Basic Statistics: Home assignment 1

Johannes Malm, 8111045996

Load libraries

```
library(tidyverse)
library(ggpubr)
```

1.1 "Hand" Calculations

```
obs <- sort(c(3.9, 5.3, 6.1, 4.9, 9.1, 2.8, 3.5, 3.2, 2.6, 5.9))
n = 10

mean_obs <- (3.9+5.3+6.1+4.9+9.1+2.8+3.5+3.2+2.6+5.9)/n

var_calc <- tibble(obs = obs) %>%
  mutate(mean = mean_obs) %>%
  mutate(xi_x2 = (obs - mean)^2) %>%
  mutate(sum_xi_x2 = sum(xi_x2)) %>%
  mutate(var = sum_xi_x2/(n-1))

knitr::kable(var_calc)
```

obs	mean	xi_x2	sum_xi_x2	var
2.6	4.73	4.5369	35.501	3.944556
2.8	4.73	3.7249	35.501	3.944556
3.2	4.73	2.3409	35.501	3.944556
3.5	4.73	1.5129	35.501	3.944556
3.9	4.73	0.6889	35.501	3.944556
4.9	4.73	0.0289	35.501	3.944556
5.3	4.73	0.3249	35.501	3.944556
5.9	4.73	1.3689	35.501	3.944556
6.1	4.73	1.8769	35.501	3.944556
9.1	4.73	19.0969	35.501	3.944556

```
median_position <- 50*(n+1)/100 # 5.5
median_interpolation <- (obs[5] + 0.5)*(obs[6] - obs[5])

print(paste0('a) mean: ', mean_obs, ', variance: ', round(var_calc$var[1],2)))
```

```
## [1] "a) mean: 4.73, variance: 3.94"
```

```
print(paste0('b) median: ', median_interpolation))
```

```
## [1] "b) median: 4.4"
```

1.2 Computer Exercise

a) Import dataset

```
cordblood <- read_delim("Data/cordblood.txt", locale = locale(encoding = "latin1"))
knitr::kable(head(cordblood, 25))
```

Hospital	Age	Sex	Measles	Parotitis	Rubella	Chickenpox
Mölndal	27	male	247	231	118.7	1751
Mölndal	37	female	3291	231	186.9	882
Mölndal	30	female	422	289	67.3	518
Mölndal	37	female	12946	8346	79.6	2911
Mölndal	29	female	1164	1235	75.5	1802
Borås	23	female	1875	2212	49.3	2070
Borås	29	female	2816	834	47.2	436
Mölndal	31	female	2252	5773	100.0	1085
Mölndal	36	male	3931	2026	70.5	297
Mölndal	32	male	5217	231	68.0	1147
Borås	22	female	1935	2521	35.0	1418
Borås	35	male	2312	2229	94.4	1601
Mölndal	27	male	1466	684	38.8	2911
Borås	27	female	2357	231	84.9	1930
Borås	33	female	3975	4954	36.3	3284
Sundsvall	37	female	1461	5313	22.0	2278
Sundsvall	32	male	4649	1553	36.7	760
Mölndal	35	male	11176	392	174.0	1057
Mölndal	30	male	3990	3952	52.9	2868
Mölndal	28	male	5374	794	24.1	1674
Sundsvall	37	female	5569	3126	94.5	1407
Borås	30	male	2859	1917	38.0	3288
Sundsvall	34	male	1555	1789	124.7	2140
Borås	36	male	10594	2843	101.5	1950
Sundsvall	29	male	3372	3037	100.7	2241

