STATISTICS: UNIVARIATE, CONTINUOUS

- Graphs via `univariate`, `boxplot`, `gchart` (barchart), `gplot` (scatterplot), `sgplot` (modern alternative)
- t-test via `ttest`

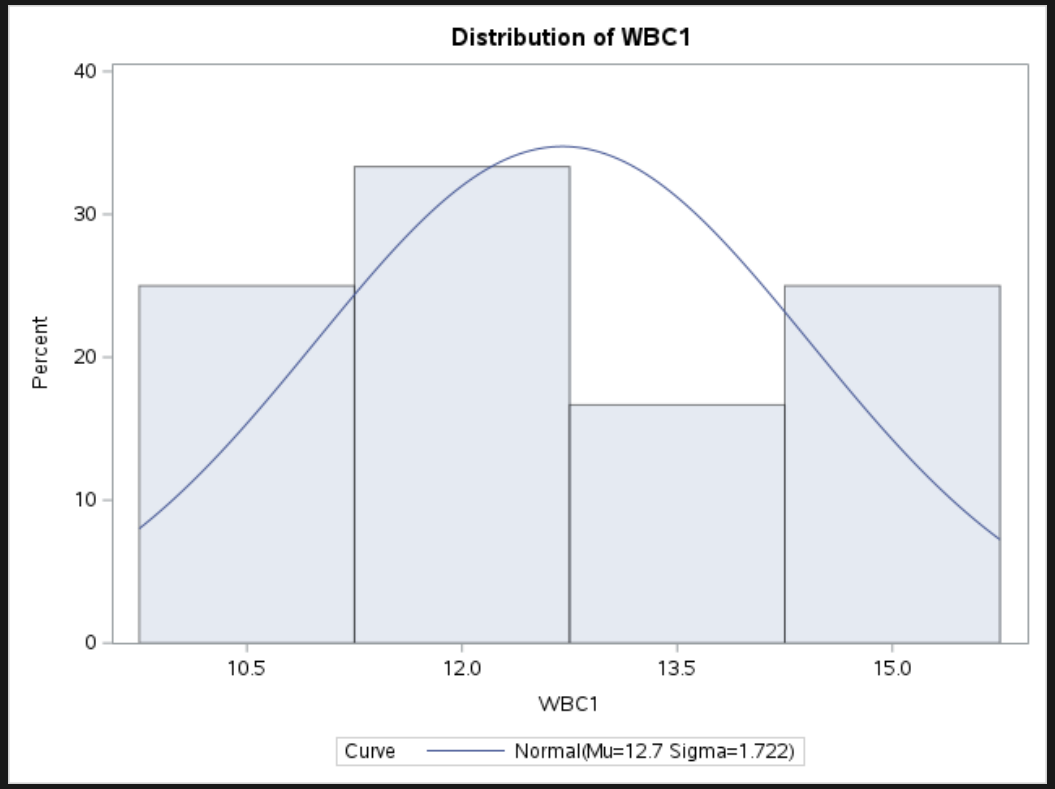
Read in white-blood-cell dataset:

```
filename myurl url "https://manctsui.github.io/datasets/WBC.xl
proc import
    datafile=myurl out=wbc dbms=xlsx replace;
    getnames = yes;
run;
```

Obs	Name	State	Bday	Hght	Wght	Sex	SexC	Educ	WBC1	WBC2	WBC3	WBC4
1	Ruth	DE	10/07/75	60	110	1	F	3	12.0	13.8	13.2	11.2
2	Jim	NJ	07/22/01	79	184	0	M	6	14.8	.	14.1	10.9
3	Mary	PA	09/03/74	70	140	1	F	8	10.5	14.5	12.7	14.6
4	Ty	DE	07/06/72	67	142	0	M	.	14.2	12.5	11.4	12.5
5	Moe	NJ	08/12/02	70	300	0	M	2	14.7	10.1	11.8	14.7
6	Ada	NJ	11/28/79	63	120	1	F	3	14.6	10.7	14.7	13.4
7	Al	NJ	08/18/57	61	190	0	M	7	12.5	10.3	14.5	14.9
8	Jill	NJ	10/07/60	61	135	1	F	2	10.0	13.2	13.8	11.2
9	Kate	DE	06/20/00	65	147	1	F	6	12.7	12.3	13.3	15.0
10	Tim	PA	08/19/99	68	225	0	M	2	11.1	10.4	14.1	13.5
11	Joe	PA	11/10/93	72	180	0	M	1	11.4	13.1	11.8	10.3
12	Anna	PA	05/26/62	59	119	1	F	6	13.9	12.5	.	12.7

HISTOGRAM WITH NORMAL CURVE OVERLAID

```
PROC UNIVARIATE DATA = wbc NOPRINT;
    HISTOGRAM wbc1 / NORMAL;
RUN;
```


Distribution of WBC1
Percent
40
30
20
10
0
10.5
12.0
13.5
15.0
WBC1
Curve
Normal(Mu=12.7 Sigma=1.722)

The UNIVARIATE Procedure
Fitted Normal Distribution for WBC1 (WBC1)

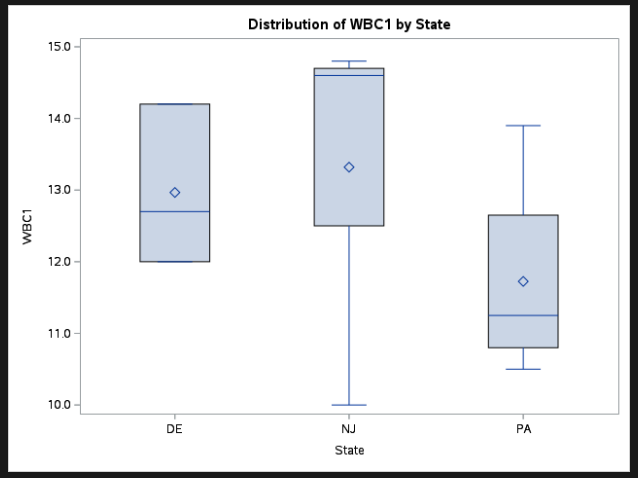
Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	12.7
Std Dev	Sigma	1.72205

