STATS 3Y03
Probability and Statistics for Engineers

Gulkaran Singh

Spring 2024

Contents

1 Introduction

Probability is quantifying the uncertainty surrounding the outcome of an experiment.

1.1 Important Terminology

- **sample space**: set of all possible outcomes (e.g. $\{1, 2, 3, 4, 5, 6\}$ for rolling a die)
- **event**: any subset of the *sample space* (e.g. $\{1, 3, 5\}$ for odd rolls)

We denote the probability of an event occurring as

$$P(E)$$

where E is the event.

The probability of any event must be between 0 and 1, and if we sum all events the probability must equal 1.

- $0 \leq P(E_i) \leq 1 \; \forall E_i$
- $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$

2 Fundamental Probability

The most basic idea of probability is represented when we have n outcomes where r satisfy some event E.

$$P(E) = \frac{r}{n} = \frac{want}{total}$$

For example, picking a red card from a deck of cards is $\frac{26}{52}$.

3 OR / Addition Rule

What happens when we want to calculate the probability of **multiple** events occurring? For starters, we can define the **generic OR rule** for two events occurring.

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Notice how we subtract the union of these events, so we only consider the probability once since we are separately adding the probabilities.

3.1 Mutually Exclusive Events

These are events that can't occur at the same time. Events E and F are mutually exclusive if

$$E \cap F = \emptyset$$

If event E occurs, then event F cannot occur (and vise versa). Here we can simplify that OR rule to

$$P(E \cup F) = P(E) + P(F)$$

since the intersection between those events is 0.

Example: A card is selected from a well shuffled pack. What is the probability of it being a jack or a 5?

Answer: Since a card can't be a jack **and** a 5, these events are mutually exclusive.

$$P(J \cup 5) = P(J) + P(5) = \frac{4}{52} + \frac{4}{52} = \frac{2}{13}$$

4 AND Rule

To calculate the probability of two events occurring simultaneously, we can define the AND rule which essentially boils down to calculating the probability of the first event and the second event **given** the first event occurred.

For events E and F that are not mutually exclusive,

$$P(E, F) = P(E \cap F) = P(E) \cdot P(F|E)$$

where $F|E$ means the probability of F given E has occurred.

Example: Two cards are selected from a well shuffled pack. What is the probability that they are both jacks?

Answer: We assume *no replacement* (the experiment is not reset),

$$P(J1 \cap J2) = P(J1) \cdot P(J2|J1) = \frac{4}{52} \cdot \frac{3}{51}$$

4.1 Independent Events

These are events where the outcome of one has no effect over the outcome of the other. Therefore, we can say two events are **statistically independent** if

$$P(E \cap F) = P(E)P(F)$$

Notice how we have simplified $P(F|E)$ to $P(F)$ since E has no effect on E.

Example: There are 3 red, 4 green, 7 blue socks in a bag. Three socks are randomly chosen, what is the probability that the first is red, second green, and third blue?

Answer: We can apply the same rule for more than 2 events. We can even logically think this one out.

$$\begin{aligned} P(R \cap G \cap B) &= P(R) \cdot P(G|R) \cdot P(B|G, R) \\ &= \frac{3}{14} \cdot \frac{4}{13} \cdot \frac{7}{12} \end{aligned}$$

5 Inverse Rule

This is a very useful trick where we can calculate the inverse probability of an event occurring to find the probability of what we actually want.

$$P(E) = 1 - P(\overline{E})$$

Example: A coin is flipped 20 times, what is the probability that it shows heads at least once?

Answer: It would be easy to compute if heads was never flipped, so we can use the inverse trick

$$\begin{aligned} P(\text{atleast one } H) &= 1 - P(\text{no}H) \\ &= 1 - \left(\frac{1}{2}\right)^{20} \end{aligned}$$

6 Counting

These topics are related to answering questions revolving around calculating the number of layouts or combinations that are possible.

6.1 Arranging Items – Factorials

To arrange n **distinct** objects in a line is simply $n!$. When we get objects where n_1 are of one type, n_2 of another type and so on, we can calculate the arrangements with

$$\frac{n!}{n_1!n_2!\cdots n_r!}$$

Example: A hospital needs to schedule 3 knee surgeries and 2 hip surgeries in a day. How many different ways can this be done?

Answer: We have $n = 5$ total surgeries, with $n_1 = 3$ knee type and $n_2 = 2$ hip type. Therefore

$$\frac{5!}{3!2!} = 10$$

6.2 Permutations

We can define a way to choose r items from n items where order matters using permutations

$$^nP_r = \frac{n!}{(n-r)!}$$

What does this conceptually look like? Imagine we have 3 runners and we want to **choose** 2 of them. We can create all the possible *permutations* as the following,

$$\{A, B, C\} \to AB, AC, BA, BC, CA, CB$$

So for examples where the order matters (e.g. runners placing 1st, 2nd, 3rd) we use permutations.

6.3 Combinations

We can also choose r items from n items where order doesn't matter (more aligned with set theory). Since we can consider AB and BA to be the same, we can use the following formula to eliminate those extra counts

$$^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Example: How many ways can a committee of 6 people be chosen from 10?

