PM520 - Week 3

Classical statistical inference

- H_o: θ=1 versus H_a: θ>1
- Classical approach
- Calculate a test statistic
  - P-value = P(Test stat. value (or "more extreme") | H_o)
  - P-value is NOT P(Null hypothesis is true)
  - Can construct confidence interval for θ [a, b] : what does it mean?
- But scientist wants to know:
  - P(θ=1 | Data)
  - P(H_o is true) = ?
- Problem
  - θ "not random"

  - *This is what you are doing in assignment 1, but you are asked to estimate the p-value via simulation.*

Bayes Theorem

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$= \frac{P(B \mid A)P(A)}{P(B)}$$

$$= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^C)P(A^C)}$$

Thomas Bayes was an English statistician, philosopher and Presbyterian minister who is known for having formulated a specific case of the theorem that bears his name: Bayes' theorem. Wikipedia

Born: 1702, London, United Kingdom

Died: April 7, 1761, Royal Tunbridge Wells, United Kingdom

(wikipedia)

Example Application of Bayes' theorem

- Population has 10% liars
- Lie Detector gets it "right" 90% of the time.
- Let A = {**Actual** Liar},
- Let L = {Lie Detector reports you are **Liar**}
- Lie detector reports suspect is a liar. What is probability that suspect actually is a liar?

$$P(A \mid L) = \frac{P(L \mid A)P(A)}{P(L \mid A)P(A) + P(L \mid A^C)P(A^C)}$$

$$= \frac{(.90)(.10)}{(.90)(.10) + (.10)(.90)} = \frac{1}{2}!!!!!$$

Bayesian statistics

- Paradigm shift in statistical philosophy
  - θ assumed to be a realization of a random variable
  - Allows us to assign a probability distribution for θ based on *prior* information
  - 95% “confidence” interval [$1.34 < \theta < 2.97$] means what we “want” it to mean: e.g., $P(1.34 < \theta < 2.97) = 95\%$

Bayesian modeling/statistics

- Three General Steps for Bayesian Modeling
  - I. Specify a probability model for unknown parameter(s), that includes your prior knowledge about the parameters (if available).
  - II. Update knowledge about the unknown parameters by conditioning this probability model on the observed data, by using Bayes' Theorem.
  - III. Evaluate the fit of the model to the data *and the sensitivity of the conclusions to the assumptions (i.e. the prior)*.

Bayesian statistics

- Let θ represent parameter(s)
- Let X represent data

$$f(\theta \mid X) = f(X \mid \theta) f(\theta) / f(X)$$

- Left-hand side is a function of θ
- Denominator on right-hand side does not depend on θ

$$f(\theta \mid X) \propto f(X \mid \theta) f(\theta)$$

- Posterior distribution $\propto$ Likelihood x Prior distribution
- Posterior dist'n = Constant x Likelihood x Prior dist'n
- Goal: Explore the posterior distribution of θ

Prior distributions

- The prior distribution reflects what you knew about θ, the model parameter(s), before you did the experiment.

- Where do priors come from?
  - Previous studies, published work.
  - Researcher intuition.
  - Substantive Experts.
  - Convenience (conjugacy, vagueness).

Simple example

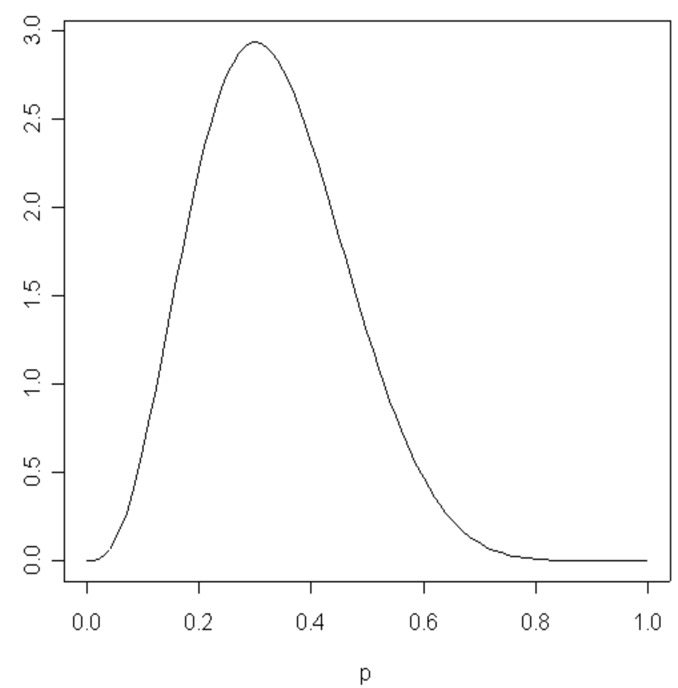
- 'Biased coin' estimation: *P(*Heads*)* = *p* = *?*
- $X_1, \ldots, X_n$ i.i.d. ordered Bernoulli(*p*) trials
- Let X be the sequence of 'heads' and 'tails' in the *n* trials
- Likelihood is $f(X \mid p) = p^X (1-p)^{n-X}$
- For prior distribution, could use *uninformative* prior
  - Uniform distribution on (0,1): *f(p)* = 1
- So posterior distribution is
- *f(p|X)* ∝ $p^X (1-p)^{n-X}$

Simple example (continued)

- Posterior density of the form $f(p)=Cp^x(1-p)^{n-x}$
- In fact, posterior distn is Beta distribution: Parameters x+1 and n-x+1
- Note that the Beta(1,1) is a Uniform(0,1) distribution.

