DEFINITIONS

Introduction

→ A **distribution** identifies the possible values of a variable and their associated frequencies or probabilities.

Experimental Design

→ **Population:** a group of objects or individuals the researcher will study.
→ **Parameter**: a numeric characteristic of a population.
→ **Sample:** a subset of a population.
→ **Statistic:** a numeric characteristic of a sample, used to estimate a parameter.
→ **Simple random sample** (SRS)
→ **Random assignment**
→ **Random**: the result of a chance process.
→ **(Controlled) experiment:** a study in which a researcher imposes a treatment on subjects and measures responses.
→ **Observational study:** a study in which a researcher measures variables of interest but does not impose intervention.

Data Summary

→ **Categorical (nominal, qualitative) data** have values that correspond to categories or types.
→ **Numerical (quantitative) data** have integer or real number values.
→ **Bar chart:** graphic with bar-heights proportional to the number of observations in each category.
→ **Pie chart:** circular graphic with wedges proportional to the number of observations in each category.
→ **Histogram**: a graphic in which bar areas are proportional to corresponding percentages.
→ **Boxplot:** a graphic composed of a box and 'whiskers' representing the minimum, lower quartile, median, upper quartile and maximum of the data.
→ **Scatterplot:** graphical summary for bivariate numeric data, each point represents measurements on one subject.
→ **Explanatory variable:** The independent variable (x).
→ **Response variable:** The dependent variable (y).
→ **Correlation** describes the *strength* and *direction* of a linear relationship between two numeric variables.

Probability

→ In the study of probability, we examine random processes or '**experiments**'.
→ The **outcomes** ($o_1, o_2,\dots,o_n$) of an experiment are the distinct things that could happen.
→ A **sample space,** denoted S, is a set consisting of all possible outcomes, denoted $\{o_1, o_2,\dots,o_n\}$, of a random process.
→ An **event**, A, is a collection of outcomes (that is, a subset) of the sample space S. We write $A \subseteq S$.
→ An event is **simple** if it consists of exactly one outcome and **compound** if it consists of more than one.
→ A **Venn Diagram** is used to illustrate relationships between events.
→ The **complement** of event A, A', consists of all outcomes in S that are not in A.
→ The **union** of events A and B, $A \cup B$, consists of all outcomes in either A or B or both.
→ The **intersection** of events A and B, $A \cap B$, consists of all outcomes in both A and B.
→ Two events A and B are **mutually exclusive** (disjoint) if $A \cap B = \emptyset$.
→ The **conditional probability** of event A given B is the probability that event A occurs given that event B occurs.
→ A **probability tree** is a diagram for working with conditional probabilities.
→ **Product Rule**: If an experiment takes place in *K* stages, where at the *i*th stage there are n_i possible outcomes, for *i* = 1,…,*K*. Then there are $n_1 \times n_2 \times \dots \times n_k$ possible experimental outcomes.
→ A **permutation** consists of *k* ordered or distinguishable objects chosen from *n* total objects.
→ A **combination** consists of *k* unordered or indistinguishable objects chosen from *n* total objects.

Random Variables

→ **Random variable:** a rule or function that associates a number with each experimental outcome.
→ A **discrete random variable** has a countable number of outcomes.
→ A **continuous random variable** takes on values over an interval or intervals.
→ A **Probability mass function (pmf),** P(X=x), describes the distribution of a discrete random variable.
→ The **expected value** or mean of a discrete random variable, X, with pmf p(x) is $E(X) = \sum_x x \cdot p(x)$.

→ A **Cumulative distribution function** describes the distribution of a random variable with cumulative probabilities. $F(X) = P(X \leq x) = \sum_{y:y \leq x} p(y)$

→ The **variance,** σ_x^2, of a random variable X describes how the values of X vary around their mean.

→ A **linear combination** of random variables is formed by adding random variables and/or multiplying them by scalars.

→ **Parameter** of a distribution: a quantity that can be assigned any of a number of possible values that each determine a different probability distribution.

→ A **family of distributions**: the collection of probability distributions for different values of the parameter.

→ A **binomial random variable** counts successes in n independent trials each resulting in success or failure.

→ A **Poisson random variable** counts independent, rare events occurring within a specified unit of time or space.

→ A **geometric random variable** counts independent trials needed until a success occurs.

