Stats 3Y03 Summary

Note: R might be on the final :$

# Chapter 1

**Categorical variable**: qualitative variable, such as funny; limited number of options

* e.g. Blood type, Political party
* It can still be a number if the number doesn’t describe a quantity
* **Ordinal**: Values that can be ordered, such as academic grade
* **Nominal**: Values that cannot be ordered, such as brand name

## Types of variables

**Numerical variable**: quantitative variable, such as position

* **Continuous**: decimals
* **Discrete**: integer

**Univariate Data**: single variable

**Bivariate Data**: 2 variables (not required in this course)

**Multivariate Data**: more than 2 variables

**Probability**: average of population is from average of sample

**Inferential statistics**: average of sample is from average of population

**Sampling Frame**: list of things in a list that can be sampled

* telemarketers' sample frame is the people with a phone number in the phone book/phone archive of the company
* when doing a culture study of farms, the sample frame could even be a map

**Enumerative study**:

* identifiable goal
* well-defined, unchanging sample frame
* enumerate (explain, evaluate, describe) a condition that exists with the existing population

**Analytic study**:

* focused on improvement of the process which created the results and which will continue creating results in the future
* no well-defined sampling frame

## Target population

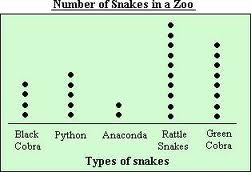
* population you want to be collecting data from
* **sample population** is the population you are collecting data from
* sample population is usually subset of target population
* sample population is useful when the target population is too large
* sometimes it is not the same as the sample population
  + e.g., when informing factory workers that their productivity is being observed, they'll act differently

**Simple random sample**: from entire population

**Stratified random sample**: from a sub-population (1 from each row)

**Convenience sample**: not entirely random; what is easy to obtain (first row)

**Dot plot**: quantifying increments and representing them by dots



**Mean**: average

**Median**: middle value; if length of set is even, average of (n+1)/2 and n/2; if length of set is odd, (n+1)/2

**Mode**: common number

**Unimodal**: 1 peak

**Bimodal**: 2 peaks

**Multimodal**: more than 2 peaks

Graphs can also be **symmetric** or **asymmetric**, which is when the top half of the boxplot looks similar to the bottom half.

**Left skew**: mostly on right side

**Right skew**: mostly on left side

Graphs can also be **unskewed**.

**Outliers**:

* values that must be mistakes or abstract exceptions
* > 1.5 × forth spread (see below) beyond closest quartile
* **extreme outlier** is > 3 × forth spread

Each data set is split up into 4 **quartiles**.

Q1: median of bottom half (includes middle number if odd length)

Q2:

Q3: median of top half (includes middle number if odd length)

**The Five-Number Summary**:

1. Minimum
2. Q1
3. Q2
4. Q3
5. Maximum

The range, minimum, and maximum can include outliers

**range**: max − min

## Variance

**Variance**: distribution of range

*N* is [target population](#_Target_population) size

*xi* are the values

*n* is [sample population](#_Target_population) size

**Trimmed mean**: mean calculated by trimming away a given percentage of elements (relative to number of elements) from the top and bottom. If the percentage gives a non-discrete number of elements, you have to calculate multiple trimmed means and find the mean of the 2 trimmed means

**Population mean**: expected outcome of mean of [target population](#_Target_population), i.e. average given a theoretically infinite amount of measurements; a.k.a. true mean, expected value



**Sample mean**: average given finite number of inputs; an estimate of the population mean



**Sample median**:

**Sample variance**: 

**Population variance**: 

**Spread**: interquartile range

## Standard Deviation

**s.d.**

* Average distance from the mean
* Larger s.d. means more spread
* i.e. when all values are the same, s.d. = 0
* Square root of variance = 

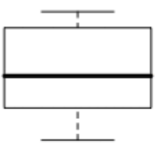
**Degrees of freedom**: n – 1

Another measure of spread is **interquartile range** or **forth spread**. (Q3 – Q1)

**Whiskers**: minimum and maximum points of the range that does not include outliers

**Boxplot**:

* Top and bottom lines are whiskers
* Box surrounds forth spread
* Middle line is median
* Can be vertical or horizontal
* Outliers are still placed on boxplots, using circles (o) or stars (\*)
* (a.k.a. Boxplot-and-whisker plot)



# Chapter 2

This is similar to the logic course [SFWR ENG 2FA3](https://docs.google.com/document/d/12b0YsgYtB3cnhiu39eKqpl6b2guSE04wUk4Gg8G9nt4/edit?pli=1).

