LeaRning about Statistics -YEAH!!!

STATISTICAL TESTS!

Zachary Thompson
Moffitt Cancer Center

June 30, 2021

What you will learn to run

- Review of tables with janitor package
- Chi square test
- Correlation
- Two-sample tests (t-test, wilcoxon test)

Review of tables

```
janitor::tabyl(tcga, radiation_therapy, vital_status )
```

```
radiation_therapy Alive Dead
               NO   217   84
              YES   190   95
             <NA>   243  214
```

Review of tables

```
janitor::tabyl(tcga, radiation_therapy, vital_status,
               show_na = FALSE )
```

```
radiation_therapy Alive Dead
               NO   217   84
              YES   190   95
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE )
```

```
smoking FEMALE MALE
Current     40  151
 Former     54  180
  Never     72   88
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col"))
```

```
smoking FEMALE MALE Total
Current     40  151   191
 Former     54  180   234
  Never     72   88   160
  Total    166  419   585
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col")) %>%
  adorn_percentages(denominator = "col")
```

```
smoking    FEMALE      MALE     Total
Current 0.2409639 0.3603819 0.3264957
 Former 0.3253012 0.4295943 0.4000000
  Never 0.4337349 0.2100239 0.2735043
  Total 1.0000000 1.0000000 1.0000000
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col")) %>%
  adorn_percentages(denominator = "col") %>%
  adorn_pct_formatting(digits = 0)
```

```
smoking FEMALE MALE Total
Current    24%  36%   33%
 Former    33%  43%   40%
  Never    43%  21%   27%
  Total   100% 100%  100%
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col")) %>%
  adorn_percentages(denominator = "col") %>%
  adorn_pct_formatting(digits = 0) %>%
  adorn_ns(position = "front")
```

smoking	FEMALE		MALE		Total	
Current	40	(24%)	151	(36%)	191	(33%)
Former	54	(33%)	180	(43%)	234	(40%)
Never	72	(43%)	88	(21%)	160	(27%)
Total	166	(100%)	419	(100%)	585	(100%)

Pearson's chi-squared test

The chi squared test is a non-parametric test that can be applied to contingency tables with various dimensions. The name of the test originates from the chi-squared distribution, which is the distribution for the squares of independent standard normal variables. This is the distribution of the test statistic of the chi squared test, which is defined by the sum of chi-square values for all cells arising from the difference between a cell's observed value and the expected value, normalized by the expected value.

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- χ^2 = chi square statistic
- O_{ij} = observed value
- E_{ij} = expected value

Pearson’s chi-squared test

The null hypothesis of the Chi-Square test is that no relationship exists between the categorical variables in the population; they are independent.

smoking	FEMALE		MALE		Total	
Current	40	(24%)	151	(36%)	191	(33%)
Former	54	(33%)	180	(43%)	234	(40%)
Never	72	(43%)	88	(21%)	160	(27%)
Total	166	(100%)	419	(100%)	585	(100%)

Pearson’s chi-squared test

The null hypothesis of the Chi-Square test is that no relationship exists between the categorical variables in the population; they are independent.

```
tcga %>% tabyl( smoking, gender , show_na = FALSE ) %>%
  chisq.test()
```

```
    Pearson's Chi-squared test

data:  .
X-squared = 30.182, df = 2, p-value = 0.0000002793
```

The p-value is very low so we reject the null hypothesis that there is no association between the variables.

Pearson's chi-squared test

More women are never smokers and more men are current smokers.

R Code | Plot

```
ggplot(tcga %>% filter(!is.na(smoking)),
       aes(x = gender, fill = smoking )) +
  geom_bar(position = "fill") +
  labs(y = "proportion")
```

Pearson’s chi-squared test

More women are never smokers and more men are current smokers.

R Code | Plot

Correlation

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables.

Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics. They all assume values in the range from −1 to +1, where ±1 indicates the strongest possible agreement and 0 no agreement.

Pearson

Spearman

Pearson Correlation

The Pearson product-moment correlation coefficient, also known as r or Pearson's r, is a measure of the strength and direction of the linear relationship between two variables that is defined as the covariance of the variables divided by the product of their standard deviations. Pearson's r is the best-known and most commonly used type of correlation coefficient. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{(\sum_{i=1}^{n}(x_i-\overline{x})^2)}\sqrt{(\sum_{i=1}^{n}(y_i-\overline{y})^2)}}$$

- n is the sample size
- x_i and y_i are the individual sample points indexed with i
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the smaple mean of x (similarly y)

Pearson Correlation

```
r1 <- tcga %>% correlation::cor_test("DUOXA1_exp",
                                     "DUOX1_exp",
                                     method = c("pearson") )

r2 <- tcga %>% correlation::cor_test("BUB1_exp",
                                     "C10orf32_exp",
                                     method = c("pearson") )

r3 <- tcga %>% correlation::cor_test("BRAF_exp",
                                     "DTL_exp",
                                     method = c("pearson") )
```

