Some R functions

In R, there are 4 main functions relating to any built-in distribution. We already saw d/p binom and pnorm. Now I want to formalise these functions. I will use the $X \sim \mathcal{B}(n, p)$ as an example.

- **r**: `rbinom(size, n, prob)` randomly generate observations from X.
- **p**: `pbinom(x, n, prob)` calculates $P(X \leq x) =?$.
- **q**: `qbinom(p, n, prob)` calculates $P(X \leq ?) = p$.
- **d**: `dbinom(x, n, prob)` calculates $P(X = x) =?$.

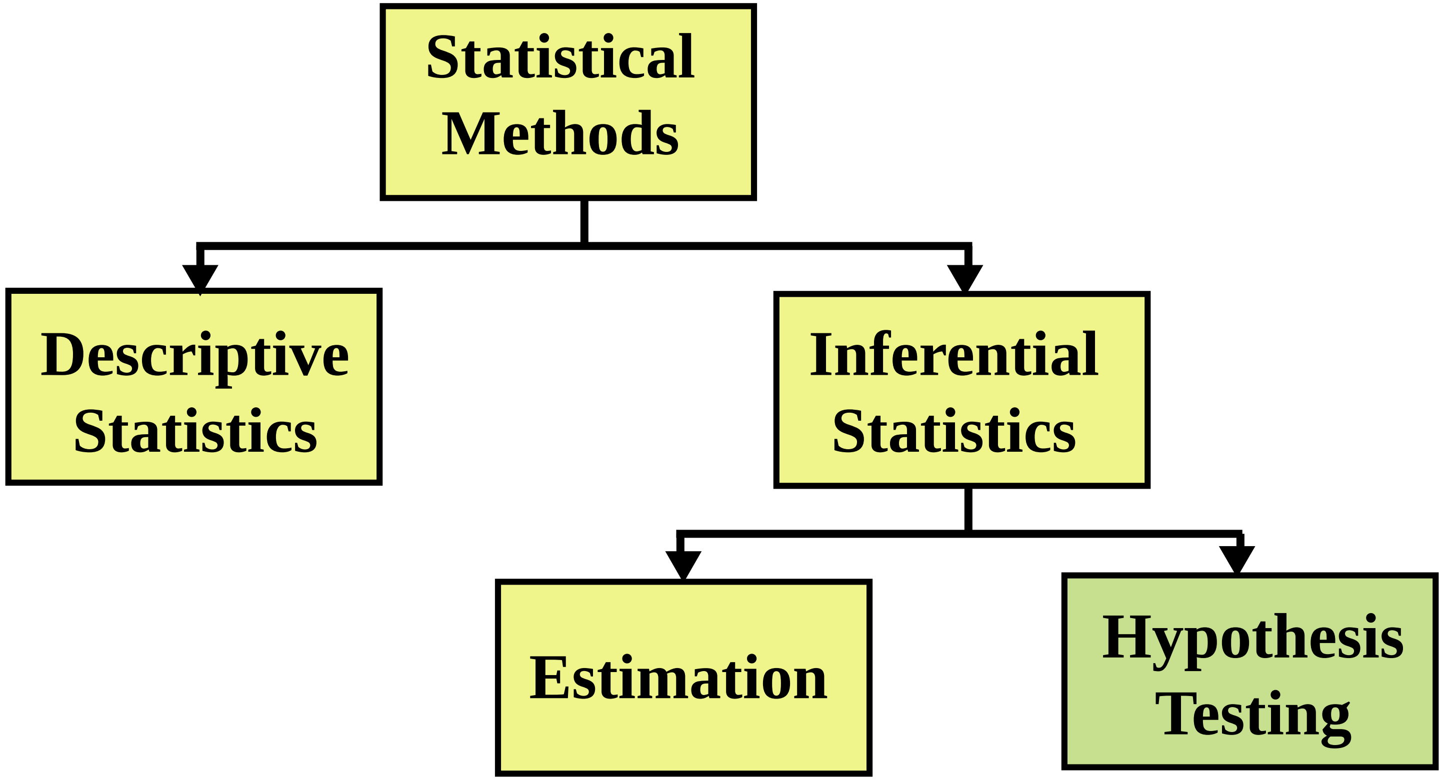
These definitions can be readily transferred to other distributions. The only exception is `dnorm`, which calculates the height of the probability density function at x.

Warm up questions

Given $X \sim \mathcal{B}(n = 20, p = 0.4)$ and $Y \sim \mathcal{N}(\mu = 20, \sigma^2 = 16)$ are two independent random variables:

1. Calculate $P(X = 10)$.
2. Calculate $P(X \leq 10)$.
3. Calculate $P(6 \leq X \leq 10)$.
4. Calculate $P(|X - 8| < 2)$.
5. Calculate $P(Y \leq 15)$.
6. Calculate $P(Y < 15)$.
7. Calculate $P(Y > 15)$.
8. Calculate $P(12 \leq Y \leq 15)$.
9. Calculate $P(X \leq 10, Y < 15)$.
10. Let $Z = X - 2Y$, calculate $E(Z)$ and $Var(Z)$.

Statistical Methods



How to Formulate a Decision Rule

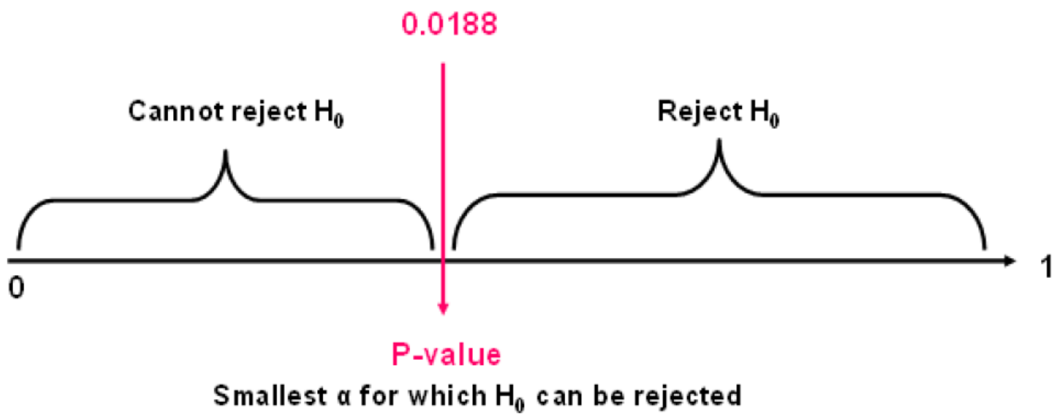
- **Critical value approach**
  - Reject H_0 if the test statistic falls in the rejection region
- **P-value approach**
  - Reject H_0 if the p-value $<$ α
- **Confidence interval approach**
  - Reject H_0 if the null value (value specified in H_0) lies **outside** the confidence interval

