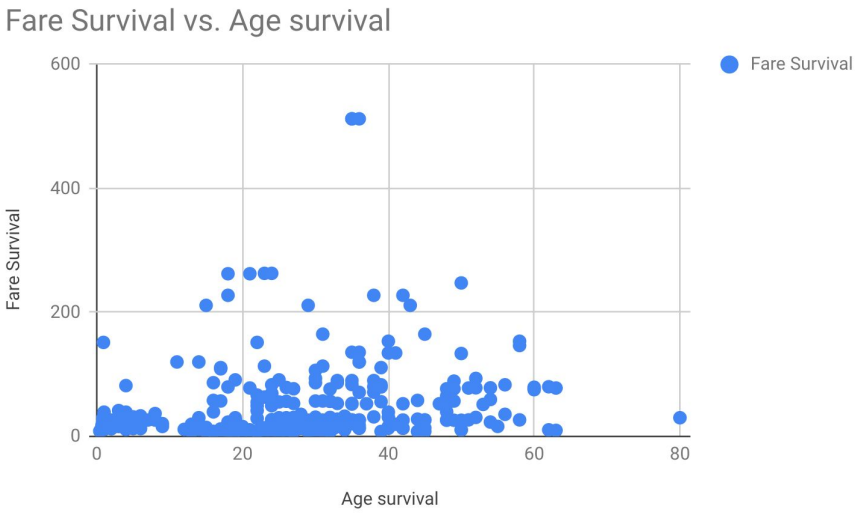
Beskrivende statistik
Roman Jurowetzki, HA 2018

Beskrivende Statistik 2018 (mini-pensum)



Datasæt kan hentes her, og en grundig beskrivelse af data er her.

Dette dokument beskriver processen af vores dataanalyse og introducerer til de forskellige statistiske begreber og koncepter, der bruges. Jeg vil bruge engelske begreber, fordi det er nemmest at finde hjælp på nettet/bøger etc., hvis man søger på engelsk.

Vores datasæt er *cross-sectional* (på dansk tværsnit). Det betyder, at der ikke findes en tidsdimension.

Generelt er der *3 typer af datasæt*.

- Cross sectional (dansk: tværsnit)
  Data beskriver forskellige personer, virksomheder, byer etc. De enkelte variabler er enten bundet til ét tidspunkt (indbyggertal i 2015, person x's mening om noget i 2018) eller noget fast, som køn (måske ikke det bedste eksempel 😉)

- Time series
  Observation af noget over tid. Fx. aktiekurser i år 2014, 2015, 2016, osv.

- Panel
  Flere tidspunkter, flere variabler, flere personer/objekter (se tabellen nedenfor).

Beskrivende statistik
Roman Jurowetzki, HA 2018

person	year	income	age	sex
1	2001	1300	27	1
1	2002	1600	28	1
1	2003	2000	29	1
2	2001	2000	38	2
2	2002	2300	39	2
2	2003	2400	40	2

Der findes flere typer af variabler. De vigtigste for os er følgende:

- **Text** (eller *string*): Kan indeholde kategorier (male/female)
- **Tal**
- **Boolean**: `TRUE/FALSE` ~ 1/0 - Man kan faktisk regne med dem og fx `=AVERAGE(B)` hvis kolonne B indeholder TRUE/FALSE værdier
- **Dato** (ikke her) men vi kommer til at arbejde med det senere

Simple beregninger og betingelser

Vi starter vores analyse med at tælle, hvor mange mænd og kvinder der er på Titanic (i vores data).

`=COUNTA()` bruges til at tælle alle rækker

Vi skriver `=COUNTA(data!E2:E)` fordi vores data er i en *sheet,* der hedder data.
Det er en god ide at lave en ny fane til beregninger/logics. Til den her beregning kan vi bruge hver kolonne, hvor der ikke er *missing values* dvs., hvor vi har data. Vi ved, at vi ved noget om køn på alle.

=COUNT() kan også bruges men kun med numeriske værdier.

Derfor kan vi sige, at:
Der er 981 personer om bord

Vi bruger `=UNIQUE(data!E2:E)` for at finde ud af, hvilke kategorier der er i en kolonne, som viser `male` og `female`

Vi kan bruge `=COUNTIF()` for at returnere en betinget optælling i et område.
Hvor mange mænd er der på Titanic? `=COUNTIF(data!E2:E,"male")`

Der er: 577 mænd og 314 kvinder

Nu skal vi videre og finde ud af, hvor mange der overlevede.

Datasættet indeholder variablen `survived` i kolonne `B`, som kun har 1 og 0 som værdier.
Variablen er numerisk, men kan betragtes som en boolean. 1 = ja, 0 = nej.
At beregne gennemsnittet for sådan en variable vil give os *sandsynligheden* for at overleve.

Beskrivende statistik
Roman Jurowetzki, HA 2018

Vi bruger `=AVERAGE(data!B2:B)` som returnerer 0.38 eller 38.38%
Husk at du kan klikke på % i menuen for at formatere din celle.

