Math 11: Calculus-Based Introductory Probability and Statistics

Course Note

Prepared by: Thu Nguyen - thn003@ucsd.edu

Last updated on: June 25, 2020

Notes

1. This note is intended to be used as a quick reference only, and thus does not contain all concepts/examples introduced in lectures and the textbook. Please use the lecture notes and the textbook as your main sources.
2. This note is still a work in progress and will thus be updated when appropriate.
3. When updating, I may not create new sections for new materials, but instead put them under existing sections, where appropriate.
4. If you spot any mistakes (typos or conceptual mistakes), please let me know. Many thanks!
5. This note is for your own use only, please do not distribute this note.

Contents

1 Exploring and Understanding Data **3**

1.1 Describing Data . . . 3

1.2 Visualizing Data . . . 4

2 Exploring Relationships between Variables **5**

2.1 Correlation . . . 5

2.1.1 z-score . . . 5

2.2 Linear Regression . . . 6

2.3 Model Evaluation . . . 7

2.3.1 Numerically - R^2 . . . 7

2.3.2 Visually - Residual plot . . . 7

3 Randomness & Probability **9**

3.1 Definitions . . . 9

3.1.1 Conditional Probability . . . 9

3.2 Probability Rules . . . 9

3.3 Bayes' Rule . . . 10

3.3.1 Probability Tree . . . 10

3.4 Law of Total Probability . . . 11

3.5 Random Variables . . . 12

3.5.1 Properties of Expected Value . . . 13

3.5.2 Properties of Variance . . . 13

3.6 Models & Distributions . . . 14

3.6.1 Discrete Random Variables . . . 14

3.6.2 Continuous Random Variables . . . 16

4 From the Data at Hand to the World at Large **18**

4.1 Statistical Inference . . . 18

4.1.1 Law of Large Number . . . 18

4.1.2 Central Limit Theorem . . . 19

4.2 Confidence Interval . . . 20

4.2.1 Qualitative Data . . . 20

4.2.2 Quantitative Data . . . 20

4.3 Hypothesis Testing . . . 22

4.4 Regression Inference . . . 25

4.4.1 Testing for Regression Coefficient . . . 25

4.4.2 Confidence Interval vs. Prediction Interval . . . 26

4.5 Chi-squared tests . . . 27

1 Exploring and Understanding Data

1.1 Describing Data

Given some sample (data), say:

$$Y = \{y_1, y_2, \ldots, y_n\}$$

for simplicity, we assume that Y has been ordered in increasing order, i.e. $y_1 \leq y_2 \leq \cdots \leq y_n$. To describe Y, we have a number of measures:

Measure	Definition and Formula
mean	the average of all values: $$\mu_Y = \hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
median	the value which is larger than the bottom 50% and less than the top 50%: $$Q_2 = \begin{cases} y_j, & (\text{if } n \text{ is odd, and } j = \frac{n+1}{2}) \\ \frac{1}{2}(y_j + y_{j+1}), & (\text{if } n \text{ is even, and } j = \frac{n}{2}) \end{cases}$$
IQR	interquartile range: covering the middle 50% of all data $$IQR = Q_3 - Q_1$$
variance	measures the spread of the data $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y})^2$$
standard deviation	$$s = \sqrt{variance}$$

Table 1: Some statistical measures.

Some intuitions to think about those measures are:

1. *mean, median,* Q_3, and Q_1 refer to some particular values that "partition" our sample into "blocks"
2. *IQR* and *variance* tell us how much of a gap there is between the smaller and the larger values

1.2 Visualizing Data

Given a sample (data), we have a variety of plots to visualize the data, with each giving a unique perspective. Figure 1 shows a few examples.

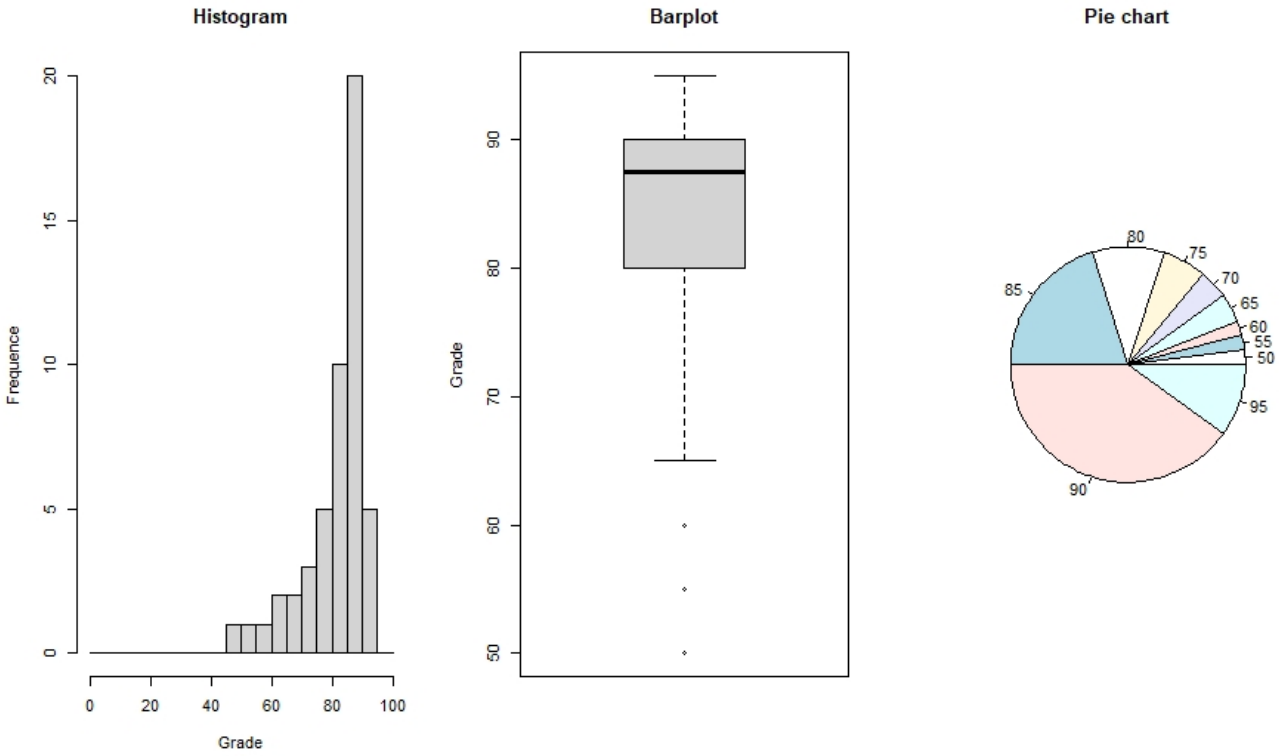


Figure 1: Examples of types of plots to visualize data.