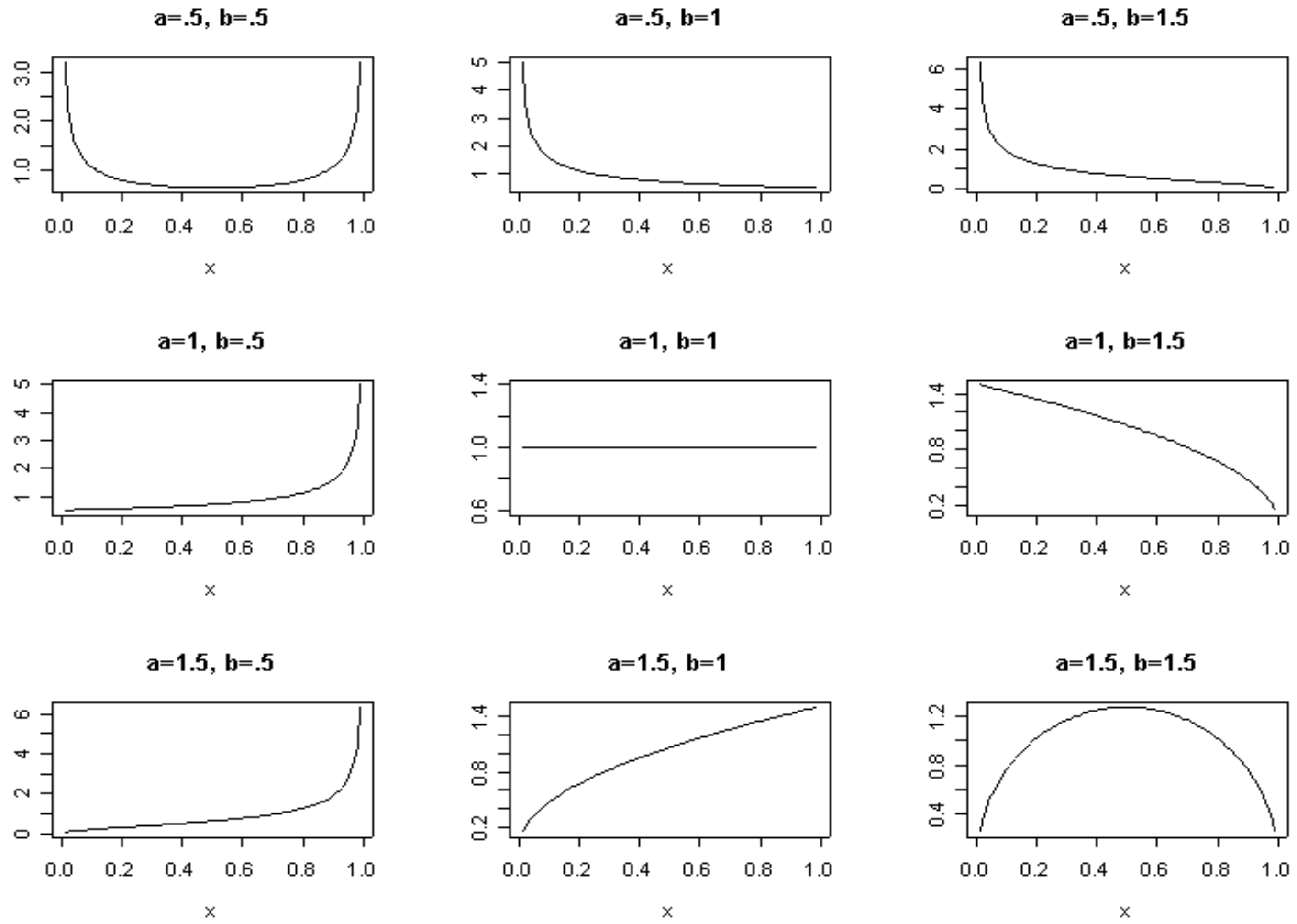
- *Example:* Data: 0, 0, 1, 0, 0, 0, 0, 1, 0, 1
- *n=10*
- *Use uniform [Beta(1,1,)] prior*

- Posterior dist'n is Beta(3+1,7+1) = Beta(4,8)
  - Mean: 0.33
  - Mode: 0.30
  - Median: 0.3238
  - 95% *credible* interval for *p is [0.11, 0.61]*
    - P(0.11 < *p* < 0.61 | *X*) = 0.95



