GAUTHAM NARAYAN

AST496: FOUNDATIONS OF DATA SCIENCE

RECAP

- We saw that Bayes' theorem is a direct result of **the axioms of probability**
- We cast our goal in this course as estimating the posterior distribution P(H|D)
- We've got data (you can plot it, be it scattered points in some high-D space, spectra, images whatever)
- We're now working on defining models for the data - i.e. the hypothesis H
- We've started looking at common **distributions** and their **moments**

RECAP

Posterior

How probable is the hypothesis given the data we observed

Likelihood

How probable is the data given the hypothesis is true

Prior

How probable was the hypothesis before we observed anything

$$p(\text{Hypothesis}|\text{Data}) = \frac{p(\text{Data}|\text{Hypothesis})p(\text{Hypothesis})}{p(\text{Data})}$$

Evidence

How probable is the data over all possible hypotheses

Expected Value

$$E(x) = \langle x \rangle = \int_X x \cdot p(x) dx$$

$$E(f(x)) = \langle f(x) \rangle = \int_X f(x) \cdot p(x) dx$$

Variance

$$\mathrm{Var}(x) = E([x - \langle x \rangle]^2)$$

nth moment (non-central)

$$\mu_n(x) = E(x^n)$$

nth moment (central)

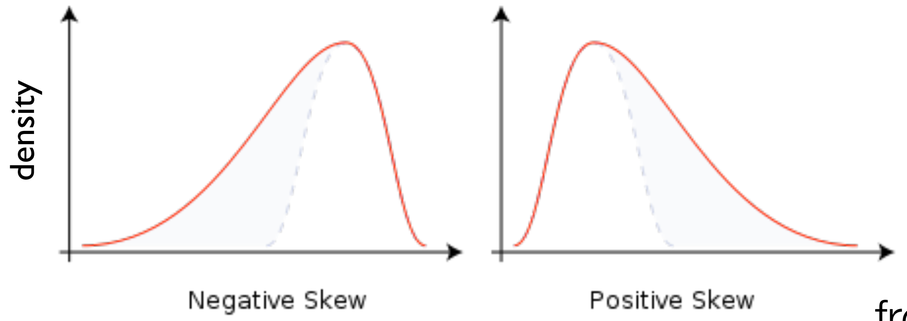
$$\widetilde{\mu}_n(x) = E([x - \langle x \rangle]^n)$$

3d and 4th moments of a distribution

- Skewness, asymmetry

$$\mu_3/\sigma^3 = \int_{-\infty}^{\infty} (x-\mu)^3 f(x) dx / \sigma^3$$

Normal: 0
Poisson: 1/√λ



from wikipedia

- Kurtosis

$$\mu_4/\sigma^4 = \int_{-\infty}^{\infty} (x-\mu)^4 f(x) dx / \sigma^4$$

$$\mu_4/\sigma^4 - 3$$

Normal: 0
Poisson: 1/λ

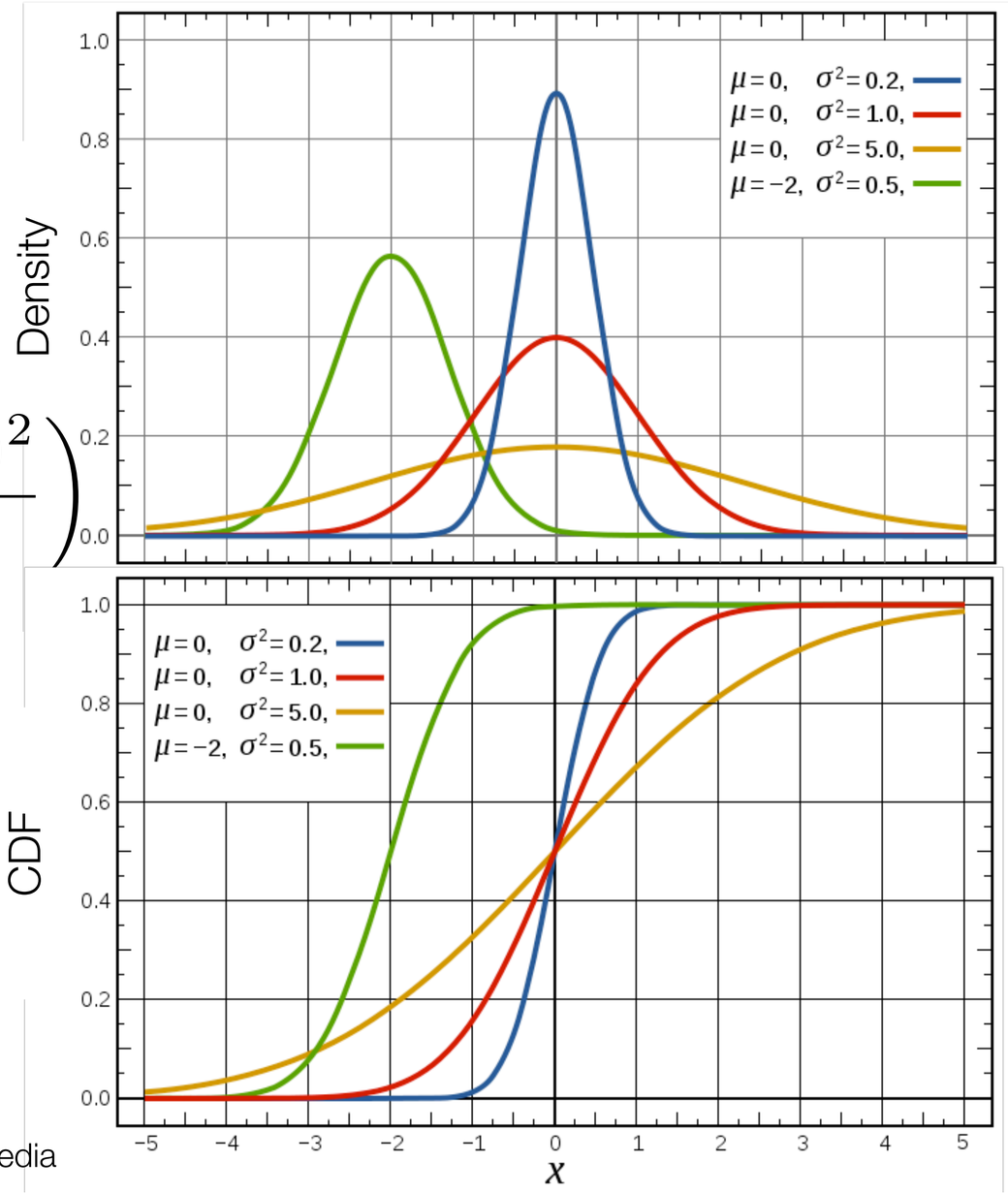
Example: Gaussian / Normal distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu \in \mathbb{R}, \sigma > 0$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Mean	μ
Median	μ
Mode	μ
Standard deviation	σ



