

Random Variables🡪 Quant/ Qualitative

Quant->Discrete (Can be integers, can’t be fractions) ( number you could use to count things)/Continuous (Any Real numbers/can be fractions)

Qualitative-> Nominal (Order doesn’t matter)/Ordinal

Data collection: Randomization, Sampling, Observational Study, Randomized experiment

To learn about characteristics of a pop: Entire sample/ subset of sample

To take representative subset of sample-> Use random mechanism-> Randomized Experiment (we create differences in the explanatory variable and then examine the results- creating treatments and assigning subjects randomly)/ Observational Study (we observe differences in the explanatory variables)

Types of Sampling:

**Simple random sampling**

Sample of size *n* from a population of size *N*

Equal probability of selection

**Stratified random sampling**

Select a random sample from each strata

e.g. proportional allocation

Reduces error

**Cluster random sampling**

Select a random sample from each cluster

Reduces cost, but increases error

**Systematic random sampling**

Simple random sampling in multiple

Simple design and administration

Experimental Design Features:

Blocking (stratification) Improves Power and reduces error

Non-Graphical Exploratory Data Analysis:

Measure of central tendency- Mean, Median, Mode

measure of spread- Variance, Standard Deviation

shape of distribution

existence of outliers

|  |  |
| --- | --- |
| ow variability | igh variability |

Standard deviation= sqrt (variance)

Outliers: Scatterplots/Box[plots🡪While fitting models such as linear regression, we also use the residuals plots and analysis to address potential influential points, e.g. leverage plots, Cook's distance, etc...

Probability and Distribution

* Review of Discrete Probability
  + Conditional Probability
  + Independence
* Random Variables
  + Continuous random variables
  + Chi-squared distribution
  + Discrete random variables
  + Joint Distributions
* Moments
  + Mean and Variance of a random variable
  + Mean and Variance of a sum of random variables

The basic problem we study in **probability**:

Given a data generating process, what are the properties of the outcomes?

The basic problem of **statistical inference**:

Given the outcomes, what can we say about the process that generated the data?

Space-> Events🡪 Outcomes

Discrete Probability:

1. Mutually Exclusive
2. Conditional Probability
3. Independence

*P*(*A* or *B*) = *P*(*A*) + *P*(*B*) − *P*(*A* and *B*)

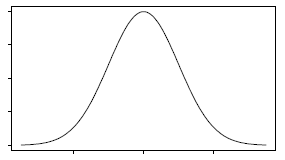
Conditional Prob : *P*(*A*|*B*) = *P*(*A* and *B*)/*P*(*B*)

Independence: *P*(*A* and *B*) = *P*(*A*) *P*(*B*)--🡪 *P*(*A*|*B*) = *P*(*A*) and *P*(*B*|*A*) = *P*(*B*). Independence means that knowing *A* has occurred provides no information about whether or not *B* has occurred and vice-versa

***Continuous random variables*** are described by probability density functions (PDF).

Noraml Distro/ Chi-square distro

For example, a normally distributed random variable has a bell-shaped density function like this:



The probability that *X* falls between any two particular numbers, say *a* and *b*, is given by the area under the density curve*f*(*x*) between *a* and *b*, (Integrable)

Chi-squared Distribution:

The "degrees-of-freedom" (*df*), completly specify a chi-squared distribution.

A χ2 random variable takes values between 0 and ∞

The mean of a chi-squared distribution equals to its df

The variance of a chi-squared distribution equals to 2df, and the standard deviation is √2df, where √ is a symbol for square root.

The shape of the distribution is skewed to the right.

As df increase, the mean gets larger and the distribution spreads more.

As df increase, the distribution becomes more bell-shaped, like a normal, e.g., df → ∞, χ2df → Normal.