Answer: Since we don't care who was chosen first, second, etc (order doesn't matter), we can use combinations.

$$\binom{10}{6} = 210$$

6.3.1 Multiplication Rule for Combinations

If we have n ways of doing a task and m ways of doing a second task, then the number of ways of doing both tasks is $n \cdot m$.

Example: A bin of 50 total parts has 3 defective. A sample of 6 is chosen, how many different samples contain exactly 2 defective parts?

Answer: We can use the multiplication rule to first choose a sample of 2 from 3 defective parts and second choose 4 from 47 non-defective parts.

$$\binom{3}{2} \cdot \binom{47}{4}$$

This represents 2 defective × 4 clean.

7 Conditional Probability

We've already seen the AND rule, but we can rewrite it to derive a new formula by simply dividing.

$$P(E \cap F) = P(E) \cdot P(F|E)$$

$$P(F|E) = \frac{P(E \cap F)}{P(E)}$$

Note that $E \cap F$ and $F \cap E$ are equivalent, so we can substitute that in and rewrite that in terms of the AND rule again.

$$P(F|E) = \frac{P(F) \cdot P(E|F)}{P(E)}$$

Key Takeaway: Whenever we see the word *given*, we want to use conditional probability!

8 Partition Theorem

When we have multiple partitions, (e.g. in the form $P(E) = P(E \cap F) + P(E \cap \overline{F})$ or $P(E|F)$ and $P(E|\overline{F})$), we can use this formula based off the AND rule to calculate the probability

$$P(E) = \sum_{i=1}^{n} P(F_i) \cdot P(E|F_i)$$

essentially summing the intersections.

Example: A bin of 50 total parts has 3 defective. Two are selected (without replacement), what is the probability that the second is defective.

Answer: We need to consider when the first part is defective and when it is not. Let E be the event where the second part is defective, and F the event where the first is defective. Therefore, we need to consider $P(E|F)$ and $P(E|\overline{F})$.

$$P(E) = P(F) \cdot P(E|F) + P(\overline{F}) \cdot P(E|\overline{F})$$
$$P(E) = \frac{3}{50} \cdot \frac{2}{49} + \frac{47}{50} \cdot \frac{3}{49}$$

9 Random Variables

We have two types of data,

- **continuous** – all values inside an interval (e.g. volume of water in a glass)
- **discrete** – finite (countable) number of values (e.g. number of bacteria)

A **random variable** is function from the sample space (all possible outcomes) to $\mathbb{R}$. For example, if we have an experiment of tossing a coin 6 times, we can let X be a discrete random variable to observe the number of heads observed. Then we can ask what is $P(X = 4)$ which means what is the probability of heads being tossed 4 times.

9.1 Discrete Probability Distribution

If we have a discrete r.v. X that takes values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$ where they all exist and sum to 1, then this defines a discrete probability distribution. Note the set of all possible values of X is called the **support**, denoted by $\mathcal{S}_X$.

Example: Let X represent the number of tails landed from 2 fair coin tosses. X can take values 0, 1, or 2 with probabilities $P(X = 0) = 0.25, P(X = 1) = 0.5, P(X = 2) = 0.25$, then this is a discrete probability distribution.

10 Probability Mass Function

To represent that discrete probability distribution as a cumulative function, we define the **probability mass function** (pmf) for every x value in X by $p(x) = P(X = x)$. The original assumptions still hold true where

- $f(x_i) \geq 0$ (probabilities exist and are greater than 0)
- $\sum_{i=1}^{n} f(x_i) = 1$ (probabilities all sum to 1)
- $f(x_i) = P(X = x_i)$

We can also define the **cumulative distribution** (cdf) as $F(x)$ which is the probabilities summed up to a value x so $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$. It is directly constructed from the summatation of the pmf to x. One useful tip here is if we see $P(X \geq x)$, we can use the inverse rule to make it defined as a cdf so $1 - P(X \leq x)$.

11 Mean, Variance, Standard Deviation (Discrete)

The **expected value** (mean) of a discrete random variable X is given by

$$\mu = \mathbb{E}[X] = \sum_x x \cdot P(X = x)$$

The **variance** of a discrete random variable is given by

$$\sigma^2 = \mathbb{V}ar[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

The **standard deviation** of a discrete random variable is given by

$$\sigma = SD[X] = \sqrt{\mathbb{V}ar[X]}$$

12 Bernoulli Trials

These are experiments that have only 2 outcomes: success and failure. We use the notation X Bernoulli(p) to say X is Bernoulli distributed.

- $P(success) = P(X = 1) = p$
- $P(failure) = P(X = 0) = 1 - p$

so p is the probability of success.

13 Binomial Distribution

The binomial probability distribution is the result of running n independent Bernoulli trials, each with a **constant** success probability (also called a parameter) of p. We define getting x successes from n trials by

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

This makes intuitive sense as

- there are ${}^nC_x$ ways of getting x successes from n trials
- p^x is the probability of success x times

- $(1-p)^{n-x}$ is the probability of failure $n-x$ times

Example: A fair coin is tossed 5 times, what is the probability of tossing *at least* 4 heads.

