Chapter 2

Cases: rows of the data table **Respondents:** individuals who answer a survey **Subjects:** people who we experiment on. **Experimental Units:** companies, websites and other inanimate subjects **Records:** rows in a database **Relational Database:** two or more separate data tables are linked so information can be merged across them **Identifier Variable:** ex: student number **Nominal Variables:** categorical values used only to name categories **Ordinal Variable:** individually ordered **Cross Sectional Data:** several variables single point in time

Chapter 3

Voluntary Response Bias: only those who respond are counted
Convince Sampling: only those who are covenant counted.
Sampling Error: cannot be avoided (sample to sample differences)
Non-response bias: when large amount doesn't respond
Sampling Frame: list from which sample is drawn

Sample: small group of population

Statistic: example % of sample who prefer X

Population

Parameter: % of population who prefer X

Simple Random Sample	• every sample of the size has equal chance of being selected • Chosen with random numbers
Stratified Sampling	• Split population into like groups (strata) than use SRS within each • Used to guarentee proportion
Cluster Sampling	• Splitting the population into clusters that each represent population • Preform census within each
Systematic Sampling	• Select every 10th person on a list

Chapter 4

Frequency Table: tells us how frequently we find something
Relative Frequency Table: percentage of how many there are relative to total it represents **Bar Chart:** bars should be same width, space between bars of whole **Contingency Table:** shows how data distributed along each variable cases that satisfy condition of another variable **Segmented Bar Charts:** treats each bar as a whole and divides it proportionately into segments corresponding to percentages is the group **Area Principle:** area occupied by graph should correspond to magnitude of value **Pie Charts:** whole group of cases as a circle, slice is proportionate to the fraction **Conditional Distribution:** shows distribution of one variable for those

Frequency Table

Province	Corporate Stores
Newfoundland	13
PEI	4
Nova Scotia	34
New Brunswick	21
Quebec	223
Ontario	155

Relative Frequency Table

Province	Corporate Stores (%)
Quebec	49.56
Ontario	34.44
Nova Scotia	7.56
Other	8.44
Total	100.00

Contingency Table

	Graduated	Failed to Graduate	Total
Exper	73	12	85
Control	43	39	82
Total	116	51	167

Mode
- humps in a graph
- **Unimodal:** one main hump
- **Bimodal:** two humps
- **Multimodal:** three or more
- **Uniform:** no modes approx. all same height

Symmetry
- **Symmetric** if can be split in two parts that look the same
- **Tails:** thinner ends of distribution
- **Skewed:** if one tail stretches further than other
- **Outliers:** stand off from the body of distribution

Chapter 5

Steam and Leaf Display
- Like histograms but also give individual values
- To display changes 2.06, 2.22, 2.44, 3.28 and 3.34
- 2|124 3|33

Histograms
- Plots bin counts as height of bars, displays entire distribution of price changes, no gaps
- Width of bins is important; n data points use log2n bins

Centre
- Average data to find center
- y=all of the values
- n=number of values
- **mean** = y/n
- find median by counting in from ends to middle
- (n+1)/2

Spread
- **Range** = max-min
- **Quartiles:** values that frame middle 50% of data
- One quarter of data lies below lower quartile Q1 and one quarter lies above upper quartile Q3
- **IQR:** Q3 – Q1

How to find quartiles
- Find median of set of data values (if odd use middle number in both halves
- Divide both sections into two again, lower half is Q1 and upper is Q3

Variance
- $s^2 = \frac{(y-mean)^2}{n-1}$

Standard Deviation
- square root of variance

Grouped Data
- $s^2 = (y - mean)^2 p$
- p = percent of the population

5 Number Summary
- **Max**
- **Upper Quartile Q3**
- **Median**
- **Lower Quartile Q1**
- **Min**

How to Make Boxplot
1. Draw axis spanning extent of data
2. Draw short horizontal lines at lower, upper quartiles and median. Then connect to make box
3. Fences (dotted line) at 1.5 IQR above and 1.5 IQR below
4. Grow "whiskers" draw lines to each of most extreme values within the fence
5. Now show any outliers by displaying data

