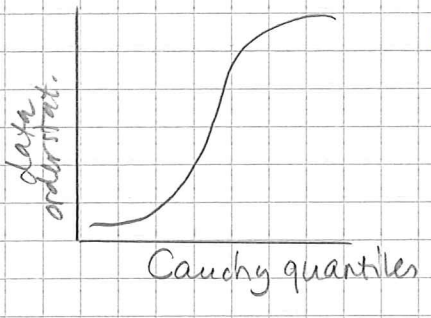
# Exam Statistical Data Analysis 03.07.14

**1.** a.) shape: rather uniform, no extremes, quite symmetric
location: median $\approx 1.2$
scale: IQR $\approx 2.7 - 0.2 = 2.5$
min $\approx -1$, max $\approx 4$

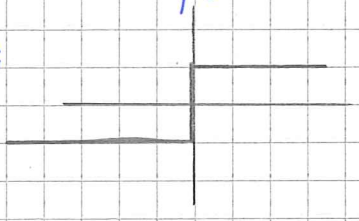
b.) Uniform looks best, since it resembles best a straight line

c.) Uniform distributions are characterized by the min and max of the support. Here: uniform $[-1, 4]$.
From plot $y = a + bx$: $b = 5$, $a = -1$.

d.) 

QQ-plot against Cauchy should have S-shape, because Cauchy has heavier tails than uniform.

**2.** a.) Incorrect. The number of observations expected under $H_0$ in each interval should equal at least 5.

b.) Correct. Sketch:  proportional to the sign function, thus bounded.

c.) Correct. In a permutation test the distribution under $H_0$ of the test statistic is simulated by resampling, as in bootstrap tests. (Perhaps also different answers possible)

d.) Correct. For $H_0: \beta_j = 0$ for a single explanatory variable one can use $T = \frac{\hat{\beta}_j}{\hat{s}_{\hat{\beta}_j}}$ which has indeed a t-distribution under $H_0$ in the given situation.

3. a.) Simulation set 1: x, y have approximately the same median, but a different scale. That is a different slope in the empirical distribution. Hence, this corresponds to the bottom right plot. Simulation set 2: x, y have a different location, but approximately the same scale. This is a horizontal shift in the empirical distributions, the top right plot.

b.) Since the Wilcoxon rank sum test is mostly suited for shift alternatives, and not so much for scale differences, this test should be applied to simul. set 2. The KS-test has good power in both simulation sets, since the differences in both emp. distr. functions are clear. Moreover there seem to be no ties (which cause problems for KS)

c.) Multiple answers possible.
simul set 1: KS, or permutation with eg. $T = sd(X) - sd(Y)$
simul set 2: Wilcoxon or KS, or permutation with eq $T = \bar{X} - \bar{Y}$.

4. a.)
* Estimate $P_\theta$ by $P_{\hat{\theta}}$, a parametric distribution
(Hence: estimate distribution $Q_P$ of T by $Q_{P_{\hat{\theta}}}$)
* Estimate $Q_{P_{\hat{\theta}}}$ of T by the empirical distribution of a sample $T_1^*, \ldots, T_B^*$ from $Q_{P_{\hat{\theta}}}$.
* Estimate the sd of $T_n$ by the sample sd of $T_1^* \ldots T_B^*$.

b) In first step: $P \approx P_{\hat{\theta}}$, in second step $Q_{P_{\hat{\theta}}} \approx$ emp. dist($T^*$)
first error depends on data, and is unavoidable
second error depends on B: larger B makes this error smaller

c) The first error would be $P \approx \hat{P}_n$, the emp. dist. of the data. The character of the errors remains the same.

d) (i) this is the sample sd of $T^*$'s when the mean is bootstrapped
(ii) this is the sample mean of $T^*$'s when the sd is bootstrapped.

**4.** (ctd)

d) (continued) (i) shows the accuracy of the sample mean, while (ii) shows an estimate of the sample sd.

e) (i) estimates $sd(T_n)$ as in part (a) and (c)
(ii) estimates the $sd(X)$ in a complicated way.

**5.** a) Model: one sample of 76 people, categorized into 4 classes, one multinomial sample from multinom$(76, p_{11}, p_{12}, p_{21}, p_{22})$ distribution with $p_{11}+p_{12}+p_{21}+p_{22}=1$.

$H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j}$ vs $H_1: p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$ with $p_{\cdot j} = p_{1j} + p_{2j}$

$H_0$ states that the two "variables" are independent $p_{i\cdot} = p_{i1} + p_{i2}$

b) Rule of thumb: $EN_{ij} \geq 5$ in all cells in a $2 \times 2$ table.

$EN_{ij} = \frac{N_{i\cdot}\, N_{\cdot j}}{n}$ with $N_{i\cdot} = N_{i1} + N_{i2}$, $N_{\cdot j} = N_{1j} + N_{2j}$, $n = 76$

table of $EN_{ij}$:

| | |
|---|---|
| 2.4 | 28 |
| 5.5 | 41.4 |

So the rule of thumb is not fulfilled.

c) the Fisher test with test statistic $N_{11}$.
Under $H_0$ is has a hypergeometric distribution with parameters $(76, 31, 6)$.
↑ total ↑ total white balls ← size of draw.

**6.** a) $Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i$ where

$Y_i$ – $i^{th}$ response
$\beta_0 \ldots \beta_p$ – unknown constants
$x_{ij}$ – value of $j^{th}$ expl. var. in $i^{th}$ observation
$e_i$ – error in $i^{th}$ observation

Assumption: $e_i \sim N(0, \sigma^2)$ i.i.d. with $\sigma^2 > 0$ unknown,

b.) Linearity - from scatter plots $(Y, X_j)$ for $j = 1, \dots p$ and added variable plots.

independence of $e_i$ - from context.

normality of errors - from QQ-plots of residuals

constant error variance - from scatter plots of $(Y, \hat{e})$ and $(\hat{Y}, \hat{e})$

c.) outliers (means: outlying values in crime): yes, since in the top row of plots, clearly one observation is outlier

leverage points (means: outlying values in other variables): yes, since in murder there is one outlier

collinearity (means: linear relationships between expl. var): yes, since the scatter between pcths and poverty is very linear.

methods: outliers → mean shift outlier test

leverage → Cook's distance (hat values)

collinearity → VIF's, condition indices and variance decomposition