Answer: Let X be the r.v. representing number of heads tossed. We have $n = 5$ trials, with $x = 4$ successes and a **constant** probability of success of $p = 0.5$.

$$P(X \geq 4) = P(X = 4) + P(X = 5)$$
$$P(X \geq 4) = \binom{5}{4}(0.5)^4(1-0.5)^1 + \binom{5}{5}(0.5)^4(1-0.5)^0$$

Here we have ${}^5C_4$ meaning out of 5 total coin tosses, how many possible ways could we choose 4 tosses? We get 5. From these 5 ways of getting 4 tosses, what's the chance of x tosses being a success times the chance of $n-x$ being a failure.

13.1 Mean, Variance, Standard Deviation

The **expected value** (mean) of a binomial random variable X is given by

$$\mu = E(X) = np$$

The **variance** of a binomial random variable is given by

$$\sigma^2 = V(X) = np(1-p)$$

The **standard deviation** of a binomial random variable is given by

$$\sigma = S(X) = \sqrt{np(1-p)}$$

14 Geometric Distribution

For this distribution, the r.v. X equals the number of trials **until the first success**

$$f(x) = (1-p)^{x-1}p$$

Example: A couple keeps having kids until they get a boy. What is the probability of them having 5 kids?

Answer: Here the 1st success is on the 5th trial.

$$f(5) = (1-0.5)^4(0.5)$$

This makes intuitive sense as well since we fail on the first $x-1$ trials ($1-p$ probability), then succeed on the xth trial which is why we multiply by p afterwards.

14.1 Mean, Variance, Standard Deviation

The **expected value** (mean) of a geometric random variable X is given by

$$\mu = E(X) = \frac{1}{p}$$

The **variance** of a geometric random variable is given by

$$\sigma^2 = V(X) = \frac{(1-p)}{p^2}$$

The **standard deviation** of a geometric random variable is given by

$$\sigma = S(X) = \sqrt{\frac{(1-p)}{p^2}}$$

15 Negative Binomial Distribution

This distribution is similar to the binomial, however here r.v. X is the number of trials until r successes occur. **In other words, the xth trial, is the rth success.** Normal binomial just states x successes in n trials total.

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

where $x \geq r$.

Example: The probability of being infected when exposed to a disease is 30%. What is the probability the 10th child exposed is the 4th to catch it?

Answer: The $x = 10$th child is the $r = 4$th 'success'.

$$P(X = 10) = \binom{10-1}{4-1} (0.3)^4 (0.7)^{10-4}$$

16 Hypergeometric Distribution

The previous three distributions are only applicable when we sample with replacement. If we have a case where we sample **without** replacement, we use the hypergeometric distribution.

Imagine a scenario with N total units with M defective ($M \leq N$). If we sample n units, the probability that x of those units are defective is given by

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Notice how this is the multiplication rule from combinations! We choose x defects from M and choose $n-x$ non-defects from $N-M$.

Example: An athlete hides 2 performance enhancing pills in a bottle containing 8 total pills. If we sample 3 of these pills, what is the probability cheating will be detected?

Answer: We have $N=8$ total pills, $M=2$ are defective, $n=3$ are sampled, and $x \geq 1$ results in cheating being detected.

$$\begin{aligned} P(X \geq 1) &= 1 - P(X=0) \\ &= 1 - \frac{\binom{2}{0}\binom{8-2}{3-0}}{\binom{8}{2}} \\ &= 0.53 \end{aligned}$$

17 Poisson Distribution

This distribution is most useful when dealing with the frequency of events (μ) occurring in a specific interval. It can also act as an approximation to the binomial distribution when you take $\mu = np$.

$$P(X=x) = \frac{e^{-\mu}\mu^x}{x!}$$

18 Continuous Random Variables

A random variable X is **continuous** if its cdf takes the form

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

where we go from some lower bound (-∞) to x. It will satisfy the following conditions

$$\begin{aligned} &P(X=x) = 0 \quad \text{(integrating under a point is 0)} \\ &P(X \leq x) = P(X < x) \\ &P(a \leq x \leq b) = F(b) - F(a) = \int_a^b f(x)\ dx \end{aligned}$$

We can also call $f(x)$ the pdf of X if it similarly satisfies

$$\begin{aligned} f(x) \geq 0\ \forall x \\ \int_{-\infty}^{\infty} f(x)\ dx = 1 \end{aligned}$$

Key tips are, if given the $f(x)$ (pdf), simply integrate to find the $F(x)$ and if given the $F(x)$, derive it to get the $f(x)$. Any $P(X < x)$ is just $F(x)$.

19 Mean, Variance (Continuous)

The **expected value** (mean) of a continuous random variable X is given by

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\ dx$$
$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\ dx$$

The **variance** of a continuous random variable is given by

$$\sigma^2 = \mathbb{V}ar[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ dx = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Note: when you calculate the variance, to calculate $\mathbb{E}[X^2]$, you treat $g(x) = X^2$ so use the second formula for mean. If it's a piece-wise function, you can use the non-zero values for the part you integrate.

20 Normal (Gaussian) Distribution

This is characterized by a bell-shaped curve and is the most frequently used continuous probability distribution.

