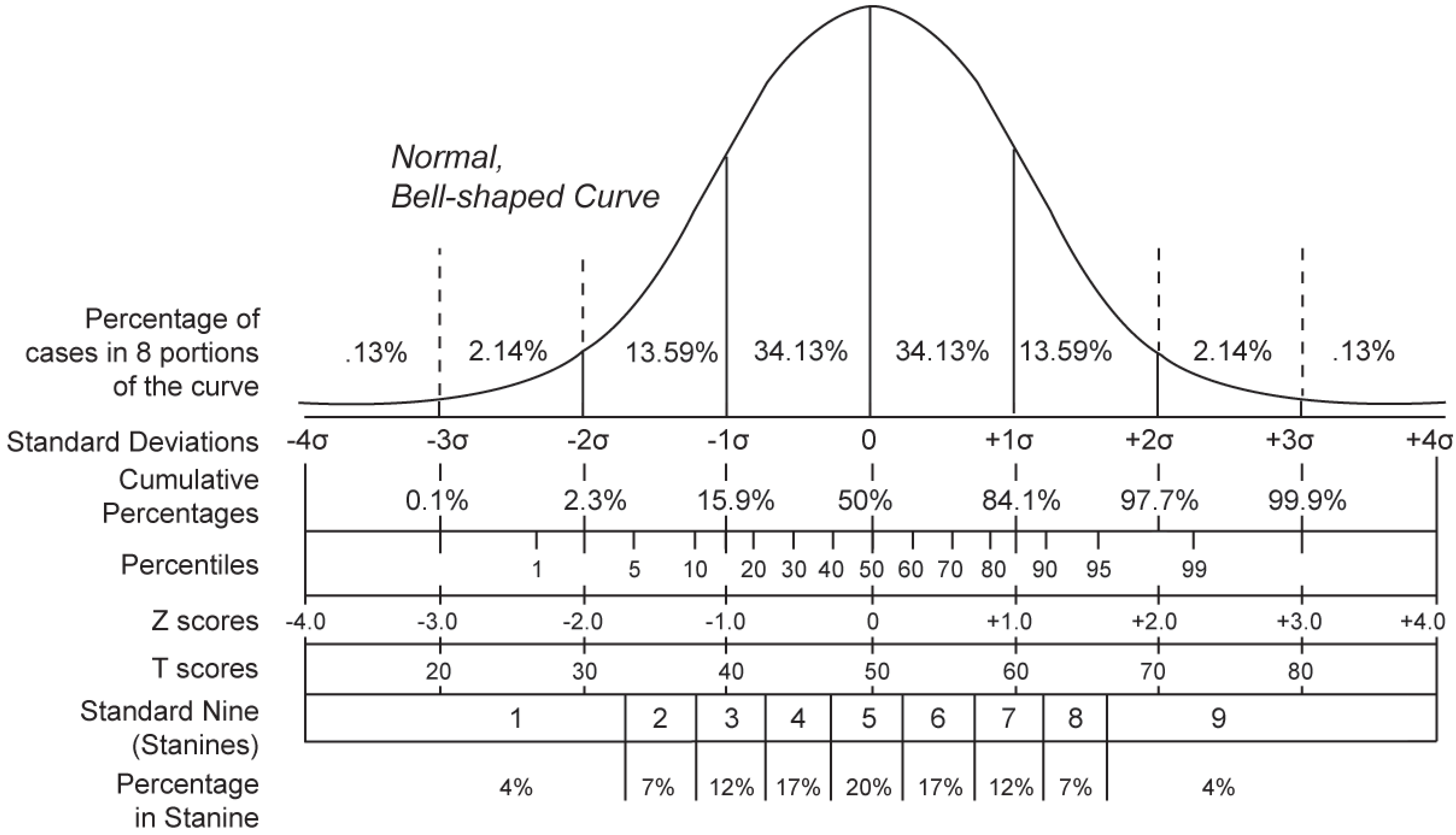
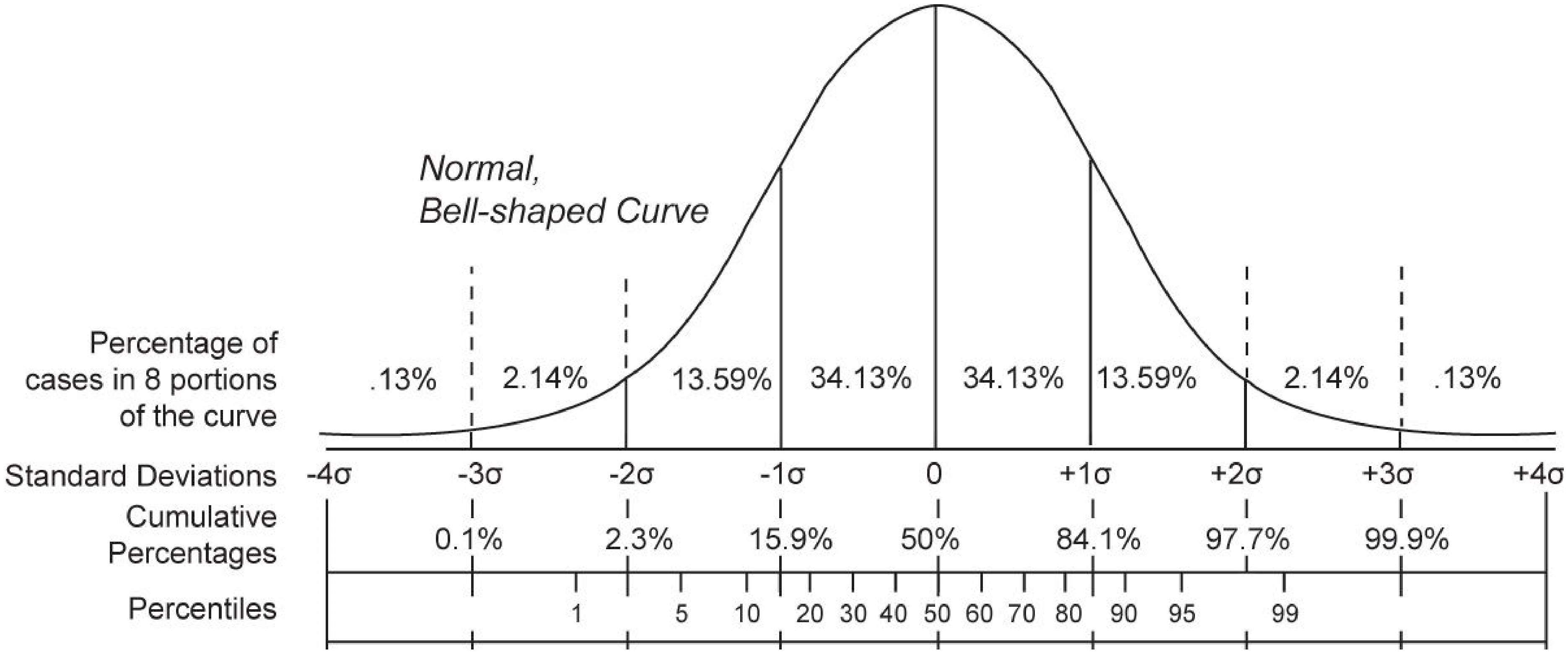
Experimental Humanities II (HUMB002) 2016
STATISTICAL ANALYSIS

