Biostat Office in RI

- **Director**: Lili Zhao

https://sph.umich.edu/faculty-profiles/zhao-lili.html

- **Statistician**: Karen, Julie, Ray, Judy, Chen
- **Health economics**: Mohammad
- **Data query**: Shirley

We provide statistical support for research activities at Corewell East!

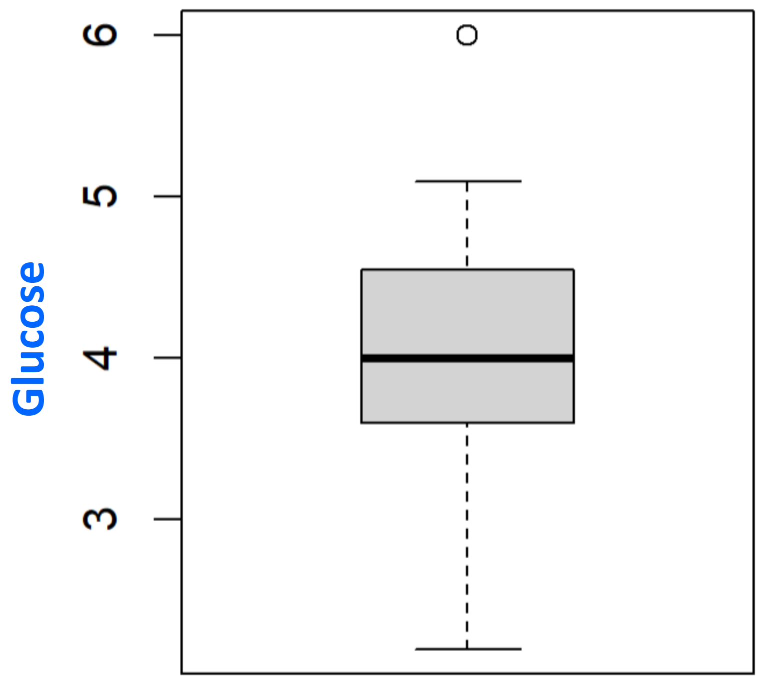
Review for the Last Lecture

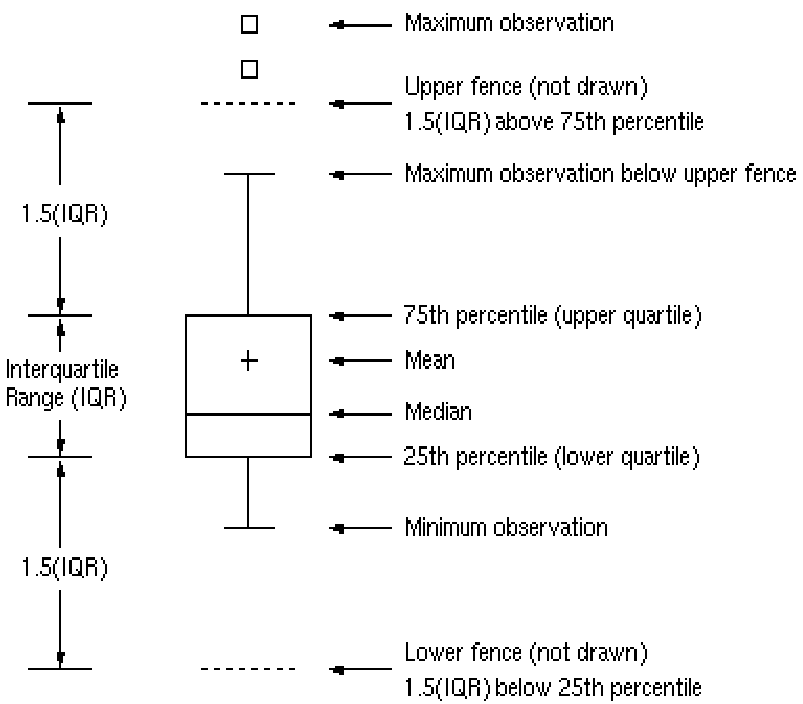
- Summary stats for a <u>categorical</u> variable
  - Frequency/percentage
  - Graph: bar plot, pie plot

- Summary stats for a <u>continuous</u> variable
  - Center (mean & median)
  - Spread (range, Q1, Q3, standard deviation)
  - Graph: boxplot, **histogram/density curve**

Review: Boxplot

R code:
Boxplot (Glucose, **outline** = TRUE)
*if **outline** is not true, the outliers are not drawn*