The Normal Distribution

→ A **normal random variable**, $X \sim N(\mu, \sigma^2)$, has density function $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$

→ A **standard normal random variable**, $Z \sim N(0,1)$, has mean 0 and variance 1.

→ **Standard units** indicate how many sd's a value is from the mean.

→ **Standardizing** is the process of converting a value to standard units**.**

→ **The Central Limit Theorem**: If X1, X2, … Xn are independent random variables with E(Xi) = μ and Var (Xi) = σ^2 and *n is large* $\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$. If $X \sim Binomial(n, p)$ and $\hat{p}$ =X/n, $\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$ If np ≥ 10 and n(1-p) ≥ 10.

Sampling Distributions

→ **Point estimate:** The observed value of a statistic.

→**Sampling distribution:** The distribution of a statistic.

→ When E($\hat{\theta}$) = θ, $\hat{\theta}$ is an **unbiased** estimator of θ. Otherwise, the bias is E($\hat{\theta}$) – θ.

Inference

→ **Statistical inference** is the science of deducing properties of an underlying probability distribution from a sample.

Confidence Intervals

→ A (1-α)100% **confidence interval** (CI) for an unknown parameter is a set of plausible values of the parameter.

→ A **critical value**, tα/2 denotes the point under the curve such that the area to the right of it is α/2.

Hypothesis Tests:

→ A **hypothesis test** is a procedure for using sample data to decide whether to reject the null hypothesis.

→ The **null** (H_0) and **alternative** (H_A) **hypotheses** are statements about the parameter of interest.

→ The **null** (H_0) **hypothesis**: a claim that is initially assumed to be true**.**

→ The **alternative** (H_A) **hypothesis**: a complementary claim**.**

→ A **test statistic** is a summary of the data that quantifies how different the data are from what is expected under H_0.

→ The **significance level** (α) is the probability of rejecting the null hypothesis when it is true.

→ The **p-value** is the probability of obtaining the data we observed or data more extreme if the null hypothesis is true.

→ A **Type I error** occurs when a true null hypothesis is rejected.

→ A **Type II error** occurs when we fail to reject .

→ Power: the probability of rejecting a false null hypothesis.

→ A **two sample problem** occurs when a comparison is made between two populations.

→ **Paired Samples** result when two measurements are made on the same or related samples.

ANOVA

→ **Between-group variability** describes how the individual group means vary around the overall mean.

→ **Within-group variability** summarizes how observations within each group vary around the group means.

→ **Pairwise comparisons** can be performed after an ANOVA to indicate which means are different and how.

Regression

→ A **scatterplot** is a graphical summary for numeric, bivariate data in which each point represents one subject.

→ **Explanatory variable**: The 'x' variable, also called the independent or explanatory variable.

→ **Response variable**: The 'y' variable, also called the dependent variable.

→ **Linear regression** is used to find a line that summarizes the linear relationship between two variables.

→ The **slope** parameter β_1 represents the change in the average of y for every one unit increase in x.

→ The **intercept** β_0 represents the average value of y when x is zero.

Chi-square Tests

→ **Goodness-of-fit test**: compare proportions in multiple categories to hypothesized values

→ **Test of independence**: determine whether there is an association between two categorical variables.

FORMULAS

For population values $x_1, x_2, \ldots x_N$ and sample values $X_1, X_2, \ldots X_n$

Parameter	Symbol	Definition
Generic parameter	θ	
Population mean	μ	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
Population variance	σ^2	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
Population SD	σ	$\sqrt{\sigma^2}$
Population proportion	p	$p = \frac{1}{N}\sum_{i=1}^{N} x_i$ where $x_i = 0$ or 1

Statistic	Symbol	Definition
Generic statistic	$\hat{\theta}$	
Sample mean	$\bar{X}$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Sample variance	S^2	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Sample SD	S	$\sqrt{S^2}$
Sample Proportion	$\hat{p}$	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $X_i = 0$ or 1

Probability

1. For any event A, $0 \le P(A)$
2. $P(S) = 1$
3. If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint events then $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$
4. $P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k) = \sum_{i=1}^{k} P(A_i)$

$P(\emptyset) = 0$

$P(A) = 1 - P(A')$

$P(A) \le 1$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A|B) = P(A \cap B)/P(B)$ provided $P(B) \neq 0$.