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.17371648	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05373622	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.36732902	Pr > A-Sq	>0.250

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	10.00000	8.69391
5.0	10.00000	9.86748
10.0	10.50000	10.49310
25.0	11.25000	11.53850
50.0	12.60000	12.70000
75.0	14.40000	13.86150
90.0	14.70000	14.90690
95.0	14.80000	15.53252
99.0	14.80000	16.70609

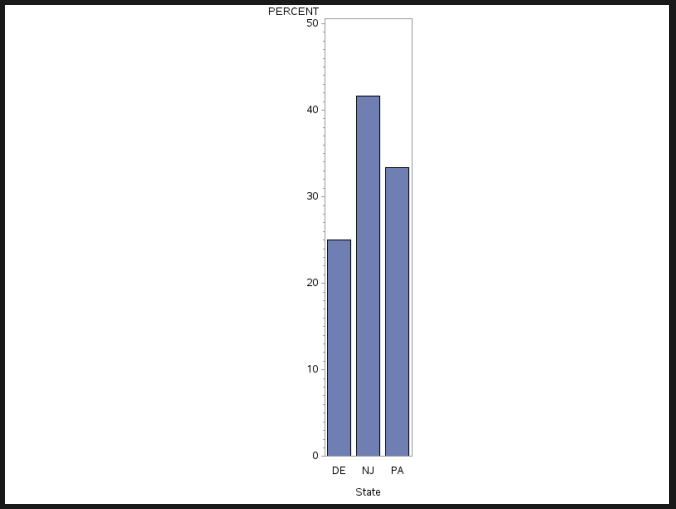
BOXPLOT

```
/* First sort by grouping variable. Then PROC BOXPLOT. */
PROC SORT DATA = wbc;
    by state;
RUN;
PROC BOXPLOT DATA = wbc;
    /* Choose "schematic" style for the modified boxplot. */
    PLOT wbc1*state / boxstyle=schematic;
RUN;
```



BAR CHART

```
/* Use HBAR for horizontal bar chart. */
PROC GCHART DATA=wbc;
    VBAR state / TYPE=PERCENT;
RUN;
```

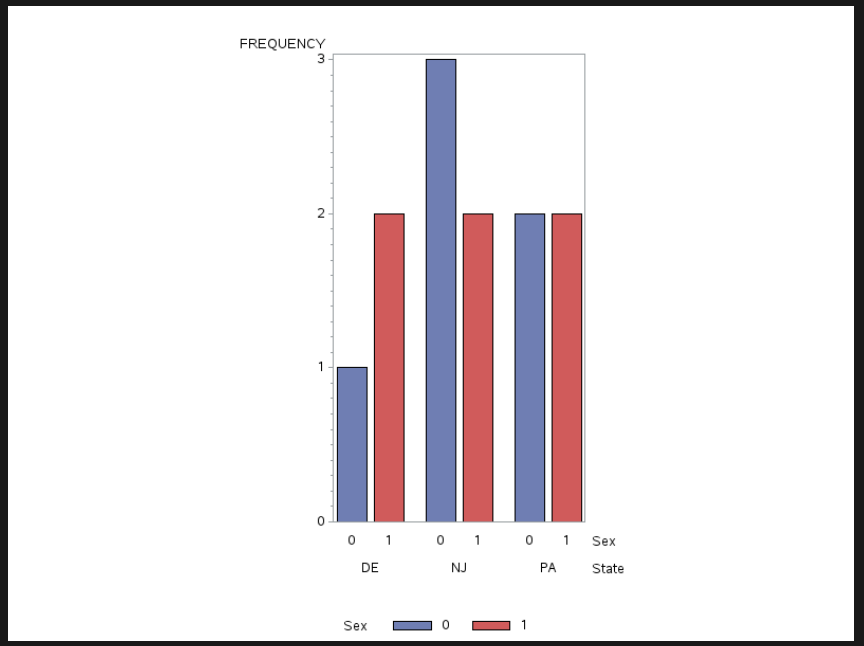


BAR CHART

```
/* Bar chart on numerical variable: use DISCRETE */
PROC GCHART DATA=wbc;
    VBAR sex/ DISCRETE TYPE=PERCENT;
RUN;
```

BAR CHART

```
/* Grouped by US State. */
PROC GCHART DATA=wbc;
    VBAR sex/ discrete
        group=state subgroup=sex;
RUN;
```


FREQUENCY
3
2
1
0
0 1 0 1 0 1 Sex
DE NJ PA State
Sex 0 1

SGPLOT is a modern alternative:

