STAT1011
Introduction to Statistics

Tutorial 5 Supplement

Chan Chun Ho, Benjamin[*†]

October 31, 2017

Abstract

It aims to supplement proofs of results used in normal distribution. The proofs are beyond the scope of this course and they serve the purpose of completeness. The target readers are students who are curious about the rigorous aspect of Statistics. Some materials are extracted from classic textbook "Statistical Inference" (Casella and Berger) used in STAT4003. Appendix contains formal definitions and discussion on related topics. More details can be found in tutorial 6.

1 Theorems and (Optional) Proofs

Theorem 1. If $X \sim N(\mu, \sigma^2)$, define $Z = \frac{X-\mu}{\sigma}$. Since X is a random variable, Z is also a random variable. Then, $Z \sim N(0, 1)$.

It therefore follows that all normal probabilities can be calculated in terms of the standard normal. The cdf of $N(\mu, \sigma^2)$ cannot be obtained in closed form. We have to transform it into $N(0, 1)$ in order to find the cumulative probabilities using a standard normal table.

Proof.

$$\begin{aligned}
P(Z \leq z) &= P\big(\frac{X-\mu}{\sigma} \leq z\big) \\
&= P(X \leq z\sigma + \mu) \\
&= \int_{-\infty}^{z\sigma+\mu} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \ , \ (\text{substitute } t = \frac{x-\mu}{\sigma})
\end{aligned}$$

showing that $P(Z \leq z)$ is the standard normal cdf. □

[*]For enquiry, please email to 1155049861@link.cuhk.edu.hk.
[†]Personal profile: www.linkedin.com/in/benjamin-chan-chun-ho

Theorem 2. If $Z \sim N(0,1)$, define $X = \mu + \sigma Z$. Since Z is a random variable, X is also a random variable. Then, $X \sim N(\mu, \sigma^2)$.

Proof.

$$\begin{aligned} P(X \le x) &= P(\mu + \sigma Z \le x) \\ &= P(Z \le \frac{x-\mu}{\sigma}) \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \text{ , (substitute } t = \mu + \sigma z) \end{aligned}$$

showing that $P(X \le x)$ is $N(\mu, \sigma^2)$ cdf. □

Theorem 3. $P(|X-\mu| \le \sigma) = P(|Z| \le 1) = 0.68$, $P(|X-\mu| \le 2\sigma) = P(|Z| \le 2) = 0.95$ and $P(|X-\mu| \le 3\sigma) = P(|Z| \le 3) = 0.997$.

Proof.

$$\begin{aligned} P(|X-\mu| \le \sigma) &= P(|Z| \le 1) \quad \text{, by Theorem 1} \\ &= P(-1 \le Z \le 1) \\ &= P(Z \le 1) - P(Z \le -1) \\ &= 0.8413 - 0.1587 \\ &= 0.6826 \\ &\approx 0.68 \end{aligned}$$

$$\begin{aligned} P(|X-\mu| \le 2\sigma) &= P(|Z| \le 2) \quad \text{, by Theorem 1} \\ &= P(-2 \le Z \le 2) \\ &= P(Z \le 2) - P(Z \le -2) \\ &= 0.9772 - 0.0228 \\ &= 0.9544 \\ &\approx 0.95 \end{aligned}$$

$$\begin{aligned} P(|X-\mu| \le 3\sigma) &= P(|Z| \le 3) \quad \text{, by Theorem 1} \\ &= P(-3 \le Z \le 3) \\ &= P(Z \le 3) - P(Z \le -3) \\ &= 0.9987 - 0.0013 \\ &= 0.9974 \\ &\approx 0.997 \end{aligned}$$

□

Theorem 4. If $Z \sim N(0,1)$, then $E(Z) = 0$ and $Var(Z) = 1$.

Proof.

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \left[-\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

$$\begin{aligned} E(Z^2) &= \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \left[-z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \quad \text{(integration by parts)} \\ &= 1, \end{aligned}$$

since the first integrand is 0 and the second integrand is 1 ($\int_{-\infty}^{\infty} f(z)dz = 1$ by Theorem 6).

$$\begin{aligned} Var(Z) &= E(Z^2) - [E(Z)]^2 \\ &= 1 - 0^2 \\ &= 1 \end{aligned}$$

□

Theorem 5. If $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$ and $Var(X) = \sigma^2$.

Proof. Using transformation of random variable,

$$\begin{aligned} Z &= \frac{X-\mu}{\sigma} \sim N(0,1) \quad \text{by Theorem 1} \\ \Rightarrow X &= \mu + \sigma Z \end{aligned}$$

$$\begin{aligned} E(X) &= \mu + \sigma E(Z) \\ &= \mu + \sigma \cdot 0 \qquad \text{by Theorem 4} \\ &= \mu \end{aligned}$$

$$\begin{aligned} Var(X) &= Var(\mu + \sigma Z) \\ &= E(\mu + \sigma Z - \mu)^2 && \text{by definition} \\ &= \sigma^2 E(Z^2) && \text{pull out non-random } \sigma^2 \\ &= \sigma^2 Var(Z) && \text{by } Var(Z) = E(Z^2) - [E(Z)]^2 = E(Z^2) \\ &= \sigma^2 \cdot 1 && \text{by Theorem 4} \\ &= \sigma^2 \end{aligned}$$

□

Theorem 6. If $Z \sim N(0,1)$ and $f(z)$ is the pdf of Z, then $\int_{-\infty}^{\infty} f(z)dz = 1$.

