MATH 321: Final Study Guide

Lecture 14 – Bivariate Distributions (4.1 and 4.4)

Joint pmf and pdf

- Discrete definition: The 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. $0 \le f_{X,Y}(x, y) \le 1 \quad$ for all x, y
  2. $\displaystyle\sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = 1$
  3. Let A be any subset of $\mathbb{R}^2$, then $P((X, Y) \in A) = \sum \sum_A f(x, y)$
- Continuous definition: The joint pdf is a function $f(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ such that
  1. $f_{X,Y}(x, y) \ge 0 \quad$ for all x, y
  2. $\displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y\,\mathrm{d}x = 1$
  3. For $A \subset \mathbb{R}^2$, $\displaystyle P((X, Y) \in A) = \int \int_A f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int \int_A f(x, y)\,\mathrm{d}y\,\mathrm{d}x$

Marginal distributions

- Discrete definition: Let (X, Y) have joint pmf $f(x, y)$. Then, the marginal pmfs are given by

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

- Continuous definition: Let (X, Y) have joint pdf $f(x, y)$. Then the marginal pdfs are defined by:

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

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 $f(x, y)$,

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

Special expectations

- Definitions: Let (X_1, X_2) be a bivariate random vector with joint pmf / pdf $f(x_1, x_2)$.
  i) If $g(X_1, X_2) = X_1$, then $E[g(X_1, X_2)] = E(X_1) = \mu_{X_1}$
  ii) If $g(X_1, X_2) = (X_1 - \mu_1)^2$, then $E[g(X_1, X_2)] = E[(X_1 - \mu_1)^2] = \sigma^2_{X_1}$
  iii) If $g(X_1, X_2) = \mathrm{e}^{tX_1}$, then $E[g(X_1, X_2)] = E(\mathrm{e}^{tX_1}) = M_{X_1}(t)$

Expected value of $X + Y$ and XY

- Theorem: Expected value of a sum of two random variables

  If $g(X, Y) = X + Y$, then $E(X + Y) = E(X) + E(Y)$
- Generalized theorem: If $g_1(X, Y)$ and $g_2(X, Y)$ are two functions and a, b and c are constants, then

  $E[ag_1(X, Y) + bg_2(X, Y) + c] = aE[g_1(X, Y)] + bE[g_2(X, Y)] + c$
- Theorem: Expected value of a product of two random variables

  If $g(X, Y) = XY$ and $X \perp\!\!\!\perp Y$, then $E(XY) = E(X) \cdot E(Y)$

Lecture 15 – Conditional Distributions (4.3)

Conditional pmf / pdf

- Definition: Let (X, Y) be a bivariate random vector with joint pmf / pdf $f(x, y)$ and marginal pmfs / pdfs $f_X(x)$ and $f_Y(y)$.
  (a) Given x such that $f_X(x) > 0$, $\quad f(y \mid x) = \dfrac{f(x, y)}{f_X(x)}$
  (b) Given y such that $f_Y(y) > 0$, $\quad f(x \mid y) = \dfrac{f(x, y)}{f_Y(y)}$

Probabilities

- For $A \subset \mathbb{R}^2$,

  Discrete: $P(X \in A \mid Y = y) = \displaystyle\sum_{x \in A} P(X = x \mid Y = y) = \sum_{x \in A} f(x \mid y)$

  Continuous: $P(X \in A \mid Y = y) = \displaystyle\int_A f(x \mid y)\,\mathrm{d}x$

Relationship between joint pmf and conditional pmfs

- Theorem: For bivariate random vector (X, Y) with joint pmf / pdf $f(x, y)$ and x and y such that $f_X(x) > 0$ and $f_Y(y) > 0$,

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

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 given by

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

- 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

  $$E(Y \mid X = x) = \sum_y y\, f(y \mid x) = \mu_{Y|X}$$

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

  $$E[(Y - \mu_{Y|X})^2 \mid X = x] = \sum_y (y - \mu_{Y|X})^2\, f(y \mid x) = \sigma^2_{Y|X}$$

Lecture 16 – Independence and the Correlation Coefficient (4.1, 4.2, and 4.4)

Independence for random variables

- Definition: Let (X, Y) be a bivariate random vector with joint pdf / pmf $f(x, y)$ and marginal pdfs / 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 $y \in \mathbb{R}$,

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

- Checking 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

Conditional distributions and independence

- Theorem: If X and Y are independent, $f(x \mid y) = f_X(x) \qquad \text{and} \qquad f(y \mid x) = f_Y(y)$