We say $X \sim N(\mu, \sigma^2)$ to say r.v. X follows a normal distribution with mean μ and variance σ^2. The mean, median, and mode are all the same for normal distributions.

20.1 Standard Normal Distribution

This is a normal distribution with $\mu = 0$ and $\sigma = 1$. For both the normal and standard, 95% of the density lies between ±1.96 of the mean. We can use Z tables to look up probabilities, but first, we need to convert from Normal to Standard Normal. We let $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$, then

$$Z = \frac{X - \mu}{\sigma}$$

and from that, we can compute

$$P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma})$$

and consult the Z-tables. If the Z value we need to lookup is negative, we can compute 1 minus the positive value since the normal distribution is symmetric.

For $P(a \leq X \leq b)$, it is the same logic as $F(b) - F(a)$ where we can do $P(X \leq b) - P(X \leq a)$ and transform those to the standard normal to use Z-tables. Note we always round up to the first 2 decimals when using the Z-table, so 0.666... would be 0.67 when looked up.

21 Binomial Approximation

We can approximate the binomial distribution with the standard normal. This works well with large n and small p. We apply a continuity correction,

$$\begin{aligned} P(X \leq x) = P(X \leq x + 0.5) &\approx P\left(Z \leq \frac{x + 0.5 - \mu}{\sigma}\right) \\ P(X \geq x) = P(X \geq x - 0.5) &\approx P\left(Z \geq \frac{x - 0.5 - \mu}{\sigma}\right) \end{aligned}$$

where $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. Remember, to calculate $P(X \geq x)$, you get $P(Z \geq z)$ but to look that z value up, you need to do $1 - P(Z < z)$.

22 Exponential Distribution

This helps describe time between events in a Poisson process – events that occur continuously, independently, and at a constant average rate.

$$\begin{aligned} f(x) &= \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \lambda > 0 \\ 0 & \text{otherwise} \end{cases} \\ P(X \leq x) &= F(x) = 1 - e^{-\lambda x} \\ P(X \geq x) &= e^{-\lambda x} \end{aligned}$$

where $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$. Typically, these questions give a mean where you can isolate to find λ. Note, the limits are related to λ, so they must be in that unit of time.

23 Joint Probability Distributions

In many situations, we are interested in 2 random variables. Consider continuous random variables X and Y, the function $f_{X,Y}(x, y)$ is called the joint probability density function if

$$\begin{aligned} &f_{X,Y}(x, y) \geq 0 \quad \forall (x, y) \in \mathbb{R}^2 \\ &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\ dx\ dy = 1 \end{aligned}$$

where we can calculate the probability if we use the restrictions as the limits.

If we don't have a simple rectangular domain, then we have

$$P(X,Y) = \int_a^b \int_{c(x)}^{d(x)} f_{X,Y}(x,y)\ dy\ dx = 1$$

where $c(x)$ and $d(x)$ are limits of Y that it can take on in terms of X.

Example: For a function $65e^{-5x-8y}$ where $0 < y < x$, find $P(X < \frac{1}{8}, Y < 2)$.

Answer: Since we know the x bounds are from 0 to $\frac{1}{8}$, we just need to find the limits for Y. Since Y depends on x and is bound by x, it can take on the value that is $\min(x, 2) = x$ since x goes to 1/8.

$$\int_0^{1/8} \int_0^x 65e^{-5x-8y}\ dy\ dx$$

23.1 Marginal PDF

If you want the probability distribution of each variable separately, you integrate with the respect of the other variable.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\ dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\ dx$$

We can say random variables X and Y are independent iff

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

24 Mean, Variance (Joint)

The **expected value** (mean) of joint continuous random variables X, Y is given by

$$\mathbb{E}[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f_{X,Y}(x,y)\ dx\ dy$$

where $h(X,Y)$ is a scalar function. The **covariance** of joint continuous random variables is given by

$$\text{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Notice the similarities to $\mathbb{E}[X^2] - \mathbb{E}[X]^2$. Covariance in general is a measure of how X and Y are related.

24.1 Correlation

This is the strength of the linear relationship between random variables.

$$\text{Corr}(X,Y) = \rho = \frac{\text{Cov}(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

We can also say these are independent if we consider scalar functions g and h,

$$E[g(X),h(Y)] = E[g(X)]E[h(X)]$$

It follows that if X and Y are independent, then their correlation coefficient $\rho = 0$.

24.2 Coefficient Rules

$$E[aX+bY] = aE[X]+bE[Y]$$
$$Var[aX \pm bY] = a^2Var[X]+b^2Var[Y] \pm 2ab\text{Cov}[X,Y]$$

Note, $2ab\text{Cov}[X,Y] = 0$ if X,Y independent.

25 Sample Mean, Standard Deviation, and IQR

We typically take a small **sample** (in the form $x_1, x_2, \ldots, x_n$) from a larger **population**.

The **mean** of a sample (different than population mean which is μ) is given by,

$$\overline{x} = \frac{\sum_{i=1}^{x} x_i}{n}$$

The **standard deviation** of a sample (different than σ) is given by,

$$s = \sqrt{\frac{\sum_{i=1}^{x}(x_i - \overline{x})^2}{n}}$$

The **interquartile range** is the size of the gap between the first and third quartile.

$$Q_1 = \frac{n+1}{4}, \quad Q_3 = \frac{3(n+1)}{4}, \quad IRQ = Q_3 - Q_1$$

Note, these equations give us the position of Q_1, Q_3 in a sorted sample.

