $r =$ the number of independent standard normals that we are adding together.

Return to theorem for $\bar{X}$ and S^2

(c) $\dfrac{(n-1)}{\sigma^2}S^2 = \displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$ has a chi squared distribution with $n-1$ degrees of freedom.

- We will not prove this, but we can understand the pieces of the theorem (the coefficients in front of S^2 and the chi-square result). Recall $X_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \qquad\qquad \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Example

- Let $X_1, \ldots, X_4$ be a random sample from of size 4 from a normal distribution $N(\mu = 76, \sigma = 383)$ and

$$U = \sum_{i=1}^{4}\left(\frac{X_i-76}{383}\right)^2 \qquad \text{and} \qquad W = \sum_{i=1}^{4}\left(\frac{X_i-\bar{X}}{383}\right)^2$$

Compute $P(0.7 < U < 7.8)$ and $P(0.7 < W < 7.8)$.

Derived distributions: Student's *t* and Snedecor's *F*

Derivation of the t distribution

- We are studying the properties of $\bar{X}$ in order to make inferences about μ.
- If $X_1, \ldots, X_n$ are a random sample for a $N(\mu, \sigma^2)$, we know that the quantity

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

- If we knew the value of σ and we measured $\bar{X}$, then we could use this as a basis for inference about μ, since μ would then be the only unknown quantity.

  However, most of the time σ is unknown. So, Student (W. S. Gosset) looked at the distribution of

$$\frac{\bar{X} - \mu}{/\sqrt{n}}$$

  as a quantity that could be used as a basis for inference about μ when σ was unknown.
- When deriving this, we have the form of the statistic, but don't know its pdf (distribution). So here is the logic to get this statistic into a form that we are then able to find the actual equation for the pdf.

- Thus, the distribution of interest can be found by solving the simplified problem of finding the distribution of $\dfrac{Z}{\sqrt{U/r}}$, where $Z \sim \text{Normal}\,(0,1)$, $U \sim \chi^2\,(r)$ and $Z \perp\!\!\!\perp U$.

  There are two ways to go from here, but we will not show this.

Definition of the t distribution

- Let $X_1, \ldots, X_n$ be a random sample form a $N(\mu, \sigma^2)$ distribution. If

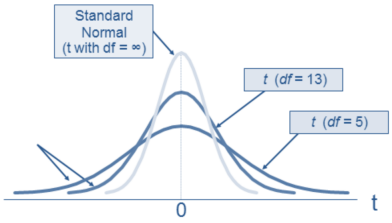
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \qquad \text{then} \qquad T \sim t_{\,n-1}$$

  Equivalently, a random variable T has **Student's *t* distribution** with r degrees of freedom, and we write $T \sim t_{\,r}$ if it has pdf

$$f_T(t) = \frac{\Gamma(\frac{r+1}{2})}{\frac{1}{\sqrt{r\pi}}\Gamma(\frac{r}{2})}\left(\frac{1}{(1+t^2/r)^{(r+1)/2}}\right), \qquad -\infty < t < \infty$$

Notes about the t distribution

- Density curve, relationship to the standard normal distribution, and probabilities.



  - The t-distribution has the "shape" of a normal distribution (bell-shaped and symmetric about zero), but it has heavier tails. This means there is more probability in the further from the center (and less around the center, zero).
  - As the degrees of freedom r increases, more probability shifts towards the center. Theoretically, as $r \to \infty$, t-distribution tends to the standard normal.
  - Example: Let $T \sim \text{t}_{\,10}$ and $Z \sim \text{N}\,(0,1)$. Compare the following probabilities (t probabilities must be found using software or a t-table):

    (a) Central interval probability: $\quad P(-1 < Z < 1) \quad$ and $\quad P(-1 < T < 1)$

    (b) Tail probability: $\quad P(Z > 2) \quad$ and $\quad P(T > 2)$.

- Mean and variance.
  - If $T \sim t_r$, then

$$E(T) = 0 \quad \text{if } r > 1 \quad \text{Only exists if 2 or more } df$$

$$V(T) = \frac{r}{r-2} \quad \text{if } r > 2 \quad \text{Only exists if 3 or more } df$$

- Moments and mgf.
  - In general, if there are r degrees of freedom, then there are only $r-1$ moments (recall first moment is $E(X)$ and $V(X)$ is the second central moment).
  - Student's t distribution does not have an mgf.
  - Informal theorem for existence of mgf:

    If the mgf exists, then all moments exist. But the opposite is not true (i.e. all moments existing doesn't always mean that the mgf exists).

    If any moment doesn't exist, then the mgf doesn't exist.
- Use in statistics.
  - The t-distribution is very important in inferential statistics and is used in one sample tests and confidence intervals of populations means, as well as simple linear regression for testing a population slope.

Derivation of the F distribution

- Another important derived distribution is Snedecor's F, whose derivation is quite similar to that of Student's t.
- Setup: Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu_X, \sigma_X^2)$ population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent $N(\mu_Y, \sigma_Y^2)$ population.
- Goal: If we were interested in comparing the variability of the populations, one quantity of interest would be the ratio of population variances $\frac{\sigma_X^2}{\sigma_Y^2}$.

  We can estimate this with
- The F distribution allows us to compare these quantities by giving us a distribution of

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2}$$

- Again we have the form of the statistic and need to rearrange until we have the form of some familiar distributions (and then the pdf can be found).

- The distribution of the above can be founded by solving the simplified problem of finding the distribution of $\dfrac{X_1/r_1}{X_2/r_2}$, where $X_1 \sim \chi^2(r_1)$, $X_2 \sim \chi^2(r_2)$ and $X_1 \perp\!\!\!\perp X_2$. Again we will not show this.

Definition of the F distribution

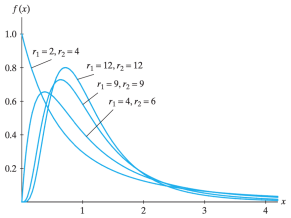
- Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu_X, \sigma_X^2)$ population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent $N(\mu_Y, \sigma_Y^2)$ population. If

$$W = \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \qquad \text{then} \qquad W \sim F(n-1, m-1)$$

  In general, if W has an F distribution with r_1 numerator degrees of freedom and r_2 denominator degrees of freedom, then we write $W \sim F(r_1, r_2)$.

Notes about the F distribution

- Density curve (more probability gets centered around 1 as *df* increase; range $0 < F < \infty$).



- Pdf, mean, variance and mgf (which doesn't exist).
  - We aren't going to worry about the pdf (it is more ugly than the t pdf) or ever calculate means and variances for this distribution.
- Probabilities.
  - Again, we need software such as R to calculate these.
  - Example: Let $X_1 \sim F(8, 4)$ and $X_2 \sim F(16, 8)$. Find the following probabilities.

    $P(X_1 > 5) =$

    $P(X_2 > 5) =$

- Relationship to other distributions.
  - Theorem:

    (a) If $X \sim F(r_1, r_2)$ then $1/X \sim F(r_2, r_1)$.

    The reciprocal of an F variable is again an F variable.

    (b) If $X \sim t_r$ then $X^2 \sim F(1, r)$.

    The square of a t random variable is an F variable with 1 and r *df*.

- Use in statistics.
  - The F-distribution is also very important in inferential statistics and is used in ANOVA when comparing means of two populations and also has lots of applications in regression.

Lecture 2 – Order Statistics

MATH 321: Mathematical Statistics

Lecture 2: Order Statistics

Chapter 6: Point Estimation (6.3)

Order statistics

Introduction

- Sample values such as the smallest, largest, or middle observation from a random sample can provide additional summary information. For example, the median price of houses sold during the previous month might be useful for estimating the cost of living.

Definition

- The **order statistics** of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$.

  The order statistics are random variables that satisfy $X_{(1)} \leq \cdots \leq X_{(n)}$. In particular

$$\begin{aligned} X_{(1)} &= \min_{1 \leq i \leq n} X_i, \\ X_{(2)} &= \text{second smallest } X_i \\ &\vdots \\ X_{(n)} &= \max_{1 \leq i \leq n} X_i. \end{aligned}$$

- The formulas for the pdfs of the order statistics for a random sample from a continuous population will be the main topic in this section.
- Notes:
  - The distribution of $X_{(j)}$ is not the same as the distribution of X_j
  - The range / support of is always the same as the random variable you are sampling from.

Bivariate case, min and max of two random variables

- Before we generalize order statistics to n random variables, we will look at the bivariate case.

  This means we are studying the min and max of two random variables.

- Derivation of the distributions of the functions $min(X_1, X_2)$ and $max(X_1, X_2)$.

  Setup: Let $X_i \overset{iid}{\sim} f(x)$ for $i = 1, 2$. We also have $F_X(x)$ and $S_X(x) = 1 - F_X(x)$.

  - Minimum: Using the above notation with $n = 2$, let $X_{(1)} = min(X_1, X_2)$.
  - We need to set up a probability statement that will make finding the distribution of $min(X_1, X_2)$ easier.

  - Now we can find the distribution.

  - Maximum: Let $X_{(2)} = max(X_1, X_2)$.

- Examples:

  1. Let $X_i \overset{\perp\!\!\!\perp}{\sim} \text{Exp}(\lambda_i)$. Find the distribution of $min(X_1, X_2)$.

     - Adding context: Suppose X_1 and X_2 are independent waiting times for accidents in two towns where $X_i \overset{iid}{\sim} \text{Exp}(\lambda = 1)$. Then $Min = min(X_1, X_2) \sim$
     - This can be interpreted in a natural way. In each of two separate towns we are waiting for the first accident in a process where the average number of accidents is 1 per month. When we study the accidents of both towns we are waiting for the first accident in the process where the average number of accidents is a total of 2 per month.

  2. Let $X_i \overset{iid}{\sim} \text{Uniform}(0, 10)$ for $i = 1, 2$.
     (a) Find the distribution of $min(X_1, X_2)$ and $max(X_1, X_2)$.

(b) Find the expected value of $min(X_1, X_2)$ and $max(X_1, X_2)$.

Generalizing order statistics

- Example: Let X be a random variable with pdf $f(x) = 2x, \quad 0 < x < 1$ and let $X_1, \ldots, X_5$ be a random sample from X.

  (a) Find the pdf of $X_{(1)}$, the first order statistic.

  Note: One strategy is to find cdf first, and then take derivative to find pdf.

  $f_{X_{(j)}}(x) = F'_{X_{(j)}}(x)$.

  Also we are going to frame everything using cdfs rather than survival functions like with the bivariate case.

(b) Find the pdf of $X_{(4)}$, the fourth order statistic.

(c) Find $P(X < 1/2)$, $P(X_{(1)} < 1/2)$, and $P(X_{(4)} < 1/2)$.

Order statistics distribution theorem

- Theorem: Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the **cdf of $X_{(j)}$** is

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}$$

and the **pdf of $X_{(j)}$** is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} [F_X(x)]^{j-1} f_X(x) [1 - F_X(x)]^{n-j}$$

- Walk through for proof of theorem:
  - Obtaining the pdf for the jth order statistic is the main goal. To do this, we first find the cdf for $X_{(j)}$ and then differentiate it to get the pdf.

    The equation for the cdf of the jth order statistic is closely related to the cdf of the binomial distribution.
  - $X_{(j)}$ represents the jth smallest value. So the cdf is

    $$F_{X_{(j)}}(x) = P(X_{(j)} \leq x)$$

    * Interpretation: This is the probability that at least j of X_is are less than or equal to x.

      So, we are essentially just counting something, specifically the number of random variables in our random sample less than x.
    * Example: Let $X_1, \ldots, X_5$ be a random sample from $f(x)$. We are interested finding in $P(X_{(3)} \leq 4)$.

      Note: If $X_{(j)} \leq x$, then $X_{(j-1)} \leq x$ must be true. But $X_{(j+1)} \leq x$ can also be true.

Data	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
(a)	2	3	4	5	6
(b)	2	3	4	4	5
(c)	2	3	4	4	4
(d)	2	4	4	5	6
(e)	2	5	4	5	6

  - Thus, we can use the binomial distribution to find the cdf of $X_{(j)}$.

    We can define the event of success as $\{X_j \leq x\}$, because we are counting how many of the original sample $X_1, \ldots, X_n$ are less than x.
  - Let Y be a random variable that counts the number of $X_1, \ldots, X_n$ less than or equal to x.

    Then, we see that $Y \sim$

- Then $f_{X_{(j)}}(x) = \frac{d}{dx} F_{X_{(j)}}(x)$.

  This derivation is not straightforward, but it can be intuitively understood.

- Concept: The pdf assigns our n random variables to three groups of sizes:

  Recall a **partition** of n objects into k groups of sizes $n_1, \ldots, n_k$ equals $\frac{n!}{n_1! \cdot \ldots \cdot n_k!}$.

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!\,1!\,(n-j)!} \left[P(X \leq x)\right]^{j-1} f_X(x) \left[P(X > x)\right]^{n-j}$$

- It is worth noting the special cases for the extreme order statistics (then show for bivariate case):

  Smallest: $f_{X_{(1)}}(x) = nf(x)[1 - F(x)]^{n-1} \rightarrow$

  Largest: $f_{X_{(n)}}(x) = n[F(x)]^{n-1} f(x) \rightarrow$

Specific order statistics and functions of order statistics

- Several very common statistics are actually order statistics.

  The importance of order statistics has increased because of more frequent use of non-parametric inferences and robust procedures.

- Sample median:

  - The sample median, which we will denote by M, is a number such that approximately one-half of the observations less than M and one-half are greater.
  - In terms of the order statistics, M is defined by

$$M = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \left[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right]/2 & \text{if } n \text{ is even} \end{cases}$$

    So it is the sole middle observation or the average of the two middle observations.

  - The median is a measure of location that might be considered as alternative to the sample mean.
  - Sample mean vs sample median.
    * Sample mean: Can be **more efficient** (i.e. more accurate in some sense because it's using all of the data ($\sum X_i$)).

      But **less robust** because it can be affected by outliers when using all of the data.
    * Sample median: **Less efficient** because it only uses the first half of the data (e.g. if $\boldsymbol{x} = \{1, 2, 3, 4, 5\}$, it is only using 1, 2 and 3 to find the median (starts from left and stops when it gets to the median)).

      But **more robust** because it is only using half the data.

- Sample range, $R = X_{(n)} - X_{(1)} = max(X_1, \ldots, X_n) - min(X_1, \ldots, X_n)$.

  This is a measure of spread which gives the distance spanned by the entire sample.

- $IQR = Q_3 - Q_1$.

  In terms of order statistics, given an even $2m$ or odd $2m + 1$ random variables:

  $$\begin{array}{rcl} Q_1 & = & \text{median of the smallest } m \text{ values} \\ Q_3 & = & \text{median of the largest } m \text{ values} \end{array}$$

  This is a measure of spread that might be considered as alternative to the standard deviation. It is better for skewed data or when there is outliers.

- Midrange $= \dfrac{X_{(1)} + X_{(n)}}{2}$.

  This is a measure of location like the sample mean or median. It is found by averaging (or taking the midpoint) of the min and max of the random sample.

- To find the distributions of functions of order statistics, e.g. involving more than than one statistic such as the sample range R or midrange, we have two options:

  a) Find the pdf of multiple ordered statistics (i.e. multivariate transformation).

  b) OR we can use simulation! (like we did in our sampling distribution R notes)

  - First we could simulate the sampling distribution of the statistic of interest.

    Then use those results to approximate any quantity we need!
  - For example, suppose we have 10,000 values for $\hat{R} = max(x_1, \ldots, x_n) - min(x_1, \ldots, x_n)$.

    $E(R) \approx$

    If we want to estimate $P(R > x)$. Let I be an indicator variable such that

    $$I = \begin{cases} 1 & = & \text{if } R > x \\ 0 & = & \text{if } R \leq x \end{cases}$$

    $P(R > x) \approx$

- Simulation is a very powerful tool that allows researchers to study things that don't have theoretical solutions.

- Examples:

  1. Continuing previous example:

     $X_1, \ldots, X_5 \overset{iid}{\sim} f(x) = 2x$ and $F(x) = x^2 \quad 0 < x < 1.$

     (a) Find the cdf of the sample median $X_{(3)}$.

     (b) Find the pdf of the sample median $X_{(3)}$.

2. Wind damage to insured homes are independent random variables with common pdf and cdf

$$f(x) = \frac{3}{x^4} \quad x > 1 \qquad \longrightarrow \qquad F(x) = 1 - \frac{1}{x^3} \quad x > 1$$

where x is in thousands of dollars. Find the expected value of the largest of three such claims.

Order statistics as estimators of population percentiles

- Expected value of the "position" of order statistics.

  Theorem: Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample of size n from a continuous population with cdf $F_X(x)$. Then

$$E[F_X(X_{(j)})] = \frac{j}{n+1}, \quad j = 1, \ldots, n$$

- Breaking down theorem:
  - For our population distribution, $F_X(x) = P(X \leq x) = p$ represents the cumulative probability up to and including x, or equivalently the area under $f(x)$ less than x.

    Recall that probability is a function, so here we are inputting a constant, particular x value and getting the corresponding probability p as a result (which is also a constant).

- Now if we input the jth order statistic $X_{(j)}$ (which is a random variable) into $F_X(x)$, the output is a random area ($\approx$ random variable p), which represents the probability X is less than or equal to $X_{(j)}$:

$$F_X(X_{(j)}) = P(X \leq X_{(j)})$$

- Because it is a random variable, we can find the expected value.

$$E[F_X(X_{(j)})] = \frac{j}{n+1}$$

Example: Let $n = 9$ and $j = 6$. Find $E[F_X(X_{(6)})]$.

- Using this theorem:
  - Recall for $0 \leq p \leq 1$ the **$100p^{th}$ percentile of X** is the number x_p defined by

$$P(X \leq x_p) = F(x_p) = p$$

  - Thus, we can use $X_{(j)}$ as an estimator of x_p, where $p = j/(n+1)$.

