- We study non-parametric methods to try an avoid specific functional forms. By doing this we expect that the resulting approach becomes useful for a wider class of datasets.
- discrete data: pmf = $\hat{f}(x)$, CDF = $F(x)$; continous data: pdf = $f(x)$, CDF = $F(x)$ for continous $f(x)$ = on intervals area, $F(x)$ = all area to the left of x.
- Emperical pmf $\hat{f}_n(x)$ is used to elect Emperical CDF($\hat{F}_n(x)$). $\hat{F}_n(x)$ is an unbiased estemate of $F(x)$. It's saying that as we collect more samples $\hat{F}_n(x)$ will = $F(x)$.
- Kernel density estemation: Solves the problem of values from our continous dataset only showing probability of the set. To solve this we add bandwidth to our set of numbers so they can give us the probability of getting numbers exactly as x and the neighborhood of x. neighborhood = $(x_i - h, x_i + h)$
  - Uniform Kernel: $\frac{(x_1 + x_2 + x_3 \dots x_n)}{n}$ where if x_i neighborhood contains $f(x)$ then $x_i = \frac{1}{2n}$. Gives us the simple average of information at $\hat{f}_{K,n}(x)$ Uniform will give a blocky piecewise constant function, as there is no decay from the x_i to $x_i - h$.
  - Triangular Kernel: $K_h(x, x_i) = \frac{1}{h} - \frac{|x - x_i|}{h^2}$; if x is within neighborhood, else 0. add up the product and divide by h to get answer. This function will give a triangle esque graph.
  - Quadratic Kernel Function: $K_h(x, x_i) = \frac{3}{4h} - \frac{3}{3}\left(\frac{(x - x_i)^2}{h^3}\right)$ if x is within the neighborhood more smooth of a graph than previous, due to better decay.
  - GAUSsian Kernel: h now plays a role in determining how quickly the information provided by the observation decays as we move away from said observation in either direction. $K_h(x, x_i) = \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{1}{2}\left(\frac{x - x_i}{h^2}\right)^2}$ The graph will now look like a normal distribution, where a larger h will increase std and smaller h will decrease std. When h is smaller the graph will show local peaks and valleys. We want to smooth enough to not show non-existant peaks.
  - To find optimized bandwith use cross-validation methods where we remove parts of the data and estimate the rest of the data to examine how well the density fits the left out data.
- Histogram density Estimate: create a set of bins that cover the data range, In the bin that held x_i calculate # of observed datat within bin / (# of total observations times h)
- Goodness of Fit: used to test if we can use parametric model with CDF $F_0(x)$.
  - χ^2 gof (chi squared): $H_0: F = F_0$ $H_1: F \neq F_0$. Table constructed bellow with formulas.

values/interval of values Ex: $I_1, I_2 \dots$	number of data points within said intervals Ex: $n_1, n_2 \dots$	Probability of interval I_K divided by total $I = P_K$	discrepancy between data and parametric model: $d_K = n_K - nP_K$

Required criteria: $I_1, I_2, \dots, I_K$ don't overlap., sum of entries in second colum is equal to sample size n., Sum of entries in 3rd column are equal to 1., sum of entries in 4th column equals $\emptyset$. - Recomended criteria: $nP_1, nP_2, \dots, nP_K$ are all ≥ 5. test stat = $T = \frac{d_1^2}{nP_1} + \frac{d_2^2}{nP_2} + \frac{d_3^2}{nP_3} + \dots + \frac{d_K^2}{nP_K}$ low values of T mean there is a low discrepancy between the data and CDF F_θ p-value = right tail prob of value of T under χ^2_{df} - A limitation of χ^2 gof is that the answers can vary depending on how many cells used, too many and it will magnify discrepancy.
  - Uniform, $DU(a,b)$, no parameter estimate, Normal, $N(\mu, \sigma^2)$, 2 parameter estimates $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = s_x^2$. - Poisson, $Poi(\lambda)$, 1 Parameter estimate $\hat{\lambda} = \bar{x}$. - Exponential, $Exp(\lambda)$, 1 Parameter est $\hat{\lambda} = 1/\bar{x}$.
  - problems with the χ^2 gof: the user decides the bins which can change results of the test. You can get varried results by messing with bins. Secondly, we are not using the actual observations, we are instead categorizing them within their relative bins and computing said bins frequency.
- Lilliefors Test for Normality: doesn't depend on the user to select bins & uses the exact values of the observations. This test states that if a dataset is from a normal distribution the the CDF $F(x)$ must be similar to the corresponding normal distributions CDF. $H_0: F = N(\mu, \sigma^2)$ CDF, $H_1: F \neq N(\mu, \sigma^2)$ we test this by looking at the maximum absolute difference between the two CDF's. Table is constructed bellow.