***Discrete random variables*** are described by probability mass functions (PMF)

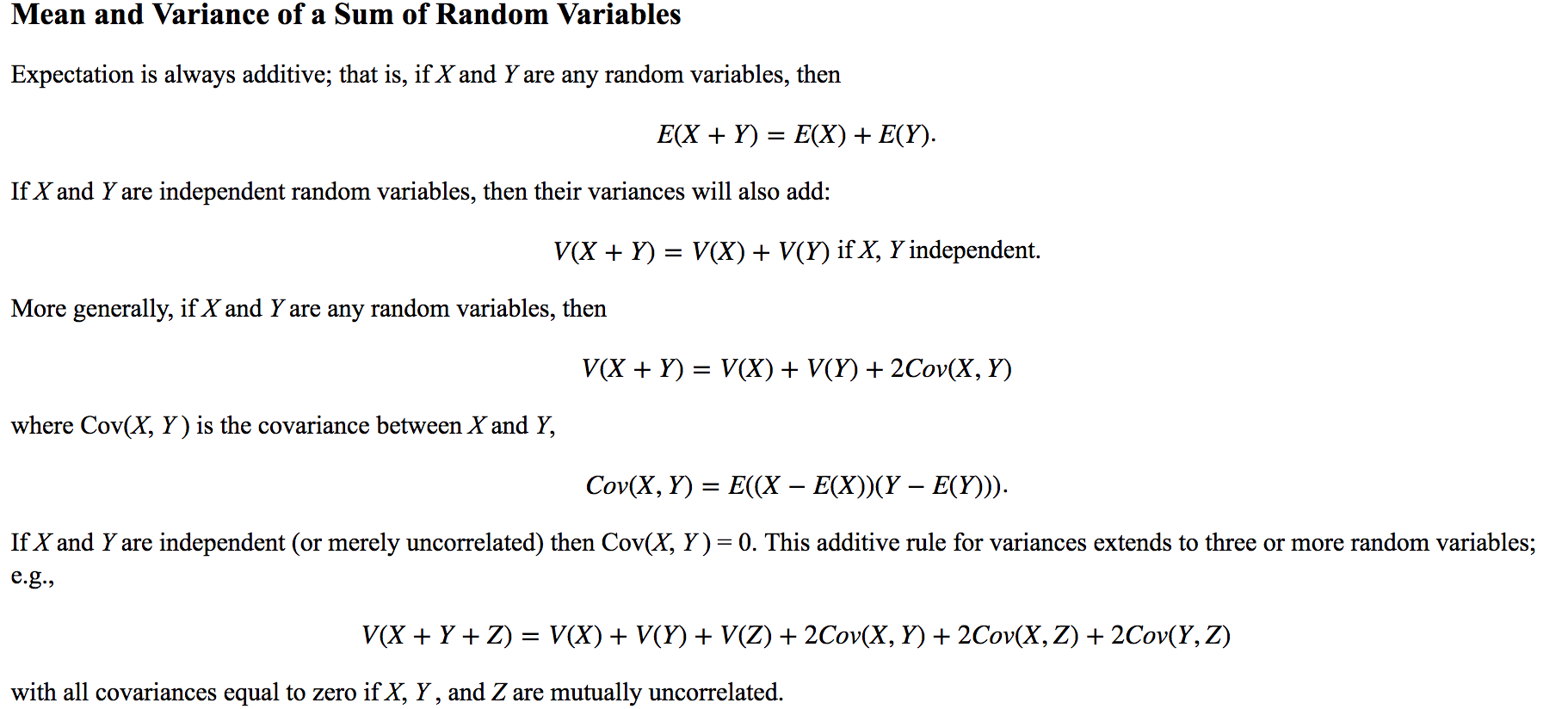
Joint Ditributions:

In particular, suppose that the random variables *X*1,*X*2, . . . , *Xn* are independent and identically distributed (*iid*). Then *X*1 = *x*1, *X*2 = *x*2, . . . , *Xn* = *xn* are independent events, and the joint distribution is

Mean: E(x)= sum(x\*f(x))

Variance V(X)=E((X-E(X))2)

*V* (*X*) is the average squared distance between *X* and its mean. Variance is a measure of dispersion, telling us how “spread out” a distribution is.



**statistical inference** aims at learning characteristics of the population from a sample; the population characteristics are *parameters* and sample characteristics are *statistics*.

A **statistical model** is a representation of a complex phenomena that generated the data.

* It has mathematical formulations that describe relationships between random variables and parameters.
* It makes assumptions about the random variables, and sometimes parameters.
* A general form: data = model + residuals
* Model should explain most of the variation in the data
* Residuals are a representation of a lack-of-fit, that is of the portion of the data unexplained by the model.

Point Estimation, Interval Estimation and Hypothesis Testing-🡪 learning about the population parameter from the sample statistic.

**Estimation** represents ways or a process of learning and determining the population parameter based on the model fitted to the data.

