Lecture 9 (Ch. 2)

We are in The process of defining measures of width and spread for hists and dists. And This is where we are:

histogram location

Sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

~ typical x (based on sample)

histogram spread

Sample variance:

$$s^2 = \frac{1}{n-1}\boxed{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad S_{xx}$$ defn.

= computational formula

Sample std. dev. = s.

~ typical deviation/spread (based on sample).

Other/related measures of spread:

s^2 itself

$$S_{xx} = \sum_{i}^{n}(x_i-\bar{x})^2$$

= numerator of s^2.

distribution/population location

dist./pop mean, or $E[x]$

$$\mu_x \equiv E[x] \equiv \sum_x x\,p(x), \quad \int_{-\infty}^{\infty} x f(x)\,dx$$

~ typical x (based on pop./dist)

Last time, we found The mean of $N(\mu,\sigma)$ and $Poiss(\lambda)$.

$\mu_x = \mu$ $\qquad$ $\mu_x = \lambda$

Today, we find The μ_x of $Binomial(n,\pi)$, and Then

dist. spread

?
(below)

E.g. # of coins tossed, or Sample size, ...

Mean of Binomial (n, π)?

$$E[x] = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} \pi^{x} (1-\pi)^{n-x} \cdot x$$

or μ_x

x=0 contributes zero to the sum

relabel $\sum_x$ and note that $\frac{x}{x!} = \frac{1}{(x-1)!}$

$$= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} \pi^{x} (1-\pi)^{n-x}$$

$y = x-1$

$$= \sum_{y=0}^{n-1} \frac{(n-1)!\, n}{y!(n-y-1)!} \pi^{y+1} (1-\pi)^{n-y-1}$$

$$= \underline{n\pi} \sum_{y=0}^{n-1} \frac{(n-1)!}{y!(n-y-1)!} \pi^{y} (1-\pi)^{n-y-1}$$

$m = n-1$

$$= (m+1)\pi \underbrace{\sum_{y=0}^{m} \frac{m!}{y!(m-y)!} \pi^{y} (1-\pi)^{m-y}}_{=1 = \sum_{y=0}^{n} p(y)}$$

$$= \underbrace{(m+1)}_{n} \pi$$

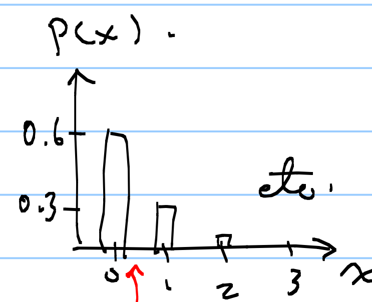
$$\boxed{E[X] = n \cdot \pi}$$

2 params of binomial Note

Note $E[x]$ is not a function of x.

E.g. 1.23:

# of Bads out of 100 → x	0	1	2	3	4
$p(x)$	.6058	.3044	.0757	.0124	---



Hard way.

$$E[x] = \sum_{x=0}^{100} x\, p(x) = 0(.6058) + 1(.3044) + \ldots$$

$$= n\pi = 100(.005) = 0.5$$

Easy way.

On avg. 0.5 out of 100 (ie. 1 out of 200) computers are defective.

Now, let's finish the 2x2 table:

histogram location

Sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

~ typical x (based on sample)

distribution/population location

dist./pop mean, or $E[x]$

$$\mu_x \equiv E[x] = \sum_x x\,p(x) \;,\; \int_{-\infty}^{\infty} x f(x)\,dx$$

~ typical x (based on pop./dist)

histogram spread

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad \text{defn.}$$

= computational formula

Sample std. dev. = s.

~ typical deviation/spread (based on sample).

Other/related measures of spread:

s^2 itself

$$S_{xx} = \sum_i^n (x_i-\bar{x})^2$$

= numerator of s^2.

distr./pop. spread

NEW dist./pop. variance

$$\sigma_x^2 \equiv V[x]$$

$$= \begin{cases} \sum_x (x - E[x])^2\, p(x) \\ \int_{-\infty}^{\infty} (x - E[x])^2 f(x)\,dx \end{cases}$$

Don't drop this x, like the book does.

dist./pop. standard dev. = σ_x

~ typical dev./spread (based on pop/dist.)

