# Test Statistics for BIT and TCS (Module 6 -201700269)

January 22, 2018, 13.45-16.00 hrs. Lecturer Dick Meijer, module-coordinator Dennis Reitsma.

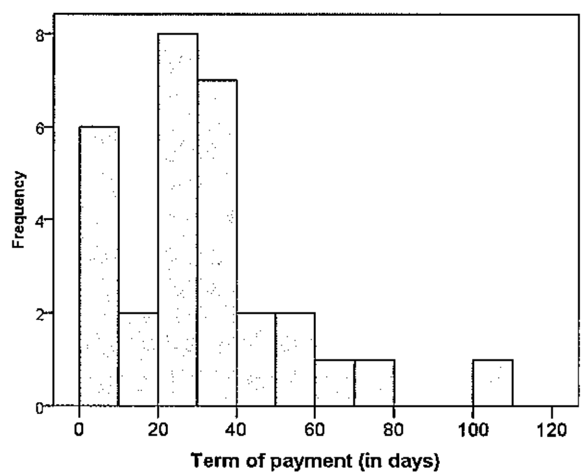
This test consists of 5 exercises and a formula sheet. The probability tables are provided separately. Use of a simple scientific calculator is allowed, use of a programmable calculator (GR) is not.

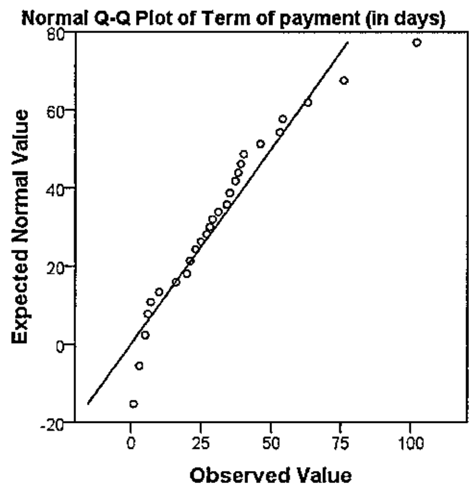
1. A company deals with varying payment behaviour of its customers and would like to construct a probability model for the term of payment in days of a random customer. The administration supplied a random sample of 30 (ordered) terms of payment of customers, presented below.
A numerical summary and two graphs, generated with SPSS, are given as well.
Note that SPSS reports the "kurtosis − 3".

| | | | | | |
|---|---|---|---|---|---|
| 1 | 3 | 5 | 5 | 6 | 7 |
| 10 | 16 | 20 | 21 | 21 | 23 |
| 25 | 27 | 28 | 29 | 31 | 34 |
| 35 | 35 | 37 | 38 | 39 | 40 |
| 46 | 53 | 54 | 63 | 76 | 102 |

Numerical summary:

| | |
|---|---|
| Sample size | 30 |
| Sample mean | 31.00 |
| Sample standard deviation | 22.71 |
| Sample variance | 515.59 |
| Sample skewness coefficient | 1.20 |
| Sample kurtosis | 2.17 |




a. Using the results of the numerical summary, determine the interval $(\overline{x} - 2s, \overline{x} + 2s)$.
How many observations are contained in the interval and how many observations did you expect to be contained in the interval according to the Empirical rule?

b. Examine whether there are any outliers according to the $1.5 \times IQR$-rule.
Determine the $z$-score for each of the outliers.

c. Determine the 80th percentile of these observations.

d. Assess whether a normal model is suitable for these observations, considering
   1. the numerical summary,
   2. the histogram and
   3. the QQ-plot,

To obtain a final conclusion about the assumption of normality, we determined, using SPSS, Shapiro Wilk's $W = 0.914$.

e. Give the value of the coefficient $a_3$ in the formula of $W$ and the rejection region for $\alpha = 5\%$.
f. What conclusion can be drawn about the assumption of a normal distribution for the terms of payment and what does this imply for the construction of a confidence interval for the expected term of payment, based on the sample?