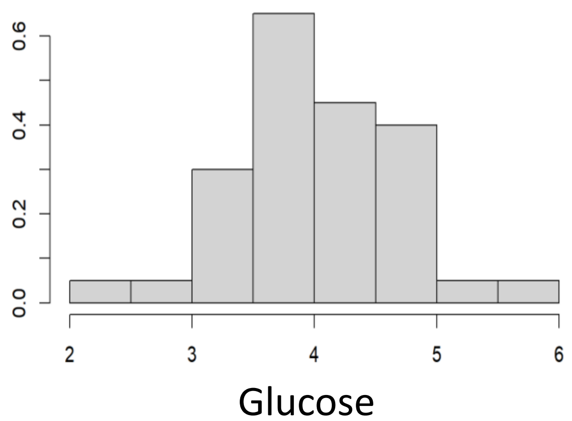


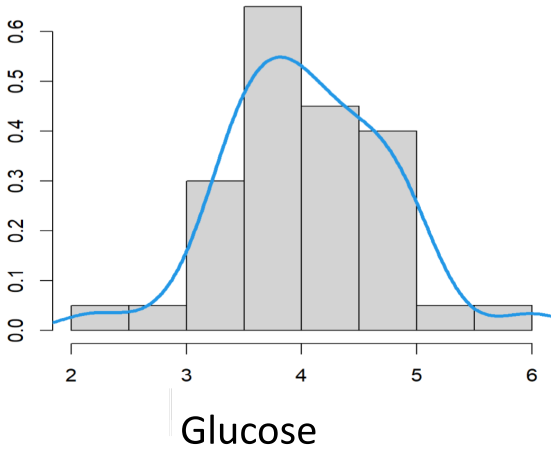
Histogram and Density Curve

Histogram and Density for Glucose

R code:
```
hist(Glucose, freq=FALSE)
```

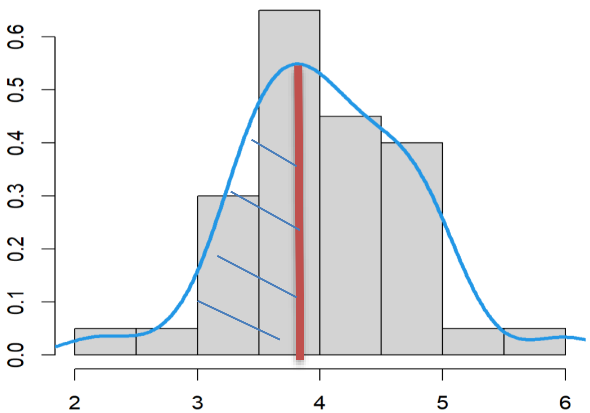
R code:
```
hist(Glucose, freq=FALSE)
lines(density(Glucose),col=4)
```





- **Density curve** is a smooth approximation of the irregular bars of a histogram
  - It describes the overall pattern of the data: center, shape and spread
  - It tells us the proportion of subjects in a certain range

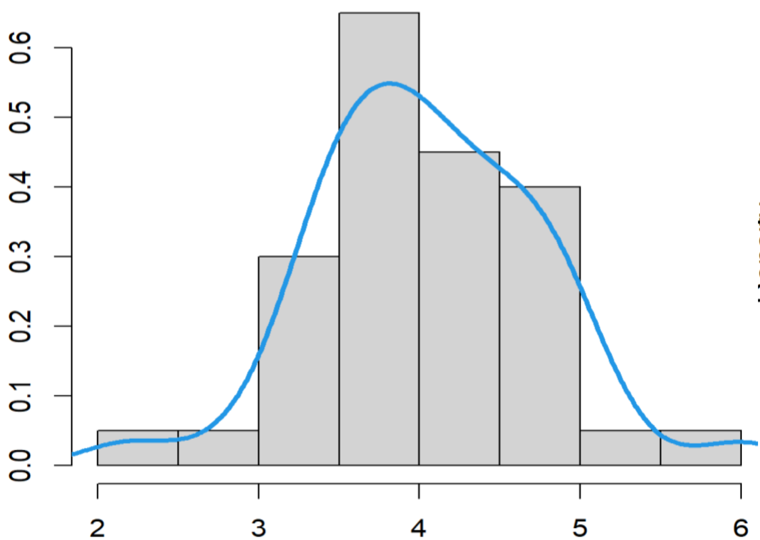
Density Plot for Glucose



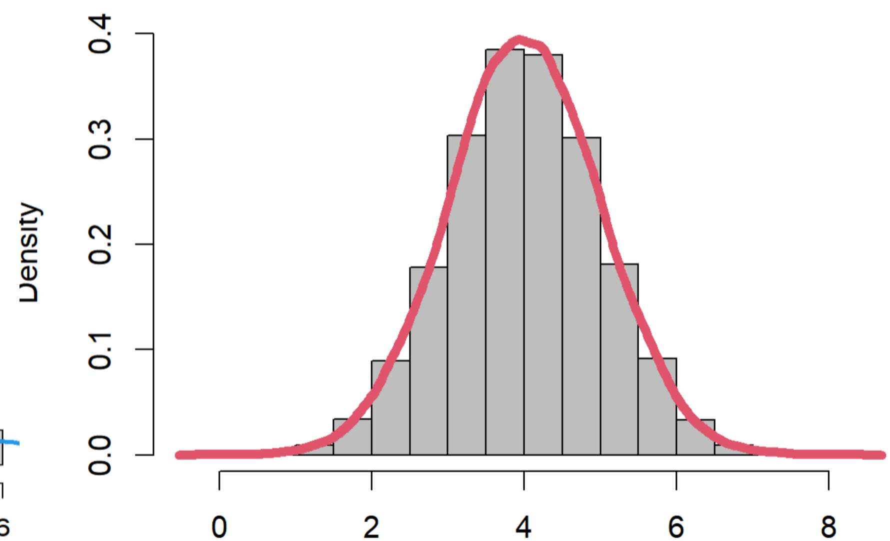
The area to the left of the red bar represents the proportion of subjects in the observed data that are less than or equal to 3.9 (low)

