Lecture 15: Nonparametric Stats and Distribution Comparisons

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2018

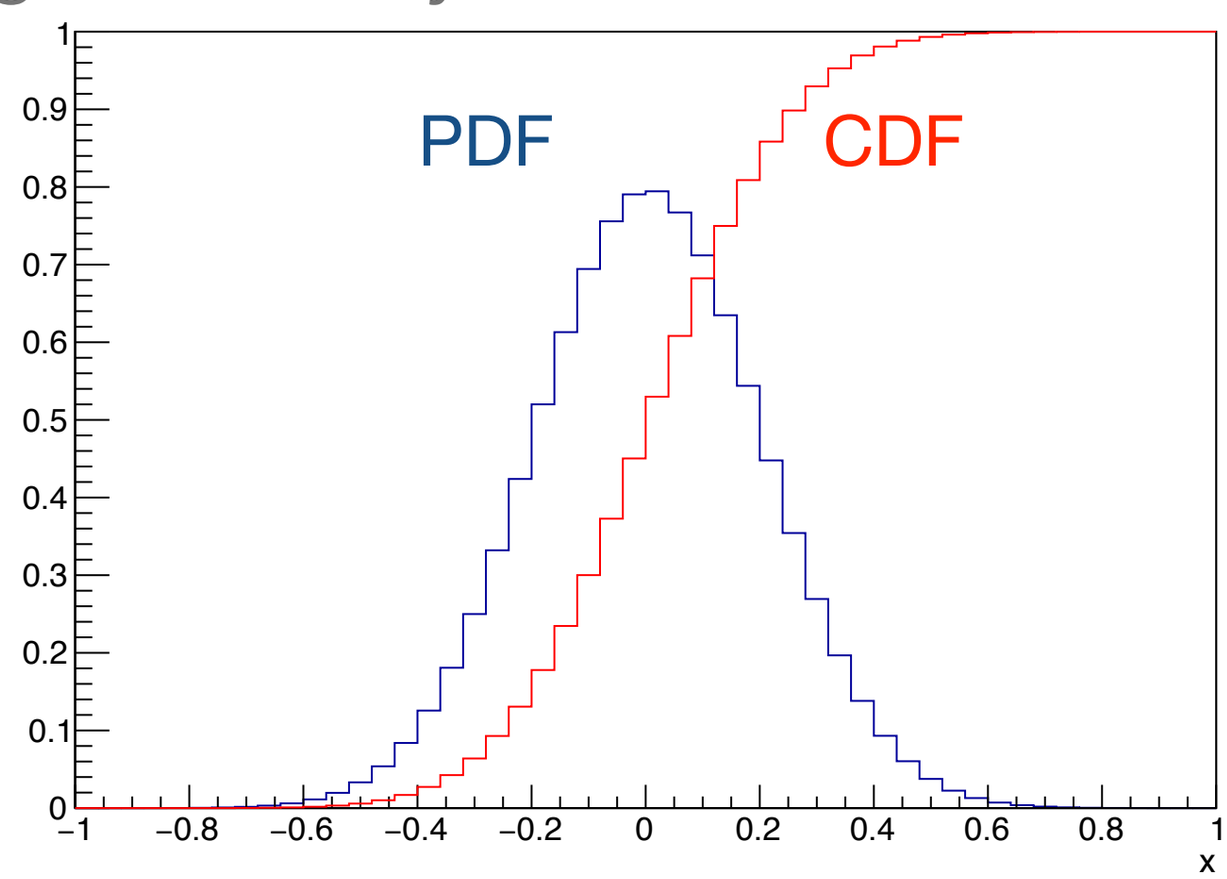
Photo by Howard Jackman

Comments

- I include these slides to be reviewed at your leisure. They are straightforward and will not **explicitly** be on the exam, but they may be useful in your research or as part of exam solutions
  - By explicit, I mean that these non-parametric tests will not be directly part of a question, e.g. “Use a Mann-Whitney test to compare distributions A and B.” will not be on the exam
  - Even so, the Kolmogorov-Smirnov test **is** something we covered briefly **and** can be used as a goodness-of-fit test statistic

Cumulative

- For smooth and parametric scenarios, i.e. those with an explicit form including parameters, there is an underlying cumulative distribution function (CDF)
- Jason: Do an example here of a Gaussian curve and the underlying CDF. Okay.



