Statistics fundamentals toolkit

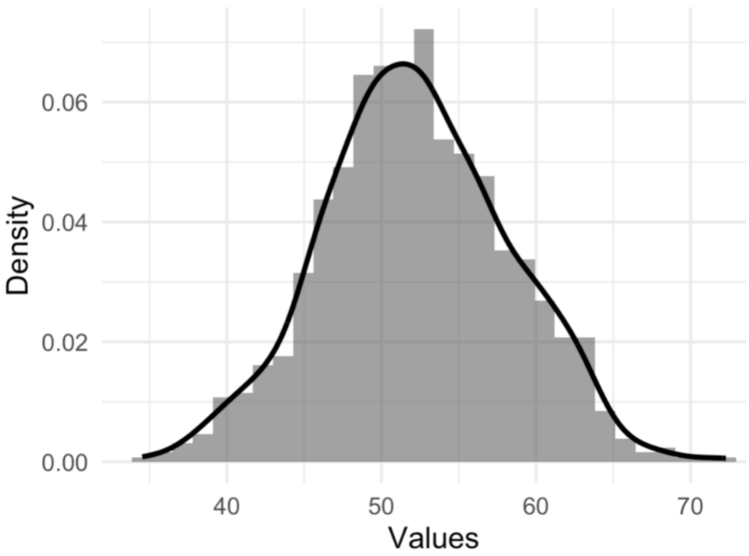
Descriptive stats • Contrasts • Correlations

PLS 206

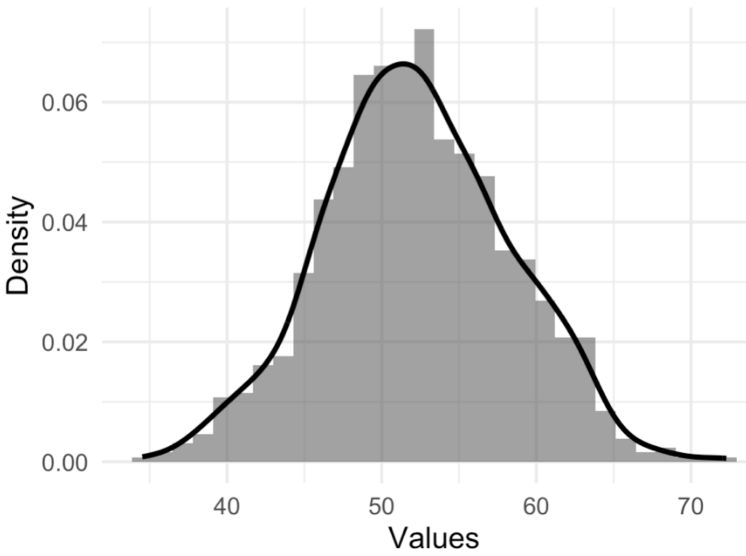
Goals for Today

- Summaries (mean, SD, SE, CI)
- Group comparisons (t-tests, ANOVA; non-parametric)
- Correlation (Pearson, Spearman)
- Test choice guidelines

```
> x
  [1] 48.63715 50.61894 61.35225 52.42305 52.77573 62.29039 54.76550
  [8] 44.40963 47.87888 49.32603 59.34449 54.15888 54.40463 52.66410
 [15] 48.66495 62.72148 54.98710 40.20030 56.20814 49.16325 45.59306
 [22] 50.69215 45.84397 47.62665 48.24976 41.87984 57.02672 52.92024
 [29] 45.17118 59.52289 54.55879 50.22957 57.37075 57.26880 56.92949
 [36] 56.13184 55.32351 51.62853 50.16422 49.71717 47.83176 50.75250
 [43] 44.40762 65.01374 59.24777 45.26135 49.58269 49.20007 56.67979
 [50] 51.49979 53.51991 51.82872 51.74278 60.21161 50.64537 61.09882
 [57] 42.70748 55.50768 52.74313 53.29565 54.27784 48.98606 50.00076
 [64] 45.88855 45.56925 53.82117 54.68926 52.31803 57.53360 64.30051
 [71] 49.05381 38.14499 58.03443 47.74480 47.87195 58.15343 50.29136
 [78] 44.67569 53.08782 51.16665 52.03459 54.31168 49.77604 55.86626
 [85] 50.67708 53.99069 58.58103 54.61109 50.04441 58.89285 57.96102
 [92] 55.29038 53.43239 48.23256 60.16391 48.39844 65.12400 61.19566
 [99] 50.58580 45.84147
```



```
> x
 [1] 48.63715 50.61894 61.35225 52.42305 52.77573 62.29039 54.76550
 [8] 44.40963 47.87888 49.32603 59.34449 54.15888 54.40463 52.66410
[15] 48.66495 62.72148 54.98710 40.20030 56.20814 49.16325 45.59306
[22] 50.69215 45.84397 47.62665 48.24976 41.87984 57.02672 52.92024
[29] 45.17118 59.52289 54.55879 50.22957 57.37075 57.26880 56.92949
[36] 56.13184 55.32351 51.62853 50.16422 49.71717 47.83176 50.75250
[43] 44.40762 65.01374 59.24777 45.26135 49.58269 49.20007 56.67979
[50] 51.49979 53.51991 51.82872 51.74278 60.21161 50.64537 61.09882
[57] 42.70748 55.50768 52.74313 53.29565 54.27784 48.98606 50.00076
[64] 45.88855 45.56925 53.82117 54.68926 52.31803 57.53360 64.30051
[71] 49.05381 38.14499 58.03443 47.74480 47.87195 58.15343 50.29136
[78] 44.67569 53.08782 51.16665 52.03459 54.31168 49.77604 55.86626
[85] 50.67708 53.99069 58.58103 54.61109 50.04441 58.89285 57.96102
[92] 55.29038 53.43239 48.23256 60.16391 48.39844 65.12400 61.19566
```




```
> mean(x)
[1] 52.23827
> sum(x)/length(x)
[1] 52.23827
> sd(x)
[1] 6.027895
> sqrt(sum((mean(x)-x)^2)/(length(x)-1))
[1] 6.027895
> (se=sd(x)/sqrt(length(x)))
[1] 0.1906188
> var(x)
[1] 36.33552
> sum((x - mean(x))^2) / (length(x) - 1)
[1] 36.33552
> median(x)
[1] 52.0056
> min(x)
[1] 34.52097
> max(x)
[1] 72.28588
> quantile(x, probs = c(0, 0.25, 0.5, 0.75, 1))
      0%      25%      50%      75%     100% 
34.52097 48.25772 52.00560 56.30350 72.28588 
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  34.52   48.26   52.01   52.24   56.30   72.29 
```