Using independence

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

  (a) For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$

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

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

Definition, theorems and properties of covariance

- Definition: The covariance of X and Y is the number defined by: $\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$
- If (X,Y) is discrete, then $E[(X-\mu_X)(Y-\mu_Y)] = \sum_x \sum_y (x-\mu_x)(y-\mu_y f(x,y)$
- Alternate calculation for covariance: $\text{Cov}(X,Y) = E(XY) - E(X)\cdot E(Y)$
- Variance is a special case of covariance: $V(X) = \text{Cov}(X,X)$
- Order in covariance does not matter (i.e. symmetric): $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
- Covariance of a random variable and a constant is zero: If c is a constant, then $\text{Cov}(X,c) = 0$
- Can factor out coefficients in covariance: $\text{Cov}(aX, bY) = ab\cdot\text{Cov}(X,Y)$
- Can factor out coefficients, but added constants disappear: $\text{Cov}(aX+c, bY+d) = ab\cdot\text{Cov}(X,Y)$
- Distributive property of covariance: $\text{Cov}(X, Y+Z) = \text{Cov}(X,Y) + \text{Cov}(X,Z)$
- Independence and covariance theorem: If $X \perp\!\!\!\perp Y$ then $\text{Cov}(X,Y) = 0$

Correlation definition and properties

- Definition: $\rho_{XY} = \text{Corr}(X,Y) = \dfrac{\text{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \dfrac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$
- Theorem: For any random variable X and Y,
  i) $-1 \le \rho_{XY} \le 1$
  ii) $\rho_{XY} = 1$ if and only if there exist numbers $a > 0$ and b such that $P(Y = aX+b) = 1$.
  iii) $\rho_{XY} = -1$ if and only if there exist numbers $a < 0$ and b such that $P(Y = aX+b) = 1$.
  iv) When $\rho_{XY} = 0$, X and Y are uncorrelated.

Variance of $X+Y$

- Theorem: Variance of a sum of two random variables

  $V(X+Y) = V(X) + V(Y) + 2\,\text{Cov}(X,Y)$

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

Lecture 17 – Several Random Variables (5.3 and 5.4)

Definitions and theorems

- Joint distributions
  - Discrete definition: 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

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

    Then for any $A \subset \mathbb{R}^n$,

    $$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x})$$
  - Continuous definition: 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

  $$E[g(\mathbf{X})] = \quad \overbrace{\sum_{\mathbf{x} \in \mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x})}^{\text{Discrete}} \qquad \overbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}^{\text{Continuous}} g(\mathbf{x}) f(\mathbf{x}) \mathrm{d}x_1 \cdots \mathrm{d}x_n$$
- 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.
- 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.

Independence

- 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

  $$f(X_1, \ldots, X_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$
- Conditional distributions: If $X_1, \ldots, X_n$ are mutually independent, 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.
- 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

  $$E[g_1(X_1) \cdots g_n(X_n)] = \prod_{i=1}^{n} E[g_i(x_i)]$$

Linear functions of random variables

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

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

- Expected value of a linear function of random variables

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

- Variance of a linear function of random variables

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

  If $X_1, \ldots, X_n$ are mutually independent,

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

Mgf of sums of independent 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 $Y = X_1 + \cdots + X_n$.

  $$M_Y(t) = M_{X_1+\cdots+X_n}(t) = M_{X_1}(t) \cdots M_{X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

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

  $$M_Y(t) = \big[M_X(t)\big]^n$$

Sums of linear combinations of 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 $a_1, \ldots, a_n$ 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

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

- Sum of linear function of normals theorem: Let $X_1, \ldots, X_n$ be mutually independent random variables with $X_i \sim \mathrm{Normal}\,(\mu_i, \sigma_i^2)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Then,

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

Lecture 1 – Random Samples and Common Statistics (5.5)

Basic concepts of random samples

- Random sample definition: $X_1, \ldots, X_n$ are a random sample of size n from the population $f(x)$ if they are *iid* random variables.
- Statistic (estimator) definition: The random variable / vector for any function of a random sample $Y = T(X_1, \ldots, X_n)$ is called a statistic, and it's distribution is called a sampling distribution.

