Lecture 18

Probability Plots.

- Imagine we have some sample $\{x_1, \ldots, x_n\}$. (from some larger population)

- We may want to assume the data/population is distributed in a particular way.

  ↳ How can we know if this is a reasonable assumption? eg: given some data, can we assume a normal distribution?

- Assumption more formally: we assume that the population consists of values of a random var X with cdf $F(x)$.

- we use a probability plot to test this assumption.

- MiniTab (Graphs > probability plot ...)

Step 1: Order the sample in increasing order and rename:

$$\{x_1, \ldots, x_n\} \longrightarrow x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

i.e. $x_{(1)}$ is the smallest, $x_{(n)}$ is largest.

Now / Idea: want $x_{(i)} \sim 100\left(\frac{i}{n}\right)^{th}$ percentile.

↳ eg: if $\{x_{(1)} \leq \dots \leq x_{(11)}\}$ is our sample, then

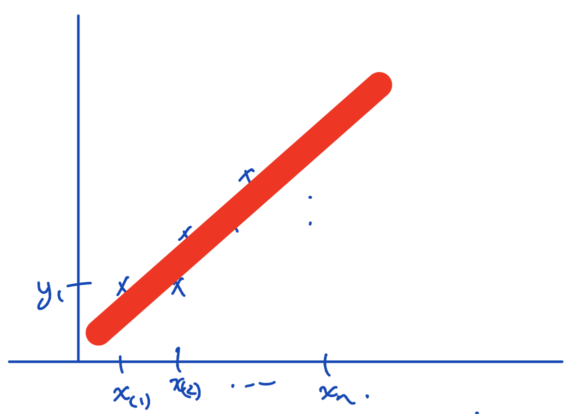
we want $x_{(6)} \sim$ median.

Step 2: Find some points y_i such that:

$$\underline{P(X \leq y_i) = F(y_i)} = \frac{i - 0.5}{n}$$

correction factor

Step 3: Plot $(x_{(i)}, y_i)$.



Step 4: Draw a straight line of best fit.

Conclusion: Are all the points on or very near the line?

→ yes: your assumption is good. i.e. the distribution of X is a good reasonable fit to the data.

→ no: not so good.

most of the time we will be concerned with the normal distribution.

↳ here, $F(x) = \Phi(x)$.

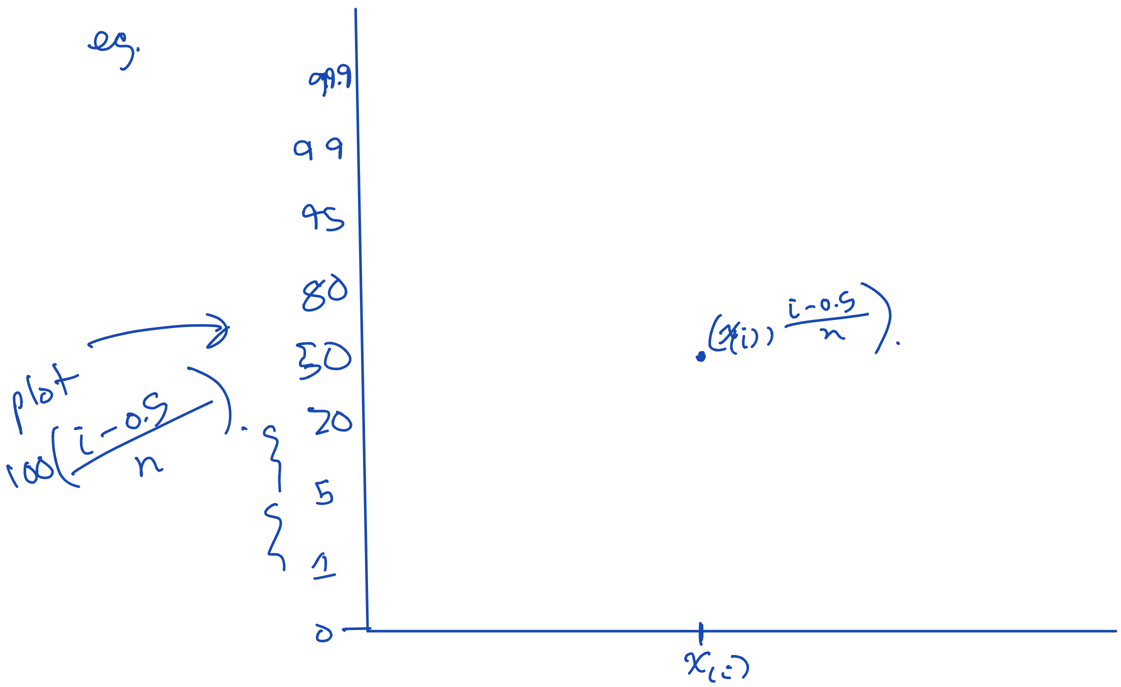
↳ to find y_i's, use Z-score tables.