Hvis vi skal beregne en værdi kun for mænd og kun for kvinder, så skal vi bruge =AVERAGEIF(), der returnerer gennemsnittet af et område (i excel eller sheets) afhængigt af kriterier.

Vi bruger `=AVERAGEIF(data!E2:E, "male", data!B2:B)`
`data!E2:E = område for betingelse`
`"male" = betingelse`
`data!B2:B = gennemsnit af hvad`
Mænd: 18.89%
Kvinder: 74.20%

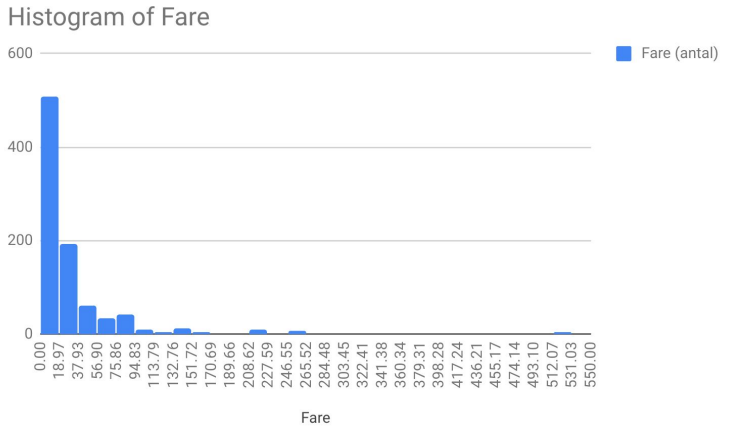
Næsten 3 ud af 4 kvinder overlever, mens mindre end 1 ud af 5 mænd overlever.

Er der forskelle i overlevelsen mellem rige og fattige passager?

Vi har ikke eksplicitte oplysninger om deres økonomiske status, men vi kan bruge billetprisen som en *proxy,* fordi vi kan **antage,** at billetprisen `Fare` afspejler, hvor fattig eller rig en passager er.

Variablen `Fare` er numerisk og vi kan se, at billetpriser var meget forskellige (nogle dyere end andre, nogle billigere end andre). For at få en bedre fornemmelse af fordelingen af variablen, kan vi beregne `=AVERAGE()`, `=MEDIAN()`, `=MIN()`, `=MAX()` og vise forskellige percentiler fx `=PERCENTILE(J:J, 0.6)`

Vi kan også bruge et histogram ved at indsætte et diagram. Sheets genkender datatypen og laver et histogram helt automatisk.



Beskrivende statistik
Roman Jurowetzki, HA 2018

Denne visualisering er meget praktisk, hvis vi skal inddele vores passagerer i flere kategorier. Der er mange komplekse løsninger for, hvordan en fordeling som den her kan inddeles, men her træffer vi bare en besluttning om at lave en ny kolonne ved at bruge

```
=IFS(J2<20, "poor", J2<40,  "OK", J2 < 60, "rich", J2 >= 60, "mega rich")
```

Denne formel evaluerer flere betingelser og returnerer en værdi, der svarer til den første sande betingelse. Hvis billetprisen er under 20, så antager vi, at personen er fattig osv. Det er en god ide, at have betingelser, der ikke modsiger hinanden.

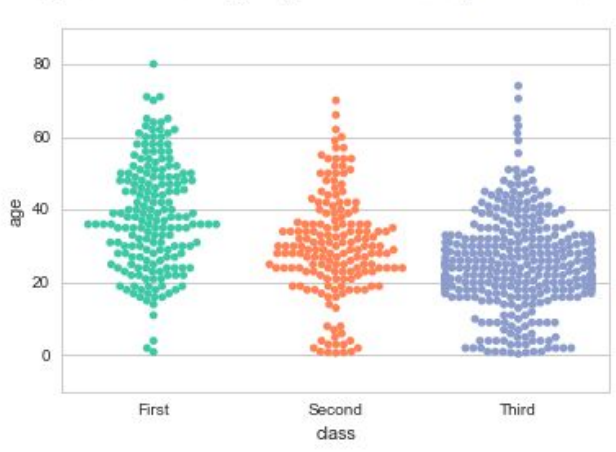
For at finde ud af om der er *korrelation* (**IKKE kausal sammenhæng og ikke teknisk korrelation - det kommer senere**) mellem økonomisk status og overlevelse, kan vi igen bruge `=AVERAGEIF()`. Denne gang bruger vi dog ikke køn men vores økonomi-kategorier i den nye kolonne K.

Vi kan se, at overlevelses-% går op med (beregnet) økonomisk status.