Empirical Distribution Function

- What happens when the underlying CDF is unknown or non-parametric?
- Using the data which all come from a common CDF $F_n(t)$ we can produce the empirical distribution function (EDF) $\hat{F}_n$

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i}^{n} \mathbb{1}_{x_i \leq t}$$

$$\hat{F}_n(t) = \frac{\text{number of elements } \leq t}{n}$$

Indicator Function,
which is sometimes shown as capital i

$$\mathbb{1}_{x_i} \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{if } x_i \notin A \end{cases}$$

Examples

- Jason: Have them use the CDF from earlier and their random number generator to create a data set for constructing the EDF. Okay.
- Exercise #0.5
  - Use a random number generator and sample from the underlying PDF to generate an empirical distribution function, and compare to the CDF function used for the PDF, i.e. compare the EDF constructed from the discrete sampling to the CDF which is smooth and analytic.

Discussion

- Unbinned or Histogrammed construction of EDF?
  - Nominally unbinned, because of the loss of information
  - Sometimes data arrives histogrammed and thankfully some tests (Kolmogorov-Smirnov, etc.) have forms which can incorporate binned data
  - KS-test **should** be unbinned
  - For KS-test “We can apply a useful rule: As long as the bin width is small compared with any significant physical effect (for example the experimental resolution) then the binning cannot have an important effect” - Jan Conrad & Fred James

Hypothesis Testing for KS-test

- With a specific model, commonly the null-hypothesis H_0 or F_0, we can test the max divergence with data through the EDF with the expectation from the model (or another EDF, which we'll do later)
- Math bits: The Kolmogorov-Smirnov statistic is the supremum of the point-wise EDF ($F_n(x)$) with the model CDF (F(x))

$$D_n = \sup_x |F_n(x) - F(x)|$$

- Note that the KS-test is shape-dependent. It is mostly insensitive to any normalization differences