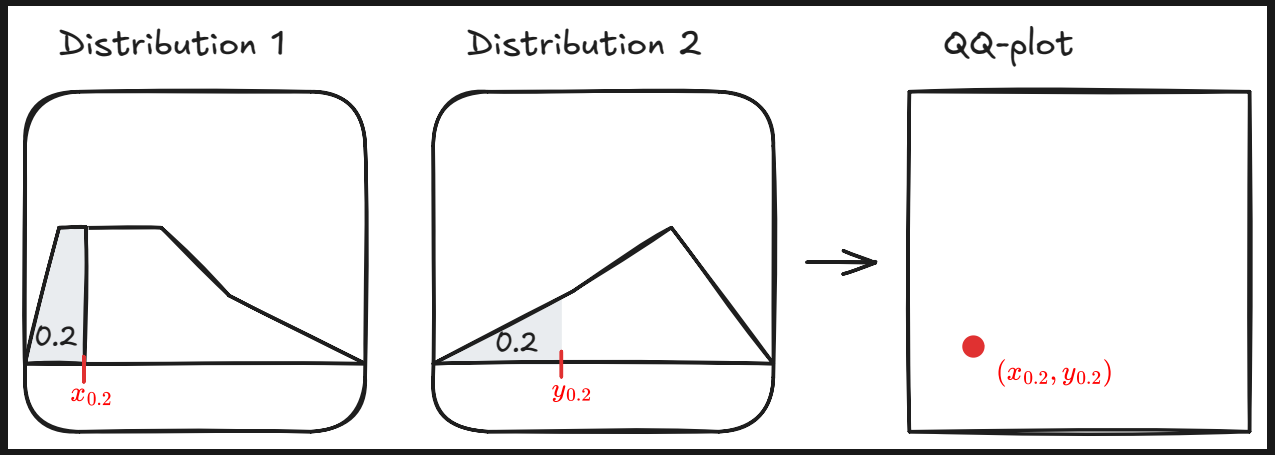
```
proc sgplot data=wbc; /* Histogram */
    histogram wbc1;
    density wbc1 / type=normal;

proc sgplot data=wbc; /* Vertical boxplot */
    vbox wbc1 / category=state;

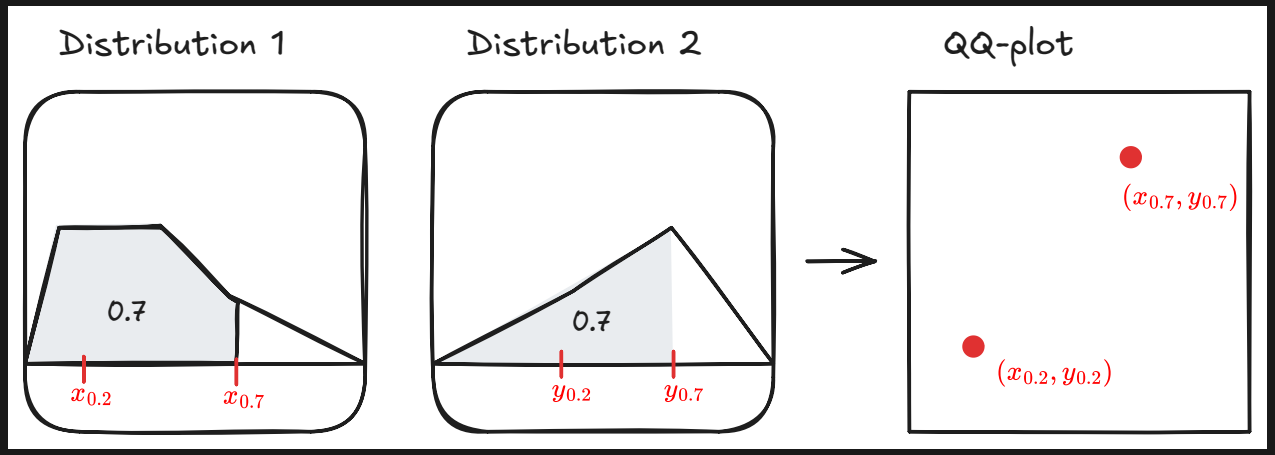
proc sgplot data=wbc; /* Vertical bar chart, optionally groupe
    vbar state / group=sex groupdisplay=cluster;

proc sgplot data=wbc; /* Scatterplot */
    scatter x=wbc1 y=wbc2;
```

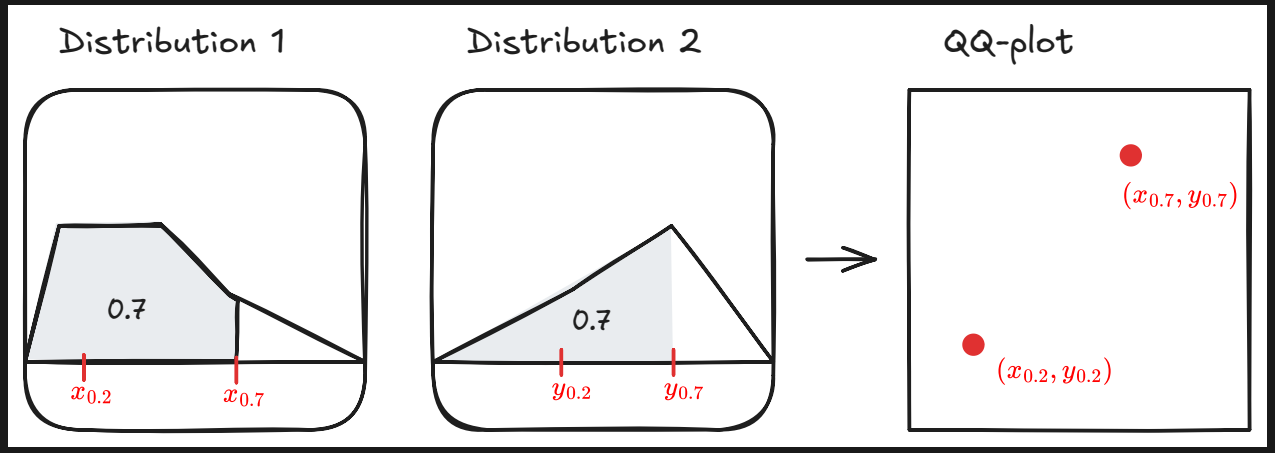
QUANTILE-QUANTILE PLOT (QQPLOT)



QUANTILE-QUANTILE PLOT (QQPLOT)

