14.381 Statistics: Handout #4

José Tessada*

October 22, 2004

1 Random Sampling

Definition 1 *The random variables $X_1, ..., X_n$ are called a random sample of size n from the population $f(x)$ if $X_1, ..., X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, ..., X_n$ are called independent and identically distributed random variables with pdf or pmf $f(x)$. This is commonly abbreviated to iid random variables.*

Notes: observations are taken in a way such that $X_1, ..., X_n$ are *mutually independent*. Then,

$$f(x_1, ..., x_n|\theta) = f(x_1|\theta) f(x_2|\theta) \cdot ... \cdot f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \qquad (1)$$

The previous formula assumes that the random variables are members of a parametric family.

Note some of the implicit assumptions used in the derivation of (1). As Casella and Berger (C&B) explain in page 209, when you have a finite sample we need to specify if we draw with or without replacement.

Now, dealing with the entire random sample can be cumbersome if n is very large, so we might be interested in summarizing the info we really care about. *Statistics* are functions of random variables which help us to summarize the info contained in a random sample.

Definition 2 *Let $X_1, ..., X_n$ be a random sample of size n from a population and let $T(x_1, ..., x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, ..., X_n)$. Then the random variable or random vector $Y = T(X_1, ..., X_n)$ is called a* ***statistic****. The probability distribution of a statistic is called the* ***sampling distribution*** *of Y.*

Now, this definition is very broad. However, it does rule out any function $T(\cdot)$ which is also a function of a parameter, you should be very clear about this.

*Usual disclaimers apply. Comments are welcome. Email: tessada@mit.edu

Definition 3 *Some popular and useful statistics:*

1. *The sample mean is the arithmetic average of the values in a random sample.*

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2}$$

2. *The sample variance is the statistic defined by*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2. \tag{3}$$

3. *The sample standard deviation is defined by $S = \sqrt{S^2}$, where S^2 is defined as in (3).*

It is clear that all these statistics are function only of $\overrightarrow{X} = (X_1, ..., X_n)$ as it is stated in Definition 2.

Theorem 4 *Let $x_1, ..., x_n$ be any numbers and $\overline{x} = (x_1 + \cdots + x_n)/n$. Then,*

1. $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$,
2. $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2$.

Proof. I will proof part 2 only; you can check the proof for part 1 in the book (C&B page 212).

The first equality in part 2 is trivial, it comes from the definition of s^2. To prove the second part, start from (3)

$$\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}\left(X_i^2 - 2\cdot X_i \cdot \overline{X} + \overline{X}^2\right) \\
&= \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\overline{X}^2 - 2\sum_{i=1}^{n} X_i \cdot \overline{X} \\
&= \sum_{i=1}^{n} X_i^2 + n\overline{X}^2 - 2\overline{X}\underbrace{\sum_{i=1}^{n} X_i}_{=n\overline{X}} \\
&= \sum_{i=1}^{n} X_i^2 + n\overline{X}^2 - 2n\overline{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2
\end{aligned}$$

Proving part 2 of the Theorem. ■

Theorem 5 *Let $X_1, ..., X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

1. $E\left(\overline{X}\right) = \mu$,
2. $Var\left(\overline{X}\right) = \frac{\sigma^2}{n}$,
3. $E\left(S^2\right) = \sigma^2$.

The proof of the theorem can be found in the book (C&B, page 214).

Part 3 of Theorem 3 shows the effect of dividing by $(n-1)$ in the definition of S^2 In fact, if we define $\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ we can show that

$$\begin{aligned} E\left(\widetilde{S}^2\right) &= \frac{1}{n}E\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}E\left(X_i^2\right) - nE\left(\overline{X}^2\right)\right] = \frac{1}{n}\left[\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\ &= \left(\sigma^2+\mu^2\right) - \left(\frac{\sigma^2}{n}+\mu^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \frac{(n-1)}{n}\sigma^2 \neq \sigma^2. \end{aligned}$$

So $\widetilde{S}^2$ is not an *unbiased estimator* of the variance. Unbiasedness is a concept you will see soon. Dividing by $(n-1)$ reflects the loss of one *degree of freedom* because instead of using μ we use $\overline{X}$. Intuitively, we have n data points. We first calculated $\overline{X}$ which places one constraint upon the n data points $\left(\sum X_i = n\overline{X}\right)$. This leaves $n-1$ unconstrained observations with which to estimate the sample variance[1]. If we know μ, then we can compute $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$ and we divide by n because we can use all the information contained in the data points.

2 Random Sampling from the Normal Distribution

Theorem 6 *Let $X_1, ..., X_n$ be a random sample from a $N\left(\mu, \sigma^2\right)$ distribution, and let $\overline{X}$ and S^2 be defined as in (2) and (3). Then*

1. *$\overline{X}$ and S^2 are independent random variables,*
2. *$\overline{X}$ has a $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution,*
3. *$(n-1)S^2/\sigma^2$ has a χ^2 distribution with $(n-1)$ degrees of freedom.*

Instead of a proof, I will illustrate this results by simulating a series of random samples from a $N(0, 10)$ distribution.