    Note that p is a function of j and $n \Longrightarrow$ We are figuring out which percentile, x_p, $X_{(j)}$ estimates.

$$F(x_p) = p \quad \longrightarrow \quad F(x_{j/(n+1)}) = \frac{j}{n+1}$$

q–q plots

- Extension of previous theorem:
  - Now let's consider the previous order statistic $X_{(j-1)}$ as well, which is of course another random variable.

    $F_X(X_{(j)}) - F_X(X_{(j-1)})$ represents the probability (area under curve) between two adjacent order statistics $X_{(j)}$ and $X_{(j-1)}$. The expected value of this random area is

$$E[F_X(X_{(j)}) - F_X(X_{(j-1)})] =$$

  - We could also show the area below the first order stat and the area above the last order stat:

$$E[F_X(X_{(1)})] = \frac{1}{n+1} \qquad \text{and} \qquad E[1 - F_X(X_{(n)})] = \frac{1}{n+1}$$

– This means the order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ partition the range of X into $n+1$ parts and thus create $n+1$ areas under $f(x)$ and above the x-axis.

On average, each of the $n+1$ areas equals $1/(n+1)$.

- So, we can use the relationships shown above to test whether a random variable X has a certain distribution by "matching up" the sample order statistics with the theoretical percentiles. This is the process to get the numbers used in a q–q plot.

  (1) Compute the percentiles $x_{\frac{1}{n+1}}, \ldots, x_{\frac{n}{n+1}}$ of the population distribution we are testing.

  (2) Compare (1) to the observed sample order statistics $x_{(1)}, \ldots, x_{(n)}$.

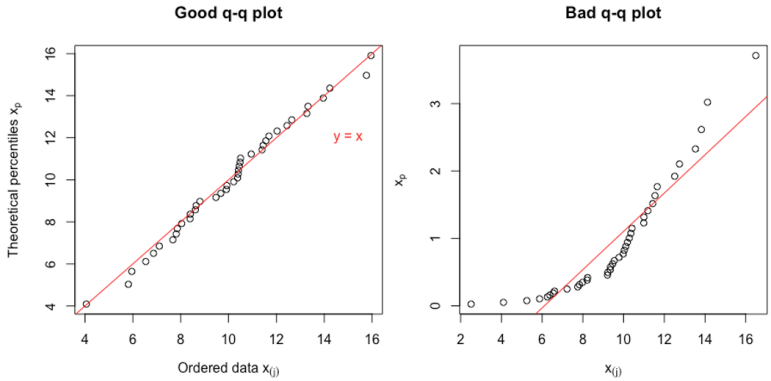
  If the theoretical distribution is a good model for the observations, then we should see

$$x_{(1)} \approx x_{\frac{1}{n+1}} \quad , \quad \cdots \quad , \quad x_{(n)} \approx x_{\frac{n}{n+1}}$$

- Definition: Let X be a random variable, $x_{(1)}, \ldots, x_{(n)}$ be the observed sample order statistics of a random sample of size n, and $x_{\frac{1}{n+1}}, \ldots, x_{\frac{n}{n+1}}$ be the percentiles from some particular distribution. A plot of the points

$$(x_{(1)}, x_{\frac{1}{n+1}}) \ , \ \cdots \ , \ (x_{(n)}, x_{\frac{n}{n+1}})$$

is known as a **quantile–quantile plot**, or more simply a **q–q plot**.



- Interpretation of a q–q plot.
  - If we picked a good model (i.e. X has the particular distribution), then $x_{(j)} \approx x_{\frac{j}{n+1}}$ and the q–q plot should be nearly a straight line through the origin with slope = 1 (i.e. diagonal line).
  - Conversely, a strong deviation from this line is evidence that the distribution did not produce the data.
  - Sidenote: It's called a quantile–quantile plot because the sample order statistics $x_{(1)}, \ldots, x_{(n)}$ associated with the sample $x_1, \ldots, x_n$ are called the **sample quantiles of order $j/(n+1)$** and the percentile x_p of a theoretical distribution is the **quantile of order p**, and we are using $p = j/(n+1)$ to match them up.

- Using q–q plots.
  - Usually we are not trying to see if the data come from a particular distribution, but rather from a parametric family of distributions (such as the normal, uniform, or exponential, etc.).

    We are usually forced into this situation because we don't know the parameters. So typically, the next step, after the q–q plot, may be to estimate the parameters, which we will learn how to do later.

- q–q plots for the normal distribution.
  - q–q plots are often used to test whether a random sample is from a normal distribution.
  - When creating the plot, we of course need to calculate the theoretical percentiles. This requires specifying μ and σ^2 when using invNorm() or qnorm(); but as mentioned these are usually unknown.

    So we have two strategies:

    1. We could use the sample statistics as best guess of the population parameters ($\bar{X} \to \mu$ and $S^2 \to \sigma^2$), as we know these are unbiased estimators.

       If we do this, the q–q plot should follow the diagonal line.

    2. If we don't want to make this assumption. We can make use of the relationship to the standard normal distribution:

Thus if we vary p and plot (x_p, z_p), we get a straight line with slope $1/\sigma$.

This means we can still test if a random sample came from a normal distribution without having to know / guess the mean and standard deviation.

So if we plot $(x_{(1)}, z_{\frac{1}{n+1}})$, ... , $(x_{(n)}, z_{\frac{n}{n+1}})$, which has our ordered sample data on the x-axis now as estimates of the population percentiles, we should see an approximately straight line. If so, then $\frac{1}{\text{slope}}$ is an approximation of σ.

Lecture 4 – Point Estimation

MATH 321: Mathematical Statistics

Lecture 4: Point Estimation

Chapter 6: Point Estimation (5.8 and 6.4)

Introduction

The process, where we have been

- Suppose we were given a dataset and went through the EDA where we created lots of summary statistics, histograms, box plots, etc. By this point, we would have a good "feel" for the data.
- If we were focusing on determining (to the best of our ability) the population distribution that a variable came from, we would have used the shape of the sample distribution to guide our selection of a few potential models to test with q–q plots.

  Usually we are not trying to see if the data come from a particular distribution, but rather from a parametric family of distributions (such as the normal, uniform, or exponential, etc.). We are usually forced into this situation because we don't know the parameters.
- Suppose we find a good model, what next? Typically, the next step may be to estimate the parameters, which is what this section is all about.

  This section / topic can be divided into two parts:

  1. Evaluating estimators.
  2. Methods of finding estimators.
- In general, these two activities are intertwined. Often the methods of evaluating estimators will suggest new ones. We will focus mainly on finding estimators.

Point estimation

- Rationale
  - The rational behind point estimation is quite simple. When sampling from a population described by a pdf or pmf $f(x \mid \theta)$, knowledge of θ yields knowledge of the entire distribution.

    Hence, it is natural to seek a method of finding a good estimator of the point θ. For example, if we assume that the population is normally distributed and we know μ and σ^2, then we know everything about the distribution.
  - It may also be the case that some function of θ, say $\tau(\theta)$, is of interest. The second method described in this section can be used to obtain estimation of $\tau(\theta)$.

- Point estimator
  - Definition: A **point estimator** is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator.
  - Notes about definition:
    - Makes no mention of any correspondence between the estimator and the parameter to be estimated.

      If this were a part of the definition, it would restrict the available set of estimators.

      So, any statistic $\rightarrow$ We could use the sample _____ as a point estimator for the population _____, but it would be a bad estimator because we get no insight about _____
    - Also, there is no mention in the definition of the range of the statistic $W(X_1, \ldots, X_n)$.

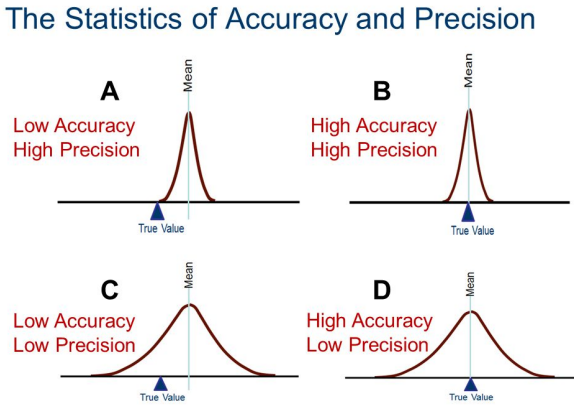
      While, in principle, the range of the statistic should coincide with that of the parameter, this is not always the case. For example, if we need $\mu > 0$ but get $\bar{x} = -5$ based on the observed data, this is bad...
  - So, at this point, we want to be careful not to eliminate any candidates from consideration.

- Estimator vs Estimate
  - An **estimator** is a function of the sample; so it is a _____ (i.e. because it is a function of *iid* random variables $X_1, \ldots, X_n$).
  - An **estimate** is the _____ of an estimator that is obtained when the sample is actually taken; so it is just a number (because it is a function of the realized values $x_1, \ldots, x_n$).

Evaluating estimators

Introduction

- There are two ways that we will evaluate estimators. In other words, there are two criteria we apply to determine how "good" an estimator is.



- Some estimators will be good at one aspect and poor in another, so there is often a tradeoff between accuracy and precision.
- Now we will formalize the theoretical ideas of accuracy and precision.

Unbiasedness

- This criteria deals with the location of the sampling distribution of a statistic.
- Definition: Let $X_1, \ldots, X_n$ be a random sample from X and let θ be a parameter of the pdf (or pmf).

  If $W(X_1, \ldots, X_n)$ is some function of $X_1, \ldots, X_n$ and $E[W(X_1, \ldots, X_n)] = \theta$, then $W(X_1, \ldots, X_n)$ is an **unbiased estimator** of θ. Otherwise it is said to be **biased**.
- Specific examples
  - If $\mu = E(X) = \theta$ is a parameter of the pdf (or pmf) of X, then $E(\bar{X}) = \mu = \theta$ and thus $\bar{X}$ is always an unbiased estimator of μ.

    Ex) For $X \sim$ Poisson (λ):
  - If $\sigma^2 = V(X) = \theta$ is a parameter of the pdf (or pmf) of X, then $E(S^2) = \sigma^2 = \theta$ and thus the sample variance S^2 is always an unbiased estimator of σ^2.

    Ex) If $X \sim$ Normal (μ, σ^2):

Consistency

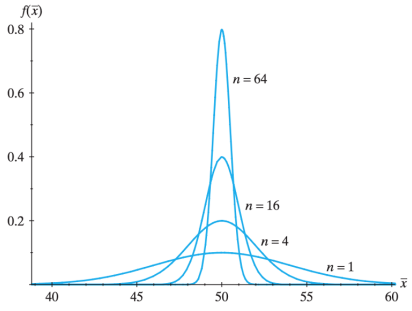
- This criteria deals with the variance of the sampling distribution of a statistic. Before we can formalize this, we need to learn another concept called convergence in probability and some associated theorems.

Convergence in probability idea

- The idea
  - When studying the mean $\bar{X}$ of a random sample of size n from a distribution with mean μ and variance $\sigma^2 > 0$, we saw that is a random variable with the following properties

$$E(\bar{X}) = \mu \qquad \text{and} \qquad V(\bar{X}) = \frac{\sigma^2}{n}$$

  - Thus, as the sample size n increases, the variance of $\bar{X}$ decreases.



pdfs of means of samples from $N(50, 16)$

  - We can see that as n increases, the probability becomes concentrated in a small interval centered at μ.

    That is, as n increases, $\bar{X}$ tends to converge to μ, or $(\bar{X} - \mu)$ tends to converge to 0 in a probability sense.

- Convergence in statistics
  - Convergence in statistics is very different from that in mathematics.
  - In mathematics, a sequence of **constants** $a_1, a_2, \ldots$ converges to a constant:

$$\lim_{n\to\infty} a_n = a \qquad \text{ex. } \lim_{n\to\infty} \frac{1}{n} = 0$$

  - But in statistics, a sequence of **random variables** $X_1, X_2, \ldots$ converges to a random variable:

$$\lim_{n\to\infty} X_n = X$$

    (Note: it can also converge to a constant, depending on the situation.)

  - Three types of convergence in statistics:
    1. Convergence in probability.
    2. Almost sure convergence.
    3. Convergence in distribution.

    We will focus on number one, mention number three and ignore number two.

Convergence in probability definition

- This type of convergence is one of the weaker types and, hence, is usually quite easy to verify.
- Definition: A sequence of random variables, $Y_1, Y_2, \ldots,$ **converges in probability** to a random variable Y if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0 \qquad \text{or, equivalently,} \qquad \lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$$

Breakdown of definition

- Notation
  - $Y_1, Y_2, \ldots$ represent statistics that depend on the subscript (i.e. functions of a random sample). More specifically, Y_n is a statistic defined with the original *iid* variables $X_1, \ldots, X_n$.
  - So $Y_n = T(X_1, \ldots, X_n)$.

    For example, if Y_n is the sample mean, then

$$Y_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - So the distribution of Y_n changes as the subscript changes and it converges to some limiting distribution as n becomes large.

- Understanding $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon)$
  - For a given (fixed) n, is $P(|Y_n - Y| < \epsilon)$ is just a regular probability; so it is just a constant.
  - So $P(|Y_n - Y| < \epsilon) = a_n$ and consequently

$$\lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = \lim_{n\to\infty} a_n$$

    * This is the convergence that we are familiar with in mathematics (more specifically in real analysis).
    * Rigorously, $\lim_{n\to\infty}$ can only be used with a sequence of constants and cannot be used with a sequence of random variables.

      It doesn't make sense to find the limit of a random variable (we can't find the pattern if each number is random and has a pattern of its own).
    * But probability is a constant number that we can find a limit of. Thus, using $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon)$ notation makes sense.
  - So, because Y_n is a random variable, we cannot find its limit directly, but we can find its limit in probability or distribution.

- Interpretations
  - The event $|Y_n - Y|$ is the difference between Y_n and Y.
  - $P(|Y_n - Y| < \epsilon)$ is the probability that the difference between Y_n and Y is smaller than ϵ.

    In definition, "for every $\epsilon > 0$" means that we can pick any really tiny number (e.g. $\frac{1}{100000000}$).
  - Putting it all together: $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$

    Even though we choose a really tiny ϵ, the probability that the difference between Y_n and Y is less than the small number converges to one as n goes to ∞.
  - In other words, the probability that there is no difference between Y_n and Y goes to one as n approaches ∞.

    Thus, we can conclude that Y_n converges to Y **in probability.**

- Correct notation (note that we need the "in probability" part in all of these):
  - $Y_n \xrightarrow{p} Y$.
  - $Y_n \to Y$ in probability.
  - $\lim_{n\to\infty} Y_n = Y$ in probability.

(Weak) Law of Large Numbers (WLLN)

- Theorem
  - Frequently, statisticians are concerned with situations in which the limiting random variable is a constant and the random variables in the sequence are sample means (of some sort). The most famous result of this type is the following.
  - **WLLN** Theorem: Let $X_1, X_2, \ldots$ be *iid* random variable with $E(X_i) = \mu$ and $V(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$,

    $$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1 \qquad \lim_{n\to\infty} P(|\bar{X}_n - \mu| \quad \epsilon) =$$

    that is, $\bar{X}_n$ converges in probability to μ (notation: $\bar{X} \xrightarrow{p} \mu$).

- Notes about WLLN
  - Comparison
    - Convergence in probability definition: $\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0$
  - Summary of theorem:
    - The Weak Law of Large Numbers (WLLN) quite elegantly states that, under general conditions, the sample mean approaches the population mean as $n \to \infty$.

      This is because the probability associated with the distribution of $\bar{X}$ becomes concentrated in an arbitrarily small interval centered at μ as n increases.
    - Needed conditions: *iid* random variables and the first and second moments (i.e. the mean and a finite variance). And do not need any distributional assumption.
  - Consistency
    - The property summarized by the WLLN, that is a sequence of the "same" sample quantity approaches a constant as $n \to \infty$, is known as **consistency**.
    - Showing consistency is the same as showing convergence in probability.

      Thus, it can be said that $\bar{X}_n$ is a consistent estimator of μ.

  - WLLN for transformations of X (extension of theorem).
    - Additionally, it can be used for statistic of the form $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$, where $g(X)$ is non negative transformation that still has a mean and a finite variance.
    - So, now instead of converging to μ, $\bar{X}_n$ converges to $E[g(X)]$ (in probability).

- Proof of WLLN
  - Want to show: $\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$

- Application of WLLN
  - Example: Suppose we are planning a poll to figure out which is better, R or Excel. Let

$$X_i = \begin{cases} 1 & \text{if R} \\ 0 & \text{if Excel} \end{cases}$$

    and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

    1. If $n = 400$, find a lower bound on $P(|\bar{X}_{400} - p| < 0.05)$.

    2. If $n = 400$ and $p = 7/10$, find a lower bound on $P(|\bar{X}_{400} - 0.70| < 0.05)$.

    3. If $n = 500$ and $p = 7/10$, find a lower bound on $P(|\bar{X}_{500} - 0.70| < 0.05)$.

Summary of consistency and unbiasedness

- Comparison of unbiasedness vs consistency.
  - Unbiasedness → This tells us the mean of a statistic, regardless of n. So we can drop the inference on n. To be unbiased, the expected value of the statistic must equal the parameter of interest.

  - Consistency → This is all about the limit of the random variable as $n \to \infty$. If a statistic is consistent, then as $n \to \infty$, there is no variation in what the statistic converges to; the entire distribution converges to a constant.

- Examples of the difference (shown through counter examples):
  1. Let $Y_n \sim \mathrm{N}(\mu, \sigma^2)$.

     So it still has some variation, whereas a constant has no variation (it is always the same).

     $E(Y_n) = \mu$ _____ $Y_n \xrightarrow{p} Y$.

  2. Now let $Y_n \sim \mathrm{N}(\mu + \frac{1}{n}, \frac{\sigma^2}{n})$

     So, the mean of the distribution converges to μ and the variance disappears.

     $Y_n \xrightarrow{p} \mu$ _____ $E(Y_n) = \mu$.