Pearson Correlation

```
knitr::kable(bind_rows(r1,r2,r3 ), format = 'html', digits = 3) %>%
 kable_styling(font_size = 12)
```

Parameter1	Parameter2	r	CI	CI_low	CI_high	t	df_error	p	Method	n_Obs
DUOXA1_exp	DUOX1_exp	0.918	0.95	0.908	0.927	74.789	1041	0.000	Pearson	1043
BUB1_exp	C10orf32_exp	-0.600	0.95	-0.638	-0.560	-24.227	1041	0.000	Pearson	1043
BRAF_exp	DTL_exp	0.020	0.95	-0.041	0.080	0.642	1041	0.521	Pearson	1043

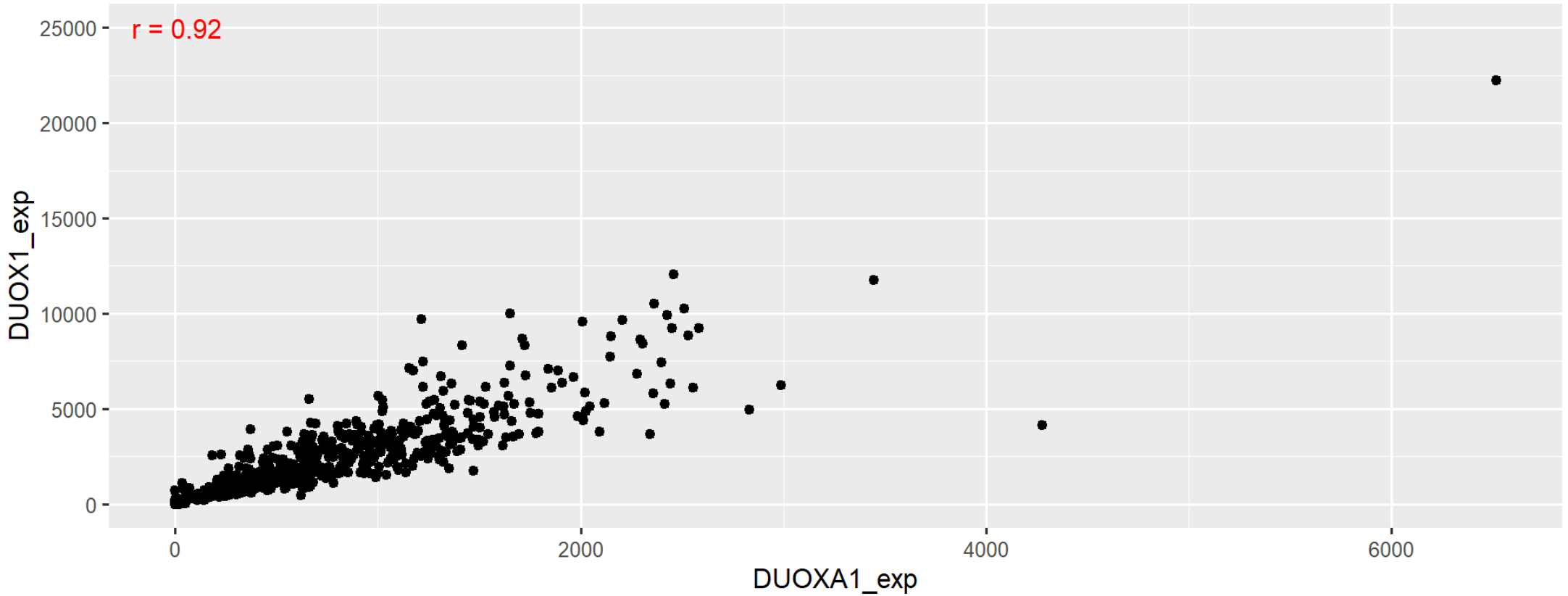
Pearson Correlation Scatter Plot

R Code | Plot

```
ggplot(tcga) +
  aes(DUOXA1_exp, DUOX1_exp) +
  geom_point() +
  annotate(geom = "text", x = 10, y = 25000,
           label = paste("r = ", round(r1$r, 2), sep = ""),
           color = "red")
```