Of all, probably the most popular type is histograms , where (usually, though not always the case) the numerical data values are plotted on the x-axis and the respective frequencies on the y-axis (the frequency can be either the absolute count, or the relative percentage). Depending on the shape of the histogram, it is preferable to report certain statistical measures:

Histogram shape	Preferred statistical measures
symmetric	mean & standard deviation/variance
skewed	median & IQR

In particular, we have some observations regarding a skewed histogram:

left skewed	mean $<$ median
right skewed	mean $>$ median

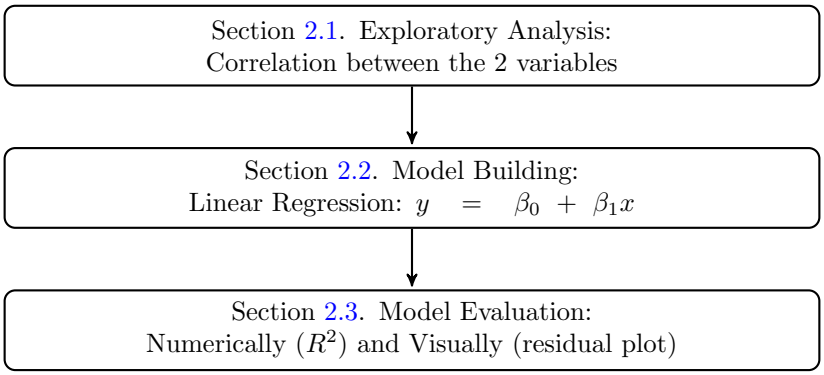
Additionally, report any outliers and modes, if appropriate.

2 Exploring Relationships between Variables

Imagine we are given a dataset which consists of more than 2 variables (in principle, the following concepts generalize naturally into the case of k variables, where k can be any number), say:

$$\{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$$

Our goal is to analyze the dataset, and to build some model relating x and y. This figure gives an overview:



2.1 Correlation

Given the above dataset, a natural question to ask would be how related are the x and y. A way to quantify that relatedness is the correlation , denoted by r:

$$Corr[x, y] = r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where $\bar{x}, \bar{y}$ are the averages for x and y data points, and s_x, s_y the respective standard deviations. In particular:

$$-1 \leq r \leq 1$$

Figure 2 shows what r can show the relatedness. However, always keep in mind that *correlation does not imply causation*: i.e. just because the correlation coefficient is high, it does not necessarily imply that one variable causes the other variable.

Note that, we can, in principle, generalize the concept of *correlation* to between any pair of *quantitative* (numerical) variables, in the case of multi-variate data.

2.1.1 z-score

Given a sample (data), say

$$Y = \{y_1, y_2, \dots, y_n\}$$

a natural question is investigating how far each data point is from the average, measured by the z-score:

$$z - \text{score}(y_i) = \frac{y_i - \bar{y}}{s_y}$$

where $\bar{y}$ and $_y$ are the average and the standard deviation.

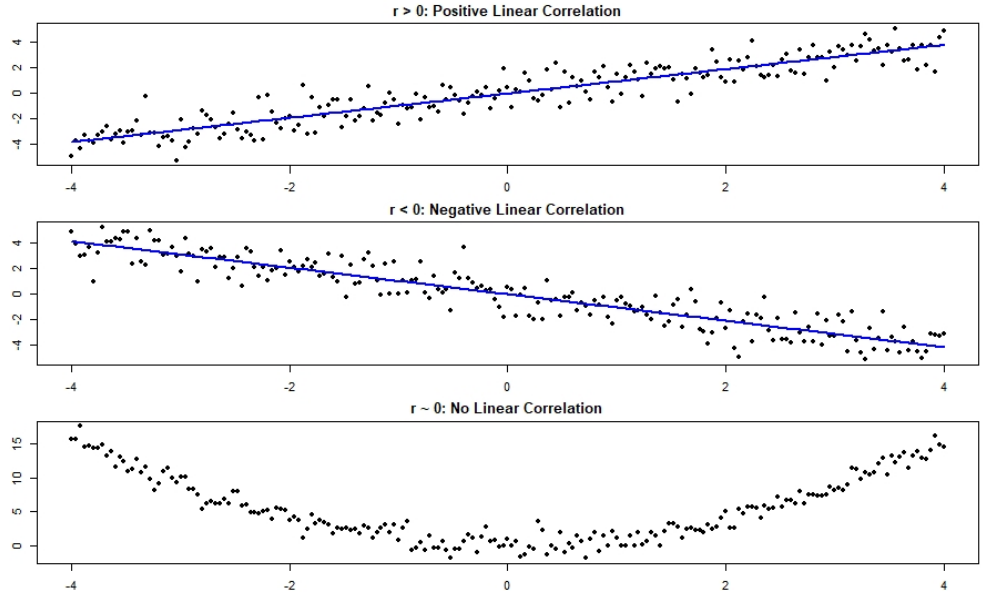


Figure 2: Examples of different types of linear correlation. The best fit line (estimated via *linear regression* - refer to section 2.2) is shown where appropriate.

2.2 Linear Regression

Similar to the setting from section 2.1, given:

$$\{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$$

suppose we want to model y as a response given variable x. In particular, we assume they are linearly related, which gives the formulation:

$$y = \beta_0 + \beta_1 x$$

Although there are multiple ways to figure out (estimate) (β_0, β_1), the most common method is via the *least squares* estimate, where we want to:

$$\min \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)^2 \tag{1}$$

Suppose we have done the computation and find out the best parameters, denoted $(\hat{\beta}_0, \hat{\beta}_1)$, these give the residuals :

$$\epsilon_i = y_i - \hat{y}_i$$

for each $i = 1, 2, \dots, n$, and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Solving 1 gives us the formula for $(\hat{\beta}_0, \hat{\beta}_1)$:

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

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

2.3 Model Evaluation

Now that we have already built and computed the optimal values $(\hat{\beta}_0, \hat{\beta}_1)$, a follow-up question could be how good our model is. We can analyze this question either numerically via the R^2 value, or visually via the *residual plot*. One way to think about the 2 methods is:

1. R^2: confirms if a model is *good*
2. *Residual plot*: confirms if a model is *not* so good

2.3.1 Numerically - R^2

R^2 is a measure of how much variation in the data could be explained by our model, intuitively:

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

Intuitively, a good model would contain almost all the information (the relationship between x and y) within it, in which case, the *Explained Variation* is approximately close to the *Total Variation*, and hence $R^2 \approx 1$. The general rule of thumb is:

1. The higher the R^2 value, the "better" our model is.
2. When comparing models, assuming all else are equal, a model with higher R^2 value would be a better model.

In particular, given our dataset above, the formula is:

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{\sum(y_i - \bar{y})^2}$$

where:

1. $\sum(y_i - \bar{y})^2$ is the total variation
2. $\sum\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$ is the variation not explained by our model

2.3.2 Visually - Residual plot

Alternatively, we could look at the plot of the residuals. Contrary to the numerical check above, generally, there is only 1 type of residual plot which could confirm the goodness-of-fit of our model, while the other types would generally indicate that either some assumptions might have been violated.

To plot a residual plot, suppose we have already computed $(\hat{\beta}_0, \hat{\beta}_1)$, the steps are:

	x_i	x_1	x_2	$\dots$	x_n
Step 1	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$	$\hat{y}_1$	$\hat{y}_2$	$\dots$	$\hat{y}_n$
Step 2	$e_i = y_i - \hat{y}_i$	e_1	e_2	$\dots$	e_n

Now that we have $\{(x_1, e_1), \dots, (x_n, e_n)\}$, we plot e_i against x_i for all i. Figure 3 is an example that could confirm that our model is a good one, while figure 4 gives examples of a variety of ways things could go wrong.