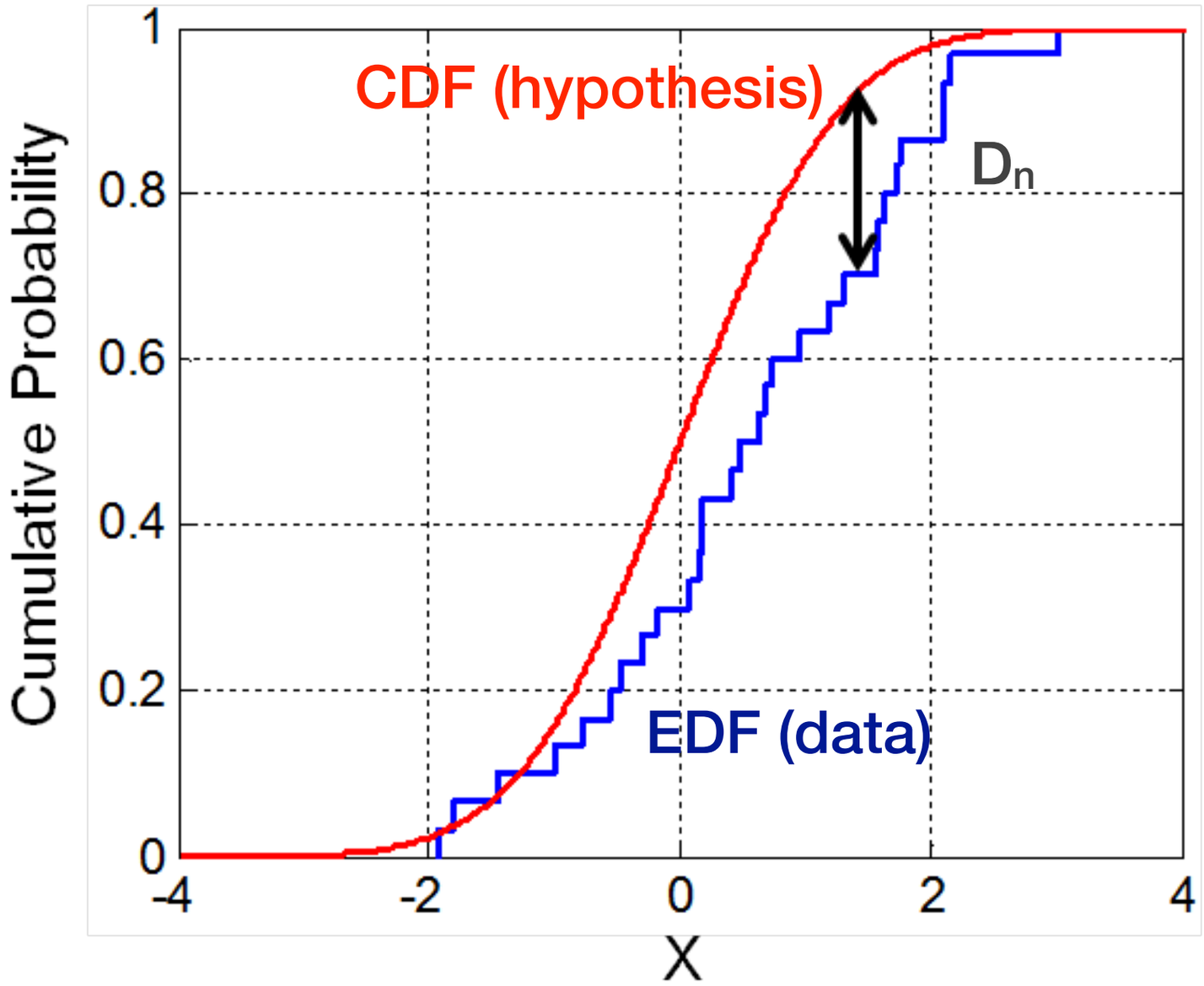
KS-Test features

- Model being tested (and parameters) should not be drawn from the data set to which the model is being compared
- If the value of the KS statistic is out in the tails, be wary, you are dealing in low-statistics and low-likelihood regimes
  - Thankfully this suggests that the two distributions are similar
  - But, actual differences in tails of distributions are unlikely to be identified by the KS-test
- Only valid for continuous distributions
- "The distribution of the KS statistic is also not distribution-free when the dataset has two or more dimensions" -Babu & Feigelson
- Works better **without** binning

Graphical KS-test

*wikipedia

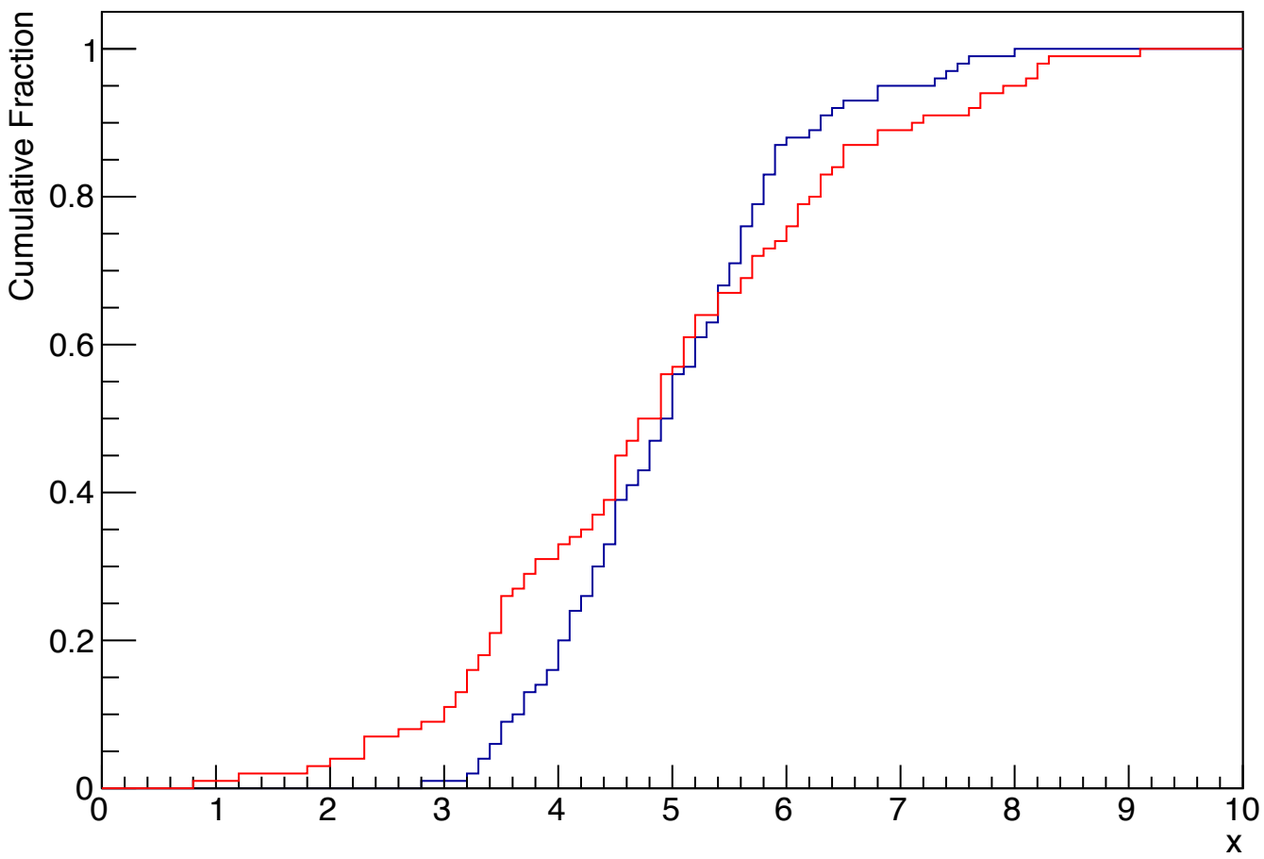
- Compare the supremum, i.e. largest difference, for the two cumulative distributions