Pearson Correlation Scatter Plot

R Code | Plot



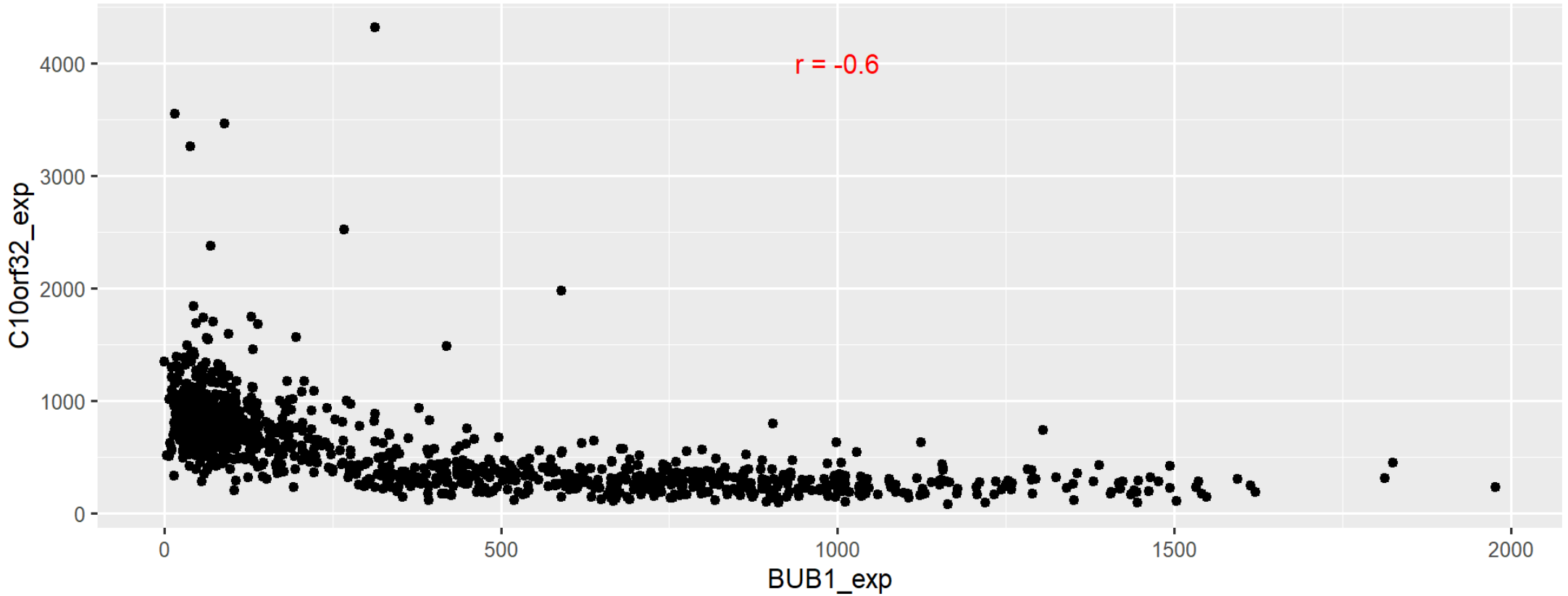
Pearson Correlation Scatter Plot

R Code | Plot

```
ggplot(tcga) +
  aes(BUB1_exp, C10orf32_exp) +
  geom_point() +
  annotate(geom = "text", x = 1000, y = 4000,
           label = paste("r = ", round(r2$r, 2), sep = ""),
           color = "red")
```

