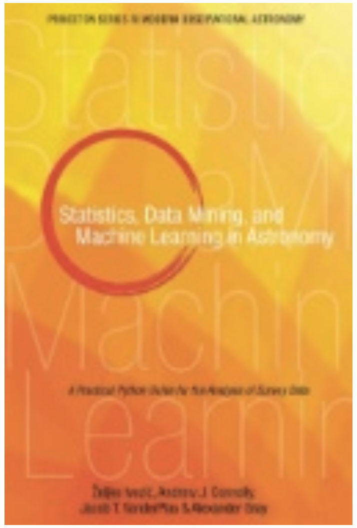
MAXIMUM LIKELIHOOD ESTIMATIONS II: HYPOTHESIS TESTING, MODEL COMPARISON, NON-PARAMETRIC ANALYSIS (SECTIONS 4.6 - 4.8)



a.k.a. ICVG

Adele Plunkett
MCMC Coffee @ ESO
Season 1, episode 4
6 October 2016

Outline of this session

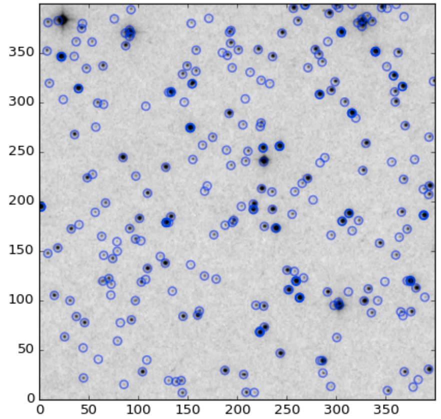
- 4.6 Hypothesis testing
  - P-value
  - Simple classification
  - Some hypothetical cases

- 4.7 Comparison of distributions
  - Regression towards the mean

- 4.8 Non-parametric analysis
  - Parametric vs. non-parametric
  - Examples
  - Histograms

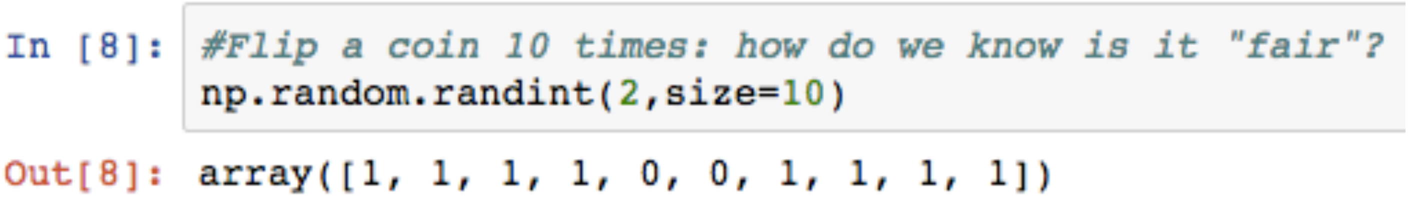
4.6 Hypothesis testing

- Statistics language: *"Whether a value x_i, or the whole set $\{x_i\}$, is consistent with being drawn from a Gaussian distribution $N(\mu,\sigma)$."* (ICVG p. 144)
- Astronomy example: Source detection or background noise?
- Astronomy example in statistics language: *"Here, the null hypothesis is that the measured brightness in a given resolution element is due to background, and when we can reject it, we have a source detection."* (ICVG p. 144)

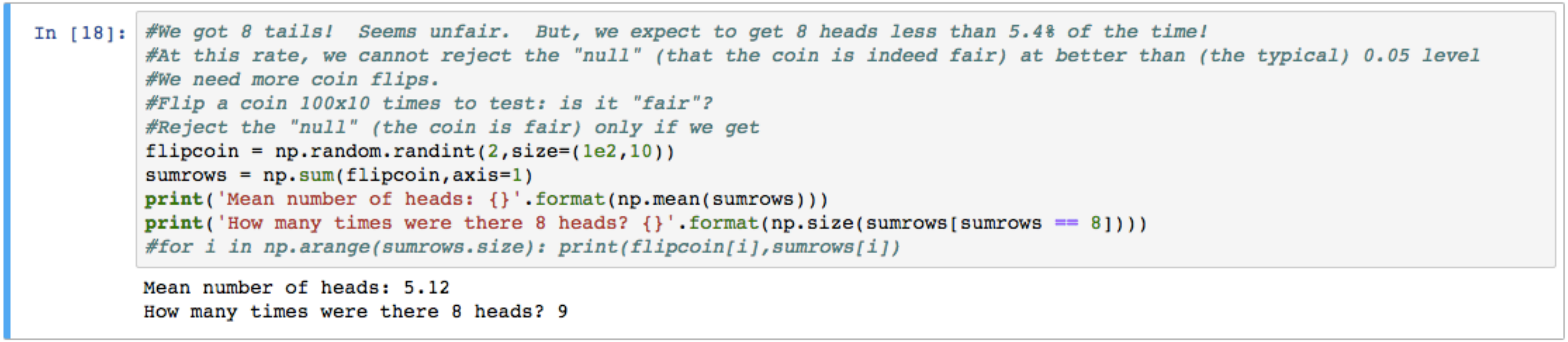


P-value

- Statisticians warning: If you “fail” to reject the hypothesis (i.e. you are not sure of your detection), it does not mean that we prove its correctness. Maybe the sample is just not large enough.
- Example:



```
In [8]:  #Flip a coin 10 times: how do we know is it "fair"?
         np.random.randint(2,size=10)

Out[8]:  array([1, 1, 1, 1, 0, 0, 1, 1, 1, 1])
```



In [18]:

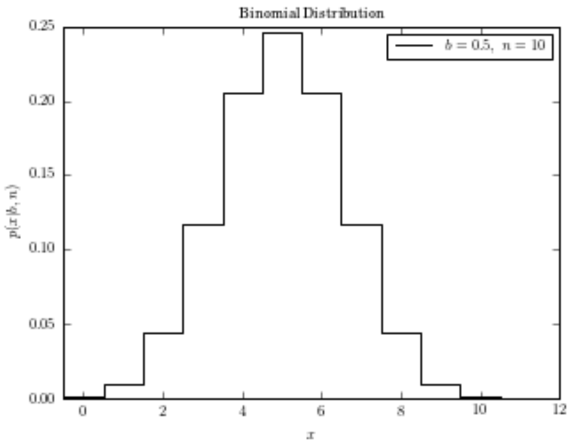
```python
#We got 8 tails!  Seems unfair.  But, we expect to get 8 heads less than 5.4% of the time!
#At this rate, we cannot reject the "null" (that the coin is indeed fair) at better than (the typical) 0.05 level
#We need more coin flips.
#Flip a coin 100x10 times to test: is it "fair"?
#Reject the "null" (the coin is fair) only if we get
flipcoin = np.random.randint(2,size=(1e2,10))
sumrows = np.sum(flipcoin,axis=1)
print('Mean number of heads: {}'.format(np.mean(sumrows)))
print('How many times were there 8 heads? {}'.format(np.size(sumrows[sumrows == 8])))
#for i in np.arange(sumrows.size): print(flipcoin[i],sumrows[i])
```

```
Mean number of heads: 5.12
How many times were there 8 heads? 9
```