unique values Smallest to largest	$F_N(\hat{\mu}, \hat{\sigma}^2)$ value	$\hat{F}_n$ value	Differences of 1st kind	$\hat{F}_n$ Previous	Differences of 2nd kind
I_1	Given by R	I_1/n	$F_1(\hat{\mu}, \hat{\sigma}^2) - \hat{F}_n$ 1st	$\emptyset$	$F_1(\hat{\mu}, \hat{\sigma}^2) - \hat{F}_1$ Previous
I_2	⋮	I_2/n	$F_2(\hat{\mu}, \hat{\sigma}^2) - \hat{F}_n$ 1st	I_1/n	$F_2(\hat{\mu}, \hat{\sigma}^2) - \hat{F}_2$ previous
I_K	⋮	I_K/n	$F_K(\hat{\mu}, \hat{\sigma}^2) - \hat{F}_K$ 1st	I_2/n	$F_3(\hat{\mu}, \hat{\sigma}^2) - \hat{F}_3$ Previous

* If p-value > α you fail to reject the null hypothesis
$H_0: F = DU(0,3)$ CDF
Ex: for poisson pwilbe provided
F_0 is one of the parameters CDF
df = K − 1 − # of estimated parameters
Parameter estimates
$\hat{F}_n$ previous is F_n shifted down
pmf essque

Test stat

Lilliefors Test

T-Stat = Maximum magnitude of 4th or 5th columns. Ignore sign of value and choose one. A large T-Statistic indicates a strong discrepancy between the data we have and the expected Normal CDF. A small T-stat indicates low discrepancy. A low discrepancy is what we want to confirm H_0. P-value is obtained through the use of the probability table.

KS2

- **KS2 Test**: goal of this test is to check if two different two different populations on the same continous random variable have identical probability distributions. To test if $F_1 = F_2$ we first construct an Empirical pmf (CDF) we then create the table.

Empirical CDF denoted as $\hat{F}_{n_1}$ & $\hat{F}_{n_2}$

Pooled observations (Smallest to largest)	which dataset	Empirical CDF of dataset 1	Empirical CDF of dataset 2	$\hat{F}_{n_1} - \hat{F}_{n_2}$ at pooled values

T-Statistic - largest magnitude ignoring sign in the $\hat{F}_{n_1} - \hat{F}_{n_2}$ column. P-value is calculated on table

- **Sign Test for Median**: - To check if median is smaller than a certain threshold, $H_0: M \geq M_0$, $H_1: M < M_0$, T = # of observations that are at or above M_0, if $n < 20$ then $\text{Bin}(n, 0.5)$ using R, if $n \geq 20$ then $\frac{T + 0.5 - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$. - To check if the median is a certain value, $H_0: M = M_0$, $H_1: M \neq M_0$, If $n < 20$ or $n \geq 20$ then calculate the previous two p-values. Whichever one of the is the min multiply by 2 to get your answer.