Pearson Correlation Scatter Plot

R Code | Plot



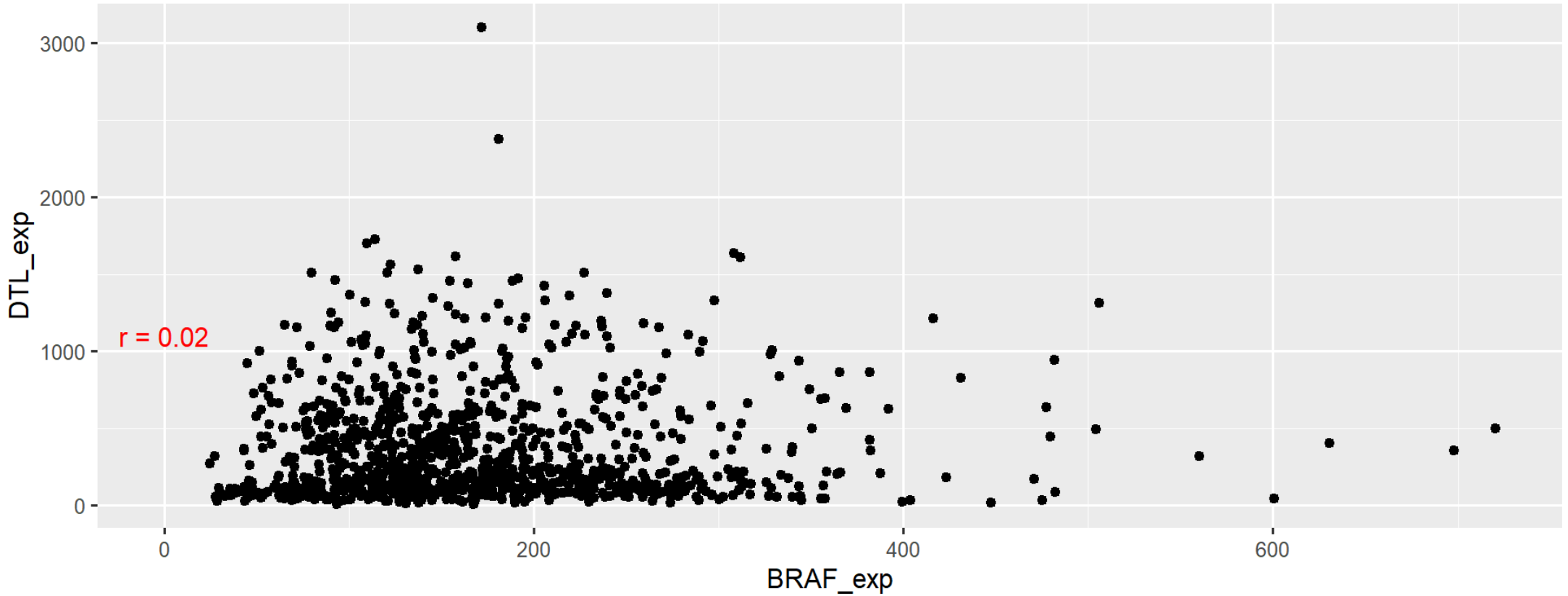
Pearson Correlation Scatter Plot

R Code | Plot

```
ggplot(tcga) +
  aes(BRAF_exp, DTL_exp) +
  geom_point() +
  annotate(geom = "text", x = 0, y = 1100,
           label = paste("r = ", round(r3$r, 2), sep = ""),
           color = "red")
```

Pearson Correlation Scatter Plot

R Code | Plot



Correllation Matrix

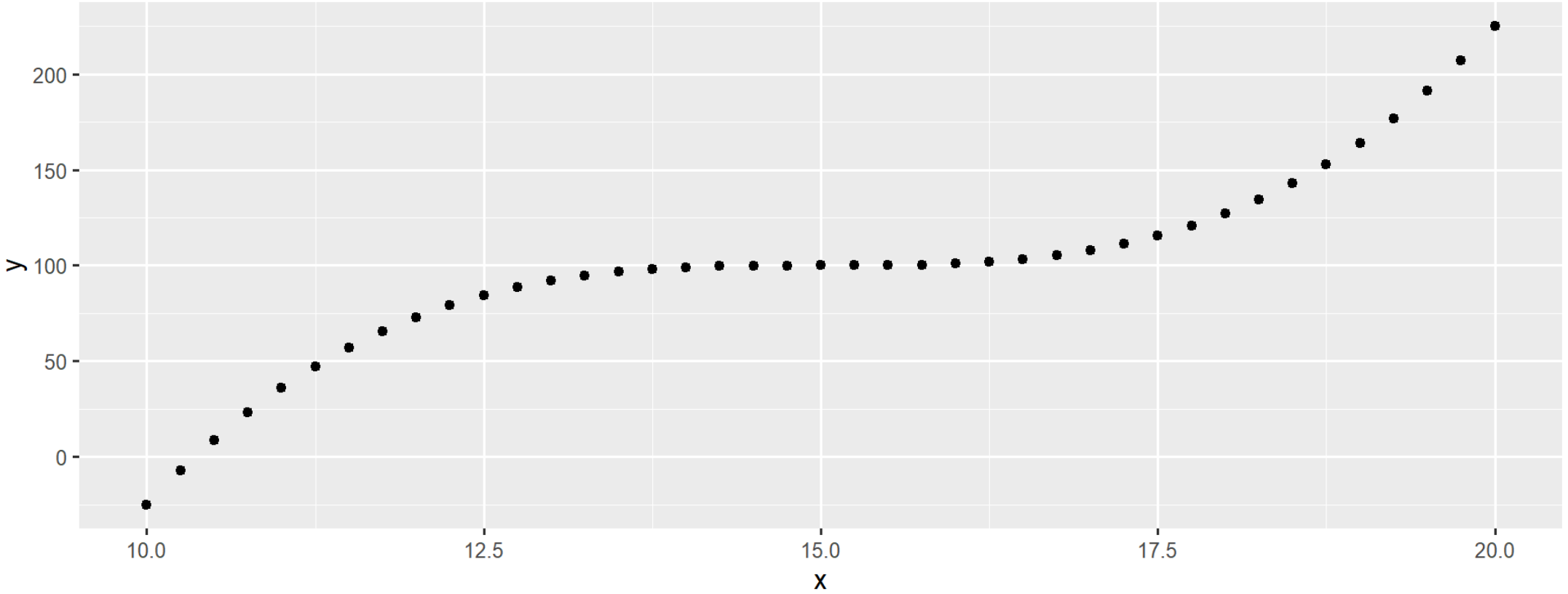
```
tcga %>%
  select(DUOXA1_exp, DUOX1_exp, BUB1_exp, BRAF_exp, DTL_exp ) %>%
  cor() %>% round(.,2)
```

	DUOXA1_exp	DUOX1_exp	BUB1_exp	BRAF_exp	DTL_exp
DUOXA1_exp	1.00	0.92	0.40	-0.07	0.30
DUOX1_exp	0.92	1.00	0.38	-0.05	0.30
BUB1_exp	0.40	0.38	1.00	0.01	0.78
BRAF_exp	-0.07	-0.05	0.01	1.00	0.02
DTL_exp	0.30	0.30	0.78	0.02	1.00

Spearman Correlation

The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. It is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