Return to methods of finding estimators

- Now that we have covered how to evaluate estimators, we can look at how to find estimators.
- In many cases, there will be an obvious or a natural candidate for a point estimator of a particular parameter. For example:
  - Population mean $\mu \rightarrow$
  - If $X \sim \text{Uniform}\,(0, \theta) \rightarrow$
  - If $X \sim \text{Gamma}\,(\text{shape } \alpha, \text{rate } \beta) \rightarrow$
- For more complicated models, intuition may not work and can often have bad results (e.g. gamma(α, β), there is no obvious estimators for the shape and scale parameters).
- Therefore, it is useful to have some techniques (more methodical ways of estimating parameters) that will at least give us some reasonable candidates for consideration.

  These still must be evaluated before their worth is established. Ideally, point estimators will provide insight and information about the unknown parameter θ.
- Now we will go into two methods for finding estimators.

Method of moments

Method of Moments (MME)

- The method of moments is a very simple procedure for finding an estimator for one or more population parameters.
- Types of moments:
  - $\boldsymbol{k}^{\text{th}}$ **(population) moment** of the distribution (about the origin)

    $$\mu_k' = E(X^k)$$

  - The corresponding **sample moment** is the average

    $$m_k' = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- The method of moments logic
  - Based on the intuitively appealing idea that sample moments should provide good estimates of the corresponding population moments.
  - Population moments $\mu_1', \ldots, \mu_k'$ are usually functions of the population parameters, so we can equate corresponding population and sample moments and solve for the desired estimators.

- Official statement of **Method of Moments**:

  Choose as estimates those values of the parameters that are solutions of the equations $\mu'_k = m'_k$, for $k = 1, 2 \ldots, t$, where t is the number of parameters to be estimated.

- Steps to find MME:

  1. Write $E(X^k)$ as a function of the parameters of interest.

     Note: Might have to do some integration or summation to get $E(X^k)$.

     Example: If $X \sim \text{Normal}(\mu, 1) \rightarrow$

  2. Then estimate the parameter of interest by equating the population moment with the sample moment and solving for the parameter.

     Example continued:

Examples

1. Let $X_1, \ldots, X_n$ be a random sample from Uniform$(0, \theta)$, where θ is unknown. Use the method of moments to estimate the parameter θ.

2. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Find the method of moments estimator for λ.

3. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}(-\theta, \theta)$. Find the MME for θ. Recall $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$ if $X \sim \text{Uniform}(a, b)$.

4. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. Find the MMEs for μ and σ^2.

   Note: There are two unknown parameters. So we will have to setup and solve a system of equations.

Summary of method of moments

- Pros
  - Simple to find and fairly intuitive (simply matching the properties of a sample to that of the population distribution).
  - Nonparametric method.

    So it works without the distributional information about the population (think back to the normal MME example, those results for estimators of $E(X)$ and $V(X)$ are true for regardless of what we start with.

    Example: Suppose $X_i \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$, then $\bar{X}$ is an estimator for α/β and v is an estimator for α/β^2.

    This means that we don't have to assume the population distribution, which is a useful property when the population distribution information is unclear.
  - Consistent estimators most of the time.

    Sample moments are consistent estimators of the corresponding population moments (can show with the (Weak) Law of Large Numbers).
- Cons
  - Nonparametric method.

    Because of this, MME information is only based on the data and doesn't give us any information about that relationship with the parameter of interest.
  - Often biased (so the center of the distribution of the estimator doesn't line up with θ).
  - May be inefficient (i.e. large variance of the distribution of $\hat{\theta}$).

Maximum likelihood estimation

Context

- We just saw one way to get estimators, but noted that there are some disadvantages of that method. One reason for this is that it is a very general method because it is non-parametric. So how can we improve upon it?
- Parametric statistics: Assume the distribution of X and estimate the parameters that determine the distribution.

  Example: Know $X \sim$ Normal , estimate μ and σ^2.
- The method of maximum likelihood is, by far, the most popular technique for deriving estimators.

Motivating (conceptual) example

- Suppose that we are confronted with a box that contains three balls. We know that each of the balls may be red or white, but we do not know the total number of either color. However, we are allowed to randomly sample two of the balls without replacement.
- If our random sample yields two red balls, what would be a good estimate of the total number of red balls in the box?

Likelihood function

- **Parameter space** definition: Given pdf (or pmf) $f(x \mid \theta_1, \ldots, \theta_k)$ the set of all possible values for $\theta_1, \ldots, \theta_k$ is known as the parameter space.

  We denote the parameter space with Θ (capital "theta").
- Examples:

  If $X \sim \text{Normal}\,(\mu, \sigma^2) \rightarrow \Theta =$

  If $X \sim \text{Poisson}\,(\lambda) \rightarrow \Theta =$

- Review: Joint pdf of $X_1, \ldots, X_n$ (if X_i's are continuous, *iid* random variables) is given by

  Pdf $f(x_i)$ is a function of X_i given the parameters $\boldsymbol{\theta}$: $f(X_i \mid \boldsymbol{\theta})$.

- **Likelihood function** definition: Let $f(\mathbf{x} \mid \boldsymbol{\theta})$ denote the joint pdf or pmf of the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by
$$L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x} \mid \boldsymbol{\theta})$$
  is called the likelihood function.

- Notes about the likelihood function
  - The only distinction between the likelihood function and the joint pdf or pmf is which variable is considered fixed and which is varying.

    In other words, the likelihood function is the same thing as the joint density of the data, but from a different point of view (i.e. different information is known).
    - For the joint density of the data, θ is fixed, while $\boldsymbol{X}$ can vary.

      This is used to answer probability questions: we know the _____ and want to figure out the _____.
    - For the likelihood function, $\mathbf{X}$ is fixed, while θ can vary.

      This is used to answer statistics questions: we have data and want to figure out the most likely _____.
  - Because both $\mathbf{x}$ and θ are in the formula, this gives us information about the **relationship** between the data and the parameter.
- We can find the likelihood function with

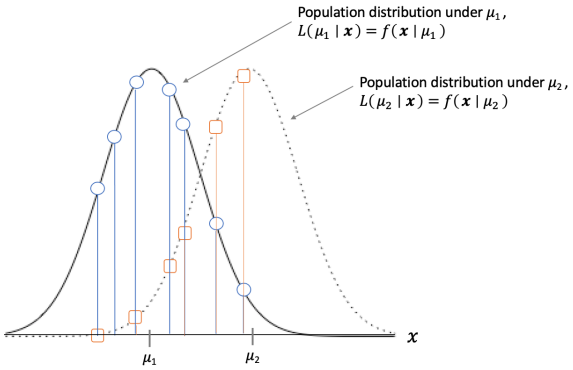
- Comparing likelihood functions (this is what exactly we did with the colored balls example!)
  - If $\mathbf{X}$ is a discrete random vector, then $L(\theta \mid \mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x})$. If we compare the likelihood at two parameter points and find that
$$P_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1 \mid \boldsymbol{x}) > L(\theta_2 \mid \mathbf{x}) = P_{\theta_2}(\boldsymbol{X} = \boldsymbol{x})$$
    then we interpret this as follows:
  - The sample $\mathbf{x}$ we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$ (with the same data).

    This can be interpreted as saying that θ_1 is a more plausible value for the true value of θ than θ_2.

Maximum likelihood estimation (MLE) definition and concept

- Definition: For each sample point $\mathbf{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which $L(\theta \mid \mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

- Notes about the definition
  - Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely.
  - In general, the MLE is a good point estimator, possessing some of the optimality properties such as consistency.
- MLE conceptualized



  - The likelihood function is the product of the density curve heights at the observed xs.
  - So for this example, _____ is a more plausible value for μ.

How to find an MLE

- Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Find the maximum likelihood estimator for λ. How do we do this?
- Start with the likelihood function:

- Then it's an optimization problem: To find the maximum of a function, we use calculus and derivatives.

  If the likelihood function is differentiable, we can solve for the points at which the first derivatives equals zero:

$$L'(\theta \mid \mathbf{x}) = \frac{d}{d\theta} L(\theta \mid \mathbf{x}) = 0$$

- Log likelihood
  - It is easier to work with the natural logarithm of $L(\theta \mid \mathbf{x})$ than it is to work with $L(\theta \mid \mathbf{x})$ directly. This is known as the **log likelihood**:

$$\ell(\theta \mid \mathbf{x}) = \ln[L(\theta \mid \mathbf{x})]$$

  - This transformation is valid since the natural log is a strictly increasing function on $(0, \infty)$ (so it's a one-to-one function), which means it's equivalent to maximize the natural log of the likelihood function.

  Continuing example:

  At this point, the solution $\hat{\theta}$ is only a **possible candidate** for the MLE of (θ).

  First derivative being zero is only a necessary condition, but not a sufficient condition because points at may be local or global minimum / maximum, or inflection points.

- So we have to check the second derivatives at $\hat{\theta}$ to ensure they are global maximum:

$$L''(\theta \mid \mathbf{x}) = \frac{d^2}{d\theta^2} L(\theta \mid \mathbf{x}) \qquad \rightarrow \qquad L''(\hat{\theta} \mid \mathbf{x}) \quad 0$$

  If this is true, then we know that we have found $\hat{\theta}_{MLE}$. Continuing example:

- Summary
  - Simply put, the likelihood function is hill shaped with the highest point at the MLE.
  - Note that the likelihood function not always differentiable, which adds in some extra complexity when finding the MLE.

    When this is the case, we can try to numerical maximization.
- The process that was just demonstrated was for univariate θ. It is the same process for a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, except we have to work with partial derivatives.

Steps to find MLEs

1. Write the likelihood function (i.e. joint density function) and the log-likelihood,

$$L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x} \mid \theta) \qquad \rightarrow \qquad \ell(\theta) = \ln[L(\theta \mid \mathbf{x})]$$

2. Optimize the log-likelihood function by taking the derivatives with respect to the parameter of interest.

   Set to zero and solve for the parameter of interest.

$$\ell'(\theta) = \frac{d}{d\theta}\ell(\theta) = 0 \qquad \rightarrow \qquad \hat{\theta} = \text{potential MLE}$$

3. Verify that the global maximum of the log-likelihood function occurs at $\theta = \hat{\theta}$.

   Find the second derivative of the log-likelihood function, then plug in $\hat{\theta}$ and see if less than zero.

$$\ell''(\theta) = \frac{d^2}{d\theta^2}\,\ell(\theta) \qquad \rightarrow \qquad \ell''(\hat{\theta}) \overset{?}{<} 0$$

   If so, then we have $\hat{\theta}_{MLE}$.

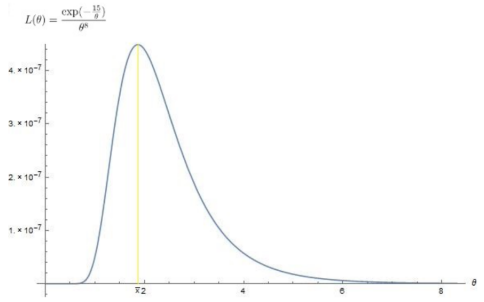
Examples

1. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Geometric}\,(p)$. Find the maximum likelihood estimator for p.

   (a) Find the likelihood function and log-likelihood function for p.

   (b) Optimize the log-likelihood function and solve for $\hat{p}$.

   (c) Perform second derivative test to confirm if $\hat{p}$ is the MLE for p.



2. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Normal}\,(\mu, \sigma^2)$. Find the MLEs for μ and σ^2.

   Note: Trying to find MLEs for two parameters, so will have to take partial derivatives.

When working with partial derivatives, the second derivative test checks to see if the determinant of the matrix of the second partial derivatives (called the Hessian matrix) is less than zero.

For this scenario, the solutions do provide a maximum.

Finding MLEs for functions of parameters

- We mentioned this before in the overview of point estimation that we be interested in some function of θ, say $\tau(\theta)$,

  A useful property of MLE is know as the invariance property of MLE.

- **(Invariance property of MLEs)**: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.
- Notes about this property
  - This type of theorem usually only holds with continuous functions, but this one works with ANY function. So we don't have to check any conditions.
  - This means if we want to find the MLE for $\tau(\theta)$:
    1. Find the MLE of θ.
    2. Simply apply the invariance property to get the MLE of $\tau(\theta)$.

- Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Geometric}\,(p)$.
  Find the MLE for for $V(X) = \dfrac{1-p}{p^2} = \tau(p)$.

Miscellaneous notes about MLEs

- Optimal properties
  - Results shows that under general conditions, MLEs are consistent estimators of their parameters and asymptotically efficient (small variance in the limiting distribution (think: convergence in distribution).
  - This means it is a method of finding an estimator that guarantees optimal properties, asymptotically.
- Maximation
  - The possibility of maximizing $L(\theta \mid \mathbf{x})$ is one of the most important features of MLEs.
  - Example: Let $X \sim \text{Gamma}\,(\alpha, \beta)$. Find the MLEs for α and β.

  - Turn to **numerical maximization**, which simply put essentially means plugging in a ton of numbers and seeing when the result is the largest.

    If a model (likelihood) can be written down, then there is some hope of maximizing it numerically and hence finding the MLEs of the parameters.

    When this is done, there is still always the question of whether a local or global maximum has been found.

- Numerical sensitivity
  - When we use numerical methods, we have to pay careful attention to a potential problem of **numerical sensitivity**. That is, how sensitive is the estimate to small changes in the data?
  - This situation arises when the MLE cannot be solved for explicitly (i.e. there is no closed form solution, perhaps because the derivative doesn't exist). This occurrence happens when the likelihood function is very flat in the neighborhood of its maximum or when there is not a finite max.
  - Example: The MLEs of n and p (both unknown) in binomial sampling can be highly unstable. Five realizations of a Binomial(n, p) experiment are observed.

    The first data set is (16, 18, 22, 25, **27**) and the MLE of n is $\hat{n} = \mathbf{99}$.

    The second data set is (16, 18, 22, 25, **28**) and the MLE of n is $\hat{n} = \mathbf{190}$.
  - So it is often wise to spend a little extra time investigating the stability of the solution.
  - If the MLE can be solved for explicitly, this is usually not a problem.

Test 3

Contents

Lecture 5 – The Central Limit Theorem

MATH 321: Mathematical Statistics

Lecture 5: The Central Limit Theorem

Chapter 5: Distributions of Functions of Random Variables (5.6 and 5.7)

The Central Limit Theorem (CLT)

Introduction

- The sample mean is one statistic whose large-sample behavior is quite important. In particular, we want to investigate its limiting distribution. This is summarized in one of the most important in statistics, the central limit theorem (CLT).
- In MATH 320, we introduced the CLT and thought about it as a sum of random variables.
- **Central Limit Theorem**: Let $X_1, \ldots, X_n$ be independent random variables, all of which have the same probability distribution and thus the same mean μ and variance σ^2. If n is large, the sum

  $$S = X_1 + X_2 + \cdots + X_n$$

  will be approximately normal with mean $n\mu$ and variance $n\sigma^2$.

- Written succinctly: If $X_i \overset{iid}{\sim} f(x)$ with mean μ and variance σ^2, then

  $$S = \sum_{i=1}^{n} X_i \overset{approx}{\sim} \text{Normal}\left(n\mu, n\sigma^2\right) \quad \text{if } n \text{ is large}$$

- We used it to solve problems like this for example:

  Suppose the number of claims filed on for a particular policy follow a Poisson distribution with a mean of 2 claims per year and the company has a portfolio of 500 active policies this year, which are assumed to be independent.

  Find the distribution of the total number of filed claims for the entire portfolio.

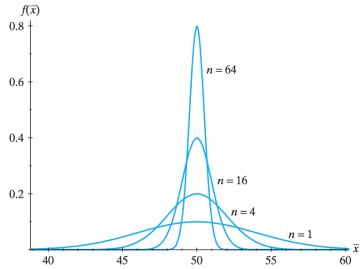
CLT - Different perspective

- Now we will think about the CLT from a convergence (in distribution) point of view.
- Build up to CLT / convergence in distribution.

Let $X_i \overset{iid}{\sim} \text{Normal}\,(\mu, \sigma^2)$ and let $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

1. For a fixed n:

2. As $n \to \infty$:



pdfs of means of samples from $N(50, 16)$

The variance decreases until all probability is a single point.

3. Suppose we "center the distribution" with $\bar{X}_n - \mu$:

Even with the location adjustment, the variance of $\bar{X}_n$ disappears when the sample size n increases (without bound).

4. We want to stop (or slow down) the "decay" of the variance, or we can think about this as spreading out the probability, so that when $n \to \infty$, $\bar{X}_n$ (and $\bar{X}_n - \mu$) does not converge to a constant, but rather distribution that still has some variation.

To do this, we multiply the quantity of interest by a factor of n:

Now this result doesn't converge to a constant because the variance doesn't depend on n and remains "in tact" when $n \to \infty$.

5. Then, we standardize the variance (adjusting the scale) by dividing by σ:

6. Lastly, it turns out that regardless of the distribution of X_i (so we are dropping the normal assumption), this result is always true!

   This is the CLT.

- Convergence in distribution
  - Definition: A sequence of random variables, $Y_1, Y_2, \ldots,$ **converges in distribution** to a random variable Y if

$$\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$$

    at all points y where $F_Y(y)$ is continuous (notation: $Y_n \xrightarrow{d} Y$).

  - Although we talk of a sequence of random variables converging in distribution, it is really **the cdfs that converge, not the random variable (or statistic)**.

    So for the CLT, we technically have:

$$\lim_{n\to\infty} \qquad F_{Y_n}(y) \qquad = \qquad F_Y(y)$$

Restating CLT

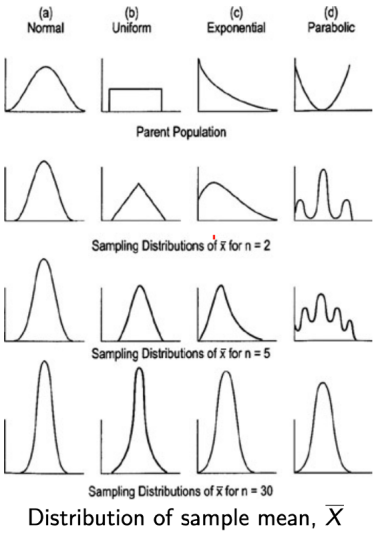
- **Theoretical result**

  Central Limit Theorem: Let $X_i \overset{iid}{\sim} f(x)$ with $E(X) = \mu$ and $V(X) = \sigma^2 > 0$. Then the distribution of

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0, 1) \quad \text{as } n \to \infty$$

- **In practice**

  This means for **any** random variable X with $E(X) = \mu$ and $V(X) = \sigma^2 > 0$, as n gets larger the distribution of $W = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ can be more closely approximated by the standard normal distribution.