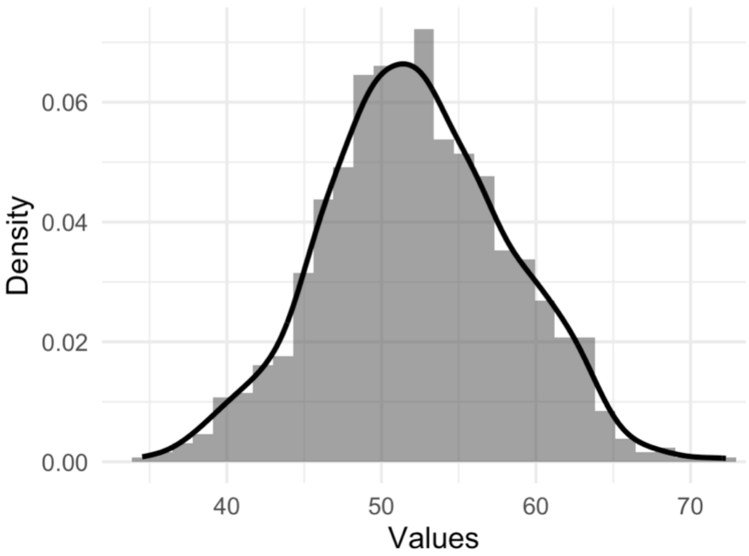
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$SE = \frac{s}{\sqrt{n}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- $\bar{x}$ = sample mean
- μ_0 = hypothesized mean
- s = sample standard deviation
- n = sample size

```
> x
  [1] 48.63715 50.61894 61.35225 52.42305 52.77573 62.29039 54.76550
  [8] 44.40963 47.87888 49.32603 59.34449 54.15888 54.40463 52.66410
 [15] 48.66495 62.72148 54.98710 40.20030 56.20814 49.16325 45.59306
 [22] 50.69215 45.84397 47.62665 48.24976 41.87984 57.02672 52.92024
 [29] 45.17118 59.52289 54.55879 50.22957 57.37075 57.26880 56.92949
 [36] 56.13184 55.32351 51.62853 50.16422 49.71717 47.83176 50.75250
 [43] 44.40762 65.01374 59.24777 45.26135 49.58269 49.20007 56.67979
 [50] 51.49979 53.51991 51.82872 51.74278 60.21161 50.64537 61.09882
 [57] 42.70748 55.50768 52.74313 53.29565 54.27784 48.98606 50.00076
 [64] 45.88855 45.56925 53.82117 54.68926 52.31803 57.53360 64.30051
 [71] 49.05381 38.14499 58.03443 47.74480 47.87195 58.15343 50.29136
 [78] 44.67569 53.08782 51.16665 52.03459 54.31168 49.77604 55.86626
 [85] 50.67708 53.99069 58.58103 54.61109 50.04441 58.89285 57.96102
 [92] 55.29038 53.43239 48.23256 60.16391 48.39844 65.12400 61.19566
 [99] 50.58580 45.84147
```

