STAT 340 Topic Review

This is an attempt to summarize the main topics covered in the course.

**Chapter 1**: Probability and Random Variables The basic rules of probability: outcome spaces, events the concept of a random variable Families of discrete random variables: Bernoulli, binomial, geometric, Poisson and uniform Families of continuous random variables: Gaussian (normal), exponential and uniform The concept of expected value PMF, PDF and CDF computing probabilities some applications of random variables

Pr[E}∈[0,1] E1∩E2=0 -> Pr[E1UE2]= Pr[E1]+Pr[E2] **mutually disjoint**

**Independent** Pr[E1∩E2]= Pr[E1]Pr[E2] learning information about

one of them doesn’t tell you anything about the other

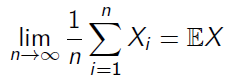
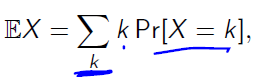
**Discrete**

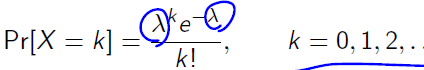
**Bernouli x~Bernoulli(p)** p probability success

**Binomial** x~Binomial(n,p) p probability success n # trials

**(rbinom(n, size, prob).** n-number generated r.v , prob=p , size=# trials p-probability of success

**Geometric rgeom(1,prob=p) Pr[X=k]=(1-p)kp k=3 FFFS k=0 -> Pr[k]=p**

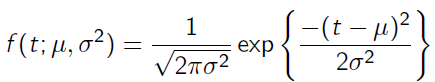
** LLN  **

**Poisson** *λ* **=#events/time  rpois(1, lamda=10.5)**

**Discrete uniform sample(x=6,size=20,replace=TRUE)**

**X~Uniform(a,b), Pr(X=k)=1/(b-a+1) k=a,aa=1,…,b E(X)=(a+b)/2**

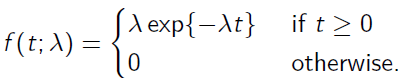
**Continues density function with d, CDF with p, random sample with r**

**Standard normal μ =0 σ=1 x<-seq(-4,4,0.1) dnorm(x,0,1) probability density function  empirical 1 sd, 0. 68, 2sd 0.95, 3sd 0.99**

**rnorm(1,mean=μ, sd=μ)**

**CDF FX(t)=Pr[X<=t] pnorm(t) standard CDF**

**Uniform f(t)=c runif(n=1, min=0,max=1}**

**Exponential  dexp(x,rate=1**

**Examples binomial asset pricing model, election model**

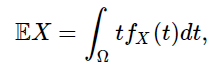
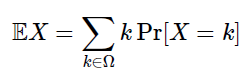
Type1 = False Alarm, Type2 = iss

Chapter 2: Independence, Conditional Probability and Bayes' Rule The concept of independent events Independent random variables Definition of variance Expectation of a linear combination of r.v.s Variance of a linear combination of r.v.s Covariance and correlation Relationship between correlation and independence When the independence assumption is reasonable Conditional probability & the general multiplication rule Bayes’ rule

Independence density functions

Discrete Pr[*X* = *k*, *Y* = ℓ] = Pr[*X* = *k*] Pr[*Y* = ℓ].

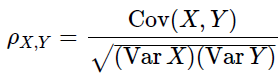
Continuous fX,Y(s, t) = fX(s) fY (t)

Expectation   **properties E(*aX* + *b*) = *a*E*X* + *b*. E(*X* + *Y* ) = E*X* + E*Y***.

Variance **Var*X* = E(*X* – E*X)2*= EX2− E2X Var(*aX* + *b*) = *a*2 Var(*X*).**

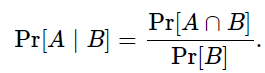
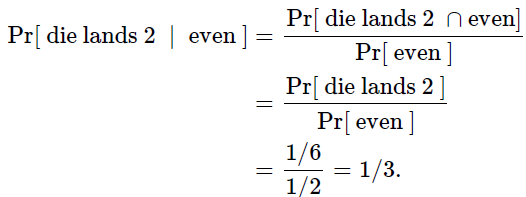
Var(*X* + *Y* ) = Var *X* + **2E(*X* − E*X*)(*Y* − E*Y* )** + Var *Y* . Cov(*X*, *Y* ) = E(*X* − E*X*)(*Y* − E*Y* ).