Spearman Correlation example of monotonic relationship



Spearman Correlation

```
r1 <- tcga %>% correlation::cor_test("DUOXA1_exp",
                                     "DUOX1_exp",
                                     method = c("spearman") )

r2 <- tcga %>% correlation::cor_test("BUB1_exp",
                                     "C10orf32_exp",
                                     method = c("spearman") )

r3 <- tcga %>% correlation::cor_test("BRAF_exp",
                                     "DTL_exp",
                                     method = c("spearman") )
```

Spearman Correlation

```
knitr::kable(bind_rows(r1, r2, r3), format = 'html', digits = 3) %>%
 kable_styling(font_size = 12)
```

Parameter1	Parameter2	rho	CI	CI_low	CI_high	S	p	Method	n_Obs
DUOXA1_exp	DUOX1_exp	0.905	0.95	0.892	0.915	18053045	0.00	Spearman	1043
BUB1_exp	C10orf32_exp	-0.796	0.95	-0.818	-0.772	339618298	0.00	Spearman	1043
BRAF_exp	DTL_exp	0.015	0.95	-0.048	0.077	186283211	0.63	Spearman	1043

Two sample tests

- T - test
- Mann Whitney U Test (Wilcoxon Rank Sum Test)

T-test

A t-test is a method used to determine if there is a significant difference between the means of two groups based on a sample of data.

The test relies on a set of assumptions for it to be interpreted properly and with validity. Among these assumptions, the data must be randomly sampled from the population of interest and that the data variables follow a normal distribution.

T-test assumptions

1. The data are continuous (not discrete).

2. The data follow the normal probability distribution.

3. The variances of the two populations are equal. (If not, the Aspin-Welch Unequal-Variance test is used.)

4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired t-test).

5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

T-test formula

Test Statistic (equal variances):

$$T = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{(1/N_1 + 1/N_2)}}$$

where: $S_p^2 = \frac{(N_1-1)s_1^2 + (N_1-1)s_1^2}{N_1+N_2-2}$

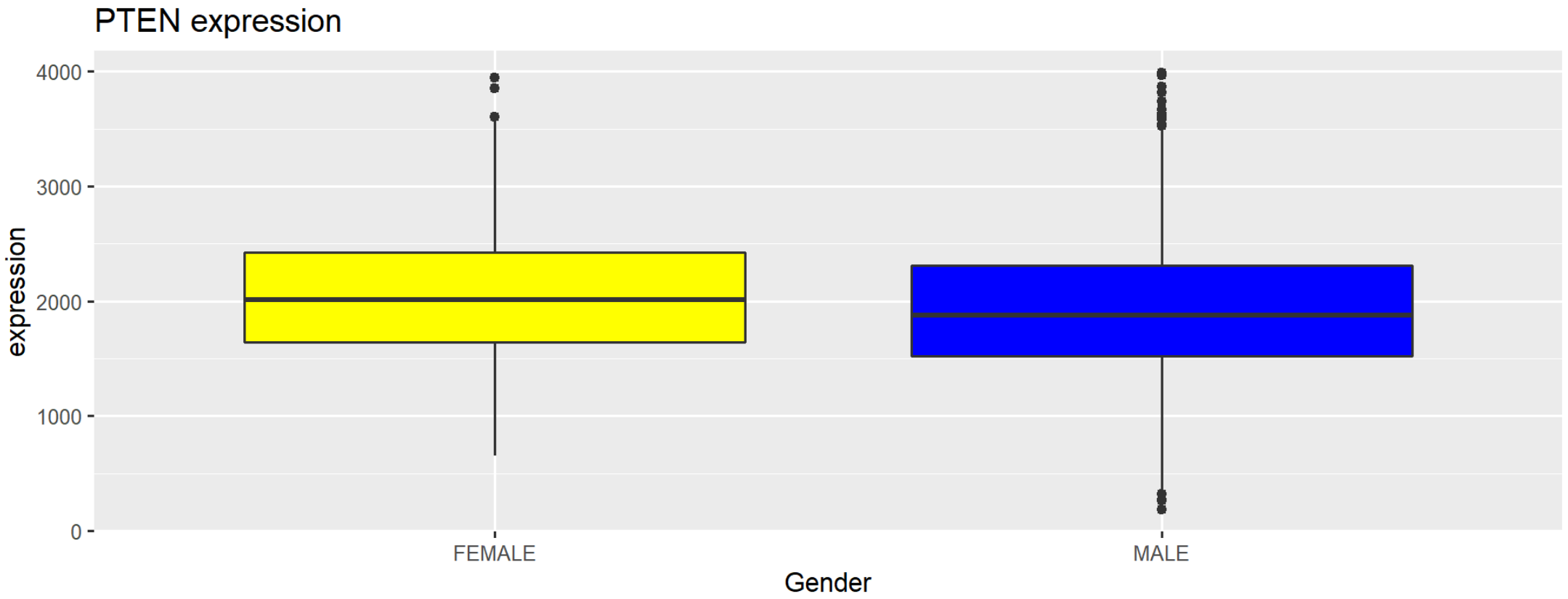
T-test visulize boxplots

R Code | Plot

```
ggplot(tcga , aes(gender, PTEN_exp, fill = gender))  +
  geom_boxplot() +
  scale_fill_manual( values = c("yellow", "blue")) +
  theme(legend.position = "none") +
  labs(title = "PTEN expression" , x = "Gender", y = "expression")
```

