

Statistics: The Beauty and the Beast

Matthieu Gilson

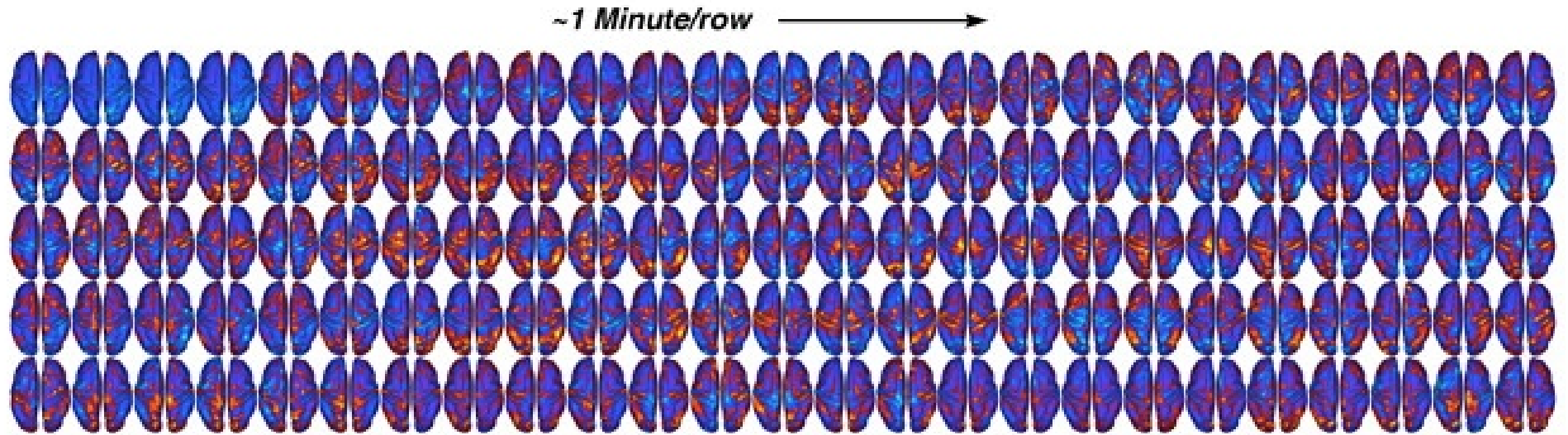
Statistics in Neuroscience

- Why use statistics?

Statistics in Neuroscience

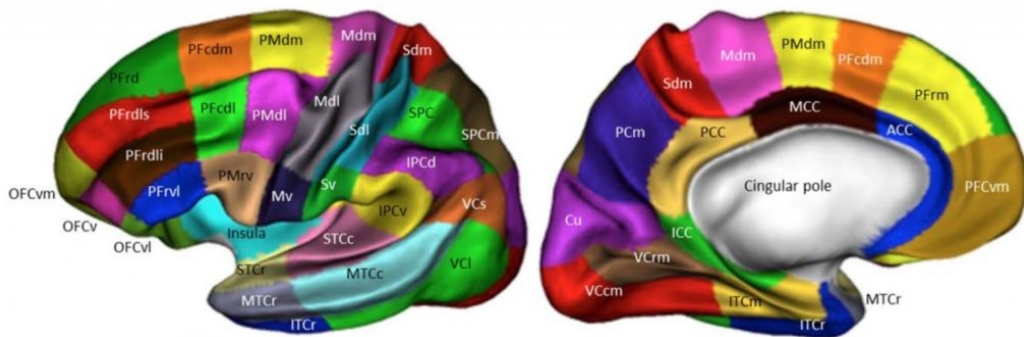
- Why use statistics?
- Uncertainty in measurements
- Intrinsic biological variability (noise?)
- Fit models to data

Example of Neuroimaging Data (fMRI)

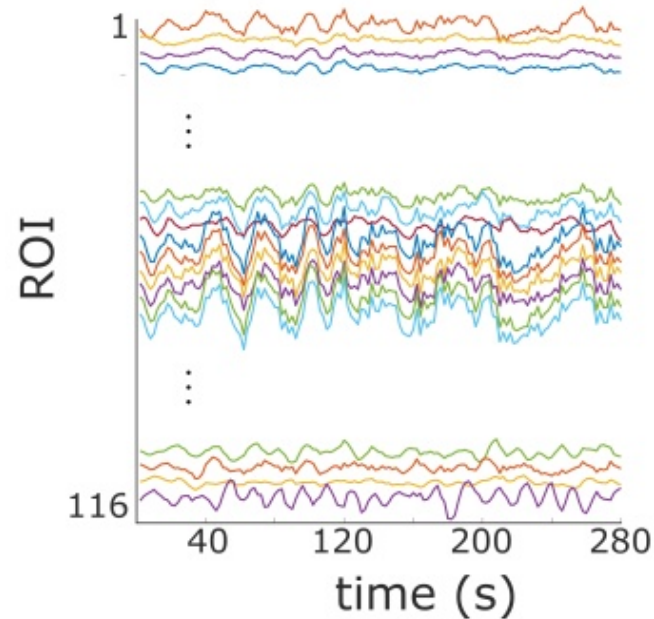


Example of Neuroimaging Data (fMRI)

Brain atlas

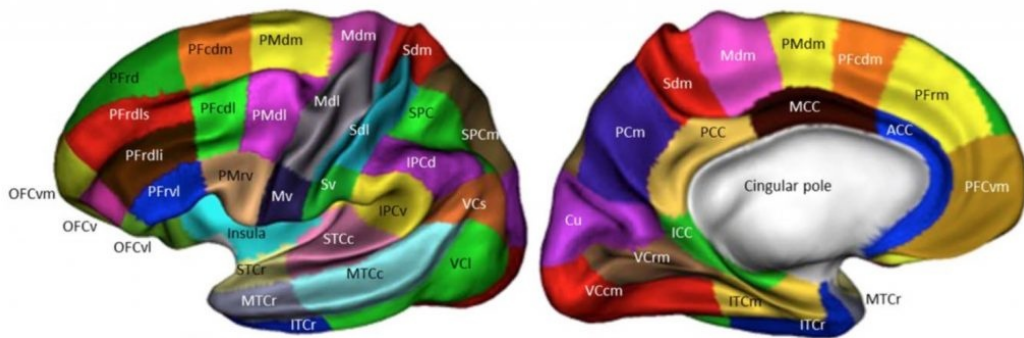


Time series of regional activity



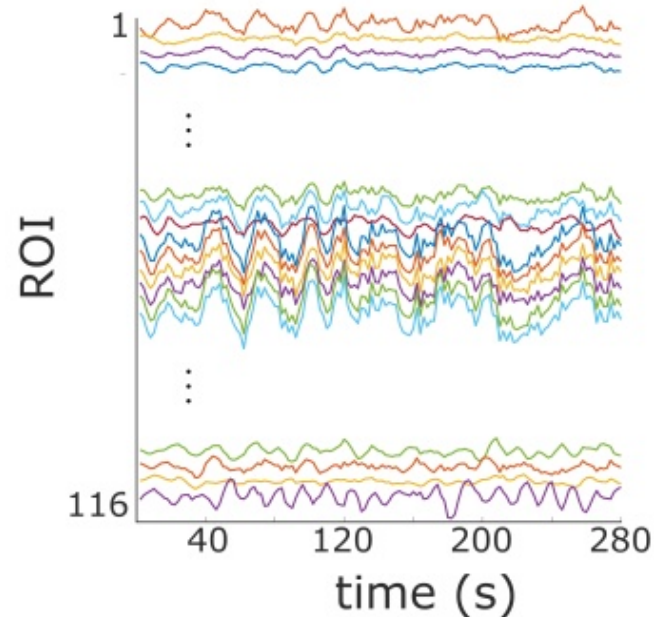
Example of Neuroimaging Data (fMRI)

Brain atlas



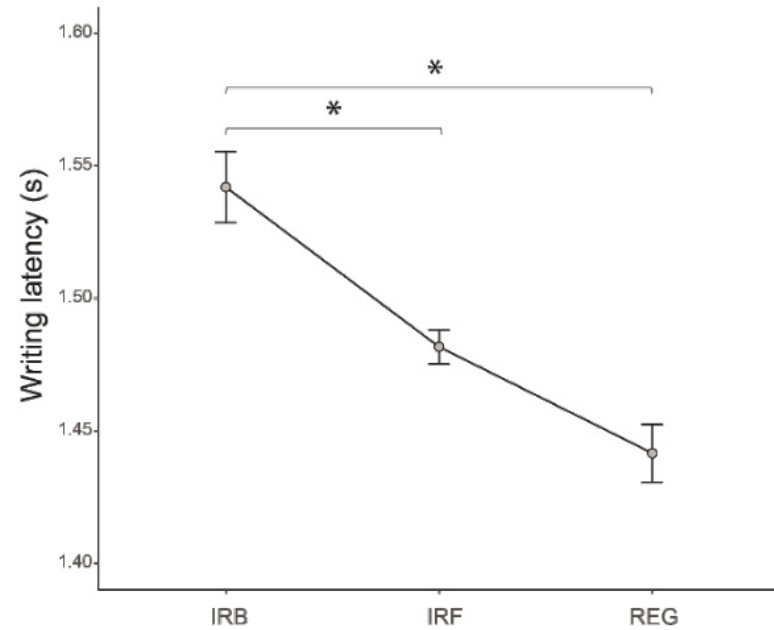
- Functional connectivity as a proxy for brain communication
- Are 2 regions correlated?
- Significantly correlated?

Time series of regional activity



Example of Handwriting Speed as Behavioral Measurement

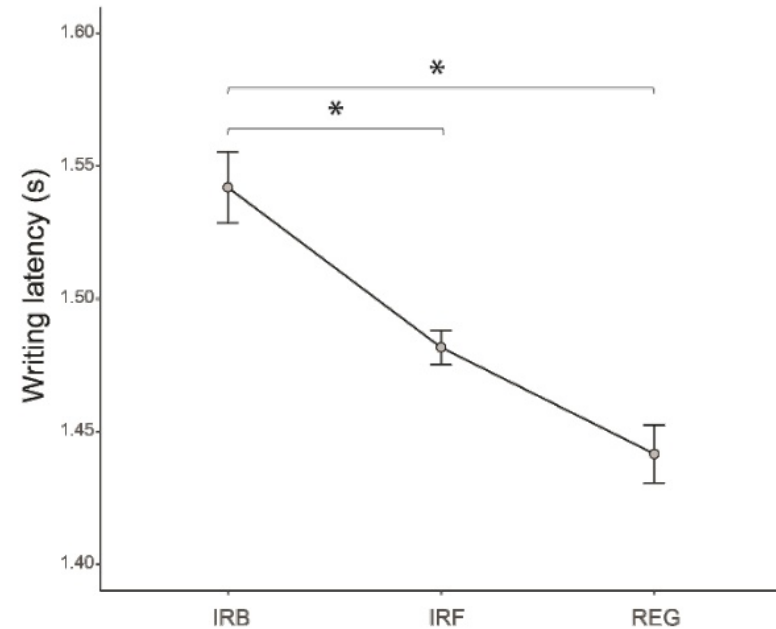
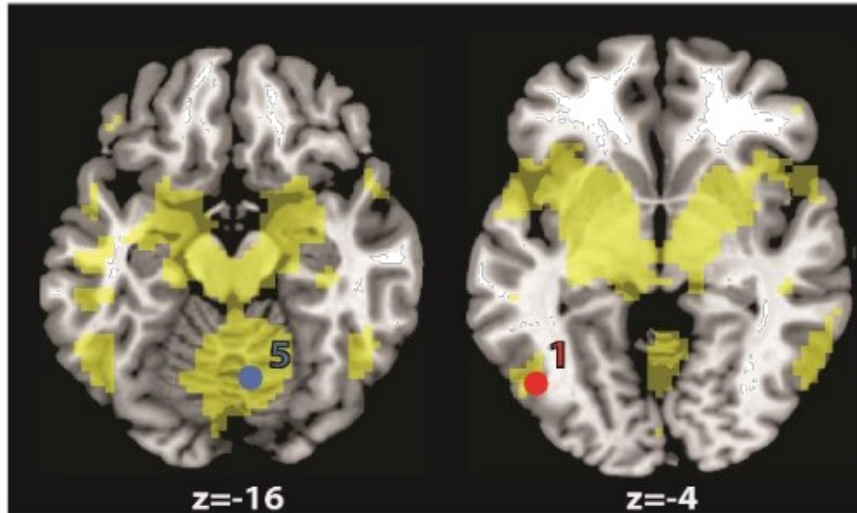
- IRB word: pharmacie
- IRF word: manuscrit
- REG word: other words



- Statistical difference across categories?

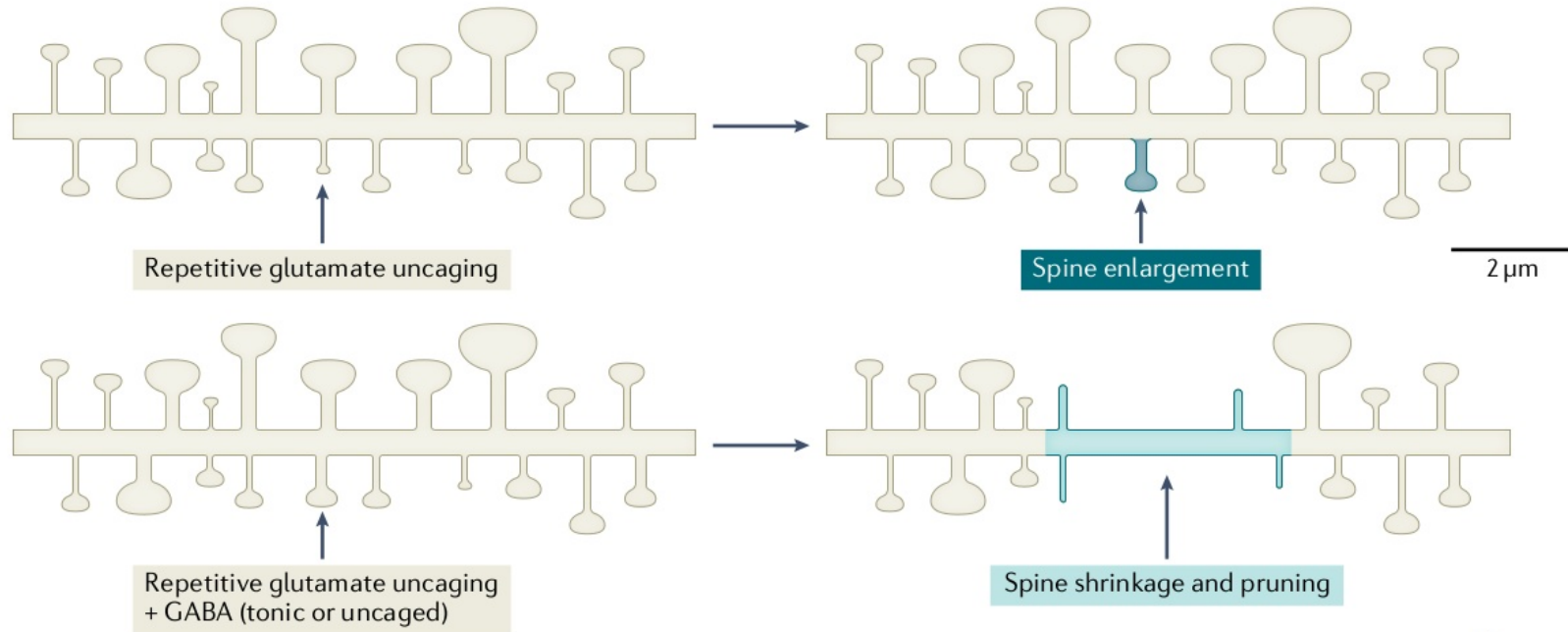
Example of Handwriting Speed as Behavioral Measurement

- IRB word: pharmacie
- IRF word: manuscrit
- REG word: other words

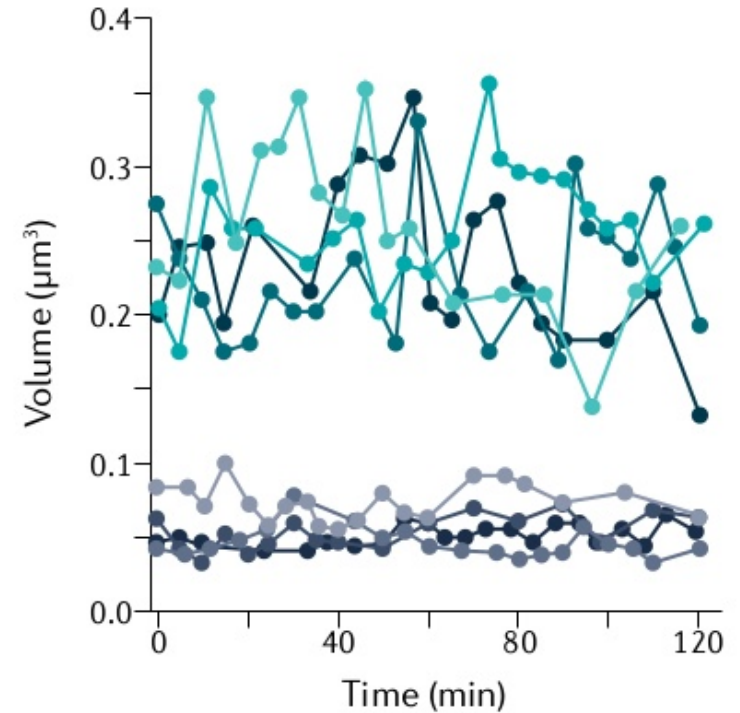
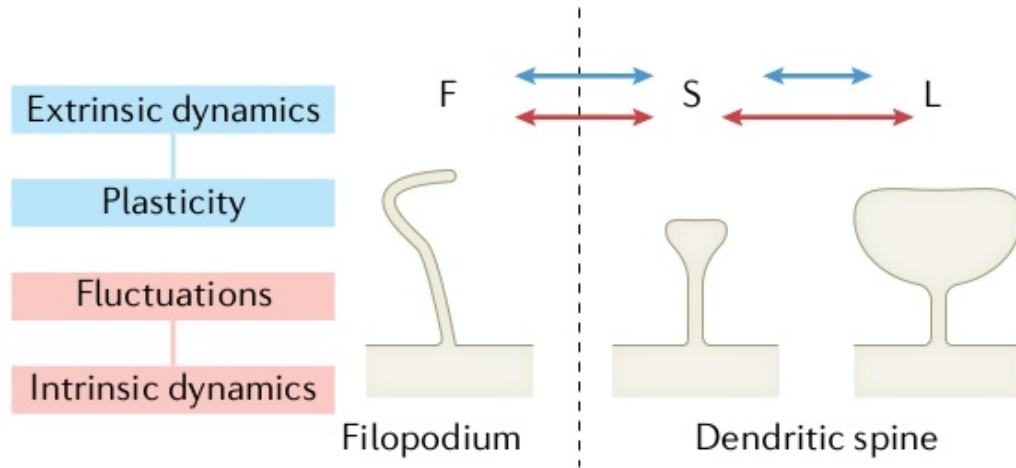


- Statistical difference across categories?
- Statistical relationships between two types of measurements?

Example of Synaptic Spine Fluctuations

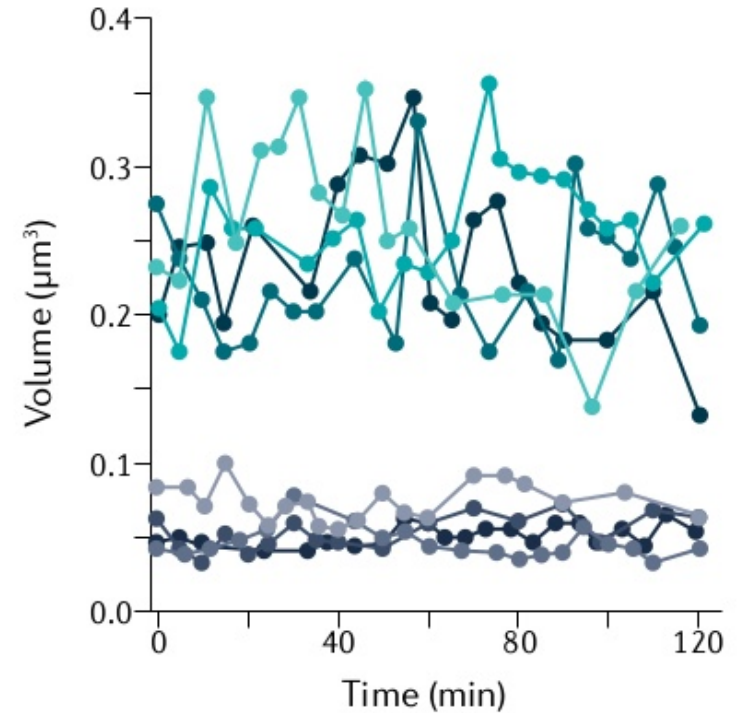
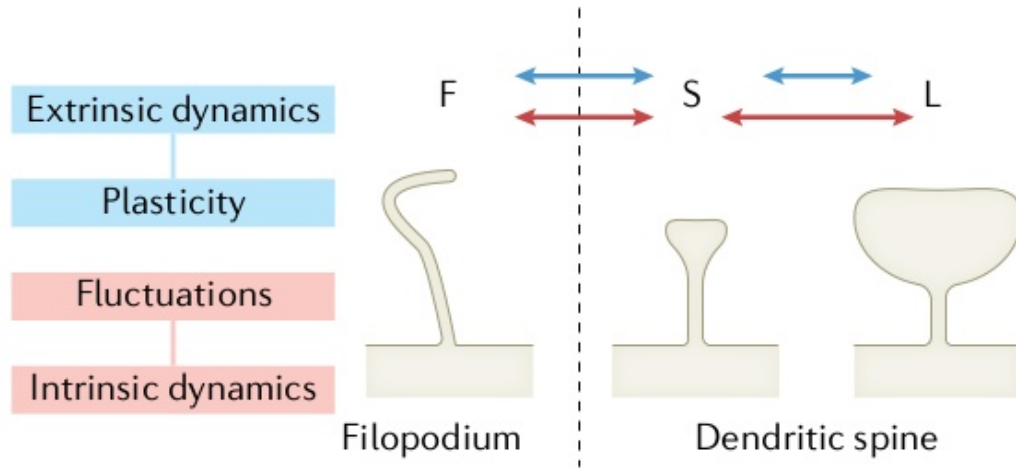


Example of Synaptic Spine Fluctuations

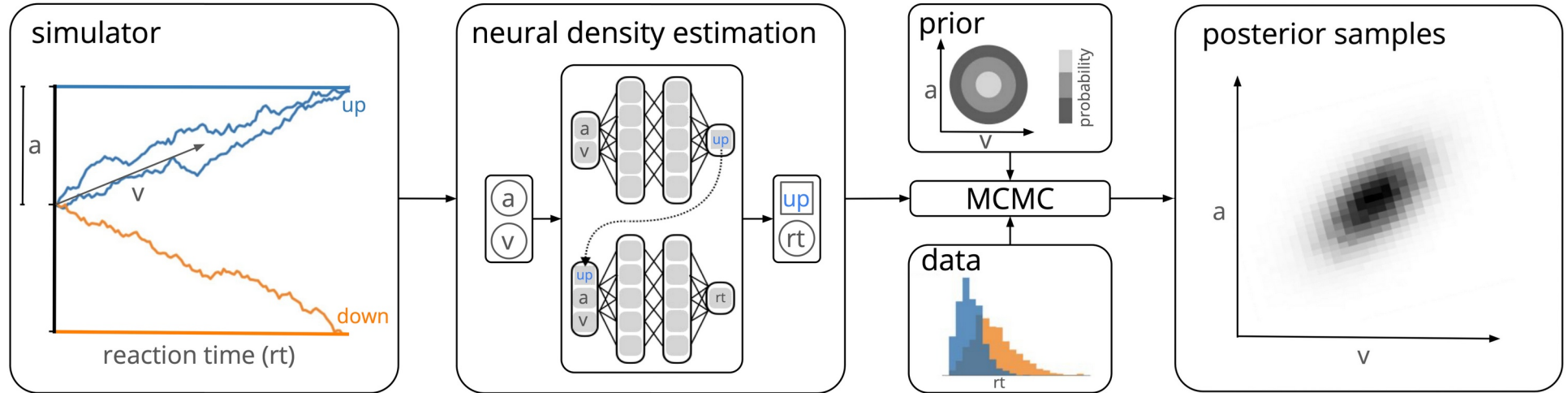


Example of Synaptic Spine Fluctuations

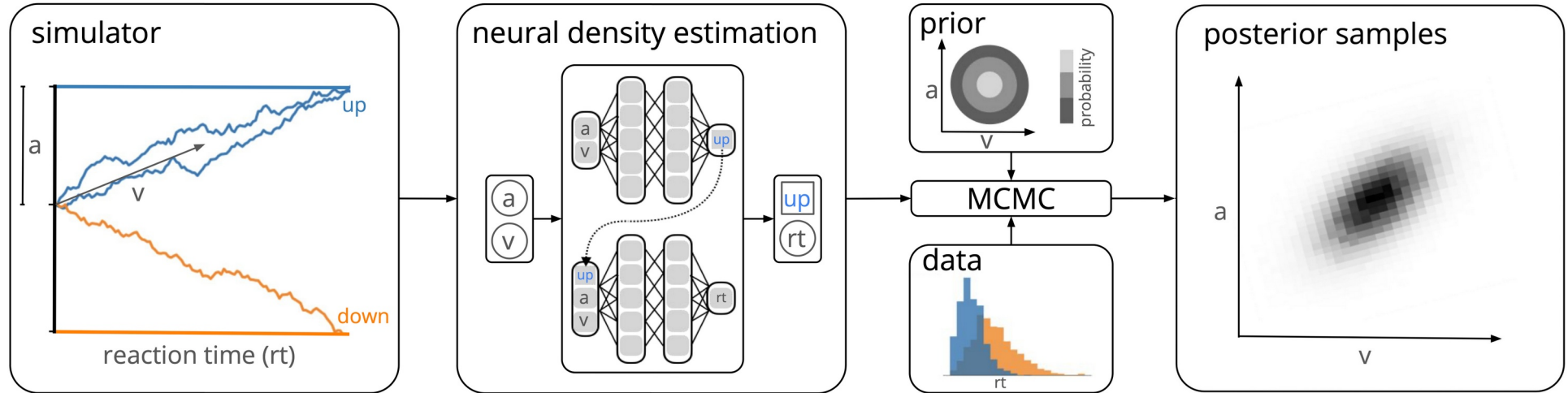
- How do fluctuations shape the distributions of spine volumes?



