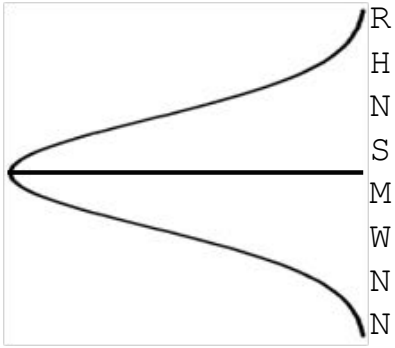
# Do the GT R at $p_{13}$ significantly impact the PT?

- We also need the approximated expected standard deviation:

$$\sigma_{p,a}^{exp} = \sqrt{\frac{(N - n_{p,a})(N+1) \cdot t_c}{12 \cdot n_{p,a}}}$$

- Where t is the tie-correction factor accounting for tied ranks. If none, then t = 1, hence we get:

- √( (20-5)·(20+1)·1 / (12·5) ) = 2.3



R 1
R 2
K 3
K 4
R 5
R 6
R 7
H 8
N 9
S 10
M 11
W 12
N 13
N 14
H 15
D 16
D 17
E 18
P 19
D 20

$$H_0: \quad \mu_{p,a}^{exp} = \overline{x}_{p,a}^{obs} \qquad H_1: \quad \mu_{p,a}^{exp} \neq \overline{x}_{p,a}^{obs}$$

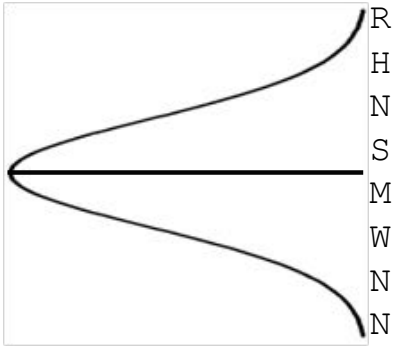
# Do the GT R at $p_{13}$ significantly impact the PT?

- Now we have what we need to calculate the z-score

$$z_{p,a} = \frac{\mu_{p,a}^{exp} - \overline{x}_{p,a}^{obs}}{\sigma_{p,a}^{exp}}$$

- Using our previous calculations, we get:

- z = (10.5 – 4.2) / 2.3 = 2.75

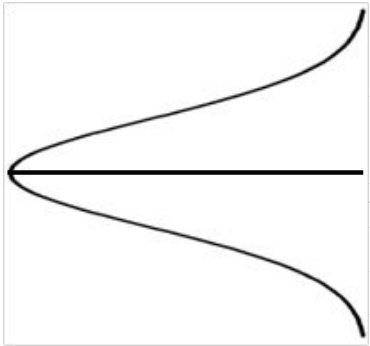


R 1
R 2
K 3
K 4
R 5
R 6
R 7
H 8
N 9
S 10
M 11
W 12
N 13
N 14
H 15
D 16
D 17
E 18
P 19
D 20

$$H_0: \quad \mu_{p,a}^{exp} = \overline{x}_{p,a}^{obs} \qquad H_1: \quad \mu_{p,a}^{exp} \neq \overline{x}_{p,a}^{obs}$$

# Do the GT R at $p_{13}$ significantly impact the PT?

- We can convert the z-score to a p-value
- So $z_{13,R} = 2.75$ corresponds to $p_{13,R} = 0.\dot{0}06$
- Which means that we reject the null-hypothesis at a level of significance of 95%
- Conclusion: The genotype: "amino acid arginine at position 13" in the MSA is significantly associated with the protein phenotype!



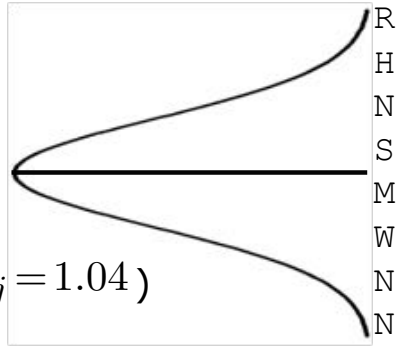
R 1
R 2
K 3
K 4
R 5
R 6
R 7
H 8
N 9
S 10
M 11
W 12
N 13
N 14
H 15
D 16
D 17
E 18
P 19
D 20

$$H_0: \quad \mu_{p,a}^{exp} = \overline{x}_{p,a}^{obs} \qquad H_1: \quad \mu_{p,a}^{exp} \neq \overline{x}_{p,a}^{obs}$$