Sample mean and variance

- Definitions
  - Sample mean: The arithmetic average of the values in a random sample

    $$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
  - Sample variance: The statistic defined by $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2$
  - Sample standard deviation: 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 with mean μ and variance $\sigma^2 < \infty$. Then

  (a) $\mu_{\bar{X}} = E(\bar{X}) = \mu$ (b) $\sigma^2_{\bar{X}} = V(\bar{X}) = \frac{\sigma^2}{n}$ (c) $E(S^2) = \sigma^2$
- Sampling distribution of $\bar{X}$ from random sample $X_1, \ldots, X_n$

  Theorem: Mgf of the sample mean is $M_{\bar{X}}(t) = [M_X(t/n)]^n$

Sampling from the normal distribution

- Let $X_1, \ldots, X_n$ be a random sample of size n from a Normal (μ, σ^2) distribution. Then

  (a) $\bar{X} \perp\!\!\!\perp S^2$ (b) $\bar{X} \sim \text{Normal}\,(\mu, \frac{\sigma^2}{n})$ (c) $\dfrac{(n-1)}{\sigma^2}S^2 \sim \chi^2\,(n-1)$

Chi-square random variables

- If $Z \sim \text{Normal}\,(0,1)$, then $Z^2 \sim \chi^2\,(1) \rightarrow \left(\dfrac{\bar{X}-\mu}{\sigma}\right)^2 = Z^2 \sim \chi^2\,(1)$
- Additive *df*: If $X_1, \ldots, X_n$ are mutually independent and $X_i \sim \chi^2\,(r_i)$ for $i = 1, \ldots, n$, then $Y = X_1 + \cdots + X_n \sim \chi^2\,(r_1 + \cdots + r_n)$
- 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

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

t distribution

- Definition: Let $X_1, \dots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. Then $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{\,n-1}$
- Derivation: $\dfrac{Z}{\sqrt{{\chi^2}_r/r}} \sim t_{\,r}$

F distribution

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

  $W = \dfrac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2}$ then $W \sim F\,(n-1, m-1)$. In general, $W \sim F\,(r_1, r_2)$.
- Derivation: $\dfrac{{\chi^2}_{r_1}/r_1}{{\chi^2}_{r_2}/r_2} \sim F\,(r_1, r_2)$
- Relationship to other distributions theorem

  (a) If $X \sim F\,(r_1, r_2)$ then $1/X \sim F\,(r_2, r_1)$ (b) If $X \sim t_{\,r}$ then $X^2 \sim F\,(1, r)$

Lecture 2 – Order Statistics (6.3)

Order statistics definition and distributions

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

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

- Distribution theorems
  - Cdf:

$$\begin{aligned} F_{X_{(j)}}(x) = P(X_{(j)} \le x) &= \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k} \\ &= P(Y \le j), \quad \text{where} \quad Y \sim \text{Binomial}\,(n, p = P(X \le x) = F_X(x)) \end{aligned}$$

  - Pdf:

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!} \, [F_X(x)]^{j-1} \, f_X(x) \, [1 - F_X(x)]^{n-j} \\ &= [\text{multinomial coefficient}] \times [j-1 \text{ RVs } \le x] \times [1 \text{ RV} \approx x] \times [n-j \text{ RVs } > x] \end{aligned}$$

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

- Extreme order stats

  Min $\rightarrow$ $F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n$; $f_{X_{(1)}}(x) = nf_X(x)[1 - F_X(x)]^{n-1}$

  Max $\rightarrow$ $F_{X_{(n)}}(x) = [F_X(x)]^n$; $f_{X_{(n)}}(x) = n[F_X(x)]^{n-1}f_X(x)$

Specific order statistics and functions of order statistics

- Sample median M

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

- Sample range, $R = X_{(n)} - X_{(1)} = max(X_1, \ldots, X_n) - min(X_1, \ldots, X_n)$
- $IQR = Q_3 - Q_1$
- Midrange $= \dfrac{X_{(1)} + X_{(n)}}{2}$

Order statistics as estimators of population percentiles

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

  If $X_{(1)}, \ldots, X_{(n)}$ are order statistics, then $E[F_X(X_{(j)})] = \dfrac{j}{n+1}, \quad j = 1, \ldots, n$

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

q–q plots

- Expected probability between two adjacent order statistics theorem:

  $E[F_X(X_{(j)}) - F_X(X_{(j-1)})] = \frac{1}{n+1}$; $E[F_X(X_{(1)})] = \frac{1}{n+1}$; $E[1 - F_X(X_{(n)})] = \frac{1}{n+1}$
- q–q plot definition: Let $x_{(1)}, \ldots, x_{(n)}$ be the observed sample order statistics and $x_{\frac{1}{n+1}}, \ldots, x_{\frac{n}{n+1}}$ be the percentiles from some particular distribution. A q–q plot is a plot of the points

  $(x_{(1)}, x_{\frac{1}{n+1}})$, $\ldots$, $(x_{(n)}, x_{\frac{n}{n+1}})$
- Interpretation of a q–q plot

  Good model $\rightarrow$ Follows $y = x$ line.