Choice of prior

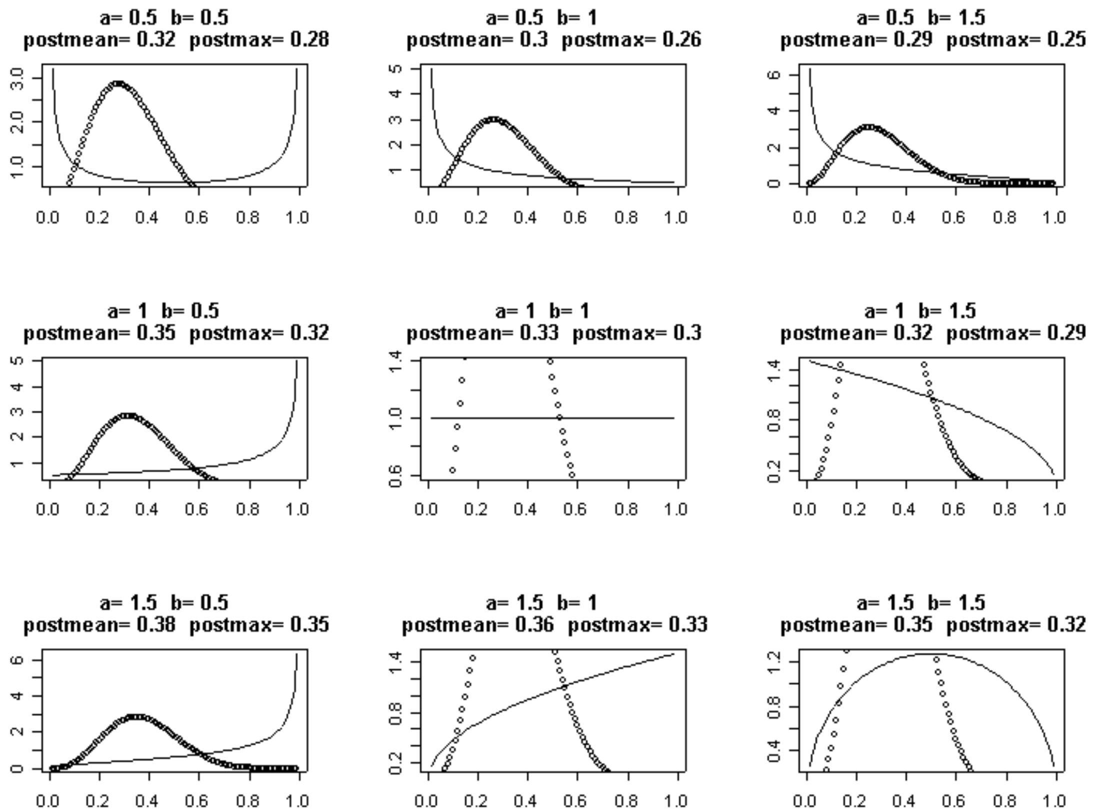
- Could use some other Beta distribution:



  - http://www.people.carleton.edu/~rdobrow/courses/275w05

Choice of prior

- Would get these posterior dist'ns:



[Priors that result in posteriors of the same form (i.e. same distributional family) are called *conjugate priors*.]

- http://www.people.carleton.edu/~rdobrow/courses/275w05

Differences Between Bayesians and Frequentists

- Frequentist:
  - The parameters of interest are fixed and unchanging under all realistic circumstances.
  - No information prior to the model specification.

- Bayesian:
  - View the world probabilistically, rather than as a set of fixed phenomena that are either known or unknown.
  - Prior information abounds and it is important and helpful to use it. *But results are now sensitive to priors.*

http://www.stat.ufl.edu/archived/casella/Talks/BayesRefresher.pdf

Generation of random variables, probability distributions, etc. (chapters 14/18)

- R, random number function:

```
runif(n,x,y)
```

- Generates n continuous random numbers distributed uniformly between x and y. e.g.,

```
> runif(10,0,5)
   [1] 0.7615195 2.6318839 1.4084045 1.3250196 3.6335061 4.8151777 2.7537802 2.7817826 1.3949205 0.7280294
```

To turn that into integers use the ceiling() function (ceiling(x) returns the smallest integer bigger than x):

```
> WhichBall<-ceiling(runif(10,0,NumberOfBalls))
> WhichBall
 [1]  7  7  3 10  3  3  2  2 10  2

> sample(x, size, replace = FALSE, prob = NULL)
> sample(1:10,10,TRUE)
 [1]  9  4  7  2  2  6 10  8  6  1
```

Sampling from Distributions - Discrete random variables

- Random variable [rv], X, takes a value x ∊ Ω (the state-space).
- e.g. X ~ coin toss: Ω = {"Head", "Tail"}
- e.g. Y ~ roll 6-sided die: Ω = {1,2,3,4,5,6}.
- In R, e.g., sample(c("H","T"),5,replace=TRUE)

- P(X=x) is a map from Ω to the unit interval [0,1], that gives the probability of the outcome x. [N.B., rv= capital letter; outcome= lower-case letter.]
- e.g. P(X="head") = 1/2
- e.g. P(Y=5) = 1/6.

Sampling from Distributions - Discrete random variables

- $F(x) = P(X \leq x)$ is the *Cumulative Distribution Function*.
- $F(x) = \sum_{y \leq x} P(X=y)$, for discrete random variables
- It follows that $P(a<X \leq b)=F(b)-F(a)$.