Example: Consider the following data, 19, 27, 41, 44, 44, 47, 51, 51, 56, 57, 60, 60, 60. What is the IQR?

Answer: First start by sorting the sample. Then we calculate Q_3 and Q_1. Notice how we have to interpolate for decimals!

$$Q_1 = \frac{13+1}{4} = 3.5$$
$$Q_3 = \frac{3(13+1)}{4} = 10.5$$

The data in position 3 is 41 and the data in position 10 is 57. Now we interpolate which means find the decimal portion of position 3 → 4 and 10 → 11.

$$Q_1 = 41 + 0.5(44 - 41) = 42.5$$
$$Q_3 = 57 + 0.5(60 - 57) = 58.5$$
$$\therefore IQR = 58.5 - 42.5 = 16$$

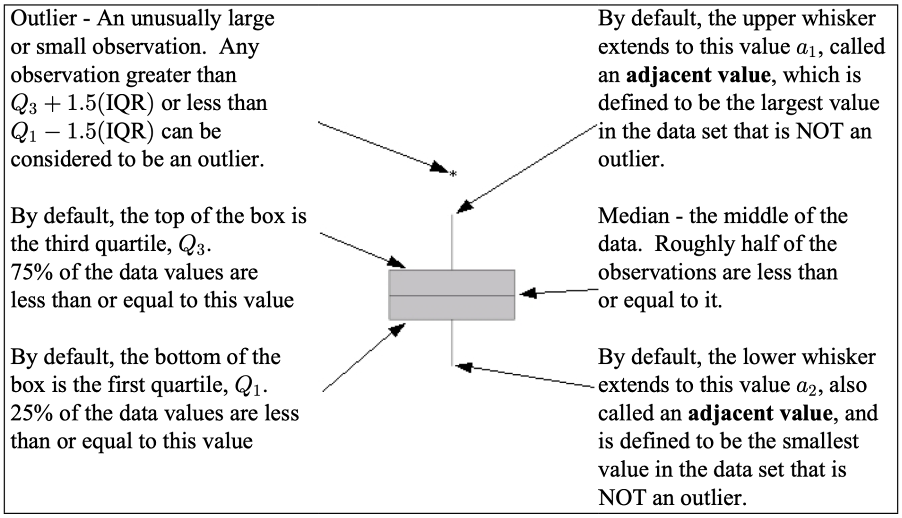
26 Histograms

A way to display data in equal bins where the sum of the heights = 1 (density = 1). We typically choose the number of bins to be the square root of the number of observations ($\sqrt{n}$)

27 Box Plots

Box and whisker plots show us the IQR range, outliers, adjacent values, range, median, symmetry, and min/max of a dataset.

- center box is the IQR
- data points outside of $1.5IQR$ are **outliers**
- lines (whiskers) are drawn from the outside of the box to the min/max values that are not outliers (within $1.5IQR$)
- the line in the middle box is the **median**
- typically drawn vertically where the top of the box is Q_3 (75% of the data is below this point) and the bottom of the box is Q_1 where 25% of the data is below that line.



Outlier - An unusually large or small observation. Any observation greater than $Q_3 + 1.5(\text{IQR})$ or less than $Q_1 - 1.5(\text{IQR})$ can be considered to be an outlier.

By default, the top of the box is the third quartile, Q_3.
75% of the data values are less than or equal to this value

By default, the bottom of the box is the first quartile, Q_1.
25% of the data values are less than or equal to this value

By default, the upper whisker extends to this value a_1, called an **adjacent value**, which is defined to be the largest value in the data set that is NOT an outlier.

Median - the middle of the data. Roughly half of the observations are less than or equal to it.

By default, the lower whisker extends to this value a_2, also called an **adjacent value**, and is defined to be the smallest value in the data set that is NOT an outlier.

28 Normal Probability Plots

We draw these plots to see if data is normally distributed. We can create one by plotting z-scores! We plot the sample points (x_j) on the x-axis and the corresponding z_j values on the y-axis.

$$\frac{j - 0.5}{n} = P(Z \leq z_j)$$

After calculating the probability, you need to find the corresponding z_j value that would result in that probability from the z-table and plot that z_j value on the y-axis.

29 Estimators

In the real world, it is unlikely we know the population mean, variance, or standard deviation. This is why we have samples! So if we have a parameter θ which can be $\mu, \sigma, \sigma^2, \ldots,$ we want to find ways we can **estimate** these from sample data.

Denote $\hat{\theta}$ a **point estimator** of θ, then $\hat{\theta}$ can be an *unbiased* estimator if

$$\mathbb{E}[\hat{\theta}] = \theta$$

but we can get biases (things added or subtracted to θ). The bias estimate is

$$\mathbb{E}[\hat{\theta}] - \theta = bias$$

We now define two axioms which we can use to prove other estimates in the future.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is an unbiased estimator of μ. So the $\mathbb{E}[\overline{x}] = \mu$.