  Bad model $\rightarrow$ Strong deviation from this line.
- q–q plots for the normal distribution.

  If plot $(x_{(1)}, z_{\frac{1}{n+1}})$, $\ldots$, $(x_{(n)}, z_{\frac{n}{n+1}})$, then $\frac{1}{\text{slope}} \approx \sigma$

Lecture 3 – Exploratory Data Analysis (6.2)

Univariate EDA

- Descriptive statistics: Goal is to summarize a whole dataset wtih a single or few measures
  - Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}v$
  - Data (or population) variance: $v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Displaying data
  - Frequency tables: Data is grouped into intervals of equal length (bins)

    Freq = count of observations in each; Relative freq = proportion of observations in each bin = Freq / n
  - Histograms: Shape and summary stats

Right-skewed:	mean	$>$	median	$>$	mode
Symmetric:	mean	$\approx$	median	$\approx$	mode
Left-skewed:	mean	$<$	median	$<$	mode
  - Density histograms: Estimate underlying pdf

    For constants c_1 and c_2, $P(c_1 \le X < c_2) \approx \frac{\text{Freq}}{n}$ on $(c_1, c_2]$

    Height of bar $h(x) = \frac{\text{Freq}}{n(c_2 - c_1)}$
- Empirical rule:
  1. $\approx 68\%$ of data is in $(\bar{x} - s, \bar{x} + s)$.
  2. $\approx 95\%$ of data is in $(\bar{x} - 2s, \bar{x} + 2s)$.
  3. $\approx 99.7\%$ of data is in $(\bar{x} - 3s, \bar{x} + 3s)$.
- Order statistics:
  - 5 number summary
    1. Sample minimum $x_{(1)}$
    2. Lower quartile or First (lower) quartile $q_1 = \hat{x}_{0.25}$
    3. Median (second quartile) $m = \hat{x}_{0.5}$
    4. Third (upper) quartile $q_3 = \hat{x}_{0.75}$
    5. Sample maximum $x_{(n)}$
  - Other statistics

    Sample range, $R = x_{(n)} - x_{(1)}$; $IQR = q_3 - q_1$; Midrange $= \frac{x_{(1)} + x_{(n)}}{2}$
  - Boxplots: Visual of 5-number summary, also used to identify outliers

    Suspected outlier $\rightarrow$ Below $q_1 - 1.5 \times IQR$ (low outlier) or above $q_3 + 1.5 \times IQR$

    Outlier $\rightarrow$ Below $q_1 - 3 \times IQR$ (low outlier) or above $q_3 + 3 \times IQR$

- Another way to identify outliers: Three-sigma rule

  Outlier if outside $(\bar{x} - 3s, \bar{x} + 3s)$
- q–q plots can be used to test potential models

Bivariate EDA

- Goal: Examine pairwise relationships between variables
- Visualizing dependence: Scatterplots can be used to look for positive, negative or no association.
- Quantifying linear dependence:

  Sample correlation $r = \dfrac{1}{n-1} \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \, s_y}$

Lecture 4 – Point Estimation (5.8 and 6.4)

Point estimators

- Definition: A point estimator is any function $\hat{\theta} = W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator
- An estimator is a random variable (a function of the sample); an estimate is the realized value of the random variable once data is collected

Evaluate estimators

- Unbiased definition: Point estimator $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$; otherwise it is biased.

  This tells us the mean of a statistic, regardless of n.
- Consistency definition: The property summarized by the WLLN that says if a sequence of the "same" sample quantity approaches a constant as $n \to \infty$, then it is consistent.

  In other words, ff 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.
  - Convergence in probability
    - Definition: A sequence of random variables, $Y_1, Y_2, \ldots$, converges in probability to a random variable Y if, for every $\epsilon > 0$,

      $$\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0 \qquad \text{or, equivalently,} \qquad \lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$$
    - Notation: $Y_n \xrightarrow{p} Y$
  - (Weak) Law of Large Numbers (WLLN)
    - 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 $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$,

      $$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1 \qquad\qquad \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

      that is, $\bar{X} \xrightarrow{p} \mu$.