p-value

- Probability of obtaining a test statistic more extreme ($\leq$ or $\geq$) than actual sample value, given H_0 is true
- p-value = P(observing a **test statistic** in the tail area(s))
- Called **observed level of significance** / **observed α**
  - **Smallest value of α** for which H_0 can be rejected
  - Say, H_1: $\mu > 15$, $z_{stat} = 2.08$
    p-value = $P(Z > 2.08) = 0.0188$
- Used to make rejection decision:
  reject H_0 if p-value $< \alpha$

Level of Significance, α

- α = P(Type I error)
- Designated α (alpha)
  - Typical values are 0.01, 0.25, 0.05, 0.10
  - Default $\alpha = 0.05$
- Defines unlikely values of sample statistic if H_0 is true called **rejection region** of sampling distribution
- Selected by researcher at start



Relationship between Hypothesis Tests and Confidence Intervals		
Lower-tailed Test: H_0: **=** H_1: **<**	**Upper-tailed Test:** H_0: **=** H_1: **>**	**Two-tailed Test:** H_0: **=** H_1: **≠**
A **one-sided (left) C.I.** (**upper confidence bound**) of the form **(-∞, U]** corresponds to the retention region of a lower-tailed test.	A **one-sided (right) C.I.** (**lower confidence bound**) of the form **[L, ∞)** corresponds to the retention region of an upper-tailed test.	A **two-sided C.I.** of the form **[L, U]** corresponds to the retention region of a two-tailed test.
Decision rule: Reject H_0 if the null value lies **outside** the upper confidence bound.	Decision rule: Reject H_0 if the null value lies **outside** the lower confidence bound.	Decision rule: Reject H_0 if the null value lies **outside** the two-sided confidence interval.

Steps in Hypothesis Testing

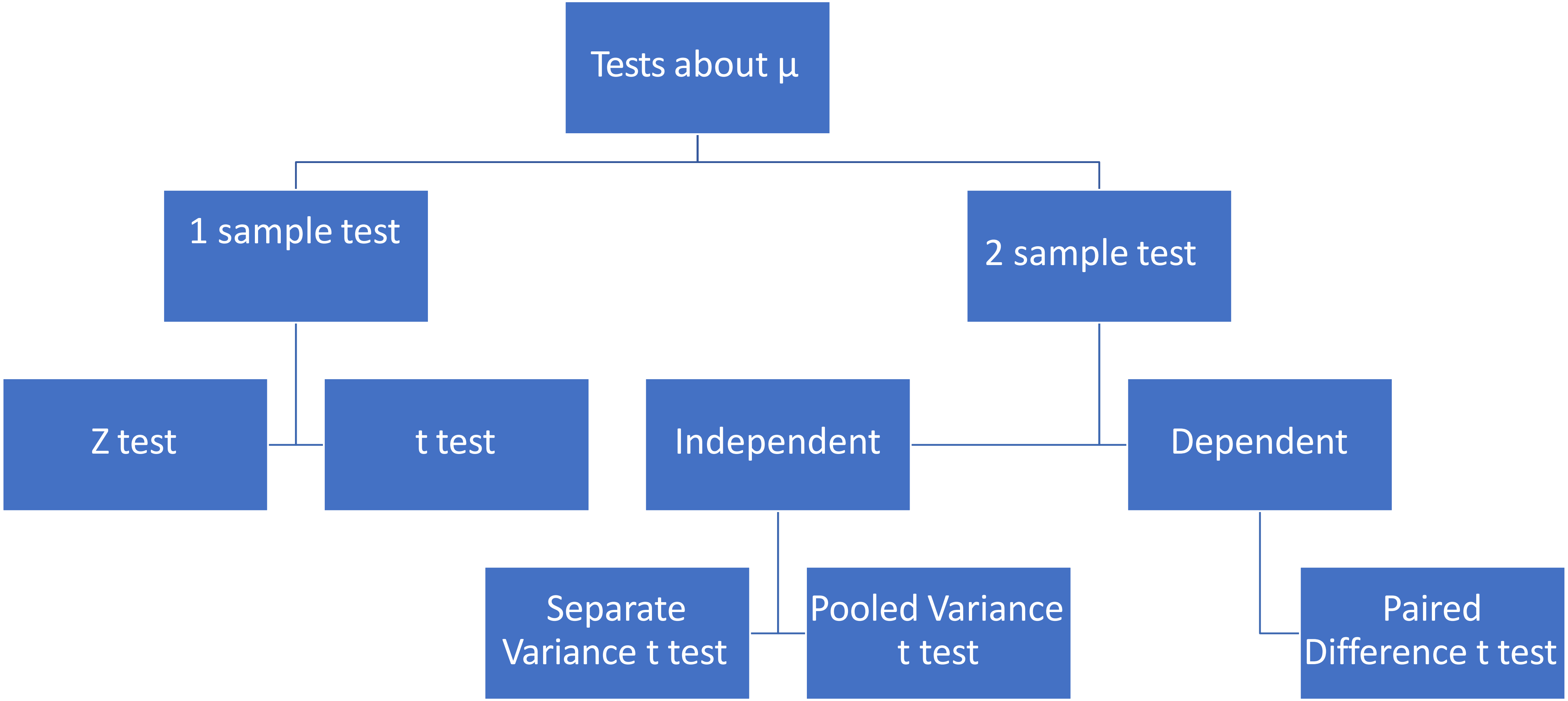
- H: Set up the 2 hypotheses: H_0 and H_1
- A: State the assumption(s) of the test, and justify whether they are valid from the sample.
- T: State the test statistic and specify the sampling distribution of the test statistic under H_0

  State what values argue against H_0

  Find the observed value of the test statistic.
- P: Calculate the p-value, which represents the probability of observing this sample (or more extreme) assuming H_0 is true.
- C: Weigh up the conclusion, based on the size of the p-value

Decision	Conclusion
Reject H_0	There is sufficient evidence to show that .... (**refer to H_1**)..........
Retain H_0	There is not sufficient evidence to show that .... (**refer to H_1**) ....