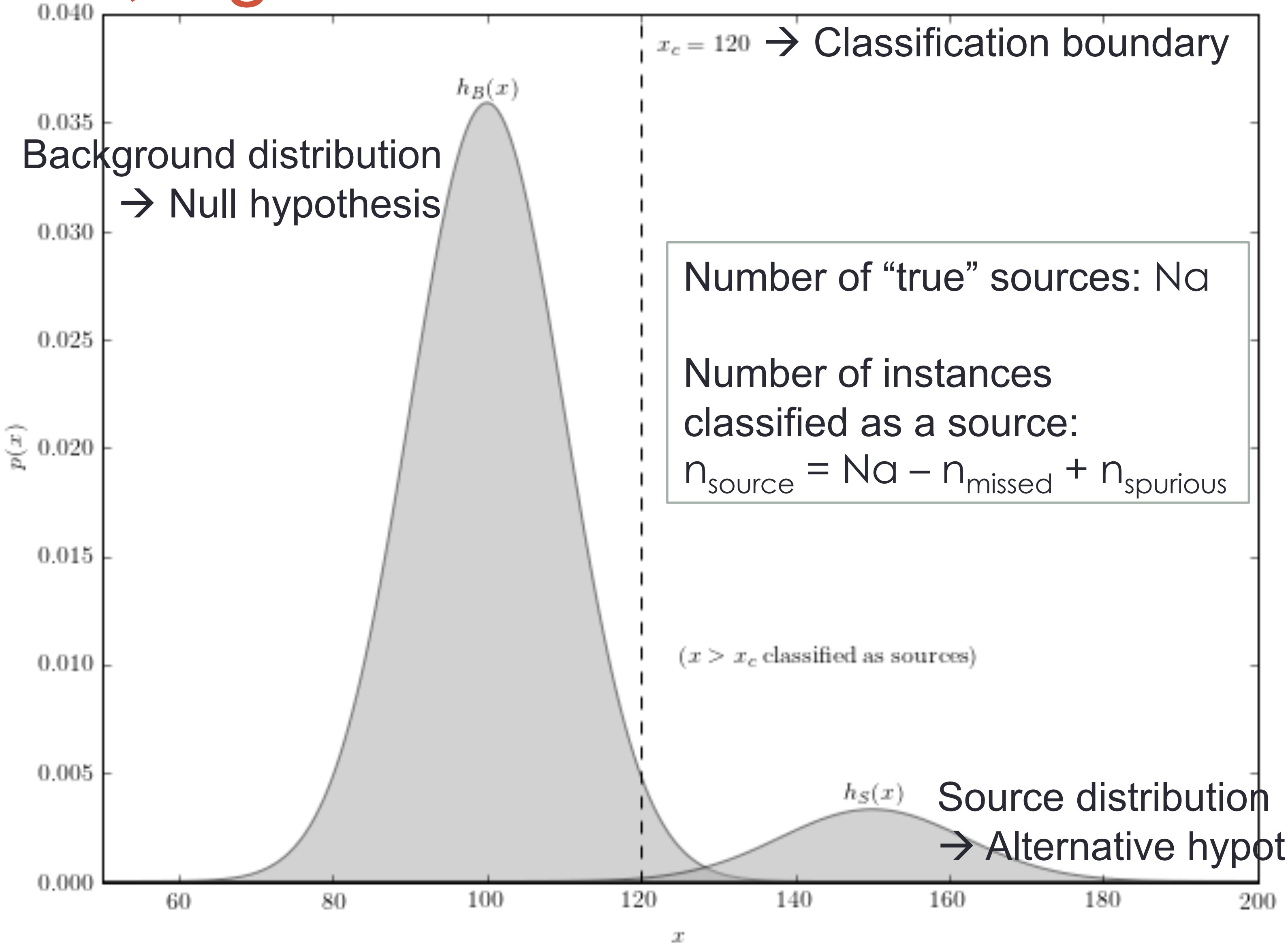
```
In [18]: #We got 8 tails!  Seems unfair.  But, we expect to get 8 heads less than 5.4% of the time!
         #At this rate, we cannot reject the "null" (that the coin is indeed fair) at better than (the typical) 0.05 level
         #We need more coin flips.
         #Flip a coin 100x10 times to test: is it "fair"?
         #Reject the "null" (the coin is fair) only if we get
         flipcoin = np.random.randint(2,size=(1e2,10))
         sumrows = np.sum(flipcoin,axis=1)
         print('Mean number of heads: {}'.format(np.mean(sumrows)))
         print('How many times were there 8 heads? {}'.format(np.size(sumrows[sumrows == 8])))
         #for i in np.arange(sumrows.size): print(flipcoin[i],sumrows[i])
```

```
In [17]: from scipy import stats
         dist = stats.binom(10,0.5) #N=10,b=0.5 (e.g. coin is fair)
         r = dist.rvs(10) #the outcome of 10 random flips
         p = dist.cdf(7) #probability to get >7 heads, k=8,9,or10 successes
         print('Probability to get >=8 heads: {}'.format(1.-p))
         #PLOT
         x = np.arange(-1, 200)
         fig, ax = plt.subplots(figsize=(5, 3.75))
         plt.plot(x, dist.pmf(x), ls='-', c='black', label=r'$b=%.1f,\ n=%i$' %(0.5,10), linestyle='steps-mid')

         plt.xlim(-0.5, 12)
         plt.ylim(0, 0.25)

         plt.xlabel('$x$')
         plt.ylabel(r'$p(x|b, n)$')
         plt.title('Binomial Distribution')

         plt.legend()
         plt.show()
```

```
Probability to get >=8 heads: 0.0546875
```



Simple Classification
ICVG, Figure 4.5



→ Classification boundary

Background distribution
→ Null hypothesis

Number of “true” sources: Nα

Number of instances classified as a source:
n_{source} = Nα – n_{missed} + $n_{spurious}$

Source distribution
→ Alternative hypothesis

A few “hypothetical” cases (ICVG)

- Care about false negatives:
  - If null hypothesis is “this undergrad student would do great in grad school”
  - else if we reject a good student (false negative), ☹
  - else if we accept a bad student (false positive), no big deal 😐
- Care about false positives:
  - If null hypothesis is “my parachute is good”
  - else if it’s bad, but we accept it as good (false positive) == disaster ☹
  - else if it’s good, but we reject it as bad, fine 😐