2. A mortgage provider wants to determine the prices for mortgage consultations. One of the things he examines for this purpose, is the length of the conversations. The results of a random sample of 20 consultations are as follows (in minutes per conversation):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 67.5 | 56.0 | 40.4 | 49.3 | 57.0 | 62.3 | 50.5 | 51.5 | 44.6 | 55.0 |
| 52.3 | 54.7 | 44.8 | 57.7 | 72.7 | 50.4 | 60.5 | 66.8 | 46.8 | 38.9 |

a. Determine (using your calculator) the sample mean and the sample standard deviation.
b. Give a 95%-confidence interval for the expected conversation time.
First clearly state the assumptions on which this interval is based.
c. Someone interprets the interval, calculated in b., as follows:
"If we repeat the experiment often (observing the conversation times of 20 random customers), then in about 95 of the 100 repetitions the sample mean is included in the interval, as you determined in b." Is this a correct interpretation? Why (not)?
d. Give an interval estimate for the standard deviation of the service times at a 95% confidence level.

Based on the interval determined in d., a standard deviation of $\sigma = 10$ minutes seems to be a reasonable assumption. Now suppose that we want to test the null hypothesis that the expected conversation time is at most 50 minutes ($H_0: \mu \leq 50$), against the alternative that the expected conversation time is greater than 50 minutes ($H_1: \mu > 50$). We decide to reject the null hypothesis if the sample mean $\overline{X}$ is at least 55 minutes. Solve for this testing problem the following questions, assuming a known $\sigma = 10$ and a $N(\mu, 10^2)$-distribution of the conversation times:

e. Determine the significance level of this test. For this purpose, also sketch the distribution of the sample mean $\overline{X}$ if $\mu = 50$ and shade the significance level.
f. Calculate both the power of the test and the probability of a type II error if $\mu = 58$.
*(Hint: first sketch in the graph, determined in e., the distribution of* $\overline{X}$ *if* $\mu = 58$ *and shade the power of the test given that* $\mu$ *truly has this value).*
g. Determine the p-value for the observed mean (see a.): what does this p-value mean?

3. In order to gain some insight into the difference between the yields of two wheat varieties, an experiment on different acres of land was carried out under similar conditions.
The results are as follows (yields in bushels per acre):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variety A | 36.0 | 31.6 | 35.3 | 40.1 | 35.7 | 33.0 | 37.2 | 31.9 | 34.3 | | |
| Variety B | 34.1 | 37.8 | 39.0 | 38.4 | 35.6 | 42.1 | 42.8 | 38.8 | 39.4 | 45.9 | 37.6 |

Give, for both a. and b., 1. the hypotheses,
2. the formula of the test statistic and
3. the distribution under $H_0$
(*only these 3 points, so you do not have to calculate anything!*),
if you want to base your decision about the differences between the expected yields of the two wheat varieties on:

**a.** a parametric test
**b.** a non-parametric test (so in case the normality assumption for the yields turns out to be unreasonable)?

**4.** In an opinion poll in 2014 Dutch voters were asked which party they voted for at the most recent elections and whether or not they support the policy of the current VVD-PvdA-government. The results are provided in the table below:

| | | Support | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Political preference | VVD | 90 | 130 | 220 |
| | PvdA | 90 | 100 | 190 |
| | Remaining parties | 120 | 270 | 390 |
| | Total | 300 | 500 | 800 |

Examine whether the 3 political groups of voters think differently about the current government policy using a suitable test with α = 1%.
First argue which test you choose (and then apply the complete testing procedure).

**5.** As you know, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is an unbiased estimator of $\sigma^2$. But is it the best estimator? To figure this out, we consider a random sample, drawn from the normal distribution with unknown parameters $\mu$ and $\sigma^2$: we consider two estimators of $\sigma^2$: $T_1 = S^2$ and $T_2 = \frac{n-1}{n+1}S^2 = \frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.
Note that we can then show that $var(S^2) = \frac{2\sigma^4}{n-1}$: you may use this information to solve this exercise.