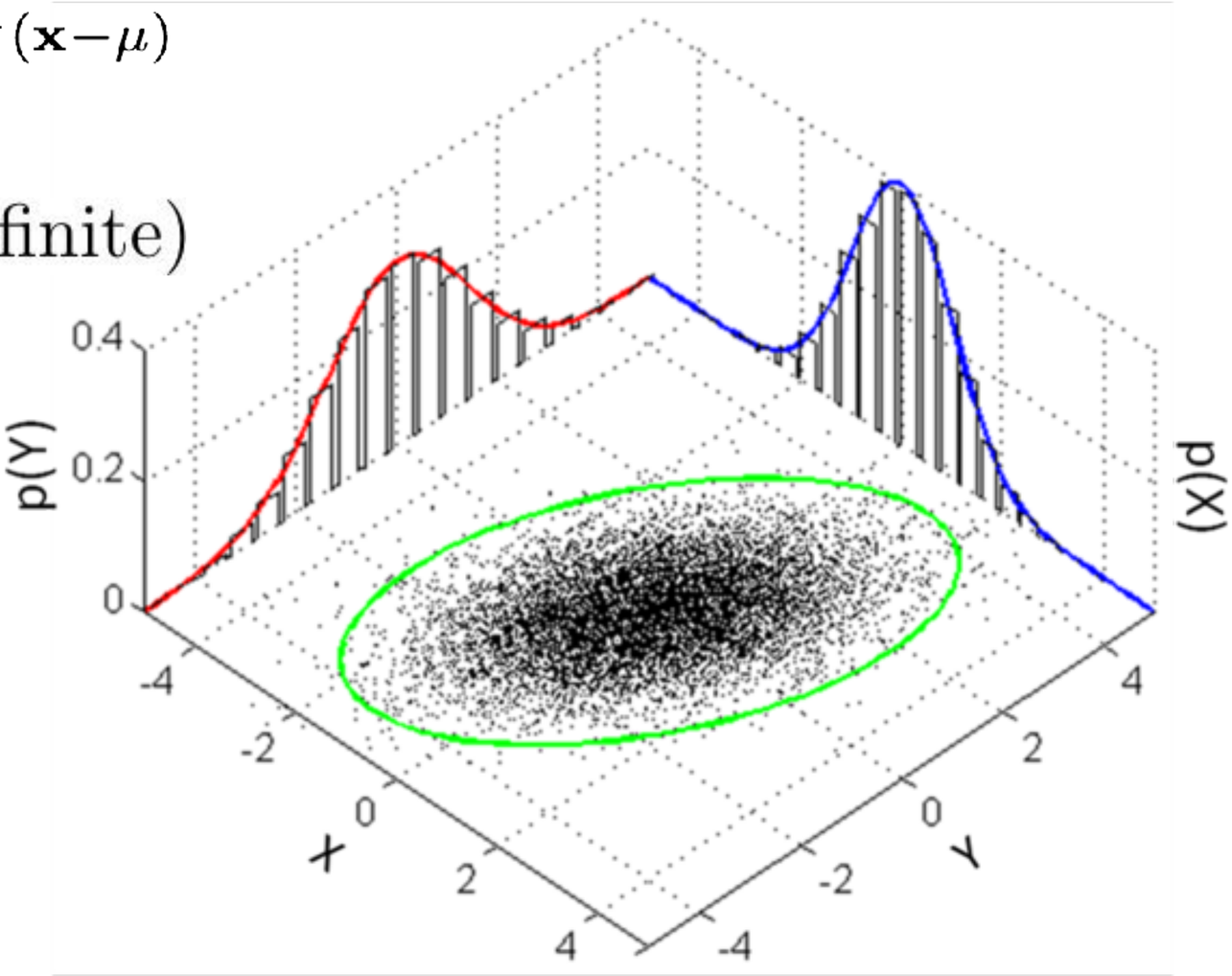
from wikipedia

THE MULTIVARIATE NORMAL DISTRIBUTION

$$\phi(\mathbf{x};\mu,\mathbf{\Sigma}) = \frac{1}{\sqrt{2\pi\det(\mathbf{\Sigma})}} e^{\frac{-1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

$$\mu \in \mathbb{R}^n, \mathbf{\Sigma} \in \mathbb{R}^{n\times n} \text{(and is postive definite)}$$

Mean	μ
Median	μ
Mode	μ
Standard deviation	$\mathbf{\Sigma}$



2.1

ESTIMATORS

Data, samples

- Usually we have observations, e.g. additive process

$$y_i = f(t_i) + \epsilon_i \qquad\qquad i = 1, \ldots, n$$

Deterministic random variable

- We want a characterisation of the deterministic and random parts

- Suppose something about the random variable, often normality: $\mathcal{N}(0, \sigma^2)$

- Assumption of models

- Estimate the parameters of a distribution, moments

- Exercise 1: Sample mean: $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E(\bar{X}) = \mu$

- Exercise 2: Sample variance (bias):

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad E(\hat{\sigma^2}) = \frac{n}{n-1}\sigma^2$$

redefine

$$\hat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Sample quantiles are estimators of quantiles

IN CLASS EXERCISE

- Load sample_stats.csv (pandas or astropy.table)
- You'll find multiple bivariate datasets
- Estimate the sample mean and sample standard deviation for each
- Now plot them… (matplotlib/seaborn)

Distribution derived from Normal distribution

1) Chi square distribution

Modified from Maria Suveges, Laurent Eyer

If $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ ⟹ $\sum_{i=1}^{k} X_i^2 \sim \chi_k^2$

iid= Independent identically distributed

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp(-x/2)$$

mean: k

variance: 2 k

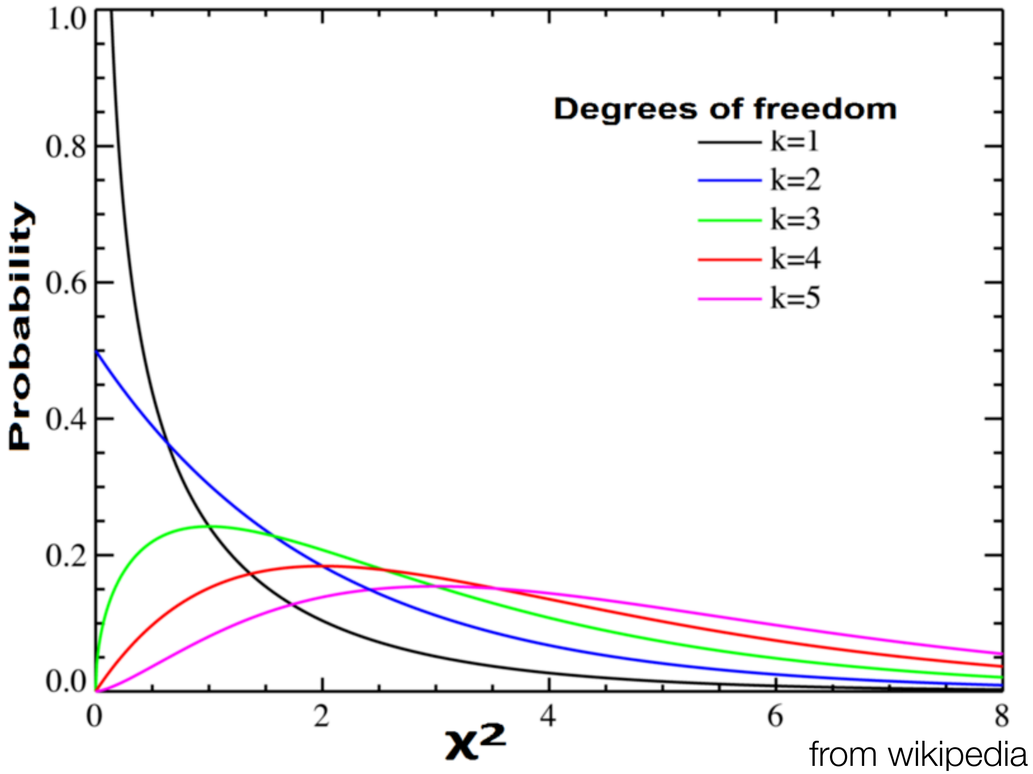
skewness: √(8/k)

kurtosis: 12/k

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$$

$$\sum_{i=1}^{k} (X_i - \bar{X})^2/\sigma^2 \sim \chi_{k-1}^2$$

When k is large χ_k^2 approximates a $\mathcal{N}(k, 2k)$



from wikipedia

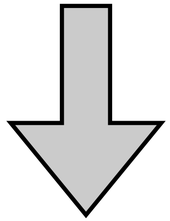
Central limit theorem

The distribution of the mean of a sufficiently large number of random variables can be approximated by a Gaussian distribution!