Now, let's find the var. of some of our named distributions

$E[X]=\mu$

Normal(μ,σ): $\sigma_x^2 = V[X] = \int (x-\mu)^2 f(x)\,dx = \cdots = \sigma^2$ (Normal)

which is why the param. σ^2 is called (distr.) variance.

μ_x of Binomial, $p(x)$ of Bin.

Binomial(n,π) : $\sigma_x^2 = V[X] = \sum_x (x-n\pi)^2\, p(x) = \cdots = n\pi(1-\pi)$.

See hw and/or prelab for better understanding This formula!

Poisson(λ): $\sigma_x^2 = V[X] = \sum (x-\lambda)^2\, p(x) = \cdots = \lambda$ (poisson)

Recall $E[X]=\lambda$ ⟵ same! ⟶

$\boxed{\text{Summary}}$ (Some of These are done in hw)

	binomial(n,π)	poisson(λ)	$N(\mu,\sigma)$	Unif(a,b)	Exp(λ)
$E[X]=\mu_x$	$n\pi$	λ	μ	$(a+b)/2$	$1/\lambda$
$V[X]=\sigma_x^2$	$n\pi(1-\pi)$	λ	σ^2	$(b-a)^2/12$	$1/\lambda^2$

Again note The diff. between μ_x $(\equiv E[X])$ and μ.

mean of any dist.

μ param. of $N(\mu,\sigma)$.

READ!

Jargon

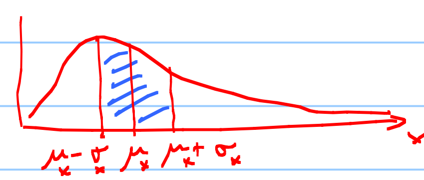
We now have measures of location & spread for hists & dists.

histograms		vs.	distributions / pop.	
Sample mean	$\bar{x}$	vs.	distr. mean	$E[x] \equiv \mu_x$
" variance	s^2	vs.	" variance	$V[x] \equiv \sigma_x^2$
" std. dev.	s	vs.	" std. dev.	σ_x
sample "statistic"			pop. "parameters"	

One says that $\bar{x}$ is a point estimate of μ_x; Etc.
(This statement may seem obvious, because of the similarity of the names (e.g. sample mean and pop. mean); but it's not!

Let's not forget "areas". Before Ch. 2 we used to say things like
For $N(\mu, \sigma)$: 68% of the area is within 1 σ of μ.
But now, that translates to ← or 2, or 2.5, ...
68% of the area is within 1 std. dev. of the mean.

And now we can say things like that
for any distr. e.g. Poisson, ...



Computing "areas" like this will (eventually) allow us to provide some measure of confidence as to what μ_x is, based on observed data.

Recall,

For hists: "area" = $\boxed{\text{prop. of times}}$ (probability of) x is observed to be within ...

For dists: "area" = $\boxed{\text{prop. of times}}$ x is expected to be within ...

[because we don't know the pop./dist. But if we assume the dist. that describes the pop., then we would expect ...]

Switching gears

A very powerful tool!

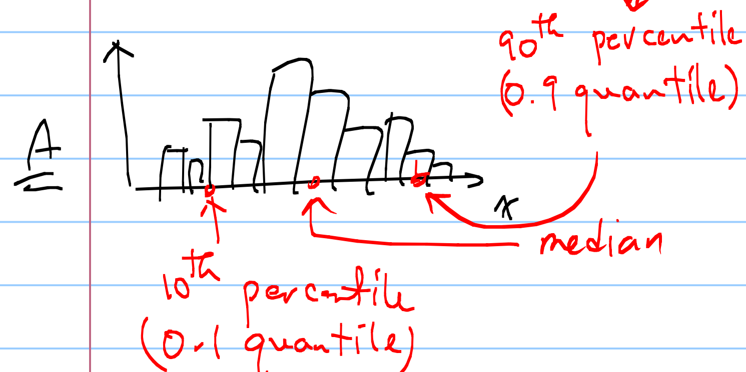
(qq plots)

