4. Principles of statistical inference. The normal distribution.

Introduction to Medical Statistics

OUCRU, Ho Chi Minh City

Mar 23-27, 2026

Van Thuan Hoang

and the biostatistics crew

Program for this session

- The normal distribution
- Statistical inference
  - confidence intervals
  - hypothesis tests and p-values

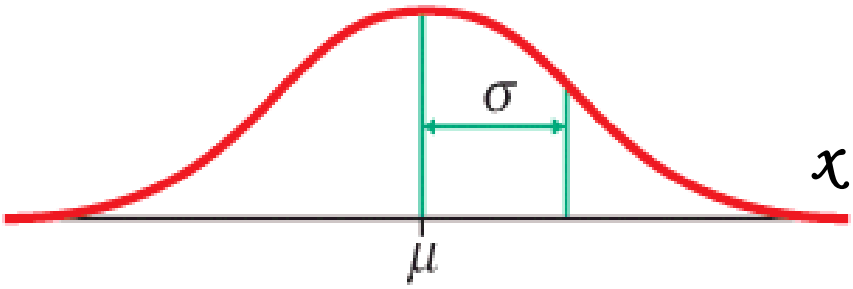
Probability distributions

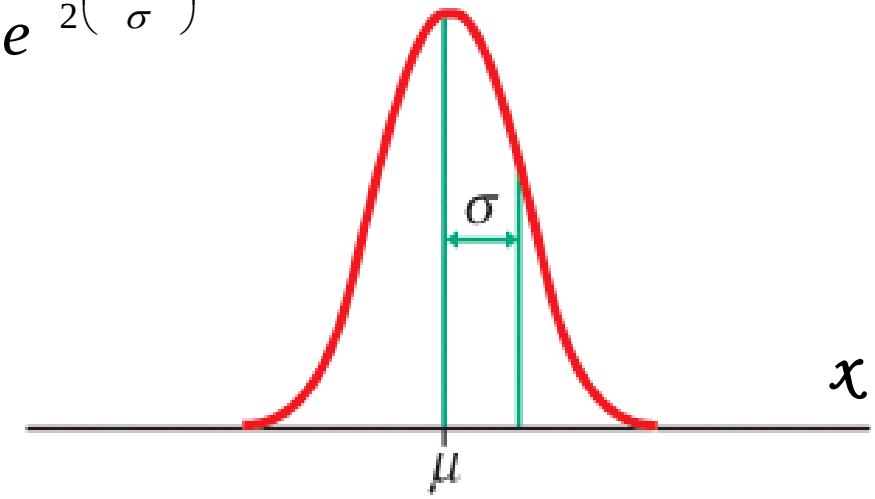
- Distribution of values of variable in sample (e.g. as displayed by the histogram): empirical distribution
- Distribution of values of variable in population: probability distribution
  - Often assumed to belong to some theoretical family of distributions, e.g.
    - Normal distribution
    - Log-normal distribution
    - Binomial distribution
    - Poisson distribution

Normal distribution

Normal – or Gaussian – distribution: family of symmetrical, bell shaped density curves defined by a mean μ (=median) and a standard deviation σ: $N(\mu,\sigma^2)$.

$$\text{Density function}: f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