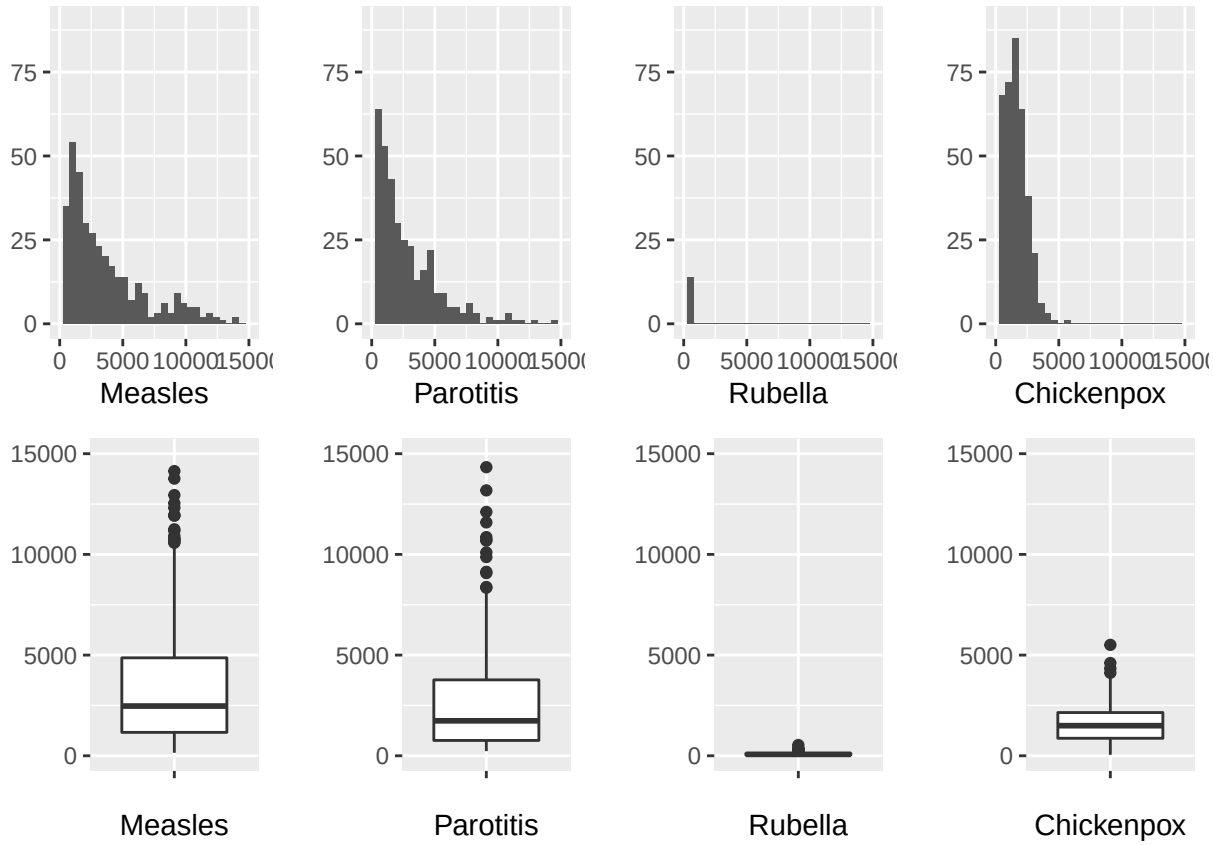
b) Graphs and computations

1 Histograms and boxplots

```
create_hist <- function(data, name) {
  return(
    ggplot(cordblood, aes(data)) +
      geom_histogram() +
      labs(x = name, y = '') +
      xlim(0,15000) +
      ylim(0, 90)
  )
}

create_bp <- function(data, name) {
  return(
    ggplot(cordblood, aes(x = '', data)) +
      geom_boxplot() +
      labs(x = name, y = '') +
      ylim(0, 15000)
  )
}

var_names <- c('Measles', 'Parotitis', 'Rubella', 'Chickenpox')
ggarrange(create_hist(cordblood$Measles, var_names[1]),
          create_hist(cordblood$Parotitis, var_names[2]),
          create_hist(cordblood$Rubella, var_names[3]),
          create_hist(cordblood$Chickenpox, var_names[4]),
          create_bp(cordblood$Measles, var_names[1]),
          create_bp(cordblood$Parotitis, var_names[2]),
          create_bp(cordblood$Rubella, var_names[3]),
          create_bp(cordblood$Chickenpox, var_names[4]),
          ncol = 4, nrow = 2)
```



