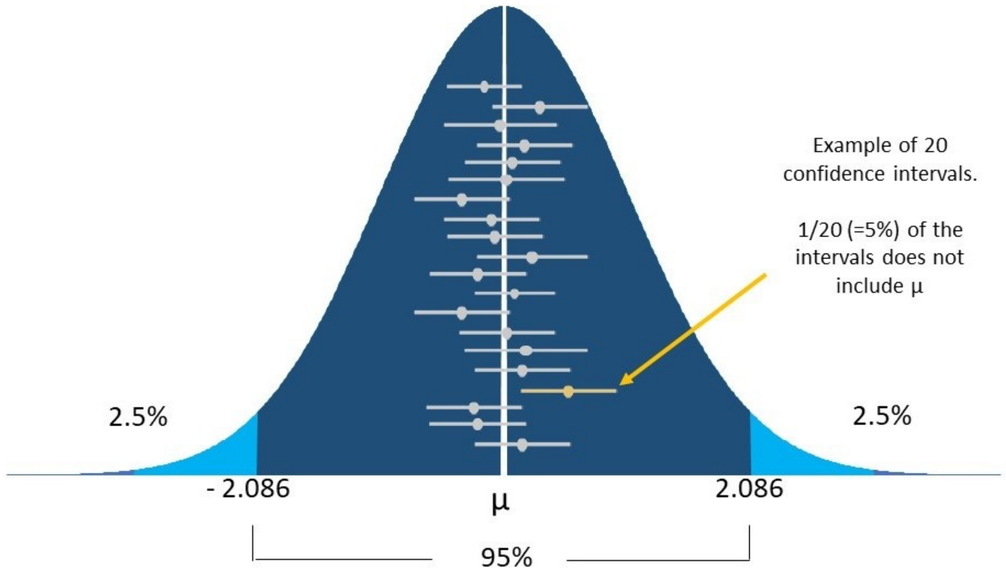
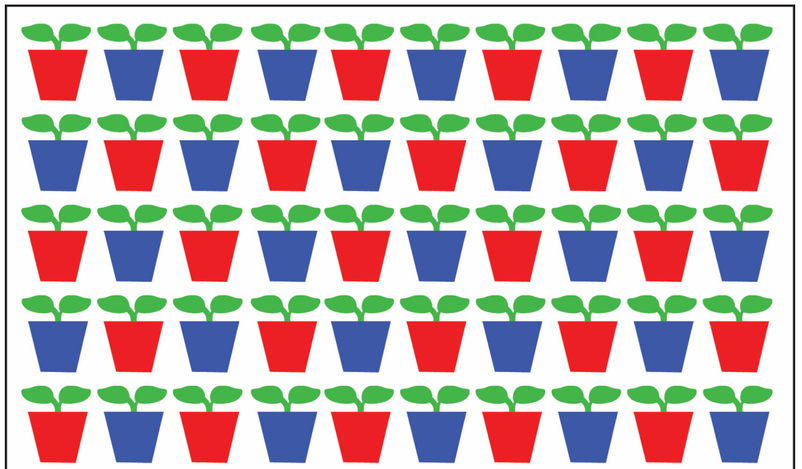
Foundational Statistics

Confidence Intervals / Experimental Design





One Measurement Per Pot	Three Measurements Per Pot
Total Sample Size (N) = 50 Experimental Units = 50 Treatments = 2 Treatment Size (n)= 25 Replication (r) = 25	Total Sample Size (N) = 150 Experimental Units = 50 Treatments = 2 Treatment Size (n)= 25 Replication (r) = 25

