MATH 324: Statistics

Julian Lore

Last updated: February 12, 2018

Rough notes from Wackerly's Mathematical Statistics with Applications (7^{th} edition).

Contents

Information from Previous Chapters

Some formulas from probability:

$$Var(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$$

Distributions:

Distribution	Probability Function	Mean	Variance	MGF
Binomial	$p(y) = \binom{n}{y} p^y (1-p)^{n-y}$	np	$np(1-p)$	$[pe^t + (1-p)]^n$
Geometric	$p(y) = p(1-p)^{y-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Hypergeometric	$p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$	$\frac{nr}{N}$	$n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$	No closed form
Poisson	$p(y) = \frac{\lambda^y e^{-\lambda}}{y!}$	λ	λ	$e^{\lambda(e^t-1)}$
Negative binomial	$p(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^r$
Uniform	$f(y) = \frac{1}{\theta_2-\theta_1}$	$\frac{\theta_1+\theta_2}{2}$	$\frac{(\theta_2-\theta_1)^2}{12}$	$\frac{e^{t\theta_2}-e^{t\theta_1}}{t(\theta_2-\theta_1)}$
Normal	$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}\right)(y-\mu)^2}$	μ	σ^2	$e^{\mu t + \frac{t^2\sigma^2}{2}}$
Exponential	$f(y) = \frac{1}{\beta} e^{-\frac{y}{\beta}}$	β	β^2	$(1-\beta t)^{-1}$
Gamma	$f(y) = \left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right) y^{\alpha-1} e^{-\frac{y}{\beta}}$	$\alpha\beta$	$\alpha\beta^2$	$(1-\beta t)^{-\alpha}$
Chi-square	$f(y) = \frac{y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}$	ν	2ν	$(1-2t)^{-\frac{\nu}{2}}$
Beta	$f(y) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) y^{\alpha-1}(1-y)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	No closed form

6.7 Order Statistics

Y_i, with $i = 1, \ldots, n$ independent continuous random variables. Denote ordered random variables by: $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$. We can then define:

$$Y_{(1)} = \min(Y_1, Y_2, \ldots, Y_n)$$
$$Y_{(n)} = \max(Y_1, Y_2, \ldots, Y_n)$$

In order to find the pdf of $Y_{(1)}$ and $Y_{(n)}$, we use the method of distribution functions.

$$\begin{aligned} F_{Y_{(n)}}(y) &= P(Y_{(n)} \leq y) = P(Y_1 \leq y, Y_2 \leq y, \ldots, Y_n \leq y) \\ &\overset{\text{ind}}{=} P(Y_1 \leq y)P(Y_2 \leq y)\ldots p(Y_n \leq y) = [F(y)]^n \end{aligned}$$

Derive for the density:

$$g_{(n)}(y) = n[F(y)]^{n-1} f(y)$$

$$\begin{aligned} F_{Y_{(1)}}(y) &= P(Y_{(1)} \leq 1) = 1 - P(Y_{(1)} > y) = 1 - 1 - P(Y_1 > y, Y_2 > y, \ldots Y_n > y) \\ &\overset{\text{ind}}{=} 1 - [P(Y_1 > y)P(Y_2 > y)\ldots P(Y_n) > y] = 1 - [1 - F(y)]^n \end{aligned}$$

Derive for the density:

$$g_{(1)}(y) = n[1 - F(y)]^{n-1} f(y)$$

Other Useful Information

$$\begin{aligned}
E(\overline{Y}) &= E(Y) \\
Var(\overline{Y}) &= \frac{\sigma^2}{n} \\
Var(c) &= 0, \text{ where } c \text{ is a constant} \implies Var(X + c) = Var(X) \\
E[aX + bY] &= aE[X] + bE[Y] \\
Var(aX) &= a^2 Var(X) \\
Var(aX + bY) &= a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y) \\
Var(aX - bY) &= a^2 Var(X) + b^2 Var(Y) - 2abCov(X, Y) \\
\frac{(n-1)^2 S^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sim \chi^2_{n-1} \text{ (Theorem 7.3)}
\end{aligned}$$

8 Estimation

8.1 Introduction

Point of statistics is to use sample information to infer data about the population. Populations are characterized by numbers (*parameters*) and we often want to estimate the value of parameter(s). Parameters include the proportion p, population mean μ, variance σ^2 and standard deviation σ.

Definition 1. The parameter of interest in an experiment is called the *target parameter*.

Definition 2. A *point estimate* is a type of estimate where we use a single value/point to estimate a parameter. If we estimate a parameter by saying that it might fall between two numbers, this is an *interval estimate*. We can use information from the sample to calculate these estimates, which are done using an estimator.

Definition 3. An *estimator* is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample.