Comment: In lack of knowledge when it comes to level of antibodies in the cord blood; I made the assumpation that a certain level of antibodies is equally good/bad across all diseases in the dataset. With that assumption I set the limits on the X- and Y-axis to fixed limits accross the graphs for better comparison. The historgrams show that Measles and Parotitis distribution of level of antibodies are clearly skewed to the left and also indicates a larger variance than the other two. Rubella shows low levels of antibodies compared to the other three. When it comes to chickenpox the histogram shows a distribution which is more narrow than Measles and Parotitis which could be interpreted that the spread is lower.

2 Mean, median, variance, interquartile

```
calc_values <- function(data, name) {
  df <- tibble(name = name,
               mean = mean(data),
               median = median(data),
               variance = var(data),
               iqr = IQR(data)
               )
  return(df)
}

res_measles <- calc_values(cordblood$Measles, "Measles")
res_parotitis <- calc_values(cordblood$Parotitis, "Parotitis")
res_rubella <- calc_values(cordblood$Rubella, "Rubella")
res_chickenpox <- calc_values(cordblood$Chickenpox, "Chickenpox")

bind_res <- bind_rows(res_measles, res_parotitis, res_rubella, res_chickenpox)
bind_res$mean <- round(bind_res$mean, 1)
bind_res$median <- round(bind_res$median, 1)
bind_res$variance <- round(bind_res$variance, 1)
bind_res$iqr <- round(bind_res$iqr, 1)

knitr::kable(bind_res)
```

name	mean	median	variance	iqr
Measles	3545.5	2471.0	9789270.6	3711.0
Parotitis	2577.8	1737.0	5999382.9	3009.0
Rubella	90.1	70.5	5196.4	80.5
Chickenpox	1587.2	1496.0	796967.0	1276.0

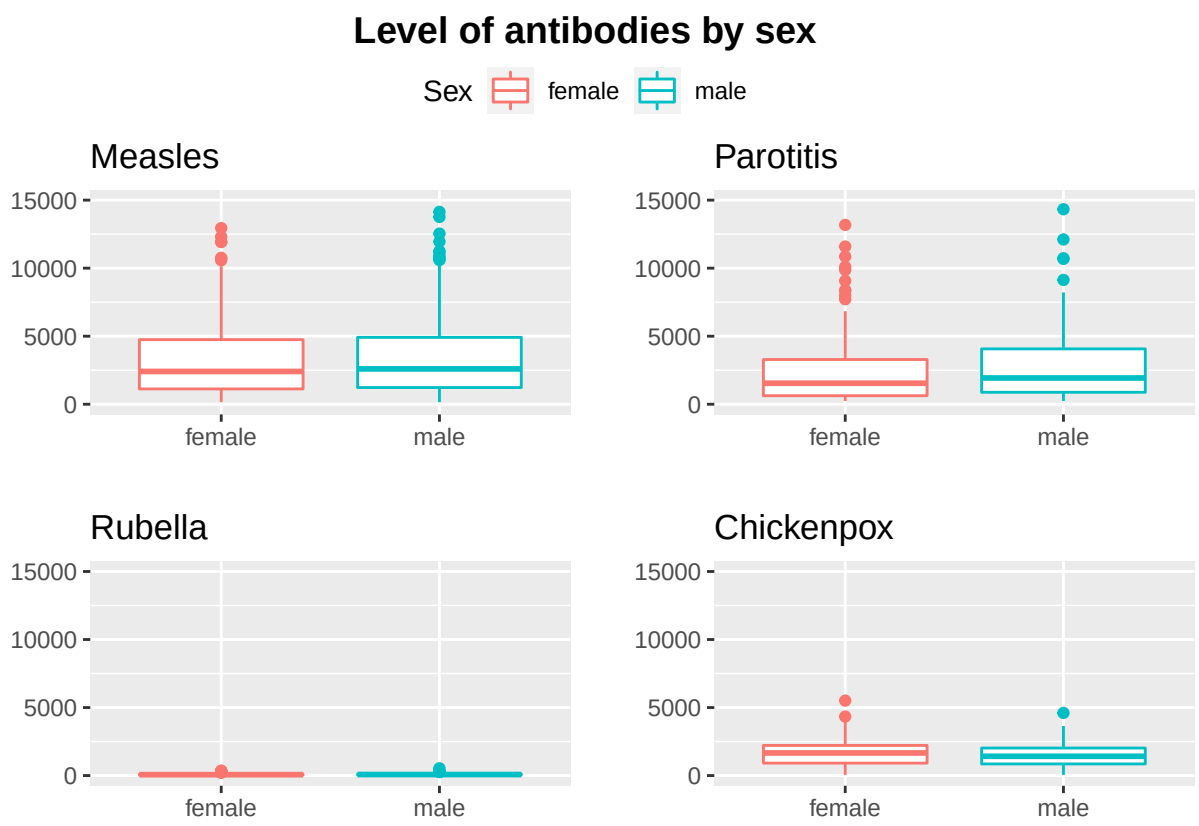
Comment: It depends on what you are after. To me its more meaningful to use the median and IQR. In the graphs we observed a skewness in the distributions for Measles, Parotitis and Chickenpox. With skewness we know that median is not the same as mean. The mean could be influenced by outliers which the median is not that sensitive to in a dataset of this size. The skewness and outliers also pushes up the variance but by using IQR we might get more meaningful picture of the spread.

