W203 Unit Summary

Unit 1. Descriptive Statistics

Terms and Definitions

Population: a well-defined collection of objects in an area of interest (investigation)

Census: an occurance of availability of desired information for all objects in the population

Sample: a subset of the population, selected in some prescribed manner

Characteristic: an attribute of the data, which may be categorical or numerical

Variable: any characteristic whose value may change from one object to another in the population

Univariate: a data set that consists of observations on a single variable

Multivariate: a data set that consists of observations on multiple variables

Bivariate: a data set that consists of observations on two variables specifically

Descriptive Statistics: the branch of statistics that summarizes and describes important features of the data

Inferential Statistics: the branch of statistics that draws a form of conclusion about the population from the sample

Probability: the bridge between descriptive and inferential techniques. In probability, properties of the population are assumed known, and questions are posed and answered from a representative sample.

Discrete and Continuous Variables: A numerical variable is *discrete* if its set of possible values is at most countable. A numerical value is *continuous* if its set of possible values is an uncountable set.
Probability: population → sample
Stats: sample → population

Measures of Location

For observations $x_1, x_2, \cdots, x_n$

Sample Mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Sample Median: $\tilde{x} = \begin{cases} \left(\frac{n+1}{2}\right)^{th} \text{ if n is odd} \\ \text{avg of } \left(\frac{n}{2}\right)^{th} \& \left(\frac{n}{2}+1\right)^{th} \text{ if n is even} \end{cases}$

Trimmed Mean: a compromise between $\tilde{x}$ and $\bar{x}$ by removing a percentage of the smallest and largest observations

Measures of Variability

Range: the difference betwen the largest and smallest sample variables

Sample Variance: $\sigma^2 = \frac{S_{xx}}{n-1} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

Sample Standard Deviation: $\sigma = \sqrt{\sigma^2}$

Properties: Let $x_1, x_2, \ldots, x_n$ be a sample and c be any nonzero constant.
1. If $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $\sigma_y^2 = \sigma_x^2$ and
2. If $y_1 = cx_1, \ldots, y_n = cx_n$, then $\sigma_y^2 = c^2 s_x^2$, $\sigma_y = |c| s_x$
where σ_x^2 is the sample variance of the x's and σ_y^2 is the sample variance of the y's.

Box Plots

Upper Fourth: the median of the largest half of observations

Lower Fourth: the median of the smallest half of observations

Fourth Spread: f_s = upper fourth - lower fourth

Outlier: any observation farther than $1.5f_s$ from the closest fourth (considered mild if less than $3f_s$)

Extreme Outlier: an observation farther than $3f_s$ from the nearest fourth

Sample Covariance: The simplest measure of association between two variables.

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_1 - \bar{y})}{n-1}$$

Sample Correlation: Measures the linear relationship between two variables.

$$r_{x,y} = \frac{cov(x, y)}{s_x s_y}$$

Properties:
1. Adding c to variable x or variable y doesn't change $r_{x,y}$
2. Multipling a non-zero c to variable x or variable y doesn't change $r_{x,y}$

Unit 3. Probability Theory

Terms and Definitions

Experiment: any activity or process whose outcome is subject to uncertainty

Sample Space: denoted by $\mathcal{S}$, is the set of all possible outcomes of an experiment

Event Space: denoted by $\mathcal{F}$, is the subset of outcomes from the sample space. It includes subsets of $\mathcal{S}$, the empty set $\emptyset$ and $\mathcal{S}$ itself.

Axioms and Properties of Probability

Probability Rule is a function, P, from the set of events to the real numbers

Probability Space is the triple of $(\mathcal{S}, \mathcal{F}, P)$]

Axioms of Probability

1. $P(A) >= \emptyset$ for any event A in $\mathcal{F}$
2. $P(\mathcal{S}) = 1$
3. For any countably infite set of disjoint events $\{A_1, A_2, \ldots\}$, $P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$

Properties of Probability
For any events A, B and C

1. $P(A) + P(!A) = 1$, from which $P(A) = 1 - P(!A)$
2. $P(A) \leq 1$
3. if $A \subset B$, then $P(A) \leq P(B)$

   Addition Rule:
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(A \cup B \cup C) =$
   $P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$
6. $P(A \cup B \cup C) = P(A) + P(!A \cap B) + P(!A \cap !B \cap !C)$

Counting

Discrete Uniform Probability Law If the sample space consists of n possible outcomes which are equally likely, then the probability of any event A is given by $P(A) = \frac{\text{number of elements of } A}{n} = \frac{k}{n}$

Product Rule for k-Tuples: In an ordered collection of k elements, and the first element can be selected in n_1 ways, and the second in n_2 ways, and so on, then there are $n_1 n_2 \cdots n_k$ possible k-tuples.