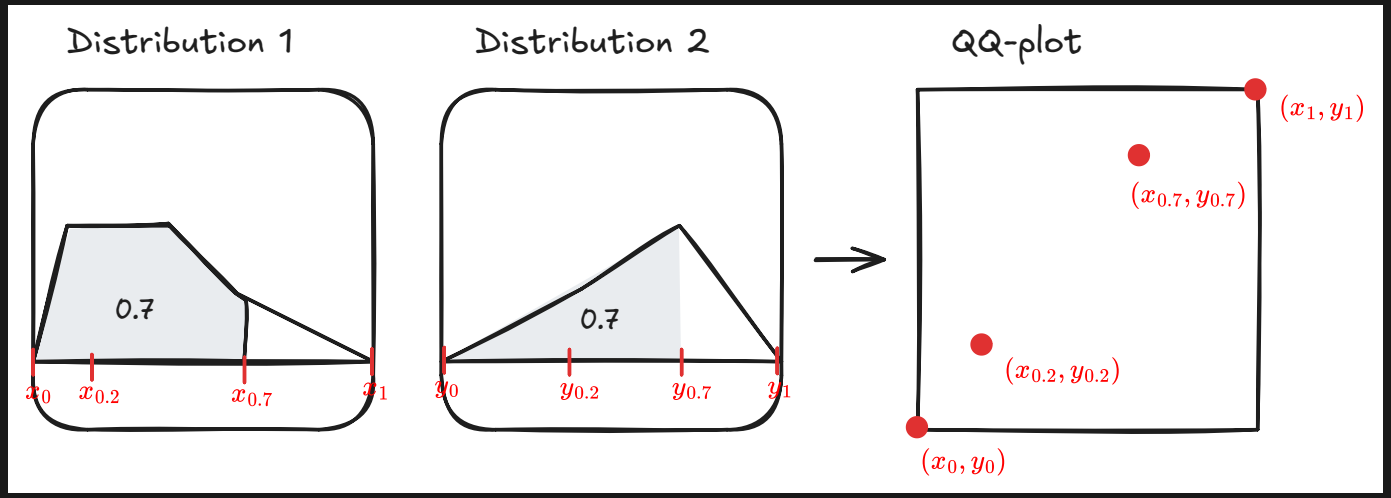


QUANTILE-QUANTILE PLOT (QQPLOT)



QQ-plot is parametrized by the quantile aka left area under each distribution.

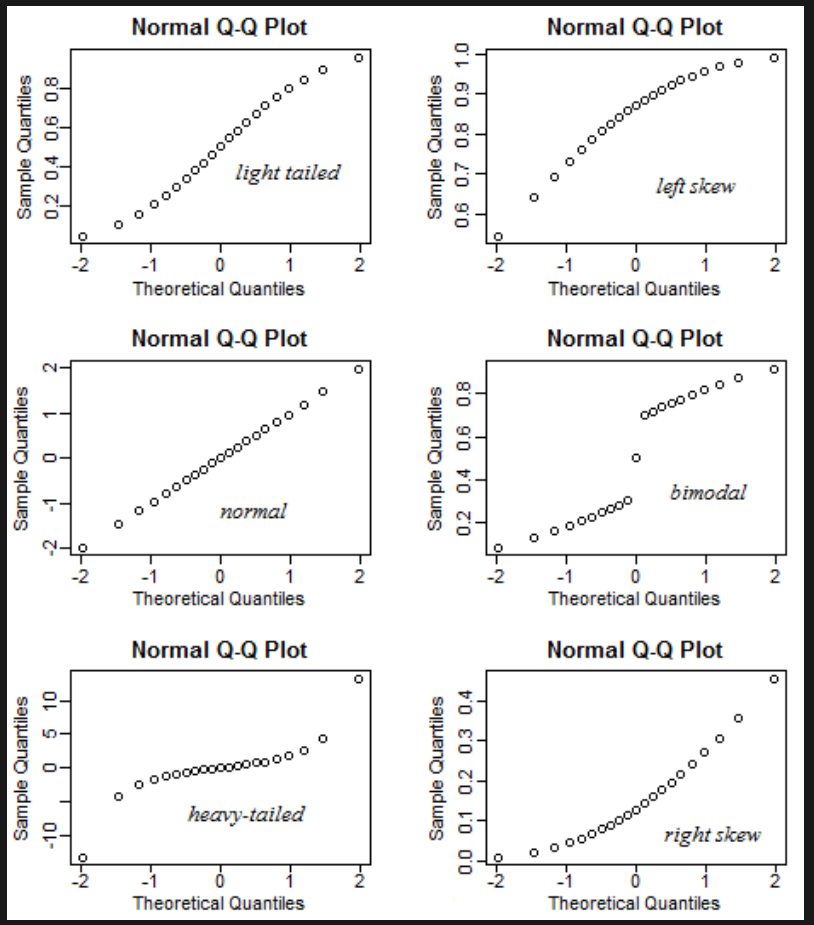
QUANTILE-QUANTILE PLOT (QQPLOT)



Plot is often stretched & translated so (x_0, y_0) & (x_1, y_1) are at square's corners.

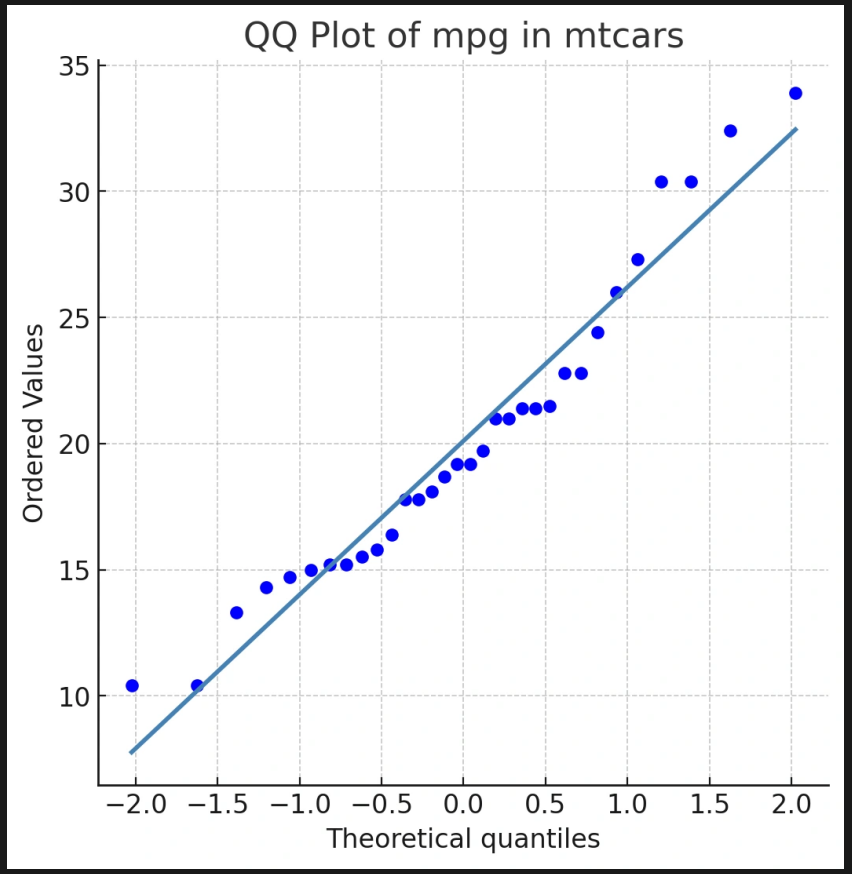
QUANTILE-QUANTILE PLOT (QQPLOT)

To assess for normality, we compare to a normal distribution, e.g. $N(0, 1)$.