3 Mean, median, variance, interquartile grouped by sex

```
create_by_sex <- function(data, name) {
  return(ggplot(cordblood, aes(x=Sex, y=data, color=Sex)) +
           geom_boxplot() +
           labs(x="", y="", title = name) +
           theme(legend.position="none") +
           ylim(0,15000)
  )
}

plot <- ggarrange(create_by_sex(cordblood$Measles, var_names[1]),
          create_by_sex(cordblood$Parotitis, var_names[2]),
          create_by_sex(cordblood$Rubella, var_names[3]),
          create_by_sex(cordblood$Chickenpox, var_names[4]),
          nrow = 2, ncol = 2, common.legend = TRUE, legend = "top")

annotate_figure(plot, top = text_grob("Level of antibodies by sex",
               color = "black", face = "bold", size = 14))
```




```
calc_by_sex <- cordblood %>%
  select(Sex:Chickenpox) %>%
  group_by(Sex) %>%
  summarise(Me_md = median(Measles),
```

```
            Me_iqr = IQR(Measles),
            Pa_md = median(Parotitis),
            Pa_iqr = IQR(Parotitis),
            Ru_md = median(Rubella),
            Ru_iqr = IQR(Rubella),
            Ch_md = median(Chickenpox),
            Ch_iqr = IQR(Chickenpox),

            )
    knitr::kable(calc_by_sex)
```

Sex	Me_md	Me_iqr	Pa_md	Pa_iqr	Ru_md	Ru_iqr	Ch_md	Ch_iqr
female	2458.0	3674.5	1542.0	2661.00	64.6	73.2	1662.0	1300.5
male	2595.5	3689.0	1929.5	3189.25	75.7	82.9	1416.5	1166.0