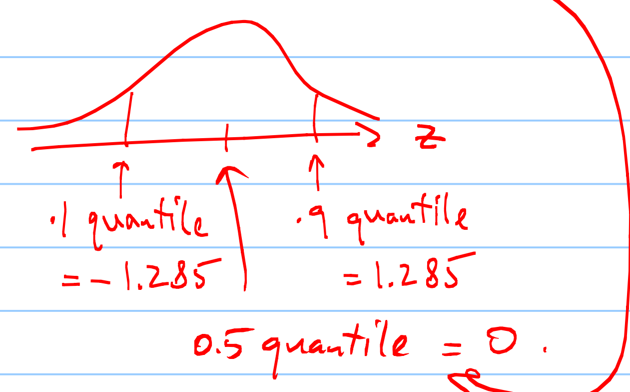
The business of estimating pop. params from sample stats refers to any distr. E.g., one says that $\bar{x}$ and s provide point estimates of μ and σ of the normal dist. IF the data come from a normal dist. to begin with.

Q: But, how do we know if our data come from a Normal dist?

Easier Q: How do we know if our data come from std. Normal?

A: compare sample quantiles (of data) with distr. (or theoretical) quantiles.





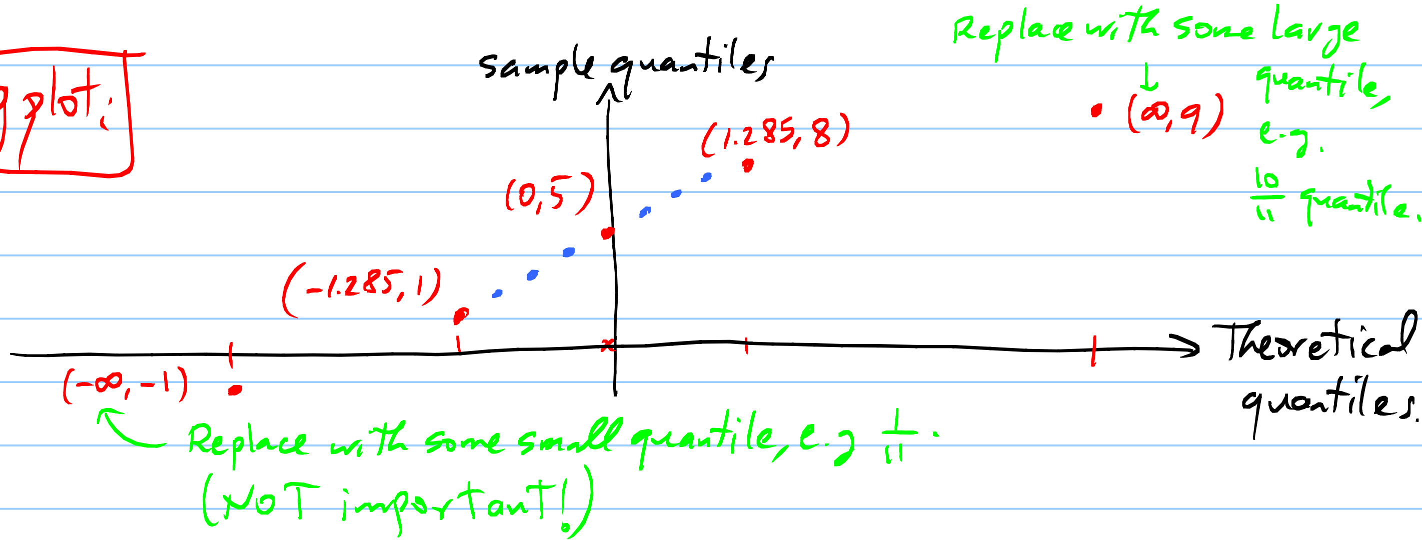
Example: (Very Crude!) Here is (sorted) data:

Data =	-1,	+1,	3,	4,	4.5,	5,	5.5,	6,	6.5,	8,	9
quantile prob. =	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Theoretical quantile =	$-\infty$	-1.285	---			0	---			$+1.285$	∞

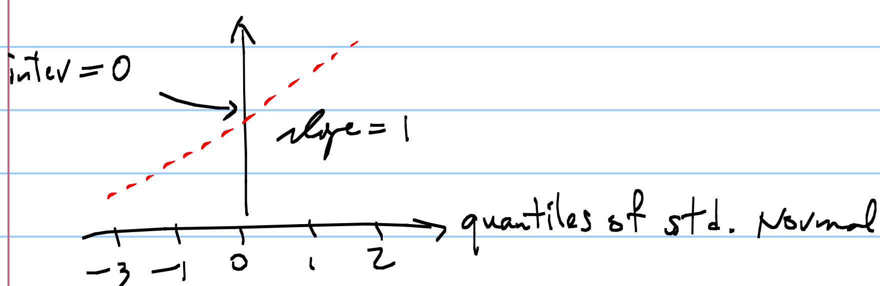
Now, we want to see if the sample quantiles "line-up" with the distribution quantiles. In Ch.3 we will learn that the best way of seeing if 2 columns of numbers "line-up" is to plot one vs. the other, i.e.

	0	0.1	0.2	...	0.9	1.0
y = sample quantile	-1	+1	3		8	9
x = dist. quantile	$-\infty$	-1.285		...	1.285	$+\infty$

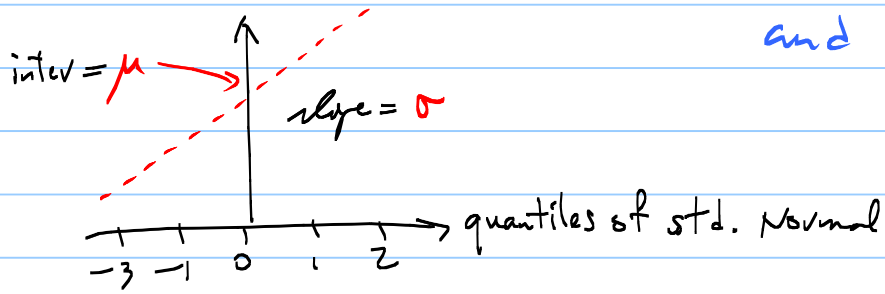
qq plot:



If the histogram is consistent with a std. Normal, then the quantiles/percentiles of data should be equal/comparable to those of the distr.. Then the qq plot should be a straight diagonal line (intercept = 0, slope = 1).



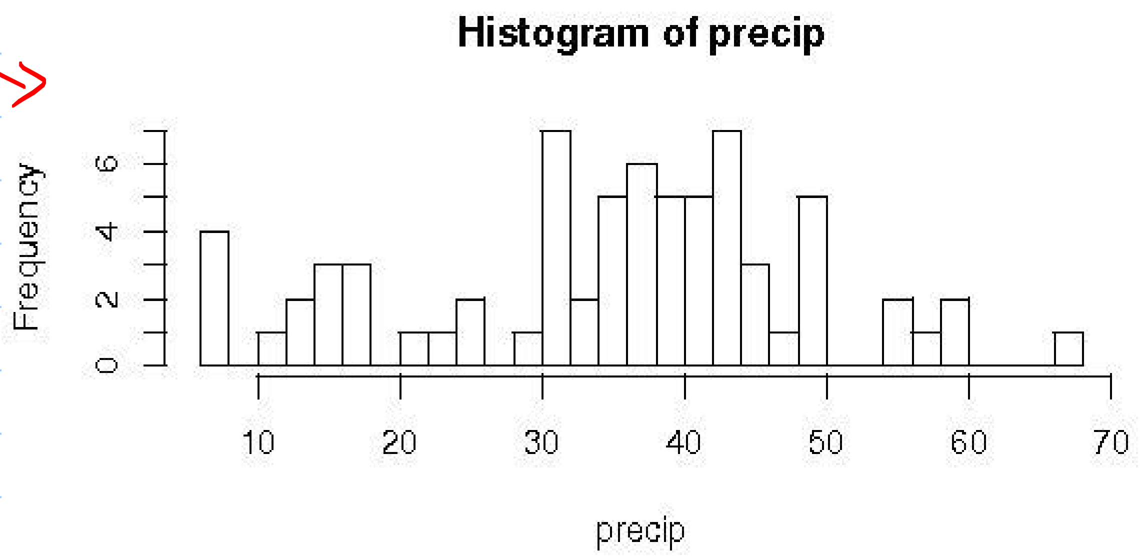
If the data are $\boxed{\text{not}}$ from std. normal, but from $\boxed{N(\mu, \sigma)}$, the only thing that changes is that the slope becomes σ, and the intercept becomes μ. The proof is easy, but later. For now, focus on the concept and the use of qqplots