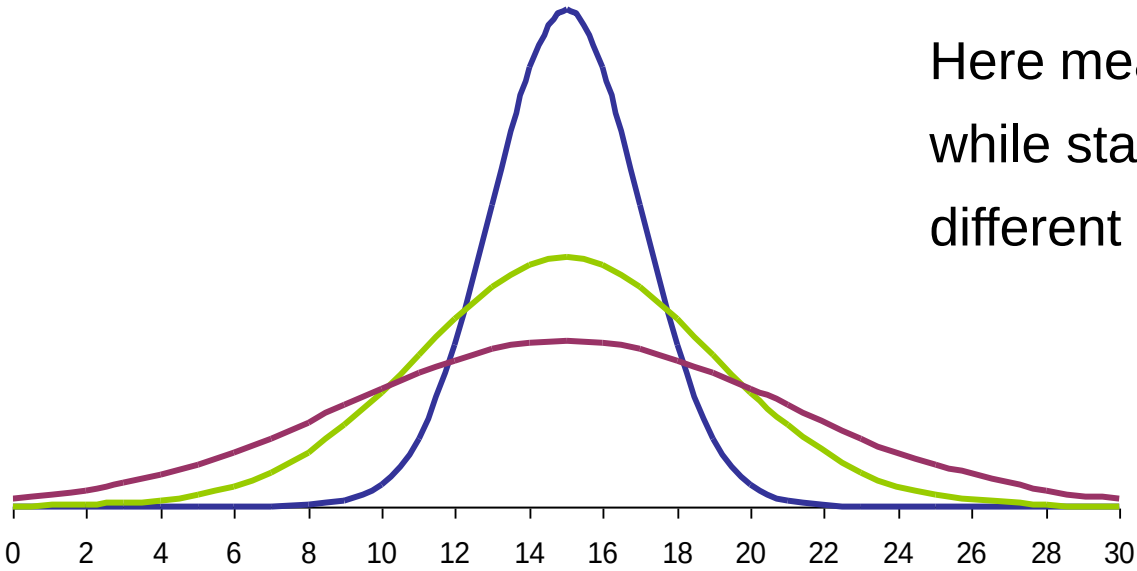




e = 2.71828… The base of the natural logarithm

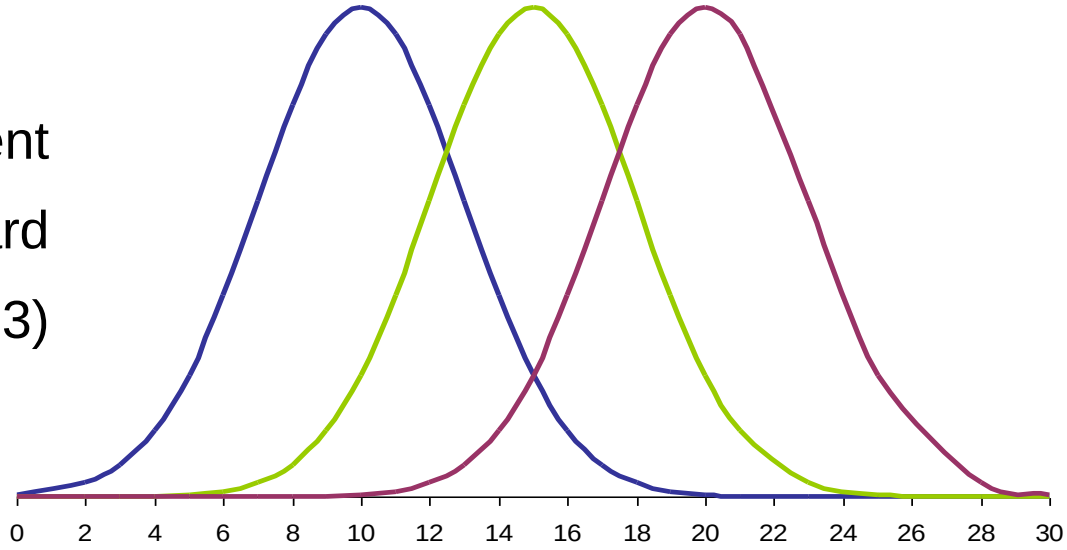
π = pi = 3.14159…

Density curves



Here means are the same ($\mu = 15$) while standard deviations are different ($\sigma = 2$, 4, and 6).

Here means are different ($\mu = 10$, 15, and 20) while standard deviations are the same ($\sigma = 3$)



Importance of the normal distribution

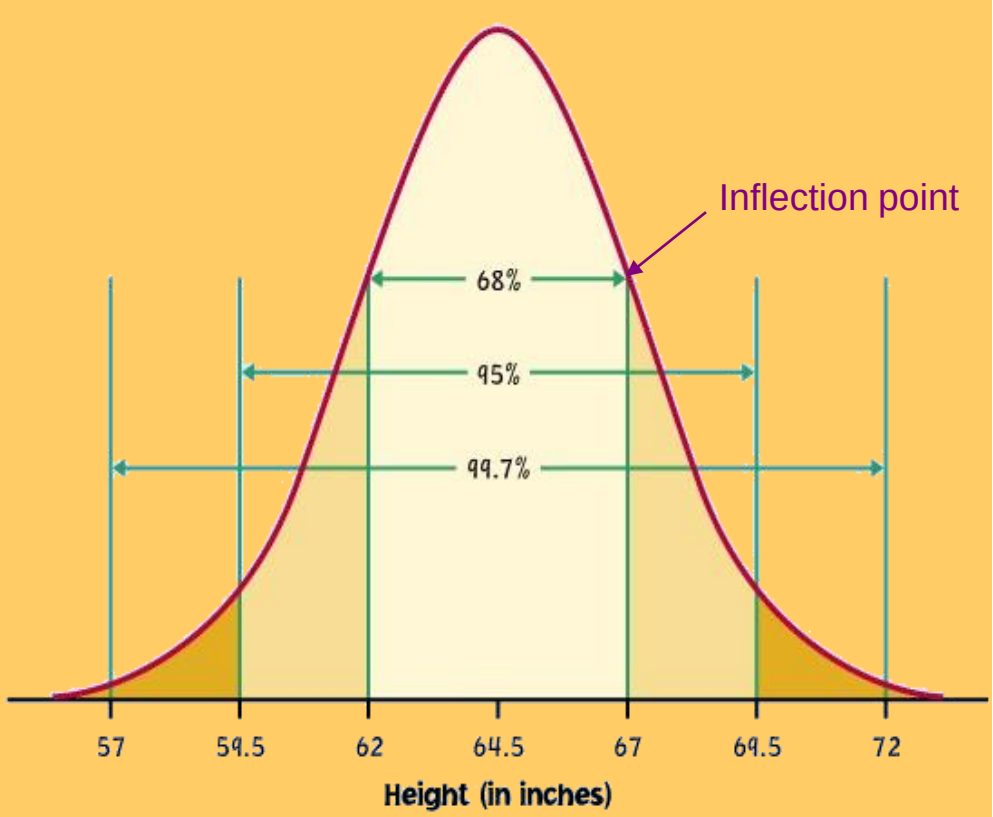
- Normal distribution central position in statistics
- Reasons:
  - Often data approximately normally distributed (On the original scale or after suitable transformation)
  - Even if the data is not normally distributed, statistics computed from the data (e.g. the sample mean) are usually approximately normally distributed if the sample size is large (based on “central limit theorem”)

The normal distribution

https://www.youtube.com/watch?v=6YDHBFVIvIs

Normal distribution: the 68-95-99.7 rule

- About 68% of all observations are within 1 standard deviation (σ) of the mean (μ).
- About 95% of all observations are within 2 σ of the mean μ.
- Almost all (99.7%) observations are within 3 σ of the mean.