$$X_i,\ i = 1, \ldots, n \ \text{ iid } \ \text{with } \ \mathrm{E}(\mathrm{X_i}) = \mu \ \ \mathrm{Var}(\mathrm{X_i}) = \sigma^2$$

iid= Independent identically distributed

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{follows approximately} \quad \mathcal{N}(0,1)$$



One reason why
the Gaussian distribution is so important

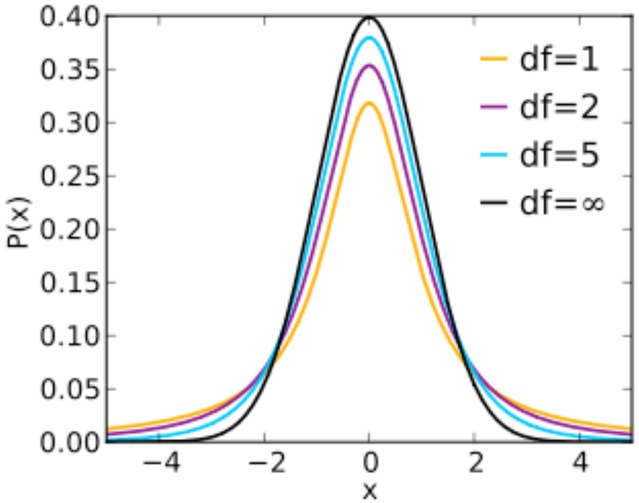
Distribution derived from Normal distribution
2) Student distribution

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

Note $t_\infty = \mathcal{N}(0,1)$



mean: 0 n>1
NaN n=0,1

variance: n/(n-2) n>2
∞ 1<n≤2
otherwise NaN

skewness: 0 n>3

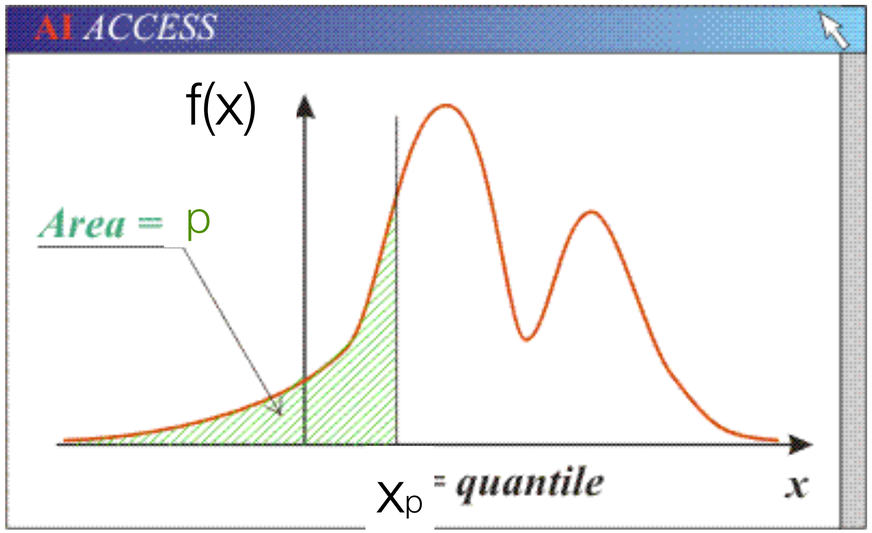
kurtosis: 6/(n-4) n>4

SO WE DEFINITELY NEED A QUANTITATIVE WAY OF ASSESING THE SIMILARITY OF TWO DISTRIBUTIONS

Quantiles

- x_p : p-quantiles of f(x)

$$p = \int_{-\infty}^{x_p} f(x)dx$$



from www.aiacces.net

- Measure of location: Median

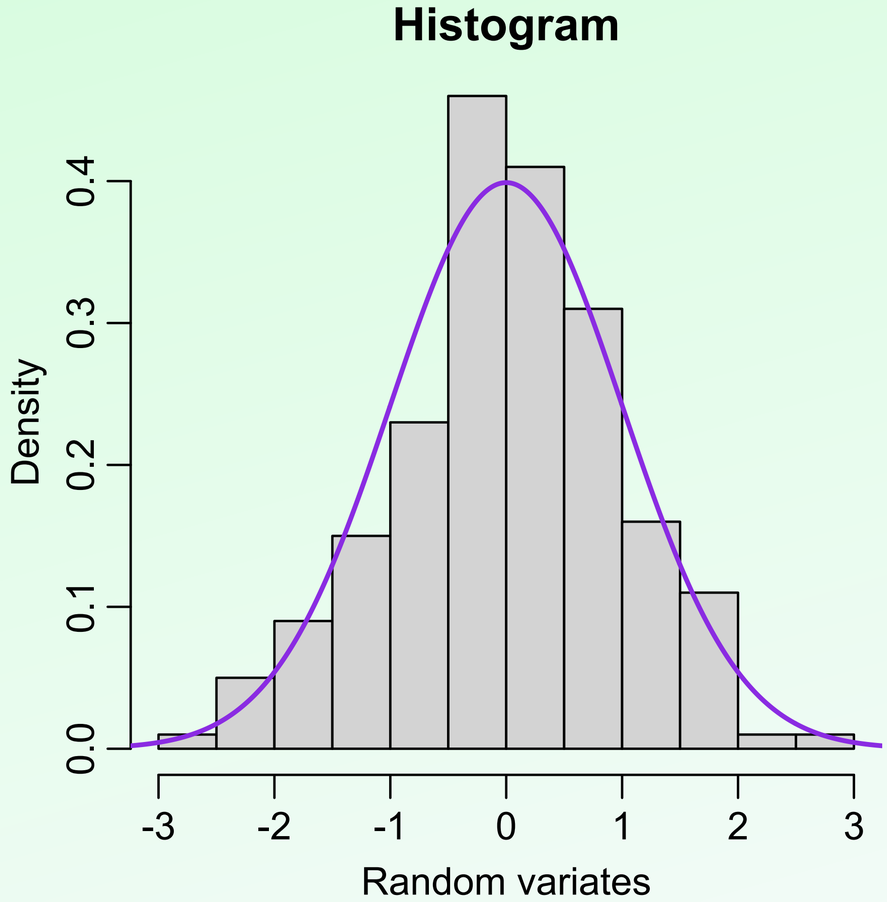
$$1/2 = \int_{-\infty}^{x_{1/2}} f(x)dx$$