Method of moments

- Types of moments:
  - k^{th} (population) moment of the distribution (about the origin) $= \mu'_k = E(X^k)$
  - The corresponding sample moment is the average $= m'_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$
- 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 (may have to integrate)
  2. Then estimate the parameter of interest by equating the population moment with the sample moment and solving for the parameter

Maximum Likelihood Estimation

- Needed items:
  - Parameter space: Set of all possible values for $\theta_1, \ldots, \theta_k$ in pdf (or pmf) $f(x \mid \theta_1, \ldots, \theta_k)$
  - Likelihood function: $L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta})$

    Equivalent to the joint pdf or pmf of the data, just with different information considered known.
- MLE 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 $\hat{\theta}(\mathbf{X})$.
- 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 $\hat{\theta}_{MLE}$.
- Finding MLEs for functions of parameters

  Invariance property of MLEs: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$

Lecture 5 – The Central Limit Theorem (5.6 and 5.7)

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: $Y_n \xrightarrow{d} Y$).

CLT

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

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0, 1) \quad \text{as } n \to \infty$$

- Normal mgf theorem: If $Z \sim \text{N}\,(\mu = 0, \sigma^2 = 1)$, and μ and $\sigma > 0$ are constants, then

  $X = \sigma Z + \mu \sim \text{N}\,(\mu, \sigma^2)$
- Results of CLT
  (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.

t, Z, and the CLT

- If $X_1, \ldots, X_n$ are a random sample for a $N(\mu, \sigma^2)$, as $n \to \infty$, $t_{n-1} \xrightarrow{d} Z$
- If $X_1, \ldots, X_n$ are not normal random variables, when the sample size is large

  $$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{approx}{\sim} \text{Normal}\,(0, 1) = Z \quad \text{by CLT}$$

Normal approximation to discrete distributions

- Continuity correction: If $X \sim \text{Discrete}$ with corresponding $S \sim \text{Normal}$ by the CLT, then for integers $a \leq b$:

  $P(X = a) = P(a - 0.5 \leq S \leq a + 0.5)$ and $P(a \leq X \leq b) = P(a - 0.5 \leq S \leq b + 0.5)$
- Normal approximation to binomial
  - Result: If $X \sim \text{Binomial}\,(n, p) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = np, \sigma^2 = npq)$
  - Conditions: $np \geq 5$ and $nq = n(1 - p) \geq 5$
- Normal approximation to Poisson
  - Result: If $X \sim \text{Poisson}\,(\lambda) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = \lambda, \sigma^2 = \lambda)$
  - Condition: $\lambda \geq 10$

Central interval probabilities

- Empirical rule: If $X \overset{approx}{\sim}$ Normal, then
  1. Approximately 68% of data is within $\mu \pm \sigma$
  2. Approximately 95% of data is within $\mu \pm 2\sigma$
  3. Approximately 99.7% of data is within $\mu \pm 3\sigma$

<u>**Lecture 6 – Confidence Intervals**</u> (7.1 - 7.4)

Interval estimators / confidence intervals

- Definition: An interval estimator or confidence interval for how to calculate endpoints of an interval from sample data: $[L(\mathbf{X}), U(\mathbf{X})]$

  Once $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made.
- Goals of CIs: (1) Capture the target parameter θ (2) Be relatively narrow
- Confidence coefficients definition:

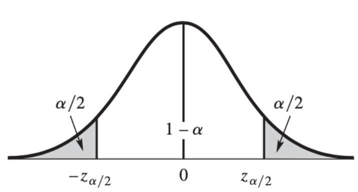
  Probability that a CI captures $\theta \rightarrow P\big(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\big) = 1 - \alpha$ for significance level α
- $100(1-\alpha)\%$ CI for $\theta = [L(\mathbf{X}), U(\mathbf{X})]$

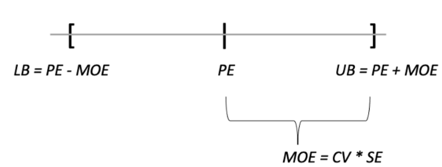
Constructing confidence intervals

- Setup: $\hat{\theta}$ = unbiased point estimator for parameter θ;
  $\sigma_{\hat{\theta}}$ = standard deviation of the sampling distribution of $\hat{\theta}$ (i.e. standard error of $\hat{\theta}$)

  If $\hat{\theta} \sim \text{Normal}\,(\theta, \sigma_{\hat{\theta}})$ (or approximately normal) $\Longrightarrow Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \sim \text{Normal}\,(0,1)$
- To find interval for θ with confidence coefficient equal to $1-\alpha$, need critical values $-z_{\alpha/2}$ and $z_{\alpha/2}$ such that $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$. Then

$$
\begin{aligned}
1 - \alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\
&= P(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}) \\
&= P(\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}) \\
&\Longrightarrow 100(1-\alpha)\% \text{ CI } = [\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}] \\
&= \hat{\theta} \pm z_{\alpha/2}\,\sigma_{\hat{\theta}}
\end{aligned}
$$

