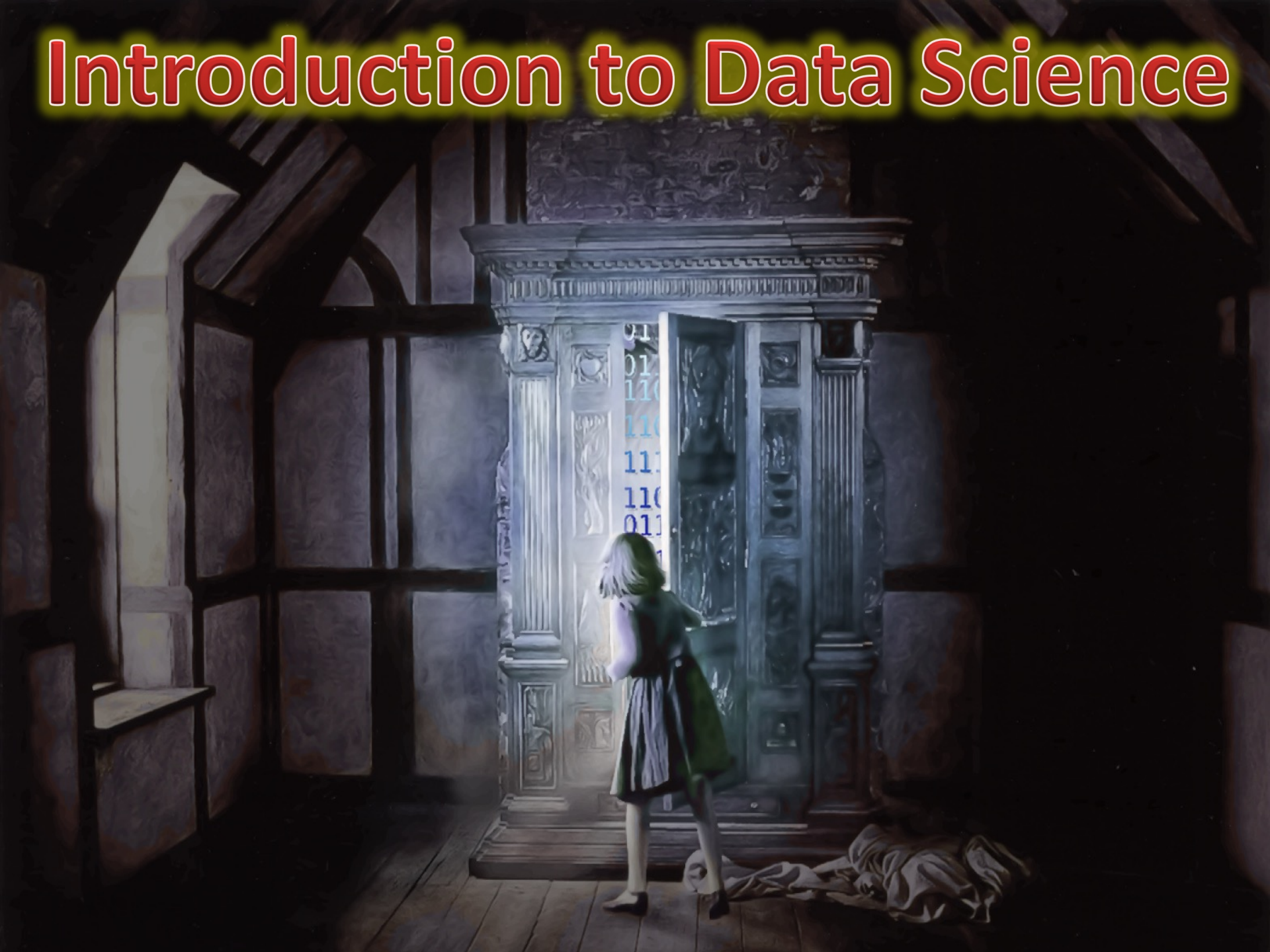
Smallest font

Please turn off and put away your cell phone

Calibration slide

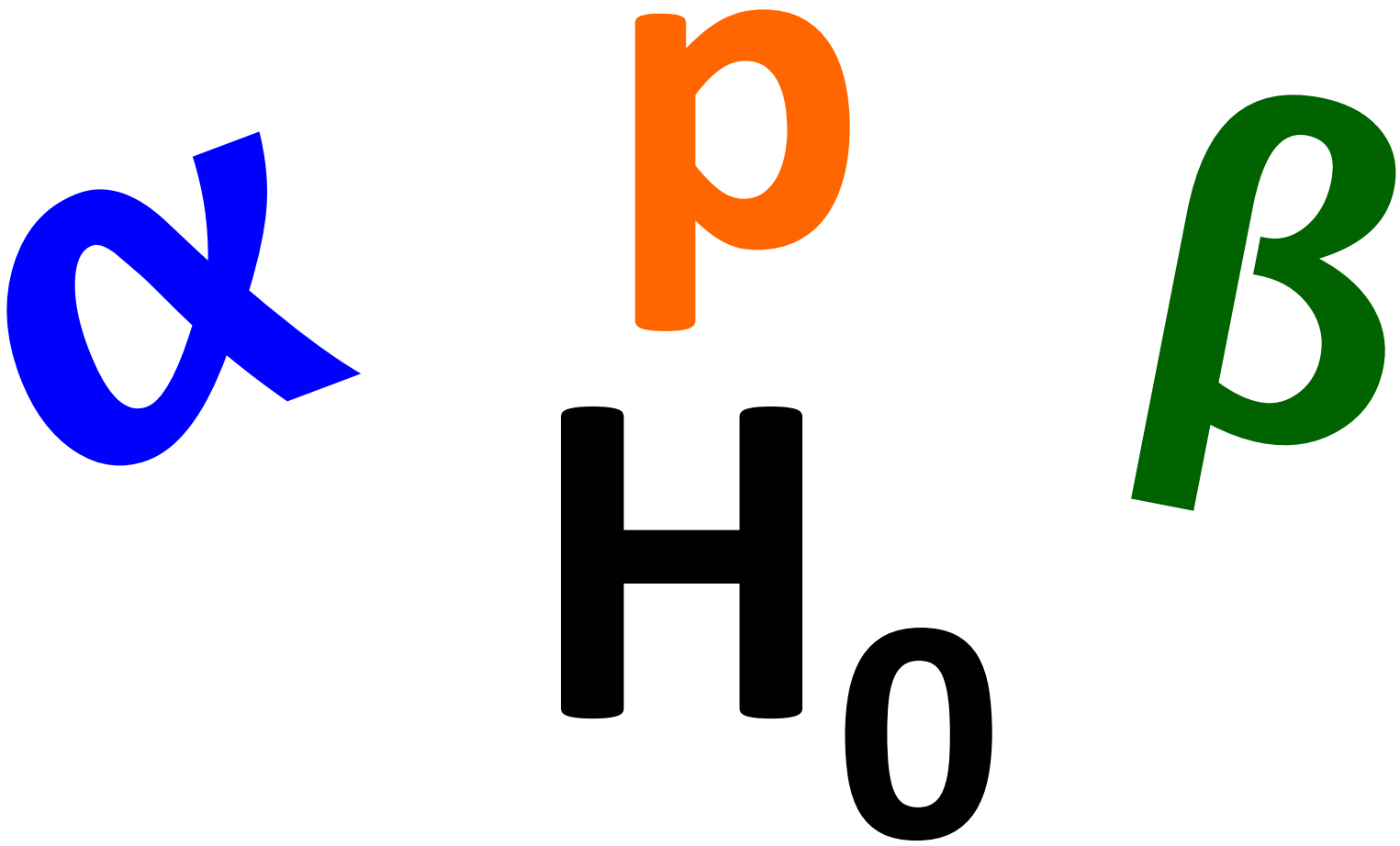
Smallest font

Introduction to Data Science



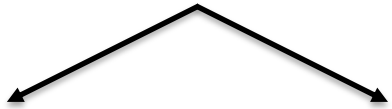
Significance tests

- focusing on A/B test designs -



The approach we introduced last time has been turned into a recipe of how to test hypotheses:

1. Start by proposing a hypothesis, e.g. that a "treatment" is effective (e.g. by implementing an A/B test).
2. Assume that the treatment has no effect (the null hypothesis)
3. Administer the treatment to the treatment group
4. Measure the outcomes of the treatment group and those of the control group.
5. Compare the outcomes in the two groups. Recognize that any difference could be due to chance (sampling variability)



6a. If the difference in outcomes is too large to be plausibly consistent with chance, reject the null hypothesis (and conclude that the treatment had an effect)

6b. If the difference in outcomes is plausibly consistent with chance alone, we don't conclude anything about the treatment (we already assumed it doesn't work)

This framework has become (and still is) the predominant approach to test hypotheses in science and industry

What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?

John P. A. Ioannidis

https://doi.org/10.1080/00031305.2018.1447512

PUBLISHED ONLINE:
20 March 2019

The results presented in this section are based on a survey of the entire biomedical literature published during the quarter-century from 1990 to 2015. Text mining was used to assess the presence of *P* values in the abstracts of 16.2 million items (13.0 million of which had an abstract). Similar text mining was performed in PubMed Central (PMC) for 844,000 full-text articles. For details, see ref. 3. Across this large

Overall (all papers)	51.1%
Articles published in core clinical journals	78.4%
Meta-analyses	82.8%
Randomized controlled trials	76.0%
Other clinical trials (excluding randomized controlled trials)	75.7%

However, there are many different use cases, which necessitate specific different tests.

Few assumptions, based on counting: The **sign test**

- John Arbuthnot (1710!) *wanted to know* whether male and female births are equally likely or not.
- Null hypothesis H_0: They are equally likely – p = 0.5.
- What did he *do*? He looked at London birth records from 1629 to 1710, involving ~0.5 million births over 82 years.
- What did he *find*? Male births were more common than female births in all 82 years of records.
- Assuming a probability of 0.5 (as per the null hypothesis), the probability of observing this outcome is 0.5^{82} ($p<2.1e-25$)
- Arbuthnot's *interpretation*: It is not plausible that this outcome is due to chance alone, so we can reject the H_0 – instead, Arbuthnot concludes that "god provides".

John Arbuthnot (1710):

"An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes"

	Christened.	
Anno.	*Males.*	*Females.*
1629	5218	4683
30	4858	4457
31	4422	4102
32	4994	4590
33	5158	4839
34	5035	4820
35	5106	4928
36	4917	4605
37	4703	4457
38	5359	4952
39	5366	4784
40	5518	5332
41	5470	5200
42	5460	4910
43	4793	4617
44	4107	3997
45	4047	3919
46	3768	3536
47	3796	3536

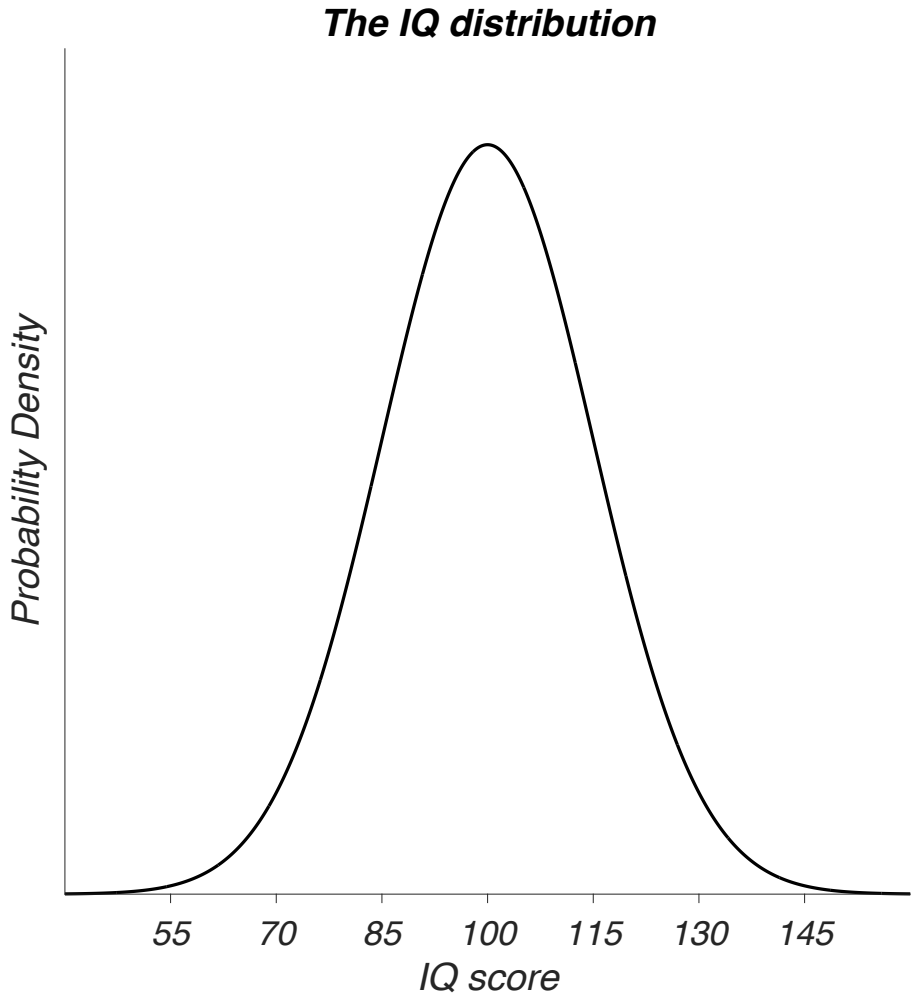
	Christened.	
Anno.	*Males.*	*Females.*
1648	3363	3181
49	3079	2746
50	2890	2722
51	3231	2840
52	3220	2908
53	3196	2959
54	3441	3179
55	3655	3349
56	3668	3382
57	3396	3289
58	3157	3013
59	3209	2781
60	3724	3247
61	4748	4107
62	5216	4803
63	5411	4881
64	6041	5681
65	5114	4858
66	4678	4319

	Christened.	
Anno.	*Males.*	*Females.*
1667	5616	5322
68	6073	5560
69	6506	5829
70	6278	5719
71	6449	6061
72	6443	6120
73	6073	5822
74	6113	5738
75	6058	5717
76	6552	5847
77	6423	6203
78	6568	6033
79	6247	6041
80	6548	6299
81	6822	6533
82	6909	6744
83	7577	7158
84	7575	7127
85	7484	7246
86	7575	7119
87	7737	7214
88	7487	7101

	Christened.	
Anno.	*Males.*	*Females.*
1689	7604	7267
90	7909	7302
91	7662	7392
92	7602	7316
93	7676	7483
94	6985	6647
95	7263	6713
96	7632	7229
97	8062	7767
98	8426	7626
99	7911	7452
1700	7578	7061
1701	8102	7514
1702	8031	7656
1703	7765	7683
1704	6113	5738
1705	8366	7779
1706	7952	7417
1707	8239	7623
1708	8239	7623
1709	7840	7380
1710	7640	7288

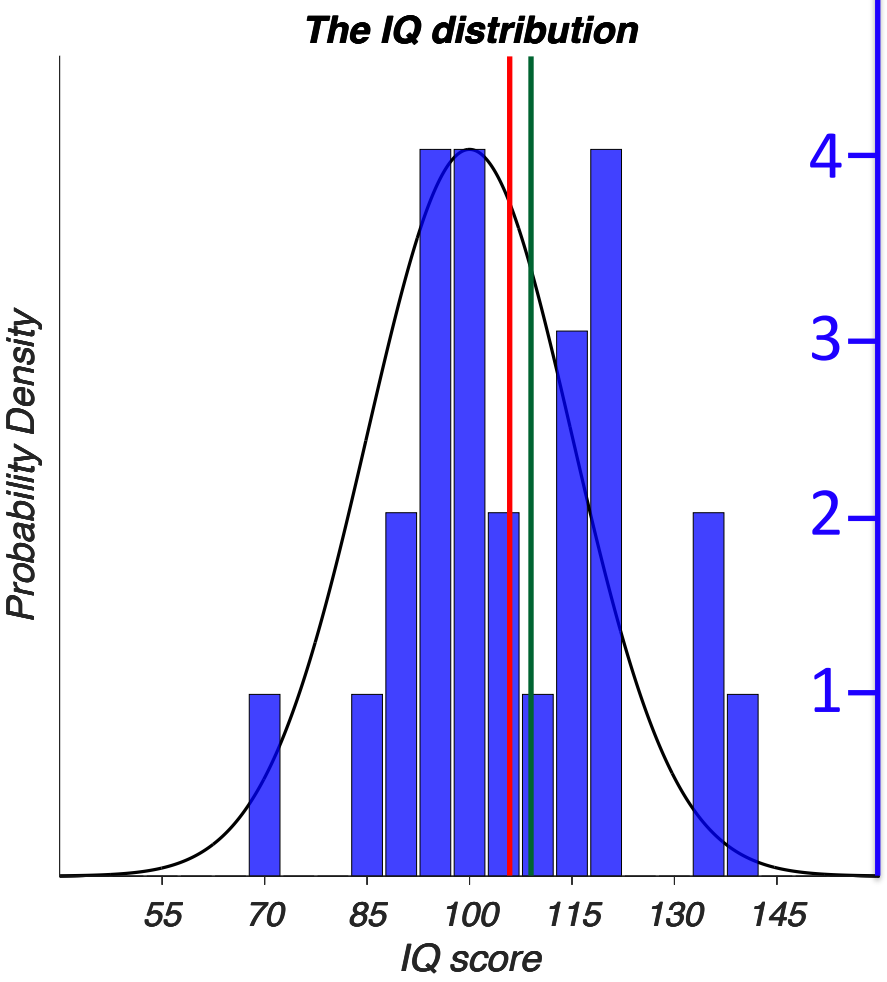
But not everything is countable – introducing the **z-test**

- Continuing with our toy example:
- Q: Does NZT improve IQ?
- H_0: NZT does not improve IQ
- In the general population, IQ is *known* to be distributed *normally* with
- $\mu = 100$
- $\sigma = 15$
- We give the drug to 25 people and test their IQ.



A possible outcome…

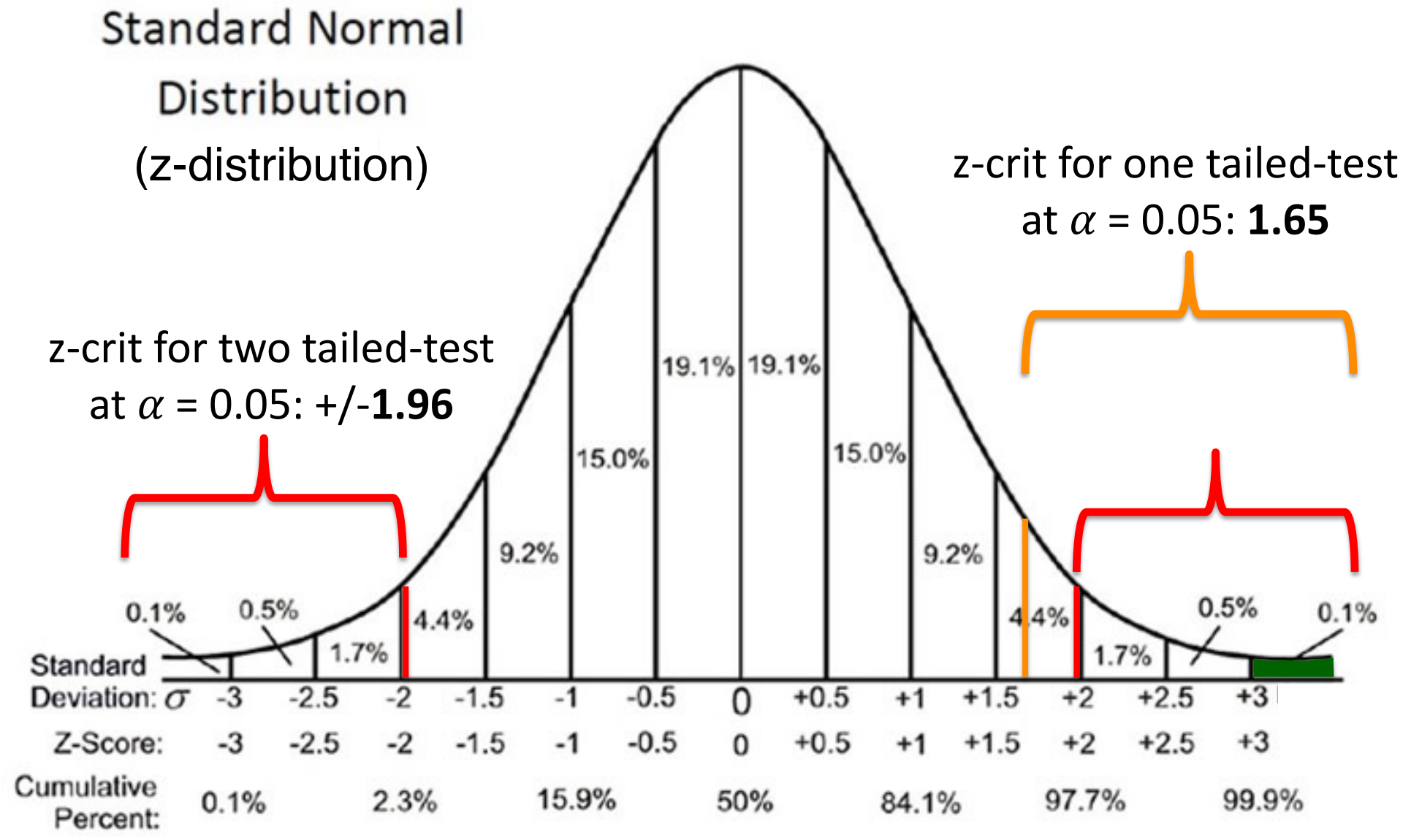
Was the sample drawn from this known distribution?



- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $n = 25$ (Sample contains scores from 25 participants)
- $\bar{x} = 109.0$ (Sample mean)
- $z = (\bar{x} - \mu)/SEM = (109\text{-}100)/SEM$ (Standardized score)
- $SEM = \sigma / \sqrt{n} = 15/\sqrt{25} = 15/5 = 3.0$
- $z = 9/3 = 3$
- Statistically significant?
- Is the sample mean larger or smaller than the critical value that corresponds to the significance level?
- In fact, here $p < 0.01$**

Where does the p value come from?

- We have to transform the z-value to a probability.
- The z-distribution is the standard normal distribution (μ = 0, σ = 1):



z-scores as a test statistic

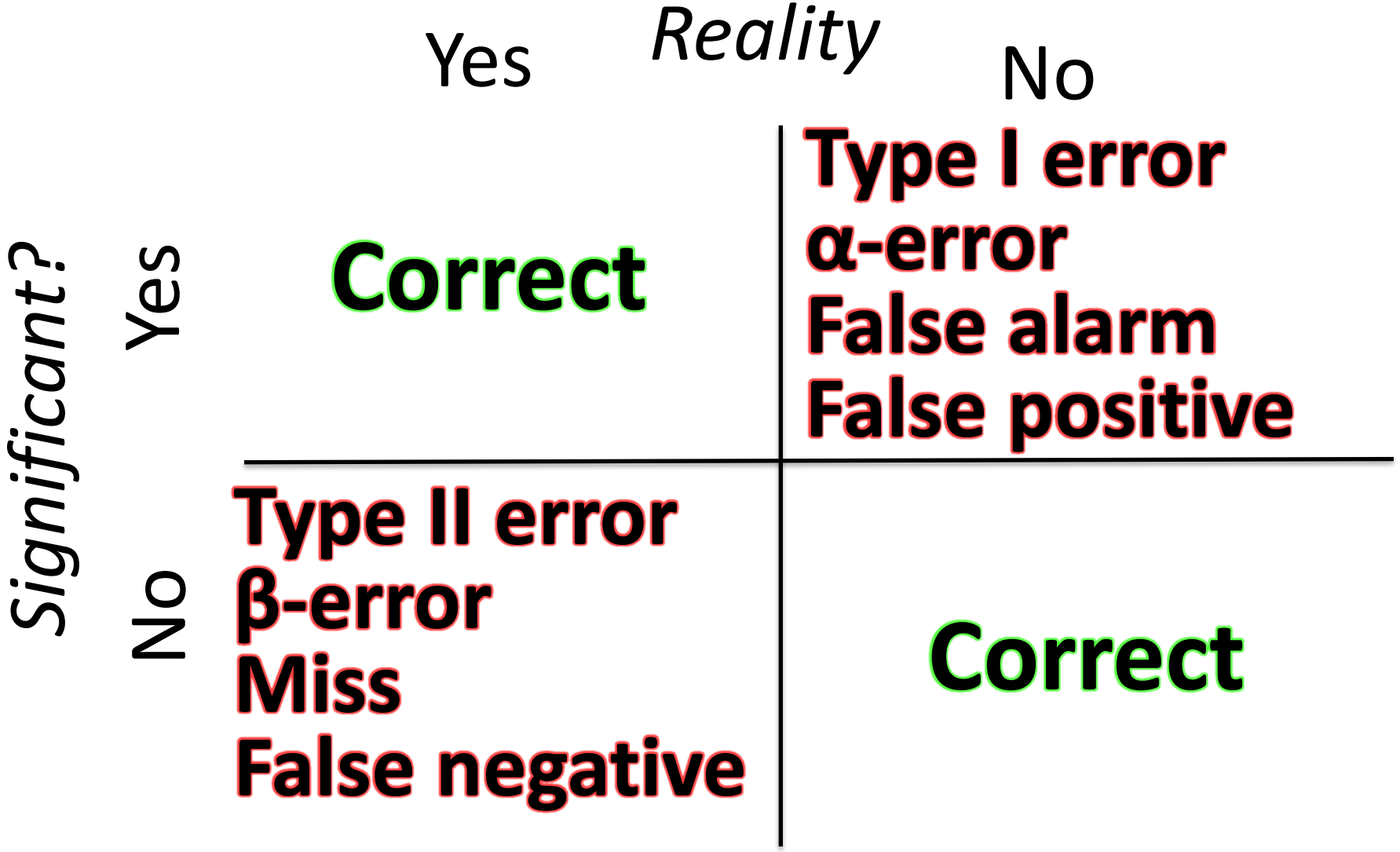
$$z = \frac{x - \mu}{\sigma}$$

- z-scores are a standard way to normalize data.
- Here, we use z-scores as a **test statistic**.
- How far is the **observed** sample mean from its **expected value** (the population mean if the null hypothesis is true), in units of standard error of the mean?

$$z = \frac{\bar{x} - \mu}{SEM}$$

- The standard normal distribution (z-distribution) is well known, so we can readily convert the z-value to a p-value and compare it to the significance level α.

Determining whether NZT works to improve IQ scores or not is a **decision**:



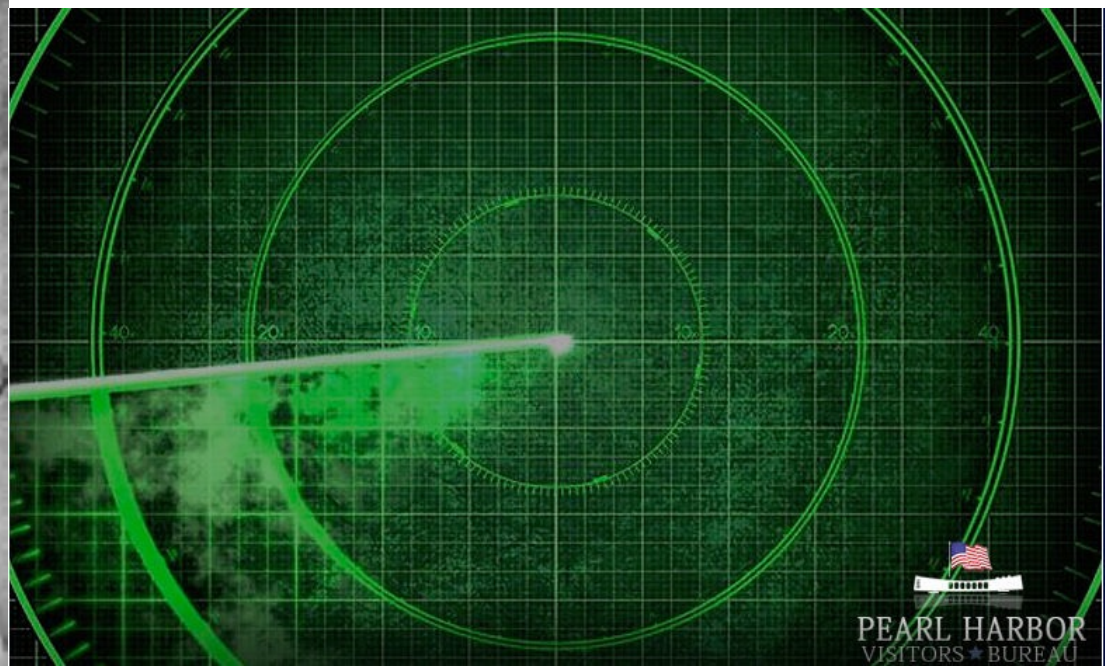
Decisions can be wrong

False positive

False negative

“Don’t worry about it”

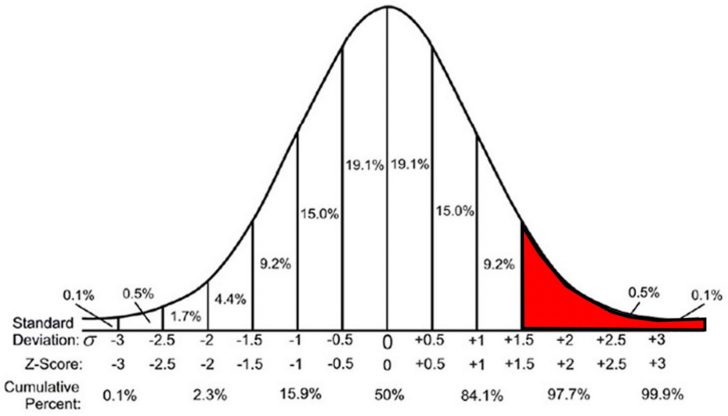
Lt. Kermit Tyler, commanding officer of the Hawaii aircraft tracking center, in response to an unusually large blip on the radar, at 7 am on 12/07/1941



Summary of this approach so far

- We always want to know how likely it was to observe our sample, assuming chance alone.
- To calculate this probability, we convert the sample mean to a **test statistic** with a known **null distribution** (distribution given H_0 is true).
- The p value is the area under this null distribution cut off by the test statistic.
- We then compare this value to the decision criterion α

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$



p $< \alpha$?

Why can't we just always use z?

- There are several issues:
- First, it relies on known population parameters:

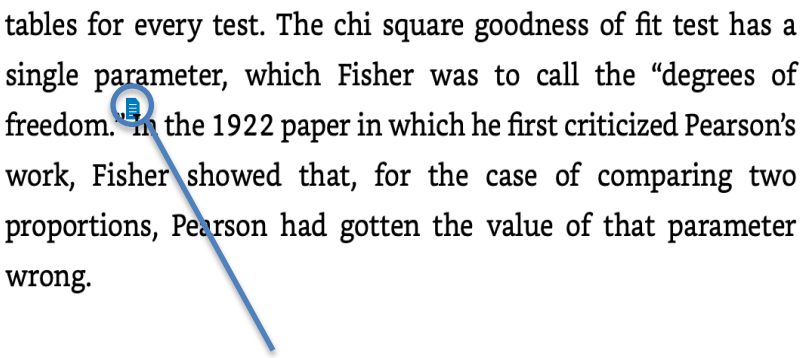
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

unknown

- Second, it relies on the central limit theorem.
- In other words, it doesn't work for small samples.
- Using t as a test statistic avoids both of these problems. The t-test relies on t as a test statistic.
- Thus, we often use - in practice – the **t-test**. It is one of the most commonly used parametric tests.

Before we can understand the t-test, we need to talk…

- About degrees of freedom.
- One of the most mysterious concepts in all of statistics.
- Statistics courses tend to just kind of gloss over it.

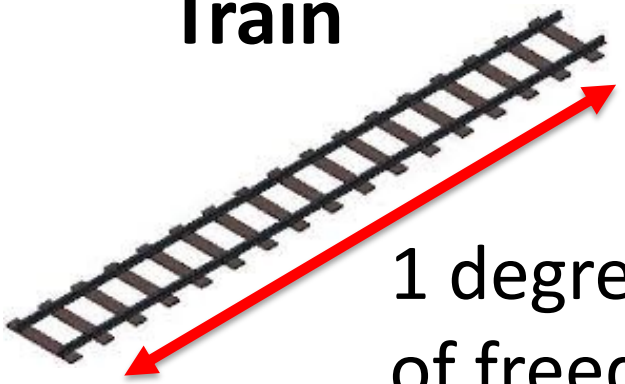

tables for every test. The chi square goodness of fit test has a single parameter, which Fisher was to call the "degrees of freedom." In the 1922 paper in which he first criticized Pearson's work, Fisher showed that, for the case of comparing two proportions, Pearson had gotten the value of that parameter wrong.

"Ok, so what are they?"
(From "The Lady tasting tea")

- It is actually quite straightforward, once understood.
- Confusing because 4 concepts are conflated here: What they are (number of independent pieces of information), why that matters, why we "lose" some and why it is called that (sounds obscure)?

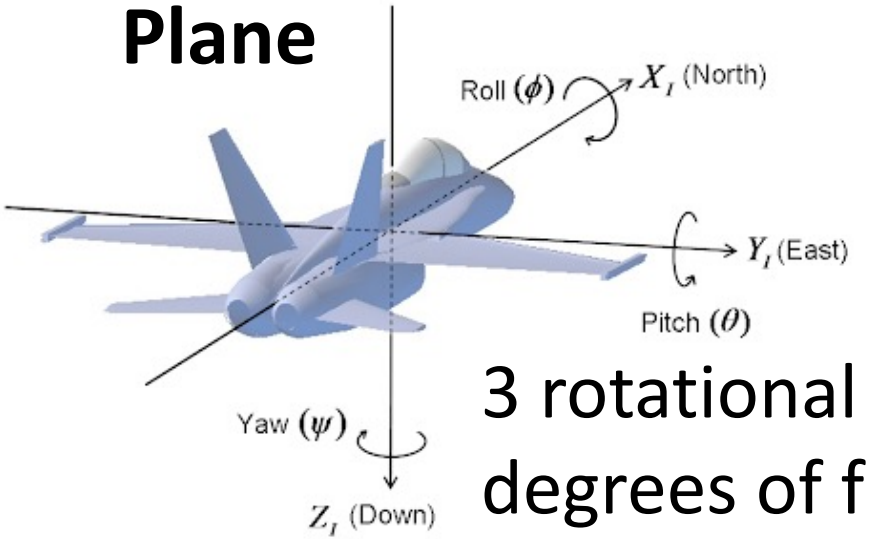
First, to clarify the analogy: Degrees of freedom in mechanics

Train



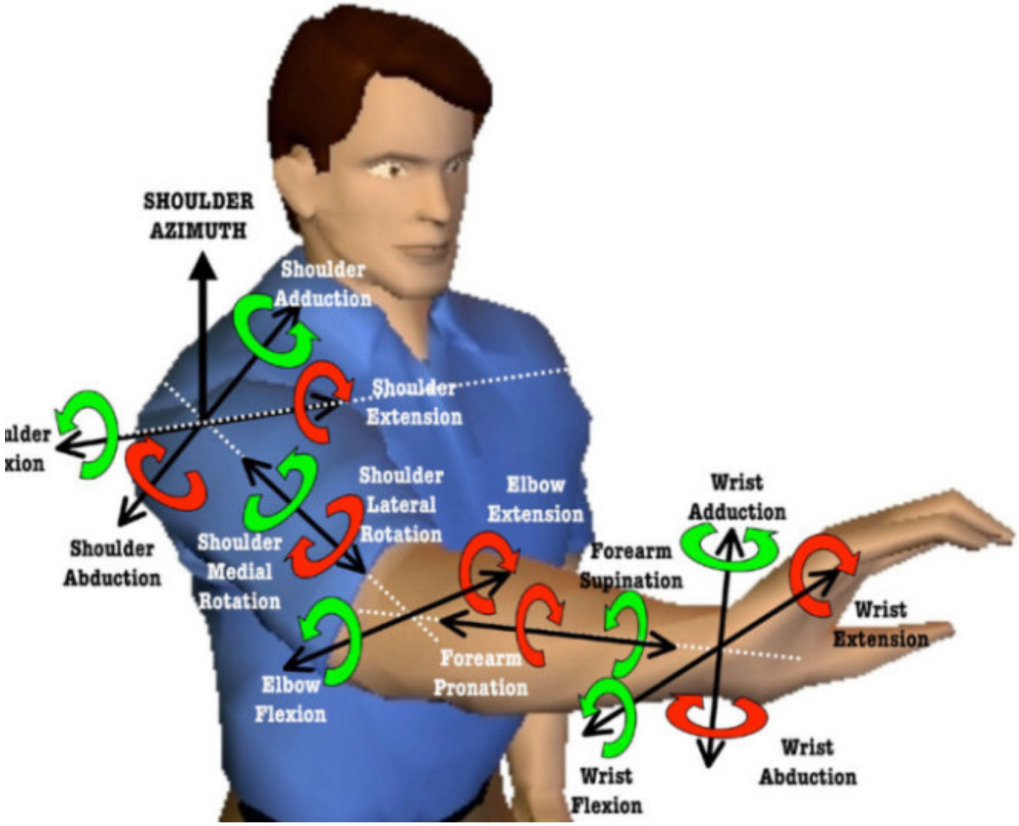
1 degree of freedom

Plane



3 rotational degrees of freedom

Human arm



7 degrees of freedom

Degrees of freedom (to vary) in statistics

- The number of independent pieces of information (numbers, measurements, datapoints) in a dataset that a parameter calculation is based on.
- The higher the number of degrees of freedom, the more evidence it is based on, the more stable the parameter estimate is.
- Example: On RMP, professor A has an average rating of 5.0 based on 2 ratings, whereas professor B has an average rating of 4.5 based on 200 ratings.
- Is A necessarily a better professor than B?

How are df lost?

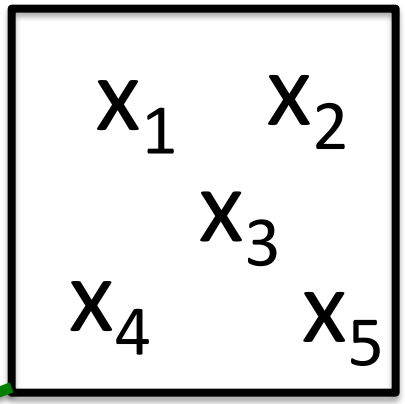
- To assess how stable the estimation of a population parameter from a sample is, we need to take into account how much evidence (*independent* pieces of information) is provided by the sample.
- This brings us to the 2nd property of degrees of freedom: Why some are "lost" if we estimate parameters from the sample.
- First, it is important to recognize that calculations do not create new, independent information (only measurements do)
- We might even lose some, if we estimate a parameter from the sample itself, then use this result to estimate other parameters from the same sample, as there are fewer independent pieces remaining.
- Once it is such constrained – by being used in a calculation, it is no longer independent because it can no longer vary freely.
- No double-dipping.

Representing the state of the warehouse in your Kontor with a database

of values unknown

	Apples	Bananas

If we calculate the sample SD as an estimate of population SD, we have to adjust the equation for population SD because we already used the *same* data to calculate the sample mean:



$$\bar{x} = \frac{\sum_{1}^{n} x_i}{n}$$

Population SD: $$\sigma = \sqrt{\frac{\sum_{1}^{n} (x_i - \mu)^2}{n}}$$?

Sample SD: $$s = \sqrt{\frac{\sum_{1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Is df = n-1?

- No.
- It is **n – k**.
- n = number of independent pieces of information in the sample.
- k = how many parameters are estimated from the sample itself.
- For instance, we might calculate k = 9 sample means from a sample size n = 90 in an ANOVA.
- After doing that, we will have 81 = 90 – 9 df left.
- Now back to t and how it is tested

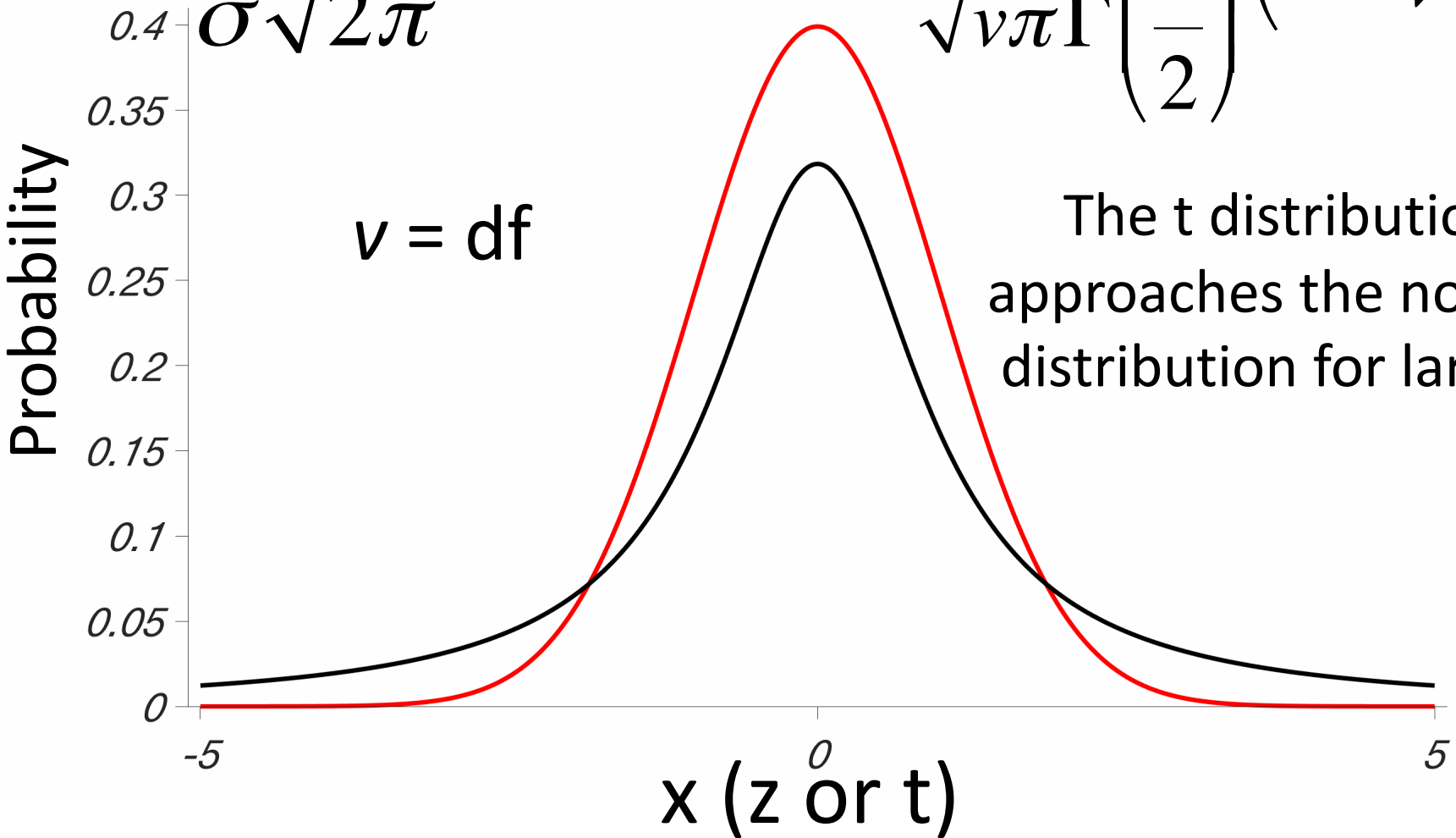
The t-distribution

Normal:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

t:

$$y = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}$$

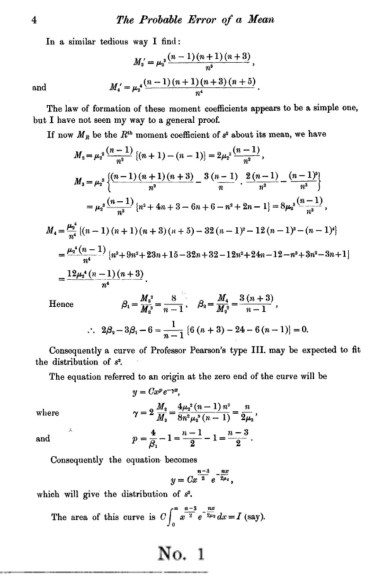


v = df

The t distribution approaches the normal distribution for large v

So what - how does this help? Enter Student's t-test

- "Student" = William Gosset
- Employed by Guinness brewery.
- Worked out a test for small sample sizes and unknown population parameters.
- Intended for quality control, a trade secret.
- Not allowed to publish under his actual name.



4 *The Probable Error of a Mean*

In a similar tedious way I find:

$$M_3' = \mu_s^3 \frac{(n-1)(n+1)(n+3)}{n^3},$$

and

$$M_4' = \mu_s^4 \frac{(n-1)(n+1)(n+3)(n+5)}{n^4}.$$

The law of formation of these moment coefficients appears to be a simple one, but I have not seen my way to a general proof.

If now M_R be the R^{th} moment coefficient of s^2 about its mean, we have

$$M_1 = \mu_s^2 \frac{(n-1)}{n^2}\{(n+1)-(n-1)\} = 2\mu_s^2 \frac{(n-1)}{n^2},$$

$$M_2 = \mu_s^2 \left\{\frac{(n-1)(n+1)(n+3)}{n^3} - 3\frac{(n-1)}{n} \cdot \frac{2(n-1)}{n^2} - \frac{(n-1)^2}{n^2}\right\}$$

$$= \mu_s^2 \frac{(n-1)}{n^3}\{n^2+4n+3-6n+6-n^2+2n-1\} = 8\mu_s^2 \frac{(n-1)}{n^3},$$

$$M_4 = \frac{\mu_s^4}{n^4}\{(n-1)(n+1)(n+3)(n+5) - 32(n-1)^2 - 12(n-1)^2 - (n-1)^2\}$$

$$= \frac{\mu_s^4 (n-1)}{n^4}\{n^3+9n^2+23n+15-32n+32-12n^2+24n-12-n^2+3n^2-3n+1\}$$

$$= \frac{12\mu_s^4 (n-1)(n+3)}{n^3},$$

Hence $$\beta_1 = \frac{M_3^2}{M_2^3} = \frac{8}{n-1}, \quad \beta_2 = \frac{M_4}{M_2^2} = \frac{3(n+3)}{n-1},$$

$$\therefore\ 2\beta_2 - 3\beta_1 - 6 = \frac{1}{n-1}\{6(n+3) - 24 - 6(n-1)\} = 0.$$

Consequently a curve of Professor Pearson's type III. may be expected to fit the distribution of s^2.

The equation referred to an origin at the zero end of the curve will be

$$y = Cs^p e^{-\gamma s},$$

where $$\gamma = 2\frac{M_2}{M_3} = \frac{4\mu_s^2 (n-1)n^3}{8n^3\mu_s^2 (n-1)} = \frac{n}{2\mu_s},$$

and $$p = \frac{4}{\beta_1} - 1 = \frac{n-1}{2} - 1 = \frac{n-3}{2}.$$

Consequently the equation becomes

$$y = Cs^{\frac{n-3}{2}} e^{-\frac{ns}{2\mu_s}},$$

which will give the distribution of s^2.

The area of this curve is $C\int_0^\infty s^{\frac{n-3}{2}} e^{-\frac{ns}{2\mu_s}} ds = I$ (say).

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SEM}$$

The logic of the t-test

- In the z-test, we reduced the sample to its sample mean, a sample statistic.
- We then asked how far this sample mean is from what we would expect given the population it came from (the population mean).
- If it is too far, we consider it implausible that it came from that population – the treatment must effectively have created a new population.
- For most things we care about, we do not know the population mean.
- So where to get our expectations from?
- If we draw two samples, and the two sample means are too far from each other, it is unlikely that they came from the same underlying population:
- They probably came from 2 populations with different means

General assumptions of parametric tests

- They assume that sample parameters like the mean can be interpreted meaningfully.
- Data is distributed normally.
- To meaningfully interpret a difference in the samples by comparing their means, we have to assume "**Homogeneity of variance**" - that variability within each sample is similar – only the means differ.
- These assumptions are often violated, but the t-test is often considered as "robust" to violations of these assumptions.

Violating homogeneity of variance

High mean wealth, high variation in wealth

Low mean wealth, low variation in wealth

There are several versions of the t-test

- Which one to use depends on the number of **independent** groups involved:
- 2 independent groups: t-test for independent groups (or samples) – ”**between subject**” designs.
- Usually comparing a treatment group and a control group or two treatment groups, often A/B tests.
- What makes them independent?
- Participants are randomly assigned to each group, every participant is only in one group.
- degrees of freedom: $df = n_1 + n_2 - 2$
- 1 independent group: t-test for dependent groups (or paired samples or correlated groups).

t-test for Correlated Groups (paired samples / dependent samples)

- In some experiments, the *same* people are tested twice. This is called a "**within-subjects**-design".
- Of course, the observations are no longer independent in this case (the same people are in both groups).
- So we need to use a test for correlated groups.
- In this case, we work with difference scores:
- This changes the df: df = n - 1
- Heuristic: If there is one measure of a given variable for each person in the group, use the t-test for independent groups. If you have two (like in repeated measures designs), use the one for correlated groups.

$$t = \frac{\bar{D}}{SEM_{\bar{D}}}$$

A practical use case: Toy example

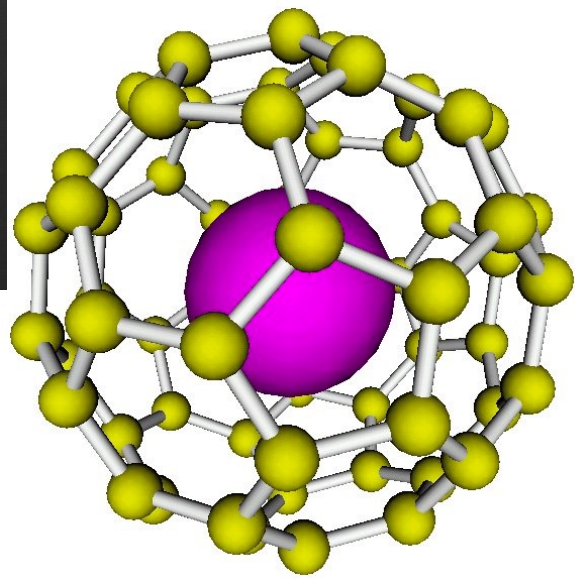
You work for the TriOptimum Corporation as a data scientist.

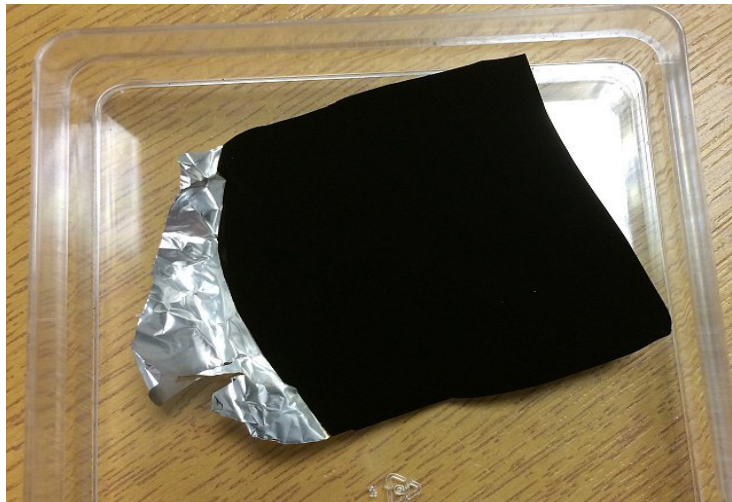
The project is to make a drug that decreases reaction time.

You suspect it to be very effective, on the order of boosting mental speed by 3 standard deviations.

But it is very expensive to make, as it involves *endohedral fullerenes* wrapped in *Vantablack* covered with forests of *carbon nanotubes*, so you could only convince your CEO to make a test batch of 3 of them.

What kind of test should you do in order to determine whether it works?





Let's try a t-test for independent groups

- The data?
- n?
- df?

Group 1: Without	Group 2: With
1700 ms	1100 ms
2000 ms	1500 ms
2300 ms	1900 ms
G1 mean: 2000 ms	**G2 mean: 1500 ms**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SEM}$$

$\sigma_{pooled} = 353.5$

$SEM_{pooled} = 288.7$

$$t = \frac{2000 - 1500}{288.7}$$

$$t = \frac{500}{288.7}$$

t = 1.73

$t_{crit}(4) = 2.78$

p = 0.16 n.s.

← Inter-individual variation

Let's try a t-test for paired groups

- The data?
- n?
- df?

Without drug	With drug
1700 ms	1100 ms
2000 ms	1500 ms
2300 ms	1900 ms
"without" mean: 2000 ms	**"with" mean: 1500 ms**

$$t = \frac{\bar{D} - 0}{SEM_{\bar{D}}}$$

$$s_D = 100$$

$$SEM_{\bar{D}} = \frac{s_D}{\sqrt{n}} = 57.7$$

$$t = \frac{500}{57.7}$$

t = 8.66

$t_{crit}(2) = 4.30$

$p < 0.05$*

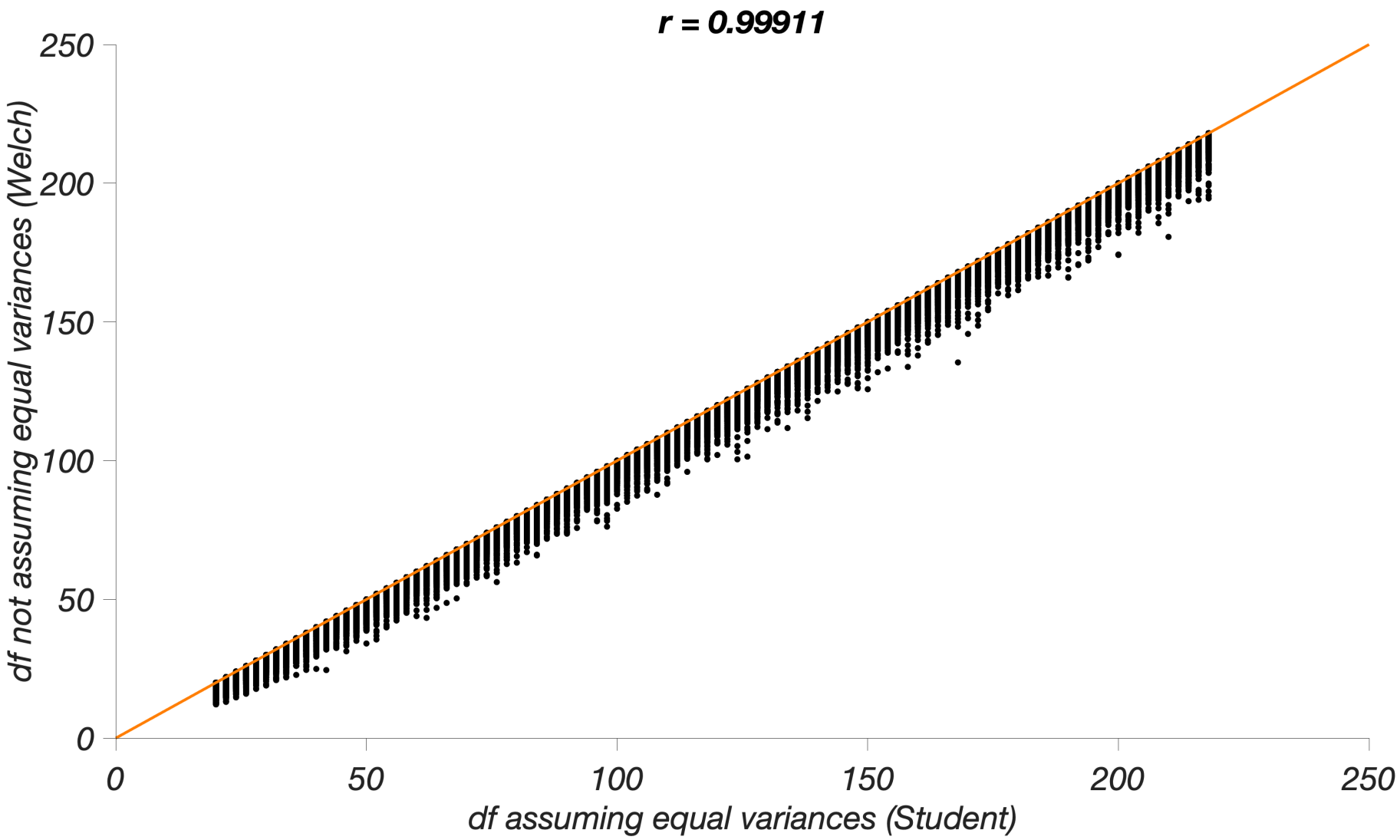
Welch's t-test

- In most real use cases, variances won't be homogeneous, but Student's t-test assumes that they are.
- Welch created a modified version of the independent samples t-test that does not assume homogeneity of variance (but keeps the other assumptions).
- Biggest difference: Degrees of freedom are different. There are fewer effective df, and they are estimated using the Satterthwaite approximation:

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

- Insight: Effective df rely on the standard deviations here.
- This procedure frequently leads to fractional df.

df vs. effective df in an independent samples t-test



What if you want to compare more than two groups?

- Then you should do an ANOVA.
- Click here to see as to why and how it works
- Briefly: ANOVA extends the logic of the t-test to more than 2 groups.
- It is mathematically equivalent to multiple regression.
- So far, our significance tests relied on reducing our data to sample means.
- What if that is sometimes not a reasonable thing to do?

Why some data doesn't lend itself to be reduced to (sample) means

- 1) **Categorical** data: When numbers are simply used as labels.
- Remember: The California police code assigns the number "213" to the use of illegal explosives and the number "217" to an assault with the intent to murder.
- Caution: Python *will* allow you to take the mean of categorical data. But just because you can doesn't mean you should:
- If we have 2 crimes, one 213 and one 217, it doesn't mean that - on average - we have two "215"s – the code for carjacking.
- In other words, taking the mean is mean-ingless here.
- Only thing we can do: **Count** how often a category occurs in a sample/dataset ("cat counting").

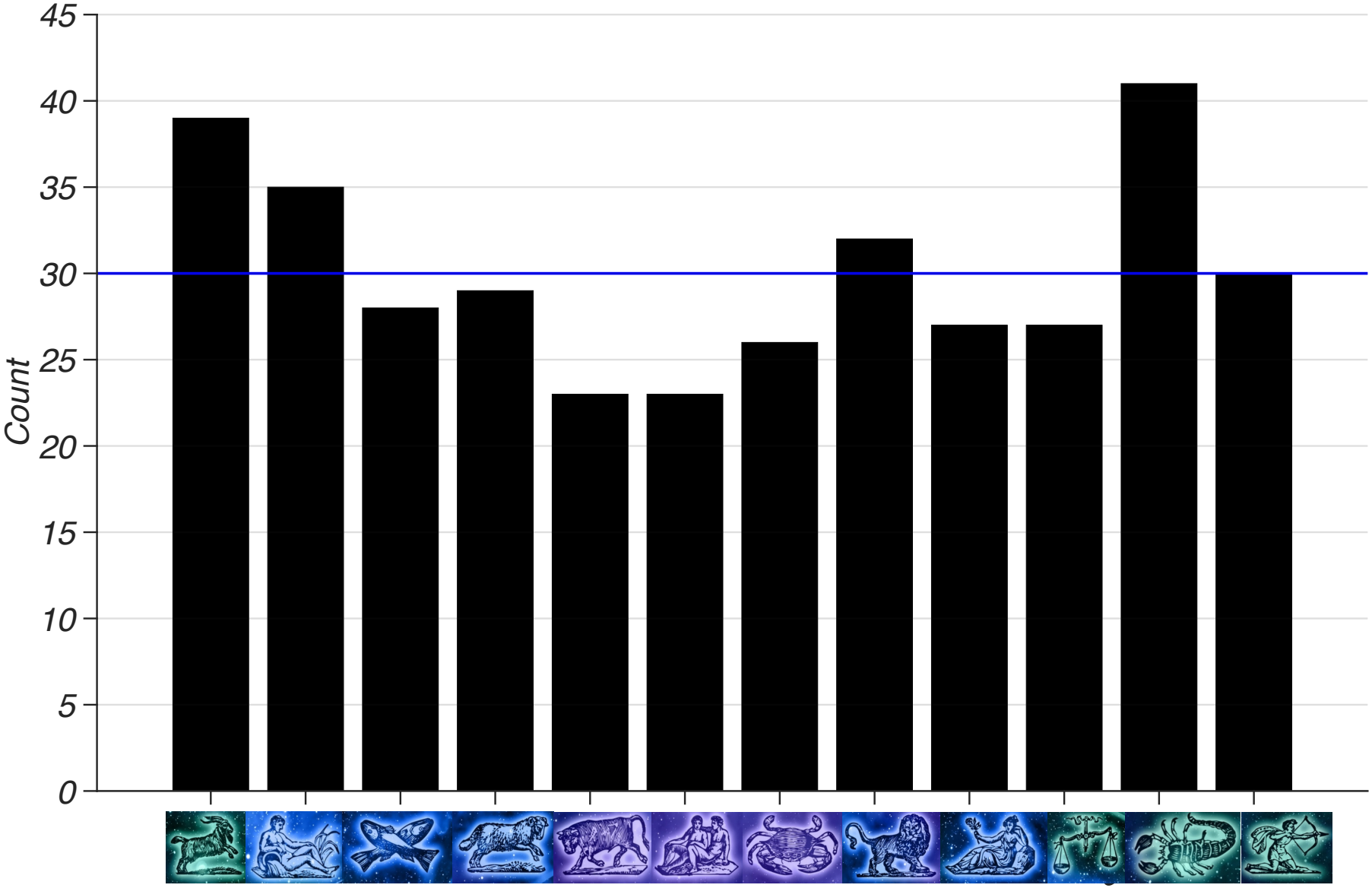
What can we do with category counts (frequencies) alone?

- A χ^2 test (introduced by Karl Pearson), for starters.
- A χ^2 test is a significance test for categorical data.
- Instead of reducing a dataset to sample statistics (like the mean), it relies on category counts alone.
- Counting is the basic operation to have data at all, so a $\chi 2$ test can always been done ("radical freedom")
- As such, it is a paradigmatic case of a non-parametric test.
- Instead of getting expectations from population parameters, non-parametric tests use distributions.
- In general, non-parametric tests make fewer assumptions, so they are more versatile.
- Drawback: They are usually less powerful than their parametric counterparts.

Is Zodiac sign predictive of being a serial killer?

- Popular culture has it that serial killers are likely Scorpios.
- If this was the case in reality, it would lend empirical support to the theoretical framework of astrology.
- Whereas time is continuous, zodiac signs are understood to be *categorically* different. For instance, someone born on April 19th is an Aries, supposedly impulsive and stubborn because they are influenced primarily by Mars, a "masculine" sign whereas someone born on April 21st is a Taurus, supposedly determined a stable, because they are influenced primarily by Venus, a "feminine" sign.
- To test the hypothesis that serial killers are Scorpios, students in a statistics class entered the birthday information of 360 (= the marginal) known serial killers.

The empirical distribution in this sample
Now what?



How the χ^2 test works:

- We compare expected category counts and observed category counts, square the differences and sum them up.
- This is the test statistic, χ^2:

$$\chi^2 = \sum \frac{(observedCount - expectedCount)^2}{expectedCount}$$

Applied to our Zodiac killer example:

Sign	Observed Count	Expected Count (n/12)	Deviation	Plugged in to equation	Contribution to χ^2
Capricorn	39	30	9	$9^2/30$	2.7
Aquarius	35	30	5	$5^2/30$	0.83
Pisces	28	30	-2	$-2^2/30$	0.13
Aries	29	30	-1	$-1^2/30$	0.03
Taurus	23	30	-7	$-7^2/30$	1.63
Gemini	23	30	-7	$-7^2/30$	1.63
Cancer	26	30	-4	$-4^2/30$	0.53
Leo	32	30	2	$2^2/30$	0.13
Virgo	27	30	-3	$-3^2/30$	0.3
Libra	27	30	-3	$-3^2/30$	0.3
Scorpio	41	30	11	$11^2/30$	4.03
Sagittarius	30	30	0	$0^2/30$	0

$\chi^2 = 2.7+0.83+0.13+0.03+1.63+1.63+0.53+0.13+0.3+0.3+4.03+0$

$\chi^2 = 12.24$

How does χ^2 distribute?

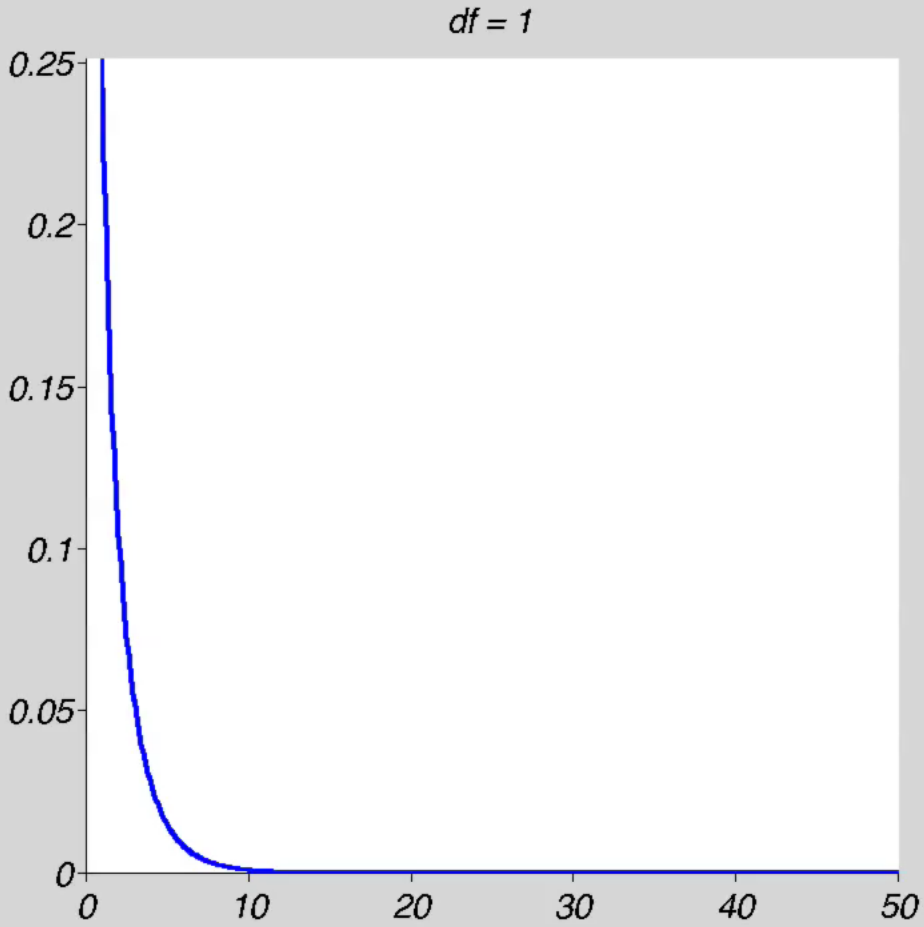
$$f(x,k) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$$

If we are not taking sample means, what are the degrees of freedom in the Chi square distribution?

The number of *categories* minus the marginal

So df = c – 1 (c = number of categories) = k

What does this look like?
The chi-square distribution



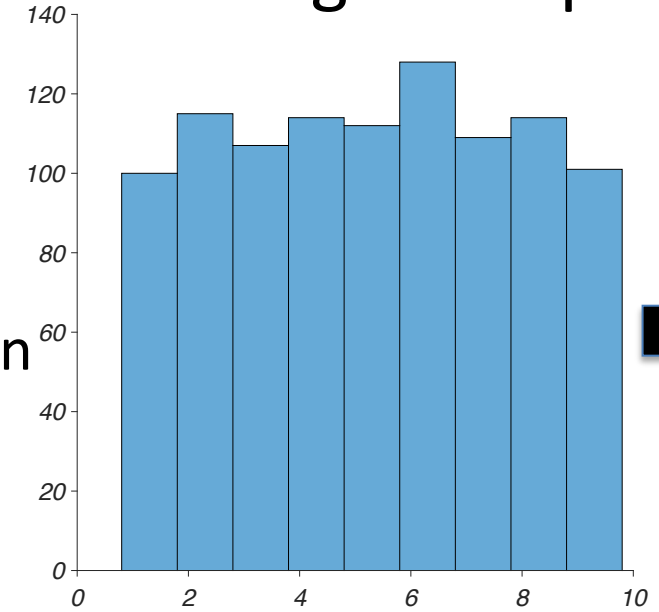
$\chi^2 = 12.24$, $df = 11$, $p = 0.34$, n.s.

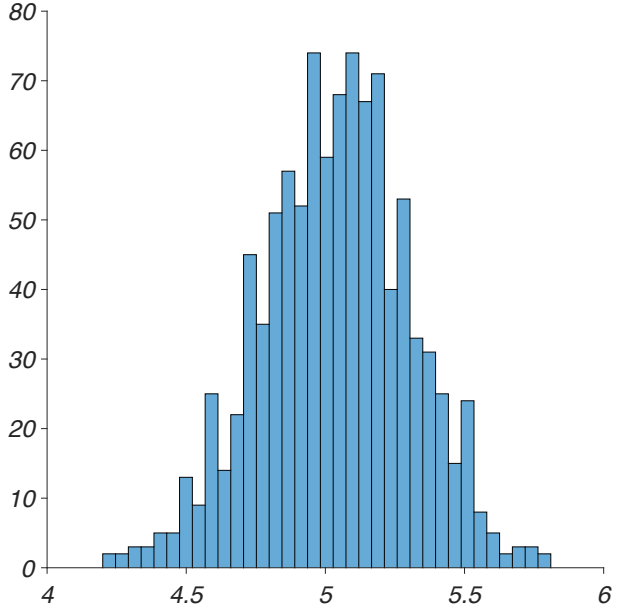
Beyond categorical tests: There are other situations where it is inappropriate to reduce a dataset to its mean

- Key example: Any situation that involves user ratings (movie, music, satisfaction, etc.)
- Why?
- Because the sample mean is a normalized sum.
- A sum presumes that the units of the items being summed are equal in order to be meaningful.
- Example: Two movies are rated as 3 and 4 stars by one user, and as 2 and 5 stars by another user.
- Is it fair to say that the mean rating is 3.5 for both?
- Only if 2 = 1 + 1, 3 = 1 + 1 + 1, and so on.
- But is this plausible for movies – that the *psychological* distance between a rating of 2 and 3, 3 and 4 and 4 and 5 is identical?

If the population contains extreme values, the mean is also not a good representation of the data:

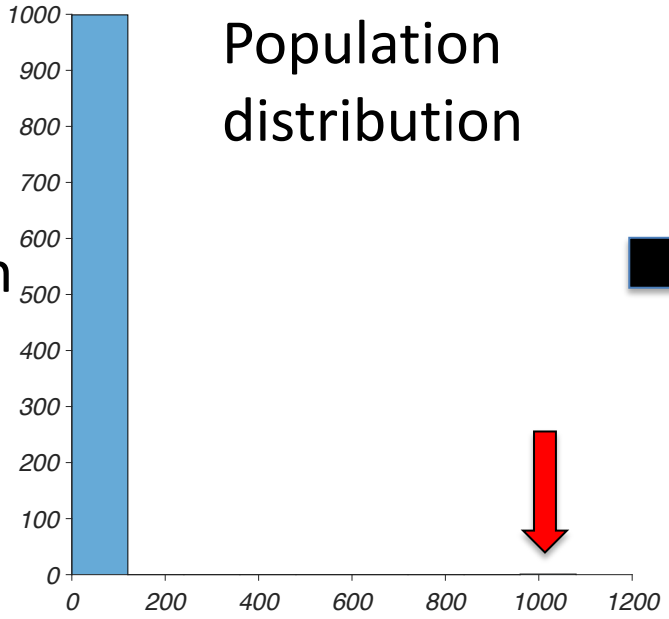
Uniform distribution



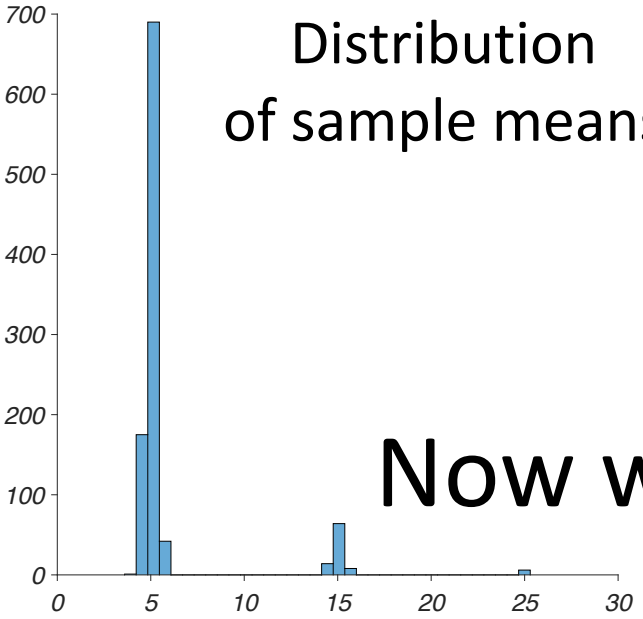


Uniform distribution plus extreme outlier

Population distribution



Distribution of sample means

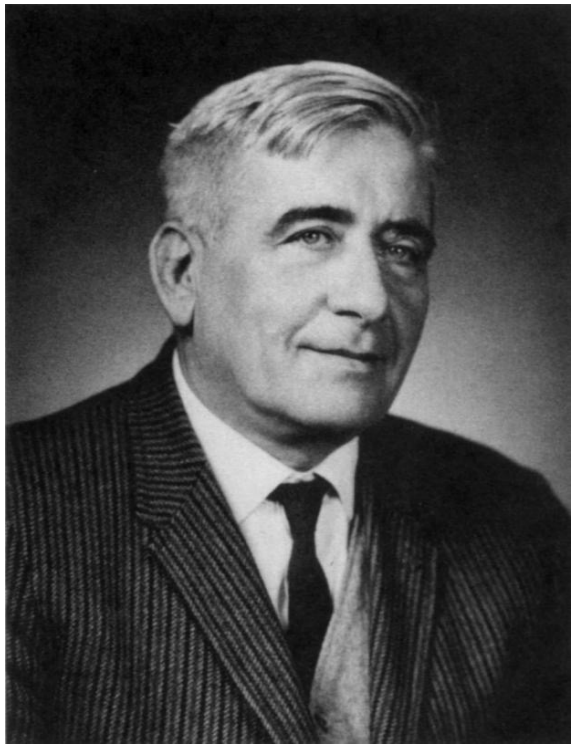


Now what?

Enter the Mann-Whitney U test

- Aka Wilcoxon rank-sum test

Frank Wilcoxon



Henry Mann

Ransom Whitney

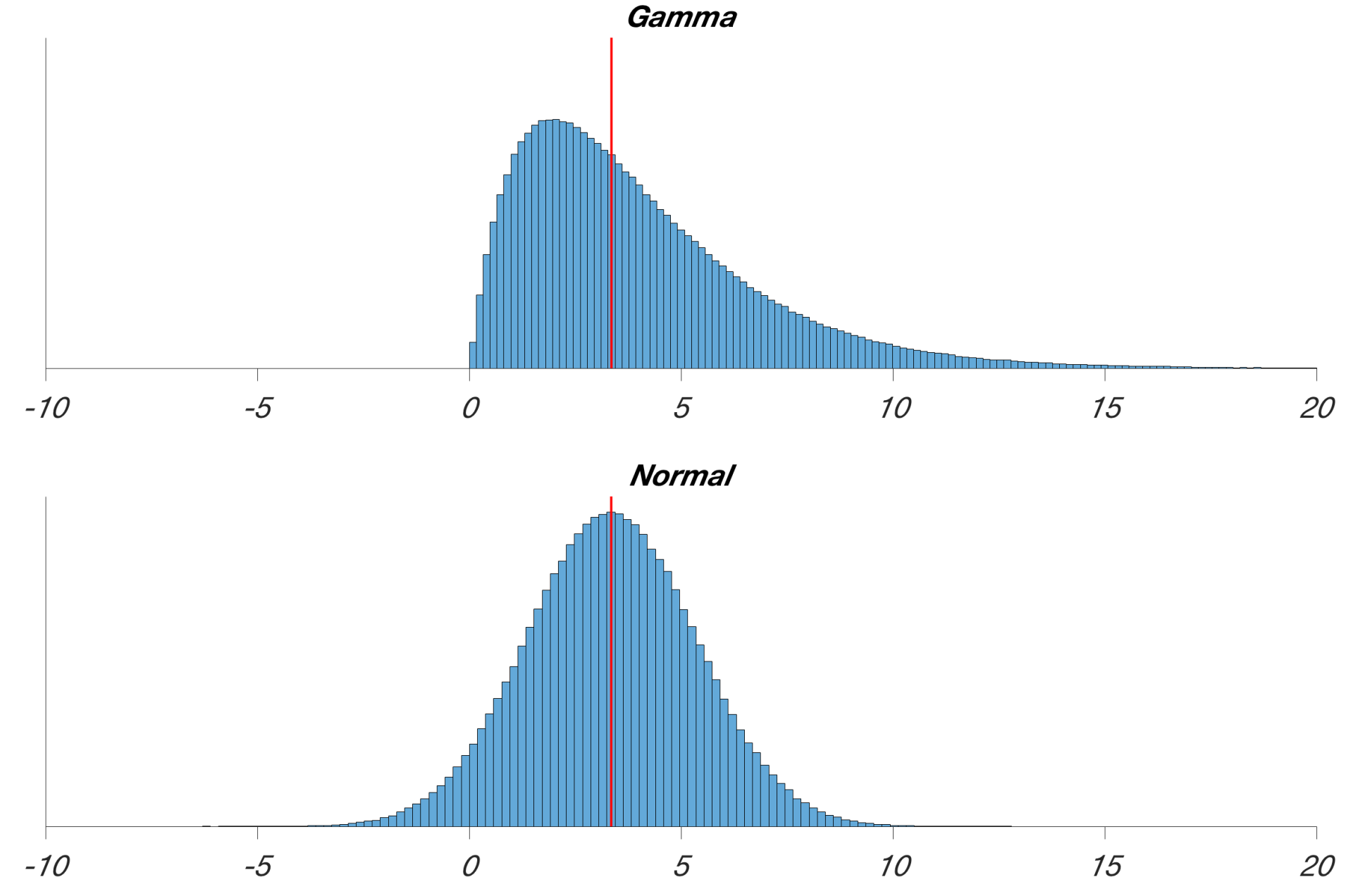
The Mann-Whitney-Wilcoxon ranksum test

- Analogy to the t-test (which tests whether two samples have the same mean).
- Tests whether two samples come from populations with the same median.
- As the median is robust to outliers – the data can come pretty much from any distribution of data and equal unit size is also not important, as the median depends only on ordering operations, not sums.
- Test statistic: U

The basic idea

- Say we have ratings data from two samples.
- We arrange them all in rank order from the smallest to the largest value, regardless of which sample they came from.
- If both samples come from the same underlying distribution/population, there should be random mixing and the sum of the ranks should be similar for both samples (assuming there is an equal n).
- Otherwise, there should be a difference in the sum of these ranks (and if n of the two samples is unequal).
- In the interest of time, we're skipping the calculation – you will never do this by hand.

But even if two samples have the same median, they can have a different distribution…



Enter the Kolmogorov-Smirnov test

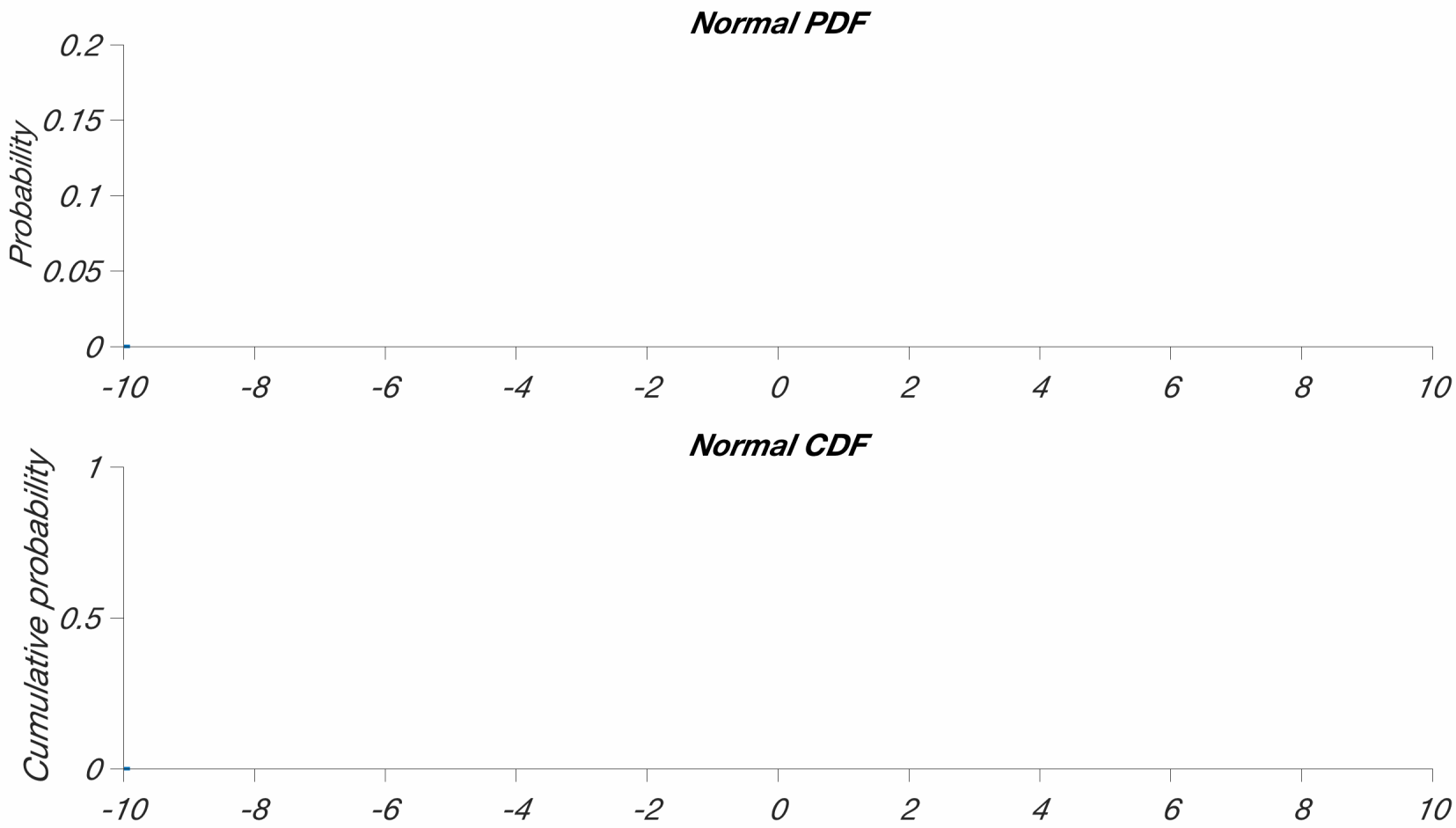
Andrey Kolmogorov

Nikolai Smirnov

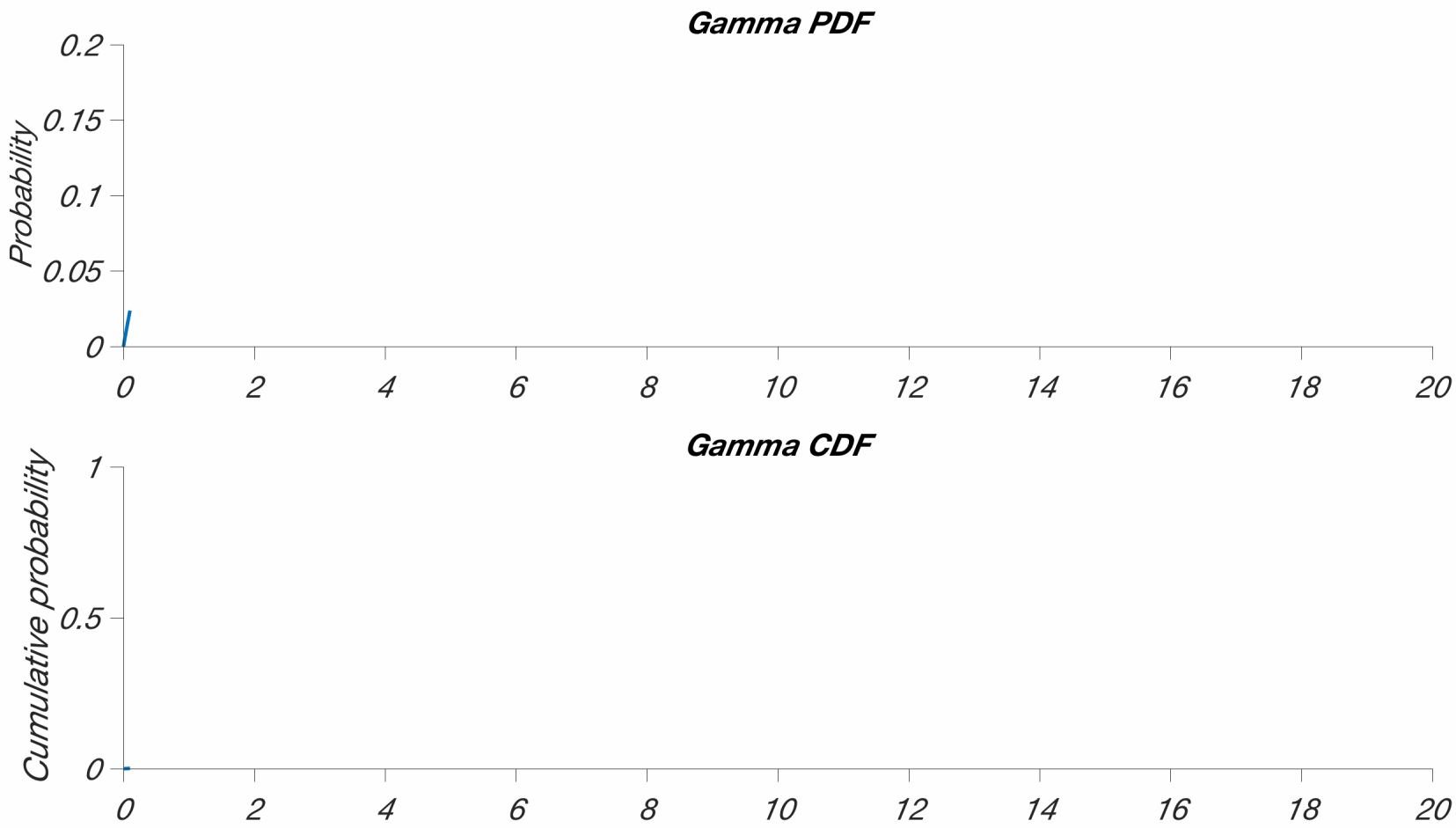
Idea

- No comparisons of means *or* medians.
- Tests whether the underlying distributions are the same (whatever they might be)
- Simply comparing the cumulative distribution function of samples.
- In that sense, the Kolmogorov-Smirnov (or KS) test is a **goodness-of-fit** test.

Probability density functions (PDF) and cumulative probability density functions (CDF): **Normal**

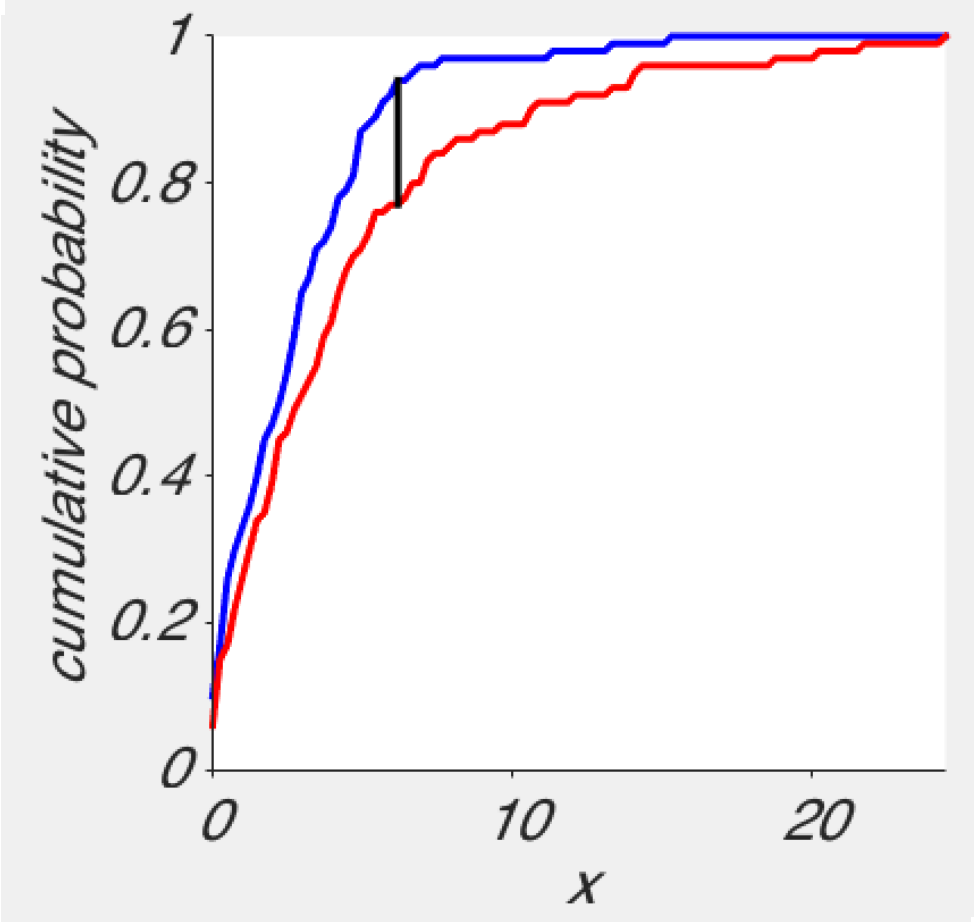


Probability density functions (PDF) and cumulative probability density functions (CDF): **Gamma**

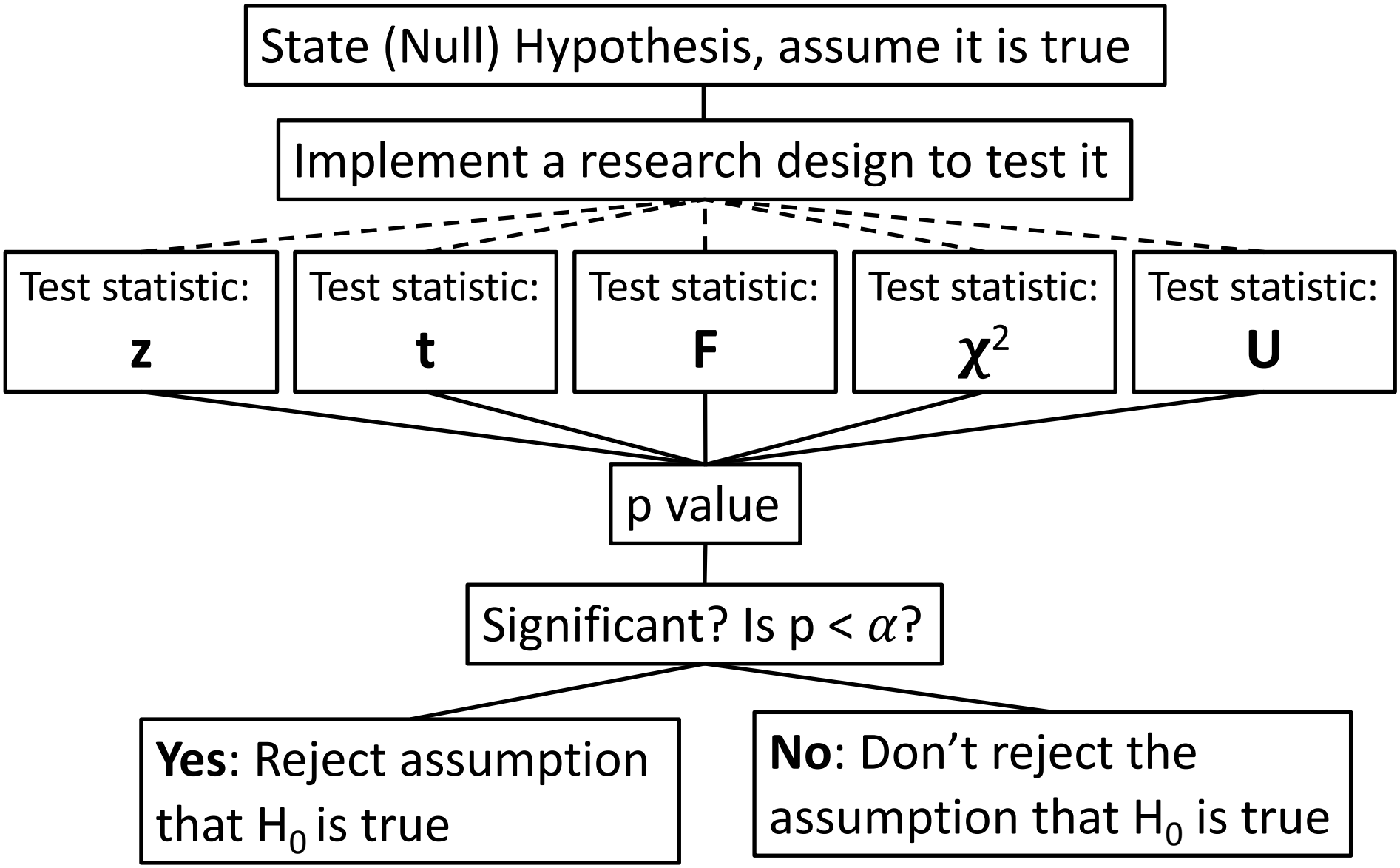


Comparing cumulative probability distributions

- Plot the cumulative distribution functions for the two empirical samples.
- Find the point of largest separation ("D").
- This is the test statistic.
- The D distribution under the null is known, → convert that to a p-value.
- Note: This uses eCDFs - empirical cumulative density functions.