- Measure of dispersion: Inter-quantile range

$$\mathrm{IQR} = x_{3/4} - x_{1/4}$$

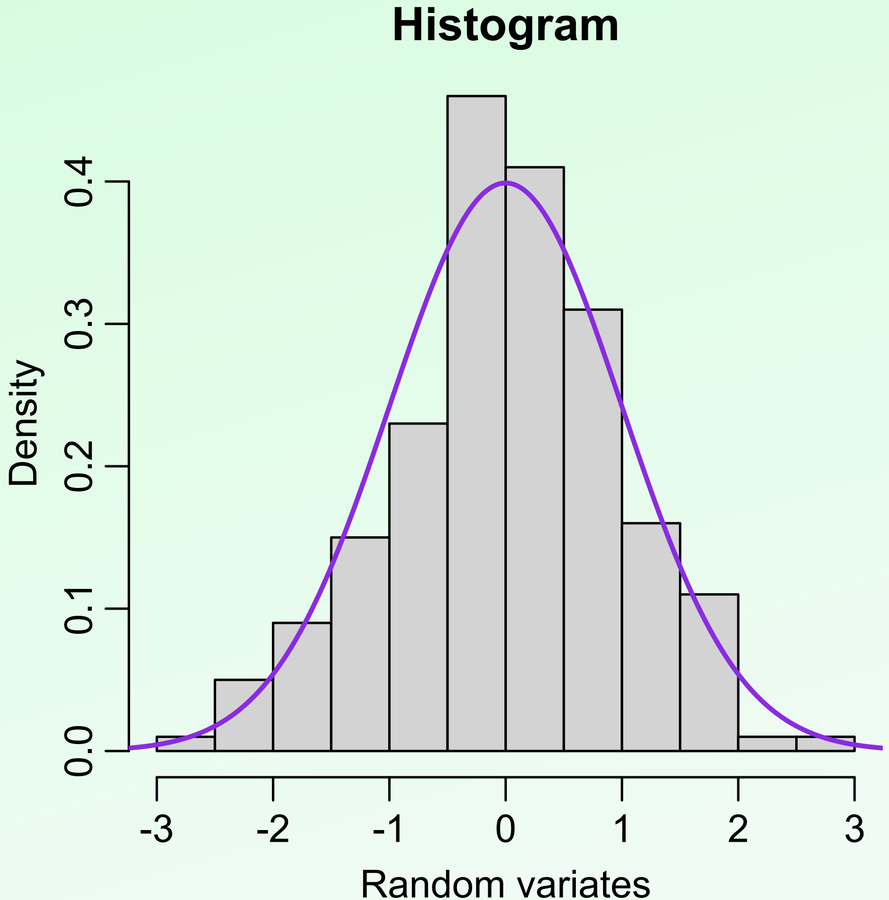
Diagnostics: the QQ plot

- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)



Diagnostics: the QQ plot

- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)



Statisticians' preference:
quantile-quantile (QQ) plot.

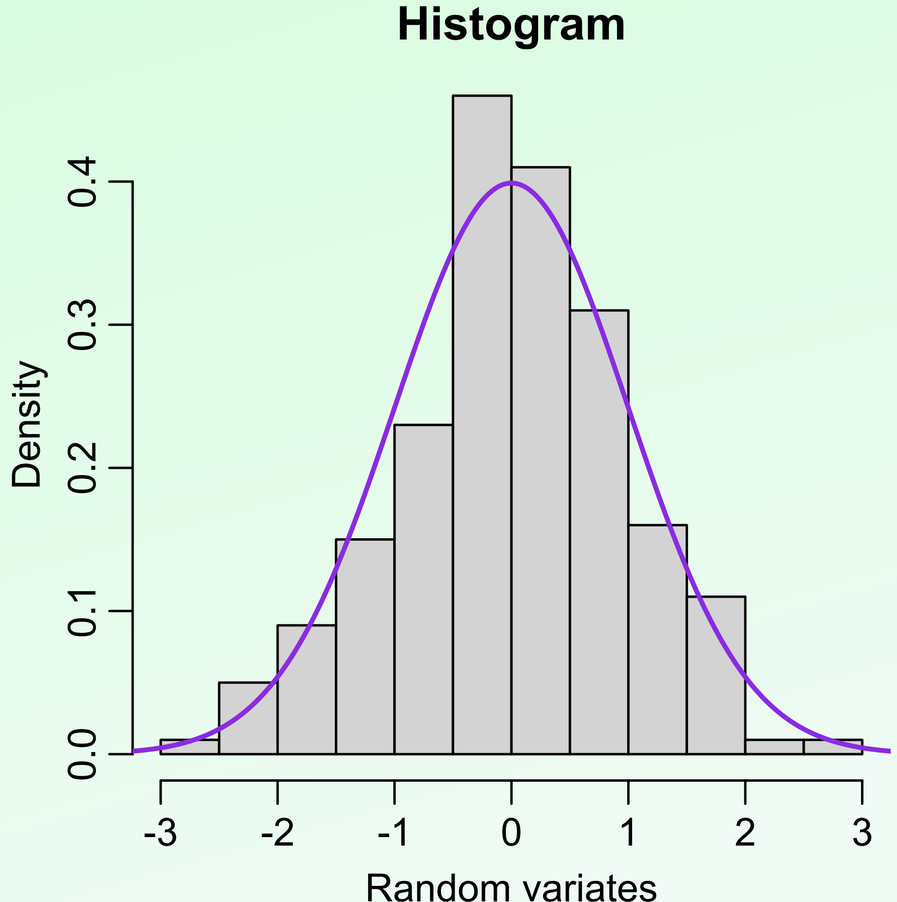
It consists of the pairs

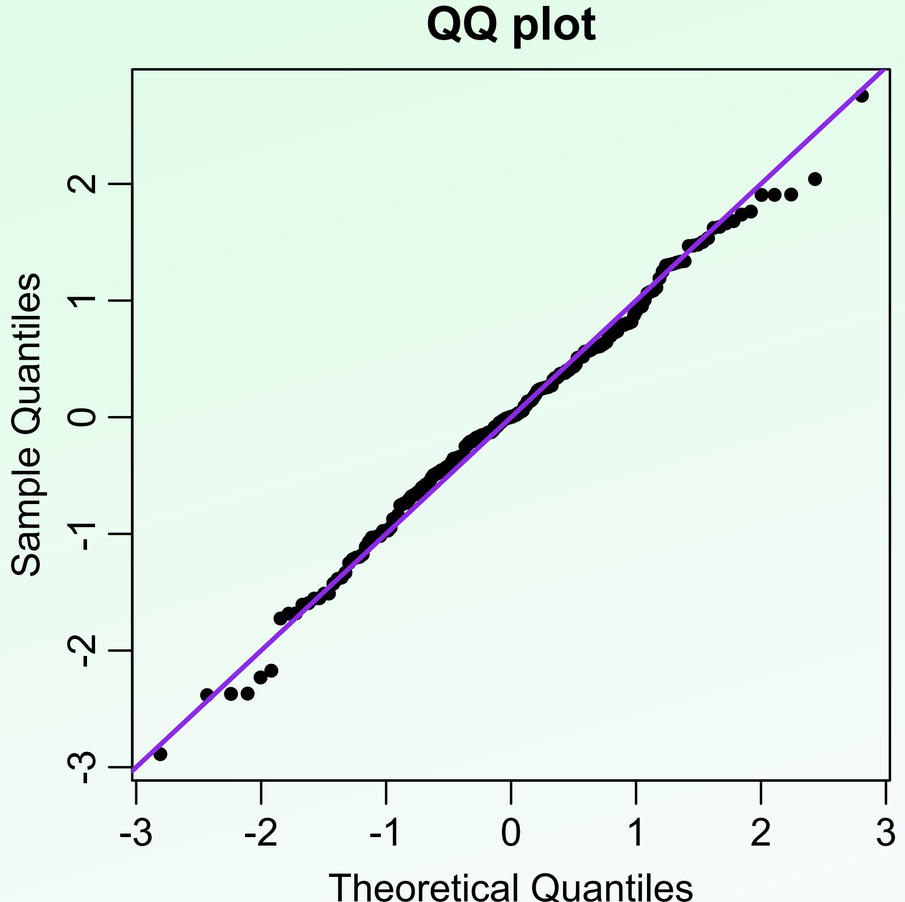
$$\left\{ \Phi^{-1}\left(\frac{j}{n+1}\right), Y_{(j)} \right\},$$

where
$\Phi^{-1}\left(\frac{j}{n+1}\right)$ is the inverse of the standard normal distribution, and
$Y_{(j)}$ is the ordered sample.

