p-Values and introduction to correlations

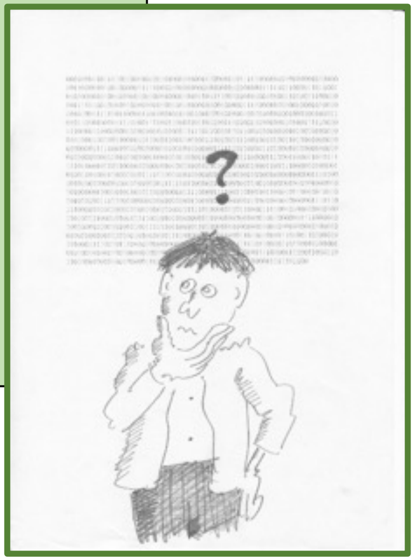
Statistics and data analysis

Ben Galili

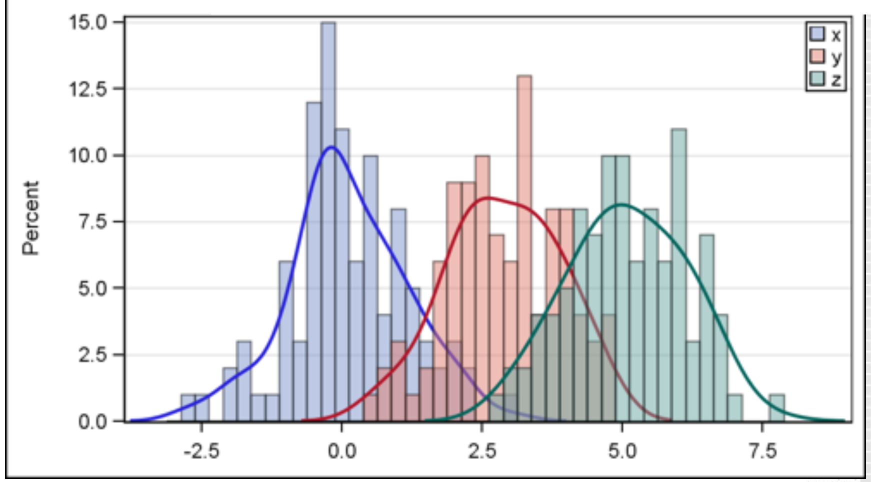
Leon Anavy

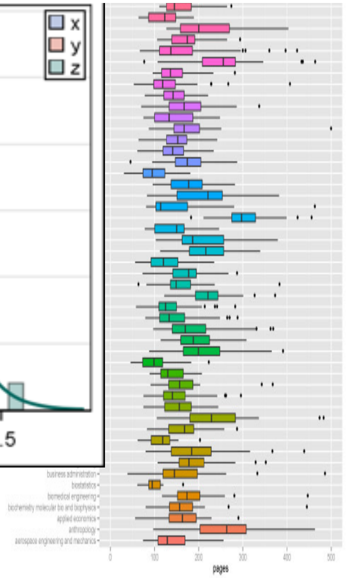
Zohar Yakhini

IDC, Herzeliya



0010011101010100101010100100100010
1010100010101111101011010011001001
1110101010011010010110010110110010





Outline

- Back to p-Values: definition and examples
- More about covariance
- Pearson correlation
- Spearman Correlation
- Testing for significance – empirical approaches
- Applications and approximation schemes

p-Value defined

In a randomized experiment the p-value is the probability that randomization alone, under some <u>stated null model</u>, led to a test statistic that is as extreme or more extreme than the <u>one observed</u>.

p-Value caveats

P-value $\neq$ $\Pr(H_0$ is true$)$!

$\Pr(A|B) \neq \Pr(B|A)$

"Probability of B given A"

$\Pr(\text{data}|H_0) \neq \Pr(H_0|\text{data})$

$\Pr(\text{clouds}|\text{rain}) \neq \Pr(\text{rain}|\text{clouds})$

p-Value caveats

No matter what question you *wish* the test would answer, a hypothesis test or null model testing *only answers one question*. Possible answers are

Not “This model is probably not true.”

Not “The effect of the drug is probably large.”

Not “female employees get paid more than male employees.”

Q: *Are the data consistent with the model (such that the observed results could <u>reasonably</u> have happened by chance)?*
A: *Yes (or No)*

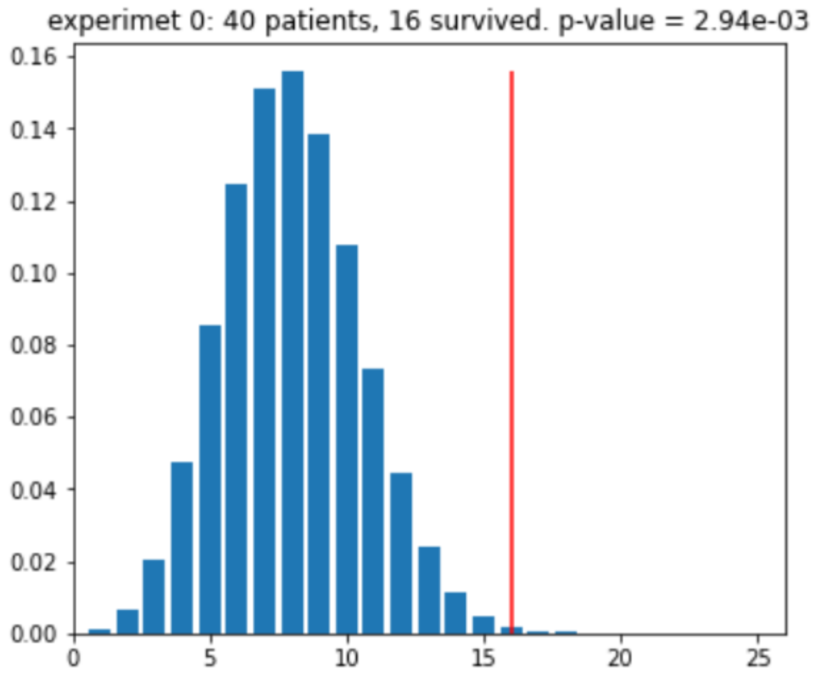
Sample size matters

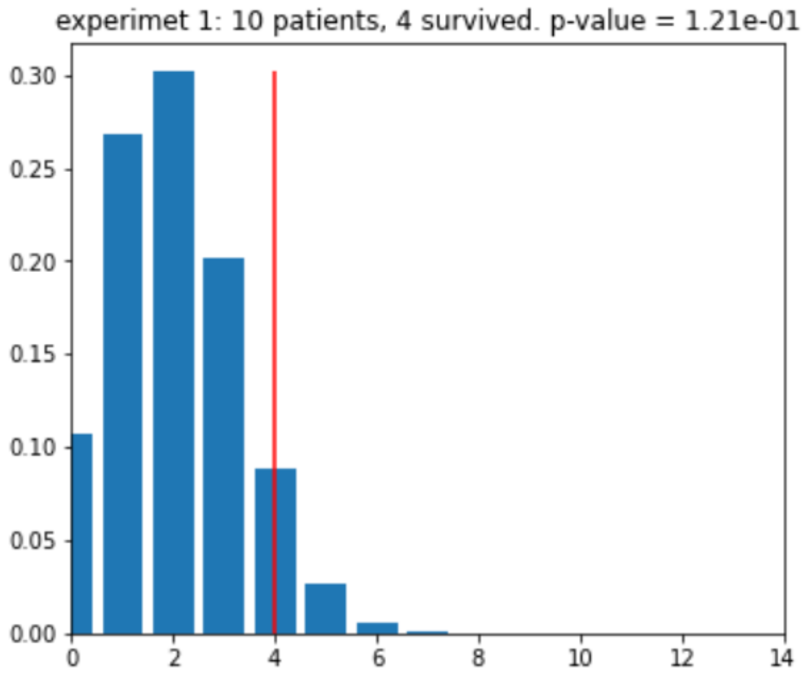
$$X \sim Binomial(n,p) \rightarrow X \dot\sim N\big(np, np(1-p)\big)$$

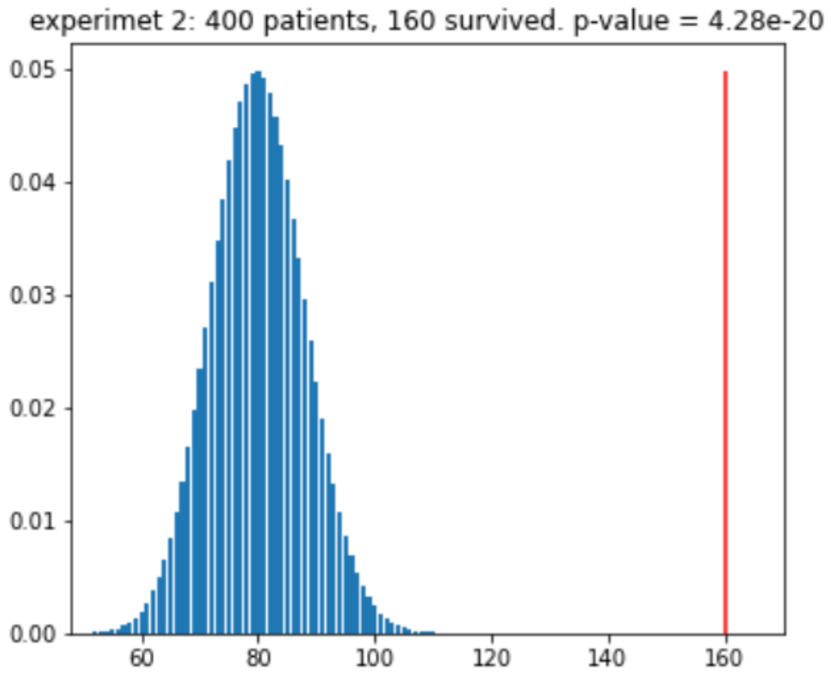
$$X \dot\sim N(8{,}6.4)$$
$$P(X \geq 16) \cong$$
$$P(Z > 2.96) = 0.0015$$

$$X \dot\sim N(2{,}1.6)$$
$$P(X \geq 4) \cong$$
$$P(Z > 1.19) = 0.12$$

$$X \dot\sim N(80{,}64)$$
$$P(X \geq 160) \cong$$
$$P(Z > 9.94) < 10^{-22}$$







Sample size matters

Say we want to test whether the percentage of American households having a cat equals to p=1/3

H_0: p =1/3 vs H_A: p $\neq$ 1/3.

We collected a sample of n=1000 households and calculate $\hat{p}$=0.31
Under our null model we drew n=1000 instances of a Bernoulli random variable with p=1/3. We then **averaged** them and received 0.31

$$P\left(\bar{X}_n - \frac{1}{3} \leq 0.31 - \frac{1}{3}\right) = ?$$

Sample size matters

Say we want to test whether the percentage of American households having a cat equals to p=1/3

H_0: p =1/3 vs H_A: p ≠ 1/3.

We collected a sample of n=1000 households and calculate $\hat{p}$=0.31
Under our null model we drew n=1000 instances of a Bernoulli random variable with p=1/3. We then **averaged** them and received 0.31

$$P\left(\bar{X}_n - \frac{1}{3} \leq 0.31 - \frac{1}{3}\right) = P\left(\frac{\sqrt{n}}{\sigma}\left(\bar{X}_n - \frac{1}{3}\right) \leq \frac{\sqrt{n}}{\sigma}\left(0.31 - \frac{1}{3}\right)\right) \approx$$

Sample size matters

H_0: $p = 1/3$ vs H_A: $p \neq 1/3$.

sample	n	Z	P-value
0.31	1000	-1.56	0.117
0.31	2000	-2.21	0.027
0.31	3000	-2.71	0.0067
0.31	10000	-4.9	0.00000007

Very strong evidence for H_A, but do we care that 31% ≠ 1/3?

$$\frac{\sqrt{n}}{\sigma}\left(0.31 - \frac{1}{3}\right)$$

So we should also care about effect size!
That's a topic for another day …

The difference between “significant” and “not significant” is not, in itself, statistically significant
(Gelman and Stern, 2006)

Correlations

Cauchy - Schwartz inequality

1. $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$
2. $\mathrm{Cov}(X,Y) = E((X - E(X))(Y - E(Y))$
3. $Var(X) = E\left(\left(X - E(X)\right)^2\right)$

For random variables U and V we have

$$[E(UV)]^2 \leq E(U^2)E(V^2)$$

Setting $U = X - EX$ and $V = Y - EY$ we get

$$[\mathrm{Cov}(X,Y)]^2 \leq V(X)V(Y)$$

And therefore

$$-1 \leq \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{V(X)^{1/2}V(Y)^{1/2}} \leq 1$$

Cauchy-Schwarz : proof

For any a we have:

$$0 \leq E\big((U - aV)^2\big) = E(U^2) - 2aE(UV) + a^2E(V^2)$$

Specifically, use this for $a = \frac{E(UV)}{E(V^2)}$:

$$0 \leq E(U^2) - 2\frac{\big(E(UV)\big)^2}{E(V^2)} + \frac{\big(E(UV)\big)^2}{E(V^2)} = E(U^2) - \frac{\big(E(UV)\big)^2}{E(V^2)}$$

QED …

Pearson correlation

Population:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

Sample, for a particular realization $(\boldsymbol{x},\boldsymbol{y})$ of n repeated sampling from (X,Y)

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n}\big(x_i-\mu(\boldsymbol{x})\big)\big(y_i-\mu(\boldsymbol{y})\big)}{\sqrt{\left(\sum_{i=1}^{n}\big(x_i-\mu(\boldsymbol{x})\big)^2\right)\left(\sum_{i=1}^{n}\big(y_i-\mu(\boldsymbol{y})\big)^2\right)}}$$

Multivariate Normal Distributions

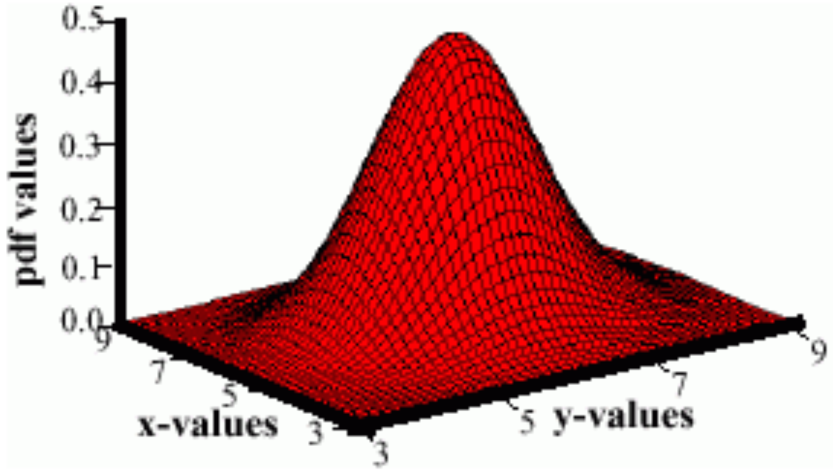
- **A multivariate normal distribution is defined by its pdf:**

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^{d} \cdot Det\,\Sigma}}\, exp\left(-\frac{1}{2}\cdot\langle\vec{x}-\vec{\mu}, \Sigma^{-1}(\vec{x}-\vec{\mu})\rangle\right)$$

where μ represents the mean (vector) and ∑ represents the covariance matrix.

In two dimensions

$$\Sigma = \begin{pmatrix} V(X) & Cov(X,Y) \\ Cov(X,Y) & V(Y) \end{pmatrix}$$

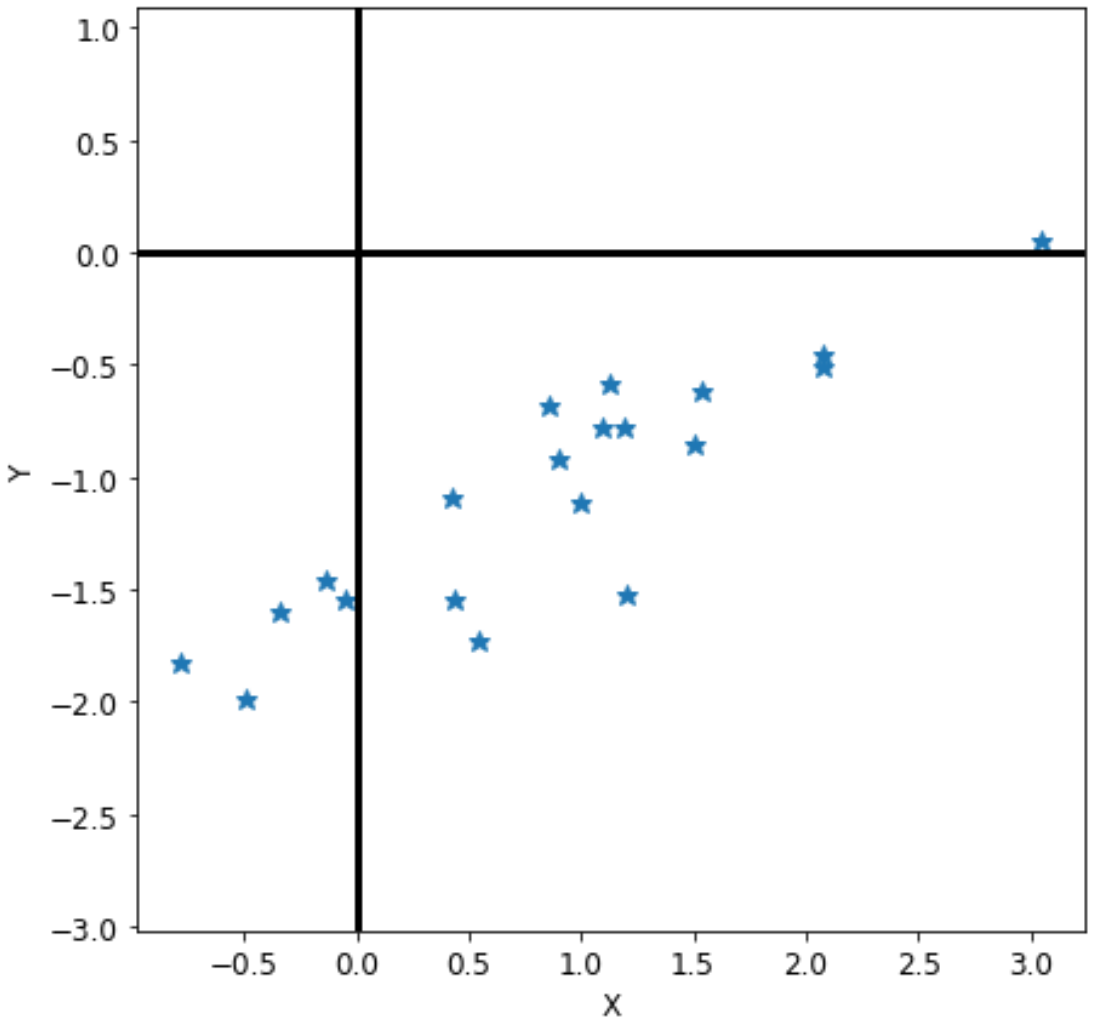


The sampling distribution of the Pearson correlation

```
from scipy.stats import multivariate_normal as mvarnorm
from scipy.stats import pearsonr
mu = [1,-1]
sigma = [[0.8,0.5],[0.5,0.4]]
X = mvarnorm(mu,sigma)
r = X.rvs(size = 20)
plt.figure(figsize=[7,7])
plt.plot(r[:,0],r[:,1],'*',markersize = 9)
plt.axhline(0,color='black')
plt.axvline(0,color='black')
plt.xlim([-1,4])
plt.ylim([-2.5,1.5])
plt.xlabel('X')
plt.ylabel('Y')
plt.axis('equal')
print(f'rho = {pearsonr(r[:,0],r[:,1])[0]:.2f}')
```

```
rho = 0.89
```

$$\rho(x,y) = 0.89$$

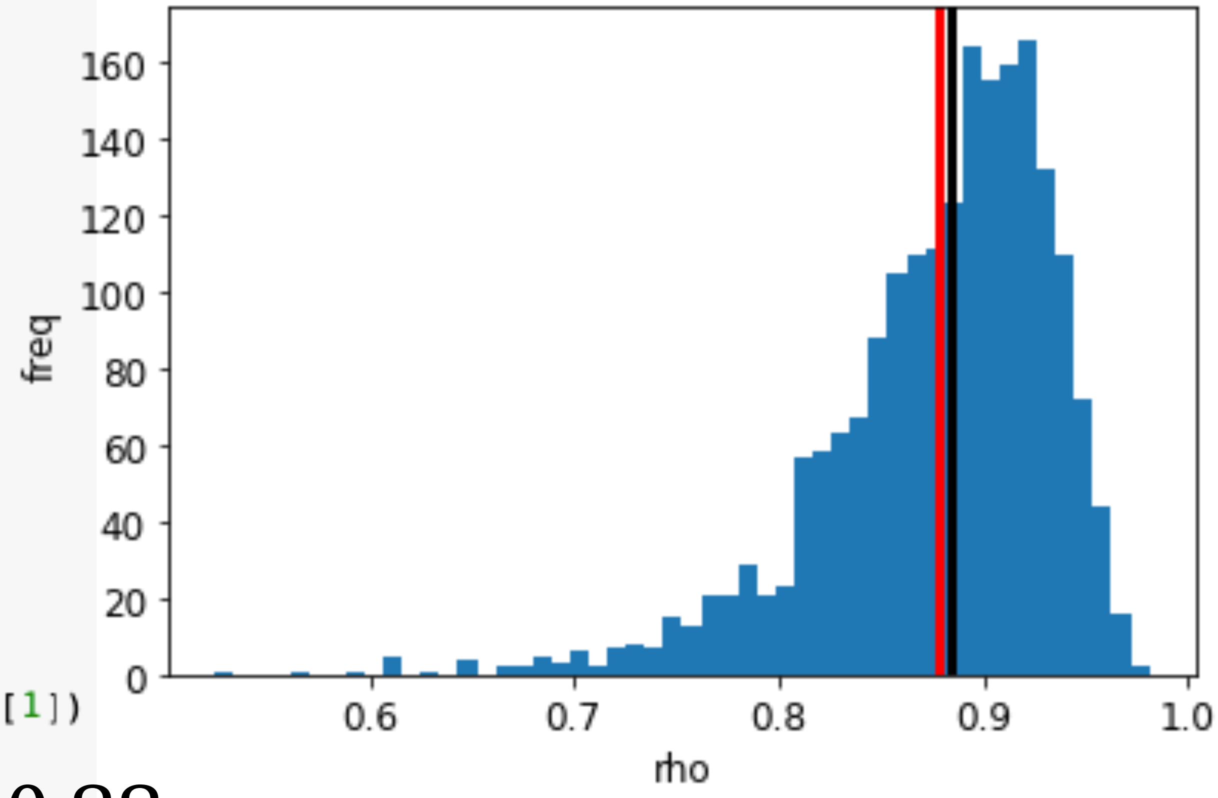


The sampling distribution of the Pearson correlation

```python
N = 2000
n = 20
mu = [1,-1]
sigma = [[0.8,0.5],[0.5,0.4]]
# sigma = [[0.8,0.3],[0.3,0.4]]
X = mvarnorm(mu,sigma)
rhos = np.empty([N,1])
for i in range(N):
    r = X.rvs(size = n)
    rho = pearsonr(r[:,0],r[:,1])
    rhos[i] = rho[0]

plt.hist(rhos,bins = 50)
plt.axvline(np.mean(rhos),color='red')
plt.xlabel('rho')
plt.ylabel('freq')
print(f'mean(rho) = {np.mean(rhos):.2f}')
rho_real = sigma[0][1] / np.sqrt(sigma[0][0]*sigma[1][1])
print(f'real(rho) = {rho_real:.2f}')
plt.axvline(rho_real,color='black')
```

```
mean(rho) = 0.88
real(rho) = 0.88
```



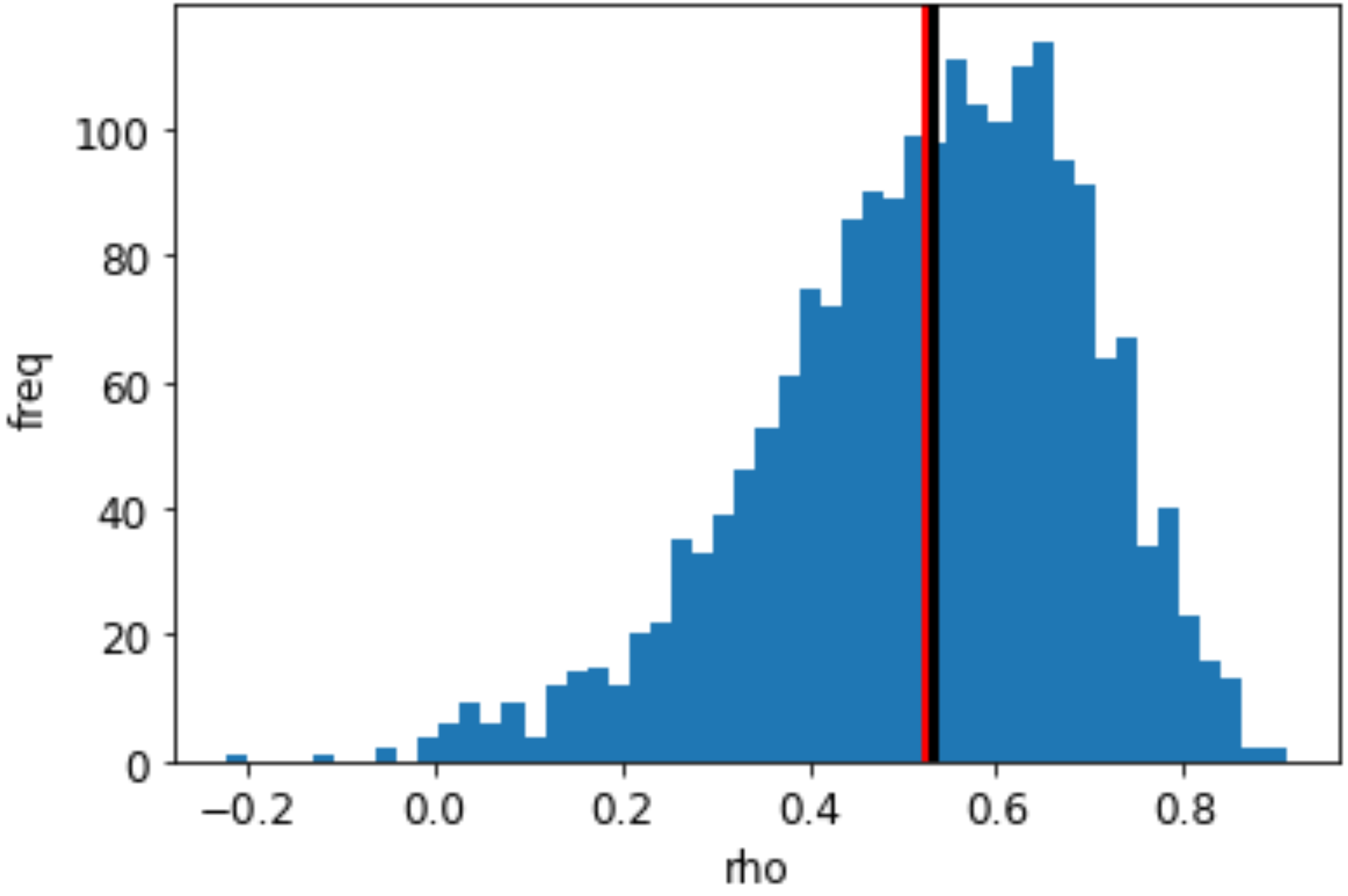
$$\rho(X,Y) = 0.88$$

$$\bar{\rho} = 0.88$$

The sampling distribution of the Pearson correlation

sigma = [[0.8,0.3],[0.3,0.4]]
n = 20

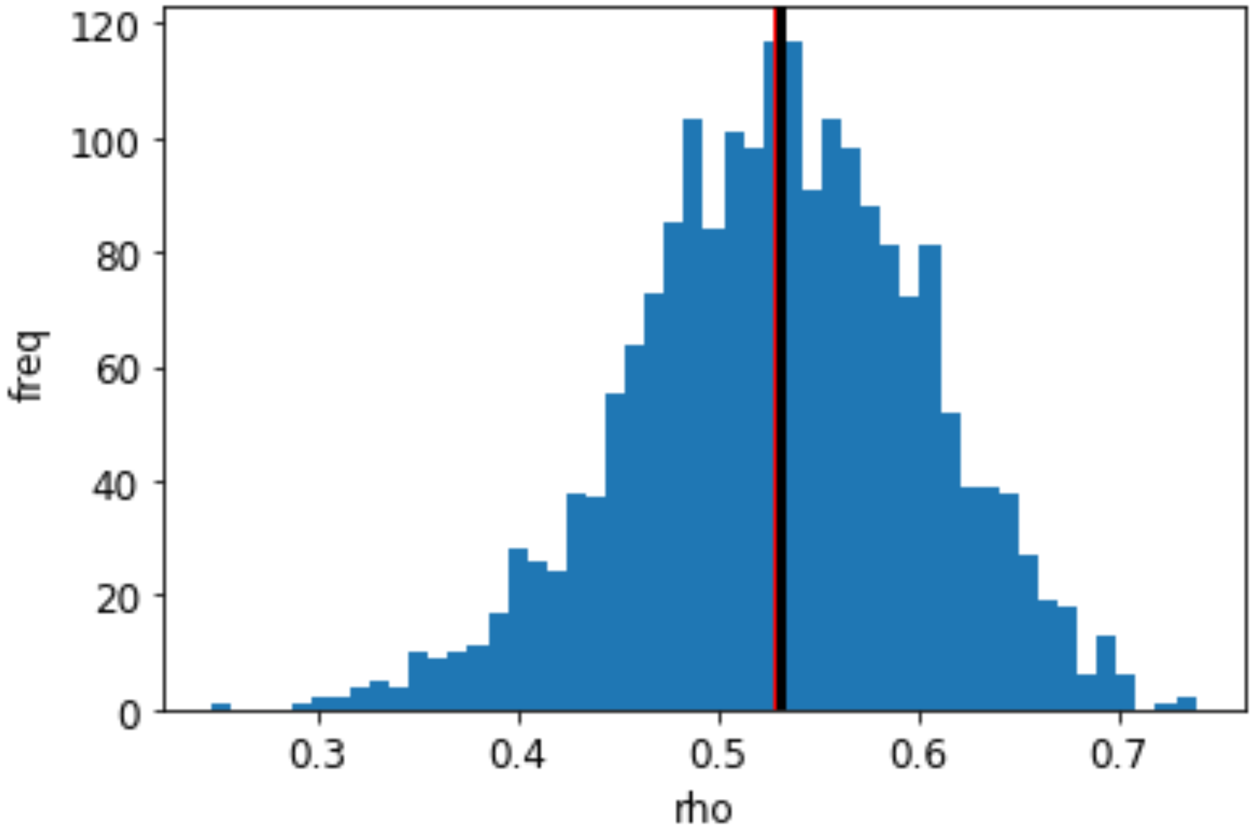
$$\rho(X,Y) = 0.53$$
$$\bar{\rho} = 0.52$$



The sampling distribution of the Pearson correlation

sigma = [[0.8,0.3],[0.3,0.4]]
n = 100

$$\rho(X,Y) = 0.53$$
$$\bar{\rho} = 0.53$$



Pearson correlation – when is it significant?

Is the sampling distribution of $\hat{\rho}$ normal?

Assess significance

- Empirically – draw permutations
- Exhaustively – draw all permutations
- Fisher Transform (next slide)
- Python, Matlab – actualy using the Fisher Transform

The Fisher Transform

$$F(r) = 0.5\ \ln\frac{1+r}{1-r}$$

Ronald Fisher
1890-1962

Thm (Fisher 1921):
If we start with (X,Y) that are close to bivariate normal then $F(\widehat{\rho_n})$, for i.i.d sampling, is normally distributed with mean $F\left(\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}\right)$ and a standard deviation of $\frac{1}{\sqrt{n-3}}$.

Assess significance

$$F(r) = 0.5\ \ln\frac{1+r}{1-r}$$

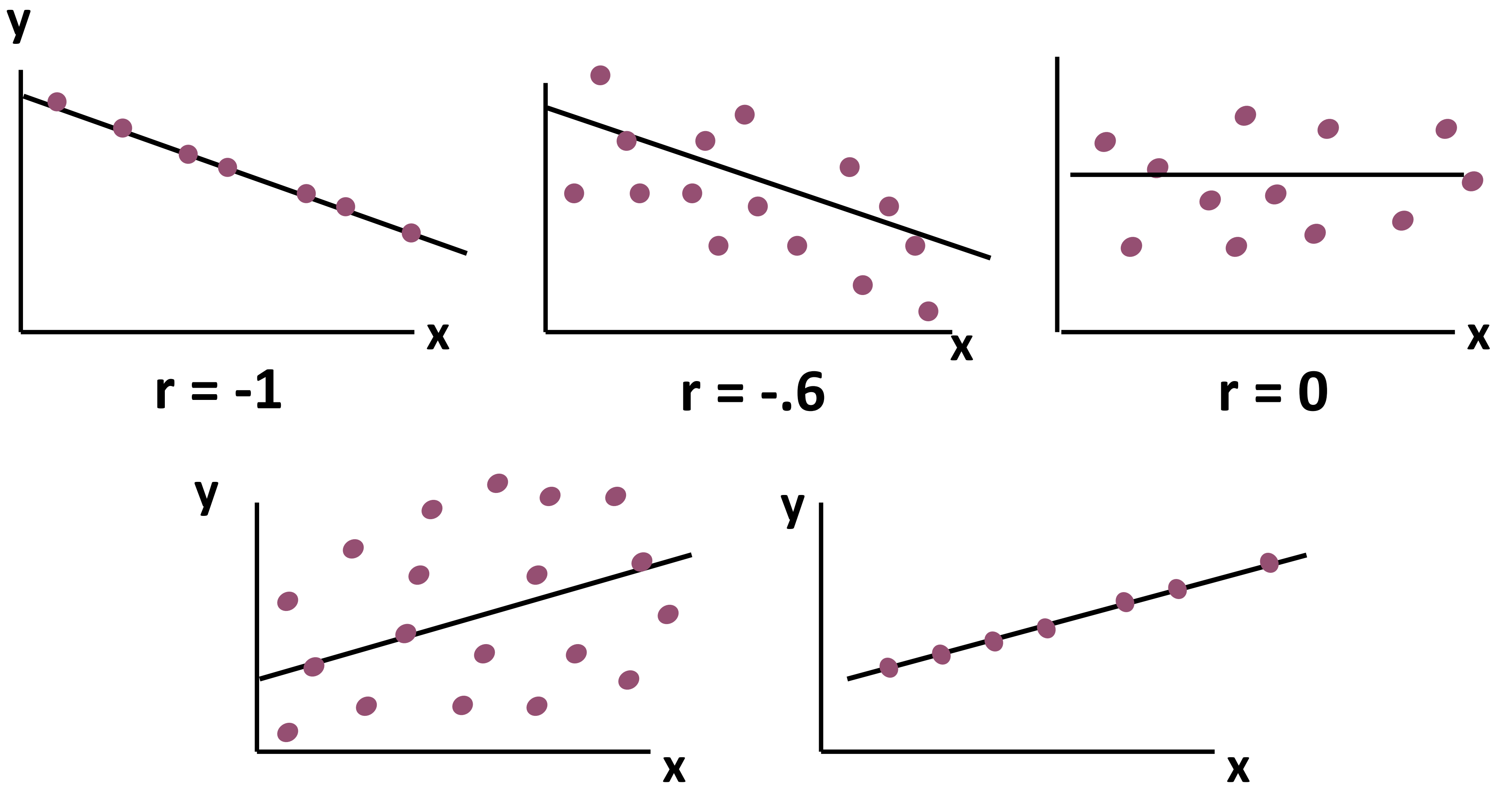
$$X \sim N\left(F(\rho), \sigma = \frac{1}{\sqrt{n-3}}\right)$$

$H_0: \rho = 0$

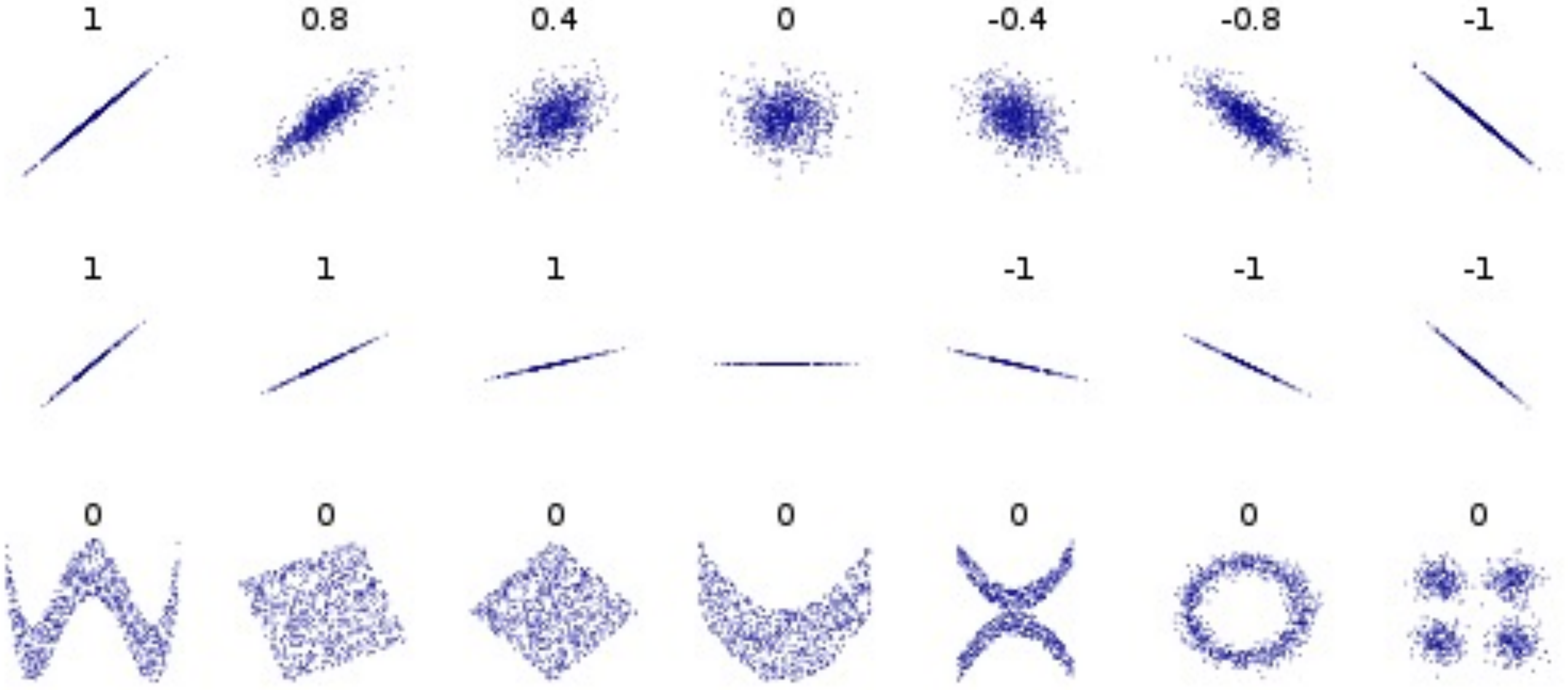
$H_1: \rho > 0$

P-value = $\mathrm{P}\left(\mathrm{X} \geq F(\widehat{\rho_n})\right) = \mathrm{P}\left(\frac{\mathrm{X}-0}{\frac{1}{\sqrt{n-3}}} \geq \frac{F(\widehat{\rho_n})-0}{\frac{1}{\sqrt{n-3}}}\right) = 1 - \Phi\left(\frac{F(\widehat{\rho_n})-0}{\frac{1}{\sqrt{n-3}}}\right)$

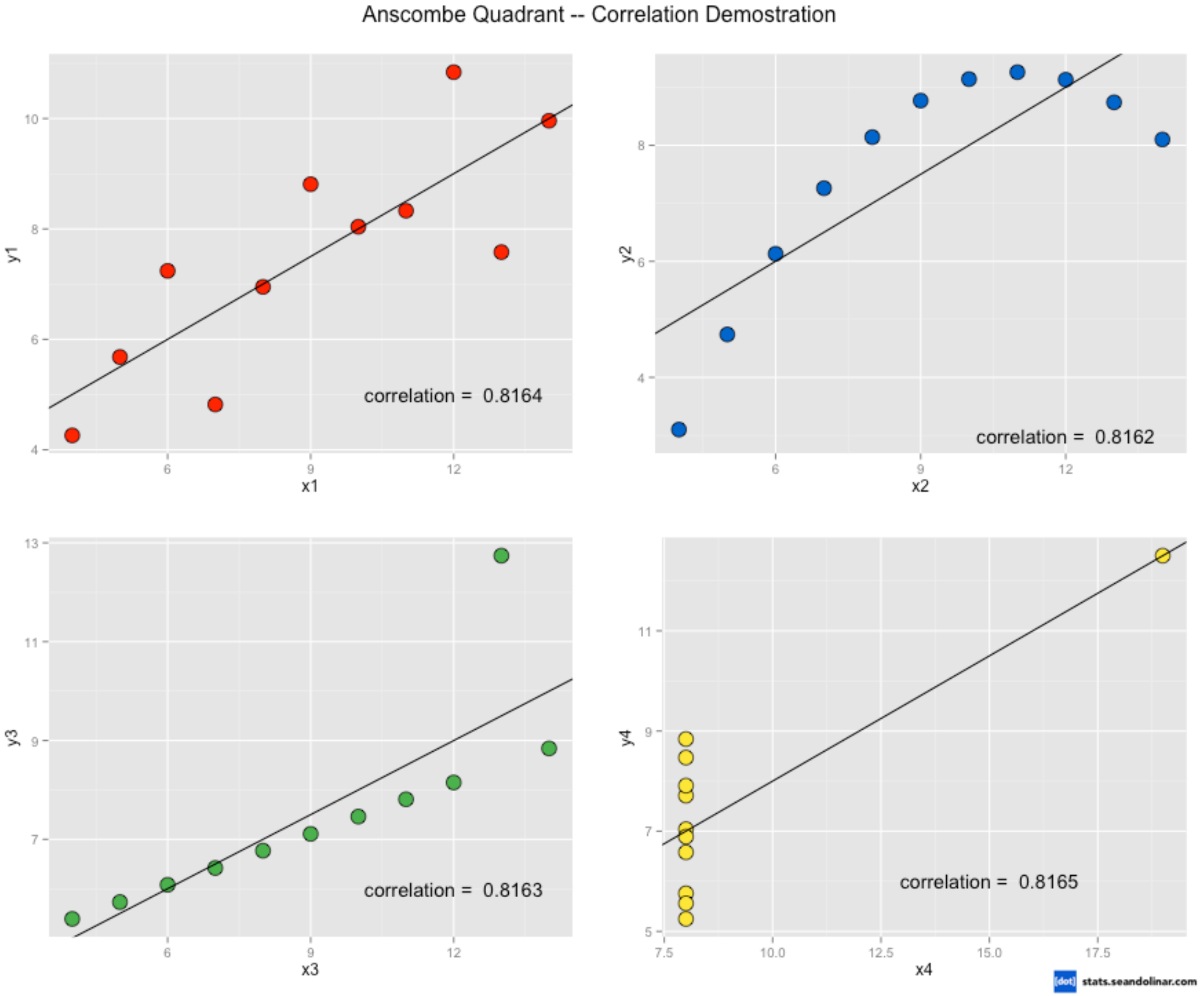
Examples of r Values

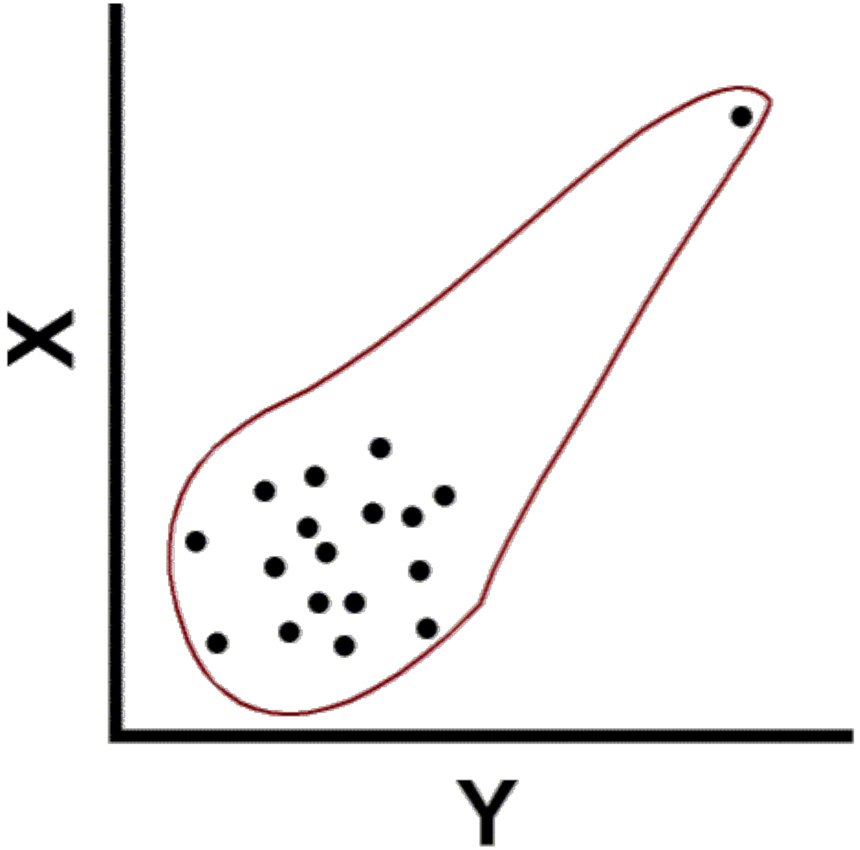


Examples of r Values

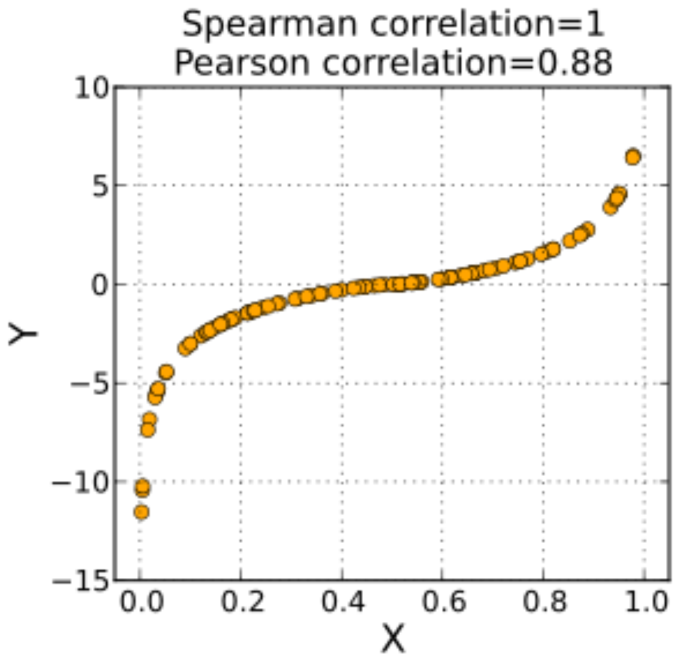


Examples of r Values





Spearman's Rank Correlation Coefficient



Pearson:

$$\rho(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mu(\boldsymbol{x}))(y_i - \mu(\boldsymbol{y}))}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \mu(\boldsymbol{x}))^2\right)\left(\sum_{i=1}^{n}(y_i - \mu(\boldsymbol{y}))^2\right)}}$$

Spearman:

$$\mathrm{SP}(x, y) = \frac{\sum_{i=1}^{n}(u_i - \frac{n+1}{2})(v_i - \frac{n+1}{2})}{\sum_{i=1}^{n}\left(u_i - \frac{n+1}{2}\right)^2}$$

$$u_i = rank(x_i)$$
$$v_i = rank(y_i)$$

The denominator?

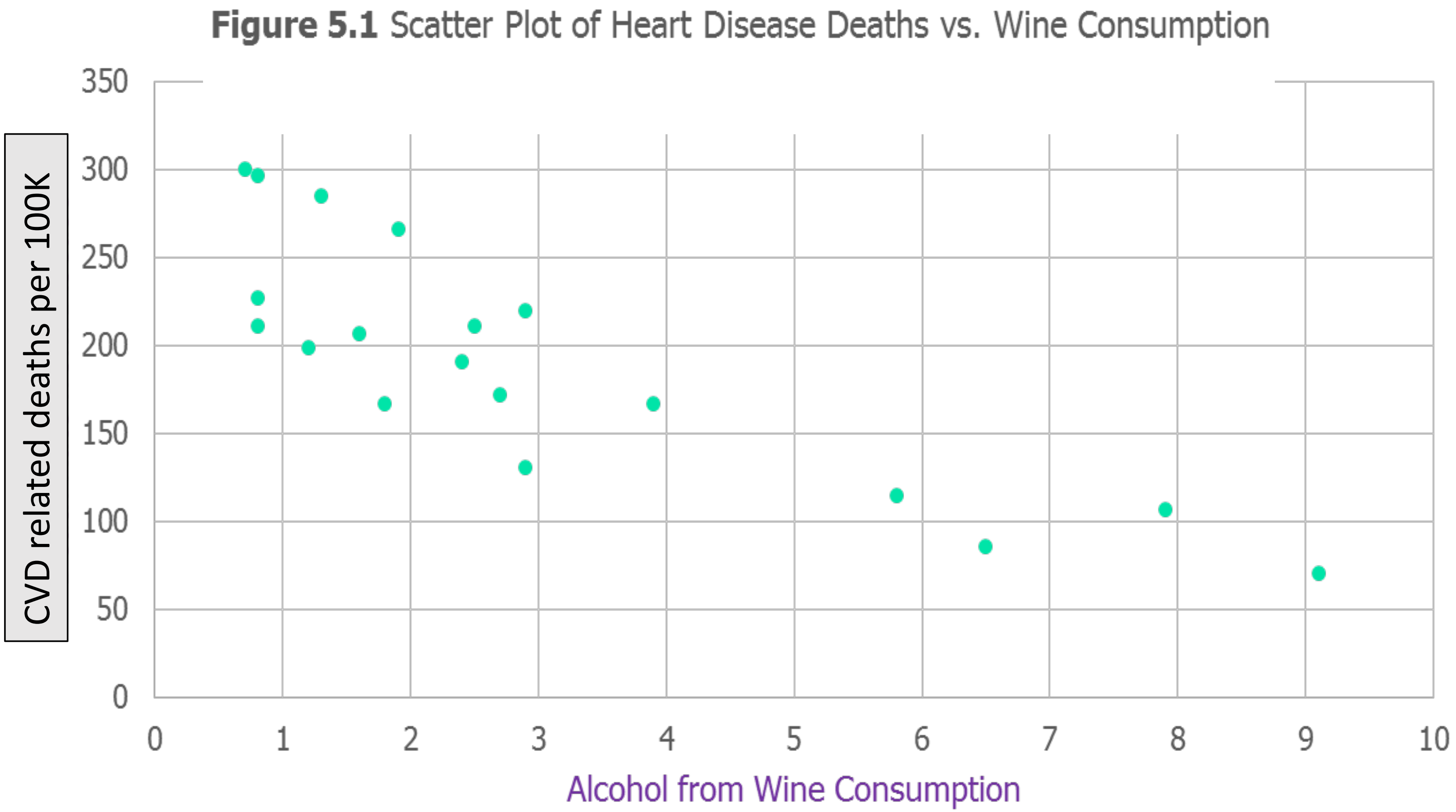
Is, in fact $\frac{n(n+1)(n-1)}{12}$, regardless of the actual $\boldsymbol{x}$ and $\boldsymbol{y}$

Example

Wine consumption and CVD mortality

Country	Alcohol from Wine	Heart disease deaths	Country	Alcohol from Wine2	Heart disease deaths3
Australia	2.5	211	Netherlands	1.8	167
Austria	3.9	167	New Zealand	1.9	266
Belgium	2.9	131	Norway	0.8	227
Canada	2.4	191	Spain	6.5	86
Denmark	2.9	220	Sweden	1.6	207
Finland	0.8	297	Switzerland	5.8	115
France	9.1	71	U.K.	1.3	285
Iceland	0.8	211	U.S.	1.2	199
Ireland	0.7	300	W. Germany	2.7	172
Italy	7.9	107			

Scatter plot



An alternative formula

If all ranks are distinct then we can use the formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where

$$d_i = rank(x_i) - rank(y_i)$$

Example - cont

- Convert values to ranks.
- Use average ranks for ties.

Ranks of wine consumption and CVD deaths

Alcohol

CVD deaths

Country i	$u_i = rank(x_i)$	$v_i = rank(y_i)$	$d_i = u_i - v_i$	Country i	$u_i = rank(x_i)$	$v_i = rank(y_i)$	$d_i = u_i - v_i$
1	11	12.5	-1.5	11	8	6.5	1.5
2	15	6.5	8.5	12	9	16	-7.0
3	13.5	5	-8.5	13	3	15	-12.0
4	10	9	1.0	14	17	2	15.0
5	13.5	14	-0.5	15	7	11	-4.0
6	3	18	-15.0	16	16	4	12.0
7	19	1	18.0	17	6	17	-11.0
8	3	12.5	-9.5	18	5	10	-5.0
9	1	19	-18.0	19	12	8	4.0
10	18	3	15.0				

Example - cont

We get $\hat{\rho}$ = - 0.826

Statistical assessment

- The NULL MODEL
- The p-value of the observation under the null model

The verbal question:

Are wine consumption and CVD death rates related?

The statistical question:

If we draw two independent and uniform permutations π and σ in S_{19},

then what is the probability of them having a Spearman ρ which is as

extreme as we observed here?

Statistical assessment

- The NULL MODEL
- The p-value of the observation under the null model

The verbal question:

Are wine consumption and CVD death rates negatively related?

The statistical question:

If we draw two independent and uniform permutations π and σ in S_{19},

then what is the probability of them having a Spearman ρ which is as

negative as we observed here?

Assess significance

- Empirically – draw permutations
- Exhaustively – draw all permutations
- Fisher Transform (next slide)
- Python, Matlab – actualy using the Fisher Transform

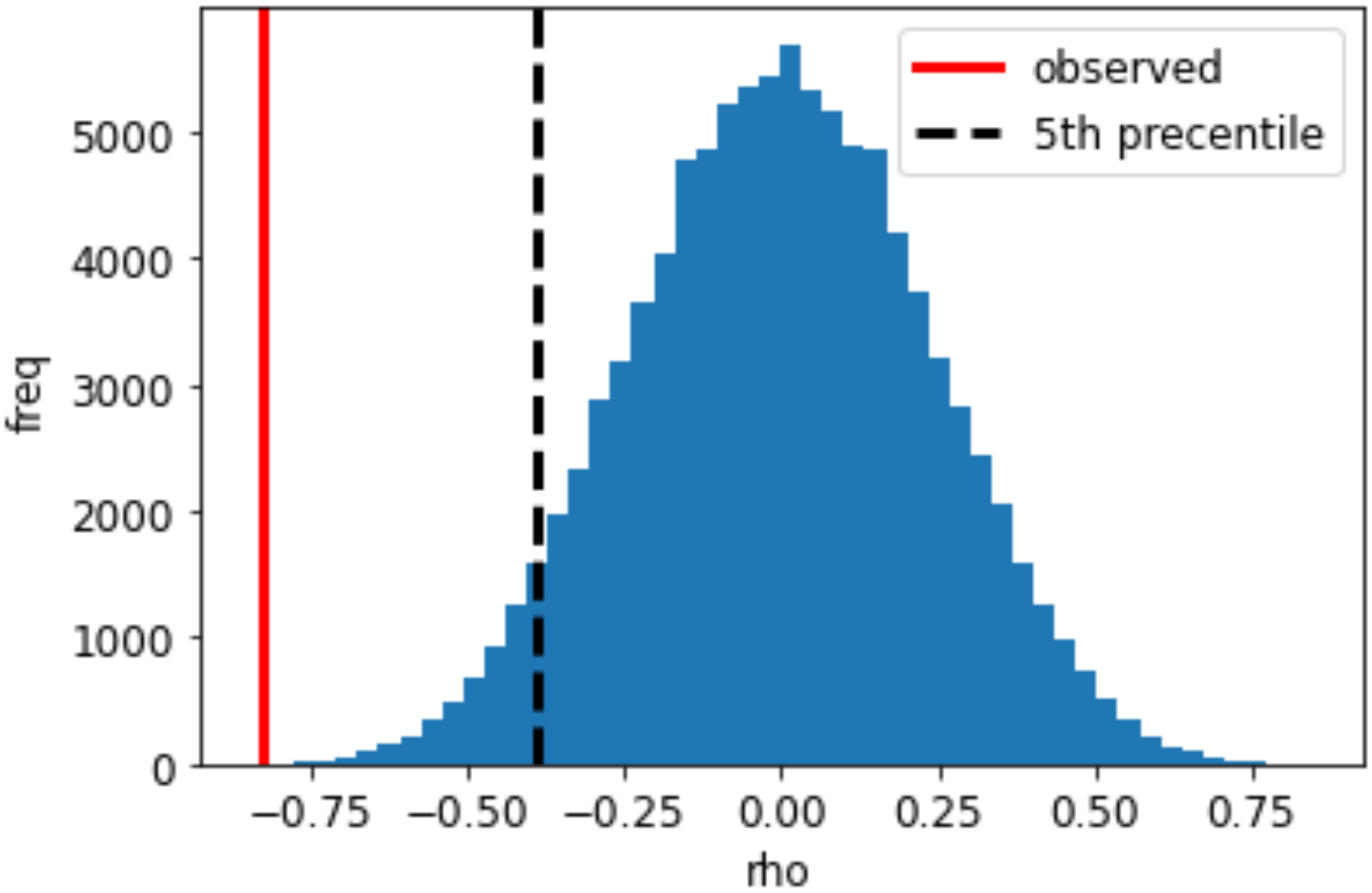
Empirically Assess significance

```python
N = 100000
n = 19
r = -0.826
rhos = np.empty([N,1])
for i in range(N):
    x = np.random.permutation(n)
    rho = spearmanr(range(n),x)
    rhos[i] = rho[0]
plt.hist(rhos,bins = 50)
plt.axvline(r,color='red',label = 'observed')
pval = sum(rhos[:,0] < r) / N
print(f'p-value = {pval:.2e}')
plt.xlabel('rho')
plt.ylabel('freq')

alpha_thresh = np.percentile(rhos,5)
plt.axvline(alpha_thresh,linestyle = '--',color='black',label = '5th precentile')
plt.legend()
```

```
p-value = 1.00e-05
```

Empirically Assess significance



Spearman p-values

Let σ, π be uniformly drawn permutations in S_n.

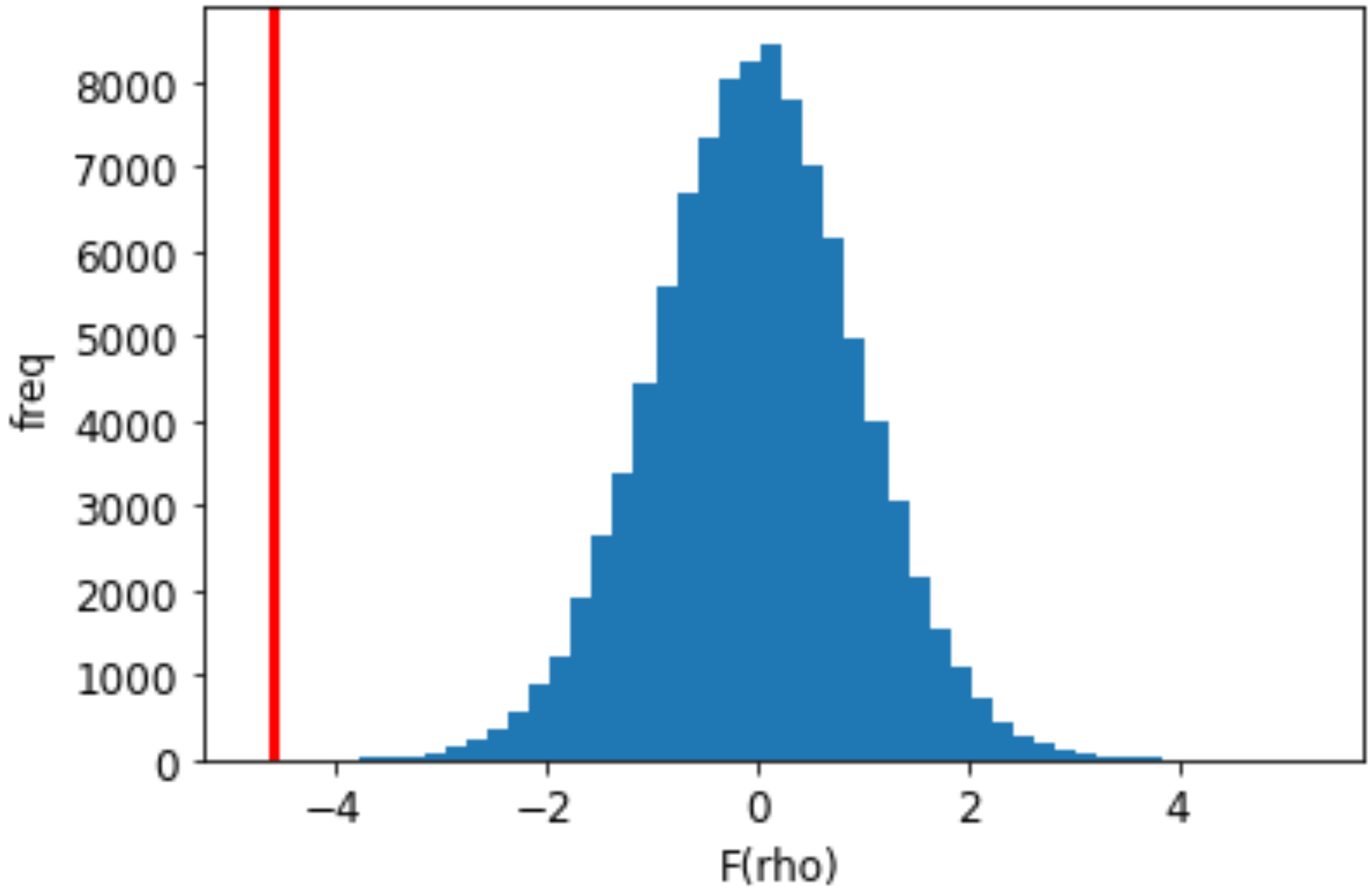
Let $\rho = \rho(\sigma, \pi)$ be their Spearman correlation.