T-test visulize boxplots

R Code | Plot



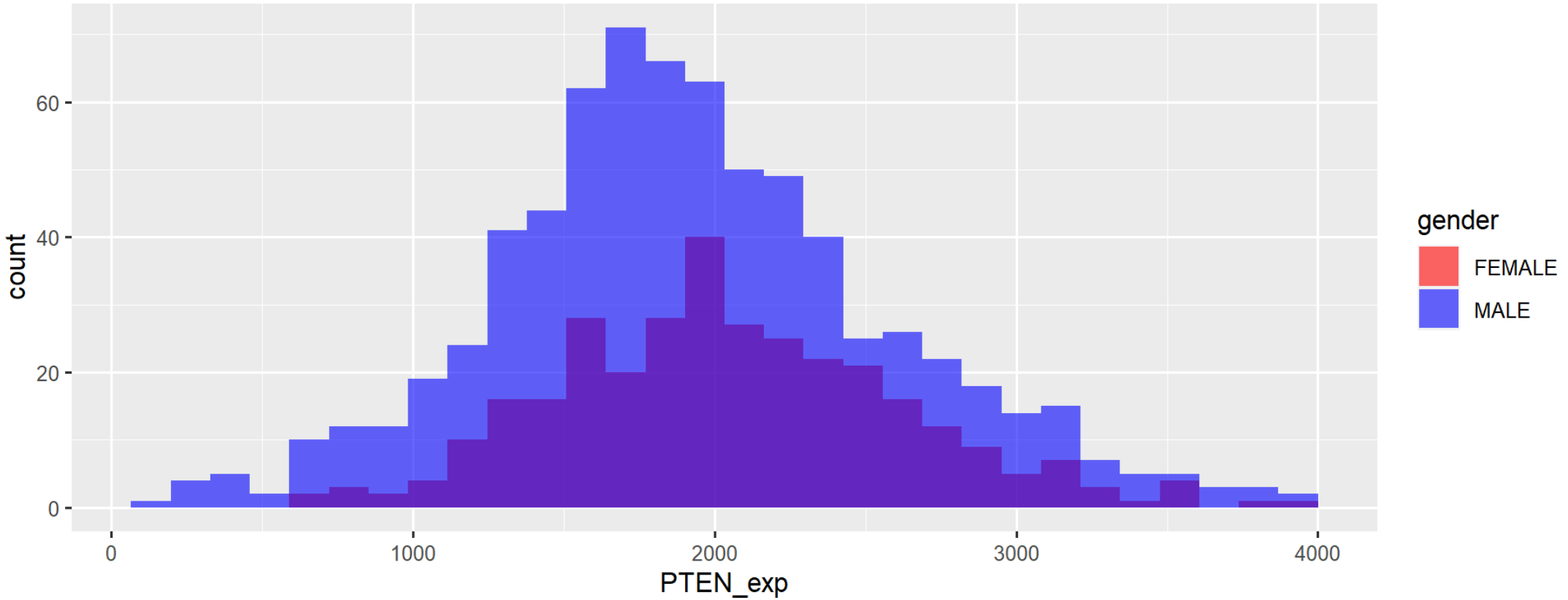
T-test Visualize data Histograms

R Code | Plot

```
ggplot(tcga, aes(x=PTEN_exp, fill=gender)) +
scale_fill_manual( values = c("red", "blue")) +
geom_histogram( alpha=0.6, position="identity")
```

T-test Visualize data Histograms

R Code | Plot



T-test R call

```
t.test( PTEN_exp ~ gender, data = tcga,
        alternative = c("two.sided"))
```

```
    Welch Two Sample t-test

data:  PTEN_exp by gender
t = 3.1297, df = 702.6, p-value = 0.001822
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  46.83934 204.52754
sample estimates:
mean in group FEMALE   mean in group MALE
            2057.036             1931.353
```