Note that both data sets can be EDFs, there is no strict requirement that both sets cannot be actual data, or sampled sets (e.g. finite statistics Monte Carlo)

Running KS-Test

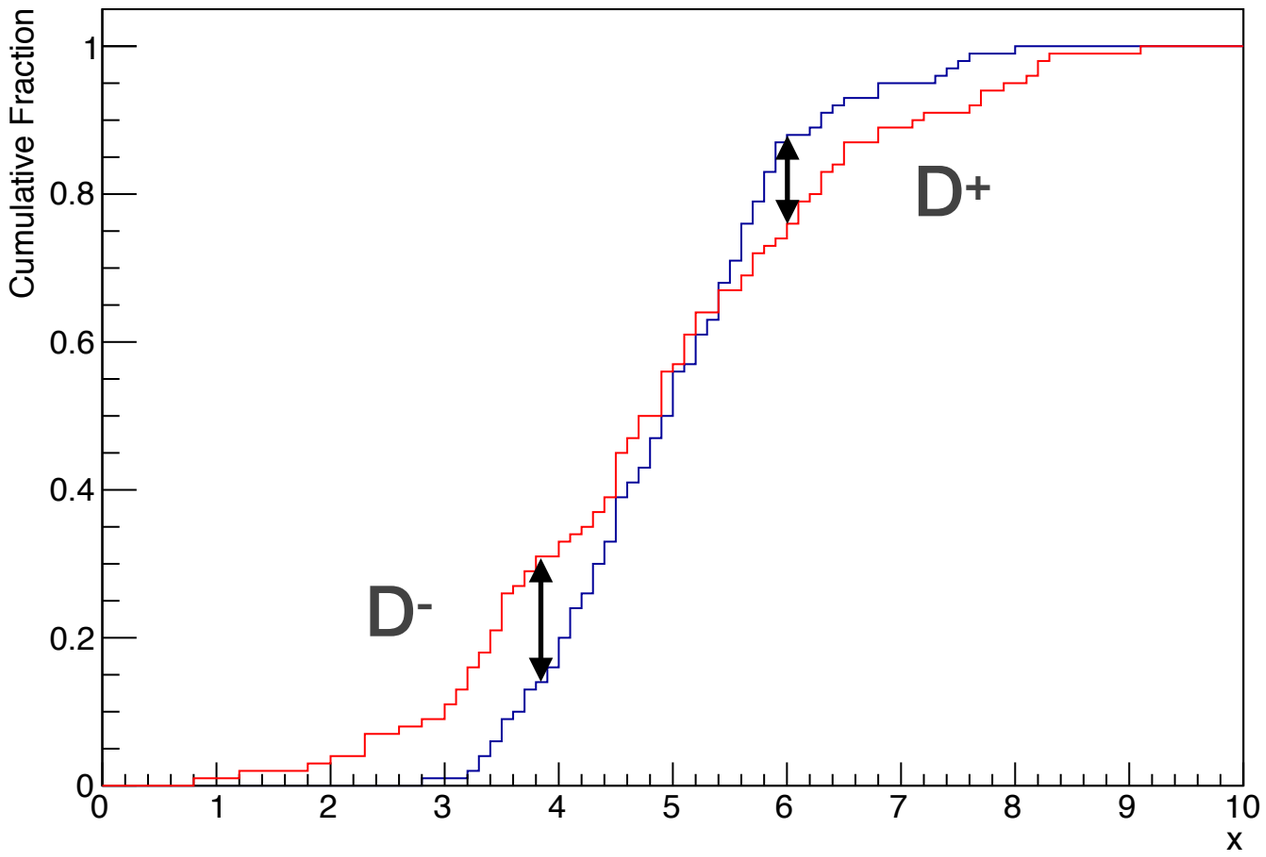
- Using a combination of the number of data points and the largest difference between the cumulative distributions, it is possible to calculate a p-value, or better yet, use computers to do it



Here, the comparison is between two data samples, i.e. two EDFs

One/Two Sided Test

- Most often we want to know about the greatest difference between the two distributions/samples, regardless of the sign (+/-) of the deviation. This is a two-sided or two-tailed test.
- A one-sided test is where we want to know about deviations in only a single direction, i.e. + or - deviations.