- Using this practical result
  - Needed theorem: If $Z \sim \mathrm{N}(0,1)$, and μ and $\sigma > 0$ are constants, then

$$X = \sigma Z + \mu \sim \mathrm{N}(\mu, \sigma^2)$$

    Proof:

  - Results

(a) $\frac{\sigma}{\sqrt{n}} W + \mu = \bar{X}$ can be approximated by
$\frac{\sigma}{\sqrt{n}} Z + \mu \sim \text{Normal}\,(\mu, \frac{\sigma^2}{n})$ for "large" n.

(b) $n\bar{X} = X_1 + \ldots + X_n = S$ can be approximated by
$(\sigma\sqrt{n})Z + n\mu \sim \text{Normal}\,(n\mu, n\sigma^2)$ for "large" n.

- How large must n be?
  - Although CLT gives us a useful general approximation, we have no automatic way of knowing how good the approximation is in general. In fact, the goodness of the approximation is a function of the original distribution, and so much be checked case by case.
  - The more the distribution of X (population distribution) is "like" a normal distribution (symmetric, unimodal, continuous, etc.), the smaller the n needed for $\bar{X}$ to be approximated well by a normal distribution.
  - **The rule $n \geq 30$ is a lie!!** But for "school" purposes, we can just use this rule of thumb as our check.
  - With the current availability of cheap, plentiful computing power, the importance of approximation like the CLT is somewhat lessened. However, despite its limitation, it is still a marvelous result.

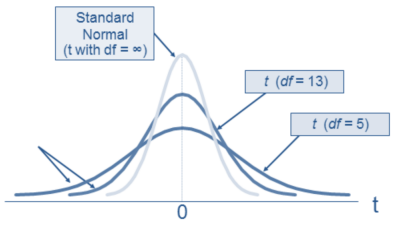
Examples

1. Let $\bar{X}$ be the mean of a random sample of size 36 from Exp $(\lambda = 1/3)$. Approximate $P(2.5 \leq \bar{X} \leq 4)$.

2. Let $X_1, \ldots, X_{20}$ denote a random sample of size 20 from continuous Uniform $(2, 8)$. If $S = X_1 + \ldots + X_{20}$, approximate $P(S < 95)$.

t, Z, and the CLT

- Previously, we learned the following
  1. If $X_1, \ldots, X_n$ are a random sample for a $N(\mu, \sigma^2)$, we know that the quantity
  $$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
  2. (Building on 1.) However, if σ is unknown, we substitute S then
  $$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$
  3. (Building on 2.) As $n \to \infty$, $t_{n-1} \to Z$



  4. If $X_1, \ldots, X_n$ are **not normal** random variables, when the sample size is large
  $$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

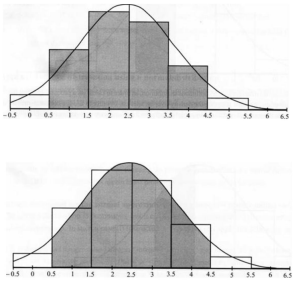
Approximations for discrete distributions

Continuity correction

- Motivation: Now we are going to use the CLT as an approximation tool when sampling from **discrete distributions**.

  Specifically, we will discuss a way to improve our approximations to account for the discrepancy created from using a continuous distribution / probability methods (integral to calculate area under curve) on originally discrete distributions.

- This is called the (half unit) **continuity correction** and is demonstrated below.



- Estimate the following probabilities using the continuity correction:

  (a) $P(1 \leq X \leq 4) \approx$

  (b) $P(X = 2) \approx$

  (c) $P(1 \leq X < 4) =$

  (d) $P(X > 2.5) =$

- In general, if X is the original discrete random variable of interest, S is the corresponding normal random variable based on the CLT, and a, b are some integers ($a \leq b$), then we can summarize the adjustments for the **continuity correction** with:

$$P(X = a) = P(a - 0.5 \leq S \leq a + 0.5)$$

$$P(a \leq X \leq b) = P(a - 0.5 \leq S \leq b + 0.5)$$

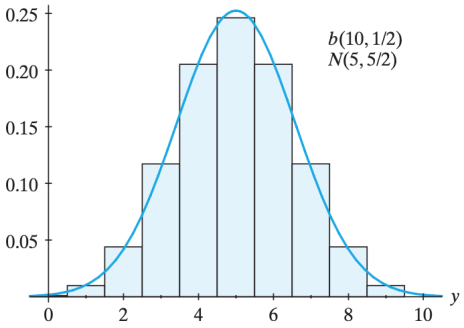
  Just need to take care to decide if we are want to include or exclude a or b (so can rewrite strict inequalities $< >$ as inclusive $\leq \geq$ and then use rule).

Normal approximation to the binomial distribution

- The most common scenario when applying the normal approximation is to the binomial distribution.
- Recall if $X \sim \text{Binomial}\,(n, p)$

- This means for "large n", $X =$

  "Rule of thumb" is that n is sufficiently large if $np \geq 5$ and $n(1-p) \geq 5$.
  We will discuss reasoning behind this after some examples.



- Examples→ Good use of the normal approximation

  1. Suppose that a multiple choice exam has 40 questions, each with 5 possible answers. A student feels that he has a probability of 0.55 of getting any particular question correct, with independence from one question to another.

     Approximate the probability of the student getting at least 25 correct.

     (a) With continuity correction:

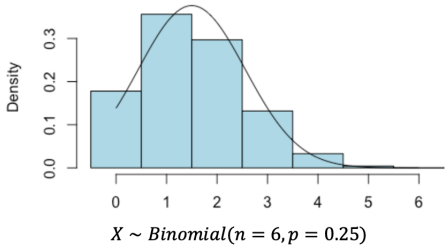
     (b) Without continuity correction:

     (c) Exact answer using binomial distribution:

- Bad example: Suppose we change the scenario in example 1 so that there are only 20 questions and the probability of getting any particular question correct is now 0.10.

  Compare the approximate answer and exact answer for $P(X \geq 3)$.

- Why is the approximation bad??
  - Lets take a look at the histogram and overlaid normal pdf for a similar scenario:



  - This illustrates the mismatch between the skewed probability histogram for and the symmetric pdf of the normal distribution. In order to do a good job of approximating the binomial distribution, the normal curve must have the bulk of its own distribution between legitimate outcomes for the Binomial distribution $[0, n]$.
  - How do we apply / check this: Based on the empirical rule, the central 95% of any normal distribution lies within two standard deviations of its mean.

  - Thus, as long as we ensure that _____, the normal approximation to the a binomial distribution will be good.

    Contextually, this condition means that we must expect (expected value) the number of success (np) and failures (nq) to be at least 5.

- This relationship between a large sample binomial and normal is important for confidence intervals and hypothesis tests of population proportions which we will cover next.

Normal approximation to the Poisson distribution

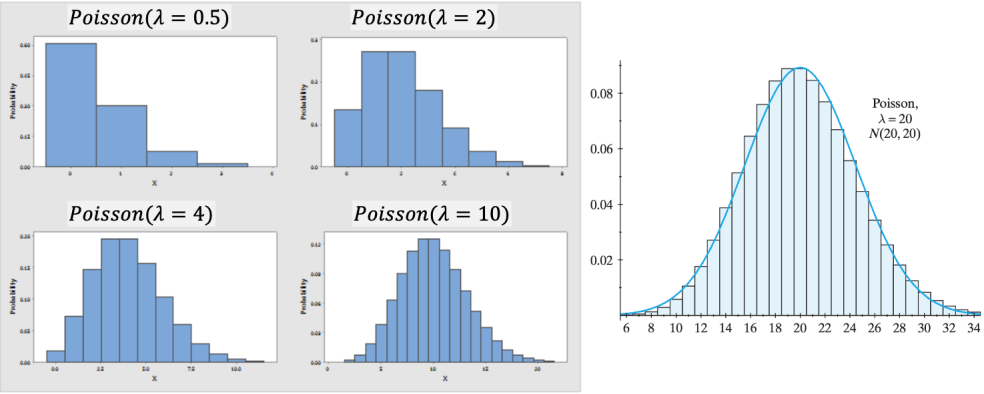
- A Poisson distribution with large enough mean can also be approximated with the use of a normal distribution.

  Let $X \sim \text{Poisson}(\lambda)$, with $E(X) = V(X) = \lambda$, where $\lambda = 1, 2, \ldots$ (in general, we just need $\lambda > 0$, but for demonstration lets assume λ is a positive integer).

- We can rewrite X as a sum of Poisson random variables:

- This means for "large n", $X =$

  "Rule of thumb" is that n is sufficiently large if $\lambda \geq 10$ (doesn't need to be an integer).



Poisson probability histograms

- Example

  Let X equal the number of alpha particles emitted by barium-133 per second and counted by a Geiger counter. Assume that $X \sim \text{Poisson}(\lambda = 49)$.

  Approximate $P(45 \leq X < 60)$.

Summary of normal approximation to the binomial and Poisson distributions

- Suppose n is large and $a = 0, 1, \ldots, n$
  (a) CLT:

  If $X \sim \text{Binomial}\,(n, p) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = np, \sigma = \sqrt{npq})$

  If $X \sim \text{Poisson}\,(\lambda) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = \lambda, \sigma = \sqrt{\lambda})$

  (b) Continuity correction:

  $P(X \leq a) \approx \text{Normalcdf(lower} = 0, \text{upper} = a + 0.5)$

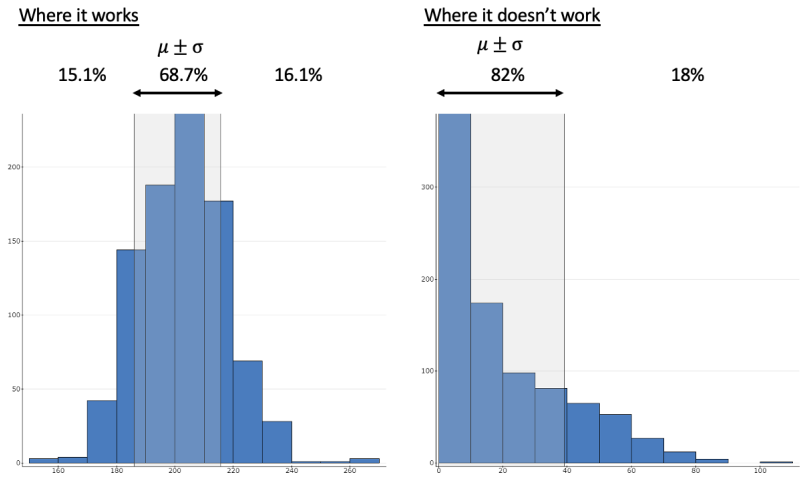
  $P(X < a) \approx \text{Normalcdf(lower} = 0, \text{upper} = a - 0.5)$

- Final note: In practice, if you have technology / software, just compute discrete probabilities exactly. However, it is important to learn how to apply the central limit theorem.

Central interval probabilities

Empirical rule

- Motivation: Because the normal distribution can be used in so many scenarios due to the CLT, there are common generalizations that are made about **central interval** probabilities for distributions that are approximately bell-shaped.
- First, lets calculate these exactly for the standard normal curve. These will of course apply to any normal distribution X with mean μ and standard deviation σ because we can standardize to get Z.
  1. $P(-1 \leq Z \leq 1) =$
  2. $P(|Z| \leq 2) =$
  3. $P(|Z| \leq 3) =$
- Not all data is exactly normally distributed of course, but because of the CLT many distributions can be approximated by a normal distribution. So we can use the exact probabilities above to make generalizations about these distributions that have a similar shape.
- The **empirical rule** states that for approximately normal distribution:
  1. Approximately _____ of data falls within _____ standard deviation of the mean.
  2. Approximately _____ of data falls within _____ standard deviations of the mean.
  3. Approximately _____ (nearly all) of data falls within _____ standard deviations of the mean.



- Example: Suppose that the scores on an achievement test are known to have, approximately, a normal distribution with mean $\mu = 64$ and standard deviation $\sigma = 10$.

  (a) Find the scores probability scores are between 54 and 74.

  (b) Find which two values lies the central 95%?

  (c) Find the percent of scores above 94.

- Thus, knowledge of the mean and the standard deviation gives us a fairly good picture of the frequency distribution of scores when the bell-shape is present (or assumed).

Lecture 6 – Confidence Intervals

MATH 321: Mathematical Statistics

Lecture 6: Confidence Intervals

Chapter 7: Interval Estimation (7.1 - 7.4)

Introduction

Estimating parameters

- Point estimates
  - Using a point estimator $\hat{\theta}$ to estimate a parameter θ.
  - It is our single best guess.
  - Usually the point estimates do not equal the parameter because of sampling variability.
- Interval estimates
  - Give a range for what we think the population parameter is.
  - Takes into account sampling variability.

Constructing confidence intervals

Interval estimators / confidence intervals

- Definition: An **interval estimator** or **confidence interval** is a rule specifying the method for using the sample data to calculate two numbers that form the endpoints of the interval.

$$[L(\mathbf{X}), U(\mathbf{X})]$$

  Once $\mathbf{X} = \mathbf{x}$ is observed, the **interval estimate** is then $L(\mathbf{x})$ and $U(\mathbf{x})$ and the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made.

- Ideally, the resulting interval will have two properties:
  1. It will contain the target parameter θ.
  2. It will be relatively narrow.

- Notes about properties:
  - The endpoints $L(\mathbf{X})$ and $U(\mathbf{X})$ (called the **lower and upper confidence limits**) of the interval are functions of the sample, which means they will vary randomly from sample to sample.

    Thus, the length and location of the interval are random quantities.

  - Because of this, we cannot be certain that the (fixed) target parameter θ will fall between the endpoints of any single interval calculated from a single sample.

    This being the case, our objective is to find an interval estimator capable of generating narrow intervals that have a high probability of enclosing θ.

- The probability that a (random) confidence interval will enclose θ (a fixed quantity) is called the **confidence coefficient**:

$$P\big(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\big) = 1 - \alpha$$

  where α is called the **significance level**.

  Thus $[L(\mathbf{X}), U(\mathbf{X})]$ is called a **$100(1-\alpha)\%$ confidence interval for θ**.

Constructing confidence intervals

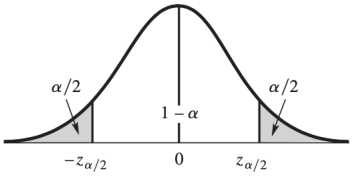
- All of the confidence intervals we will build start from this general setup and use properties of normal distributions or the central limit theorem to get the final interval of interest.
- Setup: Let $\hat{\theta}$ be an unbiased point estimator for parameter θ and $\sigma_{\hat{\theta}}$ be the standard deviation of the sampling distribution of $\hat{\theta}$ (this is often called the **standard error** of $\hat{\theta}$).

  Based on the scenario, if $\hat{\theta}$ is normally distributed, the quantity

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim \text{Normal}\,(0, 1)$$

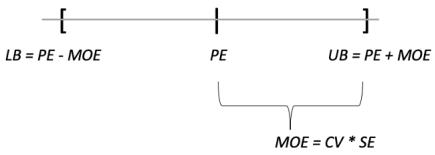
- Then to find a confidence interval for θ that possesses a confidence coefficient equal to $1 - \alpha$, we just need to select two values in the tails of this distribution, $-z_{\alpha/2}$ and $z_{\alpha/2}$ (these are called **critical values**), such that $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$.

  Then because we seek an interval estimator for θ, we just have to substitute in $\frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ for Z and rearrange to isolate θ in the middle.



- Thus, we can summarize any (two-sided) confidence interval with

$$\text{CI} = \text{Point Estimate} \pm \text{Margin of Error}$$



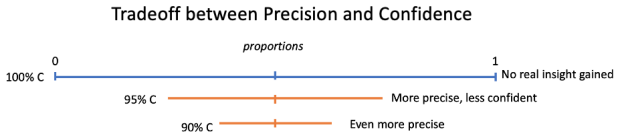
  - Point Estimate (PE) is the best guess; at the center of the interval.
  - Margin of Error (MOE) = Critical Value (CV) $\times$ Standard Error (SE).

    SE (standard deviation of the statistic) measures sampling error.

    % Confident is determined by confidence level set and incorporated via the critical value (CV).

- Recall the two goals of confidence intervals: (1) capture the parameter of interest and (2) be precise (smaller MOE = narrower interval).
  - The location (center) of the interval is determined by the _____
  - The precision (MOE) is determined by the _____ (via the standard error) AND by the _____ (via the confidence level).

- All else equal, here is how the researcher can affect the precision of intervals:
  - Larger sample size $n \rightarrow$ _____ interval
  - More confident $\rightarrow$ _____ interval



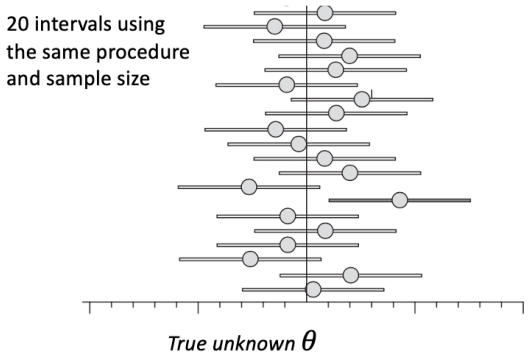
Interpreting confidence intervals

- Interpretation
  - General Structure

    I am % confident that the true/population parameter + context is between lower bound and upper bound.
  - Example: Suppose 95% CI = [24, 30]

    I am 95% confident that the true (population) mean of all Indiana ACT test scores is between 24 and 30.
  - When interpreting CIs: Make sure to mention what θ represents in context and keep in mind that the interval is giving us the range of plausible values for θ.