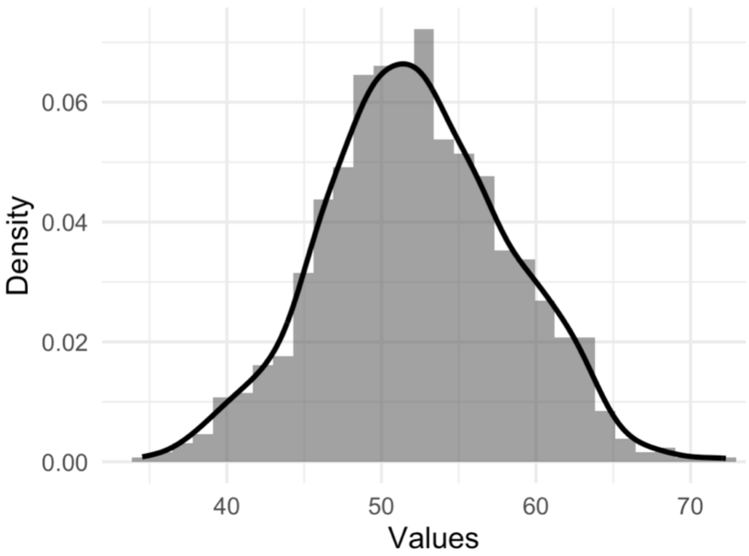


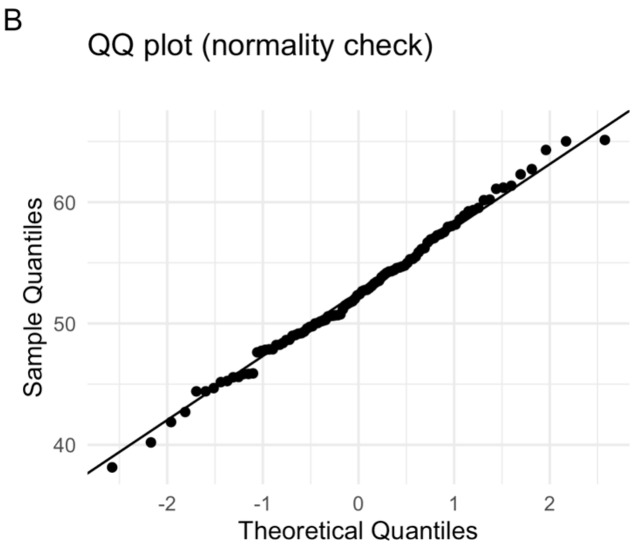
```
> x
  [1] 48.63715 50.61894 61.35225 52.42305 52.77573 62.29039 54.76550
  [8] 44.40963 47.87888 49.32603 59.34449 54.15888 54.40463 52.66410
 [15] 48.66495 62.72148 54.98710 40.20030 56.20814 49.16325 45.59306
 [22] 50.69215 45.84397 47.62665 48.24976 41.87984 57.02672 52.92024
 [29] 45.17118 59.52289 54.55879 50.22957 57.37075 57.26880 56.92949
 [36] 56.13184 55.32351 51.62853 50.16422 49.71717 47.83176 50.75250
 [43] 44.40762 65.01374 59.24777 45.26135 49.58269 49.20007 56.67979
 [50] 51.49979 53.51991 51.82872 51.74278 60.21161 50.64537 61.09882
 [57] 42.70748 55.50768 52.74313 53.29565 54.27784 48.98606 50.00076
 [64] 45.88855 45.56925 53.82117 54.68926 52.31803 57.53360 64.30051
 [71] 49.05381 38.14499 58.03443 47.74480 47.87195 58.15343 50.29136
 [78] 44.67569 53.08782 51.16665 52.03459 54.31168 49.77604 55.86626
 [85] 50.67708 53.99069 58.58103 54.61109 50.04441 58.89285 57.96102
 [92] 55.29038 53.43239 48.23256 60.16391 48.39844 65.12400 61.19566
 [99] 50.58580 45.84147
```

Null hypothesis (H_0): The data are normally distributed.

Alternative hypothesis (H_1): The data are not normally distributed.

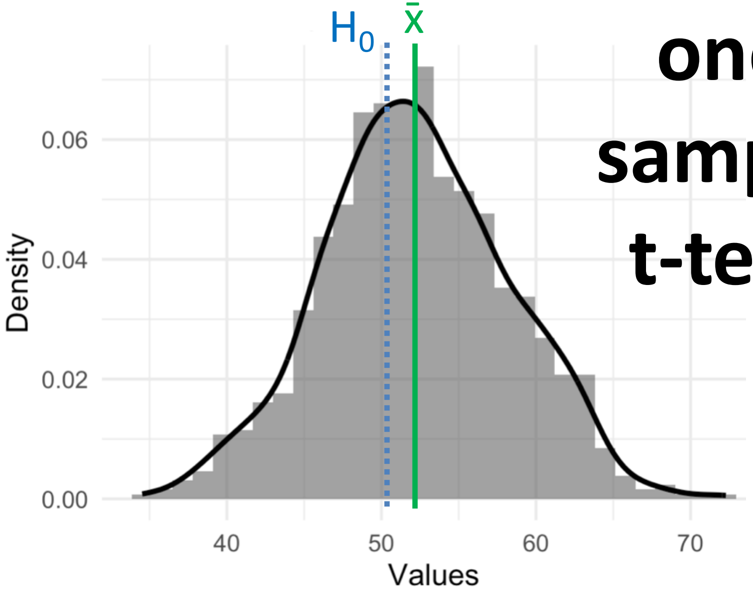
```
> shapiro.test(x)

	Shapiro-Wilk normality test

data:  x
W = 0.99388, p-value = 0.9349
```





one sample t-test

```
> x
  [1] 48.63715 50.61894 61.35225 52.42305 52.77573 62.29039 54.76550
  [8] 44.40963 47.87888 49.32603 59.34449 54.15888 54.40463 52.66410
 [15] 48.66495 62.72148 54.98710 40.20030 56.20814 49.16325 45.59306
 [22] 50.69215 45.84397 47.62665 48.24976 41.87984 57.02672 52.92024
 [29] 45.17118 59.52289 54.55879 50.22957 57.37075 57.26880 56.92949
 [36] 56.13184 55.32351 51.62853 50.16422 49.71717 47.83176 50.75250
 [43] 44.40762 65.01374 59.24777 45.26135 49.58269 49.20007 56.67979
 [50] 51.49979 53.51991 51.82872 51.74278 60.21161 50.64537 61.09882
 [57] 42.70748 55.50768 52.74313 53.29565 54.27784 48.98606 50.00076
 [64] 45.88855 45.56925 53.82117 54.68926 52.31803 57.53360 64.30051
 [71] 49.05381 38.14499 58.03443 47.74480 47.87195 58.15343 50.29136
 [78] 44.67569 53.08782 51.16665 52.03459 54.31168 49.77604 55.86626
 [85] 50.67708 53.99069 58.58103 54.61109 50.04441 58.89285 57.96102
 [92] 55.29038 53.43239 48.23256 60.16391 48.39844 65.12400 61.19566
 [99] 50.58580 45.84147
```



Null hypothesis (H_0): The population mean = some value (μ_0)

Alternative hypothesis (H_1): The population mean ≠ μ_0 (two-sided), or > / < (one-sided)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$df = n - 1$$

```
> t.test(x, mu = 50)

        One Sample t-test

data:  x
t = 4.8026, df = 99, p-value = 5.569e-06
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 51.74424 54.20026
sample estimates:
mean of x
 52.97225
```