Lecture 3

STATISTICAL INFERENCE
CONFIDENCE INTERVALS

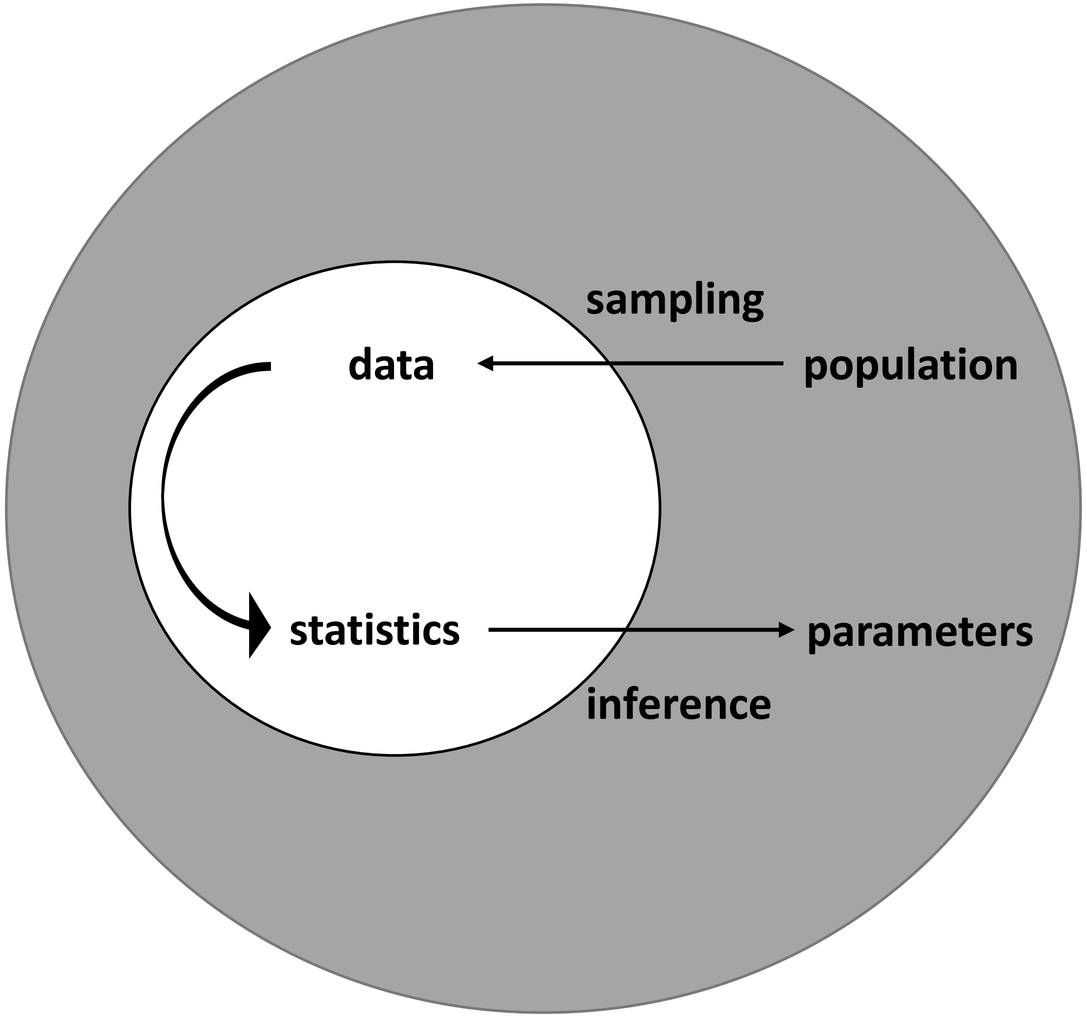

Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ -3σ -2σ -1σ 0 +1σ +2σ +3σ +4σ
Cumulative Percentages
0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
Percentiles
1 5 10 20 30 40 50 60 70 80 90 95 99
Z scores
-4.0 -3.0 -2.0 -1.0 0 +1.0 +2.0 +3.0 +4.0
T scores
20 30 40 50 60 70 80
Standard Nine (Stanines)
1 2 3 4 5 6 7 8 9
Percentage in Stanine
4% 7% 12% 17% 20% 17% 12% 7% 4%



		Reading			Math
	Wechsler IQ	3. grade	5. grade	8. grade	5. grade
m	100	3.0	5.0	8.0	5.0
s	15	1.0	1.4	1.9	1.1

- Percentiles to z-scores: =NORM.S.INV(p)
- Z-scores to percentiles (probability of a z-score): =NORM.S.DIST(z;1)

From description to inference



- Data description, parameters estimation
- Statistical inference (cz: usuzování, inference, indukce)
- Random sampling
  - every subject has the same probability of being included in the sample
  - If we don’t have random sampling, how is our sample different?