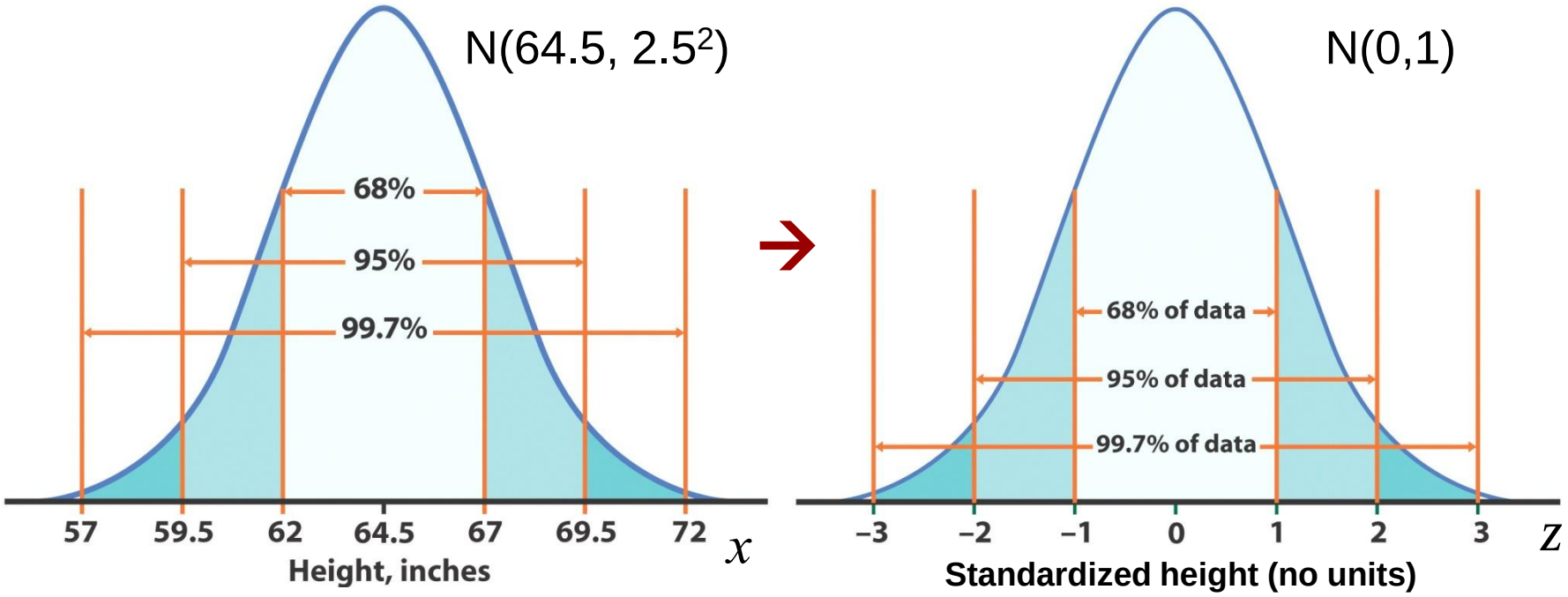


mean μ = 64.5 standard deviation σ = 2.5

$N(\mu, \sigma^2) = N(64.5, 2.5^2)$

The standard normal distribution

Because all normal distributions share the same properties, we can standardize any normal curve $N(\mu,\sigma^2)$ into standard normal curve N(0,1).



For each x we calculate a new value z (called a z-score).

Standardizing: calculating z-scores

- A z-score measures the number of standard deviations that a data value x is from the mean μ

$$z = \frac{(x - \mu)}{\sigma}$$

- z= 1: x is 1 standard deviation larger than the mean
- z= 2: x is 2 standard deviation larger than the mean.
- z=-1: x is 1 standard deviation lower than the mean.

– Reported in output of statistical model

Calculating normal probabilities with R

- `pnorm` function gives the cumulative probability of the normal distribution.
- $X \sim N(64.5, 2.5^2)$
  - $P(X \leq 67) = 0.84$

    ```
    pnorm(67,mean=64.5,sd=2.5)=
    pnorm((67-64.5)/2.5, 0, 1)=pnorm(1,0,1)
    ```

  - $P(62 \leq X \leq 67) = 0.68$

    ```
    pnorm(67,mean=64.5,sd=2.5)-
      pnorm(62,mean=64.5,sd=2.5)
    ```

Statistical inference (numerical variables)