Calculate Percentiles
- put the data in ascending order
- Suppose we want to calculate 80th percentile. If 12 data values we calculate 80% of 12 which is 9.6.
- YOU CAN EITHER
- Round to 10 and find 10th value
- Suppose we want 50% of 12 which is 6. Take average of 6 and 7th values. Median is 50th percentile

Z Score
- $z = \frac{SUMy - mean}{s}$
- To find how many standard deviations from the mean
- To identify outliers z>3 or z<-3

Covariance
Cov(x,y)=r*s(x)*s(y)
s = standard deviation
r = corelation

Correlation
$r = \frac{SUM(x-mean)(y-mean)}{(n-1)*s(x)*s(y)}$
- x and y mean a single value

Probability Trees

Reverse Conditioning
$P(case|fixed) = \frac{P(case\ and\ fixed)}{P(fixed)}$
$P(case|fixed) = \frac{0.48}{0.48-0.15-0.01} = 0.75$

Chapter 8

Empirical Probability: based on repeatedly observing the events outcome
Theoretical Probability: whenever outcomes equally likely $P(A) = \frac{\#\ of\ outcomes\ in\ A}{Total\ \#\ of\ outcomes}$
Subjective Probability: your personal assessment of an event

Conditional Probability
$P(B|A) = \frac{P(AandB)}{P(A)}$
-to find probability of B given A

Probability Rules
Complement Rule:
$P(A) = 1 - P(A^c)$
Multiplication Rule:
P(A and B)=P(A)*P(B)
Addition Rule:
P(A or B)=P(A) + P(B)
General Addition Rule:
P(AorB)=P(A)+P(B)-P(AandB)

Bernoulli Trials
- only two outcomes
- probability same for each trial
- Trials are independent

Chapter 9

Expected Value
E(X)=SUM(x*P(x))
Deviation
(x-E(x))
Variance
Expected value of those deviations replacing x
Standard Deviation: SD(x)= $\sqrt{variance}$

Adding/Subtracting Random Variables
E(X+/-Y)=E(X)+/-E(Y)
Var(X+/-Y)=Var(X)+Var(Y) –always add

Dealing with Correlation
Var(E-I)=Var(E)+Var(I)-2*SD(E)*SD(I)*r
SD(E-I)= $\sqrt{Variance}$

Binomial Distribution
- number of successes in n trials
- k=successes
- n=trials
- p=probability of success
- q=(1-p)
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- P(k successes in 5 trials)= $\binom{n}{k} p^k q^{n-k}$
- E(Y) or MEAN=np
- SD(Y)=$\sqrt{npq}$

Poisson Probability
λ= mean number of occurrences
X=number of occurrences
P(X=x)=$\frac{e^{\lambda}\lambda^{x}}{x!}$
Expected value: E(X) = λ
Standard Deviation: SD(X) = $\sqrt{\lambda}$
e=2.71828
if λ is average hits per min, than x is how many hits you want to find P of

Uniform Distribution
Density function $\begin{cases} \frac{1}{b-a} & if\ a \le x \le b \\ 0 & otherwise \end{cases}$
E(X)=$\frac{a+b}{2}$
Var(X)=$\frac{(b-a)^2}{12}$

Normal Distribution
z=$\frac{y-\mu}{\sigma}$ then look in chart for probability
- approx. normal if less than 3 SD from mean

Exponential Distribution
$F(x)=\lambda e^{-\lambda x}$
To find probability between two values s and t
$P(s \le X \le t) = e^{-\lambda s} - e^{-\lambda t}$

Sampling Distribution Model
-distribution of proportions over many independent samples from same population
mean is p, the true proportion
n = sample size
SD($\hat{p}$) = $\sqrt{\frac{pq}{n}}$

Central Limit Theorem
- mean of a random sample has a sampling distribution whose shape can be approximated by a normal model. The larger the sample, the better approximation will be

SE($\hat{p}$)=$\sqrt{\frac{\hat{p}\hat{q}}{n}}$ SE($\bar{y}$) = $\frac{s}{\sqrt{n}}$
$\hat{p}$ sample proportion

Sampling Distribution of the Mean
When random sample is drawn from any population with mean μ and its sample mean $\bar{y}$ has sampling distribution with same mean, but SD(x) $\frac{\sigma}{\sqrt{n}}$
-sampling distribution approx. large enough when, larger the better