Probability is between 0 and 1

**Sample space**: all possible outcomes

The size of the sample space is: outcomesevents.

N: number of outcomes for an event

N(A): number of outcomes in sample space, A

**Relative frequency probability**: events that occur frequently, such as rolling dice or buying lottery tickets

**Relative frequency** of a value

**Personal probability**: events that cannot be repeated or non-random events with unknown quantities that is based on belief of an individual

**Coherent**: personal probability of one event does not contradict personal probability of another

Sometimes you can have an **infinite number of possible outcomes**. For example, if you are testing something until failure, you will repeat testing until success {S, FS, FFS, …}



If there are a given number of outcomes, such as 1 through 6 for a dice, and a sample space, A, such as containing all odd outcomes, A’, the **complement**, contains everything A does not, such as all even outcomes. Therefore, P(A) + P(A’) = 1

**Simple Event**: Only one way to get each outcome

**Compound Event**: Multiple ways to get the same outcome

## Replacement

**With replacement**: e.g. if you are picking names out of a hat and you put the names back after each pick; independent

**Without replacement**: when you use each option only once; dependent

## Mutually-Exclusive Events

**Mutually exclusive** (a.k.a. disjoint event): 2 outcomes cannot occur simultaneously; A ∩ B = Ø; e.g. rolling a dice can either be 3 or 5−not both, whereas it being 3 or odd is not mutually exclusive

The **probability** is the sum of the probability of each individual event:



For ordered pairs, number of possible arrangements is: *N*!

**Permutations** are ordered sequences that are made up by *k* elements that are a subset of a set of *n* elements.

The notation for **number of permutations** is: 

**Bayes’s Theorem**: 

## Non-mutually exclusive events

For non-mutually exclusive events, there can be overlap, so:

For ordered pairs, number of possible arrangements for k events is:

Unordered permutations are known as **combinations** (n choose k). They are denoted:

For unordered pairs, number of combinations is:, where n is the number of objects and k is the size of the group (pick k, 5, players for the team from n, 8 people. number of permutations?)

**Dependent**: you can’t put it back

**Independent**: you can put it back

**Conditional probability**: Probability of A given B: 

# Chapter 3

## Random variables

**rv**

* function whose domain is the sample space and whose range is the set of real numbers, but is subject to random variations
* denoted by a capital letter, whereas its values have the same letter as the rv, but lower-case
* can either be [continuous or discrete](#_Types_of_variables)
* *x* is a particular value of a [random variable](#_Random_variables)

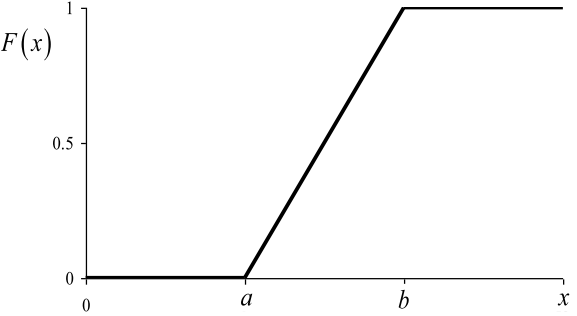
**Bernoulli**: binary output; can only be either a 0 or a 1

**Probability Mass Function (pmf)**: a function that gives the probability that a [discrete](#_Types_of_variables) [random variable](#_Random_variables) is exactly equal to some value

## Cumulative Distribution Function

**CDF**: add up all probabilities within a given range





## Expected Value

* mean using probability of [discrete](#_Types_of_variables) rv’s
* gives same result as population mean
* use if you’re not given data, but given probability
* 
* [Variance](#_Variance): 
* **General Expectation formula**:

[**Variance**](#_Variance) **of** [**CDF**](#_CDF)