Permutations: denoted by $P_{k,n}$, is an `ordered` subset containing the number of permutations of size k that can be formed from a set of n elements

$$P_{k,n} = (n)(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

Combinations: denoted by $\binom{n}{k}$, is an `unordered` subset containing the number of permutations of size k that can be formed from a set of n elements $\binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$

Permutations: denoted by $\binom{n}{n_1, n_2, \cdots, n_r}$, is a partition of n objects into r groups, with the ith group having n_i objects

$$\binom{n}{n_1, n_2, \cdots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

Independence

Disjoint Events - **A** and **B** are disjoint when they cannot happen simultaneously, or

$$P(\mathbf{A} \cap \mathbf{B}) = 0$$
$$\mathbf{A} \cap \mathbf{B} = \emptyset$$

Independent Events If events A and B are `independent` from one another, then probability of A given B is the same as the probability of A, e.g. $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Multiplication Rule:

1. For any two independent events A and B, $P(A \cap B) = P(A) \cdot P(B)$
2. This can be generalized to any n mutually independent events.

Conditional Independence A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$. Conditional independence does not imply independence, and independence does not imply conditional independence.

Conditional Probability

Joint Probability
$P(A \cap B)$ or $P(A, B)$ – Probability of A and B.

Marginal (Unconditional) Probability
$P(A)$ – Probability of A.

Conditional Probability
$P(A|B) = \frac{P(A \cap B)}{P(B)}$ – Probability of A, given that B occurred.

Multiplication Rule:

1. For any two events A and B, $P(A \cap B) = P(B) \cdot P(A|B)$
2. For any three events A, B, and C, $P(A \cap B \cap C) = P(B) \cdot P(A|B) \cdot P(C|A \cap B)$

Conditional Probability *is* Probability
$P(A|B)$ is a probability function for any fixed B. Any theorem that holds for probability also holds for conditional probability.

Law of Total Probability (LOTP)
Let $A_1, A_2, \ldots, A_k$ be mutually exclusive and exhaustive events (that partition the sample space). Then for any other event B

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)$$
$$= \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

Using Complement and Multiplication Rules

1. $P(B) = P(B|A) \cdot P(A) + P(B|!A) \cdot P(!A)$
2. $P(B) = P(B \cap A) + P(B \cap !A)$
3. $P(A \cap B) = P(A|!B)P(!B)$

Bayes' Rule

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

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} \text{ for } j = 1, \cdots, k$$

Unit 4.1 Discrete Random Variables

Terms and Definitions

Discrete Set: is a set of disconnected points with no continuous intervals

$$O = \{0, 1, 2, \ldots\}$$

Discrete Random Variable (d.r.v.): is a function whose domain is the sample space $\mathcal{S}$ and whose range is the discrete set of real numbers in the probability space $(\mathcal{S}, \mathcal{F}, P)$

$$X : \mathcal{S} \to \mathbb{R}$$

$$X(\omega) = x$$

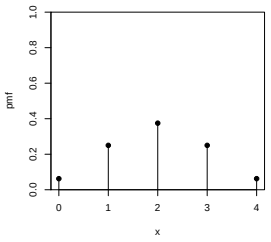
Bernoulli Random Variable: any random variable whose only possible values are 0 and 1

Bernoulli Distribution: Binary outcomes with values $O = \{0, 1\}$ where 1 represents success and has probability p