Tests for Means (Z and t tests)



One-sample z Test for μ

- Assumptions
  - **σ known**
  - Population is **normally distributed** (or **$n \geq 30$**)
  - A random sample is selected from a population
- z-test statistic:

$$\mathbf{z_{stat} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \text{ under } H_0}$$

One-sample t test for μ

- Assumptions
  - **σ unknown**
  - Population is **normally distributed** or (**$n \geq 30$**)
  - A random sample is selected from a population
- t test statistic:

$$\mathbf{t_{stat} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \text{ under } H_0}$$

Example1: Comm Bank's online service

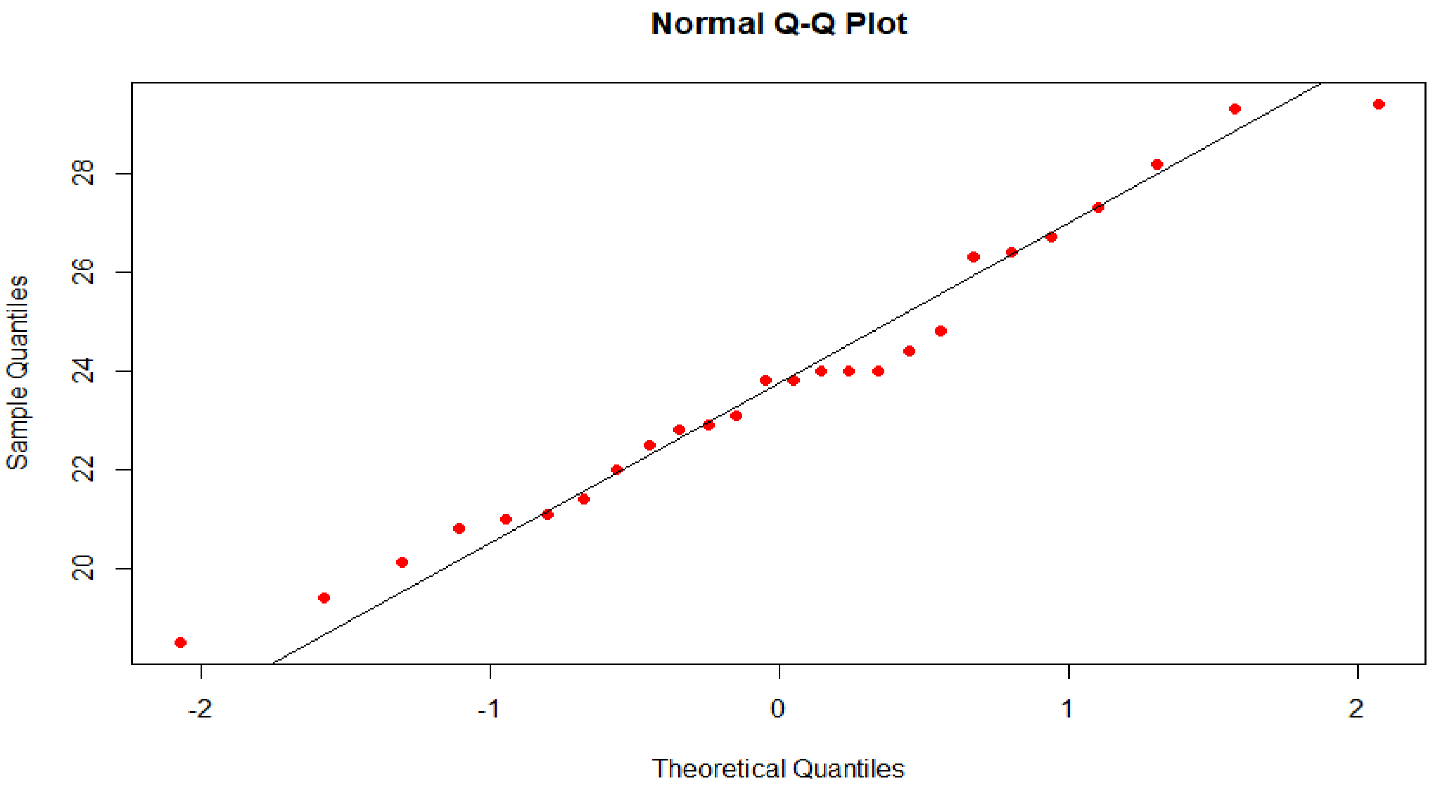
CommBank claims that an online personal loan application takes between 15-20 minutes to complete online. There has been complaints that the applications take longer. A random sample of 26 customers results in the following data. Assume that $\sigma = 5$.

```
x=c(29.3,23.1,18.5,23.8,24.8,23.8,22.5,26.3,20.8,21.1,21.4,24.0,22.0,28.2,27.3,19.4,20.1,26.4,
24.4, 24.0,21.0,22.8,29.4,22.9,26.7,24.0)
```

Would this evidence contradict the company's claim?

▸ Comm Bank FAQ

```
> x=c(29.3,23.1,18.5,23.8,24.8,23.8,22.5,26.3,20.8,21.1,21.4,24.0,22.0,28.2,27.3,19.4,20.1,26.4,24.4,24,21,22.8,29.4,22.9,26.
7,24)
> qqnorm(x, col="red", pch=19)
> qqline(x)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.97146, p-value = 0.6615
```



Example

CommBank

Do CommBank online personal loan applications take longer than advertised?

Let $\mu =$ Mean time for CommBank online personal loan application. Note this is a 1 sided test as we are testing whether the time is 'longer'.

$\boxed{\text{H}}$ We will use the upper bound (most conservative) of the claim, so $H_0 : \mu = 20$ vs $H_1 = \mu > 20$.

$\boxed{\text{A}}$ The $n = 26$ people in the survey are sampled randomly. Here $\sigma = 5$ is known.

$\boxed{\text{T}}$

- $\tau = Z = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ under H_0.
- Large values of z will argue against H_0 for H_1.
- As $\bar{x} = 23$ and we have assumed $\sigma = 5$, the observed value is $z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{23-20}{\frac{5}{\sqrt{26}}} \approx 3.06$.