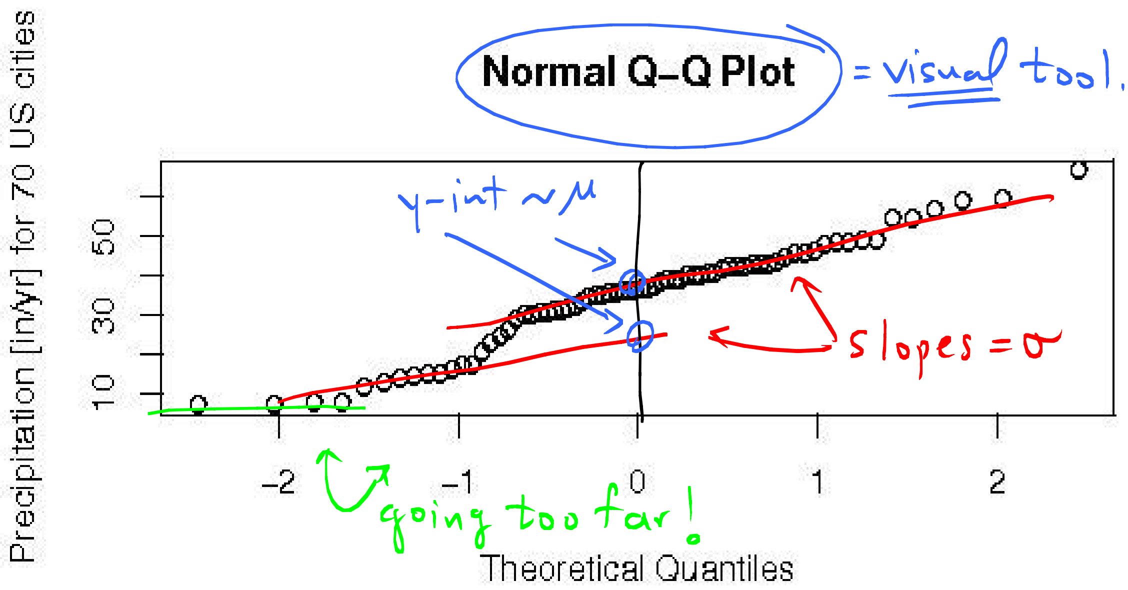


In R: qqnorm(x) where x is the vector of data.

Example

From The histogram, it's hard to tell if The data come from a normal dist., especially because hists depend on bin size.