4.7 Comparison of distributions

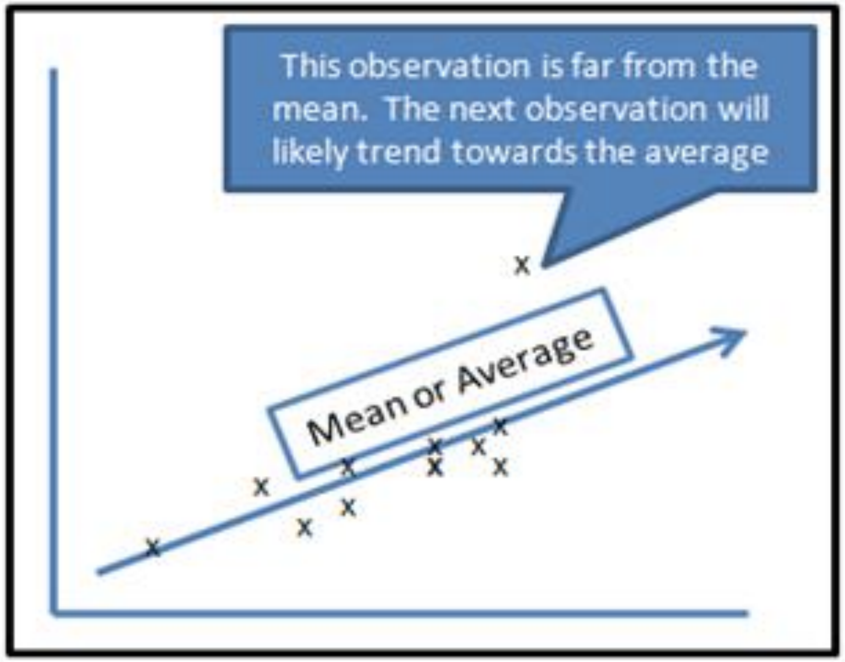
- Statistics language:
  - *"…whether two samples are drawn from the same distribution…"*
  - *"…whether two sets of measurements imply a difference in the measured quantity."*
  - *"…whether a sample is consistent with being drawn from some known distribution."* *(ICVG p. 149)*
- Astronomy example: Measure the mass of the same planet/galaxy/whatever using two methods (with different measurement errors). Do the measured masses agree?

Regression toward the mean

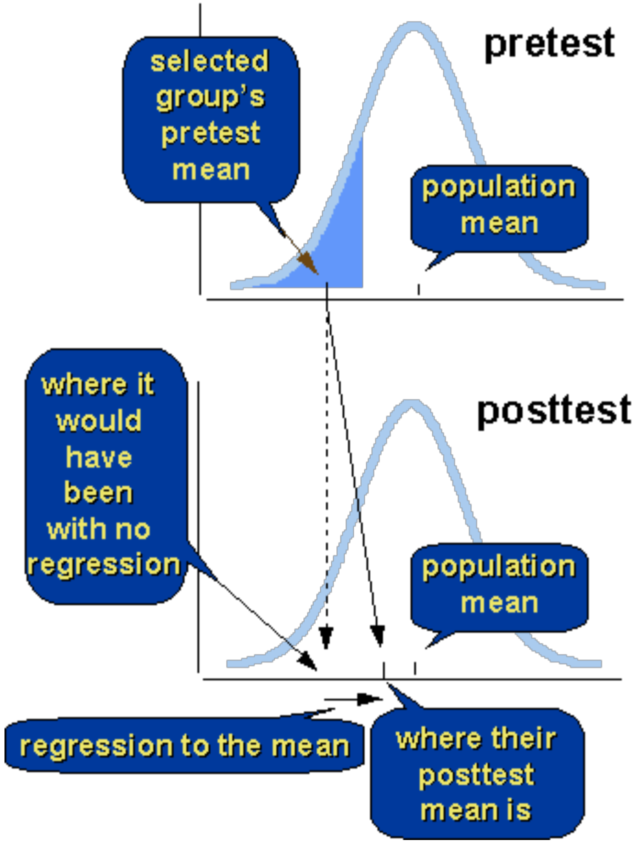
The psychologist Daniel Kahneman, winner of the 2002 Nobel prize in economics, pointed out that regression to the mean might explain why rebukes can seem to improve performance, while praise seems to backfire. (Wikipedia)

> “ I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning. When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand … He said, “On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don’t tell us that reinforcement works and punishment does not, because the opposite is the case.” This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them. ”

Regression toward the mean

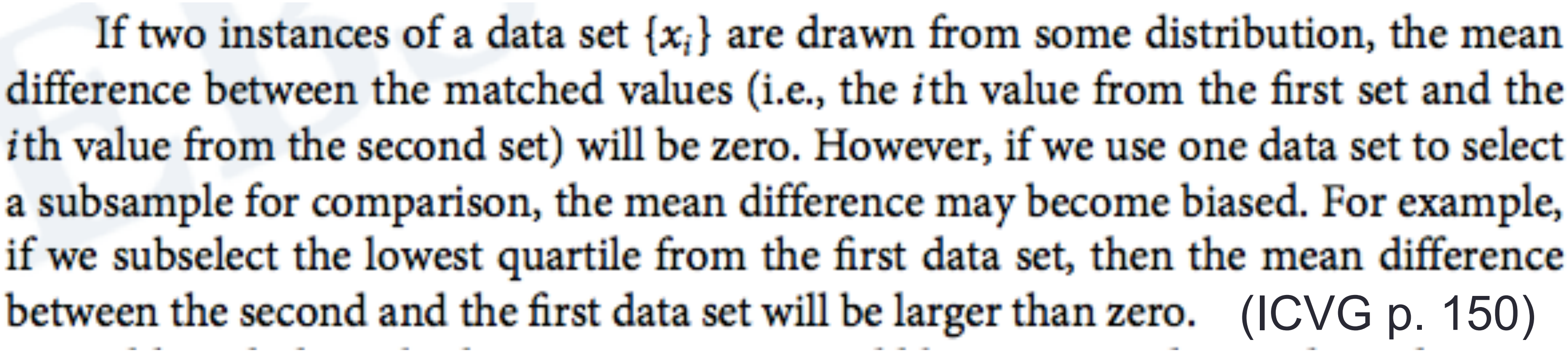

This observation is far from the mean. The next observation will likely trend towards the average
x
Mean or Average

http://www.americanthinker.com/articles/2009/08/the_200809_housing_crisis_and.html


selected group's pretest mean
pretest
population mean
where it would have been with no regression
posttest
population mean
regression to the mean
where their posttest mean is