Inference-Understanding characteristics of population from sample

Estimation-Determining population parameter based on model fitted to data

**Point estimation** = a single value that estimates the parameter. Point estimates are single values calculated from the sample

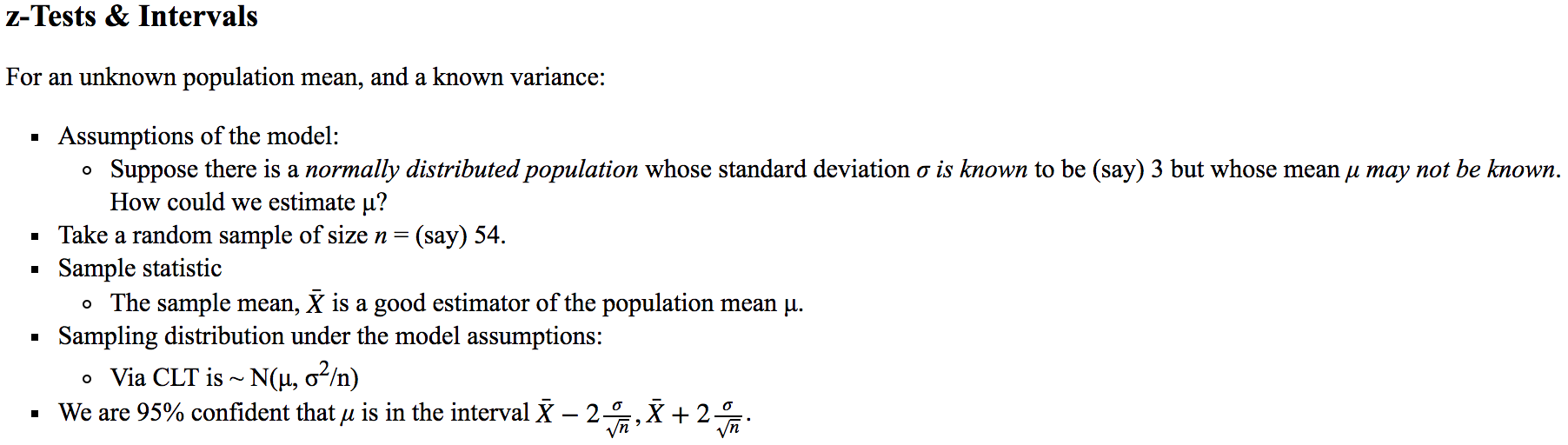
**Confidence Intervals** = gives a range of values for the parameter Interval estimates are intervals within which the parameter is expected to fall, with a certain degree of confidence.

* estimate ± critical value × std.dev of the estimate
* estimate ± margin of error

CI’s differ based on parameter of interest, design of sample and Confidence levels as well!!

A 95% CI for a population parameter DOES NOT mean that the interval has a probability of 0.95 that the true value of the parameter falls in the interval. The probability is associated with the process that generated the interval.

**Hypothesis tests** = tests for a specific value(s) of the parameter.



**Basic approach to hypothesis testing**

1. **State a model** describing the relationship between the explanatory variables and the outcome variable(s) in the population and the nature of the variability. **State all of your assumptions**.
2. **Specify the null and alternative hypotheses** in terms of the parameters of the model.
3. Invent a **test statistic** that will tend to be different under the null and alternative hypotheses.
4. Using the assumptions of step 1, **find the theoretical sampling distribution** of the statistic under the null hypothesis of step 2. Ideally the form of the sampling distribution should be one of the “standard distributions”(e.g. normal, *t*, binomial..)
5. **Calculate a *p*-value**, as the area under the sampling distribution more extreme than your statistic. Depends on the form of the alternative hypothesis.
6. **Choose your acceptable type 1 error rate** (alpha) and **apply the decision rule**: reject the null hypothesis if the p-value is less than alpha, otherwise do not reject.

In two-sided we only care there is a difference, but not the direction of it. In one-sided we care about a particular direction of the relationship. We want to know if the value is strictly larger or smaller.

The *p*-value represents **how likely** we would be to observe such an extreme sample if the null hypothesis were true. **a probability** computed assuming the null hypothesis is true, that the test statistic would take a value as extreme or more extreme than that actually observed.