Sampling from Distributions - Continuous random variables

- $F(y)=P(Y\leq y)=\int_{-\infty}^{y} f(u)du$ [the Cumulative Distribution Function [or CDF]].
- $f(y)=dF(y)/dy$, is the *probability density function.*
  - e.g. exponential distribution

    $f(x) = \lambda\exp(-\lambda x), \quad x\geq 0$

    $F(x) = P(X<x) = 1-\exp(-\lambda x), \quad x\geq 0 \quad [0\leq F(x)\leq 1]$

Empirical Density Estimates of Probabilities for Discrete rvs

- Simulate N, independent and identically distribution random variables, $X_1, X_2, \ldots, X_N \sim X$
- $f(x) = P(X=x) \simeq \Sigma_{i=1,\ldots,N}\, \mathrm{I}(X_i=x)/N$
- $F(x) = P(X \leq x) \simeq \Sigma_{i=1,\ldots,N}\, \mathrm{I}(X_i \leq x)/N$

where I is an indicator variable (takes the value 1 if true; 0 otherwise)

$\boxed{\text{“Monte Carlo” estimates}}$

Empirical Density Estimates of Continuous Variables

- Simulate N, independent and identically distributed random variables, $X_1, X_2, \ldots, X_N \sim X$
- $F(x) \simeq \Sigma_{i=1,\ldots,N}\, \mathrm{I}(X_i \leq x)/N$

  where I is an indicator variable (1 if true; 0 otherwise)

- Cannot estimate continuous f(x) using the same strategy as for discrete random variables (why?). Instead, we will (informally speaking) use histograms to estimate f(x).

Simulating discrete random variables - Chapter 18

- To generate a random variable X=x, with density f() and cdf F():
- Sample u from F(x) [i.e., sample u from Unif[0,1]].
- x= $F^{-1}(u)$. (discrete r.v.: find the smallest x such that $u \le F(x)$)
- e.g.

```
 set.seed(1473)
# sample from Unif[0,1,]
u<-runif(1,0,1)

# sample X from some distribution F (on the non-negative integers, say)
X<-0
while (F(X)<u){
   X <- X+1
   }
```

Example: Binomial random variables (c.f. page 335 of text)

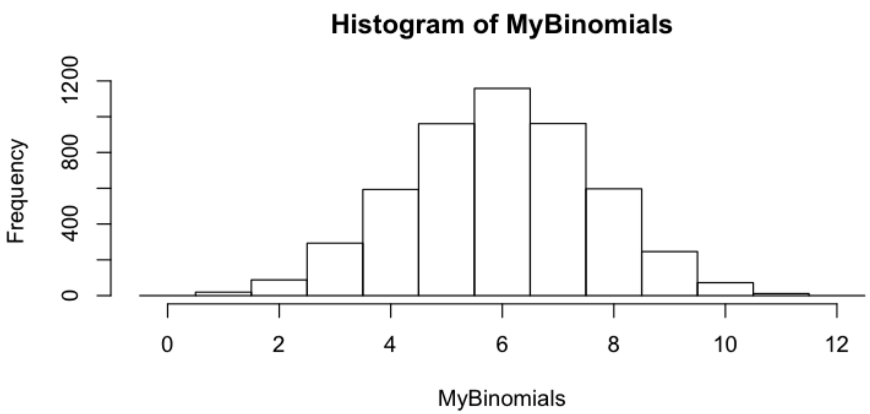
```
set.seed(1473)

binom.cdf<-function(x,n,p){
       Fx<-0
       for (i in 0:x){
           Fx <- Fx + choose(n,i)*p^i*(1-p)^(n-i)
       }
       return (Fx)
}
cdf.sim<-function(F,...){
       X <- 0
       U <- runif(1) # defaults to bounds of 0 and 1
       while (F(X,...)<U){
           X <- X+1
       }
       return (X)
}

MyBinomials<-numeric()
for (i in 1:5000){
           MyBinomials[i]<-cdf.sim(binom.cdf,12,0.5)
}

MyBreaks<-seq(0,13,1)
MyBreaks<-MyBreaks-0.5
BinHist<-hist(MyBinomials,breaks=MyBreaks)
```

... = unspecified number of other arguments



[In repo ‘Week2 - Binomials’ onGithub]

Continuous Random Variable Example: Exponential random variables

- $f(x) = \lambda \exp(-\lambda x)$
- $F(x)=P(X<x)= 1 - \exp(-\lambda x)$ $[0 \le F(x) \le 1]$
- $u \sim U[0,1] \sim F(x)$. So $x \sim F^{-1}(u)$.