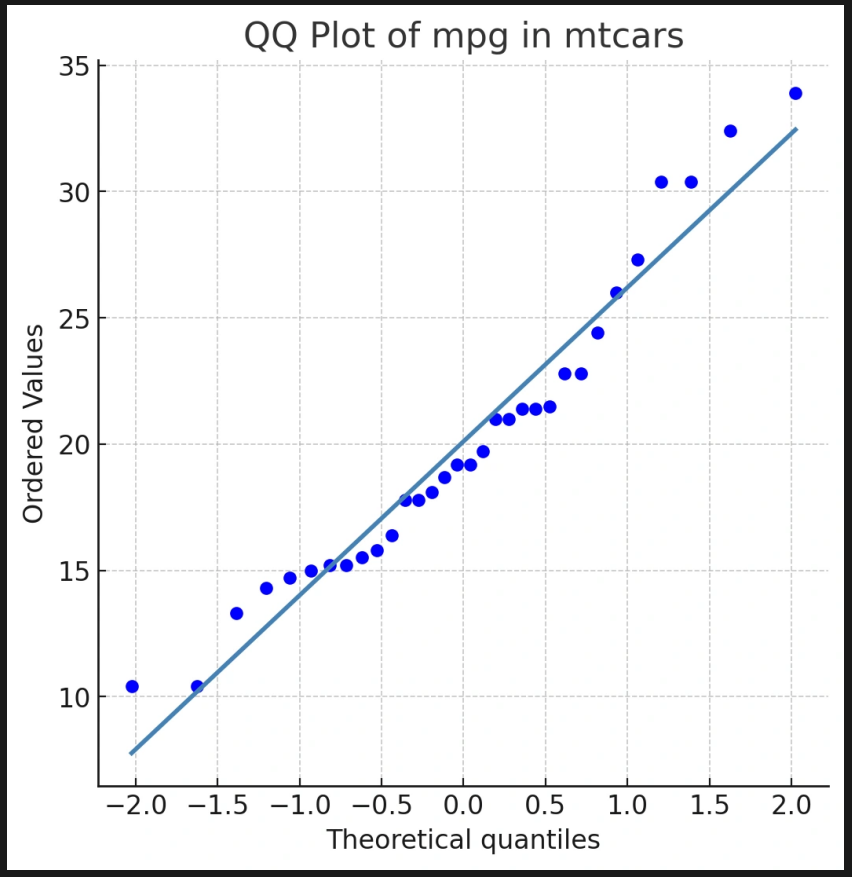


Stack Exchange

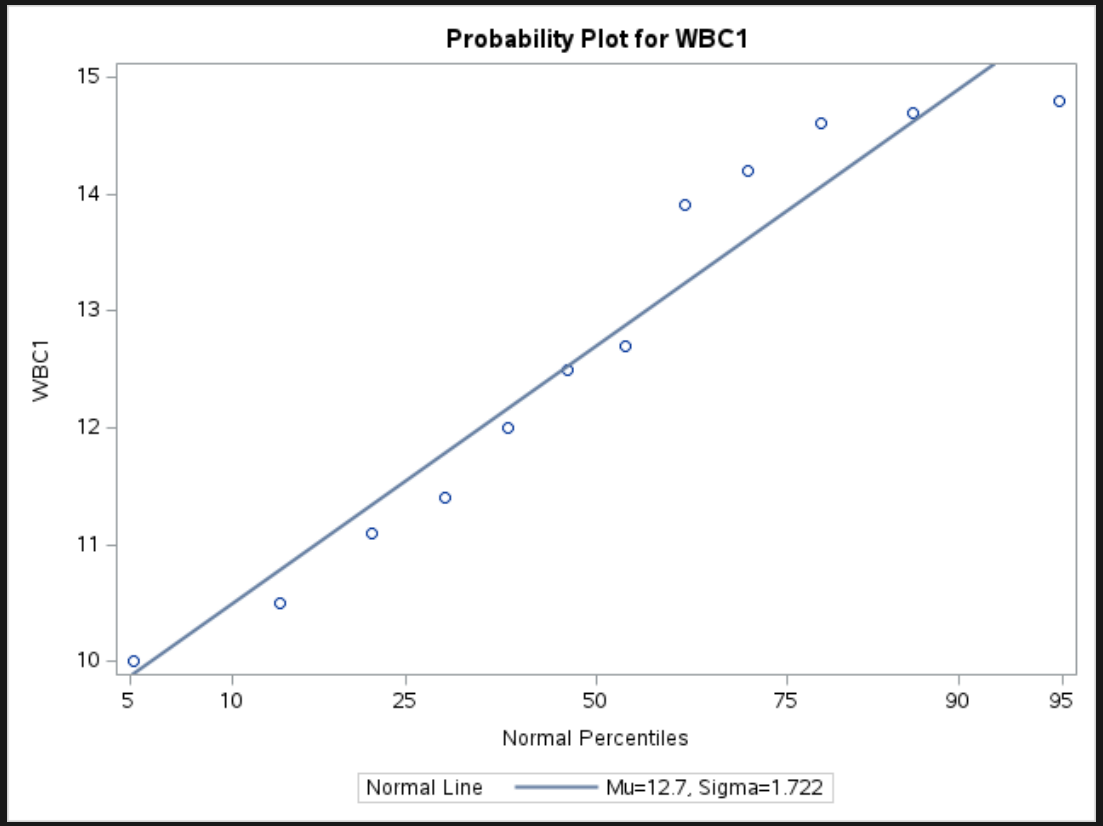
```r
# Example in R
qqnorm(mtcars$mpg)
qqline(mtcars$mpg)
```



```r
# Example in R: Fairly normal.
# Systematic pattern so normal distribution is not best model.
qqnorm(mtcars$mpg)
qqline(mtcars$mpg)
```



```
/* In SAS */
PROC UNIVARIATE DATA = wbc NOPRINT;
    PROBPLOT wbc1/NORMAL (MU=est SIGMA=est);
RUN;
```



We consider a distribution roughly normal if...