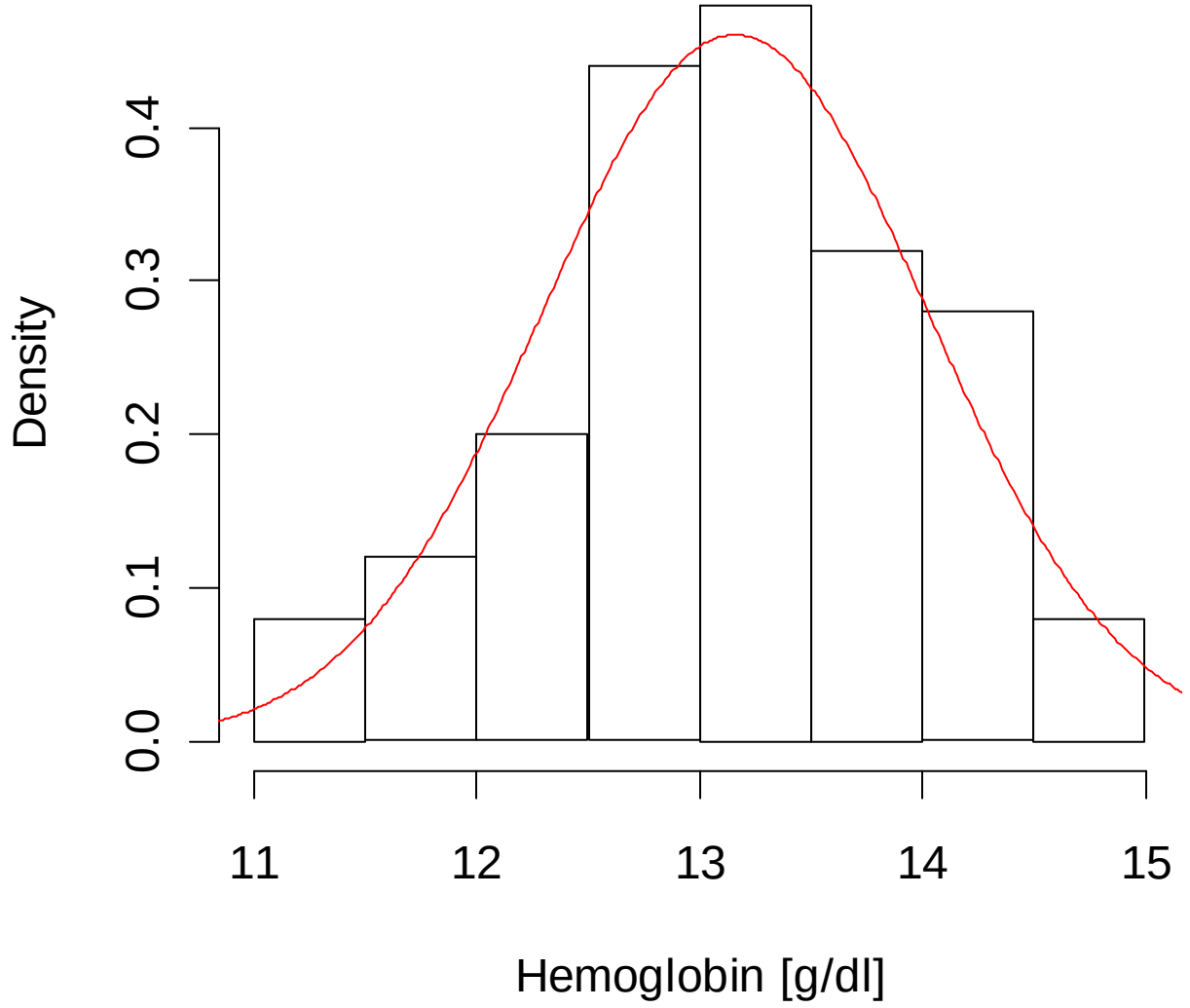
The statistical framework

1. Define a probability model for the population
   - most important (unknown) characteristics: mean μ and standard deviation σ
2. View the study sample as a random sample of n observations from the probability model. Based on data from the sample
   - decide on the most plausible population mean μ given the data → Estimation
   - decide on range of possible population means which are plausible given the data → Confidence interval
   - test whether the population mean could be equal to a pre-defined value → Statistical test

Example

- Goal: Assess average hemoglobin [Hgb] levels in children with dengue within 3 days of illness onset (presenting to Hospital for Tropical Diseases, HCMC, Vietnam)
- Study design: Assess a sample of 50 children from the target population and measure Hgb in all of them
- Observed descriptive statistics:
  - Sample mean: 13.16
  - Sample sd: 0.87

Histogram of Hgb from 50 children (Best fitting normal curve overlaid)



Statistical questions; probability model

- Some basic statistical questions:
  - What's the best estimate of the true (population) mean Hgb level μ? → Estimation
  - Which range of true mean population Hgb levels is plausible given the data? → Confidence interval
- Probability model

  The measurements are independent draws from the population distribution with unknown population mean μ and population standard deviation σ:

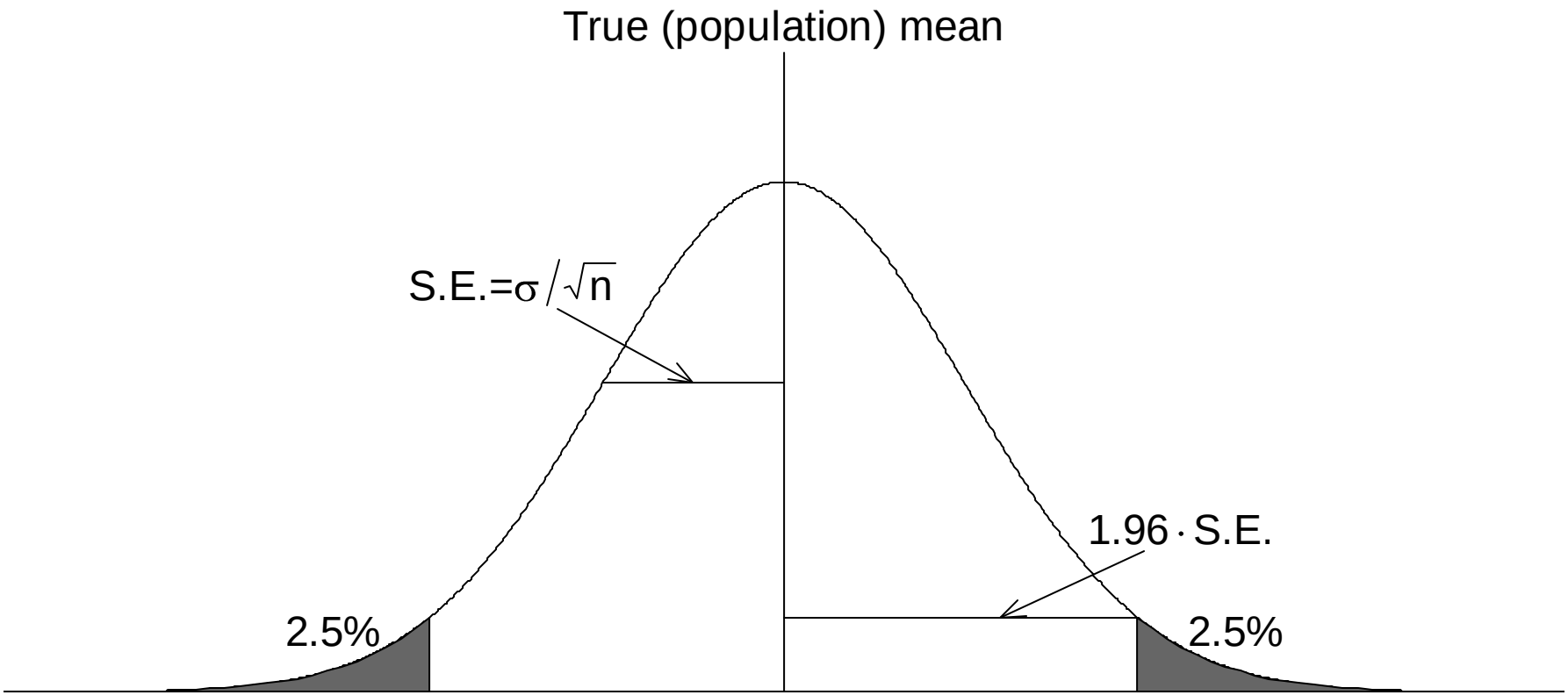
  $X_1,\ldots,X_{50} \sim \text{Prob}(\mu,\sigma^2)$ independent
- Distribution of sample mean $\overline{X}$ ≈ normal:

  $\overline{X} \sim \text{Norm}(\mu,SE^2)$

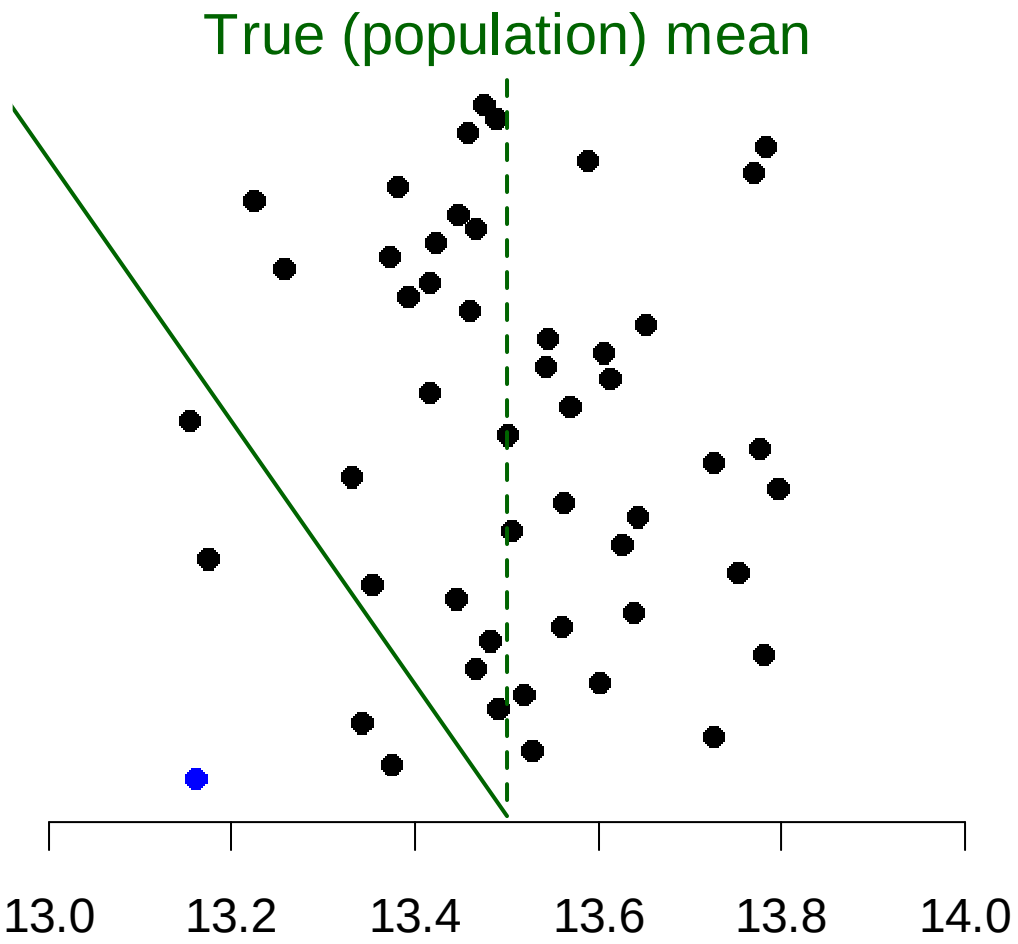
Standard error of the mean

- Quantifies the precision of the sample mean as estimate of the population mean
- Defined as the standard deviation of the sample means if the study were repeated many times
- Mathematical formula: $$S.E. = \frac{\sigma}{\sqrt{n}}$$
- Can be estimated from single study by: $$S.E. = \frac{sd}{\sqrt{n}}$$
- Hence S.E. depends on
  - Variation in population (sd)
  - Sample size

Distribution of observed sample means



Assume μ=13.5. Sample mean if we repeat the study 50 times.