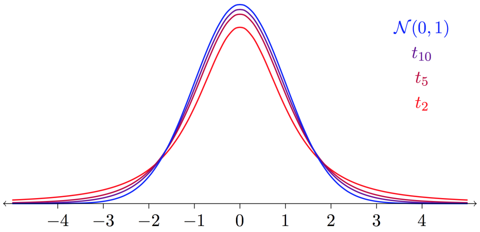
$\boxed{\text{P}}$ P-value $= P(Z \geq 3.06) \approx 0.001.$

```
1-pnorm(3.06)

## [1] 0.001106685

1-pnorm(23,20,5/sqrt(26))

## [1] 0.001108861
```

$\boxed{\text{C}}$ As the P-value is so small, we would question whether the claim is true.

t_n-distribution

- A t_n-distribution is determined uniquely by its **degree of freedom**, which is the subscript n.
- As $n \to \infty$, $t_n \to \mathcal{N}(0,1)$. t_n is essentially the normal distribution with a "small-sample-adjustment".
- The tails of t_n is a bit larger than $\mathcal{N}(0,1)$, which reflects the additional variability since we are estimating σ by S.



Example

CommBank

Do CommBank online personal loan applications take longer than advertised?

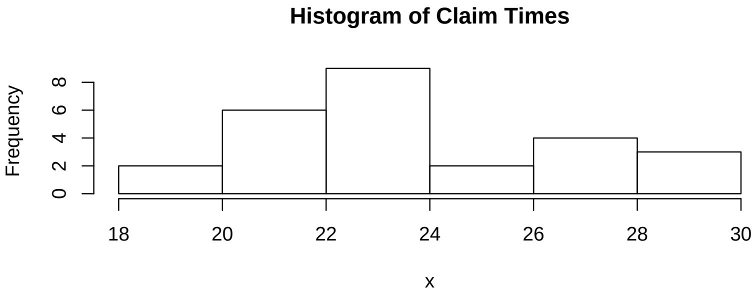
We now use the T test. This is preferable as the previous usage of Z test required assuming that $\sigma = 5$. However, here we don't know whether the distribution is Normal, so this is a untested assumption. From the sample of 26 application times $\{23.6, 26.7, 22.9, \ldots, 24.3\}$, we have $\bar{x} \approx 23.8$ and $s \approx 2.92$.

```
mean(x)

## [1] 23.76923

sd(x)

## [1] 2.928176
```

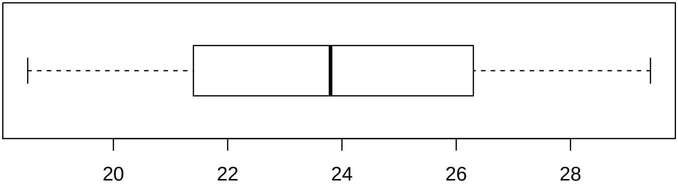
$\boxed{\text{H}}$ $H_0 : \mu = 20$ vs $H_1 = \mu > 20$.

$\boxed{\text{A}}$ We assume that the population of claim times is Normally distributed.

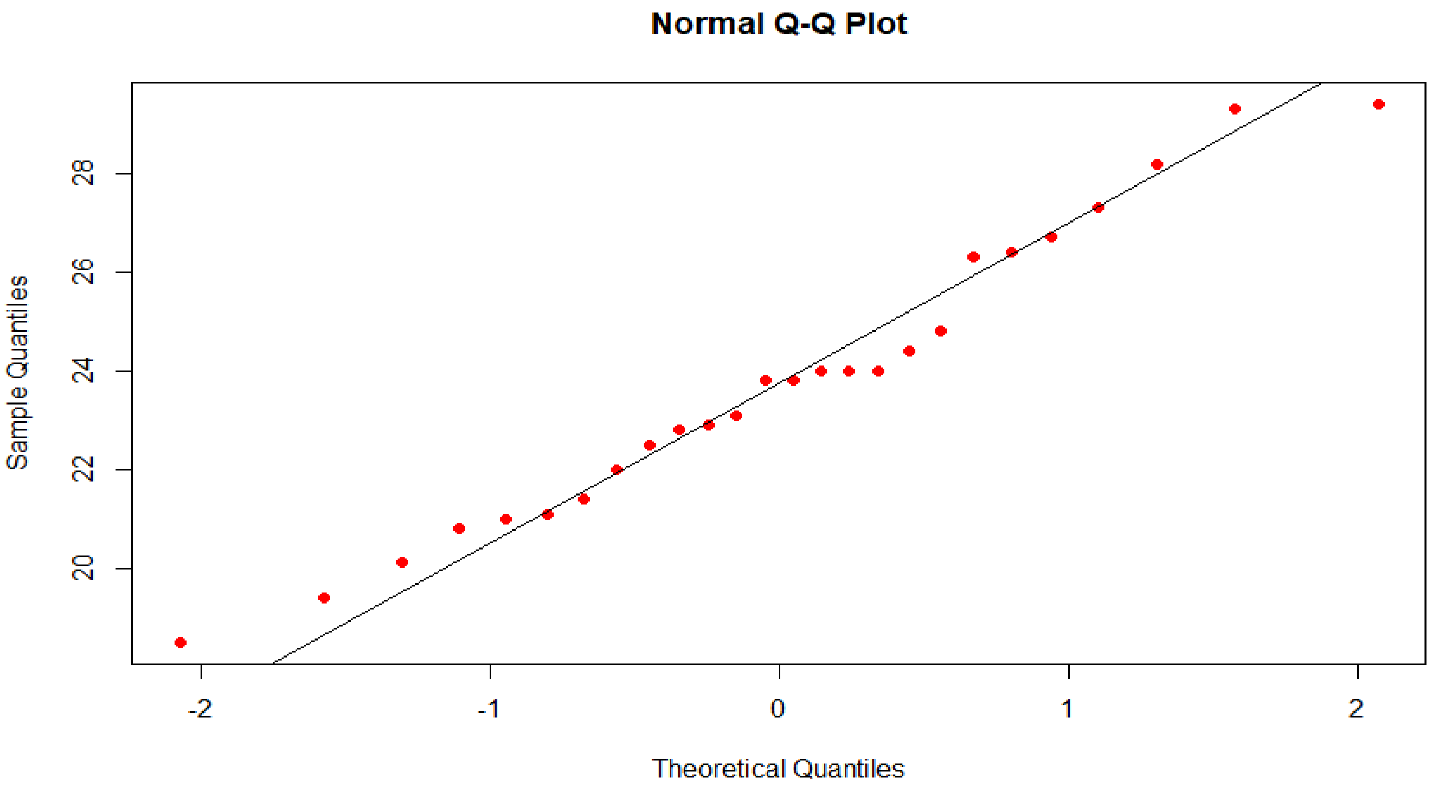
```
hist(x,main="Histogram of Claim Times")
```



```
boxplot(x,horizontal=T)
```



```
> x=c(29.3,23.1,18.5,23.8,24.8,23.8,22.5,26.3,20.8,21.1,21.4,24.0,22.0,28.2,27.3,19.4,20.1,26.4,24.4,24,21,22.8,29.4,22.9,26.7,24)
> qqnorm(x, col="red", pch=19)
> qqline(x)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.97146, p-value = 0.6615
```