Suppose $X \sim \mathit{Binomial}$ with known n and unknown p,

$$\hat{p} = \frac{X}{n}$$

is an unbiased estimator of p. Therefore, $\mathbb{E}[\hat{p}] = p$.

The unbiased estimator of σ^2 is s^2,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Something to consider is that among all unbiased estimators of θ, we choose the one with the smallest variance. This is called the Minimum Variance Unbiased Estimation.

The **standard error** is the standard deviation but is denoted by $s_{\hat{\theta}}$.

$$SE(\overline{X}) = \hat{\sigma_{\overline{x}}} = \frac{s}{\sqrt{n}}$$

Example: is $\hat{\theta} = \frac{x_1+\cdots+x_7}{7}$ a unbiased estimator of μ?

Answer: For these questions, we need the expected value of the expression to simplify to μ.

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{x_1+\cdots+x_7}{7}\right] \\
&= \frac{1}{7}(\mathbb{E}[x_1]+\cdots+\mathbb{E}[x_7]) \\
&= \frac{1}{7}(7\mu) \\
&= \mu
\end{aligned}$$

Note when we take constants out of $\mathbb{V}ar$, the constants get squared out.

30 Central Limit Theorem

If we have a sample, this theorem states $\overline{X}$ will be normally distributed as n grows.

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

We can use this as normal for probability questions that come from a sample! If we have two sample means from independent samples, then

$$Z = \frac{\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

31 Confidence Intervals

This is how likely the population mean is in a given interval we calculate using our sample mean.

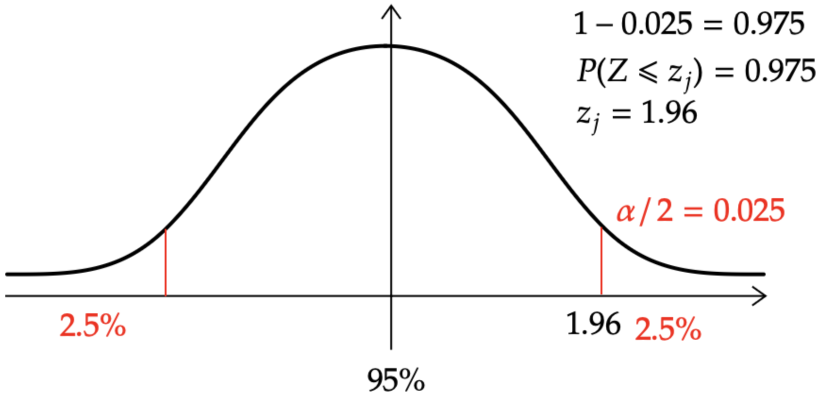
We know the distribution of a sample mean is $\overline{X} \sim N(\mu, \sigma^2/n)$. If we know σ, then the interval (centered around $\overline{x}$) is given by

$$P\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

where $1 - \alpha$ is the confidence level we want (e.g. $\alpha = 0.05$ means 0.95 or 95% confidence μ is in that interval).

31.1 Calculating $z_{\alpha/2}$

Say we want to construct a C.I for μ with 95% confidence. This means $1 - \alpha = 0.95$ and $\frac{\alpha}{2} = 0.025$. To find $z_{0.025}$, we add α back onto 0.95 and find z_j when $P(Z \leq z_j) = 0.975$. Since this is a 2-way interval, 2.5% of the region is so $1 - 0.025 = 0.975$.



This means μ lands in the confidence interval 95% of the time as μ never changes. These are the most common C.I's and z-values.

$1-\alpha$	0.80	0.90	0.95	0.98	0.99
$z_{\alpha/2}$	1.28	1.645	1.96	2.326	2.576

The width of a confidence interval is given by

$$w = 2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Note, $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is called the **margin of error** and is half the C.I. We can make the width smaller by making n larger. Say we want to calculate a new n for a new width we'll call q. Then

$$2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq q$$

$$\vdots$$

$$n \geq \left(2 \cdot z_{\alpha/2} \frac{\sigma}{q}\right)^2$$

Rearranging this is just swapping q and $\sqrt{n}$ and square. Always remember to **round up** for choosing n.

31.2 1-Sided Intervals

For these we simply replace $z_{\alpha/2}$ with z_α and read off the table. So 1 sided C.I of 90% is simply 1.26.

32 Student t-Table

In many situations, we will not know what σ is, so we estimate it with S but we use t-tables instead. As n gets large, the T distribution is indistinguishable from the Normal distribution. It has one paramter called the degrees of freedom where $v = n - 1$.

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

When reading this table, it is defined for $P(T > t)$ so it finds the area to the **right**. This is the opposite of z-tables so be careful! We read the t value directly from the middle of the table and we get an α value as a result.

Example: What is $P(t < 1.711)$ when $n = 25$?

Answer: Inverse the probability and use degrees of freedom!

$$v = 25 - 1 = 24$$
$$\alpha = 1 - P(t > 1.711)$$
$$\alpha = 1 - 0.05 = 0.95$$

32.1 Confidence Interval

$$\overline{x} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

33 Proportion Confidence Interval

For these we always use z-tables, but we can construct a C.I for $\hat{p} = \frac{X}{n}$.