- Confidence coefficient
  - "95% confident": This tells us that in repeated sampling, approximately 95% of all intervals of the form $\hat{\theta} \pm 1.96\,\sigma_{\hat{\theta}}$ include θ.



  - Be careful with using "confidence" and "probability" interchangeably.

    AFTER collecting data: BEFORE collecting data:

  - For a particular sample, this interval either does or does not contain the parameter θ, but we never know.
  - However, we are "95% confident" that the interval contains the parameter because the procedure that generated it yields intervals that do capture the true parameter in approximately 95% of the time that the procedure is used.

Confidence intervals for proportions

Introduction

- Often we want to estimate population proportions or the difference in proportions. For example
  - Proportion of voters in favor of an issue, proportion of students that graduate college, proportion of the population in a certain interval of values (success / fail perspective on a numeric variable), etc.
  - Difference in polling position for two candidates, difference in graduation rates for students involved in clubs vs not, etc.
- We can compute confidence intervals for one proportion or the difference in two proportions.

Confidence intervals for one proportion

- Setup: If we observe n independent Bernoulli trials, each with success probability p, then

  Thus X represents the number of successes in the n trials.
- Now we are interested in the parameter $\theta =$

  The unbiased estimator is the sample proportion _____
- Main result to form the interval (just need to meet conditions):

  Conditions:

  1) Normal approximation to the binomial 2) CLT for Bernoulli mean

- Thus we can construct the approximate $100(1-\alpha)\%$ confidence interval for p with

Note that the unknown parameter p appears in both endpoints of the interval, so we do the obvious thing and substitute $\hat{p}$ into the standard error $\sigma_{\hat{p}}$.

- Example: Let p equal the proportion of triathletes who suffered an overuse injury during the past year. Out of 330 triathletes who responded to a survey, 167 indicated that they suffered such an injury during the past year.

  (a) Use these data to give a point estimate of p and to find an approximate 90% confidence interval for p.

  (b) Do you think that the 330 triathletes who responded to the survey may be considered a random sample from the population of triathletes?

  This is an example of _____ (aka voluntary response bias). There is a whole branch of statistics related to surveys and how to collect data while minimizing bias in the sample.

Confidence intervals for difference of two proportions

- Setup: Same setup as for one proportion, now just two samples:

- Now we are interested in the parameter $\theta =$

  The unbiased estimator is difference in sample proportions _____

- Just need to check the conditions first before constructing the desired interval.
- Forming the interval $\rightarrow$ Conditions:

  Result:

- Thus we can construct the $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ with

Again, we will estimate the standard error using the respective sample proportions.

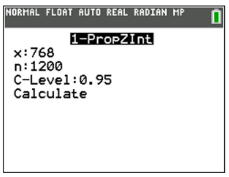
- Example: Two detergents were independently tested for their ability to remove stains of a certain type. An inspector judged the first detergent to be successful on 83 out of 100 independent trials and the second one to be successful on 42 out of 79 independent trials.

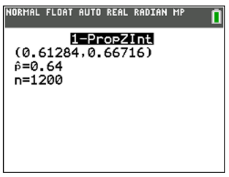
  Find a 98% confidence interval for the difference in the probability in removing stains of the two detergents and state the conclusion.

Calculator session

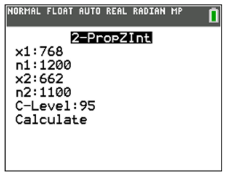
STAT > TESTS >

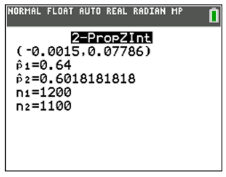
One proportion CI: p





Two proportion CI: $p_1 - p_2$





Confidence intervals for means

Confidence intervals for means

- Now we are interested in the parameter $\theta =$

  All we have to do is use the unbiased estimator for _____ and the correct standard error $\sigma_{\hat{\theta}}$, then apply those to the final confidence interval shown at the beginning.

- Variables that affect the formation of our confidence intervals:
  - Sample size (large or small)
  - Population distribution X (normal or not normal)
  - Population variance σ^2 (known or unknown)
- We will simplify the scenarios and just think about large or small samples, and add notes about when the intervals are approximate or exact.

Large sample confidence intervals

- Suppose $X_1, \ldots, X_n$ are a random sample with **"large" n** from some distribution X with unknown variance σ^2.

- How large must n be / goodness of the approximation:

  Because of the unknowns (distribution and variance), we have approximately $100(1-\alpha)\%$ confidence intervals. As more assumptions are introduced, confidence coefficients for the intervals become more exact (i.e. closer to $100(1-\alpha)\%$ level).

  1. (Least best scenario) If X is badly skewed or has outliers, then prefer to have even larger sample sizes like $n \geq 50$, and even that may not produce good results.
  2. (Most likely in practice) If starting from Normal or like Normal (unimodal, symmetric, and continuous), then need $n \geq 30$ for the CLT to work with the unknown σ^2.
  3. (Best case scenario) If assume Normal and known variance σ^2, then this procedure even works for $n << 30$.

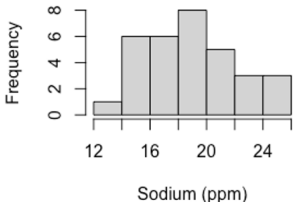
- Examples:

  1. Example: Let X equal the life of a a 60-watt light bulb marketed by a certain manufacturer with $X \sim \text{Normal}\,(\mu, \sigma^2 = 1296)$. Suppose a random sample of size 27 from this distribution yields $\bar{x} = 1478$.

     Construct 90% and 95% confidence intervals for $E(X) = \mu$.

  2. Lake Macatawa, an inlet lake on the east side of Lake Michigan, is divided into an east basin and a west basin. To measure the effect on the lake of salting city streets in the winter, students took 32 random samples of water from the west basin and measured the amount of sodium in parts per million in order to make a statistical inference about the unknown mean μ. They obtained the following data:

13.0	18.5	16.4	14.8	19.4	17.3	23.2	24.9
20.8	19.3	18.8	23.1	15.2	19.9	19.1	18.1
25.1	16.8	20.4	17.4	25.2	23.1	15.3	19.4
16.0	21.7	15.2	21.3	21.5	16.8	15.6	17.6



     Construct a 95% confidence interval for μ the mean amount of sodium in the west basin.

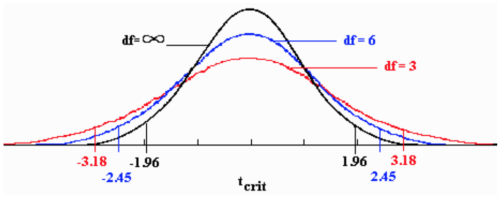
  3. Example: Let X be the amount of orange juice (in grams per day) consumed by an American. Suppose $V(X) = \sigma^2 = 96$. To estimate μ, an orange growers' association took a random sample of $n = 576$ and found $\bar{x} = 133$.

     Construct a 98% confidence interval for μ.

Small sample confidence intervals

- Suppose $X_1, \ldots, X_n$ are a random sample with **"small" n** from **Normal**(μ, σ^2), with **unknown variance σ^2**.

- Effect of converting to a t-interval
  - All else equal, t-intervals are wider than the corresponding Z-intervals because we are approximating σ with s (estimating another parameter in addition to μ).



  - However, the length of t-intervals are very much dependent on the value of the observed sample standard deviation s.

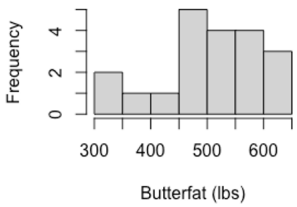
    If the observed s is smaller than σ, we can get a narrower using a t-interval compared to a Z-interval. But on average, $\bar{x} \pm z_{\alpha/2}\,(\sigma/\sqrt{n})$ is the shorter of the two confidence intervals.
  - When n gets larger ($n \geq 30$), then $t_{n-1} \approx Z$, which is why we can just use the Z critical values and the approximate Z-interval.

- What if data is not Normal?
  - Generally, this procedure works well when underlying distribution is symmetric, unimodal, and continuous and is still quite good (i.e. it is robust) for many non-normal distributions.
  - However it is not good (i.e. dangerous to use) if the distribution is highly skewed. If this is the case, safer to use certain nonparametric methods for finding a confidence interval for the median of the distribution (we will not cover this).

- Examples:

  1. Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves. Assume that the distribution of $X \sim \text{Normal}(\mu, \sigma^2)$. To estimate μ, a farmer measured the butterfat production for $n = 20$ cows and obtained the following data:

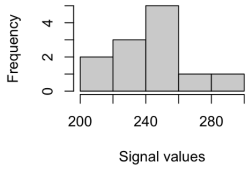
481	537	513	583	453	510	570	500	457	555
618	327	350	643	499	421	505	637	599	392



Construct a 97% confidence interval for μ.

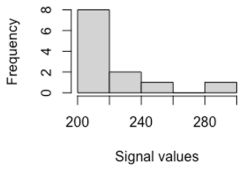
  2. In nuclear physics, detectors are often used to measure the energy of a particle. To calibrate a detector, particles of known energy are directed into it. The values of signals from 12 different detectors, for the same energy, are shown below.

260	216	259	206	265	284
232	250	225	242	240	252



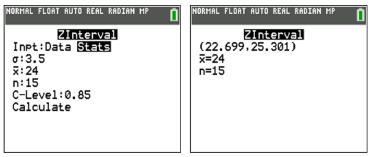
Construct a 95% confidence interval for μ.

  3. Continuing example: Suppose the data looked like this (still $n = 12$):

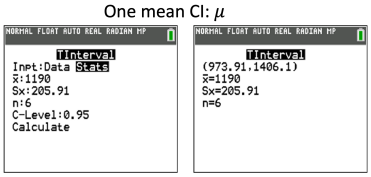


Calculator session

STAT > TESTS > One mean CI: μ



One mean CI: μ



One-sided confidence intervals

- Can also create one-sided confidence intervals if interested in the probability θ is larger or smaller than a certain number.
- By the same arguments as shown above, we can determine that **$100(1-\alpha)\%$ one-sided confidence interval** are given by

  Lower bound Upper bound

- Suppose that we compute both a $100(1-\alpha)\%$ lower bound and a $100(1-\alpha)\%$ upper bound for θ. We then decide to use both of these bounds to form a CI for θ.

  What will be the confidence coefficient of this interval?

- For one mean $\theta = \mu$, we will use the same criteria discussed above to determine Z vs t (knowing if exact vs approximate), then just use either the lower or upper bound interval if interested in a one-sided CI.
- Example: Using the Lake Macatawa data ($n = 32$ with unknown distribution and unknown variance), construct a 95% lower CI for μ and a 95% upper CI for μ.

  Then combine to form a two-sided interval and compare to the 95% two-sided interval found earlier.

  Lower bound Upper bound Combined

Confidence intervals for the difference of two means

Introduction

- Often we want to compare means of two different populations. For example, compare: average heights of male vs females for a species, average GPA of students in different school districts, mean response for two different treatments in an experiment, etc.
- We can compute confidence intervals for difference in means.

Confidence intervals for difference in means

- Now we are interested in the parameter $\theta =$

  The unbiased estimator is _____

  Again, we just need to us the correct standard error _____
- Variables that affect the formation of our confidence intervals:
  - Independent or dependent samples
  - Sample sizes n_1 and n_2 (large or small)
  - Population distributions X_1 and X_2 (normal or not normal)
  - Population variances σ_1^2 and σ_2^2 (known or unknown and ratio of variances)
- Similar logic (large vs small sample) can be used for two samples with regards to the form of the interval once we decide on independent vs dependent samples.

Independent, large sample confidence intervals

- Suppose we have **independent, large** random samples from some distributions X_1 and X_2 with sizes n_1 and n_2, respectively.

- Again, we need larger sample sizes with more unknowns in order to still have good approximations. Additionally, if we are starting from Normal or the variances are assumed known, then the approximations are better (or exact).

- Examples:

  1. A ecological study was conducted to compare rates of growth of trees at two sites by measuring leaf lengths of trees planted the previous year. It is known that the lengths of these leaves are normally distributed regardless of the conditions in which they grow and the variance of the lengths is $\sigma_1^2 = 1.69$ cm^2 at site 1 and $\sigma_1^2 = 2$ cm^2 at site 2. Two independent random samples of leaf lengths from the two sites are observed as below:

	Site Leaf Length (cm)							
Site 1	5.18	1.48	1.82	2.35	3.04	5.49	1.03	4.04
Site 2	7.45	7.27	4.06	5.75	3.31	8.19	6.4	

     Construct an 80% CI for $\mu_1 - \mu_2$ where μ_1, μ_2 are the mean leaf lengths of sites 1 and sites 2, respectively.

     – When interpreting confidence intervals for the difference in parameters, there are three scenarios for intervals:

       Below zero Contains zero Above zero

     – For example, suppose interval is $[-1, 3]$. This contains zero and we would conclude there is no difference in θ_1 and θ_2; however, keep in mind that 3 is also a "believable" value for the difference.

  2. A comparison of the durability of two types of automobile tires was obtained by road testing samples of $n_1 = n_2 = 100$ tires of each type. The number of miles until wear-out was recorded, where wear-out was defined as the number of miles until the amount of remaining tread reached a pre-specified small value. The measurements for the two types of tires were obtained independently, and the following means and variances were computed (in miles):

     $$\bar{x}_1 = 26,400, \quad s_1^2 = 1,440,000 \qquad \text{and} \qquad \bar{x}_2 = 25,100, \quad s_2^2 = 1,960,000$$

     Construct a 90% CI for $\mu_1 - \mu_2$.

Independent, small sample confidence intervals

- Suppose we have **independent, small** random samples from $X_1 \sim \textbf{Normal}\,(\mu_1, \sigma_1^2)$ and $X_2 \sim \textbf{Normal}\,(\mu_2, \sigma_2^2)$ with n_1 and n_2, and with **unknown common variance** $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$.

  - The standard error for this comes from the usual unbiased estimator of the common variance σ^2, which is obtained by pooling the sample data to obtain the pooled estimator S_p^2. This is just a weighted average of S_1^2 and S_2^2 with larger weight given to the sample variance associated with the larger sample size.
  - Proof:

    From above, we know

    $$Z = \frac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim \text{Normal}\,(0, 1)$$

    and from earlier theorems

    $$\frac{(n_1 - 1)}{\sigma^2} S_1^2 \sim \chi^2{}_{n_1 - 1} \qquad \text{and} \qquad \frac{(n_2 - 1)}{\sigma^2} S_2^2 \sim \chi^2{}_{n_2 - 1}$$

    Thus,

    $$U = \frac{(n_1 - 1)}{\sigma^2} S_1^2 + \frac{(n_2 - 1)}{\sigma^2} S_2^2 \quad \sim$$

    Combining all of this, we can form the following

– Example: Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where σ^2 is unknown. If a random sample of $n_1 = 9$ students from large high schools and a random sample from $n_2 = 15$ small high school yielded

$$\bar{x}_1 = 81.31, \quad s_1^2 = 60.76 \qquad \text{and} \qquad \bar{x}_2 = 78.61, \quad s_2^2 = 48.24$$

Construct a 95% CI for $\mu_1 - \mu_2$.

- If we don't assume a common variance, we can still do the above procedure if the sample variances are close enough.

  As a rule of thumb, if the ratio of S_1^2/S_2^2 is between 0.5 and 2 (i.e., if one variance is no more than double the other), then we can use the pooled formula.

- Now we have all the assumptions from above scenario, (independent, small samples, both Normal with unknown variances), except we **cannot assume a common variance** AND they are drastically different. There are two cases

  1. We **do know the ratio of variances $\boldsymbol{\sigma_1^2/\sigma_2^2 = d}$.**

     Can still construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ using a modified s_p (will not cover this one).

  2. We **do not know the ratio of the variances** and yet suspect that the unknown σ_1^2 and σ_2^2 **differ by a a lot**.

     Can still construct a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ using a similar interval to one with large samples and unknown variances, except we use t-critical values (due to the unknown variances) with adjusted degrees of freedom to provide a larger MOE (will not cover this one either).

Dependent samples confidence intervals

- All of the previous intervals required independent samples as one of the assumptions. This is often not the case in practice.
- Independent vs dependent samples
  – Independent: Groups are unrelated, no connection, no relationship

    This is often not the case in practice, sometimes by design.

  – Dependent: Groups have some relationship between one another, can link the two; PAIRS

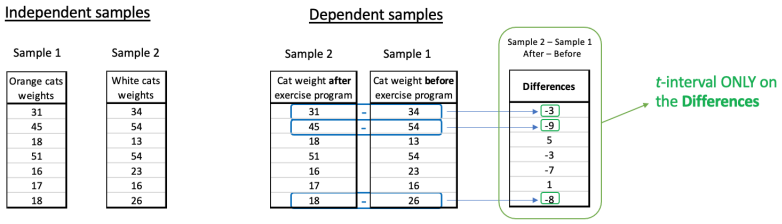
- If samples are dependent, they can be dependent in one of two ways.

  The interval that we construct is that same for both, but nonetheless it is important to know the structure of our data and how it was obtained.

  - **Paired**: Two values from the SAME subject.
  - **Matched**: Two values from DIFFERENT subjects connected in some way.
  - Examples) Determine if the following samples are independent or dependent (and matched or paired).

    1. Comparing the blood pressure of MATH 321 students before the final exam and after completing the final exam.
    2. Are brothers or sisters smarter? A researcher studied ACT scores of 8 brother and sister pairs.
    3. A study is conducted to see what effect a new drug has on dexterity. A random sample of 30 students is chosen. They are given a series of tasks to perform and a score reflecting their performance. A dose of the drug is given to the 30 students and they again perform similar tasks and are scored again.
    4. Looking to see if there is a difference in the price of the same Video Game Consoles at Target or Walmart.
    5. Seeing if the height of Faculty is shorter than the undergraduate population.

- Independent vs dependent strategy



- If X_1 and X_2 may be **dependent random variables**, then we cannot use the t-statistics and confidence intervals that we just developed, because they were based on the assumption of independence.