$\boxed{\text{T}}$

- $\tau = T = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}} \sim t_{25}$ under H_0.
- Large values of T will argue against H_0 for H_1.
- The observed value is $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{23.8-20}{\frac{2.92}{\sqrt{26}}} \approx 6.64$

$\boxed{\text{P}}$ P-value $= P(t_{25} \geq 6.64) \approx 0.0000007.$

```
1-pt(6.64,25)

## [1] 2.93773e-07
```

$\boxed{\text{C}}$ As the P-value is so small, again we question whether the claim is true.

```
> t.test(x,mu=20, alternative="greater")

        One Sample t-test

data:  x
t = 6.5636, df = 25, p-value = 3.544e-07
alternative hypothesis: true mean is greater than 20
95 percent confidence interval:
 22.78831      Inf
sample estimates:
mean of x
 23.76923
```

Independent & Related Populations

Independent

1. Different data sources
   - Unrelated
   - Independent
2. Use difference between the two sample means
   - $\bar{X}_1 - \bar{X}_2$

Related

1. Same data source
   - Paired or matched
   - Repeated measures (before/after)
2. Use difference between each pair of observations
   - $d_i = x_{1i} - x_{2i}$

Difference Between Two Means

Population means, independent samples *

- σ_1 and σ_2 unknown, assumed equal
- σ_1 and σ_2 unknown, not assumed equal

Goal: Test hypothesis or form a confidence interval for the difference between two population means, $\mu_1 - \mu_2$

The point estimate for the difference is

$$\bar{X}_1 - \bar{X}_2$$

Difference Between Two Means: Independent Samples

- Different data sources
  - Unrelated
  - Independent
    - Sample selected from one population has no effect on the sample selected from the other population

Population means, independent samples *

σ_1 and σ_2 unknown, not assumed equal → Use S_1 and S_2 to estimate unknown σ_1 and σ_2. Use a **Separate-variance t test**

σ_1 and σ_2 unknown, assumed equal → Use S_p to estimate unknown σ. Use a **Pooled-Variance t test.**

Hypothesis Tests for $\mu_1 - \mu_2$

Two Population Means, Independent Samples

Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 < 0$	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 > 0$	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$

Hypothesis Tests for $\mu_1 - \mu_2$

Two Population Means, Independent Samples

Lower-tail test:

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 < 0$

Upper-tail test:

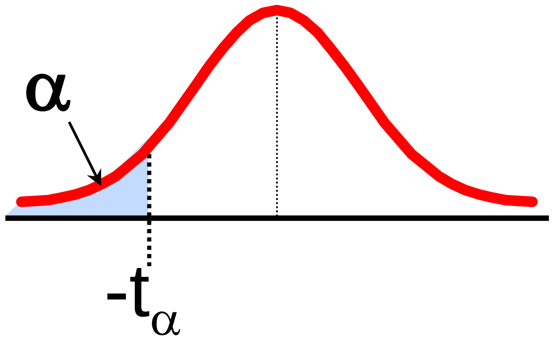
$H_0: \mu_1 - \mu_2 = 0$

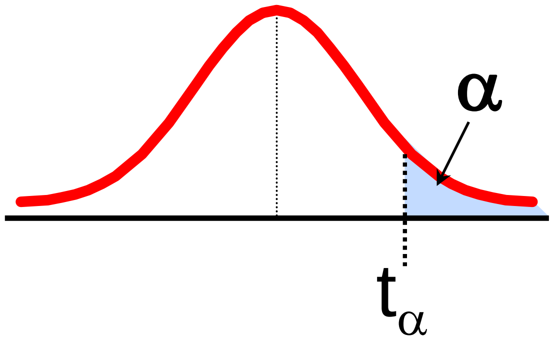
$H_1: \mu_1 - \mu_2 > 0$

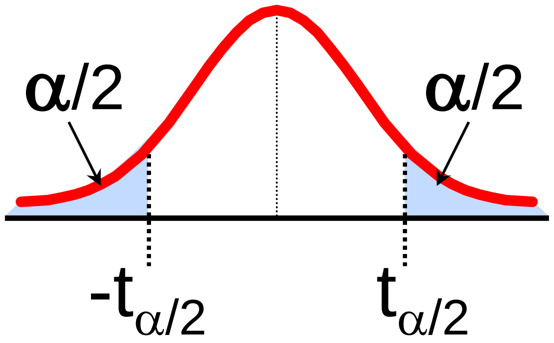
Two-tail test:

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$







Reject H_0 if $t_{STAT} < -t_{\alpha}$

Reject H_0 if $t_{STAT} > t_{\alpha}$

Reject H_0 if $t_{STAT} < -t_{\alpha/2}$ or $t_{STAT} > t_{\alpha/2}$

Separate-Variance t Test

Population means, independent samples

σ_1 and σ_2 unknown, not assumed equal *

σ_1 and σ_2 unknown, assumed equal

Assumptions:

- Populations are normally distributed or both sample sizes are at least 30.
- Population variances are unknown and cannot be assumed to be equal.
- Samples are randomly and independently drawn.

Separate-Variance t Test – cont'd

Population means, independent samples

- σ_1 and σ_2 unknown, not assumed equal *
- σ_1 and σ_2 unknown, assumed equal

The test statistic is:

$$t_{STAT} = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Rule-of-Thumb – Homogeneity of Variance

- Simulation can be used to show that this assumption does not actually matter if $n_1 = n_2$.
- The only time you run into problem is when both $\frac{n_1}{n_2}$ and $\frac{\sigma_1}{\sigma_2}$ are very different from 1 (they differ from 1 by at least a factor of 2, say). In this situation, we have a problem.