- Histogram is visually close to overlaid normal curve.
- (Modified) boxplot is symmetric without many outliers.
- Q-Q plot is straight line without systematic deviations.

z-TEST

To test if a population mean μ is some claimed number μ_0, and we do know the population standard deviation σ, we compute the z-statistics $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ which is distributed like $N(0, 1)$.

This assumes $\bar{x}$ follows a normal distribution. Phrased in terms of population distribution:

ASSUMPTIONS TO CHECK BEFORE USING z-TEST

- Data is interval scale data (not merely categorical or ordinal)
- Data is a simple random sample. If sampling without replacement, ensure the population size N is at least 10x the size of the sample size n (10% Condition).
- Normal Pop./Large Sample Condition: Either population is normal or $n \geq 30$.

z-TEST FOR MEANS

- **State hypothesis:** $H_0 : \mu = \mu_0, H_a : \mu \neq \mu_0$.
- **Check assumptions:** Sampling is SRS (10% condition if sampling without replacement),
- **Compute:** $P = P(X \geq \bar{x} \mid H_0 \text{ is } \mathrm{true})$ via area under the (normal) sampling distribution.
- **Conclude:** If $P < \alpha$, we reject H_0. If $P \geq \alpha$, we fail to reject H_0.

Ex. Car company claims their cars accelerate from 0 to 60 mph in less than 6 seconds. A random sample of 33 cars has a mean acceleration of 7.2 seconds. At `$\alpha = 0.05$` level, can you reject the claim? (Assume pop. standard deviation of 2.5 seconds.)

Ex. Car company claims their cars accelerate from 0 to 60 mph in less than 6 seconds. A random sample of 33 cars has a mean acceleration of 7.2 seconds. At `$\alpha = 0.05$` level, can you reject the claim? (Assume pop. standard deviation of 2.5 seconds.)

A. State hypotheses: $H_0 : \mu \leq 6, \quad H_a : \mu > 6$.

Ex. Car company claims their cars accelerate from 0 to 60 mph in less than 6 seconds. A random sample of 33 cars has a mean acceleration of 7.2 seconds. At `$\alpha = 0.05$` level, can you reject the claim? (Assume pop. standard deviation of 2.5 seconds.)

A. State hypotheses: $H_0 : \mu \leq 6, \quad H_a : \mu > 6$. Then check:

- 10% Condition: There are $N \geq 10 \cdot 33$ cars made by that company. OK.
- Normal Pop./Large Sample Condition: $n = 33 \geq 30$. OK.

Ex. Car company claims their cars accelerate from 0 to 60 mph in less than 6 seconds. A random sample of 33 cars has a mean acceleration of 7.2 seconds. At `$\alpha = 0.05$` level, can you reject the claim? (Assume pop. standard deviation of 2.5 seconds.)

A. State hypotheses: $H_0 : \mu \leq 6, \quad H_a : \mu > 6$. Then check:

- 10% Condition: There are $N \geq 10 \cdot 33$ cars made by that company. OK.
- Normal Pop./Large Sample Condition: $n = 33 \geq 30$. OK.

Compute: $P = 1 - \text{pnorm}(7.2,\ \text{mean} = 6, \text{sd} = \frac{2.5}{\sqrt{33}}) = 0.0029$.

Ex. Car company claims their cars accelerate from 0 to 60 mph in less than 6 seconds. A random sample of 33 cars has a mean acceleration of 7.2 seconds. At `$\alpha = 0.05$` level, can you reject the claim? (Assume pop. standard deviation of 2.5 seconds.)

A. State hypotheses: $H_0 : \mu \leq 6, \quad H_a : \mu > 6$. Then check:

- 10% Condition: There are $N \geq 10 \cdot 33$ cars made by that company. OK.
- Normal Pop./Large Sample Condition: $n = 33 \geq 30$. OK.

Compute: $P = 1 - \text{pnorm}(7.2,\ \text{mean} = 6, \text{sd} = \frac{2.5}{\sqrt{33}}) = 0.0029$.

Conclude: Since $P = 0.0029 < \alpha = 0.05$, we reject H_0. There is convincing evidence that the mean acceleration of that company's car is $>$ 6 seconds.

t-TEST

To test if a population mean μ is some claimed number μ_0, and we only have sample standard deviation s_x, we compute the t-statistic $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$ which follows the t-distribution with degree of freedom $df = n - 1$.

This assumes $\bar{x}$ follows a normal distribution. Phrased in terms of population distribution:

ASSUMPTIONS TO CHECK BEFORE USING t-TEST

- Data is interval scale data (not merely categorical or ordinal).
- Data is a simple random sample. If sampling without replacement, ensure the population size N is at least 10x the size of the sample size n (10% Condition).
- Population is approx. normally distributed if $n < 30$ and possibly moderately skewed if $n \geq 30$.

Example. Used car dealer claims the mean price of used cars sold in the past 12 months is at least \$21,000, but your SRS of 14 cars has a mean price of \$19,200 and standard deviation of \$3,000. Do you have enough evidence to reject the dealer's claim at `$\alpha = 0.05$` level? Assume the population is normally distributed.

Answer. State hypotheses: $H_0 : \mu \geq 21000, \quad H_a : \mu < 21000.$

Example. Used car dealer claims the mean price of used cars sold in the past 12 months is at least \$21,000, but your SRS of 14 cars has a mean price of \$19,200 and standard deviation of \$3,000. Do you have enough evidence to reject the dealer's claim at `$\alpha = 0.05$` level? Assume the population is normally distributed.