- **Matched-pair (dependent) t-interval for $\mu_1 - \mu_2$**

  Suppose we have **dependent** random samples from of size n from X_1 and X_2 (which can think of as ordered pairs $(X_{1,i}, X_{2,i})$).

  Let $D = X_1 - X_2$. This can be thought of as a random sample from $D \sim \text{Normal}\,(\mu_D, \sigma_D^2)$, where μ_D and σ_D^2 are the mean and variance of the difference in each pair.

  Then we can construct a $100(1-\alpha)\%$ for D in the same way as the as we did for one mean:

- Example: To compare the wearing of two types of automobile tires, A and B, a tire of type A and of type B are randomly mounted on the wheels of each of 8 automobiles. The automobiles are operated for a certain number of miles, and the amount of wear recorded for each tire below. Assuming the difference in wears of the tires are normally distributed, construct a 95% CI and a 95% lower-bounded CI for the difference in the mean wear of the two type of tires.

Car	1	2	3	4	5	6	7	8
Tire A	10.5	9.8	12.3	9.7	13.2	8.8	11	11.3
Tire B	10.4	9.6	12	9.3	12.8	8.3	10.4	10.6

Finding the minimum sample size

Motivation

- In statistical consulting, the first question frequently asked is, "How large should the sample size be to estimate a mean?"

  Determining the sample size is an important step when planning a study because of the following considerations:

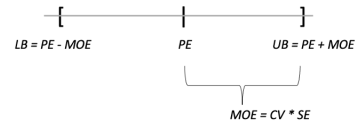
  - If n is too large, it is a waste of resources (studies are expensive, time and $$$).
  - If n is too small, they are less confident in the results (i.e. too imprecise); no real insight is gained.

- In the context of estimation, researchers want to figure out how large their sample needs to be to yield a confidence interval with a predetermined width.

  In doing so, they are controlling the precision!

Margin of error (MOE) revisited

- Recall: MOE is what you add and subtract from your point estimate to get your upper bound (UB) and lower bound (LB) of your confidence interval.



- If you are given an interval, your margin of error is the following:

$$MOE = \frac{UB - LB}{2} = \frac{Width}{2} \qquad \rightarrow \qquad Width = 2 \times MOE$$

- This is what we are controlling in the process of selecting the minimum sample size!

  For example, suppose a mathematics department wishes to evaluate a new method of teaching calculus with a computer. At the end of the course, the evaluation will be made on the basis of standard test, in which they would like to estimate μ, the mean score for students in the new class.

  In planning this course, they wish to determine how many students should take the course in order to be fairly confident that $\bar{x} \pm 1$ contains the unknown test mean μ.

- More formal definition: The **error in estimation ϵ** is the distance between an estimator and its target parameter. That is

  Typically, we are given a **maximum error in estimation**, which means we want the margin of error to be no more than ϵ (or "within" ϵ), less than is okay.

Finding minimum sample size

- The process for finding the minimum sample size for a given a maximum error in estimation ϵ is the same regardless of what type of interval we are using.

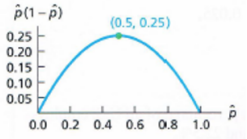
  Just start with the formula for Margin of Error and rearrange to solve for n.

- Here are the derivations / calculations for some of the different intervals that we have discussed. For each situation, we want the $100(1-\alpha)\%$ confidence interval for θ, $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$, to be no longer than that given by $\hat{\theta} \pm \epsilon$.
- One proportion

  - Again, we do not know the value p obviously and we cannot substitute $\hat{p}$ like before because we haven't collected data yet. So we have two options for specifying $p = p^*$:

    1. Set p^* based on previous research or experience.
    2. If no prior information is available, set $p^* = 0.5$.

       This results in the largest n for a specific MOE, so it is a safe (conservative) estimate. So to achieve a maximum error of estimate of at most ϵ, use the following:



- Example: The unemployment rate in a certain country has been about 8%. This rate has changed by small amount and economists wish to update their estimate of the unemployment rate p in order to make decisions about national policy. Find the sample size needed to achieve a maximum error of the estimate of

  1. $\epsilon = 0.001$ for a 95% CI for p

  2. $\epsilon = 0.01$ for a 99% CI for p

  3. $\epsilon = 0.01$ for a 95% CI for p

  4. $\epsilon = 0.01$ for a 95% CI for p, except assume now assume that we have no prior information about p.

- One mean

  - Researchers have to make an assumption about the value of σ in order to do sample size calculations, which can be tricky. And resulting estimates for minimum sample sizes can change drastically based on how much variability is in the process they are studying. So there are a few options:

    * Assume a value for σ^2.
    * Use the best approximation available such as an estimate s obtained from a previous sample.
    * Use an upper bound on σ^2 if available.
    * Use knowledge of the range of the measurements in the population.

- Examples

  1. (Continuing the math department example) Given past experience it is believed scores on such a common final are normally distributed with standard deviation of 15. Using $\bar{x}$ as an estimate, find the sample size needed to achieve a maximum error of the estimate of

     (a) $\epsilon = 1$ for a 95% CI for μ

     (b) $\epsilon = 2$ for a 95% CI for μ

     (c) $\epsilon = 2$ for a 90% CI for μ

     (d) $\epsilon = 2$ for a 90% CI for μ, except with $\sigma = 22.5$

  2. Continuing math example: Suppose test grades typically range between 60 and 95. Based on the empirical rule, 95% of data is between 2σ of the population mean μ. We can use this fact to get an approximate sample size, for say $\epsilon = 1$.

- Two means

  - If we make two simplifying assumptions, then we can get sample size estimates for this scenario as well (else it becomes like solving a system).

    Equal sample sizes: $n_1 = n_2 = n$ and equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

- Example: An experimenter wishes to compare the effectiveness of two methods of training. The selected participants are to be divided into two groups of equal size, the first receiving training method 1 and the second receiving training method 2. After training, each participant will a task and have their time recorded.

  The goal is to estimate the mean difference in times within 1 minute with 95% confidence, assume $\sigma_1^2 = \sigma_2^2 = 2$.

  Note we could use the range strategy to get an estimate of common σ if we did not have the assumption it equaled 2.

- Observations: When estimating μ with $\bar{x}$, all else equal:

  Larger margin of error $\epsilon \rightarrow$ _____ sample size n

  More confident (smaller α) $\epsilon \rightarrow$ _____ sample size n

  Larger variance $\sigma^2 \rightarrow$ _____ sample size n

Lecture 7 – Hypothesis Tests

MATH 321: Mathematical Statistics

Lecture 7: Hypothesis Tests

Chapter 8: Tests of Statistical Hypotheses (8.1 - 8.3)

Introduction

- Recall that the objective of statistics often is to make inferences about unknown population parameters based on information contained in sample data.

  These inferences are phrased in one of two ways:

  - As estimates of the respective parameters (point estimation / confidence intervals)
  - Or as tests of hypotheses about their values

- Hypothesis tests are essentially the scientific method viewed through statistics.
  - The scientist poses a hypothesis concerning one or more population parameters (e.g. that they equal specified values).
  - Then samples the population and compares observations with the hypothesis.
  - If the observations disagree with the hypothesis, the scientist rejects it.

    If not, the scientist concludes either that the hypothesis is true or that the sample did not detect the difference between the real and hypothesized values of the population parameters.

- Hypothesis tests are done in almost all fields where we are testing theory against observation. Examples:
  - A medical researcher may hypothesize that a new drug is more effective than another in combating a disease.

    To test her hypothesis, she randomly selects patients infected with the disease and randomly divides them into two groups: Group A gets the current drug and Group B gets the new drug.

    Then, based on the number of patients in each group who recover from the disease, the researcher must decide whether the new drug is more effective than the old.
  - A quality control engineer may hypothesize that a new assembly method produces only 5% defective items.
  - An educator may claim that two methods of teaching reading are equally effective.

- Statistics and what we will learn is what measures to take on the sample, how do make the decision of accept vs reject, what are the probabilities we made the correct / incorrect decision, etc.

Elements of a statistical test

Hypothesis test overview

- Definition: A **hypothesis testing procedure or hypothesis test** is a rule that specifies
  - For which sample value the decision is made to reject H_0 in favor of H_A.
  - For which sample value the decision is made to "not reject" H_0 in favor of H_A.

- Any statistical test of hypotheses works in exactly the same way and is composed of the same essential elements.
  1. Null hypothesis H_0 and Alternative hypothesis H_A
  2. Test Statistic TS and Rejection Region RR
  3. Conclusion

- Example setup: Let X equal the breaking strength of a steel bar. A company uses process I to manufacture steel bars and it is known that under process I, $X \sim \text{Normal}\,(\mu = 50, \sigma^2 = 36)$.

  The company wishes to test a new process, process II, and it is hoped that under process II $X \sim \text{Normal}\,(\mu = 55, \sigma^2 = 36)$.

- Hypotheses
  - Definition: A **hypothesis** is a statement about a population parameter.
  - The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.
    - The **Null hypothesis H_0** is an assumption about θ that is assumed to be _____.
    - The **Alternative hypothesis H_A** (or H_1, also called research hypothesis) is the _____ of the null hypothesis. The goal is generally to obtain evidence in favor of this.
  - Continuing steel bar example:
  - These are called **simple hypotheses** because each completely specifies the distribution of X. Could test H_0 against a **composite hypotheses**, which contains many possible alternative distributions.
  - In general, we have the following hypotheses:

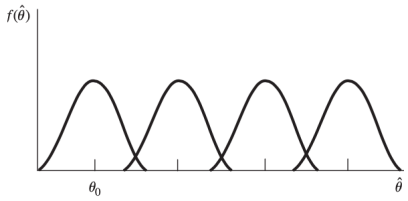
- Examples: (1) Define the parameter of interest and (2) state the null and alternative hypotheses and the directionality of the test (two-tailed, left-tailed or right-tailed) for the following scenarios:

  (a) A company reports that last year 40% of their reports in accounting were on time. From a random sample this year, they want to know if that proportion has changed.

  (b) Last year, 42% of the employees enrolled in at least one wellness class at the company's site. Using a survey from randomly selected employees, they want to know if a greater percentage is planning to take a wellness class this year.

  (c) There are two political candidates, and one wants to know from the recent polls if she is going to win a majority of votes in next week's election.

- Test statistic and rejection region
  - These are all about distributions of estimators based on assumptions from the hypotheses.

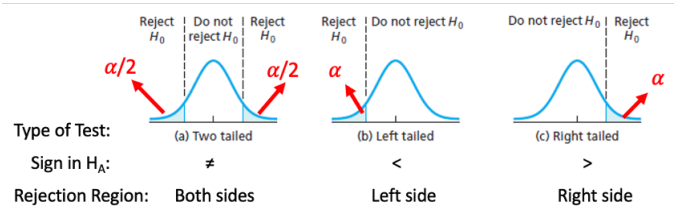


  - For example, for a right-tailed test
    - If $\hat{\theta}$ is close to θ_0, it seems reasonable to accept H_0.
    - If in reality $\theta > \theta_0$, then $\hat{\theta}$ is more likely to be large.

      Consequently, large values of $\hat{\theta}$ (relative to θ_0) favor rejection of $H_0 : \theta = \theta_0$ and acceptance of $H_A : \theta > \theta_0$.

- Simply stated, we have to determine when there is or is not enough _____ against the _____ based on our _____.

  In other words, which tail do we make the conclusion of reject, which comes from the direction in the H_A, and how large is the area.



- The hypothesis test is specified in terms of the test statistic and the corresponding rejection region.
  - **Test statistic (TS)** is a function of the sample $W(X_1, \ldots, X_n)$, think of this as the point estimator $\hat{\theta}$.
  - **Rejection Region (RR)** (or critical region) is the subset of the sample space (range of sample) for which H_0 will be rejected. RR is defined with the TS (these two parts are always together).
- Once these are defined, hypothesis tests are really easy; we then just observe data and see where it falls.
- In general, we can state the rejection region as

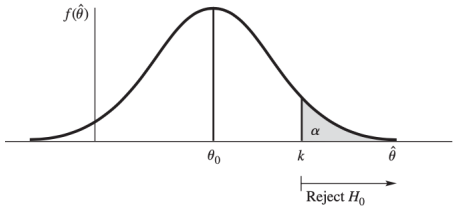
  $RR = \{\text{set of } (x_1, \ldots, x_n) \text{ such that (some math statement about TS } W(X_1, \ldots, X_n)\}$

- Continuing steel bar example: Suppose $n = 16$ bars were tested, intuitively we could choose a RR where larger values lead to rejecting H_0, say $RR = \{\bar{x} : \bar{x} \geq 53\}$.

- But how did we choose the value of k? More generally, how can we find some objective criteria for deciding which value of k specifies a good rejection region of the form $\{\bar{x} \geq k\}$?

- Significance level
  - The **significance level** α of the test is what determines how large the RR is and represents the probability of rejecting the null hypothesis.

    The actual value of k is chosen by fixing this and finding k accordingly.
  - Recall under the null hypothesis, the distribution of $\hat{\theta}$ is known. So we can find k such that (for example with a right-tailed test):



- The significance level is chosen before running the test. Setups will say something similar to: "Determine if there is enough evidence at the 5% significance level."

Building hypothesis tests

Hypothesis test setup

- Just like with confidence intervals, all of the hypothesis tests we will build start from this general setup and use properties of normal distributions or the central limit theorem to get the test statistic and rejection region of interest.

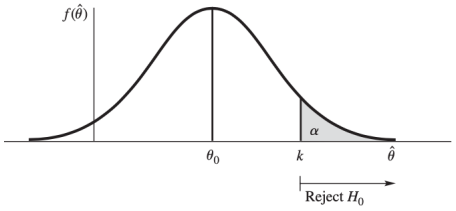
- For hypothesis tests, we will consider same variables that affect the formation of our confidence intervals:
  - Independent or dependent samples
  - Sample sizes n_1 and n_2 (large or small)
  - Population distributions X_1 and X_2 (normal or not normal)
  - Population variances σ_1^2 and σ_2^2 (known or unknown and ratio of variances)

Large sample tests

- Setup: Suppose we want to test a set of hypotheses concerning a parameter θ based on a random sample(s) $X_1, \ldots, X_n$. Additionally, let the estimator $\hat{\theta}$ have an (approximately) normal sampling distribution with mean θ and standard error $\sigma_{\hat{\theta}}$.

- Then we have the following:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta > \theta_0$$
$$TS :$$
$$RR :$$



- Defining the RR (i.e. finding k)

  Assuming H_0 is true, if we desire an α-level test, then

- Thus, an equivalent form of the test, with level α is:

$$H_0 : \theta = \theta_0$$
$$H_A : \theta > \theta_0$$
$$TS :$$
$$RR :$$

- We can use this generalization for all of the tests that large sample tests we will do, and we can state the test statistic as

$$Z = \text{————}$$

  and thus they all have equivalent form of the rejection region (because the TS has been standardized).

- Conclusions and interpretations
  - Conclusions and interpretations (two steps) for hypothesis tests can follow a general format:

    Because our test statistic (COMPARISON of TS and RR) (IS or IS NOT) in the rejection region we (REJECT or FAIL TO REJECT) the null hypothesis.
    At the (ALPHA) significance level, there (IS or IS NOT) sufficient evidence to conclude (THE ALTERNATIVE HYPOTHESIS).

- Examples

  1. A honey farmer collects 55 ml of honey on average from each of his hives during summer months. Further, he knows that the amount collected from each hive is normally distributed with a variance of $\sigma^2 = 100$. This summer he is feeding his bees a new type of pollen and he suspects that it is causing them to produce more honey. A random sample of $n = 52$ hives yields $\bar{x} = 57.25$. Test the farmer's hypothesis at a significance level of $\alpha = 0.05$.

  2. A vice president in charge of sales for a large corporation claims that salespeople are averaging no more than 15 sales contacts per week. (He would like to increase this figure.) As a check on his claim, $n = 36$ salespeople are selected at random, and the number of contacts made by each is recorded for a single randomly selected week. The mean and variance of the 36 measurements were 17 and 9, respectively. Does the evidence contradict the vice president's claim? Use a test with level $\alpha = 0.025$.

3. A machine in a factory must be repaired if it produces more than 10% defectives among the large lot of items that it produces in a day. A random sample of 100 items from the day's production contains 15 defectives, and the supervisor says that the machine must be repaired. Does the sample evidence support his decision? Use a test with level $\alpha = 0.01$.

- Here is a summary of the large-sample α-level hypothesis tests:

Large-Sample α-Level Hypothesis Tests

$H_0 : \theta = \theta_0$.

$$H_a : \begin{cases} \theta > \theta_0 & \text{(upper-tail alternative).} \\ \theta < \theta_0 & \text{(lower-tail alternative).} \\ \theta \neq \theta_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$.

$$\text{Rejection region: } \begin{cases} \{z > z_\alpha\} & \text{(upper-tail RR).} \\ \{z < -z_\alpha\} & \text{(lower-tail RR).} \\ \{|z| > z_{\alpha/2}\} & \text{(two-tailed RR).} \end{cases}$$

- In any particular test, only one of the listed alternatives H_A is appropriate. Whatever alternative hypothesis that we choose, we must be sure to use the corresponding rejection region.

  The correct one depends on the research question / goal: what are we trying to show or find evidence for?

- More examples

4. A psychological study was conducted to compare the reaction times of men and women to a stimulus. Independent random samples of 50 men and 50 women were employed in the experiment. The results are shown below. Do the data present sufficient evidence to suggest a difference between true mean reaction times for men and women? Use $\alpha = 0.10$.

Men	Women
$n_1 = 50$	$n_2 = 50$
$\bar{y}_1 = 3.6$ seconds	$\bar{y}_2 = 3.8$ seconds
$s_1^2 = .18$	$s_2^2 = .14$

5. A car manufacturer aims to improve the quality of the products by reducing the defects and also increase the customer satisfaction. Therefore, he monitors the efficiency of two assembly lines in the shop floor. In line A there are 18 defects reported out of 200 samples. While the line B shows 25 defects out of 600 cars. At $\alpha = 5\%$, are the differences between two assembly procedures significant?