Dietary effects of high-fibre breakfast cereals

Despite some controversy, scientists generally agree that high-fibre cereals reduce the likelihood of various forms of cancer. However, one scientist claims that people who eat high-fibre cereal for breakfast will consume, on average, fewer kilojoules for lunch than people who do not eat high-fibre cereal for breakfast. If this is true, high-fibre cereal manufacturers will be able to claim another advantage of eating their product – potential weight reduction for dieters.

As a preliminary test of the claim, 30 people were randomly selected and asked what they regularly ate for breakfast and lunch. Each person was identified as either a consumer or a non-consumer of high-fibre breakfast cereal, and the number of kilojoules consumed at lunch was measured and recorded. These data are listed below.

Dietary effects of high-fibre breakfast cereals

Kilojoules consumed at lunch

Consumers of high-fibre cereal:

c = c(2560, 2420, 2116, 2364, 2384, 2256, 2460, 2240, 2540, 2492)

Non-consumers of high-fibre cereal:

nc = c(2008, 2812, 2940, 2828, 2092, 2136, 3072, 2504, 2480, 2356, 2944, 2260, 2744, 2116, 2528, 3804, 2976, 2528, 2372, 3388)

At the 5% significance level, test the scientists' claim that people who eat high-fibre cereal for breakfast will consume, on average, fewer kilojoules for lunch than people who don't eat high-fibre cereal for breakfast.

```
> c = c(2560, 2420, 2116, 2364, 2384, 2256, 2460, 2240, 2540, 2492)
> nc = c(2008, 2812, 2940, 2828, 2092, 2136, 3072, 2504, 2480, 2356,
+        +         2944, 2260, 2744, 2116, 2528, 3804, 2976, 2528, 2372, 3388)
> shapiro.test(c)

        Shapiro-Wilk normality test

data:  c
W = 0.9493, p-value = 0.6602

> shapiro.test(nc)

        Shapiro-Wilk normality test

data:  nc
W = 0.94479, p-value = 0.2948

> var.test(c,nc,alternative="two.sided")

        F test to compare two variances

data:  c and nc
F = 0.095214, num df = 9, denom df = 19, p-value = 0.00106
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.03305981 0.35070530
sample estimates:
ratio of variances
        0.09521399
```

```
> t.test(c, nc,alternative = "less", var.equal=FALSE)

        Welch Two Sample t-test

data:  c and nc
t = -2.3143, df = 25.011, p-value = 0.01457
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -68.41904
sample estimates:
mean of x mean of y 
   2383.2    2644.4 
```

Pooled-Variance t Test

Population means, independent samples

σ_1 and σ_2 unknown, not assumed equal

σ_1 and σ_2 unknown, assumed equal *

Assumptions:

- Populations are normally distributed or both sample sizes are at least 30
- Population variances are unknown but assumed equal
- Samples are randomly and independently drawn

Pooled-Variance t Test – cont'd

Population means, independent samples

- σ_1 and σ_2 unknown, assumed equal *
- σ_1 and σ_2 unknown, not assumed equal

- The pooled variance is:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

- The test statistic is:

$$t_{STAT} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- where t_{STAT} has d.f. = $(n_1 + n_2 - 2)$

Why do we use the pooled variance?

As with the 1 sample t test, we need an estimate of the unknown variance σ^2 (here by assumption, common for both populations). Options for the estimate include:

- The sample variance of Population X: s_x^2.
- The sample variance of Population Y: s_y^2.
- The average of the sample variances: $\frac{s_x^2+s_y^2}{2}$.
- The weighted average of the sample variances: s_p^2. This takes into account both sample variances and the size of the samples. s_p^2 is always between s_x^2 and s_y^2.

Example2: ADHD in children in Taiwan

A study from 2013 looked at the long-term effects of stimulants on neurocognitive performance of Taiwanese children with attention deficit hyperactivity disorder (ADHD) using the Wechsler Intelligence Scale (WISC-III).

"In Taiwan, a high prevalence rate of ADHD was noticed about ten years ago, but there is still little research comparing neurocognitive function between children with ADHD and healthy children."
"Due to the nature of the populations sampled, diagnostic criteria used, cultural differences, and methodological limitations, the prevalence of ADHD in various cultures varies. The prevalence is estimated to be about 8.4–11.7% in Taiwan; 2.4% in Australia; and 4% in Japan."

It was found that the 47 children in the control group (without ADHD) had a BMI (body mass index) of 18.8 with standard deviation 3.3, and the group of 171 children with ADHD had a BMI of 18.5 with standard deviation 3.7.

Is there evidence that children with ADHD have a different BMI than the general population?

▸ Abstract ▸ TableT1 ▸ TableT2 ▸ BMI

Example

ADHD in children in Taiwan

Is there evidence that children with ADHD have a different BMI to the general population?

$\boxed{\text{Preparation}}$

Let Population X = Control and Population Y = ADHD.

μ_X = Mean BMI of children in general population and μ_Y= Mean BMI of children with ADHD.

$n_x = 47$, $\bar{x} = 18.8$, $s_x = 3.3$, $n_y = 171$, $\bar{y} = 18.5$, $s_y = 3.7$.

So $s_p = \sqrt{\frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}} = 3.618522.$

```
n_x= 47
xbar=18.8
s_x=3.3
n_y=171
ybar=18.5
s_y=3.7
sp = sqrt( ((n_x-1)*s_x^2 + (n_y-1)*s_y^2)/(n_x+n_y-2) )
sp

## [1] 3.618522

t = (xbar-ybar)/(sp*sqrt(1/n_x + 1/n_y))
t

## [1] 0.5033948
```

H $H_0 : \mu_X - \mu_Y = 0$ vs $H_1 = \mu_X - \mu_Y \neq 0$.

A We assume that both population are Normally distributed with common variance, and that the 2 samples are independent. We do not have the raw data to be able to do histograms as a diagnostic.

T

- $\tau = T = \frac{\bar{X} - \bar{Y} - c}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{n_x + n_y - 2} \sim t_{47+171-2} = t_{216}$ under H_0.
- Large and small values of T will argue against H_0 for H_1.
- The observed value is $t = 0.5033948$

P P-value $= 2P(t_{216} \geq 0.5033948) \approx 0.615$.

```
2*(1-pt(0.5033948,216))

## [1] 0.6151997
```

$\boxed{\text{C}}$ As the P-value is so big, there does not appear to be a difference in the BMIs of children with and without ADHD.