## Binomial experiment

1. fixed trial
2. 2 outcomes−success or failure
3. Trials are independent ([with replacement](#_Replacement))
4. Probability of each outcome is the same for each trial

* If the sample size is at most 5% of the population size, the experiment can be analyzed as though it were a binomial experiment ([with replacement](#_Replacement)).
* *n*: repetitions of trials
* *p* = P(success in single trial)
* *q* = P(fail in single trial)
* *x*: total number of successes
* 
* Note: the above notation can be read, where *x* is a variable in *b* and *n* and *p* are constants

**Hypergeometric** (H.D.): same as [binomial](#_Binomial_experiment), but dependent ([without replacement](#_Replacement))

* *N*: number of items in population
* *M*: number of successes in population
* *n*: number of items in sample
* *x*: number of successes in sample
* 
*  (same as binomial)
* 

**Negative Binomial Distribution**:

* n is fixed in [*binomial*](#_Binomial_experiment), whereas *here*, n is random
* trials repeated until success we want
* *r* is the number of successes you want
* If r = 1, this is known as a **geometric distribution**

**Poisson distribution**:

* discrete pdf
* number of occurrences of an event in a given interval, given average rate and time (independent), since last event
* 
* *x*: you are determining the probability that *x* things will happen
* *λ* (or *μ*): average occurrences given population (multiply average rate by population)
* mean = variance = λ, so [S.D.](#_Standard_Deviation) = √λ
* *α* – expected number of events during unit interval
* *t* – time interval length
* *λ = αt*
* 

**Exponential**: time between events, whereas poisson is more the number of events; continuous distribution

Expected value: 



For ranges, 

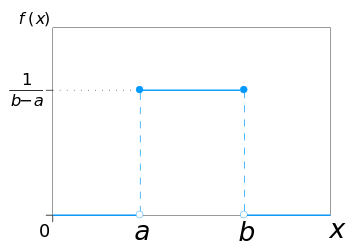
# Chapter 4

## Probability Density Function

**PDF**: a function that gives the probability that a [continuous](#_Types_of_variables) random variable is exactly equal to some value, such that:

Area under whole curve = 1

**Uniform Distribution**: if a [continuous](#_Types_of_variables) random variable, *X*, has a [pdf](#_PDF), f(x; a, b):

[](https://upload.wikimedia.org/wikipedia/commons/9/96/Uniform_Distribution_PDF_SVG.svg)

Note: *a* and *b* do not represent the entire range of the [PDF](#_PDF). Just look at the f(x) formula above!

To get [pdf](#_PDF) from [cdf](#_CDF), take the derivative of the [cdf](#_CDF).



## Percentile

**Percentile**: percentage of data below you; relative to other data in the range

* *p*: percentile
* *η*: percentile function
* 



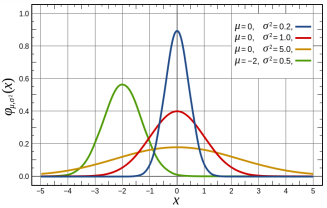
This can be used to determine the probability

## Normal Distribution

A.k.a. population normality

symmetric; mean = median = mode



[](https://en.wikipedia.org/wiki/File:Normal_Distribution_PDF.svg)

**Bell curve (a.k.a. Gaussian curve)**: normal curve, normal distribution; **Central Limit Theory** says the sampling distribution of sample means will be bell-shaped; s.d. = population s.d./√sample size

## Z-Tables

A.K.A. Standard Normal Cumulative Probability Table

**Z-function**: a standardized [cdf](#_Cumulative_Distribution_Function) that you use to predict data

* zc: critical value; this is also the area of the graph from 0 to c, where c is a point on the z-graph
* 
* It’s horizontal units are s.d.’s

If you’re given a probability (or percentile), you find the value on the z-table, where the probability represents α and choose the values at the location. If you cannot find the value on the z-table, find the two closest ones and find the weighted average.