Fitting a Model on Data (e.g. Decision Making)



Fitting a Model on Data (e.g. Decision Making)

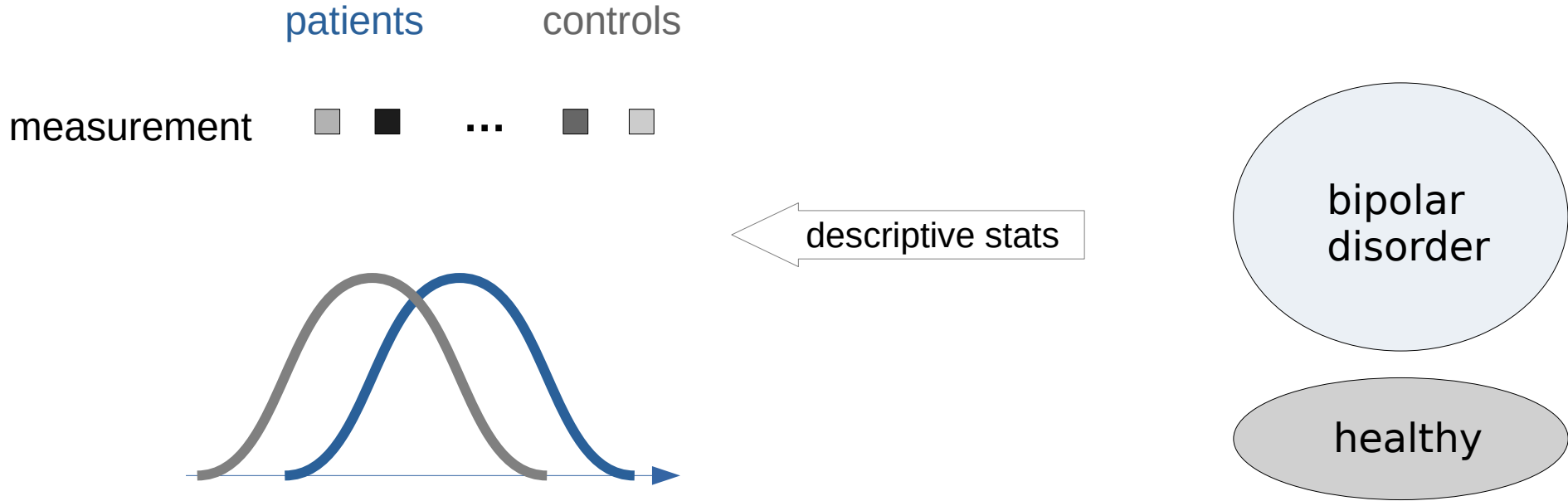


- How do model output depend on model inputs and parameters?
- What are the best parameters to fit empirical data?

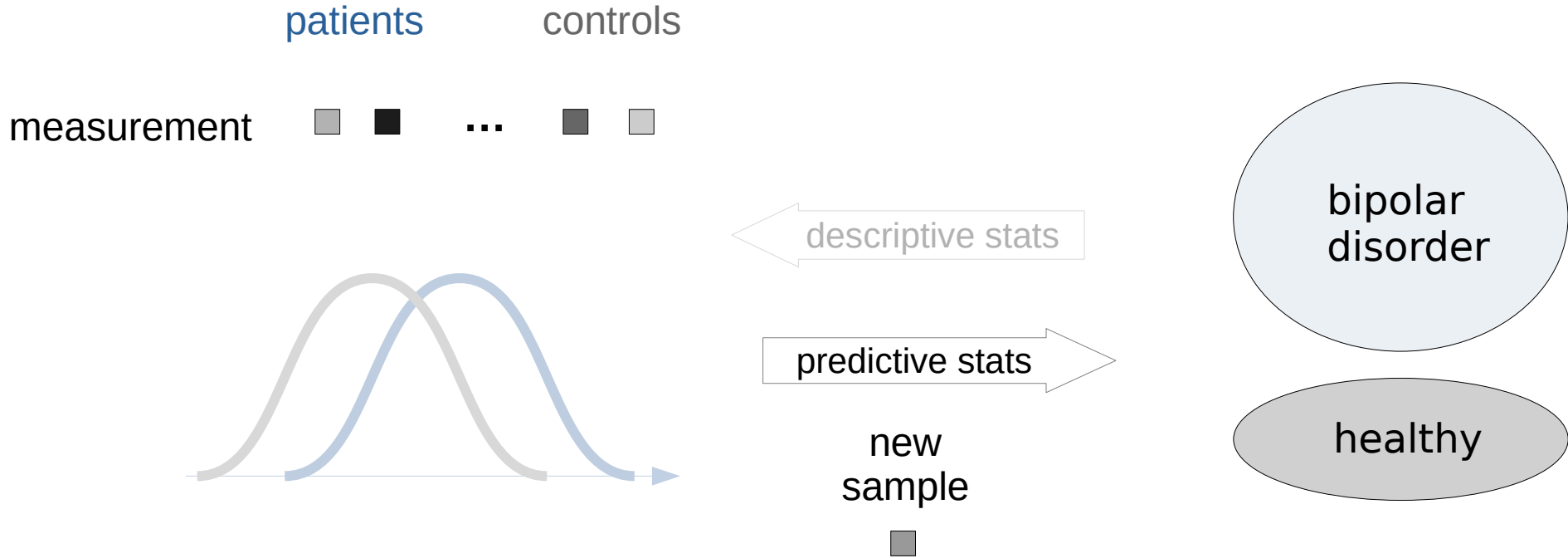
Statistical Analysis versus Machine Learning



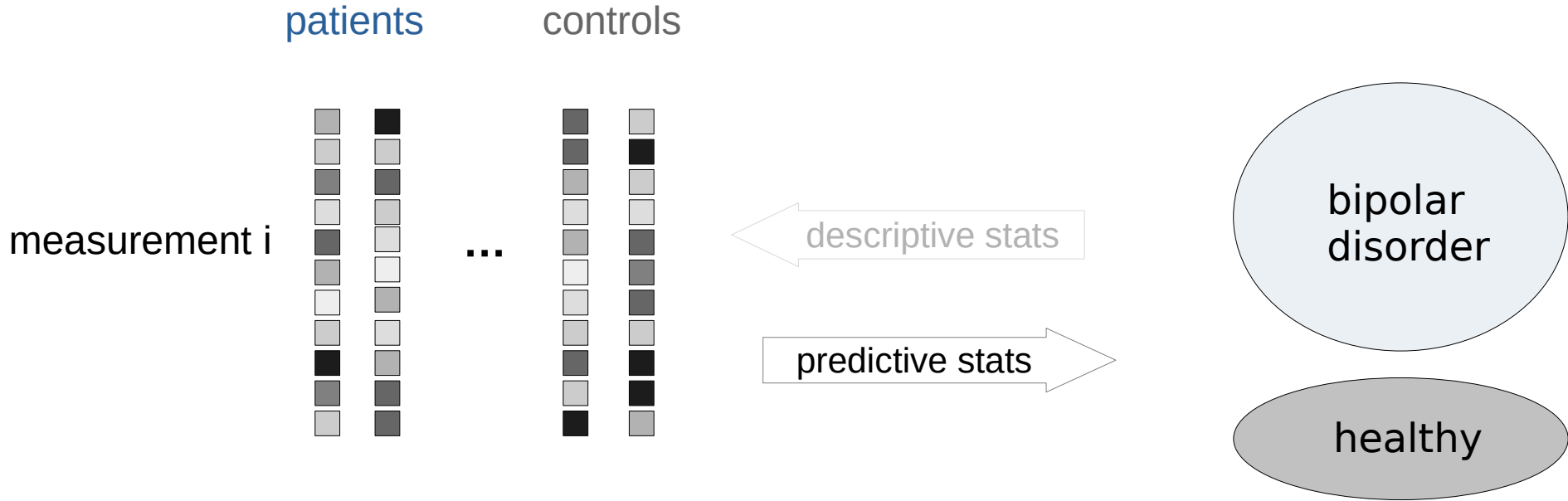
Statistical Analysis versus Machine Learning



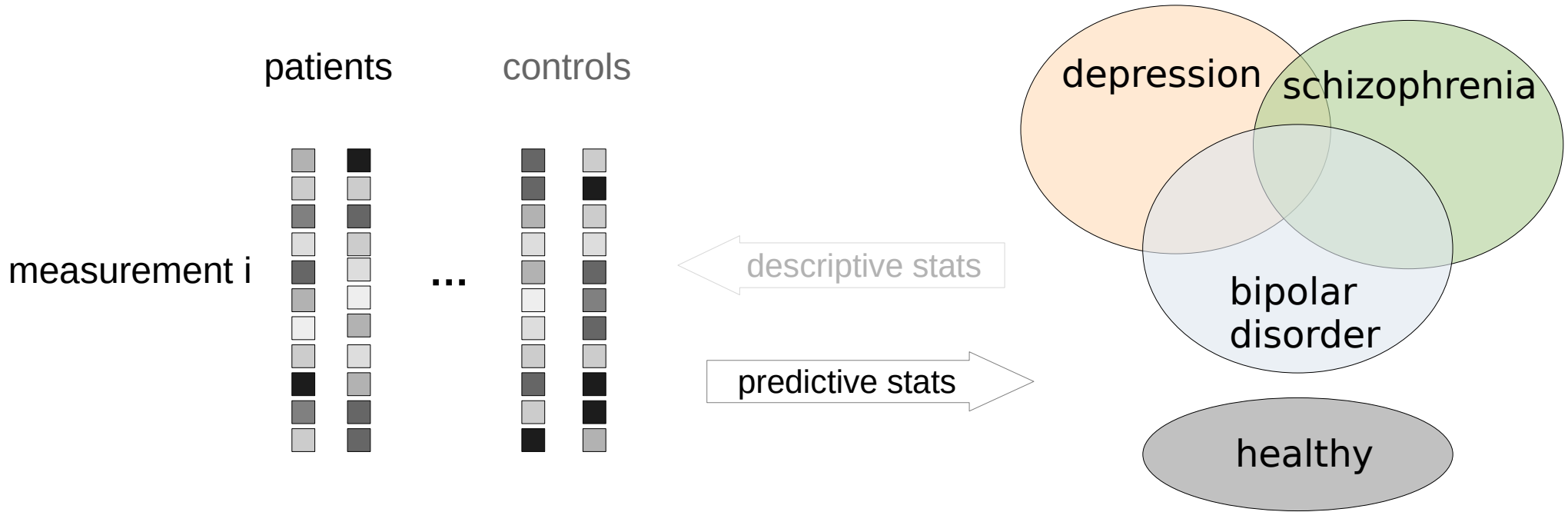
Statistical Analysis versus Machine Learning



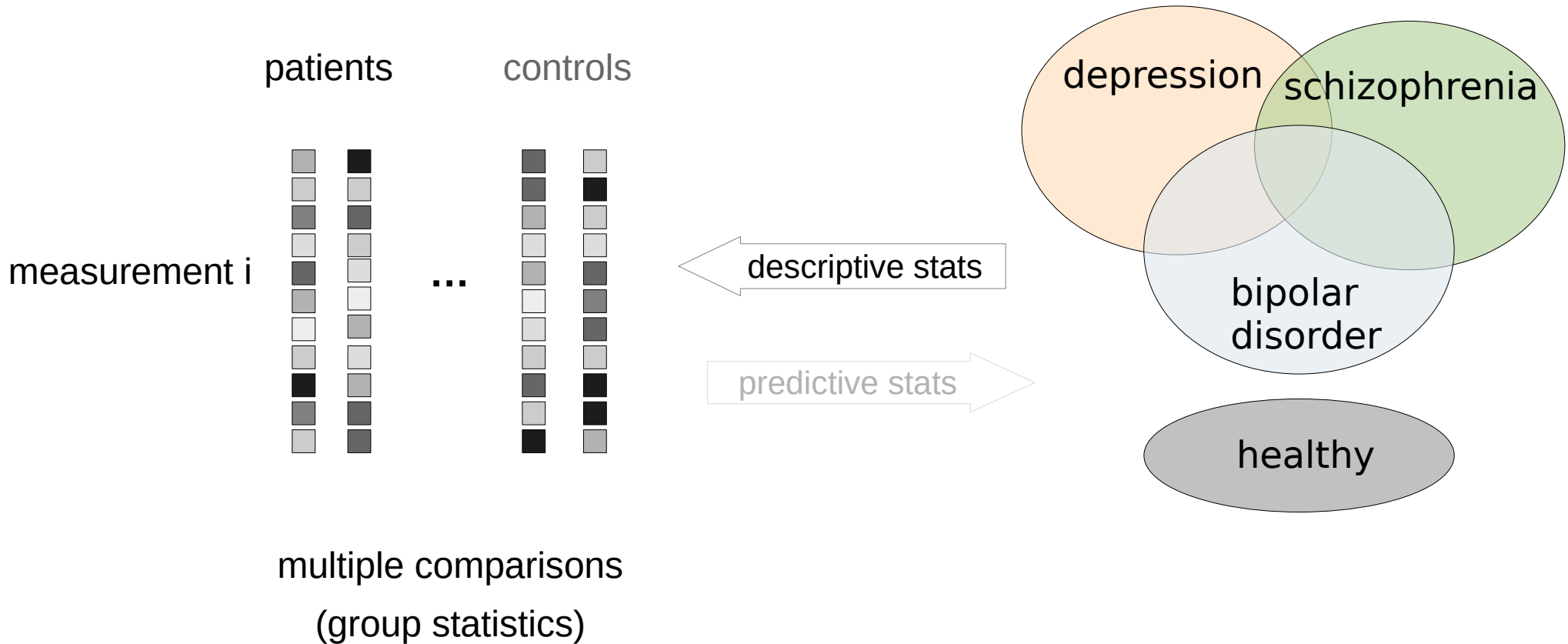
Statistical Analysis versus Machine Learning



Statistical Analysis versus Machine Learning



Statistical Analysis versus Machine Learning



Statistics in Python: Selection of Material

- Python packages:
 - scipy.stats
 - statsmodels
 - patsy (to use R formula)
- Repository: https://etulab.univ-amu.fr/gilson.m/compneuro_course/-/tree/main/stats
 - conda installation
 - environment installation: environment.yml file
 - jupyter notebooks
- References (with R language): <https://www.math.univ-toulouse.fr/~besse/Wikistat/>

Statistical Analysis in Python

- Probabilities, distributions
 - **univariate**, multivariate
 - **summary statistics**
- Parametric and non-parametric testing
- Regressions
- Bayesian inference

Observed Samples

- Probability P of random variable X

$$P(X) \stackrel{\text{def}}{=} P(X = x)$$

- Probability density function f

$$P(X \in [a, b]) = \int_a^b f(x) dx$$



Observed Samples

- Probability P of random variable X

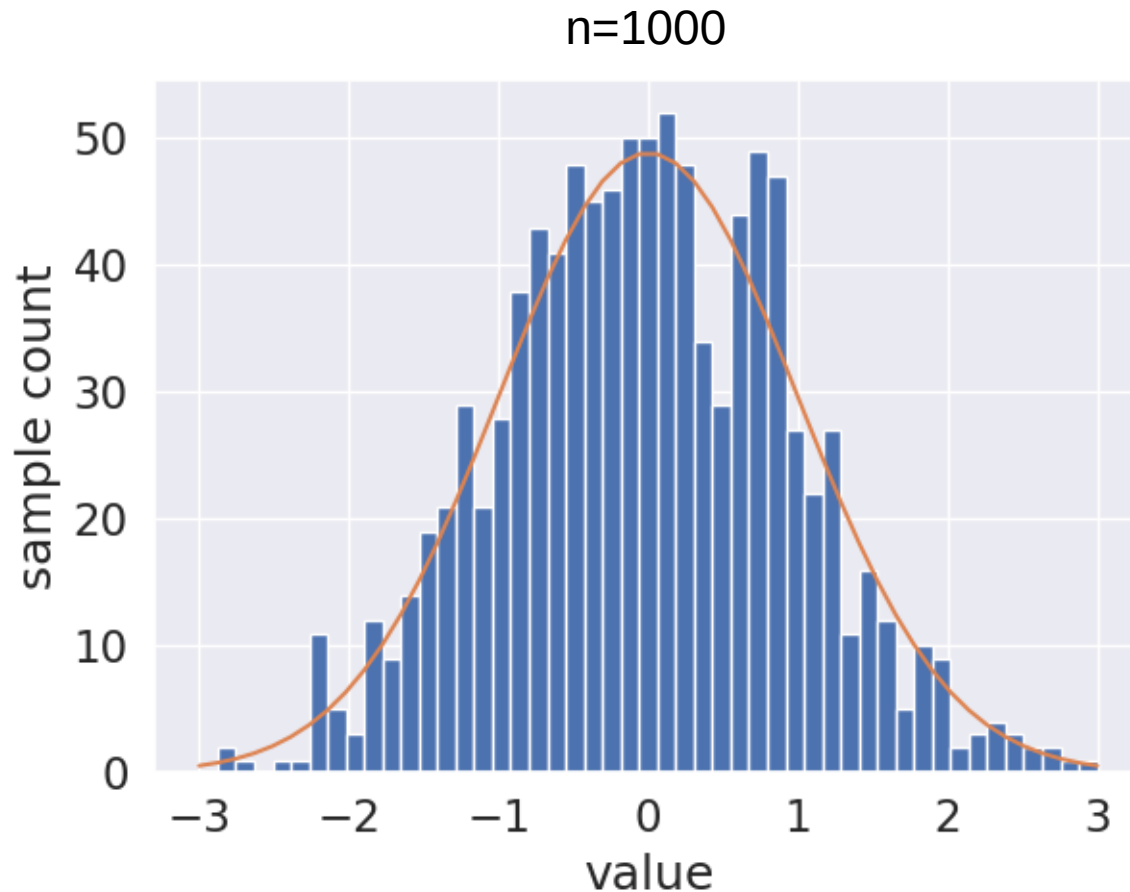
$$P(X) \stackrel{\text{def}}{=} P(X = x)$$

- Probability density function f

$$P(X \in [a, b]) = \int_a^b f(x) dx$$

- Histogram of samples

$$\sum_i 1[b_k < x_i < b_{k+1}]$$



Observed Samples

- Probability P of random variable X

$$P(X) \stackrel{\text{def}}{=} P(X = x)$$

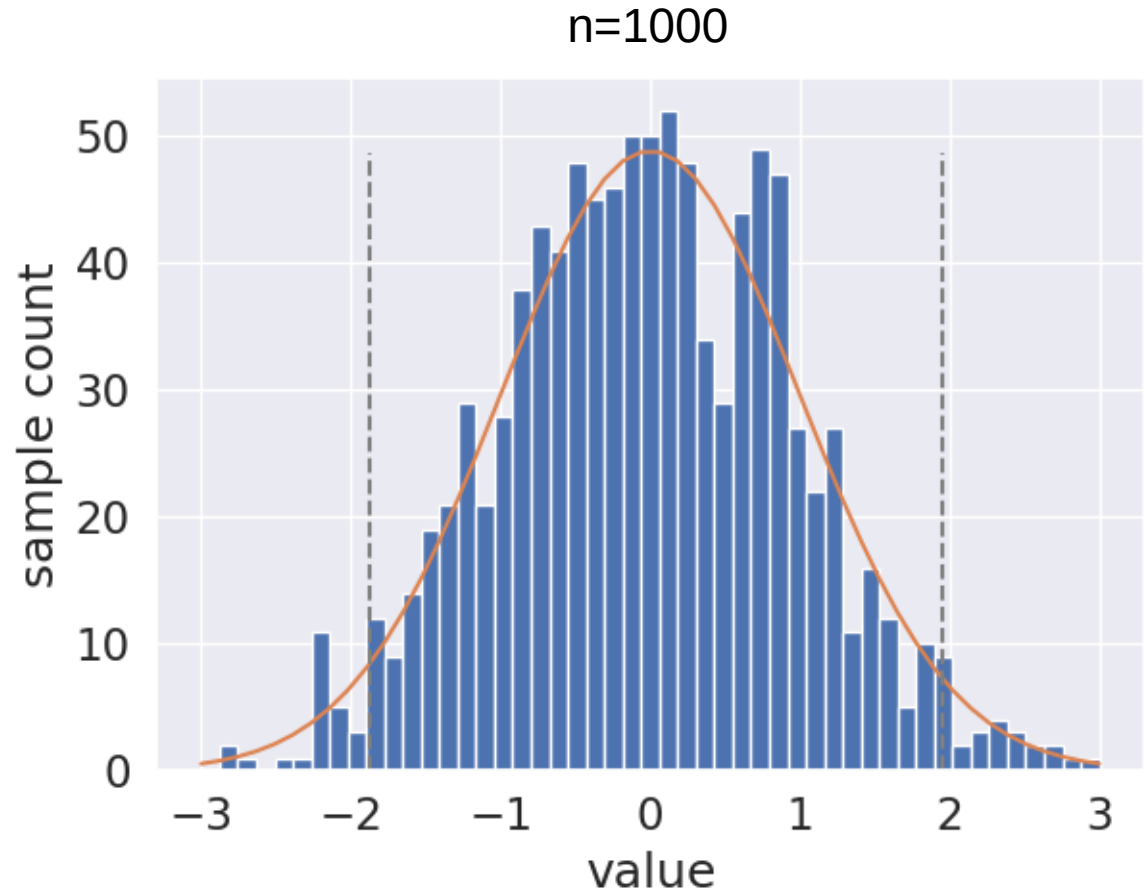
- Probability density function f

$$P(X \in [a, b]) = \int_a^b f(x) dx$$

- Histogram of samples

$$\sum_i 1[b_k < x_i < b_{k+1}]$$

- Percentiles top/bottom 2.5%
 - from samples (gray)



Observed Samples

- Probability P of random variable X

$$P(X) \stackrel{\text{def}}{=} P(X = x)$$

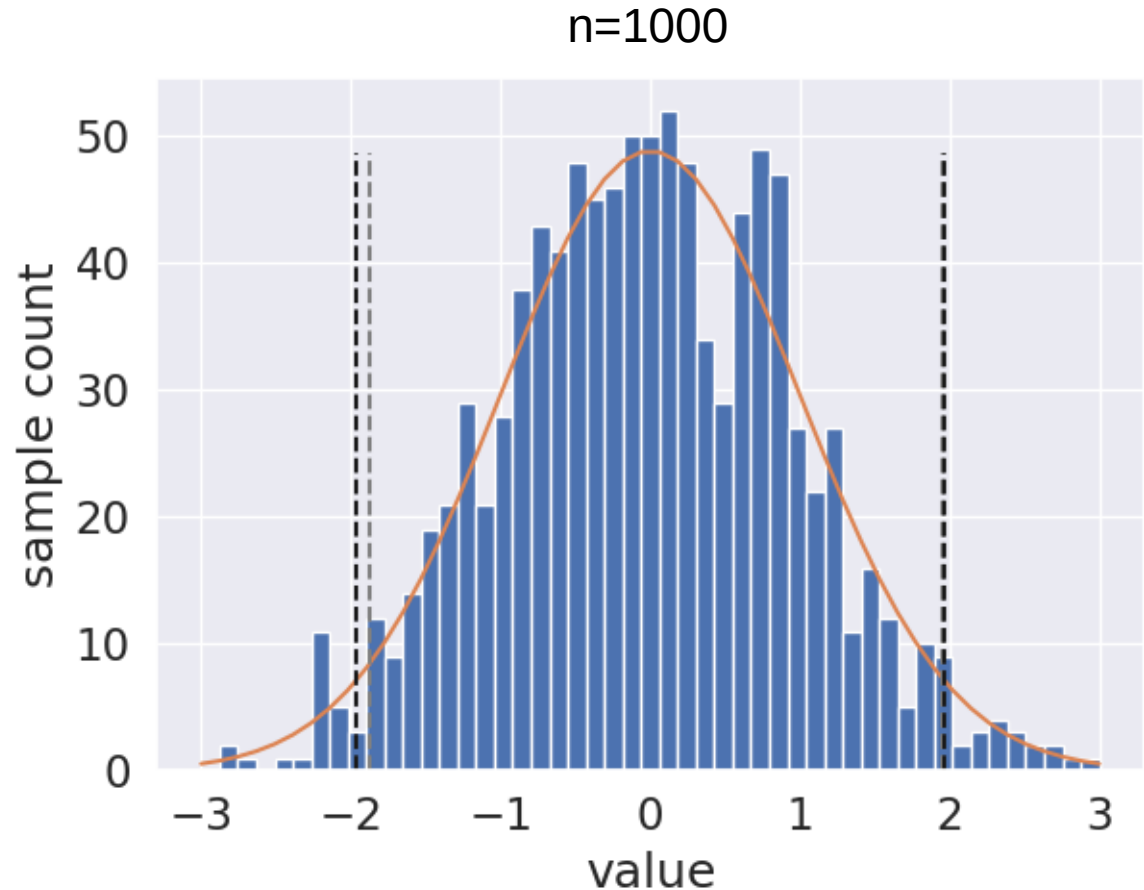
- Probability density function f

$$P(X \in [a, b]) = \int_a^b f(x) dx$$

- Histogram of samples

$$\sum_i 1[b_k < x_i < b_{k+1}]$$

- Percentiles top/bottom 2.5%
 - from samples (gray)
 - from theoretical curve (black)