Let $A_1, A_2, \ldots, A_k$ be mutually exclusive events such that $A_1 \cup A_2 \cup \ldots \cup A_k = S$. Then for any other event B

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(B \mid A_i)P(A_i)$$

Let $A_1, A_2, \ldots, A_k$ be mutually exclusive events such that $A_1 \cup A_2 \cup \ldots \cup A_k = S$. Then for any other event B where $P(B) > 0$

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)}, \quad j = 1, \ldots k$$

The number of permutations of n objects chosen k at a time is $P_k^n = \frac{n!}{(n-k)!}$.

The number of permutations of n objects chosen k at a time is $C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$.

Random Variables

$E(X) = \mu_x = \sum_x x \cdot p(x)$, $E(X) = \mu_x = \int_L^U x \cdot f(x)dx$

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

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

$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i)$ for independent random variables

X~Bin(n,p), $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, \ldots, n$, *E(X)=n, Var(X)=np(1-p)*

X~Poisson(μ), $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$; x=0, 1, 2, …

X~Geometric(p), $P(X = x) = p(1-p)^{x-1}, x = 1, 2, 3, \ldots$ $E(X) = \frac{1}{p}$ and $Var(X) = \frac{1-p}{p^2}$

The Normal Distribution

If $X_1, X_2, \dots X_n$ are independent $N(\mu,\sigma^2)$ random variables, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Sampling Distributions

Conditions	Statistic	Distribution
$X_1, X_2, \dots, X_n$ iid $X_i \sim N(\mu, \sigma^2)$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	N(0,1)
$X_1, X_2, \dots, X_n$ iid, n large, $E(X_i) = \mu$, $V(X_i) = \sigma^2$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	Approximately N(0,1)
$X_1, X_2, \dots, X_n$ iid $X_i \sim N(\mu, \sigma^2)$	$\frac{\bar{X}-\mu}{S/\sqrt{n}}$	t_{n-1}
$X \sim B(n,p)$, n large, $\hat{p} = \frac{X}{n}$, $np \geq 10$ and $n(1-p) \geq 10$	$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$	Approximately N(0,1)

Inference

Estimated $s.e.(\bar{X}) = s_{\bar{X}} = \frac{s}{\sqrt{n}}$

Estimated $s.e.\ (\hat{p}) = \hat{\sigma}_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

A (1-α)100% confidence interval for μ is $\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$

An approximate (1-α)100% confidence interval for p when n is large is $\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

P(Type 1 error) = P(Reject H_0 | H_0 is true) = α

P(Type 2 error) = P(Fail to Reject H_0 | H_0 is false) = β

Power = $1-\beta$

A (1-α) level confidence interval for $\mu_A - \mu_B$ is given by $\left(\bar{X} - \bar{Y} - t_{\alpha/2,v}\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}, \bar{X} - \bar{Y} - t_{\alpha/2,v}\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}\right)$

When making calculations by hand, we will use the convention v = min(n-1, m-1).

To implement a hypothesis test for H_0: $\mu_A - \mu_B = \delta$, the test statistic is $T = \frac{\bar{X}-\bar{Y}-\delta}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}} \sim t_v$.

ANOVA

MSTr = SSTr/(k – 1)

MSE = SSE/(n_T – k)

F = $\frac{SSTr/(k-1)}{SSE/(n_T-k)} = \frac{MSTr}{MSE} \sim F_{k-1, n_T-k}$

Source	DF	Sum of Squares	Mean Squares	F-Statistic	p-value
Treatment	k-1	SSTr	MSTr	F	P(F>f)
Error	n_T-k	SSE	MSE		
Total	n_T-1	SST			

Regression

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r\frac{S_y}{S_x}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

Mean Square Error: $\hat{\sigma}^2 = \frac{\sum(y_i-\hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$

A (1 - α)100% confidence interval for β_1 is $\hat{\beta}_1 \pm t_{\frac{\alpha}{2},n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$

Chi-square Tests

Goodness of fit test: $X^2 = \sum_{i=1}^{n} \frac{(x_i - e_i)^2}{e_i} \sim \chi^2_{k-1}$, where $e_1 = np_i$

Test of independence: $X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(x_{ij} - e_{ij}\right)^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}$, where $\mathrm{e_{ij}} = (\text{column j total}) \times \frac{\text{row i total}}{\text{grand total}}$