Statistics vs. parameters

- From sample (data) we compute statistics
- The value of a statistics in whole population is called parameter
  - For parameters we used Greek letters
  - e.g. mean: statistics *m*, parameter μ, *correlation: statistics r, parameter* ρ, standard deviation: statistics *s,* parameter σ
- Statistics are parameter **estimates**
  - They are always burdened with error – sampling error (cz: výběrová chyba)
  - Random error (cz: náhodné chyby) – we can compute them, if we know sampling distribution (cz: výběrové rozložení)
  - Systematic errors – biased measurement, bad sampling and other methodological problems
- How good are our estimates?

Sampling distribution and standard error

- If we compute the same statistics on many independent samples from a population, we get many <u>different</u> parameter estimates
- These estimates have some distribution – **sampling distribution (cz: výběrové rozložení)**
- **http://onlinestatbook.com/stat_sim/sampling_dist/index.html**
- Sampling distribution can be described by:
  - Mean – sampling distribution mean is close to parameter
  - Standard deviation – in sampling distribution called **standard error (cz: směrodatná chyba,** také střední chyba či výběrová chyba)
  - The higher is the number of samples (statistics estimates), the lower is the standard error

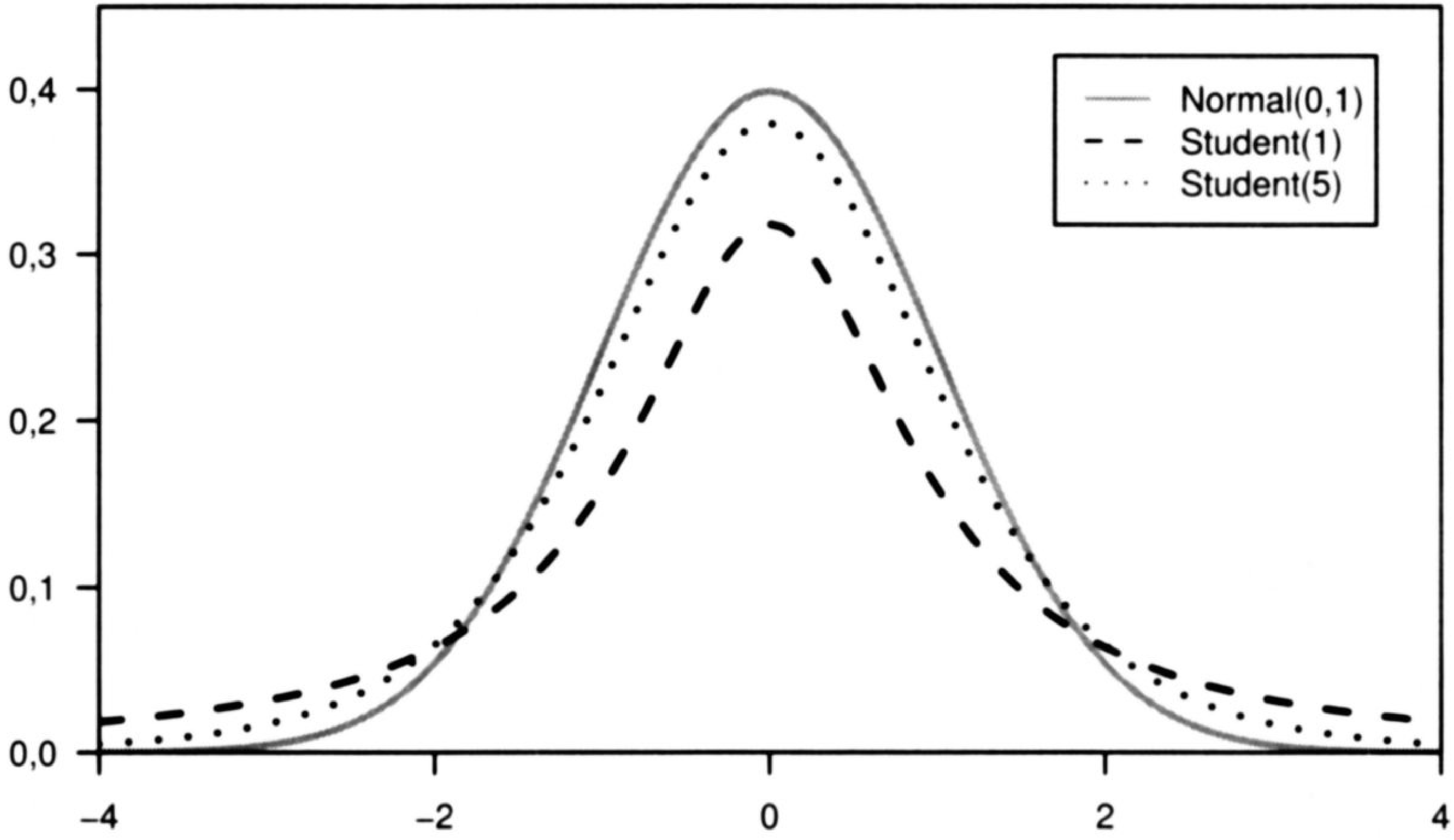
Sampling distribution of mean (estimate)