If σ, π are independently drawn then

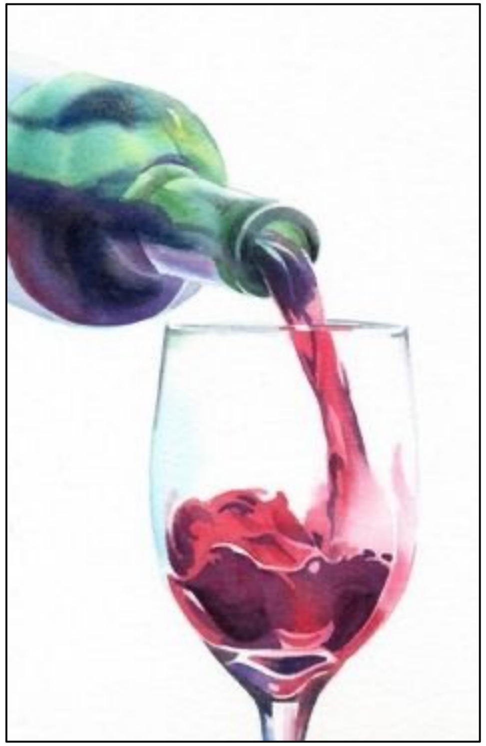
$$Z = F(\rho)\sqrt{\frac{n-3}{1.06}} \sim N(0,1)$$

where F is the Fisher Transform: $F(r) = \frac{1}{2}\ln\frac{1+r}{1-r}$

Spearman p-values



Wine consumption – conclusions …



In our case

$Z \approx -4.7$

so, using the Fisher transform we reject the hypothesis of independence or positive dependence at a p-value of $\approx 1.3\cdot10^{(-6)}$ …

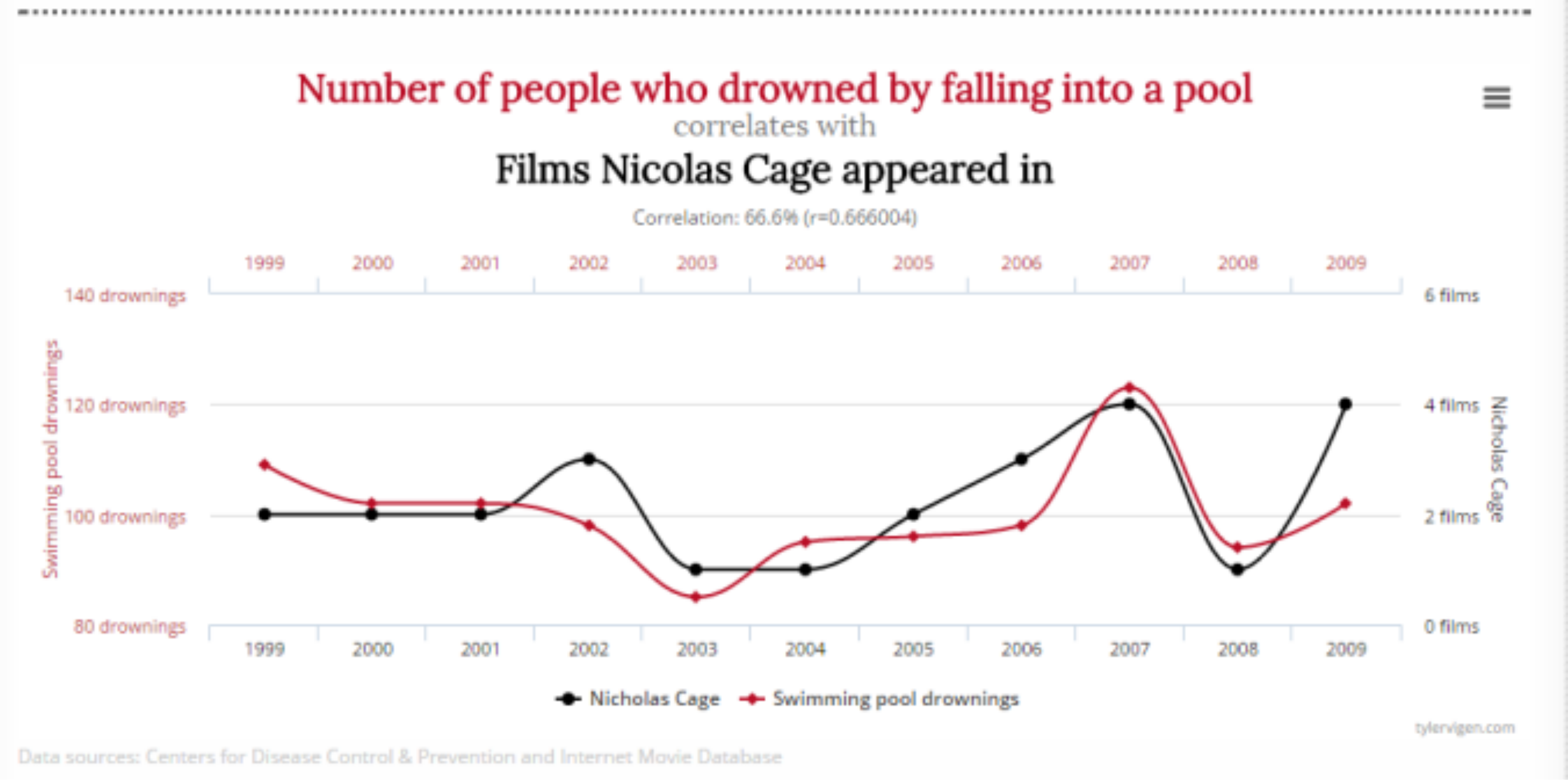
LeHaim!

CAVEAT
Exact characteristics of the data
Causality???
Confounding variables
Accuracy of measurement …

Strength AND Significance

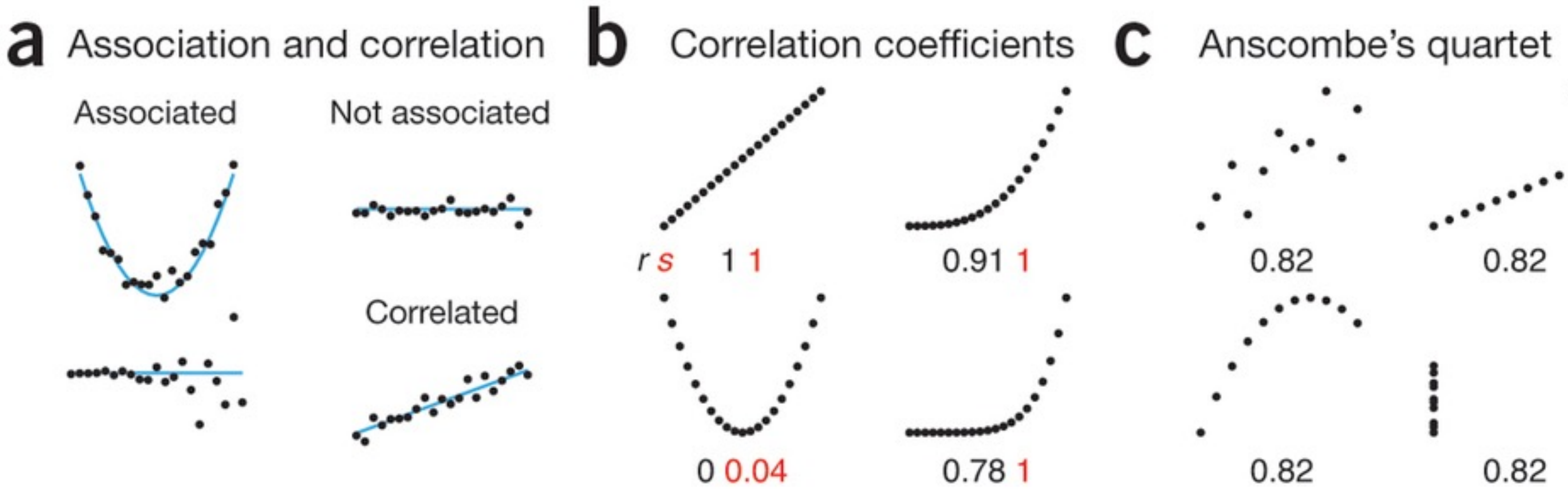
- We introduced mathematical measures of the STRENGTH of a relationship between two variables
- We discussed assessing statistical SIGNIFICANCE.
- Note that a relationship can be **strong** and yet **not** significant
- Conversely, a relationship can be **weak** but **significant**
- The key factor is the **size of the sample**.
- For small samples, it is easy to produce a strong correlation by chance and one must pay attention to significance to avoid jumping to spurious conclusions.
- For large samples, it is easy to achieve significance, and one must pay attention to the strength of the correlation to determine if the relationship explains much about the data.

Correlation is NOT (necessarily) causation



Modern association measures

- Reshef et al, Detecting Novel Associations in Large Data Sets, Science 2011
- Altman and Krzywinski, Points of Significance: Association, correlation and causation, Nature Methods 2015



Summary

- Correlation measures are mathematical tools that help quantify relationships between different aspects of observed data
- While quantitative, they still need to be followed by statistical assessment to yield significance under a model
- We can often compute p-values for observed correlations/associations
- Pearson correlation is the classical and most popular correlation coefficient. It's a sample version of the (normalized) population covariance
- Parameter free rank-based measures of correlation are often cleaner
- And finally – always assess correlations with a good measure of skepticism

Next week

- Correlations for multinomials
- Kendall correlation
- Kendall-Knight algorithm
- More on modern measures of association
- More examples