Observed Samples

- Probability P of random variable X

$$P(X) \stackrel{\text{def}}{=} P(X = x)$$

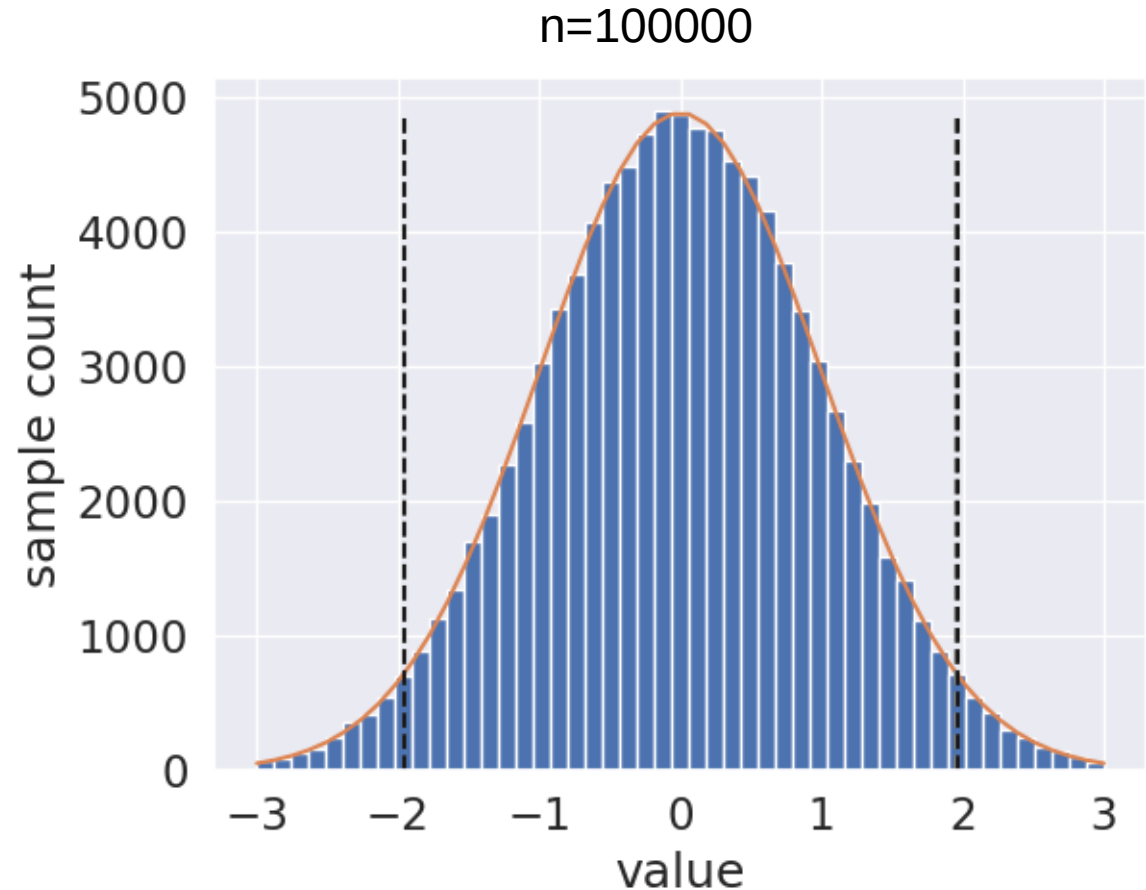
- Probability density function f

$$P(X \in [a, b]) = \int_a^b f(x) dx$$

- Histogram of samples

$$\sum_i 1[b_k < x_i < b_{k+1}]$$

- Percentiles top/bottom 2.5%
 - from samples (gray)
 - from theoretical curve (black)



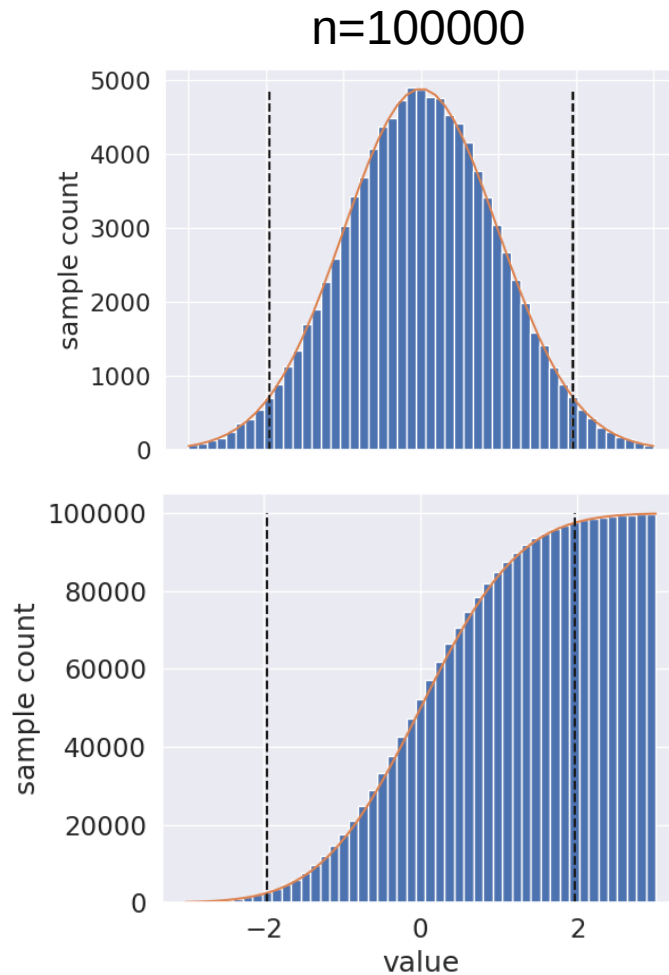
Observed Samples

- Probability density function (pdf)

$$f(x)$$

- Cumulative density function (cdf)

$$\phi(x) = \int_{-\infty}^x f(u) du$$



Observed Samples

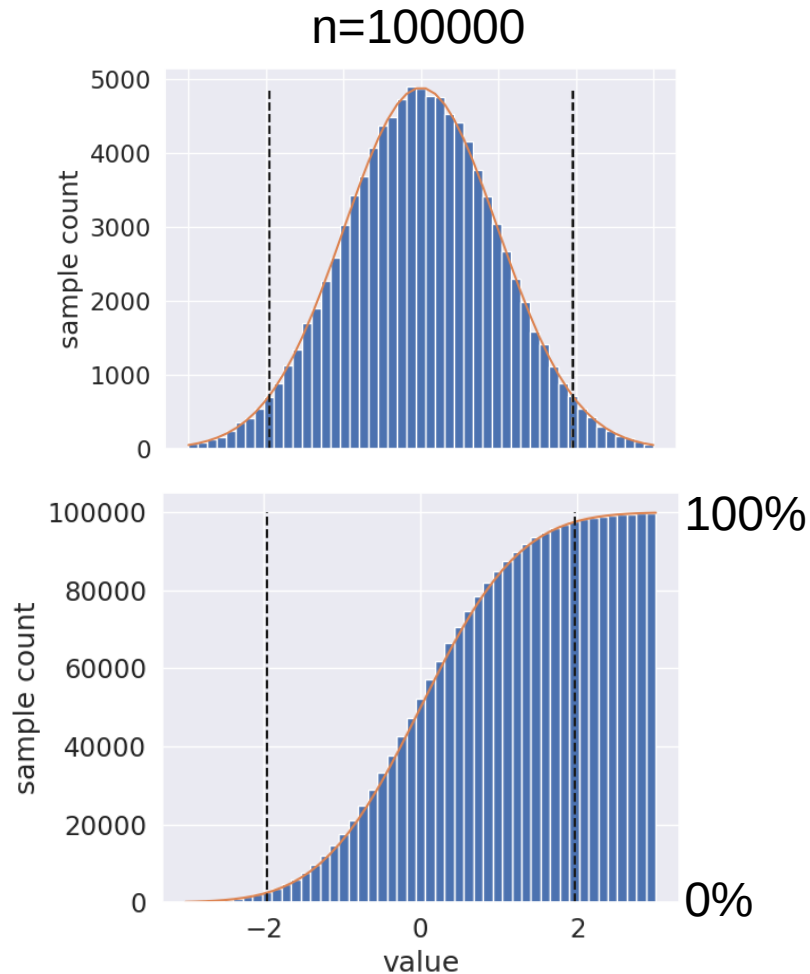
- Probability density function (pdf)

$$f(x)$$

- Cumulative density function (cdf)

$$\phi(x) = \int_{-\infty}^x f(u) du$$

- varies from 0 to 1



Observed Samples

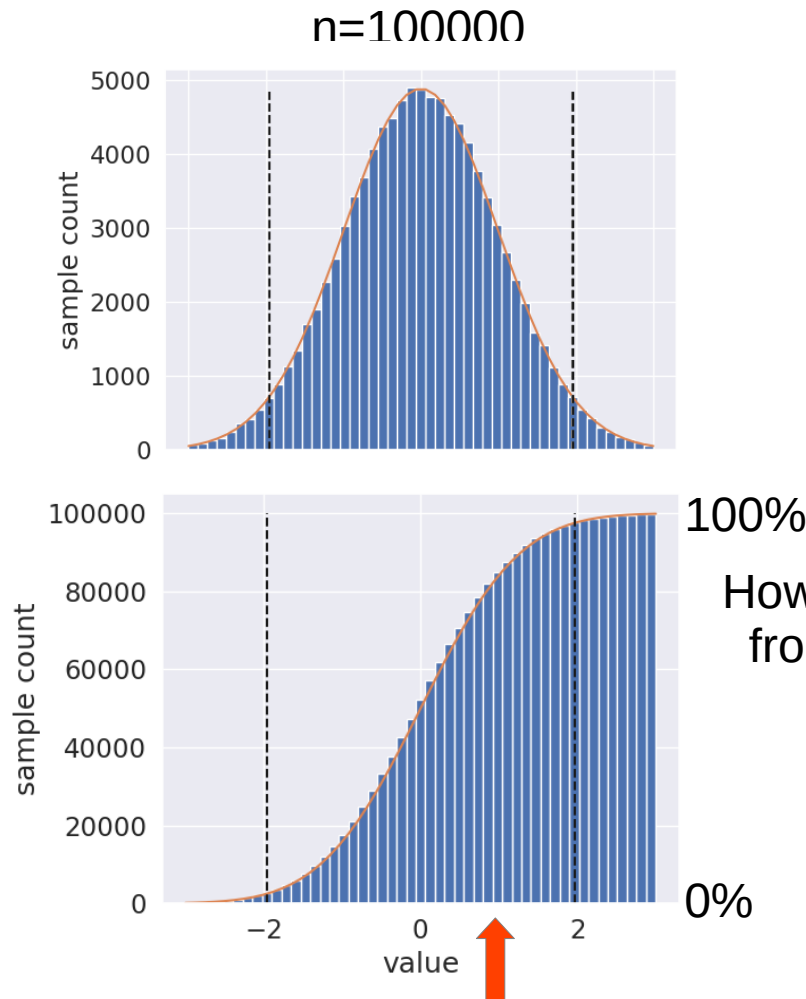
- Probability density function (pdf)

$$f(x)$$

- Cumulative density function (cdf)

$$\phi(x) = \int_{-\infty}^x f(u) du$$

- varies from 0 to 1



Observed Samples

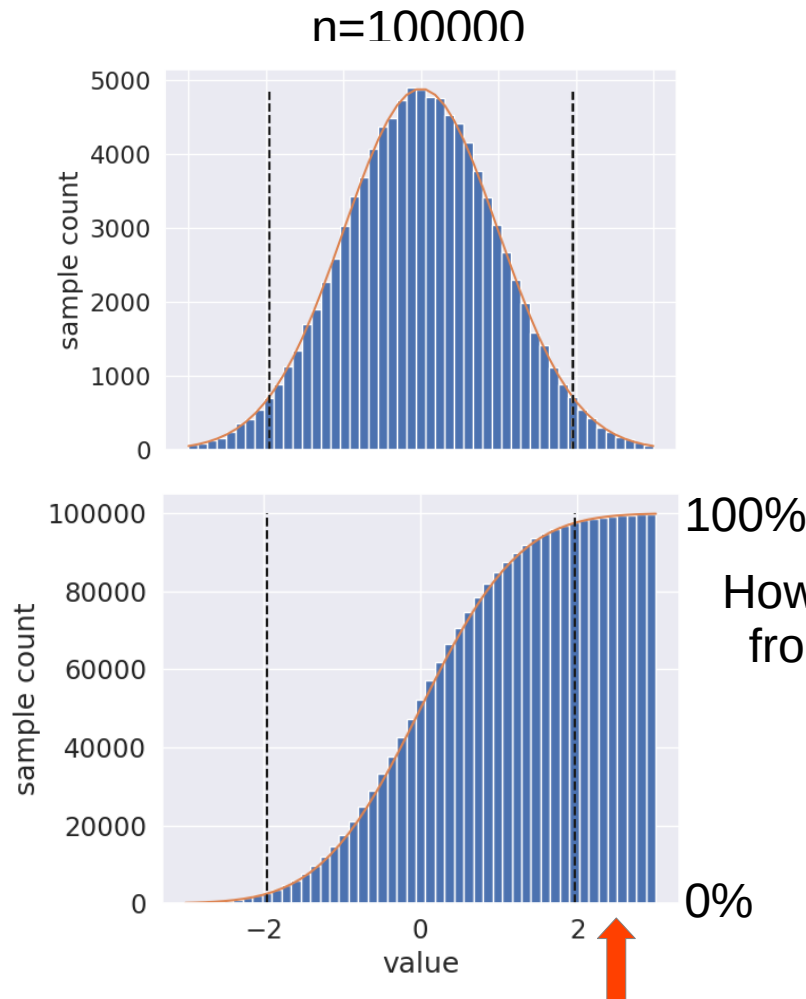
- Probability density function (pdf)

$$f(x)$$

- Cumulative density function (cdf)

$$\phi(x) = \int_{-\infty}^x f(u) du$$

- varies from 0 to 1



Observed Samples

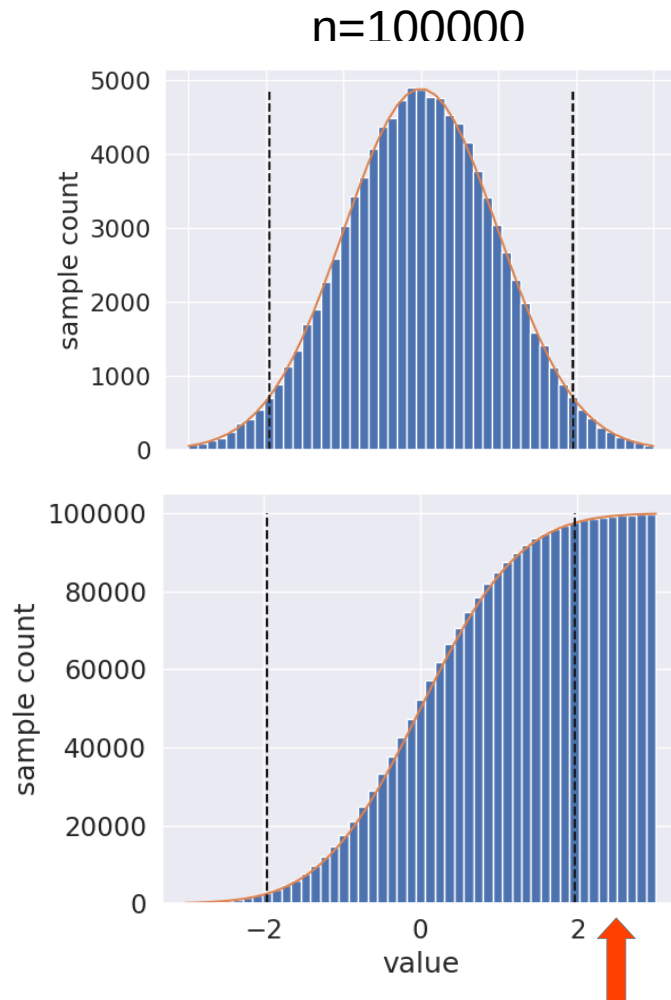
- Probability density function (pdf)

$$f(x)$$

- Cumulative density function (cdf)

$$\phi(x) = \int_{-\infty}^x f(u) du$$

- varies from 0 to 1



97.5%

How likely is a sample x
from this distribution?

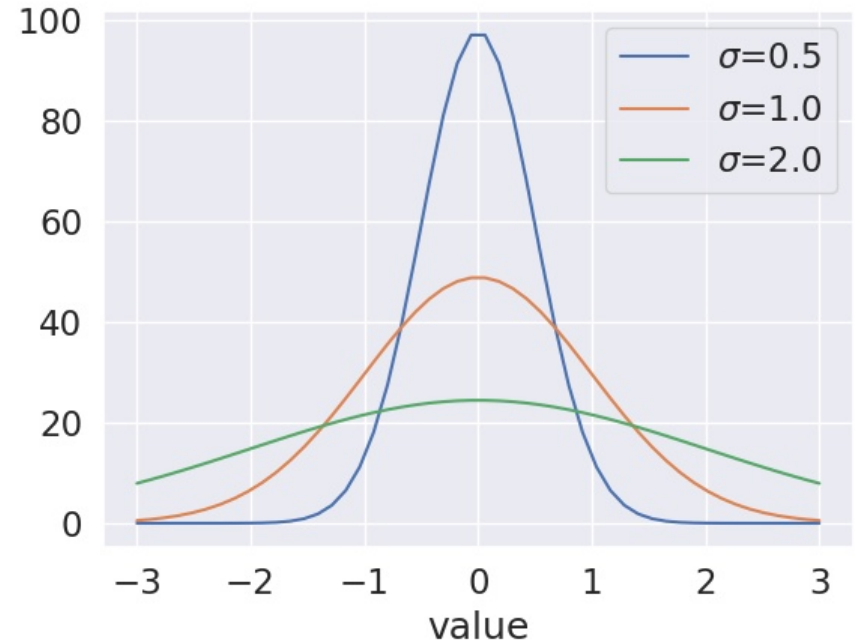
p-value = 0.05

$$p/2 < \phi(x) < 1 - p/2$$

2.5%

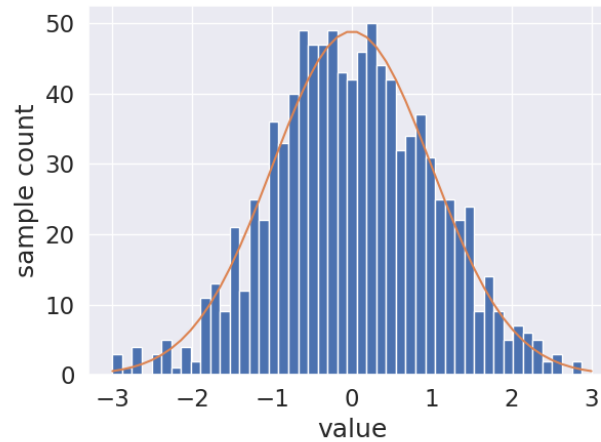
Observed Samples

- Sample x
 - mean μ and standard deviation σ
- z-score to normalize values $z = \frac{x - \mu}{\sigma}$
 - ♦ $|z| > 2$ corresponds to p-value < 0.05
 - ♦ $|z| > 3$ corresponds to p-value < 0.002



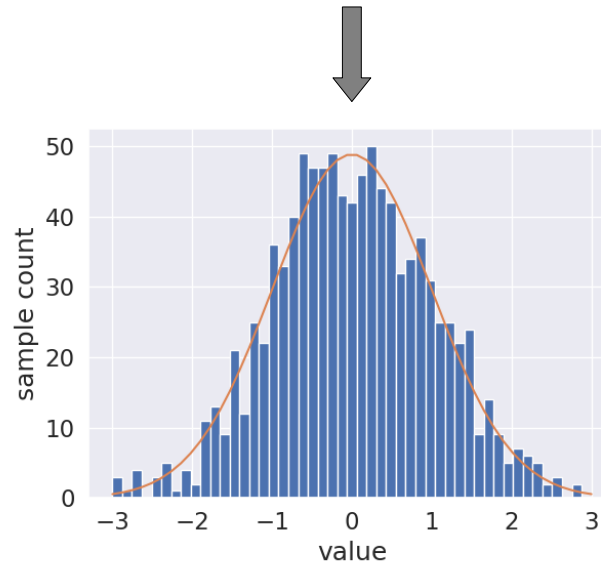
Summary Statistics and Empirical Estimates

- PDF $f(x)$



Summary Statistics and Empirical Estimates

- PDF $f(x)$
- Mean $E[X] = \int x f(x) dx = \mu$



Summary Statistics and Empirical Estimates

- PDF

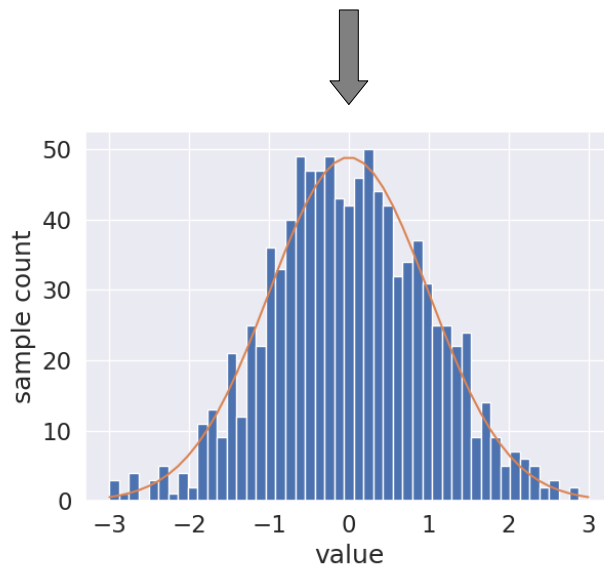
$$f(x)$$

sample estimate

- Mean

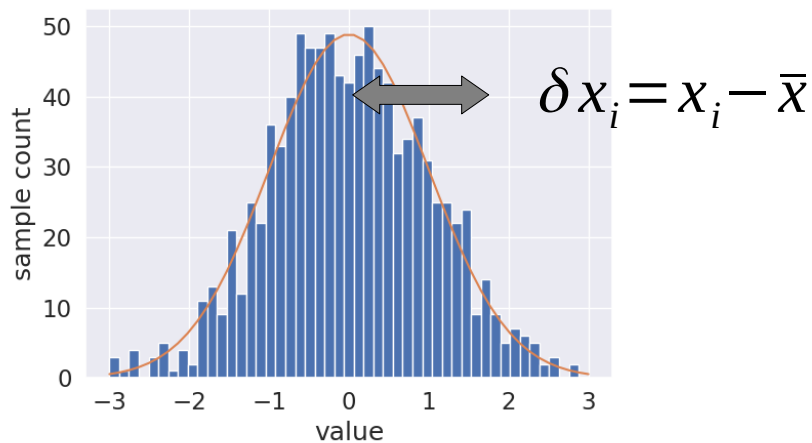
$$E[X] = \int x f(x) dx = \mu$$

$$\bar{x} = \frac{1}{n} \sum_i x_i$$



Summary Statistics and Empirical Estimates

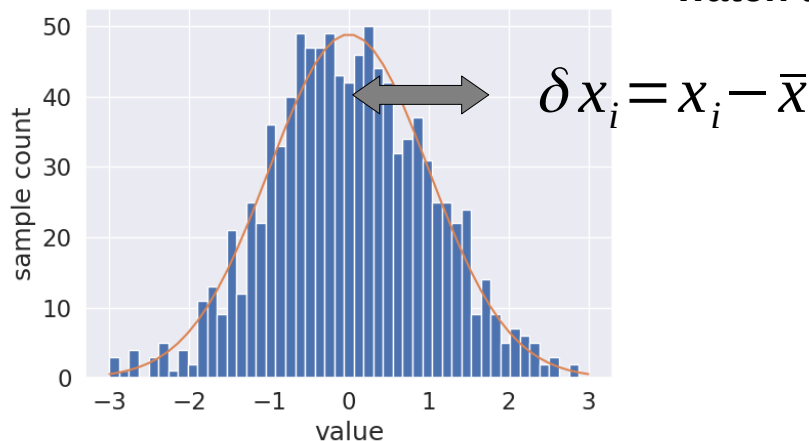
- PDF $f(x)$ sample estimate
- Mean $E[X] = \int x f(x) dx = \mu$ $\bar{x} = \frac{1}{n} \sum_i x_i$
- Variance $E[(X - \mu)^2] = \int (x - \mu)^2 f(x) dx = \sigma^2$ $\bar{s}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$



Summary Statistics and Empirical Estimates

- PDF $f(x)$ sample estimate
- Mean $E[X] = \int x f(x) dx = \mu$ $\bar{x} = \frac{1}{n} \sum_i x_i$
- Variance $E[(X - \mu)^2] = \int (x - \mu)^2 f(x) dx = \sigma^2$ $\bar{s}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$

watch the bias here!