Common parameters and their sample-based estimate calculations

Table 2.1 | Common population parameters and sample statistics

Parameter	Statistic	Formula
Mean (μ)	$\bar{y}$	$\frac{\sum_{i=1}^{n} y_i}{n}$
Median	Sample median	$y_{(n+1)/2}$ if n odd $(y_{n/2} + y_{(n/2)+1})/2$ if n even
Variance (σ^2)	s^2	$\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$
Standard deviation (σ)	s	$\sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}}$
Median absolute deviation (MAD)	Sample MAD	$\text{median}[\lvert y_i - \text{median}\rvert]$
Coefficient of variation (CV)	Sample CV	$\frac{s}{\bar{y}} \times 100$
Standard error of $\bar{y}$ ($\sigma_{\bar{y}}$)	$s_{\bar{y}}$	$\frac{s}{\sqrt{n}}$
95% confidence interval for μ **(for sample sizes <= 30)**		$\bar{y} - t_{0.05(n-1)}\frac{s}{\sqrt{n}} \le \mu \le \bar{y} + t_{0.05(n-1)}\frac{s}{\sqrt{n}}$

Parametric confidence intervals for an estimated mean

1. **The "standard normal approximation" 95% CI**

$$95\%\ \text{CI} = \bar{x} \pm 1.96 * \frac{s}{\sqrt{n}}$$

2. **The "*t* approximation" CI for smaller samples (*n* <= 30)**

$$95\%\ \text{CI} = \bar{x} \pm t_{1-0.05/2} * \frac{s}{\sqrt{n}}$$

Value from *t* distribution, depends on *n*

In R:

```
> ## a small sample (n=25) from a normal distribution
> x_var <- rnorm(n=25, mean=10, sd=2)
>
> ## get 95% confidence interval based on t distribution
> t.test(x_var)$conf.int
[1]  9.481744 11.044291
attr(,"conf.level")
[1] 0.95
```

Bootstrap resampling (non-parametric) confidence intervals for an estimated mean

- Use the computer to take a random sample of individuals from the original data, with replacement.

- Calculate the estimate using the measurements in the bootstrap sample (step 1). The first **bootstrap replicate estimate**.

- Repeat steps 1 and 2 a large number of times (1000 times is reasonable).

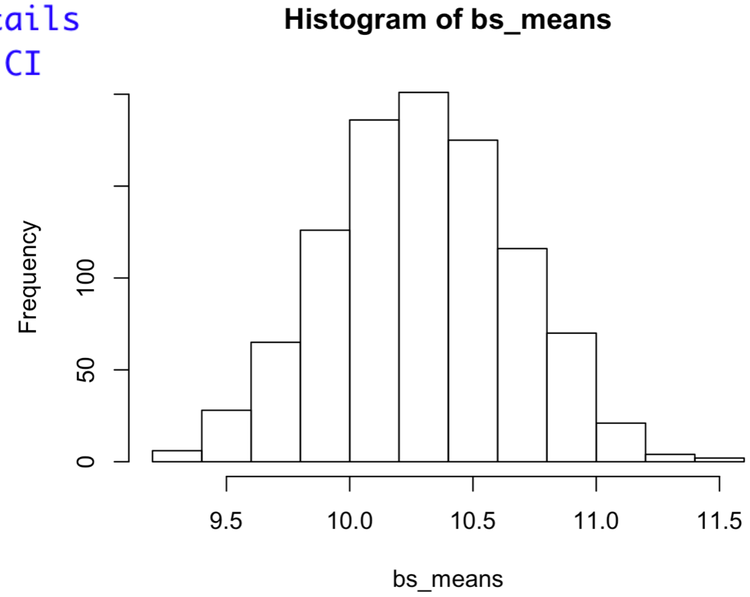
- Treat the distribution of resampled means like a probability distribution, and find the values of *x* that mark the bottom and top 2.5% of observations in the distribution. These values effectively approximate a **bootstrap confidence interval**.