**a.** Express the mean squared error of $T_1 = S^2$ in $n$ and σ.
**b.** Determine $E(T_2)$, $var(T_2)$ and the mean squared error of $T_2 = \frac{n-1}{n+1}S^2 = \frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.
**c.** Which of the two estimators ($T_1$ and $T_2$) is the best estimator of $\sigma^2$ (for $n = 2, 3, \ldots$)?

---

Grade $= 1 + \frac{number\ of\ points}{49} \times 9$

| 1 | | | | | | 2 | | | | | | | 3 | | 4 | 5 | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | b | c | d | e | f | a | b | c | d | e | f | g | a | b | | a | b | c | |
| 2 | 4 | 1 | 3 | 2 | 2 | 1 | 3 | 1 | 3 | 2 | 3 | 2 | 3 | 3 | 8 | 1 | 3 | 2 | 49 |

# Formula sheet "Statistics for Engineers"

**Rules Probability Theory:** $var(X) = E(X^2) - (EX)^2$

$E(aX + b) = aE(X) + b$ and $var(aX + b) = a^2 var(X)$

$E(X \pm Y) = E(X) \pm E(Y)$

If $X$ and $Y$ are independent: $var(X \pm Y) = var(X) + var(Y)$

**Bounds for Confidence Intervals:**

* $\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$

* $\overline{X} \pm c\frac{S}{\sqrt{n}}$, with $P(T_{n-1} \geq c) = \frac{1}{2}\alpha$

* $\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$, with $P(\chi^2_{n-1} \leq c_1) = P(\chi^2_{n-1} \geq c_2) = \frac{\alpha}{2}$

* $\overline{X} - \overline{Y} \pm c\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, with $S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}S_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}S_Y^2$

and $P\left(T_{n_1+n_2-2} \geq c\right) = \frac{1}{2}\alpha$

or: $\overline{X} - \overline{Y} \pm c\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$

* $\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, with $\Phi(c) = 1 - \frac{1}{2}\alpha$

**Testing procedure in 8 steps**

1. Give a probability model of the observed values (the statistical assumptions).
2. State the null hypothesis and the alternative hypothesis, using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic if $H_0$ is true.
5. Compute (give) the observed value of the test statistic.
6. State the test and **a.** Determine the rejection region or
   **b.** Compute the p-value.
7. State your statistical conclusion: reject or fail to reject $H_0$ at the given significance level.
8. Draw the conclusion in words.

**Test statistics and distributions under $H_0$:**

* Binomial test: $X \sim B(n, p_0)$: $P(X = x) = \binom{n}{x}p_0^x(1-p_0)^{n-x}$ or use the binomial table,

or for large $n$ approximately $N\left(np_0, np_0(1-p_0)\right)$

* $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$

* $S^2$, where $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$
* $T = \frac{(\overline{X}-\overline{Y})-\Delta_0}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$ (and $S^2$ as given above) or $Z = \frac{(\overline{X}-\overline{Y})-\Delta_0}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}} \sim N(0,1)$
* $Z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)$, with $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$
* $F = \frac{S_X^2}{S_Y^2} \sim F^{n_1-1}_{n_2-1}$

**Analysis of categorical variables**

* 1 row and $k$ columns: $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_0N_i)^2}{E_0N_i}$ $(df = k-1)$
* $r \times c$ – cross table: $\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{(N_{ij}-\hat{E}_0N_{ij})^2}{\hat{E}_0N_{ij}}$, with $\hat{E}_0N_{ij} = \frac{\text{row total} \times \text{column total}}{n}$

  and $df = (r-1)(c-1)$.

**Non-parametric tests**

* Sign test: $X \sim B\left(n, \frac{1}{2}\right)$ under $H_0$
* Wilcoxon`s Rank sum test: $W = \sum_{i=1}^{n_1} R(X_i)$,

  under $H_0$ with: $E(W) = \frac{1}{2}n_1(N+1)$ and $var(W) = \frac{1}{12}n_1n_2(N+1)$

**Test on the normal distribution**

* Shapiro – Wilk`s test statistic: $W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$