Probability Distribution or Probability Mass Function (pmf): The pmf p of a d.r.v. X in the problem space $(\mathcal{S}, \mathcal{F}, P)$ is described by

$$p_X(x) = P(X = x) = P(\{w \in \mathcal{S} : X(\omega) = x\})$$

where the outcomes $O_x = \{x \in \mathbb{R} : p_X(x) > 0\}$



The PMF satisfies

$$p_X(x) \geq 0 \text{ and } \sum_x p_X(x) = 1$$

Parameter: a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution in $p_X(x; \alpha)$

Family of probability distributions: The collection of all probability distributions for different values of the parameter

Example of a family of distributions for a *bernoulli random variable*, X, with $O_x = \{0, 1\}$

parameter for $P(X = 1) = \alpha = p_X(x; \alpha)$

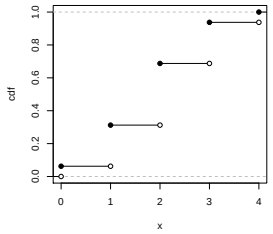
parameter for $P(X = 0) = 1 - \alpha = p_X(x; \alpha)$

family of distributions for $p_X(x; \alpha) = \begin{cases} 1-\alpha & ; \text{ x=0;} \\ \alpha; & ; \text{ x=1;} \\ 0 & ; \text{ otherwise} \end{cases}$

Cumulative Distribution Function (cdf) : The cdf F of a d.r.v. X with probability mass function p_X is described by

$$F_X(x) = P(X \leq x) = \sum_{y \in O_x : y \leq x} p_X(y) \text{ (the sum of probability masses)}$$

For any number x, $F_X(x)$ is the probability that the observed value of X will be at most x.



The CDF is an increasing, right-continuous function with

$$F_X(x) \to 0 \text{ as } x \to -\infty \text{ and } F_X(x) \to 1 \text{ as } x \to \infty$$

Independence Two random variables are independent if knowing the value of one gives no information about the other. Discrete r.v.s X and Y are independent if for *all* values of x and y

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Expectation

Expected Value (a.k.a. *mean*, *expectation*, or *average*) is a weighted average of the possible outcomes of our random variables.
Mathematically, if $x_1, x_2, x_3, \ldots$ are all of the distinct possible values that X can take, the expected value of X is

$$E(X) = \mu_X = \sum_i x_i P(X = x_i) = \sum_i x_i p_X(x_i)$$

X	Y	$X+Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	−3	−2
1	0	1
5	9	14
4	1	5
...	...	...
$\frac{1}{n}\sum_{i=1}^{n} x_i$	+ $\frac{1}{n}\sum_{i=1}^{n} y_i$	= $\frac{1}{n}\sum_{i=1}^{n}(x_i + y_i)$
$E(X)$	+ $E(Y)$	= $E(X+Y)$

Functions of Random Variables Let X be a d.r.v with pmf p and let g be a function from the real numbers to the real numbers, $g : \mathbb{R} \to \mathbb{R}$. The function g on X, $g(X)$, is also a r.v. with pmf $p(x)$

$$E[g(x)] = \sum g(x)P[g(X) = x] = \sum_x g(x)p(x)$$

In addition, it can be shown that

$$E[g(x)] = g[E(X)]$$

Linearity For any r.v.s X and Y, and constants a, b, c,

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

Short proof: Let $g(X) = aX + b$

$$E[g(X)] = \sum_{x \in O_x}(ax + b)p(x) = \sum_x axp(x) + bp(x)$$
$$= a\sum xp(x) + b\sum p(x) = aE(X) + b = g[E(X)]$$

Uniform Random Variables For r.v. X where all outcomes have the same probability, $p(x_i) = p(x_j) \forall i, j \in k$

$$E(X) = \tfrac{1}{k}\sum_j x_j = \tfrac{k+1}{2}$$

Bernoulli Random Variables For a Bernoulli r.v. X having binary outcomes, where α is the probability $X = 1$ and $1 - \alpha$ is the probability $X = 0$,

$$p(x; \alpha) = \begin{cases} 1-\alpha & ; \text{ x=0;} \\ \alpha; & ; \text{ x=1;} \\ 0 & ; \text{ otherwise} \end{cases}$$

$$E(X) = \sum_{x \in 0,1} x \cdot p(x) = 0 \cdot (1 - \alpha) + 1 \cdot \alpha = \alpha$$