Exercise #1

- For 1000 random samples from two gaussians w/ a mean of 5 for both and widths of 1.2 and 1.6, respectively calculate the two-sided p-value for the agreement between the two samples. FYI, it should be low.
- Strangely, using the unbinned ROOT function and scipy.stats.ks_2samp(), the answers I get are slightly different
  - Both are in the large N regime, i.e. enough data
  - Both (I think) are ‘two-sided’ tests

```
ROOT:
 Kolmogorov Probability = 9.07999e-05, Max Dist = 0.1
scipy 2sample:
Ks_2sampResult(statistic=0.10000000000000009, pvalue=8.1175573157381319e-05)
```

What about the Tails?

- If you know or hypothesize the expected **form** of the underlying sample (e.g. gamma, exponential, gaussian, etc.) the Anderson-Darling test is designed to be more sensitive to deviations in the tails of distributions, by dividing by the distribution (not a typo) via:

$$\int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \frac{1}{F(x)(1-F(x))} dF(x)$$

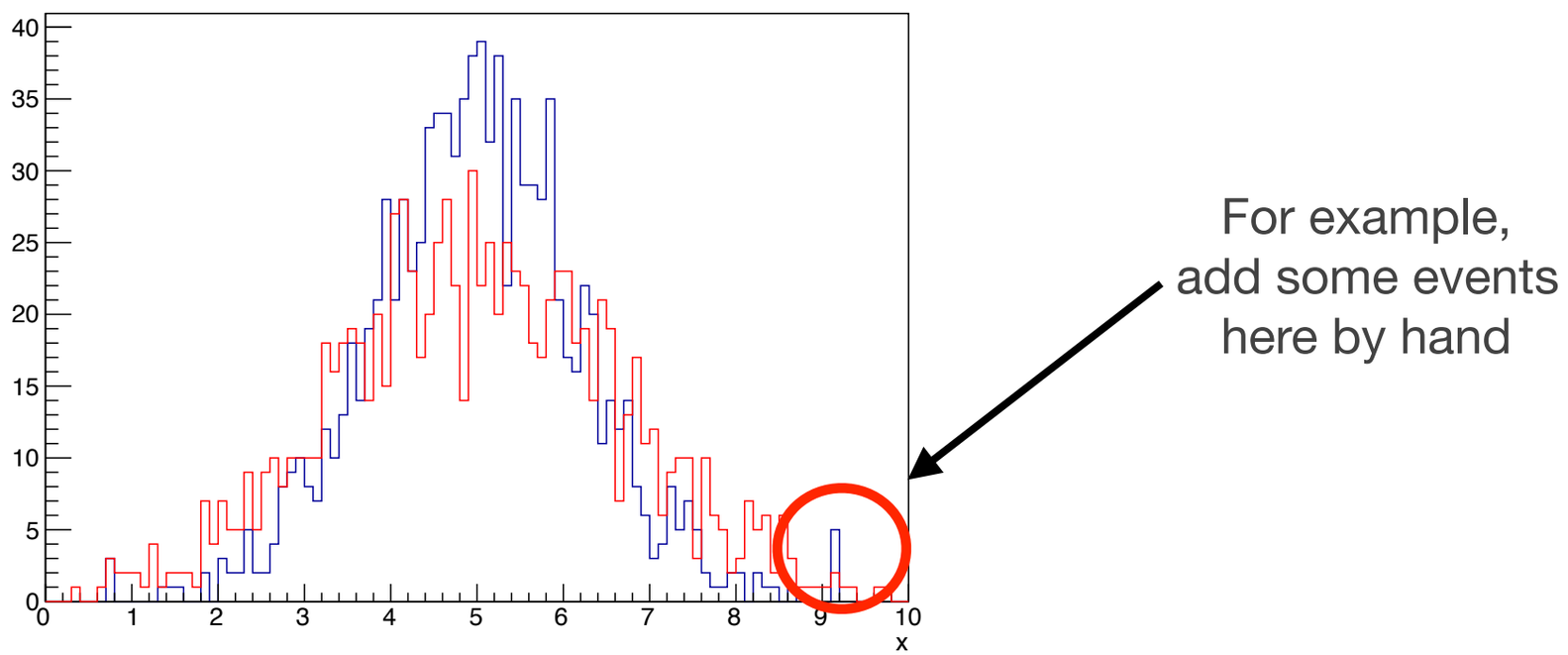
$$A^2 = \frac{[\sum_{i=1}^{n} (2i-1)(\ln(u_i) + \ln(1-u_{n+1-i}))]}{n} - n \qquad u_i = F(x_i)$$