Small sample tests for μ and $\mu_1 - \mu_2$

- If we are testing one or two population means and the sample size is not large enough so that $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}} \overset{approx}{\sim} \text{Normal}\,(0, 1)$, then we need a different procedure.
- Just like with confidence intervals, we can switch to procedures based on the t-distribution when sampling from Normal distribution(s) (assuming unknown equal variances of both populations).

  The process is the same as the large sample Z-tests shown previously. We are just standardizing the point estimator and rearranging to get the rejection region, except it is based on t critical values now.
- If $H_0 : \mu = \mu_0$ is tested against $H_A : \mu < \mu_0$ then

- Here is a summary of the small-sample α-level tests for μ

A Small-Sample Test for μ

Assumptions: $Y_1, Y_2, \ldots, Y_n$ constitute a random sample from a normal distribution with $E(Y_i) = \mu$.

$H_0 : \mu = \mu_0$.

$$H_a: \begin{cases} \mu > \mu_0 & \text{(upper-tail alternative).} \\ \mu < \mu_0 & \text{(lower-tail alternative).} \\ \mu \neq \mu_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $T = \dfrac{\overline{Y} - \mu_0}{S/\sqrt{n}}$.

$$\text{Rejection region:} \begin{cases} t > t_\alpha & \text{(upper-tail RR).} \\ t < -t_\alpha & \text{(lower-tail RR).} \\ |t| > t_{\alpha/2} & \text{(two-tailed RR).} \end{cases}$$

t_α, with df $= n - 1$

- If we are testing two independent means $\mu_1 - \mu_2$ and assume both Normal distributions with common unknown variance σ^2, then we use the pooled variance S_p^2 as the estimator for σ^2 in the standard error $\sigma_{\bar{X}_1 - \bar{X}_2}$. Then

Small-Sample Tests for Comparing Two Population Means

Assumptions: Independent samples from normal distributions with $\sigma_1^2 = \sigma_2^2$.

$H_0 : \mu_1 - \mu_2 = D_0$.

$$H_a : \begin{cases} \mu_1 - \mu_2 > D_0 & \text{(upper-tail alternative).} \\ \mu_1 - \mu_2 < D_0 & \text{(lower-tail alternative).} \\ \mu_1 - \mu_2 \neq D_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $T = \dfrac{\overline{Y}_1 - \overline{Y}_2 - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$.

$$\text{Rejection region: } \begin{cases} t > t_\alpha & \text{(upper-tail RR).} \\ t < -t_\alpha & \text{(lower-tail RR).} \\ |t| > t_{\alpha/2} & \text{(two-tailed RR).} \end{cases}$$

Here, $P(T > t_\alpha) = \alpha$ and degrees of freedom $\nu = n_1 + n_2 - 2$.

- If we are testing dependent means $\mu_1 - \mu_2$, then

  This is called a **paired *t*-test**.

- Examples

  1. X is the growth in an induced tumor in type of lab mice. It is known that the mean growth of the tumor without treatment is 4.0 mm and that the distribution of $X \sim \mathrm{N}(\mu, \sigma^2)$. Scientists believe a new type of enzyme will have an effect on the growth of the tumor. Scientists apply the enzyme to a random sample of $n = 9$ lab mice with the induced tumor and observe $\bar{x} = 4.2824$ and $s = 1.2$.

     Test the scientists' hypothesis at a significance level of $\alpha = 0.10$.

2. Data on the length of time required to complete an assembly procedure using each of two different training methods is shown below. Is there sufficient evidence to indicate a difference in true mean assembly times for those trained using the two methods? Test at the $\alpha = .05$ level of significance.

Standard Procedure	New Procedure
$n_1 = 9$	$n_2 = 9$
$\bar{y}_1 = 35.22$ seconds	$\bar{y}_2 = 31.56$ seconds
$s_1^2 = 24.445$	$s_2^2 = 20.0275$

P-values

- So far, we have only made the decision to reject or fail to reject based on whether or not the test statistic falls in the rejection region ($TS \stackrel{?}{\in} RR$). This is called the **traditional method**.
- Lets review some examples:
  - Example 1 (average honey): $TS : z = 1.622$ and $RR : \{Z > 1.645\}$ for $\alpha = 0.05$.
  - Example 3 (proportion of defectives): $TS : z = 1.667$ and $RR : \{Z > 2.362\}$ for $\alpha = 0.01$.
- In both of these, we made the conclusion to _____, but were "closer" to rejecting H_0 example ___. We can think of this as being a "stronger" result (i.e. more evidence against H_0 (just not enough), but we need a way to quantify the "strength" of the result independent of the significance level.

- Definition: A **p-value** is the probability that under the null hypothesis the test statistic will be at least as "extreme" as the observed value.

- Notes:
  - At least as "extreme" just means in the direction of the alternative hypothesis.

$$H_A: \qquad \theta < \theta_0 \qquad \theta > \theta_0 \qquad \theta \neq \theta_0$$

– Interpretation of p-values: The smaller the p-value becomes, the more compelling is the evidence that the null hypothesis should be rejected.

  For small p-values, think: If θ_0 was true, the result we got had such a tiny probability to occur. So the original assumption of θ_0 must not actually be true.

– Making the decision to reject or fail to reject based on whether or not the p-value is less than the significance level is called the **p-value method**.

– More formally, a p-value represents the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected; p-value is the **attained (observed) significance level**.

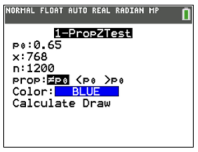
  Treating it like this leaves it up to the reader to evaluate the extent to which the observed data disagree with the null hypothesis and make their own choice in α in deciding whether or not to reject H_0. This is one advantage of p-values, and is why most scientific journals require p-values for all of their studies.

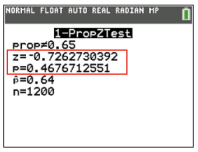
  (often $\alpha = 0.1, 0.05, 0.01$ are chosen out of convenience rather than a well-thought out choice.)

Calculator session

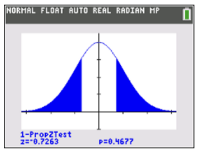
STAT > TESTS >

One proportion test: p





Calculate Output
prop = Alternative hypothesis
z = TS
p = p-value
$\hat{p}$= sample proportion
n = sample size

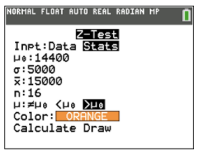


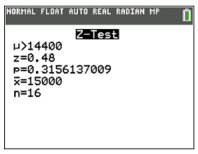
Draw Output
Plot (and displays values) of p = p-value and z = TS on the standard normal curve

H_0: $p = 0.65$
H_A: $p \neq 0.65$

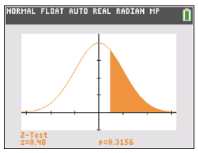
*This does NOT give us the Critical Value $z_{\alpha/2}$, we have to figure that out ourselves

One mean test: μ





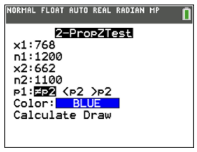
Calculate Output
μ= Alternative hypothesis
z = TS
p = p-value
$\bar{x}$= sample mean
n = sample size

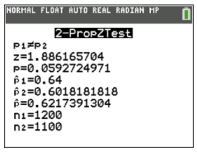


H_0: $\mu = 14{,}400$
H_A: $\mu > 14{,}400$

STAT > TESTS >

Two proportions test: $p_1 - p_2$





$H_0: p_1 - p_2 = 0$

$H_A: p_1 - p_2 \neq 0$

Calculate Output

$p_1 \neq <> p_2$ Alternative hypothesis

z = TS

p = p-value

$\hat{p}_1$ = sample proportion 1

$\hat{p}_2$ = sample proportion 2

$\hat{p}$ = pooled sample proportion → $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$

n_1 = sample size 1

n_2 = sample size 2

Relationship between confidence intervals and hypothesis tests

- For every confidence interval, there is an equivalent hypothesis test (and vice versa); two different ways of looking at the same thing.
- Demo for two-sided CIs and two-tailed tests:

  $100(1-\alpha)\%$ CI for θ $\qquad$ α-level test $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$

- The complement of the rejection region is the **acceptance region AR**:

- Thus, we "accept" $H_0 : \theta = \theta_0$ if θ_0 falls _____ the $100(1-\alpha)\%$ CI and reject if _____.

  So the confidence interval can be thought of as the set of values of θ_0 for which $H_0 : \theta = \theta_0$ is "acceptable" at level α.

  Based on this perspective, we can see that it is a range, not *one specific acceptable value* for the parameter. This is why we prefer to say "fail to reject" rather than "accept" the null hypothesis.

- One sided intervals and tests

  For $H_0 : \theta = \theta_0$ with level α and $100(1-\alpha)\%$ CIs

  - Upper tail test: $H_A : \theta > \theta_0$ $\qquad$ Reject if outside

  - Lower tail test: $H_A : \theta < \theta_0$

Errors in hypothesis tests

- In deciding to reject or fail to reject H_0, an experimenter might be making a mistake (we can never really know what the truth is, just like with confidence intervals).

  Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes.

- For any fixed rejection region, two types of errors can be made in reaching a decision.
  - **Type I error** is made if H_0 is rejected when H_0 is true.

    The probability of a type I error is denoted by α, the **significance level** of the test.

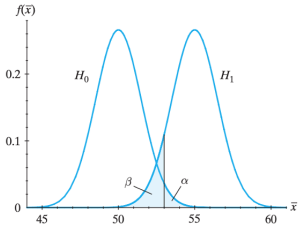
  - **Type Il error** is made if H_0 is accepted when H_A is true.

    The probability of a type II error is denoted by β.

- We can think of α and β as measuring the risks associated with the two possible incorrect decisions that might result from a statistical test. Because of this, they provide a very practical way to measure the goodness of a test.
- In calculating these error probabilities, we want tests that minimize both quantities while at the same time maximizing the **power** $= 1 - \beta$.

  The power of a test represents the probability of correctly rejecting a false null hypothesis (given a particular alternative hypothesis).

- Example: Find the probabilities of a type I error, type II error, and power for the breaking strength example (testing $H_0 : \mu = 50$ vs $H_A : \mu = 55; n = 16, \sigma^2 = 36$).



pdf of $\overline{X}$ under H_0 and H_1

- Note that the value of β depends on the true value of the parameter θ in the alternative hypothesis (needed to assume $\mu = 55$ in the type II probability calculation).

  The larger the difference is between θ and the (null) hypothesized value of $\theta = \theta_0$, the smaller is the likelihood that we will fail to reject the null hypothesis.

  Example: Find the new type II error probability and power if $H_A : \mu = 57$ and if $H_A : \mu = 51$.

- This example shows that the test using $RR = \{\bar{x} \geq 53\}$ guarantees a low risk of making a type I error _____, but it does not offer adequate protection against a type II error (high βs with some alternative hypotheses).
- Typically, in practice the type I error probability (significance level) is controlled, and then we choose a test that minimizes the type II error probability (and thus maximizing the power).

  However, there is often some give and take with these: as one error likelihood decreases, the other often increases (i.e. α and β are inversely related).
- So how can we improve our test? One way is to balance α and β by changing the rejection region, specifically we can enlarge the RR.

  This will lead us to reject H_0 more often, which means accept H_0 less often.

- Often we have to think about the consequences of committing each type of error and determine which error is more severe and therefore how to minimize its probability.
- Example: Write the consequences of each type of error and determine which is more severe.

  All commercial elevators must pass yearly inspections. An inspector has to choose between certifying an elevator as safe (no repairs needed) or saying that the elevator is not safe (repairs are needed). There are two hypotheses:

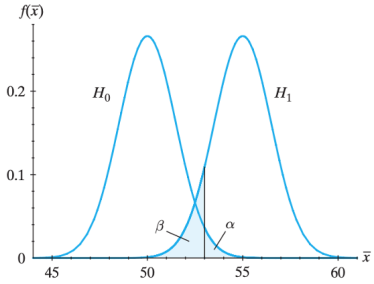
  $$H_0 : \text{The elevator is not safe (repairs are needed)}$$
  $$H_A : \text{The elevator is safe (no repairs needed)}$$

  – Consequences of Type I error:

  – Consequences of Type II error:

- How can we reduce both? For almost all statistical tests, if α is fixed at some acceptably small value, β decreases as the sample size increases.

  Intuitively obvious, collect more data!



After Test 3

Contents

Lecture 8 – Regression

MATH 321: Mathematical Statistics

Lecture 8: Regression

Applied Linear Statistical Models: Chapters 1-3

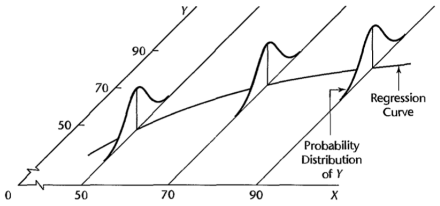
Introduction

Regression overview

- Goal is to determine _____ one variable is related to a set of other variables.
- Variables
  - Response variable, denoted Y, represents an outcome whose variation is being studied.
  - Explanatory variable, denoted X, represents the causes (i.e. potential reasons for variation).
- Two types of relationships
  - Functional (deterministic): There is an exact relation between two variables (have the form _____).
  - Statistical (probabilistic): There is not an exact relation because there are other variables that affect the relationship (have the form _____).

Regression models and their uses

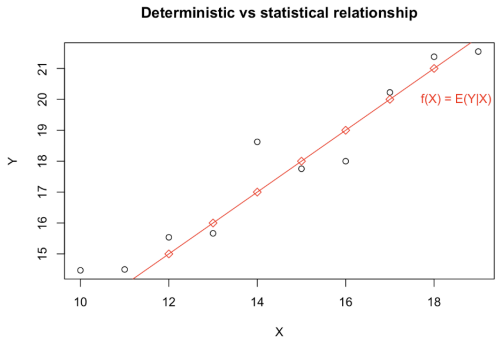
- Statistical models quantify the relationship between a response variable (i.e. a random variable) and explanatory variables, which are usually assumed to be deterministic (i.e. known exactly).
- Elements of a statistical regression model
  - In general, observations do not fall directly on the curve of a relationship.
    - $Y \mid X$ has a probability distribution.
    - $E(Y \mid X)$ varies deterministically with X.



- So the statistical model is:

$$\begin{aligned} Y &= E(Y \mid X) + \epsilon \\ &= f(X) + \epsilon, \qquad \text{where } \epsilon \text{ has some distribution} \end{aligned}$$

- Two components of a statistical model:
  1. $f(X) = E(Y \mid X)$: Defines relationship between Y and X; explains the _____ of the response.
  2. ϵ: An element of randomness (i.e. error). This contains the _____ that $f(X)$ cannot explain and/or that is of no interest.
- This means $f(X) = E(Y \mid X)$ will be the same for all samples with the same X values. The only thing that changes is the random error ϵ and as a result Y. Example $Y = 3 + 1X + \epsilon$:

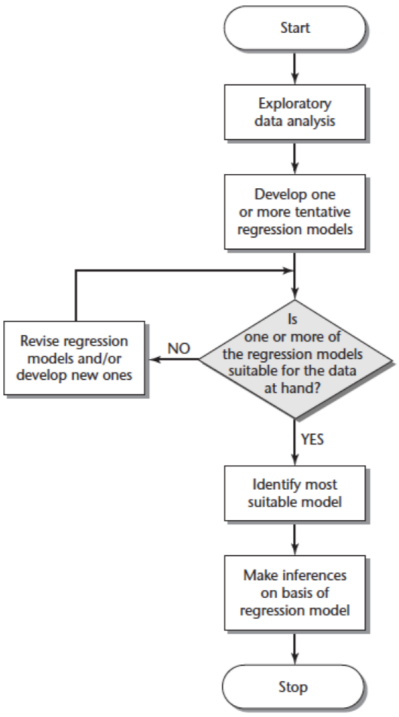


Construction of statistical regression models

1. Selection of predictor variables (how to decide which ones?).
   - Use of outside information, historical knowledge, and/or experience.
   - Exploratory data analysis.
   - Variable selection techniques: Find a subset of important variables (i.e. practical and easy to find).
2. Functional form of the regression relation (what is form of $f(X)$?).
   - < based on same info as (1) >
   - If there is an abundance of data, maybe start with more complex models and then simplify.
3. Scope of model (when is the model useful?).
   - When the model best predicts or describes the relationship between response and predictor variables.

Uses of statistical regression models

1. Determining whether an X "affects" Y or not.
2. Estimation of impact of a given X on the Y.
3. Estimation of the mean of Y for a given X value.
4. Prediction of a single value of Y for a given X value.



Simple linear regression (SLR)

Goal of SLR

- Investigate the relationship between Y and a single numeric independent variable X, assuming that, in the population, the mean of Y is linearly related to the value of X.
- Population relationship: $Y = \beta_0 + \beta_1 X + \epsilon$.
- Sample relationship: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Data structure

- Both X and Y on a random sample of n individuals are collected from the population of interest. The resulting data has the form $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Model statement: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- Y_i: Dependent (or response) variable value. These are independent, but not identically distributed (_____).
- X_i: Independent (or predictor) variable value. These are **not random variables**, rather _____.
- ϵ_i: Random error term, **assumed** to have mean zero and variance σ^2.
  $\text{Cov}(\epsilon_i, \epsilon_j) = \text{Corr}(\epsilon_i, \epsilon_j) = 0$ for all $i, j : i \neq j$. Often, the ϵ_i are assumed to be *iid*.
- β_0 and β_1: _____ regression parameters that need to be estimated.
- σ^2: Another parameter that needs estimated, but it is technically not a "regression" parameter since it does not determine the relationship between Y and X (i.e. it only deals with randomness).
- Note that Y_i and ϵ_i are random variables and therefore have distributions. Thus, discussing their mean and variances are appropriate.

Some implications of above

- Mean of Y_i for given X_i Variance of Y_i for given X_i

Interpretation of regression parameters (β_0, β_1)

- β_0: Y-intercept of the regression line and gives Y's mean when $X = 0$

- β_1: Slope of the regression line and indicates the change in Y's **mean** when X increases by one unit

  - Determines whether a relationship exists between Y and X.
  - Note that regression **does not** substantiate or prove a **cause-effect** relationship. Rather it gives evidence that Y and X are related (but not that X "causes" the value of Y).