Geometric Random Variables Let X be geometric r.v. for the number of babies born until the first girl, where G is the event the baby is a girl, $P(G_i) = g \forall i$

$$f(1) = P(G_1) = g$$
$$f(2) = P(!G_1 \cap G_2) = (1 - g) \cdot g$$
$$f(3) = (1 - g)^2 \cdot g$$
$$f(x; g) == \begin{cases} (1-g)^{x-1} \cdot g & ; \quad x \in \{1, 2, \ldots\}; \\ 0 & ; \text{ otherwise} \end{cases}$$

$$E(X) = \sum_{x \in \{1,2\ldots\}} x \cdot (1 - g)^{x-1} \cdot g = \tfrac{1}{g}$$

Same distribution implies same mean If X and Y have the same distribution, then $E(X) = E(Y)$ and, more generally,

$$E(g(X)) = E(g(Y))$$

Conditional Expected Value is defined like expectation, only conditioned on any event A.

$$E(X|A) = \sum_x xP(X = x|A)$$

Variance and Standard Deviation Let X be a random variable and let $\mu = E(X)$.
The variance of X is defined as $var(X) = E[(X - \mu)^2]$
The standard deviation of X is defined as $\sigma_x = \sqrt{var(x)}$

$$var(X) = E[X^2 - 2X\mu + \mu^2] = E(x^2) - E(2X\mu) + E(\mu^2)$$
$$= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2$$
$$= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$$

Properties
Let $c \in \mathbb{R}$.

1. $var(X + c) = E[(X + c - E(X + c))^2] =$
   $E[(X + c - E(X) - c)^2] = var(X)$
2. $var(cX) = E[(cX - E(cX)^2] = E[(cX - cE(X))^2] =$
   $E[c^2(X - E(ZX))^2] = c^2E[(x - E(x))^2] = c^2var(X)$

Discrete Distributions

Binomial Probability Distribution

Binomial Experiment is defined by the following characteristics:

1. The experiment consists of a sequence of n smaller experiments called *trials*, where n is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes (dichotomous trials), which we generically denote by success (S) and failure (F).
3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
4. The probability of success P(S) is constant from trial to trial; we denote this probability by p.

Rule Consider sampling without replacement from a dichotomous population of size N. If the sample size (number of trials) n is at most 5% of the population size, the experiment can be analyzed as though it were exactly a binomial experiment.

Binomial Random Variable A binomial r.v. X associated with a binomial experiment consisting of n trials with success probability p is defined as

$$X \sim Bin(n, p)$$

pmf of a Binomial r.v. Because the pmf of a binomial rv X depends on the two parameters n and p, the pmf is denoted by $b(x; n, p)$.

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & ; \quad x = 0, 1, 2, \ldots, n \\ 0 & ; \quad otherwise \end{cases}$$

cdf of a Binomial r.v. For $X \sim Bin(n, p)$, the cdf is denoted by

$$B(x; n, p) = P(X \le x) = \sum_{y=0}^{x} b(y; n, p) \quad x = 0, 1, \ldots, n$$

Mean, Variance and SD of Binomial r.v. If $X \sim Bin(n, p)$, then

$$E(X) = np,\ V(X) = np(1-p) = npq, \text{ and}$$
$$\sigma_x = \sqrt{npq} \text{ (where } q = 1 - p)$$

Unit 4.2 Continous Random Variables

Terms and Definitions

Continuous Set: is a set of connected points with a continuous interval

Continuous Random Variable (c.r.v.): is a function whose domain is the sample space $\mathcal{S}$ and whose range is a continuous interval or set of intervals of real numbers in the probability space $(\mathcal{S}, \mathcal{F}, P)$, , but the probability of any one value is zero.