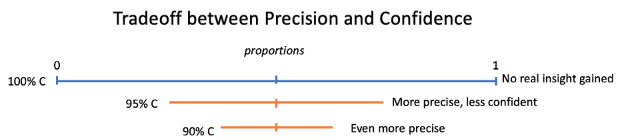




- Summary of CIs
  - Point Estimate (PE) is the best guess; at the center of the interval.
  - Margin of Error (MOE) = Critical Value (CV) $\times$ Standard Error (SE).

    SE (standard deviation of the statistic) measures sampling error.

    % Confident is determined by confidence level set and incorporated via the critical value (CV).
- All else equal, here is how the researcher can affect the precision of intervals:
  - Larger sample size $n \to$ smaller interval (smaller SE)
  - More confident $\to$ larger interval (larger CV)



- Interpretation general structure:

  I am <u>% confident</u> that the true/population <u>parameter + context</u> is between <u>lower bound</u> and <u>upper bound</u>.

Types of intervals

- Variables that affect the form of 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 confidence intervals

  If n is large $\Longrightarrow \hat{\theta} \overset{approx}{\sim} \text{Normal}\,(\theta, \sigma_{\hat{\theta}}) \Longrightarrow 100(1-\alpha)\%$ CI $= \hat{\theta} \pm z_{\alpha/2}\,\sigma_{\hat{\theta}}$

  Conditions: for means $n_i \geq 30$; for proportions $n_i p_i \geq 5$ and $n_i(1-p_i) \geq 5$

θ	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	
μ	$\bar{X}$	$\dfrac{\sigma}{\sqrt{n}}$	Estimate σ^2 with s^2 if unknown
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$	Estimate σ_i^2 with s_i^2 if unknown
p	$\hat{p}$	$\sqrt{\dfrac{p(1-p)}{n}}$	Estimate p with $\hat{p}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$	Estimate p_i with $\hat{p}_i$

  If starting from $X_i \sim$ Normal and known variances, then intervals are exact; if not then $X_i \approx$ Normal by CLT and confidence coefficients are approximate.

- Small sample confidence intervals for means ($n_i < 30$)

  If n is small $\Longrightarrow$ $100(1-\alpha)\%$ CI $= \hat{\theta} \pm t_{\alpha/2}\, \sigma_{\hat{\theta}}$ $\qquad$ t crit values $>$ z crit values

  Conditions: for one sample $X \sim$ Normal with unknown σ^2; for two samples $X_1 \perp\!\!\!\perp X_2$ and $X_1, X_2 \sim$ Normal with unknown common variance σ^2

Parameter	*Confidence Interval* ($\nu = df$)
μ	$\overline{Y} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$, $\quad \nu = n - 1.$
$\mu_1 - \mu_2$	$(\overline{Y}_1 - \overline{Y}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where $\nu = n_1 + n_2 - 2$ and $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ (requires that the samples are independent and the assumption that $\sigma_1^2 = \sigma_2^2$).

  If not starting from $X_i \sim$ Normal then confidence coefficients are approximate and work well as long as not badly skewed with outliers.

- Dependent samples CI for $\mu_1 - \mu_2$

  Simplifies to a one sample CI for the differences $\mu_1 - \mu_2 = \mu_D$ shown above if n is small

- One-sided CI

  Lower bound (at least)

  $P(\hat{\theta} - z_\alpha\, \sigma_{\hat{\theta}}) = 1 - \alpha$

  $\Longrightarrow [\hat{\theta} - z_\alpha\, \sigma_{\hat{\theta}}, \infty)$

  Upper bound (at most)

  $P(\hat{\theta} + z_\alpha\, \sigma_{\hat{\theta}}) = 1 - \alpha$

  $\Longrightarrow (-\infty, \hat{\theta} + z_\alpha\, \sigma_{\hat{\theta}}]$

Margin of error (MOE) revisited

- $MOE = \frac{UB-LB}{2} = \frac{Width}{2} \quad \rightarrow \quad Width = 2 \times MOE$
- The **error in estimation ϵ** is the distance between an estimator and its target parameter:

  $[\hat{\theta} - \epsilon, \hat{\theta} + \epsilon] \Longrightarrow |\hat{\theta} - \theta| = \epsilon$