http://www.socialresearchmethods.net/kb/regrmean.php

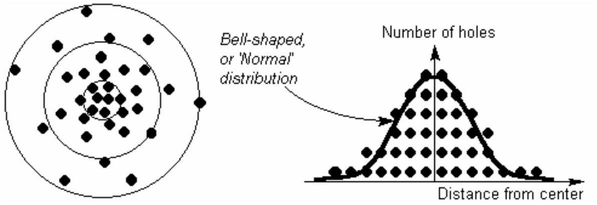
Regression toward the mean

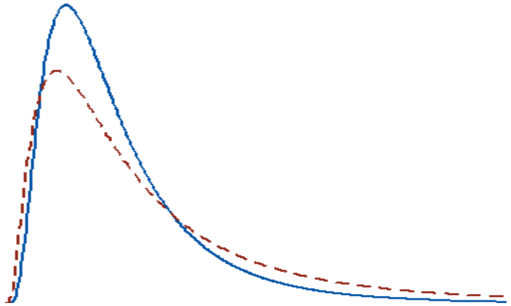


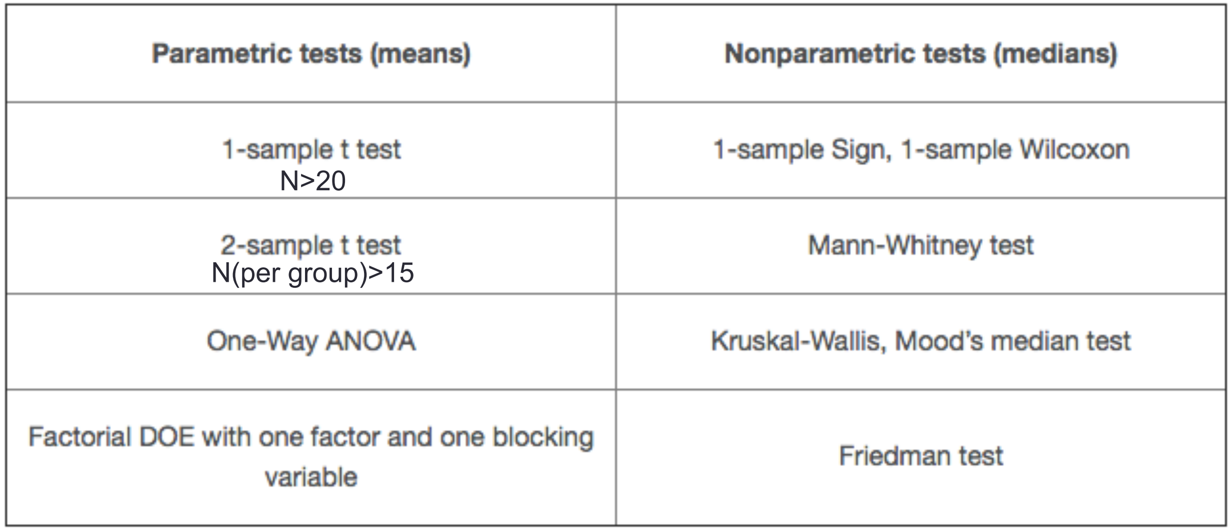
If two instances of a data set $\{x_i\}$ are drawn from some distribution, the mean difference between the matched values (i.e., the ith value from the first set and the ith value from the second set) will be zero. However, if we use one data set to select a subsample for comparison, the mean difference may become biased. For example, if we subselect the lowest quartile from the first data set, then the mean difference between the second and the first data set will be larger than zero. (ICVG p. 150)

- Examples:
  - Sir Francis Galton, 19th Century: “offspring of parents who lie at the tails of the distribution will tend to lie closer to the centre, the mean, of the distribution” (wikipedia)
  - Measuring the impact after efforts to improve students’ test scores (ICVG)
  - Financial time series “returns can be very unstable in the short run but very stable in the long run” (wikipedia)
  - Astronomical site testing (ICVG)
- Takeaway: When comparing models, make a randomly selected sample to detect a possible difference.

4.8 Non-parametric analysis







Parametric tests (means)	Nonparametric tests (medians)
1-sample t test N>20	1-sample Sign, 1-sample Wilcoxon
2-sample t test N(per group)>15	Mann-Whitney test
One-Way ANOVA	Kruskal-Wallis, Mood's median test
Factorial DOE with one factor and one blocking variable	Friedman test

Reasons to Use Nonparametric Tests

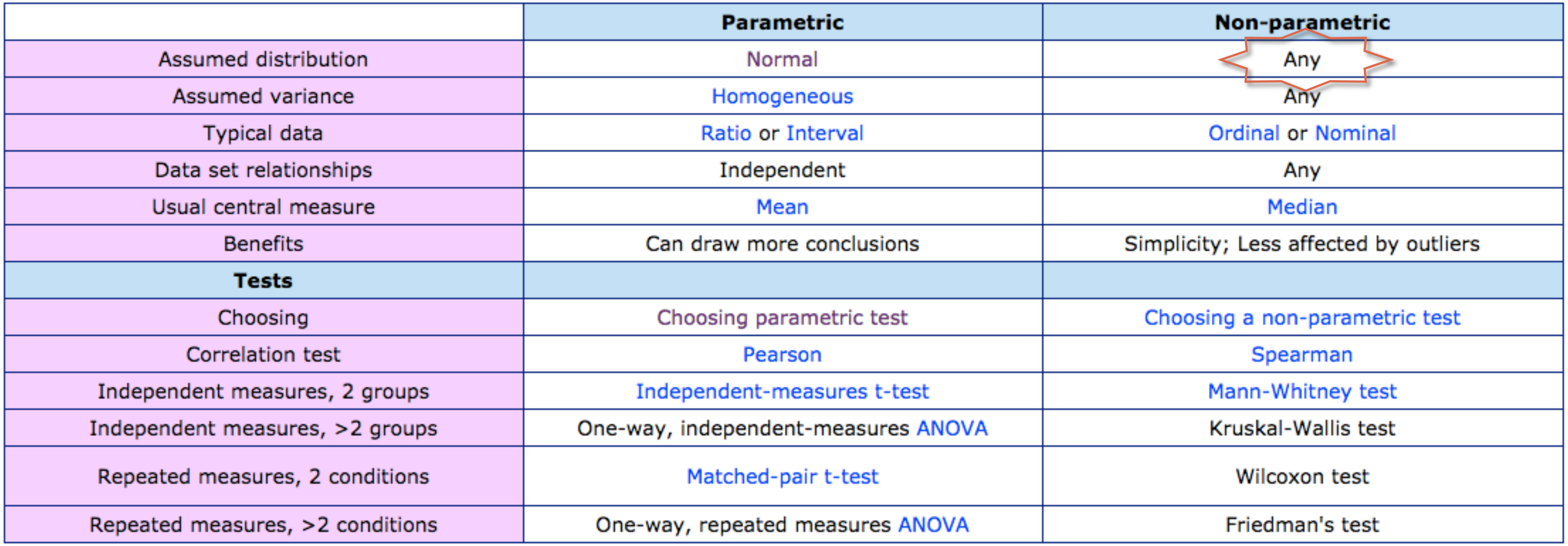
Reason 1: Your area of study is better represented by the median (example: income with some billionaire outliers)