$$X : \mathcal{S} \to \mathbb{R}$$

$$X(\omega) = 0$$

Probability Density Function (pdf): The pdf f of a c.r.v. X in the problem space $(\mathcal{S}, \mathcal{F}, P)$ is described by

$$P(a \le X \le b) = \int_{x=a}^{b} f(x)dx$$

Cumulative Distribution Function (cdf): The cdf $F(x)$ for a c.r.v X, is the area under f from $-\infty$ to x:

$$F(x) = P(X \le x) = \int_{y=-\infty}^{x} f(y)dy$$

Properties

1. The derivative of the cdf F is the pdf f

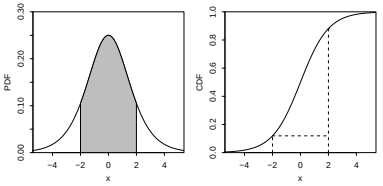
$$F'(x) = f(x)$$

2. A pdf is nonnegative and integrates to 1.

$$\int_{y=-\infty}^{\infty} f(y)dy = 1$$

3. By the fundamental theorem of calculus, we can integrate pdf to get back to cdf:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



To find the probability that a c.r.v. takes on a value in an interval, integrate the pdf over that interval.

$$F(b) - F(a) = \int_{a}^{b} f(x)dx$$

Expectation

Expected value of an r.v. The expected value of X is defined as:

$$E(X) = \sum_{x} xP(X = x) \text{ (for discrete } X)$$

Expected value of a function of an r.v. A function of a r.v. is also a r.v. If $h : \mathbb{R} \to \mathbb{R}$ then $h(x)$ is a random variable and

$$E(h(x)) = \int_{x=-\infty}^{\infty} h(x)f(x)dx$$

The **Law of the Unconscious Statistician (LOTUS)** states that you can find the expected value of a *function of a random variable*, $g(X)$, in a similar way, by replacing the x in front of the PMF/PDF by $g(x)$ but still working with the PMF/PDF of X:

$$E(g(X)) = \sum_{x} g(x)P(X = x) \text{ (for discrete } X)$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \text{ (for continuous } X)$$

Properties

1. if $h(x) = aX + b$ then $E(h(x)) = aE(X) + b$

Variance and Standard Deviation Let X have pmf $p(x)$ and expected value μ. Then the $V(X)$ or σ_X^2 is

$$V(X) = \sum_{D} (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The standard deviation (SD) of X is $\sigma = \sqrt{\sigma}$
Alternatively,

$$V(X) = \sigma^2 = [\sum_{D} x^2 \cdot p(x)] - \mu^2 = E(X^2) - [E(X)]^2$$

Properties

1. $V(aX + b) = a^2 \cdot \sigma^2$
2. In particular, $\sigma_{aX} = |a| \cdot \sigma_x$
3. $\sigma_{X+b} = \sigma_X$

Continuous Distributions

Uniform Distribution

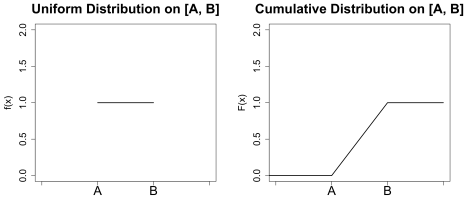
pdf of Uniform Random Variable X has a uniform distribution on $[A, B]$ if it has probably density function

$$f(x) = \begin{cases} \frac{1}{B-A} & ; \quad A \le x \le B \\ 0 & ; \quad \text{otherwise} \end{cases}$$

When X is between A and B,

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_{A}^{x} \frac{1}{B-A}dy = \frac{y}{B-A}\Big|_{a}^{x} = \frac{x-A}{B-A}$$

$$F(x) = \begin{cases} 0 & ; \quad \text{x<A;} \\ \frac{x-A}{B-A}; & ; \quad A \le x \le B; \\ 1 & ; \quad B \le x \end{cases}$$



Expectation

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{A}^{B} x \cdot \frac{1}{B-A}dx$$
$$= \frac{x^2}{2 \cdot (B-A)}\Big|_{A}^{B} = \frac{B^2 - A^2}{2 \cdot (B-A)} = \frac{A+B}{2}$$