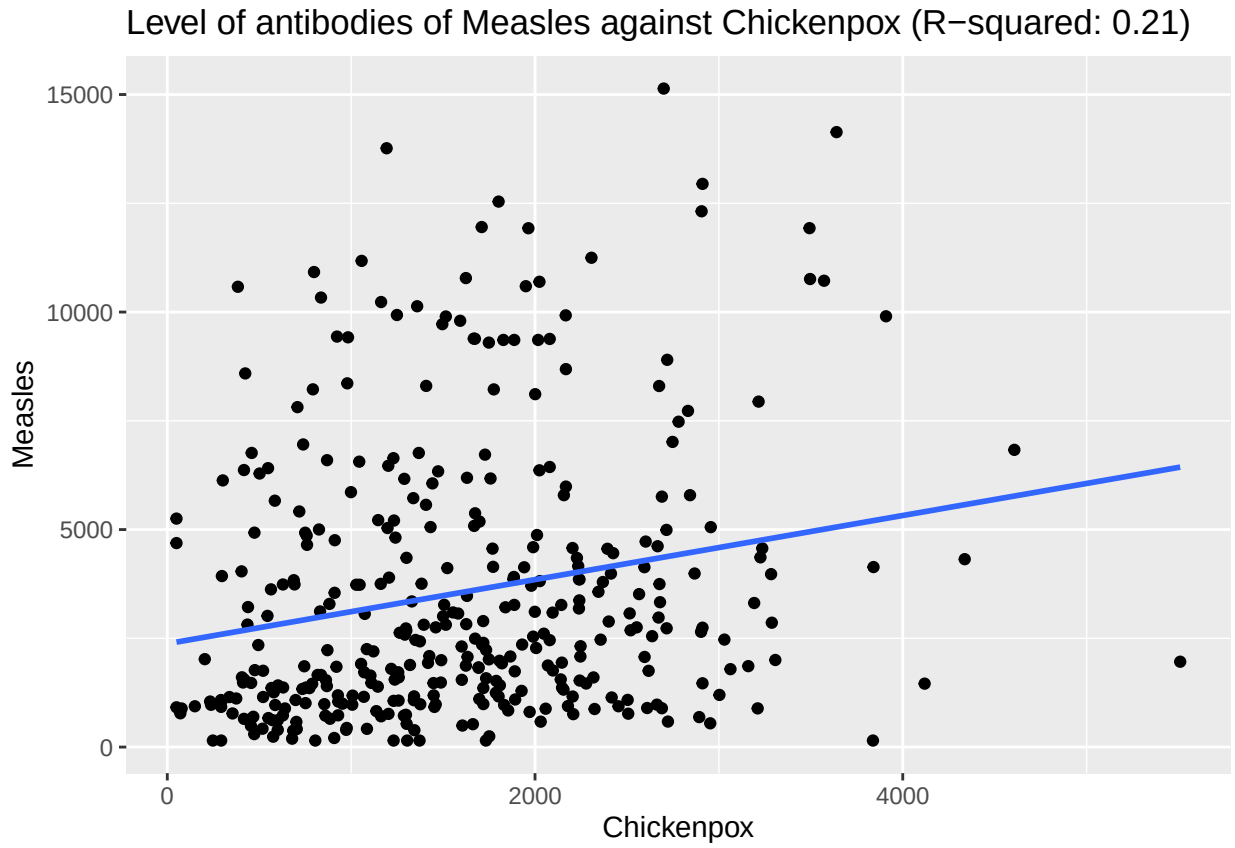
Comment: When looking at median the level of antibodies differ the most between male and female when it comes to parotites and chickenbox. With no assumption that a certain level is equally good/bad across all diseases it could have been intresting to take The Coefficient of Variation into consideration when comparing the level of antibodies.

4 Scatterplot of Measles against Chickenpox

```
corr_value <- cor(cordblood$Measles, cordblood$Chickenpox)
corr_value
```

```
## [1] 0.210237
```

```
ggplot(cordblood, aes(x = Chickenpox, y = Measles)) +
  geom_point() +
  labs(title = paste0("Level of antibodies of Measles against Chickenpox (R-squared: ", round(corr_valu
  geom_smooth(method = "lm", se=FALSE)
```



Comment: The R-squared value of 0.21 indicates low level of correlation between Measles and Chickenpox. The trend line is positive which indicates higher levels of antibodies of Measles are associated with higher levels of antibodies of Chickenpox.