Distribution of observed sample means

- Note that

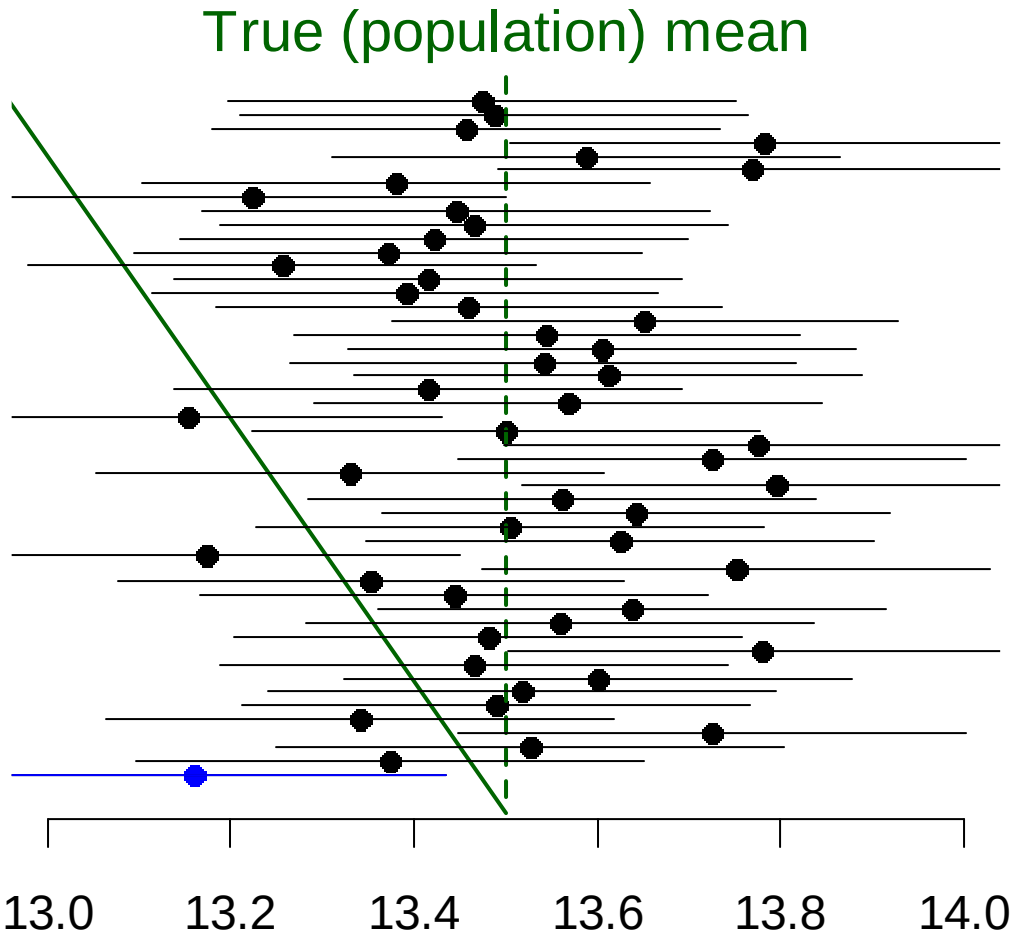
$$P(\bar{x} \text{ is in the range } \mu - 1.96 \cdot SE \text{ to } \mu + 1.96 \cdot SE) = 95\%$$

is the same as saying

$$P(\mu \text{ is in the range } \bar{x} - 1.96 \cdot SE \text{ to } \bar{x} + 1.96 \cdot SE) = 95\%$$

→ 95% confidence interval

Observed sample means for 50 repetitions of the study (with 95% CI)



Confidence interval (CI)

- Quantifies the range of plausible population means given the result from the sample
- Often 95% CI is chosen
- If we perform many studies, the 95% CI will contain the true effect in 95% of the studies
  - "We are 95% confident that the true population mean is contained in the 95% CI"

Approximate formulas for CI of the population mean (for large samples)

$$90\%\,\mathrm{CI}:\ \bar{x} \pm 1.64 \cdot S.E.$$

$$95\%\,\mathrm{CI}:\ \bar{x} \pm 1.96 \cdot S.E.$$

$$99\%\,\mathrm{CI}:\ \bar{x} \pm 2.58 \cdot S.E.$$

Confidence interval of the mean based on the *t* distribution

- Corrects for the fact that s.d. is estimated. Uses the quantiles from the Student's *t*-distribution with n-1 degrees of freedom instead of the 1.96 from the normal distribution. (n: sample size)
  - Practical consequence
    - n= 5 → 1.96 replaced by 2.78
    - n=10 → 1.96 replaced by 2.26
    - n=20 → 1.96 replaced by 2.09
    - n=50 → 1.96 replaced by 2.01

Confidence interval - example

- Sample mean Hgb in the dengue study (n=50):
  13.16 with standard deviation sd=0.87
- Standard error of the mean:

  $sd/\sqrt{50} = 0.12$
- Approximate 95% CI for the population mean:
  (13.16 – 1.96*0.12 to 13.16+1.96*0.12) =
  (12.92 to 13.39)
- More precise interval based on t-distribution:

  (13.16 – 2.01*0.12 to 13.16+2.01*0.12) =

  (12.92 to 13.40)

Distribution of variable not normal

- Confidence intervals based on the normal distribution are approximately correct if
  - n is large or the variable distribution is not too different from the normal distribution (i.e. roughly symmetric distribution, not too many outliers)
- In case of a highly asymmetric distribution
  - Try to transform first (e.g. log transformation)
  - If still not normal, use another method (class 4.)

SD and SE

SD: variation in the values of a variable in sample. May be used in descriptive statistics to show the spread in the sample data

SE: precision of the sample mean (as estimate of the population mean). Usually provided by statistical program in output of statistical model. Basis for CI and p-value

In publications use CI rather than SE

Interpretation of confidence intervals (CI)

- CI used to express uncertainty of estimate of population parameter
- CI should be provided for estimated parameters (not for purely descriptive statistics)
- The confidence interval is random, the population parameter is fixed (but unknown)
  - "Probability that the true mean Hgb is in the interval (12.92 to 13.40) is 95%" is not correct.
  - If we want to make probability statements, need to refer to (hypothetical) repetitions of the study
    - "If the study were repeated 100 times, approximately 95 of the calculated 95% CIs would contain the true parameter."