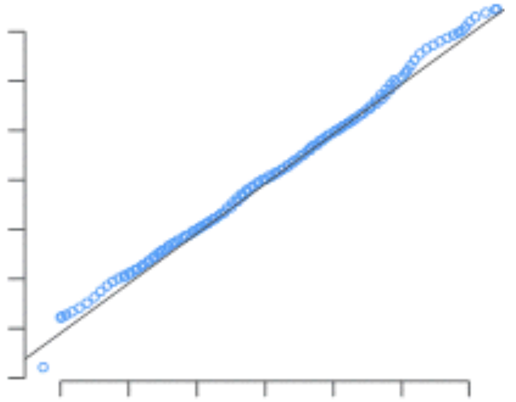
Diagnostics: the QQ plot

- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)

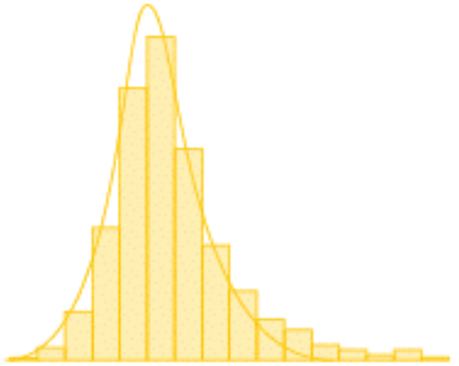


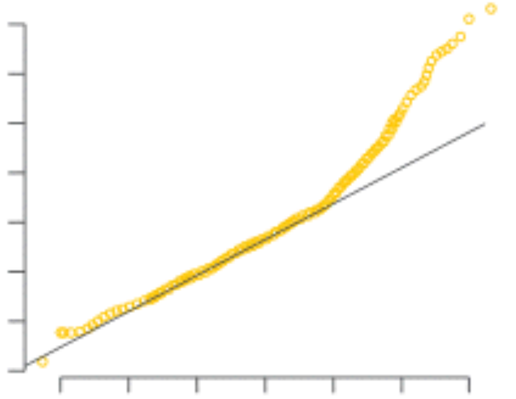


Normally distributed data

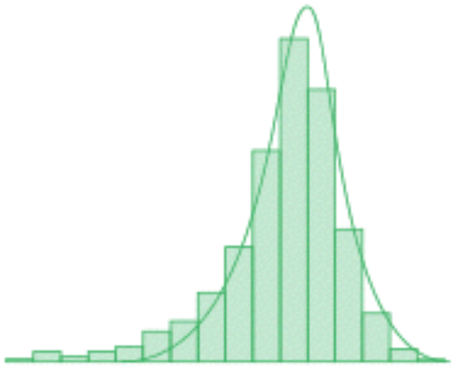


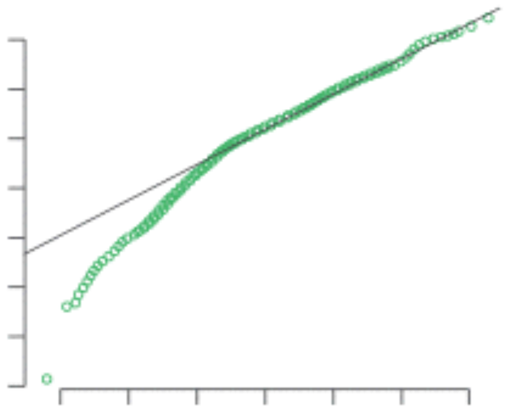
Right-skewed data





Left-skewed data



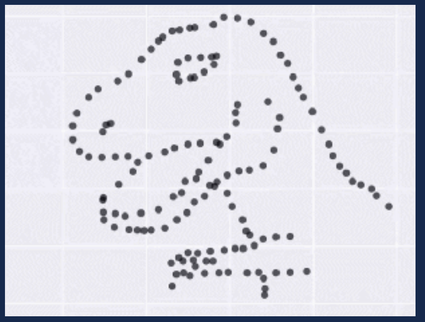


IN CLASS EXERCISE

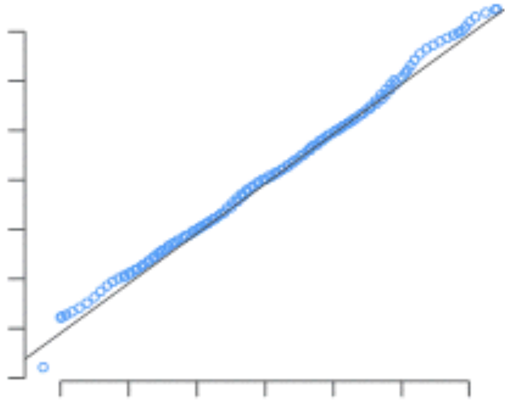
- Now that you know how to generate points from a distribution, we can use the QQ plot to compare distributions to each other, or to a normal distribution
- Use scipy.stats to generate some random numbers from a normal, uniform, and Cauchy distribution
- Use statsmodels.api.qqplot to produce a qq plot of these distributions
- Now generate random numbers from two different normal distributions (different locations and variances) and concatenate them
- Again check the QQ plot

RECAP

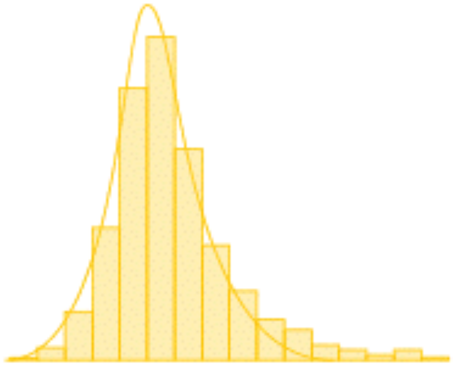
- Random variables (distribution.rvs() in scipy.stats) are just samples from distributions
- You can characterize distributions by moments - and it's fairly typical in sciences to just report the first two moments ($\mu \pm \sigma$)
  - But this doesn't tell you much about the underlying distribution
  - **There's a tacit assumption that it is Gaussian becauseof the Central Limit Theorem, even when the CLT doesn't hold**
- So we began discussing how to compare distributions with a Gaussian to check our assumption
  - Basic: look at the data if you have enough of it
  - Use the QQ plot to look for skewness