Answer. State hypotheses: $H_0 : \mu \geq 21000, \quad H_a : \mu < 21000.$

Check

- 10% Condition: There are $N \geq 10 \cdot 14 = 140$ used cars sold in last 12 months. OK.
- Normal/Large Sample Condition: Population is normal. OK.

Example. Used car dealer claims the mean price of used cars sold in the past 12 months is at least \$21,000, but your SRS of 14 cars has a mean price of \$19,200 and standard deviation of \$3,000. Do you have enough evidence to reject the dealer's claim at `$\alpha = 0.05$` level? Assume the population is normally distributed.

Answer. State hypotheses: $H_0 : \mu \geq 21000, \quad H_a : \mu < 21000.$

Check

- 10% Condition: There are $N \geq 10 \cdot 14 = 140$ used cars sold in last 12 months. OK.
- Normal/Large Sample Condition: Population is normal. OK.

Compute: $P = \operatorname{pt}\left(\dfrac{19200 - 21000}{3000/\sqrt{14}},\ \mathrm{df} = 13\right) = 0.021.$

Example. Used car dealer claims the mean price of used cars sold in the past 12 months is at least \$21,000, but your SRS of 14 cars has a mean price of \$19,200 and standard deviation of \$3,000. Do you have enough evidence to reject the dealer's claim at `$\alpha = 0.05$` level? Assume the population is normally distributed.

Answer. State hypotheses: $H_0 : \mu \geq 21000, \quad H_a : \mu < 21000.$

Check

- 10% Condition: There are $N \geq 10 \cdot 14 = 140$ used cars sold in last 12 months. OK.
- Normal/Large Sample Condition: Population is normal. OK.

Compute: $P = \text{pt}\left(\dfrac{19200 - 21000}{3000/\sqrt{14}},\ \text{df} = 13\right) = 0.021.$

Conclude: Since $P = 0.021 < \alpha = 0.05$, we reject H_0. There is convincing evidence that mean price of used cars sold in the last 12 months is $<$ \$21,000.

t-test in SAS:

```
proc ttest data=wbc H0=10 alpha=0.05;
var wbc1;
run;
```

The TTEST Procedure

Variable: WBC1 (WBC1)

N	Mean	Std Dev	Std Err	Minimum	Maximum
12	12.7000	1.7220	0.4971	10.0000	14.8000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12.7000	11.6059	13.7941	1.7220	1.2199	2.9238

DF	t Value	Pr > |t|
11	5.43	0.0002

The TTEST Procedure

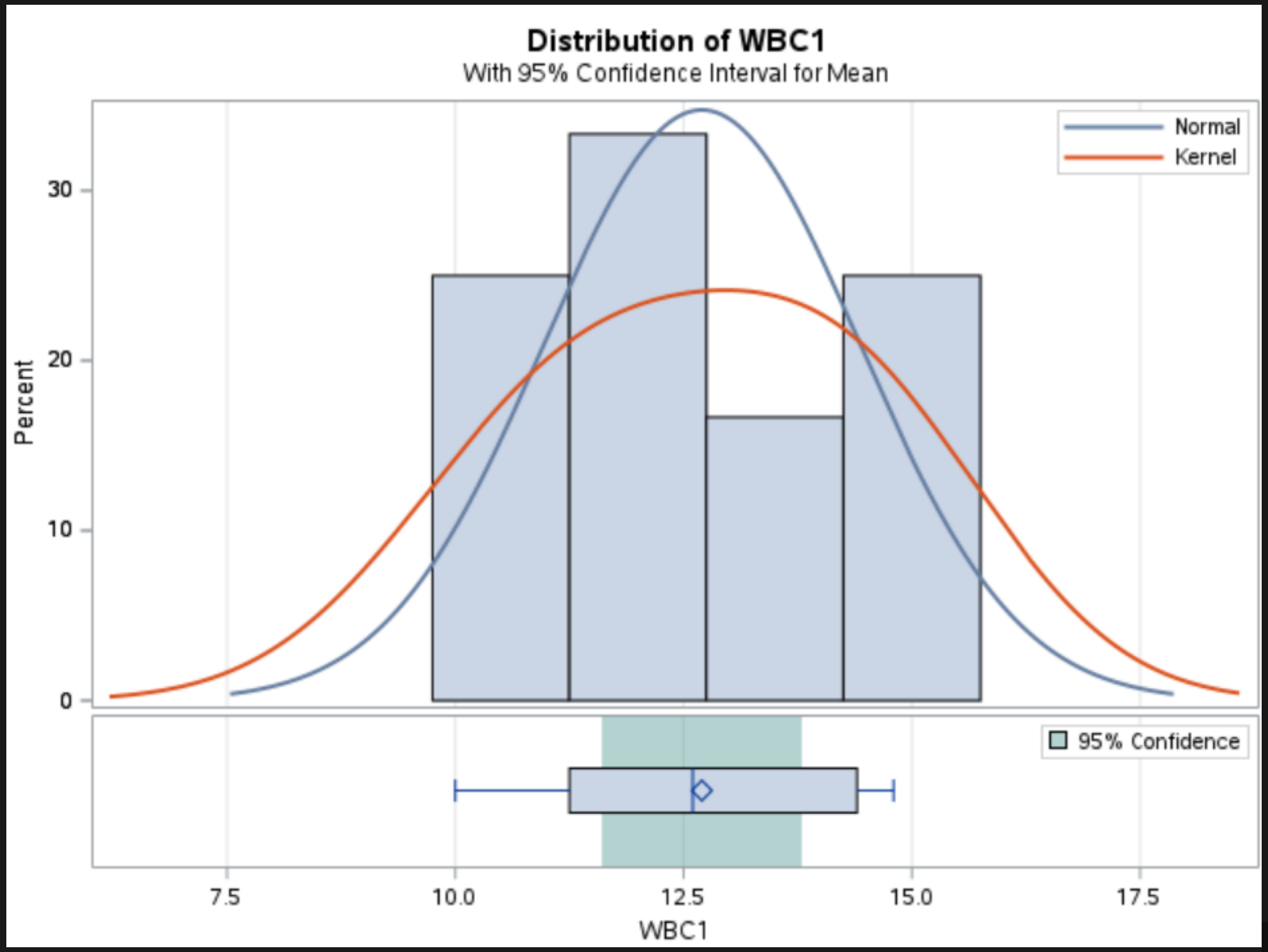
Variable: WBC1 (WBC1)