Finding minimum sample size

- 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 $\hat{\theta} \pm \epsilon$, then for
  - One mean μ with $V(X) = \sigma^2$ known and $X \sim$ Normal or assume going to have "large" n:

    $$n \geq \frac{z_{\alpha/2}^2 \sigma^2}{\epsilon^2}$$

    If σ^2 is unknown, use best approximation available.
  - One proportion p: $n \geq \dfrac{z_{\alpha/2}^2\, p^*(1-p^*)}{\epsilon^2}$

    If there is prior knowledge, use $p^* = \hat{p}$, else set $p^* = 0.5$

<u>Lecture 7 – Hypothesis Tests</u> (8.1 - 8.3)

Hypothesis test

- Definition: A 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_A in favor of H_A.

- Elements of a hypothesis test

  1. Null hypothesis H_0 and Alternative hypothesis H_A

     Definitions:

     – Hypotheses are statements about population parameters

     – The Null hypothesis H_0 is an assumption about θ that is assumed to be true

     – The Alternative hypothesis H_A is the complement of H_0

  2. Test statistic (TS) and Rejection Region RR

     TS: Function of the sample $W(X_1, \ldots, X_n)$, think of this as the point estimator $\hat{\theta}$

     RR: Subset of the sample space (range of sample) for which H_0 will be rejected

  3. Conclusion / interpretation

     General structure: Because our test statistic <u>(COMPARISON of TS and RR)</u> <u>(IS / IS NOT)</u> in the rejection region we <u>(REJECT or FAIL TO REJECT)</u> the null hypothesis.
     At the <u>(ALPHA)</u> significance level, there <u>(IS or IS NOT)</u> sufficient evidence to conclude <u>(THE ALTERNATIVE HYPOTHESIS)</u>.

Large sample tests

- If n is large, then $\hat{\theta} \sim \text{Normal}\,(\theta, \sigma_{\hat{\theta}})$ (or approximately normal) $\Longrightarrow Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \sim \text{Normal}\,(0,1)$

- Using the same parameters θ, point estimates $\hat{\theta}$, and standard errors $\sigma_{\hat{\theta}}$ as shown in confidence intervals, all of the large sample α-level tests can be summarized with

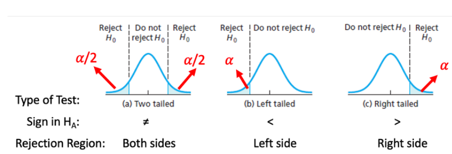
Large-Sample α-Level Hypothesis Tests

$H_0 : \theta = \theta_0$.

$$H_a : \begin{cases} \theta > \theta_0 & \text{(upper-tail alternative).} \\ \theta < \theta_0 & \text{(lower-tail alternative).} \\ \theta \neq \theta_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$.

$$\text{Rejection region: } \begin{cases} \{z > z_\alpha\} & \text{(upper-tail RR).} \\ \{z < -z_\alpha\} & \text{(lower-tail RR).} \\ \{|z| > z_{\alpha/2}\} & \text{(two-tailed RR).} \end{cases}$$



- For proportions

  – One sample: In the standard error, use $p_0 \Longrightarrow \sigma_{\hat{p}} = \sqrt{\dfrac{p_0(1-p_0)}{n}}$.

  – Two sample: In the standard error, use $p_1 = p_2 = p$ and estimate with $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} \Longrightarrow \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})[1/n_1 + 1/n_2]}$.

- In any particular test, only one of the listed alternatives H_A is appropriate, which will be based on the research question. Then use the corresponding rejection region.

Small sample tests for means

- If n is small, then need to switch to t-tests. For these we start with $X \sim$ Normal
- 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: $T = \dfrac{\overline{Y} - \mu_0}{S/\sqrt{n}}$.

$$\text{Rejection region: } \begin{cases} t > t_\alpha & \text{(upper-tail RR).} \\ t < -t_\alpha & \text{(lower-tail RR).} \\ |t| > t_{\alpha/2} & \text{(two-tailed RR).} \end{cases} \quad t_\alpha\text{, with df} = n-1$$

- 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 have paired t-test

  $\mu_1 - \mu_2 = \mu_D \qquad \rightarrow \qquad \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} = T \sim \mathrm{t}_{n-1}, \qquad$ one sample test on differences

  If not starting from $X_i \sim$ Normal then t-tests are approximately α-level and work well as long as not badly skewed with outliers.

p-values

- 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.
- Two ways to make conclusion for hypothesis tests:

  Traditional method: $TS \overset{?}{\in} RR$

  p-value method: Reject H_0 if p-value $\leq \alpha$ and Fail to reject H_0 if p-value $> \alpha$

Relationship between confidence intervals and hypothesis tests

- Confidence interval = Acceptance region = Complement of RR
- Decisions based on CI: For $H_0 : \theta = \theta_0$ and

  Two-tailed $H_A : \theta \neq \theta_0$: "Accept" H_0 if θ_0 falls within the $100(1-\alpha)\%$ CI and reject if outside.