Summary Statistics and Empirical Estimates

- Central moments
 - order 1 (always zero) $E[(X - \mu)]$
 - order 2 (variance) $E[(X - \mu)^2]$
 - order k $E[(X - \mu)^k]$

Summary Statistics and Empirical Estimates

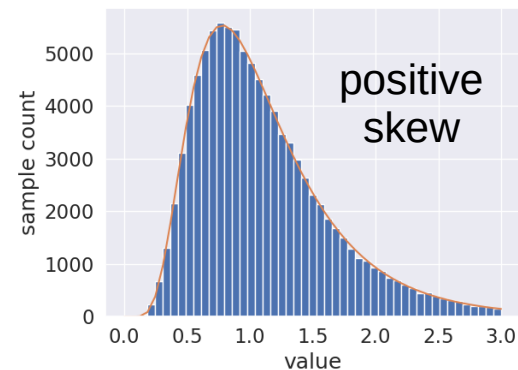
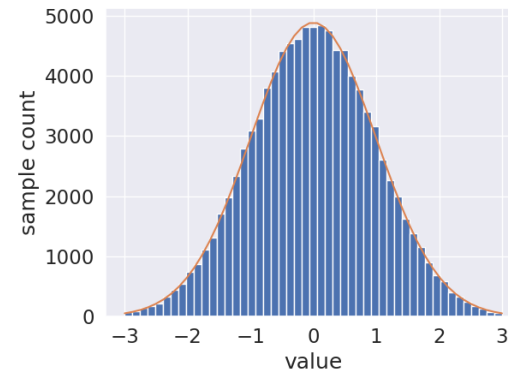
- Central moments with generating function $E[\exp(\xi X)] = \int f(x) e^{\xi x} dx = \Phi_X(\xi)$
 - order 1 (always zero) $E[(X - \mu)] = \dot{\Phi}_{X-\mu}(0)$
 - order 2 (variance) $E[(X - \mu)^2] = \ddot{\Phi}_{X-\mu}(0)$
 - order k $E[(X - \mu)^k] = \Phi^{(n)}_{X-\mu}(0) = m_k$

Summary Statistics and Empirical Estimates

- Mean $E[X] = \mu$
- Variance $E[(X - \mu)^2] = \sigma^2$
- Skew $E\left[\frac{(X - \mu)^3}{\sigma^3}\right] = \frac{m_3}{\sigma^3}$
- Kurtosis $E\left[\frac{(X - \mu)^4}{\sigma^4}\right] = \frac{m_4}{\sigma^4}$

Summary Statistics and Empirical Estimates

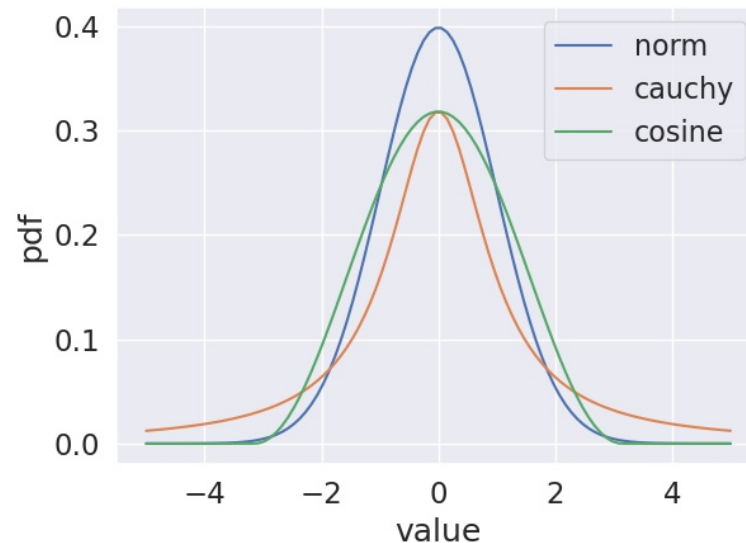
- Mean $E[X] = \mu$
- Variance $E[(X - \mu)^2] = \sigma^2$
- **Skew** $E\left[\frac{(X - \mu)^3}{\sigma^3}\right] = \frac{m_3}{\sigma^3}$
- Kurtosis $E\left[\frac{(X - \mu)^4}{\sigma^4}\right] = \frac{m_4}{\sigma^4}$



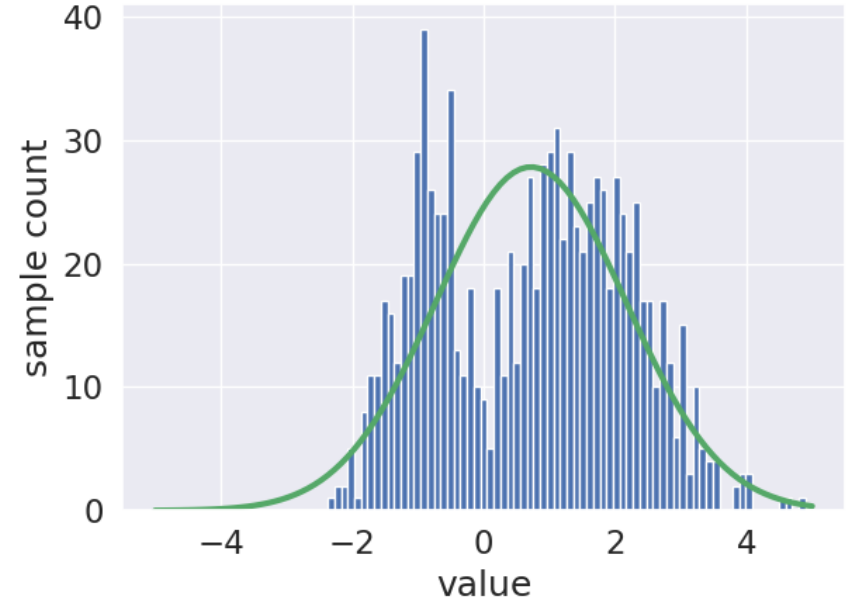
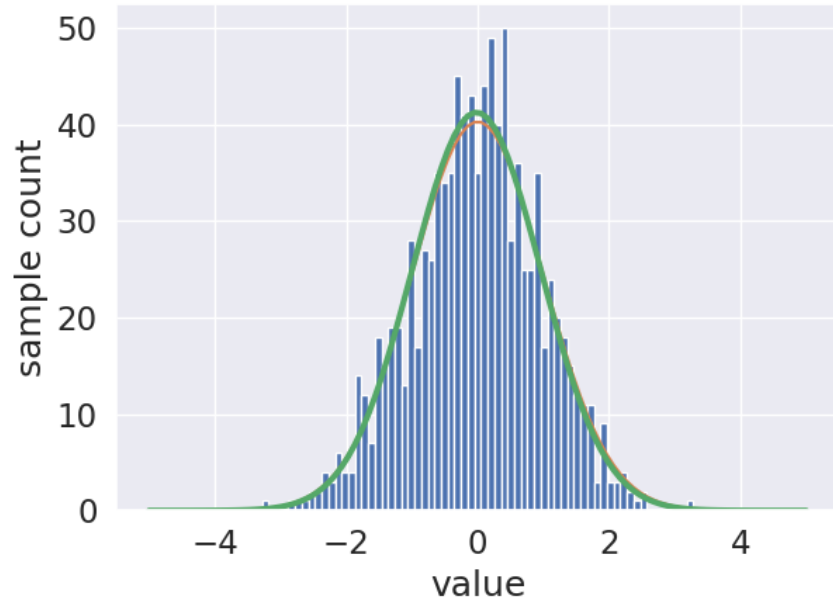
Summary Statistics and Empirical Estimates

- Mean $E[X] = \mu$
- Variance $E[(X - \mu)^2] = \sigma^2$
- Skew $E\left[\frac{(X - \mu)^3}{\sigma^3}\right] = \frac{m_3}{\sigma^3}$
- **Kurtosis** $E\left[\frac{(X - \mu)^4}{\sigma^4}\right] = \frac{m_4}{\sigma^4}$

- normal: 3.0
- Cauchy: ∞
- Cosine: 2.4



Unimodal versus Multimodal Distributions



PDF RECONSTRUCTION FROM MEAN, VAR, ETC.

Central Limit Theorem (+ Law of Large Numbers)

- For n independent random variables X_i following a same distribution $P(X_i) = P(X)$
- This theorem is about the sample mean $\bar{x} = \frac{1}{n} \sum_i x_i$
- It follows a normal distribution with
 - mean $\mu = E[X]$
 - variance $\frac{\sigma^2}{n}$ where $\sigma^2 = E[(X - \mu)^2]$
- Importantly, the underlying distribution $P(X)$ doesn't have to be normal!

Bessel's Correction for Variance Estimate

- Assumption: **the real mean μ is not known**, so we have to rely on the sample mean \bar{x}

$$\sigma^2 \simeq \frac{1}{n} \sum_i (x_i - \mu)^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 + \frac{1}{n} \sum_i (\bar{x} - \mu)^2$$

Bessel's Correction for Variance Estimate

- Assumption: **the real mean μ is not known**, so we have to rely on the sample mean \bar{x}

$$\begin{aligned}\sigma^2 &\simeq \frac{1}{n} \sum_i (x_i - \mu)^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 + \underbrace{\frac{1}{n} \sum_i (\bar{x} - \mu)^2}_{= (\bar{x} - \mu)^2} = \frac{\sigma^2}{n} \quad (\text{according to CLT})\end{aligned}$$

Bessel's Correction for Variance Estimate

- Assumption: **the real mean μ is not known**, so we have to rely on the sample mean \bar{x}

$$\begin{aligned}\sigma^2 &\simeq \frac{1}{n} \sum_i (x_i - \mu)^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 + \underbrace{\frac{1}{n} \sum_i (\bar{x} - \mu)^2}_{= (\bar{x} - \mu)^2 = \frac{\sigma^2}{n}} \quad (\text{according to CLT})\end{aligned}$$

- So we obtain $\frac{n-1}{n} \sigma^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$

Bessel's Correction for Variance Estimate

- Assumption: **the real mean μ is not known**, so we have to rely on the sample mean \bar{x}

$$\begin{aligned}\sigma^2 &\simeq \frac{1}{n} \sum_i (x_i - \mu)^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 + \underbrace{\frac{1}{n} \sum_i (\bar{x} - \mu)^2}_{= (\bar{x} - \mu)^2 = \frac{\sigma^2}{n}} \quad (\text{according to CLT})\end{aligned}$$

- And finally
$$\sigma^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = \bar{s}^2$$

Statistical Analysis in Python

- Probabilities, distributions
- **Parametric and non-parametric testing**
 - **test if observed values are off zero**
 - **test normality of distribution**
 - **compare distributions**
- Regression, fixed effect model, mixed effect model
- Bayesian inference

Statistical Testing

- Are the observed values coming from a distribution not centered at 0?

Statistical Testing

- Are the observed values coming from a distribution not centered at 0?
- The null hypothesis H_0 is that the observed values being distributed around $\mu=0$
- If H_0 is true, from CLT the sample mean is normally distributed $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
 - here, the mean and variance of the underlying distribution are μ and σ^2

Statistical Testing

- Are the observed values coming from a distribution not centered at 0?
- The null hypothesis H_0 is that the observed values being distributed around $\mu=0$
- If H_0 is true, from CLT the sample mean is normally distributed $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
 - here, the mean and variance of the underlying distribution are μ and σ^2
- The t statistic corresponds to the probability of the sample mean being close to the hypothesized mean of the underlying distribution

$$t = \frac{\bar{x} - \mu}{\bar{s}/\sqrt{n}} \quad \text{with the sample variance estimate } \bar{s}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \simeq \sigma^2$$

Statistical Testing

- Are the observed values coming from a distribution not centered at 0?
- The null hypothesis H_0 is that the observed values being distributed around $\mu=0$
- If H_0 is true, from CLT the sample mean is normally distributed $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
 - here, the mean and variance of the underlying distribution are μ and σ^2
- The t statistic corresponds to the probability of the sample mean being close to the hypothesized mean of the underlying distribution

$$t = \frac{\bar{x} - \mu}{\bar{s}/\sqrt{n}} \quad \text{with the sample variance estimate } \bar{s}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \simeq \sigma^2$$

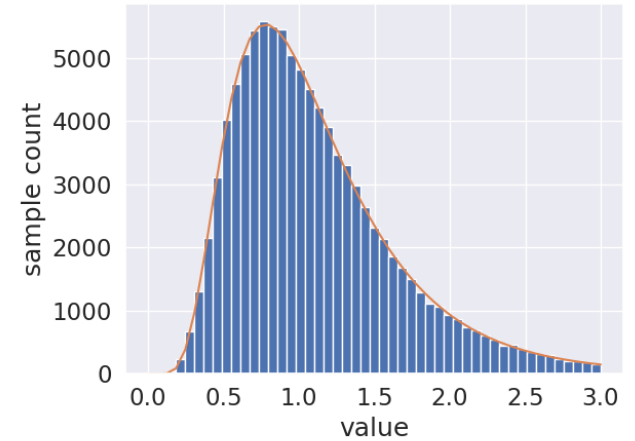
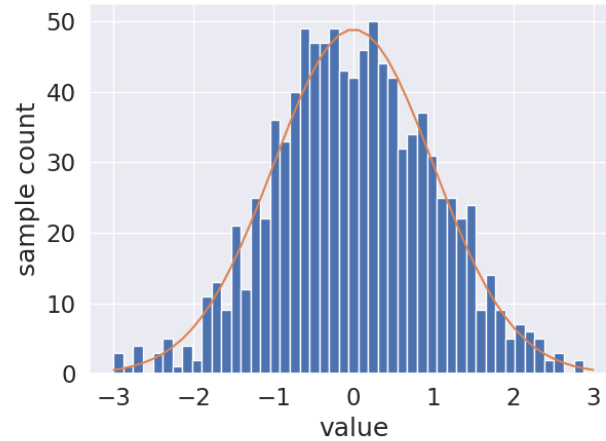
- the p-value corresponds to the tail(s) of the probability distribution above this value (with respect to the mean, possibly including symmetry)

Statistical Testing

- Is a distribution normal?

Statistical Testing

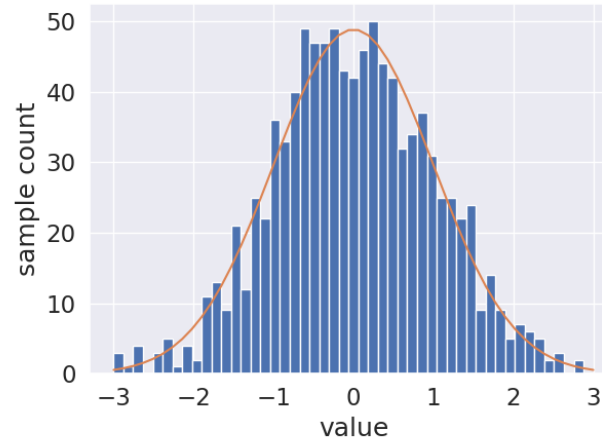
- Is a distribution normal?



Statistical Testing

- Is a distribution normal?

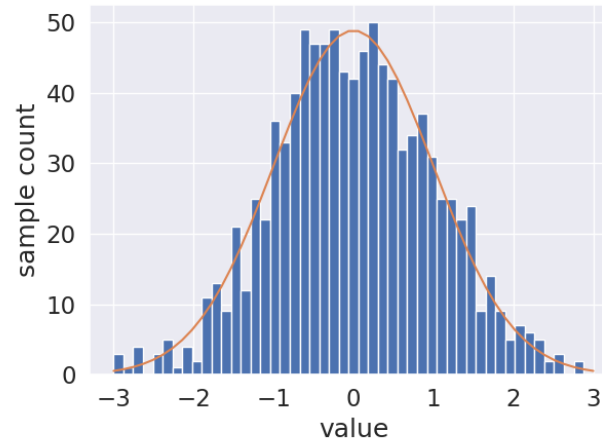
properties?



Statistical Testing

- Is a distribution normal?

properties? skew and excess kurtosis are zero
(kurtosis for normal is 3, taken as a reference)

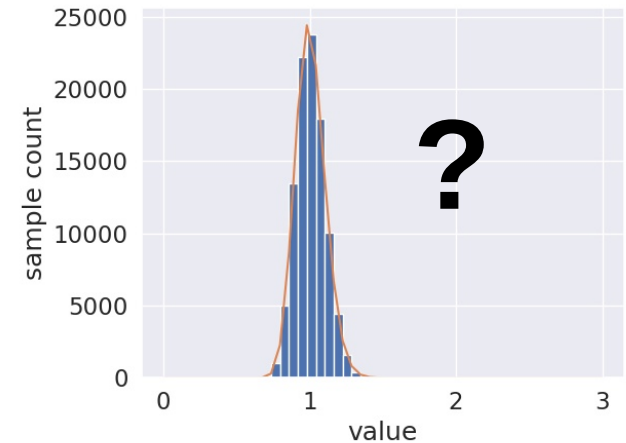


$$E\left[\frac{(X-\mu)^3}{\sigma^3}\right]=0$$
$$E\left[\frac{(X-\mu)^4}{\sigma^4}\right]=3$$

Statistical Testing

- Is a distribution normal?
 - sample estimates of skew and kurtosis

$$\{x_i\} \rightarrow \frac{m_3}{\sigma^3}, \frac{m_4}{\sigma^4}$$



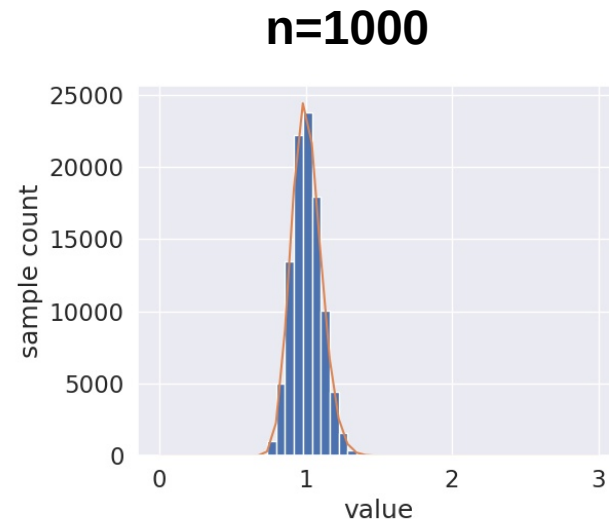
Statistical Testing

- Is a distribution normal?
 - sample estimates of skew and kurtosis
- Test if skew and kurtosis are close to predicted values: but how close?
 - what is the distribution of skew and kurtosis for a normal distribution with given number of samples?

$$\{x_i\} \rightarrow \frac{m_3}{\sigma^3}, \frac{m_4}{\sigma^4}$$

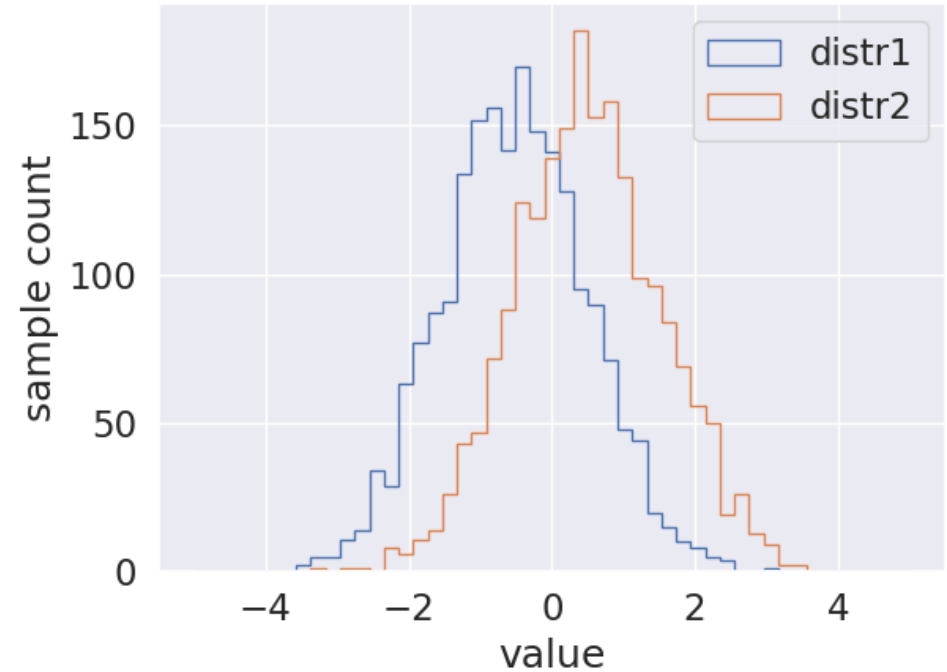
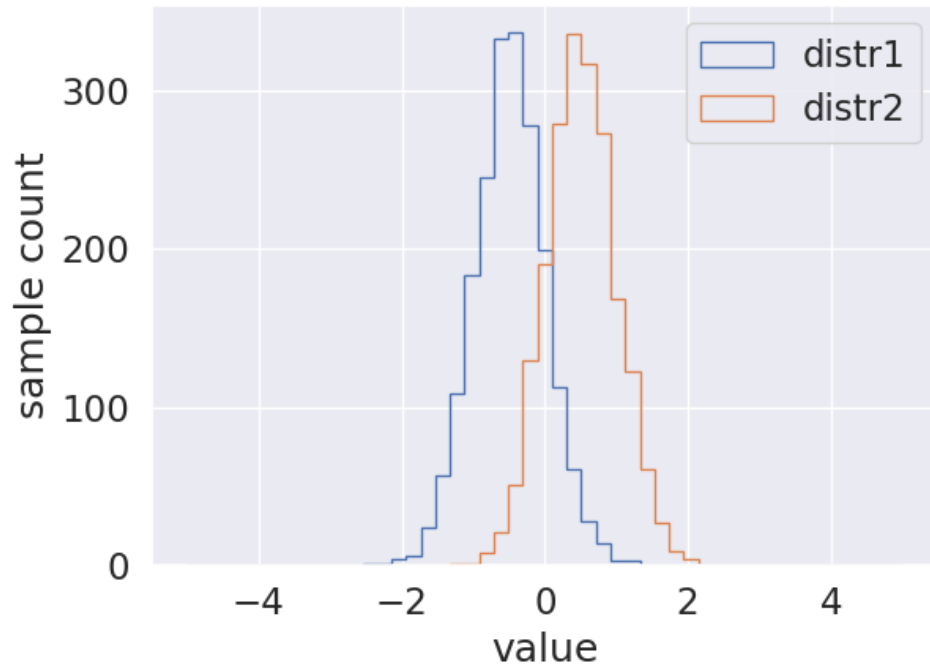
$$P_{X \sim N(0,1)} \left[\frac{m_3}{\sigma^3} \right]$$

$$P_{X \sim N(0,1)} \left[\frac{m_4}{\sigma^4} \right]$$



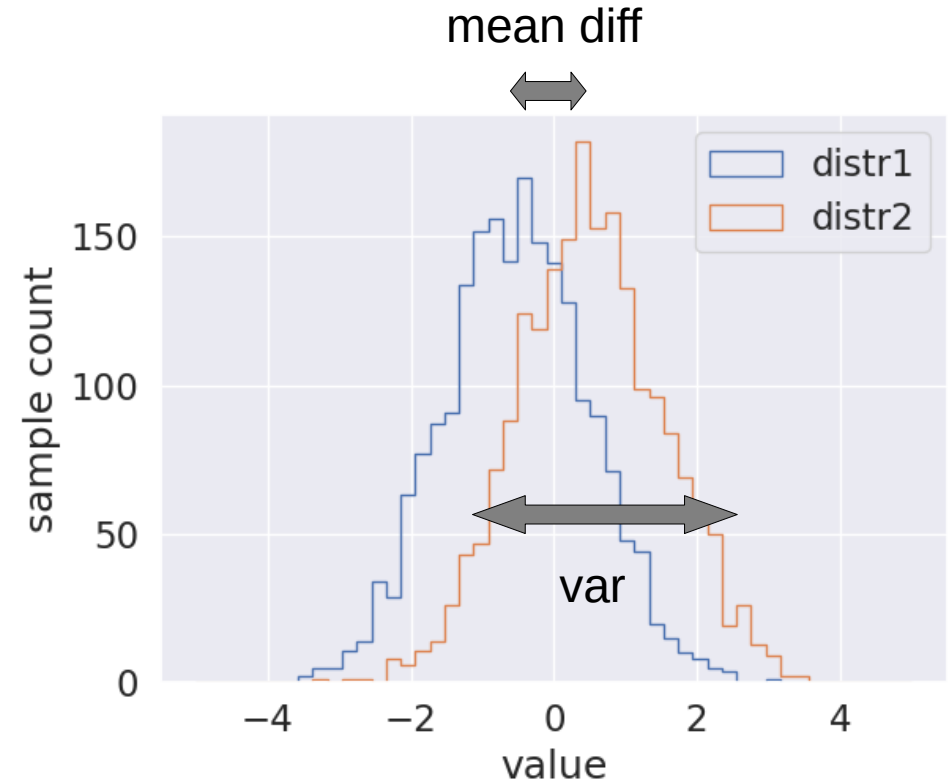
Statistical Testing

- How different are two distributions?



Statistical Testing

- How different are two distributions?
 - compare means



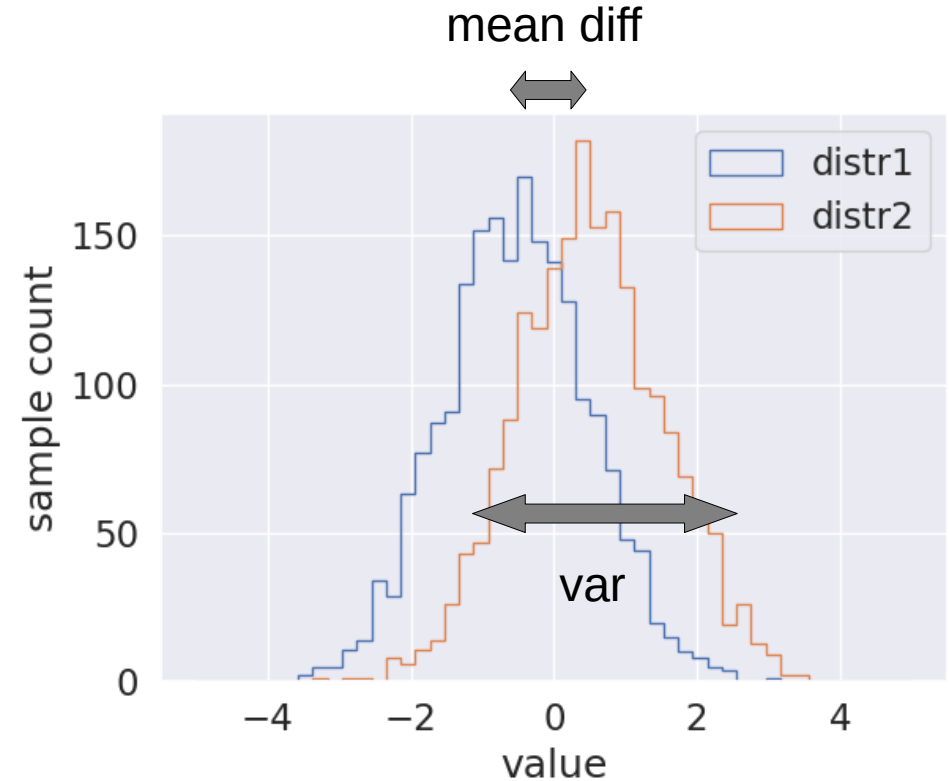
Statistical Testing

- How different are two distributions?
 - compare means
 - variance matters

effect size $\frac{\bar{x}_1 - \bar{x}_2}{\bar{s}}$

t-statistic $\frac{\bar{x}_1 - \bar{x}_2}{\bar{s} \sqrt{2/n}}$

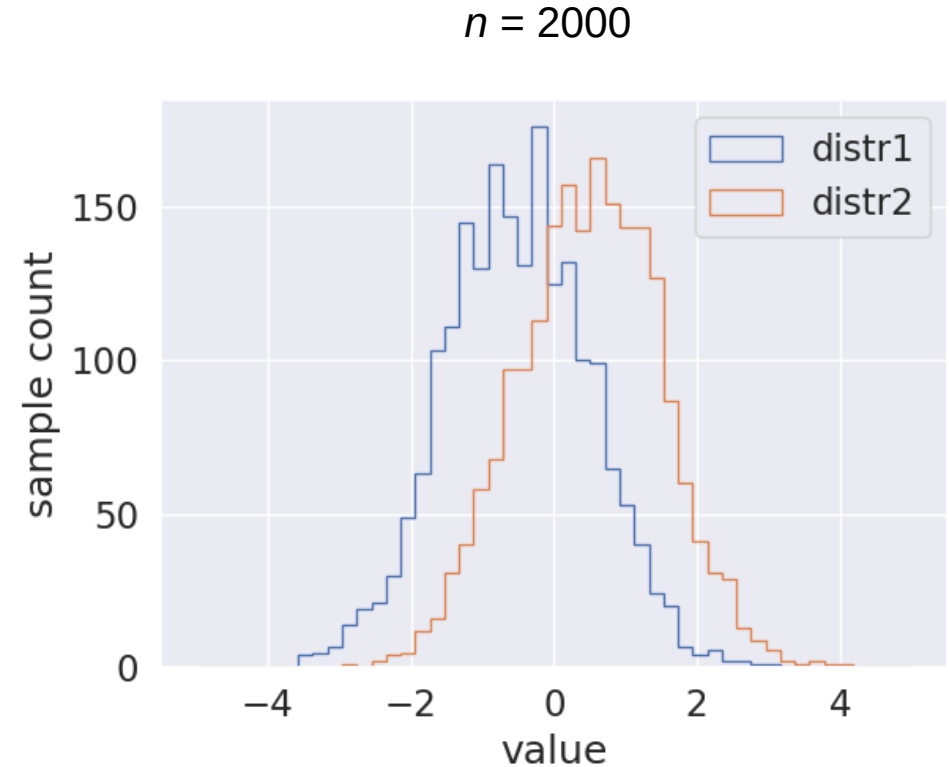
(each distribution with n samples)



Statistical Testing

- How different are two distributions?
 - compare means
 - variance matters
 - number of samples matters!