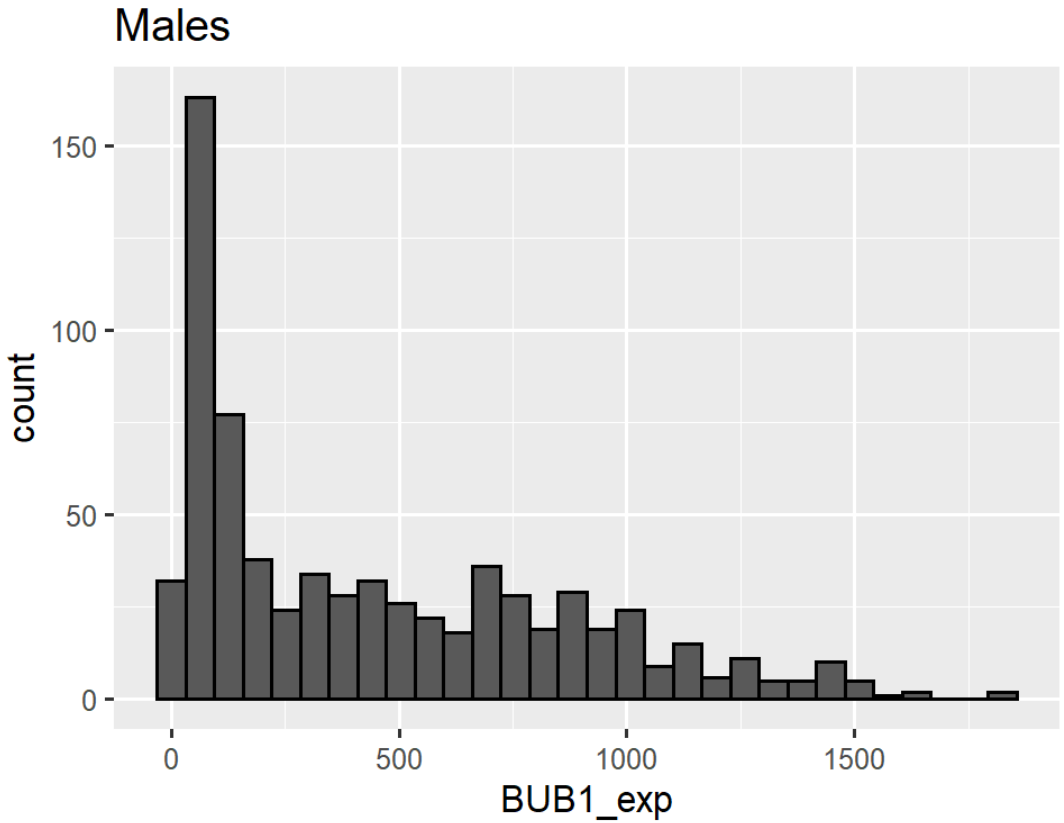
Mann Whitney U Test (Wilcoxon Rank Sum Test)

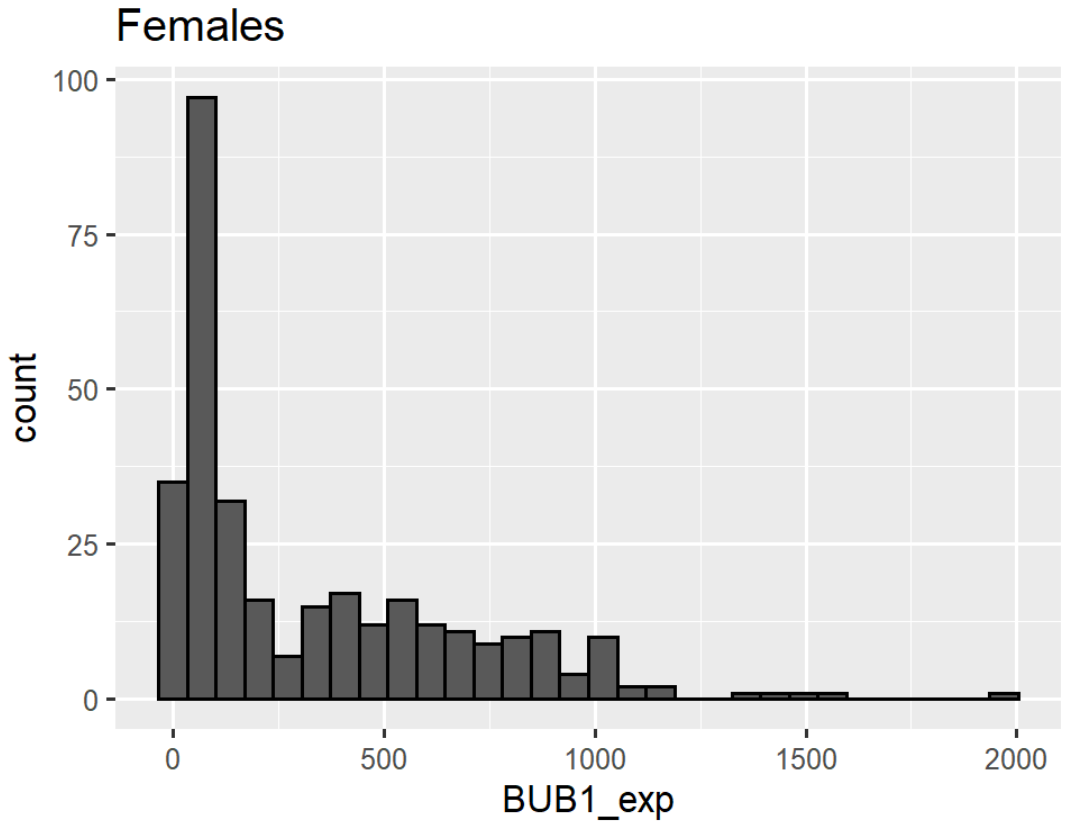
A popular non-parametric test to compare outcomes between two independent groups is the Mann Whitney U test. The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape).