$\bar{x}$ = sample mean

μ_0 = hypothesized mean

s = sample standard deviation

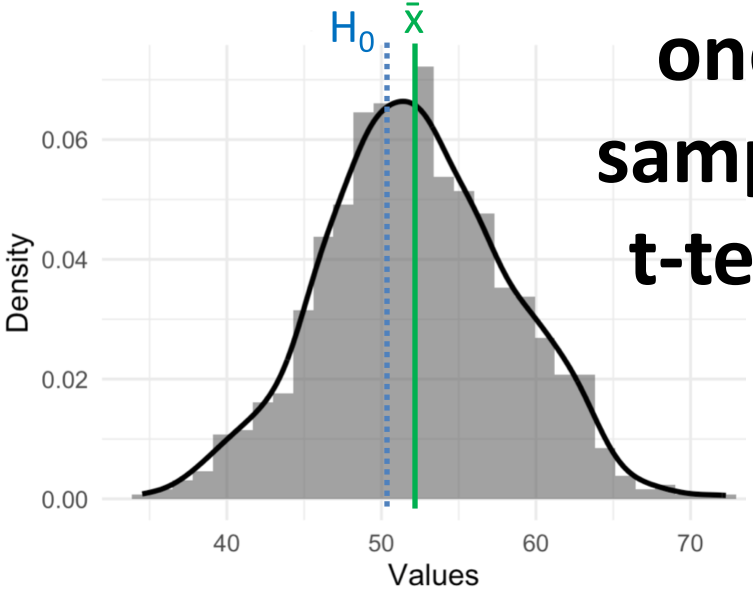
n = sample size

```
> t<-(mean(x)-50)/(sd(x)/sqrt(length(x))) ##t
> t
[1] 4.802567
> df<-length(x)-1
> pval <- 2 * (1 - pt(abs(t), df))
> pval
[1] 5.568592e-06
```

```
> x
  [1] 48.63715 50.61894 61.35225 52.42305 52.77573 62.29039 54.76550
  [8] 44.40963 47.87888 49.32603 59.34449 54.15888 54.40463 52.66410
 [15] 48.66495 62.72148 54.98710 40.20030 56.20814 49.16325 45.59306
 [22] 50.69215 45.84397 47.62665 48.24976 41.87984 57.02672 52.92024
 [29] 45.17118 59.52289 54.55879 50.22957 57.37075 57.26880 56.92949
 [36] 56.13184 55.32351 51.62853 50.16422 49.71717 47.83176 50.75250
 [43] 44.40762 65.01374 59.24777 45.26135 49.58269 49.20007 56.67979
 [50] 51.49979 53.51991 51.82872 51.74278 60.21161 50.64537 61.09882
 [57] 42.70748 55.50768 52.74313 53.29565 54.27784 48.98606 50.00076
 [64] 45.88855 45.56925 53.82117 54.68926 52.31803 57.53360 64.30051
 [71] 49.05381 38.14499 58.03443 47.74480 47.87195 58.15343 50.29136
 [78] 44.67569 53.08782 51.16665 52.03459 54.31168 49.77604 55.86626
 [85] 50.67708 53.99069 58.58103 54.61109 50.04441 58.89285 57.96102
 [92] 55.29038 53.43239 48.23256 60.16391 48.39844 65.12400 61.19566
 [99] 50.58580 45.84147
```

Null hypothesis (H_0): The population mean = some value (μ_0)

Alternative hypothesis (H_1): The population mean $\neq \mu_0$ (two-sided), or > / < (one-sided)



one sample t-test

```
> t.test(x, mu = 50)

        One Sample t-test

data:  x
t = 4.8026, df = 99, p-value = 5.569e-06
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 51.74424 54.20026
sample estimates:
mean of x
 52.97225
```

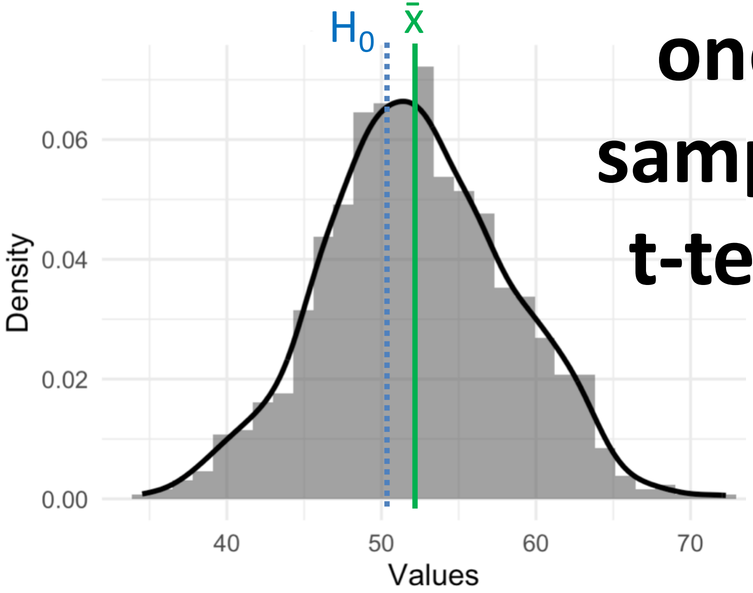
$$CI = \bar{x} \pm t^* \times SE$$

```
> alpha <- 0.05
> df <- length(x) - 1
> t_star <- qt(1 - alpha/2, df)   # two-tailed critical value
> mean(x)-sd(x)/sqrt(length(x))*t_star
[1] 51.74424
> mean(x)+sd(x)/sqrt(length(x))*t_star
[1] 54.20026
```

```
> x
  [1] 48.63715 50.61894 61.35225 52.42305 52.77573 62.29039 54.76550
  [8] 44.40963 47.87888 49.32603 59.34449 54.15888 54.40463 52.66410
 [15] 48.66495 62.72148 54.98710 40.20030 56.20814 49.16325 45.59306
 [22] 50.69215 45.84397 47.62665 48.24976 41.87984 57.02672 52.92024
 [29] 45.17118 59.52289 54.55879 50.22957 57.37075 57.26880 56.92949
 [36] 56.13184 55.32351 51.62853 50.16422 49.71717 47.83176 50.75250
 [43] 44.40762 65.01374 59.24777 45.26135 49.58269 49.20007 56.67979
 [50] 51.49979 53.51991 51.82872 51.74278 60.21161 50.64537 61.09882
 [57] 42.70748 55.50768 52.74313 53.29565 54.27784 48.98606 50.00076
 [64] 45.88855 45.56925 53.82117 54.68926 52.31803 57.53360 64.30051
 [71] 49.05381 38.14499 58.03443 47.74480 47.87195 58.15343 50.29136
 [78] 44.67569 53.08782 51.16665 52.03459 54.31168 49.77604 55.86626
 [85] 50.67708 53.99069 58.58103 54.61109 50.04441 58.89285 57.96102
 [92] 55.29038 53.43239 48.23256 60.16391 48.39844 65.12400 61.19566
 [99] 50.58580 45.84147
```

Null hypothesis (H_0): The population mean = some value (μ_0)