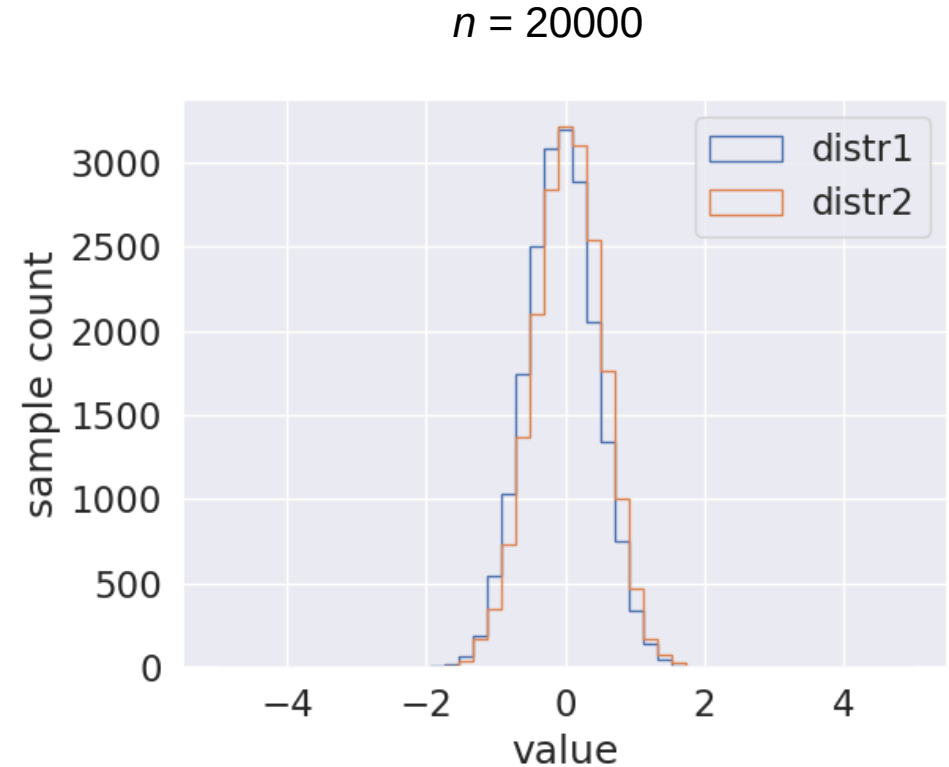
statistic = -31.68
p-value = 1.19e-197



Statistical Testing

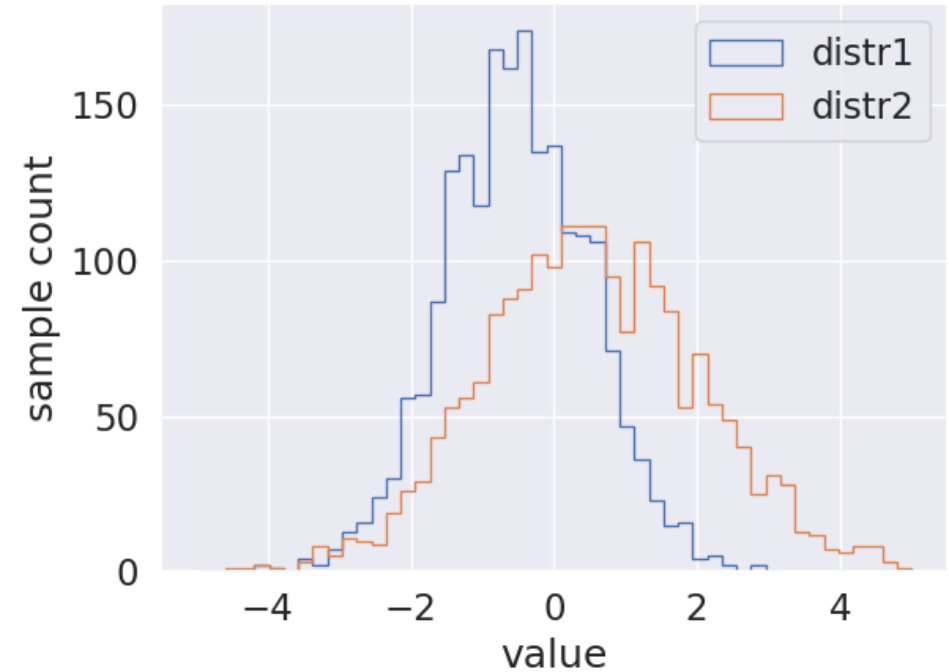
- How different are two distributions?
 - compare means
 - variance matters
 - number of samples matters!

statistic = -19.27
p-value = 2.10e-82



Statistical Testing

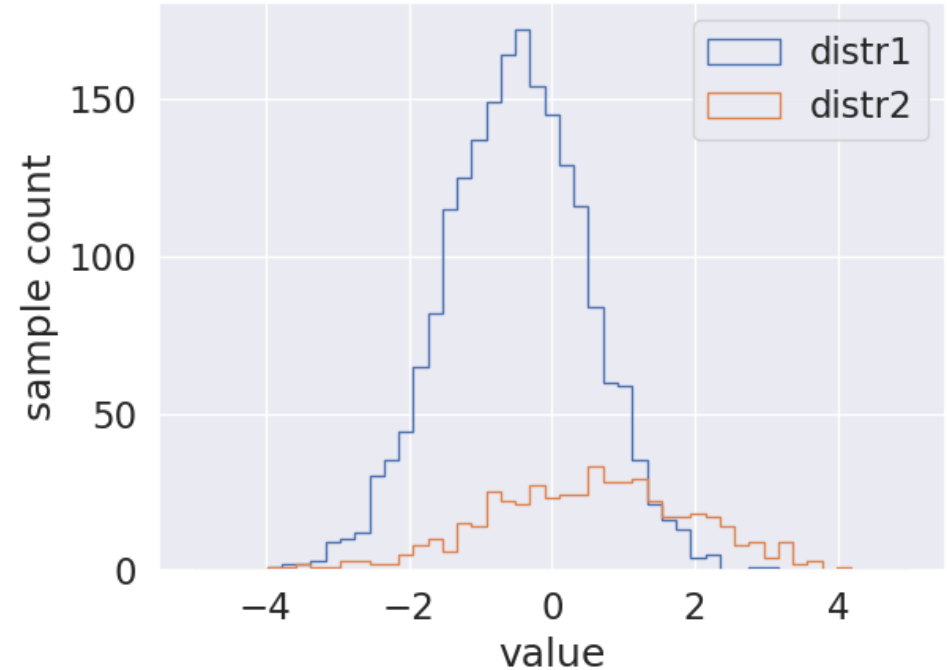
- How different are two distributions?
 - compare means
 - variances matter!
 - number of samples matters!



Statistical Testing

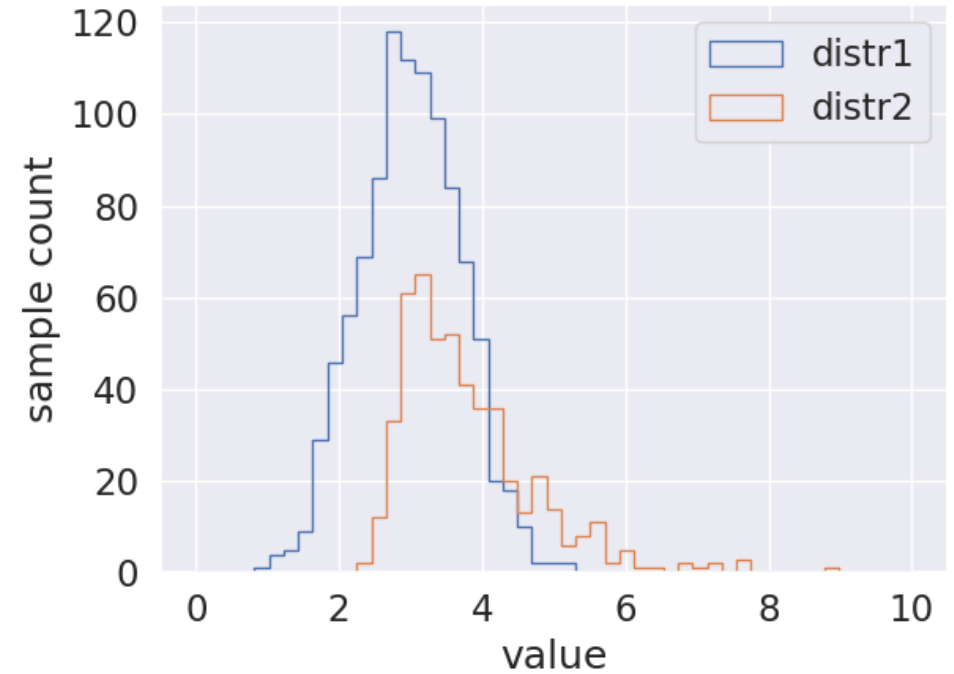
- How different are two distributions?
 - compare means
 - variances matter!
 - number of samples matters!

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\bar{S}_1^2}{n_1} + \frac{\bar{S}_2^2}{n_2}}}$$



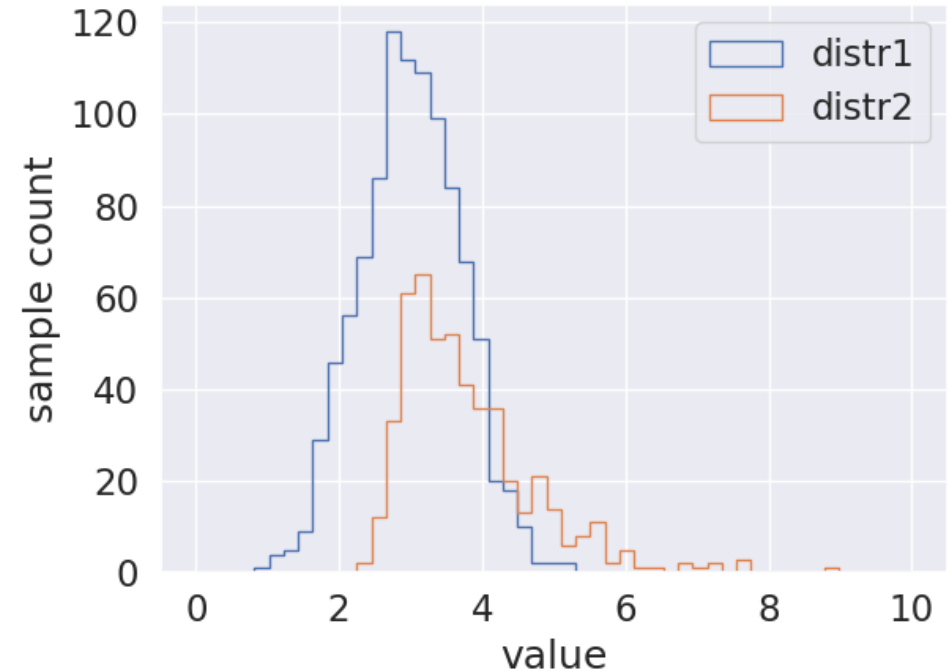
Statistical Testing

- How different are two distributions?



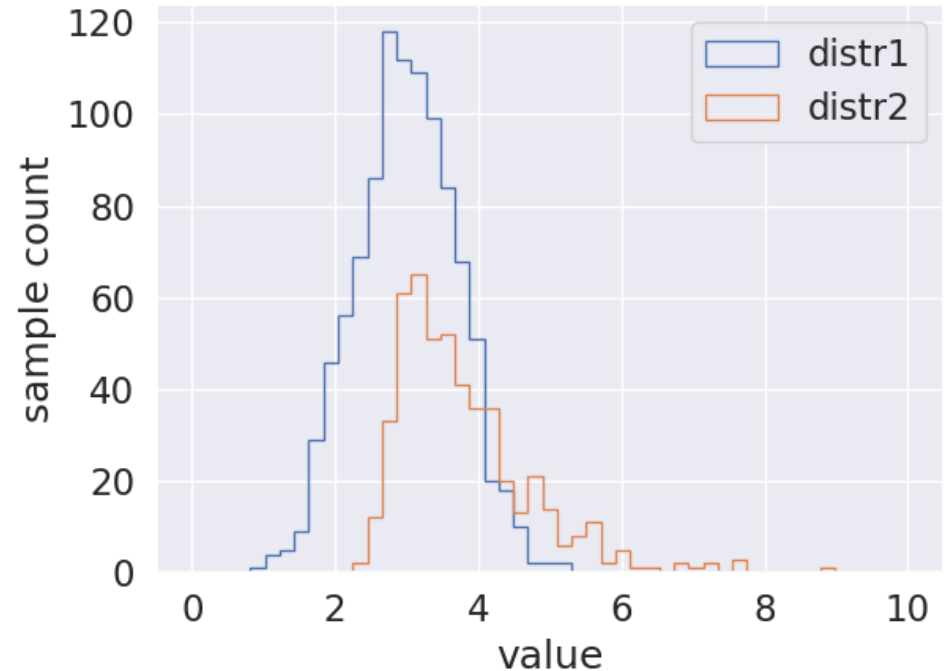
Statistical Testing

- How different are two distributions?
- let's compare normal and lognormal distributions
- they have been aligned (same mean) and normalized to have equal variance
- `Ttest_indResult(statistic=1.88, pvalue=0.06)`



Statistical Testing

- How different are two distributions?
- Transform values into (discrete) rankings
 - group index in ranked list
 - $x1=(0.0, 0.6, 0.5) \rightarrow (0,3,2)$ and $x2=(1.0, 1.2, 0.2) \rightarrow (4,5,1)$
- Check whether mean ranking for each distribution is invariant by permutation
 - statistics of rank sum
- `MannwhitneyuResult(statistic=23461.0, pvalue=0.002)`



Statistical Testing

- Parametric:
 - Student t-test `stt.ttest_ind(x1, x2)`
 - Welch t-test `stt.ttest_ind(x1, x2, equal_var=False)`
- Non-parametric:
 - `stt.ttest_ind(x1, x2, permutations=1000)`
 - Mann-Whitney u `stt.mannwhitneyu(x1, x2)`

Statistical Testing

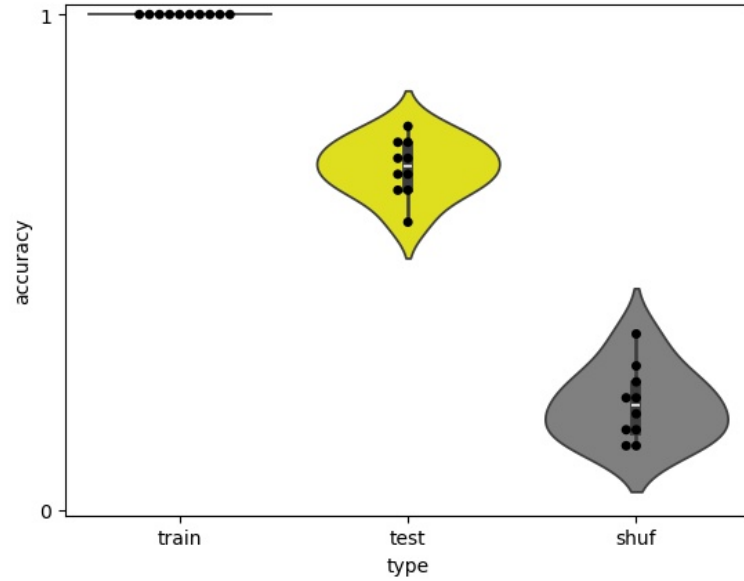
- Parametric:
 - Student t-test `stt.ttest_ind(x1, x2)`
 - Welch t-test `stt.ttest_ind(x1, x2, equal_var=False)`
- Non-parametric:
 - `stt.ttest_ind(x1, x2, permutations=1000)`
 - Mann-Whitney u `stt.mannwhitneyu(x1, x2)`
- **CHECK THE OPTIONS AND CONDITIONS OF APPLICATIONS!!!**
- **BY DEFAULT, PREFER NON-PARAMETRIC STATISTICS**

Statistical Testing

ASA statement on statistical significance... The American Statistician (2016) 70: 131-133

- P-values can indicate how incompatible the data are with a specified statistical model.
- A p-value does not measure the size of an effect or the importance of a result.
- Proper inference requires full reporting and transparency.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Example of Analysis of Classification Results

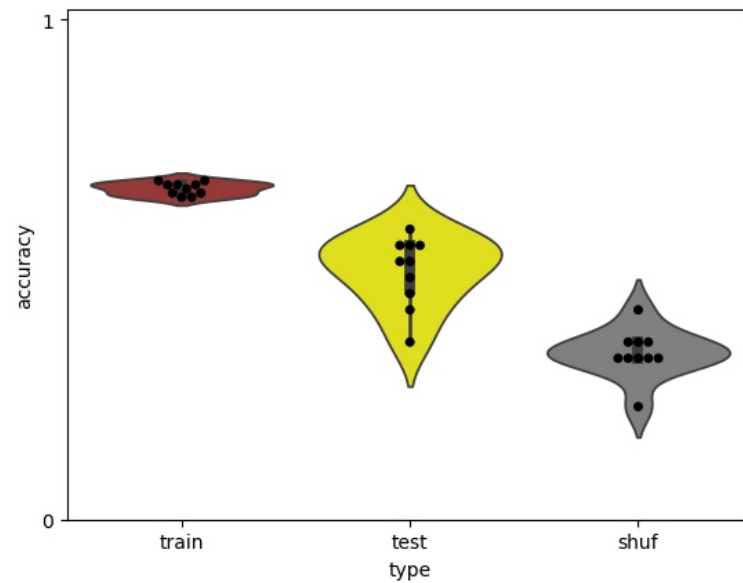
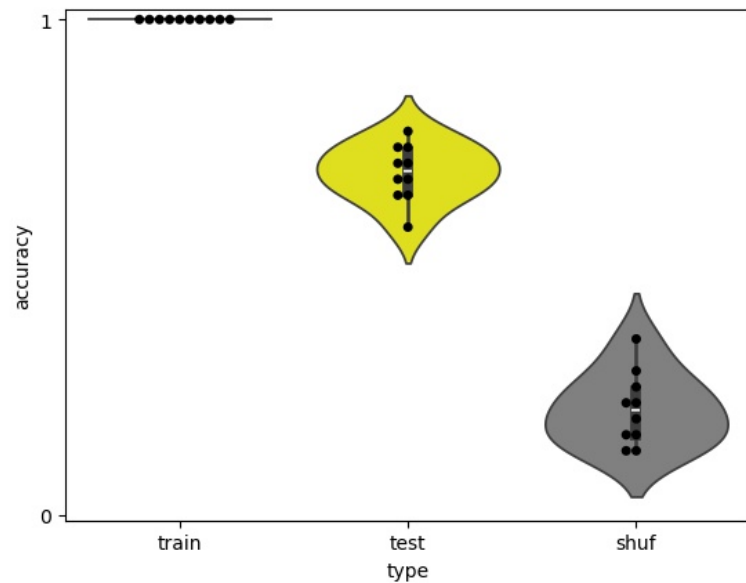


Did the classifier
train well?

Does it generalize
to "new" data?

How does a random
classifier perform?

Example of Analysis of Classification Results

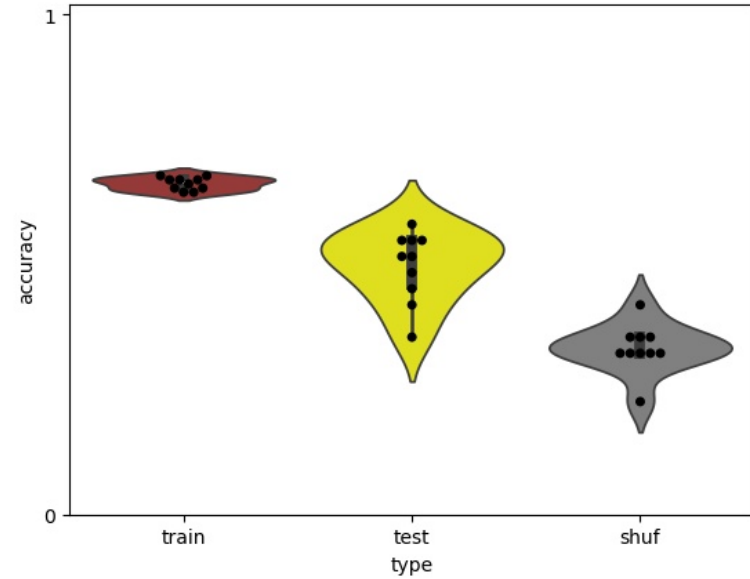
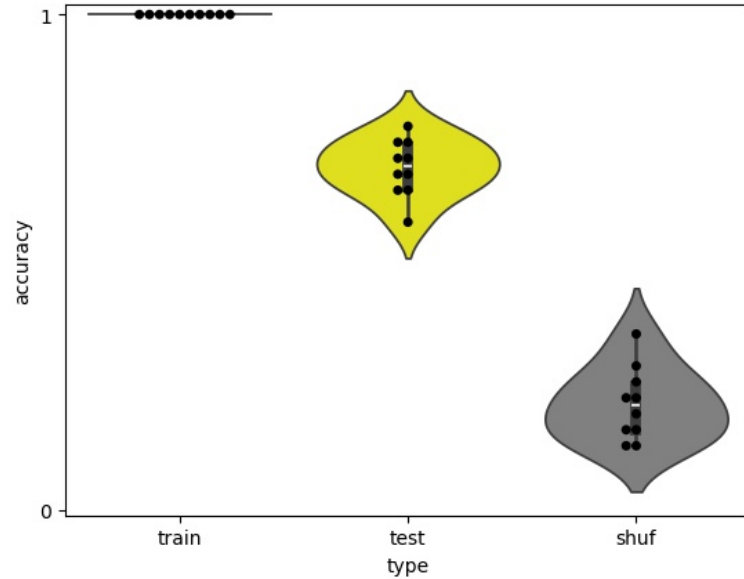


Did the classifier
train well?

Does it generalize
to "new" data?

How does a random
classifier perform?

Example of Analysis of Classification Results



- Statistical difference:

$$t = \frac{\Delta \bar{x}}{\bar{s}'}$$

- Effect size (Jacob Cohen d):

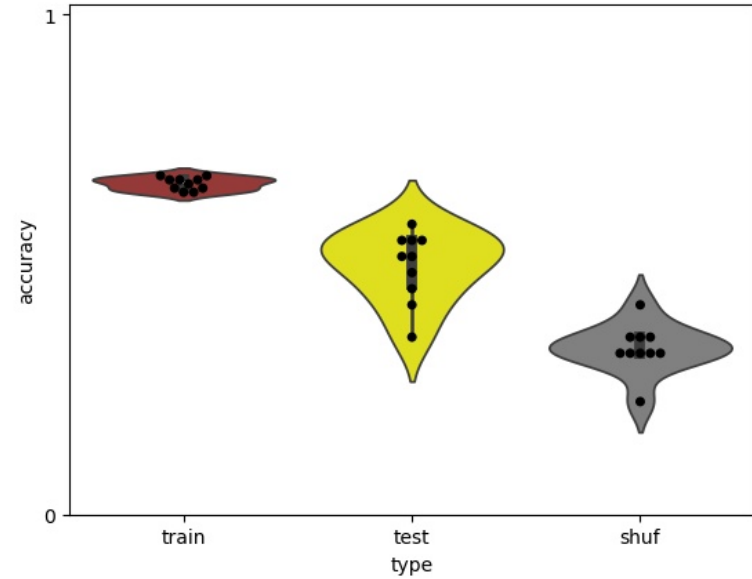
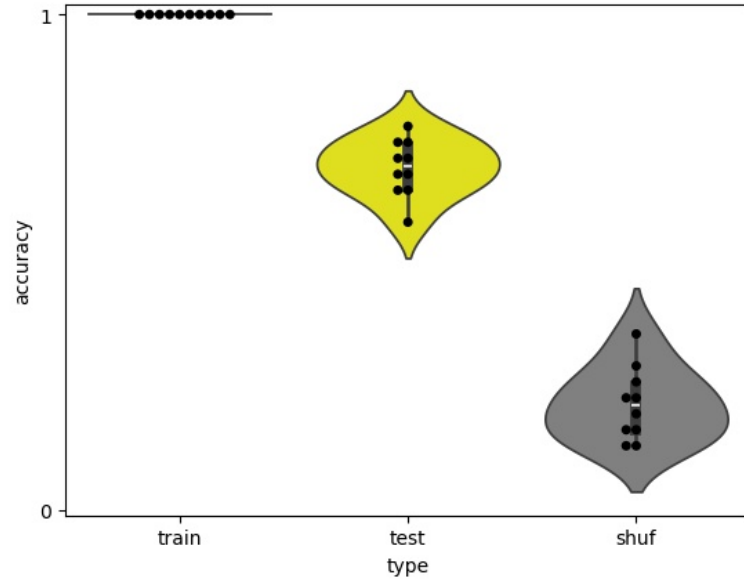
$$d' = \frac{\Delta \bar{x}}{\bar{s}''}$$

“weighted” variances (n splits)

$$\bar{s}' = \sqrt{\frac{\bar{s}_1^2 + \bar{s}_2^2}{n}}$$

$$\bar{s}'' = \sqrt{\frac{\bar{s}_1^2 + \bar{s}_2^2}{2}}$$

Example of Analysis of Classification Results



- Statistical difference:

$$t = \frac{\Delta \bar{x}}{\bar{s}'} \simeq d' \sqrt{n/2}$$

- Effect size (Jacob Cohen d): $d' = \frac{\Delta \bar{x}}{\bar{s}''}$

“weighted” variances (n splits)

$$\bar{s}' = \sqrt{\frac{\bar{s}_1^2 + \bar{s}_2^2}{n}}$$

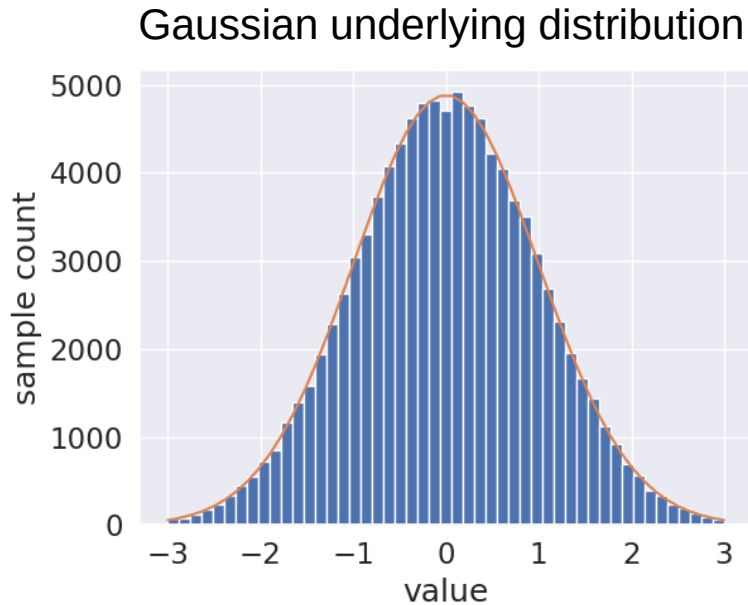
$$\bar{s}'' = \sqrt{\frac{\bar{s}_1^2 + \bar{s}_2^2}{2}}$$

Multiple Comparisons

- Let's make $m=1$ independent tests with p-value threshold $\alpha=0.05$
- What is the probability to have at least one false positive?

