Single Sample Inference

August 29, 2017

Inferences on Mean

$X_1, X_2, \ldots, X_n$ is a random sample from some distribution with unknown mean μ. Interest is on estimating μ or on testing a hypothesis about μ of the form $H_0 : \mu \in \Theta$

Inferences on Mean - Normal population

Case 1: Suppose the sample is from a normal population: $X_1, X_2, \ldots, X_n$ iid $N(\mu, \sigma^2)$, where μ is the mean of the distribution and σ^2 is the variance.

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ is the maximum likelihood estimator of μ.

Inferences will use $\bar{X}$ as a point estimator for μ and to construct confidence intervals and hypothesis tests.

Inferences on Mean - Normal population - σ^2 known

Inferences depend on whether σ^2 is known or unknown.

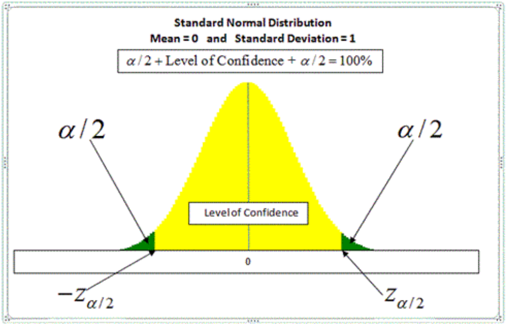
If σ^2 is known then $\bar{X} \sim N(\mu, \sigma^2/n)$ and $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

$$
\begin{aligned}
P(-z_{\alpha/2} < \tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) &= 1 - \alpha \\
P(-z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(-\bar{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(\bar{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}) &= 1 - \alpha
\end{aligned}
$$

A $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Inferences on Mean - Normal population - σ^2 known



Inferences on Mean - Normal population - σ^2 known

Hypothesis testing: one-tailed or two tailed tests

- One-tailed tests
  - $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$
  - $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$
- Two tailed tests
  - $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$

Inferences on Mean - Normal population - σ^2 known

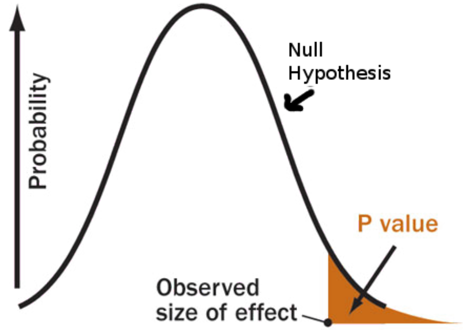
The test statistic for each case is the same - the only difference is how the p-value - P(getting something equal to or more extreme than what you observe given H_0 is true).

Test statistic: If $\bar{x}$ is the value of the sample mean calculated from the sample

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Inferences on Mean - Normal population - σ^2 known

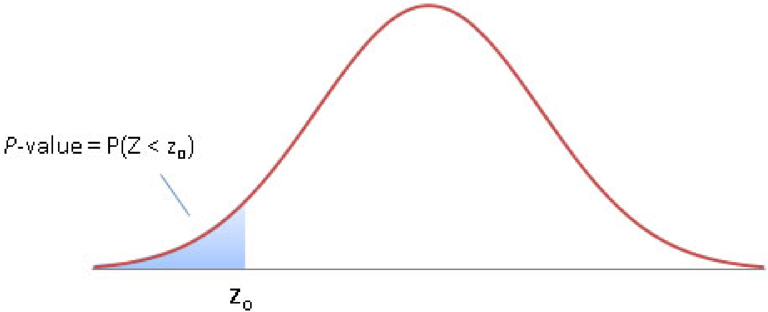
$H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$



$$p - value = P(Z > z_0)$$

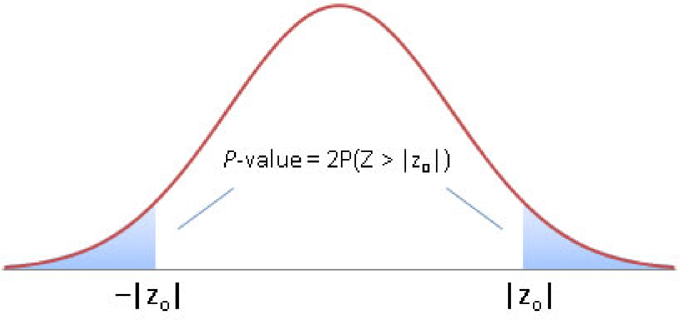
Inferences on Mean - Normal population - σ^2 known

$H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$



$$p - value = P(Z < z_0)$$

$H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$



$$p - value = 2P(Z > |z_0|)$$

Inferences on Mean - Normal population - σ^2 unknown

When σ^2 is unknown we estimate it with the sample variance:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

and

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Inferences are the same with the standard normal distribution replaced by the t-distribution with $n-1$ degrees of freedom.

Inferences on Mean - Normal population - σ^2 unknown

A $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

Test statistic is

$$t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Null Hypothesis	Alternate Hypothesis	p-value
$H_0 : \mu \leq \mu_0$	$H_1 : \mu > \mu_0$	$P(T_{n-1} > t_0)$
$H_0 : \mu \geq \mu_0$	$H_1 : \mu < \mu_0$	$P(T_{n-1} < t_0)$
$H_0 \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$2P(T_{n-1} > \|t_0\|)$

Note the p-value for a two tailed test is twice what it would be for a one-tailed test. For this reason you should not assume a one tailed test unless you have strong evidence or knowledge that suggests a one-tailed test. You should decide on a hypothesis to test **before** you do the analysis otherwise you are guilty of *data snooping*.