Paired Difference t Test

Related samples

Tests Means of 2 Related Populations

- Paired or matched samples
- Repeated measures (before/after)
- Use difference between paired values:

$$\boxed{d_i = X_{1i} - X_{2i}}$$

- Eliminates Variation Among Subjects
- Assumptions:
  - Population of differences is normal or $n \geq 30$ with σ_d unknown
  - The differences are randomly selected from the population of difference.

Paired Difference t Test

Related samples

The i^{th} paired difference is d_i , where

$$d_i = X_{1i} - X_{2i}$$

The point estimate for the paired difference population mean μ_d is $\bar{d}$

The sample standard deviation is S_d

n is the number of pairs in the paired sample

Paired Difference t Test:

Paired samples

- The test statistic for μ_d is:

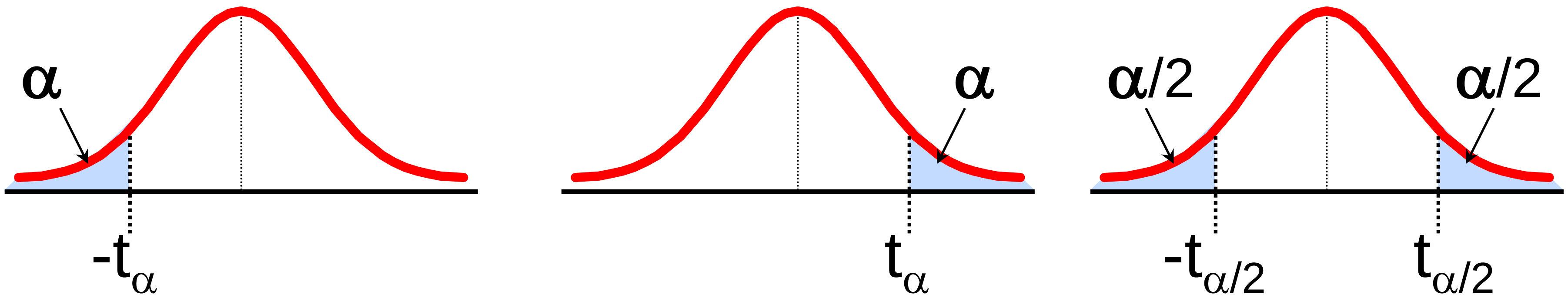
$$t_{STAT} = \frac{\bar{d} - \mu_d}{\frac{S_d}{\sqrt{n}}}$$

- where t_{STAT} has $n - 1$ d.f.

The Paired Difference Test: Possible Hypotheses

Paired Samples

Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \mu_d = 0$	$H_0: \mu_d = 0$	$H_0: \mu_d = 0$
$H_1: \mu_d < 0$	$H_1: \mu_d > 0$	$H_1: \mu_d \neq 0$



Reject H_0 if $t_{STAT} < -t_\alpha$

Reject H_0 if $t_{STAT} > t_\alpha$

Reject H_0 if $t_{STAT} < -t_{\alpha/2}$ or $t_{STAT} > t_{\alpha/2}$

where t_{STAT} has $n - 1$ d.f.