Variance

$$Var(X) = E(X^2) - [E([X)]^2$$

$$[E([X)]^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx = \int_{A}^{B} x^2 \cdot \frac{1}{B-A}dx$$
$$= \frac{x^3}{3 \cdot (B-A)}\Big|_{A}^{B} = \frac{B^3 - A^3}{3 \cdot (B-A)} = \frac{A^2 + 2AB + B^2}{3}$$

$$Var(X) = \frac{A^2+2AB+B^2}{3} - (\frac{A+B}{2})^2 = \frac{A^2+2AB+B^2}{3} - \frac{A^2+2AB+B^2}{4}$$
$$= \frac{A^2-2AB+B^2}{12} = \frac{(B-A)^2}{12}$$

Normal Distribution

Normal Distribution Random Variable The statement that X is normally distributed with parameters μ and σ^2 is often abbreviated $X \sim N(\mu, \sigma^2)$.

pdf of a Normal Distribution r.v. A c.r.v X is said to have a normal distribution with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$ when the pdf of X is:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$$

Properties

1. μ, the location parameter, is both the mean and the median.
2. σ, the scale parameter, stretches the curve horizontally.

Mean, Variance and SD of a Normal Distribution r.v. It can be shown that $E(X) = \mu$ and $V(X) = \sigma^2$, so the parameters are the mean and the standard deviation of X.
Properties

1. 68% of the area is within 1 standard deviation of the mean.

2. 95% of the area is within 2 standard deviations of the mean.
3. 99.7% of the area is within 3 standard deviations of the mean.

Standard Normal Distribution defined as Z, or $\phi(z)$, is when $\mu = 0$ and $\sigma = 1$

$$\text{pdf}: \quad \phi(z) = \tfrac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{cdf}: \quad \Phi(z) = \int_{-\infty}^{z} \phi(u)du$$

Standardizing a Normal Variable is when we take a normal variable that is not standard normal, and adjust it. The new random variable becomes $\frac{(X-\mu)}{\sigma}$, and we can express probabilities involving X in terms of the z-distribution.

Deviations from Normality
Skewness

1. Negative Skew - tail is to the left
2. Positive Skew - tail is to the right

Kurtosis

1. Platykurtic - flatter and smooshed down
2. Leptokurtic - much more bunched together and peek up.

Unit 5: Joint Distributions

Terms and Definitions

Joint Range: Let $X : \mathcal{S} \to \mathbb{D}_1$ and $Y : \mathcal{S} \to \mathbb{D}_2$ be 2 rvs with a common sample space. We define the joint range of the vector (X, Y) of the form

$$\mathbb{D} = \mathbb{D}_1 \times \mathbb{D}_2 = \{(x, y) : x \in \mathbb{D}_1, y \in \mathbb{D}_2\}$$

Random Vector: A 2-D random vector(X, Y) is a function from $\mathcal{S} \to \mathbb{R}^2$. It is defined $\forall \omega \in \mathcal{S}$ such that

$$(X, Y)(\omega) = (X(\omega), Y(\omega)) = (x, y) \in \mathbb{D}$$

Joint Probability Distribution or Joint Probability Mass Function: For two d.r.v.'s X and Y. The joint pmf of (X, Y) is defined $\forall (x, y) \in \mathbb{D}$

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

It must be that $p(x, y) \geq 0$ and $\sum_i \sum_j p(x_i, y_j) = 1$.

Marginal Prob Mass Function: of X and of Y, denoted $p_X(x)$ and $p_Y(y)$ respectively,

$$p_X(x) = \sum_{y:p(x,y)>0} p(x, y) \quad \forall x \in \mathbb{D}_1$$

Joint Probability Density Function: For two c.r.v.'s X and Y. The joint pdf of (X, Y) is defined $\forall A \subseteq \mathbb{R}^2$

$$P((X, Y) \in A) = \iint_A f(x, y)dxdy$$

It must be that $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$. Note also that this integration is commutative.