Bootstrap resampling (non-parametric) confidence intervals for an estimated mean

In R:

```
> ## a small sample (n=25) from a normal distribution
> x_var <- rnorm(n=25, mean=10, sd=2)

> ## get 95% confidence interval based on resampling
> bs_reps <- replicate(1000, sample(x_var, replace=TRUE))
> bs_means <- apply(bs_reps, MARGIN=2, FUN=mean)
>
> hist(bs_means)
>
> ## use the quantile function to get x values at 0.025 tails
> ## these are the lower and upper boundaries of the 95% CI
> quantile(bs_means, probs=c(0.025,0.975))
     2.5%     97.5%
 9.558875 11.008505
```



Mean-variance relationship and CV

1. Variables with large means tend to have large variances, relative to otherwise similarly shaped distributions with smaller means.

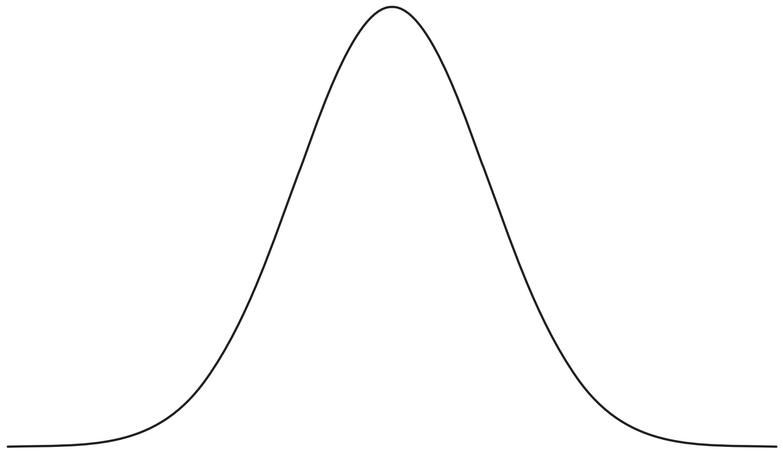
2. How do we make meaningful comparisons of variation for these variables that differ only by scale?

Yikes!

Coefficient of variation (CV)

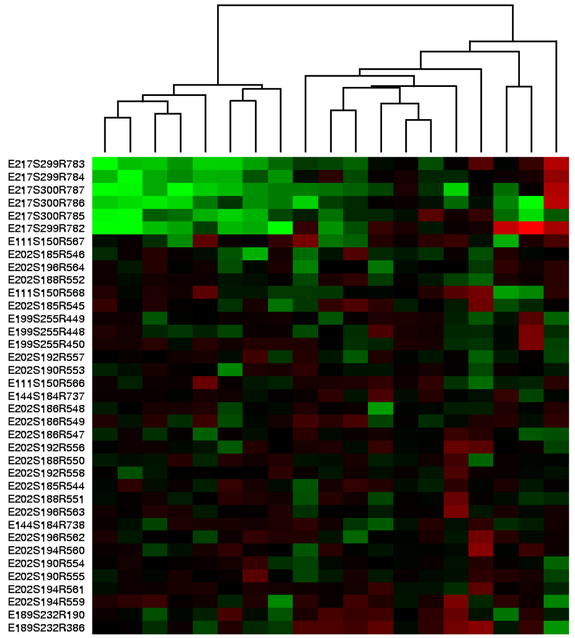
$$= \frac{s}{\bar{y}} \times 100$$

A word on the standard normal (*z*) distribution and ***z*-scores**



We can modify any normal distribution to have a **mean of 0** and a **standard deviation of 1** (aka a standard normal distribution) by changing each value into a ***z*-score**

$$z_i = \frac{(y_i - \bar{y})}{s}$$



e.g. rows= genes

Experimental/Study Design - Some initial considerations

1. **What are the main goals?**
   - Am I trying to estimate a parameter(s)?
   - Am I comparing groups? (e.g. testing a hypothesis)
   - If so, how many groups should I be comparing?

2. **Are there expectations from previous knowledge?**
   - e.g. pilot studies, the literature, etc.
   - These can help with handles on expected variance, etc.

3. **How many replicates am I going to need?**
   - Can do simulations to find out
   - Can do power analysis