- Mean estimate has approximately **normal distribution**
  - With mean μ and standard error: $$\sigma_x = \frac{\sigma}{\sqrt{N}}$$
  - This applies even when the distribution is not normal – thanks to **central limit theorem**
  - The problem is that we usually don't know σ
- If we don't know σ, we have to use *s*
  - Mean is still μ and standard error is now: $$S_x = \frac{s}{\sqrt{N}}$$
  - The sampling distribution is not normal, but Student's t-distribution

Student's t-distribution

- Like normal distribution, but with „heavier ends“
  (cz: Studentovo t-rozložení)
- *t* in Student's distribution is the same as *z* in normal distribution
- It has different shapes for different N:
  - It is characterized by degrees of freedom: df = N-1 (also ν – „ný“)
    (cz: stupně volnosti)
- The higher is N, the more t-distribution approximates normal distribution

Student's t-distribution



Sampling distribution of other statistics

- For every statistics we need to know its theoretical sampling distribution
  - Relative frequencies: approximately normal distribution
  - Pearson’s r – after Fisher transformation normal distribution

Statistics estimation quality

TABLE 5.1 The Expected Values of the Range, s^2, and s as a Function of Sample Size of n Observations from a Random Sample from a Normal Distribution in which $\sigma = 10$

If $\sigma = 10$ *n*	*Expected Value of the Range*	*Expected Value of s^2*	*Expected Value of s*	*Expected Value of Range/s*
2	11	100	8.0	1.4
5	23	100	9.4	2.4
10	31	100	9.73	3.2
20	37	100	9.87	3.7
50	45	100	9.95	4.5
100	50	100	9.97	5.0
200	55	100	9.987	5.5
500	61	100	9.993	6.1
1,000	65	100	9.997	6.5