Confidence interval for a single proportion

Estimation

- **Problem**: Out of a sample of 50 individuals with DENV infection, 2 developed severe dengue. What can we say about the risk of severe dengue in the population?
- **Statistical model:**
  - Assume: the n observations are independent with the same probability p of severe dengue [p unknown]
    → Number severe dengue in sample binomial distribution B(n,p)
  - Estimate population proportion by sample proportion.
    Note: proportion can be seen as mean value of 0's ("fail") and 1's ("success")
  - Standard error of the sample proportion (use formula for mean):

$$SE = \frac{sd}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Confidence interval for a proportion-example

- 2/50 patients developed severe dengue
  - Sample proportion: 0.04
  - Standard error of sample proportion: sqrt(0.04*0.96/50) = 0.0277
- Approximate 95% CI (note the negative value!):
  (0.04–1.96*0.0277 to 0.04+1.96*0.0277) = (-0.014 to 0.094)
- Confidence intervals with R via `prop.test` (Wilson) and `binom.test` (Exact)
- Note: `binconf` in library Hmisc gives all:

```
> binconf(x=2,n=50,method="all")
           PointEst     Lower     Upper
Exact          0.04   0.00488    0.1371
Wilson         0.04   0.01104    0.1346
Asymptotic     0.04  -0.01432    0.0943
```

Hypothesis testing – general principles for the comparison of groups

Steps in hypothesis testing I

1. Define study hypothesis for population
   "What do we want to prove"

Treatment A better cure rate than treatment B. ("one-sided")

Adults and children have different mean recovery times. ("two-sided")

2. Define null hypothesis H_0
   Usually: "No effect"

No difference in cure rates between treatment A and B

Children and adults have the same mean recovery time

3. Plan study and collect data

Steps in hypothesis testing II

4. Calculate summary statistics (estimated difference)

Cure rate difference between treatments A and B

Difference in mean recovery time between children and adults

5. Calculate test statistics

Often in “standardized” form

(Observed summary statistic – Expected if null hypothesis is true) / se(Observed)

Which test statistic (test, method) to use depends on:

- question asked
- type of data (binary/categorical/continuous)
- related/paired or independent groups
- number of groups

e.g. chi-square test, t-test, Mann-Whitney test,….

Steps in hypothesis testing III

6.Derive p-value

P-value: probability that the observed data (or more extreme) are obtained if the null hypothesis were true

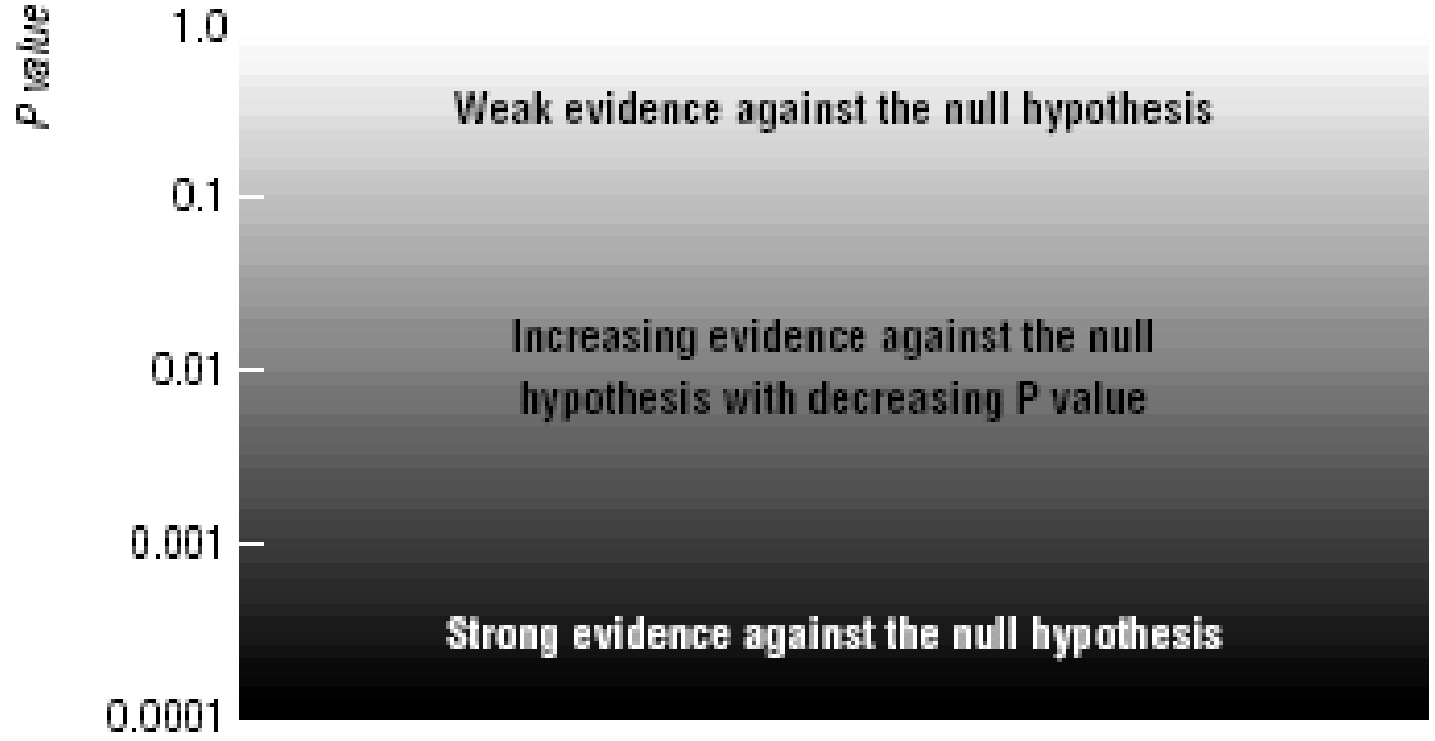
Test statistic ≈ follows some well characterized distribution, so this probability can be calculated. Usually done by statistical program, but can also be obtained via simulation (see Friday exercise)