Alternative hypothesis (H_1): The population mean ≠ μ_0 (two-sided), or > / < (one-sided)

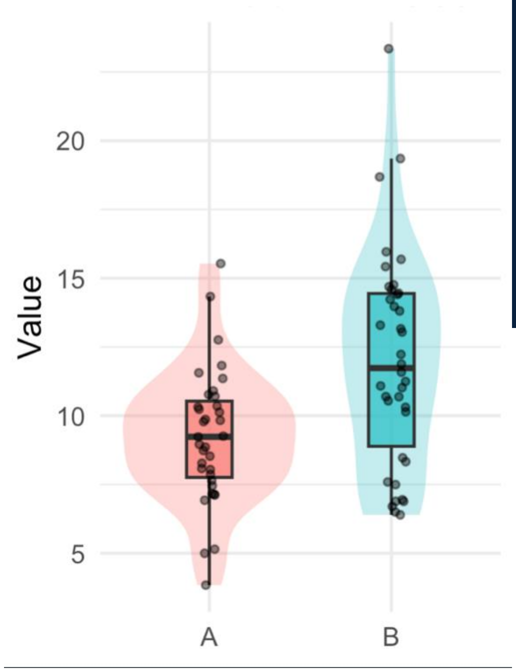


one sample t-test

```
> t.test(x, mu = 50)

	One Sample t-test

data:  x
t = 4.8026, df = 99, p-value = 5.569e-06
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 51.74424 54.20026
sample estimates:
mean of x
 52.97225
```

Two sample t-test

```
> df_two
   group     value
1      B 15.684490
2      B 10.696464
3      A  5.146352
4      A  7.145144
5      B 11.082309
6      A  9.231723
7      A  8.859320
8      A  9.786076
9      B  6.883855
10     A 12.756990
11     A 10.706160
12     B  6.698663
13     A  4.996174
14     B 14.589816
15     B 11.246168
16     A  7.645287
17     B 13.043797
18     B  7.589457
```




```
> tt_two <- t.test(value ~ group, data = df_two)
> tt_two

        Welch Two Sample t-test

data:  value by group
t = -3.6613, df = 62.644, p-value = 0.0005172
alternative hypothesis: true difference in means between group
A and group B is not equal to 0
95 percent confidence interval:
 -4.280188 -1.257436
sample estimates:
mean in group A mean in group B
       9.243104       12.011917
```

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

$\bar{x}_1, \bar{x}_2$ = sample means

s_1^2, s_2^2 = sample variances

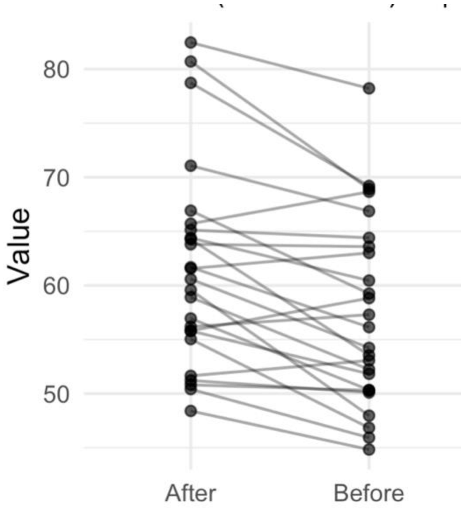
n_1, n_2 = sample sizes

Alternate syntax note:

```
t.test(x1, x2, var.equal = FALSE)
```

Paired t-test

```
> df
      after    before
1  65.70457 68.65839
2  56.94951 50.33579
3  59.58114 47.96439
4  56.18829 57.29763
5  63.80180 63.58607
6  55.03589 46.83742
7  82.45742 78.21918
8  55.79342 51.82677
9  66.91447 59.24740
10 50.41310 45.92759
11 61.68081 56.14247
12 61.56274 63.00493
13 64.39138 60.44929
14 55.81577 58.82193
15 51.21642 50.10190
```

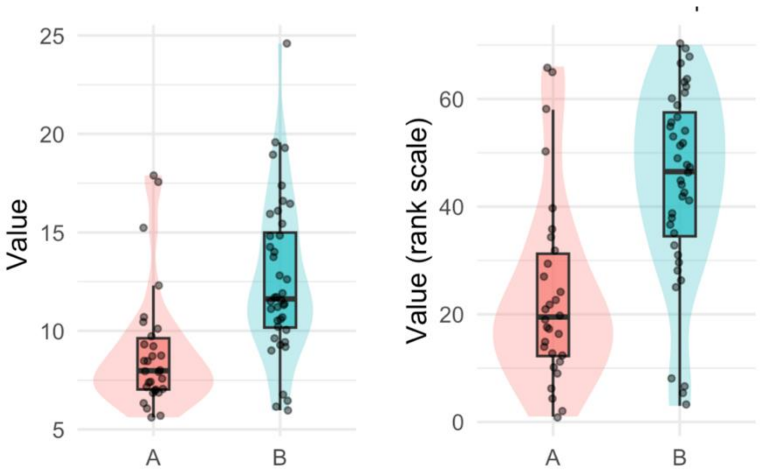



```
> df<-data.frame(after, before)
> tt_paired <- t.test(df$after, df$before, paired = TRUE)
> tt_paired

        Paired t-test

data:  df$after and df$before
t = 4.5693, df = 24, p-value = 0.0001242
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 2.231368 5.907615
sample estimates:
mean difference
       4.069491
```

(Same as one sample t-test on difference)

```
> t.test(df$after- df$before)

        One Sample t-test

data:  df$after - df$before
t = 4.5693, df = 24, p-value = 0.0001242
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.231368 5.907615
sample estimates:
mean of x
 4.069491
```

```
> df_wilx
   group     value
1      B  5.960864
2      A  5.607224
3      B  6.446158
4      B  9.428198
5      A  7.946408
6      B 11.691042
7      A  9.315445
8      B 16.458500
9      B 10.593346
10     A  8.711629
11     A  6.969304
12     B 10.044467
13     B 16.091317
14     A  8.001500
```