**Independence Cov(X,Y)=0**

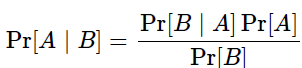
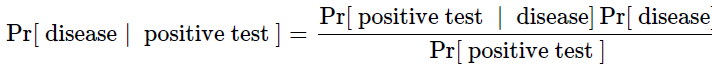
**Correlation  independence => correlation 0 cov(X,Y)=0**

**Uncorrelated not mean independent e.g Y=X2  cor is almost 0, not independent**

Independent assumption important e.g taking measure of pulse before and after training are not 2 independent r.v

Conditional probability  Pr[B]!=0 

Die lands 2 if it lands even

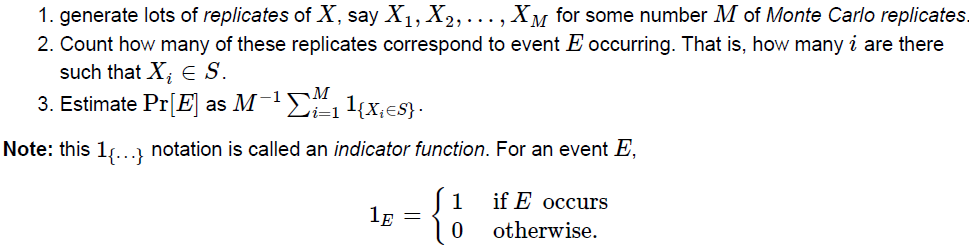
Bayes rule  

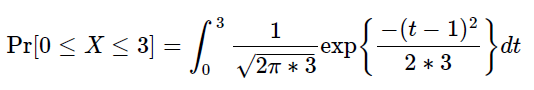


Pr[ positive test]= Pr[ positive test ∣ disease] Pr[ disease+ Pr[ positive test ∣ no disease] Pr[ disease]

Chapter 3: Monte Carlo Simulations The basic steps of a Monte Carlo simulation Estimating the expected value of a random variable using MC Estimating a probability using MC Estimating a definite integral using MC Why Fx(X)∼Uniform(0,1) for any r.v. X Simulating an aribitrary random variable using MC Psuedorandom number generation & the use of set.seed() in R

Pr[E]



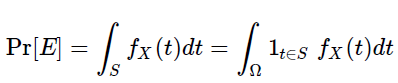
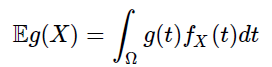


**pnorm(3, mean=1, sd=sqrt(3)) - pnorm(0, mean=1, sd=sqrt(3))**

0.594042

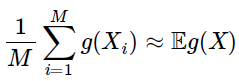
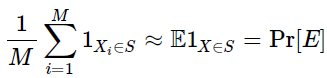
|  |  |
| --- | --- |
| ***X* ∼ N(*μ* = 1, *σ*2= 3)**  event\_E\_happened <- **function**( x ) {  **if**( 0 <= x & x <= 3 ) {  **return**( TRUE ) *# The event happened*  } **else**  { **return**( FALSE ) *# The event DIDN'T happen*  }  } | NMC <- 1000; *# 1000 seems like "a lot"*  results <- rep( 0, NMC ); *# We're going to record outcomes here*  **for**( i **in** 1:NMC) { *# Generate a draw from the normal, and then...*  X <- rnorm( 1, mean=1, sd=sqrt(3) ); *# ...record whether or not our event of interest happened.*  results[i] <- event\_E\_happened(X);} *# Now, compute what fraction of our trials were "successes" (i.e., E happened)*  sum( results )/NMC |