What “more extreme” means depends on whether we select a two- or a one-sided study hypothesis.
Usually: two-sided tests

Small p-value → evidence against the null hypothesis
e.g. p=0.003: if null hypothesis were true, the probability of observing the summary statistic at least of this size is 3 in 1000 = very unlikely

Steps in hypothesis testing IV

7. Draw conclusion in words



Suggested interpretation of P values from published medical research

Common practice (to be avoided)

p-value ≤ 0.05

- Called “statistically significant”
- Decision: reject null hypothesis
- Interpretation: proof that there is a difference between the groups

p-value> 0.05

- Not significant
- Do not reject the null hypothesis
- Often seen as “accept the null hypothesis”:

“there is no difference between the groups”, but

“absence of proof” is not “proof of absence”

There may be low power to detect an existing difference because

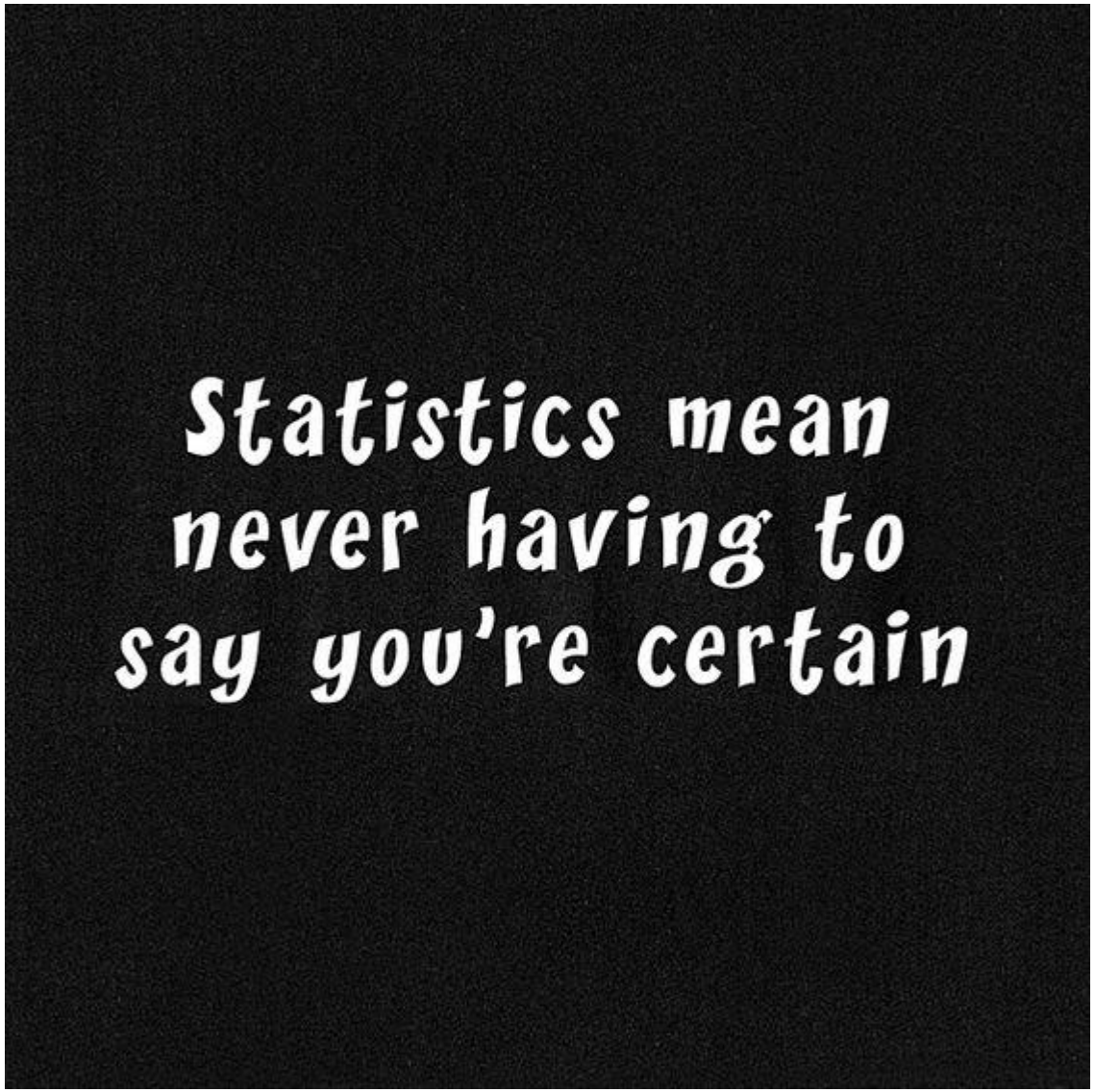
- The true difference between the groups is small
- Sample size is small

Don’t use p<0.05 as criterion

P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Summary

- Sample vs. population
- SD of variable vs. SE of the mean
- Point estimate (population mean) vs. interval estimate (95% CI of the mean)
- Hypothesis testing using:
  - p-values (plausibility of H_0)
  - CI (more clinical information)
- We will never learn the truth with 100% certainty


Statistics mean
never having to
say you're certain