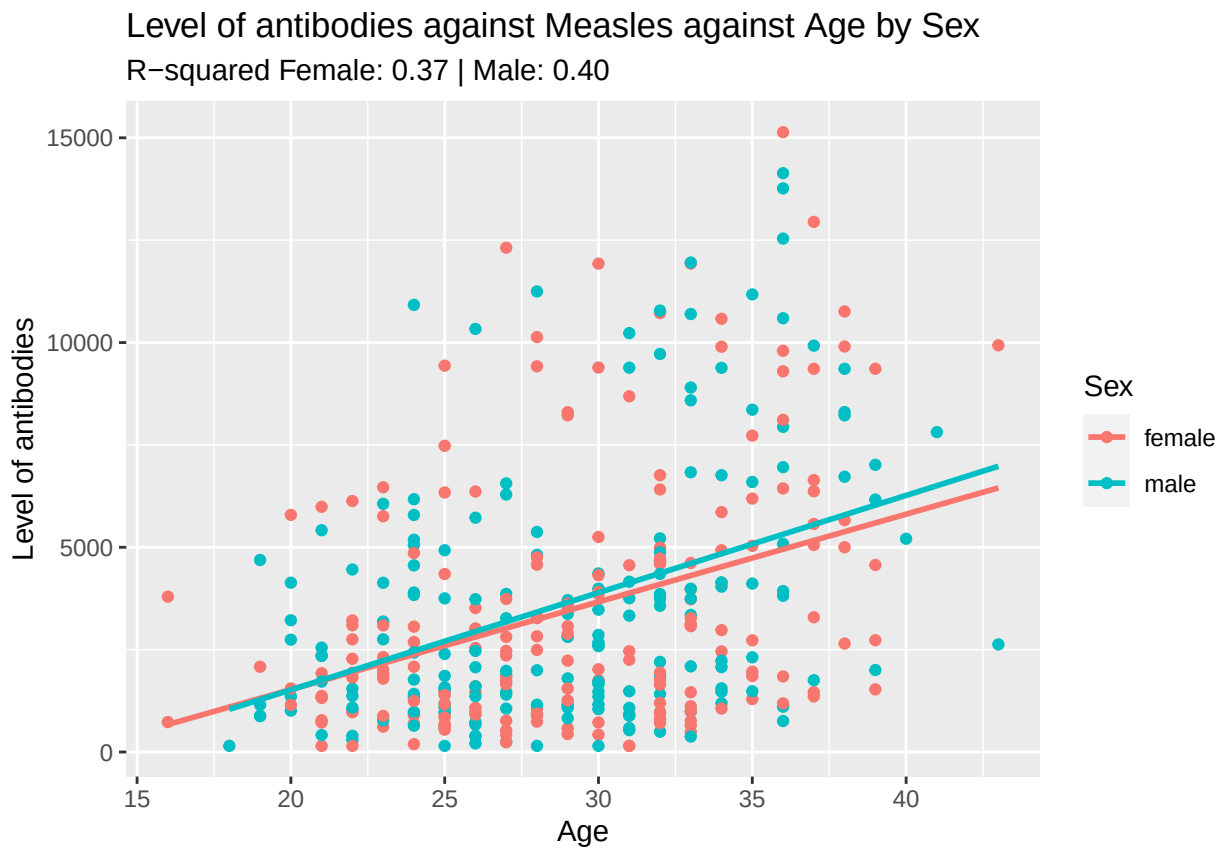
5 Scatterplot of Measles, Age and Sex

```
corr_measles_age <- cordblood %>%
  group_by(Sex) %>%
  summarise(corr_measles = cor(Measles, Age))

knitr::kable(corr_measles_age)
```

Sex	corr_measles
female	0.3670488
male	0.4034094

```
ggplot(cordblood, aes(x = Age, y = Measles, color = Sex)) +
  geom_point() +
  labs(y = "Level of antibodies", title = "Level of antibodies against Measles against Age by Sex", sub
  geom_smooth(method = "lm", se=FALSE)
```