It can also be used on related samples or matched samples to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test when the distribution of the difference between two samples' means cannot be assumed to be normally distributed.

Wilcoxon test: Histograms

Plot R Code





Wilcoxon test: Histograms

Plot | R Code

```
mplot <- ggplot(tcga %>% filter(gender == "MALE"),
                aes(x = BUB1_exp )) +
  geom_histogram(  colour = "black", position = "dodge") +
  ggtitle("Males")

wplot <- ggplot(tcga  %>% filter(gender == "FEMALE"),
                aes(x = BUB1_exp)) +
  geom_histogram(  colour= "black", position = "dodge")  +
  ggtitle("Females")

grid.arrange(mplot,wplot, ncol=2)
```

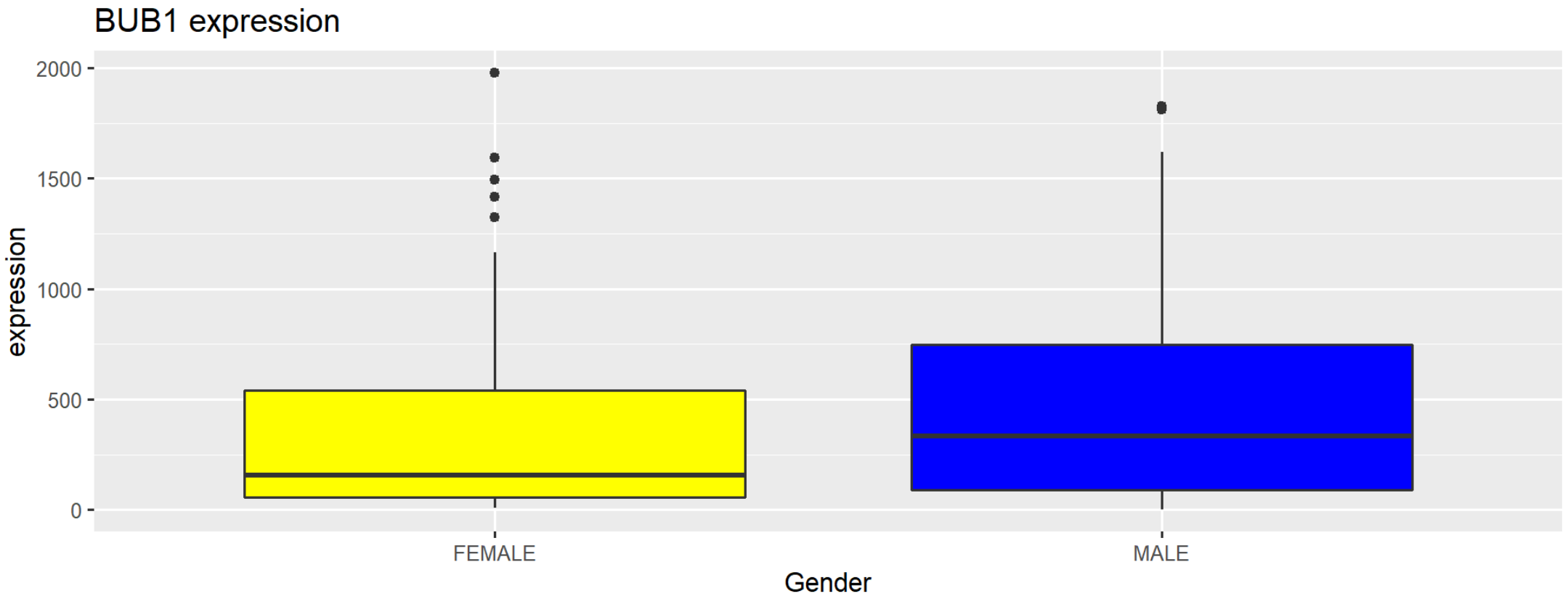
Wilcoxon test visulize boxplots gender

R Code | Plot

```
ggplot(tcga , aes(gender, BUB1_exp, fill = gender))  +
  geom_boxplot() +
  scale_fill_manual( values = c("yellow", "blue")) +
  theme(legend.position = "none") +
  labs(title = "BUB1 expression" , x = "Gender", y = "expression")
```

Wilcoxon test visulize boxplots gender

R Code | Plot



Wilcoxon test gender

```
wilcox.test(BUB1_exp ~ gender, data = tcga,
                    alternative = c("two.sided"))
```

```
    Wilcoxon rank sum test with continuity correction

data:  BUB1_exp by gender
W = 93990, p-value = 0.0000007221
alternative hypothesis: true location shift is not equal to 0
```

The p-value is very low so we reject the null hypothesis of no difference (0 location shift).

Thank you!

- The end