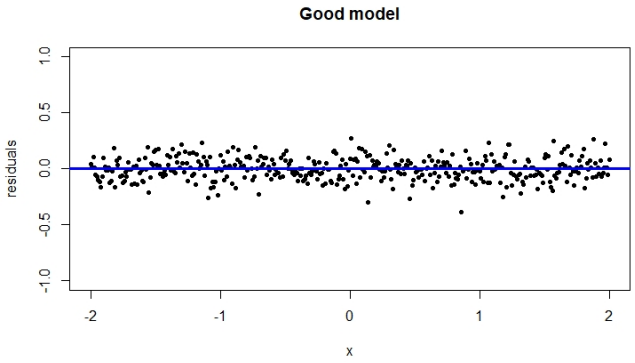


Figure 3: 2 observations can be made: (1): the magnitude of each residual is small, almost all are less than .5, and (2) the signs of the residuals are random between positive and negative.

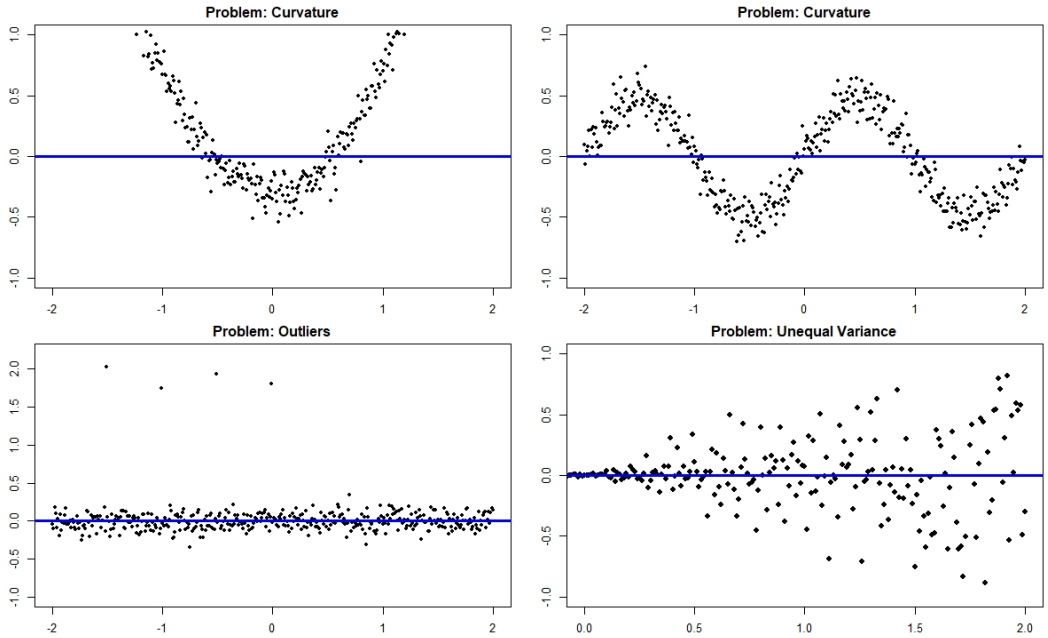


Figure 4: Some examples of what a residual plot could tell regarding our model and dataset.

3 Randomness & Probability

Now that we have been discussing what to do given some data, the study of which is *Statistics*, we are going to take a break and instead explore how those data can be generated given some "rules", which is the study of *Probability*.

3.1 Definitions

1. Probability space: which takes in 3 inputs: $(\Omega, \mathcal{F}, \mathbb{P})$:
   - Ω: the sample space, consisting of all the possible outcomes
   - $\mathcal{F}$: a collection of events, each of which is a set of some outcomes
   - $\mathbb{P}$: a probability measure, which gives each event a probability
2. Complement: given an event A, the *complement* of A, denoted A^C, is the set of all events not A.
3. Disjoint: two events A and B are *disjoint* if they cannot both happen.
4. Independent: an event A is *independent* of an event B if knowing B gives no information on determining A.
5. Random variable: a quantity whose value is determined by the outcome of a random experiment.
   (a) Discrete r.v.: takes on values in some sequence $\{x_n\}_{n=1} = \{x_1, x_2, \dots\}$
   (b) Continuous r.v.: takes on any value in some interval.

Notes: When possible, it always helps to draw pictures of all the events in question, and how they interact with one another.

3.1.1 Conditional Probability

We say A conditioned on B, denoted $A|B$, when we want to determine A given that B has already happened.

3.2 Probability Rules

There are 3 important rules when working with probability, which are true for all probability spaces:

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

where $A \cup B$ is the event that either A or B or possibly both happen, and $A \cap B$ is the event that both A and B happen. In particular, if A and B are *disjoint*, we have:

$$\mathbb{P}(A \cap B) = 0 \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

3.3 Bayes' Rule

$$\mathbb{P}(A \cap B) \;=\; \mathbb{P}(A|B)\mathbb{P}(B) \;=\; \mathbb{P}(B|A)\mathbb{P}(A) \tag{1}$$

so that if $\mathbb{P}(B) > 0$, we have:

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

In particular, if A and B are *independent*, we have:

$$\mathbb{P}(A|B) \;=\; \mathbb{P}(A) \quad \iff \quad \mathbb{P}(A \cap B) \;=\; \mathbb{P}(A) \cdot \mathbb{P}(B)$$

We can generalize the formula (1) to a case of many events:

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) \;=\; \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \ldots \mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

3.3.1 Probability Tree

Sometimes, it can be useful to illustrate a chain of events with a probability tree. Figure 5 gives a simple example of how different events can be linked, chained, and related to each other. Let $A_1, A_2, B_1, B_2, B_3, B_4$ be some events. Take the example of drawing cards from a standard deck, then A_1 is the event that the card drawn is red, and A_2 black; B_1 a spade, and so on.

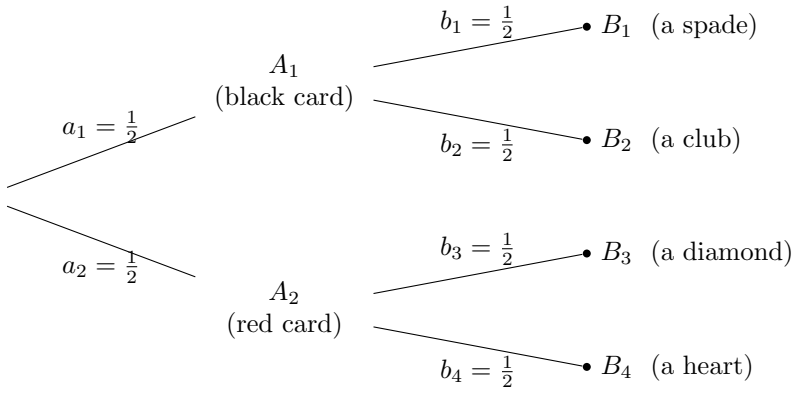


Figure 5: A simple probability tree, with an example of drawing cards from a standard card deck.