A Density Curve (sample vs population)

Glucose data from 40 medical students (**sample**)



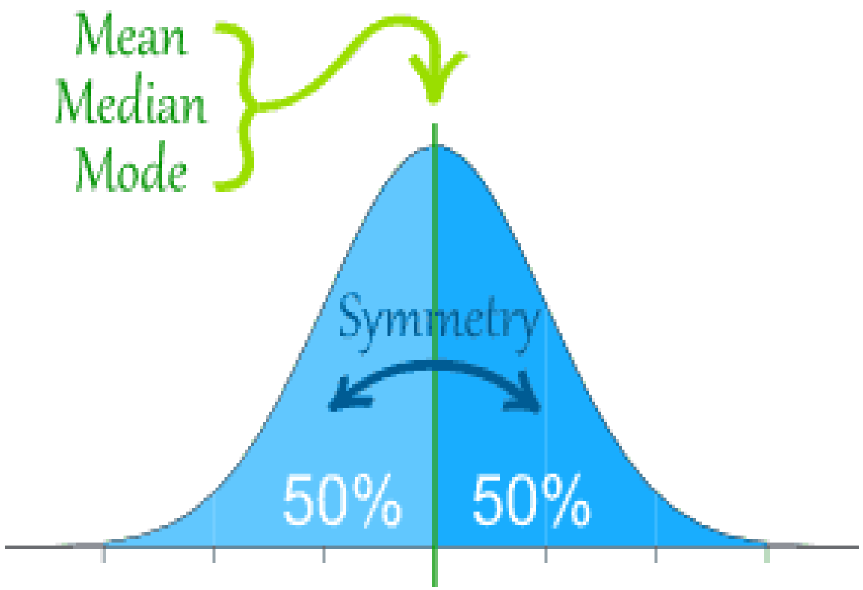
Glucose data from 1000 medical students (**population**)



Normal Distributions

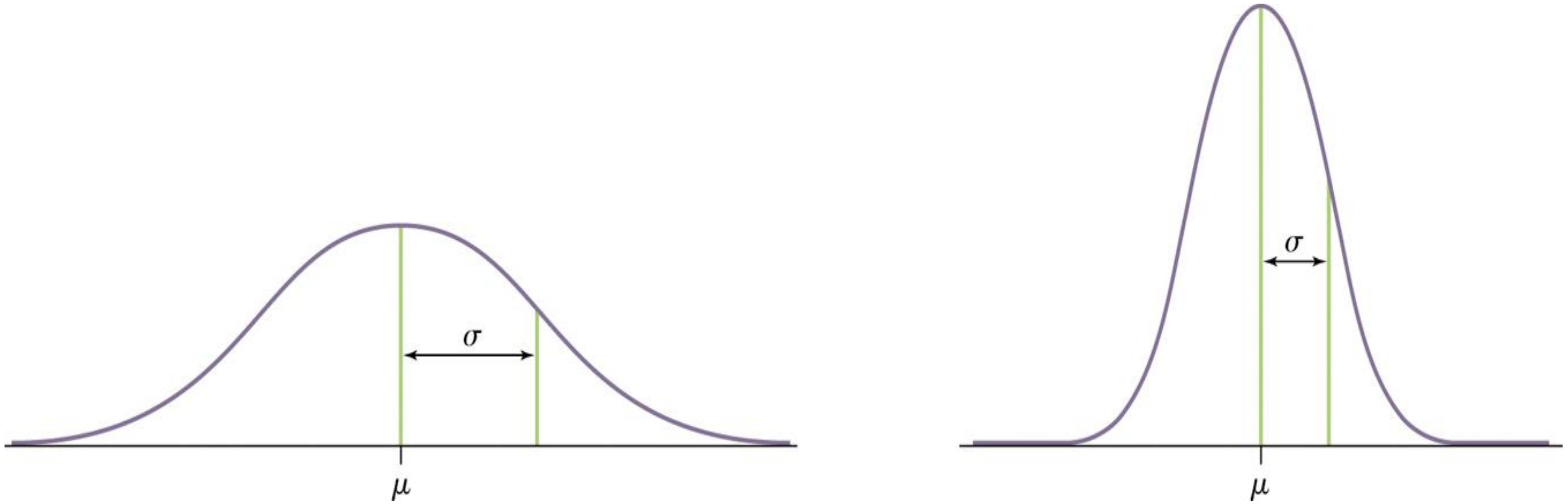
One particularly important class of density curves is the class of Normal curves, which describe Normal distributions.

- All Normal curves are symmetric, single-peaked, and bell-shaped.