Reason 2: You have a small sample size (but parametric has more “statistical power”)

Reason 3: You have ordinal non-continuous data, ranked data, or outliers that you can’t remove

http://blog.minitab.com/blog/adventures-in-statistics/choosing-between-a-nonparametric-test-and-a-parametric-test

Parametric vs. non-parametric



	Parametric	**Non-parametric**
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Typical data	Ratio or Interval	Ordinal or Nominal
Data set relationships	Independent	Any
Usual central measure	Mean	Median
Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Choosing	Choosing parametric test	Choosing a non-parametric test
Correlation test	Pearson	Spearman
Independent measures, 2 groups	Independent-measures t-test	Mann-Whitney test
Independent measures, >2 groups	One-way, independent-measures ANOVA	Kruskal-Wallis test
Repeated measures, 2 conditions	Matched-pair t-test	Wilcoxon test
Repeated measures, >2 conditions	One-way, repeated measures ANOVA	Friedman's test

http://changingminds.org/explanations/research/analysis/parametric_non-parametric.htm

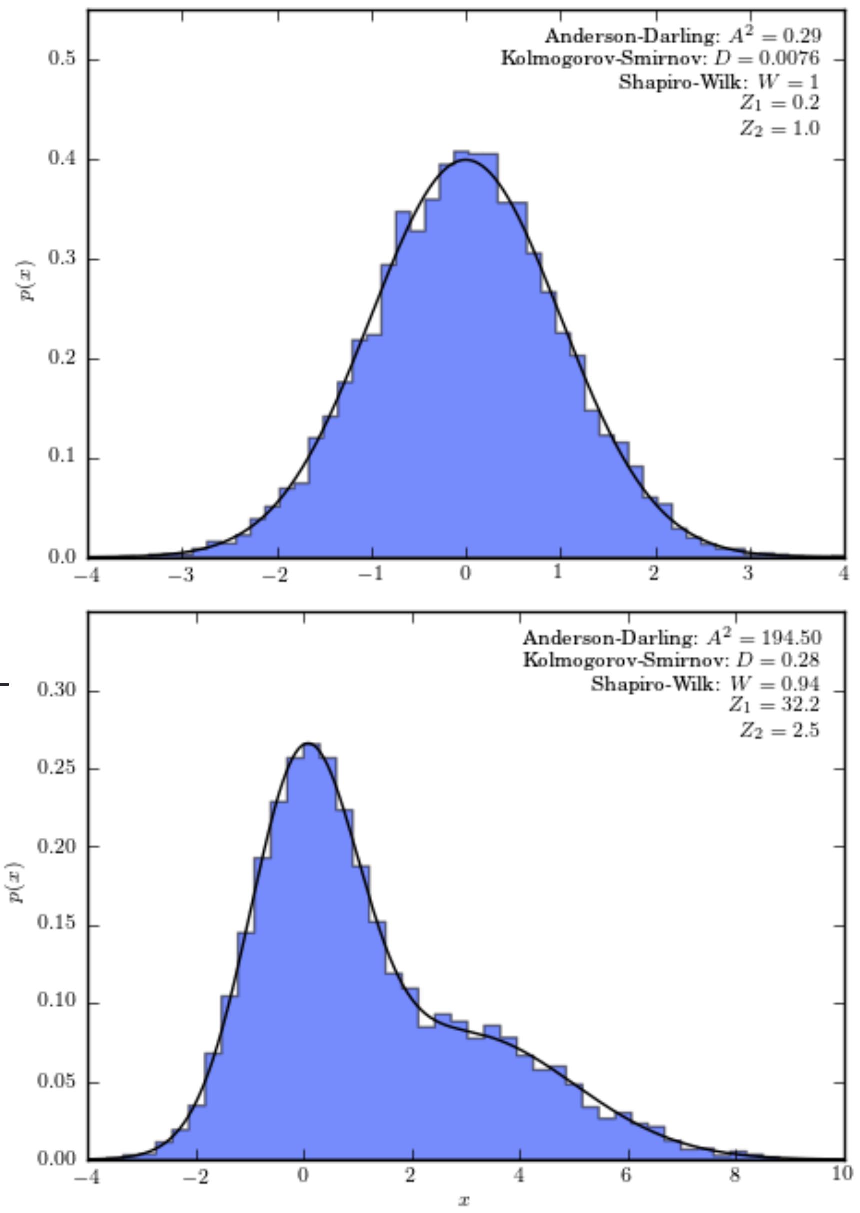
Now a few examples (from ICVG 4.7)

- Anderson-Darling
- Kolmogorov-Smirnov
- Shapiro-Wilk
- U test
- Wilcoxon test
- F-test

ICVG FIGURE 4.7

```
Kolmogorov-Smirnov test: D = 0.0076  p = 0.6
  Anderson-Darling test: A^2 = 0.29
    significance  | critical value
    --------------|----------------
    0.58          | 15.0%
    0.66          | 10.0%
    0.79          | 5.0%
    0.92          | 2.5%
    1.09          | 1.0%
  Shapiro-Wilk test: W = 1 p = 0.59
  Z_1 = 0.2
  Z_2 = 1.0
```

```
Kolmogorov-Smirnov test: D = 0.28  p = 0
  Anderson-Darling test: A^2 = 1.9e+02
    significance  | critical value
    --------------|----------------
    0.58          | 15.0%
    0.66          | 10.0%
    0.79          | 5.0%
    0.92          | 2.5%
    1.09          | 1.0%
  Shapiro-Wilk test: W = 0.94 p = 0
  Z_1 = 32.2
  Z_2 = 2.5
```



Anderson-Darling, K-S, Shapiro-Wilk tests
Is the distribution Gaussian?