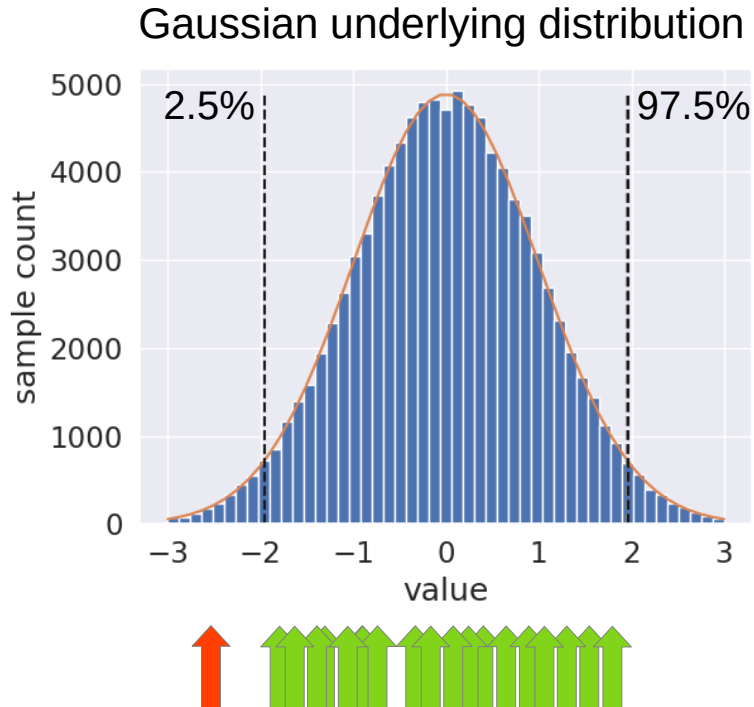
Multiple Comparisons

- Let's make $m=1$ independent tests with p-value threshold $\alpha=0.05$
- What is the probability to have at least one false positive?



Multiple Comparisons

- Let's make $m=1$ independent tests with p-value threshold $\alpha=0.05$
- What is the probability to have at least one false positive?



one error every 20 repetitions
rate = $1 - \alpha$

Multiple Comparisons

- Let's make **$m=20$** independent tests with p-value threshold $\alpha=0.05$
- What is the probability to have at least one false positive?

Multiple Comparisons

- Let's make $m=20$ independent tests with p-value threshold $\alpha=0.05$
- What is the probability to have at least one false positive?

$$1 - (1 - \alpha)^m \simeq 64 \%$$

Multiple Comparisons

- Let's make $m=20$ independent tests with p-value threshold $\alpha=0.05$
- What is the probability to have at least one false positive?

$$1 - (1 - \alpha)^m \simeq 64 \%$$

- Bonferroni correction: much stricter threshold $\alpha' = \frac{\alpha}{m}$
 - desired threshold applies to the series of “experiments”

$$1 - (1 - \alpha')^m \simeq \alpha$$

Multiple Comparisons

- Let's make $m=20$ independent tests with p-value threshold $\alpha=0.05$
- What is the probability to have at least one false positive?

$$1 - (1 - \alpha)^m \simeq 64 \%$$

- Bonferroni correction: much stricter threshold $\alpha' = \frac{\alpha}{m}$
 - desired threshold applies to the series of “experiments”

$$1 - (1 - \alpha')^m \simeq \alpha$$

- False discovery rate (FDR) correction: XXX

Experimental Design

Practice

notebook distributions.ipynb

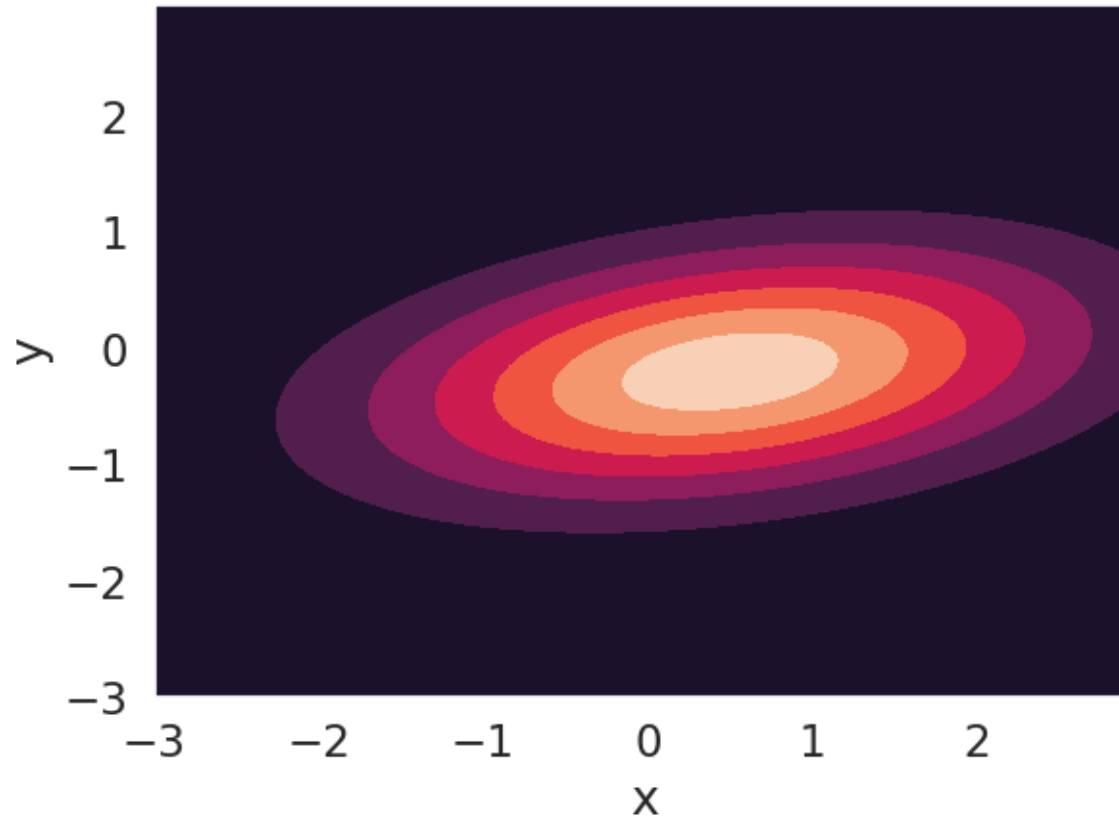
Practice

How to build non-parametric test equivalent to parametric test?

notebook **corr_param_nonparam.ipynb**

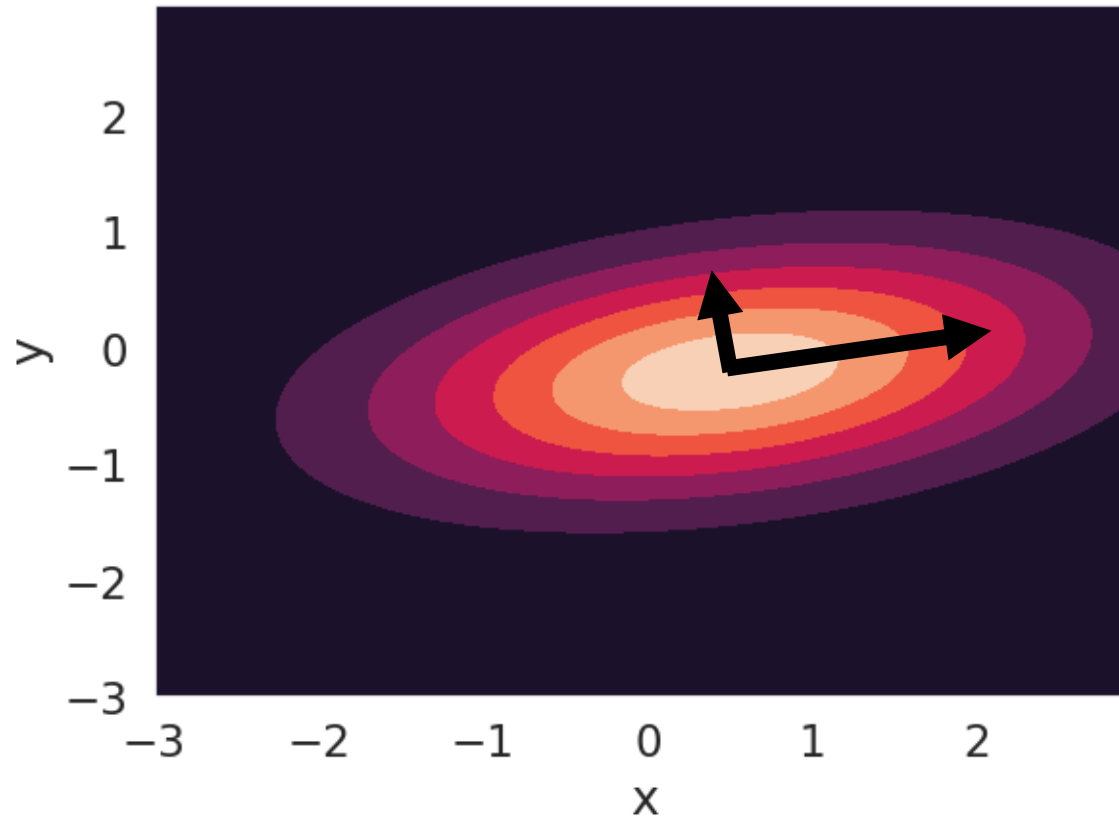
Multivariate Normal Distributions

$$p(x, y) = N(\mu, \Sigma)$$



Multivariate Normal Distributions

$$p(x, y) = N(\mu, \Sigma)$$

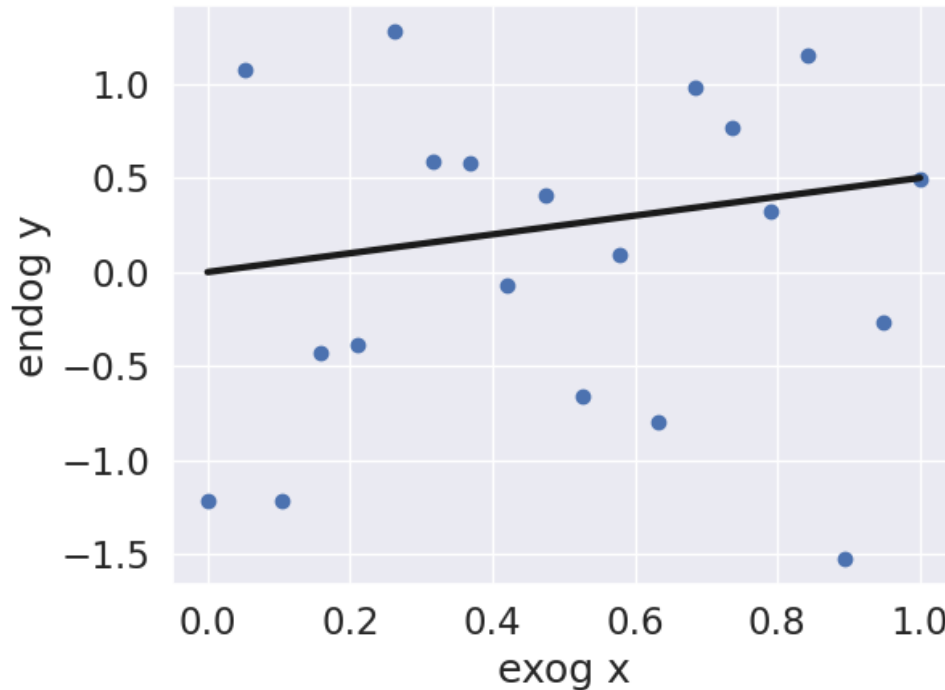


Statistical Analysis in Python

- Probabilities, distributions
- Parametric and non-parametric testing
- **Regressions**
 - **linear regression**
 - **fixed-effect model**
 - **mixed-effect model**
- Bayesian inference

Linear Regression

- Simple case of 1 response variable, we check its dependency on the predictor



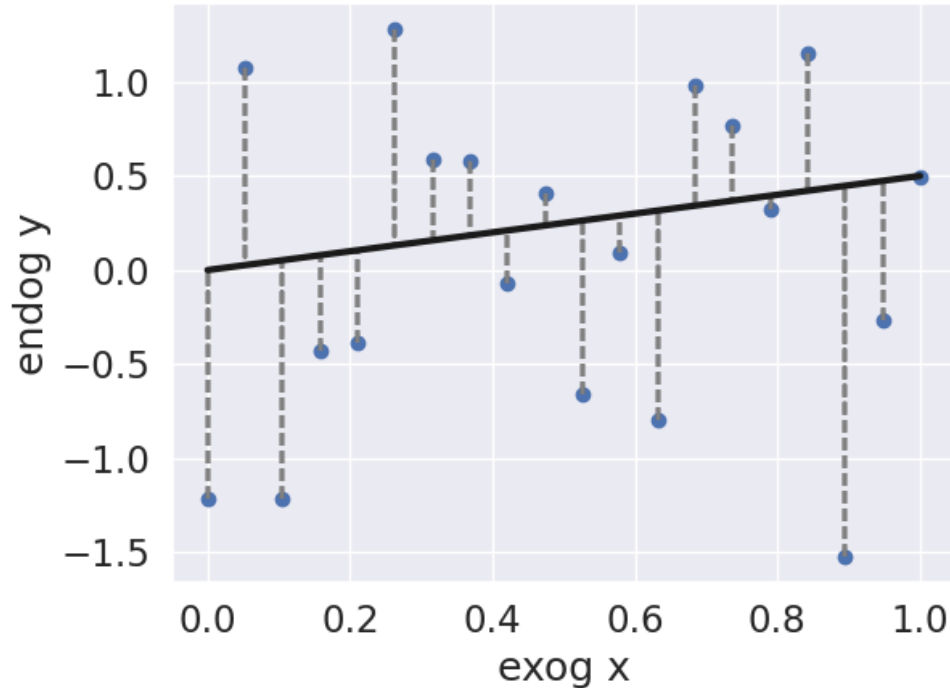
in black: real dependency

$$y = ax + \epsilon$$

normally
distributed
noise

Linear Regression

- Simple case of 1 response variable, we check its dependency on the predictor

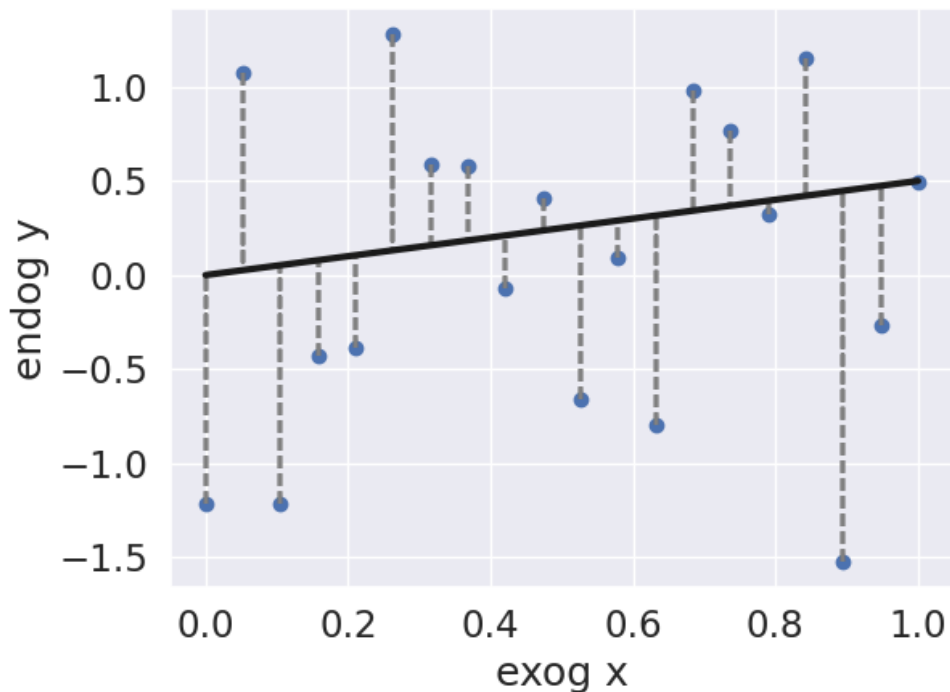


error due to noise

$$\epsilon_i = y_i - a x_i$$

Linear Regression

- Simple case of 1 response variable, we check its dependency on the predictor
- Minimize sum of squared errors, assumed to be normally distributed

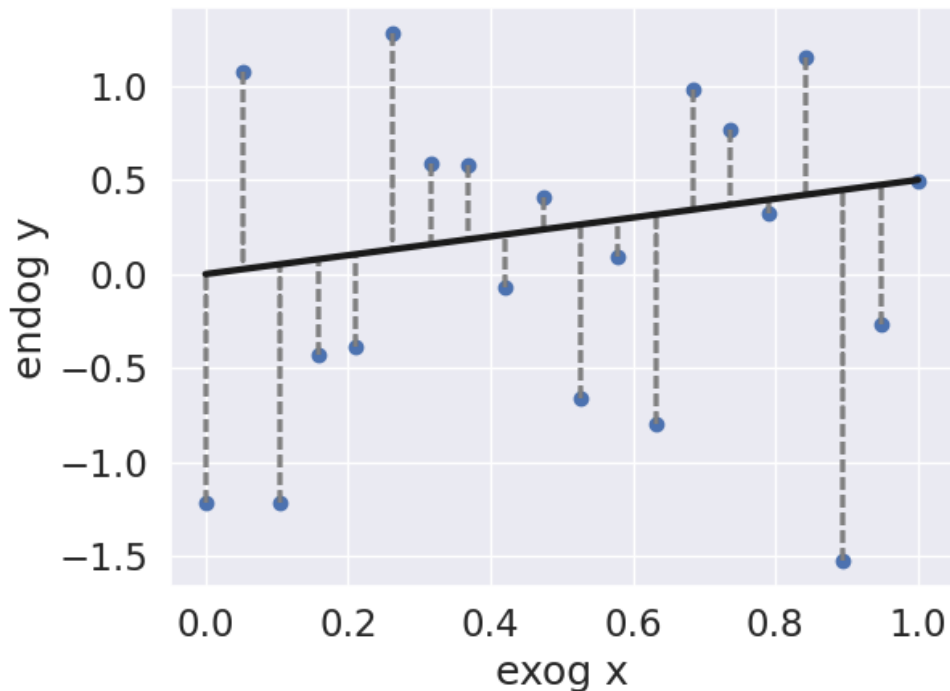


error due to noise

$$\epsilon_i = y_i - a x_i$$

Linear Regression

- Simple case of 1 response variable, we check its dependency on the predictor
- Minimize sum of squared errors, assumed to be normally distributed

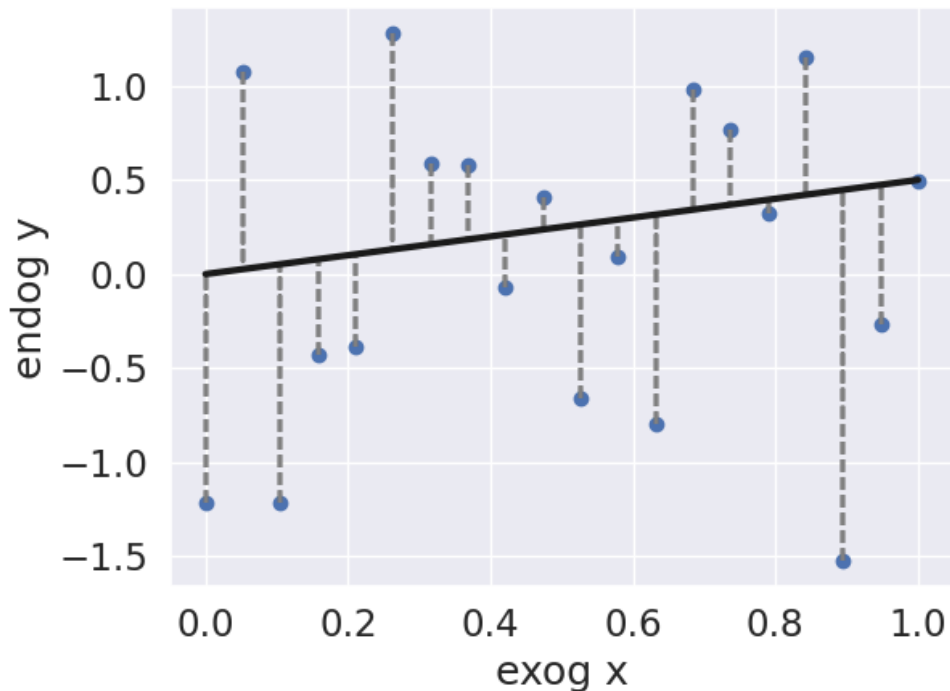


$$\epsilon_i = y_i - a x_i$$

$$\sum_i x_i \epsilon_i = \sum_i x_i y_i - a \sum_i x_i x_i$$

Linear Regression

- Simple case of 1 response variable, we check its dependency on the predictor
- Minimize sum of squared errors, assumed to be normally distributed

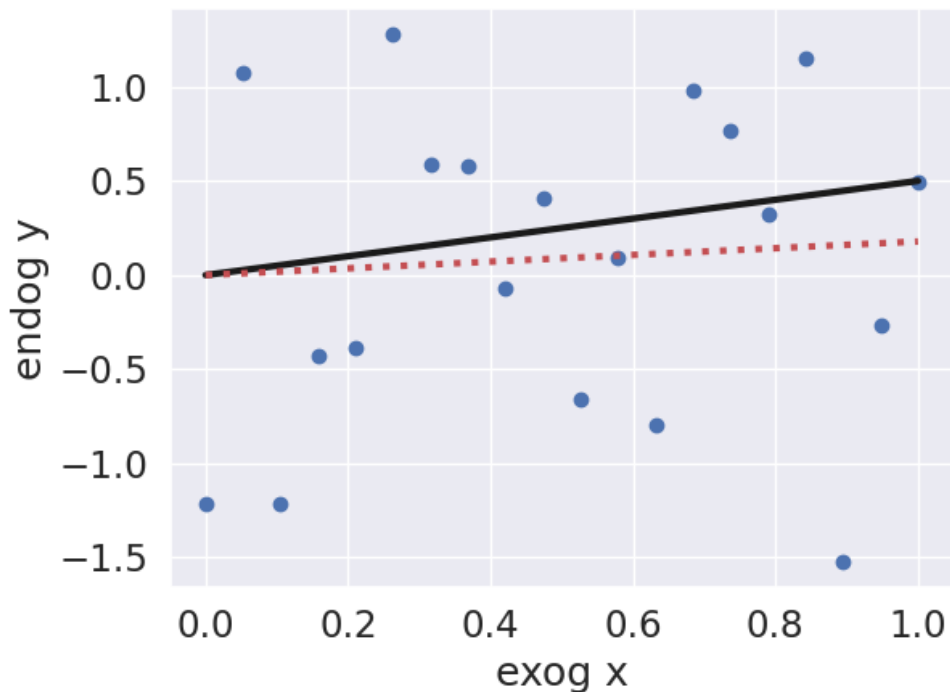


$$\epsilon_i = y_i - a x_i$$

$$0 = \sum_i x_i y_i - a \sum_i x_i x_i$$

Linear Regression

- Simple case of 1 response variable, we check its dependency on the predictor
- Minimize sum of squared errors, assumed to be normally distributed