**Standardized Score**: a.k.a. “z-score” 

*α*-level is the area of the graph of a normal distribution curve

Zα: for the standard normal distribution

When trying to find the α based on a z, make sure you round to the preferred sig figs

**Empirical rule**: you can identify that your data has normal distribution by using the rule that:

* 68% of data is within 1 s.d.’s from mean
* 95% of data is within 2 s.d.’s from mean
* 99.7% is within 3 s.d.’s from mean
* there are 3 [s.d.](#_Standard_Deviation)’s from the mean

# Chapter 5

p ← discrete

f ← continuous



Mean of sum of joint pdf (discrete):

Mean of sum of joint pmf (continuous):

**Covariance**: variance for multiple variables; 

Independent and Identically Distributed (IID):

* form a simple, random sample of size *n*
* Xi’s are independent r.v.’s
* Xi’s all have same probability distribution

**Multinomial distribution**: represented by the pmf,

**Marginal pdf** (continuous):

Conditional probability of joint pdf:

**Correlation coefficient**: 

# Chapter 6

 represents the parameter of interest

: variable with a hat means it is an estimate

error of estimation

* a function of the sample, i.e. rv

**Point estimate**: mean from multiple estimate(s), using the standard error, where θ represents parameter of interest (e.g. μ or σ), where you estimate .

**Bias of**:

**Unbiased**:

**Estimator**: the formula

* Should be unbiased (0 avg. error)
  + 
  + 
* Should have minimum variance (i.e. little spread)
* Summary for good estimators: Minimum Variance Unbiased Estimator (MVUE)
  + Unbiased is not always better than minimum variance

**Estimate**: value obtained from the formula after data has been inputted

What is point estimate for each θ:

* 
* Estimated chance of success

**True value**: mean of the population (instead of sample)

Trimmed means will result in **robust estimator**.

**Robust estimators** are less affected by outliers

Standard error of an estimator, is its standard deviation,

Estimated standard error: or

## Bootstrapping

**Bootstrapping**: fabricating multiple samples from one sample with replacement

* Only works for independent, equally-distributed, random samples
* Not useful if small data set, lots of outliers (remove outliers first), dependence structures (data based on changing time, etc.)
* n\* depends on computing capacity, type of problem, and complexity
* Computed bootstrap value is indicative of the accuracy of your sample. If it is higher than sample, sample is probably higher than actual; if lower than sample, sample is probably lower than actual

1. Compute , which is from x, sampled with replacement
2. Computefrom
3. Estimate standard error,

### e.g. Bootstrapping

Given a sample of:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Given | 0.5 | 1.5 | 2.5 | 3.5 | 4.5 |
| Fabricated  Range | 0-1 | 1-2 | 2-3 | 3-4 | 4-5 |

Now that you’ve established a range, you use a random number generator to generate 5 new points.

I randomly generated: 4.5290, 0.6349, 4.5669, 3.1618, 0.4877

Now tally how many are within each range:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Quantity | 2 | 0 | 0 | 1 | 2 |

Multiply this quantity by the initial value of the range, pretend that’s the new point, and add it up:

μboot = 2×0.5 + 0×1.5 + 0×2.5 + 1×3.5 + 2×4.5

= 13.5/5

= 2.7

Whereas, the sample average was actually 2.5

**Parametric bootstrap**: note: parameter refers to the population

## Point Estimation

**Point estimation**: 2 main methods: a method of inferring a value for a large population, θ, based on a small IID random sample, X, by calculating standard error

* [Method of moments](#_Method_of_Moments)
* [Maximum Likelihood estimation](#_Maximum_Likelihood_estimation)

Minimum Variance Unbiased Estimator (MVUE)

Estimators

|  |  |
| --- | --- |
| Population mean, μ | Sample mean |
| Population s.d., σ | Sample s.d., s |

### Method of Moments

α and β are unknown parameters that yield the estimator

kth sample moment of f(x) is

**Method of Moments Estimator (MME)**:

### -Maximum Likelihood Estimation

**Joint pdf**: pdf governing occurrences of A & B, not just one (i.e. pdf of occurrences of multiple potential events), like for regular pdf’s

Likelihood function: PMF = PDF = P(X;θ) = f(X;θ) [f ⬄ P]



More popular, easier

Results in normal distribution

**Maximum Likelihood Estimator**: (MLE) is the random sample, X, with the highest probability of being an appropriate estimator for the population, θ

How to find:

1. Find.
2. Find.
3. Equate to 0.
4. Solve for θ.

#### e.g.

e.g. for exponential distribution Note that at \*, the product becomes a summation of the xi values

# Chapter 7

Most important chapter for the midterm!