Some observations are:

- $a_1 + a_2 \;=\; b_1 + b_2 \;=\; b_3 + b_4 \;=\; 1$
- $a_1 \;=\; \mathbb{P}(A_1) \;=\; 1 - \mathbb{P}(A_2) \;=\; 1 - a_2$
- $b_1 \;=\; \mathbb{P}(B_1|A_1) = \dfrac{\mathbb{P}(B_1 \cap A_1)}{\mathbb{P}(A_1)} \quad \Longrightarrow \quad \mathbb{P}(B_1 \cap A_1) = a_1 b_1$

3.4 Law of Total Probability

Suppose we want to compute the probability of an event A, say $\mathbb{P}(A)$. Instead of direct computation, we can break A into smaller events, similar to the Venn diagram below:

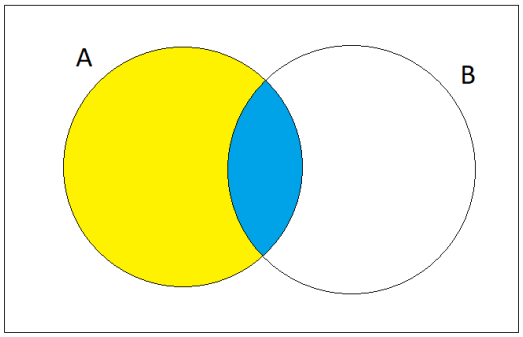


Figure 6: Here, the blue area is the event of both A and B happening, i.e. $A \cap B$, while the yellow area is that of A and not B, i.e. $A \cap B^C$, where B^C is the complement of B. Notice that $A = (A \cap B) \cup (A \cap B^C)$.

Law of Total Probability:

$$\mathbb{P}(A) \;=\; \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^C) \;=\; \mathbb{P}(A|B)\;\mathbb{P}(B) + \mathbb{P}(A|B^C)\;\mathbb{P}(B^C) \tag{2}$$

We can generalize the law: given a finite partition or countably infinitely partition $\{B_n : n = 1, 2, \dots\}$ of the sample space Ω, we have:

$$\mathbb{P}(A) \;=\; \sum_i \mathbb{P}(A \cap B_i) \;=\; \sum_i \mathbb{P}(A|B_i)\;\mathbb{P}(B_i)$$

Recall the Bayes' rule, if $\mathbb{P}(A) \neq 0$, we can reexpress 2 as:

$$\mathbb{P}(B|A) \;=\; \frac{\mathbb{P}(A|B)\;\mathbb{P}(B)}{\mathbb{P}(A)} \;=\; \frac{\mathbb{P}(A|B)\;\mathbb{P}(B)}{\mathbb{P}(A|B)\;\mathbb{P}(B) + \mathbb{P}(A|B^C)}$$

3.5 Random Variables

Random variables can be classified into 2 classes:

- discrete r.v.: which takes on single values $\{x_1, x_2, \dots\}$, each with *probability mass function* p_X such that:

$$\forall x_i \in \Omega : p_X(x_i) \geq 0; \qquad \text{and} \qquad \sum_{x_i \in \Omega} p_X(x_i) = 1$$

- continuous r.v.: which takes on values in some intervals, for example $[x_1, x_2]$ (note that in general, it does not matter if the intervals are inclusive of the end points), each of which has *probability density function* f_X:

$$\forall x \in \Omega : f_X(x) \geq 0; \qquad \text{and} \qquad \int_{x \in \Omega} f_X(x)\,\mathrm{d}x = 1$$

Analogous to the different measures introduced earlier in descriptive statistics, we define the probability over a range of values, *expected value*, and *variance* similarly:

	Discrete r.v.	*Continuous r.v.*
$\mathbb{P}(a \leq x \leq b)$	$\sum_{a \leq x_i \leq b} p_X(x_i)$	$\int_a^b f_X(x)\,\mathrm{d}x$
$\mathbb{E}[X] = \mu_X$	$\sum_{x_i \in \Omega} x_i p_X(x_i)$	$\int_\Omega x f_X(x)\,\mathrm{d}x$
$Var[X] = \sigma_X^2$	$\sum_{x_i \in \Omega} (x_i - \mu_X)^2 p_X(x_i)$	$\int_\Omega (x - \mu_X)^2 f_X(x)\,\mathrm{d}x$

Similarly, given $Var[X]$, the standard deviation of X is $sd[X] = \sqrt{Var[X]}$. Regarding the *variance*, it is sometimes more convenient to use this alternative formula when computing:

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = \mathbb{E}[X^2] - \mu_X^2$$

Note the use of notations. Traditionally, we use μ_X to denote the (theoretical) expected value of a random variable X, and $\bar{x}$ to denote the average value of the data points X.

3.5.1 Properties of Expected Value

Suppose X is a random variable (true for both *discrete* and *continuous* cases), and real numbesrs a, b, c:

$$\begin{aligned} \mathbb{E}[X+c] &= \mathbb{E}[X]+c \\ \mathbb{E}[cX] &= c\mathbb{E}[X] \\ \mathbb{E}[X+Y] &= \mathbb{E}[X]+\mathbb{E}[Y] \\ \mathbb{E}[aX+bY+c] &= a\mathbb{E}[X]+b\mathbb{E}[Y]+c \qquad \text{(Linearity of Expectation)} \end{aligned}$$

3.5.2 Properties of Variance

Suppose X is a random variable (true for both *discrete* and *continuous* cases), and real numbesrs a, b, c:

$$\begin{aligned} Var[X+c] &= VarE[X] \\ Var[cX] &= c^2 Var[X] \qquad (\implies SD[cX] = |c|SD[X]) \\ Var[X+Y] &= Var[X]+Var[Y] \qquad \text{(if } X \text{ and } Y \text{ are independent)} \\ Var[aX+bY+c] &= a^2 Var[X]+b^2 Var[Y] \qquad \text{(if } X \text{ and } Y \text{ are independent)} \end{aligned}$$

3.6 Models & Distributions

3.6.1 Discrete Random Variables

Some of the most common distributions for *discrete r.v.* are:

1. *Bernoulli* distribution: models a binary event of either success (with probability p) and failure otherwise;
2. *Geometric* distribution: models a sequence of independent *Bernoulli* trials, in which we are interested in the number of trials it takes to get to the 1^{st} success;
3. *Binomial* distribution: models a sequence of independent *Bernoulli* trials, in which we want k success out of a total of n trials ($n \geq k$);
4. *Poisson* distribution: models the number of occurrences in a time interval given the average number of occurrences in a unit time interval;

Notation	PMF	$\mathbb{E}[X]$	$Var[X]$
Bernoulli $X \sim Bern(p)$	$\mathbb{P}(X = 1) = p$	p	$p(1-p)$
Geometric $X \sim Geom(p)$	$\mathbb{P}(X = k) = (1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial $X \sim Bino(n, p)$	$\mathbb{P}(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$	np	$np(1-p)$
Poisson $X \sim Pois(\lambda)$	$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$	λ	λ

Table 2: Summary of the most common *discrete* r.v. distributions.