↳ want y_i s.t.

$$\Phi(y_i) = \frac{i-0.5}{n}.$$

<u>Remark:</u> Sometimes prob plots are done on particular graphing paper

e.g.



Chapter 7. Point ~~of~~ Parameters.

Recall: the general goal of ~~statistics~~/statistical inference is to make predictions/draw conclusions about population, especially based on limited data.

- a major component of statistical inference is called parameter estimation.

  eg: maybe you have some data set and you want to estimate the mean or variance.

- In general, a parameter (usually denoted by a lowercase θ) is any numerical property/feature of the data being studied.

- Recall that we assume a sample $\{x_1, \ldots, x_n\}$ is a particular instance of independent and identically distributed random variables $\{X_1, \ldots, X_n\}$.

- A statistic is any function of random variables.

  eg. - $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ ← sample mean
  - S^2
  - S

- Given a particular parameter, θ, an estimator for θ is a statistic $\wedge$

$\hat{\Theta} = h(X_1, \dots, X_n)$ ← capital theta w/ hat

used to estimate θ.

↳ e.g. $\bar{X}$ is a estimator for μ.

↑ statistic. ↑ parameter

- If $\hat{\Theta} = h(X_1, \dots, X_n)$ is an estimator for θ, then a particular value $h(x_1, \dots, x_n) = \hat{\theta}$ ← a point estimate for θ.

Descriptive Statistics: Sample

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Sample	20	0	82.500	2.620	11.718	55.560	73.610	86.110	88.890	100.000

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	Test 1 Scores	Sample										
1	72.22	77.78										
2	61.11	55.56										
3	66.67	88.89										
4	83.33	83.33										
5	83.33	94.44										
6	94.44	66.67										
7	88.89	94.44										
8	88.89	88.89										
9	61.11	66.67										
10	66.67	77.78										
11	55.56	88.89										
12	83.33	100.00										
13	94.44	83.33										
14	72.22	83.33										
15	77.78	88.89										
16	72.22	94.44										
17	77.78	66.67										
18	38.89	88.89										
19	83.33	88.89										
20	61.11	72.22										
21	100.00											
22	83.33											

Ex: The mean for test 1 was $\mu = 79.15$.

Taking a sample of size 20.

$\bar{x} = 82.5 \sim 79.15$.

point estimate for μ.

Similarly: the sample variance

$\hat{\sigma}^2 = S^2$ is an estimator for σ^2.

Sample Distributions & Central Limit Theorem.

- Recall that a statistic is a function of random variables, $h(X_1, \dots, X_n)$

- So a statistic is itself a random variable.
  ↳ therefore, each statistic has a prob. distrib.
    ↳ such distributions are called sampling distributions.
  ↳ eg: - $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ ← sampling distribution of the mean (μ)

  - S^2 is the sample distribution of σ^2.

<u>Next:</u> Central limit theorem.