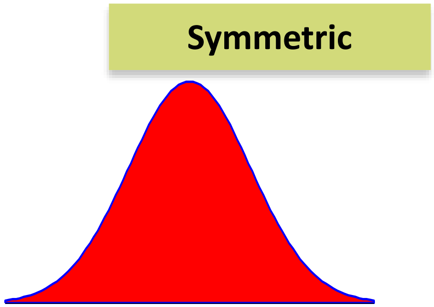
Normal Distributions

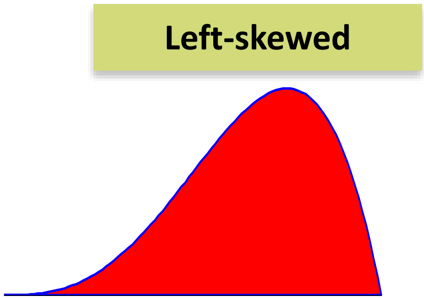
- Any particular Normal distribution **is completely specified by two numbers**: its mean μ and standard deviation σ
- We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu,\sigma)$

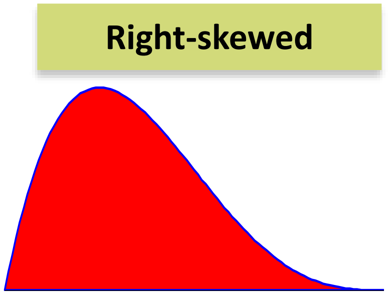


Shape

- A distribution is **symmetric** if the right and left sides of the graph are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the graph is much longer than the left side.
- It is **skewed to the left** if the left side of the graph is much longer than the right side.