Proof. The target is to show that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1.$$

Notice that the integrand is symmetric around 0, implying that the integral over $(-\infty, 0)$ is equal to the integral over $(0, \infty)$. We need only to show that

$$\int_{0}^{\infty} e^{-\frac{z^2}{2}} dz = \frac{\sqrt{2\pi}}{2} = \sqrt{\frac{\pi}{2}}.$$

The function $e^{-\frac{z^2}{2}}$ does not have an antiderivative that can be written in closed form, so we cannot perform the integration directly. Since both sides are positive, the equality will hold if we establish that the squares are equal.

$$\begin{aligned}
\Big(\int_{0}^{\infty} e^{-\frac{z^2}{2}} dz\Big)^2 &= \Big(\int_{0}^{\infty} e^{-\frac{t^2}{2}} dt\Big)\Big(\int_{0}^{\infty} e^{-\frac{u^2}{2}} du\Big) \\
&= \int_{0}^{\infty} \int_{0}^{\infty} e^{-\frac{t^2+u^2}{2}} dt du.
\end{aligned}$$

To convert to polar coordinates. Define

$$t = r\cos\theta \quad \text{and} \quad u = r\sin\theta.$$

Then $t^2 + u^2 = r^2$ and $dtdu = rd\theta dr$. The limits of integration becomes $0 < r < \infty$ and $0 < \theta < \frac{\pi}{2}$ (the upper limit on θ is $\frac{\pi}{2}$ because t and u are restricted to be positive). So,

$$\begin{aligned}
\int_{0}^{\infty} \int_{0}^{\infty} e^{-\frac{t^2+u^2}{2}} dt du &= \int_{0}^{\infty} \int_{0}^{\frac{\pi}{2}} r e^{-\frac{r^2}{2}} d\theta dr \\
&= \frac{\pi}{2} \int_{0}^{\infty} r e^{-\frac{r^2}{2}} dr \\
&= \frac{\pi}{2} \Big[-e^{-\frac{r^2}{2}}\Big]_{0}^{\infty} \\
&= \frac{\pi}{2},
\end{aligned}$$

which establishes the result. $\square$

Theorem 7. If $X \sim N(\mu, \sigma^2)$ and $f(x)$ is the pdf of X, then $\int_{-\infty}^{\infty} f(x)dx = 1$.

Proof. Using transformation of random variable,

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \text{ by Theorem 1}$$

Let $f(z)$ be the pdf of transformed Z.

$$\int_{-\infty}^{\infty} f(z)dz = 1 \text{ by Theorem 6}$$

It follows that

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

$\square$

2 Appendix

2.1 Normal Approximation to Binomial Distribution

From tutorial 4, if $X \sim \text{binomial}(n, p)$, then $E(X) = np$ and $Var(X) = np(1-p)$. If n is large and p is not extreme (near 0 or 1), the distribution of X can be approximated by that of a normal random variable with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. Here n should be large so that there are enough discrete values of X to make an approximation by a continuous distribution reasonable; p should be "in the middle" so the binomial is nearly symmetric, as is the normal. A conservative rule to follow is that the approximation will be good if $\min(np, n(1-p)) \geq 5$.

In general, the normal approximation with the continuity correction is far superior to the approximation without the continuity correction. Please refer to the diagram drawn in class for intuition behind.

In summary, if $X \sim \text{binomial}(n, p)$ and $Y \sim N(np, np(1-p))$, then we approximate

$$P(X \leq x) \approx P(Y \leq x + 0.5),$$

$$P(X \geq x) \approx P(Y \geq x - 0.5).$$

2.2 The Central Limit Theorem

(Optional) Formal definition: Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables with $E(X_i) = \mu$ and $0 < Var(X_i) = \sigma^2 < \infty$ for $i = 1, 2, ...$ Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $G_n(x)$ denote the cdf of $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$. Then, for any x, $-\infty < x < \infty$,

$$\lim_{n \to \infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy;$$

that is, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ has a limiting standard normal distribution.

Intuitive version: If $X_1, ..., X_n$ have the same distribution (not necessarily normal) and sample size n is large enough (usually for $n \geq 30$), the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately distributed as $N(0, 1)$.

Starting from virtually no assumptions (other than independence and finite variances), we end up with normality! (Optional) The point here is that normality comes from sums of "small" (finite variance), independent disturbances. The assumption of finite variances is essentially necessary for convergence to normality.

(Optional) Interested students may search "law of large numbers", "convergence in distribution" and "Central Limit Theorem".