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

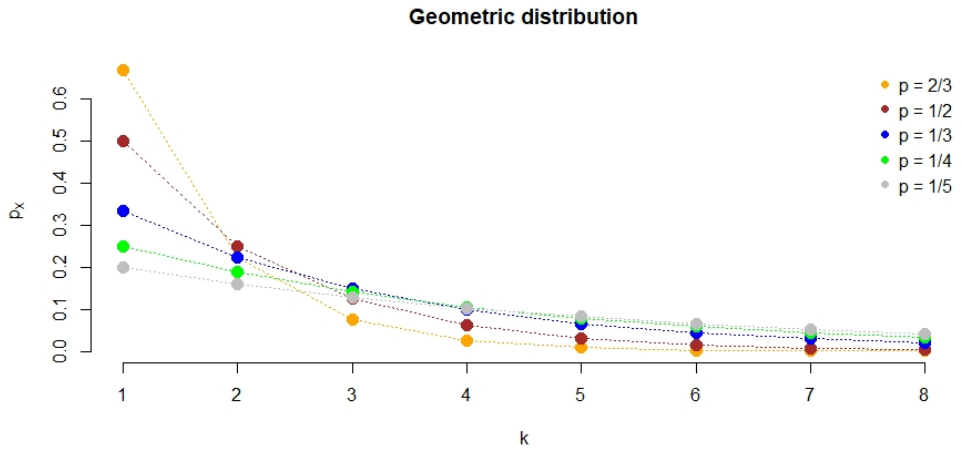
Note that in those distributions, once we know the parameters $p, n, \lambda, \ldots$, we know the distribution completely, i.e. we can compute the probability of any events, the expectation, ...

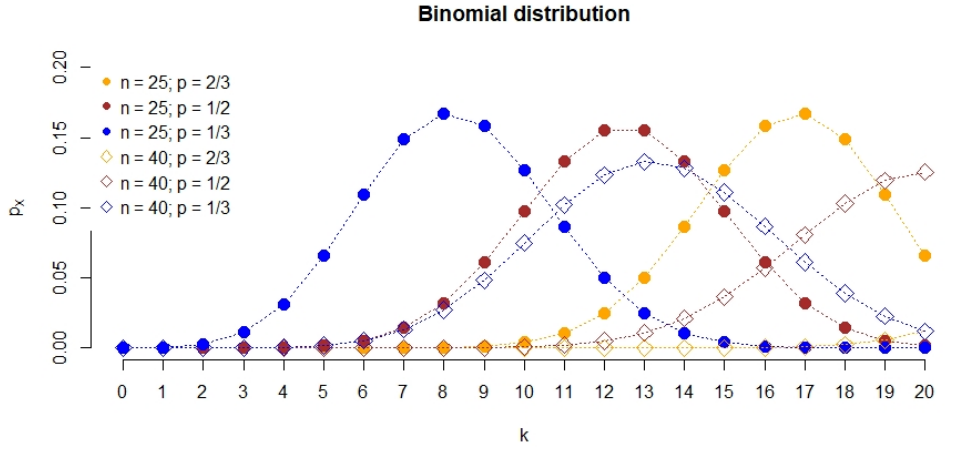
Remark: when n is large and p is small, then the *Binomial* distribution is well approximated by the *Poisson* distribution via

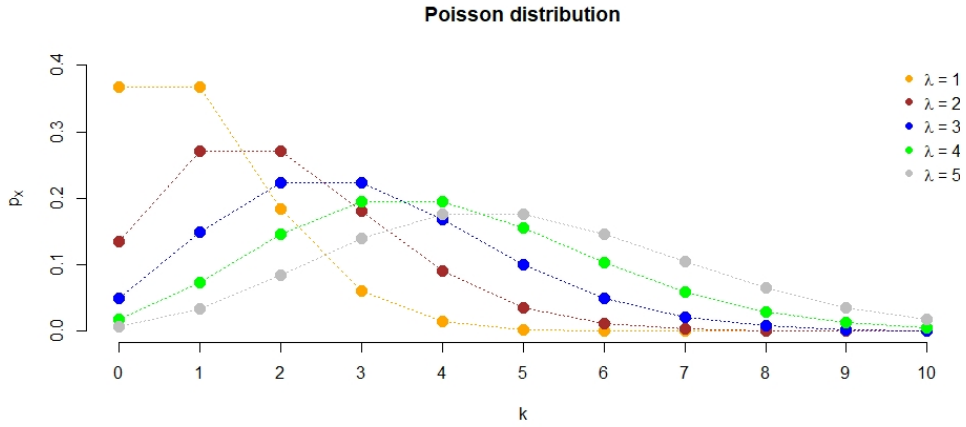
$$\lambda = np$$

an example is when modeling rare events such as the occurrences of earthquakes, tsunami, ...

The following figures give the plots of the density of those distributions, i.e. the *probability mass function* p_X. Note that the above distributions take on integer values only, i.e. $k \in \mathbb{Z}^{\geq 0}$, and are not defined for anything in between integers. As such, the dotted lines in the plots are merely to show the "trend" (or behavior) of such distributions.


Geometric distribution
p = 2/3
p = 1/2
p = 1/3
p = 1/4
p = 1/5
p_X
0.0 0.1 0.2 0.3 0.4 0.5 0.6
1 2 3 4 5 6 7 8
k


Binomial distribution
n = 25; p = 2/3
n = 25; p = 1/2
n = 25; p = 1/3
n = 40; p = 2/3
n = 40; p = 1/2
n = 40; p = 1/3
p_X
0.00 0.05 0.10 0.15 0.20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
k


Poisson distribution
λ = 1
λ = 2
λ = 3
λ = 4
λ = 5
p_X
0.0 0.1 0.2 0.3 0.4
0 1 2 3 4 5 6 7 8 9 10
k

3.6.2 Continuous Random Variables

Some of the most common distributions for *continuous r.v.* are:

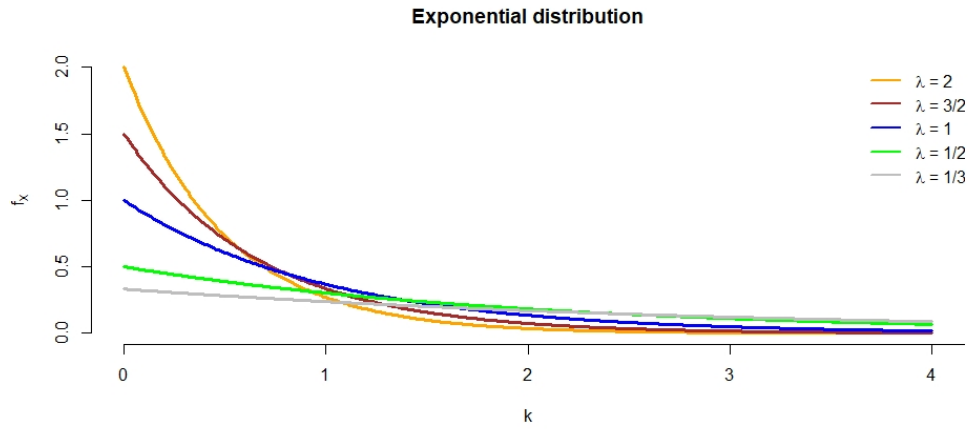
1. *Uniform* distribution: models an event which has equal chances over the entire interval;
2. *Exponential* distribution: models the wait time between 2 consecutive occurrences, given the average number of occurrences in a unit time interval;
3. *Normal* distribution: also known as the bell curve, which is repeatedly observed in real life;