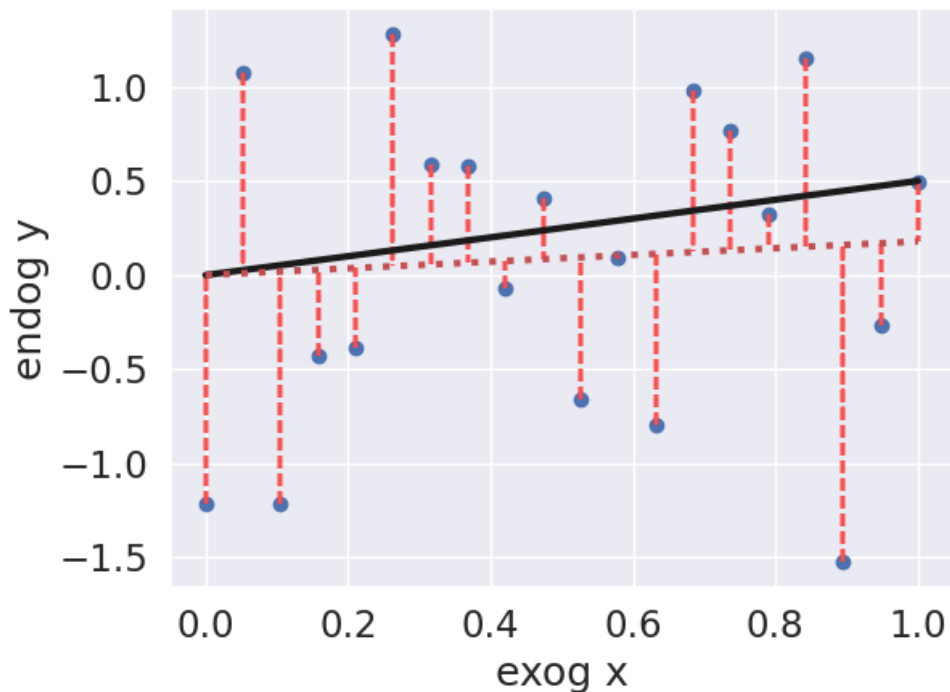


$$\epsilon_i = y_i - a x_i$$

$$\hat{a} = \frac{\sum_i x_i y_i}{\sum_i x_i x_i} \quad \text{estimate}$$

Linear Regression

- Simple case of 1 response variable, we check its dependency on the predictor
- Minimize sum of squared errors, assumed to be normally distributed



$$\hat{\epsilon}_i = y_i - \hat{a} x_i \quad \text{residuals}$$

$$\hat{a} = \frac{\sum_i x_i y_i}{\sum_i x_i x_i} \quad \text{estimate}$$

Linear Regression

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared (uncentered):      0.057
Model:                  OLS    Adj. R-squared (uncentered):    0.007
Method:                 Least Squares    F-statistic:          1.144
Date:                   Thu, 20 Jul 2023    Prob (F-statistic):    0.298
Time: 14:47:50    Log-Likelihood:      -25.629
No. Observations:      20    AIC:          53.26
Df Residuals:          19    BIC:          54.25
Df Model:              1
Covariance Type: nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
x1	0.3656	0.342	1.069	0.298	-0.350	1.081

```
=====
Omnibus:                1.359    Durbin-Watson:          2.199
Prob(Omnibus):          0.507    Jarque-Bera (JB):       0.364
Skew:                   0.283    Prob(JB):               0.834
Kurtosis:               3.340    Cond. No.:              1.00
=====
```

Quality and Confidence in Estimation

Quality and Confidence in Estimation

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared (uncentered):      0.069
Model:                  OLS    Adj. R-squared (uncentered):    0.020
Method:                 Least Squares    F-statistic:          1.409
Date:                  Thu, 20 Jul 2023    Prob (F-statistic):    0.250
Time:                  15:40:40    Log-Likelihood:       -30.071
No. Observations:      20      AIC:                  62.14
Df Residuals:          19      BIC:                  63.14
Df Model:               1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
x1	0.5067	0.427	1.187	0.250	-0.387	1.400

```
=====
Omnibus:                2.071    Durbin-Watson:          2.495
Prob(Omnibus):           0.355    Jarque-Bera (JB):       0.700
Skew:                   0.358     Prob(JB):               0.705
Kurtosis:                3.573    Cond. No.                1.00
=====
```

Quality and Confidence in Estimation

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared (uncentered):      0.069
Model:                  OLS    Adj. R-squared (uncentered):    0.020
Method:                 Least Squares    F-statistic:          1.409
Date:                  Thu, 20 Jul 2023    Prob (F-statistic):    0.250
Time:                  15:40:40    Log-Likelihood:       -30.071
No. Observations:      20    AIC:                  62.14
Df Residuals:          19    BIC:                  63.14
Df Model:              1
Covariance Type:       nonrobust
=====
```

```
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
x1             0.5067      0.427        1.187      0.250      -0.387      1.400
=====
```

```
=====
Omnibus:                2.071    Durbin-Watson:                2.495
Prob(Omnibus):          0.355    Jarque-Bera (JB):            0.700
Skew:                   0.358    Prob(JB):                    0.705
Kurtosis:               3.573    Cond. No.                     1.00
=====
```

Quality and Confidence in Estimation

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared (uncentered):      0.069
Model:                  OLS    Adj. R-squared (uncentered):    0.020
Method:                 Least Squares    F-statistic:          1.409
Date:                   Thu, 20 Jul 2023    Prob (F-statistic):    0.250
Time:                   15:40:40    Log-Likelihood:       -30.071
No. Observations:      20    AIC:                  62.14
Df Residuals:          19    BIC:                  63.14
Df Model:               1
Covariance Type:       nonrobust
=====
```

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i \hat{\epsilon}_i^2}{\sum_i (y_i - \bar{y})^2}$$

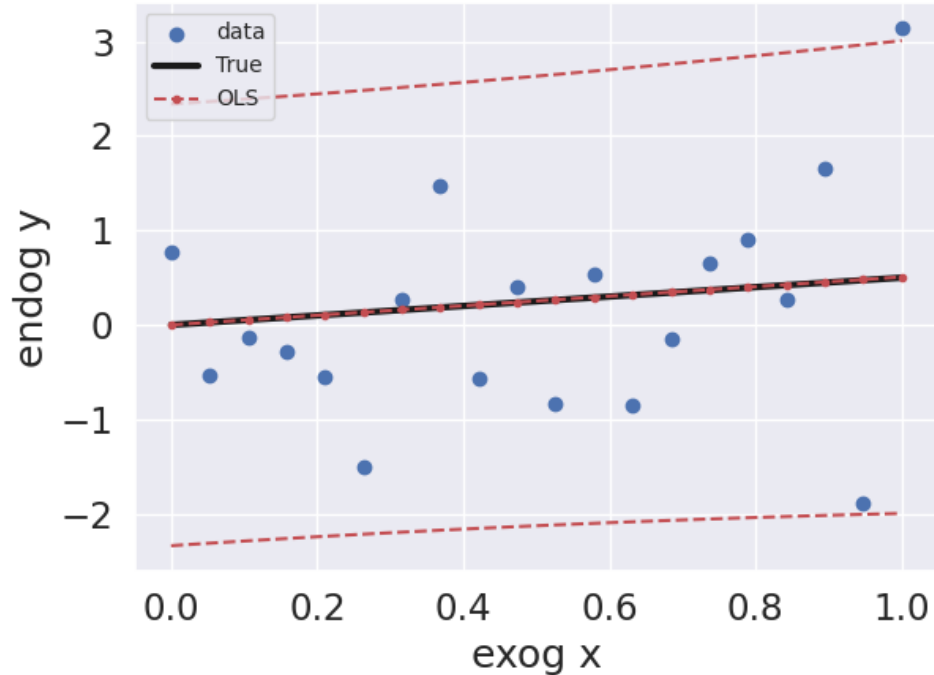
explained variance
total variance

prediction $\hat{y}_i = \hat{a} x_i$

residuals $\hat{\epsilon}_i = y_i - \hat{a} x_i = y_i - \hat{y}_i$

Quality and Confidence in Estimation

$y = a x + \epsilon$ generative model with high noise

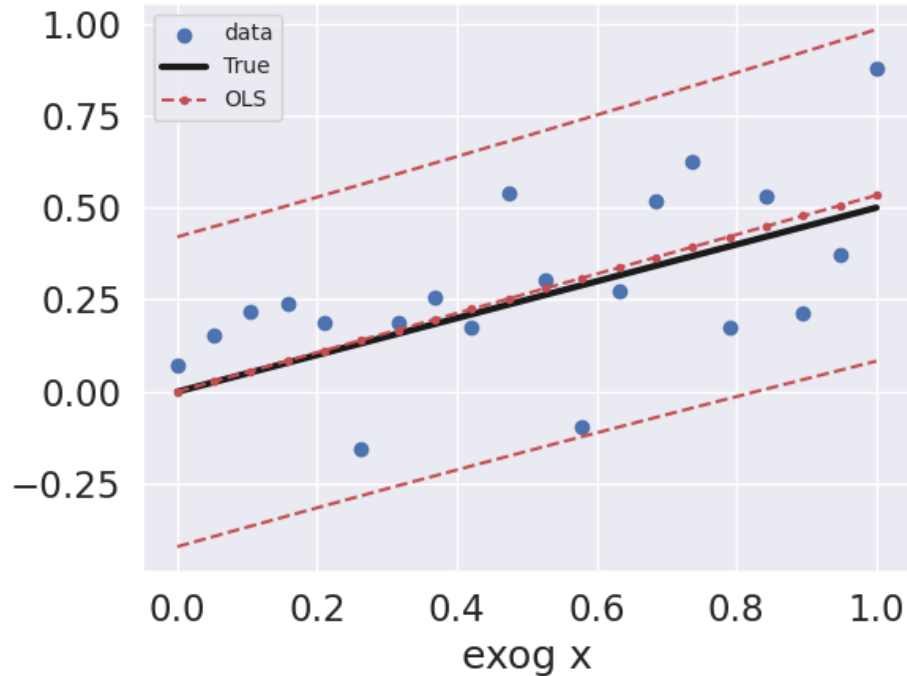


R-squared (uncentered): 0.069
F-statistic: 1.409
Prob (F-statistic): 0.250
Log-Likelihood: -30.071

```
=====
                        coef      std err      [0.025      0.975]
-----
x1      0.5067      0.42      -0.387      1.400
=====
```

Quality and Confidence in Estimation

$y = a x + \epsilon$ generative model with low noise



R-squared (uncentered): 0.717
F-statistic: 48.25
Prob (F-statistic): 1.28e-06
Log-Likelihood: 4.2313

=====				
	coef	std err	[0.025	0.975]

x1	0.5335	0.077	0.373	0.694
=====				

Quality and Confidence in Estimation

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared (uncentered):          0.057
Model:                  OLS    Adj. R-squared (uncentered):       0.007
Method:                 Least Squares    F-statistic:          1.144
Date:                  Thu, 20 Jul 2023    Prob (F-statistic):    0.298
Time:                  14:47:50    Log-Likelihood:       -25.629
No. Observations:      20    AIC:                  53.26
Df Residuals:          19    BIC:                  54.25
Df Model:               1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
x1	0.3656	0.342	1.069	0.298	-0.350	1.081

```
=====
Omnibus:                1.359    Durbin-Watson:          2.199
Prob(Omnibus):          0.507    Jarque-Bera (JB):       0.364
Skew:                   0.283    Prob(JB):               0.834
Kurtosis:               3.340    Cond. No.                1.00
=====
```

tests on residuals (normality, homoscedasticity, etc.)

Quality and Confidence in Estimation

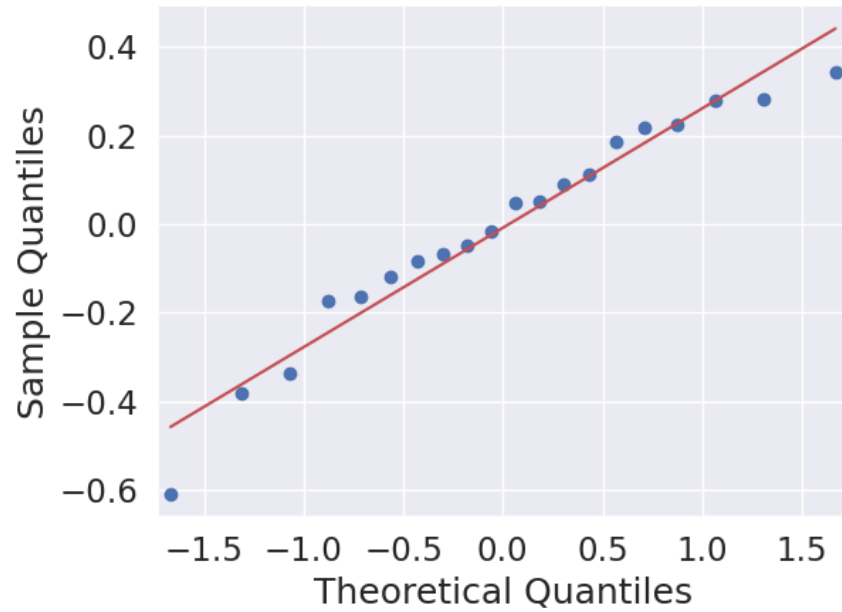
- Residuals distributed normally: omnibus test, Jarque-Bera test
- Independent residuals: Durbin-Watson test
 - target value 2 = absence of autocorrelation (< 2 means positive autocorrelation)
 - if statistic between 1 and 3: OK
- Homogeneous residuals (homoscedasticity)

Omnibus:	1.359	Durbin-Watson:	2.199
Prob (Omnibus) :	0.507	Jarque-Bera (JB) :	0.364
Skew:	0.283	Prob (JB) :	0.834
Kurtosis:	3.340	Cond. No.	1.00

tests on residuals (normality, homoscedasticity, etc.)

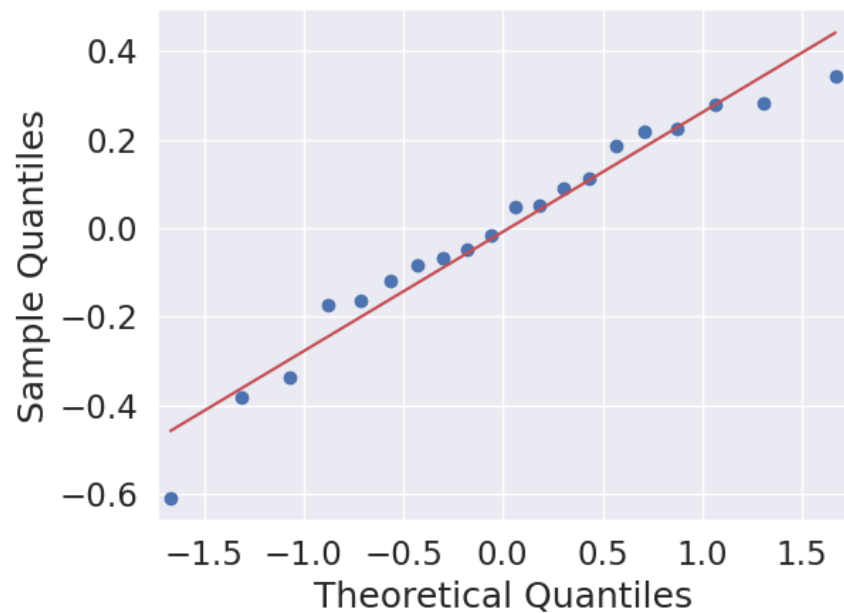
Quality and Confidence in Estimation

normality of residuals: QQ plot

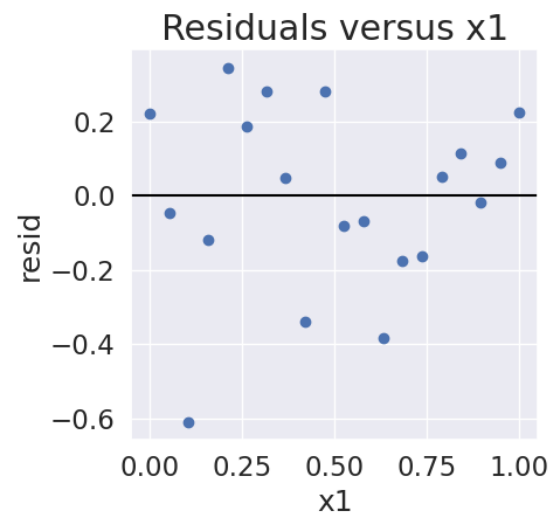
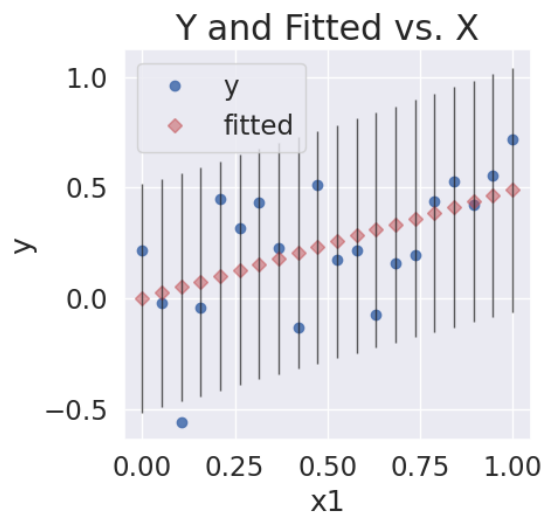


Quality and Confidence in Estimation

normality of residuals: QQ plot



Regression Plots for x1



Model Comparison

Model Comparison

OLS Regression Results

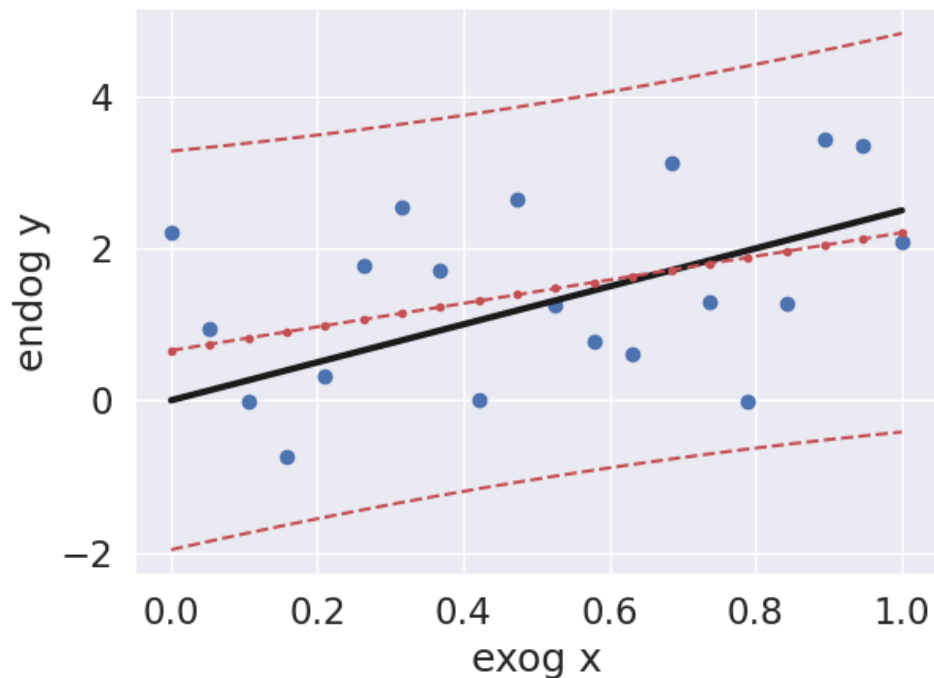
```
=====
Dep. Variable:          y      R-squared:          0.158
Model:                  OLS    Adj. R-squared:       0.111
Method:                 Least Squares  F-statistic:       3.366
Date:                  Tue, 18 Jul 2023  Prob (F-statistic):  0.0831
Time:                  16:37:00  Log-Likelihood:     -30.061
No. Observations:      20      AIC:               64.12
Df Residuals:          18      BIC:               66.11
Df Model:               1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	0.6566	0.494	1.329	0.200	-0.381	1.695
x1	1.5500	0.845	1.835	0.083	-0.225	3.325

```
=====
Omnibus:                4.121    Durbin-Watson:       1.652
Prob(Omnibus):           0.127    Jarque-Bera (JB):     1.412
Skew:                   0.002    Prob(JB):             0.494
Kurtosis:               1.698    Cond. No.             4.18
=====
```

Model Comparison

polynomial regression $y = \hat{a}_0 + \hat{a}_1 x$

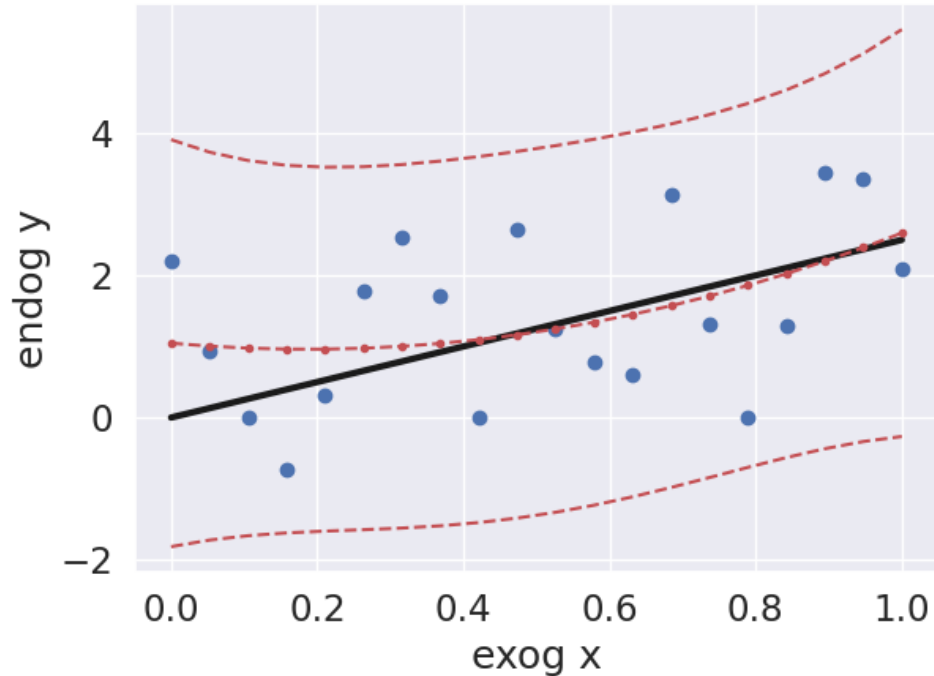


R-squared (uncentered): 0.158
F-statistic: 3.366
Prob (F-statistic): 0.0831
Log-Likelihood: -30.061
BIC: 66.11

	coef	std err	t	P> t
x1	1.5500	0.845	1.835	0.083

Model Comparison

polynomial regression $y = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2$

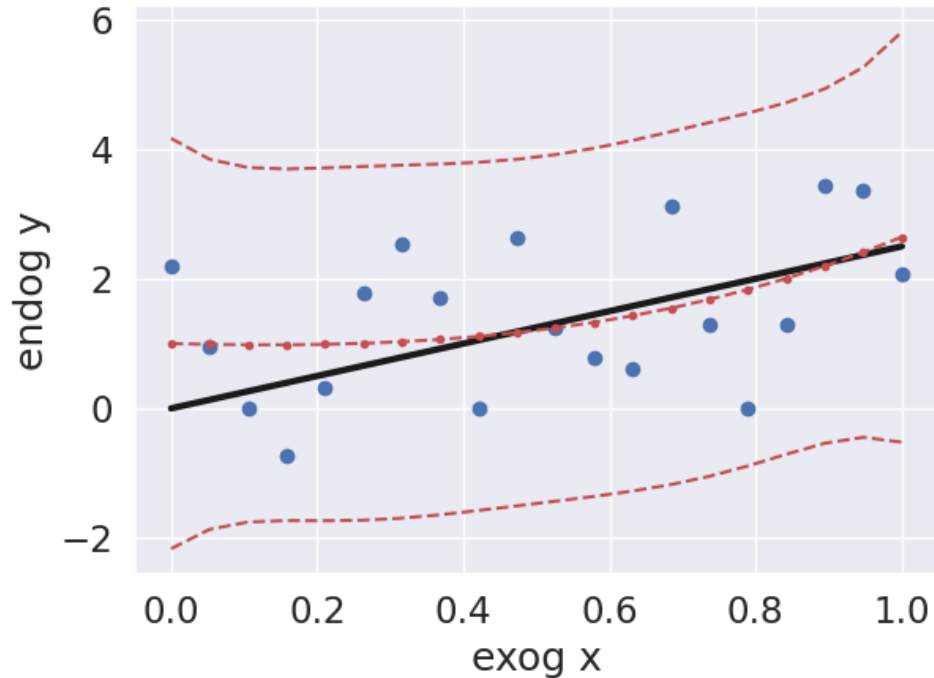


R-squared (uncentered): 0.187
F-statistic: 1.953
Prob (F-statistic): 0.172
Log-Likelihood: -29.707
BIC: 68.40

	coef	std err	t	P> t
x1	-0.9210	3.272	-0.282	0.782
x2	2.4710	3.158	0.782	0.445

Model Comparison

polynomial regression $y = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2 + \dots$

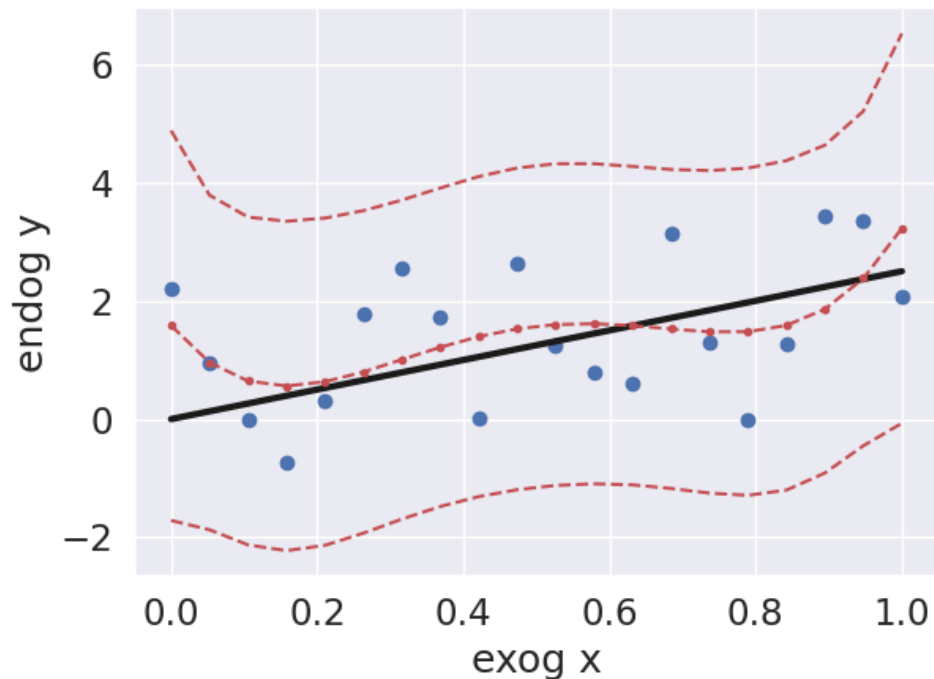


R-squared (uncentered): 0.187
F-statistic: 1.229
Prob (F-statistic): 0.332
Log-Likelihood: -29.702
BIC: 71.39

	coef	std err	t	P> t
x1	-0.2800	7.962	-0.035	0.972
x2	0.8267	18.785	0.044	0.965
x3	1.0963	12.334	0.089	0.930

Model Comparison

polynomial regression $y = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2 + \dots$

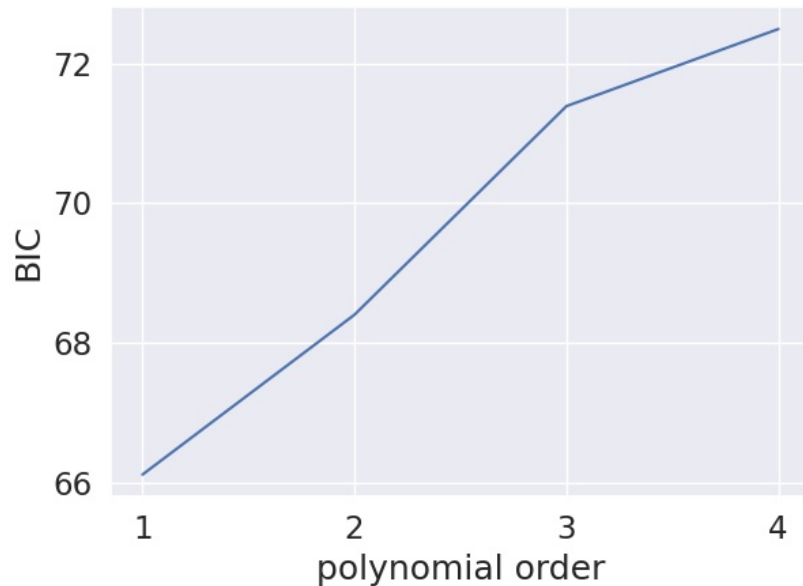
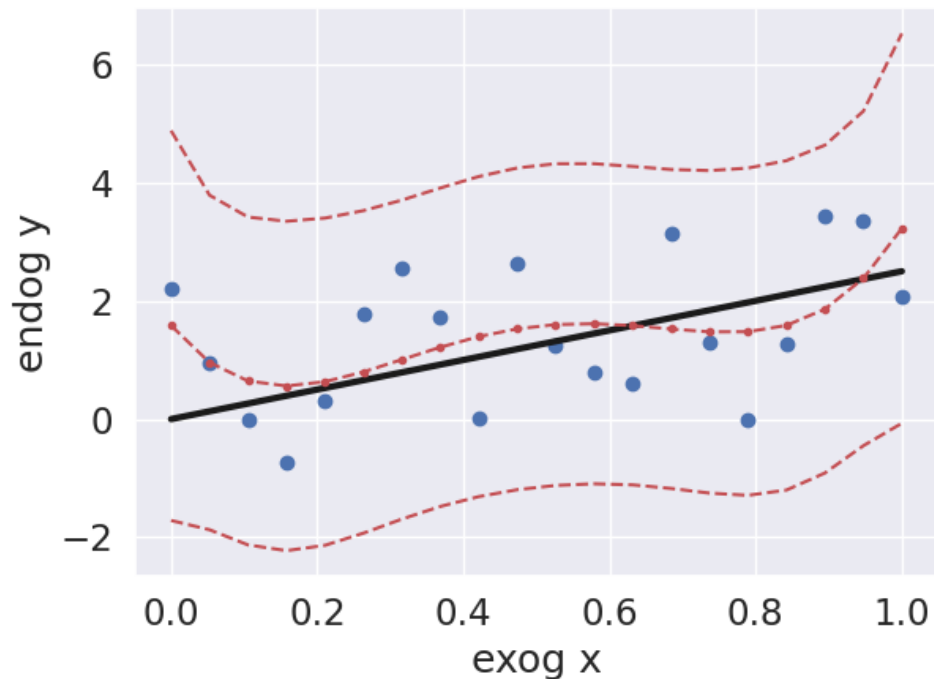


R-squared (uncentered): 0.261
F-statistic: 1.322
Prob (F-statistic): 0.307
Log-Likelihood: -28.757
BIC: 72.49

	coef	std err	t	P> t
x1	-15.4003	14.669	-1.050	0.310
x2	72.9756	61.976	1.177	0.257
x3	-112.9609	94.292	-1.198	0.250
x4	57.0286	46.753	1.220	0.241

Model Comparison

polynomial regression $y = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 x^2 + \dots$



LOWEST BIC VALUE INDICATES BETTER MODEL
(not R2, nor likelihood, etc.)

Building a Model with Formula

Building a Model with Formula

- The *patsy* package facilitates the construction of design matrices using R's formula
- The models can be built using *statsmodels.formula.api*

dataframe

	x1	x2	y
0	0.226948	0.358764	0.847936
1	0.531380	0.930131	-0.908091
2	1.484127	1.642451	0.428391
3	0.363713	0.023218	0.355144
4	-0.032484	-0.006205	0.577427

$$y = a_1 x_1 + a_2 x_2 + \epsilon$$

Building a Model with Formula

- The *patsy* package facilitates the construction of design matrices using R's formula
- The models can be built using *statsmodels.formula.api*

dataframe

	x1	x2	y
0	0.226948	0.358764	0.847936
1	0.531380	0.930131	-0.908091
2	1.484127	1.642451	0.428391
3	0.363713	0.023218	0.355144
4	-0.032484	-0.006205	0.577427

'y ~ x1 + x2'

design matrix

	Intercept	x1	x2
0	1.0	0.226948	0.358764
1	1.0	0.531380	0.930131
2	1.0	1.484127	1.642451
3	1.0	0.363713	0.023218
4	1.0	-0.032484	-0.006205

Building a Model with Formula

true values