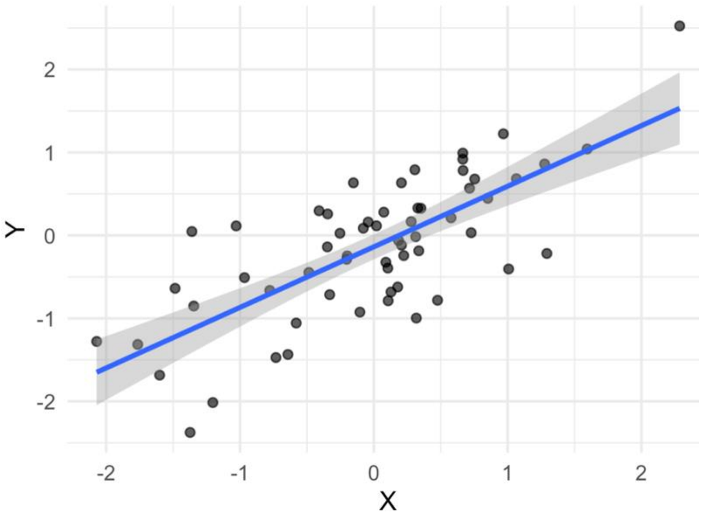
Wilcoxon test

(“*non parametric t-test*”)

```
> wx_rs <- wilcox.test(value ~ group, data = df_wilx)
> wx_rs

	Wilcoxon rank sum exact test

data:  value by group
W = 249, p-value = 1.596e-05
alternative hypothesis: true location shift is not equal to 0
```

Pearson correlation

```
> df
           xC          yC
1  -0.34778693 -0.13858920
2   0.66671889  0.78298571
3   0.66478866  0.99223687
4  -1.60079686 -1.68439178
5  -0.20319713 -0.28725030
6  -0.15295493  0.63394353
7   0.07441540  0.28013303
8   1.59382310  1.04195056
9  -2.07130493 -1.27841449
10  0.20650953 -0.11749842
11 -0.10440737 -0.92468670
12 -1.48556503 -0.63764525
13 -0.19935287 -0.24739595
14 -1.02915229  0.11441361
```



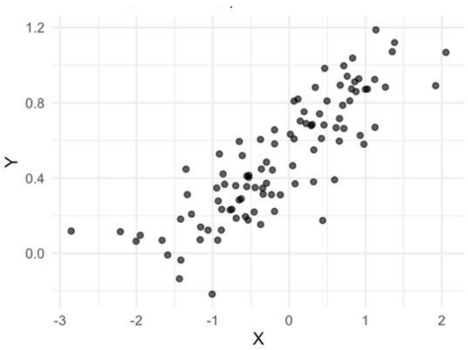

```
> ct_pear <- cor.test(xC, yC, method = "pearson")
> ct_pear

        Pearson's product-moment correlation

data:  xC and yC
t = 8.2438, df = 58, p-value = 2.393e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5908006 0.8330748
sample estimates:
      cor
0.7345317
```

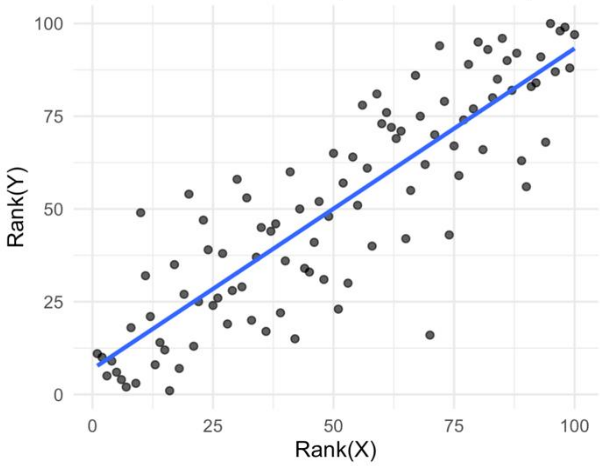
Spearman correlation (non-parametric)

```
> cor_spear <- cor.test(df_cor_s$x, df_cor_s$y, method = "spearman")
> cor_spear

	Spearman's rank correlation rho

data:  df_cor_s$x and df_cor_s$y
S = 22610, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho 
0.8643264 
```

```
> df_cor_s
             x           y rank_x rank_y
1  -0.65299176 0.594404400     30     58
2   1.35050938 1.070701180     97     98
3  -0.52991148 0.402799808     37     44
4  -0.92826260 0.278389803     19     27
5  -0.36435588 0.448172637     43     50
6   0.14544443 0.702384925     60     73
7   0.30126605 0.684225086     64     71
8  -2.85561802 0.118934330      1     11
9   1.13652272 1.187260760     95    100
10  0.65905478 0.715439229     77     74
```

Ranked >>>





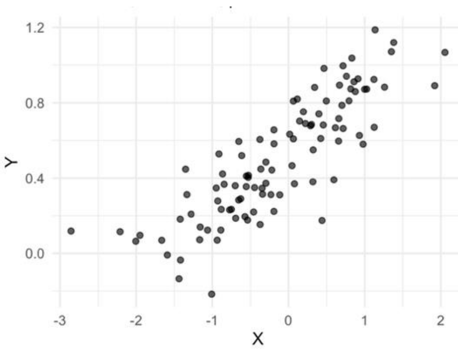
Spearman correlation (non-parametric)