PN PY

AN TN FP

AY FN TP

**true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.

**true negatives (TN):** We predicted no, and they don't have the disease.

**false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

**false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

Type -2 error more dangerous

**True Positive Rate/Senstivity:** When it's actually yes, how often does it predict yes? (TP/AY)

**False Positive Rate:** When it's actually no, how often does it predict yes? (FP/AN)

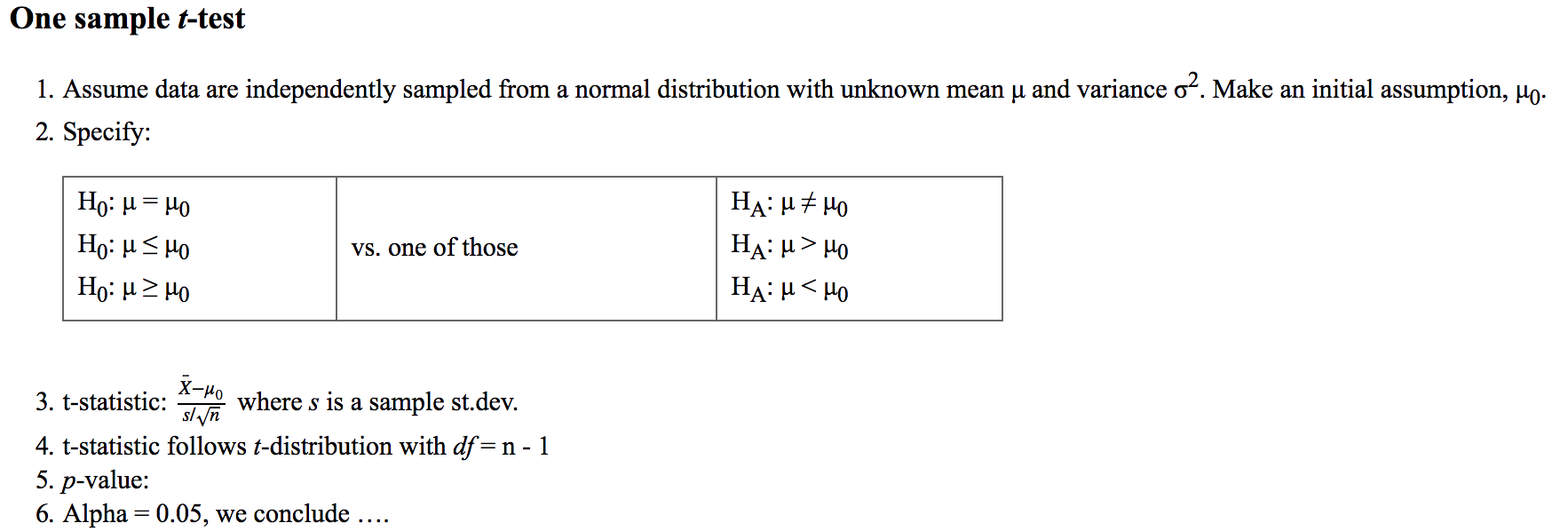
**Specificity/ True Negative Rate:** When it's actually no, how often does it predict no?TN/AN

**Precision:** When it predicts yes, how often is it correct? (TP/PY)

Power of a test:

The power of a statistical test is its probability of rejecting the null hypothesis if the null hypothesis is false. That is, power is the ability to correctly reject H0 and detect a significant effect.

If the sample is very large, we can treat σ as known by assuming that σ = *s*. According to the law of large numbers, this is not too bad a thing to do. But if the sample is small, the fact that we have to estimate both the standard deviation and the mean adds extra uncertainty to our inference. In practice this means that we need a larger multiplier for the standard error.



Inference about unknown population parameter🡪 Sample Statistics🡪 Sampling distribution🡪Model Assumptions/ Standard Error

**Central Limit theorem**

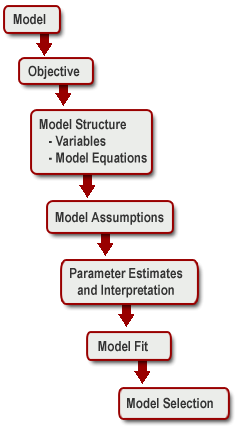
If numerous samples of size n are taken, the frequency curve of the sample means from those various samples is approximately bell shaped with mean μ and standard deviation, i.e. standard error (Normally distributed)

Holds if:

* *X* is normally distributed
* *X* is NOT normal, but *n* is large (e.g. *n* >30) and μ finite.
* For continuous variables

**Objective**

State what the objective is for this model. For instance, "Estimate the probability that a characteristic is present given the value of the explanatory values are ... "

**Model Structure**

State the **important variables in the model**. What is the response variable Y? What is included in the set of explanatory variables?

State the **Equation for the Model**.

**Model Assumptions**

State the assumptions for the model that you are using. Are the data independently distributed? Do linear relationships exist between the dependent and independent variables? Is the variance homogeneous? Are errors independent and normally distributed?

**Parameter Estimation and Interpretation**

What are the odds that the objective characteristic is present? What evidence will you use to establish the estimate?

**Model Fit**

How is goodness of fit determined? Pearson chi-square statistic? Deviance? Likelihood ratio test? What does the analysis of the residuals show?

**Model Selection**

Choosing a single ”best” model among the presence of more than one reasonable model involves some subjective judgment. We seek a parsimonious model that is as simple as possible and adequately explains the phenomena of interest.

**Analysis of Discrete Data**

* Understand the basic principles of likelihood-based inference and how to apply it to tests and intervals regarding population proportions