- Set $u=F(x)=1-\exp(-\lambda x)$

  So $\exp(-\lambda x)=1-u$

  and $x=(-1/\lambda)\log(1-u)=F^{-1}(u)$

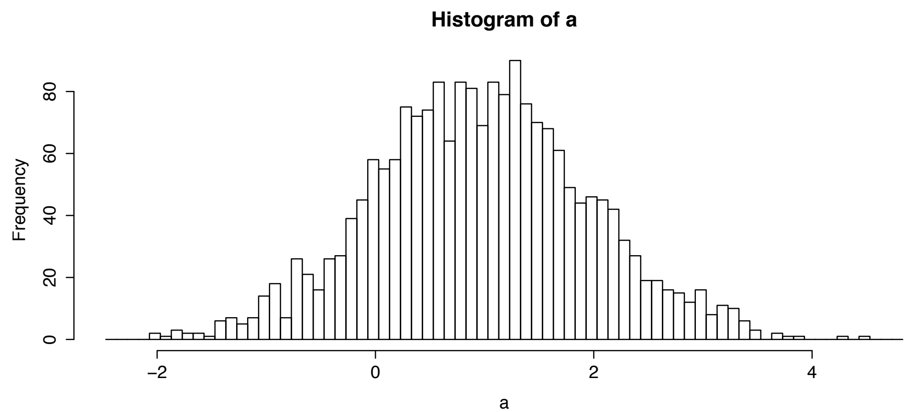
```
lambda <- 1.1   # the (example) parameter for the exponential distn
 u <-runif(1000,0,1)
 ExpRVs <- (-1/lambda)*log(1-u)  # note that this works even when u is a vector
 hist(ExpRVs)
 # or if U~U(0,1), so is 1-U, so....
lambda <- 1.1   # the parameter for the exponential distn
u <- runif(1000,0,1)
ExpRVs <- (-1/lambda)*log(u)
hist(ExpRVs)
```

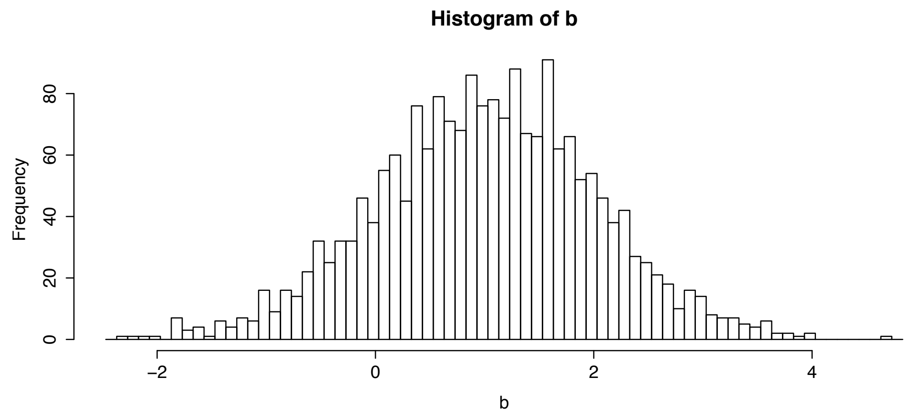
QQ plots

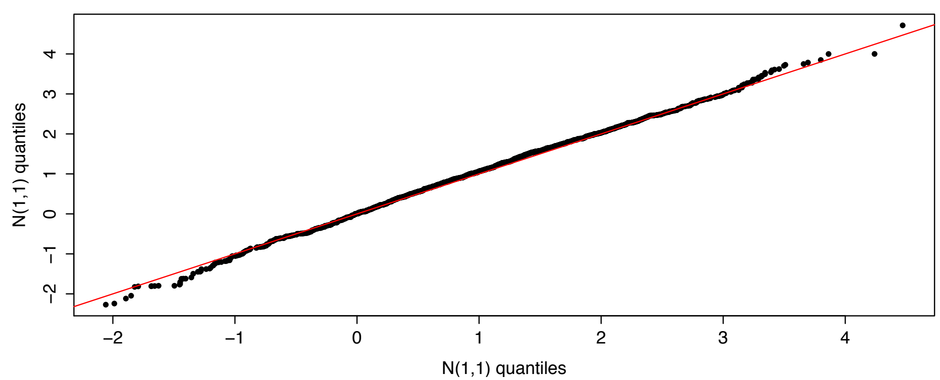
- Used to compare samples, $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$:
  - Order the data points in each sample from low to high, to get $X_{[1]},\ldots,X_{[n]}$ and $Y_{[1]},\ldots,Y_{[n]}$
  - Plot $X_{[1]}$ against $Y_{[1]}$, $X_{[2]}$ against $Y_{[2]}$, $X_{[3]}$ against $Y_{[3]}$, etc.
  - If the distributions are the same, you should see a straight line (for large samples)
  - 'qqplot' in R