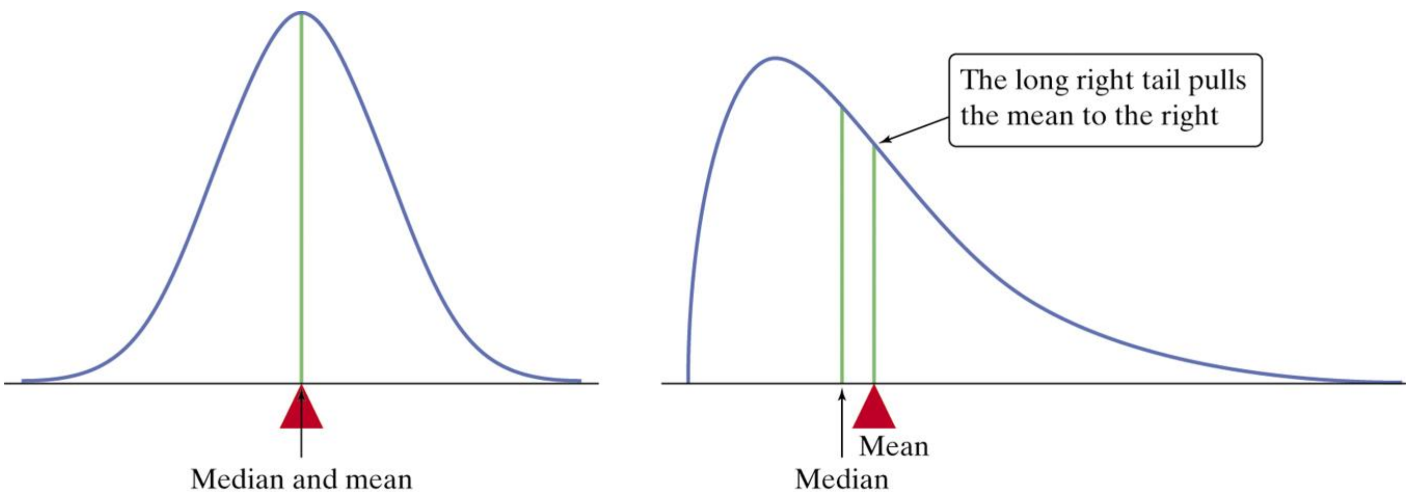






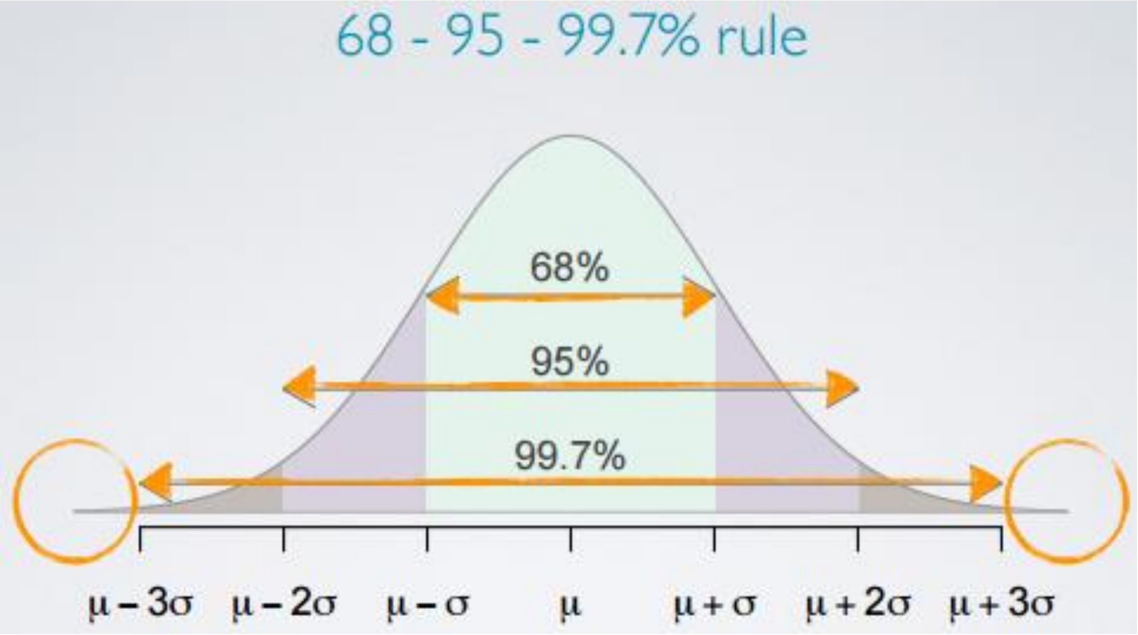
Mean and Median of a Density Curves

- The **median** of a density curve is the point that divides the area under the curve in half.
- The median and the mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.



Use Mean for a symmetric distribution and median for a skewed distribution

Normal Distributions



In the Normal distribution with mean ***μ*** and standard deviation ***σ***:

- Approximately **68%** of the observations fall within σ of μ.
- Approximately **95%** of the observations fall within 2σ of μ.
- Approximately **99.7%** of the observations fall within 3σ of μ.

Motivating Example

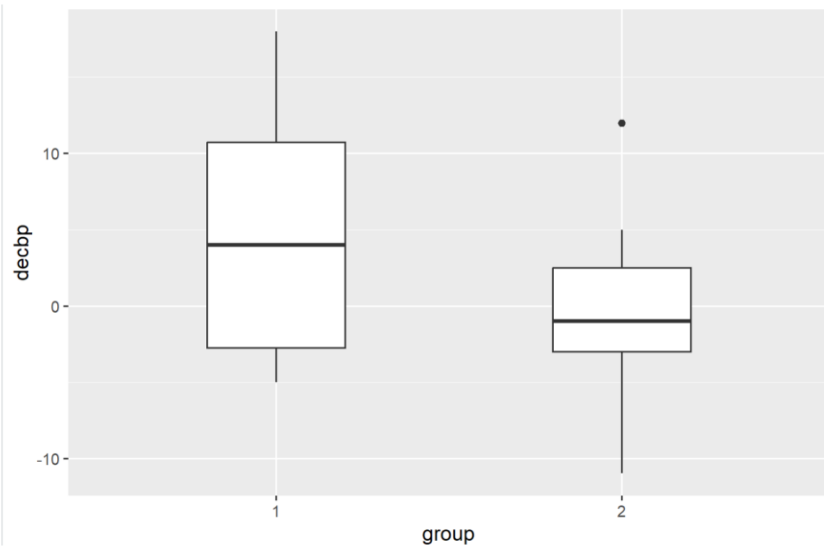
21 subjects were randomly assigned to two groups: 10 of them received a calcium supplement for 12 weeks, while the control group of 11 men received a placebo pill that looked identical. The response variable is the decrease in systolic (top number) blood pressure for a subject after 12 weeks, in millimeters of mercury.
The **goal** is to find out whether the calcium has effect on the systolic blood pressure.

Data:

Group 1 (calcium):	7	−4	18	17	−3	−5	1	10	11	−2	
Group 2 (placebo):	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

Obtain Estimates from Two Groups

```
ggplot(data=cal, aes(x=group, y=decbp)) +
geom_boxplot(width=0.4)
```




```
tapply(cal$decbp, cal$group, summary)
```

```
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -5.00   -2.75    4.00    5.00   10.75   18.00

$`2`
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-11.0000  -3.0000  -1.0000  -0.2727   2.5000  12.0000
```

Is the difference statistically significant (not due to chance)?