Point vs. Interval estimates

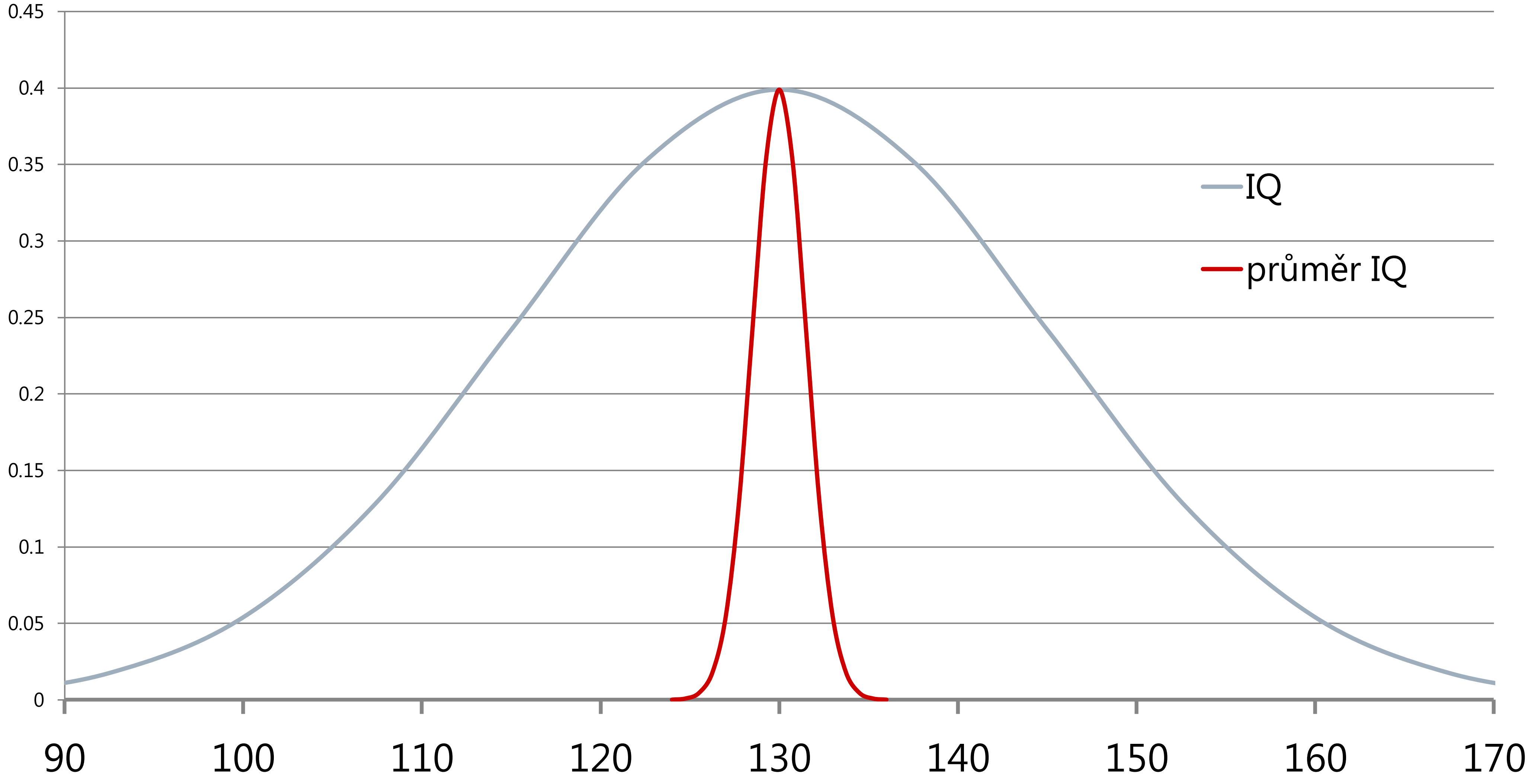
- Parameter can be estimated by:
  - Point estimate – that we're trying to estimate the parameter value itself (e.g. mean)
  - Interval estimate – estimating an interval in which the parameter lies with certain probability
    - The result is **confidence interval (cz: interval spolehlivosti), CI**
    - Confidence interval can be computed from point estimate and its sampling distribution (point $\pm$ deviation)
    - Interval estimate is better – we have more information

$$_{(1-\alpha)}CI = \bar{X} \pm z_{1-\alpha/2}\,\sigma_{\bar{X}}$$

    - α – error probability, (1 – α) is **confidence level (cz: hladina spolehlivosti)**
    - We typically use 95% or 99% confidence level, then it means that the parameter lies in the confidence interval with 95% probability (where α is 0.05 = 5% error probability, (1 – α) = (1 – 0.05))

Computing confidence interval for mean 1

- In a sample of 100 children with multicolored eyes we computed mean IQ 130 and we know that σ =15.
  - Point parameter estimate (μ) is 130
  - Interval parameter estimate:
    - **Sampling distribution of mean is normal**…
    - …with centre in μ. We don’t know μ, so we use our **point interval m = 130**.
    - …with **standard error of mean $s_m = \sigma / \sqrt{N}$** = 15/ √100 = 1,5.
    - We **choose our confidence level**: 1-α = 95%
    - Then we find **z-score between which lies 1-α % of normal distribution**:
      95% of normal distribution lies between z-scores -1.96 and 1.96
      in other words: ${}_{1-\alpha/2}z = {}_{0,975}z = 1.96$
      Excel: **=NORM.S.INV(0.975)**
    - Confidence interval: **(m – 1.96s_m; m + 1,96s_m)** = (130 – 1.96*15; 130 + 1.96*15) = (127.1 ; 132.9)
    - **That is: with 95% probability 127.1 ≤ μ ≤ 132.9**