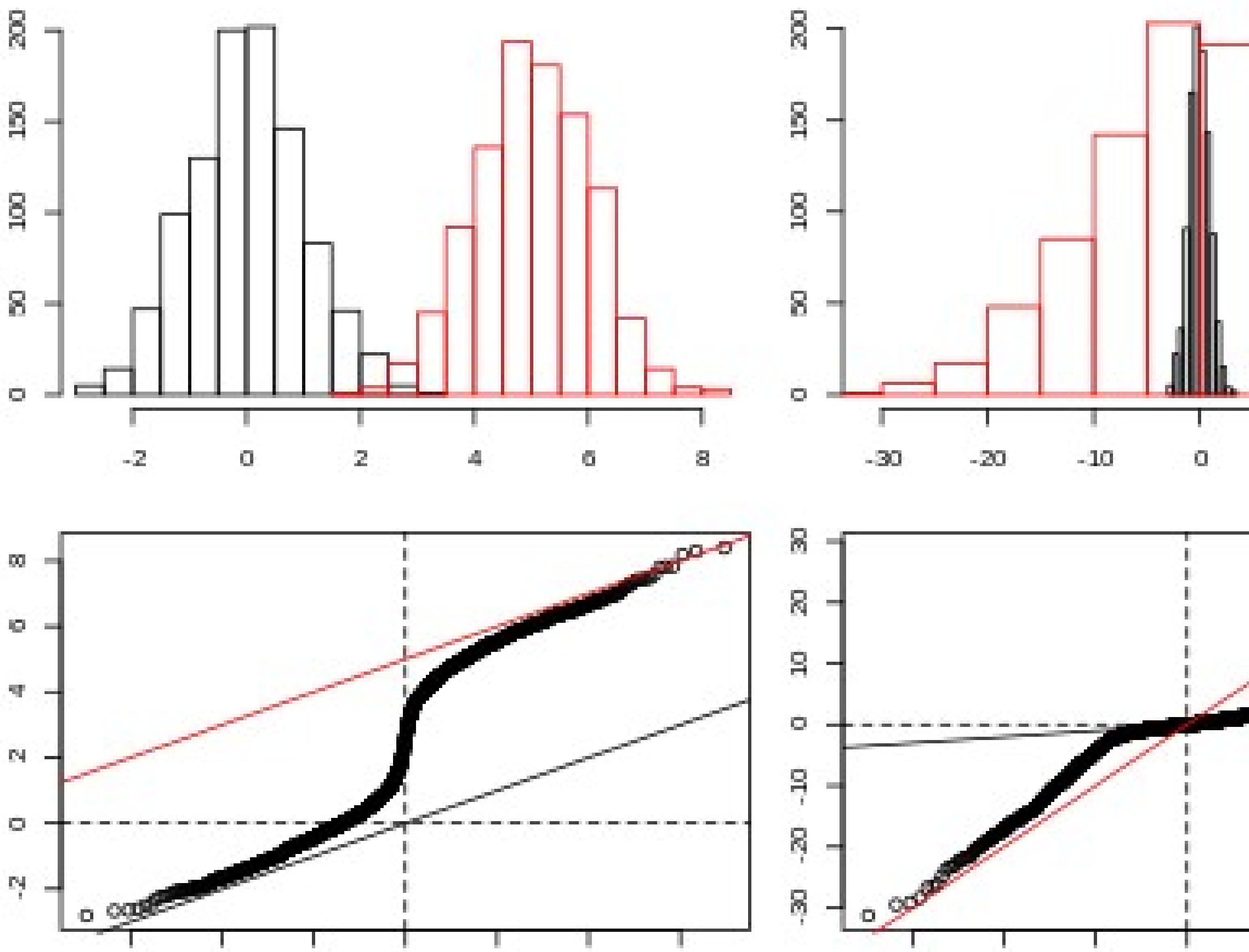


$\Rightarrow$ The qq-plot looks mostly linear $\Rightarrow$ Data are consistent with Normal dist.