Marginal Prob Density Function: of X and of Y, denoted $f_X(x)$ and $f_Y(y)$ respectively,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \quad \forall x \in \mathbb{D}_1$$

Note that if $f(x, y)$ is the joint density of the random vector (X, Y) and $A \in \mathbb{R}^2$ is of the form $A = [a, b] \times [c, d]$ we have that

$$P((X, Y) \in A) = \int_c^d \int_a^b f(x, y)dxdy = \int_a^b \int_c^d f(x, y)dxdy$$

Independence: Two rvs are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad f(x, y) = f_X(x)f_Y(y)$$

Conditional Distribution(discrete): For two discrete rv's X and Y with joint pmf $p(x_i, y_j)$ and marginal X pmf $p_X(x)$, then for any realized value x in the range of X, the conditional mass function of Y, given that $X = x$ is

$$p_{Y|X}(y|x) = \frac{p(x_i, y_j)}{p_X(x)}$$

Conditional Distribution(continuous): For two continuous rv's X and Y with joint pdf $f(x, y)$ and marginal X pdf $f_X(x)$, then for any realized value x in the range of X, the conditional density function of Y, given that $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Expected Values, Covariance & Correlation

Expected value: The expected value of a function $h(X, Y)$ of two jointly distributed r.v.'s is

$$E(h(X, Y)) = \sum_x \sum_y h(x, y)P(X = x, Y = y) \text{ for discrete r.v.}$$

$$E(h(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)f_{X,Y}(x, y)dxdy \text{ for continuous r.v.}$$

Covariance: Measures the strength of the relation btwn 2 RVs, however very

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Shortcut Formula:

$$Cov(X, Y) = E(XY) - \mu_x\mu_y$$

The defect of the covariance however is that its value depends critically on the units of measurement.

Correlation: Cov after standardization. Helps interpret Cov.

$$\rho = \rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

Has the property that $Corr(aX + b, cY + d) = Corr(X, Y)$
and that for any rvs X, Y $-1 \leq \rho \leq 1$.
Note also that ρ is independent of units, the larger $|\rho|$ the stronger the linear association, considered strong linear relationship if $|\rho| \geq 0.8$.
Caution though: if X and Y are independent then $\rho = 0$ but $\rho = 0$ does not imply that X,Y are independent.
Also that $\rho = 1$ or -1 iff $Y = aX + b$ for some a, b with $a \neq 0$.

Unit 6: Sampling and the Central Limit Theorem

Terms and Definitions

Statistic: Any quantity whose value can be calculated with sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

Sampling Distribution: The probability distribution of a statistic. It describes how the statistic varies in value across all samples that might be selected and how much uncertainty we have in our statistics

1. If the sampling distribution is wide, we could get very different values each time we draw a new sample, so we have little confidence in the numbers we get.
2. If the sampling distribution is narrow, we would get similar values each time we repeat the experiment.
   We only get one number, but we believe that it's representative (i.e., more meaningful).
3. A centerpiece of statistics is estimating how much uncertainty exists within the statistics collected.
   Distinguishes statistics from other fields like machine learning

Unit of analysis: the singular unit that defines all the different cases in a given study.

Population: the entire set of all units of analysis, every single one.

Selected Sample: is who we target and who we try to reach.

Actual Sample: is who responds.

Sampling Design: our specific strategy for obtaining a sample, which is presumably going to be representative of the larger population. Examples include quota sampling, random sampling, snowball sampling, stratified random sampling, and convenient sampling.

Problems with Sampling Design: 1) getting a representative sample 2) one or two outliers skewing the data 3) controlling bias in the sampling method

Important issue in samples: All other things being equal, a small simple random sample of 100 people is preferable to a convenience sample of 10,000 people from the same population. A small random sample will give better estimates of the population despite the smaller sample size.

Examples of Bias: 1) Non-voters who do not respond. 2) Those who acre the most are more likely to respond 3) Interviewer inconsistencies in selecting who to interview 4) Self-selection sampling - waiting for people to decide whether to take it versus recruiting people randomly

The point here is that we need to anticipate bias. We need to find it even when we didn't anticipate it, and we need to deal with it. So it's always better, always better to acknowledge potential bias, rather than to assume it does not exist at all

The chance error: is the stuff that we expect. That's the stuff that we know is part of just doing statistics. We can at least estimate how much error that is, how much chance error there is.

Statistic (what we know) = parameter(what we want to know) + bias + chance error.

Try to use procedures that take biases out of the equation

Sampling Frame: A list of units of analysis from which you take a sample: Directories, Local census, Registered users of an online system

Simple Random Samples The researcher has no discretion over who is included. The procedure for selecting a sample is definite; it involves planned use of chance.