F_n is the EDF (sample)
F is the hypothesis function

- The Anderson-Darling statistic (A) is then compared to critical values for a desired significance (**α**) specific to a distribution type (gaussian, etc.)

Exercise #2

- Start playing with KS-Test and Anderson-Darling
  - Try gaussian distributions, beta, exponential, etc.
  - Try one-sided KS versus two-sided
- What happens when ‘outliers’ are introduced to the samples?
  - Generate some pseudo-random data points from known distributions, and add in some additional data in the tails



Two-Sample Tests

- So with the idea of a robust non-parametric comparison between two independent samples there are some nice options
- Mann-Whitney-Wilcoxon a.k.a. Mann-Whitney U test

Calculating "Ranks"

- Instead of making comparisons between set X^a and X^b based on the underlying distributions as a function of a single variable, look at their ordering and throw some math at their 'ranks'
- Ranks: The ordering of a data set

$$X_i^a = (1, 4, 5, 5, 6, 9, 10) \qquad n_a = 7$$
$$R(X_i^a) = (1, 2, 3.5, 3.5, 5, 6, 7)$$
$$X_j^b = (1.1, 3, 4, 4, 6, 9, 9.8, 12) \qquad n_b = 8$$

Mann-Whitney U test

- Also known as the Mann-Whitney-Wilcoxon test we calculate the sum of the ranks of X^a in a merged set of X^a and X^b
  - Has some resistance to distortion from outliers
  - U_a is then compared to the critical value to get out the hypothesis test

$$X_i^a = (1, 4, 5, 5, 6, 9, 10) \qquad n_a = 7$$
$$X_j^b = (1.1, 3, 4, 4, 6, 9, 9.8, 12) \qquad n_b = 8$$

$$U_a = R_a - \frac{n_a(n_a + 1)}{2}$$

R_a is the sum of the ranks of X^a in the merged set X^a and X^b

Mann-Whitney U test

- Also known as the Mann-Whitney-Wilcoxon test we calculate the sum of the ranks of X^a in a merged set of X^a and X^b

$$X_i^a = (1, 4, 5, 5, 6, 9, 10)$$
$$X_j^b = (1.1, 3, 4, 4, 6, 9, 9.8, 12)$$
$$X^{(a+b)} = (1, 1.1, 3, 4, 4, 4, 5, 5, 6, 6, 9, 9, 9.8, 10, 12)$$
$$R(X^{(a+b)}) = (1, 2, 3, 5, 5, 5, 7.5, 7.5, 9.5, 9.5, 11.5, 11.5, 13, 14, 15)$$
$$R_a(X^{(a+b)}) = (1, 5, 7.5, ...)$$

*notation is a bit sloppy/confusing, but should be illustrative

Mann-Whitney U test

- For sample sizes >~ 20, we can start to use the 'normal' approximation, where

$$m_U = \frac{n_a n_b}{2}$$

m_u is the mean of the U statistic

$$\sigma_U = \sqrt{\frac{n_a n_b (n_a + n_b + 1)}{12}}$$

- Which assumes no ties, but can be used to produce values related to significance
- What is the uncertainty for data sets w/ ties?

Exercise #3

- Using the values on slide 17 and 18, calculate the value of the Mann-Whitney U statistics and p-values.
- Using the same sample what happens when you only take the first 1/2 of the values and recalculate
  - You can decide what to do with the sample that has 7 events, i.e. take 3 or 4.

Journal Article Reading

- Statistical Tools for Classifying Galaxy Group Dynamics
  http://arxiv.org/abs/0908.0938

Resources

- Wikipedia
  - https://en.wikipedia.org/wiki/Kolmogorov–Smirnov_test
  - https://en.wikipedia.org/wiki/Empirical_distribution_function
  - https://en.wikipedia.org/wiki/Anderson–Darling_test