Lecture 9 (Ch. 2)

So far, we have introduced the following measures of location and spread of

Histogram/Sample/Data

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

sample mean

Important

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

sample variance, s = sample std. dev.

~ typical deviation

Distribution/population

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

dist./pop. mean

?

next.

Now, a popular measure of distr. spread is (distr.) variance:

$$V[X] = \sigma_X^2 = \begin{cases} \sum_{\forall x} (x - E[X])^2\, p(x) \\ \int_{-\infty}^{\infty} (x - E[X])^2 f(x)\, dx \end{cases}$$

$E[X] = \sum x\, p(x)$

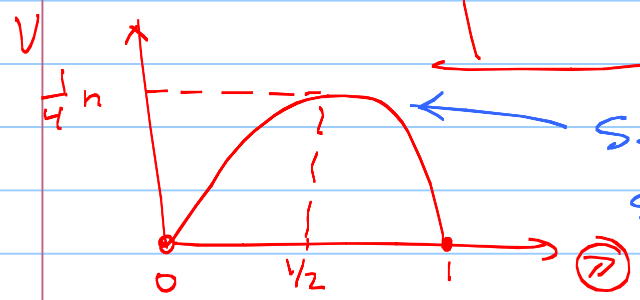
The book drops the x, but then this σ can be confused with the σ of the Normal distr.

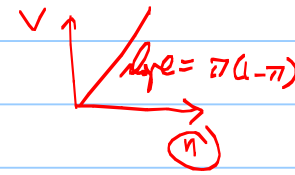
As with sample std. dev., σ_x = distr. std. dev. ← same units as x.

Let's compute $V[X] = \sigma_x^2$ for some of our special distr.:

$E[X] = n\pi$

Binomial(n,π) : $\boxed{V[X] \equiv \sigma_x^2 = \sum_x (x-n\pi)^2 \underset{\text{"binomial"}}{p(x)} = \cdots = n\pi(1-\pi)}$





So the largest possible standard deviation is $\sqrt{\frac{1}{4}n} = \frac{1}{2}\sqrt{n}$, at $\pi = \frac{1}{2}$.

I.e. If I'm trying to estimate the number of heads out of n tosses, the largest "uncertainty" is $\frac{1}{2}\sqrt{n}$ ($\boxed{\text{at } \pi = \tfrac{1}{2}}$).

Poisson(λ) : $V[X] \overset{=\sigma_x^2}{} = \sum_x (x-\lambda)^2 \underset{}{\overset{\text{"poisson"}}{p(x)}} = \cdots = \lambda$

Recall $E[X] = \lambda$ ← same →

$E[X] = \mu$

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

which is why the param. σ^2 is called variance.

$\boxed{\text{Q1:}}$ Let $f(x) = 2x$, $0 < x < 1$. In the last qz, we found $E[X] \equiv \mu_x = \frac{2}{3}$. Now, find $V[X] \equiv \sigma_x^2$?

A) 1 B) $\frac{1}{2}$ $\boxed{\text{C) } \frac{1}{18}}$ D) This $f(x)$ has no variance.

$$V[X] \equiv \sigma_x^2 = \int_{-\infty}^{\infty} (x-\mu_x)^2 f(x)\,dx = \int_0^1 \left(x-\frac{2}{3}\right)^2 2x\,dx = 2\int_0^1 \left(x^3 - \frac{4}{3}x^2 + \frac{4}{9}x\right)dx$$

$$= 2\left[\frac{1}{4}x^4 - \frac{4}{3}\frac{1}{3}x^3 + \frac{4}{9}\frac{1}{2}x^2\right]_0^1 = 2\left[\frac{1}{4} - \frac{4}{9} + \frac{2}{9}\right] = 2\left(\frac{1}{4} - \frac{2}{9}\right) = \frac{2}{36} = \frac{1}{18}$$

By now, you should be familiar with the meaning of

histograms vs. distributions

Sample mean $\bar{x}$ vs. distr. mean $E[x] = \mu_x$

" Variance s^2 vs. " Variance $V[x] = \sigma_x^2$

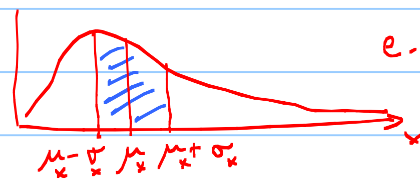
" std. dev. s vs. " std. dev. σ_x

Finally, given that we can compute all of the above quantities, you can then compute the proportion of times x is expected to be within some std. dev. of its mean, for ANY distr.