Examples Birth problem, Buffon’s needle, MC to estimate pi

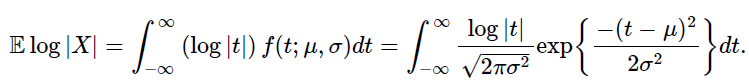
 

Expectation

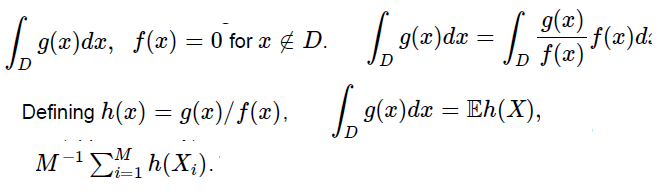
 

**Integration by darts**

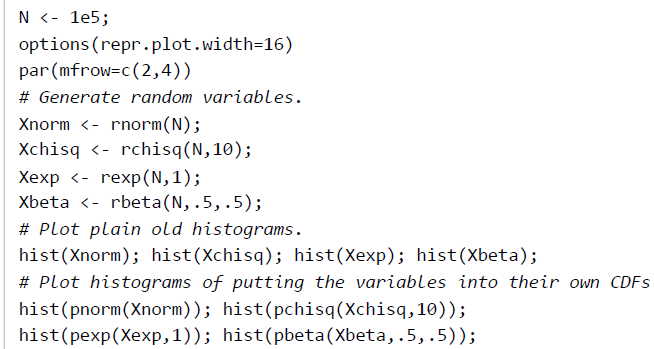
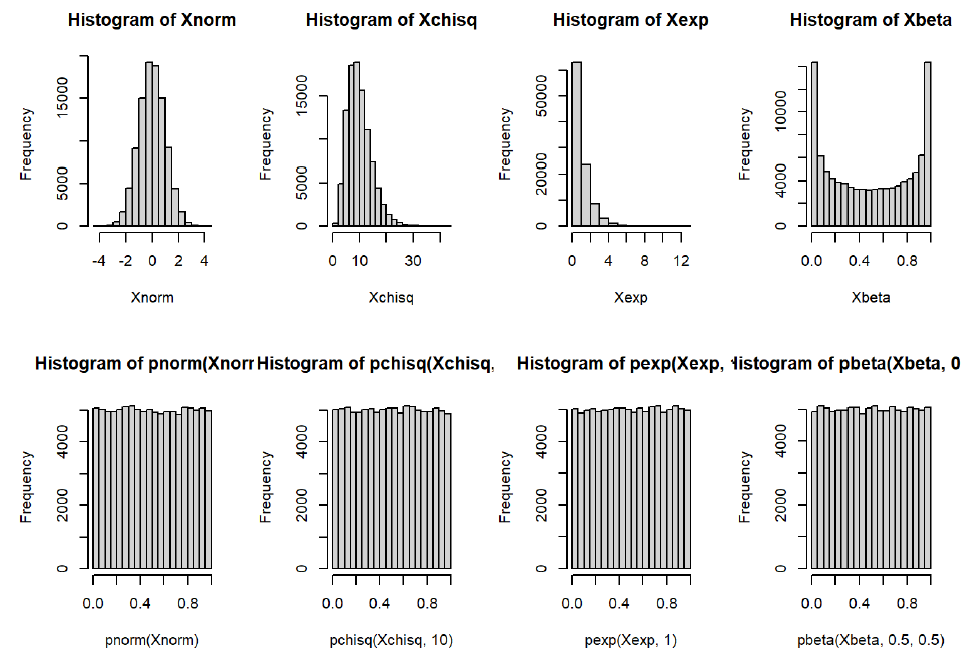


Draw lots X1,…,XM from normal distribution. With mean and sigma



**Important**

**For *any* random variable X with CDF F ,F(X) is distributed as a uniformrandom variable on interval [0.1].**

|  |  |
| --- | --- |
| *# reset Rmarkdown print width for readability*  options(repr.plot.width=7)  *# generate a bunch of Unif(0,1) RVs*  unifs <- runif(1e5)  *# Apply the built-in inverse CDF function using qnorm*  *# that is, this is computing F\_X^{-1}(U).*  **x.normal.mc = qnorm(unifs,mean=1,sd=1)**  *# plot results*  hist(x.normal.mc) |  |
| 2 distributions are similar if Q-Q plot is line  *# plot our sample quantiles against theoretical quantiles for comparison*  plot(sort(**x.normal.mc**),qnorm(ppoints(1e5,1),mean=1,sd=1)) |  |

Chapter 4: Testing (Part 1)

Formulating a decision into a null and alternative hypothesis

Randomization tests (model-based)

Permutation tests (two-sample)

Simulating a distribution of test statistics

Calculation of one or two-tailed p

-values from simulated distribution of test statistics

Interpretation of a p -value

**Lady testing tea**, generate string of cups and count hits with of lady’sstring

Repeate NMC 2000 times

Count ncorect/NMC

H0 random

H1 not random gueses

1. We start with a null hypothesis, and we determine what kinds of observations we would expect to see if that null hypothesis were true.

2. We collect our data.

3. We compare our observed data to what we would expect to be true under the null.

4. If our observed data is “unlikely” or “unusual” under the null, then this is evidence against the null hypothesis.

Vaccine efficacy

In the case of the vaccine trial, the null hypothesis was that subjects in the treatment and control groups were equally likely to get infected with COVID.

T- test statistic, actually observed data d within test statistic T(d), and T(D0) random copy of data generated under H0.