  Right-tailed $H_A : \theta > \theta_0$: Reject if outside lower bound CI

  Right-tailed $H_A : \theta < \theta_0$: Reject if outside upper bound CI

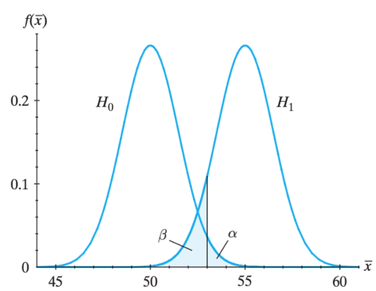
Type I and Type II errors

- Type I: Incorrectly rejecting H_0

  $\alpha = P(\text{Type I error}) = P(\text{Reject when } H_0 \text{ is true}) = P(TS \in RR \mid H_0)$
- Type II: Incorrectly failing to reject H_0

  $\beta = P(\text{Type II error}) = P(\text{Fail to reject when } H_0 \text{ is false}) = P(TS \notin RR \mid H_A)$

Example for one mean



- Power: Correctly rejecting H_0

  Power $= 1 - \beta = P(\text{Reject } H_0 \text{ when } H_0 \text{ is false}) = P(TS \in RR \mid H_A)$

Lecture 8 – Regression (new textbook)

Initial model statement: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- Pieces
  - Y_i: Dependent (or response) variable value.
  - X_i: Independent (or predictor) variable value.
  - 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 $i, j : i \neq j$.
  - 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.
  - σ^2: Error variance.
- Implications
  - Mean of Y_i for given $X_i \to E(Y_i) = \beta_0 + \beta_1 X_i$
  - Variance of Y_i for given $X_i \to V(Y_i) = \sigma^2$

Estimation of the regression function

- Method of least squares
  - Goal: Minimize function of errors $Q = \displaystyle\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$
  - Results:

$$
\begin{array}{rcccccc}
\text{Intercept} & \hat{\beta}_0 & = & \frac{1}{n}\sum Y_i + \hat{\beta}_1 \frac{1}{n}\sum X_i & = & \bar{Y} - \hat{\beta}_1 \bar{X} & \\
\text{Slope} & \hat{\beta}_1 & = & \dfrac{\sum X_i Y_i - \frac{1}{n}\sum X_i Y_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2} & = & \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} & = \dfrac{S_{XY}}{S_{XX}}
\end{array}
$$

Residuals and estimating error variance

- Residuals are the known, observable estimate of the unobservable model error: $\hat{\epsilon}_i = e_i = Y_i - \hat{Y}_i$

  These are very useful for studying whether the given regression model is appropriate for the data.
- Estimating error variance:

  Error (residual) mean square $\quad S^2 = MSE = \dfrac{SSE}{n-2} = \dfrac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2} = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2}$
  - Unbiased: $E(MSE) = \sigma^2$

  and $S = \sqrt{MSE}$

Normal error regression model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \text{where} \quad \epsilon_i \overset{iid}{\sim} \text{Normal}\,(0, \sigma^2)$
- MLEs of regression parameters and error variance

Parameter	MLE	
β_0	$\hat{\beta}_0$	Same as LSE
β_1	$\hat{\beta}_1$	Same as LSE
σ^2	$\hat{\sigma}^2 = \dfrac{\sum(Y_i - \hat{Y}_i)^2}{n}$	

Inference (tests on β_1)

- Sampling distribution of standardized $\hat{\beta}_1$: $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1}$

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{XX}}} \sim \text{t}_{\,n-2}$$

- 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

- Coefficient of determination R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}, \qquad \text{range:} \quad 0 \leq R^2 \leq 1$$

Diagnostics

- Diagnostics = assumption checking

  Residuals should model the properties of $\epsilon_i \overset{iid}{\sim}$ Normal $(0, \sigma^2)$

  - LINE → Linearity, Independence, Normality, Equal variance

- Ideal residual plots if all assumptions on the error terms are met: