Statistics

Philippe Fanaro

Heavily influenced by outliers

mean → "average" $\frac{\sum_{i=1}^{n} x_i}{n}$

median → middle value

mode → most common value

no spread-out description ✗

no description of the shape ✗

Central Tendency

Range → (max − min)

Variance → $\frac{\sum(x_i-\mu)^2}{n-1} = s^2$ (sample) | $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$ (population)

depends on the context it's not always everything

Bessel's correction

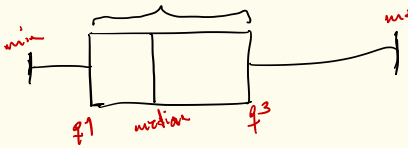
standard deviation → $\sqrt{\sigma^2}$ → same units as the main variable

Dispersion

Quantiles — Divide in half → Divide in half → 1st quantile → value that splits the 1st 25% of the data

if it isn't in the data, get the mean

IQR →



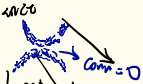
Fence and Outliers → Commonly, we Fence at 1,5 x IQR beyond q1 and q3 → outliers are drawn away from the whiskers

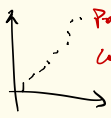
Determined by the Data, not a prior definition!

Bivariate Data and Covariance

scatter plots may uncover correlation → but correlation is not necessarily causation

Weird structures cannot be captured





Pearson Corr Coefficient

↳ Normalized Covariance

Covariance → Normalize the data in order to be able to interpret the results

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\sum(x-\bar{x})^2}{n}}\sqrt{\frac{\sum(y-\bar{y})^2}{n}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})$$

Permutation → $P_n = n!$ PERMUT(n,n)

→ Out of a subset → ${}_nP_r = \frac{n!}{(n-r)!}$ (no repetition) PERMUTATIONS(n,r)

→ with repetition → n^r

arrangement → $A_{n,r}$

Combinations → Unordered

→ ${}_nC_r = C_{n,r} = \frac{n!}{r!(n-r)!}$ COMBIN(n,r)

→ With repetition → $C_{n+r-1,r} = \frac{(n+r-1)!}{r!(n-1)!}$ COMBINA(n,r)

Intersections, Unions and Complements

→ $A \cap B \Rightarrow$ AND

→ $A \cup B \Rightarrow$ OR → $P(A) + P(B) - P(A \cap B)$

could happen! $A \cup B \neq U$

→ $\bar{A} = U - A \Rightarrow 1 - P(A)$

$P(R_1 \cap R_2) =$

$= P(R_1) \cdot P(R_2|R_1) =$

$= \frac{3}{5} \cdot \frac{2}{5} = 0,3$

Conditional Prob and Bayes Theorem → $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$P(B|A) = \frac{P(A \cap B)}{P(A)}$

Bayes Theorem:

$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$, $P(A), P(B) > 0$

$P(Def|Pos) = \frac{P(Pos|Def)\,P(Def)}{P(Pos)} =$

$= \frac{0,99 \cdot 0,002}{P(T \cap Pos) + P(F \cap Pos)} = \frac{0,99 \cdot 0,002}{P(Pos|Def) \cdot P(Def) + P(Pos|\overline{Def}) \cdot P(\overline{Def})} =$

$= \frac{0,99 \cdot 0,002}{0,99 \cdot 0,002 + 0,04 \cdot 0,998} = 0,165$

$P(Def|Pos|Pos)$

$P(Def|Pos) = \frac{0,99 \cdot 0,165}{0,99 \cdot 0,165 + 0,04 \cdot 0,835} = 0,951$

If a second test

Discrete Distributions
- Uniform
- Binomial
- Poisson

Bernoulli Trial — 2 possible outcomes, independent trials

successes

$$P(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

from scipy.stats import binom
binom.pmf(3, 16, 1/6)

Normal Distribution

$\mu = np$

$\sigma^2 = np(1-p)$

Poisson
- Bernoulli → successes out of n trials
- Poisson → successes per unit of time (continuous unit)

$$\lambda = \frac{\#\text{ occurrences}}{\text{interval}} = \mu = E(X) = \sigma^2$$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \rightarrow \text{Prob of seeing } x \text{ occurrences per interval}$$

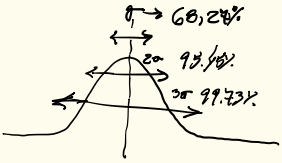
$$cdf(x) = P(X \le x) = \sum_{i=0}^{N} \frac{\lambda^i e^{-\lambda}}{i!}$$

Assumes that that prob during small time interval is proportional to the entire length, i.e.

$\lambda_{min} = \frac{\lambda_{hour}}{60}$

Continuous Distributions
- Normal
- Exponential
- Beta

Normal → Standard Normal Distribution $N(0, 1)$ (μ, σ)



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

Z-Scores

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

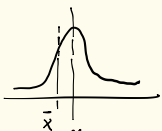
from scipy import stats
stats.norm.cdf(z=0.4)
stats.norm.ppf(p=0.95)

Sampling
- Selection Bias — the samples chosen don't quite reflect the population
  - Undercoverage
  - Self-selection
  - Healthy-user
  - Survivorship — Hollywood: head injuries ↑, head statistics ↓
- Types
  - random
  - stratified random
  - cluster — less precise

>30 usually quite good

If you wish to know how similar $F_n(x)$ and $\Phi(x)$ are, use the Berry-Esseen Inequality

Central Limit Theorem (CLT)
- Mean values will be normally distributed even if the population itself is not normally distributed



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

Standard Error = $\frac{\sigma_{pop}}{\sqrt{n}} \approx \frac{\sigma_{sample}}{\sqrt{n}}$

Hypothesis Testing
- Start with a Null Hypothesis
  - Accept / fail to reject
  - Reject → Assume another mutually exclusive hypothesis
- H_0 usually should be the opposite of what you want to "prove"
- We never prove a hypothesis
- Null hypothesis should contain an equality ($\geq$, $\leq$, $=$)
  - $>$ → right-tail
  - $<$ → left-tail
  - $=$ → two-tail
- level of significance α → tail of the null hypothesis

Testing: Means vs Proportions

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad z = \frac{\hat{P} - P}{\sqrt{\frac{P \cdot q}{n}}} = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

- P-value test
  - take test statistic
  - use it to determine the P-value
  - Compare the P-value to the level of significance α

P value is low → Null must go → reject H_0

P value is high → Null must fly → fail to reject H_0

Example 1: Load time // $\mu = 3,125$ s, $\sigma = 0,4$ s

Desired Confidence on improvement: 99% → $\alpha = 0,01$

$n = 40$

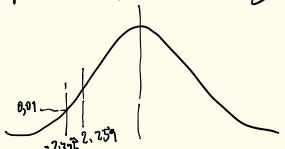
New average → $\bar{x} = 2,875$ s

$H_0: \mu \geq 3,125$ → left tail

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = -2,259$

$\alpha \rightarrow z = -2,325$ → fail to reject H_0

p-value → 0,0119 > 0,01 → fail to reject H_0

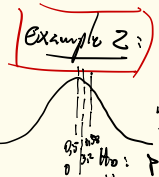


With 99% confidence the new average is not an improvement

Example 2: $n = 100$

58% → teenagers

Are most customers teenagers?



$P \geq 0,5$

$\alpha = 0,05$ (right-tail) → 1,645

$z = \frac{0,58 - 0,5}{\sqrt{\frac{0,5(1-0,5)}{100}}} = \frac{0,08}{0,025} = 3,2$

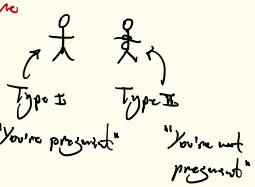
$3,2 > 1,645$ → reject H_0

Type I and II Errors

Decision about H_0	H_0 is: True	H_0 is: False
Fail to Reject	True Negative ($p = 1-\alpha$)	Type II ($p = \beta$)
Reject	Type I ($p = \alpha$)	True Positive ($p = 1-\beta$)

False Negative

False Positive



Student's T-Distribution

↳ William Gossett → pseudonym Student
Guinness Brewery

→ Select the best barley from small samples where the standard deviation was unknown.

t-table / t-statistic → t-test determines if there is significant difference between two sets of data

(1): $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ → t-table

- Degrees of freedom
- Chosen significance level
- t → t-statistic
  - $t_{n-1,\alpha}$ = t-critical
  - $n-1$ = degrees of freedom
  - α = significance level

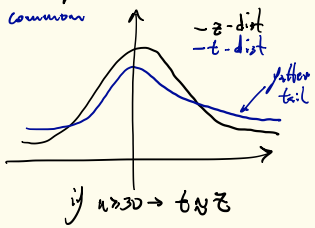
s: sample variance

t-test → signal / noise

(1) One sample t-test example — $\mu = \mu$?
(2) Independent 2 sample test — $\mu_1 \approx \mu_2$?
(3) Dependent paired sample t-test
↳ Ex.: test scores before and after a prep course

(2):
- equal sample sizes, equal variance
- unequal sample sizes, equal variance
- unequal or equal sample sizes, unequal variance
  ↳ most common

$t = \frac{\text{signal}}{\text{noise}} = \frac{\text{difference in means}}{\text{sample variability}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$



degrees of freedom:

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

if $s_1 = s_2$:

$df = n_1 + n_2 - 2 = (n_1 - 1) + (n_2 - 1)$

Example:

Two plants, same car
Which one to close?