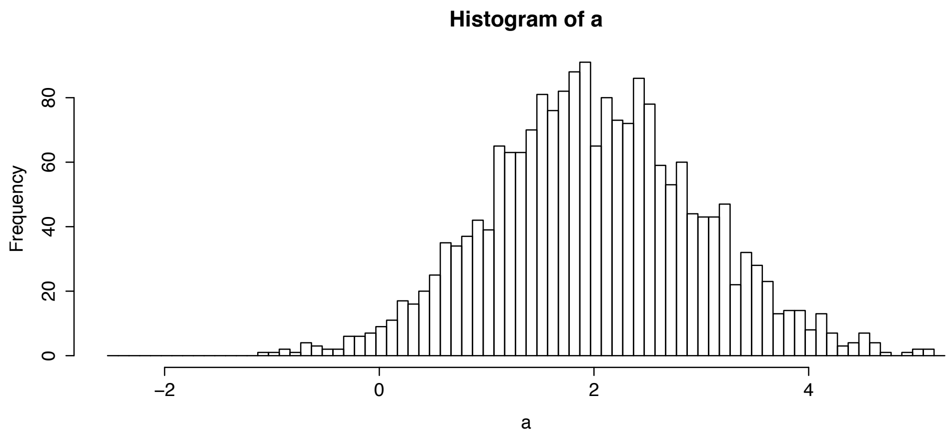
- Can do the same with one sample and Normal random deviates (qqnorm in R)

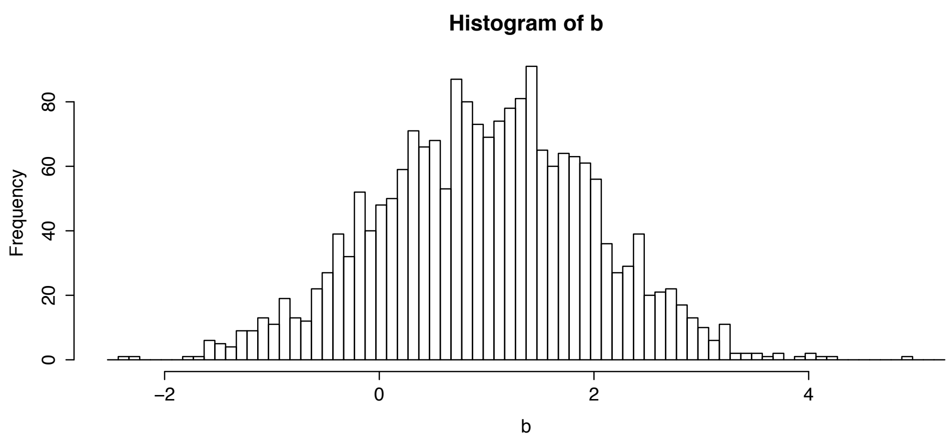
- Formal tests: Kruskal-Wallis test or ANOVA.