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

Notice that the standard deviation is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (it was $\frac{\sigma}{\sqrt{n}}$ in the other cases).

Example: 680 out of 1000 Canadians said cruises are good use of money. Calculate a 95% C.I for the proportion of Canadians who feel this way.

Answer: Calculate $\hat{p}$ first, then fill in the rest!

$$\hat{p} = \frac{X}{n} = \frac{680}{1000} = 0.68$$
$$0.68 \pm 1.96(0.01475)$$

Choosing a sample size with an error is given by the following,

$$n = \left[\frac{z_{\alpha/2}}{E}\right]^2 \cdot \hat{p}(1-\hat{p})$$

when exactly confident but if no estimate of the sample proportion is calculated or to maximize n, use $\hat{p}(1-\hat{p}) = 0.50$.

34 Hypothesis Testing

We can test a theory on a sample of data by conducting hypothesis testing on some parameter, so we can reject or fail to reject (accept) it.

- H_0 – null hypothesis (of no effect / status quo)
- H_a – alternative hypothesis (result the researcher is hoping to show)

With these two hypothesis, we can find evidence to **reject** H_0 in favour of H_a or **fail to reject** H_0. Note, we never say accept H_0 so for exam answers, that is never the case.

$$\begin{aligned} \text{Two tail test} - H_0 : \mu = \mu_0 \quad & H_a : \mu \neq \mu_0 \\ \text{Upper tail test} - H_0 : \mu = \mu_0 \quad & H_a : \mu > \mu_0 \\ \text{Lower tail test} - H_0 : \mu = \mu_0 \quad & H_a : \mu < \mu_0 \end{aligned}$$

So how do we conduct a hypothesis test? First we calculate the **test statistic** from a sample. If this value is highly unusual (falls in our rejection region), we use it as proof to reject H_0.

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

Then we can calculate our **critical value** which defines where our **rejection region** starts. Our rejection region is the set of all values of the test statistic that causes us to reject H_0. To find the critical value, we use the significance value (α)

$$z_\alpha \text{ or } z_{\alpha/2}$$

Now we can reject H_0 if z falls in the rejection region. This will depend on if its an upper, lower, or two-tailed test.

34.1 p-value

The p value is the probability of getting a more extreme test statistic. So it is the probability to the left of the test statistic. If the $|p| < \alpha$, we reject H_0. For the two tail test, the p value is split between both ends, so we multiply by 2.

34.2 Student t Test

In the same way for confidence intervals, if σ is not known, we use s and use t tables. So the test statistic is now

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

and the critical value is t_{n-1}^{α} or $t_{n-1}^{\alpha/2}$ for two tailed test.

34.3 Proportions

If we have a proportion $\hat{p} = x/n$, we can work out a test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

34.4 Connection with Confidence Intervals

We reject H_0 if u_0 is not in the C.I.

34.5 Errors in Hypothesis Testing

- Type 1 error (α): reject the H_0 when it's actually true (calculate probabilities left and right of critical values)

- Type 2 error (β): fail to reject (accept) H_0 when it's actually false (calculate the acceptance region)
- Power: $1 - P(Type\ 2\ error)$

35 Inference on Difference of Means (Variance Unknown)

Assuming we have two independent samples where the data is normally distributed and there is equal variance ($\sigma_A^2 = \sigma_B^2$), we can test hypothesis about $\mu_A - \mu_B = 0$.

35.1 Confidence Intervals

To compute the standard error, we define

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

so the confidence interval (always two tails) is

$$(\overline{x}_A - \overline{x}_B) \pm t_{n_A + n_B - 2}^{\alpha/2} \cdot s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

If the C.I does not contain 0, we reject the H_0. For hypothesis testing, the test statistic is

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

If variances are unequal, there are different formulas to use so look at the formula sheet!

36 Simple Linear Regression

We use a linear line to predict values. We fit a simple linear regression (SLR) model by using the Least Squares Technique where we minimize the sum of the squared vertical distance from each point to the line.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \forall i = 1, \ldots, n$$

where β_0 is the y-intercept and β_1 is the slope. We don't know the true values of β_0 and β_1, so we use estimates b_0, b_1. This gives us

$$\hat{Y}_i = b_0 + b_1 X_i$$

We can define the vertical distances (observed errors) as

$$e_i = Y_i - \hat{Y}_i$$

These estimates of b_0, b_1 minimize the squared vertical distance and are called the least square estimates.

$$\begin{aligned} b_0 &= \overline{Y} - b_1\overline{X} \\ b_1 &= \frac{s_{xy}}{s_{xx}} = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} \end{aligned}$$

For every additional unit increase in X, the variable Y increases by b_1 units. The b_0 can mean something, but it can also have no practical meaning. Here are some other important formulas not on the formula sheet,

$$\begin{aligned} SS_T &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\ SS_E &= SS_T - b_1 s_{xy} \end{aligned}$$