Indtil videre har vi mest brugt middelværdi som indikator. Det kan dog blive et problem, hvis vores data er fordelt som i tilfældet med vores billetpriser. Her er gennemsnitet 32.20. Dog kan vi se i vores histogram, at de fleste om bord har billigere billetter. Gennemsnittet tager ikke højde (eller det gør de faktisk men det hjælper ikke) for "outliere". Det er her, hvor medianen (eller den midterste værdi) kan hjælpe. For Fare i Titanice datasæt ligger den ved 14.45 dvs. halvdelen af personer om bord betalte 14.45 eller mindre. Medianen giver tit mere information om fordelingen af en kontinuert variabel end middelværdien.
Ligeledes kan vi beregne `=QUARTILE(A2:A100, 3)` og `=PERCENTILE(A2:A100, 0.95)`.

Spredning og Standardafvigelse

En vigtig beregning i beskrivende statistik er spredning eller standardafvigelse.
For at beregne SD (engl. standard deviation), skal vi dog først forestå varians.



Dette (swarm)plot viser alder vs klasse. Som vi allerede ved (forelæsning), er folk på 1. klasse ældre end på 2. og 3. (hhv. 38, 30, 25). For at forestå spredning, starter vi med at måle afstanden fra hver punkt i forhold til gennemsnittet for vores variable.

Beskrivende statistik
Roman Jurowetzki, HA 2018

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n}$$

σ^2 = varians
μ_x = gennemsnit af vores variable
x_i = én observation af vores variable
n = antal af observationer

I sheets bruger vi `=VARP(B:B)` for at beregne varians for hele populationen (kun `VAR` beregner varians for en stikrøve - ikke del af vores forelæsning)

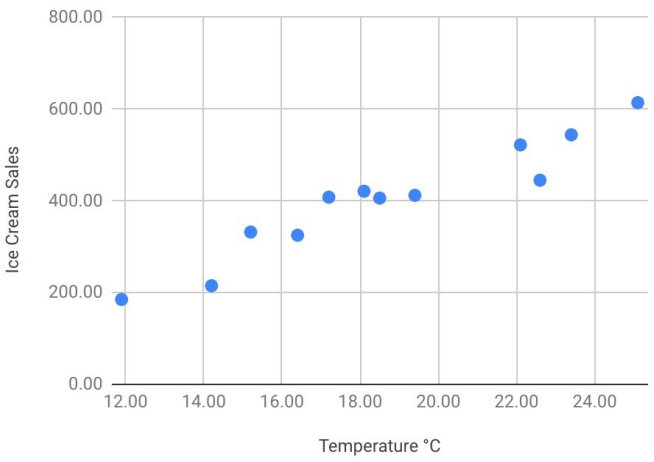
Bemærk at spredningen ofte betegnes standardafvigelse - deraf forkortelsen SD, som står for den engelske betegnelse standarddeviation.

$$SD = \sqrt{\sigma^2}$$

Hvad med at se på sammenhængen mellem 2 variabler?

Temperatur vs Is solgt

```
14.20 - 215.00
16.40 - 325.00
11.90 - 185.00
15.20 - 332.00
18.50 - 406.00
22.10 - 522.00
19.40 - 412.00
25.10 - 614.00
23.40 - 544.00
18.10 - 421.00
22.60 - 445.00
17.20 - 408.00
```



Hvis vi skal anlysere sammenhæng mellem 2 variabler, så er den første ting, vi gør at lave et scatterplot. Her kan vi fx se, at der bliver solgt mere is, når temperaturen stiger. Ændringer i de to variabler følges ad - de to variabler er korelleret. Men vi kan også måle korrelation. Det er en god ide, hvis vi skal analysere mange variabler samtidigt. Det mest brugte mål for korrelation er Pearsons r eller r-værdi (eller meget nørdet Pearson's produkt-moment korrelationskoefficient).

Lad os starte med at sige at værdien af Pearsons r går fra -1 til 1.

1 indikerer en perfekt korrelation. 0 ingen korrelation (vi vil bare se tilfældigepunkter i et scatterplot). -1 indikerer perfekt negativ korrelation. Is vs temperatur har fx en r-værdi af 0.96 og er dermed næsten perfekt korreleret.

I sheets kan vi bruge `=CORREL(G5:G16,H5:H16)` for at beregne vores r-værdi. Men det er godt lige at have en forståelse af , hvad der sker “under the hood”.

Man kan tale om, at der er en korrelation (sammenhæng), hvis 2 variabler varierer “sammen” - den ene går op og den anden følger ved enten at gå op eller ned: Det hedder også kovarians.

Beskrivende statistik
Roman Jurowetzki, HA 2018

$$cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{n}$$

I sheets kan vi bruge `=COVAR(A:A;B:B)` hvor A:A og B:B er to lige lange kolonner (vektorer)

Ligesom varians for 1 variable, bare med en $(y_i - \mu_y)$ term i tælleren i stedet for $(x_i - \mu_x)^2$

r-værdi kan nu beregnes ved at dividere $cov(x,y)$ med $\sigma_x \sigma_y$:

$$\frac{cov(x,y)}{\sigma_x \sigma_y}$$