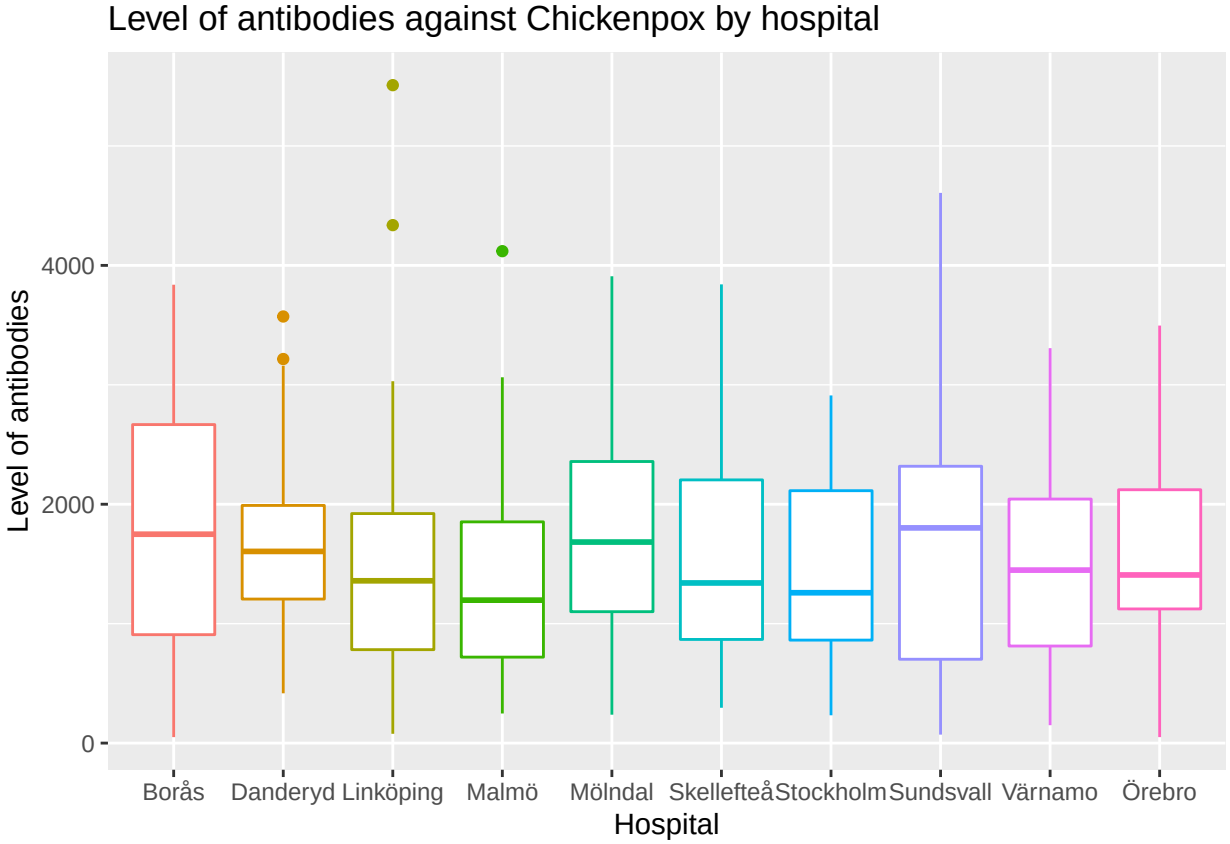
Comment: The rendlines implies when the age of the mother increases the level of antibodies against Measles increases across both sexes. The trendlines indicates small differences in level of antibodies againts Measles between the sexes. Trendline fit female: 0.37, male: 0.40.

6 Boxplots of Chickenpox by hospital

```
Sys.setlocale("LC_ALL","sv_SE.UTF-8")
```

```
## [1] "sv_SE.UTF-8/sv_SE.UTF-8/sv_SE.UTF-8/C/sv_SE.UTF-8/en_US.UTF-8"
```

```
ggplot(cordblood, aes(x=Hospital, y=Chickenpox, color=Hospital)) +
  geom_boxplot() +
  theme(legend.position="none") +
  labs(title = "Level of antibodies against Chickenpox by hospital", y = "Level of antibodies")
```



Comment: From the scatter plot we can see that none of the medians is outside any of the other boxes which implies there are not significant differences of level of antibodies against Chickenpox between hospital observations. Note the two outliers in Lidköping and a fairly large max whisker in Sundsvall.