↑ 1, 1.96, 2, ---

For examples, for the normal dist. we can now say that 68% of x's fall within 1 std. dev. of the mean.

But now we can say things like that for any distr. even skewed ones:



e.g. Poisson, ---

Computing areas like this will eventually enable us to provide some measure of confidence when we try to estimate a population parameter, later.

Summary

Single summary of histogram location

Sample mean:

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

$\sim$ typical x/obs.

Single summary of histogram spread

Sample variance:

recall computational formula, too. $\rightarrow$

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

Sample std. dev. $= s$.

$\sim$ typical deviation/spread

Recall why $\frac{1}{n}\sum(x_i - \bar{x})'$ will not do!

Single summary of distribution/population location

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

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

Single summary of dist/pop. spread

dist/pop. var. or $V[X]$

$$\sigma_x^2 \equiv V[X] = \sum_x (x-\mu_x)^2 p(x)$$

$$\int_{-\infty}^{\infty} (x-\mu_x)^2 f(x)\,dx$$

Eg. binomial(n,π): $\mu_x = n\pi$ $\quad \sigma_x^2 = n\pi(1-\pi)$

poisson(λ): $\mu_x = \lambda$ $\quad \sigma_x^2 = \lambda$

Normal(μ,σ): $\mu_x = \mu$ $\quad \sigma_x^2 = \sigma^2$

uniform(a,b): $\mu_x = \frac{a+b}{2}$ $\quad \sigma_x^2 = \frac{(b-a)^2}{12}$

Exponential(λ): $\mu_x = \frac{1}{\lambda}$ $\quad \sigma_x^2 = \left(\frac{1}{\lambda}\right)^2$

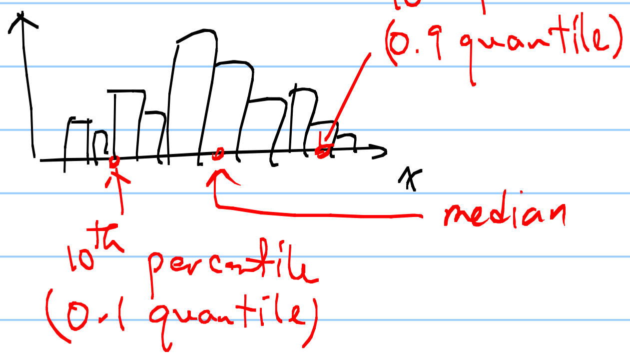
One more Thing in Ch 2.

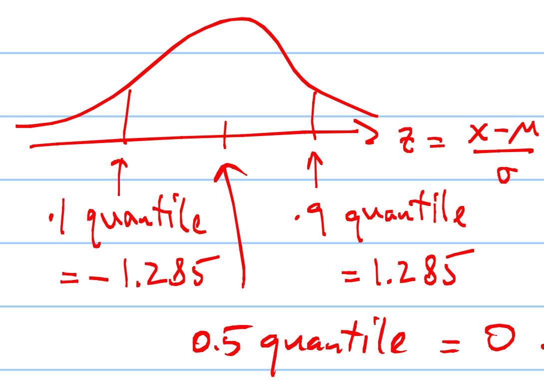
This business of estimating pop. dist. parameters refers to **any** parameter. Specifically, $\bar{x}$ and s provide **point estimates** for μ, σ, respectively, of the Normal distr. if the data come from a Normal distr. to begin with.

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

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

A





Example: (Very Crude!)

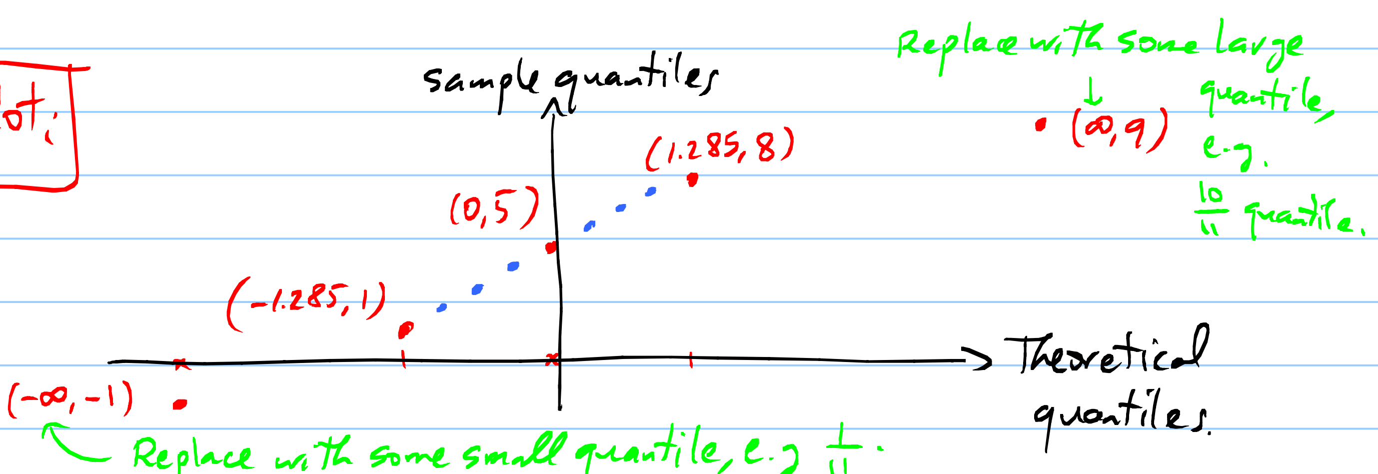
Data, sorted:

−1,	1,	3,	4,	4.5,	5,	5.5,	6,	6.5,	8,	9
0th quantile	0.1 quantile	---			0.5 quantile		---		0.9 quantile	1.0 quantile

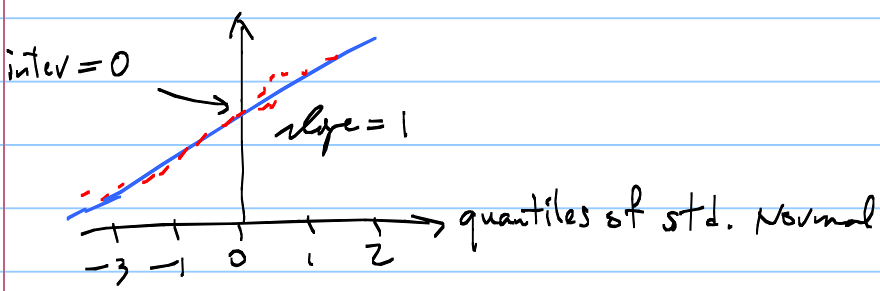
→ Theoretical quantiles:

$-\infty$	-1.285	---	0	---	$+1.285$	∞

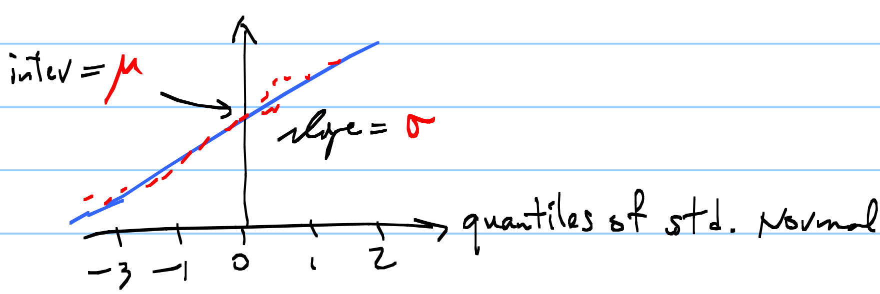
$\boxed{\text{qq plot:}}$



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).



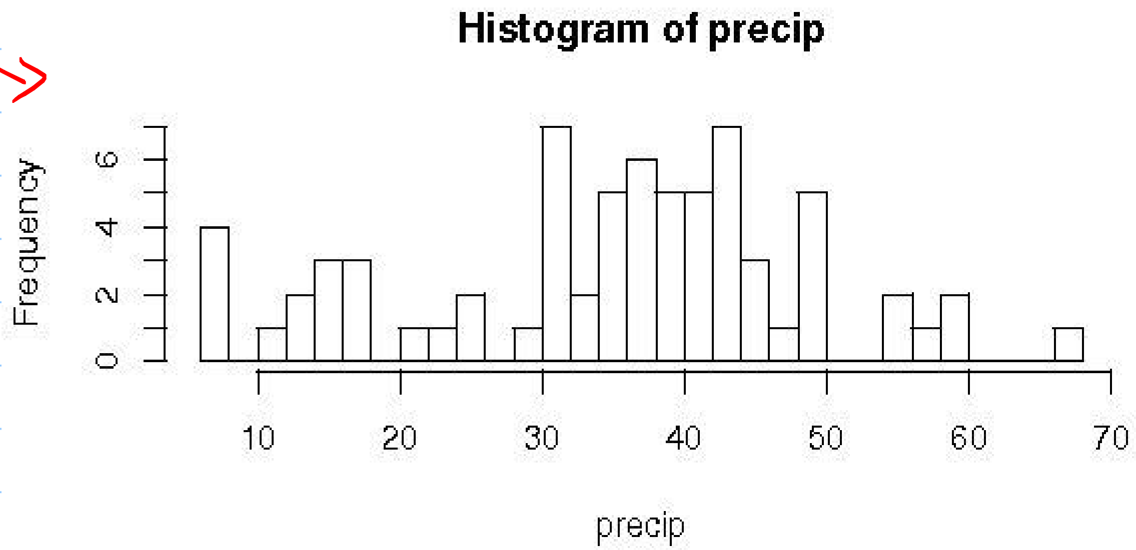
If The data are $\boxed{\text{not}}$ from std. normal, but from $\boxed{N(\mu, \sigma)}$, the only Thing That changes is That The slope becomes σ, and The intercept becomes μ. NOT too obvious, but pf. in book.