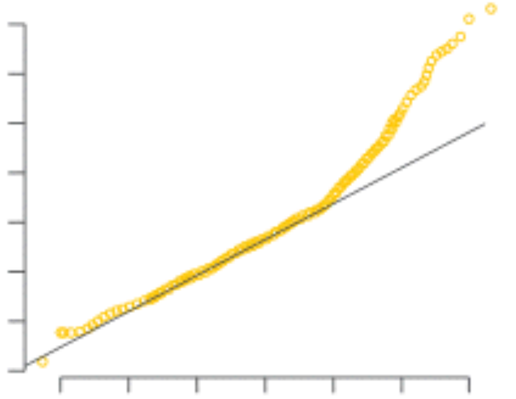


Normally distributed data

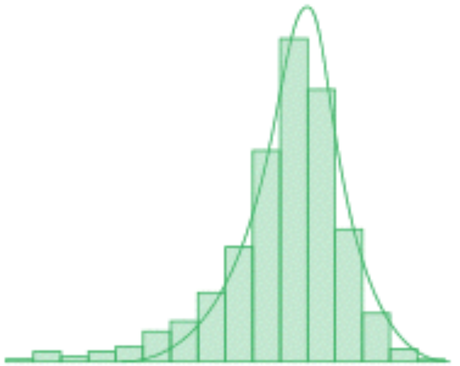


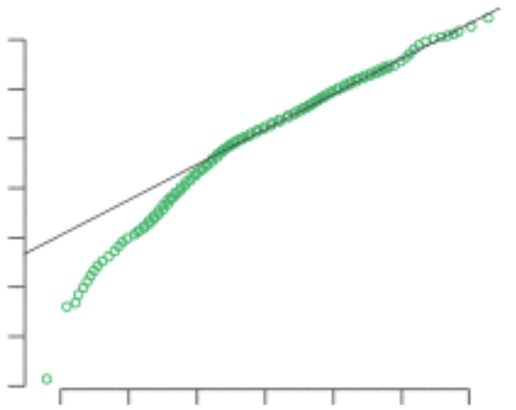
Right-skewed data





Left-skewed data





IN CLASS EXERICSE 4: QUANTILES LET YOU DO A LOT MORE THAN DIAGNOSE A PROBLEM

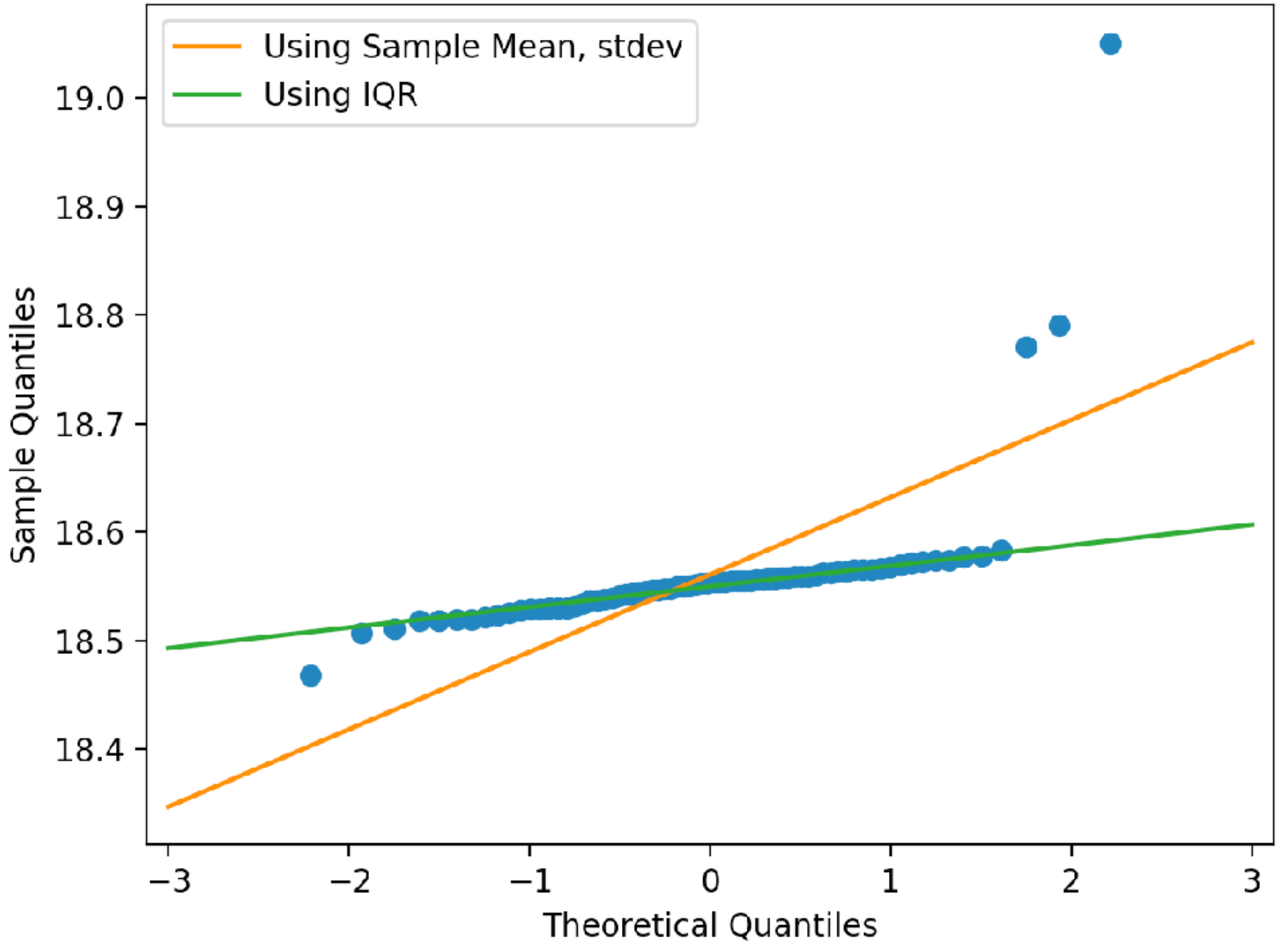
- I've provided a data file + notebook in the github repo
  - The notebook already reads the data file for you, so it's faster to get started
- Get the suspiciously named "mag.outlier" column from the file and estimate the sample mean μ and standard deviation σ
- Since the regular Q-Q plot just has the data values on the y-axis and theoretical quantiles on the x-axis (typically over some range like -3 to +3) you can just plot a line with $y = \mu + \sigma \cdot x$
- Next, compute the IQR - the 75th percentile - the 25th percentile of mag.outlier (numpy.percentile is your friend) - those are two new y-values.
- Over plot the line with you'd get if you had described the distribution by the IQR instead. You'll need to get the x values for the 25th and 75th percentiles - i.e. given an area of 0.25 what is the x-location for a normal distribution that encloses this area (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.rv_continuous.ppf.html might be of use)
  - You'll also need to remember what the equation of a line with two points is :-)

HOPEFULLY YOU GET SOMETHING THAT LOOKS LIKE THIS:

- You'll notice the IQR is pretty robust against outliers
- For a normal distribution:

$$\sigma = \mathrm{IQR}/2\sqrt{2} \cdot \mathrm{erf}^{-1}(\tfrac{1}{2}) \quad (\approx 1.349)$$

  - so if your data is mostly Gaussian with a few outliers that are throwing off the standard deviation, you can get a better estimate by using the IQR.



- Another alternative is the median absolute deviation:

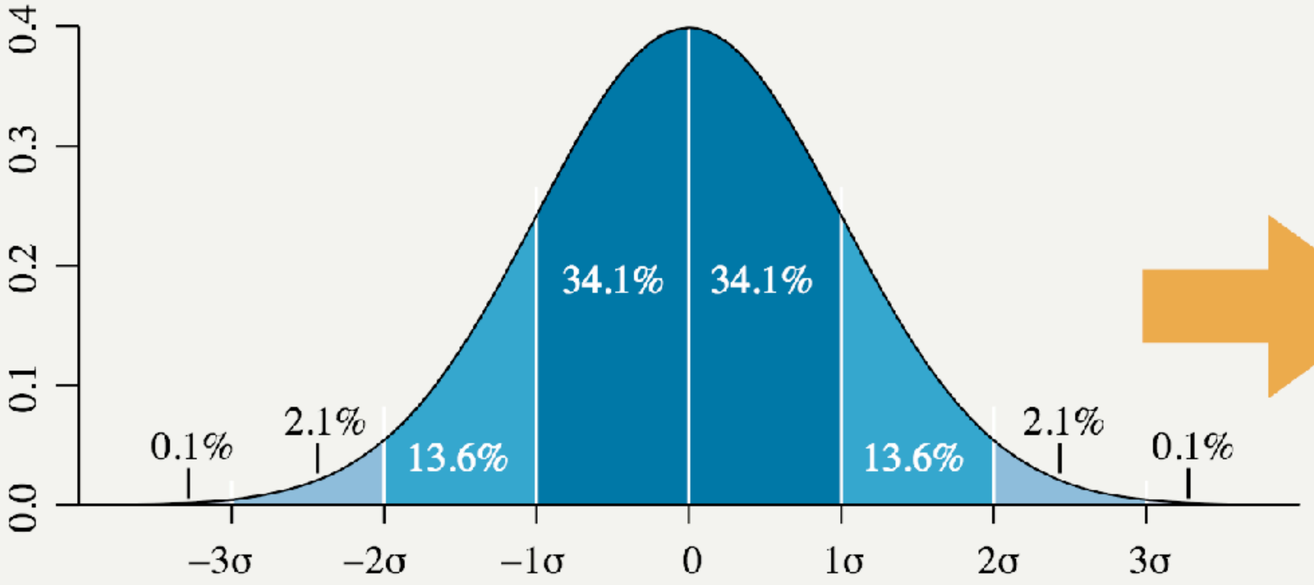
$$\mathrm{MAD} = \mathrm{median}\left(\left|x_i - x_{\frac{1}{2}}\right|\right) \quad \sigma \approx 1.4826 \cdot \mathrm{MAD}$$

L-ESTIMATORS AND OUTLIER REJECTION

- Both the IQR and MAD are the basic version to the most common method astronomers use to deal with outliers - sigma clipping (astropy.stats.sigma_clip)
  - Assume the data is normally distributed
  - Estimate a mean and standard deviation for the data
  - Identify points that are more than some <theoretical quantile>-sigma away from the mean (e.g. 4σ) and "clip" those points
  - Repeat the process: recompute mean and sigma, until they change by less than some small amount or for a preset number of iterations
    - This process is ad-hoc and can be painfully slow with large amounts of data - in those cases, just use the IQR/MAD directly.
- IQR and MAD are also examples of what are called L-estimators (L because they are linear combinations of moments/ordered statistics). **LINEAR IS GOOD. LINEAR IS FAST.**
  - Another example - the Winsorized mean - replace values in the tails of the distribution (say 5% and 95% with the 5th and 95th percentile values) and then take the mean.