$\Rightarrow$ Looking deeper, infact it looks like data may have come from a bi-modal distr. (ie. 2 normals with same σ but different μ's)

(However, FYI, see next page)

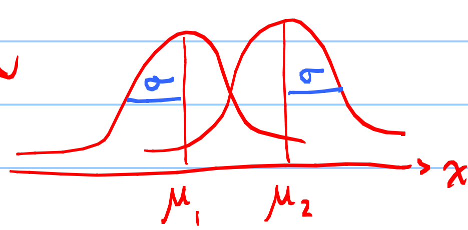
FYI

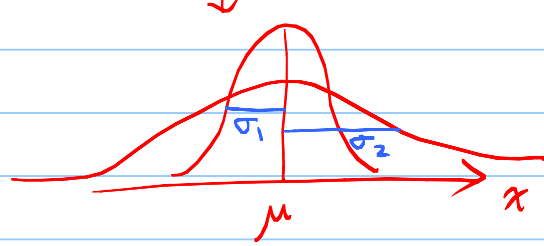


n = 1000
x = rnorm(n, 0, 1)
y = rnorm(n, 5, 1)
hist(x, xlim=range(x,y))
hist(y, add=T, border=2)

n = 100
x = rnorm(n
y = rnorm(n,
hist(x, xlim=ra
hist(y, add=T, b

A "bend" in the qq-plot can be interpreted as
1) two normals with the different mu's, and the same sigma (if the tail ends of qqplot are parallel)
2) two normals with the same mu, and different sigma's.

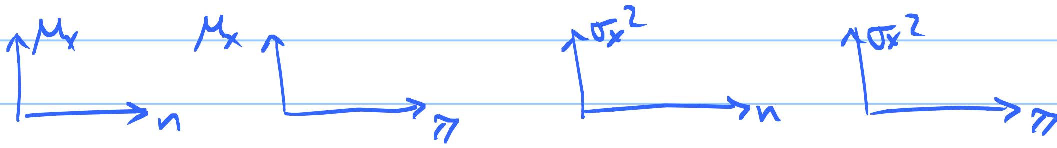




hw_lect9-1

a) Consider the binomial dist. with params n,pi. Draw four figures that show qualitatively how its mean (mu_x) and variance (sigma_x^2) vary with n and pi.

ie.



Suppose we toss n=l00 unfair coins, with an unknown pi.
b) What is the expected number of heads out of n? (The answer depends on pi),
c) What is the typical deviation in the number of heads out of n? (The answer depends on pi).
d) what is the largest typical deviation of the number of heads out of n? (The answer is a number!)
Hint: Consult your graph of variance vs. pi, in part a).

hw-lect9-2

For the exponential distribution with parameter lambda, find the variance.Hint: You may use this integral:

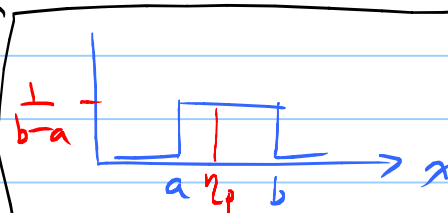
$$\int_0^\infty (y-1)^2 e^{-y}\, dy = 1$$

hw_lect9_3: Find the prob that x is within 1 standard deviation of the mean, for

a) binomial ($n=20$, $\pi = \frac{1}{4}$)
b) poisson ($\lambda = 5$)
c) Normal ($\mu = 5$, $\sigma = 1$)

hw-lect9_4

It can be shown that the p^{th} percentile of Unif (a,b) is given by $\eta_p(a,b) = a + (b-a)\frac{p}{100}$



$$\frac{(\eta_p - a)}{b-a} = \frac{p}{100}$$

$$\eta_p = a + (b-a)\frac{p}{100}$$

a) What's the p^{th} percentile of Unif$(0,1)$, ie. $\eta_p(0,1)$?

b) Write $\eta_p(a,b)$ in terms of $\eta_p(0,1)$.

c) What will the plot of $\eta_p(a,b)$ vs. $\eta_p(0,1)$ look like? What are the slope & y-intercept?

Later, check the soln to see the moral

hw-lect9_5

Do a qq-plot of each of the 2 cont. vars. in the data you collected. By R. Describe/Interpret the result.
Note: If you find out that there is not much you can say about the qq-plot, it may be that your data is not appropriate. This is another chance to correct the error, because later you will be doing more hw problems using your data. So, see me, if you are not sure.