Make Comparisons Using R

```
t.test(decbp[group==1], decbp[group==2])
```

```
        Two Sample t-test

data:  decbp[group == 1] and decbp[group == 2]
t = 1.6341, df = 19, p-value = 0.1187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.48077 12.02622
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

What is a p-value?

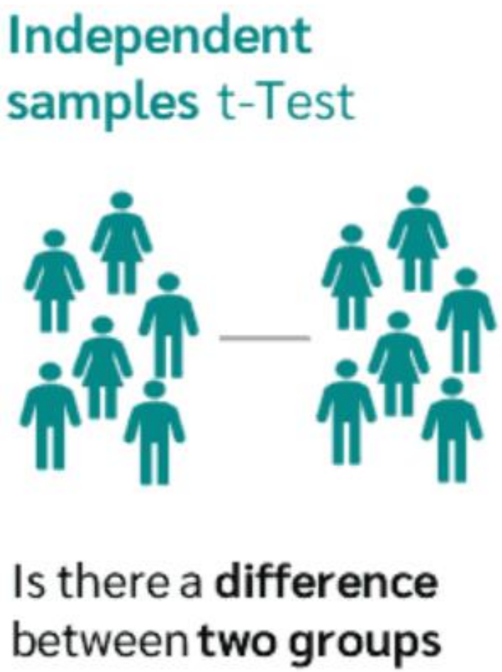
How to interpret the confidence interval?

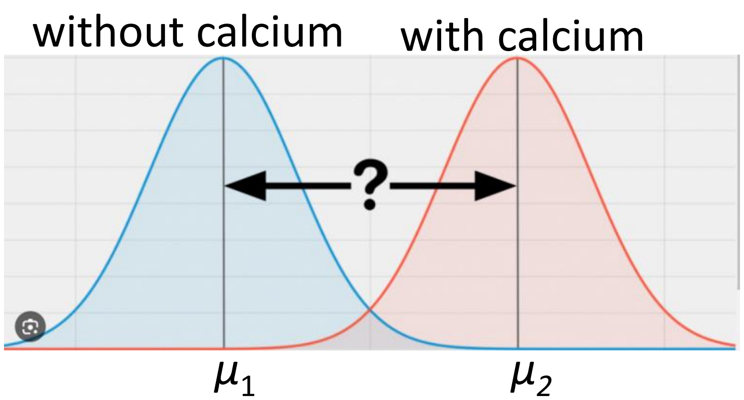
Statistical Inference

Confidence interval: uncertainty of the sample estimate

Tests of significance (p-value): assess evidence in the data about some claim concerning a population.

Compare Two Population

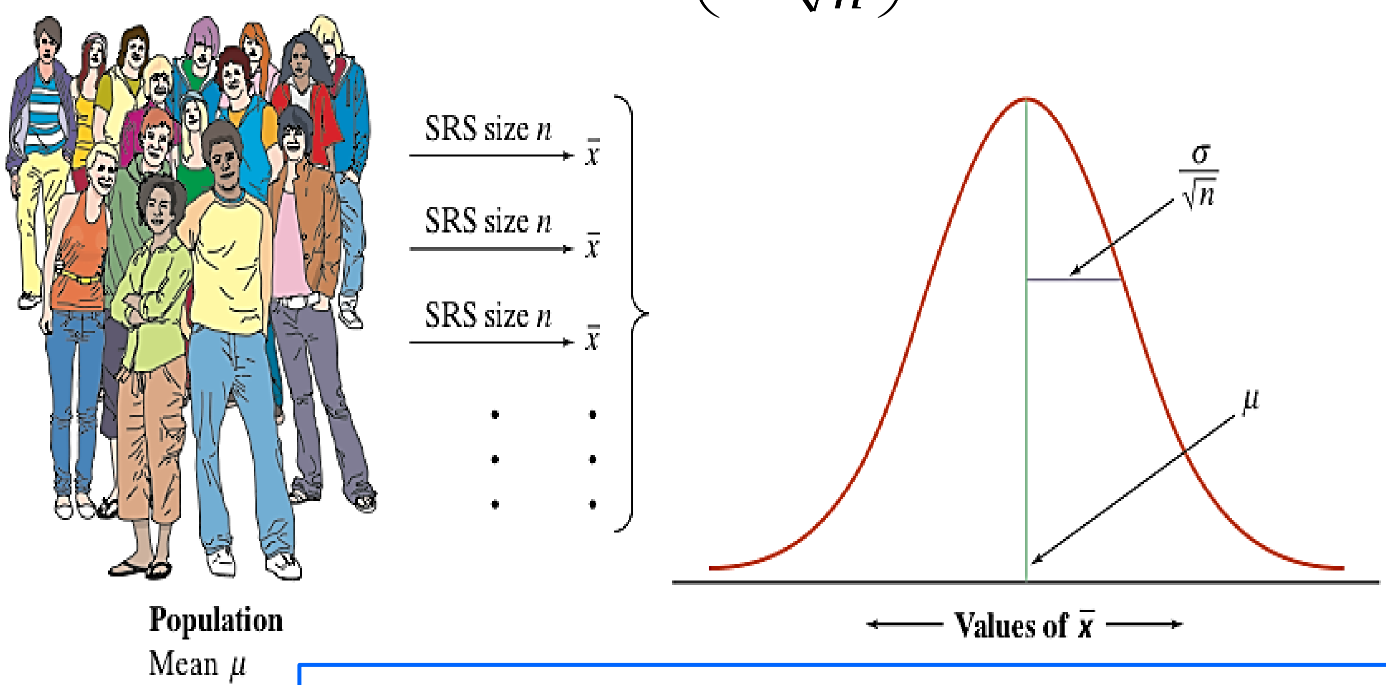




The Distribution of a Sample Mean (Theory)

If a population has a **Normal distribution** with mean **μ** and standard deviation of **σ** , then the sample mean also has a Normal distribution:

$$\bar{x} \text{ is } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



We use sample mean $\bar{x}$ to estimate population mean **μ**
The larger the sample size n, the more accurate to estimate **μ**

18

Test of Significance

Two-sample t Test

The claim is about a population parameter μ !

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

$$\Updownarrow$$

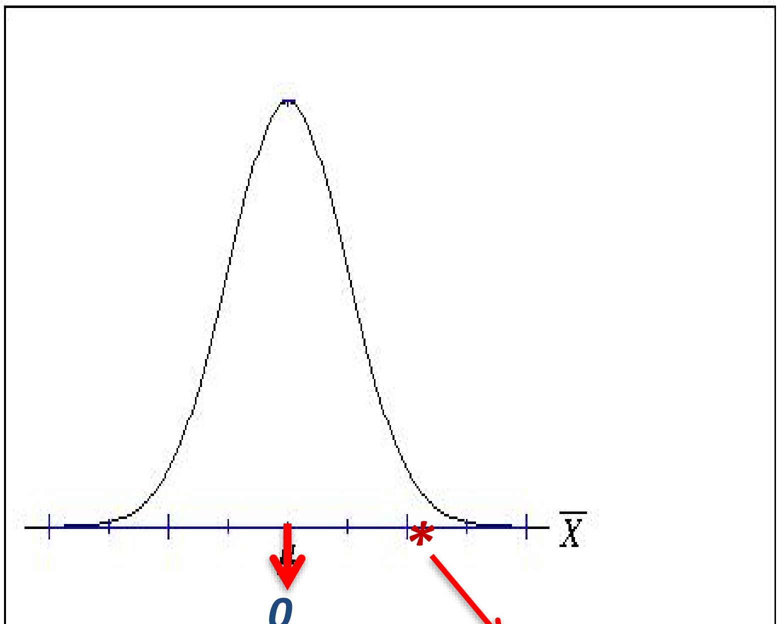
$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

Null hypothesis (H_0) is the claim, which you seek to *disprove*
Alternative hypothesis (H_a) is the claim, which we're trying to find evidence

- A **test of significance** is for comparing observed data with the null hypothesis and is designed to quantify the strength of the evidence against the null hypothesis
- We express the results of a significance test in terms of a probability, called the *p*-value, that measures how well the data and the null claim agree.

The Reasoning of p-value

Distribution of difference (diff) **under the Null**



Assuming that the difference is 0 (null), if the observed statistic is very unlikely (at the left or right tail), the null must be wrong.

observed diff (“standardized”) ⟺ **test statistic (ts)**

Test Statistic in General

A **test statistic** calculated from the sample data measures how far the data diverge from what we would expect if the null hypothesis H_0 were true.

$$ts = \frac{\text{estimate - hypothesized value in Null}}{\text{sd(estimate)}}$$

Large values of the statistic show that the data are not consistent with H_0.