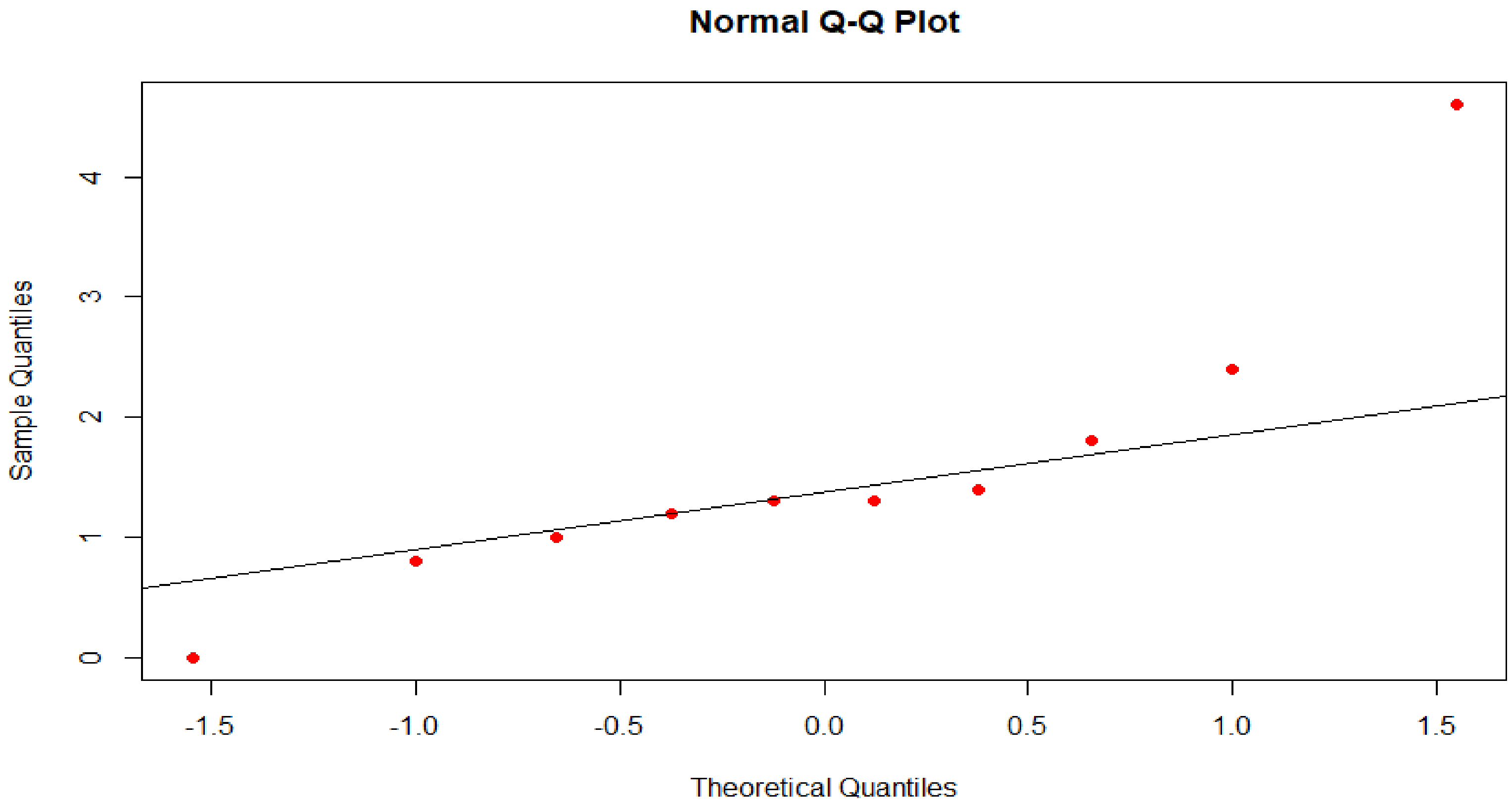
Paired T Test

The T Test can also be applied to paired data. While the Sign Test only requires the population is continuous, the T Test requires a stronger assumption: that the population of differences are Normal.

Sleep Study

Is there a difference between the affect of drugs on sleep?

```
> a=c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0)
> b=c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4)
> diff=b-a
> qqnorm(diff, col="red", pch=19)
> qqline(diff)
> shapiro.test(diff)

        Shapiro-Wilk normality test

data:  diff
W = 0.82987, p-value = 0.03334
```



```r
a=c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0)
b=c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4)
diff=b-a
mean(diff)

## [1] 1.58

sd(diff)

## [1] 1.229995

tobs = (mean(diff)-0)/(sd(diff)/sqrt(10))
2*(1-pt(tobs,9))

## [1] 0.00283289
```

$\boxed{\text{H}}$ $H_0 : \mu = 0$ vs $H_0 : \mu \neq 0$, where μ is the population mean of the differences $B - A$.

$\boxed{\text{A}}$ The set of differences is Normal.

- $\tau = T = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}} \sim t_9$ under H_0.
- Large and small values of T will argue against H_0 for H_1.
- The observed value is $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{1.58-0}{\frac{1.229995}{\sqrt{10}}} \approx 4.06$

$\boxed{\text{P}}$ P-value $= 2P(t_9 \geq 4.06) \approx 0.003$.

```
2*(1-pt(4.06,9))

## [1] 0.002841947
```

$\boxed{\text{C}}$ As the P-value is so small, again we would question whether the drugs are equivalent.

```
> t.test(diff, mu=0)

        One Sample t-test

data:  diff
t = 4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.7001142 2.4598858
sample estimates:
mean of x
     1.58
```

Paired t-test vs two-sample t-test

- Data structure:

  PTT assumes a natural pairing of samples.

  TSTT No pairings are assumed.

Paired t-test vs two-sample t-test

- Data structure:
  - PTT: assumes a natural pairing of samples.
  - TSTT: No pairings are assumed.
- Assumptions:
  - PTT: $D_i = X_i - Y_i$ are assumed to be $D_i \overset{iid}{\sim} \mathcal{N}(\mu_D, \sigma_D^2)$
  - TSTT: $X \overset{iid}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ and $Y \overset{iid}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$. X and Y are independent and share a common, unknown variance σ^2.

Paired t-test vs two-sample t-test

- Data structure:
  - PTT assumes a natural pairing of samples.
  - TSTT No pairings are assumed.
- Assumptions:
  - PTT $D_i = X_i - Y_i$ are assumed to be $D_i \overset{iid}{\sim} \mathcal{N}(\mu_D, \sigma_D^2)$
  - TSTT $X \overset{iid}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ and $Y \overset{iid}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$. X and Y are independent and share a common, unknown variance σ^2.
- Hypothesis:
  - PTT $H_0 : \mu_D = 0$, $H_A : \mu_D \neq 0$
  - TSTT $H_0 : \mu_X = \mu_Y$, $H_A : \mu_X \neq \mu_Y$.

Paired t-test vs two-sample t-test

- Data structure:

  PTT assumes a natural pairing of samples.

  TSTT No pairings are assumed.

- Assumptions:

  PTT $D_i = X_i - Y_i$ are assumed to be $D_i \overset{iid}{\sim} \mathcal{N}(\mu_D, \sigma_D^2)$

  TSTT $X \overset{iid}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ and $Y \overset{iid}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$. X and Y are independent and share a common, unknown variance σ^2.

- Hypothesis:

  PTT $H_0 : \mu_D = 0$, $H_A : \mu_D \neq 0$

  TSTT $H_0 : \mu_X = \mu_Y$, $H_A : \mu_X \neq \mu_Y$.

- Test statistic:

PTT:

$$\frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$$

TSTT:

$$\frac{\bar{X} - \bar{Y} - 0}{S_{\text{pooled}}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \overset{H_0}{\sim} t_{n_x+n_y-2}$$

2 Sample T Test

The T test (or Z Test) can be generalised to cover 2 populations and samples as follows.

$\boxed{\text{Context}}$ Consider 2 populations with unknown means μ_X and μ_Y and unknown common variance σ^2. We take 2 independent samples. We want to test a hypothesis about $\mu_X - \mu_Y$.

$\boxed{\text{H}}$ $H_0 : \mu_X - \mu_Y = c$ vs $H_1 : \mu_X - \mu_Y < c$. (Note: Often $c = 0$.)

$\boxed{\text{A}}$ The 2 populations are Normal with common σ^2. The 2 samples are independent.

T

- $\tau = T = \frac{\bar{X}-\bar{Y}-c}{s_p\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \sim t_{n_x+n_y-2}$ (under H_0), where the pooled standard deviation is $s_p = \sqrt{\frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}}$.
- Small values of T will argue against H_0 for H_1.
- The observed value is t.

P P-value $= P(t_{n_x+n_y-2} \leq t)$.

C Weigh up size of P-value.

Breaking down two-sample t-test

$$\frac{\bar{X} - \bar{Y} - 0}{S_{\text{pooled}}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}},$$

- $\bar{X} - \bar{Y} - 0$. Difference in the sample means. Discrepancies between data and the null hypothesis,
- $\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$. A term that accounts the sample sizes, and hence, certainty of our inferences.
- ${S_{\text{Pooled}}}^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$. Recall that we assumed X and Y have the same common variance σ^2. S^2_{Pooled} collects both estimates of σ^2, and "pool" those together to achieve a more sensible and stable estimate. It accounts for variabilities in the data.