# Do the GT R at $p_{13}$ significantly impact the PT?

- We can convert the z-score to a p-value
- So $z_{13,R} = 2.75$ corresponds to $p_{13,R} = 0.006$
- However, we are performing one test per amino acid per position, so we need to adjust the p-value for multiple testing
- If we are performing e.g. a total of 50 tests, then

$$p_{adj}^{Bonf.} = min(1, p \cdot n_{tests}) = min(1, 0.006 \cdot 50) = 0.3 \quad (z_{adj} = 1.04)$$

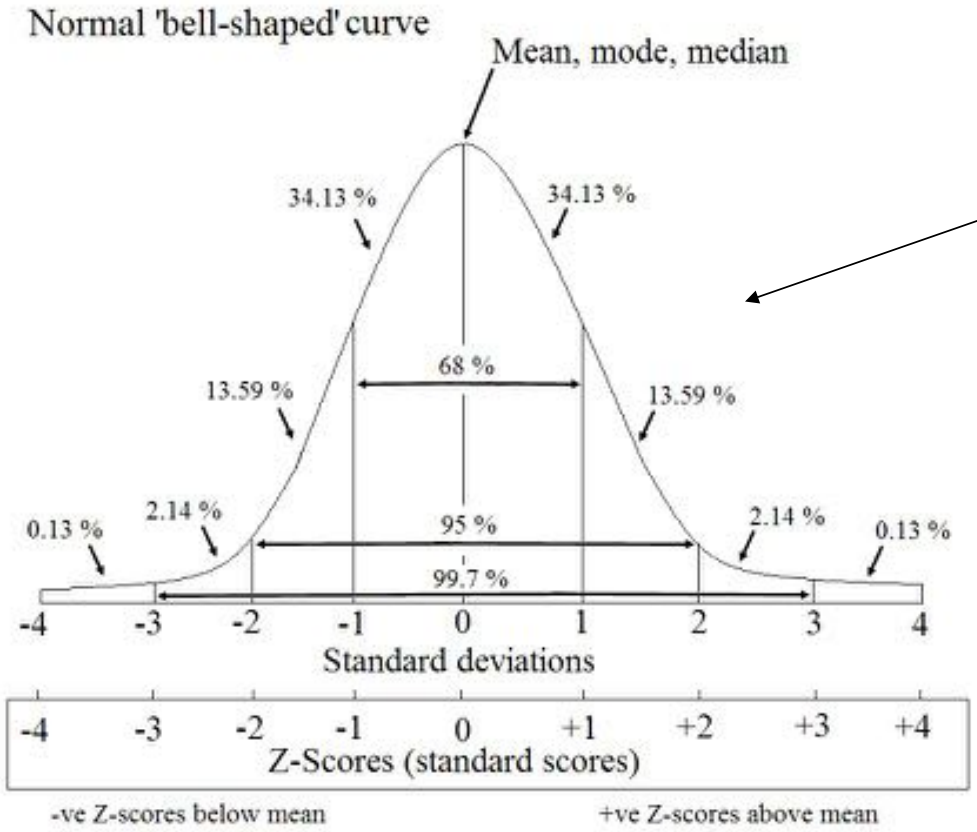


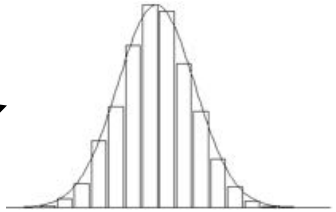
- Conclusion: We cannot reject the null-hypothesis at a 95% significance level

$$H_0: \quad \mu_{p,a}^{exp} = \overline{x}_{p,a}^{obs} \qquad H_1: \quad \mu_{p,a}^{exp} \neq \overline{x}_{p,a}^{obs}$$

```
R 1
R 2
K 3
K 4
R 5
R 6
R 7
H 8
N 9
S 10
M 11
W 12
N 13
N 14
H 15
D 16
D 17
E 18
P 19
D 20
```

# From z-score to p-values



A density curve is "just" a smoothed histogram

# So, let me know if…

# Now: Exercises!

# If you are curious for more

- Method paper

*W286–W291* *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 12 June 2013*
*doi:10.1093/nar/gkt497*

## *SigniSite*: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments

**Leon Eyrich Jessen[1], Ilka Hoof[2], Ole Lund[1] and Morten Nielsen[1,3,*]**

[1]Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark, [2]Department of Molecular Biology and Biotech Research and Innovation Centre (BRIC), Bioinformatics Centre, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen, Denmark and [3]Instituto de Investigaciones Biotecnologicas, Universidad Nacional de San Martin, San Martin, B 1650 HMP, Buenos Aires, Argentina

- And server implementation available at
  - http://www.cbs.dtu.dk/services/SigniSite/