Class Notes: STAT 501

Nonparametrics & Log-Linear Models

Review: Hypothesis Testing

Da Kuang

University of Pennsylvania

1 Big Picture

The statistical analysis is consist of three key components.

- Descriptive analysis
    - mean, median, variance, standard deviation, quantile, IQR
    - tables
    - graphs
- Statistical inference
    - Point estimation
    - Hypothesis testing
    - Confidence interval
- Model diagnostics.

In this lecture, we will focus on reviewing Hypothesis Testing.

2 Parametric Method

We associate parametric distributions with populations. For instance, $N(\mu, \sigma^2)$, binomial(n, p). So we make inference about the unknown parameters.

For example, we can estimate the population mean and variance by the sample.

- $X_i, i = 1, \ldots, n$, iid from $N(\mu, \sigma^2)$.
- $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

2.1 Population variance and sample variance

Why we have $n - 1$ in the denominator for sample variance?

Answers:

- Concise but not helpful answer: Bessel's correction.
- Intuitive but over-simplified answer: overcome underestimation.
  We are trying to use the sample variance to estimate the population variance. The estimator of variance is consistent but biased because it is based on the estimator of mean. The freedom of the variance is actually not n but $n - 1$. So we can overcome the underestimation by dividing by $n - 1$.
- To have a more thorough explanation, check the Wikipedia of Variance.

2.2 `pnorm` and `qnorm`

- `pnorm(-1.645) # 0.04998491`

- `qnorm(0.05) # -1.645`
- lower.tail: logical; if TRUE (default), probabilities are $P[X \leq x]$ otherwise, $P[X > x]$.

In the following figure, on the left is the lower quantile and on the right is the upper quantile.

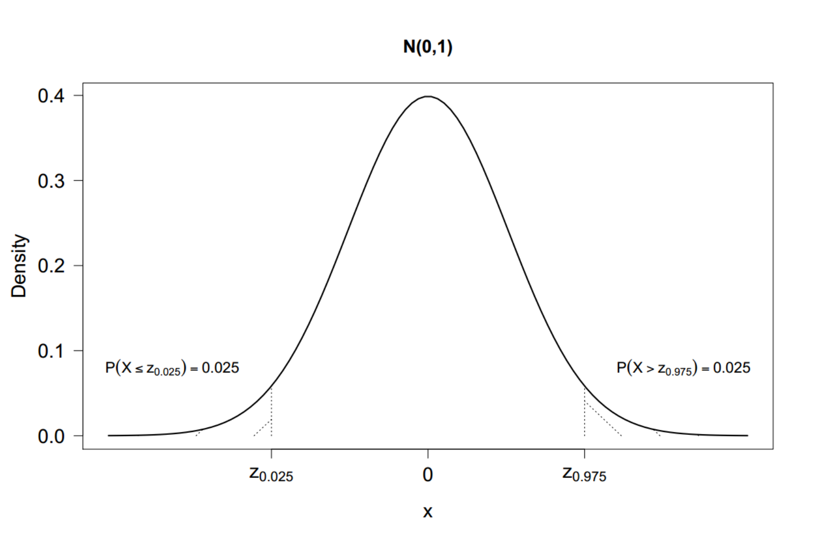

Figure 1

3 Hypothesis Testing

Here suppose our null hypothesis is $H_0 : \mu = \mu_0$ Random variable X is sampled from normal distribution with known mean and unknown variance, therefore the statistics $T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\hat{\sigma}}$ is from Student's t-distribution with $n - 1$ degree of freedom.

3.1 One-sided H_1

- The alternative hypothesis is $H_1 : \mu > \mu_0$.
- Reject H_0 if $T > t_{n-1,1-\alpha}$.
- Type I error: $\alpha = P_{\mu_0}(T > t_{n-1,1-\alpha})$
- Acceptance region: $T < t_{n-1,1-\alpha}$.

$$\begin{aligned} 1-\alpha &= P(\frac{(\sqrt{n}\bar{X}-\mu_0)}{\hat{\sigma}} \leq t_{n-1,1-\alpha}) \\ &= P(\bar{X} - t_{n-1,1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu_0) \end{aligned}$$

3.2 Two-sided H_1

- The alternative hypothesis is $H_1 : \mu \neq \mu_0$.
- Reject H_0 if $|T| > t_{n-1,1-\alpha}$: $T > t_{n-1,1-\alpha}$ or $T < -t_{n-1,1-\alpha}$.
- Acceptance region: $\frac{\sqrt{n}|\bar{X}-\mu_0|}{\hat{\sigma}} < t_{n-1,1-\alpha}$.

$$\begin{aligned} 1-\alpha &= P(\frac{\sqrt{n}|\bar{X}-\mu_0|}{\hat{\sigma}} \leq t_{n-1,1-\alpha}) \\ &= P(-t_{n-1,1-\alpha} \leq \frac{\sqrt{n}(\bar{X}-\mu_0)}{\hat{\sigma}} \leq t_{n-1,1-\alpha}) \\ &= P(\bar{X} - t_{n-1,1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_{n-1,1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}}) \end{aligned}$$