- To check if median is larger than a certain threshold then $H_0: M \leq M_0$, $H_1: M > M_0$, T = # of observations that are at or above M_0, If $n < 20$ then Right tail prob of $\text{Bin}(n, 0.5)$, if $n \geq 20$ $\frac{T - 0.5 - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$

- Type of prob & how to compute using Z table:
- Right tail prob with a positive value, compute directly from table. - Left tail prob with a positive value, Find its right tail prob then subtract 1 - said prob. - Left tail prob for a negative value, Find the corresponding positive value and compute directly from table. Right tail probability of a negative value, pretend its not negative and calculate the prob with said prob do 1 - prob to get correct answer.

KS2 for Medians

- **MWW Test**: This test compares two datasets medians, similar to KS2. Firstly, to use this test we must assume f_1 & f_2 are in a "location family". This basically states that the shape of f_1 and f_2 are identical, only their locations are different. Also the population with a smaller n is population one or f_1 and population with a bigger n is f_2.

- To test if first median is smaller than the second median, $H_0: M_1 \geq M_2$, $H_1: M_1 < M_2$ T = sum of ranks for observations for population 2, If at least one sample $n \leq 10$ then Left tail probability of the value of T using table for MWW test. If both $n > 10$ then use a "+" sign on the $\pm$. - To check if the first median is larger than the second median, $H_0: M_1 \leq M_2$ $H_1: M_1 > M_2$, T = same as above, if at least one of the samples $n \leq 10$ then use the MWW table, if both sample $n > 10$ then use a "-" for the $\pm$ of formula.

$$\frac{T \pm 0.5 - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

- To check if the two medians are the same/coincide, $H_0: M_1 = M_2$, $H_1: M_1 \neq M_2$, T = same as above, If one of the samples $n \leq 10$ then find the minimum between the left & right tail probabilities then multiply by 2 by the MWW table. If both samples $n > 10$ the do the same except with the above formula. To find T-stat for above:

Sorted pooled data	Is it from 1st pop?	order of pooled data from 1 to $n_1 + n_2$	Add order # of 1st pop data

The last column's sum will give the T-stat.

Population ≥ 3

omit ranked ties

upgraded MWW Test

small T

If P-val < α Reject H_0

K = # of pop

Tells us if positive negative or no association occurs and tells us how strong it is.

Nature of association

H_0: X is not associated with Y

H_1: X is associated with Y

- **one way ANOVA**: used to check if multiple (on the same random variable) populations have the same median for the probability distribution of said random variable.
$H_0: M_1 = M_2 = \ldots = M_K$ where M_K is median of the distribution of populations.
H_1: At least one of the medians is different from the rest. First we must make an assumption being, that $f_1, f_2, \ldots, f_K$ are all shifted versions of eachother, AKA all populations are members of a location family.

- **KW one way ANOVA**: Table constructed bellow:
| Pooled data sorted small to large | which population it comes from | Rank; 1, 2, 3, ... and so on | →
| Calculation of dataset-specific sum of Ranks, $R_1, R_2, R_3, \ldots, R_K$ where R = sum of Rank collumn for each population. |

Test stat $= \frac{12}{n \cdot (n+1)} \cdot \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ldots + \frac{R_K^2}{n_K}\right) - (3 \cdot (n+1))$ where large T supports H_1, small T lacks support for H_1.

P-value: if each of K datasets has more than 5 observations, then P-value can be approximated P-value = RT probability of T under the chi squared distribution, df = K − 1.

- **Multivariate datasets**: stats with datasets containing more than one variable. Univariate dataset: stats pertaining to datsets on only one variable. Bivariate dataset: ..... two variables.
- **Nonparametric measures of association**: Dependence/Independence between two variables. AKA if one variable increases should we expect an increase/decrease/no change in the other variable.
- **Nonparametric regression**: Prediction, If we have the value for one of the two variables can we predict the value of the other variable for the same observation with reasonable accuracy.
- **Measuring association**: we cannot use the correlation coefficent r, because it can only calculate linear associations which cannot tell us if X & Y are independent. Thus we have to assume a normal parametric model. ↓

- **Spearman's Rank correlation coefficient r_s**: Assume no ties, AKA repititions in X and Y values.
| Sortered pairs (based on X values from smallest to largest | Rank from 1, 2, 3, ..., n^Y for Y values |
| Rank from 1, 2, 3, ..., n^X for X values (allready sorted) | Difference between Ranks $Y_{ranks} - X_{ranks}$ for each row |
Accuracy check, sum of last column must = 0. The final colum is known as $D_1, D_2, D_3, \ldots, D_n$

$r_s = 1 - \frac{6 \cdot (D_1^2 + D_2^2 + \ldots + D_n^2)}{n \cdot (n^2 - 1)}$ (r_s can never be higher than 1 and lower than −1.)