## Confidence interval

**Confidence level**: measures reliability of confidence interval; most popular confidence levels: 90, 95, and 99%; the percent of all samples that will give correct results, CL = P(CI)

**Confidence interval**: interval where certain where data is reliable

* Precision is width of confidence interval
* First determine confidence level
* use the [z tables](#_Z-Tables) to find
* In order for this to work:
  + Population distribution is [normal](#_Normal_Distribution)
  + [s.d.](#_Standard_Deviation) given
* Actual meandoes not necessarily have to be in the interval even if the estimated mean *is* in it

Sample mean ± 1.96 standard errors



The bound of the error is half the width, i.e. if estimate is within 1% of the true percentage, the 1% represents the bound of the error, so the width is 0.01×2.

Sample size:OR

CI:

A larger sample size gives a narrower confidence interval.

A smaller sample size gives a wider confidence interval.

**Standard error**: conversion of standard deviation (total population) to sample distribution (sample population) = 

**Statistical Inference**: a method of inferring certain statistical characteristics of a population based off a smaller sample, where characteristics could include things, such as sample mean or sample portion

**Sampling variability**: a concept in statistical inference, where even though you are inferring from a sample, each sample’s inferred population characteristics can vary from sample-to-sample; the smaller the standard error, the less the sampling variability; the larger the sample size, the smaller the standard error of the mean

**T-Table**: Z-Table, but for s, instead of σ, but you can still use z if sample size > 40; uses 2 parameters: degrees of freedom and probability level

# Chapter 8

The point of this to see if the error in the sample mean is low enough to make the sample valid/satisfactory.

**Statistical hypothesis**: assumption about a population characteristic; 2 types:

* [Null Hypothesis](#_Null_Hypothesis)
* [Alternative Hypothesis](#_Alternative_Hypothesis)
* Choose the hypothesis based on the [level of significance](#_Errors)
  + for lower level, choose Type I / Null
  + for higher level, choose Type II / Alternative

## Null Hypothesis

* H0: μ = μ0, where μ0 is the given value of μ
* proof by contradiction
* assume it is the thing you think it isn’t and prove that wrong
* think *equality*
* If you reject it, the evidence is **statistically significant**

## Alternative Hypothesis

* HA OR Ha OR H1
  + μ > μ0, z ≥ zα
  + μ < μ0, z ≤ zα
  + μ ≠ μ0, …
* specified *range*
* think >, <, or ≠
* if you only choose one inequality, it is called a **one-sided** [**hypothesis test**](#_Hypothesis_Test)

## Errors

* **Type I**: say something is right when it’s wrong
  + **Level of significance** (α): P(Type I error)
  + Proving [null hypothesis](#_Null_Hypothesis) true
  + Since null hypothesis is a value, P has one value
* **Type II**: say something is wrong when it’s right
  + P(Type II error) = β
  + Proving [alternative hypothesis](#_Alternative_Hypothesis) true
  + Since alternative hypothesis is a range, P is a range

### Case I

σ given (not *s*), normal distribution



### Case II

For large n (i.e. n > 40), s is close to σ



### Case III

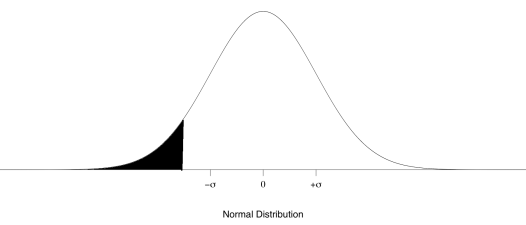
normal dist, *s* given

Test statistic value:

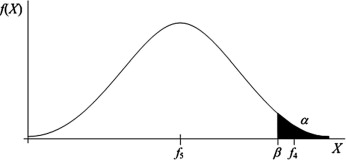
## Hypothesis Test

**One tail**:

Use lower-tail when the alternative hypothesis is: μ < Ha

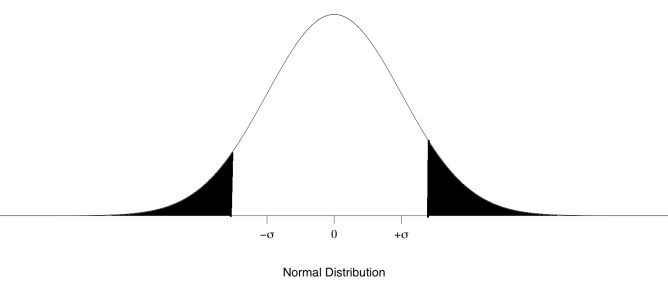


Use upper-tail when the alternative hypothesis is: μ > Ha

[](http://kld.metalbykevin.com/wp-content/uploads/2013/02/uppertailpdf.jpg)

**Two tails**: zα/2

Use this when alternative hypothesis is μ ≠ Ha



The rejection region is the dark part of these graphs. If in rejection region, reject the null hypothesis.

## P-Value

**P-Value**: observed level of significance

**Level of significance (α)**: a percentage or decimal that represents the cut-off value

* If P-value < *α* , reject the null hypothesis and accept the alternative hypothesis
* If p-value > *α* , don’t reject the null hypothesis and there is not enough information to determine whether or not to accept the alternative hypothesis
  + Just because one thing is wrong, doesn’t make the null hypothesis right. Thus, instead of accepting the null hypothesis, you refuse to reject the null hypothesis
* It is different for each region
  + 
  + Upper tailed: P = 
  + Lower tailed: P = 
  + Two-tailed: P = 

# Chapter 9 – Test Statistics

**Degrees of Freedom**: number of samples − 1

For normal populations with known variances, test statistic value:

Δ0 is usually 0

**Null hypothesis**: 

**Alternative Hypothesis**: 

3 Cases (conditions stay the same as before):

Note: n’s must be the same, use, instead of , use instead of μ; , and 

1. [Case I](#_Case_I)
2. [Case II](#_Case_II)
3. [Case III](#_Case_III)

Round down to the nearest integer

Pooled *t* happens when 

**Margin of Error**: 

***f* distribution**:

* pdf distribution is too difficult, so we will work with tables
* Assumptions:
  + 2 populations independent
  + Simple random samples
  + Normally distributed
  + Test statistic for test hypothesis, given two variances is: 
* Demonstrates the difference between the two variances
* Determines whether or not the rejection region is too high or not
* Inputs:
  + Significance level, α
* Null Hypothesis:
* Alternative Hypothesis:

# Chapter 12

**Correlation**: a measure of the relationship of your independent variable to your dependent variable

* Defines the relationship between the data points and the equation  
  only true within −1 to +1
* **−1**: implies perfectly negative correlation, i.e. inversely proportional
* **0**: no linear correlation
* **1**: proportional relationship
* **Weak correlation**: −0.5 ≤ r ≤ 0.5
* **Strong correlation**: r ≤ −0.8 OR r ≥ 0.8
* **Moderate**: −0.8 < r < −0.5 OR 0.5 < r < 0.8

**Causation**: Correlation does not mean causation.

Determine a line with the least variance

**Deterministic Relationship**:

one variable can be found in terms of the other variable

**Linear**: a first order polynomial example of a deterministic relationship (i.e. y = mx + b)

**Statistical**: non-deterministic; relies on probability

**Regression Analysis**: looks at correlations between two things by removing other variables

**Model equation**:

ε: measure of variation; error in data

**Principle of least squares**: gives minimum error



**Point Prediction**: plugging in values of *x* into the regression equation

**Residual**: error; vertical deviation from estimated line (y − y0)

**Extrapolation**: usually doesn’t work, though

library (MASS)

summary() gives 5-number summary

**Sum of Squares for Errors (SSE)**: 

# Chapter 10

**ANOVA**:

**Factor**:

**levels of the factor**:

The number of populations being compared is *I*.

represents the random variable for the *j*th experiment for the *i*th population

## Hypothesis



