Chapter 8 Summary and Review (draft: 2019/12/09-23:39:07)

CLT, LLN, χ^2, and T

Summary of topics and terminology:

- For X_i $i = 1, 2, ..., n$ i.i.d. (independent and identically distributed) with mean $\mathrm{E}(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$, the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has mean μ and variance $\frac{\sigma^2}{n}$, that is $\mathrm{E}(\overline{X}_n) = \mu$ and variance $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$
- If the X_i are normally distributed, then $\overline{X}_n$ is normally distributed.
- Law of large numbers (LLN): Using Chebyshev's inequality $P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$ we get that
  $P(|\overline{X}_n - \mu| < k\frac{\sigma}{\sqrt{n}}) \geq 1 - \frac{1}{k^2}$
  thus letting $k = a\frac{\sqrt{n}}{\sigma}$ for some $a > 0$ we get
  $P(|\overline{X}_n - \mu| < a) \geq 1 - \frac{\sigma^2}{a^2 n}$
  So we can see that for any fixed a $P(|\overline{X}_n - \mu| < a)$ goes to zero as n gets large.
  This is called *convergence in probability.* We say that the sample mean converges to μ in probability as the sample size goes to infinity. This is why with a large number of coin flips, we expect the proportion of heads to be extremely close to 50%, for example.
- CENTRAL LIMIT THEOREM (CLT): If the X_i are **not** normally distributed, then $\overline{X}_n$ is approximately normally distributed as long as n is large enough.
  - ➠ $\overline{X}_n \overset{\text{approx}}{\sim} \mathrm{N}(\mu, \frac{\sigma^2}{n})$.
  - ➠ $\sum_{i=1}^{n} X_i \overset{\text{approx}}{\sim} \mathrm{N}(n\mu, n\sigma^2)$.
  - ➠ Usually $n > 30$ makes the approximation somewhat reasonable, but for distributions that have extremely large higher order moments (or infinite ones!) it can require extremely large n values.
- If $Z \sim \mathrm{N}(0, 1)$ then $Z^2 \sim \chi^2(\nu = 1)$ chi-square distribution with parameter $\nu = 1$.
- Now we assume $X_i \sim \mathrm{N}(\mu, \sigma^2)$
  - ➠ $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(\nu = n - 1)$ (chi-square distribution with $n - 1$ degrees of freedom)
  - ➠ $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim T(\nu = n - 1)$ (t-distribution with $n - 1$ degrees of freedom)
  - ➠ We use these fact to calculate probabilities about the sample variance, and the smaple means standardized by the sample variance.
- Our discussion of chapter 8 stopped here, we didn't discuss finite populations, the F-distributions, or order statistics.

Example problems:

1. Let $X_1, X_2, ..., X_{25}$ be i.i.d. normal with mean 100 and variance 1.
   (a) Calculate the probability that the sample mean is less than 99.4.
   (b) Calculate the probability that the sample variance is greater than 1.5.
   (c) Calculate the probability that $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} > 1$. Note that this can be rewritten as $\overline{X} > \mu + \frac{S}{5}$.
   (d) How large of a sample size would we need to guarantee that $\overline{X}_n$ is within 0.01 of 100 with probability of at least 99.7%?

Solution:

(a) $\overline{X}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ thus $\overline{X} \sim N(100, \frac{1}{25})$. $100 - 99.4 = 0.6 = 3/5$ this is 3 standard deviations below the mean for $\overline{X}$. Using the 68-95-99.7 rule the answer is approximately 0.15%.

So if our sample mean happened to be so low, then we either have an extremely rare sample, or maybe we are incorrect in our knowledge that the X_i's are coming from this distribution. This last statement is a lead in to statistical reasoning.

(b) $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(\nu = n-1)$ thus $24S^2 \sim \chi^2(\nu = 24)$ and
$P(S^2 > 1.5) = P(24S^2 > 36) = \int_{36}^{\infty} f(x; 24)dx$ where $f(x; 24)$ is the chi-square distribution with $\nu = 24$. This is not an integral you need to do by hand. You would look this up in a table or use software.

In R we can calculate this by `1-pchisq(36,df=24)` $\approx .0549$. So it would be fairly rare for our sample variance to be larger than 1.5.

(c) Calculate the probability that $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} > 1$. Note that this can be rewritten as $\overline{X} > \mu + \frac{S}{5}$.
For this problem $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\overline{X} - 100}{S/5}$
$P(\frac{\overline{X}-100}{S/5} > 1) = P(T_{24} > 1)$ where T_{24} is a t-distributed random variable with $\nu = 24$ degrees of freedom.
$P(T_{24} > 1) = \int_1^{\infty} f(t; 24)dt$ where $f(t; 24)$ is the t-distribution pdf with $\nu = 24$. Again, this is not an integral you can do by hand. You would look this up in a table or use software.

In R this can be calculated by `1-pt(1,24)` ≈ 0.1636.

(d) We just require 0.01 to be 3 standard deviations, since we know $\overline{X}$ is normally distributed with standard deviation $\frac{1}{\sqrt{n}}$. S we just set $0.01 = \frac{1}{\sqrt{n}}$ and solve to get $n = (\frac{1}{0.01})^2 = 100^2 = 10$ thousand.

2. How many fair coin flips would we need to do to guarantee that the proportion of heads is between 0.499 and 0.501 with at least 95% probability?

   Solution:

   In this case the X_i are our i.i.d. Bernoulli random variables and the sample mean is the proportion of heads. By the normal approximation to the binomial distribution we can see that the sample mean $\overline{X}$ is approximately normally distributed with mean $\mu = 0.5$ and variance $\sigma^2/n = \frac{1}{4n}$.

   Using again the 68-95-99.7 rule we see that we are asking to be within 2 standard deviations from the mean. So we want 0.001 to be twice the standard deviation of the sample mean, thus $0.001 = 2 \cdot \sqrt{\frac{1}{4n}} = \frac{1}{\sqrt{n}}$ thus $n = (\frac{1}{0.001})^2 = 1$ million.

   Intuitively, if after 1 million coin flips, you can a proportion of heads sufficiently far from 0.5, you might start to question whether the coin is actually fair. Again, this is getting into ideas we will explore in statistics.