Interpretation of r_s: 0 = independence, $\pm(0, 0.3)$ = weak, $\pm(0.3, 0.7)$ = moderate, $\pm(0.7, 1)$ = strong, 1 = perfect

- **hypothesis testing**: To check if X & Y have association. $H_0: P \geq 0$ $H_1: P < 0$, $T = r_s$, if $n \leq 30$ then LT probability using Spearmens rank table. If $n > 30$ then LT probability of $T \cdot \sqrt{n-1}$ using z table. To check if X & Y have a positive association. $H_0: P \leq 0$ $H_1: P > 0$, $T = r_s$. If $n \leq 30$ then RT prob of spearmens rank table, If $n > 30$ the RT prob of $T \cdot \sqrt{n-1}$ using z table. To check if X & Y are associated or independent $H_0: P = 0$ $H_1: P \neq 0$, $T = r_s$, If $n \leq 30$ 2 · the min value between left & right tail probs. If $n > 30$ then 2 · min of RT & LT prob of $T \cdot \sqrt{n-1}$ using z table.

- **Regression**: The use of one variable to predict the other, X is the covariate, Y is the response, reffered to as the regression of Y on X.
- **Simple linear Regression: RT Line**: $y = a + bX + e$ where a is the intercept, b is the slope, and e is the error in the regression equation. The median of the distribution always has $e = 0$, thus median of $Y = a + bX$. This makes the line linear. Thus, we must assume median of $Y = a + bX$ to use this test. To find b AKA the slope:

Pairs sorted by X values	(x_1, y_1)	(x_2, y_2)	(x_3, y_3)	...	(x_n, y_n)
(x_1, y_1)		$\frac{y_2 - y_1}{x_2 - x_1}$	$\frac{y_3 - y_1}{x_3 - x_1}$		$\frac{y_n - y_1}{x_n - x_1}$
(x_2, y_2)			$\frac{y_3 - y_2}{x_3 - x_2}$		$\frac{y_n - y_2}{x_n - x_2}$
(x_3, y_3)					$\frac{y_n - y_3}{x_n - x_3}$
⋮					$\frac{y_n - y_{\ldots}}{x_n - x_{\ldots}}$
(x_n, y_n)					

b turns to b now

↓ Find $\tilde{b}$

- find the middle most value, if b is odd. If b is even find the 2 most middle and avrg.

Estimation of a, the intercept:

$a_1 = y_1 - \hat{b}x_1$ | $a_2 = y_2 - \hat{b}x_2$ | $a_3 = y_3 - \hat{b}x_3$ | ..... | $a_n = y_n - \hat{b}x_n$

Similarly, we then find the median of the collection, a turns to $\hat{a}$

→ we now are able to predict the median of the distribution y_{new}

TOH for simple linear regression KT test! Testing if or if not the linear regression is meaningful.

$H_0: b = 0$ $H_1: b \neq 0$, $T = (\#\text{ of positive pairwise slopes}) - (\#\text{ of negative pairwise slopes})$.

If sample size $n \leq 30$ then $2 \cdot \min(\text{LT prob of } T, \text{RT prob of } T)$ using the KT table. If $n > 30$ first compute: $\frac{T \cdot \sqrt{18}}{\sqrt{n \cdot (n-1) \cdot (2n+5)}}$ then $2 \cdot \min(\text{LT prob, RT prob})$ using standard Z score table.

* interpretation of $\hat{b}$: if x increases by 1 unit then median y is estimated to increase/decrease by $\hat{b}$.