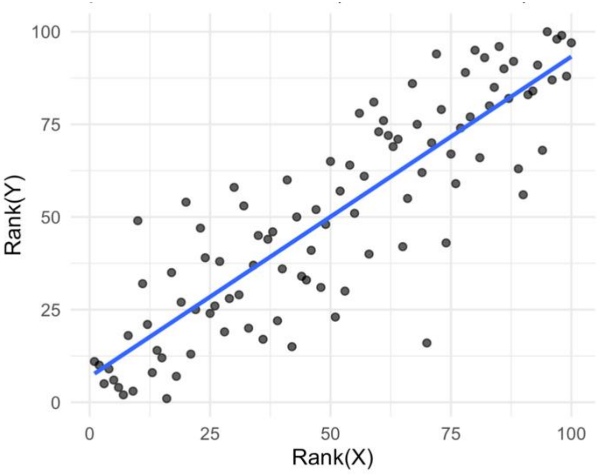
```
> cor_spear <- cor.test(df_cor_s$x, df_cor_s$y, method = "spearman")
> cor_spear

        Spearman's rank correlation rho

data:  df_cor_s$x and df_cor_s$y
S = 22610, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8643264
```

```
> cor.test(df_cor_s$rank_x, df_cor_s$rank_y, method = "pearson")

        Pearson's product-moment correlation

data:  df_cor_s$rank_x and df_cor_s$rank_y
t = 17.013, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8044853 0.9067980
sample estimates:
      cor
0.8643264
```

```
> df_cor_s
             x           y rank_x rank_y
1  -0.65299176 0.594404400     30     58
2   1.35050938 1.070701180     97     98
3  -0.52991148 0.402799808     37     44
4  -0.92826260 0.278389803     19     27
5  -0.36435588 0.448172637     43     50
6   0.14544443 0.702384925     60     73
7   0.30126605 0.684225086     64     71
8  -2.85561802 0.118934330      1     11
9   1.13652272 1.187260760     95    100
10  0.65905478 0.715439229     77     74
```

Ranked >>>

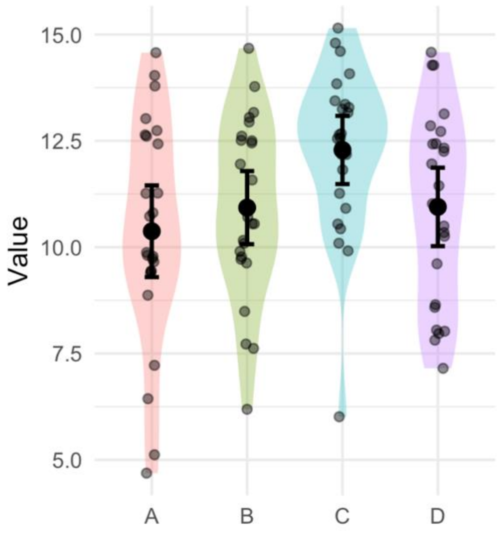




One way analysis of variance (ANOVA)

```
> df_aov
      group     value
1         A 13.790387
2         D  9.609565
3         C 11.265531
4         C 12.370461
5         A  9.442422
6         B 10.160362
7         C 14.605085
8         B 13.050642
9         B 10.934007
10        A  9.874572
```

```
> str(df_aov)
'data.frame':	100 obs. of  2 variables:
 $ group: Factor w/ 4 levels
"A","B","C","D": 1 4 3 3 1 2 3 2 2 1 ...
 $ value: num  13.79 9.61 11.27 12.37 9.44
```

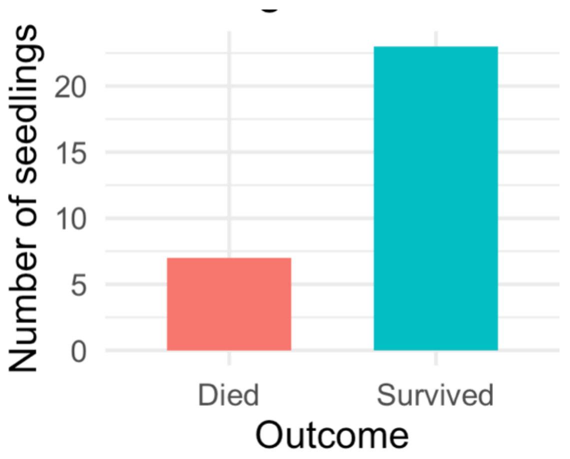



```
> anovafit <- aov(value ~ group, data = df_aov)
> summary(anovafit)
            Df Sum Sq Mean Sq F value Pr(>F)
group        3   49.5  16.487   3.308 0.0234 *
Residuals   96  478.5   4.985
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> fit
Call:
   aov(formula = value ~ group, data = df_aov)

Terms:
                  group Residuals
Sum of Squares  49.4604  478.5241
Deg. of Freedom       3        96

Residual standard error: 2.232628
Estimated effects may be unbalanced
```

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = value ~ group, data = df_aov)

$group
           diff        lwr       upr     p adj
B-A  0.55539226 -1.0956860 2.2064706 0.8154664
C-A  1.91041309  0.2593348 3.5614914 0.0165497
D-A  0.57157795 -1.0795003 2.2226563 0.8021317
C-B  1.35502083 -0.2960575 3.0060991 0.1462149
D-B  0.01618569 -1.6348926 1.6672640 0.9999939
D-C -1.33883514 -2.9899134 0.3122432 0.1540804
```

Binomial test

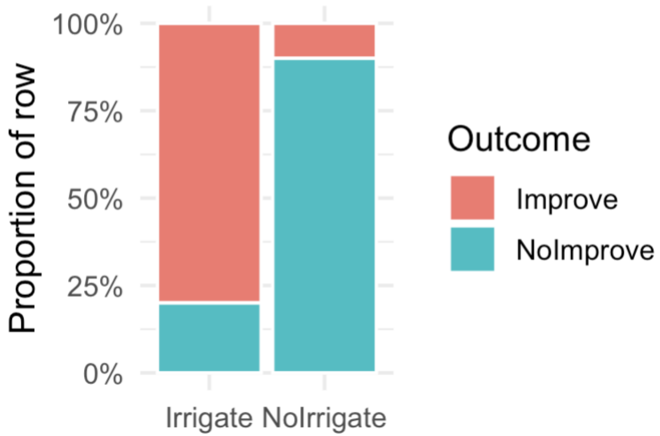
```r
# survival data
survived <- 23
total <- 30
```




```
> binom.test(survived, total, p = 0.5, alternative = "two.sided")