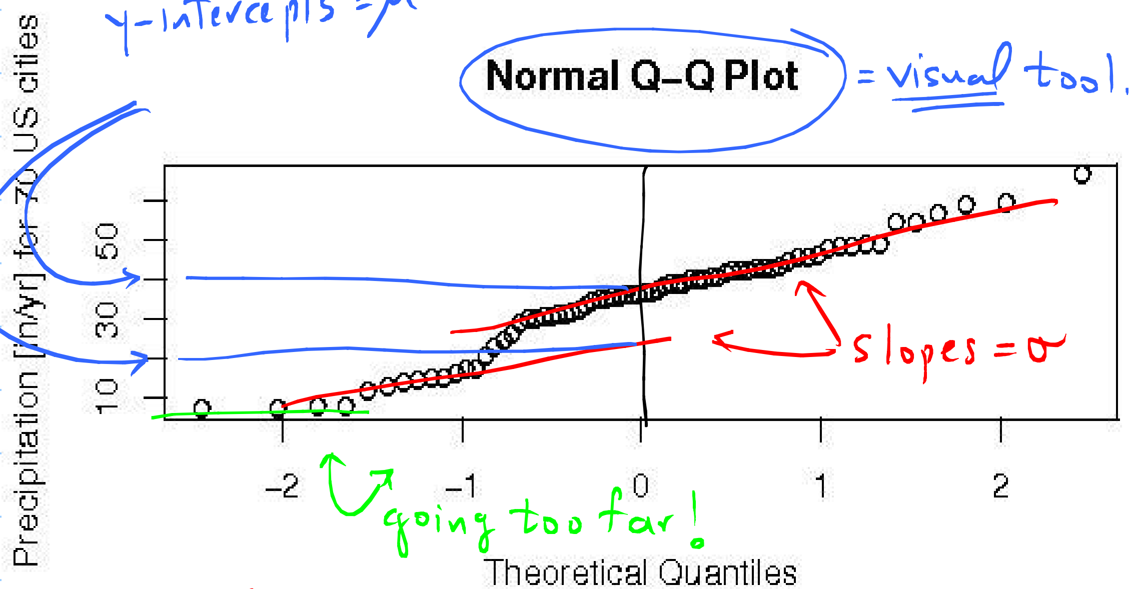
In R: qqnorm(x) where x is The vector of data.

Example

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



y-intercepts = μ



The plot looks linear, mostly!
So, data are consistent with a Normal.
In fact, it looks like 2 different normals (Bimodal) with diff μ's, same σ (slope).

hw-lect11-1 For The uniform distr. between a,b, show $V(X) = \frac{1}{12}(b-a)^2$

hw-lect11-2 Find The variance of The exp. dist. with param. λ.

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

hw-lect11-3 In Example 1.22 (in Text and in Lect), we found That on the average, out of 100 computers, 0.5 computers are defective.

a) What is The typical deviation we expect to see from This number (still out of 100)?

b) Suppose we do not know That The proportion of defective computers is 0.005. Then out of 100 computers, what is The maximum value we expect to see for typical deviation?

hw-lect11-4 Find The area within $\mu_x \pm \sigma_x$ for

a) binomial ($n=20$, $\pi = \frac{1}{4}$)

b) poisson ($\lambda = 5$)

c) Normal ($\mu = 5$, $\sigma = 1$)

hw-lect11-5 Do a qq plot of each of The 2 cont. vars. in The data from hw-lect1. By R. Describe/Interpret The results.

Note: If you find out That There is not much you can say about The qqplot, 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.