Null hypothesis testing step by step



How about vedic astrology?

VEDICFEED

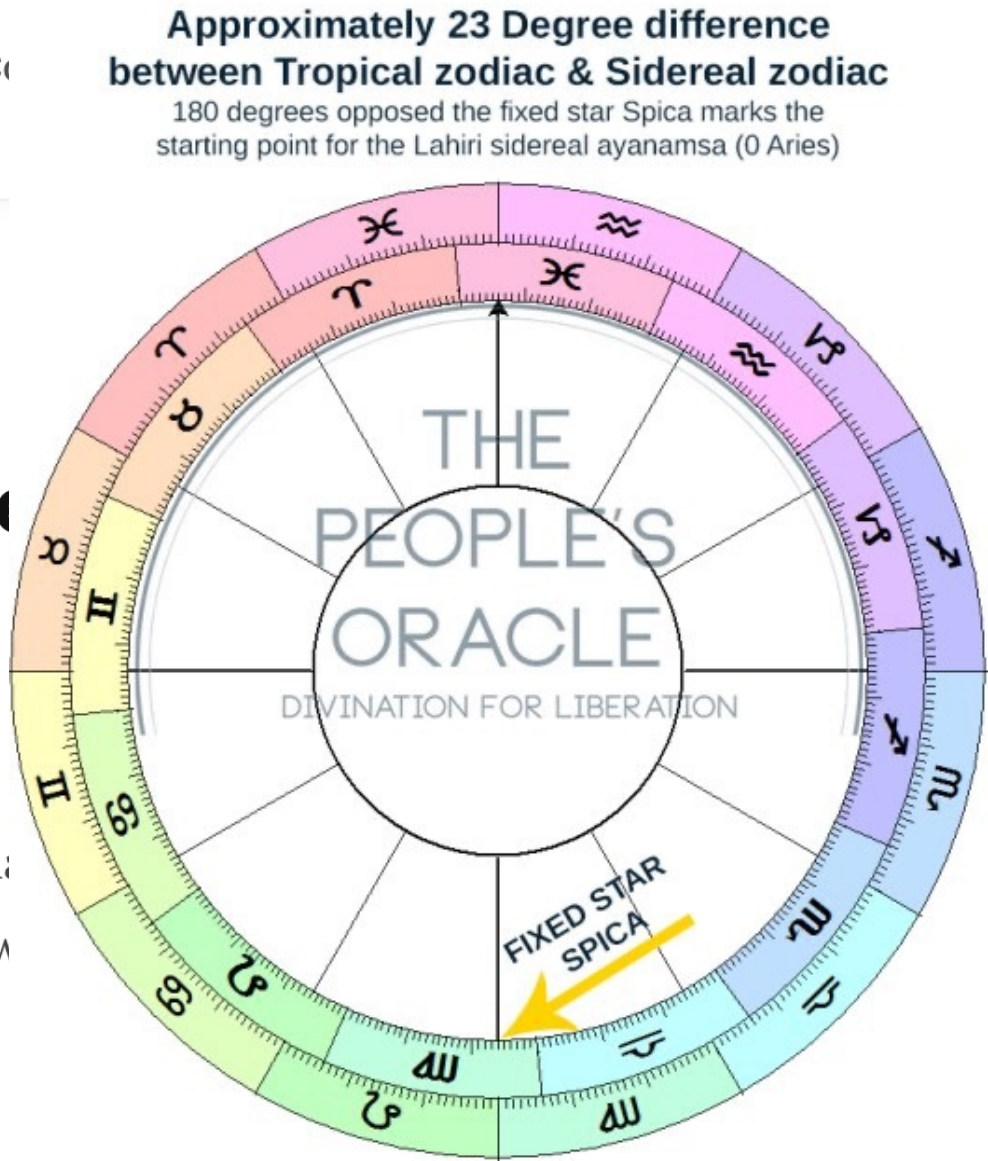
Astrology C

ASTROLOGY

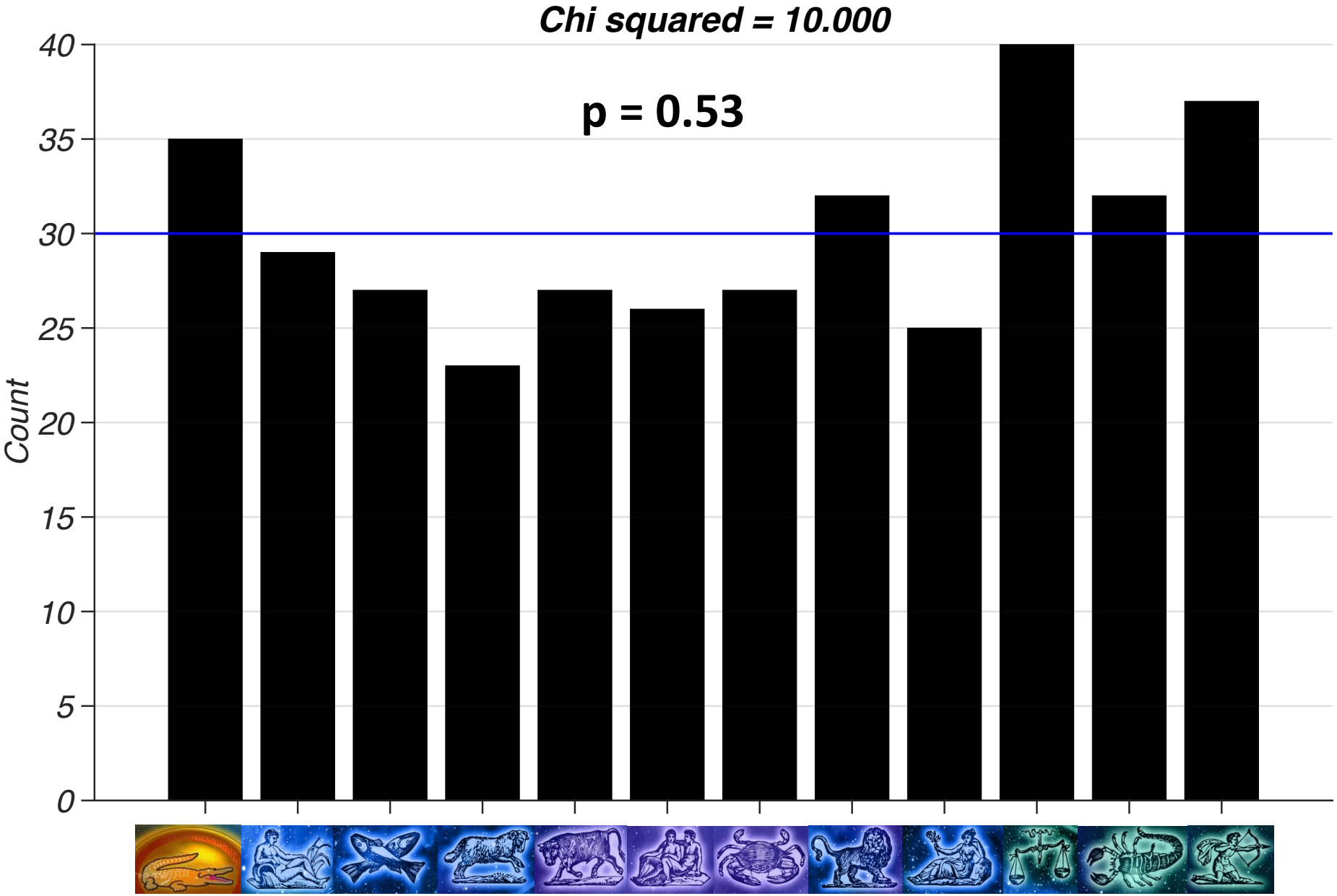
5 Basic Differences Betwe and Western Astrology

BY MONA

Astrology is the savior of suffering huma can eradicate sorrows and fill our lives w

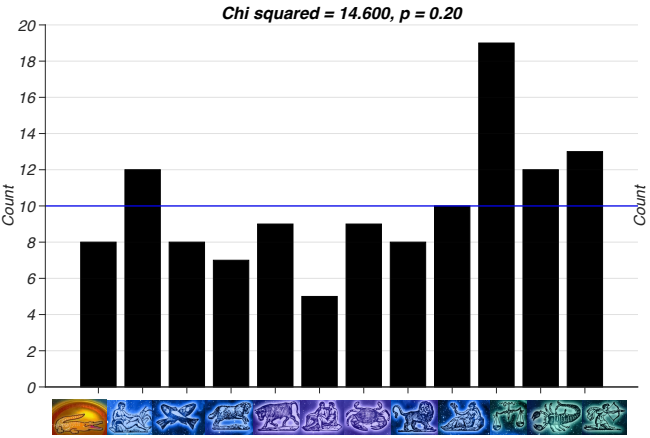


The same data through the vedic lens

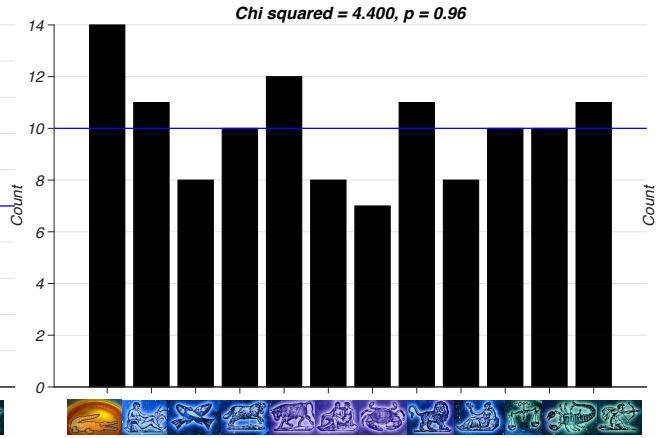


How about subdividing the sample into 3 batches of 120?

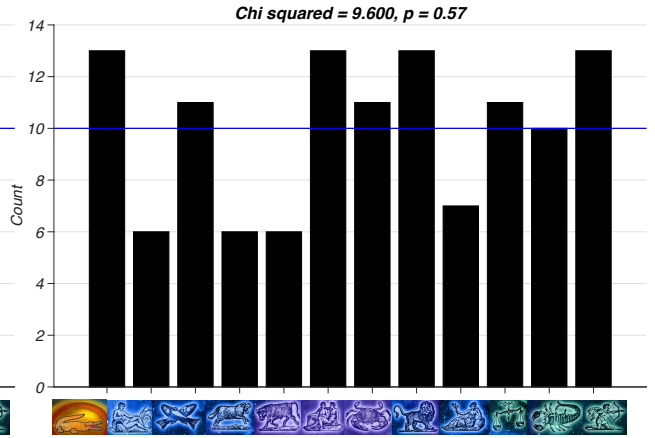
n: 1-120



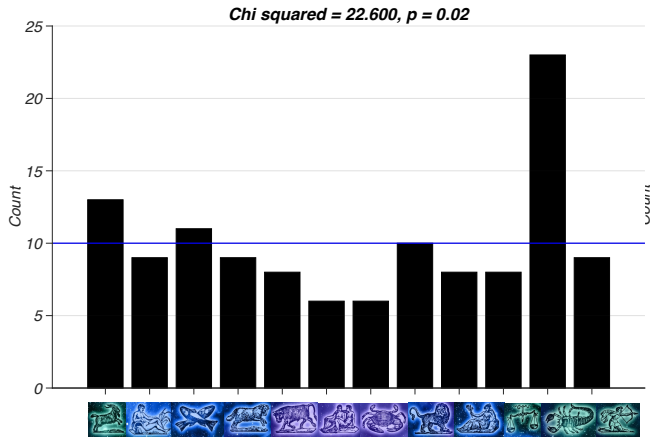
n: 121-240



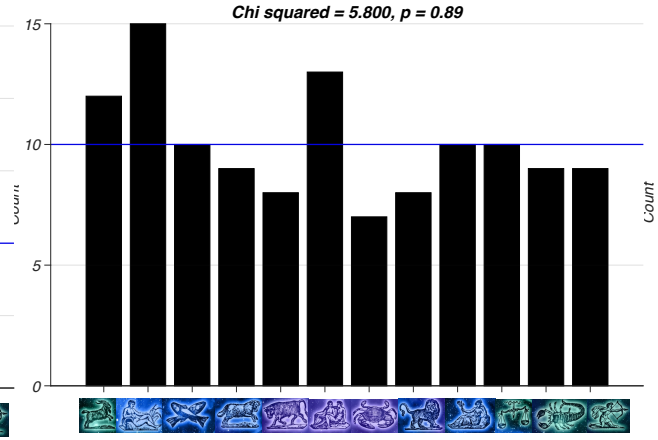
n: 241-360



n: 1-120



n: 121-240



n: 241-360