$n_A = 10$ $\bar{x}_A = 1282$
$n_B = 10$ $\bar{x}_B = 1186$ → Is A statistically different from B?

H_0: $x_A \leq x_B$ one tailed test $s_A^2 = 1288$
H_1: $x_A > x_B$ $df = (10+10-2) = 18$ $s_B^2 = 1256$

$t = \frac{|\bar{x}_A - \bar{x}_B|}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = 2,28$

↓

Critical Value: 1,734

↓

$2,28 > 1,734$

↓

Reject H_0 → Plant A produces more than B

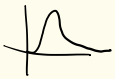
t-table		
conf. prob	t.10	t.05
one tail	0.1	0.05
two tail	0.2	0.1
17		
18		
19		
⋮		

ANOVA

(analysis of variance)

— Previously: z- and t-distributions

↓

Now: F-distribution 

↳ What is the probability that two samples or more come from a population with the same variance

→ If you have 3 samples, you can make 3 pairings and t-test them.

↳ But the overall confidence will drop: 0,95·0,95·0,95 = 0,857

μ_{Total}

$\mu_A, \mu_B, \mu_C, \dots$

→ Variance: between groups / within groups

$$F = \frac{\text{var between groups}}{\text{var within groups}} = \frac{SS_B / df_{groups}}{SS_E / df_{error}}$$

SS_B: $n_{samples} \cdot \sum_{groups} (\mu_i - \mu_{tot})^2$

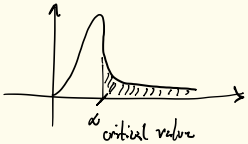
df_{groups}: $\#groups - 1$

SS_E: $\sum (x_i - \mu_i)^2$

df_{error}: $(n_{samples} - 1) \cdot \#groups$

$$s = \frac{\sum (x_i - \bar{x})^2}{n-1}$$ ← sum of squares (SS); ← df

→ F-Distribution



F-table Upper Tail Area of 0.05

denominator df \ numerator df	1	2	3	...
25				
26				

Example

Give Discounts to different customers → does it affect them? (make them pay early) → H_0: discounts yield the same behavior (distribution)

days elapsed until payment	2% Disc	1% Disc	No Disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
μ	12	17	16
μ_{tot}	15		
$(\mu_i - \mu_{tot})^2$	9	4	1 → $\sum (\mu_i - \mu_{tot})^2 = 14$
$(x_i - \mu_i)^2$	34	106	58 → $\sum (x_i - \mu_i)^2 = 198$

$$F = \frac{14 \cdot 3}{3-1} \Big/ \frac{198}{(5-1)\cdot 3} = 2,121 \Rightarrow F_{critical} = 3,885$$

$F = 2,121 < 3,885 = F_{critical}$

⇓

Fail to reject H_0

Two-Way ANOVA → More variables / e.g. 3 invoices for $50, 3 invoices for $100

different amounts may yield different conversion rates

	2%	1%	0%	$\bar{x}_{block}$
$50	16	23	21	20
$100	14	21	16	17
$150	11	16	18	15
$200	10	15	14	13
$250	9	10	11	10
$\bar{x}_{group}$	12	17	16	15	$\bar{x}_{total}$
$(\bar{x}_{group} - \bar{x}_T)^2$	9	4	1	14	28

- each row is a block
- You want to isolate and remove any variance contributed by the blocks, to better understand the variance in the groups.
  - ↳ $\bar{x}_{block}$, SSBlock

$SS_{Error} = SS_{Total} - SS_{Groups} - SS_{Blocks}$

$df_{groups} = n_{groups} - 1$

$df_{error} = (n_{blocks} - 1)(n_{groups} - 1)$

$F = \frac{SS_{groups}/df_{groups}}{SS_E/df_{error}}$

				Σ
$(\bar{x}_{block} - \bar{x}_{total})^2$	25	4	0	4	58

$SS_E = 14 \cdot (5) = 70$ ($n_{samples\ in\ group}$)

$df_{Groups} = 3 - 1 = 2$

$SS_B = 58 \cdot (3) = 174$ ($n_{samples\ in\ block}$)

$SS_T = 268$

$SS_E = SS_T - SS_E - SS_B = 24$

$df_{error} = (5-1)\cdot(3-1) = 8$

$F_{groups} = \frac{70/2}{24/8} = \frac{35}{3} = 11{,}67$

$\alpha = 0{,}05$ $df_{num} = 2$ $df_{denom} = 8$

$F_{critical} = 4{,}46$

$F_{critical} = 4{,}46 < 11{,}67 = F_{groups}$

⇓

reject H_0 ⇒ giving discounts yields statistically different results

Two-Way ANOVA with Replication

Heights of Plants / types of fertilizers A, B, C (groups) — H_0: no statistical difference

temperatures warm and cold (blocks)