Estimation of the regression function

- Setup
  - For each point we have an observed value Y_i, a fitted value $\hat{Y}_i$ and a residual $\hat{\epsilon}_i$.
  - Fitted regression function: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, where $\hat{\beta}_0$ and $\hat{\beta}_0$ are estimators of β_0 and β_1, respectively.
- Goal
  - Goal is to estimate the two "regression" parameters β_0 and β_1.
  - There are several methods to do this.

Method of least squares

- Overview
  - For each observation (X_i, Y_i), this method considers the model error term, which is the deviation of Y_i from its expected value:

$$\epsilon_i = Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$$

  - Then we minimize the sum of some function of these errors:

$$\begin{aligned} Q &= \sum_{i=1}^{n} \text{function of } \epsilon_i \\ &= \sum_{i=1}^{n} \text{function of } \big(Y_i - E(Y_i)\big) \\ &= \sum_{i=1}^{n} \text{function of } \big(Y_i - (\beta_0 + \beta_1 X_i)\big) \qquad < \text{for SLR} > \end{aligned}$$

  - For least squares method specifically, we consider the sum of the n squared errors (deviations). Thus we have:

$$\begin{aligned} Q &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

  - And the point estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that achieve the minimum Q. These can be found analytically.

- Results

$$\begin{array}{rrcccccc} \text{Intercept} & \hat{\beta}_0 & = & \frac{1}{n}\sum Y_i + \hat{\beta}_1 \frac{1}{n}\sum X_i & = & \bar{Y} - \hat{\beta}_1 \bar{X} & & \\ \text{Slope} & \hat{\beta}_1 & = & \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i Y_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2} & = & \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} & = & \frac{S_{XY}}{S_{XX}} \end{array}$$

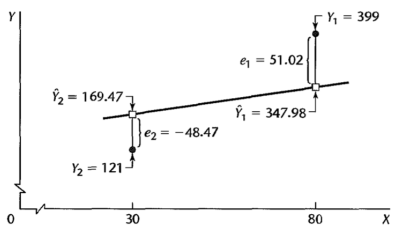
- Derivation

  - Note: We did not have to assume any distribution of the error term. These are the LSE estimators for any SLR model.

- Least squares estimators contain some optimal properties (Best Linear Unbiased Estimator), similar to how MLEs did.

Residuals and estimation of the error terms variance (useful for inference on model)

- Residuals
  - $\hat{\epsilon}_i = e_i = Y_i - \hat{Y}_i$: This is a known, observable estimate of the unobservable model error. Measures the deviation of the observed value from the fitted regression function.



  - Residuals are very useful for studying whether the given regression model is appropriate for the data.

- Error terms variance
  - Need to estimate the variance σ^2 of the error terms ϵ_i in a regression model to get an indication of the variability of the probability distributions of Y.
  - Motivation: Very similar to variance of a single population $S^2 = \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}$, except we use the residuals as the deviations because each Y_i comes from a different probability distribution with different X (depends on the X_i level).

$$\text{Error (residual) sum of squares} \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

  - Then divide by the $df = n-2$ to the mean square (two dfs are lost when because β_0 and β_1 need to be estimated when getting the estimated means $\hat{Y}_i$).

$$\text{Error (residual) mean square} \quad S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

  - It can be shown that MSE is an unbiased estimator for σ^2: $E(MSE) = \sigma^2$.

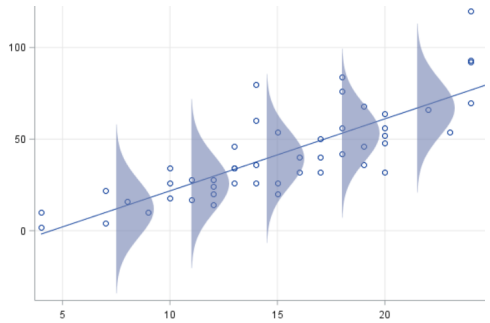
    An estimator of the standard deviation is simply $S = \sqrt{MSE}$.

Normal error regression model

- These assumptions on ϵ_i are needed to set up interval estimates and make tests.
- The standard assumption is that the error terms are normally distributed. This greatly simplifies the theory of regression analysis and is justifiable in many real-world situations where regression analysis is applied.

$$\text{New regression model:} \qquad Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \text{where} \quad \epsilon_i \overset{iid}{\sim} \text{Normal}\,(0, \sigma^2)$$

- This model means $Y_i \overset{\perp\!\!\!\perp}{\sim}$ Normal with $E(Y_i) = \beta_0 + \beta_1 X_i$ and $V(Y_i) = \sigma^2$.



- Justification of the normality assumption
  - Error terms frequently represent the effects of factors omitted from the model that affect the response to some extent and that vary at random without reference to the variable X.
  - These random effects have a degree of mutual independence, the composite error term representing all these factors tends to normal as the number of factors becomes large (by the CLT).
  - Also, the estimation and testing procedures shown later are based on the t distribution and are usually only sensitive to large departures from normality. So, unless the departures from normality are serious, particularly with respect to skewness, the actual confidence coefficients and risks of errors will be close to the levels for exact normality.

Estimation of parameters by method of maximum likelihood

- We can also estimate the parameters β_0, β_1, and σ^2 using maximum likelihood estimation.

Parameter	MLE	
β_0	$\hat{\beta}_0$	Same as LSE
β_1	$\hat{\beta}_1$	Same as LSE
σ^2	$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n}$	

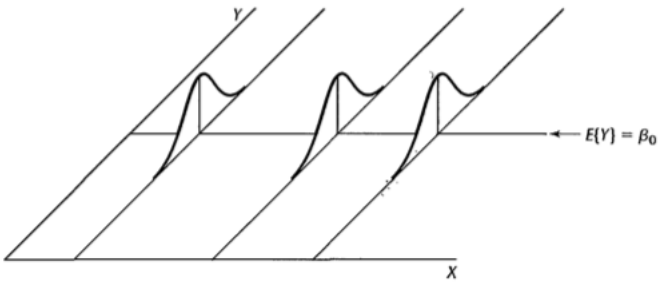
- Because of these results, $\hat{\beta}_0$ and $\hat{\beta}_1$ also possess the optimal properties of MLEs like consistency and minimum variance in the class of unbiased estimators.

Inference

- For the rest of this section, assume the normal error regression model from above is applicable.

Inferences concerning β_1

- Overview
  - We often want to make inferences about β_1. A common test on β_1 has the form below.
  - If $\beta_1 = 0 \implies$ Regression line in horizontal, which means there is no linear association between Y and X, and even more no relation of any type because all probability distributions of Y are identical at all levels of X: normal with $E(Y) = \beta_0 + (0)X = \beta_0$ and variance σ^2.



- Sampling distribution of $\hat{\beta}_1$
  - Refers to distribution of $\hat{\beta}_1$ from repeated sampling when the levels of the predictor variable X are held constant from sample to sample.
  - Recall $\hat{\beta}_1 = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$; this is the point estimator.
  - Distribution of $\hat{\beta}_1$ is Normal with mean and variance:

$$E(\hat{\beta}_1) = \beta_1$$
$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

  - Then we can estimate the variance by replacing the parameter σ^2 with MSE, the unbiased estimator of σ^2. This gives us $S^2_{\hat{\beta}_1}$, which is an unbiased estimator for the variance of the sampling distribution of $\hat{\beta}_1$. And we can take the positive square root to give us $S_{\hat{\beta}_1}$, which is the point estimator of $\sigma_{\hat{\beta}_1}$.

$$S^2_{\hat{\beta}_1} = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{MSE}{S_{XX}} \qquad \longrightarrow \qquad s_{\hat{\beta}_1} = \sqrt{\frac{MSE}{S_{XX}}} = \frac{S}{\sqrt{S_{XX}}}$$

  - Thus, $S^2_{\hat{\beta}_1}$ is an unbiased estimator for the variance of the sampling distribution of $\hat{\beta}_1$.

- Sampling distribution of standardized $\hat{\beta}_1$: $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1}$

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{XX}}} \sim \mathrm{t}_{\,n-2}$$

Tests on β_1

- The test shown below is called a test of utility of the model for SLR.
- If reject: We conclude that X does contribute information for the prediction of Y when using the straight-line model.

  If fail to reject: Then we conclude there is no linear relationship between Y and X (horizontal model). But keep in mind:

  - Additional data might indicate that β_1 differs from zero.
  - A more complex relationship between Y and X may exist, which would require fitting a model other than the straight-line model.
  - All assumptions about the error terms ($\epsilon_i \overset{iid}{\sim} \text{Normal}\,(0, \sigma^2)$) should be satisfied.
- Two-tailed test (most common)
  - Hypotheses

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

  - Test statistic

$$TS = t^* = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{MSE/S_{XX}}}$$

  - Rejection region and p-value

$$RR = \{|t| > t_{\alpha/2, n-2}\}$$
$$p\text{-value} = 2 \cdot P(t_{n-2} \geq |t|)$$

  - Decision
    - Reject H_0 and conclude H_A if $\quad TS \in RR \quad \Longleftrightarrow \quad p\text{-value} \leq \alpha$
    - Fail to reject H_0 if $\quad TS \notin RR \quad \Longleftrightarrow \quad p\text{-value} > \alpha$
    - Can also look at the $100(1-\alpha)\%$ CI for β_1 to see if contains 0.
  - Conclusion / Interpretation

    At the α significance level, we < have / do not have > sufficient evidence of a significant linear relationship between < Y context > and < X context >. < if yes... > This is a < positive / negative > linear relationship, indicating that as < X context > increases, < Y context > < increases / decreases >, on average.

Descriptive measures of linear association between X and Y

- Overview
  - There is no one single measure to completely describe the usefulness of a regression model for a particular application.
  - If the goal is estimation of parameters and means and predicting new observations, usefulness of estimates or predictions depends upon the width of the interval and the user's needs for precision. This can vary from one application to another.
  - Rather than making inferences, goals could be to describe the degree of linear association between Y and X.
- **Coefficient of determination R^2**
  - A very common measure because of its simplicity is the coefficient of determination R^2, which is a measure of the effect of X in reducing the uncertainty in predicting Y. This reduction in sum of squares ($SSTO - SSE = SSR$) gets expressed as a proportion:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}, \qquad \text{range:} \quad 0 \leq R^2 \leq 1$$

  - Interpretation: $< R^2 * 100 >\%$ of the variation in $< Y$ context $>$ can be explained by the linear relationship between $< Y$ context $>$ and $< X$ context $>$.
    - So, the larger R^2 is, the more the total variation of Y is reduced by introducing the predictor variable $X \Longleftrightarrow$ greater degree of linear association between Y and X.
    - Practically, this indicates the quality of the fit by measuring the proportion of variability explained by the fitted model.
  - Facts about R^2:
    - $0 \leq R^2 \leq 1$, which ranges from horizontal regression line to perfect fit.
    - Usefulness in prediction: A high coefficient of determination does not necessarily indicate that useful (precise) predictions can be made.
    - Overfitting: R^2 can be artificially inflated by including additional model terms (adding extra predictors). This is because SSR always increases with more predictors, even if they are completely unrelated to the response variable.

Diagnostics

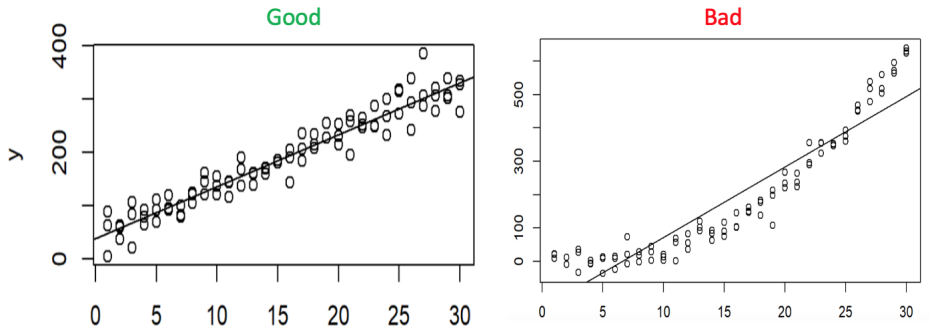
Overview

- Diagnostics are methods to check whether our model is reasonable for our data and representative of the system that we are studying (i.e. assumption checking).
- Why do we need to check the model?
  - The goal of building a model is to **learn something** about the real world or **predict outcomes** in the real world.
- To use a model successfully, we need to know its limitations:
  - Does it adequately describe the functional relationship of interest?
  - Is there reason to worry that inferences about the parameters might be flawed?
  - Is the error distribution appropriate?
- All of these are checked via **residual analysis**, whose goal is to assess the aptness of a statistical model.
- Why residuals?
  - Direct diagnostic plots for the response variable Y are ordinarily not too useful in regression analysis because the response variable observations are a function of the level of predictor variable.
  - So, instead we look at diagnostics for Y indirectly by examining the residuals.
- For our regression model, we assume $\epsilon_i \overset{iid}{\sim} \text{Normal } (0, \sigma^2)$.

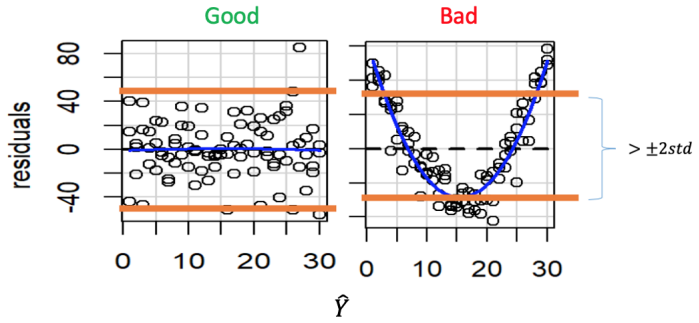
  So if the model is appropriate for the data at hand, the residuals should reflect these properties.

Residual analysis (LINE)

- Linearity
  - Can look at the scatterplot of Y vs X from the initial EDA to see if a linear model is appropriate:



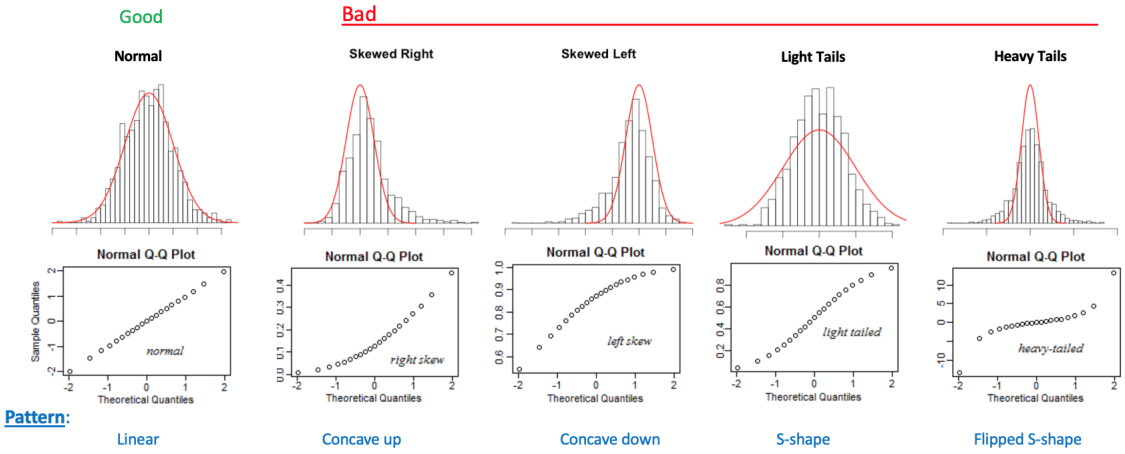
  - The preferred plot is the **residual plot against the fitted values** plot.



    * When a linear regression model is appropriate, the residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative (randomly scattered around 0).
    * When the linearity assumption is violated, there are systematic deviations.

- Independence
  - Ideally, any potential source of dependence is handled at the experimental design stage (or the sampling scheme), so that it is either eliminated by randomization or explicitly included in the data and we have one observation per subject.
  - There is a plot to look at this, but we will not cover it.

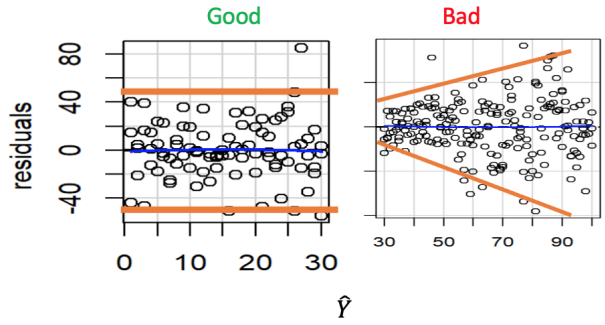
- Normality
  - Small departures from normality do not create any serious problems, but major departures should be of concern.
  - We can check the normality of error terms in a variety of ways:
    - Normal probability (QQ) plot of residuals.



    - Distribution plots of residuals: Boxplots should be symmetric and histograms should be roughly normal.
  - Difficulties in assessing normality: The analysis for model departures regarding normality is often more difficult than departures of other types because...
    - Random variation can be particularly mischievous when studying the nature of a probability distribution unless the sample size is quite large.
    - Even worse, other types of departures can and do affect the distribution of the residuals.

      e.g. Residuals may appear to be not normally distributed because an inappropriate regression function is used or because the error variance is not constant.
  - So, it is usually a good strategy to investigate these other types of departures first, before assessing the normality of the error terms.

- Equal variance
  - We can again look at the residual plot against the fitted values.



    * When there is a constant error variance, points again should fall within a horizontal band. So there is a constant spread of the residuals as move across the scope of fitted (or X) values.
    * "Tipped over tornado" effect of the points indicates a non-constant variance (e.g. as the fitted values increase, the residuals vary more, or vice versa).

      A nonconstant variance in called *heteroscedasticity* (the assumption is a *homoscedastic* error variance).