Definition 4. *Sample mean:*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This is an example point estimator of μ.

There can be different estimators for the same population parameter. Some estimators are considered good and others are bad.

8.2 The Bias and Mean Square Error of Point Estimators

We cannot measure how good a point estimation procedure is with a single estimate, we need to observe the procedure many times. We create a frequency distribution to measure the goodness of a point estimator.

Point Estimators For a population parameter θ, the estimator of θ is called $\hat{\theta}$.

Definition 5. Ideally, we'd want $E(\hat{\theta}) = \theta$. Point estimators that satisfy this are called *unbiased*. Otherwise, they are called *biased*, where the *bias* is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

In addition, we'd also like the estimator $V(\hat{\theta})$ to be as small as possible, since a smaller variance guarantees a higher fraction of estimators to be "close" to θ. If two estimators are unbiased and everything else is equal other than variance, we prefer the one with smaller variance.

Definition 6. Another way to characterize goodness of a point estimator is via its *mean square error*,

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Which is the average of the square of the distance between the estimator and its target parameter. It can be shown that:

$$MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$$

8.3 Some Common Unbiased Point Estimators

Definition 7. $\sigma^2_{\hat{\theta}}$ denotes the variance of the sampling distribution of the estimator $\hat{\theta}$. The standard deviation $\sigma_{\hat{\theta}} = \sqrt{\sigma^2_{\hat{\theta}}}$ is called the *standard error* of the estimator $\hat{\theta}$.

Common Point Estimators If random samples are independent we have:

Target Parameter θ	Sample Size(s)	Point Estimator $\hat{\theta}$	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
μ	n	$\overline{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1, n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1, n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

Note, these are all **unbiased** estimators. Note that the expected values and standard errors for $\overline{Y}$ and $\overline{Y}_1 - \overline{Y}_2$ are valid for any distribution of the population and all four estimators are approximately normal for large n (see CLT).

Definition 8. The sample variances,

$$S^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$$

$$S'^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n}$$

are unbiased (S^2) and biased (S'^2) estimators for σ^2.

8.4 Evaluating the Goodness of a Point Estimator

Definition 9. The *error of estimation* ε is the distance between an estimator and its target parameter:

$$\varepsilon = |\hat{\theta} - \theta|$$

Note that $\hat{\theta}$ is a random variable, so the error of estimation is a random quantity and thus we can make probability statements about it (rather than saying exactly how big it is for some estimate).

The probability $P(\varepsilon < b)$ represents the fraction of times, in repeated sampling that $\hat{\theta}$ falls within b units of θ.

Sometimes we want to find a value of b such that $P(\varepsilon < b) =$ some value, say 0.90.

If $\hat{\theta}$ is unbiased we can find a bound on ε by expressing b as a multiple of the standard error. Take for example, $b = k\sigma_{\hat{\theta}}$, where $k \geq 1$, then by Tchebysheff's we know that $P(\varepsilon < k\sigma_{\hat{\theta}}) \geq 1 - \frac{1}{k^2}$. $b = 2\sigma_{\hat{\theta}}$ is a good approximate bound on ε in practice. By Tchebysheff we know that $P(\varepsilon < 2\sigma_{\hat{\theta}}) \geq .75$.

8.5 Confidence Intervals

Definition 10. An *interval estimator* is a rule specifying the method for using sample measurements to calculate two numbers (endpoints) of the interval. We want the interval to contain θ and for it to be **narrow**. One or both of the endpoints are functions of the sample measurements and thus vary randomly from sample to sample, i.e. cannot be sure that θ falls in the interval, so we want an interval estimator with a high probability of containing θ.

Definition 11. Interval estimators are commonly called *confidence intervals*.

Definition 12. Upper/lower endpoints are called *upper* and *lower confidence limits*, sometimes denoted $\hat{\theta}_L$ and $\hat{\theta}_U$.

Definition 13. Probability that a random confidence interval will enclose θ is called the *confidence coefficient*, denoted by $(1-\alpha)$. In practice, this coefficient describes the fraction of time that random/repeated sampling intervals will contain θ.

We have:

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

Definition 14. $[\hat{\theta}_L, \hat{\theta}_U]$ is called a *two-sided confidence interval.* A *one-sided confidence interval* satisfies $[\hat{\theta}_L, \infty)$, an upper one-sided confidence interval is $(-\infty, \hat{\theta}_U]$.

Definition 15. The *pivotal method* is very useful for finding confidence intervals. We find a pivotal quantity that satisfies:

1. Being a function of sample measurements and θ, where θ is the only unknown quantity.
2. Probability distribution does not depend on θ.