NULL HYPOTHESIS REJECTION TEST (NHRT)

- Looking at your outliers is important - read the short article on Moodle about Millikan, confirmation bias, and determining the actual charge of the electron
- When you want to claim a discovery is significant, you need to compare it to something - a “null hypothesis”
- You have data drawn from some unknown population
- Assume a model for that population
- This allows you to make a prediction for the data you have
- "Under the null hypothesis" = if the model is a good description of the data then the data should be highly probable
- Define a test such that your significant observation has a very low probability of happening
- Given your data, run your test and see if your low-probability event is like what you’ve seen. If so, then you can reject your model
- If it hasn't happened, your model still holds **but that doesn't tell you it's a good description of the data**



rejected at 99.7%

3σ 0.003 p-value

0.3% confidence

data

does not falsify → **model still holds**

falsifies → **model rejected**

low probability event happened

NULL HYPOTHESIS REJECTION TEST (NHRT)

- So when you hear "x is a 2-σ detection!"
- 2-σ: confidence level
- 0.05: p-value
- 95%: threshold
- Unstated here is that there is an assumption of some model, typically a Gaussian.

M-ESTIMATORS

- L-estimators and derived techniques like sigma-clipping are still used by astronomers, but largely rejected by statisticians in favor of “M-estimators”
  - M for maximum or minimum - these are going to be using extremum
- These give us a less ad-hoc way of incorporating the uncertainties and dealing with outliers
  - As before, assume a form for the empirical PDF $f(x;\theta)$
  - Define a **loss function** $\rho(u)$ for some quantity u e.g. $x-\mu$
  - Minimize this loss over the sample $\sum_{i=1}^{N} \rho(u_i)$

M-ESTIMATORS

- If ($x1$,$x2$...,xN) is a set of i.i.d random variables from some distribution p(x) which we don't know, but want to estimate.
- What we're trying to do is to build an estimator for moments out of the sample.
- If we define:

$$\rho(x, \theta = \mu) = \frac{(x - \mu)^2}{2}$$

- How do you go about minimizing this with respect to θ?

- <insert math here>

- For $u=x-\mu$, there's several potential cost functions:

$$\rho(u) = u^2$$

- is an old friend to all of you. It's the sum of squared residuals or what we sometimes call the $L2$ norm.
- Of the loss functions, this is the nicest behaved - it's convex and differentiable
- **Differentiability implies continuity. Continuity does not guarantee differentiability.**
- Most M-estimators aren't nicely differentiable, but often have other properties - such as resistance to outliers.
- You've already seen another loss function: $\rho(u) = |u|$
- This is the sum of absolute residuals and is related directly to the MAD. You might see it called the $L1$norm, particularly in machine learning literature.

Estimators: method of moments (robustness)

Examples of loss functions:

- sum of squared residuals

  $\rho(u) = u^2$

- sum of absolute residuals

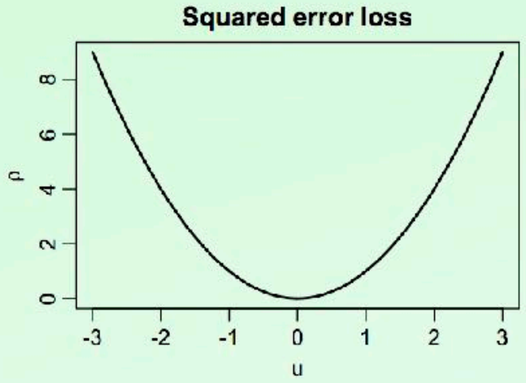
  $\rho(u) = |u|$

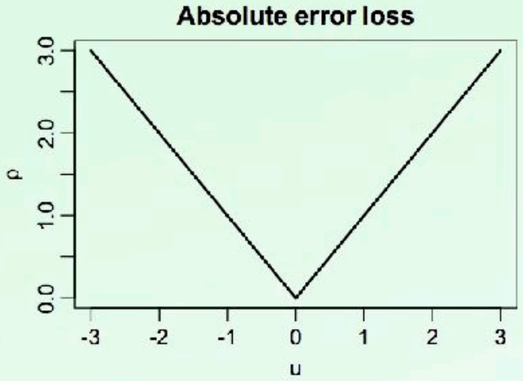
- Huber's loss function

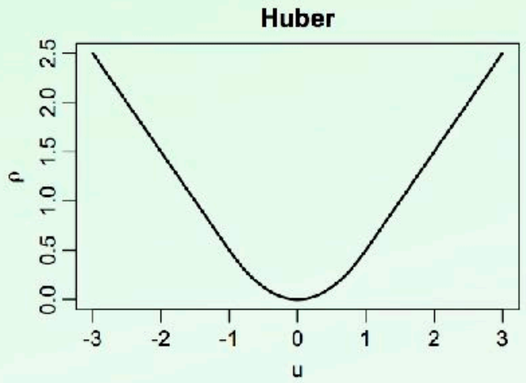
$$\rho(u) = \begin{cases} \frac{1}{2}u^2) & \text{if } |u| \leq \delta \\ \delta(|u| - \frac{1}{2}\delta) & \text{if } |u| > \delta \end{cases}$$

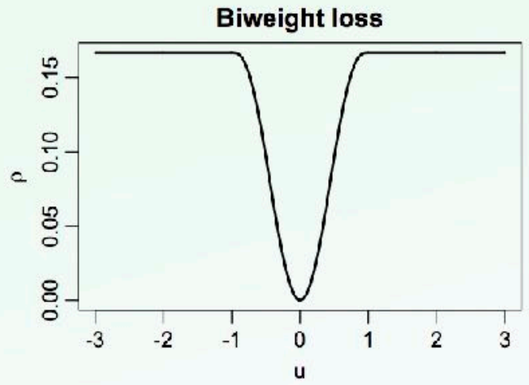
- Tukey's biweight

$$\rho(u) = \begin{cases} \frac{1}{6}[1 - (1 - u^2)^3] & \text{if } |u| \leq 1 \\ \frac{1}{6} & \text{if } |u| > 1 \end{cases}$$









AND NOW WE HAVE SOME MATH.

GIRD YOURSELVES.