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
90
100
110
120
130
140
150
160
170
IQ
průměr IQ

Computing confidence interval for mean 2

- In a sample of 100 children with multicolored eyes we computed mean IQ 130 and s = 15.
  - Point parameter estimate (μ) is 130
  - Interval parameter estimate:
    - We don't know σ, so the sampling distribution of mean is not normal, but **Student's t-distribution with df = N-1 = 99**
    - The distribution centre will again be our **point interval m = 130**.
    - **Standard error of mean** is $s_m = s/\sqrt{n} = 15/\sqrt{100} = 1{,}5$
    - We **choose our confidence level**: $1-\alpha = 95\%$
    - Then we find **t-score between which lies 1-α % of t-distribution**:
      95% of t-distribution with df=99 lies between t-scores -1.98 and 1.98
      in other words: ${}_{1-\alpha/2}t\ (df) = {}_{0,975}t\ (99) = 1.98$
    - Excel: **=T.INV(p;df), here =T.INV(0.975;99)**
    - Confidence interval: $\boldsymbol{(m - 1.98 s_m;\ m + 1{,}98 s_m)}$ = (130 – 1.98*15; 130 + 1.98*15) = (127.0 ; 133.0)
    - **That is: with 95% probability 127.0 $\leq \mu \leq$ 133.0**

Confidence intervals interpretation

- 95% confidence interval means that in 95% of such interval constructions (measurements) the parameter will fall into this interval, that is in 95% of measurements the parameter estimate will lie in the interval
- We have 95% subjective confidence that the parameter lies in the interval
- But – the parameter value doesn’t change, only our estimates are always a bit different

Sampling distribution of relative frequencies *p*

- …is approximately normal with mean *p* and standard error $\sqrt{p(1-p)/n}$
- (1-α)% confidence interval thus is:

$$\left(p - z_{1-\alpha/2}\sqrt{p(1-p)/n};\, p + z_{1-\alpha/2}\sqrt{p(1-p)/n}\right)$$

Sampling distribution of Pearson's correlation *r*

- We don't know sampling distribution of correlation…
- …but we know sampling distribution of correlation after Fisher transformation:
  in Excel: *Z* = FISHER(r)
- Sampling distribution of the Fisher *Z* is approximately normal with mean *Z* and standard error s_Z=1/√(n-3)
- (1−α)% CI for *Z*: $\left(Z - z_{1-\alpha/2} s_Z ; Z + z_{1-\alpha/2} s_Z\right)$
- Then, we need to transform Fisher *Z* back to Pearson's *r:*
  in Excel: =FISHERINV(z)

$$\left(FISHERINV\left(Z - z_{1-\alpha/2} s_Z\right); FISHERINV\left(Z + z_{1-\alpha/2} s_Z\right)\right)$$

Confidence interval for correlation

On a sample of 20 children we found correlation between number of hours spend by reading per week and score in a creativity test r=0.45. Compute 90% confidence interval for the correlation.

- Transform Pearson's correlation to Fisher Z in Excel: =FISHER(0.45) = 0.48
- Sampling distribution of Fisher Z is approximately normal
- Compute standard error for Fisher Z: s_Z = 1/√(20-3) = 0.24
- Compute border z-scores for 90% confidence interval:
  =NORM.S.INV(1-0.1/2) =NORM.S.INV(0.95) = 1.64
- 90% CI for Fisher Z: (0.48 – 1.64*0.24; 0.48 + 1.64*0.24) = (0.09; 0.88)
- Tranform the results back to Pearson's r:
  90% CI for Pearson's r: (=FISHERINV(0.09); =FISHERINV(0.88)) = (0.09; 0.71)