We can then say:

$$\begin{aligned} P(a \leq Y \leq b) = p &\implies P(ca \leq cY \leq cb) = p \\ &\implies P(a + d \leq Y + d \leq b + d) = p \end{aligned}$$

i.e. A change of scale or translation of Y does not change the probability. So knowing the probability distribution of a pivotal quantity, we can use these operations and get the interval.

8.6 Large-Sample Confidence Intervals

If $\theta = \mu, p, \mu_1 - \mu_2$ or $p_1 - p_2$ then for large samples

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \overset{approx}{\sim} Normal(0, 1)$$

Also, Z is (approximately) a pivotal quantity.

We also have:

$$\begin{aligned} 100(1-\alpha)\% \text{ lower bound for } \theta &= \hat{\theta} - z_\alpha \sigma_{\hat{\theta}} \\ 100(1-\alpha)\% \text{ upper bound for } \theta &= \hat{\theta} + z_\alpha \sigma_{\hat{\theta}} \\ \implies 100(1-2\alpha)\%\theta &= \hat{\theta} \pm z_\alpha \sigma_{\hat{\theta}} \end{aligned}$$

The important formula of this section is

$$\hat{\theta} \pm z_{\frac{\alpha}{2}} \sigma_{\hat{\theta}}$$

When $\theta = \mu, \hat{\theta} = \overline{Y}$ and $\sigma_{\hat{\theta}}^2 = \frac{\sigma^2}{n}$, then σ^2 should be used if known, but if it is not known and n is large, we can use s^2 instead. Similarly for σ_1^2 and σ_2^2, for $\theta = \mu_1 - \mu_2$. For $\theta = p$, we can replace p by $\hat{p}$ for large n (to be justified in Section 9.3).

8.7 Selecting the Sample Size

We want to obtain information at minimum cost. Note that the sample size n controls the amount of relevant information in a sample. How many measurements should we include in our sample? We first need to know how much information we'd like to obtain. Specifically, how **accurate** should the estimate be?

For example, if we'd like to estimate μ within 5 units with probability .95: since approximately 95% of the sample means will lie within $2\sigma_{\overline{Y}}$ of μ, then we want

$$2\sigma_{\overline{Y}} = 5 \implies \frac{2\sigma}{\sqrt{n}} = 5 \implies n = \frac{4\sigma^2}{25}$$

8.8 Small-Sample Confidence Intervals for μ and $\mu_1 - \mu_2$

We assume in this section that our sample has a normal distribution. We want to approximate the population mean when we don't know the variance or the sample size is too small to apply large-sample techniques. But we can use:

$$T = \frac{\overline{Y} - \mu}{\frac{S}{\sqrt{n}}}$$

which has a t distribution with $(n-1)$ df. T is a pivotal quantity for μ and we can get:

$$P(-t_{\frac{\alpha}{2}} \leq T \leq t_{\frac{\alpha}{2}}) = 1 - \alpha$$

The t distribution is similar to the normal distribution, except the tails are thicker.

Important Parameters and Confidence Intervals

Parameter	Confidence Interval
μ	$\overline{Y} \pm t_{\frac{\alpha}{2}} \left(\frac{S}{\sqrt{n}}\right), df = n-1$
$\mu_1 - \mu_2$	$(\overline{Y}_1 - \overline{Y}_2 \pm t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}), df = n_1 + n_2 - 2$

Where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

Note that as n gets large, so does df and eventually, the t distribution can be approximated by the standard normal distribution (the small n intervals become nearly equivalent to large n samples when $df > 30$).

8.9 Confidence Intervals for σ^2

Recall that

$$\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

has a χ^2 distribution with $n-1$ df.

Using the pivotal method, we get:

$$P\left[\chi_L^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_U^2\right] = 1 - \alpha$$

Since the χ^2 density function is not symmetric, we get to chose χ_L^2 and χ_U^2. We usually choose points that cut off equal tail areas:

$$P\left[\chi^2_{1-(\frac{\alpha}{2})} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{(\frac{\alpha}{2})}\right] = 1 - \alpha$$

Obtaining:

$$P\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-(\frac{\alpha}{2})}}\right] = 1 - \alpha$$

$100(1-\alpha)\%$ **Confidence Interval for** σ^2

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

8.10 Summary

The main objective of most statistical investigations is to make inferences about population parameters based on sample data. These inferences are often estimated (point or interval estimates). Unbiased estimators with small variances are preferred. Goodness of unbiased $\hat{\theta}$ is measured via $\sigma_{\hat{\theta}}$ because error of estimation is usually smaller than $2\sigma_{\hat{\theta}}$. The MSE is small if the variance and bias is small.

Interval estimates of many parameters like μ and p can be derived from the normal distribution for large n because of CLT. If n is small, we must assume normality of the population and use the t distribution for confidence intervals.

If sample measurements are from a normal distribution, we can get a confidence interval for σ^2 through the χ^2 distribution.