```
=====
Dep. Variable:                y      R-squared:                0.484
Model:                        OLS    Adj. R-squared:           0.462
Method:                      Least Squares  F-statistic:             22.01
Date:                        Thu, 06 Jul 2023  Prob (F-statistic):      1.80e-07
Time:                        00:36:28    Log-Likelihood:          -62.306
No. Observations:            50      AIC:                     130.6
Df Residuals:                47      BIC:                     136.3
Df Model:                    2
Covariance Type:              nonrobust
=====
```

```
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
0.0 Intercept    -0.0413      0.125     -0.332      0.741     -0.292      0.209
-0.5 x1          -0.4097      0.118     -3.468      0.001     -0.647     -0.172
0.7 x2           0.5906      0.136      4.335      0.000      0.316      0.865
=====
```

```
=====
Omnibus:            1.032    Durbin-Watson:           2.003
Prob(Omnibus):      0.597    Jarque-Bera (JB):        0.986
Skew:               0.167    Prob(JB):                0.611
Kurtosis:           2.398    Cond. No.:               1.47
=====
```

Building a Model with Formula

- The *patsy* package facilitates the construction of design matrices using R's formula
- The models can be built using *statsmodels.formula.api*

dataframe

	x1	x2	y	y2
0	-0.598903	-1.107940	-0.983083	-0.717664
1	1.484895	-0.308404	-2.069647	-2.252826
2	-1.279121	-0.071924	0.927597	0.964397
3	0.796557	-1.757192	-0.702349	-1.262231
4	0.772527	0.049753	0.529516	0.544891

$$y_2 = a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 + \epsilon$$

Building a Model with Formula

- The *patsy* package facilitates the construction of design matrices using R's formula
- The models can be built using *statsmodels.formula.api*

dataframe

	x1	x2	y	y2
0	-0.598903	-1.107940	-0.983083	-0.717664
1	1.484895	-0.308404	-2.069647	-2.252826
2	-1.279121	-0.071924	0.927597	0.964397
3	0.796557	-1.757192	-0.702349	-1.262231
4	0.772527	0.049753	0.529516	0.544891

'y ~ x1 + x2'

design matrix

	Intercept	x1	x2
0	1.0	0.226948	0.358764
1	1.0	0.531380	0.930131
2	1.0	1.484127	1.642451
3	1.0	0.363713	0.023218
4	1.0	-0.032484	-0.006205

$$y_2 = a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 + \epsilon$$

Building a Model with Formula

true values
0.0
-0.5
0.7
0.4

```
=====
Dep. Variable:                y2      R-squared:                0.347
Model:                        OLS      Adj. R-squared:           0.319
Method:                        Least Squares      F-statistic:           12.48
Date:                        Thu, 06 Jul 2023      Prob (F-statistic):     4.48e-05
Time:                        00:36:52      Log-Likelihood:        -67.852
No. Observations:            50      AIC:                    141.7
Df Residuals:                47      BIC:                    147.4
Df Model:                    2
Covariance Type:              nonrobust
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept    -0.1561      0.139      -1.121      0.268     -0.436      0.124
x1           -0.3311      0.132      -2.509      0.016     -0.597     -0.066
x2            0.5106      0.152       3.354      0.002      0.204      0.817
=====
Omnibus:            0.529      Durbin-Watson:           2.231
Prob(Omnibus):      0.768      Jarque-Bera (JB):        0.634
Skew:              -0.016      Prob(JB):                0.728
Kurtosis:           2.449      Cond. No.                1.47
=====
```

Building a Model with Formula

- The *patsy* package facilitates the construction of design matrices using R's formula
- The models can be built using *statsmodels.formula.api*

dataframe

	x1	x2	y	y2
0	-0.598903	-1.107940	-0.983083	-0.717664
1	1.484895	-0.308404	-2.069647	-2.252826
2	-1.279121	-0.071924	0.927597	0.964397
3	0.796557	-1.757192	-0.702349	-1.262231
4	0.772527	0.049753	0.529516	0.544891

'y ~ x1 * x2'

design matrix

	Intercept	x1	x2	x1:x2
0	1.0	-0.598903	-1.107940	0.663548
1	1.0	1.484895	-0.308404	-0.457947
2	1.0	-1.279121	-0.071924	0.092000
3	1.0	0.796557	-1.757192	-1.399704
4	1.0	0.772527	0.049753	0.038436

Building a Model with Formula

```
=====
Dep. Variable:                y2      R-squared:                0.477
Model:                        OLS      Adj. R-squared:           0.443
Method:                        Least Squares      F-statistic:           14.01
Date:                          Thu, 06 Jul 2023      Prob (F-statistic):       1.28e-06
Time:                          00:36:41      Log-Likelihood:          -62.280
No. Observations:              50      AIC: 132.6
Df Residuals:                  46      BIC: 140.2
Df Model:                      3
Covariance Type:               nonrobust
=====
```

true values

0.0
-0.5
0.7
0.4

```
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept    -0.0335      0.131      -0.256      0.799      -0.297      0.230
x1            -0.4150      0.122     -3.405      0.001      -0.660     -0.170
x2             0.5960      0.140      4.259      0.000      0.314      0.878
x1:x2         0.4273      0.126      3.389      0.001      0.174      0.681
=====
```

```
=====
Omnibus:                0.909      Durbin-Watson:           1.989
Prob(Omnibus):           0.635      Jarque-Bera (JB):         0.917
Skew:                    0.164      Prob(JB):                 0.632
Kurtosis:                2.423      Cond. No.                 1.71
=====
```

Building a Model with Formula

MORE COMPLEX MODEL, BUT BETTER THAN WITHOUT INTERACTION!

```
=====
Dep. Variable:          y2      R-squared:          0.477
Model:                  OLS      Adj. R-squared:       0.443
Method:                 Least Squares  F-statistic:        14.01
Date:                   Thu, 06 Jul 2023  Prob (F-statistic):    1.28e-06
Time:                   00:36:41    Log-Likelihood:      -62.280
No. Observations:      50      AIC: 132.6
Df Residuals:          46      BIC: 140.2
Df Model:              3
Covariance Type:       nonrobust
=====
```

true values

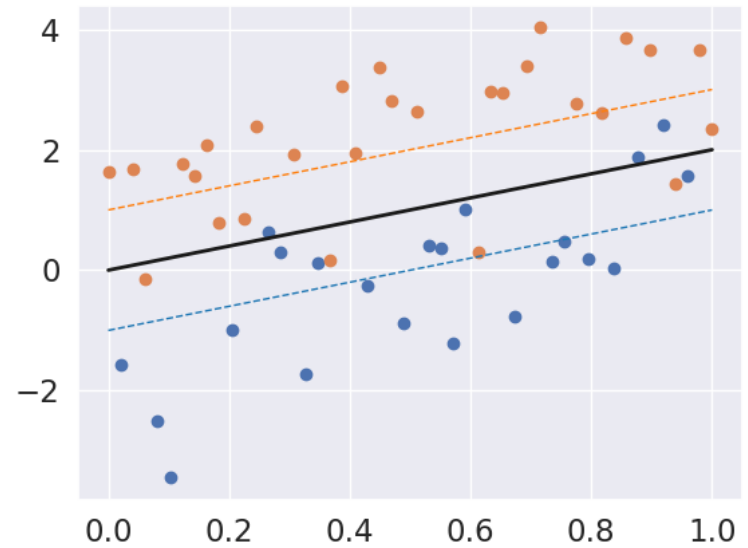
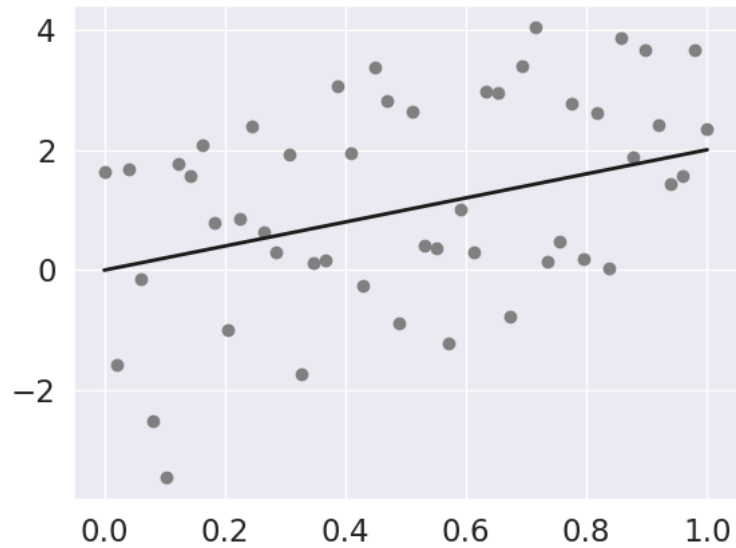
0.0
-0.5
0.7
0.4

```
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept    -0.0335      0.131      -0.256      0.799      -0.297      0.230
x1           -0.4150      0.122     -3.405      0.001     -0.660     -0.170
x2            0.5960      0.140      4.259      0.000      0.314      0.878
x1:x2         0.4273      0.126      3.389      0.001      0.174      0.681
=====
```

```
=====
Omnibus:      0.909      Durbin-Watson:      1.989
Prob(Omnibus): 0.635      Jarque-Bera (JB):      0.917
Skew:         0.164      Prob(JB):      0.632
Kurtosis:     2.423      Cond. No.      1.71
=====
```

Mixed Model

- Example of distinct baseline for 2 groups



Practice

notebook DESU_regression

Statistical Analysis in Python

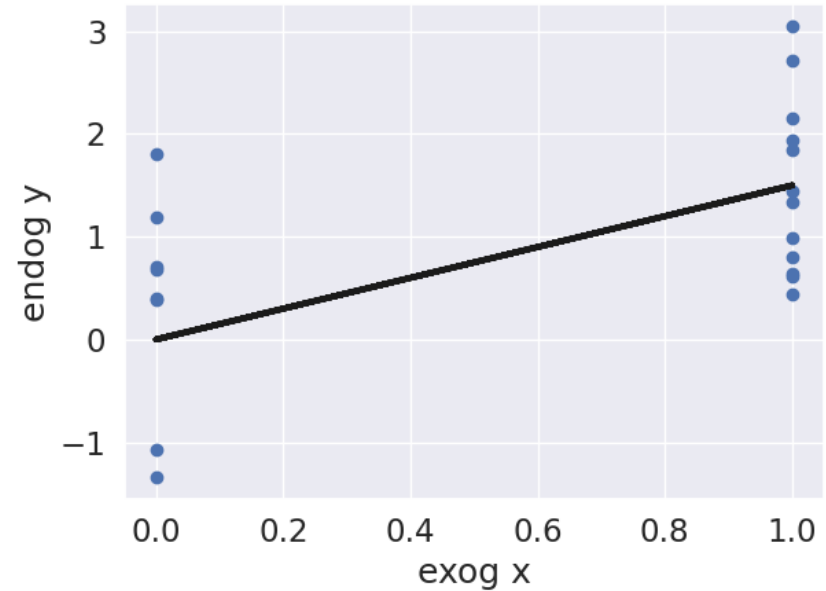
- Probabilities, distributions
- Parametric and non-parametric testing
- **Regressions**
 - **categorical predictor variables**
 - **ANOVA**
- Bayesian inference

Linear Regression with Categorical Variables

- Back to simple case
- Same number of samples in each group (n)

$$y = ax + \epsilon$$

$x \in \{0, 1\}$
group index



Linear Regression with Categorical Variables

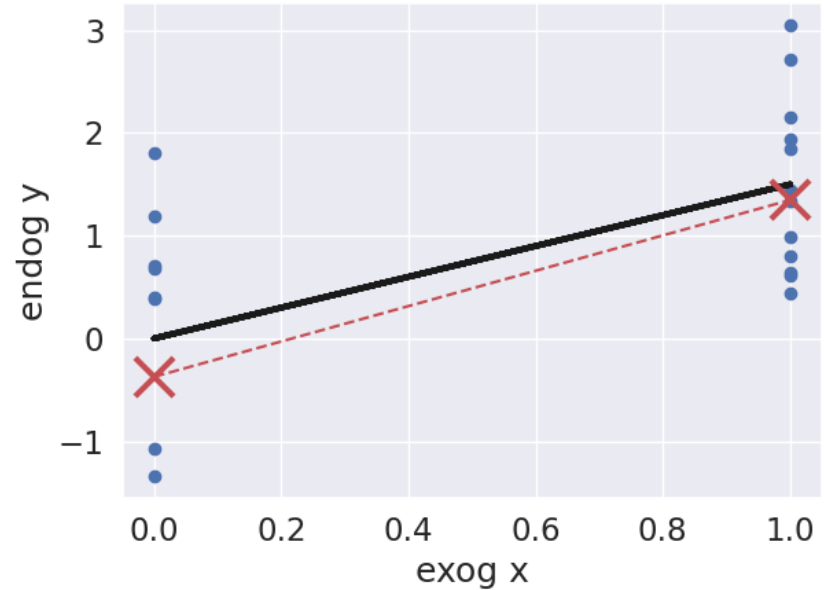
- Back to simple case
- Same number of samples in each group (n)

$$y = ax + \epsilon$$

$x \in \{0, 1\}$
group index

$$y_i = ax_i + \epsilon_i \quad \longrightarrow \quad \hat{a} = \bar{y}_1 - \bar{y}_0$$

\bar{y}_x average of group x



Linear Regression with Categorical Variables

- Back to simple case
- Same number of samples in each group (n)

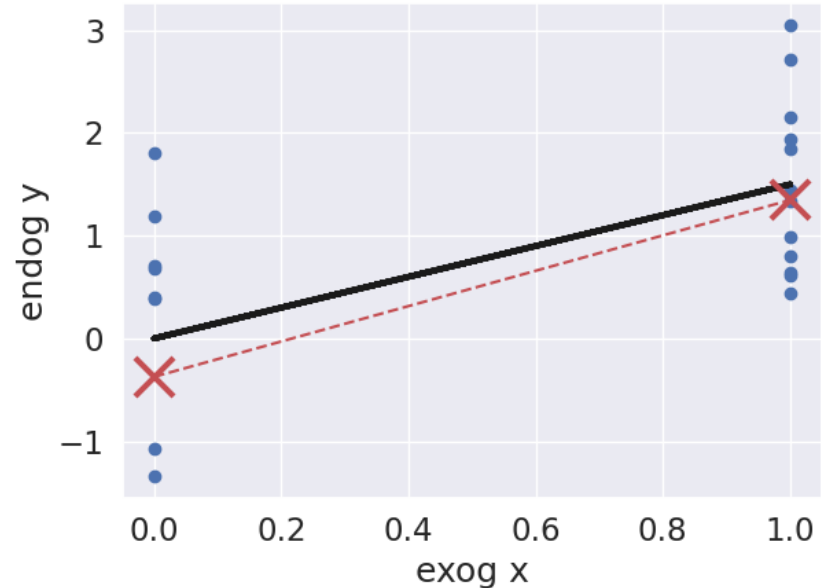
$$y = a x + \epsilon$$

$x \in \{0, 1\}$
group index

$$y_i = a x_i + \epsilon_i \quad \longrightarrow \quad \hat{a} = \bar{y}_1 - \bar{y}_0$$

\bar{y}_x average of group x

**ESTIMATE REFLECTS MEAN
DIFFERENCE BETWEEN GROUPS**



Linear Regression with Categorical Variables

- Back to simple case
- Same number of samples in each group (n)

$$y = a x + \epsilon$$

$$x \in \{0, 1\}$$

group index

t statistic for significance
(slope estimate non zero)

$$t = \frac{\hat{a} - 0}{s.e.} = \frac{\bar{y}_1 - \bar{y}_0}{\sigma_{est} / \sqrt{n}}$$

same as Student's t test with equal variance...

s.e. = standard error of estimate
(function of variance of estimate)

$$\sigma_{est} = var(\bar{y}_0) + var(\bar{y}_1)$$

$$= \frac{1}{n-1} \sum_{i \in x} (y_i - \bar{y}_x)^2$$

Linear Regression with Categorical Variables

- Back to simple case
- Same number of samples in each group (n)

$$y = a x + \epsilon$$

$$x \in \{0, 1\}$$

group index

F statistic for ANOVA:

$$F = \frac{\sum_x n (\bar{y}_x - \bar{y})^2 / 1}{\sum_{x,i} (y_i - \bar{y}_x)^2 / (2n - 2)}$$

BSS

WSS

with $\bar{y} = \frac{\bar{y}_1 + \bar{y}_0}{2}$

- explained sum of squares (between-group sum of squares, BSS)
- unexplained sum of square (within-group sum of squares, WSS)
- total sum of squares (TSS = BSS + WSS)
- note degrees of freedom

Linear Regression with Categorical Variables

- Back to simple case $y = a x + \epsilon$ $x \in \{0, 1\}$
- Same number of samples in each group (n) group index
- In fact, we have $F = t^2$ and the two tests give the same p-value!
 - BSS relates to \hat{a}
 - WSS relates to σ_{est}

Typology of Statistical Tests

- Sum of squares of normal random variables (e.g. residuals)
 - chi-square distribution (cf. degrees of freedom)

$$\sum_i \epsilon_i^2$$

- Ratio of sum of squares (e.g. explained versus unexplained variance)
 - F statistic

$$\frac{\sum_i \epsilon_i^2}{\sum_i \xi_i^2}$$

- Ratio of estimate by its variability (standard error) to test if
 - t statistic (and distribution)

$$\frac{\hat{a}}{s.e.(\hat{a})} = \frac{\hat{a}}{\sigma_{\hat{a}}/\sqrt{n}}$$

Statistical Analysis in Python

- Probabilities, distributions
- Parametric and non-parametric testing
- Regressions
- **Bayesian inference**
 - **model inversion**
 - **likelihood, posterior, prior**

Bayes' Rule

- Conditional probability definition for events A and B $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Corresponds to the probability of A knowing that B occurs (is true)
- The conditional probabilities are linked by Bayes' rule

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

- The same holds for random variables and pdf

Model, Parameter, Inference

- We consider a generative model that produce a signal x and has a parameter θ
 - what is the best model (best parameter) to explain my observations?
- Likelihood: probability to observe data x for model parameter
$$P(x|\theta)$$
- Posterior: probability that a model is parametrized by θ given that it has produced x
$$P(\theta|x)$$
- Prior: initial knowledge about the parameter distribution (e.g. physiological range)
$$P(\theta)$$

Model, Parameter, Inference

- We consider a generative model that produce a signal x and has a parameter θ
 - what is the best model (best parameter) to explain my observations?
- Likelihood: probability to observe data x for model parameter
- Posterior: probability that a model is parametrized by θ given that it has produced x
- Prior: initial knowledge about the parameter distribution (e.g. physiological range)

$$P(\theta)$$

- Bayes rule:
- $$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

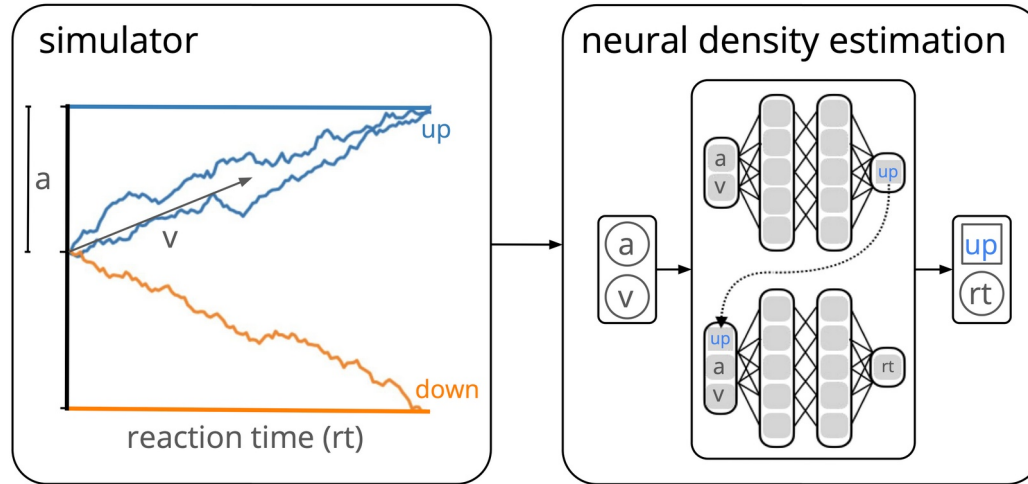
Parameter Inference

- Maximize posterior = likelihood x prior

$$\max_{\theta} [P(\theta|x)] = \max_{\theta} \frac{P(x|\theta) P(\theta)}{P(x)}$$

- in practice the denominator is a scaling factor that is ignored

Example of Simulation-Based Inference



model parameters
 a and v

model outputs
decision (up/down)
and reaction time (rt)

artificial neural
network to map
parameters to
observed outputs

Example of Simulation-Based Inference

