Lecture 14 chi-square test, P-value

• Measurement error (review from lecture 13)
• Null hypothesis; alternative hypothesis
• Evidence against null hypothesis
• Measuring the Strength of evidence by P-value
• Pre-setting significance level
• Conclusion
• Confidence interval
Some general thoughts about hypothesis testing

- A claim is any statement made about the truth; it could be a theory made by a scientist, or a statement from a prosecutor, a manufacture or a consumer.
- Data cannot prove a claim however, because there may be other data that could contradict the theory.
- Data can be used to reject the claim if there is a contradiction to what may be expected.
- Put any claim in the null hypothesis $H_0$.
- Come up with an alternative hypothesis and put it as $H_1$.
- Study data and find a hypothesis testing statistics which is an informative summary of data that is most relevant in differentiating $H_1$ from $H_0$. 
• Testing statistics is obtained by experience or statistical training; it depends on the formulation of the problem and how the data are related to the hypothesis.

• Find the strength of evidence by P-value: from a future set of data, compute the probability that the summary testing statistics will be as large as or even greater than the one obtained from the current data. If P-value is very small, then either the null hypothesis is false or you are extremely unlucky. So statistician will argue that this is a strong evidence against null hypothesis.

If P-value is smaller than a pre-specified level (called significance level, 5% for example), then null hypothesis is rejected.
Back to the microarray example

- $H_0$: true SD $\sigma = 0.1$ (denote 0.1 by $\sigma_0$)
- $H_1$: true SD $\sigma > 0.1$ (because this is the main concern; you don’t care if SD is small)

Summary:
- Sample SD $(s) = \sqrt{\frac{\text{sum of squares}}{(n-1)}} = 0.18$
- Where sum of squares $= (1.1-1.3)^2 + (1.2-1.3)^2 + (1.4-1.3)^2 + (1.5-1.3)^2 = 0.1$, $n=4$
- The ratio $s/\sigma = 1.8$, is it too big?

The P-value consideration:
- Suppose a future data set ($n=4$) will be collected.
- Let $s$ be the sample SD from this future dataset; it is random; so what is the probability that $s/\sigma$ will be
- As big as or bigger than 1.8? $P(s/\sigma > 1.8)$
• \( P(s/ \bar{s}_0 > 1.8) \)
• But to find the probability we need to use chi-square distribution:
• Recall that sum of squares/ true variance follow a chi-square distribution;
• Therefore, equivalently, we compute
• \( P(\text{future sum of squares}/ \bar{s}_0^2 > \text{sum of squares from the currently available data}/ \bar{s}_0^2) \), (recall \( \bar{s}_0 \) is
• The value claimed under the null hypothesis);
Once again, if data were generated again, then Sum of squares/ true variance is random and follows a chi-squared distribution with n-1 degrees of freedom; where sum of squares= sum of squared distance between each data point and the sample mean.

Note: Sum of squares = \((n-1)\) sample variance = \((n-1)(\text{sample SD})^2\)

\[P\text{-value} = P(\text{chi-square random variable} > \text{computed value from data}) = P(\text{chisquare random variable} > 10.0)\]

For our case, \(n=4\); so look at the chi-square distribution with \(df=3\); from table we see:

The value computed from available data = \(\frac{.10}{.01}=10\)

(note sum of squares = .1, true variance = .1^2)

P-value is between .025 and .01, reject null hypothesis at 5% significance level.
Confidence interval

• A 95% confidence interval for true variance $\sigma^2$ is
• (Sum of squares/$C_2$, sum of squares/$C_1$)
• Where $C_1$ and $C_2$ are the cutting points from chi-square table with d.f=n-1 so that
• $P(\text{chisquare random variable} > C_1) = .975$
• $P(\text{chisquare random variable} > C_2) = .025$
• This interval is derived from
• $P( C_1 < \text{sum of squares}/ \sigma^2 < C_2) = .95$

For our data, sum of squares = .1 ; from d.f=3 of table, $C_1 = .216$, $C_2 = 9.348$; so the confidence interval of $\sigma^2$ is 0.1017 to .4629; how about confidence interval of $\sigma$?