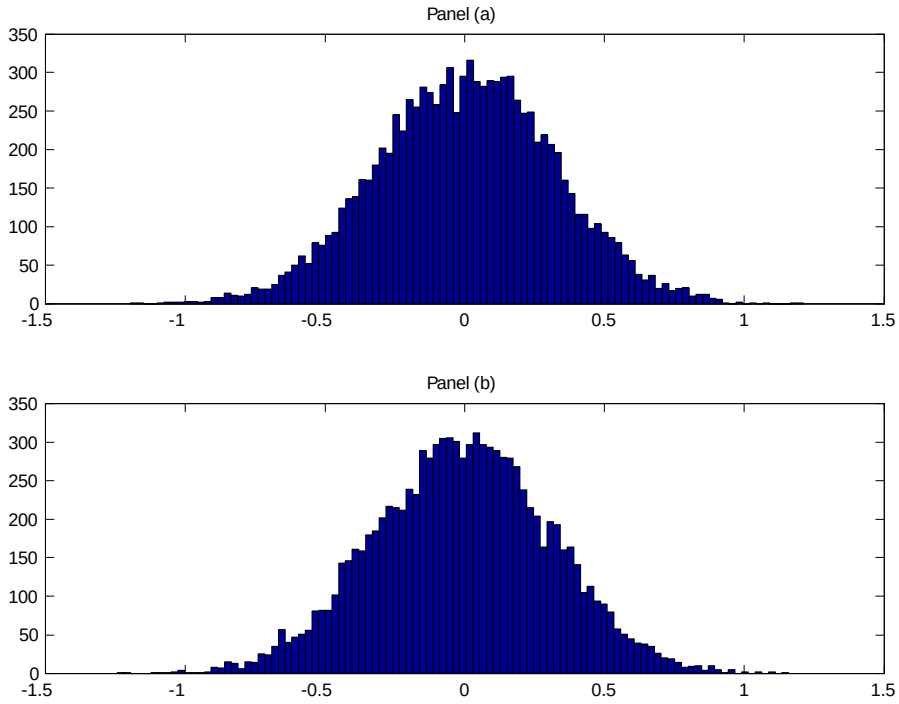

Figure 1: Panel (a) shows the histogram of S^2 from the simulated series. Panel (b) shows the histogram of a random sample of the same size drawn from the theoretical distribution of S^2.

Lemma 7 *Use χ_p^2 to denote a chi squared random variable with p degrees of freedom.*

1. *If $Z \sim N(0,1)$, then $Z^2 \sim \chi_1^2$; that is, the square of a standard normal random variable is a chi squared random variable.*
2. *If $X_1, ..., X_n$ are independent $\chi_{p_i}^2$, then $\sum_{i=1}^n X_i \sim \chi_{p_1+\cdots+P_n}^2$; that is, independent χ^2 variables add to a χ^2 variable, and the degrees of freedom also add.*

Part (2) can be proved by the fact that the χ_p^2 is a particular case of the Γ distribution (χ_p^2 is a $\Gamma(p/2, 2)$). Part (1) can be shown by deriving the distribution of the square if a normal standard random variable.

2.1 The Student's t and Snedecor's F Distributions

One very useful statistic is

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

but we do not know σ^2, so we need to use an estimate of σ^2. A natural candidate is S^2, so we compute

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}, \tag{4}$$

Now, a distribution for that number is needed!

Definition 8 *Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. The quantity $\left(\overline{X} - \mu\right) / \left(S/\sqrt{n}\right)$ has Student's t distribution with $n-1$ degrees of freedom. Equivalently, a random variable T has Student's t distribution with p degrees of freedom, and we write $T \sim t_p$ if it has pdf*

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1 + t^2/p)^{(p+1)/2}}, \qquad -\infty < t < \infty \tag{5}$$

[2]

Definition 9 *Let $X_1, ..., X_n$ be a random sample from a $N(\mu_X, \sigma_X^2)$ population, and let $Y_1, ..., Y_m$ be a random sample from a $N(\mu_Y, \sigma_Y^2)$ population. The random variable $F = \left(S_X^2/\sigma_X^2\right) / \left(S_Y^2/\sigma_Y^2\right)$ has Snedecor's F distribution with $n-1$ and $m-1$ degrees of freedom. Equivalently, the random variable F has the Fdistribution with p and q degrees of freedom if it has pdf*

[1] Taken from section notes by Ziad Nejmeldeen.

[2] In order to see why this applies to (4) we need to use the theorem for independence of functions of independent random variables. Moshe was right when he corrected me during the recitation stating the more general result of the theorem, thanks.

p

$$f_F(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}\frac{x^{(p/2)-1}}{(1+(p/q)\,x)^{(p+q)/2}}, \qquad 0<t<\infty \qquad (6)$$

Theorem 10 *1. If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$.*

2. If $X \sim t_q$, then $X^2 \sim F_{1,q}$.

3. If $X \sim F_{p,q}$, then $(p/q)\,X/\left(1+(p/q)\,X\right) \sim beta\,(p/2, q/2)$.