36.1 Testing Linear Relationship

If there is a linear relationship, the slope between X and Y will be nonzero, so we can do a hypothesis test on this! The test statistic is calculated by

$$\begin{gathered} H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0 \\ t = \frac{b_1}{\sqrt{\sigma^2 / s_{xx}}} \end{gathered}$$

and the critical value is calculated by $t_{n-2}^{\alpha/2}$

36.2 F-Test

For the ANOVA procedure

$$\begin{aligned} &\sum_{i-1}^{n} (Y_i - \overline{Y})^2 = \sum_{i-1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i-1}^{n} (Y_i - \hat{Y}_i)^2 \\ &SS_T = SS_R + SS_E \end{aligned}$$

We can use the F-Test for significance of regression. It leads to the same result as a the t-test and can only be used for 2-tailed tests. The test statistic is

$$F_0 = \frac{SS_R}{SS_E/(n-2)} = \frac{MS_R}{MS_E}$$

where we reject H_0 if $F_0 > F_{1,n-2}^{\alpha}$ (critical value). Note we always set the first degree of freedom to 1. MS stands for Mean Square and is calculated from SS divided by $d.f.$

36.3 Confidence Intervals

If we want to build a 100(1-α)% C.I. for β_1, it has the form

$$b_1 \pm t_{n-2}^{\alpha/2} \cdot \sqrt{\hat{\sigma}^2/s_{xx}}, \quad \hat{\sigma}^2 = \frac{SS_E}{n-2}$$

36.4 Prediction Intervals

This interval is 100(1-α)% confident the true value of Y falls in that interval when we predict using our SLR model. For a new observation Y at $X = X_h$,

$$\hat{Y}_h \pm t_{n-2}^{\alpha/2} \cdot \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{s_{xx}}\right)}$$

where $\hat{Y}_h$ is the predicted value at X_h.

36.5 Co-efficient of Determination

We use this to judge the adequacy of a regression model.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

37 Anova Tables

We've compared two means, however, what if we want to compare more than two? We can use ANOVA (analysis of variances) to test means. This is a more generalized version of the previous version. Say we have $i = 1, \ldots, I$ groups producing data and suppose each group has $j = 1, \ldots, J$ observations. We get the same formula as before, but SS_R is now SS_{Trt}.

$$SS_T = SS_E + SS_{Trt}$$
$$\sum_{ij}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = \sum_{ij}(Y_{ij} - \overline{Y}_{i\cdot})^2 + \sum_{ij}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

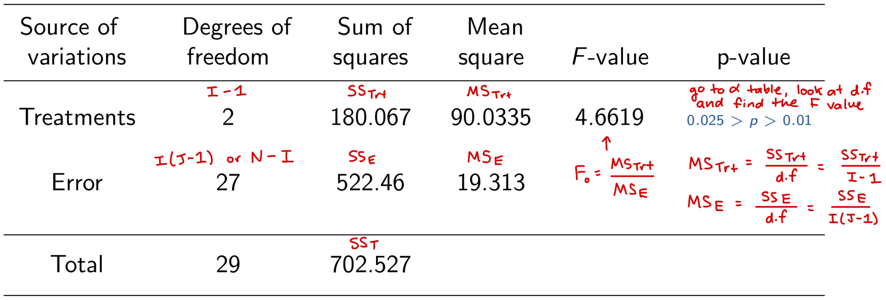
The $\cdot$ notation is the summation over the subscript that it replaces. Note that $N = IJ$ (total number of observation).

$$Y_{i\cdot} = \sum_{j=1}^{J} Y_{ij}$$
$$\overline{Y}_{\cdot\cdot} = \frac{Y_{\cdot\cdot}}{n}, \quad Y_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$$

The F-Test associated with ANOVA is given by $H_0 : \mu_i = \mu \ \forall i$ vs H_a : not all μ_i are the same.

$$F_0 = \frac{MS_{Trt}}{MS_E}$$

and we compare this to a critical value of $F^{\alpha}_{I-1,I(J-1)}$. Rejecting H_0 means there is evidence that atleast two means are significantly different from eachother.



Source of variations	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Treatments	$I-1$ 2	SS_{Trt} 180.067	MS_{Trt} 90.0335	4.6619	go to α table, look at d.f and find the F value $0.025 > p > 0.01$
Error	$I(J-1)$ or $N-I$ 27	SS_E 522.46	MS_E 19.313	↑ $F_0 = \frac{MS_{Trt}}{MS_E}$	$MS_{Trt} = \frac{SS_{Trt}}{d.f} = \frac{SS_{Trt}}{I-1}$ $MS_E = \frac{SS_E}{d.f} = \frac{SS_E}{I(J-1)}$
Total	29	SS_T 702.527			

Note $F = T^2$. Also J is split evenly by the groups so we would have to divide N total observations by I groups and if it's not even, we use $N - I$. Note, if given summary statistics, and we have the observations per group, we can total them to get us N total observations, but in our table, it is $N - 1$ total.

38 Fisher's Least Significant Different (LSD)

If we reject H_0, we know there is a difference in means, but we don't know where/which group that difference is in. We can calculate all the pairwise confidence intervals,

$$(\overline{x}_i - \overline{x}_j) \pm t^{\alpha/2}_{n-I} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Similarly, if the C.I contains 0, we reject the H_0 and this suggests there is evidence that means are significantly different.