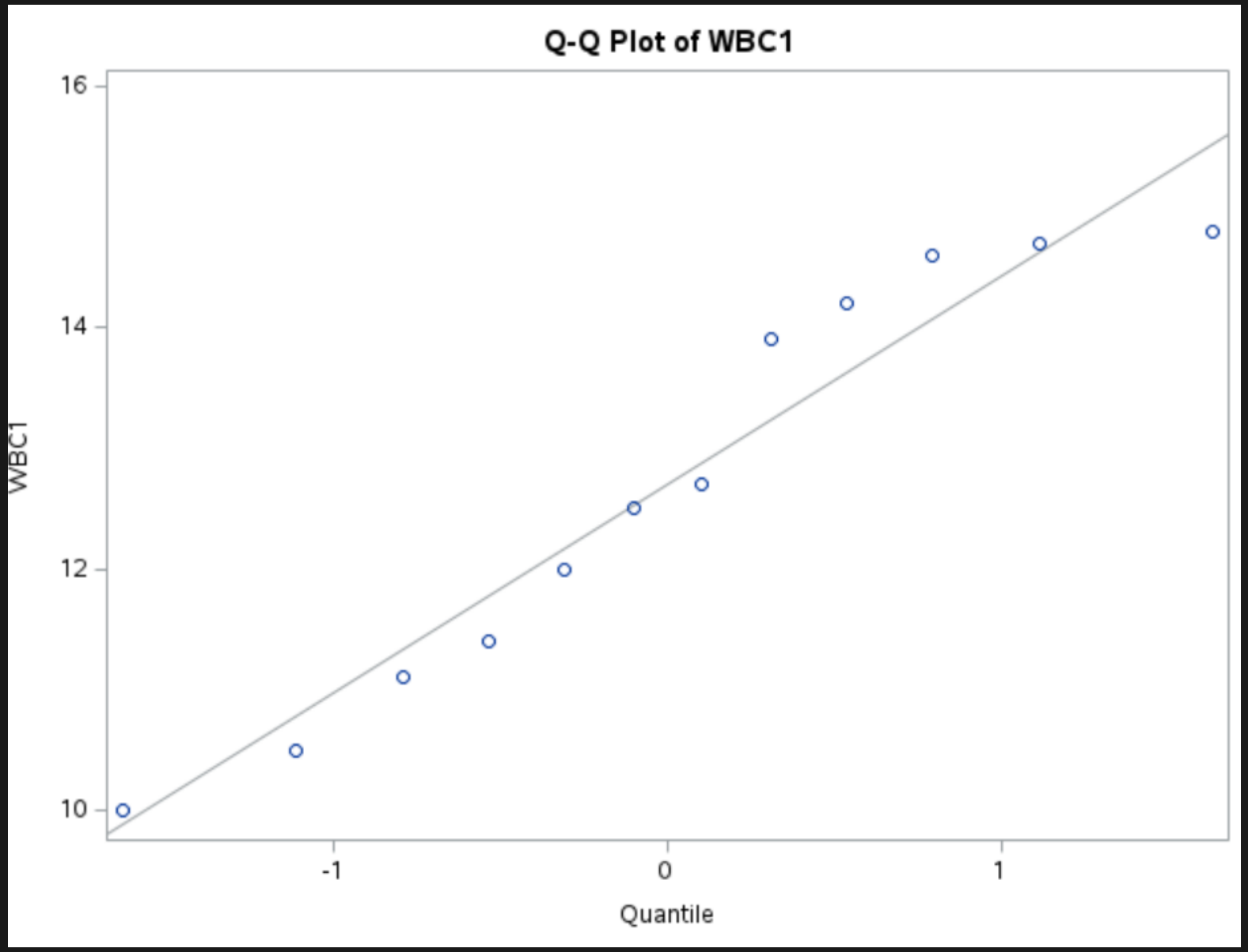
N	Mean	Std Dev	Std Err	Minimum	Maximum
12	12.7000	1.7220	0.4971	10.0000	14.8000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12.7000	11.6059	13.7941	1.7220	1.2199	2.9238

DF	t Value	Pr > |t|
11	5.43	0.0002

For one-sided test, divide the value of `Pr > |t|` by 2.


Distribution of WBC1
With 95% Confidence Interval for Mean
Normal
Kernel
Percent
30
20
10
0
95% Confidence
7.5
10.0
12.5
15.0
17.5
WBC1


Q-Q Plot of WBC1
16
14
12
10
WBC1
-1
0
1
Quantile

ASSESSING HYPOTHESIS TESTING

STATISTICAL SIGNIFICANCE VS PRACTICAL SIGNIFICANCE (EFFECT SIZE)

- H_0 : Proportion p_H of heads in coin flips is 0.5.
- Many physical coins have $p_H \approx 0.506$.
- With many coin flips (n large), we can reject H_0.
- Measure practice significance by effect size like
  - Cohen's $d = \frac{\bar{x} - \mu_0}{\sigma}$
  - Cohen's $d = \frac{\bar{x} - \mu_0}{s_x}$ if pop. standard deviation is unknown.

TYPE I AND II ERRORS

- **Type I error** or false positive rate (FPR) is
  $\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true}) = \text{“}P(H_a \mid H_0)\text{”}$
- **Type II error** or false negative rate (FNR) is
  $\beta = P(\text{Fail to reject } H_0 \mid H_a \text{ true})$
  $= \text{“}P(H_0 \mid H_a)\text{”}$