Pr[T(D0)>=T(d):H0]

Drug trials

Control, treatment data H0: treatment has no effect

**Parametric testing**

H0 : = μcontrol = μtreatment

t-test **assumptions**

data came from a normal distribution, but with an unknown variance

|  |  |
| --- | --- |
| **Parametric testing model with parameters**  test statistic  t-test Not known S.D on population n-1 degrees of freedom  asumptions: data from specific dristibution  d.f. *n*1 + *n*2 − 2 = 78,  1 **- pt**( t, df=**length**(control)**+length**(treatment)**-**2) | Welch’s t-test unequal variances and possibly unequal  sample sizes |
| Not parametric testing  Permutation test H0: control and treated sequence  e.g difference. Calculate mean diference  1. Reshuffles the control and treatment data randomly  2. Assigns the shuffled data to control and treatment groups  of the same sizes as the original ones  3. Computes our test statistic on the reshuffled data (in this case, we’re taking the test statistic to be the difference of means, but we could make a different choice if we really wanted)  sum(test\_statistics >= Tobsd)/NMC;  0.0469  this is evidence against the null hypothesis that the control and treatment groups had the same distribution. |  |

Chapter 5: Testing (Part 2)

Type 1 and Type 2 Errors

Balance between error types

Significance Level α

Rejection rules (critical value)

One-sided vs two-sided tests

Rejection region

Power of a test statistic

Power function (curve)

Comparing test statistics

Choosing α

Thus, there is some probability that, even if the null hypothesis is true, we observe very unlikely data and, as a result, reject the null hypothesis incorrectly.

Even if the null hypothesis is false, the data that we observe may not be “weird” enough to constitute sufficient evidence against the null hypothesis, and we may conclude, incorrectly, that the null hypothesis is true.

|  |  |
| --- | --- |
|  | * A Type I error corresponds to rejecting the null hypothesis when it is in   fact true (“**false alarm**”).   * A Type II error corresponds to accepting the null hypothesis when it   is not true **(“miss**”).  H0: p=1/2 fair coin |
| A Type I error would correspond to the case where Bristol is guessing completely randomly (i.e., cannot tell the milk-first cups from the milk-second cups), but we conclude, incorrectly, that she can tell the difference.  A Type II error would correspond to the case where Bristol really can tell the difference, but we incorrectly conclude that she is guessing at random.  **Parametric model**  **Coin test statistic T=# of H n=200 independent Bernoulli r.v.**  **Binomial r.v size parametae=200 succes probability p**  **Eg** qbinom(0.95, size=200, prob=0.5) 112  pbinom( 112, size=200, prob=0.5) 0.961 | **Type I error is called the *level* or *size* of the test, and is denoted α. Threshold for rejecting H0 when it is true. α =0.05**  α could be lower 0.01 or even 0.001 lower risk of rejecting H0      **Find t solving FT(t)=1−α**,  **reject H0 if T≥t.**  **We have CDF invers will give t value qnorm, qbin F(t)=q q∈[0,1]** |
| **α =0.05 incorrectly reject H0 5% times chance to reject true** | **Two sided** |
| x≤tα,1  or x<tα,1x | choose tα,1 tα,2 so that Pr[T<tα,1]=α/2, Pr[T>tα,2]=α/2    choose tα,1 tα,2 so that Pr[T<tα,1]=α/2, Pr[T>tα,2]=α/2 |
|  | One side test is more powerful  0ne sided t 1 < two sided t2  p=0.05 p=0.025  smaller error I -> greater error II  -> smaller power |
| One sided vs two sided test |  |
| Greater # of Monte Carlo simulation, **more precise estimation**  Accuracy, stability convergence  Higher repetition more precise results | 0nly for p=0.05 H0 is true, for greater p H0 is false  one-sided test is more likely to (correctly) reject compared to two-sided test when p>1/2.  Error type I  Error type II (yellow with stripes)  Power (yellow) |
| Other statistics for flipping coins:  Max lengths  longestRun = function(x,target){ max(0,with(rle(x), lengths[values==target])) } | If error II more important than Error I, take greater α, ? |
|  |  |
|  |  |
| ODS (odds) running ratio  Odds Ratio: The odds ratio is a statistic that quantifies the strength of the association between two events. It compares the odds of an event occurring in one group to the odds of it occurring in another group.  Odds Ratio (OR)} = Odds of event in group 1/Odds of event in group 2}}  Odds = Probability of event/Probability of no event | OR > 1: Greater odds of the event occurring in the first group.  OR = 1: No difference in odds between the groups.  OR < 1: Lower odds of the event occurring in the first group |
| ODS ratio eg. basket YNNNNYYN -> YN NN NN NN NY YY YN 4 types  YN NN YY NY  2 3 1 1 |  |

Type1 = False Alarm (False Positive), Type2 = miss (True Negative)

Pr(Type1) = called level or size = a

Power = Pr(Reject H0 | H0 is False)

1 sided test is more powerful and is more likely to correctly detect

And reject H0 hypothesis.

With higher number of MCarlo simulations we get better estimate

Of the power, but it does not change the power.

Geom -