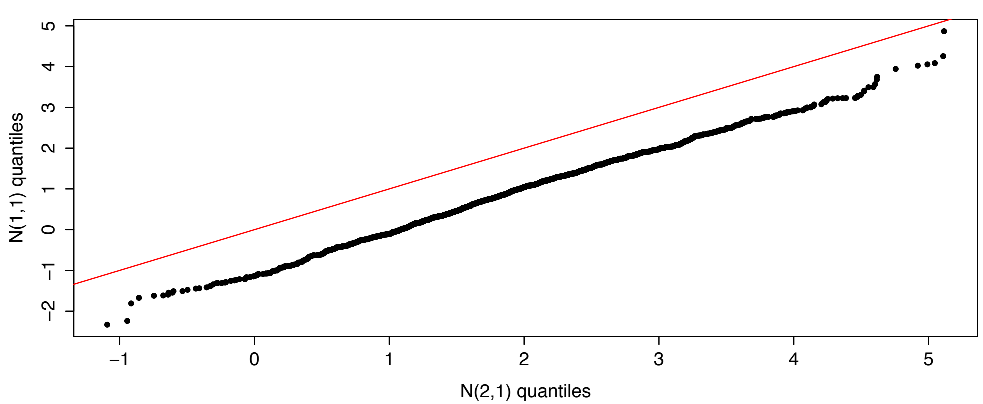

Histogram of a
Frequency
0
20
40
60
80
−2
0
2
4
a

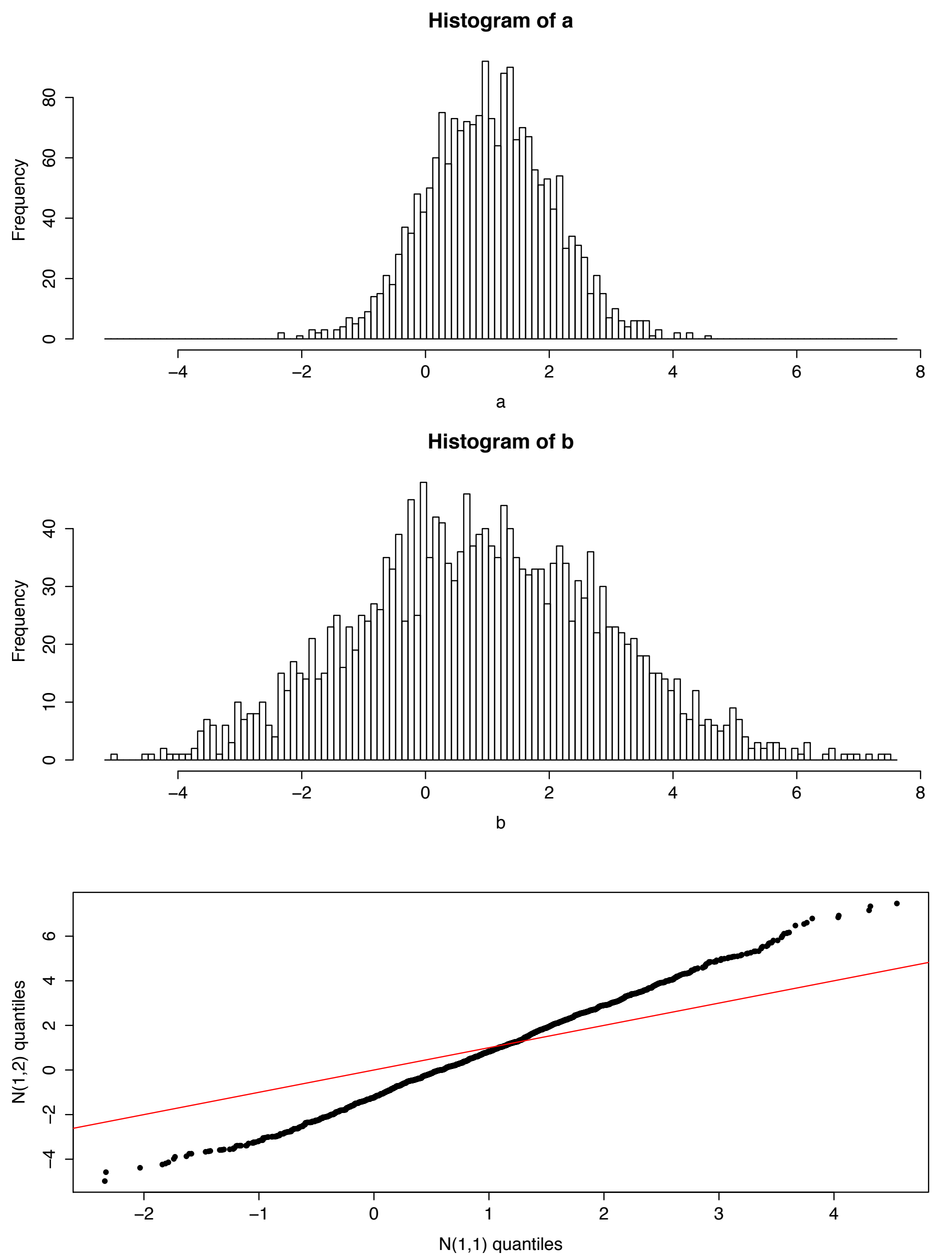

Histogram of b
Frequency
0
20
40
60
80
−2
0
2
4
b


N(1,1) quantiles
−2
−1
0
1
2
3
4
N(1,1) quantiles

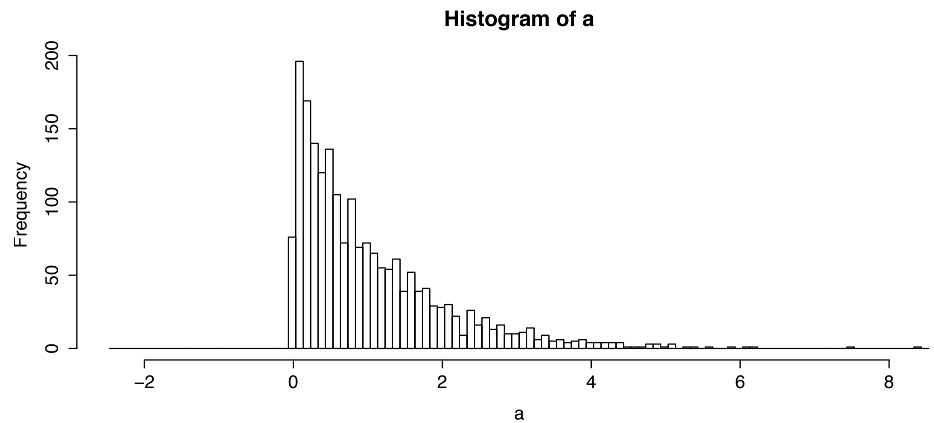

Histogram of a
Frequency
0
20
40
60
80
−2
0
2
4
a


Histogram of b
Frequency
0
20
40
60
80
−2
0
2
4
b

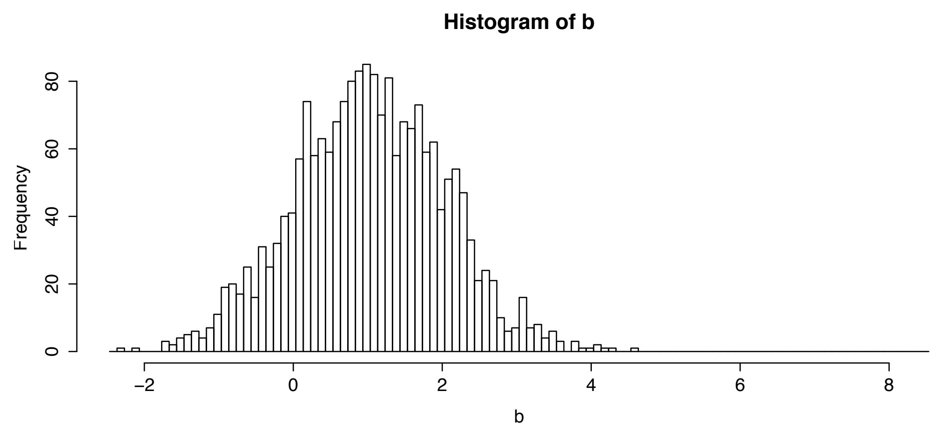

N(1,1) quantiles
−2
−1
0
1
2
3
4
5
N(2,1) quantiles



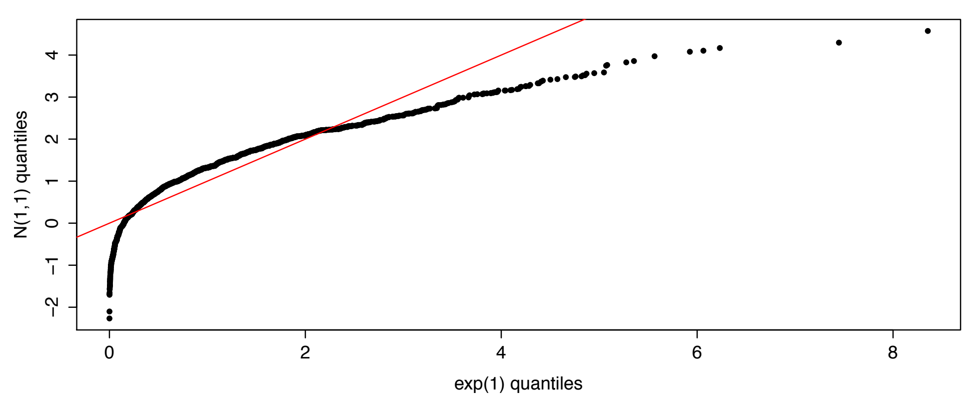
Histogram of a
Frequency
a
Histogram of b
Frequency
b
N(1,2) quantiles
N(1,1) quantiles



Histogram of a
Frequency
0
50
100
150
200
−2
0
2
4
6
8
a



Histogram of b
Frequency
0
20
40
60
80
−2
0
2
4
6
8
b



N(1,1) quantiles
−2
−1
0
1
2
3
4
0
2
4
6
8
exp(1) quantiles

But R has many built-in functions

binom

geom

pois

unif

exp

chisq

gamma

norm

t

...

In-class exercise: Exponential task 1

- Generate 1000 Exp(λ) rvs. conditional on them each being greater than y, for some y (Try λ=1, y=1, say). Let's call those r.v.s X.

- Plot a histogram showing the distribution of (x-y), for y=1 and λ=1, and compare it to 1000 Exp(1) rvs. [or superimpose the exponential density function using the command curve(λ*exp(-λ*x)]

- How do we generate exponential rvs conditional on them being greater than y?

Simple rejection method

- To simulate 1000 exponential r.v.s:

  $u \sim U[0,1] \sim F(x)$. So $x \sim F^{-1}(u)$.

  Set $u=F(x)=1-\exp(-\lambda x)$

  $\exp(-\lambda x)=1-u$

  $x=(-1/\lambda)\log(1-u)=F^{-1}(u)$ [or, $x=(-1/\lambda)\log(u)$]

```
u<-runif(1000,0,1)