Simple Random Sample: Procedure

1. Requires numbering all potential participants in a given sampling frame (N)
2. Pulls random numbers from any source and uses the random numbers to create the sample (n)
3. Has an equal chance of each value being selected
4. Issues:
   1) True randomization (seeding random number generators)
   2) Replacement in the field (e.g., door-to-door problem)

Stratified Random Sampling:

1. Key issue: representation of salient subpopulations
   Maximizes between-group variance while minimizing within-group variance
2. For proportionate samples:
   1) Do you know the key independent variables?
   2) If not, you may be better off avoiding stratification

3. For disproportionate samples:
   Sample weighting is used because some groups are much smaller or larger than others.
4. Example - Obtaining a sample of participants:
   1) We will attempt to obtain an equal number of survey participants (100 maximum per strata) from five different editing strata.
   2) We oversample those who edit more because they make up a smaller percentage of the population.

Cluster Sampling:

1. Randomly selected clusters are specifically sampled from
2. Used when there are no available sampling frames
3. Based on areas, institutions, or "clusters"

Respondent-Driven Sampling:

1. Combines snowball sampling with a mathematical model that weights the sample
2. Compensates for the nonrandom collection of the sample

Non-probability Sampling:

1. Quotas
   1) Pick key groups of interest and find individuals to fill specific goals.
   2) Quotas are fulfilled without using random sampling.
2. Purposive sampling Find key groups and only study them.
3. Convenience sampling Taking anyone you can get to participate
4. Snowball sampling Finding a starting point and using these individuals to get next participant

Survey Weight: is a numeric value that is assigned to each case in our dataset.

Weighted Value: indicates how much each case (each unit of analysis) will count in any type of statistical analysis or procedure that is run.

Types of Weights:

1. Design weights
   (a) Used to correct for oversampling or undersampling specific cases
   Example: Smaller groups in the population may be oversampled, such as Linux users compared to PC and Mac users.
   (b) Correct proper representation in the population
2. Poststratification weights
   (a) Used to correct for the fact that some types of cases are less likely to show up in the sample
   Example: People of different age groups differ in how likely they are to respond to surveys.
   (b) Used to correct for actual proportions in the population

Weighting With Multiple Characteristics

1. Weight the sample by the first characteristic.
2. Then generate the frequency for the second characteristic from this weighted data.
3. Calculate the second characteristic weight.
4. Generate a final weight by multiplying the two weights.

Random Sampling: The rv's $X_1, X_2, \cdots X_n$ are said to form a (simple) random sample of size n if

1. The Xi's are independent rv's.
2. Every Xi has the same probability distribution.
3. Think of this as an ideal case.
   Real-world sampling may deviate from this, but we usually try to get as close to random sampling as we can.
4. We can often rephrase this by saying the Xi's are independent and identically distributed (iid).
5. If we sample randomly from a population with replacement, we satisfy the conditions for random sampling.
6. If we sample from an infinite population (a generating process that gives us as independent number of independent draws), we also have random sampling.
7. Random sampling is approximately satisfied if sampling is without replacement, yet the sample size n is much smaller than the population size N, and each unit of observation has the same probability of being sampled.
   In practice, as long as we draw 5% of the population or less, and each unit of observation has the same probability of being sampled, we usually proceed as though we have a random sample.

The Distribution of the Sample Mean: Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables with mean μ and deviation σ. Then

1. $E(\bar{X}) = \mu_{\bar{X}} = \mu$
2. $Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$
3. $\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$

Central Limit Theorem: Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\bar{X}$ has approximately a normal distribution with $\mu_{\bar{X}}$ and $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$. The larger the value of n, the better the approximation. When to apply:

1. The degree to which the population distribution is non-normal determines whether to use the CLT.
2. A common rule of thumb in statistics says to use the CLT if n > 30.
3. 30 is enough for the vast majority of distributions you'll encounter.
4. It's still worth looking at your data before applying the CLT.
5. If you have a very unusual distribution
   E.g., something extremely skewed
   E.g., it can take n = 100, or perhaps even n = 1000 to achieve a normal distribution
6. Very unusual distributions are rare.