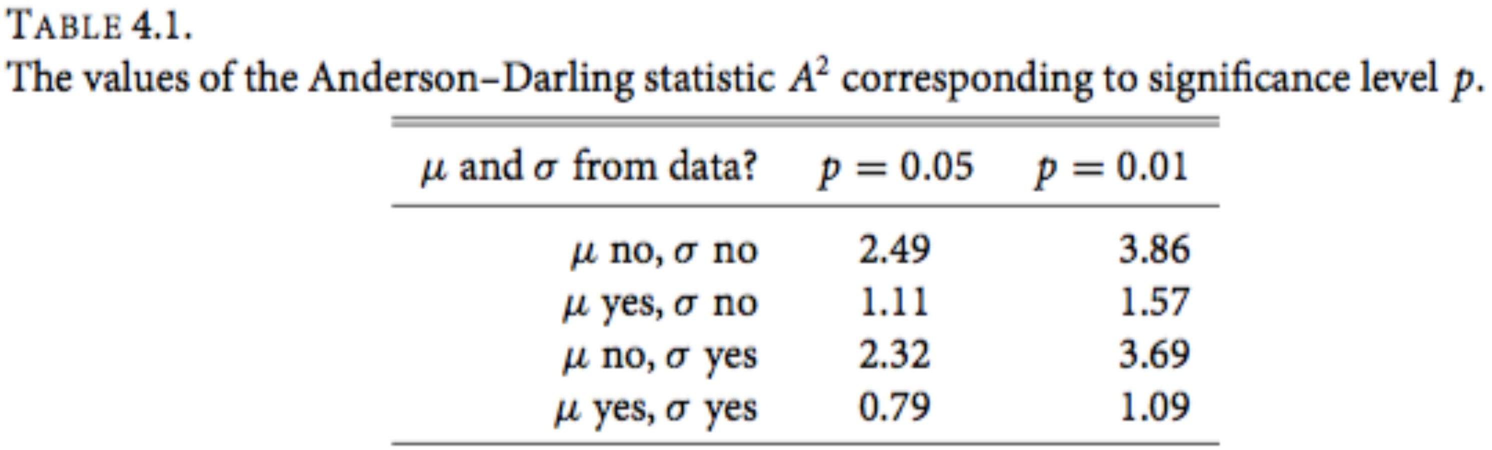
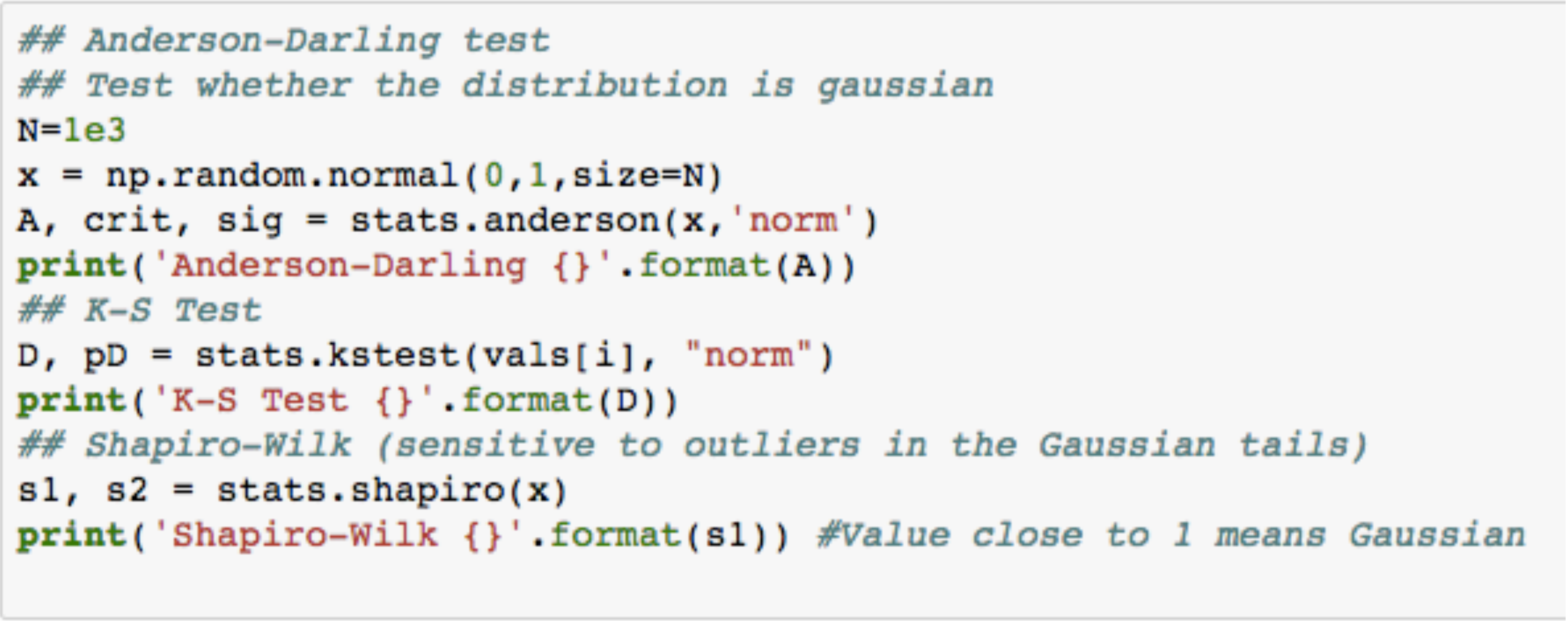


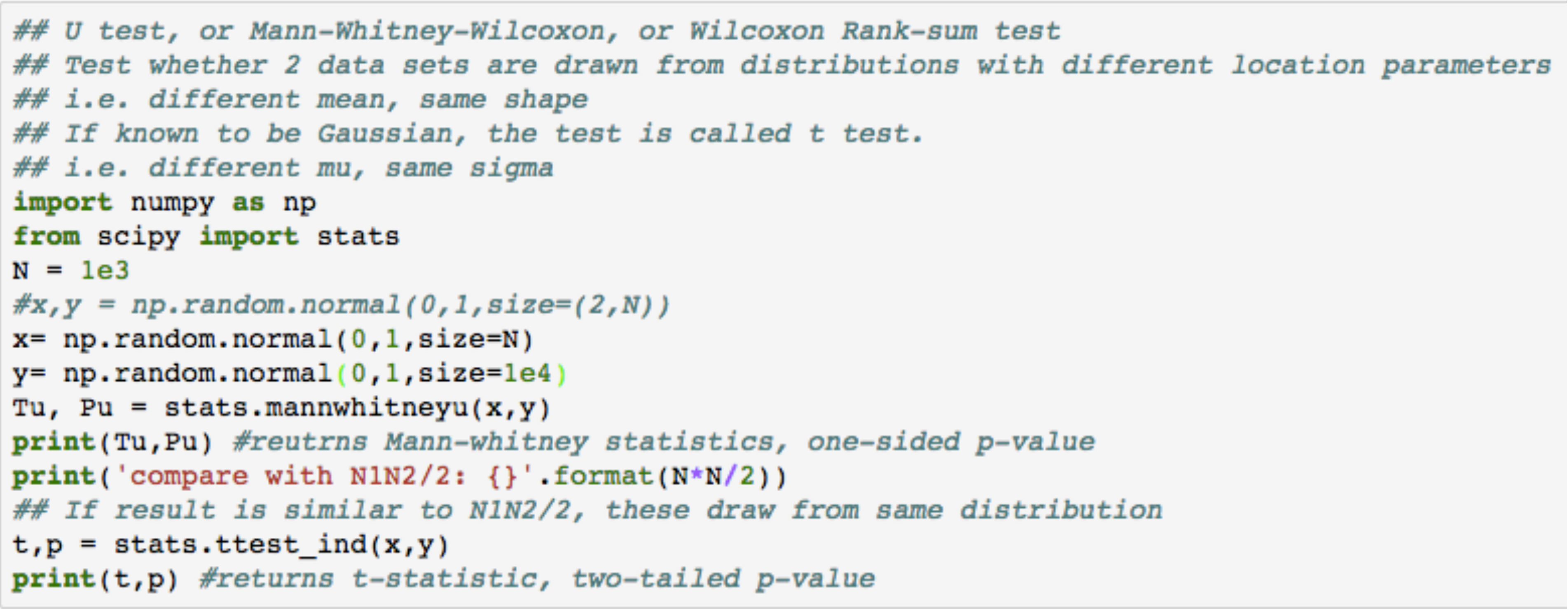
TABLE 4.1.
The values of the Anderson–Darling statistic A^2 corresponding to significance level p.