Inferences on Mean - Non-normal population

In this case $X_1, X_2, \ldots, X_n$ is a random sample from a population with unknown mean μ that is not normally distributed.

- if n is small:
  - Use a bootstrap or other resampling method to estimate the mean with a point estimate or confidence interval.
  - Use a non-parametric test such as the Wilcoxon test to conduct a hypothesis test.
- if n is large (how large depends on how far from normality the population is)
  - the central limit theorem $\bar{X} \to N(\mu, \sigma^2/n)$ where μ is the population mean and σ^2 is the population variance. So for large n, an approximate analysis assuming normality is appropriate.

Inferences on Mean - Non-normal population

Estimation using resampling:

- Draw a sample with replacement from the original sample. Compute the value of $\bar{X}$ in that sample
- Repear N times.
- Use the N results to
  - calculate the estimate of variance ($Var(\bar{X})$
  - produce a histogram or boxplot to look at the shape of the sampling distribution of $\bar{X}$
  - construct a confidence interval - order the N $\bar{X}$'s and use as lower limit the value that has $N * \alpha/2$ values below it and the upper limit is the value that has $N * \alpha/2$ values above it.

Inferences on Mean - Non-normal population

Wilcoxon test: (a rank based procedure) For testing $H_0 : \mu = \mu_0$ versus a two-tail or one-tail alternative.

- For each sample point compute $D_i = X_i - \mu_0$, the difference of each point from the hypothesized mean, μ_0.
- Obtain the ranks of differences, R_i is the rank of D_i
- Define

$$R_i = \begin{cases} R_i & \text{if } D_i > 0 \\ -R_i & \text{if } D_i < 0 \end{cases}$$

- The test statistic is $T^+ = \sum(\text{all possible } R_i\text{'s})$.
- Tables for the distribution of T^+ are available for small n
- For larger $n > 50$ use a normal approximation

$$T = \frac{\sum_{i=1}^n R_i}{\sqrt{\sum_{i=1}^n R_i^2}} \sim N(0,1)$$

under H_0.

Inference on Means

- In general procedures based on the normal distribution will result in shorter confidence intervals and more power for the hypothesis tests. Power is the probability of rejecting the null hypothesis when it is false. In general, procedures with higher power are preferred.
- In general, inferences on means are robust to non-normality - meaning that the methods will give good results even for non-normal populations
- Inferences on means are **not** robust to outliers

Inference on variances - Normal Populations

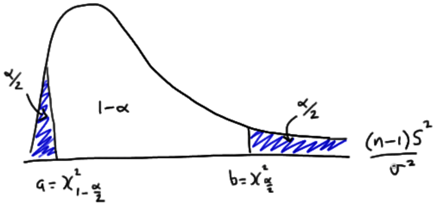
$X_1, X_2, \ldots, X_n$ is a random sample from a normal population $N(\mu, \sigma^2)$. Then for $S^2 = \dfrac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ we have

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

We will use this sampling distribution to construct confidence intervals and perform hypothesis testing.

Confidence intervals

$$P\left(x^2_{n-1,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < x^2_{n-1,\alpha/2}\right) = 1-\alpha$$



Inference on variances - Normal Populations

The resulting $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

This is not a symmetric confidence interval because the chi-square distribution is not symmetric.

Hypothesis Testing: As before we can have one tailed or two tailed tests for σ^2.

- One-tailed tests
  - $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$
  - $H_0 : \sigma^2 \geq \sigma_o^2$ versus $H_1 : \sigma^2 < \sigma_0^2$
- Two-tailed test
  - $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$

Inference on variances - Normal Populations

We use test statistic

$$T = \frac{(n-1)S^2}{\sigma_0^2}$$

which has a chi-square distribution with $n-1$ degrees of freedom under the null hypothesis.

Inference on variances - Normal Populations

Null Hypothesis	Alternative Hypothesis	p-value
$H_0 : \sigma^2 \leq \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	$P(\chi^2_{n-1} > T)$
$H_0 : \sigma^2 \geq \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$P(\chi^2_{n-1} < T)$
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$	$2 * \min(P(\chi^2_{n-1} > T), P(\chi^2_{n-1} < T))$

Inference on variance

- The inference on variance is not robust to non-normality. If the population is not normally distributed, T does not have a chi-square distribution and there is no central limit theorem for variances.
- If your population is not normal, use a bootstrap for confidence intervals and non-parametric hypothesis tests.

Transformations

If your data is not normally distributed but you really want to use a method that assume normality, sometimes a transformation will result in variates that appear more normal - especially if your data is skewed.
Box-Cox Transformation

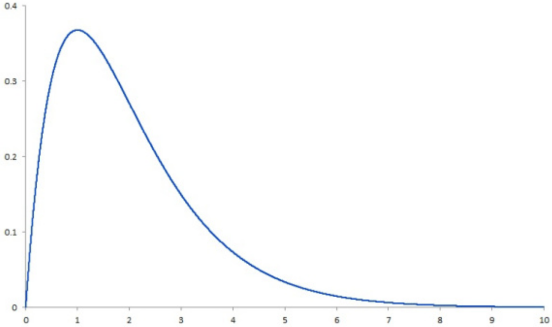
$$Y^\lambda = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

Choose the value of λ that maximizes

$$l(\lambda) = -\frac{n}{2} \ln S_\lambda^2 + (\lambda - 1) \sum_{i=1}^{n} \ln(x_i)$$

where S_λ^2 is the sample variance of the transformed data.

Right skewed - try $\ln(x), x^{-1}, x^{-2}$, etc.



Transformations

Left skewed - try y^2, y^3, etc.