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- If we know the distribution from which our data were drawn (or make a hypothesis about it), then we can compute the probability of our data being generated.
- For example, for the Gaussian distribution probablity of getting a specific value of x is given by:

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

- If we want to know the total probability of our *entire* data set (as opposed to one measurement) then we must compute the *product* of all the individual probabilities:

$$L \equiv p(\{x_i\} \mid H(\theta)) = \prod_{i=1}^{n} p(x_i \mid H(\theta))$$

- H refers to the hypothesis and θ refers collectively to the k parameters of the model, which can generally be multi-dimensional.
- In words, **this is the probability of the data given the model parameters.**

REMEMBER INDEPENDENCE BEFORE MULTIPLYING PROBABILITIES WILLY NILLY

- Note that this implicitly assumes that the measurements in your sample are, as always, i.i.d
- Recall (from the axioms of probability) that $P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$
- and iff $P(A \cap B) = P(A) \cdot P(B)$
- then A **and** B **are independent.**

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- We can write this out as:

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right)$$

- Which simplifies to:

$$L = \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\sum\left[\frac{-(x_i - \mu)}{\sigma}\right]^2\right)$$

  - where we have written the product of the exponentials as the exponential of the sum of the arguments, which will make things easier to deal with later.

$$\prod_{i=1}^{n} A_i \exp(-B_i) = (A_i A_{i+1} \ldots A_n) \exp[-(B_i + B_{i+1} + \ldots + B_n)]$$

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- The argument to the exponential

$$L = \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum \left[\frac{-(x_i - \mu)}{\sigma} \right]^2 \right)$$

- This should look vaguely familiar from Monday…

Distribution derived from Normal distribution

1) Chi square distribution

Modified from Maria Suveges, Laurent Eyer

If $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ ⟹ $\sum_{i=1}^{k} X_i^2 \sim \chi_k^2$

iid= Independent identically distributed

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp(-x/2)$$

mean: k

variance: 2 k

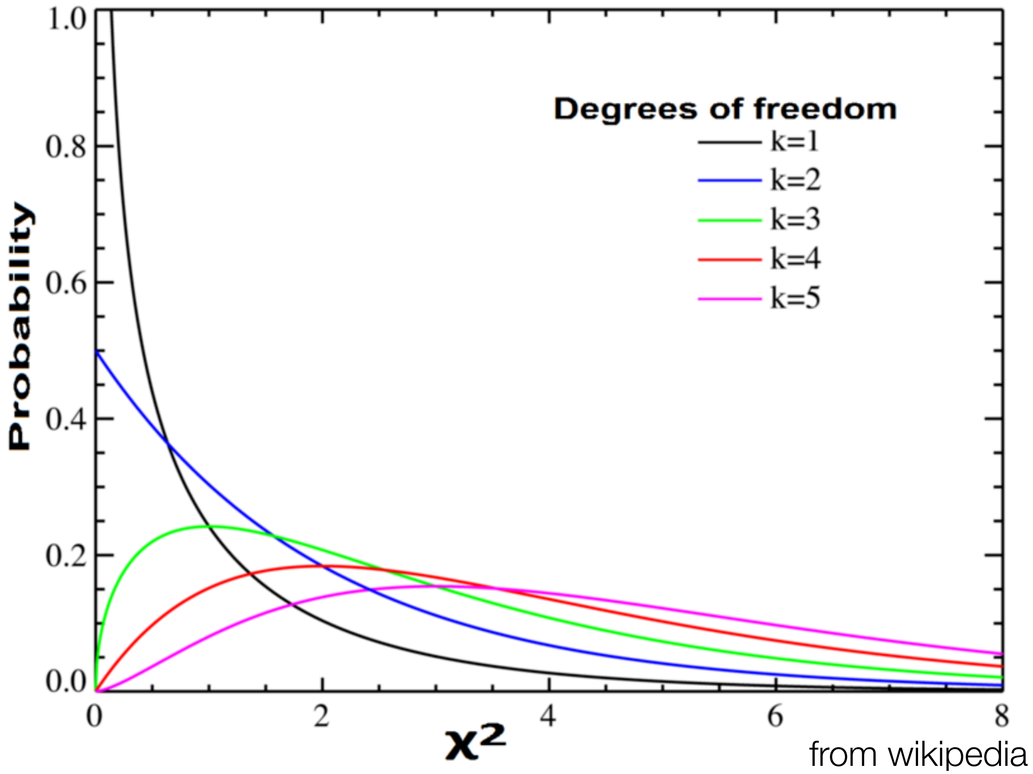
skewness: √(8/k)

kurtosis: 12/k

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$$

$$\sum_{i=1}^{k} (X_i - \bar{X})^2/\sigma^2 \sim \chi_{k-1}^2$$

When k is large χ_k^2 approximates a $\mathcal{N}(k, 2k)$



from wikipedia

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- The argument to the exponential

$$L = \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum \left[\frac{-(x_i - \mu)}{\sigma} \right]^2 \right)$$

- is just

$$\exp\left(-\frac{\chi^2}{2} \right)$$

- where

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- For practical reasons, it's better to work with the natural logarithm of the likelihood (we'll get to why in a few slides)
- We define the log-likelihood function as lnL=ln[$L(\theta)$]. The maximum of this function happens at the same place as the maximum of L. Given all that, we have:

$$\ln \mathrm{L} = \text{constant} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

- We then determine the maximum in the same way that we always do. It is the parameter set for which the derivative of lnL is zero:

$$\left. \frac{d \ln \mathrm{L}(\mu)}{d\mu} \right|_{\hat{\mu}} \equiv 0$$

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- We then determine the maximum in the same way that we always do. It is the parameter set for which the derivative of lnL is zero:

$$\left.\frac{d\,\mathrm{lnL}(\mu)}{d\mu}\right|_{\hat{\mu}} \equiv 0$$

- That gives

$$\sum_{i=1}^{N} \frac{(x_i - \hat{\mu})}{\sigma^2} = 0$$

- (note: we should also check that the 2nd derivative is negative, to ensure this is the maximum of L)

- (also note: any constants in ln$L$ disappear when differentiated, so constant terms can typically be ignored - i.e. you can add any constant to the log-likelihood to make the numbers reasonable.

CHI-SQ MINIMIZATION IS JUST A SPECIAL CASE OF AN M-ESTIMATOR

- So, maximizing the likelihood is the same as minimizing $\chi 2$:
- Maximizing the likelihood is solving for the extremum of:

$$L \sim \exp\left(-\frac{\chi^2}{2}\right)$$

- is the same as Maximizing the natural logarithm of the likelihood: (because the log is a monotonically increasing function)

$$\ln(L) \sim -\frac{\chi^2}{2}$$

- and therefore is the same as minimizing the negative log likelihood:

$$-\ln(L) \sim \frac{\chi^2}{2}$$

IF YOUR ERRORS ARE CONSTANT

- i.e. σ=constant, then

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \hat{\mu} = N\hat{\mu}$$

- then:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- which is just the arithmetic mean of all the measurements.

THE SAMPLE MEAN AS AN M-ESTIMATOR

- The mean of observations drawn from a $N(\mu,\sigma=const)$ distribution is a maximum-likelihood estimator of the distribution's μ parameter.
- We'd used this and guessed this intuitively, but this derivation clarifies our choice: as an estimator of the real value of μ, we adopt the value $\hat{\mu}$ with which it's maximally likely for the measured data set to occur.
- It also exposes the assumptions behind this conclusion; namely homoscedasticity and gaussianity of uncertainties. For example, if our uncertainties were Cauchy-distributed the mean (or any higher moment) is not defined
- The Core Idea Behind Maximum Likelihood Estimators - Let’s say that we know that some data were drawn from a Gaussian distribution, but we don't know the $\theta=(\mu,\sigma)$ values of that distribution (i.e., the parameters).
  - Then **Maximum Likelihood Estimation** method tells us to think of the likelihood as a function of the unknown model parameters, and find those that maximize the value of L
  - **Those will be our Maximum Likelihood Estimators for for the true values of the model.**

WHY MINIMIZE THE NLL INSTEAD OF MAXIMIZE THE LIKELIHOOD?

- You will generally see people minimizing the negative log likelihood rather than maximizing the likelihood directly
- The reason is fairly simple: While the components of L may be normalized pdfs, their product is not.
- The PDFs often yield very small numbers, which you are then multiplying together, making even smaller numbers.
- i.e. 10^−24 is pretty small - you'll run into floating point precision issues.
- log10(10^−24)= −24 is not!