μ and σ from data?	$p = 0.05$	$p = 0.01$
μ no, σ no	2.49	3.86
μ yes, σ no	1.11	1.57
μ no, σ yes	2.32	3.69
μ yes, σ yes	0.79	1.09

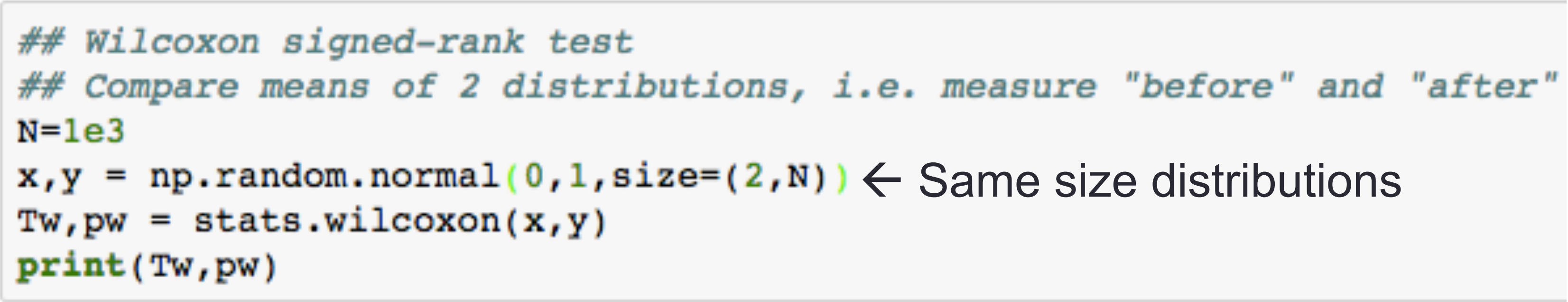


```
## Anderson-Darling test
## Test whether the distribution is gaussian
N=1e3
x = np.random.normal(0,1,size=N)
A, crit, sig = stats.anderson(x,'norm')
print('Anderson-Darling {}'.format(A))
## K-S Test
D, pD = stats.kstest(vals[i], "norm")
print('K-S Test {}'.format(D))
## Shapiro-Wilk (sensitive to outliers in the Gaussian tails)
s1, s2 = stats.shapiro(x)
print('Shapiro-Wilk {}'.format(s1)) #Value close to 1 means Gaussian
```

U test and Wilcoxon test
Test “locations” of distributions



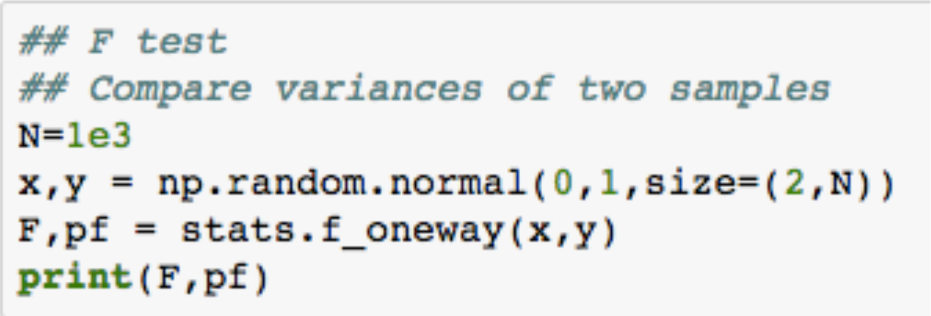
```
## U test, or Mann-Whitney-Wilcoxon, or Wilcoxon Rank-sum test
## Test whether 2 data sets are drawn from distributions with different location parameters
## i.e. different mean, same shape
## If known to be Gaussian, the test is called t test.
## i.e. different mu, same sigma
import numpy as np
from scipy import stats
N = 1e3
#x,y = np.random.normal(0,1,size=(2,N))
x= np.random.normal(0,1,size=N)
y= np.random.normal(0,1,size=1e4)
Tu, Pu = stats.mannwhitneyu(x,y)
print(Tu,Pu) #reutrns Mann-whitney statistics, one-sided p-value
print('compare with N1N2/2: {}'.format(N*N/2))
## If result is similar to N1N2/2, these draw from same distribution
t,p = stats.ttest_ind(x,y)
print(t,p) #returns t-statistic, two-tailed p-value
```



```
## Wilcoxon signed-rank test
## Compare means of 2 distributions, i.e. measure "before" and "after"
N=1e3
x,y = np.random.normal(0,1,size=(2,N))
Tw,pw = stats.wilcoxon(x,y)
print(Tw,pw)
```

← Same size distributions

F-test
Compare the variances



```
## F test
## Compare variances of two samples
N=1e3
x,y = np.random.normal(0,1,size=(2,N))
F,pf = stats.f_oneway(x,y)
print(F,pf)
```

Histograms, how to do them “right”

- “simplest non-parametric method to analyze 1D data” (ICVG p. 163)
- Bin width is the important tuning parameter
- More data, more bins; fewer data, fewer bins