	Exact binomial test

data:  survived and total
number of successes = 23, number of trials = 30, p-value = 0.005223
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5771635 0.9006621
sample estimates:
probability of success
             0.7666667
```

Fishers Test

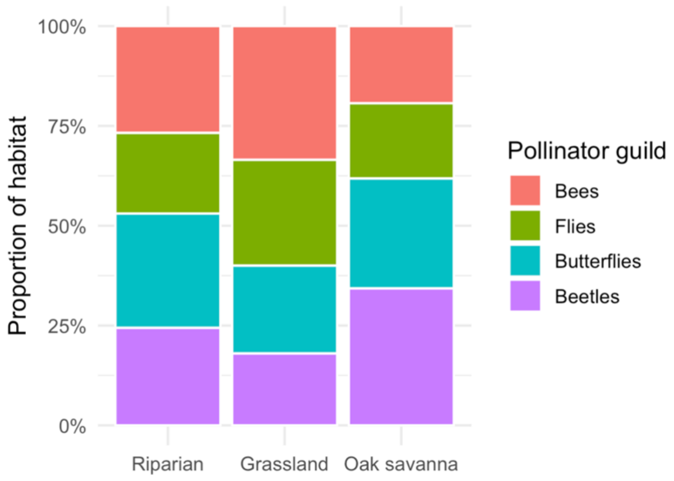
```
> tab_2x2
           Outcome
Treatment    Improve NoImprove
  Irrigate         8         2
  NoIrrigate       1         9
```




```
> fish_2x2 <- fisher.test(tab_2x2)
> fish_2x2

        Fisher's Exact Test for Count Data

data:  tab_2x2
p-value = 0.005477
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
    2.057999 1740.081669
sample estimates:
odds ratio
  27.32632
```

Chi-squared test

```
> tab_rc
             Guild
Habitat       Bees Flies Butterflies Beetles
  Riparian      57    43          61      52
  Grassland     67    53          44      36
  Oak savanna   40    39          57      71
```




```
> chi_rc  <- chisq.test(tab_rc)
> chi_rc

	Pearson's Chi-squared test

data:  tab_rc
X-squared = 23.561, df = 6, p-value = 0.0006289
```

Parametric tests

Assume your data follow a certain distribution (i.e. **normal**).

They use the **actual values** (means, variances).

Examples:

- **t-test** (compares means between groups)
- **Pearson correlation** (measures strength of a straight-line relationship)

✅ Advantages: More powerful if the assumptions hold (they can detect smaller effects).

⚠️ Drawback: Can give misleading results if assumptions are badly violated (e.g., very skewed data, outliers, small sample sizes).

Non-parametric tests

Do not assume normality.

They use **ranks or order of the data** instead of raw values.

Examples:

- **Wilcoxon tests** (rank-based alternative to the t-test)
- **Spearman correlation** (rank-based alternative to Pearson)

✅ Advantages: Work with skewed data, outliers, small samples, or ordinal data (e.g., ratings 1–5).

⚠️ Drawback: Less powerful than parametric tests when data actually *are* normal.

Correlation (Pearson vs. non-normal data)

- **Pearson correlation** assumes bivariate normality (both variables normally distributed with a linear relationship).
- **Violations** (skew, heavy tails, outliers):
  - Outliers can **inflate correlation** spuriously → **false positives** (detecting significance when no real relationship).
  - Nonlinear but monotonic relationships can give **low Pearson r** even if association is strong → **false negatives**.
- **Spearman correlation** (rank-based) is robust to non-normality and monotonic-but-nonlinear patterns.

Side-by-side

t-test vs. Wilcoxon

- *t-test*: compares group means, assumes data ≈ normal.
- *Wilcoxon*: compares group ranks/medians, no normality assumption.

Pearson vs. Spearman

- *Pearson*: checks linear relationship, sensitive to outliers, needs numeric data with normal-like distribution.
- *Spearman*: checks monotonic relationship (just whether values go up or down together), based on ranks, robust to outliers.

🔷 t-test (two groups)

- Assumes each group is normally distributed with equal variances (Welch's version relaxes equal variance).
- **Violations**:
  - With **large samples** (n > ~30 per group), the Central Limit Theorem helps → type I error (false positives) stays near 5%.
  - With **small samples**:
    - **Heavy tails / outliers**: can inflate type I error → **false positives**.
    - **Skewness**: reduces power → **false negatives** (harder to detect real differences).
- In practice: non-normality is more likely to cause **loss of power** (false negatives) unless outliers dominate, in which case **false positives** can rise.

🔷 ANOVA (more than 2 groups)

- Assumes **normal residuals** and **homogeneity of variances**.
- **Violations**:
  - **Unequal variances + unequal group sizes** → especially dangerous, can lead to **inflated type I error (false positives)**.
  - Pure **non-normality with equal variances** → mainly reduces power → **false negatives**.
- That's why we often check with **Levene's / Fligner–Killeen tests** for variance equality, and use **Kruskal–Wallis** when assumptions fail.

Rough summary and guide (green discussed in lecture):

Goal	Parametric test	Non-parametric alternative	When to use parametric	When to use non-parametric
Compare **one sample mean** vs. a value	One-sample **t-test**	**Wilcoxon signed-rank**	Data ≈ normal	Data skewed, ordinal, outliers
Compare **two independent groups**	Two-sample **t-test**	**Mann–Whitney U** (a.k.a. Wilcoxon rank-sum)	Normal-ish distribution, equal variances	Skewed, ordinal data, unequal variances, outliers
Compare **two paired samples** (before/after)	Paired **t-test**	**Wilcoxon signed-rank**	Normal-ish differences	Non-normal differences, small N, outliers
Compare **>2 groups**	**ANOVA**	**Kruskal–Wallis**	Normal-ish, equal variances	Skewed/ordinal, unequal variances
Compare **>2 repeated measures**	Repeated-measures **ANOVA**	**Friedman test**	Normal-ish, sphericity assumption	Non-normal, ordinal, violations of sphericity
Test **correlation**	**Pearson** correlation	**Spearman** correlation	Linear relationship, normal data	Monotonic (not necessarily linear), skewed/outliers, ordinal
Test **association in counts**	**Chi-squared** test	**Fisher's exact test**	Large sample, expected counts >5	Small sample, expected counts <5
Test a **sample proportion** vs. expected probability	Large-sample z-test for proportions (normal approx)	**Binomial test**	Large n (so normal approximation to binomial is valid, expected successes/failures > 5)	Small n, skewed proportions, exact inference needed