When H_0 is true, we expect the estimate to be near the parameter value specified in H_0. Values of the estimate far from the parameter value specified by H_0 give evidence against H_0

T-test for Independent Samples

Goal: To infer the difference between two populations: μ_1- μ_2

Summary statistics

Group	Sample size	Sample mean	Sampl e SD
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

Which statistics can be used to estimate the population difference ?

$$\bar{x}_1 - \bar{x}_2 \Longrightarrow \mu_1 - \mu_2$$

Two-sample t Test (Theory)

$H_0 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 \neq 0$

- Assume **unequal** variances, i.e. $x_{i1} \sim N(\mu_1, \sigma_1)$ and $x_{i2} \sim N(\mu_2, \sigma_2)$

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

has **approximately** *t* distribution. We can use software to determine degrees of freedom

- Assume **equal** variances, i.e. $x_{i1} \sim N(\mu_1, \sigma)$ and $x_{i2} \sim N(\mu_2, \sigma)$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has the *t* distribution with degrees of freedom $df = n_1 + n_2 - 2$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

P-value

- **P-value** is the **probability**, under H_0 , that the **test statistic** would take **as extreme or more extreme values** than the one actually observed.

- If p-value is **small**, it serves as **an evidence against H_0**. It is unlikely under the null to get the data we already have. Then we reject the H_0 in favor of the alternative.

- Note that failing to reject the H_0 does NOT mean that we have clear evidence that H_0 is true.

Decision rule

- We need a **cut-off point** that we can compare our p-value to and draw a conclusion or make a decision.
- This cut-off point is the **significance level**. It is announced in advance and serves as a standard on how much evidence against H_0 we need to reject H_0. Usually denoted by $\boldsymbol{\alpha}$.
- Typical values of α: **0.05, 0.01**.

Statistical Significance

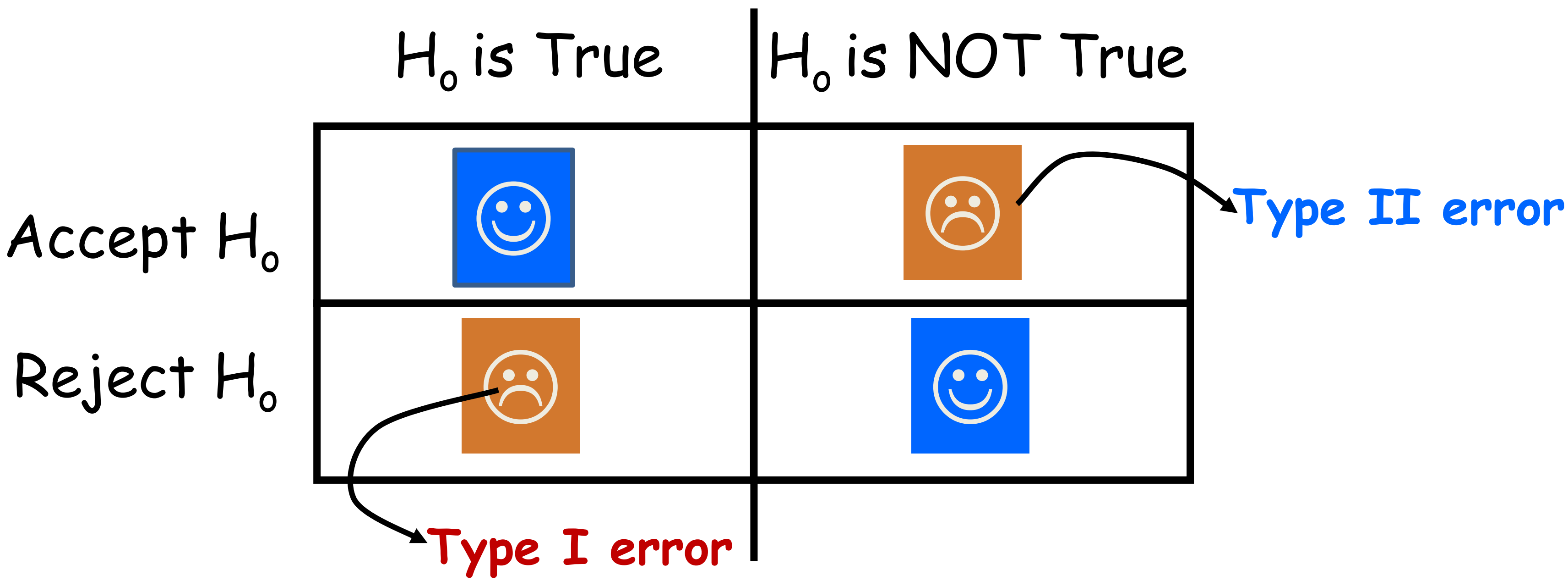
- When **p-value ≤ α**, we say that the data are **statistically significant** at level α i.e. we have significant evidence against the null hypothesis.

Note:

- data with a p-value of 0.02 are statistically significant at level 0.05, but not at level 0.01
- Failing to find evidence again H_0 (i.e. $p>\alpha$) can't show that H_0 is true

Two Errors in Making Decisions

Making Decisions: Test of Hypothesis



α = probability of Type I error (level of significance)
β = probability of Type II error
$1-\beta$ =Power

The Risks of Making Decisions