Confidence interval for correlation

- Let's continue with the exercise from the previous slide. Imagine we measured the same correlation 0.45, but now in sample size N=100. What computation in the confidence interval will change with sample size?
- Only standard error for correlation will change. Will the standard error be higher, or lower than in the sample of N=20?
- It will be lower, because we divide 1 by higher number. Will the confidence interval get wider, or narrower with bigger sample?
- It will get narrower because we have lower standard error. Compute again 90% confidence interval for correlation 0.45 and N=100 and see how the CI changed.
  - $s_Z = 1/\sqrt{(100-3)} = 0.10$
  - 90% for Pearson's r: (=FISHERINV(0.48 – 1.64*0.10); =FISHERINV(0.48 + 1.64*0.10)) = (0.31; 0.57)
- See that with 5 times bigger sample the confidence interval got substantially narrower (e.g.: more precise interval estimate)

Confidence interval for relative frequency

A survey on 1000 people discovered that approximately 12% women experienced a depression episode, whereas in men it was 7%. Compute 99% confidence interval for the difference in probability of having depression between woman and men.

- Compute the probability difference: p = 0.12 – 0.07 = 0.05
- Sampling distribution of relative frequency is approximately normal
- Compute standard error for probability: $s_p = \sqrt{(0.05*(1-0.05))/1000} = 0.007$
- Compute border z-scores for 99% confidence interval: =NORM.S.INV(1-0.01/2) =NORM.S.INV(0.995) = 2.58
- 99% CI for p: (0.05 – 2.58*0.007; 0.05 + 2.58*0.007) = (0.032; 0.068) = (3.2%; 6.8%)
- Compute confidence interval for the same data, but now compute 96% CI. Will the interval be narrower, or wider?
- It will be narrower: =NORM.S.INV(1-0.04/2) =NORM.S.INV(0.98) = 2.05
- 96% CI for p: (0.05 – 2.05*0.007; 0.05 + 2.05*0.007) = (0.036; 0.064) = (3.6%; 6.4%)

General procedure for computing CIs

1. **Determine sampling distribution of given statistics:**
   - for mean with known σ: normal distribution (z-scores)
   - for mean with unknown σ: Student's t-distribution (t-scores) with df=N-1
   - for relative frequency: normal distribution (z-scores)
   - for Pearson's correlation: normal after Fisher transformation (then z-scores)
2. **If needed, transform the statistics:**
   - from the above only for correlation, Excel: =FISHER(r)
3. **Determine standard error for given sampling distribution:**
   - for mean with known σ: $s_m = \sigma / \sqrt{N}$
   - for mean with unknown σ: $s_m = s / \sqrt{N}$
   - for relative frequency: $s_p = \sqrt{p(1-p)/N}$
   - for Fisher Z: $s_Z = 1 / \sqrt{(N-3)}$

General procedure for computing CIs

4. Determine point estimate for given sampling distribution:

- m, p, r

5. Choose confidence level – typically 95% or 99% (theoretically any):

- for 95%: α = 0.05 = 5% error probability
  $1 - \alpha = 1 - 0.05 = 0.95$ = 95%,
  $1 - \alpha/2 = 1 - 0.025 = 0.975$

6. Find boundary scores between which lies ($1 - \alpha$) % of given distribution:

- for normal distribution (in Excel): =NORM.S.INV(1 – α/2)
- for t-distribution (in Excel): =T.INV(1 – α/2; df)

5. Compute confidence interval:

- CI = point estimate ± boundary score*standard error
- normal distribution: CI = point estimate ± ${}_{1-\alpha/2}z$*standard error
- t-distribution: CI = point estimate ± ${}_{1-\alpha/2}t(df)$*standard error

Exercise

- A researcher studies reading efficiency in college students. He measured number of words read in one minute in 6 students: 200, 240, 300, 410, 450, 600.
- Compute mean and standard deviation.
- What sampling distribution will we use for confidence interval construcion? Why?
- Construct 95% confidence interval for mean.
- Will 99% confidence interval for mean be narrower or wider than 95% interval? Construct 99% confidence interval.