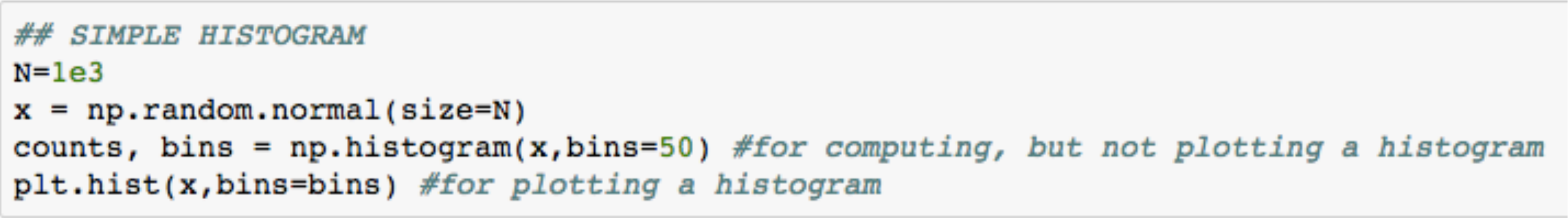
Bin width:

- Scott’s rule $\Delta_b = \frac{3.5\sigma}{N^{1/3}},$

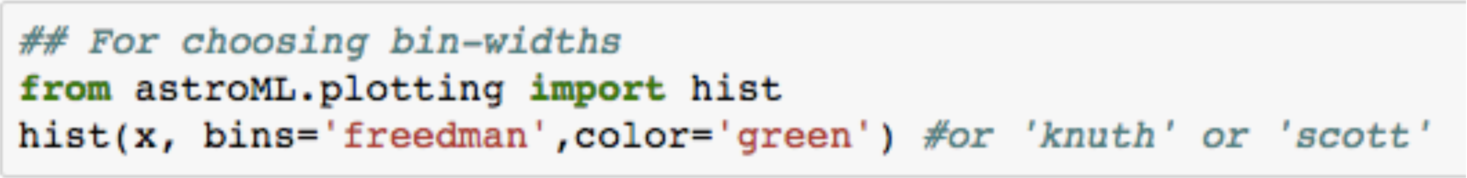
- Generalized to non-Gaussian distributions,

Freedman-Diaconis rule: $\Delta_b = \frac{2(q_{75} - q_{25})}{N^{1/3}} = \frac{2.7\sigma_G}{N^{1/3}},$

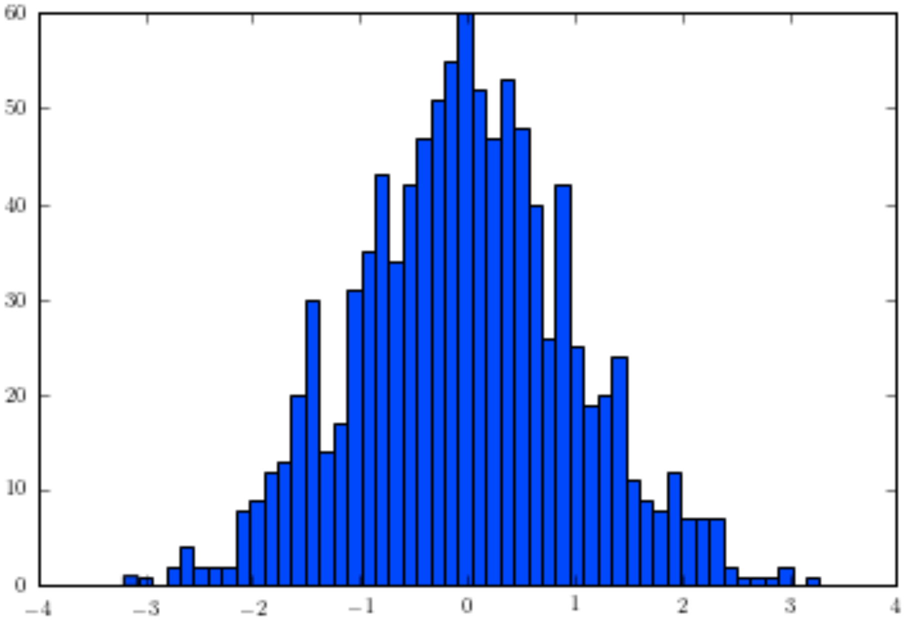
Histogram examples

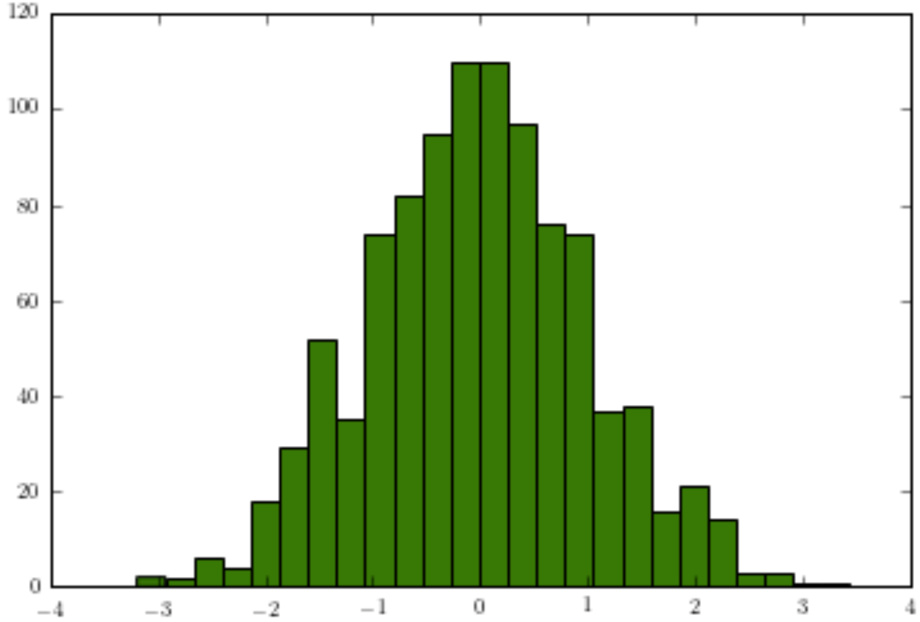


```
## SIMPLE HISTOGRAM
N=1e3
x = np.random.normal(size=N)
counts, bins = np.histogram(x,bins=50) #for computing, but not plotting a histogram
plt.hist(x,bins=bins) #for plotting a histogram
```



```
## For choosing bin-widths
from astroML.plotting import hist
hist(x, bins='freedman',color='green') #or 'knuth' or 'scott'
```





Summary of tools

- `from scipy import stats`

Note: For many more stat related functions install the software R and the interface package rpy.

- Look at book figures:

http://www.astroml.org/book_figures/index.html

- `from astroML.plotting import hist`

- https://github.com/adeleplunkett/MCMC/blob/master/161006_MLE2_Adele.ipynb