- Type I error. Falsely reject the null hypothesis. Significance level (**α**) is the chance of this happening.

e.g. setting α=0.05 means we allow 5% error of rejecting the null when null is true.

- Type II error. Fail to accept the alternative when it is true.

  (1-Type II error) is **power**, the chance to correctly reject the null hypothesis.

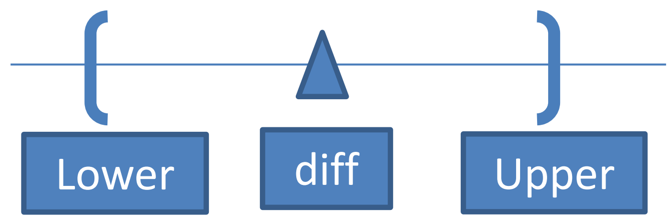
Recap for Decision Using P-value

Reject H_0 when the P-value is smaller than significance level α.

Do not reject otherwise.

Confidence Interval

Confidence Interval (CI) is **Point estimate ± Margin of error**



Interpretation:

CI is a range of values that is likely to contain the value of an unknown population parameter

For example, a 95% CI of (3, 10) suggests you can be 95% confident that the population mean is between 3 and 10.

Confidence Interval for Two-sample Means

Confidence Interval is **difference ± Margin of error**

Standard error

$$CI = (\bar{x}_1 - \bar{x}_2) \pm \underbrace{t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}_{\text{Margin of error}}$$

The <u>CI gets narrower</u> when:

- Standard deviation (s1 and s2) are smaller
- Sample size (n1 & n2) are larger

Two-sample t Test (Theory)

95% Confidence Interval :

- Assume **unequal** variances, i.e. $x_{i1} \sim N(\mu_1, \sigma_1)$ and $x_{i2} \sim N(\mu_2, \sigma_2)$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^{\ 2}}{n_1} + \frac{s_2^{\ 2}}{n_2}}$$

- Assume **equal** variances, i.e. $x_{i1} \sim N(\mu_1, \sigma)$ and $x_{i2} \sim N(\mu_2, \sigma)$

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

Two-sided Test and Confidence Intervals

A level α **two-sided** significance test rejects H_0: $\mu=\mu_0$ exactly when μ_0 falls outside a level $1-\alpha$ confidence interval for μ.

95% CI includes μ_0 $\Longleftrightarrow$ do not reject H_0 at $\alpha=0.05$

95% CI doesn't include μ_0 $\Longleftrightarrow$ reject H_0 at $\alpha=0.05$

Note: This is only true under a two-sided test

Revisit: Make Comparisons Using R

```
t.test(decbp[group==1], decbp[group==2])
```

P>0.05

95% CI include 0

```
        Two Sample t-test

data:  decbp[group == 1] and decbp[group == 2]
t = 1.6341, df = 19, p-value = 0.1187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.48077 12.02622
sample estimates:
 mean of x  mean of y
 5.0000000 -0.2727273
```

Conclusion: There is no statistically significant difference between the group with and without calcium in decreasing the systolic blood pressure (the difference is 5.27; 95% CI is from -1.48 to 12.03; p=0.12)