ExpRVs<- (-1/lambda)*log(u)
```

- To generate $X \sim \exp(\lambda)$, conditional on $(X>y)$:

```
x <-0
while (x<y){
    # Generate x~exp(λ)
    }
```

The x-value that results has the correct distribution. So repeat that process 1000 times. We'll return to rejection sampling later in the course.

Pseudocode (Week2-ConditionedExponentials repo on Github)

```
set.seed(99999)
# repeat the following until you have 1000 conditioned exponential rvs.
u<-runif(1,0,1)
y<-1    # suppose we want to condition on the rv being bigger than 1
lambda<-2   # suppose we want exponentials with parameter 2
ConditionedExpRV<- (-1/lambda)*log(u)
while (ConditionedExpRV < y){
  u<-runif(1,0,1)
  ConditionedExpRV<- (-1/lambda)*log(u)
}
#Store the value of ConditionedExpRV
```

Exponential: Memoryless property

- Memoryless property:
  - If X is Exp(λ), then $f(x+y|X>y)=f(x)$ (i.e., x-y is still Exp(λ)).

In-class exercise: Exponential task 2: Waiting for a bus

- Suppose times between bus arrivals are distributed as T~exp(1).

1. Suppose we arrive at a bus-stop at some fixed time during the day (say after 10 hours). How long, on average, do we have to wait for a bus? [What if we arrive at a random time each day?]
2. If we get off one bus and wait for the next one to arrive on the same route, how long, on average, do we have to wait?
3. Continuing part 1., how long on average was the time between the arrival of the bus we caught and the one before it.
4. What is the expected time between any two buses?

Note: the mean of an exp(λ) r.v. is 1/λ.

See 'Week2-BusWaitingTimesExercise on Github

END