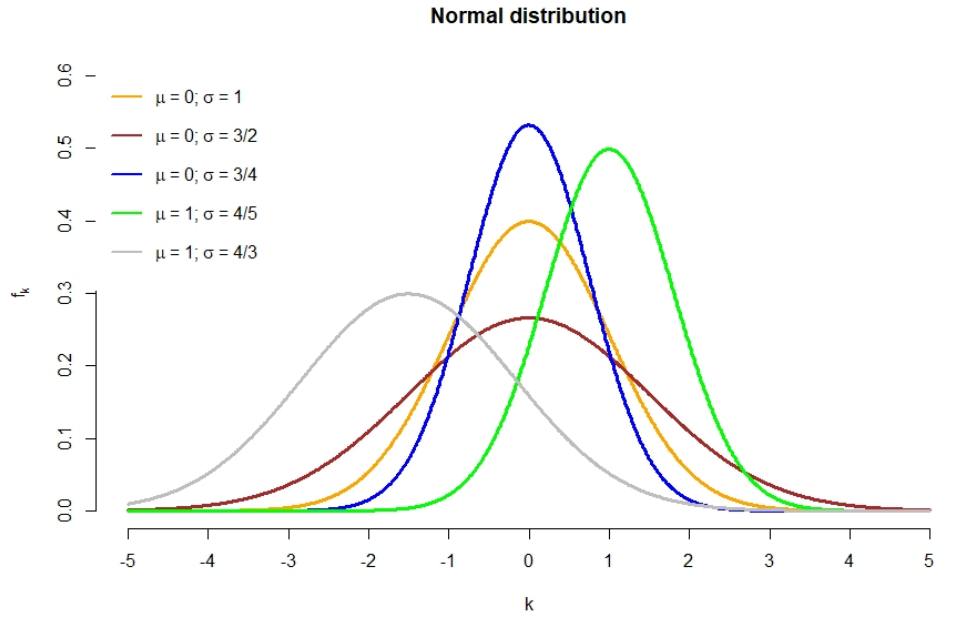
Notation	PDF	$\mathbb{E}[X]$	$Var[X]$
Uniform $X \sim Unif[a,b]$	$f_X(x) = \frac{1}{b-a}$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$
Exponential $X \sim Exp(\lambda)$	$f_X(x) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal $X \sim \mathcal{N}(\mu, \sigma)$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2

Table 3: Summary of the most common *continuous* r.v. distributions.

Remarks:

1. *Memorylessness property* in the *Exponential* distribution: if $X \sim Exp(\lambda)$, then
$$\mathbb{P}(X \geq t+s | X \geq s) = \mathbb{P}(X \geq t) = \int_t^\infty \lambda e^{-\lambda x}\, \mathrm{d}x = e^{-\lambda t}$$
2. A *standard Normal* distribution is when $\mu = 0, \sigma = 1$, in particular, if $Z \sim \mathcal{N}(0,1)$, then
$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$
3. If $X \sim \mathcal{N}(\mu, \sigma)$, then $Z = \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.


Exponential distribution
λ = 2
λ = 3/2
λ = 1
λ = 1/2
λ = 1/3
f_X
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
k


Normal distribution
μ = 0; σ = 1
μ = 0; σ = 3/2
μ = 0; σ = 3/4
μ = 1; σ = 4/5
μ = 1; σ = 4/3
f_k
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-5
-4
-3
-2
-1
0
1
2
3
4
5
k

4 From the Data at Hand to the World at Large

4.1 Statistical Inference

Motivation: Suppose we want to describe statistics about a population. However, as usually observed in real life, there is a multitude of reasons that obtaining the population statistics is practically impossible, instead, what we have are samples of the population. Our objective is to infer the population statistics from the sample statistics.

Now, suppose that we have a sample of $\{x_1, x_2, \ldots, x_n\}$, where each of x_i is *iid* (independent and identically distributed) sampled from the population. There are 2 types of statistics:

proportion (qualitative data)	mean (quantitative data)
$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=1}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

where $\mathbb{1}_{x_i=1}$ is an indicator r.v. of if $x_i = 1$ (a success). In other words, we are taking the naive estimates that p is the proportion of successes in our sample, and $\bar{x}$ the average of our sample data.

Importantly, we have the *expected value* and the *variance* of those estimates:

Estimates	*Expectation*	*Variance*	
$\hat{p}$	$\mathbb{E}[\hat{p}] = p$	$Var[\hat{p}] = \frac{p(1-p)}{n}$	$\hat{p} \sim \left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
$\bar{x}$	$\mathbb{E}[\bar{x}] = \mu$	$Var[\bar{x}] = \frac{\sigma^2}{n}$	$\bar{x} \sim \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Table 4: Expectation and Variance of the estimates $\hat{p}$ and $\bar{x}$.

where the last column is a concise expression of the *expectation* and *standard deviation* (note that though they may look similar to the notation for the *Normal* distribution, they are not and do not infer that the estimates are normally distributed).

4.1.1 Law of Large Number

As shown in table 4, the expected values of the estimates are the true (population) statistics. Indeed, as the sample size gets larger, the sample estimates will converge to the true statistics:

$$\lim_{n\to\infty} \hat{p} = p \qquad \text{and} \qquad \lim_{n\to\infty} \bar{x} = \mu$$

4.1.2 Central Limit Theorem

Idea: the sampling distribution of *any mean* becomes more *Normal* as the sample size increases.

The precise statement is:

- Suppose we have a sample of $\{x_1, x_2, \ldots, x_n\}$, where each x_i is *iid*. Regardless of the true distribution of the r.v. X, the sample mean is approximately *Normal* distributed with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \frac{Var[X]}{n}$ as n goes to infinity.

In other words, if we let $S_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, then:

$$S_n \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}\Big(\ \mu, \frac{\sigma}{\sqrt{n}}\ \Big)$$

where $\xrightarrow[n\to\infty]{\mathbb{P}}$ means as n goes to infinity, the distribution of S_n converges to.

Remarks:

1. In determining how *large* n needs to be before CLT applies, in general:
   (a) for *qualitative data*, we want both $np \geq 10$ and $n(1-p) \geq 10$
   (b) for *quantitative data*, we want $n \geq 30$; and larger the more severe the skewness of the data
2. CLT applies to the *mean*, and not any individual observations.

Put very loosely:

1. *Law of Large Number* implies that when the sample size gets large, our naive estimates get closer and closer to the true parameters.
2. *Central Limit Theoem* implies that when the sample size gets large enough, our naive estimates approximately have a normal distribution.

4.2 Confidence Interval

Now that we have the estimates for the statistics, it is natural to construct some intervals which would (hopefully) contain the true statistics.

4.2.1 Qualitative Data

Recall that we estimate p via $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=1}$, it follows that the *standard error* of $\hat{p}$ is $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

In building a confidence interval, we start with some tolerance α (for example $\alpha = .05$), which in turn gives us the critical value z^* (based on the standard normal distribution); then the $100 \cdot (1-\alpha)\%$ CI is:

$$\Big(\hat{p} - z^* \cdot SE(\hat{p}),\ \hat{p} + z^* \cdot SE(\hat{p}) \Big) = \left(\hat{p} - z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Additionally, we define the *margin of error* as $ME = z^* \cdot SE(\hat{p})$ (note that ME is a function of α, or our CI).

Remarks:

1. **Interpretation of CI:**
   (a) Suppose we are working with a sample of size n, and built a 95% CI, this means that if we were to repeat this experiment many times, each of sample size n, then we would expect that 95% of the CI built would contain the true population statistics p.
   (b) In particular, it will be wrong to say that p has a 95% chance of being in our CI since p, the true statistics, is not a r.v. and hence it does not make sense to give a p a distribution.
2. The large the sample size n, the better the CI.