Fertilizer	A	B	C	$\bar{x}_{block}$
Warm	13	21	18	16
	14	19	15	
	12	17	16	
Cold	16	14	15	14
	18	11	13	
	14	14	8	
$\bar{x}_W$	13	19	16	
$\bar{x}_C$	17	13	12	

($\bar{x}_{sm}$ in sample means)

$\sum (\bar{x}_B - \bar{x}_T)^2 = (16-15)^2 + (14-15)^2 = 2$

$SS_B = 2 \cdot (9) = 18$ ($n_{samples\ in\ block}$)

$\sum (\bar{x}_G - \bar{x}_T)^2 = (15-15)^2 + (16-15)^2 + (14-15)^2 = 2$

$SS_E = 2 \cdot (6) = 12$ ($n_{samples\ in\ group}$)

$\bar{x}_T$	15

$\bar{x}_{col}$	15	16	14

$df_{groups} = 3 - 1 = 2$

$df_{blocks} = 2 - 1 = 1$

$SS_T = 164$

$SS_E = 164 - 18 - 12 - 84 = 50$

$df_{error} = blocks \cdot groups \cdot (observ. - 1) = 2 \cdot 3 \cdot (3-1) = 12$ (observ. in sample)

- SS Statistic (interaction)

$SS_I = \sum_{sm} (\bar{x}_{sm} - \bar{x}_{block} - \bar{x}_{col} + \bar{x}_T)^2 =$

$= (13 - 16 - 15 + 15)^2 +$

$(19 - 16 - 16 + 19)^2 +$

$\vdots$

$= SS_I = 28 \cdot (3) = 84$ (within in sample)

fail to reject H_0 ← $F = \frac{12/2}{50/12} = 1{,}44$

$\alpha = 0{,}05$

$F_{crit} = 3{,}885$

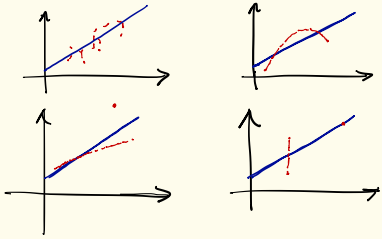
Regression

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \rho_{x,y} \frac{\sigma_y}{\sigma_x} \qquad \rho_{x,y} = \text{Pearson Corr coefficient} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Limitations — Anscombe Quartet illustrates the pitfalls of relying on pure calculation (1973)



↳ Same Regression ↳ totally different underlying functions

Multiple Regression

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots$$

$$b_1 = \frac{\sum (x_2 - \bar{x}_2)^2 \sum (x_1 - \bar{x}_1)(y - \bar{y}) - \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \sum (x_2 - \bar{x}_2)(y - \bar{y})}{\sum (x_1 - \bar{x}_1)^2 \sum (x_2 - \bar{x}_2)^2 - \left(\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)\right)^2}$$

$$b_2 = \cdots$$

- Avoid using factors without correlation → they will just be noise

χ^2 Chi-Squared analysis / Karl Pearson (1900)

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(12-9)^2}{9} + \frac{(6-9)^2}{9} = 2$$

(observed; expected; coin example: heads)

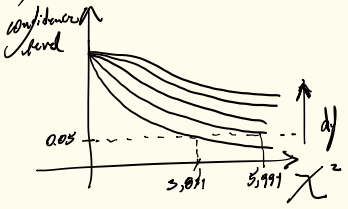
how much our observations diverge from the expected

$$Q = \sum_{i=1}^{k} z_i^2 \sim \chi^2(k)$$

(k: degrees of freedom)

the product of two independent Gaussians is a chi-squared distribution

χ^2-distribution



$df_{coins} = 2 - 1 = 1$

$\chi^2 = 2 < 3,841 = \chi^2_{crit}$

⇓

fail to reject H_0

(H_0: 12 heads out of 18 flips was reasonable)

Example:

Based on the following data, can we assume that servers fail at the same rate?

Assumptions: • failures are independent. • no "degrees of failure", either fail or not.

$E = \frac{240}{6} = 40$

servers	observed	expected	$(O-E)^2/E$
A	46	40	0,9
B	36	40	0,4
C	52	40	3,6
D	26	40	4,9
E	42	40	0,1
F	38	40	0,1
Σ	240		

$\chi^2 = \sum \frac{(O-E)^2}{E} = 10$

$\alpha = 0,05$

$df = 6 - 1 = 5$

$\chi^2_{crit} = 11,07$

$\chi^2 = 10 < 11,07 = \chi^2_{crit}$

⇓

fail to reject H_0

(98% confidence that they converge to the expected)