1 **Z-score** = $\frac{(x-\mu)}{\sigma}$

2 **Z-score** = $\frac{(x-\mu)}{\sigma/\sqrt{n}}$

3 **Z-score** = $\frac{(x-\mu)}{[\sigma/\sqrt{n}]*\sqrt{\frac{N-n}{N-1}}}$

4 **T-score** = $\frac{(x-\mu)}{s/\sqrt{n}}$ df = n-1

8 **Normal Approximation**
Z-score = = (x-μ +/– CC)/σ
Correction for Continuity = CC = 0.5
if <= or >, add 0.5 if >= or <, subtract 0.5

9 **Z-score** = $\frac{(\hat{p} - p)}{\sqrt{\frac{p(1-p)}{n}}}$ $\hat{p}$ = x/n

10 **Z-score** = $\frac{(\hat{p} - p)}{\sqrt{\frac{p(1-p)}{n}} * \sqrt{\frac{N-n}{N-1}}}$ $\hat{p}$ = x/n

Binomial -

5 **Uniform** - all outcomes are equally likely
f(x) = 1/ (b-a) b = upper limit a = lower limit
E(x) = u= (b+a)/2 VAR(x) = (b-a)^2/12
SD(x) = σ=SQRT(VAR(x))
Formulas:
P(X>x) = (b - x) *1/ (b - a) P(X<x) = (x - a) *1/ (b - a)

6 **Exponential** - must be given a rate of success
P(t=t) = e^[-λ(t)] mean = 1/λ variance = 1/λ^2

7 **Poisson** - must be given a constant rate of success
P(x success in t periods) = [e^(-λt)*(λt)^x]/x!
E(x) = λt VAR(x) = λt
-"x" can have multiple values. We must apply the formula for all values and add the probabilities.

11 **Binomial** - 2 outcomes, success or failure, trials must be independent, "p" and "q" remain constant.
P(X=x) = [n!]/[x!(n-x)!] * p^x * q$^{(n-x)}$ = nCx * p^x * q$^{(n-x)}$
-"x" can have multiple values like in the poisson.
-if np>=10 and nq>=10, use normal approximation.
E(x)= np VAR(x) = n*p*(1-p) SD(x) = σ=SQRT(VAR(x))

NOTES - Z-score and T-score represents the # of standard errors a data point is away from the mean.
- Once Z-scores (T-scores) are calculated you can convert to a probability using the Z Table (T Table)
- RED FLAG - Occurs when the probability of an outcome(s) is less than 0.05. This can suggest that the pop mean or proportion is not correct.

Variables
x=mean, #off successes, etc.
μ=population mean
σ=population SD
n=sample size
s=sample SD
p=population prop.
q=1-p
$\hat{p}$=sample prop.
$\hat{q}$=1-$\hat{p}$
$Z_{\alpha/2}$=from Z-table
$T_{\alpha/2}$=from T-table
k=# of SD
e=exp. function
t=# of periods
λ=rate of success
N=population size
M=# of x available for sel.
C="nCr" on calculator

For Example Working with the sampling distribution of the mean

Suppose that the weights of boxes shipped by a company follow a unimodal, symmetric distribution with a mean of 12 kg and a standard deviation of 4 kg. Boxes are shipped in pallets of 10 boxes. The shipper has a limit of 150 kg per pallet for such shipments.

Question: What's the probability that a pallet will exceed that limit?

Answer: Asking the probability that the total weight of a sample of 10 boxes exceeds 150 kg is the same as asking the probability that the *mean* weight exceeds 15 kg. First we'll check the conditions. We will assume that the 10 boxes on the pallet are a random sample from the population of boxes and that their weights are mutually independent. We're told that the underlying distribution of weights is unimodal and symmetric, so a sample of 10 boxes should be large enough. And 10 boxes is surely less than 10% of the population of boxes shipped by the company.

Under these conditions, the CLT says that the sampling distribution of $\bar{y}$ has a Normal model with mean 12 and standard deviation

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{10}} = 1.26 \text{ and } z = \frac{\bar{y} - \mu}{SD(\bar{y})} = \frac{15 - 12}{1.26} = 2.38$$

$$P(\bar{y} > 15) = P(z > 2.38) = 0.0087.$$

So the chance that the shipper will reject a pallet is only .0087—less than 1%.