4.2.2 Quantitative Data

The case for quantitative data is a bit different. First we introduce the t-distribution.

A r.v. X is said to follow a *t-distribution* with *degree of freedom* ν, denoted $X \sim T_\nu$, if

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $\Gamma(x)$ is the Gamma function. (Don't worry about this. You will not be responsible for knowing the details of the t-distribution.) The significance of the t-distribution is that as ν gets larger, the t-distribution gets closer to the Normal distribution:

$$T_\nu \xrightarrow[\nu\to\infty]{} \mathcal{N}(0,1)$$

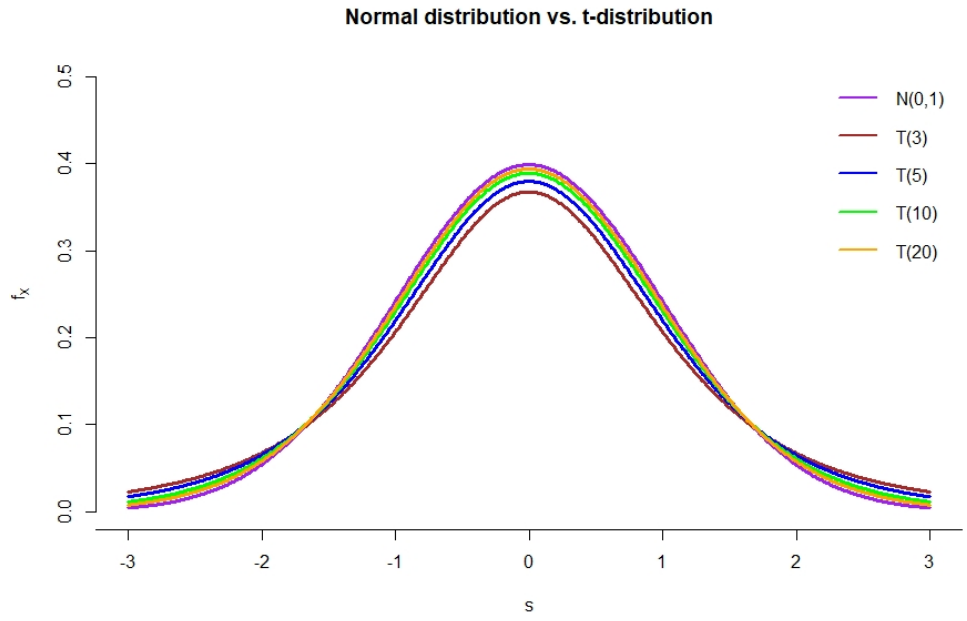


Figure 7: Observe that as n increases, the t-distribution is approximately standard Normal distribution.

Recall that regardless of the distribution of x_i, we always have that

$$\bar{x} \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}\Big(\ \mu, \frac{\sigma}{\sqrt{n}}\ \Big) \tag{1}$$

However, (1) is only an approximation. Now, if we impose a condition that $x_i \sim^{iid} \mathcal{N}(\mu, \sigma)$, then

$$\frac{\bar{x}-\mu}{s/\sigma} \ \sim\ T_{n-1}, \qquad \text{where}\ \ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

which has the exact distribution for all n: t-distribution with $(n-1)$ degrees of freedom.

Now, when building a CI, similar to the procedure described in the case of *qualitative data*, we start with some tolerance α, which in turn gives us the critical value t^*_{n-1} (based on the t-distribution with $(n-1)$ d.f.); then the $100\cdot(1-\alpha)\%$ CI is:

$$100\cdot(1-\alpha)\%\ \text{CI} = \left(\ \bar{x} - t^*_{n-1}\cdot\frac{s}{\sqrt{n}},\ \bar{x} + t^*_{n-1}\cdot\frac{s}{\sqrt{n}}\ \right)$$

Similarly, we define the *margin of error* as $ME = t^*_{n-1}\cdot\frac{s}{\sqrt{n}}$ (note that ME is a function of α, or our CI).

4.3 Hypothesis Testing

Motivation: Same type of problem (inferring the population from the available samples) but we are now looking from a different perspective: we want to test if our current belief about the population is true, or more precisely, if the observed sample statistics is statistically significantly different from the belief.

We divide into 4 main steps:

1. State the null hypothesis, H_0, and the alternative hypothesis, H_1
2. Calculate the test statistics, $\hat{p}$ if qualitative data and $\bar{x}$ if quantitative
3. Calculate p-value given the test statistics
4. Give a conclusion: either (1) Reject H_0, or (2) Fail to Reject H_0

Remark: In principle, we have the choices to compare either the test statistics or the induced p-value, both will return the same answer. Depending on the context, 1 approach may be more straight-forward.

We consider the problem of hypothesis testing in 3 different contexts:

1. *qualitative data:* test if a proportion $p = p_0$ where p_0 is the current belief;
2. *quantitative data:*
   (a) 1-sample t-test: for if the mean $\mu_X = \mu_0$, where μ_0 is the current belief;
   (b) 2-sample t-test: for if there are differences between the means of 2 different populations in question, $\mu_X - \mu_Y = 0$;

We note that in the 2-sample context, if we can pair them 1-to-1, then it is reduced to the problem of 1-sample t-test. As an example, suppose we have the sample on the number of hours students spend on studying during a regular week and during the finals week:

Regular week	30	52	28	15	32
Finals week	32	46	34	20	32

We can formulate hypothesis testing questions such as:

1. 1-sample t-test: it is believed that students spend on average 30 hours per week on studying, however UCSD students study more than that, we want to know if this is true.
2. 2-sample t-test: it is believed that students spend on average the same number of hours per week on studying regardless of if they are having finals, however UCSD students tend to study more than during finals week, we want to know if this is true.

While the null hypothesis H_0 will mostly be an equality, there are 3 options for the alternative hypothesis H_1. We use the example of qualitative data, testing for proportion p, the other 2 cases are similar with the appropriate test statistics.

$$\begin{aligned} H_0 &: \quad p = p_0 \\ H_1 &: \quad p > p_0; \quad \text{or} \quad p < p_0; \quad \text{or} \quad p \neq p_0 \end{aligned}$$

To draw conclusions, we follow:

H_1	Conclusion 1: Reject H_0	Conclusion 2: Fail to Reject H_0
$H_1 : p > p_0$	$\mathbb{P}(Z \geq z^*) \leq \alpha$	$\mathbb{P}(Z \geq z^*) > \alpha$
$H_1 : p < p_0$	$\mathbb{P}(Z \geq z^*) \geq 1 - \alpha$	$\mathbb{P}(Z \geq z^*) > 1 - \alpha$
$H_1 : p \neq p_0$	$\mathbb{P}(\lvert Z \rvert \geq z^*) \leq \frac{\alpha}{2}$	$\mathbb{P}(\lvert Z \rvert \geq z^*) > \frac{\alpha}{2}$

Table 5: Decision table for hypothesis testing of proportion.

To better demonstrate where we are on the normal curve, these figures below illustrate the hypothesis testing situation. Note that the first 2 figures correspond to the 1-sided test, while the last one corresponds to the 2-sided test.

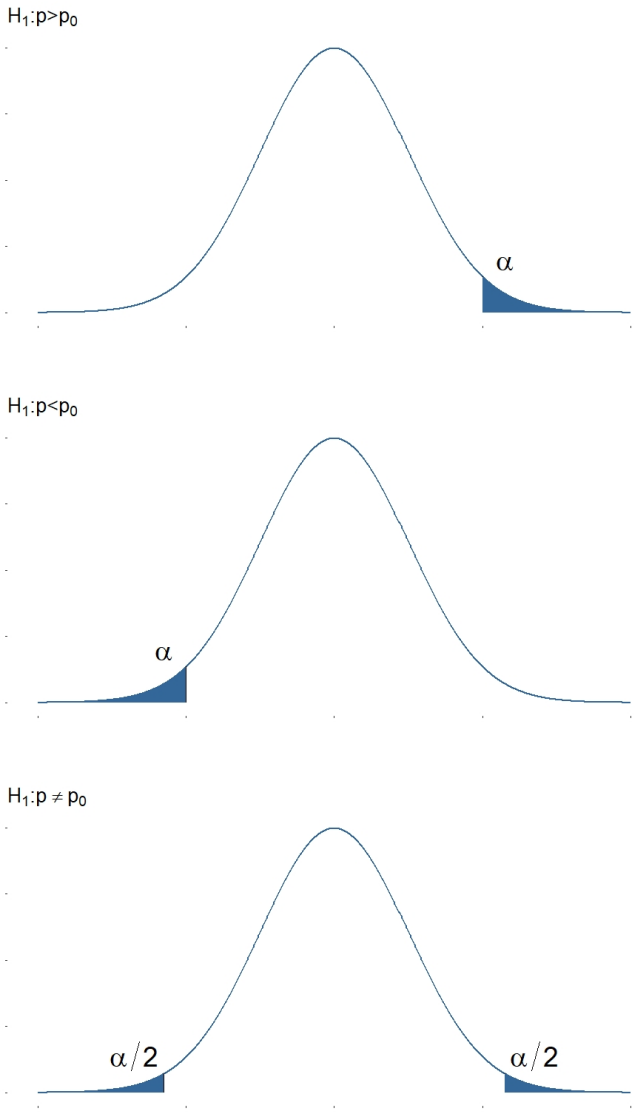


Figure 6 summarizes of the different test statistics and the corresponding distributions.

Data type	Parameter	Test statistics	Distribution
Proportion	$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=1}$	$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\mathcal{N}(0,1)$
1-*sample t-test*	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$t^* = \frac{\bar{x} - \mu_X}{s/\sqrt{n}}$	approximately T_{n-1}
2-*sample t-test*	$d = \bar{x} - \bar{y}$	$t^* = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$	$T_{\min\{n_X-1, n_Y-1\}}$

Table 6: Summary of different test statistics and the corresponding distributions.

Recall that when carrying out hypothesis testing, the conclusion is either "Reject H_0" or "Fail to Reject H_0" (be careful that it would not be right to conclude that H_0 is true, since in so doing, we would have confirmed that H_0 was true). A natural question to follow up is how likely that we are right, or rather how likely is it that we are wrong. We define 2 types of errors:

	Reject H_0	Fail to Reject H_0
H_0 is True	Type 1 Error	Correct
H_0 is False	Correct	Type 2 Error

In particular, we have direct control over the probability of *Type* 1 *Error* by the choice of α, whereas the probability of *Type* 2 *Error* is indirectly influenced by α. Here, we define the power of a hypothesis test to be:

$$power = 1 - \beta$$

where β is the probability of *Type* 2 *Error*. In practice, while we can make assumptions on the null hypothesis H_0, we do not know the alternative hypothesis H_1. As such, we are somewhat limited in our ability to completely control the power of our hypothesis test.

4.4 Regression Inference

4.4.1 Testing for Regression Coefficient

Recall the setting of building *linear regression model* which was introduced at the beginning of the quarter. Suppose we have some data points

$$\big\{(x_i, y_i)\big\} \quad \text{for } i = 1, 2, \ldots, n$$

We modeled this as

$$y_i \;=\; \beta_0 + \beta_1 x_i + \epsilon_i$$

The "best" estimates (in terms of least squares) are

$$\hat{\beta}_0 \;=\; \bar{y} - \hat{\beta}_1 \bar{x}; \qquad\qquad \hat{\beta}_1 \;=\; r \cdot \frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Motivation: Suppose we have now built the model (ie. calculating $\hat{\beta}_0$ and $\hat{\beta}_1$). We want to take a step back and investigate if there was indeed a relationship between x and y. In particular

$$\hat{\beta}_1 \;\neq\; 0 \quad \Longrightarrow \quad x \text{ and } y \text{ were indeed linearly related}$$

We formulate this as a hypothesis test. For example, we want to test

$$H_0 :\; \beta_1 = 0 \qquad \text{vs.} \qquad H_A :\; \beta_1 \neq 0$$

or H_A other one-sided inequality, depending on the context.

We first define some quantities:

$$\begin{aligned} s^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ s_x^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ SE\left(\hat{\beta}_1\right) &= \frac{s}{s_x \sqrt{n-1}} \end{aligned}$$

where $SE(\hat{\beta}_1)$ is the standard error of our estimate $\hat{\beta}_1$. The appropriate test statistics is

$$T \;=\; \frac{\hat{\beta}_1 - \beta_1^{(0)}}{SE\left(\hat{\beta}_1\right)} \;\sim\; T_{n-2}$$

where $\beta_1^{(0)}$ is the current belief for $\hat{\beta}_1$, here $\beta_1^{(0)} = 0$.

4.4.2 Confidence Interval vs. Prediction Interval

Now, given a new value of x, suppose we have already calculated the predicted new value of $\hat{y}$. Given $(1-\alpha)\%$, we can find the appropriate critical t^* value. We then have:

1. *Confidence Interval.* the Margin of Error is:

$$ME_{CI} = t^* \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

   which gives the Confidence Interval:

$$(1-\alpha)\%\ \text{CI} = (\hat{\mu}_y \pm ME_{CI})$$

2. *Prediction Interval.* the Margin of Error is:

$$ME_{PI} = t^* \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

   which gives the Prediction Interval:

$$(1-\alpha)\%\ \text{PI} = (\hat{y} \pm ME_{PI})$$

Remark. The only difference between the ME for Confidence interval and for Prediction interval is that the ME for Prediction interval has a $+1$ term in the square root, which accounts for the error of any particular observation.

4.5 Chi-squared tests

Given observations and expectations, the test statistics is

$$D \;=\; \sum \frac{(Obs - Exp)^2}{Exp}$$

1. **Goodness-of-fit test:** very straightforward where the expected count for group i is

   $$\mathrm{Exp}_i = np_i$$

   where p_i is the default probability of being in group i. The df is $df = k - 1$, where k is the number of groups.

2. **Independence test:** in this case, it is a little more involved to calculate the expected counts. Suppose our data is presented with I groups and J categories, put in a form of a I-row, J-column table, then the expected count of group i and category j is

   $$\mathrm{Exp}_{i,j} = \frac{R_i C_j}{n}$$

   where n is the total sample size, R_i the total number of row i and C_j is the total of column j. The df is $df = (I-1)(J-1)$, ie. the product of $\#(rows) - 1$ and $\#(columns) - 1$.